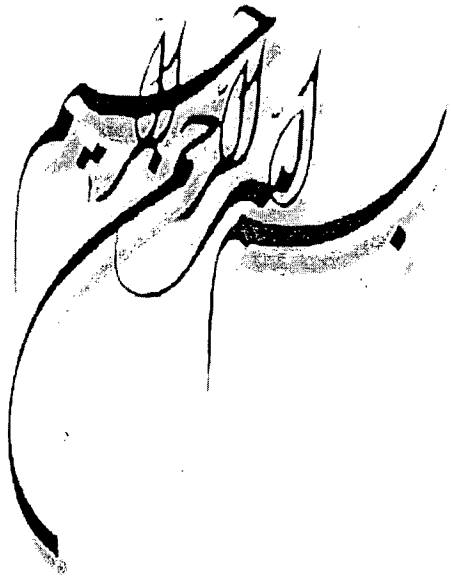
بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان نامه‌ی کارشناسی ارشد رشته‌ی آمار گرایش آمار ریاضی

مباحثی در تحلیل مدل‌های آمیخته: توزیع‌های نرمال – چوله و t– چوله

استاد راهنما:

دکتر ایرج کاظمی

پژوهشگر:

محبوبه خلفی

اسفند۱۳۸۸

دانشگاه اصفهان

دانشکده علوم

گروه آمار

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی آمارگرایش آمار ریاضی

خانم محبوبه خلفی

تحت عنوان

مباحثی در تحلیل مدلهای آمیخته :توزیعهای نرمال – چوله و t – چوله

در تاریخ**۱۲ /۸۸/۱۲** توسط هیأت داوران زیر بررسی با درجه **عالی** به تصویب نهایی رسید.

۱- استاد راهنمای اول پایان‌نامه دکتر ایرج کاظمی با مرتبه‌ی علمی استادیار امضاء

۲- استاد داور داخل گروه پایان‌نامه دکتر محمد بهرامی با مرتبه‌ی علمی استاد یار امضاء

۳- استاد داورخارج از گروه دکتر سعید پولاد ساز با مرتبه‌ی علمی دانشیار امضاء

امضای مدیر گروه

تقدیم و تشکر از استاد بزرگوار،

جناب آقای دکتر ایرج کاظمی

که راهنمایی‌هایشان در مسیر این پژوهش تمامی موانع را هموار ساخت

و حمایتشان باعث پیشرفت و ارتقاء علمی اینجانب شد،

قطعاً بی‌لطف و عنایت ایشان امکان‌پذیرنبود.

تقدیم به

پدر و مادر عزیزم،

که با نثار فداکاری‌ها و مهربانی‌هایشان به وسعت

دریاها و کوه‌ها

به من استقامت، انسانیت و دوست داشتن قلب‌ها

را آموختند.

تقدیم به همسر عزیزم،

صابر

که عشق ناتمامش، امید روزهای زندگی‌ام است.

چکیده

توزیع‌های آمیخته در بسیاری از تحقیقات کاربردی مانند ژنتیک، زیست شناسی و اقتصاد مورد توجه محققان قرار گرفته است. این توزیع‌ها در برازش مدل‌های رگرسیونی سهم بسزایی دارند. مطالعات فراوانی براساس توزیع‌های آمیخته صورت گرفته و روش‌های برآوردیابی مختلفی برای برآورد پارامترهای این توزیع‌ها مورد بررسی قرار گرفته است. در این رساله، توزیع‌های نرمال،t ، نرمال-چوله و t-چوله آمیخته را برای برازش خط رگرسیونی مورد استفاده قرار می‌دهیم. از آنجایی که استفاده از روش‌های مناسب برآوردیابی پارامترهای این مدل‌ها دارای اهمیت ویژه‌ای در حصول نتایج است؛ در این پایان‌نامه، روش‌های برآوردیابی را از دو دیدگاه فراوان‌گرا و بیزی مورد مطالعه قرار می‌دهیم. برآوردهای ماکسیمم درستنمایی با توجه به تعمیم‌هایی از الگوریتم EM و برآوردهای بیزی توسط نمونه‌گیرگیبز انجام می‌شود. با استفاده از روش‌های بیزی، برآورد تعداد مؤلفه‌های یک توزیع آمیخته را بیان و یک چارچوب کلی برای آن مطرح می‌کنیم. همچنین، ابتدا به بررسی تحقیقات انجام شده در کاربرد الگوریتم گیبز در تحلیل بیزی توزیع t با درجه‌آزادی ثابت پرداخته، آنگاه با در نظر گرفتن توزیع پیشین مناسب برای درجه آزادی توزیع‌های t و t-چوله، استنباط بیزی آن‌ها را توسعه می‌دهیم. علاوه برآن، یک چارچوب برای برآورد تعداد مؤلفه‌های این توزیع‌ها، به کمک الگوریتم پرش‌های معکوس‌پذیر پیشنهاد کرده‌ایم. در آخر، کاربرد توزیع‌های آمیخته را به عنوان توزیع جملات خطا در مدل رگرسیون چندگانه بررسی، مدل برتر را انتخاب و پارامترهای آن را برآورد کرده‌ایم.

کلید واژه: الگوریتم EM، نمونه‌گیرگیبز، توزیع نرمال-چوله آمیخته، توزیع t-چوله آمیخته، روش‌های بیزی، داده‌افزایی

فهرست مطالب

عنوان **صفحه**

صفحه

عنوان

صفحه

عنوان

فصل پنجم: استنباط بیزی توزیع‌های آمیخته با تعداد مؤلفه مجهول

عنوان صفحه

فهرست شکل‌ها

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱ موضوع تحقیق

توزیع‌های آمیخته به دلیل کاربرد زیاد در علوم مختلف مانند ژنتیک، زیست‌شناسی و اقتصاد، در چند دهه اخیر بسیار مورد توجه محققان قرار گرفته است. این توزیع‌ها در مبحث برازش مدل‌های رگرسیونی سهم بسزایی دارند.

توزیع آمیخته گوسین با قابلیت تقریب توابع چگالی پیوسته یک متغیره و انعطاف‌پذیری زیاد، از مشهورترین‌ها آنها است. با وجود سودمند بودن توزیع نرمال در این تحقیقات، گاه مشاهدات پرنفوذ در مجموعه داده‌ها منجر به خدشه‌دار شدن استنباط درباره پارامترهای مدل خواهند شد. یک مدل جایگزین معقول، استفاده از توزیع‌های دم سنگین‌تر از نرمال، مانند توزیع‌های آمیخته t است. توزیع t با داشتن دم‌های سنگین‌تر باعث تقلیل تعداد مؤلفه‌ها و در نتیجه تعداد پارامترها می‌گردد.

در بسیاری از مسائل کاربردی، به ویژه وقتی با مشاهداتی با توزیع نامتقارن سروکار داریم، استفاده از توزیع‌های نرمال آمیخته و t آمیخته در برازش داده‌ها مناسب نبوده و منجر به گمراهی در نتایج خواهند شد. به عنوان مثال، توزیع نرمال

آمیخته به منظور جبران چولگی موجود در داده‌هایی با توزیع نامتقارن، مؤلفه‌های اضافی (شبه مؤلفه‌ها[1]) را جایگزین می‌کند که منجر به ناکارآمدی و ایجاد پارامترها و مؤلفه‌های اضافی می‌گردد.

به عنوان توزیعی مقاوم[2]، خانواده جدید توزیع‌های نرمال–چوله معرفی شد که بر ضعف ذاتی موجود غلبه کرد. علاوه بر آن، یک ساختار مقاوم‌تر از توزیع نرمال–چوله، توزیع t–چوله نام داردکه خانواده توزیع نرمال، t–استیودنت و نرمال – چوله را در خود جای داده است. این توزیع، با داشتن دم‌های سنگین‌تر و چولگی زیاد معمولاً دارای قابلیت برازندگی بهتری نسبت به سایر توزیع‌های مطرح شده است.

۱-۲ پیشینه تحقیق

توزیع‌های آمیخته در بسیاری از تحقیقات کاربردی مورد استنباط قرار می‌گیرند. آنچه در این پایان‌نامه ارائه می‌شود، مبحث برآوردیابی در توزیع‌های نرمال آمیخته، t آمیخته، نرمال – چوله آمیخته و t – چوله آمیخته است. اغلب محاسبات پیچیده در برآوردیابی پارامترها در توزیع‌های آمیخته نشان داده است که روش‌های نوین عددی مانند الگوریتم ای ام (EM [3]) و مونت کارلوی زنجیر مارکفی (MCMC [4]) نیاز است.

دمپستر و همکاران (۱۹۷۷)، خواص جالبی از برآوردیابی پارامترهای توزیع نرمال آمیخته را توسط الگوریتم EM مطرح کردند. اما گاه وجود یک یا چند گروه مشاهدات دورافتاده استفاده از این توزیع‌ها را مختل نموده و برازش مدل مناسب به داده‌ها مرهون استفاده از روش‌های مقاوم‌تر است. در دهه‌های اخیر کاربرد توزیع‌های آمیخته t در استنباط پارامتری به عنوان یک مدل مقاوم بسیار مورد توجه محققان قرار گرفته است. پیل و مک‌لاچلان (۲۰۰۰) توزیع t آمیخته را به عنوان یک روش مقاوم معرفی کردند. در این مطالعات آنان با استفاده از الگوریتم EM که روش عددی برای یافتن برآوردهای ماکزیمم درستنمایی بخصوص در توزیع‌های آمیخته است، برازش مدلها را به روش داده‌افزایی انجام دادند. ونگ و همکاران (۲۰۰۴) پارامترهای توزیع t آمیخته را توسط الگوریتم EM برآورد کرده و یک روش برای تعیین درجات آزادی ارائه دادند.

---

[1] Pseudo – components
[2] Robust
[3] Expectation Maximization algorithm
[4] Markov Chain Monte Carlo

لین و همکاران (۲۰۰۷) با به کارگیری توزیع‌های نرمال- چوله که آزالینی در سال ۱۹۸۵ به عنوان کلاس جدیدی از توابع وابسته به پارامتر شکل معرفی کرده بود، تحلیل داده‌ها با ساختار توزیع‌های نامتقارن را معرفی کردند. دراین مطالعات تعدیل‌های مختلفی از الگوریتم EM، به عنوان روش‌های عددی، در یافتن برآوردهای ماکزیمم درستنمایی (ML [1]) معرفی شده است. همچنین، از دیدگاه بیزی استنباط پارامترها براساس روش MCMC انجام شده است. در تحقیقی دیگر، لین و همکاران (۲۰۰۷) روش‌های به کار رفته برای آمیخته‌های نرمال را بسط داده و یک چارچوب کلی برای توزیع t- چوله ، به منظور تحلیل داده‌هایی با توزیع دم سنگین ارائه دادند. علاوه بر آن، برآورد پارامترها را توسط تعمیم‌هایی از الگوریتمEM به عنوان روش‌های محاسباتی جدید انجام دادند.

یک روش نمونه‌گیری برای مدل آمیخته بیز با تعداد مؤلفه‌های مجهول در ونگ و فو (۲۰۰۷) آمده است. آنان با شبیه‌سازی پارامترها از توزیع‌های پسین شرطی کامل و استفاده از الگوریتم نمونه‌گیرگیبز استنباط بیزی پارامترها را انجام داده‌اند.

یک مسئله اساسی مطرح شده در بسیاری از تحقیقات در ارتباط با توزیع‌های آمیخته متناهی، تعیین تعداد مؤلفه‌های آن‌ها است.

اولین بار گرین (۱۹۹۵-۱۹۹۴) الگوریتم پرش‌های معکوس‌پذیر مونت کارلوی زنجیرمارکفی (RJMCMC [2]) را به عنوان روشی برای برآورد همزمان پارامترهای مدل و تعداد مؤلفه‌ها معرفی کرد. ایده گرین بر این اساس است که حرکت بین ابعاد مختلف پارامترها را توسط روش MCMC معرفی و ابعاد فضای پارامتر و نیز پارامترهای هر مؤلفه را برآورد کند.

ریچاردسون و گرین (۱۹۹۷) الگوریتم RJMCMC را جهت برآورد بیزی پارامترهای توزیع نرمال آمیخته با تعداد مؤلفه‌های نامعلوم به کار بردند. آنها الگوریتم را به حالت چند متغیره نیز توسعه دادند. استیفنز (۲۰۰۰) روش دیگری را بر اساس امکان تغییر فضای پارامتر ارائه داد. او با طرح فرآیند زاد و مرگ مونت کارلوی زنجیر مارکفی (BDMCMC [3]) این موضوع را برای مشاهداتی با ابعاد زیاد نیز گسترش داد. از طرفی دیگر، ونگ و همکاران (۲۰۰۴) یک الگوریتم RJMCMC را برای خانواده‌ای خاص از آمیخته‌های نرمال چند متغیره معرفی کردند، که در آن تمامی ماتریس‌های کواریانس موجود دارای مقادیر ویژه یکسانی بودند. در بسیاری از تحقیقات اخیر، استنباط در

---

[1] Maximum Likelihood
[2] Reversible Jump Markov Chain Monte Carlo
[3] Birth and Death Markov Chain Monte Carlo

مورد تعداد مؤلفه‌های توزیع‌های آمیخته هدف اصلی بوده است و یا حداقل به عنوان اولین گام در تحلیل این روش به کار رفته است (چانگ، ۲۰۰۵). هو(۲۰۰۵) بسط دیگری برای توزیع‌های آمیخته نرمال بیان کرده اند که درآن با تغییری مفید در الگوریتم RJMCMC برازش مدل‌های خطی با اثرات آمیخته را مطرح شده است.
کاپ و.همکاران (۲۰۰۳) مقایسه‌ای بین دو روش RJMCMC و BDMCMC انجام دادند. آنها در این تحقیقات به این نتیجه رسیدند که یک الگوریتم RJMCMC که تنها شامل گام‌های زاد و مرگ الگوریتم BDMCMC است، می‌تواند کاراترین الگوریتم باشد.

۱-۳ محاسبات آماری

تحلیل و برازش مدل‌ها با استفاده از نرم افزارهای SAS و OpenBugs (WinBugs) انجام شده است. نرم افزار OpenBugs به همراه مثال‌های متنوع از سایت http://www.mrc.bsu.com.ac.uk/bugs قابل دانلود می‌باشد.
نسخه مورد استفاده از این دو نرم افزار OpenBugs 3.03 و SAS 9.1 است. در فصل ۶ برآورد پارامترهای مدل رگرسیون خطی چندگانه به روش بیزی توسط نرم‌افزار OpenBugs و به روش ماکزیمم درستنمایی از طریق نرم‌افزار SAS و به کمک روش Proc nlmixed صورت گرفته است.

۱-۴ اهمیت و کاربرد نتیجه‌های تحقیق

توزیع‌های آمیخته در بسیاری از تحقیقات به خصوص در ژنتیک ، زیست‌شناسی و اقتصاد به کار برده می‌شوند. از آنجا که روش‌های معمول برآوردیابی، یافتن مدل مناسب و بسیاری مسائل دیگر در این توزیع‌ها به سهولت انجام‌پذیر نیست، این پایان‌نامه به بررسی این توزیع‌ها و مسائل مربوط به آنها می‌پردازد.
یک مسئله مهم در تحلیل داده‌ها، روش مناسب برآوردیابی است به طوری که برآوردها خواص مطلوب و در عین حال راه حلی صریح داشته باشند. این پایان‌نامه به بررسی این روش‌ها در توزیع‌های آمیخته پرداخته است. در اکثر مواقع الگوریتم گیبز روش ساده و معمول است. مسئله مهم دیگر در تحقیقات تجربی انتخاب مدل می‌باشد، به حدی

که گاه برازش مدلی نامناسب موجب گمراهی در نتایج خواهد شد. این تحقیق با در نظر گرفتن توزیع‌های متفاوت و در غالب مثال‌های تجربی با استفاده از معیارهای مناسب به انتخاب مدل برتر از بین توزیع‌های نامبرده می پردازد.

۱-۵ اهداف تحقیق

هدف از انجام این تحقیق، مطالعه‌ای بر مدل‌های آمیخته با مؤلفه‌های نرمال، t، نرمال- چوله و t- چوله و مقایسه آنها است. به منظور برآورد پارامترهای این مدل‌ها، از روش‌های مختلف ماکزیمم درستنمایی و الگوریتم EM استفاده شده است. علاوه بر آن، در این تحقیق از رهیافت‌های MCMC به عنوان روشی مناسب برای برآورد پارامترهای مدل‌ها در استنباط بیزی بهره گرفته شده است. با توجه به آنکه تعداد مؤلفه‌های توزیع آمیخته به عنوان موضوعی مهم در بسیاری از تحقیقات مطرح است، با استفاده از روش‌های بیزی به حل این موضوع پرداخته شده است. استنباط‌های آماری با در نظر گرفتن توزیع‌های آمیخته در مدل‌های خطی بسط داده شده و با گنجاندن چند مثال تجربی و برازش بر داده‌های واقعی، روش‌های عددی جهت برآورد پارامترهای مدل‌ها به کار برده شده است.

۱-۶ ساختار پایان‌نامه

ساختار این پایان‌نامه به شرح زیر است. در فصل ۲، به طور مختصر توزیع‌های آمیخته معرفی شده و الگوریتم EM، نمونه‌گیرگیبز و روش‌های MCMC، الگوریتم پرش‌های معکوس‌پذیر و فرآیند زاد و مرگ توضیح داده شده است. در فصل ۳، با فرض آنکه تعداد مؤلفه‌ها معلوم هستند، ویژگی‌های توزیع‌های نرمال آمیخته و t آمیخته توسط الگوریتم EM و روش‌های MCMC بررسی شده است. در فصل ۴، مدل‌های آمیخته نرمال – چوله و t – چوله مورد مطالعه قرار گرفته است. در فصل ۵، کاربرد الگوریتم RJMCMC برای توزیع‌های آمیخته مطرح و الگوریتم زاد و مرگ به عنوان روشی دیگر معرفی شده است. در فصل ۶، نقش این توزیع‌ها در تحلیل داده‌های واقعی روشن‌تر خواهد شد.

فصل دوم

تعاریف و مفاهیم مقدماتی

۲-۱ مقدمه

در این فصل به اختصار مباحثی را که بسیار در بخش‌های مختلف این پایان‌نامه مورد استفاده قرار می‌گیرند، شرح می‌دهیم.

پیچیدگی محاسبات در روند برآوردیابی پارامترهای توزیع‌های آمیخته، نیاز به استفاده از روش‌های پیشرفته بهینه‌سازی توابع را دوچندان کرده است. عموماً روش‌های عددی تکراری مانند الگوریتم EM (دمپستروهمکاران، ۱۹۷۷) و روش MCMC (دیابلت و رابرت، ۱۹۹۰) جهت برآورد پارامترها استفاده می‌شود. هر یک از این الگوریتم‌ها با توجه به کاربردی که در حل مسائل دارند، در انواع مختلفی طراحی شده‌اند. این الگوریتم‌ها معمولاً زمانی مفید هستند که تعداد مؤلفه‌های توزیع‌های آمیخته معلوم باشد.

یک موضوع اساسی در بسیاری از تحقیقات مرتبط با توزیع‌های آمیخته متناهی، تعیین تعداد مؤلفه‌های آن است. الگوریتم‌هایی مانند الگوریتم پرش‌های معکوس‌پذیر مونت کارلوی زنجیر مارکفی (ریچاردسون و گرین، ۱۹۹۵) و

الگوریتم زاد و مرگ مونت کارلوی زنجیر مارکفی (استیفنز، ۲۰۰۰) به عنوان روش‌های بیزی در برآورد تعداد مؤلفه‌ها همزمان با پارامترهای توزیع آمیخته، بسیار کاربرد دارند.

ساختار فصل حاضر به این صورت است. در بخش ۲-۲، توزیع‌های آمیخته متناهی معرفی می‌شوند. در بخش ۲-۳، الگوریتم EM شرح داده می‌شود. در بخش ۲-۴، الگوریتم MCMC را به عنوان یک روش بیزی در برآوردیابی مطرح می‌کنیم. در بخش ۲-۵، یک چارچوب کلی برای تعیین تعداد مؤلفه‌های یک توزیع آمیخته با استفاده از الگوریتم RJMCMC و الگوریتم زاد و مرگ ارائه می‌شود.

۲-۲ توزیع‌های آمیخته متناهی

تابع چگالی آمیخته متناهی از تابع چگالی‌های $f_1, \cdots, f_k$ ترکیبی محدب به صورت

$$f(y_j) = \sum_{i=1}^{k} \pi_i f_i(y_j), \qquad (۲-۱)$$

است که در آن $\sum_{i=1}^{k} \pi_i = 1$ و $k > 1$ با فرض این که f_i متعلق به یک خانواده از توزیع‌ها با پارامتر(های) نامعلوم θ_i باشد، یک مدل آمیخته پارامتری به صورت

$$f(y_j|\Theta) = \sum_{i=1}^{k} \pi_i f_i(y_j|\theta_i), \quad \sum_{i=1}^{k} \pi_i, \quad k > 1, \qquad (۲-۲)$$

تعریف می‌شود، که در آن $\Theta = (\theta_1, \cdots, \theta_k)'$ پارامترهای مدل است. در حالت خاص، اگر برای هر $i = 1, \cdots, k$ $f_i(y_j|\theta_i)$ تابع چگالی نرمال و θ_i شامل میانگین و واریانس این توزیع باشد، آنگاه شکل‌های متفاوتی براساس kهای مختلف به دست می‌آید (شکل ۲-۱).

با توجه به عبارت (۲-۲) گشتاور مرتبه mام مربوط به توزیع آمیخته از رابطه

$$E[Y^m] = \sum_{i=1}^{k} \pi_i E_{f_i}[Y^m]$$

به دست می‌آید. در این حالت همان طور که ملاحظه می‌شود، کافی است تنها گشتاور f_iها معلوم باشند.

توزیع‌های آمیخته که به صورت (۲-۲) تعریف می‌شوند، نسبت به برآوردگرهای ماکسیمم درستنمایی (در صورت وجود) و برآوردگرهای بیز بسیار آسیب‌پذیر هستند. برای درک این مطلب، یک نمونه تصادفی nتایی در نظر بگیرید. تابع درستنمایی (۲-۲) به صورت

$$L(\Theta, \pi|y) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(y_i|\theta_j) \qquad (۲-۳)$$