
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

تحلیل بیزی رگرسیون چگالی برای پاسخ‌های گسسته

نگارنده: حسن محمدی

استاد راهنما

دکتر حسین باغیشنی

بهمن ۱۳۹۶

شماره:
تاریخ:

باسمه‌تعالی

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

فرم شماره (۳): صورت‌جلسه نهایی دفاع از پایان‌نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان‌نامه کارشناسی ارشد آقای **حسن محمدی** با شماره دانشجویی **۹۴۳۸۹۶۴** رشته **آمار** گرایش **آمار ریاضی** تحت عنوان **تحلیل بیزی رگرسیون چگالی برای پاسخ‌های گسسته** که در تاریخ **۱۳۹۶/۱۱/۱۰** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: خیلی خوب.....) ☑ **مردود** ☐

نوع تحقیق: **نظری** ☐ **عملی** ☐

عضو هیات داوران	نام و نام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استاد راهنمای اول	دکتر حسین باغیشنی	استادیار	
۲- استاد راهنمای دوم	----------	----------	
۳- استاد مشاور	----------	----------	
۴- نماینده تحصیلات تکمیلی	دکتر محمدرضا ربیعی	استادیار	
۵- استاد ممتحن اول	دکتر محمد آرشی	دانشیار	
۶- استاد ممتحن دوم	دکتر احمد نزاکتی	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده:

وزارت علوم، تحقیقات و فناوری

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می‌تواند از پایان‌نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به

تمام کسانی که علم و دانش را سرلوحه زندگی خود قرار دادند.

نه ثروت، نه قدرت و نه هیچ چیز دیگر...

سپاس گزاری...

سپاس خدای را که سخنوران در ستودن او بمانند، شمارندگان شمردن نعمت‌های او ندانند و کوشندگان حق او را گزاردن نتوانند.
بدون شک جایگاه و منزلت معلم بالاتر از آن است که در مقام قدردانی از زحمات آن ها، با زبان قاصر و دست ناتوان چیزی بنگاریم؛ اما بر حسب وظیفه بر خود واجب می دانم از استاد شایسته و فرهیخته جناب آقای دکتر حسین باغیشنی که در کمال سعه صدر، با حسن خلق و فروتنی از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این پایان نامه را بر عهده گرفتند، کمال تشکر و قدردانی را به جا آورم. همچنین از اساتید داور جناب آقای دکتر محمد آرشی و جناب آقای دکتر احمد نزاکتی و تمامی اساتید فرهیخته گروه آمار دانشگاه صنعتی شاهرود تشکر و قدردانی می نمایم. در انتها وظیفه خود می دانم از پدر و مادر دلسوز و فداکارم، برادران و خواهران عزیز و مهربانم و تمامی دوستانی که در طول دوران تحصیل لحظات تکرار نشدنی را در کنارشان گذراندم صمیمانه سپاس گزاری کنم.
باشد که این خرد ترین، بخشی از زحمات آنان را سپاس گوید.

حسن محمدی
بهمن ۱۳۹۶

تعهد نامه

اینجانب **حسن محمدی** دانشجوی کارشناسی ارشد رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **تحلیل بیزی رگرسیون چگالی برای پاسخ‌های گسسته** ، تحت راهنمایی **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارائه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

حسن محمدی
بهمن ۱۳۹۶

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

در این پایان‌نامه مدل بیزی رگرسیون چگالی برای پاسخ‌های گسسته معرفی می‌شود. استفاده از مدل‌های آمیخته رهیافتی مناسب برای برآورد چگالی شرطی را فراهم می‌سازد. برای برآورد تابع چگالی چندمتغیره در مدل‌های آمیخته، متغیرهای گسسته شامل متغیر پاسخ و متغیرهای تبیینی، به صورت متغیرهای پیوسته پنهان که در نقاط برش گسسته‌سازی شده‌اند به‌کار برده می‌شوند. مدل‌های متفاوتی جهت محاسبه نقاط برش، برای پاسخ‌های شمارشی با ویژگی‌های پیش‌پراکنش یا کم‌پراکنش، معرفی و مقایسه می‌شوند. همچنین مدل آمیخته فرآیند دیریکله با هسته گاوسی چندمتغیره برای مشاهدات و متغیرهای پیوسته پنهان به‌کار برده می‌شود. با توجه به پیچیده بودن توزیع پسین از الگوریتم‌های مونت کارلوی زنجیر مارکوفی برای نمونه‌گیری از توزیع پسین استفاده می‌شود. همچنین عملکرد مدل با استفاده از داده‌های شبیه‌سازی‌شده و واقعی ارزیابی می‌شود.

کلمات کلیدی: فرآیند دیریکله، متغیرهای پنهان، بیش‌پراکنش، کم‌پراکنش، توزیع پسین، الگوریتم مونت کارلوی زنجیر مارکوفی.

پیش‌گفتار

مدل‌های رگرسیونی در موقعیت‌های کاربردی بسیاری از علوم از اهمیت ویژه‌ای برخوردار هستند. هدف اصلی در مدل‌های رگرسیونی، بررسی تأثیر برداری از متغیرهای تبیینی $\mathbf{x}$ بر روی یک متغیر پاسخ y است. معمولاً بنای مدل‌های رگرسیونی بر اساس یکی از شاخص‌های توزیع شرطی $y|\mathbf{x}$ پی‌ریزی می‌شود. برای مثال، تابع رگرسیون خطی ساده، میانگین توزیع شرطی پاسخ به شرط متغیرهای تبیینی است؛ یا در رگرسیون چندکی، چندک‌های توزیع شرطی مورد نظر به‌عنوان تابعی از متغیرهای تبیینی مدل‌بندی می‌شود. اخیراً یک مدل جدید رگرسیونی مطرح شده‌است که مدل‌بندی کل تابع چگالی را مدنظر قرار داده و تمام شاخص‌های توزیع را پوشش می‌دهد. در این پایان‌نامه، رگرسیون تابع چگالی شرطی، زمانی که متغیر پاسخ گسسته است در نظر گرفته می‌شود. تحلیل این مدل رگرسیونی بر پایه دیدگاه بیزی صورت گرفته و از روشهای بیزی ناپارامتری استفاده می‌شود. با توجه به این مقدمه ساختار پایان‌نامه به‌صورت زیر است:

- در فصل اول، مفاهیم و تعاریف مورد نیاز برای ورود به موضوع اصلی پایان‌نامه را مطرح می‌کنیم.
- در فصل دوم، استنباط بیزی ناپارامتری را شرح می‌دهیم.
- در فصل سوم، به استنباط بیزی مدل رگرسیون چگالی و نمونه‌گیری از توزیع پسین آن توسط الگوریتم‌های مونت کارلوی زنجیر مارکوفی می‌پردازیم.
- در فصل چهارم، با مطالعه شبیه‌سازی، عملکرد مدل را مورد ارزیابی قرار می‌دهیم. همچنین کاربرد مدل را در یک مثال واقعی نمایش خواهیم داد.
- پیوست آ نیز شامل بخشی از کدهای نوشته‌شده در محیط R برای بازتولید مثال‌های موجود در پایان‌نامه است.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مفاهیم و تعاریف پایه‌ای

۱.۱ مقدمه

آمار علمی است که به گردآوری، سازمان‌دهی، تحلیل و استخراج نتایج حاصل از داده‌ها می‌پردازد. هدف آمار توصیفی، سازمان‌دهی و طبقه‌بندی مجموعه داده‌ها و هدف اصلی استنباط آماری، تحلیل و استخراج نتایج حاصل از داده‌ها است. هدف اصلی این پایان‌نامه، استنباط آماری حاصل از داده‌ها است.

دامنه تحلیل‌های آماری دربردارنده تحلیل‌های یک تا چند متغیره است. علم آمار شامل روش‌های مختلف برای تعیین رابطه بین متغیرها است؛ که تحلیل مدل‌های رگرسیونی، از جمله مهم‌ترین و پرکاربردترین روش‌ها برای رسیدن به این هدف هستند. تحلیل‌های رگرسیونی در علوم متعددی مانند مهندسی، فیزیک، اقتصاد، مدیریت، علوم زیستی و علوم اجتماعی به‌کار گرفته می‌شوند. اصلی‌ترین هدف در تحلیل رگرسیون، تعیین بهترین مدل به‌صورت

$$y = f(\mathbf{x}) + \epsilon$$

است که بتواند ارتباط بین یک متغیر پاسخ y، با یک یا چند متغیر تبیینی $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$، را تعیین کند. به‌طور معمول فرض می‌شود جمله خطای ϵ دارای توزیع نرمال است که با فرض تصادفی نبودن $\mathbf{x}$، متغیر پاسخ y نیز دارای توزیع نرمال خواهد بود. در کاربردهای مختلفی

ممکن است متغیر پاسخ نرمال نباشد. به‌عنوان مثال، متغیرهای پاسخ رسته‌ای[۱] یا حضور داده‌های پرت در مشاهدات، پذیره نرمال بودن توزیع پاسخ را ضعیف می‌کند. در این موارد میانگین پاسخ را نمی‌توان به صورت مستقیم، به‌عنوان تابعی از متغیرهای تبیینی، مدل‌سازی نمود، بلکه تبدیلی از آن با استفاده از تابعی به‌نام تابع پیوند[۲] مورد استفاده قرار می‌گیرد. در این موارد مدلی که به داده‌ها برازش می‌شود، عضوی از خانواده مدل‌های خطی تعمیم‌یافته[۳] (GLM) است (مک‌کلا و نلدر، ۱۹۸۹). یک حالت خاص از این مدل‌ها، مدل رگرسیون خطی است (جکمن، ۱۹۸۹).

در برخی از موارد، متغیر پاسخ از نوع رسته‌ای دو سطحی[۴] با سطوح موفقیت (۱) و شکست (۰) است. چنین پاسخ‌هایی در کاربردهای متعددی ظاهر می‌شوند. به‌عنوان نمونه، در علوم بانکداری مدیران به دنبال شناخت نرخ بازپس‌دهی وام‌های واگذارشده به مشتریان برای برنامه‌ریزی‌های آینده هستند. در این مثال شناخت مشتریان خوش‌حساب، به‌عنوان سطح موفقیت در نظر گرفته می‌شود. به‌عنوان مثال دیگر، در علم پزشکی، محققین به دنبال کشف بیماران مبتلا به سرطان در یک منطقه خاص هستند. معمولاً برای تحلیل این نوع پاسخ‌ها از مدل‌های رگرسیونی پرابیت[۵] و لجستیک[۶] استفاده می‌شود.

معمولاً بنای مدل‌های رگرسیونی بر اساس یک شاخص خاص از توزیع شرطی $y|\mathbf{x}$ پی‌ریزی می‌شود. برای مثال، تابع رگرسیونی خطی ساده، میانگین توزیع شرطی پاسخ به شرط متغیرهای تبیینی است. یا در رگرسیون چندکی[۷]، چندک‌های توزیع شرطی مورد نظر به‌عنوان تابعی از متغیرهای تبیینی مدل‌بندی می‌شود. اخیرا یک مدل جدید رگرسیونی مطرح شده است که مدل بندی کل تابع چگالی را مد نظر قرار می‌دهد. این نوع مدل رگرسیونی که به رگرسیون چگالی[۸] معروف شده است، در مقایسه با مدل‌های معمول رگرسیونی، بسیار منعطف‌تر عمل کرده و تمام شاخص‌ها را پوشش می‌دهد.

هدف اصلی در رگرسیون چگالی برآورد کردن چگالی شرطی پاسخ y به شرط بردار متغیرهای تبیینی $\mathbf{x}$، یعنی همان $f(y|\mathbf{x})$ است. با برآورد این چگالی شرطی مقادیر مورد علاقه دیگر مانند میانگین شرطی، میانه و چندک‌ها نیز به‌دست خواهند آمد. بنابراین مسئله اصلی پیش رو

[۱] Categorical
[۲] Link Fanction
[۳] Generalized Linear Models
[۴] Binary
[۵] Probit Regression
[۶] Logistic Regression
[۷] Quantil Regression
[۸] Density Regression

برآورد چگالی شرطی $f(y|\mathbf{x})$ است. استفاده از مدل‌های آمیخته[۹] رهیافتی بسیار مطلوب برای برآورد چگالی شرطی $f(y|\mathbf{x})$ را فراهم می‌آورد. با استفاده از این رهیافت، چگالی شرطی را می‌توان به صورت زیر نوشت:

$$f_p(y|\mathbf{x}) = \int k(y, \mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}P(\theta)$$

که در آن $k(\cdot, \cdot; \cdot)$ یک تابع هسته[۱۰] و $P(\cdot)$ یک اندازه احتمال است. برای بررسی موضوع از دیدگاه بیزی، نیازمند تعریف توزیع پیشین مناسب برای اندازه احتمال مورد نظر و نیز پارامترهای تابع هسته هستیم که در فصل‌های بعد معرفی می‌شوند.

مبنای اصلی استنباط آماری در دیدگاه بیزی، توزیع پسین[۱۱] است و می‌تواند انتخابی مناسب برای استنباط مدل رگرسیون چگالی باشد. توزیع پسین در یک مدل بیزی، به‌روز شده باور و دیدگاه اولیه در مورد پارامتر با توجه به مشاهدات است. در این روش، اولین مرحله این است که توزیع پیشین[۱۲] را مشخص کنیم. با ورود اطلاعات جدید، توزیع پیشین به توزیع پسین تبدیل می‌شود. ابزار اساسی این تبدیل، قضیه بیز[۱۳] است (گلمن و همکاران، ۲۰۰۳). تحلیل بیزی نیاز به حل انتگرالی دارد که معمولاً به راحتی نمی‌توان آن را به‌دست آورد. در این صورت می‌توان از روش‌های مبتنی بر شبیه‌سازی مونت کارلویی[۱۴] برای تقریب آن استفاده کرد. همچنین، از آن‌جایی که نمونه‌گیری مستقیم از توزیع پسین نیز به سادگی ممکن نیست، می‌توان از روش‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی[۱۵] (MCMC) استفاده کرد. روش‌های MCMC شامل روش‌هایی است که برای انتگرال‌گیری به روش مونت کارلویی، اغلب در مسائل مربوط به استنباط بیزی، به‌کار برده می‌شوند. با استفاده از این روش‌ها می‌توان از توزیع پسین نمونه تولید کرد (برانسکام و همکاران، ۲۰۰۷). از جمله روش‌های MCMC می‌توان به الگوریم متروپلیس ـ هستینگز[۱۶] (متروپلیس و همکاران، ۱۹۵۳ و هستینگز، ۱۹۷۰) و همچنین الگوریتم نمونه‌گیر گیبز[۱۷] (گمان و گمان، ۱۹۸۴ و گلفاند و اسمیت، ۱۹۹۰) اشاره کرد که به‌عنوان ابزار بسیار مفیدی برای تحلیل مدل‌های آماری پیچیده به‌وجود آمده‌اند.

---

[۹] Mixture Models
[۱۰] Kernel
[۱۱] Posterior Distribution
[۱۲] Prior Distribution
[۱۳] Bayes Theorem
[۱۴] Monte Carlo Simulation
[۱۵] Markov Chain Monte Carlo
[۱۶] Metropolis-Hastings Algorithm
[۱۷] Gibbs Sampler

۲.۱ مدل‌های رگرسیونی

موضوع رگرسیون نخستین بار در سال ۱۸۸۵ توسط گالتن مطرح گردید. او می‌خواست رابطه‌ای بین قد والدین و فرزندانشان بیابد. اعتقاد او بر این بود که والدین قدبلند معمولاً فرزندان قدبلندی خواهند داشت و والدین کوتاه‌قد صاحب فرزندان کوتاه‌قدتری هستند. او این مطلب را به‌صورت یک مدل‌بندی ریاضی مطرح کرد. برای نشان دادن رابطه بین دو متغیر، می‌توان از ضریب همبستگی نیز استفاده کرد. اما باید این نکته را توجه کرد که ضریب همبستگی نمی‌تواند اثر یک متغیر بر متغیر دیگر را نشان دهد. از طرفی ضریب همبستگی نمی‌تواند تخمینی از تغییر در یک متغیر را با تغییر در یک متغیر دیگر ارائه دهد. بنابراین، برای پاسخ‌گویی به موارد فوق می‌توان از مدل رگرسیونی استفاده کرد (گلمن و همکاران، ۲۰۰۶). تحلیل رگرسیونی کاربردهای وسیعی دارد. امروزه می‌توان از آن در علوم مختلفی مانند: اقتصاد، بازرگانی، هواشناسی، مهندسی و علوم اجتماعی بهره برد.

در حالت کلی یک مدل رگرسیونی به‌صورت زیر نمایش داده می‌شود:

$$y = f(\mathbf{x}) + \epsilon$$

که در آن y متغیر پاسخ نامیده می‌شود و از سایر متغیرها تأثیر می‌پذیرد. متغیر یا متغیرهایی که بر پاسخ اثر می‌گذارند، یعنی $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$، متغیرهای تبیینی (توضیحی یا پیش‌بین) نام دارند. مؤلفه ϵ نیز خطای مدل می‌باشد.

مدل رگرسیون خطی، ساده‌ترین مدل رگرسیونی است. در رگرسیون خطی به دنبال رابطه خطی متغیر پاسخ y با متغیرهای تبیینی $x_1, x_2, \ldots, x_p$ هستیم. الگوی رگرسیون خطی را به صورت زیر نمایش می‌دهند:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \qquad (۱.۱)$$

که در آن $\beta_0, \beta_1, \ldots, \beta_p$ ضرایب مدل رگرسیونی هستند. این ضرایب نامعلوم بوده و باید مستقیم برآورد شوند. متغیر تصادفی ϵ جمله خطای این معادله است که نوسانات تصادفی، خطای اندازه‌گیری یا اثر عوامل خارج از کنترل را نشان می‌دهد. معادله (۱.۱) را به این دلیل خطی گویند که در آن میانگین پاسخ، یک تابع خطی از پارامترهای نامعلوم $\beta_0, \beta_1, \ldots, \beta_p$ است. معمولاً مدل‌های رگرسیونی برای

- پیش‌گویی
- توصیف یا تفسیر داده‌ها
- کنترل پاسخ

به‌کار می‌روند.

برای مدل‌های رگرسیونی پذیره‌های بنیادی در نظر می‌گیرند که از طریق خطاها، قابل آزمون می‌باشند. این پذیرها به‌صورت زیر هستند:

- مؤلفه تصادفی خطا دارای میانگین صفر است؛ یعنی برای همه مشاهدات $i = ۱, ۲, \ldots, n$، $E(\epsilon_i) = ۰$ است.
- واریانس مؤلفه خطاها، ثابت می‌باشد؛ یعنی برای $i = ۱, ۲, \ldots, n$، $Var(\epsilon_i) = \sigma^۲$ است.
- خطاها ناهمبسته‌اند.
- مؤلفه تصادفی خطا، از متغیرهای تبیینی x مستقل است.

بدون بررسی این پذیره‌ها، هیچ‌گاه نباید مدل‌های رگرسیونی را به‌کار برد. زیرا در صورت رعایت نکردن این پذیره‌ها اطلاعات به‌دست آمده از رگرسیون، گمراه‌کننده خواهند بود. می‌توان برآورد پارامترها در این مدل را به روش‌های کمترین توان‌های دوم یا ماکسیمم درستنمایی به‌دست آورد (نیرومند، ۱۳۸۷). حالت‌های دیگری وجود دارند که امکان استفاده از رگرسیون خطی فراهم نیست. مدل‌های خطی تعمیم‌یافته رده بزرگتری از مدل‌های خطی است که رهیافت مناسبی برای این حالت‌ها فراهم می‌سازد.

۳.۱ مدل‌های خطی تعمیم‌یافته

در روش‌های استنباطی مربوط به الگوهای رگرسیون خطی و غیرخطی[۱۸]، یکی از پذیره‌های معمول این است که متغیر پاسخ دارای توزیع نرمال است. اما حالت‌هایی وجود دارند که پذیره نرمال بودن متغیر پاسخ، برقرار نیست. به‌عنوان نمونه، زمانی که متغیر پاسخ دوسطحی یا شمارشی است، پذیره نرمال بودن پاسخ، کاملاً غیر واقعی به نظر می‌رسد. در چنین مواقعی، استفاده از مدل‌های خطی مناسب نیست. به‌علاوه، مدل‌های خطی محدودیتی را که در میانگین پاسخ برای این گونه داده‌ها وجود دارد، در نظر نمی‌گیرند.

برای مثال، در توزیع پواسن که معمولاً برای داده‌های شمارشی به‌کار می‌رود، میانگین توزیع پاسخ باید نامنفی باشد. اگر در این حالت از مدل خطی استفاده کنیم، ممکن است برآورد میانگین پاسخ منفی شود؛ در نتیجه، برای این گونه از داده‌ها، عموماً تبدیلی به‌نام تابع پیوند لگاریتمی را به‌کار می‌برند. رده مدل‌های خطی تعمیم‌یافته، تعمیمی از مدل‌های خطی هستند که توسط نلدر و ودربرن (۱۹۷۲) معرفی شد. اساس این تعمیم مبتنی بر دو جنبه است به‌طوری که

[۱۸] Nonlinear

- متغیرهای پاسخ، توزیعی غیر از توزیع نرمال می‌توانند داشته باشند. باید به این نکته توجه داشت که توزیع در نظر گرفته‌شده، باید متعلق به خانواده توزیع‌های نمایی[۱۹] باشد.
- میانگین پاسخ به‌صورت مستقیم مدل‌سازی نمی‌شود، بلکه تبدیلی از آن با استفاده از تابع پیوند مدل‌سازی می‌شود.

در ادامه، نخست به معرفی خانواده توزیع‌های نمایی پرداخته و سپس مدل‌های خطی تعمیم‌یافته را بیان می‌کنیم.

۱.۳.۱ خانواده توزیع‌های نمایی

خانواده توزیع‌های نمایی، بخش مهمی در مدل‌های خطی تعمیم‌یافته به‌شمار می‌رود. این خانواده توزیع‌های مهمی مانند نمایی، نرمال، هندسی، پواسن، گاما، دوجمله‌ای منفی و نرمال وارون را شامل می‌شود. تابع چگالی خانواده توزیع‌های نمایی، به‌صورت زیر است:

$$f(y;\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y;\phi)\right\}$$

که در رابطه، $a(\cdot)$، $b(\cdot)$ و $c(\cdot)$ توابعی مشخص هستند. پارامتر θ پارامتر مکانی[۲۰] است و پارامتر ϕ را اغلب پارامتر پراکندگی[۲۱] می‌نامند. در این خانواده میانگین و واریانس به‌صورت زیر هستند (بیکل و داکسوم، ۲۰۰۱):

$$E(Y)=\mu=\frac{db(\theta)}{d\theta}=b'(\theta)$$

و

$$Var(Y)=\frac{d^2b(\theta)}{d\theta^2}a(\phi)=b''(\theta)a(\phi).$$

به‌عنوان مثال، توزیع پواسن عضوی از خانواده توزیع‌های نمایی است. همان‌طور که می‌دانیم تابع احتمال توزیع پواسن به‌صورت زیر است:

$$f(y;\mu)=\frac{e^{-\mu}\mu^y}{y!}.$$

اگر بخواهیم تابع احتمال فوق را به‌صورت خانواده توزیع‌های نمایی بنویسیم داریم:

$$f(y;\mu)=\exp\left\{y\ \ln\ \mu-\mu-\ln\ y!\right\}.$$

بنابراین با انتخاب‌های $\theta=\ln\mu$، $b(\theta)=e^\theta$، $a(\phi)=1$ و $c(y;\phi)=\ln(y!)$ نتیجه به‌دست می‌آید.

[۱۹] Exponential Family of Distributions
[۲۰] Location Parameter
[۲۱] Dispersion Parameter

۲.۳.۱ مؤلفه‌های مدل‌های خطی تعمیم‌یافته

مدل‌های خطی تعمیم‌یافته سه مؤلفه اساسی دارند:

- مؤلفه اصلی: این مؤلفه دارای یک پیش‌گوی خطی[۲۲] است که شامل متغیرهای تبیینی $\mathbf{x}$ است و می‌توان آن را به‌صورت زیر نوشت:

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

- مؤلفه تصادفی: این مؤلفه، توزیع پاسخ Y را مشخص می‌کند. نکته مهم این‌که این مؤلفه باید عضوی از خانواده توزیع‌های نمایی باشد.
- تابع پیوند: این مؤلفه، پیش‌گوی خطی را به میانگین متغیر پاسخ مرتبط می‌سازد. تابع پیوند یک تابع یکنوای مشتق‌پذیر است، به‌طوری که

$$\eta_i = g(E(Y_i|\mathbf{x_i})) = \mathbf{x_i}'\boldsymbol{\beta}$$

که در آن، تابع $g(\cdot)$، تابع پیوند است. چون تابع پیوند معکوس‌پذیر نیز هست، بنابراین می‌توان آن را به‌صورت زیر نوشت:

$$E\left[Y_i|\mathbf{x_i}\right] = g^{-1}(\eta_i) = g^{-1}(\mathbf{x_i}'\boldsymbol{\beta}).$$

تابع پیوند انواع گوناگونی دارد که از جمله آن‌ها می‌توان به توابع پیوند زیر اشاره کرد:

- تابع پیوند همانی[۲۳]: $g(t) = t$
- تابع پیوند لگاریتمی: $g(t) = \log(t)$
- تابع پیوند لجیت[۲۴]: $g(t) = \log(\frac{t}{1-t})$
- تابع پیوند پرابیت[۲۵]: $g(t) = \Phi^{-1}(t)$
- تابع پیوند Cloglog[۲۶]: $g(t) = \log(-\log(1-t))$

بسته به انتخاب نوع تابع پیوند $g(\cdot)$، مدل خطی تعمیم‌یافته می‌تواند یک الگوی غیرخطی را شامل شود. بنابراین، الگوهای خطی تعمیم‌یافته را می‌توان به‌عنوان یکسان‌سازی الگوهای خطی و غیرخطی به‌حساب آورد که خانواده غنی توزیع‌های پاسخ نرمال و غیرنرمال را با هم متحد می‌کند (نیرومند، ۱۳۸۴). اکنون به معرفی الگوهایی می‌پردازیم که برای متغیر پاسخ دوسطحی کاربرد بسیاری دارند.

---

[۲۲] Linear Predictor
[۲۳] Identity Link Function
[۲۴] Logit Link Function
[۲۵] Probit Link Function
[۲۶] Complementary Link Function

۳.۳.۱ رگرسیون پرابیت

رگرسیون پرابیت برای مدل‌بندی متغیرهای پاسخ دوسطحی با فرض این‌که تابع پیوند $g(\cdot)$ پرابیت است، به‌کار می‌رود. این مدل نسبت به مدل لجستیک دارای انعطاف‌پذیری کمتری است و به اندازه آن کاربرد ندارد.

هنگامی که متغیر پاسخ دارای توزیع برنولی با احتمال موفقیت π باشد و تابع پیوند، معکوس تابع توزیع تجمعی نرمال استاندارد، $\eta = \Phi^{-1}(\pi)$، اختیار شود، در این صورت مدل خطی تعمیم‌یافته، مدل پرابیت است. بنابراین در مدل پرابیت

$$E(Y|\mathbf{x}) = \pi(\mathbf{x}) = \Phi(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)$$

که در آن تابع $\Phi(\cdot)$ همان تابع توزیع نرمال استاندارد است (آلبرت و چیپ، ۱۹۹۳). در این مدل، پارامترها به روش درستنمایی ماکسیمم برآورد کرد.

۴.۳.۱ رگرسیون لجستیک

رگرسیون لجستیک نیز یکی از اعضای رده مدل‌های خطی تعمیم‌یافته است که از تابع پیوندی به‌نام تابع پیوند لجیت برای مدل‌بندی متغیرهای پاسخ دودویی(دوسطحی) با فرض مستقل بودن خطاها استفاده می‌کند. این تابع پیوند از توزیع لجستیک نتیجه می‌شود. اگر $F(\cdot)$ تابع توزیع لجستیک با پارامتر مکان μ و پارامتر مقیاس[۲۷] σ باشد، آنگاه

$$F(x) = \frac{1}{1 + \exp\left\{\frac{(x-\mu)}{\sigma}\right\}}.$$

بنابراین، زمانی که متغیر پاسخ دارای توزیع برنولی با احتمال موفقیت π باشد و تابع پیوند لجیت، $\eta = \log\frac{\pi}{1-\pi}$، اختیار شود، آن‌گاه مدل رگرسیونی حاصل، مدل رگرسیون لجستیک است. بنابراین

$$E(Y|\mathbf{X} = \mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}) = \frac{\exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}$$

و

$$P(Y = 0|\mathbf{X} = \mathbf{x}) = 1 - \pi(x) = \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}.$$

[۲۷] Scale Parameter

مدل رگرسیون لجستیک دارای محدودیت زیر است:

$$۰ \leq E(Y|\mathbf{X} = \mathbf{x}) \leq ۱.$$

در رگرسیون لجستیک ارتباط بین متغیر پاسخ با متغیرهای تبیینی، غیرخطی و شکل تابعی آن مانند S است که به‌طور یکنواخت افزایش (کاهش) می‌یابد. معمولاً برای برآورد پارامترها در این مدل، از روش درستنمایی ماکسیمم استفاده می‌شود (مونت گومری، ۲۰۱۵).

۵.۳.۱ رگرسیون پواسن

رگرسیون پواسون یکی دیگر از اعضای خانواده مدل‌های خطی تعمیم‌یافته است و برای تحلیل داده‌های شمارشی به‌کار می‌رود. اگر متغیر تصادفی Y دارای توزیع پواسن با پارامتر μ باشد، در این صورت

$$E(Y|\mathbf{x}) = \mu(\mathbf{x}).$$

با در نظر گرفتن تابع پیوند

$$g(\mathbf{x}) = \ln(\mathbf{x})$$

محدویت مثبت بودن میانگین پوشش داده شده و مدل رگرسیونی پواسن به صورت زیر نتیجه می‌شود:

$$E(Y|\mathbf{x}) = e^{\mathbf{x}'\boldsymbol{\beta}}.$$

۴.۱ مدل‌های رگرسیونی خارج از میانگین

۱.۴.۱ رگرسیون چندکی

همان‌طور که می‌دانید، رگرسیون خطی ضابطه بین میانگین شرطی یک متغیر پاسخ به شرط یک یا چند متغیر تبیینی را بیان می‌کند. گاهی اوقات رگرسیون خطی بر روی میانگین توزیع شرطی، عملکرد ضعیفی در تحلیل داده‌ها دارد. برای مثال در حالتی که توزیع خطا غیرنرمال است یا در صورتی که ناهمسانی واریانس وجود دارد، برآوردگرهای کمترین توان‌های دوم نسبت به داده‌های پرت حساس بوده و به برآوردهای اریب منجر می‌شوند. در این حالت‌ها می‌توان از رگرسیون چندکی استفاده کرد که این نوع مشکلات را مرتفع می‌کند.

رگرسیون چندکی یک مدل آماری با توانایی محاسبه و رسم منحنی‌های رگرسیونی متفاوت و منطبق با نقاط صدکی متفاوت توزیع شرطی $y|\mathrm{x}$ است که ضمن بیان تصویری کامل‌تر از داده‌ها امکان سنجش ارتباط متغیرهای تبیینی با چندک‌های مورد نظر متغیر پاسخ را بدون نیاز به نرمال بودن داده‌ها و حتی در حضور نقاط پرت فراهم می‌کند. به‌عبارت دیگر، این رگرسیون

نسبت به داده‌های پرت تنومند است. از طرف دیگر بر خلاف رگرسیون حداقل مربعات که روی میانگین شرطی متمرکز است، رگرسیون چندکی استراتژی منظمی را برای تعیین چگونگی تأثیر متغیرهای تبیینی روی مکان، مقیاس و شکل توزیع پیشنهاد می‌کند.

رگرسیون چندکی در علوم پزشکی کاربردهای فراوانی دارد. هدف اصلی در رگرسیون چندکی، ارائه مدلی است که امکان دخالت متغیرهای تبیینی نه تنها در مرکز داده‌ها، بلکه در تمام قسمت‌های توزیع به‌ویژه در دم‌های ابتدایی و انتهایی را فراهم می‌کند. در واقع بدون پذیره‌ها و محدودیت‌های رگرسیون خطی، مدلی را می‌توان ارائه کرد که نسبت به داده‌های پرت تنومند بوده و حتی اگر پذیره ناهمسانی واریانس‌ها وجود داشته باشد، این مدل برآورد مناسبی از ضرایب رگرسیون ارائه خواهد داد. به‌علاوه در رگرسیون خطی، با می‌نیمم کردن مجموع توان‌های دوم باقیمانده‌ها، برآورد پارامترها به‌دست می‌آید در حالی که در رگرسیون چندکی از می‌نیمم کردن نامتقارن قدر مطلق موزون باقیمانده‌ها برای برآورد پارامترهای مدل استفاده می‌شود.

۲.۴.۱ رگرسیون چگالی

همه مدل‌های رگرسیونی که تا به حال معرفی شدند یک شاخص خاص از توزیع شرطی $f(y|\mathbf{x})$ را مدنظر قرار می‌دهند. اما اگر هدف بررسی ویژگی‌هایی از مدل، مانند میانگین، میانه، چندک‌ها و واریانس باشد که کل شاخص‌های توزیع را پوشش دهد، مدل‌های رگرسیونی که تا به حال معرفی شده‌اند پاسخ‌گو نخواهند بود. اخیراً مفهوم جدیدی از رگرسیون به نام رگرسیون چگالی مطرح شده‌است که این مسئله را مرتفع نموده و تمام ویژگی‌های توزیع را پوشش می‌دهد. موضوع رگرسیون چگالی سابقه چندان طولانی نداشته و اخیراً توسط عده‌ای از آماردان‌های سرشناس در حال گسترش و به‌کارگیری است. برای مثال می‌توان به دانسون و همکاران (۲۰۰۷)، بایلی و همکاران (۲۰۰۹)، کانل و دانسون (۲۰۱۵) و پاپاجئورجیو و همکاران (۲۰۱۵) اشاره کرد.

هدف اصلی در این مدل، برآورد کردن تابع جرم یا چگالی احتمال شرطی متغیر پاسخ y به شرط بردار متغیرهای تبیینی $\mathbf{x}$، یعنی برآورد $f(y|\mathbf{x})$ است. این برآورد به ما اجازه می‌دهد تا چگونگی تأثیر متغیرهای تبیینی $\mathbf{x}$ بر روی متغیر پاسخ y را ملاحظه کنیم. استفاده از مدل‌های آمیخته به‌صورت زیر، یک رهیافت مناسب برای برآورد تابع چگالی شرطی $f(y|\mathbf{x})$ را فراهم می‌سازد:

$$f_P(y,\mathbf{x}) = \int k(y,\mathbf{x};\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta})$$

نکته دیگر که باید به آن توجه داشت این است که با برآورد کردن تابع چگالی شرطی $f(y|\mathbf{x})$، مقادیر دیگری مانند میانگین شرطی، میانه و چندک‌ها نیز به‌دست می‌آیند. در فصل‌های بعد

این موضوع را به‌طور کامل شرح می‌دهیم.

۵.۱ رهیافت‌های استنباط آماری

مجموعه روش‌های آماری که برای به‌دست آوردن دانش، در مورد ویژگی‌های موردنظر داده‌ها، به‌کار می‌روند را استنباط آماری گویند. کسب اطلاعات، بر اساس داده‌های مشاهده‌شده در مورد مسئله مورد بررسی، مهم‌ترین هدف استنباط آماری است. استنباط آماری دو رهیافت اساسی را شامل می‌شود:

- رهیافت کلاسیک[۲۸]، که به‌طور عمده مبتنی بر روش درستنمایی[۲۹] است.
- رهیافت بیزی[۳۰].

۱.۵.۱ رهیافت کلاسیک استنباط آماری

در رهیافت کلاسیک استنباط آماری، فرض می‌شود پارامترهای مدل، کمیت‌هایی ثابت ولی معمولاً نامعلوم هستند. در آمار کلاسیک، احتمال یک پیشامد برابر با فراوانی نسبی وقوع آن پیشامد تعریف می‌شود. این بدان معناست که احتمال برای پیشامدهای قابل تکرار در نظر گرفته می‌شود که در آن عدم قطعیت به دلیل تصادفی بودن پیشامدها تعریف می‌شود. در صورتی که عدم قطعیت وقوع پیشامد به دلیل عدم تکرار آن پیشامد باشد، خواص احتمالی پارامترها را نمی‌توان به‌دست آورد. از این جهت اصل تکرار پذیری[۳۱] در این رهیافت، یکی از اصول پایه‌ای به حساب می‌آید. برای مثال، در رهیافت کلاسیک اگر یک فاصله اطمینان ۹۵ درصد برای میانگین نامعلوم جامعه $[-1,1]$ باشد، به این معنا نیست که احتمال این‌که میانگین در این فاصله قرار گیرد برابر ۹۵ درصد است؛ بلکه بدان معنا است که اگر به دفعات مکرر نمونه‌گیری شده و یک فاصله اطمینان ۹۵ درصد ساخته شود، در ۹۵ درصد موارد انتظار بر این است این فاصله‌ها شامل میانگین باشند. در ادامه، رهیافت استنباط آماری مبتنی به روش درستنمایی را شرح می‌دهیم.

روش درستنمایی ماکسیمم

یکی از روش‌های معمول برای برآورد پارامترهای نامعلوم و استنباط در مدل‌های آماری، روش درستنمایی ماکسیمم است که به اختصار آن را MLE می‌نامند. شاید این روش متداول‌ترین روش در بین کسانی باشد که از آمار استفاده می‌کنند. روش درستنمایی ماکسیمم نخستین بار

[۲۸] Classical Approach
[۲۹] Likelihood-Based
[۳۰] Bayesian Approach
[۳۱] Repeated Measure Principle

توسط گاوس (۱۹۲۱) به‌کار گرفته شد و پس از آن توسط فیشر (۱۹۲۵) به‌صورت گسترده‌تری در انتقاد از روش گشتاوری[۳۲] استفاده شد. برای معرفی این روش، ابتدا تابع درستنمایی را معرفی می‌کنیم.

تعریف ۱.۵.۱ (تابع درستنمایی). برای هر بردار مشاهده‌شده $\mathbf{X} = \mathbf{x}$، تابع درستنمایی $\mathbf{X}$ را تابع چگالی احتمال توأم (تابع احتمال توأم) $\mathbf{X}$، یعنی $f(\mathbf{x}|\boldsymbol{\theta})$ می‌نامیم که به‌صورت تابعی از پارامتر $\boldsymbol{\theta}$ در نظر گرفته می‌شود و آن را با نماد $L(\boldsymbol{\theta})$ نمایش می‌دهیم. پس

$$L(\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}).$$

در ادامه به ذکر چند نکته که از اهمیت زیادی برخوردار هستند، می‌پردازیم:

- تابع درستنمایی $L(\boldsymbol{\theta})$ لزوماً نسبت به $\boldsymbol{\theta}$ مشتق‌پذیر نیست.
- تابع درستنایی $L(\boldsymbol{\theta})$ یک تابع چگالی احتمال (تابع احتمال) نیست، بلکه متناسب با تابع چگالی احتمال است.
- اگر $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ یک نمونه تصادفی nتایی از توزیعی از خانواده چگالی‌های

$$\{f(\mathbf{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

باشد، آن‌گاه

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x_i}|\boldsymbol{\theta}).$$

تعریف ۲.۵.۱ (برآورد درستنمایی ماکسیمم). فرض کنید $\delta(\mathbf{X})$ برآوردگری برای $\boldsymbol{\theta}$ باشد، به‌طوری که

$$i) \qquad P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) \in \Theta) = 1 \qquad \forall\, \boldsymbol{\theta} \subseteq R^p$$

$$ii) \qquad L(\delta(\mathbf{X})) \geq L(\boldsymbol{\theta}) \qquad \forall\, \boldsymbol{\theta} \subseteq R^p$$

در این صورت $\delta(\mathbf{X})$ به‌عنوان یک برآوردگر درستنمایی ماکسیمم $\boldsymbol{\theta}$ تعریف می‌شود. به‌طور معمول برآوردگر درستنمایی ماکسیمم $\boldsymbol{\theta}$ را با $\hat{\theta}$ نشان می‌دهند و لازم نیست که یکتا باشد. همچنین طبق تعریف $\hat{\theta}$ (پارسیان، ۱۳۹۱) داریم

$$L(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}).$$

روش‌های متداول در رهیافت کلاسیک، روش‌های مبتنی بر درستنمایی هستند که برای محاسبه تابع درستنمایی در رده مدل‌های پیچیده، مانند مدل‌های آمیخته خطی[۳۳] و همچنین

[۳۲] Method of Moments
[۳۳] Linear Mixed Models

مدل‌های آمیخته خطی تعمیم‌یافته[۳۴] ، شامل حل انتگرال‌هایی با بعد بالا هستند و محاسبه آن‌ها زمان‌بر و دشوار است. بنابراین این روش که به‌صورت شهودی مورد توجه قرار می‌گیرد، بهترین روش نیست و نمی‌توان همیشه از آن بهره برد. به همین جهت رهیافت بیزی استنباط آماری که استنباط در آن مبتنی بر توزیع پسین صورت می‌گیرد، به‌عنوان جایگزینی مناسب معرفی می‌گردد.

۲.۵.۱ رهیافت بیزی استنباط آماری

یک رهیافت کاربردی و منعطف در استنباط آماری، رهیافت بیزی است. در رهیافت بیزی پارامتر واقعی و نامعلوم $\boldsymbol{\theta}$، به‌عنوان تحققی از یک توزیع احتمالی به‌نام توزیع پیشین در نظر گرفته می‌شود. این در حالی است که آماردان‌های کلاسیک بر این باور هستند که پارامتر نامعلوم را به دشواری می‌توان به‌صورت یک متغیر تصادفی به حساب آورد. همچنین تعیین توزیع احتمالی برای چنین متغیری نیز به نظر و عقیده آماردان‌ها بستگی دارد. آماردان‌های بیزی بر این باور هستند که با ترکیب اطلاعات شخصی توسط توزیع پیشین و اطلاعات نهفته در داده‌ها توسط تابع درستنمایی، می‌توان به واقعیت نزدیک شد. اگر در استنباط بیزی، میزان اطلاعات حاصل از داده‌ها افزایش یابد، به‌طور مثال اگر حجم نمونه زیاد شود، اثر توزیع پیشین ناچیز می‌شود. پس انتظار بر این است که برآوردگرهای بیزی و کلاسیک، عملکرد یکسانی داشته باشند. اما اگر نمونه کوچک باشد، ممکن است اختلاف چشمگیری حاصل شود. در این‌صورت استنباط بیزی می‌تواند بسیار قوی‌تر از استنباط کلاسیک ظاهر شود. به‌علاوه در این دیدگاه می‌توان با استفاده از داده‌ها، نسخه به‌روزتری از توزیع پیشین ساخت و توزیع احتمال جدید را که توزیع پسین نام دارد، به‌دست آورد.

قضیه ۱.۵.۱ (قضیه بیز)**.** فرض کنید $A_1, A_2, \ldots, A_n$ یک افراز از فضای نمونه و E یک پیشامد دلخواه باشد. در این‌صورت برای هر $i = 1, 2, \ldots, n$ داریم

$$P(A_i|E) = \frac{P(E|A_i)P(A_i)}{\sum_{i=1}^{n} P(E|A_i)P(A_i)}.$$

این قضیه یک روش رسمی برای به‌روز کردن شانس A_i از $P(A_i)$ به $P(A_i|E)$ وقتی E مشاهده شده است، می‌باشد. $P(A_i|E)$ را احتمال پسین و $P(A_i)$ را احتمال پیشین می‌نامند (جفریز، ۱۹۳۶).

فرض کنید $\mathbf{x}$ یک بردار تصادفی از خانواده توابع چگالی‌های $f(\mathbf{x}|\boldsymbol{\theta})$ باشد. در این صورت قضیه بیز برای تابع چگالی $f(\mathbf{x}|\boldsymbol{\theta})$ به‌صورت زیر است:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}.$$

[۳۴] Generalized Linear Mixed Models

در رابطه بالا، $\pi(\boldsymbol{\theta})$ تابع (چگالی) احتمال توزیع پیشین و $\pi(\boldsymbol{\theta}|\mathbf{x})$ تابع (چگالی) احتمال توزیع پسین $\boldsymbol{\theta}$ است. به‌دلیل این‌که مقدار $\int_\Theta f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ مستقل از $\boldsymbol{\theta}$ است و مقدار آن یک عدد ثابت است، برای برای محاسبه $\pi(\boldsymbol{\theta}|\mathbf{x})$ معمولاً از محاسبه انتگرال خودداری کرده و توزیع پسین را به‌صورت زیر می‌نویسند:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

توزیع پیشین

توزیع پیشین $\boldsymbol{\theta}$، $\pi(\boldsymbol{\theta})$، بیان‌کننده اطلاعات و باورهای آماردان‌ها در مورد مقدار نامعلوم $\boldsymbol{\theta}$ است. به‌عبارت دیگر، قبل از مشاهده و گردآوری هر گونه داده، اطلاعات و دانسته‌های قبلی آماردان‌ها، آن‌ها را متقاعد می‌کند این باور را داشته باشند که شانس قرار گرفتن مقادیر $\boldsymbol{\theta}$ در فضای Θ چگونه می‌تواند باشد. فرض می‌شود چنین باورهایی می‌توانند در قالب یک توزیع بیان شوند. بنابراین، مجموعه همه این اطلاعات در قالب یک توزیع که همان توزیع پیشین است خلاصه می‌شوند. هنگامی که توزیع پیشین دارای واریانس بسیار بزرگ باشد، در برابر درستنمایی، به‌این دلیل که هیچ اطلاعاتی در مورد پارامتر در اختیار آماردانان نمی‌گذارد، مغلوب[۳۵] خواهد شد. در ادامه برخی از مفاهیم پایه‌ای توزیع‌های پیشین را مطرح می‌کنیم.

تعریف ۳.۵.۱ (توزیع‌های پیشین سره[۳۶]). توزیع پیشین با تابع چگالی احتمال $\pi(\boldsymbol{\theta})$ را سره گوییم، هرگاه

$$\int_\Theta \pi(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} < \infty.$$

تعریف ۴.۵.۱ (توزیع‌های پیشین ناسره[۳۷]). توزیع پیشین با تابع چگالی احتمال $\pi(\boldsymbol{\theta})$ را ناسره گوییم، هرگاه

$$\int_\Theta \pi(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} = +\infty.$$

اگر توزیع پیشین سره باشد، توزیع پسین حتما سره خواهد بود؛ اما اگر توزیع پیشین ناسره باشد، ممکن است توزیع پسین سره شود. توجه داشته باشید زمانی می‌توان از توزیع پیشین ناسره برای $\boldsymbol{\theta}$ استفاده کرد که توزیع پسین به‌دست آمده سره باشد. در غیر این‌صورت استنباط آماری ممکن نخواهد بود (رابرتز، ۲۰۰۷). بنابراین اگر در مدل بیزی از یک توزیع پیشین ناسره استفاده کنیم، بررسی سره بودن به یک مسئله کلیدی تبدیل می‌شود.

تعریف ۵.۵.۱ (توزیع‌های پیشین مزدوج[۳۸]). توزیع پیشین $\pi(\boldsymbol{\theta})$ را برای $f(\mathbf{x}|\boldsymbol{\theta})$ مزدوج گوییم، هرگاه توزیع پسین آن متعلق به خانواده توزیع پیشین $\pi(\boldsymbol{\theta})$ باشد (رابرتز، ۲۰۰۷).

[۳۵] Dominated
[۳۶] Proper
[۳۷] ImProper
[۳۸] Conjucate Prior

تعریف ۶.۵.۱ (توزیع‌های پیشین ناآگاهی‌بخش[۳۹]). توزیع پیشین ناآگاهی‌بخش، یکی از مهم‌ترین پیشین‌هایی است که کاربرد دارد. این توزیع پیشین در مواردی که اطلاع چندانی در مورد پارامتر موجود نباشد، استفاده می‌شود. در این حالت سعی آماردانان بر این است که احتمال برابری را برای همه پیشامدهای ممکن اختصاص دهند. چنین توزیعی را ناآگاهی‌بخش می‌گویند. پیشین ناآگاهی‌بخش را به‌صورت $\pi(\boldsymbol{\theta}) = k$، $k > 0$، تعریف می‌کنند که در این پیشین انتخاب k اختیاری است. اگر از توزیع پیشین ناآگاهی‌بخش استفاده کنیم، توزیع پسین به‌صورت زیر خواهد بود:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto k f(\mathbf{x}|\boldsymbol{\theta}) = kL(\boldsymbol{\theta}). \qquad (۲.۱)$$

اگر متغیر مورد بررسی روی دامنه فضای پارامتر نامحدود باشد، در این‌صورت اختصاص دادن یک مقدار ثابت یکنواخت به‌عنوان توزیع پیشین، توزیع پسین ناسره ایجاد می‌کند. چگونگی انتخاب k در رابطه (۲.۱) به روش‌های متفاوتی انجام می‌شود (برگر و برنارد، ۱۹۹۲). برای مثال، پیشین ناآگاهی‌بخش یکنواخت[۴۰] و پیشین ناآگاهی‌بخش جفریز[۴۱] را می‌توان نام برد.

پیشین ناآگاهی‌بخش یکنواخت

برای نخستین بار لاپلاس (۱۸۱۲) با انتخاب $k = 1$ پیشین ناآگاهی‌بخش را به‌صورت $\pi(\boldsymbol{\theta}) = 1$ استفاده کرد که به پیشین ناآگاهی‌بخش یکنواخت معروف است.

پیشین ناآگاهی‌بخش جفریز

جفریز (۱۹۶۱) توزیع پیشین ناآگاهی‌بخش را که بر اساس ماتریس اطلاع فیشر[۴۲] پایه‌ریزی شده بود، معرفی کرد. فرض کنید

$$I(\boldsymbol{\theta}) = (I_{ij}(\boldsymbol{\theta}))$$

ماتریس اطلاع فیشر باشد، که در آن

$$I_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\mathbf{x}|\boldsymbol{\theta})\right).$$

در این‌صورت توزیع پیشین جفریز به‌صورت

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}$$

تعریف می‌شود.

---

[۳۹] Noninformative Prior

[۴۰] Noninformative Uniform Prior

[۴۱] Noninformative Jeffreys Prior

[۴۲] Fisher Information

در عمل انتخاب توزیع پیشین یکی از اصلی‌ترین بخش‌های استنباط بیزی محسوب می‌شود. آماردان‌های بیزی مختلف روش‌های متعددی را برای تعریف توزیع پیشین پیشنهاد کرده‌اند که برای اطلاع بیشتر در مورد آن‌ها می‌توانید به رابرتز (۲۰۰۷) مراجعه کنید.

توزیع پسین

محاسبه توزیع پسین، حتی در ساده‌ترین مسائل استنباط آماری، ممکن است نیازمند حل انتگرال‌های چندگانه باشد. همان‌طور که می‌دانیم محاسبه بسیاری از انتگرال‌های چندگانه، کار دشواری است و روش متعارفی برای حل آن‌ها وجود ندارد. بنابراین راه حل معمول، حل تقریبی آن‌ها است. یک رده کلی از روش‌های تقریب توزیع پسین، روش‌های مبتنی بر نمونه‌گیری، مانند روش‌های رد و پذیرش[۴۳] (رابرتز و کسلا، ۲۰۱۰)، نمونه‌گیری نقاط مهم[۴۴] و نمونه‌گیری MCMC است. با به‌کارگیری این روش‌ها، جهت انجام استنباط‌های بیزی مبتنی بر توزیع پسین، دیگر نیازی به بهینه‌سازی توابع پیچیده و محاسبه گشتاورهای پیچیده (انتگرال‌های پیچیده) نیست. در ادامه، به اختصار برخی از این روش‌های نمونه‌گیری را شرح می‌دهیم.

۶.۱ روش‌های تقریب توزیع پسین مبتنی بر نمونه‌گیری

در بسیاری از کاربردها مانند مدل‌های بیزی پیچیده، محاسبه توابع چگالی نامعلوم، تنها به کمک شبیه‌سازی امکان‌پذیر است و در این میان روش‌های مونت کارلویی از اهمیت بالایی برخوردار هستند. روش‌های شبیه‌سازی مونت‌کارلویی، مبتنی بر نمونه‌گیری مکرر از توزیع هدف هستند.

استفاده از روش مونت کارلویی به‌جای روش‌های عددی در استنباط بیزی، به چند دلیل برتری دارد. نخست این‌که روش مونت کارلویی، مشکل روش‌های عددی در حل انتگرال‌ها را ندارد. به‌بیان دیگر، روش‌های مونت کارلویی به نواحی چگال‌تر تابع زیر انتگرال، وزن بیشتری در محاسبه تقریب می‌دهد، در حالی که در روش‌های عددی، همه نواحی زیر تابع انتگرال با وزن یکسان در نظر گرفته می‌شوند. دلیل دوم این‌که، بعد فضای پارامتر در اغلب مسائل بیزی بسیار بالا است و برای اینکه بتوانیم یک تقریب قابل قبول به‌دست آوریم، باید روش‌های عددی بسیار پیچیده‌ای را به‌کار بگیریم. دلیل سوم، دقت روش‌های عدی با افزایش بعد کاهش می‌یابد؛ درحالی‌که روش‌های مونت کارلویی از نظر دقت، به بعد پارامتر بستگی ندارند (رابرت و کسلا، ۲۰۰۴؛ ۲۰۱۰). ایده اصلی روش‌های مونت کارلویی این است که انتگرال را به یک امید ریاضی، بر اساس یک تابع چگالی احتمال مشخص تبدیل می‌کند. سپس با استفاده

[۴۳] Accept-Reject Methods

[۴۴] Importance Sampling

از قانون قوی اعداد بزرگ[۴۵] (گات، ۲۰۰۵) این امید ریاضی را تقریب می‌زند. در واقع مسئله اصلی این است که انتگرال زیر را محاسبه کنیم:

$$J = E_f[h(X)] = \int_{\chi} h(x) f(x)\, \mathrm{d}x$$

که در آن، χ تکیه‌گاه متغیر تصادفی X است که می‌تواند یک یا چند بعدی باشد. همچنین تابع $f(\cdot)$ یک تابع چگالی دارای شکل بسته یا به‌طور جزئی بسته است و تابع $h(\cdot)$ نیز یک تابع حقیقی مقدار است. الگوریتم روش مونت کارلویی به‌صورت زیر است:

- یک نمونه تصادفی mتایی از تابع چگالی $f(\cdot)$ تولید کنید.
- با جایگذاری این مقادیر در تابع $h(\cdot)$، مقدار عبارت زیر را حساب کنید:

$$\frac{1}{m}\sum_{j=1}^{m} h(x_j).$$

طبق قانون قوی اعداد بزرگ، مقدار فوق همان برآورد امید ریاضی است. به‌عبارت دیگر

$$\bar{h}_m = \frac{1}{m}\sum_{j=1}^{m} h(x_j) \xrightarrow{a.s.} E_f[h(x)].$$

بنابراین به محض این‌که نمونه $(x_1, x_2, \ldots, x_m)$ از توزیع $f(\cdot)$ تولید شوند، از این تقریب می‌توان برای محاسبه کمیت‌های مورد علاقه توزیع، با تابع چگالی احتمال $f(x)$ استفاده کرد. در روش مونت کارلویی، مسأله اصلی تولید کردن نمونه از توزیع دلخواه است که در اغلب موارد قادر به تولید مستقیم نمونه از تابع $f(x)$ نیستیم. بنابراین باید از روش‌های غیر مستقیم برای تولید نمونه استفاده کنیم (چن و همکاران، ۲۰۰۰). الگوریتم متروپلیس‌ـ‌هستینگز (MH) و الگوریتم نمونه‌گیر گیبز از جمله روش‌هایی هستند که برای تولید غیرمستقیم نمونه به‌کار برده می‌شوند.

۱.۶.۱ روش‌های زنجیر مارکوف مونت کارلویی

یک چارچوب کلی از روش‌هایی که برای انتگرال‌گیری توسط متروپلیس و همکاران (۱۹۵۳) معرفی شد، روش‌های MCMC است. سپس این روش‌ها توسط هستینگز (۱۹۷۰) به منظور تمرکز روی مسائل آماری، تطبیق و تعمیم یافت. روش‌های MCMC همواره در کنار افزایش قدرت پردازش و سرعت محاسبات کامپیوترها، تکامل و گسترش یافته‌اند. روش‌های نمونه‌گیری از زنجیر مارکوف با توزیع مانای $f(\cdot)$[۴۶]، برای تولید نمونه غیر مستقیم از $f(\cdot)$ طراحی شده‌اند که برای بعدهای بالا نیز کارا هستند. در این روش‌ها، توزیعی مانند $f(\cdot)$، که اغلب آن را

[۴۵] Strong Low of Large Number

[۴۶] Stationary Distribution

توزیع هدف[۴۷] می‌گویند، وجود دارد. اگر تولید نمونه به‌صورت مستقیم از توزیع $f(\cdot)$ سخت باشد، یعنی $f(\cdot)$ پیچیده باشد، یک روش نمونه‌گیری غیرمستقیم، ساختن یک زنجیر مارکوف تحویل‌ناپذیر[۴۸] و بازگشتی[۴۹] (ایلدریم، ۲۰۱۲) با توزیع مانای $f(\cdot)$ است که با نمونه‌گیری از این توزیع و استفاده از قضیه ارگودیک[۵۰] (بیرخوف، ۱۹۴۲) می‌توان به برآورد تابع دلخواه از توزیع مورد نظر پرداخت.

تعریف ۱.۶.۱ (زنجیر مارکوف)**.** یک زنجیر مارکوف $\{X^{(t)}\}$ دنباله‌ای از متغیرهای تصادفی وابسته $(X^{(۰)}, X^{(۱)}, \ldots, X^{(t)}, \ldots)$ است، به گونه‌ای که توزیع احتمال $X^{(t)}$ به شرط مفروض بودن متغیرهای گذشته، تنها به $X^{(t-۱)}$ وابسته است. یعنی

$$X^{(t)}|X^{(۰)}, X^{(۱)}, \ldots, X^{(t-۱)} \sim K(X^{t}|X^{(t-۱)}). \qquad (۳.۱)$$

در رابطه (۳.۱) توزیع احتمال شرطی $K(\cdot|\cdot)$، هسته مارکوف[۵۱] نامیده می‌شود.

نمونه تولیدشده بر اساس زنجیر مارکوف فاقد کیفیت یک نمونه مستقل با حجم مشابه است؛ زیرا نمونه تولیدشده بر اساس زنجیر مارکوف وابسته است. کیفیت چنین نمونه‌ای، با استفاده از کمیتی به‌نام حجم نمونه موثر[۵۲] (ESS) قابل ارزیابی است. هرچه مقدار ESS محاسبه‌شده برای هر پارامتر به تعداد نمونه‌های تولیدشده نزدیک‌تر شود، نمونه‌های تولیدشده از کیفیت بهتری برخوردار هستند.

تعریف ۲.۶.۱. فرض کنید n تعداد نمونه‌های تولیدشده از زنجیر باشد، آن‌گاه ESS برای پارامتر مورد نظر به‌صورت زیر محاسبه می‌شود:

$$ESS = \frac{n}{۱ + ۲\sum_{k=۱}^{\infty} \rho_k}.$$

که در آن ρ_k مقدار خودهمبستگی در گام kام را نشان می‌دهد. در واقع ESS کمیتی است که تعداد نمونه‌های مستقل به‌دست آمده از زنجیر را تخمین می‌زند.

اولین قدم در الگوریتم MCMC با فرض داشتن چگالی هدف $f(\cdot)$، به‌دست آوردن هسته مارکوف $K(\cdot|\cdot)$ با توزیع مانای f است. دومین قدم تولید زنجیر مارکوف $X^{(t)}$ با استفاده از این زنجیر که توزیع حدی زنجیر $X^{(t)}$ چگالی هدف $f(.)$ است و با افزایش t زنجیر به توزیع هدف همگرا می‌شود. در این حالت زنجیر تولیدشده نسبت به توزیع هدف شرایطی را باید داشته باشد که آن‌ها را در زیر بیان می‌کنیم:

---

[۴۷] Target Distribution
[۴۸] Irreducible Markov Chain
[۴۹] Recurrent
[۵۰] Ergodic Theorem
[۵۱] Markov Kernel
[۵۲] Effective Sample Size

- زنجیر تولیدشده نسبت به توزیع هدف باید بازگشتی باشد، به این معنا که زنجیر به هر مجموعه دلخواه غیرقابل اغماض، به دفعات نامتناهی باز می‌گردد.
- زنجیر تولیدشده باید تحویل‌ناپذیر باشد، یعنی هسته $K(\cdot|\cdot)$ اجازه حرکت بر روی تمام وضعیت‌های فضای حالت زنجیر را داشته باشد. به بیانی ساده‌تر، با صرف‌نظر کردن از مقدار اولیه $X^{(0)}$، دنباله $\{X^{(t)}\}$ با احتمال مثبت به هر ناحیه از فضای حالت می‌تواند برود. وجود توزیع مانا، یکی از مهم‌ترین ویژگی‌های زنجیرهای مارکف است، به این معنا که توزیع احتمال f وجود دارد به‌طوری که

$$X^{(t)} \sim f$$

آن‌گاه

$$X^{(t+1)} \sim f.$$

این ویژگی شرط تحویل‌ناپذیری را به هسته ماکوف $K(\cdot|\cdot)$ تحمیل می‌کند.

- زنجیر تولیدشده باید ارگودیک باشد، یعنی به ازای هر مقدار اولیه $X^{(0)}$، توزیع حدی $X^{(t)}$، برابر با توزیع مانای زنجیر f است.

در صورتی که زنجیر تولیدشده این ویژگی‌ها را دارا نباشد، زنجیر مفیدی نیست؛ چون تقریب‌های به‌دست آمده از این زنجیر در شرایطی اعتبار دارند و به مقادیر نظری خود همگرا هستند که زنجیر ارگودیک، تحویل‌ناپذیر و بازگشتی باشد. در این صورت در مراحل نهایی زنجیر به یک توزیع مانا می‌رسیم. پس با توجه به شرط ارگودیک بودن زنجیر مارکوف، اگر $t \longrightarrow \infty$، آن‌گاه

$$\frac{1}{T}\sum_{t=1}^{T} h(X^{(t)}) \longrightarrow E_f(h(X))$$

که به آن قضیه ارگودیک نیز گفته می‌شود و طبق قانون قوی اعداد بزرگ، برای رهیافت MCMC نیز می‌توان آن را به‌کار گرفت (رابرت وکسلا، ۲۰۰۴). در ادامه دو روش مهم را که از توزیع‌های پیچیده با کمک زنجیرهای مارکوف، نمونه‌گیری می‌کند را معرفی می‌کنیم.

الگوریتم متروپلیس ـ هستینگز

در روش متروپلیس ـ هستینگز، اساس کار این‌گونه است که با فرض داشتن چگالی هدف، $f(\cdot)$، هسته مارکف $K(\cdot|\cdot)$ را با توزیع مانای f می‌سازیم. سپس یک زنجیر مارکوف $\{X^{(t)}\}$ را با استفاده از این هسته طوری تولید می‌کنیم که تابع چگالی توزیع حدی $X^{(t)}$، f باشد و پس از آن کمیت‌های مورد نظر مانند میانگین و واریانس که در قالب انتگرال بیان می‌شوند، با استفاده از قضیه ارگودیک به‌دست می‌آوریم. معمولاً ساختن هسته مارکوف $K(\cdot|\cdot)$ که به چگالی دلخواه $f(.)$ مربوط است، کار دشواری است. روش‌هایی وجود دارند که به کمک

آن‌ها می‌توان هسته‌هایی که شکل شناخته‌شده‌تری دارند را برای چگالی هدف $f(.)$ محاسبه کرد. از جمله این روش‌ها، می‌توان به الگوریتم متروپلیس–هستینگز اشاره کرد. در روش متروپلیس–هستینگز برای تولید نمونه از چگالی هدف، چگالی شرطی $q(y|x)$ را که چگالی پیشهادی[۵۳] نامیده می‌شود، انتخاب می‌کنیم. توزیعی را که به‌عنوان چگالی پیشنهادی معرفی می‌کنیم باید شکل واضح و صریحی داشته باشد تا اگر خواستیم با استفاده از روش‌های معمول از آن شبیه‌سازی کنیم، دشوار نباشد. همچنین تکیه‌گاه آن باید تا حدی وسیع باشد که بتواند کل تکیه‌گاه f را پوشش دهد. این الگوریتم به‌صورت زیر است:

- برای مقدار جاری[۵۴] $x^{(t)}$، یک مقدار پیشنهادی از چگالی پیشنهادی $q(y|x)$ تولید کنید؛ مثلاً
$$y_t \sim q(y|x).$$
- عبارت زیر را به‌دست آورید:
$$\frac{f(y_t)}{f(x^{(t)})}\frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}.$$
همان‌طور که مشاهده می‌شود این الگوریتم فقط به نسبت‌های $\frac{f(y_t)}{f(x^{(t)})}$ و $\frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}$ وابسته است.
- مقدار y_t را با احتمال $\rho(x^{(t)}, y_t)$ به‌عنوان مقدار بعدی زنجیر بپذیرید و قرار دهید
$$X^{(t+1)} = y_t$$
در غیر این‌صورت قرار دهید
$$X^{(t+1)} = X^{(t)}$$
که در آن
$$\rho(x^{(t)}, y_t) = \min\left\{1, \frac{f(y_t)}{f(x^{(t)})}\frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}\right\}.$$

در عبارت فوق $\rho(x^{(t)}, y_t)$ را احتمال پذیرش متروپلیس–هستینگز می‌گویند. هنگامی که توزیع پیشنهادی متقارن باشد، به‌بیان دیگر وقتی که $q(y|x) = q(x|y)$، آن‌گاه
$$\rho(x^{(t)}, y_t) = \min\left\{1, \frac{f(y_t)}{f(x^{(t)})}\right\}.$$
بنابراین احتمال پذیرش، مستقل از توزیع پیشنهادی $q(\cdot|\cdot)$ می‌باشد. عملکرد الگوریتم را معمولاً بر مبنای نرخ پذیرش آن ارزیابی می‌کنند که نرخ پذیرش الگوریتم، میانگین احتمال پذیرش بر روی تمامی تکرارها است. یعنی
$$\bar{\rho} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\rho(X^{(t)}, Y_t) = \int \rho(x,y)f(x)q(x|y)\,\mathrm{d}x\,\mathrm{d}y.$$

[۵۳] Proposal Density
[۵۴] Current Value

کارایی روش MH به تولید نمونه از تابع چگالی $q(x|y)$ بستگی دارد. در حقیقت تولید یک نمونه خوب زمانی امکان پذیر است که تابع چگالی پیشنهادی، یک تقریب مناسب برای تابع چگالی هدف باشد (رابرتز و کسلا، ۲۰۱۰). در ادامه دو حالت خاص از الگوریتم متروپلیس‌ـ هستینگز، به نام‌های الگوریتم متروپلیس‌ـ هستینگز مستقل[۵۵] والگوریتم متروپولیس‌ـ هستینگز قدم زدن تصادفی[۵۶] را معرفی می‌کنیم.

الگوریتم متروپلیس‌ـ هستینگز مستقل

در الگوریتم MH زمانی که تابع چگالی پیشنهادی، مستقل از $X^{(t)}$ باشد، الگوریتم را الگوریتم MH مستقل می‌نامند. در این حالت

$$q(x|y) = g(y).$$

با شرط داشتن $X^{(t)}$،

- $y_t \sim g(y)$ را تولید کنید.
- مقدار y_t را به‌عنوان مقدار جدید $X^{(t+1)}$ با احتمال

$$\rho(x^{(t)}, y_t) = \min\left\{1, \frac{f(y_t)}{f(x^{(t)})}\frac{q(x^{(t)}|y_t)}{q(y_t|x^{(t)})}\right\}$$

بپذیرید و قرار دهید

$$X^{(t+1)} = y_t$$

در غیر این‌صورت

$$X^{(t+1)} = X^{(t)}.$$

بنابراین مقدار پیشنهادی از یک توزیع مورد علاقه که مستقل از مقدار جاری است، نتیجه می‌شود. برای اطلاعات بیشتر در مورد ویژگی‌های این نوع الگوریتم به کتاب رابرتز و کسلا (۲۰۱۰) مراجعه کنید.

الگوریتم متروپلیس‌ـ هستینگز قدم زدن تصادفی

در الگوریتم MH قدم زدن تصادفی، مقدار پیشنهادی بر طبق مقدار شبیه‌سازی‌شده قبلی تولید می‌شود؛ به‌طوری که در آن یک جستجوی محلی[۵۷] حول یک همسایگی از مقدار جاری زنجیر مارکوف انجام می‌شود. در این روش، برای تولید بعدی زنجیر، الگوریتم نمونه تولیدشده قبلی

[۵۵] Independent Metropolis-Hastings
[۵۶] Random Walk Metropolis-Hastings
[۵۷] Local Exploration

را به حساب می‌آورد. یعنی در همسایگی موقعیت جاری زنجیر، کاوش موضعی بیشتری را انجام می‌دهد. اجرای این ایده، شبیه‌سازی y_t از رابطه زیر است:

$$y_t = X^{(t)} + \epsilon_t$$

که در آن، ϵ_t یک خطای تصادفی مستقل از $X^{(t)}$ با تابع چگالی $g(\cdot)$ است. در این حالت تابع چگالی توزیع پیشنهادی به‌صورت زیر است:

$$q(y_t|x^{(t)}) = g(y_t - x^{(t)}).$$

هنگامی که تابع چگالی $g(\cdot)$ حول صفر متقارن باشد، که در بیشتر مواقع این‌گونه است، زنجیر مارکوف متناظر با آن یک قدم زدن تصادفی است. فرض کنید $X^{(t)} = x^{(t)}$ را داشته باشیم، در این صورت الگوریتم به‌صورت زیر خواهد بود:

- یک مقدار پیشنهادی را از توزیع $g(\cdot)$، برای مقدار جاری زنجیر تولید کنید:

$$y_t \sim g(y_t - x^{(t)}).$$

- نسبت زیر را به‌دست آورید:

$$\frac{f(y_t)}{f(x^{(t)})}.$$

مقدار y_t را به‌عنوان مقدار جدید با احتمال $\min\left\{۱, \frac{f(y_t)}{f(x^{(t)})}\right\}$ بپذیرید و قرار دهید:

$$X^{(t+۱)} = y_t$$

در غیر این‌صورت

$$X^{(t+۱)} = X^{(t)}.$$

نکته مهمی که باید به آن اشاره کنیم این است که احتمال پذیرش، به توزیع پیشنهادی $g(\cdot)$ بستگی ندارد. یعنی برای یک جفت مفروض $(x^{(t)}, y_t)$، احتمال پذیرش، چه y_t از توزیع نرمال تولید شود چه از توزیع کوشی، یکسان خواهد بود. اما باید توجه داشت انتخاب کردن $g(\cdot)$های مختلف، منجر به دامنه‌های مختلف مقادیر برای y_t ها خواهد شد، در نتیجه نرخ‌های پذیرش تفاوت خواهند داشت. بنابراین، می‌توان گفت انتخاب $g(\cdot)$، بر رفتار الگوریتم تأثیر دارد.

در این الگوریتم، توزیع پیشنهادی ϵ_t دارای یک پارامتر مقیاس است که انتخاب آن بر عملکرد الگوریتم تأثیر به‌سزایی دارد. انتخاب بهینه برای این پارامتر به‌گونه‌ای است که نرخ پذیرش الگوریتم در فاصله [۱۵٪, ۵٪] قرار بگیرد (گلمن و همکاران، ۲۰۰۳).

الگوریتم نمونه‌گیر گیبز

الگوریتم نمونه‌گیر گیبز یکی دیگر از روش‌های نمونه‌گیری MCMC است. نام این الگوریتم از کار معروف گمان و گمان (۱۹۸۴) برگزیده شده است؛ اما استفاده شایع از این الگوریتم با کار گلفاند و اسمیت (۱۹۹۰) آغاز شد که از حل این الگوریتم برای حل مسئله‌های پیچیده در استنباط بیزی که قبلاً بدون حل باقی مانده بودند بهره بردند. برای اجرای این روش باید بتوانیم توابع چگالی شرطی کامل[۵۸] هر مؤلفه X را به شرط بقیه مؤلفه‌ها به‌دست آوریم. این الگوریتم زمانی می‌تواند کارا باشد که شبیه‌سازی از چگالی‌های شرطی کامل، ساده و وابستگی مولفه‌ها ناچیز باشد. در ادامه دو نسخه دو مرحله‌ای و چند مرحله‌ای این الگوریتم را شرح می‌دهیم.

الگوریتم نمونه‌گیر گیبز دومرحله‌ای

الگوریتم نمونه‌گیر گیبز دومرحله‌ای[۵۹] با استفاده از توزیع توأم یک زنجیر مارکوف را تشکیل می‌دهد. اگر دو متغیر X و Y وجود داشته باشند که دارای توزیع توأم $f(x,y)$ باشند، به‌طوری که چگالی‌های شرطی کامل متناظر با آن‌ها $f(x|y)$ و $f(y|x)$ باشند، آن‌گاه الگوریتم یک زنجیر مارکوف (X_t, Y_t) را با شرط مفروض بودن $X_0 = x_0$، برای $t = 1,2,\ldots$، طبق مراحل زیر تولید می‌کند:

- $Y_t \sim f_{Y|X}(.|x_{t-1})$
- $X_t \sim f_{X|Y}(.|y_t)$

الگوریتم نمونه‌گیر گیبز چندمرحله‌ای

به سادگی می‌توان الگوریتم نمونه‌گیر گیبز دومرحله‌ای را به الگوریتم نمونه‌گیر گیبز چندـ مرحله‌ای[۶۰] نسبت داد. فرض کنید متغیر تصادفی X به‌صورت $X = (X_1, X_2, \ldots, X_m)$ تعریف شود. مؤلفه‌های X_i ممکن است یک یا چند بعدی باشند. الگوریتم نمونه‌گیر گیبز تحت انتقال از $X^{(t)}$ به $X^{(t+1)}$ با تکرار $t = 1,2,3,\ldots$ و با فرض $X^{(t)} = (X_1^{(t)}, X_2^{(t)}, \ldots, X_m^{(t)})$ طبق مراحل زیر اجرا می‌شود:

۱. تولید $X_1^{(t+1)} \sim f_1(x_1|X_2^{(t)}, \ldots, X_m^{(t)})$

۲. تولید $X_2^{(t+1)} \sim f_2(x_2|X_1^{(t+1)}, X_3^{(t)}, \ldots, X_m^{(t)})$

⋮

[۵۸] Full Conditionall Densities

[۵۹] Two-Stage Samples

[۶۰] Multi-Stage Gibbs Sampling

m. تولید $X_m^{(t+1)} \sim f_m(x_m|X_1^{(t+1)},\ldots,X_{m-1}^{(t+1)})$

از مهم‌ترین ویژگی‌های نمونه‌گیر گیبز این است که چگالی‌های با بعد کوچکتر از چگالی توأم را جهت شبیه‌سازی به‌کار می‌گیرد. پس، مسائلی با ابعاد بالا را می‌توان به چند مسئله با ابعاد کوچکتر تقسیم کرد و این مزیت مهمی است که روش نمونه‌گیر گیبز از آن برخوردار است.

۷.۱ توزیع‌های احتمالی مورد نیاز

توزیع پواسن

در آمار توزیع پواسن یک توزیع گسسته است و احتمال این‌که یک حادثه به تعداد مشخصی در فاصله زمانی یا مکانی ثابتی رخ دهد را شرح می‌دهد. تابع جرم احتمال توزیع پواسن به شکل زیر است:

$$f_Y(y) = \frac{\exp(-H\lambda)(H\lambda)^y}{y!} \qquad y = 0,1,2,\ldots$$

که در آن، H پارامتر بی‌اثر[۶۱] و میانگین و واریانس توزیع برابر با $H\lambda$ است.

توزیع پواسن تعمیم‌یافته

استفاده از توزیع پواسن هنگامی که داده‌ها دارای عامل بیش‌پراکنش یا کم‌پراکنش هستند، می‌تواند نتایج گمراه‌کننده‌ای را به دنبال داشته باشد. برای این منظور توزیع‌های متعددی را پیشنهاد کرده‌اند که توزیع پواسن تعمیم‌یافته[۶۲] از جمله آن‌ها است. متغیر تصادفی Y دارای توزیع پواسن تعمیم‌یافته است، هرگاه تابع جرم احتمال آن به صورت زیر باشد:

$$P(Y=y) = \theta(\theta+\lambda y)^{y-1}\exp\left\{\frac{-\theta-\lambda y}{y!}\right\} \qquad y = 0,1,2,\ldots \qquad (۴.۱)$$

که در آن $\theta > 0$ و $\max(-1, \frac{-\theta}{4}) \leq \lambda \leq 1$.

این توزیع بسته به اینکه $\lambda > 0$ یا $\lambda < 0$ باشد به ترتیب می‌تواند بیش‌پراکنش و کم‌پراکنش را نمایش دهد. میانگین و واریانس توزیع برابر هستند با:

$$E(Y) = \mu$$

$$Var(Y) = \mu(1-\lambda\mu).$$

[۶۱] Offset

[۶۲] Generalized Poisson

بنابراین وقتی $\lambda = ۰$، این توزیع برابر با توزیع پواسن است. مدل (۴.۱) را می‌توان به‌صورت زیر بازنویسی کرد:

$$P(Y=y)=\mu(\mu+(\phi-۱)y)^{y-۱}\phi^{-y}\frac{\exp\left\{\frac{-(\mu+(\phi-۱)y)}{\phi}\right\}}{y!} \qquad y=۰,۱,۲,\dots \qquad (۵.۱)$$

توزیع (۵.۱) هنگامی که $\phi = ۱$ است، برابر توزیع پواسن می‌باشد. همچنین وقتی $\phi > ۱$ بیش‌پراکنش، و وقتی $\frac{۱}{۲} < \phi < ۱$ و نیز $\mu > ۲$، کم‌پراکنش توسط این توزیع نمایش داده می‌شود.

توزیع کانوی ـ مکس‌ول ـ پواسن (COM−پواسن)

این توزیع جزء خانواده توزیع‌های پواسن است و با توجه به دارا بودن دو پارامتر، نسبت به توزیع پواسن که یک پارامتر دارد، بسیار منعطف‌تر عمل می‌کند. گوییم متغیر تصادفی Y دارای توزیع $COM-$پواسن[۶۳] است هرگاه دارای تابع جرم احتمال زیر باشد:

$$P(Y=y)=\frac{\lambda^y}{Z(\lambda,\nu)(y!)^\nu} \qquad y=۰,۱,۲,\dots$$

که در آن

$$Z(\lambda,\nu)=\sum_{j=۰}^{\infty}\frac{\lambda^j}{(j!)^\nu}.$$

در این توزیع اگر $\nu = ۱$ و $\lambda = \exp(\lambda)$، توزیع پواسن نتیجه می‌شود. همچنین اگر $\nu = ۰$، $\lambda < ۰$ و $Z(\lambda,\nu)$ به‌صورت زیر باشد:

$$Z(\lambda,\nu)=\sum_{j=۰}^{\infty}\lambda^j=\frac{۱}{۱-\lambda}$$

می‌توان توزیع هندسی را از آن نتیجه گرفت. یعنی

$$P(Y=y)=\lambda^y(۱-\lambda) \qquad y=۰,۱,۲,\dots$$

گشتاورهای توزیع نیز از رابطه زیر قابل محاسبه هستند:

$$E(Y^{r-۱})=\begin{cases}\lambda E(Y+۱)^{۱-\nu} & r>۰\\ \lambda\frac{d}{d\lambda}E(Y^r)+E(Y)E(Y^r) & r<۰\end{cases}.$$

بنابراین میانگین توزیع برابر است با

$$E(Y)=\lambda\frac{\mathrm{d}\left[\log\{Z(\lambda,\nu)\}\right]}{\mathrm{d}\lambda}\approx\lambda^{\frac{۱}{\nu}}-\frac{\nu-۱}{۲\nu}$$

[۶۳] Conway Maxwell Poisson

یا

$$E(Y)=\sum_{j=0}^{\infty}\frac{j\lambda^{j}}{(j!)^{\nu}Z(\lambda,\nu)}.$$

واریانس توزیع نیز به‌صورت زیر است:

$$Var(Y)=\sum_{j=0}^{\infty}\frac{j^{2}\lambda^{j}}{(j!)^{\nu}Z(\lambda,\nu)}.$$

این توزیع را به اختصار به‌صورت $COM-$پواسن می‌نویسند و بسته به پارامترهایش می‌تواند بیش‌پراکنش و کم‌پراکنش داده‌ها را تحت پوشش قرار دهد (شمولی و همکاران، ۲۰۰۵؛ سلرز و شمولی،۲۰۱۰).

توزیع ابرپواسن

یکی دیگر از توزیع‌هایی که برای نشان دادن عامل بیش‌پراکنش و کم‌پراکنش کاربرد دارد، توزیع ابرپواسن[۶۴] است. این توزیع نیز به خانواده توزیع‌های پواسن تعلق دارد و به دلیل دارا بودن پارامترهای بیشتر نسبت به توزیع پواسن می‌تواند انعطاف‌پذیری بالایی از خود نشان دهد. این توزیع برای اولین بار توسط باردول و کرو (۱۹۶۴) معرفی شد که دارای تابع احتمال زیر است:

$$P(Y=y)=\frac{1}{{}_1F_1(1,\gamma,\lambda)}\frac{\lambda^{y}}{(\gamma)_y}\qquad y=0,1,2,\ldots$$

که در آن، $\gamma>0$، $\lambda>0$ و ${}_1F_1(a,c,z)$ تابع فوق هندسی تعمیم‌یافته بوده و مقدار آن برابر است با

$${}_1F_1(a,c,z)=\sum_{r=0}^{\infty}\frac{(a)_r z^{r}}{(c)_r r!}$$

و

$$(a)_r=a(a+1)\ldots(a+r-1)=\frac{\Gamma(a+r)}{\Gamma(a)}\qquad r>0$$

میانگین و واریانس توزیع به ترتیب برابر هستند با

$$\mu=\lambda-(\gamma-1)(1-f_0)=\lambda-(\gamma-1)\frac{{}_1F_1(1,\gamma,\lambda)}{{}_1F_1(1,\gamma,\lambda)}$$

$$\sigma^{2}=\lambda+(\lambda-(\gamma-1))\mu+\mu^{2}.$$

این توزیع با توجه به دارا بودن پارامترهای بیشتر نسبت به توزیع پواسن انعطاف‌پذیری بیشتری دارد و می‌تواند برای $\gamma>0$ بیش‌پراکنش و برای $\gamma<0$ کم‌پراکنش را پوشش دهد. همچنین این توزیع هنگامی که $\gamma=1$ باشد، برابر با توزیع پواسن است.

[۶۴] Hyper Poisson

توزیع پیرسون سه‌پارامتری مختلط (CTP)

توزیع پیرسون سه‌پارامتری مختلط[۶۵] نیز برای مرتفع کردن مشکل عدم پوشش عوامل بیش‌پراکنش و کم‌پراکنش در توزیع پواسن، توسط رودریگز–آوی (۲۰۰۴) پیشنهاد شد. این توزیع با توجه به دارا بودن سه پارامتر، بسیار منعطف است. متغیر تصادفی Y دارای توزیع پیرسون سه‌پارامتری مختلط است هرگاه تابع احتمال آن به صورت زیر باشد:

$$P(Y=y)=f_{\circ}\frac{(\alpha+\beta_i)_y(\alpha-\beta_i)_y}{(\gamma)_y y!}\qquad y=\circ,۱,۲,\ldots$$

که در آن $\beta>\circ$، $\gamma>\circ$ و $-\infty<\alpha<\frac{\gamma}{۲}$. میانگین و واریانس توزیع به ترتیب برابر هستند با

$$\mu=\frac{\alpha^۲+\beta^۲}{\gamma-۲\alpha-۱}$$

$$\sigma^۲=\frac{(\alpha^۲+\beta^۲)\left[(\gamma-\alpha-۱)\beta^۲\right]}{(\gamma-۲\alpha-۱)^۲(\gamma-۲\alpha-۱)}.$$

بیش‌پراکنش، کم‌پراکنش و پراکنش متعادل، بسته به شرایط α که در زیر آمده است، توسط این توزیع نمایش داده می‌شود:

- بیش‌پراکنش

$$\alpha>\frac{۲\gamma-۴-\sqrt{(۲\gamma-۴)^۲+۱۲(\beta^۲+\gamma-۱)}}{۶}.$$

- کم‌پراکنش

$$\alpha<\frac{۲\gamma-۴-\sqrt{(۲\gamma-۴)^۲+۱۲(\beta^۲+\gamma-۱)}}{۶}.$$

- پراکنش متعادل

$$\alpha=\frac{۲\gamma-۴-\sqrt{(۲\gamma-۴)^۲+۱۲(\beta^۲+\gamma-۱)}}{۶}.$$

توزیع دوجمله‌ای

در نظریه احتمال و آمار توزیع دوجمله‌ای، توزیعی گسسته است که از تعداد موفقیت‌ها در دنباله‌ای شامل n آزمایش مستقل برنولی با احتمال موفقیت π نتیجه می‌شود. متغیر تصادفی Y دارای توزیع دوجمله‌ای با احتمال موفقیت π است هرگاه

$$P(Y=y)=\binom{n}{x}\pi^y(۱-\pi)^{n-y}$$

[۶۵] Complex Triparametric Pearson

که در آن، $n \geq ۰$ و $۰ \leq \pi \leq ۱$. میانگین و واریانس این توزیع به ترتیب برابر هستند با

$$E(Y) = n\pi$$

$$Var(Y) = n\pi(۱ - \pi).$$

توزیع بتا دوجمله‌ای

متغیر تصادفی Y دارای توزیع بتا دوجمله‌ای[۶۶] است، هرگاه تابع جرم احتمال آن به‌صورت زیر باشد:

$$P(Y = y) = \binom{n}{y} \frac{Beta(y + a, n - y + \beta)}{Beta(\alpha, \beta)} \quad y = ۰, ۱, ۲, \ldots, n$$

میانگین و واریانس آن به ترتیب برابر است با:

$$E(Y) = \frac{n\alpha}{(\alpha + \beta)}$$

$$Var(Y) = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^{۲}(\alpha + \beta + ۱)}.$$

این توزیع از آمیختن یک توزیع بتا برای پارامتر موفقیت π در توزیع دوجمله‌ای نتیجه می‌شود. به‌عبارت دیگر اگر

$$Y|\pi \sim Binimial(n, \pi)$$

و

$$\pi \sim Beta(\alpha, \beta)$$

آن‌گاه

$$Y \sim Beta-Binomial(n, \alpha, \beta).$$

توزیع دوجمله‌ای منفی

در آزمایش‌های برنولی، گاهی می‌خواهیم احتمال x موفقیت را از n آزمایش بدانیم که این مورد به توزیع دوجمله‌ای[۶۷] مربوط است. اما گاهی نیاز به دانستن تعداد آزمایش‌هایی داریم که در آن‌ها k موفقیت رخ داده است. در این‌صورت از توزیع دوجمله‌ای منفی استفاده می‌کنیم. متغیر Y دارای توزیع دوجمله‌ای منفی است، هرگاه تابع احتمال آن به شکل زیر باشد:

[۶۶] Beta Binomial

[۶۷] Negative Binomial

$$P(Y=y)=\binom{y+r-۱}{y}\pi^{y}(۱-\pi)^{r}\quad y=۰,۱,۲,\dots$$

که در آن $r>۰$ بیانگر تعداد شکست‌ها و $۰<\pi<۱$ احتمال موفقیت در هر آزمایش است. میانگین و واریانس این توزیع برابر است با:

$$E(Y)=\frac{r\pi}{۱-\pi}$$

$$Var(Y)=\frac{r\pi}{(۱-\pi)^{۲}}.$$

همان‌گونه که ملاحظه می‌شود، واریانس این توزیع از میانگین آن بیشتر است. بنابراین این توزیع می‌تواند برای داده‌های شمارشی، مفهوم بیش‌پراکنش را به خوبی پوشش دهد. اما برای مدل‌بندی داده‌های کم‌پراکنش گزینه مناسبی نیست.

۸.۱ سایر تعاریف مورد نیاز

تعریف ۱.۸.۱ (متغیرهای پنهان). در آمار متغیرهای پنهان[۶۸] مقابل متغیرهای مشاهده‌شده هستند. متغیرهای پنهان به‌صورت مستقیم در مدل قابل مشاهده نیستند و تنها می‌توانند از میان سایر متغیرهایی که قابل مشاهده هستند توسط یک مدل آماری به‌کار گرفته شوند. در مدل‌بندی آماری، از این مفهوم با توجه به زمینه‌ای که به‌کار برده می‌شود، برای بیان متغیرهای تصادفی و پارامترهای مدل استفاده می‌شود (بورسبوم و همکاران، ۲۰۰۳).

تعریف ۲.۸.۱ (تشخیص‌پذیری). در تشخیص‌پذیری[۶۹] مدل، هدف اصلی بررسی یکتا بودن پارامترها است. به عبارت دیگر، برای بررسی مقدار یگانه هر یک از پارامترهای آزاد در مدل از روی داده‌های مشاهده‌شده، از مفهوم تشخیص‌پذیری مدل استفاده می‌شود. در عمل برای بررسی تشخیص‌پذیری مدل از چندین مجموعه مقادیر اولیه برای برآورد پارامترها استفاده می‌شود. چنانچه مقادیر متفاوت اولیه به برآورد مشابهی برای پارامترها منجر شوند، مدل تشخیص‌پذیر است (دایتون و مک ریدی، ۱۹۸۸). به بیانی دقیق‌تر، پارامتر β تشخیص‌پذیر است، هرگاه به ازای هر $\beta_۱$ و $\beta_۲$ که $\mu(\beta_۱)=\mu(\beta_۲)$، حتما تساوی $\beta_۲=\beta_۲$ برقرار باشد.

تعریف ۳.۸.۱ (بیش‌پراکنش). در داده‌های شمارشی، وقتی داده‌ها از توزیع پواسن تبعیت می‌کنند، انتظار بر این است که در این حالت میانگین و واریانس برابر باشد. اما در مواردی پذیره برابری میانگین و واریانس برقرار نیست و واریانس بیشتر از میانگین است. در این حالت پیش‌فرض توزیع پواسن برقرار نیست و اصطلاحاً می‌گویند داده‌ها دارای بیش‌پراکنش هستند.

[۶۸] Latent Variables
[۶۹] Identifiability

تعریف ۴.۸.۱ (کم‌پراکنش)**.** با توجه به تعریف قبل، وقتی داده‌ها شمارشی و دارای توزیع پواسن هستند، در مواردی ممکن است واریانس داده‌ها کمتر از میانگین آن‌ها باشد. در این حالت می‌گویند داده‌ها کم‌پراکنش هستند.

فصل ۲

استنباط بیزی ناپارامتری

۱.۲ مدل بیزی ناپارامتری

برای شناخت فرآیندی که داده‌ها از آن تولید شده‌اند، از مدل‌های آماری استفاده می‌کنیم. فرض کنید داده‌ها تحققی از متغیرهای تصادفی $Y_1, Y_2, \ldots, Y_n$ هستند. معمولاً فرض می‌شود که $Y_1, Y_2, \ldots, Y_n$ به‌طور مستقل از توزیع احتمالی با اندازه احتمال زیر تولید شده‌اند:

$$\mu(B) = P(Y \in B)$$

که در آن B زیرمجموعه‌ای اندازه‌پذیر از فضای نمونه و $P(A)$ احتمال پیشامد A است. فرض کنید $f(x)$ تابع چگالی متناظر با اندازه احتمال $\mu(\cdot)$ باشد. در مدل‌های پارامتری معمولاً فرض می‌شود که $f(x)$ به‌صورت $f_{\boldsymbol{\theta}}(x)$ است و در واقع یک عضو از خانواده توزیع‌های $F = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ روی فضای متناهی $\Theta \subseteq \mathbb{R}^p$ است که در آن p یک عدد مثبت است. اصلی‌ترین هدف در مدل‌های پارامتری، برآورد یک مقدار قابل قبول برای پارامتر مدل، یعنی همان $\boldsymbol{\theta}$ است. همان‌طور که می‌دانیم در مدل‌های پارامتری، فرض محدودکننده معلوم بودن شکل توزیع، ممکن است در استنباط نتایج گمراه‌کننده‌ای را منجر شود. اما می‌توان با کنار گذاشتن فرض‌های پارامتری، یک مدل منعطف‌تر ارائه داد. این دسته از مدل‌ها حالتی را در نظر می‌گیرند که تابع چگالی در فضای بزرگ‌تر و همچنین با بعد نامتناهی پارامترها، قابل دستیابی هستند.

تعریف ۱.۱.۲ (فضای پارامتر با بعد نامتناهی و پارامتر با بعد نامتناهی). فرض کنید $\Theta \subseteq S$ باشد. در این‌صورت فضای خطی S که شامل تعداد نامتناهی عنصر است، فضای پارامتری با بعد نامتناهی[۱] و پارامترهای عضو این فضا را، پارامترهای با بعد نامتناهی می‌گویند.

مدل‌هایی که فقط پارامترهای با بعد نامتناهی را در نظر می‌گیرند، به‌عنوان مدل‌های ناپارامتری شناخته می‌شوند (گوش و رامامورتی، ۲۰۰۳؛ تسیاتیس، ۲۰۰۶). موارد دیگری هم هستند که در آن‌ها پارامترهای مدل، به دو دسته تقسیم می‌شوند؛ به‌عبارت دیگر پارامتر $\boldsymbol{\theta}$ به‌صورت $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ است. در این نوع پارامتربندی، پارامتر $\boldsymbol{\theta}_1$ با بعد q و پارامتر $\boldsymbol{\theta}_2$ با بعد نامتناهی فرض می‌شود. این نوع مدل‌ها را به دلیل وجود یک جزء پارامتری و یک جزء ناپارامتری، مدل‌های نیمه‌پارامتری می‌نامند (تسیاتیس، ۲۰۰۶). در حقیقت مدل نیمه‌پارامتری ترکیبی از مدل‌های پارامتری و ناپارامتری است. به‌طور کلی در مدل‌های بیزی ناپارامتری، برای تحلیل داده‌های $Y_1, Y_2, \ldots, Y_n$، که به‌طور مستقل از توزیع نامعلوم G تولید شده‌اند، فرض می‌شود G توزیعی تصادفی بوده و برای این توزیع تصادفی، پیشینی به‌صورت $\pi(G)$ در نظر می‌گیرند. در واقع

$$Y_1, Y_2, \ldots, Y_n | G \overset{iid}{\sim} G$$

$$G \sim \pi(G) .$$

بنابراین با توجه به آن‌چه گفته شد، در مدل‌های بیزی پارامتری توزیع مشاهدات مشخص است و برای پارامترهای نامعلوم این توزیع، پیشین در نظر گرفته می‌شود؛ ولی در مدل‌های بیزی ناپارامتری، اطلاعاتی پیرامون توزیع مشاهدات در دسترس نبوده و به همین جهت توزیع را تصادفی در نظر می‌گیرند و پیشین مناسبی روی این توزیع تصادفی تعریف می‌کنند. شیوه‌های مختلفی برای تعریف پیشین برای توزیع تصادفی وجود دارند (دی و همکاران، ۱۹۹۸) که رایج‌ترین آن‌ها بر مبنای فرآیندهای تصادفی است. فرآیندهای تصادفی روی خانواده‌ای از توابع توزیع تعریف شده و می‌تواند به‌عنوان یک پیشین مناسب برای توزیع تصادفی در نظر گرفته شود. به دلیل پیچیدگی محاسبات در روش‌های بیزی ناپارامتری، انتخاب نوع فرآیند تصادفی خود مسئله بسیار مهمی است. به همین جهت، فرآیندی باید انتخاب شود که انعطاف‌پذیر بوده و محاسبات بر پایه آن چندان دشوار نباشد. به‌عبارت دیگر پیشین انتخاب‌شده، پیشینی مزدوج باشد. از جمله این فرآیندها، فرآیند دیریکله[۲] است.

فرگوسن (۱۹۷۳) فرآیند دیریکله را به‌عنوان توزیع پیشین روی تمام فضای توابع توزیع تعریف کرد که موجب راحتی محاسبات و همچنین استنباط‌های دقیق‌تر می‌شود. گستردگی روش‌های محاسباتی بیزی، مانند روش‌های MCMC باعث شد تا فرآیند دیریکله به‌عنوان توزیع

[۱] Infinite Dimentional Space

[۲] Direchlet Process

پیشین در گستره وسیعی از مسائل بیزی ناپارامتری، رگرسیون ناپارامتری، برآورد توابع چگالی ناپارامتری و مدل‌های با اثرات تصادفی به‌کار برده شود. در این فصل، فرآیند دیریکله را به‌عنوان یک پیشین شناخته‌شده معرفی می‌کنیم. برای این منظور ابتدا توزیع دیریکله[۳] را معرفی می‌کنیم و سپس به معرفی فرآیند دیریکله و ویژگی‌های آن می‌پردازیم. همچنین در بخش‌های بعد ضمن معرفی مدل‌های مبتنی بر فرآیند دیریکله، با ویژگی‌ها و کاربردهای هر یک از آن‌ها آشنا می‌شویم.

۲.۲ توزیع دیریکله

توزیع دیریکله یک توزیع پیوسته است. در واقع این توزیع یک توزیع چندپارامتری تعمیم‌یافته از توزیع بتا می‌باشد. توزیع دیریکله معمولاً به‌عنوان توزیع پیشین در استنباط‌های بیزی به‌کار برده می‌شود؛ چرا که این توزیع یک پیشین مزدوج برای پارامترهای توزیع چندجمله‌ای برای داده‌های رسته‌ای است.

تعریف ۱.۲.۲ (توزیع دیریکله)**.** فرض کنید $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\}$ یک توزیع احتمال روی فضای گسسته $\boldsymbol{\chi} = \{\chi_1, \ldots, \chi_n\}$ و X یک متغیر تصادفی روی $\boldsymbol{\chi}$ باشد، به‌طوری که $P(X = \chi_i) = \theta_i$. توزیع دیریکله روی Θ با پارامترهای $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\}$ دارای تابع چگالی احتمال به‌صورت زیر است:

$$P(\boldsymbol{\Theta}|\boldsymbol{\beta}) = \frac{\Gamma(\beta_0)}{\Pi_{i=1}^{n}\Gamma(\beta_i)} \prod_{i=0}^{n} \theta_i^{\beta_i - 1}$$

که در آن $\theta_i > 0$، $\sum_{i=1}^{n} \theta_i = 1$ و همچنین $\beta_i > 0$ و $\sum_{i=1}^{n} \beta_i = \beta_0$. گشتاورهای توزیع به صورت زیر هستند:

$$E(\theta_i) = \frac{\beta_i}{\beta_0}$$

$$Var(\theta_i) = \frac{\beta_i(\beta_0 - \beta_i)}{\beta_i^2(\beta_0 + 1)}$$

$$Cov(\theta_i, \theta_j) = \frac{-\beta_i\beta_j}{\beta_i^2(\beta_0 + 1)}.$$

نکته‌ای که باید به آن توجه داشت، این است که کواریانس غیرصفر نشان‌دهنده وجود وابستگی بین داده‌های تولیدشده از این توزیع است. یک پارامتربندی جایگزین برای بردار پارامتر β به‌صورت αm_i است که در آن $\alpha = \beta_0$ پارامتر تمرکز[۴] بوده و پراگندگی توزیع حول میانگین را کنترل می‌کند و $m_i = \frac{\beta_i}{\beta_0}$ یک اندازه پایه است. در این حالت

$$P(\boldsymbol{\Theta}|\alpha m_1, \ldots, \alpha m_n) = \frac{\Gamma(\alpha)}{\Pi_{i=1}^{n}\Gamma(\alpha m_i)} \prod_{i=0}^{n} \theta_i^{\alpha m_i - 1}. \qquad (۱.۲)$$

[۳] Direchlet Distribution

[۴] Concentration Parameter

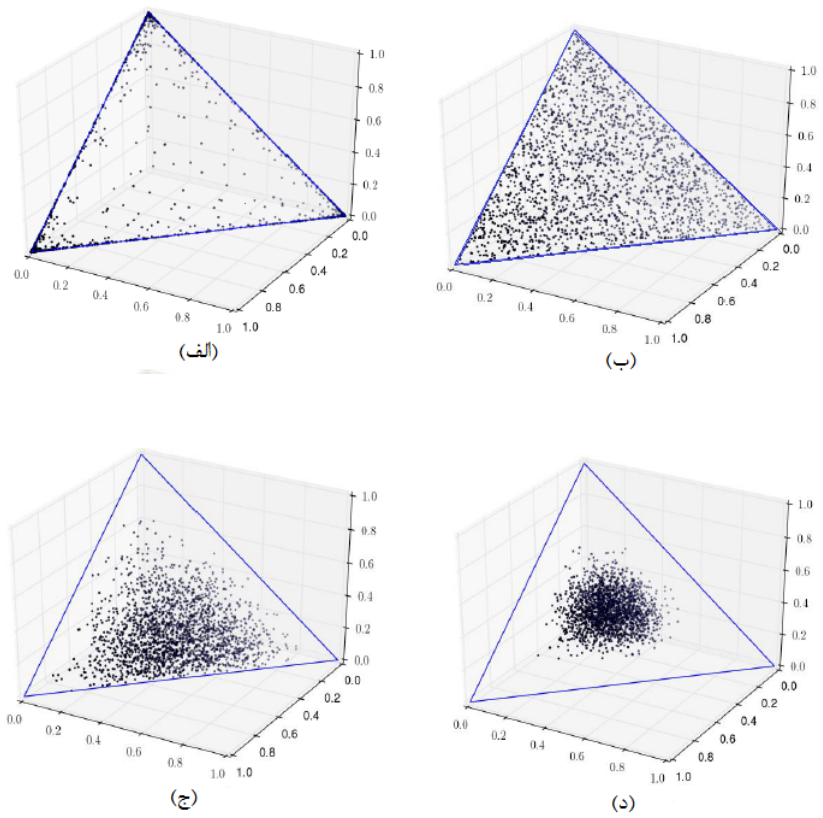
برای این توزیع از نماد $\Theta \sim Dir(\alpha M)$ استفاده می‌کنیم که در آن α پارامتر تمرکز بوده و هرچه $\alpha \longrightarrow \infty$ واریانس کاهش یافته و داده‌ها اطراف میانگین جمع می‌شوند. همچنین $M = \{m_1, \ldots, m_n\}$ با ویژگی $\sum_{i=1}^{n} m_i = 1$ است. در این حالت

$$E(\theta_i) = m_i$$

$$Var(\theta_i) = \frac{m_i(1 - m_i)}{\alpha(\alpha + 1)}.$$

با توجه به رابطه واریانس واضح است که با افزایش مقدار α، واریانس کاهش یافته و داده‌ها حول میانگین متمرکز می‌شوند.

شکل ۱.۲ بر اساس شبیه‌سازی با $n = 2000$ نمونه از یک توزیع دیریکله سه متغیره با پارامترهای تمرکز $\alpha = 0.1, 1, 4, 10$ به تصویر کشیده شده است. همان گونه که در تصویر ملاحظه می‌شود با افزایش پارامتر تمرکز یعنی همان α، داده‌ها حول میانگین (m) متمرکز می‌شوند.



شکل ۱.۲: توزیع دیریکله سه متغیره با پارامترهای تمرکز (الف) $\alpha = 0.1$، (ب) $\alpha = 1$، (ج) $\alpha = 4$ و (د) $\alpha = 10$ و $n = 2000$.

دو ویژگی معروف توزیع دیریکله عبارت‌اند از:

- برای $n = ۲$، توزیع دیریکله همان توزیع بتا است (فارو ملکلم، ۲۰۰۸).
- توزیع دیریکله یک توزیع پیشین مزدوج برای پارامترهای توزیع چندجمله‌ای است. به بیانی دیگر، فرض کنید N مشاهده $x_۱, x_۲, \ldots, x_N$ از یک توزیع گسسته با پارامتر $\boldsymbol{\Theta} = \{\theta_۱, \theta_۲, \ldots, \theta_n\}$ داشته باشیم. اگر n_i، $i = ۱, ۲, \ldots, n$، تعداد دفعاتی باشد که x_i در N مشاهده رخ می‌دهد و $\sum_{i=۱}^{n} n_i = N$، آن‌گاه با در نظر گرفتن توزیع پیشین دیریکله برای Θ، توزیع پسین آن نیز با استفاده از قانون بیز به‌صورت زیر خواهد بود:

$$\begin{aligned} P(\boldsymbol{\Theta}|\alpha, M, X_{۱:N}) &\propto \prod_{i=۱}^{n} P(X_i|\alpha, M, \boldsymbol{\Theta}) P(\boldsymbol{\Theta}|\alpha, m) \\ &\propto \prod_{i=۱}^{n} \boldsymbol{\theta}_i^{n_i} \prod_{i=۱}^{n} \boldsymbol{\theta}_i^{\alpha M_i - ۱} \\ &\propto \prod_{i=۱}^{n} \boldsymbol{\theta}_i^{\alpha M_i + n_i - ۱} \\ &\propto Dir(\alpha M_۱ + n_۱, \ldots, \alpha M_n + n_n). \end{aligned}$$

بنابراین همان‌طور که ملاحظه می‌شود، با انتخاب توزیع پیشین دیریکله، توزیع پسین نیز دیریکله است.

۱.۲.۲ نمونه‌گیری از توزیع دیریکله

یک روش تولید نمونه از توزیع دیریکله استفاده از ظرف پولیا[۵] است. برای تشریح آن ابتدا تعریف زیر را ارائه می‌دهیم.

تعریف ۲.۲.۲ (تابع دلتای دیراک). تابع دلتای دیراک[۶] یک تابع تعمیم‌یافته روی محور اعداد حقیقی است که در همه جا به جز در صفر، مقدار صفر دارد. این تابع که با حرف یونانی δ نمایش داده می‌شود، در نقطه $x = ۰$ مقداری برابر بی‌نهایت و در سایر نقاط مقداری برابر با صفر دارد و در نتیجه انتگرال آن روی بازه $(-\infty, +\infty)$ برابر با یک است. به بیانی دیگر

$$\delta(x) = \begin{cases} \infty & x = ۰ \\ ۰ & x \neq ۰ \end{cases}$$

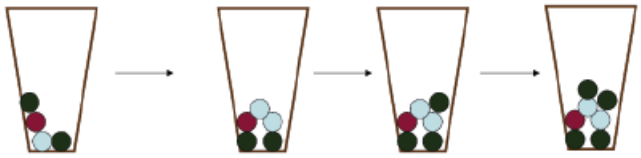
بنابراین

$$\int_{-\infty}^{+\infty} \delta(x) \mathrm{d}x = ۱$$

[۵] Polya Urn

[۶] Dirac Delta Function

فرض کنید می‌خواهیم از توزیع دیریکله $\Theta \sim Dir(\alpha M)$ نمونه تولید کنیم. برای شروع فرض کنید α گوی با رنگ‌های مختلف در ظرفی وجود داشته باشد؛ به‌طوری که αm_1 گوی در ظرف اول، αm_2 گوی در ظرف دوم و به همین ترتیب αm_n گوی در ظرف nام باشند. به‌طور تصادفی یک گوی از ظرف خارج می‌کنیم، سپس یک گوی هم‌رنگ آن به ظرف برمی‌گردانیم. بلکول و مک‌کوییین (۱۹۷۳) نشان دادند که با تکرار این روش در دفعات زیاد، نسبت رنگ هر گوی در کیسه، به توزیع دیریکله همگرا می‌شود.



شکل ۲.۲: مراحل تولید داده از توزیع دیریکله بر اساس ظرف پولیا.

اگر احتمال این‌که در مرحله i گوی به رنگ j خارج شود را با $P(\theta_i = j)$ نشان دهیم، در این صورت

$$P(\theta_1 = j) = \frac{\alpha m_j}{\sum_{i=1}^{n} \alpha m_j}$$

$$P(\theta_2 = j|\theta_1) = \frac{\alpha m_j + \delta(\theta_1 = j)}{\sum_{i=1}^{n} \alpha m_j}$$

$$\vdots$$

$$P(\theta_{N+1} = j|\theta_{1:N}) = \frac{\alpha m_j + \sum_{i=1}^{n} \delta(\theta_i = j)}{\alpha + N}$$

که در آن، N تعداد مشاهدات و δ تابع دلتای دیراک است. شکل ۲.۲ سه مرحله از فرآیند تولید داده از ظرف پولیا با $\alpha = 4$ را نشان می‌دهد.

۳.۲ فرآیند دیریکله

همان‌طور که در قبل گفته شد، از فرآیند دیریکله برای تعیین توزیع پیشین در استنباط‌های بیزی ناپارامتری استفاده می‌شود. در واقع فرآیند دیریکله یک توسیع از توزیع دیریکله در فضای پیوسته است. فرض کنید χ یک فضای نمونه، β سیگما میدان بورل مجموعه‌های χ و F فضای تمامی اندازه‌های احتمال تعریف‌شده روی (χ, β) باشد.

تعریف ۱.۳.۲ (فرآیند دیریکله). فرض کنید α یک مقدار حقیقی مثبت، G_0 یک اندازه متناهی، نامنفی و غیر صفر روی فضای اندازه‌پذیر (χ, β) و $G(\cdot)$ یک فرآیند اندیس‌گذاری شده به وسیله

عناصر β باشد. در این صورت گوییم G یک فرآیند دیریکله با پارامتر $(\alpha, G_\circ)$ است و با نماد $DP(\alpha, G_\circ)$ نمایش می‌دهیم، هرگاه دارای شرایط زیر باشد:

- $G(B)$ برای $B \in \beta$، یک متغیر تصادفی باشد؛ به‌طوری که مقادیرش را در بازه $[\circ, ۱]$ اختیار کند.
- هر تحقق از G، یک اندازه احتمال روی فضای اندازه‌پذیر (χ, β) است.
- به ازای هر تعداد افرازهای متناهی اندازه‌پذیر $B_۱, B_۲, \ldots, B_k$ برای $k = ۱, ۲, \ldots$ از χ، (B_iها اندازه‌پذیر و مجزا هستند و $\bigcup_{i=۱}^{k} B_i = \chi$) بردار تصادفی $(G(B_۱), G(B_۲), \ldots, G(B_k))$ دارای توزیع دیریکله با پارامترهای $(\alpha G_\circ(B_۱), \alpha G_\circ(B_۲), \ldots, \alpha G_\circ(B_k))$ است.

در این تعریف α را پارامتر دقت[۷] و $G_\circ$ را اندازه مرکزی[۸] یا توزیع پایه[۹] فرآیند دیریکله می‌نامند. وجود فرآیند دیریکله به‌عنوان یک فرآیند تعریف‌شده روی فضای (χ, β) از طریق قضیه سازگاری کلموگروف (کلموگروف، ۱۹۳۳) توسط فرگوسن اثبات شد. همچنین بلکول (۱۹۷۳) اثبات دیگری را برای این موضوع ارائه داد.
از آن‌جا که فرآیند دیریکله توزیعی روی اندازه احتمال تصادفی G است، برای هر $B \in \beta$، $G(B)$ یک متغیر تصادفی است. پس با توجه به تعریف فرآیند می‌توان گفت

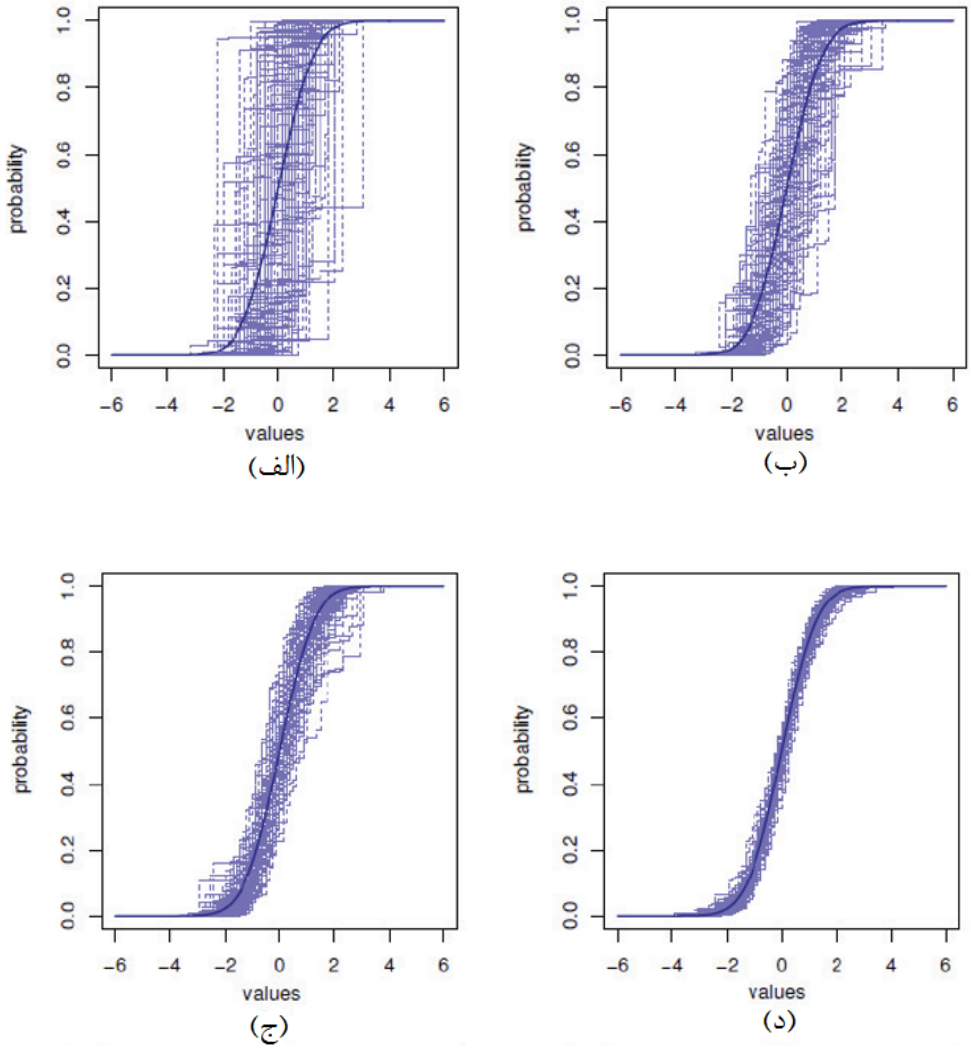
$$G(B) \sim Beta\Big(\alpha G_\circ(B)\ ,\ \alpha(۱ - G_\circ(B))\Big).$$

بنابراین

$$E\left[G(B)\right] = G_\circ(B)$$
$$Var\left[G(B)\right] = \frac{G_\circ(B)(۱ - G_\circ(B))}{۱ + \alpha}.$$

در فرآیند دیریکله پارامتر دقت یعنی همان α، میزان انحراف G از $G_\circ$ را کنترل می‌کند. یعنی هر چه قدر α افزایش یابد، G بیشتر اطراف $G_\circ$ متمرکز می‌شود. در شکل ۳.۲ یک نمونه صد تایی از فرآیند دیریکله با پارامتر $\alpha G_\circ$ تولید شده است. در این فرآیند توزیع مرکزی را نرمال استاندارد در نظر گرفته‌ایم. همچنین در این فرآیند پارامتر دقت را برای شکل‌های (الف)، (ب)، (ج) و (د) به ترتیب برابر با $\alpha = ۱$، $\alpha = ۵$، $\alpha = ۲۰$ و $\alpha = ۱۰۰$ انتخاب شده است. در شکل (الف) مقدار پارامتر دقت برابر با ۱ است که با توجه به این موضوع، فرآیند در اطراف توزیع پایه با فاصله بیشتر قرار دارد. اما در شکل (د) پارامتر دقت برابر با ۱۰۰ است، بنابراین فرآیند به سمت توزیع $G_\circ$ متمرکز شده است. این موضوع در شکل ۳.۲ به تصویر در آمده است.

---
[۷] Precision Parameter
[۸] Center Measure
[۹] Base Distribution



شکل ۳.۲: نمودار ۱۰۰ داده تولید شده از $DP(\alpha G_\circ)$ با توزیع مرکزی نرمال استاندارد و پارامتر دقت (الف) $\alpha = ۱$، (ب) $\alpha = ۵$، (ج) $\alpha = ۲۰$ و (د) $\alpha = ۱۰۰$.

نکته مهمی که باید به آن توجه داشت این است که تکیه‌گاه فرآیند دیریکله همان تکیه‌گاه توزیع مرکزی، یعنی همان $G_\circ$، است. در ادامه به بررسی برخی از ویژگی‌های فرآیند دیریکله می‌پردازیم که به‌صورت چند قضیه عنوان شده‌اند. هر تحقق از فرآیند دیریکله یک توزیع گسسته است. این موضوع، در قضیه زیر بیان می‌شود.

قضیه ۱.۳.۲ (فرگوسن، ۱۹۷۳). فرض کنید $G \sim DP(\alpha, G_\circ)$، در این صورت G تقریباً همه جا $(a.s.)$ یک اندازه احتمال گسسته است.

این قضیه به روش‌های مختلفی توسط فرگوسن (۱۹۷۳) و بلکول (۱۹۷۳) اثبات شد. فرگوسن این قضیه را از طریق یک نمایش گاما ثابت نمود. بلکول برای اثبات آن از طرح ظرف پولیا استفاده کرد. همچنین این قضیه باعث ایجاد توجهی خاص به فرآیند دیریکله به‌عنوان یک

توزیع پیشین روی توزیع‌های گسسته شد. بنابراین با فرض این‌که $X_1, X_2, \ldots, X_n$ نمونه‌هایی از یک توزیع تحقق‌یافته G از یک فرآیند دیریکله باشند، احتمال وجود X_iهای مشابه وجود دارد.

فرض کنید $n^* < n$ تعداد خوشه‌های متفاوتی باشد که مشاهدات x_i، $i = 1, 2, \ldots, n$، در آن خوشه‌ها قرار می‌گیرند. مجموعه مقادیر متفاوت با $X^* = (X_1^*, X_2^*, \ldots, X_{n^*}^*)$ نشان داده می‌شوند. همچنین فرض کنید n_j، $j = 1, 2, \ldots, n^*$ تعداد خوشه‌های هر نمونه را نمایش دهد. بلکول و مک‌کویین (۱۹۷۳) نشان دادند که برای بردار نمونه $X_1, X_2, \ldots, X_n$ از یک توزیع تحقق‌یافته فرآیند دیریکله، توزیع پیش‌گو به‌صورت زیر است:

$$P(X_{n+1}|X_1, X_2, \ldots, X_n) \propto \sum_{j=1}^{n^*} n_j \delta_{X_j^*} + \alpha G_0. \qquad (۲.۲)$$

رابطه (۲.۲) بیانگر این است که نمونه جدید با احتمال متناسب با n_j، یک نمونه مشابه نمونه قبلی است یا با احتمال متناسب با α، یک نمونه جدید از G_0 است. این رابطه با انتگرال‌گیری نسبت به توزیع تصادفی G به‌دست می‌آید. در این حالت نمونه‌ها وابسته و دارای توزیع حاشیه‌ای G_0 هستند. مدل (۲.۲) را اصطلاحاً نمایش فرآیند دیریکله از طریق ظرف پولیا می‌نامند.

قضیه ۲.۳.۲ (فرگوسن، ۱۹۷۳). فرض کنید $X_1, X_2, \ldots, X_n|G \overset{iid}{\sim} G$ و $G|\alpha G_0 \sim DP(\alpha, G_0)$، در این‌صورت توزیع پسین G به شرط $X_1, X_2, \ldots, X_n$ به‌صورت زیر است

$$(G|X_1, X_2, \ldots, X_n, \alpha, G_0) \sim DP(\alpha G_0 + \sum_{i=1}^{n} \delta_{X_i}).$$

یک نتیجه مهم این قضیه که در محاسبات بیزی کاربرد دارد به صورت زیر است:
فرض کنید $X_1, X_2, \ldots, X_n, X_{n+1}|G \overset{iid}{\sim} G$ و $G|\alpha G_0 \sim DP(\alpha, G_0)$، آن‌گاه

$$X_{n+1}|\alpha, G_0, X_1, X_2, \ldots, X_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}. \qquad (۳.۲)$$

ستورامن (۱۹۹۴) تصویر دیگری از فرآیند دیریکله را به‌صورت قضیه زیر مطرح کرد. او بر اساس این قضیه ثابت کرد که با احتمال یک، فرآیند دیریکله یک اندازه احتمال گسسته است.

قضیه ۳.۳.۲ (ستورامن، ۱۹۹۴). فرض کنید شرایط زیر را داشته باشیم

۱. برای متغیرهای تصادفی مستقل و هم‌توزیع $\nu_1, \nu_2, \ldots, \nu_n$ داشته باشیم

$$\nu_1, \nu_2, \ldots, \nu_n \overset{iid}{\sim} Beta(1, \alpha).$$

۲. برای متغیرهای تصادفی مستقل و هم‌توزیع $X_1, X_2, \ldots, X_n$ داشته باشیم:

$$X_1, X_2, \ldots, X_n \sim G_0.$$

در این صورت

$$G \overset{d}{=} \sum_{i=1}^{\infty} \pi_i \delta_{X_i}. \qquad (۴.۲)$$

مدل (۴.۲)، یک فرآیند دیریکله با پارامترهای (α, G_0) است که در آن π_iها توابعی نزولی هستند که از فرآیند شکست چوب[۱۰] به‌دست می‌آیند. در واقع، $\pi_1 = \nu_1$ و برای $i \geq 2$

$$\pi_i = \nu_i \prod_{j=1}^{i-1} (1 - \nu_j).$$

این قضیه بیان می‌کند که اندازه تصادفی G در مدل (۴.۲)، با احتمال یک، اندازه احتمالی گسسته است؛ حتی اگر G_0 یک توزیع پیوسته باشد، یک نتیجه مهم این قضیه آن است که اگر شرایط

۱. $G \sim DP(\alpha, G_0)$

۲. $X \sim G_0$

۳. $U \sim Beta(1, \alpha)$

۴. G، X و U همگی مستقل باشند.

برقرار باشند، آن‌گاه

$$U\delta_X(.) + (1 - U)G(.) \sim DP(\alpha, G_0).$$

یعنی $U\delta_X(\cdot) + (1 - U)G(\cdot)$ نیز یک فرآیند دیریکله $DP(\alpha, G_0)$ است.

۴.۲ مدل‌های مبتنی بر فرآیند دیریکله

همان‌طور که می‌دانیم مدل‌های آماری برای شناخت فرآیندی که داده‌ها از آن تولید شده‌اند کاربرد دارند. در بیشتر مدل‌ها، فرض بر این است که متغیرهای تصادفی $X_1, X_2, \ldots, X_n$ نمونه تصادفی از توزیع F هستند که F متعلق به یک رده از خانواده توزیع‌های پارامتری است. اما در عمل همیشه نمی‌توان انتظار این را داشت که مدل پارامتری برای توصیف داده‌ها مناسب باشد. در این صورت با کنار گذاشتن مدل‌های پارامتری، می‌توان مدل‌های منعطف‌تر و نیرومندتر ناپارامتری را برای تحلیل داده‌ها استفاده کرد. در چارچوب روش‌های بیزی ناپارامتری، پیشین فرآیند برای انواع مختلفی از مدل‌های آماری به‌کار برده می‌شود که به معرفی برخی از این مدل‌ها می‌پردازیم.

[۱۰] Stick Breaking

۱.۴.۲ آمیخته‌ای از فرآیندهای دیریکله

آمیخته‌ای از فرآیندهای دیریکله[۱۱] (MPD) اولین بار توسط آنتونیاک (۱۹۷۴) معرفی شد. این مدل در مسائل داده‌های سانسورشده و مدل‌های سلسله‌مراتبی که در آن یک سطح شامل فرآیند دیریکله است، کاربرد دارد. به‌طور کلی اگر توزیع پایه $G_\circ$ و پارامتر دقت α تصادفی باشد، یعنی

$$G|\eta \sim DP(\alpha_\eta, G_{\circ\eta})$$

$$\eta \sim \pi(\eta)$$

آن‌گاه G، آمیخته‌ای از فرآیندهای دیریکله با پارامتر دقت α_η، توزیع پایه $G_{\circ\eta}$ و توزیع آمیختگی $\pi(\cdot)$ است. در این حالت می‌نویسیم

$$G \sim \int DP(\alpha_\eta, G_{\circ\eta})\ \mathrm{d}\pi(\eta).$$

۲.۴.۲ فرآیند دیریکله آمیخته

اگرچه انتخاب فرآیند دیریکله به‌عنوان پیشین یک توزیع نامعلوم، مناسب است و ویژگی‌های بسیار خوبی به همراه دارد، اما هر تحقق از این فرآیند یک توزیع گسسته است و انتخاب آن به عنوان پیشینی برای توزیعی که می‌دانیم پیوسته است، چندان منطقی به نظر نمی‌رسد. یکی از روش‌هایی که برای برطرف کردن این محدودیت فرآیند دیریکله مورد استفاده قرار می‌گیرد، مدل فرآیند دیریکله آمیخته[۱۲] (DPMM) است. در این مدل، پیچشی[۱۳] از توزیع گسسته G و یک هسته پیوسته که به‌عنوان توزیع داده‌ها انتخاب می‌شود، مورد استفاده قرار می‌گیرد (فرگوسن، ۱۹۸۳؛ لو، ۱۹۸۴). فرض کنید $y_1, y_2, \ldots, y_n$ نمونه‌های تصادفی مستقل از توزیع نامعلوم F باشند. در این‌صورت پیشین فرآیند دیریکله آمیخته برای توزیع F به‌صورت زیر است:

$$y_i|G \sim \int P(y_i|\boldsymbol{\theta})\ G(d\boldsymbol{\theta})$$

$$G \sim DP(\alpha, G_\circ). \qquad (۵.۲)$$

در مدل (۵.۲)، $P(y_i|\boldsymbol{\theta})$، یک توزیع پارامتری با پارامتر با بعد متناهی $\boldsymbol{\theta}$ به‌عنوان هسته مدل آمیخته است. برای مثال، اگر فرآیند دیریکله آمیخته با هسته نرمال مبتنی بر پارامتر میانگین، μ، را بخواهیم بنویسیم، مدل به‌صورت زیر است:

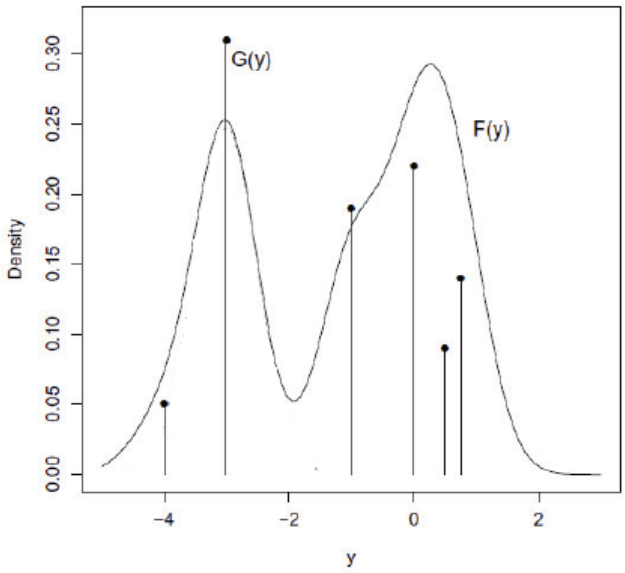
$$y_i|G, \sigma^2 \sim \int N(y_i|\mu, \sigma^2)\ G(d\mu)$$

[۱۱] Mixture of Dirichlet Processess

[۱۲] Dirichlet Process Mixture Model

[۱۳] Convolution

شکل ۴.۲ یک مدل آمیخته با هسته نرمال و $G \sim DP(\alpha, G_\circ)$ را نشان می‌دهد.



شکل ۴.۲: فرآیند دیریکله آمیخته، به‌صورت پیچشی از اندازه احتمال تصادفی G و هسته نرمال.

مدل (۵.۲) را می‌توان به‌صورت سلسله‌مراتبی

$$y_i|\boldsymbol{\theta}_i \overset{iid}{\sim} P(y_i|\boldsymbol{\theta}_i)$$

$$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n \overset{iid}{\sim} G$$

$$G \sim DP(\alpha, G_\circ) \qquad (۶.۲)$$

نوشت. مدل سلسله‌مراتبی (۶.۲) نیز بر ماهیت خوشه‌بندی $\boldsymbol{\theta}_i$های تولیدشده از اندازه احتمال گسسته G، در یک مدل آمیخته تأکید می‌کند (نیل، ۲۰۰۰). با حاشیه‌سازی نسبت به G، داریم

$$y_i|\boldsymbol{\theta}_i \overset{iid}{\sim} P(y_i|\boldsymbol{\theta}_i)$$

$$(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) \sim P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n).$$

که در آن توزیع توأم $P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n)$ با توجه به رابطه (۶.۲) به‌دست می‌آید (بلکول، ۱۹۷۳). مک‌ایچر و مولر (۱۹۹۸) با قرار دادن نمونه‌های مشابه در یک خوشه و نمونه‌های متفاوت در خوشه‌های مختلف و معرفی مجموعه مقادیر یا خوشه‌های متفاوت با $\boldsymbol{\theta}^* = (\theta^1, \theta^2, \ldots, \theta^{n^*})$ و نیز $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ یک بردار نشانگرهای پیکربندی که $\xi_i = j$ اگر و تنها اگر $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_j$، نشان دادند

$$y_i|\xi_i, \boldsymbol{\theta}^*_j \sim P(y_i|\boldsymbol{\theta}^*_{\xi_j})$$

$$\boldsymbol{\theta}^*_۱, \boldsymbol{\theta}^*_۲, \ldots, \boldsymbol{\theta}^*_{n^*} \overset{iid}{\sim} G_\circ$$

$$P(\xi_۱, \xi_۲, \ldots, \xi_n | \alpha, n^*) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^{n^*} \prod_{j=۱}^{n^*} \Gamma(n_j)$$

که در آن، n^* تعداد خوشه‌ها و $n_j = \sum_i I(\xi_i = j)$ اندازه خوشه jام را نشان می‌دهد. این شکل جدید از فرآیند دیریکله آمیخته به دلیل خارج کردن توزیع با بعد نامتناهی G و لحاظ نمودن بحث خوشه‌بندی، کاربردهای فراوانی دارد. خصوصیت خوشه‌بندی فرآیند دیریکله سبب شده است که این توزیع آمیخته در مسائلی که خوشه‌بندی داده‌ها مورد نظر است و تعداد خوشه‌ها مشخص نیست، به‌طور گسترده مورد استفاده قرار گیرد. مدل (۶.۲) را می‌توان بر اساس ساختاری که ستورامن (۱۹۹۴) برای فرآیند دیریکله معرفی کرد، به‌صورت زیر نوشت:

$$y_i | \pi_h, \boldsymbol{\theta}_h \sim \sum_{h=۱}^{\infty} \pi_h P(y_i | \boldsymbol{\theta}_h) \qquad (۷.۲)$$

به‌طوری که

$$\boldsymbol{\theta}_h \overset{iid}{\sim} G_\circ$$

$$\pi_h = \nu_h \prod_{k<h} (۱ - \nu_h)$$

$$\nu_h \overset{iid}{\sim} Beta(۱, \alpha)$$

$$G = \sum_{h=۱}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}.$$

مدل (۷.۲) به این نکته اشاره دارد که یک مدل فرآیند دیریکله آمیخته را می‌توان به‌صورت آمیخته گسسته‌ای از توزیع‌های $P(y_i|\boldsymbol{\theta}_h)$ با وزن‌های π_h در نظر گرفت. یکی از نتایج مدل فوق آن است که مدل DPMM بین مشاهدات یک خوشه‌بندی ایجاد می‌کند و α تعداد این خوشه‌ها را کنترل می‌کند. یعنی اگر $\alpha \longrightarrow \circ$، تعداد خوشه‌ها کاهش می‌یابد، به این معنی که تمامی $\boldsymbol{\theta}_i$ها رفتار مشابه دارند و مدل به یک مدل پارامتری تبدیل می‌شود که $\boldsymbol{\theta} \sim G_\circ$ و داده‌ها به‌طور تصادفی و مستقل از $P(y|\boldsymbol{\theta})$ تولید می‌شوند. در مقابل هنگامی که $\alpha \longrightarrow \infty$، تعداد خوشه‌ها بسیار زیاد می‌شود به این معنی که هر $\boldsymbol{\theta}_i$ در یک خوشه قرار می‌گیرد و رفتاری مستقل از بقیه $\boldsymbol{\theta}_i$ها دارد. بنابراین

$$y_i \overset{iid}{\sim} \int P(y_i|\boldsymbol{\theta}_i)\, G(d\boldsymbol{\theta}_i).$$

۳.۴.۲ فرآیند دیریکله متناهی

با توجه به ساختار معرفی‌شده توسط ستورامن (۱۹۹۴) شبیه‌سازی یک اندازه احتمال تصادفی G با استفاده از فرآیند دیریکله، نیازمند شبیه‌سازی تعداد نامتناهی از متغیرهای تصادفی است. این موضوع منجر به پیچیدگی‌های محاسباتی خواهد شد و عملکرد روش بیزی ناپارامتری را

تحت تأثیر قرار می‌دهد. به‌علاوه با توجه به این‌که وزن‌ها دنباله‌ای نزولی تشکیل می‌دهند، در نظر گرفتن تعداد نامتناهی از متغیرهای تصادفی غیر ضروری به نظر می‌رسد. با توجه به این مسائل، مولیر و تاردلا (۱۹۹۸) فرآیند دیریکله بریده‌شده[۱۴] (TDP) را تحت عنوان فرآیند ϵـ دیریکله به‌صورت زیر تعریف کردند

تعریف ۱.۴.۲ (فرآیند ϵـ دیریکله)**.** فرض کنید $\epsilon \in (۰, ۱)$. در این‌صورت توزیع تصادفی G^ϵ یک فرآیند ϵـ دیریکله است، هرگاه داشته باشیم

$$G^\epsilon(\cdot) = \sum_{h=۱}^{H_\epsilon} \pi_h \delta_{\boldsymbol{\theta}_h}(\cdot) + \left[۱ - \sum_{h=۱}^{H_\epsilon} \pi_h \right] \delta_{\boldsymbol{\theta}_{H_\epsilon+۱}}(\cdot) \qquad (۸.۲)$$

که در آن $\boldsymbol{\theta}_h$ به‌طور تصادفی و مستقل از توزیع پایه $G_۰$ نمونه‌گیری می‌شود. همچنین

$$\pi_h = \nu_h \prod_{k<h} (۱ - \nu_h)$$

$$\nu_h \overset{iid}{\sim} Beta(۱, \alpha)$$

$$H_\epsilon = \inf \left\{ m \in N_n : \sum_{h=۱}^{m} \pi_h \geq ۱ - \epsilon \right\}.$$

ساختار تعریف‌شده در مدل (۸.۲)، همانند ساختار فرآیند دیریکله است. اما این ساختار پس از تعداد تصادفی از جملات متوقف می‌شود و جرم احتمال باقیمانده به نقطه‌ای تصادفی در χ که به‌طور مستقل از $G_۰$ تولید شده، اختصاص می‌یابد. در تعریف فرآیند ϵـ دیریکله نقش زمان توقف، یعنی همان H_ϵ، این است که اجازه دهد احتمال تصادفی تا حد مورد نظر نزدیک به فرآیند دیریکله تولید شود. مولیر و تاردلا (۱۹۹۸) نشان دادند

$$H_\epsilon \sim \text{Poisson}(-\alpha \log \epsilon) .$$

همچنین آن‌ها نشان دادند اگر

$$G(\cdot) = \sum_{h=۱}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}$$

$$G^\epsilon(\cdot) = \sum_{h=۱}^{H_\epsilon} \pi_h \delta_{\boldsymbol{\theta}_h}(\cdot) + \left[۱ - \sum_{h=۱}^{H_\epsilon} \pi_h \right] \delta_{\boldsymbol{\theta}_{H_\epsilon+۱}}(\cdot)$$

آن‌گاه

$$\sup_B \{|G(B) - G^\epsilon(B)|\} \leq \epsilon .$$

در واقع هنگامی که $\epsilon \longrightarrow ۰$، نمونه‌های تولیدشده از فرآیند ϵـ دیریکله به فرآیند دیریکله همگرا می‌شوند. به‌عبارت دیگر، با در نظر گرفتن یک مقدار $\epsilon > ۰$ مناسب، می‌توان تقریبی مناسب از

[۱۴] Truncated Dirichlet Process

فرآیند دیریکله در اختیار داشت. لازم به ذکر است که ایشواران و جیمز (۲۰۰۱) تعریف دیگری از فرآیند دیریکله بریده‌شده را ارائه کردند. در این تعریف تقریب دیگری از فرآیند دیریکله به‌صورت

$$\sum_{i=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h} \approx \sum_{i=1}^{N} \pi_h \delta_{\boldsymbol{\theta}_h} \qquad (۹.۲)$$

به‌عنوان فرآیند دیریکله بریده‌شده معرفی شد و همچنین آن‌ها نشان دادند

$$G \sim TDP(\alpha G_0, N) .$$

در این تقریب، $\pi_N = 1 - \sum_{i=1}^{n-1} \pi_i$ پارامتر N با پارامتر α، رابطه مستقیم دارد که کنترل خوشه‌ها در بین واحدهای آزمایشی است. هنگامی که N کوچک باشد، محاسبات ساده است اما تقریب مورد نظر، خیلی دقیق نیست. رویکردی که برای انتخاب N ارائه داده‌اند بر این اساس است که مقدار N باید به‌گونه‌ای انتخاب شود که مقدار مورد انتظار π_N بسیار کوچک باشد. به‌عبارت دیگر:

$$E(\pi_N) \approx \epsilon .$$

ایشواران و جیمز (۲۰۰۱) نشان دادند

$$N \approx 1 + \frac{\log \epsilon}{\log(\frac{\alpha}{1+\alpha})}.$$

شکل ۵.۲ نمودار α در برابر N هنگامی که $\alpha = 0/01$ انتخاب شده است را نشان می‌دهد.

شکل ۵.۲: نمودار α در برابر N هنگامی که $E(\sum_{i=1}^{N-1} \pi_i) > 0/99$ برگرفته از ایشواران و جیمز (۲۰۰۱).

همان‌طور که در شکل ۵.۲ دیده می‌شود، بین پارامترهای α و N رابطه‌ای تقریباً خطی برقرار است.

۴.۴.۲ فرآیند دیریکله در ساختار مدل متغیرهای پنهان

متغیر پنهان، در مقابل متغیر آشکار، متغیرهایی هستند که به صورت مستقیم مشاهده نمی‌شوند، اما به واسطه متغیرهای دیگر که قابل مشاهده هستند، توسط یک مدل آماری در استنباط به‌کار برده می‌شوند. به‌بیانی دیگر، متغیر پنهان، متغیری است که به‌صورت مستقیم مشاهده نمی‌شود اما منبعی است که موجب تغییرات پنهان در متغیرهای آشکار مدل یا پارامترهای مدل می‌شود. به‌عنوان مثال، فرض کنید که هدف مطالعه و بررسی تأثیر یک دارو بر روی تعدادی از کودکان بیمار در یک جامعه باشد. در این جامعه علاوه بر متغیرهای آشکار مانند سن، جنسیت، و همچنین وزن که قابل اندازه‌گیری هستند، یک مجموعه از عوامل ژنتیکی در نتیجه آزمایش اثر دارند که به‌عنوان متغیر پنهان می‌توان آن‌ها را در مدل در نظر گرفت.

معمولاً فرض بر این است که توزیع پیشین متغیرهای پنهان نرمال است. این در حالی است که به دلیل عدم مشاهده این متغیرها صحت چنین فرضی کمی سؤال‌برانگیز است. برای حل این مسئله می‌توان از فرآیند دیریکله به‌عنوان یک رهیافت بیزی ناپارامتری برای مدل‌بندی متغیرهای پنهان استفاده نمود. مدل سلسله‌مراتبی زیر را در نظر بگیرید:

$$Y_i|b_i,\eta \overset{iid}{\sim} f(y_i|b_i,\eta)$$

$$b_i|G \overset{iid}{\sim} G$$

$$G \sim DP(\alpha, G_\circ). \qquad (۱۰.۲)$$

در مدل (۱۰.۲)، Y_i به شرط پارامتر η و متغیرهای پنهان b_i از خانواده توزیع پارامتری f آمده‌اند. بردار η، بردار پارامترهای مدل و b_iها متغیرهای پنهان هستند که از توزیع نامعلوم G تولید شده‌اند. G تحققی از یک فرآیند دیریکله با پارامترهای α و $G_\circ$ است.

بلکول و مک‌کویین (۱۹۷۳) نشان دادند که با حاشیه‌سازی و انتگرال‌گیری روی G در مدل (۱۰.۲)، b_iها از توزیع شرطی زیر تبعیت می‌کنند:

$$b_i|b_1, b_2, \ldots, b_{n-1}, \alpha, G_0 \sim \frac{1}{\alpha+i-1}\sum_{h=1}^{i-1}\delta_{b_h} + \frac{\alpha}{\alpha+i-1}G_0. \qquad (11.2)$$

مدل (۱۱.۲) در حقیقت بیانگر این موضوع است که اثرات تصادفی با احتمال $\frac{1}{\alpha+i-1}$ مشابه یکی از مقادیر قبلی است و با احتمال $\frac{\alpha}{\alpha+i-1}$ یک مقدار جدید از G_0 را اختیار می‌کند. همان‌گونه که مطرح شد، استفاده از این ساختار به دلیل عدم وجود توزیع با بعد متناهی G، سادگی محاسبات را به دنبال دارد. اما مفهوم خوشه‌بندی به‌عنوان یک ویژگی مهم فرآیند دیریکله را شکل نمی‌دهد. بنابراین مشابه آن‌چه که در قبل بیان شد، با استفاده از ویژگی خوشه‌بندی فرآیند دیریکله، می‌توان یک مدل ساده‌تر و کاراتر ارائه نمود (جردن و همکاران، ۲۰۰۶).

یکی از کاربردهای مدل متغیرهای پنهان، در تحلیل داده‌های دودویی است. در این چارچوب، داده‌های دودویی به یک متغیر پنهان پیوسته ارتباط داده می‌شوند. یعنی با استفاده از یک مقدار آستانه متغیر پیوسته در دو رده گسسته شده و متغیر پاسخ معرفی می‌شود. به‌طور کلی در اغلب متون تحقیقات انجام‌شده، فرض می‌شود توزیع پیشین متغیر پنهان نرمال است و به دنبال آن مدل به‌دست آمده مدل پرابیت نامیده می‌شود. همان‌گونه که اشاره شد ممکن است چنین فرضی برقرار نباشد. برای حل این مسئله، رهیافت‌های متفاوتی معرفی شده‌اند. یک رهیافت پیشنهادشده بر مبنای اتخاذ رهیافت بیزی نیمه‌پارامتری[۱۵] است که در آن پارامترهای مدل متغیر پنهان به دو دسته تقسیم می‌شوند و برای حل بعضی از آن‌ها روش پارامتری و برای بقیه رویکردی ناپارامتری بر اساس فرآیند دیریکله انتخاب می‌شود. به‌این ترتیب یک مدل پرابیت تعمیم‌یافته با متغیر پنهان از فرآیند دیریکله آمیخته شکل می‌گیرد. در مدل‌های آمیخته خطی تعمیم‌یافته[۱۶] نیز برای مدل‌بندی اثرات تصادفی اغلب از توزیع نرمال استفاده می‌شود، اما ممکن است این فرض برقرار نباشد. به همین دلیل استفاده از پیشین فرآیند دیریکله برای توزیع نامعلوم اثرات تصادفی، منجر به استنباط دقیق‌تر و کارایی بیشتر مدل می‌شود.

تا این‌جا فرآیند دیریکله و ویژگی‌های آن به‌عنوان یک پیشین مناسب در مدل‌های بیزی ناپارامتری معرفی شد. همان‌گونه که بیان شد، این فرآیند ضمن این‌که به‌عنوان یک پیشین مزدوج در سرعت بخشیدن به محاسبات بیزی نقش به سزایی ایفا می‌کند، با ویژگی خوشه‌بندی داده‌ها ناهمگنی بین آن‌ها را کنترل می‌کند و این امر منجر به بهبود و کارایی مدل می‌شود.

[۱۵] Bayesian Semiparametric Model
[۱۶] Generalized Linear Mixed Models

این ویژگی‌های فرآیند دیریکله در سال‌های اخیر موجب محبوبیت آن در گستره وسیعی از مسائل پزشکی مانند پردازش تصاویر و اپیدمیولوژی شده است. یکی دیگر از کاربردهایی که می‌توان از این فرآیند در نظر داشت، استفاده از این فرآیند در مدل‌بندی بیزی ناپارامتری داده‌های فضایی است.

۵.۴.۲ رگرسیون بیزی ناپارامتری

در سال‌های اخیر مدل‌های آماری مختلفی بر مبنای روش‌های بیزی ناپارامتری مورد بررسی قرار گرفته‌اند. دراین قسمت روش بیزی ناپارامتری برای یک مدل رگرسیونی را شرح می‌دهیم. یک مدل رگرسیونی با متغیر پاسخ y_i و متغیرهای تبیینی x_i را به‌صورت زیر در نظر بگیرید:

$$y_i = f(x_i) + \epsilon_i \qquad i = ۱, ۲, \ldots, n$$

$$\epsilon_i \sim P(\epsilon_i).$$

در مدل‌های رگرسیون پارامتری، توزیع خطا و $f(x_i)$، پارامتری در نظر گرفته می‌شوند. اما در مدل‌های رگرسیون بیزی ناپارامتری، آن‌ها را ناپارامتری در نظر می‌گیرند که منتهی به سه حالت زیر می‌شود:

۱. مدل خطای ناپارامتری: در این مدل فرض می‌شود که ϵ_i دارای توزیع تصادفی G است، یعنی $\epsilon_i \sim G$، که G دارای پیشین $\pi(G)$ است و $f(x_i)$ یک تابع پارامتری در نظر گرفته می‌شود.

۲. مدل تابع میانگین ناپارامتری: در این مدل برای تابع میانگین توزیع تصادفی در نظر گرفته می‌شود و پیشینی برای این توزیع انتخاب می‌شود. یعنی

$$f \sim \pi(f)$$

که در آن $\pi(f)$ یک توزیع پیشین است. در واقع پیشین مرسوم برای f، فرآیندهای گاوسی[۱۷] است (مولر، ۲۰۱۳).

۳. رگرسیون ناپارامتری کامل: در این مدل هم تابع میانگین و هم توزیع خطاها تصادفی در نظر گرفته می‌شوند و برای آن‌ها توزیع‌های پیشین معرفی می‌شود.

[۱۷] Gaussian Process

همان‌گونه که مطرح شد، شیوه‌های مختلفی برای تعریف پیشین برای توزیع تصادفی وجود دارند. اما باید روشی را اتخاذ کنیم که با توجه به پیچیدگی محاسبات در روش بیزی ناپارامتری، دارای این محدودیت نباشد. در واقع استفاده از فرآیند دیریکله جهت تعریف پیشین روی توزیع تصادفی، روشی است که موجب راحتی محاسبات و همچنین استنباط دقیق و کارا می‌شود. مدل‌های آماری دیگری نیز بر اساس روش‌های بیزی ناپارامتری معرفی شده‌اند.

برای دیدن جزئیات و مثال‌های بیشتر پیرامون روش‌های بیزی ناپارامتری، می‌توان به مولر و میترا (۲۰۱۳) مراجعه کرد.

فصل ۳

مدل رگرسیون چگالی بیزی

۱.۳ بیان مسئله

در این فصل مدل بیزی رگرسیون چگالی را با در نظر گرفتن این موضوع که متغیر پاسخ از نوع گسسته است، معرفی خواهیم کرد. مدل رگرسیونی زیر را در نظر بگیرید:

$$y = f(\mathbf{x}) + \epsilon \ .$$

در واقع هدف اصلی برآورد کردن تابع جرم احتمال شرطی متغیر پاسخ y و بردار متغیرهای $\mathbf{x}$، یعنی برآورد $f(y|\mathbf{x})$ است. این برآورد به ما اجازه می‌دهد تا چگونگی تغییر توزیع متغیر پاسخ با حضور متغیرهای تبیینی را مطالعه کنیم. لازم به ذکر است که با برآورد کردن چگالی شرطی $f(y|\mathbf{x})$ مقادیر مورد علاقه دیگری مانند میانگین شرطی، میانه و چندک‌ها نیز به‌دست می‌آیند. رهیافتی را که برای برآورد چگالی شرطی $f(y|\mathbf{x})$ به‌کار خواهیم گرفت، بر مبنای برآورد چگالی چندمتغیره $f(y, \mathbf{x})$ است که متغیرهای پاسخ به‌صورت مقیاس آمیخته[۱] (پیوسته ـ گسسته) هستند. همان‌طور که می‌دانید تابع چگالی شرطی $f(y|\mathbf{x})$ به‌صورت زیر به‌دست می‌آید:

$$f(y|\mathbf{x}) = \frac{f(y, \mathbf{x})}{\int f(y, \mathbf{x}) \mathrm{d}y}. \qquad (۱.۳)$$

[۱] Mixed Scale

استفاده از مدل‌های آمیخته یک رهیافت بسیار مطلوب برای برآورد چگالی شرطی را در دسترس قرار می‌دهد. شکل کلی مدل آمیخته چگالی توأم $(y, \mathbf{x})$ به‌صورت زیر است:

$$f_p(y, \mathbf{x}) = \int k(y, \mathbf{x}; \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) \qquad (۲.۳)$$

که در آن، $k(\cdot, \cdot; \cdot)$ یک هسته احتمال با بردار پارامتر $\boldsymbol{\theta}$ و $P(\cdot)$ یک اندازه احتمال روی فضای پارامتر Θ است. بنابراین تعیین مدل اصلی، نیازمند تعیین توزیع پیشین برای اندازه احتمال $P(\cdot)$ و نیز هسته $k(\cdot, \cdot; \cdot)$ است. تعریف پیشین برای پارامترهای $k(\cdot, \cdot; \cdot)$ به سادگی انجام می‌شود، اما در تعریف پیشین برای اندازه احتمال $P(\cdot)$ روش‌های متفاوتی وجود دارند. روش‌هایی مانند فرآیند دیریکله، رستوران چینی[۲] و درخت پولیا[۳] . در این پایان‌نامه، از رهیافت بیزی ناپارامتری فرآیند دیریکله که در فصل دوم معرفی گردید، استفاده شده است. اکنون با توجه به تعریف فرآیند دیریکله، رابطه (۲.۳) را می‌توان به‌صورت زیر بازنویسی کرد:

$$f_p(y, \mathbf{x}) = \sum_{h=۱}^{\infty} \pi_h(\mathbf{x}) k(y, \mathbf{x}; \boldsymbol{\theta})$$

که این رابطه همان فرآیند دیریکله برای مدل‌های آمیخته (DPMM) است. همچنین مدل آمیخته فرآیند دیریکله برای تابع چگالی شرطی رابطه (۱.۳) به شکل زیر است:

$$f_p(y|\mathbf{x}) = \sum_{h=۱}^{\infty} \pi_h(\mathbf{x}) k(y|\mathbf{x}; \boldsymbol{\theta})$$

که در آن

$$\pi_h(\mathbf{x}) = \frac{\pi_h g(\mathbf{x}; \boldsymbol{\theta})}{f_p(\mathbf{x})}.$$

این رهیافت بیانگر این موضوع است که چگونه یک مدل منعطف از رگرسیون چگالی، به‌وسیله فرآیند دیریکله مدل‌های آمیخته نتیجه می‌شود (دانسون و همکاران، ۲۰۰۷).
هدف اصلی در این‌جا، تأکید روی مدل $f_p(y|\mathbf{x}) = \sum_{h=۱}^{\infty} \pi_h(\mathbf{x}) k(y|\mathbf{x}; \boldsymbol{\theta})$ است که در آن متغیر پاسخ y گسسته و بردار متغیرهای تبیینی $\mathbf{x}$ گسسته و پیوسته هستند. ایده پیش رو این است که هسته $k(\cdot, \cdot; \cdot)$ با فرض این‌که متغیرهای گسسته، نسخه گسسته‌شده متغیرهای پیوسته پنهان[۴] هستند، به‌دست آورده می‌شود (موثن، ۱۹۸۴). سپس یک هسته گوسی چندمتغیره برای همه متغیرها در نظر گرفته می‌شود. گسسته کردن متغیرهای پیوسته پنهان و مدل‌بندی آن‌ها با استفاده از نقاط برش[۵] انجام می‌شود. سپس با استفاده از مدل آمیخته فرآیند دیریکله، مدل نهایی به‌دست می‌آید.

---

[۲] Chinese Restaurant
[۳] Polya Tree
[۴] Latent Continuous Variables
[۵] Cut-Points

۲.۳ تعیین مدل بیزی

فرض کنید Y متغیر پاسخ گسسته، با تکیه‌گاه اعداد صحیح نامنفی باشد. همچنین متغیر تبیینی $\mathbf{X} = (\mathbf{X}_d^T, \mathbf{X}_c^T)^T$ که شامل تعداد p متغیر گسسته و پیوسته است را در نظر بگیرید که در آن $\mathbf{X}_d$ برداری شامل متغیرهای تبیینی گسسته با بعد p_d و $\mathbf{X}_c$ برداری شامل متغیرهای تبیینی پیوسته با بعد p_c می‌باشد. در این حالت $p_d + p_c = p$. هدف اصلی در این قسمت یافتن مدل مشترک $f \in \mathcal{F}$ حاصل از $Z = (Y, \mathbf{X}^T)^T$ است که $\mathcal{F}$ بیانگر مجموعه همه چگالی‌ها است. رهیافت پیش‌رو برای مدل‌بندی متغیرهای گسسته استفاده از متغیرهای پنهان است. در واقع متغیرهای گسسته شامل متغیرهای پاسخ و نیز متغیرهای تبیینی، فرض می‌شود که گسسته‌سازی شده‌اند (موثن، ۱۹۸۴). با توجه به الگوریتمی که در زیر مشاهده می‌شود متغیر پیوسته پنهان، گسسته‌سازی می‌شود.

اگر متغیر گسسته را با Z و متغیر پیوسته پنهان را با Z^* نمایش دهیم قاعده گسسته‌سازی به شکل زیر است:

$Z = z$ اگر و فقط اگر

$$z^* \in R(z) = (c_{z-1}, c_z), \quad z = 0, 1, 2, \ldots \tag{۳.۳}$$

که در را بطه (۳.۳) $R(z)$ فاصله بین نقاط برش را مشخص می‌کند. همچنین $c_{-1} = -\infty$ و برای $q \geq 0$ داریم

$$c_q = c_q(\boldsymbol{\lambda}) = \Phi^{-1}\{F(., \boldsymbol{\lambda})\}$$

که در آن، $\Phi(\cdot)$ تابع توزیع تجمعی نرمال استاندارد و $F(\cdot, \boldsymbol{\lambda})$ توزیعی است که متناسب با داده‌ها انتخاب می‌شود. فرض کنید متغیرهای پنهان به‌طور مستقل از توزیع $N(0, 1)$ هستند. بر همین اساس و با توجه به قضیه زیر، به راحتی ثابت می‌شود که توزیع حاشیه‌ای Z، $F(z; \boldsymbol{\lambda})$ است. یعنی

$$P(Z \leq z) = P(Z^* < c_z) = P(Z^* < \Phi^{-1}\{F(z, \boldsymbol{\lambda})\}) = P\{\Phi(Z^*) < F(z; \boldsymbol{\lambda})\} = F(z; \boldsymbol{\lambda}).$$

قضیه ۱.۲.۳ (قضیه تبدیل انتگرال احتمال)**.** اگر متغیر تصادفی X دارای توزیع پیوسته $F(x)$ باشد، آن‌گاه متغیر تصادفی $U = F(X)$ دارای توزیع $U(0, 1)$ است.

برهان. چون $F(x) = P(X \leq x)$ تابعی غیرنزولی است و $0 \leq F(x) \leq 1$، با توجه به پیوسته بودن $F(x)$ داریم

$$F_U(u) = P(U \leq u) = P(F(X) \leq u) = \begin{cases} 0 & u \leq 0 \\ 1 & u \geq 1 \end{cases}$$

برای $0 < u < 1$ فرض می‌کنیم

$$z = \sup \{x : F(x) = u\} .$$

البته اگر F اکیداً صعودی باشد، دارای وارون F^{-1} بوده و می‌توانیم بنویسیم

$$z = F^{-1}(u).$$

حال با توجه به پیوسته بودن $F(x)$، دو پیشامد $x \leq z$ و $F(x) \leq u$ دارای احتمال‌های مساوی هستند. بنابراین

$$F_U(u) = P(U \leq u) = P(F(X) \leq u) = P(X \leq z) = F(z) = u .$$

□

همان‌طور که اشاره شد، مدل‌بندی نقاط برش بر اساس تابع توزیع $F(\cdot, \boldsymbol{\lambda})$ صورت می‌پذیرد. اما انتخاب تابع توزیع نیز خود مسئله مهمی است. بسته به این‌که داده‌ها از نوع شمارشی، دوجمله‌ای یا دودویی باشند، می‌توان توزیع $F(\cdot, \boldsymbol{\lambda})$ مناسبی در نظر گرفت.

اگر داده‌ها از نوع شمارشی باشند، برای مدل کردن $F(., \boldsymbol{\lambda})$، از توزیع $\text{Poisson}(H, \xi)$ استفاده می‌کنیم. پارامترهای این توزیع به صورت $\boldsymbol{\lambda} = (H, \xi)^T$ است که ξ همان پارامتر توزیع پواسن و H پارامتر بی‌اثر است. انتخاب این توزیع محدودیت‌هایی دارد. از جمله این محدودیت‌ها، عدم پوشش مفاهیمی چون بیش‌پراکنش[۶] و کم‌پراکنش[۷] در داده‌های شمارشی است. برای محاسبه عامل بیش‌پراکنش، به جای توزیع پواسن، می‌توان از توزیع $\text{NB}(H, \xi_1, \xi_2)$ استفاده کرد که به دلیل برخورداری از دو پارامتر، انعطاف‌پذیری بیشتری دارد. همچنین برای پوشش دادن هر دو موضوع بیش‌پراکنش و کم‌پراکنش، از توزیع‌های زیر استفاده می‌کنیم:

- پواسن تعمیم‌یافته
- توزیع ابرپواسن
- توزیع $COM-$پواسن
- توزیع پیرسون سه‌پارامتری مختلط

این توزیع‌ها به دلیل وجود تعداد پارامترهای بیشتر نسبت به توزیع پواسن، بسیار منعطف‌تر عمل می‌کنند. برای مدل‌بندی داده‌های دوجمله‌ای، توزیع $F(\cdot, \boldsymbol{\lambda})$ را، توزیع $\text{Binimial}(H, \xi)$ انتخاب می‌کنیم که در این توزیع، H تعداد آزمایش‌ها و ξ احتمال موفقیت را نشان می‌دهند. یک توزیع مناسب‌تر برای $F(\cdot, \boldsymbol{\lambda})$ استفاده از توزیع $\text{Beta Binomial}(H, \xi_1, \xi_2)$ به جای توزیع دوجمله‌ای است. این توزیع با دارا بودن بردار پارامتر $\boldsymbol{\lambda} = (H, \xi_1, \xi_2)^T$، بسیار منعطف‌تر است.

[۶] Overdispersion
[۷] Underdispersion

دسته دیگر از داده‌ها، داده‌های دودویی هستند. این داده‌ها حالتی خاص از داده‌های دوجمله‌ای با $H = ۱$ هستند که به‌صورت زیر مدل‌بندی می‌شوند:

$$Z = ۰$$

اگر و فقط اگر $z^* < ۰$، به‌طوری که

$$z^* \sim N(\mu_z^*, ۱).$$

با این رهیافت، تولید نمونه از توزیع پسین راحت‌تر خواهد بود. جدول ۱.۳ توزیع‌های معرفی‌شده و پیشین‌های مناسب برای پارامتر آن‌ها را نشان می‌دهد.

جدول ۱.۳: توزیع‌های مختلف برای مدل‌بندی نقاط برش و ویژگی‌های آن‌ها برگرفته از پاپاجئورجیو و ریچاردسون (۲۰۱۶).

هسته توزیع	تابع احتمال	مؤلفه بی‌اثر	میانگین	پارامتر
پواسن	$\exp\left\{\frac{(-H\lambda)(H\lambda)^y}{y!}\right\}$	H	$H\lambda$	λ
پواسن تعمیم‌یافته	$\mu\{\mu+(\phi-۱)y\}^{y-۱}\phi^{-y}$ $\times\exp\left\{\frac{\left(\frac{-[\mu+(\phi-۱)y]}{\phi}\right)}{y!}\right\}$	H	$H\mu$	μ, ϕ
ابرپواسن	$\frac{۱}{{}_۱F_۱(۱,\gamma,\lambda)}\frac{\lambda^y}{(\gamma)_y}$	H	$\lambda-(\gamma-۱)$ $\times\frac{{}_۱F_۱(۱,\gamma,\lambda)-۱}{{}_۱F_۱(۱,\gamma,\lambda)} = H\beta_۰$	$\beta_۰, \gamma$
$COM-$پواسن	$\frac{\lambda^y}{Z(\lambda,\nu)(y!)^\nu}$	H	$\lambda\frac{d\log(Z)}{d\lambda} = H\beta_۰$	$\beta_۰, \nu$
CTP	$f_۰\frac{(\alpha+\beta_i)_y(\alpha-\beta_i)_y}{(\gamma)_y y!}$	H	$\frac{\alpha^۲+\beta^۲}{\gamma-۲\alpha-۱} = H\beta_۰$	$\beta_۰, \gamma, \alpha$
دوجمله‌ای	$\binom{N}{y}\pi^y(۱-\pi)^{N-y}$	N	$N\pi$	π
دوجمله‌ای منفی	$\frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+۱)}(\frac{\beta}{H+\beta})^\alpha(\frac{H}{H+\beta})^y$	H	$H\frac{\alpha}{\beta}$	α, β
بتا دوجمله‌ای	$\binom{N}{y}\frac{\text{Beta}(y+\alpha,N-y+\beta)}{\text{Beta}(\alpha,\beta)}$	N	$\frac{N\alpha}{(\alpha+\beta)}$	α, β

اکنون برای مدل‌بندی تابع چگالی، ابتدا تعریف زیر را ارائه می‌دهیم.

تعریف ۱.۲.۳ (کانل و دانسون، ۲۰۱۵). فرض کنید $A^{(j)} = \left\{A_۱^{(j)}, A_۲^{(j)}, \ldots, A_q^{(j)}\right\}$ یک افراز از $\mathbb{R}$ به مجموعه‌های دوبه‌دو ناسازگار باشد. برای هر $h < l$، $A_h^{(j)}$ را قبل از $A_l^{(j)}$ قرار دهید. همچنین فرض کنید $A_{y_۲} = \left\{y_۲^* : y_{۲,j}^* \in A_{y_۲}^{(j)}, j = ۱, \ldots, p_۲\right\}$. در این صورت چگالی توأم f برابر است با

$$f(y) = g(f^*) = \int_{A_{y_۲}} f^*(y^*)\mathrm{d}y^*$$

که در آن، $g : \mathcal{F}^* \longrightarrow \mathcal{F}$ یک نگاشت از فضای تمام چگالی‌های $\mathcal{F}^*$ که با اندازه لبگ روی $\mathbb{R}^p$ رابطه دارند به فضای تمام چگالی‌های توأم $\mathcal{F}$ است.

حال فرض کنید $\mathbf{Z} = (Y, \mathbf{X}_d^T, \mathbf{X}_c^T)^T$، پیشین Π روی $\mathcal{F}$ به‌صورت زیر تعریف می‌شود:

$$f_p(z) = \int k(z|\boldsymbol{\theta})\ \mathrm{d}P(\boldsymbol{\theta}) \qquad (۴.۳)$$

که در آن، $k(z|\boldsymbol{\theta})$، یک هسته پارامتری و $P(.)$، توزیع آمیخته تصادفی است. اگر هسته $k(z|\boldsymbol{\theta})$ را یک هسته گاوسی با بعد q، برای $\mathbf{Z}^* = (Y^*, (\mathbf{X}_d^*)^T, \mathbf{X}_c^T)^T$ در نظر بگیریم، آن‌گاه طبق تعریف ۱.۲.۳ و رابطه (۴.۳) داریم

$$k(z|\boldsymbol{\theta}) = \int_{R(y)} \int_{R(x_d)} N_q(z^*|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\ \mathrm{d}y^* \mathrm{d}\mathbf{x}_d^*. \qquad (۵.۳)$$

بنابراین $q = ۱ + p$. همچنین پارامترهای توزیع هسته به‌صورت $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ است.
به دلیل این‌که برخی از پارامترهای مکان و نیز پارامترهای مقیاس توزیع مربوط به متغیرهای پنهان (به استثنای پارامترهای مکان و مقیاس متغیرهای پنهان مربوط به متغیرهای دودویی) تشخیص‌ناپذیر هستند، میانگین $\boldsymbol{\mu}^*$ و کواریانس $\boldsymbol{\Sigma}^*$ به‌صورت زیر در نظر گرفته می‌شوند:

$$\boldsymbol{\mu}^* = \begin{pmatrix} ۰ \\ \boldsymbol{\mu} \end{pmatrix} \qquad \boldsymbol{\Sigma}^* = \begin{pmatrix} \mathbf{C} & \boldsymbol{\nu}^T \\ \boldsymbol{\nu} & \boldsymbol{\Sigma} \end{pmatrix}$$

که در آن $\mathbf{C}$، ماتریس کواریانس مربوط به متغیرهای پنهان است و عناصر روی قطر اصلی آن برابر یک هستند. لذا ماتریس همبستگی است. همچنین $\boldsymbol{\Sigma}$ ماتریس کواریانس نامحدود شامل متغیرهای پیوسته مشاهده‌شده است.

در چارچوب روش‌های بیزی ناپارامتری (تادی و کوتاس، ۲۰۱۰؛ دانسون و باتاچاریا، ۲۰۱۱) و به‌کار بردن ساختار شکست‌ـچوب معرفی شده توسط ستورامن (۱۹۹۴) برای فرآیند دیریکله، داریم

$$P(\cdot) = \sum_{h=۱}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}(\cdot) \qquad (۶.۳)$$

با ترکیب روابط (۴.۳) و (۶.۳) مدل DPMM برای متغیر $\mathbf{z}^* = (y^*, (\mathbf{x}_d^*)^T, \mathbf{x}_c^T)^T$ به‌صورت زیر حاصل می‌شود:

$$f_p(\mathbf{z}) = \sum_{h=۱}^{\infty} \pi_h k(z|\boldsymbol{\theta}_h). \qquad (۷.۳)$$

در رابطه (۷.۳) وزن‌های $\pi_h, h \geq ۱$، با استفاده از فرآیند شکست‌ـچوب به‌دست می‌آیند. یعنی

$$\begin{aligned} \pi_۱ &= \nu_۱ \\ \pi_l &= \nu_l \prod_{h=۱}^{l-۱} (۱ - \nu_h) \quad l \geq ۲ \\ \nu_k &\sim Beta(۱, \alpha) \quad k \geq ۱\ . \end{aligned}$$

توزیع پایه $G_۰$ در رابطه (۷.۳) با تعیین توزیع‌های پیشین برای عناصر $\boldsymbol{\theta}_h = (\boldsymbol{\xi}_h, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h^*)$، $h \geq ۱$ تعریف می‌شود.

نخست توزیع پیشین $\boldsymbol{\mu}_h$ که قسمت تشخیص‌پذیر $\boldsymbol{\mu}_h^*$ است را مشخص می‌کنیم. در واقع این پیشین، یک توزیع نرمال چندمتغیره با بردار میانگین برابر با میانگین نمونه مربوط به متغیرهای توضیحی و ماتریس کواریانس قطری با واریانس‌های بزرگ است.

پیشین مربوط به پارامتر $\{\xi_k\}$، بسته به این‌که چه توزیعی برای $F(\cdot, \boldsymbol{\lambda})$ در نظر گرفته شود، از جدول ۲.۳ انتخاب می‌کنیم:

جدول ۲.۳: توزیع پیشن برای پارامترهای توزیع $F(\cdot;\cdot,\cdot)$.

هسته	پیشین
پواسن	$\lambda \sim \text{Gamma}(۱, ۱)$
پواسن تعمیم‌یافته	$\lambda \sim \text{Gamma}(۱, ۱)$
	$\phi \sim N(۱, ۱)\ I\left[\phi > \max\left\{۰/۵, \frac{۱-\mu}{۴}\right\}\right]$
ابرپواسن	$\beta_۰ \sim \text{Gamma}(۱, ۱)$
	$\gamma \sim \text{Gamma}(۱, ۱)\ I(\gamma > ۰/۱)$
$COM-$پواسن	$\beta_۰ \sim \text{Gamma}(۱, ۱)$
	$\nu \sim \text{Gamma}(۱, ۱)\ I\left[\nu > \max\left\{۰/۳, \frac{۱}{(۱+۲\beta_۰)}\right\}\right]$
CTP	$\beta_۰, \gamma \sim \text{Gamma}(۱, ۱)$
	$\alpha \sim N(۰, ۱۰۰)\ I\left[\alpha < \frac{\gamma}{۲-۱}\right]$
دوجمله‌ای	$\pi \sim \text{Beta}(۱, ۱)$
دوجمله‌ای منفی	$\alpha, \beta \sim \text{Gamma}(۱, ۱)$
بتا دوجمله‌ای	$\alpha, \beta \sim \text{Gamma}(۱, ۱)$

در نهایت توزیع پیشین مربوط به ماتریس کواریانس $\boldsymbol{\Sigma}_h^*$ بر اساس روش ارائه‌شده توسط ژانگ و همکاران (۲۰۰۶) و بارنارد و همکاران (۲۰۰۰) معرفی می‌شود.
با اضافه کردن پارامترهای واریانس به مدل، که آن‌ها به وسیله داده‌ها تشخیص‌ناپذیراند، باید پارامترهای تشخیص‌پذیر و تشخیص‌ناپذیر از یکدیگر تفکیک شوند. برای این منظور باید از ماتریس همبستگی به جای ماتریس کواریانس استفاده کنیم. در واقع باید $\boldsymbol{\Sigma}_h^*$ ماتریس همبستگی باشد.

ابتدا برای ماتریس کواریانس $\mathbf{E}_h$ توزیع پیشین $\text{Wishart}_q(\mathbf{E}_h; \eta, \mathbf{H})$ را معرفی می‌کنیم. در این حالت

$$P(\mathbf{E}_h|\eta, \mathbf{H}) \propto |\mathbf{E}_h|^{\frac{\eta-q-1}{2}} \text{etr}\left\{\frac{-\mathbf{H}^{-1}\mathbf{E}_h}{2}\right\} \quad h \geq 1$$

که در آن

$$\text{etr}\{\cdot\} = \exp\{\text{tr}(\cdot)\}$$

و منظور از $\text{tr}(A)$، اثر ماتریس A است. همچنین

$$\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12} & \mathbf{H}_{22} \end{pmatrix}$$

که در آن، $\mathbf{H}_{11}$ ماتریس همبستگی $(1+p_d)\times(1+p_d)$، $\mathbf{H}_{22}$ ماتریس کواریانس $p_c \times p_c$ و $\mathbf{H}_{12}$ ماتریس حاصل از کواریانس‌ها با بعد $(1+p_d \times p_c)$ است.

اکنون با استفاده از تجزیه طیفی ماتریس $\mathbf{E}_h$ را به‌صورت زیر تجزیه می‌کنیم:

$$\mathbf{E}_h = \mathbf{D}_h^{\frac{1}{2}} \, \mathbf{\Sigma}_h^* \, \mathbf{D}_h^{\frac{1}{2}}$$

به‌طوری که

$$\mathbf{D}_h = \text{Diag}(d_{h,1}^2, d_{h,2}^2, \ldots, d_{h,1-p_d}^2, 1, \ldots, 1)$$

یک ماتریس قطری شامل $(1+p_d)$ پارامتر واریانس تشخیص‌ناپذیر و تعداد p_c، یک است که مربوط به واریانس‌های تشخیص‌پذیر است. بنابراین $\mathbf{\Sigma}_h^*$ ماتریس همبستگی است که محدودیت تشخیص‌پذیری پارامترها را برطرف می‌کند. ژاکوبین تبدیل نیز به‌صورت زیر محاسبه می‌شود:

$$J(\mathbf{E}_h \longrightarrow \mathbf{D}_h, \mathbf{\Sigma}_h^*) = \prod_{j=1}^{(1+p_d)} d_{hj}^{q-1} = |\mathbf{D}_h|^{\frac{q-1}{2}}.$$

بنابراین توزیع پیشین برای $(\mathbf{D}_h, \mathbf{\Sigma}_h^*)$ به‌صورت زیر تعریف می‌شود:

$$P(\mathbf{D}_h, \mathbf{\Sigma}_h^*|\eta, \mathbf{D}) \propto |\mathbf{E}_h|^{\frac{\eta-q-1}{2}} \text{etr}\left\{\frac{-\mathbf{H}^{-1}\mathbf{E}_h}{2}\right\} J(\mathbf{E}_h \longrightarrow \mathbf{D}_h, \mathbf{\Sigma}_h^*) \quad h \geq 1$$

که در آن η برابر با q است و $\mathbf{H}$ باید قطری باشد.

۳.۳ توزیع پسین مدل رگرسیون چگالی بیزی

در بخش ۲.۳ با در نظر گرفتن ساختار معرفی‌شده توسط ستورامن (۱۹۹۴) برای فرآیند دیریکله، مدل زیر نتیجه شد:

$$f_p(\mathbf{z}) = \sum_{h=1}^{\infty} \pi_h k(z|\boldsymbol{\theta}_h).$$

همان‌طور که ملاحظه می‌شود حد بالای سیگما تا بی‌نهایت است و تولید نمونه از آن عملی نیست. بنابراین باید مجموع را تا یک مقدار شمارا متناهی کنیم. با متناهی کردن مدل و در نظر گرفتن پارامتر تخصیص δ برای T مولفه، تابع درستنمایی به‌صورت زیر تبدیل می‌شود:

$$\mathbf{z}|\boldsymbol{\theta}, \delta = l \ \sim \ k(\mathbf{z}|\boldsymbol{\theta}_l)$$

$$P(\delta = l|\alpha) = \pi_l \ , \quad l = ۱, ۲, \ldots, T$$

$$\ell(\boldsymbol{\theta}, \alpha|\mathbf{z}, \boldsymbol{\delta}) = \ell(\boldsymbol{\theta}, \alpha|\mathbf{z}_i, \delta_i = l_i \ , i = ۱, ۲, \ldots, n) = \prod_i k(\mathbf{z}_i|\boldsymbol{\theta}_{l_i})\pi_{l_i}. \qquad (۸.۳)$$

می‌دانیم $\mathbf{z}_i$ شامل متغیر پاسخ $\mathbf{y}_i$، متغیرهای تبیینی گسسته $\mathbf{x}_{d,i} = (x_{d,i,۱}, \ldots, x_{d,i,p_d})^T$ و همچنین متغیرهای تبیینی پیوسته $\mathbf{x}_{c,i} = (x_{c,i,۱}, \ldots, x_{c,i,p_c})^T$ است که $i = ۱, ۲, \ldots, n$. با افزودن متغیرهای پنهان y_i^* و $\mathbf{x}_{d,i}^* = (x_{d,i,۱}^*, \ldots, x_{d,i,p_d}^*)^T$ به رابطه (۸.۳)، تابع درستنمایی به‌صورت زیر به‌دست می‌آید:

$$\ell(\boldsymbol{\theta}, \alpha|\mathbf{z}, \boldsymbol{\theta}, y^*, \mathbf{x}_d^*) =$$

$$\prod_i \left\{ I[y^* \in R(y_i)] \left[\prod_{m=۱}^{p_d} I[\mathbf{x}_{d,i,m}^* \in R(\mathbf{x}_{d,i,m})] \right] N_q(y_i^*, \mathbf{x}_{d,i}^*, \mathbf{x}_{c,i}|\boldsymbol{\mu}_{l_i}^*, \boldsymbol{\Sigma}_{l_i}^*)\pi_{l_i} \right\} \qquad (۹.۳)$$

برای تولید نمونه از مدل قبل، به جای متناهی کردن می‌توان از نمونه‌گیری لایه‌ای[۸] استفاده کرد (والکر، ۲۰۰۷؛ پاپاسپیلیوپولوس، ۲۰۰۸). این کار با اضافه کردن متغیر u_i، $i = ۱, \ldots, n$، که دارای توزیع $U(۰, ۱)$ هستند، انجام می‌گیرد. با استفاده از این روش، تابع درستنمایی به صورت زیر حاصل می‌شود:

$$\prod_i \Big\{ I[y^* \in R(y_i)] \left[\prod_{m=۱}^{p_d} I[\mathbf{x}_{d,i,m}^* \in R(\mathbf{x}_{d,i,m})] \right] N_q(y_i^*, \mathbf{x}_{d,i}^*, \mathbf{x}_{c,i}|\boldsymbol{\mu}_{l_i}^*, \boldsymbol{\Sigma}_{l_i}^*) \\ I\left[u_i < \pi_{l_i}\right] \Big\} \qquad (۱۰.۳)$$

بنابراین توزیع پسین به‌صورت زیر خواهد بود:

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{y}^*, \alpha|\mathbf{y}, \mathbf{x}) &\propto g_۱(\mathbf{y}, \mathbf{x}_d|\mathbf{y}^*, \mathbf{x}_d^*, \boldsymbol{\delta}, \boldsymbol{\theta}) \ g_۲(\mathbf{y}^*, \mathbf{x}_d^*, \mathbf{x}_c|\boldsymbol{\delta}, \boldsymbol{\theta}) \ g_۳(\boldsymbol{\delta}|\alpha) \ g_۰(\boldsymbol{\theta}, \alpha) \\ &\propto \ell(\mathbf{z}, \boldsymbol{\delta}, \mathbf{y}^*, \mathbf{x}_d^*) \ g_۰(\boldsymbol{\theta}, \alpha) \end{aligned} \qquad (۱۱.۳)$$

که در آن $g_۱(\cdot)$ توزیع پیشین مربوط به متغیرهای گسسته، $g_۲(\cdot)$ توزیع پیشین مربوط به متغیرهای پیوسته، $g_۳(\cdot)$ توزیع پیشین مربوط به پارامتر تخصیص δ و $g_۰(\cdot)$ توزیع پیشین مربوط به پارامتر هسته $k(\cdot, \cdot)$ در رابطه (۷.۳) است.

[۸] Slice Sampling

۴.۳ برازش مدل با استفاده از نمونه‌گیری MCMC

همان‌طور که می‌دانیم تمامی استنباط‌ها در دیدگاه بیزی بر اساس توزیع پسین صورت می‌گیرد. فرض کنید $\Pi(.|\mathbf{D})$ توزیع پسین روی $\mathcal{F}$ و ϕ مجموعه همه پارامترها باشد. با استفاده از ساختار معرفی‌شده برای فرآیند دیریکله در رابطه (۳.۲)، توزیع پسین به‌صورت زیر است:

$$f(\mathbf{z}|\boldsymbol{\phi}) = \frac{1}{n+\alpha}\sum_{h} n_h k(\mathbf{z}|\boldsymbol{\theta}_h) + \frac{\alpha}{n+\alpha}\int k(\mathbf{z}|\boldsymbol{\theta})\mathrm{d}G_\circ(\boldsymbol{\theta}). \qquad (۱۲.۳)$$

حال فرض کنید ϕ_i ، $i = 1, \ldots, S$ نمونه‌هایی از توزیع پسین باشند. در این‌صورت تابع چگالی پیش‌گو به‌صورت زیر برآورد می‌شود:

$$\begin{aligned} f(\mathbf{z}|\mathbf{D}) &= \int f(\mathbf{z}|\boldsymbol{\phi})\Pi(\boldsymbol{\phi}|\mathbf{D})\mathrm{d}\phi \\ &= E_{\phi|D}\left[f(\mathbf{z}|\boldsymbol{\phi})|\mathbf{D}\right] \\ &= \frac{1}{S}\sum_{i=1}^{S} f(\mathbf{z}|\boldsymbol{\phi}_i) \end{aligned} \qquad (۱۳.۳)$$

در ادامه نحوه به‌دست آوردن چگالی شرطی $f(y|\mathbf{x},\phi)$ و نیز استنباط‌هایی پیرامون آن را شرح می‌دهیم.

تابع چگالی شرطی $f(y|\mathbf{x},\phi)$ را می‌توان به‌صورت زیر نوشت:

$$f(y|\mathbf{x},\phi) = \frac{f(\mathbf{z}|\phi)}{f(\mathbf{x}|\phi)}$$

که در آن $f(\mathbf{z}|\phi)$ با استفاده از رابطه (۱۲.۳) به‌دست می‌آید و $f(\mathbf{x}|\phi)$ نیز مشابه $f(\mathbf{z}|\phi)$ به‌دست می‌آید با این تفاوت که به جای چگالی توأم $\mathbf{z}$ از چگالی حاشیه‌ای $\mathbf{x}$ استفاده می‌کنیم. برای هر نمونه از این توزیع شرطی، تابعی‌های[۹] مورد نظر مانند میانگین، میانه و چندک‌ها نیز قابل محاسبه هستند.

همان گونه که ملاحضه می‌شود، توزیع پسین پارامترها با در نظر گرفتن پیشین‌های مورد نظر فرم بسته‌ای ندارد. یکی از روش‌های متداول برای تقریب این گونه توزیع‌های پیچیده، استفاده از الگوریتم‌های MCMC است. اگر بخواهیم از الگوریتم نمونه‌گیر گیبز برای تولید نمونه استفاده کنیم، باید توزیع‌های شرطی کامل را بدانیم که در ادامه نحوه به‌دست آوردن آن‌ها را بیان می‌کنیم.

[۹] Functionals

۱. توزیع شرطی کامل ν_h، بتا است. درواقع

$$\nu_h|\ldots \sim \text{Beta}(n_h+۱, n-\sum_{l=۱}^{h} n_h+\alpha)$$

که در آن n_h تعداد مشاهداتی است که در خوشه hام قرار می‌گیرد.

برهان.

$$\begin{aligned} f(\nu_h|\ldots) &\propto \prod_i \left\{ \nu_{l_i} \prod_{h=۱}^{l_i-۱}(۱-\nu_h) \right\} \nu_h^{۱-۱}(۱-\nu_h)^{\alpha-۱} \\ &\propto \nu_h^{n_h}(۱-\nu_h)^{n-\sum_{l=۱}^{h} n_l}(۱-\nu_h)^{\alpha-۱} \\ &= \nu_h^{n_h+۱-۱}(۱-\nu_h)^{n-\sum_{l=۱}^{h} n_l+\alpha-۱} \\ &= \text{Beta}(n_h+۱, n-\sum_{l=۱}^{h} n_h+\alpha). \end{aligned}$$

□

۲. توزیع شرطی کامل توأم $(\mathbf{D}_h, \mathbf{\Sigma}_h^*)$ به‌صورت زیر است:

$$f(\mathbf{D}_h, \mathbf{\Sigma}_h^*|\ldots) \propto |\mathbf{D}_h|^{\frac{\eta}{۲}-۱}|\mathbf{\Sigma}_h^*|^{\frac{\eta-q-۱-n_h}{۲}} \text{etr}\left\{ -\frac{\mathbf{H}^{-۱}\mathbf{E}_h + \mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h}{۲} \right\}$$

که در آن

$$\mathbf{S}_h = \sum_{i:\delta_i=h} (\mathbf{z}_i^* - \boldsymbol{\mu}_h^*)(\mathbf{z}_i^* - \boldsymbol{\mu}_h^*)^T \qquad \text{etr}(\cdot) = \exp(tr(\cdot))$$

برهان.

$$\begin{aligned} f(\mathbf{D}_h, \mathbf{\Sigma}_h^*|\ldots) &\propto |\mathbf{E}_h|^{\frac{\eta-q-۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h)} \text{e}^{-\frac{۱}{۲}tr\left(\sum_{i:\delta_i=h}(\mathbf{z}_i^*-\boldsymbol{\mu}_h^*)^T\mathbf{\Sigma}_h^{*^{-۱}}(\mathbf{z}_i^*-\boldsymbol{\mu}_h^*)\right)} \\ &= |\mathbf{E}_h|^{\frac{\eta-q-۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h)} \text{e}^{-\frac{۱}{۲}tr\left(\mathbf{\Sigma}_h^{*^{-۱}}\sum_{i:\delta_i=h}(\mathbf{z}_i^*-\boldsymbol{\mu}_h^*)(\mathbf{z}_i^*-\boldsymbol{\mu}_h^*)^T\right)} \\ &= |\mathbf{E}_h|^{\frac{\eta-q-۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h+\mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h)} \\ &= |\mathbf{D}_h^{\frac{۱}{۲}}\ \mathbf{\Sigma}_h^*\ \mathbf{D}_h^{\frac{۱}{۲}}|^{\frac{\eta-q-۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h+\mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h)} \text{J}(\mathbf{E}_h \longrightarrow \mathbf{D}_h, \mathbf{\Sigma}_h^*) \\ &= |\mathbf{D}_h^{\frac{۱}{۲}}\ \mathbf{\Sigma}_h^*\ \mathbf{D}_h^{\frac{۱}{۲}}|^{\frac{\eta-q-۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h+\mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h)} |\mathbf{D}_h|^{\frac{q-۱}{۲}} \\ &= |\mathbf{D}_h^{\frac{۱}{۲}}\ \mathbf{\Sigma}_h^*\ \mathbf{D}_h^{\frac{۱}{۲}}|^{\frac{\eta}{۲}-\frac{q}{۲}-\frac{۱}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h+\mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h)} |\mathbf{D}_h|^{\frac{q}{۲}-\frac{۱}{۲}} \\ &= |\mathbf{D}_h|^{\frac{\eta}{۲}-۱} \quad |\mathbf{\Sigma}_h^*|^{\frac{\eta-q-۱-n_h}{۲}} \text{e}^{-\frac{۱}{۲}tr(\mathbf{H}^{-۱}\mathbf{E}_h+\mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h)} \\ &= |\mathbf{D}_h|^{\frac{\eta}{۲}-۱} \quad |\mathbf{\Sigma}_h^*|^{\frac{\eta-q-۱-n_h}{۲}} \text{etr}\left\{ -\frac{\mathbf{H}^{-۱}\mathbf{E}_h + \mathbf{\Sigma}_h^{*^{-۱}}\mathbf{S}_h}{۲} \right\} \end{aligned}$$

□

همان‌گونه که ملاحظه می‌شود این توزیع شرطی کامل شکل شناخته‌شده‌ای ندارد. بنابراین برای تولید نمونه از این توزیع از روش متروپلیس‌ـ هستینگز استفاده می‌کنیم. در این روش توزیع پیشنهادی برای $\mathbf{E}_h$ را به‌صورت زیر در نظر می‌گیریم:

$$\mathbf{E}_h^{(p)} \sim \text{Wishart}_s(\mathbf{E}_h^{(p)}; \psi, \frac{\mathbf{E}_h^{(t)}}{\psi}) \qquad \mathbf{E}_h^{(t)} = \mathbf{D}_h^{(t)^{\frac{1}{2}}} \mathbf{\Sigma}_h^{*(t)} \mathbf{D}_h^{(t)^{\frac{1}{2}}}$$

که در آن $\mathbf{\Sigma}_h^{*(t)}$ و $\mathbf{D}_h^{(t)}$ تحقق‌هایی هستند که در تکرار قبلی الگوریتم MH به‌دست آمده اند. فرض می‌شود که مقادیر پیشنهادی برای $\mathbf{\Sigma}_h^*$ و $\mathbf{D}_h$ با تجزیه $\mathbf{E}_h^{(p)} = \mathbf{D}_h^{(p)^{\frac{1}{2}}} \mathbf{\Sigma}_h^{*(p)} \mathbf{D}_h^{(p)^{\frac{1}{2}}}$ و با احتمال پذیرش زیر تولید می‌شوند:

$$\alpha = \min \left\{ \frac{f(\mathbf{D}_h^{(p)}, \mathbf{\Sigma}_h^{*(p)}|\ldots)}{f(\mathbf{D}_h^{(t)}, \mathbf{\Sigma}_h^{*(t)}|\ldots)} \frac{t(\mathbf{D}_h^{(t)}, \mathbf{\Sigma}_h^{*(t)}|\mathbf{D}_h^{(p)}, \mathbf{\Sigma}_h^{*(p)})}{t(\mathbf{D}_h^{(p)}, \mathbf{\Sigma}_h^{*(p)}|\mathbf{D}_h^{(t)}, \mathbf{\Sigma}_h^{*(t)})} \ , \ ۱ \right\} .$$

چگالی پیشنهادی با در نظر گرفتن

$$t(\mathbf{D}_h^{(p)}, \mathbf{\Sigma}_h^{*(p)}|\mathbf{D}_h^{(t)}, \mathbf{\Sigma}_h^{*(t)}) = Wishart_s(\mathbf{E}_h^{(p)}; \psi, \frac{\mathbf{E}_h^{(t)}}{\psi}) J(\mathbf{E}_h^{(p)} \longrightarrow \mathbf{D}_h^{(p)}, \mathbf{\Sigma}_h^{*(p)})$$

به‌دست می‌آید. پارامتر ψ، پارامتری است که در بازه $[۰/۲, ۰/۲۵]$ انتخاب می‌شود (رابرتز و روزنتال، ۲۰۰۱).

۳. در این قسمت چگونگی تولید نمونه از $\boldsymbol{\mu}_h$، $h \geq ۱$ را شرح می‌دهیم. همان‌طور که به یاد دارید $\boldsymbol{\mu}_h^*$ با بعد q به صورت زیر است:

$$\boldsymbol{\mu}_h^* = \begin{pmatrix} ۰ \\ \boldsymbol{\mu}_h \end{pmatrix}$$

که قسمت صفر آن مربوط به متغیرهای گسسته است که با $z_۱$ نمایش داده می‌شوند و بعد آن‌ها $p_۱$ است. همچنین قسمت $\boldsymbol{\mu}_h$ آن مربوط به میانگین متغیرهای دودویی و پیوسته است که با $z_۲$ نشان داده می‌شوند و بعد آن‌ها برابر $p_۲$ است. همچنین $p_۱ + p_۲ = q$. بنابراین توزیع $(z_۱, z_۲)$ برابر است با

$$(z_۱, z_۲)|(\boldsymbol{\mu}_h^*, \mathbf{\Sigma}_h^*) \sim N_q \left(\boldsymbol{\mu}_h^* = \begin{bmatrix} ۰ \\ \boldsymbol{\mu}_h \end{bmatrix} \quad , \quad \mathbf{\Sigma}_h^* = \begin{bmatrix} \mathbf{\Sigma}_{۱۱h} & \mathbf{\Sigma}_{۱۲h} \\ \mathbf{\Sigma}_{۱۲h} & \mathbf{\Sigma}_{۲۲h} \end{bmatrix} \right)$$

پس

$$f(\boldsymbol{\mu}_h|\ldots) \propto \prod_{i:l_i=h} \left\{ N_{p_۲}(\mathbf{z}_{۲_i}|\boldsymbol{\mu}_h + \mathbf{\Sigma}_{۱۲h}\mathbf{\Sigma}_{۲۲h}^{-۱}\mathbf{z}_{۱_i} \ , \ \mathbf{\Sigma}_{۱۱h} - \mathbf{\Sigma}_{۱۲h}\mathbf{\Sigma}_{۲۲h}^{-۱}\mathbf{\Sigma}_{۱۲h}^T) \right\} g_۰(\boldsymbol{\mu}_h) \ .$$

برهان. با در نظر گرفتن توزیع پیشین $N_{p_۲}(\bar{\mathbf{x}}, \mathbf{D})$ برای $g_۰(\boldsymbol{\mu}_h)$ که در آن، $\bar{\mathbf{x}}$ میانگین متغیرهای تبیینی و $\mathbf{D}$ ماتریس قطری متناظر با متغیرهای تبیینی است (ریچاردسون

و گرین، ۱۹۹۷)، داریم

$$\begin{aligned}
f(\boldsymbol{\mu}_h|\ldots) \;\propto\; & \exp\Big\{-\frac{1}{2}\sum_{i-1}^{n}\Big[(\mathbf{z}_{2i}-\boldsymbol{\mu}_h-\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\mathbf{z}_{1i})'\mathbf{W}_h^{-1}(\mathbf{z}_{2i}-\boldsymbol{\mu}_h \\
& -\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\mathbf{z}_{1i})\Big]\Big\}\times\exp\Big\{-\frac{1}{2}(\boldsymbol{\mu}_h-\bar{x})'\mathbf{D}^{-1}(\boldsymbol{\mu}_h-\bar{\mathbf{x}})\Big\} \\
\propto\; & \exp\Big\{-\frac{1}{2}\Big[\sum_{i-1}^{n}(-\mathbf{z}_{2i}\mathbf{W}_h^{-1}\boldsymbol{\mu}_h-\boldsymbol{\mu}_h'\mathbf{W}_h^{-1}\mathbf{z}_{2i}+\boldsymbol{\mu}_h'\mathbf{W}_h^{-1}\boldsymbol{\mu}_h \\
& +\boldsymbol{\mu}_h'\mathbf{W}_h^{-1}\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\mathbf{z}_{1i}+\mathbf{z}_{1i}'\boldsymbol{\Sigma}_{22h}^{-1}\boldsymbol{\Sigma}_{12h}\mathbf{W}_h^{-1}\boldsymbol{\mu}_h) \\
& +\boldsymbol{\mu}_h'\mathbf{D}^{-1}\boldsymbol{\mu}_h-\boldsymbol{\mu}_h'\mathbf{D}^{-1}\bar{\mathbf{x}}-\bar{\mathbf{x}}'\mathbf{D}^{-1}\boldsymbol{\mu}_h\Big]\Big\} \\
\propto\; & \exp\Big\{-\frac{1}{2}\Big[2(-\mathbf{W}_h^{-1}\sum_{i=1}^{n}\mathbf{z}_{2i})\boldsymbol{\mu}_h+2\boldsymbol{\mu}_h'(\mathbf{W}_h^{-1}\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\sum_{i=1}^{n}\mathbf{z}_{1i}) \\
& -2\boldsymbol{\mu}_h(\bar{\mathbf{x}}'\mathbf{D}^{-1})+\boldsymbol{\mu}_h'(\mathbf{D}^{-1})\boldsymbol{\mu}_h+\boldsymbol{\mu}_h'n_h\mathbf{W}_h^{-1}\boldsymbol{\mu}_h\Big]\Big\} \\
\propto\; & \exp\Big\{-\frac{1}{2}\Big[2(-\mathbf{W}_h^{-1}\sum_{i=1}^{n}\mathbf{z}_{2i})\boldsymbol{\mu}_h+2\boldsymbol{\mu}_h'(\mathbf{W}_h^{-1}\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\sum_{i=1}^{n}\mathbf{z}_{1i}) \\
& -2\boldsymbol{\mu}_h(\bar{\mathbf{x}}'\mathbf{D}^{-1})+\boldsymbol{\mu}_h'(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})\boldsymbol{\mu}_h\Big]\Big\} \\
\propto\; & \exp\Big\{(\boldsymbol{\mu}_h-A)'(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})\Big\} \\
\propto\; & \exp\Big\{-\boldsymbol{\mu}_h'(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}A-A'(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})\boldsymbol{\mu}_h\Big\} \\
\propto\; & \exp\Big\{-2\boldsymbol{\mu}_h'(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})A\Big\}
\end{aligned}$$

که در آن A به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}
A = & (\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}(\mathbf{D}^{-1}\bar{\mathbf{x}})+(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}\mathbf{W}_h^{-1}\sum_{i=1}^{n}\mathbf{z}_{2i} \\
& -(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}(\mathbf{W}_h^{-1}\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\sum_{i=1}^{n}\mathbf{z}_{1i}) \\
= & (\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}\Big\{\mathbf{W}_h^{-1}\Big[\sum_{i=1}^{n}(\mathbf{z}_{2i}-\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\mathbf{z}_{1i})\Big]+\mathbf{D}^{-1}\bar{\mathbf{x}}\Big\}.
\end{aligned}$$

بنابراین

$$\boldsymbol{\mu}_h|\ldots \;\sim\; N_{p_2}\left(A,(\mathbf{D}^{-1}+n_h\mathbf{W}_h^{-1})^{-1}\right)$$

که در آن

$$\mathbf{W}_h^{-1}=\boldsymbol{\Sigma}_{12h}-\boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\boldsymbol{\Sigma}_{12h}^{T}\,.$$

□

۴. توزیع شرطی $\boldsymbol{\xi}_h$ عبارت است از

$$f(\boldsymbol{\xi}_h|\ldots) \propto \prod_{i:\delta_i=h} \left[\Phi(\frac{c_{y_i}(\lambda_i) - E_i^*}{sd_i^*}) - \Phi(\frac{c_{y_i-1}(\lambda_i) - E_i^*}{sd_i^*})\right] g_0(\boldsymbol{\xi}_h)$$

برهان.

$$\begin{aligned} P(Z = c_{y_i}(\lambda_i)) &= P(Z \le c_{y_i}(\lambda_i)) - P(Z \le c_{y_i-1}(\lambda_i)) \\ &= P(\frac{c_{y_i}(\lambda_i) - E_i^*}{sd_i^*}) - P(\frac{c_{y_i-1}(\lambda_i) - E_i^*}{sd_i^*}) \\ &= \left[\Phi(\frac{c_{y_i}(\lambda_i) - E_i^*}{sd_i^*}) - \Phi(\frac{c_{y_i-1}(\lambda_i) - E_i^*}{sd_i^*})\right] \end{aligned}$$

که در آن

$$E_i^* \equiv E(\mathbf{y}_i^*|\mathbf{x}_{d_i}^*, \mathbf{x}_c)$$

$$sd_i^* \equiv \sqrt{\mathrm{Var}(\mathbf{y}_i^*|\mathbf{x}_{d_i}^*, \mathbf{x}_c)}.$$

همچنین $g_0(\boldsymbol{\xi}_h)$ تابع چگالی توزیع پیشین برای $\boldsymbol{\xi}$ است که طبق جدول ۲.۳ تعیین می‌شود. □

با در نظر گرفتن پیشین‌های مناسب، چگالی‌های به‌دست آمده شکل شناخته شده‌ای نخواهند داشت. بنابراین برای تولید نمونه از آن‌ها باید از الگوریتم متروپلیس ـ هستینگز استفاده کنیم. در زیر به برخی از این توزیع‌ها اشاره می‌کنیم.

الف. اگر $F(\cdot|\boldsymbol{\lambda})$ پواسن انتخاب شود، چگالی زیر را برای $\boldsymbol{\xi}_h^{(p)}$ در نظر می‌گیریم (رابرتز و روزنتال، ۲۰۰۱):

$$\boldsymbol{\xi}_h^{(p)} \sim \mathrm{Gamma}(\tau\boldsymbol{\xi}_h^{(t)}\boldsymbol{\xi}_h^{(t)}, \tau\boldsymbol{\xi}_h^{(t)})$$

که در آن

$$E(\boldsymbol{\xi}_h^{(p)}) = \boldsymbol{\xi}_h^{(t)}$$

$$Var(\boldsymbol{\xi}_h^{(p)}) = \frac{1}{\tau}.$$

همچنین احتمال پذیرش برابر است با

$$\min\left\{1, \frac{f(\boldsymbol{\xi}_h^{(p)}|\ldots)\mathrm{Gamma}(\boldsymbol{\xi}_h^{(t)}; \tau\boldsymbol{\xi}_h^{(p)}\boldsymbol{\xi}_h^{(p)}, \tau\boldsymbol{\xi}_h^{(p)})}{f(\boldsymbol{\xi}_h^{(t)}|\ldots)\mathrm{Gamma}(\boldsymbol{\xi}_h^{(p)}; \tau\boldsymbol{\xi}_h^{(t)}\boldsymbol{\xi}_h^{(t)}, \tau\boldsymbol{\xi}_h^{(t)})}\right\}.$$

ب. اگر $F(\cdot|\boldsymbol{\lambda})$ دوجمله‌ای انتخاب شود، $\boldsymbol{\xi}_h^{(p)}$ به‌صورت زیر است:

$$\boldsymbol{\xi}_h^{(p)} \sim \mathrm{Beta}(\boldsymbol{\xi}_h^{(p)}; a^{(t)}, b^{(t)})$$

که در آن

$$a^{(t)} = b^{(t)}\frac{\boldsymbol{\xi}_h^{(t)}}{1 - \boldsymbol{\xi}_h^{(t)}}$$

$$b^{(t)} = \boldsymbol{\xi}_h^{(p)} - ۱ + \boldsymbol{\xi}_h^{(t)}(۱ - \boldsymbol{\xi}_h^{(t)})^۲\tau_۰$$

همچنین

$$E(\boldsymbol{\xi}_h^{(p)}) = \boldsymbol{\xi}_h^{(t)}$$
$$Var(\boldsymbol{\xi}_h^{(p)}) = \frac{۱}{\tau}.$$

احتمال پذیرش در این حالت نیز برابر است با

$$\min\left\{۱\ ,\ \frac{f(\boldsymbol{\xi}_h^{(p)}|\ldots)\text{Gamma}(\boldsymbol{\xi}_h^{(t)};\tau\boldsymbol{\xi}_h^{(p)}\boldsymbol{\xi}_h^{(p)},\tau\boldsymbol{\xi}_h^{(p)})}{f(\boldsymbol{\xi}_h^{(t)}|\ldots)\text{Gamma}(\boldsymbol{\xi}_h^{(p)};\tau\boldsymbol{\xi}_h^{(t)}\boldsymbol{\xi}_h^{(t)},\tau\boldsymbol{\xi}_h^{(t)})}\right\}.$$

ج. اگر توزیع دوجمله‌ای منفی را برای $F(\cdot|\boldsymbol{\lambda})$ در نظر بگیریم، در این صورت پارامتر به‌صورت $\boldsymbol{\xi}_h = (\xi_{۱h}, \xi_{۲h})^T$ است که تولید نمونه از آن در یک مرحله انجام می‌شود. توزیع پیشنهادی برای تولید نمونه از $\xi_{۱h}^{(p)}$ و $\xi_{۲h}^{(p)}$ از بردار پارامتر $\boldsymbol{\xi}_h^{(p)} = (\xi_{۲h}^{(p)}, \xi_{۲h}^{(p)})$ همانند قسمت (الف) برابر است با

$$\boldsymbol{\xi}_h^{(p)} \ \sim\ \text{Gamma}(\tau\boldsymbol{\xi}_h^{(t)}\boldsymbol{\xi}_h^{(t)})\ ,\ \tau\boldsymbol{\xi}_h^{(t)}).$$

۵. متغیرهای پنهان y_i^* و $\mathbf{x}_{d,i}^* = (x_{d,i,۱}^*, \ldots, x_{d,i,p_d}^*)^T$، $i = ۱,۲,\ldots,n$، با استفاده از تابع چگالی شرطی زیر تولید می‌شوند:

(۱۴.۳)

$$(y_i^*, \mathbf{x}_{d,i}^*)|\mathbf{x}_{c,i} \sim N_{۱+p_d}(\boldsymbol{\mu}_{۱h}+\boldsymbol{\Sigma}_{۱۲h}^*(\boldsymbol{\Sigma}_{۲۲h}^*)^{-۱}(\mathbf{x}_{c,i}-\boldsymbol{\mu}_{۲h})\ ,\ \boldsymbol{\Sigma}_{۱۱h}^*-\boldsymbol{\Sigma}_{۱۲h}^*(\boldsymbol{\Sigma}_{۲۲h}^*)^{-۱}(\boldsymbol{\Sigma}_{۱۲h}^*)^T)$$

برهان. همان‌طور که می‌دانیم اگر

$$(x_{p_۱}, y_{p_۲}) \sim N_p\left(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{xy} & \Sigma_{yy}\end{bmatrix}\right)$$

آن‌گاه تابع چگالی شرطی $x|y$ به‌شکل زیر است:

$$x|y \sim N_{p_۱}\left(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-۱}(y-\mu_y) \quad , \quad \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-۱}\Sigma_{xy}^T\right)$$

بنابراین مشابه استدلال بالا داریم

$$(y_i^*, \mathbf{x}_{d,i}^*)|\mathbf{x}_{c,i} \sim N_{۱+p_d}(\boldsymbol{\mu}_{۱h}+\boldsymbol{\Sigma}_{۱۲h}^*(\boldsymbol{\Sigma}_{۲۲h}^*)^{-۱}(\mathbf{x}_{c,i}-\boldsymbol{\mu}_{۲h})\ ,\ \boldsymbol{\Sigma}_{۱۱h}^*-\boldsymbol{\Sigma}_{۱۲h}^*(\boldsymbol{\Sigma}_{۲۲h}^*)^{-۱}(\boldsymbol{\Sigma}_{۱۲h}^*)^T)_۰$$

□

۶. برای تولید نمونه از چگالی شرطی کامل δ_i، $i = ۱,۲,\ldots,n$، از روش زیر استفاده می‌کنیم (پاپاجئورجیو و ریچاردسون، ۲۰۱۶):

$$P(\delta_i = h) \propto \int\int\cdots\int g(y_i^*\mathbf{x}_{d,i}^*|\mathbf{x}_{c,i})\mathrm{d}y_i^*\mathrm{d}\mathbf{x}_{d,i}^* N_{p_c}(\mathbf{x}_{c,i}|\boldsymbol{\mu}_{۲h},\boldsymbol{\Sigma}_{۲۲h}^*)\pi_{hi}$$

که در آن $g(y_i^*\mathbf{x}_{d,i}^*|\mathbf{x}_{c,i})$ تابع چگالی نشان داده شده در (۱۴.۳) است.

۷. حرکت تعویض برچسب[۱۰] (پاپاسپیلیپولوس و رابرتز، ۲۰۰۸).

الف. دو خوشه ناتهی به نام‌های a و b به تصادف انتخاب می‌کنیم و سپس نام آن‌ها را عوض می‌کنیم. احتمال پذیرش این حرکت برابر است با

$$\min\left\{۱\ ,\ (\frac{\pi_b}{\pi_a})^{n_a-n_b}\right\}.$$

اگر این حرکت پذیرفته شد، متغیرهای تخصیص و پارامترهای تعیین خوشه را تغییر می‌دهیم.

ب. یک خوشه به نام a انتخاب می‌کنیم و نام خوشه‌های a و $a+۱$ را تغییر می‌دهیم. همچنین ν_a و $\nu_{a+۱}$ را در زمان‌های برابر تغییر می‌دهیم. نام خوشه a به تصادف با خوشه‌هایی با نام‌های $۱,۲,\ldots,n^*-۱$ عوض می‌شود. که در آن n^* خوشه ناتهی با بزرگترین نام است. احتمال پذیرش این حرکت برابر است با

$$\min\left\{۱\ ,\ \frac{(۱-\nu_{a+۱})^{n_a}}{(۱-\nu_a)^{n_{a+۱}}}\right\}.$$

اگر این حرکت پذیرفته شد، متغیرهای تخصیص و پارامترهای تعیین خوشه را تغییر می‌دهیم.

۸. تولید نمونه از پارامتر α که به آن پارامتر تمرکز نیز می‌گویند بر طبق فرآیند زیر انجام می‌شود (اسکوبار و وست، ۱۹۵۸). با در نظر گرفتن توزیع پیشین $\mathrm{Gamma}(\alpha|a,b)$ با میانگین $(\frac{a}{b})$، توزیع پسین آمیخته‌ای از دو توزیع گاما است. یعنی

$$\alpha|\eta,k \sim \pi_h\mathrm{Gamma}(a+k, b-\log(\eta)) + (۱-\pi_h)\mathrm{Gamma}(a+k-۱, b-\log(\eta))$$

که در آن، k شماره خوشه ناتهی است و

$$\pi_h = (\frac{a+k-۱}{[a+k-۱+n(b-\log(\eta))]}).$$

همچنین باید نشان دهیم

$$\eta|\alpha,k \sim \mathrm{Beta}(\alpha+۱, n).$$

[۱۰] Label Swichings

برهان.

$$f(\alpha|k) = f(\alpha)f(k|\alpha)$$

$$f(k|\alpha) = C_n(k)n!\alpha^k\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}$$

که در آن $k = ۱,۲,\ldots,n$ و $C_n(k) = f(k|\alpha = ۱, n)$. حال قرار می‌دهیم

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+۱,n)}{\alpha\Gamma(n)}$$

که در آن $\beta(\cdot,\cdot)$ تابع بتا است. بنابراین

$$f(\alpha|k) \propto f(\alpha)\alpha^{k-۱}(\alpha+n)\beta(\alpha+۱,n) \propto f(\alpha)\alpha^{k-۱}(\alpha+n)\int_۰^۱ x^\alpha(۱-x)^{n-۱}\mathrm{d}x.$$

در واقع $f(\alpha|k)$ چگالی حاشیه‌ای از توزیع توأم $f(\alpha,\eta|k)$ است. پس

$$f(\alpha,\eta|k) \propto f(\alpha)\alpha^{k-۱}(\alpha+n)\eta^\alpha(۱-\eta)^{n-۱} \quad ,\alpha > ۰, ۰ < \eta < ۱.$$

برای به‌دست آوردن $f(\alpha|\eta,k)$ داریم

$$\begin{aligned} f(\alpha|\eta,k) &\propto \alpha^{a+k-۲}(\alpha+n)\mathrm{e}^{-\alpha(b-\log(\eta))} \\ &\propto \alpha^{a+k-۱}\mathrm{e}^{-\alpha(b-\log(\eta))} + n\alpha^{a+k-۲}\mathrm{e}^{-\alpha(b-\log(\eta))}. \end{aligned}$$

بنابراین

$$\alpha|\eta,k \sim \pi_h\mathrm{Gamma}(a+k, b-\log(\eta)) + (۱-\pi_h)\mathrm{Gamma}(a+k-۱, b-\log(\eta))$$

که در آن

$$\pi_\eta = \frac{a+k-۱}{n(b-\log(\eta))}.$$

$f(\eta|\alpha,k)$ نیز به‌صورت زیر است:

$$f(\eta|\alpha,k) \propto \eta^\alpha(۱-\eta)^{n-۱} \qquad ۰ < \eta < ۱.$$

در نتیجه

$$\eta|\alpha,k \sim \mathrm{Beta}(\alpha+۱,n).$$

□

تولید نمونه از این شرطی‌های کامل در بسته BNSP در نرم‌افزار R پیاده‌سازی شده است (پاپاجئورجیو، ۲۰۱۵).

فصل ۴

ارزیابی مدل رگرسیون چگالی بیزی

۱.۴ مطالعه شبیه‌سازی

در این فصل نتایج حاصل از یک مطالعه شبیه‌سازی را مورد بررسی قرار می‌دهیم. هدف اصلی، بررسی تأثیر انتخاب هسته‌های مختلف بر روی نتایج حاصل از رگرسیون چگالی است. در طول این شبیه‌سازی سه مکانیسم برای تولید داده ارائه می‌شود. با استفاده از هر کدام از این مکانیسم‌ها، با توجه به اینکه متغیر پاسخ شمارشی و متغیر تبیینی پیوسته است، داده‌ها را تولید کردیم. برای تولید داده، حجم نمونه را $n = ۱۵۰$ انتخاب کردیم. ابتدا متغیر تبیینی X را از توزیع $\text{Uniform}(۰, ۲۰)$ با میانگین μ_x تولید کردیم:

$$\mu_x = E(Y|X = x) = ۲۰ \left\{ N(x; ۵, ۲/۵^۲) + N(x; ۱۵, ۵^۲) \right\}$$

که در آن $N(x; \mu, \sigma^۲)$ چگالی نرمال با میانگین μ و واریانس $\sigma^۲$ است. سه مکانیسمی که برای تولید متغیر پاسخ در نظر گرفتیم، به صورت زیر است:

$$Y|X = x \sim \begin{cases} \text{round}(H\mu_x + \epsilon_۱) & (M_۱) \\ \text{Poisson}(H\mu_x) + \text{round}(\epsilon_۲) & (M_۲) \\ \text{Poisson}(H\mu_x) & (M_۳) \end{cases}$$

در مکانیسم‌های فوق، H پارامتر بی‌اثر است که آن را به‌صورت زیر تولید کردیم:

$$H \sim \text{Uniform}(۱, ۳۰).$$

تابع round() تابعی است که آرگمان‌ها را به نزدیک‌ترین عدد صحیح گرد می‌کند. همچنین خطای تصادفی ϵ_i، $i = ۱, ۲$، را به‌صورت زیر تولید کردیم:

$$\epsilon_۱ \sim N(۰, ۲)$$

$$\epsilon_۲ \sim N(۰, ۱۰)$$

از آن‌جا که در طول تولید داده‌ها از مکانیسم‌های $(M_۱)$ و $(M_۲)$ ممکن بود تحقق‌های منفی حاصل شوند، بنابراین متغیر پاسخ y را به‌صورت زیر در نظر گرفتیم:

$$y = \max(۰, y).$$

مکانیسم‌های $(M_۱)$، $(M_۲)$ و $(M_۳)$ به ترتیب نمایانگر وجود کم‌پراکنش، بیش‌پراکنش و پراکنش متعادل در داده‌های تولیدشده هستند. برای هر مجموعه داده شبیه‌سازی‌شده مدل زیر را برای داده‌ها برازش دادیم:

$$f(y_i^*, x_i) = \sum_{h=۱}^{\infty} \pi_h N_۲ \left(\begin{bmatrix} ۰ \\ \mu_h \end{bmatrix}, \begin{bmatrix} ۱/۰ & \sigma_{۱۲} \\ \sigma_{۲۱} & \sigma_{۲۲} \end{bmatrix} \right).$$

در مدل بالا، y_i^*، $i = ۱, ۲, \ldots, n$، متغیرهای پیوسته پنهان مربوط به متغیرهای پاسخ y_i است که به صورت زیر هستند:
$Y_i = y_i$ اگر و فقط اگر

$$c_{y_i - ۱} < y_i^* < c_{y_i}$$

که در آن c_{y_i} از قاعده زیر به‌دست می‌آید:

$$c_{y_i} = \Phi^{-۱} \{F(y_i; H, \boldsymbol{\xi})\}.$$

برای تابع توزیع $F(\cdot; \cdot, \cdot)$ نیز از هسته‌های معرفی شده در جدول ۱.۳ استفاده کردیم. برای این مجموعه داده و هر انتخاب برای تابع توزیع $F(\cdot; \cdot, \cdot)$، تعداد ۸۰۰۰۰ نمونه از توزیع پسین تولید کردیم. سپس تعداد ۴۰۰۰۰ از نمونه تولید شده را کنار گذاشتیم. از ۴۰۰۰۰ نمونه باقیمانده نیز از هر ۱۰ نمونه یکی را انتخاب کردیم. همچنین برای این شبیه‌سازی، ۳۰ مقدار برای x از چگالی شرطی $f(y|\mathbf{x})$ در فاصله $[۰/۱, ۱۹/۹]$ و دوره مؤلفه بی‌اثر $H = ۱۵$ تولید کردیم. مقدار میانگین نظری را محاسبه و به پایین گرد کردیم. برای هر نمونه از چگالی شرطی نیز مقادیر میانگین، صدک دهم و صدک نودم را محاسبه کردیم. سپس مقادیر میانگین، صدک دهم

و صدک نودم به‌دست آمده از نمونه‌ها را با مقادیر واقعی، به‌وسیله محاسبه کردن میانه‌های پسین توان‌دوم خطای کل مقایسه کردیم. توان‌دوم خطای کل به‌صورت زیر محاسبه می‌شود:

$$\sum_{c=۱}^{۳۰}(p_c - p_c^s)^۲ \quad s = ۱, ۲, \ldots, ۴۰۰۰$$

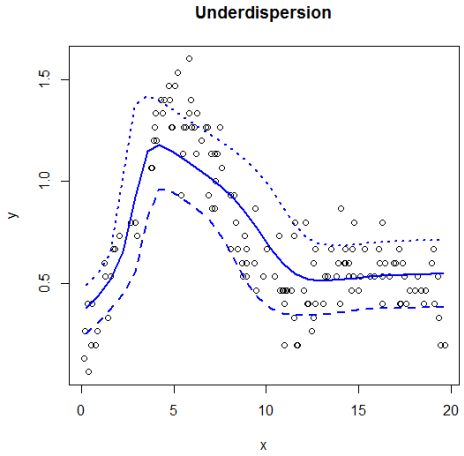
که در آن، p_c مقدار واقعی پارامتر یعنی همان میانگین، صدک دهم و صدک نودم است و p_c^s مقادیر به‌دست آمده از تعداد s نمونه MCMC می‌باشد. جدول ۱.۴ شامل خطاهای کل مدل فرآیند دیریکله آمیخته حاصل از انتخاب هسته‌های مختلف برای $k(\cdot;\cdot,\cdot)$ است.

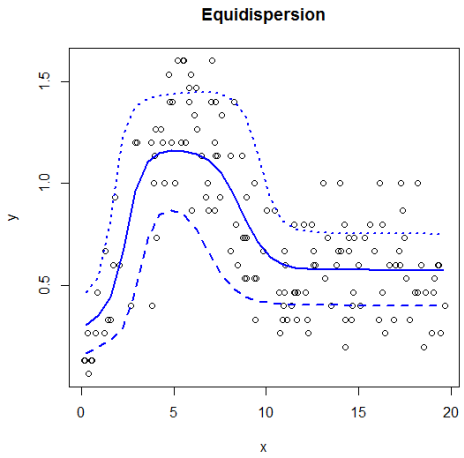
جدول ۱.۴: نتایج شبیه‌سازی: میانه‌های پسین حاصل از توان‌دوم خطای کل برای سه پارامتر میانگین، صدک دهم و صدک نودم تحت سه مکانیسم تولید داده و با انتخاب هسته‌های متفاوت.

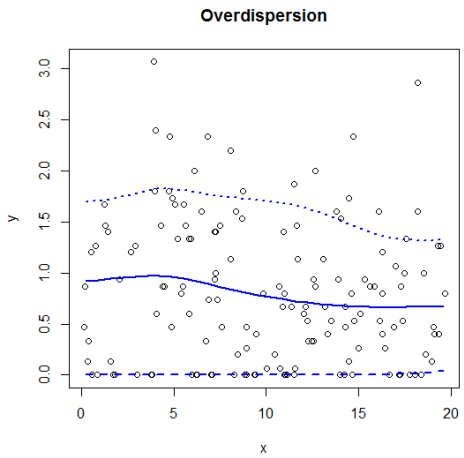
هسته	کم‌پراکنش			بیش‌پراکنش			پراکنش متعادل		
	میانگین	$Q_{۱۰}$	$Q_{۹۰}$	میانگین	$Q_{۱۰}$	$Q_{۹۰}$	میانگین	$Q_{۱۰}$	$Q_{۹۰}$
پواسن	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰	۱۰۰/۰۰
دوجمله‌ای منفی	۶۴/۷۲	۵۳/۵۵	۱۰۲/۲۱	۸۷/۳۷	۸۷/۹۷	۱۵۹/۹۰	۱۲۸/۲۴	۹۹/۱۳	۱۳۵/۷۴
پواسن تعمیم‌یافته	۶۰/۰۲	۵۰/۴۳	۹۶/۴۳	۷۵/۴۰	۶۵/۵۷	۸۵/۰۱	۸۴/۰۲	۸۷/۷۷	۱۱۲/۹۳
COM–پواسن	۵۶/۰۶	۴۵/۶۷	۵۸/۷۷	۶۶/۲۳	۶۳/۲۳	۷۰/۶۹	۱۰۳/۷۴	۹۰/۳۹	۹۲/۷۷
ابرپواسن	۹۷/۱۶	۱۰۰/۵۳	۹۷/۴۸	۷۳/۲۴	۷۸/۰۹	۶۵/۳۱	۱۰۵/۳۰	۱۰۳/۶۴	۱۰۳/۴۸
CTP	۱۵۷/۶۸	۶۴/۲۸	۲۷۷/۷۳	۱۱۳/۵۰	۱۴۸/۷۵	۲۷۶/۱۹	۲۱۶/۸۲	۱۴۴/۹۸	۴۸۴/۰۸

با به‌کار بردن مدل آمیخته فرآیند دیریکله نتایج جالبی به‌دست آمدند. اول این‌که با انتخاب هسته CTP، بدترین نتیجه نسبت به انتخاب سایر هسته‌ها در شبیه‌سازی ملاحظه می‌شود. فقط در یک مورد CTP بهتر از هسته پواسن عمل می‌کند و آن هم در برآورد چندک دهم برای داده‌هایی است که از مکانیسم کم‌پراکنش تولید شده‌اند. برآورد با انتخاب هسته‌های پواسن تعمیم‌یافته و COM-پواسن به‌طور اساسی بهبود می‌یابد. با انتخاب هسته ابرپواسن نیز برآورد بهبود داده می‌شود اما بهبود آن نسبت به هسته‌های پواسن تعمیم‌یافته و COM-پواسن کمتر است. با انتخاب هسته دوجمله‌ای منفی برآورد پارامتر مورد علاقه را تحت مکانیسم‌های بیش‌پراکنش و کم‌پراکنش بهبود می‌دهد و تحت مکانیسم پراکنش متعادل، فقط نسبت به هسته CTP بهتر عمل می‌کند.

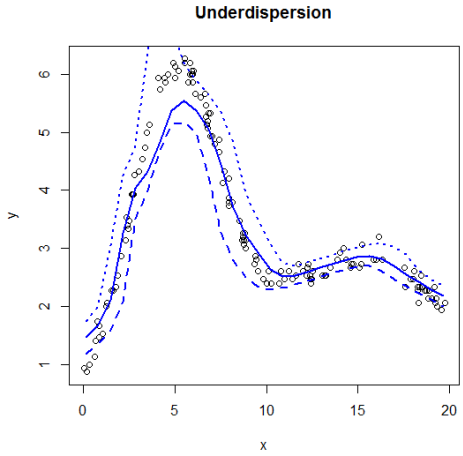
نتایج حاصل از شبیه‌سازی برای سه مکانیسم تولید داده و انتخاب هسته‌های مختلف در شکل‌های ۱.۴ تا ۸.۴ داده شده‌اند. در همه شکل‌ها، خط نقطه‌چین مربوط به صدک نودم، خط ممتد مربوط به میانگین و خط‌چین مربوط به صدک دهم است.

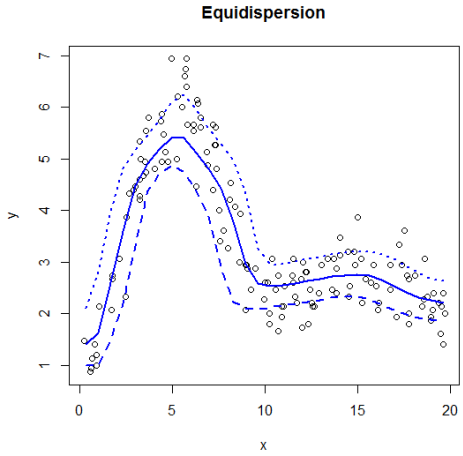


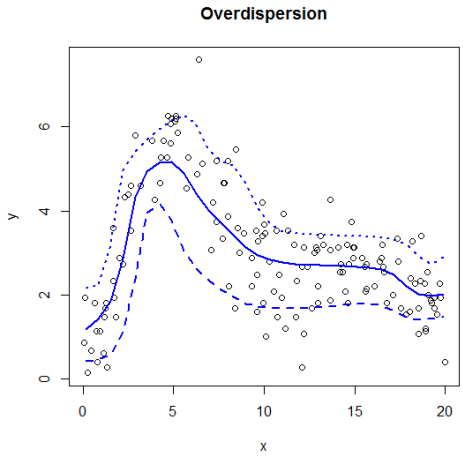




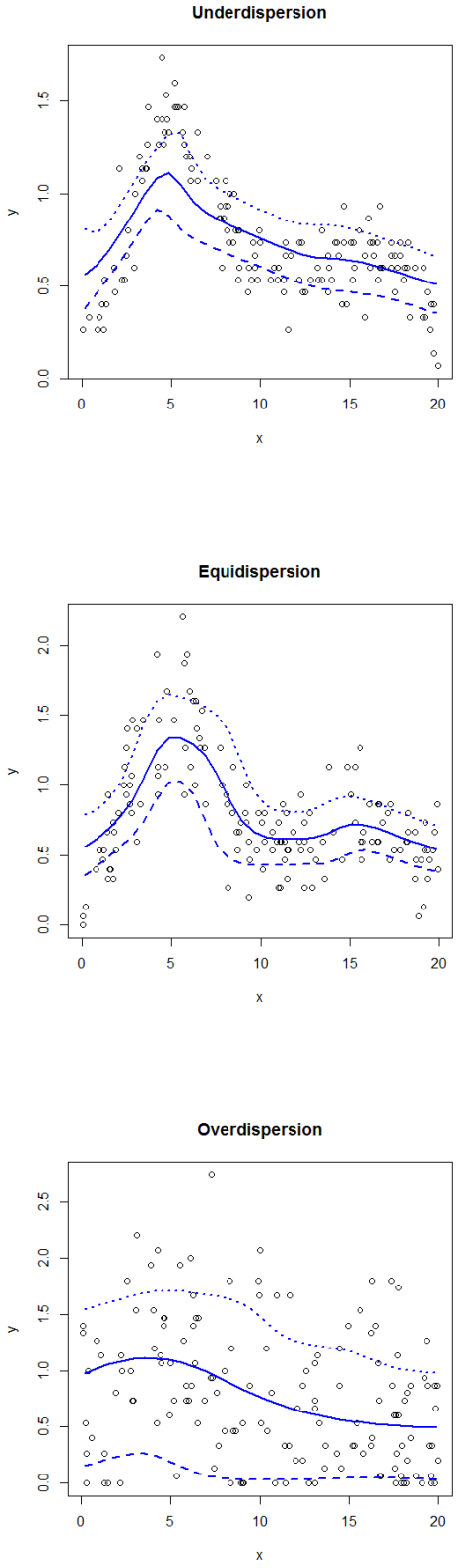
شکل ۱.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته پواسن و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.



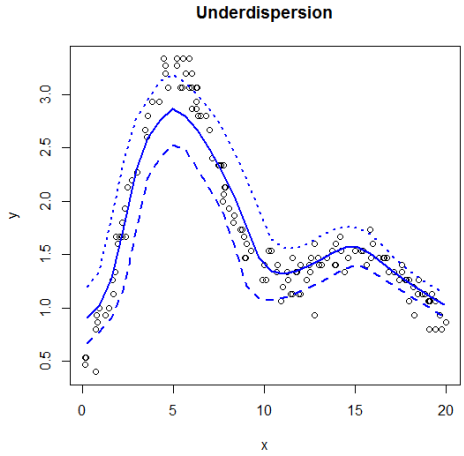


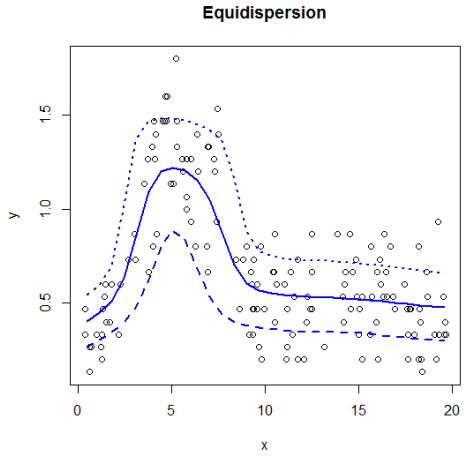


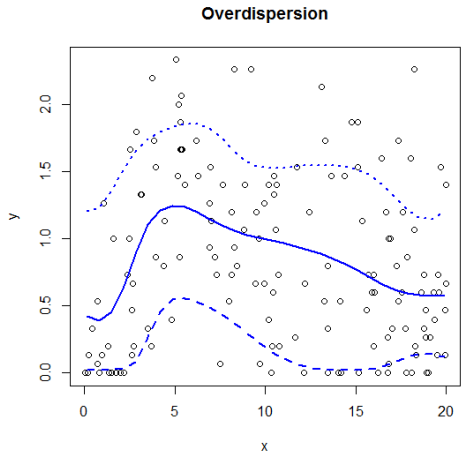
شکل ۲.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته پواسن تعمیم‌یافته و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.



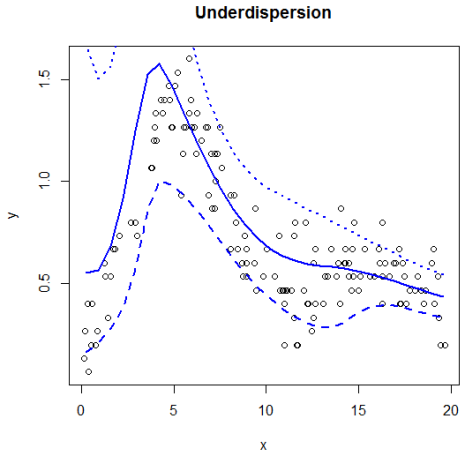
شکل ۳.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته ابرپواسن و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.

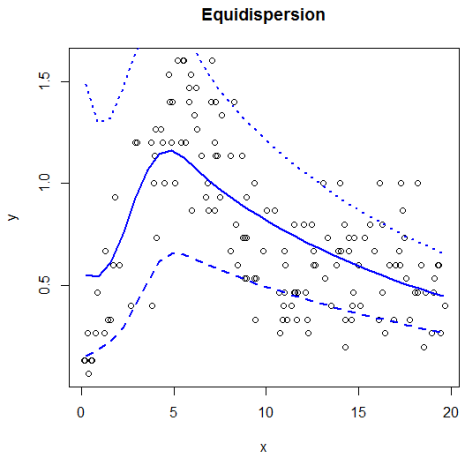


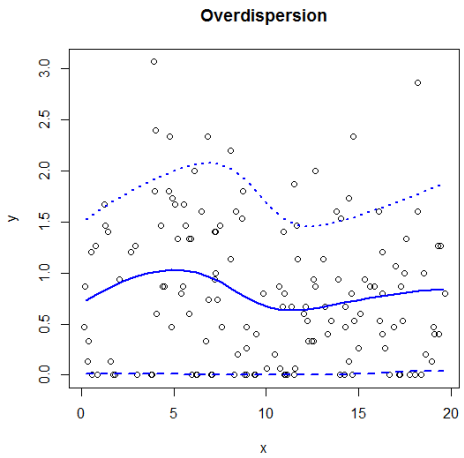




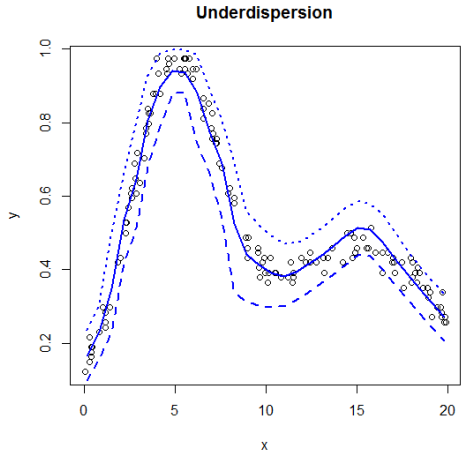
شکل ۴.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته $COM-$پواسن و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.

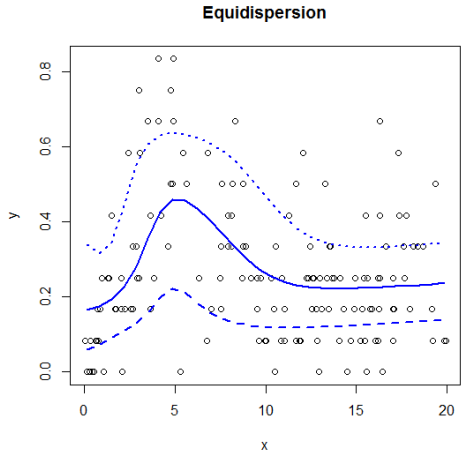


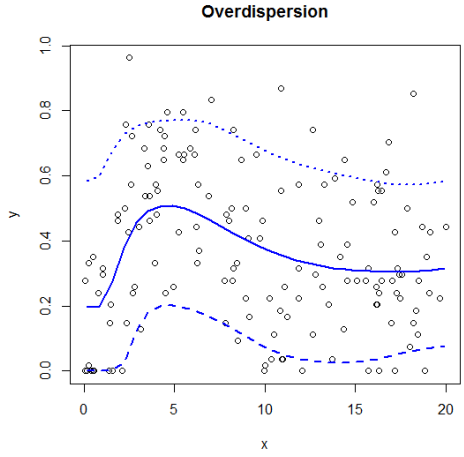




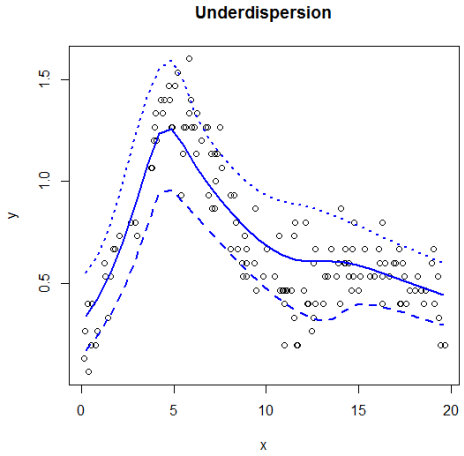
شکل ۵.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته CTP و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.

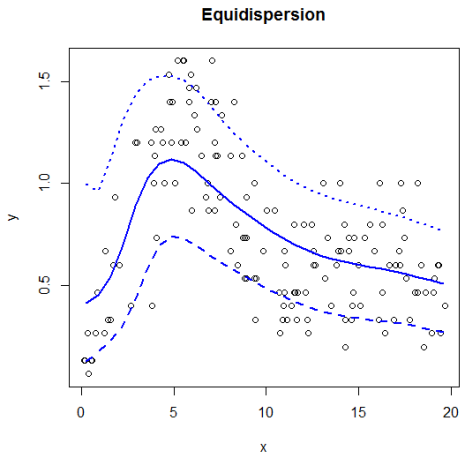


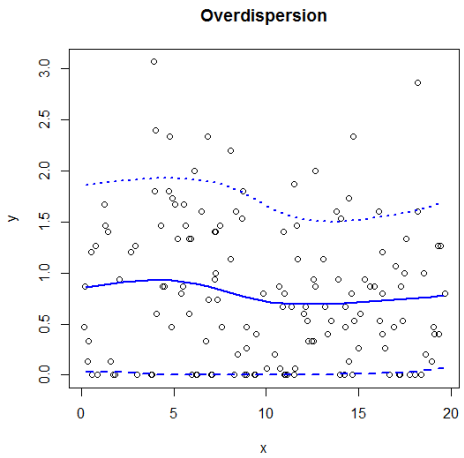




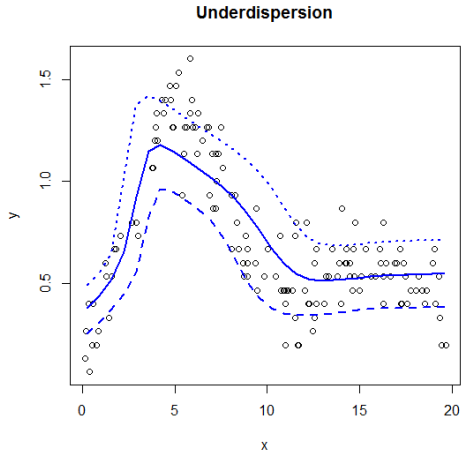
شکل ۶.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته دوجمله‌ای و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.

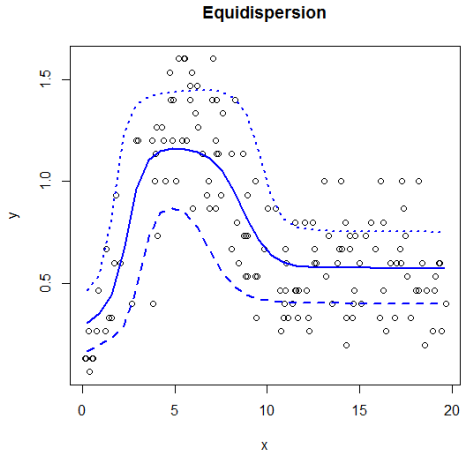


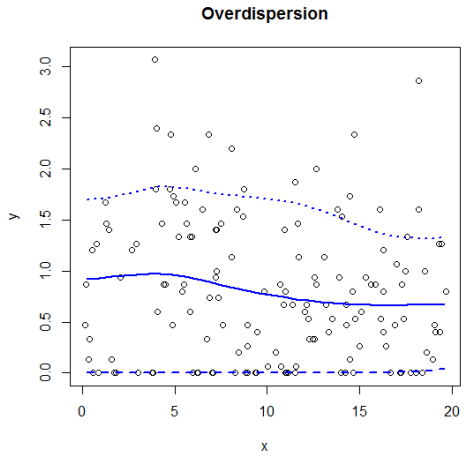




شکل ۷.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته دوجمله‌ای منفی و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.



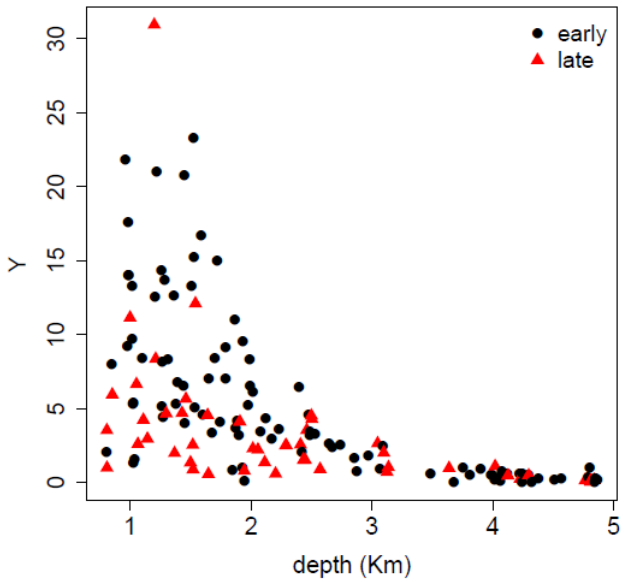




شکل ۸.۴: نتایج حاصل از شبیه‌سازی با انتخاب هسته بتا دوجمله‌ای و تولید داده از سه مکانیسم به ترتیب کم‌پراکنش، پراکنش متعادل و بیش‌پراکنش.

۲.۴ کاربرد مدل رگرسیون چگالی برای داده‌های واقعی

در این بخش تحلیل بیزی رگرسیون چگالی را بر روی یک مجموعه داده واقعی تشریح می‌کنیم. رگرسیون چگالی برای این مجموعه داده توسط بایلی و همکاران (۲۰۰۹) انجام شده است. داده‌ها مربوط به تأثیر تجارت ماهیگیری بر روی جمعیت ماهی‌ها در دریای آتلانتیک شمالی است. برای دسترسی به این داده‌ها می‌توان به هیلب (۲۰۱۴) مراجعه کرد. داده‌ها شامل $n = ۱۴۷$ مشاهده بر روی تورهای ماهیگیری در دو دوره زمانی متفاوت بین سال‌های ۱۹۸۹ـ ۱۹۷۷ با برچسب اولیه و ۱۹۹۷-۲۰۰۲ با برچسب ثانویه و در اعماق ۴/۸-۰/۸ کیلومتر از سطح دریا است. برای هر یک از تورهای ماهیگیری، متغیر پاسخ y جمعیت ماهی‌های در هر تور است. متغیرهای توضیحی نیز شامل $x_۱$، دوره‌های زمانی و $x_۲$، میانگین عمق تورها در دریا بر حسب کیلومتر است. همچنین هر تور شامل مؤلفه بی‌اثر H است که بر مبنای اندازه منطقه جاروب‌شده بر حسب کیلومتر مربع است. نمودار فراوانی داده‌ها در شکل ۹.۴ نمایش داده شده است.



شکل ۹.۴: نمودار فراوانی داده‌ها با متغیر پاسخ Y (جمعیت ماهی‌ها) و متغیرهای توضیحی $x_۱$ (دوره‌های زمانی) و $x_۲$ (میانگین عمق تورها در دریا).

برای این داده‌ها مدل فرآیند دیریکله آمیخته با هسته نرمال سه‌متغیره را به شکل زیر برازش دادیم:

$$f(y_i^*, \mathbf{x}_{i1}^*, \mathbf{x}_{i2}) = \sum_{h=1}^{\infty} \pi_h N_3 \left(\begin{bmatrix} 0 \\ \mu_{1h} \\ \mu_{2h} \end{bmatrix}, \begin{bmatrix} 1/0 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & 1/0 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \right).$$

مشاهدات و متغیرهای پنهان، با توجه به دو مرحله زیر، وارد مدل شدند:

- $X_{i1} = 0$ اگر و فقط اگر

$$x_{i1}^* < 0/0$$

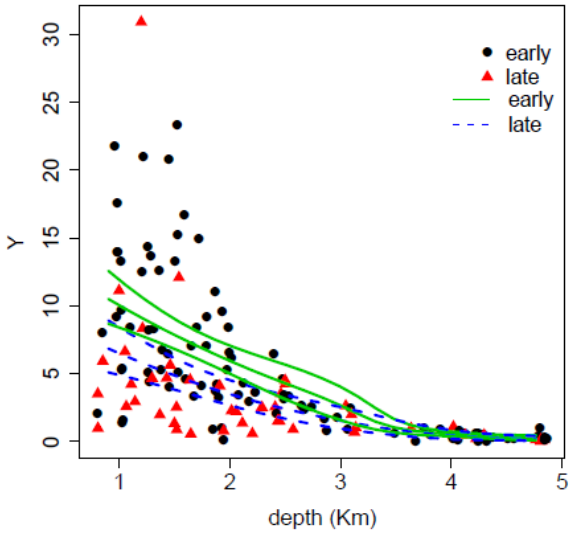
- $Y_i = y_i$ اگر و فقط اگر

$$c_{y_i-1} < y_i^* < c_{y_i}$$

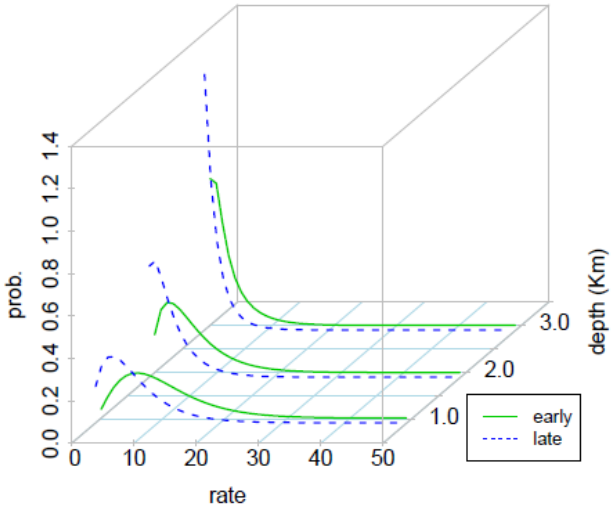
به‌طوری که

$$c_{y_i} = \Phi^{-1}\{F(y_i; H, \boldsymbol{\xi})\}.$$

در این‌جا تابع توزیع $F(\cdot;\cdot,\cdot)$ که برای مدل‌بندی نقاط برش از آن استفاده کردیم، توزیع دوجمله‌ای منفی در نظر گرفته شد. همان‌طور که قبلاً هم اشاره شد، توزیع دوجمله‌ای منفی زمانی که داده‌ها دارای بیش‌پراکنش هستند، کاربرد دارد. نتایج برازش در شکل‌های ۱۰.۴ و ۱۱.۴ نمایش داده شده‌اند.



شکل ۱۰.۴: نتایج حاصل از رگرسیون میانگین به همراه فواصل اطمینان ۹۰ درصدی برای داده‌های ماهیگیری.



شکل ۱۱.۴: توابع چگالی برازش‌شده برای داده‌های ماهیگیری به تفکیک مقادیر متغیرهای تبیینی.

شکل ۱۰.۴ نتایج حاصل از رگرسیون میانگین به همراه فاصله‌های اطمینان ۹۰ درصدی نقاط را نمایش می‌دهد که در آن خط‌های ممتد نتایج حاصل از دوره زمانی اولیه و خط‌چین‌ها نتایج حاصل از دوره زمانی ثانویه هستند. همان‌طور که مشاهده می‌شود، جمعیت ماهی‌ها برای هر دو دوره زمانی، با افزایش عمق کاهش می‌یابد. تا عمق ۳/۵ کیلومتری، میانگین جمعیت ماهی‌ها در دوره زمانی ثانویه معمولاً کمتر از میانگین جمعیت ماهی‌ها در دوره زمانی اولیه است. برای عمق بیشتر از ۲/۱ کیلومتر نیز، فاصله‌های اطمینان برای میانگین جمعیت ماهی‌ها تقریباً با هم هم‌پوشانی دارند.

شکل ۱۱.۴ نتایج حاصل از رگرسیون چگالی را نشان می‌دهد که در آن، خط‌های ممتد برآورد چگالی برای دوره اولیه و خط‌چین‌ها برآورد چگالی برای دوره ثانویه هستند. همان گونه که ملاحظه می‌شود تا عمق ۳ کیلومتر، چگالی احتمال برای دوره اولیه بیشتر از دوره ثانویه است. برای عمق‌های بالاتر از ۳ کیلومتر، چگالی احتمال برای دو دوره تقریباً بر هم منطبق می‌شوند.

۳.۴ بحث و نتیجه‌گیری

در این پایان‌نامه مدل بیزی رگرسیون چگالی وقتی که متغیر پاسخ گسسته است، ارائه شد.

با استفاده از روشی که به‌کار برده شد، متغیرهای گسسته به‌صورت متغیرهای پیوسته پنهان که گسسته‌سازی شده‌اند، در نظر گرفته شدند و با استفاده از مدل آمیخته فرآیند دیریکله با هسته گاوسی، چگالی توأم حاصل از مشاهدات و متغیرهای پنهان معرفی گردید.

مدل‌بندی نقاط برش بر اساس $\Phi[F(\cdot;\cdot;\cdot)]$ انجام می‌شود که در آن $\Phi(\cdot)$ تابع توزیع نرمال استاندارد و $F(\cdot;\cdot,\cdot)$ متناسب با داده‌ها انتخاب می‌شود. همچنین می‌توان تصور کرد که تابع توزیع $F(\cdot;\cdot,\cdot)$ همانند تابعی برای پیوند بین مشاهدات و متغیرهای پنهان عمل می‌کند و نقش مشابه تابع پیوند در مدل‌های خطی تعمیم‌یافته را ایفا می‌کند.

برای برازش مدل به دلیل پیچیدگی محاسبه توزی پسین، از الگوریتم‌های MCMC استفاده کردیم. با به‌کار بردن توابع توزیع مختلف برای $F(\cdot;\cdot,\cdot)$ و استفاده از الگوریتم‌های MCMC مدل نهایی برازش داده می‌شود. انجام این کار با استفاده از بسته BNSP (پاپاجئورجیو، ۲۰۱۵) در نرم‌افزار R صورت می‌گیرد.

۴.۴ پیشنهادات برای آینده تحقیق

در این پایان‌نامه مدل بیزی رگرسیون چگالی برای پاسخ‌های گسسته معرفی و به‌کار برده شد. استفاده از این مدل هنگامی که متغیرها دارای وابستگی فضایی هستند می‌تواند موضوع جالب توجهی برای پژوهش‌های بیشتر باشد.

مراجع

[۱] پارسیان، ا. (۱۳۹۱)، **"مبانی آمار ریاضی"** چاپ دهم، مرکز نشر دانشگاه اصفهان.

[۲] نیرومند، ح. (۱۳۸۴)، **"تحلیل رگرسیون خطی"** چاپ اول، موسسه چاپ و انتشارات دانشگاه فردوسی مشهد.

[۳] نیرومند، ح. (۱۳۸۷)، **"الگوهای خطی تعمیم‌یافته با کاربردهای آن در مهندسی و علوم"** چاپ اول، موسسه چاپ و انتشارات دانشگاه فردوسی مشهد.

[4] Albert, J. H. & Chib, S. (1993). Bayesian analaysis of binary and polychotomous responso data, *Journal of the American Statistical Association*, **88**(422), 669-679.

[5] Antoniak, C. E. (1974). Mixture of Dirichlet Processes with Applications to Bayesian Non-parametric Problems, *The Annals of Statistics*, **2**, 1152-1174.

[6] Bailley, D. Collins, M., Gordon, J., Zuur, A., & Priede, I. (2009). Long-term changes in deep-water fish papulations in the northeast atlantic: a deeper reaching effect of fisheries? *Proceedings of the Royal Society of London B: Biological Sciense*.

[7] Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage, *Statistica Sinica*, **10**(4), 1281-1311.

[8] Barron, A., Schervish, M.J., & Wasserman, L. (1999). Posterior distributions in nonparametric problems, *Annals of Statistics*, **27**, 536–561.

[9] Berger, J. O. & Bernardo, J.M. (1992). On the development of refernce priors, *Bayesian Statistic*, **4**(4), 35-60.

[10] Bickel, P. J. & Doksum K.A. (2001). *Mathematical Statistics: basics ideas and selected topics, CRC Press. New York.*

[11] Birkhof, G. D. (1942). What is the ergodic theorem? *American Mathematical Monthly*, **49**(4), 222-226.

[12] Blackwell, D. (1973). Discreteness of Ferguson Selection, *The Annals of Statistics*, **1**, 356-358.

[13] Blackwell, D. & MacQueen, J. (1973). Ferguson Distributions Via Polya urn Schemes, *The Annals of Statistics*, **1**, 353-355.

[14] Borsboom. D., Mellenbergh, G.J. & Van Herden, J. (2003). The theorical status of latent variables, *Psychlogical Peview,* **110**(2), 203.

[15] Branscum, A. J., Jonson, W.O. & Thurmond, M.C. (2007). Bayesian beta regression: application to household data and genetic distanse between foot-and-mouth disease viruses, *Australian & New Zealand Journal of Statistics*, **49**(3), 287-301.

[16] Caffo, B., An, M. & Rohde, C. (2007). Flexible Random Intercept Models for Binary Outcoms Using Mixtures of Normal, *Journal of Computational Statistics and Data Analysis*, **51**, 5220-5235.

[17] Canale, A. & Dunson, D.B. (2015). Bayesian Multivariate mixed-scale density estimation, *Statistics and its Inference*, **8**(2), 195-201.

[18] Cansul, P.C. & Famoye, F. (1992). Generalized Poisson regression model, *Communications in Statistics Theory and methods*, **21**(1), 89-109.

[19] Casella, G. & Berger, R.L. (2002).*Statistical inference,* Duxbury, theorem learning, New York.

[20] Chen, M. H., Shao, Q.M. & Ebrahim, J. (2000). *Monte Carlo Methods in Mayesian Computation*, Springer, New York.

[21] Dayton, C. M. & Macready, G.B. (1988). Concomitant-variable latent-class models, *Journal of the American statisticall Assosiation*, **83**, 173-178.

[22] Dey, D., Muller, P. & Sinha, D. (1998). Practical Nonparametric and semiparmetric Bayesian Statistics, New York, *Springer-Verlag*.

[23] DeYoreo, M. & Kottas, A. (2014). *Bayesian nonparametric modeling for multivariate ordinal regression*, Technical report, University of California, Santa Cruz, $http$: $//arxiv.org/abs/$۱۴۰۸/۱۰۲۷.

[24] DeYoreo, M. & Kottas, A. (2015). A fully nonparametric modelling approach to binary regression, *Bayesian Analysis*, **10**(4), 821-847.

[25] Dunson, D. & Bhattacharya, A. (2011). Nonparametric Bayes regression and classifcation through mixtures of product kernels, In *Bayesian Statistics 9, Proceedings of the Ninth Valencia International Meeting*: Oxford University Press.

[26] Dunson, D. B., Pillai, N., & Park, J.-H. (2007). Bayesian density regression, *Journal of the Royal Statistical Society: Series B*, **69**(2), 163-183.

[27] Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, **90**, 577-588.

[28] Fisher, R. A. (1925). Theory of statistical estimation, *In Mathematical Proseeding of the Cambridg Philosophical Society*, **22**, 700-725.

[29] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**(2), 209-230.

[30] Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions, In Recent Advances in Statistics, ads. J. Rutage and G. G. Rizvi. *New York: Academic Press*, 287-302.

[31] Fisher, R. A. & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *In Mathenatical Proseeding of the Cambridg Philosophical Society*, **24**, 180-190.

[32] Galton, F. (1885). Photographic composites, *The Photographic News*, **29**, 243-245.

[33] Gauss, C. F. (1821). *Theoria Combinationis Observationum Erroribus Minimise Obnoxiae*, Werk 4, Gottingen, New York.

[34] Gelfand, A. & Smith, A. (1990). Sampling based approaches to calculating marginal densities, *Journal of the American statisticall Assosiation*, **85**, 398-409.

[35] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D.B. (2003). *Bayesian Data Analysis*, Chapman and Hall, New York.

[36] Gelman, A. & Hill, J. (2006). *Data Analysis Regression and Multilevel Hierarchical Models*, Cambridge University Press, New York.

[37] Ghosh, J. K. & Romamoorthi, R. V. (2003). Bayesian Nonparametric, *Springer Series in Statistics*, New York.

[38] Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). Makov Chain Monte Carlo in Practice, *Chapman and Hall*, London.

[39] Gut, A. (2005). *Probability: A Graduate Course*, Springer Texts in Statistics, New York.

[40] Hastings, W. (1970). Monte Carlo sampling methods using Marcov Chans and their applications, *Biometrica*, **57**(1), 97-109.

[41] Hilbe, J. M. (2014). COUNT: *Functions, data and code for count data*, R package version 1.3.2.

[42] Hosking, J. R. M. (1985). Maximum likelihood estimation of the parameters of the generalized extreme value distributions, *Journal of the Royal Statistical Society (Series C)*, **34**(3), 301-310.

[43] Kolmogorov, A. N. (1933). Foundations of the Theory of Probability, 2nd Ed, *New York*.

[44] Kottas, A., Mller, P., & Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data, *Journal of Computational and Graphical Statistics*, **14**(3), 610-625.

[45] Kyung, M. Guil, J. & Casella, G. (2011). Sampling Schemes for Generalized Linear Dirichlet Process Random Effect Modell, *Journal of Statistical Methods and Applications*, **20**, 303-304.

[46] Ishwaran, H. & James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors, *Journal of American Statistical Association*, **96**, 161-173.

[47] Jackman, S. (1989). *Generalized Linear Models*, Stanford University, New York.

[48] Jara, A. Garc Ia-Zattera, M. J. & Lesaffere, E. (2007). A Diriclet Process Mixture Model for the Analysis of Correlated Binary Response, *Computational Statistics and Data Analysis*, **51**, 5402-5415.

[49] Jeffreys, H. (1936). Further significance tests, *Mathematical Proceedings of the Cambridge Philosophical society*, **32**(3), 416-445.

[50] Jeffreys, H. (1946). An invariant form for the prior probability in estimations problems, *Proceedings of the Royal Society*, **186**, 453-461.

[51] Jordan, M. I. Teh, Y. W. Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet Processes, *American Statistical Association*, **101**(476), 1566-1581.

[52] Laplace, P. S. (1812). *Thorie Analatique Des Probabilities*, Courcier, Paris.

[53] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, **12**(1), 351-357.

[54] MacEacher, S. N. & Muller, P. (1988). Estimating Mixture of Dirichlet Process Models, *Junal of Computation and Grafical Statistics*, **2**, 223-238.

[55] Malcolm, F. (2008). Bayesian Statistics. *Newcastle University*.

[56] McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models (2nd edition)*, London: Chapman & Hall.

[57] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E., (1953). Equatin of state calculation by fast computing machines, *The Journal of Chemical Physics*, **21**(6), 1087-1091.

[58] Metropolis, N. & Ulam, S. (1949). The Monte Carlo methods. *Journal of the American Statistical Association*, **44**, 335-341.

[59] Montgomery, D. C., Peck, E. A. & Vining, G.G. (2015). *Intruduction to Linear Regression Analysis*, John Wiley & Sons, Hoboken, New York.

[60] Muliere, P. & Tardella, L. (1998). Aproximating Distributions of Functionals of Ferguson-Dirichlet Process, *Canadian Jounal of Statistics*, **30**, 269-283.

[61] Muller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures, *Biometrika*, **83**(1), 67-79.

[62] Muller, P. & Mitra, R. (2013). Bayesian nonparametric inference-Why and How, *Bayesian Analysis*, **8**. 1-34.

[63] Muller, P. & Rodriguez, A. (2013). *Nonparametric Bayesian Inference*, IMS-CBMS Lecture Notes. IMS.

[64] Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika*, **49**(1), 115-132.

[65] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[66] Papageorgiou, G. (2015). BNSP: *Bayesian non- and semi-parametric model fitting*, R package version 1.1.0.

[67] Papageorgiu, G., and Richardson, S. (2016). Bayesian dencity regression for discrete outcomes. *arXiv preprint arXiv*: 1603, 09706.

[68] Papageorgiou, G., Richardson, S., & Best, N. (2015). Bayesian nonparametric models for spatially indexed data of mixed type, *Journal of the Royal Statistical Society: Series B*, 973-999.

[69] Papaspiliopoulos, O. (2008). *A note on posterior sampling from Dirichlet mixture models*, Technical report, University of Warwick.

[70] Papaspiliopoulos, O. & Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, *Biometrika*, **95**(1), 169-186.

[71] Prescott, P. & Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalized extreme vlue distributions, *Biometrica*, **67**(3), 723-724.

[72] Richardson, S. & Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society: Series B*, **59**, 731-792.

[73] Roberts, C. P. (2009). Simulation of truncated normal variables, *Statistics and Computing*, **5**(2), 121-125.

[74] Roberts. C. P. (2007). *Bayesian Choice*, Springer Texts in Statistics.

[75] Roberts, C. P. & Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*, Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[76] Roberts, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer, New York.

[77] Roberts, C. & Casella, G. (2010). *Introdusing Monte Carlo Methods With R*, Springer, New York.

[78] Roberts, G. O. & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms, *Statistical Science*, **16**(4), 351-367.

[79] Rodriguez-Avi, J., Conde-Sanchez, A., Saez-Castillo, A. J., & Olmo-Jimenez, M. J. (2004). A triparametric discrete distribution with complex parameters, *Statistical Papers*, **45**(1), 81-95.

[80] Saez-Castillo, A. & Conde-Sanchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data, *Computational Statistics & Data Analysis*, **61**, 148-157.

[81] Sellers, K. F. & Shmueli, G. (2010). A exible regression model for count data, *The Annals of Applied Statistics*, **4**(2), 943-961.

[82] Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639-650.

[83] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution, *Journal of the Royal Statistical Society: Series C*, **54**(1), 127-142.

[84] Taddy, M. A. & Kottas, A. (2010). A Bayesian nonparametric approach to inference for quantile regression, *Journal of Business & Economic Statistics*, **28**(3), 357-369.

[85] Tsiatis, A. A. (2006). Semiparametric Theory and Missing Data, *Springer, New York*.

[86] Vehtary, A. & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142-228.

[87] Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices, *Communications in Statistics Simulation and Computation*, **36**(1), 45-54.

[88] Wang, X. & Dey, D. K. (2010). Generalized extreme value regression for binary response data: an application to B2B electronic payments system adoption, *The Annals of Applied Statistics*, **4**(4), 867-897.

[89] Yang, M. and Dunson, D. B. (2010). Bayesian Semiparametric Structural Equation Models with Latent Variables, *Psychometrika*, **75**.

[90] Yildirim, I. (2012). *Bayesian Inference: Metropolis-Hastings Sampling*, Rochester, New York.

[91] Zhang, X., Boscardin, J. W., & Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables, *Journal of Computational & Graphical Statistics*, **15**(4), 880-896.

پیوست آ

کدهای نرم‌افزار R برای بازتولید قسمتی از نتایج پایان‌نامه

در این پیوست کدهای مرتبط با شبیه‌سازی و نمودارها فصل چهارم گزارش شده‌اند.

آ.۱ کدهای مربوط به شبیه‌سازی مدل با هسته پواسن

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))
```

```
y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۱ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for poisson #################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۱ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for poisson #################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(y.equi ~ x, family="poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۱ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for poisson #################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))
```

```
plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲ کدهای شبیه‌سازی مدل با هسته پواسن تعمیم‌یافته

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۲ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for generalized poisson#################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="generalized poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۲ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for generalized poisson#################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)
```

```
fit.y.equi <- bnpglm(y.equi ~ x, family="generalized poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۲ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for generalized poisson#################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="generalized poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳ کدهای مربوط به شبیه‌سازی مدل با هسته ابرپواسن

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۳ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for hyper-poisson#################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="hyper-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
```

```
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۳ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for hyper-poisson#################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(y.equi ~ x, family="hyper-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۳ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for hyper-poisson#################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="hyper-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۴ کدهای مربوط به شبیه‌سازی مدل با هسته $COM-$پواسن

```
rm(list=ls())
library(BNSP)

################# data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))
```

```
y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۴ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for com-poisson#################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="com-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۴ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for com-poisson#################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(y.equi ~ x, family="com-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۴ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for com-poisson#################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="com-poisson",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
```

```
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۵ کدهای مربوط به شبیه‌سازی مدل با هسته CTP

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۵ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for ctpd#################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="ctpd",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۵ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for ctpd#################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(y.equi ~ x, family="ctpd",
```

```
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۵ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for ctpd#################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="ctpd",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۶ کدهای مربوط به شبیه‌سازی مدل با هسته دوجمله‌ای

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۶ برای مکانیسم کم‌پراکنش

```
######################## fit under dispersin for binomial ################

E.under<- max((dat$y.under))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(cbind(y.under,(E.under-y.under))~x, family="binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
```

```
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,E.under),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۶ برای مکانیسم پراکنش متعادل

```
######################## fit equi dispersin for binomial ################

E.equi <- max((dat$y.equi))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(cbind(y.equi,(E.equi-y.equi))~x, family="binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,E.equi),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۶ برای مکانیسم بیش‌پراکنش

```
######################## fit over dispersin for binomial ###############

E.over <- max((dat$y.over))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(cbind(y.over,(E.over-y.over))~x, family="binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,E.over),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۷ کدهای مربوط به شبیه‌سازی مدل با هسته دوجمله‌ای منفی

```
rm(list=ls())
library(BNSP)

################# data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
```

```
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۷ برای مکانیسم کم‌پراکنش

```
################ fit under dispersin for negative binomial#################

E.under <- max(dat$y.under)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(y.under ~ x, family="negative binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)),sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,15),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۷ برای مکانیسم پراکنش متعادل

```
################ fit equi dispersin for negative binomial#################

E.equi <- max(dat$y.equi)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(y.equi ~ x, family="negative binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,15),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۷ برای مکانیسم بیش‌پراکنش

```
################ fit over dispersin for negative binomial#################

E.over <- max(dat$y.over)+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(y.over ~ x, family="negative binomial",
```

```
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
offset=rep(15,length(x)), sampler="truncated", prec=c(200),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,15),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۸ کدهای به مربوط شبیه‌سازی مدل با هسته بتا دوجمله‌ای

```
rm(list=ls())
library(BNSP)

################## data generate ################################

data.gen <- function(n=150){
H = runif(1,1,30)
x =runif(n,0,20)
mu.x = 20*(dnorm(x,5,2.5)+dnorm(x,15,5))

y.equi = rpois(n, H*mu.x)
y.under = pmax(round(H*mu.x+rnorm(n,0,2)),0)
y.over = pmax(rpois(n, H*mu.x)+round(rnorm(n,0,10)),0)
df <- data.frame(cbind(x,y.equi,y.under,y.over))
return(df)
}
dat <- data.gen(n=150)
dat
head(dat)
```

آ.۱.۸ برای مکانیسم کم‌پراکنش

```
######################## fit under dispersin for beta binomial#################

E.under<- max((dat$y.under))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.under <- bnpglm(cbind(y.under,(E.under-y.under))~x, family="beta binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.under)/with(dat,E.under),
xlab="x",ylab="y",main="Underdispersion")

lines(pred,fit.y.under$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.under$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.under$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۲.۸ برای مکانیسم پراکنش متعادل

```
######################## fit equi dispersin for beta binomial#################

E.equi <- max((dat$y.equi))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.equi <- bnpglm(cbind(y.equi,(E.equi-y.equi))~x, family="beta binomial",
```

```
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.equi)/with(dat,E.equi),
xlab="x",ylab="y",main="Equidispersion")

lines(pred,fit.y.equi$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.equi$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.equi$meanReg/rep(30,npred),col=4,lwd=2)
```

آ.۳.۸ برای مکانیسم بیش‌پراکنش

```
######################## fit over dispersin for beta binomial################

E.over <- max((dat$y.over))+2

pred <- seq(with(dat,min(x))+0.1, with(dat,max(x))-0.1, length.out = 30 )
npred <- length(pred)

fit.y.over <- bnpglm(cbind(y.over,(E.over-y.over))~x, family="beta binomial",
data=dat, ncomp=30, sweeps=80000, burn=40000, thin=10,
sampler="slice", prec=c(2000),
Xpred=pred, offsetPred=rep(30,npred))

plot(with(dat,x),with(dat,y.over)/with(dat,E.over),
xlab="x",ylab="y",main="Overdispersion")

lines(pred,fit.y.over$Q10Reg/rep(30,npred),col=4,lwd=2,lty=2)
lines(pred,fit.y.over$Q90Reg/rep(30,npred),col=4,lwd=2,lty=3)
lines(pred,fit.y.over$meanReg/rep(30,npred),col=4,lwd=2)
```

Abstract

We develop Bayesian models for density regression with emphasis on discrete responses. The problem of density regression is approached by considering methods for multivariate density estimation of mixed scale wariables, and obtaining conditional density estimation from the latter. The approach to multivariate mixed scale outcome density estimation that we describe represents discrete variables, either responses or covariates, as continuse latent variables that are thresholded into discrete ones. We present and compare several models for obtaining these thresholds in the challenging context of count data analysis where the response may be over-dispersed or under-dispersed in som of the regions of the covariate space. We utilize a nonparametric mixture of mutivariate Gaussians to model the directly observed and the latent continuouse variables. We present a Markov chain Monte Carlo algorithm for postorior sampling and provide illustrations on density, mean and quantile regression utilizing simulated and real datasets.

Keywords: Dirichlet process, latent variables, over-dispersion, under-dispersion, posterior distribution, Markov chain Monte Carlo algorithm.

Shahrood University of Technology

Faculty of Mathematical Sciences

MSc Thesis in: Statistics

Bayesian analysis of density regression for discrete responses

By: Hassan Mohammadi

Supervisor

Dr. Hossein Baghishani

January 2018