نمایی بر مدل‌های رده پنهان برای مدل‌بندی توأم اندازه‌گیری‌های طولی و داده‌های بقا

پروانه مهدی‌زاده[۱]، تابان باغفلکی[۲]، مهدی اسماعیلیان[۱]

تاریخ دریافت: ۱۴۰۰/۰۶/۱۳
تاریخ پذیرش: ۱۴۰۰/۰۸/۰۱

چکیده:

مدل‌های توأم در مطالعات پیگیری شونده برای بررسی ارتباط بین اندازه‌گیری‌های نشانگر طولی و یک پیشامد بقا استفاده می‌شود و به وضعیت‌هایی با چند نشانگر طولی و یا مخاطره‌های رقابتی تعمیم‌یافته است. بسیاری از دستاوردهای آماری در زمینه مدل‌بندی توأم در مدل‌های پارامتر مشترک متمرکزشده است که شامل مشخصه‌هایی از نشانگر طولی به‌عنوان متغیرهای تبیینی در مدل بقا در نظر گرفته می‌شود. یک رهیافت کمتر شناخته‌شده، مدل رده پنهان توأم است، این مدل با فرض اینکه ارتباط بین نشانگرهای طولی و خطر رخداد با یک ساختار رده پنهان کاملاً مشخص می‌شود، بناشده است. مدل رده پنهان به دلیل انعطاف‌پذیری در مدل‌بندی ارتباط بین نشانگرهای طولی و زمان تا رخداد پیشامد و همچنین توانایی در برگرفتن متغیرهای تبیینی به‌ویژه برای پیش‌بینی مناسب است. در این مقاله یک نمای کلی از مدل رده پنهان توأم و تعمیم‌های آن ارائه می‌دهیم، در این راستا، ابتدا مروری بر مدل‌های بحث شده انجام می‌شود و سپس برآورد پارامترهای مدل موردبحث قرار می‌گیرد. در بخش کاربرد، دو مجموعه‌ی داده‌ی واقعی مورد تحلیل و بررسی قرار می‌گیرند.

واژه‌های کلیدی: الگوریتم EM، برآوردگر درستنمایی ماکسیمم، داده‌های طولی، مدل بقا، مدل رده پنهان، مخاطره‌های رقابتی.

۱ مقدمه

در برخی مطالعات طولی، علاقه‌مند به جمع‌آوری هر دو اندازه‌های مکرر طولی و پیشامد زمان بقا هستیم. مدل‌های توأم ازآنجایی‌که یک تحلیل واحد با مدنظر گرفتن ارتباط بین خطر یک رخداد و تغییرات نشانگر طولی در طی زمان، ارائه می‌دهند، بسیار حائز اهمیت هستند و استفاده از آن‌ها در چنین مطالعاتی متداول است [۳۸، ۳۵]. به‌عنوان‌مثال، در مطالعات HIV، تعداد $CD4$ها و زمان ابتلا به ایدز و در مطالعه سرطان پروستات، اندازه‌گیری‌های آنتی‌ژن مخصوص پروستات[۳] (PSA) و خطر عود سرطان بررسی می‌شوند. ازآنجاکه در چنین مثال‌هایی، دو برآمد همبسته هستند، مدل بندی توأم آن‌ها برای در نظر گرفتن پیوند دو برآمد لازم و ضروری است. همچنین، ممکن است علاقه‌مند به پیش‌بینی نشانگر طولی در طی زمان باشیم، زیرا در برخی کاربردها مانند مطالعه سرطان پروستات، روند PSA بر عود بیماری حائز اهمیت است و پیش‌بینی آن می‌تواند کمک شایان‌ذکری برای درمان بیماری محسوب شود.

با استفاده از مدل‌بندی توأم می‌توان اریبی ناشی از خطاهای اندازه‌گیری تصادفی و اندازه‌گیری متناوب نشانگر طولی را اصلاح کرد [۳۲]. در برخی کاربردهای دیگر، هدف، بررسی روند نشانگر طولی در طی یک بیماری است و مدل‌بندی توأم اریبی ناشی از وقوع رخداد را اصلاح می‌کند. در نهایت، مدل‌های توأم زمانی موردنیاز است که درک ارتباط بین داده‌های نشانگر طولی و خطر وقوع رخداد از اهمیت بالایی برخوردار باشد. روش مدل‌بندی توأم شامل یک مدل طولی معمولاً یک مدل آمیخته و یک مدل زمان تا رخداد پیشامد معمولاً یک مدل مخاطره متناسب و ارتباط بین این دو مدل است که با استفاده از یک ساختار اشتراکی پنهان، در قالب اثرهای تصادفی، توصیف می‌شود. منابع شایان‌ذکری در این زمینه وجود دارد که هر دو استنباط کلاسیک و بیزی را برای برآورد پارامترهای مدل توأم به‌کاربرده‌اند [۱۰، ۴۰، ۱۶، ۱۵، ۱۹، ۳۸، ۸، ۴۱، ۱۷، ۳۶، ۳۹، ۳، ۴، ۵، ۱۲، ۱۳].

چنین مدل توأمی امکان مدل‌بندی داده‌های همگن و تنها با یک الگو را فراهم می‌سازد و استفاده از آن در تحلیل داده‌هایی با زیرگروه‌هایی با الگوهای مختلف از نمیرخ‌های پاسخ[۴] معقول نیست زیرا در چنین شرایطی فرض همگن بودن جامعه موردمطالعه نمی‌تواند واقع‌گرایانه باشد. برای مثال در کارآزمایی‌های بالینی چندمرکزی انتظار می‌رود مراکز مختلف به دلیل جمعیت‌های

[۱] گروه آمار و علوم کامپیوتر، دانشگاه محقق اردبیلی
[۲] گروه آمار، دانشگاه تربیت مدرس، (نویسنده مسئول: t.baghfalaki@modares.ac.ir)

[3] Prostate Specific Antigen
[4] Response Profiles

بیمار متفاوت و الگوهای ارجاعی مختلف، عملکردهای متفاوتی نسبت به بقا اولیه داشته باشند. بنابراین، باید زیرگروه‌های بالقوه اندازه‌گیری‌های طولی و نتایج بقا را که ممکن است توزیع‌های مختلف را دنبال کنند، در نظر بگیریم.

یک تعمیم استاندارد از مدل‌های مخاطره نسبی که چنین ساختاری را در برمی‌گیرند، در نظر گرفتن چندین رده است. به‌طور خاص، فرض بر این است که بیماران به رده‌های مختلفی تقسیم می‌شوند که هر رده تابع مخاطره پایه مربوط به خودش را دارد، اما مقادیر ضرایب رگرسیونی بین رده‌ها مشترک است.

به عبارتی تحت یک مدل رده‌بندی شده، تابع مخاطره بیمار i متعلق به رده k به‌صورت زیر بیان می‌شود:

$$h_{ik}(t) = h_{\circ k}(t)\, exp\{\boldsymbol{\gamma}'\boldsymbol{w_i} + \alpha m_i(t)\}, \quad (۱)$$

که در آن $h_{\circ k}(t)$ نشانگر تابع مخاطره پایه رده kام، $\boldsymbol{w_i}$ بردار متغیرهای تبیینی، γ بردار پارامترهای متناظر با $\boldsymbol{w_i}$ و α ضریب رگرسیونی است.

مدل‌های توأم با یک زیر مدل مخاطره نسبی رده‌بندی شده در بسته JM، تحت عملکرد تابع مخاطره پایه تقریبی Bـاسپلاین موجود است.

بر اساس فرمول استاندارد مدل‌های مخاطره نسبی رده‌بندی شده، تأثیر هر متغیر در رده‌های مختلف ثابت در نظر گرفته می‌شود. درصورتی‌که این فرضیه همیشه منطقی نیست، زیرا در بسیاری از موارد برخی از متغیرها ممکن است اثر متفاوتی برای هر رده داشته باشند. به‌عنوان‌مثال در یک کارآزمایی بالینی چندمرکزی با جمعیت بیمار مشابه در هر مرکز، ممکن است منطقی باشد که فرض کنیم اثر سن در مراکز مختلف یکسان است، اما نمی‌توان ادعا کرد که اثر درمان در تمام مراکز یکنواخت باشد. برای رفع مشکل در چنین مواردی مدل (۱) را با در نظر گرفتن اثرات متغیرها برای هر رده به‌صورت زیر تعمیم می‌دهیم:

$$h_{ik}(t) = h_{\circ k}(t)\, exp\{\boldsymbol{\gamma_k}'\boldsymbol{w_i} + \alpha_k m_i(t)\}, \quad (۲)$$

که در آن نه‌تنها تابع مخاطره پایه $h_{\circ k}(t)$ بلکه ضرایب رگرسیونی α_k و γ_k هم به رده k بستگی دارد.

مدل‌های توأم رده پنهان که به مدل توأم رده‌بندی شده مرتبط می‌باشد، برای رسیدگی به چنین فرایندهای ناهمگن استفاده می‌شود [۲۳، ۲۴، ۲۹]. انگیزه و هدف این نوع مدل‌های توأم نیز در نظر گرفتن ناهمگنی احتمالی در جمعیت است. بااین‌حال، و برخلاف مدل توأم رده‌بندی شده، مدل‌های توأم رده پنهان فرض می‌کنند که زیرجامعه‌هایی که جامعه را تشکیل می‌دهند، پنهان هستند، به این معنی که ناهمگنی توسط هیچ‌یک از متغیرهای مشاهده‌شده قابل‌دستیابی نیست.

لین و همکاران [۲۴] یک مدل رده پنهان برای تحلیل توأم نشانگرهای طولی و داده‌های بقا پیشنهاد کردند. آن‌ها احتمال عضویت رده را توسط یک مدل لوژستیک چندجمله‌ای مشخص کردند و همچنین فرض کردند که فرایندهای طولی و پیشامد بقا در هر رده پنهان مستقل هستند. بر این اساس، پارامترهای مدل را با استفاده از روش درستنمایی ماکسیمم و از طریق الگوریتم EM برآورد کردند. همچنین، مدل خود را برای مطالعه داده‌های مربوط به سرطان پروستات به کار بردند. مدل آن‌ها چهار زیرگروه را نشان می‌دهد که بیان‌کننده سطوح مختلف نشانگر آنتی‌ژن مخصوص پروستات و خطر ابتلا به سرطان پروستات است.

لارسن [۲۱]، یک مدل رده پنهان را برای تحلیل هم‌زمان پیشامد بقا و چندین نشانگر دودویی در مطالعه سلامتی زنان و پیری[5] به کار برد. بر اساس این مدل، سه رده پنهان کشف شد که نشان‌دهنده توجه به مسئله تحرک است. در پژوهش [۲۳] یک مدل آمیخته الگوی پنهان[6] برای داده‌های گمشده متناوب آگاهی‌بخش در مطالعات طولی ارائه شد.

هان و همکاران [۱۴] مدل رده پنهان با یک رخداد که در تحقیق [۲۴] معرفی‌شده بود را به چند رخداد تعمیم دادند به‌طوری‌که مدل‌های آمیخته الگوی پنهان معرفی‌شده برای داده‌های گمشده متناوب آگاهی‌بخش در منبع [۲۳] را به‌عنوان حالتی خاص تحت یک الگوی پارامتری در برگرفته است. شین و همکاران [۳۴] مدل رده پنهان تعدیل‌یافته برای بررسی داده‌های رسته‌ای با انصراف آگاهی‌بخش را برای تحلیل داده‌های ترک سیگار ارائه کردند. همچنین، یک مدل توأم رده پنهان برای بررسی ارتباط بین نشانگر طولی آنتی‌ژن مخصوص پروستات و عود سرطان پروستات در پژوهش [۳۳] پیشنهادشده است.

بیشتر روش‌های استفاده‌شده در مقالات ذکرشده، فرض استقلال شرطی پیشامدهای طولی و بقا برای هر رده پنهان را در نظر گرفته‌اند. چنین فرض استقلال شرطی، ممکن است برای به دست آوردن همبستگی بین دو پیشامد طولی و بقا ناکافی باشد، به این صورت که ممکن است منجر به کم‌برآوردی قدرت ارتباط و عدم قطعیت ساختار پیوند بین این دو فرایند شود [۲۵]. البته یک آزمون امتیاز توسط جاکمین‌ـگادا و همکاران [۲۰] برای ارزیابی اینکه آیا فرض استقلال شرطی در مدل‌های رده پنهان توأم برقرار است یا خیر، پیشنهادشده است.

گری و همکاران [۱۱] برای مدل‌بندی توأم داده‌های طولی و بقا،

[5] Women's Health and Aging Study

[6] Latent Pattern Mixture Model

یک مدل نقطه تغییر رده پنهان توأم با دو رده پنهان پیشنهاد کردند که در آن پیشامد طولی رده پنهان یک را با استفاده از یک مدل اثرات تصادفی تنها-عرض از مبدأ[7] و پیشامد طولی رده پنهان دو را با استفاده از یک مدل نقطه تغییر تصادفی تقسیم‌شده مدل‌بندی کردند و یک روش بیزی برای برازش مدل به کار بردند. بینکنز و همکاران [۶]، مدل رده پنهان را برای تحلیل داده‌های طولی ناقص به کار بردند. به‌طوری‌که بدون در نظر گرفتن پیشامد بقا، مدل لوژستیک را برای وضعیت گمشدگی استفاده کردند.

لیو و همکاران [۲۵]، مدل اثرات تصادفی توأم را در چارچوب مدل رده پنهان ادغام کردند تا بهتر بتوانند به ناهمگنی موجود در هر دو فرایند طولی و بقا با استفاده از یک روش متداول دست پیدا کنند. آن‌ها فرض کردند که در هر رده پنهان، یک مدل توأم متمایز از نقاط انتهایی بقا و طولی با اثرات تصادفی مشترک وجود داشته باشد و نشان دادند که مدل آن‌ها جامع‌تر است، از این نظر که می‌تواند، عامل‌هایی که بر عضویت رده پنهان تأثیر می‌گذارند را شناسایی و ارتباط بین اندازه‌های طولی و فرایند بقا را مدل‌بندی کند. علاوه بر این، با فراهم کردن یک ساختار کلی می‌تواند اثرات رده-مشخص متغیرها را روی مدل‌های بقا و طولی در نظر بگیرد. لازم به ذکر است که، در سال‌های اخیر استفاده از مدل رده پنهان برای مدل‌بندی توأم داده‌های طولی و زمان بقا بیشتر موردتوجه قرارگرفته و توسعه‌یافته است. همچنین، کاربردهایی از این نوع مدل‌بندی در آزمایش‌های بالینی مطرح‌شده است.

در برخی از آسیب‌شناسی‌ها، علاوه بر تعداد وقایع بالینی، چندین نشانگر یا رخداد در طول زمان جمع‌آوری می‌شوند و معمولاً در مواردی رخ می‌دهند که انجام فرایند طولی موردنظر از طریق نتایج مشاهده‌شده دشوار باشد. برای مثال، در بیماری آلزایمر (AD)[8] عملکرد شناختی[9] با آزمایش‌های روان‌سنجی متعدد اندازه‌گیری می‌شود تا شناخت بنیادی بهتر تخمین زده شود [۲۸]. همچنین به‌عنوان‌مثال دیگر، کیفیت زندگی توسط مجموعه‌ای از نشانگرها اندازه‌گیری می‌شود یا وابستگی عملکردی[10] توسط مجموعه‌ای از فعالیت‌های روزمره زندگی تعریف می‌شود [۲۷]. ازاین‌رو، مدل رده پنهان توأم، ابتدا با تمرکز بر یک نشانگر طولی و یک رخداد بالینی مطالعه شده است و سپس به حالت‌های چندین نشانگر طولی (پیوسته یا گسسته-ترتیبی) و یا چند رخداد به مفهوم علت-مشخص برای مخاطره‌های رقابتی تعمیم‌یافته است.

در پژوهش [۲۹]، یک روش رده پنهان غیرخطی برای مدل‌بندی توأم اندازه‌های طولی چند متغیره و زمان بقا پیشنهادشده است. همچنین پروست-لیما و همکاران [۳۲] یک مقاله مروری از مدل‌های رده پنهان توأم داده‌های طولی و زمان تا رخداد پیشامد ارائه کردند. انتینک و همکاران [۹]، یک مدل آمیخته معرفی کردند که ناهمگنی موجود در مسیرهای توسعه‌ای[11] را به دست آورد، درحالی‌که توسط مدل پاسخ چند سطحی[12] و سایر متغیرهای تبیینی به‌طور کامل قابل توصیف نبود. به‌منظور مدل‌بندی یک الگوی غیرخطی برای پیشامد طولی، چن و هوانگ [۷]، با دیدگاه بیزی یک ترکیب متناهی از مدل‌های توأم اثرات آمیخته نیمه پارامتری با توزیع چوله-t برای اندازه‌های طولی به دست آوردند. همچنین، هوانگ و همکاران [۱۸] برای مدل‌بندی داده‌های طولی چندمتغیره ناهمگن غیرخطی، ترکیبی متناهی از مدل‌های اثرات آمیخته غیرخطی برای مدل‌بندی رده پنهان نشانگر طولی در نظر گرفتند. به‌علاوه آن‌ها، مدل‌بندی چندین پیشامد رخداد را با استفاده از مخاطرات متناسب پیشنهاد کردند که تابع مخاطره برای هر رده پنهان به‌صورت تابع پله‌ای تعریف می‌شود. رونت و همکاران [۳۷]، یک مدل رده پنهان توأم بیماری-مرگ برای رخدادهای سانسور شده فاصله‌ای نیمه رقابتی و یک نشانگر طولی پیشنهاد کردند. آن‌ها دو تفسیر از مدل را توسعه دادند، یکی، مارکوفی و دیگری، نیمه مارکوفی و همچنین یک تعمیم برای تحلیل توأم از نشانگرهای طولی چندگانه ارائه کردند.

پروست-لیما و همکاران [۲۷]، نشانگرهای طولی و مخاطرات مرگ طبیعی و زوال عقل را با استفاده از رده‌های پنهان مشترک به هم پیوند زدند. این رده‌های پنهان، زیرجامعه‌های پنهانی که با نمایش معینی از تغییرات برای نشانگر و مخاطره رخدادها مشخص‌شده‌اند را فرمول‌بندی می‌کند. به‌این‌ترتیب، علاوه بر جداسازی دو نوع همبستگی و مدل‌بندی انعطاف‌پذیر وابستگی، یک ناهمگنی مورد انتظار در منحنی‌های طولی و خطرات رخداد را نیز مدل می‌کنند.

همچنین، [۳۰، ۳۱] یک بسته نرم‌افزاری R با نام **lcmm** معرفی کردند که مجموعه‌ای از توابع را برای برآورد مدل‌های آماری بر اساس تئوری مدل آمیخته خطی فراهم می‌کند. از جمله، تابع **Jointlcmm** برای برآورد مدل‌های آمیخته رده پنهان توأم برای یک پیشامد طولی (گوسی یا خم‌خطی[13]) و یک پیشامد بقا که می‌تواند چپ-بریده شده یا راست-سانسورشده باشد و در یک

[7] Intercept-only random-effects
[8] Alzheimer's Disease
[9] Cognitive Functioning
[10] Functional Dependency
[11] Developmental Trajectories
[12] Multilevel Item Response Model
[13] Curvilinear

ساختار رقابتی تعریف شود. لی و همکاران [۲۲]، یک روش رده پنهان برای مدل‌بندی توأم پیشامد زمان تا رخداد و چند نشانگر طولی پیشنهاد کردند درحالی‌که اندازه‌گیری‌های نشانگر سانسور شده را به دلیل محدودیت‌های تشخیص[۱۴] به‌حساب آورده‌اند. آندرینوپولو و همکاران [۲] برای غلبه به محدودیت‌های انتخاب تعداد رده بهینه، با استفاده از یک رویکرد کاملاً بیزی، رده‌های پنهان را در مدل توأم پارامتر مشترک ادغام کردند و برای انتخاب تعداد رده‌های بهینه یک ساختار مدل آمیخته معرفی کردند.

این مقاله به‌این‌ترتیب سازمان‌دهی شده است. در بخش دوم از این مقاله، مدل رده پنهان توأم را معرفی می‌کنیم که شامل مدل‌بندی توأم داده‌های طولی و زمان بقا و مدل‌بندی توأم اندازه‌های طولی و داده‌های مخاطره‌های رقابتی است. در بخش بعد به کاربرد مدل توصیف‌شده خواهیم پرداخت. این مقاله با بحث و نتیجه‌گیری خاتمه می‌یابد.

۲ مدل‌های توأم رده پنهان

۱.۲ مدل‌بندی توأم داده‌های طولی و زمان بقا

برای تعریف مدل‌های توأم رده پنهان، فرض می‌کنیم که G تعداد زیرجامعه‌هایی باشد که جامعه اصلی ما را تشکیل می‌دهند و نشانگر رده مشاهده نشده (پنهان) را به‌صورت $c_i = g$، $g = ۱, ..., G$ تعریف می‌کنیم که عضویت iامین مؤلفه را در رده پنهان g نشان می‌دهد.

با در نظر گرفتن δ_i به‌عنوان نشانگر رخداد، دوتایی (T_i, δ_i) زمان رخداد مشاهده‌شده را توصیف می‌کند، همچنین بردار $\boldsymbol{y}_i$ نشان‌دهنده اندازه‌های طولی و $\boldsymbol{\theta}$ بردار کل پارامترهای مدل توأم است، در این صورت این مدل‌ها تحت مفروضات استقلال شرطی زیر برقرار هستند:

$$\begin{aligned} P(T_i, \delta_i, \boldsymbol{y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) &= P(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times P(\boldsymbol{y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}), \\ P(\boldsymbol{y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) &= \prod_j P\{y_i(s_{ij}) | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}\}. \end{aligned} \quad (۳)$$

در این مدل‌ها فرض بر این است که همبستگی بین اندازه‌گیری‌های مکرر در فرایند طولی توسط اثرات تصادفی $\boldsymbol{b}_i$ بیان می‌شود، درحالی‌که، ارتباط بین زمان رخداد و فرایند طولی توسط شاخص رده پنهان مشترک c_i توضیح داده می‌شود. استفاده از دو مؤلفه پنهان از این نظر سودمند است که امکان ایجاد ساختارهای انعطاف‌پذیرتر برای اشتراک دو زیر مدل طولی و زمان بقا، در مقایسه با مدل توأم کلاسیک که تنها مجموعه اثرات تصادفی $\boldsymbol{b}_i$ را برای هر دو نوع ارتباط در نظر می‌گیرد، فراهم می‌کند.

تحت فرض‌های استقلال شرطی مذکور، یک تعریف کلی از مدل توأم رده پنهان شامل زیرمدل‌های زیر خواهد بود:

(۴)

$$\begin{cases} (\boldsymbol{Y}_i(s) | c_i = g, \boldsymbol{b}_i; \boldsymbol{\beta}_g) = X_i'(s)\boldsymbol{\beta}_g + Z_i'(s)\boldsymbol{b}_{ig} + \epsilon_i(s), \\ \qquad \epsilon_i(s) \sim N(۰, \sigma^۲) \\ h_i(t | c_i = g; \boldsymbol{\gamma}_g) = h_{۰g}(t) exp(\boldsymbol{w}_i{}' \boldsymbol{\gamma}_g) \\ \pi_{ig} = Pr(c_i = g; \boldsymbol{\lambda}_g) = \frac{exp(\boldsymbol{u}_i{}' \boldsymbol{\lambda}_g)}{\sum_{L=۱}^{G} exp(\boldsymbol{u}_i{}' \boldsymbol{\lambda}_g)}. \end{cases}$$

به‌طوری‌که فرض می‌کنیم بیماران در رده‌های پنهان متفاوت هم تکامل طولی متفاوت و هم تابع مخاطره مختلفی برای وقوع یک رخداد دارند. معادله اول رابطه (۴) تحت رده پنهان g، یک مدل آمیخته خطی استاندارد برای بردار اندازه‌های تکرارشونده نشانگر طولی، $\boldsymbol{Y}_i(s) = (Y_i(s_{i۱}), ..., Y_i(s_{in_i}))'$، در زمان‌های متفاوت اندازه‌گیری s_{ij}، $j = ۱, ..., n_i$، را در نظر می‌گیرد که در آن X_i و Z_i به ترتیب ماتریس طرح $p \times n_i$ و $q \times n_i$ بعدی متناظر با بردار p بعدی اثرات ثابت $\boldsymbol{\beta}_g$ و بردار q بعدی اثرات تصادفی $\boldsymbol{b}_{ig}$ برای فرد iام در رده g است. به‌طوری‌که qـبردار اثرات تصادفی رده مشخص به‌صورت، $\boldsymbol{b}_{ig} = \boldsymbol{b}_i | _{c_i = g} \sim N_q(\boldsymbol{\mu}_g, D_g)$ در نظر گرفته می‌شود که در آن، $D_g = \sigma_g^۲ D$. به‌عبارت‌دیگر، بردار اثرات تصادفی به‌صورت $\boldsymbol{b}_i \sim \sum_{g=۱}^{G} \pi_{ig} N_q(\boldsymbol{\mu}_g, D_g)$ مشخص می‌شود که در آن π_{ig} احتمال عضویت در رده پنهان gام است. ملاحظه می‌شود که وابستگی ماتریس کوواریانس اثرات تصادفی به رده c_i، تنها از طریق ضریب پارامتر واریانس یعنی $\sigma_g^۲$، بیان می‌شود و به‌منظور شناساپذیری $\sigma_G = ۱$ در نظر گرفته می‌شود.

معادله دوم رابطه (۴)، یک مدل خطر متناسب کاکس با عملکرد خطی و پارامترهای رده‌ـمشخص را توصیف می‌کند که در آن $h_{۰g}(t)$ تابع خطر پایه رده g و $\boldsymbol{w}_i$ بردار متغیر کمکی برای فرد iام در زمان t متناظر با بردار پارامتر γ_g است.

آخرین معادله رابطه (۴)، یک زیر مدل چندجمله‌ای برای احتمال عضویت در رده پنهان g، $g = ۱, ..., G$ را مشخص می‌کند، که در آن $\boldsymbol{u}_i$ بیانگر برداری از متغیرهای کمکی مرتبط با این احتمالات برای فرد iام و $\boldsymbol{\lambda}$ بردار پارامتر رده‌ـمشخص متناظر با $\boldsymbol{u}_i$ هستند. $\boldsymbol{\lambda}' = (\boldsymbol{\lambda}_۱', ..., \boldsymbol{\lambda}_G')$ و برای شناساپذیری فرض $\boldsymbol{\lambda}_G = ۰$ را در نظر می‌گیریم.

یک نتیجه سودمند حاصل از مدل توأم معرفی‌شده این است که

[14] Detection Limits

می‌توان لگاریتم درستنمایی تحت این مدل را به فرم زیر بیان کرد:

(۵)

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log P(T_i, \delta_i, \boldsymbol{Y}_i) \\ &= \sum_{i=1}^{n} \log \Bigg\{ \sum_{g=1}^{G} \Pr(c_i = g; \boldsymbol{\theta}) h_i(t_i | c_i = g; \boldsymbol{\theta})^{\delta_i} \\ &\quad \times S_i(t_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times \int [\prod_j P\{Y_i(s_{ij}) | c_i = g, b_i; \boldsymbol{\theta}\}] P(b_i | c_i = g; \boldsymbol{\theta}) db_i \Bigg\} \\ &= \sum_{i=1}^{n} \log \Bigg\{ \sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) h_i(t_i | c_i = g; \boldsymbol{\theta})^{\delta_i} \\ &\quad \times S_i(t_i | c_i = g; \boldsymbol{\theta}) \times P(\boldsymbol{Y}_i | c_i = g; \boldsymbol{\theta}) \Bigg\}\end{aligned}$$

که در مقایسه با مدل توأم کلاسیک به‌راحتی قابل‌دستیابی است، چون برای محاسبه تابع بقا و درستنمایی نیاز به روش‌های عددی برای محاسبه انتگرال‌ها ندارد. به‌خصوص، انتگرال موجود در تعریف تابع بقا یک فرم بسته دارد (البته به تابع مخاطره پایه هم‌بستگی دارد)، چون هیچ مؤلفه وابسته به زمان از فرایند طولی را شامل نمی‌شود و همچنین انتگرال نهایی به‌شرط اثرات تصادفی دارای فرم بسته‌ای است، زیرا تنها شامل مدل طولی شرطی $P(y_i|c_i = g, b_i)$ و چگالی اثرات تصادفی $P(b_i|c_i = g)$ است که تحت فرض نرمال هر دو عبارت منجر به یک توزیع گوسی چندمتغیره می‌شود. بااین‌وجود، مسئله چالش‌برانگیز در مدل‌های توأم رده پنهان این است که ممکن است لگاریتم تابع درستنمایی منجر به چندین ماکسیمم موضعی شود، که در این صورت توصیه می‌شود با چندین بار استفاده از مجموعه‌های متفاوت از مقادیر اولیه مدل را اصلاح کنیم و همگرایی را بررسی کنیم. در حقیقت واضح است که این مشکل، تقریباً مزیت محاسباتی نداشتن تقریب عددی انتگرال‌ها را باطل می‌کند، زیرا اصلاح مجدد مدل، حتی اگر ساده‌تر از برازش مدل توأم کلاسیک باشد، همچنان دارای پیچیدگی محاسباتی است. این مسئله همچنین با توجه به اینکه از ابتدا تعداد مناسب رده‌های پنهان مشخص نیست، افزایش می‌یابد. بنابراین، لازم است چندین مدل با تعداد رده‌های فزاینده برازش داده شوند و ازنظر آماری، تعداد رده‌ها را به‌گونه‌ای انتخاب کنیم که بهترین برازش را روی داده‌ها انجام داده باشد. این انتخاب معمولاً بر اساس معیارهای انتخاب مدل مانند AIC یا BIC انجام می‌شود. مسئله دیگر در مورد مدل‌های توأم رده پنهان این است که تفسیر در مورد ساختار اشتراک زیر مدل طولی و زمان بقا ساده نیست. تحت فرمول‌بندی رده پنهان، هیچ مجموعه‌ای از پارامترها وجود ندارد که به‌طور مستقیم قدرت پیوند دو فرایند طولی و خطر یک رخداد را اندازه‌گیری کند، که در بسیاری از موارد موضوع اصلی است. بنابراین، در اصل این نوع مدل‌بندی توأم، زمانی که علاقه‌مند به بهبود ناهمگنی جامعه هدف هستیم، مفید است و نه زمانی که علاقه‌مند به ارائه یک تفسیر ساده از مکانیسم ارتباط دو فرایند باشیم [۳۵].

با استفاده از مدل برازش داده‌شده، همچنین می‌توان طبقه‌بندی پسین بیماران را در نمونه به دست آورد که با استفاده از ماکسیمم احتمالات پسینی حاصل می‌شود:

$$\hat{\pi}_{ig}^{(Y,T)} = \Pr\left(c_i = g \mid T_i, \delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\theta}}\right) = \frac{\pi_{ig}(\hat{\boldsymbol{\theta}}) h_i\left(T_i | c_i = g; \hat{\boldsymbol{\theta}}\right)^{\delta_i} \mathcal{S}_i\left(T_i | c_i = g; \hat{\boldsymbol{\theta}}\right) P\left(\boldsymbol{Y}_i | c_i = g; \hat{\boldsymbol{\theta}}\right)}{\sum_{l=1}^{G} \pi_{ig}(\hat{\boldsymbol{\theta}}) h_i\left(T_i | c_i = l; \hat{\boldsymbol{\theta}}\right)^{\delta_i} \mathcal{S}_i\left(T_i | c_i = l; \hat{\boldsymbol{\theta}}\right) P\left(\boldsymbol{Y}_i | c_i = l; \hat{\boldsymbol{\theta}}\right)},$$

بنابراین، رابطه $\hat{c}_i = \arg\max_g \left\{\Pr\left(c_i = g \mid T_i, \delta_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\theta}}\right)\right\}$ نتیجه می‌دهد که فرد iام در رده g طبقه‌بندی‌شده است، که ازنظر مفهومی مشابه برآورد بیز تجربی اثرات تصادفی است.

۱.۱.۲ استفاده از الگوریتم EM برای برآورد پارامترهای مدل رده پنهان در داده‌های طولی و برآمد بقا

ماکسیمم کردن مستقیم رابطه (۵) با جمع رده‌های پنهان در لگاریتم و با صدها پارامتر بالقوه، ممکن است دارای پیچیدگی‌های محاسباتی بسیار باشد، ازاین‌رو، بسیاری از محققین حوزه مدل‌بندی توأم داده‌های طولی و زمان بقا بر اساس مدل رده پنهان، از الگوریتم EM برای برآورد پارامترهای مدل توأم استفاده کرده‌اند. بر اساس رابطه (۵)، تابع امتیاز لگاریتم درستنمایی داده‌های مشاهده‌شده را می‌توان به‌صورت زیر به دست آورد:

$$\begin{aligned}S(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}'} \log P(T_i, \delta_i, \boldsymbol{Y}_i; \boldsymbol{\theta}) \qquad (۶) \\ &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}'} \log \int \sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta}) d\boldsymbol{b}_i \\ &= \sum_i \frac{1}{P(T_i, \delta_i, \boldsymbol{Y}_i; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}'} \int \sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta}) d\boldsymbol{b}_i \\ &= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{Y}_i; \boldsymbol{\theta})} \int \frac{\partial}{\partial \boldsymbol{\theta}'} \{\sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta})\} d\boldsymbol{b}_i \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}'} \log\{\sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta})\} \\ &\quad \times \frac{\sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta})}{p(T_i, \delta_i, \boldsymbol{Y}_i; \boldsymbol{\theta})} \\ &\quad \times \frac{p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta})}{1} d\boldsymbol{b}_i \\ &= \sum_i \int \frac{\partial}{\partial \boldsymbol{\theta}'} \log\{\sum_{g=1}^{G} \pi_{ig}(\boldsymbol{\theta}) p(T_i, \delta_i | c_i = g; \boldsymbol{\theta}) \\ &\quad \times p(\boldsymbol{Y}_i | c_i = g, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | c_i = g; \boldsymbol{\theta})\} \\ &\quad \times p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{Y}_i, c_i = g; \boldsymbol{\theta}) d\boldsymbol{b}_i.\end{aligned}$$

بنابراین، ملاحظه می‌کنیم که بردار امتیاز داده‌های مشاهده‌شده به‌صورت مقدار امید ریاضی بردار امتیاز داده‌های کامل نسبت به توزیع پسین اثرات تصادفی به دست می‌آید. به عبارتی اگر عضویت رده c_i و اثرات تصادفی b_iها را به‌عنوان داده‌های گمشده در نظر بگیریم، رابطه حاصل برای $S(\boldsymbol{\theta})$ نقش دوگانه ایفا می‌کند. به‌طور خاص، اگر معادلات امتیاز را نسبت به $\boldsymbol{\theta}$ با فرض اینکه $p(\boldsymbol{b}_i|T_i,\delta_i,\boldsymbol{Y}_i,c_i=g;\boldsymbol{\theta})$ در مقدار تکرار قبلی از $\boldsymbol{\theta}$ ثابت باشد، حل کنیم، یک الگوریتم EM خواهیم داشت، درحالی‌که اگر معادلات امتیاز را نسبت به $\boldsymbol{\theta}$ و با فرض اینکه $p(\boldsymbol{b}_i|T_i,\delta_i,\boldsymbol{Y}_i,c_i=g;\boldsymbol{\theta})$ تابعی از $\boldsymbol{\theta}$ است، حل کنیم، متناظر با ماکسیمم سازی مستقیم لگاریتم درستنمایی داده‌های مشاهده‌شده خواهد بود. اگر $\breve{X}_i$ نشان‌دهنده بردار ترکیب متغیرهای کمکی X_i، Z_i، w_i و u_i تعریف‌شده در رابطه (۴) باشد، در این صورت لگاریتم درستنمایی داده‌های کامل $(\boldsymbol{Y}_i,T_i,\delta_i,\boldsymbol{b}_i,c_i)$ به‌صورت زیر قابل‌محاسبه است:

(۷)
$$\ell_{com}(\boldsymbol{\theta};\boldsymbol{Y},T,\delta,\breve{X},\boldsymbol{b},c)=\sum_{i=1}^{n}\Big\{\log Pr\left(c_i\mid\breve{X}_i\right)+\log P(\boldsymbol{b}_i|c_i)+\log P\left(\boldsymbol{Y}_i\mid b_i,c_i,\breve{X}_i\right)+\log P\left(T_i,\delta_i\mid c_i,\breve{X}_i\right)\Big\},$$

بسط هر یک از عبارت‌های رابطه (۷) را می‌توان به‌صورت زیر محاسبه کرد:

(۸)
$$\log Pr\left(c_i\mid\breve{X}_i\right)=\sum_{g=1}^{G}c_{ig}\left\{\boldsymbol{u}_i{}'\boldsymbol{\lambda}_g-\log\left[\sum_{L=1}^{G}exp(\boldsymbol{u}_i{}'\boldsymbol{\lambda}_g)\right]\right\},$$

$$\begin{aligned}\log P\left(\boldsymbol{b}_i\mid c_i\right) &= \sum_{g=1}^{G}c_{ig}\left(\log(\pi_{ig})+\log\phi_q(\boldsymbol{b}_{ig}\mid\boldsymbol{\mu}_g,D_g)\right)\\ &= \sum_{g=1}^{G}c_{ig}\left\{\boldsymbol{u}_i{}'\boldsymbol{\lambda}_g-\log\left[\sum_{L=1}^{G}exp(\boldsymbol{u}_i{}'\boldsymbol{\lambda}_g)\right]\right.\\ &\quad -\frac{q}{2}\log(2\pi)-\frac{1}{2}\log|D_g|\\ &\quad \left.-\frac{(\boldsymbol{b}_{ig}-\boldsymbol{\mu}_g)'D_g^{-1}(\boldsymbol{b}_{ig}-\boldsymbol{\mu}_g)}{2}\right\},\end{aligned}\qquad (۹)$$

$$\begin{aligned}\log P\left(Y_i\mid\boldsymbol{b}_i,c_i,\breve{X}_i\right) &= \sum_{g=1}^{G}c_{ig}\log(\boldsymbol{Y}_i(s)\mid\boldsymbol{b}_i,c_i,\breve{X}_i)\qquad (۱۰)\\ &= \sum_{g=1}^{G}c_{ig}\left\{-\frac{n_i}{2}\log(2\pi)-\frac{1}{2}\log(\sigma^2)\right.\\ &\quad -\frac{1}{2}\left(\boldsymbol{Y}_i(s)-X_i'\boldsymbol{\beta}_g-Z_i'\boldsymbol{b}_{ig}\right)'\\ &\quad \left.\times\sigma^{-2}I_{n_i}\left(\boldsymbol{Y}_i(s)-X_i'\boldsymbol{\beta}_g-Z_i'\boldsymbol{b}_{ig}\right)\right\},\end{aligned}$$

$$\begin{aligned}\log P\left(T_i,\delta_i\mid c_i,\breve{X}_i\right) &= \sum_{g=1}^{G}c_{ig}\left\{\delta_i\log h_i(t_i\mid c_i,\breve{X}_i)\right.\qquad (۱۱)\\ &\quad \left.+\log S_i(t_i|c_i,\breve{X}_i)\right\}\\ &= \sum_{g=1}^{G}c_{ig}\left\{\delta_i(\log h_{\circ g}(t_i)+\boldsymbol{w}_i{}'\boldsymbol{\gamma}_g)\right.\\ &\quad \left.+\log\left\{\int_{t_i}^{\infty}h_{\circ g}(t)\,\exp(\boldsymbol{w}_i{}'\boldsymbol{\gamma}_g)\,dt\right\}\right\}.\end{aligned}$$

تابع $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ را به‌عنوان مقدار امید ریاضی تابع لگاریتم درستنمایی داده‌های کامل نسبت به توزیع شرطی داده‌های مشاهده نشده به‌شرط مشاهدات و مقدار پارامتر در مرحله rام به‌صورت زیر تعریف می‌کنیم:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})=E_{\boldsymbol{b},c|\boldsymbol{Y},T,\delta,\breve{X};\boldsymbol{\theta}^{(r)}}[\ell_c(\boldsymbol{\theta};\boldsymbol{Y},T,\delta,\breve{X},\boldsymbol{b},c)],\quad (۱۲)$$

بنابراین با جایگذاری معادلات (۸) در تابع $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ خواهیم داشت:

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= \sum_{i=1}^{n}E_{\boldsymbol{b},c|\boldsymbol{Y},T,\delta,\breve{X};\boldsymbol{\theta}^{(r)}}\left[\ell_{com}(\boldsymbol{\theta};\boldsymbol{Y}_i,T_i,\delta_i,\breve{X}_i,\boldsymbol{b}_i,c_i)\right]\\ &= \sum_{i=1}^{n}E_{\boldsymbol{b},c|\boldsymbol{Y},T,\delta,\breve{X};\boldsymbol{\theta}^{(r)}}\left[\log Pr\left(c_i\mid\breve{X}_i\right)\right.\\ &\quad +\log P(\boldsymbol{b}_i|c_i,\breve{X})+\log P\left(\boldsymbol{Y}_i\mid b_i,c_i,\breve{X}_i\right)\\ &\quad \left.+\log P\left(T_i,\delta_i\mid c_i,\breve{X}_i\right)\right]\\ &= \sum_{i=1}^{n}\left[2\sum_{g=1}^{G}\left\{E(c_{ig}|\boldsymbol{Y}_i,T_i,\delta_i,\breve{X};\boldsymbol{\theta}^{(r)})\right.\right.\\ &\quad \left.\times\left(\boldsymbol{u}_i'\boldsymbol{\lambda}_g-\log\left(\sum_{L=1}^{G}exp(\boldsymbol{u}_i{}'\boldsymbol{\lambda}_g)\right)\right)\right\}\\ &\quad +\sum_{g=1}^{G}E(c_{ig}|\boldsymbol{Y}_i,T_i,\delta_i,\breve{X};\boldsymbol{\theta}^{(r)})\\ &\quad \times\left(-\frac{q+n_i}{2}\log(2\pi)-\frac{1}{2}\log|D_g|-\frac{1}{2}\log(\sigma^2)\right)\\ &\quad +\sum_{g=1}^{G}\left\{-\frac{1}{2}E\left(c_{ig}\boldsymbol{b}_{ig}'D_g^{-1}\boldsymbol{b}_{ig}-2c_{ig}\boldsymbol{\mu}_g'D_g^{-1}\boldsymbol{b}_{ig}\right.\right.\\ &\quad \left.\left.+c_{ig}\boldsymbol{\mu}_g'D_g^{-1}\boldsymbol{\mu}_g|\boldsymbol{Y}_i,T_i,\delta_i,\breve{X};\boldsymbol{\theta}^{(r)}\right)\right\}\\ &\quad +\sum_{g=1}^{G}\left\{-\frac{1}{2}E\left(\sigma^{-2}c_{ig}\left(\boldsymbol{Y}_i(s)-X_i'\boldsymbol{\beta}_g\right)'\left(\boldsymbol{Y}_i(s)-X_i'\boldsymbol{\beta}_g\right)\right.\right.\\ &\quad -2\sigma^{-2}c_{ig}\left(\boldsymbol{Y}_i(s)-X_i'\boldsymbol{\beta}_g\right)'Z_i'\boldsymbol{b}_{ig}\\ &\quad \left.\left.+\sigma^{-2}c_{ig}\boldsymbol{b}_{ig}'Z_iZ_i'\boldsymbol{b}_{ig}|\boldsymbol{Y}_i,T_i,\delta_i,\breve{X};\boldsymbol{\theta}^{(r)}\right)\right\}\\ &\quad +\sum_{g=1}^{G}E(c_{ig}|\boldsymbol{Y}_i,T_i,\delta i,\breve{X};\boldsymbol{\theta}^{(r)})\left((\log h_{\circ g}(t_i)+\boldsymbol{w}_i{}'\boldsymbol{\gamma}_g)\right.\\ &\quad \left.+\log\left\{\int_{t_i}^{\infty}h_{\circ g}(t)\,\exp(\boldsymbol{w}_i{}'\boldsymbol{\gamma}_g)\,dt\right\}\right)+\sum_{g=1}^{G}E\left(c_{ig}\right.\\ &\quad \left.\left.\log\left\{\int_{t_i}^{\infty}h_{\circ g}(t)\,\exp(\boldsymbol{w}_i{}'\boldsymbol{\gamma}_g)\,dt\right\}|\boldsymbol{Y}_i,T_i^{obs},\breve{X};\boldsymbol{\theta}^{(r)}\right)\right]\end{aligned}$$

گام E الگوریتم EM: محاسبه امید ریاضی‌های شرطی

با فرض این‌که، $\tilde{x}=E\left(x\mid y,T,\delta,\breve{X}\right)\equiv\tilde{E}(x)$ نشان‌دهنده امید ریاضی شرطی (پسینی) متغیر تصادفی x به‌شرط داده‌های مشاهده‌شده باشد، می‌توان امید ریاضی‌های شرطی را به‌صورت زیر به دست آورد.

ابتدا امید شرطی عضویت رده را به‌صورت زیر محاسبه می‌کنیم:

(۱۳)
$$\begin{aligned}\tilde{c}_{ig}&=E\left(c_{ig}\mid\boldsymbol{Y}_i,T_i,\delta_i,\breve{X}_i;\boldsymbol{\theta}^{(r)}\right)\\ &=\frac{\pi_{ig}P(\boldsymbol{Y}_i\mid c_{ig}=1,\breve{X}_i;\boldsymbol{\theta}^{(r)})P(T_i,\delta_i\mid c_{ig}=1,\breve{X}_i;\boldsymbol{\theta}^{(r)})}{\sum_{g=1}^{G}\pi_{ig}P(Y_i\mid c_{ig}=1,\breve{X}_i;\boldsymbol{\theta}^{r})P(T_i,\delta_i\mid c_{ig}=1,\breve{X}_i;\boldsymbol{\theta}^{(r)})},\end{aligned}$$

که در آن

$$P(Y_i \mid \breve{X}_i, c_{ig}=1) = \mathrm{MVN}_{n_i}\left(X_i\beta, Z_i\sigma_g^2 D Z_i^{'} + \sigma^2 \mathrm{I}_{n_i}\right) \quad (14)$$

و $\mathrm{MVN}_{n_i}(\boldsymbol{m}u, \Sigma)$ نشان‌دهنده چگالی نرمال چندمتغیره n_i بعدی با میانگین $\boldsymbol{\mu}$ و واریانس Σ است.

همچنین، $P(T_i, \delta_i \mid c_{ig}=1, \breve{X}_i; \boldsymbol{\theta}^{(r)})$ از رابطه زیر به دست می‌آید:

$$P(T_i, \delta_i \mid c_{ig} = 1, \breve{X}_i; \boldsymbol{\theta}^{(r)}) = h_{\circ g}(t)\, \exp(\boldsymbol{w}_i{}^{'}\boldsymbol{\gamma}_g{}^{(r)}) \times \left\{\int_{t_i}^{\infty} h_{\circ g}(t)\, \exp(\boldsymbol{w}_i{}^{'}\boldsymbol{\gamma}_g{}^{(r)})\, dt\right\}.$$

توزیع پسینی b_{ig}ها که در پیوست محاسبه‌شده است به‌صورت نرمال چندمتغیره به ترتیب با میانگین $\boldsymbol{\mu}_g^*$ و واریانس V_i به دست می‌آید:

$$\boldsymbol{b}_{ig} \mid c_{ig}, \boldsymbol{Y}_i, \breve{X}_i; \boldsymbol{\theta} \sim MVN_q(\boldsymbol{\mu}_g^*, V_i),$$

که در آن $V_i = \left(Z_i Z_i^{'}\sigma^{-2} + D_g^{-1}\right)^{-1}$ و $\boldsymbol{\mu}_g^* = V_i\left(Z_i(\boldsymbol{Y}_i - X_i^{'}\boldsymbol{\beta}_g)\sigma^{-2} + D_g^{-1}\boldsymbol{\mu}_g\right)$.

بر این اساس، سایر امیدهای شرطی را به‌صورت زیر محاسبه می‌کنیم. لازم به ذکر است که جزییات بیشتر و نحوه به دست آوردن این امید ریاضی‌های شرطی در پیوست آورده شده است:

$$\begin{aligned}
&\tilde{\boldsymbol{b}_{ig}} = V_i\left(Z_i(\boldsymbol{Y}_i - X_i^{'}\boldsymbol{\beta}_g^{(r)})\sigma^{-2(r)} + D_g^{-1}\boldsymbol{\mu}_g{}^{(r)}\right), \quad (15)\\
&\widetilde{\boldsymbol{b}_{ig}\boldsymbol{b}_{ig}^{'}} = V_i + \tilde{\boldsymbol{b}_{ig}}\left(\left((\boldsymbol{Y}_i - X_i^{'}\boldsymbol{\beta}_g^{(r)})^{'} Z_i^{'}\sigma^{-2(r)} + \boldsymbol{\mu}_g{}^{'(r)} D_g^{-1}\right) V_i\right),\\
&\widetilde{c_{ig}\boldsymbol{b}_{ig}} = \tilde{c_{ig}}\left(\tilde{\boldsymbol{b}_{ig}}\right),\\
&\widetilde{c_{ig}\boldsymbol{b}_{ig}^{'} D_g^{-1}\boldsymbol{b}_{ig}} = tr\left(\tilde{c}_{ig} D_g^{-1}\left(\widetilde{\boldsymbol{b}_{ig}\boldsymbol{b}_{ig}^{'}}\right)\right),\\
&\widetilde{c_{ig}\boldsymbol{b}_{ig}^{'} Z_i Z_i^{'}\boldsymbol{b}_{ig}} = tr\left(\tilde{c}_{ig} Z_i Z_i^{'}\left(\widetilde{\boldsymbol{b}_{ig}\boldsymbol{b}_{ig}^{'}}\right)\right),
\end{aligned}$$

در نهایت برای محاسبه

$$E(c_{ig}\log\left\{\int_{t_i}^{\infty} h_{\circ g}(t)\, \exp(\boldsymbol{w}_i{}^{'}\boldsymbol{\gamma}_g)\, dt\right\} | \boldsymbol{Y}_i, Ti^{obs}, \delta_i, \breve{X}; \boldsymbol{\theta}^{(r)}$$

روش‌های عددی اتخاذ می‌گردد که در آن T_i^{obs} نشان‌دهنده زمان‌های مشاهده‌شده با نشانگر $\delta_i = 1$ است.

گام M الگوریتم EM: بیشینه‌سازی

در این مرحله، بر اساس امیدهای شرطی حاصل در گام قبلی، تابع $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ را در هر تکرار از الگوریتم بیشینه می‌کنیم و مقدار جدید برآوردگر درستنمایی ماکسیمم پارامترها را به دست می‌آوریم. عبارت‌های زیر با فرم بسته برای پارامترها به دست می‌آید:

$$\begin{aligned}
\hat{D}_g^{(r+1)} &= \arg\max_{D_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\} \quad (16)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\tilde{\boldsymbol{b}_{ig}} - \boldsymbol{\mu_g}\right)\left(\tilde{\boldsymbol{b}_{ig}} - \boldsymbol{\mu_g}\right)^{'}\right\}\\
\hat{\boldsymbol{\beta}}_g^{(r+1)} &= \arg\max_{\boldsymbol{\beta}_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\}\\
&= \left(\sum_{i=1}^{n} X_i X_i^{'}\right)^{-1}\sum_{i=1}^{n} X_i\left(\boldsymbol{Y}_i - Z_i\tilde{\boldsymbol{b}}_{ig}\right)^{'}\\
\hat{\sigma^2}^{(r+1)} &= \arg\max_{\sigma^2}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\}\\
&= \frac{1}{\sum_{i=1}^{n} n_i}\sum_{i=1}^{n}\left\{(\boldsymbol{Y}_i - X_i^{'}\hat{\boldsymbol{\beta}}_g)^{'}\left(\boldsymbol{Y}_i - X_i^{'}\hat{\boldsymbol{\beta}}_g - 2Z_i^{'}\tilde{b_{ig}}\right)\right.\\
&\quad \left. + b_{ig}^{'}\widetilde{Z_i Z_i^{'}} b_{ig}\right\}\\
\hat{\boldsymbol{\mu}}_g^{(r+1)} &= \arg\max_{\boldsymbol{\mu}_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\}\\
&= \sum_{i=1}^{n}\widetilde{c_{ig}\boldsymbol{b}_{ig}}\,(\tilde{c_{ig}})^{-1},
\end{aligned}$$

همچنین، برای پارامترهای زیر فرم بسته وجود ندارد و از روش‌های عددی برای برآورد آن‌ها استفاده می‌شود:

$$\hat{\boldsymbol{\lambda}}_g^{(r+1)} = \arg\max_{\boldsymbol{\lambda}_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\},\ \hat{\gamma}_g^{(r+1)} = \arg\max_{\gamma_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\}.$$

گام‌های E و M الگوریتم را تا جایی ادامه می‌دهیم که معیار همگرایی زیر برقرار گردد:

$$\ell(\boldsymbol{\theta}^{(r+1)}) - \ell(\boldsymbol{\theta}^{(r)}) < \epsilon,$$

که در آن ϵ مقدار مثبت بسیار کوچکی در نظر گرفته می‌شود و تابع $\ell(\boldsymbol{\theta})$ لگاریتم درستنمایی داده‌های مشاهده‌شده می‌باشد.

۲.۲ مدل‌بندی توأم اندازه‌های طولی و داده‌های مخاطره‌های رقابتی

برای تعریف مدل رده پنهان به‌منظور مدل‌بندی توأم اندازه‌های طولی و داده‌های مخاطره‌های رقابتی، مشابه بخش قبلی فرض می‌کنیم که جامعه اصلی از G زیرجامعه تشکیل‌شده باشد و نشانگر رده پنهان را به‌صورت $c_i = g$، $g = 1, ..., G$ تعریف می‌کنیم که عضویت iامین مؤلفه را در رده پنهان g نشان می‌دهد. فرض می‌کنیم T_{ik}^* زمان رخداد عامل kام، $k = 1, ..., K$ و $\tilde{T}_i$ زمان سانسور برای مؤلفه iام باشد، در این صورت $T_i = min(\tilde{T}_i, T_{i1}^*, ..., T_{iK}^*)$ زمان رخداد مشاهده‌شده خواهد بود. δ_i را به‌عنوان نشانگر رخداد در نظر می‌گیریم که $\delta_i = k$ نشان‌دهنده این است که عامل kام زودتر از بقیه رخ‌داده است و $\delta_i = 0$ اگر فرد iام سانسور شده باشد. در ادامه سه زیرمدلی که مشخص‌کننده مدل رده پنهان توأم است را معرفی می‌کنیم. یک مدل آمیخته خطی برای نشانگر طولی پیوسته با پارامترهای رده مشخص

که توزیع شرطی $\boldsymbol{Y}_i$ها را نسبت به $c_i = g$ توصیف می‌کند را به‌صورت زیر در نظر می‌گیریم:

$$\begin{aligned} \boldsymbol{Y}_i(s) &= X_i'(s)\boldsymbol{\beta}_g + Z_i'(s)\boldsymbol{b}_{ig} + \epsilon_i(s), \\ \epsilon_i(s) &\sim N(۰, \sigma^۲),\ i = ۱, ..., n, \end{aligned} \qquad (۱۷)$$

اندازه‌های تکرارشونده نشانگر طولی به‌صورت $\boldsymbol{Y}_i(s) = (Y_i(s_{i۱}), ..., Y_i(s_{in_i}))'$ در زمان‌های متفاوت اندازه‌گیری $i = ۱, ..., n$، s_{ij}، $j = ۱, ..., n_i$، در نظر گرفته می‌شود که در آن X_i بردار ضرایب متناظر با اثرات ثابت $\boldsymbol{\beta}_g$ و Z_i بردار ضرایب متناظر با اثرات تصادفی $\boldsymbol{b}_{ig}$ برای فرد iام در رده g است. بردار $\boldsymbol{b}_{ig}$ دارای توزیع نرمال با میانگین و ماتریس کوواریانس رده مشخص به‌صورت $\boldsymbol{b}_{ig} \sim N(\boldsymbol{\mu}_g, \sigma_g^۲ D)$، در نظر گرفته می‌شود و در آن $\sigma_G = ۱$. همچنین یک مدل مخاطره متناسب رده‌ـ مشخص برای مخاطره‌های رقابتی تعریف می‌کنیم. در این مدل فرض می‌شود که مخاطره رخداد برای هر رده همگن است و بر اساس متغیرهای کمکی شرح داده می‌شود. مخاطره رخداد عامل k در رده g برای فرد iام به‌صورت زیر در نظر گرفته می‌شود:

$$\begin{aligned} h_{ik}(t|c_i = g) = h_{۰kg}(t, \nu_{kg}) exp(\boldsymbol{w}_{ik}{}' \gamma_{kg}), \\ k = ۱, ..., K,\ g = ۱, ..., G, \end{aligned} \qquad (۱۸)$$

که در آن $\boldsymbol{w}_i$ بردار متغیرهای کمکی متناظر با مخاطره رخداد عامل k در رده g با بردار پارامترهای γ_{kg} است. تابع مخاطره پایه $h_{۰kg}$ برای هر رده و هر عامل مشخص است و به‌صورت پارامتری (با ν_{kg}) با استفاده از یک توزیع بقا استاندارد مثل وایبول یا گامپرتز و یا با استفاده از تعداد کمی از توابع پله‌ای یا Mـ اسپلاین‌ها برای انعطاف بیشتر، مدل‌بندی می‌شود. سرانجام یک مدل رگرسیون چندجمله‌ای بر اساس متغیرهای کمکی برای مدل احتمال عضویت رده پنهان به‌صورت زیر در نظر می‌گیریم:

$$\pi_{ig} = Pr(c_i = g) = \frac{exp(\boldsymbol{u}_i{}' \boldsymbol{\lambda}_g)}{\sum_{L=۱}^{G} exp(\boldsymbol{u}_i{}' \boldsymbol{\lambda}_g)}, \qquad (۱۹)$$

که در آن $\boldsymbol{u}_i$ برداری از متغیرهای کمکی مرتبط با بردار پارامترهای رده‌ـ مشخص $\boldsymbol{\lambda}_g$ است. $\boldsymbol{u}_i$ شامل عرض از مبدأ است و در عمل، زمانی که هیچ پیش‌بینی‌کننده‌ای برای عضویت رده فرض نمی‌شود به عرض از مبدأ کاهش می‌یابد. برای شناساپذیری، یک رده مرجع لازم است که رده G را انتخاب می‌کنیم، در نتیجه $\boldsymbol{\lambda}_G = ۰$.

مدل رده پنهان توأم اندازه‌های طولی و داده‌های مخاطره‌های رقابتی با ماکسیمم کردن لگاریتم درستنمایی برای تعداد معینی از رده پنهان G برآورد می‌شود، سپس مدل‌هایی با مقادیر متفاوتی از G توسط برخی معیارها، مانند BIC، باهم مقایسه می‌شوند. برداری که شامل تمام پارامترهای سه زیر مدل تعریف‌شده (۱)،(۱۸) و (۱۹) با G رده پنهان است را با $\boldsymbol{\theta}$ نمایش می‌دهیم. با در نظر گرفتن مفروضات استقلال شرطی (۳) که در بخش قبلی معرفی شدند، لگاریتم درستنمایی را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=۱}^{n} \log \sum_{g=۱}^{G} \{ & P(\boldsymbol{Y}_i \mid c_i = g; \boldsymbol{\theta}) \\ & \times P(T_i, \delta_i \mid c_i = g; \boldsymbol{\theta})\, Pr(c_i = g \mid \boldsymbol{\theta}) \}, \end{aligned} \qquad (۲۰)$$

در این رابطه، $Pr(c_i = g \mid \boldsymbol{\theta}) = \pi_{ig}$ از فرمول (۱۹) به دست می‌آید و تابع چگالی مدل مخاطره‌های رقابتی به‌صورت زیر بیان می‌شود:

$$\begin{aligned} P(T_i, \delta_i \mid c_i = g; \boldsymbol{\theta}) = \exp\left\{ -\sum_{k=۱}^{K} H_k(T_i \mid c_i = g; \boldsymbol{\theta}) \right\} \\ \times \prod_{k=۱}^{K} h_k(T_i \mid c_i = g; \boldsymbol{\theta})^{I(\delta_i = k)}, \end{aligned}$$

که در آن $h_k(T_i \mid c_i = g; \theta)$ مخاطره لحظه‌ای علت‌ـ مشخص تعریف‌شده در رابطه (۱۸) و $H_k(T_i \mid c_i = g; \theta)$ تابع مخاطره تجمعی متناظر با آن است. همچنین،

$$\begin{aligned} P(\boldsymbol{Y}_i \mid c_i = g; \boldsymbol{\theta}) = \int [\prod_j P(Y_i(s_{ij}) | c_i = g, b_i; \boldsymbol{\theta})] \\ \times P(b_i | c_i = g; \boldsymbol{\theta}) db_i,\ j = ۱, ..., n_i. \end{aligned}$$

همچنین، احتمالات پسینی را می‌توان از مدل توأم به دست آورد. احتمالات پسینی عضویت رده به‌شرط تمام مشاهدات داده‌شده برای نشانگر طولی و رخدادها را که به‌صورت $\pi_{ig}^{(Y,T)}(\boldsymbol{\theta}) = Pr(c_i = g | \boldsymbol{Y_i}, T_i, \delta_i; \boldsymbol{\theta})$ تعریف می‌کنیم را برای طبقه‌بندی پسینی افراد مورداستفاده قرار می‌دهیم، که در آن

$$\begin{aligned} \hat{\pi}_{ig}^{(Y,T)} &= \Pr\left(c_i = g \mid Y_i, \tilde{T}_i, \delta_i; \hat{\boldsymbol{\theta}} \right) \\ &= \frac{\hat{\pi}_{ig} P(Y_i \mid c_i = g; \hat{\boldsymbol{\theta}}) P(T_i, \delta_i \mid c_i = g; \boldsymbol{\theta})}{\sum_{l=۱}^{G} \hat{\pi}_{il} P(Y_i \mid c_i = l; \hat{\boldsymbol{\theta}}) P(T_i, \delta_i \mid c_i = g; \boldsymbol{\theta})}, \end{aligned}$$

و

$$\hat{c}_i = \arg\max_g \left\{ \Pr\left(c_i = g \mid \boldsymbol{Y}_i, T_i, \delta_i; \hat{\boldsymbol{\theta}} \right) \right\}.$$

۱.۰.۲ استفاده از الگوریتم EM برای برآورد پارامترهای مدل رده پنهان در داده‌های طولی و برآمدهای مخاطره‌های رقابتی

مشابه آنچه در زیر بخش قبلی برای برآورد پارامترهای مدل توأم رده پنهان داده‌های طولی و بقا با استفاده از الگوریتم EM شرح دادیم، در اینجا نیز می‌توان با استفاده از این الگوریتم پارامترهای مدل رده پنهان را برای مدل‌بندی توأم اندازه‌های طولی و داده‌های مخاطره رقابتی به دست آورد. تنها با این تفاوت که توزیع زمان رخداد مشاهدات با استفاده از مخاطره‌های رخداد K عامل تعریف‌شده در مدل به دست می‌آید. بنابراین خواهیم داشت:

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(r)}) &= \sum_{i=۱}^{n} E_{\boldsymbol{b}, c | \boldsymbol{Y}, T, \delta, \breve{X}; \boldsymbol{\theta}^{(r)}} \left[\ell_{com}(\boldsymbol{\theta}; \boldsymbol{Y}_i, T_i, \delta_i, \breve{X}_i, \boldsymbol{b}_i, c_i) \right] \\ &= \sum_{i=۱}^{n} E_{\boldsymbol{b}, c | \boldsymbol{Y}, T, \delta, \breve{X}; \boldsymbol{\theta}^{(r)}} \left[\log Pr\left(c_i \mid \breve{X}_i \right) \right. \\ &\quad + \log P(\boldsymbol{b}_i | c_i, \breve{X}) + \log P\left(\boldsymbol{Y}_i \mid b_i, c_i, \breve{X}_i \right) \\ &\quad \left. + \log P\left(T_i, \delta_i \mid c_i, \breve{X}_i \right) \right] \end{aligned} \qquad (۲۱)$$

که در آن

$$\log P\left(T_i,\delta_i \mid c_i,\breve{X}_i\right) = -\sum_{k=1}^{K} H_k\left(T_i \mid c_i=g;\boldsymbol{\theta}\right) + \sum_{k=1}^{K} I(\delta_i=k)\log\left\{h_k\left(T_i \mid c_i=g;\boldsymbol{\theta}\right)\right\}$$

و بقیه عبارت‌ها در رابطه (۲۱)، مشابه موارد متناظر در زیر بخش قبلی محاسبه می‌شوند. در نتیجه، برآوردگرهای درستنمایی ماکسیمم با استفاده از الگوریتم EM را می‌توان به دست آورد.

$$\hat{\boldsymbol{\theta}}_g^{(r+1)} = \arg\max_{\boldsymbol{\theta}_g}\{Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})\}.$$

۳ کاربرد

در نرم‌افزار R مدل‌های توأم رده پنهان با استفاده از تابع **Jointlcmm**() از بسته نرم‌افزاری **lcmm** برازش داده می‌شوند [۳۰]. برخلاف تابع **JointModel**() که ابتدا لازم است به‌طور جداگانه مدل‌های اثرات آمیخته خطی و بقا را برازش دهیم، تابع **Jointlcmm**() تنها نیاز به ارائه فرمول جداگانه دارد که قسمت‌های مختلف مدل را مشخص می‌کند.

به‌عنوان‌مثال، یک مدل رده پنهان از مجموعه داده ایدز برای مدل‌بندی توأم داده‌های طولی و زمان بقا و همچنین یک مدل رده پنهان برای داده‌های صرع به‌منظور مدل‌بندی توأم داده‌های طولی و مخاطره‌های رقابتی را بررسی می‌کنیم.

۱.۳ داده‌های ایدز

در مجموعه داده‌های ایدز، $n = ۴۶۷$ بیمار آلوده به ویروس نقص ایمنی انسانی (HIV) به‌طور تصادفی تحت درمان با دو داروی دیدانوسین (ddI) و زالتیسابین (ddC) قرار می‌گیرند. در این مطالعه، اندازه‌های مکرر تعداد سلول‌های $CD۴$ هر دو ماه از ابتدا مطالعه تا ۲۰ ماه ثبت می‌شوند. همچنین، پیشامد زمان تا مرگ نیز در نظر گرفته می‌شود. در این مطالعه، ارتباط بین تعداد $CD۴$ها و پیشامد مرگ، و همچنین تأثیر داروهای ddI و ddC بر روی این دو پیشامد موردتوجه قرار می‌گیرد. جزئیات بیشتر در مورد طراحی این مطالعه را می‌توان در [۱] یافت.

بنابراین، بر اساس این مجموعه داده، برای بخش طولی، در قسمت ثابت تأثیر اصلی زمان و درمان و در قسمت تصادفی عرض از مبدأ تصادفی و شیب تصادفی را در نظر می‌گیریم:

(۲۲)

$$y_i(s)_{|c_i=g} = \beta_{0g} + \beta_{1g}s + \beta_{2g}\mathrm{ddI}_i + b_{i0} + b_{i1}s + \varepsilon_i(s),$$
$$i = 1,...,n, \quad g = 1,...,G.$$

برای سادگی، فرض می‌کنیم که هر دو متغیر در قسمت اثرات ثابت به رده وابسته هستند، ازاین‌رو، $(b_{i0}, b_{i1})' \sim N_2(0, \boldsymbol{D})$ و $\epsilon_i(s)\sim N(0,\sigma^2)$ در نظر گرفته می‌شوند. برای زیرمدل بقا، فرض می‌کنیم که رده پنهان تابع مخاطره پایه و اثر درمان را مشخص می‌کند:

$$h_i\left(t \mid c_i=g\right) = h_{0g}(t)\exp\left(\gamma_g\ \mathrm{ddI}_i\right) \qquad (۲۳)$$

توابع خطر پایه رده‌ـ مشخص با پارامتر $\boldsymbol{\nu}_g$ به‌صورت تکه‌ای با شش گره و در صدک‌های مربوط به توزیع زمان رخداد فرض می‌شود، البته در نرم‌افزار R توابع پایه دیگری نیز موجود است.

سرانجام، در مدل چندجمله‌ای عضویت پنهان، فرض می‌کنیم که احتمال پیشین هر بیمار متعلق به یک رده خاص به درمان بستگی دارد، یعنی:

$$\Pr\left(c_i=g\right) = \frac{\exp\left(\lambda_{0g}+\lambda_{1g}\ \mathrm{ddI}_i\right)}{\sum_{l=1}^{G}\exp\left(\lambda_{0l}+\lambda_{1l}\mathrm{ddI}_i\right)}. \qquad (۲۴)$$

چهار مدل توأم را به ترتیب با دو، سه، چهار و پنج رده پنهان برازش می‌دهیم. مقادیر AIC و BIC برای $g = ۲,...,۵$ در جدول ۱ آورده شده است.

جدول ۱: مقادیر لگاریتم درستنمایی و معیارهای اطلاع برای تعداد رده پنهان مدل توأم برازش داده‌شده برای داده‌های ایدز

تعداد رده‌ها	loglik	AIC	BIC
۲	−۴۲۵۳٫۳۷	۸۵۵۴٫۷۵	۸۶۵۴٫۲۶
۳	−۴۲۱۷٫۶۶	۸۵۰۵٫۳۲	۸۶۵۰٫۴۴
۴	−۴۱۹۳٫۱۵	۸۴۷۸٫۳۰	۸۶۶۹٫۰۳
۵	−۴۱۸۷٫۲۷	۸۴۸۸٫۵۳	۸۷۲۴٫۳۷

اختلاف‌نظر شدیدی بین دو معیار اطلاعات مشاهده می‌شود. به این صورت که، AIC به‌طور پیوسته‌ای کاهش می‌یابد مادامی‌که تعداد رده‌ها افزایش می‌یابد و مدل چهار رده‌ای را توصیه می‌کند، درحالی‌که مدل بهینه مطابق با BIC، مدل سه رده است. مطالعات تجربی در این زمینه نشان داده است که BIC، اغلب مدل با تعداد صحیح زیرگروه‌های پنهان را پیشنهاد می‌کند [۲۳، ۲۹]. به دنبال این توصیه، مدل با سه رده پنهان را انتخاب می‌کنیم. همچنین، نتایج مدل رده پنهان توأم با سه رده پنهان برای این مجموعه داده، در جداول ۲ و ۳ خلاصه‌شده است.

بر اساس مقادیر آماره والد و پی‌ـمقدار به‌دست‌آمده از جدول۲، چنانچه پی‌ـمقدار کمتر از ۰٫۰۵ و قدر مطلق آماره والد بزرگ‌تر از ۱٫۹۶ (مقدار توزیع نرمال استاندارد در سطح۰٫۹۵) باشد، ضریب متغیر مستقل معنی‌دار خواهد بود. بنابراین، مشاهده می‌کنیم که برای برخی پارامترها، مثل پارامتر زمان در قسمت طولی تفاوت بین طبقات قابل‌توجه است، به این صورت که، اثر زمان به جزء رده ۳ همه‌جا معنادار است. همه رده‌ها دارای عرض از مبدأ هستند و اثر درمان در هیچ رده‌ای معنادار نیست. تمام پارامترهای مربوط به تابع خطر پایه مدل بقا در رده ۱ معنادار هستند، از طرفی داروی ddI در هیچ رده‌ای روی بقا مؤثر نیست. همچنین، برای مدل احتمال، عرض از مبدأ و اثر درمان برای رده ۱ معنادار هستند.

۲.۳ داده‌های صرع

داده‌هایی که در این بخش تحلیل می‌شوند، مربوط به مطالعه داروهای استاندارد و جدید ضد صرع $(SANAD)$[15] است، که در آن، بیماران مبتلا به صرع را به آزمایش AED[16]های جدید متقاعد می‌کنند.

در اینجا، تنها بیمارانی که با CBZ (کاربامازپین) یا LTG (لاموتریژین) تحت درمان هستند، در نظر گرفته می‌شوند. زمان تا رخداد عدم موفقیت درمان به‌عنوان پیشامد بقا ثبت می‌شود و تحلیل مخاطره‌های رقابتی برای عدم موفقیت درمان شامل عدم موفقیت درمان ناشی از کنترل نامناسب تشنج (ISC) یا اثرهای جانبی غیرقابل‌قبول (UAE) در نظر گرفته می‌شود [۲۶].

این مجموعه داده شامل ۶۰۵ بیمار است که به‌طور تصادفی ۲۹۲ نفر با CBZ و ۳۱۳ نفر با LTG تحت درمان هستند. ۹۴ بیمار به دلیل UAE و ۱۲۰ نفر به دلیل ISC طی حداکثر زمان پیگیری ۶٫۶ سال (میانگین ۲٫۸ سال)، از مطالعه خارج می‌شوند. خروج به دلایل دیگر در این مطالعه سانسور در نظر گرفته می‌شود. برای مقایسه AEDها پس از تنظیم نرخ تیتراسیون، ابتدا کالیبراسیون دوز با استاندارد کردن دوز هر دو دارو نسبت به نقطه میانی دامنه نگهداری آن انجام‌شده است. این دوزهای کالیبره شده به‌عنوان اندازه‌گیری‌های طولی در مدل توأم مخاطره‌های رقابتی در نظر گرفته می‌شود.

به‌طور متوسط ۴٫۶ اندازه‌گیری طولی از ۱ تا ۱۵ برای بیماران ثبت‌شده است.

مدل رده پنهان توأم را برای تحلیل این داده‌ها به کار می‌بریم. مدل طولی را به‌صورت زیر در نظر می‌گیریم:

$$y_i(s)_{|c_i=g} = \beta_{0g} + \beta_{1g}s + \beta_2 LTG_i + \beta_{3g}LTG_i s + b_{i0} + b_{i1}s + \epsilon_i(s); \quad i = 1,\dots,n; j = 1,\dots,n_i, \qquad (۲۵)$$

که در آن $(b_{i0}, b_{i1})' \sim N(0, \mathbf{D})$ (به عبارتی برای سادگی، $\sigma_g^2 = 1, g = 1,...,G$) و $\epsilon_i \sim N(0,\sigma^2)$. همچنین، LTG_i به‌عنوان اثر درمان در مدل به‌صورت دودویی تعریف می‌شود که مقدار ۱ را می‌گیرد اگر بیمار به‌طور تصادفی داروی LTG را دریافت کند و مقدار صفر را می‌گیرد اگر بیمار داروی CBZ را دریافت کند.

مدل مخاطره علت‌ـمشخص را به‌صورت زیر در نظر می‌گیریم:

$$h_{ikg}(t) = h_{0k}(t)exp\{\gamma_{0kg} + \gamma_{1k}LTG_i\}, \quad k = 1,2,\ g = 1,...,G \qquad (۲۶)$$

که در آن γ_{11} همان γ_{ISC} و γ_{12} برابر γ_{UAE} و $\gamma_{0kG} = 0$. توابع خطر پایه به‌صورت تکه‌ای با پنج گره و با پارامتر $\boldsymbol{\nu}_k$ در نظر گرفته می‌شوند. همچنین،

$$\Pr(c_i = g) = \frac{exp\{\lambda_g\}}{1 + \Sigma_{g=1}^{G-1} exp\{\lambda_g\}}; g = 1,\dots,G. \qquad (۲۷)$$

لازم به ذکر است که برای انتخاب تعداد رده‌ها، $G = 1,\dots,6$ در نظر گرفته‌شده است. ازاین‌رو، با مقایسه مقادیر BIC بهترین G را انتخاب می‌کنیم. مقادیر لگاریتم درستنمایی، BIC و همچنین احتمال‌های مربوط به هر رده برای مقادیر مختلف G در جدول ۴ ارائه‌شده است. با توجه به اینکه معیار انتخاب را BIC قرار داده‌ایم، از آوردن مقادیر AIC در جدول خودداری نموده‌ایم. مشاهده می‌شود که بهترین مقدار BIC برای $G = 5$ حاصل‌شده است. مقادیر برآورد پارامترها برای $G = 5$ نیز در جدول ۵ گزارش‌شده است.

[15] Standard and New Anti-epileptic Druge's

[16] Anti Epileptic Druge

جدول ۲: نتایج مدل رده پنهان توأم اندازه‌های طولی و داده‌های زمان بقا برای مجموعه داده ایدز با $g = ۳$

پارامتر	رده	برآورد	خطای استاندارد	آماره والد	پی‌ـمقدار
اثرهای ثابت مدل طولی					
$\beta_{۰۱}$	رده۱	۴٫۵۵۲	۰٫۲۱۰	۲٫۶۹۹	۰٫۰۰۰
$\beta_{۰۲}$	رده۲	۱۱٫۹۷۳	۰٫۶۱۲	۱۹٫۵۷۷	۰٫۰۰۰
$\beta_{۰۳}$	رده۳	۱۵٫۳۸۵	۰٫۵۰۴	۳۰٫۵۰۳	۰٫۰۰۰
$\beta_{۱۱}$	رده۱	−۰٫۱۳۵	۰٫۰۱۷	−۷٫۷۵۰	۰٫۰۰۰
$\beta_{۱۲}$	رده۲	−۰٫۳۳۴	۰٫۰۴۴	−۷٫۵۶۰	۰٫۰۰۰
$\beta_{۱۳}$	رده۳	۰٫۰۰۸	۰٫۰۳۲	۲٫۲۴۳	۰٫۸۰۸
$\beta_{۲۱}$	رده۱	۰٫۱۹۲	۰٫۲۷۲	۰٫۷۰۵	۰٫۴۸۱
$\beta_{۲۲}$	رده۲	−۰٫۳۶۴	۰٫۹۰۳	−۰٫۴۰۴	۰٫۶۸۶
$\beta_{۲۳}$	رده۳	−۱٫۰۳۴	۰٫۵۷۰	−۱٫۸۱۵	۰٫۰۶۹
پارامترهای مدل بقا					
$\nu_{۱۱}$	رده ۱	۰٫۱۷۵	۰٫۰۱۶	۱۰٫۶۱۲	۰٫۰۰۰
$\nu_{۲۱}$	رده۱	۰٫۱۷۸	۰٫۰۱۷	۱۰٫۶۲۸	۰٫۰۰۰
$\nu_{۳۱}$	رده۱	۰٫۲۰۶	۰٫۰۱۹	۱۰٫۵۲۴	۰٫۰۰۰
$\nu_{۴۱}$	رده۱	۰٫۲۸۴	۰٫۰۲۶	۱۰٫۱۰۶	۰٫۰۰۰
$\nu_{۵۱}$	رده۱	۰٫۲۲۰	۰٫۰۲۲	۱۰٫۱۰۸	۰٫۰۰۰
$\nu_{۱۲}$	رده ۲	۰٫۰۷۵	۰٫۰۴۳	۱٫۷۶۸	۰٫۰۷۷
$\nu_{۲۲}$	رده۲	۰٫۰۰۰	۰٫۰۷۷	۰٫۰۰۴	۰٫۹۹۷
$\nu_{۳۲}$	رده۲	۰٫۰۰۰	۰٫۰۶۵	۰٫۰۰۶	۰٫۹۹۵
$\nu_{۴۲}$	رده۲	۰٫۰۰۰	۰٫۰۹۴	۰٫۰۰۰	۱٫۰۰۰
$\nu_{۵۲}$	رده۲	۰٫۱۷۸	۰٫۰۵۳	۳٫۳۳۳	۰٫۰۰۱
$\nu_{۱۳}$	رده ۳	۰٫۰۱۴	۰٫۰۲۹	۰٫۴۶۶	۰٫۶۴۱
$\nu_{۲۳}$	رده۳	−۰٫۰۱۲	۰٫۰۲۶	−۰٫۴۵۹	۰٫۶۴۶
$\nu_{۳۳}$	رده۳	۰٫۰۲۲	۰٫۰۴۷	۰٫۴۷۷	۰٫۶۳۳
$\nu_{۴۳}$	رده۳	۰٫۰۰۰	۰٫۰۱۴	۰٫۰۰۰	۱٫۰۰۰
$\nu_{۵۳}$	رده۳	۰٫۰۰۰	۰٫۰۱۳	۰٫۰۰۰	۱٫۰۰۰
$\gamma_{۱}$	رده۱	۰٫۲۸۱	۰٫۱۵۹	۱٫۷۶۹	۰٫۰۷۷
$\gamma_{۲}$	رده۲	−۰٫۳۹۳	۰٫۹۵۱	−۰٫۴۱۳	۰٫۶۸۰
$\gamma_{۳}$	رده۳	۳٫۶۰۳	۴٫۲۰۶	۰٫۸۵۷	۰٫۳۹۲
اثرات ثابت مدل احتمال عضویت رده					
$\lambda_{۰۱}$	رده۱	۱٫۸۹۱	۰٫۲۳۶	۸٫۰۲۶	۰٫۰۰۰
$\lambda_{۰۲}$	رده۲	۰٫۳۷۸	۰٫۳۱۲	۱٫۲۱۰	۰٫۲۲۶
$\lambda_{۱۱}$	رده۱	−۰٫۶۱۷	۰٫۳۰۴	−۲٫۰۳۲	۰٫۰۴۲
$\lambda_{۱۲}$	رده۲	−۰٫۸۳۳	۰٫۴۳۴	−۱٫۹۲۱	۰٫۰۵۵

جدول ۳: ادامه نتایج مدل رده پنهان توأم برای $g = ۳$ بر اساس داده‌های ایدز

ماتریس واریانس_کواریانس اثرات تصادفی (مستقل از رده‌ها)	
پارامتر	برآورد
$D_{۱۱}$	۴٫۹۹۸
$D_{۱۲}$	−۰٫۲۴۰
$D_{۲۲}$	۰٫۰۱۷
خطای استاندارد مانده	
پارامتر	برآورد
σ	۱٫۷۳۴

خطای استاندارد
۰٫۰۴۷

جدول ۴: مقادیر لگاریتم درستنمایی ، معیار اطلاع و احتمال‌های عضویت هر رده، برای ۶ رده پنهان مدل توأم برازش داده‌شده بر اساس داده‌های صرع

تعداد رده(G)	loglik	BIC	% رده ۱	% رده ۲	% رده ۳	% رده ۴	% رده ۵	% رده ۶
۱	−۳۵۷۶٫۵۸	۷۲۶۸٫۴۵	۱۰۰	۰	۰	۰	۰	۰
۲	−۳۴۷۳٫۵۶	۷۱۰۰٫۸۶	۳۲٫۰۷	۶۷٫۹۳	۰	۰	۰	۰
۳	−۳۴۳۸٫۷۰	۷۰۶۹٫۵۵	۷۰٫۵۸	۱۰٫۰۸	۱۹٫۳۴	۰	۰	۰
۴	−۳۴۰۶٫۹	۷۰۴۴٫۳۹	۶٫۴۵	۷۴٫۷۱	۱۳٫۸۸	۴٫۹۶	۰	۰
۵	−۳۳۶۰٫۸۹	۶۹۹۰٫۸۰	۶٫۲۸	۱۵٫۲۱	۶۶٫۴۵	۷٫۱۱	۴٫۹۶	۰
۶	−۳۳۹۸٫۲	۷۱۰۳٫۸۵	۰	۷٫۲۷	۷۲٫۲۳	۱۶٫۳۶	۲٫۹۸	۱٫۱۶

نتایج جدول ۵ نشان می‌دهد که همه رده‌ها جزء رده ۱ برای مدل طولی دارای عرض از مبدأ هستند. اثر زمان فقط برای رده‌های ۲ و ۴ معنی‌دار است. اثر متقابل تیمار و زمان برای همه رده‌ها، جزء رده ۱ و ۳ معنی‌دار است. اثر مخاطره رقابتی اول فقط برای رده ۳ معنی‌دار نیست درصورتی‌که اثر مخاطره رقابتی دوم تنها در رده اول معنی‌دار است. همچنین، اثر درمان مخاطره دوم (γ_{UAE}) معنی‌دار است.

جدول ۵: نتایج مدل رده پنهان توأم اندازه‌های طولی و داده‌های مخاطره رقابتی برای $g = ۵$ بر اساس داده‌های صرع

پارامتر	رده	برآورد	خطای استاندارد	آماره والد	پی‌ـ مقدار
		اثرهای ثابت مدل طولی			
$\beta_{۰۱}$	رده۱	۰/۲۵۳	۰/۱۹۶	۱/۲۸۸	۰/۱۹۸
$\beta_{۰۲}$	رده۲	۱/۸۴۹	۰/۰۹۷	۱۹/۰۰۴	۰/۰۰۰
$\beta_{۰۳}$	رده۳	۱/۸۸۷	۰/۰۵۲	۳۶/۶۴۲	۰/۰۰۰
$\beta_{۰۴}$	رده۴	۲/۲۱۲	۰/۱۴۵	۱۵/۲۰۶	۰/۰۰۰
$\beta_{۰۵}$	رده۵	۳/۸۷۱	۰/۱۹۴	۱۹/۹۰۹	۰/۰۰۰
$\beta_{۲}$	–	−۰/۱۶۴	۰/۰۶۱	−۲/۷۰۲	۰/۰۰۷
$\beta_{۱۱}$	رده۱	−۰/۴۹۲	۱/۱۹۴	−۰/۴۱۲	۰/۶۸۰
$\beta_{۱۲}$	رده۲	۰/۵۱۶	۰/۰۹۱	۵/۷۰۱	۰/۰۰۰
$\beta_{۱۳}$	رده۳	۰/۰۳۸	۰/۰۲۵	۱/۴۸۳	۰/۱۳۸
$\beta_{۱۴}$	رده۴	۱/۶۷۳	۰/۲۲۴	۷/۴۷۱	۰/۰۰۰
$\beta_{۱۵}$	رده۵	۰/۱۰۶	۰/۱۰۲	۱/۰۳۷	۰/۳۰۰
$\beta_{۳۱}$	رده۱	۱/۲۳۷	۱/۱۲۲	۱/۱۰۳	۰/۲۷۰
$\beta_{۳۲}$	رده۲	۰/۴۳۲	۰/۱۰۴	۴/۱۳۹	۰/۰۰۰
$\beta_{۳۳}$	رده۳	۰/۰۶۵	۰/۰۳۴	۱/۹۰۴	۰/۰۵۷
$\beta_{۳۴}$	رده۴	۰/۷۶۱	۰/۲۷۰	۲/۸۱۴	۰/۰۰۵
$\beta_{۳۵}$	رده۵	۰/۲۶۵	۰/۱۲۳	۲/۱۵۳	۰/۰۳۱
		مدل مخاطره‌های رقابتی			
$\nu_{۱۱}$	$log(piecewise۱)$	−۳/۸۲۵	۰/۶۲۶	−۶/۱۱۲	۰/۰۰۰
$\nu_{۲۱}$	$log(piecewise۲)$	−۲/۹۸۳	۰/۵۹۶	−۵/۰۰۷	۰/۰۰۰
$\nu_{۳۱}$	$log(piecewise۳)$	−۲/۴۶۸	۰/۵۷۴	−۴/۳۰۱	۰/۰۰۰
$\nu_{۴۱}$	$log(piecewise۴)$	−۲/۳۳۳	۰/۵۷۵	−۴/۰۶۱	۰/۰۰۰
$\gamma_{۰۱۱}$	رده ۱	۳/۶۱۵	۰/۷۴۴	۴/۸۶۱	۰/۰۰۰
$\gamma_{۰۱۲}$	رده ۲	۱/۷۱۵	۰/۵۷۸	۲/۹۶۷	۰/۰۰۳
$\gamma_{۰۱۳}$	رده ۳	−۱/۲۴۵	۰/۷۰۸	−۱/۷۵۸	۰/۰۷۹
$\gamma_{۰۱۴}$	رده ۴	۳/۳۲۷	۰/۶۳۲	۵/۲۶۷	۰/۰۰۰
$\nu_{۱۲}$	$log(piecewise۱)$	−۲/۳۱۶	۰/۷۸۰	−۳/۶۱۴	۰/۰۰۰
$\nu_{۲۲}$	$log(piecewise۲)$	−۲/۹۲۳	۰/۷۷۹	−۳/۷۵۴	۰/۰۰۰
$\nu_{۳۲}$	$log(piecewise۳)$	−۳/۷۱۷	۰/۸۱۷	−۴/۵۵۱	۰/۰۰۰
$\nu_{۴۲}$	$log(piecewise۴)$	−۴/۷۷۹	۱/۰۱۸	−۴/۶۹۷	۰/۰۰۰
$\gamma_{۰۲۱}$	رده ۱	۴/۲۱۱	۰/۸۰۷	۵/۲۱۷	۰/۰۰۰
$\gamma_{۰۲۲}$	رده ۲	۰/۵۸۴	۱/۲۶۹	۰/۴۶۰	۰/۶۴۵
$\gamma_{۰۲۳}$	رده ۳	۰/۴۴۷	۰/۷۹۹	۰/۵۵۹	۰/۵۷۶
$\gamma_{۰۲۴}$	رده ۴	۰/۷۷۲	۱/۲۱۰	۰/۶۳۸	۰/۵۲۳
γ_{ISE}	–	−۰/۱۴۲	۰/۲۳۲	−۰/۶۰۹	۰/۵۴۲
γ_{UAE}	–	−۰/۶۹۳	۰/۲۴۸	−۲/۷۹۴	۰/۰۰۵
		پارامترهای مدل احتمال			
$\lambda_{۱}$	رده۱	۰/۱۸۱	۰/۳۱۶	۰/۵۷۳	۰/۵۶۶
$\lambda_{۲}$	رده۲	۱/۳۳۱	۰/۳۲۱	۴/۱۴۸	۰/۰۰۰
$\lambda_{۳}$	رده۳	۲/۴۱۰	۰/۲۷۷	۸/۶۹۸	۰/۰۰۰
$\lambda_{۴}$	رده۴	۰/۳۱۹	۰/۳۴۸	۰/۹۱۷	۰/۳۵۹

جدول ۶: ادامه نتایج مدل رده پنهان توأم برای $g = ۵$ بر اساس داده‌های صرع

ماتریس واریانس_کواریانس اثرات تصادفی (مستقل از رده‌ها)		
پارامتر	برآورد	
$D_{۱۱}$	۰/۲۸۶	
$D_{۱۲}$	۰/۰۲۷	
$D_{۲۲}$	۰/۰۲۸	
خطای استاندارد مانده		
پارامتر	برآورد	خطای استاندارد
σ	۰/۴۳۲	۰/۰۰۷

۴ بحث و نتیجه‌گیری

در سال‌های اخیر، مدل‌های رده پنهان برای تحلیل توأم اندازه‌های طولی و بقا مورداستفاده قرارگرفته است. این مدل‌ها، فرض می‌کنند که جمعیت در G رده پنهان تقسیم می‌شوند که با توابع مخاطره مختلف برای رخداد پیشامد و نمایش‌های متفاوت از نشانگر طولی توسط یک مدل آمیخته برای هر رده قابل توصیف است. در این مقاله، مطالعاتی که در دو دهه اخیر در این زمینه صورت گرفته است را مرور کردیم و یک نمای کلی از مدل‌های رده پنهان توأم ارائه دادیم. همچنین، کاربردهایی از مدل توأم رده پنهان را بر اساس دو مجموعه داده ایدز و داروهای ضد صرع بیان کردیم. ملاحظه کردیم که مدل‌های رده پنهان ابتدا بر اساس یک نشانگر طولی و یک زمان رخداد مطالعه شده‌اند، سپس به حالت‌های پیچیده‌تری تعمیم‌یافته‌اند، حالت‌هایی که ممکن است چندین نشانگر طولی یا متغیر طولی چندمتغیره شامل متغیر پیوسته، گسسته ترتیبی یا رسته‌ای و یا چند رخداد به مفهوم علت_مشخص برای مخاطره‌های رقابتی را در نظر بگیرند. بیش‌تر روش‌های استفاده‌شده در این زمینه مدل‌بندی، فرض استقلال شرطی بین فرایند طولی و زمان تا رخداد پیشامد برای هر رده را در نظر گرفته‌اند درصورتی‌که لیو و همکاران [۲۵]، از این فرض استفاده نکردند و مدل جدیدی را با ادغام مدل اثرات تصادفی توأم در چارچوب رده پنهان ارائه کردند. بیشترین روش استفاده‌شده در منابع، برای برآورد پارامترهای مدل رده پنهان توأم، روش درستنمایی ماکسیمم و به‌ویژه استفاده از الگوریتم EM بوده است، البته برخی هم روش استنباط بیزی را به‌کاربرده‌اند [۱۱، ۲]. لازم به ذکر است که بیشتر منابع، شامل کاربردهایی از آزمایش‌های بالینی است.

اگرچه در سال‌های اخیر، استفاده از مدل رده پنهان برای مدل‌بندی توأم داده‌های طولی و زمان تا رخداد پیشامد بیشتر موردتوجه قرارگرفته است، اما در نظر گرفتن متغیرهای ورودی بیشتری در زیر مدل بقا به‌منظور لحاظ کردن اطلاعات اضافی در خصوص بیماران و همچنین کاربردهای بیش‌تر در زمینه پژوهش‌های بالینی، معرفی پکیج‌های کاربردی جهت پردازش سریع، روش‌های انتخاب رده بهینه و استفاده از فرض‌های توزیعی دیگر برای اثرهای تصادفی به‌جای فرض توزیعی نرمال، می‌تواند موضوعات قابل‌بحث در پژوهش‌های آینده باشد.

مراجع

[1] Abrams, D. I., Goldman, A. I., Launer, C., Korvick, J. A., Neaton, J. D., Crane, L. R., and Terry Beirn Community Programs for Clinical Research on AIDS. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, **330**, 657-662.

[2] Andrinopoulou, E. R., Nasserinejad, K., Szczesniak, R., and Rizopoulos, D. (2020). Integrating latent classes in the Bayesian shared parameter joint model of longitudinal and survival outcomes. *Statistical methods in medical research*, **29**, 3294-3307.

[3] Baghfalaki, T., Ganjali, M. and Berridge, D. (2013). Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions: a Bayesian approach. *Biometrical Journal*, **55**, 844-865.

[4] Baghfalaki, T., Ganjali, M., and Hashemi, R. (2014). Bayesian joint modeling of longitudinal measurements and time-to-event data using robust distributions. *Journal of biopharmaceutical Statistics*, **24**, 834-855.

[5] Baghfalaki, T. and Ganjali, M. (2015). A Bayesian approach for joint modeling of skew-normal longitudinal measurements and time to event data. *REVSTAT–Statistical Journal*, **13**, 169-191.

[6] Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics*, **64**, 96-105.

[7] Chen, J., and Huang, Y. (2015). A Bayesian mixture of semiparametric mixed-effects joint models for skewed-longitudinal and time-to-event data. *Statistics in medicine*, **34**, 2820-2843.

[8] Diggle, P. J., Sousa, I. and Chetwynd, A. G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: the fourth Armitage lecture. *Statistics in Medicine*, **27**, 2981-2998.

[9] Entink K. R. H., Fox, J. P., and van den Hout, A. (2011). A mixture model for the joint analysis of latent developmental trajectories and survival. *Statistics in medicine*, **30**, 2310-2325.

[10] Faucett, C.L and Thomas, D. C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685.

[11] Garre, F. G., Zwinderman, A. H., Geskus, R. B., and Sijpkens, Y. W. (2008). A joint latent class changepoint model to improve the prediction of time to graft failure. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**, 299-308.

[12] Gould, A. L., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S. and Bois, F. Y. (2015) Joint modeling of survival and longitudinal non-survival data: current methods and issues: Report of the DIA Bayesian joint modeling working group. *Statist. Med.*, **34**, 2181–2195.

[13] Guler, I., Faes, C., Cadarso-Suárez, C., Teixeira, L., Rodrigues, A. and Mendonca, D. (2017). Two stage model for multivariate longitudinal and survival data with application to nephrology research. *Biometrical Journal*, **59**, 1204-1220.

[14] Han, J. , Slate, EH., and PeÃ´sa EA. (2007). Parametric latent class joint model for a longitudinal biomarker and recurrent events. *Statistics in Medicine*, **26**, 5285–5302.

[15] Henderson, R., Diggle, p. and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics*, **4**, 465–480.

[16] Hogan, J. and Laird, N. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239 – 258.

[17] Huang, Y., Dagne, G. and Wu, L. (2011). Bayesian inference on joint models of HIV dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, **30**, 2930–2946.

[18] Huang, Y., Lu, X., Chen, J., Liang, J., and Zangmeister, M. (2018). Joint model-based clustering of nonlinear longitudinal trajectories and associated time-to-event data analysis, linked by latent class membership: with application to AIDS clinical studies. *Lifetime data analysis*, **24**, 699-718.

[19] Ibrahim, J. G., Chen, M. and Sinha, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials, *Statistica Sinica*, **14**, 863–883.

[20] Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J. M., and Commenges, D. (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics*, **66**, 11-19.

[21] Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, **60**, 85-92.

[22] Li, M., Lee, C. W., and Kong, L. (2020). A latent class approach for joint modeling of a time-to-event outcome and multiple longitudinal biomarkers subject to limits of detection. *Statistical methods in medical research*, **29**, 1624-1638.

[23] Lin, H., McCulloch, C., and Rosenheck, R. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, **60**, 295 – 305.

[24] Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, **97**, 53–65.

[25] Liu, Y., Liu, L., and Zhou, J. (2015). Joint latent class model of survival and longitudinal data: An application to CPCRA study. *Computational Statistics and Data Analysis*, **91**, 40-50.

[26] Marson, A. G., Al-Kharusi, A. M., Alwaidh, M., Appleton, R., Baker, G. A., Chadwick, D. W., Cramp, C., Cockerell, O. C., Cooper, P. N., Doughty, J., Eaton, B., Gamble, C., Goulding, P. J., Howell, S. J. L., Hughes, A., Jackson, M., Jacoby, A., Kellett, M., Lawson, G. R., Leach, J. P., Nicolaides, P., Roberts, R., Shackley, P.,Shen, J., Smith, D. F., Smith, P. E. M., Tudur-Smith, C., Vanoli, A. and Williamson, P. R. (2007). The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. *Lancet*, **369**, 1000–1015.

[27] Proust-Lima, C., Dartigues, J.F., and Jacqmin-Gadda, H. (2016). Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: A latent process and latent class approach. *Statistics in Medicine*, **35**, 382–398.

[28] Proust, C., Jacqmin-Gadda, H., Taylor, J. M., Ganiayre, J., and Commenges, D. (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*, **62**, 1014-1024.

[29] Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics and Data Analysis*, **53**, 1142–1154.

[30] Proust-Lima, C., and Liquet, B. (2011). LCMM: an R package for estimation of latent class mixed models and joint latent class models. *In The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*, (p. 66)

[31] Proust-Lima, C., Philipps, V., Diakite, A., Liquet, B., Proust, M. C., and Lima, P. (2021). Package 'lcmm'.

[32] Proust-Lima, C., S´ene, M., Taylor, J. M. G., and Jacqmin- Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medicine Research*, **23**, 74–90.

[33] Proust-Lima, C., and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post treatment PSA: a joint modeling approach. *Biostatistics*, **10**, 535-549.

[34] Qin, L., Weissfeld, L. A., Shen, C., and Levine, M. D. (2009). A two-latent-class model for smoking cessation data with informative dropouts. *Communications in Statistics—Theory and Methods*, **38**, 2604-2619.

[35] Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*, Vol. 6. Boca Raton:Chapman and Hall.

[36] Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Computational Statistics and Data Analysis*, **56**, 491 – 501.

[37] Rouanet, A., Joly, P., Dartigues, J. F., Proust-Lima, C., and Jacqmin-Gadda, H. (2016). Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics*, **72**, 1123-1135.

[38] Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data : An overview. *Statistica Sinica*, **14**, 809–834.

[39] Wu, L., Liu, W., Yi, G. Y. and Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*.

[40] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-339.

[41] Ye, W., Lin, X. and Taylor, J. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data a two stage regression calibration approach. *Biometrics*, **64**, 1238-1246.

A view on latent class models for joint modeling of longitudinal measurements and survival data

Parvaneh Mehdizadeh[1], Taban Baghfalaki[2], Mahdy Esmailian[1]

Abstract:

Joint models are used in follow-up studies to investigate the relationship between longitudinal marker and a survival outcome and have been generalized to analyze multiple markers or competing risks data. Many statistical achievements in the field of joint modeling focus on shared random effects models which include characteristics of longitudinal markers as explanatory variables in the survival model. A less-known approach is the joint latent class model, assuming that a latent class structure fully captures the relationship between the longitudinal markers and the event risk. The latent class model may be appropriate because of the flexibility in modeling the relationship between the longitudinal marker and the time of event, as well as the ability to include explanatory variables, especially for predictive problems. In this paper, we provide an overview of the joint latent class model and its generalizations. In this regard, first a review of the discussed models is introduced and then the estimation of the model parameters is discussed. In the application section, two real data sets are analyzed.

Keywords: Competing risks, Latent class model, Longitudinal data, Maximum likelihood estimator, Survival model, The EM Algorithm.

[1]**Department of Statistics and Computer Sciences, Faculty of Sciences, University of Mohaghegh Ardabili**

[2]**Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University**