Journal of Statistical Sciences, Spring and Summer, 2025
Vol. 19, No. 1, pp 57-79
DOI: 10.61186/jss.19.1.7

Modeling Error Distribution in Accelerated Failure Time Models Using a Dirichlet Processes Mixture Model with the Burr XII Distribution as the Kernel

Haji Joudaki, B.[1], Khazaei, S.[2], Hashemi, R.[2]
[1]Department of Statistics, Lorestan University, Khorramabad, Iran.
[2]Department of Statistics, Razi University, Kermanshah, Iran.

Corresponding author: S. Khazaei, s.khazaei@razi.ac.ir

Received: 19/9/2024 **Revised:** 6/3/2025 **Accepted and Published Online:** 8/3/2025.

Introduction

One of the key challenges in survival data analysis is censoring, which occurs when data is only partially observed. This can introduce bias in estimates and reduce the accuracy of traditional statistical models. A common scenario in survival analysis is the availability of auxiliary variables in addition to potentially censored survival data. To address this challenge, various models can be employed, including the accelerated failure time model, the Cox proportional hazards model, and the proportional odds model. From a Bayesian perspective, Bayesian nonparametric models, such as those based on Dirichlet or Gaussian processes, offer powerful tools for analyzing censored data. These models provide several advantages, including high flexibility, the ability to manage uncertainty through probability distributions, and a decreased reliance on restrictive assumptions concerning the shape of the data distribution. In this context, the Dirichlet process is utilized as the prior distribution. In this paper, we explore a Dirichlet process mixture model utilizing a burr distribution type 12 kernel for modeling the survival distribution within an accelerated failure time framework. We compare the performance of the proposed model with that of a Pólya tree mixture model using both simulated and real datasets. The results indicate that the proposed model outperforms the alternative approach.

Material and Methods

The CODA package, version 4.4.1, of R software was used to evaluate the convergence of the aforementioned methods, the convergence of the Gibbs

sampling algorithms is confirmed by the results. To compare the performance of the proposed model with other Bayesian nonparametric mixture models, the spBayesSurv package in version 4.4.1 of R software, was utilized.

Results and Discussion
In this paper, a Dirichlet process-based mixture model with a kernel on Burr distribution type 12 is utilized to model the error term distribution in accelerated failure time models. The proposed model can flexibly estimate the error term distribution without requiring restrictive assumptions about a specific distribution shape by leveraging the properties of the Dirichlet process. This approach enables modeling hidden data structures and is particularly suitable for survival data analysis with varying levels of censoring. To evaluate the performance of the proposed model, both simulated and real datasets were analyzed. The simulated data section examined various scenarios, including censored and heterogeneous data, to assess the model's ability to handle common survival data challenges. Additionally, real datasets were employed to test the model's performance in practical settings.

Conclusion
The analysis results indicate that the proposed model can accurately estimate probability density and survival functions. The model has also successfully identified the data's latent structures and clustering patterns with high accuracy. Overall, this study's findings suggest that the proposed model, by leveraging the Dirichlet process's flexibility and modeling power, is a powerful tool for survival data analysis. This model exhibits strong potential in estimating density and survival functions and is an effective and reliable approach for analyzing survival data across various conditions.

Keywords: Mixture model, Accelerated failure time model, Dirichlet process, Burr Distribution type 12, Censored data, MCMC.

Mathematics Subject Classification (2010): 62F15, 62J05.

مجله علوم آماری، بهار و تابستان ۱۴۰۴
جلد ۱۹، شماره ۱، ص ۵۷ -- ۷۹
DOI: 10.61186/jss.19.1.7
مقاله پژوهشی

مدل‌بندی توزیع خطا در مدل‌های زمان شکست شتابیده با استفاده از مدل آمیخته‌ای از فرایندهای دیریکله با هسته بِر نوع ۱۲

بهرام حاجی جودکی[۱]، سلیمان خزائی[۲]، سیدرضا هاشمی[۲]

[۱]گروه آمار، دانشکده علوم پایه، دانشگاه لرستان
[۲]گروه آمار، دانشکده علوم، دانشگاه رازی

نویسنده مسئول : سلیمان خزائی، s.khazaei@razi.ac.ir

تاریخ دریافت: ۱۴۰۳/۶/۲۹ تاریخ بازنگری: ۱۴۰۳/۱۲/۱۶ تاریخ پذیرش و انتشار: ۱۴۰۳/۱۲/۱۸

چکیده: مدل‌های زمان شکست شتابیده یکی از مدل‌هایی است که در تحلیل بقا هنگامی که داده‌ها سانسور شده‌است، بویژه زمانی‌که همراه با متغیرهای کمکی است مورد استفاده قرار می‌گیرد. هنگامی‌که مدل‌های مورد نظر به پارامتر مجهول وابسته است یکی از روش‌هایی که می‌توان بکاربرد روش بیزی، بویژه روش بیز ناپارامتری است که فضای پارامتر را بینهایت بعدی در نظر می‌گیرد. در این چارچوب مدل‌های آمیخته فرایند دیریکله نقش مهمی را ایفا می‌کنند. در این مقاله یک مدل آمیخته فرایند دریکله با هسته بِر ۱۲ برای مدل‌سازی توزیع بقا زمان شکست شتابیده درنظر گرفته می‌شود. سپس با استفاده از روش‌های مونت کارلوی زنجیر مارکف از توزیع پسین نمونه تولید می‌شود. عملکرد مدل پیشنهادی با مدل آمیخته فرایند درخت پولیا براساس داده‌های شبیه‌سازی شده و واقعی مقایسه می‌شود. نتایج به‌دست آمده نشان می‌دهد که مدل پیشنهادی عملکرد بهتری دارد.

واژه‌های کلیدی: مدل آمیخته، مدل زمان شکست شتابیده، فرایند دیریکله، توزیع بِر ۱۲، داده‌های سانسور شده، زنجیره مارکف مونت کارلوی .
کد موضوع‌بندی ریاضی (۲۰۱۰): 62F15، 62J05.

۱ مقدمه

در تحلیل داده‌های بقا، هدف اصلی بررسی زمان وقوع یک رویداد خاص (مانند مرگ، خرابی دستگاه، یا عود بیماری) است. یکی از چالش‌های کلیدی در این زمینه، سانسور[1] است. سانسور به این معنا است که اطلاعات کامل در مورد زمان وقوع رویداد برای همه نمونه‌ها در دسترس نیست. انواع اصلی سانسور شامل، سانسور از راست (زمانی رخ می‌دهد که تا پایان مطالعه، رویداد موردنظر برای برخی از نمونه‌ها اتفاق نیفتاده باشد)، سانسور از چپ (این حالت زمانی رخ می‌دهد که رویداد موردنظر قبل از شروع مطالعه رخ داده است و تنها می‌دانیم که زمان وقوع رویداد کمتر از یک مقدار مشخص بوده است) و سانسور فاصله‌ای (این نوع سانسور زمانی اتفاق می‌افتد که زمان دقیق وقوع رویداد مشخص نیست، اما می‌دانیم که این رویداد در یک بازه زمانی خاص رخ داده است.) است (لاولس ، ۲۰۱۱؛ کالبفلایچ و پرنتیس ، ۲۰۱۱). سانسور یک چالش کلیدی است، زیرا عدم مشاهده کامل داده‌ها می‌تواند منجر به سوگیری در برآوردها و کاهش دقت مدل‌های آماری سنتی شود. به همین دلیل، روش‌هایی که بتوانند داده‌های سانسور شده را به‌درستی مدیریت کنند، اهمیت زیادی دارند. وضعیتی که در تحلیل بقا بسیار رخ می‌دهد زمانی است که علاوه بر داده‌های بقا احتمالاً سانسور شده، اطلاعات مجموعه‌ای از متغیرهای کمکی نیز در دسترس باشد. لذا ممکن است علاقمند باشیم که با وجود داده‌های سانسور شده تاثیر این متغیرهای کمکی را روی توزیع طول‌عمر بررسی کنیم. برای این منظور، می‌توان از مدل‌های مختلفی مانند مدل زمان شکست شتابیده[2] (AFT)، مدل مخاطره متناسب کاکس[3] (PH) و یا مدل نسبت متناسب[4] (PO) استفاده کرد. مدل AFT که یکی از مدل‌های پُرکاربرد در تحلیل داده‌های بقا است بر خلاف مدل‌هایی مانند PH و PO دارای پیش فرض خاصی نیست. بنابراین در وضعیت‌هایی که امکان استفاده از دو مدل اخیر نیست، می‌توان از آن استفاده کرد. مدل AFT به‌صورت

$$\log(T_i) = -\boldsymbol{x}_i\boldsymbol{\eta} + \epsilon_i, \quad i = ۱, \ldots, n, \tag{۱}$$

تعریف می‌شود، که در آن T_i و $\boldsymbol{x}_i$ به ترتیب زمان بقا و مقادیر بردار متغیرهای کمکی برای فرد i ام، $\boldsymbol{\eta}$ بردار ضرایب رگرسیونی و ϵ_i نیز خطای تصادفی مدل است. یک منبع بسیار عالی برای مطالعه تحقیقات انجام شده در مورد تحلیل داده‌های بقا با استفاده از مدل‌های AFT و دیگر مدل‌های پارامتری، نیمه‌پارامتری و ناپارامتری مقاله گومز و همکاران (۲۰۰۹) است. مدل AFT در (۱) را می‌توان به‌صورت

$$T_i = \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}V_i, \quad V_i = \exp\{\epsilon_i\}, \quad i = ۱, \ldots, n \tag{۲}$$

---

[1] Censoring
[2] Accelerated Failure Time
[3] Cox Proportional Hazard
[4] Proportional Odds

بازنویسی نمود. چون $V_i > ۰$ بنابراین، اگر بتوان برای V_i هر توزیع متعلق به خانواده توزیع‌های طول‌عمر را در نظر گرفت؛ آنگاه مدل AFT به یک مدل پارامتری تبدیل می‌شود. مدل‌های بیزی ناپارامتری، مانند مدل‌هایی مبتنی بر فرایند دیریکله یا فرایند گاوسی، ابزارهای قدرتمندی برای تحلیل داده‌های سانسورشده هستند. این مدل‌ها چند مزیت کلیدی دارند:

- انعطاف‌پذیری بالا در مدل‌سازی داده‌های پیچیده و غیرخطی.
- توانایی مدیریت عدم قطعیت به‌صورت صریح با استفاده از توزیع‌های احتمالی.
- نیاز کمتر به فرضیات محدود کننده در مورد شکل توزیع داده‌ها.

مدل‌های بیزی ناپارامتری مبتنی بر فرایند دیریکله، یکی از ابزارهای قدرتمند در آمار بیزی ناپارامتری هستند که به تحلیل داده‌ها با ویژگی‌های پیچیده و ساختار نامشخص کمک می‌کنند. در این مدل‌ها، از فرایند دیریکله (DP)[۱] به عنوان توزیع پیشین استفاده می‌شود. این فرایند یک توزیع احتمالی روی توزیع‌ها است. به‌عبارت دیگر، اگر یک متغیر تصادفی مانند G از یک فرایند دیریکله پیروی کند، G خود یک توزیع احتمال خواهد بود و به‌صورت $G \sim (\nu, G_۰)$ تعریف می‌شود که در آن $\nu > ۰$ پارامتر تمرکز است و میزان انحراف G از $G_۰$ را کنترل می‌کند. مقدار زیاد ν باعث می‌شود G به $G_۰$ (توزیع پایه، به‌عنوان میانگین دیریکله) نزدیک‌تر شود. یکی از رایج‌ترین کاربردهای فرایند دیریکله استفاده آن در مدل‌های آمیخته بیزی است. این مدل‌ها برای خوشه‌بندی و مدل‌سازی داده‌ها با تعداد نامعلوم خوشه‌ها استفاده می‌شوند و به‌صورت

$$
\begin{aligned}
&X_i \overset{ind.}{\sim} F(.|\theta_i), \quad i = ۱, \ldots, n, \\
&\theta_i | G \overset{i.i.d.}{\sim} G, \\
&G \sim DP(\nu, G_۰).
\end{aligned}
$$

تعریف می‌شود که در آن G توزیع پنهانی است که توسط فرایند دیریکله مدل می‌شود و θ_i پارامتر پنهان مربوط به داده iام است. یک راهکار دیگر در کار کردن با مدل‌های AFT بکارگیری دیدگاه روش‌های بیز ناپارامتری است. در دیدگاه بیز ناپارامتری بجای در نظر گرفتن توزیع مشخصی برای V_iها، فضای همه توزیع‌های طول‌عمر ممکن برای V_iها در نظر گرفته می‌شود. این فضا به عنوان یک پارامتر بینهایت بعدی رفتار می‌شود که هر مولفه آن یک تابع توزیع بقا است(بخش ناپارامتری). از دیدگاه بیزی ناپارامتری باید برای این فضای بینهایت بُعدی توزیع پیشینی تعریف نماییم. با توجه به نامتناهی بودن این فضای پارامتر نوعاً توزیع‌های پیشین چنین پارامترهایی از نوع فرایندهای تصادفی هستند. برای این منظور می‌توان از فرایند دیریکله[۲] (DP) استفاده کرد که در چارچوب بیز ناپارامتری از اهمیت خاصی برخوردار است. پس از معرفی فرایند DP بوسیله فرگوسن (۱۹۷۳) استفاده از این فرایند در تحلیل داده‌های بقا شروع شد. از جمله تحقیقات اولیه‌ای که در این زمینه انجام شد می‌توان به سوسارلا و ون رایزین (۱۹۷۶)،

[1] Dirichlet Process
[2] Dirichlet Process

فرگوسن و فادیا (۱۹۷۹) و لاهیری و پارک (۱۹۹۱) اشاره کرد. کریستینسن و جانسن (۱۹۸۸) و جانسن و کریستینسن (۱۹۸۹) برای مدل‌بندی توزیع بقا پایه در مدل‌های AFT از پیشین DP استفاده کردند. از آنجایی که DP یک فرایند گسسته است، بنابراین تقریب زدن توابع پیوسته با استفاده از آن خالی از اشکال نیست. آنتونیاک (۱۹۷۴) پیشنهاد می‌کند که برای رفع این مشکل از مدل‌های آمیخته فرایندهای دیریکله[1] (DPMM) استفاده شود. DPMM یک مدل بیزی ناپارامتری است که امکان یادگیری خودکار تعداد خوشه‌ها یا ساختار پنهان داده‌ها را فراهم می‌کند. این مدل برخلاف مدل‌های پارامتری که تعداد پارامترها یا ساختار ثابت دارند و یا انتخاب تعداد خوشه‌ها یا شکل توزیع‌ها معمولاً نیازمند آزمون و خطای زیادی است، تعداد خوشه‌ها را به‌طور خودکار از داده‌ها یاد می‌گیرد و مسئله را به‌صورت خودکار مدیریت می‌کند. در داده‌هایی که ساختار پیچیده دارند یا اطلاعات کافی برای تعیین مدل وجود ندارد (مانند داده‌های بقا با سانسور زیاد) عملکرد بهتری دارد و قادر است زمان‌های بقا و نرخ سانسور را بهتر مدل‌سازی کند و توزیع احتمال زمان بقا را با دقت بیشتری تخمین می‌زند. در مدل DPMM، به دلیل ساختار پیچیده و ناپارامتری بویژه ساختارهای سلسله مراتبی محاسبات تحلیلی دشوار است و روش‌های MCMC[2] به ما امکان نمونه‌گیری از توزیع پسین را می‌دهند. یکی از رایج‌ترین الگوریتم‌های MCMC، نمونه‌گیری گیبز [3] است که به‌صورت گسترده در DPMMها استفاده می‌شود (گوش و راماورتی ، ۲۰۰۳).

کاوا و مالیک (۱۹۹۷)، هنسن و جانسن (۲۰۰۴) و هنسن (۲۰۰۶) برای مدل‌بندی توزیع بقا پایه در مدل‌های AFT، DPMM با هسته‌های به ترتیب نرمال، لگ‌نرمال و گاما را برای تحلیل داده‌های سانسور شده بکار بردند. گوش و گسال (۲۰۰۶) برای مدل‌بندی تابع توزیع بقا، DPMM با هسته وایبل را بکار بردند. آن‌ها پارامتر مقیاس توزیع وایبل را به‌صورت $\beta_i = \mu_i \exp\{\boldsymbol{x}_i\boldsymbol{\eta}\}$ در نظر گرفتند و برای μ_i پیشین DP تعریف کردند. واکر و مالیک (۱۹۹۹)، هنسن و جانسن (۲۰۰۲)، هنسن (b۲۰۰۶) و هنسن (۲۰۱۴) برای مدل‌بندی توزیع جمله خطا در مدل‌های AFT، فرایند درخت پولیا[4] (PT) و یا آمیخته‌ای از فرایندهای درخت پولیا[5] (PTMM) با هسته‌های نرمال، لگ‌نرمال و وایبل را بکار بردند. بِر (۱۹۴۲) یک خانواده از توزیع‌های احتمال را معرفی نمود که به خانواده توزیع‌های بِر معروف شدند. توزیع بِر ۱۲ مهم‌ترین عضو خانواده توزیع‌های بِر است که به دلیل انعطاف‌پذیری بالایی که دارد اخیراً در حوزه‌هایی مانند تحلیل بقا، قابلیت اعتماد، آمار بیمه و غیره مورد توجه محققین قرارگرفته‌است. حاجی جودکی و همکاران (۲۰۲۲) برای تحلیل داده‌های بقای سانسور شده از راست، DPMM با هسته بِر ۱۲سه پارامتری را معرفی کردند و برخی از پتانسیل‌های مدل مذکور را برای برآورد توابع چگالی، بقا و نرخ‌خطر نشان دادند. در این مقاله هدف این است که DPMM با هسته بِر ۱۲سه پارامتری را برای مدل‌بندی تابع بقای پایه در مدل‌هایAFT بکار ببریم و برخی از پتانسیل‌های این مدل را برای مدل‌بندی داده‌های بقای چند مُدی نشان دهیم. بخش‌های بعدی این مقاله به‌صورت ذیل هستند. در بخش دوم ابتدا مدل را معرفی کرده و سپس در مورد انتخاب

[1] Dirichlet Process Mixture Model
[2] Markov Chain Monte Carlo
[3] Gibbs Sampling
[4] Polya Tree
[5] Polya Tree Mixture Model

توزیع پیشین پارامترها صحبت می‌شود. محاسبه توزیع پسین توام پارامترها از دیگر مباحث مطرح شده در بخش دوم مقاله است. در انتهای این بخش نیز جزئیات مربوط به اجرای الگوریتم‌های شبیه‌سازی MCMC این مدل‌ها ارائه می‌شود. بخش سوم شامل شبیه‌سازی لازم برای بررسی نتایج به‌دست آمده است. در بخش چهارم مدل پیشنهادی برای تحلیل مجموعه داده‌های واقعی بکار برده می‌شود. در بخش پایانی مقاله برخی نتایج به‌دست آمده بیان شده است و پیشنهاداتی برای تحقیقات آینده ارائه می‌شود.

۲ معرفی مدل

در چارچوب DPMM تابع چگالی متغیر V_i در (۲) به‌صورت

$$f(v_i;G)=\int k(v_i|\alpha,\beta,\lambda)dG(\alpha,\beta,\lambda) \qquad (۳)$$

مدل‌بندی می‌شود که در آن $k(v|\alpha,\beta,\lambda)$ تابع چگالی توزیع برِ ۱۲است. توزیع برِ ۱۲سه پارامتری دارای توابع چگالی، بقا و نرخ‌خطر[1] به‌صورت

$$k(t|\alpha,\beta,\lambda)=\alpha\lambda\beta^{-\alpha}t^{-\alpha-1}(1+(\frac{t}{\beta})^{-\alpha})^{-(\lambda+1)},\quad t>0,\alpha>0,\beta>0,\lambda>0$$
$$S(t|\alpha,\beta,\lambda)=(1+(\frac{t}{\beta})^{\alpha})^{-\lambda},\qquad t>0$$
$$h(t|\alpha,\beta,\lambda)=\alpha\lambda\beta^{-\alpha}t^{\alpha-1}(1+(\frac{t}{\beta})^{\alpha})^{-1},\quad t>0$$

است که در آن α و λ پارامترهای شکل و β پارامتر مقیاس است. از این پس DPMM با هسته برِ۱۲با نماد DPBM نشان داده می‌شود. فرض می‌شود $G(\alpha,\beta,\lambda)$ تابع توزیع پیشین بردار (α,β,λ) کمیتی تصادفی است و برای آن فرایند دیریکله به‌عنوان توزیع در نظر گرفته می‌شود و با نماد $G\sim DP(\nu,G_0)$ نشان داده می‌شود. در واقع G نقش وزن‌های مدل آمیخته را ایفا می‌کند. چون $E(G)=G_0$ است. معمولاً از G_0 به‌عنوان مرکز فرایند دیریکله یاد می‌شود. میزان نزدیکی G به G_0 به‌وسیله پارامتر ν کنترل می‌شود و از آن به‌عنوان پارامتر دقت یاد می‌شود (گوش و راماورتی ، ۲۰۰۳). با قرار دادن $\beta_i^*=\beta_i\exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}$ از روابط (۲) و (۳) داریم:

$$\begin{aligned} f(t_i;G,\boldsymbol{x}_i,\boldsymbol{\eta}) &= \exp\{\boldsymbol{x}_i\boldsymbol{\eta}\}\int k(t_i\exp\{\boldsymbol{x}_i\boldsymbol{\eta}\}|\alpha_i,\beta_i,\lambda_i)dG(\alpha,\beta,\lambda)\ , \\ &= \int k(t_i|\alpha_i,\beta_i^*,\lambda_i)dG(\alpha,\beta,\lambda),\quad i=1,\ldots,n, \end{aligned} \qquad (۴)$$

[1] Hazard Rate

۲.۱ انتخاب توزیع پیشین پارامترها

در روش‌های بیزی ناپارامتری، به‌ویژه در تحلیل داده‌های بقا، ابرپارامترها[1] نقش مهمی در تعیین ویژگی‌ها و رفتار مدل دارند. استفاده از توزیع‌های پیشین برای این ابرپارامترها یک ضرورت است که دلایل متعددی دارد. زیرا مقادیر ثابتی نیستند، بلکه به‌عنوان متغیرهای تصادفی در نظر گرفته می‌شوند که عدم قطعیت درباره آن‌ها باید مدل‌سازی شود. توزیع‌های پیشین روی ابرپارامترها، مدل را انعطاف‌پذیرتر و از سوگیری ناشی از انتخاب مقادیر ثابت جلوگیری می‌کنند. در تحلیل بقا، ممکن است توزیع زمان‌های بقا یا خوشه‌بندی زیربازه‌ها نامشخص باشد. با تعریف توزیع‌های پیشین بر روی ابرپارامترها، مدل می‌تواند به‌صورت پویا خود را با داده‌های مشاهده‌شده تطبیق دهد. با تخصیص توزیع‌های پیشین برای ابرپارامترها، مدل قادر است پیچیدگی خود را به‌صورت خودکار با توجه به حجم و ساختار داده‌ها تنظیم کند. این توزیع‌های پیشین باید به‌گونه‌ای انتخاب شوند که هم اطلاعات پیشین را بازتاب دهند و هم انعطاف کافی برای یادگیری از داده‌ها داشته باشند. بنابراین انتخاب توزیع پیشین مناسب برای ابرپارامترها بسیار مهم است. معمولاً برای پارامترهایی که مقادیر مثبت دارند و یا مقادیرشان بین ۰ و ۱ قرار دارد به ترتیب توزیع‌های گاما و بتا به عنوان توزیع پیشین انتخاب می‌شوند. بنابراین انتخاب توزیع نمایی به‌عنوان توزیع پیشین برای پارامترهای β، λ و ν بدلیل غیرمنفی بودن آن‌ها و مطابقت با فضای دامنه تغییرات توزیع نمایی، فاقد حافظه بودن، سادگی آن (در محاسبات بیزی سادگی و سازگاری خوبی دارد) است. توزیع یکنواخت یکی دیگر از انتخاب‌های رایج برای پیشین‌های بیزی است، به‌ویژه هنگامی که اطلاعات خاصی درباره پارامتر در دسترس نیست یا بخواهیم یک پیشین ناآگاهی‌بخش[2] داشته باشیم. انتخاب توزیع یکنواخت به‌عنوان توزیع پیشین برای α علاوه بر کنترل پذیری بازه مقادیر ممکن از در نظر گرفتن هرگونه فرض اضافی جلوگیری می‌کند. برای پارامتر ν مشابه اسکوبار و وست (۱۹۹۵) توزیع پیشین گاما، $\Gamma(a_\nu, b_\nu)$، در نظر گرفته می‌شود. فرض کنید $g_\circ$ تابع چگالی متناظر با تابع توزیع $G_\circ$ است. برای پارامترهای α، β و λ به ترتیب توزیع‌های پیشینی $\alpha|\phi \sim U(\circ, \phi)$، $\beta|\delta \sim Exp\{\frac{۱}{\delta}\}$ و $\lambda|\gamma \sim Exp\{\frac{۱}{\gamma}\}$ در نظر گرفته می‌شود که در آن نمادهای Exp و U به ترتیب برای نمایش توزیع‌های نمایی و یکنواخت بکار برده می‌شوند. با فرض استقلال این پیشین‌ها تابع چگالی پیشین توام پارامترها به‌صورت $g_\circ(\alpha, \beta, \lambda|\delta, \phi, \gamma) = \pi(\alpha|\phi)\pi(\beta|\delta)\pi(\lambda|\gamma)$ خواهد بود. برای ابرپارامترهای ϕ، δ و γ به ترتیب توزیع‌های پیشین مزدوجی به‌صورت $Par(a_\phi, b_\phi)$، $IG(a_\delta, b_\delta)$ و $IG(a_\gamma, b_\gamma)$ انتخاب شده‌است که در آن نمادهای IG و Par به‌ترتیب برای نشان دادن توزیع‌های گامای وارون و پاراتو به‌کار می‌روند. اگر $a_\phi = a_\gamma = a_\delta = ۲$، آنگاه $\mathrm{Var}(\delta) = \mathrm{Var}(\gamma) = \mathrm{Var}(\phi) = \infty$ و در نتیجه توزیع‌های پیشین مذکور ناآگاهی‌بخش خواهند بود. برای برآورد مقادیر b_ϕ، b_δ و b_γ از رویکرد بیز تجربی کوتاس (a۲۰۰۶) شده است. پیشین بردار ضرایب رگرسیونی به‌صورت $\pi(\boldsymbol{\eta}) = ۱$ انتخاب می‌شود که یک توزیع پیشین ناسرهٔ ناآگاهی‌بخش است.

[1] Hyperparameters
[2] Non-informative

۲.۲ توزیع پسین پارامترها

سانسور مورد نظر در این تحقیق سانسور فاصله‌ای است که تحت حالت‌های خاصی سانسورهای راست و چپ را شامل می‌شود. تحت این نوع سانسور مجموعه داده‌ها به‌صورت $(t_i, \boldsymbol{x}_i, r_i); i = ۱, \ldots, n$ هستند. t_i و $\boldsymbol{x}_i$ به ترتیب مقادیر طول عمر و بردار متغیرهای کمکی برای فرد iام هستند. r_i وضعیت سانسور فرد iام را مشخص می‌کند. اگر $r_i = ۱$ باشد طول عمر فرد iام مشاهده شده در زمان t_i و اگر $r_i = ۰$ باشد طول عمر فرد iام سانسور شده در بازه (L_i, U_i) است. با توجه به رابطه (۴) و توزیع‌های پیشین تعریف شده پارامترها در زیر بخش قبلی، DPMM با هسته بِر نوع ۱۲ را می‌توان در شکل سلسله مراتبی به‌صورت زیر نوشت:

$$
\begin{aligned}
&t_i|\alpha_i, \beta_i, \lambda_i, \boldsymbol{x}_i, \boldsymbol{\eta} \overset{i.d.}{\sim} \{k(t_i|\alpha_i, \beta_i^*, \lambda_i)\}^{r_i} \{S(L_i|\alpha_i, \beta_i^*, \lambda_i) - S(U_i|\alpha_i, \beta_i^*, \lambda_i)\}^{۱-r_i},\\
&(\alpha_i, \beta_i, \lambda_i)|G \overset{i.i.d.}{\sim} G, \quad i = ۱, \ldots, n,\\
&G|\phi, \delta, \gamma \sim DP(\nu, G_۰(\cdot|\phi, \delta, \gamma)), \qquad g_۰(\alpha, \beta, \lambda|\phi, \delta, \gamma) = \pi(\alpha|\phi)\pi(\beta|\delta)\pi(\lambda|\gamma),\\
&\alpha|\phi \sim U(۰, \phi), \qquad \beta|\delta \sim Exp\{\frac{۱}{\delta}\}, \qquad \lambda|\gamma \sim Exp\{\frac{۱}{\gamma}\}, \quad \phi \sim Par(a_\phi, b_\phi),\\
&\delta \sim IG(a_\delta, b_\delta), \quad \gamma \sim IG(a_\gamma, b_\gamma), \quad \nu \sim \Gamma(a_\nu, b_\nu), \quad \pi(\boldsymbol{\eta}) = ۱
\end{aligned}
$$

با توجه به ارائه سلسله مراتبی از مدل پیشنهادی واضح است که $(\alpha_i, \beta_i, \lambda_i)$ پارامتر مربوط به داده iام و $\{(\alpha_i, \beta_i, \lambda_i); i = ۱, \ldots, n\}$ نیز مجموعه پارامترهای مربوط به n داده است. اگر n^* تعداد بردارهای متمایز در مجموعه اخیر فرض شود آنگاه مجموعه بردارهای متمایز به‌صورت $\{(\alpha_j^*, \beta_j^*, \lambda_j^*); j = ۱, \ldots, n^*\}$ خواهد بود. در واقع n^* تعداد خوشه‌ها و $(\alpha_j^*, \beta_j^*, \lambda_j^*)$ پارامتر مربوط به خوشه jام است. با توجه به رابطه بین مقادیر $(\alpha_i, \beta_i, \lambda_i)$ و $(\alpha_j^*, \beta_j^*, \lambda_j^*)$ وضعیت خوشه‌بندی n داده مشخص می‌شود. فرض کنید بردار $\boldsymbol{w} = (w_۱, \ldots, w_n)$ وضعیت خوشه‌بندی n داده را نشان دهد. معمولاً از این بردار به‌عنوان بردار پیکربندی[1] یاد می‌شود. داده iام در خوشه jام قرار می‌گیرد، یعنی $w_i = j$ است اگر و تنها اگر رابطه $(\alpha_i, \beta_i, \lambda_i) = (\alpha_j^*, \beta_j^*, \lambda_j^*)$ برقرار باشد. برای $\boldsymbol{\alpha} = (\alpha_۱, \ldots, \alpha_n)$، $\boldsymbol{\beta} = (\beta_۱, \ldots, \beta_n)$ و $\boldsymbol{\lambda} = (\lambda_۱, \ldots, \lambda_n)$، فرض کنید $\boldsymbol{\alpha}_{-i}$، $\boldsymbol{\beta}_{-i}$ و $\boldsymbol{\lambda}_{-i}$ بردارهای $\boldsymbol{\alpha}$، $\boldsymbol{\beta}$ و $\boldsymbol{\lambda}$ بعد از حذف i امین مولفه آن‌ها هستند و n^{*-i} نیز تعداد بردارهای متمایز در مجموعه $\{\boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}\}$ است. توزیع پسین توام پارامترها، $[G, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \phi, \delta, \gamma, \nu|\boldsymbol{t}, \boldsymbol{x}]$، را می‌توان به‌صورت

$$
[G, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \phi, \delta, \gamma, \nu|\boldsymbol{t}, \boldsymbol{x}] \propto [G|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \phi, \delta, \gamma, \nu]\, [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \phi, \delta, \gamma, \nu|\boldsymbol{t}, \boldsymbol{x}]
$$

[1] Configuration vector

نوشت، که در آن

$$[G(.)|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\eta},\phi,\delta,\gamma,\nu] \sim DP(\nu+n,\frac{\nu}{\nu+n}G_{\circ}(.|\phi,\delta,\gamma)+\frac{\sum_{i=1}^{n}I_{(\alpha_i,\beta_i,\lambda_i)}(\cdot)}{\nu+n}) \quad (۵)$$

$$[\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\eta},\phi,\delta,\gamma,\nu|\boldsymbol{t},\boldsymbol{x}] \propto \prod_{i=1}^{n}\{\{k(t_i|\alpha_i,\beta_i^*,\lambda_i)\}^{r_i}\{S(L_i|\alpha_i,\beta_i^*,\lambda_i)-$$

$$S(U_i|\alpha_i,\beta_i^*,\lambda_i)\}^{1-r_i}\pi(\alpha_i|\phi)\pi(\beta_i|\delta)\pi(\lambda_i|\gamma)\}\pi(\phi)\pi(\delta)\pi(\gamma)\pi(\nu)\pi(\boldsymbol{\eta}) \quad (۶)$$

توزیع پسین توام پارامترها شکل پیچیده‌ای دارد و نمی‌توان به‌سادگی از آن توزیع‌های پسین کناری پارامترها را به‌دست آورد. بنابراین به ناچار باید با استفاده از روش‌های شبیه‌سازی MCMC از آن نمونه تولید کرد. براساس نمونه‌های تولید شده می‌توان توابع مورد علاقه مانند چگالی، بقا و نرخ‌خطر را برآورد کرد. الگوریتم نمونه‌گیری گیبز یکی از روش‌های MCMC است که می‌توان برای تولید نمونه از توزیع پسین توام پارا مترها بکار برد. در زیر بخش بعدی جزئیات مربوط به اجرای الگوریتم نمونه‌گیری گیبز داده شده‌است.

۲.۳ الگوریتم نمونه‌گیری گیبز

برای اجرای الگوریتم گیبز باید توزیع‌های شرطی کامل پارامترها را محاسبه کرد و سپس به‌صورت متوالی از آن‌ها نمونه تولید کرد. توزیع‌های شرطی کامل پارامترها به‌صورت

a. $[\alpha_i,\beta_i,\lambda_i|\boldsymbol{\alpha}_{-i},\boldsymbol{\beta}_{-i},\boldsymbol{\lambda}_{-i},\boldsymbol{\eta},t_i,\boldsymbol{x}_i,\nu,\phi,\delta,\gamma]\,, \quad i=1,\ldots,n,$

b. $\left[\alpha_j^*,\beta_j^*,\lambda_j^*|\boldsymbol{w},\boldsymbol{\eta},\nu,\phi,\delta,\gamma,\boldsymbol{t},\boldsymbol{x}\right], \quad j=1,\ldots,n^*,$

c. $[\boldsymbol{\eta}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{x}]\,,$

d. $[\nu|n^*]\,,[\phi|\boldsymbol{\alpha}^*,n^*]\,,[\delta|\boldsymbol{\beta}^*,n^*]\,,[\gamma|\boldsymbol{\lambda}^*,n^*]$

e. $[G|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\eta},\phi,\delta,\gamma,\nu]$

هستند، که در آن برای سادگی کار از نماد [ ] برای بیان توزیع‌های شرطی استفاده شده است. در ادامه فرایند تولید نمونه از شرطی‌های اخیر توضیح داده می‌شود. توزیع شرطی در (a) به‌صورت

$$[\alpha_i,\beta_i,\lambda_i|\boldsymbol{\alpha}_{-i},\boldsymbol{\beta}_{-i},\boldsymbol{\lambda}_{-i},\boldsymbol{\eta},\nu,\phi,\delta,\gamma,t_i,\boldsymbol{x}_i] \propto \{\nu g_{\circ}(\alpha_i,\beta_i,\lambda_i|\phi,\delta,\gamma)k(t_i|\alpha_i,$$

$$\beta_i\exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\},\lambda_i)\}+\sum_{j=1}^{n^{*-i}}n_j^{-i}k(t_i|\alpha_i,\beta_i\exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\},\lambda_i)I_{(\alpha_j^*,\beta_j^*,\lambda_j^*)}(\alpha_i,\beta_i,\lambda_i)$$

است، که در آن n_j^{-i} برابر با تعداد اعضای از مجموعه $\{\boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}\}$است که با مقدار بردار $(\alpha_j^*, \beta_j^*, \lambda_j^*)$ برابر هستند. بعد از جایگزینی مقادیر توابع چگالی k و توزیع پیشین $g_\circ$ در توزیع شرطی اخیر و بعد از مرتب‌سازی، توزیع شرطی $[\alpha_i, \beta_i, \lambda_i|\boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}, \boldsymbol{\eta}, \nu, \phi, \delta, \gamma, t_i, \boldsymbol{x}_i]$ را می‌توان به‌صورت

$$p_\circ h_\circ(\alpha_i, \beta_i, \lambda_i|\phi, \delta, \gamma, t_i, \boldsymbol{x}_i, \boldsymbol{\eta}) + \sum_{j=1}^{n^{*-i}} p_j I_{(\alpha_j^*, \beta_j^*, \lambda_j^*)}(\alpha_i, \beta_i, \lambda_i)$$

نوشت که در آن

$$\begin{aligned}
p_\circ &= \frac{q}{q + \sum_{j=1}^{n^{*-i}} n_j^{-i} k(t_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*)} \\
p_j &= \frac{n_j^{-i} k(t_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*)}{q + \sum_{j=1}^{n^{*-i}} n_j^{-i} k(t_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*)}, \quad j = 1, \ldots, n^{*-i} \\
q &= \frac{\nu}{\phi\delta\gamma t_i} \int_\circ^\phi \int_\circ^\infty \alpha(1 + (\frac{t_i}{\beta \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}})^{-\alpha})^{-1} \exp\{-\frac{\beta}{\delta}\} \\
&\quad \times (\frac{1}{\gamma} + \log(1 + (\frac{t_i}{\beta \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}})^\alpha))^{-2} d\alpha d\beta,
\end{aligned}$$

چون انتگرال موجود در q فرم بسته‌ای ندارد، باید با روش‌های عددی مقدار آن را تقریب زد. نمونه تولید شده از توزیع شرطی $[\alpha_i, \beta_i, \lambda_i|\boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}, \boldsymbol{\eta}, \nu, \phi, \delta, \gamma, t_i, \boldsymbol{x}_i]$ یا یک مقدار جدید است که از $h_\circ(\alpha_i, \beta_i, \lambda_i|\phi, \delta, \gamma, t_i, \boldsymbol{x}_i, \boldsymbol{\eta})$ تولید می‌شود و یا برابر با یکی از مقادیر بردار $(\alpha_j^*, \beta_j^*, \lambda_j^*)$، $j = 1, \ldots, n^{*-i}$ است. توزیع $h_\circ(\alpha_i, \beta_i, \lambda_i|\phi, \delta, \gamma, t_i, \boldsymbol{x}_i, \boldsymbol{\eta})$ متناسب با عبارت

$$g_\circ(\alpha, \beta, \lambda|\phi, \delta, \gamma) k(t_i|\alpha, \beta \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda)$$

است. بعد از جایگزینی مقادیر $g_\circ$ و k در عبارت اخیر $h_\circ$ را می‌توان به‌صورت

$$h_\circ(\alpha_i, \beta_i, \lambda_i|\phi, \delta, \gamma, t_i, \boldsymbol{x}_i, \boldsymbol{\eta}) \propto [\alpha_i|\phi, t_i][\beta_i|\alpha_i, \delta, t_i, \boldsymbol{x}_i, \boldsymbol{\eta}][\lambda_i|\alpha_i, \beta_i, \gamma, t_i, \boldsymbol{x}_i, \boldsymbol{\eta}],$$

نوشت، که در آن

$$[\alpha_i|\phi, t_i] \propto \alpha_i t_i^{\alpha_i} I_{(\circ,\phi)}(\alpha_i),$$

$$[\beta_i|\alpha_i, \delta, t_i, \boldsymbol{x}_i, \boldsymbol{\eta}] \propto \frac{1}{\delta} \exp\{-\frac{\beta_i}{\delta}\}(t_i^{\alpha_i} + \beta_i^{*\alpha_i})^{-1}, \beta_i^* = \beta_i \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\},$$

$$[\lambda_i|\alpha_i,\beta_i,\gamma,t_i,\boldsymbol{x}_i,\boldsymbol{\eta}] \sim \Gamma(۲,\frac{۱}{\gamma}+\log(۱+(\frac{t_i}{\beta_i^*})^{\alpha_i})).$$

برای تولید نمونه از $h_۰$، ابتدا با استفاده از الگوریتم نمونه‌گیری برشی دی‌می‌ان و همکاران (۱۹۹۹) از توزیع‌های شرطی $[\beta_i|\alpha_i,\delta,t_i,\boldsymbol{x}_i,\boldsymbol{\eta}]$ و $[\alpha_i|\phi,t_i]$ نمونه تولید می‌شود. سپس با توجه به اینکه توزیع شرطی $[\lambda_i|\alpha_i,\beta_i,\gamma,t_i,\boldsymbol{x}_i,\boldsymbol{\eta}]$ یک توزیع گاما است و می‌توان از آن نمونه تولید نمود. در حالتی که مشاهده iام سانسور شده در بازه (L_i,U_i) باشد آنگاه توزیع شرطی در (a) را می‌توان به‌صورت مدل آمیخته

$$p_۰ h_c(\alpha_i,\beta_i,\lambda_i|\phi,\delta,\gamma,L_i,U_i,\boldsymbol{x}_i,\boldsymbol{\eta})+\sum_{j=۱}^{n^{*-i}} p_j I_{(\alpha_j^*,\beta_j^*,\lambda_j^*)}(\alpha_i,\beta_i,\lambda_i)$$

نوشت، که در آن

$$p_۰ = \frac{q}{q+\sum_{j=۱}^{n^{*-i}} n_j^{-i}\{S(L_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*)-S(U_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*)\}}$$

$$p_j = \frac{n_j^{-i}\{S(L_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*,\boldsymbol{\eta})-S(U_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*)\}}{q+\sum_{j=۱}^{n^{*-i}} n_j^{-i}\{S(L_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*)-S(U_i|\alpha_j^*,\beta_j^*\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_j^*)\}}$$

و $q=\nu\{K(L_i,\phi,\delta,\gamma,\mathbf{x}_i,\boldsymbol{\eta})-K(U_i,\phi,\delta,\gamma,\mathbf{x}_i,\boldsymbol{\eta})\}$

$$K(Z,\phi,\delta,\gamma,\mathbf{x}_i,\boldsymbol{\eta})=\int_۰^{\phi}\int_۰^{\infty}\alpha\exp\{-\frac{\beta}{\delta}\}(\frac{۱}{\gamma}+\log(۱+(\frac{Z}{\beta\exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}})^{\alpha}))^{-۱}d\alpha d\beta$$

انتگرال اخیر فرم بسته‌ای ندارد، بنابراین باید با روش‌های عددی مقدار آن را تقریب زد. همچنین $h_c(\alpha_i,\beta_i,\lambda_i|\phi,\delta,\gamma,L_i,U_i,\boldsymbol{x}_i,\boldsymbol{\eta})$ متناسب با عبارت

$$\{S(L_i|\alpha_i,\beta_i\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_i)-S(U_i|\alpha_i,\beta_i\exp\{-\mathbf{x}_i\boldsymbol{\eta}\},\lambda_i)\}\times g_۰(\alpha_i,\beta_i,\lambda_i|\phi,\delta,\gamma)$$

است. بعد از جایگزینی مقادیر $S(L_i|.)$، $S(U_i|.)$ و $g_۰$ در h_c، h_c شکل پیچیده‌ای خواهد داشت و برای تولید نمونه از آن باید از الگوریتم‌های مانند متروپولیس–هستینگز استفاده کرد. برای بهبود آمیختگی زنجیره‌های مارکف، دومین گام (b) از الگوریتم گیبز اجرا می‌شود. توزیع شرطی در (b) یعنی $\left[\alpha_j^*,\beta_j^*,\lambda_j^*|\boldsymbol{W},\boldsymbol{\eta},\nu,\phi,\delta,\gamma,\boldsymbol{t},\boldsymbol{x}\right]$

برای هر مقدار $j = ۱, \ldots, n^*$ متناسب با عبارت

$$\prod_{i:W_i=j} \{\{S(L_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*) - S(U_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*)\}^{۱-r_i}\}$$
$$\{k(t_i|\alpha_j^*, \beta_j^* \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_j^*)\}^{r_i} \times g_\circ(\alpha_j^*, \beta_j^*, \lambda_j^*|\phi, \delta, \gamma)$$

است. بعد از جایگزینی مقادیر توابع k، S و $g_\circ$ در عبارت اخیر، توزیع پسین شرطی $(\alpha_j^*, \beta_j^*, \lambda_j^*)$ شکل پیچیده‌ای خواهد داشت و بنابراین باید با استفاده از الگوریتم‌های مانند متروپولیس–هستینگز یا برشی از آن نمونه تولید کرد. در سومین گام الگوریتم گیبز، باید از توزیع شرطی $\pi[\boldsymbol{\eta}|\alpha, \beta, \lambda, \phi, \delta, \gamma, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{r}]$ نمونه تولید کرد. این توزیع شرطی متناسب با عبارت

$$\prod_{i=۱}^{n} \{\{S(L_i|\alpha_i, \beta_i \exp\{-\mathbf{x}_i\boldsymbol{\eta}\}, \lambda_i) - S(U_i|\alpha_i, \beta_i \exp\{-\mathbf{x}_i\boldsymbol{\eta}\}, \lambda_i)\}^{۱-r_i}$$
$$\{k(t_i|\alpha_i, \beta_i \exp\{-\boldsymbol{x}_i\boldsymbol{\eta}\}, \lambda_i)\}^{r_i}\}\pi(\boldsymbol{\eta})$$

است، که در آن $\pi(\boldsymbol{\eta}) = ۱$. توزیع شرطی اخیر شکل بسته‌ای ندارد. بنابراین با استفاده از الگوریتم متروپولیس–هستینگز گام تصادفی از آن نمونه تولید می‌شود. برای بروزرسانی پارامتر ν در الگوریتم گیبز از الگوریتم ارائه شده در اسکوبار و وست (۱۹۹۵) استفاده می‌شود. توزیع‌های شرطی ابرپارامترهای ϕ، δ و γ شکل بسته دارند. بعد از انجام برخی عملیات جبری این توزیع‌ها به‌صورت

$$[\phi|\alpha^*, n^*] \sim Pareto(a_\phi + n^*, \max(b_\phi, \max_{۱\le j\le n^*} \alpha_j^*)),$$
$$[\delta|\beta^*, n^*] \sim IG(shape = a_\delta + n^*, scale = b_\delta + \sum_{j=۱}^{n^*} \beta_j^*),$$
$$[\gamma|\lambda^*, n^*] \sim IG(shape = a_\gamma + n^*, scale = b_\gamma + \sum_{j=۱}^{n^*} \lambda_j^*),$$

حاصل می‌شوند، که در آن $\boldsymbol{\alpha}^*$، $\boldsymbol{\beta}^*$ و $\boldsymbol{\lambda}^*$ به ترتیب به‌صورت $\boldsymbol{\alpha}^* = \{\alpha_j^*; j = ۱, \ldots, n^*\}$، $\boldsymbol{\beta}^* = \{\beta_j^*; j = ۱, \ldots, n^*\}$ و $\boldsymbol{\lambda}^* = \{\lambda_j^*; j = ۱, \ldots, n^*\}$ هستند. توزیع شرطی (e) در رابطه (۵) داده شده است. برای تولید نمونه از این توزیع شرطی ارائه ستورامان (۱۹۹۴) از DP بکار برده می‌شود. ارائه مذکور به صورت $G(\alpha, \beta, \lambda)|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu \approx \sum_{j=۱}^{J} \omega_j I_{(\alpha_j, \beta_j, \lambda_j)}(\alpha, \beta, \lambda)$ است. برای تعیین J می‌توان از روش ارائه شده در کوتاس (a۲۰۰۶) استفاده کرد یا مشابه چنگ و یوان (۲۰۱۳) مقدار ثابتی مانند ۴۰۰۰ برای آن در نظر

گرفت. فرایند تولید مقادیر ω_jها و $(\alpha_j, \beta_j, \lambda_j)$ ها در الگوریتم زیر داده می‌شود.

$$I) \quad (\alpha_j, \beta_j, \lambda_j) \overset{i.i.d.}{\sim} \frac{\nu g_0(\alpha_j, \beta_j, \lambda_j|\phi, \delta, \gamma) + \sum_{i=1}^{n^*} I_{(\alpha_i^*, \beta_i^*, \lambda_i^*)}(\alpha_j, \beta_j, \lambda_j)}{\nu + n},$$

$$II) \quad z_j \overset{i.i.d.}{\sim} beta(1, \nu + n), \quad j = 1, \ldots, J-1,$$

$$III) \quad \omega_1 = z_1, \omega_j = z_j \prod_{k=1}^{j-1}(1 - z_k), j = 2, \ldots, J-1, \omega_J = 1 - \sum_{j=1}^{J-1} \omega_j.$$

در تکرار bام الگوریتم گیبز برمبنای مقادیر تولید شده

$$\{\boldsymbol{\theta}_j^{(b)} = (\alpha_j^{(b)}, \beta_j^{(b)} e^{\boldsymbol{x}_0' \boldsymbol{R}^{(b)}}, \lambda_j^{(b)}); j = 1, \ldots, J\}$$

و مقادیر وزن‌ها $\omega_j^{(b)}; j = 1, \ldots, J$، توابع چگالی و بقا در تکرار bام الگوریتم گیبز در نقطه t به‌صورت $f^{(b)}(t; \boldsymbol{x}_0) \approx \sum_{j=1}^{J} \omega_j^{(b)} k(t|\boldsymbol{\theta}_j^{(b)})$ و $S^{(b)}(t; \boldsymbol{x}_0) \approx \sum_{j=1}^{J} \omega_j^{(b)} S(t|\boldsymbol{\theta}_j^{(b)})$ تقریب زده می‌شوند که در آن $\boldsymbol{x}_0$ مقدار معلوم و مشخصی از بردار متغیرهای کمکی است.

۳ مطالعه شبیه‌سازی

در این مطالعه شبیه‌سازی فرض می‌شود که در مدل AFT ارائه شده در رابطه (۲)، دو متغیر کمکی x_1 و x_2 وجود دارد. مقادیر این متغیرها به ترتیب از توزیع‌های برنولی با پارامتر ۰/۵ و نرمال استاندارد تولید می‌شود. همچنین مقادیر متغیر V در مدل (۲) از توزیع آمیخته

$$f(v) = 0.5N(v|\mu = 1, \sigma = 0/15) + 0.5N(v|\mu = 3, \sigma = 0/15) \qquad (۷)$$

تولید می‌شود، که در آن نماد $N(v|\mu, \sigma)$ بیانگر توزیع نرمال با میانگین μ و انحراف معیار σ است. با توجه به رابطه $T = \exp\{-\eta_1 x_1 - \eta_2 x_2\}V$ یک نمونه از Tها به‌دست می‌آید. مقادیر T در فاصله دو عدد صحیح متوالی، سانسور فاصله‌ای و مقادیر $T > 10$ سانسور از راست می‌شوند، یعنی ۱۰۰٪ داده‌ها سانسور شده‌اند. حجم نمونه $n = 100$ و $n = 500$ در نظر گرفته شده‌است. مقادیر واقعی ضرایب رگرسیونی نیز به‌صورت $\eta_1 = 1$ و $\eta_2 = -1$ انتخاب شده‌اند. با توجه به رابطه (۷) توابع چگالی و بقای T به ترتیب به‌صورت ذیل خواهند بود، $f(t) = 0.5N(t|\mu_1, \sigma_1) + 0.5N(t|\mu_2, \sigma_1)$ و $S(t) = 0.5\Phi(\frac{\mu_1 - t}{\sigma_1}) + 0.5\Phi(\frac{\mu_2 - t}{\sigma_1})$ که در آن $cons = \exp\{\eta_1 x_1 + \eta_2 x_2\}$، $\mu_1 = \frac{1}{cons}$، $\mu_2 = \frac{3}{cons}$ و $\sigma_1 = \frac{0/15}{cons}$. توزیع پیشین پارامتر ν به‌صورت $\Gamma(1, 0/001)$ انتخاب می‌شود. در فرایند اجرای الگوریتم گیبز، تعداد کل تکرارهای الگوریتم گیبز، تعداد

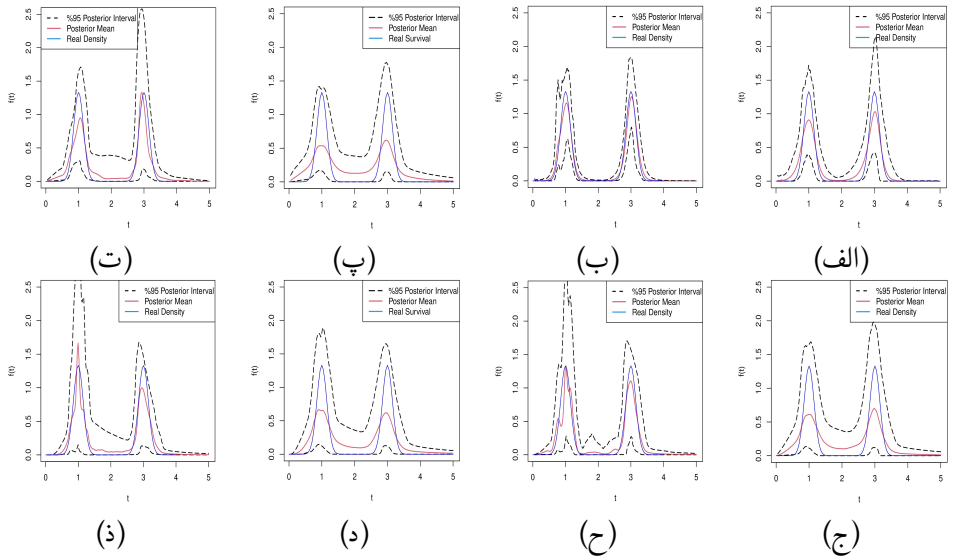
تکرارهای دوره داغیدن و فاصله تاخیر به ترتیب برابر با مقادیر $N = ۱۵۰۰۰$، $B = ۵۰۰۰$ و ۲۰ تعیین شده‌است و بنابراین حجم نمونه نهایی در استنباط ۵۰۰ است. برای بررسی همگرایی الگوریتم‌های نمونه‌گیری گیبز از روش‌های گِوک، رفتری–لوییس و هیدلبرگ–ولچ استفاده شده است (گلمن و همکاران ، ۲۰۱۳). برای دستیابی به نتایج روش‌های مذکور، بسته CODA در نسخه ۴.۰.۱ نرم‌افزار R مورد استفاده قرار گرفته‌است. نتایج به‌دست آمده همگرایی الگوریتم‌های نمونه‌گیری گیبز را تایید می‌کنند. برای اینکه بتوان عملکرد مدل پیشنهادی را با عملکرد دیگر مدل‌های آمیخته بیز ناپارامتری مقایسه کرد بسته نرم‌افزاری spBayesSurv در نسخه ۴.۰.۱ نرم‌افزار R بکار گرفته می‌شود(بولته و همکاران ، ۲۰۲۱). با استفاده از این بسته PTMM با هسته‌های وایبل، لگ‌نرمال و لگ‌لوژستیک نیز به مجموعه داده‌های شبیه‌سازی شده و واقعی برازش داده‌می‌شود. در ادامه PTMM با هسته‌های وایبل، لگ‌نرمال و لگ‌لوژستیک به ترتیب با نمادهای PTWM، PTLNM و PTLGM نشان داده می‌شود (هنسن و جانسن ، ۲۰۰۲؛ هنسن ، ۲۰۰۶b؛ هنسن و یانگ ، ۲۰۰۷؛ هنسن ، ۲۰۱۴). در مجموعه داده‌های شبیه‌سازی شده برای اینکه بتوان نتایج مربوط به برآورد توابع چگالی و بقا را با مقادیر واقعی این توابع مقایسه کرد از شاخص‌های اقلیدسی (d_E)، میانگین قدر مطلق خطا[1] (d_{MAE}) و هلینگر[2] (d_H) استفاده شده‌است. در جدول ۱ مقادیر شاخص‌های مذکور برای مدل‌های DPBM، PTWM، PTLNM و PTLGM و تحت دو اندازه نمونه ۱۰۰ و ۵۰۰ داده شده‌است.

جدول ۱. شاخص‌های انحراف در برآورد توابع چگالی و بقای مدل AFT آمیخته نرمال با استفاده از مدل‌های آمیخته بیز ناپارامتری گوناگون.

		$n = ۱۰۰$			$n = ۵۰۰$		
تابع	مدل	d_E	d_{MAE}	d_H	d_E	d_{MAE}	d_H
چگالی	DPBM	۱٫۶۱۴۱	۰٫۰۶۷۵	۱٫۲۸۴۵	۰٫۸۱۵۵	۰٫۰۳۱۴	۰٫۶۳۵۲
	PTWM	۳٫۴۰۸۳	۰٫۱۶۲۷	۲٫۶۹۶۰	۱٫۸۱۹۱	۰٫۰۷۸۷	۱٫۶۱۹۷
	PTLNM	۳٫۰۸۴۸	۰٫۱۴۱۹	۲٫۳۹۱۲	۱٫۶۰۲۴	۰٫۰۶۹۲	۱٫۴۱۷۹
	PTLOM	۳٫۰۳۲۵	۰٫۱۴۰۸	۲٫۴۰۳۵	۱٫۵۸۹۵	۰٫۰۶۱۹	۱٫۳۰۲۹
بقا	DPBM	۰٫۳۰۸۴	۰٫۰۱۷۲	۰٫۶۰۴۸	۰٫۱۳۶۲	۰٫۰۰۵۵	۰٫۲۳۵۴
	PTWM	۰٫۸۲۶۹	۰٫۰۴۸۲	۱٫۱۳۸۷	۰٫۳۴۸۶	۰٫۰۱۸۹	۰٫۵۵۶۳
	PTLNM	۰٫۷۵۷۶	۰٫۰۴۵۱	۱٫۲۶۸۷	۰٫۳۲۰۲	۰٫۰۱۷۰	۰٫۵۷۹۳
	PTLOM	۰٫۷۶۹۹	۰٫۰۴۶۶	۱٫۲۷۰۶	۰٫۲۷۳۹	۰٫۰۱۵۹	۰٫۷۰۵۱

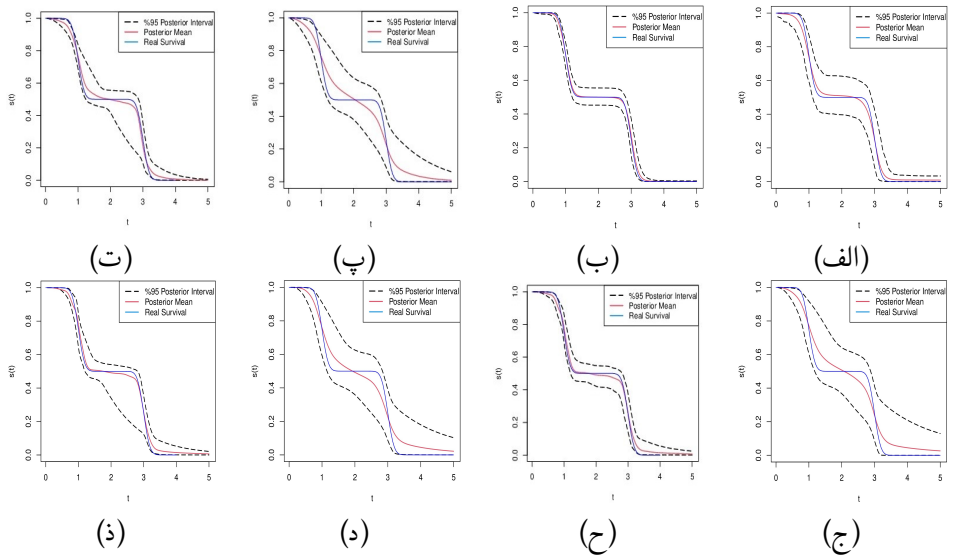
[1] Mean Absolute Error
[2] Hellinger

با توجه به جدول ۱ ملاحظه می‌شود که مقادیر شاخص‌ها برای مدل DPBM در مقایسه با دیگر مدل‌ها مقادیر خیلی کوچکتری هستند و بنابراین می‌توان نتیجه‌گیری کرد که مدل DPBM منحنی‌های واقعی (چگالی و بقا) را نسبت به مدل‌های دیگر بهتر برآورد کرده‌است. همچنین در کل با افزایش اندازه نمونه مقادیر شاخص‌ها نیز کاهش‌یافته است. نتایج نموداری مربوط به مدل‌های DPBM، PTWM، PTLNM و PTLGM در برآورد منحنی‌های واقعی $f(t)$ و $S(t)$ به ترتیب در شکل‌های ۱ و ۲ نشان داده شده است. در هر یک از شکل‌های مذکور نتایج مربوط به دو اندازه نمونه ۱۰۰ و ۵۰۰ به ترتیب در نمودارهای سمت راست و چپ نشان داده شده‌اند. در این نمودارها میانگین‌های پسینی[1] و کران‌های فاصله اعتبار ۹۵٪ بیزی[2] منحنی‌های واقعی در نقاط مختلف به ترتیب با رنگ‌های قرمز و مشکی بریده شده نشان داده شده‌اند. همچنین در این نمودارها منحنی‌های واقعی با رنگ آبی نمایش داده می‌شود. با مقایسه نمودارهای ارائه شده در دو سمت چپ و راست شکل‌های ۱ و ۲ با یکدیگر مشاهده می‌شود که با افزایش اندازه نمونه مقادیر برآوردها(میانگین‌های پسینی) به مقادیر واقعی منحنی‌های واقعی (چگالی و بقا) نزدیک‌تر شده‌اند و همچنین طول فواصل اعتبار ۹۵٪ بیزی منحنی‌های واقعی نیز کاهش‌یافته است. در کل نتایج ارائه شده بوسیله مدل DPBM نسبت به دیگر مدل‌ها خیلی بهتر هستند.



شکل ۱. منحنی برآوردهای نقطه‌ای (قرمز) و فاصله‌ای (تیره) از تابع چگالی واقعی (آبی) با مدل‌های DPBM (الف- $n = ۱۰۰$ و ب- $n = ۵۰۰$)، PTWM(پ- $n = ۱۰۰$ و ت- $n = ۵۰۰$)، PTLGM(ج- $n = ۱۰۰$ و ح- $n = ۵۰۰$) و PTLNM(د- $n = ۱۰۰$ و ذ- $n = ۵۰۰$).

[1]Posterior Mean
[2]Bayesian Credible Interval

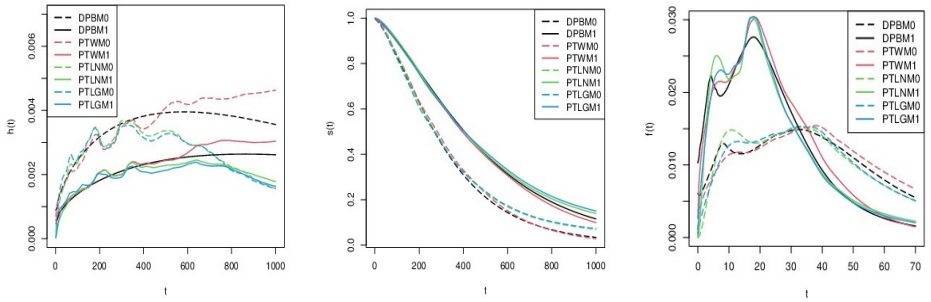


شکل ۲. منحنی برآوردهای نقطه‌ای (قرمز) و فاصله‌ای(تیره) از تابع بقای واقعی (آبی) با مدل‌های DPBM(الف- $n = ۱۰۰$ و ب- $n = ۵۰۰$)، PTWM(پ- $n = ۱۰۰$ و ت- $n = ۵۰۰$)، PTLGM(ج- $n = ۱۰۰$ و ح- $n = ۵۰۰$) و PTLNM(د- $n = ۱۰۰$ و ذ- $n = ۵۰۰$).

۴ تحلیل مجموعه داده واقعی

در این بخش، یک مجموعه داده واقعی تحلیل می‌شود. این مجموعه داده تحت عنوان lung از مجموعه داده‌های بسته survival در نسخه 1.4.4 نرم‌افزار R است که مربوط به یک مجموعه از بیماران مبتلا به سرطان پیشرفته ریه است. در این مجموعه داده اطلاعاتی مانند زمان بقا، سن، جنسیت، وضعیت سانسور، میزان کالری مصرف شده، وزن از دست رفته بیمار در ۶ ماه گذشته و غیره برای بیماران حاضر در تحقیق داده شده‌است. حدود ۲۸٪ داده‌های بقا سانسورشده از راست هستند. در این مثال هدف این است که تاثیر متغیرهای کمکی سن و جنسیت روی توزیع طول عمر بیماران بررسی شود. در واقع هدف مقایسه مقادیر بقا و نرخ خطر بیماران مرد و زن در یک سن مشخص است. این سن مشخص میانگین سن کل بیماران در نظر گرفته شده‌است. میانگین سن بیماران حدود ۶۲ سال است. بنابراین می‌خواهیم بقا و نرخ‌خطر بیماران مرد و زن ۶۲ ساله را با هم مقایسه کنیم. در این بخش برآنیم تا علاوه بر برازش مدل DPBM، مدل‌های PTWM، PTLNM و PTLGM نیز به مجموعه داده مذکور برازش داده شود. در شکل ۳ نتایج مربوط به برآورد توابع چگالی، بقا و نرخ‌خطر برای بیماران مرد و زن با استفاده از ۴ مدل مذکور نشان داده‌است. در این نمودارها نتایج ارائه شده بوسیله مدل‌های DPBM، PTWM، PTLNM و PTLGM به ترتیب با رنگ‌های سیاه، قرمز، سبز و آبی نشان داده شده‌است. نتایج مربوط به هر مدل برای بیماران مرد و زن به ترتیب با خطوط بریده شده و پُر رنگ نشان داده شده‌اند. در نمودار (الف) شکل ۳ مشاهده

می‌شود که نمودار چگالی احتمال ارائه شده بوسیله ۴ مدل برای داده‌های طول‌عمر بیماران مرد و زن ۶۲ ساله دو قله‌ای هستند. وجود دو قله (یا چندین قله) در نمودار چگالی داده‌ها معمولاً به وجود دو گروه متمایز یا بیشتر در مجموعه داده‌ها اشاره دارد. این وضعیت می‌تواند دلایل مختلفی داشته باشد. ممکن است داده‌ها از دو توزیع مختلف نشأت گرفته باشند(وجود گروه‌های متنوع با ویژگی‌های متفاوت). گاهی اوقات وجود چند قله می‌تواند به ناهنجاری‌های خاصی در داده‌ها اشاره کند که نیاز به بررسی بیشتری دارند. همچنین ممکن است متغیرهای کمکی مختلف بر روی داده‌ها تأثیر گذاشته و باعث ایجاد الگوهای متفاوت شوند. به‌طور کلی، شناسایی وجود دو یا چند قله در نمودار چگالی داده‌ها می‌تواند نشان دهنده تنوع و پیچیدگی بیشتر در ساختار داده‌های مورد مطالعه باشد و نیازمند تحلیل دقیق‌تری برای فهم بهتر آن است. در نمودارهای (ب) و (پ) شکل ۳ به ترتیب نتایج مربوط به برآورد توابع بقا و نرخ‌خطر برای بیماران مرد و زن ۶۲ ساله بوسیله ۴ مدل مختلف نشان داده شده‌است. براساس نتایج ارائه شده در این نمودارها مشاهده می‌شود بیماران مردان ۶۲ ساله در مقایسه با بیماران زنان ۶۲ ساله دارای مقادیر بقاهای کوچکتری و نرخ‌خطرهای بزرگتری هستند. این به معنای این است که بیماران مردان ۶۲ ساله نسبت به بیماران زنان ۶۲ ساله طول‌عمر کمتری دارند و احتمال مرگ آن‌ها بیشتر است. برای اینکه بتوان عملکرد ۴ مدل را در برآورد توابع چگالی، بقا و نرخ‌خطر داده‌ها با هم مقایسه کرد شاخص لگاریتم درستنمایی کناری کاذب[1] (LPML) بکار برده می‌شود (گای‌سر و ادی ، ۱۹۷۹). براساس این شاخص مدلی بهترین است که بیشترین مقدار LPML را داشته باشد. مقادیر LPML مدل‌های DPBM، PTWM، PTLNM و PTLGM به ترتیب ۱۱۵۲٫۱۳۳−، ۱۱۵۴٫۰۵۹−، ۱۱۶۰٫۸۶۸− و ۱۱۵۷٫۲۳۹− هستند. مشاهده می‌شود که بیشترین مقدار LPML مربوط به مدل DPBM است و بنابراین می‌توان گفت مدل DPBM در مقایسه با دیگر مدل‌ها برازش بهتری به داده‌ها دارد.



(پ) (ب) (الف)

شکل ۳. برآورد توابع الف–چگالی، ب–بقا و پ–نرخ‌خطر بیماران زن(خطوط پُر رنگ) و مرد(خطوط بریده) با مدل‌های DPBM(سیاه)، PTWM(قرمز)، PTLNM(سبز) و PTLGM(آبی).

[1] Log Pseudo Marginal Likelihood

۵ بحث و نتیجه‌گیری

برای مدل‌سازی توزیع جمله خطا در مدل‌های زمان شکست شتابیده از یک مدل آمیخته مبتنی بر فرایندهای دیریکله با هسته برِ ۱۲ بهره گرفته شده‌است. مدل پیشنهادی قادر است توزیع جمله خطا را بدون نیاز به فرضیات محدودکننده در مورد شکل توزیع خاص و با استفاده از خواص فرایند دیریکله، به‌صورت انعطاف‌پذیر برآورد کند. این رویکرد، امکان مدل‌سازی ساختارهای پنهان داده‌ها را فراهم کرده و برای تحلیل داده‌های بقا با سطوح مختلف سانسور مناسب است. برای ارزیابی عملکرد مدل پیشنهادی، مجموعه داده‌های شبیه‌سازی‌شده و واقعی مورد تحلیل قرار گرفت. در بخش داده‌های شبیه‌سازی‌شده، حالت‌های مختلفی شامل داده‌های سانسورشده و داده‌های ناهمگن بررسی شدند تا توانایی مدل در مواجهه با چالش‌های متداول داده‌های بقا ارزیابی شود. علاوه بر این، مجموعه داده‌های واقعی برای آزمون عملکرد مدل در شرایط واقعی بکار گرفته شد. نتایج تحلیل‌ها نشان می‌دهند که مدل پیشنهادی توانایی بالایی در برآورد توابع چگالی احتمال و بقا دارد. این مدل توانسته است ساختارهای نهفته و خوشه‌بندی موجود در داده‌ها را با دقت بالایی شناسایی کند. به‌طور کلی، یافته‌های این مقاله نشان می‌دهد که مدل پیشنهادی با بهره‌گیری از انعطاف‌پذیری فرایند دیریکله و قدرت مدل‌سازی آن، ابزار قدرتمندی برای تحلیل داده‌های بقا محسوب می‌شود. این مدل نه‌تنها پتانسیل بالایی در برآورد توابع چگالی و بقا دارد، بلکه می‌تواند به‌عنوان یک رویکرد مؤثر و قابل اعتماد برای تحلیل داده‌های بقا در شرایط مختلف بکار گرفته شود. یکی از چالش‌های موجود، در این زمینه، نیاز به زمان محاسباتی بالا برای همگرایی الگوریتم‌های نمونه‌گیری گیبز در چارچوب MCMC است. این مسئله، به‌ویژه در تحلیل داده‌های بزرگ و پیچیده، می‌تواند عملکرد محاسباتی را محدود کند. مدل پیشنهادی به انتخاب توزیع‌های پیشین برای پارامترها حساس است. انتخاب نامناسب توزیع‌های پیشین می‌تواند به تخمین‌های نادرست و یا کاهش دقت مدل منجر شود. اگرچه مدل DPBM انعطاف‌پذیری بالایی دارد، ممکن است در مواردی که داده‌ها دارای ساختارهای چندلایه یا همبستگی‌های پیچیده‌تر هستند، عملکرد آن محدود شود. پیشنهاداتی که می‌تواند مسیر تحقیقات آینده را در این حوزه روشن‌تر و قابلیت‌های مدل DPBM را در تحلیل داده‌های بقا تقویت کند عبارتند از

- طراحی الگوریتم‌های سریع‌تر و کارامدتر برای کاهش زمان محاسبات.
- گسترش مدل DPBM برای داده‌های بقا و متغیرهای وابسته به زمان یا ساختارهای مکانی–زمانی.
- ترکیب مدل‌های بیزی ناپارامتری با روش‌های یادگیری ماشین برای داده‌های حجیم و غیرساختاریافته.

تقدیر و تشکر

نویسنده‌های مقاله از راهنمایی‌ها و پیشنهادهای مفید و موثر داوران، سردبیر و ویراستار محترم مجله علوم آماری که سبب ارتقای کیفی مقاله شد، کمال تشکر و قدردانی را دارند. به دلیل طولانی بودن زمان اجرای برنامه‌های کامپیوتری از سرور مرکز شیخ بهایی دانشگاه صنعتی اصفهان استفاده کردیم. نویسنده‌های مقاله لازم می‌دانند از مدیریت مرکز مذکور کمال تشکر و قدردانی را داشته باشند.

مراجع

Al-Hussaini, E. K., and Jaheen, Z. F. (1992), Bayesian Estimation of the Parameters, Reliability and Failure Rate Functions of the Burr Type XII Failure Model, *Journal of Statistical Computation and Simulation*,**41**(1-2), 31–40.

Antoniak, C.E. (1974), Mixtures of Dirichlet Processes with Applications to Nonparametric Problems, *The Annals of Statistics*, **2**, 1152-1174.

Blackwell, D., and MacQueen, J. B. (1973), Ferguson Distributions via Polya Urn Schemes, *The Annals of Statistics*,**1**(2), 353–355.

Bolte, B., Schmidt, N., Béjar, S., Huynh, N., and Mukherjee, B. (2021), BayesSPsurv: An R Package to Estimate Bayesian (Spatial) Split-Population Survival Models, *R Journal*, **13**(1).

Burr, I. W. (1942), Cumulative Frequency Functions, *The Annals of mathematical statistics*, **13**(2), 215–232.

Cheng, N., and Yuan,T. (2013), Nonparametric Bayesian Lifetime Data Analysis Using Dirichlet Process Lognormal Mixture Model, *Naval Research Logistics*, **60**(3), 208–221.

Christensen, R., and Johnson, W. (1988), Modelling Accelerated Failure Time with a Dirichlet Process, *Biometrika*, **75**(4), 693–704.

Damien, P., Wakefield, J., and Walker, S. (1999), Gibbs Sampling for Bayesian Nonconjugate and Hierarchical Models by Using Auxiliary Variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(2), 331–344.

Escobar, M.D., and West, M. (1995), Bayesian Density Estimation and Inference Using Mixtures, *Journal of the American Statistical Association*, **90**(430), 577-588.

Ferguson,T.S. (1973), Bayesian Analysis of Some Nonparametric Problems, *The Annals of Statistics*, **1**(2), 209–230.

Ferguson,T.S., and Phadia, E. G. (1979), Bayesian Nonparametric Estimation Based on Censored Data, *The Annals of Statistics*, 163-186.

Geisser, S., and Eddy, W. (1979), A Predictive Approach to Model Selection, *Journal of the American Statistical Association*, **74**, 153–160.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Ghosh,J.K., and Ramamoorthi,R.V. (2003), *Bayesian Nonparametrics*, Springer Series in Statistics, Springer-Verlag, New York.

Ghosh, S. K., and Ghosal, S. (2006), Semiparametric Accelerated Failure Time Models for Censored Data, *Bayesian statistics and its applications*, **15**, 213–229.

Gómez, G., Calle, M. L., Oller, R., and Langohr, K. (2009), Tutorial on Methods for Interval-censored Data and Their Implementation in R, *Statistical Modelling*, **9**(4), 259–297.

Haji Joudaki, B., Hashemi, R., and Khazaei, S. (2022), Survival Analysis Using Dirichlet Process Mixture Model with Three-Parameter Burr XII Distribution as Kernel, *Communications in Statistics-Simulation and Computation*, **53**(5), 2406–2424.

Hanson, T., and Johnson, W. (2002), Modeling Regression Error with a Mixture of Polya Trees, *Journal of the American Statistical Association*, **97**(460), 1020–1033.

Hanson, T., and Johnson, W. (2004), A Bayesian Semiparametric AFT Model for Interval-censored Data, *Journal of Computational and Graphical Statistics*, **13**(2), 341–361.

Hanson,T. (2006), Modeling Censored Lifetime Data Using a Mixture of Gammas Baseline, *Bayesian Analysis*, **1**(3), 575–594.

Hanson,T. (2006), Inference for Mixtures of Finite Polya Tree Models, *Journal of the American Statistical Association*, **101**(476), 1548–1565.

Hanson, T., and Yang, M. (2007), Bayesian Semiparametric Proportional Odds Models, *Biometrics*, **63**(1), 88–95.

Hanson,T. (2014), Polya Trees and their Use in Reliability and Survival Analysis, *Wiley StatsRef: Statistics Reference Online*.

Johnson, W., and Christensen, R. (1989), Nonparametric Bayesian Analysis of the Accelerated Failure Time Model, *Statistics & Probability Letters*, **8**(2), 179–184.

Kalbfleisch, J. D., and Prentice, R. L. (2011), *The Statistical Analysis of Failure Time Data*, John Wiley and Sons, New York.

Kottas, A. (2006), Nonparametric Bayesian Survival Analysis Using Mixtures of Weibull Distributions, *Statistical Planning and Inference*, **136**(3), 578–596.

Kottas, A. (2006), Dirichlet Process Mixtures of Beta Distributions, with Applications to Density and Intensity Estimation, *In Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning (ICML)*, **47**.

Kuo, L., and Mallick, B. (1997), Bayesian Semiparametric Inference for the Accelerated Failure Time Model, *Canadian Journal of Statistics*, **25**(4), 457–472.

Lahiri, P., and Park, D.H. (2000), Nonparametric Bayes and Empirical Bayes Estimators of Mean Residual Life at Age t, *Journal of Statistical Planning and Inference*, **29**(1-2), 125–136.

Lawless, J. F. (2011), *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York.

Neal, R.M. (2000), Markov Chain Sampling Methods for Dirichlet Process Mixture Models, *Journal of Computational and Graphical Statistics*, **9**(2), 249–265.

Okasha, M.K., and Matter, M.Y. (2015), On the Three-parameter Burr Type XII Distribution and its Application to Heavy Tailed Lifetime Data, *Journal of Advances in Mathematics*, **10**(4), 3429–3442.

Sethuraman, J. (1994), A Constructive Definition of Dirichlet Priors, *Statistica Sinica*, 639–650.

Susarla,V., and Van Ryzin, J. (1976), Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations, *Journal of the American Statistical Association*, **71**, 897–902.

Walker, S., and Mallick, B. (1999), A Bayesian Semiparametric Accelerated Failure Time Model, *Biometrics*, **55**(2), 477–483.

Zhao, L., Hanson, T., and Carlin, B. P. (2009), Mixtures of Polya Trees for Flexible Spatial Frailty Survival Modelling, *Biometrika*, **96**(2), 263–276.