Journal of Statistical Sciences, Autumn and Winter, 2023
Vol. 17, No. 2, pp 275-298
DOI: 10.52547/jss.17.2.03

Analysis of Censored Data Using Dirichlet Process Mixture Model with Generalized Inverse Weibull Distribution as Kernel

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

Corresponding author: B. Haji Joudaki, Bahram600631@gmail.com

Received: 23/4/2023 **Revised:** 23/7/2023 **Accepted and Published Online:** 31/7/2023.

Introduction

Evaluating complex data distribution, such as multimodal cases, requires more complex statistical models. The complexity of a statistical model is related to the number of model parameters. To achieve a more complex model and consequently achieve better flexibility in statistical inference, one can use an infinite-dimensional family of probability models. Using the Bayesian approach for the infinite-dimensional parameter need a suitable prior distribution. Typically, the prior distribution of such parameters are stochastic processes. Such priors are called nonparametric Bayes priors. The most important nonparametric Bayes prior is the Dirichlet process, first introduced by Ferguson (1973). Due to the discreteness of the Dirichlet process, using a mixed model of Dirichlet processes (DPMM) (Antoniak, 1974) is preferable. The ability to cover multimodal data distribution and perform data clustering are two advantages of DPMMs. In this article, we profit from these two advantages of DPMMs. Kernel selection is an essential issue in working with DPMMs. According to under study data, a flexible distribution should be selected as the kernel of DPMM. For lifetime data, distributions such as Weibull, lognormal, or other lifetime distributions can be chosen. The generalized inverse Weibel distribution is a flexible distribution introduced by de Gusmão, et al. (2011). In this article, the generalized inverse Weibel distribution is considered the kernel of DPMM.

Material and Methods

In fitting a DPMM, prolongation of the execution time of Markov chain

Monte Carlo simulation algorithms is challenging. Dirichletprocess package in R software is a resolution. The mentioned package has a high ability to fit DPMMs and other non-parametric Bayes models. After obtaining posterior samples, survival density and hazard rate functions can be easily estimated using this package. Finally, by analyzing several real and simulated data sets, the performance of the proposed model is evaluated.

Results and Discussion
We designed a simulation study to evaluate the performance of the proposed model under different prior distributions for the accuracy parameter of the Dirichlet process based on sample sizes of 100 and 1000. Bayesian and interval estimations of the survival, density, and hazard rate functions of a mixed model of two generalized inverse Weibull distributions reported. The results show that the proposed model has a high potential in estimating the mentioned functions. The proposed model was also used to analyze several real multimodal data sets. The results show that the proposed model performed better compared to other methods. The proposed model is also applied for real data clustering and simulation.

Conclusion
A mixed model of Dirichlet processes with a generalized inverse Weibull kernel is used to analyze right-censored survival data. The performance of the proposed model is evaluated based on Several simulated and real data sets. Achieved results in this paper show that the proposed model has good potential for estimating density, survival, and hazard rate functions in survival data. Another advantage of the proposed model is its high potential for data clustering.

Keywords: Mixture model, Dirichlet process, Generalized inverse Weibull Distribution, Right censoring, Markov chain monte carlo.

Mathematics Subject Classification (2010): 62F10, 62N01.

مجله علوم آماری، پاییز و زمستان ۱۴۰۲
جلد ۱۷، شماره ۲، ص ۲۷۵ -- ۲۹۸
DOI: 10.52547/jss.17.2.03
مقاله پژوهشی

تحلیل داده‌های سانسور شده با استفاده از مدل آمیخته فرایند دیریکله با هسته وارون وایبل تعمیم‌یافته

بهرام حاجی جودکی[۱]، سیدرضا هاشمی[۲]، سلیمان خزایی[۲]
[۱]گروه آمار، دانشکده علوم پایه، دانشگاه لرستان
[۲]گروه آمار، دانشکده علوم، دانشگاه رازی

نویسنده مسئول : بهرام حاجی جودکی، bahram600631@gmail.com
تاریخ دریافت: ۱۴۰۱/۰۰/۰۰ تاریخ بازنگری: ۱۴۰۲/۰۰/۰۰ تاریخ پذیرش و انتشار: ۱۴۰۲/۰۰/۰۰

چکیده: در این مقاله یک مدل آمیخته فرایند دیریکله جدید با هسته وارون وایبل تعمیم‌یافته پیشنهاد شده است. پس از تعیین توزیع پیشین پارامترها در مدل پیشنهادی، برای نمونه‌گیری از توزیع پسین توام پارامترها از روش‌های مونت کارلوی زنجیره مارکف استفاده شده است. عملکرد مدل پیشنهادی با تحلیل چندین مجموعه داده واقعی و شبیه‌سازی شده مورد بررسی قرار گرفته است. در مجموعه داده‌های واقعی برخی از داده‌ها سانسور شده از راست هستند. همچنین در این مقاله پتانسیل مدل پیشنهادی برای خوشه‌بندی کردن داده‌ها بکار گرفته شده است. نتایج بدست آمده نشان دهنده عملکرد مطلوب مدل پیشنهادی است.
واژه‌های کلیدی: مدل آمیخته، فرایند دیریکله، وارون وایبل تعمیم‌یافته، سانسور از راست، مونت کارلوی زنجیره مارکف.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62F10، 62N01.

۱ مقدمه

یکی از مهم‌ترین دغدغه‌های یک آماردان، انتخاب مدلی برای انجام استنباط آماری است. مسئله انتخاب مدل از دیدگاه نظری و عملی یک مسئله مهم اما دشوار است. اکثر آماردانان بر این باور هستند که عمده مشکلاتی که در استنباط آماری هست را می‌توان مشکلات مربوط به مدل‌سازی آماری در نظر گرفت. در واقع نحوه ترجمه از

مسئله موضوعی به مدل آماری اغلب حیاتی‌ترین بخش یک تحلیل آماری است. در علم آمار برای اینکه بتوان به راحتی به مجموعه‌ای از استنباط‌های آماری دست یافت، معمولاً یک خانواده از مدل‌های احتمال انتخاب می‌شود که این خانواده از مدل‌ها با مجموعه‌ای از پارامترهای که بُعد متناهی دارند اندیس‌گذاری می‌شوند. حال اگر برای پارامترهای مدل، توزیع‌های پیشین‌های انتخاب شود آنگاه می‌توان توزیع پسین پارامترها را محاسبه کرد. در ادامه هر نوع استنباطی در مورد پارامترها برمبنای توزیع پسین پارامترها خواهد بود که از این رویکرد به عنوان رویکرد بیز پارامتری یاد می‌شود. مدل‌های احتمال ساده، معمولاً توانایی پوشش دادن توزیع داده‌های پیچیده همانند داده‌های چند مُدی را ندارند، بنابراین اگرچه کار کردن با این مدل‌ها راحت‌تر است ولی ممکن است نتایج بدست آمده بر اساس این مدل‌ها با نتایج واقعی اختلاف اساسی داشته باشند. برای پوشش دادن توزیع داده‌های پیچیده، به ناچار باید از مدل‌های آماری پیچیده‌تر استفاده کرد. در کل پیچیدگی یک مدل آماری با تعداد پارامترهای مدل مرتبط است. برای دستیابی به یک مدل پیچیده‌تر و در نتیجه دستیابی به نتایج بهتر، معقول است راهکاری برای انجام استنباط آماری برگزیده شود که در چارچوب این راهکار مجاز به داشتن خانواده‌ای بینهایت بُعدی از مدل‌های احتمال باشیم. در این صورت یک کلاس غنی‌تر و بزرگ‌تر از مدل‌های احتمال در دسترس خواهد بود. برای اینکه بتوان رویکرد بیزی را برای چنین مدل احتمالی پیاده‌سازی کرد باید برای پارامتر بینهایت بُعدی مدل مذکور توزیع پیشینی انتخاب شود. با توجه به نامتناهی بودن بُعد چنین پارامتری نمی‌توان برای آن هر توزیع پیشینی را در نظر گرفت و نوعاً توزیع پیشین چنین پارامترهای از نوع فرایندهای تصادفی هستند که چنین پیشین‌های را اصطلاحاً پیشین‌های بیز ناپارامتری می‌نامند. چنانچه برای انجام استنباط آماری از چنین رویکردی استفاده شود در این صورت گوییم از رویکرد بیزی ناپارامتری استفاده شده است. مهم‌ترین پیشین بیز ناپارامتری فرایند دیریکله[1](DP) است که اولین بار توسط فرگوسن (۱۹۷۳) معرفی و برخی از ویژگی‌های مهم آن بیان و اثبات شد. سوسارلا و ون رایزین (۱۹۷۶)، فرگوسن و فادیا (۱۹۷۹)، لاهیری و پارک (۱۹۹۱) و کائو و همکاران (۱۹۹۲) برای مدل‌بندی تابع توزیع بقا از توزیع پیشین DP استفاده کردند. علیرغم جذابیت‌های توزیع پیشین DP، اما این فرایند یک فرایند گسسته است بنابراین تقریب زدن توابع پیوسته‌ای مانند چگالی، بقا و نرخ‌خطر با آن خالی از اشکال نخواهد بود. با پیچش توزیع پیشین DP با برخی توزیع‌های پیوسته می‌توان مشکل مذکور را رفع کرد که این منجر به معرفی مدل‌های آمیخته‌ای از فرایندهای دیریکله[2](DPMM) می‌شود (آنتونیاک ، ۱۹۷۴). توانایی در پوشش دادن توزیع داده‌های چندمُدی و انجام خوشه‌بندی داده‌ها از مهم‌ترین مزایایDPMMها است. در این مقاله این دو مزیت DPMMها مورد استفاده قرار می‌گیرد. برای دستیابی به اطلاعات بیشتر در مورد پیشین‌های بیز ناپارامتری به فرگوسن (۱۹۷۳)، فرگوسن (۱۹۷۴) و قوش و رامامورتی (۲۰۰۳) رجوع شود. انتخاب هسته موضوع مهمی در کار کردن با DPMMها است. در واقع با توجه به نوع داده‌های تحت مطالعه توزیعی انعطاف‌پذیر به عنوان هسته DPMM انتخاب می‌شود. برای داده‌های حقیقی مقدار، معمولاً هسته توزیع نرمال یا نرمال‌چوله انتخاب می‌شود. برای داده‌های طول‌عمر، توزیع‌هایی مانند وایبل، گاما، لگ‌نرمال، بور ۱۲ و

[1]Dirichlet Process
[2]Dirichlet Process Mixture Model

یا دیگر توزیع‌های طول‌عمر انتخاب‌های مناسب‌تری برای هسته هستند. برای داده‌های با مقادیر در بازه $(۰,۱)$، معمولاً هسته توزیع بتا است. هنسن (۲۰۰۶)، کوتاس (a۲۰۰۶)، کوتاس (b۲۰۰۶)، چنگ و یوان (۲۰۱۳)، بهلوری حجار و خزائی (۲۰۱۸) و حاجی جودکی و همکاران (۲۰۲۲) DPMM با هسته‌های به ترتیب گاما، وایبل، بتا، لگ‌نرمال، بور۱۲ دو پارامتری و بور۱۲ سه پارامتری را برای تحلیل داده‌های بقای سانسور شده از راست بکار بردند. اگرچه DPMM با هسته‌های مذکور مدل‌های انعطاف‌پذیری را ارائه می‌دهند اما هیچ کدام از این مدل‌ها لزوماً همیشه نتایج خوبی را فراهم نمی‌کنند. با انتخاب دیگر توزیع‌های انعطاف‌پذیر به عنوان هسته DPMM می‌توان مدل‌های آمیخته جدیدی معرفی کرد که این مدل‌ها نیز ممکن است در برخی مجموعه داده‌ها در مقایسه با دیگر DPMM ها نتایج بهتری ارائه دهند. یکی از توزیع‌های انعطاف‌پذیری که اخیراً در تحلیل داده‌های بقا و قابلیت اعتماد فراوان بکار می‌رود توزیع وارون وایبل تعمیم‌یافته[1] (GIW) است که توسط دی گاس مائو و همکاران (۲۰۱۱) معرفی شد. این توزیع همانند توزیع لگ‌نرمال دارای تابع نرخ‌خطر تک مُدی است. برای دستیابی به اطلاعات بیشتر در مورد توزیع GIW به دی گاس مائو و همکاران (۲۰۱۱) و کائو و همکاران (۲۰۱۸) رجوع شود. در این مقاله هدف این است که یک مدل DPM جدید با هسته GIW معرفی شود و برخی از پتانسیل‌های این مدل جدید خصوصاً در انجام خوشه‌بندی و مدل‌بندی داده‌های بقای چند مُدی نشان داده شود.

سانسور پدیده‌ای رایج در تحلیل داده‌های بقا است که خود دارای انواع مختلفی مانند سانسور از راست، سانسور از چپ، سانسور فاصله‌ای و غیره است. سانسور متداول در داده‌های بقا، سانسور از راست است که خود دارای دو نوع سانسور زمان (نوع ۱) و سانسور شکست (نوع۲) است. سانسور مورد علاقه در این مقاله سانسور زمان است. یک منبع مفید برای مطالعه انواع سانسور لاولس (۲۰۱۱) است.

در بخش ۲ ابتدا برخی تعاریف را ارائه داده و آنگاه در مورد توزیع GIW صحبت می‌شود. در بخش ۳، DPMM با هسته GIW معرفی می‌شود و فرایند اجرای نمونه‌گیری از توزیع پسین توام پارامترها شرح داده می‌شود. در بخش‌های ۴ و ۵ مدل پیشنهادی برای تحلیل مجموعه داده‌های شبیه‌سازی شده و واقعی بکار برده می‌شود و در نهایت در بخش پایانی نتایج بدست آمده ارائه شده است.

۲ تعاریف

تعریف ۱. فرض کنید سه‌گانه (A,B,P) یک فضای احتمال، P_0یک اندازه احتمال روی فضای A و ν نیز یک مقدار عددی مثبت است. اگر برای هر افزار اندازه‌پذیر متناهی دلخواه $B_1,\cdots,B_k$ از A رابطه $(P(B_1),\cdots,P(B_k)) \sim Dir(\nu P_0(B_1),\cdots,\nu P_0(B_k))$ برقرار باشد، آنگاه P یک فرایند دیریکله با پارامترهای ν و P_0 است و با نماد $P \sim DP(\nu,P_0)$نشان داده می‌شود که در آن نماد Dir برای نشان دادن توزیع دیریکله بکار رفته است.

[1] Generalized Inverse Weibull

معمولاً برای مدل‌بندی توزیع داده‌های (بقای) چند مُدی از مدل‌های آمیخته استفاده می‌شود. یکی از مدل‌های آمیخته‌ای که برای این منظور می‌تواند مورد استفاده قرار گیرد مدل DPM است. این مدل بوسیله آنتونیاک (۱۹۷۴) معرفی شد. در حالت کلی یک مدل DPM به صورت

$$f(t:P)=\int_{\Theta}k(t|\boldsymbol{\theta})dP(\boldsymbol{\theta})$$

تعریف می‌شود، که در آن k هسته مدل آمیخته، $\boldsymbol{\theta}$ بردار پارامتری و Θ فضای پارامتری، P تابع توزیع پیشین بردار پارامتری $\boldsymbol{\theta}$ است که نقش وزن‌های مدل آمیخته را ایفا می‌کند. در این چارچوب این اندازه احتمال تصادفی است و فرض می‌شود $P\sim DP(\nu,P_{\circ})$. در بخش قبلی اشاره شد که در این مقاله هدف این است که توزیع GIW به عنوان هسته DPMM بکار برده شود. توابع چگالی، بقا و نرخ خطر[1] توزیع GIW به ترتیب به صورت

$$\begin{aligned}
k(t|\alpha,\beta,\lambda)&=\beta\lambda\alpha^{\beta}x^{-\beta-1}\exp(-\lambda(\frac{\alpha}{t})^{\beta}),\quad t>0,\alpha>0,\beta>0,\lambda>0\\
S(t|\alpha,\beta,\lambda)&=1-\exp(-\lambda(\frac{\alpha}{t})^{\beta}),\quad t>0\\
h(t|\alpha,\beta,\lambda)&=\frac{\beta\lambda\alpha^{\beta}x^{-\beta-1}\exp(-\lambda(\frac{\alpha}{t})^{\beta})}{1-\exp(-\lambda(\frac{\alpha}{t})^{\beta})},\quad t>0
\end{aligned}$$

هستند، که در آن α پارامتر مقیاس و β و λ نیز پارامترهای شکل توزیع هستند. توزیع GIW با پارامترهای مذکور با نماد $GIW(\alpha,\beta,\lambda)$ نشان داده می‌شود. اگر $\lambda=1$ باشد آنگاه توزیع GIW به توزیع وارون وایبل تبدیل می‌شود. در شکل ۱ نمودار توابع چگالی و نرخ خطر توزیع GIW به ازای مقادیر مختلف پارامترهای α، β و λ رسم شده است. یکی از ویژگی‌های مهم توزیع‌های طول‌عمر که معمولاً در تحلیل داده‌های بقا مورد توجه است شکل تابع خطر است. تابع خطر توزیع GIW تک مُدی است که این وضعیت در نمودار ب شکل ۱ به وضوح مشاهده می‌شود.

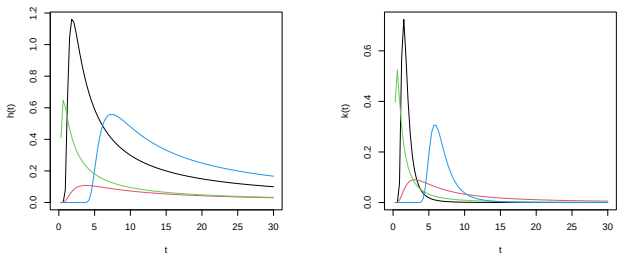
۳ معرفی مدل

مدل آمیخته فرایند دریکله (DPMM) با هسته GIW به صورت

$$f(t:P)=\int_{\alpha}\int_{\beta}\int_{\lambda}k(t|\alpha,\beta,\lambda)dP(\alpha,\beta,\lambda)$$

تعریف می‌شود، که در آن k تابع چگالی توزیع GIW و P نیز دارای توزیع پیشین $DP(\nu,P_{\circ})$ است. در این مدل پارامتر ν را کمیتی تصادفی در نظر گرفته و برای آن مشابه اسکوبار و وست (۱۹۹۵) توزیع پیشین گامای $G(a_{\nu},b_{\nu})$

[1] Hazard Rate



(ب) (الف)

شکل ۱. الف- نمودار تابع چگالی، ب- نمودار تابع نرخ خطر توزیع GIW به ازای مقادیر مختلف پارامترهای α، β و λ.

در نظر گرفته شده است. فرض کنید $p_\circ$ تابع چگالی متناظر با تابع توزیع $P_\circ$ است. برای پارامترهای α، β و λ به ترتیب توزیع‌های پیشینی $\alpha|\delta \sim Exp\{\frac{1}{\delta}\}$، $\beta|\phi \sim U(\circ,\phi)$ و $\lambda|\gamma \sim Exp\{\frac{1}{\gamma}\}$ در نظر گرفته می‌شود که در آن نمادهای Exp و U به ترتیب برای بیان توزیع‌های نمایی و یکنواخت بکار رفته‌اند. با فرض استقلال این پیشین‌ها تابع چگالی پیشین توام پارامترها به صورت $p_\circ(\alpha,\beta,\lambda|\delta,\phi,\gamma) = \pi(\alpha|\delta)\pi(\beta|\phi)\pi(\lambda|\gamma)$ خواهد بود. برای ابرپارامترهای δ، ϕ و γ به ترتیب توزیع‌های پیشین مزدوجی به صورت $\delta \sim IG(a_\delta, b_\delta)$، $\phi \sim Par(a_\phi, b_\phi)$ و $\gamma \sim IG(a_\gamma, b_\gamma)$ انتخاب شده است که در آن نمادهای IG و Par به ترتیب برای نشان دادن توزیع‌های گامای وارون و پاراتو بکار 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.

سانسور مورد نظر در این تحقیق سانسور از راست است. تحت این نوع سانسور مجموعه داده‌ها به صورت $(r_i, t_i^*); i = ۱,\cdots,n$ هستند. $t_i^* = \min(t_i, C_i)$ زمان بقای(طول عمر) مشاهده شده فرد iام است به طوریکه t_i و C_i به ترتیب زمان بقای واقعی و زمان سانسور فرد iام هستند. همچنین 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 $r_i = \circ$ باشد طول عمر فرد iام سانسور شده از راست در زمان C_i است. نمایش سلسله مراتبی DPMM با هسته GIW را می‌توان به صورت

$$
\begin{aligned}
&t_i^*|\alpha_i,\beta_i,\lambda_i \overset{i.d.}{\sim} \{k(t_i^*|\alpha_i,\beta_i,\lambda_i)\}^{r_i}\{S(t_i^*|\alpha_i,\beta_i,\lambda_i)\}^{۱-r_i},\\
&(\alpha_i,\beta_i,\lambda_i)|P \overset{i.i.d.}{\sim} P, \quad i = ۱,\cdots,n,\\
&P|\delta,\phi,\gamma \sim DP(\nu, P_\circ(\cdot|\delta,\phi,\gamma)), \qquad p_\circ(\alpha,\beta,\lambda|\delta,\phi,\gamma) = \pi(\alpha|\delta)\pi(\beta|\phi)\pi(\lambda|\gamma),\\
&\alpha|\delta \sim Exp\{\frac{1}{\delta}\}, \quad \beta|\phi \sim U(\circ,\phi), \quad \lambda|\gamma \sim Exp\{\frac{1}{\gamma}\}, \quad \delta \sim IG(a_\delta, b_\delta),\\
&\phi \sim Par(a_\phi, b_\phi), \quad \gamma \sim IG(a_\gamma, b_\gamma), \quad \nu \sim G(a_\nu, b_\nu)
\end{aligned}
$$

نوشت. با توجه به ارائه سلسله مراتبی از مدل پیشنهادی واضح است که $(\alpha_i, \beta_i, \lambda_i)$ پارامتر مربوط به داده iام و $\{(\alpha_i, \beta_i, \lambda_i); i = ۱, \cdots, n\}$ نیز مجموعه پارامترهای مربوط به n داده است. اگر n^* تعداد بردارهای متمایز در مجموعه اخیر فرض شود آنگاه مجموعه بردارهای متمایز به صورت $\{(\alpha_j^*, \beta_j^*, \lambda_j^*); j = ۱, \cdots, n^*\}$ 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 $\boldsymbol{L} = (L_۱, \cdots, L_n)$ وضعیت خوشه‌بندی n داده را نشان دهد. معمولاً از این بردار به عنوان بردار پیکربندی[1] یاد می‌شود. داده iام در خوشه jام قرار می‌گیرد، یعنی $L_i = j$ است اگر و تنها اگر رابطه $(\alpha_i, \beta_i, \lambda_i) = (\alpha_j^*, \beta_j^*, \lambda_j^*)$ برقرار باشد. برای ادامه کار بردارهای $\boldsymbol{\alpha}$، $\boldsymbol{\beta}$، $\boldsymbol{\lambda}$، $\boldsymbol{r}$ و $\boldsymbol{t}^*$ به ترتیب به صورت $\boldsymbol{\alpha} = (\alpha_۱, \cdots, \alpha_n)$، $\boldsymbol{\beta} = (\beta_۱, \cdots, \beta_n)$، $\boldsymbol{\lambda} = (\lambda_۱, \cdots, \lambda_n)$، $\boldsymbol{r} = (r_۱, \cdots, r_n)$ و $\boldsymbol{t}^* = (t_۱^*, \cdots, t_n^*)$ تعریف می‌شوند. همچنین $\boldsymbol{\alpha}_{-i}$، $\boldsymbol{\beta}_{-i}$ و $\boldsymbol{\lambda}_{-i}$ بیانگر بردارهای $\boldsymbol{\alpha}$، $\boldsymbol{\beta}$ و $\boldsymbol{\lambda}$ بدون i امین مولفه آنها هستند. با توجه به آنتونیاک (۱۹۷۴) توزیع پسین توام پارامترها به صورت

$$\pi\left[P, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu | \boldsymbol{t}^*, \boldsymbol{r}\right] \propto \pi\left[P | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu\right] \pi\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu | \boldsymbol{t}^*, \boldsymbol{r}\right] \quad (۱)$$

است، که در آن

$$P(\cdot) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu \sim DP(\nu + n, \frac{\nu}{\nu + n} P_\circ(\cdot | \phi, \delta, \gamma) + \frac{۱}{\nu + n} \sum_{i=۱}^{n} I_{(\alpha_i, \beta_i, \lambda_i)}(\cdot))$$

$$\pi\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu | \boldsymbol{t}^*, \boldsymbol{r}\right] \propto \pi\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda} | \phi, \delta, \gamma, \nu, \boldsymbol{t}^*, \boldsymbol{r}\right] \pi\left[\phi\right] \pi\left[\delta\right] \pi\left[\gamma\right] \pi\left[\nu\right]$$

$$\pi\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda} | \phi, \delta, \gamma, \nu, \boldsymbol{t}^*, \boldsymbol{r}\right] \propto \prod_{i=۱}^{n} \pi\left[\alpha_i, \beta_i, \lambda_i | \boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}, \nu, \phi, \delta, \gamma\right]$$
$$\times \{k(t_i | \alpha_i, \beta_i, \lambda_i)\}^{r_i} \{S(C_i | \alpha_i, \beta_i, \lambda_i)\}^{۱ - r_i}$$

توزیع شرطی $\pi\left[\alpha_i, \beta_i, \lambda_i | \boldsymbol{\alpha}_{-i}, \boldsymbol{\beta}_{-i}, \boldsymbol{\lambda}_{-i}, \nu, \phi, \delta, \gamma\right]$ نیز در بلکول و مک‌کوین (۱۹۷۳) داده شده است. با توجه به رابطه (۱) فرایند تولید نمونه از توزیع پسین توام پارامترها دو مرحله‌ای است. درگام اول با استفاده از الگوریتم نمونه‌گیری گیبز از توزیع $\pi\left[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu | \boldsymbol{t}^*, \boldsymbol{r}\right]$ نمونه‌های پسینی پارامترها را بدست آورده و آنگاه در گام دوم با توجه به اینکه توزیع شرطی $P(\cdot) | (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu)$ یک DP است با بکار بردن ارائه ستورامان (۱۹۹۴) از DP از آن نمونه تولید می‌شود. ارائه مذکور به فرم $P(\cdot) | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \phi, \delta, \gamma, \nu \approx \sum_{j=۱}^{J} \omega_j I_{(\alpha_j, \beta_j, \lambda_j)}(\cdot)$ است که در آن J دنباله مقادیر وزن‌های مدل است که برای محاسبه این مقادیر از روش چوب

[1] Configuration vector

شکنی[۲] استفاده می‌شود. مقادیر $(\alpha_j, \beta_j, \lambda_j); j = ۱, \cdots, J$ نیز از توزیع آمیختهٔ

$$\frac{\nu}{\nu+n} P_۰(\cdot|\phi, \delta, \gamma) + \frac{۱}{\nu+n} \sum_{i=۱}^{n} I_{(\alpha_i, \beta_i, \lambda_i)}(\cdot)$$

تولید می‌شوند. برای تعیین J (نقطه برش) 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. یکی از مشکلاتی که در برازش DPMM ها با آن روبرو هستیم طولانی بودن زمان اجرای الگوریتم‌های مونت کارلوی زنجیره مارکف[۱] (MCMC) مربوط به این مدل‌ها است. در سال‌های گذشته چندین بسته مانند Nimble، Dpweibull، DPackage و غیره برای این منظور طراحی شده است. بسته مفیدی که بوسیله راس و مارک‌ویک (۲۰۱۸) برای برازش DPMMها در نرم‌افزار R معرفی شد بسته Dirichletprocess است. اگر چه در این بسته، به صورت پیش‌فرض توزیع وایبل به عنوان هسته DPMM انتخاب شده است، اما می‌توان هر توزیع دلخواهی را نیز به عنوان هسته DPMM انتخاب کرد. همچنین در بسته مذکور می‌توان با داده‌های سانسور شده نیز کار کرد. در نیل (۲۰۰۰) برای اجرای الگوریتم‌های MCMC در مدل‌های DPM تعداد ۸ الگوریتم داده شده است که الگوریتم شماره ۸ مهم‌تر و پُرکاربردتر از بقیه الگوریتم‌ها است. در بسته Dirichletprocess می‌توان دو الگوریتم ۴ و ۸ مقاله نیل (۲۰۰۰) را بکار برد که در این مقاله از الگوریتم ۸ استفاده شده است. الگوریتم ۸ نیل (۲۰۰۰) بر اساس نمایش کیسه پولیا[۲] از DP است. برای دستیابی به اطلاعات بیشتر در مورد این ارائه از DP به بلکول و مک‌کوین (۱۹۷۳) رجوع شود. در الگوریتم ۸ نیل (۲۰۰۰) بروزرسانی مولفه‌های بردار پیکربندی و پارامترهای خوشه‌ها تا زمانی ادامه می‌یابد که همگرایی زنجیره‌های مارکف حاصل شود. در فرایند اجرای الگوریتم نمونه‌گیری گیبز ۴ گام زیر تکرار می‌شود.

الگوریتم ۱. نمونه‌گیری گیبز :
گام ۱- بروزرسانی مولفه‌های بردار پیکربندی.
گام ۲- بروزرسانی پارامترهای خوشه‌ها بر مبنای داده‌هایی که در هر خوشه داریم.
گام ۳- بروزرسانی ابرپارامترهای δ، ϕ و γ.
گام ۴- بروزرسانی پارامتر ν.

برای دستیابی به جزئیات اجرای دو گام اول الگوریتم نمونه‌گیری گیبز به الگوریتم ۸ مقاله نیل (۲۰۰۰) رجوع شود. در فرایند اجرای گام ۳ الگوریتم گیبز باید از توزیع‌های شرطی ابر پارامترهای δ، ϕ و γ نمونه تولید کرد. این

[2] Stick-breaking
[1] Markov Chain Monte Carlo
[2] Polya urn

توزیع‌های شرطی بعد از برخی محاسبات و ساده‌سازی‌ها به صورت

$$\delta|\boldsymbol{\alpha}^*, n^* \sim IG(\text{shape} = a_\delta + n^*, \text{scale} = b_\delta + \sum_{j=1}^{n^*} \alpha_j^*),$$
$$\phi|\boldsymbol{\beta}^*, n^* \sim Par(a_\phi + n^*, \max(b_\phi, \max_{1\le j\le n^*} \beta_j^*)),$$
$$\gamma|\boldsymbol{\lambda}^*, n^* \sim IG(\text{shape} = a_\gamma + n^*, \text{scale} = b_\gamma + \sum_{j=1}^{n^*} \lambda_j^*).$$

هستند، که در آن $\boldsymbol{\alpha}^*$، $\boldsymbol{\beta}^*$ و $\boldsymbol{\lambda}^*$ به ترتیب به صورت $\boldsymbol{\alpha}^* = \{\alpha_j^*; j = 1, \cdots, n^*\}$، $\boldsymbol{\beta}^* = \{\beta_j^*; j = 1, \cdots, n^*\}$ و $\boldsymbol{\lambda}^* = \{\lambda_j^*; j = 1, \cdots, n^*\}$ هستند. برای اجرای گام ۴ الگوریتم گیبز از الگوریتم ارائه شده در اسکوبار و وست (۱۹۹۵) استفاده می‌شود. در تکرار bام الگوریتم گیبز بر مبنای مقادیر تولید شده $\{(\alpha_j^{*(b)}, \beta_j^{*(b)}, \lambda_j^{*(b)}); j = 1, \cdots, n^{*(b)}\}$ و مقادیر وزن‌ها $\omega_j^{(b)}; j = 1, \cdots, n^{*(b)}$، توابع چگالی، بقا و نرخ‌خطر در نقطه t به صورت

$$f(t) \approx \sum_{j=1}^{n^{*(b)}} \omega_j^{(b)} k(t|\boldsymbol{\theta}_j^{*(b)}), \quad S(t) \approx \sum_{j=1}^{n^{*(b)}} \omega_j^{(b)} S(t|\boldsymbol{\theta}_j^{*(b)}), \quad h(t) \approx \frac{f(t)}{S(t)}.$$

تقریب زده می‌شوند، که در آن $\boldsymbol{\theta}_j^{*(b)} = (\alpha_j^{*(b)}, \beta_j^{*(b)}, \lambda_j^{*(b)})$ و $n^{*(b)}$ نیز تعداد خوشه‌ها در تکرار b ام الگوریتم گیبز است.

۴ مطالعه شبیه‌سازی

برای ارزیابی عملکرد مدل پیشنهادی یک مطالعه شبیه‌سازی طراحی می‌شود. برای این منظور از مدل

$$f(t) = ۰/۶ f_{GIW}(t|\alpha_1, \beta_1, \lambda_1) + ۰/۴ f_{GIW}(t|\alpha_2, \beta_2, \lambda_2), \qquad (۲)$$

که آمیخته‌ای از دو توزیع GIW است، نمونه‌های در اندازه‌های $n = ۱۰۰, ۱۰۰۰$ تولید می‌شود. برای پارامترهای مدل آمیخته (۲) دو مجموعه به صورت

$$I: \alpha_1 = ۳, \ \beta_1 = ۲, \ \lambda_1 = ۶, \ \alpha_2 = ۱۵, \ \beta_2 = ۸, \ \lambda_2 = ۴$$
$$II: \alpha_1 = ۱, \ \beta_1 = ۴, \ \lambda_1 = ۸, \ \alpha_2 = ۵, \ \beta_2 = ۸, \ \lambda_2 = ۱۰$$

نظر گرفته شده است. در ادامه برای سادگی، از مدل (۲) با دو مجموعه I و II از پارامترهای این مدل به ترتیب به عنوان مدل‌های I و II یاد می‌شود. برای نشان دادن عدم حساسیت استنباط نسبت به انتخاب پارامترهای توزیع پیشین پارامتر ν یعنی (a_ν, b_ν) دو مجموعه از مقادیر $(a_\nu, b_\nu) = (۲, ۲), (۳, ۰/۰۵)$ در نظر گرفته شده‌اند. همچنین با روش آزمایش و خطا این نتیجه حاصل شده است که با انتخاب مقادیر $b_\delta = ۱۰۰۰۰$، $b_\phi = ۱۰$ و $b_\gamma = ۱۰۰$ نتایج خوبی بدست می‌آید. در فرایند اجرای الگوریتم گیبز، تعداد کل تکرارهای الگوریتم گیبز، تعداد تکرارهای دوره سوخت و فاصله تاخیر به ترتیب برابر با مقادیر $N = ۳۰۰۰۰$، $B = ۱۰۰۰۰$ و ۲۰ تعیین شده است و بنابراین حجم نمونه نهایی در استنباط ۱۰۰۰ 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 نرم‌افزار R مورد استفاده قرار گرفته است. نتایج بدست آمده همگرایی الگوریتم‌های نمونه‌گیری گیبز را تایید می‌کنند. در مجموعه داده‌های شبیه‌سازی شده برای اینکه نتایج برآورد توابع چگالی، بقا و نرخ خطر با مقادیر واقعی این توابع مقایسه شود از شاخص‌های اقلیدسی، میانگین قدر مطلق خطا[1] و هلینگر[2] استفاده شده است. در جدول‌های ۱ و ۲ سه شاخص مذکور به ترتیب با نمادهای d_E، d_{MAE} و d_H نشان داده شده‌اند. مقادیر شاخص‌های انحراف در برآورد منحنی‌های واقعی(چگالی، بقا و نرخ خطر) مدل‌های I و II به ترتیب در جدول‌های ۱ و ۲ داده شده است. مقادیر انحراف‌های مذکور تحت دو اندازه نمونه۱۰۰ و ۱۰۰۰ و تحت دو توزیع پیشین پارامتر ν ارائه شده‌اند. با توجه به مقادیر داده شده در جدول ۱ مشاهده می‌شود که برای هر یک از دو اندازه نمونه ۱۰۰ و ۱۰۰۰ مقادیر شاخص‌های انحراف تحت دو توزیع پیشین پارامتر ν تقریباً یکسان هستند. همین نتیجه با تحلیل مقادیر داده شده در جدول ۲ نیز حاصل می‌شود. در کل با افزایش اندازه نمونه مقادیر شاخص‌ها کاهش‌یافته است. نتایج نموداری برآورد منحنی‌های واقعی (چگالی، بقا و نرخ‌خطر) مدل I تحت پیشین‌های $G(۳, ۰/۰۵)$ و $G(۲, ۲)$ از پارامتر ν به ترتیب در شکل‌های ۲ و ۳ نشان داده شده است. به طور مشابه نتایج برای مدل II در شکل‌های ۴ و ۵ ارائه شده‌اند. در هر یک از شکل‌های مذکور نتایج مربوط به دو اندازه نمونه ۱۰۰ و ۱۰۰۰ به ترتیب در نمودارهای سمت راست و چپ نشان داده شده‌اند. در این نمودارها میانگین‌های پسینی [3] و کران‌های فاصله اعتبار ۹۵٪ بیزی[4] منحنی‌های واقعی در نقاط مختلف به ترتیب با رنگ‌های قرمز و مشکی بریده شده نشان داده شده‌اند. همچنین در این نمودارها منحنی‌های واقعی هر دو مدل I و II با رنگ آبی نمایش داده می‌شود. با مقایسه نمودارهای ارائه شده در دو سمت چپ و راست شکل‌های ۲، ۳، ۴ و ۵ با یکدیگر مشاهده می‌شود که با افزایش اندازه نمونه مقادیر برآوردها(میانگین‌های پسینی) به مقادیر واقعی منحنی‌های واقعی (چگالی، بقا و نرخ‌خطر) مدل‌های I و II نزدیک‌تر شده‌اند و همچنین طول فواصل اعتبار ۹۵٪ بیزی منحنی‌های واقعی نیز کاهش‌یافته است. در ادامه هدف مقایسه نتایج نموداری هر یک از دو مدل I و II نسبت به دو توزیع پیشین انتخاب شده برای پارامتر ν است. نتایج نموداری مدل I تحت دو توزیع پیشین $G(۳, ۰/۰۵)$ و $G(۲, ۲)$ پارامتر ν

[1] Mean Absolute Error
[2] Hellinger
[3] Posterior Mean
[4] Bayesian Credible Interval

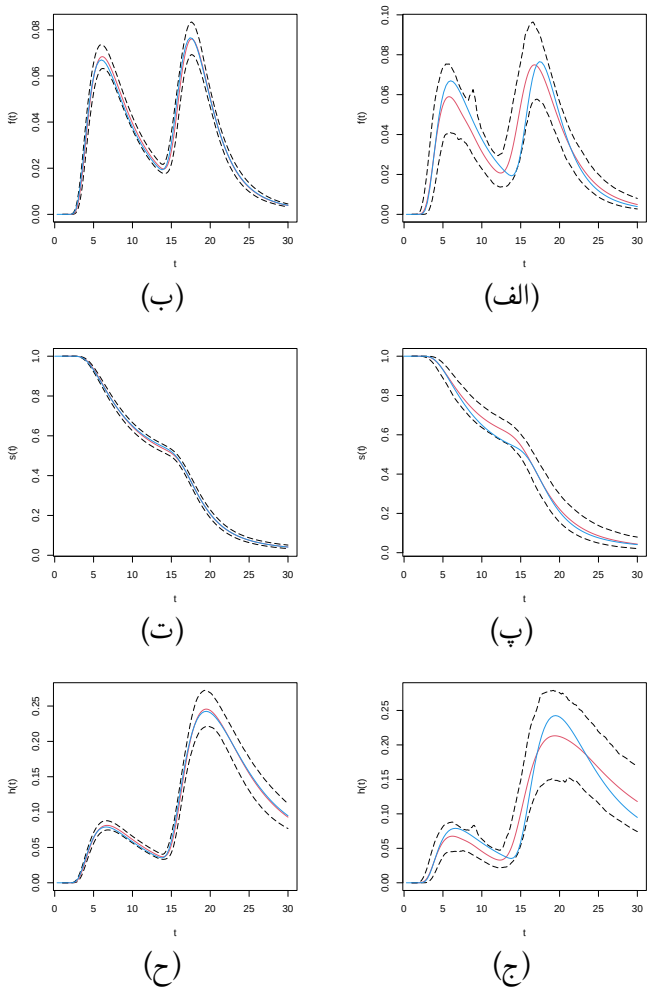
جدول ۱. نتایج محاسبه شاخص‌های انحراف در مدل I تحت دو اندازه نمونه ۱۰۰ و ۱۰۰۰ و دو توزیع پیشین پارامتر ν.

		$n = ۱۰۰$			$n = ۱۰۰۰$		
توزیع پیشین	تابع	d_E	d_{MAE}	d_H	d_E	d_{MAE}	d_H
$G(۲, ۲)$	چگالی	۰/۱۰۰۲	(۰/۰۰۴۹)	۰/۱۷۹۷	۰/۰۲۱۵	۰/۰۰۱۰	۰/۰۳۹۸
	بقا	۰/۳۲۶۹	۰/۰۱۶۴	۰/۱۷۴۴	۰/۰۴۹۹	۰/۰۰۲۲	۰/۰۲۳۲
	نرخ خطر	۰/۲۳۹۴	۰/۰۱۳۸	۰/۲۶۷۷	۰/۰۳۷۱	۰/۰۰۱۹	۰/۰۴۹۹
$G(۳, ۰/۰۵)$	چگالی	۰/۱۰۰۲	۰/۰۰۴۹	۰/۱۷۸۸	۰/۰۲۱۸	۰/۰۰۱۱	۰/۰۳۹۹
	بقا	۰/۳۳۶۷	۰/۰۱۷۱	۰/۱۸۰۶	۰/۰۵۱۱	۰/۰۰۲۳	۰/۰۲۳۹
	نرخ خطر	۰/۲۳۸۱	۰/۰۱۳۷	۰/۲۶۵۱	۰/۰۴۱۴	۰/۰۰۲۴	۰/۰۵۲۳

جدول ۲. نتایج محاسبه شاخص‌های انحراف در مدل II تحت دو اندازه نمونه ۱۰۰ و ۱۰۰۰ و دو توزیع پیشین پارامتر ν.

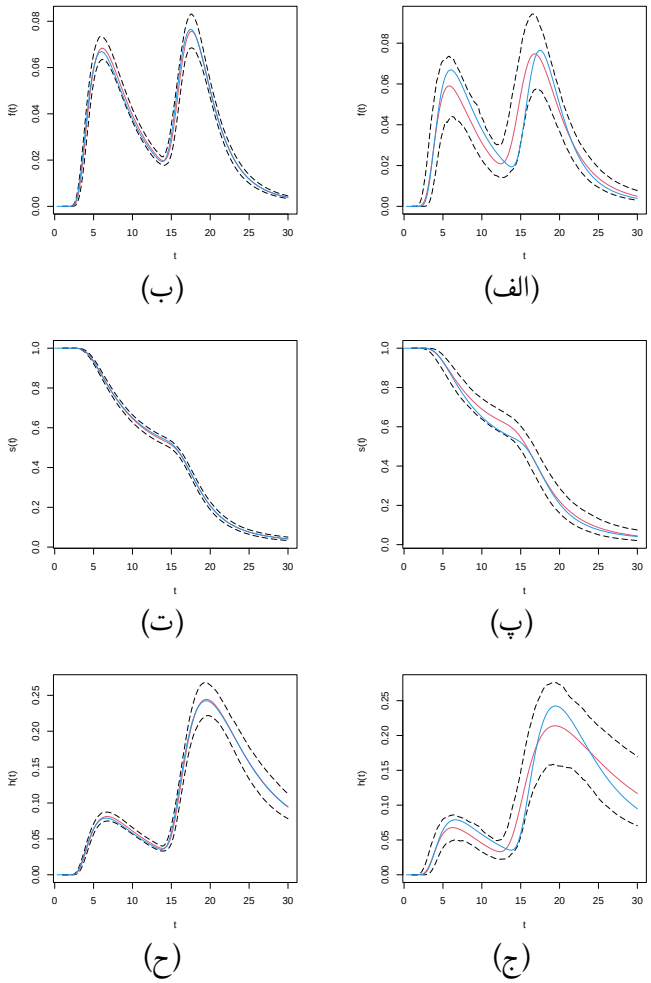
		$n = ۱۰۰$			$n = ۱۰۰۰$		
توزیع پیشین	تابع	d_E	d_{MAE}	d_H	d_E	d_{MAE}	d_H
$G(۲, ۲)$	چگالی	۰/۱۲۷۲	۰/۰۰۴۵	۰/۱۱۹۴	۰/۰۴۶۵	۰/۰۰۱۷	۰/۰۵۳۶
	بقا	۰/۱۱۹۴	۰/۰۰۶۴	۰/۰۹۹۵	۰/۰۶۵۳	۰/۰۰۳۲	۰/۰۴۴۲
	نرخ خطر	۰/۲۲۹۹	۰/۰۱۲۹	۰/۱۵۵۸	۰/۱۳۲۴	۰/۰۰۷۳	۰/۰۸۶۰
$G(۳, ۰/۰۵)$	چگالی	۰/۱۳۶۲	۰/۰۰۴۶	۰/۱۲۶۶	۰/۰۴۴۷	۰/۰۰۱۸	۰/۰۵۳۹
	بقا	۰/۱۱۷۰	۰/۰۰۶۲	۰/۰۹۷۱	۰/۰۶۳۳	۰/۰۰۳۰	۰/۰۴۳۴
	نرخ خطر	۰/۲۴۰۲	۰/۰۱۳۵	۰/۱۶۴۴	۰/۱۷۸۲	۰/۰۱۰۲	۰/۱۰۲۸

به ترتیب در شکل‌های ۲ و ۳ داده شده‌اند. با مقایسه نمودارهای در دو شکل اخیر مشاهده می‌شود که استنباط‌های پسینی تحت دو توزیع پیشین پارامتر ν تقریباً یکسان هستند. همین نتیجه در مقایسه شکل‌های ۴ و ۵ مربوط به مدل II نیز برقرار است و بنابراین استنباط‌های پسینی نسبت به توزیع پیشین انتخاب شده برای پارامتر ν نوعی عدم حساسیت را نشان می‌دهند. در جدول ۳ استنباط‌های پسینی تعداد خوشه‌های (n^*) مدل I تحت دو توزیع پیشین $\nu \sim G(۲, ۲)$ و $\nu \sim G(۳, ۰/۰۵)$ ارائه شده است. این استنباط‌ها شامل مقادیر احتمال‌های تجربی، میانه، میانگین، انحراف‌معیار و فواصل اعتبار ۹۵٪ پسینی n^* است که تحت اندازه نمونه‌های ۱۰۰، ۳۰۰، ۵۰۰ و ۱۰۰۰ بدست آمده‌اند. نتایج موجود در جدول اخیر بعد از کنار گذاشتن ۱۰۰۰۰ تکرار اولیه در الگوریتم گیبز حاصل شده‌اند. در جدول ۳ با توجه به مقادیر احتمال تجربی n^* که تحت توزیع پیشین $\nu \sim G(۲, ۲)$ بدست آمده‌اند می‌توان



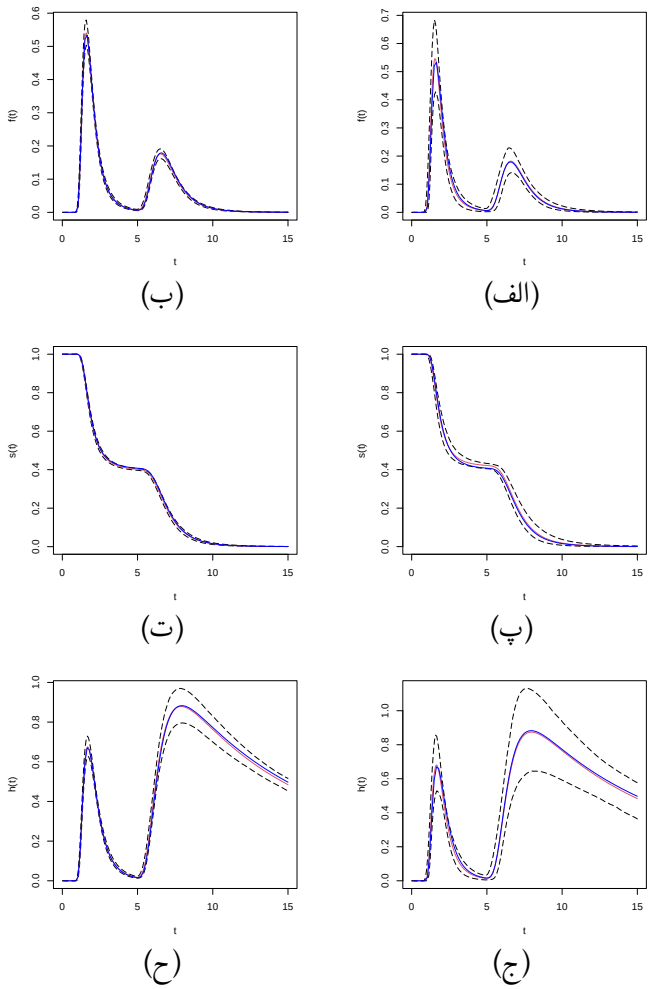
شکل ۲. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای(تیره رنگ) از توابع چگالی (سطر اول برای الف- $n = ۱۰۰$ و ب- $n = ۱۰۰۰$)، بقا (سطر دوم پ- $n = ۱۰۰$ و ت- $n = ۱۰۰۰$) و نرخ‌خطر (سطر سوم ج- $n = ۱۰۰$ و ح- $n = ۱۰۰۰$) مدل I در نقاط مختلف و تحت پیشین $G(۳, ۰/۰۵)$. منحنی‌های واقعی با رنگ آبی نشان داده شده است.

مشاهده کرد که بیشترین مقدار احتمال تجربی $n^* = ۲$ است. البته با افزایش اندازه نمونه مقدار احتمال مذکور کاهش و مقدار بقیه احتمال‌ها افزایش یافته است. اما در کل بیشترین احتمال تجربی مربوط به $n^* = ۲$ است و بنابراین برآورد ماکسیمم درستنماییn^* برابر با ۲ یعنی تعداد خوشه‌های واقعی داده‌ها است. نتایج مربوط به محاسبه احتمال تجربی مقادیر مختلف n^* که تحت توزیع پیشین $\nu \sim G(۳, ۰/۰۵)$ بدست آمده‌اند مشابه نتایج تحت توزیع



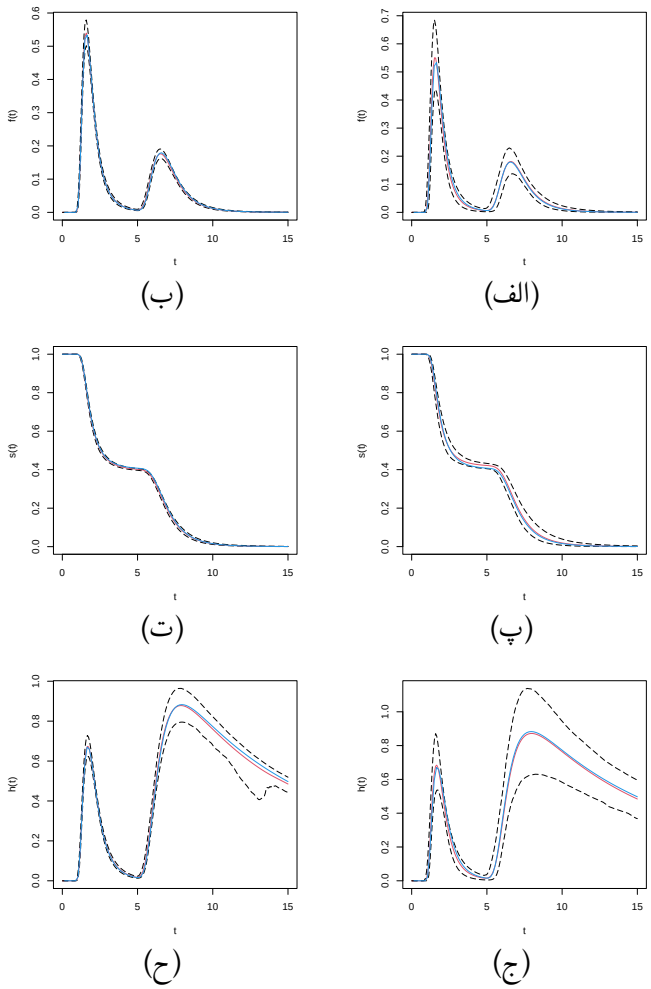
شکل ۳. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای(تیره رنگ) از توابع چگالی (منحنی‌های سطر اول برای الف- $n = ۱۰۰$ و ب- $n = ۱۰۰۰$)، بقا (سطر دوم پ- $n = ۱۰۰$ و ت- $n = ۱۰۰۰$) و نرخ‌خطر (سطر سوم ج- $n = ۱۰۰$ و ح- $n = ۱۰۰۰$) مدل I در نقاط مختلف و تحت پیشین $G(۲, ۲)$. منحنی‌های واقعی با رنگ آبی نشان داده شده است.

پیشین $\nu \sim G(۲, ۲)$ هستند اما در حالت کلی مقدار احتمال $P(n^* = ۲)$ تحت توزیع پیشین $\nu \sim G(۲, ۲)$ در مقایسه با توزیع پیشین $\nu \sim G(۳, ۰/۰۵)$ مقدار بزرگتری است ولی مقادیر بقیه احتمال‌های تجربی تحت توزیع پیشین $\nu \sim G(۳, ۰/۰۵)$ بزرگتر هستند. درکل تحت هر دو توزیع پیشین با افزایش اندازه نمونه مقادیر میانگین‌ها و انحراف معیارهای نمونه پسینی n^* افزایش یافته است اما مقادیر میانه‌های پسینی و فواصل اعتبار ۹۵٪ پسینی



شکل ۴. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای(تیره رنگ) از توابع چگالی (سطر اول برای الف- $n = ۱۰۰$ و ب- $n = ۱۰۰۰$)، بقا (سطر دوم پ- $n = ۱۰۰$ و ت- $n = ۱۰۰۰$) و نرخ‌خطر (سطر سوم ج- $n = ۱۰۰$ و ح- $n = ۱۰۰۰$) مدل II در نقاط مختلف و تحت پیشین $G(۳, ۰/۰۵)$. منحنی‌های واقعی با رنگ آبی نشان داده شده است.

n^* تقریباً ثابت هستند. در ادامه این بخش با استفاده از مدل پیشنهادی نتایج خوشه‌بندی داده‌های تولید شده از مدل I ارائه می‌شود. این نتایج تحت دو توزیع پیشین پارامتر ν و تحت اندازه نمونه‌های گوناگون در جدول ۴ داده شده است. نتایج مذکور بر مبنای خوشه‌بندی داده‌ها در آخرین تکرار الگوریتم نمونه‌گیری گیبز بدست آمده‌اند. تعداد خوشه‌های واقعی مدل I برابر ۲ است و بنابراین در فرایند شبیه‌سازی از این مدل هر داده تولید شده یا متعلق به



شکل ۵. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای(تیره رنگ) از توابع چگالی (سطر اول برای الف- $n = ۱۰۰$ و ب- $n = ۱۰۰۰$)، بقا (سطر دوم پ- $n = ۱۰۰$ و ت- $n = ۱۰۰۰$) و نرخ‌خطر (سطر سوم ج- $n = ۱۰۰$ و ح- $n = ۱۰۰۰$) مدل II در نقاط مختلف و تحت پیشین $G(۲, ۲)$. منحنی‌های واقعی با رنگ آبی نشان داده شده است.

خوشه ۱ و یا خوشه ۲ است. مدل پیشنهادی نیز هر داده را یا در خوشه واقعی خود قرار می‌دهد و یا در خوشه‌ای غیر از خوشه واقعی داده قرار می‌دهد. نتایج متناظر با توزیع پیشین $\nu \sim G(۳, ۰/۰۵)$ داخل پرانتز نشان داده شده است. در ادامه نتایج متناظر با $n = ۱۰۰$ و توزیع پیشین $\nu \sim G(۲, ۲)$ تحلیل می‌شوند. تحلیل نتایج به ازای توزیع پیشین $\nu \sim G(۳, ۰/۰۵)$ و بقیه اندازه نمونه‌ها مشابه است. در نمونه $n = ۱۰۰$ تایی تولید شده از مدل I

جدول ۳. استنباط‌های پسینی n^* تحت دو توزیع پیشین ν و اندازه نمونه‌های مختلف در مدل I.

پیشین ν	n	فاصله اعتبار ۹۵٪	انحراف معیار	میانگین	میانه	$P(n^*=۵)$	$P(n^*=۴)$	$P(n^*=۳)$	$P(n^*=۲)$
$G(۲,۲)$	۱۰۰	$[۲,۳]$	۰٫۱۹۸	۲٫۰۴۰	۲	۰٫۰۰۰	۰٫۰۰۱	۰٫۰۳۸	۰٫۹۶۱
	۳۰۰	$[۲,۳]$	۰٫۳۵۱	۲/۱۲۵	۲	۰٫۰۰۱	۰٫۰۰۶	۰٫۱۱۲	۰٫۸۸۱
	۵۰۰	$[۲,۳]$	۰٫۳۲۲	۲/۱۰۲	۲	۰٫۰۰۰	۰٫۰۰۵	۰٫۰۹۲	۰٫۹۰۳
	۱۰۰۰	$[۲,۴]$	۰٫۵۷۳	۲/۴۸۲	۲	۰٫۰۰۲	۰٫۰۳۴	۰٫۴۰۹	۰٫۵۵۵
$G(۳,۰٫۰۵)$	۱۰۰	$[۲,۳]$	۰٫۲۲۹	۲/۰۵۳	۲	۰٫۰۰۰	۰٫۰۰۱	۰٫۰۵۱	۰٫۹۴۸
	۳۰۰	$[۲,۳]$	۰٫۴۳۳	۲/۱۸۹	۲	۰٫۰۰۱	۰٫۰۱۳	۰٫۱۵۹	۰٫۸۲۷
	۵۰۰	$[۲,۳]$	۰٫۴۳۴	۲/۱۸۸	۲	۰٫۰۰۱	۰٫۰۱۵	۰٫۱۵۵	۰٫۸۲۹
	۱۰۰۰	$[۲,۴]$	۰٫۵۶۶	۲/۶۱۴	۳	۰٫۰۰۲	۰٫۰۳۶	۰٫۵۳۶	۰٫۴۲۶

تعداد داده‌های خوشه‌های اول و دوم به ترتیب ۵۵ و ۴۵ است که مدل پیشنهادی برای ۵۵ داده خوشه اول در ۴۵ مورد خوشه واقعی داده‌ها را به درستی خوشه ۱ پیش‌بینی کرده است و در ۱۰ مورد به اشتباه خوشه داده‌ها را ۲ پیش‌بینی کرده است. اما مدل پیشنهادی برای ۴۵ داده مربوط به خوشه دوم در ۳۸ مورد خوشه واقعی داده‌ها را به درستی پیش‌بینی کرده است و در ۷ مورد به اشتباه خوشه‌بندی را انجام است. مدل پیشنهادی برای۱۰۰ داده در ۸۳ مورد به درستی خوشه‌بندی را انجام داده است. بنابراین نرخ درست پیش‌بینی کردن خوشه داده‌ها برابر با ۸۳٪ است. مقدار مذکور تحت توزیع پیشین $\nu \sim G(۳, ۰٫۰۵)$ برابر با ۸۲٪ است. به همین ترتیب بقیه اعداد و ارقام جدول ۵ می‌تواند تحلیل شود.

جدول ۴. نتایج خوشه‌بندی داده‌های تولید شده از مدل I تحت دو توزیع پیشین ν و اندازه نمونه‌های مختلف.

پیشین ν	n	فاصله اعتبار ۹۵٪	انحراف معیار	میانگین	میانه	$P(n^*=۵)$	$P(n^*=۴)$	$P(n^*=۳)$	$P(n^*=۲)$
$G(۲,۲)$	۱۰۰	$[۲,۳]$	۰٫۱۹۸	۲/۰۴۰	۲	۰٫۰۰۰	۰٫۰۰۱	۰٫۰۳۸	۰٫۹۶۱
	۳۰۰	$[۲,۳]$	۰٫۳۵۱	۲/۱۲۵	۲	۰٫۰۰۱	۰٫۰۰۶	۰٫۱۱۲	۰٫۸۸۱
	۵۰۰	$[۲,۳]$	۰٫۳۲۲	۲/۱۰۲	۲	۰٫۰۰۰	۰٫۰۰۵	۰٫۰۹۲	۰٫۹۰۳
	۱۰۰۰	$[۲,۴]$	۰٫۵۷۳	۲/۴۸۲	۲	۰٫۰۰۲	۰٫۰۳۴	۰٫۴۰۹	۰٫۵۵۵
$G(۳,۰٫۰۵)$	۱۰۰	$[۲,۳]$	۰٫۲۲۹	۲/۰۵۳	۲	۰٫۰۰۰	۰٫۰۰۱	۰٫۰۵۱	۰٫۹۴۸
	۳۰۰	$[۲,۳]$	۰٫۴۳۳	۲/۱۸۹	۲	۰٫۰۰۱	۰٫۰۱۳	۰٫۱۵۹	۰٫۸۲۷
	۵۰۰	$[۲,۳]$	۰٫۴۳۴	۲/۱۸۸	۲	۰٫۰۰۱	۰٫۰۱۵	۰٫۱۵۵	۰٫۸۲۹
	۱۰۰۰	$[۲,۴]$	۰٫۵۶۶	۲/۶۱۴	۳	۰٫۰۰۲	۰٫۰۳۶	۰٫۵۳۶	۰٫۴۲۶

۵ تحلیل مجموعه داده‌های واقعی

در این بخش دو مجموعه داده واقعی تحلیل می‌شود. اولین مجموعه داده تحت عنوان acidity از مجموعه داده‌های بسته نرم‌افزاری BNPdensity در نرم‌افزار R است که مربوط به اندازه‌گیری میزان اسید در نمونه‌های از آب دریاچه‌ها است. در ادامه از این مجموعه داده به عنوان مجموعه داده اسید یاد می‌شود. دومین مجموعه داده تحت عنوان mouse از مجموعه داده‌های نرم‌افزار SPSS است که مربوط به زمان‌های بقای یک مجموعه از موش‌ها است

جدول ۵. نتایج خوشه‌بندی داده‌های تولید شده از مدل I تحت دو توزیع پیشین ν و اندازه نمونه‌های مختلف. مقادیر داخل پرانتز متناظر با پیشین $G(۳, ۰/۰۵)$ است.

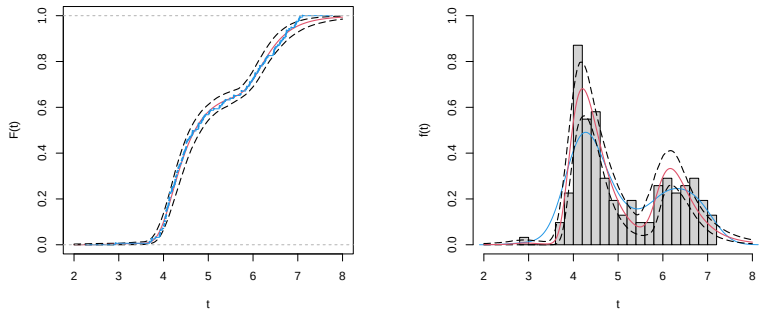
n	خوشه واقعی	۱	۲	درصد پیش‌بینی درست خوشه داده‌ها
۱۰۰	۱	۴۵(۴۳)	۱۰(۱۲)	۰/۸۳(۰/۸۲)
	۲	۷(۶)	۳۸(۳۹)	
۳۰۰	۱	۱۵۴(۱۵۰)	۱۵(۱۹)	۰/۸۶(۰/۸۷)
	۲	۲۷(۱۹)	۱۰۴(۱۱۲)	
۵۰۰	۱	۲۴۷(۲۳۳)	۳۵(۴۵)	۰/۸۱(۰/۸۲)
	۲	۵۹(۴۰)	۱۵۹(۱۷۸)	
۱۰۰۰	۱	۵۴۲(۵۲۴)	۷۸(۹۶)	۰/۸۴(۰/۸۳)
	۲	۸۶(۷۹)	۲۹۴(۳۰۱)	

که برخی از آنها سانسور شده از راست هستند. از این مجموعه داده به عنوان مجموعه داده بقای موش‌ها یاد می‌شود.

۵.۱ تحلیل مجموعه داده‌های اسید

داده‌های این مثال مربوط به اندازه‌گیری میزان اسید در نمونه‌های از آب ۱۵۵ دریاچه در ایالت ویسکانسین آمریکا است. در این مثال در فرایند اجرای الگوریتم نمونه‌گیری گیبز، تعداد کل تکرارها، تکرارهای دوره سوخت و فاصله تاخیر به ترتیب $N = ۵۰۰۰۰$، $B = ۱۰۰۰۰$ و ۲۰ در نظر گرفته شده‌اند. بنابراین حجم نمونه نهایی که در استنباط استفاده می‌شود برابر با ۲۰۰۰ است. در شکل ۶ برآوردهای بیزی (میانگین پسینی) و کران‌های فواصل اعتبار ۹۵٪ پسینی از توابع چگالی و توزیع تجمعی برای مجموعه داده‌های اسید به ترتیب با رنگ‌های قرمز و تیره بریده شده نشان داده شده‌اند. در نمودار سمت راست شکل ۶، بافت نگار داده‌ها به همراه استنباط‌های پسینی برای تابع چگالی رسم شده است. همچنین در این نمودار منحنی آبی رنگ نتایج برآورد چگالی داده‌ها با استفاده از دستور density در نرم‌افزار R را نشان می‌دهد. بافت نگار مجموعه داده‌های اسید دو قله‌ای (مُدی) است و بنابراین توزیع‌های پارامتری معمول که عمدتاً تک قله‌ای هستند نمی‌توانند توزیع چنین داده‌های را پوشش دهند، اما شکل مذکور به وضوح نشان می‌دهد که DPMM با هسته GIW توزیع این مجموعه داده را به خوبی پوشش داده‌است.

در نمودار سمت چپ شکل ۶ نتایج برآورد تابع توزیع تجمعی داده‌ها ارائه شده است. در این نمودار منحنی آبی رنگ نتایج برآورد تابع توزیع داده‌ها با روش تابع توزیع تجربی است. نتایج ارائه شده در این نمودار نشان می‌دهد که DPMM با هسته GIW و روش تابع توزیع تجربی تقریباً نتایج یکسانی ارائه داده‌اند. همچنین در این نمودار مشاهده می‌شود کران‌های فواصل اعتبار ۹۵٪ پسینی برآورد تابع توزیع تجمعی در نقاط مختلف خیلی نزدیک بهم هستند. مدل DPM با هسته GIW مجموعه داده‌های اسید را با احتمال‌های $P(n^* = ۲) = ۰/۰۰۰۶$، $P(n^* = ۳) = ۰/۹۹۷۳$ و $P(n^* = ۴) = ۰/۰۰۲۱$ به ترتیب در دو، سه و چهار خوشه قرار داده‌است. با توجه



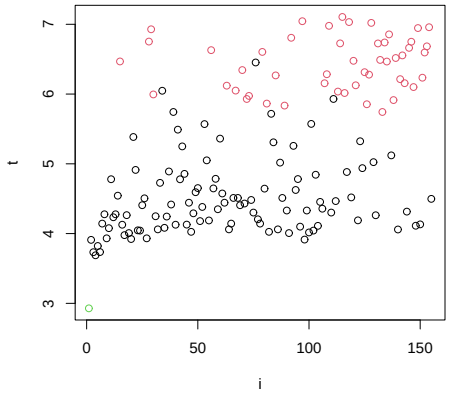
(ب) (الف)

شکل ۶. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای(تیره رنگ) از تابع الف– چگالی در نقاط مختلف به همراه بافت‌نگار مجموعه داده‌های اسید و ب–توزیع تجمعی به همراه نتایج روش تابع توزیع تجربی(رنگ آبی) برای مجموعه داده‌های اسید در نقاط مختلف با مدل پیشنهادی و پیشین $G(۲, ۲)$.

به مقادیر مذکور محتمل‌ترین مقدار (برآورد ماکسیمم درستنمایی) برای تعداد خوشه‌ها ۳ است یعنی مدل پیشنهادی داده‌ها را در سه خوشه قرار می‌دهد. در شکل ۷ نتایج خوشه‌بندی داده‌ها با استفاده ازمدل پیشنهادی نشان داده شده است. در این نمودار شماره داده‌ها با حرف لاتین i روی محور x ها نشان داده می‌شود. با توجه به نتایج ارائه شده در این نمودار مشاهده می‌شود که داده‌ها در سه خوشه قرار گرفته‌اند. داده‌های خوشه‌های اول، دوم و سوم به ترتیب با رنگ‌های مشکی، قرمز و سبز نشان داده شده‌اند. تعداد داده‌های سه خوشه مذکور به ترتیب به صورت ۱۰۵، ۴۹ و ۱ هستند که البته با توجه به نمودار خوشه‌بندی داده‌ها مشاهده می‌شود که در خوشه سوم تنها داده شماره ۱ هست.

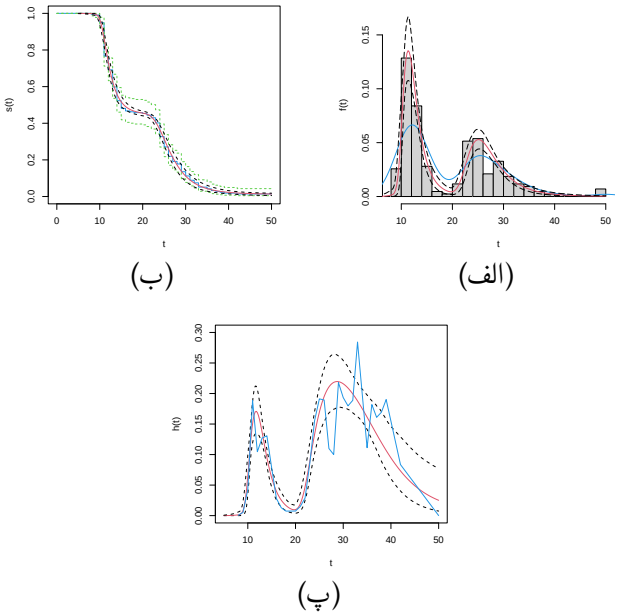
۵.۲ تحلیل مجموعه داده‌های بقای موش‌ها

در این زیربخش هدف این است که یکی از مجموعه داده‌های نرم‌افزار SPSS تحت عنوان مجموعه داده‌های بقای موش‌ها تحلیل شود. در این مجموعه داده اطلاعات متغیرهایی مانند زمان بقا، وضعیت سانسور شدن داده‌ها و غیره برای ۲۱۴ موش داده شده است که تنها ۳ تا از داده‌های بقا، سانسور شده از راست هستند. در این مثال در فرایند اجرای الگوریتم نمونه‌گیری گیبز، تعداد کل تکرارها، تکرارهای دوره سوخت و فاصله تاخیر به ترتیب $N = ۲۰۰۰۰$، $B = ۱۰۰۰۰$ و ۲۰ انتخاب شده‌اند. در شکل ۸ برآوردهای بیزی (میانگین پسینی) و همچنین فواصل اعتبار ۹۵٪ بیزی از توابع چگالی، بقا و نرخ خطر داده‌ها در نقاط مختلف زمانی به ترتیب با رنگ‌های قرمز و مشکی بریده شده نشان داده می‌شود. بافت‌نگار زمان بقای موش‌ها همراه با استنباط‌های پسینی برای توابع چگالی در سمت راست سطر بالای شکل ۸ نشان داده شده‌اند. همچنین در این نمودار منحنی آبی رنگ نتایج برآورد چگالی داده‌ها با استفاده از دستور density را نشان می‌دهد. این نمودارها به وضوح نشان می‌دهند که مدل پیشنهادی توزیع داده‌ها را به



شکل ۷. خوشه‌بندی مجموعه داده‌های اسید با استفاده از مدل پیشنهادی و پیشین $G(۲, ۲)$.

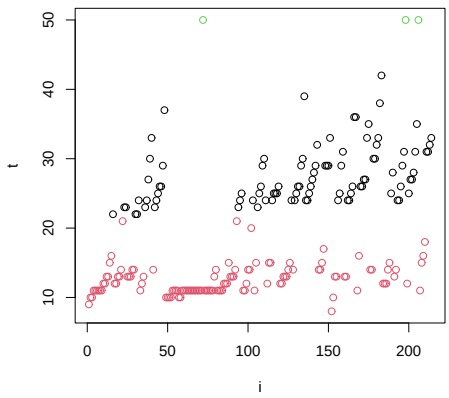
خوبی پوشش داده است. استفاده از روش‌های ناپارامتری مانند کاپلان–مه‌یر در تحلیل داده‌های بقا متداول است. در صورتی که حجم زیادی از داده‌های بقا سانسور نشده باشند مانند وضعیت موجود در این مثال، روش کاپلان–مه‌یر نتایج خوبی در برآورد تابع بقا ارائه می‌دهد (لاولس ، ۲۰۱۱). در نمودار سمت چپ سطر بالایی شکل ۸ علاوه بر استنباط‌های پسینی تابع بقا، مقادیر برآورد کاپلان–مه‌یر به همراه کران‌های فواصل اطمینان ۹۵٪ ناپارامتری تابع بقا ارائه شده‌اند. با توجه به این نمودار مشاهده می‌شود که در کل نتایج روش‌های پیشنهادی و کاپلان–مه‌یر در برآورد تابع بقا خیلی نزدیک بهم هستند. اگرچه طول فواصل اعتبار ۹۵٪ بیزی تابع بقا در مقایسه با طول فاصله اطمینان ۹۵٪ ناپارامتری خیلی کوتاه‌تر است. در سال‌های اخیر در نرم‌افزار R برای برآورد تابع نرخ‌خطر بسته‌های نرم‌افزاری مختلفی توسعه داده شده است که یکی از این بسته‌ها bshazard است. در این بسته نرم‌افزاری برای دستیابی به یک برآورد ناپارامتری از تابع نرخ‌خطر در چارچوب مدل‌های آمیخته خطی تعمیم‌یافته از B –اسپلاین‌ها استفاده می‌شود (ربورا و همکاران ، ۲۰۱۴). از آنجایی که بسته مذکور معمولاً نتایج خوبی را در برآورد تابع نرخ‌خطر ارائه می‌کند. بنابراین از این بسته نیز برای برآورد نرخ‌خطر داده‌ها استفاده شده است. استنباط‌های پسینی مربوط به برآورد تابع نرخ‌خطر که با DPMM با هسته GIW بدست آمده‌اند، در آخرین نمودار شکل ۸ ارائه شده‌اند. همچنین در این نمودار نتایج بسته نرم‌افزاری bshazard با رنگ آبی نشان داده شده‌اند. مقایسه نتایج این دو روش نشان می‌دهند که DPMM با هسته GIW نتایج بهتری فراهم کرده‌است. همچنین در این نمودار مشاهده می‌شود که طول فاصله اعتبار ۹۵٪ بیزی تابع نرخ‌خطر با افزایش زمان به سرعت افزایش می‌یابد. در شکل ۹ نتایج خوشه‌بندی داده‌های بقای موش‌ها با روش پیشنهادی آورده شده است. این نتایج بر اساس وضعیت خوشه‌بندی داده‌ها در آخرین تکرار الگوریتم گیبز بدست آمده‌اند. همانطور که ملاحظه می‌شود داده‌ها در سه خوشه قرار گرفته‌اند.



شکل ۸. منحنی برآوردهای نقطه‌ای (قرمز رنگ) و فاصله‌ای (تیره رنگ) از تابع الف- چگالی به همراه بافت‌نگار داده‌ها، ب-بقا به همراه برآوردهای کاپلان-مه‌یر(آبی رنگ) و برآوردهای فاصله‌ای ناپارامتری(سبز رنگ) از تابع بقا و پ-نرخ خطر به همراه نتایج بسته bshazard (آبی رنگ) در برآورد نرخ خطر برای مجموعه داده‌های بقای موش‌ها با مدل پیشنهادی و پیشین $G(۲, ۲)$.

۶ بحث و نتیجه‌گیری

در این مقاله برای تحلیل داده‌های بقای سانسور شده از راست، مدل آمیخته‌ای از فرایندهای دیریکله با هسته وارون وایبل تعمیم‌یافته بکار برده شده است. با تحلیل چندین مجموعه داده شبیه‌سازی شده و واقعی، عملکرد مدل پیشنهادی مورد بررسی قرار گرفت. نتایج داده شده در این مقاله نشان می‌دهد مدل پیشنهادی از پتانسیل خوبی برای برآورد توابع چگالی، بقا و نرخ‌خطر داده‌های بقا برخوردار است. از دیگر مزیت‌های مدل پیشنهادی می‌توان به پتانسل بالای آن برای خوشه‌بندی کردن داده‌ها اشاره کرد. عدم توانایی مدل پیشنهادی در بررسی تاثیر متغیرهای کمکی روی توزیع طول‌عمر از جمله نقایص این مدل است. بنابراین در پژوهش‌های آینده هدف نویسندگان این مقاله توسعه دادن مدل پیشنهادی به منظور بر طرف کردن نقیصه مذکور است.



شکل ۹. نتایج خوشه‌بندی داده‌های بقای موش‌ها با استفاده از مدل پیشنهادی و پیشین $G(۲, ۲)$.

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.

مراجع

Antoniak, C.E. (1974), Mixtures of Dirichlet Processes with Applications to Non-parametric 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.

Bohlouri Hajjar, S. and Khazaei, S. (2018), Bayesian Nonparametric Survival Analysis Using Mixture of Burr XII Distributions, *Communications in Statistics-Simulation and Computation*, **47**(9), 2724-2738.

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

De Gusmão, F. R., Ortega, E. M. and Cordeiro, G. M. (2011), The Generalized Inverse Weibull Distribution, *Statistical Papers*, **52**, 591-619.

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. (1974), Prior Distributions on Spaces of Probability Measures, *The Annals of Statistics* , **2**(4), 615-629.

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

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*, 1–19. https://doi.org/10.1080/03610918.2022.2076868

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

Kaur, K., Mahajan, K.K. and Arora, S. (2018), Bayesian and Semi-Bayesian Estimation of the Parameters of Generalized Inverse Weibull Distribution, *Journal of Modern Applied Statistical Methods*, **17**(1), 1–32.

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**.

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.

Kuo, L., Smith, A.F., MacEachern, S. and West, M. (1992), *Bayesian Computations in Survival Models via the Gibbs Sampler*, Springer Netherlands, 11–24.

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

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.

Rebora,P., Salim, A. and Reilly, M. (2014), Bshazard: A Flexible Tool for Nonparametric Smoothing of the Hazard Function, *The R Journal*, **6**(2), 114–122.

Ross, G.J. and Markwick, D. (2018), Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models.

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.