برآورد پارامترهای توزیع فوق‌هندسی تعمیم‌یافته تولید شده به‌وسیلهٔ فرایند تولد-مرگ

داود فربد*
دانشگاه صنعتی قوچان، گروه ریاضی

دریافت ۹۷/۱۲/۲۰ پذیرش ۹۸/۰۳/۲۷

چکیده

در این مقاله یک توزیع فوق‌هندسی تعمیم‌یافته که به‌کمک فرایند تولد-مرگ و برای مدل‌بندی داده‌های بیوانفورماتیک ساخته شده است را در نظر می‌گیریم. تحت برقراری بعضی شرایط، یک سیستم معادلات درست‌نمایی را به‌دست می‌آوریم که جواب حاصل از آن منطبق بر برآوردگرهای بیشینهٔ درست‌نمایی پارامترهای مورد نظر است. یک روش تقریبی همراه با بررسی شبیه‌سازی برای برآورد پارامترها ارائه می‌شود. هم‌چنین به‌منظور ارائه کاربردهای این توزیع، سه نوع داده واقعی در بیوانفورماتیک را با توزیع مورد نظر برازش داده و نتایج را با استفاده از شاخص‌های آماری با چهار توزیع گسسته دیگر مقایسه می‌کنیم که بر این اساس ملاحظه می‌شود توزیع فوق‌هندسی تعمیم‌یافته نسبت به چهار توزیع گسسته دیگر مدل مناسب‌تری است.

واژه‌های کلیدی: توزیع فوق‌هندسی تعمیم‌یافته، فرایند تولد-مرگ، بیوانفورماتیک، برآورد بیشینهٔ درست‌نمایی، روش مونت کارلوی زنجیر مارکفی (MCMC[1]).
کد موضوع بندی ریاضی (۲۰۱۰): ۶۲F۱۰، ۶۲F۳۰، ۶۲P۱۰، ۶۰J۲۸

مقدمه

توزیع‌های آماری مختلفی برای مدل‌بندی داده‌های بیوانفورماتیک معرفی شده‌اند. در واقع با توجه به پیچیدگی، تنوع و گستردگی داده‌های بیوانفورماتیک، نمی‌توان یک توزیع آماری واحد و یکسان برای مدل‌بندی این نوع داده‌ها ارائه کرد. از این رو توزیع‌های آماری متنوعی معرفی شده‌اند. دو روش اساسی برای ایجاد چنین توزیع‌هایی ارائه شده است [۱]:

الف) بر اساس مجموعه داده‌های واقعی و مشاهدات تجربی که می‌توان به توزیع توانی[2] یک‌پارامتری و توزیع شبه-پارتو دوپارامتری[3] اشاره کرد.

ب) بر اساس استفاده از روش‌های ریاضی که می‌توان به توزیع‌های مانای تولید شده به‌وسیلهٔ فرایند تولد-مرگ، گسسته‌سازی توابع چگالی توزیع‌های پایدار و استفاده از توابع ریاضی خاص اشاره کرد.

یکی ازمهم‌ترین روش‌های ایجاد توزیع‌های آماری مورد نیاز در بیوانفورماتیک، استفاده ازفرایند تولد-مرگ است که براین اساس تاکنون توزیع‌های آماری متعددی معرفی شده‌اند. برای‌مثال می‌توان به توزیع درحال رشد دوپارامتری[4]،

*نویسنده مسئول: d.farbod@qiet.ac.ir

1. Markov Chain Monte Carlo (MCMC)
2. Power-law distribution
3. Two-parameters Pareto-like distribution
4. Distribution with moderate growth

[DOR: 20.1001.1.25882546.1399.6.3.15.8]
[DOI: 10.52547/mmr.6.3.417]

توزیع ورینگ دوپارامتری[۱]، توزیع شبه-پارتو تعمیم‌یافته چهارپارامتری[۲] اشاره کرد. یکی از توزیع‌های آماری که به‌کمک فرایند تولد-مرگ و به‌عنوان تعمیمی از توزیع ورینگ دوپارامتری ارائه شد، توزیع فوق‌هندسی تعمیم‌یافته[۳] سه‌پارامتری است که به‌وسیلهٔ دانیلیان و آستولا (۲۰۰۶) معرفی شد [۴]. آستولا و همکاران (۲۰۰۷) [۱]، بعضی خواص آماری این توزیع سه‌پارامتری را بررسی کردند، اما با توجه به فرم پیچیده تابع جرم احتمال و تابع توزیع آن، تا کنون کاربردهای این توزیع و هم‌چنین برآورد پارامترهای آن بررسی نشده است. از این رو هدف اصلی این مقاله، برآورد پارامترهای توزیع فوق‌هندسی تعمیم‌یافته و هم‌چنین ارائه مثال‌های واقعی برای بیان کاربردهای این توزیع است.

۱. توزیع فوق‌هندسی تعمیم‌یافته تولید شده به‌وسیلهٔ فرایند تولد-مرگ

می‌دانیم در فرایند تولد-مرگ با ضرایب λ_{k-1} و μ_k شرط لازم و کافی برای وجود توزیع‌های مانا عبارت است از:

$$\sum_{x=1}^{\infty}\prod_{k=1}^{x}\frac{\lambda_{k-1}}{\mu_k}<\infty \qquad (۱)$$

اگر رابطهٔ (۱) برقرار باشد آن‌گاه برای $x\in\mathbb{N}$ داریم:

$$\begin{cases} p_x = p_0\prod_{k=1}^{x}\frac{\lambda_{k-1}}{\mu_k} \\ p_0 = \left(1+\sum_{y=1}^{\infty}\prod_{k=1}^{y}\frac{\lambda_{k-1}}{\mu_k}\right)^{-1} \end{cases} \qquad (۲)$$

فرض کنید ضرایب λ_{k-1} و μ_k بدین‌صورت باشد:

$$\lambda_{k-1}=(p_1+k)(p_2+k),\quad \mu_k=(1+k)(q+k) \qquad (۳)$$

با جای‌گذاری (۳) در (۲)، توزیع آماری (۴) که یک خانواده از توزیع‌های مانای تولید شده به‌وسیلهٔ فرایند تولد-مرگ با ضرایب (۳) است را به‌دست می‌آوریم. این توزیع به توزیع فوق‌هندسی تعمیم‌یافته تولید شده به‌وسیلهٔ فرایند تولد-مرگ معروف است. برای جزئیات بیش‌تر به [۱] و [۴] مراجعه شود.

توزیع فوق‌هندسی تعمیم‌یافته سه‌پارامتری (۴) به‌وسیلهٔ فرایند تولد-مرگ و برای مدل‌بندی داده‌های موجود در بیوانفورماتیک ارائه شد [۱]. تابع جرم احتمال آن بدین‌صورت است [۴]:

$$\begin{cases} p_x(\theta) = p_0(\theta)\prod_{k=0}^{x-1}\frac{(p_1+k)(p_2+k)}{(1+k)(q+k)} & x=1,2,3,\dots \\ p_0(\theta) = \left(1+\sum_{y=1}^{\infty}\prod_{k=0}^{y-1}\frac{(p_1+k)(p_2+k)}{(1+k)(q+k)}\right)^{-1} \end{cases} \qquad (۴)$$

که در آن $\theta=(p_1, p_2, q)$. $0<p_1<\infty$، $0<p_2<\infty$ و $0<q<\infty$ پارامترهای این توزیع هستند و هم‌چنین $q-p_1-p_2>0$.

مدل (۴) دارای خواص مهمی از قبیل *تک مدی بودن*[۴]، *محدب بودن*[۵] و *به‌طور منظم در حال تغییربودن در بی‌نهایت*[۶] است [۱]. به‌عبارت دیگر هر توزیع آماری که در سه خاصیت مذکورصدق کند، می‌تواند برای توصیف سیستم‌های بیوانفورماتیک استفاده شود. برای جزئیات بیش‌تر به [۱] و [۲] مراجعه شود.

یکی از مشکلات مدل (۴) عدم وجود فرم بسته برای آن است. از این‌رو، استنباط آماری برای پارامترهای آن به‌طور جدی بررسی نشده است. به‌عبارتی دانیلیان و آستولا (۲۰۰۶) [۴]، ضمن معرفی این توزیع، بررسی بعضی خواص

1. Two-parameters Waring distribution
2. Four-parameters generalized Pareto-like distribution
3. Three-parameters Generalized Hypergeometric distribution
4. Unimodality
5. Convexity
6. Regularly varying at infinity

آماری آن و هم‌چنین ارائه فرم‌هایی برای گشتاورهای آن [۳]، به‌طور اساسی به برآوردهای پارامترهای آن نپرداختند و مضافاً از جنبه‌های کاربردی نیز این توزیع بررسی نشد. ساختار مقاله حاضر در ۴ بخش بعدی به‌شرح ذیل خواهد بود. در بخش ۲ روش بیشینهٔ درست‌نمایی برای برآورد پارامترهای مدل (۴) بررسی می‌شود. در بخش ۳ و با توجه به ساختار پیچیده تابع جرم احتمال و تابع توزیع تجمعی، روشی تقریبی برای برآورد پارامترها ارائه می‌شود که در ادامه آن شبیه‌سازی نیز به‌کمک MCMC انجام می‌شود. در بخش ۴ به‌منظور ارائه کاربردهای این توزیع، سه نوع داده واقعی را با مدل (۴) برازش می‌دهیم. بخش ۵ به مقایسهٔ توزیع فوق‌هندسی تعمیم‌یافته (۴) با چهار توزیع آماری گسسته دیگر می‌پردازد که با استفاده از سه سری داده‌ی واقعی برتری این توزیع نسبت به چهار توزیع دیگر نشان داده می‌شود.

برآورد بیشینهٔ درست‌نمایی

در این بخش، تحت برقراری بعضی شرایط یک سیستم معادلات درست‌نمایی را به‌دست می‌آوریم که جواب حاصل از آن منطبق بر برآوردگرهای بیشینهٔ درست‌نمایی پارامترهای مورد نظر است. در واقع از حل بعضی معادلات گشتاوری برآوردگرهای بیشینهٔ درست‌نمایی به‌دست می‌آیند (دراین سیستم برآوردگرهای بیشینهٔ درست‌نمایی منطبق با بعضی برآوردگرهای گشتاوری هستند).

فرض کنید $X^n=(X_1,\dots,X_n)$ با مشاهدات $x^n=(x_1,\dots,x_n)$، یک نمونه متناظر با متغیر تصادفی گسسته ξ با توزیع آماری (۴) باشد (برای $x\in\mathbb{N}$).

قبل از بیان قضایای اصلی، لم ۱ را ارائه می‌کنیم.

لم ۱. برای مدل (۴) داریم:

$$E_\theta[t(\xi;\theta)]<\infty\,,\ E_\theta[s(\xi;\theta)]<\quad E_\theta[g(\xi;\theta)]<\infty, \tag{۵}$$

که $t(\xi;\theta)=\sum_{k=0}^{x-1}\frac{1}{p_1+k}$, $s(\xi;\theta)=\sum_{k=0}^{x-1}\frac{1}{p_2+k}$, $g(\xi;\theta)=\sum_{k=0}^{x-1}\frac{-1}{q+k}$.

اثبات. به‌کمک تعریف امید ریاضی و مشابه لم ۱ (فربد، ۲۰۱۵، ص. ۴۵، [۵]) اثبات رابطهٔ (۵) بدیهی است.

اکنون، از لم ۱ قضیهٔ ۲ را می‌توان بیان کرد.

قضیهٔ ۲. برآوردگر بیشینهٔ درست‌نمایی پارامتر $\theta=\left(p_1\ ،p_2\ ،q\right)$ به‌کمک معادلات گشتاوری (۶) به‌دست می‌آید:

$$\begin{cases}E_\theta[t(\xi;\theta)]=\overline{t^n(\theta)}\\ E_\theta[s(\xi;\theta)]=\overline{s^n(\theta)}\\ E_\theta[g(\xi;\theta)]=\overline{g^n(\theta)}\end{cases} \tag{۶}$$

که در آن $\overline{t^n(\theta)}=\frac{1}{n}\sum_{i=1}^n t(x_i;\theta)$ ، $\overline{s^n(\theta)}=\frac{1}{n}\sum_{i=1}^n s(x_i;\theta)$ ، $\overline{g^n(\theta)}=\frac{1}{n}\sum_{i=1}^n g(x_i;\theta)$.

اثبات. می‌دانیم که شرط لازم برای وجود برآوردگرهای بیشینهٔ درست‌نمایی عبارت‌است از:

$$\frac{\partial l(X^n;\theta)}{\partial\theta_i}=0,\qquad i=1,2,3,\qquad \theta_1=p_1,\ \theta_2=p_2,\ \theta_3=q.$$

که $l(X^n;\theta)$ لگاریتم تابع درست‌نمایی است.

لگاریتم تابع درست‌نمایی را بدین‌صورت به‌دست می‌آوریم:

$$\begin{aligned}l(X^n;\theta)&=\ln\prod_{i=1}^n\left(p_0(\theta)\cdot\prod_{k=0}^{x_i-1}\frac{(p_1+k)(p_2+k)}{(1+k)(q+k)}\right)\\ &=-n\ln p_0(\theta)+\sum_{i=1}^n\sum_{k=0}^{x_i-1}\ln\left(\frac{(p_1+k)(p_2+k)}{(1+k)(q+k)}\right),\end{aligned} \tag{۷}$$

با مشتق‌گیری از رابطهٔ (۷) نسبت به پارامتر p_1 داریم:

$$\frac{\partial l(X^n;\theta)}{\partial p_1} = -n\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_1} + \sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{1}{p_1+k},$$

که $\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_1} = E_\theta\left[\sum_{k=0}^{x_i-1}\frac{1}{p_1+k}\right] = E_\theta[t(\xi;\theta)]$. از رابطهٔ $\frac{\partial l(X^n;\theta)}{\partial p_1} = 0$ به‌راحتی می‌توان نتیجه گرفت:

$$E_\theta[t(\xi;\theta)] = \overline{t^n(\theta)}.$$

هم‌چنین، با مشتق‌گیری از رابطهٔ (۷) نسبت به پارامتر p_2 داریم:

$$\frac{\partial l(X^n;\theta)}{\partial p_2} = -n\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_2} + \sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{1}{p_2+k},$$

که $\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_2} = E_\theta\left[\sum_{k=0}^{x_i-1}\frac{1}{p_2+k}\right] = E_\theta[s(\xi;\theta)]$. از رابطهٔ $\frac{\partial l(X^n;\theta)}{\partial p_2} = 0$ می‌توان نتیجه گرفت:

$$E_\theta[s(\xi;\theta)] = \overline{s^n(\theta)}.$$

در نهایت، با مشتق‌گیری از رابطهٔ (۷) نسبت به پارامتر q داریم:

$$\frac{\partial l(X^n;\theta)}{\partial q} = -n\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial q} + \sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{-1}{q+k}.$$

که $\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial q} = E_\theta\left[\sum_{k=0}^{x_i-1}\frac{-1}{q+k}\right] = E_\theta[g(\xi;\theta)]$. در نتیجه از $\frac{\partial l(X^n;\theta)}{\partial q} = 0$ داریم:

$$E_\theta[g(\xi;\theta)] = \overline{g^n(\theta)}\,.$$

اثبات قضیهٔ ۲ کامل می‌شود.

اکنون، می‌خواهیم ثابت کنیم که جواب $\hat{\theta} = \hat{\theta}_i^n = \left(\hat{\theta}_i^n\right)_{i=1}^3$ از سیستم (۶) برآورد بیشینهٔ درست‌نمایی برای پارامتر θ است. از این رو کافی است نشان دهیم ماتریس

$$\hat{Q}_n = \begin{pmatrix} \hat{Q}_{1,1}^n & \hat{Q}_{1,2}^n & \hat{Q}_{1,3}^n \\ \hat{Q}_{2,1}^n & \hat{Q}_{2,2}^n & \hat{Q}_{2,3}^n \\ \hat{Q}_{3,1}^n & \hat{Q}_{3,2}^n & \hat{Q}_{3,3}^n \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 l(X^n,\theta)}{\partial p_1^2} & \frac{\partial^2 l(X^n,\theta)}{\partial p_1\partial p_2} & \frac{\partial^2 l(X^n,\theta)}{\partial p_1\partial q} \\ \frac{\partial^2 l(X^n,\theta)}{\partial p_2\partial p_1} & \frac{\partial^2 l(X^n,\theta)}{\partial p_2^2} & \frac{\partial^2 l(X^n,\theta)}{\partial p_2\partial q} \\ \frac{\partial^2 l(X^n,\theta)}{\partial q\partial p_1} & \frac{\partial^2 l(X^n,\theta)}{\partial q\partial p_2} & \frac{\partial^2 l(X^n,\theta)}{\partial q^2} \end{pmatrix}_{\theta=\hat{\theta}} \qquad (۸)$$

یک ماتریس معین نامنفی است ($\hat{Q}_{i,j}^n = Q_{i,j}^n(\hat{\theta}) = \frac{\partial^2 l(X^n,\theta)}{\partial\theta_i\partial\theta_j}$ $\mathrm{i,j} = 1,2,3$). بدین‌منظور قضیهٔ ۳ را ارائه می‌کنیم.

قضیهٔ ۳. فرض کنید جواب $\hat{\theta}$ از سیستم (۶) در این شرایط برقرار باشد:

$$\begin{cases} \mathrm{E}_\theta[\mathrm{h}(\xi;\theta)] = \overline{\mathrm{h^n}(\theta)} \\ \mathrm{E}_\theta[\mathrm{c}(\xi;\theta)] = \overline{\mathrm{c^n}(\theta)} \\ \mathrm{E}_\theta[\mathrm{m}(\xi;\theta)] = \overline{\mathrm{m^n}(\theta)} \end{cases} \qquad (۹)$$

که در آن $\mathrm{h}(\xi;\theta) = \sum_{k=0}^{x-1}\frac{-1}{(p_1+k)^2}$، $\mathrm{c}(\xi;\theta) = \sum_{k=0}^{x-1}\frac{-1}{(p_2+k)^2}$، $\mathrm{m}(\xi;\theta) = \sum_{k=0}^{x-1}\frac{1}{(q+k)^2}$. آن‌گاه داریم:

الف) اعضای ماتریس $\widehat{\mathrm{Q}}_\mathrm{n}$ بدین‌صورت است:

$$\widehat{\mathrm{Q}}_{11}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Var}_{\hat{\theta}}(\mathrm{t}(\xi;\theta)), \qquad \widehat{\mathrm{Q}}_{12}^{\mathrm{n}} = \widehat{\mathrm{Q}}_{21}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Cov}_{\hat{\theta}}(\mathrm{t}(\xi;\theta),\mathrm{s}(\xi;\theta)),$$
$$\widehat{\mathrm{Q}}_{13}^{\mathrm{n}} = \widehat{\mathrm{Q}}_{31}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Cov}_{\hat{\theta}}(\mathrm{t}(\xi;\theta),\mathrm{g}(\xi;\theta)), \qquad \widehat{\mathrm{Q}}_{23}^{\mathrm{n}} = \widehat{\mathrm{Q}}_{32}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Cov}_{\hat{\theta}}(\mathrm{s}(\xi;\theta),\mathrm{g}(\xi;\theta)),$$
$$\widehat{\mathrm{Q}}_{22}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Var}_{\hat{\theta}}(\mathrm{s}(\xi;\theta)). \qquad \widehat{\mathrm{Q}}_{33}^{\mathrm{n}} = -\mathrm{n}\,\mathrm{Var}_{\hat{\theta}}(\mathrm{g}(\xi;\theta)).$$

ب) تحت برقراری شرایط (۹)، ماتریس $\widehat{\mathrm{Q}}_{\mathrm{n}}$ یک ماتریس معین نامنفی است.

اثبات. الف) از فرمول (۸) مؤلفه‌های ماتریس $\widehat{\mathrm{Q}}_{\mathrm{n}}$ به‌ترتیب بدین‌صورت است:

$$Q_{11}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial p_1^2}=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial p_1^2}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_1}\right)^2\right)+n\sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{-1}{(p_1+k)^2}$$
$$=-\mathrm{n}\big(\mathrm{E}_\theta[h(\xi;\theta)]+\mathrm{E}_\theta[t(\xi;\theta)]^2-\mathrm{E}_\theta^2[t(\xi;\theta)]\big)+n\,\overline{h^n(\theta)}$$
$$=-n\,Var_\theta(t(\xi;\theta))-n\big(E_\theta[h(\xi;\theta)]-\overline{h^n(\theta)}\big),\qquad(۱۰)$$

$$Q_{12}^{n}(\theta)=Q_{21}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial p_1\,\partial p_2}=\frac{\partial^2 l(X^n;\theta)}{\partial p_2\,\partial p_1}$$
$$=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial p_1\,\partial p_2}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_1}\right)\cdot\left(\frac{1}{p_0(\theta)}\frac{\partial p_0(\theta)}{\partial p_2}\right)\right)$$
$$=-\mathrm{n}\,(\mathrm{E}_\theta[t(\xi;\theta)\cdot s(\xi;\theta)]-\mathrm{E}_\theta[t(\xi;\theta)]\cdot\mathrm{E}_\theta[s(\xi;\theta)])$$
$$=-n\,Cov_\theta(t(\xi;\theta),s(\xi;\theta)),\qquad(۱۱)$$

$$Q_{13}^{n}(\theta)=Q_{31}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial p_1\,\partial q}=\frac{\partial^2 l(X^n;\theta)}{\partial q\,\partial p_1}$$
$$=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial p_1\,\partial q}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_1}\right)\cdot\left(\frac{1}{p_0(\theta)}\frac{\partial p_0(\theta)}{\partial q}\right)\right)$$
$$=-\mathrm{n}\,(\mathrm{E}_\theta[t(\xi;\theta)\cdot g(\xi;\theta)]-\mathrm{E}_\theta[t(\xi;\theta)]\cdot\mathrm{E}_\theta[g(\xi;\theta)])$$
$$=-n\,Cov_\theta(t(\xi;\theta),g(\xi;\theta)),\qquad(۱۲)$$

$$Q_{22}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial p_2^2}=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial p_2^2}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_2}\right)^2\right)+n\sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{-1}{(p_2+k)^2}$$
$$=-\mathrm{n}\big(\mathrm{E}_\theta[c(\xi;\theta)]+\mathrm{E}_\theta[s(\xi;\theta)]^2-\mathrm{E}_\theta^2[s(\xi;\theta)]\big)+n\,\overline{c^n(\theta)}$$
$$=-n\,Var_\theta(s(\xi;\theta))-n\big(\,E_\theta[c(\xi;\theta)]-\,\overline{c^n(\theta)}\big),\qquad(۱۳)$$

$$Q_{23}^{n}(\theta)=Q_{32}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial p_2\,\partial q}=\frac{\partial^2 l(X^n;\theta)}{\partial q\,\partial p_2}$$
$$=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial p_2\,\partial q}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial p_2}\right)\cdot\left(\frac{1}{p_0(\theta)}\frac{\partial p_0(\theta)}{\partial q}\right)\right)$$
$$=-\mathrm{n}\,(\mathrm{E}_\theta[s(\xi;\theta)\cdot g(\xi;\theta)]-\mathrm{E}_\theta[s(\xi;\theta)]\cdot\mathrm{E}_\theta[g(\xi;\theta)])$$
$$=-n\,Cov_\theta(s(\xi;\theta),g(\xi;\theta)),\qquad(۱۴)$$

$$Q_{33}^{n}(\theta)=\frac{\partial^2 l(X^n;\theta)}{\partial q^2}=-n\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial^2 p_0(\theta)}{\partial q^2}-\left(\frac{1}{p_0(\theta)}\cdot\frac{\partial p_0(\theta)}{\partial q}\right)^2\right)+n\sum_{i=1}^{n}\sum_{k=0}^{x_i-1}\frac{1}{(q+k)^2}$$
$$=-\mathrm{n}\big(\mathrm{E}_\theta[m(\xi;\theta)]+\mathrm{E}_\theta[g(\xi;\theta)]^2-\mathrm{E}_\theta^2[g(\xi;\theta)]\big)+n\,\overline{m^n(\theta)}$$
$$=-n\,Var_\theta(g(\xi;\theta))-n\big(\,E_\theta[m(\xi;\theta)]-\,\overline{m^n(\theta)}\big)\qquad(۱۵)$$

اکنون با جای‌گذاری (۹) در روابط (۱۰)-(۱۵)، اثبات *قسمت الف* کامل می‌شود.

اثبات ب) برای اثبات *قسمت ب* کافی است نشان دهیم $\hat{Q}_{11}^{n}<0$ و $\det(\hat{Q}_n)_{i,j=1}^{2}>0$، $\det\big(\hat{Q}_n\big)<0$. از *قسمت الف* این قضیه، واضح است $\hat{Q}_{11}^{n}<0$. برای اثبات این‌که $\det(\hat{Q}_n)_{i,j=1}^{2}>0$ داریم:

$$\det(\hat{Q}_n)_{i,j=1}^{2}=\hat{Q}_{1,1}^{n}\cdot\hat{Q}_{2,2}^{n}-\big(\hat{Q}_{1,2}^{n}\big)^2$$
$$=n^2\left[Var_{\hat{\theta}}\big(t(\xi;\hat{\theta})\big)\cdot\,Var_{\hat{\theta}}\big(s(\xi;\hat{\theta})\big)-Cov_{\hat{\theta}}^{2}\big(t(\xi;\hat{\theta}),s(\xi;\hat{\theta})\big)\right]$$

که به‌کمک نامساوی کوشی- شوارتز اثبات کامل می‌شود.

هم‌چنین برای اثبات این‌که $\det\big(\hat{Q}_n\big)<0$ خواهیم داشت:

$$\det(\hat{Q}_n)=\begin{pmatrix}\hat{Q}_{1,1}^n & \hat{Q}_{1,2}^n & \hat{Q}_{1,3}^n\\ \hat{Q}_{2,1}^n & \hat{Q}_{2,2}^n & \hat{Q}_{2,3}^n\\ \hat{Q}_{3,1}^n & \hat{Q}_{3,2}^n & \hat{Q}_{3,3}^n\end{pmatrix}=-n\det(\hat{A}_3),\qquad \hat{A}_3=\left(\hat{a}_{i,j}\right)_{i,j=1}^3$$

که $\hat{a}_{i,j}=Cov_{\hat{\theta}}(\xi_i,\xi_j)$، $\xi_1=t(\xi;\theta),\quad \xi_2=s(\xi;\theta),\quad \xi_3=g(\xi;\theta)$. $\hat{A}_3$ یک ماتریس واریانس- کوواریانس است. با توجه به این‌که $\det(\hat{A}_3)\geq 0$ از این‌رو، $\det(\hat{Q}_n)\leq 0$. اما از طرفی ثابت‌هایی مانند $d_i(\theta)\neq 0\ \ i=1,2,3$ و $e(\theta)\in\mathbb{R}$ وجود ندارند که

$$\sum_{i=1}^3 d_i(\theta)\,\xi_i=e(\theta).$$

بنابراین با احتمال تقریباً مطمئن $\det(\hat{Q}_n)\neq 0$ که اثبات به پایان می‌رسد.

از قضیهٔ ۳، نتیجهٔ ۴ به‌دست می‌آید.

نتیجهٔ ۴. اگر جواب سیستم (۶) در شرایط (۹) صدق کند، آن‌گاه این جواب یک برآوردگر بیشینهٔ درست‌نمایی برای پارامتر θ است.

روش تقریبی برای برآورد بیشینهٔ درست‌نمایی

از آن‌جا که از حل معادلات (۶) نمی‌توان به‌راحتی فرم مشخصی برای برآوردهای بیشینهٔ درست‌نمایی پارامترهای مورد نظر به‌دست آورد، از این‌رو، برای حل مشکل یک روش تقریبی ارائه می‌کنیم. با مقایسهٔ با مقاله فربد و گاسپاریان (۲۰۱۳) [۶]، روش تجمعی اسکورینگ فیشر[1] به‌عنوان یک روش تقریبی پیشنهاد می‌شود (برای جزئیات بیش‌تر در ارتباط با این روش ایوچنکو و مدودف، ۱۹۹۰، ص. ۸۰ را ببینید).

فرض کنید $\theta(0)=(\mathrm{p}_1(0),\mathrm{p}_2(0),\mathrm{q}(0))$ مقادیر آغازین و سازگار از پارامترهای $\theta=(\mathrm{p}_1,\mathrm{p}_2,\mathrm{q})$ باشد. به‌طور بازگشتی، تقریب $(r+1)_{th}$ بدین‌صورت به‌دست می‌آید:

$$\theta_j(r+1)=\theta_j(r)+\frac{\mathrm{T_j}(\theta(\mathrm{r}))}{n\cdot\det I(\theta(r))},\qquad j=1,2,3;\quad r=0,1,2,\ldots, \tag{۱۶}$$

که در آن $I(\cdot)$ اندازهٔ اطلاع فیشر است. فرمول (۱۶) را به‌صورت ۱۷ می‌نویسیم:

$$\begin{cases}p_1(r+1)=p_1(r)+\dfrac{\mathrm{T_{p_1}}(\theta(\mathrm{r}))}{n\cdot\det I(\theta(r))},\\ p_2(r+1)=p_2(r)+\dfrac{\mathrm{T_{p_2}}(\theta(\mathrm{r}))}{n\cdot\det I(\theta(r))},\\ q(r+1)=q(r)+\dfrac{\mathrm{T_q}(\theta(\mathrm{r}))}{n\cdot\det I(\theta(r))}\end{cases} \tag{۱۷}$$

که در آن

$$\mathrm{T_{p_1}}(\theta(\mathrm{r}))=\begin{pmatrix}\mathrm{U}_1(\theta) & I_{12}(\theta) & I_{13}(\theta)\\ \mathrm{U}_2(\theta) & I_{22}(\theta) & I_{23}(\theta)\\ \mathrm{U}_3(\theta) & I_{32}(\theta) & I_{33}(\theta)\end{pmatrix},$$

$$\mathrm{T_{p_2}}(\theta(\mathrm{r}))=\begin{pmatrix}I_{11}(\theta) & \mathrm{U}_1(\theta) & I_{13}(\theta)\\ I_{21}(\theta) & \mathrm{U}_2(\theta) & I_{23}(\theta)\\ I_{31}(\theta) & \mathrm{U}_3(\theta) & I_{33}(\theta)\end{pmatrix},$$

$$\mathrm{T_q}(\theta(\mathrm{r}))=\begin{pmatrix}I_{11}(\theta) & I_{12}(\theta) & \mathrm{U}_1(\theta)\\ I_{21}(\theta) & I_{22}(\theta) & \mathrm{U}_2(\theta)\\ I_{31}(\theta) & I_{32}(\theta) & \mathrm{U}_3(\theta)\end{pmatrix}$$

1. Fisher's Scoring accumulation method

و $U_1(\theta)=\frac{\partial l(X^n;\theta)}{\partial p_1}$، $U_2(\theta)=\frac{\partial l(X^n;\theta)}{\partial p_2}$ و $U_3(\theta)=\frac{\partial l(X^n;\theta)}{\partial q}$ توابع کمکی هستند. پس از انجام بعضی محاسبات می‌توان نوشت:

$$U_1(\theta) = -n\,E_\theta[t(\xi;\theta)] + n\,\overline{t^n(\theta)},$$
$$U_2(\theta) = -n\,E_\theta[s(\xi;\theta)] + n\,\overline{s^n(\theta)},$$
$$U_2(\theta) = -n\,E_\theta[g(\xi;\theta)] + n\,\overline{g^n(\theta)}.$$

الگوریتم ۱ را برای مدل (۴) ارائه می‌کنیم:

الگوریتم ۱.

۱. اعداد تصادفی را بر اساس روش MCMC از مدل (۴) تولید می‌کنیم؛

۲. به‌منظور محاسبهٔ $\theta_j(r)$، $j=1,2,3;\ r=0,1,2,\ldots$، فرمول (۱۷) را استفاده می‌کنیم؛

۳. اگر $|\theta_j(r+1)-\theta_j(r)|<\varepsilon$ (ε بعضی ثابت کوچک مثبت است) آن‌گاه $\theta_j(r+1)=\hat{\theta}$ برآورد بیشینهٔ درست‌نمایی θ است، درغیر این‌صورت به گام دوم می‌رویم.

۱. بررسی شبیه‌سازی

از آن‌جا که تاکنون فرم‌های بسته و مشخصی برای تابع توزیع تجمعی مدل (۴) ارائه نشده است نمی‌توان به‌کمک تابع توزیع تجمعی آن، اعداد تصادفی را تولید کرد. بدین‌منظور روش MCMC پیشنهاد می‌شود. برای جزئیات بیش‌تر در ارتباط با روش MCMC [۱۱] را ببینید.

به‌منظور انجام محاسبات عددی، مدل (۴) را به‌صورت بریده شده در نظر می‌گیریم (با مقایسه با مقالاتی مانند: کوزنتسف، (۲۰۰۲، ص. ۳۹۹)، [۱۰]؛ فربد و گاسپاریان (۲۰۱۳)، [۶]. در قسمت شبیه‌سازی مقادیر x و y را تا ۱۰۰ در نظر می‌گیریم. بر اساس الگوریتم ۱ و نرم افزار آماری R نتایج عددی در جدول ۱ ارائه می‌شوند.

نکته ۱. مقادیر $\theta=(p_1=0.6,\ p_2=1.5,\ q=4.5)$ به‌عنوان مقادیر آغازین پارامترهای مدل (۴) در نظر گرفته می‌شود.

نکته ۲. بررسی شبیه‌سازی برای M=۱۰۰۰ مرتبه انجام می‌شود (M تعداد تکرارها است) تا رفتار برآوردگرهای بیشینهٔ درست‌نمایی را مشخص کنند. نمونه‌های N=۵۰ و N=۱۰۰ (N اندازه نمونه است) و هم‌چنین ε=۰/۰۰۰۵ در نظر گرفته می‌شود.

جدول ۱. میانگین برآوردها و خطای مربع میانگین برای مدل (۴)

	N=۵۰		N=۱۰۰	
	Mean	MSE	Mean	MSE
$p_1=$ ۰/۶	۱/۲۴۶۷۳۹	۰/۴۱۸۲۷۱	۰/۹۱۸۲۳۶	۰/۱۰۱۲۷۴
$p_2=$ ۱/۵	۲/۳۱۸۵۵۲	۰/۶۷۰۰۲۷	۱/۸۵۹۲۵۳	۰/۱۲۹۰۶۳
$q=$ ۴/۵	۵/۷۸۴۳۰۵	۱/۶۴۹۴۳۹	۴/۹۱۷۹۳۴	۰/۱۷۴۶۶۹
Iteration	۱۸۳	---	۱۱۴	---

میانگین برآوردها و خطای مربع میانگین[1] ارائه شده‌اند. از جدول ۱ می‌بینیم که با افزایش اندازهٔ نمونه، خطای مربع میانگین و تعداد تکرارها کاهش می‌یابند.

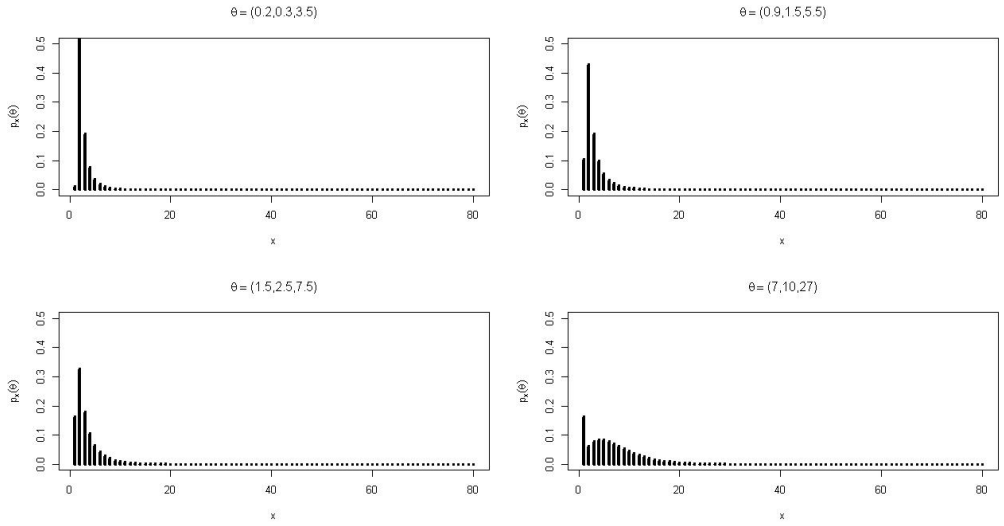
1. Mean Square Error (MSE)

مثال‌های کاربردی

در این بخش به‌کمک نرم‌افزار آماری R، ابتدا نمودار مدل (۴) برای مقادیر مختلف پارامترها رسم می‌شود، سپس به برازش این مدل با بعضی داده‌های واقعی در بیوانفورماتیک وعلوم زیستی پرداخته می‌شود. برای نشان دادن میزان دقت برازش مدل با داده‌های واقعی، به برازش توزیع فوق‌هندسی تعمیم‌یافته (۴) در مقابل مقادیر برآورد شده به‌روش بیشینهٔ درست‌نمایی می‌پردازیم.

۱. نمودار مدل

در شکل ۱، نمودار تابع جرم احتمال (۴) برای مقادیر مختلف پارامترهای p_1، p_2، q مشخص شده است.



شکل ۱. نمودار تابع جرم احتمال (۴) برای مقادیر مختلف $\boldsymbol{\theta = (p_1, p_2, q)}$

۲. تحلیل داده‌های واقعی

در این زیر بخش، به‌منظور بیان کاربردهای توزیع فوق‌هندسی تعمیم‌یافته (۴) سه مثال را ارائه می‌کنیم.

مثال ۱. تعداد بقایا در بعضی پروتئین‌ها[1] را در جدول ۲ در نظر می‌گیریم [۱۵].

جدول ۲

۱۰۷	۱۲۵	۱۷۲	۴۲	۱۸۴	۳۷۳	۳۵	۸۲	۵۹	۴۵	۶۲	۱۳۵	۱۷۶
۴۲	۲۱۷	۵۳	۱۰۶	۲۹۵	۹۵	۳۳۳	۱۵۲	۱۸۲	۸۴	۷۰	۱۶۳	۴۸۰

برای داده‌های جدول ۲، برآورد بیشینهٔ درست‌نمایی پارامترها و پی-مقدار[2] مدل (۴) عبارت‌است از:

$\hat{p}_1 = ۱۴/۹۷۱۶۸$ $\hat{p}_2 = ۱۵/۳۶۵۰۵$ $\hat{q} = ۳۲/۸۶۶۱۵$

و هم‌چنین داریم: $p - value = ۰/۹۹۴۸$ $-2\ln L = ۳۰۰/۸۳۲۶$

مثال ۲. تعداد چنبره[3] (سیم پیچ[4]) در ساختار دوم پروتئین را در جدول ۳ در نظر می‌گیریم [۱۳].

1. Number of residues in some proteins
2. p-value
3. Coil

٤. یک ناحیهٔ ساختار ثانویه که یک مارپیچ، ورق یا نوبه خود قابل تشخیص نیست، یک سیم پیچ نامیده می‌شود

جدول ۳

۱۰۸	۳۰۶	۸۵	۸۵	۱۰۳	۱۰۳	۱۱۲	۱۳۴	۸۲	۵۴
۹۸	۱۳۸	۵۴	۱۲۵	۹۹	۱۲۳	۱۶۴	۱۲۹	۱۴۲	۱۲۴
۴۲۸	۱۰۷	۱۵۳	۱۳۶	۲۸۷	۱۴۸	۱۵۳	۲۳۰	۲۴۶	۲۳۷
۱۶۲	۱۵۱	۴۶	۷۱	۶۲	۲۵۶	۵۶	۵۸	۱۰۷	۱۱۴

برای داده‌های جدول ۳، برآورد بیشینهٔ درست‌نمایی پارامترهای مدل (۴) عبارت‌است از:

$$\hat{p}_1 = ۲۴/۱۸۸۷۰ \qquad \hat{p}_2 = ۲۷/۵۳۱۱۰ \qquad \hat{q} = ۵۷/۳۲۴۱۷$$

و هم‌چنین داریم: $-2\ln L = ۴۴۲/۳۵۶۷ \qquad p-value = ۰/۷۵۷۷$

مثال ۳. تعداد باقی‌مانده‌ها در ۳۶ پروتئین کروی[1] را در جدول ۴ در نظر می‌گیریم [۷].

جدول ۴

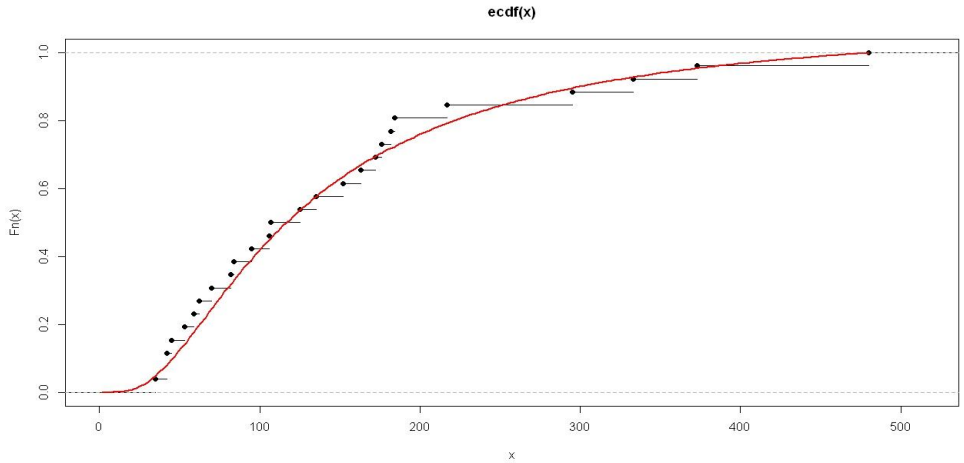
۱۸۲	۶۱	۲۷۴	۷۳	۷۷	۶۱	۱۷۰	۴۴	۶۷	۵۰	۲۳	۵۱
۱۰۲	۶۸	۳۹	۱۴۵	۱۲۵	۹۶	۴۵	۷۲	۱۵	۹۲	۵۰	۴۳
۱۱۷	۱۳۵	۶۶	۴۰	۲۵۵	۱۸۵	۲۷	۱۷۰	۸۱	۶۷	۱۴۲	۲۸۴

برای داده‌های جدول ۴، برآورد بیشینهٔ درست‌نمایی پارامترهای مدل (۴) عبارت‌است از:

$$\hat{p}_1 = ۱۲/۲۷۴۳۶ \qquad \hat{p}_2 = ۱۴/۳۲۳۰۲ \qquad \hat{q} = ۲۸/۷۳۰۳۲$$

و هم‌چنین داریم: $-2\ln L = ۳۸۶/۳۸۸۱ \qquad p-value = ۰/۹۴۹۳$

از مثال‌های ۱، ۲، ۳ و براساس مقادیر p- مقدار می‌بینیم که مدل (۴) به‌خوبی روی داده‌ها برازش شده است. به‌عنوان نمونه از سه مثال ارائه شده، نمودار شکل ۲ برای داده‌های جدول ۳ (مثال ۲) را رسم می‌کنیم. نقاط دنباله‌دار نشان‌دهندهٔ تابع توزیع تجربی داده‌ها و خط ممتد شکسته نشان دهنده برازش مدل (۴) است که با استفاده از آزمون کولموگوروف- اسمیرنوف این دو مقدار در مقابل هم‌دیگر آزمون شدند که نتایج آزمون نشان از نبود تفاوت معنی‌دار $p-value = ۰/۷۵۷۷$ است، یعنی مدل (۴) به‌خوبی روی داده‌ها برازش شده است.



شکل ۲. نقاط دنباله‌دار نشان‌دهندهٔ تابع توزیع تجربی داده‌ها و خط ممتد شکسته نشان‌دهندهٔ برازش مدل (۴) با داده‌های جدول ۳ است.

1. Number of residues in globular protein

مقایسهٔ با بعضی توزیع‌های آماری

در این بخش، نتایج ارائه شده برای داده‌های جدول‌های ۲، ۳، ۴ را با بعضی توزیع‌های آماری (توزیع توانی یک پارامتری، توزیع شبه-پارتو دوپارامتری، توزیع ورینگ دوپارامتری و توزیع شبه-پارتو تعمیم‌یافته چهارپارامتری) که قبلاً برای مدل‌بندی داده‌های بیوانفورماتیک معرفی شده‌اند، مقایسه می‌کنیم. خلاصه نتایج شامل p- مقدار، $\ln L$، معیار اطلاعاتی آکائیکه ($AIC = -2\ln L + 2a$)[1] و معیار اطلاعاتی آکائیکه تصحیح شده ($AICc = AIC + \frac{2a(a+1)}{n-a-1}$)[2] در جدول‌های ۵، ۶ و۷ ارائه می‌شوند (a نشان‌دهندهٔ تعداد پارامترهای مدل آماری است).

جدول ۵. نتایج برای داده‌های جدول ۲

مدل	ln L	AIC	AICc	p − value
توزیع توانی یک پارامتری	−۱۵۸/۴۷۸۲	۳۱۸/۹۵۶۴	۳۱۹/۱۱۰۲	۰/۰۲۹۰۱
توزیع شبه- پارتو دو پارامتری	− ۱۵۵/۳۷۲۵	۳۱۴/۷۴۵	۳۱۵/۲۲۵	۰/۲۷۰۷
توزیع ورینگ دوپارامتری	− ۱۵۶/۱۹۵۲	۳۱۶/۳۹۰۴	۳۱۶/۸۷۰۴	۰/۲۷۶۲
توزیع شبه-پارتو تعمیم‌یافته چهار پارامتری	−۱۵۰/۸۰۶۷	۳۰۹/۶۱۳۳	۳۱۱/۳۵۲۴	۰/۹۵۸۶
توزیع فوق‌هندسی تعمیم‌یافته سه پارامتری (۴)	− ۱۵۰/۴۱۶۳	۳۰۶/۸۳۲۶	۳۰۷/۸۳۲۶	۰/۹۹۴۸

جدول ۶. نتایج برای داده‌های جدول ۳

مدل	ln L	AIC	AICc	p − value
توزیع توانی یک پارامتری	− ۲۴۱/۰۸۶۵	۴۸۴/۱۷۳	۴۸۴/۲۷۳	۰/۰۰۰۱۰۶
توزیع شبه- پارتو دو پارامتری	− ۲۳۶/۶۲۹۱	۴۷۷/۲۵۸۱	۴۷۷/۵۶۵۸	۰/۰۱۵۹
توزیع ورینگ دوپارامتری	− ۲۳۷/۸۰۲۶	۴۷۹/۶۰۵۱	۴۷۹/۹۱۲۸	۰/۰۱۶
توزیع شبه-پارتو تعمیم‌یافته چهارپارامتری	− ۲۲۱/۹۸۲۹	۴۵۱/۹۶۵۸	۴۵۳/۰۴۶۹	۰/۵۸۶۸
توزیع فوق‌هندسی تعمیم‌یافته سه‌پارامتری (۴)	−۲۲۱/۱۷۸۴	۴۴۸/۳۵۶۷	۴۴۸/۹۸۸۲	۰/۷۵۷۷

جدول ۷. نتایج برای داده‌های جدول ۴

مدل	ln L	AIC	AICc	p − value
توزیع توانی یک پارامتری	−۲۰۲/۰۰۳۹	۴۰۶/۰۰۷۸	۴۰۶/۱۱۸۹	۰/۰۹۰۵
توزیع شبه- پارتو دو پارامتری	−۱۹۹/۲۰۷۶۵	۴۰۲/۴۱۵۳	۴۰۲/۷۵۸۱۶	۰/۱۳۹۸
توزیع ورینگ دوپارامتری	− ۱۹۹/۸۱۷۳	۴۰۳/۶۳۴۶	۴۰۳/۹۷۷۴۶	۰/۱۵۶۱
توزیع شبه-پارتوتعمیم‌یافته چهارپارامتری	− ۱۹۳/۷۶۴۵	۳۹۵/۵۲۹	۳۹۶/۷۴۱۱۲	۰/۸۲۲۳
توزیع فوق‌هندسی تعمیم‌یافته سه‌پارامتری (۴)	−۱۹۳/۱۹۴۱	۳۹۲/۳۸۸۱	۳۹۳/۰۹۳۹۸	۰/۹۴۹۳

در جدول‌های ۵، ۶ و ۷ بعضی شاخص‌ها از قبیل p- مقدار، $\ln L$، AIC و AICc محاسبه شده‌اند. چنان‌که ملاحظه می‌شود توزیع فوق‌هندسی تعمیم‌یافته سه‌پارامتری (۴) نسبت به چهار توزیع ذیل، مدل بهتری برای برازش این نوع داده‌ها می‌باشد. فرمول‌های تابع‌های جرم احتمال این چهار توزیع عبارت‌است از:

تابع جرم احتمال توزیع توانی یک‌پارامتری (رژتسکی و گومز، ۲۰۰۱، ص. ۹۸۹، فرمول (۱)):

$$p_x(\rho) = \frac{x^{-\rho}}{\sum_{y=1}^{\infty} y^{-\rho}} \qquad x = 1,2,\ldots, \quad \rho > 1$$

1. Akaike Information Criterion (AIC)
2. AIC with correction (AICc)

تابع جرم احتمال توزیع شبه‌پارتو دوپارامتری (کوزنتسف، ۲۰۰۱، ص. ۲۸۷، فرمول (۲) [۸]):

$$p_x(\theta) = \frac{(x+b)^{-\rho}}{\sum_{y=1}^{\infty}(y+b)^{-\rho}} \qquad x = 1,2,\dots, \quad \theta = (\rho, b),\ \rho > 1,\ b > -1$$

تابع جرم احتمال توزیع ورینگ دوپارامتری (فربد، ۲۰۱۵، ص. ۴۴، فرمول (۱) [۵]):

$$\begin{cases} p_x(\theta) = p_0(\theta)\prod_{k=0}^{x-1}\frac{(p+k)}{(q+k)} & x = 1,2,\dots, \quad \theta = (p,q),\ p > 0,\ q > 0 \\ p_0(\theta) = \left(1 + \sum_{y=1}^{\infty}\prod_{k=0}^{y-1}\frac{(p+k)}{(q+k)}\right)^{-1} \end{cases}$$

تابع جرم احتمال توزیع شبه-پارتو تعمیم‌یافته چهارپارامتری (فربد و گاسپاریان، ۲۰۱۳، ص. ۲۱۳، فرمول (۳) [۶]):

$$\begin{cases} p_x(\alpha) = p_0(\alpha)\frac{\theta^x}{(x+b)^{\rho}}\prod_{k=0}^{x-1}\left(1 + \frac{c-1}{(k+b)^{\rho}}\right) & x = 1,2,\dots, \\ p_0(\alpha) = \left(1 + \sum_{y=1}^{\infty}\frac{\theta^y}{(y+b)^{\rho}}\prod_{k=0}^{y-1}\left(1 + \frac{c-1}{(k+b)^{\rho}}\right)\right)^{-1} \end{cases}$$

که در آن $\alpha = (\theta, c, b, \rho)$ و $0 < \theta < 1, c > 0, b > 0, \rho > 1$.

تقدیر و تشکر

از داوران محترم مجله که نظرات ارزنده آن‌ها موجب بهبود مقاله شد قدردانی می‌شود.

منابع

1. Astola J., Danielian E., "Frequency Distributions in Biomolecular Systems and Growing Networks, Tampere International Center for Signal Processing (TICSP)", Series no. 31, Tampere, Finland (2007).
2. Astola J., Danielian E., Arzumanyan S., "Frequency distributions in bioinformatics", A Review, Proceedings Yerevan State University: Phys. Math. Sci., 223(3), (2010) 3-22.
3. Astola J., Gasparian K., Danielian E., "Moments estimators for hypergeometric distributions, Proceedings of the TISCP Workshop on Spectral Methods and Multirate Signal Processing (SMMSP) | Moscow", Russia, (2007) 233-234.
4. Danielian E., Astola J., "On regularly varying hypergeometric distributions", In Astola et al. (eds.), Proceedings International TICSP Workshop on Spectral Methods and Multirate Signal Processing, Florence, Italy, 2-3 Sept. 2006. TICSP Series no. 34, (2006) 127-132.
5. Farbod D., "On the parameters estimators for two frequency distributions arising in bioinformatics", Bulletin of the Georgian National Academy of Sciences, 9(1), (2015) 44-50.
6. Farbod D., Gasparian K., "Maximum likelihood estimators for some generalized Pareto-like frequency distribution", Journal of the Iranian Statistical Society (JIRSS), 12(2) (2013) 211-233.
7. Kabsch W., Sander C., "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", Biopolymers, 22(12) (1983) 2577-2637.
8. Kuznetsov V. A., "Distributions associated with stochastic processes of gene expression in a single eukaryotic cell", EURASIP Journal on Applied Signal Processing, 4, (2001) 258-296.

9. Kuznetsov V. A., "Family of skewed distributions associated with the gene expression and proteome evolution", Signal Processing, 33(4) (2003) 889-910.
10. Kuznetsov V. A., Pickalov V. A., Senko O. V., Knott G. D., "Analysis of the evolving proteomes: Predictions of the number for protein domains in nature and the number of genes in eukaryotic organisms", Journal of Biological Systems, 10(4) (2002) 381-407.
11. Givens G. H., Hoeting J. A., "Computational Statistics, Wiley and Sons" (2005).
12. Ivchenko G. I., Medvedev Yu., "Mathematical Statistics, Mir Press", Moscow (1990), translated from original Russian edition.
13. Qian N., Sejnowski T. J., "Predicting the secondary structure of globular proteins using neural network models", Journal of Molecular Biology, 202, (1988) 865-884.
14. Rzhetsky A., Gomez Sh. M., "Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome", Bioinformatics, 17(10) (2001) 988-996.
15. Yang J., "Protein secondary structure prediction based on neural network models and support vector machines", CS229 Final Project, Dec (2008).