مجله‌ی بررسی‌های آمار رسمی ایران
سال۲۴، شماره‌ی ۲، پاییز و زمستان ۱۳۹۲، صص ۱۴۵-۱۶۵

چهار روش براورد پارامترهای یك مدل آمیخته‌ی گاوسی

دنیا رحمانی و عادل محمدپور*

دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)

چکیده: مدل آمیخته‌ی گاوسی پرکاربردترین مدل آمیخته‌ی متناهی است. خاصیت مهم این مدل انعطاف‌پذیری آن نسبت به توزیع‌های پیوسته با اشکال گوناگون است. از آن‌جا که مهم‌ترین بخش برازش یک مدل، براورد پارامترهای آن می‌باشد، در این جا برآنیم تا پارامترهای مدل آمیخته‌ی گاوسی دومؤلفه‌ای را از طریق چهار روش براورد کنیم. ابتدا مدل آمیخته‌ی گاوسی را در حالت دومؤلفه‌ای بیان می‌کنیم، سپس پارامترهای مدل را از دو روش گشتاوری و ماکسیمم درستنمایی با عنوان حل تحلیلی و عددی براورد می‌کنیم. در ادامه براورد پارامترها را با استفاده از الگوریتم EM به دست آورده و در انتها نیز از الگوریتم نمونه‌گیر گیبز برای یافتن براوردها استفاده کرده‌ایم. در بخش نتیجه‌گیری، نتایج به دست آمده از روش‌ها را با یکدیگر مقایسه می‌کنیم. سعی ما بر این است که یک مسئله‌ی براورد را با چهار روش مرسوم حل کرده و برتری‌ها و محدودیت‌های هر یک را برای کاربران مشخص کنیم.

واژگان کلیدی: حل تحلیلی و عددی؛ الگوریتم EM؛ نمونه‌گیر گیبز؛ مدل آمیخته‌ی گاوسی.

۱- مقدمه

مدل آمیخته‌ی متناهی اولین بار در قرن نوزدهم وارد ادبیات آمار شد [۹]. از جمله مدل‌های آمیخته‌ی متناهی می‌توان به مدل آمیخته‌ی پواسن برای گروه‌بندی اسناد در امر بازیابی اطلاعات و مدل آمیخته‌ی فیشر برای تحلیل متون و آزمایش‌های ژنی اشاره کرد. مشهورترین مدل آمیخته، مدل آمیخته‌ی گاوسی می‌باشد [۴] و [۱۳]. مدل‌های آمیخته‌ی

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۰/۸/۹، پذیرش: ۱۳۹۳/۹/۳.

متناهی در خوشه‌بندی، براورد تابع چگالی، تحلیل مؤلفه‌ای، تحلیل تصاویر و دیگر موارد کاربرد بسزایی دارد. یک مدل آمیخته‌ی متناهی به‌صورت زیر تعریف می‌شود:
فرض کنید $x_1, \dots, x_n$ مشاهداتی از نمونه‌ی تصادفی مستقل و هم‌توزیع $X_1, \dots, X_n$ با بردار پارامتر $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ باشند، آنگاه تابع چگالی آمیخته‌ی مشاهدات به‌صورت زیر خواهد بود:

$$f(x_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k f_k(x_i; \theta_k), \qquad i = 1, \dots, n \tag{1}$$

$f(x_i; \boldsymbol{\theta})$ را تابع چگالی آمیخته‌ی Kمؤلفه‌ای نیز گویند. منظور از مؤلفه همان زیرجامعه‌های تشکیل‌دهنده‌ی جامعه می‌باشد که تعدادشان را با K نشان می‌دهند. θ_k پارامتر مربوط به زیر جامعه‌ی k می‌باشد و α_k ضریب وزنی یا ضریب آمیخته‌ی مؤلفه‌ی kاُم است که در شرایط زیر صدق می‌کند:

$$\sum_{k=1}^{K} \alpha_k = 1 \qquad 0 \leq \alpha_k \leq 1.$$

$f_k(x_i; \theta_k)$ چگالی مؤلفه‌ی kاُم با پارامتر θ_k می‌باشد [۸].

۲- مدل آمیخته‌ی گاوسی

مدل آمیخته‌ی گاوسی برای مشاهدات مستقل و هم‌توزیع $x_1, \dots, x_n$ مجموع وزن‌دار K مؤلفه، با تابع چگالی گاوسی است، که با معادله‌ی زیر نشان داده می‌شود:

$$f(x_i; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} \alpha_k \phi(x_i; \mu_k, \sigma_k^2)$$

در این‌جا $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ و $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)$ بردار پارامترهای مدل آمیخته‌ی (۱) می‌باشند. $\phi(x_i; \mu_k, \sigma_k^2)$.تابع چگالی گاوسی با پارامترهای μ_k و σ_k^2 مربوط به مولفه‌ی kاُم می‌باشد که از رابطه‌ی زیر به دست می‌آید:

$$\phi(x_i;\mu_k,\sigma_k^{۲})=\frac{۱}{\sqrt{۲\pi\sigma_k^{۲}}}\exp\left[-\frac{(x_i-\mu_k)^{۲}}{۲\sigma_k^{۲}}\right].$$

مثال ۱: برای مشاهدات $x_۱,\ldots,x_n$، که ترکیبی از دو زیرجامعه‌ی گاوسی به‌ترتیب با میانگین‌های $\mu_۱$، $\mu_۲$ و واریانس‌های $\sigma_۱^{۲}$، $\sigma_۲^{۲}$ باشند، آنگاه تابع چگالی آمیخته‌ی آن‌ها با ضرایب وزنی $\alpha_۱=\alpha$ و $\alpha_۲=۱-\alpha$ به‌صورت زیر خواهد بود:

$$\sum_{k=۱}^{۲}\alpha_k\phi\left(x_i;\mu_k,\sigma_k^{۲}\right)=\frac{\alpha}{\sqrt{۲\pi\sigma_۱^{۲}}}\exp\left[-\frac{(x_i-\mu_۱)^{۲}}{۲\sigma_۱^{۲}}\right]+\frac{۱-\alpha}{\sqrt{۲\pi\sigma_۲^{۲}}}\exp\left[-\frac{(x_i-\mu_۲)^{۲}}{۲\sigma_۲^{۲}}\right]$$

می‌توان مشاهداتی از مدل آمیخته‌ی گاوسی دومؤلفه‌ای را بر اساس دستور ارایه‌شده در بخش پیوست در نرم‌افزار R تولید کرد. مقادیر در نظر گرفته‌شده برای پارامترها بسته به نظر کاربر انتخاب می‌شود.

همان‌طور که می‌دانیم مهم‌ترین مسئله در ارتباط با یک مدل، براورد پارامترهای آن می‌باشد. از این رو در بخش ۳ این مقاله پارامترهای مدل آمیخته‌ی گاوسی دومؤلفه‌ای را از دو روش گشتاوری و ماکسیمم درستنمایی به‌عنوان روشی تحلیلی بیان کرده، سپس از روش عددی استفاده می‌کنیم. در ادامه نیز از الگوریتم EM و الگوریتم نمونه‌گیر گیبز برای یافتن براورد پارامترها استفاده کرده‌ایم.

۳- براورد پارامترهای مدل آمیخته‌ی گاوسی به روش تحلیلی

در این بخش ابتدا روش گشتاوری و سپس روش ماکسیمم درستنمایی، را برای براورد پارامترهای مدل آمیخته‌ی گاوسی به کار می‌بریم. در ادامه براورد ماکسیمم درستنمایی به روش عددی و الگوریتم EM را محاسبه کرده و با براورد بیز که از الگوریتم گیبز به دست آمده، مقایسه می‌کنیم.

۱-۳- روش گشتاوری

پی‌یرسون در سال ۱۸۹۴ برای برازش مدل به داده‌هایی که نسبت طول پیشانی به طول بدن ۱۰۰۰ خرچنگ را بیان می‌داشتند و توسط ولدن در سال ۱۸۹۲ جمع‌آوری شده بودند، از مدل آمیخته‌ی گاوسی تک‌متغیره استفاده کرد، [۹]. او داده‌ها را به ۲۹ بازه با فراوانی y_i برای $i = ۱, \dots, ۲۹$، تقسیم‌بندی کرد وبرای مدل‌بندی آن‌ها دو تابع چگالی نرمال با میانگین‌های $\mu_۱$ و $\mu_۲$ و واریانس‌های $\sigma_۱^۲$ و $\sigma_۲^۲$ را با ضرایب وزنی α و $۱ - \alpha$ با یکدیگر آمیخت. او برای براورد پارامترهای مدل از روش گشتاوری استفاده کرد. پی‌یرسون، در ابتدا گشتاورهای مرکزی مرتبه‌ی اول تا پنجم مشاهدات $\tilde{\mu}_i\ , i = ۱, \dots, ۵$ را به دست آورد. سپس با استفاده از روش گشتاوری به ۵ معادله برای براورد ۵ پارامتر مجهول رسید. با حل دستگاه معادلات به معادله‌ی درجه‌ی ۹ زیر دست یافت:

$$۲۴p_۲^۹ - ۲۸\lambda_۴ p_۲^۷ + ۳۶\tilde{\mu}_۳^{\,۲} p_۲^۶ - (۲۴\tilde{\mu}_۳\lambda_۵ - ۱۰\lambda_۴^۲)p_۲^{\,۵}$$
$$-(۱۴۸\tilde{\mu}_۳^{\,۲}\lambda_۴ + ۲\lambda_۵^۲)p_۲^{\,۴} + (۲۸۸\tilde{\mu}_۳^{\,۴} - ۱۲\lambda_۴\lambda_۵\tilde{\mu}_۳ - \lambda_۴^{\,۳})p_۲^{\,۳}$$
$$+(۲۴\tilde{\mu}_۳^{\,۳}\lambda_۵ - ۷\tilde{\mu}_۳^{\,۲}\lambda_۴^۲)p_۲^۲ + ۳۲\tilde{\mu}_۳^{\,۴}\lambda_۴ p_۲ - ۲۴\tilde{\mu}_۳^{\,۶} = ۰$$

به طوری‌که

$$p_۱ = \mu_۱ + \mu_۲, \quad p_۲ = \mu_۱\mu_۲ \tag{۲}$$
$$\lambda_۴ = ۹\tilde{\mu}_۲^۲ - ۳\tilde{\mu}_۴,$$
$$\lambda_۵ = ۳۰\tilde{\mu}_۲\tilde{\mu}_۳ - ۳\tilde{\mu}_۵.$$

پی‌یرسون با حل معادله‌ی درجه‌ی ۹ بالا مقدار $p_۲$ را به دست آورد. بعد از به دست آوردن مقدار $p_۲$ مقدار متناظر با آن، یعنی $p_۱$، از رابطه‌ی زیر حاصل شد:

$$p_۱ = \frac{۲\tilde{\mu}_۳^۳ - ۲\tilde{\mu}_۳\lambda_۴ p_۲ - \lambda_۵ p_۲^۲ - ۸\tilde{\mu}_۳ p_۲^۳}{p_۲\left(۴\tilde{\mu}_۳^۲ - \lambda_۴ p_۲ + ۲p_۲^۳\right)}$$

سپس با توجه به رابطه‌ای (۲) μ_1 و μ_2 را معادل ریشه‌های معادله‌ی (۳) در نظر گرفت:

$$\mu^2 - p_1\mu + p_2 = 0 \qquad (۳)$$

و همچنین α و $1-\alpha$ را معادل ریشه‌های معادله‌ی (۴) قرار داد:

$$\alpha^2 - \alpha - \frac{p_2}{{p_1}^2 - 4p_2} = 0 \qquad (۴)$$

در انتها نیز، ${\sigma_1}^2$ و ${\sigma_2}^2$ از رابطه‌های زیر به دست می‌آیند:

$$(\mu_1\sigma_1)^2 = \frac{\tilde{\mu}}{\mu_1} - \frac{1}{3}\cdot\frac{\tilde{\mu}_3}{\mu_1\mu_2} - \frac{1}{3}(\mu_1+\mu_2) + \mu_2$$

$$(\mu_2\sigma_2)^2 = \frac{\tilde{\mu}}{\mu_2} - \frac{1}{3}\cdot\frac{\tilde{\mu}_3}{\mu_1\mu_2} - \frac{1}{3}(\mu_1+\mu_2) + \mu_1.$$

بعد از حل معادله‌ی درجه‌ی ۹، دو مدل برای داده‌ها به دست می‌آید. پی‌یرسون با رسم نمودارهای دو مدل، مشاهده کرد که هر دو مدل برای داده‌ها مناسب است. به‌عنوان معیاری برای مقایسه‌ی دو مدل، گشتاور مرتبه‌ی ششم دو مدل را به دست آورد و نتیجه گرفت که مدل ۱ نسبت به مدل ۲ بهتر است، زیرا گشتاور مرتبه‌ی ششم کم‌تری دارد. همان‌طور که مشاهده می‌شود روشی که پی‌یرسون برای براورد گشتاوری پارامترهای مدل در نظر گرفته است، برای داده‌های چندمتغیره، نیاز به محاسبات زیادی دارد که در عمل کاربرد چندانی نخواهد داشت. بدین منظور از روش ماکسیمم درستنمایی برای براورد پارامترهای مدل استفاده می‌شود. براوردگر ماکسیمم درستنمایی تحت برقراری شرایط نظم، کاراتر از براوردگر گشتاوری است. قابل ذکر است که معادلات بالا را می‌توان با استفاده از نرم‌افزارهای محاسبات جبری نظیر متمتیکا (Mathematica) حل کرد و نتیجه‌ی مشابه پی‌یرسون گرفت.

۲-۳- روش ماکسیمم درستنمایی

روشی متداول در آمار برای براورد پارامتر، روش ماکسیمم درستنمایی است. در این روش

پارامترها به گونه‌ای براورد می‌شوند که تابع درستنمایی مدل آمیخته‌ی گاوسی را ماکسیمم کند. تابع درستنمایی برای مشاهدات یک نمونه‌ی تصادفی $x_1, \dots, x_n$ با تابع چگالی (یا تابع جرم احتمال) $f(x_i;\boldsymbol{\theta})$ با بردار پارامتر $\boldsymbol{\theta}$ به‌صورت زیر تعریف می‌شود:

$$L(\boldsymbol{\theta}|x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

در بیش‌تر موارد برای سهولت کار $\ln L(\boldsymbol{\theta}|x_1, \dots, x_n)$ را ماکسیمم می‌کنند. تابع درستنمایی توزیع آمیخته‌ی گاوسی تک‌متغیره (مثال ۱) تحت تبدیل لگاریتمی به‌صورت زیر است:

$$l(\boldsymbol{\theta}) = \ln L\left(\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \middle| x_1, \dots, x_n\right)$$

$$= \sum_{i=1}^{n} \ln\left[\frac{\alpha}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{(1-\alpha)}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right\}\right] \quad (۵)$$

برای به دست آوردن براوردگر ماکسیمم درستنمایی از معادله‌ی (۵)، نسبت به $\boldsymbol{\theta} = (\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ مشتق گرفته برابر با صفر قرار داده و ۵ معادله به دست می‌آوریم. معادلات به دست آمده توابعی غیر خطی از پارامترها می‌باشند و ماکسیمم کردن آن‌ها به روش مستقیم امکان‌پذیر نمی‌باشد. همچنین برای براورد پارامترها، فرم تحلیلی وجود ندارد. از این رو این براوردگرها نیز همیشه رفتار خوبی ندارند [۳].

۱-۲-۳- براورد پارامترهای مدل آمیخته‌ی گاوسی به روش عددی

از آن‌جا که برای دستیابی به براورد ماکسیمم درستنمایی محاسبات زیادی مورد احتیاج است، می‌توان از روش‌های عددی، مقداری تقریبی برای براورد پارامترهای مدل را محاسبه کرد. دستور optim در نرم‌افزار R، با مینیمم کردن قرینه‌ی تابع درستنمایی، به روش عددی Nelder-Mead پارامترهای مدل را براورد می‌کند. می‌توان برای براورد پارامترهای مدل به روش عددی از دستور بخش پیوست استفاده کرد.

با توجه به این‌که روش‌های عددی نسبت به مقدار اولیه حساس هستند، از این رو نمی‌توان جواب‌های قابل قبولی از آن‌ها انتظار داشت. برای رفع این حساسیت، روش‌های استوارتری نسبت به روش‌های عددی ارائه شده که در بخش‌های بعد به معرفی آن‌ها می‌پردازیم.

۲-۲-۳- براورد پارامترهای مدل آمیخته‌ی گاوسی با الگوریتم EM

دمپستر و همکارانش در سال ۱۹۷۷ الگوریتم EM را ارایه دادند. این الگوریتم روشی برای محاسبه‌ی براوردگر ماکسیمم درستنمایی است، هنگامی که داده‌ی گمشده وجود داشته باشد یا روش‌های ساده‌ی بهینه‌سازی با شکست مواجه شوند [۲]. از مهم‌ترین کاربردهای الگوریتم EM یافتن براورد پارامترهای مدل آمیخته متناهی می‌باشد.

برای یافتن پارامترهای مدل آمیخته گاوسی در این روش علاوه بر مجموعه‌ی مشاهدات، از متغیر تصادفی برنولی Z_i با احتمال

$$\alpha = P(Z_i = ۱)$$

$$۱ - \alpha = P(Z_i = ۰)$$

استفاده می‌شود. به Z_i متغیر پنهان یا برچسب گفته می‌شود. به عبارت ساده‌تر با متناظر کردن یک برچسب به مشاهده‌ی x_i، می‌توان نشان داد که این مشاهده به کدام زیرجامعه تعلق دارد.

الگوریتم EM با در نظر گرفتن متغیرهای پنهان از چرخه‌ی مکرر برای براورد پارامترها استفاده می‌کند. این الگوریتم با در نظرگرفتن مقدار اولیه برای پارامترهای مدل شروع می‌شود، که به این مرحله، مرحله‌ی آغازین گویند. در گام بعد که مرحله‌ی تکرار نامیده می‌شود، این پارامترها به روز می‌شود و چرخه تا جایی تکرار می‌شود که الگوریتم همگرا شود. مرحله‌ی تکرار از دو گام محاسبه‌ی امید ریاضی و ماکسیمم‌سازی تشکیل می‌شود. در گام اول به جای محاسبه‌ی مستقیم لگاریتم تابع درستنمایی، امید ریاضی آن بر حسب بردار متغیرهای پنهان $\mathbf{Z} = (Z_۱, \dots, Z_n)$ به صورت زیر محاسبه می‌گردد:

$$E_{f(\mathbf{Z}|x_۱,\dots,x_n,\boldsymbol{\theta}^{(t)})}[\ln f(x_۱, \dots, x_n, \mathbf{Z}|\boldsymbol{\theta})].$$

در گام بعد پارامترهایی انتخاب می‌شوند که بر اساس آن‌ها امید ریاضی به دست آمده از مرحله‌ی قبل ماکسیمم مقدار شود یا به عبارتی:

$$\boldsymbol{\theta}^{(t+1)} = \sup E_{f(\mathbf{Z}|x_1,\ldots,x_n,\boldsymbol{\theta}^{(t)})}[\ln f(x_1,\ldots,x_n,\mathbf{Z}|\boldsymbol{\theta})]$$

در این‌جا منظور از $\boldsymbol{\theta}^{(t)}$ براورد $\boldsymbol{\theta}$ در تکرار tاُم می‌باشد. از آن‌جا که مقدار تابع درستنمایی در هر تکرار افزایش می‌یابد، از این رو این الگوریتم، همگراست. بنا بر این براوردهای به دست آمده از این روش به مقدار ماکسیمم درستنمایی آن‌ها میل می‌کند.

اگر $\mu_1^{(t)}, \mu_2^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}, \alpha^{(t)}$ براوردهای به دست آمده از مرحله‌ی tاُم الگوریتم باشند، امید تابع درستنمایی مثال ۱ را با $Q = Q(\alpha^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$ نشان می‌دهیم، و می‌توان آن را به‌صورت زیر نوشت:

$$Q = E\left[\left\{\frac{\alpha^{(t)}}{\sqrt{2\pi\sigma_1^{2(t)}}}\exp\left(-\frac{(x_i-\mu_1^{(t)})^2}{2\sigma_1^{2(t)}}\right)\right\}^{Z_i} \times \left\{\frac{1-\alpha^{(t)}}{\sqrt{2\pi\sigma_2^{2(t)}}}\exp\left(-\frac{(x_i-\mu_2^{(t)})^2}{2\sigma_2^{2(t)}}\right)\right\}^{1-Z_i}\right]$$

$$= \sum_{i=1}^{n} E\left(Z_i \middle| x_i, \mu_1^{(t)}, \sigma_1^{2(t)}\right)\left(\ln\alpha^{(t)} - \frac{1}{2}\ln\left(2\pi\sigma_1^{2(t)}\right) - \frac{(x_i-\mu_1^{(t)})^2}{2\sigma_1^{2(t)}}\right)$$

$$+\left[1-E\left\{Z_i|x_i,\mu_2^{(t)},\sigma_2^{2(t)}\right\}\right]\left[\ln\left\{1-\alpha^{(t)}\right\}-\frac{1}{2}\ln\left\{2\pi\sigma_2^{2(t)}\right\}-\frac{\left\{x_i-\mu_2^{(t)}\right\}^2}{2\sigma_2^{2(t)}}\right] \quad (۶)$$

سپس $E\left\{Z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)}\right\}$ محاسبه می‌شود:

$$E\left\{Z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)}\right\}=f\left(Z_i=1|x_i,\mu_1^{(t)},\sigma_1^{2(t)}\right)$$

$$=\frac{\alpha^{(t)}\phi\left(x_i;\mu_1^{(t)},\sigma_1^{2(t)}\right)}{\alpha^{(t)}\phi\left(x_i;\mu_1^{(t)},\sigma_1^{2(t)}\right)+(1-\alpha^{(t)})\phi\left(x_i;\mu_2^{(t)},\sigma_2^{2(t)}\right)}$$

و در نهایت از رابطه‌ی (۶) نسبت به پارامترها مشتق می‌گیریم:

$$\frac{\partial Q}{\partial\alpha^{(t)}}=\frac{\sum_{i=1}^{n}E\left\{Z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)}\right\}-n\alpha^{(t)}}{\alpha^{(t)}(1-\alpha^{(t)})}$$

$$\frac{\partial Q}{\partial\mu_1^{(t)}}=\frac{\sum_{i=1}^{n}E\left\{Z_i|x_i,\mu_1^{(t)},\sigma_1^{2(t)}\right\}\left(x_i-\mu_1^{(t)}\right)}{2\sigma_1^{2(t)}}$$

$$\frac{\partial Q}{\partial\mu_2^{(t)}}=\frac{\sum_{i=1}^{n}\left[1-E\left\{Z_i|x_i,\mu_2^{(t)},\sigma_2^{2(t)}\right\}\right](x_i-\mu_2^{(t)})}{2\sigma_2^{2(t)}}$$

$$\frac{\partial Q}{\partial \sigma_{۱}^{۲(t)}} = \frac{\sum_{i=۱}^{n} E\left\{Z_i | x_i, \mu_{۱}^{(t)}, \sigma_{۱}^{۲(t)}\right\}\left\{(x_i - \mu_{۱}^{(t)})^{۲} - \sigma_{۱}^{۲(t)}\right\}}{۲\sigma_{۱}^{۴(t)}}$$

$$\frac{\partial Q}{\partial \sigma_{۲}^{۲(t)}} = \frac{\sum_{i=۱}^{n} \left[۱ - E\left\{Z_i | x_i, \mu_{۲}^{(t)}, \sigma_{۲}^{۲(t)}\right\}\right]\left\{(x_i - \mu_{۲}^{(t)})^{۲} - \sigma_{۲}^{۲(t)}\right\}}{۲\sigma_{۲}^{۴(t)}}$$

می‌توان نشان داد ماتریس مشتقات دوم پارامترها، معین منفی است و در نتیجه از برابر صفر قرار دادن مشتق Q براورد پارامترها، که در هر مرحله به‌روز می‌شوند، را محاسبه کرد. به گونه‌ی ساده‌تر می‌توان الگوریتم EM را برای مثال ۱، به‌صورت زیر بیان کرد:

الگوریتم EM برای مدل آمیخته‌ی گاوسی دومولفه‌ای:

گام اول: انتخاب مقادیر اولیه برای پارامترهای مدل $\left(\mu_{۱}^{(t)}, \mu_{۲}^{(t)}, \sigma_{۱}^{۲(t)}, \sigma_{۲}^{۲(t)}, \alpha^{(t)}\right)$ به ازای $t = ۰$.

گام دوم: محاسبه‌ی امید ریاضی. در این مرحله احتمال متعلق بودن مشاهده‌ی iاُم به مؤلفه‌ی اول (که آن را $\gamma_i^{(t)}$ نامیده‌ایم) محاسبه می‌شود.

جدول ۱- توزیع پیشین و توزیع پسین برای پارامترهای مدل آمیخته‌ی گاوسی $(k = ۱, ۲)$

پارامترها	توزیع پیشین	توزیع شرطی کامل
α	$D(\delta_{۱}, \delta_{۲})$	$D(\delta_{۱} + n, \delta_{۲} + n)$
μ_k	$N(\mu_{۰}, \tau^{۲})$	$N\left(\frac{\tau^{۲}\sum_{i=۱;Z_i=k}^{n} x_i + \mu_{۰}\sigma_k^{۲}}{n_k\tau^{۲} + \sigma_k^{۲}}, \frac{۱}{n_k\tau^{۲} + \sigma_k^{۲}}\right)$
$\sigma_k^{۲}$	$IG(\omega_{۰}, \beta_{۰})$	$IG\left(\omega_{۰} + \frac{۱}{۲}n_k, \beta_{۰} + \frac{۱}{۲}\sum_{i=۱;Z_i=k}^{n}(x_i - \mu_k)^{۲}\right)$

$$\gamma_i^{(t)} = \frac{\alpha^{(t)}\phi_{\theta_۱^{(t)}}(x_i)}{\alpha^{(t)}\phi_{\theta_۱^{(t)}}(x_i) + (۱-\alpha^{(t)})\phi_{\theta_۲^{(t)}}(x_i)}, \quad i = ۱, \ldots, n$$

گام سوم: ماکسیمم‌سازی. در این مرحله پارامترهای مدل حسب $\gamma_i^{(t)}$ که از گام دوم به دست آمده، محاسبه می‌شود.

$$\mu_۱^{(t+۱)} = \frac{\sum_{i=۱}^n \gamma_i^{(t)} x_i}{\sum_{i=۱}^n \gamma_i^{(t)}}, \qquad \mu_۲^{(t+۱)} = \frac{\sum_{i=۱}^n (۱-\gamma_i^{(t)}) x_i}{\sum_{i=۱}^n (۱-\gamma_i^{(t)})},$$

$$\sigma_۱^{۲(t+۱)} = \frac{\sum_{i=۱}^n \gamma_i^{(t)} (x_i - \mu_۱^{(t)})^۲}{\sum_{i=۱}^n \gamma_i^{(t)}},$$

$$\sigma_۲^{۲(t+۱)} = \frac{\sum_{i=۱}^n (۱-\gamma_i^{(t)}) (x_i - \mu_۲^{(t)})^۲}{\sum_{i=۱}^n (۱-\gamma_i^{(t)})},$$

$$\alpha^{(t+۱)} = \sum_{i=۱}^n \frac{\gamma_i^{(t)}}{n}.$$

گام چهارم: تکرار گام ۲ و ۳ تا رسیدن به همگرایی در براورد پارامترهای مدل.

برای براورد پارامترهای مدل بر اساس الگوریتم EM از کدهای نوشته شده در بخش پیوست می‌توان استفاده کرد.

۳-۲-۳- براورد پارامترهای مدل آمیخته‌ی گاوسی با روش بیزی

در آمار بیز، باید اطلاعات موجود در خصوص پارامترهای مجهول را به‌صورت یک توزیع آماری، که به آن توزیع پیشین گفته می‌شود، بیان کرد. یکی از سراست‌ترین انتخاب‌ها برای توزیع پیشین، توزیع مزدوج است [۱]. یعنی خانواده‌ای از توزیع‌ها که اگر توزیع پارامتر به شرط مشاهدات را محاسبه کنیم، هم خانواده‌ی توزیع پیشین باشد. توزیع به

دست آمده توزیع پسین نام دارد. اساس استنباط بیزی بر پایه‌ی توزیع پسین و تابع زیانی است که در نظر گرفته می‌شود. به عبارت دیگر براوردگر بیز با توجه به تابع زیان در نظر گرفته شده تغییر می‌کند. به‌عنوان مثال اگر تابع زیان را به‌صورت توان‌های دوم خطا در نظر گرفته شود، براوردگر بیز برابر میانگین توزیع پسین می‌شود. در بسیاری از مسائل محاسبه‌ی این امید ریاضی به‌صورت تحلیلی وجود ندارد. روش مرسوم برای محاسبه‌ی آن با استفاده از روش شبیه‌سازی است، که به روش مونت کارلو (Monte Carlo) شهرت دارد. در این روش با تولید اعداد تصادفی از توزیع پسین، میانگین اعداد تولیدشده را به‌عنوان براورد پارامتر توزیع پسین در نظر می‌گیرند، که پشتوانه‌ی صحت و کارا بودن آن قانون ضعیف اعداد بزرگ است. با توجه به این‌که در مسئله‌ی پیش‌رو، ما نیاز به براورد هم‌زمان چند پارامتر داریم، و تولید اعداد تصادفی از توزیع توأم کار ساده‌ای نیست، لذا روش‌های گوناگونی برای تسهیل تولید اعداد تصادفی از توزیع‌های توأم، ارایه شده است. روش نمونه‌گیر گیبز از جمله روش‌های مونت کارلوی زنجیر مارکوفی می‌باشد که بر اساس توزیع شرطی مشاهدات، زنجیر مارکوفی از آن‌ها تولید می‌کند. این روش اولین بار در سال ۱۹۸۴ در مقاله‌ای توسط برادران گمن برای مدل‌های پردازش تصویر بیان شد، اما الگوریتمی که امروزه به‌عنوان الگوریتم گیبز در مسائل آماری از آن استفاده می‌کنیم، در سال ۱۹۹۰ توسط گلفند و اسمیت ارایه شد. نمونه‌گیر گیبز روشی برای تولید متغیرهای تصادفی بر اساس توزیع شرطی آن‌ها است. از سوی دیگر، از لحاظ نظری صحت و کارا بودن این روش براورد اثبات شده و پیاده‌سازی آن نیز دشواری زیادی ندارد [۱۲]. در این روش توزیع تک‌تک پارامترها به شرط بقیه متغیرها محاسبه می‌شود که توزیع شرطی کامل مشهور است. می‌توان نشان داد، با تولید اعداد تصادفی از توزیع‌های شرطی کامل، از آن‌ها بجای اعداد تصافی از توزیع توأم در براورد بیز بهره برد.

به پارامترهای توزیع پیشین ابرپارامتر گفته می‌شود، بنا بر این اولین گام برای اجرای الگوریتم گیبز تعیین مقادیر اولیه برای ابرپارامترها می‌باشد. هرچند که روش‌های متفاوتی برای رهایی از تعیین این مقادیر اولیه وجود دارد، این وابستگی به مقادیر اولیه تأثیری در همگرایی روش نمونه‌گیر گیبز ندارد ولی زمان همگرایی را تغییر می‌دهد. مرسوم‌ترین روش برای تعیین ابرپارامترها استفاده از روش بیز تجربی پارامتری است. در این روش با در نظر گرفتن یک فرض اضافی روی مشاهدات، باعث استقلال توزیع حاشیه‌ای مشاهدات شده و به روش‌های آمار کلاسیک این ابرپارامترها را براورد می‌کند. هرچند از روش‌های

بیز سلسله مراتبی یا میانگین‌گیری بیزی نیز می‌توان استفاده کرد، روش بیز تجربی ساده‌تر و در عین حال کارایی مطلوبی دارد. در این مقاله ما بر اساس تجربه این مقادیر را تعیین کرده‌ایم. اگر این مقادیر را تغییر دهید مشاهده خواهید کرد که تغییر چندانی در نتایج شبیه‌سازی رخ نمی‌دهد.

در جدول ۱ توزیع پیشین مزدوج برای مثال ۱ آورده شده است [۵]. همان‌طور که می‌بینیم $\boldsymbol{\delta} = (\delta_1, \delta_2), \omega_0, \beta_0, \nu_0, \tau^2$ ابرپارامترها برای توزیع‌های پیشین در نظر گرفته شده‌اند. بعد از تولید متغیرهای تصادفی بر اساس توزیع پیشین، وارد مرحله‌ی تکرار می‌شویم. در اولین گام از تکرار tاُم، احتمال متعلق بودن هر مشاهده به دو زیرجامعه، بر حسب پارامترهای قبل‌تر یعنی مرحله‌ی $(t-1)$اُم محاسبه می‌گردد. سپس بر اساس این اطلاعات به دست آمده، پارامترها بر اساس توزیع شرطی کامل‌شان از جدول ۱ تولید می‌گردند. حال به کمک روش‌های بیزی و جدول ۱ الگوریتم نمونه‌گیر گیبز را برای براورد پارامترهای مدل آمیخته‌ی گاوسی (مثال ۱)، می‌توان به‌صورت ساده‌تر بیان کرد:

الگوریتم نمونه‌گیر گیبز برای مدل آمیخته‌ی گاوسی دو مولفه‌ای:

مرحله‌ی آغازین: در این مرحله مقادیر اولیه برای ابرپارامترهای توزیع پیشین $\boldsymbol{\delta}, \omega_0, \beta_0, \mu_0, \tau^2$ انتخاب می‌شوند، سپس بر اساس این ابرپارامترها از توزیع پیشین داده تولید می‌کنیم. توجه کنید که در این‌جا مقدار اولیه برای ابرپارامترهای هر دو زیرجامعه یکسان در نظر گرفته شده است.

مرحله‌ی تکرار: این مرحله برای هرt، $t = 1, 2, \ldots, T$ که T تعداد تکرار الگوریتم و بسته به نظر کاربر تعریف می‌شود، در دو گام انجام می‌شود:

گام اول: ابتدا Z_i از توزیع چند جمله‌ای با احتمال زیر تولید می‌شود.

$$P\left(Z_i = 1 | \mu_1^{(t-1)}, \sigma_1^{2(t-1)}, \alpha^{(t-1)}\right) = \frac{\alpha^{(t-1)} \phi_{\theta_1^{(t-1)}}(x_i)}{\alpha^{(t-1)} \phi_{\theta_1^{(t-1)}}(x_i) + \left\{1 - \alpha^{(t-1)}\right\} \phi_{\theta_2^{(t-1)}}(x_i)}$$

در ایـن جـا ذکـر ایـن نکتـه لازم اسـت کـه $\mu_1^{(t-1)}, \sigma_1^{2(t-1)}, \alpha^{(t-1)}$ در تکـرار اول یعنـی $t = 1$ همـان متغیرهـای تصـادفی تولیـدشـده از توزیـع پیشـین در مرحلـه‌ی آغـازین می‌باشند.

گام دوم: $\alpha^{(t)}$، $\mu_1^{(t)}$، $\mu_2^{(t)}$، $\sigma_1^{2(t)}$ و $\sigma_2^{2(t)}$ بر اساس توزیع شرطی کامل‌شان در جدول ۱ تولید می‌شوند.

برای براورد پارامترهای مدل با استفاده از الگوریتم گیبز، می‌توان کدهای بخش پیوست را به کار برد. نقطه‌ی داغیدن در این الگوریتم ۵۰۰ در نظر گرفته شده است.

در انتها، در بخش نتیجه‌گیری، نتـایج بـه دسـت آمـده از روش‌هـای مختلـف را بـا یکـدیگر مقایسه و مورد ارزیابی قرار می‌دهیم و برتری‌ها و محدودیت‌های هر یک را بـرای کـاربران مشخص می‌کنیم.

۴- نتیجه‌گیری

نتایج به دست آمده در جدول‌های ۲، ۳ و ۴ مربوط بـه داده‌هـای شبیه‌سـازی شـده از مـدل آمیخته‌ی گاوسی دو مؤلفه‌ای به حجم ۱۰۰۰ در نرم‌افزار R می‌باشد. همان‌طور که مـی‌بینیم پارامترها، مقدار واقعی آن‌ها، مقدار اولیه‌ی پارامترها، متوسط مقدار براوردشـده، انحـراف معیـار و میـانگین مربـع خطـای نمونـه‌ای در سـتون‌های ایـن جـداول نشـان داده شـده‌اند. میانگین مربع خطای نمونه‌ای به‌عنوان معیاری برای مقایسه‌ی دقت روش‌های شبیه‌سازی، در آمار کلاسیک، در نظر گرفته شده است، که از رابطه‌ی زیر به دست می‌آید:

$$\frac{1}{1000}\sum_{i=1}^{1000} (\hat{\theta}_i - \theta)^2$$

که θ مقدار واقعی و $\hat{\theta}_i$ مقادیر براوردشـده در هـر الگـوریتم می‌باشـد. تعـداد تکـرار در هـر روش ۱۰۰۰ در نظر گرفته شده و متوسط اختلاف مقدار واقعی و مقدار براوردشـده مبنـای مقایسه‌ی روش‌هاست. همان‌گونه کـه انتظـار داشـتیم روش گیبـز از EM و EM از عـددی پاسخ بهتری بـه مـا می‌دهـد. امـا زمـان محاسـباتی روش EM از عـددی، و گیبـز از EM بیش‌تر، ولی شرایط همگرایـی آن‌هـا کـم‌تـر اسـت. در حالـت کلـی بـراورد بیـز پاسـخ قابـل قبول‌تری در مقایسه با دو روش دیگر ارایه داده است.

در پایان نیز برای قابل لمس بودن موارد استفاده از مدل‌های آمیخته کاربردهایی از آن، ارایه شده است.

کاربردها

در سال‌های اخیر برای شناسایی گوینده (تشخیص صدا)، در متون مستقل، از مدل‌های آمیخته‌ی گاوسی استفاده می‌کنند. شناسایی گوینده، هنگامی که هیچ پیش‌فرضی از آن‌چه گوینده به زبان می‌آورد وجود ندارد، اصطلاحاً شناسایی گوینده با متون مستقل گفته می‌شود. برای هر گوینده مدل آمیخته‌ی گاوسی به گونه‌ای در نظر گرفته می‌شود که تابع احتمال پسین‌اش ماکسیمم مقدار شود. رینولد و روز در سال ۱۹۹۵ در مقاله‌ی [۱۱] نشان دادند که مدل آمیخته‌ی گاوسی برای شناسایی گوینده در متون مستقل، مدلی استوار می باشد.

علاوه بر تشخیص گوینده، از مدل آمیخته‌ی گاوسی برای شناسایی چهره‌ی افراد نیز استفاده می‌شود [۱۰]. مهم‌ترین مشکل در تشخیص چهره‌ی افراد و بازیابی آن، سایه روشن‌ها، تغییرات نور و پس‌زمینه‌های همرنگ است. تشخیص چهره در مسائل امنیتی، تشخیص تغییرات در افراد و فهرست‌گذاری در تصاویر ویدئویی، کاربرد دارد. برای مدل‌بندی رنگ چهره‌ی (پوست) افراد نیز، می‌توان از مدل آمیخته‌ی گاوسی استفاده کرد.

در روش خوشه‌بندی مبتنی بر مدل، که برای مشاهدات، مدلی احتمالاتی در نظر گرفته می‌شود، از مدل آمیخته‌ی گاوسی استفاده می‌شود. بدین صورت که در این روش، هر خوشه به وسیله‌ی یک توزیع پارامتری نشان داده می‌شود. آن‌گاه مدلی که برای کل داده‌ها ارایه می‌شود ترکیب آمیخته‌ی متناهی از این توزیع‌ها، می‌باشد. با استفاده از مدل آمیخته‌ی گاوسی، اطلاعات کامل‌تری درباره‌ی خوشه‌ها به دست می‌آوریم.

به‌دلیل انعطاف‌پذیری مدل آمیخته‌ی گاوسی برای انواع مختلفی از توزیع‌ها، در یافتن الگوهایی برای امور مالی تجربی نیز، از مدل آمیخته‌ی گاوسی استفاده می‌شود. در مدل‌سازی مالی و کاربردهای آن، توزیع نرخ سود (بازده) در دارایی‌های مالی نقش مهمی دارد. متداول‌ترین فرض این است که نرخ سود دارایی‌ها، توزیع گاوسی دارد و از آن‌جا که دیگر توزیع‌ها نیز می‌توانند به خوبی با یک مدل آمیخته‌ی گاوسی متناهی تقریب زده شوند، این مدل در امور مالی مورد توجه بسیار قرار گرفته است.

همچنین مدل آمیخته‌ی گاوسی در نجوم، زیست‌شناسی، پزشکی و مهندسی نیز کاربرد بسیاری دارد که برای جزییات بیش‌تر می‌توان به [۴]، [۶]، [۷] و [۱۳] مراجعه کرد.

جدول ۲- براوردماکسیمم درستنمایی با روش عددی

پارامترها	مقدار واقعی	مقدار اولیه	مقدار براوردشده	انحراف معیار	میانگین مربع خطای نمونه‌ای
α_1	۰٫۷	۰٫۵	۰٫۵۷۳۳۱۸۹	۰٫۲۴۵۸۲۱۲	۰٫۱۴۷۷۶۹۸
α_2	۰٫۳	۰٫۵	۰٫۴۲۶۶۸۱۱	۰٫۲۴۵۸۲۱۲	۰٫۱۴۷۷۶۹۸
μ_1	۳	۴	۲٫۹۳۰۶۹	۰٫۳۳۵۹۰۱۴	۰٫۱۰۱۸۳۳۴
μ_2	۱	۱٫۶	۱٫۵۱۵۷۶۸	۰٫۹۰۸۲۸۲۵	۰٫۹۱۵۳۴۲۲
σ_1	۱	۰٫۳	۰٫۸۷۵۷۶۲۹	۰٫۲۲۶۰۸۹۵	۰٫۱۶۱۴۳۹۶۸
σ_2	۱٫۲	۰٫۸	۱٫۲۵۵۱۶۸۸	۰٫۳۵۳۷۲۸۷	۰٫۸۷۷۷۲۲۶۸

جدول ۳- براورد ماکسیمم درستنمایی بر اساس الگوریتمEM

پارامترها	مقدار صحیح	مقدار اولیه	مقدار براوردشده	انحراف معیار	میانگین مربع خطای نمونه‌ای
α_1	۰٫۷	۰٫۵	۰٫۵۲۶۳۹۶	۰٫۱۹۴۳۷۷	۰٫۰۷۰۴۳۳۳۵
α_2	۰٫۳	۰٫۵	۰٫۴۷۳۶۰۴	۰٫۱۹۴۳۷۷	۰٫۰۷۰۴۳۳۳۵
μ_1	۳	۴	۳٫۱۰۸۸۳۶	۰٫۲۴۳۸۲۳۱	۰٫۰۶۵۳۵۰۰۳
μ_2	۱	۱٫۶	۱٫۸۲۳۷۵۶	۰٫۳۱۹۹۵۳۱	۰٫۷۷۰۷۰۷۵۵
σ_1	۱	۰٫۳	۰٫۷۲۶۲۶۳۲	۰٫۳۳۴۴۰۵۷	۰٫۱۲۸۴۹۵۱
σ_2	۱٫۲	۰٫۸	۱٫۹۸۶۶۵۲۱	۰٫۲۱۳۹۵۶۳	۰٫۷۳۱۸۶۴۶

جدول ۴- براورد بیز بر اساس الگوریتم گیبز

پارامترها	مقدار صحیح	مقدار براوردشده	انحراف معیار	میانگین مربع خطای نمونه‌ای
α_1	۰/۷	۰/۶۳۷۲۹۱۵	۰/۱۱۶۴۴۱۷	۰/۰۴۲۳۴۱۱۵
α_2	۰/۳	۰/۳۶۲۷۰۸۵	۰/۱۱۶۴۴۱۷	۰/۰۴۲۳۴۱۱۵
μ_1	۳	۲/۹۴۸۳۰۵	۰/۱۴۶۳۴۳۴	۰/۰۴۵۰۳۸۲
μ_2	۱	۱/۲۷۹۸۰۰	۰/۲۱۶۹۶۳۴	۰/۳۴۶۹۳۰۳
σ_1	۱	۱/۰۸۴۸۲۸	۰/۱۰۹۲۲۳۱	۰/۰۴۳۲۵۲۸۸
σ_2	۱/۲	۱/۳۷۷۲۹۶	۰/۲۹۵۵۳۸۳	۰/۶۸۲۷۹۸۳۱

قدردانی

نویسندگان بر خود واجب می‌دانند از پیشنهادات، نظرات و تصحیح‌های داوران محترم که باعث بهبود مقاله شده است، تشکر کنند. همچنین از آقای حسین هشیارمنش به‌خاطر خواندن متن نهایی و تصحیح برخی اشتباه‌های تایپی نویسندگان تشکر می‌شود.

مرجع‌ها

[1] Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46, 167-174.

[2] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. journal of the Royal Statistical Society, 39, 1-38.

[3] Everitt, B. S. and Hand, D. J. (1981). Finite Mixture Distributions. Chapman and Hall, New York.

[4] Greenspan, H., Goldberger, J. and Eshe, I. (2001). Mixture model for face-color modeling and segmentation. Pattern Recognition Letters, 22, 1525-1536.

[5] Gonzalez, D.S., Kuruoglu, E. and Ruiz, D.P. (2010). with mixture of symmetric stable distributions using Gibbs sampling. Modelling Signal Processing, 90, 774–783.

[6] Kon, S. (1984). Models of stock returns a comparison. The Journal of Finance, 39, 147–165.

[7] McKenna, S., Gong, S. and Raja, Y. (1998). Modelling facial color and identity with Gaussian mixtures. Pattern Recognition, 31, 1883–1892.

[8] Melnykov, V. and Maitra, R. (2010). Finite mixture models and model–based clustering. Statistics Surveys, 4, 80–116.

[9] Pearson, K. (1894). Contributions to the mathematical theory of evolution source , Philosophical Transactions of the Royal Society of London, 185, 71–110.

[10] Reynolds, D.A., Quatieri, T.F. and Dunn R.B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10, 19–41.

[11] Reynolds, D.A. and Rose, R.C. (1995), Robust text–independent speaker identification using Gaussian mixture speaker models, IEEE Transaction on Speech Audio Process, 3, 72–83.

[12] Robert, C.P and Casella, G. (2004). Monte Carlo Statistical Methods. Springer, New York.

[13] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). Statistical Analysis of Finite of Mixture Distributions. Willey, New York.

پیوست

کد تولید اعداد تصادفی مدل آمیخته گاوسی

```
#Generating random sample from Gaussian mixture of 2
components
#n: sample size
```

```
#alpha1: mixing proportion of the first component
#mean1: mean of the first component
#mean2: mean of the second component
#sigma1: sd of the first component
#sigma2: sd of the second component
n=1000
alpha=0.7
mean1=3
mean2=1
sigma1=1
sigma2=1.2
z=rbinom(n,1,alpha)
x=rnorm(n,
ifelse(z==1,mean1,mean2),ifelse(z==1,sigma1,sigma2))
```

کد براورد پارامترهای مدل آمیخته گاوسی با روش عددی

```
#At first, it is defined the likelihood function for
maximizing
#par: parameters of Gaussian mixture of 2 components
#par=c(mean1,mean2,sigma1,sigma2,alpha)
#x: univariate Gaussian mixture of 2 components
Likelihood=function(par,x){
f=par[5]*dnorm((x-par[1])/par[3])/par[3]+
(1-par[5])*dnorm((x-par[2])/par[4])/par[4]
if(any (f<=0)) Inf
else -sum(log(f))}
#initial values
intpar=c(4,1.6,0.3,0.8,0.5)
optim(intpar,Likelihood,x=x)$par
```

کد براورد پارامترهای مدل آمیخته گاوسی با روش الگوریتم EM

```
#EM Algorithm:
Em=function(par){
#stage 1: (E-step)
gamma1=NULL
gamma1=par[5]*dnorm(x,par[1],sqrt(par[3]))/((1-par[5])*
dnorm(x,par[2],sqrt(par[4]))+par[5]*dnorm(x,par[1],sqrt(
par[3])))
#stage 2: (M-step)
par[5] = (mean(gamma1))
par[1] = sum(gamma1*x)/sum(gamma1)
par[2] = sum((1-gamma1)*x)/sum(1-gamma1)
par[3] = sum(gamma1*((x-par[1])^2))/sum(gamma1)
par[4] = sum((1-gamma1)*((x-par[2])^2))/sum(1-gamma1)
```

```
c(par[1],par[2],par[3],par[4],par[5])}
# Initial values for EM algorithm:
par0 = c(4,1.6,0.3,0.8,0.5)
# Running the EM algorithm
dis = 1
iter = 1
while (dis > 0.01 || iter <= 200){
iter = iter+1
param = Em(par0)
dis = max(abs(par0-param))
par0 = param
}
par0
```

کد براورد پارامترهای مدل آمیخته گاوسی با روش الگوریتم گیبز

```
#iteration: number of iterations of the algorithm.
##define initial values for parameters##
#Packages “rgenoud” and “multinomRob” should be installed.
library(rgenoud)
library(multinomRob)
k=2
iteration=1000
mix.new=mu.new= var.new=matrix(0,iteration,k)
z.new=matrix(0,length(x),k)
mix=mu=var=NULL
##define hyperparameter for prior distibution###
delta=mu.0= omega.0= betta.0=rep(1,k)
tau2=rep(9,k)
##generate random sample from the prior distributions##
for(i in 1:k){
mix[i]<-rgamma(n=1,shape=delta[i],rate=1)
mu[i]=rnorm(1,mu.0[i],sqrt(tau2[i]))
var[i]=1/rgamma(1,omega.0[i],betta.0[i])
}
mix=mix/sum(mix)
numer=matrix(0,nrow=length(x),ncol=k)
## Iteration step#####
for(it in 1:iteration)
{
## find the latent variable z#####
for(i in 1:k) {
numer[,i]=(mix[i]*dnorm(x,mean=mu[i],sd=sqrt(var[i])))
}
prob=numer/matrix(rep(rowSums(numer),k),ncol=k,byrow=F)
z=matrix(0,length(x),k)
for(j in 1:length(x)){
```

```
z[j,]=t(rmultinomial(1,prob[j,]))
}
n.mix=apply(z,2,sum)
## generate parameters from Posterior distribution ####
for(i in 1:k) {
mix[i]=rgamma(1,shape=delta[i]+n.mix[i],rate=1)
mu[i]=rnorm(1,(tau2[i]*sum(x[z[,i]==1])+mu.0[i]*var[i])/
(n.mix[i]*tau2[i]+var[i]),sqrt((var[i]*tau2[i])/
(n.mix[i]*tau2[i]+var[i])))
var[i]=1/rgamma(1,shape=omega.0[i]+.5*n.mix[i],rate=
betta.0[i] + .5*sum(z[,i]*(x-mu[i])^2))
}
mix=mix/sum(mix)
##Save###
z.new=z.new+z
mix.new[it,]=mix
mu.new[it,]=mu
var.new[it,]=var
}
##END OF ITERATION STAGE###
##Compute mean between estimated parameter #####
apply(mix.new[(iteration/2):(iteration-10),],2,mean)
apply(mu.new[(iteration/2):(iteration-10),],2,mean)
apply(var.new[(iteration/2):(iteration-10),],2,mean)
```

دنیا رحمانی
فوق لیسانس آمار
تهران، خیابان حافظ، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، دانشکده‌ی ریاضی و علوم کامپیوتر، گروه آمار.

عادل محمدپور
دکتری آمار
تهران، خیابان حافظ، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، دانشکده‌ی ریاضی و علوم کامپیوتر، گروه آمار.
رایانشانی: adel@aut.ac.ir