مجله مدل‌سازی پیشرفته ریاضی
سال ۱۳۹۹،دوره۱۰، شماره۲، ص ۲۶۶-۲۴۵
شاپا چاپی: ۸۰۸۸-۲۲۵۱
DOI: 10.22055/JAMM.2020.30812.1753

تحلیلی بر مساله انتخاب متغیرهای کمکی در مدل گاوسی با استفاده از ماکسیمم احتمال پسین و رهیافت‌های فراوانی‌گرا و بیزی

امیرحسین قطاری[۱*]، مجتبی گنجعلی[**]

* گروه آمار، دانشگاه صنعتی امیرکبیر، تهران، ایران

** گروه آمار، دانشگاه شهید بهشتی، تهران، ایران

تاریخ دریافت: ۱۳۹۸/۰۶/۰۴ تاریخ پذیرش: ۱۳۹۸/۱۲/۲۲

چکیده: مسئله‌ی انتخاب مناسب‌ترین مدل جهت برازش بر روی داده‌ها همواره چالش برانگیز بوده است. روش ماکسیمم احتمال پسین از جمله روش‌های انتخاب مدل است که در هر دو رهیافت فراوانی‌گرا و بیزی کاربرد دارد. به‌علاوه، مطلوبیت مدل نیز یکی از ابزارهای مورد استفاده برای سنجش عملکرد روش‌های انتخاب مدل است. در این مقاله، روش ماکسیمم احتمال پسین برای مدل گاوسی استاندارد بیزی موردمطالعه قرار گرفته و عملکرد آن با حالت فراوانی‌گرا مقایسه می‌شود. همچنین، یک صورت جبری برای برآورد مطلوبیت مدل ارائه خواهد شد. در ادامه، مطالعه‌ی شبیه‌سازی روی مدل گاوسی عملکرد بهتر رهیافت بیزی را هم از دیدگاه مطلوبیت و هم با بررسی میانگین توان‌های دوم خطای مدل (MSE) تایید می‌کند. با این وجود، هر دو رهیافت با افزایش اندازه‌ی نمونه، کمتر در معرض بیش‌برازش قرار می‌گیرند. همچنین با افزایش ضریب همبستگی میان متغیرهای کمکی، MSE در هر دو رهیافت افزایش یافته حال آن‌که تمایل به انتخاب مدل با تعداد متغیر کمتر، افزایش می‌یابد. مطالعه بر روی داده‌های واقعی نشان می‌دهد که در هر دو رهیافت با افزایش اندازه‌ی نمونه، MSE مدل‌های انتخاب شده، کاهش می‌یابد.

واژه‌های کلیدی: انتخاب مدل، روش ماکسیمم احتمال پسین، بیش‌برازش، مطلوبیت مدل.

رده‌بندی ریاضی (۲۰۱۰): ۶۲J۰۵، ۶۲J۱۲

۱- آدرس الکترونیکی نویسنده مسئول مقاله: a.h.ghatari@aut.ac.ir

۱- مقدمه

انتخاب مدل یکی از مسائل بنیادی در مدل‌بندی‌های آماری است. دیدگاه مورد پذیرش در انتخاب مدل، یافتن سودمندترین مدل از نقطه نظر پیشگویی است. معمولاً سودمندی یک مدل به‌وسیله‌ی توانایی آن در پیشگویی مقادیر آینده یا درون‌یابی نقاط مشاهده نشده سنجیده می‌شود. با اینکه پیشگویی مهمترین قسمت مربوط به مسائل مدل‌بندی نیست، توانایی پیشگویی یک سنجش طبیعی برای فهمیدن این است که آیا مدل مناسب است یا خیر؛ بنابراین یافتن بهترین مدل از دیدگاه پیشگویی یکی از ملزومات مدل‌بندی‌های آماری است.

فرض کنید مجموعه‌ای از مدل‌ها که لزوماً ابعاد بردار پارامتری آن‌ها یکسان نیست دراختیار باشند و بخواهیم تصمیم‌گیری کنیم کدام یک از آن‌ها بهترین برازش را بر روی داده‌ها دارند. همچنین برای متغیر پاسخ چند متغیر کمکی می‌توان انتخاب کرد که پاسخ با آن‌ها به‌صورت تک به تک وابستگی دارد و یا اینکه در حضور مابقی متغیرهای کمکی همبستگی جزئی قابل رویت است؛ بنابراین هدف، انتخاب یک مدل از میان تمام مدل‌های ممکن است به‌گونه‌ای که مناسب‌ترین عملکرد را با توجه به انتخاب متغیرهای کمکی بهینه داشته باشد. در چنین شرایطی مسئله‌ی انتخاب مدل و انتخاب متغیر مطرح می‌شود.

انتخاب مدل بر پایه‌ی توانایی پیشگویی ترجیحاً به کمک اعتبارسنجی با حذف یک‌به‌یک (LOO-CV[1]) [۱] و یا معیار اطلاع پرکاربرد (WIAC[2]) [۲] انجام می‌شود. هردوی این‌ها به‌عنوان ارائه دهنده‌ی برآوردهای به‌طور تقریبی ناریب برای توانایی پیشگویی یک مدل داده شده، شناخته شده‌اند. از جمله‌ی معیارهای دیگر پیشگویی نیز با توابع زیان مختلف می‌توان به معیار اطلاع کیبش (DIC[3]) [۳] و اندازه‌های مختلف بر مبنای توان دوم خطای پیشگویی [۴، ۵، ۶] اشاره کرد. در گذشته استفاده از روش‌های بیزی به دلیل بهره بردن از الگوریتم‌های پیچیده برای رسیدن به نتیجه، سخت و بعضاً ناممکن بود؛ اما در قرن بیست و یکم پیشرفت تکنولوژی و نرم‌افزارهای کاربردی پیاده‌سازی این الگوریتم‌ها را با دقت بالاتری ممکن کرده است؛ به همین دلیل شاهد افزایش روز افزون استفاده از روش‌های بیزی در سال‌های اخیر هستیم. در مسائل پیشگویی، کمیت کلیدی برآمده از نظریه بیز، توزیع پیشگوی پسین بوده که همان توزیع مشاهدات استفاده نشده در مدل‌بندی به‌شرط داده‌های موجود برای برازش مدل است.

یک روش نسبتاً مطلوب در ادبیات بیزی برای انتخاب یک مدل، روش ماکسیمم احتمال پسین است. [۷] نشان دادند که استفاده از یک پیشین یکنواخت بر روی فضای مدل در این روش،

1- Leave-One Out Cross-Validation
2- Widely Applicable Information Criterion
3- Deviance Information Criterion

بیشینه‌سازی درستنمایی حاشیه‌ای را نتیجه می‌دهد. همچنین در زمینه‌ی انتخاب متغیرهای تبیینی، احتمال‌های حاشیه‌ای نیز مورد استفاده قرار می‌گیرند. [۸] و [۹] به‌صورت تفصیلی در این مورد فعالیت کرده‌اند. تاکنون روش‌های بسیاری برای انتخاب مدل پیشگوی بیزی و ارزیابی آن ارائه شده که [۱۰] به مقایسه عملکرد آن‌ها پرداخته است

در این مقاله، به دنبال آن هستیم که عملکرد روش ماکسیمم احتمال پسین در انتخاب مدل را در حالت‌های فراوانی‌گرا (توزیع پیشین تخت) و مورد سنجش و ارزیابی قرار دهیم. به این منظور، در بخش ۲، مطلوبیت مدل و برآورد آن را به‌عنوان رویکردی برای سنجش توانایی پیشگویی مدل و مقایسه‌ی عملکرد روش‌های انتخاب مدل، در نظر گرفته‌ایم. همچنین نقصانی در برآورد مطلوبیت به نام بیش‌برازش مطرح شده است که وجود آن باعث غیرقابل اعتماد بودن برآورد مطلوبیت می‌شود [۱۰].

در بخش ۳، روش ماکسیمم احتمال پسین، به‌عنوان روش انتخاب مدل مورد استفاده در این مقاله مطرح شده و از مدل استاندارد گاوسی برای مطالعات تحلیلی و عددی استفاده خواهد شد. همچنین، برآوردی برای مطلوبیت مدل براساس توزیع پسین بردار ضرایب رگرسیونی به‌صورت جبری ارائه خواهد شد. در بخش ۴، به کمک مطالعه‌ی شبیه‌سازی، عملکرد روش ماکسیمم احتمال پسین برپایه‌ی تعداد متغیرهای کمکی انتخاب شده، مطلوبیت و MSE در رهیافت‌های فراوانی‌گرا و بیزی مقایسه خواهد شد. به‌علاوه، یک مجموعه از داده‌های واقعی نیز با استفاده از روش ماکسیمم احتمال پسین، مورد تحلیل و بررسی قرار خواهد گرفت. در انتها نتیجه‌گیری و پیشنهادها ارائه می‌شود.

۲- مطلوبیت معیاری برای مقایسه‌ی مدل‌ها

همان‌طور که در بخش اول گفته شد، سنجش عملکرد (توانایی پیشگویی) مدل یکی از معیارهای قضاوت در مورد مناسب بودن مدل‌هاست. محاسبه‌ی مطلوبیت[1] مدل یکی از ابزارهای مقایسه‌ی مدل‌هاست. از دیگر سو، مطلوبیت مدل می‌تواند معیاری برای مقایسه‌ی عملکرد دو روش در انتخاب مدل باشد و روشی که مطلوبیت بالاتری داشته باشد می‌تواند به‌عنوان روش بهتر مدنظر قرار گیرد. توابع مطلوبیت مناسب در مسائل پیش‌بینی احتمال، توابع امتیاز لگاریتمی و تابع صفر-یکی هستند که خواص آن‌ها توسط [۱۱] و [۱۲] به‌صورت تفصیلی بررسی شده است.

1- Utility

۲-۱- مطلوبیت لگاریتمی

توانایی پیشگویی مدل در قالب تابعی لگاریتمی از مطلوبیت به نام **مطلوبیت لگاریتمی** ارزیابی می‌شود [۱۴]. یک استفاده‌ی غالب از توزیع پیشگویی مدل نامزد شده‌ی M، استفاده از تابع مطلوبیت به‌صورت زیر است:

$$u(M,\tilde{y}) = \ln p(\tilde{y}\,|D,M) \qquad (۱)$$

که در آن $\tilde{y}$ پاسخ‌های مشاهده نشده و D مجموعه‌ی داده‌های مورد استفاده در برازش مدل M است. ازآنجاکه $\tilde{y}$ نامعلوم است، پس $u(M,\tilde{y})$ قابل‌محاسبه نیست؛ بنابراین می‌توان از امید ریاضی آن نسبت به توزیع داده‌ها به‌صورت

$$\bar{u}(M) = E[\ln p(\tilde{y}|D,M)] = \int_R p(\tilde{y}) \ln p(\tilde{y}|D,M) d\tilde{y} \qquad (۲)$$

استفاده کرد که در آن $p(\tilde{y})$ نشان دهنده‌ی توزیع تولید داده‌ها است. با توجه به آنچه گفته شد (۲) را می‌توان به‌عنوان مطلوبیت تعمیم‌یافته‌ی عملکرد پیشگویی مدل M معرفی کرد. $p(\tilde{y})$ در برخی موارد در دسترس نیست و در صورت در دسترس بودن نیز در بسیاری موارد عبارت داخل انتگرال پیچیده بود و محاسبه‌ی انتگرال به‌صورت تحلیلی ممکن نیست؛ بنابراین، مسئله‌ی برآورد معیار (۲) به‌عنوان ملاکی برای مقایسه‌ی مدل‌ها مطرح می‌شود. در مقایسه‌ی دو مدل، مدلی با مطلوبیت بیشتر دارای عملکرد بهتری در نظر گرفته می‌شود. بدیهی است در مسئله‌ی انتخاب مدل از میان تمام مدل‌های کاندید، مدلی با مطلوبیت بیشتر می‌تواند انتخاب شود [۱۰].

۲-۲- برآورد مطلوبیت تعمیم‌یافته

فرض کنید D (متشکل از بردار مشاهداتی از متغیر پاسخ Y و ماتریس X شامل مشاهداتی از متغیرهای کمکی متناظر) مجموعه داده‌های آموزشی[1] باشد که مدل M بر آن‌ها برازش داده شده است. همچنین $\tilde{y}^* = (\tilde{y}_1^*, \ldots, \tilde{y}_n^*)$ و X^* مجموعه داده‌های آزمون[2] (داده‌هایی که توانایی پیشگویی مدل را با آن‌ها ارزیابی می‌کنند) را تشکیل دهند. با این مفروضات، به‌منظور برآورد مطلوبیت تعمیم‌یافته، [۱۰] میانگین لگاریتم چگالی پیشگو ($MLPD$[3]) را به‌صورت

$$MLPD(M) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} ln\left(P(\tilde{y}_i^* \,|x_i^*, D, M)\right), \qquad (۳)$$

1- Training Data
2- Test Data
3- Mean of Logarithm of Predictive Density

ارائه کردند که در آن X_i^* سطر iام ماتریس X^* است. از (۳) این‌گونه برداشت می‌شود که [۱۰] این برآورد را براساس تقریب‌های مونت‌کارلویی از توزیع پسین مدل (به‌منظور محاسبه‌ی یک تقریب برای $\overline{u}(M)$) ارائه کرده‌اند. با این توضیح که برای محاسبه این برآورد، مقدار مشاهدات موردنظر را در چگالی پسین تقریب زده شده توسط الگوریتم‌های نمونه‌گیری وابسته به روش‌های مونت‌کارلویی (مانند روش گیبز) قرار داده و سپس میانگین لگاریتم مقادیر چگالی به‌دست‌آمده را به‌عنوان تقریبی از مطلوبیت تعمیم‌یافته (۲) استفاده می‌کنند.

۲-۳- بیش‌برازش

روش‌های زیادی می‌توانند برای برآورد مطلوبیت در نظر گرفته شوند. اعتبارسنجی [۱۱]، MLPD و ... . در این بخش به معرفی مفهومی به نام بیش‌برازش که حاصل از واریانس زیاد برآوردگر بوده و باعث نقصان در عملکرد برآوردهای مطلوبیت می‌شود، اشاره خواهیم کرد. در حضور واریانس بالا بیشینه‌سازی برآورد مطلوبیت ممکن است منجر به انتخاب مدلی شود که به‌صورت معنی‌داری از مدل بهینه دور باشد. به پدیده‌ی انتخاب مدل نابهینه در روش‌های برآورد مطلوبیت به دلیل واریانس بالا، **بیش‌برازش** در انتخاب مدل گفته می‌شود. می‌توان گفت که این پدیده باعث نقصان در برآورد مطلوبیت می‌شود، بنابراین انتظار می‌رود که مدل انتخاب شده مطلوبیتی نابهینه داشته باشد. برای مشاهده مثال و توضیحات بیشتر به [۱۳] مراجعه شود.

۳- روش انتخاب مدل موردبررسی

در مسائل انتخاب مدل رویکردها و دیدگاه‌های مختلفی وجود دارند که [۱۴] آن‌ها را بررسی کرده‌اند. یکی از این دیدگاه‌ها مربوط به روش‌هایی می‌شود که در آن مجموعه‌ی زیر مدل‌های[1] ممکن (از مدلی تنها شامل عرض از مبدأ آغاز و به مدل کامل ختم می‌شود) متناهی است و فرض می‌کنیم یکی از مدل‌های موردنظر مدل درست تولید داده باشد. روش ماکسیمم احتمال پسین که موردنظر ما در این مقاله است؛ در این دسته جای می‌گیرد.

۳-۱- انتخاب مدل به روش ماکسیمم احتمال پسین

رهیافت بیزی راه‌کاری برای توصیف عدم قطعیت با توجه به مشخصه‌های مدل مورد استفاده فراهم می‌کند. به‌شرط یک فهرست جامع از مدل‌های موردنظر، $\{M_\ell\}_{\ell=1}^{L}$ و توزیع بر روی فضای مدل می‌توان نوشت:

1- Sub-models

$$p(M|D) \propto p(D|M)p(M), \qquad (۴)$$

بنابراین پیشگویی‌ها می‌توانند از طریق روش متوسط‌گیری مدل بیزی (BMA [۱]) به‌صورت:

$$p(\tilde{y}|D) = \sum_{\ell=1}^{L} p(\tilde{y}|D.M_\ell)\, p(M_\ell|D),$$

در اینجا فرض می‌کنیم یکی از مدل‌های موردنظر مدل واقعی تولید داده‌ها است. از دیدگاه انتخاب مدل می‌توان مدلی را انتخاب کرد که (۴) را بیشینه کند که به آن مدل **ماکسیمم احتمال پسین** (MAP [۲]) می‌گوییم. در ادامه برای انتخاب متغیر و به دنبال آن انتخاب مدل برای احتمال وقوع مدل‌ها این روش را در نظر می‌گیریم و با استفاده از احتمالات پسین حاصل مدل موردنظر را انتخاب می‌کنیم. تحت توزیع پیشین $p(M) \propto 1$ ، مسئله به بیشینه‌سازی تابع درستنمایی حاشیه‌ای در روش فراوانی‌گرا تقلیل خواهد یافت.

۳-۲- مدل رگرسیونی موردبررسی

در این مقاله و به‌منظور تحلیل‌های نظری و مطالعات عددی، مدل گاوسی استاندارد

$$\begin{aligned} Y|X,\beta,\sigma^2 &\sim N(X\beta,\sigma^2), \\ \beta|\sigma^2 &\sim N(0.\sigma^2 I). \end{aligned} \qquad (۵)$$

مطالعه می‌شود که در آن Y متغیر پاسخ، X ماتریسی شامل متغیرهای کمکی بوده، β بردار ضرایب و σ^2 واریانس نوفه و پارامتری ثابت است. عبارت عرض از مبدأ را به‌واسطه‌ی کسر کردن مقدار میانگین هریک از ستون‌های X از آن، کنار می‌گذاریم؛ بنابراین بردار ضرایب رگرسیونی در مدل (۵) به‌صورت $\beta = (\beta^1, \beta^2, ..., \beta^p)$ است که p در آن تعداد متغیرهای کمکی است.

۳-۳- برآورد مطلوبیت برای مدل گاوسی استاندارد

با توجه به آنچه در بخش دوم درباره‌ی برآورد مطلوبیت ارائه شده توسط [۱۰] مطرح شد، می‌توان گفت که با داشتن توزیع پسین بردار ضرایب رگرسیونی می‌توان رابطه‌ی (۳) را به‌صورت جبری و نه به کمک تقریب‌هایی مانند روش‌های مونت‌کارلویی حساب کرد. در ادامه با ارائه‌ی یک لم و قضیه‌ی بعد از آن، به‌صورت جبری میانگین لگاریتم چگالی پیشگو را به‌عنوان برآورد مطلوبیت،

1- Bayesian Model Averaging
2- Maximum a Posteriori

برای مدل گاوسی استاندارد ارائه خواهیم کرد که تاکنون تنها به‌صورت تقریب عددی محاسبه شده و مورد استفاده بوده است.

لم ۳.۱- فرض کنید $\underset{\sim}{y} = y_1, \ldots y_n$ و $X_{n\times p}$ داده‌های آموزشی برای برازش مدل گاوسی استاندارد (۵) باشد؛ آنگاه توزیع پسین بردار وزن β به‌صورت زیر است:

$$\beta \Big| \underset{\sim}{y}, X, \sigma^2 \sim N_p(\beta_p, \sigma^2 \sum_p^{-1})$$

که در آن $\beta_p = (X^T X + I)^{-1}(X^T y)$ و $\sum_p = (X^T X + I)$.

اثبات. براساس فرض مسئله برای توزیع پیشین بردار β داریم:

$$\beta \sim N_p(0, \sigma^2 I_{p\times p}) \Rightarrow \pi(\beta \Big| \sigma^2) \propto exp\{-\frac{1}{2\sigma^2}(\beta^T \beta)\}$$

همچنین برای توزیع پسین بردار β می‌توان نوشت:

$$\pi(\beta \mid \underset{\sim}{y}, X, \sigma^2) \propto f(\underset{\sim}{y} \Big| \beta, X, \sigma^2)\ \pi(\beta \Big| \sigma^2)$$

بنابراین با توجه به توزیع مشاهدات حاصل از متغیر پاسخ $\underset{\sim}{y}$ می‌توان نوشت:

$$\pi(\beta \mid \underset{\sim}{y}, X, \sigma^2.) \propto exp\{-\frac{1}{2\sigma^2}(\underset{\sim}{y} - X\beta)^T(\underset{\sim}{y} - X\beta)\} exp\{-\frac{1}{2\sigma^2}(\beta^T \beta)\}$$

فرمایر و کنایب [۱۵] (فرمول‌های ۳.۴۰ و ۳.۴۱) نشان دادند که عبارت حاضر در توان را می‌توان به‌صورت زیر به فرم درجه دوم از بردار β نوشت:

$$(\underset{\sim}{y} - X\beta)^T(\underset{\sim}{y} - X\beta) + (\beta^T \beta) = (\beta - \beta_p)^T \sum_p (\beta - \beta_p) + \underset{\sim}{y}^T \underset{\sim}{y} + \beta_p{}^T \sum_p \beta_p$$

در نتیجه برای توزیع پسین بردار β داریم:

$$\pi(\beta \Big| \underset{\sim}{y}, X, \sigma^2) \propto exp\{-\frac{1}{2\sigma^2}\Big((\beta - \beta_p)^T \sum_p (\beta - \beta_p) + \underset{\sim}{y}^T \underset{\sim}{y} + \beta_p{}^T \sum_p \beta_p\Big)\}$$

$$\propto exp\{-\frac{1}{2\sigma^2}\Big((\beta - \beta_p)^T \sum_p (\beta - \beta_p)\Big)\}$$

بنابراین می‌توان نتیجه گرفت:

$$\beta \Big| \underset{\sim}{y}, X, \sigma^2 \sim N_p(\beta_p, \sigma^2 \sum_p^{-1})$$ □

اکنون با بهره‌گیری از مفروضات و نتایج لم قبل که توزیع پسین بردار ضرایب مدل رگرسیونی را در اختیار می‌گذارد؛ در قالب یک قضیه، برآورد مطلوبیت را برای موردمطالعه ارائه می‌شود

قضیه ۳.۲- فرض کنید $\tilde{y}^* = (\tilde{y}_1^*,...,\tilde{y}_{\tilde{n}}^*)$ و $X^*_{\tilde{n}\times p}$ داده‌های آزمون برای سنجش عملکرد مدل M باشند. میانگین لگاریتم چگالی پیشگو $MLPD$ برای مدل M به‌صورت

$$MLPD^*(M) = -\frac{1}{\tilde{n}}\left(\sum_{i=1}^{\tilde{n}}\frac{1}{\gamma_i^2}(\tilde{y}_i^* - \mu_i^*)^2 + ln(\gamma_i)\right), \qquad (۶)$$

محاسبه می‌شود که در آن μ_i^* در واقع i امین عضو بردار $X\beta_p$ و γ_i^2 عضو i ام قطر اصلی ماتریس $X\sum_p^{-1}X^T$ است (بردار $X\beta_p$ و ماتریس $\sum_p^{-1}$ از لم قبل به‌وسیله‌ی داده‌های آموزشی برای برازش مدل M به‌دست آمده‌اند).

اثبات. با توجه به حکم لم قبل داریم:

$$\beta_{posterior} = \beta \Big| \underset{\sim}{y}, X, \sigma^2 \sim N_p(\beta_p, \sigma^2\sum_p^{-1}),$$

از طرفی $\eta^* = E(\mathrm{Y}^*_{posterior}\,|X,\mathrm{D},\mathrm{M}) = X^*\beta_{posterior}$ را به‌عنوان پیشگوی (خطی) مدل M تعریف می‌کنیم. η^* ترکیب خطی از یک توزیع نرمال است؛ پس خود نیز از توزیع نرمال تبعیت خواهد کرد که میانگین و واریانس آن به‌صورت

$$E(\eta^*) = E(X^*\beta_{posterior}) = X^*E(\beta_{posterior}) = X^*\beta_p$$

و

$$Var(\eta^*) = Var(X^*\beta_{posterior}) = X^*Var(\beta_{posterior})X^{*T} = \sigma^2 X\sum_p^{-1}X^{*T}$$

از طرفی دیگر، توزیع حاشیه‌ای یک نرمال چند متغیره، خود توزیع نرمال تک متغیره خواهد شد با میانگین و واریانسی متناظر با اعضای بردار میانگین و قطر اصلی ماتریس واریانس-کوواریانس توزیع چند متغیره‌ی اصلی یا به عبارت دیگر:

$$\eta_i^* \sim N\left(\left(X^*\beta_p\right)_i, \left(\sigma^2 X^*\sum_p^{-1}X^{*T}\right)_{ii}\right)$$

که مطابق فرض مسئله می‌توان میانگین و واریانس آن‌ها را با μ_i^* و γ_i^2 نمایش داد. مطابق رابطه (۳)، توابع چگالی η_i^* ها را به‌عنوان چگالی توزیع پیشگوی مدل در نظر می‌گیریم. درباره‌ی چگالی پیشگو برای هر η_i^* داریم:

$$\ln\left(P_{\eta_i^*}(\tilde{y}_i^* \mid \tilde{x}_i^*, D, M)\right) = \ln\left(\frac{1}{\gamma_i\sqrt{2\pi}}\exp\{-\frac{1}{2\gamma_i^2}(\tilde{y}_i^* - \mu_i^*)^2\}\right)$$

$$= -\ln(\gamma_i) - \ln(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2\gamma_i^2}(\tilde{y}_i^* - \mu_i^*)^2$$

به این دلیل که در مسئله‌ی بررسی عملکرد مدل براساس مطلوبیت، مدلی با بیشینه برآورد مطلوبیت انتخاب می‌شود؛ بنابراین عبارت ثابت $\ln(\frac{1}{\sqrt{2\pi}})$ تأثیری در پاسخ نداشته و قابل چشم‌پوشی است. با توجه به آنچه تاکنون به‌دست آمده است؛ میانگین لگاریتم چگالی پیشگو برای مدل گاوسی استاندارد به‌صورت زیر نوشته می‌شود:

$$MLPD^*(M) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} ln\left(P_{\eta_i^*}(\tilde{y}_i^* \mid \tilde{x}_i^*, D, M)\right)$$

$$= -\frac{1}{n}(\sum_{i=1}^{n}\frac{1}{\gamma_i^2}(\tilde{y}_i^* - \mu_i^*)^2 + ln(\gamma_i)) \qquad \square$$

شایان‌ذکر است که بر اساس رویه‌ی لم و قضیه قبل می‌توان بررسی کرد که اگر $\pi(\beta) = 1$، آنگاه:

$$\beta_{posterior} = \beta \mid \tilde{y}^*, X, \sigma^2 \sim N_p((X^{*T}X^*)^{-1}X^{*T}\tilde{y}^*, \sigma^2(X^{*T}X^*)^{-1}),$$

این توزیع پسین به‌دست آمده برای بردار ضرایب رگرسیونی، تبیین کننده‌ی این اصل است که در مسائل بیزی، با قرار دادن توزیع پیشین تخت (ناآگاهی بخش) بر روی پارامتر، به مسئله‌ای در حالت فراوانی‌گرا می‌رسیم. بنابر آنچه مطرح شد؛ مقدار برآورد مطلوبیت مدل M، برای حالت فراوانی‌گرا می‌تواند به‌صورت زیر محاسبه شود:

$$MLPD^*(M) = -\frac{1}{\tilde{n}}\left(\sum_{i=1}^{n}\frac{1}{\theta_i^2}(\tilde{y}_i^* - \hat{y}_i)^2 + ln(\theta_i)\right)$$

که در آن $\hat{y}_i = \left(X^*(X^TX)^{-1}X^Ty\right)_i$ و $\theta_i = \sigma^2\left(X^*(X^TX)^{-1}X^{*T}\right)_{ii}$. در ادامه، روش ماکسیمم احتمال پسین را در قالب مطالعات عددی بر پایه‌ی میانگین توان دوم خطا و برآورد مطلوبیت ارائه شده در این بخش ارزیابی می‌شود.

۴- مطالعات عددی

در این بخش، روش ماکسیمم احتمال پسین برای حالت بیزی و فراوانی‌گرا را در قالب آزمایش‌های عددی برپایه‌ی مطلوبیت مدل و MSE، مورد ارزیابی قرار خواهیم داد. ابتدا در بخش اول مثال‌هایی تشریحی با استفاده از داده‌های شبیه‌سازی شده از مدل گاوسی استاندارد را موردبررسی قرار خواهیم داد. سپس در قسمت بعد، داده‌های واقعی را برای بررسی عملکرد دو روش در قبال مدل استاندارد گاوسی و محاسبه‌ی توانایی مدل نهایی انتخاب شده به‌کار خواهیم برد. محاسبات آماری در این بخش با استفاده از نرم‌افزار R انجام شده است.

۴-۱- شبیه‌سازی مسئله‌ی رگرسیونی

به‌منظور انجام بررسی‌ها و مطالعات عددی، ابتدا یک مسئله‌ی انتخاب مدل شبیه‌سازی شده را معرفی می‌کنیم که در آن تعدادی از مفاهیم مهم و اساسی درباره‌ی روش انتخاب مدل موردبررسی تشریح شده است. مدل

$$X \sim N(0, R), R \in \mathbb{R}^{p\times p},$$
$$Y \,|\, X = \underset{\sim}{x} \sim N(\beta^T \underset{\sim}{x}, \sigma^2)\ ,\ \sigma^2 = 1,$$

را درنظر بگیرید. تعداد متغیرهای کمکی را $p = 30$ قرار می‌دهیم. متغیرها در گروه‌های ۳ تایی دسته‌بندی شده‌اند. هر متغیر X_j دارای توزیع نرمال استاندارد است؛ همچنین این متغیر با هر یک از متغیرهای گروه خود دارای همبستگی با ضریب ρ بوده و نسبت به متغیرهای سایر گروه‌ها ناهمبسته است؛ در واقع ماتریس واریانس-کوواریانس R ماتریسی قطری-بلوکی به‌صورت زیر است:

$$R = \begin{bmatrix} r & o & o & \cdots & o \\ o & r & o & \cdots & o \\ \vdots & & \ddots & & \\ o & o & o & \cdots & r \end{bmatrix}_{10\times 10}$$

که در آن:

$$r = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}_{3\times 3}$$

و $O_{۳\times۳}$ ماتریسی مربعی با ابعادی مشابه r با درایه‌های صفر است. همچنین بردار β که تبیین کننده‌ی وزن متغیرهای کمکی است به‌صورت

$$\beta^T = (\xi,\xi,\xi,\frac{\xi}{۲},\frac{\xi}{۲},\frac{\xi}{۲},\frac{\xi}{۴},\frac{\xi}{۴},\frac{\xi}{۴},O^T)$$

بوده که در آن $O = [۰]_{۲۱\times۱}$ و این بدان معناست که تنها ۹ متغیر کمکی مرتبط با تولید متغیر پاسخ در مسئله بوده و سایر متغیرها دارای وزنی برابر صفر هستند. ثابت ξ سیگنال نرخ نوفه در داده‌ها را تنظیم می‌کند. برای رسیدن به نتایج قابل‌مقایسه به ازای مقادیر مختلف ρ و ξ را به‌گونه‌ای قرار می‌دهیم که رابطه‌ی

$$\frac{\sigma^۲}{Var(Y)} = ۰/۲۵ \qquad (۷)$$

برقرار باشد. برای به دست آوردن مقادیر ξ طبق رابطه‌ی (۷) داریم:

$$\frac{\sigma^۲}{Var(Y)} = \frac{\sigma^۲}{Var(\beta^T X+\varepsilon)} = \frac{\sigma^۲}{\beta^T R\beta+\sigma^۲} = ۰/۲۵. \qquad (۸)$$

که در آن

$$R\beta = \xi(۱+۲\rho)(۱,۱,۱,\frac{۱}{۲},\frac{۱}{۲},\frac{۱}{۲},\frac{۱}{۴},\frac{۱}{۴},\frac{۱}{۴},O)^T.$$

در نتیجه خواهیم داشت:

$$\beta^T R\beta = ۳\xi^۲(۱+۲\rho)+\frac{۳\xi^۲(۱+۲\rho)}{۴}+\frac{۳\xi^۲(۱+۲\rho)}{۱۶}, \qquad (۹)$$

با توجه به اینکه می‌دانیم $\sigma^۲ = ۱$ است؛ با جایگذاری (۹) در رابطه‌ی (۸) معادله‌ای به‌صورت زیر خواهیم داشت:

$$\frac{۱}{\beta^T R\beta+۱} = \frac{۱}{۳\xi^۲(۱+۲\rho)(۱+\frac{۱}{۴}+\frac{۱}{۱۶})+۱} = ۰/۲۵ \qquad (۱۰)$$

با بازنویسی رابطه‌ی (۱۰) برحسب ξ خواهیم داشت:

$$\xi^۲ = \frac{۱}{(۱/۳۱۲۵)(۱+۲\rho)}. \qquad (۱۱)$$

با قرار دادن مقادیر ρ = ۰،۰/۵،۰/۹ در معادله‌ی (۱۱)، مقادیر ξ = ۰/۸۷، ۰/۶۲، ۰/۵۲ طوری به‌دست می‌آیند که رابطه‌ی (۷) برقرار باشد.

آزمایش‌ها با تغییر دادن اندازه‌ی داده‌های آموزشی n =۱۰۰،۲۰۰،۴۰۰ و مقادیر ضرایب همبستگی ذکر شده انجام می‌شوند. بدین ترتیب، ۵۰ تکرار از هریک از اندازه‌های گفته شده به‌عنوان داده‌های آموزشی تولید می‌کنیم و با استفاده از آن‌ها روش ماکسیمم احتمال پسین را برای انتخاب مدل به‌کار می‌بریم. به‌عنوان روش انتخاب متغیر، روش استاندارد پیشرو را برمی‌گزینیم که در آن از مدلی با تک متغیر کمکی که بین تمام مدل‌های مشابه خود بیشترین احتمال پسین را دارد؛ شروع کرده و سپس در هر گام متغیری که بیشترین افزایش در احتمال پسین را نتیجه می‌دهد به مدل خواهیم افزود. سپس مدل‌های به‌دست‌آمده بر روی یک مجموعه داده‌ی آزمون مستقل ۱۰۰۰تایی با محاسبه‌ی MSE و مطلوبیت سنجیده می‌شوند.

به‌عنوان برآورد مطلوبیت تعمیم‌یافته، از میانگین لگاریتم چگالی پیشگو (۶) که صورت تحلیلی آن در بخش سوم ارائه شد؛ استفاده می‌کنیم. به‌منظور کاهش واریانس ناشی از استفاده از داده‌های مختلف، مطلوبیت زیرمدل‌های انتخاب شده‌ی M را با در نظر گرفتن مدل مرجع M_* (مدل کامل که شامل تمام متغیرهای کمکی حاضر در مسئله است) به‌صورت زیر گزارش می‌کنیم:

$$\Delta MLPD^*(M) = MLPD^*(M) - MLPD^*(M_*). \qquad (۱۲)$$

مقدار صفر در این معیار، مشخص کننده‌ی این است که عملکرد مدل M با مدل مرجع یکی است. مقدار منفی بیانگر عملکرد بهتر مدل مرجع و مقدار مثبت بیانگر عملکرد بهتر مدل M است.

جدول (۱): میانگین تعداد متغیرهای کمکی انتخاب شده به روش ماکسیمم احتمال پسین به ازای ۵۰ تکرار (مقدار داخل پرانتز انحراف معیار تعداد متغیرهای انتخاب شده است).

روش	بیزی			فراوانی‌گرا		
ρ / n	۱۰۰	۲۰۰	۴۰۰	۱۰۰	۲۰۰	۴۰۰
۰	۱۵/۴ (۳/۸)	۱۵/۷ (۳/۷)	۱۷/۵ (۴/۱)	۱۲/۱ (۲/۸)	۱۲/۸ (۱/۷)	۱۲/۷ (۲/۱)
۰/۵	۱۱/۹ (۳/۶)	۱۴/۳ (۳/۳)	۱۳/۸ (۴/۲)	۱۰/۸ (۲/۵)	۱۲ (۲/۲)	۱۱/۵ (۱/۶)
۰/۹	۹ /۳ (۲/۵)	۱۱/۲ (۳/۵)	۱۲ (۳/۴)	۷/۳ (۲)	۷/۴ (۱/۸)	۸/۴ (۱/۶)

جدول ۱ متوسط تعداد متغیرهای کمکی انتخاب شده به روی ۵۰ تکرار را به‌ازای مقادیر همبستگی مختلف میان متغیرهای کمکی و اندازه نمونه‌های مختلف نشان می‌دهد. همان‌گونه که مشاهده می‌شود. با افزایش همبستگی تعداد متغیرهای کمکی انتخابی کمتر و همچنین انحراف معیار کاهش می‌یابد. با افزایش حجم نمونه، تعداد متغیرهای انتخاب شده نیز افزایش می‌یابد. در واقع هردو رهیافت که مبتنی بر رگرسیون خطی هستند، با افزایش حجم نمونه مایل به انتخاب متغیرهای بیشتری هستند. با توجه به نتایج حاصل از جدول برای دو حالت فراوانی‌گرا و بیزی، مشاهده می‌شود که روش فراوانی‌گرا تعداد متغیرهای کمکی کمتری نسبت به روش بیزی در سطوح متناظر دارد. انتخاب تعداد متغیر کمتر لزومی بر برتری نیست، اگرچه می‌تواند نقطه‌ی قوتی برای یک روش انتخاب مدل باشد. در واقع، این نکته هنگامی مزیت است که قدرت پیش‌بینی مدل‌های انتخاب شده نیز در مقایسه با حالت بیزی کمتر باشند. در ادامه با استفاده از میانگین توان دوم خطا و همچنین برآورد مطلوبیت به بررسی قدرت پیش‌بینی مدل‌ها در دو حالت فراوانی‌گرا و بیزی می‌پردازیم.

جدول ۲ میانگین MSE مدل‌ها را براساس ۵۰ تکرار نشان می‌دهد. این مقادیر با استفاده از داده‌های آزمون محاسبه شده و این مزیت را دارد که MSE مدل‌های گوناگون را با تنها یک مجموعه داده محاسبه کرده است؛ بنابراین این مقادیر قابل‌مقایسه بوده و هرچه به صفر نزدیک‌تر باشد بیانگر بهتر بودن عملکرد روش و اندازه‌های مختلف است.

جدول (۲): مقادیر مربوط به MSE در روش ماکسیمم احتمال پسین.

روش	بیزی			فراوانی‌گرا		
ρ / n	۱۰۰	۲۰۰	۴۰۰	۱۰۰	۲۰۰	۴۰۰
۰	۱/۲۴۸	۱/۰۹۹	۱/۰۳۲	۱/۳۴۶	۱/۱۱۵	۱/۰۳۳
۰/۵	۱/۶۲۹	۱/۴۶۲	۱/۳۷۱	۱/۷۱۱	۱/۵۵۹	۱/۵۱۲
۰/۹	۲/۰۰۵	۱/۸۳۱	۱/۶۹۵	۲/۴۵۹	۲/۰۱۲	۱/۸۸۳

براساس نتایج موجود در جدول ۲، می‌توان گفت که MSE در حالت بیزی کمتر از حالت فراوانی گراست و با افزایش همبستگی میان متغیرهای کمکی، به‌وضوح MSE افزایش می‌یابد. به‌عنوان مثال، وقتی $n = ۱۰۰$ باشد؛ در حالت بیزی اختلاف MSE بین ضرایب همبستگی صفر و ۰/۹ تقریباً ۰/۸ است حال آنکه برای $n = ۴۰۰$ این مقدار تقریباً ۰/۶۵ است. همچنین به‌صورت عمومی و در هردو حالت فراوانی گرا و بیزی با افزایش اندازه‌ی مجموعه داده‌ی آموزشی برای برازش مدل، مقدار MSE نیز کاهش می‌یابد.

شایان‌ذکر است که در عمل داده‌هایی که کاملاً مستقل باشند به‌ندرت یافت می‌شوند. همین مسئله برای داده‌هایی با همبستگی خیلی زیاد (۰/۹) نیز صادق است؛ با توجه به دلایل مذکور، در ادامه برآورد مطلوبیت را برای همبستگی $\rho = 0/5$ بررسی می‌کنیم که نزدیکی بیشتری با مسائل واقعی دارد.

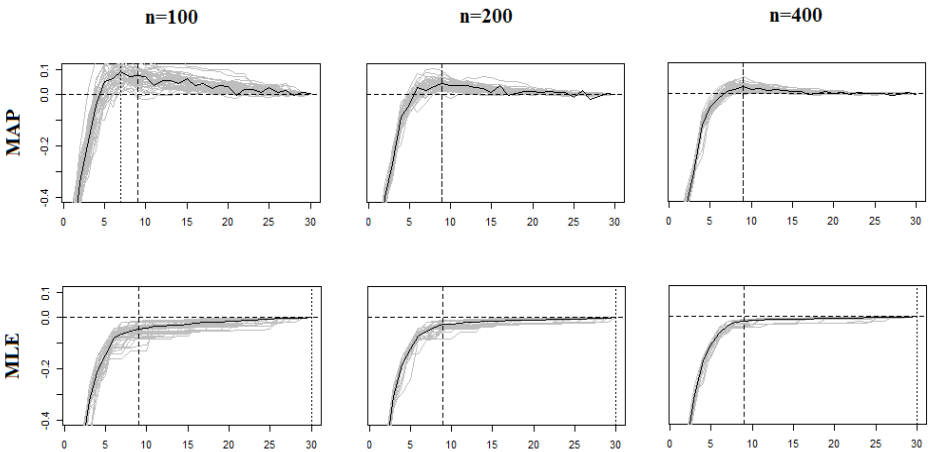
پیش از ادامه‌ی تحلیل‌ها و تفسیر نتایج، نیاز به معرفی مفهومی به نام مسیر جستجو داریم. در رویه‌ی انتخاب متغیر پیشرو، متغیرها به ترتیب اثرگذاری در مدل با توجه به ملاک موردبررسی (در اینجا احتمال پسین) وارد مدل شده و مرتب می‌شوند؛ اما این ورود متغیرها در نقطه‌ای که مدلی با تعداد متغیر بهینه از دید ملاک موردنظر (احتمال پسین) انتخاب شده است متوقف می‌شود؛ زیرا مدل‌های بعدی با یک یا چند متغیر بیشتر عملکرد بهتری نسبت به مدل فعلی ندارند. اکنون فرض کنید که این شرط توقف ورود متغیرها برداشته شده و مدل‌های بعدی با متغیرهای کمکی اضافه شده تنها بین مدل‌هایی با شرایط مشابه (یعنی مدل‌هایی که متغیرهای کمکی قبل از آن یکسان باشند) بهترین عملکرد را داشته باشند. در چنین شرایطی تمامی متغیرهای کمکی موجود با چینش جدیدی مرتب شده‌اند. این ترتیب جدید متغیرهای کمکی تشکیل یک مسیر جستجو را می‌دهند.

شکل ۱ نمایش‌دهنده‌ی تغییرات برآورد مطلوبیت با حرکت بر روی مسیرهای جستجو و به‌تبع آن افزایش تعداد متغیرهای کمکی حاضر در مدل است. در این شکل، خطوط خاکستری نشان دهنده‌ی مطلوبیت به‌ازای هر یک از تکرارهای ۵۰ گانه است و خطوط مشکی میانگین این ۵۰ تکرار را بیان می‌کنند. همان‌گونه که در شکل ۱ مشاهده می‌شود، هردو حالت فراوانی‌گرا و بیزی، با افزایش اندازه‌ی داده‌های آموزشی کمتر در معرض بیش‌برازش قرار می‌گیرد (واریانس کمتر می‌شود). همچنین در حالت بیزی، برآورد مطلوبیت در جایی بیشینه می‌شود که مطابق رابطه‌ی (۱۲) که مطلوبیت مدل منتخب از مدل مرجع بیشتر بوده و درواقع پیشگویی بهتری را در مقایسه با مدل مرجع ارائه می‌دهد. نکته‌ی دیگر این است که در حالت فراوانی‌گرا، برآورد مطلوبیت مدل مرجع همواره بیشتر بوده و خطوط مشکی رنگ نیز بیشینه‌ی خود را در انتهای بازه اختیار می‌کنند که این یعنی پیشنهاد حالت فراوانی‌گرا مدل کامل است.

شایان‌ذکر است که نتایج به‌دست‌آمده به‌وسیله‌ی MLPD^{*} (۶)، برتری محسوسی نسبت به مطالعات عددی مشابه که توسط [۱۰] انجام شده است، دارد. نتایج حاصل در [۱۰] برای روش ماکسیمم احتمال پسین تبیین کننده‌ی آن است که مدل‌های انتخابی مطلوبیت بیشتری نسبت به مدل مرجع ندارند. همچنین برآورد عددی استفاده شده در [۱۰]، منجر به پدیده‌ی کم‌برازش[1]

1- Under-fitting

و انتخاب مدل‌هایی شده است که تعداد متغیرهای حاضر در آن‌ها از تعداد متغیرهای مرتبط نیز کمتر هستند. حال آن‌که با توجه به نتایج شکل ۱، در اغلب موارد برآورد مطلوبیت عیناً در مدلی با تعداد ۹ متغیر بیشینه شده است. با توجه به تعریف مسیر جستجو متغیرها، فارغ از مسئله‌ی انتخاب برخی متغیرها، تمامی متغیرهای کمکی به ترتیب اولویت ورود به مدل مرتب می‌شوند؛ بنابراین، نقطه‌ی بیشینه‌ی مطلوبیت در هر مسیر جستجو، لزوماً در نقطه‌ای که برابر تعداد متغیرهای انتخاب شده در مسئله‌ی انتخاب متغیراست، رخ نمی‌دهد. براساس توضیحات مطرح شده، در شکل ۱، برای مسیرهای جستجو مطلوبیت در نقاطی بیشینه شده است که مدل انتخاب شده مطلوبیت بیشتری از مدل مرجع دارد.

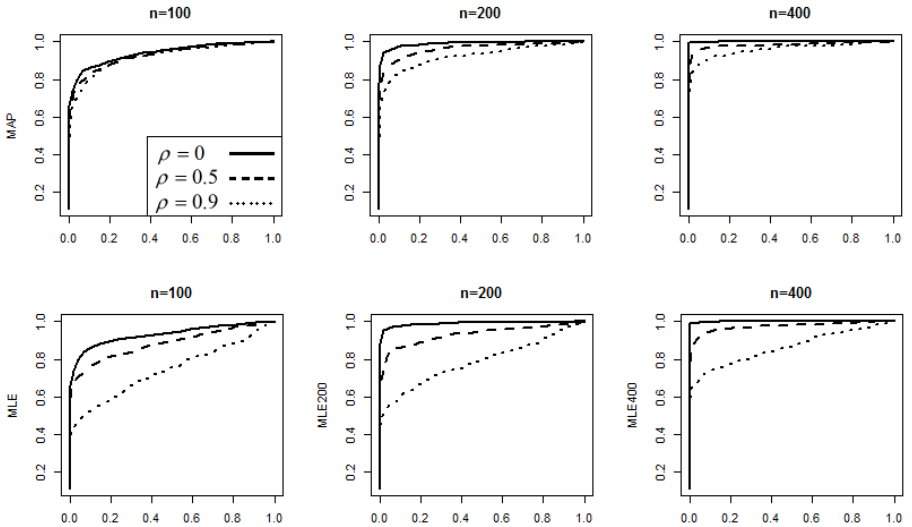


شکل (۱): پراکندگی برآورد مطلوبیت به روی مسیرهای جستجو (که در آن محور افقی تعداد متغیرهای حاضر در مدل و محور عمودی اختلاف برآورد مطلوبیت از مدل مرجع است. همچنین، خطوط عمودی خط‌چین تعداد متغیرهای کمکی مرتبط "عدد ۹" را نمایش داده و خطوط عمودی نقطه‌چین بیان‌گر نقطه‌ای هستند که خطوط مشکی بیشینه مقدار "بیشینه برآورد مطلوبیت" خود را اخذ کرده‌اند).

شکل ۲ نرخ ورود متغیرهای مرتبط (۹ ستون اول ماتریس X) به مدل در مقابل نرخ ورود متغیرهای نامربوط را در مسیرهای جستجو نشان می‌دهد. در مورد نحوه‌ی به دست آمدن منحنی‌های حاضر در شکل باید گفت که محور عمودی آن متوسط نرخ ورود متغیرهای مرتبط و محور افقی آن متوسط نرخ ورود متغیرهای نامرتبط به‌ازای ۵۰ تکرار موجود است. به‌طور مثال اگر مسیر جستجو به‌صورت $3,11,10,4,1,\ldots$ باشد آنگاه نرخ ورود متغیرهای مرتبط به‌صورت $\frac{1}{3},\frac{2}{9},\frac{1}{9},\frac{1}{9},\frac{1}{9},\ldots$ است. این نرخ با رسیدن به هر متغیر مرتبط در مسیر جستجو، به‌اندازه‌ی

$\frac{۱}{۹}$ افزایش می‌یابد و رسیدن به عدد ۱ به معنای این است که مدل در مسیر جستجو، به تمام متغیرهای مرتبط رسیده است. همچنین، هرچقدر دیرتر نرخ به عدد یک برسد، متغیرهای نامرتبط بیشتری توسط روش انتخاب مدل دارای اولویت حضور در مدل هستند.

به‌طور مشابه، نرخ ورود متغیرهای نامرتبط به‌صورت $\frac{۲}{۲۱},\frac{۲}{۲۱},\frac{۲}{۲۱},\frac{۱}{۲۱},\frac{۰}{۲۱},\ldots$ خواهد بود. این نرخ با رسیدن به هر متغیر نامرتبط در مسیر جستجو به‌اندازه‌ی $\frac{۱}{۲۱}$ افزایش می‌یابد و وقتی به یک برسد یعنی تمام متغیرهای نامرتبط در مدل حضور دارند. از دیدگاه نرخ‌های معرفی شده، برای یک مسیر جستجو هرچه سرعت ورود متغیرهای مرتبط به مدل‌بالاتر باشد و محور افقی زودتر به عدد ۱ نزدیک شود، گواهی بر عملکرد بهتر آن مسیر جستجو است.



شکل (۲): نسبت متغیرهای مرتبط انتخاب شده در مقابل نسبت متغیرهای نامرتبط.

از شکل ۲ این‌گونه برداشت می‌شود که به‌صورت عمومی، با افزایش تعداد داده‌های آموزشی برای برازش مدل، نرخ ورود متغیرهای مرتبط افزایش می‌یابد. همچنین با افزایش ضریب همبستگی گرایش به ورود متغیرهای مرتبط کاهش‌یافته و منحنی‌ها دیرتر به عدد ۱ می‌رسند. به‌علاوه می‌توان گفت که به‌طور محسوسی در حالت بیزی سرعت ورود متغیرهای مرتبط بیشتر است؛ تا جایی که برای $n = ۴۰۰$ و ضریب همبستگی صفر، تقریباً ابتدا متغیرهای مرتبط وارد می‌شوند و سپس باقی متغیرها در مسیرهای جستجو گنجانده می‌شوند.

به‌طور خلاصه، براساس نتایج حاصل از مطالعات شبیه‌سازی می‌توان گفت که روش ماکسیمم احتمال پسین در حالت بیزی عملکرد بهتری ازلحاظ تعداد متغیر انتخاب شده، MSE، برآورد مطلوبیت بر روی مسیرهای جستجو و نرخ ورود متغیرهای مرتبط دارد.

۴-۲- تحلیل داده‌های واقعی

در این بخش به بررسی عملکرد روش ماکسیمم احتمال پسین بر روی مجموعه داده‌ی واقعی جرم‌شناسی می‌پردازیم که در دسترس عموم قرار دارد. مجموعه داده‌های جرم‌شناسی[1]، به‌عنوان مجموعه داده‌ی واقعی در این مقاله مورد استفاده و تحلیل قرار می‌گیرد. متغیر پاسخ در این داده‌ها، سرانه‌ی انواع جرائم خشونت‌آمیز از قبیل قتل، تجاوز، سرقت مسلحانه و... در ایالات‌متحده است که به‌منظور تسهیل در مدل‌بندی رگرسیونی به‌صورت لگاریتمی نرمال‌سازی شده است. تعداد متغیرهای کمکی در این مجموعه داده ۱۰۲، اندازه‌ی نمونه ۱۹۹۲ و توزیع‌های پیشین موردنیاز در مدل به‌صورت $\beta_{\circ} \sim N(\circ,\Delta)$, $\underset{\sim}{\beta} \sim N_{۱۰۲}(\circ,\sigma^{۲}\mathrm{I})$ است که در آن منظور از $\hat{\sigma}^{۲}$ ، در واقع MSE مدل کامل است که با استفاده از برازش تمام داده‌ها به‌دست آمده است و به‌عنوان مدل مرجع در این بخش مورد استفاده قرار می‌گیرد.

جدول (۴): میانگین تعداد متغیرهای کمکی انتخاب شده به روش ماکسیمم احتمال پسین به ازای ۵۰ تکرار (مقدار داخل پرانتز انحراف معیار تعداد متغیرهای انتخاب شده است).

n	روش	
	بیزی	فراوانی‌گرا
۱۰۰	۱۱/۸۶ (۴/۲۵)	۲۴/۸۴ (۱۸)
۲۰۰	۱۱/۲ (۴/۵)	۱۷/۱ (۵/۷۷)
۴۰۰	۲۰/۵ (۶/۲)	۲۰/۸ (۵/۷)

برای مسئله‌ی رگرسیون خطی از مدل (۵) استفاده می‌کنیم. برای برآورد مطلوبیت مدل‌های انتخاب شده، فرایند انتخاب مدل را چندین بار تکرار می‌کنیم (مانند بخش شبیه‌سازی قسمتی از داده‌ها به‌عنوان داده‌ی آموزشی و قسمتی دیگر به‌عنوان داده‌ی آزمون مورد استفاده قرار می‌گیرد). در مجموعه داده‌های جرم‌شناسی، برای قسمت‌بندی به دو دسته‌ی داده‌های آموزشی و داده‌های آزمون نمونه به‌اندازه‌ی کافی بزرگ است. همانند بخش شبیه‌سازی هر بار ۵۰ نمونه به اندازه‌های ۱۰۰، ۲۰۰ و ۴۰۰ به‌عنوان داده‌های آموزشی تولید می‌کنیم. سپس مسئله‌ی انتخاب

1- https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

مدل را برای هرکدام از این مجموعه‌ها انجام می‌دهیم و از مابقی داده‌ها به‌عنوان داده‌های آزمون استفاده می‌کنیم. این کار اجازه می‌دهد که اثر اندازه‌های مختلف داده‌های آموزشی بر نتایج را نیز بررسی کنیم. در ادامه تحلیل نتایج به‌دست‌آمده از داده‌های واقعی را مشاهده می‌کنیم.

نتایج جدول ۴ حاکی از آن است که به‌طورکلی، با افزایش تعداد داده‌های آموزشی هم تعداد متغیرهای انتخابی و هم انحراف معیار آن‌ها روند خاصی ندارند. در واقع، برای حالت فراوانی‌گرا می‌توان گفت که با افزایش حجم نمونه، انحراف معیار کاهش می‌یابد. همچنین روش بیزی تمایل به انتخاب مدل‌هایی با تعداد متغیرهای کمکی کمتری در مقایسه با روش فراوانی‌گرا دارد.

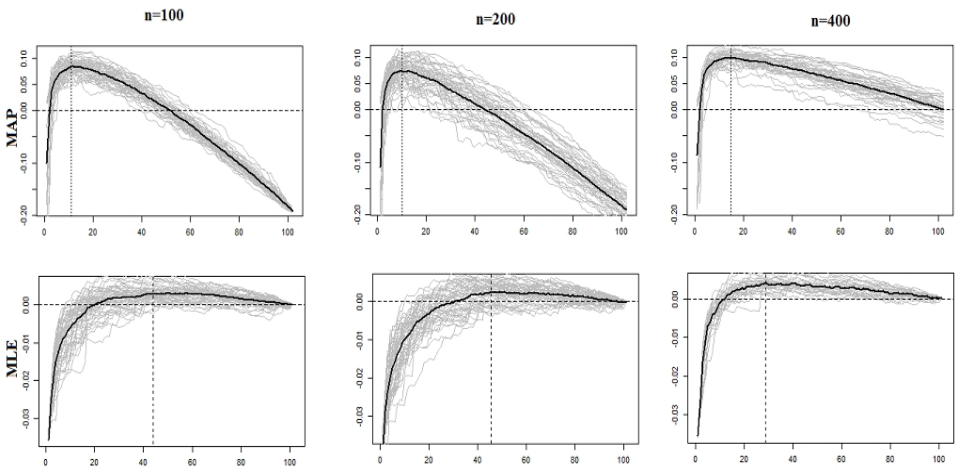
جدول ۵ متوسط MSE را برای اندازه‌های مختلف داده‌های آموزشی در دو حالت بیزی و فراوانی گرا نشان می‌دهد. به مانند مطالعات شبیه‌سازی، در اینجا نیز روش بیزی به‌وضوح خطای کمتری دارد. تا جایی که به‌ازای تعداد داده‌ی آموزشی ۱۰۰، MSE حالت فراوانی‌گرا تقریباً ۶/۵ برابر حالت بیزی است. همچنین در اینجا نیز به مانند مطالعات شبیه‌سازی شده، با افزایش اندازه‌ی داده‌های آموزشی متوسط MSE کاهش می‌یابد.

جدول (۵): مقادیر مربوط به متوسط میانگین توان دوم خطا در روش ماکسیمم احتمال پسین.

n	روش	
	بیزی	فراوانی‌گرا
۱۰۰	۰/۴۱۱	۲/۶۶۷
۲۰۰	۰/۳۷۹	۰/۴۸۸
۴۰۰	۰/۳۶۵	۰/۴۰۸

شکل ۳ تغییرات برآورد مطلوبیت را بر روی مسیرهای جستجو نشان می‌دهد. منحنی‌ها و خطوط با همان تعریف مورد استفاده در مطالعات شبیه‌سازی شده رسم شده‌اند. همچنین مقادیر به‌دست‌آمده براساس رابطه‌ی (۱۲) و با استفاده از مدل مرجع به‌دست آمده‌اند.

در شکل ۳ می‌توان مشاهده کرد که در هر دو حالت فراوانی‌گرا و بیزی، برآورد مطلوبیت در جایی بیشینه مقدارش را اتخاذ کرده است که مطلوبیت مدل از مطلوبیت مدل مرجع بیشتر است. همچنین در حالت بیزی، برآورد مطلوبیت تمایل به انتخاب متغیرهای کمتری در مقایسه با حالت فراوانی‌گرا داشته که با نتایج قبلی مطابقت دارد.



شکل (۳): پراکندگی برآورد مطلوبیت به روی مسیرهای جستجو

۵- بحث و نتیجه‌گیری

با قرار دادن توزیع پیشین نرمال در برابر زمانی که پیشین تخت مورد استفاده قرار می‌گیرد، در واقع یکی از روش‌های بیزی را با روش فراوانی‌گرا مورد مقایسه قرار دادیم. براساس نتایج حاصل از برآورد مطلوبیت جدید معرفی شده در این مقاله، می‌توان گفت روش ماکسیمم احتمال پسین، در مسائل رگرسیون خطی عملکرد بهتری داشته و تمایل کمتری به انتخاب متغیرهای نامرتبط با متغیر پاسخ دارند. همچنین به‌طورکلی هرچه اندازه داده‌های آموزشی بیشتر باشد؛ دو روش بیزی و فراوانی‌گرا عملکرد نزدیک‌تری به یکدیگر خواهند داشت؛ اما آنچه از نتایج مشاهده می‌شود، این است که برای نمونه‌های با حجم کم، روش ماکسیمم احتمال پسین در حالت بیزی عملکرد مناسب‌تری دارد. بررسی و محاسبه‌ی مقدار تحلیلی برآورد مطلوبیت برای مدل‌های خطی تعمیم‌یافته می‌تواند هدف مطالعات آینده باشد.

تشکر و قدردانی

نویسندگان مقاله از زحمات سردبیر محترم نشریه و همچنین داوران محترم که با نظرات و پیشنهادهای خود باعث بهبود مطالب مندرج در مقاله شدند؛ کمال تشکر و قدردانی را به عمل می‌آورند.

منابع

[1] Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association*, **74**(365), 153–160.

[2] Watanabe, S. (2009). *Algebraic geometry and statistical learning theory (Vol. 25)*, Cambridge University Press.

[3] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, **64**(4), 583-639.

[4] Laud, P.W. and Ibrahim, J.G. (1995). Predictive model selection, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 247–262.

[5] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach, *Biometrika*, **85**(1), 1–11.

[6] Marriott, J.M. and Spencer, N.M. and Pettitt, A.N. (2001). A bayesian approach to selecting covariates for prediction, *Scandinavian Journal of Statistics*, **28**(1), 87–97.

[7] Kass, R.E. and Raftery, A.E. (1995). Bayes factors, *Journal of the American Statistical Association,* **90**, 773-795.

[8] O'Hara, R.B. and Sillanpää, M.J. (2009). A review of bayesian variable selection methods: what, how and which, *Bayesian Analysis*, **4**(1), 85–117.

[9] Narisetty, N.N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors, *The Annals of Statistics*, **42**(2), 789–817.

[10] Piironen, J. and Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection, *Statistics and Computing,* **27**, 711-735.

[11] Bernardo, J. M. and Smith, A.F.M. (1994). *Bayesian Theory,* Wiley, New York.

[12] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association,* **102**(477), 359-378.

[13] Cawley, G.C. and Talbot, N.L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research,* **2**, 2079-2107.

[14] Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys,* **6**, 142-228.

[15] Fahrmeir, L., and Kneib, T., Lang, S. (2009). Regression: Modelle, Methoden und Anwendungen, Berlin Heidelberg: Springer.

JAMM, (2020), 10(2), 245-266
DOI: 10.22055/JAMM.2020.30812.1753
Shahid Chamran University of Ahvaz
ISSN: 2251-8088
Jamm.Scu.ac.ir

An Analysis on Covariates Selection Problem for Gaussian Model by Maximum a Posteriori Criterion Using Frequentist and Bayesian Approaches

Amirhossein Ghatari*, Mojtaba Ganjali**

* Department of Statistics, Amirkabir University of Technology, Tehran, Iran

** Department of Statistics, Shahid Beheshti University, Tehran, Iran

Received: August 26 2019 Accepted for publication: March 12 2020
Corresponding author: a.h.ghatari@aut.ac.ir

Abstract

Choosing the most suitable fitted model on data is one of the common challenges in statistical modeling. Maximum a posteriori (MAP) criterion is a method used in both frequentist and Bayesian approaches. Additionally, the utility of the model is used as a tool to compare the performances of methods. In this paper, the MAP method is applied for the Gaussian model and its performance is compared to that of frequentist approach. Also, an analytical form of utility estimation is proposed. Besides, using simulation studies, it is shown that the Gaussian model has better performance, based on both utility and mean of squared errors (MSE) criteria, when it is used by the Bayesian approach. However, both frequentist and Bayesian approaches avoid over-fitting by increasing the sample size. Also, by increasing correlation among covariates, MSE increases, while the tendency of choosing fewer covariates is raised. Eventually, the study on a real dataset is shown that in both frequentist and Bayesian approaches, MSE of selected models decreases when the size of sample increases.

Keywords: Model selection, Maximum a posteriori, Over-fitting, Utility of The Model.

Mathematics Subject Classification (2010): 62J05, 62J12.