چهارمین سمینار آمار فضایی و کاربردهای آن
4th Seminar on Spatial Statistics and Its Applications
۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

برآورد مدل سری زمانی اتورگرسیو چندکی با استفاده از الگوریتم EM

علی آقامحمدی، علی م. مصمم، محمد بهمنی
گروه آمار، دانشگاه زنجان

چکیده: در این مقاله مدل سری زمانی اتورگرسیو چندکی معرفی شده و سپس پارامترهای مدل با استفاده از الگوریتم EM که یک روش تکراری برای محاسبه برآوردهای ماکزیمم درستنمایی است، برآورد می‌شوند. تابع درستنمایی برای مدل اتورگرسیو چندکی بر اساس توزیع لاپلاس نامتقارن بیان شده و برای برآورد پارامترهای مدل به روش الگوریتم EM از شکل آمیخته مقیاس این توزیع استفاده میشود. با استفاده از مطالعه شبیه سازی و همچنین تحلیل داده‌های واقعی کارایی و کاربست روش ارائه شده مورد بررسی قرار میگیرد.

واژه‌های کلیدی: مدل اتورگرسیو چندکی، الگوریتم EM، توزیع لاپلاس نامتقارن،مدل‌های سری زمانی چندکی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08، 62M10، 62F10.

۱ مقدمه

تحلیل سری‌های زمانی به طور نظری و عملی از سال‌های ۱۹۷۰ میلادی به بعد برای پیش‌بینی و کنترل، به سرعت بسط و توسعه پیدا کرد. در سالهای اخیر نیز تحلیل‌ها و مدل‌های سری‌های زمانی کاربرد زیادی در علوم مختلف نظیر زیست شناسی، مطالعات زیستی، اقتصاد، محیط‌زیست و ... پیدا کرده است، زیرا تحلیل سری‌های زمانی معمولا به داده‌هایی مربوط می‌شود که مستقل نبوده و بطور متوالی به‌هم وابسته‌اند یا به عبارت دیگر مشاهدات در طول زمان جمع آوری می‌شوند و یک وابستگی بین آنها موجود است. در واقع همین وابستگی بین مشاهدات متوالی است که مورد توجه قرار می‌گیرد و بیشتر کاربرد آن نیز در پیش‌بینی خواهد بود. این پیش‌بینی در مورد حوادثی است که در آینده احتمال وقوع دارند و معمولا براساس رویدادهایی صورت می‌گیرند که در گذشته اتفاق افتاده‌اند. یکی از مدل‌های مهم تحلیل‌های سری‌های زمانی، مدل اتورگرسیو است. فرایندهای اتورگرسیو کاربرد زیادی در پردازش سیگنال، مخابرات، رادار، اقتصاد سنجی و ... دارند. مدل‌های اتورگرسیو به اختصار با نماد $AR(p)$ نمایش داده می‌شود، که در آن p مرتبه مدل اتورگرسیو است که آن را می‌توان

* نام وآدرس پست الکترونیکی ارائه‌دهنده مقاله: علی آقامحمدی، Aghamohammadi.ali@znu.ac.ir

بر اساس توابع خودهمبستگی و خودهمبستگی جزئی مشاهدات تعیین کرد. این مدل از دیدگاه آمار بیزی و آمار فراوانی‌گرا مورد مطالعه قرار گرفته است. همچنین به دلیل ماهیت این مدل که در آن مشاهده در زمان t (y_t) روی مقادیر گذشته سری یعنی $y_{t-p}, \ldots, y_{t-1}$ رگرسیون شده و در واقع یک الگوی رگرسیون چندگانه را تشکیل می‌دهد، به روش مدل رگرسیون میانگین و مدل رگرسیون چندکی نیز مورد مطالعه قرار گرفته است. برای مثال کوئنکر و شیائو (۲۰۰۶) مدل اتورگرسیو چندکی را ارائه و مورد مطالعه قرار دادند. انگل و منگالی (۲۰۰۴) مدل اتورگرسیو شرطی را از طریق رگرسیون چندکی در تجزیه و تحلیل ارزش در معرض خطر برای مدیریت ریسک بررسی کردند. همان طور که در کوئنکر و شیائو (۲۰۰۶) آمده است، برآورد پارامترها در مدل‌های اتورگرسیو دارای اهمیت بسیار زیادی است، معمولا برآوردگرهایی که بر میانگین تمرکز دارند و به روش رگرسیون میانگین به دست می‌آیند دارای محدودیت‌هایی هستند، که ممکن است عملکرد ضعیفی در تحلیل داده‌ها داشته باشند. در واقع این مدل‌ها نسبت به مشاهدات دورافتاده استوار نیستند و زمانی که توزیع داده‌ها غیرنرمال است از کارایی کمتری برخوردارند (کوئنکر و شیائو ، ۲۰۰۶). از این رو مدل‌هایی که توزیع مشاهدات را در چندک‌های مختلف مورد بررسی قرار می‌دهند، و نسبت به داده‌های دورافتاده نیز استوار هستند، می‌توانند جایگزینی مناسب برای مدل‌های میانگین در نظر گرفته شوند. در این مقاله ابتدا در بخش دوم مدل اتورگرسیو چندکی براساس توزیع لاپلاس نامتقارن [۱] (ALD) بیان شده و در بخش سوم روش برآورد پارامترها در سطوح متفاوت چندکی با استفاده از الگوریتم EM مورد بررسی قرار می‌گیرند. در بخش چهارم نیز مطالعه شبیه‌سازی و تحلیل داده‌های واقعی برای ارزیابی کارایی روش پیشنهادی ارائه شده است.

۲ مدل

مدل اتورگرسیوخطی مرتبه p به صورت

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t = y_t - \sum_{j=1}^{p} y_{t-j}\phi_j + \varepsilon_t, \quad t = p+1, ..., n, \qquad (۱.۲)$$

است، که در آن y_t مشاهده در زمان tام، $\Phi = (\phi_1, \ldots, \phi_p)'$ پارامترهای مدل اتورگرسیو و ε_t فرایند تصادفی محض در زمان t با میانگین صفر و واریانس σ^2 [۲] است. در رابطه (۱.۲) برای اطمینان از ایستایی فرض می‌کنیم تمام ریشه‌های چندجمله‌ای $1 - \sum_{j=1}^{p} \phi_j z^j$ خارج دایره واحد هستند. با توجه به مدل رگرسیون چندکی ارائه شده در کوئنکر و باست (۱۹۷۸) و همچنین ایده بیان شده در کوئنکر و شیائو (۲۰۰۶)، برآورد پارامترها در مدل اتورگرسیو چندکی برای چندک τام بصورت

$$\hat{\Phi} = \arg\min_{\Phi} \sum_{t=p+1}^{n} \rho_\tau \left(y_t - \sum_{j=1}^{p} \phi_j y_{t-j} \right), \qquad (۲.۲)$$

به دست می‌آیند، که در آن $\rho_\tau(u)$ تابع زیان و بصورت $\rho_\tau(u) = \frac{|u| + (2\tau - 1)u}{2}$ تعریف می‌شود. یو و مویید (۲۰۰۱) از طریق تعریف توزیع لاپلاس نامتقارن بصورت

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\rho_\tau \left(\frac{y-\mu}{\sigma} \right) \right\},$$

[1] Asymmetric Laplace Distribution
[2] Maximum Likelihood

که در آن μ پارامتر مکان، σ پارامتر مقیاس و τ پارامتر شکل است، برآورد پارامترها را در مدل رگرسیون چندکی به روش ماکزیمم درستنمایی [۲] محاسبه کردند. تابع درستنمایی مدل اتورگرسیو چندکی مرتبه p براساس این توزیع بصورت

$$\mathrm{L}(\Theta|\boldsymbol{y}) = \prod_{t=q+1}^{n} \left[\frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\rho_\tau \left(\frac{y_t - \sum_{j=1}^{p} \phi_j y_{t-j}}{\sigma} \right) \right\} \right], \tag{۳.۲}$$

است، که در آن $\Theta = (\Phi', \sigma)'$ پارامترهای مدل را نشان می‌دهد. توجه کنیم که مینیمم کردن تابع هدف در رابطه (۲.۲) معادل ماکزیمم کردن تابع درستنمایی در رابطه (۳.۲) نسبت به پارامتر Φ در حضور پارامتر مزاحم σ است(توجه کنیم در رگرسیون میانگین تابع زیان درجه دوم را داریم که از طریق توزیع نرمال برای خطاها می‌توان تابع درستنمایی معادل را تشکیل داد). به همین دلیل در مطالعه رگرسیون چندکی از توزیع ALD استفاده می‌شود. بنابراین استنباط آماری برای مدل اتورگرسیو چندکی رابطه (۱.۲) را می‌توان براساس توزیع ALD انجام داد. اما به دلیل وجود تابع قدرمطلق نمی‌توان بصورت تحلیلی تابع درستنمایی را نسبت به پارامترها ماکزیمم کرد و فرم بسته‌ای برای برآوردها به دست آورد. برای حل این مشکل ابتدا با استفاده از شکل آمیخته مقیاس توزیع ALD، مدل را بصورت آمیخته‌ای از توزیع نرمال بازنویسی می‌کنیم (کازومی و کوبایاشی ، ۲۰۱۱). در نتیجه با استفاده از شکل آمیخته توزیع ALD رابطه (۱.۲) را می‌توان بصورت

$$y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \theta_1 v_t + \sqrt{\theta_2 \sigma v_t}.e_t, \quad t = q+1, \ldots, n, \tag{۴.۲}$$

در نظر گرفت، که در آن $v_t \sim \mathrm{Exp}(\frac{1}{\sigma})$، $e_t \sim N(0,1)$ و e_t و v_t مستقل از هم هستند. در عبارت فوق $\theta_1 = \frac{1-2\tau}{\tau(1-\tau)}$، $\theta_2 = \frac{2}{\tau(1-\tau)}$ است و v_t همان متغیر آمیخته کننده را نشان می‌دهد. بنابراین تابع درستنمایی شرطی توام برای داده‌های کامل $\{\boldsymbol{y}, \boldsymbol{v}\}$ بصورت زیر به دست می‌آید.

$$\mathrm{L_C}(\Theta|\boldsymbol{y}, \boldsymbol{v}) = \prod_{t=p+1}^{n} \left[\frac{1}{\sqrt{2\pi}\sqrt{\theta_2 \sigma v_t}} \exp\left\{ -\frac{\left(y_t - \sum_{j=1}^{p} \phi_j y_{t-j} - \theta_1 v_t\right)^2}{2\theta_2 \sigma v_t} \right\} . \frac{1}{\sigma} \exp\left\{ -\frac{1}{\sigma} v_t \right\} \right] \tag{۵.۲}$$

توجه کنیم که در تابع درستنمایی فوق متغیر v_t، متغیر پنهان است. لذا می‌توان برای برآورد پارامترهای مدل از الگوریتم EM استفاده کرد.

۳ روش برآورد

الگوریتم EM که اولین بار توسط دمپستر و همکاران (۱۹۹۷) ارائه شد ابزاری قدرتمند برای محاسبه برآورد ماکزیمم درستنمایی پارامترهای یک مدل آماری در حضور داده‌های ناقص، متغیرهای پنهان ، توزیع‌های بریده شده، مشاهدات گروه‌بندی و یا سانسور شده است. با توجه به وجود متغیرهای پنهان v_t در تابع درستنمایی رابطه (۵.۲)، الگوریتم EM برای برآورد پارامترها روش کارایی است. در این قسمت با استفاده از این روش برآورد پارامترهای مدل عنوان شده را به دست می‌آوریم. الگوریتم EM یک روش تکراری برای محاسبه برآوردهای ماکزیمم درستنمایی است. هر تکرار از این الگوریتم شامل دو مرحله است: مرحله E و مرحله M ، که با یک مقدار اولیه مانند $\Theta^{(0)}$ آغاز می‌شود. در مرحله E که به آن مرحله محاسبه متوسط مقدار گویند، متوسط مقدار لگاریتم (طبیعی) تابع درستنمایی نسبت به توزیع شرطی متغیرهای پنهان به شرط مشاهدات (متغیر پاسخ) محاسبه می‌شود. مرحله M که به آن مرحله ماکزیمم سازی گویند، عبارت حاصل در مرحله E نسبت به پارامترهای مجهول یعنی Θ ماکزیمم می‌شود تا Θی جدید به دست آید. به همین ترتیب مراحل E و M تا رسیدن به همگرایی قابل قبول ادامه پیدا کرده و Θی محاسبه شده در مرحله همگرایی، همان برآورد ماکزیمم درستنمایی Θ است. حال برای مدل اتورگرسیو چندکی مرتبه p، در تکرار hام لگاریتم تابع درستنمایی برای داده‌های کامل به صورت

$$\log\left[L_C\left(\Theta|\boldsymbol{y},\boldsymbol{v}\right)\right] \propto -\frac{(n-p)}{۲}\log(\theta_۲) - \frac{۳\,(n-p)}{۲}\log\sigma - \frac{۱}{۲}\sum_{t=p+۱}^{n}\log v_t$$
$$-\sum_{t=p+۱}^{n}\frac{۱}{\sigma}\left[\frac{\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)^{۲}}{۲\theta_۲}v_t^{-۱}+\frac{\theta_۱^{۲}+۲\theta_۲}{۲\theta_۲}v_t-\frac{\theta_۱\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)}{\theta_۲}\right] \qquad (۱.۳)$$

است. مرحله E : در این مرحله متوسط مقدار لگاریتم تابع درستنمایی فوق نسبت به توزیع شرطی متغیر پنهان به شرط مشاهدات محاسبه می‌شود، یعنی

$$\mathrm{Q}\left(\Theta|\Theta^{(h-۱)}\right)=\mathrm{E}\left[\log\mathrm{L_C}\left(\Theta|\boldsymbol{y},\boldsymbol{v}\right)|\Theta^{(h-۱)},\boldsymbol{y}\right]=-\frac{(n-p)}{۲}\log\left(\theta_۲\right)-\frac{۳(n-p)}{۲}\log\sigma$$
$$-\frac{۱}{۲}\sum_{t=p+۱}^{n}\mathrm{E}\left(\log v_t|\Theta^{(h-۱)},y_t\right)-\sum_{t=q+۱}^{n}\frac{۱}{\sigma}\left[\frac{\eta_t^{۲}}{۲\theta_۲}\mathrm{E}\left(v_t^{-۱}|\Theta^{(h-۱)},y_t\right)+\frac{\theta_۱^{۲}+۲\theta_۲}{۲\theta_۲}\mathrm{E}\left(v_t|\Theta^{(\mathrm{h}-۱)},y_t\right)-\frac{\theta_۱\eta_t}{\theta_۲}\right], \qquad (۲.۳)$$

که در آن $\eta_t = y_t - \sum_{j=۱}^{p}\phi_j y_{t-j}$ است. در رابطه (۲.۳) امید ریاضی متغیرهای تصادفی v_t، $v_t^{-۱}$ و $\log(v_t)$ به شرط مشاهدات باید محاسبه شوند. تابع چگالی شرطی v_t به شرط y_t به صورت

$$f\left(v_t|y_t\right)\propto\frac{۱}{\sqrt{v_t}}\exp\left\{-\frac{۱}{۲}\left[\frac{\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)^{۲}}{\theta_۲\sigma}v_t^{-۱}+\frac{\theta_۱^{۲}+۲\theta_۲}{\theta_۲\sigma}v_t\right]\right\}$$
$$\sim GIG\left(\frac{۱}{۲},\frac{\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)^{۲}}{\theta_۲\sigma},\frac{\theta_۱^{۲}+۲\theta_۲}{\theta_۲\sigma}\right),\quad t=p+۱,...,n, \qquad (۳.۳)$$

به دست می‌آید، که در آن $GIG(.,.,.)$ توزیع گاوسی وارون تعمیم‌یافته است (زو و همکاران ، ۲۰۱۴). با توجه به ویژگی‌های این توزیع داریم (باراندوف و نیلسن ، ۲۰۰۱):

$$\delta_t = E\left(\log v_t|\hat{\Theta}^{(h-۱)}\right)=\left.\frac{dE\left(v_t^{\alpha}|\hat{\Theta}^{(h-۱)}\right)}{d\alpha}\right|_{\alpha=۰},\quad \alpha\in\mathbb{R}$$
$$\gamma_t = E\left(v_t^{-۱}|\hat{\Theta}^{(h-۱)}\right)=\frac{\sqrt{\theta_۱^{۲}+۲\theta_۲}}{\left|y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right|},$$
$$\lambda_t = E\left(v_t|\hat{\Theta}^{(h-۱)}\right)=\frac{\theta_۲\sigma}{\theta_۱^{۲}+۲\theta_۲}+\frac{\left|y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right|}{\sqrt{\theta_۱^{۲}+۲\theta_۲}}.$$

حال با جایگذاری عبارات فوق در رابطه (۲.۳)، در مرحله M این رابطه نسبت به $\Theta=(\Phi',\sigma)'$ ماکزیمم می‌شود. برای ماکزیمم سازی از رابط (۲.۳) نسبت به پارامترها مشتق می‌گیریم. برای پارامتر مقیاس σ در تکرار hام داریم:

$$\frac{\partial Q}{\partial\sigma}=-\frac{۳(n-p)}{۲\sigma}+\frac{۱}{\sigma^{۲}}\sum_{t=p+۱}^{n}\left[\frac{\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)^{۲}}{۲\theta_۲}\gamma_t+\frac{\theta_۱^{۲}+۲\theta_۲}{۲\theta_۲}\lambda_t-\frac{\theta_۱\left(y_t-\sum_{j=۱}^{p}\phi_j y_{t-j}\right)}{\theta_۲}\right]=۰ \qquad (۴.۳)$$

لذا،

$$\hat{\sigma}^{(h)}=\frac{۲}{۳(n-p)}\sum_{t=p+۱}^{n}\left[\frac{\left(y_t-\sum_{j=۱}^{p}\hat{\phi}_j^{(h-۱)}y_{t-j}\right)^{۲}}{۲\theta_۲}\gamma_t+\frac{\theta_۱^{۲}+۲\theta_۲}{۲\theta_۲}\lambda_t-\frac{\theta_۱\left(y_t-\sum_{j=۱}^{p}\hat{\phi}_j^{(h-۱)}y_{t-j}\right)}{\theta_۲}\right]. \qquad (۵.۳)$$

همچنین برای بردار ضرایب Φ در تکرار hام داریم:

$$\frac{\partial Q}{\partial\phi_l}=-\frac{۱}{\sigma\theta_۲}\sum_{t=p+۱}^{n}\left[-\gamma_t\left(y_t-\sum_{j=۱}^{p}y_{t-j}\phi_j\right)y_{t-l}+\theta_۱y_{t-l}\right]=۰,\quad l=۱,\ldots,p, \qquad (۶.۳)$$

از رابطه (۶.۳)، برآورد Φ در تکرار hام به صورت

$$\hat{\Phi}^{(h)} = \left(E' W^{(h-1)} E\right)^{-1} . \left[E' W^{(h-1)} \left(e - \left(W^{(h-1)}\right)^{-1} d\right)\right] \qquad (۷.۳)$$

به دست می‌آید، که در آن

$$E' = \begin{pmatrix} y_p & y_{p+1} & \cdots & y_{n-1} \\ y_{p-1} & y_p & \cdots & y_{n-2} \\ \vdots & \vdots & & \vdots \\ y_1 & y_2 & \cdots & y_{n-p} \end{pmatrix}_{p\times(n-p)} , \qquad d^{(h)} = \begin{pmatrix} \theta_1 \\ \theta_1 \\ \vdots \\ \theta_1 \end{pmatrix}_{(n-p)\times 1}$$

و $W^{(h-1)} = \text{daig}\left(\gamma_{p+1}^{(h-1)}, \ldots, \gamma_n^{(h-1)}\right)$ و $e = (y_{p+1}, \ldots, y_n)^T$ است. حال گام‌های E و M را تا رسیدن به همگرایی مطلوب ادامه داده و برآوردها را به دست می‌آوریم.

۴ مطالعه شبیه‌سازی

در این بخش برای مقایسه کارایی روش ارائه شده (EMQA) با مدل اتورگرسیو چندکی بر اساس تابع زیان (QA) ارائه شده در کوئنکر و شیائو (۲۰۰۶) به مطالعه شبیه‌سازی می‌پردازیم. برای محاسبه برآوردهای مدل اتورگرسیو چندکی بر اساس تابع زیان، از تابع $qr()$ در بسته $quantreg$ نرم‌افزار R استفاده می‌کنیم. یک مجموعه داده به حجم $n = ۱۰۰$ از مدل

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \varepsilon_t, \qquad t = ۵, ..., ۱۰۰ \qquad (۱.۴)$$

با فرض $(\phi_1, \phi_2, \phi_3, \phi_4) = (۰/۳, -۰/۰۵, ۰/۱, ۰/۱)$، $N(۰, ۱)$، Laplace$(۰, ۱)$ و $t_{(۳)}$ برای ε_tها تولید می‌کنیم. برای همگرایی سریع الگوریتم EM مقادیر آغازین پارامترها را برابر برآورگرهای حداقل مربعات قرار می‌دهیم. به منظور لحاظ کردن تغییرپذیری نتایج حاصل از شبیه‌سازی، ۵۰۰ مرتبه روند شبیه‌سازی را تکرار می‌کنیم. برای مقایسه‌ی کارایی روش‌های عنوان شده از معیار خطای مدل (ME) که به صورت

$$\text{ME}\left(\hat{\Phi}\right) = \left(\hat{\Phi} - \Phi^{true}\right)' \left(\hat{\Phi} - \Phi^{true}\right)$$

تعریف می‌شود، استفاده می‌کنیم. درواقع مقدار ME$(\hat{\Phi})$ را ۵۰۰ مرتبه تحت هر مدل محاسبه کرده و سپس متوسط مقدار و انحراف استاندارد آنها به عنوان معیار مقایسه گزارش می‌شوند. نتایج شبیه‌سازی برای مدل‌های مختلف در سه سطح چندکی $\tau = (۰/۲۵, ۰/۵, ۰/۷۵)$، در جدول ۱ و ۲، آمده است. در جدول ۱ برآورد پارامترهای مدل (۱.۴) از طریق روش ارائه شده آمده است (متوسط مقدار برآورد پارامتر در ۵۰۰ تکرار شبیه‌سازی). با توجه به این جدول واضح است که مقادیر برآورد به مقادیر واقعی آنها نزدیک است.

جدول ۱: برآورد پارامترهای مدل با استفاده از روش EMQA

	Error												
	$N(۰, ۱)$				laplace$(۰, ۱)$				$t_{(۳)}$				
τ	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_1	ϕ_2	ϕ_3	ϕ_4	
$True$	۳/۰	-۰۵/۰	۱/۰	۱/۰	۳/۰	-۰۵/۰	۱/۰	۱/۰	۳/۰	-۰۵/۰	۱/۰	۱/۰	
۲۵/۰	۲۹۰/۰	-۰۶۲/۰	۰۸۸/۰	۰۸۶/۰	۲۸۹/۰	-۰۶۶/۰	۱۰۲/۰	۰۷۹/۰	۲۹۶/۰	-۰۵۵/۰	۰۹۳/۰	۰۸۲/۰	
۵/۰	۲۹۰/۰	-۰۶۶/۰	۰۸۴/۰	۸۳/۰	۲۸۷/۰	-۰۶۶/۰	۰۹۹/۰	۰۷۷/۰	۲۹۲/۰	-۰۵۶/۰	۰۹۱/۰	۰۸۰/۰	
۷۵/۰	۲۹۳/۰	-۰۶۸/۰	۱۰۲/۰	۰۷۹/۰	۲۸۸/۰	-۰۶۷/۰	۱۰۲/۰	۰۷۸/۰	۲۹۳/۰	-۰۵۳/۰	۰۹۴/۰	۰۸۲/۰	

در جدول ۲ ، متوسط مقدار خطای مدل و انحراف استاندارد آنها در ۵۰۰ تکرار شبیه‌سازی آمده است. با توجه به این جدول مشاهده می‌شود که متوسط خطای مدل روش پیشنهادی نسبت به روش QA تمام سطوح چندکی کوچکتر بوده و لذا کارایی آن بیشتر از مدل QA است.

جدول ۲: مقادیر ME و sd تحت مدل‌های مختلف

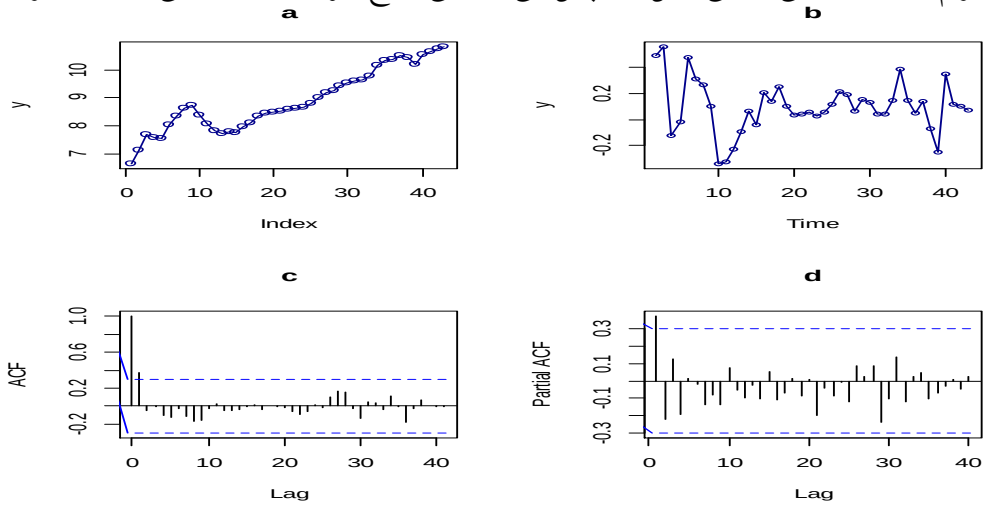
$EMQA_{۰/۷۵}$	$EMQA_{۰/۵}$	$EMQA_{۰/۲۵}$	$QA_{۰/۷۵}$	$QA_{۰/۵}$	$QA_{۰/۲۵}$		Error
۰۶۰۵/۰	۰۵۳۰/۰	۰۶۱۰/۰	۰۸۷۹/۰	۰۷۱۴/۰	۰۸۴۹/۰	ME	$N(۰,۱)$
۰۵۰۰/۰	۰۴۳۳/۰	۰۴۸۷/۰	۰۶۳۹/۰	۰۵۷۴/۰	۰۶۹۵/۰	sd	
۰۳۳۹/۰	۰۳۱۴/۰	۰۳۴۵/۰	۰۶۴۰/۰	۰۳۳۱/۰	۰۶۱۸/۰	ME	Laplace$(۰,۱)$
۰۲۷۴/۰	۰۲۵۸/۰	۰۲۹۲/۰	۰۵۲۵/۰	۰۲۸۲/۰	۰۵۰۵/۰	sd	
۰۳۵۳/۰	۰۳۱۹/۰	۰۳۵۸/۰	۰۵۴۱/۰	۰۳۴۴/۰	۰۵۱۹/۰	ME	$t(۳)$
۰۲۸۱/۰	۰۲۶۵/۰	۰۲۹۲/۰	۰۴۶۸/۰	۰۳۰۶/۰	۰۴۵۶/۰	sd	

۵ تحلیل داده‌های واقعی

در این بخش، هدف برازش مدل‌های عنوان شده به داده‌های واقعی و ارزیابی کارایی آنها است. مجموعه داده‌ها از وب‌سایت info@energyinformation.ir، گرفته شده است. داده‌ها انتشار گاز آلاینده $۲CO$ ناشی از سوختن روغن‌های صنعتی در سراسر کره زمین در ۴۳ سال مختلف را از سال ۱۹۷۱ تا ۲۰۱۳ میلادی بر حسب میلیون تن نشان می‌دهد. مشاهدات را ابتدا بر عدد ۱۰۰۰ تقسیم می‌کنیم که نمودار سری زمانی آن در شکل ۱ (a) آمده است. مشخص است که داده‌ها دارای روند بوده و لذا سری ایستا نیست. برای ایستایی یک بار سری را تفاضلی می‌کنیم که نمودار آن در شکل ۱ (b) آمده است و حاکی از ایستایی سری است. نمودارهای ACF و PACF در شکل ۱ (c,d) آمده است با توجه به این دو نمودار می‌توان گفت داده‌های تفاضلی شده از مدل $AR(۱)$ پیروی می‌کنند. لذا مدل را به صورت

$$۲CO_t = \phi_۰ + \phi_۱ ۲CO_{t-۱} + \varepsilon_t, \quad t = ۲, ..., ۴۳,$$

در نظر می‌گیریم. جهت ارزیابی کارایی روش‌ها از پیش‌بینی بازگشتی خارج نمونه‌ای با مبدا زمانی $t_۰ = ۳۰$ برای هر روش



شکل ۱: ردیف اول نمودارهای پراکندگی مشاهدات قبل و بعد از ایستاسازی و ردیف دوم نمودارهای ACF و PACF را نشان می‌دهد.

ستفاده می‌کنیم. یعنی ابتدا مدل‌ها را به بخشی از داده‌ها $(n = ۳۰)$ برازش می‌دهیم و سه گام بعدی را پیش‌بینی می‌کنیم سپس سه گام به جلو رفته و قرار می‌دهیم $(n = ۳۳)$ و دوباره سه گام بعدی را پیش‌بینی می‌کنیم. این روند را تا زمان $t = ۳۹$ ادامه می‌دهیم. در واقع با این روند ۱۲ مورد پیشی‌بینی تحت هر روش انجام می‌شود. مقادیر متوسط قدر مطلق

خطای پیش‌بینی (MAP) و انحراف استاندارد (SDP) را در چندک‌های $\tau = (۰/۱, ۰/۲۵, ۰/۵, ۰/۷۵)$، به صورت

$$\text{MAP} = \frac{۱}{۱۲}\sum_{i=۱}^{۱۲}|\hat{y}_{t.+i} - y_{t.+i}| = \frac{۱}{۱۲}\sum_{i=۱}^{۱۲}|\hat{e}_{t.+i}|, \qquad \text{SDP} = \sqrt{\frac{۱}{۱۲}\sum_{i=۱}^{۱۲}(|\hat{e}_{t.+i}| - \text{MAP})^{۲}},$$

محاسبه می‌کنیم. واضح است که هر چه مقادیر MAP و SDP تحت یک مدل کمتر باشد، کارایی آن مدل بهتر است. مقادیر MAP و SDP در جدول ۳، گزارش شده است. با توجه به این جدول، هر چند روش QA در سطح چندک ۵/۰ تا حدودی عملکرد بهتری نسبت به روش EMQA دارد. اما برای چندک‌های بالایی و پایینی عملکرد EMQA خیلی بهتر است. لذا روش پیشنهادی به لحاظ پیش‌بینی نیز عملکرد بهتری نسبت به روش QA دارد. بنابراین پیشنهاد می‌شود برای تحلیل داده‌های سری‌های زمانی به روش مدل رگرسیون چندکی از روش پیشنهادی استفاده شود.

جدول ۳: مقادیر MAP و SDP تحت مدل‌های مختلف

$EMQA_{۰/۷۵}$	$EMQA_{۰/۵}$	$EMQA_{۰/۲۵}$	$EMQA_{۰/۱}$	$QA_{۰/۷۵}$	$QA_{۰/۵}$	$QA_{۰/۲۵}$	$QA_{۰/۱}$	
۱۵۰۵/۰	۱۳۹۹/۰	۱۴۱۱/۰	۱۴۹۸/۰	۲۳۰۷/۰	۱۳۷۳/۰	۱۵۳۷/۰	۲۵۳۵/۰	MAP
۱۲۱۶/۰	۱۲۷۰/۰	۱۳۸۲/۰	۱۵۳۶/۰	۱۵۲۶/۰	۱۱۴۶/۰	۱۳۷۹/۰	۱۹۹۹/۰	SDP

مراجع

Dempester, A., Lird, N., Rubin, D. (1977), Maximum Likelihood from Incomplete Data via the EM algorithm, *Journal of the Royal Statistical Society*, **39**, 1-38.

Koenker, R., Xiao, Z. (2006), Quantile autoregression, *Journal of the American Statistical Association* , **101** , 980-990.

Koenker, R., Bassett, J. (1978), Regression Quantiles, *Econometrica*, **46**, 33-50.

Zau, H., Yuan, M. (2008), Composite quantile regression and the oracle model selection theory, *The Annals of Statistics*, **36**, 1108-1126

Yu, K., Moyeed, R.A. (2001), Bayesian quantile regression, *Statistics and Probability Letters*, **54**, 437-447.

Kozumi, H., Kobayashi, G. (2011), Gibbs sampling methods of Bayesian quantile regression, *Journal of Statistical Computation and Simulation*, **81**, 1565-1578.

Engle, R.F., Manganelli, S. (2004), CAViaR:Conditional autoregressive value at risk by regression quantiles, *Journal of Business Economic Statistics*, **22**, 367-381.

Barndorff, O., Nielsen, E., (2001), Non-Gassian Ornstein-Uhlenbeck Based Models some of their uses in financial Economics, *Journal of the Royal Statistical Society: Series B* , **63** , 167-241.

Zhou, Y.H., Ni, Z.X., Li, Y. (2014), Quantile Regression via the EM Algorithm, *Communications in Statistics-Simulation and Computation* , **43**, 2162-2172.