مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۲، دوره ۱۳، شماره ۳، ص ۳۱۶-۳۳۱

شاپا چاپی: ۲۲۵۱-۸۰۸۸
DOI:10.22055/JAMM.2023.42541.2119

جانهی داده‌های گمشده با استفاده از ترکیب روش تحلیل مجموعه‌ی مقدار تکین و الگوریتم پالایش کالمن و مقایسه با روش‌های جانهی یک متغیره

مسعود یارمحمدی ،[(۱)] [۱] رضا ذبیحی مقدم[(۱)]

[(۱)] گروه آمار دانشگاه پیام نور، صندوق پستی ۴۶۹۷-۱۹۳۹۵ تهران، ایران

دبیر مسئول: رحیم چینی‌پرداز

تاریخ دریافت: ۱۴۰۱/۹/۲۴ تاریخ پذیرش: ۱۴۰۲/۶/۲۵

چکیده: مقادیر گمشده در داده‌های سری زمانی، یکی از مشکلاتی است که گاهی اوقات در تحلیل سری‌های زمانی به وجود می‌آیند. هر چقدر جانهی این مقادیر دقت بیشتری داشته باشد، درک بهتری از ساختار سری زمانی به دست آمده و در نتیجه، تشخیص الگوی آن و پیش‌بینی مقادیر آینده نیز دقیق‌تر خواهند بود. از این رو انتخاب یک روش مناسب جانهی، بخش مهمی از یک تحلیل سری زمانی را تشکیل می‌دهد. در این مقاله، به معرفی روش جدید جانهی داده‌های گمشده از روش تحلیل مجموعه‌ی مقدار تکین با استفاده از الگوریتم پالایش کالمن می‌پردازیم. در ادامه روش‌های جانهی مقادیر گمشده در سری‌های زمانی یک متغیره معرفی شده و سپس به مقایسه روش‌های مذکور با استفاده از داده‌های شبیه‌سازی شده در مدل‌های ساختاری و داده واقعی می‌پردازیم. نتایج مقایسه بر اساس معیار ریشه میانگین مربعات خطا و میانگین قدر مطلق انحرافات نشان می‌دهد که جانهی مقادیر گمشده بر اساس روش تحلیل مجموعه‌ی مقدار تکین با استفاده از الگوریتم پالایش کالمن، عملکرد بهتری نسبت به سایر روشهای جانهی داشته و روش نما نیز بدترین روش است.

واژه‌های کلیدی: سری زمانی، مقادیر گمشده، جانهی، روش تحلیل مجموعه‌ی مقدار تکین، معادلات فضای حالت، الگوریتم پالایش کالمن، مدل‌های ساختاری.

رده‌بندی ریاضی: 62M10; 62M20

[۱] نویسنده مسئول مقاله

(R. Zabihi Moghadam) Rezazm63@gmail.com, (M. Yarmohammadi) masyar@pnu.ac.ir

۱ مقدمه

یکی از مشکلاتی که گاهی اوقات در تحلیل سری‌های زمانی ممکن است رخ دهد، وجود مقادیر گمشده[۲] می‌باشد. این مقادیر به دلایل مختلفی از جمله عدم امکان اندازه‌گیری متغیر مورد نظر در برخی زمان‌ها، خرابی دستگاه‌های اندازه‌گیری، خطای انسانی و یا از بین رفتن اطلاعات، به وجود می‌آیند. نادیده گرفتن و حذف مقادیر گمشده، موجب از دست دادن اطلاعات نهفته در این مقادیر شده و این امر نیز استنباط‌های اشتباه در مورد الگوی سری زمانی و تشخیص نادرست ساختار همبستگی آن را به دنبال خواهد داشت. برآورد مقادیر گمشده که جانهی[۳] نامیده می‌شود، بخشی مهم از فرآیند آماده سازی داده‌ها برای تحلیل سری زمانی را تشکیل می‌دهد [۳]. جانهی اشتباه مقادیر گمشده می‌تواند دقت پیش‌بینی مقادیر آینده سری زمانی را به شدت تحت تأثیر قرار دهد [۲۳]. بنابراین هر چقدر جانهی این مقادیر دقت بیشتری داشته باشد، درک بهتری از ساختار سری زمانی حاصل شده و در نتیجه، تشخیص الگوی سری زمانی و پیش بینی مقادیر آینده نیز دقیق‌تر خواهند بود. همچنین با توجه به این که اغلب روش‌های تحلیل سری‌های زمانی را فقط برای داده‌های کامل می‌توان به کار برد، جانهی مقادیر گمشده با مقادیری معقول و منطقی، از اهمیت ویژه‌ای برخوردار است. حتی برخی از محققین بر این باورند که اهمیت جانهی مقادیر گمشده می‌تواند بیش از انتخاب یک روش پیش‌بینی باشد [۳].

جانهی مقادیر گمشده بخش وسیعی از پژوهش‌های مربوط به سری‌های زمانی را به خود اختصاص داده و روش‌های مختلفی برای سری‌های زمانی یک متغیره و چند متغیره معرفی شده‌اند. به عنوان مثال برخی از روش‌های قدیمی را می‌توان در [۱، ۷، ۱۴، ۱۸] یافت. در [۹] یک روش جانهی برای سری‌های زمانی چند متغیره نرمال بر اساس الگوریتم EM معرفی شده و بسته‌ای نیز با نام mtsdi برای انجام محاسبات مربوطه در نرم افزار R تهیه شده است[۱۰]. روش فضای حالت[۴] یا پالایش کالمن[۵] نیز یکی دیگر از روش‌های مناسب برای جانهی مقادیر گمشده در یک سری زمانی است و برای اطلاع از جزئیات این روش میتوان از [۶] بهره گرفت. یکی دیگر از این روش‌های جانهی، روش تحلیل مجموعه‌ی مقدار تکین[۶] (SSA) است که یک روش قدرتمند ناپارامتری است و از چهار گام تکمیلی تشکیل شده است و نیازی به فرض مانایی سری زمانی و نرمال بودن باقیمانده‌ها ندارد. روش SSA اولین بار توسط برومهید و کینگ [۲] معرفی گردید و از آن زمان به بعد مورد توجه بسیاری از محققین در علوم مختلف قرار گرفته است.

روش‌های جانهی بر اساس روش SSA را می‌توان به دو دسته‌ی کلی روش‌های مبتنی بر بازسازی و روش‌های مبتنی بر پیش‌بینی تقسیم کرد. در روش مبتنی بر بازسازی، ابتدا مقادیر گمشده با مقادیری ثابت جایگزین می‌شوند. سپس سری زمانی را بازسازی کرده و به جای مقادیر اولیه، مقادیر بازسازی شده‌ی آنها جایگزین می‌شوند. این فرایند آنقدر تکرار می‌شود تا یک همگرایی حاصل شود. آخرین مقادیر بازسازی شده به عنوان جانهی از مقادیر گمشده در نظر گرفته می‌شوند. برای اطلاعات بیشتر می‌توان به کوندراشوف و گیل [۱۳] و هوی زان و همکاران[۸] درباره روش جانهی برای سری‌های زمانی مانا مراجعه کرد.

در روش مبتنی بر پیش‌بینی، می‌توان از مقادیر پیش بینی‌ها به عنوان جایگزین نسبتاً مناسبی برای مقادیر گمشده استفاده کرد. در رودریگز و کاروالهو [۱۹] سری زمانی به دو بخش قبل و بعد از مقادیر گمشده تقسیم شده و این مقادیر به وسیله‌ی این دو زیر سری و با استفاده از روش پیش‌بینی بازگشتی در SSA، برآورد می‌شوند. سپس دو نوع برآورد به دست آمده به صورت وزن‌دار با هم ترکیب شده و در نهایت برآوردی برای مقادیر گمشده به دست می‌آید. محمودوند و رودریگز [۱۵] نیز با ارائه‌ی نوع جدیدی از وزن‌دهی مبتنی بر بوت استرپ، این روش را بهبود بخشیدند.

در این مقاله روش ترکیبی برای بهبود عملکرد روش جانهی SSA مبتنی بر پیش‌بینی با استفاده از الگوریتم پالایش کالمن پیشنهاد می‌شود. در ادامه روش‌های جانهی مقادیر گمشده درون‌یابی، هموارسازی کالمن، جانهی با مشاهده‌ی قبلی، جانهی با مشاهده‌ی بعدی، میانگین متحرک موزون، میانگین کل، میانه و نما در سری‌های زمانی یک متغیره معرفی شده و با استفاده از داده‌های شبیه‌سازی شده و واقعی مربوط به تعداد افراد مبتلا به آنفولانزا در کشور فرانسه، بر اساس معیار ریشه میانگین مربعات خطا[۷] و میانگین قدر مطلق انحرافات[۸] مورد مقایسه قرار می‌گیرند. برای داده‌های شبیه‌سازی شده نیز از مدل‌های ساختاری نامانا که اغلب با نویز قابل توجهی همراه هستند، استفاده شده است.

در بخش ۲ روش SSA و مراحل آن معرفی می‌شود. در بخش ۳ به معرفی مدل‌های ساختاری، مدل فضای حالت و الگوریتم پالایش کالمن می‌پردازیم. در بخش ۴ روش پیش‌بینی بازگشتی SSA با استفاده از الگوریتم پالایش کالمن ارائه می‌شود. در بخش ۵ روش‌های جانهی درون‌یابی، هموارسازی کالمن، جانهی با مشاهده‌ی قبلی، جانهی با مشاهده‌ی بعدی، میانگین متحرک موزون، میانگین کل، میانه، نما، SSA و SSA با استفاده از الگوریتم پالایش کالمن مورد بحث و بررسی قرار می‌گیرد. در بخش ۶ با انجام شبیه‌سازی و استفاده از داده‌های واقعی توانایی عملکرد روش‌های جانهی نشان داده شده است.

---

[2] Missing Values
[3] Imputation
[4] State Space
[5] Kalman Filter
[6] Singular Spectrum Analysis
[7] Root Mean Squared Error
[8] Mean Absolute Deviations

۲ مروری بر روش SSA

روش SSA یک روش ناپارامتری برای تجزیه و تحلیل سری‌های زمانی بوده که می‌تواند مشاهدات سری زمانی را به چندین مؤلفه تقسیم کرده و در ادامه برای انجام پیش‌بینی مورد استفاده قرار گیرد. این روش از دو مرحله تجزیه و بازسازی تشکیل شده است که هر یک از این مراحل شامل دو گام می‌باشند.

مرحله **۱**. تجزیه

مرحله تجزیه از دو گام تعبیه کردن[۹] و تجزیه مقدار تکین[۱۰] (SVD) تشکیل شده است.

گام **۱**. تعبیه کردن

در گام اول از مرحله تجزیه یعنی تعبیه کردن، ابتدا مشاهدات سری زمانی $Y_N = \{y_1, ..., y_N\}$ به صورت ماتریس $\mathbf{X} = \{X_1, ..., X_K\}$ که $X_i = (y_i, ..., y_{i+L-1})' \in \mathbb{R}^L$ نوشته می‌شود. لازم به ذکر است به این ماتریس، ماتریس مسیر می‌گویند، که در آن L طول پنجره به صورت $2 \le L \le N/2$ و $K = N - L + 1$ می‌باشد. همچنین عناصر در قطرهای فرعی این مسیر با هم برابر هستند، که چنین ماتریسی را ماتریس هنکل می‌گویند.

گام **۲**. تجزیه مقدار تکین

در این گام ماتریس مسیر به مجموع ماتریس‌هایی با رتبه یک تجزیه می‌شود. فرض کنید $\lambda_1, ..., \lambda_L$ مقادیر ویژه ماتریس $\mathbf{XX}'$ باشند که به صورت نزولی $(\lambda_1 \ge ... \ge \lambda_L \ge 0)$ مرتب شده‌اند و $U_1, ..., U_L$ نیز بردارهای یکه متعامد ویژه متناظر با مقادیر ویژه $\lambda_1, ..., \lambda_L$ و $d = max\{i, \ \lambda_i > 0\} = rank(\mathbf{X})$ باشند. در این صورت SVD ماتریس X به صورت $\mathbf{X} = \mathbf{X}_1 + ... + \mathbf{X}_d$ نوشته می‌شود، که $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i'$ و $V_i = \mathbf{X}' U_i / \sqrt{\lambda_i}$ برای $i = 1, ..., d$ می‌باشند. همچنین سه تایی $(\sqrt{\lambda_i}, U_i, V_i)$ را سه تایی ویژه می‌نامند.

مرحله **۲**. بازسازی

این مرحله از دو گام گروه بندی و میانگین گیری قطری تشکیل شده است.

گام **۱**. گروه بندی

در این گام، هدف تشخیص مؤلفه‌های سیگنال و نویز می‌باشد. پس از محاسبه SVD، ماتریس‌های ابتدایی $\mathbf{X}_i$ به چندین گروه تقسیم شده و در هر گروه با هم جمع می شوند. بعنوان نمونه فرض کنید، اندیس‌های $i_1 ... i_r$ متناظر با گروه سیگنال و $I_r = \{i_1 ... i_r\}$ باشد، در نتیجه ماتریس $\mathbf{X}_{I_r}$ متناظر با گروه I_r یا گروه سیگنال‌ها به صورت $\mathbf{X}_{I_r} = \mathbf{X}_{i_1} + ... + \mathbf{X}_{i_r}$ تعریف می‌شود.

گام **۲**. میانگین گیری قطری

در این گام، ماتریس‌های بدست آمده از مرحله گروه‌بندی به ماتریس هنکل تبدیل می‌شوند و سپس به صورت یک سری زمانی نوشته می‌شوند.

۱.۲ پارامترهای روش SSA

دو پارامتر مهم در روش SSA وجود دارند. اولین پارامتر طول پنجره L است که در گام نشانیدن برای تشکیل ماتریس مسیر مورد نیاز است. انتخاب نادرست این پارامتر به تجزیه و گروه‌بندی نامناسب سری زمانی منجر خواهد شد. متاسفانه روشی یکتا برای تعیین L وجود ندارد ولی به هر حال، مجموعه‌ای از اصول و قواعد کلی که از پشتوانه نظری و کاربردی خوبی برخوردار بوده وجود دارند که می‌توانند در انتخاب مناسب پارامتر L راهگشا باشند. دومین پارامتری که نقش اساسی و کلیدی در بازسازی سیگنال یک سری زمانی ایفا می‌کند، تعداد سه تایی‌های ویژه‌ای است که تبیین کننده سیگنال بوده و در بازسازی آن به کار می‌روند. معمولاً این پارامتر را با r نشان داده و پارامتر بازسازی می‌نامند. برای تعیین r می‌توان از اطلاعات نهفته در مقادیر ویژه و بردارهای ویژه متناظر استفاده کرد. روش SSA پس از انجام مراحل تجزیه و بازسازی، قابلیت محاسبه پیش‌بینی را با استفاده از چندین روش پیش‌بینی دارد. در ادامه پیش‌بینی با $SSA - R$ معرفی می‌شود. علاقمندان برای کسب اطلاعات بیشتر در مورد انتخاب پارامترهای L و r و نیز آشنایی با سایر روش‌های پیش‌بینی SSA می‌توانند به گولیاندینا و ژیگلجاوسکی [۵] و صانعی و حسنی [۲۰] مراجعه نمایند.

۲.۲ پیش‌بینی بازگشتی SSA

به منظور انجام پیش‌بینی با روش SSA برای مشاهدات سری زمانی $Y_N = \{y_1, ..., y_N\}$، فرض کنید $I_r = \{i_1, ..., i_r\}$ مجموعه سه تایی ویژه متناظر با سیگنال‌ها، r تعداد سه تایی ویژه اول، $U_i \in \mathbb{R}^L$ بردارهای ویژه متناظر با مقادیر ویژه از سه تایی ویژه اول، $\underline{U}_i \in \mathbb{R}^{L-1}$، $L - 1$ مؤلفه از بردارهای ویژه U_i، π_i آخرین مؤلفه از بردارهای ویژه U_i و $\tilde{Y}_N = \{\tilde{y}_1, ..., \tilde{y}_N\}$ سری بازسازی

[9] Embedding
[10] Singular Value Decomposition

شده براساس I_r باشند. در این صورت پیش‌بینی مشاهدات جدید سری زمانی با استفاده از روش SSA به صورت زیر به دست می‌آید:

$$z_i = \begin{cases} \tilde{y}_i, & i = ۱, \ldots, N \\ \sum_{j=۱}^{L-۱} \phi_j z_{i-j}, & i = N+۱, \ldots, N+h. \end{cases} \quad (۱.۲)$$

که $z_{N+۱}, \ldots, z_{N+h}$ پیش‌بینی‌های بدست آمده تا h گام به جلو و $\phi_۱, \ldots, \phi_{L-۱}$ ضرایب پیش‌بینی روش بازگشتی SSA هستند، که به صورت زیر به دست می‌آیند:

$$R = (\phi_{L-۱}, \ldots, \phi_۱)' = \frac{۱}{۱ - \nu^۲} \sum_{i \in I_r} \pi_i \underline{U}_i \quad (۲.۲)$$

که $\nu^۲ = \sum_{i \in I_r} \pi_i^۲$ می‌باشد.

۳ مدل‌های ساختاری، مدل‌های فضای حالت و الگوریتم پالایش کالمن

۱.۳ مدل‌های ساختاری سری زمانی

در یک مدل ساختاری مجموعه‌ای از مؤلفه‌های مشاهده نشده در مدل سری زمانی قرار می‌گیرند که هر کدام تفسیری خاص از عوامل تأثیرگذار را نشان می‌دهند. استفاده از مدل‌های ساختاری باعث می‌شود که سری مشاهدات را به مؤلفه‌های غیر قابل مشاهده که باعث درک بهتر مشخصات دینامیکی سری زمانی شده، تجزیه نموده و پایه‌ای مؤثر برای تعدیلات فصلی ایجاد کرد. در ادامه می‌توان پیش‌بینی‌های بهینه‌ای را با استفاده از الگوریتم‌های خاص به دست آورد.
شکل مدل‌های ساختاری برای مشاهدات $\{y_t : t = ۱, \ldots, n\}$ به‌صورت زیر تعریف می‌شود:

$$y_t = \mu_t + \psi_t + \gamma_t + \epsilon_t, \qquad \epsilon_t \sim N(۰, \sigma_\epsilon^۲)$$

که در آن y_t مشاهده در لحظه t، μ_t مؤلفه روند، ψ_t مؤلفه دوره، γ_t مؤلفه اثرات فصلی و ϵ_t مؤلفه تغییرات نامنظم هستند و می‌توانند به صورت معادلات زیر نوشته شوند:

$$\mu_t = \mu_{t-۱} + \beta_{t-۱} + \eta_t, \qquad \eta_t \sim N(۰, \sigma_\eta^۲)$$
$$\beta_t = \beta_{t-۱} + \zeta_t, \qquad \zeta_t \sim N(۰, \sigma_\zeta^۲)$$

$$\begin{pmatrix} \psi_t \\ \psi_t^* \end{pmatrix} = \rho \begin{pmatrix} \cos \lambda_c & \sin \lambda_c \\ -\sin \lambda_c & \cos \lambda_c \end{pmatrix} \begin{pmatrix} \psi_{t-۱} \\ \psi_{t-۱}^* \end{pmatrix} + \begin{pmatrix} \kappa_t \\ \kappa_t^* \end{pmatrix}$$

$$\sum_{j=۰}^{s-۱} \gamma_{t-j} = \omega_t, \qquad \omega_t \sim N(۰, \sigma_\omega^۲)$$

در معادلات فوق β_t شیب، λ_c فرکانس مؤلفه دوره، ρ عامل میرایی دوره، s تعداد فصل‌ها در یک دوره، η_t، ζ_t، κ_t، κ_t^* و ω_t عامل‌های اغتشاشی و این عامل‌های اغتشاشی مستقل از یکدیگر و همچنین مستقل از اغتشاش ϵ_t هستند. همچنین مؤلفه‌های روند، دوره و اثرات فصلی مستقل از یکدیگرند. برای اطلاعات بیشتر درباره مدل‌های ساختاری به کاماندر و کاپمن [۴] و شاموی و استوفر [۲۱] مراجعه شود.

۲.۳ مدل فضای حالت و الگوریتم پالایش کالمن

مدل‌های فضای حالت که اولین بار توسط کالمن[۱۱] و کالمن و بوسی[۱۲] معرفی شدند، مدل‌های گسترده‌ای هستند، که بسیاری از مدل‌های خطی و غیرخطی را در بر می‌گیرند و مهم‌ترین الگوریتم آن یعنی الگوریتم پالایش کالمن برای برآورد پارامترها و پیش‌بینی نیازی به شرط مانایی و وارون‌پذیری برای تحلیل‌های آماری ندارد.
به طور خلاصه مدل فضای حالت برای مشاهدات سری زمانی $\{y_t : t = ۱, \ldots, N\}$ شامل یک معادله اندازه و یک معادله انتقال

مرتبط به معادله اندازه از داده های مشاهده شده به یک بردار حالت[11] است، که این بردار حالت در معادله انتقال از یک فرآیند مارکوف به دست می‌آید. معادلات فضای حالت برای $t = ۱, \dots, N$ به صورت زیر تعریف می‌شوند:

$$y_t = Z_t\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(۰, \sigma_\epsilon^۲) \tag{۱.۳}$$
$$\alpha_t = \mathrm{T}_t\alpha_{t-۱} + \eta_t, \qquad \eta_t \sim N(۰, \mathrm{Q}_t) \tag{۲.۳}$$

که در این معادله y_t متغیر مشاهده شده، Z_t یک بردار $۱ \times m$، $\boldsymbol{T}_t$ ماتریس انتقال با بعد ماتریسی $m \times m$، α_t بردار حالت یک بردار $m \times ۱$، ϵ_t و η_t به ترتیب اغتشاش‌های معادله اندازه و معادله انتقال می‌باشند. معادلات فضای حالت با بردار حالت اولیه $\alpha_۰$ شروع می‌شود و دارای میانگین و واریانس به‌صورت

$$E(\alpha_۰) = a_۰, \qquad Var(\alpha_۰) = \boldsymbol{p}_۰$$

است. در معادلات فضای حالت اغتشاش‌های ϵ_t و η_t از یکدیگر مستقل هستند. این اغتشاش‌ها از بردار حالت اولیه $\alpha_۰$ نیز ناهمبسته هستند. یعنی

$$E(\epsilon_t\eta_s') = ۰, \qquad E(\eta_t\alpha_۰') = ۰, \qquad E(\epsilon_t\alpha_۰') = ۰, \qquad s, t = ۱, \dots, n$$

در معادلات فضای حالت ماتریس‌های $\boldsymbol{T}_t$، $\boldsymbol{Q}_t$ و Z_t ماتریس‌های سیستمی نامیده می‌شود. اگر این ماتریس‌ها برحسب زمان ثابت باشند، مدل زمان–همگن نامیده می‌شود که مدل‌های مانا را می‌پوشانند.

مثال **۱.۳**. ساده‌ترین مدل ساختاری مدل روند خطی موضعی است که با استفاده از معادلات تعریف شده به‌صورت

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, & \epsilon_t &\sim N(۰, \sigma_\epsilon^۲) \\ \mu_t &= \mu_{t-۱} + \beta_{t-۱} + \eta_t, & \eta_t &\sim N(۰, \sigma_\eta^۲) \\ \beta_t &= \beta_{t-۱} + \zeta_t, & \zeta_t &\sim N(۰, \sigma_\zeta^۲) \end{aligned}$$

نوشته می‌شود، که در آن شکل مدل فضای حالت به‌صورت

$$\alpha_t = \begin{pmatrix} \mu_t \\ \beta_t \end{pmatrix}, \qquad Z_t = \begin{pmatrix} ۱ & ۰ \end{pmatrix},$$

$$\boldsymbol{T}_t = \begin{pmatrix} ۱ & ۱ \\ ۰ & ۱ \end{pmatrix}, \qquad \boldsymbol{Q}_t = \begin{pmatrix} \sigma_\eta^۲ & ۰ \\ ۰ & \sigma_\zeta^۲ \end{pmatrix},$$

است. در ادامه به معرفی الگوریتم پالایش کالمن که مهم‌ترین الگوریتم در تجزیه و تحلیل مدل‌های فضای حالت است، پرداخته می‌شود.

پالایش کالمن

از معروف‌ترین الگوریتم‌ها برای تجزیه و تحلیل معادلات فضای حالت، الگوریتم پالایش کالمن است. در مدل فضای حالت گاوسی، پالایش کالمن مجموعه‌ای از روش‌های بازگشتی برای بدست آوردن برآورد بردار α_t، پالایش کردن و پیش‌بینی داده‌های سری زمانی براساس مشاهدات $Y_t = \{y_۱, \dots, y_t\}$ است. به منظور اجرا کردن الگوریتم پالایش کالمن فرض کنید $Y_t = \{y_۱, ..., y_t\}$ و $a_{t-۱}$ بهترین برآوردکننده بر اساس مشاهدات $Y_{t-۱}$ برای $\alpha_{t-۱}$ به صورت

$$a_{t-۱} = a_{t-۱|t-۱} = E(\alpha_{t-۱}|Y_{t-۱}),$$

باشد و $\boldsymbol{p}_{t-۱}$ نیز نشان‌دهنده ماتریس کوواریانس با بعد $m \times m$ برای خطای این برآوردکننده باشد. یعنی

$$\boldsymbol{p}_{t-۱} = \boldsymbol{p}_{t-۱|t-۱} = Var\left[\alpha_{t-۱}|Y_{t-۱}\right],$$

[11] State Vector

حال اگر a_{t-1} و $\boldsymbol{p}_{t-1}$ در اختیار باشند، آنگاه برآورد α_t و ماتریس کوواریانس برآورد $\boldsymbol{p}_t$ با توجه به اطلاعات موجود تا زمان $t-1$ به‌صورت

$$a_{t|t-1} = E(\alpha_t|Y_{t-1}) = \boldsymbol{T}_t E(\alpha_{t-1}|Y_{t-1}), \quad (۳.۳)$$
$$\boldsymbol{p}_{t|t-1} = Var(\alpha_t|Y_{t-1}) = \boldsymbol{T}_t Var(\alpha_{t-1}|Y_{t-1})\boldsymbol{T}'_t + \boldsymbol{Q}_t. \quad (۴.۳)$$

بوده و معادلات پیش‌بینی نامیده می‌شوند. بنابراین مقدار پیش‌بینی y_t در لحظه $t-1$ یا مقدار y_t بدون نویز به‌صورت

$$\hat{y}_{t|t-1} = E(y_t|Y_{t-1}) = Z_t a_{t|t-1}, \quad (۵.۳)$$

به‌دست می‌آید. آن‌گاه υ_t خطای پیش‌بینی یا نوساز در لحظه $t-1$ عبارت است از:

$$\upsilon_t = y_t - \hat{y}_{t|t-1} = Z_t(\alpha_t - a_{t|t-1}) + \epsilon_t, \quad t = 1, \ldots, n$$

و ماتریس کوواریانس این نوسازهای متعامد به‌صورت

$$F_t = Var(\upsilon_t|Y_{t-1}) = Z_t \boldsymbol{p}_{t|t-1} Z'_t + {\sigma_t}^2,$$

می‌باشد. وقتی مشاهدات y_t در لحظه t مشاهده شود برآوردگر α_t و ماتریس کوواریانس آن به‌صورت

$$a_t = a_{t|t} = E(\alpha_t|Y_t) = a_{t|t-1} + \boldsymbol{p}_{t|t-1} Z'_t F_t^{-1} \upsilon_t, \quad (۶.۳)$$
$$\boldsymbol{p}_t = \boldsymbol{p}_{t|t} = \boldsymbol{p}_{t|t-1} - \boldsymbol{p}_{t|t-1} Z'_t F_t^{-1} Z_t \boldsymbol{p}'_{t|t-1}. \quad (۷.۳)$$

به‌هنگام می‌شود. معادلات پیش‌بینی و معادلات به‌هنگام تواماً الگوریتم پالایش کالمن را می‌سازند. برای توضیحات کامل‌تر به شاموی و استوفر [۲۱] مراجعه شود. در ادامه روش پیش‌بینی بازگشتی SSA با استفاده از الگوریتم پالایش کالمن معرفی می‌شود.

۴ روش ترکیبی پیش‌بینی بازگشتی $SSA-R$ با استفاده از الگوریتم پالایش کالمن

در این بخش روش پیش‌بینی بازگشتی SSA با استفاده از الگوریتم پالایش کالمن معرفی می‌شود. فرض کنید سری زمانی شامل دو مؤلفه سیگنال و نویز به صورت زیر باشد:

$$y_t = s_t + \epsilon_t. \quad (۱.۴)$$

که $\{y_t : t = 1, \ldots, N\}$ مشاهدات، s_t مؤلفه سیگنال و ϵ_t مؤلفه نویز می‌باشند. همان طوری که اشاره شد، برای پیش‌بینی مشاهدات در یک سری زمانی با طول پنجره ثابت L، تعداد $L-1$ ضریب $\phi_1, \ldots, \phi_{L-1}$ وجود دارد، که با استفاده از بردارهای ویژه ماتریس مسیر X بدست می‌آیند که نقش مهمی در پیش‌بینی دارند. در روش SSA، اگر سری زمانی شامل نویز نباشد، یعنی $y_t = s_t$، مقدار دقیق این ضرایب را می‌توان بدست آورد، اما در داده‌های واقعی با توجه اینکه مشاهدات معمولاً دارای نویز هستند، ماتریس مسیر و بردارهای ویژه بدست آمده از آن نیز نویزی می‌باشند. بنابراین ضرایب پیش‌بینی نیز شامل نویز بوده و از دقت کافی برخوردار نمی‌باشند. بنابراین برای دستیابی به داده های با نویز کمتر و بهبود پیش‌بینی روش $SSA-R$ می‌توان از معادلات فضای حالت و الگوریتم های پالایش کالمن بهره جست. با استفاده از این ایده، روش ترکیبی برای پیش‌بینی SSA براساس پالایش کالمن، که با نماد $(KF-SSA-R)$ نشان داده می‌شود، به دست می‌آید.

فرض کنید $\{y_t : t = 1 \ldots N\}$ مشاهدات اولیه سری زمانی باشند که معادلات فضای حالت برای آن‌ها به صورت معادلات (۱.۳) و (۲.۳) تعریف شده باشند و $\{\hat{y}_t : t = 1 \ldots N\}$ یک سری زمانی با نویز کم باشند که به وسیله الگوریتم پالایش کالمن معادلات فضای حالت در معادله (۵.۳) تولید شده باشند. سپس با جایگزینی سری زمانی کم نویز $\{\hat{y}_t : t = 1 \ldots N\}$ به جای مشاهدات $\{y_t : t = 1 \ldots N\}$ در روش SSA اصلی، فرض کنید $\tilde{\mathbf{X}}$ ماتریس مسیر بدست آمده برای سری زمانی $\{\hat{y}_t : t = 1 \ldots N\}$ از گام تعبیه کردن مرحله تجزیه روش SSA اصلی، $\tilde{U}_1 \ldots \tilde{U}_L$ و $\tilde{\lambda}_1 \ldots \tilde{\lambda}_L$ به ترتیب مقادیر ویژه و بردارهای ویژه ماتریس $\tilde{\mathbf{X}}\tilde{\mathbf{X}}'$ باشند. همچنین فرض کنید I_r مجموعه سه تایی ویژه انتخاب شده باشد، در این صورت ضرایب پیش‌بینی روش $KF-SSA-R$ با استفاده از معادله (۲.۲) به صورت زیر به‌دست می‌آیند:

$$R = (\tilde{\phi}_{L-1}, \ldots, \tilde{\phi}_1)' = \frac{1}{1-\tilde{\nu}^2} \sum_{i \in I_r} \tilde{\pi}_i \underline{\tilde{U}}_i, \quad (۲.۴)$$

که بردار $\underline{\tilde{U}}_i$ شامل $L-1$ مؤلفه از بردار $\tilde{U}_i$، $\tilde{\pi}_i$ آخرین مؤلفه از بردار $\tilde{U}_i$ و $\tilde{\nu}^2 = \sum_{i\in I_r} \tilde{\pi}_i^2$ می‌باشند. در نتیجه پیش‌بینی h گام به جلو برای روش $KF-SSA-R$ با استفاده از معادله (۱.۲) به صورت زیر تعریف می‌شود:

$$z_i = \begin{cases} \tilde{\tilde{y}}_i, & i = 1, \dots, N. \\ \sum_{j=1}^{L-1} \tilde{\phi}_j z_{i-j}, & i = N+1, \dots, N+h. \end{cases} \quad (۳.۴)$$

که $\tilde{\tilde{y}}_i$ برای $i = 1, ..., N$ سری بازسازی شده بوسیله مجموعه سه تایی ویژه I_r می‌باشد. لازم به ذکر است که نویز زیاد در داده‌های سری زمانی موجب نادقیق بودن در انتخاب سه تایی ویژه نیز شده و استفاده از روش $KF-SSA-R$ می‌تواند به دقیق‌تر شدن سه تایی ویژه کمک نماید. همچنین روش $KF-SSA-R$ باعث تجزیه بهتر مشاهدات سری به مؤلفه‌های غیر قابل مشاهده شده و در نتیجه می‌تواند موجب بهبود پیش‌بینی شود. در بخش بعدی به معرفی روش جانهی داده‌های گمشده با استفاده از روش $KF-SSA$ و چندین روش معروف جانهی دیگر پرداخته می‌شود.

۵ روش‌های جانهی

فرض کنید $\hat{y}_i$، مقدار جانهی شده‌ی iامین مقدار گمشده باشد. روش‌های جانهی در سری‌های زمانی یک متغیره که در این بخش مورد استفاده قرار خواهند گرفت عبارتند از:

۱- درون‌یابی: در این روش از سه نوع درون‌یابی خطی، اسپلاین و استینمن برای جانهی استفاده می‌شود.

۲- هموارسازی کالمن : در این روش از هموارسازی کالمن برای نمایش فضای حالت یک مدل ARIMA یا مدل‌های ساختاری استفاده می‌شود.

۳- جانهی گمشده با مشاهده‌ی قبلی **(LOCF)**: در این روش، هر مقدار گمشده با مقدار بلافاصله قبل از خود، جانهی می‌شود. به عبارت دیگر: $\hat{y}_i = y_{i-1}$.

۴- جانهی گمشده با مشاهده‌ی بعدی **(NOCB)**: در این روش، هر مقدار گمشده با مقدار بلافاصله بعد از خود، جانهی می‌شود. به عبارت دیگر: $\hat{y}_i = y_{i+1}$.

۵- میانگین متحرک موزون: در این روش، هر مقدار گمشده با میانگین متحرک موزون جانهی می‌شود. برای محاسبه‌ی این نوع میانگین، تعداد مشاهدات یکسانی در دو سمت چپ و راست مقدار گمشده به کار می‌روند. به عنوان مثال از مشاهدات $y_{i-2}, y_{i-1}, y_{i+1}, y_{i+2}$ در محاسبه‌ی میانگین متحرک به طول ۴ (۲ مشاهده در سمت چپ و ۲ مشاهده در سمت راست) استفاده می‌شود. در این مقاله، از میانگین متحرک به طول ۸ (۴ مشاهده در سمت چپ و ۴ مشاهده در سمت راست)، در سه نوع مختلف زیر استفاده خواهد شد:

- میانگین متحرک ساده **(SMA)**: تمام مشاهدات به کار رفته در محاسبه‌ی این نوع میانگین، وزن یکسانی دارند. بنابراین مقدار جانهی برابر است با:

$$\hat{y}_i = \frac{1}{8} \sum_{\substack{j=-4 \\ j\neq 0}}^{4} y_{i-j}$$

- میانگین متحرک موزون خطی **(LWMA)**: وزن مشاهداتی که در محاسبه‌ی این نوع میانگین به کار می‌روند، با فاصله گرفتن از مقدار گمشده به طور خطی کاهش می‌یابد. نزدیک‌ترین مشاهدات به iامین مقدار گمشده یعنی y_{i-1} و y_{i+1}، هرکدام وزن $\frac{1}{4}$ دارند. دو مشاهده‌ی دورتر یعنی y_{i-2} و y_{i+2}، هرکدام وزن $\frac{1}{3}$ دارند؛ دو مشاهده‌ی y_{i-3} و y_{i+3}، هرکدام وزن $\frac{1}{4}$ و الی آخر. در این حالت مقدار جانهی برابر است با:

$$\hat{y}_i = \sum_{\substack{j=-4 \\ j\neq 0}}^{4} \frac{y_{i-j}}{|j|+1}$$

- میانگین متحرک موزون نمایی **(EWMA)**: در این نوع میانگین، با دور شدن مشاهدات از مقدار گمشده، وزن‌های مربوطه به طور نمایی کاهش می‌یابند. به عبارت دیگر مشاهدات y_{i-1} و y_{i+1} هرکدام وزن $\frac{1}{2^1}$ دارند. مشاهداتy_{i-2} و y_{i+2} ، هر کدام وزنی برابر با $\frac{1}{2^2}$ داشته و مشاهدات y_{i-3} و y_{i+3} ، هر کدام وزنی برابر با $\frac{1}{2^3}$ و الی آخر. در این حالت مقدار جانهی برابر است با:

$$\hat{y}_i = \sum_{\substack{j=-4 \\ j\neq 0}}^{4} \frac{y_{i-j}}{2^{|j|}}$$

۶- میانگین کل: در این روش، هر مقدار گمشده با میانگین حسابی سایر مشاهدات جانهی می‌شود.

۷- میانه: در این روش، هر مقدار گمشده با میانه‌ی سایر مشاهدات جانهی می‌شود.

۸- نما: در این روش، هر مقدار گمشده با مشاهده‌ای که بیشترین فراوانی را داشته باشد، جانهی می‌شود. اگر دو یا چند مشاهده فراوانی یکسانی داشته باشند، کوچکترین آنها انتخاب می‌شود.

۹- روش **SSA**: در این روش از SSA که مبتنی بر پیش‌بینی می‌باشد برای جانهی مقادیر گمشده استفاده می‌شود. در این روش ابتدا داده‌های سری زمانی به دو زیر سری زمانی قبل و بعد از مشاهده گمشده تقسیم می‌شوند و سپس برای هر دو زیر سری زمانی با استفاده از روش SSA پیش‌بینی صورت می‌گیرد که میانگین این دو پیش‌بینی به به عنوان مقدار جانهی در نظر گرفته می‌شود.

۱۰- روش **KF-SSA**: در این روش از KF-SSA برای جانهی مقادیر گمشده استفاده می‌شود. در این روش ابتدا داده‌های سری زمانی به دو زیر سری زمانی قبل و بعد از مشاهده گمشده تقسیم می‌شوند و سپس برای هر دو زیر سری زمانی با استفاده از روش KF-SSA پیش‌بینی صورت می‌گیرد که میانگین این دو پیش‌بینی به عنوان مقدار جانهی در نظر گرفته می‌شود.

همچنین لازم به ذکر است برای اطلاعات بیشتر درباره‌ی روش جانهی SSA به رودریگز و کاروالهو (۲۰۱۳)، روش جانهی هموارسازی کالمن به گمز و مارول[۶] و در مورد سایر روش‌های جانهی استفاده شده به پنا و همکاران [۱۷] مراجعه شود. در بخش بعدی کارایی و دقت روش‌های معرفی شده فوق با استفاده از معیار ریشه میانگین مربعات خطا و میانگین قدر مطلق انحرافات مورد مقایسه قرار می‌گیرد.

۶ بررسی کارایی روش‌های جانهی

فرض کنید $Y_N^{(i)} = \{y_1, \ldots, y_{i-1}, \star, y_{i+1}, \ldots, y_N\}$ یک سری زمانی به طول N بوده به طوری که iامین مقدار آن گمشده باشد. نماد $\star$ نشان دهنده‌ی مقدار گمشده است و مقدار جانهی شده‌ی آن را با $\hat{y}_i$ نشان داده می‌شود. در این بخش با استفاده از شبیه‌سازی، روش‌های مختلف جانهی بر اساس معیارهای ریشه میانگین مربعات خطا RMSE و میانگین قدر مطلق انحرافات MAD که به صورت زیر تعریف می‌شود، مقایسه خواهند شد:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2},$$

$$MAD = \frac{1}{N}\sum_{i=1}^{N} |e_i|,$$

که $e_i = y_i - \hat{y}_i$ برابر با خطای جانهی در مکان iام $(i = 1, \ldots, N)$ است. برای مقایسه‌ی بهتر روش KF-SSA با سایر روش‌ها، از نسبت‌های زیر استفاده شده است:

$$RRMSE = \frac{KF\text{-SSA روش } \text{بر اساس } RMSE}{\text{بر اساس روش دیگر } RMSE},$$

$$RMAD = \frac{KF\text{-SSA روش } \text{بر اساس } MAD}{\text{بر اساس روش دیگر } MAD},$$

اگر هر یک از نسبت‌های فوق کمتر از یک باشند آن‌گاه می‌توان نتیجه گرفت که روش KF-SSA بهتر از روش دیگر است. لازم به ذکر است که برای انجام محاسبات مربوط به روش SSA، پالایش کالمن و روش‌های جانهی، از محیط برنامه نویسی نرم افزار R و بسته‌های $Rssa$، $Stats$ و `imputeTS` استفاده شده است. برای کسب اطلاعات بیشتر در مورد بسته‌ی `imputeTS` به موریتس [۱۶] مراجعه کنید.

۱.۶ مطالعات شبیه‌سازی

در مطالعات شبیه‌سازی، ۲۰۰ داده با استفاده از مدل‌های مختلف ساختاری تولید شده است. همچنین به منظور داشتن مقادیر گمشده در سری‌های زمانی شبیه‌سازی شده، دو مشاهده در زمان‌های ۱۰۰ و ۱۰۱ حذف شده است. برای مقایسه روش‌های جانهی، شبیه‌سازی برای هر یک از مدل‌ها در ۱۰۰۰ تکرار انجام شده و میانگین $RRMSE$ و RMAD برای نسبت واریانس سیگنال به نویز $SNR = ۱۰$ محاسبه شده است. همچنین به منظور ارزیابی تأثیر سطح نویز و طول سری زمانی بر عملکرد روش‌های جانهی SSA و $KF-SSA$، از انواع مختلف نسبت واریانس سیگنال به نویز (SNR) و طول سری زمانی مختلف استفاده شده و بدین منظور مقادیر $SNR = ۰/۱۲۵, ۰/۵, ۱, ۱/۵, ۱۰$ و $N = ۵۰, ۱۰۰, ۲۰۰, ۳۰۰$ درنظر گرفته شده و دو مشاهده وسط سری زمانی به عنوان دو مشاهده گمشده حذف شده است. لازم به ذکر است در روش‌های SSA و $KF-SSA$، تعداد سه تایی ویژه یعنی (r)، برای بازسازی و پیش‌بینی بر اساس رتبه ماتریس مسیر به دست آمده است.

مثال **۱.۶**. در این مثال سری زمانی تولید شده از مدل قدم تصادفی همراه با اغتشاش زیر را در نظر بگیرید:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, & \epsilon_t &\sim N(۰, \sigma_\epsilon^۲) \\ \mu_t &= \mu_{t-۱} + \eta_t, & \eta_t &\sim N(۰, \sigma_\eta^۲) \end{aligned}$$

که در آن $\mu_۰ = ۲/۵$، ϵ_t مؤلفه نویز و η_t مؤلفه اغتشاشی و مستقل از ϵ_t می‌باشند. برای جانهی بر اساس روش‌های SSA و $KF-SSA$ داده‌ها به دو زیر سری زمانی قبل و بعد از مقادیر گمشده تقسیم شده و برای هر دو زیر سری زمانی نیاز به پارامترهای L و r می‌باشد که این پارامترها به ترتیب برابر ۴۸ و تعداد سه تایی ویژه با توجه به رتبه ماتریس مسیر برابر با ۱ در نظر گرفته شده است. در جدول ۱ روش جانهی $KF-SSA$ با سایر روش‌های مختلف جانهی بر حسب معیارهای $RRMSE$ و $RMAD$ مقایسه شده‌اند. نتایج نشان می‌دهند که روش KF-SSA عملکرد بسیار بهتری در مقایسه با سایر روش‌ها دارد. روش SSA اصلی نیز در رتبه‌ی بعدی قرار داشته و بدترین روش جانهی روش نما است. در جدول ۲ روش جانهی $KF-SSA$ با روش جانهی SSA بر حسب معیارهای $RRMSE$ و $RMAD$ مقایسه شده‌اند. نتایج نشان می‌دهند که روش KF-SSA عملکرد بسیار بهتری در مقایسه با روش جانهی SSA برای انواع مختلف نسبت واریانس سیگنال به نویز در طول سری‌های زمانی مختلف دارد.

جدول ۱: مقایسه روش جانهی $KF-SSA$ با سایر روش‌های مختلف جانهی در سری زمانی مربوط به مدل قدم تصادفی همراه با اغتشاش

روش	RRMSE	RMAD
SSA اصلی	۰/۴۲	۰/۴۱
هموارسازی کالمن	۰/۳۹	۰/۳۸
درون‌یابی خطی	۰/۳۷	۰/۳۸
درون‌یابی اسپلاین	۰/۳۴	۰/۳۶
درون‌یابی استینمن	۰/۳۶	۰/۳۷
LOCF	۰/۱۴	۰/۱۵
NOCB	۰/۱۴	۰/۱۵
SMA	۰/۲۵	۰/۲۷
LWMA	۰/۲۷	۰/۲۹
EWMA	۰/۳۲	۰/۳۲
میانگین کل	۰/۰۹	۰/۰۸
میانه	۰/۰۹	۰/۰۸
نما	۰/۰۵	۰/۰۵

جدول ۲: مقایسه روش جانهی $KF-SSA$ با روش جانهی SSA در سری زمانی مربوط به مدل قدم تصادفی همراه با اغتشاش

n=300		n=200		n=100		n=50		
RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	SNR
۰/۴۱	۰/۴۲	۰/۴۱	۰/۴۲	۰/۴۶	۰/۴۶	۰/۵۵	۰/۵۷	۱۰
۰/۴۰	۰/۴۰	۰/۴۰	۰/۴۰	۰/۴۳	۰/۴۴	۰/۵۵	۰/۵۵	۱/۵
۰/۳۹	۰/۴۰	۰/۳۹	۰/۴۰	۰/۴۳	۰/۴۵	۰/۵۴	۰/۵۵	۱
۰/۳۹	۰/۳۹	۰/۳۹	۰/۳۹	۰/۴۲	۰/۴۴	۰/۵۴	۰/۵۵	۰/۵
۰/۳۹	۰/۳۹	۰/۳۹	۰/۳۹	۰/۴۲	۰/۴۳	۰/۵۴	۰/۵۵	۰/۱۲۵

مثال **۲.۶**. در این مثال سری زمانی تولید شده از مدل روند خطی موضعی زیر را در نظر بگیرید:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, & \epsilon_t &\sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t &\sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \zeta_t, & \zeta_t &\sim N(0, \sigma_\zeta^2) \end{aligned}$$

که در آن $\mu_0 = 2/5$، $\beta_0 = 0/005$، ϵ_t مؤلفه نویز، η_t و ζ_t مؤلفه‌های اغتشاشی مستقل از یکدیگر می‌باشند. برای جانهی بر اساس روش‌های *SSA* و $KF-SSA$ داده‌ها به دو زیر سری زمانی قبل و بعد از مقادیر گمشده تقسیم شده و برای هر دو زیر سری زمانی نیاز به پارامترهای L و r می‌باشد که این پارامترها به ترتیب برابر ۴۸ و تعداد سه تایی ویژه با توجه به رتبه ماتریس مسیر برابر با ۲ در نظر گرفته شده است. در جدول ۳ روش جانهی $KF-SSA$ با سایر روش‌های مختلف جانهی بر حسب معیارهای *RRMSE* و *RMAD* مقایسه شده‌اند. نتایج نشان می‌دهند که روش KF-*SSA* عملکرد بسیار بهتری در مقایسه با سایر روش‌ها دارد. روش *SSA* اصلی نیز در رتبه‌ی بعدی قرار داشته و بدترین روش جانهی روش نما است. در جدول ۴ روش جانهی $KF-SSA$ با روش جانهی SSA بر حسب معیارهای *RRMSE* و *RMAD* مقایسه شده‌اند. نتایج نشان می‌دهند که روش KF-*SSA* در مقایسه با روش جانهی SSA برای انواع مختلف نسبت واریانس سیگنال به نویز در طول سری‌های زمانی مختلف از عملکرد بسیار بهتری برخوردار می‌باشد.

جدول ۳: مقایسه روش جانهی $KF-SSA$ با سایر روش‌های مختلف جانهی در سری زمانی مربوط به مدل روند خطی موضعی

RMAD	RRMSE	روش
۰/۴۲	۰/۴۳	SSA اصلی
۰/۴۰	۰/۴۱	هموارسازی کالمن
۰/۳۹	۰/۳۸	درون‌یابی خطی
۰/۳۶	۰/۳۵	درون‌یابی اسپلاین
۰/۳۷	۰/۳۷	درون‌یابی استینمن
۰/۱۵	۰/۱۴	LOCF
۰/۱۵	۰/۱۴	NOCB
۰/۲۷	۰/۲۶	SMA
۰/۲۹	۰/۲۷	LWMA
۰/۳۴	۰/۳۳	EWMA
۰/۰۷	۰/۰۸	میانگین کل
۰/۰۷	۰/۰۸	میانه
۰/۰۴	۰/۰۴	نما

جدول ۴: مقایسه روش جانهی $KF-SSA$ با روش جانهی SSA در سری زمانی مربوط به مدل روند خطی موضعی

n=300		n=200		n=100		n=50		
RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	SNR
۰/۴۱	۰/۴۳	۰/۴۲	۰/۴۳	۰/۴۶	۰/۴۸	۰/۵۸	۰/۵۹	۱۰
۰/۴۰	۰/۴۱	۰/۴۱	۰/۴۱	۰/۴۷	۰/۴۷	۰/۵۶	۰/۵۸	۱/۵
۰/۳۹	۰/۴۰	۰/۳۹	۰/۴۰	۰/۴۵	۰/۴۶	۰/۵۵	۰/۵۸	۱
۰/۴۰	۰/۴۰	۰/۴۰	۰/۴۰	۰/۴۵	۰/۴۴	۰/۵۵	۰/۵۷	۰/۵
۰/۳۹	۰/۳۹	۰/۳۹	۰/۳۹	۰/۴۳	۰/۴۴	۰/۵۴	۰/۵۶	۰/۱۲۵

مثال **۳.۶**. در این مثال سری زمانی تولید شده از مدل فصلی ساختگی سه ماهه زیر را در نظر بگیرید:

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \epsilon_t, && \epsilon_t \sim N(0, \sigma_\epsilon^2) \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, && \eta_t \sim N(0, \sigma_\eta^2) \\ \beta_t &= \beta_{t-1} + \zeta_t, && \zeta_t \sim N(0, \sigma_\zeta^2) \\ \sum_{j=0}^{3} \gamma_{t-j} &= \omega_t, && \omega_t \sim N(0, \sigma_\omega^2) \end{aligned}$$

در معادلات فوق γ_t مؤلفه اثرات فصلی ، $\mu_0 = 2/5$، $\beta_0 = 0/005$ ، η_t، ζ_t و ω_t عامل‌های اغتشاشی مستقل از یکدیگر و نیز مستقل از اغتشاش ϵ_t هستند. همچنین مؤلفه‌های روند، دوره و اثرات فصلی مستقل از یکدیگرند. برای جانهی بر اساس روش‌های SSA و $KF-SSA$ داده‌ها به دو زیر سری زمانی قبل و بعد از مقادیر گمشده تقسیم شده زیر سری زمانی اول مشاهدات ۱ تا ۹۹ و زیر سری زمانی دوم مشاهدات ۱۰۲ تا ۲۰۰ به طور برعکس می‌باشد. همچنین برای هر دو زیر سری زمانی نیاز به پارامترهای L و r می‌باشد که این پارامترها به ترتیب برابر ۴۸ و تعداد سه تایی ویژه با توجه به رتبه ماتریس مسیر برابر با ۵ در نظر گرفته شده است. در جدول ۵ مقادیر *RRMSE* و *RMAD* مربوط به روش‌های مختلف جانهی، ارائه شده است. نتایج نشان می‌دهند که همانند سری به دست آمده از مدل روند خطی موضعی، روش KF-*SSA* عملکرد بسیار خوبی نسبت به سایر روش‌ها دارد. در اینجا نیز روش *SSA* اصلی، هموار ساز کالمن و درون‌یابی خطی در رتبه‌های بعدی قرار داشته و بدترین روش جانهی روش نما است. در جدول ۶ روش جانهی $KF-SSA$ با روش جانهی SSA بر حسب معیارهای *RRMSE* مقایسه شده‌اند. نتایج نشان می‌دهند که روش KF-*SSA* عملکرد بسیار بهتری در مقایسه با روش جانهی SSA برای انواع مختلف نسبت واریانس سیگنال به نویز در طول سری‌های زمانی مختلف برخوردار می‌باشد.

جدول ۵: مقایسه روش جانهی $KF-SSA$ با سایر روش‌های مختلف جانهی در سری مربوط به مدل فصلی ساختگی سه ماهه

RMAD	RRMSE	روش
۰/۶	۰/۶۱	SSA اصلی
۰/۵۸	۰/۵۹	هموارسازی کالمن
۰/۵۸	۰/۵۷	درون‌یابی خطی
۰/۵۶	۰/۵۴	درون‌یابی اسپلاین
۰/۵۷	۰/۵۶	درون‌یابی استینمن
۰/۳۱	۰/۲۴	LOCF
۰/۳۱	۰/۲۴	NOCB
۰/۵۱	۰/۴۹	SMA
۰/۵۲	۰/۵۱	LWMA
۰/۵۶	۰/۵۴	EWMA
۰/۰۹	۰/۱	میانگین کل
۰/۰۹	۰/۱	میانه
۰/۰۵	۰/۰۶	نما

جدول ۶: مقایسه روش جانهی $KF-SSA$ با روش جانهی SSA در سری زمانی مربوط به مدل فصلی ساختگی سه ماهه

n=300		n=200		n=100		n=50		
RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	SNR
۰/۶۰	۰/۶۰	۰/۶۰	۰/۶۱	۰/۶۶	۰/۶۶	۰/۷۴	۰/۷۳	۱۰
۰/۵۹	۰/۶۰	۰/۵۹	۰/۵۹	۰/۶۳	۰/۶۴	۰/۷۱	۰/۷۱	۱/۵
۰/۵۹	۰/۵۷	۰/۵۹	۰/۵۸	۰/۶۳	۰/۶۵	۰/۷۰	۰/۷۰	۱
۰/۵۷	۰/۵۷	۰/۵۷	۰/۵۸	۰/۶۲	۰/۶۴	۰/۶۹	۰/۶۸	۰/۵
۰/۵۷	۰/۵۷	۰/۵۸	۰/۵۷	۰/۶۲	۰/۶۳	۰/۶۷	۰/۶۸	۰/۱۲۵

نتایج به دست آمده از مطالعات شبیه‌سازی برای سری‌های زمانی با طول ۲۰۰ و $SNR = ۰/۱۲۵, ۰/۵, ۱, ۱/۵, ۱۰$ در جدول ۷ ارائه شده است. همانطوری که مشاهده می‌شود برای سه مدل با ساختارها و سه تایی ویژه متفاوت با کاهش مقدار SNR روش جانهی $KF-SSA$ از عملکرد بهتری برخوردار است.

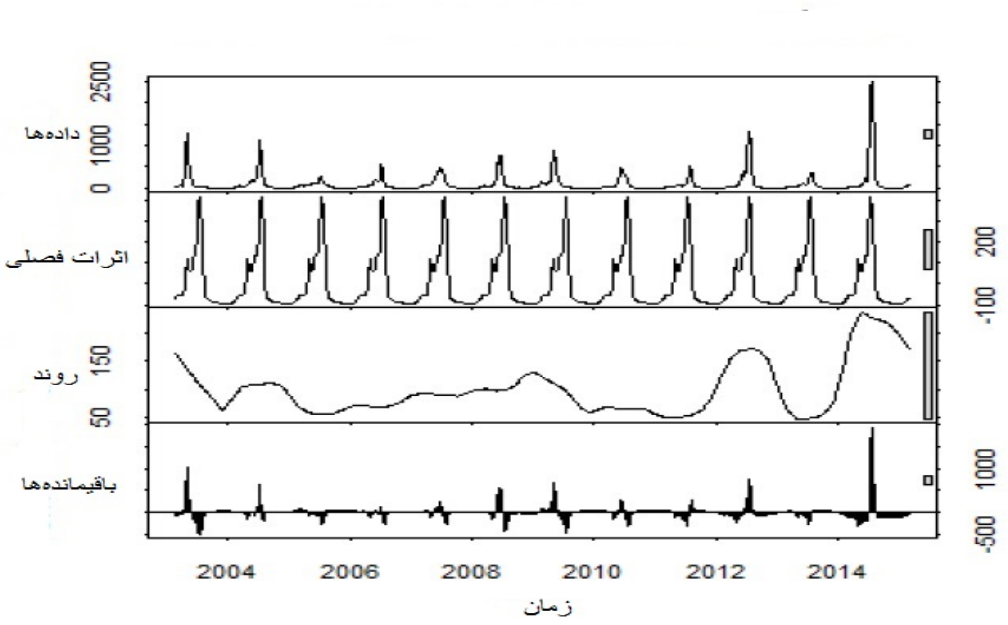
جدول ۷: مقایسه روش جانهی $KF-SSA$ با روش جانهی SSA در سری‌های زمانی شبیه‌سازی شده

مدل فصلی ساختگی		مدل روند خطی موضعی		مدل قدم تصادفی		
L=48 r=5		L=48 r=2		L=48 r=1		
RMAD	RRMSE	RMAD	RRMSE	RMAD	RRMSE	SNR
۰/۶۰	۰/۶۱	۰/۴۲	۰/۴۳	۰/۴۱	۰/۴۲	۱۰
۰/۵۹	۰/۵۹	۰/۴۱	۰/۴۱	۰/۴۰	۰/۴۰	۱/۵
۰/۵۹	۰/۵۸	۰/۳۹	۰/۴۰	۰/۳۹	۰/۴۰	۱
۰/۵۷	۰/۵۸	۰/۴۰	۰/۴۰	۰/۳۹	۰/۳۹	۰/۵
۰/۵۸	۰/۵۷	۰/۳۹	۰/۳۹	۰/۳۹	۰/۳۹	۰/۱۲۵

۲.۶ داده‌های واقعی تعداد افراد مبتلا به آنفولانزا در فرانسه

در این بخش روش‌های مختلف جانهی با استفاده داده‌های سری زمانی مربوط به تعداد مبتلایان به بیماری آنفولانزا در کشور فرانسه که توسط [۲۲] به صورت هفتگی از ماه سپتامبر ۲۰۰۳ تا ماه سپتامبر ۲۰۱۵ جمع آوری شده مورد بحث و بررسی قرار می‌گیرد. شکل ۱ نمودار سری زمانی این داده‌ها را نشان می‌دهد. در شکل ۲ نمودار تجزیه سری زمانی به مؤلفه‌های روند، اثرات فصلی و نویز با استفاده از بسته $Stats$ در برنامه R ارائه شده است. با توجه به این نمودارهای ارائه شده در این شکل وجود روند و اثرات فصلی قابل ملاحظه است.

شکل ۱: نمودار سری زمانی داده‌های تعداد آنفولانزا در فرانسه از ماه سپتامبر سال ۲۰۰۳ تا ماه سپتامبر سال ۲۰۱۵



شکل ۲: نمودار تجزیه سری زمانی به مؤلفه‌های روند، اثرات فصلی و باقیمانده‌ها

همانند سری‌های زمانی شبیه‌سازی شده، به منظور ساختن مقادیر گمشده، دو مشاهده در زمان‌های ۳۱۰ و ۳۱۱ حذف شده است. برای جانهی بر اساس روش‌های SSA و $KF-SSA$ مشاهدات به دو زیر سری زمانی به صورت مشاهدات ۱ تا ۳۰۹ و مشاهدات ۳۱۲ تا ۶۲۰ به طور برعکس تقسیم شده است و مقادیر $L=۱۵۲$ و با توجه به رتبه ماتریس مسیر $r=۷$ انتخاب شده‌اند. در جدول ۸ روش‌های مختلف جانهی بر حسب معیارهای RRMSE و RMAD مقایسه شده‌اند. نتایج حاکی از این هستند که همانند سری‌های زمانی شبیه‌سازی شده، روش KF-SSA بهتر از سایر روش‌ها است. توجه کنید که طبق معیار RRMSE، روش‌های SSA اصلی، هموار سازی کالمن، درون یابی خطی و EWMA به ترتیب در رتبه‌های بعدی و بدترین روش جانهی نما در پایین‌ترین رتبه قرار می‌گیرد.

جدول ۸: مقایسه روش‌های جانهی در داده‌های آنفولانزا در فرانسه

روش	RRMSE	RMAD
SSA اصلی	۰/۸۹	۰/۸۰
هموارسازی کالمن	۰/۸۷	۰/۸۳
درون‌یابی خطی	۰/۸۵	۰/۸۲
درون‌یابی اسپلاین	۰/۸۰	۰/۷۶
درون‌یابی استینمن	۰/۸۲	۰/۸۰
LOCF	۰/۵۸	۰/۵۷
NOCB	۰/۵۸	۰/۵۷
SMA	۰/۷۹	۰/۸۱
LWMA	۰/۸۰	۰/۸۲
EWMA	۰/۸۳	۰/۸۴
میانگین کل	۰/۴۲	۰/۳۴
میانه	۰/۳۸	۰/۳۵
نما	۰/۳۸	۰/۲۳

۷ نتیجه‌گیری

در این مقاله به منظور بهبود جانهی داده‌های گمشده روش SSA مبتنی بر پیش‌بینی هنگامی که مشاهدات آلوده به نویز هستند، روش مبتنی بر معادلات فضای حالت و الگوریتم پالایش کالمن پیشنهاد گردید. همچنین سایر روش‌های جانهی یک متغیره نظیر SSA، درون‌یابی، هموارسازی کالمن، $NOCB$، $LOCF$، میانگین متحرک موزون، میانگین، میانه و نما معرفی شدند. سپس برای ارزیابی روش $KF-SSA$ با سایر روش‌های جانهی از معیار $RMSE$ و $RMAD$ در مدل‌های ساختاری از مطالعات شبیه‌سازی شده و داده واقعی استفاده شد. همچنین در این ارزیابی برای بررسی کارایی و تأثیر نویز بر نتایج جانهی روش‌های $KF-SSA$ و SSA از SNR مختلف استفاده شد. نتایج نشان می‌دهد، هنگامی که مشاهدات آلوده به نویز زیاد و دارای مؤلفه‌های روند و اثرات فصلی هستند، روش $KF-SSA-R$ در مقایسه با روش SSA و سایر روش‌های جانهی از کارایی بهتری برخوردار می‌باشد. دلیل کاراتر بودن روش $KF-SSA-R$ تجزیه بهتر مؤلفه‌ها و در نتیجه بهبود عملکرد روش جانهی داده‌های گمشده است. در موارد بررسی شده، روش‌های جانهی میانگین کل، میانه و نما به دلیل عدم استفاده از خود همبستگی از عملکرد ضعیفی در مقایسه با سایر روش‌ها برخوردارند.

تقدیر و تشکر

نویسندگان مقاله از سردبیر محترم مجله مدل‌سازی پیشرفته ریاضی و داوران محترم همچنین ویراستار گرامی که باعث ارتقای مقاله شدند، کمال تشکر و قدردانی را دارند.

فهرست منابع

[1] Abraham, B., 1981. Missing observations in time series, *Communications in Statistics-Theory and Methods*, **10** (16), 1643-1653. doi: 10.1080/03610928108828138

[2] Broomhead, D. and King, G., 1986b. On the qualitative analysis of experimental dynamical systems, Nonlinear Phenomena and Chaos.

[3] Chatfield, C., 2000. *Time-Series Forecasting*, Chapman & Hall/CRC.

[4] Commandeur, J. F. and Koopman, S. J. (2007) *An Introduction to State Space Time Series Analysis*, Oxford University Press Inc, New York.

[5] Golyandina, N. and Zhigljavsky, A., 2013. *Singular Spectrum Analysis for Time Series*, Springer.

[6] Gomez, V. and Maravall, A., 1994. Estimation, Prediction, and Interpolation for Nonstationary Series with the Kalman Filter, *Journal of the American Statistical Association*, **89** (426), 611-624. doi: 10.1080/01621459.1994.10476786

[7] Harvey, A. C. and Pierse, R. G., 1984. Estimating Missing Observations in Economic Time Series, *Journal of the American Statistical Association*, **79** (385), 125-131. doi: 10.1080/01621459.1984.10477074

[8] Hui-zan, W., Rein, Z., Wei, L., Gui-hua, W. and Bao-gang, J., 2008. Improved interpolation method based on singular spectrum analysis iteration and its application to missing data recovery, *Applied Mathematics and Mechanics (English Edition)*, **29**, 1351–1361. doi: 10.1007/s10483-008-1010-x

[9] Junger, W. L., De Leon, A. P. and Santos, N., 2003. Missing Data Imputation in Multivariate Time Series via EM Algorithm, *Cadernos do IME*, **15**, 8-21.

[10] Junger, W. L., De Leon, A. P., 2012. mtsdi: Multivariate Time Series Data Imputation. http://CRAN.R-project.org/package=mtsdi.

[11] Kalman, R. E., 1960. A new approach to linear filtering and prediction problems, *J. of Basic Engineering*, **83**, 35-45. doi: 10.1115/1.3662552

[12] Kalman, R. E. and Bucy, R. S., 1961. New results in linear filtering and prediction theory, *J. of Basic Engineering*, **83**, 95-108. doi: 10.1115/1.3658902

[13] Kondrashov, D. and Ghil, M., 2006. Spatio-temporal filling of missing points in geophysical data sets, *Nonlinear Processes in Geophysics*, **13**, 151–159. doi: 10.5194/npg-13-151-2006

[14] Ljung, G. M., 1989. A Note on the Estimation of Missing Values in Time Series, *Communications in Statistics-Simulation and Computation*, **18** (2), 459-465. doi: 10.1080/03610918908812770

[15] Mahmoudvand, R. and Rodrigues, P. C., 2016. Missing value imputation in time series using singular spectrum analysis, *International Journal of Energy and Statistics*, **4**(1), 1650005. doi: 10.1142/S2335680416500058

[16] Moritz, S., 2016. imputeTS: Time Series Missing Value Imputation. https://CRAN.Rproject.org/package=imputeTS. R package version 1.8. doi: 10.32614/rj-2017-009

[17] Pena, D., Tiao, G. C. and Tsay, R. S., 2011. A Course in Time Series Analysis, chap. Outliers, Influential Observations and Missing Data, *John Wiley & Sons*, 136–170. doi: 10.1002/9781118032978.ch6

[18] Pourahmadi, M., 1989. Estimation and Interpolation of missing values of a stationary time series, *Journal of Time Series Analysis*, **10** (2), 149-169. doi: 10.1111/j.1467-9892.1989.tb00021.x

[19] Rodrigues, P. C. and Carvalho, M. D., 2013. Spectral modeling of time series with missing data, *Applied Mathematical Modelling*, **37**, 4676–4684. doi: 10.1016/j.apm.2012.09.040

[20] Sanei, S. and Hassani, H., 2016. *Singular Spectrum Analysis of Biomedical Signals*. Taylor & Francis/CRC.

[21] Shumway, R. H. and Stoffer, D. S., 2011. *Time Series Analysis and Application*, 3rd ed, Springer, New York.

[22] The Google Flu and Dengue Trends Team, 2015. 'Google Flu Trends'. URL: http://www.google.org/flutrends

[23] Wu, S. F., Chang, C. Y. and Lee, S. J., 2015. Time Series Forecasting with Missing Values. 1st *International Conference on Industrial Networks and Intelligent Systems (INISCom)*, 151-156. doi: 10.4108/icst.iniscom.2015.258269

Imputation of Missing Data Using the Combination of Singular Spectrum Analysis Method and Kalman Filter Algorithm and Comparison with Univariate Imputation Methods.

Masoud Yarmohammadi,[(1)] [12] Reza Zabihi Moghadam[(1)]

[(1)] Department of Statistics, Payame Noor University, Tehran 19395-4697, Iran.

Communicated by: Rahim Chinipaedaz

Received: 15 December 2022 Accepted: 16 September 2023

Abstract: Missing values in time series data are one of the problems that sometimes arise in time series analysis. The more accurate imputation of these values, the better understanding of the structure of the time series will be obtained, and as a result, the recognition of its pattern and the prediction of future values will be more accurate. Therefore, choosing an appropriate method of imputation is an important part of a time series analysis. In this paper we introduce the new missing data imputation method from the singular spectrum analysis procedure, using the Kalman filter algorithm. Then other methods of imputation of missing values in univariate time series are introduced, and will be compared the mentioned methods using simulated data in structural models and real data. The results of the comparison based on the criteria of root mean square error and mean absolute deviations show that the method of imputation of missing values based on singular spectrum analysis approach using the Kalman filter algorithm has a better performance than the other imputation methods and the mode method is the worst.

Keywords: Time Series, Missing Values, Singular Spectrum Analysis, State Space Form, Kalman Filter, Structural Models.

©2024 Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonComertial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

[12] Corresponding author.
E-mail addresses: masyar@pnu.ac.ir (M. Yarmohammadi), Rezazm63@gmail.com, (R. Zabihi Moghadam).