مقاله پژوهشی

بهبود کیفیت مدل‌سازی و پیش‌بینی سری‌های زمانی با استفاده از تحلیل مجموعه مقادیر تکین چندمتغیره استوار

طاهره امینی
دانشجوی دکترای آمار دانشگاه پیام نور، تهران، ایران. tahereamini90@yahoo.com

مسعود یارمحمدی*
(نویسنده‌ی مسؤل)، دانشیار گروه آمار دانشگاه پیام نور، تهران، ایران. masyar@pnu.ac.ir

علی شادرخ
دانشیار گروه آمار دانشگاه پیام نور، تهران، ایران. a.shadrokh@pnu.ac.ir

مهدی کلانتری
استادیار گروه آمار دانشگاه پیام نور ، تهران، ایران. kalantarimahdi@pnu.ac.ir

چکیده: در تحلیل سری‌های زمانی نادیده گرفتن نقاط دورافتاده منجر به شناسایی نادرست مدل، برآورد اریب پارامترها و در نتیجه پیش‌بینی‌های ضعیف و به عبارتی کاهش کیفیت و دقت مدل‌سازی می‌شود. یکی از روش‌های ناپارامتری معتبر در پیش‌بینی و بهبود کیفیت مدل‌سازی سری‌های زمانی چند متغیره، روش مجموعه مقادیر تکین چند متغیره[1] (MSSA) است که نیازمند هیچ‌گونه فرض اولیه‌ای نیست. از آنجایی که وجود نقاط دورافتاده کارایی روش MSSA را کاهش داده و نُرم ماتریسی فروبنیوس به کار رفته در آن را متأثر ساخته و به عبارتی غیر استوار می‌سازد، در این تحقیق، نسخه‌ی جدیدی از روش MSSA بر اساس نُرم L_1 پیشنهاد می‌شود. در ادامه با استفاده از مطالعات شبیه‌سازی و نیز استفاده از داده‌های واقعی، عملکرد روش تحلیل مجموعه مقادیر تکین چندمتغیره بر اساس هر دو نُرم مورد مقایسه قرار می‌گیرد. معیارهای مورد استفاده شامل ریشه‌ی میانگین توان دوم خطاها و میانگین قدرمطلق خطاها، برتری روش تحلیل مجموعه مقادیر تکین چندمتغیره بر اساس نُرم L_1 را در بازسازی و پیش‌بینی سری زمانی نشان می‌دهند.

واژه‌های کلیدی: تحلیل مجموعه مقادیر تکین چند متغیره، استوارسازی، نقاط دورافتاده، نُرم L_1، نسبت میانگین توان دوم خطا، نسبت میانگین قدرمطلق خطا.

۱- مقدمه

به دلیل عدم اطمینان روزافزونی که با آن مواجه هستیم، تحلیل و پیش‌بینی سرهای زمانی در تمامی حوزه‌های صنعتی، کشاورزی، اقتصادی و... از اهمیت بسیار بالایی برخوردار است. علت این امر آن است که مدل‌های پیش‌بینی دقیق، توانایی ما را در تصمیم‌گیری، برنامه‌ریزی و مدیریت ریسک افزایش می‌دهند. رسیدن به چنین منظوری تا حد زیادی به ماهیت روش پیش‌بینی (از دیدگاه پارامتری یا ناپارامتری)، دقت پیش‌بینی‌ها و توانایی پیش‌بینی صحیح جهت تغییر[2] (DC)، بستگی دارد [۱]. از این‌رو روش‌های متفاوتی برای پیش‌بینی سری‌های زمانی وجود دارد که با توجه به نوع و ماهیت داده‌ها می‌توان از آنها استفاده کرد. البته اغلب به‌منظور پیش‌بینی داده‌ها با استفاده از روش‌های کلاسیکی مانند ARIMA، فرض‌هایی خطی بودن مدل و نرمال بودن باقیمانده‌ها محدودیت‌هایی را برای رسیدن به نتایج مطلوب فراهم می‌کنند. یکی از روش‌های ناپارامتری مورد استفاده توسط محققین برای تحلیل و پیش‌بینی سری‌های زمانی به‌خصوص در زمینه‌های مهندسی و بهینه‌سازی مدل‌های خطی، روش تحلیل مجموعه مقادیر تکین[3] (SSA) بوده که به دلیل عدم نیاز به برقراری فرض‌های محدودکننده و اولیه‌ای مانند مانایی داده‌ها، نرمال بودن باقیمانده‌ها و خطی بودن مدل، مورد توجه بسیاری از محققین در این حوزه قرار گرفته است. روش تحلیل مجموعه

تاریخ دریافت: ۱۴۰۲/۰۵/۲۶ تاریخ پذیرش: ۱۴۰۲/۰۹/۰۷
دوره ۱۲/ شماره ۴
صفحات ۵۲۰-۴۹۵

*Corresponding author: masyar@pnu.ac.ir

[1] Multivariate Singular Spectrum Analysis (MSSA)
[2] Direction of Change (DC)
[3] Singular Spectrum Analysis (SSA)

مقادیر تکین برای اولین‌بار به وسیله بروم‌هد و کینگ[۴] معرفی شد. بعد از آن گیل و همکاران[۵] و گولیاندینا و استپانف[۶] ویرایش‌های دیگری از روش‌های تحلیل مجموعه مقادیر تکین یک و چند متغیره را برای پیش‌بینی مورد بحث و بررسی قرار دادند [۲]، [۳] و [۴].

روش تحلیل مجموعه مقادیر تکین چندمتغیره (MSSA) و یک متغیره (SSA) دارای الگوریتمی دو مرحله‌ای و هر مرحله شامل دو گام است. در مرحله‌ی تجزیه، سری زمانی به‌صورت یک ماتریس خاص که به آن ماتریس مسیر[۷] گفته می‌شود در می‌آید. سپس در مرحله‌ی بازسازی مشاهدات سری زمانی در ساده‌ترین حالت تجزیه خود به دو مؤلفه‌ی نوفه و سیگنال تقسیم شده و متناسب با سری بازسازی‌شده از مؤلفه‌ی سیگنال، مقادیر پیش‌بینی با استفاده از یک رابطه‌ی خطی بازگشتی[۸] (LRR) محاسبه می‌شود.

در هر بار استفاده از روش‌های تحلیل مجموعه مقادیر تکین یک و چندمتغیره و قبل از تعیین ماتریس مسیر، کاربر ملزم به تعیین «پارامتر»، طول پنجره[۹] [۵] و پس از آن در گام دوم، تعیین پارامتر بازسازی یا تعداد سه‌تایی‌های ویژه[۱۰] مورد استفاده برای بازسازی است [۶]. تعیین پارامتر بازسازی، متناسب با ساختار ماتریس مسیر سری زمانی بوده که در حالت چند متغیره بنا به تشخیص کاربر، ماتریس مسیر به دو حالت افقی و عمودی بلوک‌بندی می‌شود. برای توضیحات کامل‌تر به [۵] مراجعه شود. اگر چه هر دو روش SSA و MSSA رویکردی بهینه برای مدل‌سازی و پیش‌بینی ساختارهای متفاوت سری زمانی اعم از خطی یا غیر خطی، مانا یا نامانا و همچنین سری‌های زمانی با ساختارهایی ناشناخته و پنهان دارند [۷]، اما بر خلاف روش SSA، درباره‌ی جنبه‌های نظری MSSA تحقیقات کمتری انجام شده است، MSSA بدون داشتن پیچیدگی‌های خاصی، همانند SSA کاربردهای زیادی مانند نوفه‌زدایی، استخراج روند، علیت و پیش‌بینی دارد [۸] و [۹]. از طرفی در سال‌های اخیر به دلیل پیشرفت فناوری و قابلیت‌های ذخیره‌سازی ماشین‌های روزمره، توجه بیشتری به سمت سری‌های زمانی چندمتغیره جلب شده است. همچنین موارد بسیاری از کاربردهای MSSA را در زمینه‌هایی مانند تجارت [۱۰]، ترافیک خطوط هوایی [۱۱]، صنعت [۱۲]، [۸]، [۱۳] و [۱۴] و پیش‌بینی مرگ و میر [۱۵]، می‌توان نام برد. از طرف دیگر روش MSSA کارایی زیادی در بازسازی و بهینه‌سازی مدل‌های آماری و پیش‌بینی داده‌های مالی و اقتصادی، صنعتی و تجاری دارد و برتری این روش در برابر روش SSA و حتی روش‌های کلاسیک پارامتری و اقتصادسنجی در مطالعاتی نشان داده شده است [۱۳]، [۱۵]، [۱۶]، [۱۷] و [۱۸]. از این‌رو در این مقاله به موضوع تحلیل مجموعه مقادیر تکین چندمتغیره پرداخته خواهد شد.

تحلیل سری‌های زمانی همراه با نقاط دورافتاده یک بحث اساسی و مهم در علم آمار است. این داده‌ها اغلب به دلیل عواملی خارجی، مانند جنگ‌ها و بحران‌های سیاسی و اقتصادی، خرابی دستگاه‌ها، وقوع جریانات سیاسی، اعتصاب‌ها و... به وجود می‌آیند و باعث گسست‌های ساختاری در سری‌های زمانی نیز می‌شوند. وجود داده‌های دورافتاده بر تحلیل سری‌های زمانی و نتایج پیش‌بینی تأثیرات منفی بسیار زیادی گذاشته، می‌توانند باعث انحراف نتایج شده و در نتیجه باعث مشکلات عدیده‌ای شوند. اگر چه بر اساس مطالعات صورت‌گرفته، حساسیت مدل‌های ناپارامتری به داده‌های دورافتاده کمتر از مدل‌های پارامتری است اما به دلیل کارایی بیشتر مدل‌های ناپارامتری در سری‌های زمانی، بررسی و مطالعه در زمینه‌ی کاهش اثر منفی داده‌های دورافتاده بر این مدل‌ها می‌تواند قابل تأمل باشد [۱۹]. با توجه به این‌که ساختار روش مجموعه مقادیر تکین چندمتغیره بر اساس نُرم فروبنیوس است، از این‌رو نسبت به داده‌های دورافتاده حساس بوده و کارایی مناسبی ندارد. به‌منظور کاهش اثرات نقاط دورافتاده بر روی این مدل، در این تحقیق نُرم ماتریسی L_1 به جای نُرم فروبنیوس جایگزین و نسخه‌ی جدیدی از MSSA معرفی شده که در ادامه‌ی متن با نماد L_1 − MSSA نمایش داده می‌شود. در این راستا ابتدا روشی برای محاسبه‌ی ماتریس سیگنال بلوکی بر اساس نُرم L_1 معرفی شده و سپس روش هنکل‌سازی با استفاده از این نُرم برای حالت عمودی و افقی ماتریس مسیر به‌صورت یک قضیه بیان خواهد شد. البته لازم است تا ماتریس مسیر هر دو حالت افقی و عمودی به دقت مورد واکاوی قرار گرفته و تغییرات لازم در شیوه‌ی محاسبه‌ی ماتریس سیگنال برای ساخت ماتریس مسیر بلوک‌بندی هنکل‌سازی‌شده، ارائه شود. در ادامه با استفاده

[4] Broomhead, D. S. and King, G. P. (1986 b).
[5] Ghil, M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, et al. (2002)
[6] Golyandina, N., and Stepanov, D. (2005)
[7] Trajectory matrix
[8] Linear Recurrence Relations (LRR)
[9] Window length
[10] eigentriples

از مطالعات شبیه‌سازی عملکرد روش تحلیل مجموعه مقادیر تکین چندمتغیره بر اساس هر دو نُرم مورد مقایسه قرار می‌گیرد و با به کارگیری این روش برای سری‌های زمانی واقعی، معیارهای مورد استفاده شامل ریشه‌ی میانگین توان دوم خطاها و میانگین قدرمطلق خطاها، برتری روش تحلیل مجموعه مقادیر تکین چندمتغیره بر اساس نُرم L_1 را نسبت به نسخه‌ی قبلی، در بازسازی و پیش‌بینی سری زمانی نشان می‌دهد. در ادامه در بخش دوم علاوه بر معرفی اجمالی MSSA، راهکارهای محاسباتی روش تحلیل مجموعه مقادیر تکین چند متغیره بر اساس نرم L_1 ارائه می‌شود پس از آن با استفاده از نتایج شبیه‌سازی و به کارگیری داده‌های واقعی، کارایی روش جدید مورد ارزیابی قرار داده می‌شود.

۲- روش‌شناسی

تحلیل مجموعه مقادیر تکین چند متغیره (MSSA) یکی از روش‌های ناپارامتری برای تحلیل سری‌های زمانی چندمتغیره بوده و مانند حالت یک متغیره (SSA) شامل دو مرحله **تجزیه**[11] و **بازسازی**[12] است، که هر مرحله خود به دو گام **نشانیدن**[13] **و تجزیه‌ی مقادیر تکین** (SVD)[14] در مرحله‌ی اول و **گروه‌بندی**[15] **و میانگین‌گیری قطری**[16] در مرحله‌ی دوم دسته‌بندی می‌شوند. علاقه‌مندان برای مطالعه‌ی جزییات مدل MSSA می‌توانند به [۹] مراجعه کنند. در روش MSSA، تعریف ماتریس مسیر به دو حالت مختلف افقی و عمودی است. در مرحله‌ی تجزیه هر متغیر که یک سری زمانی می‌باشد، به طور جداگانه به چند مؤلفه‌ی قابل تفسیر همچون روند، مولفه‌ی نوسانی و نوفه تجزیه می‌شود و بعد از گروه‌بندی مناسب، مؤلفه‌های بازسازی‌شده مربوط به هر سری در یک ماتریس معین، درست به همان ترتیب و چیدمان قبل از تجزیه در کنار هم قرار می‌گیرند. به این ترتیب مؤلفه‌های بازسازی شده برای پیش‌بینی مشاهدات جدید مربوط به هر سری مورد استفاده قرار می‌گیرند. روش MSSA همانند SSA به انتخاب دو پارامتر اساسی طول پنجره و پارامتر ساختاری مرتبط با نحوه گروه‌بندی بستگی دارد و ساختار این روش بر اساس نُرم L_2 بوده که نسبت به نقاط دورافتاده استوار نیست. حسنی و همکاران (۲۰۱۴)، نشان دادند که وجود نقاط دور افتاده در یک سری زمانی می‌تواند مرتبه ماتریس مسیر، بزرگی مقادیر ویژه، بردارهای ویژه و در نتیجه تجزیه، بازسازی و پیش‌بینی مقادیر آینده سری زمانی را تحت تأثیر قرار دهند [۱۹]. لذا بر خلاف تمام مزیت‌هایی که برای روش MSSA بیان شد، این روش نسبت به نقاط دور افتاده استوار نبوده و یک نقطه ضعف برای آن محسوب می‌شود. حال با توجه به این‌که وجود نقاط دورافتاده در سری‌های زمانی بر تحلیل‌ها از جمله پیش‌بینی مقادیر آینده سری زمانی و برآورد پارامترهای مدل تأثیر منفی دارد، لزوم شناسایی و کاهش اثرات آنها از اهمیت بسیاری برخوردار است. از طرفی دیگر به دلیل کارا بودن روش MSSA نسبت به روش SSA و به‌منظور استوارسازی MSSA کلاسیک با استفاده از نُرم L_1 به جای L_2 در روش MSSA کلاسیک، نسخه جدیدی از این روش ارائه می‌شود که استفاده از این نُرم می‌تواند یکی از راه‌های استوارسازی و بهبود کارایی در این روش باشد.

همانند روش کلاسیک MSSA، روش L_1- MSSA نیز دارای دو مرحله‌ی مکمل بوده و هر مرحله هم شامل دو قدم اصلی است، گام‌های نشانیدن و گروه‌بندی در روش L_1- MSSA، همانند روش کلاسیک MSSA است اما گام‌های تجزیه‌ی مقادیر تکین و هنکل‌سازی در روش L_1- MSSA تفاوتی اساسی با این گام‌ها در روش کلاسیک MSSA دارند که در ادامه شرح داده خواهند شد.

۲-۱- تجزیه

این مرحله شامل دو گام است: **نشانیدن** و **تجزیه‌ی مقادیر تکین**

الف) نشانیدن

در روش SSA تعبیه کردن به‌عنوان یک نگاشت برای تبدیل یک سری زمانی مانند $\boldsymbol{Y_N}=(y_1, y_2, \dots, y_N)$ به طول N به‌سری‌های زمانی $X_i=(y_1, y_2, \dots, y_{i+L-1})^T \in R^L$ به‌صورت بردارهای $X_1, X_2, \dots, X_k$ در نظر گرفته می‌شود. در واقع در این گام سری زمانی Y_N به k زیرسری تبدیل می‌شود، وقتی که عدد صحیح L به‌عنوان طول پنجره تعریف شود ($2 \le L \le N$) و $K=N-L+1$. دقت کنید بردار $\mathbf{X_i}$، یک زیرسری متشکل از یک بردار ستونی با L مؤلفه است و گاهی

[11] Decomposition
[12] Reconstruction
[13] Embedding
[14] Singular Value Decomposition (SVD)
[15] Grouping
[16] Diagonal Averaging

اوقات بردار L-تأخیر[17] نامیده می‌شود. نتیجه‌ی مرحله‌ی نشانیدن، ماتریس مسیر $\boldsymbol{X} = [\mathbf{X_1}, \mathbf{X_2}, \dots, \mathbf{X_K}]$ است که تشکیل یک ماتریس $\mathrm{L} \times \mathrm{K}$ را می‌دهد و به‌صورت رابطه‌ی (۱) نمایش داده می‌شود.

$$\boldsymbol{X} = \left(x_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ y_3 & y_4 & y_5 & \cdots & y_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_{L+K-1} \end{pmatrix} \quad (۱)$$

همچنین: $X_{ij} = y_{i+j-1}$. ماتریس مسیر ساخته شده ($\boldsymbol{X}$) در این مرحله را ماتریس هنکل[18] نیز می‌نامند، زیرا عناصر روی قطرهای فرعی این ماتریس با هم برابرند [۲۰]. یکی از تفاوت‌های مراحل ساخت SSA با MSSA و $\mathrm{L_1} - \mathrm{MSSA}$ در ساختن ماتریس مسیر است. روش‌های ساخت ماتریس مسیر در مدل $\mathrm{L_1} - \mathrm{MSSA}$ به دو حالت افقی و عمودی است که در ادامه به‌صورت زیر بیان خواهد شد.

فرض کنید $y_j = \{y_j^{(1)}, y_j^{(2)}, \dots, y_j^{(p)}\}$ مشاهده‌ی jام یک نمونه‌ی به طول N از یک سری زمانی p- متغیره بوده $X^{(i)}$ $(i = 1, \dots, p)$، ماتریس مسیر متناظر با سری زمانی iام $Y_N^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \dots, y_N^{(i)}\}$ باشد. در این صورت ماتریس مسیر چندمتغیره را می‌توان با کنار هم قراردادن ماتریس‌های مسیر هر سری زمانی (به‌صورت افقی) یا زیر هم قراردادن آنها (به‌صورت عمودی) به دست آورد. در مرحله‌ی تعبیه کردن ماتریس‌ها به ترتیب در حالت افقی و عمودی (از چپ به راست) به‌صورت رابطه (۲) در خواهند آمد:

$$\boldsymbol{X_V} = \begin{bmatrix} \boldsymbol{X^{(1)}} \\ \vdots \\ \boldsymbol{X^{(p)}} \end{bmatrix} \quad \text{یا} \quad \boldsymbol{X_H} = [\boldsymbol{X^{(1)}} : \boldsymbol{X^{(2)}} : \ \dots : \ \boldsymbol{X^{(p)}}] \quad (۲)$$

چنانچه همه‌ی p- تا ماتریس‌های مسیر به‌صورت افقی کنار هم قرار داده شوند ماتریس مسیر به حالت افقی در می‌آید که با $\boldsymbol{X_H}$ نامیده می‌شود. بنا بر این در رابطه (۳) نحوه‌ی شکل‌گیری ماتریس مسیر افقی نمایش داده شده است.

$$\boldsymbol{X_H} = [\boldsymbol{X^{(1)}} : \boldsymbol{X^{(2)}} : \ \dots : \ \boldsymbol{X^{(p)}}] = \begin{bmatrix} y_1^1 & y_2^1 & y_3^1 & \cdots & y_K^1 & y_1^2 & y_2^2 & y_3^2 & \cdots & y_K^2 & \cdots & y_1^p & y_2^p & y_3^p & \cdots & y_K^p \\ y_2^1 & y_3^1 & y_4^1 & \cdots & y_{K+1}^1 & y_2^2 & y_3^2 & y_4^2 & \cdots & y_{K+1}^2 & \cdots & y_2^p & y_3^p & y_4^p & \cdots & y_{K+1}^p \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L^1 & y_{L+1}^1 & y_{L+2}^1 & \cdots & y_N^1 & y_L^2 & y_{L+1}^2 & y_{L+2}^2 & \cdots & y_N^2 & \cdots & y_L^p & y_{L+1}^p & y_{L+2}^p & \cdots & y_N^p \end{bmatrix} \quad (۳)$$

و اگر همه‌ی p- تا ماتریس‌های مسیر به‌صورت عمودی زیر هم قرار داده شوند ماتریس مسیر حالت عمودی حاصل شده و با $\boldsymbol{X_V}$ نمایش داده می‌شود در رابطه‌ی (۴) نحوه‌ی شکل‌گیری ماتریس مسیر عمودی نمایش داده شده است:

$$\mathbf{X_V} = \begin{bmatrix} \mathrm{X}^{(1)} \\ \vdots \\ \mathrm{X}^{(\mathrm{P})} \end{bmatrix} = \begin{bmatrix} y_1^1 & y_2^1 & y_3^1 & \cdots & y_K^1 \\ y_2^1 & y_3^1 & y_4^1 & \cdots & y_{K+1}^1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L^1 & y_{L+1}^1 & y_{L+2}^1 & \cdots & y_N^1 \\ y_1^2 & y_2^2 & y_3^2 & \cdots & y_K^2 \\ y_2^2 & y_3^2 & y_4^2 & \cdots & y_{K+1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L^2 & y_{L+1}^2 & y_{L+2}^2 & \cdots & y_N^2 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ y_1^p & y_2^p & y_3^p & \cdots & y_K^p \\ y_2^p & y_3^p & y_4^p & \cdots & y_{K+1}^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L^p & y_{L+1}^p & y_{L+2}^p & \cdots & y_N^p \end{bmatrix}, \quad (۴)$$

دقت به این نکته حائز اهمیت است که عناصر روی قطرهای فرعی هر ماتریس $\boldsymbol{X^{(i)}}$ برای هر بلوک با هم برابرند. چنین ماتریسی را ماتریس مسیر هنکل می‌نامند. از طرفی واضح است که با در اختیار داشتن ماتریس $\boldsymbol{X^{(i)}}$، می‌توان سری زمانی $Y_N^{(i)}$ را به دست آورد. در واقع با کنار هم قرار دادن ستون اول و سطر آخر (یا سطر اول و ستون آخر) در هر بلوک، ماتریس $\boldsymbol{X^{(i)}}$، سری زمانی $Y_N^{(i)}$ حاصل می‌شود. همچنین یک تناظر یک به یک بین ماتریس مسیر (که هنکل نیز می‌باشد) و سری زمانی وجود دارد. در جدول ۱ تفاوت بین تعیین پارامترهای بازسازی و طول پنجره متناسب با تعداد داده‌های سری زمانی و عمودی یا افقی[19] بودن ماتریس مسیر بیان می‌شود.

جدول ۱- نحوه محاسبات پارامترها برای دو حالت افقی و عمودی مجموعه مقادیر تکین چند متغیره

روش	طول سری‌ها	L	K	تعداد مقادیر ویژه
$L_1 - VMSSA$	مختلف	مختلف	یکسان	$L_{sum} = \sum_{i=1}^{p} L_i$
$L_1 - HMSSA$	مختلف	یکسان	مختلف	L

ب) تجزیه‌ی مقادیر تکین (SVD) ماتریس مسیر در حالت افقی

دومین گام از مرحله‌ی اول ساختن مدل $\mathrm{L_1} - \mathrm{MSSA}$ تجزیه‌ی ماتریس مسیر بر اساس تجزیه‌ی مقادیر تکین است. با استفاده از روش SVD، ماتریس مسیر به‌صورت مجموع ماتریس‌های

[17] L-Lagged
[18] Hankel Matrix

[19] برای نمایش حالت افقی یا عمودی روش تحلیل مجموعه مقادیر تکین چندمتغیره به ترتیب از نمادهای HMSSA و VMSSA استفاده می‌شود.

مقدماتی متعامد (هر کدام با رتبه‌ی ۱) تجزیه می‌شوند. در نسخه‌ی $L_1 - MSSA$ همانند روش کلاسیک MSSA با توجه ساختار متفاوت ماتریس مسیر برای هر دو حالت افقی یا عمودی، ابعاد ماتریس $\mathbf{X}\boldsymbol{X}^{\boldsymbol{T}}$ به دلیل نوع قرار گرفتن ماتریس مسیر هر سری زمانی در کنار هم تحت تأثیر قرار می‌گیرد و لذا تفاوتی در تعداد مقادیر ویژه و بردارهای ویژه ظاهر می‌شود. اما شیوه‌ی محاسبه ماتریس‌های مقدماتی متعامد با حالت کلاسیک MSSA متفاوت است. بدین صورت که برای استوارسازی، با معرفی ماتریس قطری $\boldsymbol{W}$ به‌عنوان ماتریس وزن و ضریب قرار دادن این ماتریس در محاسبه‌ی ماتریس‌های مقدماتی متعامد، تجزیه‌ی مقادیر تکین به‌صورتی متفاوت خواهد بود. لذا در ابتدا علاوه بر توضیح کامل روش تجزیه‌ی مقادیر تکین برای روش تحلیل مجموعه مقادیر تکین کلاسیک در هر دو حالت افقی و عمودی، ماتریس قطری وزن $\boldsymbol{W}$ معرفی شده و پس از آن این مرحله برای روش L_1 − MSSA بیان خواهد شد.

برای حالت HMSSA و به‌منظور اجرای روش SVD بر روی ماتریس مسیر $\boldsymbol{X_H}$، سه‌تایی‌های ویژه $(\lambda_{H_i}, U_{H_i}, V_{H_i})$ از ماتریس $\boldsymbol{X_H X_H^T}$ به‌دست می‌آیند. فرض کنید $\lambda_{H_1}, \dots, \lambda_{H_L}$ مقادیر ویژه‌ی ماتریس $\boldsymbol{X_H X_H}^{\boldsymbol{T}}$ بوده که به‌صورت نزولی مرتب شده‌اند $(\lambda_{H_1} \geq \cdots \geq \lambda_{H_L} \geq 0)$ و $\boldsymbol{U_{H_1}, \dots, U_{H_L}}$ نیز بردارهای ویژه یکامتعامد مقادیر ویژه‌ی $\lambda_{H_1}, \dots, \lambda_{H_L}$ باشند.

$$\boldsymbol{X_H X_H^T} = [\boldsymbol{X^{(1)}} : \boldsymbol{X^{(2)}} : \ \dots : \ \boldsymbol{X^{(p)}}] \begin{bmatrix} \boldsymbol{X^{(1)}} \\ \dots \\ \boldsymbol{X^{(2)}} \\ \dots \\ \boldsymbol{X^{(p)}} \end{bmatrix} \quad (۵)$$

$$\boldsymbol{X_H X_H^T} = \boldsymbol{X^{(1)} X^{(1)^T}} + \boldsymbol{X^{(2)} X^{(2)^T}} + \cdots + \boldsymbol{X^{(p)} X^{(p)^T}}$$

در این‌صورت SVD ماتریس $\boldsymbol{X_H}$ به‌صورت زیر تعریف می‌شود:

$$\boldsymbol{X_H} = \boldsymbol{X_{H_1}} + \boldsymbol{X_{H_2}} + \cdots + \boldsymbol{X_{H_L}} \quad (۶)$$

وقتی که $\boldsymbol{X_{H_i}} = \sqrt{\boldsymbol{\lambda_{H_i}}}\boldsymbol{U_{H_i}}\boldsymbol{V_{H_i}^T}$ به‌عنوان ماتریس‌های مقدماتی متعامد و $\boldsymbol{V_{H_i}} = \boldsymbol{X_H^T U_{H_i}}/\sqrt{\lambda_{H_i}}$ به ازای $i = 1, \dots, L$. سه‌تایی $\left(\sqrt{\lambda_{H_i}}, \boldsymbol{U_{H_i}}, \boldsymbol{V_{H_i}}\right)$ را سه‌تایی ویژه و $\sqrt{\lambda_{H_i}}$ها را مقادیر تکین می‌نامند.

برای حالت VMSSA و به‌منظور اجرای SVD بر روی ماتریس مسیر $\boldsymbol{X_V}$، سه‌تایی ویژه $(\lambda_{V_i}, U_{V_i}, V_{V_i})$ را از ماتریس $\boldsymbol{X_V X_V^T}$ به‌دست می‌آید.

$$\boldsymbol{X_V X_V^T} = \begin{bmatrix} \boldsymbol{X^{(1)}} \\ \dots \\ \boldsymbol{X^{(2)}} \\ \dots \\ \boldsymbol{X^{(p)}} \end{bmatrix} [\boldsymbol{X^{(1)}} : \boldsymbol{X^{(2)}} : \ \dots : \ \boldsymbol{X^{(p)}}]$$

$$\boldsymbol{Q} = \boldsymbol{X_V X_V^T}$$

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{X_V}^{(1)}\boldsymbol{X_V^{(1)T}} & \boldsymbol{X_V}^{(1)}\boldsymbol{X_V^{(2)T}} & \dots & \boldsymbol{X_V}^{(1)}\boldsymbol{X_V^{(p)T}} \\ \boldsymbol{X_V}^{(2)}\boldsymbol{X_V^{(1)T}} & \boldsymbol{X_V}^{(2)}\boldsymbol{X_V^{(2)T}} & \dots & \boldsymbol{X_V}^{(2)}\boldsymbol{X_V^{(p)T}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{X_V}^{(p)}\boldsymbol{X_V^{(1)T}} & \boldsymbol{X_V}^{(p)}\boldsymbol{X_V^{(2)T}} & \dots & \boldsymbol{X_V}^{(p)}\boldsymbol{X_V^{(p)T}} \end{bmatrix}$$

لازم به توضیح است که SVD ماتریس مسیر $\boldsymbol{X_V}$ را می‌توان به‌صورت زیر نوشت.

$$\boldsymbol{X_V} = \boldsymbol{U_V \Sigma_V V_V^T}$$

که در آن $\boldsymbol{U_V} = \left[\boldsymbol{U_{V_1}}, \boldsymbol{U_{V_2}}, \dots, \boldsymbol{U_{V_{L_{sum}}}}\right] \in R^{L_{sum} \times L_{sum}}$،

$$\boldsymbol{V_V} = \left[\boldsymbol{V_{V_1}}, \boldsymbol{V_{V_2}}, \dots, \boldsymbol{V_{V_k}}\right] \in R^{\mathrm{K} \times \mathrm{K}}$$،

$$\boldsymbol{\Sigma} = diag\left(\sqrt{\lambda_{V_1}} \geq \sqrt{\lambda_{V_2}} \geq \cdots \geq \sqrt{\lambda_{V_{L_{sum}}}}\right)$$ و

$L_{sum} = \sum_{i=1}^{p} L_i$ مقادیر ویژه $\lambda_{V_1}, \lambda_{V_2}, \dots, \lambda_{V_{L_{sum}}}$ از ماتریس $\boldsymbol{Q}$ به دست می‌آید که $\lambda_{V_1} \geq \lambda_{V_2} \geq \cdots \geq \lambda_{V_{L_{sum}}}$ و همچنین $\boldsymbol{U_{V_1}}, \boldsymbol{U_{V_2}}, \dots, \boldsymbol{U_{V_{L_{sum}}}}$ بردارهای ویژه‌ی مرتبط با این مقادیر ویژه هستند.

در روش L_1 − MSSA تجزیه مقادیر تکین هر کدام از ماتریس مسیرها به ترتیب در حالت عمودی و افقی به‌صورت زیر به‌دست می‌آید.

$$\begin{aligned} \boldsymbol{X_V} &= \boldsymbol{X_{V_1}} + \boldsymbol{X_{V_2}} + \cdots + \boldsymbol{X_{V_{L_{sum}}}} = \sum_{i=1}^{d_V} \sqrt{\boldsymbol{\lambda_{V_i}}}\boldsymbol{U_{V_i}}\boldsymbol{W_i}\boldsymbol{V_{V_i}^T} \\ \boldsymbol{X_H} &= \boldsymbol{X_{H_1}} + \boldsymbol{X_{H_2}} + \cdots + \boldsymbol{X_{H_d}} = \sum_{i=1}^{d_H} \sqrt{\boldsymbol{\lambda_{H_i}}}\boldsymbol{U_{H_i}}\boldsymbol{W_i}\boldsymbol{V_{H_i}^T} \end{aligned} \quad (۷)$$

وقتی که $\boldsymbol{X_{V_i}} = \sqrt{\boldsymbol{\lambda_{V_i}}}\boldsymbol{U_{V_i}}\boldsymbol{W_i}\boldsymbol{V_{V_i}^T}$ و $\boldsymbol{X_{H_i}} = \sqrt{\boldsymbol{\lambda_{H_i}}}\boldsymbol{U_{H_i}}\boldsymbol{W_i}\boldsymbol{V_{H_i}^T}$ هستند. همچنین d_V و d_H به ترتیب برابر با تعداد مقادیر ویژه‌ی مثبت یا رتبه‌ی ماتریس مسیر در حالت عمودی و افقی است.

بنا بر خاصیت ساختاری روش مجموعه مقادیر تکین چندمتغیره، در حالت افقی یا حالت عمودی ماتریس سیگنال $\boldsymbol{Z} = \boldsymbol{U\Sigma_r V^T}$ در بین تمام ماتریس‌های سیگنال با رتبه‌ی $r < d$ کمترین فاصله را از ماتریس مسیر دارد یعنی $\|\boldsymbol{X} - \boldsymbol{U\Sigma_r V^T}\|_F$ حداقل مقدار را دارد. لذا از این ویژگی مهم می‌توان در محاسبه‌ی ماتریس سیگنال بر پایه‌ی نُرم L_1 استفاده کرد. برای جلوگیری از تکرار مطالب و نیز به دلیل تغییرات اساسی ابعاد و ساختار ماتریس مسیر در حالت عمودی، در این قسمت تنها به بررسی یک حالت از ماتریس سیگنال بر اساس نُرم L_1 پرداخته شده و ویژگی و ساختار ماتریس وزن در ساخت نزدیکترین ماتریس سیگنال به ماتریس مسیر حالت عمودی تشریح خواهد شد. حالت افقی شبیه

به حالت تک‌متغیره می‌باشد که علاقمندان می‌توانند برای اطلاعات بیشتر به [۲۱]، مراجعه نمایند.

فرض کنید ماتریس مسیر حالت عمودی $\boldsymbol{X_V} = \boldsymbol{X_{V_1}} + \boldsymbol{X_{V_2}} + \cdots + \boldsymbol{X_{V_{L_{sum}}}}$ دارای رتبه‌ی $d_V = r = rank(X) < L < K$ باشد. در این قسمت ماتریس قطری $\boldsymbol{W}$ به‌عنوان ماتریس وزن معرفی می‌شود.

$$\boldsymbol{W} = diag\ (\overbrace{w_1, w_2, \dots, w_r}^{r}, \overbrace{0,0,\dots,0}^{L_{sum}-r}) \in R^{L_{sum}\times L_{sum}} \quad (۸)$$

توجه داشته باشید که

$$\boldsymbol{w} = \begin{pmatrix} w_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ddots & w_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{W_r} & 0 \\ 0 & 0 \end{pmatrix},$$

که در آن $\boldsymbol{W_r} = diag\ (w_1, w_2, \dots, w_r) \in R^{r\times r}$.

برای محاسبه‌ی ماتریس سیگنال بر پایه نُرم L_1، لازم است ماتریس وزن قطری $\boldsymbol{W}_{L_{sum}\times L_{sum}}$ را به‌گونه‌ای یافت که $\left\|\boldsymbol{X_V} - \boldsymbol{U_V W \Sigma_d V_V^T}\right\|_{L_1}$ کمترین مقدار را داشته باشد. به طور دقیق‌تر، مسأله کمینه‌سازی زیر باید حل شود:

$$\min_{w}\left\|\boldsymbol{X_V} - \boldsymbol{U_V W \Sigma_d V_V^T}\right\|_{L_1}$$

توجه داشته باشید که

$$\begin{aligned}\boldsymbol{U_V W \Sigma_d V_V^T} &= [\boldsymbol{U_{V_1}} \quad \boldsymbol{U_{V_2}}]\begin{bmatrix}\boldsymbol{W_r} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma_1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{V_{V_1}^T}\\ \boldsymbol{V_{V_2}^T}\end{bmatrix} \\ &= [\boldsymbol{U_{V_1}} \boldsymbol{W_r} \quad \boldsymbol{0}]\begin{bmatrix}\boldsymbol{\Sigma_1 V_{V_1}^T}\\ \boldsymbol{0}\end{bmatrix} = \boldsymbol{U_{V_1} W_r \Sigma_1 V_{V_1}^T}\end{aligned} \quad (۹)$$

که در آن $\boldsymbol{U_{V_1}} \in R^{L_{sum}\times r}$، $\boldsymbol{V_{V_1}} \in R^{k\times r}$ و $\boldsymbol{\Sigma_1} \in R^{r\times r}$ به آسانی می‌توان نشان داد که

$$\boldsymbol{U_{V_1} W_r \Sigma_1 V_{V_1}^T} = \sum_{i=1}^{r}\sqrt{\lambda_{V_i}} w_i U_{V_i} V_{V_i}^T \quad (۱۰)$$

با استفاده از رابطه (۸) و (۹) مسئله کمینه‌سازی زیر حل می‌شود:

$$\min_{w}\left\|\boldsymbol{X_V} - \boldsymbol{U_{V_1} W_r \Sigma_1 V_{V_1}^T}\right\|_{L_1}$$

به همین منظور، فرض کنید $\boldsymbol{A_r} = \boldsymbol{U_{V_1} W_r}$ و $\boldsymbol{B_1} = \boldsymbol{\Sigma_1 V_{V_1}^T}$ که در آن $\boldsymbol{B_1} \in R^{k\times L_{sum}}$. بر اساس مشاهدات ماتریس‌های $\boldsymbol{B_1}$ و $\boldsymbol{U_{V_1}}$ معلوم هستند. هدف یافتن ماتریس $\boldsymbol{A_r}$ است به‌طوریکه $\|\boldsymbol{X_V} - \boldsymbol{A_r B_1}\|_{L_1}$ مینیمم شود؛ به عبارت دیگر، برای تقریب ماتریس سیگنال مسئله کمینه‌سازی (۱۱) حل می‌شود.

$$\min_{A_r}\|\boldsymbol{X_V} - \boldsymbol{A_r B_1}\|_{L_1} \quad (۱۱)$$

برای حل (۱۱) فرض کنید $\boldsymbol{S} = \boldsymbol{A_r B_1}$ باشد آن‌گاه $\boldsymbol{S^T} = \boldsymbol{B_1^T A_r^T}$ و

$$\boldsymbol{S_j^T} = \boldsymbol{B_1^T A_j^T} \quad (۱۲)$$

که در آن $\boldsymbol{A_j^T}$ و $\boldsymbol{S_j^T}$ به ترتیب jامین ستون ماتریس‌های $\boldsymbol{A_r^T}$ و $\boldsymbol{S^T}$ هستند. رابطه‌ی (۱۳)، با در نظر گرفتن خاصیت تعریف L_1 برای یک ماتریس و رابطه‌ی (۱۲) به صورت زیر خواهد بود.

$$\begin{aligned}\|\boldsymbol{X_V} - \boldsymbol{S}\|_{L_1} &= \|(\boldsymbol{X_V} - \boldsymbol{S})^T\|_{L_1} \\ &= \left\|\boldsymbol{X_V}^T - \boldsymbol{S^T}\right\|_{L_1} \\ &= \sum_{j=1}^{L_{sum}}\left\|\boldsymbol{X_{Vj}^T} - \boldsymbol{S_j^T}\right\|_{L_1} \\ &= \textstyle\sum_{j=1}^{L_{sum}}\left\|\boldsymbol{X_{Vj}^T} - \boldsymbol{B_1^T A_j^T}\right\|_{L_1}\end{aligned} \quad (۱۳)$$

که $\mathbf{X_{Vj}^T}$ به‌عنوان jامین ستون ماتریس $\mathbf{X_V}^{\mathbf{T}}$ در نظر گرفته شده است. رابطه (۱۳) نشان می‌دهد برای کمینه کردن عبارت $\|\boldsymbol{X_V} - \boldsymbol{S}\|_{L_1}$ کافی است عبارت $\left\|\boldsymbol{X_{Vj}^T} - \boldsymbol{B_1^T A_j^T}\right\|_{L_1}$ نسبت به $\boldsymbol{A_j^T}$ کمینه شود.

با توجه به این‌که مسئله‌ی فوق یک مسئله‌ی کمینه‌سازی شناخته‌شده در رگرسیون کمترین قدرمطلق انحرافات[20] (LAD) بوده و حل آن به‌صورت عددی وجود دارد؛ لذا با فرض این‌که $\boldsymbol{X_{Vj}^T}$، $\boldsymbol{B_1^T}$ و $\boldsymbol{A_j^T}$ را به ترتیب به‌عنوان بردار متغیر وابسته، ماتریس مشاهدات متغیرهای مستقل و بردار ضریب مجهول در نظر گرفته شوند، بردار $\boldsymbol{A_j^T}$ را می‌توان به کمک روش‌های عددی تکراری، مشابه آن‌چه که در [۲۲] و [۲۳] آمده است محاسبه کرد. بنا بر این با داشتن بردارهای $\boldsymbol{A_j^T}$، $j = 1,2,\dots,L_{sum}$، ماتریس $\boldsymbol{A_r}$ به‌دست می‌آید. برای انجام محاسبات مربوط به رگرسیون LAD از تابع rq در بسته‌ی quantreg نرم‌افزار R با آرگومان $\tau = ۰/۵$ استفاده شده است.

۲-۲- بازسازی

این مرحله شامل دو گام است: گروه‌بندی و میانگین‌گیری قطری.

الف) گروه‌بندی

گروه‌بندی یکی از مراحل مهم در تجزیه و بازسازی سری‌های زمانی به روش‌های SSA و MSSA است. در حال حاضر

[20] Least Absolute Deviation (LAD)

متداول‌ترین روش‌ها برای گروه‌بندی رسم نمودار مقادیر تکین ماتریس مسیر و استفاده از مفهوم تفکیک‌پذیری و محاسبه‌ی همبستگی بین زیرسری‌هاست. در خصوص نمودار مقادیر تکین، چنان‌چه λ_iها مقادیر نزدیک به هم داشته باشند در یک گروه قرار می‌گیرند.

گروه‌بندی برای ماتریس مسیر افقی

مرحله‌ی گروه‌بندی مربوط به افراز ماتریس‌های $\boldsymbol{X_{H_1}, X_{H_2}, \dots, X_{H_L}}$ به چندین گروه جدا از هم و جمع‌بندی ماتریس‌های درون هر گروه است. در مرحله گروه‌بندی با توجه به ماتریس مسیر حاصل‌شده، مجموعه‌ی $\{1, \dots, L\}$ به m زیرمجموعه‌ی $\{I_1, I_2, \dots, I_m\}$ جدا از هم افراز می‌شوند در این صورت رابطه‌ی $\boldsymbol{X_H = X_{H_1} + X_{H_2} + \cdots + X_{H_L}}$ را می‌توان در گام گروه‌بندی به‌صورت $\boldsymbol{X_H = X_{I_1} + X_{I_2} + \cdots + X_{I_m}}$ نوشت. شیوه‌ی انتخاب مجموعه‌های $I_1, I_2, \dots, I_m$ گروه‌بندی نامیده می‌شود. در واقع در این گام هدف یافتن سه‌تایی‌های ویژه مربوط به هر یک از اجزای سری زمانی نظیر روند، مؤلفه‌های فصلی، نوفه و غیره است. ساده‌ترین حالت گروه‌بندی برای زمانی است که تنها دو گروه وجود داشته باشند که این دو گروه می‌تواند شامل سیگنال و باقیمانده‌ها باشد $I_1 = \{1, 2, \dots, r\}$، در واقع r- تا از بزرگ‌ترین مقادیر تکین و بردارهای ویژه متناظر برای تقریب سری اصلی انتخاب شده و $I_2 = \{r+1, r+2, \dots, L\}$ بقیه مقادیر تکین به‌عنوان مؤلفه‌ی نوفه در نظر گرفته می‌شود. در متون علمی موجود از r به‌عنوان پارامتر بازسازی نیز اسم برده شده است. بنا بر این چنان‌چه در متن مقاله از پارامتر بازسازی نام برده شود منظور همان r (بزرگ‌ترین مقدارهای تکین و بردارهای ویژه متناظر برای تقریب سری اصلی انتخاب شده) است.

گروه‌بندی برای ماتریس مسیر عمودی

در مرحله گروه‌بندی با توجه به ماتریس مسیر حاصل‌شده، مجموعه‌ی $\{1, 2, \dots, L_{sum}\}$ به m زیرمجموعه‌ی $\{I_1, I_2, \dots, I_m\}$ جدا از هم افراز می‌شود. در این صورت رابطه‌ی $\boldsymbol{X_V = X_{V_1} + X_{V_2} + \cdots + X_{V_{L_{sum}}}}$ را می‌توان در گام گروه‌بندی به‌صورت $\boldsymbol{X_V = X_{I_1} + X_{I_2} + \cdots + X_{I_m}}$ در نظر گرفت. مانند حالت افقی شیوه‌ی انتخاب مجموعه‌های $I_1, \dots, I_m$، گروه‌بندی نامیده می‌شود. سایر توضیحات این قسمت همانند حالت افقی است با این تفاوت که ماتریس مسیر ترکیبی از سری‌های زمانی است که به‌صورت عمودی کنار هم قرار گرفته‌اند.

ب) میانگین‌گیری قطری

در روش MSSA، ماتریس مسیر ترکیبی از ماتریس‌هایی است که به‌صورت بلوک کنار هم قرار گرفته‌اند. از این‌رو در این مرحله میانگین‌گیری قطری برای هر بلوک به‌صورت جداگانه محاسبه می‌شود. در بخش میانگین‌گیری قطری ماتریس‌های بازسازی‌شده در مرحله‌ی گروه‌بندی خاصیت هنکلی ندارند و چنان‌چه ماتریس مسیر هنکل نباشد نمی‌توان به سری متناظر با آن ماتریس دست یافت. از این‌رو هدف اصلی در این مرحله تبدیل ماتریس بازسازی‌شده به یک ماتریس مسیر هنکل است. در واقع برای رسیدن به این هدف به جای عناصر روی قطر فرعی میانگین آنها قرار داده می‌شود. در آخر به ازای تعداد سری‌های زمانی‌ای (تعداد بلوک‌ها) که در مسئله وجود دارد، به همان تعداد هم ماتریس هنکل‌سازی شده در ماتریس مسیر وجود خواهد داشت. از این‌رو در حالت چندمتغیره ماتریس سیگنال، انباشتی از ماتریس‌های هنکل‌شده برای هر سری زمانی (هر بلوک) به تنهایی است. با این فرض که p تعداد بلوک‌ها یا تعداد متغیرهای سری زمانی باشد، در حالت کلی می‌توان ماتریس مسیر افقی ($\boldsymbol{X_H}$) و ماتریس مسیر عمودی ($\boldsymbol{X_V}$) به ترتیب با نمادهای $\mathcal{T}_{MSSA}(\mathbb{X}_H)$ و $\mathcal{T}_{MSSA}(\mathbb{X}_V)$ نمایش داده و آن‌ها را به‌صورت رابطه‌ی (۱۴) نوشت.

$$\mathcal{T}_{MSSA}(\mathbb{X}_H) = \boldsymbol{X_H} = [\boldsymbol{X^{(1)}} : \cdots : \boldsymbol{X^{(p)}}]$$
$$\mathcal{T}_{MSSA}(\mathbb{X}_V) = \boldsymbol{X_V} = \begin{bmatrix} \boldsymbol{X^{(1)}} \\ \boldsymbol{X^{(2)}} \\ \vdots \\ \boldsymbol{X^{(p)}} \end{bmatrix} \qquad (۱۴)$$

بنا بر این می‌توان ماتریس مسیر هنکل‌سازی شده را به‌صورت زیر در نظر گرفت، به طوری‌که برای حالت افقی $[\![\boldsymbol{X^{(1)}} : \cdots : \boldsymbol{X^{(p)}}]\!]$ و برای حالت عمودی $\begin{bmatrix}\!\begin{bmatrix} \boldsymbol{X^{(1)}} \\ \boldsymbol{X^{(2)}} \\ \vdots \\ \boldsymbol{X^{(p)}} \end{bmatrix}\!\end{bmatrix}$ به‌ترتیب بلوک‌های تشکیل‌دهنده ماتریس مسیر بوده و

$$\mathcal{T}^{-1}_{MSSA}(\boldsymbol{X_H}) = [\mathcal{T}^{-1}_{SSA}(\boldsymbol{X^{(1)}}) : \cdots : \mathcal{T}^{-1}_{SSA}(\boldsymbol{X^{(p)}})]$$
$$\mathcal{T}^{-1}_{MSSA}(\boldsymbol{X_V}) = \begin{bmatrix} \mathcal{T}^{-1}_{SSA}(\boldsymbol{X^{(1)}}) \\ \mathcal{T}^{-1}_{SSA}(\boldsymbol{X^{(2)}}) \\ \vdots \\ \mathcal{T}^{-1}_{SSA}(\boldsymbol{X^{(p)}}) \end{bmatrix} \qquad (۱۵)$$

نماد $\mathcal{T}^{-1}$ هنکل شدن یک ماتریس مسیر است.

در ابتدای این بحث، با توضیح مرحله‌ی میانگین‌گیری قطری که در روش MSSA کلاسیک به آن اشاره شد به جای عناصر روی قطر فرعی، میانگین آنها قرار داده می‌شود. از طرفی دیگر، یکی از روش‌های هنکل‌سازی استفاده از ملاک کمترین فاصله بر اساس نُرم فروبنیوس است. بر اساس این ملاک، ماتریس هنکل‌سازی شده‌ی $\boldsymbol{A}$ که با نماد $\boldsymbol{\mathcal{H}_A}$ نمایش داده می‌شود طوری تعیین می‌شود که $\|\boldsymbol{A}-\boldsymbol{\mathcal{H}}A\|_F$ کمترین مقدار را داشته باشد، اما در روش $\mathrm{L}_1-\mathrm{MSSA}$، هنکل‌سازی بر اساس نُرم L_1 ارائه می‌شود. همانگونه که در حالت یک متغیره ثابت شده است [۲۱]، در حالت چندمتغیره نیز با استفاده از قضیه زیر می‌توان ثابت کرد که برای هنکل‌سازی ماتریس $\mathbf{A}$ بر اساس نُرم L_1 کافی است به جای درایه‌های قطر فرعی در هر بلوک، مقدار میانه‌ی درایه‌های آن قطر را جایگزین درایه‌های موجود در آن قطر کرد.

قضیه:

فرض کنید $\mathbf{A_I}$ یک ماتریس $L_I \times \mathrm{K}$ (بلوک Iام از ماتریس $\boldsymbol{X_V}$) بوده $S_I = i_I + j_I$ و $2 \le S_I \le L_I \times \mathrm{K}$، $(I = 1,2,\dots,p)$.

اگر درایه‌های $\tilde{a}_{i_I j_I}$ ماتریس $\boldsymbol{\mathcal{H}A_I}$ به‌صورت عبارت زیر باشد،

$$\tilde{a}_{i_I j_I} = \underset{(l,k)\in A_S}{median}\, a_{l_I k_I}$$

آن‌گاه وقتی که

$$(\boldsymbol{A_S})_{\boldsymbol{I}} = \{(L_I, \mathrm{K}): L_I \times \mathrm{K} = S_I \quad ,1 \le l_I \le L_I \, , 1 \le k_I \le K_I\}$$

$\|\boldsymbol{A_I} - \boldsymbol{\mathcal{H}A_I}\|_{L_1}$ کمترین مقدار را خواهد داشت.

طبق تعریف درایه‌های $\left(b_{ij}\right)_I$ ماتریس هنکل $\boldsymbol{B} = \boldsymbol{\mathcal{H}A_I}$ باید به ازای $S_I = i_I + j_I$ و برخی مقادیر $(g_s)_I$ در شرط $\left(b_{ij}\right)_I = (g_s)_I$ صدق کند، از طرفی با در نظرگرفتن تعریف نرم L_1 یک ماتریس داریم $(I = 1,2,\dots,p)$؛

$$\|\boldsymbol{A_I} - \boldsymbol{\mathcal{H}A_I}\|_{L_1} = \|\boldsymbol{A_I} - \boldsymbol{B}\|_{L_1}$$

$$\sum_{i_I=1}^{L_I}\sum_{j_I=1}^{K_I}\left|(a_{ij})_I - (b_{ij})_I\right| = \sum_{S_I=1}^{L_I+K_I}\sum_{S_I=i_I+j_I}\left|(a_{ij})_I - (g_s)_I\right|$$

بنابر نامساوی میانه، مقادیر $(g_s)_I$ که سمت راست رابطه‌ی فوق را کمینه می‌سازند برابرند با میانه‌ی اعضای مجموعه‌ی $(\boldsymbol{A_s})_{\boldsymbol{I}}$.

علاوه بر این باید دقت کرد که اندیس I بیانگر تعداد متغیرهای (سری‌های زمانی) به کار برده شده در مدل است و باید هنکل‌سازی بر اساس نرم L_1 برای همه‌ی p متغیر صورت پذیرد تا به این صورت تمام بلوک‌ها با کمترین اختلاف با ماتریس مسیر بازسازی‌شده هنکل‌سازی شوند. همچنین چون اختلاف با هر بلوک با ماتریس مسیر هنکل‌سازی شده کمینه مقدار است بنا بر این ماتریس مسیر تولید شده در مرحله‌ی آخر نزدیک‌ترین داده‌ها را به سری زمانی تجزیه و بازسازی شده تولید خواهد کرد و داده‌های هر سری زمانی قابل دستیابی است. قضیه‌ی فوق بیان می‌کند که برای هنکل‌سازی ماتریس $\mathbf{A}$ بر اساس نُرم L_1 کافی است به جای درایه‌های قطر فرعی در هر بلوک مقدار میانه‌ی درایه‌های آن قطر را جایگزین درایه‌های موجود در آن قطر کرد.

۲-۳- پیش‌بینی بازگشتی برای سری‌های زمانی به روش $\mathbf{L_1-MSSA}$

برای پیش‌بینی سری‌های زمانی به روش تحلیل مجموعه مقادیر تکین چندمتغیره دو شیوه‌ی پیش‌بینی به روش‌های بازگشتی و برداری معرفی شده است که می‌توان برای مطالعه‌ی بیشتر به [۲۰] و [۲۴] مراجعه نمود. در این مقاله و در بخش‌های شبیه‌سازی برای HMSSA و VMSSA پیش‌بینی به روش بازگشتی خطی[21] (LRF) به کار برده خواهد شد و ضرایب رابطه‌ی خطی بازگشتی از بردارهای ویژه‌ی ماتریس مسیرها به دست می‌آیند.

سطر آخر r ستون اول، از ماتریس بردارهای ویژه که در مرحله SVD روی ماتریس $\boldsymbol{X_H X_H^T}$ به‌دست آمده بودند را بر $1-L$ سطر اول رگرسیون کرده و بردار ضرایب را به شکل $\boldsymbol{\mathcal{R}} = [\widehat{\boldsymbol{a}_1}, \widehat{\boldsymbol{a}_2}, \dots, \widehat{\boldsymbol{a}_{L-1}}]$ تعریف می‌شوند. با فرض آن‌که $\boldsymbol{U_{H_j}^{\nabla}}$ نشان‌دهنده‌ی $L-1$ مؤلفه‌ی اول بردار $\boldsymbol{U_{H_j}}$ و نماد Π_{H_j} آخرین مؤلفه بردار ویژه‌ی $\boldsymbol{U_{H_j}}$، $(j = 1,2,\dots,\mathrm{r})$ باشد، می‌توان بردار ضرایب خطی $\boldsymbol{\mathcal{R}}$ را به‌صورت رابطه‌ی (۱۶) تعریف کرد که در آن $\mathcal{V}^2 = \sum_{j=1}^{r}\Pi_{H_j}^2$:

$$\boldsymbol{\mathcal{R}} = [\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a_{L-1}}] = \frac{1}{1-\mathcal{V}^2}\sum_{j=1}^{r}\boldsymbol{\Pi_{H_j}}\boldsymbol{U_{H_j}^{\nabla}} \qquad (16)$$

اگر $\mathcal{V}^2 < 1$، آن‌گاه hگام جلوتر (پیش‌بینی hگام جلوتر) از سری زمانی چندمتغیره به روش HMSSA با استفاده از رابطه (۱۷) به‌دست می‌آید.

$$\left[\hat{y}_{j_1}^{(1)}, \hat{y}_{j_2}^{(2)}, \dots, \hat{y}_{j_p}^{(p)}\right]^T = \begin{cases}\left[\tilde{y}_{j_1}^{(1)}, \tilde{y}_{j_2}^{(2)}, \dots, \tilde{y}_{j_p}^{(p)}\right] & j = 1,2,\dots,N_i \\ \boldsymbol{\mathcal{R}^T Z_h} & j = N_i+1, \dots, N_i+\mathrm{h}\end{cases} \qquad (17)$$

در حالی که؛

[21] Linear Recurrent Formula

$$\boldsymbol{Z_h} = \left[\boldsymbol{Z_h^{(1)}},\ \boldsymbol{Z_h^{(2)}}, \dots, \boldsymbol{Z_h^{(p)}}\right]^T \text{ و}$$
$$\boldsymbol{Z_h^{(i)}} = \left[\hat{y}_{N_i-L_i+h+1}^{(i)}, \dots, \hat{y}_{N_i+\mathrm{h}-1}^{(i)}\right] \quad i = 1,2,\dots,p$$

برای حالت عمودی و پیش‌بینی hگام جلوتر از سری زمانی چندمتغیره به روش VMSSA نیز می‌توان با رگرسیون کردن سطر آخر r ستون اول، از هر کدام از ماتریس بردارهای ویژه بر $L-1$ سطرهای قبل از سطر آخر، که در واقع به‌صورت بلوک برای هر سری زمانی در مرحله SVD روی ماتریس $\boldsymbol{X_V X_V^T}$ حاصل می‌شود، به دست آورد. در این حالت، $\mathbf{U_j^{(i)\nabla}}$ به‌عنوان $\mathrm{L}-1$ مولفه اول بردار $\mathbf{U_j^{(i)}}$ و نماد $\Pi_j^{(i)}$ به‌عنوان آخرین مؤلفه بردار $\mathbf{U_j^{(i)}}$ در نظر گرفته شده و زمانی که $\mathbf{U_j^{p\nabla}} = \begin{bmatrix} U_j^{(1)\nabla} \\ U_j^{(2)\nabla} \\ \vdots \\ U_j^{(\mathrm{p})\nabla} \end{bmatrix}$ باشد، ماتریس $\mathbf{U^{p\nabla}}$ به صورت $\mathbf{U^{p\nabla}} = \left(\mathbf{U_1^{p\nabla}}, \mathbf{U_2^{p\nabla}}, \dots, \mathbf{U_r^{p\nabla}}\right)$ تعریف می‌شود.

همچنین ماتریس $\boldsymbol{\Psi}$ به‌صورت $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\pi}_1^{(1)} & \boldsymbol{\pi}_2^{(1)} & \dots & \boldsymbol{\pi}_{\mathbf{r}}^{(1)} \\ \boldsymbol{\pi}_1^{(2)} & \boldsymbol{\pi}_1^{(2)} & \cdots & \boldsymbol{\pi}_{\mathbf{r}}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\pi}_1^{(\mathrm{p})} & \boldsymbol{\pi}_1^{(\mathrm{p})} & \dots & \boldsymbol{\pi}_{\mathbf{r}}^{(p)} \end{bmatrix}$ تعریف می‌شود.

اگر ماتریس $\left(\boldsymbol{I_{p\times p}} - \boldsymbol{\Psi\Psi^T}\right)^{-1}$ وجود داشته باشد و $\mathrm{r} \le L_{sum} - p$ آن‌گاه h مرحله جلوتر VMSSA که همان پیش‌بینی hگام جلوتر است بر اساس رابطه (۱۸) به‌دست می‌آید.

$$\left[\hat{y}_{j_1}^{(1)}, \hat{y}_{j_2}^{(2)}, \dots, \hat{y}_{j_p}^{(p)}\right]^T = \begin{cases} \left[\tilde{y}_{j_1}^{(1)}, \tilde{y}_{j_2}^{(2)}, \dots, \tilde{y}_{j_p}^{(p)}\right] & j = 1,2,\dots,N_i \\ \left(\boldsymbol{I_{p\times p}} - \boldsymbol{\Psi\Psi^T}\right)^{-1}\boldsymbol{\Psi U^{(\mathrm{p})\nabla T} Z_h} & j = N_i+1,\dots,N_i+\mathrm{h} \end{cases} \quad (18)$$

در حالی که

$$\boldsymbol{Z_h} = \left[\boldsymbol{Z_h^{(1)}},\ \boldsymbol{Z_h^{(2)}}, \dots, \boldsymbol{Z_h^{(p)}}\right]^T \text{ و}$$
$$\boldsymbol{Z_h^{(i)}} = \left[\hat{y}_{N_i-L_i+h+1}^{(i)}, \dots, \hat{y}_{N_i+\mathrm{h}-1}^{(i)}\right] \quad i = 1,2,\dots,p$$

۳- کاربرد

در این بخش کارایی روش‌های کلاسیک MSSA و $\mathrm{L}-1$ MSSA با استفاده از سری‌های زمانی شبیه‌سازی شده و داده‌های واقعی مورد مقایسه قرار می‌گیرند. به همین منظور مجموعه‌ی $\{1, 2, \dots, L_{sum}\}$ به دو زیرمجموعه داده تقسیم می‌شود. برای مقایسه‌ی عملکرد هر دو روش، در مرحله‌ی بازسازی و پیش‌بینی از دو معیار ریشه‌ی میانگین توان دوم خطا[22] ($RMSE$) و میانگین قدرمطلق خطا[23] (MAE) و همچنین از نسبت معیارهای نام‌برده‌شده که به‌صورت رابطه (۱۹) تعریف می‌شوند استفاده شد.

$$RRMSE = \frac{L_1 - MSSA \text{ با استفاده از } RMSE}{MSSA \text{ با استفاده از روش کلاسیک } RMSE}$$
$$RMAE = \frac{L_1 - MSSA \text{ با استفاده از } MAE}{MSSA \text{ با استفاده از روش کلاسیک } MAE} \quad (19)$$

اگر ($RMAE < 1$ یا $RRMSE < 1$)، آن‌گاه می‌توان گفت روش $L_1 - MSSA$ نسبت به روش کلاسیک $MSSA$ از کارایی بیشتری برخوردار است.

۳-۱-مطالعات شبیه‌سازی

در مطالعات شبیه‌سازی، ۲۰۰ داده با استفاده از مدل‌های مختلف ساختاری تولید شده است. از ۱۹۰ مشاهده‌ی اول برای بازسازی و ۱۰ مشاهده‌ی بعدی برای پیش‌بینی استفاده شده است. در تمام روش‌های مطرح شده نکته‌ی قابل توجه تعیین پارامترهای طول پنجره و بازسازی است که مقدار آنها در گروه‌بندی و در نتیجه بازسازی سری‌های زمانی اهمیت زیادی داشته و انتخاب آنها به ساختار داده‌ها و نیز هدف تجزیه و تحلیل بستگی دارد. شبیه‌سازی برای هریک از مدل‌ها در ۱۰۰۰ تکرار انجام شده و دو معیار RRMSE و RMAE محاسبه می‌شوند. نتایج شبیه‌سازی نشان می‌دهد که دقت روش MSSA به مقدار زیادی به این پارامترها بستگی دارد. انتخاب نامناسب پارامترهای طول پنجره و بازسازی، باعث تغییر ساختار ماتریس مسیر و تغییر در نحوه‌ی گروه‌بندی‌ها شده و نتایج غیر قابل اتکایی پدید خواهد آمد [۲۵].

[22] Root Mean Square Error (RMSE)
[23] Mean Absolute Error (MAE)

هدف از محاسبه‌ی پارامتر بازسازی در گروه‌بندی جدا کردن نوفه و سیگنال است [۸]، که در واقع همان r- تای اول یا r- تا از بزرگ‌ترین مقدارهای تکین و بردارهای ویژه متناظر برای تقریب سری اصلی انتخاب شـده هستند. برای تعیین پارامتر طول پنجره مقدارهای نزدیک به نصف طول سری زمانی می‌تواند انتخاب مناسبی برای پارامتر طول پنجره باشد. چنان‌چه سری زمانی دارای دوره‌ی تناوب یا مؤلفه‌ی متناوب به‌صورت یک عدد صحیح باشد طول پنجره متناسب با این دوره انتخاب می‌شود [۲۱]. همچنین در روش MSSA برای حالت HMSSA و VMSSA به‌منظور دستیابی به بهترین مقدار برای پارامتر طول پنجره به ترتیب از روابط $L=\left[\frac{p(N+1)}{p+1}\right]$ و $L=\left[\frac{(N+1)}{p+1}\right]$ استفاده می‌شود [۱۳] و [۲۶]. در این روابط N طول سری زمانی و p تعداد متغیرهای موجود در سری زمانی است.

در مطالعات شبیه‌سازی شده در این مقاله طول پنجره متناسب با دوره‌ی تناوب سری زمانی و به‌صورت مضربی از آن در نظر گرفته شده است، بنا بر این برای هر ۳ مثال بیان شده مقدار طول پنجره ضرایبی از ۱۲ در نظر گرفته شده‌اند $L=$ ۱۲, ۲۴, ۳۶, ۴۸, ۶۰, ۷۲, ۸۴, ۹۶.

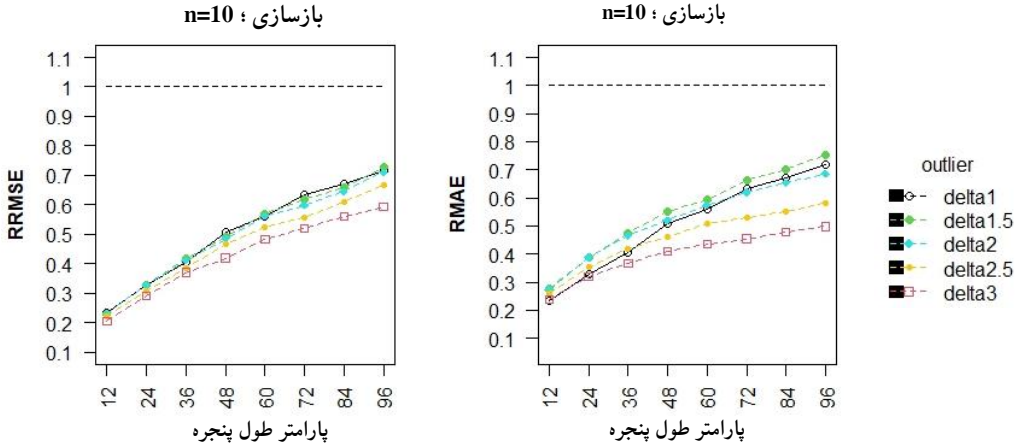
از طرفی با توجه به هدف مورد بررسی در مقاله یعنی مقایسه‌ی روش L_1-MSSA با روش کلاسیک $MSSA$ در برابر داده‌های دورافتاده از حضور داده‌های دورافتاده در سری‌های زمانی استفاده شده است. برای تولید داده‌های دورافتاده، ابتدا به تعداد n مشاهده (n=5,10) از سری زمانی شبیه‌سازی شده به طور تصادفی انتخاب کرده و سپس به جای آنها، δ برابر آنها را جایگزین شد. در تمامی مثال‌هایی که ارائه خواهند شد مقدار ضرایب δ برابر با $\{$۱, ۱٫۵ , ۲, ۲٫۵ و ۳$\}$ است. لازم به توضیح است در تمامی شکل‌ها هر نمودار به‌طور مشخص برای حجم نمونه‌های جایگزین $n=5$ یا $n=10$ رسم شده است. ضمن این‌که به هر دو سری زمانی خطا (ε) با توزیع نرمال استاندارد به‌عنوان عامل نوفه اضافه شده است. نمودار هر خط (روند) نشان‌دهنده‌ی نسبت ریشه‌ی میانگین توان دوم خطاها برای هرکدام از ۲ روش L_1-MSSA و روش کلاسیک $MSSA$ است که برای طول پنجره‌های متفاوت با پارامترهای بازسازی اختصاص داده شده به آنها با یک حجم خاص و مقدار δ تعیین شده است.

مثال اول: سری‌های زیر به‌عنوان یک سری زمانی دو متغیره برای شبیه‌سازی در نظر گرفته شده‌اند.

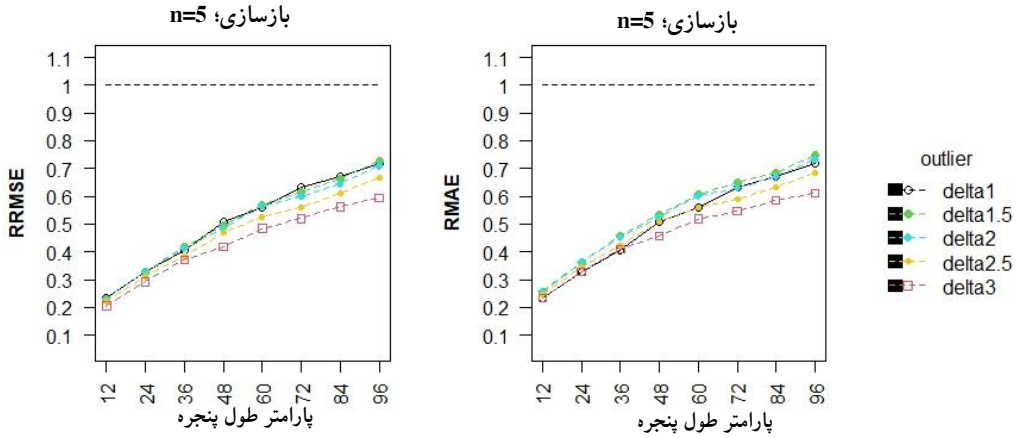
$$y_1=3\sin\left(\frac{2t\pi}{12}\right)+\varepsilon$$

$$y_2=2\sin\left(\frac{2t\pi}{12}+\frac{\pi}{4}\right)+\varepsilon$$

با توجه به این‌که مرتبه‌ی این سری زمانی برای حالت افقی ۲ در نظر گرفته شده [۴]، بنا بر این از ۲ سه‌تایی ویژه اول برای بازسازی و پیش‌بینی این سری زمانی دومتغیره در حالت افقی استفاده شده است. همچنین با توجه به راهنمایی‌های کلی در مورد نحوه‌ی انتخاب پارامتر طول پنجره که در قسمت قبل ارائه شده و نیز با توجه به این‌که پارامتر طول پنجره ضرایبی از دوره‌ی تناوب سری زمانی است، طول پنجره به‌صورت ضرایبی از دوره‌ی تناوب این سری زمانی که ۱۲ است در نظر گرفته شده است.



شکل ۱- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۱۰



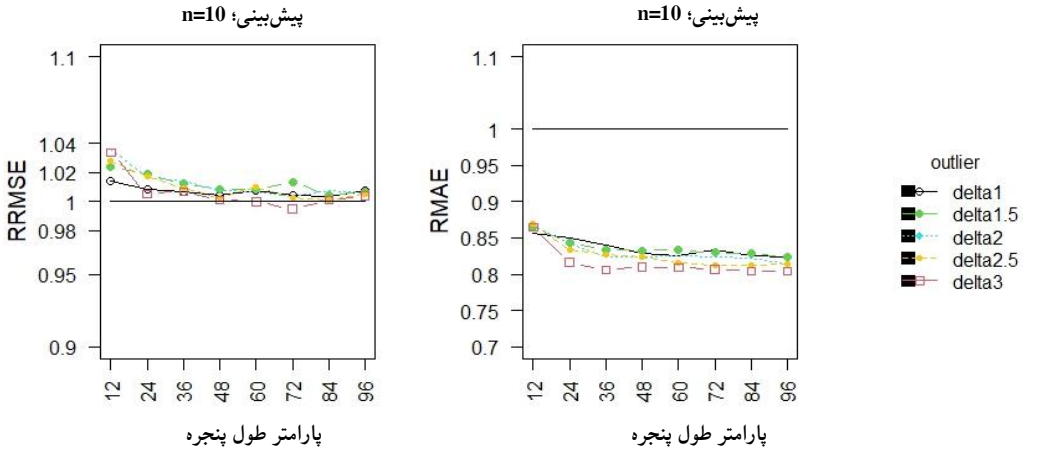
شکل ۲- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۵

شکل‌های ۱ و ۲ نمودارهایی هستند که برای مقایسه بازسازی سری‌های زمانی به دو روش $L_1 - MSSA$ و روش کلاسیک $MSSA$ رسم شده‌اند. در شکل ۱ به ازای $\delta = \{1, 1/5, 2, 2/5 \text{ و } 3\}$ انتخاب یک نمونه ۱۰- تایی برای ساختن داده‌های دورافتاده و در شکل ۲ به ازای همین ضرایب یک نمونه ۵تایی برای ساختن داده‌های دورافتاده در نظر گرفته شده‌اند. بنا بر این همان‌طور که در این شکل‌ها مشاهده می‌شود برای هر دو نمونه‌ی انتخابی نسبت خطاها یعنی RRMSE و RMAE کمتر از یک شده و این بدان معناست که با حضور داده‌های دور افتاده در سری زمانی عملکرد روش $L_1 - MSSA$ در بازسازی سری‌های زمانی بهتر از روش کلاسیک است.

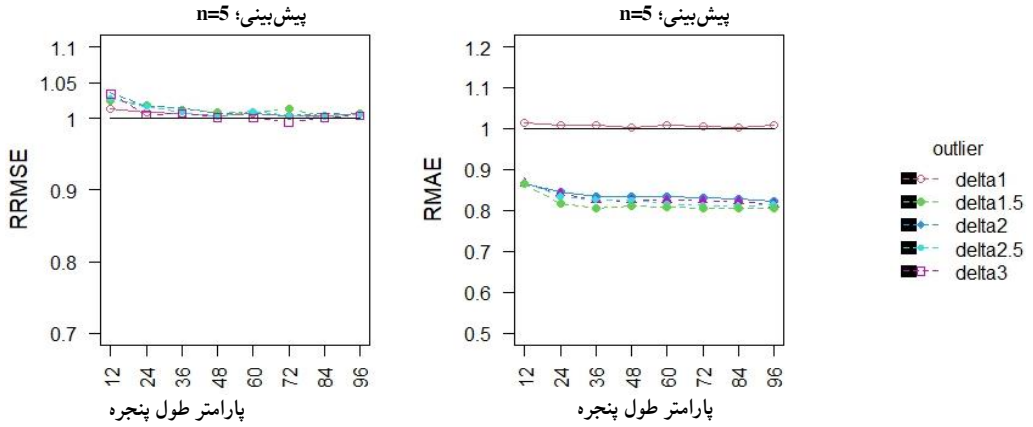
شکل‌های ۳ و ۴ نمودارهایی هستند که برای مقایسه پیش‌بینی سری‌های زمانی به دو روش کلاسیک $MSSA$ و $L_1 - MSSA$ ارائه شده‌اند. همانند قسمت بازسازی در این شکل‌ها به سری‌های زمانی داده‌های دورافتاده با ضرایب $\delta = \{1, 1/5, 2, 2/5 \text{ و } 3\}$ به‌صورت تصادفی به ترتیب با حجم نمونه ۵ و ۱۰- تایی اضافه شده است.

درهر دو حالتی که حجم نمونه‌ی انتخابی ۱۰یا ۵ بوده، با افزایش ضریب داده‌های دور افتاده، $RRMSE < 1$ شده و عملکرد روش $L_1 - MSSA$ بهتر از روش کلاسیک مجموعه مقادیر تکین چندمتغیره است.

اما برای نمودار RMAE با انتخاب نمونه ۱۰- تایی در تمام ضرایب داده‌های دورافتاده، عملکرد روش $L_1 - MSSA$ بهتر از روش کلاسیک مجموعه مقادیر تکین چندمتغیره است. بر اساس همین معیار در پیش‌بینی، زمانی که حجم نمونه ۵- تایی باشد، در حالی که مشاهده‌ی داده‌ی دورافتاده‌ای وجود نداشته باشد عملکرد هر دو روش یکسان است.



شکل ۳- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۱۰



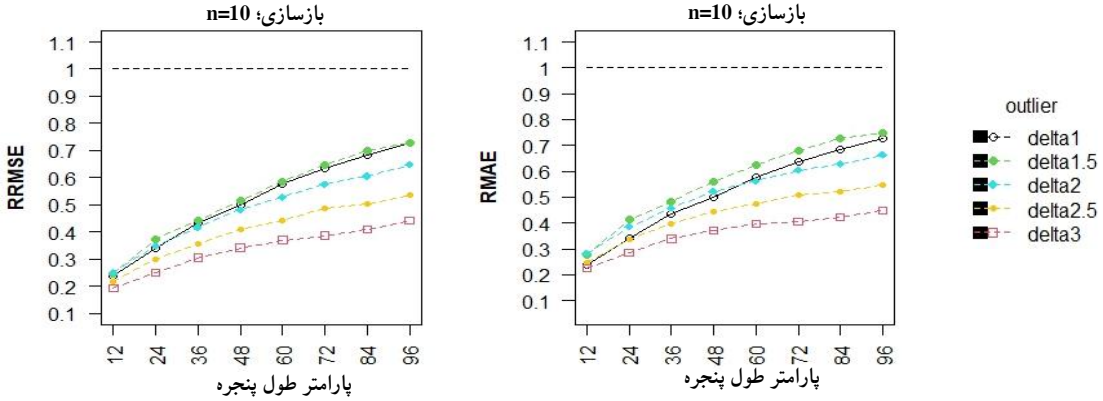
شکل ٤- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۵

مثال دوم: سری‌های زیر به‌عنوان یک سری زمانی دو متغیره برای شبیه‌سازی در نظر گرفته شده‌اند:

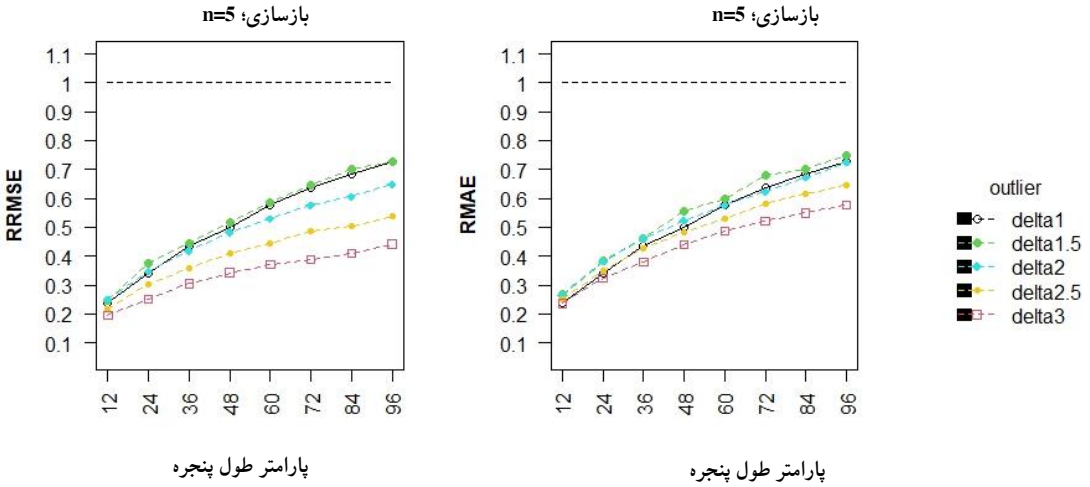
$$y_{1} = 3\,cos\left(\frac{2t\pi}{12}\right) + \varepsilon$$

$$y_{2} = 3\cos\left(\frac{2t\pi}{12} + \frac{\pi}{2}\right) + \varepsilon$$

رتبه‌ی ماتریس مسیر این سری‌های زمانی برای حالت افقی ۲ در نظر گرفته شده است [۴]. بنا بر این از ۲، سه‌تایی ویژه اول برای بازسازی و پیش‌بینی این سری زمانی دومتغیره در حالت افقی استفاده شده است. همچنین طول پنجره را به‌عنوان ضرایبی از دوره‌ی تناوب سری زمانی که ۱۲ است در نظر گرفته شده است.



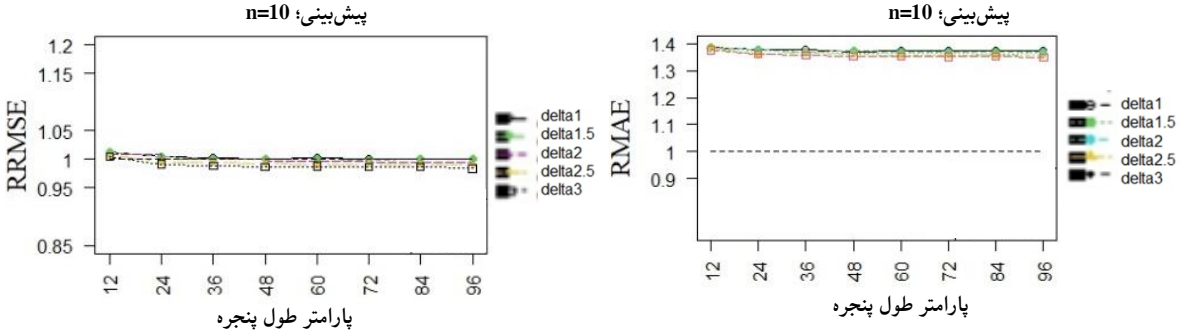
شکل ۵- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۱۰



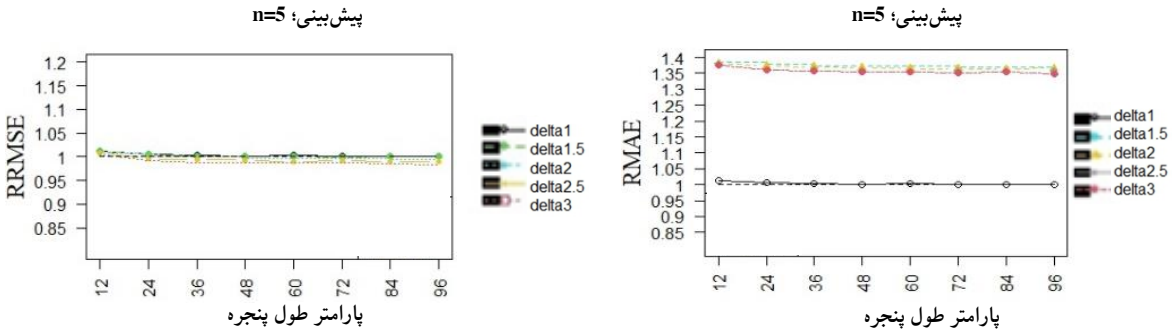
شکل ٦- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۵

شکل‌های ۵ و ۶ نمودارهایی هستند که برای مقایسه بازسازی سری‌های زمانی به دو روش L_1 – MSSA و کلاسیک $MSSA$ رسم شده‌اند. برای هر دو حالت حجم نمونه ۵ و ۱۰- تایی و تمام ضرایب تولید داده‌های دورافتاده، نسبت خطاها یعنی RRMSE و RMAE کمتر از یک شده و این بدان معناست که با حضور داده‌های دور افتاده در سری زمانی عملکرد روش L_1 – MSSA در بازسازی سری‌های زمانی بهتر از حالت روش کلاسیک $MSSA$ است.

نمودارهای RRMSE و RMAE در شکل۷ برای مقایسه پیش‌بینی سری‌های زمانی به دو روش کلاسیک $MSSA$ و L_1 – MSSA مورد استفاده قرار گرفته‌اند. در نمودار RRMSE نسبت خطای بین دو روش کمتر از یک شده و در نتیجه عملکرد روش L_1 – MSSA بهتر از عملکرد روش کلاسیک مجموعه مقادیر تکین چندمتغیره است. اما برای RMAE با انتخاب نمونه ۱۰- تایی در تمام ضرایب داده‌های دورافتاده، عملکرد روش کلاسیک $MSSA$ بهتر از روش L_1 – MSSA است.



شکل ۷- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۱۰



شکل ۸- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۵

شکل ۸ که شامل نمودارهای RRMSE و RMAE برای پیش‌بینی سری‌های زمانی دو متغیره‌ی مثال دوم است، نشان می‌دهد چنان‌چه اندازه نمونه‌ی انتخابی برای اضافه کردن داده‌های دورافتاده ۵ باشد با افزایش ضریب داده‌های دورافتاده، با توجه به معیار RRMSE عملکرد روش $L_1 - MSSA$ بهتر از عملکرد روش کلاسیک $MSSA$ است.

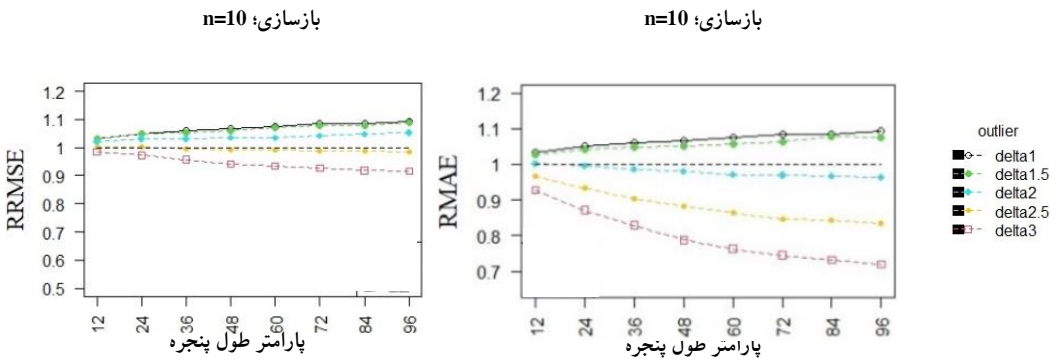
مثال سوم: سری‌های زیر به‌عنوان یک سری زمانی دو متغیره برای شبیه‌سازی در نظر گرفته شده‌اند.

$$y_{1} = 3\,cos\left(\frac{2t\pi}{12}\right) + \varepsilon$$
$$y_{2} = 2\cos\left(\frac{2t\pi}{12} + \frac{\pi}{4}\right) + \varepsilon$$

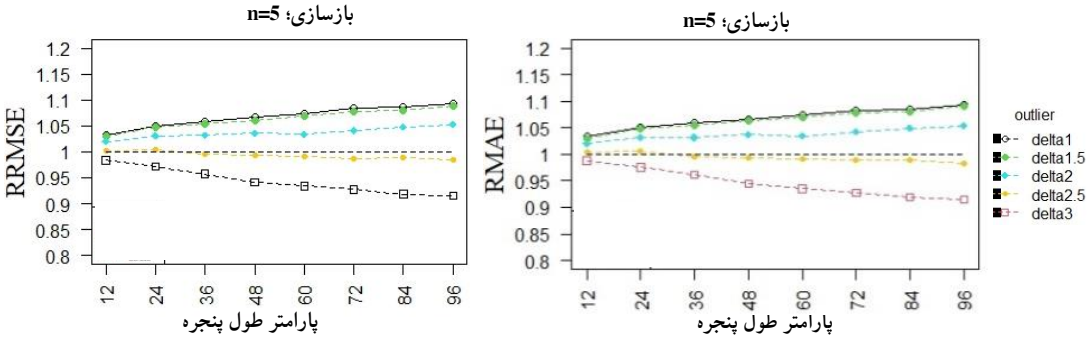
رتبه‌ی ماتریس مسیر این سری زمانی برای حالت افقی ۲ در نظر گرفته شده است [۴]. بنا بر این از ۲، سه‌تایی ویژه اول برای بازسازی و پیش‌بینی این سری زمانی دومتغیره در حالت افقی استفاده شده است.

همچنین طول پنجره به‌عنوان ضرایبی از دوره‌ی تناوب سری زمانی که ۱۲ است در نظر گرفته شده است.

در شکل ۹ نمودارهای RRMSE و RMAE در حالت بازسازی برای مثال سوم نتایجی متفاوت‌تر با خروجی دو مثال دیگر دارند. در این مثال نمودار RRMSE برای مقایسه‌ی عملکرد $L_1 - MSSA$ و روش کلاسیک $MSSA$ در حالت بازسازی با حضور ۱۰ داده‌ی دورافتاده با ضرایب $\{1.5, 2, 2.5$ و $3\}$ را نشان می‌دهد. با توجه به این نمودار متوجه می‌شویم چنان‌چه ضرایب تولید داده‌های دورافتاده بزرگتر شوند عملکرد روش $L_1 - MSSA$ بهتر از عملکرد روش کلاسیک $MSSA$ خواهد بود. بنا بر این تنها در دو حالت که ضرایب تولید داده دورافتاده ۲ و ۳ باشد، $RRMSE < 1$ خواهد بود. نمودار RMAE تفسیری شبیه نمودار RRMSE دارد و عملکرد $L_1 - MSSA$ زمانی که داده‌های دورافتاده بزرگتری در مشاهدات باشد در مقایسه با روش کلاسیک $MSSA$ بهتر خواهد بود و $RMAE < 1$ است.

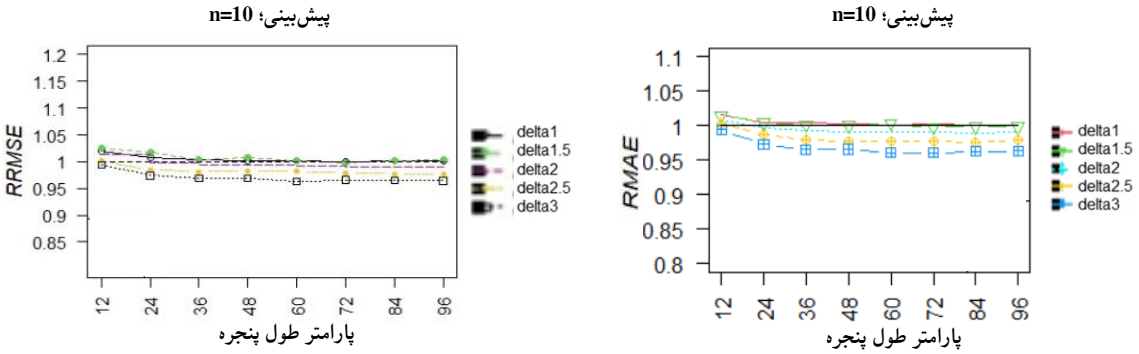


شکل ۹- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۱۰

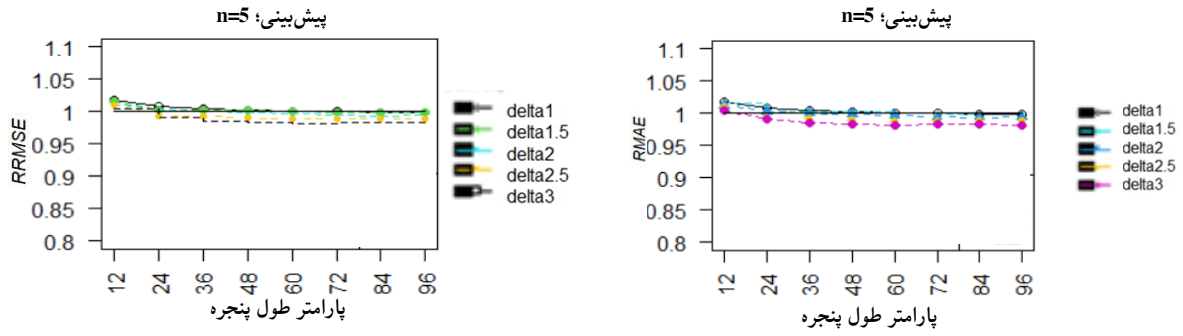


شکل ۱۰- نمودارهای RRMSE و RMAE برای بازسازی با اندازه نمونه ۵

شکل ۱۰ نمودارهای RRMSE و RMAE عملکرد دو روش L_1-MSSA و کلاسیک $MSSA$ را در مرحله‌ی بازسازی نشان می‌دهد. با توجه به این شکل زمانی که ضریب داده‌های دورافتاده $\delta=\{2/5,3\}$ در هر دو مورد مقدارهای RRMSE و RMAE افزایش می‌یابند نسبت خطاها کاهش یافته تا جایی که برای کمتر از یک شده و در نتیجه عملکرد L_1-MSSA در مقایسه با روش کلاسیک $MSSA$ در مرحله‌ی بازسازی بهتر است.



شکل ۱۱- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۱۰



شکل ۱۲- نمودارهای RRMSE و RMAE برای پیش‌بینی با اندازه نمونه ۵

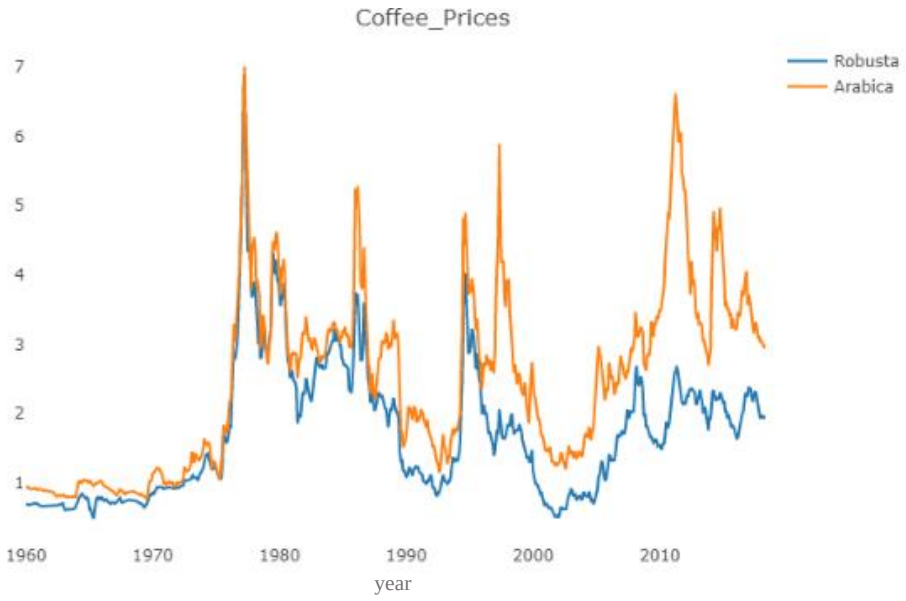
شکل‌های ۱۱ و ۱۲ به ترتیب نمودارهای RRMSE و RMAE برای حضور ۱۰ و ۵ داده‌ی دورافتاده در مشاهدات سری‌های زمانی چندمتغیره، این نمودارها بهتر بودن عملکرد L_1-MSSA در مرحله‌ی پیش‌بینی را نسبت به عملکرد روش کلاسیک $MSSA$ نشان می‌دهند. همچنین از این نمودارها می‌توان به این نتیجه دست یافت که در مرحله‌ی پیش‌بینی

چنان‌چه تعداد داده‌های دورافتاده بیشتر و بزرگتر باشند عملکرد L_1 − MSSA بهتر از عملکرد روش کلاسیک $MSSA$ است.
با مشاهده‌ی همه‌ی نمودارهای RRMSE و RMAE در هر دو حالت نمونه‌های ۵ و ۱۰تایی، این‌گونه به نظر می‌رسد که انتخاب مناسب طول پنجره در رسیدن به نتیجه مطلوب تأثیرگذار خواهد بود.

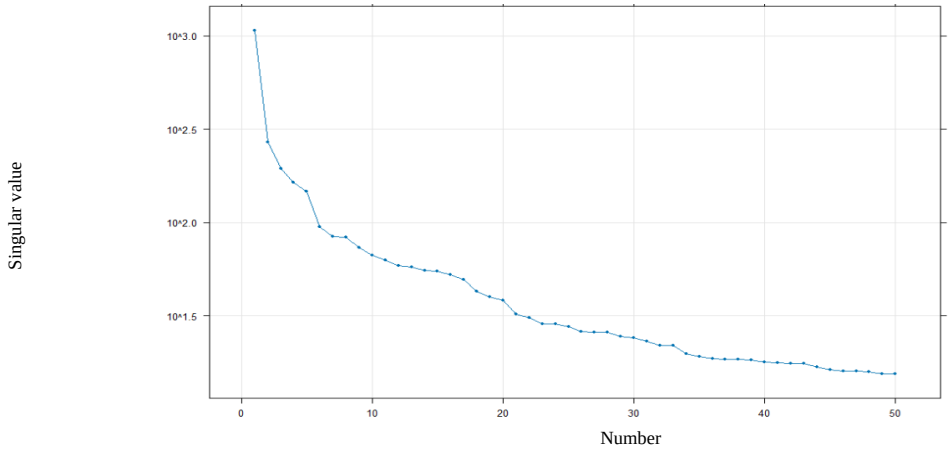
۳-۲- داده‌های واقعی

در این بخش برای بررسی عملکرد کارایی روش تحلیل مجموعه مقادیر تکین چندمتغیره‌ی استوار در حالت افقی با روش کلاسیک MSSA در حالت افقی از داده‌های قیمت قهوه که یک سری زمانی دو متغیره با ۷۰۱ مشاهده به‌صورت ماهانه از January سال ۱۹۶۰ تا May سال ۲۰۱۸می‌باشد، استفاده شده است. این داده‌ها در نرم‌افزار R و در بسته‌ی TSstudio با نام Coffee_Prices قابل دسترس هستند. دلیل استفاده از این داده‌ها اولاً داشتن نقاط دورافتاده در نمودار سری زمانی و دوماً دومتغیره بودن آنها می‌باشد (شکل ۱۳). از ۶۶۶ مشاهده‌ی اول این سری زمانی برای بازسازی و ۳۵ مشاهده‌ی بعدی برای پیش‌بینی استفاده شده است.
با توجه به این‌که برای انجام محاسبات لازم، ماتریس مسیر را به‌صورت افقی در نظر گرفته‌ایم لذا با استفاده از رفتار مقادیر ویژه ماتریس مسیر برای طول پنجره ۴۶۸ که در شکل ۱۴ نمایش داده شده و با توجه به مطالب و دلایل ارائه شده [۹] در هر دو روش ۶ مقدار اول مقادیر ویژه برای بازسازی و پیش‌بینی در نظر گرفته می‌شود.
با بررسی قدم به قدم نمودارهای «بردارهای ویژه»، «بردارهای ویژه جفتی» و «w-همبستگی» و وجود دوره‌ی تناوب ۳ و ۴ برای داده‌ها، پارامتر طول پنجره را مقداری بین ۳۶ (ضریبی از دوره تناوب داده‌ها) به‌عنوان کمترین و ۴۶۸[۲۴] به‌عنوان بیشترین طول پنجره‌ای که می‌توان انتخاب کرد، در نظر گرفته شده است. از ۳۶ تا ۴۶۸ در هر مرحله ضریبی از ۳۶ را به مقدار طول پنجره اضافه کرده و در محاسبات، به دست آوردن پارامتر بازسازی در نظر گرفته شد. با توجه به این مطلب که برای تعیین پارامتر بازسازی یا نحوه گروه‌بندی داده‌های واقعی روش ریاضی و فرموله‌شده خاصی وجود ندارد بنا بر این به ازای هر طول پنجره، نمودارهای «بردارهای ویژه»، «بردارهای ویژه جفتی» و «w-همبستگی» مجدداً رسم شده و بر اساس اطلاعات حاصل شده از نمودارها و تجربه محقق، پارامتر بازسازی متناسب با هر طول پنجره تعیین شد. با تعیین پارامتر طول پنجره و پارامتر بازسازی، دو گام گروه‌بندی و میانگین‌گیری از مرحله‌ی بازسازی داده‌ها انجام می‌شود. به همین خاطر متناسب با طول‌های پنجره متفاوت، مقادیر پارامتر بازسازی نیز تعیین شده و خروجی‌های برنامه متناسب با آنها به دست آمد، سپس بر اساس معیارهای RMAE RRMSE، عملکرد دو روش مورد مقایسه قرار گرفت. نتیجه این مقایسه‌ها نیز به‌صورت نمودارهای ۱۵ برای بازسازی و ۱۶ برای پیش‌بینی رسم شده است.

[۲۴] (بر اساس فرمول $L = \frac{p}{p+1}(N+1)$)



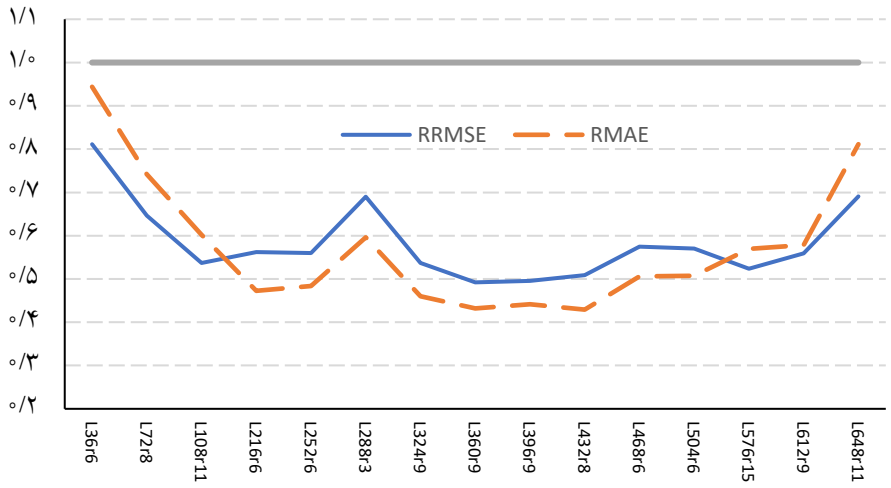
شکل ۱۳- سری‌های زمانی فروش قهوه از ماه اول سال ۱۹۶۰ تا ماه پنجم سال ۲۰۱۸



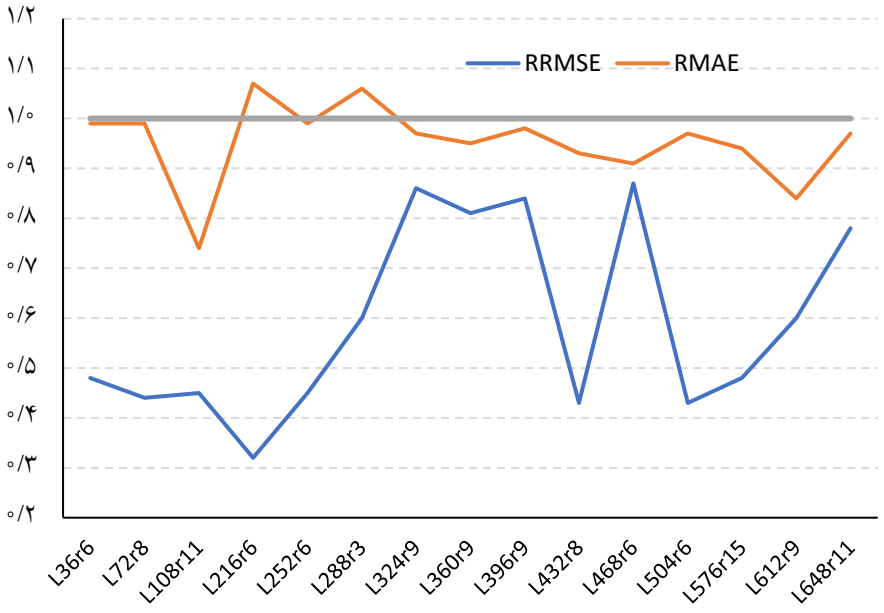
شکل ۱۴- نمودار رفتار مقادیر ویژه ماتریس مسیر برای طول پنجره ۴۶۸

شکل ۱۵ نشان می‌دهد که تقریباً با افزایش طول پنجره و پارامتر بازسازی، کارایی روش $L_1 - MSSA$ بهتر از روش کلاسیک $MSSA$ است و مقدار $RRMSE < 1$ و $RMAE < 1$ است.

شکل ۱۶ که مقایسه عملکرد دو روش $L_1 - MSSA$ و روش کلاسیک $MSSA$ را در مرحله‌ی پیش‌بینی نشان می‌دهد گویای این واقعیت می‌باشد که تقریباً با افزایش طول پنجره و پارامتر بازسازی، کارایی روش $L_1 - MSSA$ بهتر از روش کلاسیک $MSSA$ است و مقدار $RRMSE < 1$ و $RMAE < 1$ است.

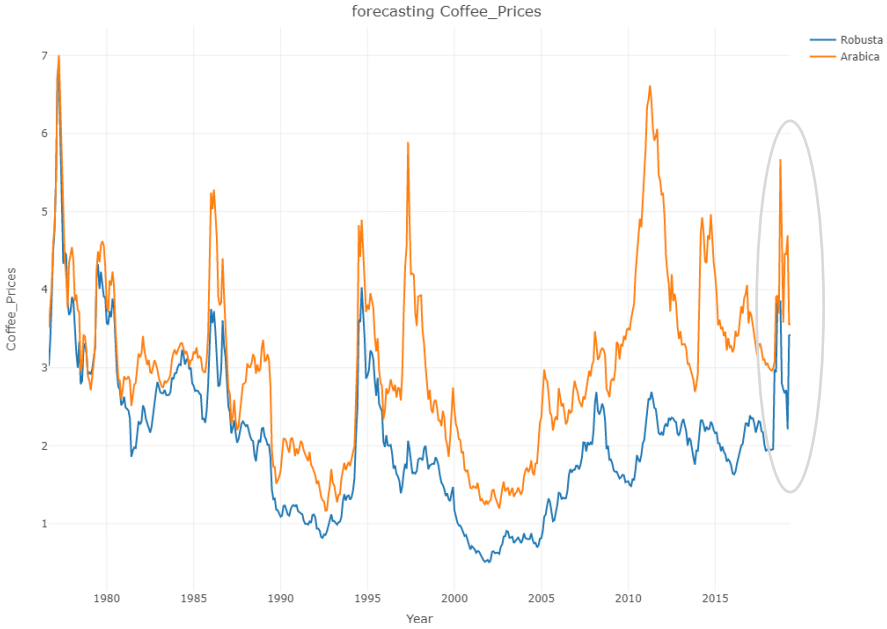


شکل ۱۵- نمودارهای RRMSE و RMAE برای مقایسه عملکرد دو روش $MSSA - L_1$ و کلاسیک $MSSA$ در بازسازی داده‌های واقعی



شکل ۱۶- نمودارهای RRMSE و RMAE برای مقایسه عملکرد روش $MSSA - L_1$ و روش کلاسیک $MSSA$ در پیش‌بینی داده‌های واقعی

در شکل ۱۷، پیش‌بینی یک‌ساله برای سری‌های زمانی فروش قهوه با استفاده از روش بازگشتی خطی (LRF) و در نظر گرفتن پارامتر طول پنجره‌ی ۴۶۸ و پارامتر بازسازی ۶ انجام شده است. در این نمودار مقادیر پیش‌بینی با کادری که به دور آنها کشیده شده مشخص شده است.



شکل ۱۷–نمودار پیش‌بینی سری‌های زمانی فروش قهوه از ماه ششم سال ۲۰۱۸ تا ماه پنجم سال ۲۰۱۹

۴–نتیجه‌گیری

در این مقاله به‌منظور بهینه‌سازی کیفیت مدل‌سازی و پیش‌بینی به دست آمده از روش کلاسیک MSSA با حضور داده‌های دورافتاده، روش استوارسازی بر اساس تغییر نُرم به کار رفته در ماتریس مسیر پیشنهاد شد. این روش که یک نسخه‌ی جدید از روش کلاسیک $MSSA$ بر اساس نُرم L_1، و تغییر ساختار ماتریس مسیر در مرحله‌ی بازسازی است به عنوان شیوه‌ای نوین در استوارسازی سری‌های زمانی برای بهبود مدل‌سازی شناخته می‌شود. لذا این روش علاوه بر بهینه‌سازی کیفیت مدل پیشنهادی، باعث افزایش کارایی تحلیل و پیش‌بینی سری‌های زمانی نیز می‌شود.

برخی محققان برای افزایش قابلیت‌های بازسازی و پیش‌بینی در روش SSA، شناسایی و حذف نقاط دورافتاده را پیشنهاد داده‌اند زیرا SSA مبتنی بر نُرم L_2 نسبت به نقاط دورافتاده حساس‌تر است[۱۹]. در حالی که نتایج مقاله حاضر نشان می‌دهد که در صورت مواجهه با سری‌های زمانی دارای نقاط دورافتاده، می‌توان با تغییر نُرم ماتریسی در $MSSA$ از L_2 به L_1 به نتایج بهتری دست پیدا کرد زیرا نیازی به تبدیل داده وجود نداشته و تضمین می‌کند که هیچ اطلاعاتی از دست نرود.

برای ارزیابی عملکرد نسخه‌ی جدید پیشنهادی ($L_1 - MSSA$) با روش کلاسیک MSSA از معیارهای RRMSE و RMAE در مطالعات شبیه‌سازی و داده‌های واقعی استفاده شد. در این ارزیابی با اضافه کردن داده‌های دورافتاده با حجم و ضرایب متفاوت در مرحله‌ی شبیه‌سازی عملکرد دو روش بیان شده در مرحله‌ی بازسازی و پیش‌بینی مورد مقایسه گرفتند. براساس مشاهدات حاصل از نتایج تحلیل شبیه‌سازی و داده‌های واقعی نشان داده شد که اگر داده‌های دورافتاده در سری زمانی وجود داشته باشند، تقریباً در تمام موارد کارایی روش ناپارامتری مجموعه مقادیر تکین استوار شده چند متغیره $L_1 - MSSA$ بهتر از حالت کلاسیک این روش است.

۵–منابع

[1] Hassani, H., Kalantari, M., and Yarmohammadi, M. (2017). An improved SSA forecasting result based on a filtered recurrent forecasting algorithm. Comptes Rendus. Mathématique, 355(9), 1026-1036.

[2] Broomhead, D. S., and King, G. P. (1986 b). On the Qualitative Analysis of Experimental Dynamical Systems. *Nonlinear Phenomena and Chaos*, 113, 114.

[3] Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., ... and Yiou, P. (2002). Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1), 3-1.
[4] Golyandina, N. E. and Stepanov, D. (2005). SSA-based approaches to analysis and forecast of multidimensional time series. Proceedings of the 5th St.Petersburg Workshop on Simulation, 293–298.
[5] Wang, R., Ma, H.-G., Liu, G.-Q. and Zuo, D.-G. (2015). Selection of window length for singular spectrum analysis, Journal of the Franklin Institute, 352, 1541–1560.
[6] Alharbia, N. and Hassani, H. (2016). A new approach for selecting the number of the eigenvalues in singular spectrum analysis, Journal of the Franklin Institute, 353, 1-16.
[7] Hassani, H. and Rua, A. and Silva, E.S. and Thomakos, D. (2019). Monthly forecasting of GDP with mixed-frequency multivariate singular spectrum analysis. International Journal of Forecasting, 35 (4). pp. 1263-1272.
[8] Patterson, K., Hassani, H., Heravi, S., and Zhigljavsky, A. (2011). Forecasting the final vintage of the industrial production series, Journal of Applied Statistics, 38, 2183–2211.
[9] Hassani, H. and Mahmoudvand. R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach, Int. J. Energy Stat. 1(1) (2013) 55–83.
[10] de Carvalho, M., Rodrigues, P.C. and Rua, A. (2012). Tracking the US business cycle with a singular spectrum analysis, Economics Letters, 114, 32–35.
[11] Rodrigues, P.C. and de Carvalho, M. (2013). Spectral modeling of time series with missing data, Applied Mathematical Modelling, 37, 4676–4684.
[12] Hasani, H., Heravi, S., and Zhigljavsky, A. (2009). Forecasting European industrial production with singular spectrum analysis. International journal of forecasting, 25(1), 103-118.
[13] Mahmoudvand, R. and Rodrigues, P.C. (2016). Missing value imputation in time series using Singular Spectrum Analysis. International *Journal of Energy and Statistics*, 4(01), 1650005.
[14] Rodrigues, P.C., and Mahmoudvand, R. (2016). Correlation analysis in contaminated data by singular spectrum analysis, Quality and Reliability Engineering International, 32, 2127–2137.
[15] Mahmoudvand, R.; Alehosseini, F.; and Rodrigues, P. C. (2015). Mortality Forecasting with Singular Spectrum Analysis. RevStat - Statistical Journal, 13, 193– 206.
[16] Mahmoudvand, R., Konstantinides, D. and Rodrigues, P.C. (2017). Forecasting Mortality Rate by Multivariate Singular Spectrum Analysis. Applied Stochastic Models in Business and Industry. DOI: 10.1002/asmb.2274.
[17] Hassani, H., Soofi, A., and Avazalipour, M. S. (2011). Forecasting GDP with aggregated and sectoral data. Fluctuation and Noise Letters, 10 (3), 249-265.
[18] Sirimal Silva, E., Hassani, H. and Heravi, S. (2018). Modeling European industrial production with multivariate singular spectrum analysis: A cross-industry analysis, Journal of Forecasting, https://doi.org/10.1002/for.2508
[19] Mahmoudvand, R. (2012). Some Theoretical Development of the Singular Spectrum Analysis. Ph.D. Thesis (In Persian), Shahid Beheshti University, Tehran, Iran.
[20] Kalantari, M., Yarmohammadi, M. and Hassani, H. (2016). Singular Spectrum Analysis Based on L_1-norm. Fluctuation and Noise Letters, 15(01), 1650009.
[21] Bloomfield, P. and W. Steiger (1983). Least Absolute Deviations: Theory, Applications and Algorithms. Boston: Birkhäuser.
[22] Birkes, D., & Dodge, Y. (1993). Alternative Methods of Regression. John Wiley&Sons. Inc., New York.
[23] Golyandina, N. and Zhigljavsky, A. (2018). Singular Spectrum Analysis for Time Series Second Edition.,Springer Briefs in Statistics.
[24] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001). Analysis of Time Series Structure: SSA and Related Techniques. London: Chapman & Hall/CRC.
[25] Rodrigues, P. C., & Mahmoudvand, R. (2018). The benefits of multivariate singular spectrum analysis over the univariate version. Journal of the Franklin Institute, 355(1), 544-564.

Improving The Quality of Time Series Modeling and Forecasting Using Robust Multivariate Singular Spectrum Analysis

Tahere Amini
Ph.D. Student in Statistics, Payame Noor University, Tehran, Iran.
tahereamini90@yahoo.com

Masoud Yarmohammadi[1]
(Corresponding Author): Associate Professor of Statistics, Payame Noor University, Tehran, Iran. masyar@pnu.ac.ir

Ali Shadrokh
Associate Professor of Statistics, Payame Noor University, Tehran, Iran.
a.shadrokh@pnu.ac.ir

Mahdi Kalantari
Assistant Professor of Statistics, Payame Noor University, Tehran, Iran.
kalantarimahdi@pnu.ac.ir

Abstract

In time series analysis, ignoring outliers leads to misidentification of the model, biased estimation of parameters, and poor predictions. One of the reliable non-parametric methods in predicting and improving the quality of multivariate time series modeling is the multivariate Singular Spectrum Analysis (MSSA) technique, which does not require any initial assumptions. The presence of outliers affects the Frobenius norm of matrix and reduces the efficiency of the MSSA method. In this research, a new version of MSSA based on the L_1- norm is proposed. Then the performance of this method is compared with basic MSSA using simulation studies and real data.

Key Words: Multivariate Singular Spectrum Analysis (MSSA), Robustification, Outliers, L_1-norm, Ratio of Root Mean Square Errors (RRMSE), Ratio of Mean Absolute Errors (RMAE).

1- Introduction

Multivariate Singular Spectrum Analysis (MSSA) is a nonparametric method to analyze signals without any assumptions of the underlying system. It can decompose the original time series into a set of components, which are recognized as either a trend, periodic or quasi-periodic signal or noise. The matrix norm used in MSSA is the Frobenius norm which is not robust to the outliers. In order to robustify the method, the Frobenius norm is replaced by the L_1 -norm and a new version of MSSA (L_1-MSSA) is proposed. In this approach; a method for computing the signal matrix and the hankelization method using the L_1- norm are proposed. Using real-world data and the simulation studies the

[1] Corresponding Author: masyar@pnu.ac.ir

performance of L_1-MSSA and basic MSSA are compared. Appling the Root Mean Square Errors (RMSE) and Mean Absolut Error (MAE) criteria, show that the L_1-MSSA outperforms basic MSSA in reconstruction and forecasting.

2- Methodology

The Multivariate Singular Spectrum Analysis (MSSA) technique consists of two complementary stages: Decomposition and Reconstruction, and both of them include two separate steps. In the first stage, we decompose the series to enable signal extraction and noise reduction. In the second stage, we reconstruct a less noisy series and use the reconstructed series for forecasting new data points [1]. In the MSSA method there are two forms for the trajectory matrix: vertical and horizontal. The vertical and the horizontal forms are called VMSSA and HMSSA, respectively. The matrix norm used in basic MSSA is the Frobenius norm which is not robust in the presence of outliers. In this paper, in order to robustify the method, the Frobenius norm is replaced by the L_1 norm and a new version of MSSA (L_1-MSSA) is proposed. In L_1-MSSA, the negative effects of outliers is reduced by utilizing the median instead of averaging for trajectory matrix hankelization in the last step of time series reconstruction (averaging) and the performance of L_1-MSSA and basic MSSA in reconstruction and forecasting are evaluated using the real-world and simulated time series that are contaminated with outliers. To measure the accuracy of forecasting results, we use the commonly adopted forecasting performance evaluation measures of Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE). Moreover, the effects of outliers on forecasting and reconstruction for two procedures are evaluated. The following ratios are used for comparing the established and newly proposed MSSA methods:

$$RRMSE = \frac{RMSE\ based\ on\ \ L_1 - MSSA}{RMSE\ based\ on\ \ basic\ MSSA}$$

$$RMAE\ = \frac{MAE\ based\ on\ \ L_1 - MSSA}{MAE\ based\ on\ \ basic\ MSSA}$$

If $RRMSE < 1$ and $RMAE < 1$, then the L_1-MSSA procedure outperforms basic MSSA. Alternatively, when $RRMSE > 1\ and\ RMAE > 1$, it would indicate that the performance of the $L_1 - MSSA$ procedure is worse than basic MSSA.

The purpose of determining the reconstruction parameter in the grouping step of MSSA is to separate noise and signal [6]. This parameter is equal to the first r singular values and the corresponding eigenvectors are chosen to approximate the original series. Selection of the proper window length (denoted by L) depends on the structure of time series and the preliminary information of the data. Theoretical results indicate that L should be large enough but not greater than $\left[\frac{N}{2}\right]$. Furthermore, if we know that the time series may have a periodic component with an integer period then to get better separability of this periodic component it is advisable to take the window length proportional to that period [3]. Regarding the window length, Hassani and Mahmoudvand (2013) showed that a value close to the $L = \left[\frac{p(N+1)}{p+1}\right]\ and\ L = \left[\frac{(N+1)}{p+1}\right]$

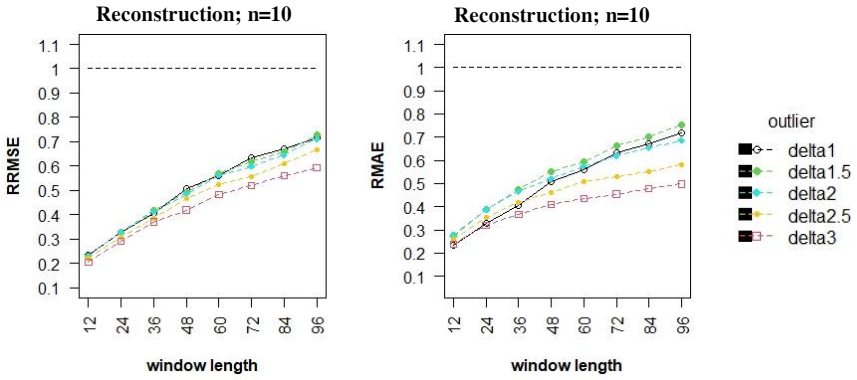
is optimal for HMSSA and VMSSA, respectively. In these formulas, N is the length of the time series and p is the number of variables in the time series [4, 5].

3- Results

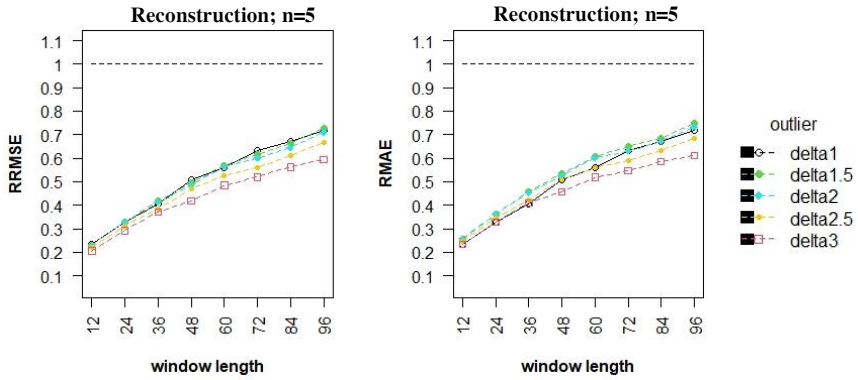
In this section, the results of the paper are presented. To generate outliers, first, n observations (n=5,10) are randomly selected from the simulated time series and then a multiplication (δ) of them is substituted for each selected observation. In all examples, the values of δ are equal to 1, 1.5, 2, 2.5, and 3. In this section, we present an example to declare the purposed method. As an example, we consider the following Sine series:

$$y_1 = 3\sin\left(\frac{2t\pi}{12}\right) + \varepsilon$$
$$y_2 = 2\sin\left(\frac{2t\pi}{12} + \frac{\pi}{4}\right) + \varepsilon \qquad t=1, 2, \ldots, 200,$$

where $\varepsilon \sim N(0,1)$ is the noise component. The first 190 observations were considered as in-sample (reconstruction) and the rest as out-of-sample series (forecasting). Since the rank of the trajectory matrix for this model (HMSSA) is equal to 2 we choose r =2. Using the recommendations given in Sec.2, we take L =12, 24, 36, 48, 60, 72, 84, and 96. In figures 1 and 2, the plots of RRMSE and RMAE for reconstruction are depicted for $n = 5\ and\ 10$. The plots of RRMSE and RMAE show that L_1-MSSA is better than basic MSSA for all values of L and δ (RRMSE<1 and RMAE<1).



Figs. 1- The plots RRMSE and RMAE for reconstruction and sample size 10



Figs. 2– The plots RRMSE and RMAE for reconstruction and sample size 5

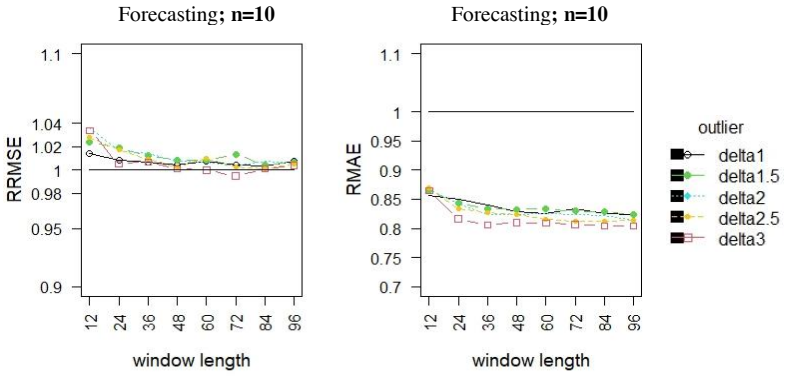


Fig. 3– The plots RRMSE and RMAE for forecasting and sample size 10

In figures 3 and 4, the plots of RRMSE and RMAE for forecasting out of samples are depicted for $n = 5, 10$. The plot of RRMSE shows that L_1–MSSA is better than basic MSSA for $L = 72\ and\ \delta = 3$.

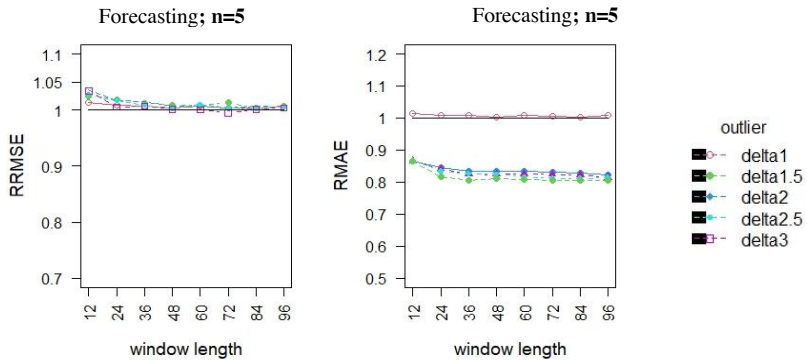


Fig.4– The plots RRMSE and RMAE for forecasting and sample size 5

For $n = 5\ and\ 10$, according to RRMSE, by increasing δ, the efficiency of $L_1 - MSSA$ tends to increase. But based on RMAE, the performance of the L_1-MSSA method is better than the basic MSSA for all values of n, L, and δ. When $\delta = 1$ (i.e. no outliers), the performance of both methods (L_1-MSSA and basic MSSA) are the same in forecasting.

4- Discussion and Conclusion

In this paper, to optimize the quality of modeling and forecasting obtained from the basic MSSA with outliers, a robust method was proposed by replacing Frobenius norm with the $L_1 - norm$. This method is a new way of using robust time series to improve modeling. In addition to optimizing the quality of the proposed model, this technique increases the efficiency of time series analysis and forecasting. In order to increase the reconstruction and forecasting performance of the SSA method, identifying and removing outliers have been suggested by some researchers, because SSA based on the L_2 norm is more sensitive to outliers [2]. The results of this paper show that in the case of encountering time series with outliers, better results can be obtained by changing the matrix norm in MSSA from L_2 to L_1 because there is no need to transform the data and it ensures that no information is lost. This study considered both RMSE and MAE criteria. In brief, the comparison of forecasting results showed that L_1-MSSA is more accurate than the basic MSSA version, further confirming the results obtained via theoretical results.

5- Reference

[1] Hassani, H., Webster, A. E., Silva, S. and Heravi, S. (2015). Forecasting U.S. tourist arrivals using optimal singular spectrum analysis, *Tour. Manage*. 46, 322–335.
[2] Hassani, H. and Mahmoudvand. R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach, *Int. J. Energy Stat.* 1(1), 55–83.
[3] Hassani, H. (2007). Singular spectrum analysis: Methodology and comparison, J. Data Sci. 5(2), 239–257.
[4] Mahmoudvand, R. and Rodrigues, P.C. (2016). Missing value imputation in time series using Singular Spectrum Analysis. *International Journal of Energy and Statistics 4*. DOI: 10.1142/S2335680416500058
[5] Mahmoudvand, R. (2012). Some Theoretical Development of the Singular Spectrum Analysis. Ph.D. Thesis (In Persian), Shahid Beheshti University, Tehran, Iran.
[6] Patterson, K., Hassani, H., Heravi, S., and Zhigljavsky, A. (2011). Forecasting the final vintage of the industrial production series, *Journal of Applied Statistics*, 38, 2183–2211.