Journal of Data Science 0 (0), 1–13
??? 2025
DOI: 10.6339/25-JDS1164
Statistical Data Science

An Innovative Method of Singular Spectrum Analysis to Conduct Gap-filling and Denoising on Time Series Data

James J. Yang[1,*] and Anne Buu[2]

[1]*Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, U.S.A.*
[2]*Department of Health Promotion and Behavioral Sciences, University of Texas Health Science Center at Houston, U.S.A.*

Abstract

Heart rate data collected from wearable devices – one type of time series data – could provide insights into activities, stress levels, and health. Yet, consecutive missing segments (i.e., gaps) that commonly occur due to improper device placement or device malfunction could distort the temporal patterns inherent in the data and undermine the validity of downstream analyses. This study proposes an innovative iterative procedure to fill gaps in time series data that capitalizes on the denoising capability of Singular Spectrum Analysis (SSA) and eliminates SSA's requirement of pre-specifying the window length and number of groups. The results of simulations demonstrate that the performance of SSA-based gap-filling methods depends on the choice of window length, number of groups, and the percentage of missing values. In contrast, the proposed method consistently achieves the lowest rates of reconstruction error and gap-filling error across a variety of combinations of the factors manipulated in the simulations. The simulation findings also highlight that the commonly recommended long window length – half of the time series length – may not apply to time series with varying frequencies such as heart rate data. The initialization step of the proposed method that involves a large window length and the first four singular values in the iterative singular value decomposition process not only avoids convergence issues but also facilitates imputation accuracy in subsequent iterations. The proposed method provides the flexibility for researchers to conduct gap-filling solely or in combination with denoising on time series data and thus widens the applications.

Keywords *heart rate; imputation; missing data; wearable device*

1 Introduction

Heart rate data collected from wearable devices – one type of time series data – could provide insights into activities, stress levels, and health (Yang et al., 2024). Yet, missing measurements commonly observed in these data could negatively impact the reliability and interpretability of subsequent analyses. Two types of issues, participant-related and device-related, contribute to missing heart rate measurements (Wu et al., 2020). Participant behavior, including improper device placement or removal during certain activities (e.g., swimming, bathing), could lead to gaps in the data. Technical issues with the device, including poor signal transmission due to environmental noise or device malfunction, could also result in missing measurements. Failing to

*Corresponding author. Email: James.J.Yang@uth.tmc.edu.

Received June 28, 2024; Accepted January 3, 2025

address this issue is likely to distort the temporal patterns inherent in the data and undermine the validity of downstream analyses, as many statistical methods for time series data assume completeness.

Missing values in heart rate data typically manifest in two distinct forms: random missingness and consecutive missing segments (i.e., gaps). Random missing data occur sporadically throughout the data collection period and are often of short duration. In contrast, gaps represent prolonged periods of missing data, ranging between minutes and hours. While simple imputation methods like linear interpolation may suffice for addressing random missingness due to the dense sampling intervals commonly adopted by modern wearable devices, the presence of gaps poses a real challenge that necessitates specialized imputation techniques.

Existing literature on time series imputation offers a range of methods, including regression-based approaches, expectation-maximization (EM) methods, and matrix factorization-based techniques (Fang and Wang, 2020). Yet, these methods may not be directly applicable to the gap-filling problem in heart rate data, particularly when dealing with extended missing segments. Singular Spectrum Analysis (SSA) is a widely adopted method for noise reduction and data reconstruction in time series data (Golyandina and Zhigljavsky, 2020) and has been shown to outperform other imputation methods (Hassani et al., 2019). Nevertheless, applying SSA to heart rate data presents challenges, primarily due to the need to pre-specify the window length and number of groups, which could be difficult to determine for complex and noisy data like heart rate measurements.

To fill the current methodology gap, we propose a novel approach for time series imputation that is based on SSA but does not require pre-specification of the window length and number of groups. The remainder of this paper is organized as follows: Section 2 provides an overview of SSA-based imputation methods and details our proposed approach for gap-filling. In Section 3, we demonstrate the effects of window length and initial value choices on results of SSA using simulations based on the sum of sinusoidal functions. We also present the findings of a simulation study based on real heart rate data to demonstrate the effectiveness of our proposed method (i.e., with smaller construction error) in comparison to SSA-based methods. Finally, Section 4 concludes the advantages of the proposed method and suggests possible directions for future research.

2 Methods

Singular Spectrum Analysis (SSA) is fundamentally a non-parametric, model-free technique that reduces the dimensionality of a time series through singular value decomposition (SVD). Unlike conventional statistical methods, SSA does not require assumptions of stationarity or normality, as demonstrated by Golyandina et al. (2001). This flexibility allows SSA to be applied broadly to explore and reconstruct the structure of time series data. Furthermore, Golyandina (2020) established that being model-free, SSA can accurately capture the signal's explicit form when the underlying time series follows a parametric model, further highlighting its versatility. Our proposed Enhanced Singular Spectrum Analysis (ESSA) and gap-filling methods inherit the same model-free framework of SSA that does not require explicit specification of a parametric model, and thus are suitable for a wide range of time series applications without imposing restrictive assumptions.

2.1 A Brief Review of Singular Spectrum Analysis (SSA)

SSA is a non-parametric method used to extract signal components from noisy time series data. It aims to produce a denoised time series of the same length as the original input. Consider a time series $y_1, \ldots, y_n$ of length n. Given a window length w (where $w < n/2$), subseries are defined as column vectors $\boldsymbol{y}_j = (y_j, y_{j+1}, \ldots, y_{j+w-1})'$, for $j = 1, \ldots, n-w+1$. The four fundamental steps of SSA are briefly reviewed below. For detailed algorithmic explanations, interested readers may refer to the following references: Sanei and Hassani (2015); Golyandina et al. (2018); and Golyandina and Zhigljavsky (2020).

The first step of SSA is *embedding* where the trajectory matrix $\boldsymbol{Y}$ is constructed by stacking these subseries: $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-w+1})$. Note that $\boldsymbol{Y}$ is a Hankel matrix of size $w \times (n-w+1)$, where anti-diagonal elements are identical.

The second step employs *Singular Value Decomposition (SVD)* on $\boldsymbol{Y}$ to decompose it into a sum of rank one matrices: $\boldsymbol{Y} = \sum_{i=1}^{w} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i'$. Here, $\lambda_1 \geqslant \cdots \geqslant \lambda_w \geqslant 0$ denote the singular values; $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_w$ represent the left singular vectors; and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_w$ are the right singular vectors.

The third step, *grouping*, involves obtaining a low-rank approximation of $\boldsymbol{Y}$, $\boldsymbol{Y}_r$, by retaining only the first r singular values and their corresponding vectors: $\boldsymbol{Y}_r = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i'$.

The fourth step – *diagonal averaging* or *hankelization* – addresses the issue that anti-diagonal elements of $\boldsymbol{Y}_r$ are not identical by replacing these elements with their average value. In this way, $\boldsymbol{Y}_r$ is transformed into a Hankel matrix. Finally, this Hankel matrix is converted into a denoised time series vector by applying reverse mapping of the embedding in the first step.

2.2 Existing Methods for Gap-Filling Based on SSA

Kondrashov and Ghil (2006) proposed the first gap-filling method for time series data based on SSA. Hassani et al. (2019) later proposed a method that is equivalent to the first method but easier to implement and thus became more widely used. The procedure of this method is reviewed as follows. First, the missing values are initially imputed using the mean of the non-missing values. Second, given the window length and the number of groups, SSA is applied to reconstruct the series. Third, the missing values in the original time series are updated with the reconstructed series. Steps 2 and 3 are iterated until the imputed values converge. In our simulation studies, we will evaluate the performance of our proposed method in comparison to this commonly adopted SSA-based method.

2.3 Enhanced Singular Spectrum Analysis (ESSA)

Applying SSA to fill gaps in time series data, as described in Section 2.2, unavoidably inherits limitations of SSA, particularly the requirement of pre-specifying the window length and the number of groups of which the optimal values are challenging to find as the underlying structure of time series data is unknown in practical settings. To address this important limitation, we propose an iterative procedure, the Enhanced Singular Spectrum Analysis (ESSA), as follows:

1. Center the series: The series is centered by the mean of non-missing values to ensure that the dominant singular value does not overwhelm the subsequent Singular Value Decomposition (SVD) step.
2. Extract series using SSA with various window lengths: Conventional SSA with the number of groups set to 2 is utilized to extract series under various window lengths (a series is extracted under each length). A geometric sequence from 2^4 to $2^{\lfloor \log_2(n/2) \rfloor}$ is adopted to balance between

short and long window lengths. The mean of all extracted series forms the final series from this step.

3. Sequential SSA with various window lengths: The residual series, which is the difference between the original series and the resulting series from the previous step, serves as the input for the previous step to further extract series.
4. Stop rule determination: The stopping criterion for ESSA is based on the singular values derived from the residual series. If the residual series contains no meaningful information, the square of the singular values of the trajectory matrix derived from it would follow a symmetric distribution. Otherwise, it would exhibit characteristics resembling a Poisson distribution (Bose and Mitra, 2002; Bryc et al., 2006). Thus, by conducting a symmetric test (Miao et al., 2006) on the distribution of the squared singular values, we could determine whether the residual series contains mere noise (e.g., a p-value > 0.05).

2.4 The Proposed Method for Gap-Filling Based on ESSA

We propose an innovative method to incorporate the ESSA into the gap-filling method described in Section 2.2 so the improved procedure does not require pre-specifying the window length, and the number of groups is determined by statistical testing. This method first initializes the missing value imputation using SSA with a large window length and then iteratively updates missing values using the ESSA procedure. The detailed imputation algorithm is described below.

1. Initialization: The mean value of the non-missing data is used to impute the missing values. Subsequently, the SSA procedure with a window length equal to half of the time series length and the number of groups set to four is applied to update the missing values until convergence.
2. ESSA Imputation:
   (a) Reconstruct the time series using ESSA with the number of groups set to two.
   (b) Update the missing values with the reconstructed values.
   (c) Repeat Steps (2a) and (2b) until the values at the missing locations converge. Designate the converged, denoised series as the reconstructed series.
   (d) Calculate the residual series as the difference between the series from Step (2a) and the reconstructed series from Step (2c).
   (e) Stop if the symmetric test result indicates that the residual series contains no signal. Otherwise, return to Step (2a) with the residual series as the input series.
   (f) The final series is the sum of all reconstructed series from Step (2c).

In Step (2c) of the ESSA imputation, the theoretical assurance is grounded in the convergence properties of the iterative imputation process for the trajectory matrix using singular spectrum analysis (SSA) based on singular value decomposition (SVD). As demonstrated by Caussinus (1986a), the iterative SVD approach can be viewed as a specific case of the expectation-maximization (EM) algorithm applied to the fixed-effect model. The convergence of the EM algorithm has been rigorously established by Dempster et al. (1977) and applies to our method. Additionally, our simulation studies confirm that the proposed method converges fairly quickly, further supporting its practical implementation.

In Step (2e) of the ESSA imputation, we utilize the singular values of the trajectory matrix to determine if the series contains no signal. If the series consists of only noise, the p-value of the symmetry test would be large, triggering a stop of the iterative procedure. Practically, the threshold for the p-value is set at 0.05, and a window length of 100 is used for creating the trajectory matrix to ensure a sufficient number of singular values for the test.

It is important to highlight two key features of the proposed method. First, the procedure results in a final series with imputed missing values and thus fills the gaps. Second, the final series is automatically denoised due to the intrinsic noise reduction capability of ESSA. In cases where the objective is to impute missing values solely, one may simply update the missing values in the original series with the imputed values obtained from the gap-filling procedure.

The proposed method based on ESSA as described above not only fills the gap but also denoises the imputed data. This dual capability ensures that the filled data serve as reliable point predictors while maintaining the integrity of the signal. Consequently, the tasks of gap-filling and denoising are inherently linked in the proposed approach, providing a comprehensive solution for handling noisy and incomplete time series data.

3 Simulations

Sections 3.1 and 3.2 describe the simulations based on the sum of sinusoidal functions to demonstrate the effects of window length on SSA's denoising and gap-filling capabilities, respectively. Section 3.3 presents a simulation study based on the sum of sinusoidal functions to evaluate the effect of initial values on gap-filling performance. Section 3.4 evaluates the performance of the proposed ESSA-based method in comparison to SSA-based methods through simulations based on real heart rate data.

3.1 The Effect of Window Length on SSA's Denoising Capability

The impact of window length on the capability of SSA to denoise time series data was examined by simulating two time series data sets with distinct characteristics: one with fixed frequencies and the other with varying frequencies.

The first time series, $y_1(t)$, was generated as $y_1(t) = x_1(t) + \epsilon(t)$, $x_1(t) = 2\cos(2\pi f_1 t) + \sin(2\pi f_2 t)$, and $\epsilon(t)$ represents random noise following a standard normal distribution. The frequencies, f_1 and f_2, are fixed at 0.05 and 0.3 respectively, with t ranging from 1 to 1000.

The second time series, $y_2(t)$, was generated similarly as $y_2(t) = x_2(t) + \epsilon(t)$, where $x_2(t)$ follows the same functional form as $x_1(t)$, but its frequencies vary across different time intervals. Specifically, the series was divided into ten disjoint sub-series each of which contains 100 consecutive data points. Within each sub-series, the frequencies f_1 and f_2 were drawn from normal distributions: $f_1 \sim \mathcal{N}(0.05, 0.01^2)$ and $f_2 \sim \mathcal{N}(0.3, 0.1^2)$. Thus, while the average frequencies remain the same, they exhibit small variation across different intervals.

For $x_1(t)$, the trajectory matrix has four positive singular values. When noise is added to generate $y_1(t)$, the trajectory matrix derived from $y_1(t)$ also exhibits four relatively large singular values. Therefore, the number of groups of SSA was set to be four in this simulation study that aims to compare the reconstruction errors across window lengths varying from 10 to 500 in increments of 10.

The performance was evaluated using the root mean square error (RMSE), defined as the square root of the mean squared difference between the reconstructed series and the true series (i.e., $x_1(t)$ or $x_2(t)$). The RMSE was chosen over the widely used criterion for prediction accuracy, the mean absolute percentage error (MAPE), because the latter depends heavily on the magnitude of the observed values, making it sensitive to the scale of the data. The results are depicted in the top panel of Figure 1, where the RMSE for the first time series is denoted by "$\circ$," and the RMSE for the second time series is denoted by "$\times$."

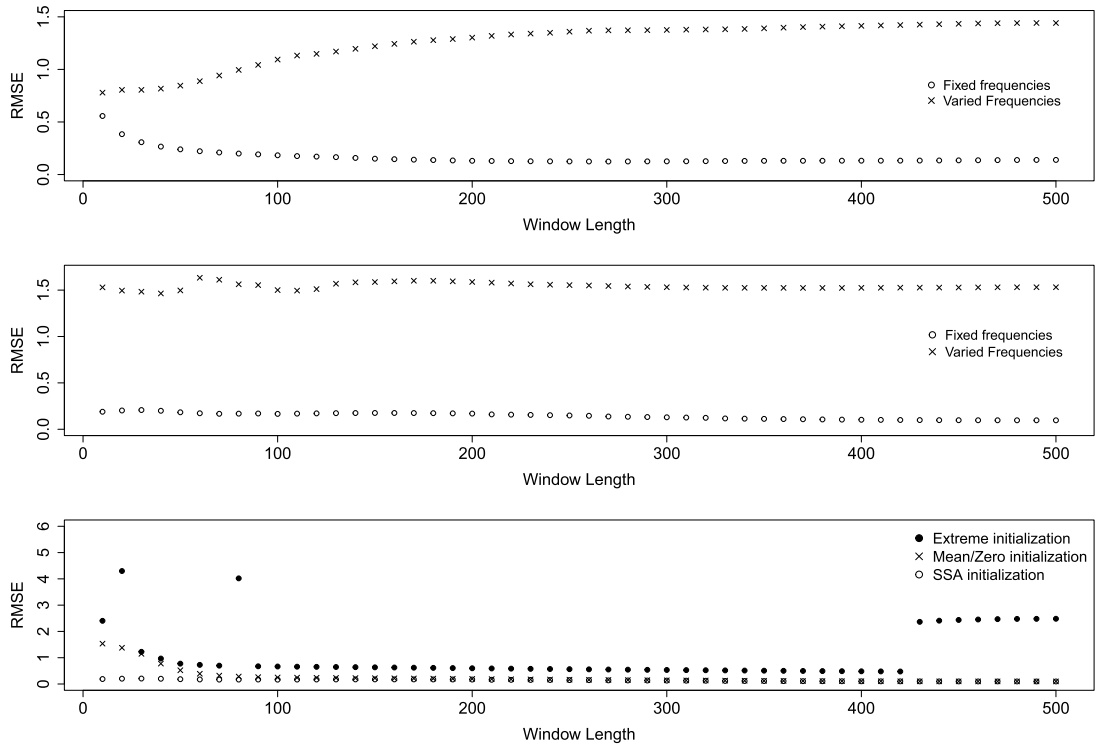

Figure 1: Top: The effect of window length on Singular Spectrum Analysis's performance on denoising time series data with fixed and varying frequencies. Middle: The effect of window length on Singular Spectrum Analysis's performance on filling gaps in time series data with fixed and varying frequencies. Bottom: The effect of three ways of initial value choice on the performance of gap-filling.

The top panel of Figure 1 shows a contrast between the times series with fixed frequencies and the one with varying frequencies. The RMSE for the time series with fixed frequencies (i.e., $y_1(t)$) decreases as the window length increases. This is consistent with findings from prior studies (Golyandina, 2010). The RMSE reaches its minimum when the window length approaches half of the total time series length, suggesting that using half the series length as the window length is optimal for such data. However, for the time series with varying frequencies (i.e. $y_2(t)$), the RMSE increases with larger window lengths. This indicates that smaller window lengths are more effective for reconstructing this type of time series data. Although the frequency variation in $y_2(t)$ is considered minimal, it significantly impacts the reconstruction accuracy.

3.2 The Effect of Window Length on SSA's Gap-Filling Capability

To evaluate the capability of SSA to fill gaps in time series data, we generated missing values in the time series data from the previous section (i.e., $y_1(t)$ and $y_2(t)$). Specifically, we removed the data points indexed from 376 to 625, resulting in 25% of the total time series data being missing. We evaluated the effect of window length on SSA's gap-filling capability using the RMSE calculated solely within the interval containing the missing data. The results are presented in the middle panel of Figure 1, where the RMSE for the first time series (with fixed frequencies) is denoted by "$\circ$," and the RMSE for the second time series (with varied frequencies) is denoted by "$\times$."

The middle panel of Figure 1 demonstrates that the impact of window length depends on whether the time series has fixed or varying frequencies. For the time series with fixed frequencies, the RMSE decreases as the window length increases. This suggests that a larger window length is beneficial for accurately imputing the missing values in the gap. For the time series with varying frequencies, in general, larger window lengths tend to produce higher RMSE values for gap-filling, whereas smaller window lengths, particularly around 50, result in RMSE values closer to the minimum. This indicates that for time series with variable frequency components, smaller window lengths are more effective in minimizing the reconstruction error within the gaps.

3.3 The Effect of Initial Values on Gap-Filling Performance

Both the SSA-based (Section 2.2) and the ESSA-based (Section 2.4) gap-filling methods are iterative processes that are likely to be impacted by the choice of initial values. Yet, this issue has rarely been investigated. To fill this knowledge gap, we conducted a simulation study using the time series data $y_1(t)$ with 25% missing values, as described in the previous section. To evaluate the effect of initial values on gap-filling performance, we examined three different scenarios:

1. Large fixed initial values: A relatively large fixed value (5 in this case) was assigned as the initial value for the missing data.
2. Mean-based initial values: The mean value calculated from the non-missing observed data was used as the initial value. This approach was adopted by the SSA-based gap-filling method.
3. SSA-derived initial values: Initially, the mean value of the non-missing data was used to impute the missing values. Then, SSA with the number of groups set to 4 was iteratively applied to refine these initial values. The window length was set to half of the total time series length. This approach, the initialization step of the proposed ESSA-based gap-filling method, ensures that the window length, coupled with the first four singular values, captures the overall structure of the time series while avoiding a trajectory matrix that contains rows with all missing values.

The effects of these initial values were compared based on the RMSE calculated over the missing intervals. The results are presented in the bottom panel of Figure 1, where the RMSE for large fixed initial values is denoted by "•"; the RMSE for mean-based initial values is denoted by "×"; and the RMSE for SSA-derived initial values is denoted by "∘."

According to the bottom panel of Figure 1, it is evident that initializing missing values with a large fixed value would result in a substantially higher RMSE, even when the initial value is within the range of observed data (like our simulation setting). Using the mean value as the initial value generally produces a relatively smaller RMSE, especially when a large window length is used. In comparison to large or mean initial values, the proposed method of SSA-derived initial values consistently results in the smallest RMSE across different window lengths.

3.4 Simulations to Evaluate the Proposed Method Based on Heart Rate Data

The imputation method proposed in this paper was motivated by heart rate data collected from young adult e-cigarette users who wore Garmin Vivosmart 5 smartwatches (Garmin, Olathe, KS, USA) 24/7 for seven days (see Yang et al., 2024 for a detailed description of the study). The smartwatch recorded beat-to-beat intervals (BBI) in milliseconds (with the mean of 0.7 seconds). These BBI data were converted into the number of heart beats at each rounded minute (i.e., heart rate) by averaging heart rate measurements in the neighborhood of 30 seconds. The resulting heart rate data tend to contain missing values that appear consecutively, forming gaps rather

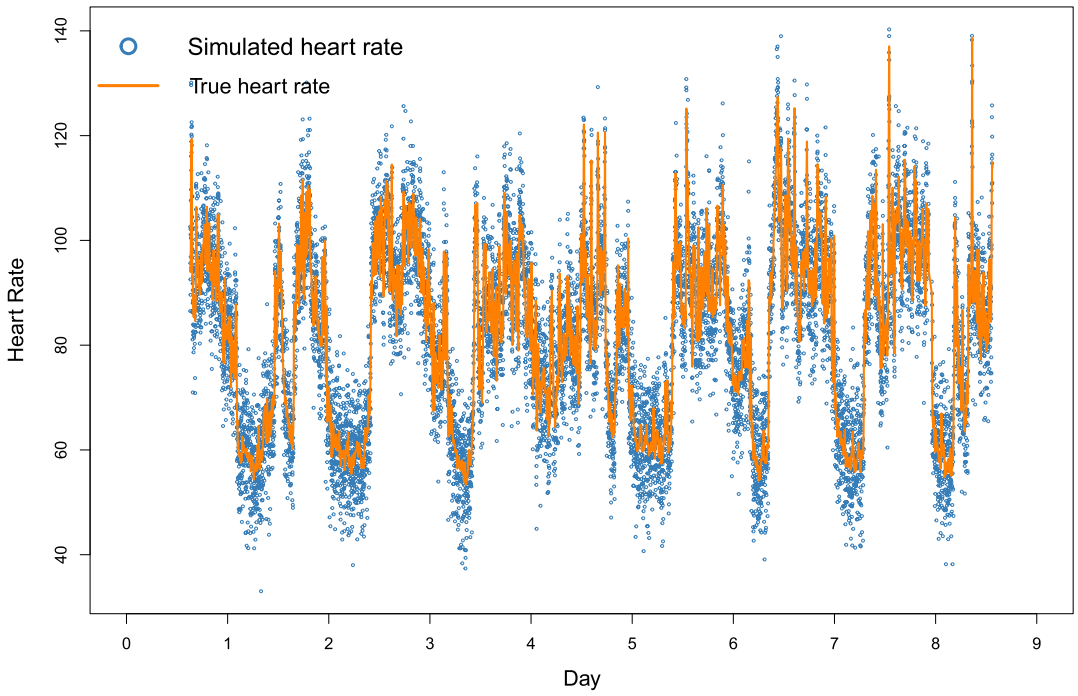

Figure 2: The true heart rate data and simulated heart rate data with added noise used in simulations.

than occurring randomly. This pattern of missing values motivated us to develop the proposed method to impute these gaps.

To evaluate performance of the proposed gap-filling method, we generated simulated data based on the feature of real heart rate data. We applied ESSA to the observed heart rate data from an individual participant, which typically contain a substantial amount of noise, to extract heart rate patterns. For this individual, a total of 11,372 heart rate points (one minute per data point) were observed. The algorithm converged after ten iterations, resulting in denoised and filled heart rate data. These data were treated in the simulation as the *t*rue heart rate data. As shown by the orange line in Figure 2, these "true" heart rate data characterize repeated day-to-day patterns with significant variation within each day. To generate simulated heart rate data, white noise from a normal distribution with the mean 0 and standard deviation 6.7 (estimated from our data) was added to the true heart rate data. The simulated data are shown as dots in Figure 2.

To manipulate the length of the gap in simulated data, we removed 5%, 10%, 20%, or 40% of consecutive heart rate data from the central portion of the time series. The SSA-based and the proposed ESSA-based methods were both used to impute these gaps for performance comparison. For the SSA-based method that requires pre-specification of the number of groups and window length, we considered a range of number of groups from 2 to 14, in combination with a range of window lengths from 16 to 4096 to encompass short to long window lengths. Since the heart rate data consist of two parts – non-missing and missing parts – we evaluated the performance of the gap-filling methods on each part separately.

For the non-missing part, we calculated the RMSE to reflect the difference between the denoised and true heart rate data. The results are shown in Figure 3. The reconstruction errors

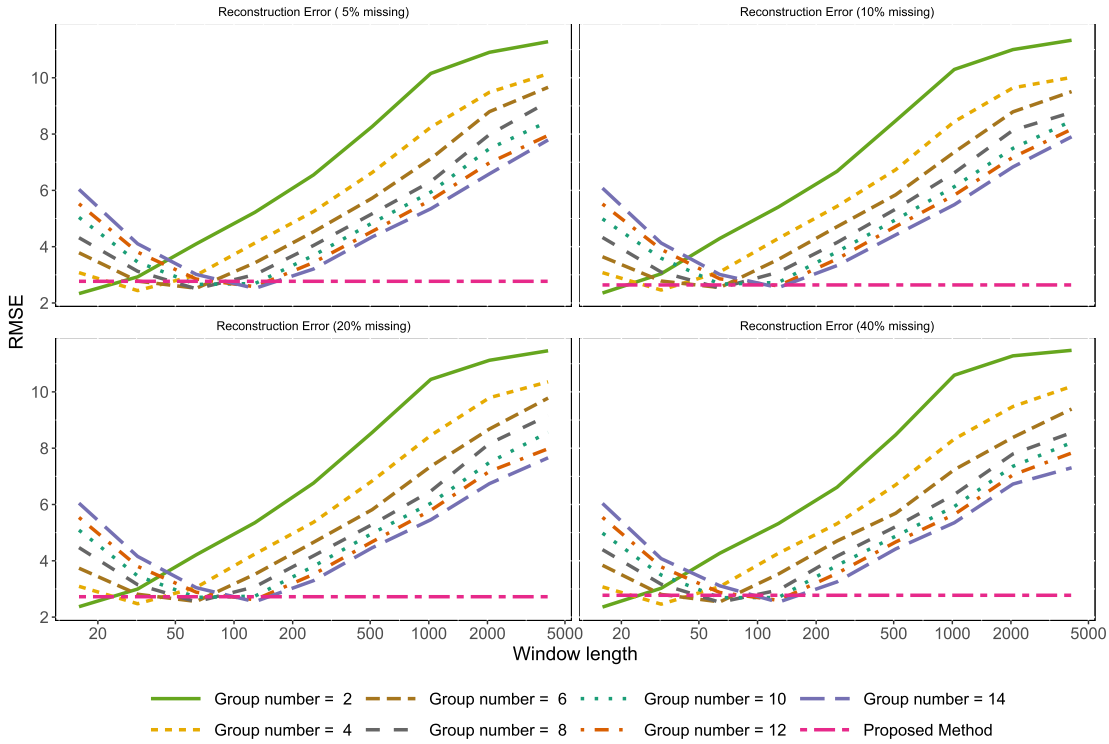

Figure 3: Evaluate the performance of the proposed ESSA method regarding reconstruction error (RMSE) for non-missing heart rate data, compared to SSA-based methods with varying window lengths and number of groups, across different percentages of missing values.

are similar across different missing rates. The performance of the SSA-based method depends on the window length and the number of groups. Given a large window length, the RMSE decreases with an increasing number of groups. For a large number of groups (e.g., 14), the RMSE decreases when the window length increases from small (16) to medium (around 100), but from that point on, the RMSE increases as the window length increases. Thus, determining the optimal window length and number of groups is crucial for the SSA-based method, and the effects of these two parameters on RMSE are not independent. On the contrary, the proposed method does not require the input of window length and number of groups. As shown by the horizontal blue dashed lines in Figure 3, the proposed method consistently achieves the smallest RMSE, demonstrating its superiority over the SSA-based method for denoising heart rate data.

For the missing part, we calculated the RMSE to reflect the difference between the imputed heart rate and the true heart rate data. The results are delineated in Figure 4. The performance of the SSA-based method depends on the missing rate, group number, and window length. The range of RMSE varies across different missing rates. For small missing rates (e.g., 5%), the RMSE could be as large as 20 using the SSA-based method, particularly for a large number of groups. As the percentage of missing values increases, the RMSE generally decreases. In this simulation setting, using the SSA-based method with small window lengths tends to produce smaller RMSE across different percentages of missing data. Compared to the SSA-based method with all combinations of the number of groups and window length, the proposed method (denoted by the horizontal blue dashed line) generally produces smaller RMSE.

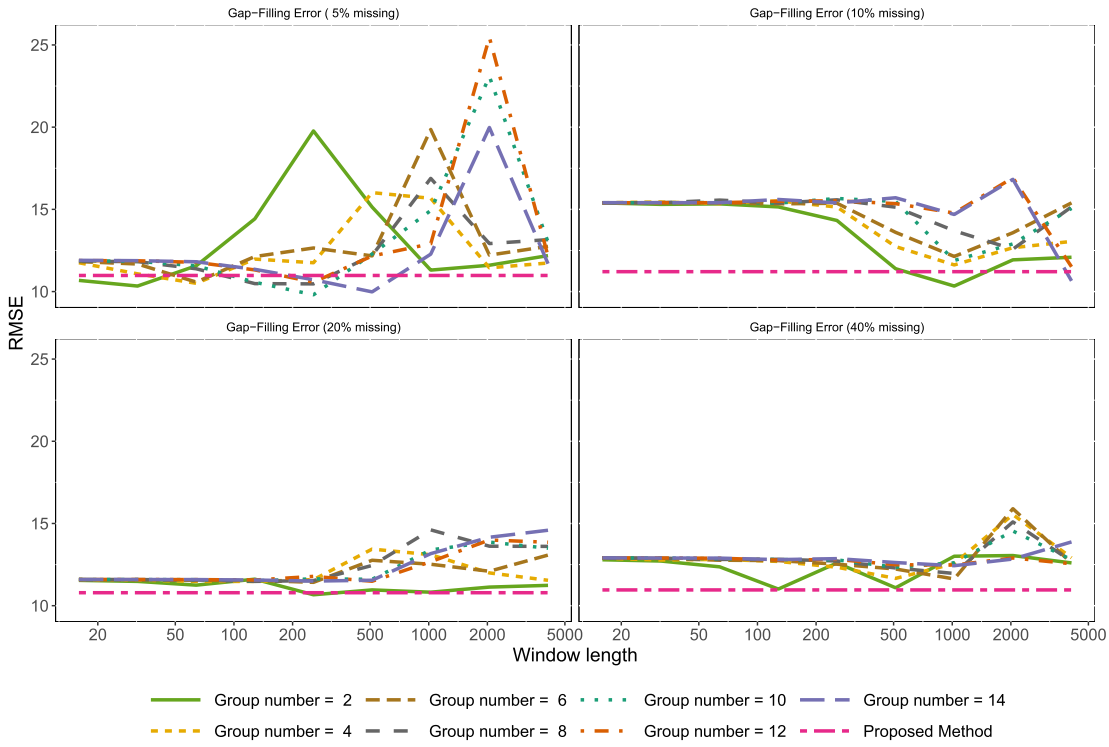

Figure 4: Evaluate the performance of the proposed ESSA method regarding gap-filling error (RMSE) for missing heart rate data, compared to SSA-based methods with varying window lengths and number of groups, across different percentages of missing values.

4 Discussion

This study proposes an innovative iterative procedure to fill gaps in time series data that capitalizes on the denoising capability of SSA and eliminates SSA's requirement of pre-specifying the window length and number of groups. Unlike SSA-based methods of which the performance depends on the choice of window length, number of groups, and the percentage of missing values, the proposed method consistently achieves the lowest rates of reconstruction error and gap-filling error across a variety of combinations of the factors manipulated in the simulations. The proposed procedure also provides the flexibility for researchers to conduct gap-filling solely or in combination with denoising and thus widens the applications.

The simulation findings highlight that the optimal window length for SSA-based gap-filling methods depends on the features of time series. For time series with fixed frequencies, a larger window length leads to smaller reconstruction and gap-filling errors. Conversely, for time series with varying frequencies, a smaller window length yields better results. Thus, the commonly recommended long window length – half of the time series length (Golyandina, 2010) – should be applied with caution when dealing with time series data exhibiting frequency variability. Physiological data such as heart rate series, of which frequency characteristics tend to vary over time, would be a perfect example.

A crucial and yet often overlooked aspect in imputation is the choice of initial values. Based on the work by Caussinus (1986b), the iterative SVD, the key step of SSA-based imputation,

can be viewed as a specific expectation-maximization (EM) algorithm for the fixed effect model. Since the EM algorithm converges to a *local minimum*, the initial values used in the iterative steps of the trajectory matrix become crucial. Through simulations, this study demonstrates that improper initial values could lead to poor performance. Although using the mean value performs well in most situations, it falls short when the window length is small. In contrast, the proposed method, which can be considered as a global initialization technique, provides better initial values. The initialization step of the proposed method that involves a large window length and the first four singular values in the iterative SVD process has multiple advantages. First, when the length of gap is less than 50% of the total length, the trajectory matrix derived with a large window length does not contain any all-missing columns and thus avoids convergence issues. Second, using a large window length combined with the first four singular values ensures that the major structure of the time series is captured, and thereby improves imputation accuracy in subsequent iterations.

The proposed method is classified as a single imputation method that is more suitable than the commonly adopted multiple imputation in the context of time series analysis. First, the multiple imputation requires specification of a data generation mechanism that tends to be complex and often unknown for time series data. Our method, on the other hand, does not rely on a parametric model for data generation and thus is less prone to biases resulting from incorrect model specification. Second, the strength of multiple imputation lies in its ability to estimate the uncertainty associated with parameter estimates which does not match well with the primary goal of time series data analysis that is to reconstruct the series and recover missing temporal patterns. Third, implementing the multiple imputation on time series data would encounter tremendous computational challenges as a large number of imputations would be required to accurately estimate the standard error of a parameter estimator when the parameter space is high-dimensional.

The primary requirement for imputing missing data using the proposed methods is that the length of any uninterrupted gap should be less than half of the total length of the time series. This ensures that the trajectory matrix retains sufficient structure for decomposition and reconstruction. Notably, our simulation studies demonstrated that the proposed ESSA-based method performed robustly even when up to 40% of the time series consisted of missing values in a contiguous block. This result underscores the method's capacity to handle substantial gaps while maintaining accurate reconstruction and imputation performance.

Our model assumes that the error term in the time series is pure white noise. If there is temporal dependence within the noise, it is implicitly addressed by our approach through the decomposition of the trajectory matrix. Temporal patterns are captured by the first few eigen-triples obtained via singular value decomposition (SVD), which are used to reconstruct the time series. The remaining components, characterized by smaller singular values, are treated as the error term. Because these residual components have relatively small magnitudes compared to the dominant eigen-triples, their influence on the overall reconstruction is minimal. As a result, the model is robust to mild temporal dependence in the noise.

When the time series is stationary (e.g., sinusoidal functions), both the SSA-based and the proposed ESSA-based method are expected to perform the best. Yet, time series data collected from human subjects, such as heart rate data, are often non-stationary. Thus, the simulation work based on the features of real heart rate data as conducted in this study has made a unique contribution to the literature. While our simulations show that the proposed method performs well on heart rate data, further studies are needed to evaluate its performance on other types of non-stationary time series such as accelerometer data that could capture activity or sleep

(Indic et al., 2011, 2012). Another potential direction for future research is to develop a way to adaptively adjust the distribution of window lengths used in the proposed method. The major challenge of this direction would be expensive computation time.

Alternative approaches to addressing challenges of applying SSA to time series with missing data have been proposed by recent studies. Ji et al. (2025) circumvented the requirement of complete data in SSA by employing a Toeplitz lagged covariance matrix. Yet, the objective of this method was to reconstruct the observed portion of the time series without explicitly imputing the missing segments. Fu et al. (2024) extended the improved SSA (ISSA) originally proposed by Groth and Ghil (2011) to multivariate time series to address the issue of degenerate eigenvectors through varimax orthogonal rotation. Their method simultaneously imputes missing data and enhances multivariate clustering. Yet, its reliance on a fixed window length may limit flexibility. Future research may integrate our adaptive window-length approach into this framework to improve robustness and applicability, particularly for complex multivariate and non-stationary time series.

Supplementary Material

The supplementary material includes the following files: (1) `README.md`, a brief explanation of all the files in the supplementary material; (2) `HR.csv`, the application dataset; (3) `GapFilling.jl`, the Julia module implementing the proposed method; and (4) `main.jl`, the demo program.

Funding

This research was supported by a grant funded by the National Institutes of Health (NIH): R01 DA049154 to A. Buu. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors declare no conflict of interest.

References

Bose A, Mitra J (2002). Limiting spectral distribution of a special circulant. *Statistics & Probability Letters*, 60(1): 111–120. https://doi.org/10.1016/S0167-7152(02)00289-4

Bryc W, Dembo A, Jiang T (2006). Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Annals of Probability*, 34(1): 1–38. https://doi.org/10.1214/009117905000000495

Caussinus H (1986a). Models and uses of principal component analysis. *Multidimensional Data Analysis*, 86: 149–170.

Caussinus H (1986b). Models and uses of principal component analysis. *Multidimensional Data Analysis*, 86: 149–170.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 39(1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Fang C, Wang C (2020). Time series data imputation: A survey on deep learning approaches. arXiv preprint: https://arxiv.org/abs/2011.11347

Fu Y, Fan X, Cao J (2024). An imputation method based on the varimax variant of multivariate singular spectrum analysis. *IEEE Access*, 12: 127749–127767. https://doi.org/10.1109/ACCESS.2024.3429292

Golyandina N (2010). On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and its Interface*, 3(3): 259–279. https://doi.org/10.4310/SII.2010.v3.n3.a2

Golyandina N (2020). Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(4): e1487. https://doi.org/10.1002/wics.1487

Golyandina N, Korobeynikov A, Zhigljavsky A (2018). *Singular Spectrum Analysis with R*. Springer Berlin, Heidelberg.

Golyandina N, Nekrutkin V, Zhigljavsky AA (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. CRC Press.

Golyandina N, Zhigljavsky A (2020). *Singular Spectrum Analysis for Time Series*, 2nd edition. Springer Berlin, Heidelberg.

Groth A, Ghil M (2011). Multivariate singular spectrum analysis and the road to phase synchronization. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, 84(3): 036206. https://doi.org/10.1103/PhysRevE.84.036206

Hassani H, Kalantari M, Ghodsi Z (2019). Evaluating the performance of multiple imputation methods for handling missing values in time series data: A study focused on East Africa, soil-carbonate-stable isotope data. *Stats*, 2(4): 457–467. https://doi.org/10.3390/stats2040032

Indic P, Murray G, Maggini C, Amore M, Meschi T, Borghi L, et al. (2012). Multi-scale motility amplitude associated with suicidal thoughts in major depression. *PLoS ONE*, 7(6): e38761. https://doi.org/10.1371/journal.pone.0038761

Indic P, Salvatore P, Maggini C, Ghidini S, Ferraro G, Baldessarini RJ, et al. (2011). Scaling behavior of human locomotor activity amplitude: Association with bipolar disorder. *PLoS ONE*, 6(5): e20650. https://doi.org/10.1371/journal.pone.0020650

Ji K, Shen Y, Wang F, Chen Q (2025). An efficient improved singular spectrum analysis for processing GNSS position time series with missing data. *Geophysical Journal International*, 240(1): 189–200. https://doi.org/10.1093/gji/ggae381

Kondrashov D, Ghil M (2006). Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(2): 151–159. https://doi.org/10.5194/npg-13-151-2006

Miao W, Gel YR, Gastwirth JL (2006). A new test of symmetry about an unknown median. In: Hsiung AC, Ying Z, Zhang CH (eds.), *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, 199–214. World Scientific.

Sanei S, Hassani H (2015). *Singular Spectrum Analysis of Biomedical Signals*. CRC Press.

Wu X, Mattingly S, Mirjafari S, Huang C, Chawla NV (2020). Personalized imputation on wearable-sensory time series via knowledge transfer. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1625–1634.

Yang JJ, Piper ME, Indic P, Buu A (2024). Statistical methods for predicting e-cigarette use events based on beat-to-beat interval (BBI) data collected from wearable devices. *Statistics in Medicine*, 43(17): 3227–3238. https://doi.org/10.1002/sim.10124