Supplementary for "Adaptive Seasonal-Trend Decomposition for Streaming Time Series Data with Transitions and Fluctuations in Seasonality"

Thanapol Phungtua-eng ✉ and Yoshitaka Yamamoto

Department of Informatics, Shizuoka University, Shizuoka, Japan
thanapol@yy-lab.info, yyamamoto@inf.shizuoka.ac.jp

1 Sliding Discrete Fourier transform (SDFT)

Here, this is proof for SDFT eq. (as shown in Eq. (2) in the main paper) [4,5]. Given an input sequence with a length of at least $(N+q+1)$, where q denotes the starting index of the DFT window, we consider a DFT of length N for the window $(x_q, x_{q+1}, \ldots, x_{q+N-1})$:

$$X_q = \sum_{n=0}^{N-1} x_{n+q} e^{-j2\pi nk/N} \tag{1}$$

Then, sliding to the next window with the starting point at the $(q+1)$-th position, we compute the DFT of length N for this new window $(x_{q+1}, x_{q+2}, \ldots, x_{q+N})$, dynamically tracking changes in the frequency domain as the window advances:

$$X_{q+1} = \sum_{n=0}^{N-1} x_{n+q+1} e^{-j2\pi nk/N} \tag{2}$$

Substituting $p = n+1$ for the range 1 to N, we have:

$$X_{q+1} = \sum_{p=1}^{N} x_{p+q} e^{-j2\pi(p-1)k/N} \tag{3}$$

Adjusting for the N-th term by subtracting and adding the $p=0$ case:

$$X_{q+1} = \sum_{p=0}^{N-1} x_{p+q} e^{-j2\pi(p-1)k/N} + x_{q+N} e^{-j2\pi(N-1)k/N} - x_q e^{j2\pi k/N} \tag{4}$$

The exponential terms can be factored as follows:

$$X_{q+1} = e^{j2\pi k/N} [\sum_{p=0}^{N-1} x_{p+q} e^{-j2\pi pk/N} + x_{q+N} e^{-j2\pi Nk/N} - x_q] \tag{5}$$

The $e^{-j2\pi Nk/N}$ term simplifies to $1 + j0$ for k is always integer values, since $e^{-j2\pi Nk/N} = 1$, leading to:

$$X_{q+1} = e^{j2\pi k/N}[\sum_{p=0}^{N-1} x_{p+q} e^{-j2\pi pk/N} + x_{q+N} - x_q] \tag{6a}$$

$$= e^{j2\pi k/N}[\sum_{n=0}^{N-1} x_{n+q} e^{-j2\pi nk/N} + x_{q+N} - x_q] \tag{6b}$$

$$= e^{j2\pi k/N}[X_q + x_{q+N} - x_q] \tag{6c}$$

Note that the summation enclosed in square brackets in Eq. (6a) represents the DFT calculated for the kth component, using p as the indexing variable rather than n. For the latest timestamp t, the DFT results from the current sliding window $(x_{t-N+1}, \ldots, x_t)$ and the previous sliding window $(x_{t-N}, \ldots, x_{t-1})$ are denoted as $\mathcal{F}_t$ and $\mathcal{F}_{t-1}$, respectively. This notation allows us to succinctly express the DFT update formula, transitioning from $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$ as follows:

$$\mathcal{F}_t(k) = e^{j2\pi k/N}[\mathcal{F}_{t-1}(k) + x_t - x_{t-N}] \tag{7}$$

2 Spectral Peak Location Estimation

Spectral peak location estimation interpolates index k_{peak}, which corresponds to the largest power in the Fourier transform result without an increase in N [6]. The k_{peak} is determined by $k_{peak} = \hat{k} + \delta$, where $\hat{k}$ denotes the index of the peak location from $\mathcal{P}(k)$, and δ denotes the residual frequency that can be positive or negative, as shown in Fig. 1. Spectral peak location estimation base on the cuvre-fiting technique. This supplementary report presents a comparison between the non-estimator and the hybrid Aboutanios-Mulgrew and q-shift estimator (HAQSE), which is utilized in our ASTD.

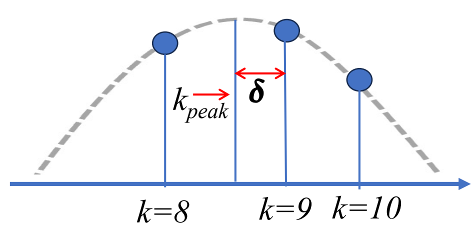

Fig. 1: Example of DFT coefficients resolution issue.

2.1 HAQSE

In this study, we utilize the HAQSE to determine k_{peak} from $\mathcal{P}(k)$ [8]. HAQSE is iterative estimator that operates at a computational cost of $O(N)$. The HAQSE algorithm processes the results of $\mathcal{P}(k)$ as follows:

1. **Peak Identification**: The peak index ($\hat{k}$) from $\mathcal{P}(k)$ is identified using $\hat{k} = \text{argmax}_k(\mathcal{P}(k))$.
2. **Initial δ_α Calculation**: The initial value of δ_α is calculated according to:

$$\delta_\alpha = \frac{N}{2\pi} \arcsin\left(\sin\left(\frac{\pi}{N}\right) \Re\left\{\frac{X_{\hat{k}+0.5} + X_{\hat{k}-0.5}}{X_{\hat{k}+0.5} - X_{\hat{k}-0.5}}\right\}\right), \tag{8}$$

   where $\Re\{\cdot\}$ denotes the real part, and X_k denotes Fourier transfrom for non-integer k values to accommodate fine-grained frequency estimation.
3. **Final δ Estimation**: The final δ is estimated using:

$$\delta = \frac{1}{c(q)} \left(\Re\left\{\frac{X_{\hat{k}+\delta_\alpha+q} + X_{\hat{k}+\delta_\alpha-q}}{X_{\hat{k}+\delta_\alpha+q} - X_{\hat{k}+\delta_\alpha-q}}\right\}\right) + \delta_\alpha, \tag{9}$$

   with $q = \frac{1}{\sqrt[3]{N}}$ and $c(q) = \frac{1-\pi q \cot(\pi q)}{q \cos^2(\pi q)}$.
4. **Frequency Estimation**: k_{peak} is estimated as $\hat{k} + \delta$. The actual frequency, which is the peak value of $\mathcal{P}(k)$, is computed as $f_{peak} = k_{peak}/N$.

Here, $X_k = \sum_{n=0}^{N-1} x_n \exp(-j2\pi nk/N)$ extends k to non-integer values, including $\hat{k} \pm 0.5$, and $\hat{k} + \delta_\alpha \pm q$, enabling HAQSE to estimate the actual frequency with higher resolution than possible with DFT's integer frequency bins. This method efficiently identifies k_{peak} by leveraging HAQSE's computational advantages, notably its $O(N)$ computation cost, which is attributed to the transformation from the time domain to the frequency domain using the twiddle factor in steps 2 and 3.

2.2 Comparison between None-estimator and HAQSE

We conducted an evaluation with a synthetic data set consisting of a sine wave with a season length of 50 instances and a total length of 500 instances. The evaluation process began with calculating $\mathcal{P}(k)$ using the data within a window, followed by identifying $\hat{k} = \text{argmax}_k(\mathcal{P}(k))$. This evaluation structured the analysis into two distinct groups.

1. **Non-estimator**: we take the reciprocal with $\hat{k}/N$ to get the season length.
2. **HAQSE estimator**: we find k_{peak} using HAQSE estimator. Then, we take the reciprocal with k_{peak}/N to get the season length.

The results are shown in Fig. 2, where the x-axis represents the window sizes and the y-axis represents the season length results provided by the estimator. Notably, N is considered to be the optimal window size for accurately determining k_{peak} if N divided by a positive integer k_{peak} equals 50. This condition ensures the most accurate determination of the peak frequency without the estimator.

The season length determined by the non-estimator is unstable and often diverges significantly from the ground truth. However, using the optimal window size can provide the correct season length, which exactly matches the ground truth. HAQSE exhibited stable results without the influence of the window size. Therefore, we used HAQSE to avoid the problem of the influence of the window size.

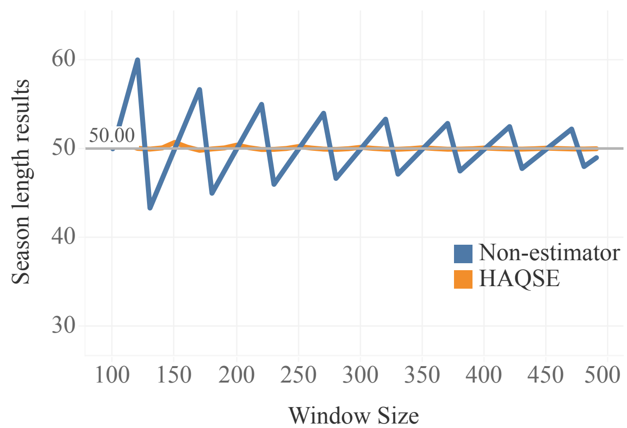

Fig. 2: Utilizing HAQSE estimator

3 Experimental Metrics for Real-world Datasets

This section provides additional details on the metrics used in our experiment on real-world datasets.

Trend Smoothness: Measures the smoothness of the trend component via the standard deviation of its first-order difference [7]. Given the trend component denoted by $T = (T_0, T_1, \ldots, T_t)$, the first-order difference of the trend component, (ΔT_i), is calculated as:

$$\Delta T_i = T_{i+1} - T_i \quad \text{for} \quad i = 0, 1, \ldots, t-1, \tag{10}$$

where t denotes the latest timestamp, indicating the total length of the trend component data. The smoothness measure, denoted as $(\sigma_{\Delta T})$, is then the standard deviation of these first-order differences:

$$\sigma_{\Delta T} = \sqrt{\frac{1}{t-1}\sum_{i=0}^{t-1}(\Delta T_i - \mu_{\Delta T})^2} \tag{11}$$

where $\mu_{\Delta T}$ is the mean of the first-order differences of trend component. Lower values of $\sigma_{\Delta T}$ idenote smoother trends.

Seasonality Presence: Measures the presence of seasonality by applying the Kruskal–Wallis test to the seasonal component [2]. Given the seasonal component denoted by $S = (S_0, S_1, \ldots, S_t)$, the Kruskal–Wallis test statistic is calculated as:

$$W = \frac{12}{N(N+1)}\sum_{j=1}^{g}\frac{U_j^2}{n_j} - 3(N+1) \tag{12}$$

where N denotes the length of S, g denotes the number of groups, n_j denotes the number of observations in the j-th group, and U_j denotes the sum of ranks in the j-th group. To determine the number of groups (g), it is set equal to the season length m, which reflects the position within the cycle [2]. For example, if we have monthly data spanning one year (with a season length of 12), we group the data into 12 groups, with each group corresponding to one month within the cycle. Therefore, we group the observations by month, starting with January as the first group, February as the second, and so on until December, which is the twelfth group. This aligns with the season length m.

To illustrate the calculation of the sum of ranks (U_j) within each group for the Kruskal-Wallis test, consider an example with three groups, resulting in each group having its unique set of data points:

- Group 1 ($j = 1$): 5, 3, 8
- Group 2 ($j = 2$): 7, 6, 2
- Group 3 ($j = 3$): 4, 9, 1

Ranks are assigned to the original observations within each group, and U_j, the sum of ranks in the j-th group, is calculated:

- U_1 for Group 1: $5 + 3 + 8 = 16$
- U_2 for Group 2: $7 + 6 + 2 = 15$
- U_3 for Group 3: $4 + 9 + 1 = 14$

Thus, U_j denotes the sum of ranks within each group. The values are $U_1 = 16$, $U_2 = 15$, and $U_3 = 14$. After calculating the Kruskal-Wallis test statistic W, it is compared against a chi-square distribution with $g-1$ degrees of freedom. The resulting p-value is used to determine the statistical significance of the observed test statistic. Lower values suggest stable repeating cycles, indicating consistent seasonality, whereas higher values may indicate inconsistency in the seasonal component.

Randomness: Measures randomness in the residual component by applying the Ljung-Box test to the residual component [3, 2]. Given the residual component denoted by $R = (R_0, R_1, \ldots, R_t)$, the Ljung-Box test statistic is calculated as:

$$Q = N(N+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{N-k} \tag{13}$$

where N denotes the length of R, h is the number of lags being tested, and $\hat{\rho}_k$ is the autocorrelation at lag k. $\hat{\rho}_k$ is calculated as:

$$\hat{\rho}_k = \frac{\sum_{i=0}^{N-1-k}(R_i - \mu_R)(R_{i+k} - \mu_R)}{\sum_{i=0}^{N-1}(R_i - \mu_R)^2} \tag{14}$$

where μ_R denotes the the mean of the residual component (R). To determine h, it is set to $\min(2m, N/5)$, where m is the season length [3]. Lower values suggest that the residual component originates from independent and identically distributed (iid) data, indicating the successful extraction of the seasonal component.

Note that the source code for the Kruskal–Wallis and Ljung-Box tests is available in ‘evaluation/02_Real1_dataset/’ [1].

References

1. Supplementary website, https://sites.google.com/view/astd-ecmlpkdd
2. Bee Dagum, E., Bianconcini, S.: Linear Filters Seasonal Adjustment Methods: Census Method II and Its Variants, pp. 79–114. Springer International Publishing, Cham (2016)
3. Hyndman, R., Athanasopoulos, G.: Forecasting: principles and practice, 2nd edition. OTexts (2018)
4. Jacobsen, E., Lyons, R.: The sliding DFT. IEEE Signal Processing Magazine **20**(2), 74–80 (2003)
5. Jacobsen, E., Lyons, R.: An update to the sliding DFT. IEEE Signal Processing Magazine **21**(1), 110–111 (2004)
6. Jacobsen, E., Kootsookos, P.: Fast, accurate frequency estimators. IEEE Signal Processing Magazine **24**(3), 123–125 (2007)
7. Mishra, A., Sriharsha, R., Zhong, S.: OnlineSTL: Scaling time series decomposition by 100x. VLDB **15**(7), 1417–1425 (2022)
8. Serbes, A.: Fast and efficient sinusoidal frequency estimation by using the DFT coefficients. IEEE Transactions on Communications **67**(3), 2333–2342 (2019)