IEICE **TRANSACTIONS**

on Electronics

VOL. E102-C NO. 2
FEBRUARY 2019

The usage of this PDF file must comply with the IEICE Provisions on Copyright.

The author(s) can distribute this PDF file for research and educational (nonprofit) purposes only.

Distribution by anyone other than the author(s) is prohibited.

A PUBLICATION OF THE ELECTRONICS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

How to Decide Window-Sizes of Smoothing Methods: A Goodness of Fit Criterion for Smoothing Oscillation Data

Kenichi SHIBATA^{†a)}, Nonmember and Takashi AMEMIYA^{†b)}, Member

SUMMARY Organic electronics devices can be applicable to implant sensors. The noises in the acquired data can be removed by smoothing using sliding windows. We developed a new criterion for window-size decision based on smoothness and similarity (SSC). The smoothed curve fits the raw data well and is sufficiently smooth.

key words: window-size, smoothing, oscillation, data processing

1. Introduction

Organic electronics devices are mechanically compatible with three-dimensionally curved surfaces of tissues and organs, and can be applicable to bioelectronic implants because of its flexible nature [1]–[3]. Since the acquired data contain not only true variations but also noises, the noises have to be eliminated to extract the true variations.

Moving average is a widely used processing method to remove noise using sliding windows. The window-size is, however, delicate. For example, oscillatory behaviors can be eliminated if the window-size is near the period. However, there are no criterion to evaluate the goodness of fit, which make researchers troubled when they decide window-sizes of moving average. This kind of problem is general and applies to the other smoothing methods that use sliding windows such as moving median, locally weighted scatterplot smoothing (LOWESS) and locally weighted scatter plot smooth (LOESS) [4].

In the previous study, we brought the first report of unicellular glycolytic oscillations in cancer cells [5]. The cancer cells exhibited oscillations in glycolytic metabolites when glucose was added to cells that were starved of glucose. The periods were diverse with representing relative standard deviations of 22% or 30% depending on the experimental conditions. The optimum window-size for each cell is obviously different. However, again, there is no goodness of fit criterion (GFC) for smoothing oscillation data. Objective criteria are needed for the window-size decision.

Akaike's information criterion (AIC) [6] is one of the most popular criteria that is used for model selections. AIC is calculated as a mathematical product of a goodness of fit (the maximum value of the likelihood function for the model) and a penalty for the number of estimated param-

eters. Since, in the case of most smoothing methods, the number of parameter is fixed depending on the method, AIC is not applicable. Underlying concept of AIC is a balance of benefit and risk. The benefit of smoothing is obviously the smoothness of smoothed data. The risk of over-smoothing is deviation from the raw data. In other words, its risk is the decrease in a similarity between the raw data and the smoothed data. Since smoothness trades off with similarity, a product of smoothness and similarity presumptively has a peak where the smoothness-similarity balance is the best. Therefore we developed a smoothness-similarity criterion (SSC) that is a product of smoothness and similarity as a GFC for smoothing oscillation data.

In this study, we figured out the window-size dependencies of smoothness, similarity, and SSC, and thereby evaluated the availability of SSC as a GFC for smoothing oscillation data.

2. Theorem

Firstly, the symbols y and $\hat{y}$ were defined as raw data and smoothed data as a function of x , respectively:

$$\begin{cases} y = f(x) + \epsilon \\ \hat{y} = S(y), \end{cases} \quad (1a) \quad (1b)$$

where $f(x)$, ϵ , and S are true value, error, and smoothing function, respectively. Additionally, to avoid reductions of the data length, the range of a window for y_i was defined to be from $k_{1,i}$ to $k_{2,i}$ as follows, which is the default of LOESS:

$$\begin{cases} k_{1,i} = \max(1, i - \lceil (K - 1)/2 \rceil) \\ k_{2,i} = \min(i + \lfloor (K - 1)/2 \rfloor, N), \end{cases} \quad (2a) \quad (2b)$$

where K and N are a window-size and a data length of x , respectively. To allow even-numbered window-size, ceil and floor function were used in Eqs. (2a) and (2b), respectively, which is the default setting of `movmean` function of MATLAB (MATLAB R2018a; MathWorks, Natick, MA, USA).

2.1 Smoothness

Smoothness can be expressed as smallness in numbers of peaks and valleys. Since the peak and valley numbers depend on the data size, its ratio for maximum possible number of peak and valley is more appropriate for smoothness index. Thus the smoothness s can be expressed as

Manuscript received June 13, 2018.

Manuscript revised August 6, 2018.

[†]The authors are with Yokohama National University, Yokohama-shi, 240–8501 Japan.

a) E-mail: shibata-kenichi-ym@ynu.jp

b) E-mail: amemiya-takashi-jk@ynu.ac.jp

DOI: 10.1587/transele.2018OMS0003

$$s := 1 - \frac{N_{\text{peak}} + N_{\text{valley}}}{N - 2}, \quad (3)$$

where N_{peak} and N_{valley} are numbers of peaks and valleys of $\hat{y}$, respectively. The denominator, namely the maximum possible number of peaks and valleys, is $N - 2$ since the both ends of the data can be neither peak nor valley. The value of s ranges from 0 (rough) to 1 (smooth).

2.2 Similarity

Decreases in similarity between raw data and the smoothed data are not favorable. Ideally, the mean of difference between raw data and the smoothed data is zero at any part. In other words, p -values of paired Student's t -test between them should be higher (do not reject the null hypothesis of no difference between them). Therefore, firstly, we calculated the p -values (p_i) of two-sided paired Student's t -test between y and $\hat{y}$ for each sliding windows. Secondly the median of the p_i was used as the representative value, and written as p . The value of p ranges from 0 (dissimilar) to 1 (similar).

2.3 SSC and Weighted SSC

Since smoothed data should be smooth “and” — but not “or” — similar to the raw data, SSC was defined as a product, but not a sum, of s and p :

$$\text{SSC} := sp. \quad (4)$$

The value of SSC ranges from 0 (rough or dissimilar) to 1 (smooth and similar). Thus the window-size that gives the higher SSC is the more desirable. This can be generalized by weighting smoothness and similarity by following equation:

$$\text{weighted SSC} := s^{2(1-r)} p^{2r} \quad (5)$$

where r is a weight of the similarity ($0 < r < 1$). Equation (4) is a special case of Eq. (5) when r is 0.5.

3. Methods

The oscillation data of glycolytic oscillations in cancer cells [5] and two often used smoothing methods, i.e. simple moving average (SMA) and LOESS [4], were used to verify SSC as a GFC for smoothing oscillation data. These two smoothing methods require users to decide window-sizes as with the most smoothing methods.

By LOESS, $\hat{y}_i$ is computed as

$$\hat{y}_i = a_i + b_i x_i + c_i x_i^2. \quad (6)$$

The values for a_i , b_i , and c_i are found by a polynomial regression in a neighborhood of x_i minimizing W_i , where

$$\begin{cases} W_i = \sum_{k=k_{1,i}}^{k_{2,i}} w_i(x_k) (y_k - (a_i + b_i x_k + c_i x_k^2))^2 & (7a) \\ w_i(x_k) = \left(1 - \left|\frac{x_i - x_k}{d_i}\right|^3\right)^3 & (7b) \\ d_i = \max(x_i - x_{k_{1,i}}, x_{k_{2,i}} - x_i). & (7c) \end{cases}$$

By SMA, $\hat{y}_i$ is computed as

$$\hat{y}_i = \frac{1}{n_i} \sum_{k=k_{1,i}}^{k_{2,i}} y_k, \quad (8)$$

where n_i is the window-size ($k_{2,i} - k_{1,i} + 1$). The data analysis was performed using MATLAB.

4. Results and Discussion

The data of the cancer glycolytic oscillations [5] were smoothed by either LOESS or SMA with a range of window-sizes, and the SSC were computed. A typical case of SSC based window-size determination of LOESS for a oscillation data is shown in Fig. 1. The oscillation period is approximately 35 s. The smoothness exhibited an inverse exponential relationship with the window-size (Fig. 1 A, blue line). In contrasts, the similarity showed a negative relationship with the window-size (Fig. 1 A, green line). Thus, there is a concave relationship between the SSC and the window-size (Fig. 1 A, red line). When the data was smoothed by LOESS using the window-size at the maximum SSC, the smoothed curve fits the raw data well and is greatly smooth (Fig. 1 B).

Additionally, a SSC based window-size determination of SMA for the same data as Fig. 1 was performed (Fig. 2). The smoothness sharply increased when window-size is smaller than approximately 10 s, and gradually increased when window-size is larger than that (Fig. 2 A, blue line). The similarity decreased while waving (Fig. 2 A, green line). The increase in smoothness canceled the trend of similarity, and the waves remained in the SSC (Fig. 2 A, red line). The maximum SSC was obtained at the first local maximum. When the data were smoothed by SMA using the

Fig. 1 A SSC based window-size determination of LOESS for a data of the cancer glycolytic oscillations. (A) Window-size dependencies of SSC (red), similarity (green), and smoothness (blue). The red circled dot is the point of the maximum SSC, where the window-size is 32.5 s. (B) Raw data (green) and the data smoothed using the window-size of 32.5 s (red). The inset shows the third oscillatory peak.

Fig. 2 A SSC based window-size determination of SMA for the same data of Fig. 1. (A) Window-size dependencies of SSC (red), similarity (green), and smoothness (blue). The points B, C, and D are the point of the maximum SSC, the first local minimum, and the second local maximum, respectively; the window-sizes at B, C, D were 8.5 s, 25 s, and 35 s, respectively. The alphabets represent the panel labels of this figure. Dotted horizontal line represents the maximum SSC. (B–D) Raw data (green) and the data smoothed using the window-sizes at the points B–D (red). The insets show the third oscillation waves.

window-size at the maximum SSC, the smoothed curve fits the raw data well and is sufficiently smooth (Fig. 2 B). Obviously the data was oversmoothed by using the window-size at the first local minimum resulting in reduced amplitudes (Fig. 2 C). Furthermore, the window-size at the second local maximum caused eliminations of the oscillation waves because the window-size equals the oscillation period (Fig. 2 D). However, the SSC at the second local maximum is comparable to the maximum SSC (Fig. 2 A).

The following is the cause of the second local maximum of the SSC. Since a window-size which is equal to a oscillation period eliminates the waves and extracts the trends, the difference between a raw data and its trend results in waves around zero. In this study, the same sliding window was used for both the smoothing and p_i calculation, which means the window-size that is used to calculate p_i is equal to the oscillation period under this condition, and just one cycle of the oscillations is included in the sliding window. Therefore, the mean of one cycle of oscillations around zero is zero, and hence the p_i for the window is close to 1. This raise the median of p_i and caused the second local maximum of the similarity and the SSC. In the same way, this happens when a window-size is m -times a oscillation period theoretically, where m is a natural number.

There are two options to avoid this problem: (a) focus only on the first local maximum, (b) change the weight balance between smoothness and similarity by weighting.

Fig. 3 A weighted SSC based window-size determination of SMA with $r = 2/3$. (A) Window-size dependencies of SSC (red), similarity (green), and smoothness (blue). The red circled dot is the point of the maximum SSC, where the window-size is 5.5 s. Dotted horizontal line represents the maximum SSC. (B) Raw data (green) and the data smoothed using the window-size of 5.5 s (red). The insets show the third oscillation wave.

Each peak of the similarity represents an elimination of certain waves, and the first peak appears when the noise are eliminated. Thus, the option (a) is a reasonable choice. On the other hand, the maximum of the weighted SSC is relatively clearer than that of the normal SSC when r was $2/3$ (Fig. 3 A), and the smoothed data fits the raw data well and is smooth (Fig. 3 B). The option (b) is a replacement of the problem from how to decide a smoothing window-size to how to decide a weight balance between smoothness and similarity. However, since the weighted SSC is based on numerical analysis of smoothness and similarity, the weighted SSC based window-size determination is superior to a subjective decision depending on trial-and-error. Additionally, the weight balance between the two parameters is intuitively understandable. Furthermore, Eq. (5) is superior to Eq. (4) from the view point of mathematical generality and customizability.

We performed SSC based window-size determinations for generated data to investigate the limitations of SSC. SSC showed a clear peak at a window-size of 9.5 for a sinusoidal curve whose period linearly increase with t (Fig. 4 A). Although oscillations with shorter periods were eliminated with noise, a sinusoidal curve whose period linearly increase with t could be smoothed properly (Fig. 4 B). Also in the case of Lorenz chaos [7], SSC showed a peak at a window-size of 0.3125 (Fig. 4 C), and the smoothed data fits the raw data well (Fig. 4 D). Therefore, SSC based window-size determination is valid for chaotic change and data with multiple periods.

Although we developed SSC for smoothing oscillatory data, SSC is also applicable to monotonically changing data whereas there are some exceptions depending on the smoothing methods. In the case of moving average, SSC exhibits no peak when the data is horizontal line. On the other hand, in the case of LOESS, SSC exhibits no peak when the data can be fitted to a quadratic regression line. However, again, except for these cases, SSC is applicable to monotonic changes, too. Additionally, while SSC is not supposed to be used for noiseless data, SSC exhibits a local maximum value at the minimum window-size and can

Fig. 4 SSC based window-size determinations of LOESS for two types of generated oscillations with noises. (A–B) Sinusoidal curve whose period linearly increase from 0.5 to 40. (C–D) Y variable of Lorenz chaos [7]. (A and C) Window-size dependencies of SSC (red), similarity (green), and smoothness (blue). The red circled dots are the points of the maximum SSC. The window-sizes for the sinusoidal curve and Lorenz chaos were 9.5 and 0.3125, respectively. Dotted horizontal line represents the maximum SSC. (B and D) Raw data (green) and the data smoothed using the acquired window-sizes (red).

be used for noiseless data, because even slight changes that are caused by smoothing result in drastic decreases in p -value and then decreases in SSC. However, even weak noise causes a peak.

In conclusion, we proposed the (weighted) SSC as

GFC for smoothing oscillation data. When the window-size at the first local maximum of (weighted) SSC was used, a sufficiently smooth curve that fits well to the raw data was obtained. We applied the (weighted) SSC for two typical window requiring smoothing methods to determine a window-size (a parameter of a natural number). Since SSC is obtained by just a comparison between raw data and smoothed data and hence is independent of smoothing methods, this criterion might be applicable to the other types of parameters of wider range of smoothing methods such as the standard deviation of the Gaussian filter, and the filter type and maximum wavelet decomposition level of wavelet denoising depending on optimization algorithms.

References

- [1] R.M. Owens and G.G. Malliaras, "Organic electronics at the interface with biology," *MRS Bulletin*, vol.35, no.6, pp.449–456, 2010.
- [2] S. Lee, Y. Inoue, D. Kim, A. Reuveny, K. Kuribara, T. Yokota, J. Reeder, M. Sekino, T. Sekitani, Y. Abe, and T. Someya, "A strain-absorbing design for tissue-machine interfaces using a tunable adhesive gel," *Nature Communications*, vol.5, 5898, Dec. 2014.
- [3] S. Yoshimoto, T. Araki, T. Uemura, Y. Noda, and T. Sekitani, "Flexible electronics for bio-signal monitoring in implantable applications," *IEICE Electron. Express*, vol.14, no.20, 20172003, 2017.
- [4] W.S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol.74, no.368, pp.829–836, 1979.
- [5] T. Amemiya, K. Shibata, Y. Itoh, K. Itoh, M. Watanabe, and T. Yamaguchi, "Primordial oscillations in life: Direct observation of glycolytic oscillations in individual HeLa cervical cancer cells," *Chaos*, vol.27, no.10, 104602, 2017.
- [6] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Proc. 2nd International Symposium on Information Theory*, pp.267–281, 1973.
- [7] E.N. Lorenz, "Deterministic nonperiodic flow," *Journal of the Atmospheric Sciences*, vol.20, no.2, pp.130–141, 1963.