A peak fitting method for 29Si nuclear magnetic resonance spectra based on singular spectral analysis

Changjun Li 1,2, Guiliang Li 1,2, *, Gang Liu 1,2
1 School of Petroleum Engineering, Southwest Petroleum University, Chengdu 610500, China.
2 CNPC Key Laboratory of Oil & Gas Storage and Transportation, Southwest Petroleum University, Chengdu 610500, China

* Corresponding author e-mail: liguiliang@aliyun.com

Abstract. Solid-state 29Si Nuclear Magnetic Resonance (NMR) is commonly used in the detection of silicate molecular structure. However, noise in 29Si NMRS disturbs the judgment of characteristic peak, thereby affecting the determination of molecular structure types of silicates. A method of peak fitting based on Singular Spectrum Analysis (SSA) is proposed in this paper accordingly. SSA is adopted to determine the position of characteristic peaks and Gaussian fitting model (GFM) is applied to fit characteristic peak in the method, thereby realizing a quantitative analysis of 29Si NMRS. When SSA is applied, the embedding dimension \( m \) determines its accuracy. However, the methods for determining \( m \), such as correlation dimension (GP) and False Nearest Neighbors (FNN), often fail because of the differences in the shape of 29Si NMRS. Therefore, two-step greedy (TSG) is proposed to determine the embedding dimension \( m \) in the paper. The accuracy of TSG can be respectively improved by 350.7% and 366.8% compared with GP and FNN according to example verification, and the method can accurately determine the embedding dimension \( m \). The peak fitting method (TSG-SSA-GFM) is formed by combining TSG, SSA, and GFM. The experimental results show that the average error of the characteristic peak location is 0.09ppm, and goodness of fit is 99.75%. The method has good accuracy. Therefore, the study provides an important method for the quantitative analysis of 29Si NMR.

Key words: Silicate; 29Si nuclear magnetic resonance spectra; two-step greedy; singular spectral analysis; peak fitting.

1. Introduction
Silicate accounts for more than 90% contents of the earth's crust [1]. It is an important industrial raw material [2]. Solid-state 29Si nuclear magnetic resonance (NMR) is often used in the detection of silicate molecular structure, thereby studying its physical properties [3]. However, 29Si nuclear magnetic resonance spectra (NMRS) can contain noise and affect the spectral analysis due to the interference of different elements in the silicate sample as well as the changes in the environment and detection parameters of 29Si NMR equipment [4,5].
Peak fitting method is commonly used in order to determine the location and content of $^{29}$Si NMRS characteristic peak. The determination of peak position and peak profile model is the key to peak fitting. The characteristic peak position can be determined by statistical methods such as Discrete Wavelet Transform (DWT) [6], Empirical Mode Decomposition (EMD) [7] and SSA [8]. Wavelet transform originated from Fourier transform is an important harmonic analysis method. When DWT is applied, the wavelet basis function and the number of decomposed layers should be given first. Different wavelet basis functions and number of decomposed layers have great influence on the effect of the wavelet transform, thereby bringing inconvenience to the application of wavelet transform [9]. The EMD method was proposed by Huang et al. [7] in 1998 to analyze nonlinear non-stationary sequences. If the spectral diagram contains abnormal interference components when EMD is applied, namely extreme value envelope contains deterministic component envelope and noise component envelope, it is easy to cause ‘mode confusion’, thereby affecting characteristic peak determination [10].

SSA is a module-free method [11] with Singular Value Decomposition (SVD) as the core algorithm. It has strong recognition ability for feature components and good adaptability for continuous and intermittent data. It can be used for feature component extraction, trend analysis and smoothing/filtering processing of ordered sequences [12]. Alipanahi et al proposed the SVD-based peak acquisition and noise reduction methods [13] aiming at nuclear magnetic resonance spectra. Ye et al. proposed a local Capon calculator based on SVD for identifying tight spacing peaks in NMRS in the case of unknown peak numbers [14]. De et al. used SSA to remove solvent artifacts in multidimensional nuclear magnetic resonance spectra [8].

Embedding dimension $m$ is the key parameter for constructing trajectory matrix of SSA[15]. There are two methods to determine $m$: one is separate determination of $m$, such as correlation dimension (GP) and False Nearest Neighbors (FNN), etc. The other is synchronous determination with delay time $\tau$, such as C-C method [15]. Elsner et al. [16] suggests that the window length should less than $N/2(N$, spectral data length). C-C is invalidated because $\tau$ is set to be 1 in SSA. GP is vulnerable to a system error influence [17], FNN calculation results is big [18], the judgement threshold value of embedding dimension in SVD can not be determined easily. Therefore, TSG is proposed in the paper to determine the embedding dimension.

Spectral contour model is the basis of spectral quantitative analysis. Gaussian model and Lorentz model are two commonly used models in molecular spectroscopy [19]. The shape of spectral line is influenced by Doppler broadening, collision broadening, instrument parameters and other factors. The spectral line will be Gaussian if Doppler broadening is the main factor determining band width. The spectral line will be Lorentz if collision broadening is the main factor [20].C.Morrison et al. compared the fitting effect of Gaussian model and Lorentz model to NMRS by taking relative error as the evaluation criterion, and they found that Gaussian model had good consistency with an average relative error of 3.2%, and Lorentz is 3.5%[19].

A peak fitting method was proposed based on the statistical method and spectral contour model for peak fitting of $^{29}$Si NMRS: TSG-SSA-GFM.

2. Establishment of TSG- SSA- GFM fitting model

(1) **TSG:** in view of the disadvantage of high calculated quantity in greedy method. Two-step greedy method is proposed in the paper, namely TSG, thereby reducing the calculated quantity. The specific steps are shown as follows:

(a) Determine $m$ as $[2, m_{max}]$. Set $^{29}$Si NMRS as $Y=(y_1, y_2, y_3, \ldots , y_N)\ N$ as length, $m_{max}$ as $[N/2]$;

(b) Divide $[2, m_{max}]$ into $q$ subsets, set the $k$th subset as $[m_{k1}, m_{k2}](k=1,2,3 \ldots q)$, subscript the first digit $k$ representing the $k$th subset; the second digit ‘1’ and ‘2’ respectively represent the minimum and maximum embedding dimensions of the subset.

(c) Determine local optimal embedding dimension $m_{loc}$, experience $[m_{11}, m_{21}, m_{31}, \ldots , m_{q1}]$, calculate characteristic peak by SSA, and determine $m_{loc}$ ($m_{loc}=m_{k1}$).

(d) Determine the optimal embedding dimension $m_{opt}$, apply SSA to calculate characteristic peak of $^{29}$Si NMRS under $[m_{(k-2)1}, \ldots , m_{(k-2)2}, \ldots , m_{(k+2)1}, \ldots , m_{(k+2)2}]$ in the $(k-2)$ to $(k+2)$ subset, and determine $m_{opt}$.
(2) SSA: SSA includes the decomposition process and reconstruction process. Decomposition process: $^{29}$Si NMRS is mapped into a track matrix, and the singular value is used for decomposing the matrix. Reconstruction process: components with the same characteristics after decomposition are divided into characteristic peak or noise components, and then the components after division are converted into new ordered series $S_i$. Detailed steps are described in literature [21].

(3) Gaussian model: Gaussian functions are shown in many forms, while FWHM is a common form as shown in formula (1) [22].

$$y = \frac{2A}{w_G} \sqrt{\ln 2} \exp \left[ -\left( \frac{x - x_c}{w_G / 2\sqrt{\ln 2}} \right)^2 \right]$$  \hspace{1cm} (1)

Wherein, $A$ is the peak area, $w_G$ is half height width, and $x_c$ is the peak position.

Evaluation method of peak error: The mean deviation ($MD$) is introduced to evaluate the characteristic peak accuracy as shown in formula (2).

$$MD = \frac{1}{N} \sum_{i=1}^{N} |z_{\text{cal,}i} - z_{\text{ref,}i}|$$  \hspace{1cm} (2)

Wherein, $z_{\text{ref,}i}$ and $z_{\text{cal,}i}$ are respectively literature marking peak position and calculation peak position of the $i$th characteristic peak in the spectrogram.

3. Results and discussion

3.1. Data sources

$^{29}$Si NMRS containing different levels of noise is selected as shown in Figure 1 in order to study and determine the accuracy of $m$ and characteristic peak methods.

![Fig. 1 $^{29}$Si Nuclear Magnetic Resonance spectra from Ref.](image)

Note: ① Intensity is dimensionless; ② Fig. 1(1)～Fig. 1(4) from Ref. [23], while Fig. 1(5)～Fig. 1(6) from Ref. [2].

3.2. Accuracy of TSG, GP and FNN in determination of embedding dimension
(1) Determination method of characteristic peak spectrogram: Since SSA is required to verify the accuracy of TSG, GP and FNN in determining $m$, the method for SSA to determine characteristic peak spectrogram is firstly explained. Figure 1(1) is adopted as an example. After $m_{\text{opt}}$ is determined, $S_i$ spectrogram is calculated with SSA as shown in Figure 2.
Fig. 2 Spectrum of different components $S_i$

Note: ①Intensity is dimensionless; ②RC1, RC2, ..., RC6 represents the spectrogram of $S_1, S_2, ..., S_6$ of Fig. 1(1).

“◆” refers to the characteristic peak of $S_i$ spectrum of Figure 1(1). The peak position is -72.0, -80.2, -88.6, -97.1 and -106.4 ppm according to the calculation results of Figure 2(RC2), while the original spectrum is -72.1, -80.2, -88.6, -97.1 and -106.4 ppm, and the mean deviation is 0.02 ppm. Therefore, the characteristic peak position of $S_2$ spectrogram can represent the characteristic peak position of Figure 1(1).

(2) **Determination of embedding dimension with TSG;** 1) Determine the embedding dimension $m_{loc}$. Figure 1(1) is adopted as an example, $m$ is defined as 2, 7, 12, ..., 102. SSA is used for calculating $S_2$ spectrogram, and the results are shown in Figure 3.

Fig. 3 Spectra of $S_2$ composition when determining $m_{loc}$

Note: ①Intensity is dimensionless; ②Fig. 3(1)～Fig. 3(6) respectively show the $S_2$ spectra when $m$ is 32, 37, 42, 47, 52 and 57. ③Noise peaks are located within the red circle.

When $m$ is small, $S_2$ spectrogram contains noise as shown in red circle of Figure 3. The noise peak gradually decreases with the increase of $m$. The noise peak disappears when $m$ is 57, namely $m_{loc}$ is 57.

2) Determine the embedding dimension $m_{opt}$. $m$ is defined as 47, 48, 49, ..., 67 and SSA is used for calculating $S_2$ spectrogram on the basis of $m_{loc}$ determination, and the results are shown in Figure 4.

Fig. 4 Spectra of $S_2$ composition when determining $m_{opt}$
Note: ① Intensity is dimensionless; ②Fig. 4(1)～Fig. 4(6) respectively show the $S_2$ spectrum when $m$ is 52～57.

Figure 3 and Figure 4 show the trend that $m$ increase is beneficial for distinguishing characteristic peak and noise components in $^{29}\text{Si}$ NMRS. The data in space become more dispersed due to constant data of $^{29}\text{Si}$ NMRS with the increase of $m$, which is conducive to the SSA in distinguishing characteristic peak and noise components. However, if $m$ is too large, there will be ‘dimensionality disasters’ [24], thereby leading to failure of the deterministic peak. Therefore, $m$ increase is beneficial for SSA to distinguish different components in a certain range. The greedy method applies to the monotonous process of change.

(3) **Embedding dimension determination with GP:** In the paper, the minimum value and maximum value of the distance between phase points are taken as the lower limit and upper limit of $r$ value, and the step length is uniformly increased [25]. \( \ln(C(m,r)) \) and \( \ln(r) \) are calculated by GP in Figure 1. The results are shown in Figure 5.

![Fig. 5 \( \ln(C(m,r)) \) and \( \ln(r) \) under different $m$](image)

Note: ① Intensity is dimensionless; ② Colors of curves represent different $m$.

The relation of correlation dimension $D_m$ and $m$ is obtained by using the three-line method [26] as shown in Figure 6 aiming at Figure 5.

![Fig. 6 The relation between $D_m$ and $m$](image)

Figure 6 shows that $D_m$ tends to be stable, namely $m_{\text{opt}}$ is obtained when $m$ reaches 222, 217, 207, 232, 112 and 67, and then $S_2$ spectrogram is calculated with SSA, and the result is shown in Figure 7.

![Fig. 7 $S_2$ spectra determined by the combined GP and SSA](image)
Note: Intensity is dimensionless.

Figure 7 shows that SSA fails to determine characteristic peak. Relative errors of GP are respectively 344.0%, 393.2%, 430.8%, 314.3%, 387.0% and 235.0% compared with \( m \) determined by TSG. The average relative error is 350.7%. The reason for high relative error of GP possibly lies in that the determination of the scale-free interval by the three-line method is not objective, and the saturation of slope requires subjective judgment [25].

(3) **Embedding dimension determination with FNN**: The method proposed by Kennel [27] is adopted to identify false nearest neighbor points in order to avoid subjective factor interference. The data point is fully open when the proportion of false nearest neighbor points is reduced to 0, wherein \( m \) is \( m_{opt} \). FNN calculation results are shown in Figure 8 aiming at Figure 1.

![Fig. 8 The relation between \( m \) and the percentage of FNN](image)

Figure 8 shows that \( m \) is 167, 162, 160, 130, 159 and 153 respectively when the proportion of FNN is decreased to 0. SSA is used to determine \( S_2 \) spectrogram, and the results are shown in Figure 9.

![Fig. 9 \( S_2 \) spectra determined by the combined FNN and SSA](image)

Note: Intensity is dimensionless.

Figure 9 shows that SSA fails to determine characteristic peak, \( m_{opt} \) determined by FNN has errors, the errors are respectively 234.0%, 268.2%, 310.3%, 132.1%, 591.3% and 665.0% compared with TSG with an average relative error of 366.8%. \( m_{opt} \) determined by FNN is too large as a result possibly because of irrational judgment threshold of false nearest neighbor points or scale threshold value of false nearest neighbor points [18, 28].

3.3. **Application of TSG - SSA - GFM**

Gaussian model can better fit the characteristic peak[19] aiming at NMRS characteristic peak. Therefore, Gaussian model is used for fitting the characteristic peak, and TSG-SSA-GFM is formed by combining TSG and SSA. The method is utilized for peak fitting of \( ^{29}\text{Si} \) NMRS. The result is shown in Figure 10.
Fig. 10 Peak fitting of $^{29}$Si NMRS by TSG-SSA-GFM

Note: ① The spectrogram is from Ref. [2]. ② The black line is the original spectrum line, the red line is the spectrum line after the peak fitting, the pink line, the blue line and the green line are the fitting characteristic peak respectively.

Characteristic peak position and goodness of fit are shown in Table 1.

| No. | Q$^2$/ppm | Q$^3$/ppm | Q$^4$/ppm | MD/ppm | R$^2$/% |
|-----|-----------|-----------|-----------|--------|---------|
| 11  | -91.78    | -101.16   | -110.41   | 0.06   | 99.43   |
| 12  | -91.64    | -101.13   | -110.52   |        |         |
| 21  | -90.90    | -99.85    | -109.38   | 0.07   | 99.86   |
| 22  | -90.82    | -99.77    | -108.44   |        |         |
| 31  | -91.45    | -100.98   | -109.64   | 0.10   | 99.83   |
| 32  | -91.26    | -100.91   | -109.61   |        |         |
| 41  | -91.30    | -101.64   | -110.70   | 0.12   | 99.86   |
| 42  | -91.16    | -101.53   | -110.60   |        |         |

Note: in the serial number, the first number is the number of the graph, and the second number is the method for determining peak position, ‘1’ represents the marked value from the literature, and ‘2’ represents the peaks determined by TSG-SSA-GFM.

The mean deviation of TSG-SSA in determining characteristic peak is 0.09 ppm and the goodness of fit is 99.75% according to Table 1. Therefore, the method can be used for peak fitting of $^{29}$Si NMRS with noise.

4. Conclusion
A peak fitting method (TSG-SSA-GFM) is proposed in the paper in order to quantitatively analyze $^{29}$Si NMRS. The characteristic peak position can be accurately determined, and $^{29}$Si NMRS can be fit with the method, thereby realizing peak fitting of $^{29}$Si NMRS. It has active role to study properties of silicate in mineralogy, geochemistry, silicate materials and other disciplines. However, it is not easy to automatically identify characteristic peak due to the different strengths of characteristic peak and the interference of noise. Further research is needed in the aspect.

References
[1] Coey.J. Mössbauer Spectroscopy of Silicate Minerals[M]. Springer, 1984: 443-444.
[2] Halasz,I.;Kierys,A.;Goworek,J.;Liu,H.;Patterson,R.E.29Si NMR and Raman Glimpses into the Molecular Structures of Acid and Base Set Silica Gels Obtained from TEOS and Na-Silicate[J]. Journal of Physical Chemistry C, 2015, 115(2011): 24788-24799.
[3] Tognonvia,M.T.;Massiotb,D.;Lecomtec,A.;Rossignola,S.;Bonnetea,J. Identification of solvated species present in concentrated and dilute sodium silicate solutions by combined 29Si NMR
and SAXS studies[J]. Journal of Colloid & Interface Science, 2010, 352(2): 309-315.

[4] Sherriff BL,Hartman JS. Solid-state High-resolution 29Si NMR of Feldspars: Al-Si Disorder and the Effects of Paramagnetic Centres[J]. Canadian Mineralogist, 1985, 23: 205-212.

[5] Engelhardt G,Jancke H,Lippmaa E, et al. Structure Investigations of Solid Organosilicon Polymers By High Resolution Solid State 29Si NMR[J]. Journal of Organometallic Chemistry, 1981, 210: 295-301.

[6] Kim,D.;Won,Y.;Won,H. Noise Suppression in NMR Spectrum by Using Wavelet Transform Analysis [J]. Journal of the Korean Magnetic Resonance Society, 2000, 4(2): 103-115.

[7] Ghanati,R.;Hafizi,M.;Fallahsafari,M. Surface nuclear magnetic resonance signals recovery by integration of a non-linear decomposition method with statistical analysis[J]. Geophysical Prospecting, 2016, 64(2): 489-504.

[8] De,S.;Malloni,W.;Kremer,W.;Tomă,A.;Lang,E. Singular spectrum analysis for an automated solvent artifact removal and baseline correction of 1D NMR spectra[J]. Journal of Magnetic Resonance, 2011, 2(210): 177-183.

[9] Xiao Xiaobing; Liu Hongli; Ma Ziji. An empirical mode decomposition de-noising method based on singular spectrum analysis [J]. Computer Engineering and Science, 2007, 5:919-924.

[10] Dezfouli,M.P.;Dezfouli,M.;Rad,H. A Novel Approach for Baseline Correction in 1H-MRS Signals Based on Ensemble Empirical Mode Decomposition[C].Engineering in Medicine & Biology Society, 2014: 3196-3199.

[11] Golyandina,N.;Zhigljavsky,A. Singular Spectrum Analysis for Time Series[M]. Springer, 2010.

[12] Hassani H. Singular Spectrum Analysis Methodology and Comparison[J]. Journal of Data Science, 2007, 5(2): 239-257.

[13] Alipanahi,B.;Gao,X.;Karacak,E.;Donaldson,L.;Li,M. PICKY:a novel SVD-based NMR spectra peak picking method[J]. Bioinformatics, 2009, 25(12): 75-268.

[14] Ye,S.;Aboutanios,E.;Thomas,D.;Hook,J. Localised high resolution spectral estimator for resolving superimposed peaks in NMR signals[J]. Signal Processing, 2017, 130(C): 343-354.

[15] Hong-Guang,M.;Han,C. Selection of Embedding Dimension and Delay Time in Phase Space Reconstruction[J]. Frontiers of Electrical and Electronic Engineering in China, 2006, 1(1): 111-114.

[16] Elsner J. Analysis of Time Series Structure: SSA and Related Techniques[M]. United States of America: Chapman & Hall/CRC, 2001.

[17] Kennel,M.;Abarbanel,H. False neighbors and false strands: A reliable minimum embedding dimension algorithm[J]. Physical Review E, 2002, 66(2): 26209.

[18] Abarbanel,H.;Kennel,M. Local false nearest neighbors and dynamical dimensions from observed chaotic data[J]. Physical Review E Statistical Physics Plasmas Fluids & Related Interdisciplinary Topics, 1993, 47(5): 3057.

[19] Morrison,C.;Henkelman,R. A Model for Magnetization Transfer in Tissues[J]. Magnetic Resonance in Medicine, 2010, 33(4): 475-482.

[20] Petrakis,L. Spectral line shapes: Gaussian and Lorentzian functions in magnetic resonance[J]. Journal of Chemical Education, 1967, 44(8): 432-436.

[21] Hu,B.X.;Li,Q.M.;Smith,A. Noise reduction of hyperspectral data using singular spectral analysis[J]. International Journal of Remote Sensing, 2009, 30(9): 2277-2296.

[22] Stancic,A.L.;Brauns,E.B. A simple asymmetric lineshape for fitting infrared absorption spectra[J]. Vibrational Spectroscopy, 2008, 47(1): 66-69.

[23] Vidal,L.;Joussein,E.;Colas,M.;Cornette,J.;Sanz,J. Controlling the reactivity of silicate solutions- A FTIR, Raman and NMR study[J]. Colloids & Surfaces a Physicochemical & Engineering Aspects, 2016, 503(5): 101-109.

[24] Verleysen,M.;François,D. The Curse of Dimensionality in Data Mining and Time Series Prediction[C].Computational Intelligence and Bioinspired Systems, 2005: 758-770.

[25] Gao Junjie; Wang Hao. Determination of embedding dimension for phase space based on improved G-P method[J]. Computer Engineering and Applications, 2014, 50 (9) : 107-110.
[26] BQ Du, GF Deng, JJ Shi. Improved Algorithm of Scaleless Band Identification on Correlation Dimension of Vibration Signal[C]. Asia-pacific Power & Energy Engineering Conference, 2009, 1-4.

[27] MB, Kennel; R, Brown; HD, Abarbanel. A. Determining embedding dimension for phase-space reconstruction using a geometrical construction[J]. Physical Review A, 1992, 45(6): 3403-3411.

[28] Rao Guoqiang; Feng Fuzhou, Si Aiwei, Xie Jinliang. Method for optimal determination of parameters in permutation entropy algorithm. Journal of Vibration and Shock, 2014, 33(1): 188-193.