Earth, Planets and Space

FULL PAPER Open Access

Improved shift-invariant sparse coding for noise attenuation of magnetotelluric data

Guang Li[1,2], Xiaoqiong Liu[3], Jingtian Tang[2], Juzhi Deng[1], Shuanggui Hu[2*], Cong Zhou[1,2*], Chaojian Chen[4] and Wenwu Tang[1]

Abstract

Magnetotelluric (MT) method is widely used for revealing deep electrical structure. However, natural MT signals are susceptible to cultural noises. In particular, the existing data-processing methods usually fail to work when MT data are contaminated by persistent or coherent noises. To improve the quality of MT data collected with strong ambient noises, we propose a novel time-series editing method based on the improved shift-invariant sparse coding (ISISC), a data-driven machine learning algorithm. First, a redundant dictionary is learned autonomously from the raw MT data. Second, cultural noises are reconstructed using the learned dictionary and the orthogonal matching pursuit (OMP) algorithm. Finally, the de-noised MT data are obtained by subtracting the reconstructed cultural noises from the raw MT data. The synthetic data, field experimental data and measured data are tested to verify the effectiveness of the newly proposed method. The results show that our new scheme can effectively remove strong cultural noises and has better adaptability and efficiency than the predefined dictionary-based methods. The method can be used as an alternative when a remote reference station is not available.

Keywords: Magnetotelluric, Machine learning, Shift-invariant sparse coding (SISC), De-noising, Dictionary learning

Introduction

The magnetotelluric (MT) method employs natural electromagnetic fields as sources, reaching a prospecting depth of 600 km or more (Simpson and Bahr 2005; Garcia et al. 2015). It is a popular and indispensable method for deep mineral resources exploration and electrical conductivity structure probing of the Earth (Guo et al. 2019). However, natural MT signals are weak, non-stationary and broadband in frequency, and therefore susceptible to cultural noises (Cai et al. 2009; Neukirch and Garcia 2014). As the proportion of urbanization continues to increase, the distribution of cultural noises is becoming wider and wider, which severely limits the application of the MT method.

*Correspondence: hushuanggui808@csu.edu.cn; zhoucong_522@163.com
[1] State Key Laboratory of Nuclear Resources and Environment, East China University of Technology, Nanchang 330013, China
[2] Key Laboratory of Metallogenic Prediction of Non-Ferrous Metals and Geological Environment Monitor, Ministry of Education, Central South University, Changsha 410083, China
Full list of author information is available at the end of the article

Methods to deal with MT noises are mainly include remote reference (RR) technique (Gamble et al. 1979), robust statistic estimate (Egbert 1986, 1997; Garcia and Jones 2008) and signal–noise separation in time domain (Trad and Travassos 2000; Neukirch and Garcia 2014). With the development of urbanization, the construction of remote reference stations becomes more and more difficult. The remote reference technique often fails in noisy environments because it only works if the noise is incoherent between local and remote reference stations. This assumption is increasingly violated due to the large-scale construction of infrastructure (rail networks, pipelines, power systems, industrial and mining enterprises, etc.). Robust statistic estimate method requires that most of the data should be noise-free, while some types of noise are persistent and appear during the entire observation process. In this case, the robust statistic method may lead to worse results (Escalas et al. 2013; Campanya et al. 2014; Larnier et al. 2016; Tang et al. 2018).

By removing noises in time domain, the time-series editing methods can directly and effectively improve

the quality of MT data (Trad and travassos 2000; Tang et al. 2013, 2018; Neukirch and Garcia 2014; Larnier et al. 2016). The most representative time-series editing method is the wavelet transform-based scheme. Mathematical morphological filtering, empirical mode decomposition and combinations between them are also used for MT time-series editing (Cai 2016; Li et al. 2017). However, the time domain signal–noise separation methods mentioned above have the risk of losing effective signals, especially the low-frequency signal (Li et al. 2020). In other words, no matter the time-series segment being processed is clean or noisy, some components will be removed if using the methods mentioned above (Li et al. 2018).

To solve this problem, sparse representation was applied to MT signal processing (Tang et al. 2017; Li et al. 2017). By designing a redundant dictionary (or called over-complete dictionary, see details later) that matches with cultural noises but is not sensitive to useful MT signals, cultural noises can be effectively removed while retaining useful signals. However, cultural noises in the measured data are complex and diverse. Predesigned redundant dictionaries have obvious limitations. For instance, a simple redundant dictionary can only effectively deal with a certain type or a few types of noises. Besides, a complex redundant dictionary is time consuming because the number of atoms (each column in the dictionary is an atom) in the dictionary is too large.

Olshausen and Field (1996) used self-learned dictionaries for sparse representations of natural images, and since then self-learned dictionaries have received widespread attention. Shift-invariant sparse coding (SISC) is a data-driven machine learning algorithm (Blumensath and Davies 2005, 2006). It learns the feature structures (the so-called redundant dictionary) autonomously from a given sample set and uses convolution as a shift operator to conveniently represent multiple features with one atom. In other words, SISC can acquire the regular pattern of the signal autonomously. Currently, SISC has been modified several times and is successfully applied to heartbeat signal processing (Blumensath and Davies 2005), speech signal processing (Plumbley et al. 2006), and mechanical fault feature extraction (Liu et al. 2011; Zhu et al. 2016). In this paper, we extend the improved shift-invariant sparse coding (ISISC; Zhu et al. 2015; Wang et al. 2015) to noise attenuation of MT data and attempt to obtain better flexibility and efficiency by replacing predesigned redundant dictionary with self-learned redundant dictionary.

The rest of this paper is organized as follows. Section "Improved shift-invariant sparse coding" gives the theory of ISISC; Sect. "Synthetic case studies" presents the analysis of synthetic data; field experimental data and measured MT data are studied in Sect. "Real case studies"; the conclusions will be given in the final section.

Improved shift-invariant sparse coding

The shift-invariant sparse coding model

In the traditional model of sparse representation (Mallat and Zhang 1993), a signal or image is represented as a linear combination of the redundant dictionary and coefficients. Thus similar or identical feature structures at different locations in the time series require multiple atoms to represent. In addition, the redundant dictionary in the traditional sparse representation is manually predefined. In actual demand, the predefined dictionary is not flexible enough for complex signals (Jafari and Plumbley 2011; Chen 2017).

Shift-invariant sparse coding is a machine learning algorithm based on data-driven framework. It employs convolution as a shift operator to satisfy the property of shift-invariant, and represents the signal as a convolution of the dictionary and the coefficients. This allows feature atoms to be translated, flipped, and scaled anywhere in the time series, thereby facilitating the use of one atom to conveniently represent multiple features at different locations. What is more, the redundant dictionary in SISC is learned from the raw time series adaptively. In other words, SISC is able to learn the laws of signals autonomously and is effective for all kinds of morphological features, which is highly suitable for the processing of complex time series.

For the discrete signal set $Y = [y_1, y_2, \ldots y_k]^T$, shift-invariant sparse coding expresses $\mathbf{y}_k$ as the sum of convolutions of the atoms $\mathbf{d}_m$ and sparse coding coefficients $\mathbf{s}_{m,k}$:

$$\mathbf{y}_k = \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{s}_{m,k} + \boldsymbol{\varepsilon}, \tag{1}$$

where $Y_k = [y_1, y_2, \cdots y_N]^T$ is a time-series segment with N sampling points. $D = [d_1, d_2, \cdots d_M]^T \in R^{Q\times M}$ is the so-called redundant dictionary, or called over-complete dictionary, in which $\mathbf{d}_m$ is an atom in dictionary $\mathbf{D}$ and the number of atoms M is larger than N. In most cases the number of atoms M is much larger than N. In other words, signal $\mathbf{y}_k$ can be represented using the dictionary $\mathbf{D}$ in many ways. This is why dictionary $\mathbf{D}$ is called a redundant dictionary. $\boldsymbol{\varepsilon}$ stands for Gaussian white noise. "$*$" denotes the operation of convolution. $\mathbf{s}_{m,k} \in R^P$ and $\mathbf{s}_{m,k}$ is sparse (most of the elements in $\mathbf{s}_{m,k}$ are zero). $Q<N$, $P<N$ and $Q+P-1=N$.

As shown in Eq. (1), both the atom $\mathbf{d}_m$ and the coefficient $\mathbf{s}_{m,k}$ in the model of SISC are unknown. The optimization problem is non-convex and difficult to obtain a stable solution if $\mathbf{d}_m$ and $\mathbf{s}_{m,k}$ are obtained

simultaneously. Many scholars solved this problem by turning it into a convex optimization problem. The atom $\mathbf{d}_m$ and coefficients $\mathbf{s}_{m,k}$ are updated alternately in their schemes (Smith and Lewicki 2006; Plumbley et al. 2006; Aharon et al. 2006). When $\mathbf{d}_m$ is known, $\mathbf{s}_{m,k}$ can be obtained based on the convex optimization method; correspondingly, when $\mathbf{s}_{m,k}$ is constant, $\mathbf{d}_m$ can be solved based on the convex optimization method. Sparsity is the common goal of the above two optimization problems, and the cost function for evaluating the sparsity of $\mathbf{y}_k$ is (Liu et al., 2011; Wang et al., 2015):

$$\psi(\theta) = \min_{\mathbf{d},\mathbf{s}} \sum_{k=1}^{K} \left\| \mathbf{y}_k - \sum_{m=1}^{M} \mathbf{d}_m * \mathbf{s}_{m,k} \right\|_2^2 + \beta \cdot \sum_{m,k} \left\| \mathbf{s}_{m,k} \right\|_1, \tag{2}$$

where $\|.\|_F$ represents the F-order norm, β is a parameter of constraint used to balance reconstruction error and sparsity. The learned atom $\mathbf{d}_m$ usually needs to be normalized, i.e., $\left\|d_m\right\|_2^2 = 1$.

Solving the sparse coding coefficients

Keeping the dictionary unchanged, the sparse representation coefficients can be obtained by matching pursuit (MP) algorithm (Mallat and Zhang 1993) or orthogonal matching pursuit (OMP) algorithm (Pati et al. 1993). OMP is improved from MP by adding the step of orthogonalization. We use OMP to solve the coding coefficients because it has a better characteristic of convergence.

Imaging $\mathbf{y}_k$ is the signal to be represented, $\mathbf{g}_{i,u}$ is the atom obtained by shifting u points from the learned feature structure, its length is the same as signal $\mathbf{y}_k$, and $||\ \mathbf{g}_{i,u}\ ||=1$. L denotes the number of iterations, $\mathbf{r}^L$ represents the residual after the Lth iteration, $\boldsymbol{\psi}_L$ represents the selected set of atoms after the Lth iteration. The steps of the OMP algorithm are as follows:

1. Initializations, $\mathbf{r}^0 = \mathbf{y}_k$, $\boldsymbol{\psi}_0 = \emptyset$, $L = 1$;
2. Pursuit of the atom $\mathbf{g}_{i,u}$ that satisfies the following equation:

$$\left|\left\langle \mathbf{r}^L, \mathbf{g}_{i,u}^L \right\rangle\right| = \sup_{1\le i\le Q} \left(\sup_{0\le u\le P} \left|\left\langle \mathbf{r}^L, \mathbf{g}_{i,u}^L \right\rangle\right| \right); \tag{3}$$

3. Update the selected set of atoms, $\boldsymbol{\psi}_L = \boldsymbol{\psi}_{L-1} \cup \{\mathbf{g}_{i,u}^L\}$;
4. Calculate the projection coefficients according to least squares method $\mathbf{s}_L = (\tilde{\Psi}_L^T \Psi_L)^{-1} \cdot \tilde{\Psi}_L^T \tilde{y}_k$, and subsequently, residual $\mathbf{r}^L = \mathbf{y}_k - \mathbf{s}_L \Psi_L$, reconstructed signal $\hat{y}_k = \mathbf{s}_L \Psi_L$;
5. $L = L+1$ and return to step 2 if L is smaller than the maximum number of iterations. Otherwise, output the current reconstructed signal and the corresponding residual.

Dictionary learning

Make full use of the idea of K-SVD algorithm (Aharon et al. 2006), Wang et al. (2015) updated the atoms one by one, rather than all at once. They termed the new method ISISC and show that ISISC is superior to the gradient-based SISC (GSISC) in accuracy and efficiency. In the stage of dictionary learning, the atoms are updated while the coding coefficients stay unchanged. The optimization function can be simplified as:

$$\begin{aligned} \bar{\psi}(\theta) &= \min_{d} \sum_{i=1}^{N} \left\| y_k - \sum_{m=1}^{M} d_m * s_{m,k} \right\|_2^2 \\ &= \min_{d} \sum_{k=1}^{K} \left\| \left(y_k - \sum_{m\neq i}^{M} d_m * s_{m,k} \right) - d_i * s_{i,k} \right\|_2^2 \\ &= \min_{d} \sum_{k=1}^{K} \left\| E_{i,k} - d_i * s_{i,k} \right\|_2^2, \end{aligned} \tag{4}$$

where $\mathrm{E}_{i,k}$ represents the recovery error of all the atoms except the ith atom with respect to the kth signal. The update of the ith atom can be translated into solving an equation for $\mathbf{d}_i$. Since $\mathbf{d}_i * \mathbf{s}_{i,k} = \mathbf{s}_{i,k} * \mathbf{d}_i$, when only taking the kth signal into consideration, the optimization of Eq. (4) equals to the solution of the following equation (Zhu et al. 2016):

$$\begin{bmatrix} s_{i,k}^1 & & & \\ s_{i,k}^2 & s_{i,k}^1 & & \\ \vdots & s_{i,k}^2 & \ddots & \\ s_{i,k}^P & \vdots & \ddots & s_{i,k}^1 \\ & s_{i,k}^P & \ddots & s_{i,k}^2 \\ & & \ddots & \vdots \\ & & & s_{i,k}^P \end{bmatrix} \cdot \begin{bmatrix} d_m^1 \\ d_m^2 \\ \vdots \\ d_m^Q \end{bmatrix} = \begin{bmatrix} E_{i,k}^1 \\ E_{i,k}^2 \\ \vdots \\ E_{i,k}^N \end{bmatrix}. \tag{5}$$

Considering the matrix on the left side of Eq. (5) as a special Toeplitz matrix of coefficients $\mathbf{s}_{i,k}$, the above equation can be written as $Toep(\mathbf{s}_{i,k}) \cdot \mathbf{d}_i = \mathbf{E}_{i,k}$. Since the coefficient $\mathbf{s}_{i,k}$ is sparse, many of the row vectors in the matrix $\mathrm{Toep}(\mathbf{s}_{i,k})$ are 0 vectors, and these 0 vectors have no effect on the result. Remove these 0 vectors from $\mathrm{Toep}(\mathbf{s}_{i,k})$ and the corresponding row vectors from $\mathbf{E}_i$, then Eq. (5) can be written as $\mathrm{Toep}(\tilde{s}_{i,k}) \cdot \mathbf{d}_i = \widetilde{\mathbf{E}}_{i,k}$. When taking all K signals into consideration, the optimization function can be expressed as:

$$\begin{bmatrix} \mathrm{Toep}(\tilde{s}_{i,1}) \\ \mathrm{Toep}(\tilde{s}_{i,2}) \\ \vdots \\ \mathrm{Toep}(\tilde{s}_{i,K}) \end{bmatrix} \cdot \mathbf{d}_i = \begin{bmatrix} \tilde{E}_{i,1} \\ \tilde{E}_{i,2} \\ \vdots \\ \tilde{E}_{i,K} \end{bmatrix}. \tag{6}$$

Simplifying Eq. (6) to $\mathbf{S}\cdot\mathbf{d}_i=\mathbf{E}$, and according to the least squares method, the feature atom can be derived as $\mathbf{d}_i = (\mathbf{S}^T\mathbf{S})^{-1}(\mathbf{S}^T\mathbf{E})$. Since $(\mathbf{S}^T\mathbf{S}) \in \mathrm{R}^{Q\times Q}$ and in most cases, $Q \ll N$, the above equation can be transformed into a small-scale linear equation. The solution can be directly obtained by Cholesky decomposition, LU decomposition and other methods. Each atom is updated in a random order, and finally gets all the feature structures, i.e., feature atoms. The learned dictionary is obtained by normalizing all the feature atoms, i.e., $\mathbf{d}_i = \mathbf{d}_i/||\mathbf{d}_i||_2^2$.

The flow diagram of data processing

Here, we give the flow diagram of data processing for MT data:

Input: raw MT data **Y**.

Initialization: give the initial value to the dictionary **D** and the sparse representation coefficient **S** randomly, $z=0$.

Repeat the following:

{

$z=z+1$;

Solving the sparse coding coefficients;

Update the dictionary;

}

Until z reaches the maximum number of iterations.

Output: the learned dictionary **D**, sparse coding coefficients **S** and the reconstructed signal $\bar{Y}$.

The reconstructed signals are components with abnormally large amplitude or obvious regularity. These components are noise according to the characteristics of natural magnetotelluric signals. Therefore, the de-noised MT data can be obtained by subtracting the reconstructed signal from the original signal.

Synthetic case studies

Square-wave noise and impulsive noise are two common types of noises during natural field source electromagnetic exploration. These two types of noise are usually very large in amplitude and theoretically have an infinite number of harmonic components, thus affecting multiple frequency bands of the apparent resistivity and phase curves. It is very difficult to deal with these types of noise. The pseudo-random square-wave signal contains a variety of square-wave structures of different widths. The charge–discharge-like waveforms have both abrupt components similar to impulsive noises and stationary components similar to harmonics. In this section, pseudo-random square-wave noise and charge–discharge-like noises are used to verify the effectiveness and performance of different de-noising methods since they can simulate the complex noises in the real world.

Noisy MT data are obtained by adding simulated pseudo-random square-wave noises or charge–discharge-like noises to measured noise-free MT data. Signal-to-noise ratio (SNR), recovery error (E), normalized cross-correlation (NCC) and time consuming (T) are used to quantitatively evaluate the results of noise attenuation. SNR, E and NCC are defined as (Candès and Wakin 2008; Li et al. 2017; Zhang et al. 2019):

$$\mathrm{SNR} = 20\lg\frac{\|y(n)\|_2}{\|y(n)-r(n)\|_2}, \tag{7}$$

$$E = \frac{\|y(n)-r(n)\|_2}{\|y(n)\|_2}, \tag{8}$$

$$\mathrm{NCC} = \frac{\sum_{n=1}^{N} y(n)\cdot r(n)}{\sqrt{(\sum_{n=1}^{N} y^2(n))\cdot(\sum_{n=1}^{N} r^2(n))}}, \tag{9}$$

where $y(n)$ represents the original signal; $r(n)$ stands for the de-noised signal; N is the length of signal. All data processing work in this paper is completed on the laptop of ThinkPad W530 (CPU, i7-3610, 2.3 GHz; RAM, 8.00 GB).

Noise attenuation for square-wave noises

Figure 1a shows the original noise-free MT data, which are collected in Qaidam Basin using MTU (Phoenix Geophysics Ltd), *with a sampling rate of 15 Hz. Figure 1b is the synthetic noisy data which is obtained by adding simulated pseudo-random square-wave noise to noise-free MT data. Figure 1c is the de-noised signal using predefined square-wave dictionary (SD) based method (Tang et al. 2017; Li et al. 2017). In this method, an improved OMP (IOMP) algorithm is used to improve the efficiency. Please refer to related literatures for more details about the method. Figure 1d is the de-noised signal using our new method proposed in this paper. Figure 1 demonstrates that both methods completely eliminate noises. However, compared with the original spectrum, it is clear that the results obtained by the predefined dictionary-based method lose some of the low-frequency effective signals. In contrast, our new method yields no visible distortion.

As shown in Table 1, both predefined dictionary and learned dictionary-based methods have achieved good de-noising results; SNR, E and NCC are greatly

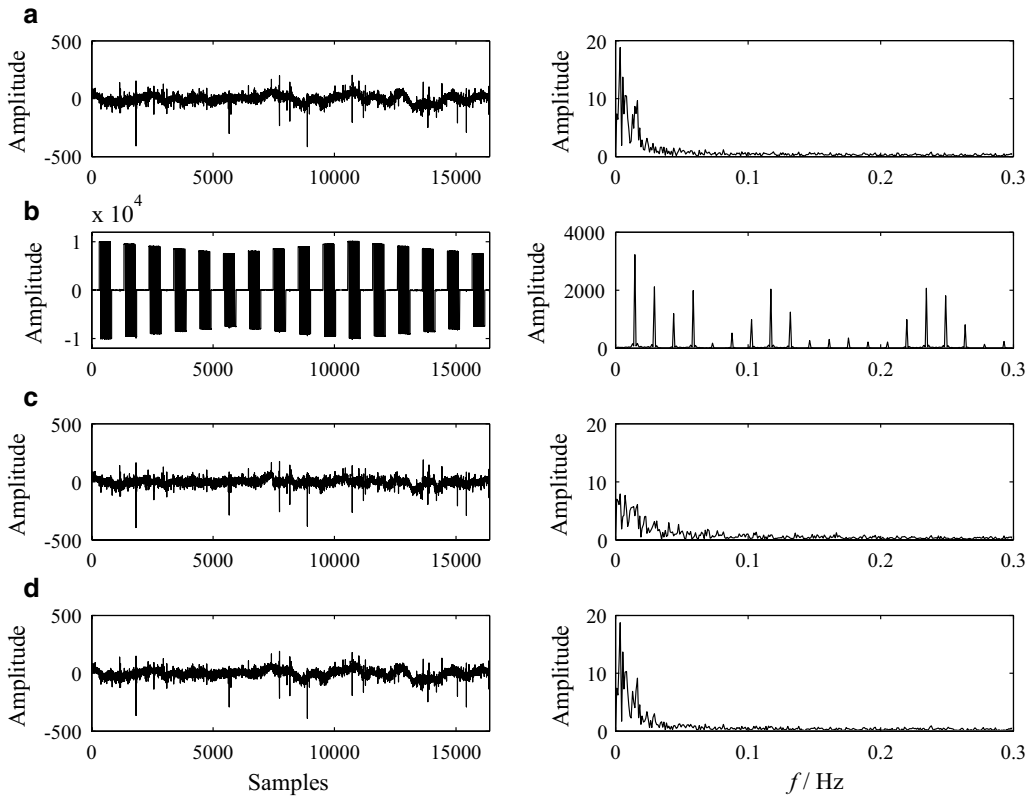

Fig. 1 Results of pseudo-random square-wave noise processing. **a** Original noise-free MT signal; **b** the synthetic noisy signal; **c** de-noised by square-wave dictionary-based method (IOMP-SD); **d** de-noised by ISISC. The left panels are time domain waveforms and the right panels are their spectra

Table 1 Quantitative evaluation of pseudo-random square-wave noise processing results

	NCC	*E*	SNR (dB)	*T* (s)
Before processing	0.0561	182.91	−45.2448	/
IOMP-SD	0.8035	0.5952	4.5057	278.7
ISISC	0.9777	0.2097	13.5657	3.7

improved over the noisy signal. However, result obtained by our method has a better signal-to-noise ratio, smaller recovery error and higher normalized cross-correlation. Especially the time consumption is much smaller than the predefined dictionary-based method.

Noise attenuation for charge–discharge-like noises

Figure 2a is the original noise-free MT data, which is collected in Qaidam Basin, with a sampling rate of 150 Hz. Figure 2b is the noisy MT signal which is obtained by adding simulated charge–discharge-like noises to noise-free MT data. Figure 2c is the de-noised signal using predefined pulse dictionary (PD, contains charge–discharge-like atoms) based method (Wang et al. 2013). In this method, the algorithm of particle swarm optimization (PSO) is used to improve the efficiency. Figure 2d is the de-noised signal using ISISC-based method proposed in this paper. Both predefined dictionary and learned dictionary-based methods effectively remove noises. However, the damage of low-frequency signal can be easily found from the spectrum obtained by the pulse dictionary-based method. In addition, the increase of the spectrum around 50 Hz indicates that pulse dictionary-based method generates some new noises.

As shown in Table 2, the pulse dictionary-based method also obtained good NCC, *E* and SNR because some atoms in the dictionary are very similar to the simulated charge–discharge-like noises. Nevertheless, the results obtained by pulse dictionary are far from satisfactory because it takes 440.2 s. The ISISC-based method adaptively learns the dictionary from the noisy data and is therefore not limited by the type of noise. The value of NCC obtained by ISISC surged up to 0.9827, recovery error decreased from 57.0588 to 0.1849, SNR increased from -35.1264 dB to 14.6597 dB, time consumption decreased from 440.2 s to 1.2 s. Obviously, the results

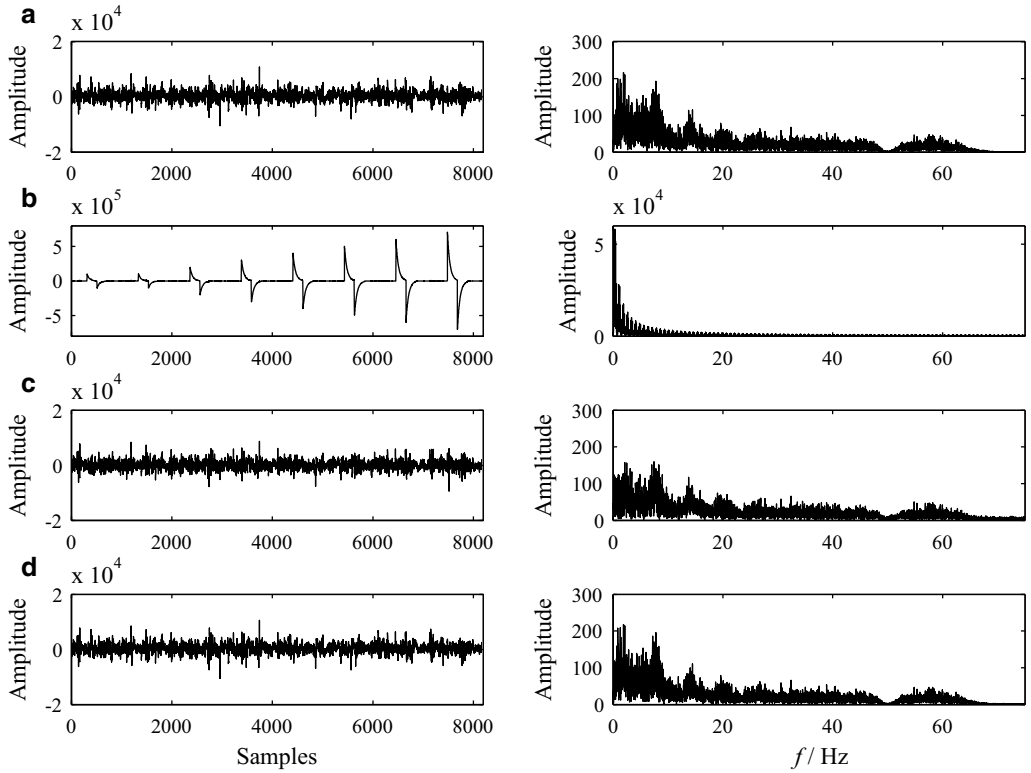

Fig. 2 Results of the charge–discharge-like noise processing. **a** Original noise-free MT signal; **b** synthetic noisy signal; **c** de-noised by pulse dictionary-based method (PSO-PD); **d** de-noised by ISISC. The left panels are time domain waveforms and the right panels are their spectra

Table 2 Quantitative evaluation of the charge–discharge-like noise processing

	NCC	*E*	SNR (dB)	*T* (s)
Before processing	0.0108	57.0588	−35.1264	/
PSO-PD	0.8254	0.5660	4.9437	440.2
ISISC	0.9827	0.1849	14.6597	1.2

achieved by ISISC-based method is significantly better than the predefined dictionary-based method.

Real case studies

Experimental data in Liangshan Prefecture, Sichuan Province

In order to verify the validity of the data-processing method and to evaluate the data processing results reasonably, we conducted a MT data observation experiment in Liangshan Prefecture, Sichuan Province, China, in 2014. The experimental instruments are the magnetotelluric data acquisition system MTU-5A (Phoenix Geophysics Ltd) and the controlled-source electromagnetic method (CSEM) transmitter GGT-30 (Zonge International INC.). The data obtained at this station is recorded as ASY0002B.

The experiment lasted for 100 min, in which the first 60 min were observed normally according to the standard procedure (the transmitter did not work), and the data observed during this time period was marked as D_1. In the next 40 min, the transmitter continuously transmits electric-dipole source signals with different frequencies. The maximum transmitting current was about 20 A. The length of the electric dipole was 1.2 km and the electric-dipole was parallel to the receiving dipole of x-direction. The transmitter was 1.2 km away from the observation station (the Rx–Tx distance was 1.2 km), and the data observed in this time period was marked as D_2. Since the vicinity of the observation site was sparsely populated and there was almost no electromagnetic noise, the D_1 data set is high-quality data. The apparent resistivity and phase curves obtained using this data set can be seen as the real geoelectric responses, so they can be used as a reference for comparison. Note that the controlled-source signals in our experiment are treated as noises because they are not plane wave. Therefore, the D_2 data set is noisy because

Table 3 Fundamental frequencies and duration of the controlled-source signal

Order	Frequency (Hz)	Duration (s)
1	0.125	640
2	0.175	220
3	0.25	320
4	0.35	390
5	0.5	220
6	0.7	150
7	1	260

it was contaminated by strong controlled-source signals.

As shown in Table 3, the transmitter sends 7 square-wave signals of different frequencies in sequence. Due to the filters in the acquisition device and the low-pass filtering effect of the earth system, the received signals are not ideal square-wave signals, but are similar to charge–discharge-like waveforms. The frequencies of the controlled-source signals range from 0.125 Hz to 1 Hz, however, both the square-wave and the charge–discharge-like signals have a large number of odd harmonics. Therefore, the frequency actually affected by the CSEM source is not limited to 0.125 Hz–1 Hz.

Figure 3 shows the MT time-series segments in the D_2 data set, the sampling rate is 150 Hz. Similar to the segments shown in Fig. 3, the entire D_2 data set was contaminated by CSEM source. These signals are standard CSEM noises in morphology and do not have the characteristics of natural magnetotelluric signals at all. When subjected to such persistent noise pollution, traditional MT data-processing methods are difficult to obtain good results. Observing the time series, it can be seen that the noise (CSEM source) is formed by shifting and flipping the same or similar structures, and thus has a characteristic of shift-invariant. These structures are perfect for feature extraction using SISC.

Figure 4 shows the feature structures that ISISC learns autonomously from the D_2 data set. These feature structures are zero-padded and then become atoms in the redundant dictionary, which can be used to accurately represent a particular type of signal. Since they are learned autonomously based on the signal to be processed, they match the CSEM noises very well. As shown in Fig. 4, the learned feature structure is neither an ideal square-wave signal nor an ideal charge–discharge signal, but a complex signal that is difficult to

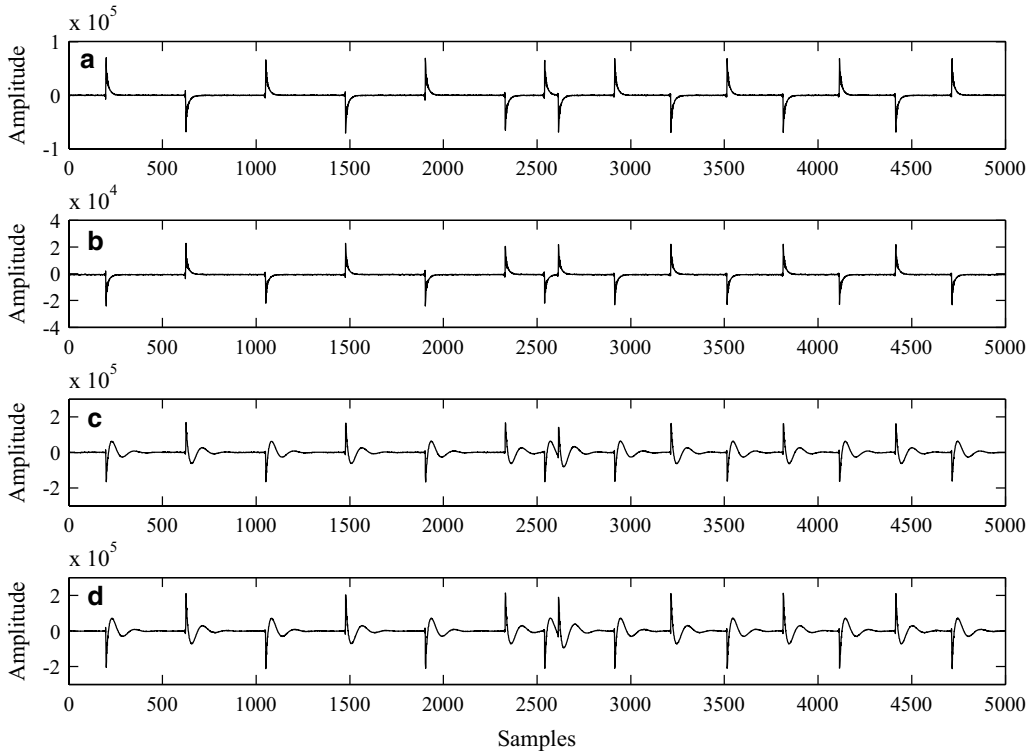

Fig. 3 Raw time-series segments from the data set D_2 of the station ASY0002B, a sampling rate of 150 Hz. **a** Ex component; **b** Ey component; **c** Hx component; **d** Hy component

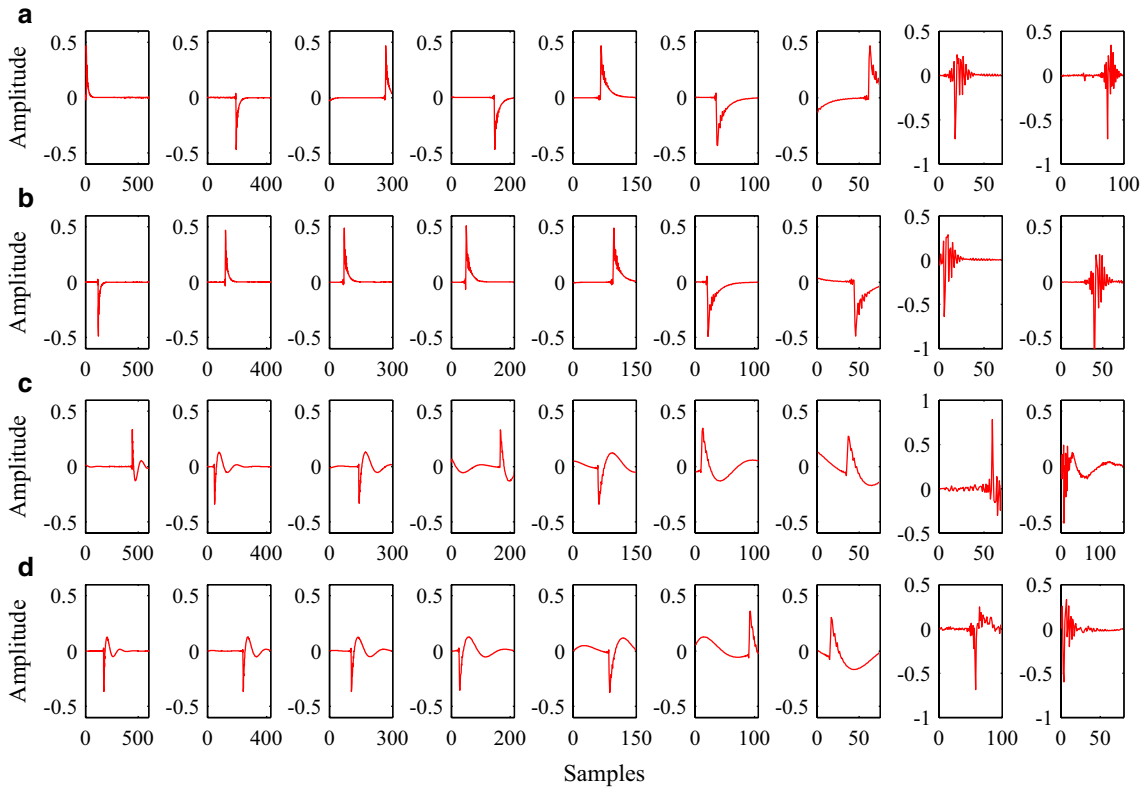

Fig. 4 Feature structures (atoms) learned from Ex (**a**), Ey (**b**), Hx (**c**) and Hy (**d**) components of the MT data set D_2 (a sampling rate of 150 Hz) using ISISC

describe. A regular zigzag waveform is produced on its rising and falling edges. As it is, the source of human noise in the measured MT signal is complex and has different forms. It is very difficult to design a predefined dictionary that exactly matches the complex and variable noise. Therefore, the predefined dictionary-based method is inevitable to reduce the decomposition accuracy or increase the time consumption.

As shown in Fig. 5, the magnetic components (Hx and Hy) obtained by PSO-PD have obvious differences between the first half and the second half. The first half of Hx and Hy components may lose some useful signals because natural MT signals usually contain some useful spikes (refer to the noise-free signals shown in Figs. 1a, 2a). The electric signals (Ex and Ey) and the second half of the magnetic signals indicate that some noise has not been removed by PSO-PD. The results obtained by ISISC are more reasonable since the noise is effectively removed and there is no obvious difference between the first and the second half.

As shown in Fig. 6, the apparent resistivity and phase curves calculated from the noise-free MT data set D_1 are continuous, smooth and vary slowly with the frequencies except for some little bias below 1 Hz. All values are in the normal range. Overall, it shows good data quality.

The apparent resistivity and phase curves obtained by raw D_1 and D_2 data sets are severely distorted below 300 Hz since the D_2 data set is contaminated by CSEM noises. There are a large number of outliers in both resistivity and phase curves, and the phase values in the xy component are mostly distributed in an unreasonable range. This result is significantly different from the apparent resistivity and phase curves calculated from the noise-free data set.

The processing of PSO-PD results in an obvious improvement over the previous. However, the MT response of xy components still reveals obvious distortion around 100 Hz and below 1 Hz. Moreover, visible bias can be found at yx components below 1 Hz.

By de-noising the D_2 data set with our new method and calculating the MT responses using data sets of D_1 and D_2, we get the curves shown as green triangles in Fig. 6. Obviously, most of the outliers disappear and the phase curves return to reasonable. The results obtained by our method are generally consistent with the results obtained from the D_1 data set except for some little bias below 1 Hz.

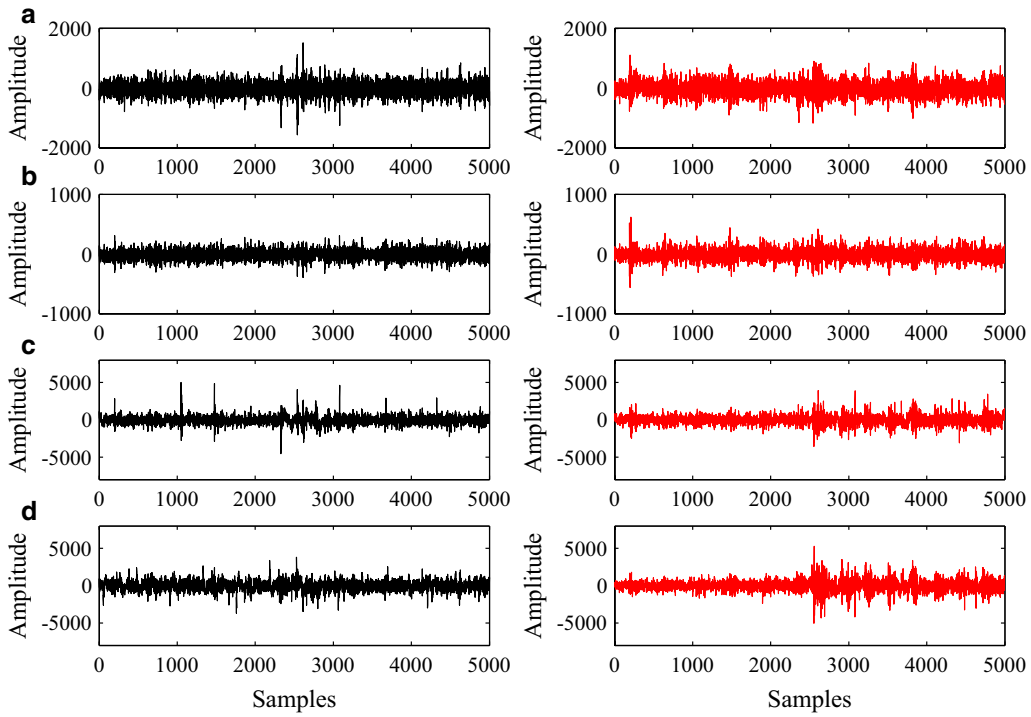

Fig. 5 De-noised time-series segments by ISISC (left) and PSO-PD (right). **a** Ex component; **b** Ey component; **c** Hx component; **d** Hy component

Experimental data in Qaidam Basin, Qinghai Province

In June 2012, we conducted another experiment in Qaidam Basin, the northeastern part of the Qinghai-Tibet Plateau. The acquisition device is an MTU (Phoenix Geophysics Ltd). The transmitting equipment is a pseudo-random signal transmitter developed by Central South University. A 1.5-km-long, *y*-oriented electric-dipole source was placed 2 km away from the receive dipole. The maximum transmitting current is about 80 A and the frequencies of the controlled-source signals range from 0.0117 Hz to 48 Hz. The observation station is named as QH401504 and located in a very remote and sparsely populated area, so there is almost no human interference. The observation time is 20 h in total, and the data collected in the first 1.5 h are marked as D_1. The D_1 data set is severely polluted by the CSEM noises. The data collected in the next 18.5 h are high-quality data because the transmitter has stopped working during this time, where the data set is marked as D_2.

Figure 7 shows the learned atoms from data set D_1 of the site QH401504. The features of pseudo-random square-wave signals can be easily found from these atoms. In fact, each atom represents a periodic time-series segment. Each component contains three atoms, which means that there are three types of pseudo-random sources. They are different in fundamental frequencies. Figure 8 shows the results of signal–noise separation by ISISC. It is clear that the raw time series shown in Fig. 8 can be easily represented by a certain atom illustrated in Fig. 7. As shown in Fig. 8a, the raw data and de-noised data have the same trend (very low-frequency signal). It means that our new method has little risk of discarding the effective low-frequency signal. As shown in Fig. 8d, both the raw signal and de-noised signal have a spike around the sampling point of 220. This is the evidence that our method can accurately identify the target noise and retain useful signals. Similar evidence can be found near the sampling point of 12,000 in Fig. 8b. The processing of the MT data set QH401504 fully illustrates that our method can accurately identify and remove relevant and persistent noises.

As shown in Fig. 9, the MT responses obtained using high-quality data set D_2 vary slowly with frequencies, while the results achieved using the noisy D_1 and noisy-free D_2 data sets are severely distorted between 2 Hz and 0.002 Hz. The results obtained after applied our ISISC-based method are greatly improved in the whole band and are highly consistent with the results obtained from high-quality data.

Real data in Lujiang-Zongyang ore district, Anhui Province

Because of the well-developed economics in the Lujiang-Zongyang ore district, the MT data collected in this region are more or less contaminated by cultural noises.

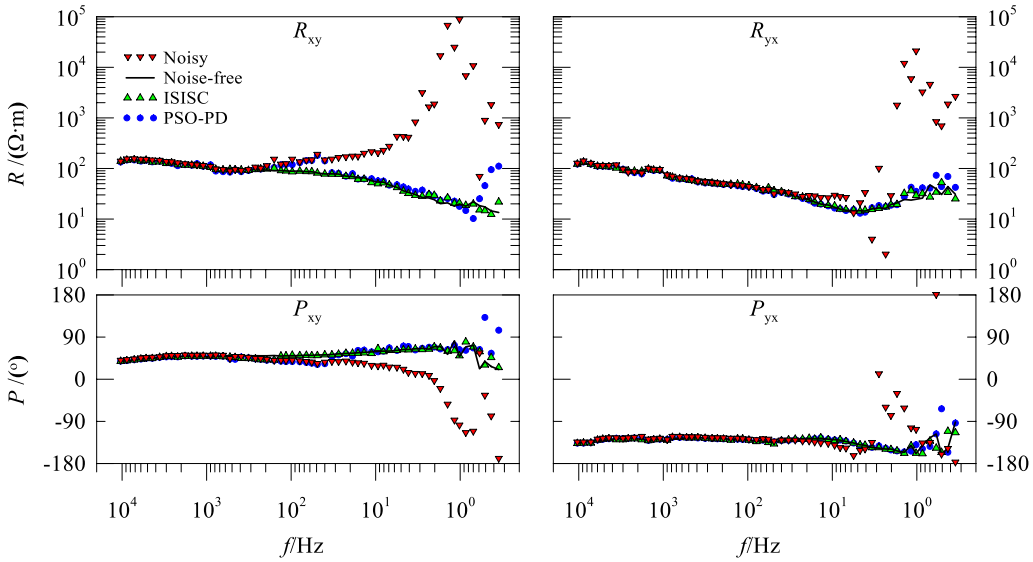

Fig. 6 Apparent resistivity and phase curves of the station ASY0002B. The upper panels are apparent resistivity curves and the bottom panels are phase curves. The red curves with inverted triangles represent the results obtained by all raw data sets D_1 and D_2; the black solid lines represent the results obtained using noise-free data set D_1; the green curves with triangles stand for the results achieved by data sets D_1 and D_2 but is de-noised by ISISC; the curves with blue circles stand for the results calculated from the data sets D_1 and D_2, but filtered by the pulse dictionary-based method

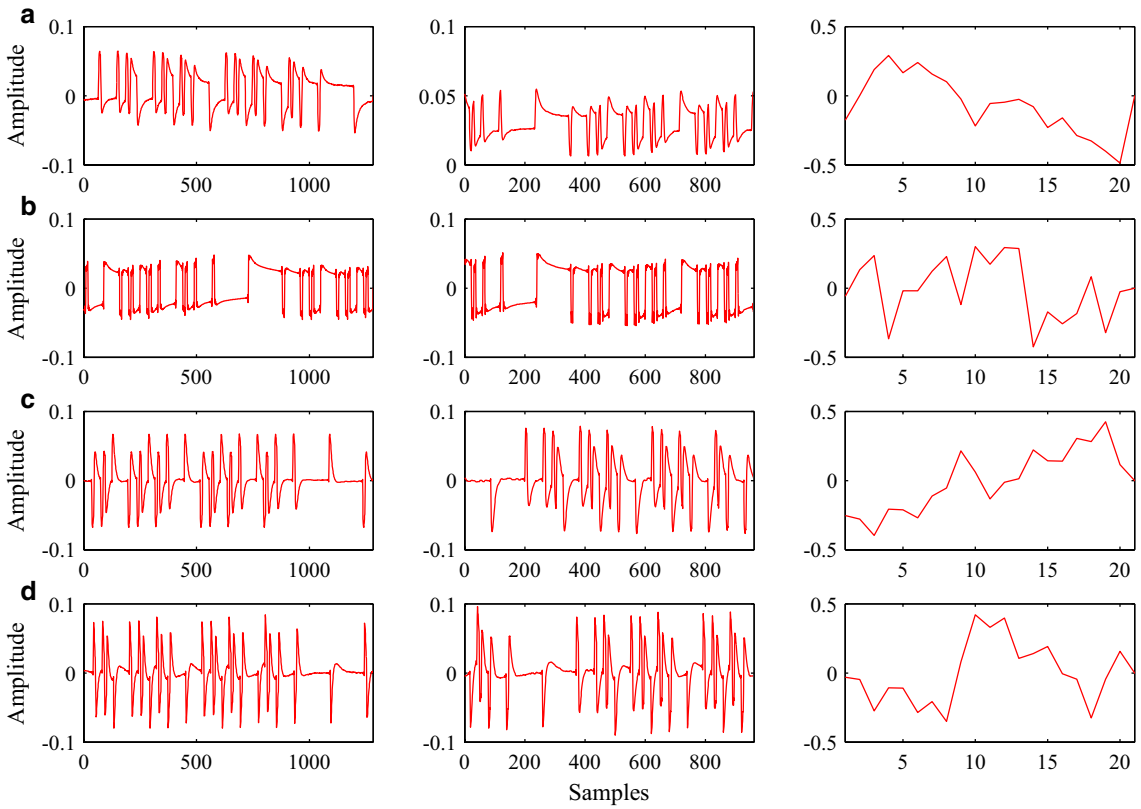

Fig. 7 Feature structures (atoms) learned from Ex (**a**), Ey (**b**), Hx (**c**) and Hy (**d**) components of the MT data set D_1 (a sampling rate of 15 Hz) using ISISC

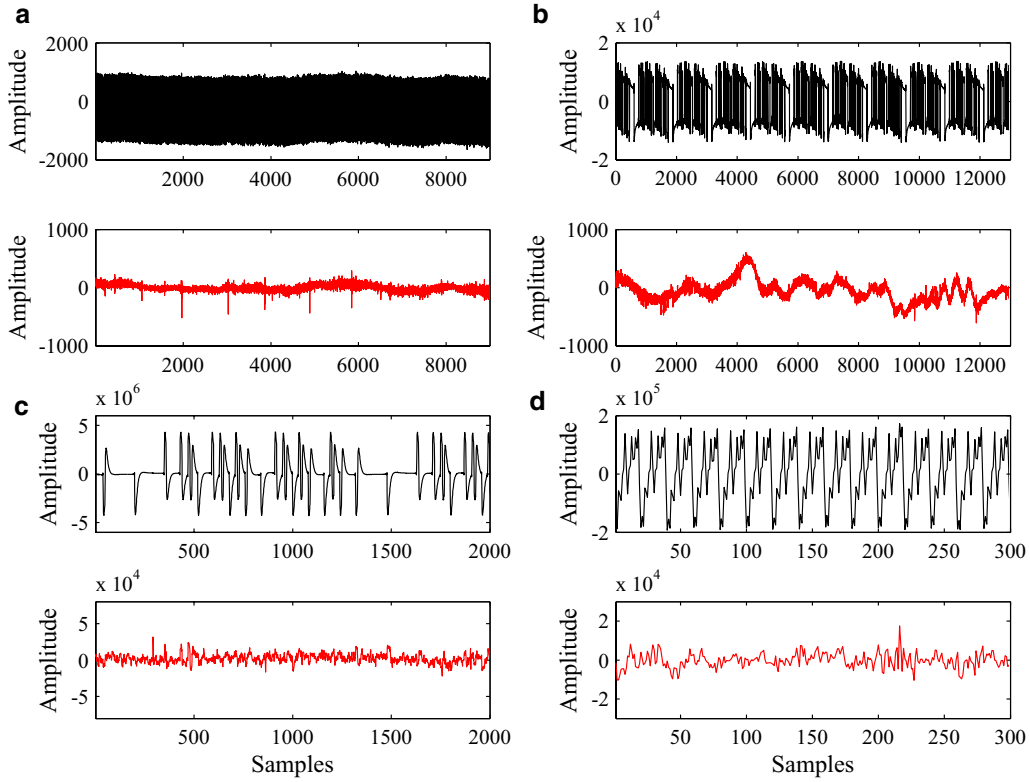

Fig. 8 Raw (black) and de-noised (read) time-series segments from the data set D_1 of the station QH401504, a sampling rate of 15 Hz. **a** Ex component. **b** Ey component. **c** Hx component. **d** Hy component

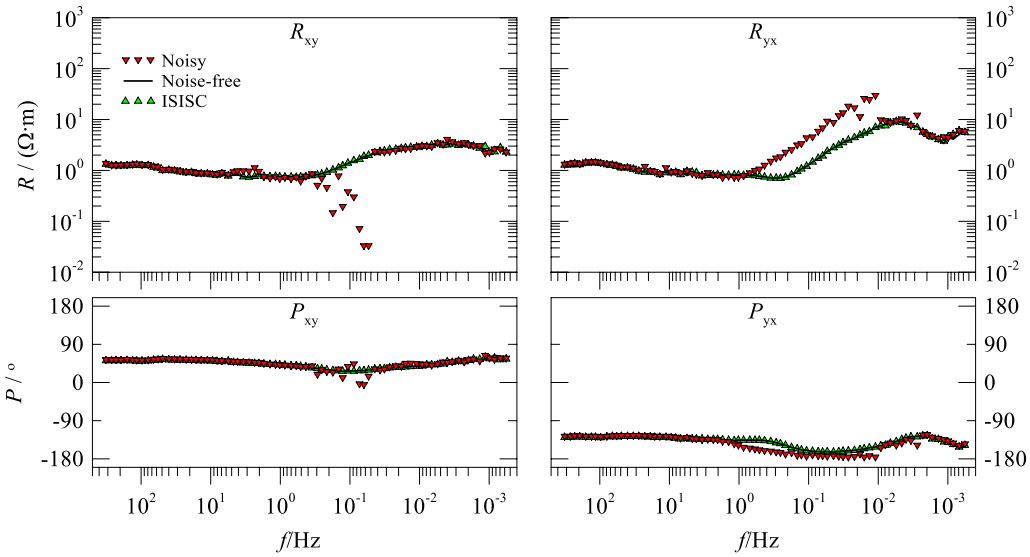

Fig. 9 Apparent resistivity and phase curves of the station QH401504. The upper panels are apparent resistivity curves and the bottom panels are phase curves. The red curves denote the results obtained by all the raw data sets D_1 and D_2; the solid lines represent the results obtained using noise-free data set D_2; the green curves stand for the results achieved by data sets D_1 and D_2, but is de-noised by our new method

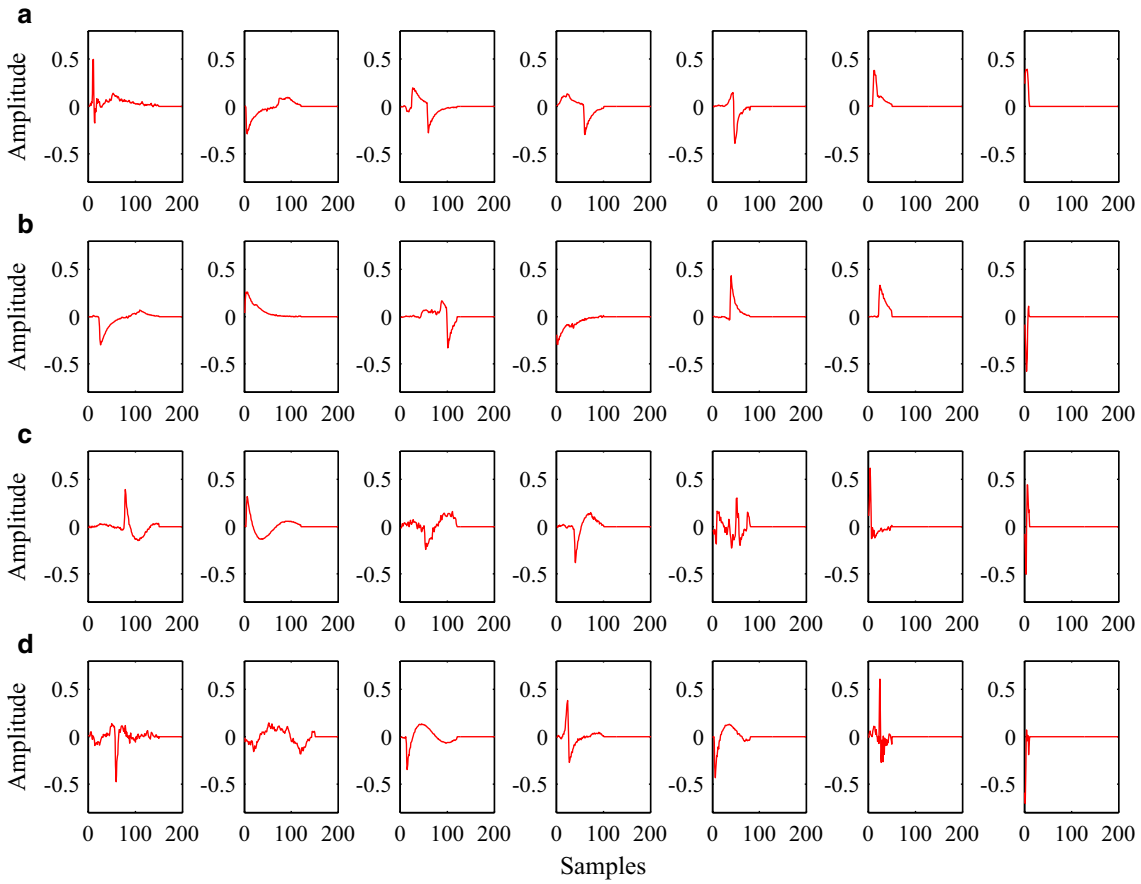

Fig. 10 Feature structures (atoms) learned from Ex (**a**), Ey (**b**), Hx (**c**) and Hy (**d**) components of the MT data set EL22192 (a sampling rate of 150 Hz) using ISISC

There are a lot of time series with abnormal large amplitudes or regular structures. According to the origin and characteristics of natural magnetotelluric signals, such anomalous and periodic structures are not effective MT signals and need to be removed. Nevertheless, the noises are relevant and persistent and therefore difficult to be effectively eliminated by traditional methods.

Figure 10 shows the atoms learned from the real data set EL22192. These structures have regular morphology and strong instantaneous energy. Obviously these structures are generated by human activities.

As shown in Fig. 11, the raw time series is decomposed into 7 independent components and a residual (the denoised signal). The reconstructed components have different characteristics and may come from independent noise sources. After removing these reconstructed components, the residual has no regular components or abnormal large amplitude structures. Judging from results of the time series, the ISISC-based method effectively removed strong cultural noises.

As shown in Fig. 12, there are numerous outliers in the curves calculated from the raw data. When the frequency is lower than 40 Hz, the apparent resistivity varying linearly with frequencies and rising up with 45° on the logarithmic coordinates, phase at the corresponding frequency approaching 0 or $\pm 180°$, which is similar to the CSAMT apparent resistivity and phase curves in the near-field. This is the so-called near-source effect (Wei and Pedersen 1991; Tang et al. 2013). It is clear that the data are contaminated by noises near the observation station, and these raw curves cannot accurately reflect the electrical structures in the subsurface.

After processing by our method, the apparent resistivity and phase curves vary smoothly with the frequencies and the near-source effect has been eliminated completely. The values of the apparent resistivity and phase are distributed in a reasonable range. The results obtained by our method are consistent with those obtained by the remote-reference processing method, except for a few data in frequency bands below 3 Hz. Case studies of MT data sets recorded in Lujiang-Zongyang ore district have

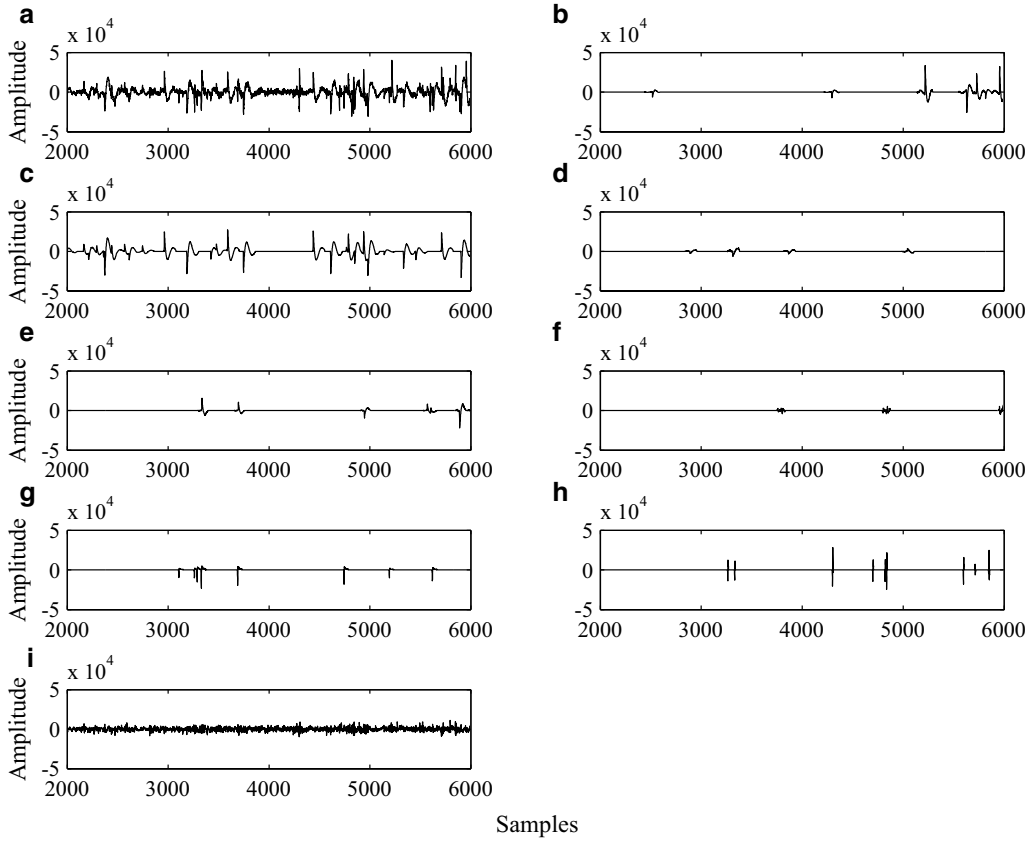

Fig. 11 Raw time-series segments (**a**), reconstructed components (**b**–**h**), and residual (**i**) of the Hx component from real site EL22192. Each component is reconstructed by a learned atom. The residual is obtained by subtracting all reconstructed components from the raw signal

shown that our new scheme can effectively remove the cultural noises and obtain comparable results to that from remote-reference processing.

Conclusions and discussion

To improve the quality of MT data, we propose a novel signal–noise separation method based on the improved shift-invariant sparse coding. The application of the proposed method to synthetic data and real MT data shows:

1. The new scheme proposed in this paper can improve the signal–noise ratio of MT data significantly and the MT responses obtained by our method are consistent with that from the remote reference method. This method can be used with or without a remote reference station and may provide superior results in two important cases: when there is no remote reference station and when the noise is coherent between the local and remote reference stations.
2. In our new scheme, the redundant dictionary is learned autonomously from the data to be processed, instead of predefined manually. Compared with the methods based on predefined redundant dictionary, the adaptability of the newly proposed method is greatly enhanced.
3. The number of atoms in the learned dictionary is greatly reduced since the dictionary is learned based on the characteristics of the data itself. This change greatly reduces the time consumption of subsequent decomposition. Noise removal using our new procedure can be completed on a laptop or a desktop computer. Nevertheless, it is worthwhile to improve the efficiency of our method since the amount of data in the magnetotelluric method is very large.

The method proposed in this paper is based on single site processing, and does not take into account the regularity of the cultural noise between different sites. In

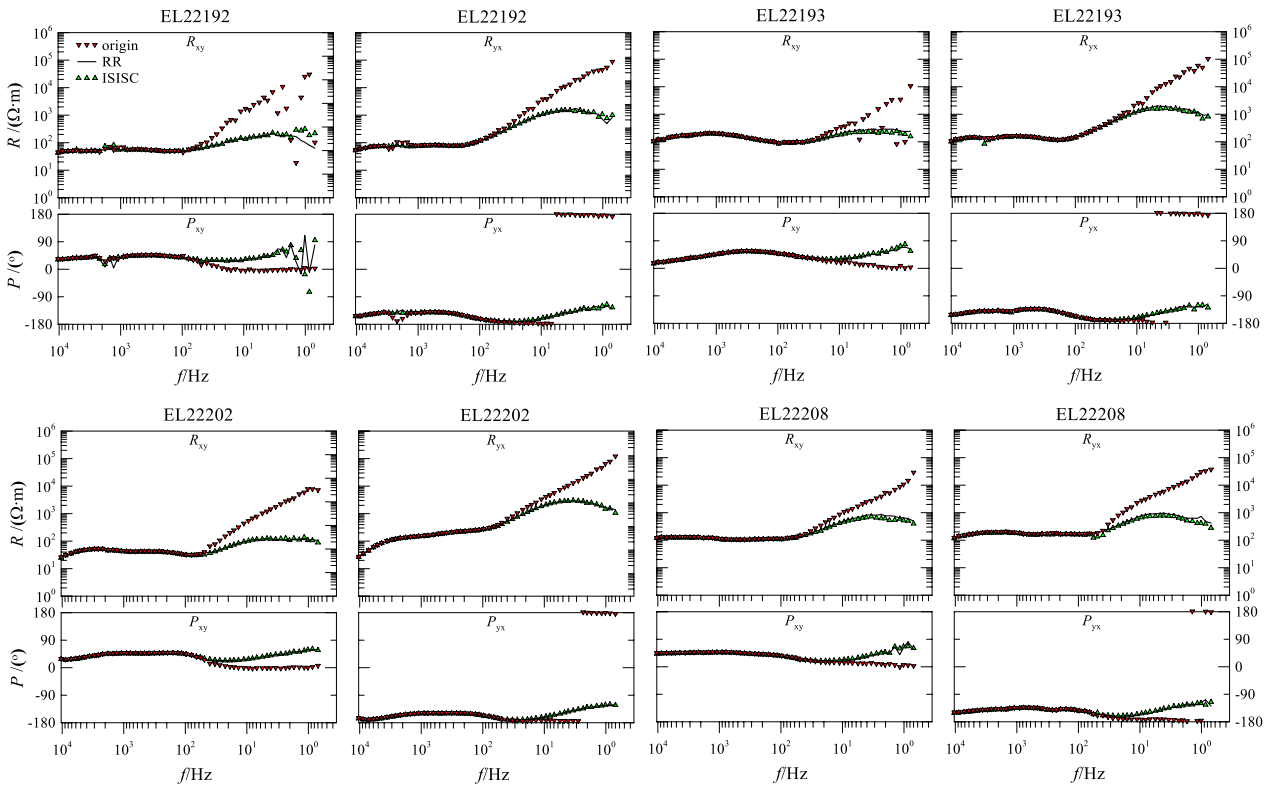

Fig. 12 Apparent resistivity and phase curves of the stations in Lujiang-Zongyang ore district. The upper panels are resistivity curves and the bottom panels are phase curves. The red curves with inverted triangles denote the results obtained using the raw noisy data, the black solid line represents the results obtained by remote reference method, the green curves with triangles stand for the results achieved by our new method

fact, cultural noises at different stations are likely to be relevant. It is possible to improve the performance of the proposed method using dictionary learning or other machine learning algorithms for multi-station processing. Besides, the new scheme proposed in this paper could be used together with remote-reference processing. The ISISC may take out the high amplitude and repetitive cultural noises while RR processing of the cleaned time series may further remove faint incoherent noises.

Abbreviations

MT: Magnetotelluric; SISC: Shift-invariant sparse coding; ISISC: Improved SISC; GSISC: Gradient-based SISC; MP: Matching pursuit; OMP: Orthogonal MP; SNR: Signal-to-noise ratio; NCC: Normalized cross-correlation; CSEM: Controlled-source electromagnetic method; CSAMT: Controlled-source audio magnetotelluric; RR: Remote reference; SD: Square-wave dictionary; IOMP: Improved OMP; PD: Pulse dictionary; PSO: Particle swarm optimization.

Acknowledgements

We are very grateful to the Geology Survey Institution of Anhui Province for providing help during the field experiments. We would also like to thank Xiao Xiao, Lincheng Zhang, Zijie Liu and other field work crew and data process team. The authors also acknowledge the anonymous reviewers for their valuable and constructive comments helpful in improving this paper.

Authors' contributions

LiG proposed the idea, wrote most of the manuscript and is one of the administrators of the projects. LiX assisted in processing MT data and wrote part of the manuscript. Professor Tang JT and Professor Deng JZ provided the MT data and technical guidance. Hu SG conducted the field experiments and assisted in analyzing MT data. Zhou C assisted in analyzing and processing MT data and is one of the administrators of the projects. Tang WW and Chen CJ revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is financially supported by the National Natural Science Foundation of China (Nos. 41904076, 41904072, 41830107 and 41674077), Natural Science Foundation of Jiangxi Province (No. 20192BAB212009), Science and Technology Project of Jiangxi Provincial Education Department (No. GJJ180368) and National Key R&D Program of China (No. 2018YFC0603202).

Availability of data and materials

The datasets and MATLAB code used during the current study are available from the corresponding authors on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1] State Key Laboratory of Nuclear Resources and Environment, East China University of Technology, Nanchang 330013, China. [2] Key Laboratory of Metallogenic Prediction of Non-Ferrous Metals and Geological Environment Monitor, Ministry of Education, Central South University, Changsha 410083, China. [3] College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China. [4] Department of Earth Sciences, Institute of Geophysics, ETH Zurich, 8092 Zurich, Switzerland.

Received: 6 July 2019 Accepted: 28 March 2020
Published online: 06 April 2020

References

Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322
Blumensath T, Davies M (2005) Shift-invariant sparse coding for single channel blind source separation. Proc SPARS 05:75–78
Blumensath T, Davies M (2006) Sparse and shift-invariant representations of music. IEEE Trans Audio Speech Lang Process 14(1):50–57
Cai JH (2016) A combinatorial filtering method for magnetotelluric data series with strong interference. Arab J Geosci 9(13):628
Cai JH, Tang JT, Hua XR, Gong YR (2009) An analysis method for magnetotelluric data based on the Hilbert-Huang Transform. Explor Geophys 40:197–205
Campanya J, Ledo J, Queralt P, Marcuello A, Jones A (2014) A new methodology to estimate magnetotelluric (MT) tensor relationships: estimation of Local transfer-functIons by Combining Interstation Transfer-functions (ELICIT). Geophys J Int 198(1):484–494
Candès EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25(2):21–30
Chen Y (2017) Fast dictionary learning for noise attenuation of multidimensional seismic data. Geophys J Int 209:21–31
Egbert GD (1997) Robust multiple-station magnetotelluric data processing. Geophys J Int 130(2):475–496
Egbert GD, Booker JR (1986) Robust estimation of geomagnetic transfer functions. Geophys J Int 87(1):173–194
Escalas M, Queralt P, Ledo J (2013) Polarisation analysis of magnetotelluric time series using a wavelet-based scheme: a method for detection and characterisation of cultural noise sources. Phys Earth Planet Inter 218(218):31–50
Gamble TD, Goubau WM, Clarke J (1979) Magnetotellurics with a remote magnetic reference. Geophysics 44(1):53–68
Garcia X, Jones AG (2008) Robust processing of magnetotelluric data in the AMT dead band using the continuous wavelet transform. Geophysics 73(6):F223–F234
Garcia X, Seillé H, Elsenbeck J (2015) Structure of the mantle beneath the Alboran Basin from magnetotelluric soundings. Geochem Geophys Geosyst 16(12):4261–4274. https://doi.org/10.1002/2015GC006100
Guo R, Liu L, Liu J, Sun Y, Liu R (2019) Effect of data error correlations on transdimensional MT Bayesian inversions. Earth Plan Space 71:134. https://doi.org/10.1186/s40623-019-1118-3
Jafari MG, Plumbley MD (2011) Fast dictionary learning for sparse representations of speech signals. IEEE J Selected Topics Signal Processing 5(5):1025–1031
Larnier H, Sailhac P, Chambodut A (2016) New application of wavelets in magnetotelluric data processing: reducing impedance bias. Earth Plan Space 68(1):70. https://doi.org/10.1186/s40623-016-0446-9
Li G, Xiao X, Tang JT, Li J, Zhu HJ, Zhou C, Yan FB (2017) Near-source noise suppression of AMT by compressive sensing and mathematical morphology filtering. Appl Geophys 14(4):581–589. https://doi.org/10.1007/s11770-017-0645-6
Li J, Zhang X, Gong JZ, Tang JT, Ren ZY, Li G, Deng YL, Cai J (2018) Signal-noise identification of magnetotelluric signals using fractal-entropy and clustering algorithm for targeted de-noising. Fractals 26(2):1840011
Li G, Liu X, Tang J, Li J, Ren Z, Chen C (2020) De-noising low-frequency magnetotelluric data using mathematical morphology filtering and sparse representation. J Appl Geophys 172(2020):103919. https://doi.org/10.1016/j.jappgeo.2019.103919
Liu H, Liu C, Huang Y (2011) Adaptive feature extraction using sparse coding for machinery fault diagnosis. Mech Syst Signal Processing 25(2):558–574
Mallat SG, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process 41(12):3397–3415
Neukirch M, Garcia X (2014) Nonstationary magnetotelluric data processing with instantaneous parameter. J Geophys Res Solid Earth 119(3):1634–1654. https://doi.org/10.1002/2013JB010494
Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583):607–609
Pati YC, Rezaiifar R, Krishnaprasad PS (1993) Orthogonal matching pursuits: Recursive function approximation with applications to wavelet decomposition. Proceedings of the 27th Asilomar Conference in Signals, Systems, and Computers. Pacific Grove, CA: IEEE
Plumbley MD, Abdallah SA, Blumensath T (2006) Sparse representations of polyphonic music. Signal Processing 86(3):417–431
Simpson F, Bahr K (2005) Practical magnetotellurics. Cambridge University Press, Cambridge
Smith EC, Lewicki MS (2006) Efficient auditory coding. Nature 439(7079):978–982
Tang JT, Zhou C, Wang X, Xiao X, Lv QT (2013) Deep electrical structure and geological significance of Tongling ore district. Tectonophysics 606:78–96
Tang JT, Li G, Xiao X, Li J, Zhou C, Zhu HJ (2017) Strong noise separation for magnetotelluric data based on a signal reconstruction algorithm of compressive sensing. Chin J Geophys 60(9):3642–3654. https://doi.org/10.6038/cjg20170928 **(In Chinese with English abstract)**
Tang JT, Li G, Zhou C, Li J, Liu XQ, Zhu HJ (2018) Power-line interference suppression of MT data based on frequency domain sparse decomposition. J Cent S Univ 25(9):2150–2163
Trad DO, Travassos JM (2000) Wavelet filtering of magnetotelluric data. Geophysics 65(2):482–491
Wang X, Zhu H, Wang D, Zhao Y, Li Y (2013) The diagnosis of rolling bearing based on the parameters of pulse atoms and degree of cyclostationarity. J Vibroeng 15(3):1560–1575
Wang X, Zhu H, Rui T, Li Y, Liu T, Liu M (2015) Shift invariant sparse coding ensemble and its application in rolling bearing fault diagnosis. J Vibroeng 17(4):1837–1848
Wei Q, Pedersen LB (1991) Industrial interference magnetotellurics: an example from the Tangshan area, China. Geophysics 56(2):265–273
Zhang H, Diao S, Chen W, Hang G, Li H, Bai M (2019) Curvelet reconstruction of non-uniformly sampled seismic data using the linearized Bregman method. Geophys Prospect 67(5):1201–1218
Zhu HJ, Wang XQ, Rui T, Li YF, Zhang HT, Zhao Y (2015) Shift invariant sparse coding for blind source separation of single channel mechanical signal. J Vibration Eng 28(4):625–632 **(In Chinese with English abstract)**
Zhu HJ, Wang XQ, Rui T, Li YF, Wang D (2016) Multi scale shift invariant sparse coding for robust machinery diagnosis. Trans Beijing Inst Technol 36(1):19–24 **(In Chinese with English abstract)**

Publisher's Note