Symplectic geometry spectrum regression for prediction of noisy time series

Hong-Bo Xie,[1,*] Socrates Dokos,[2] Bellie Sivakumar,[3,4] and Kerrie Mengersen[1]

[1]*ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane QLD 4000, Australia*

[2]*Graduate School of Biomedical Engineering, The University of New South Wales, Sydney NSW 2052, Australia*

[3]*School of Civil and Environmental Engineering, The University of New South Wales, Sydney NSW 2052, Australia*

[4]*Department of Land, Air and Water Resources, University of California, Davis, California 95616, USA*

(Received 21 October 2015; revised manuscript received 6 March 2016; published 20 May 2016)

We present the symplectic geometry spectrum regression (SGSR) technique as well as a regularized method based on SGSR for prediction of nonlinear time series. The main tool of analysis is the symplectic geometry spectrum analysis, which decomposes a time series into the sum of a small number of independent and interpretable components. The key to successful regularization is to damp higher order symplectic geometry spectrum components. The effectiveness of SGSR and its superiority over local approximation using ordinary least squares are demonstrated through prediction of two noisy synthetic chaotic time series (Lorenz and Rössler series), and then tested for prediction of three real-world data sets (Mississippi River flow data and electromyographic and mechanomyographic signal recorded from human body).

DOI: 10.1103/PhysRevE.93.052217

I. INTRODUCTION

During the past several decades, the theory of chaos has moved to center stage in many areas such as physics, ecology, hydrology, sociology, economics, finance, atmospheric sciences, and life sciences. A broad range of chaotic processes that one can observe in the above diverse areas of natural and human spheres inspires a great interest to develop chaotic models and, particularly, to predict the future evolution of a system from its past measurements. As a consequence, the ability of chaos theory–based models to achieve accurate predictions of time series has been an important area of research in recent decades. In this regard, the local approximation (LA) method proposed by Farmer and Sidorowich [1] is very popular among the local prediction methods. The idea behind such a method is to recognize that any manifold in a high-dimensional space is locally linear. In this method, after embedding a time series in a state space using delay coordinates, a local approximation based on ordinary least squares (OLS) is applied to learn the induced nonlinear mapping [1]. This method allows us to make a short-term prediction of the future behavior of a time series, using information based on past values.

According to Occam's razor, the local approximation method is simple but effective. The coefficients of the local linear model are typically estimated using ordinary least squares. Apart from potential linearization errors, the approach also suffers from the high variance of the predictions under noisy conditions [2]. These issues can significantly degrade the predictive accuracy of real systems, as such systems inevitably contain some amount of noise. Kugiumtzis *et al.* [2] showed that the regularization technique, originally derived to solve ill-posed regression problems, could give better predictions than OLS on noisy chaotic time series. In order to avoid the negative effects of noise, they considered four regularization techniques: principal component regression (PCR), partial least squares (PLS), ridge regression (RR), and truncated total least squares (TTLS). The first three regularization methods were found to provide improved prediction performance compared to OLS for synthetic noise–corrupted data from typical nonlinear systems. Similar results were also found for real-world data from the R-R intervals of ECG signals and sunspot data. Jade *et al.* [3] extended the PCR technique to kernel principal component regression (KPCR), which first maps the input data into high-dimensional space through some nonlinear function in prediction. Results obtained for the Lorenz and Mackey-Glass equations and laser data in the Sante Fe Institute prediction contest demonstrated that the KPCR, when combined with a model parameters selection method, can improve the prediction of the unseen test data, especially those with sharp singularities. However, it is well known that the kernel methods often suffer from a heavy computational burden.

All of these regularization techniques are essentially considered in Euclidean space and based on singular value decomposition (SVD) and principal component analysis (PCA). However, classical eigenvalue or subspace methods, such as PCA and SVD, can only deal with flat Euclidean structures and, thus, fail to discover the curved or nonlinear structures of the input data. Although there are various nonlinear extensions of PCA, such methods often suffer from difficulties in designing cost functions or tuning too many free parameters. Moreover, most of these methods are computationally expensive, thus severely limiting their application to high-dimensional data sets. In this paper, we present a symplectic geometry spectrum regression (SGSR) regularization technique based on symplectic geometry theory [4].

Symplectic geometry and its associated eigenvalue methods have several unique characteristics for overcoming some important limitations inherent in traditional approaches for complex data analysis, such as those based on Euclidean geometry. The symplectic transform is structure preserving, which means eigenvalues can be approximated more accurately [5]. Compared with SVD, symplectic matrix factorizations exhibit small norm and condition numbers, which is desirable for

[*] Present address: ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane, QLD 4000, Australia; hongbo.xie@qut.edu.au

improving the numerical stability and noise performance in data analysis and image processing [6,7]. In addition, the symplectic transform identifies nonlinear relations in a set of data points, while preserving global submanifold geometrical properties of the data [8]. Several studies have shown that symplectic geometry spectrum–based methods are superior to SVD-based techniques in the detection of chaos [9], estimation of the embedding dimension of a nonlinear dynamic system [10], and in the reduction of noise in nonlinear systems [11,12]. The SGSR approach proposed in this study creates the components by modeling the relationship between input and output variables while maintaining most of the information in the input data. The effectiveness of this approach is evaluated by predicting two noisy synthetic chaotic time series (Lorenz and Rössler series) and three real-world time series (Mississippi River flow data and electromyographic and mechanomyographic signal recorded from the human body). The results are also compared with those obtained using OLS.

II. METHODS

A. Local linear prediction model

Consider a univariate time series $x_1,x_2,\dots,x_n$, where n is the number of samples, generated from a D-dimensional chaotic attractor. A phase space of the attractor can be reconstructed by using delay coordinates defined as

$$X_i = \{x_i,x_{i+\tau},\dots,x_{i+(d-1)\tau}\}^T, \tag{1}$$

where d is the embedding dimension of the reconstructed phase space, τ is the delay time, and T denotes the vector transpose. The original time series can thus be mapped into a multidimensional state space, as

$$\mathbf{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_m^T \end{bmatrix} = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(d-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(d-1)\tau} \\ \vdots & \vdots & \cdots & \vdots \\ x_m & x_{m+\tau} & \cdots & x_{m+(d-1)\tau} \end{bmatrix}, \tag{2}$$

where $m = n-(d-1)\tau$ is the number of points in the d-dimensional attractor.

The above reconstruction allows predictions of the time series, through a local approximation that relates the present and future states of the system. The first step in the local linear prediction method is to find the nearest neighbor points of the current phase point X_m in the reconstructed phase space. Because of the assumption of deterministic behavior, it is reasonably expected that the evolution of the selected vector is correlated with the evolution of the neighboring vectors, which could provide predictions of the future value of X_m through an appropriate local model. Given the embedding vector, we calculate the Euclidean norm between the point X_m and all the remaining points $X_i (i = 1,2,\dots,m-1)$. The closeness is evaluated and q nearest neighbors X_m^k are selected, where $k = 1,2,\dots,q$ and $q > d$ in most cases.

The prediction of x_{n+1} involves finding an estimator of the regression function so that $\hat{x}_{n+1} = \hat{f}(X_m)$. An autoregressive model is often applied to obtain the local map function $\hat{f}$ in the local linear prediction model. The prediction value is thus a linear superposition of the d elements in the delay vector X_m, which can be represented as

$$\hat{x}_{n+1} = GZ_m = g_0 + \sum_{j=1}^{d} g_j x_{m+(j-1)\tau}, \tag{3}$$

where $G = [g_0,g_1,\dots,g_d]$ is a coefficient vector that needs to be determined, and

$$Z_m = [1,X_m]^T = [1,x_m,x_{m+\tau},\dots,x_{m+(d-1)\tau}]^T. \tag{4}$$

The local prediction method relies on the fact that a set of nearest neighbors evolves similarly in the reconstructed chaotic attractor. Thus, such models have to learn neighborhood relations from the data and map them forward in time. For phase point X_m that is similar to its q nearest neighbors X_m^k currently, the future point X_{m+1} will be close to the future point set X_{m+1}^k. The coefficient vector G can be identified from the current phase point X_m and its neighborhood X_m^k:

$$G\mathbf{X}_c = X_f, \tag{5}$$

where $X_f = [x^1_{m+(d-1)\tau+1},x^2_{m+(d-1)\tau+1},\dots,x^q_{m+(d-1)\tau+1}]^T$ is the next series value of the nearest points X_m^k $(k = 1,2,\dots,q)$, and

$$\mathbf{X}_c = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_m^1 & x_m^2 & \cdots & x_m^q \\ \vdots & \vdots & \ddots & \vdots \\ x^1_{m+(d-1)\tau} & x^2_{m+(d-1)\tau} & \cdots & x^q_{m+(d-1)\tau} \end{bmatrix}. \tag{6}$$

This linear regression problem is solvable in the ordinary least square sense and leads to $G = X_f\mathbf{X}_c^T(\mathbf{X}_c\mathbf{X}_c^T)^{-1}$. Then, the prediction $\hat{x}_{n+1}$ can be obtained from Eq. (3). The multistep prediction consists of repeating the above one-step predictions up to the desired horizon. To evaluate the prediction accuracy, central tendency estimates of forecast error, such as the normalized mean squared error (NMSE), are often adopted, while the receiver operating characteristic (ROC) is also used in exceptional conditions, such as in the prediction of extreme events [13,14]. In essence, better prediction is dependent on the optimal trade-off between, for example, bias and variance and consists of finding filter factors such that the mean square error is minimized [2].

B. Symplectic geometry spectrum regression (SGSR)

A Hamiltonian matrix $\mathbf{M} \in R^{2n\times 2n}$ has the form

$$\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{L} \\ \mathbf{Q} & -\mathbf{A}^T \end{pmatrix}, \quad \mathbf{L} = \mathbf{L}^T, \quad \mathbf{Q} = \mathbf{Q}^T, \tag{7}$$

where $\mathbf{A}$, $\mathbf{L}$, and $\mathbf{Q}$ are real $n \times n$ matrices.

A ubiquitous matrix when dealing with Hamiltonian eigenvalue problems is the skew-symmetric matrix,

$$\mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{8}$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix.

By straightforward algebraic manipulation one can show that a Hamiltonian matrix $\mathbf{M}$ is equivalently defined by the property

$$\mathbf{MJ} = (\mathbf{MJ})^T. \tag{9}$$

However, if matrix $\mathbf{N} \in R^{2n\times 2n}$ satisfies

$$(\mathbf{NJ})^T = -\mathbf{NJ}, \tag{10}$$

then it is called skew-Hamiltonian. Therefore, for a Hamiltonian matrix $\mathbf{M}$, $\mathbf{M}^2 = \mathbf{N}$ is skew-Hamiltonian.

Any matrix $\mathbf{S} \in R^{2n\times 2n}$ satisfying

$$\mathbf{S}^T\mathbf{JS} = \mathbf{SJS}^T = \mathbf{J} \tag{11}$$

is called symplectic, and since

$$(\mathbf{S}^{-1}\mathbf{HS})\mathbf{J} = \mathbf{S}^{-1}\mathbf{HJS}^{-T} = \mathbf{S}^{-1}\mathbf{J}^T\mathbf{H}^T\mathbf{S}^{-T} = [(\mathbf{S}^{-1}\mathbf{HS})\mathbf{J}]^T, \tag{12}$$

the Hamiltonian structure is preserved in symplectic similarity transformations. For a symplectic matrix $\mathbf{S}$, there is $\mathbf{S} = \mathbf{QR}$, where $\mathbf{Q}$ is a symplectic unitary matrix and $\mathbf{R}$ is an upper triangle matrix.

Similar to PCR, symplectic geometry spectrum regression starts from the reconstructed trajectory matrix. This matrix is transformed into a symmetric matrix and then a Hamiltonian matrix in symplectic space. The Hamiltonian matrix is subjected to a symplectic QR decomposition to obtain its eigenvalues and eigenvectors. Each of these eigenvectors can be inversely transformed into a reconstructed embedding vector. For the trajectory matrix in Eq. (2), a $d \times d$ autocorrelation matrix is given by $\mathbf{A} = \mathbf{X}^T\mathbf{X}$. Then, we can construct a Hamiltonian matrix, as follows:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}^T \end{bmatrix}. \tag{13}$$

For this Hamiltonian matrix $\mathbf{M}$, its eigenvalues can be evaluated by symplectic QR decomposition and the primary $2d$-dimensional space can be transformed into d dimensions [6,12,15]. After the real Hamiltonian matrix $\mathbf{M}$ is squared to form $\mathbf{M}^2 = \mathbf{N}$, a symplectic orthogonal matrix $\mathbf{P}$ is constructed such that

$$\mathbf{P}^T\mathbf{NP} = \begin{bmatrix} \mathbf{B} & \mathbf{R} \\ \mathbf{0} & \mathbf{B}^T \end{bmatrix}, \tag{14}$$

where $\mathbf{B}$ is the upper Hessenberg matrix. Various methods can be used to construct the symplectic orthogonal matrix $\mathbf{P}$. Given a Householder matrix $\mathbf{Q}$, it can easily be shown that the matrix $\mathbf{H} = [\mathbf{Q0};\mathbf{0Q}]$ is also a Householder matrix and that, furthermore, $\mathbf{H}$ is symplectic and unitary [9]. In order to simplify the computation, we can construct matrix $\mathbf{Q}$ by Schmidt orthogonalization, using $\mathbf{H}$ to replace $\mathbf{P}$ to obtain the upper Hessenberg matrix $\mathbf{B}$:

$$\begin{aligned}\mathbf{HMH}^T &= \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}^T \end{bmatrix}\begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{QAQ}^T & \mathbf{0} \\ \mathbf{0} & -\mathbf{QA}^T\mathbf{Q}^T \end{bmatrix} = \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & -\mathbf{B}^T \end{bmatrix}. \end{aligned} \tag{15}$$

The sympletic QR algorithm is used to compute the eigenvalues $\sigma(\mathbf{B}) = \{\sigma_1,\sigma_2,\ldots,\sigma_d\}$ [9,10]. If $\mathbf{A}$ is real and symmetric, then eigenvalues of $\mathbf{A}$ are equal to those of $\mathbf{B}$, and the eigenvalues $\lambda(\mathbf{X})$ of $\mathbf{X}$ can be obtained from the positive square roots of $\sigma(\mathbf{B})$ as

$$\lambda(\mathbf{X})_j = \sqrt{\sigma_j}, \quad j = 1,2,\ldots,d, \tag{16}$$

with eigenvalues in descending order,

$$\sigma_1 > \sigma_2 > ,\ldots, > \sigma_p \gg \sigma_{p+1} \geqslant ,\ldots, \geqslant \sigma_d. \tag{17}$$

The σ_j are the symplectic singular values, constituting the symplectic geometry spectra of $\mathbf{A}$ with relevant symplectic orthonormal bases. The corresponding matrix $\mathbf{Q}$ denotes the symplectic eigenvectors of $\mathbf{A}$. Those σ_j with low values are often related to the noise component in the data. Similar to the SVD-based regularized regression estimation, those smallest symplectic eigenvalues result in larger variance in the prediction. If we use only the first p eigenvectors to form the symplectic principal eigenvalue matrix while discarding the remaining eigenvectors from $p+1$ in the prediction, then the regression is called a symplectic geometry spectrum regression. The underlying qualitative assumption behind this method is that the projections of X_f onto the last d-p columns of $\mathbf{Q}$ are below the noise level and, therefore, give little or no information about the true neighbors of the next point.

C. Local prediction based on SGSR

In the case of noisy time series, even increasing the number of observations cannot ensure an effective application of the OLS algorithm because the probability that false neighbors appear and true neighbors get expelled is high, thus resulting in large variance [2]. This problem can be overcome with the use of SGSR. The specific steps in the prediction of noisy series using SGSR can be summarized as follows:

(1) Given a scalar time series $x_1,x_2,\ldots,x_n$, select the embedding dimension d and the delay time τ to construct the trajectory matrix $\mathbf{X}_{m\times d}$;

(2) Build the real $d \times d$ symmetric matrix $\mathbf{A}$;

(3) Calculate the symplectic principal components of $\mathbf{A}$ by symplectic QR decomposition, and form the Householder transform matrix $\mathbf{Q}$;

(4) Determine the regularization parameter p by finding a threshold value that represents the noise variance.

(5) Construct the corresponding symplectic principal eigenvalue matrix $\mathbf{W}$ according to the regularization parameter p, i.e., $\mathbf{W} = \mathbf{Q}(:,1:p)$;

(6) Form the transformed coefficients matrix $\mathbf{S} = \mathbf{W}^T\mathbf{X} = \mathbf{Q}^T(:,1:p)\mathbf{X}$;

(7) Form the new trajectory matrix $\mathbf{Y} = \mathbf{WS} = \sum_{i=1}^{p} W(:,i)\mathbf{S}$;

(8) Search for the q nearest neighbors X_m^k of the phase point X_m in the new trajectory matrix $\mathbf{Y}$;

(9) Fit Eq. (5) to get the prediction coefficient vector G and calculate the prediction value $\hat{x}_{n+1}$ by Eq. (3);

(10) For the new target point, repeat the above steps until the desired prediction horizon.

III. RESULTS

A. Application to synthetic chaotic data

The SGSR and the conventional OLS prediction methods are first employed to predict two noisy synthetic time series: the Lorenz and Rössler series [16,17]. The equations of the Lorenz and Rössler systems, numerical integration method and steps involved, and the appropriate embedding parameters were detailed in previous papers [18,19]. In this study, a total of

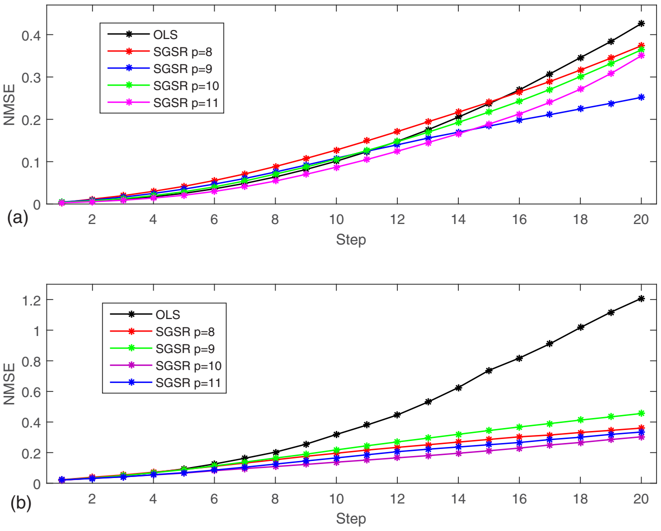

FIG. 1. Prediction with SGSR regularization technique and OLS for data generated from the Lorenz system corrupted with 2% (a) and 5% (b) NLs.

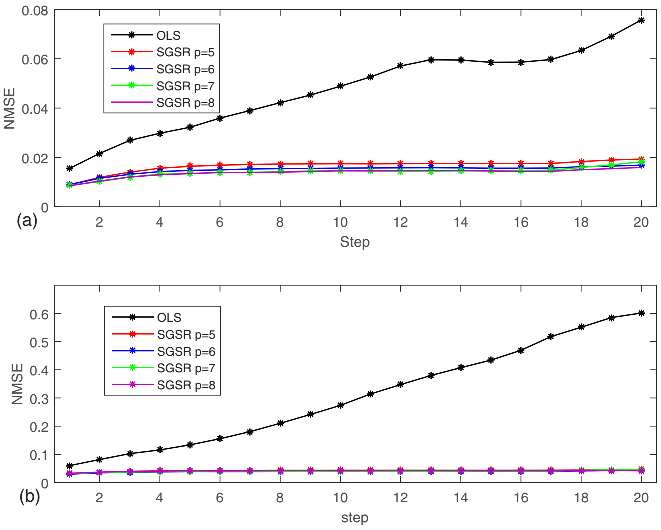

FIG. 2. Prediction with SGSR regularization technique and OLS for data generated from the Rössler system corrupted with 5% (a) and 10% (b) NLs.

3000 data points are used for analysis, with the first 2000 points used for training the model and the remaining 1000 points for testing. The prediction quality is evaluated in terms of the normalized mean squared error. For the noise-free Lorenz and Rössler time series, the predictions obtained from SGSR and OLS are very similar, indicating both methods perform equally well when the data are noise free.

We now focus on evaluating the NMSEs for the SGSR method on the Lorenz and Rössler series with superimposed noise. Noise is superimposed onto the time series through the addition of independent and identically distributed (i.i.d.) Gaussian white noise with various noise levels (NLs). For the Lorenz series, noise levels of 2% and 5% are considered, while 5% and 10% noise levels are considered for the Rössler series. For both series, predictions are made for lead time (or prediction horizon) from 1 to 20. As the number of neighbors plays an important role in predictions, we consider several different numbers of neighbors as well, from $d + 1$ to $d + 10$.

Figure 1 shows the NMSEs obtained using SGSR and OLS methods for the Lorenz series for all lead times, when the number of neighbors is $d + 1$. For both 2% and 5% NL, the NMSEs increase with increasing lead time for both SGSR and OLS methods. Since the predicted values are typically iterated to obtain future values in multistep prediction, any prediction error at a given time will propagate and accumulate in later predictions. For the 2% NL, the prediction errors of SGSR and OLS are similar when the lead time is below 10. However, when the lead time is greater than 11, the NMSEs of SGSR are lower than OLS for regularization parameter $p = 9, 10$, and 11, and SGSR consistently outperforms OLS for lead time beyond 15. For the case of 5% NL, the NMSEs of SGSR are lower than OLS for all the lead times considered. Further, for longer lead time, SGSR is significantly superior to OLS, indicating the superiority of SGSR for multistep prediction of noisy Lorenz series. It is clear, therefore, that SGSR is particularly more suitable and effective than OLS for prediction of series with high noise levels.

Figure 2 shows the NMSEs obtained using SGSR and OLS methods for the Rössler series for all the lead times, when the number of neighbors is $d + 1$. Similar to the case of Lorenz series, the prediction error of the SGSR method for Rössler series is lower than OLS in 5% NL. For the case of 10% NL, the NMSEs of SGSR are significantly lower than those of OLS, and particularly so at longer lead times. This example further validates the usefulness of SGSR and its superiority over OLS for multistep prediction of time series heavily contaminated with noise, especially for much longer lead times.

With investigation of the effect of regularization parameter p done above, the effect of the number of neighbors is now examined. Here we test the effect of nearest neighbors on the prediction performance for the SGSR. We vary the number of nearest neighbors from $d + 1$ to $d + 10$. Figure 3 shows the NMSEs for the Lorenz series with 2% and 5% NLs for $p = 9$ and 11, respectively, where the optimal regularization

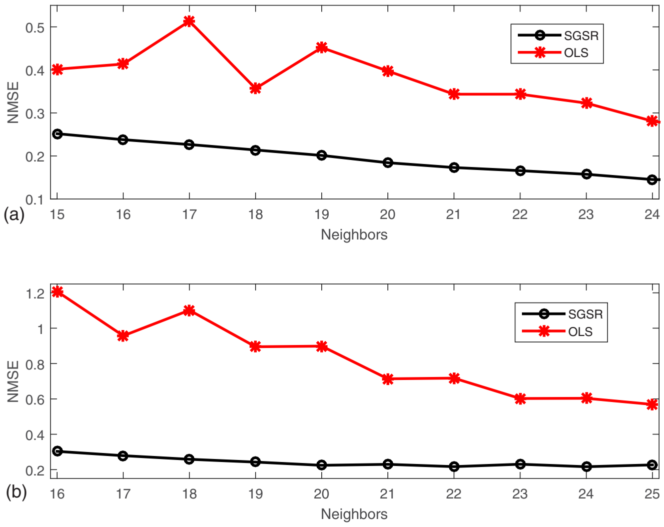

FIG. 3. The effect of nearest neighbors for 20-step prediction of Lorenz series with 2% (a) and 5% (b) NLs.

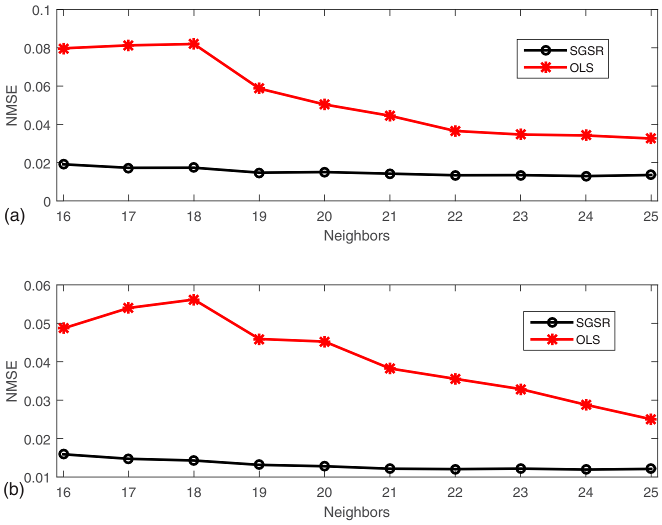

FIG. 4. The effect of nearest neighbors for 20-step prediction of Rössler series with 5% (a) and 10% (b) NLs.

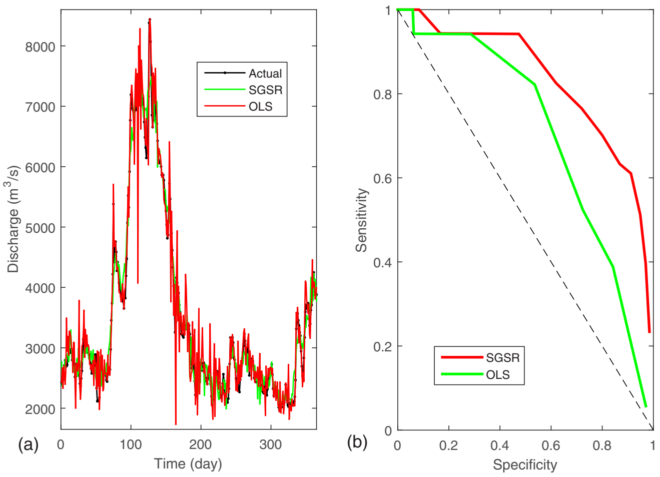

FIG. 5. Mississippi River flow over year 2016 and its 30-step prediction values by SGSR and OLS: direct time series plot (a) and receiver operating characteristic curves (b).

parameter p is chosen as the natural cutoff level in the symplectic geometry spectrum of Eq. (17). Figure 4 shows the results for the Rössler series with 5% and 10% NLs. The results indicate that the NMSEs of both SGSR and OLS decrease with increasing number of neighbors, but the decrease for OLS is far more pronounced when compared with that for SGSR. However, SGSR still has lower NMSEs for each time series and noise level, as well as number of nearest neighbors.

B. Application to real-world data

To test the applicability of SGSR for prediction of real-world data, we consider three data sets: Mississippi River flow data (representing hydrologic time series) and the electromyographic (EMG) and mechanomyographical (MMG) signals collected from human skeletal muscle (representing biomedical time series).

1. *Mississippi River flow*

Prediction of flows in rivers is crucial for planning and management of our water resources and environment. Large river basins, in particular, play key roles in the socioeconomic development at regional scales and beyond. The Mississippi River basin is one of the world's major river systems in size, habitat diversity, and biological productivity, thus meeting the water demands and improving and sustaining the socioeconomic development of millions of people in the United States and Canada. Therefore, modeling and prediction of flows in the Mississippi River have been of enormous interest. During the past decade or so, a number of studies have also investigated the chaotic behavior of flow and sediment dynamics in the Mississippi River and their prediction [20–22]. It is important to note, however, that river flow data are often contaminated by noise, both measurement and dynamic, which can influence the outcomes of chaos identification and prediction of such data. Therefore, it is worthwhile to apply the SGSR method for prediction of river flow data and test its effectiveness.

In this study, we analyze the flow data from the Mississippi River basin to test the effectiveness of the SGSR method for prediction purposes and its superiority over the OLS method. We consider daily flow data observed at the St. Louis gaging station, Missouri (US Geological Survey station no. 07010000). Studying the flow at the daily scale is particularly important for assessment of high flows (and floods) and to undertake emergency measures. We use flow data observed during 2001–2005 for training the SGSR and OLS models and make predictions for 2006.

Figure 5, for instance, shows the prediction results obtained using SGSR and OLS for the Mississippi River flow for a lead time of 30 (days), through direct time series comparisons [Fig. 5(a)] and ROC curves [Fig. 5(b)]. In this case, the number of nearest neighbors is 9 and the optimal optimization parameter is 2. As can be seen, the predicted flows from both methods match reasonably well with the observed values, but the SGSR method (with an NMSE value of 0.033) consistently outperforms the OLS method (NMSE = 0.052), as it captures both the major changes and the minor variations in flow dynamics more accurately. Moreover, the area under the ROC curve (AUC) for SGSR is larger than that for OLS, which confirms the higher accuracy of SGSR for Mississippi River flow data prediction when compared to that from OLS. Figure 6 presents a summary of prediction NMSEs for both methods with lead time from 1 to 30 days. As seen, the NMSEs increase with the increasing lead time for both methods, as normally expected. However, the error for SGSR is significantly lower than that for OLS for all the prediction horizons, indicating the consistently better performance of the former over the latter for any predictability horizon. The results, for both methods, indicate that the errors are not strictly increasing with increasing lead time but are slightly fluctuating at certain lead times. This may be due to the limited number of nearest neighbors selected in multistep prediction. However, additional evidence is needed to confirm this. We will investigate this aspect in a future study.

2. *Electromyographic (EMG) and mechanomyographical (MMG) signals*

EMG and MMG signals, recordings of electrical and mechanical activities detectable on the body surface during

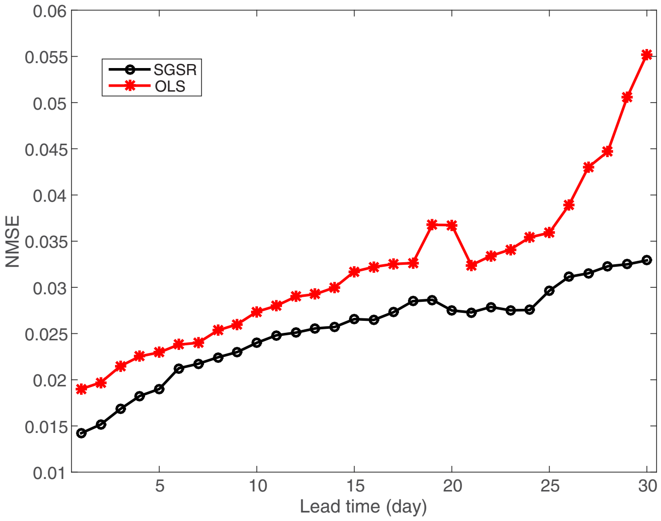

FIG. 6. The Mississippi River flow prediction NMSEs with lead time from 1 to 30 days for SGSR and OLS.

muscle contraction, have been broadly used to control different human-machine interfaces (HMIs) [23]. However, muscle fatigue often happens with sustained contraction, which can degrade the robustness and accuracy of HMIs. A real-time signal prediction scheme is, thus, essential in these systems in order to compensate for the effect of muscle fatigue. However, EMG or MMG signals are contaminated with noise due to the limb movement artifact, cross talk, and measurement system and environmental noises. Therefore, the SGSR method seems to be a suitable tool for their predictions.

In this study, we apply the current and preceding 2000 EMG and MMG data points as a training sample to perform real-time prediction at a lead time of 32. The EMG and MMG data are sampled at a rate of 1000 Hz. A detailed description of the experimental protocol and recording procedures for both signals can be found in Xie *et al.* [23,24]. The reason for considering the lead time of 32 is that the minimal length of a moving window satisfying real-time control in EMG- or MMG-based HMI is 32 in most cases [23]. Figures 7 and 8

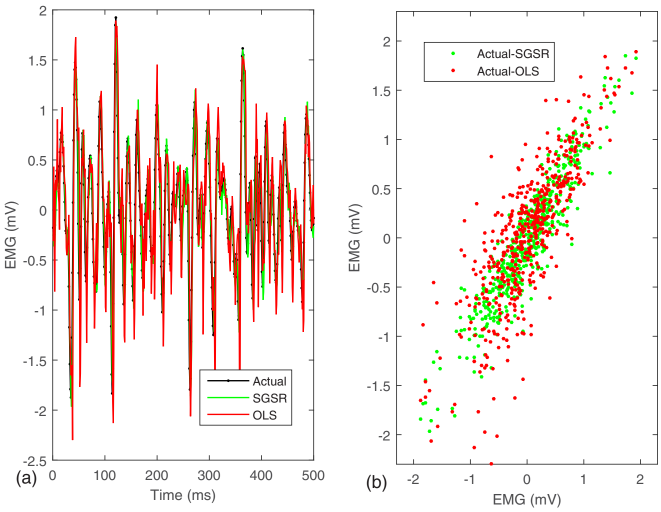

FIG. 7. A 500-point EMG segment and its 32-step prediction by SGSR and OLS: direct time series plot (a) and scatter plot (b).

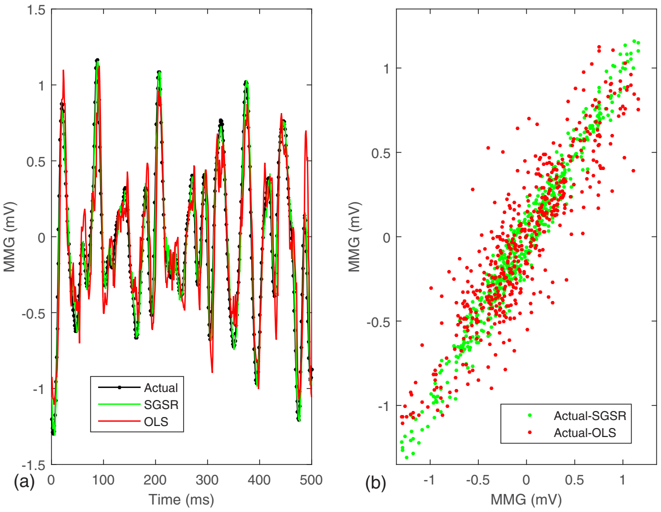

FIG. 8. A 500-point MMG segment and its 32-step prediction by SGSR and OLS: direct time series plot (a) and scatter plot (b).

show the actual and predicted values through both the time series and scatter plots using SGSR and OLS on a 500-point EMG and MMG segment, respectively. The NMSEs for EMG are 0.07 using SGSR and 0.132 using OLS, while those for MMG are 0.032 using SGSR and 0.076 using OLS. The results for the EMG and MMG time signals further indicate the superiority of SGSR over OLS for short-term prediction of noisy time series. MMG is a low-frequency biosignal with a relatively narrow main frequency band from 5 to 50 Hz, while EMG has a wide frequency band from 10 to 300 Hz. MMG's major noise source is low-frequency movement artifact, while EMG signals are mostly contaminated by both movement artifact and high-frequency noise. Many previous studies have indicated that both signals are deterministic and even chaotic [19,25]. This example further demonstrates the effectiveness of the symplectic geometry method for prediction of real-world signals with different physical and physiological characteristics.

IV. DISCUSSION AND CONCLUSIONS

We have presented a time series regression method based on symplectic geometry theory, and used it to predict both simulated and experimentally recorded noisy time series. The performance of the method on two benchmark problems in hydrology and medicine shows its superiority over the existing OLS method. Similar to PCR, the idea of SGSR is to achieve a trade-off between the prediction bias and variance. For the Mississippi River flow, though SGSR performs better than OLS, we find that the prediction error of SGSR is relatively large. This is because SGSR is also sensitive to outliers in the data (e.g. sharp peaks) that produce large errors in the sense of minimizing the squared error.

In previous studies, Xie *et al.* [11,12] developed the symplectic geometry spectrum analysis (SGSA) framework which decomposes a time series into its constituent components. The SGSA method consists of four steps, i.e., embedding, symplectic QR decomposition, grouping, and diagonal averaging [12]. Similar to principal component analysis, independent component analysis, and empirical mode decomposition, SGSA can be used as a denoising method when discarding those noise components [12]. Xie *et al.* [11,12] also extended the SGSA for time series decomposition to a hybrid prediction method, in which SGSA denoises the original time series, and the local approximation technique based on ordinary least squares is conducted to predict the denoised data. The prediction method involved in such a method can be summarized in seven steps: embedding; symplectic QR decomposition; grouping; diagonal averaging; denoising; new embedding; local approximation. The time complexity of this hybrid prediction method is $O(n^3)$. However, the SGSR mode presented in this paper can be summarized in just four steps: embedding; symplectic QR decomposition; regularization; local approximation. The algorithm presented in this study has significantly low time complexity $[O(n^2)]$, especially with four steps from "grouping" to "new embedding" in SGSA-based prediction replaced by "regularization" in SGSR. In order to further compare the prediction performance of two symplectic geometry–based methods, we applied the hybrid SGSA prediction approach to the same Mississippi River flow, EMG, and MMG data sets. The NMSEs for the river flow, EMG, and MMG data were 0.038, 0.072, and 0.033, respectively. Compared with results in Sec. III, the SGSR model has slightly lower or similar prediction NMSEs over the SGSA-based prediction method. However, the former is much more efficient than the latter due to the lower computational complexity. This is certainly a great advantage, especially when one is dealing with large data sets.

In this paper, the focal comparison between the SGSR and OLS methods is minimization of the squared error. This merits further research into robust symplectic geometry spectrum regression. Such an approach could be developed by combining the SGSR with a robust outlier detection method, thus improving the prediction performance for noisy time series with outliers.

[1] J. D. Farmer and J. J. Sidorowich, Predicting Chaotic Time-Series, Phys. Rev. Lett. **59**, 845 (1987).

[2] D. Kugiumtzis, O. C. Lingjærde, and N. Christophersen, Regularized local linear prediction of chaotic time series, Physica D **112**, 344 (1998).

[3] A. M. Jade, V. K. Jayaraman, and B. D. Kulkarni, Improved time series prediction with a new method for selection of model parameters, J. Phys. A: Math. Gen. **39**, L483 (2006).

[4] K. Feng and M. Z. Qin, *Symplectic Geometry Algorithms for Hamiltonian Systems* (Zhejiang Science & Technology Press, Hangzhou, China, 2003).

[5] H. Fassbender and D. Kressner, Structured eigenvalue problems, GAMM-Mitteilungen **29**, 297 (2006).

[6] H. Xu, An SVD-like matrix decomposition and its applications, Linear Algebra Appl. **368**, 1 (2003).

[7] S. Skare, M. Hedehus, M. E.Moseley, and T. Q. Li, Condition number as a measure of noise performance of diffusion tensor data acquisition schemes with MRI, J. Magn. Reson. **147**, 340 (2000).

[8] A. K. M. Nazimuddin and M. R. Hasan, Applications of Riemannian geometry comparing with symplectic geometry, Ann. Pure Appl. Math. **6**, 170 (2014).

[9] H. B. Xie, H. Huang, and Z. Z. Wang, Identification determinism in time series based on symplectic geometry spectra, Phys. Lett. A **342**, 156 (2005).

[10] M. Lei, Z. Z. Wang, and Z. J. Feng, A method of embedding dimension based on symplectic geometry, Phys. Lett. A **303**, 179 (2002).

[11] H. B. Xie and S. Dokos, A symplectic geometry-based method for nonlinear time series decomposition and prediction, Appl. Phys. Lett. **103**, 054103 (2014).

[12] H. B. Xie, T. Guo, B. Sivakumar, A. W. C. Liew, and S. Dokos, Symplectic geometry spectrum analysis of nonlinear time series, Proc. R. Soc. London, Ser. A **470**, 20140409 (2014).

[13] W. L. Gorr, Forecast accuracy measures for exception reporting using receiver operating characteristic curves, Int. J. Forecasting **25**, 48 (2009).

[14] M. I. Bogachev and A. Bunde, Improved risk in multifractal records: Application to the value risk in finance, Phys. Rev. E **80**, 026131 (2009).

[15] C. Van Loan, A symplectic method for approximating all the eigenvalues of a Hamiltonian matrix, Linear Algebra Appl. **61**, 233 (1984).

[16] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. **20**, 130 (1963).

[17] O. E. Rössler, An equation for continuous chaos, Phys. Lett. A **57**, 397 (1976).

[18] H. B. Xie, J. Y. Guo, and Y. P. Zheng, Using the modified sample entropy to detect determinism, Phys. Lett. A **374**, 3926 (2010).

[19] H. B. Xie and S. Dokos, A hybrid symplectic principal component analysis and central tendency measure method for detection of determinism in noisy time series with application to mechanomyography, Chaos **23**, 023131 (2013).

[20] B. Sivakumar and W. W. Wallender, Predictability of river flow and suspended sediment transport in the Mississippi River basin: a non-linear deterministic approach, Earth Surf. Processes Landforms **30**, 665 (2005).

[21] B. Sivakumar and A. W. Jayawardena, An investigation of the presence of low-dimensional chaotic behaviour in the sediment transport phenomenon phenomenon, Hydrol. Sci. J. **47**, 405 (2002).

[22] B. Sivakumar, A phase-space reconstruction approach to prediction of suspended sediment concentration in rivers, J. Hydrol. **258**, 149 (2002).

[23] H. B. Xie, J. Y. Guo, and Y. P. Zheng, Classification of the mechanomyogram signal using a wavelet packet transform and singular value decomposition for multifunction prosthesis control, Physiol Meas. **30**, 441 (2009).

[24] H. B. Xie, H Huang, J. Wu, and L. Liu, A comparative study of surface EMG classification by fuzzy relevance vector machine and fuzzy support vector machine, Physiol Meas. **36**, 191 (2015).

[25] H. B. Xie, J. Y. Guo, and Y. P. Zheng, Uncovering chaotic structure in mechanomyography signals of fatigue biceps brachii muscle, J. Biomech. **43**, 1224 (2010).