Efficient Data-Driven Geologic Feature Characterization from Pre-stack Seismic Measurements using Randomized Machine-Learning Algorithm

Youzuo Lin[1,*], Shusen Wang[2], Jayaraman Thiagarajan[3], George Guthrie[1], and David Coblentz[1]

1 : *Earth and Environment Division, Los Alamos National Laboratory, Los Alamos, NM 87545;*

2 : *Department of Statistics, University of California, Berkeley, CA 94720;*

3 : *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94550.*

Correspondence ∗: ylin@lanl.gov.

SUMMARY

Conventional seismic techniques for detecting the subsurface geologic features are challenged by limited data coverage, computational inefficiency, and subjective human factors. We developed a novel data-driven geological feature characterization approach based on pre-stack seismic measurements. Our characterization method employs an efficient and accurate machine-learning method to extract useful subsurface geologic features automatically. Specifically, we use kernel ridge regression to account for the nonlinear relationship between seismic data and geological features. We further employ kernel tricks to avoid the explicit nonlinear mapping and infinite dimension of feature space. However, the conventional kernel ridge regression can be computationally prohibitive because of the large volume of seismic measurements. We employ a data reduction technique in combination with the conventional kernel ridge regression method to improve the computational efficiency and reduce memory usage. In particular, we utilize a randomized numerical linear algebra technique, named Nyström method, to effectively reduce the dimensionality of the feature space without compromising the information content required

for accurate characterization. We provide thorough computational cost analysis to show the efficiency of our new geological feature characterization methods. We validate the performance of our method in characterizing geologic fault zones because faults play an important role in various subsurface applications. Our numerical examples demonstrate that our new characterization method significantly improves the computational efficiency while maintaining comparable accuracy. Interestingly, we show that our method yields a speed-up ratio on the order of $\sim 10^2$ to $\sim 10^3$ in a multi-core computational environment.

Key words: Geologic Feature characterization, Seismic Measurements, Machine Learning Methods, Dimensionality Reduction, Randomization Techniques, Nyström Approximation

1 INTRODUCTION

It is challenging to analyze and interpret seismic measurements for identifying prospective geological features. The challenges arise from processing of large volumes of seismic data and subjective human factors. Different geologic features play different roles in characterizing the subsurface structure. Since geologic fault is one of the most interesting features in subsurface characterization, we use that as the target to demonstrate the efficacy of our new data-driven geologic feature characterization method. The geologic fault zone is essential to various subsurface energy applications. In geothermal exploration, geologic faults provide important information for siting the drilling wells. In carbon sequestration, geologic faults can be critical to monitor the potential leaks of stored CO_2. In oil & gas production, geologic faults are used to signal reservoir boundaries or hydrocarbon traps.

In current seismic exploration workflows, both imaging/inversion techniques and human interpretation are incorporated to characterize the subsurface. The workflows start with imaging/inversion. Most seismic imaging/inversion techniques are physics dominated, meaning that the governing physics equations are well understood and utilized to describe the underlying physics of the problems of interest. A well-known example of this is the seismic full-waveform inversion (FWI) (Lin & Huang 2015a,c; Virieux et al. 2014; Virieux & Operto 2009). In FWI, an inverse problem is formulated to connect the measurements and the governing physics equations. Numerical optimization techniques are utilized to solve for the subsurface models. Similar framework and procedures can be applied to many other techniques such as seismic imaging (Zhang et al. 2015; Lin & Huang 2015b), tomography (Lin et al. 2015; Rawlinson & Sambridge 2014), etc. Even though those conventional methods have been shown great success in many applications, in some situations they can be limited because of poor

data coverage, computational inefficiency, and subjective human factors. With the imaging/inversion results available, the human interpretor will be employed to make certain modifications based on domain knowledge. However, an incorrect imaging/inversion result can either interfere the interpretation, or even mislead the human expert. Considering the critical impact of seismic imaging/inversion techniques to the final interpretation, a robust, efficient, and accurate subsurface characterization method is therefore needed.

With the advancement of data science and machine learning, there has been a recent surge in utilizing automated machine learning methods to characterize subsurface geologic features (Schnetzler & Alumbaugh 2017; Araya-Polo et al. 2017; Guillen 2015; Zhang et al. 2014; Hale 2013; Ramirez & Meyer 2011). In seismic applications, the machine learning methods can be categorized into either "learning from prestack data" or "learning from migrated/inverted model" as shown in Fig. 1. The major difference between these two types of methods is whether a machine learning method works on the pre-stack seismic datasets or migrated/inverted models. Most of the existing machine learning methods for seismic applications are based on the migrated/inverted models, meaning migrated or inverted models need to be obtained prior to the use of machine learning techniques. In Guillen (2015), migration imaging models are first obtained from seismic datasets. Machine learning method is then applied to the imaging model to automatically characterize the salt body. Similarly, in Hale (2013), a seismic image is first computed before the estimation of the geologic fault location. Despite the success of those methods, there are limitations though. Firstly, to obtain the migrated/inverted models usually involves significant amount of computation. Secondly, the success of the prediction heavily relies on the resulting migrated/inverted model obtained from the data. To avoid these limitations, another type of learning method has been recently proposed and developed, i.e., learning from pre-stack seismic data directly. In the work of Araya-Polo et al. (2017) and Zhang et al. (2014), supervised learning methods are directly applied to the pre-stack seismic data to look for patterns which indicate the geologic features. Specifically, in Araya-Polo et al. (2017) deep neural network was applied to seismic datasets to obtain geologic faults. In Zhang et al. (2014), kernel regression was used to learn a mapping between seismic data and geologic faults. In both papers, promising results have been reported.

In this work, our novel geologic feature characterization belongs to the "learning from prestack seismic data" category, meaning our algorithm detects geological features from pre-stack seismic data directly. Through our experiments, we notice that despite the success of those existing "learning from prestack seismic data" methods in controlled experiments, they are significantly limited by their computational efficiency, mostly due to the need to process large volumes of high-dimensional data. Consequently, none of the existing solutions are suitable for real-time or even near real-time characterization.

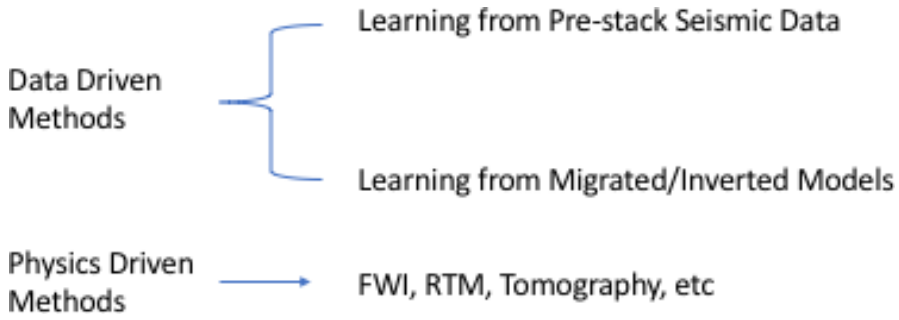


Figure 1. The category of the different seismic exploration approaches. The data driven methods can be categorized into either "learning from prestack data" or "learning from migrated model". The major difference between these two types of methods is whether a machine learning method works on the pre-stack seismic datasets or migrated/inverted models.

In typical exploratory geophysics applications, strongly rectangular data arise, which implies that the number of receivers is much smaller than the number of data points that each receiver collects. Hence, we develop a scalable geologic feature characterization technique by utilizing tools from randomized linear algebra allowing computational efficient geological feature characterization.

Randomized matrix approximation methods enable us to efficiently deal with large-scale problems by sacrificing a provably trivial amount of accuracy (Drineas & Mahoney 2016). Broadly, the underlying idea is to perform dimensionality reduction on the large-scale matrix without losing information pertinent to the considered task. The approach is to construct a sketch of the input matrix, which is usually a smaller matrix that yields a good approximation and represents the essential information of the original input (Drineas & Mahoney 2016). The sketch can be obtained by applying a random projection or selection matrix to the original data. Randomized algorithms have been successfully applied to various scientific and engineering domains, such as scientific computation and numerical linear algebra (Meng & Mahoney 2014; Drineas et al. 2011; Lin et al. 2010; Rokhlin & Tygert 2008), seismic full-waveform inversion and tomography (Moghaddam et al. 2013; Krebs et al. 2009), and medical imaging (Huang et al. 2016; Wang et al. 2015; Zhang et al. 2012).

In this paper, we developed a novel randomized geologic feature characterization method. In particular, we consider the use of kernel machines for automated feature characterization and design a scalable algorithm using the Nyström approximation (Drineas & Mahoney 2005; Gittens & Mahoney 2016). It is well known that the kernel matrix is the bottleneck for scaling up the kernel machines, because the forming, storing, and manipulating of the kernel matrix have high time and memory costs. The main idea of Nyström method is to approximate an arbitrary symmetric positive semidefinite (SPSD) kernel matrix using a small subset of its columns, and the method reduces the time complexity of many matrix operations from $\mathcal{O}(n^2)$ or $O(n^3)$ to $\mathcal{O}(n)$ and space complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$, where n is the number of data samples. There has been various work applying Nyström

approximation to improve the computational efficiency and memory usage in machine learning community. Williams & Seeger (2001) used the Nyström method to speed up matrix inverse such that the inference of large-scale Gaussian process regression can be efficiently performed. Later on, the Nyström method has been applied to spectral clustering (Li et al. 2011; Fowlkes et al. 2004), kernel SVMs (Zhang et al. 2008), and kernel PCA and manifold learning (Talwalkar et al. 2013), etc. In this work, we employ the Nyström approximation to kernel ridge regression. Instead of forming the full kernel matrix from seismic data, we generate a low-rank approximation of the full kernel matrix by using Nyström approximation. We further validate the performance of our new subsurface geologic feature characterization method using synthetic surface seismic data. Our proposed characterization method significantly improves the computational efficiency while maintaining the accuracy of the full model.

In the following sections, we first briefly describe some fundamentals of underlying geology and the governing physics of our problem of interests. We then go through the data-driven approaches – kernel ridge regression (Sec. 2). We develop and discuss our novel geologic feature characterization method based on randomized kernel ridge regression method (Sec. 3). We then apply our method to test problems using both acoustic and elastic velocity models and further discuss the results (Sec. 4). Further discussions towards future work are provided in Sec. 5. Finally, concluding remarks are presented in Sec. 6.

2 THEORY

2.1 Geologic Features of Interest: Fault Zones

Geologic fault zone provides critical information for various subsurface energy applications. As an example, in carbon sequestration, leakage of CO_2 and brine along faults at carbon sequestration sites is a primary concern for storage integrity (Zhang et al. 2009). Accurately siting the geologic fault zones is essential to monitor the CO_2 storage. We first provide some fundamentals on the geological fault.

The geological fault is a fracture or crack along which two blocks of rock slide past one another (Haakon 2010). As illustrated in Fig. 2, there are three major geological fault types depending on the relative direction of displacement between the rocks on either side of the fault: normal fault, reverse fault, and strike-slip fault. The fault block above the fault surface is called the hanging wall, while the fault block below the fault is the footwall. In this study, we focus on both normal faults and reverse faults, which are the most common fault types (Haakon 2010).

Out of various geophysical exploration methods, seismic waves are more sensitive to the acous-

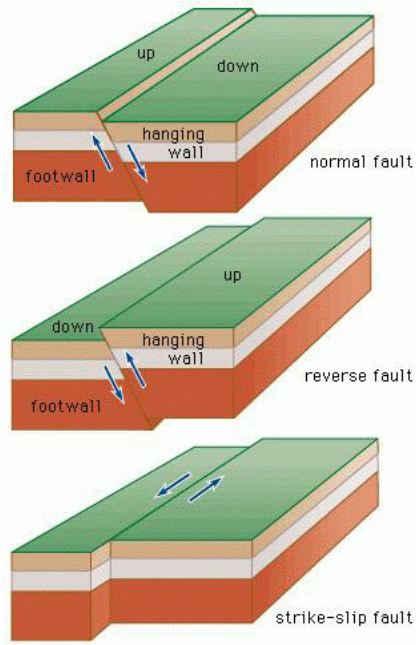


Figure 2. An illustration of the geologic fault zones (image courtesy of Encyclopaedia-Britannica (2010)). There are three major geological fault types depending on the relative direction of displacement between the rocks on either side of the fault: normal fault, reverse fault, and strike-slip fault. The fault block above the fault surface is called the hanging wall, while the fault block below the fault is the footwall.

tic/elastic impedance (which depends on the density and seismic velocity of the medium) of the subsurface than other geophysical measurements (Fig. 3a). Hence, seismic exploration has been widely used to infer changes in the media impedance, which indicates geologic structures. In the next section, we briefly cover the mathematics and governing physics of seismic exploration.

2.2 Physics-Driven Methods

The physics-driven methods (Fig. 1) are those to infer subsurface model provided with governing physics and equations. Take the seismic exploration as an example. Seismic waves are mechanical perturbations that travel in the Earth at a speed governed by the acoustic/elastic impedance of the medium in which they are traveling. In the time-domain, the acoustic-wave equation is given by

$$\left[\frac{1}{K(\mathbf{r})}\frac{\partial^2}{\partial t^2} - \nabla\cdot\left(\frac{1}{\rho(\mathbf{r})}\,\nabla\right)\right] p(\mathbf{r},t) = s(\mathbf{r},\,t), \tag{1}$$

where $\rho(\mathbf{r})$ is the density at spatial location $\mathbf{r}$, $K(\mathbf{r})$ is the bulk modulus, $s(\mathbf{r},\,t)$ is the source term, $p(\mathbf{r},t)$ is the pressure wavefield, and t represents time.

The elastic-wave equation is written as

$$\rho(\mathbf{r})\,\ddot{u}(\mathbf{r},t) - \nabla\cdot[C(\mathbf{r}):\nabla u(\mathbf{r},t)] = s(\mathbf{r},\,t), \tag{2}$$

where $C(\mathbf{r})$ is the elastic tensor, and $u(\mathbf{r}, t)$ is the displacement wavefield.

The forward modeling problems in Eqs. (1) and (2) can be written as

$$P = f(\mathbf{m}), \tag{3}$$

where P is the pressure wavefield for the acoustic case or the displacement wavefields for the elastic case, f is the forward acoustic or elastic-wave modeling operator, and $\mathbf{m}$ is the velocity model parameter vector, including the density and compressional- and shear-wave velocities. We use a time-domain stagger-grid finite-difference scheme to solve the acoustic- or elastic-wave equation. Throughout this paper, we consider only constant density acoustic or elastic media.

The inverse problem of Eq. (3) is usually posed as a minimization problem

$$E(\mathbf{m}) = \min_{\mathbf{m}} \left\{ \|\mathbf{d} - f(\mathbf{m})\|_2^2 + \lambda\, R(\mathbf{m}) \right\}, \tag{4}$$

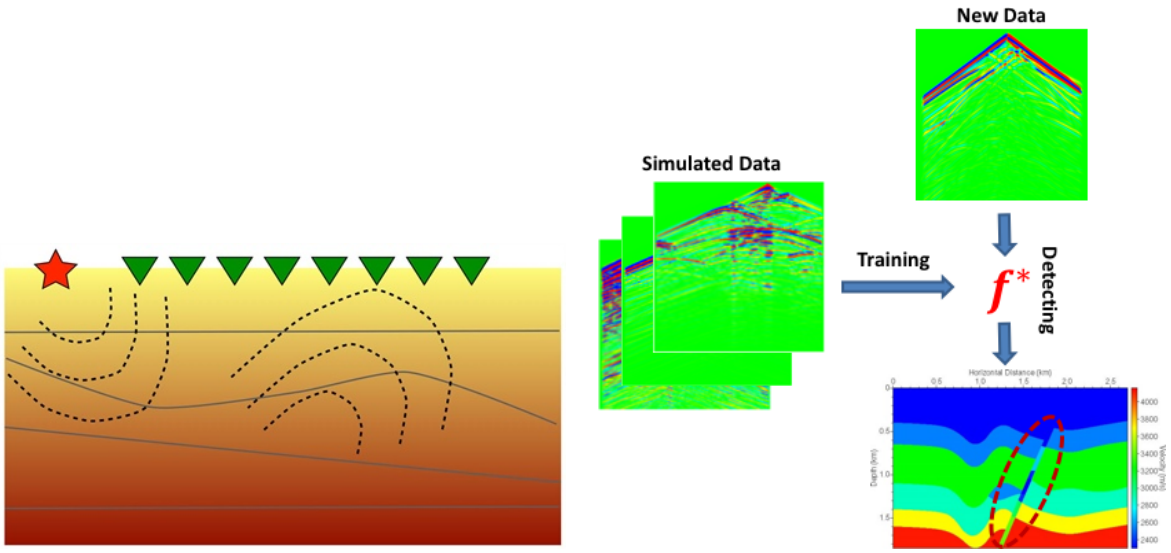
where $\mathbf{d}$ represents a recorded/field waveform dataset, $f(\mathbf{m})$ is the corresponding forward modeling result, $\|\mathbf{d} - f(\mathbf{m})\|_2^2$ is the data misfit, $||\cdot||_2$ stands for the L_2 norm, λ is a regularization parameter and $R(\mathbf{m})$ is the regularization term which is often the L_2 or L_1 norm of $\mathbf{m}$. The current technology to infer the subsurface geologic features is based on seismic inversion and imaging methods, which are computationally expensive and often yield unsatisfactory resolution in identifying small geologic features (Lin & Huang 2015a,c). In recent years, with the significantly improved computational power, machine learning and data mining have been successfully employed to various domains from science to engineering. In the next section, we provide a different perspective (data-driven approach) of extracting subsurface geological features from seismic measurements.

2.3 Data-Driven Approach for Subsurface Feature characterization

In this paper, we adopt a data-driven approach, which means that we employ machine learning techniques directly to infer the geological features and that no underlying physics is utilized as shown Fig. 1. Specifically, suppose one has n synthetic and/or historical *seismic measurement vectors* $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ and the associated *labels* (which can be the location or angle of geologic faults) $y_1, \cdots, y_n \in \mathbb{R}$. We define

$$X = [\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T]^T \in \mathbb{R}^{n\times d}, \tag{5}$$

$$\mathbf{y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^{n\times 1}. \tag{6}$$



(a) (b)

Figure 3. (a). An illustration of subsurface properties exploration by using seismic wave (image courtesy of Leeuwen (2016)). We see that wavefront propagating beneath the subsurface and a part of it reflected to the receivers. The source is denoted by a red star and the receivers are denoted by green upside-down triangular. (b). The diagram of the data-driven procedure to learn geologic features from seismic data. Simulated seismic measurements are utilized as training data sets, which are fed into the data-driven model. A mapping function, $f^\star(\mathbf{x}')$, is the outcome of the training algorithms. The function, $f^\star(\mathbf{x}')$, is the characterization function, which creates a link from the seismic measurements to the corresponding geological features.

Overall, the idea of data-driven approach independent of applications can be illustrated as

$$\textbf{Physical Measurements} \xrightarrow{f^\star} \textbf{Labels}.$$

In particular, one can build a machine learning model, such as kernel ridge regression (KRR), and *train* the model using the synthetic and/or historical seismic measurements. After training, one gets a function, $f^\star$, which takes a d-dimensional measurement vector as input and returns a prediction of its label. Then for any unseen measurement vector $\mathbf{x}' \in \mathbb{R}^d$, one can predict its label by $f^\star(\mathbf{x}')$.

As for subsurface geological feature characterization specifically, we illustrate our data-driven approach in Fig. 3b. Simulated seismic measurements are utilized as training data sets, which are fed into the data-driven model. A mapping function, $f^\star(\mathbf{x}')$, is the outcome of the training algorithms. The function, $f^\star(\mathbf{x}')$, is the characterization function, which creates a link from the seismic measurements to the corresponding geological features.

Note the difference between a data-driven model and the physical-driven models as in Eq. (4). A data-driven model, such as KRR, is generic: it can be used to predict wine quality, web page click, house price, etc. To apply a data-driven model, one need zero knowledge of the physics behind the problem; one just need to provide the historical measurement vectors and labels for training. This is in

sharp contrast to the physics-driven model in Eq. (4), which is specific to one particular problem and requires strong domain knowledge and intricate mathematical models.

The correctness of our applied data-driven approach, KRR, is ensured by machine learning theory (Friedman et al. 2001; Mohri et al. 2012). Assume that the training and test data are generated by the same model (otherwise, what is learned from the training data does not apply to the test data). As more data are used for training, the prediction error monotonically decreases. Importantly, KRR is known to be robust to noise: even if the training data are corrupted by intensive noise, the prediction is still highly accurate, provided that the number of training data is sufficiently large. The robustness is useful in practice, because the seismic measurements have noise, and the locations and angles of the geologic faults may not be exactly known.

With two different categories of methods introduced ('Data-Driven Methods' V.S. 'Physics-Driven Methods'), it is worthwhile to mention the distinct differences between these two approaches. The problem of recovering the inherent parameters of a system (i.e. inverse problem) can be posed as the problem of regressing those parameters (even thousands) from the input measurements. However, unlike conventional optimization solutions, machine learning solutions have a strong data dependency, which is more severe when the regressing parameters are statistically independent. Though in practice the parameters exhibit strong correlations, the data requirement even for that case is quite high. In contrast, physics-driven methods are usually formulated as inverse problems where a solution vector with a much larger size can be calculated, without an explicit need for training data. In order to ensure the design of robust models with reasonably limited training data, we propose to regress to a small number of critical variables from geophysical systems. Understanding the actual data requirements, even with more sophisticated machine learning techniques such as deep neural networks, in order to perform complete inversion is part of our future work.

2.4 Ridge Regression and Kernel Trick

This work proposes to learn the function in question (denote $f^\star$) using data driven techniques such as ridge regression and kernel ridge regression (KRR). Since all these regression methods are central to the proposed system, we recap their definitions in the following sections.

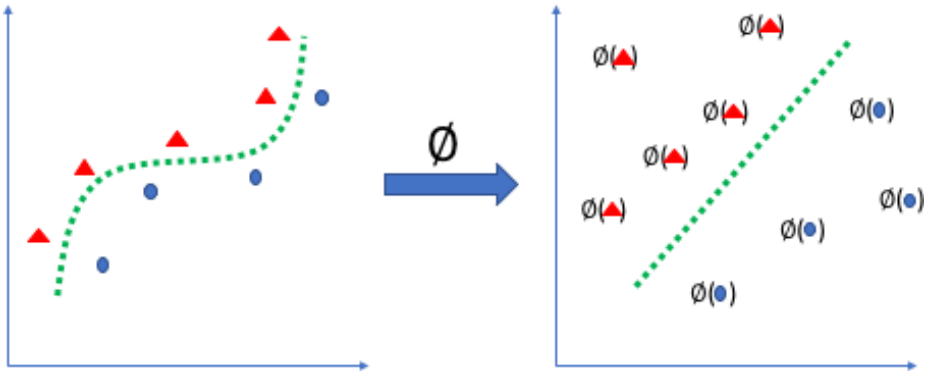


Figure 4. The illustration of the kernel function. The mapping function ϕ embeds the original data into a high dimensional feature space where the nonlinear pattern now appears linear.

2.4.1 Ridge Regression

Ridge regression is one of most popular regression methods, which models the linear dependencies between measurement vectors $\mathbf{x}$ and labels $\mathbf{y}$. Its loss function is usually posed as

$$\min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left\| y_i - \mathbf{w}^T \mathbf{x}_i \right\|_2^2 + \frac{\lambda}{2} \left\| \mathbf{w} \right\|^2 \right\}, \tag{7}$$

where the first term is the cost function and the second term is used to avoid over-fitting. The optimal solution in primal form is

$$\mathbf{w}^\star = (X^T X + \lambda I)^{-1} X^T \mathbf{y}. \tag{8}$$

A prediction can be made by

$$f^\star(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}^\star \rangle = \mathbf{x}^T \mathbf{w}^\star. \tag{9}$$

The major shortcoming of ridge regression is its limitation in modeling nonlinear data sets. In seismic applications, the relationship between the measurement vectors and the labels is nonlinear because of the governing physics as provided in Eqs. (1) and (2). We need more advanced regression techniques to model the data nonlinearity while maintaining feasible computational costs. Kernel tricks provide us with the tools (Schölkopf & Smola 2002).

2.4.2 Kernel Ridge Regression

2.4.2.1 Kernel Trick In our problem, the relationship between data and labels is nonlinear. Employing linear regression will be insufficient to detect nonlinear pattern. We therefore consider an embedding map

$$\phi : \mathbf{x} \in \mathbb{R}^m \longrightarrow \phi(\mathbf{x}) \in \mathbb{R}^M, \tag{10}$$

where the mapping function ϕ embeds the original data residing in the low dimensional space of $\mathbb{R}^m$ into a high dimensional space of $\mathbb{R}^M$, where the nonlinear pattern now appears linear as illustrated in Fig. 4. In such a way, we can use the linear regression algorithms to detect the pattern in the higher feature space. Intuitively, the form of solution to the kernel ridge regression will be the same except the replacement of $\mathbf{x}$ by $\phi(\mathbf{x})$ in Eqs. (8) and (9). However, due to the fact that $\phi(\mathbf{x})$ can reside in a very high dimensional space or even infinite space. A direct replacement of $\phi(\mathbf{x})$ becomes infeasible. Kernel trick can be therefore utilized to avoid the usage of the $\phi(\mathbf{x})$ explicitly (Schölkopf & Smola 2002). In particular, the solution to the ridge regression in Eq. (8) can be reformulated as its equivalent dual form

$$\begin{aligned} \mathbf{w} &= X^T\lambda^{-1}(\mathbf{y} - X\mathbf{w}), \\ &= X^T\boldsymbol{\alpha}, \end{aligned} \tag{11}$$

where $\boldsymbol{\alpha} = \lambda^{-1}(\mathbf{y} - X\mathbf{w})$. To solve for $\boldsymbol{\alpha}$, we have

$$\begin{aligned} \boldsymbol{\alpha} &= \lambda^{-1}(\mathbf{y} - X\mathbf{w}), \\ &= \lambda^{-1}(\mathbf{y} - XX^T\boldsymbol{\alpha}). \end{aligned}$$

Hence, we will have

$$\boldsymbol{\alpha} = (XX^T + \lambda I)^{-1}\mathbf{y}. \tag{12}$$

The dual solution to the ridge regression in Eq. (7), can be obtained through Eqs. (11) and (12). With the embedding function definied in Eq. (10), we will have the kernel ridge regression (KRR) as

$$\begin{aligned} \boldsymbol{\alpha} &= (\phi(X)\phi(X)^T + \lambda I)^{-1}\mathbf{y}, \\ &= (K + \lambda I)^{-1}\mathbf{y}, \end{aligned}$$

where $K = \phi(X)\phi(X)^T$. In practical, we do not need to obtain the $\phi(X)$ explicitly. We just need to calculate the inner product of $\phi(X)\phi(X)^T$, which is called the kernel trick. A kernel function of κ is therefore provided.

Definition A function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \quad \text{for all } \mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d \text{ and } z_1, \cdots, z_n \in \mathbb{R}.$$

In addition, a valid kernel defines such a feature map $\phi : \mathbb{R}^d \mapsto \mathcal{F}$ that $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$.

The in-equality relationship given in the Definition is guaranteed, provided with a positive semi-definite kernel (Schölkopf & Smola 2002).

2.4.2.2 Ridge Regression with Radial Basis Function Kernel In general, the kernel $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ measures the similarity between the two samples in $\mathbb{R}^d$. There are three types of kernel functions mostly used in various applications: linear function kernel, polynomial function kernel, and the radial basis function (RBF) kernel. Because of the nonlinear data pattern, linear kernel function does not fit in our problem. The complexity of the model using polynomial function kernel are limited by its polynomial degree. In contrast, the model complexity based on RBF kernel is potentially infinite. Therefore, with incremental size of data sets, model with polynomial kernel will be saturated, while model with RBF kernel will be able to represent the complex relationship. Considering these, we use RBF as the kernel which is defined as

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) \;=\; \exp\big(-\tfrac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\big), \tag{13}$$

where $\sigma > 0$ is the kernel width parameter and $\|\cdot\|_2$ is the vector Euclidean norm (ℓ_2-norm).

Suppose the seismic measurements are stored as $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. KRR uses the data for training and returns a function $\hat{f}$ which approximates $f^\star$. Given a test point $\mathbf{x}' \in \mathbb{R}^d$, KRR makes prediction $\hat{f}(\mathbf{x}')$. We directly state the dual problem of KRR without derivation; readers can refer to Campbell (2001) for the details:

$$\min_{\boldsymbol{\alpha}} \left\{ \frac{1}{2}\sum_{i=1}^{n} \big\|y_i - (K\boldsymbol{\alpha})_i\big\|_2^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^T K\, \boldsymbol{\alpha} \right\}, \tag{14}$$

where $\lambda > 0$ is the regularization parameter and should be fine tuned. As we illustrated in the previous section, problem (14) has a closed-form optimal solution

$$\boldsymbol{\alpha}^\star \;=\; (K + \lambda I_n)^{-1}\mathbf{y} \;\in\; \mathbb{R}^n, \tag{15}$$

where I_n is the $n \times n$ identity matrix. Finally, for any $\mathbf{x}' \in \mathbb{R}^d$,

$$\hat{f}(\mathbf{x}') \;=\; \sum_{i=1}^{n} \boldsymbol{\alpha}_i^\star \kappa(\mathbf{x}', \mathbf{x}_i) \tag{16}$$

is the prediction made by KRR.

Machine learning theory indicates that more training samples lead to smaller variance and thereby better prediction performance. Ideally, one can collect as many seismic measurements as desired in the quest to improve characterization. Unfortunately, the $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory costs of

KRR hinder the use of such large amounts of training data. To the best of our knowledge, these computational challenges of KRR have not been addressed by any of the prior efforts on using kernel machines for subsurface applications (Schnetzler & Alumbaugh 2017; Zhang et al. 2014; Ramirez & Meyer 2011). A practical approach to large-scale KRR is randomized kernel approximation, which sacrifices a limited amount of accuracy for a tremendous reduction in time and memory costs. In this work, we apply the Nyström method (Nyström 1930; Williams & Seeger 2001) to make large-scale KRR feasible on a single workstation. Consequently, we can easily enable the training of KRR using much larger amounts seismic measurements, thereby achieving substantially improved geologic characterization performance.

2.4.3 Multivariate Regression Model for Multiple Predictions

With the variables provided in Eqs. (5) and (6), the prediction of our data-driven model is a scalar value, which is either the location or the angle of the geologic fault zone. In order to apply our techniques to prediction of multiple geologic fault zones, multivariate regression models should be utilized (Hidalgo & Goodman 2013), which is a direct extension of the technique discussed in this paper. Further, we can incorporate additional constraints to model potential correlations between different output variables, in the multi-variate case. In particular, instead of using features and labels provided in Eqs. (5) and (6), we will have

$$X = [\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T]^T \in \mathbb{R}^{n \times d}, \tag{17}$$

$$Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times l}, \tag{18}$$

where the additional dimension of l corresponds to the size of the prediction, i.e., the number of fault zones. Correspondingly, the ridge regression provided in Eq. (7) will be modified as a least-squares minimization with multiple right-hand sides problem

$$\min_W \left\{ \frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{y}_i - W^T \mathbf{x}_i \right\|_F^2 + \frac{\lambda}{2} \left\| W \right\|_F^2 \right\}, \tag{19}$$

where $W \in \mathbb{R}^{d \times l}$, and $\left\| \cdot \right\|_F$ is the Frobenius norm. The remaining derivations for predictions can be obtained similarly. As for the focus of this paper, we concentrate on the regression model set up as Eqs. (5) and (6). Readers who are interested in multivariate regression models can utilize a straightforward extension of our techniques.

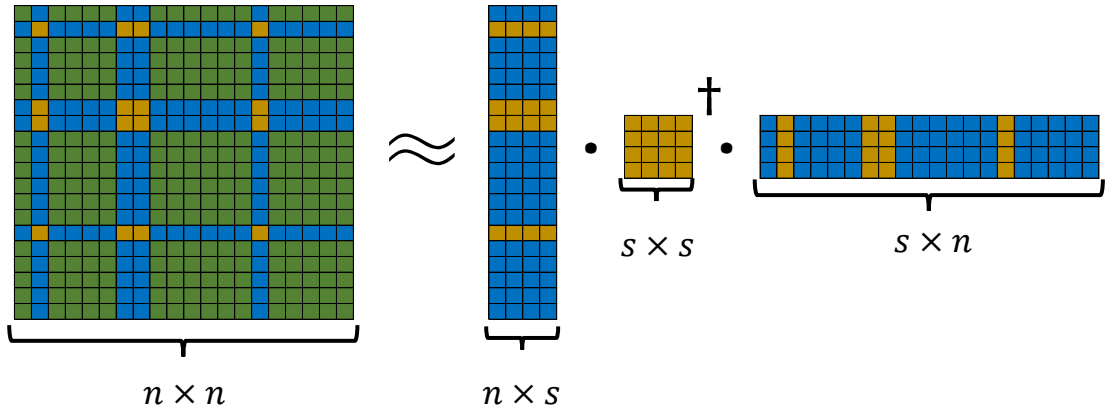


Figure 5. Illustration of the Nyström approximation $K \approx \Psi\Psi^T = CW^\dagger C^T$, where the low-rank approximation $\Psi = C(W^\dagger)^{1/2}$ can be obtained.

3 SCALABLE GEOLOGIC CHARACTERIZATION THROUGH RANDOMIZED APPROXIMATION

In this section, we introduce the Nyström method—a randomized kernel matrix approximation tool—to the geologic characterization task, aiming at solving large-scale problems using limited computational resources. Sec. 3.1 describes the Nyström method, Sec. 3.2 theoretically justifies the Nyström method and its application to KRR, Sec. 3.3 discusses the three tuning parameters, Sec. 3.4 presents the whole procedure of KRR with Nyström approximation, and finally, Sec. 3.5 analyzes the time and memory costs.

3.1 The Nyström Method

The Nyström method (Williams & Seeger 2001) is a popular and an efficient approach. In addition to its simplicity, the Nyström method is a theoretically sound approach: its approximation error is bounded (Drineas & Mahoney 2005; Gittens & Mahoney 2016); when applied to KRR, its statistical risk is also theoretically guaranteed (Alaoui & Mahoney 2015; Bach 2013).

The Nyström method computes a low-rank approximation $K \approx \Psi\Psi^T \in \mathbb{R}^{n\times n}$ in $\mathcal{O}(nds + ns^2)$ time. Here $s \ll n$ is user-specified; larger values of s leads to better approximation but incurs higher computational costs. The tall-and-skinny matrix $\Psi \in \mathbb{R}^{n\times s}$ is computed as follows: First, sample s items from $\{1, \cdots, n\}$ uniformly at random without replacement; let the resulting set be $\mathcal{S}$. Subsequently, construct a matrix $C \in \mathbb{R}^{n\times s}$ as $c_{il} = \kappa(\mathbf{x}_i, \mathbf{x}_l)$ for $i \in \{1, \cdots, n\}$ and $l \in \mathcal{S}$; let $W \in \mathbb{R}^{s\times s}$ contain the rows of C indexed by $\mathcal{S}$. Gittens & Mahoney (2016) showed that $CW^\dagger C^T$ is a good approximation to K, where $W^\dagger$ denotes the Moore-Penrose pseudo-inverse of W. The approximation is illustrated in Fig. 5. Finally, the low-rank approximation $\Psi = C(W^\dagger)^{1/2}$ is computed.

Besides the Nyström method, a number of other kernel approximation methods exist in the machine learning literature. Random feature mapping (Le et al. 2013; Rahimi & Recht 2007) is an equally popular class of approaches. However, compared to random feature mapping, several theoretical and empirical studies (Tu et al. 2016; Yang et al. 2012) are in favor of the Nyström method. Furthermore, in the recent years, alternative approaches such as the fast SPSD model (Wang et al. 2016), MEKA (Si et al. 2014), hierarchically compositional kernels (Chen et al. 2016) have been proposed to speed up KRR. Since comparing different kernel approximation methods is beyond the scope of this work, we adopt the Nyström method in our algorithm.

3.2 Theoretical Justifications of the Nyström Method

The Nyström method has been studied by many recent works (Alaoui & Mahoney 2015; Bach 2013; Drineas & Mahoney 2005; Gittens & Mahoney 2016; Wang et al. 2016, 2017), and its theoretical properties has been well understood. In the following, we first intuitively explain why the Nyström method works and then describe its theoretical properties.

Let $P \in \mathbb{R}^{n\times s}$ be a column selection matrix, that is, each column of P has exactly one nonzero entry whose position indicates the index of the selected column. We let $K = K^{\frac{1}{2}}K^{\frac{1}{2}}$ and $D = K^{\frac{1}{2}}P$. Then the matrices C and W (in Fig. 5) can be written as

$$C = KP = K^{\frac{1}{2}}K^{\frac{1}{2}}P = K^{\frac{1}{2}}D \quad \text{and} \quad W = P^TKP = P^TK^{\frac{1}{2}}K^{\frac{1}{2}}P = D^TD.$$

The Nyström approximation can be written as

$$K \approx CW^{\dagger}C^T = K^{\frac{1}{2}}\big[D(D^TD)^{\dagger}D\big]K^{\frac{1}{2}}.$$

In the extreme case where $s = n$, the matrix $D(D^TD)^{\dagger}D$ is the identity matrix I_n, and thus the Nyström approximation is exact. In general $s < n$, the matrix $D(D^TD)^{\dagger}D$ is called orthogonal projection matrix, which projects any matrix to the column space of D. Low-rank approximation theories show that if the "information" in K is spread-out, then most mass of $K^{\frac{1}{2}}$ are in the column space of a small subset of columns of $K^{\frac{1}{2}}$. Therefore, projecting $K^{\frac{1}{2}}$ to the column space of $D = K^{\frac{1}{2}}P$ loses only a little accuracy, and the Nyström approximation $K^{\frac{1}{2}}\big[D(D^TD)^{\dagger}D\big]K^{\frac{1}{2}}$ well approximates K.

Theoretical bounds (Gittens & Mahoney 2016; Wang et al. 2017) guarantee that the Nyström approximation $CW^{\dagger}C^T$ is close to K in terms of matrix norms. Let r (≥ 1) be arbitrary integer, K_r be the best rank r approximation to K, and $\|\cdot\|$ be the spectral norm, Frobenius norm, or trace norm.

If the eigenvalues of K decays rapidly and the number of samples, s, is sufficiently larger than r, then $\|K - CW^{\dagger}C^T\|$ is comparable to $\|K - K_r\|$.

Applied to the KRR problem, the quality of the Nyström method has been studied by Alaoui & Mahoney (2015); Bach (2013). The works studied the bias and variance, which directly affect the prediction error of KRR. The works showed that using the Nyström approximation, the increases in the bias is bounded, and the variance does not increase at all. Therefore, using the Nyström approximation, the prediction made by KRR will not be much affected. In addition, they showed that as the number of samples, s, increases, the performance monotonically improves.

3.3 Tuning Parameters

KRR with Nyström approximation has totally three parameters: the regularization parameter λ, the kernel width parameter σ, and the number of random samples s. We discuss the effect of the parameters.

The regularization parameter λ (≥ 0) is defined in the KRR objective function (14) and can be arbitrarily set by users. From the statistical perspective, λ trades off the bias and variance of KRR: big λ leads to small variance but big bias, and vice versa. The optimal choice of λ is the one minimizes the sum of variance and squared bias. However, such optimal choice cannot be analytically calculated; in practice, it is determined by cross-validation.*

The kernel width parameter σ is defined in (13). It defines how far the influence of a single training example reaches, with high values meaning "far" and low values meaning "close". As σ goes to zero, K tends to be identity, where the training examples do not influence each other and the KRR model is too flexible; as σ goes to infinity, K tends to be an all-one matrix (its rank is one), where the KRR model is restrictive and lacks expressive power. In practice, σ should be fine tuned; a good heuristic is setting σ to

$$\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}. \tag{20}$$

or searching σ around this value by cross-validation. Note that computing (20) costs $\mathcal{O}(n^2 d)$ time and is thereby impractical, a good heuristic is randomly sample a subset $\mathcal{J} \subset \{1, \cdots, n\}$ and approximate (20) by

$$\sqrt{\frac{1}{|\mathcal{J}|^2}\sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J}}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2},$$

which costs merely $\mathcal{O}(d|\mathcal{J}|^2)$ time.

* Cross-validation is a standard machine learning technique for tuning parameters. One can randomly split the training set into two parts, train on one part, make prediction on the other, and choose the parameter corresponding to the best prediction error.

The number of random samples s trades off the accuracy and computational costs: large s leads to good prediction but large computational costs. If the dataset has n samples of d-dimension, the total time complexity is $\mathcal{O}(nds + ns^2)$, and the space (memory) complexity is $\mathcal{O}(nd + ns)$. It is always good to set s as large as one can afford because the prediction monotonically improves as s increase.

3.4 Overall Algorithm: KRR with Nyström Approximation

Using the Nyström method, the training of KRR can be performed in $O(nds + ns^2)$ time, where the user-specified parameter s directly trades off accuracy and computational costs. Empirically speaking, setting s in the order of hundreds suffices in our application. The overall algorithm is described as follows.

Training. The inputs are $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. User specifies s, randomly select s out of the n samples, form the kernel sub-matrices $C \in \mathbb{R}^{n\times s}$ and $W \in \mathbb{R}^{s\times s}$, and compute $\Psi = CW^{-\frac{1}{2}}$. The kernel matrix K can be approximated by $\Psi\Psi^T$ according to the previous subsection. Finally, $\alpha^\star$ defined in Eq. (15) can be approximated by

$$\tilde{\boldsymbol{\alpha}} = \left(\Psi\Psi^T + \lambda I_n\right)^{-1} y \tag{21}$$

$$= \lambda^{-1} y - \lambda^{-1}\Psi(\lambda I_s + \Psi^T\Psi)^{-1}\Psi^T y \ \in \ \mathbb{R}^n, \tag{22}$$

where the latter equality follows from the Sherman-Morrison-Woodbury (SMW) matrix identity as defined in Eq. (A.1) and the detailed derivation is provided from Eq. (A.2) to Eq. (A.3). More details can be found in Wang (2015). It is worthwhile to mention that the $n \times n$ matrix of $\Psi\Psi^T$ in Eq. (21) has been replaced by the matrix of $\Psi^T\Psi$ in Eq. (22), which is a much smaller dimension of $s \times s$. This leads to the significant reduction of the computational costs.

Prediction. Let $\mathbf{x}' \in \mathbb{R}^d$ be any unseen test sample. The characterization step is almost identical to Eq. (16): we use $\tilde{\alpha}$ instead of $\alpha^\star$ and makes prediction by

$$\hat{f}(x') \ = \ \sum_{i=1}^{n} \boldsymbol{\alpha}_i^\star \kappa(\mathbf{x}', \mathbf{x}_i) \ \approx \ \sum_{i=1}^{n} \tilde{\boldsymbol{\alpha}}_i \kappa(\mathbf{x}', \mathbf{x}_i). \tag{23}$$

The location or angle of geological fault should be close to $\hat{f}(\mathbf{x}')$.

3.5 Computational and Memory Cost Analysis

The training of KRR without kernel approximation has $\mathcal{O}(n^2d+n^3)$ time complexity and $\mathcal{O}(nd+n^2)$ space (memory) complexity. The costs are calculated as follows. For most kernel functions, including the RBF kernel, the evaluation of $\kappa(x_i, x_j)$ costs $\mathcal{O}(d)$ time. One needs to keep the n data samples

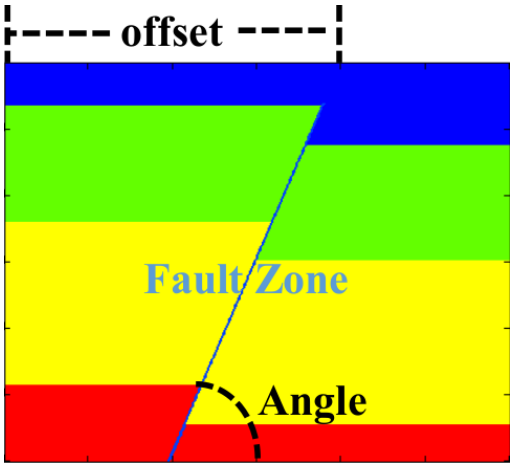


Figure 6. An illustration of the geologic fault zone. The location of a geologic fault zone can be characterized by its horizontal offset and the dipping angle (Zhang et al. 2014).

in memory to compute every entry of $K \in \mathbb{R}^{n\times n}$, which costs $\mathcal{O}(nd)$ memory and $\mathcal{O}(n^2 d)$ time. To compute $\alpha^\star$, one needs to keep K in memory and perform matrix inversion, which requires $\mathcal{O}(n^2)$ memory and $\mathcal{O}(n^3)$ time.

The training of KRR with Nyström approximation has $\mathcal{O}(nds+ns^2)$ time complexity and $\mathcal{O}(nd+ns)$ space complexity. The costs are calculated as follows. To compute $C \in \mathbb{R}^{n\times s}$ and $W \in \mathbb{R}^{s\times s}$, one only need to evaluate ns kernel functions, which requires $\mathcal{O}(nd)$ memory and $\mathcal{O}(nds)$ time. The computation of $\tilde{\alpha}$ according to Eq. (22) has $\mathcal{O}(ns)$ space complexity (because the matrices C and W need to be kept in memory) and $\mathcal{O}(ns^2)$ time complexity.

The prediction of KRR, either with or without approximation, for an unseen test sample, $\mathbf{x}'$, has $\mathcal{O}(nd)$ time complexity and $\mathcal{O}(nd)$ memory complexity. First, one keeps the n data samples in memory to evaluate the kernel functions $\kappa(\mathbf{x}', \mathbf{x}_1), \cdots, \kappa(\mathbf{x}', \mathbf{x}_n)$, which costs $\mathcal{O}(nd)$ time and $\mathcal{O}(nd)$ memory. Then, one keeps the n kernel function values and $\boldsymbol{\alpha}^\star$ (or $\tilde{\boldsymbol{\alpha}}$) in memory to make prediction, which costs merely $\mathcal{O}(n)$ time and $\mathcal{O}(n)$ memory.

4 NUMERICAL RESULTS

To validate the performance of our proposed approach, we carry out evaluations with synthetic seismic measurements to characterize the location of the geologic faults. The siting of geologic fault zones can be characterized by its horizontal offset and the dipping angle as shown in Fig. 6 (Zhang et al. 2014). We employ our new subsurface feature characterization method to estimate both the offset and angle of geologic fault zones. As for the computing environment, we run our tests on a computer with 48 Intel Xeon E5-2650 cores running at 2.3 GHz, and 64 GB memory.

The quality of training set is critical for any data-driven model. In this work, we consider velocity

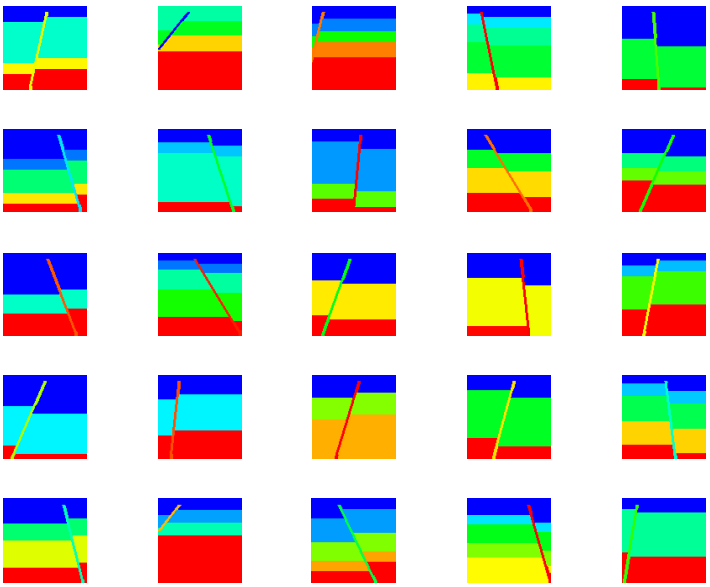


Figure 7. A database of velocity models consisting of $60,000$ models of size 100×100 grid points. The velocity models in the database are different from one another in terms of offset (ranging from 30 grids to 70 grids), dipping angle (ranging from 25° to 165°), number of layers (ranging from 3 to 5 layers), layer thickness (ranging from 5 grids to 80 grids), and layer velocity (ranging from 3000 m/s to 5000 m/s).

models consisting of horizontal reflectors with a single fault zone (layer model) to demonstrate the performance of our new geologic feature characterization method. It is straightforward to employ our method to characterize multiple fault zones. To best represent the geologically realistic velocity models, we create a database containing $n = 60,000$ velocity models of size 100×100 grid points similar to Zhang et al. (2014). The velocity models in the database are different from one another in terms of offset (ranging from 30 grids to 70 grids), dipping angle (ranging from 25° to 165°), number of layers (ranging from 3 to 5 layers), layer thickness (ranging from 5 grids to 80 grids), and layer velocity (ranging from 3000 m/s to 5000 m/s). A small portion of the training velocity models are shown in Fig. 7.

The seismic measurements are collections of synthetic seismograms obtained by implementing forward modeling on those $60,000$ velocity models. One common-shot gather of synthetic seismic data with 32 receivers is posed at the top surface of the model. The receiver interval is 15 m. We use a Ricker wavelet with a center frequency of 25 Hz as the source time function and a staggered-grid finite-difference scheme with a perfectly matched layered absorbing boundary condition to generate 2D synthetic seismic reflection data (Tan & Huang 2014; Zhang & Shen 2010). The synthetic trace at each receiver is a collection of time-series data of length $1,000$. So, the dimension of seismic measurement data is $d = 3.2 \times 10^4$. Therefore, out of $60,000$ velocity models, the total volume of synthetic seismic data is 1.92×10^9. In Fig. 8, we show a portion of the synthetic seismic data sets

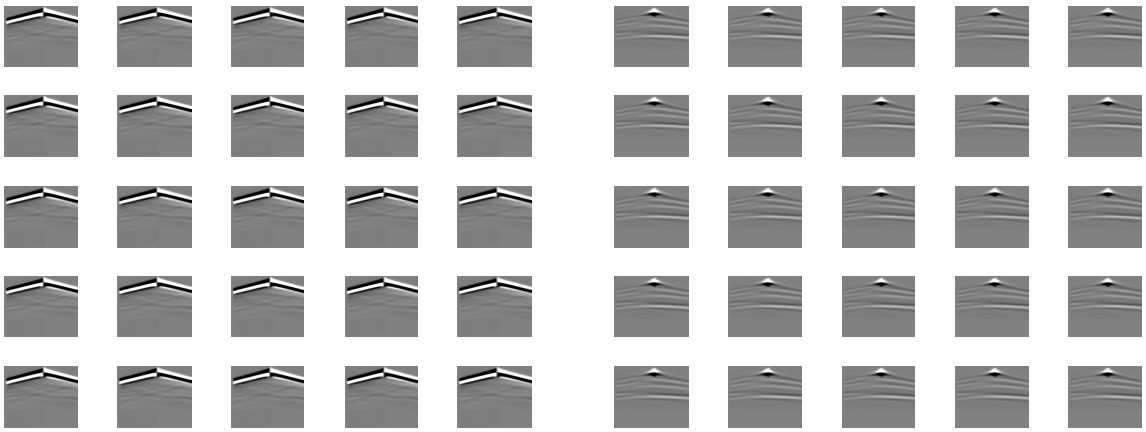


Figure 8. Synthetic seismic data sets are obtained using a staggered-grid finite-difference scheme with a perfectly matched layered absorbing boundary condition. The displacement of X direction (a) and Z direction (b) are both used as training sets. The total volume of synthetic seismic data is 1.92×10^9.

corresponding to velocity models that we generate. Specifically, the displacement in the X direction is shown in Fig. 8(a), and the displacement in the Z direction is shown in Fig. 8(b).

We employ a hold-out test to assess the efficacy of our proposed algorithm. Specifically, 75.0% of the dataset is used for training the model, while the rest is used for testing. For comparison, we use the conventional KRR method (denoted by "KRR") as the reference method. We denote our new geologic feature characterization method as "R-KRR" standing for Randomized KRR method. To evaluate the performance, we report both the accuracy and the computational efficiency of different methods. We use the mean-absolute error (MAE) metric to quantify the accuracy of a data-driven model, which is defined as

$$\mathrm{MAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}|. \tag{24}$$

We record the wall clock time to measure the computational efficiency of a method and further provide the speed-up ratio.

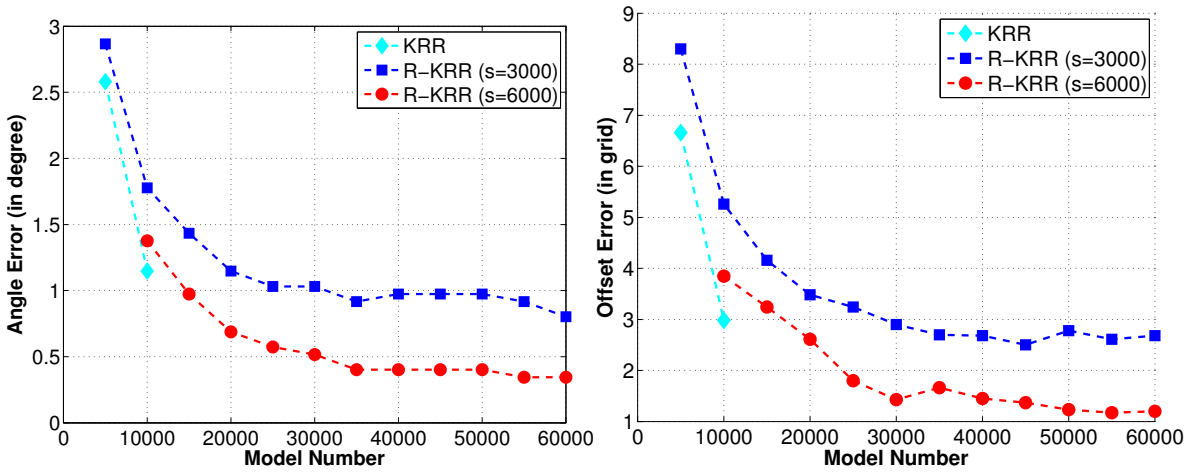
To have a comprehensive understanding our randomized geologic feature characterization methods, we provide three sets of tests. In Sec. 4.1, we provide an overall test of the characterization accuracy of our method. In Sec. 4.2, we report the performance of our method as a function of the number of random samples, s. In Sec. 4.3, we test the robustness of our method with a view on the randomness of the approximation method.

4.1 Test on Characterization Accuracy

We provide our first test on the characterization accuracy. The estimation result of the dipping angle and horizontal offset is provided in Fig. 9. We test the performances of our method using two different

Nyström approximations, $s = 3,000$ and $s = 6,000$, as well as one other characterization approach using conventional KRR method. We report the performances of those methods using different sizes of the seismic data. Specifically, we increase the seismic dataset generated from $5,000$ velocity models to $60,000$ velocity models with an incremental of $5,000$ velocity models. The corresponding MAE values are reported in Fig. 9. In particular, the results of angle estimation is provided in Fig. 9(a) and the results of the offset estimation is provided in Fig. 9(b). We notice when the dataset used for training is small, KRR method (in cyan) yields more accurate results of both angle and offset estimations. This is reasonable since all the available data sets are used for estimation. After using data from $10,000$ velocity models, KRR method becomes extremely inefficient because of the selection of the parameters using cross-validation. It is difficult to evaluate its performance given more training data. While, our method with both Nyström approximations, $s = 3,000$ and $s = 6,000$, still yields accurate results and efficient performance. In particular, our method with larger Nyström approximation, i.e., $s = 6,000$, consistently gives us better results. Our best estimate of the dipping angle on the full seismic data set is 0.5° (Fig. 9(a)). Similarly, we also report the performance of offset estimation in Fig. 9(b). The best estimate of the offset using our method on the full data set is about 1 grid.

The total computing time includes data generation, training and prediction phases. We show in Figure 10 the comparison of the computing times for the training phase using our method (in red and blue), the conventional KRR method (in cyan) and the time used to generate the required data (in magenta). In data generation, majority part of the time will be spent on seismic modeling. Based on our hardware, it takes roughly ~1.0 second to run a full seismic modeling. Provided with a 48-core processor, it will take ~104.0 seconds to generate the first $5,000$ sets of seismic data. It is reasonable to assume that the computing time in data generation scales linearly with the number of velocity models. We have all the computing time costs needed for different numbers of the velocity models as shown in Fig. 10. By comparing to training times needed for three different scenarios (conventional KRR, our methods with two different s sizes), the time in training required by the conventional KRR is much larger than the one in data generation (except for very small data size); while our methods needs much less computing time in training than the one in data generation. To further compare the training phases of all three scenarios in Fig. 10, KRR approach is much more computationally expensive and memory demanding than our method even if it provides a slightly more accurate estimation. On the other hand, our method is significantly more efficient than the conventional KRR method in training when the data sets become large. Utilizing the full dataset, it takes our method on the order of 10 seconds to train the prediction model. The speed-up ratios between our method and the conventional KRR method in the training phase are up to $1,000$. The durations of predictions for conventional KRR and our methods are comparable, which are provided in Eq. (23). In our experiments, it usually takes less than



(a) (b)

Figure 9. Estimation error for (a) dipping angle and (b) offset using conventional KRR method (in cyan), our method using $s = 3,000$ (in blue), and our method using $s = 6,000$ (in red). KRR method (in cyan) yields more accurate results of both angle and offset estimations with small size of data sets, and it fails to provide estimation when data sets becomes too large. Our method yields consistently comparable results to KRR on all sizes of data sets.

a few seconds to produce a prediction of the dipping angle or offset. To summarize, such an efficiency provided by our method would allow the possibility to characterize the geologic features in/towards real time. Though the computational time in estimating the offset is not reported in the paper, similar conclusions can be drawn on the accuracy and computational efficiency of our method.

To have a visualization of our estimation, we provide a specific example of the true model and our estimation in Fig. 11. In Fig. 11(a), we show our true velocity model with angle = 79.1° and offset = 49.0. The estimation result of our randomized characterization method is given in Fig. 11(b). The result of our estimation is angle = 79.0° and offset = 50.0. Visually, our randomized characterization method yields a rather accurate estimation compared to the ground truth.

4.2 Test on the Nyström Sample Size

The number of random Nyström sample size, s, is critical to the accuracy and efficiency of our randomized feature characterization method. The appropriate selection of the Nyström sample size value depends on the redundancy of data sets, which theoretically can be justified by the spectrum spanned by the singular vectors of the data sets. In this test, we provide our estimation results by varying the Nyström sample size, s, from from $1,000$ to $6,000$ with an incremental of 500. Besides the acoustic seismic data sets, we also generate elastic seismic data sets for testing our prediction model. The estimation results are provided in Fig. 12, where Fig. 12(a) is the estimation of horizontal offset and Fig. 12(b) is the estimation of the dipping angle. In both figures, the estimation results using acoustic data sets are plotted in red, and the results using elastic data sets are plotted in blue. We notice that in

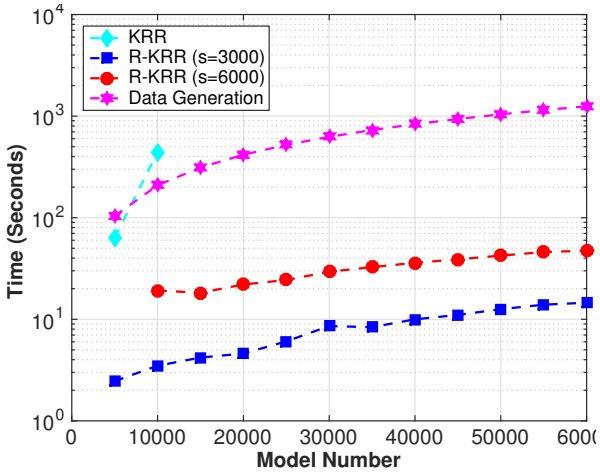
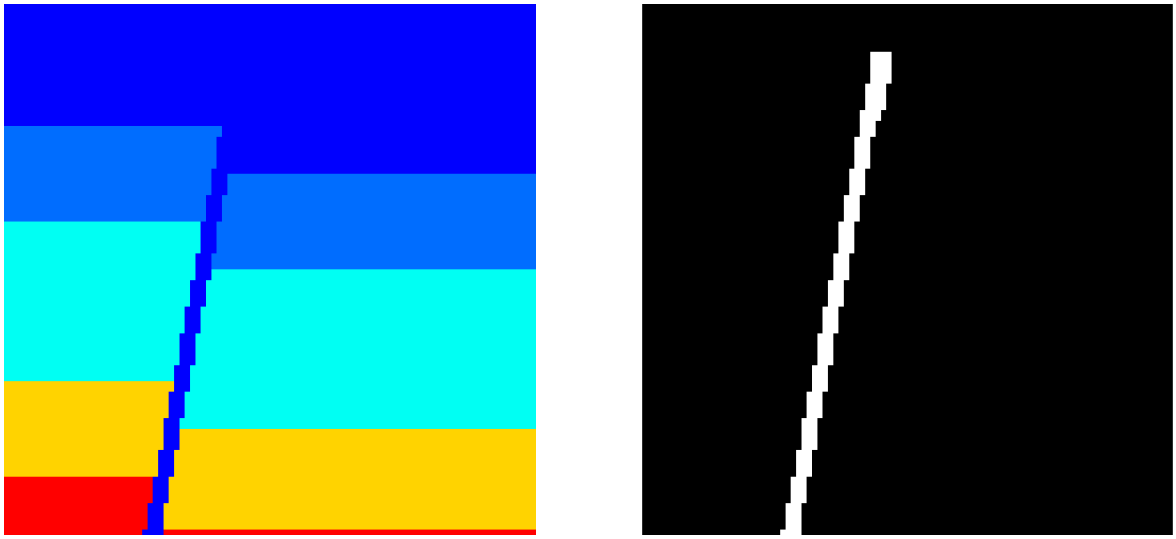


Figure 10. The comparison of the computing times for the training phase using our method (in red and blue), the conventional KRR method (in cyan) and the time used to generate the required data (in magenta). Our method yields accurate results and is computationally and memory efficient on all data points.

both figures that with the increase of the Nyström sample size, the estimation accuracy for both dipping angle and horizontal offset also increases. This is reasonable and can be explained by the fact that more information is used for generating the prediction model. Comparing the estimation results using elastic and acoustic seismic data sets, we notice that the one using acoustic seismic data sets yields consistently more accurate results. This is because the elastic models include much more parameters than the acoustic models, which is indicated by Eqs. (1) and (2). With more degree of freedom, more training data are therefore needed to achieve the same level of accuracy. To conclude on the selec-



(a) (b)

Figure 11. An synthetic model with a geologic fault in it: (a) the true model with angle = 79.1° and offset = 49.0; (b) the estimation using our new randomized characterization method. The result of our estimation is angle = 79.0° and offset = 50.0. Visually, our randomized characterization method yields a rather accurate estimation compared to the ground truth.

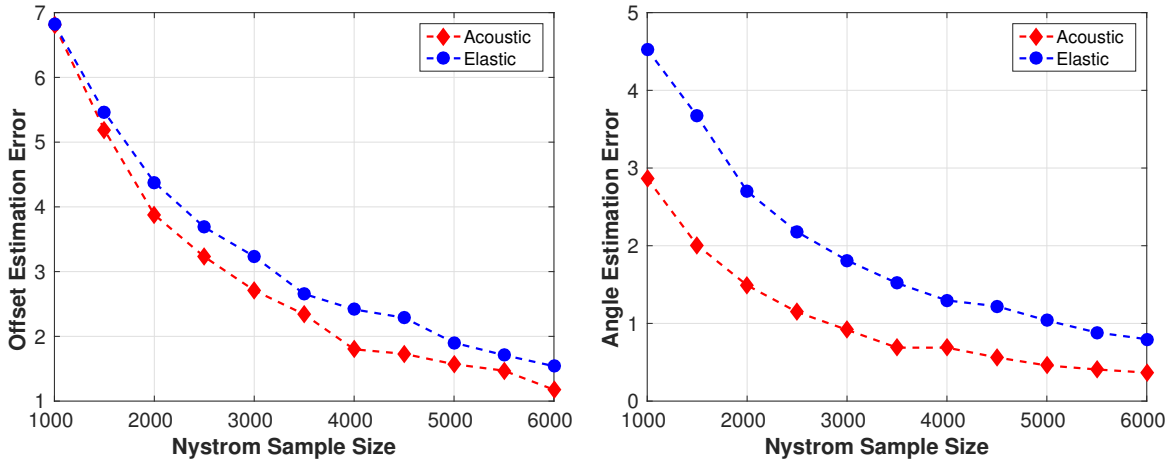


Figure 12. The estimation results by varying the Nyström sample size, s, from from $1,000$ to $6,000$ with an incremental of 500. Both the estimation results on the horizontal offset (a) and dipping angle (b) are provided. In both figures, the estimation results using acoustic data sets are plotted in red, and the results using elastic data sets are plotted in blue. We notice that in both figures that with the increase of the Nyström sample size, the estimation accuracy for both dipping angle and horizontal offset also increases, which is due to the fact that more information are utilized in generating the prediction model.

tion of the random Nyström sample size, we would suggest using a value in between $3,000$ to $6,000$ considering a balance between the accuracy and efficiency.

4.3 Test on the Randomness of the Nyström Method

The Nyström method is a randomization-based approach, where the randomness arises from the uniform sampling of the columns in generating the low-rank approximation as in illustrated Fig. 5. Here we test the randomness in the prediction made by KRR with Nyström approximation. We use the same model as in Test 1, where the size of the velocity models is 100×100 grid points. One geological fault zone is contained in the model. One common-shot gather of synthetic seismic data with 32 receivers is posed at the top surface of the model. We generate 20 different realization tests of the Nyström method. Each of them is drawn from a uniform distribution. We calculate their dipping angle and horizontal estimation errors according to Eq. (24). For all the tests, we set the Nyström sample size, $s = 3,000$, and use the full data set size. We report the randomness results in Fig. 13, where the acoustic estimation results are plotted in red and the elastic results are plotted in blue. We observe that there are two clusters of data points corresponding to the acoustic and elastic scenarios. All of the 20 different realizations lead to similar error estimations of both dipping angle and horizontal offset. Also, we notice that the estimation error of elastic cases is higher than that of the acoustic cases, and this is due to the difference of complexity in the governing equations. The elastic equation yields much higher variability than the acoustic equation, which means more training data are needed to account for the larger degree of freedom. In another word, with the same amount of data, the estimations of

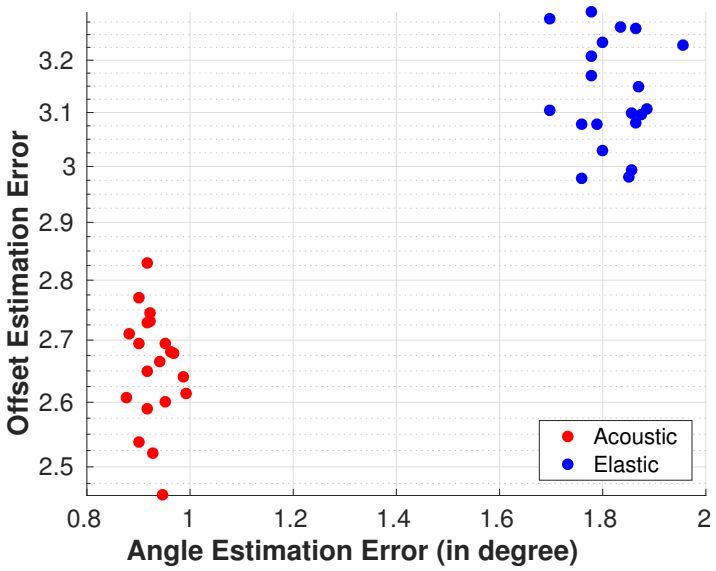


Figure 13. 20 different realization tests of the Nyström Method are generated. Each of them is drawn from a uniform distribution. Both the dipping angle and horizontal estimation errors according to Eq. 24 are calculated. For all the tests, we set the Nyström Sample Size, $s = 3,000$, and use the full data set size. We report the randomness results in Fig. 13, where the acoustic estimation results are plotted in red and the elastic results are plotted in blue. All of the 20 different realizations lead to similar error estimations of both dipping angle and horizontal offset.

acoustic case will yield higher accuracy with those of the elastic case. From this test, we conclude that our method yields robust and accurate results regardless of the randomness nature of the Nyström method.

5 DISCUSSIONS AND FUTURE WORK

In our future research work, we will address the following directions including (1). generalization to real data, (2). the incorporation of prior knowledge, and (3). detection task of geologic features and uncertainty analysis.

(i) Generalization to Real Data

It is not only challenging but also important to verify the utility of our methods with real data. One can employ a number of strategies, along with the proposed inferencing techniques, to ensure the success of these methods under more realistic settings. (1) **Data augmentation**: A straightforward strategy is to augment the training dataset with realistic variants of the synthetic data samples. Currently, we impose a small amount of Gaussian noise to the synthetic seismic data. One can instead utilize a more sophisticated Gaussian Mixture Modeling (GMM) (Stergiopoulos 2017) based noise distributions to build datasets that are more reflective of real-world measurement scenarios; (2) **Feature learning**: A common approach in machine learning to facilitate improved generalization is to engineer highly representative features that are both critical to the prediction and more broadly applicable to the entire data

distribution. The characterization method developed in this paper is an end-to-end approach, which is simpler to implement. While end-to-end learning methods are elegant, they can benefit from domain experts through the design of meaningful feature representations. Therefore, with carefully engineered features, our algorithms can be more effective with respect to real data; (3) **Domain adaptation**: Finally, when there is a mismatch between the data distribution used for training and the one used for testing, machine learning methods are known to fail. Broadly referred to as *domain adaptation*, one can adopt a gamut of techniques ranging from including a limited number real-world examples in the training set, to adjusting the pre-trained RKHS coefficients for a new dataset and finally employing subspace alignment techniques to adjust models to deal with changes to the domain.

(ii) Incorporation of Prior Knowledge

Prior knowledge is important to the generalization of machine learning models. The prior knowledge is all the auxiliary information including domain knowledge, underlying physics, and many others that can be used to guide the learning process. As discussed in Yu et al. (2007), there are in general three categories of methods of incorporating prior knowledge to learning process: 1. to design training examples using prior knowledge; 2. to initiate the learning algorithm using prior knowledge; 3. to reformulate the objective function using prior knowledge. We are most interested in approach 2 and 3, where we can modify our objective function by either designing specific kernel function, or we can learn regularization from data and use it to further facilitate the process of learning.

(iii) Others

Other important tasks of our future work include the geologic feature characterization is the detection of the existence of geologic feature, which can be formulated as a canonical classification problem. Also, uncertainty analysis provides users the confidence of the resulting predictions, which can be important in different situations. We will also investigate this direction according to some existence literature (Cawley et al. 2006).

6 CONCLUSIONS

We developed a computationally efficient, data-driven approach for subsurface geological features characterization using seismic data. Instead of detecting geological features from the migrated image or inversion, our proposed techniques are capable of detecting the geological features of interest from pre-stack seismic data sets. Our data-driven characterization methods are based on kernel ridge regression, which can be computationally intensive in training. To overcome the issues of excessive memory and computational cost that arises with kernel machines for large-scale data, we incorporated a randomized matrix sketching technique. The randomization method can be viewed as a data-reduction technique, because it generates a surrogate system that has much lower degrees of freedom than the

original problem. We show through our computational cost analysis that the proposed geologic feature characterization method achieves a significant reduction in computational and memory costs. Furthermore, we conducted several sets of experiments of detecting geological fault zone to study the performance of our method. The empirical accuracy of our method is comparable to the standard kernel ridge regression, while our method is significantly more efficient. Our data-driven characterization method presents a big advantage for the characterization of subsurface geological features. The current purpose of our technique is to complement the process of human intervention, and alleviate the chance of errors made by subjective human factors. With the improvement of computation power and the accumulation of data, we envision the replacement of human intervention by autonomous machine-learning-based subsurface characterization methods.

7 ACKNOWLEDGMENTS

This work was co-funded by the Center for Space and Earth Science (CSES) at Los Alamos National Laboratory (LANL) and the U.S. DOE Office of Fossil Energy through its Carbon Storage Program. The computation was performed using super-computers of LANL's Institutional Computing Program. J. Thiagarajanunder was supported by the U.S. DOE under Contract DE-AC52-07NA27344 to Lawrence Livermore National Laboratory.

APPENDIX A: SHERMAN-MORRISON-WOODBURY MATRIX IDENTITY

Given a square invertible $n \times n$ matrix A, an $n \times k$ matrix U, and a $k \times n$ matrix V, let B be an $n \times n$ matrix such that $B = A + UV$. Then, assuming $\left(I_k + VA^{-1}U\right)$ is invertible, we have the Sherman-Morrison-Woodbury matrix identity defined as

$$B^{-1} = A^{-1} - A^{-1}U\left(I_k + VA^{-1}U\right)^{-1}VA^{-1}. \tag{A.1}$$

By letting $A = \lambda I_n$, $U = \Psi$, and $V = \Psi^T$ and employing the above formulation to Eq. (21), we will have

$$\begin{aligned}
\tilde{\alpha} &= \left(\Psi\Psi^T + \lambda I_n\right)^{-1}\mathbf{y}, &\text{(A.2)}\\
&= \left((\lambda I_n)^{-1} - (\lambda I_n)^{-1}\Psi(I_k + \Psi^T(\lambda I_n)^{-1}\Psi)\Psi^T(\lambda I_n)^{-1}\right)\mathbf{y},\\
&= \left((\lambda^{-1} I_n - \lambda^{-2}\Psi(I_k + \lambda^{-1}\Psi^T\Psi)^{-1}\Psi^T\right)\mathbf{y},\\
&= \left((\lambda^{-1} I_n - \lambda^{-2}\lambda\Psi(\lambda I_k + \Psi^T\Psi)^{-1}\Psi^T\right)\mathbf{y},\\
&= \lambda^{-1}y - \lambda^{-1}\Psi(\lambda I_s + \Psi^T\Psi)^{-1}\Psi^T\mathbf{y} \in \mathbb{R}^n. &\text{(A.3)}
\end{aligned}$$

REFERENCES

Alaoui, A. & Mahoney, M. W., 2015. Fast randomized kernel ridge regression with statistical guarantees, in *Advances in Neural Information Processing Systems (NIPS)*.

Araya-Polo, M., Dahlke1, T., Frogner, C., Zhang, C., Poggio, T., & Hohl, D., 2017. Automated fault detection without seismic processing, *The Leading Edge*, **36**, 208–214.

Bach, F., 2013. Sharp analysis of low-rank kernel matrix approximations, in *International Conference on Learning Theory (COLT)*.

Campbell, C., 2001. An introduction to kernel methods, *Studies in Fuzziness and Soft Computing*, **66**, 155–192.

Cawley, G., Talbot, N., & Chapelle, O., 2006. Estimating predictive variances with kernel ridge regression, *Machine Learning Challenges*, pp. 56–77.

Chen, J., Avron, H., & Sindhwani, V., 2016. Hierarchically compositional kernels for scalable nonparametric learning, *arXiv preprint arXiv:1608.00860*.

Drineas, P. & Mahoney, M. W., 2005. On the nyström method for approximating a gram matrix for improved kernel-based learning, *Journal of Machine Learning Research*, **6**, 2153–2175.

Drineas, P. & Mahoney, M. W., 2016. RandNLA: Randomized numerical linear algebra, *Communications of the ACM*, **6**(6), 80–90.

Drineas, P., W., M. M., S., M., & Sarlos, T., 2011. Faster least squares approximation, *Numerische Mathematik*, **117**, 219–249.

Encyclopaedia-Britannica, 2010. Normal fault, [Online].

Fowlkes, C., Belongie, S., Chung, F., & Malik, J., 2004. Spectral grouping using the nystrom method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(2), 214–225.

Friedman, J., Hastie, T., & Tibshirani, R., 2001. *The elements of statistical learning*, vol. 1, Springer series in statistics New York.

Gittens, A. & Mahoney, M. W., 2016. Revisiting the Nyström method for improved large-scale machine learning, *The Journal of Machine Learning Research*, **17**(1), 3977–4041.

Guillen, P., 2015. Supervised learning to detect salt body, in *SEG Technical Program Expanded Abstracts*, pp. 1826–1829.

Haakon, F., 2010. *Structural Geology*, Cambridge University Press.

Hale, D., 2013. Methods to compute fault images, extract fault surfaces, and estimate fault throws from 3d seismic images, *Geophysics*, **78**(2), O33–O43.

Hidalgo, B. & Goodman, M., 2013. Multivariate or multivariable regression?, *Am J Public Health*, **13**(1), 39–40.

Huang, L., Shin, J., Chen, T., Lin, Y., Gao, K., Intrator, M., & Hanson, K., 2016. Breast ultrasound tomography with two parallel transducer arrays, in *Proc. SPIE 9783, Medical Imaging 2016: Ultrasonic Imaging, Tomography, and Therapy*, pp. 97830C–97830C–12.

Krebs, J. R., Anderson, J. E., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A., , & Lacasse, M. D., 2009.

Fast full-wavefield seismic inversion using encoded sources, *Geophysics*, **74**, WCC177–WCC188.

Le, Q., Sarlós, T., & Smola, A., 2013. Fastfood-computing hilbert space expansions in loglinear time, in *Proceedings of the 30th International Conference on Machine Learning*, pp. 244–252.

Leeuwen, T., 2016. Large-scale inversion in exploration seismology, *SIAM News*, **49**(2).

Li, M., Lian, X., Kwok, J. T., & Lu, B., 2011. Time and space efficient spectral clustering via column sampling, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, Y. & Huang, L., 2015a. Acoustic- and elastic-waveform inversion using a modified total-variation regularization scheme, *Geophysical Journal International*, **200**, 489–502.

Lin, Y. & Huang, L., 2015b. Least-squares reverse-time migration with modified total-variation regularization, in *SEG Technical Program Expanded Abstracts*.

Lin, Y. & Huang, L., 2015c. Quantifying subsurface geophysical properties changes using double-difference seismic-waveform inversion with a modified total-variation regularization scheme, *Geophysical Journal International*, **203**, 2125–2149.

Lin, Y., Wohlberg, B., & Guo, H., 2010. UPRE method for total variation parameter selection, *Signal Processing*, **90**(8), 2546–2551.

Lin, Y., Syracuse, E. M., Maceira, M., Zhang, H., & Larmat, C., 2015. Double-difference traveltime tomography with edge-preserving regularization and a priori interfaces, *Geophysical Journal International*, **201**(2), 574–594.

Meng, X. Saunders, M. A. & Mahoney, M. W., 2014. LSRN: A parallel iterative solver for strongly over- or underdetermined systems, *SIAM J. Sci. Comput.*, **36**, 95–118.

Moghaddam, P. P., Keers, H., Herrmann, F. J., & Mulder, W. A., 2013. A new optimization approach for source-encoding full-waveform inversion, *Geophysics*, **78**(3), R125–R132.

Mohri, M., Rostamizadeh, A., & Talwalkar, A., 2012. *Foundations of machine learning*, MIT press.

Nyström, E. J., 1930. Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben, *Acta Mathematica*, **54**(1), 185–204.

Rahimi, A. & Recht, B., 2007. Random features for large-scale kernel machines, in *Advances in neural information processing systems (NIPS)*, pp. 1177–1184.

Ramirez, J. & Meyer, F. G., 2011. Machine learning for seismic signal processing: Phase classification on a manifold, in *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops*.

Rawlinson, N. & Sambridge, M., 2014. Seismic travel time tomography of the crust and lithosphere, *Advances in Geophysics*, **46**, 81–197.

Rokhlin, V. & Tygert, M., 2008. A fast randomized algorithm for overdetermined linear least-squares regression, *Proc. Natl. Acad. Sci. USA*, **105**(36), 13212–13217.

Schnetzler, E. T. & Alumbaugh, D. L., 2017. The use of predictive analytics for hydrocarbon exploration in the Denver-Julesburg basin, *The Leading Edge*, **36**, 227–233.

Schölkopf, B. & Smola, A. J., 2002. *Learning with Kernels: Support Vector Machines, Regularization, Opti-*

mization, and Beyond., MIT Press.

Si, S., Hsieh, C.-J., & Dhillon, I., 2014. Memory efficient kernel approximation, in *International Conference on Machine Learning (ICML)*, pp. 701–709.

Stergiopoulos, S., 2017. *Advanced Signal Processing Handbook - Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*, Taylor and Francis Group.

Talwalkar, A., Kumar, S., Mohri, M., & Rowley, H., 2013. Large-scale SVD and manifold learning, *Journal of Machine Learning Research*, **14**, 3129–3152.

Tan, S. & Huang, L., 2014. An efficient finite-difference method with high-order accuracy in both time and space domains for modelling scalar-wave propagation, *Geophysical Journal International*, **197**(2), 1250–1267.

Tu, S., Roelofs, R., Venkataraman, S., & Recht, B., 2016. Large scale kernel learning using block coordinate descent, *arXiv preprint arXiv:1602.05310*.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics*, **74**(6), WCC1–WCC26.

Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., & Zhou, W., 2014. *Chapter 6. An introduction to full waveform inversion*, Society of Exploration Geophysicists.

Wang, K., Matthews, T., Anis, F., Li, C., Duric, N., & Anastasio, M. A., 2015. Waveform inversion with source encoding for breast sound speed reconstruction in ultrasound computed tomography, *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, **62**(3), 475–493.

Wang, S., 2015. A practical guide to randomized matrix computations with MATLAB implementations, *arXiv preprint arXiv:1505.07570*.

Wang, S., Zhang, Z., & Zhang, T., 2016. Towards more efficient SPSD matrix approximation and CUR matrix decomposition, *Journal of Machine Learning Research*, **17**(210), 1–49.

Wang, S., Gittens, A., & Mahoney, M. W., 2017. Scalable kernel k-means clustering with nystrom approximation: Relative-error bounds, *arXiv preprint arXiv:1706.02803*.

Williams, C. & Seeger, M., 2001. Using the Nyström method to speed up kernel machines, in *Advances in Neural Information Processing Systems (NIPS)*.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., & Zhou, Z.-H., 2012. Nyström method vs random fourier features: A theoretical and empirical comparison, in *Advances in Neural Information Processing Systems (NIPS)*.

Yu, T., Jan, T., Simoff, S., & Debenham, J., 2007. *Incorporating Prior Domain Knowledge into Inductive Machine Learning*, Ph.D. thesis, University of Technology, Broadway, NSW 2007, Australia.

Zhang, C., Frogner, C., Araya-Polo, M., & Hohl, D., 2014. Machine-learning based automated fault detection in seismic traces, in *Proceedings of 76th European Association of Geoscientists and Engineers Conference & Exhibition (EAGE)*.

Zhang, W. & Shen, Y., 2010. Unsplit complex frequency shifted PML implementation using auxiliary differential equation for seismic wave modeling, *Geophysics*, **75**, T141–T154.

Zhang, Y., Oldenburg, C. M., Finsterle, S., Jordan, P., & Zhang, K., 2009. Probability estimation of co2 leakage

through faults at geologic carbon sequestration sites, *Energy Procedia*, **1**, 41–46.

Zhang, Y., Duan, L., & Xie, Y., 2015. A stable and practical implementation of least-squares reverse time migration, *Geophysics*, **80**(1), V23–V31.

Zhang, Z., Tsang, I., & Kwok, J., 2008. Improved nystrm low-rank approximation and error analysis, in *International Conference on Machine Learning (ICML)*.

Zhang, Z., Huang, L., & Lin, Y., 2012. Efficient implementation of ultrasound waveform tomography using source encoding, in *Proc. SPIE 8320, Medical Imaging 2012: Ultrasonic Imaging, Tomography, and Therapy*, pp. 832003–832003–10.