Graph Spectrum Based Seismic Survey Design

Oscar López[1], *Rajiv Kumar*[2], *Nick Moldoveanu*[2] *and Felix J. Herrmann*[3]
[1]*Optimization and Uncertainty Quantification, Sandia National Laboratories,*
Albuquerque, NM, 87123
[2]*Schlumberger WesternGeco*
[3]*School of Earth and Atmospheric Sciences, Georgia Institute of Technology*

ABSTRACT

Randomized sampling techniques have become increasingly useful in seismic data acquisition and processing, allowing practitioners to achieve dense wavefield reconstruction from a substantially reduced number of field samples. However, typical designs studied in the low-rank matrix recovery and compressive sensing literature are difficult to achieve by standard industry hardware. For practical purposes, a compromise between stochastic and realizable samples is needed. In this paper, we propose a deterministic and computationally cheap tool to alleviate randomized acquisition design, prior to survey deployment and large-scale optimization. We consider universal and deterministic matrix completion results in the context of seismology, where a bipartite graph representation of the source-receiver layout allows for the respective spectral gap to act as a quality metric for wavefield reconstruction. We provide realistic survey design scenarios to demonstrate the utility of the spectral gap for successful seismic data acquisition via low-rank and sparse signal recovery.

INTRODUCTION

Low-rank matrix recovery (LRMR) and compressive sensing (CS) based seismic data acquisition and processing has successfully materialized from an academic novelty to an indispensable industrial tool (Herrmann and Hennenfent, 2008; Ma, 2008; Herrmann, 2009; Hennenfent et al., 2010; Oropeza and Sacchi, 2011; Herrmann et al., 2012; Ma, 2013; Yang et al., 2013; Kumar et al., 2015; Mosher et al., 2017). As a popular example, seismic data reconstruction from limited measurements provides an instance of the matrix completion problem (Candès and Recht, 2009; Candes and Plan, 2010; Candès and Tao, 2010; Recht, 2011). These approaches allow for simultaneous acquisition and compression of seismic data, where a substantially reduced number of field samples allow dense wavefield reconstruction via large-scale optimization. As a consequence, practitioners may achieve high-fidelity imaging and post-processing with economical surveys. In this paper, we focus on the matrix completion approach but eventually extend our scope to include sparsity based techniques.

In general, successful matrix completion based seismic data acquisition and reconstruction hinge upon the following three requisites:

1. Densely sampled seismic data (in some transform domain) reshaped as a matrix $\mathbf{X} \in$

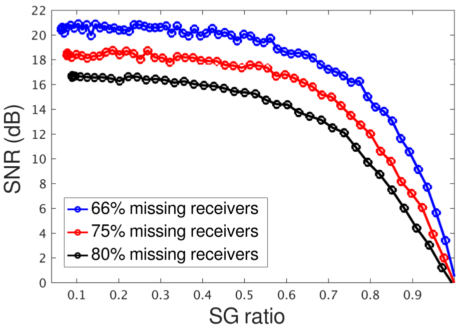
$\mathbb{C}^{n\times m}$ should exhibit *low-rank structure*, i.e., $\mathbf{X}$ is well approximated (according to some tolerance) by a rank-r matrix where $r \ll \min(n, m)$.

2. A computationally efficient program for large-scale LRMR, e.g., nuclear norm minimization (Recht et al., 2010).

3. The set of observed matrix entries $\Omega \subset \{1, 2, \cdots, n\} \times \{1, 2, \cdots, m\}$ should be suitable, e.g., uniform random sampling according to n, m and r (Candès and Recht, 2009; Candès and Tao, 2010).

To address 1), several reorganizations and transformations of seismic data tensors into matrices have been shown to expunge low-rank structure (Da Silva and Herrmann, 2013; Kumar et al., 2015; Ma, 2013). Furthermore, for 2) computationally efficient methodologies have been proposed to solve the matrix completion problem on large matrices (Aravkin et al., 2014; Kumar et al., 2017; Jain et al., 2013), including on parallel architectures (Recht and Ré, 2013; Yun et al., 2013). However, it is not clear how to fulfill 3) in a practical manner. Note that, the randomized sampling is highely desired for an optimal seismic data reconstruction since it turns aliasing into an incoherent noise, thus making interpolation a denoising problem. In the theory of matrix completion, generating appropriate sampling schemes is typically guaranteed with high probability by choosing samples uniformly at random from a dense grid (Candès and Recht, 2009; Candes and Plan, 2010; Candès and Tao, 2010; Recht, 2011). This is difficult to achieve in seismology since it corresponds to sources and receivers removed uniformly at random from equipment arrays independently for each survey. As a consequence, LRMR based acquisition design is a vague procedure that requires computational exploration of numerous criteria to decide upon an appropriate source-receiver layout. Such processes involve simulating a 5D data tensor (corresponding to a 3D seismic survey) and performing large-scale reconstruction for all possible options within physical constraints. Mosher et al. (2014) refers to this process as simulation based acquisition design where the author solves an optimization problem to determine the locations of sources and receivers for a nonuniform design, rather than relying solely on decimation, jittering, or randomization.

To mitigate this tedious process, we propose a computationally cheap metric to evaluate a given acquisition design. The novel idea comes from the matrix completion literature (Bhojanapalli and Jain, 2014; Heiman et al., 2014; Burnwal and Vidyasagar, 2020), where the authors show that matrix completion is successful if the spectral gap (SG) of the associated set of observed entries is large. To elaborate, the results consider the binary matrix $\mathbf{M} \in \{0, 1\}^{n\times m}$ whose non-zero entries specify the sampled data matrix entries and compute the gap between $\mathbf{M}$'s first and second singular values $\sigma_1(\mathbf{M}), \sigma_2(\mathbf{M})$. The main results in Bhojanapalli and Jain (2014); Burnwal and Vidyasagar (2020) state that one can recover an incoherent (i.e., “not spiky”) low-rank matrix via nuclear norm minimization if the number of samples and the SG are large enough. Intuitively, the SG can be seen as a qualitative measure informing us how well we can expect a given sampling scheme to perform matrix completion. We propose this deterministic measure to differentiate appropriate surveys at an initial design stage prior to in-field acquisition and large-scale optimization, while avoiding numerically expensive simulation based approaches. Moreover, the proposed approach does not require solving an expensive optimization problem to choose the best design criteria among the available possible candidate solutions.



a

Figure 1: Illustration of the relationship between the quality of reconstruction and the SG of the sampled matrix entries. The figure plots the average reconstruction SNR (defined in (3)) vs average SG ratio $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$ of the corresponding binary sampling matrix $\mathbf{M}$, for 3 percentages of missing receivers. The results are taken from the average of 100 independent experiments.

The relationship between the SG and quality of reconstruction is demonstrated in Figure 1. Given $\mathbf{M} \in \{0,1\}^{n\times m}$ specifying the sampled matrix entries, we constitute its SG by the ratio $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})} \in [0,1]$. Henceforth, we will refer to the term $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$ as the SG ratio and note that a large SG corresponds to a small SG ratio ($\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})} \ll 1$) and vice versa. Figure 1 illustrates the correlation between $\mathbf{M}$'s SG and the signal-to-noise ratio (SNR, defined in (3)) of reconstruction via LRMR with observed matrix entries according to $\mathbf{M}$. The plot showcases better quality of reconstruction from samples with larger SG (i.e., smaller SG ratios $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$). In each plot with a fixed percentage of missing receivers, the number of observed entries does not change and the SG is therefore mainly concerned with the distribution of these samples. The procedure to generate acquisition designs with varying spectral gap demonstarted in Figure 1 are postponed until the experimental section.

Design of optimal acquisition geometries is an ongoing area of research with several solutions in the literature relevant to our work. In Mosher et al. (2017), the authors propose a reconstruction based design procedure where a greedy method first output several optimal sampling patterns. Then the full 3D acoustic seismic modelling and imaging is performed. Based upon the quality of the imaging, the optimal design is selected from the trail optimal sampling patterns. Note that, performing full scale 3D simulation and imaging is a computationally expensive process. The work in Allouche et al. (2020) uses prior information from the low-frequency spectrum of seismic data to construct a signal and coherent noise model, followed by the evaluation of a quality metric for acquisition layout as a function of frequency. The optimal acquisition layout corresponds to the worst case value of the metric over the band. The proposed method does not require any synthetic or real data, but the metric involves computing several condition numbers (i.e, ratio of the smallest and largest singular values of the sensing matrix). This ratio is numerically expensive for sensing matrices related to large-scale 5D seismic data. In contrast, our proposed method requires only computing the two largest singular values of a single binary matrix representing source-receiver layout. This approach is numerically efficient and does not require underlying data nor a prior model.

We validate our approach with several numerical experiments that consider distinct aspects of the seismic data acquisition and reconstruction process. We begin with some background from the literature of matrix completion that formally introduces the SG. Using carefully selected numerical experiments, we first empirically demonstrate the relationship between the SG and quality of reconstruction via LRMR. We then demonstrate how the SG is sensitive to large gaps of missing data and utilize the SG to determine an appropriate density of the output interpolation grid. We also illustrate that the SG is informative in sparsity based recovery and infact not just related to the matrix completion literature. We end the paper with concluding remarks and future work.

UNIVERSAL AND DETERMINISTIC MATRIX COMPLETION

In contrast to most results in the matrix completion literature, the work in Bhojanapalli and Jain (2014); Heiman et al. (2014); Burnwal and Vidyasagar (2020) aims for *universal* and deterministic recovery guarantees. The term universal refers to the set of observed matrix entries, $\Omega \subset \{1, 2, \cdots, n\} \times \{1, 2, \cdots, m\}$, where this subset is expected to provide an accurate output via nuclear norm minimization for an entire class of signals uniformly. In mathematics, nuclear norm is defined as the sum of the singular values of teh undelrying matrix. Such results are of great value to practitioners since Ω will function for all matrices of interest, removing the need to design new sampling schemes for every independent seismic survey. Furthermore, the attached conditions that ensure the suitability of Ω inherently provide a means to quantify the source-receiver layout. This last observation is the focus of this article, where the spectral gap serves as a computationally cheap tool to decide if a provided acquisition design is suitable for reconstruction via matrix completion prior to data acquisition and large-scale optimization.

To elaborate, we begin with some matrix completion background and then discuss the work in Bhojanapalli and Jain (2014); Heiman et al. (2014); Burnwal and Vidyasagar (2020). In what follows, let $n \geq m$ without loss of generality. Given a matrix of interest $\mathbf{X} \in \mathbb{C}^{n \times m}$ and a subset of observed entries $\Omega \subset \{1, 2, \cdots, n\} \times \{1, 2, \cdots, m\}$, our collected data is given as $\mathbf{B} = P_\Omega(\mathbf{X} + \mathbf{E}) \in \mathbb{C}^{n \times m}$, where for $(k, \ell) \in \{1, 2, \cdots, n\} \times \{1, 2, \cdots, m\}$

$$P_\Omega(\mathbf{X} + \mathbf{E})_{k\ell} = \begin{cases} \mathbf{X}_{k\ell} + \mathbf{E}_{k\ell} & \text{if } (k, \ell) \in \Omega \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and $\mathbf{E}$ encompasses the noise in our observations with bounded Frobenius norm

$$\|P_\Omega(\mathbf{E})\|_F := \left(\sum_{(k,\ell)\in\Omega} |\mathbf{E}_{k\ell}|^2 \right)^{1/2} \leq \eta.$$

Matrix completion via nuclear norm minimization aims to approximate

$$\mathbf{X} \approx \mathbf{X}^\sharp := \underset{\mathbf{Z}\in\mathbb{C}^{n\times m}}{\operatorname{argmin}} \quad \|\mathbf{Z}\|_* \quad \text{s.t.} \quad \|P_\Omega(\mathbf{Z}) - \mathbf{B}\|_F \leq \eta, \tag{2}$$

where

$$\|\mathbf{Z}\|_* = \sum_{k=1}^{m} \sigma_k(\mathbf{Z})$$

is the *nuclear norm* and $\sigma_k(\mathbf{Z})$ denotes the k-th largest singular value of $\mathbf{Z}$. This method has been extensively studied Recht (2011); Candès and Recht (2009); Candes and Plan (2010); Candès and Tao (2010). In the case $\eta = 0$, standard results in the literature show that $\mathbf{X}$ can be exactly recovered via (2) with high probability under conditions of the general form:

- $\text{rank}(\mathbf{X}) = r \leq \min\{n, m\}$.
- $\mathbf{X}$ is incoherent, e.g., satisfies the strong incoherence property (see Recht (2011); Candès and Recht (2009); Candes and Plan (2010); Candès and Tao (2010); Bhojanapalli and Jain (2014)). Intuitively, this assumption ensures that the energy (i.e., Frobenius norm) of the data matrix is not concentrated on a small subset of entries.
- Ω is generated uniformly at random with $|\Omega| \sim rn\log^{\alpha}(n)$, for some $\alpha \geq 1$ and $|\Omega|$ indicating the number of observed entries. Furthermore, each matrix to be recovered requires an independent Ω to be generated.

The first condition is crucial since program (2) with $|\Omega| < nm$ is an underdetermined problem and the low-rank constraint aims to compensate for the limited number of observations. This validates the use of the nuclear norm as the objective function, in order to output $\mathbf{X}^\sharp$ that can be written as a sum of few rank 1 matrices (i.e., low-rank structure). The second assumption is a necessary condition for sampling schemes that are oblivious of the matrix structure, but can be avoided if one samples according to prior knowledge of the singular vectors (Chen et al., 2015).

However, the third condition is restrictive and does not hold in many applications. For example, in seismic acquisition the allowed structure of Ω is limited by measurement hardware and physical constraints involved in data collection. Achieving uniform random sampling independently for each survey would require a significant amount of novel sensing equipment. Instead, matrix completion results that consider achievable sampling schemes would be most informative and applicable.

To this end, we now consider results involving universal and deterministic matrix completion Bhojanapalli and Jain (2014); Heiman et al. (2014); Burnwal and Vidyasagar (2020). Let $\mathbf{1}$ be the all-ones $n \times m$ matrix and $\mathbf{M} := P_\Omega(\mathbf{1}) \in \{0,1\}^{n\times m}$. We will refer to $\mathbf{M}$ as the *sampling mask* and note that it is a binary matrix, which specifies the locations of sampled matrix entries. The main result in Bhojanapalli and Jain (2014) shows that if the spectral gap (SG) of $\mathbf{M}$ is large enough (i.e., $\sigma_2(\mathbf{M}) \ll \sigma_1(\mathbf{M})$) then one can recover any low-rank incoherent matrix via (2). Specifically, the result reads as follows:

Theorem 0.1 (Theorem 4.2 in Bhojanapalli and Jain (2014)) *Let $\boldsymbol{X} \in \mathbb{C}^{n\times m}$ be a rank-r matrix satisfying the strong incoherence property with parameter μ (see Candès and Tao (2010) and Claim 5.1 in Bhojanapalli and Jain (2014)). Let Ω be generated such that the top singular vectors of $\boldsymbol{M}$ are the all 1's vectors and*

$$|\Omega| \geq 36\frac{\sigma_2(\boldsymbol{M})^2}{\sigma_1(\boldsymbol{M})}\mu^2 r^2 n.$$

Then $\boldsymbol{X}$ is the unique optimum of (2) with $\eta = 0$.

To reiterate, the incoherence parameter μ quantifies how evenly spread the information is throughout the data matrix and ensures that we sample accordingly. The condition on the singular vectors of $\mathbf{M}$ requires Ω to contain the same number of samples in each row and column. This constraint is quite restrictive, but also required for the main results in Heiman et al. (2014); Burnwal and Vidyasagar (2020). Generating sampling schemes with this structure is not straightforward Lazebnik and Ustimenko (1995); Lazebnik and Woldar (2001) and complex to translate to field samples. However, in the experimental section we allow ourselves the liberty to violate this assumption. Our numerical experiments will demonstrate that the relationship between the SG and the success of matrix completion remains useful even when this condition does not hold.

In contrast to standard uniform random sampling requirements, $|\Omega| \sim \mathcal{O}(rn\log^{\alpha}(n))$, Theorem 0.1 requires $|\Omega| \sim \mathcal{O}(r^2 n)$. While this sample complexity is pessimistic when $r > \log^{\alpha}(n)$, we stress that the result holds deterministically and universally for all incoherent matrices. In the best case scenario we can expect $\frac{\sigma_2(\mathbf{M})^2}{\sigma_1(\mathbf{M})} \sim \mathcal{O}(1)$ (e.g., Ramanujan graph Hoory et al. (2006); Bhojanapalli and Jain (2014)). More importantly, Theorem 0.1 provides a useful relationship between the success of matrix completion and the ratio $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})} \in [0, 1]$ directly related to the SG. This ratio appears with similar implications in the related work Heiman et al. (2014); Burnwal and Vidyasagar (2020). Intuitively, a practitioner choosing amongst several sampling designs can compute the respective SG's and make an educated choice of which design will output the best quality of matrix reconstruction.

Connections with Graph Theory

Interestingly, the SG is a common tool for analyzing the connectivity of communication networks Hoory et al. (2006); Watanabe and Masuda (2010). In this context, a larger SG provides a more robust network. The foundation comes from graph theory and the study of expander graphs, where a communication network can be seen as a group of vertices (network nodes) with corresponding edges (signifying communication between nodes). Considering the respective adjacency matrix (analogous to $\mathbf{M}$ in our context), this matrix's SG gives a direct measurement of the network's efficiency in the flow of information.

A similar interpretation can be presented in our seismology scenario. The array of sources and receivers can be seen as the vertices with edges that represent the recording and transmission of pressure waves. In other words, using the matricization in the numerical experiments section (see also Kumar et al. (2015)), a source-x, receiver-x pair (vertex) will have an edge with a source-y, receiver-y pair if the receiver-x, receiver-y pair recorded the pressure wave due to the corresponding source-x, source-y pair (and vice versa with respect to the reciprocal role of transmission). Thus, in our case $\mathbf{M}$ can be seen as the corresponding bi-adjacency matrix of the bipartite graph that captures a sampling scenario with such vertices and edges. In geophysical terms, the bipartite graph represents the ray path connections between the source and receivers. Since the interpolation problem implicitly depends on well distributed samples, we can intuitively see that subsampling pattern with the well connected graph (i.e., large SG) will be more effective at infilling data between source-receiver pairs. This interpretation was first elaborated in Candès and Recht (2009), where the authors show that a fully connected graph is necessary to recover a matrix of any rank.

NUMERICAL EXPERIMENTS

To demonstrate the utility of the SG for survey design, we numerically address four specific questions related to seismic data acquisition and reconstruction organized in the following subsections as outlined below:

- What is the relationship between the SG and the quality of LRMR based reconstruction? In this first section, we numerically demonstrate that there is a clear correlation between the SG of a given sampling mask and the quality of reconstruction obtained from the set of sampled entries. Specifically we elaborate on the experiments that produced Figure 1.
- Is the SG sensitive to large gaps of missing data? We design experiments that test if the SG can detect acquisition designs with large gaps of missing samples. We argue why this property is important for seismic data reconstruction.
- Can the SG identify appropriate output interpolation grid density? In this section we consider a coil sampling acquisition design deployed by Schlumberger, with the goal of identifying an output interpolation grid onto which dense data can be reconstructed. The SG aids in detecting an output grid that is as dense as possible, while appropriate for LRMR.
- Is the SG informative for sparsity based reconstruction? We argue that SG is also useful for compressive sensing even though it is a tool derived from the matrix completion literature

We design realistic sampling scenarios and validation on geologically diverse synthetic datasets to provide strong empirical evidence that the SG can be used in early acquisition design to rule out inappropriate sampling schemes. Some of our experiments are abstract in the sense that we do not consider an explicit dataset and instead focus on the source-receiver layout. This stresses that the SG can be applied prior to data collection and without simulated data.

In what follows, we use the signal-to-noise ratio (SNR) as a metric of matrix reconstruction quality defined as

$$-20\log_{10}\left(\frac{\|\mathbf{X}^{\sharp}-\mathbf{X}\|_F}{\|\mathbf{X}\|_F}\right) \tag{3}$$

where $\mathbf{X} \in \mathbb{C}^{n\times m}$ is the desired data matrix, $\mathbf{X}^{\sharp} \in \mathbb{C}^{n\times m}$ is the recovered approximation.

Relationship between the SG and quality of reconstruction

We begin with examples that numerically demonstrate the relationship between the SG and the quality of LRMR output. We perform experiments on 5D data tensors (corresponding to 3D seismic surveys respectively). We consider the BG 3D model with 68 sources and 401 receivers along each axis to generate 5D data tensor with dimensions: time, source-x, source-y, receiver-x, and receiver-y. We work in the Fourier domain by applying the fast Fourier transform (FFT) along the time axis and reconstruct in a frequency by frequency basis (generating a sequence of 4D tensors). We use the matricization from Da Silva and

Herrmann (2013); Kumar et al. (2015) to reorganize each 4D tensor into a matrix, where the source-x, receiver-x pairs are coupled along the rows and the source-y, receiver-y pairs are coupled along the columns (with the sources serving as the outer dimensions). Note that, due to the source-receiver reciprocity, the experimental results are equally valid for the scenario where receivers are used as an outer dimension to matrizice the 4-dmension tensor. Henceforth, we will refer to this matricization as src-rec form. We restrict ourselves to a single 7.34 Hz frequency slice and the first 5 sources along each axis, generating a 2005×2005 matrix (see Figure 2a).

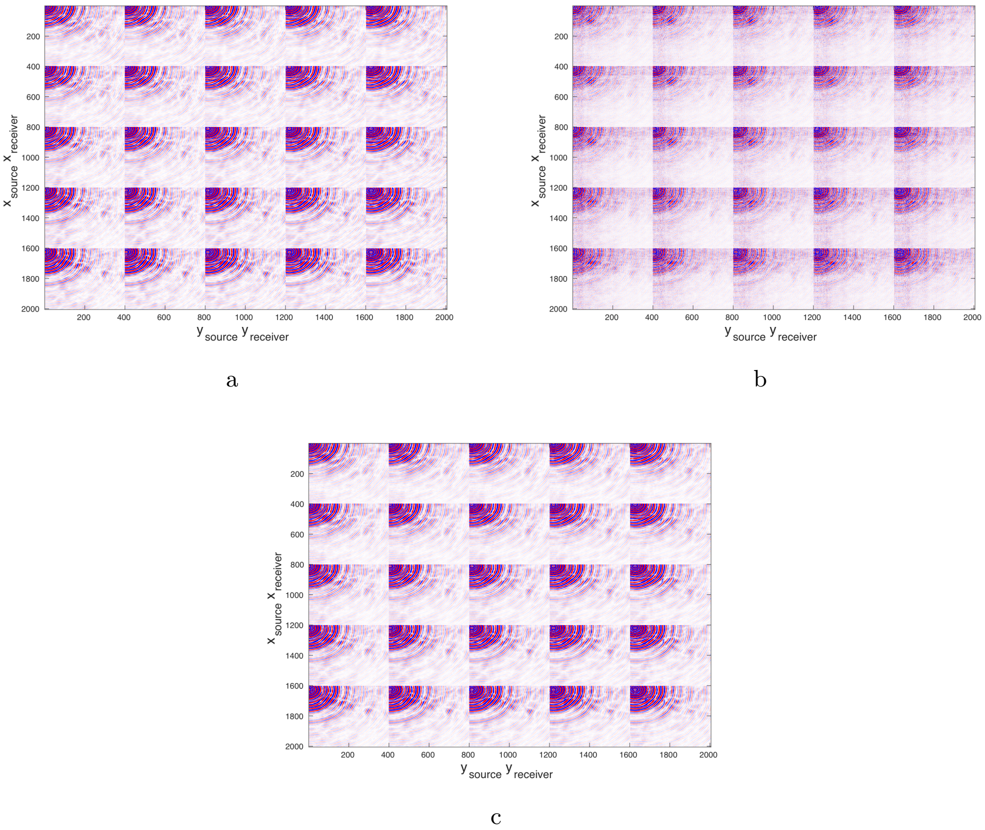
We now develop an acquisition scheme to vary the spectral gap. We begin by removing a specified number of receivers (according to subsampling percentage) in a periodic manner (e.g., keep only first receiver of every three for approximately 66% missing receivers). We then solve the nuclear norm minimization problem (2) with $\eta = 0$ applying this sampling mask and compute the respective SG ratio $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$. Subsequently, we choose a specified percentage $p \in [0, 1]$ of the observed receivers and relocate them among the unobserved receiver locations in a uniform random manner. Hence, for a fixed subsampling percentage, this generates sampling masks with varying fraction of randomly placed receivers. We now solve the nuclear norm minimization problem with the modified sampling mask and record the respective SG ratio. We repeat this procedure for several percentage p values in an increasing manner (specifically, $p \in \{0, .02, .04., \cdots, 1\}$). Note that this procedure preserve the desired subsampling percentage, it only modifies the locations of the sources and/or receivers in a random fashion.

In general, we observe a correlation between the number of randomly placed receivers and the SG of the resulting sampling mask, where masks with larger fraction of randomly placed receivers provide larger SG values (i.e., smaller SG ratios $\frac{\sigma_2}{\sigma_1}$). This whole procedure is repeated 100 times independently and we average the SG ratios and SNR values for each $p \in \{0, .02, .04., \cdots, 1\}$. To obtain Figure 1, we sort the resulting SG ratios in increasing order and plot them against the corresponding SNR of reconstruction. Furthermore, Figures 2b and 2c show two recovered data matrices from sampling masks with small and large SG, respectively. These figures imply a clear correlation between the SG and the SNR of reconstruction, i.e., a smaller spectral gap ratio results in larger signal-to-noise ratio on average.

Notice that the SNR values in Figures 1 taper off rather quickly towards the acceptable SNR value in each case. This observation implies that a practitioner need not place all receivers randomly and might achieve a similar quality of reconstruction with a hybrid layout (e.g., half periodic and half random receiver placement). Such acquisition designs provide achievable surveys with near-optimal SG ratios (i.e., close to those of a Ramanujan graph (Hoory et al., 2006)), where the resulting survey minimizes the number of receivers not placed in a convenient and patterned fashion.

Sensitivity of the SG to large gaps of missing data

The previous section implicitly demonstrated the utility of randomness for low-rank matrix recovery appropriate sampling via the SG. However, such random samples may exhibit large gaps of missing data that incur degrading artifacts on many interpolation techniques (Trad et al., 2005; Hennenfent and Herrmann, 2008). In this section, we explore the sensitivity



a

b

c

Figure 2: Examples of recovered data matrices via sampling masks with small and large SG. (a) True data (b) reconstructed data with sampling mask exhibiting SG ratio = .9828 and recovery SNR = 3.5 dB (75% missing receivers), and (c) reconstructed data with sampling mask exhibiting SG ratio = .1796 with SNR = 20.7 dB (75% missing receivers).

of the SG to such gaps. For this purpose we use jittered undersampling (Hennenfent and Herrmann, 2008; Shahidi et al., 2013), i.e., we partition the desired dense receiver grid into intervals with equal number of receivers and randomly choose a single receiver location to sample from each interval (where the number of receivers per interval depends on the subsampling ratio). These sampling techniques are also common in computer graphics where they are used to attenuate aliasing artifacts while controlling the gap size of unrendered data (Dippé and Wold, 1985; Ignjatovic and Bocko, 2005; Christensen et al., 2018).

We design an abstract 3D survey (corresponding to a 5D data tensor) with 32 sources and receivers in each axis and restrict ourselves to a single frequency slice, working in the Fourier domain and in src-rec form as in the previous section (generating a 1024 × 1024 matrix). To emphasize that our methodology does not require numerically expensive data simulation, we do not use an explicit dataset here and instead provide our argument via binary masks only.

We remove 80% of the receivers via several jitter undersampling methods that control the average gap size of missing receiver data. We do so by varying the probability distribution by which we choose the receiver location for every other interval, while the remaining intervals retain the uniform distribution (i.e., each receiver location has equal probability of being chosen). The selection distribution is modified by selecting a jitter parameter $:= \rho \in [0,1]$ and only choosing a sample among the first $\lceil \rho \cdot 5 \rceil$ receiver locations in the interval, where $\lceil z \rceil$ rounds to the nearest integer greater than or equal to z. This provides receiver locations with increasing average gap sizes (as ρ decreases). Specifically we consider $\rho \in \{.2, .4, .6, .8, 1\}$. By only modifying every other interval's selection distribution we still assure a rich sampling scheme, i.e., the uniform distribution of the unmodified intervals allows each row and column to be sampled with high probability (a crucial property for rank penalization techniques (Hennenfent and Herrmann, 2008)). Note that, the average sampling rate is fixed and only the gap size is controlled through jitter sampling.

We compute the SG ratio of each mask generated for each ρ value specified above. We repeat the experiments 200 times and plot the average SG ratio against the corresponding jitter parameter ρ in Figure 3. The experiments are repeated for 90% and 95% missing receivers. Figure 3 exhibits a clear trend between the jitter parameter ρ and the SG, where $\rho \approx 1$ provides the largest SG (smallest SG ratios $\frac{\sigma_2}{\sigma_1}$). The case $\rho = 1$ corresponds to optimally-jittered undersampling (Hennenfent and Herrmann, 2008), with smallest average gap size between receiver locations. Arguably, this type of jitter is best for general sampling and the SG can successfully make this distinction. Notice that the three plots with distinct sampling percentages deviate in SNR values for such larger values of ρ, this emphasizes that the SG is mainly concerned with the distribution of the samples rather than the number of samples. In other words, for a more complete evaluation of a given sampling scheme, the practitioner should also ensure that the percentage of samples received is sufficient (which can be achieved by consulting the theory of matrix completion Recht (2011); Candès and Recht (2009); Candes and Plan (2010); Candès and Tao (2010); Bhojanapalli and Jain (2014)). The main conclusion of this section is that the SG can be used to detect a well distributed set of samples while remaining aware of large gaps of missing data. These properties are crucial for sampling sets that achieve accurate interpolation results since smaller gaps suppress aliasing and/or blue noise in the frequency-wavenumber spectrum (Trad et al., 2005; Shahidi et al., 2013). As illustrated by the experiments, the SG is an efficient tool to quantify these properties in a given set of observed matrix entries.

Coil Sampling Experiments

To demonstrate the benefits of spectral gap to estimate the optimal interpolation grid, we consider a 3D marine coil sampled survey, which employs a type of circular shooting that introduces randomness into the acquisition geometry Moldoveanu (2010). For processing steps like interpolation, SRME and imaging, such data sets must be accurately interpolated onto a dense regular grid. We focus on this reconstruction step for a dual-coil streamer data set, deployed on a survey area of 10 × 10 km (illustrated via src-rec form in Figure 4). We consider the source-receiver layout indicated by the streamer data, with the goal of choosing an output interpolation grid that is as dense as possible while appropriate for LRMR. The experiments of this section do not utilize the actual pressure wave recordings of the survey (this data remains the property of Schlumberger). Instead, we verify the quality of reconstruction by simulating a 10 Hz frequency slice using the SEG/EAGE Overthrust model that matches the 10 × 10 km area from which the coil sampling geometry was extracted.

We adopt the same work flow from the previous sections, reconstructing sequentially on

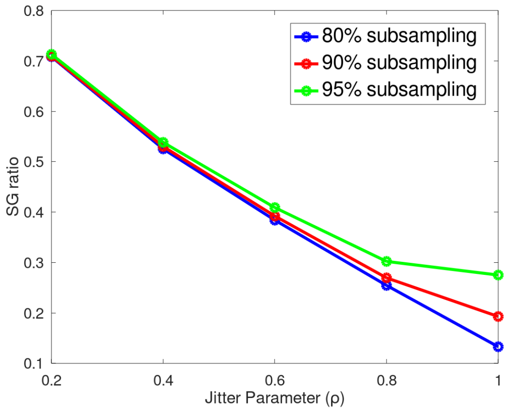


Figure 3: Plot of average SG ratio of sampling mask vs jitter sampling parameter ρ, for various subsampling percentages. The results are taken from the average of 200 independent experiments. Notice that as the average jitter sampling gap increases ($\rho \to 0$) we obtain less favorable SG values (and diminished quality of reconstruction as expected). The SG is therefore sensitive to large gaps of missing data and we expect the highest quality of reconstruction via optimally-jittered undersampling ($\rho = 1$).

4D fixed frequency tensors matricized via src-rec form. Given a desired spacing of sources and receivers, we create a corresponding sampling mask $\mathbf{M} \in \{0,1\}^{n\times m}$, with ones in the entries where streamer data was acquired and zeros assigned to all other entries. Note that, in the field acquisition, sources and receivers are placed off-the grid (i.e., not necessarily in our desired output grid). Therefore, to construct the sampling mask we *bin* the data onto the desired reconstruction grid, i.e., we assign each off-the-grid sample to its nearest grid point. We adopt this binning procedure for the acquisition design only, and highly recommend that reconstruction be achieved using LRMR methodologies that incorporate the off-the-grid samples for improved quality of reconstruction (e.g., López et al. (2016)). While our current SG methodology does not consider such off-the-grid layouts, we nonetheless found the SG tool applied in this binned scenario very useful since it still faithfully informed upon the quality of reconstruction via off-the-grid methodologies. We reiterate that we cannot share such findings here since this data remains the property of Schlumberger.

We examined many output grids by varying the source-receiver sampling between 50-200 m and showcase three examples in Table 1 along with the respective SG ratios and SNR of reconstruction (computed for the synthetic 10 Hz Overthrust frequency slice). The SG successfully distinguishes which layout will produce the highest quality output in terms of the SNR. Notice that a denser grid is not always optimal since the example in Table 1 with densest 50 m source-receiver spacing exhibits the least desirable SG ratio (and reconstruction SNR). Intuitively, this 50 m grid is relatively too dense for the given coil samples and therefore increases the subsampling percentage with large gaps of missing data that are conveniently detected via the SG.

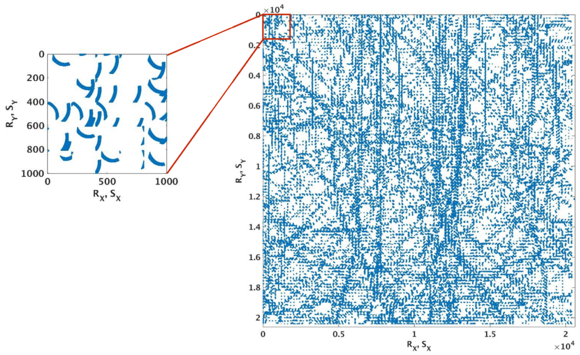


Figure 4: Coil-based marine seismic survey deployed by Schlumberger, illustrated in src-rec form. The sampling mask corresponds to single-coil survey, extracted from a dual-coil survey.

Receiver spacing (m)	Source spacing (m)	SG ratio ($\frac{\sigma_2}{\sigma_1}$)	SNR (dB)
100	200	0.77	10
50	100	0.4	16
50	50	0.88	8

Table 1: SG ratio and SNR as a function of underlying grid spacing for marine seismic data, where the sampling mask in src-rec form is extracted from a dual-coil survey deployed by Schlumberger. The SNR corresponds to LRMR experiments using a simulated 10 Hz frequency slice via the SEG/EAGE Overthrust model that matches the survey area. As we can see, lower spectral gap is tied to the best SNR value for the acquisition scenario where recieiver sampling is 50m and the source sampling is 100m.

SG for sparsity based reconstruction

Although our focus thus far has been on the matrix completion approach, graph spectrum based techniques are flexible and apply to other signal processing and imaging frameworks (including compressive sensing Lu et al. (2018)). To validate the SG on sparsity based seismic data reconstruction, we designed a 3D survey layout for an ocean-bottom node (OBN) acquisition and perform reconstruction using matching pursuit Fourier interpolation Schonewille et al. (2013). We use synthetic data simulated on the geologically complex SEAM model. To simplify the concept, we simulate a common receiver gather where the receiver is placed at the centre of the 2D surface over which we placed the sources sampled at 10m along the x- and y-directions, respectively. To simulate a realistic subsampling scenario, we sample sources at 50 and 100 m across inline and crossline directions, respectively. This subsampling will cause aliasing to appear from 15 and 7 Hz onwards across the frequency band in the inline and crossline directions. We only perform reconstruction for a common-receiver gather. We interpolate a periodic and a jittered random acquisition design Hennenfent and Herrmann (2008), comparing the resulting SG and SNR of reconstruction.

The periodic and jittered sampling masks have respective SG ratios $\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$ of 1 and 0.35, see Figure 5 for an illustration with fixed receivers and time axis. Note that the periodic and jittered sampling masks are same across the time-axis, i.e., the data is only subsampled in the space direction. For periodic sampling, we subsample data at every 5^{th} and 10^{th}

grid point along inline and crossline directions resulting in a subsampling ratio of 98%. Figure 6a shows the cross-line section whereas Figures 6b and 6c show the subsampled data using periodic and jittered sampling respectively. Figures 7a and 7b show the results for periodic and jittered subsampling with an SNR values of 20 and 24 dB, respectively. To stabilize the reconstruction, priors were introduced in the data reconstruction framework. The amplitude spectrum derived from the alias-free frequency band of small spatio-temporal windows is incorporated as weights to distinguish between the aliasing artifacts and the true events at the higher frequencies Schonewille et al. (2009); Özbek et al. (2010). Although the priors significantly improve the results for the periodic sampled data, we still loose coherent energy of complex seismic events. In comparison, the reconstruction via jittered sampling with priors is able to preserve the continuity of reflection events and complex diffraction patterns. Even though priors can stabilize the interpolation results as supported by the residuals in Figures 7c and 7d, the spectral gap at first place can illustrate the sampling scenario producing the optimal interpolation results with or without usgae of priors.

DISCUSSION AND FUTURE WORK

Though our work provides a means to alleviate survey design, the current spectral gap methodology has several limitations as a tool in seismology. These restrictions in turn galvanize directions for future work, which we identify and discuss in this section.

The main results in Bhojanapalli and Jain (2014); Heiman et al. (2014); Burnwal and Vidyasagar (2020) require that Ω be generated from a regular graph (this is the condition on the singular vectors of $\mathbf{M}$ in Theorem 0.1), which may be too restrictive. While our experiments have illustrated that the spectral gap is informative even when this condition may not hold, a general result that relaxes this requisite would yield a more flexible and informative tool for cost-effective and low-environmental impact acquisition design.

Throughout, we implicitly assume that the samples are on-the grid since our sampling masks are generated by choosing sources and receivers from a dense equispaced grid. However, in practice this is rarely the case since samples often deviate from the desired grid points and in general lie on a continuous domain (as in the dual-coil survey). Therefore, to fully consider coil sampling and other randomized acquisition designs, an analogous tool that quantifies these off-the-grid source-receiver layouts is needed. Such a generalized spectral gap would build upon the intuition of our current work but could, for example, include modeling of ocean tides to generate and quantify a realistic non-equispaced sensor layout.

Finally, this work has shown that the spectral gap can help choose among a finite set of acquisition designs. Although this is a valuable tool, practitioners must exhaustively consider many abstract sampling schemes to be compared via the spectral gap. An improved tool could utilize the concepts of this work to choose among an infinite set of layouts. This could be done, for example, by setting up a tractable optimization program that finds the sparsest mask with largest spectral gap among all masks that satisfy constraints specified by a desired sampling geometry.

CONCLUSIONS

This paper provides a computationally efficient tool that extends the utility of compressive sensing and low-rank matrix recovery for seismic data acquisition. By considering deterministic results in matrix completion, we propose the spectral gap as a means to decide if a given survey design is suitable for reconstruction via nuclear norm minimization. Our numerical experiments demonstrate a clear correlation between the spectral gap and the quality of reconstruction. We further argue that this tool is multipurpose by providing several realistic scenarios that highlight its flexibility, including its use for coil sampling based techniques. The main advantage of the spectral gap is its relatively low computational complexity, were a practitioner need only compute the first two singular values of a binary matrix. Furthermore, our approach does not require simulation based acquisition design and can be evaluated prior to in field surveys and large-scale optimization. In the end, the spectral gap renders a relatively simple tool that aids the design and execution of highly complex randomized seismic surveys.

REFERENCES

Allouche, N., K. Øzdemir, A. Özbek, and J.-F. Hopperstad, 2020, Multimeasurement aliased-noise attenuation for sparse land seismic acquisition using compressed sensing: Geophysics, **85**, V183–V200.

Aravkin, A., R. Kumar, H. Mansour, B. Recht, and F. J. Herrmann, 2014, Fast methods for denoising matrix completion formulations, with applications to robust seismic data interpolation: SIAM Journal on Scientific Computing, **36**, S237–S266.

Bhojanapalli, S., and P. Jain, 2014, Universal matrix completion: Proceedings of the 31st International Conference on Machine Learning, PMLR, 1881–1889.

Burnwal, S. P., and M. Vidyasagar, 2020, Deterministic completion of rectangular matrices using asymmetric ramanujan graphs: Exact and stable recovery: IEEE Transactions on Signal Processing, **68**, 3834–3848.

Candes, E. J., and Y. Plan, 2010, Matrix completion with noise: Proceedings of the IEEE, **98**, 925–936.

Candès, E. J., and B. Recht, 2009, Exact matrix completion via convex optimization: Foundations of Computational mathematics, **9**, 717.

Candès, E. J., and T. Tao, 2010, The power of convex relaxation: Near-optimal matrix completion: IEEE Transactions on Information Theory, **56**, 2053–2080.

Chen, Y., S. Bhojanapalli, S. Sanghavi, and R. Ward, 2015, Completing any low-rank matrix, provably: The Journal of Machine Learning Research, **16**, 2999–3034.

Christensen, P., A. Kensler, and C. Kilpatrick, 2018, Progressive multi-jittered sample sequences: Computer Graphics Forum, Wiley Online Library, 21–33.

Da Silva, C., and F. Herrmann, 2013, Hierarchical tucker tensor optimization-applications to 4d seismic data interpolation: 75th EAGE Conference & Exhibition incorporating SPE EUROPEC 2013, European Association of Geoscientists & Engineers, cp–348.

Dippé, M. A., and E. H. Wold, 1985, Antialiasing through stochastic sampling: Proceedings of the 12th annual conference on Computer graphics and interactive techniques, 69–78.

Heiman, E., G. Schechtman, and A. Shraibman, 2014, Deterministic algorithms for matrix completion: Random Structures & Algorithms, **45**, 306–317.

Hennenfent, G., L. Fenelon, and F. J. Herrmann, 2010, Nonequispaced curvelet transform for seismic data reconstruction: A sparsity-promoting approach: Geophysics, **75**, WB203–WB210.

Hennenfent, G., and F. J. Herrmann, 2008, Simply denoise: Wavefield reconstruction via jittered undersampling: Geophysics, **73**, V19–V28.

Herrmann, F. J., 2009, Sub-nyquist sampling and sparsity: getting more information from fewer samples.

Herrmann, F. J., M. P. Friedlander, and O. Yilmaz, 2012, Fighting the curse of dimensionality: Compressive sensing in exploration seismology: IEEE Signal Processing Magazine, **29**, 88–100.

Herrmann, F. J., and G. Hennenfent, 2008, Non-parametric seismic data recovery with curvelet frames: Geophysical Journal International, **173**, 233–248.

Hoory, S., N. Linial, and A. Wigderson, 2006, Expander graphs and their applications: Bulletin of the American Mathematical Society, **43**, 439–561.

Ignjatovic, Z., and M. F. Bocko, 2005, A method for efficient interpolation of discrete-time signals by using a blue-noise mapping method: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., IEEE, iv–213.

Jain, P., P. Netrapalli, and S. Sanghavi, 2013, Low-rank matrix completion using alternating minimization: Proceedings of the forty-fifth annual ACM symposium on Theory of computing, 665–674.

Kumar, R., C. Da Silva, O. Akalin, A. Y. Aravkin, H. Mansour, B. Recht, and F. J. Herrmann, 2015, Efficient matrix completion for seismic data reconstruction: Geophysics, **80**, V97–V114.

Kumar, R., O. López, D. Davis, A. Y. Aravkin, and F. J. Herrmann, 2017, Beating level-set methods for 5-d seismic data interpolation: A primal-dual alternating approach: IEEE Transactions on Computational Imaging, **3**, 264–274.

Lazebnik, F., and V. A. Ustimenko, 1995, Explicit construction of graphs with an arbitrary large girth and of large size: Discrete Applied Mathematics, **60**, 275–284.

Lazebnik, F., and A. J. Woldar, 2001, General properties of some families of graphs defined by systems of equations: Journal of Graph Theory, **38**, 65–86.

López, O., R. Kumar, Ö. Yılmaz, and F. J. Herrmann, 2016, Off-the-grid low-rank matrix recovery and seismic data reconstruction: IEEE Journal of Selected Topics in Signal Processing, **10**, 658–671.

Lu, W., W. Li, W. Zhang, and S.-T. Xia, 2018, Expander recovery performance of bipartite graphs with girth greater than 4: IEEE Transactions on Signal and Information Processing over Networks, **5**, 418–427.

Ma, J., 2008, Compressed sensing by inverse scale space and curvelet thresholding: Applied Mathematics and Computation, **206**, 980–988.

———, 2013, Three-dimensional irregular seismic data reconstruction via low-rank matrix completion: Geophysics, **78**, V181–V192.

Moldoveanu, N., 2010, Random sampling: A new strategy for marine acquisition, *in* SEG Technical Program Expanded Abstracts 2010: Society of Exploration Geophysicists, 51–55.

Mosher, C., C. Li, L. Morley, Y. Ji, F. Janiszewski, R. Olson, and J. Brewer, 2014, Increasing the efficiency of seismic data acquisition via compressive sensing: The Leading Edge, **33**, 386–391.

Mosher, C. C., C. Li, F. D. Janiszewski, L. S. Williams, T. C. Carey, and Y. Ji, 2017, Operational deployment of compressive sensing systems for seismic data acquisition: The

Leading Edge, **36**, 661–669.

Oropeza, V., and M. Sacchi, 2011, Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis: Geophysics, **76**, V25–V32.

Özbek, A., M. Vassallo, A. K. Özdemir, D. Molteni, and Y. K. Alp, 2010, Anti-alias optimal interpolation with priors, *in* SEG Technical Program Expanded Abstracts 2010: Society of Exploration Geophysicists, 3401–3405.

Recht, B., 2011, A simpler approach to matrix completion.: Journal of Machine Learning Research, **12**.

Recht, B., M. Fazel, and P. A. Parrilo, 2010, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization: SIAM review, **52**, 471–501.

Recht, B., and C. Ré, 2013, Parallel stochastic gradient algorithms for large-scale matrix completion: Mathematical Programming Computation, **5**, 201–226.

Schonewille, M., A. Klaedtke, A. Vigner, J. Brittan, and T. Martin, 2009, Seismic data regularization with the anti-alias anti-leakage fourier transform: First Break, **27**.

Schonewille, M., Z. Yan, M. Bayly, and R. Bisley, 2013, Matching pursuit fourier interpolation using priors derived from a second data set, *in* SEG Technical Program Expanded Abstracts 2013: Society of Exploration Geophysicists, 3651–3655.

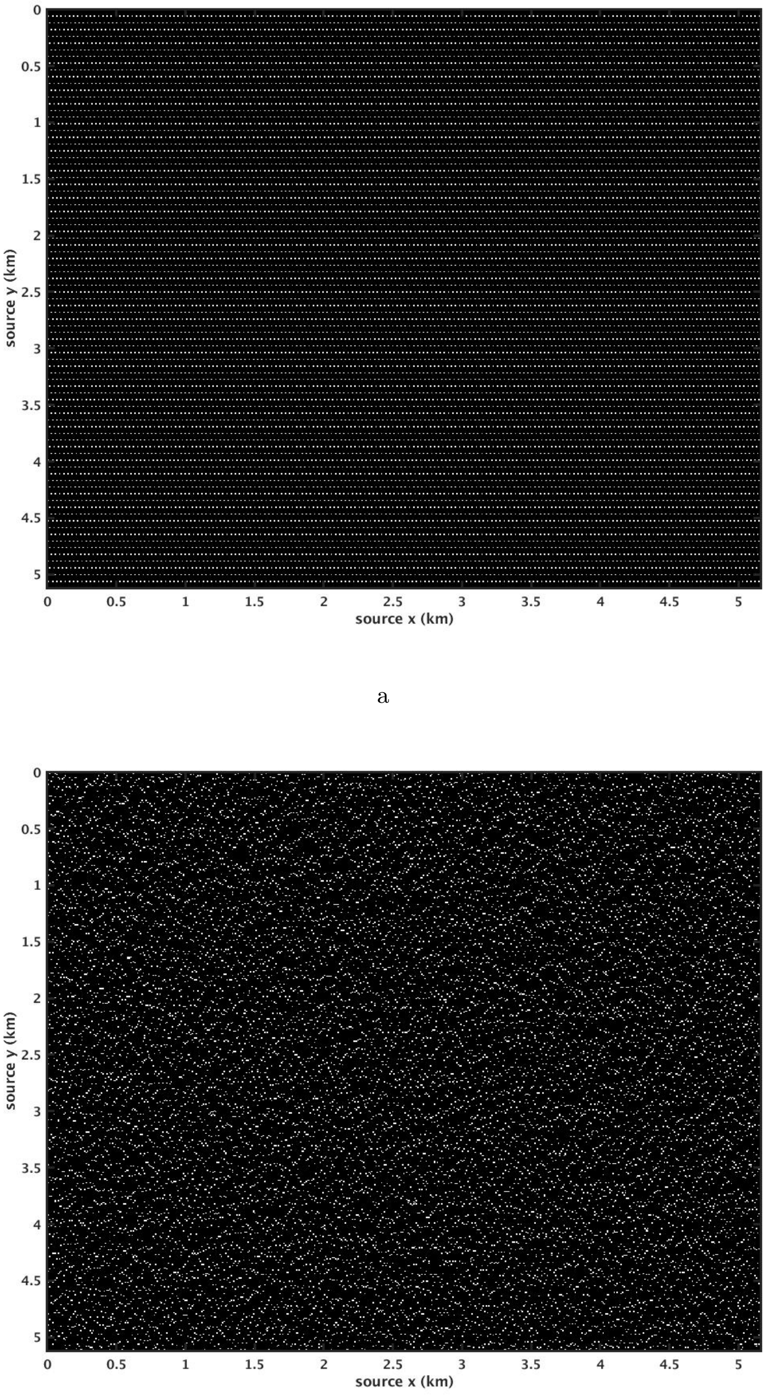
Shahidi, R., G. Tang, J. Ma, and F. J. Herrmann, 2013, Application of randomized sampling schemes to curvelet-based sparsity-promoting seismic data recovery: Geophysical Prospecting, **61**, 973–997.

Trad, D., J. Deere, S. Cheadle, et al., 2005, Challenges for land data interpolation: CSEG Annual Convention, Expanded Abstracts, Citeseer, 309–311.

Watanabe, T., and N. Masuda, 2010, Enhancing the spectral gap of networks by node removal: Physical Review E, **82**, 046102.

Yang, Y., J. Ma, and S. Osher, 2013, Seismic data reconstruction via matrix completion: Inverse Problems & Imaging, **7**, 1379.

Yun, H., H.-F. Yu, C.-J. Hsieh, S. Vishwanathan, and I. Dhillon, 2013, Nomad: Non-locking, stochastic multi-machine algorithm for asynchronous and decentralized matrix completion: arXiv preprint arXiv:1312.0193.



a

b

Figure 5: (a) Periodic and (b) jittered sampling mask design for 3D seismic data interpolation experiment. The underlying interpolation grid is 10m, whereas the data is sampled at 50m and 100m along the inline and the crossline direction, respectively.

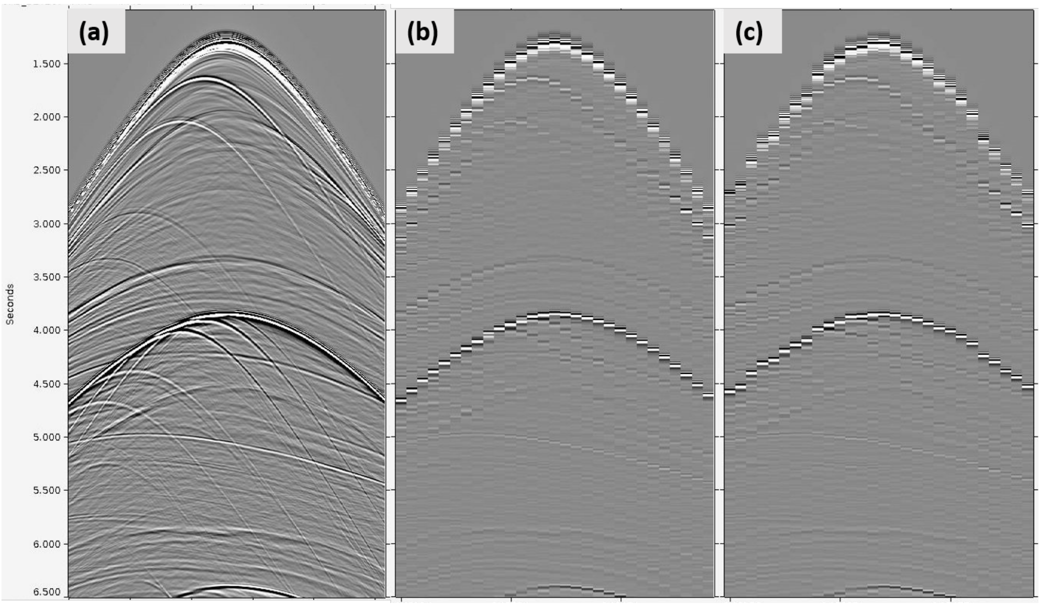


Figure 6: (a) Fully sampled crossline section extracted from the SEAM model. Subsampled crosssection using (b) periodic, and (c) jittered sampling design. The SG ratios ($\frac{\sigma_2(\mathbf{M})}{\sigma_1(\mathbf{M})}$) for (b) and (c) are 1.0 and 0.35, respectivelty. Note that the average sampling rate is same for both the periodic and jittered sampling scenario.

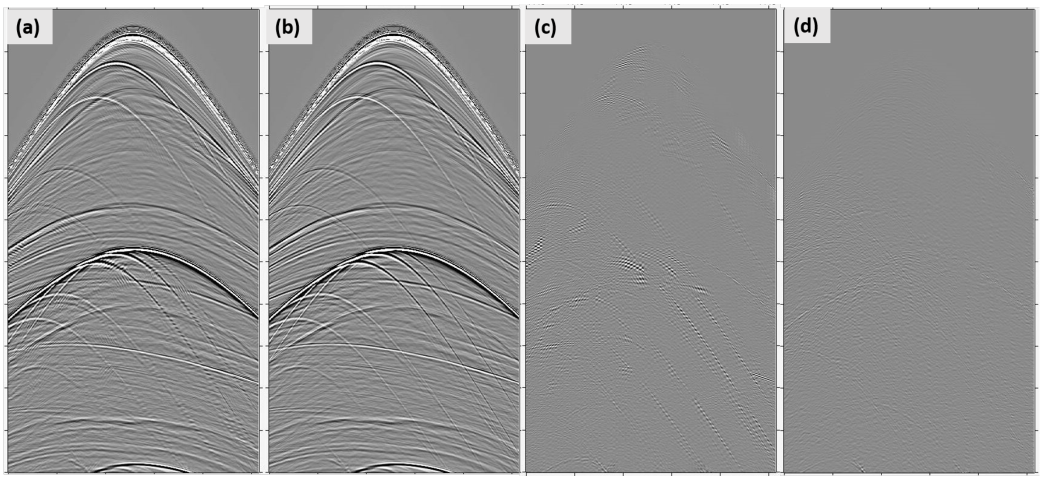


Figure 7: Crossline section from the SEAM model after seismic data reconstruction using (a) periodic, and (b) jittered subsampling, and (c, d) the corresponding residual sections. It is evident that we are able to preserve the reflections and diffraction energy quite well using the jittered sampling, which is also supported by the residual section and the SG.