Enabling large-scale seismic data acquisition, processing and waveform-inversion via rank-minimization

Rajiv Kumar
Seismic data acquisition

Objective

Dense seismic data at regular periodic grid to perform good-quality imaging and inversion

Data Organization (common-shot gather)

2D acquisition: time, source, receiver
3D acquisition: time, source-x, source-y, receiver-x, receiver-y
Acquisition Impediments

- Reduce acquisition time and cost
  - coarser sampling in sequential acquisition
  - deploy multiple-vessels for simultaneous acquisition
Sequential acquisition artifacts

Imaging (before interpolation)  Imaging (after interpolation)
Simultaneous acquisition artifacts

Stacking before source-separation

Stacking after source-separation
Contributions: Chapter 2-5

Designed *computationally scalable* and *memory efficient* large-scale low-rank based seismic data processing method for

- interpolation
- source-separation
Matrix completion

- **signal structure**
  - low rank/fast decay of singular values

- **sampling scheme**
  - missing data increase rank in “transform domain”

- recovery using *rank penalization* scheme

[Candes and Plan 2010, Oropeza and Sacchi 2011]
Low-rank structure

2-D acquisition
Singular value decay
2-D acquisition

![Graph showing the decay of singular values in acquisition and transform domains.](graph.png)
Low-rank structure
conventional 5D data, monochromatic slice, Sx-Sy matricization
Low-rank structure
conventional 5D data, monochromatic slice, Sx-Rx matricization
Matrix completion

- signal structure
  - low rank/fast decay of singular values

- sampling scheme
  - missing data increase rank in “transform domain”

- recovery using rank penalization scheme
2-D acquisition
uniform-random sampling

acquisition domain
missing columns *do not* increase rank

transform domain
missing columns *do* increase rank

(source(m), receiver(m))
Randomized sampling

singular value decay

fully sampled data

random sampled data

- number of singular values
- normalized singular value
- acquisition domain
- transform domain

- Number of singular values
- Normalized singular value
- Acquisition domain
- Transform domain

- Fully sampled data
- Random sampled data
Low-rank structure

Time-jittered data, monochromatic slice, Sx-Sy matricization
Low-rank structure
time-jittered data, monochromatic slice, Sx-Rx matricization
Matrix completion

- signal structure
  - low rank/fast decay of singular values

- sampling scheme
  - missing data increase rank in “transform domain”

- recovery using rank penalization scheme
Nuclear-norm minimization

\[ \min_{X} \|X\|_* \quad \text{s.t.} \quad \|A(X) - b\|_2 \leq \epsilon \]

sum of singular values of \( X \)

\[ X = LR^H \]

\[ X \in \mathbb{C}^{n_f \times n_{rx} \times n_{sx} \times n_{ry} \times n_{sy}} \]

\[ L \in \mathbb{C}^{n_f \times n_{rx} \times n_{sx} \times n_k} \]

\[ R^H \in \mathbb{C}^{n_f \times n_{ry} \times n_{sy} \times n_k} \]
Factorized formulation

- Upper-bound on nuclear norm is defined as

\[ \| L R^H \|_* \leq \frac{1}{2} \left\| \begin{bmatrix} L \\ R \end{bmatrix} \right\|_F^2 \]

where \( \| \cdot \|_F \) is sum of squares of all entries

- choose \( k \) explicitly & avoid costly SVD's
### Computational cost

**matrix completion v/s curvelet-based methods**

#### 2D interpolation

- 1001 time-samples, 355 sources, 355 receivers
- 50% and 75% subsampling scenario

<table>
<thead>
<tr>
<th></th>
<th>$\sigma$</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Matrix completion w/ SVD</strong></td>
<td>SNR (dB)</td>
<td>17.3</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>time (sec)</td>
<td>812</td>
<td>937</td>
</tr>
<tr>
<td><strong>Matrix completion w/o SVD</strong></td>
<td>SNR (dB)</td>
<td>17.6</td>
<td>18.4</td>
</tr>
<tr>
<td></td>
<td>time (sec)</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td><strong>Curvelet-based sparsity promotion</strong></td>
<td>SNR (dB)</td>
<td>17.4</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>time (sec)</td>
<td><strong>879</strong></td>
<td><strong>989</strong></td>
</tr>
</tbody>
</table>
Acquisition setup

- 10s temporal length
- 25 m flip-flop shooting
  - source-sampling ranges from 25 m to 175 m
  - effective 50 m source sampling for each airgun array
  - acquired 400 sources
- 10201 receivers
- Ricker wavelet with central frequency of 20 Hz
- Size of the recovered 5D seismic data volume is 0.3 TB
3D BG Compass model
Optimization information

- Parallelized factorization framework over sources and receivers
- 200 iterations, computational time 42 hours, fixed 100 rank values across frequencies
- Separation + interpolation @ 6.25 m grid, recovered 1600 sources
- SENAI Yemoja cluster: 30 nodes, 128 GB RAM each, 20-core processors
- 300 Parallel Matlab workers (10 per node), multithread - full core utilization
Conventional data
common-shot gather, @6.25 m source sampling
Observed v/s recovered

Observed data @ 25 m flip-flop (overlapping & missing shots)

Separation + Interpolation (recovered grid @ 6.25m)

Continuous data (30 s)
After Source-Separation
common-shot gather, SNR = 21dB, preserved late-arrivals energy
Residual
10x magnified

reconstructed late-arrivals
Benefits of LR framework

- **4X** up-sampling (@ 6.25m) & saving in acquisition time
- Size of final recovered data volume is **0.3 TB**
  - No need to save fully sampled seismic data volume
- Save $L$ and $R$ factors
  - Compression rate is **98%**
  - Size of final compressed 5D seismic volume is ~**7 GB**
Contributions: Chapter 6

Derive two-way wave-equation based factorization principle to form full-subsurface extended-image volume to perform

- Velocity analysis
- Targeted imaging
- QC tool for validating the velocity inversion
What is extended image-volumes?

- Mapping of seismic data recorded at the surface to each and every grid point in the subsurface

- This mapping preserves energy from data-domain to image-domain
Extended images

Given two-way wave equations, source and receiver wavefields are defined as

\[ H(m)U = P_s^T Q \]
\[ H(m)^*V = P_r^T D \]

Organize wavefields in monochromatic data matrices where each column represents a common shot gather.

Express image volume tensor for single frequency as a matrix

\[ E = UV^* \]
Extended images

Too expensive to compute \textit{(storage and computational time)}

Instead, \textit{probe} volume with \textit{tall} matrix $W = [w_1, \ldots, w_l]$

$$\tilde{E} = EW = H^{-1} P_s^T Q D^* P_r H^{-1} W$$

where $w_i = [0, \ldots, 0, 1, 0, \ldots, 0]$ represents \textit{single} scattering points
3D BG Compass model

Experimental details

- 1200 source (75 m spacing)
- 2500 receivers (50 m spacing)
- 5-12 Hz
- OBN acquisition
- peak frequency 15 Hz
- One probing vector
- **1500 times faster than conventional method**
Common-image point gather
Cross section across common-image point gather
Wave-equation based velocity analysis
**WEMVA**

conventional approach

\[ h \]

\[ E \]

\[ \ast \] stand for element-wise multiplication

Biondo & Symes, ’04, Symes 2008, Sava & Vasconcelos, ’11
\[ E \text{diag}(x) \approx \text{diag}(x)E \]

\[ * \text{ matrix-matrix multiplication } * \]
Fast WEMVA w/ randomized probing

Measure the error in some norm

$$\min_{\mathbf{m}} ||E(\mathbf{m})\text{diag}(\mathbf{x}) - \text{diag}(\mathbf{x})E(\mathbf{m})||_F^2$$

The Frobenius norm can be estimated via randomized trace estimation: Avron and Toledo, 2011

$$||A||_F^2 = \text{trace}(A^TA)$$

$$\approx \sum_{i=1}^{K} w_i^T A^T A w_i = \sum_{i=1}^{K} ||A w_i||_2^2$$

where $$\sum_{i=1}^{K} w_i w_i^T \approx I$$
Randomized probing reflection data

true model

initial model
Randomized probing reflection

- **Exact**
- **different color represents different random realization**

\[
\| E \text{diag}(x) - \text{diag}(x)E \|_F^2
\]

\[x \times 10^5\]

\(K = 10\)

\(K = 80\)
Experimental details

350 source (40 m spacing) , 700 receivers (20 m spacing)
5-25 Hz
split-spread acquisition
recording length 6s, sampling interval 4ms
peak frequency 20 Hz
25 LBFGS iterations
100 probing vectors
Marmousi model

True model

Starting model

Inverted model
Observations

100X computational and memory savings while forming the full-subsurface image volumes in 3D

Efficient way to extract informations from image volumes

Very fast (2D/3D) target-imaging tool

60X reduction in memory and computational cost in 2D WEMVA, a step closer to 3D WEMVA
Journal Publications


Thank you!

PhD Advisory Committee
Felix Hermann, Eldad Haber, Chen Greif, Ozgur Yilmaz

PhD Exam Committee
Gordon Slade, Kristopher Innanen, Brian Wetton, Tim Salcudean

Internship/Industry Colleagues
James Rickett, Can Evren Yarman, Troy Thompson, Gary Hampson, Amarjeet Kumar, Nick Moldoveanu

SLIM Group Members

SINBAD Sponsors & NSERC

Family and Friends
especially my parents, wife and son