Machine learning applications to geophysical data analysis

by

Benjamin Bryan Bougher

B.Sc. in Physics, Dalhousie University, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Geophysics)

The University of British Columbia
(Vancouver)

August 2016

© Benjamin Bryan Bougher, 2016
Abstract

The sedimentary layers of the Earth are a complex amorphous material formed from chaotic, turbulent, and random natural processes. Exploration geophysicists use a combination of assumptions, approximate physical models, and trained pattern recognition to extract useful information from complex remote sensing data such as seismic and well logs. In this thesis I investigate supervised and unsupervised machine learning models in geophysical data analysis and present two novel applications to exploration geophysics.

Firstly, interpreted well logs from the Trenton-Black River study are used to train a classifier that results in a success rate of 67% at predicting stratigraphic units from gamma ray logs. I use the scattering transform, a multiscale analysis transform, to extract discriminating features to feed a K-nearest neighbour classifier.

A second experiment frames a conventional pre-stack seismic data characterization workflow as an unsupervised machine learning problem that is free from physical assumptions. Conventionally, the Shuey model is used to fit the angle dependent reflectivity response of seismic data. I instead use principle component based approaches to learn projections from the data that improve classification. Results on the Marmousi II elastic model and an industry field dataset show that unsupervised learning models can be effective at segmenting hydrocarbon reservoirs from seismic data.
Preface

This thesis consists of original research carried out at The University of British Columbia under the supervision of Felix J. Herrmann and the Seismic Laboratory for Imaging and Modelling.

The lab work for this thesis was entirely computational and owes much gratitude to the following open source software projects:

- The MATLAB package ScatNet, used for scattering transform calculations in Chapter 2 (https://github.com/scatnet).
- Jan Thorbecke’s finite difference engine, which was used for seismic modelling in Chapter 3 (https://github.com/JanThorbecke/OpenSource).
- Madagascar framework for reproducible science, used for migration and seismic processing in Chapter 3 (https://github.com/ahay).
- scikit-learn machine learning library, used for classification in Chapter 2, and PCA and clustering computations in Chapter 3 (https://github.com/scikit-learn/scikit-learn).
- The linear operator toolbox SPOT, which was used for matrix inversions.
- The scientific python stack SciPy, which was used for general computations and data visualization throughout the thesis (https://www.scipy.org/stackspec.html)

The robust PCA optimization algorithm in Section 3.3.4 is my own bespoke development.

Versions of Chapter 2 were published as a journal article and an expanded abstract at the 2016 Geoconvention, where it received honourable mention for best student presentation.


A version of Chapter 3 was published as an extended abstract and will be presented at the 2016 SEG annual meeting.

# Table of Contents

Abstract .................................................. ii
Preface .................................................... iii
Table of Contents ......................................... v
List of Figures ............................................. vii
List of Algorithms ......................................... xi
Acknowledgments ........................................... xii
Dedication .................................................. xiii

1 Introduction .............................................. 1
   1.1 Motivation .......................................... 1
   1.2 Background ......................................... 2
   1.3 Overview ............................................ 4

2 Predicting stratigraphic units from well logs using supervised learning .................. 5
   2.1 Multifractal analysis ............................... 7
   2.2 Scattering transform ................................ 11
   2.3 Scattering transform as a convolution neural network .................. 14
   2.4 Application to the Trenton-Black River well logs .................. 16
      2.4.1 Methodology ................................. 16
      2.4.2 Results ....................................... 17
# Table of Contents

2.4.3 Discussion ........................................ 18

3 Reflection seismology analysis as an unsupervised learning problem .................................. 20
   3.1 Angle domain common image gathers by shot profile migration 21
   3.2 Overview of amplitude vs angle analysis ............................. 26
   3.3 Motivation for machine learning ........................................ 29
       3.3.1 Problem formulation ........................................... 33
   3.4 Example on Marmousi II synthetic dataset ........................... 33
       3.4.1 Seismic modelling .............................................. 34
       3.4.2 Principle component analysis .................................. 36
       3.4.3 Clustering ..................................................... 42
       3.4.4 Kernel PCA .................................................. 44
       3.4.5 Robust PCA .................................................. 48
       3.4.6 Discussion .................................................. 54
   3.5 Conclusions .................................................. 56

4 Comparison to dynamic intercept gradient inversion ......................................................... 57
   4.1 Dynamic intercept gradient inversion ................................. 57
   4.2 PCA extended DIGI .............................................. 59
   4.3 Results ......................................................... 60
   4.4 Discussion ..................................................... 61

5 Closing remarks ............................................. 66

Bibliography .................................................... 68
List of Figures

Figure 1.1 Netflix matrix completion. Putting the users along the rows and movies along the columns, the ratings form an incomplete matrix. Predictions can be made by completing the matrix via a generic convex optimization. 3

Figure 2.1 Example of a labeled gamma ray log from the Trenton-Black River dataset. 6

Figure 2.2 Flow diagram of a supervised learning problem. Features are extracted from a labelled data, and which are then split into testing and training sets. The training features and corresponding labels are used to train a classifier, which maps a set of features to a label. This mapping is applied to the testing dataset, where predictions can be compared to the true labels. 7

Figure 2.3 Examples of fractal geometries. Left: Mandelbrot set [“Mandelbrot Set” by Wolfgang Beyer is licensed under CC BY 3.0] Centre: Dendritic copper crystals at 20x magnification [licensed under CC BY SA 2.0]. Right: The coast line of Britain [“Fractal Coast” by Chris Moran (2010), used with permission from author]. 8

Figure 2.4 Singularities about 0 for $\alpha = 0$ (top), $\alpha = 1$ (middle), and $\alpha = 2$ (bottom). 9

Figure 2.5 Multifractal signals and their singularity spectrum. Gaussian white noise (top), bore hole measurement (bottom). Figure adapted from Herrmann (1997). 9
Figure 2.6  The WTMML method applied to a geophysical signal (Herrmann, 1997)). (a) A well log measurement of the compressional wave-speed. (b) The wavelet transform of (a) with modulus maxima lines overlaid. (c) The partition function formed from the modulus maxima lines. (d) The mass exponent $\tau(q)$ estimated from (c). (e) The singularity spectrum $f(\alpha)$ computed from the legendre transform of (d). ................................. 12

Figure 2.7  The scattering transform as a convolutional neural network (Andén and Mallat, 2014). ................................. 13

Figure 2.8  The action of a single window of a 2 layer scattering transform on a signal. ................................. 13

Figure 2.9  Generic 2-layer convolutional neural network (CNN) applied to a 1d-signal. A CNN extracts a feature vector through multiple stages of convolution, non-linearities, and pooling operations. The weights for each convolution filter are learned while training the classifier. ................................. 15

Figure 2.10  KNN classification. The input feature vector is compared to the training vectors and the mode of the labels from the most similar trained vectors is output as the prediction. ................................. 17

Figure 2.11  Confusion matrix of the classifier. The diagonal cells correspond to correct classifications and off-diagonal cells are misclassifications. Cell size corresponds to the number of samples of a particular class. ................................. 18

Figure 3.1  Seismic survey [CC-BY-SA-NC]. Sources and receivers on the surface measure backscattered energy from stratigraphic interfaces. ................................. 22

Figure 3.2  Angle gather migration creates scattering angle dependent images of the subsurface (bottom) from seismic shot records (top). ................................. 23

Figure 3.3  ADCIGs for multiple midpoints. ................................. 24

Figure 3.4  For each shot, source (b) and receiver wavefields (c) are extrapolated (e,f) to each depth. The extrapolated wavefields are correlated for multiple space lags creating offset-domain common-image gathers (h). The ODCIG’s are slant-stacked yielding angle-domain common-image gathers (i). The angle dependent reflectivity response (j) can be analyzed for each point in the model. 25
List of Figures

Figure 3.5  The geometric ray approximation to seismic wave scattering at an interface. .................................................. 27
Figure 3.6  AVA curves generated using the non-linear Zoeppritz equations (Equation 3.5) and the two-term Shuey approximation (Equation 3.6). .................................................. 28
Figure 3.7  IG scatter plot depicting the mudrock line of silliclastic interfaces and the outlying gas saturated sands. The seemingly positive/negative symmetry stems from the top and bottom of each layer. Every point in the scatter plot corresponds to a pixel in the seismic image. .................................................. 30
Figure 3.8  Examples of response curves of strong reflectors from real datasets. .................................................. 32
Figure 3.9  AVA characterization as unsupervised learning. ADCIGs are reshaped into a feature matrix, where each point becomes a row and the angle dependent reflectivity response becomes a column. The columns of the matrix are reduced to allow for visualization and to prevent overfitting. The low dimensional features can be clustered which results in a segmented image. . 34
Figure 3.10 Subset of the Marmousi II elastic model. ............. 35
Figure 3.11 ADCIG for the true Marmousi II reflectivity model. . 36
Figure 3.12 IG crossplot for the true Marmousi II reflectivity model. 37
Figure 3.13 ADCIG for the migrated seismic from the Marmousi II model. .................................................. 38
Figure 3.14 IG crossplot for the migrated Marmousi II model. . 39
Figure 3.15 PCA crossplot for the true Marmousi II reflectivity model. 39
Figure 3.16 Comparison of principle components from Zoeppritz modeled data, migrated data, and the Shuey terms as components. 40
Figure 3.17 PCA crossplot (a) compared to IG crossplot (b) for the migrated Marmousi II model. Note that outliers in (a) are pushed farther from the dense background clusters than in (b). ....... 41
Figure 3.18 Partitional clustering. Points are partitioned by maximizing the likelihood of membership to a distribution. ....... 42
Figure 3.19 Agglomerative clustering. Each point is initially considered to be a cluster. The geometrically closest clusters are joined into a node, which is then considered a new cluster. The algorithm steps through the hierarchy until a set number of clusters is reached. .................................................. 43
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.20</td>
<td>Classification results for physically consistent data.</td>
<td>44</td>
</tr>
<tr>
<td>3.21</td>
<td>Classification of principle components from migrated seismic data. The background trend was initially two positive/negative clusters that were joined into one for plotting purposes.</td>
<td>45</td>
</tr>
<tr>
<td>3.22</td>
<td>Kernel PCA crossplots for increasing values of $c$ for physically consistent data (top), and migrated seismic data (bottom).</td>
<td>47</td>
</tr>
<tr>
<td>3.23</td>
<td>Image segmentation results for kernel PCA applied to physically consistent data.</td>
<td>49</td>
</tr>
<tr>
<td>3.24</td>
<td>Image segmentation results for kernel PCA applied to the migrated seismic data.</td>
<td>50</td>
</tr>
<tr>
<td>3.25</td>
<td>Robust PCA applied the true Marmousi II reflectivity model.</td>
<td>53</td>
</tr>
<tr>
<td>3.26</td>
<td>Robust PCA applied to the migrated seismic from the Marmousi II model.</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Migrated data section from a real field dataset provided the BG group. There is an interpreted gas sand is located at sample 400 and x1 range 150-220.</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>Segmentation results from robust PCA.</td>
<td>61</td>
</tr>
<tr>
<td>4.3</td>
<td>Segmentation results from thresholding $EER$ values extracted using the DIGI algorithm.</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>Segmentation results from thresholding $EER$ values extracted using the DIGI algorithm with principle components replacing the Shuey terms.</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison of EER sections generated from DIGI (top) and PCA extended DIGI (bottom). The region containing the potential reservoir is highlighted in green on the trace plots.</td>
<td>64</td>
</tr>
<tr>
<td>4.6</td>
<td>Principle component vectors extracted from the BG data set (top) compared the Shuey components (bottom).</td>
<td>65</td>
</tr>
</tbody>
</table>
# List of Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>ADCIG by shot profile migration</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Component Pursuit by ADMM</td>
<td>52</td>
</tr>
</tbody>
</table>
Acknowledgments

The entire SLIM group for their support. In particular, I would not have been able to digest the mathematical aspects of my research without the help of Curt Da Silva and Dr. Rongrong Wang.

Charles Jones and James Selvage for pointing me in the direction of AVA analysis, providing their open-source analysis code for comparison, and for providing the processed field data used in Chapter 4.

Finally, I owe extreme gratitude to my supervisor Dr. Felix J. Herrmann for providing the freedom to pursue original research, the flexibility to continue my work outside of the lab, and for his progressive ideas.
Dedication

To Miranda Joyce, for taking care of everything I needed before I knew I needed it. To my roommate Maddy Kipling, for making Vancouver feel like home for 2 years. And finally, to Tim Morgan, whoever you are...
Chapter 1

Introduction

1.1 Motivation

Large scale natural phenomena exhibits complexity beyond any tractable physical model or deterministic computation. As Earth scientists, we are tasked with the impossible challenge of explaining the complex world we live in. We can use laboratories to control complexity and study individual phenomena, or impose assumptions and approximations to make inferences from tractable but simplified models. In either case, we fail to fully explain the large scale system we wish to study.

Complexity exists in many domains outside of the Earth sciences—e.g., financial markets, economics, human behaviour, and systems engineering, where technological advances in data storage and communications has created large data sets about complex systems. This has motivated a new approach to data analysis where in place of directly modelling the behaviour of a complex system, one instead learns predictions and trends directly from the data. These approaches have many names—e.g., machine learning, data science, predictive analytics, but generalize as the application of generic models to large datasets in order to extract useful information from complex systems.

As an example, consider the infamous Netflix problem: Given a large database of movies and users, and a small list of movies that each user has rated, the problem is to predict user ratings for movies they have not
seen. We can try to model a large population’s individual taste, hoping to learn much about the intrinsic relationships between people and their cinematic preferences, but will quickly realize the problem is too complex to be modelled in an objectively correct way. Alternatively, we can try a machine learning approach that uses a general model to extract useful information from the dataset. Forming a matrix with users along the rows and movies along the columns, the user ratings now form an incomplete matrix (Figure 1.1). The complex problem of predicting user taste profiles is abstracted to the general problem of matrix completion, which can be solved using low-rank optimization scheme (Candes and Tao, 2009). This method does not allow us to directly say anything about the intrinsic relationship between users and movies, but it extracts potentially useful information and predictions from the dataset.

In this thesis I study the geophysical data of the sedimentary layers of the Earth, a complex amorphous material formed from chaotic, turbulent, and random natural processes. I argue that exploration geophysics fits in the regime of machine learning models and successfully demonstrate their application to two conventional analysis methodologies:

- Determining stratigraphic units from well log data
- Exploring prestack seismic data for hydrocarbons

1.2 Background

Geophysicists analyze remote sensing data such as seismic and well logs in order to characterize geological and structural properties. Measurements are visually interpreted for patterns and correlations in order to classify regions that are similar in geological structure. This type of pattern recognition fits well into a supervised learning regime where interpreted data can be used to train a classifier, which can predict interpretations on future data sets. Convolutional neural networks (CNN) have shown success in recognizing patterns in images and signals. For example Li et al. (2010) used CNN to extract meaningful features from music for automated genre classification and an approach using CNN achieved the best classification results on the competitive MNIST hand writing benchmark (Ciresan et al., 2012).

The scattering transform (Andén and Mallat, 2014) offers an alternative approach to training a CNN. While it shares the same structure as a
CNN, the nodes are predefined as multiscale wavelet transforms and can be interpreted as multiscale measurements of the input signal. Using the scattering transform has achieved competitive classification results on both music (Andén and Mallat, 2014) and image classification problems (Bruna, 2013). Of particular interest to problems in geophysics, the multiscale nature of the transform is tied to multifractal signal analysis (Bruna et al., 2015). In Chapter 2 I use the scattering transform in the context of supervised machine learning to classify stratigraphic units from well logs.

Unsupervised learning, in the most general sense, is a subfield of machine learning that tries to infer hidden structure within an unlabeled dataset. Unsupervised methods are particularly useful when the inferred structure is lower dimensional than the original data. For example, given a list of

\[
\begin{bmatrix}
8 & 10 & 6 & 2 & \ldots \\
2 & 10 & \ldots \\
7 & 4 & 3 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ldots \\
\end{bmatrix} \rightarrow \begin{bmatrix}
8 & 6 & 10 & 7 & 9 & 6 & 2 & \ldots \\
6 & 10 & 2 & 4 & 1 & 9 & 10 & \ldots \\
8 & 2 & 7 & 7 & 6 & 4 & 2 & \ldots \\
7 & 8 & 10 & 4 & 9 & 3 & 7 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ldots \\
\end{bmatrix}
\]

**Figure 1.1** Netflix matrix completion. Putting the users along the rows and movies along the columns, the ratings form an incomplete matrix. Predictions can be made by completing the matrix via a generic convex optimization.
Chapter 1. Introduction

In patients in a hospital and their corresponding symptoms, it is unlikely that each patient-symptom combination is unique. A set of common disease-symptom combinations can be inferred from the data, where \( d \ll n, s \). Interestingly, interpreted images and geological maps produced by geoscience workflows are substantially lower-dimension than the original field data.

The structure of the sedimentary layers of the Earth is often explained with a simplified model: rock with similar physical properties is formed along relatively continuous interfaces and facies in the subsurface. For this reason, we can use a combination of physical models, local geological knowledge, and experience to reduce large seismic and well log datasets into low-dimensional models of the Earth. Unfortunately, the simplified physical and geological assumptions do not always hold true in practice, making the inferred model highly uncertain and biased. This problem can instead be reformulated using general machine learning models. Abstractly, we are inferring a low-dimensional Earth model from high-dimensional geophysical data. In Chapter 3 I apply unsupervised machine learning models to automatically classify prestack seismic data.

1.3 Overview

This thesis is organized as follows:

Chapter 2 provides an overview of fractal analysis and the motivation for using the scattering transform as a discriminating feature basis for well logs. Interpreted well logs from the Trenton-Black River dataset are used to train a classifier which can predict stratigraphic units at 67% success rate.

Chapter 3 describes reflection seismology as a scattering physics experiment and explains the assumptions and approximations used to measure seismic radiation patterns in the subsurface. I demonstrate that PCA-based approaches can successfully classify regions of the subsurface even when the physical assumptions break down.

Chapter 4 applies the methodologies from Chapter 3 to an industry field dataset where results are compared to directly to an algorithm used by the BG group.

I finish with general conclusions and outlines for future related research.
Chapter 2

Predicting stratigraphic units from well logs using supervised learning

Well logs, measurements made down bore holes, are interpreted by geophysicists in order to characterize the subsurface (Figure 2.1). Stratigraphic units are discovered and correlated across multiple wells by recognizing patterns, textures, and similarities in the data. The manual analysis of well logs is both repetitive and time consuming and would therefore benefit from automated approaches. Supervised learning provides a framework that can train a classifier to predict human interpretations of data.

In a supervised learning experiment, one uses interpreted (labelled) datasets to build a mapping from data to interpretation (Figure 2.2). Training a classifier using raw data samples is often not useful, or even possible, as raw data can be irregularly sized, incomplete, or too cumbersome to be used directly. Data samples are therefore transformed into feature vectors via a feature extraction step. Extracting meaningful features from a dataset requires significant domain expertise and experience, as finding an abstract representation of the data is the most challenging and sensitive aspect of supervised learning.

The sedimentary layers of the Earth are built up by random depositional processes occurring across different timescales. These multiscale processes form fractal relationships in the strata of the Earth, which is described by
Figure 2.1 Example of a labeled gamma ray log from the Trenton-Black River dataset.
Chapter 2. Predicting stratigraphic units from well logs using supervised learning

Figure 2.2 Flow diagram of a supervised learning problem. Features are extracted from a labelled data, and which are then split into testing and training sets. The training features and corresponding labels are used to train a classifier, which maps a set of features to a label. This mapping is applied to the testing dataset, where predictions can be compared to the true labels.

the influential work of Mandelbrot (1977). Furthermore, fractal analysis of well logs by Herrmann (1997) showed that bore hole measurements (well logs) reflect these relationships as multifractal signals. Building on this foundation, I extract multifractal features from well logs to train a classifier that predicts stratigraphic units.

2.1 Multifractal analysis

I limit the scope of this section to an introduction and conceptual overview of multifractal signal analysis and refer to original works for proofs and rigorous treatments. Multifractal analysis is heavily tied to wavelet analysis, which Mallat (1998) provides a superb treatment of.
Chapter 2. Predicting stratigraphic units from well logs using supervised learning

Fractal geometries can be defined as a natural phenomenon or mathematical set that repeats at each scale (Figure 2.3). Fractal relationships have been observed in nature and are often used as a simple way to model complex natural structures such as coast lines (Mandelbrot, 1967) and branching patterns of trees (Morse et al., 1985). The use of fractals to model nature has become ubiquitous in the special effects industry, where intricately textured landscapes can be modeled by simple parametric models (Pentland, 1984). In addition to modeling natural geometries, fractals are a powerful signal analysis tool.

![Fractal Geometries](image)

**Figure 2.3** Examples of fractal geometries. Left: Mandelbrot set [“Mandelbrot Set” by Wolfgang Beyer is licensed under CC BY 3.0] Centre: Dendritic copper crystals at 20x magnification [licensed under CC BY SA 2.0]. Right: The coast line of Britain [“Fractal Coast” by Chris Moran (2010), used with permission from author].

The fractal nature of signals can be studied by analyzing the singularities of a continuous process, where a singularity is defined as a point where the differentiability fails to be well behaved. The local differentiability of a signal $f(x)$ can be described by the Holder exponent (Estrada and Kanwal, 2000) $\alpha(x)$ where

$$|f(x) - f(y)| \leq C|x - y|^{\alpha(x)}$$  \hspace{1cm} (2.1)

for some constant $C$ and $y$ being a point at a small distance from $x$. In this regard, a signal is $\alpha$ times differentiable around the point $x$. Figure 2.4 shows examples of singularities for different values of $\alpha$.

A multifractal signal consists of singular points of differing values of $\alpha$, where the singularity spectrum $D(\alpha)$ describes the distribution of the Holder coefficients (Frisch and Parisi, 1985). Loosely, the singularity spectrum characterizes the *roughness* of a signal. Figure 2.5 shows multifractal signals with differing singularity spectrums.

The Holder exponent at a singularity can be measured by exploiting the scaling properties of the wavelet transform (Mallat and Hwang, 1992, Mallat
Figure 2.4 Singularities about 0 for $\alpha = 0$ (top), $\alpha = 1$ (middle), and $\alpha = 2$ (bottom).

Figure 2.5 Multifractal signals and their singularity spectrum. Gaussian white noise (top), bore hole measurement (bottom). Figure adapted from Herrmann (1997).
Chapter 2. Predicting stratigraphic units from well logs using supervised learning

and Zhong (1992)). The local behaviour of the wavelet transform \( T_\psi \) of a signal \( f \) behaves as

\[
T_\psi[f](x_0, a) \sim a^{\alpha(x_0)}, \quad a \to 0^+
\]

(2.2)

provided that \( n_\psi > \alpha(x_0) \) where \( n_\psi \) is the number of vanishing moments of the analyzing wavelet. The Holder exponent can thus be determined from the slope of the line from the log-log plot of the wavelet coefficients at \( x_0 \).

Directly measuring the Holder exponents at each point in a signal in order to sample \( D(\alpha) \) is not feasible. The method of wavelet transform modulus maxima lines (WTMML) (Frisch and Parisi, 1985, Mallat and Zhong (1992), Bacry et al. (1993), Muzy et al. (1991), MUZY et al. (1994), Muzy et al. (1993)) instead provides a statistical estimation of \( D(\alpha) \) from modulus maxima lines of the wavelet transform.

The scaling behaviour of the wavelet transform is approximated by the partition function

\[
Z(q, a) = \sum_{l \in L(a)} \left( \sup_{(x, a') \in l, a' \leq a} |T_\psi[f](x, a')| \right)^q
\]

(2.3)

where \( \sup_{(x, a') \in l, a' \leq a} |T_\psi[f](x, a')| \) are the modulus maxima lines of the wavelet transform. A scaling exponent \( \tau(q) \) can now be defined from the power-law

\[
Z(q, a) \propto a^{\tau(q)}
\]

(2.4)

of the partition function.

Arneodo et al. (1995) has shown this partition function has analogies to thermodynamics, where \( \tau(q) \) and \( q \) play the role of the inverse temperature and free energy. For the conjugate variables energy and entropy, one has the Holder exponent \( \alpha \) and singularity spectrum \( D(\alpha) \) which can be determined by the legendre transform

\[
D(\alpha) = \min_q (q \alpha - \tau(q)).
\]

(2.5)

Figure 2.6 shows the WTMML method applied to a multifractal signal.

Using a fractal approach to the analysis of well log data is an active field of research. Khue et al. (2002) found that texture logs formed from
multifractal measurements of microresistivity logs were indicators of heterogeneity when compared to corresponding core samples. Alvarez et al. (2003) found the scaling behaviour of wavelet coefficients distinctly different for gravel than for sands. Ouadfeul and Aliouane (2013), Aliouane and Ouadfeul (2014), Aliouane et al. (2012) have directly used the WTMML method in combination with self-organizing maps to automatically segment well-logs.

Although WTMML is a tractable method for fractal characterization and analysis, it is difficult to measure $D(\alpha)$ in local windows, which is required for a useful feature representation in a supervised learning problem.

### 2.2 Scattering transform

The scattering transform (Andén and Mallat, 2014), a cascade of wavelet transforms and modulus operators which operates on local data windows, provides an alternative method for multifractal signal analysis where Bruna et al. (2015) demonstrated the first two layers of the scattering network carry discriminating multifractal signal information. Further more, Andén and Mallat (2011) and Bruna and Mallat (2010) demonstrated the effectiveness of the scattering transform as a feature basis in image and signal classification.

Given a discrete signal $x[n] = x(n\delta t)$ for $n = 1 \ldots N$, the discrete wavelet transform $\psi_\lambda$ uses a bandpass filter bank $\lambda$ to decompose $x[n]$ into $J$ bandlimited signals. Analogous to the Fourier transform, which decomposes a signal into frequencies, the wavelet transform is considered to be a decomposition into scales and translations. The scattering transform exploits multiscale relationships in signals using a cascade of wavelet transforms, magnitude, and averaging operations (Figure 2.7). Figure 2.8 shows the action of one scattering transform window on a signal.

At the zero-level of the transform, the $S_0$ output coefficients are simply $x[n]$ averaged by a window $\phi$—i.e.,

$$S_0[n] = x[n] * \phi[n].$$  \hspace{1cm} (2.6)

The first layer of the network is formed by taking the magnitude of the wavelet transform $\Psi_\lambda_1$ of $x[n]$ and the $S_1$ output coefficients are then generated by averaging along each scale in the filter bank $\lambda_1$ by the window
Figure 2.6 The WTMML method applied to a geophysical signal (Herrmann, 1997)). (a) A well log measurement of the compressional wave-speed. (b) The wavelet transform of (a) with modulus maxima lines overlaid. (c) The partition function formed from the modulus maxima lines. (d) The mass exponent $\tau(q)$ estimated from (c). (e) The singularity spectrum $f(\alpha)$ computed from the Legendre transform of (d).

\[ \phi - \text{i.e.,} \]
\[ S_1[n, \lambda_1] = |x[n] \ast \Psi_{\lambda_1} \ast \phi[n]|. \quad (2.7) \]

Referring to Figure 2.8, each window results in $J S_1$ output coefficients, where $J$ is the number of scales in $\lambda_1$. 
Chapter 2. Predicting stratigraphic units from well logs using supervised learning

Figure 2.7 The scattering transform as a convolutional neural network (Andén and Mallat, 2014).

Figure 2.8 The action of a single window of a 2 layer scattering transform on a signal.
The second layer of the transform is realized by taking the magnitude of a second wavelet transform $\Psi_{\lambda_2}$ of each signal in the first layer of the network. Averaging is performed to make the $S_2$ output coefficients—i.e.,

$$S_2[n, \lambda_1, \lambda_2] = ||x[n] * \Psi_{\lambda_1} * \Psi_{\lambda_2} * \phi[n]||.$$  \hfill (2.8)

Again referring to Figure 2.8, the $S_2$ output contains JL coefficients where $J$ and $L$ are the number of scales in $\lambda_1$ and $\lambda_2$. Note that the second layer contains multiscale mixing coefficients $\lambda_{1,2,l}$. The notation $\lambda_{1,2,l}$ refers to $l$th scale of $\lambda_2$ operating on the $j$th scale of $\lambda_1$. This process is repeated to a defined network depth of $m$—i.e.,

$$S_m[n, \lambda_1, \lambda_2, \ldots, \lambda_m] = ||||x[n] * \Psi_{\lambda_1} * \Psi_{\lambda_2} * \cdots * \Psi_{\lambda_m} * \phi[n]||.$$  \hfill (2.9)

### 2.3 Scattering transform as a convolution neural network

Convolutional neural networks (CNNs) have achieved state of the art results in perceptual classification problems (Li et al., 2010; Garcia and Delakis, 2004; Nasse et al., 2009). Each level in a CNN consists of convolutions, non-linearities, and pooling operations (Figure 2.9). The final level of a CNN outputs a feature representation of the input signal, which is then fed into a classifier.

In each convolution layer, a filter bank of $J$ trainable kernels are fed by $K$ feature maps, where $x_i$ for $i = 1, \ldots, K$ are the feature maps output by the previous network level. For the first level, $x$ is just the raw input signal. For each kernel $k_j \in \mathbb{R}^l$ where $l$ is the length of the kernels in the filter bank, the output feature map

$$y_j = b_j + \sum_i k_j * x_i$$  \hfill (2.10)

is computed, where $b_i$ is a trainable bias parameter. The output feature maps are then fed to the non-linearity layer.

Conventionally, the non-linearity layer is simply a pointwise sigmoid function applied to every point in each feature map output by the convolution layer. However, any pointwise non-linearity function can be applied at this stage. It is common for the non-linearity function to contain a trainable
Chapter 2. Predicting stratigraphic units from well logs using supervised learning

Figure 2.9 Generic 2-layer convolutional neural network (CNN) applied to a 1d-signal. A CNN extracts a feature vector through multiple stages of convolution, non-linearities, and pooling operations. The weights for each convolution filter are learned while training the classifier.

Parameter, such as the rectified sigmoid $z_j = |g_j \cdot \tanh(y_j)|$ (LeCun et al., 2010) that contains a trainable gain parameter $g_j$. The output feature maps $z_j$ for $j = 1, \ldots, J$ are the same size as the output from the convolution layer.

In the pooling layer, a neighbourhood average $P_A(z_j)$ is computed independently on each feature map $z_j$. Implicit in the pooling layer is a downsampling operation that is applied to remove the redundancy caused by local averaging.

The output of the pooling layer in the final level of the network is concatenated into the final feature representation of the data. The feature vector is output into a classifier, which is typically a logistic classifier like softmax that assigns a probability of the signal belonging to each available class.

The filter weights and bias terms for every kernel in every level of the network become a large set of trainable parameters which are determined using stochastic gradient descent. The gradients for each level are computed using the backpropagation method, which is described in Rumelhart et al. (1986). Due to the large number of parameters in a CNN, CNNs are useful only when a very large training database is available.

The scattering transform shares much of the same architecture as a CNN, but the biggest difference is the kernel weights are predefined wavelet transforms instead of being trained. The non-linearity stages are simply magnitude operations, and the pooling operator is defined by $\phi$. The scattering
network also differs in that a feature representation is output at every level, instead of just the final level. The benefit of the scattering transform is that it has the same structure as a deep network, but does not require training and maintains a direct interpretation. The scattering transform is also useful for feature extraction for classification problems that do not have large databases of training data.

2.4 Application to the Trenton-Black River well logs

The fractal nature of well logs and the scattering transform as a practical fractal analysis tool motivates the use of the scattering transform as a feature representation of well logs in a machine learning problem.

In the early 2000s, the Trenton Black River carbonates experienced renewed interest due to speculation among geologists of potential producing reservoirs. A basin wide collaboration to study the region resulted in many data products, including well logs with corresponding stratigraphic analysis. The dataset contained 80 gamma-ray logs with corresponding stratigraphic interpretations (labels). Although the region contained more units, some were too thin and pinched out to allow for valid signal analysis. The 5 most prominent units (Black River, Kope, Ordovician, Trenton/Lexington, and Utica) were used for analysis. An example of a labelled log is shown in Figure 2.1. The dataset can be downloaded at http://www.wvgs.wvnet.edu/www/tbr/.

2.4.1 Methodology

I transform the well logs into feature vectors, where each scattering window becomes a sample and the $S_0$, $S_1$, and $S_2$ scattering coefficients form the feature vector. This study used a two-layer scattering network, as the number of coefficients becomes unmanageable for deep networks (large $m$) and Bruna (2013) found only marginal classification improvement for $m > 2$. I use a Dyadic wavelet bank of Morlet wavelets for $\Psi_{\lambda m}$ and found through trial that 256 samples was a good choice in scattering window size. The opensource MATLAB implementation ScatNet was used for calculating the scattering transforms. I apply a K-Nearest Neighbours (KNN) classifier (Figure 2.10)
that uses Euclidean distance as the measure of similarity. Cross-validation determined a K-value of 5.

**Figure 2.10** KNN classification. The input feature vector is compared to the training vectors and the mode of the labels from the most similar trained vectors is output as the prediction.

### 2.4.2 Results

The classifier had an overall prediction rate of 65%. The confusion matrix (Figure 2.11) shows the prediction and misclassification rates for each unit type. The thicker units, which inherently contained more data samples, showed the best performance. This is expected, as having more training data for a particular label increases the chances of having similar vectors and better matches in the KNN classifier. The two thickest units, Black River and Ordovician, had the best prediction rates. Figure 2.1 and Figure 2.11 show that sequential units in the stratigraphic record were the most likely to be misclassified as each other, for example Trenton Lexington and Black River.
These units may actually come from the same or similar depositional systems and would share similar multiscale structure in the well logs. More detailed assessment of the geology and the method used to interpret stratigraphy is required in order to better understand these results.

![Confusion matrix of the classifier. The diagonal cells correspond to correct classifications and off-diagonal cells are misclassifications. Cell size corresponds to the number of samples of a particular class.](image)

**Figure 2.11** Confusion matrix of the classifier. The diagonal cells correspond to correct classifications and off-diagonal cells are misclassifications. Cell size corresponds to the number of samples of a particular class.

### 2.4.3 Discussion

Assessing the results in a quantitative matter is highly uncertain, as the labels in the “truth” data set are inherently subjective. The work of Bond (2015), as well as recent work by Hall et al. (2015) has shown significant deviations of interpretations on the same section of data. Additionally, what a geologist interprets as a stratigraphic unit can be somewhat arbitrary, and may be based on subjective information other than textural patterns in the data.
The scattering transform is highly parameterized and results would likely be sensitive to choices of wavelet banks and window sizes at each level of the transform. This basic study used dyadic Morlet wavelets at each level, but the scattering transform may use different banks in each layer of the network. Multifractal analysis is highly sensitive to the number of vanishing moments in the analyzing wavelet, as the vanishing moments define the upper limit of the Hölder exponent that can be measured. The optimal analyzing wavelet at each layer can be determined in a rigorous manner by analyzing the Hölder continuity of the signal in each layer of the network. Alternatively, a pragmatic approach of using cross-validation on larger datasets could be beneficial in determining optimal parameterization.

Due to availability of data, this study was limited to the gamma ray log and did not make use of any a priori knowledge of the region. Incorporating additional log measurements (i.e. sonic, resistivity, etc) would help corroborate the classification as stratigraphic tops can be correlated on many different log measurements. A priori information such as stratigraphic ordering and thickness are reasonable assumptions that should be incorporated as features or penalties in the classification problem in order to boost classification accuracy.

There is a strong correlation between number of samples and classification accuracy, which indicates the need for more data. Generally, training a classifier requires many samples of labelled data. The availability of open labelled datasets for research has proved to be a valuable resource in many other fields. The TIMIT dataset exists for acoustic-phonetic voice classification and GTZAN dataset supports machine learning for music genre classification. The MNIST digit recognition dataset provides a popular benchmark for hand-written digit recognition, while the CURET and OUTFTEX sets provides labelled data for image texture classification. Kaggle is a popular website where datasets from many industries are made available as open machine learning competitions. The sensitive nature of interpreted geophysical makes public datasets in geophysics extremely scarce, but without useful data it is difficult to realize the same supervised learning progress experienced in other fields.
Chapter 3

Reflection seismology analysis as an unsupervised learning problem

In hydrocarbon exploration, geophysicists use reflection seismology to create images of the sedimentary layers. Using wave physics to model seismic wavefield propagation and scattering, physical rock properties can be inferred from the data that are then used to locate hydrocarbon reserves. In reality, the acquisition and processing of seismic data requires many a priori assumptions and physical approximations that inhibit calibrated measurements, making the inferred physical properties highly uncertain. As an alternative to physical models, I hypothesize that useful information can be extracted directly from the data using unsupervised learning methods.

In this chapter I describe reflection seismology as a scattering physics experiment, where a measured radiation pattern is used to characterize a material. I explain the physical assumptions required to measure and classify seismic scattering patterns at stratigraphic interfaces in the subsurface and discuss their limitations. I then propose unsupervised machine learning models that exploit low-dimensional structure directly from the data and demonstrate their application to synthetic datasets. In this experiment I segment prestack seismic data using:

1. Clustering of principle component projections extracted from conventional principle component analysis (PCA)
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

2. Clustering of principle component projections extracted from non-linear kernel principal component analysis (kPCA)

3. Direct image segmentation by matrix splitting using an extension of robust principal component analysis (rPCA).

3.1 Angle domain common image gathers by shot profile migration

Seismic reflection data is acquired by firing an impulsive source and recording the backscattered wavefield using receiver arrays placed on the surface (Figure 3.1). The geometry of a seismic survey restricts the location of sources and receivers to the surface, which prevents the direct measurement of subsurface scattering patterns. Data therefore needs to be migrated, a process that reverses the physics of propagation and maps a reflected wavefield recorded at the surface back to the point of reflection. Assuming adequate acquisition aperture and an estimate of propagation velocity, data can be migrated into reflection-angle dependent images of the subsurface (Figure 3.2) referred to as an angle-domain common-image gathers (ADCIGs). See Figure 3.3 for angle dependent images for different midpoints. These angle dependent images of reflectivity can be analyzed to characterize the rock properties of the sedimentary layers.

Consider a seismic reflection where the source wavefield propagates to a discontinuity in the Earth, scatters based on the physical properties at the discontinuity, and then the backscattered wavefield propagates to a receiver at the surface. We can use a physics model to extrapolate the recorded wavefield and source to the scattering interface. The source and receiver wavefield are related by the scattering response of the interface, and therefore correlating the wavefields provides an estimate of the reflectivity strength (Figure 3.4).

The physics of seismic wave propagation is best described by the elastic wave equation

$$\rho \ddot{u} = f + \nabla \lambda (\nabla \cdot u) + \nabla \mu : [\nabla u + (\nabla u)^T] + (\lambda + 2\mu) \nabla (\nabla \cdot u) - \mu \nabla \times (\nabla \times u)$$ (3.1)

where $\rho$ denotes the spatial-dependent density, $\lambda$ and $\mu$ are the spatial-dependent Lame parameters describing the elastic properties of the medium, $f$ is the spatial and time-dependent vector source function, and $u$ is spatial...
and time-dependent vector wavefield. Due to the computational complexity of migration, approximations to the wave physics are required to make the wavefield extrapolations feasible. Firstly, we use the density independent acoustic wave equation

\[ \nabla^2 u - \frac{1}{V_p^2} \frac{\partial^2 u}{\partial t^2} = f \]  

(3.2)

where \( u \) is the spatial and time-dependent scalar wavefield, \( V_p \) is the pressure wave velocity, and \( f \) is the source function. The computational load is further reduced by only considering one-directional wave propagation. The single square-root equation

\[ P_{z+\Delta z}(\omega, k_x, k_y) = P_z(\omega, k_x, k_y) e^{ik_z \Delta z} \]  

(3.3)
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

(a) Shot record a position $x_0$.  

(b) Shot record a position $x_1$.  

(c) Shot record a position $x_2$.  

(d) Image for scattering angle $\theta_0$.  

(e) Image for scattering angle $\theta_1$.  

(f) Image for scattering angle $\theta_2$.  

Figure 3.2 Angle gather migration creates scattering angle dependent images of the subsurface (bottom) from seismic shot records (top).

is used to extrapolate wavefields, where $P_z(\omega, k_x, k_y)$ is the Fourier transform of the wavefield $u_z(t, x, y)$ and

$$k_z = -\sqrt{\frac{\omega^2}{v(z, x, y)} - (k_x^2 + k_y^2)}$$

(3.4)

is the dispersion relationship for the one-way wave equation (Chapter 4 Biondi (1998), Chapter 6 Berkhout (1980)). The extended split-step algorithm described by Stoffa et al. (1990) and Kessinger (1992) is a computationally efficient implementation of one-way wavefield extrapolation that is valid for horizontally varying velocities.

A simple correlation for each point in the subsurface model yields an image of the zero-offset reflectivity strength. We desire the entire angular dependent reflectivity response, which requires imaging at multiple offsets. This is accomplished by extracting multiple space lags during wavefield correlation, yielding a reflectivity measurement along subsurface offsets $h$ referred
Figure 3.3 ADCIGs for multiple midpoints.

To as an offset-domain common-image gather (ODCIG). Having reflectivity measurements along subsurface offset allows us to measure $\frac{\partial z}{\partial h}$, which is geometrically related to angle via $\tan \theta = -\frac{\partial z}{\partial h}$. where $\frac{\partial z}{\partial h}$ can be calculated from a slant stack. See Algorithm 3.1 and Figure 3.4. I refer to Sava and Fomel (2003) for a full treatment of the approach.

Algorithm 3.1 ADCIG by shot profile migration

1. for each shot do
2. for each depth $z$ do
3. compute src$_z$ = forward_extrapolate(shot, $z$)
4. compute rec$_z$ = backward_extrapolate(data, $z$)
5. compute odcig[$z$]=correlate(src$_z$, rec$_z$)
6. end for
7. end for
8. compute adcig = slant_stack(odcig)
7. output: adcig
Figure 3.4 For each shot, source (b) and receiver wavefields (c) are extrapolated (e,f) to each depth. The extrapolated wavefields are correlated for multiple space lags creating offset-domain common-image gathers (h). The ODCIG’s are slant-stacked yielding angle-domain common-image gathers (i). The angle dependent reflectivity response (j) can be analyzed for each point in the model.
3.2 Overview of amplitude vs angle analysis

Given measurements of the angle-dependent reflectivity, we need to infer physical rock properties in order to classify the subsurface. Assuming that the seismic wavelength of the source is small compared to physical variations in the Earth, we can use ray theory to describe the scattering pattern at a stratigraphic interface (Figure 3.5). The angular dependence of transmitted and reflected seismic rays for an incident p-wave is described by the Zoeppritz equations (Zoeppritz, 1919)

\[
\begin{bmatrix}
R_{PP} \\
R_{PS} \\
T_{PP} \\
T_{PS}
\end{bmatrix} =
\begin{bmatrix}
-\sin \theta_1 & -\cos \phi_1 & \sin \theta_2 & \cos \phi_2 \\
\cos \theta_1 & -\sin \phi_1 & \cos \theta_2 & -\sin \phi_2 \\
\sin 2\theta \frac{V_{P1}}{V_{S1}} \cos 2\phi_1 & \frac{\rho_2 V_{S2}}{\rho_1 V_{S1}} V_{P1} \cos 2\theta_1 & \frac{\rho_2 V_{S2}}{\rho_1 V_{S1}} V_{P1} \cos 2\phi_2 \\
-\cos \phi_2 \frac{V_{S1}}{V_{P1}} \sin 2\phi_1 & \frac{\rho_2 V_{S2}}{\rho_1 V_{S1}} V_{P1} \sin 2\phi_2
\end{bmatrix}
\]

\(3.5\)

where \(R_{PP}\) and \(R_{PS}\) denote the p-wave and shear wave reflection coefficients, \(T_{PP}\) and \(T_{PS}\) denote the coefficients of transmission, \(V_S\) and \(V_P\) denote the shear and P-wave velocities, and \(\rho\) denotes the density.

The Zoeppritz equations provide a physical model that relates angle dependent reflectivity to \(V_P, V_S,\) and density. Unfortunately these are highly non-linear relationships that make direct inversion prohibitive. Shuey (1985) provides a simplification that linearizes the Zoeppritz equations for small parameter perturbations over a constant background. For scattering angles < 30 degrees (Figure 3.6), the non-linear Zoeppritz equations can be modeled by the two-term Shuey approximation

\[
R_{pp}(\theta) = i(\Delta V_P, \Delta \rho) + g(\Delta V_P, \Delta V_S, \Delta \rho) \sin^2 \theta
\]

\(3.6\)

where

\[
i(\Delta V_P, \Delta \rho) = \frac{1}{2} \left( \frac{\Delta V_P}{\langle V_P \rangle} + \frac{\Delta \rho}{\langle \rho \rangle} \right),
\]

\(3.7\)

\[
g(\Delta V_P, \Delta V_S, \Delta \rho) = \frac{1}{2} \frac{\Delta V_P}{\langle V_P \rangle} - 2 \frac{\langle V_S \rangle^2}{\langle V_P \rangle^2} \left( \frac{\Delta \rho}{\langle \rho \rangle} + 2 \frac{\Delta V_S}{\langle V_S \rangle} \right),
\]

\(3.8\)

and \(\langle V_P \rangle, \langle V_S \rangle,\) and \(\langle \rho \rangle\) are estimates of the background. The Shuey formulation is linear with respect to \(i\) and \(g,\) which can be determined from the least-squares problem.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

\[
\min_m \| A m - d \|^2_2 \tag{3.9}
\]

where

\[
A = \begin{bmatrix}
1 & \sin^2 \theta_1 \\
\vdots & \vdots \\
1 & \sin^2 \theta_k
\end{bmatrix}, \tag{3.10}
\]

\[
m = \begin{bmatrix} 1 \\ g \end{bmatrix} \tag{3.11}
\]

and \(d\) is the vectorized ADCIG. In this regard, Shuey term inversion projects an angle gather dataset \(d \in \mathbb{R}^{nk}\) where \(n\) is the number of samples in the Earth model and \(k\) is the number of angles onto two components \(i, g \in \mathbb{R}^n\).

Figure 3.5 The geometric ray approximation to seismic wave scattering at an interface.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.6 AVA curves generated using the non-linear Zoeppritz equations (Equation 3.5) and the two-term Shuey approximation (Equation 3.6).

Measurements by Castagna et al. (1985) and Gardner et al. (1974) of shales and sandstones showed the empirical relationships

\[
\frac{\Delta \rho}{\langle \rho \rangle} \propto k \frac{\Delta V_P}{\langle V_P \rangle} \tag{3.12}
\]

and

\[
V_P = mV_S + c. \tag{3.13}
\]

As a corollary, \(i\) and \(g\) are linearly related by the mudrock line

\[
g = \frac{i}{1 + k} \left[ 1 - 4 \frac{\langle V_S \rangle}{\langle V_P \rangle} \left( \frac{2}{m} + k \frac{\langle V_S \rangle}{\langle V_P \rangle} \right) \right] \tag{3.14}
\]

where \(m\), \(c\), and \(k\) are constants empirically determined from well logs or laboratory measurements. Interestingly, hydrocarbon saturated sandstones have different correlation constants than sandstones saturated with brine.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Analyzing the multivariate properties of \( i \) and \( g \) therefore provides a methodology for inferring hydrocarbon reserves directly from seismic data.

Consider the situation of imaging the sedimentary layers where reflection interfaces are primarily formed from shale and brine-saturated sandstones. Reflections from hydrocarbon-saturated sands are statistically rare. Castagna et al. (1998) exploits the multivariate correlations between \( i \) and \( g \) by crossplotting, which leads to a background trend of shale/brine-sand reflections and statistical outliers that correspond to potential sandstones with trapped hydrocarbons. Figure 3.7 shows a crossplot of Shuey terms calculated from a distribution of shale/brine-sand interfaces and shale/gas-sand interfaces. Rock properties and mudrock line parameters used are from Castagna et al. (1998) and are shown Table 3.1 and Table 3.2.

<table>
<thead>
<tr>
<th>Lithology</th>
<th>( V_P )</th>
<th>( V_S )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shale</td>
<td>3240 ± 50</td>
<td>1620 ± 50</td>
<td>2340 ± 50</td>
</tr>
<tr>
<td>Brine sand</td>
<td>2590 ± 200</td>
<td>1060 ± 200</td>
<td>2210 ± 200</td>
</tr>
<tr>
<td>Gas sand</td>
<td>2540 ± 20</td>
<td>1620 ± 20</td>
<td>2090 ± 20</td>
</tr>
</tbody>
</table>

Table 3.1 Rock settings used for Figure 3.7.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>1360 ± 136</td>
</tr>
<tr>
<td>( k )</td>
<td>0.25 ± .05</td>
</tr>
<tr>
<td>( m )</td>
<td>1.16 ± .116</td>
</tr>
</tbody>
</table>

Table 3.2 Mudrock line parameters used for Figure 3.7.

3.3 Motivation for machine learning

Machine learning should be applied to a scientific problem when:

1. There exists an underlying but unknown relationship between input data and desired output.

2. There is no known physical model to describe the relationship, or the physical model requires too many unrealistic assumptions and approximations.
In Section 3.1 and Section 3.2 I described the physics model that relates hydrocarbon-saturated sediments to seismic data. The physics relies on a simplified model of the Earth, which in reality is a complex amorphous material that has no known exact physical model. The following assumptions are imposed in order to apply a physical model to seismic data analysis:

1. Wave propagation is entirely described by the acoustic wave equation with no multiple reflections (migration approximation).
2. The sedimentary layers of the Earth are flat or show small lateral variations (one-way wave equation approximation).
3. The source wavelength is small compared variations in the Earth (ray theory approximation).
4. The survey design and receiver aperture provides a full angular illumination to each point in the subsurface.
5. Reflection amplitudes are preserved by the migration operator.

6. An accurate a priori velocity model is available.

7. The effect of limited source bandwidth can be removed (perfect deconvolution).

8. There is minimal noise.

Seismic data is collected in uncontrolled environments, such as ocean surveys, where noise and environmental factors play an important and unquantifiable role. Additionally, preprocessing seismic data to meet the assumptions listed above are each active research fields that have their own uncertainties and limitations. For example, deconvolution aims to remove the effects of the source wavelet by solving a linear inverse problem. The finite-bandwidth of the source wavelet makes an ill-posed problem that requires regularization by prior information such as sparsity (Ooe and Ulrych, 1979, Chapman and Barrodale (1983)) or smoothness (Edgar and Selvage, 2013). Both are biases that can not be fully justified. Additionally, deconvolution assumes the source is known, which in reality is estimated from the data and carries its own error. Advanced methods for removing multiple reflections from data such as surface-related multiple attenuation (Verschuur et al., 1992) or estimation of primaries by sparse-inversion (Groenestijn, 2010) provide methods for removing both internal multiples and surface-related multiples. These methods require dense and regular acquisition geometries that are not practical for the scale of modern acquisitions and thus require interpolation in order to be effective. Although anecdotal, I argue each seismic survey comes with unique acquisition and processing issues that inhibit calibrated measurements with quantifiable uncertainty. Figure 3.8 shows angle dependent curves from industry processed seismic data, which exemplify the uncertainty inherent in real world seismic data. Even the large discrepancy in scales in Figure 3.8 shows that even simple variability challenges exist between datasets. The seeming inability of conventional methods to adhere to the assumptions required by the physical model motivates new data-driven approaches.

Now reconsider the Shuey term analysis method described in Section 3.2. In the most general sense, Shuey term inversion is projecting a length $k$ vector (angle dependent reflectivity curve) onto a length 2 vector $(i, g)$. The crossplot analysis described by Castagna et al. (1998) can be abstracted as finding outlying points(hydrocarbon-saturated sandstones) against a back-
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

(a) AVA response from the Penobscot open dataset. 
(b) AVA response from the Mobil Viking dataset. 
(c) AVA response from the BG AVA dataset used in Chapter 4.

Figure 3.8 Examples of response curves of strong reflectors from real datasets.

ground trend (mudrock line). We can now use the vernacular of unsupervised learning to restate Shuey term analysis without the need of a physical model, where we apply dimensionality-reduction followed by clustering. The remainder of this section demonstrates that projections learned directly from seismic data can be successful in discriminating potential hydrocarbon reserves in seismic images.

Previous work by Hami-Eddine et al. (2012) used self-organizing maps to directly classify angle gathers without using the Shuey terms as a physical model. Instead, I closely follow the conventional analysis procedure but replace the physical model with projections learned from the data.
3.3.1 Problem formulation

Given angle-domain common-image gathers, we desire a segmented output image where each point in the Earth model, or equivalently pixel in the image, is classified according to the angle-dependent scattering response. This can be accomplished using unsupervised machine learning models (Figure 3.9).

For an unsupervised learning problem, we define a dataset in terms of samples and features, where samples are individual objects that are described by a vector of features. In the case of ADCIGs, each point in the Earth is a sample $i$ described by a feature vector $x_i \in \mathbb{R}^k$ consisting of the length $k$ angle-dependent reflectivity response. The feature vectors are formed into a matrix $X \in \mathbb{R}^{n \times k}$ with samples along the rows and features along the columns. In this regard, we simply reshape our ADCIGs into a matrix where each pixel of the image becomes a row and each angle becomes a column. Generalizing the data as a feature matrix allows us to work in an unsupervised learning framework.

Assuming the existence of a lower-dimensional representation of the columns of $X$, we can use dimensionality reduction techniques to reduce the number of columns into a new feature matrix $\hat{X} \in \mathbb{R}^{n \times m}$, $m \ll k$. The dimensionality reduced feature vectors can then be clustered based on similarity, which results in a segmented image.

3.4 Example on Marmousi II synthetic dataset

The elastic Marmousi II model (Figure 3.10) is an elastic extension to the ubiquitous Marmousi model and was developed for the specific purpose of testing amplitude vs. angle processing and analysis methodologies. I use a subset of the model that contains a gas reservoir embedded in layers of brine-saturated sand and shales in order to develop and demonstrate the following unsupervised learning methods:

1. Clustering of principle component projections extracted from conventional principle component analysis (PCA)
2. Clustering of principle component projections extracted from nonlinear kernel principle component analysis (kPCA)
### Figure 3.9
AVA characterization as unsupervised learning. ADCIGs are reshaped into a feature matrix, where each point becomes a row and the angle dependent reflectivity response becomes a column. The columns of the matrix are reduced to allow for visualization and to prevent overfitting. The low dimensional features can be clustered which results in a segmented image.

3. Direct image segmentation by matrix splitting using an extension of robust principle component analysis (rPCA).

#### 3.4.1 Seismic modelling
The Marmousi II Earth model provides physical rock properties $V_p$, $V_s$, and $\rho$. We wish to study angle dependent reflectivities and therefore need to model ADCIGs (Section 3.1). As a ground truth dataset, I use the Zoeppritz equations (Equation 3.5) to synthesize ADCIGs (Figure 3.11) directly.
from physical rock properties of the model. This dataset is ideal in that the angle dependent reflectivity responses are physically consistent with the model described in Section 3.2 and is what we would measure if we had perfect seismic acquisition and processing. Shuey terms were extracted by inverting the linear system described in Equation 3.9 and are crossplotted in Figure 3.12. The farthest outliers in the crossplot correspond to the top and bottom of the hydrocarbon reservoir.

I synthesize a more realistic dataset by first modelling a seismic acquisition via a finite-difference visco-acoustic wave equation simulation. An absorbing boundary at the water bottom was used to prevent strong surface-related multiples. ADCIGs (Figure 3.13) were then formed using the shot profile migration scheme described in Section 3.1. Note that the angle gathers are not perfectly flat, and the wavelet spreads out at greater angles. Flatness of gathers were corrected by peak picking in order to prevent phase errors dominating the angle-dependent response. The spreading of the wavelet is an artefact of the migration operator, which was not corrected for in order to keep some physical inconsistencies in this synthetic dataset. No additional processing was performed, so the data still contains
internal multiples, some diffraction artefacts, non-stationary wavelet effects, and non-uniform illumination.

Extracted Shuey terms are plotted in Figure 3.14. Visco-acoustic modelling is computationally cheaper than using a full elastic scheme, but results in subtler responses. Referring to Equations 3.6-3.8, the response will be primarily characterized by the first term (Equation 3.7). Although there is a background trend and outliers which can be related to the underlying geology, there is significant spreading due to the unmodeled contributions. I explore data projections that are robust to these unmodeled contributions.

3.4.2 Principle component analysis

PCA can be used to find data projections that will maximize the variance, which is a measure of information. Performing an eigendecomposition of the co-variance matrix
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.12 IG crossplot for the true Marmousi II reflectivity model.

\[ C = X^T X = \frac{1}{n} \sum_{i} x_i x_i^T \]  

(3.15)

where \( C \in \mathbb{R}^{d \times d} \) results in a set of eigenvectors \( v_i \) with corresponding eigenvalues \( \lambda_i \), where \( i = 1 \ldots d \). Projecting the feature matrix \( X \in \mathbb{R}^{n \times d} \) onto the eigenvectors with the \( m \) largest eigenvalues results in the dimensionally reduced matrix \( \hat{X} \in \mathbb{R}^{n \times m} \).

The use of PCA as a seismic analysis tool has been studied by Scheevel et al. (2001), where they use principle components along the time axis as features while predicting porosity from seismic amplitude in a supervised context. Hagen (1982) as well as Kaplan (2003) have used PCA as a means of denoising, which again analyzes principal components along the time axis. Alternatively, Saleh et al. (2000) has recognized the connection between PCA of scattering responses and Shuey term inversion. The author showed that crossplots of principle component coefficients and Shuey terms share similar multivariate properties in a noise free example, but the principle component projection was more robust to multiplicative noise. This is because multiplicative noise is not in the span of the Shuey vectors, where as
the data driven principle component approach will always find projections that span the data.

I now apply the same approach as Shuey term analysis to the synthetic datasets, but instead project the onto the two largest principle components instead of the Shuey vectors. For the physically consistent dataset modelled directly by the Zoeppritz equations, the crossplot geometries of PCA and Shuey methods are nearly identical (Figure 3.12 and Figure 3.15). The principle component vectors extracted are normalized versions of the Shuey vectors, which is shown in Figure 3.16. This is an interesting result, as it demonstrates that under the conditions where the physics is preserved, PCA and Shuey inversion provide interpretably equivalent crossplots.

The crossplot for the migrated physically inconsistent data showed different geometries for PCA and Shuey terms (Figure 3.12 and Figure 3.15). Most importantly, outlying clusters are spread further from the background trend in the PCA crossplot. This is an important result, as anomalous responses in this dataset correspond to hydrocarbon-saturated sediments.
Figure 3.14 IG crossplot for the migrated Marmousi II model.

Figure 3.15 PCA crossplot for the true Marmousi II reflectivity model.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.16 Comparison of principle components from Zoeppritz modelled data, migrated data, and the Shuey terms as components.

The two principle component vectors with the largest eigenvalues are compared to the vectors defined by the Shuey model(Figure 3.16). Since the principle components were extracted directly from data, unmodeled contributions such as wavelet effects, internal multiples, and the response of the migration are in the span of the projection. On the contrary, the Shuey model is static and assumes the data is explained entirely by the physics, resulting in less variance in the crossplot, which makes discriminating anomalous responses more difficult.

Although crossplots can be manually analyzed to look for background trends and anomalous clusters, we seek a fully automated approach to image segmentation. We therefore need a method to cluster the data into a finite number of classes.
Figure 3.17 PCA crossplot (a) compared to IG crossplot (b) for the migrated Marmousi II model. Note that outliers in (a) are pushed farther from the dense background clusters than in (b).
3.4.3 Clustering

Data clustering methods are typically either partitional, which classifies points by assignment to an underlying distribution, or hierarchical which clusters points based on a distance metric.

Partitional algorithms directly decompose data into disjoint clusters by minimizing a criterion function, which is usually a set of probability distribution functions. For example, Gaussian mixture models partition a dataset by maximizing probability of each point being associated with one of $K$ distributions defined by Gaussian statistics. The ubiquitous K-means algorithm is a scalable but specific case of a mixture model where each class shares identical statistics. A partitional algorithm will be sensitive to the choice of number of clusters and the initial starting guess of the distribution parameters. Partitional clustering on a toy example is shown in Figure 3.18.

![Partitional clustering](image)

**Figure 3.18 Partitional clustering.** Points are partitioned by maximizing the likelihood of membership to a distribution.

Hierarchical algorithms cluster by iteratively joining and splitting groups of points into a hierarchy based on a distance metric (Figure 3.19). Since there is no reason to assume that dimensionality reduced seismic data will be drawn from any known distribution, I use a hierarchical clustering algorithm for classification. BIRCH clustering (Zhang et al., 1996) is designed for large databases and is specifically optimized to minimize the number of passes through the dataset.

At the core of Birch clustering is the cluster feature tree. Given a cluster of $N$ samples $\{x_i\}$ where $i = 1, 2, ..., N$ and $x_i \in \mathbb{R}^m$, a cluster feature is defined as the three point vector $(N, \sum_{i=1}^{N} x_i, \sum_{i=1}^{N} x_i^2)$. A tree can be formed
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.19 Agglomerative clustering. Each point is initially considered to be a cluster. The geometrically closest clusters are joined into a node, which is then considered a new cluster. The algorithm steps through the hierarchy until a set number of clusters is reached.

by joining subclusters into nodes, where the merged cluster feature vector is a simple addition of the subcluster feature vectors. Note that cluster centroid $\bar{x} = \sum_{i=1}^{N} x_i$ and cluster radius $r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$ can be calculated from the cluster feature values.

The tree geometry is controlled by two parameters, the threshold $T$ that defines the maximum radius of any cluster, and the branching factor $B$ that defines the maximum number of samples allowed in each cluster. A cluster is split when either of these values is exceeded, and the subclusters are either left as their own leaf, or merged with other clusters. The final leaves of the tree are passed to a simple agglomerative clustering algorithm (Figure 3.19) which outputs the final classifications.

Note that the formation of the cluster feature tree requires one pass through memory and can be applied in an online fashion, which makes the algorithm highly scalable.

When applied to the principle component projections of the physically consistent data, Birch resulted in 6 distinct clusters, 2 of which correspond to the top and bottom of the hydrocarbon reservoir (Figure 3.20), one corresponds to the background trend, and 3 other clusters are related to the high amplitude reflections near the bottom of the model. Clustering principle component projections of the migrated seismic resulted in 4 distinct classes.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

(two were merged after the fact to define the background), where the top and bottom of reservoir was again delineated from the background trend (Figure 3.21).

Birch clustering provides a scalable method to automatically segment a seismic image based on principle component coefficients. The method successfully delineated a hydrocarbon reserve from a background trend in synthetic datasets. The number of clusters and branching threshold (Table 3.3) was chosen by trial to demonstrate desirable results, which I acknowledge is not a robust general solution to parameter selection. I now explore other projections with the goal of discovering multivariate geometries (locations of points on a crossplot) that are more robust to automated clustering.

Figure 3.20 Classification results for physically consistent data.

3.4.4 Kernel PCA

PCA will reduce the number of features while maximizing the variance of the data (a measure of information), but it is a linear model and may not result in the best low-dimensional representation of $X$. Instead, I seek a non-linear projection that results in advantageous multivariate geometries. Non-linear
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.21 Classification of principle components from migrated seismic data. The background trend was initially two positive/negative clusters that were joined into one for plotting purposes.

principle component projections can be calculated by employing the “kernel trick”, which implicitly transforms the data into a high-dimensional non-linear feature space (Schölkopf et al., 1997). Kernels are often employed in machine learning problems to exaggerate or linearize seperation boundaries in classification problem (Hofmann et al., 2008).

In order to apply a kernel to PCA, we need to recognize that PCA can be calculated from the inner-product matrix

\[ XX^T = \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \] (3.16)

instead of the outer product matrix \( X^T X \) (covariance). Although \( XX^T \in \mathbb{R}^{n \times n} \) compared to \( X^T X \in \mathbb{R}^{d \times d} \) and \( d \ll n \) for seismic data, \( XX^T \) is a Gramian matrix that is required by kernel methods. Now let \( U \) be the...
orthogonal matrix containing the eigenvalues of $XX^T$, $\Lambda$ be the corresponding matrix of eigenvalues, and $V$ be the orthogonal matrix containing the eigenvectors of $X^T X$. By definition, we have

$$(XX^T)U = U\Lambda \tag{3.17}$$

that we multiply both sides by $X^T$ to arrive at

$$(X^T X)(X^T)U = (X^T U)\Lambda. \tag{3.18}$$

Note this is just an eigenvalue decomposition, which shows the eigenvectors of the covariance matrix $X^T X$ can be calculated from $X^T U$. This is an important result as it shows that the inner product matrix $XX^T$ can be used to find the principle components of $X$.

Since the entries of $XX^T$ depend only on the inner product of feature vectors $\langle x_i, x_j \rangle$, we can employ a kernel $\kappa(x_i, x_j)$ that implicitly calculates the inner product $\langle \phi(x_i), \phi(x_j) \rangle$ where $\phi(x)$ defines a high-dimensional feature space. Note that by using a kernel we calculate the high-dimensional inner product in the original lower-dimensional space (Hofmann et al., 2008). Using a non-linear kernel will result in a non-linear PCA operation.

Now consider the polynomial kernel

$$\kappa(x_i, x_j) = (x^T_i x_j + b)^c. \tag{3.19}$$

It can be shown that $\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ where $\phi(x)$ is a feature transformation that contains all terms up to degree $c$. For example, the polynomial kernel with $c = 2$ and $b = 1$ corresponds to the feature transformation

$$\phi(x) = [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2] \tag{3.20}$$

when $x \in \mathbb{R}^2$ (Murphy, 2012).

Crossplots of projected components for increasing values of $c$ are shown in 3.22. The geometry changes dramatically for higher values, and outliers become linearly separable from a densely clustered background trend. This is likely partially attributed to the power-law inherent in the polynomial kernel, which will put increasing emphasis on high-amplitude features. For larger values of $c$ floating point errors become problematic, so I choose a value of 10 which pushes the background trend into a dense cluster without incurring floating-point errors.
Figure 3.22 Kernel PCA crossplots for increasing values of $c$ for physically consistent data (top), and migrated seismic data (bottom).
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

I present the results of kernel PCA as an empirical experiment, and currently do not have a rigorous explanation as to why the data forms along a seemingly smooth curve. The feature space spanned by the implicit transform $\phi(x)$ is both very large and highly non-linear, which makes the analysis of isocurves in this space beyond the scope of this thesis.

Birch clustering was applied to the projected data yielding the segmented images shown in Figure 3.23 and Figure 3.24. The reservoir is clearly delineated from the background in both of the datasets. The background formed one cluster at the apex of the curve, which allows outlying samples to be linearly separated. The farthest outlying cluster (blue) in Figure 3.24 corresponds to the high-amplitude tuning effects at the edges of the reservoir, which was not discovered as a unique cluster using linear PCA projections.

Kernel PCA becomes a computationally expensive operation for seismic data ($n \gg k$), as forming the inner product matrix $XX^T$ is $O(n^3)$ compared to $O(k^2n)$ for the outer product matrix. This limits the size of the window that the operation can be performed on, but does not inhibit the use of the method.

In summary, kernel PCA resulted in preferential multivariate geometries compared to linear PCA at the expense of computational cost. Interpreting the physical meaning of the non-linear projection is challenging and is not currently understood, which limits the appeal of the method. Although kernel PCA projections were less sensitive to clustering parameters than linear PCA, choices of the number of clusters and branching threshold (Table 3.4) are still determined by trial and will likely vary between datasets.

3.4.5 Robust PCA

Clustering of both linear and kernel principle component projections was successful at delineating the reservoir, but both rely heavily on tuning of clustering parameters. Additionally, kernel PCA significantly increases the computational cost and obscures physical interpretation. For these reasons, I develop a scalable approach that can segment an image without explicitly clustering, yet still has a direct physical interpretation.

I again state the assumption that the majority of reflections in the sedimentary layers are from brine-saturated sandstones and shales, and reflections from hydrocarbon-saturated sands are statistically rare. Additionally, I impose the assumption that the angle dependent reflectivity responses of
the background trend are similar and ideally linearly related. With these assumptions, we can state that our feature matrix $X$ is the summation of a low-rank matrix $L$ consisting of the background trend, and a sparse matrix $S$ containing the outlying responses. In this regard, we can directly segment the image into a background trend and outliers by solving the optimization problem

$$
\min_{L,S} \text{rank}(L) + \lambda \|S\|_{\text{row}-0} \quad \text{s.t.} \quad L + S = X
$$

(3.21)

where $X$ is the feature matrix, $L$ is a low-rank matrix, $S$ is a row-sparse matrix, and $\|S\|_{\text{row}-0}$ denotes the number of non-zero rows in $S$. Note that both terms are non-convex, which makes optimization a hard problem.

The $\text{rank}(L)$ can be relaxed into a convex formulation using the nuclear norm $\|L\|_*$ where

$$
\|L\|_* = \text{trace}(\sqrt{L^*L}) = \sum_i \sigma_i = \|\sigma\|_1
$$

(3.22)

Figure 3.23 Image segmentation results for kernel PCA applied to physically consistent data.
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.24 Image segmentation results for kernel PCA applied to the migrated seismic data.

and \( \sigma \) is a vector of the magnitude of the singular values of \( X \). Donoho (2004) showed that minimizing the \( \ell_1 \)-norm promotes a sparse solution, which means minimizing the nuclear norm results in a sparse set of singular values. The number of non-zero singular values corresponds to the rank of a matrix, and therefore minimizing \( \| L \|_* \) will result in a low-rank matrix.

The \( \| S \|_{row-0} \) term can be relaxed into the convex matrix norm \( \| S \|_{1,\infty} \), which is the \( \ell_1 \)-norm along rows and the \( \ell_\infty \)-norm along columns. Encouraging the vector of extrema along each row to be sparse will result in row-sparsity. Other row-sparsity promoting norms like \( \ell_{12} \) can also be used. We can now reformulate Equation 3.21 as the convex problem

\[
\min_{L,S} \| L \|_* + \lambda \| S \|_{1,\infty} \quad \text{s.t.} \quad L + S = X, \tag{3.23}
\]

which extends the robust PCA work of Candes et al. (2009) from sparse matrices to row-sparse matrices. I use the alternating direction method of multipliers (ADMM) to find a solution to Equation 3.23. ADMM methods
have well-studied convergence properties and have been applied to a wide-
range of convex optimization problems. Perhaps most relevant to seismic
data problems, it is adaptable to large distributed datasets (Boyd, 2010).

Using an augmented Lagrangian we can write Equation 3.23 as the un-
constrained loss function

\[ l(L, S, Y) = \| L \|_* + \lambda \| S \|_{1,\infty} + \langle Y, X - L - S \rangle + \frac{\mu}{2} \| X - L - S \|_F^2 \] (3.24)

where \( \mu \) can be interpreted as penalty term for the data misfit. ADMM
proceeds by iteratively finding \( L_k, S_k \) via

\[ \arg\min_{L, S} l(L, S, Y_k) \]

then updating the Lagrange multiplier

\[ Y_{k+1} = Y_k + \mu (X - L_k - S_k) \] (3.25)

until a convergence criterion is met. Following the approach used by Candes
et al. (2009), I recognize that \( \arg\min_S l(L, S, Y) \) and \( \arg\min_S l(L, S, Y) \) have
simple and efficient solutions using proximal operators. I refer to Combettes
and Pesquet (2011) for a background on proximal operators and iterative
proximal splitting methods.

Let \( S_\tau : \mathbb{R}^m \rightarrow \mathbb{R}^m \) denote the vector shrinkage operator

\[ S_\tau[x] = \max(1 - \frac{\tau}{\|x\|_\infty}, 0) x, \] (3.26)

\( T_\tau : \mathbb{R} \rightarrow \mathbb{R} \) denote the scalar shrinkage operator

\[ T_\tau[x] = \text{sgn}(x) \max(\tau - |x|, 0), \] (3.27)

and \( D_\tau \) denote the singular value thresholding operator given by

\[ D_\tau(X) = U T_\tau(\Sigma)V^* \] (3.28)

where \( X = U \Sigma V^* \) is any exact or approximate singular value decomposition.
It can be shown (Candes et al., 2009; Combettes and Pesquet, 2011) that

\[ S_k = S_\mu \tau(X - L + \mu^{-1} Y) \] (3.29)

and

\[ L_k = D_\mu \tau(X - L - \mu^{-1} Y) \] (3.30)
are proximal operators for $\arg\min_S l(L, S, Y)$ and $\arg\min_L l(L, S, Y)$ respectively. The optimization proceeds by fixing $S$ while minimizing with respect to $L$, then fixing $L$ and minimizing with respect to $S$, and finally updating the Lagrange multiplier based on the residual $X - L - S$. See algorithm 3.2.

**Algorithm 3.2 Component Pursuit by ADMM**

1. initialize $S_0 = Y_0 = 0, \mu > 0$.
2. **while** not converged **do**
   3. compute $L_{k+1} = D_\mu(M - S_k - \mu^{-1}Y_k)$;
   4. compute $S_{k+1} = S_{\lambda\mu}(M - L_{k+1} + \mu^{-1}Y_k)$;
   5. compute $Y_{k+1} = Y_k + \mu(M - L_{k+1} - S_{k+1})$;
3. **end while**
4. **output**: $L, S$.

In practice the algorithm is not run to completion as a seismic dataset can be very large, so both the stopping criteria and choice of $\mu$ play a vital role in the resulting estimates of $L$ and $S$. The values listed in Table 3.5 were determined through trial.

Robust PCA was able to exactly segment the reservoir for the physically consistent dataset (Figure 3.25). The reservoir was mostly delineated in the migrated seismic (Figure 3.26), but there is not perfect lateral continuity. Convergence of the algorithm has been proven for the case when the sparse component $S$ is random noise (high rank), but I can not assume that the outlying reflectivity responses will also be high rank. The lack of complete segmentation is therefore not surprising, but the amount delineation shown in Figure 3.26 still provides a very useful result. The fact that Robust PCA has a direct physical interpretation, does not require a priori knowledge of the number of clusters, and is scalable to large seismic datasets makes it an attractive method for seismic image segmentation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Zoeppritz data</th>
<th>Migrated seismic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of components</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 3.3** Settings for clustering PCA reduced features
Chapter 3. Reflection seismology analysis as an unsupervised learning problem

Figure 3.25 Robust PCA applied the true Marmousi II reflectivity model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Zoeppritz data</th>
<th>Migrated seismic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of components</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( c )</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>( b )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 3.4** Settings for kernel PCA

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Zoeppritz data</th>
<th>Migrated seismic data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>3000.0</td>
<td>3000.0</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>number of iterations</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

**Table 3.5** Settings for robust PCA
3.4.6 Discussion

The experiments performed in this chapter were motivated by the inability of the conventional geophysical model to adequately explain the angle dependent reflectivity response of seismic data. Seismic data acquisition is performed in an uncontrolled natural environment, and processing of the data requires many physical assumptions that prevent calibrated measurements. Seismic data can be migrated into structural images where scattering interfaces can be accurately located, but the angle dependent response will contain a combination of the true physical response as well as unmodeled contributions. For this reason, I explored unsupervised learning models that learn projections directly from the data that help delineate outlying responses from a background trend. This is important, as outlying reflectivity responses can be associated with potential hydrocarbon reserves in the sedimentary layers.

Generalizing the physical model described in Section 3.2 as a simple projection motivated the use of PCA-based approaches. Figure 3.12 and
Figure 3.15 demonstrate that in the unlikely case of a physically consistent dataset, the projections learned from PCA and the direct projections from the physical model are nearly equivalent. When applied to a more realistic dataset that contains unmodeled contributions from acquisition and processing, PCA projections provided advantageous crossplots where outliers (hydrocarbon related regions) could be more easily delineated (Figure 3.17). In this regard, projecting onto principle components should be a preferred method to using the Shuey model, as it provides similar results in an ideal scenario and is more robust to physically inconsistent data. For this simple synthetic test case, the physical inconsistencies are minor, and the principle components therefore have similar shapes (Figure 3.16). For real seismic data (Figure 3.8), the principle component vectors vary substantially from Shuey vectors indicating that the difference between PCA and Shuey term approaches will be much more pronounced.

Using a polynomial kernel provided a method to project the data onto principle components in an implicit high-dimensional, non-linear feature representation. The resulting multivariate geometry of the non-linear projection allowed outlying responses to be linearly separated from a background. Interestingly, the crossplots shown in Figure 3.22 form a distinct curve where the background densely clusters at the apex. Although this projection has properties that are useful for automatic classification, it is difficult to intuitively interpret. Further research is required to better understand the multivariate geometry created by the polynomial kernel.

Birch clustering was used to automatically classify the projected data. Results were promising, as the reservoir was segmented using kernel PCA and PCA for both physically consistent and migrated data (Figure 3.21, Figure 3.23, Figure 3.24). Automated clustering requires the desired number of clusters as a user defined input, which in practice may be an unrealistic requirement. Additionally, real world seismic data has a continuum of responses, which may not be able to be classified into a finite number of disjoint classes. Although Birch clustering was useful for segmenting this synthetic data, a less automated approach to classifying projections will likely be required for real datasets.

Alternatively, robust PCA directly segmented the reservoir from the background without the need of a clustering algorithm, which is shown in Figure 3.23 and Figure 3.24. The efficacy of the algorithm is contingent upon desirable rank properties of both the background trend and sparse outliers. The matrix of the background of angle dependent responses is certainly low
rank, but assuming that outlying responses will be significantly higher rank may not be justified. A study using real seismic data collected over a known reservoir is required to analyze the rank of hydrocarbon related reflectivity responses to gauge if use of this method can be used in practice.

Finally, these experiments were performed on simple synthetic models of the Earth. Although interesting for the development of new methodologies, the work presented is by no means a proper benchmark comparison between approaches.

3.5 Conclusions

In this chapter I used a relatively simple model to develop and demonstrate unsupervised learning methods to classifying seismic data. The novelty of this work is reformulating the conventional analysis methodology without the need of a physical model. Results on ideal data showed that projections learned directly from the data yield interpretably equivalent results to using a physical model. Results using migrated data that contains contributions outside of the span of the physical model showed that unsupervised models could provide better projections for classifying outliers from a background trend.

Future work requires testing the methodologies on seismic data acquired over a known and proven hydrocarbon reserve, which would allow proper benchmarks and firm conclusions on the effectiveness of the new methodologies.
Chapter 4

Comparison to dynamic intercept gradient inversion

Dynamic intercept gradient inversion (DIGI) is a Shuey term based method for seismic data analysis developed by Edgar and Selvage (2013). It is currently used by the BG group for finding potential hydrocarbon reserves from industrial seismic data. A migrated seismic dataset was provided to compare PCA-based analysis methods with the DIGI algorithm. A seismic section of interest (Figure 4.1) containing a potential gas reservoir was selected by a BG geologist and will serve as the segmentation target in the comparison.

The clustering based methods could only segment the water bottom/shallow segments, so I restrict the comparison to robust PCA and a modified version PCA that extends the DIGI algorithm.

4.1 Dynamic intercept gradient inversion

The Shuey model described in Section 3.2 is extended to include the bandwidth-limiting effects of the seismic wavelet, where the migrated seismic data \( d(x, t, \theta) \) is explained by the convolution model

\[
    d(x, t, \theta) = w(x, t, \theta) * r(x, t, \theta),
\]

where \( r(x, t, \theta) \) is the reflectivity response of the Earth and \( w(x, t, \theta) \) is a set of non-stationary wavelets estimated from the data. The null-space caused
by the bandwidth-limited wavelet creates the ill-posed system of equations

\[
\begin{bmatrix}
\mathbf{d} \\
\end{bmatrix} = \begin{bmatrix}
\mathbf{W} & \mathbf{W} \sin^2 \theta
\end{bmatrix} \begin{bmatrix}
i \\
\mathbf{g}
\end{bmatrix}
\]  

(4.2)

where \( \mathbf{W} \) is a Toeplitz matrix containing the estimated set of non-stationary wavelets \( w(x,t,\theta) \) and \( \mathbf{d} \) is again the migrated ADCIGs flattened into a vector. The ill-conditioned system is regularized by forcing the gradient of \( \mathbf{i}, \mathbf{g} \) to be small by the augmented system

\[
\begin{bmatrix}
\mathbf{d} \\
0
\end{bmatrix} = \begin{bmatrix}
\mathbf{W} & \mathbf{W} \sin^2 \theta \\
\lambda \nabla & \lambda \nabla
\end{bmatrix} \begin{bmatrix}
i \\
\mathbf{g}
\end{bmatrix}.
\]  

(4.3)

The system is further augmented by the minimum energy extended elastic reflectivity (J. Hicks and M. Francis, 2006)

\[
EER(\chi_{me}) = I \cos(\chi_{me}) + G \sin(\chi_{me})
\]  

(4.4)
which exploits the expected correlation between $i$ and $g$ where $EER(\chi_{me}) = 0$ for perfectly correlated terms. The value of $\chi_{me}$ is expected to vary with the shear velocity ratio $\frac{V_P}{V_S}$. Assuming this variation is dominated by compaction, $\chi_{me}$ should decrease with time below the seabed, $T_{seabed}$. The simple linear relationship

$$\chi_{me}(x, t - T_{seabed}) = m(x) \cdot (t - T_{seabed}) + c(x) \quad (4.5)$$

is used to model the compaction trend.

The final system of equations

$$\begin{bmatrix} d \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} W & W \sin^2 \theta \\ \lambda \nabla & \lambda \nabla \\ W(\theta_{me}) \cos(\chi_{me}) & W(\theta_{me}) \sin(\chi_{me}) \end{bmatrix} \begin{bmatrix} i \\ g \end{bmatrix} \quad (4.6)$$

where $\theta_{me}$ is obtained from $\sin^2 \theta_{me} = \tan(\chi_{me})$ is solved using the conjugate gradient algorithm LSQR (Paige and Saunders, 1982).

The $EER$ values calculated from $i$ and $g$ form an image where large values indicate deviations from the expected correlation. In this respect, anomalous regions can be extracted from the $EER$ image by thresholding.

### 4.2 PCA extended DIGI

In Section 3.4.2 I showed that projecting ADCIGs onto the two largest principle components had advantageous multivariate properties compared to the Shuey model. Using this knowledge, I modify the DIGI algorithm by directly replacing the Shuey terms with the two largest principle components extracted from the data – i.e.,

$$\begin{bmatrix} d \\ 0 \end{bmatrix} = \begin{bmatrix} Wc_1 & Wc_2 \\ \lambda \nabla & \lambda \nabla \\ W(\theta_{me}) \cos(\chi_{me}) & W(\theta_{me}) \sin(\chi_{me}) \end{bmatrix} \begin{bmatrix} i \\ g \end{bmatrix} \quad (4.7)$$

where $c_1$ and $c_2$ are the two largest principle components vectors. Using data driven components allows the algorithm to explain data that is outside the span of the Shuey vectors. This method is feasible only for linear PCA, as the abstraction of the kernel PCA to a higher dimension does not fit within this framework.
4.3 Results

Segmented images for robust PCA, DIGI, and PCA extended DIGI are shown in Figures 4.2-4.4. Parameters were selected to segment the potential reservoir while maintaining the least amount of misclassified clutter. Although none of the methods was able to achieve perfect segmentation, both PCA methods had significantly less clutter than the DIGI results.

The robust PCA method (Figure 4.2) primarily segmented the water bottom as well as the potential reservoir. This is a good result, as both these regions should deviate significantly from any background trend of reflections. There are some spurious classifications at the depth of the reservoir, but these misclassifications do not form coherent shapes and would likely be ignored by an experienced interpreter. The robust PCA had the least amount of classification clutter for the same amount of reservoir segmentation making it the preferred method for this test case.

The segmented image from the DIGI algorithm is shown in Figure 4.3. The image shows a significant amount of spurious classifications in addition to the potential reservoir, which would make interpreting the potential reservoir a challenge. Thresholding the DIGI results is not useful, but manually examining the EER image (Figure 4.5) for bright regions could still be effective.

The PCA extended DIGI algorithm produced a segmented image (Figure 4.4) that had significantly less clutter than the original DIGI algorithm. The interesting part of this image is that the water bottom and shallow layers were not considered as outliers, which one might expect. However, these are the highest amplitude regions, and since PCA tries to explain the most variance in the data, the extracted principle components would be biased to explain this region.

EER cross-sections created from Shuey terms and principle component vectors are shown in Figure 4.5. The image generated from the principle components is sharper, and the potential reservoir stands out higher against the background (see the green section of the trace in Figure 4.5). Principle component vectors extracted from the data are compared to the Shuey vectors in Figure 4.6. There are substantial differences between these plots, which indicates there is significant contributions in the data that out of the span of the Shuey vectors. For this reason, the PCA extended DIGI provided preferential results.
4.4 Discussion

The PCA-based methods outperformed DIGI on this particular seismic section because the data contained contributions that were outside span of the Shuey vectors. Both robust PCA and the PCA extended DIGI algorithm do not rely on a static physical model and are thus able to adapt to the dataset.

The PCA-extended DIGI method extracted principle components from the entire seismic section of data. Future research should deconvolve the data first before extracting principle components, as the convolution is part of the forward model (Equation 4.7). Additionally, studies to see how the principle components vary with spatial coordinates would be useful, as extracting principle components from windowed sections could improve results if the principle components vary spatially. Finally, the $EER$ calculation (Equation 4.4) may not be the best for determining outliers from correlations between principle component projections. Future work could assess the efficacy of thresholding on other correlation functions.

Figure 4.2 Segmentation results from robust PCA.
Chapter 4. Comparison to dynamic intercept gradient inversion

Figure 4.3 Segmentation results from thresholding $EER$ values extracted using the DIGI algorithm.

I will reserve any firm conclusions, as this was one study on a small seismic section with a very loose geological interpretation. Seismic field datasets containing proven reservoirs are required to continue research into how to find hydrocarbons directly from seismic data. Further research into feature representations and data projections will continue to be inconclusive without ground truth datasets to serve as benchmarks.
Figure 4.4 Segmentation results from thresholding EER values extracted using the DIGI algorithm with principle components replacing the Shuey terms.
Figure 4.5 Comparison of EER sections generated from DIGI (top) and PCA extended DIGI (bottom). The region containing the potential reservoir is highlighted in green on the trace plots.
Figure 4.6 Principle component vectors extracted from the BG data set (top) compared the Shuey components (bottom).
Chapter 5

Closing remarks

I this thesis I explored machine learning applications to geophysical data analysis. I made the thematic argument that geophysical data falls within the regime of machine learning models. Firstly, interpretation of geophysical data is often based on visual correlations and pattern recognition, which fits into a supervised learning framework. Geological interpretations can serve as labels which can train a classifier to make predictions on similar datasets. Secondly, I argue that the complexity of the Earth’s sedimentary layers combined with the difficulty of seismic acquisition and processing limits the use of physical models to fully explain seismic data. Using assumptions of low-dimensionality and latent models I showed that unsupervised learning can be applied to extract more useful information directly from the data.

In Chapter 2 I applied supervised learning to train a classifier to predict stratigraphic units from well logs. I learned a generic mapping from multifractal feature vectors to interpreted stratigraphy that can make useful predictions of stratigraphic units. The novel contribution was using the scattering transform as a multifractal feature representation for well logs.

In Chapter 3 I took a well-known exploration geophysics analysis methodology and presented it in as an unsupervised learning problem. I made the argument that seismic data contains contributions outside the span of the conventional two-component linear model. I showed that under ideal conditions PCA-based approaches provide the same results, and are more robust to unmodeled contributions from acquisition and migration. The main contribution of this chapter was abstracting the analysis procedure into the vernacular of unsupervised learning which is independent of a physical model.
In Chapter 4 I applied the methods developed in Chapter 3 to a field dataset and compared the results to a method currently used by the BG group for seismic data exploration. In this small example on a field data set I showed that unsupervised learning approaches could better segment a potential gas reservoir than BG’s current methodology. The novel contribution of this section is demonstrating the methodologies developed in this thesis can be effective on a real dataset.

In this thesis I developed original ideas and methodologies to use machine learning in exploration geophysics. Although several novel approaches were developed, conclusions were difficult to draw. In each experiment I lacked adequate data to do scientific benchmarks. A lot of the advances in machine learning in other fields can be owed to open datasets which allow researchers to benchmark and publish their findings on the same data. Without this same infrastructure, there are significant barriers to demonstrating the potential success of machine learning in exploration geophysics.
Bibliography


Bacry, E., J. F. Muzy, and A. Arneodo, 1993, Singularity spectrum of fractal signals from wavelet analysis: Exact results: Journal of Statistical Physics, 70, 635–674. → pages 10


Bond, C. E., 2015, Uncertainty in structural interpretation: Lessons to be learnt: Journal of Structural Geology, 74, 185–200. → pages 18


Bibliography


Edgar, J., and J. Selvage, 2013, Dynamic Intercept-gradient Inversion: Presented at the . → pages 31, 57


Frisch, U., and G. Parisi, 1985, Fully developed turbulence and intermittency: Turbulence and predictability in geophysical fluid dynamics and climate dynamics, 88, 71–88. → pages 8, 10


Hagen, D. C., 1982, The application of principal components analysis to seismic data sets: Geoexploration, 20, 93–111. → pages 37

Hall, M., B. Bougher, and E. Bianco, 2015, Pick This! Social image interpretation: , Society of Exploration Geophysicists, 1772–1775. → pages 18

Herrmann, F. J., 1997, A scaling medium representation, a discussion on well-logs, fractals and waves: [s.n]. → pages vii, viii, 7, 9, 12


J. Hicks, G., and A. M. Francis, 2006, Extended Elastic Impedance and Its Relation to AVO Crossplotting and Vp/Vs: Presented at the . → pages 58


LeCun, Y., K. Kavukcuoglu, C. Farabet, and others, 2010, Convolutional networks and applications in vision.: ISCAS, 253–256. → pages 15


Paige, C. C., and M. A. Saunders, 1982, LSQR: An algorithm for sparse linear equations and sparse least squares: ACM transactions on mathematical software, 8, 43–71. → pages 59

Pentland, A. P., 1984, Fractal-Based Description of Natural Scenes: IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 661–674. → pages 8


