Dictionary Learning and Sparse Representations for Denoising and Reconstruction of Marine Seismic Data

Pierre Turquais

February 12, 2018

Thesis submitted for the degree of Philosophiae Doctor

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1954

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Preface

This thesis was written to be submitted to the Faculty of Mathematics and Natural Sciences at the Universitetet i Oslo (UiO), in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD). The PhD project was carried out as a collaboration between the Marine Geophysics Department of Petroleum Geo-Services (PGS) and the Department of Geosciences of UiO. This collaboration was under the form of the Industrial PhD Program organized and supported by the Research Council of Norway. The research described herein was conducted under the supervision of Dr. Endrias Asgedom (PGS), Dr. Walter Söllner (PGS), Prof. Leiv-Jacob Gelius (UiO), and Prof. Valérie Maupin (UiO), between March 2015 and December 2017.

The research in my PhD studies started with an investigation of the sparsity promoting methods for seismic processing applications. During this task, I was rapidly attracted by dictionary learning methods. These methods employ complex optimization schemes to learn information embedded in the data and to derive a mathematical domain that is optimal to concisely express the signal in the data. Dictionary learning methods have recently gained high interest in signal processing as they have been shown to achieve impressive tasks in compression, denoising, data reconstruction, and data analysis. However, I found that these methods, often developed to process natural images, were not necessarily optimal for seismic data applications, and therefore, I endeavored to adapt them for such uses. This thesis summarizes this research.

Acknowledgments

The completion of these PhD studies is the contribution of many people to whom I am very grateful and I am pleased to acknowledge.

I will very naturally start with Dr. Walter Söllner and Dr. Endrias Asgedom, my supervisors at PGS, and thank them for their advices, encouragements, and support throughout those three years. They invested their time in my PhD, always available to discuss issues and way-forwards, both contributing with insightful suggestions. I would like to thank particularly Walter for his guidance regarding the ray and paraxial ray theory, and Endrias for brainstorming around sparsity-related issues and for passing his meticulous methodology on to me.

I am also indebted to Prof. Leiv-Jacob Gelius, my academic supervisor, as well as Prof. Valérie Maupin who kindly took over the role of Leiv during his absence. They both contributed in many ways to this thesis, guiding me through the various administrative procedures and providing me with advices and valuable comments during reviews of my work.

Special thanks go to Einar Otnes who suggested the topic and helped to find the funding of the PhD project. He managed to pass his enthusiasm for the topic on to me when we worked together during the six months preceding the PhD period.

This work would not have been so fruitful without the positive and sharing attitude I experienced at PGS. My colleagues openly shared their experience and knowledge on subjects that matter to my work, and willingly provided me with valuable suggestions and data examples. Hence, I hereby acknowledge all my colleagues from G&E Oslo who attended to and reviewed several of my presentations, my colleague from G&E London, Dr. Paolo Terenghi, who presented me sparsity promotion methods, and my colleagues from Imaging, Julien Oukili, Dr. Bagher Farmani, and Tony Martin, who handed me data sets required to validate various processing methods.

I would like to acknowledge PGS and the Research Council of Norway, which funded my PhD studies under the project 247292.

Those three years would very probably have been a little dull without the friends that surrounded me. These include Julien Decaen, Dr. Anna D'Annunzio, Mickael Bastard, and Dr. Alba Ordoñez. Thank you for your support at work, but also for the numerous drinks and climbing sessions outside work, and for counting me among your friends.

Additionally, I would like to give well-deserved thanks to my parents for their endless patience and support. Last but not least, I am deeply grateful to my girlfriend Aline Deloche, not only for correcting my spelling mistakes (both French and English), but more importantly, for filling my life with smiles, joys, and happiness.

Summary

Seismic signals are generally spread across many data samples of the recorded data. Applying a mathematical transformation to the data can however concentrate them on few samples only of the transform domain. Such representations are called sparse and have gained high interest in seismic processing because they build the necessary requirement to better compress or analyze the signal in the data. This thesis first assesses the effectiveness of building sparse representations for three critical seismic processing tasks, i.e., random noise attenuation, signal separation, and data reconstruction. The presented theory and the numerical experiments reveal that sparse representations can be used to achieve the aforementioned processing tasks under the condition that the signal has a high level of sparsity in the transform domain. It then follows an investigation of the transforms that can lead to sparse representations of the seismic data. A particular focus is placed on dictionary learning (DL) methods. These methods are applied to a data set to find a dictionary that can be used for sparse representation of the data set. The dictionary is a set of signals, called atoms, that represent elementary patterns of the data, and the sparse representation is found by reconstruction of the data with linear combinations of few dictionary atoms. The conventional DL methods are examined, and various modifications are implemented to develop three DL-based methods that are better adapted to each of the seismic processing tasks of interest. (1) A DL method is developed to attenuate the random noise in the seismic data. This method learns a dictionary and finds a sparse approximation of the data based on a statistical measure of the coherence in the residuals. Due to this particularity, the method is released from the need of the a priori knowledge of the noise energy. This is attractive for seismic data applications because the noise in seismic data has an intensity that is often unknown and that is varying across the data set. (2) A DL-based method is developed to separate the coherent noise from the seismic data. Some types of noise that contaminate seismic data cannot be removed with random noise attenuation methods because they appear with spatial or temporal coherency in the data. To tackle such noise, DL is combined with a statistical classification. First, DL is applied to the noise-contaminated data, which results in a dictionary of atoms representing either signal patterns or noise patterns. Using a statistical classification, the noise atoms are separated from the signal atoms, which divides the dictionary into a subdictionary of noise atoms and a subdictionary of signal atoms. Then, by finding a sparse representation of the data in the two subdictionary domains, the signal and the noise contributions in the data are identified and separated. This DL-based method has compelling advantages compared to the traditional coherent noise removal methods based on sparse representation; it does not require someone to search for adequate transforms that may sparsify the signal and the noise, and it adapts to the signal and noise in the data for an optimal separation.

(3) a DL method is developed to interpolate and regularize the seismic data. In this method, each learned atom is constrained to represent an elementary waveform that has a constant amplitude along a parabolic traveltime moveout characterized by kinematic wavefield parameters. Such a parabolic structure is consistent with the physics of the seismic wavefield propagation and it can be used to easily interpolate and extrapolate the atoms. Using this advantage, the method can interpolate and regularize the seismic data. The process consists in learning a parabolic dictionary, interpolating the atoms, and computing a sparse representation of the data in the interpolated dictionary domain. Benefiting from the parabolic structure, the sparsity promotion, and the data adaptation, this method is able to interpolate severely aliased data. The three proposed DL methods are validated with synthetic and field seismic data examples. The effectiveness of the denoising methods are also assessed in comparison to industry-standard and state-of-the-art methods. Each method is demonstrated to be valuable for seismic processing.

List of Publications

This thesis is based on three articles, and is related to three conference papers and two patents.

Articles

In this thesis, the three following articles will be referred to with Roman numerals as given below.

I Turquais, P., E. G. Asgedom, and W. Söllner, 2017c, A method of combining coherence-constrained sparse coding and dictionary learning for denoising: Geophysics, **82**, V137–V148

II Turquais, P., E. G. Asgedom, and W. Söllner, 2017a, Coherent noise suppression by learning and analyzing the morphology of the data: Geophysics, **82**, V397–V411

III Turquais, P., E. G. Asgedom, W. Söllner, and L.-J. Gelius, 2017f, Parabolic dictionary learning for seismic wavefield reconstruction across the streamers: Submitted to Geophysics

Conference papers

i Turquais, P., E. G. Asgedom, W. Söllner, and E. Otnes, 2015, Dictionary learning for signal-to-noise ratio enhancement: SEG Technical Program Expanded Abstracts 2015, 4698–4702

ii Turquais, P., E. G. Asgedom, and W. Söllner, 2016, Sparsity promoting morphological decomposition for coherent noise suppression: Application to streamer vibration related noise: SEG Technical Program Expanded Abstracts 2016, 4639–4643

iii Turquais, P., E. G. Asgedom, and W. Söllner, 2017d, Structured dictionary learning for interpolation of aliased seismic data: SEG Technical Program Expanded Abstracts 2017, 4257–4261

Patents

i Turquais, P., E. G. Asgedom, and W. Söllner, 2017b, Denoising seismic data: U.S. Patent 2017,0108,604 A1

ii Turquais, P., E. G. Asgedom, and W. Söllner, 2017e, Structured dictionary learning for interpolation of aliased seismic data: Invention disclosure PGS-17116US

Contents

Notation and Acronyms

Unless indicated otherwise, the nomenclature and acronyms in chapters 1, 2, 3 and 7 correspond to the ones given below.

Notation

In general, scalars are denoted by lowercase letters or symbols, vectors are denoted by lowercase bold letters or symbols, and matrices are denoted by uppercase bold letters or symbols. Particularly, the notations that are recurrently used are given below.

$\lvert . \rvert$	absolute value
$\lvert\lvert . \rvert\rvert_p$	ℓ_p-norm of a vector
$.^*$	adjoint of a matrix
$.^{\mathrm{T}}$	transpose of a vector or a matrix
ϵ	error threshold tolerated in an error-constrained sparse approximation
Λ	indexes of the support of a sparse representation
μ	mutual coherence
$\mathbf{a}_i$	ith atom of the dictionary
$\mathbf{D}$	dictionary
K	number of atoms in the dictionary
M	number of recording in the training set used to learn a dictionary
N	length of a recording, signal, or atom
$\mathbf{n}$	noise
$\mathbf{S}$	sampling matrix
T	ℓ_0-norm threshold imposed to a cardinality-constrained sparse approximation
$\mathbf{x}$	sparse coefficient vector (most of the cases) or spatial coordinates (section 2.3.1)
$\mathbf{y}$	signal
$\mathbf{z}$	recording

Acronyms

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
BP	basis pursuit
CDL	coherence-constrained dictionary learning
DCT	discrete cosine transform
DDTF	data-driven tight frame
DL	dictionary learning
DWT	discrete wavelet transform
FOCUSS	focal underdetermined system solver
IRLS	iterative-reweighed-least-square
K-SVD	k times singular value decomposition
MCA	morphological component analysis
MOD	method of optimal direction
MP	matching pursuit
NP	non-deterministic polynomial-time
OMP	orthogonal matching pursuit
PDL	parabolic dictionary learning
PGS	Petroleum Geo-Services
PhD	Philosophiae Doctor
S/N	signal-to-noise ratio
SGK	sequential generalization of K-means
SVD	singular value decomposition
UiO	Universitetet i Oslo

Chapter 1

Introduction

1.1 A towed-streamer marine seismic survey

This section starts with a brief description of the objective and the course of action of a typical towed-streamer marine seismic survey, then it points out some issues encountered during the acquisition of the data, and it relates these issues to the need for denoising, interpolation, and regularization of the data during seismic processing.

1.1.1 The objective and course of action of the survey

Marine seismic surveys aim to acquire knowledge about the geology below the sea floor to find and extract valuable mineral resources. For a typical three-dimensional (3D) towed-streamer survey, a seismic vessel tows below the sea one or several seismic sources and several cables called streamers containing sensors. During the survey, the vessel sails above the area of interest and the sources are triggered at desired time intervals. Each source emits a seismic wavefield that travels down through the water and into the subsurface. At each interface between different types of rocks, a portion of the wavefield is reflected toward the sea surface and may be recorded by the sensors. The recorded wavefield is different from the emitted wavefield because it has been modified and altered during its propagation in the different layers. The modifications and alterations are described by known wave equations and they depend on the composition of the rocks, the structure of the layers and formations, and on the presence and type of fluid in the rock. Hence, the recorded wavefield contains information about the geology of the subsurface. In practice, one does not deduce this information by observing the recorded data; the reasons include that the quantity of recorded data is extremely large (e.g., tens of terabytes of data are acquired to survey an area of a few thousand squared kilometers). To enlighten this information, the geophysicists apply a processing sequence to the data, which transforms the data into an interpretable image of the subsurface. This process is called processing, imaging, and inversion. Afterward, geologists can quantitatively interpret the image to deduce the geology of the subsurface and locate or monitor a reservoir of oil or gas.

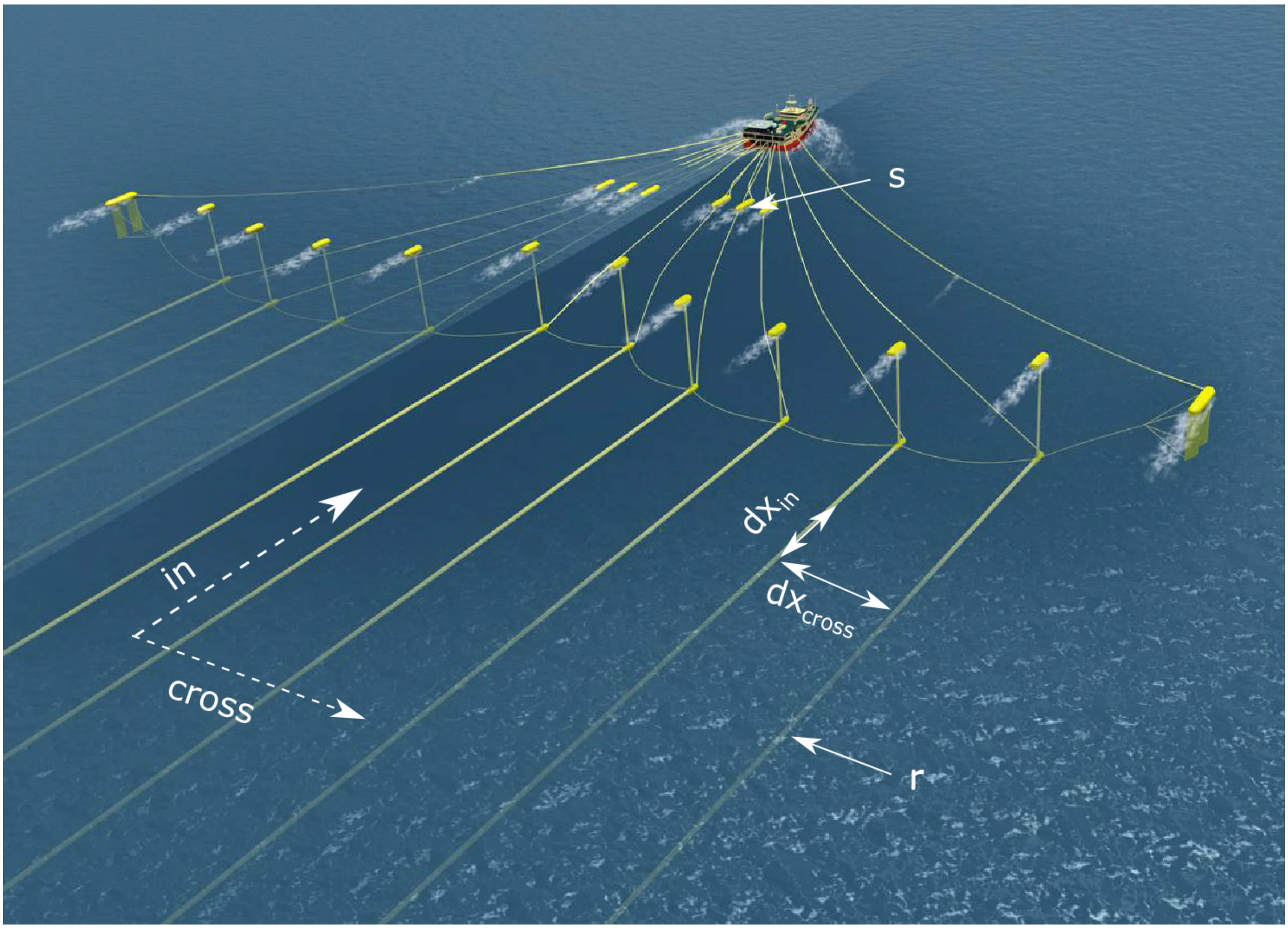

Figure 1.1: Scheme of the acquisition of a conventional towed-streamer marine survey. s: an array of airguns used to generate a source wavefield; r: a streamer containing sensors that record the wavefield; in: inline direction; cross: crossline direction; $\mathrm{dx}_{\mathrm{in}}$: spacing of the sensors in a streamer; $\mathrm{dx}_{\mathrm{cross}}$: streamer spacing.

1.1.2 Acquisition setup

Nowadays, most of the towed-streamer marine surveys are 3D. In this case, several streamers are towed by the vessel. A scheme of a conventional 3D marine acquisition setup is shown in Figure 1.1. Several parameters, for instance the number of sources, the number of streamers, the length of the streamers, the streamer spacing, and the shot interval, vary from one survey to another. These parameters vary because they are adjusted for each survey to optimize the extraction of a targeted information that is specific to the survey. I will describe one survey with standard parameters to give an idea of the scale of a marine seismic survey. The vessel sails over the area of interest along parallel straight lines. Two sources (arrays of air guns) called flip and flop are alternatively triggered. The distance between two activations of flip is 37.5 meters; the same applies for flop. The pressure and the vertical particle velocity wavefields emitted by each source is recorded over 9 s by sensors placed into 14 streamers separated by 75 m. Each streamer is 8 km long and contains a group of sensors every 12.5 m that records both the vertical particle velocity and the pressure wavefields. The depth of

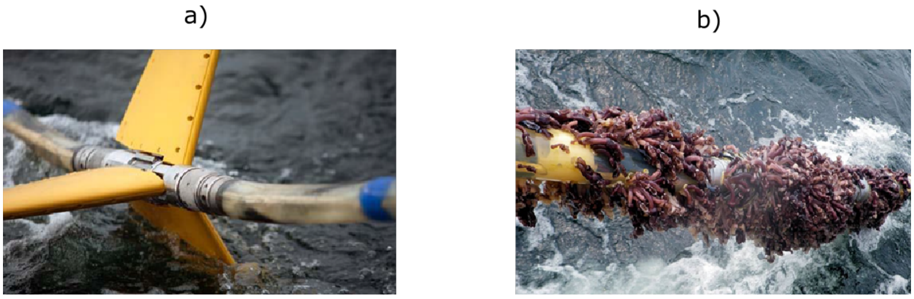

Figure 1.2: Examples of sources of mechanical noise. a) A steering device often called a "bird" and b) barnacles growing on a seismic cable.

the sources is 7 m and the depth of the receivers is 20 meters.

1.1.3 Acquisition-related issues

Unfortunately, the acquisition of the seismic data is not ideal with respect to the later processing and imaging. Two main issues are the noise contamination and the poor spatial sampling.

Noise contamination

Undesired ambient signals are recorded together with the seismic wavefield. These undesired signals, referred to as noise, additively contaminate the seismic data. A part of the noise is due to the swell. Swell momentarily changes the height of the water column above the sensors and creates hydrostatic pressure variations. Swell also leads to flow motions around the streamers and creates dynamic-pressure fluctuations on the surface of the streamer (Elboth et al., 2009b). These pressure variations are recorded by the sensors. Swell-related noise is low frequency; this characteristics facilitates its separation from the seismic signal (Elboth et al., 2008). Another type of noise, the mechanical noise, largely contaminates the seismic data. This noise is due to vibrations that propagate along the streamers. These vibrations are induced by perturbations of the movement of the streamers in the water. The two well-known causes of these perturbations are steering devices and barnacles. The steering devices are placed at regular distances along the streamers to control their position. The barnacles are crustaceans that can attach and grow on the streamers; they are particularly abundant in warm waters. Pictures of a steering device and some barnacles are shown in Figure 1.2.

Figure 1.3 presents a shot gather taken from a 3D field seismic data set to illustrate the mechanical noise in the seismic data. A common shot gather refers to the set of traces that were recorded for the same shot point. For a 3D acquisition, this set of traces forms a data cube having the dimensions time, inline offset, crossline offset, where the inline and crossline offsets are the inline and crossline components of the source-receiver offset vector, respectively. Note that a t^2 amplitude correction was applied to the data shown in Figure 1.3 to compensate for the absorption and the geometrical spreading. Also, the frequencies of the data were progressively attenuated from 15 to 5 Hz and muted below 5 Hz due to very poor signal-to-noise ratio (S/N) in this frequency range. In Figure 1.3a, an inline and a crossline slice through the shot gather are shown. An inline slice corresponds to the data

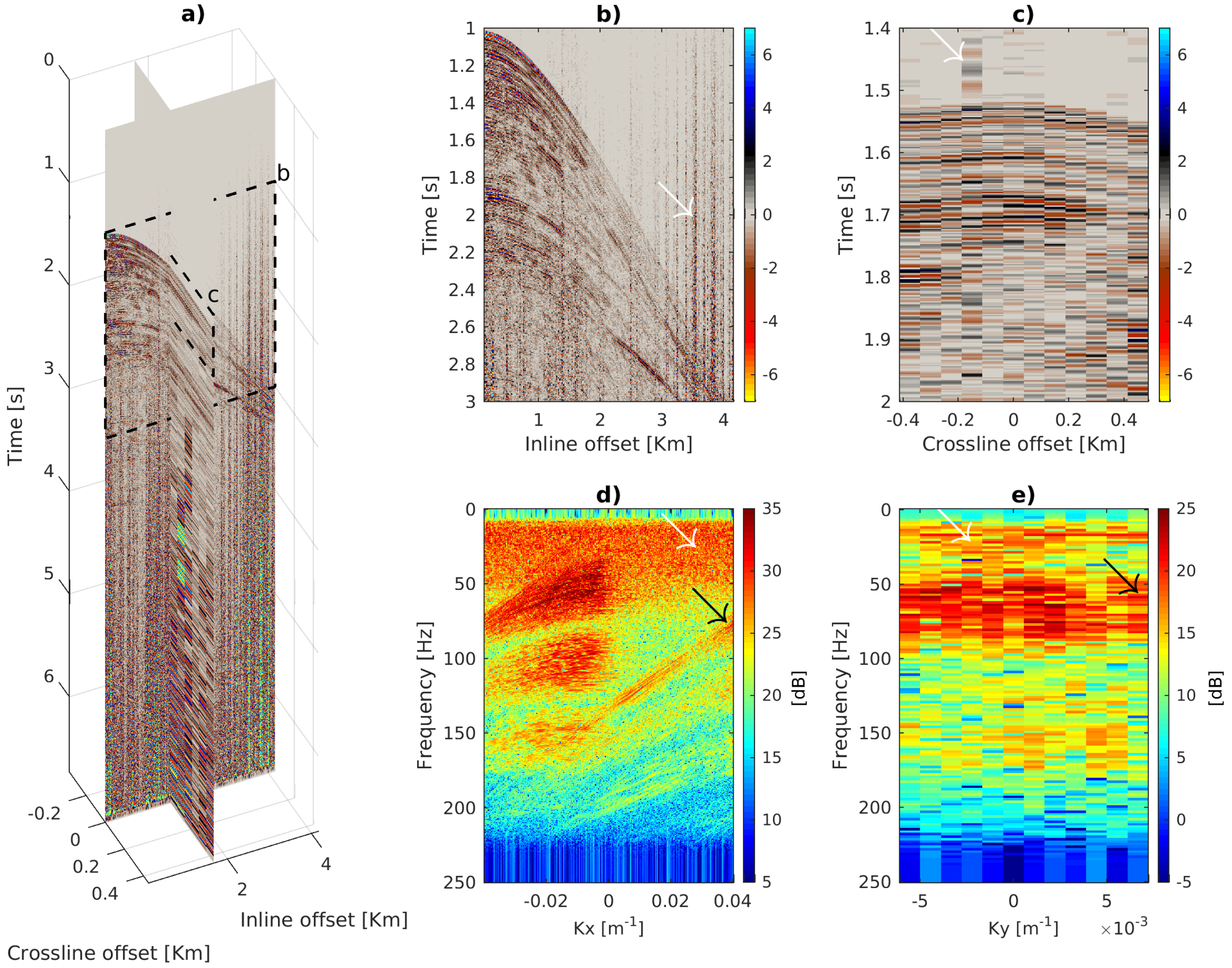

Figure 1.3: Illustration of the mechanical noise and the aliasing in the seismic data. a) an inline and a crossline slice through a shot gather. The dashed frame boxes enclose b) a window of the inline slice and c) a window of the crossline slice. d, e) the $f-k$ amplitude spectra of the inline slice window and the crossline slice window, respectively. The white arrows point out locations where the mechanical noise is observed, whereas the black arrows point out the position of aliased energy.

recorded by one streamer, whereas a crossline slice corresponds to the data recorded by the receivers of the streamers having an identical position in the streamer. A window from each slice is shown in Figure 1.3b-c. The $f-k$ (i.e., frequency-wavenumber) amplitude spectra of the data within the two selected windows are presented in Figure 1.3d-e. We can observe mechanical noise in the data, as pointed out by the white arrows. In time, the mechanical noise is localized on few neighboring traces. This behavior is to be expected considering the fact that the streamer vibrations are generated locally and their amplitudes rapidly decay with distance. In the $f-k$ domain, the mechanical noise is significant from 0 to 70 Hz and is spread across all wavenumbers. It is spread across all wavenumbers because the velocity of the vibrations is very low (lower than 100 m/s). The mechanical noise is challenging to remove because it can be coherent in space and time and it largely overlaps the seismic signal in the $f-k$ domain. There are other types of coherent noise that can contaminate marine seismic data but they will not be investigated in this thesis. A description of these types of noise is presented by Elboth (2010).

In addition to coherent noise, there is some energy that appears to have no spatiotemporal coherency. This energy is called random noise. It can be caused by association of many sources, e.g., the ambient sea signal, human activities, sensor inaccuracy, etc.,... or it can correspond to seismic energy that is not interpretable. Uninterpretable seismic energy can be weak scattering energy that is too poorly sampled to appear with spatiotemporal coherency in the data.

If the noise is not removed in an early stage of processing, it alters the data-dependent processing methods, e.g., surface-related multiple elimination, and it reduces the resolution of the final image.

Poor spatial sampling

To have a complete discrete description of a continuous signal, the sampling should satisfy the Nyquist criterion (Shannon et al., 1993). For a one-dimensional (1D) temporal signal containing energy up to a maximum frequency f_{max}, the Nyquist criterion dictates to sample the data regularly with a rate dt of at least

$$dt = \frac{1}{2f_{\text{max}}} \, . \tag{1.1}$$

In this case, the description of the signal is complete because the continuous signal can be reconstructed exactly from the discrete signal using a Fourier reconstruction. If the Nyquist criterion is not respected, the signal part lying on frequencies above the Nyquist frequency $1/(2dt)$ will be expressed with lower frequencies in the Fourier domain, and a Fourier reconstruction would be incorrect. In that case, the signal is said to be aliased, and a reconstruction is not possible unless other a priori information is available. In the seismic data, there is no aliasing issue in time. The analog signal recorded by the sensors is band limited due to the instrument response, and when it is later digitized, it is sampled at a rate that satisfies the Nyquist criterion.

To avoid aliasing in a spatial dimension x, the Nyquist criterion imposes to sample the data with the spatial sampling rate dx of at least

$$dx = \frac{1}{2k_{x,\text{max}}} \, , \tag{1.2}$$

where $k_{x,\text{max}}$ is the maximum magnitude of the wavenumbers (spatial frequency) containing signal energy. Since the seismic signal is a wavefield that follows the dispersion relationship, the magnitudes of the wavenumbers of the monochromatic plane waves that compose the wavefield are given by

$$|k_x| = \frac{|f \sin(\alpha)|}{v_r} , \tag{1.3}$$

where the wavenumber k_x is expressed in m^{-1}, f is the temporal frequency of the wave, v_r is the velocity of the wave at the receiver location, and α is the incidence angle of the wave. Considering the worst-case scenario, i.e., the incidence angle of the wave is 90°, the relationship in equation 1.3 shows that the upper bound of k_max is f_max/v_r. Therefore, to avoid spatial aliasing and ensure a complete description of a wavefield, the Nyquist criterion imposes to record the data with a spatial sampling rate of at least

$$dx = \frac{v_r}{2f_\text{max}} . \tag{1.4}$$

The frequency range of interest may go up to 200 Hz and the seismic wavefield propagates in the water with a velocity of approximately 1500 m/s, so the criterion given in equation 1.4 dictates to record the seismic wavefield with a spatial sampling rate of 3.75 m. In the seismic acquisition described in section 1.1.2, the spatial sampling rate was 12.5 m in the inline direction and 75 m in the crossline direction, and therefore, the spatial sampling is too coarse to obtain a complete description of the seismic wavefield up to a frequency of 200 Hz. The data can be aliased from 60 Hz in the inline direction, and from 10 Hz in the crossline direction. The $f-k$ amplitude spectra shown in Figure 1.3d-e illustrate how the spatial aliasing appears in the seismic data. The black arrows point out some energy that should lie on non-sampled positive wavenumbers, but that lie on negative wavenumbers instead.

The wavefield is recorded with a spatial sampling that is coarser than the one imposed by the Nyquist criterion. The spatial sampling is set coarse due to cost reasons and practical limitations. For instance, because the vessel has a limited towing capacity, towing the streamers closer to each other would reduce the crossline aperture and the covered surface per line. Hence, more lines would be required to survey the same area, which would convert into more time and more cost.

In addition, the wavefield may be sampled at locations that deviate from the intended ones because marine currents often drift the sensors from the planned positions. The lateral deviation of a streamer away from the towing direction is known as streamer feathering. In most of the cases the deviation is within few meters, but it can be larger in bad weather. Hence, the recorded traces do not lie exactly on a regular sampling grid. Besides, sensors or other elements of the recording system sometimes break unexpectedly during the survey, leading to missing data or bad traces. This missing traces leave gaps in the data set.

The aliasing and sampling irregularities cause errors if they are not corrected in an early stage of the processing. The aliasing leads to an erroneous $f - k$ representation of the seismic wavefield and therefore alters the $f - k$ domain-based processing steps such as wavefield separation, designature, demultiple, or migration. In addition, if the irregular sampling grid is approximated by a regular grid for convenience, the processing and imaging methods that depend on the location of the traces are partly mislead and are inaccurately applied. The errors induced by the aliasing and the sampling irregularities impacts the resolution and reliability of the final seismic image.

I presented two acquisition-related issues, namely, the noise contamination, and the poor spatial sampling. They were presented because data applications in this thesis aim to correct for those issues. However, note that the seismic surveys have other imperfections, e.g., the small crossline aperture or the poor spatial sampling of the sources.

1.2 Objectives and outline of the thesis

This section briefly introduces the methods that will be investigated, then it states the objectives of the thesis, and it finally presents the organization of the thesis.

1.2.1 Objectives of the thesis

The imperfections of the acquisition degrade the resolution of the final seismic image if they are not corrected in an early stage of the processing sequence. I aim to achieve (1) random noise attenuation, (2) separation of coherent noise from seismic signal, and (3) interpolation of the seismic data over a dense and regular grid, by means of sparse representations.

Signal and noise separation or interpolation beyond aliasing can be seen as an underdetermined problem. There is not enough information in the problem for the solution to be unique. For instance, in the interpolation problem, several Fourier representations fit the data at the recorded locations and are candidates to interpolate the signal. To resolve the underdetermined nature of the problem, a priori information needs to be integrated. Adding a priori information about the morphology of the data can be sufficient. Such information can be integrated to the problem using sparsity promotion. A sparsity promoting process picks the solution that has the sparsest representation in a selected transform domain. A sparse representation in the transform domain has few high amplitude coefficients, whereas the other coefficients have an amplitude close to zero. In other words, the signal is represented using only few basis vectors of the transform domain. This is possible only if these basis vectors describe the different morphological elements of the signal. Consequently, the sparse representation follows the morphology described by the selected transform domain. Hence, promoting sparsity in a transform domain whose basis vectors describe the a priori morphology of the seismic signal offers the necessary requirements to achieve the aforementioned seismic processing tasks 1-3.

There are several sparse representation-based methods that have the potential to achieve the seismic processing tasks 1-3. To attenuate the random noise, a method consists in transforming the data in a domain that describes the morphology of the seismic signal, and approximating the data with the high amplitude coefficients only. In the transform domain, the noise coefficients have low amplitudes because the noise does not significantly correlate with the basis vectors due to its random character. The signal coefficients have high amplitudes because all the signal energy is concentrated on few coefficients only. Therefore, approximating the data with the high amplitude coefficients attenuates random noise, whereas it preserves the signal to a high degree. To separate the coherent noise from the seismic signal, a method consists in representing the data in a domain where the signal and the noise are sparse and lie on two different subdomains. This is sufficient to separate the noise from the signal because the signal can be retrieved by muting the coefficients of the noise subdomain. Finally, to interpolate or regularize the data, a method consists in finding its sparse representation in a domain in which only the densely sampled wavefield is sparse, whereas other solutions are not.

For the three methods described in the preceding paragraph, there are several transform domains that may describe the a priori morphology of the seismic data, and many processes that are candidates for solving the sparse inversion problem. However, the different transform domains and sparse solvers are more or less appropriate depending on the data and the problem. The objectives of the thesis are to

(i) examine the different sparsity promoting processes,

(ii) investigate the different transform domains that can lead to sparse representation of the seismic data,

(iii) select an appropriate transform domain and sparsity promoting process for each of the processing tasks 1-3 based on the results from (i) and (ii), and implement a method that can achieve the processing task with high-quality results and a minimal human interaction.

1.2.2 Thesis outline

Chapter 2 starts by examining the different sparsity promoting problems and the algorithms that can solve them. Then, it assesses the effectiveness of sparse representations for random noise attenuation, coherent noise separation, and signal reconstruction. Finally, it investigates the transforms and dictionaries that can enable sparse representation of the seismic data. Chapter 3 explains the scientific contribution of the articles I-III. Chapters 4 contains article I, which proposes a method to attenuate the random noise in seismic data. Chapter 5 contains article II, which proposes a method to separate the mechanical noise from the seismic data. Chapter 6 contains article III, which proposes a method to reconstruct the seismic wavefield that is coarsely recorded during 3D marine acquisitions. Finally, chapter 7 gives some conclusions and presents an outlook.

Chapter 2

Scientific Background

2.1 Sparse optimization problems

This section first explains the notion of representation in a transform domain, and then it presents the sparsity promoting problems and the sparse solvers.

2.1.1 Notion of representation in a transform domain

The transform domain is defined by a dictionary. The term dictionary refers to a set of vectors called atoms. The atoms are denoted with $\mathbf{a}_1$, ..., $\mathbf{a}_K$ and their size is denoted with N. The atoms are stored in the dictionary matrix $\mathbf{D}$ such that $\mathbf{D} = [\mathbf{a}_1 \ldots \mathbf{a}_K]$. In the synthesis approach, a signal $\mathbf{y}$ of size N is reconstructed using a linear combination of the atoms. Hence,

$$\mathbf{y} = \mathbf{D}\mathbf{x} \, , \tag{2.1}$$

where the vector $\mathbf{x}$ of size K is a solution that represents the signal in the domain defined by the dictionary. If $K > N$, the dictionary is said to be redundant; in this case a solution $\mathbf{x}$ that represents a signal $\mathbf{y}$ is not unique.

In some cases, the dictionary has additional properties; for instance, the dictionary can be a frame, a tight-frame, or an orthonormal basis (Kovacevic and Chebira, 2007). The dictionary is a frame of the Hilbert space $\mathcal{H}$ if there exists two constants A and B with $B \geq A > 0$ such that

$$A||\mathbf{y}||_2^2 \leq \sum_{j=1}^{K} |\mathbf{a}_j^{\mathrm{T}}\mathbf{y}|^2 \leq B||\mathbf{y}||_2^2 \, , \forall \, \mathbf{y} \in \mathcal{H} \, . \tag{2.2}$$

A frame is said to be complete, i.e., for any signal $\mathbf{y}$, there is a solution $\mathbf{x}$ in the frame domain. If $A = B$, the frame is called a tight frame. In that case $\mathbf{D}\mathbf{D}^* = \mathbf{I}$, where * is used to denote the adjoint of a matrix and $\mathbf{I}$ is the identity matrix. This property conditions and stabilizes sparse inversions and is exploited to build quick sparse solvers. If in addition the atoms are normalized, the tight-frame is necessarily an orthonormal basis, which ensures that for any signal $\mathbf{y}$, there is a unique solution $\mathbf{x}$ in the basis domain, and $\mathbf{D}^{\mathrm{T}}\mathbf{D} = \mathbf{D}\mathbf{D}^{\mathrm{T}} = \mathbf{I}$. Hence, orthonormal bases benefit from easy forward transformation operators because a representation $\mathbf{x}$ of a signal $\mathbf{y}$ can be simply obtained by multiplying the transpose of the frame to the signal ($\mathbf{x} = \mathbf{D}^{\mathrm{T}}\mathbf{y}$).

2.1.2 The exact sparse representation problem

Strictly speaking, the sparse representation problem consists of finding the exact representation of the signal in the dictionary domain that has the least ℓ_0-norm. The ℓ_0-norm is not a proper mathematical norm, it has been redefined as the number of nonzero coefficients in a vector or an array (Donoho and Elad, 2003; Donoho, 2006). Hence, a representation that is sparse in the ℓ_0 sense is a representation that has a small number of non-zero coefficients in the transform domain. The sparse representation problem can be formally written as follows

$$\min_{\mathbf{x}} ||\mathbf{x}||_0 \text{ subject to } \mathbf{y} = \mathbf{D}\mathbf{x}\,. \tag{2.3}$$

The problem in equation 2.3 is not necessarily solvable. There might be no solution, or an infinity of solutions, that exactly represent the recording. It depends on the dictionary. If the dictionary is a basis, there is a unique solution in the dictionary domain that exactly represents the signal. This solution is hence the sparsest one for that dicitonary and is the solution to the problem in equation 2.3. If the dictionary is a redundant frame, there are an infinity of solutions that can exactly represent the recording. Yet, the solution of the sparse representation problem can still be unique if the sparsest solution is unique. One can verify if a solution is guarantied to be unique using the spark, or the mutual coherence (Elad, 2010, p.17-33). The spark of a dictionary is the smallest number of atoms of the dictionary that are linearly dependent. The calculation of the spark is quite complex, and hence, using the mutual coherence is often preferred. The mutual coherence of a dictionary quantitatively characterizes the dependence between its atoms; it is defined as the largest absolute normalized inner product between its atoms. mathematically, it reads

$$\mu(\mathbf{D}) = \max_{1\le i,j\le m, i\ne j} \frac{|\mathbf{a}_i^{\mathrm{T}}\mathbf{a}_j|}{||\mathbf{a}_i||_2||\mathbf{a}_j||_2}\,. \tag{2.4}$$

Note that the atoms of a dictionary are often already normalized, which makes the normalization in equation 2.4 unnecessary. A solution $\mathbf{x}$ of the sparse representation problem given in equation 2.3 is unique if (Bruckstein et al., 2009)

$$||\mathbf{x}||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)\,. \tag{2.5}$$

In this case, the solution is unique because the atoms are not dependent enough for a second solution to be as sparse as $\mathbf{x}$. A second solution would necessarily be denser.

2.1.3 Sparse approximation problems

When dealing with recorded data, the signal can rarely be exactly reconstructed using a representation vector that has many zero coefficients. Hence, looking for a sparse representation that exactly represents the signal leads to a solution that is not very sparse, i.e, that has no or few zero coefficients. However, it is often possible to reconstruct the signal to a high degree of accuracy using a representation vector that has many zero coefficients. Hence, when looking for a sparse representation, it is often preferred to tolerate a small representation error, in the aim of finding a solution that is sparser. In this case, we can speak of a sparse approximation problem. There are several possibilities for setting the problem. One of them consists in tolerating an error below a fixed threshold ϵ and looking for the sparsest solution (Bruckstein et al., 2009). This problem reads

$$\min_{\mathbf{x}} ||\mathbf{x}||_0 \text{ subject to } ||\mathbf{y} - \mathbf{D}\mathbf{x}||_2 \leq \epsilon \,. \tag{2.6}$$

Alternatively, one can impose a constraint on the ℓ_0-norm of the solution and search for the solution that provides the least error (Tropp, 2004). Such a problem can be mathematically expressed as

$$\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{D}\mathbf{x}||_2 \text{ subject to } ||\mathbf{x}||_0 \leq T \,, \tag{2.7}$$

where T is the threshold that fixes the maximal ℓ_0-norm of the solution. As the ℓ_0-norm of $\mathbf{x}$ is also the cardinality of the solution, this problem can be referred as the cardinality-constrained sparse approximation problem.

Similarly as for the exact sparse representation problem, there exist conditions that guaranty the stability of the solution (Elad, 2010, p.79-109). When looking for a sparse approximation, solving the problem in equation 2.6 is not equivalent to solving the problem in equation 2.7. The two problems will often lead to different solutions. Depending on the situation, one problem may be more adequate than the other. For instance, if one is concerned about signal preservation, he may want to ensure a small data misfit, which would imply a formulation as in equation 2.6. In another situation, one may have a priori information about the cardinality of the solution, which would lead to the formulation in equation 2.7.

The problems in equations 2.6 and 2.7 take the synthesis approach to find a sparse representation. This is called the synthesis approach because it seeks a reconstruction of the signal as a combination of atoms. There exists also the analysis approach where the signal is represented via its inner product with the dictionary atoms. Elad et al. (2007) and Rubinstein (2011) explain the conceptual and technical differences between the two approaches, and show how they differ for problems involving redundant dictionaries. As the synthesis approach is simpler and more intuitive, this is the approach that will be taken in this thesis.

2.1.4 Sparse solvers

If the dictionary is an orthonormal basis, the sparse approximation problem can be simply solved using hard thresholding. For instance, to solve the problem in equation 2.6, one can first apply the forward transform operator to find the exact representation of the signal ($\mathbf{x} = \mathbf{D}^{\mathrm{T}}\mathbf{y}$), and mute the smallest amplitude coefficients that jointly have a norm just below ϵ. Similarly, to solve the problem in equation 2.7, one can apply the forward transform operator to find the exact representation of the signal, and mute all the coefficients apart from the T coefficients that have the largest amplitudes.

For redundant dictionaries, solving exactly the sparse representation or approximation problems is very complex and is in general intractable. Such problems belong to the class of non-deterministic polynomial-time (NP)-hard problems (Davis et al., 1997). In practice, these problems are solved using sub-optimal processes, which include matching pursuit (MP) algorithms (Pati et al., 1993; Davis et al., 1997; Needell and Tropp, 2009; Donoho et al., 2012), Convex relaxation methods (Gorodnitsky and Rao, 1997; Chen et al., 1998; Donoho and Elad, 2003), and iterative shrinkage methods (Daubechies et al., 2004; Elad, 2006).

The MP algorithms, e.g., orthogonal matching pursuit (OMP) (Pati et al., 1993), take a greedy approach to select the few atoms that best match the elements of the signal, and then compute the representation of the signal as a linear combination of these atoms. The set of selected atoms is called the support of the sparse representation and is denoted with $\{\mathbf{a}_j\}_{j\in\Lambda}$, where Λ is the set of indexes of the atoms in the support. The approach to select the support is called greedy as the atoms of the support are selected one by one with an iterative process. At each iteration, all the atoms that are in the support are tested, and the one that can lead to the greater reduction of the representation error is selected and added to the support. Here, the representation error corresponds to the difference between the true signal and its sparse approximation given the support obtained in the previous iteration. Afterward, the representation is updated considering the new support. Hence, MP algorithms solve several local optimization problems to find the solution of the global sparse representation problem. Such an approach is fast, robust, and accurate, for relatively low-dimensional problems. This approach is however not well adapted to a high-dimensional problem, in which case it requires many iterations and a large computational effort to converge to the solution.

In convex relaxation methods, the sparse representation problem is relaxed by switching the discontinuous ℓ_0-norm with a continuous ℓ_p-norm, where p is strictly higher than 0 and lower than 1. The case $p = 1$ leads to the basis pursuit (BP) optimization problem (Chen et al., 1998). After relaxation, the problem is better conditioned and can be solved with different solvers, for instance with the focal underdetermined system solver (FOCUSS) (Gorodnitsky and Rao, 1997), which uses an iterative-reweighed-least-square (IRLS) scheme.

Iterative shrinkage methods (Daubechies et al., 2004; Elad, 2006) loop over the three steps: apply the transpose of the dictionary to the representation error, shrink the obtained coefficients, and use the result to update the representation of the signal. The rules used to shrink the coefficients and to use the result for updating the representation vary from one algorithm to another. In contrast to the greedy approach, such algorithms use global optimization schemes and are suitable for high-dimensional problems.

As sparse solvers are suboptimal to solve the sparse representation problem, they do not always find the exact solution to the problem. Only under specific conditions, they are guaranteed to find the exact solution. For instance, if the condition given in equation 2.5 is satisfied, many solvers, including OMP, are guarantied to find the exact solution to the problem in equation 2.3 (Elad, 2010, p.55-77).

In this thesis, the dictionaries used will often be of relatively low dimensions but highly redundant. In such cases, the sparse approximations will be computed using the OMP algorithm, which is fast, robust, and accurate, under those conditions.

2.2 Sparse representations and signal processing

This section investigates the effectiveness of sparse representation-based processes for random noise attenuation, coherent noise separation, and data interpolation.

2.2.1 Random noise attenuation

To formally define the random noise contamination problem, let us denote a recording $\mathbf{z} \in \mathbb{R}^N$ containing a signal of interest $\mathbf{y} \in \mathbb{R}^N$ and white Gaussian noise $\mathbf{n} \in \mathbb{R}^N$ of zero mean and σ^2 variance. Such noise is denoted with $\mathcal{N}_N(0, \sigma^2)$. In addition, let us consider a dictionary $\mathbf{D} \in \mathbb{R}^{N \times K}$ defining a domain where the signal has a sparse representation. Then, the noise contamination model reads

$$\begin{aligned} \mathbf{z} &= \mathbf{y} + \mathbf{n}\,, \\ \mathbf{y} &= \mathbf{D}\mathbf{x}\,, \\ \mathbf{n} &\sim \mathcal{N}_N(0, \sigma^2), \end{aligned} \tag{2.8}$$

where the vector $\mathbf{x}$ is assumed to contain a small number of nonzero coefficients.

Attenuating the noise using sparsity promotion can be quite simple. It can be achieved by computing a sparse approximation of the recording in the dictionary domain. Since the noise is random, it cannot be adequately represented by the sparse approximation, and it is attenuated. There are many possibilities to compute the sparse approximation, but they are more or less suitable depending on the a priori information. If the variance of the noise is known a priori, it is reasonable to solve the error-constrained problem in equation 2.6 and aim for a solution with a representation error close to the a priori norm of the noise $\sigma\sqrt{N}$. In the case in which the variance of the noise is unknown, but the ℓ_0-norm of the signal representation is known a priori, it is preferable to compute the sparse approximation using the cardinality-constrained problem in equation 2.7.

There are two critical questions that need to be risen before performing random noise attenuation: What are the conditions to preserve the signal? How much noise will be removed? In the general case, it is not possible to answer these questions. However, it is possible to answer them for certain cases. For instance, let us consider the case in which the atoms are normalized and uncorrelated, and the ℓ_0-norm of the true solution is known a priori. The noise is attenuated by solving the problem in equation 2.7, where $T = ||\mathbf{x}||_0$. In this case, the signal is preserved if the correct support is selected to compute the approximation. Roughly speaking, the correct support is selected if the signal coefficients stand out from the noise coefficients in the dictionary domain. Hence, to preserve the signal, the minimum absolute value of the projection of the recording on the true support $\min_{j\in\Lambda} |\mathbf{z}^{\mathrm{T}}\mathbf{a}_j|$ should be larger than the maximum absolute value of the projection of the noise on the rest of the dictionary atoms $\max_{j\notin\Lambda} |\mathbf{n}^{\mathrm{T}}\mathbf{a}_j|$. A close upper bound of $\max_{j\notin\Lambda} |\mathbf{n}^{\mathrm{T}}\mathbf{a}_j|$ can be found using the result derived by

Berman (1964). The authors show that the maximum absolute value of n random variables χ_1, ..., χ_n, $n \to \infty$, following an identical Gaussian distribution of mean 0 and variance σ^2, has the close upper bound

$$\max_i |\chi_i| \leq \sigma\sqrt{2\log(n)} \, . \tag{2.9}$$

This result was later used by Donoho and Johnstone (1994) and Donoho (1995) to set to an optimal threshold for noise attenuation by shrinkage in a transform domain. Similarly, this result can be used to show that there is a high probability that $\max_{j \notin \Lambda} |\mathbf{n}^{\mathrm{T}}\mathbf{a}_j| < \sigma\sqrt{2\log(K - ||\mathbf{x}||_0)}$. Therefore, if the coefficients of the signal are above $\sigma\sqrt{2\log(K - ||\mathbf{x}||_0)}$, then there is a high probability to select the correct support of the representation, and to preserve the signal. Yet, even if the correct support is selected, the noise is not entirely attenuated. The noise that is correlated to the support still remains. Considering the random nature of the noise, the norm of the remaining noise can be estimated to be $\sigma\sqrt{||\mathbf{x}||_0}$. Hence, the quantity of remaining noise is a square root function of the ℓ_0-norm of the representation of the signal. This highlights the importance of selecting a dictionary that enables a highly sparse representation of the signal.

To illustrate the effectiveness of sparse representations for random noise attenuation, I will present simple numerical experiments, in which the recording and the signal are synthesized following the model described in equation 2.8. The dictionary matrix **D** used is the discrete cosine transform (DCT) basis of size 64×64. The basis vectors of the DCT basis are defined using cosine functions. Among other applications, the DCT basis is used for compression of audio signals (e.g., MP3) and images (e.g., JPEG). Eight basis vectors from the DCT basis are shown in Figure 2.1. The signal was synthesized with a linear combination of L atoms of the dictionary. As the number L is also the ℓ_0-norm of the representation of the signal in the DCT domain, such a construction of the signal enables to control its sparsity level. The L nonzero entries of **x** are selected such that they are independently distributed, they follow the same zero-mean Gaussian distribution, and the signal is normalized. The variance of the noise is selected such that the noise is normalized. Hence, in the recording, the norm of the signal is equal to the norm of the noise. The sparse representation of the recording are obtained by solving the problem in equation 2.7, where $T = L$, using hard thresholding in the DCT domain.

Three experiments, in which the signals were synthesized with L set to 2, 5, and 10, are presented in Figure 2.2. The recordings are displayed with green solid lines and are presented in the recording domain in the left plots and in the DCT domain in the middle plots. For each recording, the L nonzero coefficients that were selected to compute the sparse approximation of the signal are depicted with blue dots in the middle plot. The true representations of the signals in the DCT domain are superimposed on the results using black dotted lines. One can observe that when the coefficients are too small, they are not selected to compute the approximation of the recording, and instead, a noise coefficient is selected. The approximations of the recordings are presented with blue lines in the plots on the right side. There, the true signals are displayed with black dotted lines. The highest quality denoising is obtained for $L = 2$.

To better assess the impact of the level of sparsity on the quality of the denoising, I repeated the experiment for L ranging from 1 to 64. For each L value, the experiment was repeated 1,000 times, and for each repetition, the signal and noise were recomputed with a reselection of the random

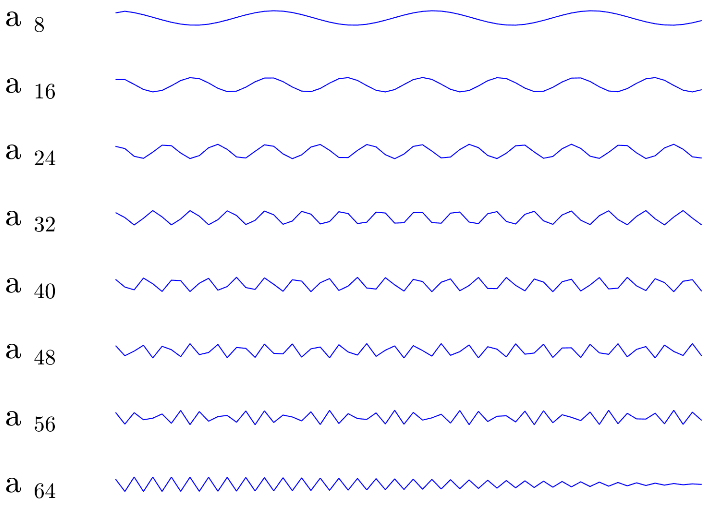

Figure 2.1: Eight basis functions of the DCT dictionary.

parameters. The error (i.e., the ℓ_2-norm of the difference between the denoised signal and the true signal) was computed as an average of the errors of the 1,000 experiments. The resulting error values are presented in Figure 2.3. The error increases as the ℓ_0-norm of the signal representation increases.

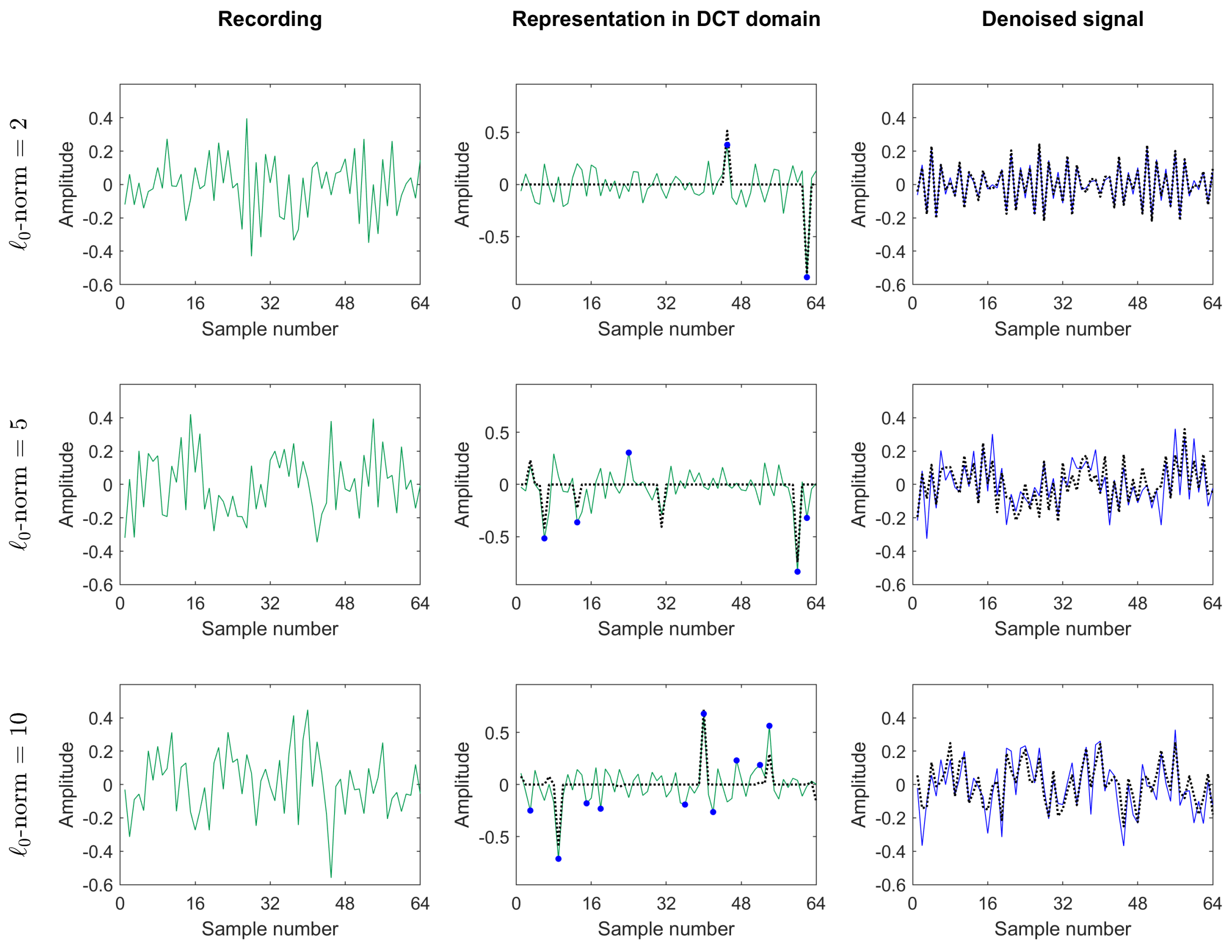

Figure 2.2: Denoising by sparse approximation in the DCT domain in the cases in which the ℓ_0-norm of the true solution is equal to 2, 5, and 10, as indicated on the left side of the plots. The noise-contaminated signals in the original domain and in the DCT domain are displayed with green lines in the left and middle plots, respectively. Each recordings was approximated with the L coefficients having the largest absolute values in the DCT domain, were L is the ℓ_0-norm of the true solution. The blue dots depict the selected coefficients. The denoised signals in the original domain are displayed with blue lines in the plots on the right side. The true signals in the DCT domain and in the original domain are superimposed on the results using black dotted lines in the middle plots and right plots, respectively.

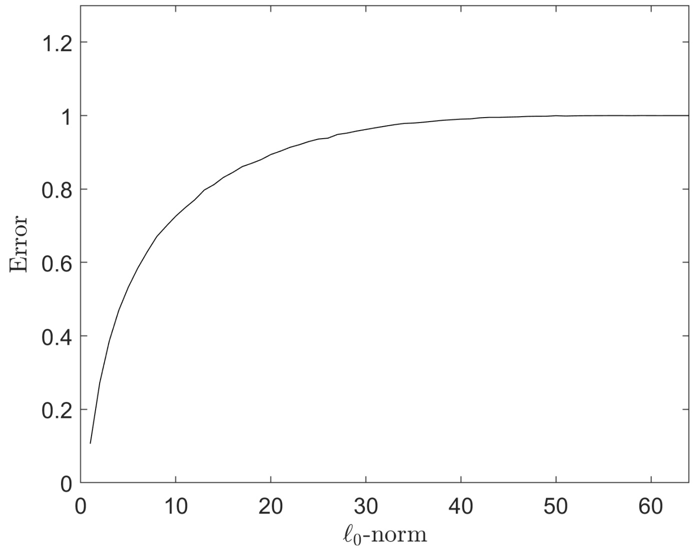

Figure 2.3: The error in noise attenuation by sparse approximation as a function of the ℓ_0-norm of the true solution. As the ℓ_0-norm of the solution increases, the error increases.

2.2.2 Coherent noise separation

In this noise contamination model, the recording **z** contains a signal **y** additively contaminated by a coherent noise **n**, and a priori information about the morphology of the signal and the noise are both known. The a priori information enables the selection of the dictionaries $\mathbf{D}_s$ and $\mathbf{D}_n$ in which the signal and the noise have a sparse representation, respectively. Such a model can be written as follows

$$\begin{aligned} \mathbf{z} &= \mathbf{y} + \mathbf{n}\,, \\ \mathbf{y} &= \mathbf{D}_s \mathbf{x}_s\,, \\ \mathbf{n} &= \mathbf{D}_n \mathbf{x}_n\,, \end{aligned} \tag{2.10}$$

where the vectors $\mathbf{x}_s$ and $\mathbf{x}_n$ are assumed to contain a small number of nonzero coefficients. In this case, the signal and the noise can be separated using morphological component analysis (MCA) (Starck et al., 2004, 2005). This method uses sparse representation as driving force to separate the different morphological components in the recording. When no representation error is tolerated, MCA seeks the sparse vector

$$\hat{\mathbf{x}}_s, \hat{\mathbf{x}}_n = \arg\min_{\mathbf{x}_s, \mathbf{x}_n} ||\mathbf{x}_s||_0 + ||\mathbf{x}_n||_0 \;\text{ subject to } \mathbf{z} = \mathbf{D}_s \mathbf{x}_s + \mathbf{D}_n \mathbf{x}_n\,, \tag{2.11}$$

and reconstructs the signal and the noise using $\mathbf{D}_s \hat{\mathbf{x}}_s$ and $\mathbf{D}_n \hat{\mathbf{x}}_n$, respectively. Note that the MCA problem in equation 2.11 is equivalent to the sparse representation problem in equation 2.3 in which the dictionary **D** is the concatenations of the dictionaries $\mathbf{D}_s$ and $\mathbf{D}_n$. Consequently, if the sum of the ℓ_0-norms of $\mathbf{x}_s$ and $\mathbf{x}_n$ is small enough to satisfy the condition given in equation 2.5, the solution of the

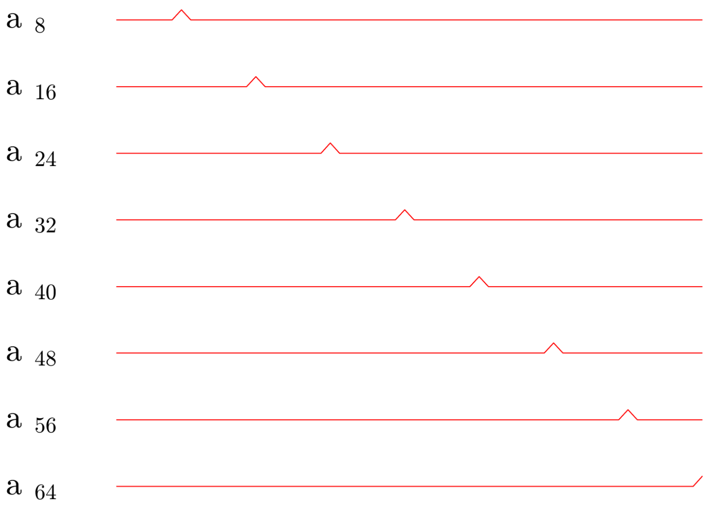

Figure 2.4: Eight basis functions of the Dirac dictionary.

problem is unique and can be retrieved using OMP; this solution leads to a perfect noise suppression.

To illustrate the capability of MCA for coherent noise suppression, I will present simple numerical experiments, in which the recording is synthesized following the model described in equation 2.10. The length of the recording is 64 samples. The signal is constructed as a linear combination of a few cosine functions, such that it is sparse in the DCT domain. The noise is impulsive and hence is sparse in the Dirac dictionary domain. The Dirac dictionary matrix is in fact an identity matrix. Ten basis functions of the DCT and Dirac dictionaries are presented in Figures 2.1 and 2.4, respectively. The total number of basis functions used to construct the signal and the noise is denoted by L, which is also the ℓ_0-norm of the solution to the signal and noise separation problem. The nonzero coefficients of the signal and noise representations in their respective dictionaries follow an identical zero-mean Gaussian distribution and both the signal and the noise are normalized. To find the solution of the MCA problem, the OMP sparse solver is used.

The results of three experiments, in which the recordings were synthesized with L set to 20, 40, and 60, are presented in Figure 2.5. The recordings are presented in the left plots. The separated noises are displayed with red lines in the middle plots and the separated signals are displayed with blue lines in the right plots. For both the signals and the noises, the truth is superimposed on the results using dotted black lines. In the case $L = 20$, the separation is exact. In the cases $L = 40$ and $L = 60$, some errors occurs in the separation.

To better assess the impact of the level of sparsity on the quality of the signal and noise separation, I repeated the experiment for L ranging from 1 to 128. For each L value, the experiment was repeated 1,000 times, and for each repetition, the signal and noise were recomputed with a reselection of the random parameters. Then, the error (i.e., the ℓ_2-norm of the difference between the denoised signal and the true signal) was computed as an average of the errors of the 1,000 experiments. The resulting

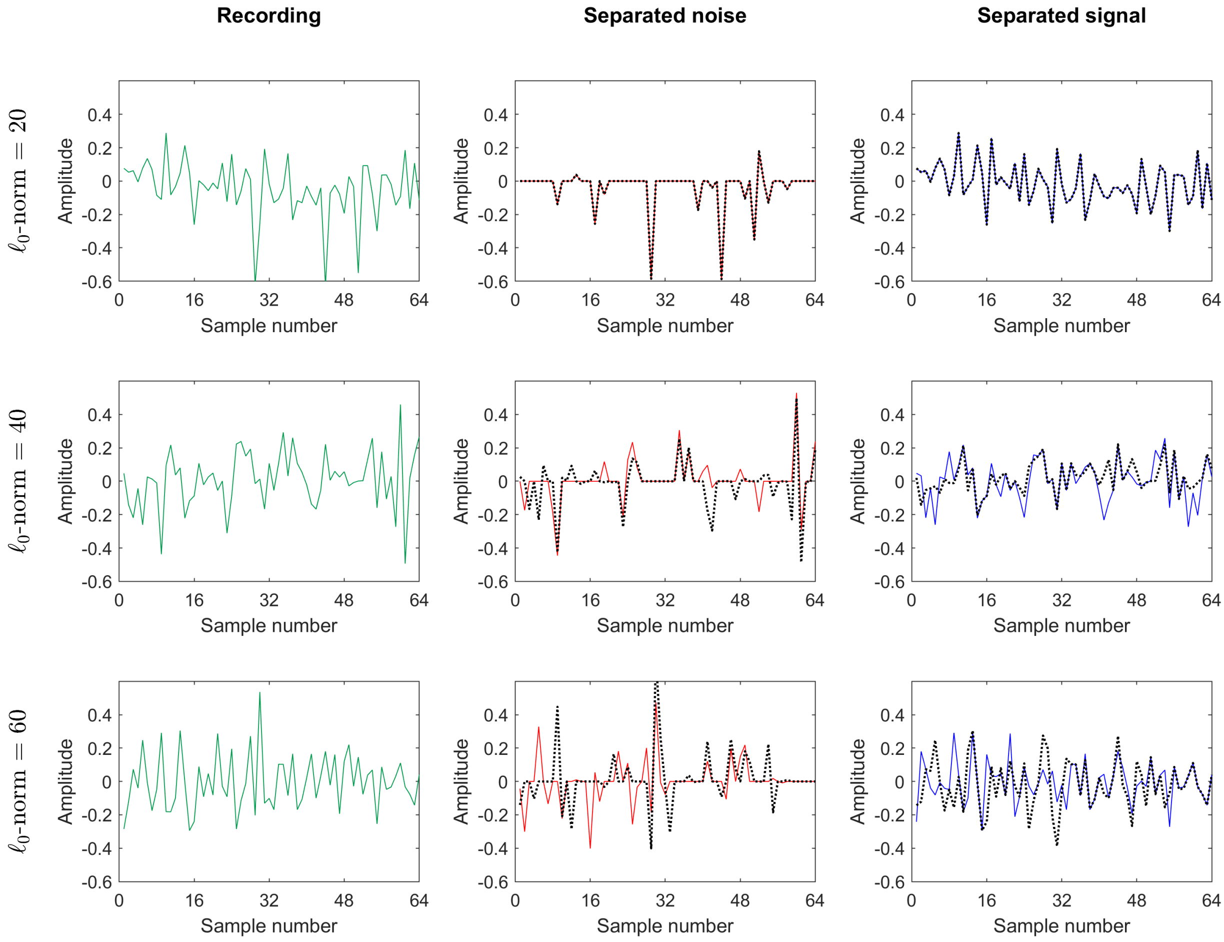

Figure 2.5: Signal and noise separation by sparse representation in the cases in which the ℓ_0-norm of the true solution is equal to 20, 40, and 60, as indicated on the left side of the plots. The recordings are displayed with green lines in the left plots, the separated noises are displayed with red lines in the middle plots, and the separated signals are displayed with blue lines in the right plots. The true noises and the true signals are displayed with black dotted lines in the middle and right plots, respectively.

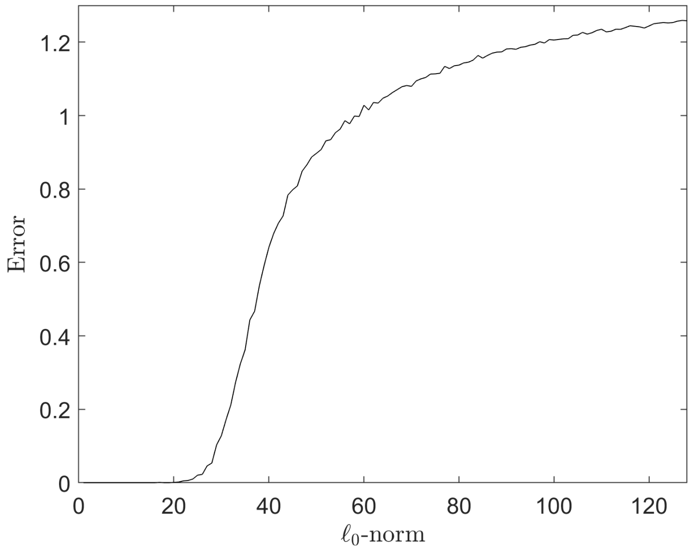

Figure 2.6: The error in signal and noise separation by sparse representation as a function of the ℓ_0-norm of the true solution. As the ℓ_0-norm of the solution increases, the error increases.

error values are presented in Figure 2.6. There, we observe that the signal and noise separation is exact, or nearly exact, when the ℓ_0-norm of the solution is smaller than 20. Also, for an ℓ_0-norm of the solution higher than 20, the error increases as the ℓ_0-norm of the solution increases.

In the experiments, the mutual coherence of the combined DCT and Dirac dictionary was 0.17. Therefore, according to the condition in equation 2.5, the solution was guaranteed to be found for $L \leq 3$. And yet, OMP was able to perfectly separate the signal from the noise for $L \leq 20$. This gives me the opportunity to point out that in many numerical applications the sparse representation problem is correctly solved beyond the theoretical bound that guaranties to find the correct solution. Note also that MCA can perfectly separate the signal from the noise, in contrast to random noise attenuation by sparse approximation. Hence, when a priori information about the noise morphology is available, using MCA is more powerful than attenuating the noise by sparse approximation of the recording.

2.2.3 Signal reconstruction

In the data reconstruction problem, the signal $\mathbf{y}$ is recorded partially. The recording $\mathbf{z}$ contains a subselection of signal samples. The position of the available samples in the vector $\mathbf{y}$ is known. Using this information, a sampling matrix $\mathbf{S}$ can be easily formed such that the recording $\mathbf{z}$ equals to the matrix multiplication of the sampling matrix $\mathbf{S}$ with the signal $\mathbf{y}$. Priors about the morphology of the signal enable the selection of a dictionary $\mathbf{D}$ that defines a domain in which the signal is sparse. Hence, the data model can be written as follows

$$\begin{aligned} \mathbf{z} &= \mathbf{Sy} \,, \\ \mathbf{y} &= \mathbf{Dx} \,, \end{aligned} \tag{2.12}$$

where the vector $\mathbf{x}$ is assumed to contain a small number of nonzero coefficients.

Given the model in equation 2.12, $\mathbf{x}$ can be recovered by finding the solution of the problem

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} ||\mathbf{x}||_0 \text{ subject to } \mathbf{z} = \mathbf{SDx} \,, \tag{2.13}$$

and reconstructing the signal with $\mathbf{D}\hat{\mathbf{x}}$ (e.g., Bruckstein et al. (2009)). The problem in equation 2.13 consists in finding a sparse representation of the recording using the dictionary $\mathbf{SD}$. Therefore $\mathbf{x}$ is retrieved under the condition that $||\mathbf{x}||_0 < (1 + 1/\mu(\mathbf{SD}))/2$ (see equation 2.5).

To illustrate the data reconstruction capability of the sparsity promoting problem presented in equation 2.13, I will present simple numerical experiments, in which the recording and the signal are synthesized following the model described in equation 2.12. The signal was synthesized with a linear combination of L atoms of the dictionary. As the number L is also the ℓ_0-norm of the representation of the signal in the DCT domain, such a construction of the signal enables to control its sparsity level. The L nonzero entries of $\mathbf{x}$ are selected such that they are independently distributed, they follow an identical zero-mean Gaussian distribution, and the signal is normalized. The recording is obtained by random selection of half of the samples of the signal. The problem in equation 2.13 is solved using OMP to reconstruct the signal.

Three experiments, in which the signals were synthesized with L set to 10, 20, and 30, are presented in Figure 2.7. The location of the available samples of the signals are indicated with green dots in the left plots. The reconstructed signals are presented on the right plots with blue lines. The true signals are superimposed on the results using black dotted lines in the left and right plots. We observe that the reconstruction of the signal is correct for L set to 10, some errors occur for L set to 20, and even more errors occur for L set to 30.

To better assess the impact of the level of sparsity on the quality of the reconstruction, I repeated the experiment for L ranging from 1 to 64. For each L value, the experiment was repeated 1,000 times, and for each repetition, the signal was recomputed with a reselection of the random parameters. The error was computed as an average of the errors of the 1,000 experiments. The resulting error values are presented in Figure 2.8. We observe that the reconstruction is exact, or nearly exact, when L, or equivalently the ℓ_0-norm of the solution, is smaller than 10. Also, for an ℓ_0-norm of the solution larger than 10, the error increases as the ℓ_0-norm of the solution increases.

When using a sparse representation for data reconstruction, the sampling scheme can be crucial. To point that out, I repeated the experiment presented in Figure 2.7 in which L was set to 10, but using uniform sampling instead of nonuniform sampling. The sampled signal and the reconstructed signal are presented in Figure 2.9. We can observe that the reconstructed signal diverges from the original signal. In contrast, there was no error when the sampling was nonuniform. The mutual coherence of **SD** is now equal to 1, whereas it was equal to 0.56 when the sampling was nonuniform. A mutual coherence that is equal to 1 means that there are at least two columns in **SD** that are identical. In that case, the problem is ill-posed. Two atoms of the dictionary can equally explain the same recorded signal. Whenever one of the two atoms is needed to reconstruct a signal, the other can be used instead. When considering only the dictionary **D**, the two atoms are however different, and using one instead of the other leads to errors when reconstructing the signal using $\mathbf{D}\hat{\mathbf{x}}$. This phenomenon is in fact related to the aliasing problem presented in section 1.1.3. A high frequency atom that is regularly sampled at a rate below the Nyquist criterion is similar to a lower frequency atom. Consequently, uniform sampling is not adapted for sparse representation-based reconstructions when using a frequency-based dictionary.

The theory and numerical results presented in this subsection demonstrate that certain signals recorded with a sampling rate that does not satisfy the Nyquist criterion can still be reconstructed using sparsity promotion in a dictionary domain. The conditions for the reconstruction to be exact concern the sampling scheme, the dictionary, and the degree of sparsity of the signal in a dictionary domain. The field of research that investigates sampling schemes and reconstruction methods to sense signals using a least number of samples is called compressed sensing or compressive sensing (Donoho, 2006; Candes and Wakin, 2008). The compressed sensing theory was used to design new seismic acquisition protocols that can sense the seismic wavefield more efficiently (Herrmann, 2010; Charles et al., 2014; Mosher et al., 2017; Kumar et al., 2017).

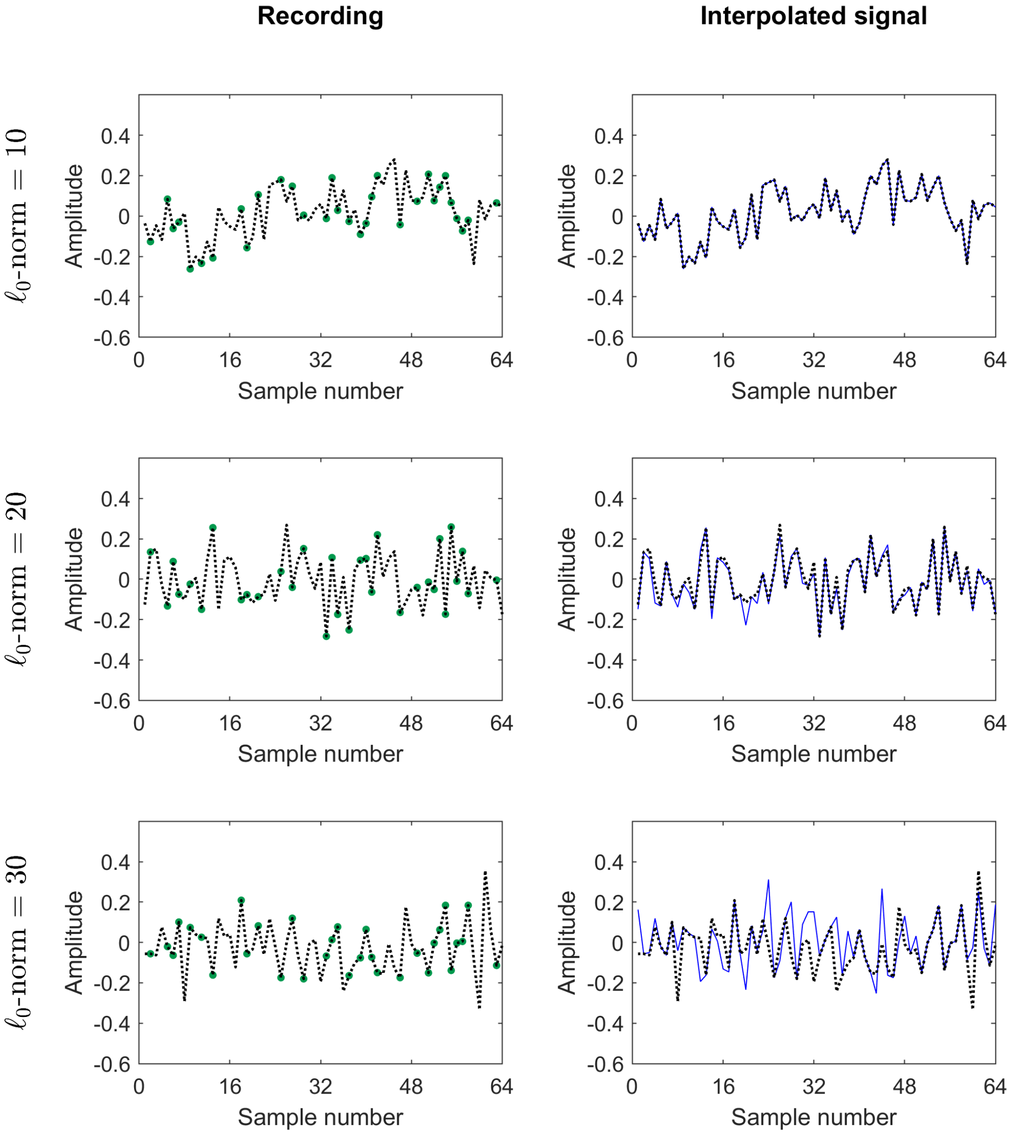

Figure 2.7: Examples of signal reconstructions using sparse representations in the cases in which the ℓ_0-norm of the true solution is equal to 10, 20, and 30, as indicated on the left side of the plots. The nonuniform sampling locations of the signals are indicated with green dots in the left plots. The reconstructed signals are displayed with blue lines in the right plots. The true signals are superimposed using black dotted lines.

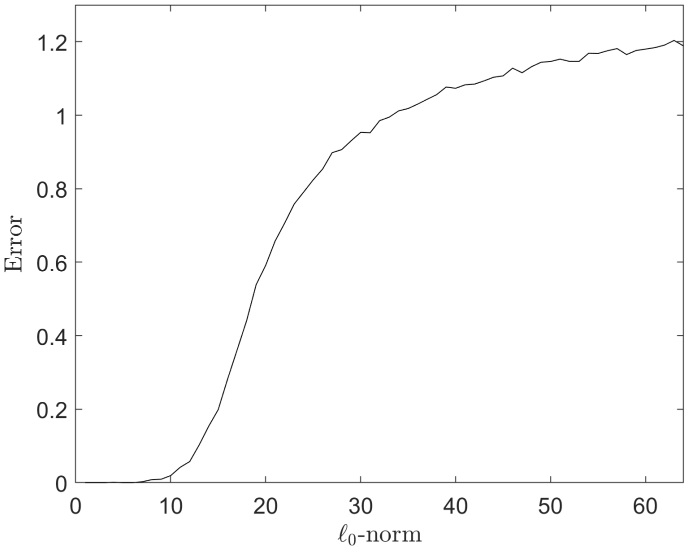

Figure 2.8: The error in signal reconstruction by sparse representation as a function of the ℓ_0-norm of the true solution. As the ℓ_0-norm of the solution increases, the error increases.

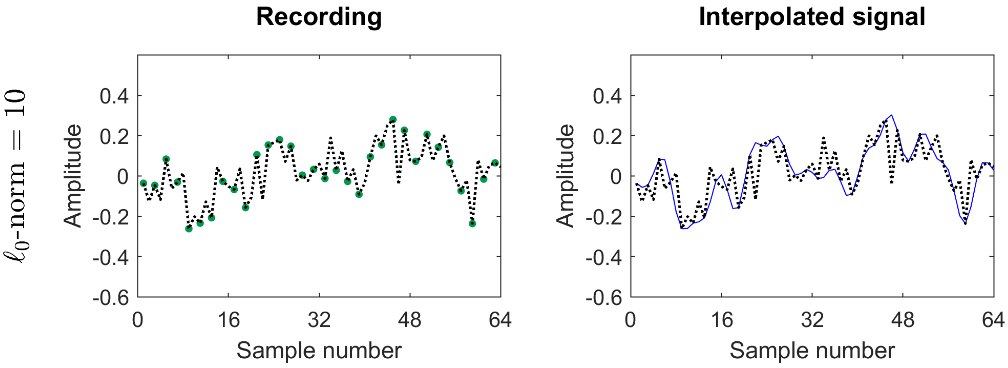

Figure 2.9: Example of signal reconstruction using a sparse representation in the case in which the ℓ_0-norm of the true solution is equal to 10 and the data is regularly sampled. The uniform sampling locations of the signal are indicated with green dots in the left plot. The reconstructed signal is displayed with a blue line in the right plot. The true signal is superimposed using a black dotted line.

2.3 Dictionaries for sparse representations of seismic data

Considering the theory and results presented in section 2.2, it is clear that the effectiveness of sparsity promoting methods relies on a high level of sparsity in the transformed domain. Therefore, this section investigates the dictionaries that can lead to sparse representations of the seismic data. Such a dictionary needs to contain atoms that describe the morphological elements of the signal. Therefore, to define a dictionary that can lead to a sparse representation of the seismic data, it is recommended to take into account the morphology of the seismic data, which is governed by the physics of the wavefield propagation. This section first refreshes the basics of the wave equation and presents possible descriptions of a wavefield from a kinematic point of view. Based on these descriptions and previous studies, it then examines predefined dictionaries that could be used to concisely represent seismic data. Finally, it presents a different approach to find a dictionary that can lead to a sparse representation of the seismic data; this second approach consists in training the dictionary on the data.

2.3.1 Descriptions of the seismic signal

In marine seismic processing, the signal of interest is an acoustic pressure or particle velocity wavefield. Such a wavefield satisfies the wave equation. The wave equation for the pressure wavefield p and the particle velocity wavefield **v** reads

$$\begin{pmatrix} \nabla^2 p \\ \nabla(\nabla \cdot \mathbf{v}) \end{pmatrix} = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\begin{pmatrix} p \\ \mathbf{v} \end{pmatrix}, \tag{2.14}$$

where ∇ is the gradient operator, ∇^2 is the Laplace operator, and c is the speed of sound at the ambient conditions.

In a homogeneous medium, the simplest solution of the wave equation is a monochromatic wave. For instance, in the global Cartesian coordinate system $\mathbf{x} = [x_1,\ x_2,\ x_3 = z]$, such a solution for the pressure wavefield can be written as

$$p(\mathbf{x}, t) = p_0\ e^{i2\pi(\mathbf{k}\cdot\mathbf{x} - ft)}\ , \tag{2.15}$$

where p_0 is the amplitude of the wave, f is the frequency, and **k** is the wavenumber vector expressed in units of m^{-1}.

In an inhomogeneous medium, describing the wavefield emitted from one source is more complex. There are several general descriptions of the wavefield that are solutions of the wave equation. Each description has its advantages and disadvantages. In the ray theory (Červený, 2001; Aki and Richards, 2002), the wavefield is seen as many high-frequency elementary waves propagating along rays of certain geometric trajectories. Considering the zero-order ray theory, the wave equation admits for each elementary pressure wave the solution

$$p(\mathbf{x}, t) = \overline{p}(\mathbf{x}) S_p[t - \mathcal{T}(\mathbf{x})]\ . \tag{2.16}$$

The amplitude factor $\overline{p}(\mathbf{x})$ determines the amplitude of the pressure, and $S_p[t]$ is the source signal. The function $\mathcal{T}$ is a real-valued function interpreted as the traveltime of the wave. Similarly, the wave equation admits for each elementary particle velocity wave the solution

$$\mathbf{v}(\mathbf{x}, t) = \overline{\mathbf{v}}(\mathbf{x}) S_v[t - \mathcal{T}(\mathbf{x})] \, . \tag{2.17}$$

The vectorial factor $\overline{\mathbf{v}}(\mathbf{x})$ determines the amplitude and polarization direction of the particle velocity vector, and $S_v[t]$ is the source signal. The vectorial amplitude factors $\overline{p}(\mathbf{x})$ and $\overline{\mathbf{v}}(\mathbf{x})$, and the function $\mathcal{T}$ are assumed to vary smoothly in space.

Moreover, using the paraxial ray theory (Hubral, 1983; Bortfeld, 1989; Červený, 2001), it is possible to locally describe an elementary wave by extrapolating the time it takes to travel along a reference ray called the central ray. Consider a central ray SG starting from the source at $S(\mathbf{x}) = [x_1^S,\ x_2^S,\ x_3^S]$ and emerging at the horizontal measurement surface at $G(\mathbf{x}) = [x_1^G,\ x_2^G,\ x_3^G]$, and a paraxial ray $S\overline{G}$ starting from the same source and emerging at the measurement surface at $\overline{G}(\mathbf{x}) = [x_1,\ x_2,\ x_3]$. The traveltime of the wave propagating along the central ray and arriving at G is denoted by $\mathcal{T}_0$ and the horizontal position of $\overline{G}$ relative to G by $\mathbf{x}' = [x_1 - x_1^G,\ x_2 - x_2^G]$. Then, in a second-order traveltime approximation, the traveltime of the wave propagating along the paraxial ray and arriving at $\overline{G}$ can be expressed as a function of $\mathcal{T}_0$ and the position of $\overline{G}$ relative to G as

$$\mathcal{T}(\mathbf{x}') = \mathcal{T}_0 + \mathbf{s} \cdot \mathbf{x}' + \frac{1}{2} \mathbf{x}' \cdot \mathbf{N}_G^S \mathbf{x}' \, . \tag{2.18}$$

The vector $\mathbf{s}$ is the gradient of the traveltime and is also called the slowness vector, and $\mathbf{N}_G^S$ is the second-derivative matrix of the traveltime. They can be written as

$$\begin{aligned} \mathbf{s} &= \begin{bmatrix} \frac{\partial \mathcal{T}}{\partial x_1} & \frac{\partial \mathcal{T}}{\partial x_2} \end{bmatrix}_{\mathbf{x}'=\mathbf{0}} \\ \mathbf{N}_G^S &= \begin{bmatrix} \frac{\partial^2 \mathcal{T}}{\partial x_1 \partial x_1} & \frac{\partial^2 \mathcal{T}}{\partial x_1 \partial x_2} \\ \frac{\partial^2 \mathcal{T}}{\partial x_2 \partial x_1} & \frac{\partial^2 \mathcal{T}}{\partial x_2 \partial x_2} \end{bmatrix}_{\mathbf{x}'=\mathbf{0}} \, . \end{aligned} \tag{2.19}$$

Equation 2.18 shows that an elementary wave recorded at the horizontal surface can be locally interpolated using a parabolic traveltime moveout (e.g., Hoecht et al. (2009); Andrade et al. (2005)). Ursin (1982) shows that it is possible to further approximate the traveltime as

$$\mathcal{T}^2(\mathbf{x}') = (\mathcal{T}_0 + \mathbf{s} \cdot \mathbf{x}')^2 + \mathcal{T}_0 \mathbf{x}' \cdot \mathbf{N}_G^S \mathbf{x}' \, . \tag{2.20}$$

Equation 2.20 shows that an elementary wave recorded at the horizontal surface can be locally interpolated using a hyperbolic traveltime moveout (e.g., Zhang et al. (2001)). The hyperbolic approximation is closely related to the parabolic approximation, though it is generally more accurate for extrapolation at larger distance from the central ray (Ursin, 1982). Considering the expressions in equations 2.16, 2.17, 2.18, and 2.20, the pressure and particle velocity wavefields recorded at the surface can be locally described by a superposition of elementary waves whose traveltime moveouts are analytically given by the first and second derivatives of the traveltime. Furthermore, along the traveltime moveout of each elementary wave, the amplitude is constant and can be described by a geometrical spreading factor that depends on the traveltime curvature at the central ray (Schleicher et al., 1993). However, considering the high frequency approximation made in the ray theory, such a description is valid for smooth earth models only, i.e, models that are not changing rapidly with respect to the wavelength of the signal.

2.3.2 Predefined dictionaries

The basis vectors of the dictionary used to compute a sparse representation of the seismic data can be predefined using analytical functions. The functions that have been used for seismic data applications include monochromatic waves, wavelets (Mallat, 2008), curvelets (Candès and Donoho, 2000, 2002), and seislets (Fomel and Liu, 2010; Liu and Fomel, 2010). Figure 2.10 presents two-dimensional (2D) basis vectors defined using monochromatic waves, Haar wavelets, and curvelets.

I will now examine different analytical functions that have been used for seismic processing, and I will evaluate their effectiveness for sparse representations of the seismic data. For convenience, I will describe 2D basis functions, and assess their capability to represent seismic data in a time-space domain, where the temporal dimension is the traveltime, and the spatial dimension is the inline component of the source-receiver offset vector. From there, the conclusions will be extendable for 3D extensions of the basis functions and their capability to represent 3D seismic data where the additional dimension is the crossline component of the offset vector.

The Fourier basis, whose basis vectors represent monochromatic waves, is largely used to reconstruct the seismic data via a sparse inversion, e.g., Zwartjes (2005); Abma and Kabir (2005); Zwartjes and Sacchi (2007); Schonewille et al. (2009); Naghizadeh and Sacchi (2010); Gao et al. (2013). These studies show that if the original sampling is not too poor, the Fourier-based method can provide an accurate reconstruction of the seismic data. In addition, a monochromatic plane wave is the simplest solution to the wave equation in a homogeneous medium (see equation 2.15). In an inhomogeneous medium whose properties are changing smoothly in space, it is reasonable to assume that the wavefield is locally representable with few monochromatic plane waves. In addition, the forward and backward Fourier transformations are fast, which leads to an efficient computation of a sparse representation. Hence monochromatic waves may be suitable to compute a sparse representation of seismic data. Though, the stationary nature of monochromatic waves can be inconvenient to concisely represent seismic data. As stationary signals are not localized in the time-space domain, they have a limited effectiveness to concisely represent the seismic events that are localized in time.

The basis vectors of a 1D discrete wavelet transform (DWT) are analytically defined with dilates and translates of a function ϕ. Each basis vector is defined using a function $\phi_{a,b}(t) = 1/\sqrt{a}\phi((t - b)/a)$, where a sets the scale of the wavelet, and b sets its local position. In the basic DWT, i.e., the Haar DWT (Haar, 1910), ϕ is a "square-shaped" function. Considering how they are constructed, one can see that DWT basis vectors are well localized in time. Hence, DWT bases can concisely represent sharp signals, which provides advantages over Fourier bases for some analysis tasks in seismic processing (Foster et al., 1994). The DWT is a separable transform; the 2D functions that define the 2D DWT basis vectors is the direct product of two functions that define the 1D DWT basis

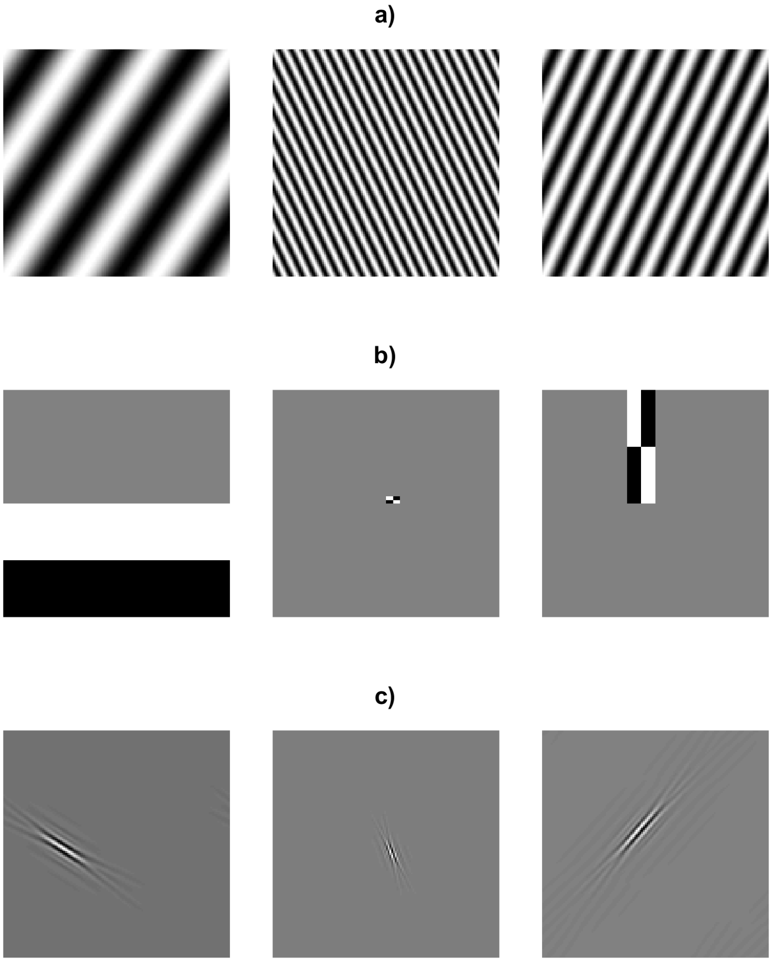

Figure 2.10: Real part of three (a) Fourier basis vectors, (b) Haar wavelet basis vectors, and (c) Curvelet basis vectors.

vectors ($\phi_{a,b,c,d}(t,x) = 1/\sqrt{ac}\phi((t-b)/a))\phi((x-d)/c)$). As the dimensions are identically treated, the wavelets are said to be isotropic, and they represent 2D patterns that are either horizontal, vertical, or diagonal. Consequently, the wavelets fail to concisely represent patterns that are curved (Candès and Demanet, 2005). As seismic events in a common shot gather can be significantly curved, wavelets may have limitations to concisely represent the seismic data.

The curvelets, an extension of the wavelets, are scaled, localized, and also directional at fine scales (see Figure 2.10). In the $f-k$ domain, a fine-scale curvelet is localized on a small oriented dyadic rectangle of length 2^j and width $2^{j/2}$, where j is the scale of the curvelet. In time, this curvelet is a narrow ridge of length $2^{-j/2}$ and width 2^{-j} pointing in a determined direction. The curvelets have been shown to be effective predefined functions to represent and denoise the seismic data (Hennenfent and Herrmann, 2006; Neelamani et al., 2008). In addition, a representation in a curvelet tight-frame of a solution satisfying the wave equation was proven to be sparse (Candès and Demanet, 2005).

An experiment was carried out to evaluate the effectiveness of the Fourier bases, the Haar wavelet bases, and the curvelet frames for sparse representation of the seismic data. The seismic data selected for this experiment is presented in Figure 2.11. It is a window of size 128 × 128 samples taken from an inline slice of a shot gather. A sparse approximation was computed using a Fourier base, a Haar wavelet base, and a curvelet tight-frame. Each sparse approximation was constrained to have an ℓ_0-norm equal to 15% of the number of samples in the original data. The sparse approximations in the Fourier and Haar wavelet bases were computed as follows: The forward transform operator was applied to the data, the 15% of the coefficients having the highest magnitudes were kept, the other coefficients were muted, and the backward transform operator was applied. The curvelet frame had 125,395 basis vectors and hence was quite redundant. The redundancy obliged a more complex process to compute the sparse approximation. First, the BP problem was solved to find a representation that had a small ℓ_1-norm, i.e, a representation that is sparse in the ℓ_1 sense. Then, the curvelets corresponding to the highest magnitude coefficients of this representation were selected as support to compute a representation that was sparse in the ℓ_0 sense. The nonzero coefficients of this sparse representation were computed using a least square inversion. The Fourier-, Haar wavelet- , and curvelet-based sparse representations are shown in Figure 2.12a-c and the residuals of the sparse representations are shown in Figure 2.13a-c. We observe a significant coherency in the residuals of the Haar wavelet-based sparse approximation. It indicates that the seismic data could not be approximated with only 15% of the Haar wavelet coefficients. In other words, the seismic data did not have a high degree of sparsity in the Haar wavelet domain. The Fourier- and curvelet-based sparse approximations appear to be more accurate, as less residuals are observed. The S/N was used to quantitatively assess the accuracy of the sparse approximations. The S/N was computed such that

$$\text{S/N} = 10\,\log_{10}\frac{\|\mathbf{d}_{\text{ref}}\|_2^2}{\|\mathbf{d}_{\text{ref}} - \mathbf{d}\|_2^2}\,, \tag{2.21}$$

where $\mathbf{d}$ is the approximation of the data and $\mathbf{d}_{\text{ref}}$ is the true data. The S/N values found for the Fourier-, Haar wavelet- , and curvelet-based sparse representations were 11.55 dB, 7.10 dB, and 14.40 dB, respectively. This concludes that the curvelet frame was more effective to compute a sparse

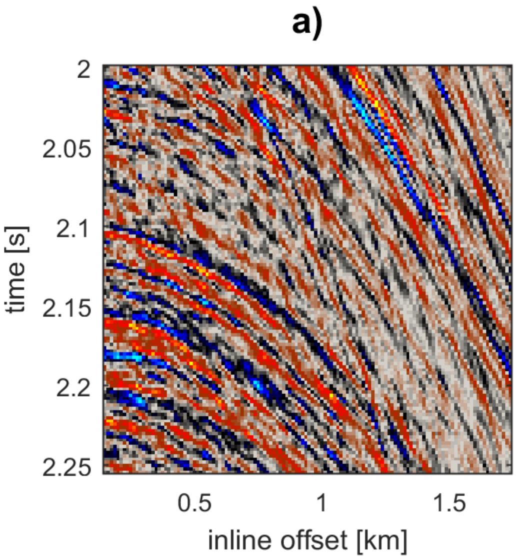

Figure 2.11: The pressure wavefield within a window of an inline shot gather.

representation of the data shown in Figure 2.11. Of course, one should keep in mind that the curvelet frame is redundant, which makes the comparison unfair. The number of curvelets and Fourier basis vectors used to compute the sparse approximations are the same, but the curvelets used are selected from a larger set of basis vectors. Finally, I note that for the three predefined dictionaries used, we can see some coherency in the residuals of the sparse approximation, which indicates that some signal could not be represented in a sparse manner with these dictionaries.

It is also possible to reconstruct the seismic wavefield using a sparse inversion in the parabolic (Herrmann et al., 2005), or hyperbolic (Ibrahim et al., 2015) Radon domain. The coefficients of the 2D parabolic and hyperbolic Radon domains are integrals of the data along lines defined by parabolic and hyperbolic functions, respectively. This suggests that the seismic wavefield can be represented in a sparse manner using parabolic or hyperbolic events. In addition, a wavefield recorded at the surface can be locally described by a sum of elementary waves whose traveltimes are analytically given by hyperbolic or parabolic moveouts, and whose amplitude is constant along these traveltime moveouts (see equations 2.16, 2.17, 2.18, and 2.20). Hence, basis vectors that represent hyperbolic or parabolic events are valid candidates to compute a sparse representation of the seismic data.

Using a predefined dictionary for sparse representation of a seismic data set leads to a global approach - the same dictionary is used regardless of the time or the offset of the data. Thus, the dictionary should be selected such that all events of the data are sparse in the transform domain. However, it is difficult to find a dictionary that can concisely describe all types of events present in the seismic data set, e.g., diffractions and reflections, shallow and deep events. In practice, the selection of the dictionary requires compromises and the dictionary is not optimal to represent complex events. Hence, the lack of adaptability that is inherent to predefined dictionaries limits their efficiency.

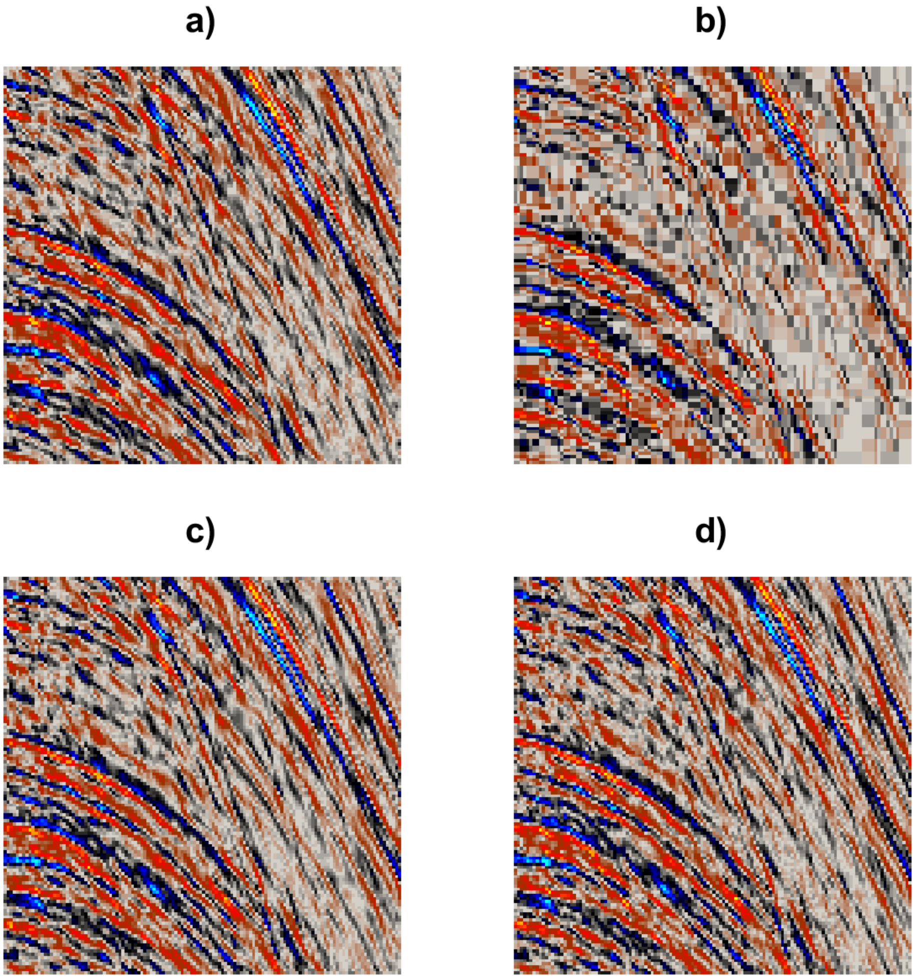

Figure 2.12: Sparse approximations of the data presented in Figure 2.11 using the a) Fourier base, b) Haar wavelet base, c) curvelet frame, and d) learned dictionary. For computing the four sparse approximations, the number of basis vectors used was set to 15% of the number of samples in the original data. The S/N of the results is as follows: Fourier base: 11.55 dB, Haar wavelet base: 7.10 dB, curvelet frame: 14.40 dB, and learned dictionary: 16.53 dB.

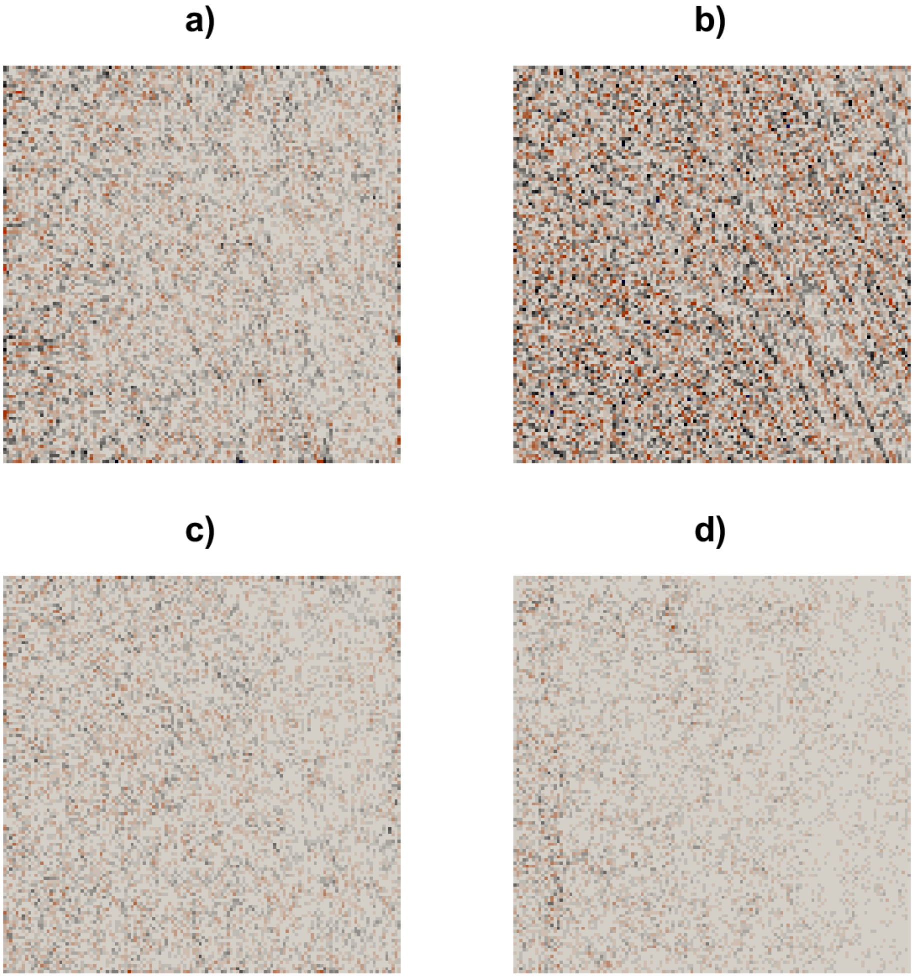

Figure 2.13: Residuals of the sparse approximations that were computed using the a) Fourier base, b) Haar wavelet base, c) curvelet frame, and d) learned dicitonary.

2.3.3 Learned dictionaries

Dictionary learning (DL) methods, e.g., the method of optimal direction (MOD) (Engan et al., 1999) or the k-times singular value decomposition (K-SVD) (Aharon et al., 2006), are alternatives to predefining the dictionary. In 2D applications, small-sized patches are extracted from a local area of the data and a dictionary is trained to optimally represent those patches. The resulting dictionary is optimal to find a sparse representation of the patches used for the training, or any data patches of similar morphology. For higher-dimensional applications, a similar procedure is used; in 3D, the training is carried out using extracted cubes, and in even higher dimensions, higher-dimensional vertices are used. The training consists of solving a sparse optimization problem. When DL is applied in 2D, M small-sized patches are extracted from the data and are vectorized to obtain a set of vectors $\mathbf{y}_1, ..., \mathbf{y}_M$ called the training set. The number of extracted patches, M, is selected to be several times larger that the desired number of dictionary atoms. Then, the DL problem generally consists of finding the dictionary $\mathbf{D} \in \mathbb{R}^{N\times K}$ and the set of sparse coefficient vectors $\mathbf{x}_1, ..., \mathbf{x}_M$ that minimize the representation error given a sparsity constraint T placed on the sparse coefficient vectors (Aharon et al., 2006). This problem is mathematically expressed as

$$\min_{\{\mathbf{x}_i\}_{i=1}^{M},\mathbf{D}} \sum_{i=1}^{M} ||\mathbf{y}_i - \mathbf{D}\mathbf{x}_i||_2^2 \text{ subject to } ||\mathbf{x}_i||_0 \leq T,\, i = 1, ..., M \, . \tag{2.22}$$

Although it is not necessarily said, the atoms of the dictionary are constrained to be normalized. There is no other constraints imposed to the dictionary in conventional DL.

The problem in equation 2.22 is very complex to solve because both the dictionary and the sparse representation vectors are variables of the problem. In practice, the optimization is carried out with a sub-optimal process that iteratively solve two simpler subproblems. The first subproblem focuses on the optimization of the sparse representations and is solved during a step called the sparse coding stage, whereas the second subproblem focuses on the optimization of the dictionary and is solved during the dictionary update stage. The kth iteration of this two-stage process can be summarized as follows:

- Sparse coding stage: The representation of each vector $\mathbf{y}_i$ of the training set is updated with

$$\hat{\mathbf{x}}_i = \arg\min_{\mathbf{x}} ||\mathbf{y}_i - \mathbf{D}^{k-1}\mathbf{x}||_2 \;\text{ subject to }\; ||\mathbf{x}||_0 \leq T \, . \tag{2.23}$$

- Dictionary update stage: The dictionary is updated with

$$\mathbf{D}^k = \arg\min_{\mathbf{D}} \sum_{i=1}^{M} ||\mathbf{y}_i - \mathbf{D}\hat{\mathbf{x}}_i||_2^2 \, . \tag{2.24}$$

The trained dictionary is not necessarily complete and is generally highly redundant. Due to the high degree of redundancy in the dictionary, the sparse approximations in the sparse coding stage need to be computed with a robust sparse solver such as OMP. This is computationally demanding.

The dictionary update stage is solved using the Moore-Penrose pseudoinverse in MOD, whereas it is solved using singular value decompositions (SVDs) in K-SVD. Solving iteratively the two stages is in general computationally demanding.

The last years have seen many developments of the DL methods to reduce their cost. Rubinstein et al. (2010) proposed a so-called double sparsity dictionary learning method. The learned atoms are constructed as sparse linear combinations of predefined atoms from a basis (i.e., $\mathbf{D} = \mathbf{X\Phi}$, where $\mathbf{X}$ is a sparse matrix and $\mathbf{\Phi}$ is a base). Such a construction provides efficient forward and adjoint operators, and results in a cheaper dictionary training. Cai et al. (2014) proposed a method called data-driven tight frame (DDTF), in which the learned dictionary is constrained to be a tight frame. The tight frame properties enable to carry out the sparse coding stage using hard thresholding instead of a matching pursuit type of algorithms, which speeds up DL.

DL methods have attracted a lot of interest in seismic processing. Using a learned dictionary was shown to be a better alternative than using a predefined dictionary for random noise attenuation or reconstruction of randomly missing traces (Beckouche and Ma, 2014; Liang et al., 2014; Yu et al., 2015; Zhu et al., 2015; Chen et al., 2016).

I will illustrate the effectiveness of DL using the data of size 128×128 samples presented in Figure 2.11. 10,000 possibly overlapping patches of size 8×8 were extracted from the data. They were arranged as vectors of size 64 samples. The 10,000 vectors were used to solve the DL problem in equation 2.22, where K and T were set to 1,000 and 9, respectively. The DL method used was K-SVD. 64 atoms of the learned dictionary are presented in Figure 2.14. The learned dictionary was used to compute a sparse approximation of the data. The data was decomposed into 256 patches of size 8×8. For each patch, OMP was used to solve the problem in equation 2.7 where T was set to 9. The sparse approximation of the patches were reassembled to form the sparse approximation of the data window. Since the threshold T was set to 9, and the number of samples in a patch was 64, the ℓ_0-norm of the sparse representation of the data is close to 15% of the number of data samples. The sparse approximation is presented in Figure 2.12d and the residuals are presented in Figure 2.13d. The absence of coherency in the residuals attests to the accuracy of the sparse approximation. The S/N of the result is 16.53 dB.

The drawback of DL is the lack of analytical expression for the dictionary. The learned atoms are defined at discrete positions and need to be physically stored. The direct consequence is that it requires a memory storage. This is a small problem because learned dictionaries are usually of low dimensions. An indirect consequence, which can be inconvenient, is the lack of analytical expression for the data representation. Since the atoms are only defined at discrete positions, a representation of the data as a linear combination of the atoms is also defined at discrete positions. This causes limitations for interpolation of a signal. The sparse representation cannot be used to interpolate the data over an arbitrary grid. In contrast, when the atoms are predefined with analytical functions, a sparse representation can be interpolated over a desired grid by representing the data as a linear combination of the atoms and taking the atoms at the desired positions using their analytical expression.

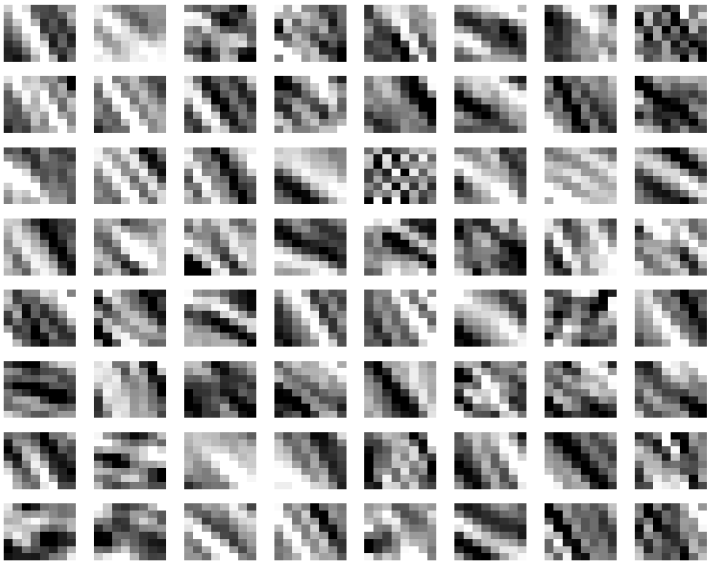

Figure 2.14: 64 atoms of the dictionary learned from the data presented in Figure 2.11.

Chapter 3

Main Scientific Contribution

3.1 Article I

Random noise is a long-standing problem in signal and image processing. It has been addressed by many studies and there is now a large range of random noise attenuation methods. However, a large part of them are designed to remove a noise of predictable energy. It is also the case for conventional DL-based denoising methods. They learn a dictionary from the data and approximate the data with an error-constrained sparse optimization process. Applications of such methods to seismic data are not optimal since the noise in the seismic data has an intensity that is often unknown and locally varying across the data.

Turquais et al. (2017c) modify the conventional DL problem to better suit the seismic data denoising problem. The authors propose to learn the dictionary and find the sparse approximation of the data using a coherence-constrained sparse optimization process instead of an error-constrained sparse optimization process. The coherence-based constraint imposes that the statistical measure of the coherence present in the removed noise is below a fix threshold. This threshold is derived to be optimal to filter out Gaussian noise and is independent of the variance of the noise. The proposed method is tested on seismic data and its effectiveness is assessed in comparison to a conventional DL method and a seismic industry-standard method. In contrast to the conventional DL method, the proposed method does not require any empirical testing and it adapts well to spatial or temporal variations in the noise variance. It also preserves better the fine structures of the signal and it removes more noise compared with both the conventional DL and the industry-standard methods.

3.2 Article II

In a towed-streamer seismic survey, the steering devices that are placed along the streamers, as well as barnacles that grow on the surface of the streamers, perturb the forward movement of the streamers in the water and lead to mechanical noise in the seismic data. The mechanical noise significantly hinders seismic processing and imaging if not or incorrectly removed. Yet, removing this noise is particularly challenging. The mechanical noise is too coherent to be properly handled with conventional random noise attenuation methods but not sufficiently predictable to be modeled and subtracted.

Turquais et al. (2017a) propose to separate the mechanical noise from the data by exploiting the morphological differences between the signal and the noise. The proposed method is a fully automatic process. First, DL is applied to the data to get a dictionary that contains some atoms describing the signal morphology and some atoms describing the noise morphology. Then, the atoms describing the noise morphology are separated from the atoms describing the signal morphology with a statistical classification. This divides the dictionary into noise and signal subdictionaries. Finally, the data is approximated with a sparse constraint in the combined noise and signal subdictionaries domain. This separates the noise from the signal as the noise cannot be sparsely represented in the signal subdictionary domain, and is therefore represented in the noise subdictionary domain, and vice versa for the signal. This noise suppression method was tested on two seismic data examples and was assessed in comparison to four industry-standard or state-of-the-art denoising methods. All examples considered, the method provided the highest quality results.

3.3 Article III

In 3D towed-streamer marine seismic surveys, the seismic wavefield is poorly sampled in the crossline direction. The crossline sampling is not sufficiently dense to meet the requirements of several seismic processing and imaging methods. Hence, the recorded data needs to be interpolated over a denser grid in an early stage of the processing. This task is very challenging and no satisfactory solution is yet found. DL methods were shown to be highly effective to reconstruct randomly missing traces in seismic data. These methods learn morphological features from the available data and fill the gaps left by missing traces using learned morphological features that match the neighboring traces. Yet, conventional DL methods cannot interpolate uniformly sampled data over a denser grid because there is no example in the data that can be used to learn densely sampled morphological features.

Turquais et al. (2017f) overpass the limitation of conventional DL methods by imposing a structure to the dictionary atoms. Each atom is constrained to represent an elementary waveform that has a constant amplitude along a parabolic traveltime moveout characterized by kinematic wavefield parameters. Among other advantages, this parabolic structure offers the possibility to easily interpolate the atoms over an arbitrary sampling grid. Once the dictionary is learned, a sparse representation of the data in the dictionary domain is computed, the atoms of the dictionary are interpolated over the desired grid, and the sparse representation of the data is taken in the interpolated dictionary domain, which interpolates the data. Three characteristics of this method, i.e., the parabolic structure, the sparsity promotion, and the adaptation to the data, strengthen robustness to noise and to aliasing and they increase the accuracy of the interpolation. Synthetic and field data examples show that the method reconstructs well the seismic wavefield across the streamers of typical 3D acquisitions. This indicates that the proposed method is reliable and could be applied in an early stage of the seismic processing sequence. This would improve the later 3D processing and imaging steps, e.g., wavefield separation, multiple removal, and migration, and it would enhance the final image.

Chapter 4

Article I

The first article is entitled "A method of combining coherence-constrained sparse coding and dictionary learning for denoising". It was published in the journal Geophysics. A manuscript was sent to the editor the 30th of March 2016, a revised manuscript was sent the 31st of October 2016, and it was published online the 27th of February 2017. The layout has been changed from the official publication to better fit the format of the thesis. The page number located in the header of each page of the article except the first one is relative to the article and is starting from A102, whereas the page number relative to the thesis is located in the footer of the page.

A method of combining coherence-constrained sparse coding and dictionary learning for denoising

Pierre Turquais[1,2], *Endrias G. Asgedom*[1], *Walter Söllner*[1]

ABSTRACT

We have addressed the seismic data denoising problem, in which the noise is random and has an unknown spatiotemporally varying variance. In seismic data processing, random noise is often attenuated using transform-based methods. The success of these methods in denoising depends on the ability of the transform to efficiently describe the signal features in the data. Fixed transforms (e.g., wavelets, curvelets) do not adapt to the data and might fail to efficiently describe complex morphologies in the seismic data. Alternatively, dictionary learning methods adapt to the local morphology of the data and provide state-of-the-art denoising results. However, conventional denoising by dictionary learning requires a priori information on the noise variance, and it encounters difficulties when applied for denoising seismic data in which the noise variance is varying in space or time. Here, we propose a coherence-constrained dictionary learning (CDL) method for denoising that does not require any a priori information related to the signal or noise. To denoise a given window of a seismic section using CDL, overlapping small 2D patches are extracted and a dictionary of patch-size signals is trained to learn the elementary features embedded in the seismic signal. For each patch, using the learned dictionary, a sparse optimization problem is solved, and a sparse approximation of the patch is computed to attenuate the random noise. Unlike conventional dictionary learning, the sparsity of the approximation is constrained based on coherence such that it does not need a priori noise variance or signal sparsity information and is still optimal to filter out Gaussian random noise. The denoising performance of the CDL method is validated using synthetic and field data examples, and is compared with the K-SVD and FX-Decon denoising. We found that CDL gives better denoising results than K-SVD and FX-Decon for removing noise when the variance varies in space or time.

INTRODUCTION

Raw seismic data are often contaminated with random noise over the entire time and frequency band. This noise obscures details and hinders seismic imaging from revealing the real subsurface structures. Attenuating such noise is well-known to be a long-standing problem (Yilmaz, 2001). Over the past decade, however, sparse and redundant representations have received a lot of attention in signal and image processing for analyzing the information in data sets and providing state-of-the-art results for compression, interpolation, and denoising (Elad, 2010). Based on the observation that the relevant information about the physical process that causes our recording is of low dimensionality (Tosic and Frossard, 2011), sparse and redundant representations may be used to express the relevant information as a linear combination of few elementary signals called atoms stored in a redundant set known as the dictionary.

[1] Petroleum Geo-Services ASA, Oslo, Norway

[2] University of Oslo, Department of Geosciences, Oslo, Norway

The atoms of a given dictionary can be predefined assuming that the signal in the data follows a given analytical model. For example, the atoms can be analytically defined as Fourier, wavelet (Mallat, 2008), curvelet (Candès and Donoho, 2000, 2002) or seislet (Fomel and Liu, 2010; Liu and Fomel, 2010) basis vectors. Alternatively, it is possible to avoid any assumption about the morphology of the signal and learn a redundant dictionary from the data with a dictionary learning (DL) method. These methods train a dictionary to be optimally adapted for representing the signal in a sparse manner. The DL methods include method of optimal direction (MOD) (Engan et al., 1999), k-means singular value decomposition (K-SVD) (Aharon et al., 2006), and data-driven tight frame (DDTF) method (Cai et al., 2014). Otherwise, a double sparsity dictionary learning method (Rubinstein et al., 2010) reconciles the fix and adaptive approaches by learning a dictionary as a sparse linear combination of a predefined dictionary.

When sparse representations are used for denoising, the recording is sparsely approximated with the part of the recording that correlates best with atoms of the dictionary. Therefore, the quality of denoising depends on the ability of the dictionary to efficiently describe the signal. Practically, Fourier-transform-based seismic denoising often performs poorly because Fourier basis functions fail to represent localized seismic events in a sparse manner. On the contrary, wavelet basis functions are well localized and therefore they can describe seismic data in a sparser manner and provide better denoising than the non-space methods (Foster et al., 1994). Curvelets can efficiently model the geometry of waveforms (Candès and Demanet, 2005), and they have proven to be some of the most suitable predefined functions to represent and denoise the seismic data (Hennenfent and Herrmann, 2006; Neelamani et al., 2008). However, using a dictionary that is constructed from a predefined transform leads to a global approach for seismic data representation - a single dictionary is used to represent all the seismic features. This lack of adaptability to the local morphology can make predefined dictionaries inefficient to represent some of the complex features in seismic data. Thus, attempting to perform denoising using predefined dictionaries might result in distorted signal output. This is why, using a learned dictionary for denoising seismic data has proven to be a better alternative than using a fixed dictionary (Beckouche and Ma, 2014). Training a redundant dictionary has a high computational cost, which results in expensive denoising methods. To reduce this cost, DDTF trains a tight frame instead of a general redundant dictionary because tight frames benefit from simpler decomposition and recomposition schemes (Liang et al., 2014; Yu et al., 2015, 2016). Also, double sparsity dictionary learning methods benefit from taking a data-driven approach and integrating prior information about the signal morphology into the problem (Zhu et al., 2015; Chen et al., 2016).

When denoising seismic data by sparse approximation subject to a fixed or learned dictionary, the key parameter is the constraint on the sparsity of the approximation because it controls the amount of energy that is removed from the data. For now, this constraint is either fixed or dictated by the variance of the noise. A constraint dictated by the variance of the noise can be optimal only for filtering noise whose variance is known a priori and constant over the data (Donoho and Johnstone, 1994). If the variance of the noise is uncertain or varying over the data, one needs to compute the sparse approximation with a variance parameter that compromises between signal losses and remaining noise because signal loss may occur when the true variance is locally lower and noise may remain in the data when the true variance is locally higher.

In this work, we implement a sparse approximation method in which the sparsity is constrained by a statistical measure of the coherence present in the removed noise. We derive a coherence threshold that is independent of the variance of the noise and is ideal for filtering out Gaussian noise. We further integrate such a sparse approximation to the K-SVD DL scheme and build a coherence-constrained dictionary learning (CDL) method, which is adapted for attenuating random noise of unknown spatiotemporally varying variance.

The rest of this paper is organized as follows: First, we formulate and evaluate the CDL denoising algorithm. Then, we assess the denoising performance of CDL in comparison with conventional K-SVD, and FX-Decon (Canales, 1984; Gulunay, 1986) methods using synthetic data contaminated with random noise having a constant variance and having a spatiotemporally varying variance. Finally, we apply CDL on a field data section to validate its noise attenuation and signal preservation capability.

METHOD

DL algorithms contain a step that performs sparse approximation. It is in this step that CDL method differs from classic DL methods, by using a coherence-constrained sparse approximation. Therefore, this section first presents the coherence-constrained sparse approximation problem and then it describes the CDL algorithm.

The ideal coherence-constrained sparse approximation

Consider a recording $\mathbf{z} \in \mathbb{R}^N$ containing signal of interest $\mathbf{y} \in \mathbb{R}^N$ and white Gaussian noise $\mathbf{n} \in \mathbb{R}^N$ of zero-mean and σ^2 variance. Formally, this data model is given by

$$\mathbf{z} = \mathbf{y} + \mathbf{n} \, . \tag{1}$$

The recording $\mathbf{z}$ can be expressed in another domain via matrix multiplication with a dictionary. This dictionary is a matrix containing an atom, i.e., a unit vector of length N, in each of its columns ($\mathbf{D} = [\mathbf{a}_1 \, \mathbf{a}_2 \, ... \, \mathbf{a}_K] \in \mathbb{R}^{N \times K}$) and it is chosen such that the signal of interest is sparse in the dictionary domain. Therefore, there exists a sparse vector $\mathbf{x} \in \mathbb{R}^K$, containing a small number L, of nonzero coefficients such that

$$\mathbf{y} = \mathbf{D}\mathbf{x} \, . \tag{2}$$

In sparse optimization problems, the number L of nonzero coefficients in the solution $\mathbf{x}$ is referred to as the cardinality. For attenuating the noise, the recording is sparsely approximated in the dictionary domain. To do so, it is popular to place a constraint on the representation error and aim for the solution that minimizes the ℓ_0-norm (Donoho et al., 2006). This problem can be formally expressed as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_0 \text{ subject to } ||\mathbf{z} - \mathbf{D}\mathbf{x}||_2 \leq \epsilon \, , \tag{3}$$

where the representation error threshold ϵ is dictated by the noise variance σ^2. After solving this problem, the sparse approximation of the recording is computed by

$$\hat{\mathbf{y}} = \mathbf{D}\hat{\mathbf{x}}. \tag{4}$$

However, if the standard deviation of the noise is unknown but the cardinality L of the solution is known, one can interchange the constraint and the objective function (Tropp, 2004). That is, switching the problem in equation 3 to the problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{z} - \mathbf{D}\mathbf{x}||_2 \text{ subject to } ||\mathbf{x}||_0 \leq T \, , \tag{5}$$

where the threshold T is dictated by the cardinality L of the solution.

In this work, we want to tackle the problem in which the cardinality of the solution and the noise variance are unknown. In this case, it is possible to use a method known as coherent denoising (Mallat, 2008, p. 656-659). In coherent denoising, the sparsity is constrained by the coherence of the residual vector $\mathbf{r}$ relative to the dictionary $\mathbf{D}$. This coherence measure is denoted by $\mu(\mathbf{r}, \mathbf{D})$ and mathematically given by

$$\mu(\mathbf{r}, \mathbf{D}) = \max_j \left| \frac{\mathbf{r}^T}{||\mathbf{r}||_2} \mathbf{a}_j \right| \,, \tag{6}$$

where $\mathbf{a}_j$ is one of the unit vectors of the dictionary as mentioned earlier. The residual vector is the difference between the recording and the sparse approximation (i.e., $\mathbf{r} = \mathbf{z} - \mathbf{D}\hat{\mathbf{x}}$). Hence, coherent denoising aims to solve the problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_0 \;\text{ subject to }\; \mu(\mathbf{z} - \mathbf{D}\mathbf{x}, \mathbf{D}) \leq \bar{\mu} \,, \tag{7}$$

where $\bar{\mu}$ is a fixed coherence threshold. Finding an approximation of a recording, which uses a minimal number of atoms from a redundant dictionary and satisfies $\mu(\mathbf{z} - \mathbf{D}\mathbf{x}, \mathbf{D}) \leq \bar{\mu}$ is a problem called "Non-deterministic Polynomial-time (NP)-hard". The NP-hardness is a class of problems whose complexity cannot be expressed as a polynomial function of its input size, but rather as an exponential function. Therefore, as problems in equations 3 and 5, the problem in equation 7 is not tractable for realistic seismic data sizes. However, an approximate solution can be obtained using coherent matching pursuit (Mallat, 2008, p. 656-659).

The coherent matching pursuit algorithm uses the iterative greedy scheme of orthogonal matching pursuit (OMP)(Pati et al., 1993), but has a different stopping criterion. The coherent matching pursuit algorithm selects the atom having the highest correlation with the current residual vector, then it updates the coefficient vector by error minimization, and finally updates the residual vector, at each iteration. The iterative process continues until the coherence $\mu(\mathbf{r}, \mathbf{D})$ reduces to become less than the threshold $\bar{\mu}$ and results in an approximate solution to the problem in equation 7.

Donoho and Johnstone (1994) show that the threshold $\sigma\sqrt{2\log(N)}$ is ideal for denoising a signal of length N contaminated by Gaussian noise of variance σ^2 when used with thresholding over a wavelet base. Following the same track, we show in Appendix B that the threshold

$$\mu_{Ideal} = \sqrt{2\frac{\log(K)}{N}} \tag{8}$$

is ideal for filtering out white Gaussian noise by coherent denoising subject to a redundant dictionary $\mathbf{D} \in \mathbb{R}^{N\times K}$. This threshold has the additional advantage of being independent of the noise variance. Consequently, using this threshold for coherent denoising provides a safe and optimal noise attenuation process and does not require testing of any parameters.

We designed an experiment to evaluate the denoising performances of a sparse approximation that is constrained with the proposed coherence threshold in comparison with the error-constrained and the cardinality-constrained sparse approximations. We first synthesized a signal vector following the model established in equation 2. Here, the dictionary $\mathbf{D}$ was a matrix of size 100×100 and its atoms were zero-mean unit vectors with identically and independently distributed Gaussian entries. The sparse coefficient vector $\mathbf{x}$ contained L nonzero coefficients. The position of the nonzero coefficients was randomly chosen and their values were fixed to the same value, α. Then, a recording $\mathbf{z}$ was synthesized by adding random noise $\mathbf{n}$ with variance σ^2 as in equation 1. We will use $\hat{\mathbf{y}}_{err}$ and $\hat{\mathbf{y}}_{car}$ to refer to the sparse approximations obtained with OMP for the error- and cardinality-constrained problems in equations 3 and 5, in which the thresholds are fixed such that $\epsilon = \sqrt{N}\sigma$ and $T = L$, respectively. Furthermore, $\hat{\mathbf{y}}_{coh}$ will refer to the coherent matching pursuit approximate solution to the problem in equation 5, for the threshold μ_{Ideal}. The sparse approximations $\hat{\mathbf{y}}_{coh}$, $\hat{\mathbf{y}}_{err}$, and $\hat{\mathbf{y}}_{car}$ of the recording $\mathbf{z}$ were computed to recover the signal of interest. Because $\mathbf{n}$ is known in this experiment, the denoising capability of a sparse approximation was measured with the mean-squared error relative to the initial noise given by

$$E = \frac{||\mathbf{y} - \hat{\mathbf{y}}||_2^2}{||\mathbf{n}||_2^2} \,. \tag{9}$$

The quantity E assesses the error reduction; the closer it is to 0, the greater is the error reduction. Note that E is independent of the amplitude of the recording, unlike the mean-squared error.

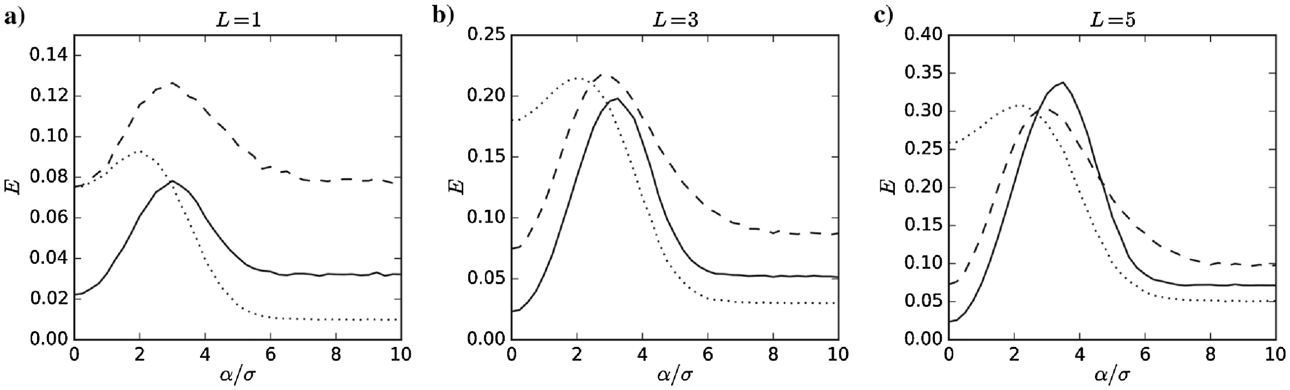

Figure 1: Denoising capability of the coherence-constrained sparse approximation $\hat{\mathbf{y}}_{coh}$ (solid curve), the error-constrained sparse approximation $\hat{\mathbf{y}}_{err}$ (dashed curve), and the cardinality-constrained sparse approximation $\hat{\mathbf{y}}_{car}$ (dotted curve) when the cardinality L of the solution is fixed at 1, 3, and 5, as indicated above each plot. The error E is presented as a function of the ratio α/σ .

The results of this experiment are presented in Figure 1 for α/σ ranging from zero to 10 and the cardinality value L, which was fixed at 1, 3, and 5, as indicated above each plot. Each E value is computed as the average of 10,000 trials of a Monte Carlo experiment. Trials of a Monte Carlo experiment are repetitions of the experiment in which the random parameters are reselected. For small values of ratio α/σ, $\hat{\mathbf{y}}_{coh}$ provides the lowest E, and for larger values, $\hat{\mathbf{y}}_{car}$ performs better. However, E does not differ significantly for the three sparse approximations. Hence, this experiment demonstrates that $\hat{\mathbf{y}}_{coh}$ can perform denoising similar to that of $\hat{\mathbf{y}}_{car}$ or $\hat{\mathbf{y}}_{err}$ without the knowledge of the cardinality of the solution or the noise variance. For the range of parameters studied, E of $\hat{\mathbf{y}}_{coh}$ varies between 0.02 and 0.33. This means that the mean squared error of $\hat{\mathbf{y}}_{coh}$ is at least 3 times smaller and up to 50 times smaller than the mean squared error of the recording. Hence, these E values prove the noise attenuation capabilities of $\hat{\mathbf{y}}_{coh}$.

Denoising by coherence-constrained DL

The sparse approximation problem presented in the subsection ”The ideal coherence-constrained sparse approximation” can be applied to each recording of data independently. For seismic data, the recording could be a trace, a window of a gather, or the entire gather, as long as the dictionary is chosen accordingly. However, DL algorithms are designed to be applied on a set containing a large number M of recording vectors, $\mathbf{z}_1$, $\mathbf{z}_2$, ..., $\mathbf{z}_M$, having the same length N. For seismic data applications, the recording vectors correspond generally to small 2D patches from a gather that have been vectorized. Here, for studying the problem, we consider a set in which the signal and noise content of each recording are given by the models in equations 1 and 2, where the dictionary $\mathbf{D}$ is the same for all the recordings. Then, for i=1, 2, ..., M, the recording $\mathbf{z}_i$ contains noise and signal components. The noise component, $\mathbf{n}_i$, is white Gaussian noise of mean zero and variance σ_i^2 and the signal component $\mathbf{y}_i$ can be expressed as a linear combination of a small number L_i of atoms of the dictionary.

Denoising by DL consists in finding a dictionary $\hat{\mathbf{D}}$ and a set of sparse coefficient vectors $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2$, ..., $\hat{\mathbf{x}}_M$ such that the sparse approximations

$$\hat{\mathbf{y}}_i = \hat{\mathbf{D}}\hat{\mathbf{x}}_i \; , \tag{10}$$

for $i = 1, ..., M$, recover the signal of interest present in the data set $\mathbf{z}_1$, $\mathbf{z}_2$, ..., $\mathbf{z}_M$.

For a data set contaminated by noise of constant variance, i.e., $\sigma_i^2 = \sigma^2$ for $i = 1, 2, ..., M$, Elad and Aharon (2006) propose to solve the problem

$$(\{\hat{\mathbf{x}}_i\}_{i=1}^M, \hat{\mathbf{D}}) = \underset{\{\mathbf{x}_i\}_{i=1}^M, \mathbf{D}}{\arg\min} \sum_{i=1}^{M} ||\mathbf{x}_i||_0 \text{ subject to } ||\mathbf{z}_i - \mathbf{D}\mathbf{x}_i||_2 \leq \epsilon \; , \; i = 1, 2, ..., M \; . \tag{11}$$

This problem consists of finding the dictionary and sparse coefficients that minimize the sparsity of the representation with the constraint that the representation error of each recording should be below the threshold ϵ. The threshold ϵ is dictated by the variance σ of the noise. For instance, a threshold $\epsilon = \sqrt{N}\sigma$ would ensure to not remove more energy from the recordings than the energy coming from the noise.

Alternatively, if the cardinality of the solution is constant, i.e., $L_i = L$ for $i = 1, 2, ..., M$, Aharon et al. (2006) propose to learn the dictionary and sparse coefficients by solving the problem

$$(\{\hat{\mathbf{x}}_i\}_{i=1}^M, \hat{\mathbf{D}}) = \underset{\{\mathbf{x}_i\}_{i=1}^M, \mathbf{D}}{\arg\min} \sum_{i=1}^{M} ||\mathbf{z}_i - \mathbf{D}\mathbf{x}_i||_2^2 \text{ subject to } ||\mathbf{x}_i||_0 \leq T \; , \; i = 1, 2, ..., M \; . \tag{12}$$

In this problem, the representation error is minimized with the constraint that each representation of recordings should have a sparsity below the threshold T that is dictated by the cardinality L of the true solution.

Here, we address the problem in which the cardinality of the solution and the noise variance are varying over the data set. In this case, none of the solutions in equations 11 and 12 are optimal for denoising. Indeed, one could solve the problem in equation 11 with a fixed error threshold, but the solution would represent noise of recordings in which the norm of the noise is above the error threshold and might not preserve the signal of the recordings in which the norm of the noise is below the error threshold. Similarly, one could solve the problem in equation 12, but the solution would distort the signal of the recordings in which the cardinality of the true solution is above the threshold and represent noise of the recordings in which the cardinality of the true solution is below the threshold. To overcome these problems, we propose to solve

$$(\{\hat{\mathbf{x}}_i\}_{i=1}^M, \hat{\mathbf{D}}) = \underset{\{\mathbf{x}_i\}_{i=1}^M, \mathbf{D}}{\arg\min} \sum_{i=1}^M ||\mathbf{x}_i||_0 \text{ subject to } \begin{cases} \min_{\mathbf{D}} \sum_{i=1}^M ||\mathbf{z}_i - \mathbf{D}\mathbf{x}_i||_2^2 \\ \mu(\mathbf{z}_i - \mathbf{D}\mathbf{x}_i, \mathbf{D}) \leq \lambda\sqrt{2\frac{\log(K)}{N}} \ , i = 1, 2, ..., M \ . \end{cases} \tag{13}$$

Here, the sparsity of the representation is minimized and subject to two constraints, (1) the dictionary should be computed to minimize the representation error, and (2) the residuals of the representation should have coherence below the threshold $\lambda\sqrt{2\frac{\log(K)}{N}}$. The gain factor λ controls the strength of the denoising. However, one should set the gain factor λ to one for proper signal preservation (cf. Appendix B).

The problem in equation 13 is highly underdetermined and cannot be solved exactly. Similarly to conventional DL problems (Engan et al., 1999; Aharon et al., 2006), it is approximated with the iterative two step process summarized as follows

I **Sparse coding step:** For each recording $\mathbf{z}_i$, $i = 1, ..., M$, use the coherent matching pursuit algorithm to solve the problem in equation 7 and find the sparse coefficient vector $\hat{\mathbf{x}}_i$. In this problem, the coherence threshold is set to $\lambda\sqrt{2\log(K)/N}$ and the dictionary used is the one found at the dictionary update step of the previous iteration.

II **Dictionary update step:** Use the K-SVD dictionary update step (Aharon et al., 2006) to find a new dictionary as the solution to the problem

$$\hat{\mathbf{D}} = \min_{\mathbf{D}} \sum_{i=1}^M ||\mathbf{z}_i - \mathbf{D}\hat{\mathbf{x}}_i||_2^2 \ , \tag{14}$$

where $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2$, ..., $\hat{\mathbf{x}}_M$ are the sparse coefficient vectors found in step 1.

For the first iteration, the dictionary is initialized with K normalized recordings randomly chosen from the data set.

For step 2, CDL borrows the dictionary update process from the K-SVD algorithm. The K-SVD dictionary update process has been chosen over the MOD dictionary update process because the K-SVD algorithm converges quicker to a solution (Aharon et al., 2006). The CDL algorithm is detailed in Appendix A.

As proposed by Elad (2010), p. 227-238, we designed an experiment to assess the dictionary recovery capability of CDL in comparison with the conventional K-SVD algorithm. The comparison is done for both cases in which K-SVD is used to solve the error-constrained DL problem stated in equation 11 (denoted K-SVD$_{err}$) and the cardinality-constrained DL problem stated in equation 12 (denoted K-SVD$_{car}$). This experiment is described in Figure 2. A dictionary of size 100×100 (N=100, K=100) was synthesized and used to construct a set of 8000 recording vectors (M=8000) following the models described in equations 1 and 2. The cardinality L of the solution and the standard deviation σ of the noise were constant over the data set in order that K-SVD$_{car}$ and K-SVD$_{err}$ could be used correctly. The L nonzero coefficients were chosen uniformly at random between five and 10 times the standard deviation of the noise. The constructed data set was given as input to CDL, K-SVD$_{car}$, and K-SVD$_{err}$. The thresholds used by K-SVD$_{err}$ and K-SVD$_{car}$ were $\epsilon = \sqrt{N}\sigma$ and $T = L$, respectively. The CDL method was used with the gain factor λ set to one. The dictionaries output

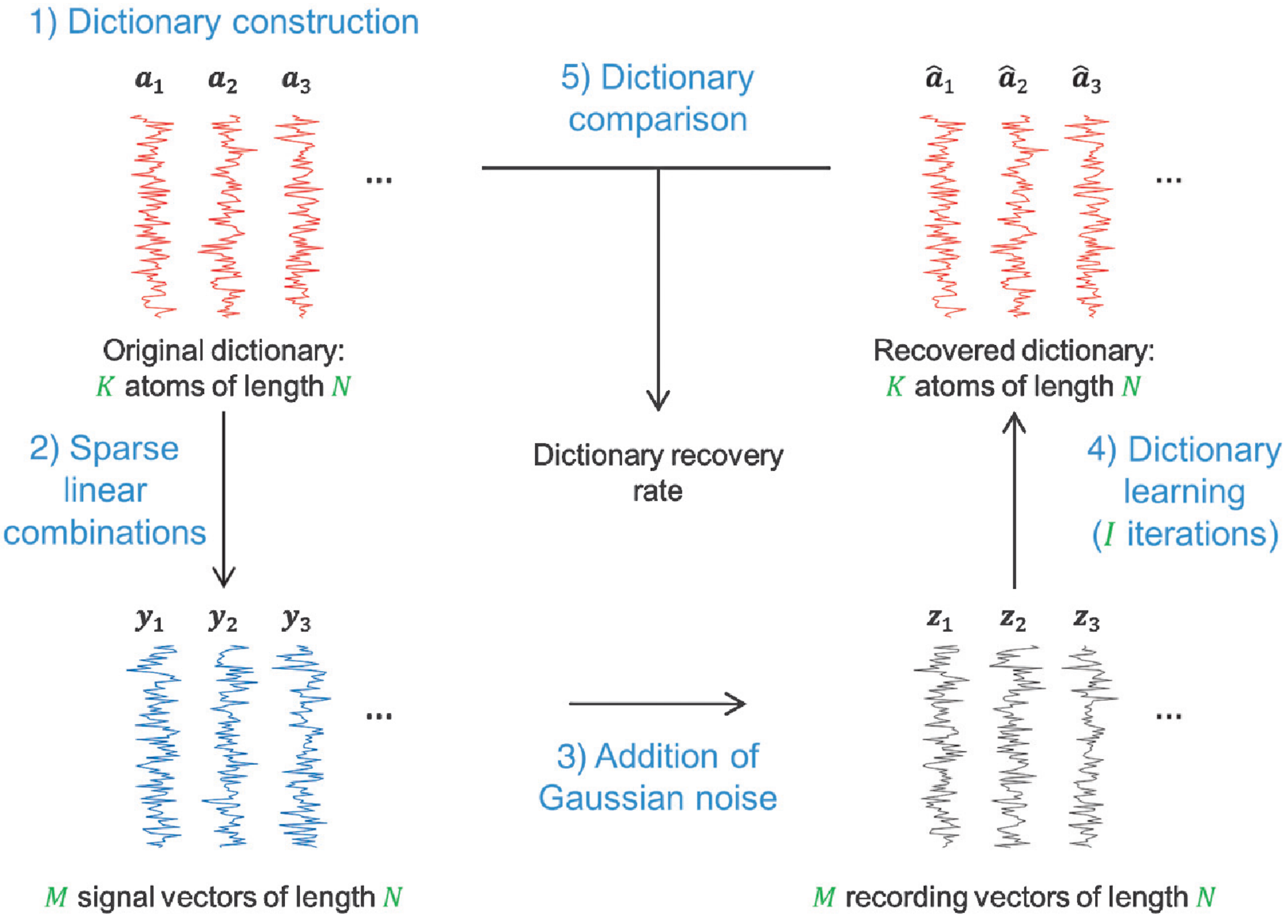
1) Dictionary construction
a_1 a_2 a_3 ...
Original dictionary:
K atoms of length N
5) Dictionary comparison
Dictionary recovery rate
$\hat{a}_1$ $\hat{a}_2$ $\hat{a}_3$ ...
Recovered dictionary:
K atoms of length N
2) Sparse linear combinations
4) Dictionary learning (I iterations)
y_1 y_2 y_3 ...
3) Addition of Gaussian noise
z_1 z_2 z_3 ...
M signal vectors of length N
M recording vectors of length N

Figure 2: An experiment that assesses the dictionary recovery capability of a DL algorithm. (1) A dictionary is generated such that its atoms are zero-mean unit vectors with identically and independently distributed Gaussian entries. (2) Each signal vector is constructed as a linear combination of L randomly chosen atoms from the dictionary. (3) White Gaussian noise of mean zero and variance σ^2 is added to the signal vectors to synthesize the recording vectors. (4) The DL algorithm is applied to learn a dictionary. (5) The original and learned dictionaries are compared to compute the dictionary recovery rate (i.e., the percentage of recovered atoms).

by the DL algorithms were compared with the dictionary that was first synthesized. The dictionary recovery capability is assessed with the percentage of recovered atoms where an atom is considered as recovered if the correlation between the atom from the true dictionary and an atom from the learned dictionary is higher than the value of 0.99.

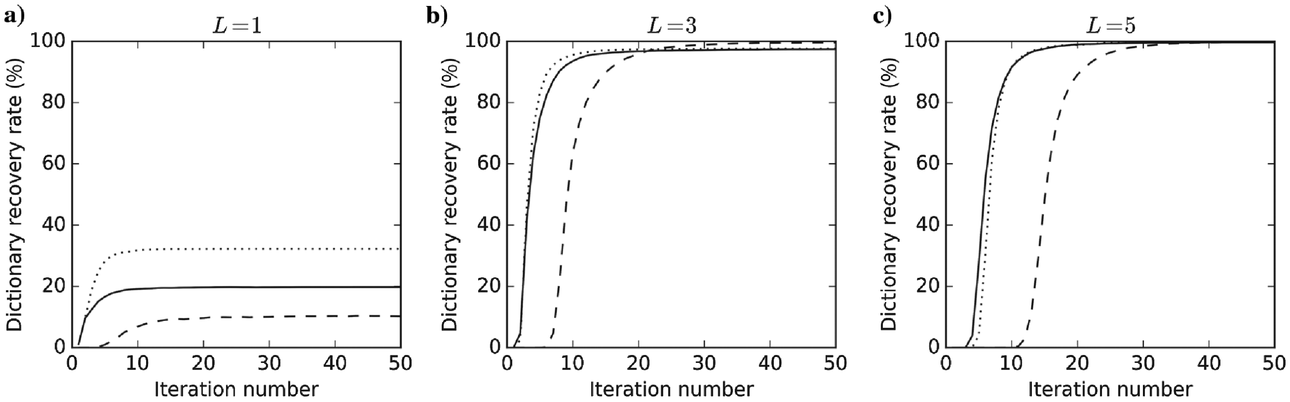

Figure 3: Dictionary recovery capability of CDL (solid curve), K-SVD$_{car}$ (dotted curve), and K-SVD$_{err}$ (dashed curve) when the cardinality of the solution L is fixed at 1, 3, and 5, as indicated above each plot. The dictionary recovery rate is presented versus the number of iterations of the DL process.

In Figure 3, the recovery rate for CDL, K-SVD$_{car}$, and K-SVD$_{err}$ are displayed at each iteration for $L = 1$, 3, and 5, as indicated above each plot. The recovery rate values are computed with 100 trials of Monte Carlo experiments. For $L = 1$, neither of the three algorithms is able to recover the dictionary, but, for $L = 3$ and 5, the algorithms recover the dictionary. An explanation for this behavior can be, as L increases, the signal-to-noise ratio (S/N) of the recordings increases, and the performance of the dictionary update step increases. For a cardinality that is equal to or higher than three, K-SVD$_{car}$ and CDL have similar dictionary recovery rates. For all tested cardinality values, K-SVD$_{err}$ needed more iterations to converge than K-SVD$_{car}$ and CDL did. However, for a cardinality equal to 3, K-SVD$_{err}$ reached a slightly higher recovery rate after convergence. Given the parameters used, after 25 iterations, the three algorithms have nearly converged. For the same parameters, if we increase the number of recordings or the S/N, the dictionary is more easily recovered by any of the three algorithms. Finally, we note that CDL performs DL similar to K-SVD without knowledge of the noise variance or the cardinality of the solution.

SYNTHETIC DATA APPLICATION

In this section, we first compare the performances of the CDL method with that of the FX-Decon and conventional K-SVD methods for a classic problem - removing Gaussian noise of constant variance from a 2D seismic data set. Then, we test the capability of CDL, FX-Decon, and K-SVD methods to adapt to the challenging and under-explored problem in which the variance of the additive noise is spatiotemporally varying.

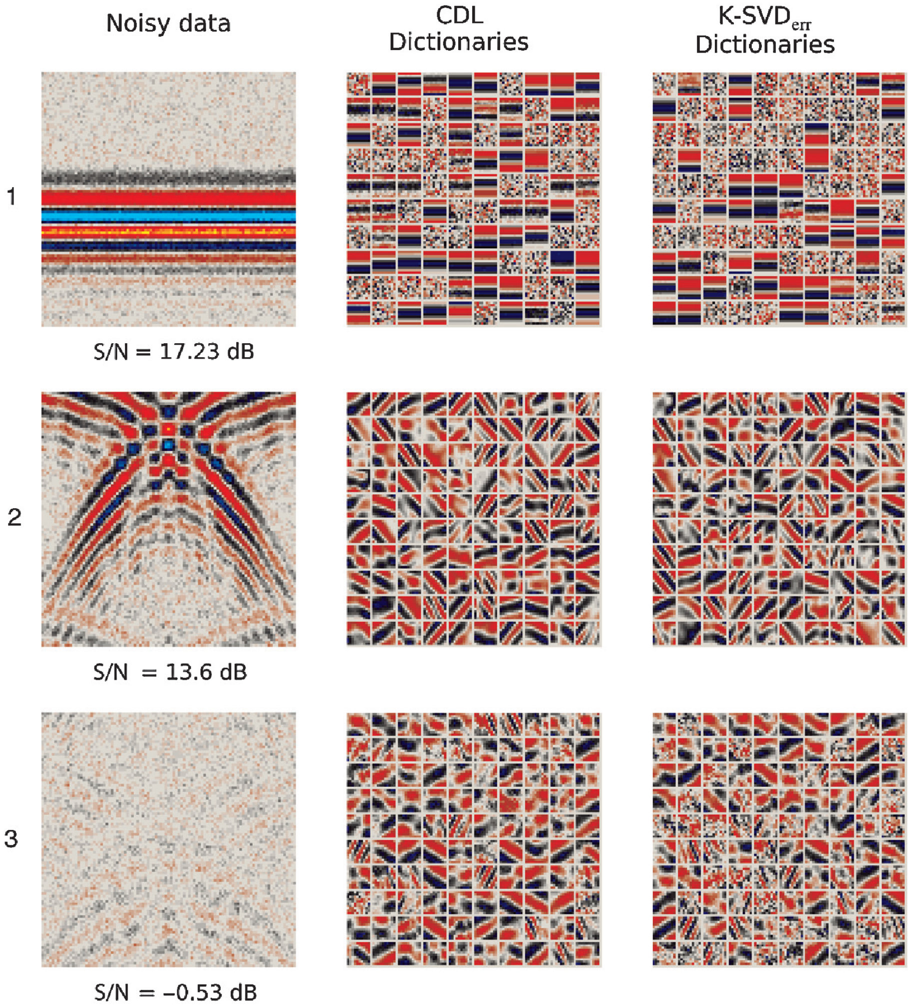

Figure 4: CDL and K-SVD$_{err}$ that were applied on three windows of synthetic data for learning dictionaries.

Noise with constant variance

A noise-free synthetic 2D data set was generated using finite difference for an earth model consisting of plane and syncline reflectors. The data were acquired for sources and receivers placed at every 10 m and with a time sampling increment of 4 ms. Then, we added zero-mean white Gaussian noise to the signal to obtain the noisy data. The method is applied on the zero offset gather. The common-offset domain (in this case, zero offset) has been selected to apply CDL because DL exploits redundancy of the features over the data set, and we expect higher redundancy to be present in common-offset domain because most of the events are flat. This section is divided into windows of size 100×100, which are individually denoised. We selected three windows of the noisy data for the study (see Figure 4, first column). In window 1, the seismic signal is a linear flat event with strong amplitude compared with the noise. In window 2, the signal is composed of nonlinear seismic events, and in window 3, the data have a poor S/N.

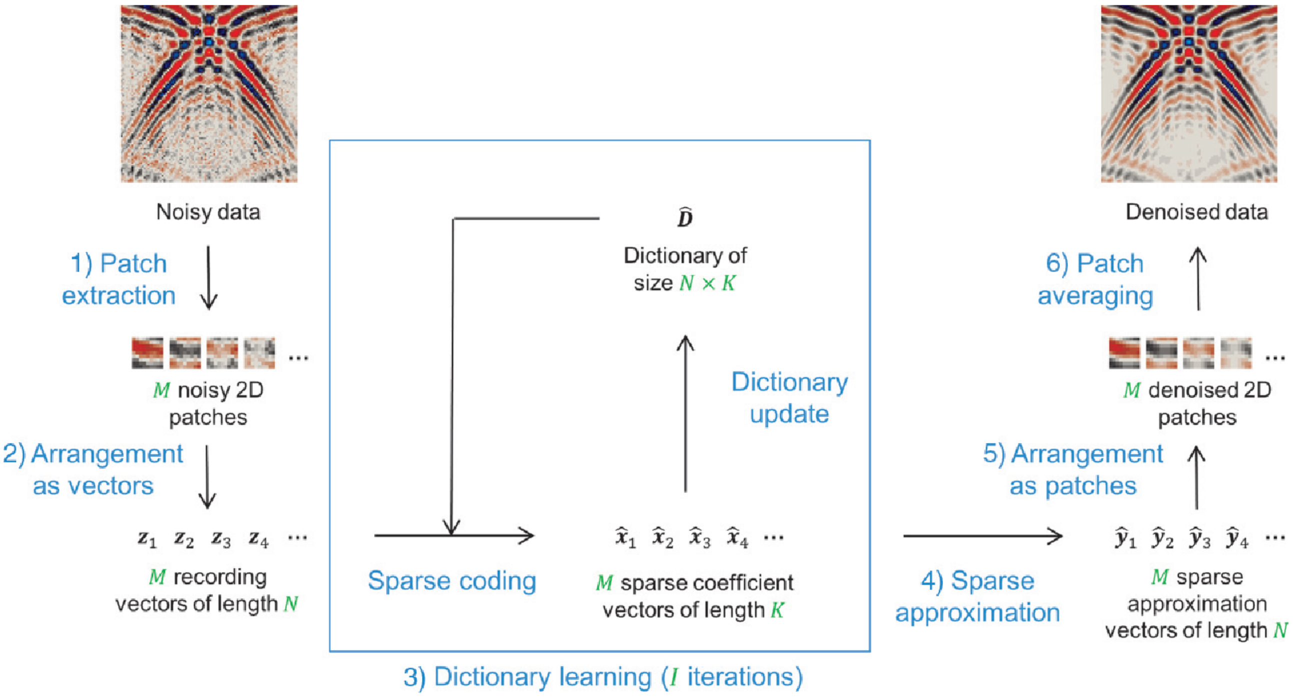
Noisy data
1) Patch extraction
M noisy 2D patches
2) Arrangement as vectors
z_1 z_2 z_3 z_4 ⋯
M recording vectors of length N
$\hat{D}$
Dictionary of size $N \times K$
Dictionary update
Sparse coding
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$ $\hat{x}_4$ ⋯
M sparse coefficient vectors of length K
3) Dictionary learning (I iterations)
4) Sparse approximation
$\hat{y}_1$ $\hat{y}_2$ $\hat{y}_3$ $\hat{y}_4$ ⋯
M sparse approximation vectors of length N
5) Arrangement as patches
M denoised 2D patches
6) Patch averaging
Denoised data

Figure 5: Flow-diagram of the CDL denoising process that is applied on a 2D seismic data window. (1) Fully overlapping 2D patches are extracted from the data window. (2) The patches are rearranged as column vectors. (3) The CDL algorithm detailed in Appendix A is applied to the extracted data set to learn a dictionary and find a sparse coefficient vector for each recording. (4) The sparse approximation vectors are computed by matrix multiplication of the learned dictionary with the sparse coefficient vectors. (5) The sparse approximation vectors are rearranged as 2D patches. (6) The patches are assembled considering their original positions and are averaged to obtain the seismic signal of the data window.

The workflow for denoising a given 2D data window with CDL is presented in Figure 5. All overlapping patches of size 10×10 ($N = 100$) are extracted from the window. In other words, the shift between the extracted patches is one sample in both dimensions. Hence, $M = 91 \times 91 = 8281$. The patches are rearranged into vectors of length 100 to generate the recordings, $\mathbf{z}_1$, ..., $\mathbf{z}_{8281}$. The

CDL algorithm is applied to the data set with 25 iterations ($I = 25$), a gain factor λ set to one, and a number of dictionary atoms fixed at 100 ($K = 100$). Each sparse coefficient vector, $\hat{\mathbf{x}}_i$, and the dictionary output from the algorithm are multiplied to generate the sparse approximation $\hat{\mathbf{y}}_i$ of the recording (see equation 10). Each sparse approximation vector $\hat{\mathbf{y}}_i$ is later rearranged as a 10×10 patch. Because the patches extracted from the noisy image were overlapping, there are several versions of the same samples in the set of denoised patches. Therefore, when assembling the denoised patches to retrieve a signal window, the multiple versions of the same samples are averaged. This averaging further attenuates the noise. The number M of patches in the data set and the number I of iterations of the algorithm have been chosen to guaranty optimal filtering according to the studies presented in Figure 3. Indeed, the results presented in Figure 3 show that the studied DL algorithms are able to recover the dictionary after 25 iterations when the extracted data set contains 8000 recordings. The size of the patches and the number K of atoms in the dictionary have been chosen empirically as a good compromise between quality of denoising and tractability of the algorithm, but how to chose these parameters is still an open question in the DL field.

We compared the CDL method with the conventional K-SVD and FX-Decon denoising methods. For this problem, we chose to apply K-SVD$_{err}$ and not K-SVD$_{car}$. Indeed, because we introduced noise of constant variance, all of the recordings present in the extracted data set contain noise having the same variance. Thus, considering as known this variance value, K-SVD$_{err}$ can be applied in optimal condition. On the other hand, because the seismic signal within a patch is more or less complex depending on where it has been extracted, it requires more or fewer features to be reconstructed; i.e, L is not constant over the data set, and K-SVD$_{car}$ would encounter difficulties. The K-SVD$_{err}$ method was applied with an error threshold ϵ set to $\sqrt{N}\sigma$, where σ is the standard deviation of the additive Gaussian noise. The rest of the parameters were the same as those used for applying CDL. FX-Decon was applied on windows of size 50×50 with 50% of overlapping in both dimensions and a filter of length 6 samples. These parameters have been selected because they have been shown to give the best denoising results on another example (Chen et al., 2016).

The dictionaries learned with CDL and K-SVD are presented in the second and third columns of Figure 4. For each learned dictionary, its 100 atoms are pictured as 10×10 patches in 10 lines of 10 atoms. We can observe the atoms represent redundant features (i.e., features present in many data patches) from the windows in which they have been learned. Because redundant features are the most efficient to sparsely represent the entire window, this attests to a successful DL for both algorithms. The results after CDL, K-SVD$_{err}$, and FX-Decon denoising are presented in Figure 6a, in the second, third, and fourth column, respectively. The denoising performance is quantitatively assessed via the S/N computed before and after the noise attenuation. For a given data $\hat{\mathbf{d}}$ and its noise-free reference $\mathbf{d}_{\text{ref}}$, the S/N expressed in decibels is defined as

$$S/N = 10\log_{10}\frac{\|\mathbf{d}_{\text{ref}}\|_2^2}{\left\|\mathbf{d}_{\text{ref}} - \hat{\mathbf{d}}\right\|_2^2}\,. \tag{15}$$

The S/N values displayed under each noisy and denoised window show that CDL is performing similar to K-SVD$_{err}$ and outperforms FX-Decon. We computed the difference between the recovered and the true signal to visually verify if some signal has been removed. These error windows are displayed in Figure 6b. For both CDL and K-SVD$_{err}$ methods, we see no significant coherence in the error windows. However, we can observe signal in the FX-Decon error window 2. This shows that FX-

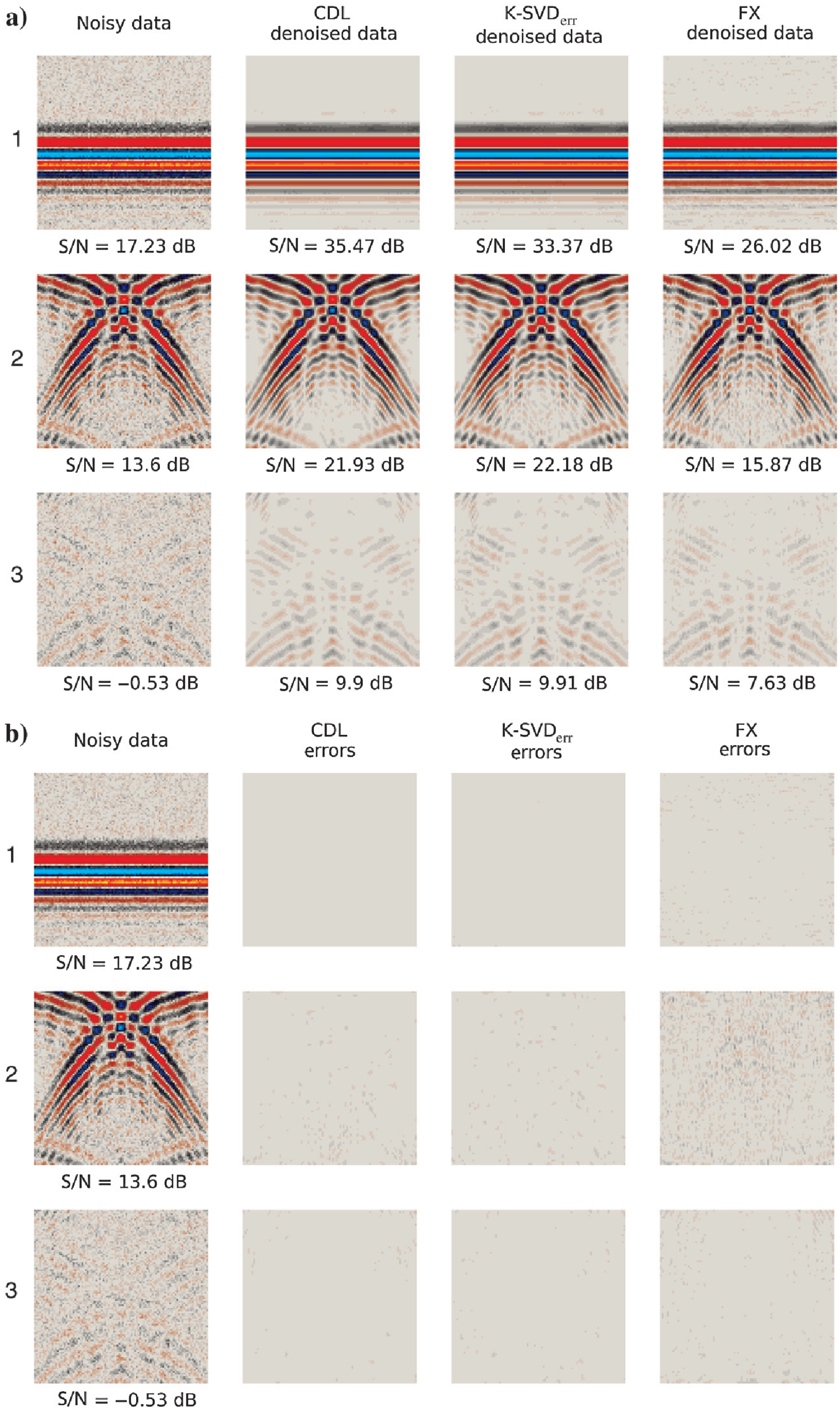

Figure 6: CDL, K-SVD$_{err}$, and FX-Decon denoising that were applied on three windows of synthetic data for attenuating random Gaussian noise of constant variance. (a) Denoised data and (b) error.

Decon denoising does not entirely preserve the signal when the input section is very complex. This signal distortion can be reduced by decreasing the window size, but this would increase the amount of remaining noise.

To summarize the results of this study, CDL performs better than FX-Decon and similarly to K-SVD in denoising seismic data contaminated by Gaussian noise of constant variance. However, in contrary to K-SVD, CDL does not require knowledge of the noise variance.

Noise with varying variance

We reproduced the experiment presented in the previous section but with random noise having a spatiotemporally varying variance. To create a noise window $\mathbf{N}_{var}$ having a spatio-temporally varying variance, we generated a matrix of size 100×100 containing zero-mean white Gaussian noise $\mathbf{N}_g$, and modulated its amplitude such that

$$\mathbf{N}_{var} = (\mathbf{W} + \mathbf{T}) \circ \mathbf{N}_g \ , \tag{16}$$

where $\circ$ denotes the element-wise multiplication. The matrices **W** and **T** are of size 100×100 and are defined such that **W** contains a 2D cosine signal varying in amplitude between 0.5 and 1.5 with a wavelength of 63 samples in both dimensions and **T** has 15% of its columns filled with the value one and the rest with the value zero. The three data windows obtained from the addition of the noise $\mathbf{N}_{var}$ to the signal are shown in Figure 7a, in the first column.

CDL denoising was applied to the three windows and compared with the K-SVD$_{car}$, K-SVD$_{err}$, and FX-Decon denoising methods. The parameters for applying the different methods were the same as the ones used during the previous experiment (see the section ”Noise with constant variance”). The K-SVD$_{car}$ methods was applied with a cardinality threshold T fixed at four, whereas K-SVD$_{err}$ was applied with an error threshold ϵ fixed at $\sqrt{N}\sigma$, where σ is the standard deviation computed on the noise $\mathbf{N}_{var}$. The learned dictionaries are not shown here. They are, however, similar to the ones learned during the previous experiment in which the data were contaminated with noise of constant variance. The windows after CDL, K-SVD$_{car}$, K-SVD$_{err}$, and FX-Decon denoising are presented in Figure 7a, as indicated above the columns. The S/N of each denoised window is displayed under it. In addition, the error windows are presented in Figure 7b.

No remaining noise is visible on the CDL-denoised windows. Moreover, the absence of signal and noise in the error windows attests that noise is highly attenuated and signal is preserved. Examining the CDL-denoised results presented in Figure 6a and comparing them to the ones presented in Figure 7a shows that CDL performs similar denoising for noise having constant variance and noise having spatiotemporally varying variance. Therefore, CDL is not affected by the variations of the noise variance.

In the top of the K-SVD$_{car}$ denoised window 1, some remaining noise is observed. In this area, there is initially very low signal, and therefore, some of the four atoms are used to reconstruct noise. Similarly, we can observe remaining noise in the K-SVD$_{err}$ denoised windows. Where the remaining noise is observed, the norm of the initial noise is locally higher than the error threshold, and the method consequently represents some noise. Therefore, the K-SVD$_{car}$ and K-SVD$_{err}$ results show that variations in the signal complexity or noise variance reduce the denoising performances of the K-SVD denoising method.

As we can observe from the denoising results, FX-Decon denoising is not affected by the vari-

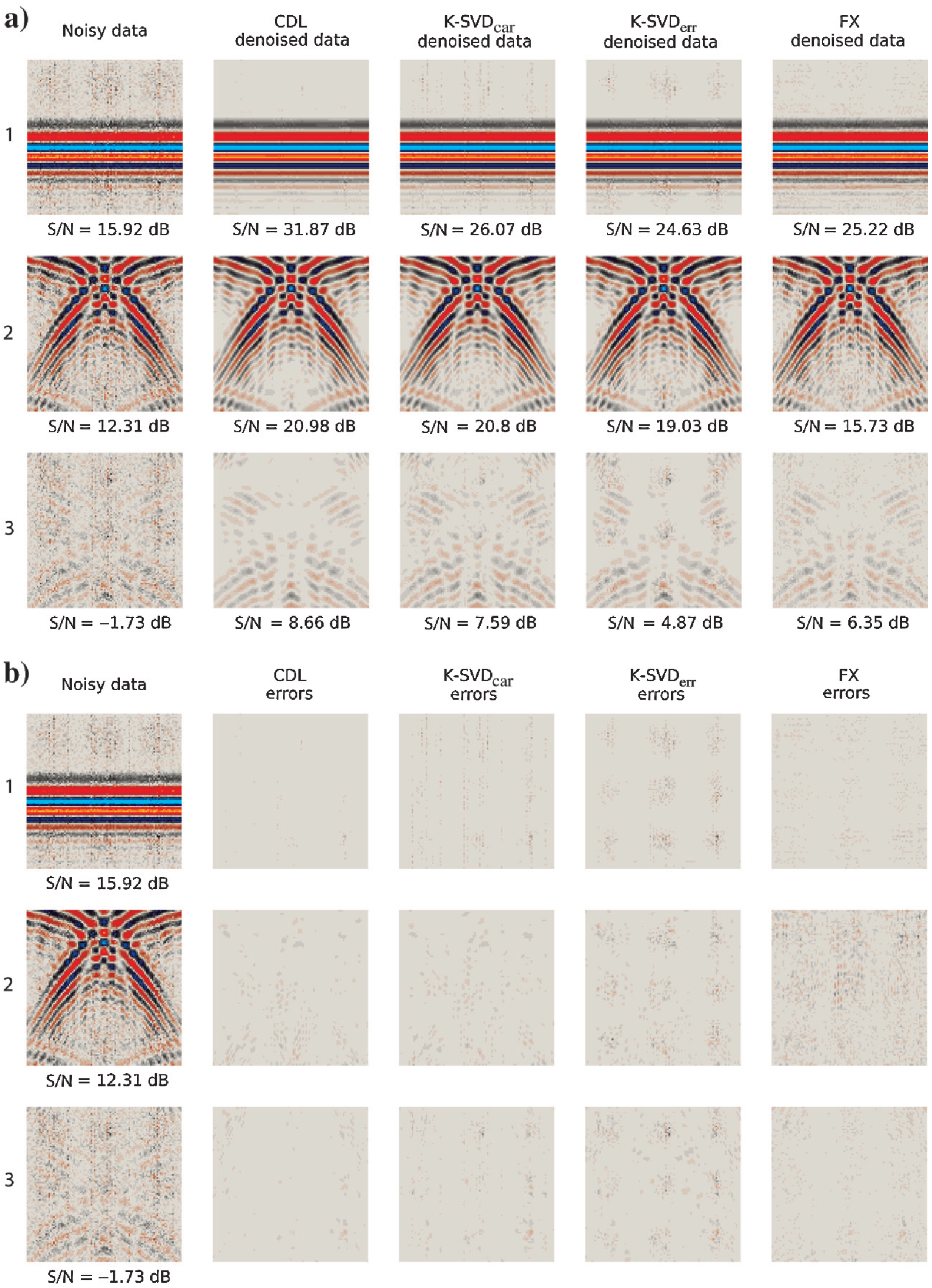

Figure 7: CDL, K-SVD$_{car}$, K-SVD$_{err}$, and FX-Decon denoising that were applied to three windows of synthetic data for attenuating spatiotemporally varying noise. (a) Denoised data and (b) error.

ations in the noise variance. However, it performs poorly in terms of noise attenuation and fails to preserve the complex signal present in window 2. Finally, the S/N values show that CDL attains higher S/N enhancement compared with K-SVD$_{err}$, K-SVD$_{car}$, and FX-Decon.

FIELD DATA APPLICATION

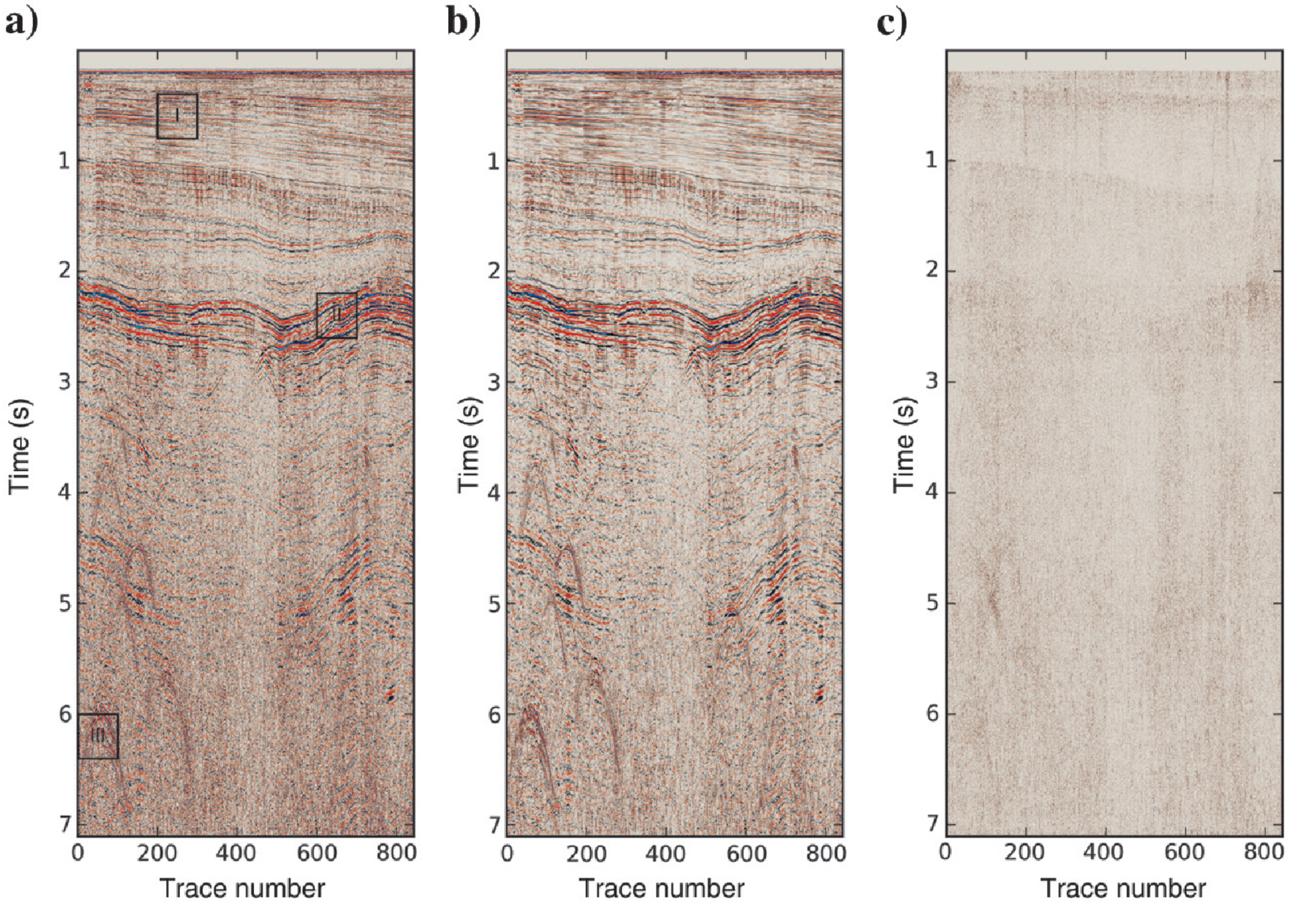

Figure 8: CDL that was applied on a common offset gather of a field data set for attenuating random noise. (a) Input data, (b) denoised data, and (c) removed noise.

A common-offset section of marine data (see Figure 8a) was selected to validate the capability of CDL to attenuate random noise present in field data sets while preserving the underlying seismic signal. The CDL method was applied on windows of size 100×100, which were overlapping on 15 samples in both dimensions. We used the same parameters as the ones selected for the study on seismic synthetics. The resulting denoised and removed-noise sections are shown in Figure 8b, and 8c, respectively.

In the removed noise, we observe a high variation in the variance. This shows that the random noise in field data has spatiotemporally varying characteristics. The absence of significant coherence in the removed noise shows that CDL can attenuate such noise while preserving the signal. We selected three windows of the section for a detailed analysis which is presented in Figure 9. Windows 1, 2, and 3 contain high-frequency flat events with poor S/N, linear high amplitude dipping events,

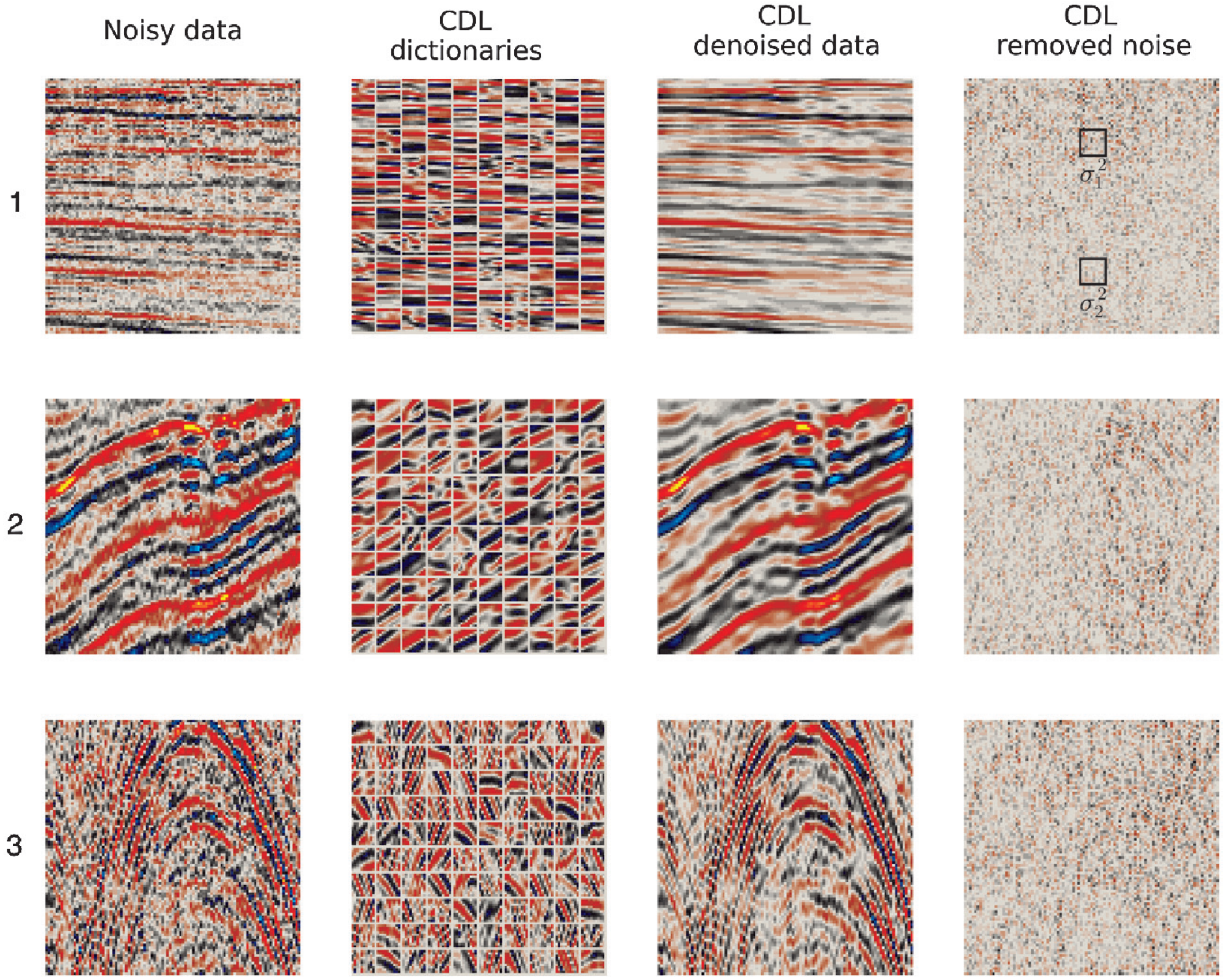

Figure 9: CDL that was applied on three windows of field data. These windows are chosen from the common offset gather presented in Figure 8a. In the removed noise window 1, two frame boxes point out two patches in which variance has been computed. The variance of the higher patch, σ_1^2, is 3.6 times larger than the variance of the lower patch, σ_2^2.

and diffractions, respectively. The 100 atoms of the learned dictionaries are presented as patches in the second column. We note that the learned atoms are free of noise. On window 1, essentially atoms representing flat linear features have been learned. However, many dips are present in the dictionary learned on the diffractions in window 3. This shows that in a dictionary trained on a data window, only the features necessary to efficiently represent the signal are captured, and it attests to an accurate DL. From the denoised windows presented in the third column, we can observe that random noise has been highly attenuated. The removed noise windows, presented in the last column, show that no significant coherent signal has been removed from the data. It is also true for the diffractions in window 3, which are especially challenging to preserve while denoising. On the removed noise window 1, we computed the variance of two 10×10 size patches pointed out with frame boxes. The variance σ_1^2 of the higher patch is 3.6 times larger than the variance σ_2^2 of the lower patch. This shows that the variance of the noise in field data is highly variable, even within a 100×100 size window, and it justifies the need for a method such as CDL that does not depend on a fixed variance parameter.

DISCUSSION

In this work, we proposed to modify the sparsity constraint of the sparse approximation used in DL-based denoising methods. This does not affect the computational complexity of DL algorithms. Here, as in the conventional K-SVD denoising method, CDL denoising requires O($NKLI$) operations per pixel, where N is the number of samples in a patch, K is the number of atoms in the dictionary, L is the number of nonzero elements in each coefficient vector, and I is the number of iterations of the algorithm (Elad and Aharon, 2006). CDL is, however, faster than K-SVD for the examples presented in this paper. For instance, filtering window 1, which is presented in Figure 6a took 101.7s for CDL and 201.2s for K-SVD on a laptop having a CORE i7vpro CPU. The CDL denoising was faster because it provided a solution with smaller L when no signal was present in a patch. Moreover, the tractability of K-SVD can be very affected by the variation of the noise variance when the problem in equation 11 is solved. For instance, denoising window 1 with K-SVD$_{err}$ was about three times longer when additive noise had varying variance (see Figure 7a) compared to when noise had constant variance (see Figure 6a). The run time of K-SVD$_{err}$ is large when removing noise with varying variance because it provides a solution in which L is large when the norm of the noise is locally higher than the error threshold. For the same example, the tractability of CDL denoising was not affected by the variation of the noise variance.

In this work, we chose the parameters of the filtering to achieve an optimal DL and denoising according to the studies performed on 1D synthetics (see Figures 3). In practice, the parameters can be modified to decrease the run time significantly but without significantly affecting the quality of the denoising results. For instance, for the same window 1, if the dictionary is learned on 20% of the data set with 5 iterations of the CDL process, and we use this dictionary to compute the sparse approximation of the complete data set, then we obtain a denoised data window having an S/N value of 35.03 dB in 9.7s. In addition, the algorithm used is a straightforward implementation of the algorithm presented in Appendix A, and therefore, is not optimized. Using the optimizations proposed by Rubinstein et al. (2008) speeds up the DL algorithms by a factor of 27 for the presented examples.

In the results presented in Figure 6a, we observe that FX-Decon is less effective in noise removal compared to the DL methods. But, it is much faster. For instance, filtering window 1 took 0.07s. The algorithm used is from SeisLab and it corresponds to the implementation proposed by Ulrych and Sacchi (2005), p. 229-232.

CDL denoising could be easily extended to higher dimensions. For instance, for 3D seismic data, it would consist in constructing the data set of recordings by extracting small 3D cubes instead of 2D patches. Then, as for 2D data, denoising of the data set would be performed with the CDL algorithm, which is presented in Appendix A. The 3D CDL denoising method would most probably perform better than the 2D CDL denoising method because it would benefit from the 3D coherency of the seismic wavefield. The complexity of the 3D CDL denoising process would be the same as the complexity of 2D CDL denoising process.

For denoising field data, Beckouche and Ma (2014) propose to estimate the variance of the noise with Median Absolute Deviation (MAD) and apply an error-constrained-DL-based denoising method. First, such a method does not adapt to variation of the noise variance within the window. Second, the MAD of the noise that is mixed with the signal is very often higher than the MAD of the noise alone because the signal is not sparse in time. Therefore, a method that is using MAD of the noise mixed with the signal very often overestimates the variance of the noise. For the field data section that is presented in Figure 8a, using MAD tends to overestimate the noise variance and would lead to substantial signal loss.

For the presented results, CDL has always been used with a gain factor set to 1 to optimize the signal preservation while attenuating the random noise. However, one could increase the gain factor to filter out noise that is slightly coherent in space and time.

In this work, we proposed to change the constraint of the DL problem, which only concerns the sparse coding step of the DL process. We used the K-SVD dictionary update step because the K-SVD algorithm has established itself as the standard DL algorithm. However, the proposed sparse coding step could be implemented using more up-to-date and more efficient algorithms, for instance, the sparse K-SVD algorithm (Rubinstein et al., 2010). The resulting denoising method could integrate a priori information about the seismic wavefield morphology and be locally adaptive to the data and to the noise variance.

CONCLUSION

Conventional DL methods are not adapted for denoising seismic data contaminated by noise with spatiotemporally varying variance because they are constrained with fixed error or cardinality thresholds. We proposed a DL method for denoising, which is constrained with a coherence measure. This method, referred to as CDL, can adapt to data in which the signal complexity and the noise variance vary in space and time. Furthermore, we derived a coherence threshold for CDL that is optimal for filtering out noise, which is locally white and Gaussian while preserving signal. Using synthetic data, we compared CDL denoising to K-SVD and FX-Decon denoising for two noise contamination cases, noise with constant variance and noise with spatiotemporally varying variance. We observed that CDL method performs similar to K-SVD for removing Gaussian noise with constant variance and has the advantage that it does not require the knowledge of the variance. Moreover, the CDL method provides better denoising results than K-SVD when the variance of the noise is spatio-temporally varying. For both cases in which the noise variance is constant and spatiotemporally varying, the CDL method outperforms FX-Decon denoising. Finally, on a field data example, we observed that noise recorded during seismic acquisition has a spatiotemporally varying variance, and that the proposed CDL method can attenuate such noise while preserving the signal.

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to the Petroleum Geo-Services (PGS) G&E team of Oslo for the fruitful discussions on the subject. We would also like to thank PGS for providing the data and giving the permission to publish this work. Finally, we wish to thank the associate editor N. Gulunay, and the reviewers R. Neelamani, A. Özbek, and J. Ma, for their valuable and constructive comments.

APPENDIX A
THE CDL ALGORITHM

The CDL algorithm described in Algorithm 1 takes a matrix $\mathbf{Z}$ containing a recording vector in each of its columns as input, and returns a dictionary $\mathbf{D}$ and a sparse coefficient matrix $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_M]$, which are approximate solutions of the problem presented in equation 13. In Algorithm 1, brackets have been used to refer to an index of a vector or a matrix; for instance, $\mathbf{x}[i]$ is the ith sample of the vector $\mathbf{x}$ and $\mathbf{D}[i, j]$ is the sample at the ith line and jth column of the matrix $\mathbf{D}$. In addition, columns inside the brackets are used to refer to all the indexes in a dimension; for instance, $\mathbf{D}[i, :]$ is the ith line of the matrix $\mathbf{D}$. The notation $\mathbf{D}^+$ has been used to denote the More-Penrose generalized inverse (Penrose, 1955) of a matrix $\mathbf{D}$. The function svd($\mathbf{D}$) applies the SVD decomposition of the matrix $\mathbf{D}$ and returns the matrices of eigenvectors $\mathbf{U}$ and $\mathbf{V}$, and of eigenvalues Δ, such that $\mathbf{D} = \mathbf{U}\Delta\mathbf{V}^T$. The symbol ← stands for assignment of the object at the right of the arrow into the location at the left of the arrow. We used {a,b} to refer to a set containing the objects a and b.

Algorithm 1 CDL

```
Input: Matrix of recordings $\mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ ...\ \mathbf{z}_M] \in \mathbb{R}^{N \times M}$
Parameter: Number of dictionary atoms: $K$, number of iterations: $I$, gain factor: $\lambda$
Initialization: Initialize the dictionary $\mathbf{D} = [\mathbf{a}_1\ \mathbf{a}_2\ ...\ \mathbf{a}_K]$ with $K$ normalized recordings randomly chosen from $\mathbf{Z}$ and allocate space for the sparse coefficient matrix $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ ...\ \mathbf{x}_M]$
Repeat $I$ times,
    • Sparse coding step: For each recording $\mathbf{z}_i$ of the data set,
        – Initialize the support $\Lambda$ (indexes of the selected atoms), the coefficient vector $\mathbf{x}_i$, and the residual vector $\mathbf{r}$ such that $\Lambda \leftarrow \{\varnothing\}$, $\mathbf{x}_i \leftarrow \mathbf{0}$, and $\mathbf{r} \leftarrow \mathbf{z}_i$.
        – Repeat iteratively until the stopping criterion is satisfied
            ∗ verify the stopping criterion:
                  $\mu(\mathbf{r}, \mathbf{D}) \leq \lambda\sqrt{2\text{log}(K)/N}$
            ∗ update support:
                  $\Lambda \leftarrow \Lambda \cup \arg\max_j \left|\mathbf{a}_j^T\ \mathbf{r}\right|$
            ∗ update solution using More-Penrose pseudoinverse:
                  $\mathbf{x}_i[\Lambda] \leftarrow (\mathbf{D}[:, \Lambda])^+\ \mathbf{z}_i$
            ∗ update residual vector:
                  $\mathbf{r} \leftarrow \mathbf{z}_i - \mathbf{D}[:, \Lambda]\ \mathbf{x}_i[\Lambda]$
    • Dictionary update step: For each atom $\mathbf{a}_j$,
        – Find recording indexes that use the atom:
              $\Omega \leftarrow \{k|\ \mathbf{X}[j, k] \neq 0\}$
        – Create temporary coefficient matrix and zero out atom coefficients:
              $\bar{\mathbf{X}} \leftarrow \mathbf{X}[:, \Omega]$
              $\bar{\mathbf{X}}[j, :] \leftarrow \mathbf{0}$
        – Apply SVD decomposition of residuals:
              $\mathbf{U}, \Delta, \mathbf{V} \leftarrow \text{svd}(\mathbf{Z}[:, \Omega] - \mathbf{D}\bar{\mathbf{X}})$
        – Update atom and nonzero coefficients:
              $\mathbf{a}_j \leftarrow \mathbf{U}[:, 1]$
              $\mathbf{x}_j[\Omega] \leftarrow \Delta[1, 1]\ \mathbf{V}[:, 1]$
Output: Dictionary $\mathbf{D}$, sparse coefficient matrix $\mathbf{X}$
```

APPENDIX B

DERIVATION OF THE IDEAL THRESHOLD FOR COHERENT DENOISING

Here, we derive an ideal upper bound of the statistical coherence of white Gaussian noise $\mathbf{n} \in \mathbb{R}^N$ of mean 0 and variance σ^2 relative to a redundant dictionary $\mathbf{D} = [\mathbf{a}_1\ \mathbf{a}_2\ ...\ \mathbf{a}_K] \in \mathbb{R}^{N\times K}$. Such a value for coherence is mathematically given by

$$\mu(\mathbf{n}, \mathbf{D}) = \max_{j\in 1,2,...,K} \left| \frac{\mathbf{n}^T}{||\mathbf{n}||_2}\mathbf{a}_j \right| . \tag{B-1}$$

Each entry of the normalized noise vector follows a Gaussian distribution of mean zero and variance $1/N$. This can be written as

$$\frac{\mathbf{n}^T}{||\mathbf{n}||_2} \sim \mathcal{N}_N(0, \frac{1}{N}) , \tag{B-2}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a notation that refers to a Gaussian distribution of mean μ and variance σ^2. Hence, the projection of the normalized noise vector on a dictionary ato, $\mathbf{a}_j^T\mathbf{n}/||\mathbf{n}||_2$ can be seen as a linear combination of mutually independent random variables following an identical Gaussian distribution of mean zero and variance $1/N$. Moreover, if χ_1, χ_2, ..., χ_n are mutually independent variables following Gaussian distributions of means μ_1, μ_2, ..., μ_n and variances σ_1^2, σ_2^2, ..., σ_n^2, then the linear combination of these variables $\sum_{j=1}^{n} c_j\chi_j$ follows a Gaussian distribution (Eisenberg and Rosemary, 2008) such that

$$\sum_{j=1}^{n} c_j\chi_j \sim \mathcal{N}\left(\sum_{j=1}^{n} c_j\mu_j, \sum_{j=1}^{n} c_j^2\sigma_j^2\right). \tag{B-3}$$

Hence, using the results in equations B-2 and B-3, it can be established that

$$\frac{\mathbf{n}^T}{||\mathbf{n}||_2}\mathbf{a}_j \sim \mathcal{N}(0, \frac{1}{N}) . \tag{B-4}$$

Using the definition in equation B-1 and the result in equation B-4, one may notice that estimating the coherence of the noise vector relative to the dictionary can be reformulated as estimating the maximum of the absolute value of K dependent but nondeterministic variables following an identical Gaussian distribution of mean zero and variance $1/N$. In addition, the maximum absolute value of n random variables $\chi_1, \chi_2, ..., \chi_n$ following an identical Gaussian distribution of mean zero and variance σ^2 has an asymptotically optimal upper bound of

$$\max_i |\chi_i| \leq \sigma\sqrt{2\log(n)} , \tag{B-5}$$

if the variables are independent (Berman, 1964; Donoho and Johnstone, 1994) or dependent but nondeterministic (Hartigan, 2014). Therefore, the coherence between the noise and the dictionary can be bounded such that

$$\mu(\mathbf{n}, \mathbf{D}) \leq \sqrt{\frac{2\log(K)}{N}} , \tag{B-6}$$

and the threshold $\mu_{Ideal} = \sqrt{2\log(K)/N}$ can be considered to be ideal for filtering Gaussian noise using coherent denoising.

REFERENCES

Aharon, M., M. Elad, and A. Bruckstein, 2006, k-svd: An algorithm for designing overcomplete dictionaries for sparse representation: IEEE Transactions on Signal Processing, **54**, 4311–4322.

Beckouche, S., and J. Ma, 2014, Simultaneous dictionary learning and denoising for seismic data: Geophysics, **79**, A27–A31.

Berman, S. M., 1964, Limit theorems for the maximum term in stationary sequences: The Annals of Mathematical Statistics, **35**, 502–516.

Cai, J.-F., H. Ji, Z. Shen, and G.-B. Ye, 2014, Data-driven tight frame construction and image denoising: Applied and Computational Harmonic Analysis, **37**, 89–105.

Canales, L., 1984, Randon noise reduction: 54th Annual International Meeting, SEG, Expanded Abstracts, 525–527.

Candès, E. J., and L. Demanet, 2005, The curvelet representation of wave propagators is optimally sparse: Communications on Pure and Applied Mathematics, **58**, 1472–1528.

Candès, E. J., and D. L. Donoho, 2000, *in* Curvelets: a surprisingly effective nonadaptive representation of objects with edges: Vanderbilt University Press, 105–120.

Candès, E. J., and D. L. Donoho, 2002, Recovering edges in ill-posed inverse problems: optimality of curvelet frames: Ann. Statist., **30**, 784–842.

Chen, Y., J. Ma, and S. Fomel, 2016, Double-sparsity dictionary for seismic noise attenuation: Geophysics, **81**, V103–V116.

Donoho, D., M. Elad, and V. Temlyakov, 2006, Stable recovery of sparse overcomplete representations in the presence of noise: IEEE Transactions on Information Theory, **52**, 6–18.

Donoho, D. L., and J. M. Johnstone, 1994, Ideal spatial adaptation by wavelet shrinkage: Biometrika, **81**, 425–455.

Eisenberg, B., and S. Rosemary, 2008, Why is the sum of independent normal random variables normal?: Mathematics Magazine, **81**, 362–366.

Elad, M., 2010, Sparse and redundant representations: From theory to applications in signal and image processing, 1st ed.: Springer Publishing Company, Incorporated.

Elad, M., and M. Aharon, 2006, Image denoising via sparse and redundant representations over learned dictionaries: IEEE Transactions on Image Processing, **15**, 3736–3745.

Engan, K., S. Aase, and J. Hakon Husoy, 1999, Method of optimal directions for frame design: Proceedings on 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2443–2446.

Fomel, S., and Y. Liu, 2010, Seislet transform and seislet frame: Geophysics, **75**, V25–V38.

Foster, D. J., C. C. Mosher, and S. Hassanzadeh, 1994, Wavelet transform methods for geophysical applications: SEG Technical Program Expanded Abstracts 1994, 1465–1468.

Gulunay, N., 1986, FX DECON and complex Wiener prediction filter: Presented at the 56th Annual International Meeting, SEG, Expanded Abstracts, Session: POS2.10.

Hartigan, J. A., 2014, Bounding the maximum of dependent random variables: Electronic Journal of Statistics, **8**, 3126–3140.

Hennenfent, G., and F. Herrmann, 2006, Seismic denoising with nonuniformly sampled curvelets: Computing in Science and Engineering, **8**, 16–25.

Liang, J., J. Ma, and X. Zhang, 2014, Seismic data restoration via data-driven tight frame: Geophysics, **79**, V65–V74.

Liu, Y., and S. Fomel, 2010, oc-seislet: Seislet transform construction with differential offset continuation: Geophysics, **75**, WB235–WB245.

Mallat, S., 2008, A wavelet tour of signal processing: The sparse way, 3rd ed.: Academic Press.

Neelamani, R., A. I. Baumstein, D. G. Gillard, M. T. Hadidi, and W. L. Soroka, 2008, Coherent and random noise attenuation using the curvelet transform: The Leading Edge, **27**, 240–248.

Pati, Y. C., R. Rezaiifar, Y. C. P. R. Rezaiifar, and P. S. Krishnaprasad, 1993, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition: Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers, 40–44.

Penrose, R., 1955, A generalized inverse for matrices: Mathematical Proceedings of the Cambridge Philosophical Society, **51**, no. 03, 406–413.

Rubinstein, R., M. Zibulevsky, and M. Elad, 2008, Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit.

Rubinstein, R., M. Zibulevsky, and M. Elad, 2010, Double sparsity: Learning sparse dictionaries for sparse signal approximation: IEEE Transactions on Signal Processing, **58**, 1553–1564.

Tosic, I., and P. Frossard, 2011, Dictionary learning: Signal Processing Magazine, IEEE, **28**, 27–38.

Tropp, J., 2004, Greed is good: Algorithmic results for sparse approximation: IEEE Transactions on Information Theory, **50**, 2231–2242.

Ulrych, T. J., and M. D. Sacchi, 2005, Information-based inversion and processing with applications, *in* Handbook of geophysical exploration, 1st ed.: Elsevier.

Yilmaz, O., 2001, Seismic data analysis: Processing, inversion, and interpretation of seismic data: SEG.

Yu, S., J. Ma, and S. Osher, 2016, Monte carlo data-driven tight frame for seismic data recovery: Geophysics, **81**, V327–V340.

Yu, S., J. Ma, X. Zhang, and M. D. Sacchi, 2015, Interpolation and denoising of high-dimensional seismic data by learning a tight frame: Geophysics, **80**, V119–V132.

Zhu, L., E. Liu, and J. H. McClellan, 2015, Seismic data denoising through multiscale and sparsity-promoting dictionary learning: Geophysics, **80**, WD45–WD57.

Chapter 5

Article II

The second article is entitled "Coherent noise suppression by learning and analyzing the morphology of the data". It was published in the journal Geophysics. A manuscript was sent to the editor the 6th of February 2017, a revised manuscript was sent the 21st of June 2017, it was published ahead of production the 15th of August 2017, and published the 9th of October 2017. The layout has been changed from the official publication to better fit the format of the thesis. The page number located in the header of each page of the article except the first one is relative to the article and is starting from A202, whereas the page number relative to the thesis is located in the footer of the page.

Coherent noise suppression by learning and analyzing the morphology of the data

Pierre Turquais[1,2], *Endrias G. Asgedom*[1], *Walter Söllner*[1]

ABSTRACT

We have developed a method for suppressing coherent noise from seismic data by using the morphological differences between the noise and the signal. This method consists of three steps. First, we applied a dictionary learning method on the data to extract a redundant dictionary in which the morphological diversity of the data is stored. Such a dictionary is a set of unit vectors called atoms that represent elementary patterns that are redundant in the data. Because the dictionary is learned on data contaminated by coherent noise, it is a mix of atoms representing signal patterns and atoms representing noise patterns. In the second step, we separate the noise atoms from the signal atoms using a statistical classification. Hence, the learned dictionary is divided into two subdictionaries; one describing the morphology of the noise, the other one describing the morphology of the signal. Finally, we separate the seismic signal and the coherent noise via morphological component analysis (MCA); it uses sparsity with respect to the two subdictionaries to identify the signal and the noise contributions in the mixture. Hence, the proposed method does not use prior information about the signal and the noise morphologies, but it entirely adapts to the signal and the noise of the data. It does not require a manual search for adequate transforms that may sparsify the signal and the noise, in contrast to existing MCA-based methods. We develop an application of the proposed method for removing the mechanical noise from a marine seismic dataset. For mechanical noise that is coherent in space and time, the results show that our method provides better denoising in comparison with the standard FX-Decon, FX-Cadzow, and the curvelet-based denoising method.

INTRODUCTION

In marine seismic surveys, the seismic wavefield is generally recorded by sensors located in the streamers that are towed by a vessel. The steering devices that are placed along the streamers, as well as barnacles that grow on the surface of the streamers, can perturb the flow of the water and cause local vibrations of the streamers. These vibrations are recorded by the motion sensors and appear in the seismic data. The recording of these vibrations is often referred to as the mechanical noise. This noise significantly hinders seismic processing and imaging if incorrectly removed.

Various analytical transforms, e.g., Fourier, wavelet (Mallat, 2008), curvelet (Candès and Donoho, 2000; Candès and Demanet, 2005), or seislet (Fomel and Liu, 2010), can be used to attenuate random noise in seismic data by sparse approximation. The key is to find the transform that represents the data into a domain in which the signal of interest is sparse, i.e., the signal can be represented with a minor part of the coefficients in the transform domain. When the data are represented with such

[1] Petroleum Geo-Services ASA, Oslo, Norway
[2] University of Oslo, Department of Geosciences, Oslo, Norway

a transform, the random noise in the data is spread along all the coefficients. The noise is then attenuated by approximating the data to its large amplitude coefficients only. The sparser the signal in the transformed domain, the higher is the noise attenuation. This strategy has been extensively studied for seismic data denoising (Foster et al., 1994; Hennenfent and Herrmann, 2006; Neelamani et al., 2008; Liu and Fomel, 2010).

Sparse representations can also be obtained via data-driven methods. For example, dictionary learning (DL) methods train redundant dictionaries to sparsify specific data. The dictionary is a set of elements called atoms and is learned such that the atoms represent the morphological structures or waveforms that compose the data. As such, it can be said that a dictionary learned on data contains the morphological diversity of the data. The data can later be approximated with a sparse linear combination of the dictionary atoms, which attenuates the random noise. Many methods for DL have been proposed, e.g., the method of optimal direction (MOD) (Engan et al., 1999), k-means singular value decomposition (K-SVD) (Aharon et al., 2006), data-driven tight frame (DDTF) (Cai et al., 2014), sparse K-SVD (Rubinstein et al., 2010), and SuKro (Dantas et al., 2017). MOD and K-SVD learn unstructured dictionaries; there is no constrain on the structure of the atoms. For more efficient training, DDTF, Sparse-KSVD, and SuKro learn structured dictionaries; in DDTF the dictionary is constrained to be a tight frame; in sparse K-SVD the atoms are constructed as sparse linear combinations of predefined basis functions; and in SuKro, the learned dictionary is a sum of Kronecker products of smaller dictionaries. The DL methods have proven to perform well for denoising seismic data (Beckouche and Ma, 2014; Liang et al., 2014; Yu et al., 2015; Zhu et al., 2015; Yu et al., 2016; Turquais et al., 2017). Another data-driven method, the Cadzow filtering method (Trickett, 2002, 2008), also called singular spectrum analysis (SSA) (Sacchi, 2009; Chen and Sacchi, 2015), uses rank reduction for denoising. This method embeds each frequency slice of the data into a Hankel matrix, mutes the low singular values, and averages the antidiagonal elements. The noise attenuation is mainly achieved by muting the low singular values as random noise is spread along all the singular values.

Although mechanical noise is generally unpredictable, it is recorded continuously in time and by several neighboring receivers. Therefore, it can appear coherent in space and time in the data. In that case, i.e., if the noise is not entirely random, the effectiveness of sparse approximation based denoising methods can be degraded, for either case in which the sparse approximation is in a fixed or a data-driven dictionary domain. In the case of a fixed dictionary, the part of the coherent noise that is described by the dictionary is also represented in the sparse approximation. In the case of a conventional DL method, the elementary patterns of the coherent noise are captured in the dictionary during the training step, and the noise is represented by the sparse approximation instead of being attenuated.

In seismic processing, coherent noise is often removed by exploiting its coherence property. For instance, the linearity of seismic interference noise helps to separate it from the signal in the $\tau - p$ domain (Yu, 2011). For swell noise, one can utilize its time-frequency characteristics, which are often different from those of the seismic signal. Vaezi and Kazemi (2016) propose to separate this noise based on non-negative matrix factorization of the power spectrum of its time-frequency representation. The coherence properties of the noise can also be exploited via morphological component analysis (MCA). MCA has been developed to decompose images into different morphological components (Starck et al., 2004, 2005). To separate a noise component from a signal component, the data are represented with a sparse combination of the elements from two dictionaries; each of the dictionaries describing the morphology of one of the components. If the two dictionaries have a low mutual coherence, and the noise and signal have highly sparse representations in their respective dictionaries, MCA separates the signal and the noise correctly (Starck et al., 2004; Bruckstein et al., 2009). For instance, ground-roll noise can be removed from land data by solving the MCA problem using a wavelet transform to represent the signal and a discrete cosine transform to represent the noise (Wang et al., 2010). However, predefining dictionaries to represent the signal and the noise components in a sparse manner is risky because the signal or the noise might not have a sparse representation in its attributed dictionary. In this case, the signal and noise separation is incomplete, and the quality of the denoising is poor.

In this paper, we propose a method that combines DL and MCA to separate coherent noise from seismic data. This method has been briefly introduced by Turquais et al. (2016). A dictionary is learned on the noise-contaminated data using the K-SVD method (Aharon et al., 2006; Rubinstein et al., 2008). The learned dictionary contains both elements representing seismic signal patterns and elements representing noise patterns; they are segregated using a statistical classification (Anderson and Bahadur, 1962). The learned dictionary is hence divided into two sub-dictionaries; one describing the morphology of the noise, and the other one describing the morphology of the signal. Both are included into an MCA problem in which the data is represented with a sparsity constraint. This sparsity constraint enables signal and noise separation because the signal cannot be sparsely represented in the noise dictionary, and is therefore represented in the signal dictionary, and vice versa for the noise.

The rest of the paper is organized as follows: the first section presents the theory and methodology, the second section illustrates the proposed method using a simple synthetic example, and the third section shows a successful application for removing the high frequency mechanical noise from marine seismic data.

METHODOLOGY

The proposed method is composed of three steps, namely, DL, atom classification, and MCA. The three steps will be presented separately, and then a workflow will describe how they are assembled to remove coherent noise.

Dictionary learning

The first step of the proposed method is using a DL algorithm to extract the morphological diversity of the data as a redundant dictionary. Practically, a dictionary $\mathbf{D}$ is a matrix containing unit vectors $\mathbf{a}_1, ..., \mathbf{a}_K$ in its columns (i.e., $\mathbf{D} = [\mathbf{a}_1 \, ... \, \mathbf{a}_K]$). These unit vectors are referred to as the atoms of the dictionary. Dictionaries are used to compute a sparse representation of a data. Computing a sparse representation of the recording $\mathbf{z} \in \mathbb{R}^N$ in a dictionary $\mathbf{D} \in \mathbb{R}^{N \times K}$ requires finding a sparse coefficient vector $\mathbf{x} \in \mathbb{R}^K$ such that $\mathbf{Dx}$ equals or closely approximates $\mathbf{z}$. The dictionary is the key element of the sparse representation problem. Its atoms need to describe the morphology of the data to be able to compute a representation that is sparse and accurate. A dictionary representing the morphology of a data set can be obtained by applying a DL algorithm on the data set or a representative sub-part of the data set. The sub-part of the data set used to learn the dictionary is referred to as the training set.

For seismic data application, DL is often applied in 2D on a gather. In this case, the training set $\mathbf{z}_1, ..., \mathbf{z}_M$ is a set of 2D patches that have been extracted from the gather and vectorized. One possibility to learn the dictionary on the training set is to find the dictionary $\mathbf{D} \in \mathbb{R}^{N \times K}$, with $K << M$, and the set of sparse coefficient vectors $\mathbf{x}_1, ..., \mathbf{x}_M$ which minimize the representation error given a sparsity constraint T placed on the sparse coefficient vectors. This minimization problem is mathematically expressed as

$$\min_{\{\mathbf{x}_i\}_{i=1}^{M},\mathbf{D}} \sum_{i=1}^{M} ||\mathbf{z}_i - \mathbf{D}\mathbf{x}_i||_2^2 \text{ subject to } ||\mathbf{x}_i||_0 \leq T \text{ , } i = 1, ..., M \text{ .} \tag{1}$$

In the proposed method, the DL problem is solved using the K-SVD method (Aharon et al., 2006). The resulting dictionary atoms describe 2D morphological structures in the t-x domain that are redundant in the training set and complementary for representing the recordings of the data set.

Atom classification

In a dictionary learned on an image, the atoms $\mathbf{a}_1, ..., \mathbf{a}_K$ of length N describe patterns when rearranged as 2D patches of size $\sqrt{N} \times \sqrt{N}$. In the case in which the image is a gather that is contaminated by coherent noise, the learned dictionary describes the morphology of the noise and the signal. If the signal and the noise are independently distributed in the window, and if their morphologies have low correlation, the two morphologies are described by different atoms of the dictionary. Hence, the atoms can be classified to create two subdictionaries; one signal dictionary $\mathbf{D}_s$ containing the atoms describing the signal morphology and one noise dictionary $\mathbf{D}_n$ containing the atoms describing the noise morphology. Below, we describe three different methods that can carry out this classification.

Attributes

To classify the atoms as signal or noise, it is necessary to use attributes. Here, an attribute is a value that is computed on the 2D pattern described by an atom. The quality of the classification depends on their capability to discriminate noise from signal. Therefore, it is necessary to select attributes that have different values for noise and signal atoms. In this work, we use textural attributes that are based on the gray-level co-occurrence matrix (GLCM) (Haralick et al., 1973). The GLCM is a discrete description of the probability of co-occurrence of two gray levels for two pixels with a given relative position in the pattern. For seismic applications, the gray-level is the dynamic range that has been rescaled and the pixels are the recorded samples. For the given relative position (Δt, Δx), the element at the ith line and jth column of the GLCM is the probability of changing from the amplitude i to j when moving by Δt samples in time and Δx samples in space. The seismic data are generally stored with 32 bits per samples and hence can have 2^{32} possible amplitude values. Because computing a GLCM for such a high number of amplitude values would be expensive, the data are rescaled prior computing the GLCM. The data are often converted to 4 or 5 bits data where each sample is rescaled to an integer value between 1 and 16 or 1 and 32 (Gao, 2003). The GLCM of a 2D array $\mathbf{A}$ can be computed as described in Algorithm 1.

Textural attributes (e.g., known as energy, homogeneity, inertia) are the weighted sum of the GLCM elements. They have proven to be successful for seismic data classification (Vinther et al., 1995; Vinther, 1997; West et al., 2002; Gao, 2003). In this study, the attributes used are the inertia for several relative positions. The inertia for the relative position (Δt, Δx) is mathematically given by

$$\text{Inertia} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (i-j)^2 \mathbf{P}[i,j] \text{ ,} \tag{2}$$

Algorithm 1 Computation of the GLCM of **A** for the relative position (Δt, Δx)

```
Input: matrix A of size $M \times N$; relative position ($\Delta t$, $\Delta x$); number $G$ of integer values.
Rescale the samples in A to a few integer values $1, 2, ..., G$.
Initialize the GLCM, P, of size $G \times G$ with zeros.
for $i = 1$ to $M$, and $j = 1$ to $N$, increment by 1 the sample at the $A[i, j]$th line and $A[i + \Delta t, j + \Delta x]$th column of P, where $A[i, j]$ stands for the value at the $i$th line and $j$th column of A.
Divide P by the sum of all its elements.
Output: P.
```

where $\mathbf{P}[i, j]$ is the element at the ith line and jth column of the GLCM computed for the relative position (Δt, Δx). The larger the probability that two samples separated by Δt samples in time and Δx samples in space have close amplitude values, the lower is the inertia. The inertia is therefore sensitive to the frequency content and the orientation of a pattern. For instance, if a pattern describes a high frequency signal, it has sharp amplitude variations in time, so the probability that its samples would conserve the same amplitude while moving in time is small, and its inertia is high for non-null Δt and null Δx. Similarly, if a pattern describes a flat linear event, it contains samples that conserve the same amplitude while moving only in space, and it has low inertia for null Δt and non-null Δx.

Supervised classification

A supervised classification analyzes the available examples in a training set to derive a law or condition that can classify new examples. For instance, if a training set contains some atoms labeled as ”signal” and some atoms labeled as ”noise”, a supervised process can be used to classify the atoms of the learned dictionary into the signal and the noise classes. The supervised processes include multivariate Gaussian classifiers (Anderson and Bahadur, 1962). To classify atoms that are either noise or signal, a multivariate Gaussian classifier assumes that the attributes of the signal and noise atoms follow two different multivariate Gaussian distributions. A multivariate Gaussian distribution is entirely defined by a mean vector $\boldsymbol{\mu}$, which gives the centroid of the distribution, and a covariance matrix $\boldsymbol{\Sigma}$, which gives the shape and orientation of the distribution. Such a distribution is denoted with $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The distribution of the signal is defined with the mean vector and the covariance matrix of the attributes computed on the atoms that are labeled as signal. Likewise, the distribution of the noise is defined using the atoms that are labeled as noise. The probability that a vector $\mathbf{f}$ containing n attributes belongs to a distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$p(\mathbf{f} \in \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \boldsymbol{\mu}) \right] . \tag{3}$$

In this classification, the formula in equation 3 is used to compute both probabilities that each atom from the dictionary belongs to the signal and noise distributions. Then, the atom is classified as signal or noise according to the highest probability.

One-class classification

If the training set contains only atoms labeled as signal or only atoms labeled as noise, the fully supervised model described above cannot be used to classify the atoms of the learned dictionary.

In this case, a one-class classification (Moya and Hush, 1996; Tax, 2001) is suitable. A one-class classification aims to identify patterns of a specific class among other patterns by learning from a training set containing only the patterns of that class. There are two possible scenarios: Either the training set contains only signal atoms, or it contains only noise atoms. From now on, we will consider that it contains only noise atoms. This does not aim to restrict the application of the method; it is to simplify the explanation and understanding. In that case, a one-class classifier uses the atoms labeled as noise to identify the noise atoms of the learned dictionary and classifies the rest of the atoms as signal. It is assumed that the distribution of the attributes computed on the noise atoms is a multivariate Gaussian distribution. This distribution is defined by the mean vector and the covariance matrix of the attributes computed on the atoms labeled as noise. An atom from the learned dictionary is identified as noise if it has a small Mahalanobis distance to the noise distribution. The Mahalanobis distance for an atom of feature vector **f** to a distribution of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by

$$d_M(\mathbf{f}) = \sqrt{(\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{f} - \boldsymbol{\mu})} \, . \tag{4}$$

The Mahalanobis distance is used because it is unitless, scale invariant, and takes into account the correlation of the attributes. If **f** follows the multivariate Gaussian distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, there is a 68.3% of probability that $d_M(\mathbf{f}) < 1$, a 95.5% of probability that $d_M(\mathbf{f}) < 2$, and a 99.7% of probability that $d_M(\mathbf{f}) < 3$. Keeping this last probability in mind, one can classify an atom as noise if the Mahalanobis distance is smaller than 3 and one can classify an atom as signal if the Mahalanobis distance is higher than 3. The signal atoms would not be misclassified if the density probability function of their attributes does not significantly overlap the one of the noise attributes.

Unsupervised classification

If there is no training set, one has to use an unsupervised classification. An unsupervised classification groups together the patterns of a data set that have the closest attributes. In the case in which the number of classes in the set is a priori known and in which multivariate Gaussian distributions are assumed for the classes, the unsupervised classification can be carried out with the k-mean clustering algorithm (MacQueen, 1967). For a given number C of classes, it finds the clusters $G_1, ..., G_C$ of atoms which minimize the sum of the squared Mahalanobis distances between each attribute vector and its closest cluster centroid. This minimization problem is written such as

$$\min_{G_1,...,G_C} \sum_{k=1}^{C} \sum_{i \in G_k} (\mathbf{f}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{f}_i - \boldsymbol{\mu}_k) \, , \tag{5}$$

where $\mathbf{f}_i$ is the attribute vector of the atom i, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and the covariance matrix of the attribute vectors in the cluster k. For the signal and noise separation problem, C is set to 2, and each of the resulting clusters, G_1 and G_2, should contain either the noise atoms or the signal atoms.

Morphological component analysis

MCA (Starck et al., 2004, 2005) is a method that uses sparse representation as the driving force to separate the different morphological components in a mixture. Consider a recording **z** that contains a signal component that can be sparsely represented in a dictionary $\mathbf{D}_s$, and a noise component that can be sparsely represented in a dictionary $\mathbf{D}_n$. The MCA problem consists in finding the sparse representation of the recording in both dictionaries. This can be done by finding the sparse vectors $\mathbf{x}_s$ and $\mathbf{x}_n$ that are the solution of the following minimization problem

$$\min_{\mathbf{x}_s,\mathbf{x}_n} ||\mathbf{z} - \mathbf{D}_s\mathbf{x}_s - \mathbf{D}_n\mathbf{x}_n||_2 \text{ subject to } ||\mathbf{x}_s||_0 + ||\mathbf{x}_n||_0 \leq T \, , \tag{6}$$

where T is the sparsity of the representation. Such a problem can be solved using orthogonal matching pursuit (OMP) (Pati et al., 1993). The resulting $\mathbf{D}_s\mathbf{x}_s$ and $\mathbf{D}_n\mathbf{x}_n$ are sparse approximations of the signal and noise components, respectively. The separation of the components is exact if the sparsity of the recording in the two dictionaries is below a threshold dictated by the mutual coherence of the dictionaries (Starck et al., 2004; Bruckstein et al., 2009). This requires the signal and noise components to both have a very sparse representation in their corresponding dictionary and the correlation between the signal and noise morphologies to be low. The sparse approximations $\mathbf{D}_s\mathbf{x}_s$ and $\mathbf{D}_n\mathbf{x}_n$ are random noise free because random noise cannot be represented sparsely. For denoising, the component $\mathbf{D}_s\mathbf{x}_s$ is of interest because it contains neither coherent noise nor random noise. If the signal is not strictly sparse in its attributed dictionary, $\mathbf{D}_s\mathbf{x}_s$ might not represent the entire signal. In this latter case, the signal can be retrieved by subtracting the reconstructed noise component from the recording, i.e., $\mathbf{z} - \mathbf{D}_n\mathbf{x}_n$. This solution contains random noise but preserves better the signal.

In the proposed method, the atoms in $\mathbf{D}_s$ and $\mathbf{D}_n$ are vectors of length N describing small 2D patterns of size $\sqrt{N} \times \sqrt{N}$. To separate the signal from the noise in a 2D gather, MCA needs to be applied to all the juxtaposed patches of size $\sqrt{N} \times \sqrt{N}$ in the gather. The resulting signal and noise patches need to be respectively assembled to generate gather sized signal and noise components. To obtain a more accurate separation result, MCA can also be applied to overlapping patches. In the latter case, the multiple versions of the same sample are averaged when the signal or noise patches are assembled.

The proposed work flow

When the proposed method is used to separate the signal and noise components of a 2D data $\mathbf{Y}$, the workflow can be summarized as follows:

I **Dictionary learning:** The K-SVD algorithm is used to learn a dictionary $\mathbf{D}$ from the data $\mathbf{Y}$.

II **Atom classification:** The signal and the noise atoms from the dictionary $\mathbf{D}$ are segregated to obtain a dictionary $\mathbf{D}_s$ containing the signal atoms and a dictionary $\mathbf{D}_n$ containing the noise atoms. The segregation of the dictionary atoms is carried out with either a supervised, one-class, or unsupervised classification, depending on the availability of labeled atoms.

III **MCA:** OMP is used to solve the problem in equation 6 for overlapping patches extracted from $\mathbf{Y}$. The resulting patch-sized signal and noise components are respectively assembled and averaged to reconstruct the data-sized signal and noise components.

SYNTHETIC EXAMPLE

In this section, we illustrate the proposed morphological decomposition method with a synthetic example. We synthesized a noisy data by adding a recording of mechanical noise to a window of a synthetic shot gather. The mechanical noise was recorded during a marine survey for the same configuration as the signal was synthesized. The resulting noisy data are of size 100×100 samples with a sampling of 2 ms in time and 12.5 m in space and with a signal-to-noise ratio (S/N) of 2.23 dB. The S/N in dB of a data $\mathbf{d}$ is given by the formula

$$\mathrm{S/N} = 10 \log_{10} \frac{\|\mathbf{d}_{\mathrm{ref}}\|_2^2}{\|\mathbf{d}_{\mathrm{ref}} - \mathbf{d}\|_2^2} , \tag{7}$$

where $\mathbf{d}_{\mathrm{ref}}$ is the noise free data reference. Here, the data do not contain the frequencies below 10 Hz because they were removed due to very poor S/N in this range. The windows of signal, noise, and noisy data are presented in Figure 1. Note that these data are atypically small-sized. We selected a small-scale example because it provides results that are easier to display, explain, and understand, but in practice the method would be more efficient on larger-sized data.

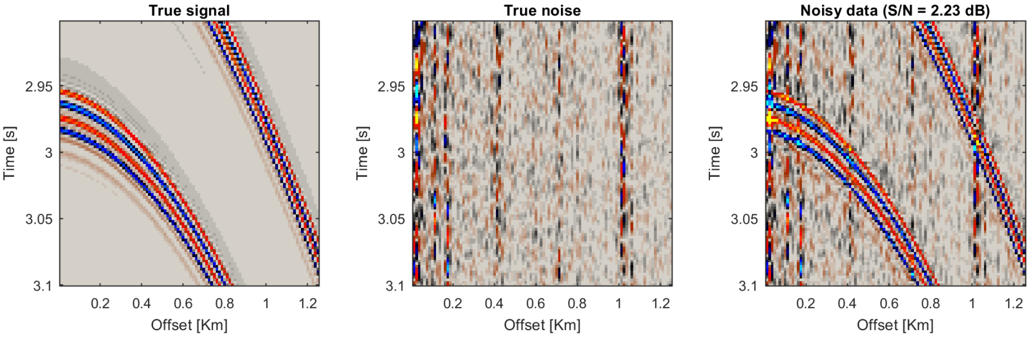

Figure 1: A window of synthetic signal (left) and a window of recorded noise (center) were summed to generate a window of noisy data (right). The S/N is indicated above the noisy data.

The K-SVD algorithm was used to learn a dictionary on the noisy data in Figure 1. The parameters were the following: for the training, 8000 patches of size 10×10 were extracted from the noisy data; the algorithm ran with 15 iterations; the number of atoms in the learned dictionary was set to 200; the sparsity threshold, T in equation 1, was set to 8. The choice of these parameters has been made considering the following guidelines. The patch size needs to be chosen considering that within a patch of chosen size, the shape of the noise needs to be different from the shape of the signal, to enable later the separation of the noise atoms from the signal atoms. The number of atoms in the dictionary should be higher than the number of samples in an atom to enforce redundancy in the dictionary. Redundant dictionaries lead to sparser representations, and from our experience, to more accurate signal and noise separation. Also, for an accurate DL, the number of patches in the training set should be several times higher than the number of dictionary atoms. Despite these constraints, we were left with many possibilities for setting the parameters. The rest of the decision was empirically-based as we aimed for a solution that compromises between tractability of the algorithm and accuracy of the representation. However, how to set optimally these parameters is open to discussion. The 200 atoms of the output dictionary have been rearranged as 10×10 patches and represented in Figure 2. In this figure, it can be observed that some of the atoms represent signal patterns whereas others represent noise patterns.

The inertia of the GLCM for the relative distances $(\Delta t = 1, \Delta x = 1)$ and $(\Delta t = 1, \Delta x = 0)$ were

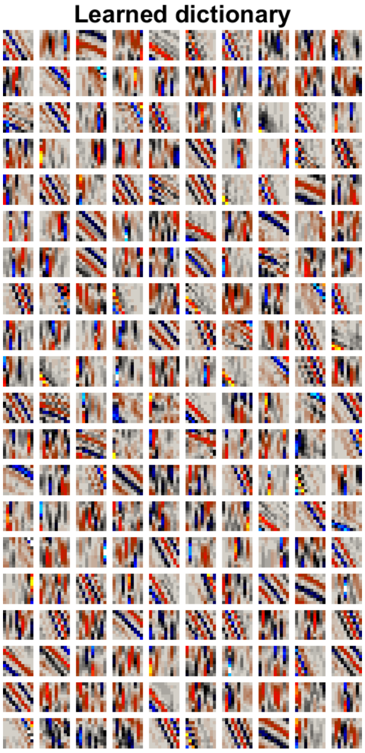

Figure 2: The dictionary that is learned on the noisy data. Each atom of the dictionary is displayed as a small 2D patch.

selected as attributes for the classification of the atoms. A look at the efficiency of these two attributes for discriminating signal and noise atoms is given in Figure 3. The two highest plots show a noise and a signal atom selected from the learned dictionary. Below each atom, is successively presented its GLCMs computed for the relative positions $(\Delta t = 1, \Delta x = 1)$ and $(\Delta t = 1, \Delta x = 0)$. The GLCMs have been computed with a number G of integer values set to 16. For the signal atom and the relative position $(\Delta t = 1, \Delta x = 1)$, the high value elements of the GLCM are around the diagonal. This indicates that the values of two samples that are separated by one sample in time and one sample in space have a high probability to be close. This is explained by the atom describing an event that dips roughly in the direction of the selected relative position. In contrast, the high value elements of the GLCM for $(\Delta t = 1, \Delta x = 0)$ are more spread. This is because the amplitude values of the signal pattern are sharply varying in time. For the noise atom, the high value elements of the GLCM are more concentrated around the diagonal for the relative position $(\Delta t = 1, \Delta x = 0)$ compared with the relative position $(\Delta t = 1, \Delta x = 1)$. This is explained by the noise pattern being smoother in time than in space. For the four GLCMs presented in Figure 3, the inertia in written below the plots. The inertia is small when the high value elements of the GLCM are close to the diagonal because the inertia is a weighted sum of the GLCM elements where a higher weight is given to the elements further away from the diagonal (cf. equation 2). Thus, a small inertia for $(\Delta t = 1, \Delta x = 1)$ may be a distinguishing characteristics of a signal atom whereas a small inertia for $(\Delta t = 1, \Delta x = 0)$ may be a distinguishing characteristics of a noise atom.

Later, we will display the location of the atoms in the attribute space, i.e., the space with the attributes as axes. As the dimensionality of this space is equal to the number of attributes, selecting

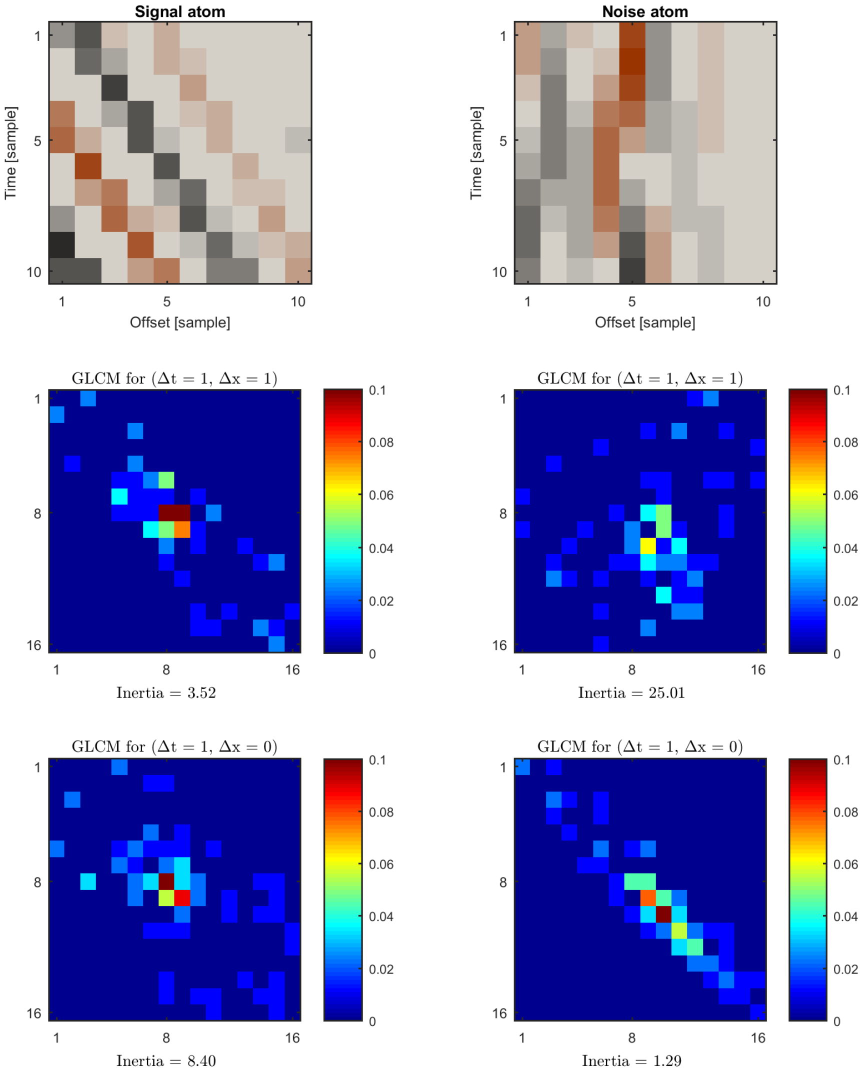

Figure 3: Inertia of the GLCM for classification of the signal and noise patterns. A signal atom (top-left) and a noise atom (top-right) were selected from the learned dictionary. Bellow each atom are presented its GLCMs for the relative positions ($\Delta t = 1$, $\Delta x = 1$) and ($\Delta t = 0$, $\Delta x = 1$). The inertia values of the GLCMs are written below them. The signal atom has a low inertia for the relative position ($\Delta t = 1$, $\Delta x = 1$), and the noise atom has a low inertia for the relative position ($\Delta t = 1$, $\Delta x = 0$).

three or four attributes would give a 3D or 4D space, which would be hard to visualize. This is why we limited the number of attributes to two in this example.

To divide the dictionary into a signal dictionary and a noise dictionary, the three types of classification presented in the section ”Atom classification” were tested. For the three possibilities, the processes applied are successively presented below:

- For the supervised classification, a noise and a signal model were used. Both models were of size 100×100 samples and taken from the same gathers where the signal and noise data presented in Figure 1 have been extracted. These models are presented in the two highest plots of Figure 4. For each model, K-SVD was used to learn a dictionary. The parameters used to learn the dictionary on the models are the same as the ones used to learn the dictionary on the noisy data except for the number K of atoms and the sparsity threshold T that are respectively set to 100 and 4. The parameters K and T are twice smaller when learning on the models compared with when learning on the noisy data because there are less data to represent in a signal or a noise model compared to in a mixture of both. The signal and noise output dictionaries are presented in the middle plots of Figure 4. For each atom of the dictionaries, the inertia for $(\Delta t = 1, \Delta x = 1)$ and $(\Delta t = 1, \Delta x = 0)$ were computed. The location of the atoms in the attribute space for both dictionaries is presented in the two lowest plots of Figure 4. We then computed the mean vector $\boldsymbol{\mu}_s$ and the covariance matrix $\boldsymbol{\Sigma}_s$ of the attributes of the 100 signal atoms, and the mean vector $\boldsymbol{\mu}_n$ and the covariance matrix $\boldsymbol{\Sigma}_n$ of the attributes of the 100 noise atoms. Both multivariate Gaussian density functions defined by ($\boldsymbol{\mu}_s$,$\boldsymbol{\Sigma}_s$) and ($\boldsymbol{\mu}_n$,$\boldsymbol{\Sigma}_n$) are shown in the two lowest plots of Figure 4 with lines of equal probability. For each atom of the learned dictionary, we computed its probability to belong to the signal class and its probability to belong to the noise class. To compute the probability that an atom belongs to the signal class, the formula in equation 3 was used with $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$. Similarly, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ were used to compute the probability that an atom belongs to the noise class. Finally, an atom was classified according to the highest probability.

- For the one-class classification, only the noise model was used. Similar to the supervised classification, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ were computed. For each atom of the learned dictionary, the Mahalanobis distance was computed using the formula in equation 4. The atoms with a Mahalanobis distance smaller than 3 were classified as noise and the rest as signal.

- For the unsupervised classification, no model was used. The selected attributes were computed on all the atoms of the learned dictionary. Then, the k-mean clustering algorithm was used to solve the problem in equation 5 with the number of clusters set to 2. The cluster of signal atoms was manually identified and its atoms were classified as signal. The atoms from the other cluster were classified as noise.

For the three types of classification, the decision map and the classification results in the attribute space are presented in Figure 5. The decision map is the class that would be given to an unlabeled atom in function of its location in the attribute space. The subspace where an unlabeled atom would be classified as signal is colored in blue whereas the subspace where an unlabeled atom would be classified as noise is colored in red. The location of the atoms from the dictionary presented in Figure 2 is superposed to the decision map. The atoms classified as signal are indicated with blue dots, whereas the atoms classified as noise are indicated with red dots. The three decision maps are very similar and the classifications of the learned atoms are the same except for five atoms.

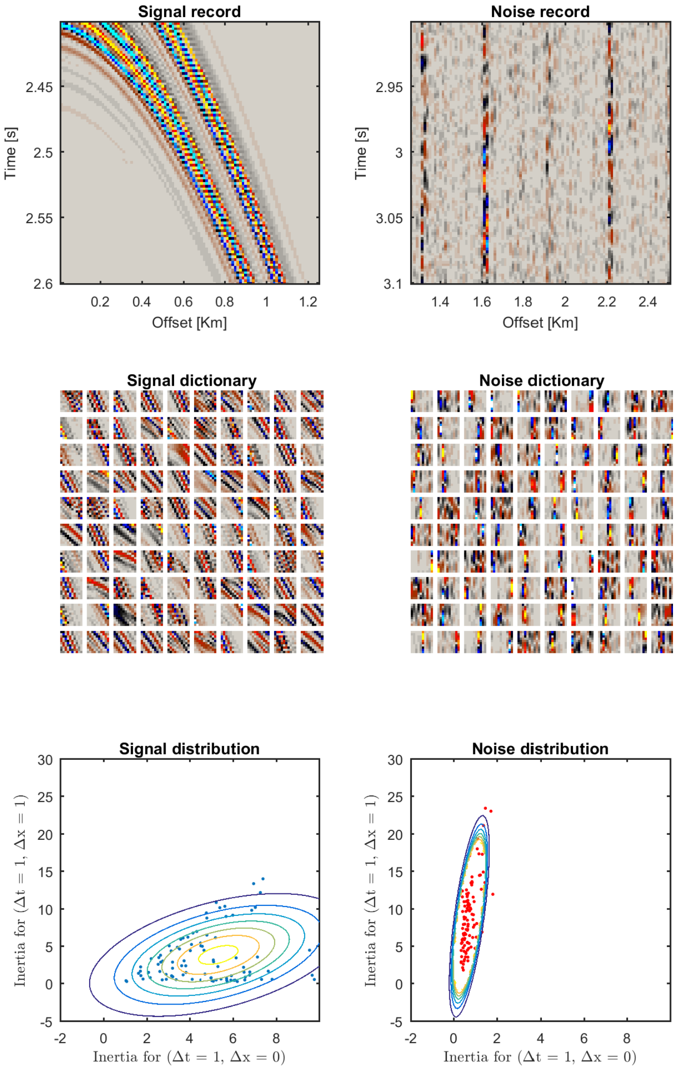

Figure 4: The signal and noise models that are used by the supervised and one-class classifiers. The highest plots are the signal and noise data models. The central plots are the dictionaries learned on the data models. The lowest plots show the location of the atoms in the attribute space. The location of the signal atoms is indicated in the left plot with blue dots and the location of the noise atoms is indicated in the right plot with red dots. The multivariate Gaussian density functions that characterize the signal and noise distributions are shown with lines of equal probability.

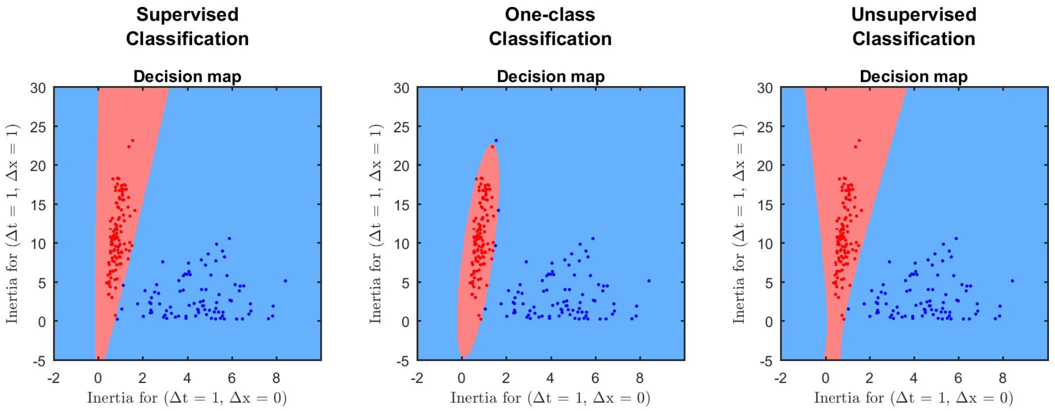

Figure 5: Decision map for the supervised, one-class, and unsupervised classifications. The blue and red areas are the sub-domains in which an atom is classified as signal and noise, respectively. The location of the dictionary atoms is indicated with blue dots for the atoms classified as signal and with the red dots for the atoms classified as noise.

The classified signal and noise dictionaries for the three types of classification are presented in Figure 6. To display the signal dictionary, the noise patches are masked with gray patches. Similarly, the signal patches are masked to display the noise dictionary. Five atoms have been classified differently by the three types of classification; their positions in the dictionaries have been indicated with black frame boxes.

Finally, MCA was applied to separate the signal and noise components. The OMP algorithm was used to solve the problem presented in equation 6 for patches of size 10×10 overlapping on 9 samples in both dimensions. The sparsity threshold T was set to 8, as when the dictionary was learned. For the three pairs of signal and noise dictionaries resulting from the classifications, the separated components are presented in Figure 7a. The S/N of the signal components is given under the higher plots. The differences between the true and recovered components are shown in Figure 7b. The three results provide high noise attenuation as well as signal preservation.

For this example, the three classifications seem to have equivalent effectiveness. The unsupervised classification has managed to place the signal atoms in one cluster and the noise atoms in the other cluster because there was a large quantity of both of them in the dictionary. If there would have been much fewer noise atoms than signal atoms, the unsupervised classification could have divided the signal atoms into two clusters to minimize the cost function in equation 5. The signal would have been separated in two parts instead of being separated from the noise. Therefore, a supervised or one-class classifier would be preferred in cases in which the method is blindly applied.

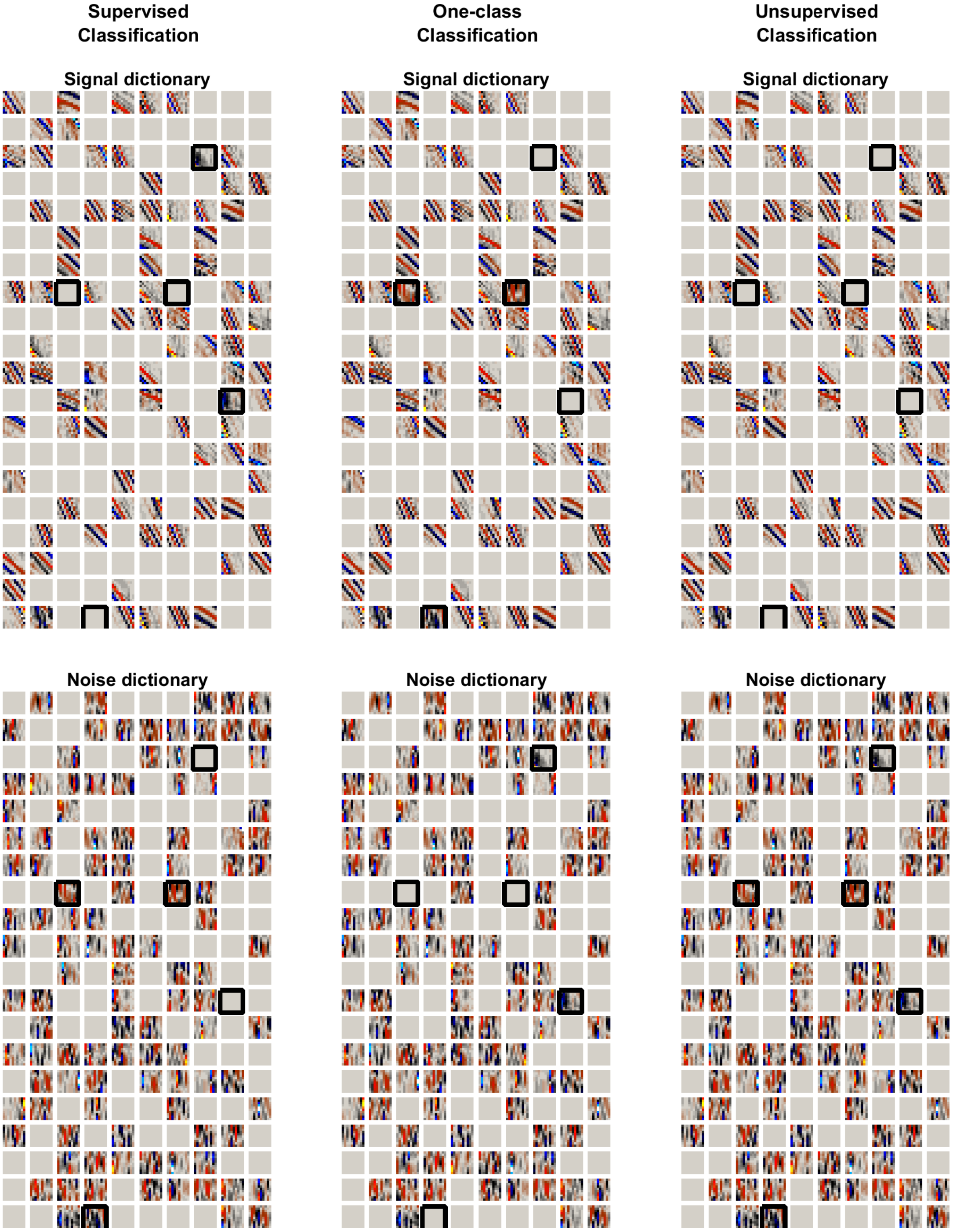

Figure 6: Classified signal and noise dictionaries for the supervised, one-class and unsupervised classifications. The black boxes point out the positions of the atoms for which the results of the classifications differ.

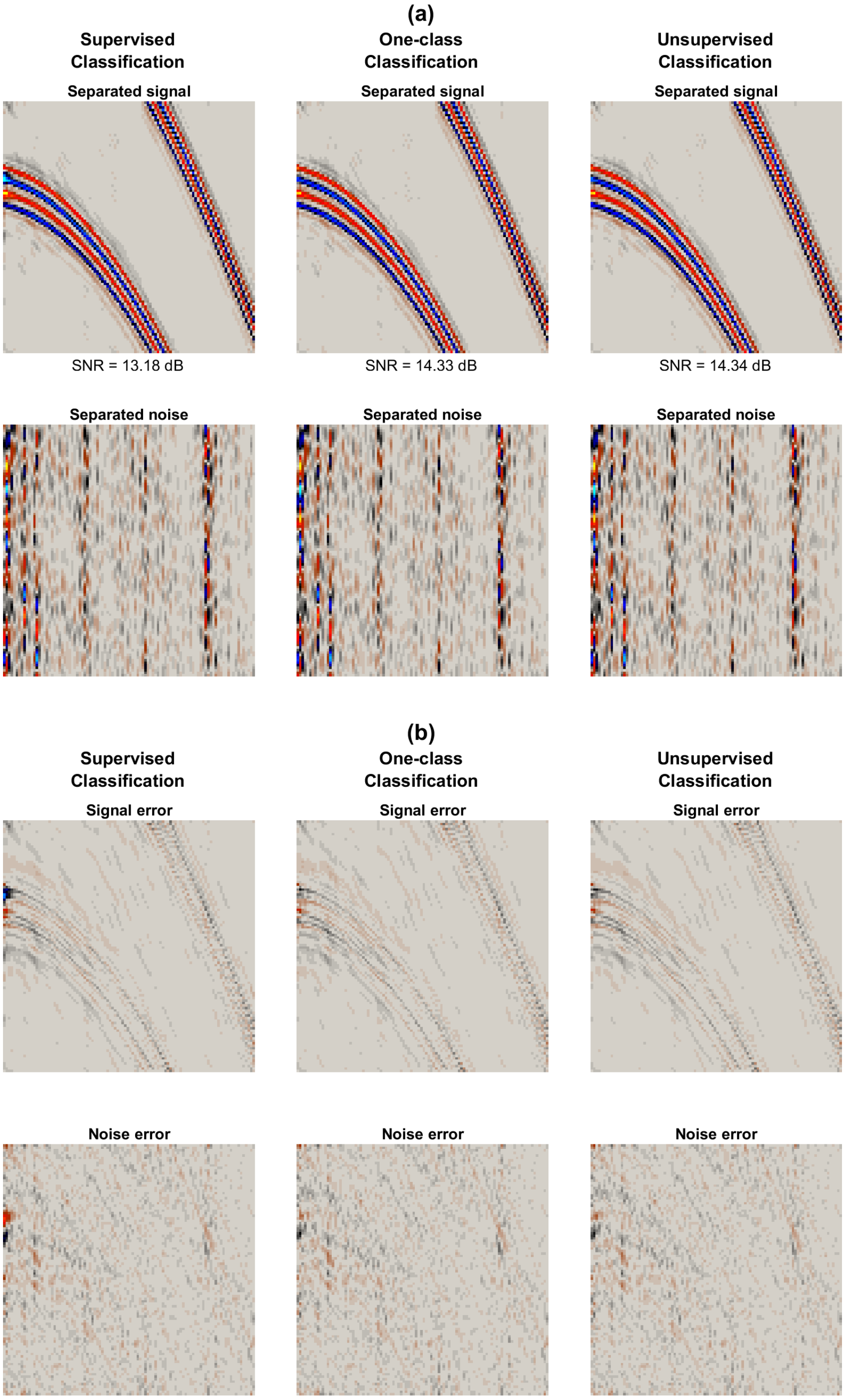

Figure 7: Results of the signal and noise separation by MCA. The results are presented for each pair of signal and noise dictionaries resulting from the three types of classification. The separated components are presented in panel a, and their differences with the true components are presented in panel b.

FIELD DATA APPLICATION

Steering device-related noise removal

We selected a raw shot gather acquired during a marine survey. The data are the vertical particle velocity sampled at 2 ms in the temporal dimension and 12.5 m in the offset dimension, and they are contaminated by steering device-related noise. The frequency range 0-10 Hz was muted due to very poor S/N. Denoising this frequency range is not in the scope of this work. The muted signal is generally retrieved using pressure measurements (Day et al., 2013). This shot gather is shown in Figure 8.

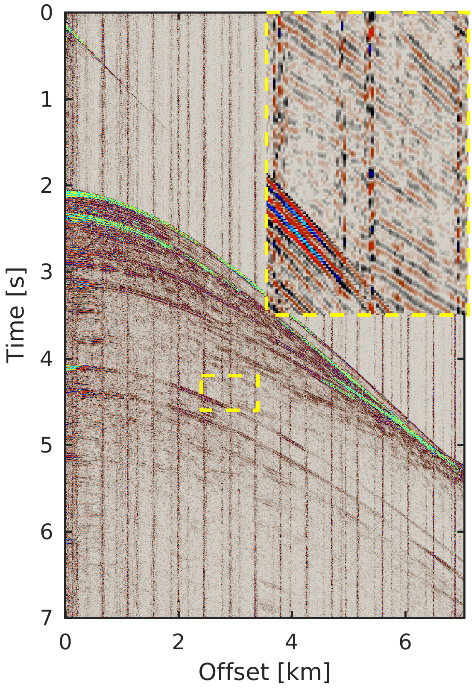

Figure 8: Shot gather contaminated by steering device-related noise. The close-up on the data framed by the yellow dashed box is shown in the top-right corner of the plot.

We applied the proposed method to the presented shot gather. We learned a dictionary of 4,000 atoms using the K-SVD algorithm. The parameters were the following: seven iterations of the algorithm were run, 40,000 patches of size 12×12 samples were used for the training, and the sparsity threshold T was set to 8. For the classification of the learned dictionary, three attributes were selected; they were the inertia for the relative positions $(\Delta t = 1, \Delta x = 0)$, $(\Delta t = 0, \Delta x = 1)$, and $(\Delta t = 1, \Delta x = 1)$. For such a field data application, a noise only model is available in the upper right part of the shot gather; the area where the first arrival of the seismic waves is not yet recorded. Due to the possibility to have a noise model, we used a one-class classifier to segregate the atoms of the learned dictionary. The noise model used was the data between 1 s and 2 s in time and 4.0 km and 6.5 km in space. The one-class classification was applied in the same way as it was applied in the synthetic example presented earlier. Because three attributes were selected, the classification was carried out given the location of the atoms in a 3D attribute space. Because it is hard to visualize the location of the atoms in this space from a 3D plot, we present the results via projections of the space onto planes; in Figure 9, each atom is located given its inertia for the relative distances (panel a) $(\Delta t = 1, \Delta x = 1)$ and $(\Delta t = 1, \Delta x = 0)$, (panel b) $(\Delta t = 0, \Delta x = 1)$ and $(\Delta t = 1, \Delta x = 0)$, and (panel c) $(\Delta t = 1, \Delta x = 1)$ and $(\Delta t = 0, \Delta x = 1)$. If an atom was classified as signal, the location is indicated with a blue dot, and if it was classified as noise, the location is indicated with a red dot. In Figure 9d-f, we give a view on the shape of the atoms depending on their location in the attribute space. The attribute spaces presented in Figure 9a-c were divided into small squares, and for each square, an atom that was located in the square was displayed. When several atoms were located in the square, the closest to the center of the square was displayed. If no atom was in the square, no atom was displayed. In addition, the atoms that were classified as noise were framed with red boxes. The two dictionaries resulting from the classification were used to separate the signal and the noise with MCA. The separation was carried out for patches of size 12×12 overlapping on 10 samples in both dimensions and with the sparsity threshold T set to 8. To well preserve the signal, the noise component was subtracted from the data to obtain the denoised data. This is equivalent to adding the residual of the sparse approximation to the signal component.

The proposed method was compared with academic implementations of FX-Decon (Gulunay, 1986), FX-Cadzow (Trickett, 2002), and curvelet (Hennenfent and Herrmann, 2006) denoising. For the FX-Decon method, the algorithm used was from SeisLab and corresponds to the implementation proposed by Ulrych and Sacchi (2005), p. 229-232, where both forward and backward error prediction filters are used. The filtering was applied to windows of size 50×50 samples overlapping on 25 samples in both dimensions and with filters of size six samples. These parameters have been selected because they perform well on this example and they have been shown to give the best denoising results on other examples (Chen et al., 2016). FX-Cadzow was applied with a rank parameter of 1 and with windows of size 130×8 samples overlapping on 65 samples in time and six samples in space (Oropeza, 2010). In the curvelet method, the curvelet coefficients of the data were calculated using wrapping curvelets, 16 angles, and a spgl1 solver for the l1 norm constraint. Then, the smallest curvelet coefficients that correspond to 4% of the data energy were muted for denoising.

The denoised data obtained with the proposed method, FX-Decon, FX-Cadzow, and the curvelet method are presented in Figure 10. There, we observe that the proposed method provides a cleaner result compared with the other methods. The removed noise gathers, i.e., the difference between the input and the denoised gathers, are presented in Figure 11. There, we observe that the proposed method preserves better the signal compared with the other methods.

We shall examine the denoising results obtained with the proposed method in the $f - k$ domain. In the $f - k$ spectrum of the noisy data presented in Figure 12a, we see that the noise is overlapping the signal between 10 and 50Hz. In the spectra of the denoised data and removed noise presented in Figure 12b and 12c, we observe that the noise has been effectively removed when the seismic reflections remain untouched. However, we see that a minor part of the direct wave has also been removed.

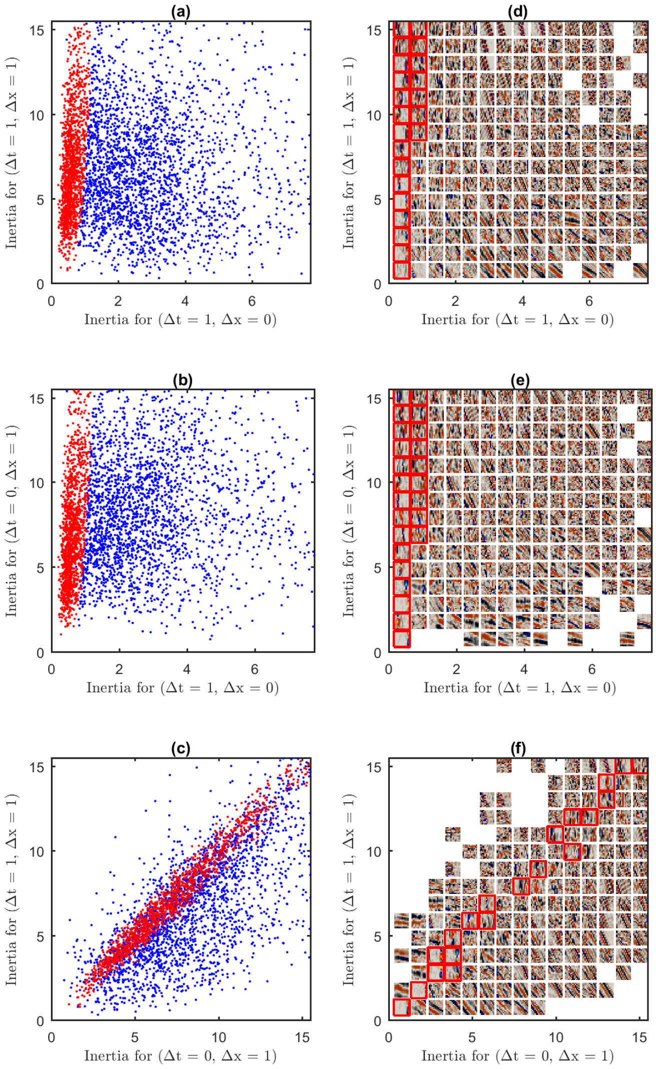

Figure 9: Classification of the dictionary learned from the data contaminated by steering device-related noise. (a-c) Location of the atoms in projections of the 3D attribute space. The location of the atoms classified as signal are indicated with blue dots and the location of the atoms classified as noise are indicated with red dots. (d-f) The shape of atoms are presented given their location in the projections of the attribute space. The atoms that were classified as noise are framed with red boxes.

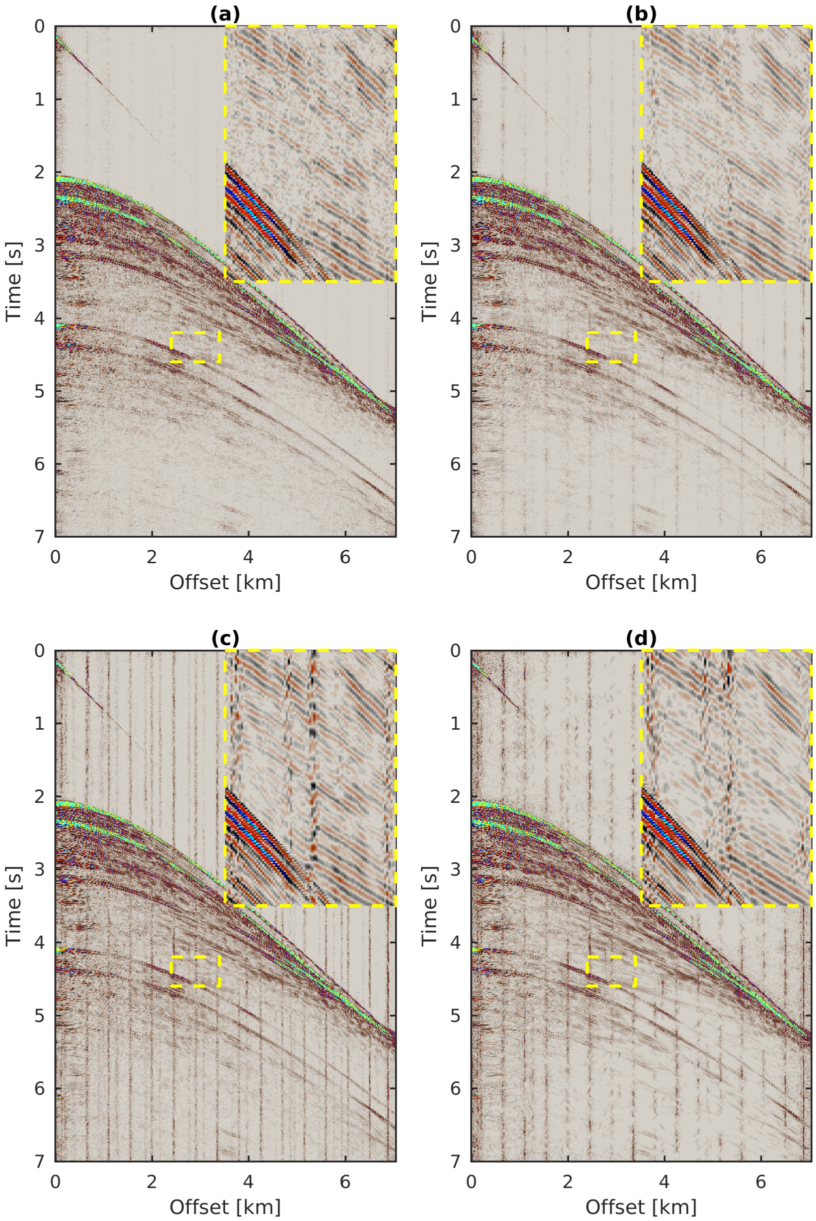

Figure 10: Denoised data that is obtained with (a) the proposed method, (b) FX-Decon, (c) FX-Cadzow, and (d) the curvelet method.

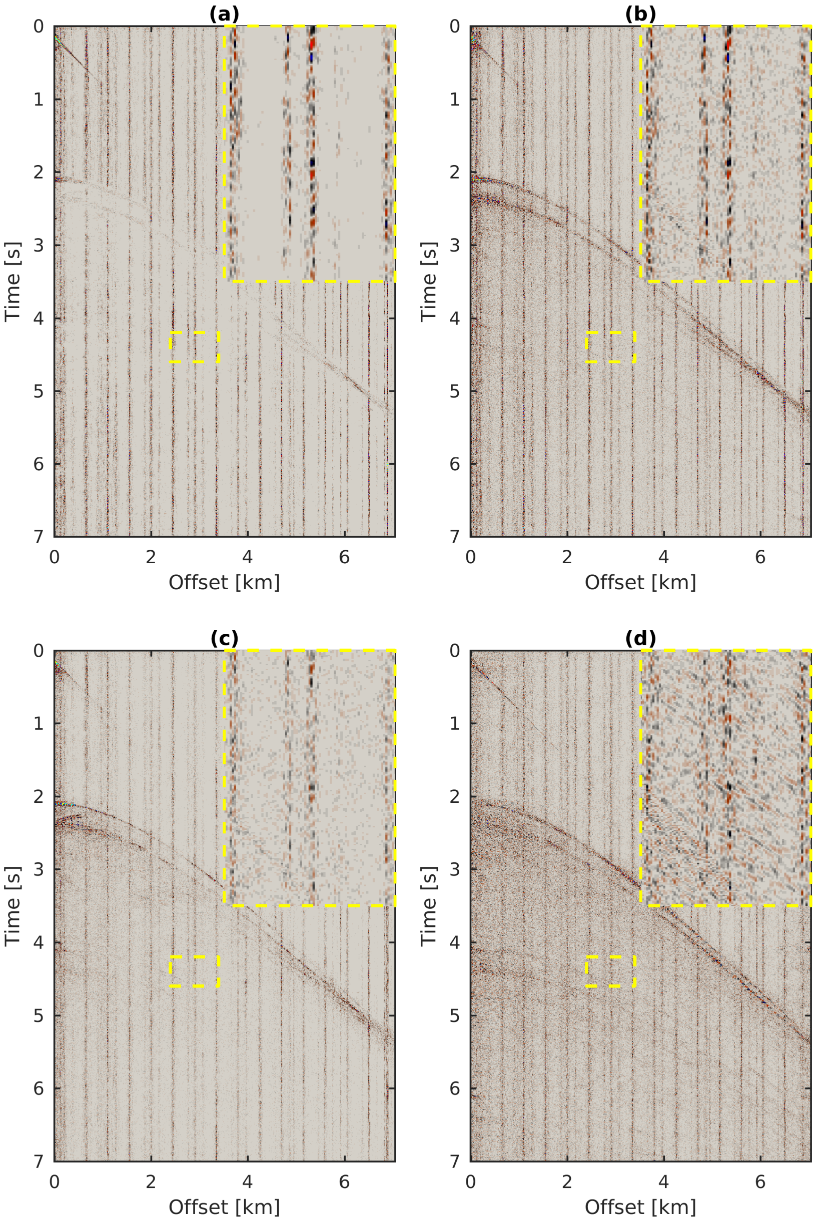

Figure 11: Noise that is removed by (a) the proposed method, (b) FX-Decon, (c) FX-Cadzow, and (d) the curvelet method.

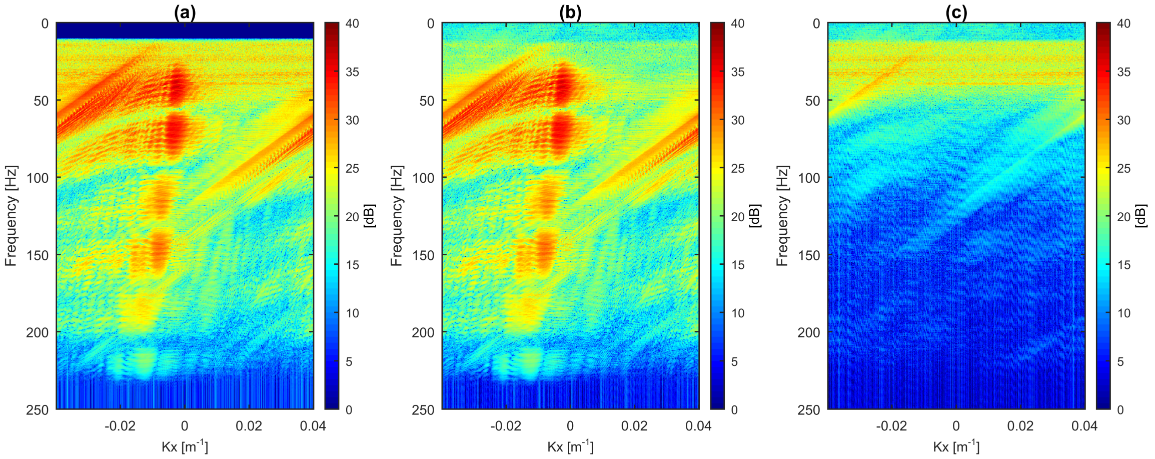

Figure 12: The $f-k$ spectra of the (a) data contaminated by steering device-related noise, (b) denoised data obtained with the proposed method, and (c) corresponding removed noise.

Barnacle-related noise removal

We selected a second raw shot gather acquired during another marine survey. The data are the vertical particle velocity sampled at 2 ms in the temporal dimension and 12.5 m in the offset dimension. We removed the frequencies below 10 Hz for the reason explained in the previous example. This time, the shot gather is contaminated by barnacle-related noise. This shot gather is presented in Figure 13.

We applied the proposed method, FX-Decon, and FX-Cadzow with the same parameters as the ones used in the previous example. In the curvelet method, the removed energy was increased to 10% because there is more noise in this shot gather compared with the previous example.

The denoised data obtained with the proposed method, FX-Decon, FX-Cadzow, and the curvelet methods are presented in Figure 14. The proposed method removes more noise than FX-Cadzow and the curvelet method but slightly less noise than FX-Decon. For the four methods, the removed noise sections are presented in Figure 15. In these gathers, we observe that the proposed method preserves better the signal compared with the other methods.

In the previous example, we observed that the proposed method was better than FX-Decon for removing steering device-related noise. For removing the barnacle related-noise of this example, it is not better. This could be explained by the different characteristics of the noises; the barnacle-related noise was less sparse and less coherent in space than the steering device-related noise. The fact that the barnacle-related noise was less sparse reduced the efficiency of the proposed method because the sparse representation is less accurate when the sparsity decreases (Bruckstein et al., 2009). The fact that the barnacle-related noise was less coherent from one trace to another benefited FX-Decon, which is hindered when the noise is linear.

The $f-k$ spectra of the data and the results obtained with the proposed method are presented in Figure 16. This time, we observe that the noise is overlapping the signal in the frequency range 10-70 Hz. In the spectra of the denoised data and removed noise, we observe that the noise has been effectively removed when the seismic signal is preserved. These results attest that the proposed method removes noise that cannot be removed with simple frequency and/or wavenumber-based muting. They also show that the noise is highly aliased and suggest that it does not affect the proposed

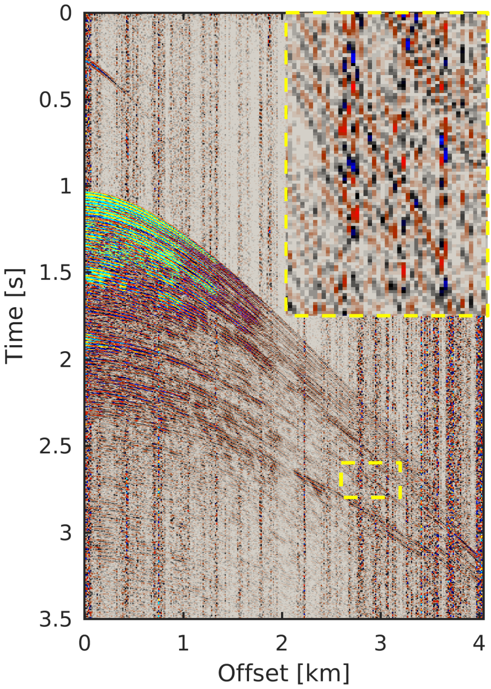

Figure 13: Shot gather contaminated by barnacle-related noise. The close up on the data framed by the yellow dashed box is shown in the top-right corner of the plot.

method. Aliasing often affects denoising methods which remove noise based on predefined dip information because aliasing leads to dip miscalculations. The proposed method, on the other hand, is not denoising based on predefined dip information.

DISCUSSION

In the proposed method, the strength of the denoising is mainly controlled by the sparsity threshold T. If the threshold is low, only the main noise features are reconstructed and separated; the amount of removed noise is hence small. On the contrary, if the threshold is high, the data is almost entirely reconstructed and separated; almost all the noise is removed. The accuracy of the signal and noise separation can be increased by simultaneously increasing the patch size, the number of patches in the training set, and the number of atoms in the dictionary. As the patch size increases, the signal and noise contents of the patches are less correlated, which improves the classification and MCA steps of the denoising process. As will be shown later in the analysis of the method complexity, increasing these parameters also increases the run time.

We shall justify the choice of the methods that were compared with the proposed method in the field data examples. The noise in these examples is not impulsive in time, so it cannot be removed with a median filter. Neither is it a white Gaussian noise, so it cannot be attenuated with conventional DL methods. It is spread on a large frequency range and has low coherency from trace to trace, so it could hardly be isolated and removed in the $f-x$, $f-k$, wavelet or curvelet domain. However, since it is not highly coherent from trace to trace, it does not appear linear in the data, and it is why

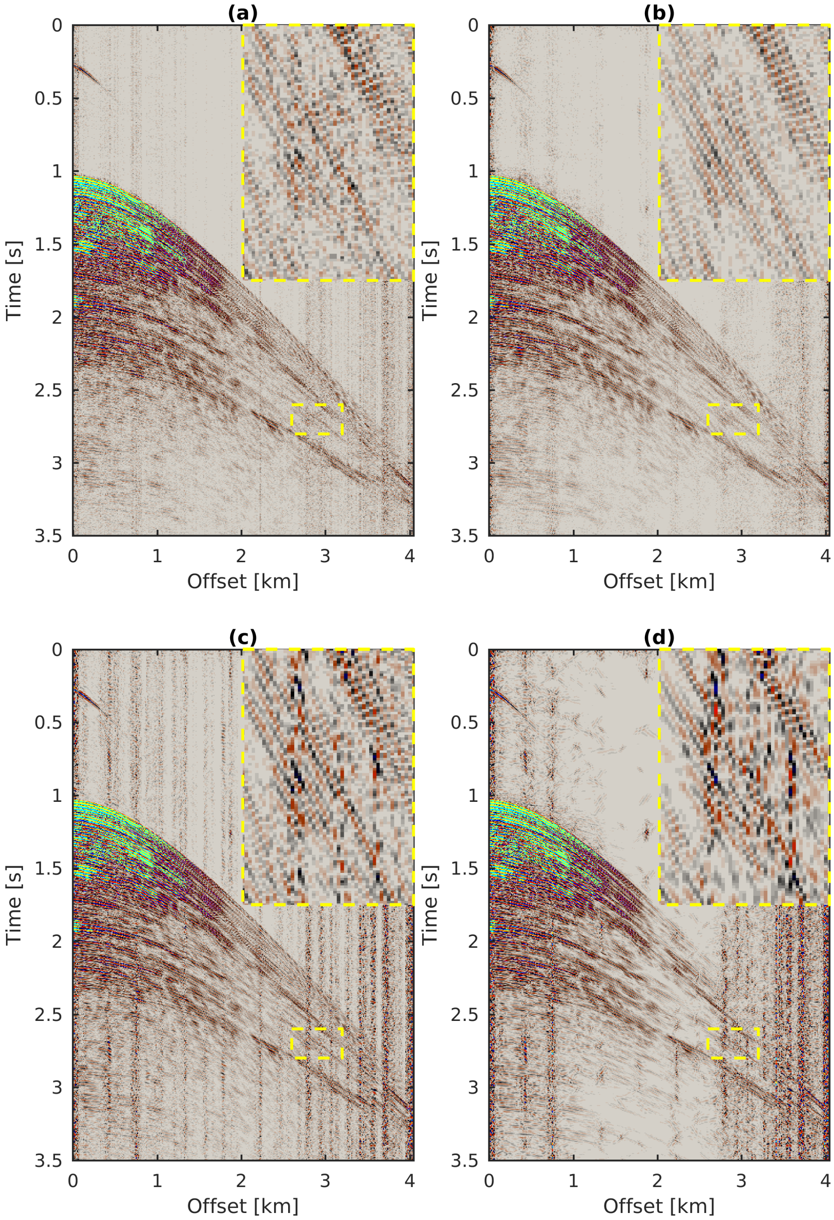

Figure 14: Denoised data that is obtained with (a) the proposed method, (b) FX-Decon, (c) FX-Cadzow, and (d) the curvelet method.

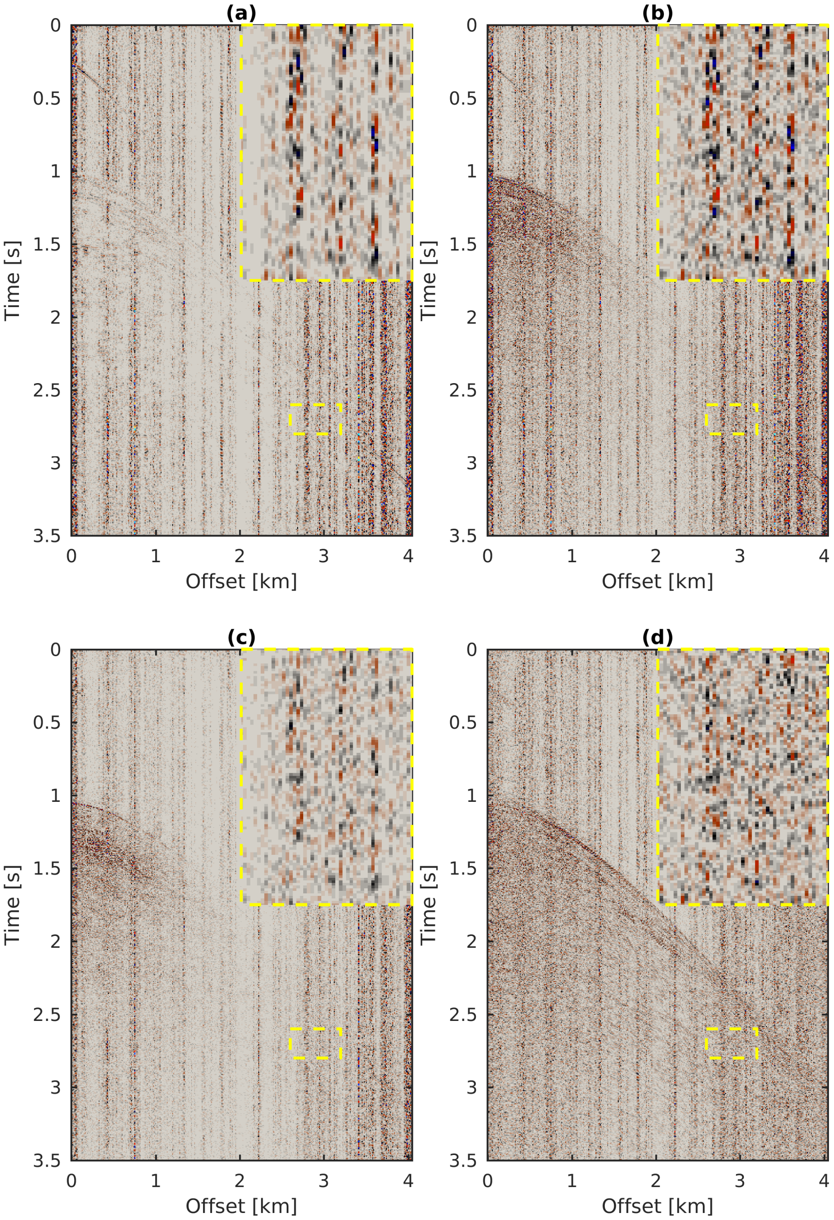

Figure 15: Noise that is removed by (a) the proposed method, (b) FX-Decon, (c) FX-Cadzow, and (d) the curvelet method.

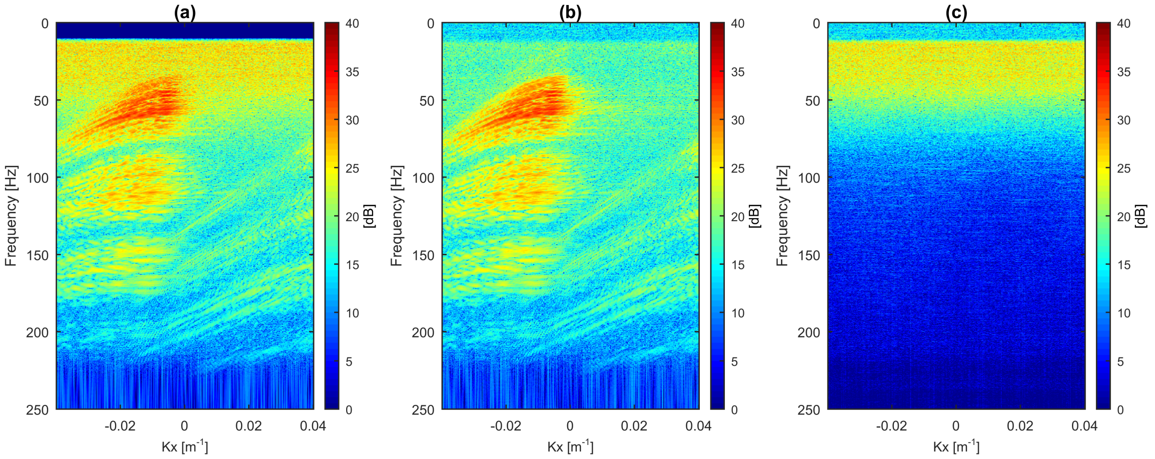

Figure 16: The $f-k$ spectra of the (a) data contaminated by barnacle-related noise, (b) denoised data obtained with the proposed method, and (c) corresponding removed noise.

it can be filtered using FX-Decon, which reconstructs only the events that are linearly predictable. Likewise, FX-Cadzow attenuates non-linear events and can be used to filter such noise. As is shown in Figures 12 and 16, the noise appears quite white within a considerable part of the $f-k$ spectrum. This indicates that it cannot be represented with few curvelets because a curvelet is localized in the f-k domain. Contrarily to that noise, the seismic wavefield is usually sparse in the curvelet domain (Hennenfent and Herrmann, 2006). Hence, the compared curvelet method is based on the following assumption: Because the signal is concentrated on few curvelet coefficients and the noise is spread on many, the signal coefficients have a high amplitude while the noise coefficients have a low amplitude. Therefore, picking only the largest coefficients may preserve the signal and attenuate the noise. The results obtained with this method suggest that the noise was too correlated with the curvelets for the method to work effectively. We note that there might be better ways to use the curvelet domain for removing such noise.

In the field data examples showed, we observe signal leakage for the proposed and compared methods. This signal leakage could be decreased by integrating additional steps to the denoising process. For instance, one could add a noise detection step (Bekara and van der Baan, 2010). This would consist in identifying the part of the data which is contaminated by noise. The identification could be in the $t-x$ or $f-x$ domain and would result in the location of the noisy samples in the $t-x$ or $f-x$ domain, correspondingly. In parallel, the denoising method would be applied to the data to get a full noise model, but only the part of the noise model that was flagged as noisy during the detection would be removed from the data. It would guaranty to preserve the signal at locations where the data are not or little noisy. Another additional step in the processing could consist in detecting eventual signal in the obtained noise model and adding it back to the denoised results. Such additional processing steps are often applied in industrial processing.

For the presented examples, the proposed method was more expensive than FX-Decon but cheaper than FX-Cadzow or the curvelet method. For instance, denoising the shot gather in Figure 10 took 8.48 min on one CPU for the proposed method when it took 1.01 min, 21.48 min, and 36.62 min for FX-Decon, FX-Cadzow, and the curvelet method, respectively. These run times were found for academic implementations of the different methods; be aware that they may not reflect run times found for industrial implementations. In the proposed method, the major part of the time is used to learn the dictionary. For this part of the method, we use the implementation of the K-SVD algorithm proposed by Rubinstein et al. (2008). The authors of this implementation evaluate the number of operations per iteration of the algorithm to approximately $M \cdot (2NK + T^2K + 7TK + T^3 + 4TN) + 5NK^2$, where M is the number of patches in the training set, N is the number of samples in a patch, K is the number of atoms in the dictionary, and T is the sparsity of each coefficient vector.

In the proposed method, we selected only three textural attributes for the classification of the atoms. They have proven to be very discriminative, and no additional attributes were needed to identify a mechanical noise pattern from a signal pattern. There is no guaranty that they would discriminate any other types of noise. This is why increasing the number of attributes for the classification would lead to a more robust method that could target many different types of noises. With more attributes, the method could be used for other applications such as removing algorithm artifacts or noise resulting from simultaneous source shooting. For an application in which there is no noise or signal model, the method could still be used with an unsupervised classification. In contrast, if both a noise and a signal model are available, the method could be applied with a supervised classification. Yet, this method would have limitations on the types of noise that it could separated from seismic data. The separation would be complete only if the morphology of the noise is different from the morphology of the signal and if the occurrence of the noise in the data is independent from the occurrence of the signal. For instance, it could not remove multiples because the atoms could not be classified as signal or noise since they would represent both, i.e., a primary and a multiple. Neither could it remove a processing artifact that would smear from the signal. This artifact would be systematically associated to the signal, and therefore, atoms that contain both the signal and the noise would be learned in the dictionary. These atoms could not be later classified as signal or noise because they would contain both.

The proposed method can be improved in several other ways. To increase the accuracy of the sparse representation, basis pursuit (Chen et al., 1998) could be used instead of OMP for solving the MCA problem. To speed up the denoising process, K-SVD could be replaced with a more efficient DL method, e.g., DDTF, or sparse K-SVD. However, we note that replacing K-SVD with sparse K-SVD could affect the effectiveness of the proposed denoising process. Sparse K-SVD learns the atoms as sparse linear combinations of fixed basis functions in contrast to K-SVD that learns unstructured atoms. Using sparse K-SVD in the proposed process would result in signal and noise dictionaries that are sparse linear combinations of the same fixed basis functions. This could increase the mutual coherence between the signal and noise dictionaries and decrease the effectiveness of MCA, which would reduce the accuracy of the signal and noise separation (Starck et al., 2004; Bruckstein et al., 2009). Finally, an extension of the method in 3D is straightforward and would benefit from a better description of the seismic wavefield.

CONCLUSION

We proposed a new sparsity promoting method for removing coherent noise from seismic data. In this method, a dictionary is learned from the data and divided into two subdictionaries; one describing the morphology of the signal and the other one describing the morphology of the noise. Then, both sub-dictionaries are used to separate the noise from the signal via MCA. For that step, the sub-dictionaries are optimal because they have been specifically trained to provide a very sparse representation of the data. Unlike conventional DL-based methods, the proposed method can remove coherent noise. In addition, this method does not require a manual search for optimal transforms that may sparsify the signal and the noise, in contrast to existing MCA-based denoising methods. We used the proposed method to remove the barnacle and steering device related noise from two field data examples and compared the results with those of FX-Decon, FX-Cadzow and the curvelet based denoising method. The proposed method provided the best results for removing the steering device-related noise. For removing the barnacle-related noise, the proposed method performed as good as FX-Decon and better than FX-Cadzow and the curvelet based denoising method.

ACKNOWLEDGMENTS

We acknowledge the SLIM Institute which provided us with the software of the curvelet based denoising method. We also thank R. Rubinstein who made available the K-SVD software that was used in the proposed method. We would like to express our gratitude to the G&E team of PGS for the fruitful discussions on the subject, and thank PGS for permission to publish this paper. Finally, we wish to thank the assistant editor J. Etgen, the associate editor I. Moore, the reviewers M. Aharchaou, N. Kazemi, and J. Ma, for valuable and constructive comments.

REFERENCES

Aharon, M., M. Elad, and A. Bruckstein, 2006, K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation: IEEE Transactions on Signal Processing, **54**, 4311–4322.

Anderson, T. W., and R. R. Bahadur, 1962, Classification into two multivariate normal distributions with different covariance matrices: The Annals of Mathematical Statistics, **33**, 420–431.

Beckouche, S., and J. Ma, 2014, Simultaneous dictionary learning and denoising for seismic data: Geophysics, **79**, A27–A31.

Bekara, M., and M. van der Baan, 2010, High-amplitude noise detection by the expectation-maximization algorithm with application to swell-noise attenuation: Geophysics, **75**, V39–V49.

Bruckstein, A. M., D. L. Donoho, and M. Elad, 2009, From sparse solutions of systems of equations to sparse modeling of signals and images: SIAM Review, **51**, 34–81.

Cai, J.-F., H. Ji, Z. Shen, and G.-B. Ye, 2014, Data-driven tight frame construction and image denoising: Applied and Computational Harmonic Analysis, **37**, 89–105.

Candès, E. J., and L. Demanet, 2005, The curvelet representation of wave propagators is optimally sparse: Communications on Pure and Applied Mathematics, **58**, 1472–1528.

Candès, E. J., and D. L. Donoho, 2000, *in* Curvelets: a surprisingly effective nonadaptive representation of objects with edges: Vanderbilt University Press, 105–120.

Chen, K., and M. D. Sacchi, 2015, Robust reduced-rank filtering for erratic seismic noise attenuation: Geophysics, **80**, V1–V11.

Chen, S. S., D. L. Donoho, and M. A. Saunders, 1998, Atomic decomposition by basis pursuit: SIAM Journal on Scientific Computing, **20**, 33–61.

Chen, Y., J. Ma, and S. Fomel, 2016, Double-sparsity dictionary for seismic noise attenuation: Geophysics, **81**, V103–V116.

Dantas, C. F., M. N. da Costa, and R. d. R. Lopes, 2017, Learning dictionaries as a sum of kronecker products: IEEE Signal Processing Letters, **24**, 559–563.

Day, A., T. Klüver, W. Söllner, H. Tabti, and D. Carlson, 2013, Wavefield-separation methods for dual-sensor towed-streamer data: Geophysics, **78**, WA55–WA70.

Engan, K., S. Aase, and J. Hakon Husoy, 1999, Method of optimal directions for frame design: Proceedings on 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2443–2446.

Fomel, S., and Y. Liu, 2010, Seislet transform and seislet frame: Geophysics, **75**, V25–V38.

Foster, D. J., C. C. Mosher, and S. Hassanzadeh, 1994, Wavelet transform methods for geophysical applications: SEG Technical Program Expanded Abstracts 1994, 1465–1468.

Gao, D., 2003, Volume texture extraction for 3D seismic visualization and interpretation: Geophysics, **68**, 1294–1302.

Gulunay, N., 1986, FX decon and complex Wiener prediction filter: Presented at the 56th Annual International Meeting, SEG, Expanded Abstracts, Session: POS2.10.

Haralick, R. M., K. Shanmugam, and I. Dinstein, 1973, Textural features for image classification: IEEE Transactions on Systems, Man, and Cybernetics, **SMC-3**, 610–621.

Hennenfent, G., and F. Herrmann, 2006, Seismic denoising with nonuniformly sampled curvelets: Computing in Science and Engineering, **8**, 16–25.

Liang, J., J. Ma, and X. Zhang, 2014, Seismic data restoration via data-driven tight frame: Geophysics, **79**, V65–V74.

Liu, Y., and S. Fomel, 2010, OC-seislet: Seislet transform construction with differential offset continuation: Geophysics, **75**, WB235–WB245.

MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, 281–297.

Mallat, S., 2008, A wavelet tour of signal processing: The sparse way, 3rd ed.: Academic Press.

Moya, M. M., and D. R. Hush, 1996, Network constraints and multi-objective optimization for one-class classification: Neural Networks, **9**, 463 – 474.

Neelamani, R., A. I. Baumstein, D. G. Gillard, M. T. Hadidi, and W. L. Soroka, 2008, Coherent and random noise attenuation using the curvelet transform: The Leading Edge, **27**, 240–248.

Oropeza, V. E., 2010, The singular spectrum analysis method and its application to seismic data denoising and reconstruction: Master's thesis, University of Alberta, Edmonton, Alberta.

Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad, 1993, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition: Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers, 40–44.

Rubinstein, R., M. Zibulevsky, and M. Elad, 2008, Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit.

Rubinstein, R., M. Zibulevsky, and M. Elad, 2010, Double sparsity: Learning sparse dictionaries for sparse signal approximation: IEEE Transactions on Signal Processing, **58**, 1553–1564.

Sacchi, M. D., 2009, FX singular spectrum analysis: Presented at the CSPG CSEG CWLS Convention.

Starck, J.-L., M. Elad, and D. Donoho, 2004, Redundant multiscale transforms and their application for morphological component separation: Advances in Imaging and Electron Physics, **132**, 287–348.

Starck, J.-L., Y. Moudden, J. Bobin, M. Elad, and D. Donoho, 2005, Morphological component analysis: Wavelets XI, 209–223.

Tax, D., 2001, One-class classification: concept learning in the absence of counterexamples: PhD thesis, Technische Universiteit Delft.

Trickett, S., 2002, F-x eigen noise suppression: Presented at the CSEG Geophysics.

Trickett, S., 2008, F-xy Cadzow noise suppression: SEG Technical Program Expanded Abstracts 2008, 2586–2590.

Turquais, P., E. Asgedom, and W. Söllner, 2016, Sparsity promoting morphological decomposition for coherent noise suppression: Application to streamer vibration related noise: SEG Technical Program Expanded Abstracts 2016, 4639–4643.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017, A method of combining coherence-constrained sparse coding and dictionary learning for denoising: Geophysics, **82**, V137–V148.

Ulrych, T. J., and M. D. Sacchi, 2005, Information-based inversion and processing with applications, *in* Handbook of geophysical exploration, 1st ed.: Elsevier.

Vaezi, Y., and N. Kazemi, 2016, Attenuation of swell noise in marine streamer data via nonnegative matrix factorization: SEG Technical Program Expanded Abstracts 2016, 4633–4638.

Vinther, R., 1997, Seismic texture classification applied to processed 2D and 3D seismic data: SEG Technical Program Expanded Abstracts 1997, 721–724.

Vinther, R., K. Mosegaard, K. Kierkegaard, I. Abatzis, C. Andersen, O. V. Vejbaek, F. If, and P. H. Nielsen, 1995, Seismic texture classification: A computeraided approach to stratigraphic analysis: SEG Technical Program Expanded Abstracts 1995, 153–155.

Wang, W., W. Chen, J. Lei, and J. Gao, 2010, Ground roll separation by sparsity and morphological diversity promotion: SEG Technical Program Expanded Abstracts 2010, 3705–3710.

West, B. P., S. R. May, J. E. Eastwood, and C. Rossen, 2002, Interactive seismic facies classification using textural attributes and neural networks: The Leading Edge, **21**, 1042–1049.

Yu, M. C., 2011, Seismic interference noise elimination a multidomain 3d filtering approach: SEG Technical Program Expanded Abstracts 2011, 3591–3595.

Yu, S., J. Ma, and S. Osher, 2016, Monte Carlo data-driven tight frame for seismic data recovery: Geophysics, **81**, V327–V340.

Yu, S., J. Ma, X. Zhang, and M. D. Sacchi, 2015, Interpolation and denoising of high-dimensional seismic data by learning a tight frame: Geophysics, **80**, V119–V132.

Zhu, L., E. Liu, and J. H. McClellan, 2015, Seismic data denoising through multiscale and sparsity-promoting dictionary learning: Geophysics, **80**, WD45–WD57.

Chapter 6

Article III

The third article is entitled "Parabolic dictionary learning for seismic wavefield reconstruction across the streamers". It was submitted to the journal Geophysics on the 20$^{\text{th}}$ of September 2017 for publication. The page number located in the header of each page of the article except the first one is relative to the article and is starting from A302, whereas the page number relative to the thesis is located in the footer of the page.

Chapter 7

Conclusions and Outlook

7.1 Conclusions

First, I analyzed sparse representation-based methods for random noise attenuation, signal separation, and signal reconstruction, and I assessed their effectiveness using simple experiments that were tailored to control the sparsity level of the signal in the dictionary domain. I made the following observations:

- A sparse approximation of the recording attenuates the random noise, and the smaller the ℓ_0-norm of the signal representation is, the smaller is the error (see Figure 2.3).
- A sparse representation of the recording in a domain that comprises signal and noise subdomains can be used to separate the coherent noise from the signal (see equation 2.11). If the ℓ_0-norm of the representation of the recording in the transform domain is under a threshold dictated by the dictionary (see equation 2.4), then the signal and noise separation is exact. Above this threshold, the error increases as the ℓ_0-norm of the representation increases (see Figure 2.6).
- A sparse representation of the recording can be used to reconstruct missing samples of the signal (see equation 2.13). If the ℓ_0-norm of the fully sampled signal in the transform domain is below a threshold dictated by the sampling scheme and the dictionary, then the reconstruction of the signal is exact. Above this threshold, the error increases as the ℓ_0-norm of the representation increases (see Figure 2.7).

I studied predefined dictionaries that define a domain in which the seismic data may have a sparse representation. A particular attention was given to the Fourier bases, the DWT bases, and the curvelet frames. I investigated their capability to represent a wavefield and I tested on a field data example if they can be used to compute a sparse approximation of the seismic data that preserves the signal to a high degree (see Figures 2.11, 2.12, and 2.13). The investigation showed that the wavelets of a DWT dictionary have a limited capability to describe a wavefield due to their isotropic nature. Besides, the sparse approximation of the seismic data in the DWT domain distorted significantly the signal. The basis functions of a Fourier base are suited to describe a wavefield, as they are solutions to the wave equation in a homogeneous medium. In addition, the sparse approximation of the seismic data in the Fourier domain represented a large part of the signal. The curvelets are also capable to concisely represent a wavefield because a solution of the wave equation was proven to be sparse in the curvelet domain (Candès and Demanet, 2005). From the three sparse approximations that were based on predefined dictionaries, the curvelet-based sparse approximation was the closest to the original data. Yet, a minor part of the seismic signal was not represented in the sparse approximation. Since predefined dictionaries do not adapt to the data, they have a limited effectiveness to represent the complex seismic events in a sparse manner.

I then investigated DL methods. DL locally adapts the dictionary to the data without requiring human interaction. This is quite appealing for seismic data application because seismic data are large, high-dimensional, and contain a signal whose morphology varies across the data. In addition, I observed that DL was more effective to provide a sparse representation of the seismic data compared to the Fourier base and the curvelet frame; for the same compression ratio, it provided a representation that was more accurate. However, for random noise attenuation, coherent signal separation, and seismic data reconstruction, I found that the conventional DL methods have the following limitations:

(i) To attenuate the random noise, conventional DL methods need the a priori variance of the noise. This is inconvenient because the random noise in the seismic data has a variance that is often unknown and spatiotemporaly varying.

(ii) When DL is applied to data contaminated by noise that has some spatial or temporal coherency, the learned dictionary contains both atoms representing the signal morphology and atoms representing the noise morphology. Hence, a sparse approximation of the signal in the dictionary domain does not remove the noise.

(iii) The learned dictionaries do not have analytical expressions, they are numerically defined. Due to this particularity, a sparse representation in the dictionary domain cannot be used to interpolate the data over an arbitrary grid.

To overcome the limitation in (i), Turquais et al. (2017c) developed the coherence-constrained dictionary learning (CDL) method. The CDL method learns the dictionary and attenuates the noise using a coherence-based constraint such that the noise variance is not needed. Yet, CDL is ideal to filter out Gaussian noise. If the seismic data are contaminated by Gaussian noise of constant variance, CDL is as efficient as the conventional DL method that uses the knowledge of the noise variance. Furthermore, if the seismic data are contaminated by Gaussian noise of spatiotemporally varying variance, CDL provides a better denoising compared to conventional DL. In both cases in which the noise variance is constant and spatiotemporally varying, the CDL method is more effective than the industry-standard FX-Decon denoising method (Canales, 1984; Gulunay, 1986).

To overcome the limitation in (ii), Turquais et al. (2017a) combined DL, MCA, and a statistical classification. DL combined with the statistical classification uses the noise contaminated data to learn a dictionary that defines a domain in which the signal is sparse, and a dictionary that defines a domain in which the noise is sparse. Then, MCA uses the two dictionaries to separate the signal and the noise. Because the signal and the noise dictionaries are learned from the data, MCA is applied under optimal conditions. The method can be used to separate coherent noise from the signal as long as the morphology of the noise is different from the morphology of the signal and the occurrence of the noise in the data is independent from the occurrence of the signal. Field data examples showed that the proposed method is more effective than FX-Decon (Canales, 1984; Gulunay, 1986), FX-Cadzow (Trickett, 2002), and the curvelet denoising method (Hennenfent and Herrmann, 2006) to remove mechanical noise.

To overcome the limitation in (iii), Turquais et al. (2017f) developed a parabolic dictionary learning (PDL) method. In this method, each learned atom is constrained to represent an elementary waveform that has a constant amplitude along a parabolic traveltime moveout characterized by kinematic wavefield parameters. Imposing such a parabolic structure is quite appropriate for seismic data application because a wavefield can be locally approximated by a superposition of parabolic events (see equations 2.16, 2.17, and 2.18). In addition, it can be used to easily interpolate or extrapolate the atoms. Using this advantage, the method can interpolate and regularize the seismic data. Benefiting from the parabolic structure, the sparsity promotion, and the data adaptation, the PDL method is able to operate beyond aliasing. Synthetic and field data examples validated that PDL can interpolate the recorded seismic data between the streamers of typical 3D acquisition configurations, and hence to reconstruct the 3D seismic wavefield.

7.2 Outlook

7.2.1 Assessing the impact of the proposed methods on the final image

The next step in my research will be to assess the impact of the proposed methods on the migrated image. Assessing this impact is crucial because migrated images are the data that are interpreted. Although it has been shown that denoising, regularization, and interpolation improve the image (Elboth et al., 2008, 2009a; Herrmann, 2010; Zhang and Wang, 2015; Charles et al., 2014; Mosher et al., 2017), it is important to quantify the image quality uplift that results from applying each of the proposed methods.

First, for each application, an appropriate data set will need to be selected. A data set that is particularly noisy could be a good choice to test if the proposed denoising methods are effective where conventional methods fail. To test the proposed wavefield reconstruction method, a shallow water data set, or a data set acquired in a complex geological environment, would be preferred. Such a data set is more likely to contain aliased energy, and the processing to suffer from a poor spatial sampling. After selection of the data set, the impact on the final image of applying the proposed methods could be assessed by comparing the images obtained with the three following workflows: (1) The industry-standard seismic processing and imaging workflow; (2) The industry-standard seismic processing and imaging workflow in which the standard noise attenuation / wavefield reconstruction process was removed; (3) The industry-standard seismic processing and imaging workflow in which the standard noise attenuation / wavefield reconstruction process was replaced with the proposed method. Comparing the resolution of the image from the workflow 3 with the image from workflow 1 will reveal the improvement resulting from applying the proposed method in comparison to the standard method. Also, computing the differences between the images resulting from workflows 2 and 3 would point out a signal attenuation or would validate that the signal is untouched.

7.2.2 Multi-scale dictionary learning and processing

As proposed by Ophir et al. (2011), and implemented for seismic data denoising by Zhu et al. (2015), the DL methods proposed by Turquais et al. (2017c,f) could be integrated into a scheme in which DL and denoising/reconstruction is applied independently to each scale of the data. The 1D DWT would be used to separate the seismic data into different scales containing different temporal frequency subbands of the signal. Then, DL and denoising/reconstruction would be applied independently to each scale, and the results at the different scales would be combined using the inverse DWT. Such a process has two main advantages. First, it preserves the integrity of the frequency spectrum of the data. In some cases, using DL methods directly on data may lead to partial attenuation of one of the frequency subbands of the signal. This is due to the ℓ_2-norm minimization of the data misfit in the cost function that is used to learn the dictionary and to denoise/reconstruct the data. When one frequency subband of the signal has a very small ℓ_2-norm compared with another subband of the data, the cost function places a higher focus toward representing the subband having the largest ℓ_2-norm. This leads to the least squared error, but also to an error that is unequally spread over the spectrum. In contrast, when the DL method is applied independently to each scale, as much atoms are learned to represent each scale, which ensures an equal focus among the frequency subbands of the signal. As a second advantage of multi-scale processing, some parameters, e.g., the size of the patches, can be adjusted to each scale. When using DL methods, it is advised to set the size of the patches to the size of the elementary patterns in the data. The low frequency elementary patterns are typically larger than the high frequency elementary patterns. Using a multi-scale scheme enables setting a larger patch size to process the scales containing the low frequency signal, and a smaller patch size to process the scales containing the high frequency signal.

It would also be interesting to implement a multi-scale approach in which the dictionaries at different scales are learned with a cross-scale cooperative learning (Chen and Chau, 2015). The dictionaries would be learned such that the atoms from the dictionary at one scale would have the same shape as the atoms from the dictionaries at the other scales, but would have different time-space scales. For instance, an atom at a given scale would represent a pattern that has the same shape as the corresponding atom of a subsequent finer scale, but would be twice longer in time and twice larger in space, such that it would represent a signal of lower frequency and wavenumber. In contrast to the multi-scale scheme without collaboration, the cross-scale cooperative learning approach could learn information at a given scale and use it to represent a signal at another scale. Such a characteristic could be advantageous for interpolation. An atom characterizing the kinematic of the wavefield could be learned from the well sampled low frequency signal and be used to reconstruct the aliased higher frequency signal.

7.2.3 DL and processing in other dimensions

It would be straightforward to apply the proposed methods in other dimensions. This could be used to resolve other seismic processing challenges. For instance, the PDL method (Turquais et al., 2017f) could be applied to common receiver gathers instead of common shot gathers. A common receiver gather is a set of traces that have been recorded for the same receiver position and for different source positions. As a common shot gather, a common receiver gather contains a wavefield that can be locally described by a superposition of elementary waves whose traveltimes follow parabolic moveouts (Hubral, 1983; Bortfeld, 1989; Červený, 2001). Therefore, it is reasonable to assume that PDL would interpolate in the common receiver gathers with the same accuracy as it interpolates in the common shot gathers. It would increase the spatial sampling of the common receiver gathers, which would increase the source sampling of the seismic data set. This would be valuable because the seismic data are acquired with a coarse source spacing such that they are aliased is the source dimension. This aliasing is an issue for processing methods that are applied to common receiver gathers, e.g., source deghosting methods.

7.2.4 DL and processing in more dimensions

The proposed methods could be extended to more dimensions. Extending the methods in more dimensions would consist to learn cubes or higher-dimensional vertices, and correspondingly, to process cubes or vertices from the data (Yu et al., 2015). For instance, a 3D extension of PDL could be used to learn cubes of dimensions time, crossline offset, inline offset, and to simultaneously interpolate in the inline and crossline dimensions of common shot gathers. Such a method could use information learned in the inline direction to interpolate in the crossline direction, and vice versa. Hence, the 3D extension is expected to improve the interpolation. On the other hand, the process would have a higher computational cost compared with using the 2D PDL method to first interpolate in the inline direction and then interpolate in the crossline direction. The quality uplift would need to be quantified to judge if it would justify the additional computational effort.

7.2.5 Computational optimization

The proposed methods were in general computationally demanding. For instance, using one CPU, the wavefield reconstruction method took 49 min per 3D shot gather (Turquais et al., 2017f), and the coherent noise separation method took 8.48 min to process one inline shot gather (Turquais et al., 2017a). A computational optimization is required to use the methods on a standard basis in seismic processing. A modification of the DL process could reduce the computational cost of the methods. The process used is iterative and alternates over a sparse coding stage and a dictionary update stage. The dictionary update stage is the most time-consuming stage. To update the dictionary, we have used SVDs or approximate SVDs. The computational cost could be reduced by using the sequential generalization of K-means (SGK) instead (Sahoo and Makur, 2013).

The computational optimization could be simultaneously tackled on the software and hardware sides. The proposed methods were coded using dynamic types of languages and have run on CPUs.

They could be rewritten using a static language, which is generally several times faster. In addition, the codes could be run on GPUs, which can be 10-15 times faster for tasks that require a lot of computing power but relatively few memories.

Bibliography

Abma, R., and N. Kabir, 2005, 3d interpolation of irregular data with a pocs algorithm: SEG Technical Program Expanded Abstracts 2005, 2150–2153.

Aharon, M., M. Elad, and A. Bruckstein, 2006, k-svd: An algorithm for designing overcomplete dictionaries for sparse representation: IEEE Transactions on Signal Processing, **54**, 4311–4322.

Aki, K., and P. G. Richards, 2002, Quantitative seismology, 2nd ed.: University Science Books.

Andrade, L., G. Hoecht, E. Landa, and S. Spitz, 2005, QC of a marine seismic trace reconstruction technique: SEG Technical Program Expanded Abstracts 2005, 2166–2169.

Beckouche, S., and J. Ma, 2014, Simultaneous dictionary learning and denoising for seismic data: Geophysics, **79**, A27–A31.

Berman, S. M., 1964, Limit theorems for the maximum term in stationary sequences: The Annals of Mathematical Statistics, **35**, 502–516.

Bortfeld, R., 1989, Geometrical ray theory: Rays and traveltimes in seismic systems (second-order approximations of the traveltimes): Geophysics, **54**, 342–349.

Bruckstein, A. M., D. L. Donoho, and M. Elad, 2009, From sparse solutions of systems of equations to sparse modeling of signals and images: SIAM Review, **51**, 34–81.

Cai, J.-F., H. Ji, Z. Shen, and G.-B. Ye, 2014, Data-driven tight frame construction and image denoising: Applied and Computational Harmonic Analysis, **37**, 89–105.

Canales, L., 1984, Randon noise reduction: 54th Annual International Meeting, SEG, Expanded Abstracts, 525–527.

Candès, E. J., and L. Demanet, 2005, The curvelet representation of wave propagators is optimally sparse: Communications on Pure and Applied Mathematics, **58**, 1472–1528.

Candès, E. J., and D. L. Donoho, 2000, Curvelets: a surprisingly effective nonadaptive representation of objects with edges: Curve and surface fitting, Vanderbilt University Press, Vanderbilt University Press, 105–120.

Candès, E. J., and D. L. Donoho, 2002, Recovering edges in ill-posed inverse problems: optimality of curvelet frames: Ann. Statist., **30**, 784–842.

Candes, E. J., and M. B. Wakin, 2008, An introduction to compressive sampling: IEEE Signal Processing Magazine, **25**, 21–30.

Červený, V., 2001, Seismic ray theory: Cambridge University Press.

Charles, M., L. Chengbo, M. Larry, J. Yongchang, J. Frank, O. Robert, and B. Joel, 2014, Increasing the efficiency of seismic data acquisition via compressive sensing: The Leading Edge, **33**, 386–391.

Chen, J., and L. P. Chau, 2015, Multiscale dictionary learning via cross-scale cooperative learning and atom clustering for visual signal processing: IEEE Transactions on Circuits and Systems for Video Technology, **25**, 1457–1468.

Chen, S. S., D. L. Donoho, and M. A. Saunders, 1998, Atomic decomposition by basis pursuit: SIAM Journal on Scientific Computing, **20**, 33–61.

Chen, Y., J. Ma, and S. Fomel, 2016, Double-sparsity dictionary for seismic noise attenuation: Geophysics, **81**, V103–V116.

Daubechies, I., M. Defrise, and C. De Mol, 2004, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint: Communications on Pure and Applied Mathematics, **57**, 1413–1457.

Davis, G., S. Mallat, and M. Avellaneda, 1997, Adaptive greedy approximations: Constructive Approximation, **13**, 57–98.

Donoho, D. L., 1995, De-noising by soft-thresholding: IEEE Transactions on Information Theory, **41**, 613–627.

Donoho, D. L., 2006, Compressed sensing: IEEE Transactions on Information Theory, **52**, 1289–1306.

Donoho, D. L., and M. Elad, 2003, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization: Proceedings of the National Academy of Sciences of the United States of America, **100**, 2197–2202.

Donoho, D. L., and J. M. Johnstone, 1994, Ideal spatial adaptation by wavelet shrinkage: Biometrika, **81**, 425–455.

Donoho, D. L., Y. Tsaig, I. Drori, and J. L. Starck, 2012, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit: IEEE Transactions on Information Theory, **58**, 1094–1121.

Elad, M., 2006, Why simple shrinkage is still relevant for redundant representations?: IEEE Transactions on Information Theory, **52**, 5559–5569.

Elad, M., 2010, Sparse and redundant representations: From theory to applications in signal and image processing, 1st ed.: Springer Publishing Company, Incorporated.

Elad, M., P. Milanfar, and R. Rubinstein, 2007, Analysis versus synthesis in signal priors: Inverse Problems, **23**, 947.

Elboth, T., 2010, Noise in marine seismic data: PhD thesis, University of Oslo.

Elboth, T., F. Geoteam, and D. Hermansen, 2009a, Attenuation of noise in marine seismic data: SEG Technical Program Expanded Abstracts 2009, 3312–3316.

Elboth, T., F. Geoteam, H. H. Qaisrani, and T. Hertweck, 2008, De-noising seismic data in the time-frequency domain: SEG Technical Program Expanded Abstracts 2008, 2622–2626.

Elboth, T., B. A. Reif, and Ø. Andreassen, 2009b, Flow and swell noise in marine seismic data: Geophysics, **74**, Q17–Q25.

Engan, K., S. Aase, and J. Hakon Husoy, 1999, Method of optimal directions for frame design: Proceedings on 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2443–2446.

Fomel, S., and Y. Liu, 2010, Seislet transform and seislet frame: Geophysics, **75**, V25–V38.

Foster, D. J., C. C. Mosher, and S. Hassanzadeh, 1994, Wavelet transform methods for geophysical applications: SEG Technical Program Expanded Abstracts 1994, 1465–1468.

Gao, J., A. Stanton, M. Naghizadeh, M. D. Sacchi, and X. Chen, 2013, Convergence improvement and noise attenuation considerations for beyond alias projection onto convex sets reconstruction: Geophysical Prospecting, **61**, 138–151.

Gorodnitsky, I. F., and B. D. Rao, 1997, Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm: IEEE Trans. Signal Processing, 600–616.

Gulunay, N., 1986, FX DECON and complex Wiener prediction filter: Presented at the 56th Annual International Meeting, SEG, Expanded Abstracts, Session: POS2.10.

Haar, A., 1910, Zur theorie der orthogonalen funktionensysteme: Mathematische Annalen, **69**, 331–371.

Hennenfent, G., and F. Herrmann, 2006, Seismic denoising with nonuniformly sampled curvelets: Computing in Science and Engineering, **8**, 16–25.

Herrmann, F. J., 2010, Randomized sampling and sparsity: Getting more information from fewer samples: Geophysics, **75**, WB173–WB187.

Herrmann, P., T. Mojesky, M. Magesan, and P. Hugonnet, 2005, De-aliased, high-resolution radon transforms: SEG Technical Program Expanded Abstracts 2000, 1953–1956.

Hoecht, G., P. Ricarte, S. Bergler, and E. Landa, 2009, Operator-oriented crs interpolation: Geophysical Prospecting, **57**, 957–979.

Hubral, P., 1983, Computing true amplitude reflections in a laterally inhomogeneous earth: Geophysics, **48**, 1051–1062.

Ibrahim, A., M. D. Sacchi, and P. Terenghi, 2015, Wavefield reconstruction using a stolt-based asymptote and apex shifted hyperbolic radon transform: SEG Technical Program Expanded Abstracts 2015, 3836–3841.

Kovacevic, J., and A. Chebira, 2007, Life beyond bases: The advent of frames (part i): IEEE Signal Processing Magazine, **24**, 86–104.

Kumar, R., H. Wason, S. Sharan, and F. J. Herrmann, 2017, Highly repeatable 3d compressive full-azimuth towed-streamer time-lapse acquisition – a numerical feasibility study at scale: The Leading Edge, **36**, 677–687.

Liang, J., J. Ma, and X. Zhang, 2014, Seismic data restoration via data-driven tight frame: Geophysics, **79**, V65–V74.

Liu, Y., and S. Fomel, 2010, oc-seislet: Seislet transform construction with differential offset continuation: Geophysics, **75**, WB235–WB245.

Mallat, S., 2008, A wavelet tour of signal processing: The sparse way, 3rd ed.: Academic Press.

Mosher, C., C. Li, Y. Ji, F. D. Janiszewski, B. Bankhead, L. Williams, J. Hand, and J. Anderson, 2017, Compressive seismic imaging: Moving from research to production: SEG Technical Program Expanded Abstracts 2017, 74–78.

Naghizadeh, M., and M. D. Sacchi, 2010, On sampling functions and fourier reconstruction methods: Geophysics, **75**, WB137–WB151.

Needell, D., and J. Tropp, 2009, Cosamp: Iterative signal recovery from incomplete and inaccurate samples: Applied and Computational Harmonic Analysis, **26**, 301 – 321.

Neelamani, R., A. I. Baumstein, D. G. Gillard, M. T. Hadidi, and W. L. Soroka, 2008, Coherent and random noise attenuation using the curvelet transform: The Leading Edge, **27**, 240–248.

Ophir, B., M. Lustig, and M. Elad, 2011, Multi-scale dictionary learning using wavelets: IEEE Journal of Selected Topics in Signal Processing, **5**, 1014–1024.

Pati, Y. C., R. Rezaiifar, and P. S. Krishnaprasad, 1993, Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition: Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers, 40–44.

Rubinstein, R., 2011, Analysis and synthesis sparse modeling methods in image processing: PhD thesis, Senate of the Technion – Israel Institute of Technology.

Rubinstein, R., M. Zibulevsky, and M. Elad, 2010, Double sparsity: Learning sparse dictionaries for sparse signal approximation: IEEE Transactions on Signal Processing, **58**, 1553–1564.

Sahoo, S. K., and A. Makur, 2013, Dictionary training for sparse representation as generalization of k-means clustering: IEEE Signal Processing Letters, **20**, 587–590.

Schleicher, J., M. Tygel, and P. Hubral, 1993, 3-d true-amplitude finite-offset migration: Geophysics, **58**, 1112–1126.

Schonewille, M., A. Klaedtke, and A. Vigner, 2009, Anti-alias anti-leakage fourier transform: SEG Technical Program Expanded Abstracts 2009, 3249–3253.

Shannon, C. E., N. J. A. Sloane, and A. D. Wyner, 1993, Claude Elwood Shannon: collected papers, 1st ed.: Wiley-IEEE Press.

Starck, J.-L., M. Elad, and D. Donoho, 2004, Redundant multiscale transforms and their application for morphological component separation: Advances in Imaging and Electron Physics, **132**, 287–348.

Starck, J.-L., Y. Moudden, J. Bobin, M. Elad, and D. Donoho, 2005, Morphological component analysis: Wavelets XI, 209–223.

Trickett, S., 2002, F-x eigen noise suppression: Presented at the CSEG Geophysics.

Tropp, J., 2004, Greed is good: Algorithmic results for sparse approximation: IEEE Transactions on Information Theory, **50**, 2231–2242.

Turquais, P., E. G. Asgedom, and W. Söllner, 2016, Sparsity promoting morphological decomposition for coherent noise suppression: Application to streamer vibration related noise: SEG Technical Program Expanded Abstracts 2016, 4639–4643.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017a, Coherent noise suppression by learning and analyzing the morphology of the data: Geophysics, **82**, V397–V411.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017b, Denoising seismic data: U.S. Patent 2017,0108,604 A1.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017c, A method of combining coherence-constrained sparse coding and dictionary learning for denoising: Geophysics, **82**, V137–V148.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017d, Structured dictionary learning for interpolation of aliased seismic data: SEG Technical Program Expanded Abstracts 2017, 4257–4261.

Turquais, P., E. G. Asgedom, and W. Söllner, 2017e, Structured dictionary learning for interpolation of aliased seismic data: Invention disclosure PGS-17116US.

Turquais, P., E. G. Asgedom, W. Söllner, and L.-J. Gelius, 2017f, Parabolic dictionary learning for seismic wavefield reconstruction across the streamers: Submitted to Geophysics.

Turquais, P., E. G. Asgedom, W. Söllner, and E. Otnes, 2015, Dictionary learning for signal-to-noise ratio enhancement: SEG Technical Program Expanded Abstracts 2015, 4698–4702.

Ursin, B., 1982, Quadratic wavefront and traveltime approximations in inhomogeneous layered media with curved interfaces: Geophysics, **47**, 1012–1021.

Yu, S., J. Ma, X. Zhang, and M. D. Sacchi, 2015, Interpolation and denoising of high-dimensional seismic data by learning a tight frame: Geophysics, **80**, V119–V132.

Zhang, Y., S. Bergler, and P. Hubral, 2001, Common-reflection-surface (CRS) stack for common offset: Geophysical Prospecting, **49**, 709–718.

Zhang, Z., and P. Wang, 2015, Seismic interference noise attenuation based on sparse inversion: SEG Technical Program Expanded Abstracts 2015, 4662–4666.

Zhu, L., E. Liu, and J. H. McClellan, 2015, Seismic data denoising through multiscale and sparsity-promoting dictionary learning: Geophysics, **80**, WD45–WD57.

Zwartjes, P. M., 2005, Fourier reconstruction with sparse inversion: PhD thesis, Delft University.

Zwartjes, P. M., and M. D. Sacchi, 2007, Fourier reconstruction of nonuniformly sampled, aliased seismic data: Geophysics, **72**, V21–V32.