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

Pierre Turquais¹, Endrias G. Asgedom², and Walter Söllner²

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. We have developed 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-sized 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 it 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 few decades, 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.

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 (Candés 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 using a dictionary learning (DL) method. These methods train a dictionary to be optimally adapted for representing the signal in a sparse manner. DL methods include the method of optimal direction (MOD) (Engan et al., 1999), K-means singular value decomposition (K-SVD) (Aharon et al., 2006), and the data-driven tight frame (DDTF)
method (Cai et al., 2014). Otherwise, a double-sparcity DL method (Rubinstein et al., 2010) reconciles the fixed 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 to the part of the recording that correlates best with atoms of the dictionary. Therefore, the quality of the 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 nonspace methods (Foster et al., 1994). Curvelets can efficiently model the geometry of waveforms (Candés and Demanet, 2005), and they have proven to be some of the most suitable predefined dictionaries 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 a 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 reconstruction schemes (Liang et al., 2014; Yu et al., 2015, 2016). Also, double-sparcity DL methods benefit from taking a data-driven approach and integrating priori 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, which 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 the 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 $y \in \mathbb{R}^N$ containing signal of interest $\mathbf{x} \in \mathbb{R}^K$ and white Gaussian noise $\mathbf{n} \in \mathbb{R}^N$ of zero mean and $\sigma^2$ variance. Formally, this data model is given by

$$\mathbf{z} = \mathbf{y} + \mathbf{n}. \quad (1)$$

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 ($\mathbf{D} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K] \in \mathbb{R}^{N \times K}$), 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

$$\mathbf{y} = \mathbf{D}\mathbf{x}. \quad (2)$$

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 $\ell_0$-norm (Donoho et al., 2006). This problem can be formally expressed as

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

where the representation error threshold $\epsilon$ is dictated by the noise variance $\sigma$. After solving this problem, the sparse approximation of the recording is computed by

$$\hat{\mathbf{y}} = \mathbf{D}\hat{\mathbf{x}}. \quad (4)$$

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

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

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
In coherent denoising, the sparsity is constrained by the coherence of the residual vector \( r \) relative to the dictionary \( D \). This coherence measure is denoted by \( \mu(r, D) \) and mathematically given by
\[
\mu(r, D) = \max_j \left| \frac{r^T}{\|r\|_2} a_j \right|.
\] (6)

where \( a_j \) 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., \( r = z - Dx \)). Hence, coherent denoising aims to solve the problem
\[
\hat{x} = \arg \min_{x} \|x\|_0 \quad \text{subject to} \quad \mu(z - Dx, D) \leq \mu, \quad (7)
\]
where \( \mu \) is a fixed coherence threshold. Finding an approximation of a recording, which uses a minimal number of atoms from a redundant dictionary and satisfies \( \mu(z - Dx, D) \leq \mu \), is a problem called “nondeterministic 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 the coherent matching pursuit algorithm (Mallat, 2008, pp. 656–659).

The coherent matching pursuit algorithm uses the iterative greedy scheme of the orthogonal matching pursuit (OMP) (Pati et al., 1993), but it 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 \( \mu(r, D) \) reduces to become less than the threshold \( \mu \), and it 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 \( \sigma^2 \) when used with thresholding over a wavelet base. Following the same track, we show in Appendix B that the threshold
\[
\mu_{\text{ideal}} = \sqrt{2 \log(K) \over N} \quad (8)
\]
is ideal for filtering out white Gaussian noise by coherent denoising subject to a redundant dictionary \( D \in \mathbb{R}^{N \times K} \). 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 \( D \) was a matrix of size 100 x 100 and its atoms were zero-mean unit vectors with identically and independently distributed Gaussian entries. The sparse coefficient vector \( x \) contained \( L \) nonzero coefficients. The position of the nonzero coefficients was randomly chosen, and their values were fixed to the same value \( \alpha \). Then, a recording \( z \) was synthesized by adding random noise \( n \) with variance \( \sigma^2 \) as in equation 1. We will use \( \hat{y}_{\text{err}} \) and \( \hat{y}_{\text{car}} \) 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, \( \hat{y}_{\text{coh}} \) will refer to the coherent matching pursuit approximate solution to the problem in equation 5, for the threshold \( \mu_{\text{ideal}} \). The sparse approximations \( \hat{y}_{\text{coh}}, \hat{y}_{\text{err}}, \) and \( \hat{y}_{\text{car}} \) of the recording \( z \) were computed to recover the signal of interest. Because \( n \) 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
\[
E = \frac{\|y - \hat{y}\|^2_2}{\|n\|^2_2}. \quad (9)
\]
The quantity \( E \) assesses the error reduction; the closer it is to zero, the greater is the error reduction. Note that \( E \) is independent of the amplitude of the recording, unlike the mean-squared error.

The results of this experiment are presented in Figure 1 for \( \sigma/\sigma \) 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 \( \sigma/\sigma, \hat{y}_{\text{car}} \) provides the lowest \( E \), and for larger values, \( \hat{y}_{\text{car}} \) performs better. However, \( E \) does not differ significantly for the three sparse approximations. Hence, this experiment demonstrates that \( \hat{y}_{\text{coh}} \) can perform denoising similar to that of \( \hat{y}_{\text{car}} \) or \( \hat{y}_{\text{err}} \), without the knowledge of the cardinality of the solution or the noise variance. For the range of parameters studied, \( E \) of \( \hat{y}_{\text{coh}} \) varies between 0.02 and 0.33. This means that the mean-squared error of \( \hat{y}_{\text{coh}} \) is at least three 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 \( \hat{y}_{\text{coh}} \).

**Denoising by coherence-constrained DL**

The sparse approximation problem presented in the subsection “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, \( z_1, z_2, \ldots, z_M \), 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 \( D \) is the same for all the recordings. Then, for \( i = 1, 2, \ldots, M \), the recording \( z_i \) contains noise and signal components. The noise component \( n_i \) is white Gaussian noise of mean zero and variance \( \sigma^2 \), and the signal component \( y_i \) 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 \( \hat{D} \) and a set of sparse coefficient vectors \( \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M \) such that the sparse approximations...
\( \hat{y}_i = \hat{D} \hat{x}_i. \)

for \( i = 1, \ldots, M \), recover the signal of interest present in the data set \( z_1, z_2, \ldots, z_M \).

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

\[
\{ (\hat{x}_i)_{i=1}^M, \hat{D} \} = \arg \min_{\{x_i\}_{i=1}^M, D} \sum_{i=1}^M \|x_i\|_0
\]

subject to \( \|z_i - Dx_i\|_2 \leq \epsilon, \quad i = 1, 2, \ldots, M. \) \( (11) \)

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 \( \epsilon \). The threshold \( \epsilon \) is dictated by the variance \( \sigma^2 \) 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 \( i = 1, 2, \ldots, M \), Aharon et al. (2006) propose to learn the dictionary and sparse coefficients by solving the problem

\[
\{ (\hat{x}_i)_{i=1}^M, \hat{D} \} = \arg \min_{\{x_i\}_{i=1}^M, D} \sum_{i=1}^M \|z_i - Dx_i\|_2^2
\]

subject to \( \|x_i\|_0 \leq T, \quad i = 1, 2, \ldots, M. \) \( (12) \)

In this problem, the representation error is minimized with the constraint that each representation of recordings should have 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

\[
\{ (\hat{x}_i)_{i=1}^M, \hat{D} \} = \arg \min_{\{x_i\}_{i=1}^M, D} \sum_{i=1}^M \|x_i\|_0
\]

subject to

\[
\begin{align*}
\min_{D} \sum_{i=1}^M \|z_i - Dx_i\|_2^2 \\
\mu(z_i - Dx_i, D) \leq \lambda \sqrt{\frac{\log(K)}{N}}, \quad i = 1, 2, \ldots, M.
\end{align*}
\]

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 \( \lambda \sqrt{\frac{\log(K)}{N}} \). The gain factor \( \lambda \) controls the strength of the denoising. However, one should set the gain factor \( \lambda \) to one for proper signal preservation (see Appendix B).

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

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

2) 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

\[
\hat{D} = \min_{D} \sum_{i=1}^M \|z_i - D\hat{x}_i\|_2^2.
\]

where \( \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M \) 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.

Figure 1. Denoising capability of the coherence-constrained sparse approximation \( \hat{y}_{coh} \) (solid curve), the error-constrained sparse approximation \( \hat{y}_{err} \) (dashed curve), and the cardinality-constrained sparse approximation \( \hat{y}_{car} \) (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 \( \alpha/\sigma \).
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, pp. 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-SVDerr) and the cardinality-constrained DL problem stated in equation 12 (denoted K-SVDcar). 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-SVDcar and K-SVDerr 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-SVDcar, and K-SVDerr. The thresholds used by K-SVDerr and K-SVDcar were ϵ = √Nσ and T = L, respectively. The CDL method was used with the gain factor λ set to one. The dictionaries that are output 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.

In Figure 3, the recovery rates for CDL, K-SVDcar, and K-SVDerr 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-SVDcar and CDL have similar dictionary recovery rates. For all tested cardinality values, K-SVDerr needed more iterations to converge than K-SVDcar and CDL did. However, for a cardinality equal to three, K-SVDerr 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

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 σ² 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).

Figure 3. Dictionary recovery capability of CDL (solid curve), K-SVDcar (dotted curve), and K-SVDerr (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 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 the CDL, FX-Decon, and K-SVD methods to adapt to the challenging and underexplored problem in which the variance of the additive noise is spatiotemporally varying.

**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 $100 \times 100$, 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.

The workflow for denoising a given 2D data window with CDL is presented in Figure 5. All overlapping patches of size $10 \times 10$ ($N = 100$) 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 $N = 100$ to generate the recordings $\mathbf{z}_1, \ldots, \mathbf{z}_{8281}$. The CDL algorithm is applied to the data set with 25 iterations ($I = 25$), the gain factor $\lambda$ set to one, and the number of dictionary atoms fixed at 100 ($K = 100$). Each sparse coefficient vector $\mathbf{x}_i$ and the dictionary output from the algorithm are multiplied to generate the sparse approximation $\hat{y}_i$ of the recording (see equation 10). Each sparse approximation vector $\hat{y}_i$ is later rearranged as a $10 \times 10$ 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 guarantee 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 $\epsilon$ set to $\sqrt{N}\sigma$, where $\sigma$ 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 $50 \times 50$ with 50% of overlapping in both dimensions and a filter of length of six 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 $10 \times 10$ patches in 10 lines of 10 atoms. We can observe that 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 columns, respectively. The denoising performance is quantitatively assessed via the S/N computed before and after the noise attenuation. For a given data $\hat{d}$ and its noise-free reference $d_{ref}$, the S/N expressed in decibels is defined as

$$S/N(\hat{d}, d_{ref}) = 10 \log_{10} \frac{k d_{ref} k^2}{k d_{ref} - \hat{d} k^2}.$$  

(15)

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.
The S/N values displayed under each noisy and denoised window show that CDL is performing similarly to K-SVDerr 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 CDL and K-SVDerr methods, we see no significant coherence in the error windows. However, we can observe signal in FX-Decon error window 2. This shows that FX-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 $N_{\text{var}}$ having a spatiotemporally varying variance, we generated a matrix of size $100 \times 100$ containing zero-mean white Gaussian noise $N_g$, and we modulated its amplitude such that

$$N_{\text{var}} = (W + T) \circ N_g,$$

where $\circ$ denotes element-wise multiplication. The matrices $W$ and $T$ are of size $100 \times 100$, and they 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 $N_{\text{var}}$ 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-SVDcar, K-SVDerr, 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-SVDcar method was applied with a cardinality threshold $T$ fixed at four, whereas K-SVDerr was applied with an error threshold $\epsilon$ fixed at $\sqrt{N_p} \sigma$, where $\sigma$ is the standard deviation computed on the noise $N_{\text{var}}$. 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-SVDcar, K-SVDerr, 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 the noise is highly attenuated and the 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 in the noise variance.

In the top of the K-SVDcar denoised window 1, some remaining noise is observed. In this area, there is initially a very low signal, and therefore, some of the four atoms are used to reconstruct the noise. Similarly, we can observe remaining noise in the K-SVDerr 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 repre-
sents some noise. Therefore, the \( K\text{-SVD}_{\text{car}} \) and \( K\text{-SVD}_{\text{err}} \) results show that variations in the signal complexity or noise variance reduce the denoising performances of the \( K\text{-SVD} \) denoising method.

As we can observe from the denoising results, \( \text{FX-Decon} \) denoising is not affected by the variations in the noise variance. However, it performs poorly in terms of noise attenuation and it fails to preserve the complex signal present in window 2. Finally, the \( S/N \) values show that \( \text{CDL} \) attains a higher \( S/N \) enhancement compared with \( K\text{-SVD}_{\text{err}}, K\text{-SVD}_{\text{car}}, \) and \( \text{FX-Decon} \).

FIELD DATA APPLICATION

A common-offset section of marine data (see Figure 8a) was selected to validate the capability of \( \text{CDL} \) to attenuate random noise present in field data sets while preserving the underlying seismic signal. The \( \text{CDL} \) method was applied on windows of size 100 \( \times \) 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 section, 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 \( \text{CDL} \) can attenuate such noise while preserving the signal. We selected three windows of the section for a detailed analysis, which are presented in Figure 9. Windows 1, 2, and 3 contain high-frequency flat events with poor \( S/N \), linear high-amplitude dipping events, 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 removed noise window 1, we computed the variance of two 10 \( \times \) 10 size patches pointed out with frame boxes. The variance \( \sigma_1^2 \) of the higher patch is 3.6 times larger than the variance \( \sigma_2^2 \) of the lower patch. This shows that the variance of the noise in field data is highly variable, even within a 100 \( \times \) 100 size window, and it justifies the need for a method such as \( \text{CDL} \) 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 the DL algorithms. Here, as in the conventional K-SVD denoising method, CDL denoising requires $O(NKL)I$ 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 101.7 s for CDL and 201.2 s for K-SVD on a laptop having a CORE i7pro CPU. The CDL denoising was faster because it provided a solution with a smaller $L$ when no signal was present in a patch. Moreover, the tractability of K-SVD can be strongly affected by the variation in the noise variance when the problem in equation 11 is solved. For instance, denoising window 1 with K-SVD$_{err}$ was approximately three times longer when additive noise had a varying variance (see Figure 7a) compared to when noise had a constant variance (see Figure 6a). The run time of K-SVD$_{err}$ is long 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 in 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 Figure 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 five 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 SNR value of 35.03 dB in 9.7 s. In addition, the algorithm used is a straightforward implementation of the algorithm presented in Appendix A, and therefore, it 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 with the DL methods. But it is much faster. For instance, filtering window 1 took 0.07 s. The algorithm used is from SeisLab, and it corresponds to the implementation proposed by Ulrych and Sacchi (2005, pp. 229–232).

CDL denoising could easily be 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 the 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 in 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 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 one 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 the signal. Using synthetic data, we compared CDL denoising with K-SVD and FX-Decon denoising for two noise contamination cases: noise with constant variance and noise with spatiotemporally varying variance. We observed that the CDL method performs similar to K-SVD for removing Gaussian noise with constant variance and has the advantage that it does not require knowledge of the variance. Moreover, the CDL method provides better denoising results than K-SVD when the variance of the noise is spatiotemporally varying. For both cases in which the noise variance is constant and spatiotemporally varying, the CDL method outperforms the FX-Decon denoising method. 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 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 $Z$ containing a recording vector in each of its columns as input, and it
returns a dictionary $D$ and a sparse coefficient matrix $X = [x_1, \ldots, x_M]$, 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, $x[j]$ is the $i$th sample of the vector $x$ and $D[i, j]$ is the sample at the $i$th line and $j$th column of the matrix $D$. In addition, columns inside the brackets are used to refer to all the indexes in a dimension; for instance, $X[i]$ is the $i$th line of matrix $D$. The notation $D^+$ has been used to denote the Moore-Penrose generalized inverse (Penrose, 1955) of matrix $D$. The function $\text{svd}(D)$ applies the SVD decomposition of the matrix $D$ and returns the matrices of eigenvectors $U$ and $V$, and of eigenvalues $\Lambda$, such that $D = U \Lambda V^T$. The symbol $\leftarrow$ 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 $b$.

**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 $n \in \mathbb{R}^N$ of mean zero and variance $\sigma^2$ relative to a redundant dictionary $D = [a_1, a_2, \ldots, a_K] \in \mathbb{R}^{N \times K}$. Such a value for coherence is mathematically given by

$$\mu(n, D) = \max_{j=1,2,\ldots,K} \left| \frac{n^T a_j}{\|n\|_2} \right|. \quad (B-1)$$

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

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

where $\mathcal{N}(\mu, \sigma^2)$ is a notation that refers to a Gaussian distribution of mean $\mu$ and variance $\sigma^2$.

Hence, the projection of the normalized noise vector on a dictionary atom $a_j^T n / \|n\|_2$ 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 $X_1, X_2, \ldots, X_n$ are mutually independent variables following Gaussian distributions of means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then the linear combination of these variables $\sum_{j=1}^n c_j X_j$ follows a Gaussian distribution (Eisenberg and Rosemary, 2008) such that

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

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

$$\frac{n^T}{\|n\|_2} a_j \sim \mathcal{N} \left( 0, \frac{1}{N} \right). \quad (B-4)$$
Using the definition in equation B-1 and the result of equation B-4, one may notice that estimating the coherence of the noise vector relatively 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 $x_1, x_2, \ldots, x_n$ following an identical Gaussian distribution of mean zero and variance $\sigma^2$ has an asymptotically optimal upper bound of

$$\max_i |x_i| \leq \sigma \sqrt{\log(n)}, \quad (B-5)$$

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

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

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

REFERENCES


