Seismic Denoising with Principal Component Analysis

Keegan Lensink
Seismic Laboratory for Imaging and Modelling, University of British Columbia
EOSC 510

Abstract

Seismic surveys are an essential part of estimating the geometry of the Earth's subsurface, and the inclusion of experimental noise in recordings is unavoidable in practice. In this paper PCA is investigated as a potential method of quickly denoising seismic data volumes. Synthetic white Gaussian noise is injected into every trace of a benchmark volume, and PCA is performed on time and frequency slices in the mid point – offset domain. It is found that there is always one optimum choice of k, the number of PCs used to reconstruct the volume, and that the signal is fit by a fraction of PCs where as noise is distributed amongst the remaining PCs. Reconstruction using the optimum k value in both the time and frequency domains reduces noise by ~60%, however operating in the time domain does a better job preserving the underlying signal. The method is limited in the sense that the optimum choice of k is not clear without benchmarking against the input signal. The results implicate that this method could be used as an effective first step in traditional denoising schemes, or as a quality assessment tool during acquisition.

Introduction

Geophysicists explore the complex geometry of the Earth's subsurface by recording its response at the surface to an injected wave field. The amplitude of this response is recorded in time at regularly spaced receivers in time series called *traces*. A collection of all the traces resulting from a single source, which in the case of marine surveys is a burst of compressed air in the water column, is called a *shot gather* (Figure 1). Subsequently, the collection of all such shot gathers results in a *seismic data volume,* which encompasses the entire spatio-temporal range of the seismic survey.

These seismic surveys are used to estimate physical properties of the subsurface, and as a result have high economic value. Experimental noise is unavoidable in practice, and must be dealt with while pre-processing the data volume. Denoising seismic data has been researched extensively and the field is saturated with effective, yet complex, methods that involve transforming the data into different domains and take advantage of the wave front's behavior in that domain. This paper investigates the effectiveness of Principal Component Analysis (PCA), a comparatively simple workflow, at denoising seismic data volumes.

In this paper, PCA is used as a way of determining patterns that capture the most variance in high dimensional data. The resulting Principal Components (PCs) are orthogonal and are returned in descending order of variance. This is useful if it is assumed that the data is corrupted with white Gaussian noise, which is uncorrelated with the underlying signal. This paper investigates the assumption that the noise is distributed throughout all PCs, and that reconstructing the signal from the fraction of the PCs carrying the most variance will remove a majority of the noise while retaining the original signal.

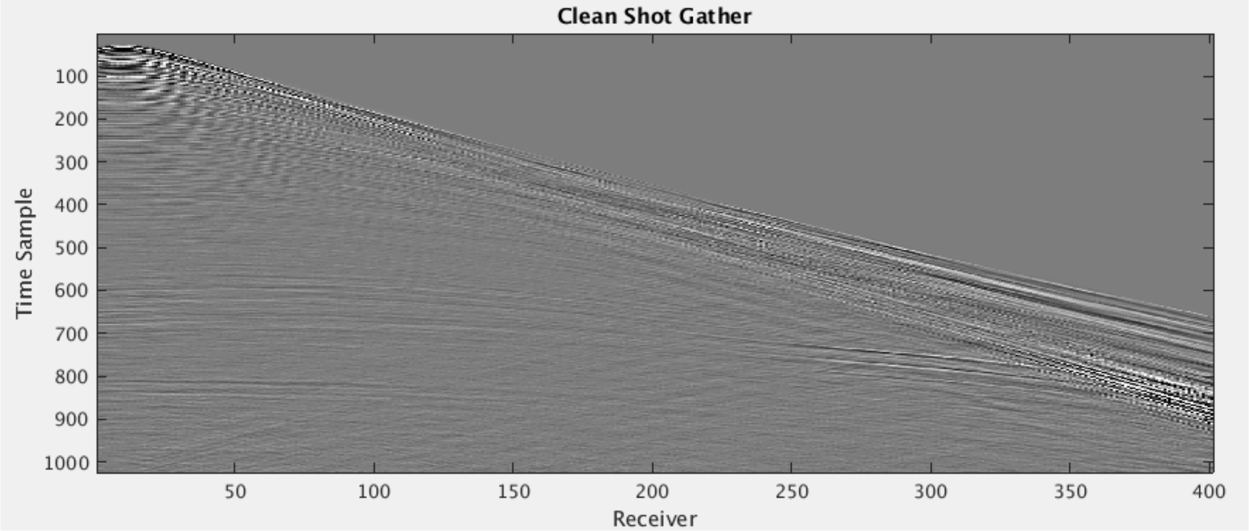


Figure 1: A noise free shot gather composed of 401 traces recording 1024 samples.

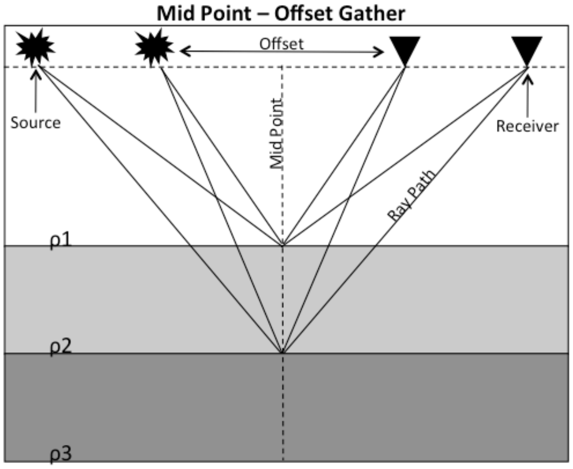
Data and Methods

The *1024 sample x 401 receiver x 401 source* seismic data volume used in this paper came from a 2D marine survey in the North Sea. In pre-processing it has been stripped of all metadata, regularized, interpolated to fit on the grid, and denoised. This PCA method is invariant of spatial and temporal resolution, so that information is unnecessary.

Starting from a noise-free data set, white Gaussian noise is added to every trace in the volume to simulate experimental noise. This allows the recovered volume to be directly compared to the original signal, which is the desired recovery.

In order to minimize the distribution of variance amongst PCs, the rank of the input to PCA must be minimized. The 3D data volume allows 3 possible directions along which PCA could be performed on 2D slices of the volume. Shot gathers are the most easily visualized method of viewing seismic data, however the diagonal spread of first arrivals means they are high rank. Due to the symmetry of acquisition, receiver gathers are identical to shot gathers. Taking time slices in the original acquisition geometry results in the majority of the signal existing along the diagonal (Figure 3 bottom left). It is apparent that in order to get the best results the data must be rearranged in such a way that the rank is minimized. Transforming the data to the midpoint-offset (MO) domain gathers source-receiver pairs that share a common midpoint by their receiver-source offset

(Figure 2). This operation has no effect on the time dimension, and effectively rotates the spatial domain 45°. Time slices in the MO domain now contain the majority of the signal in columns, minimizing the rank of the input (Figure 3).



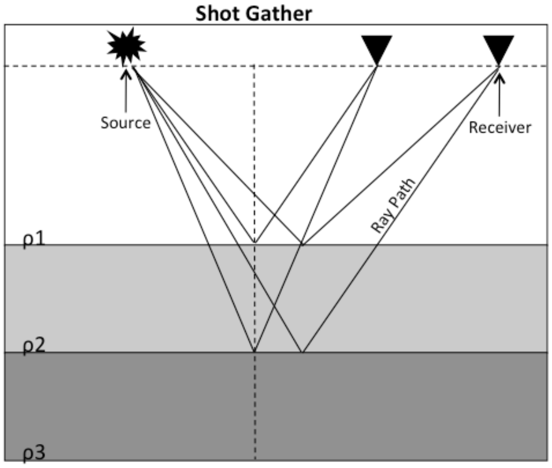


Figure 2.1 (Left): An example showing the ray path geometry in a common midpoint gather. Offset is a spatial measurement between source and receiver. Ray path is dependent on the density (ρ) of the subsurface layers.

Figure 2.2 (Right): An example showing the ray path geometry in a common shot gather.

Taking the Discrete Fourier Transform (DFT) of each trace and performing PCA on frequency slices, as opposed to time slices, has multiple benefits. It is possible to exploit symmetry in the frequency domain by only performing PCA on only the non–negative frequency slices. Additionally, data volumes are commonly too large to keep in memory, which would require time slices to be created and stored out of core before hand. This can be time consuming and is not commonly done. However, frequency slices are used regularly for other types of processing, so they have likely already been store. The negative frequency slices can be found by taking the complex conjugate of the recovered non-negative frequency slices. The complete volume of frequency slices is returned to the time domain using the Inverse Discrete Fourier Transform (IDFT) and compared to the original signal. This increases performance by reducing the PCA function calls to $\frac{n}{2}+1$.

This paper investigates the effectiveness of PCA at removing synthetic white Gaussian noise in two parallel workflows: One where time slices are input to PCA, and a second where non-negative frequency slices are input to PCA.

PCA is performed on the de-meaned 2D input by calculating the eigenvalues and eigenvectors of the covariance matrix. Eigenvalues represent the variance explained for each respective eigenvector. Reconstructing the image using only the first k eigenvectors results in an image containing only the patterns that explain the most variance.

Time Domain Workflow

- Inject White Gaussian Noise
- Switch from Source-Receiver domain to MO domain
- Perform PCA on each time slice
- Revert from MO domain to Source-Receiver domain
- Compute RMSE of original signal to recovered signal

Frequency Domain Workflow

- Inject White Gaussian Noise
- Take DFT along time dimension
- Switch from Source-Receiver domain to MO domain
- Perform PCA on non-negative frequency slices
- Revert from MO domain to Source-Receiver domain
- Take IDFT along time dimension
- Compute RMSE of original signal to recovered signal

Results and Discussion

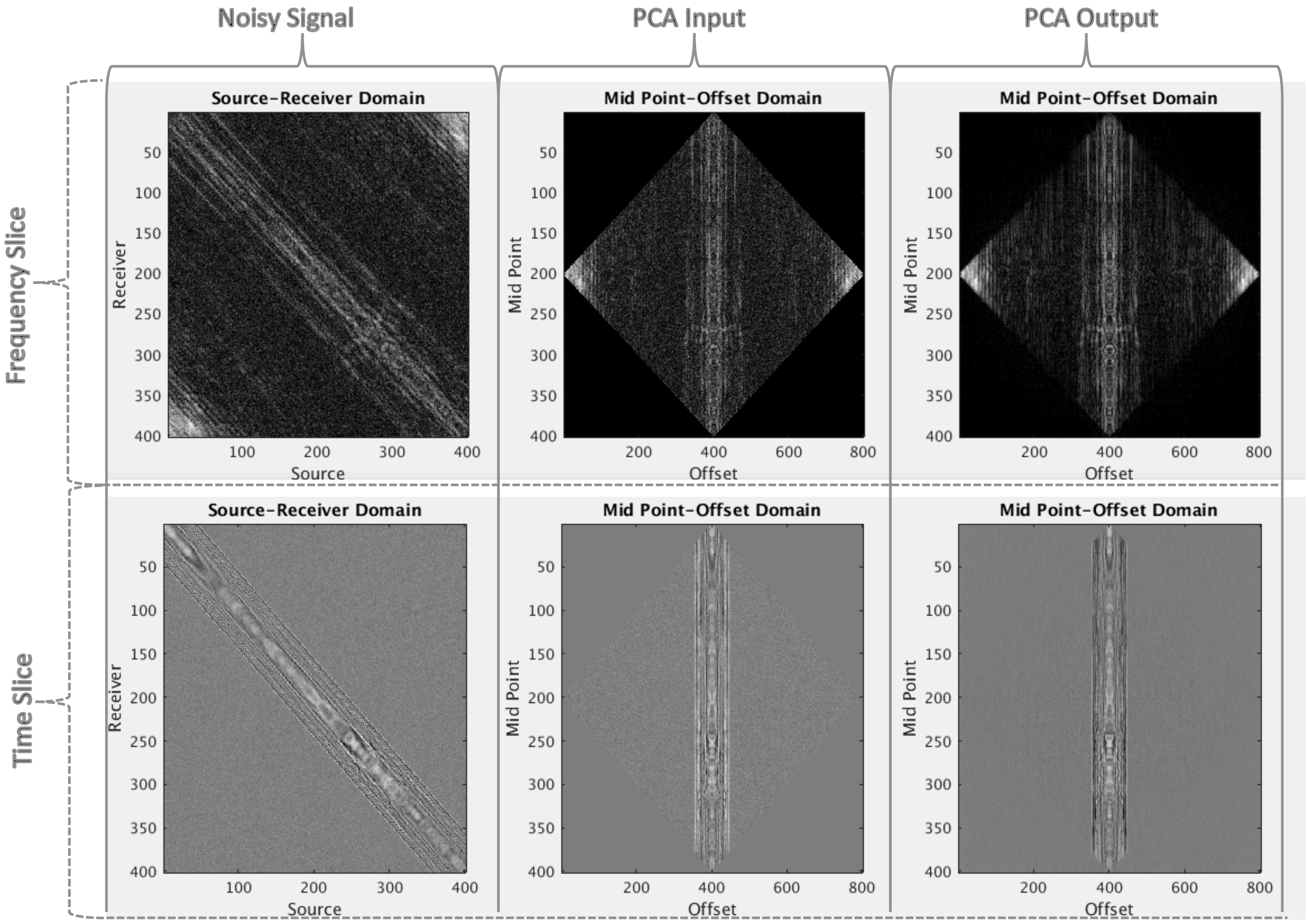


Figure 3: (Top) 100th frequency slice throughout pre-processing, and the output from PCA. (Bottom) 100th time slice throughout pre-processing, and the output from PCA. The first column shows the noisy signal before pre-processing. The second column shows the result of transforming to the MO domain, and is the input to PCA. The third column is the output from PCA.

The transformation from Source-Receiver domain to the MO domain successfully reduces the rank of the input data. The transformation constrains the signal to a fraction of the columns, as opposed to all columns (Figure 3). Since the synthetic noise has random frequency content, the effect is similar in both the time and frequency domains.

In order to calculate the optimum number of modes, k, to keep in each workflow the Root Mean Squared Error (RMSE) between the reconstructed signal and the original signal is used to measure the accuracy of recovery. The results show that, for both workflows, there is only one minimum when RMSE is calculated as a function of k. (Figure 4)

Keeping all values of k perfectly recovers the input, which is clearly not the optimum choice. Choosing k lower than the optimum value under fits the signal, while choosing k higher than the optimum value over fits the noise. Both the time and frequency domain behave similarly as k increases, with minimum at $k = 12$ and $k = 14$ respectively. Note that this represents a compression of ~97% for both domains.

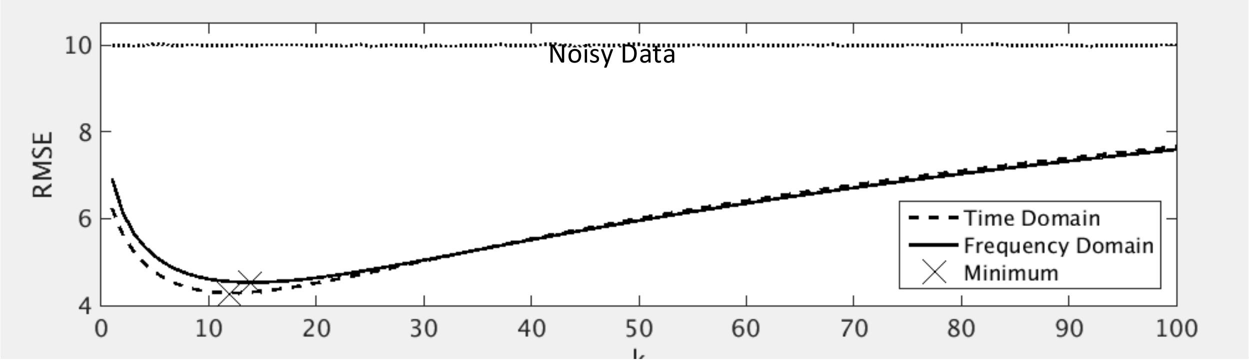


Figure 4: RMSE computed over the first 100 of 801 PCs, focusing on the minima. Error continues to increase for $k > 100$ until the noisy data is reconstructed exactly.

The resulting minima are the optimal choice of k for this level of noise input, and will be used to generate the final results. Generally if less noise in injected into the data, the optimal k value increases along with accuracy. The final results of this study compare reconstructions of shot gathers rather than time slices. While this method operates PCA on time and frequency slices in order to minimize rank, the end goal is to observe how effective the method is at recreating shot gathers.

Figure 4 shows that both methods preform similarly and are able to reduce the RMSE between the original signal and the recovered signal from 10 to ~4. Figure 4 shows that the RMSE increases gradually after the optimum choice of k. This is important because in reality the original signal is unknown, so it cannot be used to compute the optimum k. These results show that overestimating the choice of k will maintain the majority of the signal, at the cost of over fitting noise. On the other hand, erring on the low side of k under fits the signal quickly and loses information. The difference in slope of either side of the minima show that PCA is able to fit the underlying signal with fewer PCs than the noise.

Figures 5 and 6 show the optimal recovery in both the time and frequency domain. Both methods perform similarly in terms of their ability to minimize the RMSE between the recovered signal and the noise-free signal. However, the residual shows that while the methods are producing similar RMSEs, they are fitting the data differently. The residual in the frequency domain shows more energy along the diagonal and at the wave front in the top left of the image. This means that while the frequency domain workflow is twice as efficient, it is less effective at recovering the original signal than the time domain workflow. Since both methods have similar RMSE values, but upon closer inspection show different performance, this is also suggests that RMSE may not be the best metric for optimizing a choice of k. Increasing k, that is keeping more PCs, will retain more of the signal at the expense of keeping more noise along with it. This suggests that some lower amplitude behavior in the signal is not completely uncorrelated with the noise.

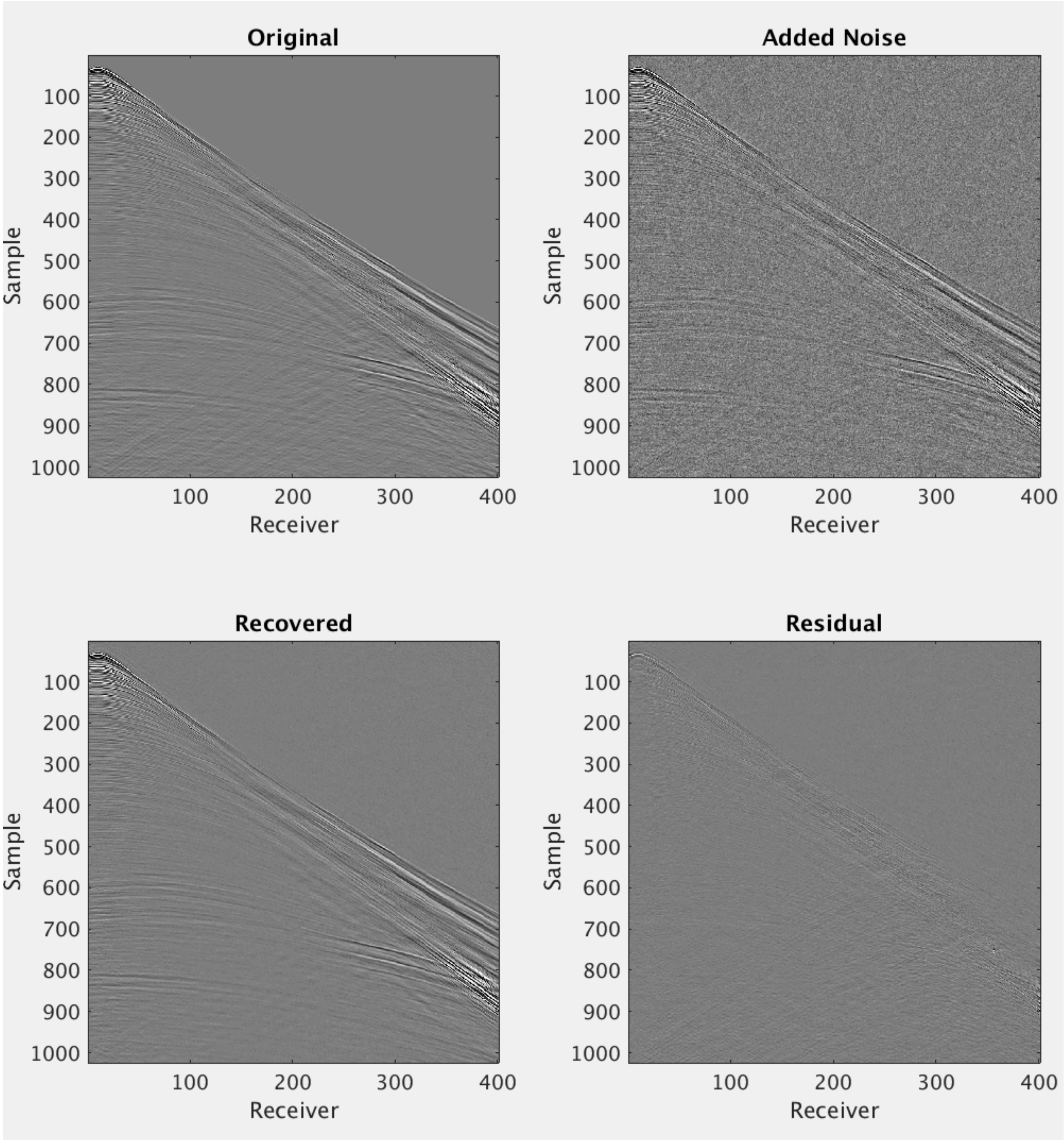


Figure 5: PCA Denoising using k=14 PCs in the **frequency** domain. (Top Left) 10th shot gather in the original noise-free volume. (Top Right): 10th shot gather after noise has been added to every trace. (Bottom Left) 10th shot gather of the recovered volume from keeping the first 14 PCs of each frequency slice. (Bottom Right) The residual between the recovered shot gather and the original noise-free shot gather.

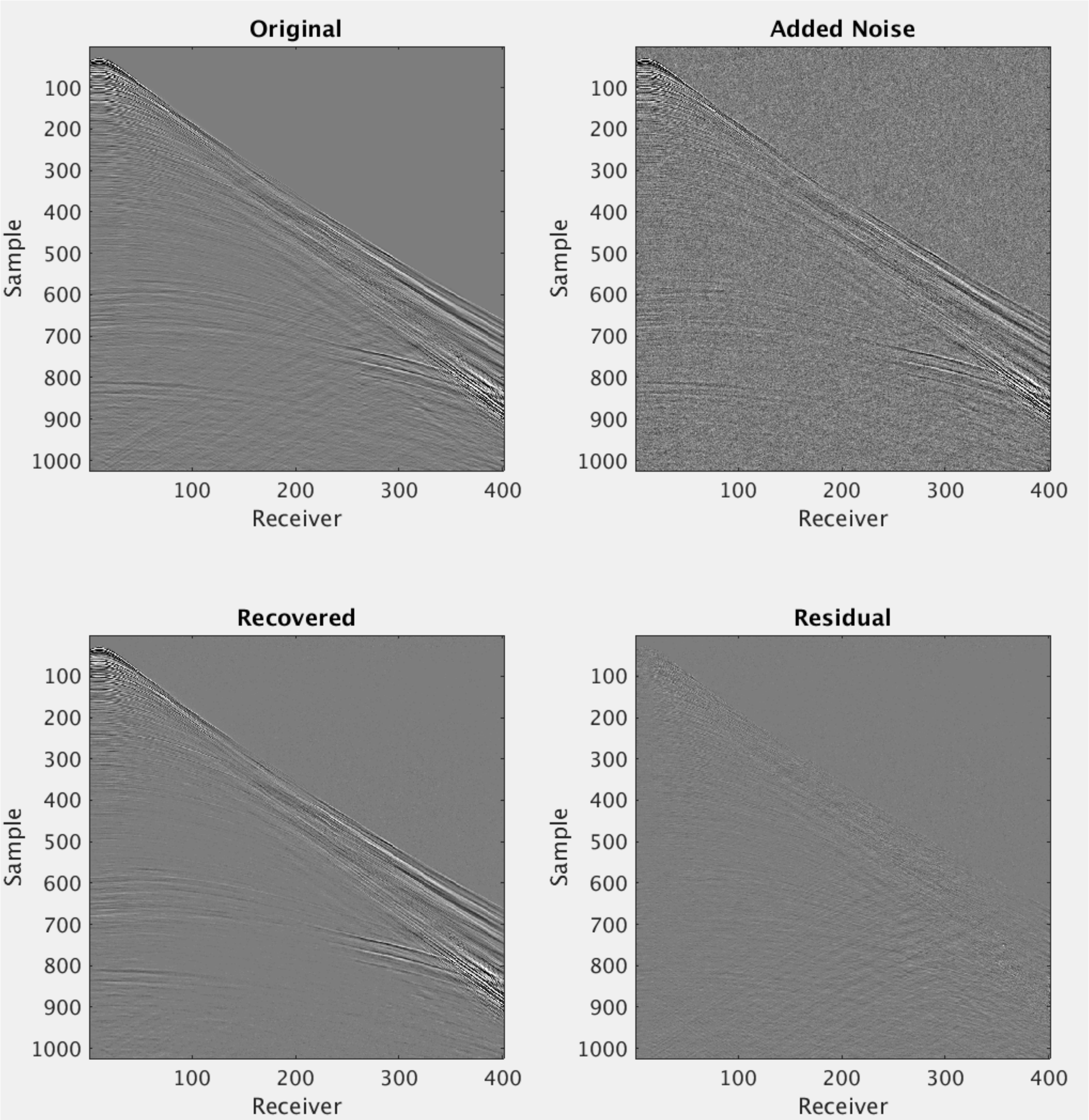


Figure 6: PCA Denoising using k=12 PCs in the **time** domain. (Top Left) 10th shot gather in the original noise-free volume. (Top Right): 10th shot gather after noise has been added to every trace. (Bottom Left) 10th shot gather of the recovered volume from keeping the first 12 PCs of each time slice. (Bottom Right) The residual between the recovered shot gather and the original noise-free shot gather.

Conclusions

PCA can be used as a relatively simple and effective method of denoising seismic data volumes. Operating on time and frequency slices reduces the rank of the input to PCA and allows the majority of the signal to be captured by a fraction of the PCs. The remaining PCs carry a small amount of energy from the signal, and a majority of the energy from the noise. Performing PCA on all time slices is less efficient, but results in a better recovery of the underlying signal from the noise. Over estimating the optimum k is safer than underestimating as the signal is fit much more rapidly than the noise is over fit. PCA was chosen due to its simplicity and its efficiency, and in that respect the method is successful. The method is able to quickly remove ~60% of noise while retaining the majority of the underlying signal.

Determining the optimum k value is the biggest limitation of this method. In reality the underlying signal isn't known and cannot be used to benchmark performance while determining the optimum value of k. Further research into the relationship between k and signal to noise ratio could provide insight into this limitation. Since this method works on time and frequency slices, the entire volume needs to be denoised in order to recover any single shot gather. Volumes too large for memory are common, so this would require time and frequency slices to be called from out of core and reduce efficiency. Future research into implementing Local Pixel Grouping when performing PCA could increase the accuracy of these workflows.

The results implicate that this method could be used as an effective first step in traditional denoising processes. Erring on the side of over estimating k means that one could confidently retain the underlying signal while drastically reducing noise before continuing on to other denoising methods. The efficiency of this method also means it could be used in the field during acquisition for quality assessment, as this doesn't require as much accuracy but needs to be done quickly.

References

Smith, Lindsay I. "A Tutorial on Principal Components Analysis." (2002): n. pag. Web. 10 Dec. 2016. <http://www.sccg.sk/~haladova/principal_components.pdf>.