Interpolation of seismic data using a principal component analysis POCS approach

Scott Keating and Kris Innanen

ABSTRACT

Projection onto convex sets, or POCS, is a simple, straightforward method of interpolation which hinges on a few basic assumptions. In order to better fulfill these assumptions, a type of POCS is proposed in which a projection onto principal components is the transform used. This method is shown to be effective both on simple synthetic data, as well as on real VSP data. The problem of acquiring the relevant principal components is addressed by using nearby intact data, and other methods are proposed if this should prove impractical. Partially incomplete data are shown to provide adequate principal components for interpolation in some cases.

INTRODUCTION

In seismic acquisition, we are frequently unable to gather all of the data we would like to. The large number of sources and receivers mean that some equipment failure is likely to occur, and surface features can prevent us from placing them everywhere we would like to. Missing data in an otherwise regular survey can adversely affect the processing and interpretation of our measurements. These data we collect are not entirely independent of each other, and from the behaviour of the data we do collect, we can gain some idea of what the missing data likely were. By interpolating, we gain an estimate of missing data based on our known data. While this will only be an approximation to the actual data, it is a significant improvement over no data at all. We are not generating new information by interpolating, but rather making explicit what the gathered data tell us about the missing points. There are many different ways to interpolate seismic data for both regular and irregular missing data, of which projection onto convex sets (Abma and Kabir (2006)), minimum weighted norm interpolation (MWNI) (Liu and Sacchi (2004)), and the F-X domain interpolation of Spitz (1991) are just a few. In this research, we have focused on the projection onto convex sets, or POCS method. In this paper, a variety of POCS in which projection onto principal components is used as a transform is proposed.

POCS

Any data with missing points which we wish to interpolate can be thought of as the sum of two functions; the 'ideal', interpolated result we hope to obtain, and some 'gapping function' which, when added to the ideal function, introduces the missing points we are attempting to interpolate. The POCS method attempts to use discrimination between these two functions in order to interpolate. At its core, POCS consists of applying a transform to the data, discriminating between the ideal data and the gapping function in the transform domain, and transforming back to the original domain (Abma and Kabir (2006)). By doing so, we obtain an estimate of the ideal function, which can then be used to replace the missing points in our measured data. There are two crucial assumptions behind POCS interpolation. The first of these is that the 'ideal' data are represented in the transform

domain with relatively few, high amplitude coefficients. The second assumption is that the 'gapping function' is not special in the transform domain and so consists of randomly distributed and relatively low amplitude coefficients in this transform domain.

The POCS method consists of the following steps. First, the data are transformed to some transform domain. Provided our assumptions are satisfied, in the transform domain we will have a few high amplitude points (from the ideal function), and an abundance of randomly scattered, low amplitude points (from the gapping function). Due to the differences between the ideal function and the gapping function in this transform domain, we now have a means of distinguishing one from the other. The high amplitude points should primarily represent the ideal function, with only small contributions from the gapping function (which is more evenly distributed). The low amplitude points contain most of the information corresponding to the gapping function, and for the ideal function only represent whatever small fraction was not well represented as high amplitude in the transform domain. We then apply a threshold function, zeroing all points in the transform domain below a certain amplitude. This should remove most of the data belonging to the gapping function, while preserving most of the data belonging to the ideal function due to their differences in amplitude. We then inverse transform, which gives us an approximation of the interpolated, ideal function. By replacing our measured data with this approximation at the missing points only, we obtain an estimate of the interpolated data. This estimate will not be perfect, as the application of the thresholding function removed those parts of the ideal function which fell below the threshold in the transform domain, as well as preserving the parts of the gapping function which were represented at points that laid above the threshold. Because our estimate is closer to the ideal function than our original data, the gapping function is now lower in amplitude than in the original data. This means that if we repeat the above procedure of forward transformation, threshold, inverse transform and replacement, we can safely lower the threshold (Abma and Kabir (2006)), as the gapping function will also be lower in amplitude. This allows us to recover more of the ideal signal, while still removing most of what remains of the gapping function. By iterating this procedure, we eventually obtain something very close to the ideal function.

The basic idea of the POCS method is illustrated in Figures 1 and 2. On the left, we see that our measured data is equal to the sum of an 'ideal function' (in this case a sine wave) and some 'gapping function' that introduces our missing points. On the right, we see the representations of each of these functions in the transform domain (in this case a Fourier transform is used). Here, we can easily distinguish between the ideal function, which by the nature of the transform chosen is high amplitude at a single point, and the gapping function, which is more evenly distributed. By inverting only the data above the threshold in the transform domain, we recover an estimate of the ideal function (Figure 2, top). We can insert this estimate into the missing portions of our measured data to obtain a first update to our measured data (Figure 2, middle). When we iterate this process, we can safely lower the threshold, as more of the gapping function is removed on each iteration (Figure 2, bottom).

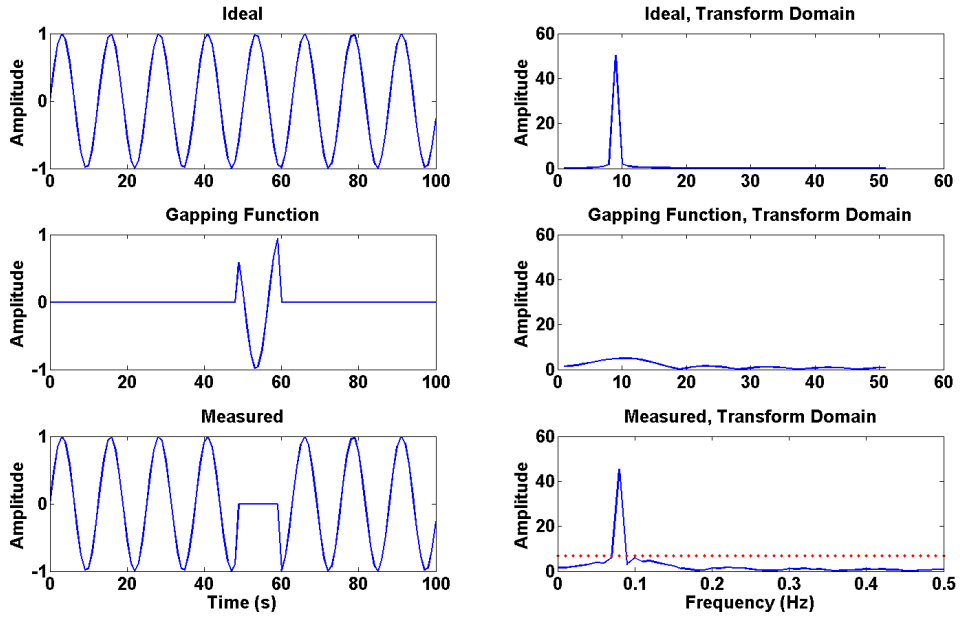

FIG. 1. Left: Measured data can be viewed as the sum of the ideal interpolation we hope to achieve, and some gapping function. Right: Ideal interpolation is easily distinguishable from the gapping function in the transform domain by means of amplitude difference. In this example, a Fourier transform is used. The red dotted line represents a possible threshold amplitude for this iteration.

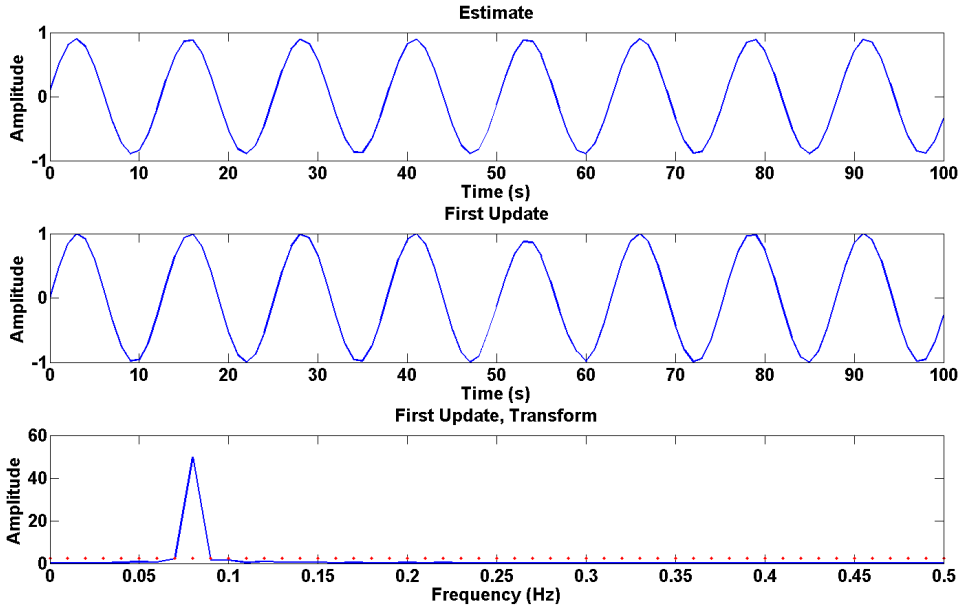

FIG. 2. Top: Estimate of the interpolated data. Middle: For the update, missing data points in the measured data are replaced with our estimate. Bottom: Transform of the estimate. As our first iteration removed much of the gapping function, we can safely lower the threshold for further iterations. The red dotted line represents a possible threshold amplitude for this iteration.

MOTIVATION FOR USING PCA

As described above, POCS is an appealing method for interpolation, due to its straightforward concept and simple implementation. It is crucial, however, that the assumptions of a sparse, high amplitude ideal function in the transform domain, and a distributed, low amplitude gapping function in the transform domain are satisfied, or the interpolation will fail. As such, it is important to ensure that the transform used satisfies these assumptions for the data being interpolated. Commonly, the Fourier transform is used for POCS interpolation. While there are many data sets for which our assumptions are valid using a Fourier transform, there are certainly cases in which the Fourier form of POCS fails. For example, if the missing data are periodic rather than random, the gapping function will be represented by a few high amplitude points in the Fourier domain, badly violating our assumptions. Furthermore, while the Fourier transform fulfills the requirements of POCS well enough to work on many data sets, there is no guarantee, or even expectation that it will be the transform which best matches these requirements. While other transforms may be appealing in specific cases, the best transform is ultimately dependent on the data to be interpolated. In this research it is proposed that such a transform should be designed based on principal component analysis.

Principal component analysis is a multivariate statistical technique that analyzes a data table, and attempts to represent it in a set of new variables, called principal components, which are created by linear combinations of the original variables (Abdi and Williams (2010)). These principal components are chosen such that the first principal component is the linear combination of the original variables which describes the maximum possible variance in the data, the second principal component is orthogonal to the first, but otherwise describes the maximum possible variance in the data, and so on, until the last principal component, which is simply orthogonal to all the others (Abdi and Williams (2010)). In this way, the first N principal components are the N linear combinations of dimensions which describe the maximum total variance. If the data are not random, this means that the first few principal components will describe the major trends in the data, while the remaining principal components will describe the deviations from these trends. Thus, for a data set with strong trends, a given data point should be well described by the first few principal components, and only have minor contributions from later principal components.

So, if we were to project a data point that closely follows the trend of a data set onto the principal components of that data set, it could be well approximated by a few high amplitude points for the first few principal components, and zeroes elsewhere. Any point that did not follow the trend of the data would not be well represented by the first few principal components, and would instead, in general, have some projection onto every principal component. This makes projection onto principal components an appealing transform to use in the POCS method; the ideal data will be high amplitude at just a few points, whereas the gapping function should be randomly distributed, provided the gaps in the data do not follow the trend of the data. As these are the conditions for the success of the POCS method, this transform would seem to be ideal. The problem which presents itself when attempting to use this transform is then attaining the necessary principal components, which may be difficult to estimate when confronted with gappy data.

For the singular value decomposition of some data matrix X,

$$X = P\Delta Q \ , \tag{1}$$

the columns of the matrix Q are the eigenvectors of the matrix $X^T X$. In the context of principal component analysis, these are called the loading vectors of the data X (Abdi and Williams (2010)). This loading matrix is important, because it is a projection matrix for the principal components of X. This means that multiplication of some vector by Q will give the projection of that vector onto the principal components of X (Abdi and Williams (2010)). The projection of some data Y onto these principal components is then

$$A = YQ \ . \tag{2}$$

When we project data onto these principal components, we know that we expect very large values for the lower order principal components, and smaller ones for the higher order principal components. Within the POCS framework, this means that the 'thresholding' step, where we evaluate in the transform domain which points correspond to the ideal function and which to the gapping function is largely redundant. Rather than performing this thresholding step, we can instead choose to project only onto the first M principal components to achieve the same end. As we iterate we will increase M, the number of principal components we project onto, effectively lowering the threshold. Our projection is then

$$A^{[M]} = XQ^{[M]} \tag{3}$$

where $Q^{[M]}$ is Q truncated to M columns. We then wish to inverse transform these projections in order to recover the approximation of our missing data. The transpose of Q is also its inverse $Q^T = Q^{-1}$, so the output of each iteration is

$$X^{[M]} = A^{[M]} Q^{[M]T} \ . \tag{4}$$

At each iteration, then, we take the input data X, and produce our M principal component estimate of the ideal function by

$$X^{[M]} = XQ^{[M]} Q^{[M]T} \ . \tag{5}$$

We then replace our input data with these new values at the missing data points, and iterate.

PRINCIPAL COMPONENTS FROM SEISMIC DATA

When we think of finding the principal components of seismic data, the idea is, at first, confusing. Suppose we wish to characterize a shot gather of M data points by its principal components. At each of these points we have some measure of the amplitude of the seismic wavefield. So, we have M points in one dimension (amplitude). PCA would seem not to apply here; as there is only one dimension, the single principal component obtained from PCA will lie along it. Alternately, we can view the shot gather as a single point, with the amplitude at each data point being its own dimension. This results in a single point in M dimensions, where again the applications of PCA are trivial. Instead of these extremes, we can divide the shot gather into P sub-windows of Q data points each. This lets us consider

our data as P points in Q dimensions, and provides us with a valid set of data from which we can construct principal components. This means that when we interpolate using these principal components, we will have to do so in these sub-windows.

In application to real data, an important question is how to determine the size of the sub-windows over which the principal components are created, and on which the interpolation occurs. If too large a window is chosen, then we will be attempting to interpolate on a scale beyond that over which the data displays predictable behaviour. That is to say, we will be interpolating the data at a given position and receiver time based in part on data sufficiently removed in position or time as to have little or no relation with the point to be interpolated. This can lead to principal components which are noisy, and the resulting interpolation will suffer from this noisiness. Alternately, if we use a very small design window, the principal components will not suffer from this noisiness as we expect always to see a strong relation between points very near in position and time. The problem that instead arises for very small windows is that our interpolation will not benefit from longer ranged behaviour. Interpolations based on these small windows are inherently blind to long scale trends which may provide useful information. Balance between these extremes is necessary in order to obtain principal components useful for interpolation.

PROBLEMS WITH USING PCA

While the use of projection onto principal components as a transform for use in POCS is appealing, there is an significant flaw in this approach as presented so far. Specifically, the problem is that the relevant principal components are those of complete data. Unfortunately, in the interpolation problem, complete data is our objective, and the data we begin with are missing certain points. The principal components of the data we begin with will then be ideally suited for representing data with gaps, and will not perform well in a POCS interpolation. Consequently, we are forced to use some other set of principal components.

It is not necessary that we use the principal components of the complete data, we simply need to use principal components which describe the same trends. We can imagine that if the data we wish to interpolate were just some subset of a larger set of similar data, the principal components of this larger set would still describe the ideal data of our interpolation in a sparse way. Our problem, then, is to identify some larger data set that we do have access to whose principal components still reflect the trend of our data.

One possibility, if we are attempting to interpolate a given shot gather with missing traces, is to create the principal components from nearby, more complete shot gathers. Given that these shot gathers will be composed of similar waves moving through the same medium as in the gather we wish to interpolate, it is reasonable to assume that these will be similar enough to share principal components with the gather we wish to interpolate. If there are a similar quantity of missing traces in nearby gathers, then another shot record in the same region, or even modeled data for the area may provide a means of obtaining principal components.

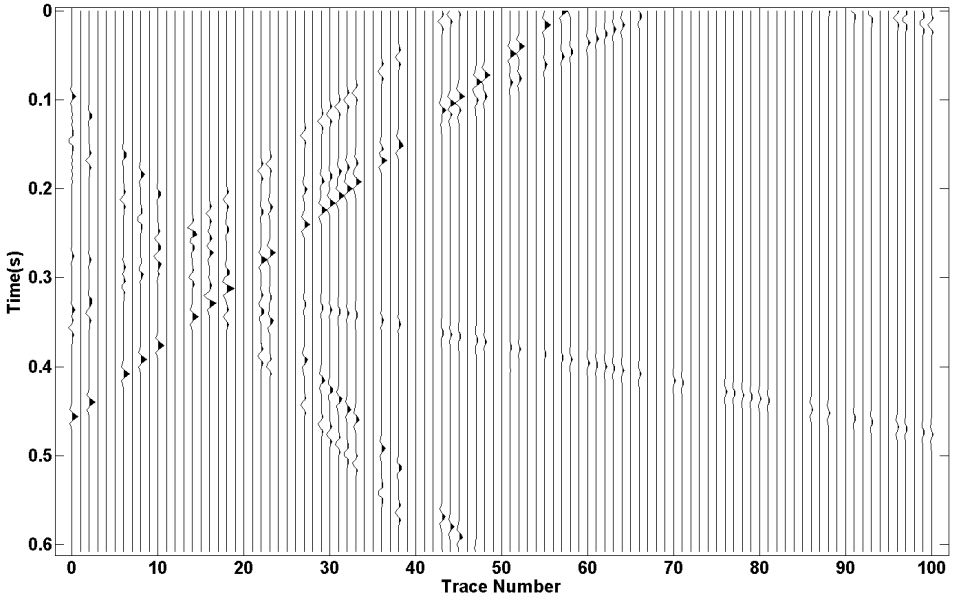

FIG. 3. Linear events with half the traces randomly removed

RESULTS

For this project, a PCA POCS method of interpolation was tested on both simple, synthetic data, as well as on real VSP data. Initially, tests were done on synthetic, linear events. First, complete linear events were generated, and the principal components of these complete events were determined. Next, traces were removed at random from these linear data. Then, a PCA POCS interpolation was applied to the data, using the principal components of the complete data. The results of this interpolation are shown in Figures 3-5. As can be seen, the interpolation, for the most part, performs well on these simple data. These linear events are very simple, however, and it is interesting to investigate how the interpolation performs on more complex data.

Another simple, synthetic test was done to investigate the interpolation of hyperbolic events. This test was performed in order to gauge the effectiveness of this interpolation on more complicated events. The same procedure was used as the one outlined above for linear events. These results can be seen in Figures 6-8. These hyperbolae are still well interpolated, which is a positive sign for the ability of this method to deal with non linear events.

This method of interpolation was also tested on VSP data from an undisclosed source and an undisclosed location. While in the synthetic examples shown so far the principal components of the complete data were used, in interpolation of real data we may safely assume that the complete data are unavailable. The challenge of using principal components that do not presuppose knowing the final product of the interpolation then presents itself. The VSP data used in this research contains multiple shot gathers, one approach is to interpolate one shot gather using principal components based on the neighbouring, complete shot gathers. When we use these principal components, we are making the assumption that

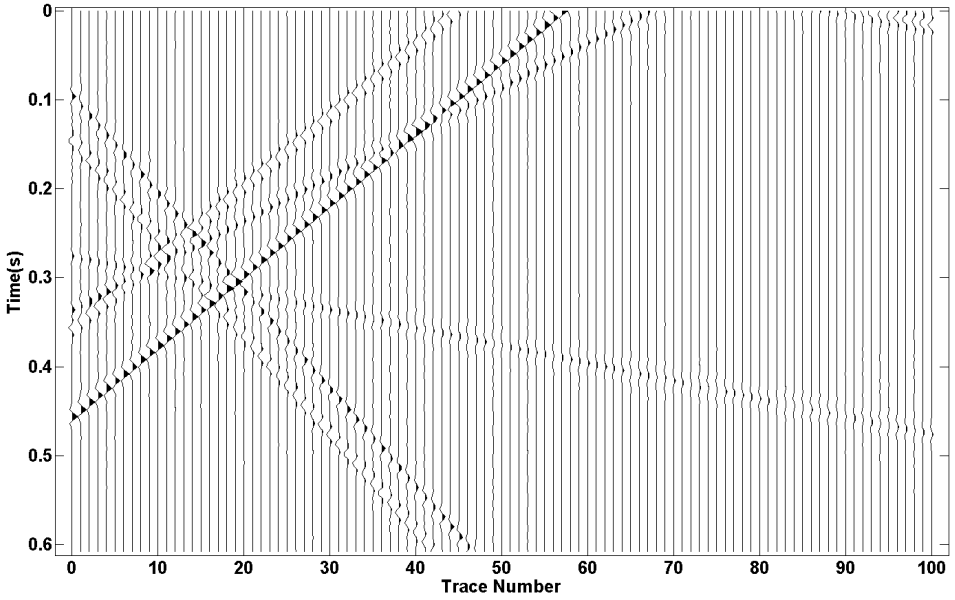

FIG. 4. Linear events after PCA POCS interpolation.

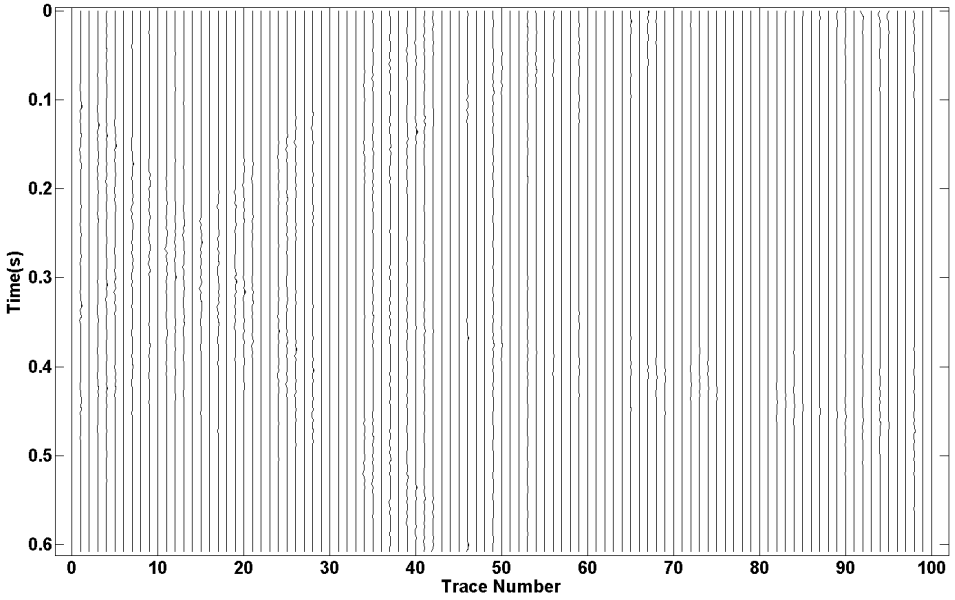

FIG. 5. Difference between the complete events and the interpolation.

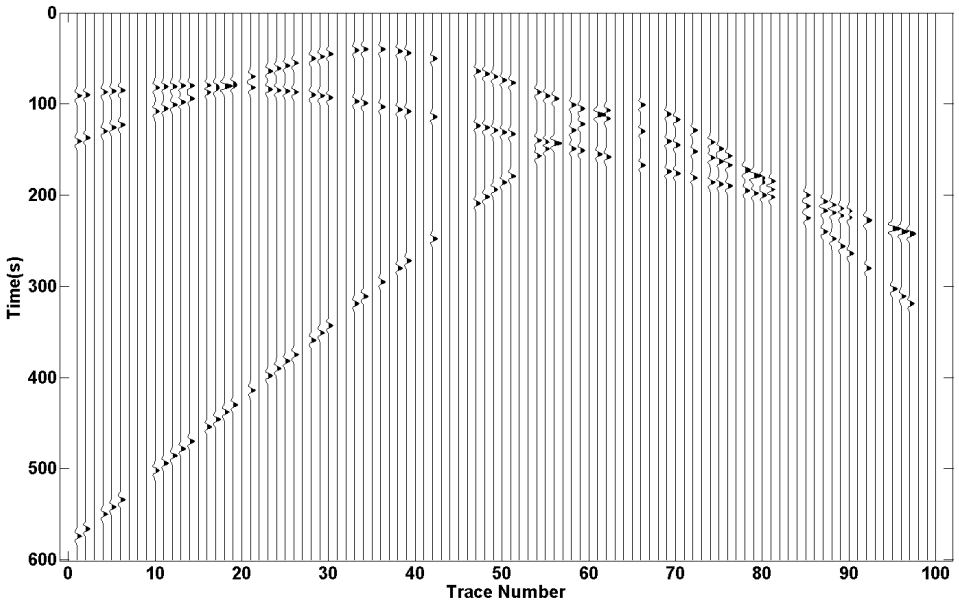

FIG. 6. Hyperbolic events with traces randomly removed

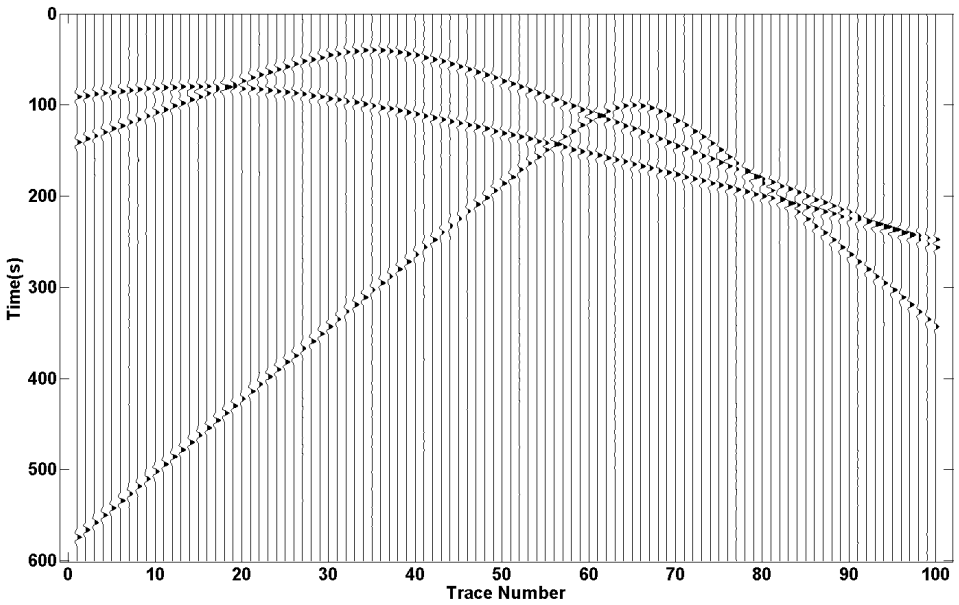

FIG. 7. Hyperbolic events after PCA POCS interpolation.

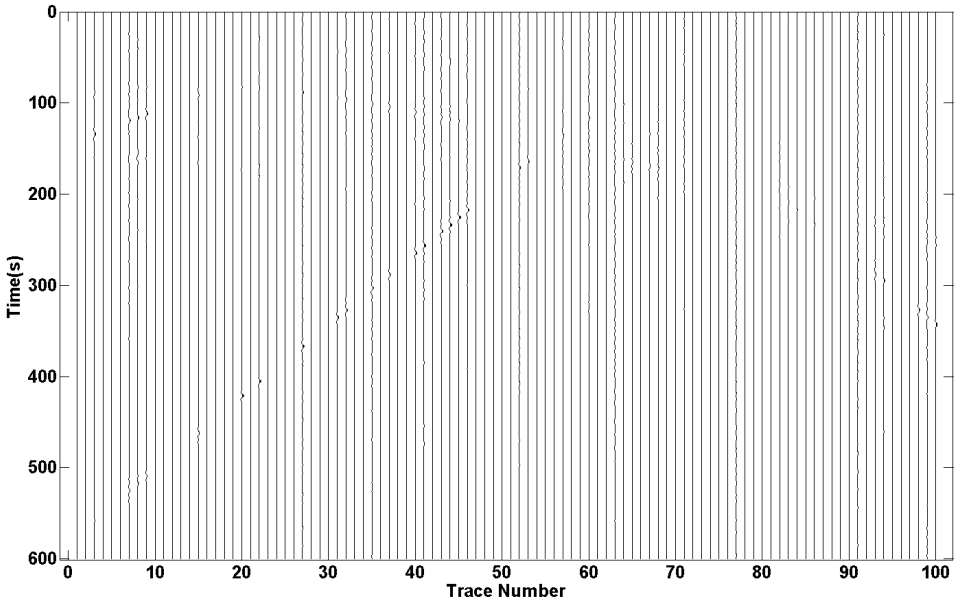

FIG. 8. Difference between the complete hyperbolic events and the interpolation.

we will see the same trends in these other shot gathers as we would in the complete data of the shot gather we hope to interpolate. It was found that using only the nearby shot gathers to create the principal components yielded better results than using all available shot gathers.

To test the interpolation on the real VSP data, principal components were created based on the four shot gathers closest to the one to be interpolated. Traces were then removed at random from the shot gather to be interpolated. A POCS PCA interpolation was then applied to the reduced data. Experimentation with the window size led to a choice of a 10 trace by 100ms window for use in the interpolation of these data. The results of this interpolation are shown in Figures 9 to 12. As shown in Figure 12, the interpolation performs well on the whole, there are few errors in the interpolation. Figures 13 and 14 show in more detail a region where the interpolation has been successful. As can be seen in Figure 14, even in noisy regions and regions where there are large amplitude variations from trace to trace, the data is still interpolated successfully. Most missing points in this data set are interpolated with a similar level of success, but there are also some isolated errors.

Figures 15 and 16 show in more detail a region in which there are noticable problems. In Figure 16 at about 0.4 ms, traces 54-56 are clearly interpolated incorrectly. At these points, the interpolation seems to have introduced data at an incorrect slope, incorrectly connecting troughs across the missing data, and cutting across a peak in the data. It is likely that the correct event here is not being adequately represented by the principal components and that the principal components corresponding to the incorrect interpolation are much lower order. This would lead to the correct interpolation being discarded, while preserving this incorrect interpolation.

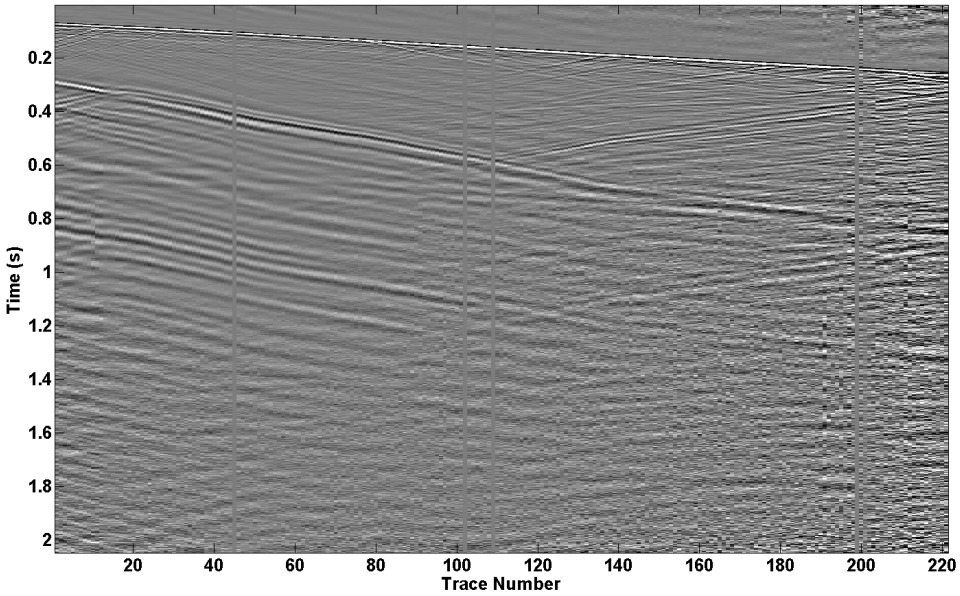

FIG. 9. Complete measured data. The missing traces were not measured, no traces have been removed.

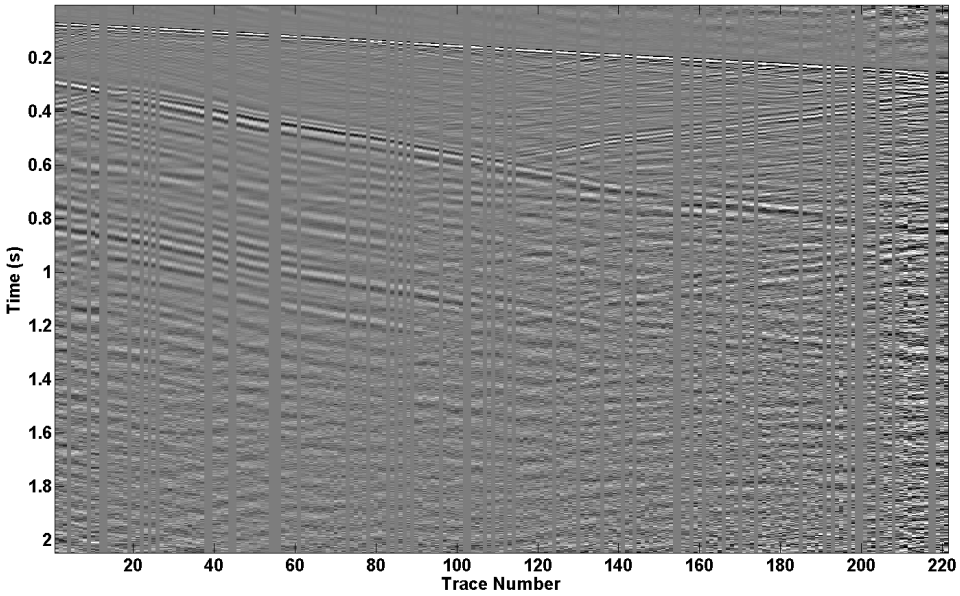

FIG. 10. Measured data with random traces removed.

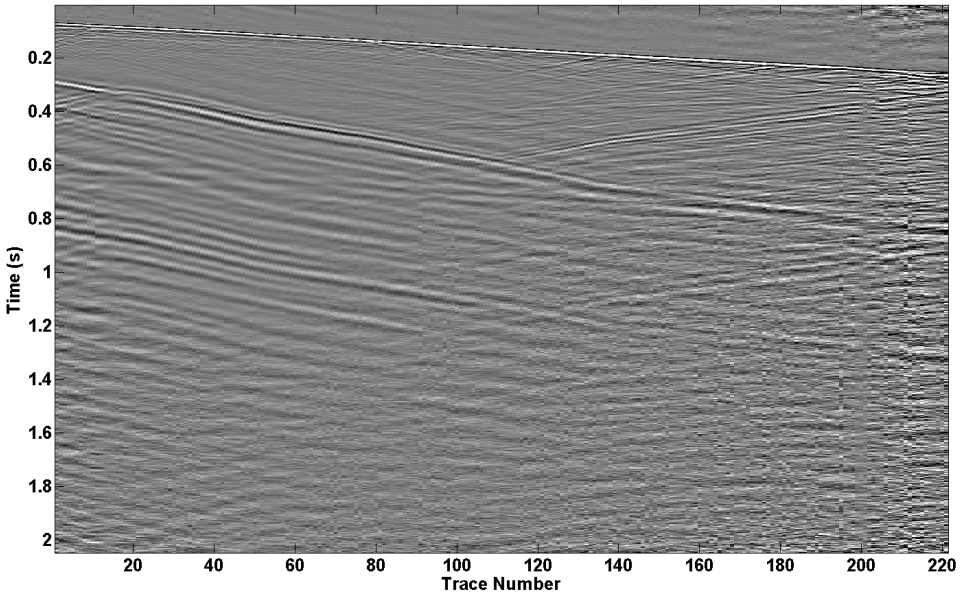

FIG. 11. Incomplete data from Figure 10 after interpolation.

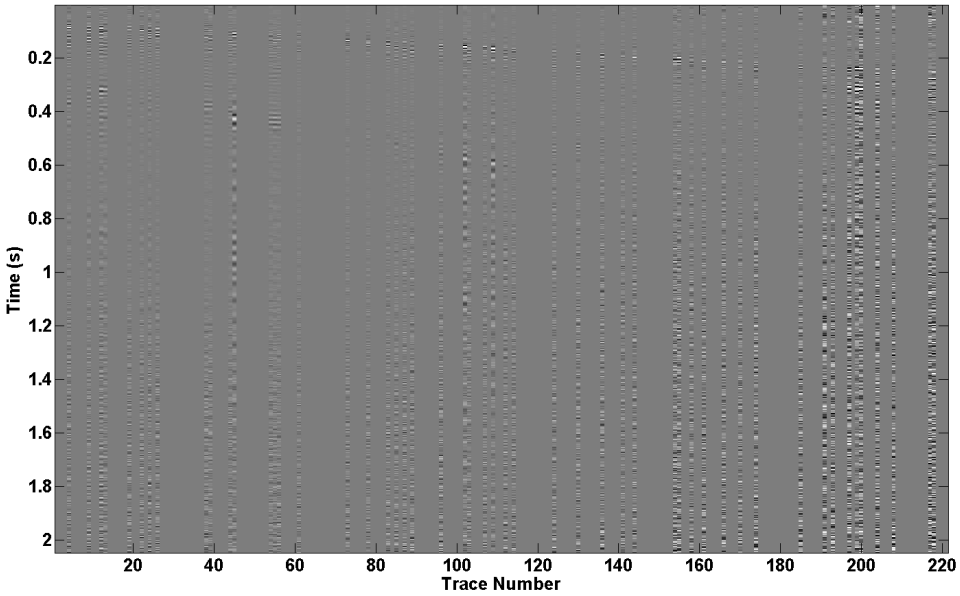

FIG. 12. Difference between the measured data (Figure 9) and complete data (Figure 11).

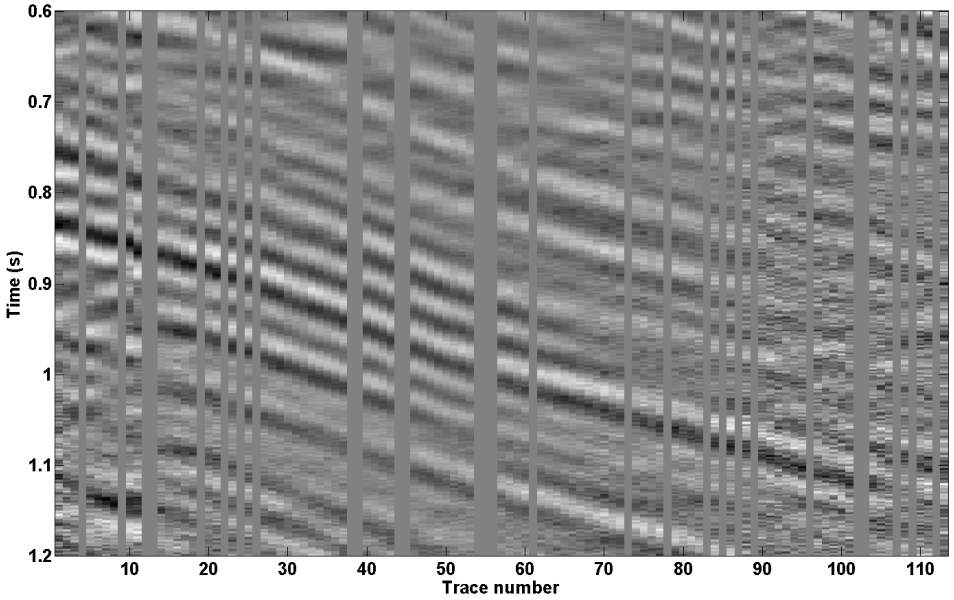

FIG. 13. Subsection of the incomplete data from Figure 10.

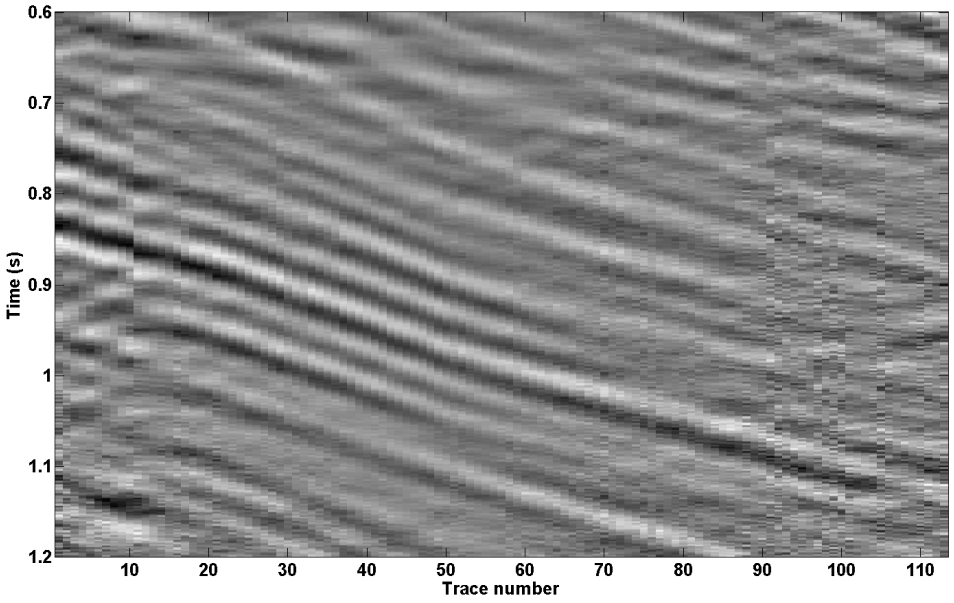

FIG. 14. Subsection of interpolated data corresponding to the missing data in Figure 13. Here the interpolation is largely successful. Even in the noisy traces on the right, the data has been interpolated successfully.

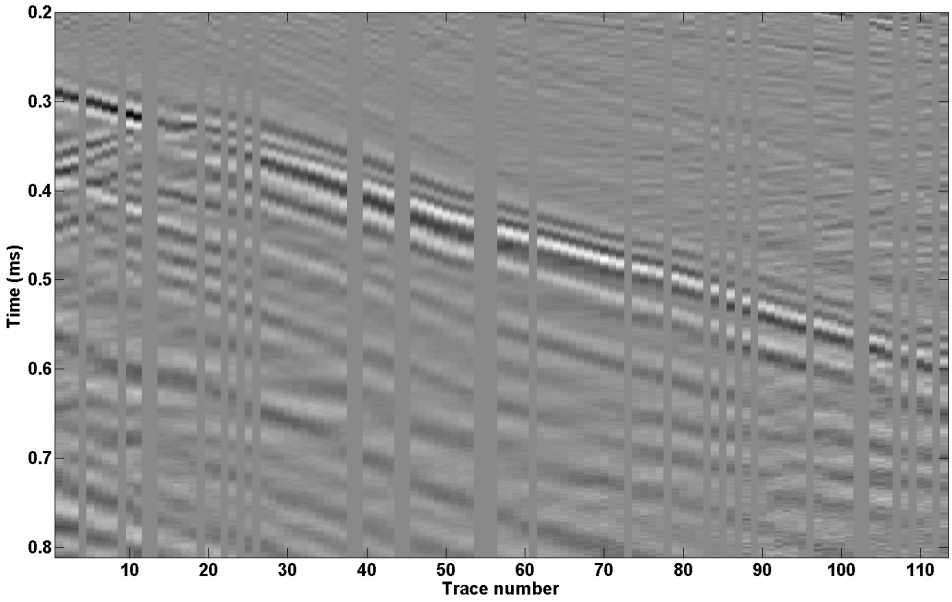

FIG. 15. Subsection of the incomplete data from Figure 10.

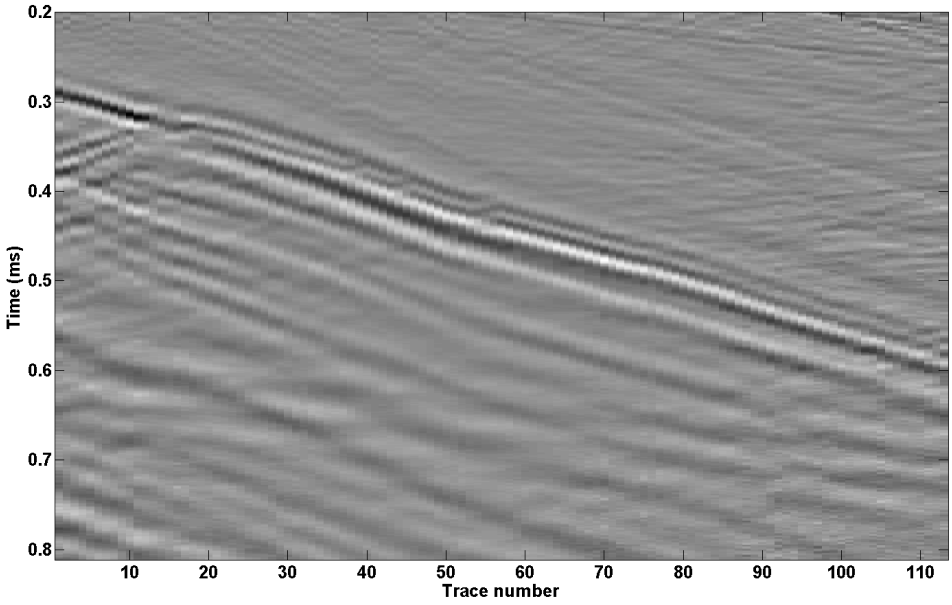

FIG. 16. Subsection of interpolation corresponding to the missing data in Figure 15. Here, some errors in the interpolation are present. Between 0.35 ms and 0.45 ms, trace 35 to 55, there are three areas where the interpolation fails.

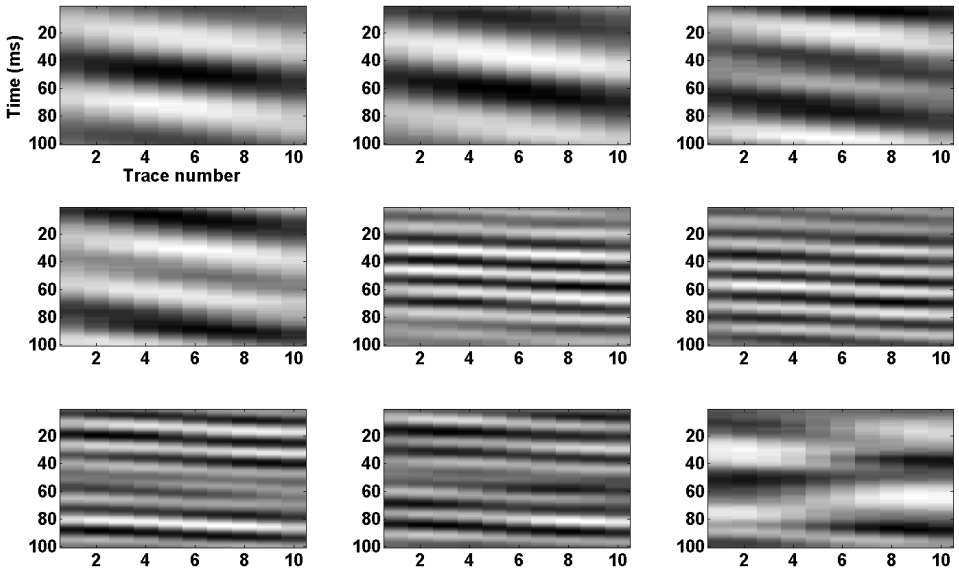

FIG. 17. First nine principal components used to interpolate the VSP data.

Looking at the principal components of the data themselves may provide us with insight about the data. For the VSP data analyzed for this research, the lowest order principal components strongly resemble plane waves (Figure 17). This supports the argument that for these particular data, a Fourier based POCS approach would yield good results.

Although the measured data from this VSP survey have been described as 'complete' thus far, there are several missing traces in these data. This is interesting, because we do not expect incomplete data to generate principal components that work well for interpolation. Only 3 of 221 traces are missing in every gather, but the fact that this interpolation method still works when the principal components are taken from incomplete data is suggestive. As a test, PCA POCS interpolation was attempted using the principal components of nearby shot gathers with traces removed. For small numbers of missing traces, results are comparable to those using principal components constructed from the complete shot gathers. For example, using 90% complete shot gathers allows for a very similar result to using complete shot gathers. As the number of missing traces from these shot gathers grow, the interpolation does decrease in quality, but it continues to interpolate correctly to a surprising extent. In Figure 18 the data from Figure 10 is interpolated, using 70% complete shot gathers to create the principal components. While close inspection shows that this interpolation is not as good as the complete principal component interpolation, it still reconstructs most of the data correctly.

DISCUSSION

While it may be unrealistic to assume that nearby shot gathers are significantly more intact than the shot gather being interpolated, the success of interpolations that use the nearby shot gathers for principal components demonstrates that it is not necessary to use the exact principal components of the data being interpolated for the interpolation to be

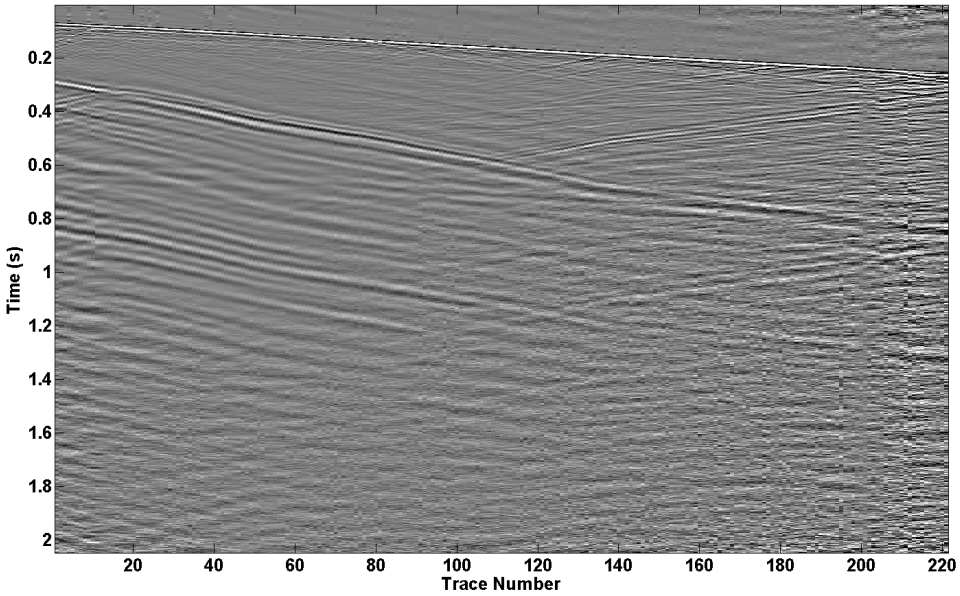

FIG. 18. Interpolation of the data from Figure 10 using principal components from 70% complete shot gathers.

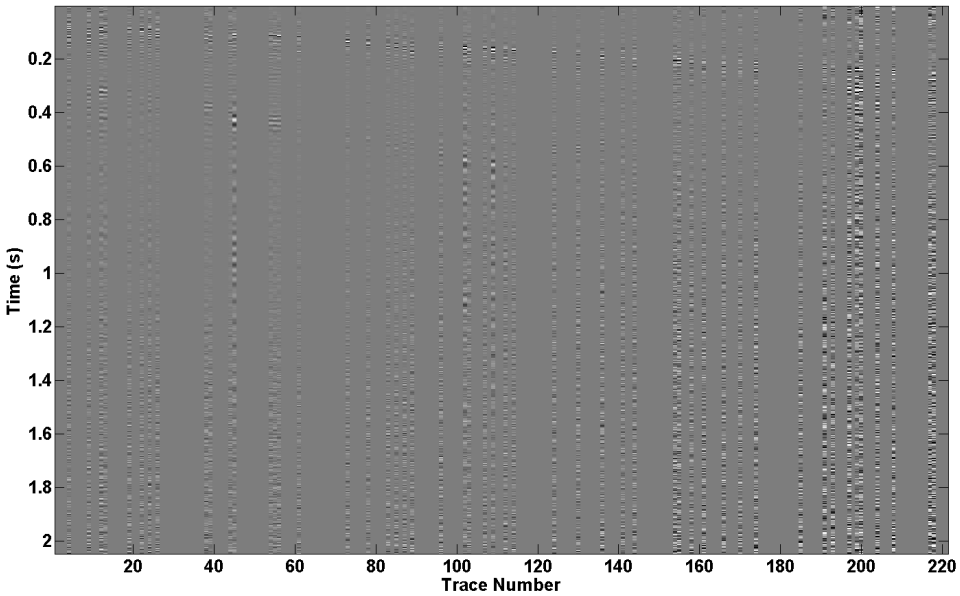

FIG. 19. Difference between the incomplete data in Figure 10 and the interpolated data in Figure 16.

successful. This means that we should be able to use any intact data that displays the same trends as we expect to see in our interpolated data for principal components. Alternatives to nearby shot gathers could be nearby seismic surveys, or even modeled data for the region.

The success of interpolations using principal components from incomplete data indicate that in areas where there are a relatively small fraction of missing traces, it may be practical to interpolate in this manner.

Errors in the interpolation are likely a consequence of inadequate representation of these regions in the principal components. If a more sophisticated method of obtaining principal components than the simple windowing employed here were used, we would likely see an improvement in the quality of the interpolation.

FUTURE WORK

One major area of work that could be further explored in this topic is the construction of the principal components from other sources. While several alternatives are proposed here, evaluating their actual effectiveness is important for gauging the usefulness of this method.

The work in this research has focussed on qualitatively evaluating the effectiveness of the interpolations. Although this is a valid way of distinguishing a very good interpolation from a very bad one, it is still very imprecise, especially when interpolations are similar in quality. A future topic of study is obtaining quantive measures of the interpolation quality. This will be especially important for evaluating the effect of missing traces on principal component creation.

CONCLUSION

POCS is a simple, straightforward method of interpolation which hinges on a few basic assumptions. In order to better fulfill these assumptions, a type of POCS is proposed in which a projection onto principal components is the transform used. This method is shown to be effective both on simple synthetic data, as well as on real VSP data. The problem of acquiring the relevant principal components is adressed by using nearby intact data, and other methods are proposed if this should prove impractical. Partially incomplete data is shown to provide adequate principal components for interpolation in some cases.

ACKNOWLEDGEMENTS

I'd like to thank Mostafa Naghizadeh for providing the codes which introduced me to POCS, the sponsors of the CREWES group for funding and support, and the undisclosed company which provided the VSP data.

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13. Scott Keating was also supported by the Queen Elizabeth II scholarship. We also thank Mostafa Naghizadeh for his initial help with the POCS algorithm. The undisclosed CREWES-sponsoring company

which provided the VSP data is also gratefully acknowledged.

REFERENCES

Abdi, H., and Williams, L., 2010, Principal component analysis: Wiley Interdisciplinary Reviews: Computational Statistics, **2**, 1–46.

Abma, R., and Kabir, N., 2006, Minimum weighted norm interpolation of seismic records: Geophysics, **69**, No. 6, E91–E97.

Liu, B., and Sacchi, M., 2004, Minimum weighted norm interpolation of seismic records: Geophysics, **69**, No. 6, 1560–1568.

Spitz, S., 1991, Seismic trace interpolation in the f-x domain: Geophysics, **56**, No. 6, 785–794.