Stanford Exploration Project, Report 124, April 4, 2006, pages 1–9

Short Note

Inversion shortcuts by model statistics

Brad Artman

INTRODUCTION

For the purpose of data analysis, various transforms have been proposed in order to provide a sparse model space with intuitive or quantitative value for interpretation. For the purposes of velocity analysis or understanding the source components captured passively by an array of geophones a sparse model space is desired that need not exactly forward model the supplied data. This is the paramount difference between data analysis and data synthesis. Because the rate of convergence of an inversion scales with the size of the model and data spaces, 3D problems supplied with large data sets and model domains, are computationally intensive operations. To assure sparsity in the model space, the situation is often exacerbated by using expensive inversion algorithms such as linear programming or BFGS.

Lloyd's algorithm (LA) is an iterative binning operation normally implemented on the histogram of values within a data space. The algorithm is used to decimate the bandwidth of signals in an optimally representative manner. It was developed to quantize/downsample the color values in images for display on graphics systems with limited memory/bandwidth.

The hypothesis of this work is to test whether the algorithm can be used to optimally select a small number of model space coordinates from an incomplete inversion. To test the hypothesis, I stop iterative inversion with linear and hyperbolic Radon transforms before convergence. I then translate the model space into a form usable by LA to select model-space coordinates that best represent the incomplete inversion. The goal is to minimally represent important model-space parameters despite the lack of focus of the incomplete inversion.

Data for Lloyd's algorithm (LA) consist of a set of N-dimensional parameters over which the algorithm optimally selects a user-supplied number, or fewer, combinations that best represent the set. The model space of a linear operator however contains a spanning parameter set differentiated by the amplitude at each location. Consider a model space defined as the Fourier transform of a trace with two sinusoids with different frequencies. The output of the transform contains two frequencies with high amplitude and many with zero amplitude. Viewing the output, an interpreter can select the two frequencies with energy and discount the rest of the model space. Only two numbers are important to know, while the rest of the transformed space can be discarded. I introduce LA to make this selection optimally and automatically.

The transforms are cast within the framework of least-squares inversion with time domain

operators. I test the hypothesis on synthetic volumes, a shot-gather from the Yilmaz data collection, the passively collected telescopic solar observation data, and the passively collected hydrophone data from the Valhall oil field in the North Sea.

IMPLEMENTATION

A set of model-space coordinates must be pre-selected for input into LA according to the amplitude (squared) at all coordinate locations. Only coordinates with an amplitude greater than a supplied percentage of the maximum amplitude in the model space are kept. The amplitude range of model space is then quantized from the minimum threshold to the maximum (squared) value. Coordinates are repeatedly selected for input into LA according to the number of quantum levels associated with its amplitude.

Figure 1 shows graphically how the data input into LA is selected from a sampled 1D function. The continuous signal is assumed to be sampled only at the tick marks on the axis and therefore has 14 amplitude values. The heavy dark line above the axis represents a threshold of approximately 25% of the maximum amplitude, shown by the uppermost dashed line. The data values have been quantized into three levels above the threshold. The dots show how many times each coordinate is selected for input into LA. The coordinate set formed from the signal is therefore $\{4,5,5,6,6,10,11,11,11,12,12,12,13,13\}$.

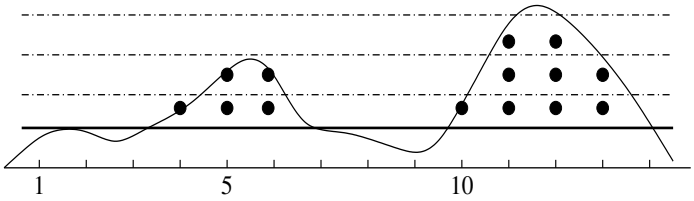

Figure 1: Dots represent coordinate selection as a function of squared signal amplitude for input into Lloyd's Algorithm. brad1-data [NR]

Assuming that the linear operator chosen for the transform is appropriate, the model space should be mostly low amplitude or zero. Therefore, the pre-selection of coordinates for LA reduces the model space by several orders of magnitude. The selected model-space coordinates are input into LA whose output is a list of optimal locations that represent the energy in the model space of the transform. These are written out to a SEP77 file and a ASCII file formatted for input into Ricksep as a picks file.

SYNTHETIC EXAMPLE

The simple adjoint transform is not sufficient to provide as input to LA. Figure 2 shows results when supplying LA with model spaces produced by adjoint plane-wave decomposition verses 40 iterations of inversion with a single synthetic planewave used as data. LA, told to choose at most 10, returned two picks for both models despite only a single planewave existing in the data. LA will not return one pick if the data supplied is not perfectly single valued. In this example the data supplied to LA are coordinates with energy in the figure times the duration of

the wavelet in/out of the plane. This leads to 3449 coordinate triples, some 500 being unique since the quantization of the amplitudes used 10 levels.

The adjoint model space has a diffuse character and diagonal streaks away from main blob of energy (due to limited rectangular surface acquisition). The model space was parameterized so that the energy would not be symmetrically located on the p_x, p_y-plane. To balance the distribution of energy, LA selected two points for the adjoint model space that an interpreter would recognize as inappropriate. After the inversion has clipped most of the acquisition tails and moved the edge of the distribution away from the boundary of the domain, the two picks returned are identical $(-0.0005006, 0.0001002)\&(-0.0004993, 9.997e-05)$. [1]

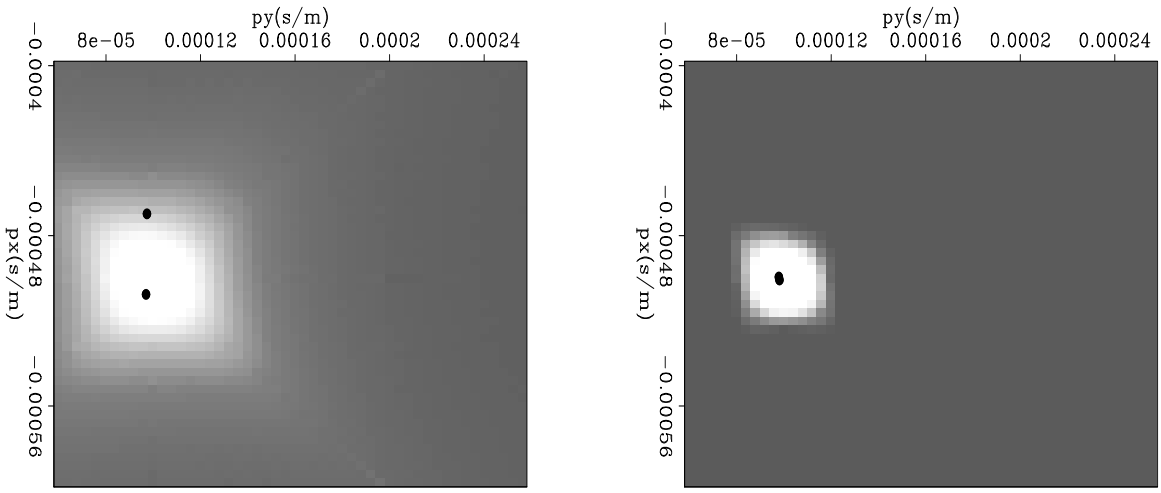

Figure 2: Coordinates selected by Lloyd's Algorithm on the adjoint plane-wave decomposition model space and on 40 iterations of least-squares inversion. brad1-adj.inv [CR]

Six planewaves were modeled in a regular 3D acquisition geometry as data. The data were inverted with the linear Radon transform. The inversion was stopped after 1, 20, and 80 iterations. These data were then supplied to the modified LA. Figure 3 shows inversion results after 20 iterations. The two panels show versions clipped for display at 99 and 100% of the maximum value. Three picks from LA are also plotted. The picks exactly overlay the maximum amplitude of the energy. LA was initialized to select 10 coordinate triples. Only six picks were returned. Note, however, that the lower right coordinates are actually two picks very close together. One planewave from the data space has not been picked. One of the planewaves had a ray parameter $p_x, p_y = (0.0005, 0.0001)$ s/m, while the range of p_x used for the inversion extended to only $p_x, p_y = (0.00045, 0.0003)$ s/m. This plane thus falls outside of the model space on the 2-axis, and the inversion is not able to focuse the energy. Unfortunately, LA does not recognize its significance either, and places an extra pick semi-randomly close to another well established pick.

The approach seems robust for noisy data as well. Figure 4 shows a section of the model

[1] LA returns coordinates representing the center of mass of the energy in the model domain not beholden to the sampling interval of the domain. This lead me to an as yet unsubstantiated hypothesis that LA could provide super-resolution.

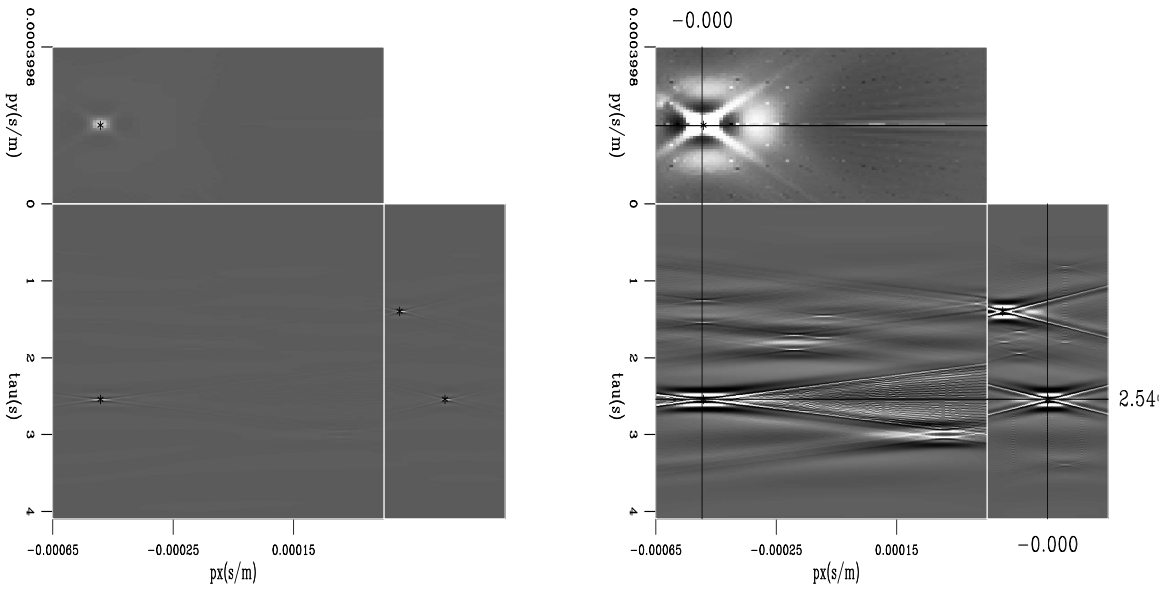

Figure 3: Ray parameter model space inverted from six synthetic planewaves. 20 iterations of least-squares inversion applied. brad1-rays [CR]

space where a uniform distribution of noise was added to the data space with the picks selected from it by LA overlain. The picks are identical to the previous runs when the variance of the added noise is less than 0.001. The threshold value used was 1% of the maximum value in the data. By increasing to higher levels (approximately 50%), the algorithm remains stable to variance values another order higher. When the level of noise is too high so that the thresholding of the data is not robust, the LA picks constantly distribute themselves roughly evenly along the one-axis and about centered across higher dimensions.

PASSIVE DATA

To understand the sources captured in a passive seismic survey, I hoped planewave decomposition/inversion could help analyze the data. However, unless identifiable events are present, analysis of passive data before cross-correlation does not produce interpretable results. After correlation, the unique character of the individual sources is lost. In effect, transforms applied to the data before correlation simply reshuffle the randomness apparent in the raw traces. The passively collected solar data (Rickett and Claerbout, 1999) was analyzed to prove this failure. Figure 5 shows the raw solar data and its autocorrelation. Clearly, there are events to be found within the raw data that are masked before correlation. Figure 6 shows the linear Radon domain inversions for data defined by the panels of Figure 5. Because the correlated data is radially symmetric from the center and only has events in the upper third of the time axis, its model space is much smaller, though sampling between the two is the same.

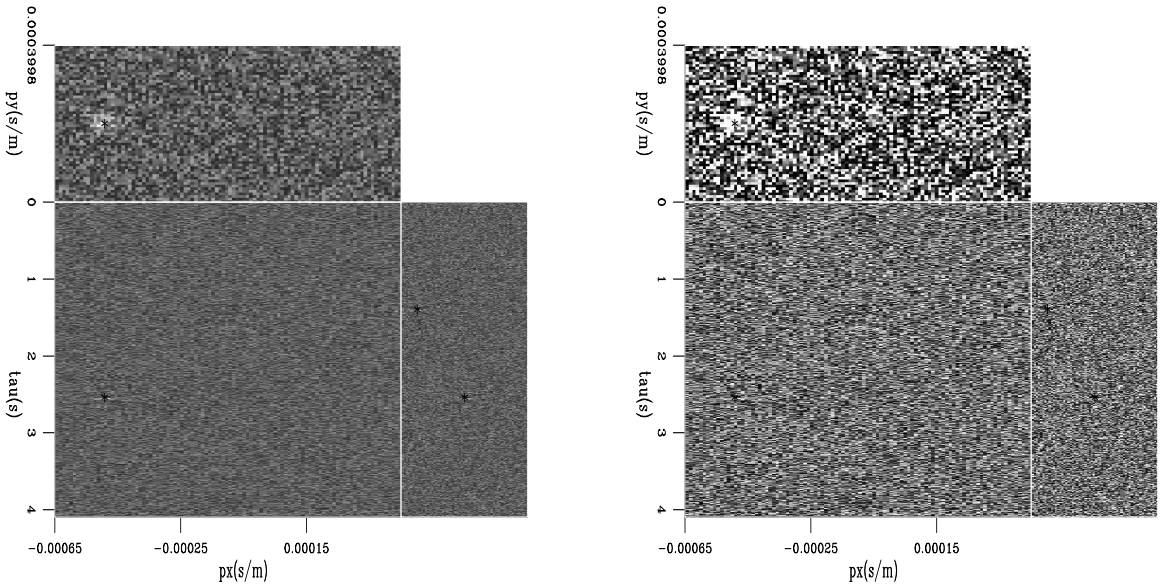

Figure 4: Uniformly distributed noise with variance 0.001 added to model space shown in Figure 3. Threshold for LA was 1% of the maximum value in the input. Output picks are still reliable at this level. brad1-noise [CR]

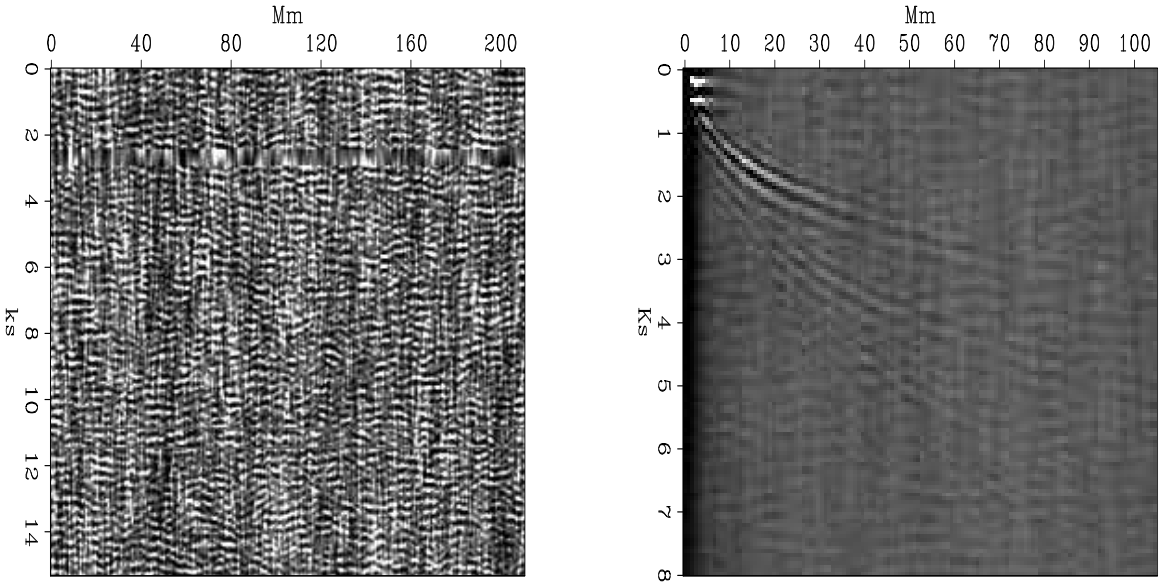

Figure 5: Passive seismic data from the sun and its autocorrelation. brad1-sun.dat [CR]

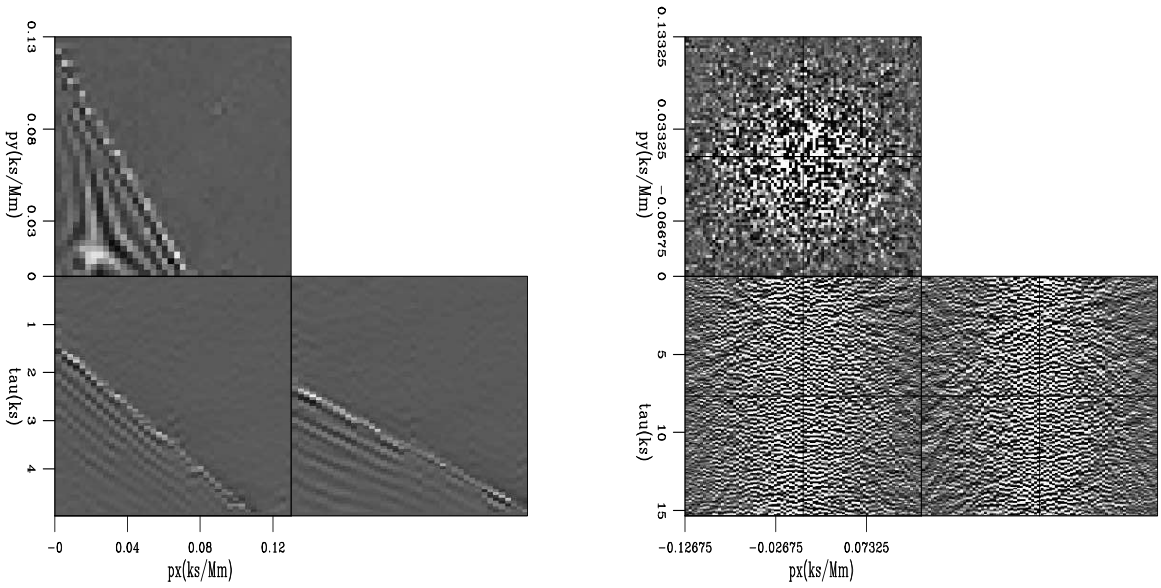

Figure 6: Model space produced with 20 iterations of planewave decomposition inversion from the data shown in Figure 5. brad1-sun.mod [CR]

VELOCITY ANALYSIS

Inversion schemes for various Radon transforms (Artman and Guitton, 2005; Artman and Sacchi, 2003; Guitton, 2004), have been proposed to provide sparse model spaces for data analysis and noise removal. Figure 7 shows the result of 20 iterations of hyperbolic Radon transform performed on gather 08 from the Yilmaz data collection. Coordinates selected by LA are overlain. I purposefully parameterized the model space to include slowness values twice those necessary. The picks remain stable with or without including this aphysical part of the model domain. With a threshold of 1% of the maximum value in the model space, the modified algorithm considered only 500 out of 43,000 coordinates for potential selection. The algorithm started with 200 evenly distributed coordinate pairs and returned with 14 optimally representative coordinates.

CONCLUSIONS

I introduce Lloyd's algorithm as a tool to optimally represent the statistics of the model space from incomplete inversions. The modified algorithm selects coordinates with high amplitude surrounded by substantial energy. Thus isolated, powerful outliers are neglected. By optimally parsing potentially large multi-dimensional model spaces, the algorithm can cut short costly inversion iterations and focus an interpreter's attention to important locations within potentially large model domains. The algorithm returns stable solutions even in the presence of substantial noise.

The algorithm is very simple, easy to modify, has few parameters, and very fast. Using the

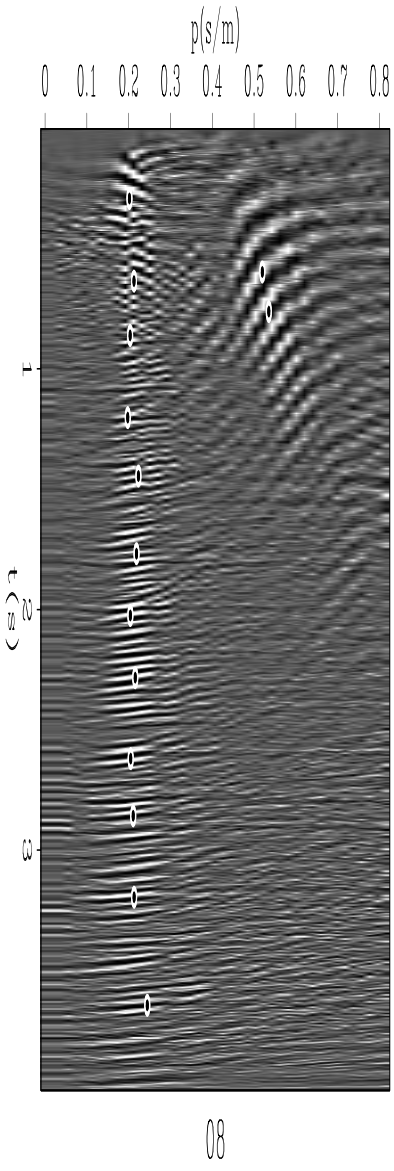

Figure 7: Hyperbolic Radon transform model domain of shot 08 from the Yilmaz data collection. 20 iterations of least-squares inversion were performed. 14 coordinates selected by LA are overlain. Without inclusion of the inappropriate high slowness values on the right side of the plot, the remaining picks remain stable. brad1-shot [CR]

algorithm depends on parameters being correlable. For multidimensional cases, uncorrelated parameters can simply be concatenated to an existing axis. Thus hypercubes of correlable and uncorrelable parameters can be evaluated simultaneously.

The next step in evaluating the effectiveness of using the algorithm to select optimal parameters would be to migrate data with a velocity model derived from RMS velocities selected by LA. This could potentially dovetail with the velocity uncertainty analysis presented in **?**.

Planewave decomposition of passive data to characterize non-obvious sources does not work. Until traces have been correlated, analysis transforms will simply redistribute the random character of the raw data. Unfortunately, correlating the wavefield destroys all the unique character of individual sources including timing, waveform, location and much of the spectral content information.

ACKNOWLEDGMENTS

Thanks to Bob Clapp for the introduction of LA to the group and good discussions with Jeff Shragge on potential uses.

REFERENCES

Artman, B. and A. Guitton, 2005, Removal of linear events with combined radon transforms: SEP–**120**, 395–406.

Artman, B. and M. Sacchi, 2003, Basis pursuit for geophysical inversion: SEP–**114**, 137–150.

Guitton, A., 2004, Bound constrained optimization: application to the dip estimation problem: SEP–**117**, 51–62.

Rickett, J. and J. Claerbout, 1999, Acoustic daylight imaging via spectral factorization: Helioseismology and reservoir monitoring: SEP–**100**, 171–180.