A Heuristic Approach to use Sparseness in Linear Inversion Problems

Adelson S. de Oliveira, José L. Carbonesi, Paulo Casañas Gomes, Thiago S. F. X. Teixeira, PETROBRAS S/A, Brazil

Abstract

The sparseness in seismic data has been successfully used to achieve many improvements in different areas such as acquisition, regularization, filtering, and imaging. The search for a sparse description for a given practical seismic (linear) problem is often conducted via the so called Iteratively Reweighted Least-Squares Inversion with an adaptive choice of a diagonal matrix, computed with a statistically derived prescription. In this paper, we discuss the role of this matrix in the inversion and propose an alternative heuristic formula which allows one to more easily constrain the solution to physical expectations and is likely to improve sparseness and accelerate convergence.

Introduction and Review

Linear inversion problems are very common in geophysical applications. Usually, a set of representative vectors are combined in a weighted way so as to fit a set of measurements. Mathematically, it is to say $\mathbf{A}\vec{x} = \vec{b}$, where $\mathbf{A}$ is a matrix with model vectors in its columns, $\vec{x}$ is a vector of weights to be determined, and $\vec{b}$ is the data vector. The choice of the model vectors is generally guided by physical demands and is not committed to fill simple mathematical properties like completeness or uniqueness of the solution $\vec{x}$. For instance, in an irregular discrete Fourier transform (IDFT), columns of $\mathbf{A}$ exhibits the values of sinusoidal functions taken at unevenly chosen positions. Thus, the matrix $\mathbf{A}$ for IDFT departs a lot from the usual discrete Fourier transform. $\mathbf{A}$ may not be invertible and the solution $\vec{x}$ may not be unique.

Least square (L2) error ($\|\mathbf{A}\vec{x}-\vec{b}\|^2$) is an option to define $\vec{x}$ for overdetermined problems. Developing a linear problem to L2 minimum error norm leads to equations like $\mathbf{A}^H\mathbf{A}\vec{x} = \mathbf{A}^H\vec{b}$ ([1]). Underdetermined and/or ill-posed problems are often handled with a dumped version of the L2 error norm (here called the **Dumped Least Square equation** or DLSE) as $(\mathbf{A}^H\mathbf{A}+\lambda\mathbf{I})\vec{x} = \mathbf{A}^H\vec{b}$ where λ is ideally chosen real, positive, and small enough to allow for a solution that keeps the error as small as possible, and $\mathbf{I}$ is the identity. At this point, it might be noticed that extending the original linear problem as,

$$\begin{bmatrix} \mathbf{A} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{b} \\ \vec{0} \end{bmatrix} \tag{1}$$

where $\vec{0}$ is a null vector, the Dumped Least Square equation remains unchanged. The least square error turns out to be $\left\|\mathbf{A}\vec{x}-\vec{b}\right\|^2 + \lambda\,\|\vec{x}\|^2$, which shows that this formulation will demand solutions of smaller L2 norm. The columns of the matrix $\sqrt{\lambda}\mathbf{I}$ forms a linearly independent set of vectors. Consequently, the extended matrix $(\mathbf{A}^T, \sqrt{\lambda}\mathbf{I})^T$, on the left hand side of equation (1) also has a set of linearly independent vectors as its columns. This is why the DLSE is invertible.

The Dumped Least Squares solution is unique but may not match all practical needs. It would be helpful to add extra constraints, and have a yet unique but also appropriate solution. In many applications, mitigating redundancy is satisfactory. Thus, one should look for the most sparse solution. Among many authors, Sacchi (1997) have a comprehensive analysis of how to impose sparseness on a linear problem. Based on assumed probability density functions (PDF) for the error and data, it consists in giving λ a dependence on the solution $\vec{x}$. This turns the linear problem into a nonlinear one and an iterative method is proposed to find the best $\vec{x}$. The DLSE (although the error might no longer be the least) assumes the form,

$$(\mathbf{A}^H\mathbf{A}+\mathbf{\Lambda})\vec{x} = \mathbf{A}^H\vec{b} \tag{2}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with elements generically given as $\Lambda_{ij} = \delta_{ij}\lambda_i(x_i)$, δ is the Kronecker delta and now λ stands for a function determined to honor a desired PDF. For instance, in a Gaussian distribution of errors and data and Cauchy distribution of weights, one has,

$$\lambda_i(x_i) = \frac{\varepsilon}{1+\frac{x_i^2}{\sigma_x^2}} \tag{3}$$

with σ_x^2 the variance of a given solution and ε a small parameter used to keep the original linear problem error small.

Now, the error or total residue is written as,

$$\left\|\mathbf{A}\vec{x}-\vec{b}\right\|^2 + \sum_i \lambda_i x_i^2 \tag{4}$$

explicitly associating the total residue [2] to the values of λ_i and the corresponding weights x_i. Manipulation of λ_i values seems to be the key for guiding the iterative process to expected goals. A varying λ_i will likely yield a solution that is no longer minimum in a L2 norm sense. This will be clear for a simple linear problem with dimension three where the L2 norm tends to increase when one updates λ_i accordingly to expression (3) in a following section.

[1] H stands for Hermitian, we assume complex valued functions.

[2] Now, $\vec{x}$ minimizes the functional $\left\|\mathbf{A}\vec{x}-\vec{b}\right\|^2 + \varepsilon\sum_i \log\left[1+\frac{x_i^2}{\sigma_x^2}\right]$

There is a vast list of authors that claims DLSE, with the implied L2 minimum norm of the solution, and even the iterative process of updating λ_i like proposed in expression (3) are not able to promote the desired degree of sparseness. It is very common to change the DLSE so as to achieve a minimum L1 norm of the solution instead (see references to the so called LASSO problem). This may be, approximately, carried out in an iterative way with a set of weights $\Lambda_{ij} = \sqrt{\varepsilon}\delta_{ij}|x_i|^{-1}$. From equation (4), it can be seen that such a choice for λ_i will indeed force convergence to a solution that minimizes,

$$\left\|\mathbf{A}\vec{x} - \vec{b}\right\|^2 + \varepsilon \sum_i |x_i| \quad .$$

However, although expressions like (3) might not help finding a properly sparse solution, different norms like L1 (or any other) are prone to promote a sparse but with no physical appealing solution. An example where the mathematical property called sparseness is of no meaning would be that of mapping a regularly sample data beyond aliasing. In this case, two or more columns of $\mathbf{A}$ would be identical and there would be no reason why sparseness would yield one particular solution.

On the other hand, a statistical choice of λ_i is not the only option. A pragmatic analysis of what a particular set of λ_i leads to in an inversion problem may offer different choices. Fast convergence may be one additional requirement to be filled in by an alternative choice.

A brief discussion of this process and a heuristic choice for λ_i is the subject of this paper.

Before going further, let's remember and emphasize that **dumping is safe** in respect to the error expressed in equation (4). It is important to notice that inadequate choices of λ_i cannot be of great harm. In fact, the error expressed by equation (4) will be small for any given set of sufficiently small λ_i's.

A geometrical view for the role of a varying λ

An analysis of the role different λ_i's play in the choice of a particular solution for the DLSE and in the rate of convergence of the iterative process is somewhat difficult if one looks only to equations (2) or (4). From (4) it may be seen that smaller values of λ_i allows for greater values of x_i and vice versa. Expression (3) exhibit this behavior but no more details are available. Moreover, the searched weights are obtained in an inversion process, which, by definition, tends to "boost small features". An attempt to predict what the inversion process will produce on the weights is presented below.

For a varying set of λ_i, equation (1) may be rewritten as,

$$\begin{bmatrix} \mathbf{A} \\ \mathbf{\Lambda} \end{bmatrix} \vec{x} = \overline{\mathbf{A}}\vec{x} = \begin{bmatrix} \vec{b} \\ \vec{0} \end{bmatrix} \quad . \tag{5}$$

For all $\lambda_i \neq 0$, corresponding columns of $\overline{\mathbf{A}}$ are linearly independent from the others, by construction. The linear independence of $\lambda_i = 0$ columns depends only on the rank of the original matrix $\mathbf{A}$. Then, if the rank of $\mathbf{A}$ is p and the dimension of $\vec{x}$ is n, the minimum and sufficient number of non zero λ_i's to have linear independence of columns is $n-p$.

The solution $\vec{x}$ is obtained by left multiplying (5) by the Hermitian of a matrix $\widetilde{\mathbf{A}}$ that has in its columns a set of vectors that are dual to the columns of $\overline{\mathbf{A}}$. It means

$$\sum_k \widetilde{A}^*_{ki}\overline{A}_{kj} = \delta_{ij} \quad . \tag{6}$$

A dual matrix $\widetilde{\mathbf{A}}$ always exist if at least $n-1$ values of λ_i is greater than zero. A dual matrix do exist if at least $n-p$ values of $\lambda_i \neq 0$, provided that a proper choice of i's is used.

In the general case, the columns of $\mathbf{A}$ may not represent a frame. That is, not all possible data vectors $\vec{b}$ can be written as $\mathbf{A}\vec{x}$. Let $\vec{\beta}$ be the component of $\vec{b}$ that can be fully described as a linear combination of the columns of $\mathbf{A}$. Since no choice of λ_i may overcome this limitation, in this case, it is only true that,

$$\vec{x} = \widetilde{\mathbf{A}}^H (\vec{\beta}^T, \vec{0}^T)^T \quad . \tag{7}$$

Finally, we can understand the role of a particular choice of λ's as follows. Consider the simplest case of a 2X3 matrix $\mathbf{A} = (\vec{u}\,\vec{v}\,\vec{w})$ with rank 2 (see figure 1). Setting $\lambda_2 = \varepsilon$ for the second column, makes the corresponding column of $\overline{\mathbf{A}}$ to move out of the plane spanned by the other two columns. Since relation (6) must hold, the second column of $\widetilde{\mathbf{A}}$ is forced to be perpendicular to the other two columns of $\overline{\mathbf{A}}$. Since $\vec{\beta}$ is a combination of these two remaining columns of $\mathbf{A}$, the weight x_2 is set to zero. Extending to matrices of higher dimensions is straightforward, if the number of nonzero values of λ_i is strictly $n-p$, one can choose different solutions for the linear problem $\vec{x}$ by simply choosing which p columns will be assigned the zero values of λ_i. It is a binary filter that establishes which weight will and will not be zeroed out.

Apart from numerical precision limitations, this binary selection will take place for very small values of λ_i. Thus, it can be done with little influence to the error given in equation (4).

A choice of λ_i's to match specific purposes, in a binary form as proposed, would require an analysis of possible clusterings of orthogonal subspaces of matrix $\mathbf{A}$. Such an analysis is of little practical interest due to inherent CPU demands. Furthermore, this analysis itself would provide more reliable choices of $\vec{x}$ than Reweighted Least Square approaches.

There is another interesting note to be made about the matrix $\widetilde{\mathbf{A}}$. If $\mathbf{A}$ is a $m X n$ matrix and its columns forms a frame for the space of vectors of dimension m, then the first m lines of $\widetilde{\mathbf{A}}$ forms also a frame that is dual to the frame in $\mathbf{A}$. Let $\widehat{\mathbf{A}}$ be the matrix formed by the first m lines of $\widetilde{\mathbf{A}}$. Hence for a frame $\mathbf{A}$ we must have $\mathbf{A}\widehat{\mathbf{A}}^H = \mathbf{I}$.

A Boolean-like distribution of λ's

The safety in dumping and the discussion in the last section encourages a search for a set of λ_i's that could have a Boolean-like result at the choice of the weights $\vec{x}$. This means finding the most sparse solution under a given physical constraint. The idea is to set to zero all λ_i's corresponding to weights x_i's to be used in the modeling. Without a priori information or a physical constraint, the iterative Reweighted Least Square method assumes that the model

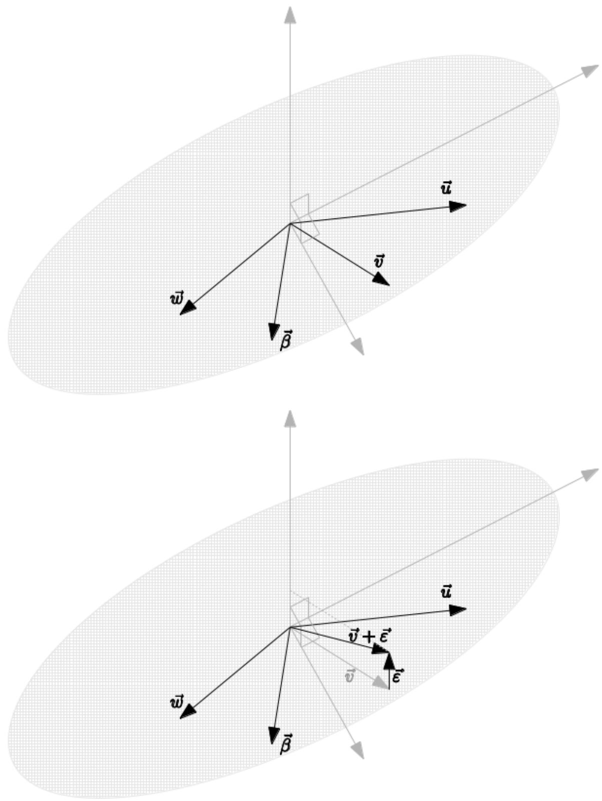

Figure 1: *A graphical representation of three columns of $\overline{\mathbf{A}}$; At the top, linearly dependent, all $\lambda_i = 0$. At the bottom, linearly independent, only $\lambda_2 \neq 0$.*

"prefer" one weight to another, in any iteration, so that the absolute value of the weights can be used to define the set of λ's to the next iteration.

In any case, one can think of a function that "flags" x_i according to its degree of rejection it has over the others in the $k-1$ iteration, $\alpha_i = \mathscr{F}(x_i^{(k-1)})$, $0 \leq \alpha_i \leq 1$. Then, one could get an almost Boolean distribution of λ_i's, for example, with a heuristic relation like,

$$\lambda_i \;=\; \varepsilon \begin{cases} \alpha_{ref}\left(\frac{\alpha_i}{\alpha_{ref}}\right)^{\gamma}; & \text{if } \alpha_i \leq \alpha_{ref} \\ 1-(1-\alpha_{ref})\left(\frac{1-\alpha}{1-\alpha_{ref}}\right)^{\gamma}; & \text{if } \alpha_i > \alpha_{ref} \end{cases} \quad , \qquad (8)$$

where ε, α_{ref}, and γ were introduced to control, respectively, the greatest value of λ_i, the value of α_i to trigger the selection, and the "degree of acceptance" to be applied.

Figure 2 have six representations of equation (8) for a hypothetical distribution $\alpha = 0.5(1-\cos\theta)$; $0 \leq \theta \leq \pi$, where α_{ref} is 0.5 (50% of the values) and 0.75 (about 66% of the values), and γ is 1, 2, and 5.

The parameter α_{ref} is chosen as the p-th value after sorting all n values of α's in increasing order. To know p may require a considerable amount of CPU which would make all of this work worthless. Fortunately, it is often possible to estimate p under "reasonable" assumptions, and, most important, dumping is safe (small ε) so we can try and adapt.

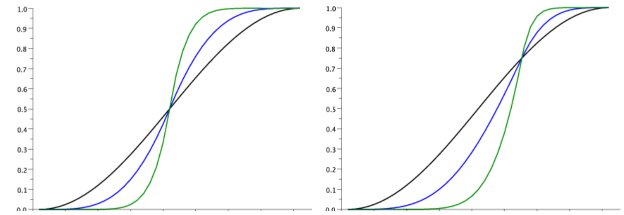

Figure 2: *Six representations of λ_i for $\varepsilon = 1$. Left: $\alpha_{ref} = 0.5$, γ is 1 (black), 2 (blue), and 5(green); Right: $\alpha_{ref} = 0.75$, γ is 1 (black), 2 (blue), and 5(green).*

Synthetic Examples

A 2X3 simple problem:

A simple but clear example could be that of a 2X3 matrix as follows,

$$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 2/\sqrt{5} \\ 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1.5 \end{bmatrix} .$$

In this case it is clear that the columns of the matrix forms a frame for vectors of dimension 2. It is also clear that the rank of the matrix is 2. Thus, the most sparse solution of the problem requires that only one $\lambda_i \neq 0$. Beginning with the usual DLSE equation, using a constant value for λ_i of 0.001, one gets $\vec{x} \approx (1.193, -0.091, 0.829)^T$. Using expression (3) for λ_i's, after the first iteration the solution changes to $\vec{x} \approx (1.299, -0.141, 0.672)^T$ and after 10 iterations the solution has converged (variations smaller than 0.001) to $\vec{x} \approx (1.402, -0.189, 0.518)^T$. The final solution is good up to errors of the order of λ but can be considered as not much sparse. Changes in the value of the variance used in (3) would likely yield another final solution. Adopting a heuristic approach, using the first DLSE solution as starting point to derive a set of λ_i's (no a priori physical constraints), after the first iteration the solution is $\vec{x} \approx (1.001, -0.0005, 1.116)^T$. The last solution is sparse and was obtained with just one iteration.

For this example, the heuristics was simply,

$$\lambda_i \;=\; \varepsilon \left(\frac{\max(x) - x_i}{\max(x) - \min(x)} \right)^{\gamma} \quad , \qquad (9)$$

with $\gamma = 5$ and $\varepsilon = 0.001$.

A spatially irregular $f-k$ transform:

Another interesting example is that of obtaining the IDFT beyond Nyquist. As pointed out in Hennenfent et al. (2007), aliasing is not as severe for fully irregularly sampled data as for regular ones. In their paper, sparseness is used to allow extending the Fourier spectra beyond an estimated Nyquist frequency. Since the "degree of regularity" is not predictable in current applications, a technique dependent only on sparseness may fail. As commented above, it is not possible to achieve a sparse solution for the problem if two or more columns of the transforming matrix are equal.

The $f-k$ transform of linear events exhibits two distinct zones with respect to spatial aliasing. It is well known that linear events in the time-space $(t-\chi)$ domain transform as linear events in the $f-k$ domain as well. It is also known that spatial aliasing in linear events is related to the velocity or dip of the event. The higher the temporal frequency f the smaller the spatial Nyquist frequency for a given dip. This means that, for a given dip, there is a highest temporal frequency f_{knyq} for which no spatial aliasing is observed. Most applications of irregular $f-k$ transforms uses this fact to develop a strategy to handle spatial aliasing. The idea is to compute the weights for all pairs (f,k), $f < f_{knyq}$ and, under the linear events hypothesis, to estimate expected weights for frequencies $f \geq f_{knyq}$. This holds for linear events only but, borrowing properties of curvelet analysis, it is expected to be a good approximation for locally linear events as well.

Predicting weights for higher temporal frequencies is similar to have a priori physically acceptable constraints. The role of a constraint here is to allow for deciding which component of the matrix to use even when alias is severe. A physical constraint may be incompatible with a general statistical principle. This is where a heuristic set of λ_i's is supposed to help. The tests below were made using the heuristics discussed above (equation (8)) only.

Figure 3 has a synthetic seismogram with three linear events in regular but gapped spatial sampling and the corresponding estimate of its $f-k$ transform (constant λ, no sparseness) with a maximum wavenumber three times the expected Nyquist $k \leq 3k_{nyq}$. The imbedded spatial regularity was used to get a more difficult to handle aliasing example.

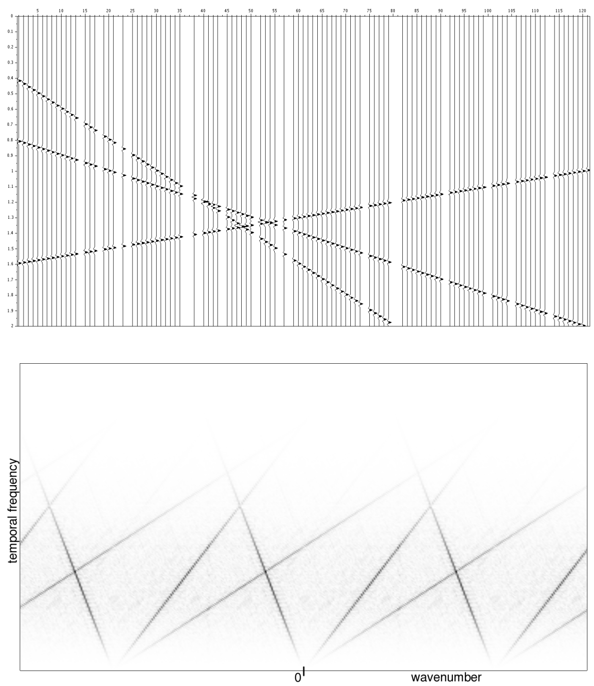

Figure 3: *A seismogram with three linear events. Regular spatial sampling but with gaps. Above, the time-space domain; below, the frequency-wavenumber domain up to three times Nyquist* $k \leq 3k_{nyq}$.

Fourier periodicity and aliasing are clearly present in figure 3. The same type of configuration is found in figure 4 where the gaps were filled without requesting sparseness. In this case, typically Fourier "predicts" null traces in the gaps.

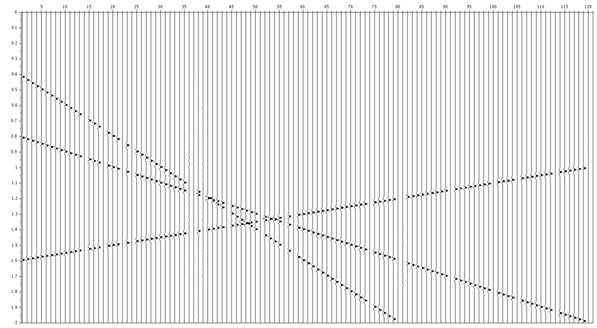

Figure 4: *A seismogram with three linear events. Regular spatial sampling after Fourier interpolation. No sparseness assumptions used. Note that Fourier typically predicts null traces in the gaps.*

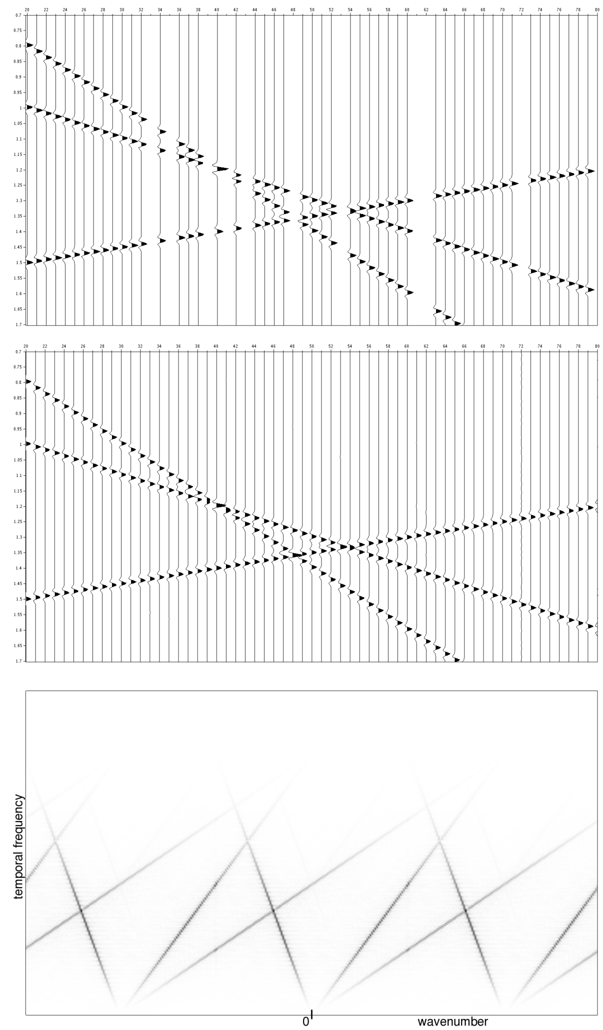

Figure 5: *A zoom of the regular but gapped seismograms before (above) and after (middle) Fourier interpolation with sparseness requirements but no physical constraint used. The* $f-k$ *domain (bellow) is totally similar to that shown in figure 3 with aliasing and periodicity still observed.*

Figure 5 has another regular but gapped seismogram after interpolation requiring sparseness but still with no physical constraint used. Now, for better appraisal of the results, figures are zoomed. This time gaps were filled properly but the $f-k$ domain still exhibits periodic events. The $f-k$ domain looks somewhat cleaner than that shown at figure 3 (better observed in a digital version of this paper). This is what one would have if the original data had no gaps at all. The problems with this estimate of the $f-k$ domain would be noticed if spatial resample is done.

Figure 6 shows the $f-k$ spectrum obtained in a regular and gapped sampling but this time the physical assumption of linear events were used so that it is possible to discard periodic manifestations of an event based on its behavior for smaller temporal frequencies.

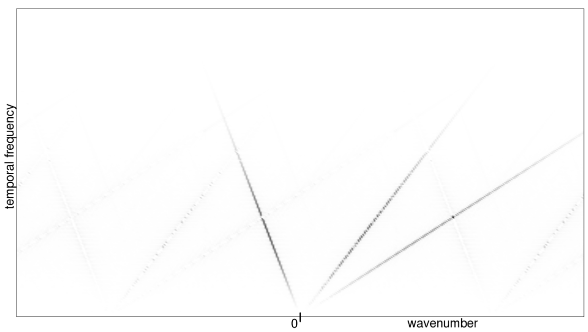

Figure 6: *The $f-k$ spectrum obtained in a regular but gapped spatially sampled seismic section but using a physical constraint. This constraint allows to mitigate the usual periodicity and aliasing.*

The irregularity impact on the aliasing mitigation can be observed at figure 7. In this case, with or without physical constraints, results are more or less the same. The use of a physical constraint in a spatial resampling application is not tested here since it seems out of the scope of this paper. We limit to remark that statistical assumptions are not enough to guide a $f-k$ domain estimate. A physical constraint, with some kind of heuristics for mitigating redundancy, is required.

In figures 3, 4, and 5, the first inversion was made with a constant $\lambda_i = 0.001$ (DLSE). In figures 4 and 5, the DLSE inversion was followed by 5 iterations of inversion using the heuristic expression (8), with $\gamma = 8$, and an estimate of p/n of 0.1. The relation between α_i and the $f-k$ weights was like equation (9) but with $\gamma = 1$. In figures 6 and 7, the first inversion was made with an estimate of α_i from some $f-k$ weights of smaller frequencies, according to the linear event hypothesis. In all synthetic examples shown above, inversion was carried out with conjugated gradient in Scilab software.

A spatially irregular $f-k$ transform application in the regularization of Real 3D Streamer Data

The application of this heuristic choice of λ_i's in the regularization of real seismic data brings about some issues not discussed in details in this paper. Particularly, the 3D nature and noise are important aspects that are addressed in the regularization of real 3D shot records. Here we show a small set of traces from a real streamer marine 3D shot. For comparison, we selected traces that have receivers along a 3D seismic line. Before 3D regularization (figure 8), due to feathering, the set of traces appear discontinuous. After 3D regularization (figure 9), there is a trace for each common

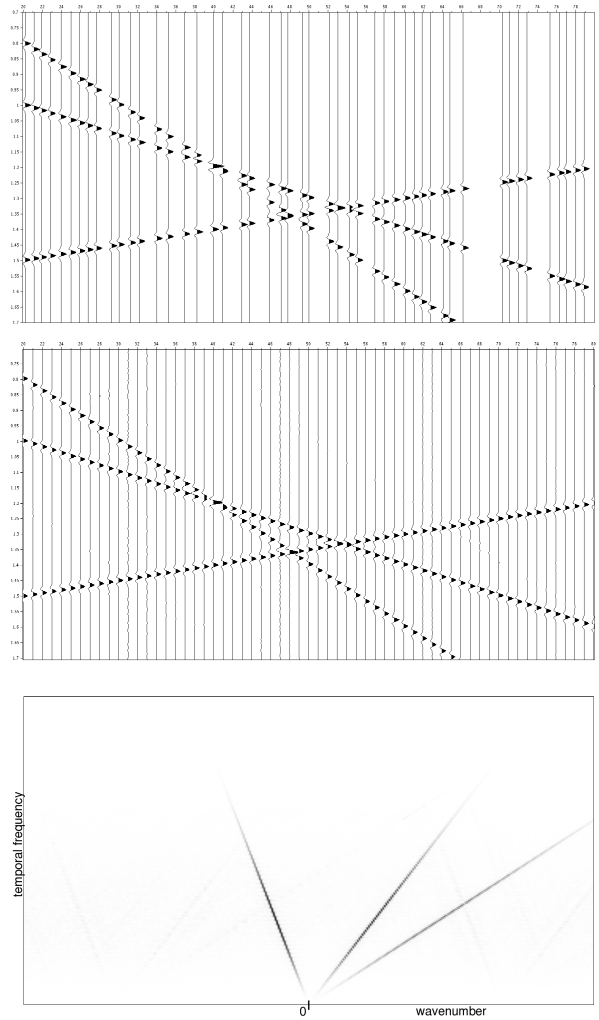

Figure 7: *A zoom at an irregularly sampled seismic section. Above, before Fourier interpolation; Middle, after interpolation; Below, the corresponding $f-k$ spectrum. Note that periodicity/aliasing is almost absent and the spectrum is quite cleaner than that shown in figure 6.*

mid point and the discontinuity disappeared.

Apart from the different amount of traces, it can be observed that random and swell-like noise have been attenuated as a result of the Fourier limitations from the original to the regularized data. It can also be observed that amplitudes were reconstructed with a great degree of confidence.

Finally, marine shot records are free of high spatial frequency discontinuities generated by variations in the velocity of sound in water and tidal effects. Also azimuth is a smooth function of receiver positions. Thus, in marine streamer surveys, the shot domain seems to be appropriate for Fourier interpolation.

Summary, Comments and Conclusions

We have analyzed the role of varying dumping factors for stabilization and definition of physically acceptable solutions in linear inverse problems. A heuristic approach, independent of statistical properties of the data and the physical

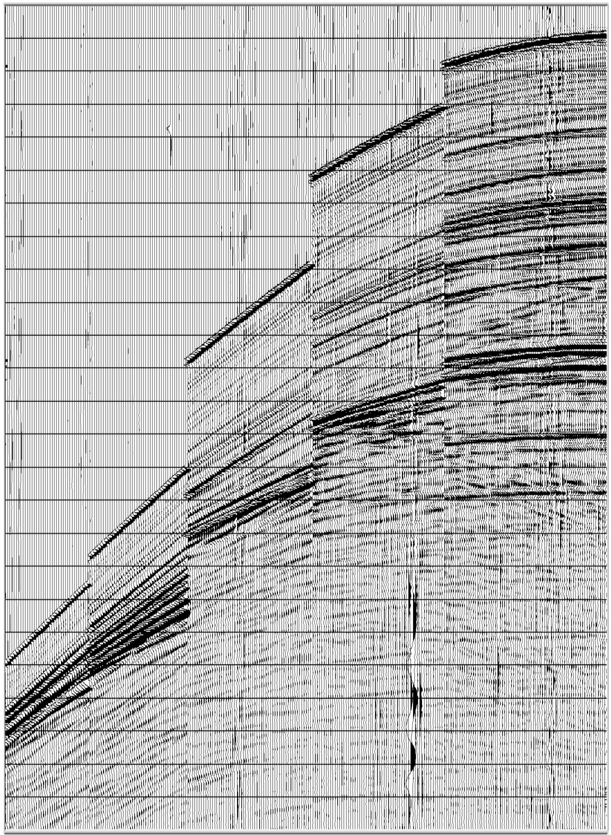

Figure 8: *Original shot. Traces colected along a 3D line. Discontinuities come from feathering.*

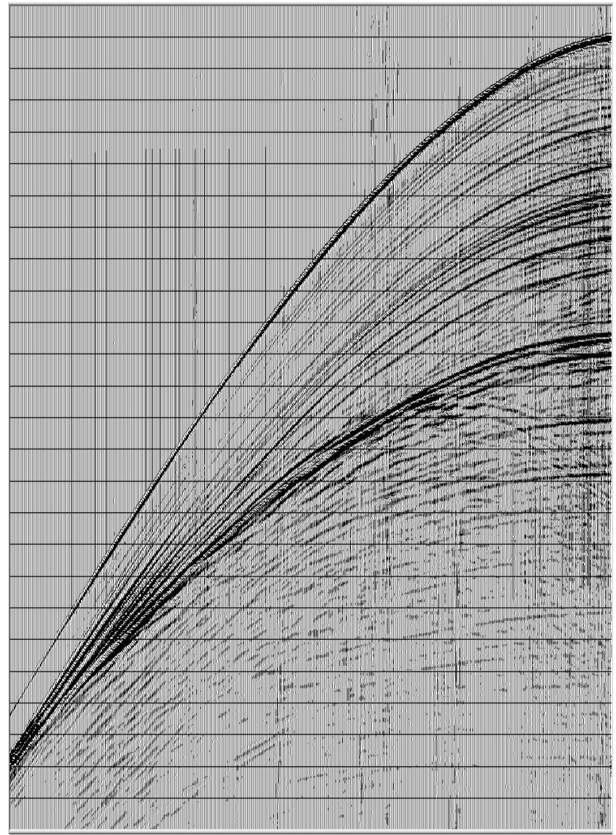

Figure 9: *Regularized shot. Traces colected along a 3D line. All mid points have a trace. The feathering in no longer visible.*

model, was proposed and applied with relative advantage in what fidelity to the model and convergence are a concern. The a priori knowledge of redundancy on specific applications is shown to be of direct use on the definition of an appropriate heuristic approach. Applications in seismic regularization with the $f-k$ transform was used.

This paper did not considered random noise because of limitations on paper's number of pages but, with some considerations on the heuristics for defining a "degree of acceptance", acceptable signal-to-noise ratios may be treated as well. The real 3D data shown somehow filled this gap in the content of the paper.

References

Sacchi, M. D., 1997, Reweighting strategies in seismic deconvolution: Geophys. J. Int., 129,651-656.

Sacchi, M.D., Ulrych, T.J, and Walker, C., 1998, Interpolation and extrapolation using a high resolution discrete Fourier transform: IEEE Trans. on Signal Processing, 46, No. 1, 31-38.

Zwartjes, P. M., and Duijndam, A. J. W., 2000, Optimizing Reconstruction for Sparse Spatial Sampling: 70th Ann. Internat. Mtg., Soc. Expl. Geophys., 2162–2165.

Zwartjes, P. M., and Sacchi, M. D.; 2004, Fourier reconstruction of non-uniformly sampled, aliased data: SEG, Expanded Abstracts, 23 , no. 1, 1997-2000.

Hennenfent, G. and Herrmann, F. J.; 2007, Random sampling: new insights into the reconstruction of coarsely-sampled wavefields. In the proceedings of the Society of Exploration Geophysicists International Exposition and Annual Meeting (SEG).

Acknowledgments

We would like to thank Paulo M. de Carvalho for reviewing and commenting this paper. We also would like to thank PETROBRAS for permission to publish this work.