Mitigating data gaps in the estimation of primaries by sparse inversion without data reconstruction

Tim T.Y. Lin and Felix J. Herrmann
Seismic Laboratory for Imaging and Modeling (SLIM), University of British Columbia

SUMMARY

We propose to solve the Estimation of Primaries by Sparse Inversion problem from a sesimic record with missing near-offsets and large holes without any explicit data reconstruction, by instead simulating the missing multiple contributions with terms involving auto-convolutions of the primary wavefield. Exclusion of the unknown data as an inversion variable from the REPSI process is desireable, since it eliminates a significant source of local minima that arises from attempting to invert for the unobserved traces using primary and multiple models that may be far-away from the true solution. In this talk we investigate the necessary modifications to the Robust EPSI algorithm to account for the resulting non-linear modeling operator, and demonstrate that just a few auto-convolution terms are enough to satisfactorily mitigate the effects of data gaps during the inversion process.

INTRODUCTION

Multiple removal is a crucial aspect of seismic signal processing that constantly face a difficult quad-lemma between accuracy, robustness, low computational complexity, and full-azimuthal sampling. Current prediction-subtraction methods such as Surface-Related Multiple Removal (Verschuur, 1992) face limits in accuracy and robustness when confronted with undersampled data of limited quality, prompting recent developments in whole- wavefield inversion/deconvolution techniques to decrease dependence on practitioner guesswork and QC. In recent work Lin and Herrmann (2014) proposed a multiscale strategy that reduces the computational burden by using coarser spatial sampling grids while exploiting the unique way in which REPSI (Lin and Herrmann, 2013) mitigates spatial aliasing. While this approach successfully addressed some of the computational costs associated with "data-driven" techniques typified by the Raleigh-Helmholtz reciprocity relationship between the field-measured wavefield data and its multiple-free version (Fokkema and van den Berg, 1993; Frijlink et al., 2011), these methods all rely on dense wide-azimuthal samplings that include near offset information.

This reliance on dense samplings has and continues to be challenging and has called for intricate on-the-fly trace interpolations as part of SRME predictions or extensions of EPSI to include missing data as unknowns (van Groenestijn and Verschuur, 2009), forcing the algorithm to alternate between estimating the source, the surface-free Green's function and this missing data. While the initial results of the latter technique in EPSI has been successful, these explicit inversion schemes do not exploit possibilities to extend the multiple-prediction operator to include recursive terms that model the imprint on missing data. Our main contribution in this work is to come up with a formulation where the effects of missing data–i.e., a mask acting on the data matrix zeroing entries where data is missing, are incorporated in the forward model of EPSI explicitly through auto-convolution terms.

Our work is organized as follows. First, we briefly summarize the EPSI formulation as an alternating optimization problem inverting for the source and surface-free Green's by promoting sparsity via the ℓ_1-norm on the latter. We make the dependence of EPSI on the fully sampled upward going wavefield explicit to emphasize the dependence of the formulation on dense sampling. Next, we discuss the method proposed by van Groenestijn and Verschuur (2009), which extends EPSI to the situation where information on the upgoing wavefield is missing, followed by our method where we incorporate convolutional terms in the forward modeling operator that account for missing (e.g., near offsets) traces in the upgoing wavefield. The different terms in these expansion predict the impact of missing data on the prediction of first and higher multiples. We conclude by demonstrating the efficacy of this method on both synthetic and field data sets.

THEORY

The basic assumption of surface-related multiple removal with EPSI is that, with noiseless and perfectly sampled up-going seismic wavefield $\mathbf{P}$ at the Earth's surface (with seismic reflectivity operator $\mathbf{R}$, typically close to $-\mathbf{I}$) due to a finite energy source wavefield $\mathbf{Q}$, there exists an operator $\mathbf{G}$ for every frequency in the seismic bandwidth such that the relation $\mathbf{P} = \mathbf{GQ} + \mathbf{RGP}$ holds true. Here we use the "detail-hiding" notation (Berkhout and Pao, 1982) where all upper-case bold quantities are monochromatic data-matrices, with the row index corresponding to the discretized receiver positions and column index the source positions. Moreover, $\mathbf{G}$ is interpreted as the (surface) multiple-free subsurface Green's function. The term $\mathbf{GQ}$ is interpreted as the primary wavefield, while the term $\mathbf{RGP}$ contains all surface-related multiples.

Since only $\mathbf{P}$ can be measured, inverting the above relation for $\mathbf{GQ}$ admits non-unique solutions without additional regularization. Based on the argument that a discretized physical representation of $\mathbf{G}$ resemble a wavefield with impulsive wavefronts, the Robust EPSI algorithm (REPSI, Lin and Herrmann, 2013) attempts to find the *sparsest* possible $\mathbf{g}$ in the physical domain (from here on, all lower-case symbols indicate discretized physical representations of previously defined quantities). Specifically, it solves the following optimization problem in the space-time domain:

$$\min_{\mathbf{g},\,\mathbf{q}} \|\mathbf{g}\|_1 \quad \text{subject to} \quad f(\mathbf{g},\mathbf{q};\mathbf{p}) \leq \sigma, \tag{1}$$
$$f(\mathbf{g},\mathbf{q};\mathbf{p}) := \|\mathbf{p} - M(\mathbf{g},\mathbf{q};\mathbf{p})\|_2,$$

where the forward-modeling function M can be written in terms of the data-matrix notation $M(\mathbf{G},\mathbf{Q};\mathbf{P}) := \mathbf{GQ} + \mathbf{RGP}$. Prob-

lem (1) essentially asks for the sparsest (via minimizing the ℓ_1-norm) multiple-free impulse response that explains the surface multiples in $\mathbf{p}$, while ignoring some amount of noise as determined by σ.

To consider the parts of data $\mathbf{P}$ that are not sampled, we now introduce a masking matrix $\mathbf{K}$ that is has same dimensions as the data-matrices. The elements of mask $\mathbf{K}$ has value 0 where we do not have data at the corresponding source-receiver position pair, and the value 1 at sampled positions. Thus, we can segregate parts of the data that are sampled as $\mathbf{P}' := \mathbf{K} \circ \mathbf{P}$, where the symbol $\circ$ denotes the matrix Hadamard product. Similarily, we introduce the complement mask $\mathbf{K_c}$ (a "stencil"), such that the unknown parts of the data can be written as $\mathbf{P}'' := \mathbf{K_c} \circ \mathbf{P}$ with $\mathbf{P}' + \mathbf{P}'' = \mathbf{P}$. The presence of these gaps are a significant source of error in the calculation of M (see, for example, Verschuur, 2006).

Explicit data inversion

As first proposed by van Groenestijn and Verschuur (2009), we can augment the optimization problem (1) to explicitly reconstruct $\mathbf{p}''$ from intermediate estimates of $\mathbf{g}$ such that the accuracy of M can be improved as the algorithm converges. We now overload the definition of M with a more general modeling operator

$$M(\mathbf{G}, \mathbf{Q}, \mathbf{P}''; \mathbf{P}') := \mathbf{GQ} + \mathbf{RG}(\mathbf{P}' + \mathbf{P}''), \tag{2}$$

which defines a more complicated inversion problem that has $\mathbf{p}''$ as an inversion variable

$$\min_{\mathbf{g},\,\mathbf{q},\,\mathbf{p}''} \|\mathbf{g}\|_1 \quad \text{subject to} \quad f(\mathbf{g}, \mathbf{q}, \mathbf{p}''; \mathbf{p}') \leq \sigma, \tag{3}$$
$$f(\mathbf{g}, \mathbf{q}, \mathbf{p}''; \mathbf{p}') := \|\mathbf{p}' + \mathbf{p}'' - M(\mathbf{g}, \mathbf{q}, \mathbf{p}''; \mathbf{p}')\|_2.$$

A major pitfall in solving (3) via an alternating optimization strategy (or cyclic coordinate descent) is that $\mathbf{g}$ and $\mathbf{p}''$ are in fact tightly coupled, because we can write the cyclic relation

$$\mathbf{P}'' = \mathbf{K_c} \circ \mathbf{P} = \mathbf{K_c} \circ \left[\mathbf{GQ} + \mathbf{RGP}' + \mathbf{RGP}''\right] \tag{4}$$

where it is evident that $\mathbf{P}''$ can be almost completely characterized by $\mathbf{G}$. We therefore would expect that $\partial \mathbf{p}''/\partial \mathbf{g}$ and $\partial \mathbf{g}/\partial \mathbf{p}''$ cannot be ignored when updating $\mathbf{p}''$ and $\mathbf{g}$. However, while $\partial \mathbf{p}''/\partial \mathbf{g}$ is straightforward to compute, the term $\partial \mathbf{g}/\partial \mathbf{p}''$ is convoluted and necessarily involves $\mathbf{Q}^{-1}$ (deconvolving the source function $\mathbf{q}$). Motivated by this observation, we propose to remove $\mathbf{p}''$ as an inversion variable all-together, and instead model its multiple contribution with terms that ultimately involve auto-convolutions of $\mathbf{g}$.

Alternative: inclusion of auto-convolution terms

Substituting expression (4) recursively into (2) results in a new forward-modeling operator into the range of observed data locations with infinitely many terms in a series expansion

$$\begin{aligned}
\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}') =& \mathbf{K} \circ \big[\mathbf{GQ} + \mathbf{RGP}' \\
&+ \mathbf{RGK_c} \circ (\mathbf{GQ} + \mathbf{RGP}') && (5)\\
&+ \mathbf{RGK_c} \circ \big(\mathbf{RGK_c} \circ (\mathbf{GQ} + \mathbf{RGP}')\big) && (6)\\
&+ \mathscr{O}(\mathbf{G}^4)\big] \\
:=& \mathbf{K} \circ \sum_{n=0}^{\infty} (\mathbf{RGK_c}\circ)^n \left(\mathbf{GQ} + \mathbf{RGP}'\right). && (7)
\end{aligned}$$

In this expression we slightly abuse the notation and write matrix Hadamard products as linear operators (valid as long as the corresponding rules of associativity are obeyed). Since high-order multiple decay, we know from physical arguments that $\|\mathbf{G}\| < 1$, and therefore (7) converges to $\mathbf{K} \circ (\mathbf{I} - \mathbf{RGK_c}\circ)^{-1} (\mathbf{GQ} + \mathbf{RGP}')$. In a noise-free setting with perfect data sampling outside of the holes, we expect $\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}')$ to exactly match the observed data $\mathbf{P}' = \mathbf{K} \circ \mathbf{P}$. Thus we have the following relation:

$$\mathbf{K} \circ \mathbf{P} = \mathbf{K} \circ (\mathbf{I} - \mathbf{RGK_c}\circ)^{-1} \left(\mathbf{GQ} + \mathbf{RGP}'\right). \tag{8}$$

The physical interpretation of expression (8) is clear: if we have access to the total data $\mathbf{P}$, we can exactly derive the multiple contribution due to $\mathbf{P}''$ by $\mathbf{RGK_c} \circ \mathbf{P} = \mathbf{RGP}''$. Since $(\mathbf{K}\circ)^{-1}$ and $(\mathbf{K_c}\circ)^{-1}$ does not exist, expression (8) cannot be directly turned into a practical inversion problem. However, it does serve to validate our new forward model $\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}')$.

To invert this operator, we propose to invert an approximate $\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}')$ that only includes the first few terms of (7). Figure 1 demonstrates a justification of this approach with shot-gather representations of the different terms in (7) for a synthetic dataset with missing near-offset traces (up to 75 m). Comparing panels **(d)** and **(g)**, it is evident that just the first three terms of (7) is enough to model most of the significant multiple contributions from the missing data $\mathbf{p}''$.

Algorithms

With the new forward-modeling operator $\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}')$ we can again redefine a new optimization problem based on (3):

$$\min_{\mathbf{g},\,\mathbf{q}} \|\mathbf{g}\|_1 \quad \text{subject to} \quad f(\mathbf{g}, \mathbf{q}; \mathbf{p}') \leq \sigma, \tag{9}$$
$$f(\mathbf{g}, \mathbf{q}; \mathbf{p}') := \|\mathbf{p}' - \widetilde{M}(\mathbf{g}, \mathbf{q}; \mathbf{p}')\|_2.$$

Compared to the EPSI problem with fully-sampled data (1), the new problem (9) is no longer a linear misfit problem in terms of $\mathbf{g}$. In Lin and Herrmann (2013), we relied on the bilinear property of $f(\mathbf{g}, \mathbf{q}; \mathbf{p}')$ in terms of $\mathbf{g}$ and $\mathbf{q}$ in order to solve problem (1). Below we will discuss two possible approaches to modify our solution strategy to account for the additional non-linear auto-convolution terms in $\widetilde{M}(\mathbf{G}, \mathbf{Q}; \mathbf{P}')$.

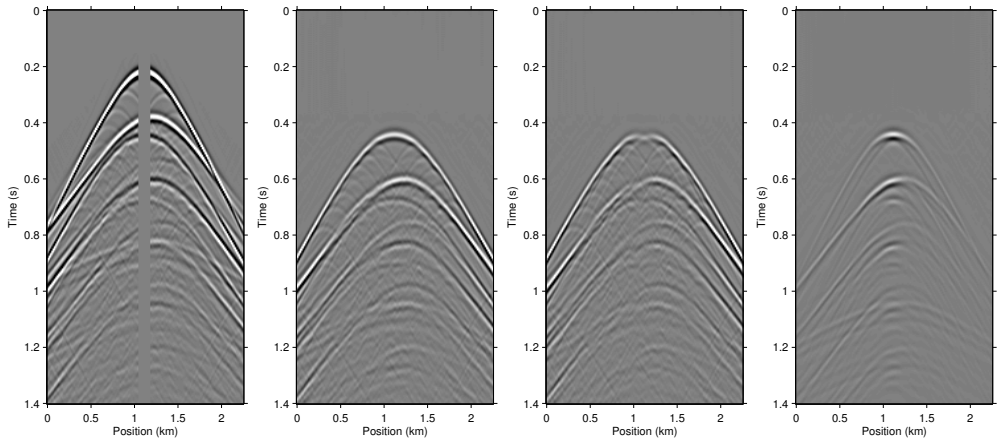
Modified Gauss-Newton

Several existing works on regularized inversion of auto-convolution functions rely on either the Gauss-Newton method or more generally the Levenberg-Marquardt method (Fleischer et al., 1999; Fleischer and Hofmann, 1996). In the same vein, we can adopt a modified Gauss-Newton method introduced in Li et al. (2012) to heuristically obtain sparse solutions to (9).

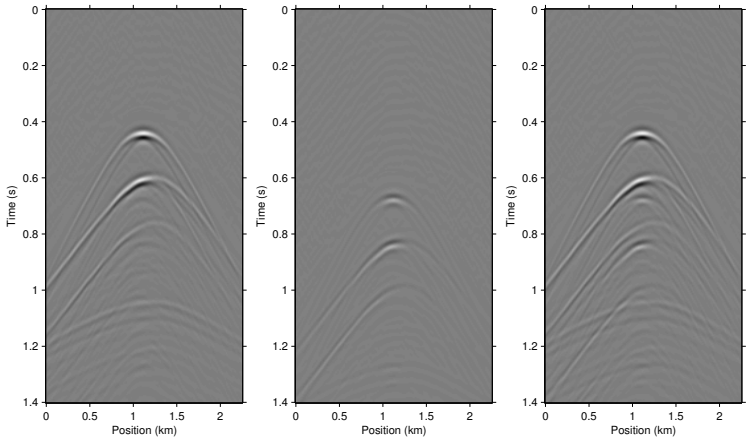
The crux of this approach is to always ensure that the model updates $\Delta\mathbf{g}$ are the sparsest possible for any given amount of decrease in $f(\mathbf{g}, \mathbf{q}; \mathbf{p}')$. This is achieved most effectively by taking as updates the solution to a Lasso problem (Tibshirani, 1996) around the current Jacobian:

$$\Delta\mathbf{g}_{k+1} = \underset{\mathbf{g}}{\operatorname{argmin}} \nabla\widetilde{M}_{g_k}\mathbf{g} \quad \text{s.t.} \quad \|\mathbf{g}\|_1 \leq \tau_k, \tag{10}$$

where k is the Gauss-Newton iteration count, and τ_k is an iteration-dependent ℓ_1-norm constraint determined in closed-form from the Pareto curve associated with (10) (c.f. line 5 of



(a) data $\mathbf{P}'$ (b) exact multiple $\mathbf{GP}$ (c) inexact multiple $\mathbf{GP}'$ (d) multiple contribution from $\mathbf{P}''$ (difference of b and c)



(e) expression (5) (f) expression (6) (g) sum of (e) and (f)

Figure 1: Shot-gathers of various multiple contribution terms in the auto-convolution based forward modeling operator $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$ (shown in expression 7) when applied to a synthetic dataset with missing near-offsets in $\mathbf{p}''$, and the correct values for $\mathbf{g}$ and $\mathbf{q}$. Panels **(e)** and **(f)** are respectively the first two terms of $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$ involving auto-convolutions with $\mathbf{g}$ (expressions 5 and 6). Comparing panels **(d)** and **(g)**, it is evident that just the first three terms of (7) is enough to model most of the significant multiple contributions from the missing data $\mathbf{p}''$.

Algorithm 1 in Li et al., 2012). Although this method does not claim to solve (9) exactly, it has been demonstrated to work well at giving sparse solutions to non-linear problems such as full-waveform inversion.

Re-linearization

Another heuristical method is to adapt the same approach used in Lin and Herrmann (2013) for the fully-sampled EPSI problem by simply re-linearizing the forward modeling operator at each iteration of the REPSI algorithm (c.f. line 9 of Algorithm 1 in Lin and Herrmann, 2013) by

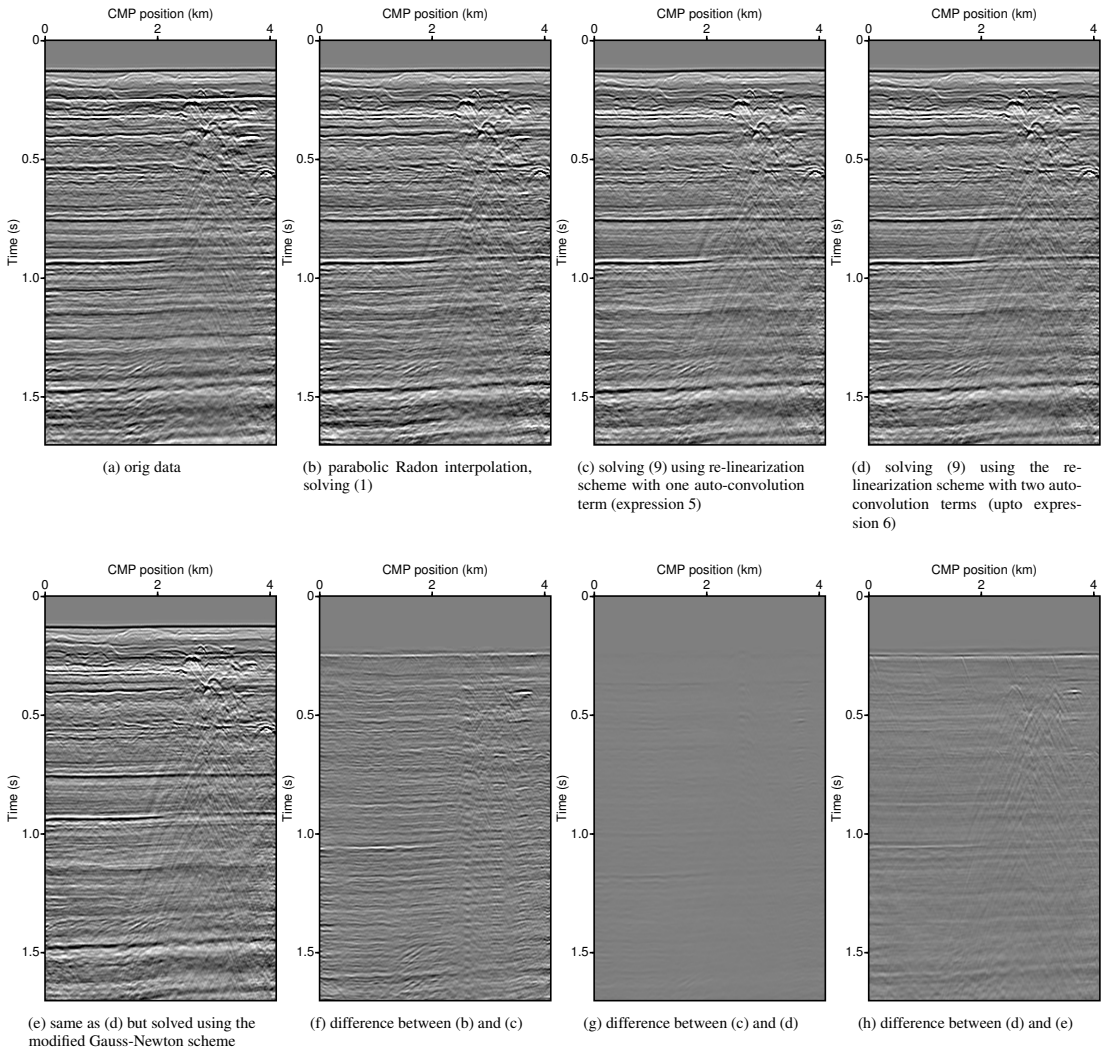
$$\begin{aligned}\widetilde{M}_{\mathbf{G}_k}(\mathbf{G}) =& \mathbf{K}\circ\big[\mathbf{G}\mathbf{Q}_k+\mathbf{R}\mathbf{G}\mathbf{P}' \\ &+\mathbf{R}\mathbf{G}_k\mathbf{K}_{\mathbf{c}}\circ(\mathbf{G}\mathbf{Q}_k+\mathbf{R}\mathbf{G}\mathbf{P}')+\ ...\big].\end{aligned}$$

Compared to the modified Gauss-Newton, this approach avoids computing the action of the Jacobian of $f(\mathbf{g},\mathbf{q};\mathbf{p}')$, which saves the computational cost of one wavefield convolution per term used in $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$. As we see below in numerical experiments on real data, the two approaches gives comparable performance, and both approaches outperform pre-EPSI interpolation by parabolic Radon.

EXPERIMENTS

We now demonstrate the effectiveness of the auto-convolution based problem using a shallow-water marine dataset with 100 m of missing near-offset data. The data has been pre-processed via up-down decomposition to be the upgoing wavefield, and a 3D-to-2D correction of $\sqrt{t}$ has been applied to the data.

Figure 2 show the REPSI results on this field dataset with missing near-offsets using various methods. Figure 2b shows REPSI results by pre-interpolation using parabolic Radon, which is known to under-estimate near-offset amplitudes, while Fig-



(a) orig data

(b) parabolic Radon interpolation, solving (1)

(c) solving (9) using re-linearization scheme with one auto-convolution term (expression 5)

(d) solving (9) using the re-linearization scheme with two auto-convolution terms (upto expression 6)

(e) same as (d) but solved using the modified Gauss-Newton scheme

(f) difference between (b) and (c)

(g) difference between (c) and (d)

(h) difference between (d) and (e)

Figure 2: NMO stacks of various REPSI on real marine streamer data. Compared to parabolic Radon interpolation, our approach succeeds in removing more coherent multiple energy from the data. We also see relatively minor differences between the modified Gauss-Newton and the re-linearization schemes to solve (9), and virtually no difference by using more auto-convolution terms in the modified modeling operator $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$. See the text for more detail.

ures 2c and 2d show results obtained by solving $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$ using the re-linearization strategy. Looking at the difference plot 2f, our approach is much more successful at removing coherent multiple energy. Figure 2g shows that very little additional accuracy is gained by incorporating successively higher-order auto-convolution terms in $\widetilde{M}(\mathbf{G},\mathbf{Q};\mathbf{P}')$. Finally, 2e and 2h shows a modified Gauss-Newton solution to the problem. We see relatively minor differences that is mostly phase errors.

SUMMARY

We have presented a variation of the EPSI problem that accounts for data gaps by implicitly account for it in the inversion model without explicit data reconstruction steps. We proposed two methods to solve this new non-linear problem, and have demonstrated its efficacy on real marine streamer data. We note that this scheme does not wholly account for *under*sampling issues in the data, which causes more severe aliasing problems in the multiple step which cannot be wholly mitigated reliably using the approaches shown here.

Acknowledgements

We would like to acknowledge Gert-Jan van Groenestijn and Eric Verschuur for discussion with led to our involvement in EPSI. Thanks to PGS for permission to use the real dataset. This work was financially supported in part by the NSERC of Canada Discovery Grant (RGPIN 261641-06) and the CRD Grant DNOISE II (CDRP J 375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, ConocoPhillips, CGG, ION GXT, Petrobras, PGS, Statoil, Total SA, WesternGeco, and Woodside.

REFERENCES

Berkhout, A. J., and Y. H. Pao, 1982, Seismic Migration - Imaging of Acoustic Energy by Wave Field Extrapolation: Journal of Applied Mechanics, **49**, 682.

Fleischer, G., R. Gorenflo, and B. Hofmann, 1999, On the Autoconvolution Equation and Total Variation Constraints: ZAMM, **79**, 149–159.

Fleischer, G., and B. Hofmann, 1996, On inversion rates for the autoconvolution equation: Inverse Problems, **12**, 419–435.

Fokkema, J. T., and P. M. van den Berg, 1993, Seismic applications of acoustic reciprocity: Elsevier Science.

Frijlink, M. O., R. G. van Borselen, and W. Söllner, 2011, The free surface assumption for marine data-driven demultiple methods: Geophysical Prospecting, **59**, 269–278.

Li, X., A. Y. Aravkin, T. van Leeuwen, and F. J. Herrmann, 2012, Fast randomized full-waveform inversion with compressive sensing: Geophysics, **77**, A13–A17.

Lin, T. T. Y., and F. J. Herrmann, 2013, Robust estimation of primaries by sparse inversion via one-norm minimization: Geophysics, **78**, R133–R150.

———, 2014, Multilevel acceleration strategy for the robust estimation of primaries by sparse inversion: Presented at the 76th EAGE Conference & Exhibition.

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso: Journal of the Royal Statistical Society Series B Methodological, **58**, 267–288.

van Groenestijn, G. J. A., and D. J. Verschuur, 2009, Estimating primaries by sparse inversion and application to near-offset data reconstruction: Geophysics, **74**, A23–A28.

Verschuur, D. J., 1992, Adaptive surface-related multiple elimination: Geophysics, **57**, 1166.

———, 2006, Seismic multiple removal techniques: Past, present and future: EAGE Publications.