Robust full-waveform inversion using the Student's t-distribution

Aleksandr Aravkin[1], *Tristan van Leeuwen*[1] *and Felix Herrmann*[1]

[1] *Dept. of Earth and Ocean sciences University of British Columbia Vancouver, BC, Canada*

SUMMARY

Full-waveform inversion (FWI) is a computational procedure to extract medium parameters from seismic data. Robust methods for FWI are needed to overcome sensitivity to noise and in cases where modeling is particularly poor or far from the real data generating process. We survey previous robust methods from a statistical perspective, and use this perspective to derive a new robust method by assuming the random errors in our model arise from the Student's t-distribution. We show that in contrast to previous robust methods, the new method progressively down-weighs large outliers, effectively ignoring them once they are large enough. This suggests that the new method is more robust and suitable for situations with very poor data quality or modeling. Experiments show that the new method recovers as well or better than previous robust methods, and can recover models with quality comparable to standard methods on noise-free data when some of the data is completely corrupted, and even when a marine acquisition mask is entirely ignored in the modeling. The ability to ignore a marine acquisition mask via robust FWI methods offers an opportunity for stochastic optimization methods in marine acquisition.

INTRODUCTION

Full-waveform inversion (FWI) is a data-fitting procedure based on full wavefield modeling designed to extract medium parameters (velocity and density) from seismograms. FWI is often formulated as a nonlinear least squares optimization problem (Virieux and Operto (2009)). Specifically, the parameters m are recovered by solving the following problem:

$$\min_m \quad \phi(m) := \|D - PH[m]^{-1}Q\|_F^2 \ , \qquad (1)$$

where m is the vector of velocity parameters or, more typically, 'slowness squared' parameters on a 2D or 3D grid, $\|\cdot\|_F^2$ is the Frobenius norm, $D \in \mathbf{C}^{k\times l}$ contains time harmonic data from l source experiments (as k-dimensional columns), $H[m]$ is a discretization of the Helmholtz operator with boundary conditions, $Q \in \mathbf{C}^{p\times l}$ specifies l source experiments, $H^{-1}[m]Q$ describes the solution of the Helmholtz equation for the sources Q, and P is a restriction of this solution to the surface where the data was observed.

It is important to consider the statistical interpretation of this classical approach. The statistical model corresponding to (1) is

$$D = PH[m]^{-1}Q + \varepsilon\ , \quad \varepsilon \sim (0, I)\ . \qquad (2)$$

In the standard approach, FWI finds the maximum *a posteriori* (MAP) solution for the statistical model (2) where all errors are assumed to be independent identically distributed (i.i.d.) Gaussians. To see this, note that the Gaussian likelihood $\mathbf{p}(D,m)$ of observing the data in model (2) satisfies

$$\mathbf{p}(D,m) \propto \exp\left(-\frac{1}{2}\sum_i (D - PH[m]^{-1}Q)_i^2\right)\ ,$$

where i runs over all the entries in the data structure D. Once D has been observed, the MAP value of m can be found by minimizing $-\log(\mathbf{p}(m|D))$, given by (1).

Gaussian models (and correspondingly, least squares formulations) are known to be sensitive to large errors or artifacts in the data. 'Artifacts' refer to coherent or systematic (non-random) events that arise because of shortcomings in the forward modeling. The distinction between 'errors' and 'artifacts' can become blurred—for example, multiples in the data, which motivated robust methods in tomography (Bube and Langan (1997)), can actually aid recovery once modeling tools are available to appropriately use them in the inversion process. Other examples of such systematic events include out of plane scattering when working with 2D seismic data, and using acoustic PDEs for the forward model where the elastic or anisotropic models could be applied. Finally, in marine acquisition, accounting for the mask precludes stochastic optimization techniques, and so we may want to ignore it. Overall, it is of great value to develop inversion methods that still recover well when modeling assumptions are violated by real-world data, since some degree of violation is unavoidable.

Past approaches have addressed the issue of robustness to errors in the data by using a variety of misfit functionals that pay less attention to large outliers than formulation (1). Bube and Langan (1997) consider a hybrid misfit applied to the residuals ε_i in (2):

$$H_\sigma(\varepsilon_i) = \sqrt{1 + (r_i/\sigma)^2} - 1\ . \qquad (3)$$

This function is strictly convex, yielding unique solutions for full rank linear systems, and was minimized with the IRLS algorithm. In a later paper (Bube and Nemeth (2007)), the LBFGS and NLCG methods are proposed for this same approach for overdetermined linear systems, and exact line search techniques are derived to speed up convergence of these methods. Guitton and Symes (2003) exploit the Huber norm applied to ε in (2). A (somewhat modified) Huber model is given below:

$$H_k(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2 & |\varepsilon| \le k \\ k|\varepsilon| - \frac{1}{2}k^2 & |\varepsilon| > k \end{cases}\ . \qquad (4)$$

Finally, Brossier et al. (2010) compare the standard approach (1) with the Huber and ℓ_1 misfits applied to ε in (2), again using LBFGS.

Each of the approaches described above can be interpreted in the statistical framework already established in (1) and (2). Specifically, a robust method that relies on a misfit $\rho(\cdot)$ and minimizes $\sum_i \rho(\varepsilon_i)$ across residuals in the forward model also recovers the MAP estimate of the parameters according to a statistical model with density for ε in (2) satisfying

$$\mathbf{p}(m|D) \propto \exp(-\sum_i \rho(\varepsilon_i))\ .$$

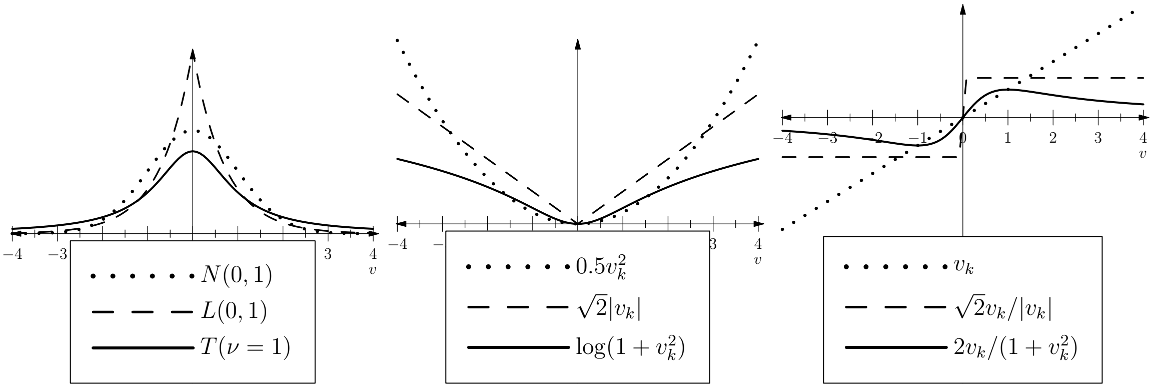


Figure 1: Gaussian, Exponential, and Student's t- Densities, Corresponding Negative Log Likelihoods, and Influence Functions (for scalar error v_k).

Previous approaches can be recovered by taking $\rho(\varepsilon) = \frac{1}{2}\varepsilon^2$ (standard), $\rho(\varepsilon) = H_\sigma(\varepsilon)$ (hybrid), $\rho(\varepsilon) = H_k(\varepsilon)$ (Huber), and $\rho(\varepsilon) = |\varepsilon|$ (ℓ_1). Then the recovery is obtained by solving

$$\min_m \quad \phi_\rho(m) := \sum_i \rho\left(D - PH[m]^{-1}Q\right)_i . \qquad (5)$$

Note that all of these objectives are convex, and the first two are strictly convex. Convexity is a desirable property for both the theory and practice of optimization. In particular, any local minimum of a convex function is a global minimum, and strict convexity guarantees uniqueness of global minima.

Unfortunately, as is well known for (1), the composition of even a strictly convex misfit with the nonlinear forward model difference $D - PH[m]^{-1}Q$ fails to be convex in (5), and this is reflected in the effort that has been dedicated to regularization of (1) and to the standard practice in FWI, where frequency sweeps are done from low to high frequencies (see Virieux and Operto (2009) and sources within, as well as Vogel and Oman (1996)).

All previous robust approaches are derivable from distributions whose tail probabilities are no thicker than that of the exponential, $\exp(-|x|)$. This is precisely because they preserve convexity of the resulting penalty. In this paper, we forego this aim, as the composite problems are no longer convex anyway. In particular, we model the distributions of ε in (2) as arising from a Student's t-distribution, whose tails go to zero as x^{-k}, and are thus much thicker than exponential tails. Figure 1 compares the Gaussian, Exponential, and Student's t-densities (left frame), along with the negative log likelihoods (center frame) and their derivatives, known as 'influence functions' in statistical literature (right frame). Please see (Huber (2004); Hampel et al. (1986)) for more details on robust statistical methods and the role of influence functions.

Note that the Student's t-negative log likelihood is not convex. As a result, the influence function (derivative of this log likelhood) is 're-descending', meaning that it goes to 0 as the size of an error term increases. In the new method, the impact of a residual is therefore down-weighed by its size, whereas in previous robust methods (ℓ_1, Huber, and hybrid) extremely large residuals still maintain a fixed influence on the recovered model . In (1), the influence grows with the size of the residual.

The paper is organized as follows. In the next section, we present the multivariate Student's t-density, and formulate a MAP optimization problem for the model parameters m that arises from assuming Student's t-errors ε in (2). We then make several observations about this problem and methods for its solution. Finally, we present several examples comparing the standard approach (1) with the Huber approach and the Student's t-approach in FWI in several scenarios that involve randomly generated outliers and systematic artifacts that arise from unexplained systematic events in the data.

STUDENT'S T-DENSITY AND FWI FORMULATION

For a vector $\varepsilon \in \mathbf{R}^n$ and any positive definite matrix $M \in \mathbf{R}^{n\times n}$, let $\|\varepsilon\|_M := \sqrt{\varepsilon^{\mathrm{T}} M \varepsilon}$. We use the following generalization of the Student's t-distribution

$$\mathbf{p}(\varepsilon|\mu, M) = \frac{\Gamma(\frac{s+l}{2})}{\Gamma(\frac{s}{2})\det[\pi s M]^{1/2}} \left(1 + \frac{1}{s}\|\varepsilon - \mu\|^2_{M^{-1}}\right)^{\frac{-(s+l)}{2}} , \qquad (6)$$

where μ is the mean parameter, s is the degrees of freedom, and l is the dimension of the vector ε. The distribution is symmetric, and hence has mean μ, and the reader can easily verify that it has covariance M.

For our purposes, the distribution is determined up to knowledge of the degrees of freedom parameter s. Having free parameters is a disadvantage shared by the hybrid and Huber approach, though not by the Laplace (ℓ_1) methodology or the standard approach (1). As $s \to \infty$, the Student's t-density approaches the Gaussian, as can be easily seen by observing (from basic calculus) that

$$\lim_{s\to\infty}\left(1 + \frac{1}{s}\|\varepsilon - \mu\|^2_{M^{-1}}\right)^{-(s+l)/2} = \exp\left(-\frac{1}{2}\|\varepsilon - \mu\|^2_{M^{-1}}\right) .$$

However, we would like the tails as thick as possible, so we typically take the degrees of freedom parameter s to be small, for example 2 or 3. The quality of our results are not influenced by which particular small s we pick, and we will focus on methods for estimation of s in future work.

If we assume ε_i in (2) are i.i.d, as in the standard approach, then the MAP optimization problem we must solve to recover the model m is given by

$$\min_m \; \phi_T(m) := \frac{s+l}{2}\sum_i \log\left(s + \left(D - PH[m]^{-1}Q\right)_i^2\right) \; , \quad (7)$$

where i runs through the data structure D, so in particular $l = 1$ since each data point is a scalar. Define

$$F[m;Q] = PH[m]^{-1}Q \; . \quad (8)$$

The objective in (7) is differentiable, with gradient given by

$$\nabla_m \phi_T(m) = \frac{s+l}{2}\sum_i \frac{\nabla_m F[m;Q]_i^T (F[m;Q]-D)_i}{s + (D - F[m;Q])_i^2} \; . \quad (9)$$

This makes it quite easy to implement the LBFGS method to solve it. Moreover, it is very interesting to compare the gradient (9) with the least squares gradient

$$\nabla_m \phi(m) = \sum_i \nabla_m F[m;Q]_i^T (F[m;Q]-D)_i \; . \quad (10)$$

The i-th contribution to the gradient (9) is the same as the i-th contribution to (10), but down-weighed by the magnitude of the i-th residual $s + (D-F)_i^2$. This is an explicit demonstration of the Student t-influence function, shown in the leftmost panel of figure 1. Notice that large residuals contribute very little to the search direction as the optimization proceeds as a result of this down-weighing. Functionally, such down-weighing strongly resembles the IRLS method (see for example Bube and Langan (1997)); however, in our context the down-weighing arises naturally, deriving from the gradient of the MAP optimization problem (7).

There is one more benefit to using the formulation (7). Note that the Hessian $\nabla_m^2 \phi_T(m)$ exists, and is given by

$$\frac{s+l}{2}\sum_i \frac{\nabla^2 F_i^T (F-D)_i + \nabla F_i^T \nabla F_i}{s+(D-F)_i^2} + \frac{\nabla F_i (\nabla F_i^T (F-D)_i}{(s+(D-F)_i^2)^2} \; , \quad (11)$$

where $F[m;Q]$ has been replaced by F for readability. Recall that a function g is Lipshitz continuous if its gradient can be controlled via $\|\nabla g(m') - \nabla g(m)\| \leq L\|m' - m\|$. One can see from the expression (11) that the Lipshitz continuity of the Hessian $\nabla_m^2 \phi_T(m)$ at the solution $\bar{m}$ is ensured by the Lipshitz continuity of $\nabla_m^2 F[m;Q]$ and $\nabla_m F[m;Q]$ at $\bar{m}$, since $\nabla_m^2 \phi_T(m)$ is a linear combination of such terms down-weighed by residual norms, where the weights are always greater than 1. This observation is important because if we have Lipshitz continuity of $\nabla_m^2 \phi_T(m)$, (Nocedal and Wright, 1999,Theorem 8.6) guarantees super-linear convergence of the BFGS method. This nice result is a consequence of the smoothness of the negative log likelihood function, and is not valid for the Huber or ℓ_1 approaches, though it can be easily shown to hold for the hybrid approach using (3). In practice, we of course still use the LBFGS method with a small number of iterations.

RESULTS

We test the proposed approach on a synthetic dataset. We generate data for the model depicted in figure 2 using a frequency-domain finite difference method. The dataset contains 151 shots with a 40 m spacing and 301 receivers with a 10 m spacing. We use an LBFGS method to invert the data in 4 separate frequency bands : [3,5,7], [9,11,13],[15,17,19], [21,23,25] (cf. Bunks et al., 1995). We do a fixed number of 20 iterations per frequency band. We consider two scenarios to illustrate the robustness of the Huber and Student's t-misfits. In the first scenario, we follow standard approaches to testing robust methods, and add large outliers to the dataset. In the second scenario, we ignore known information (mask from marine acquisition), to illustrate robustness against wrong modeling.

Outliers. We add background Gaussian noise to the data, by generating a standard Gaussian matrix of the same size as the data, and then normalizing it to 1% of the energy of the data. To simulate outliers, we corrupt 10% of the traces with Gaussian noise of the *same energy level* (100%) as the data. A slice of the data is depicted in figure 3 (a). This emulates poor preprocessing of the data, for example. The result for the LS, Huber and Student's t-misfit are depicted in figure 4. The standard method (LS) fails to recover anything useful, while both of the robust methods do equally well, with performance comparable to the standard method on noiseless data.

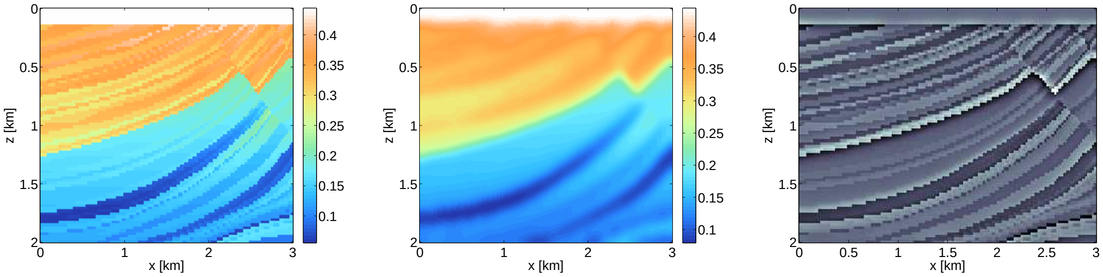
Missing data. Instead of using a full acquisition, we emulate a marine acquisition, where not all sources are sampled by the same receivers. The acquisition mask is depicted in figure 3 (b). In practice, one would take this into account in the inversion. However, we attempt to recover the model while *ignoring* the mask, and using robust methods. Note that the LS method breaks completely if the mask is ignored, but both of the robust methods still recover nicely, with the Student's t-method exhibiting qualitatively superior performance: in figure 5, the Huber method produces extraneous artifacts.

CONCLUSION

A robust method, derived from the MAP estimate of the Student's t-density is presented as an alternative to previous robust methods. In contrast to previous methods, the new method diminishes the influence of large outliers, which suggests that the new method may be a better choice when the quality of the data is extremely poor or models poorly explain the data. This conclusion is supported by simulated experiments, where a) very high proportions of outliers are added, and b) a marine acquisition mask is entirely ignored. This latter experiment also suggests that robust methods open a promising direction to design simultaneous source FWI methods (cf. Krebs et al., 2009; Haber et al., 2010; van Leeuwen et al., 2010) for marine acquisition. It is currently an open question how to use such methods for cases where we do not have full acquisition; specifically, the need to deal with the acquisition mask breaks these methods. Ignoring the mask, but using a robust method, paves the way for the design of fast stochastic methods for new and more realistic acquisition scenarios.

ACKNOWLEDGMENTS

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BP, Chevron, ConocoPhillips, Petrobras, Total SA, and WesternGeco.

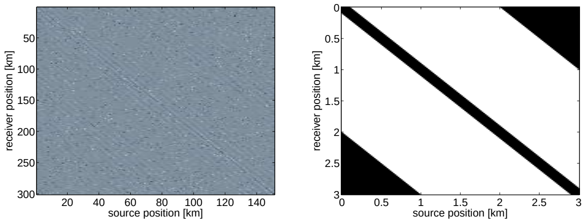


(a)

(b)

(c)

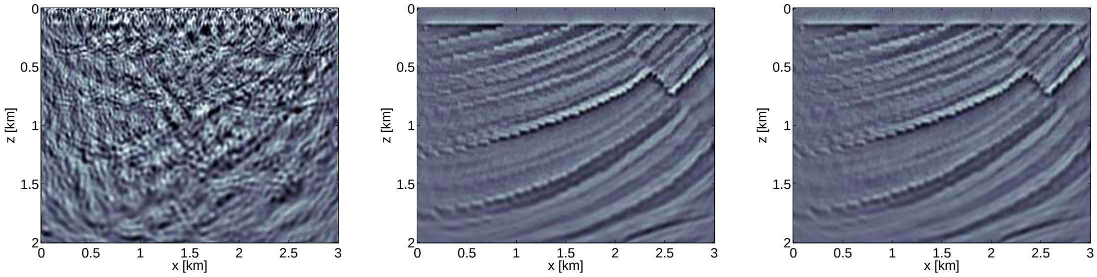
Figure 2: (a) true model $[s^2/km^2]$. (b) initial model $[s^2/km^2]$ and (c) true reflectivity.



(a)

(b)

Figure 3: (a) data slice at 15 Hz with spiky noise, (b) source-receiver mask.

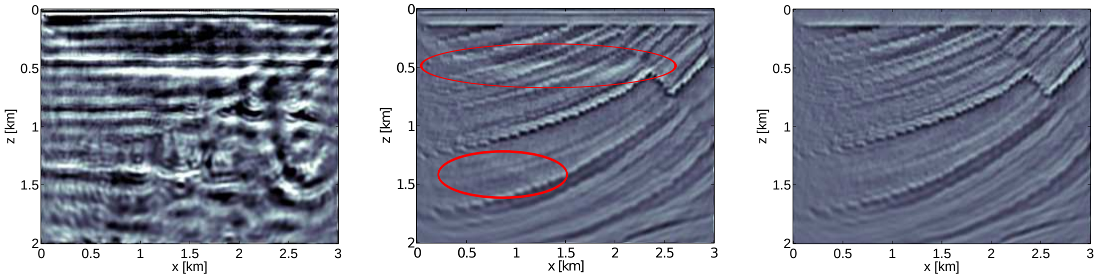


(a)

(b)

(c)

Figure 4: Inversion result (difference with initial model) with least-squares (a), Huber (b) and Students T (c) misfit for data with spiky noise.



(a)

(b)

(c)

Figure 5: Inversion result (difference with initial model) with least-squares (a), Huber (b) and Students T (c) misfit for incomplete data. Both robust approaches recover well. The Huber result shows some minor artifacts that are not present in the more robust Students T formulation.

REFERENCES

Brossier, R., S. Operto, and J. Virieux, 2010, Which data residual norm for robust elastic frequency-domain full waveform inversion?: Geophysics, **75**, R37–R46.

Bube, K. P., and R. T. Langan, 1997, Hybrid ℓ_1/ℓ_2 minimization with applications to tomography: Geophysics, **62**, 1183–1195.

Bube, K. P., and T. Nemeth, 2007, Fast line searches for the robust solution of linear systems in the hybrid ℓ_1/ℓ_2 and Huber norms: Geophysics, **72**, A13–A17.

Bunks, C., F. Saleck, S. Zaleski, and G. Chavent, 1995, Multiscale seismic waveform inversion: Geophysics, **60**, 1457–1473.

Guitton, A., and W. W. Symes, 2003, Robust inversion of seismic data using the Huber norm: Geophysics, **68**, 1310–1319.

Haber, E., M. Chung, and F. J. Herrmann, 2010, An effective method for parameter estimation with PDE constraints with multiple right hand sides: Technical Report TR-2010-4, UBC-Earth and Ocean Sciences Department.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, 1986, Robust statistics: John Wiley and Sons.

Huber, P. J., 2004, Robust statistics: John Wiley and Sons.

Krebs, J. R., J. E. Anderson, D. Hinkley, R. Neelamani, S. Lee, A. Baumstein, and M.-D. Lacasse, 2009, Fast full-wavefield seismic inversion using encoded sources: Geophysics, **74**, WCC177–WCC188.

Nocedal, J., and S. J. Wright, 1999, Numerical optimization: Springer. Springer Series in Operations Research.

van Leeuwen, T., A. Aravkin, , and F. J. Herrmann, 2010, Seismic waveform inversion by stochastic optimization: Technical Report TR-2010-5, UBC-Earth and Ocean Sciences Department.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, 127–+.

Vogel, C. R., and M. E. Oman, 1996, Iterative methods for total variation denoising: SIAM J. Sci. Comput., **17**, no. 1, 227–238.