Robust FWI using Student's T & Robust Source Estimation

Aleksandr Aravkin*
saravkin@eos.ubc.ca

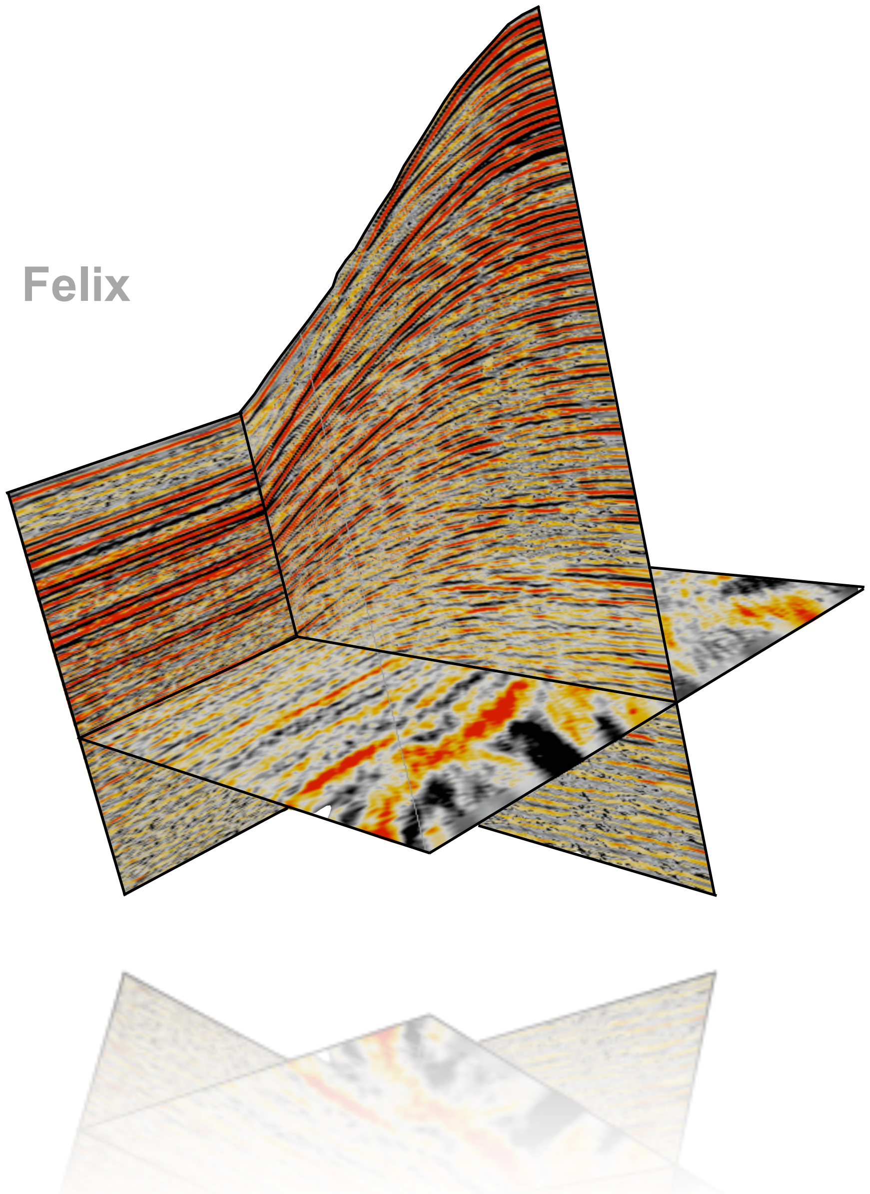
Joint work with

Tristan van Leeuwen*, Michael P. Friedlander*, Felix J. Herrmann*.

Implementations and tests at Total with

Henri Calandra, Bertrand Denel

***Seismic Laboratory for Imaging & Modeling**
Department of Earth & Ocean Sciences
The University of British Columbia



Sinbad Consortium
Whistler, 2011

Outline

- Robust FWI
  - Motivation and statistical insight
  - Robust FWI formulations, with focus on **Student's t**
  - Results on synthetic data (including implementation at Total)

- Robust Source Estimation
  - General formulation (includes Student's t, Huber, hybrid, etc formulations).
  - Generalized Variable Projection Approach
  - Specific implementations and examples
  - Proof of concept numerics

Nonlinear Least Squares Formulation

- We consider inverse problems of the form

$$\mathbf{D} = \mathcal{F}(\mathbf{m};\mathbf{Q}) + \boldsymbol{\epsilon}$$

$\mathbf{D}$	$n \times m$ matrix of observations
$\mathbf{Q}$	$l \times m$ array of source parameters
$\mathbf{m}$	parameters to be recovered
$\mathcal{F}(\mathbf{m};\mathbf{Q})$	Forward model (calculated data)
$\boldsymbol{\epsilon}$	Model for error, typically Gaussian i.i.d.

- Choice of Gaussian error leads to least squares formulation:

$$\min_{\mathbf{m}} \Phi(\mathbf{m}) = \| \underbrace{\mathbf{D} - \mathcal{F}(\mathbf{m};\mathbf{Q})}_{\mathbf{R}(\mathbf{m})} \|_F^2 = \sum_{i=1}^{m} \| \underbrace{\mathbf{d}_i - \mathcal{F}(\mathbf{m})\mathbf{q}_i}_{\mathbf{r_i}(\mathbf{m})} \|_2^2$$

Statistical Perspective for Least Squares

- The NLLS formulation is equivalent to the following statistical model:

$$\begin{aligned} \mathbf{D} &= \mathcal{F}[\mathbf{m};\mathbf{Q}] + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathbf{N}(0, I) \end{aligned}$$

- Equivalence follows from maximum likelihood estimate for model parameters:

$$\mathcal{L}(\mathbf{m}) \quad \propto \quad \exp\left(-\frac{1}{2}\left\|\mathbf{D} - \mathcal{F}[\mathbf{m};\mathbf{Q}]\right\|_F^2\right)$$

- Minimizing the negative log likelihood is exactly the FWI problem.

- Statistical perspective explains why least squares are sensitive to outliers and artifacts in the data!

Sensitivity to Outliers in Gaussian Regime

- Large deviations from the mean are VERY unlikely in the Gaussian model:

	Gaussian
$p(\|x\| > 4\sigma)$	6.3×10^{-5}
$p(\|x\| > 8\sigma)$	1.3×10^{-15}
$p(\|x\| > 8\sigma \mid \|x\| > 4\sigma)$	2.1×10^{-11}

- Observations more than 4 standard deviations away from the mean occur less than .006 percent of the time.

- Even when we KNOW we have an outlier 4 standard deviations away, we still believe it is impossible for the outlier to be more than 8 standard deviations away!

- Low likelihood values correspond to HIGH penalties for outliers.

Motivation for Robust Formulations

- Errors in measurement, e.g. equipment malfunction
- Missing data: measurement instruments may fail to record
- Even more important: unexplained “artifacts” in the data! A lot of effort is routinely devoted to
  - Data cleaning to remove unexplained artifacts
  - Complex forward model design to explain such artifacts e.g. acoustic vs. elastic vs. anisotropic
- Why not use robust fitting methods with cheaper modeling?

Modeling from the Statistical Perspective

- We can alter the assumptions on the model error:

$$\mathbf{D} \quad = \quad \mathcal{F}[\mathbf{m};\mathbf{Q}] + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \text{ has density } \mathbf{p}$$

	Gaussian	$L(\lambda = 1)$	$T(k = 3)$
$p(\lvert x\rvert > 4\sigma)$	6.3×10^{-5}	$\mathbf{0.02}$	$\mathbf{0.6 \times 10^{-2}}$
$p(\lvert x\rvert > 8\sigma)$	1.3×10^{-15}	3.3×10^{-4}	8.1×10^{-4}
$p(\lvert x\rvert > 8\sigma \mid \lvert x\rvert > 4\sigma)$	2.1×10^{-11}	$\mathbf{0.02}$	$\mathbf{0.14}$

- Laplace/Huber have heavier tails than the Gaussian...

- But Student's t-density (3rd column) is **heavy tailed**.

Some Previous Work

- Robust statistical work has a long history (I've seen references to 1930's). A few useful 'Robust statistics' books:
  - Huber 1981
  - Hampel et al (2003)
  - Marona et al, (2006)
- For robust penalties in Seismic, see
  - Huber: Guitton & Symes, 2003
  - Huber and L1: Brossier, Operto, Virieux 2009, 2010
  - Hybrid: Bube, 2007.
- We are particularly interested in Student's t distribution. See
  - Lange 1989, general paper applying student's t formulations to regression
  - Fahrmeir 1998, Robust kalman smoothing using Student's t
- In our experience, Student's t works well for structured inverse problems in nonlinear Kalman smoothing, computer vision applications, and FWI.

From the Statistics to the Formulation

- Formulate *maximum a posteriori* (MAP) problem:

$$\begin{aligned} \mathbf{D} &= \mathcal{F}[\mathbf{m};\mathbf{Q}] + \boldsymbol{\epsilon}\,, \quad \boldsymbol{\epsilon} \text{ has density } \mathbf{p} \\ f(\mathbf{R}) &= -\log(\mathbf{p}) \end{aligned}$$

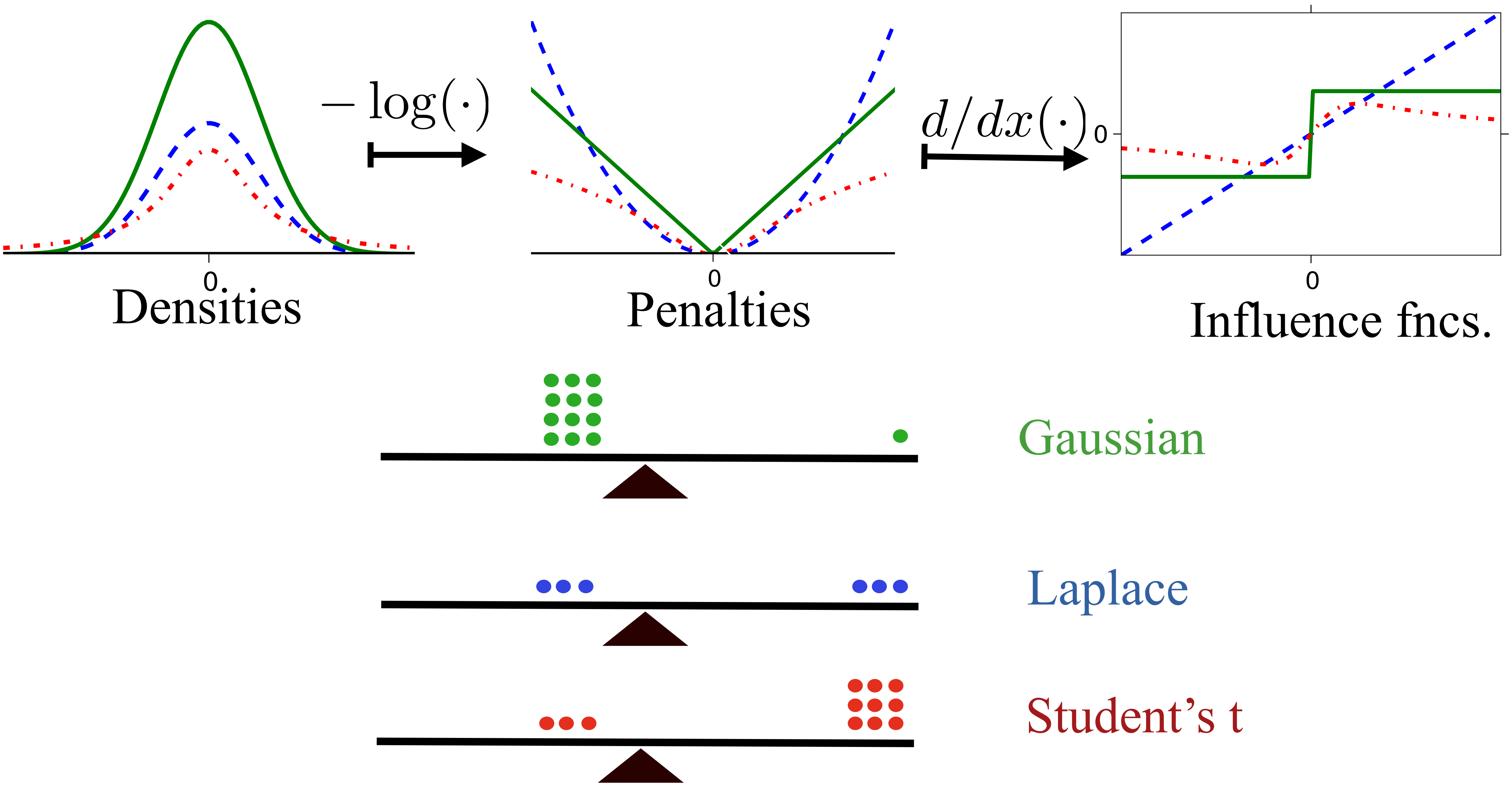
- MAP solution can be found by solving

$$\min_{\mathbf{m}} \Phi(\mathbf{m}) := f\left(\mathbf{D} - \mathcal{F}[\mathbf{m};\mathbf{Q}]\right)$$

  – NLLS: $\boldsymbol{\epsilon} \sim \mathbf{N(0,I)} \iff \Phi(\mathbf{m}) = \|\mathbf{D} - \mathcal{F}[\mathbf{m};\mathbf{Q}]\|_F^2$

- Theorem: **heavy tailed** densities correspond to **nonconvex f!**
  – Aravkin, Friedlander, Herrmann, and van Leeuwen, 2011.

Densities, Penalties, and Influence Functions



FWI Using Student's t-distribution

DENSITY:

$$\mathbf{p}(\epsilon|\mu,\sigma,k) = \frac{\Gamma(\frac{k+1}{2})}{\sigma\Gamma(\frac{k}{2})\sqrt{\pi k}}\left(1+\frac{(\epsilon-\mu)^2}{k\sigma^2}\right)^{\frac{-(k+1)}{2}}$$

FOR FWI:

$$\mathbf{p}(\epsilon|\mu=0,\sigma=1,k) \propto \left(k+\epsilon^2\right)^{\frac{-(k+1)}{2}}$$

ROBUST OBJECTIVE:

$$\min_{\mathbf{m}} \quad \mathbf{\Phi}_{St}(\mathbf{m}) := \frac{k+1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\log\left(k+(\mathbf{r_{ij}})^2\right)$$

Gradient Comparison

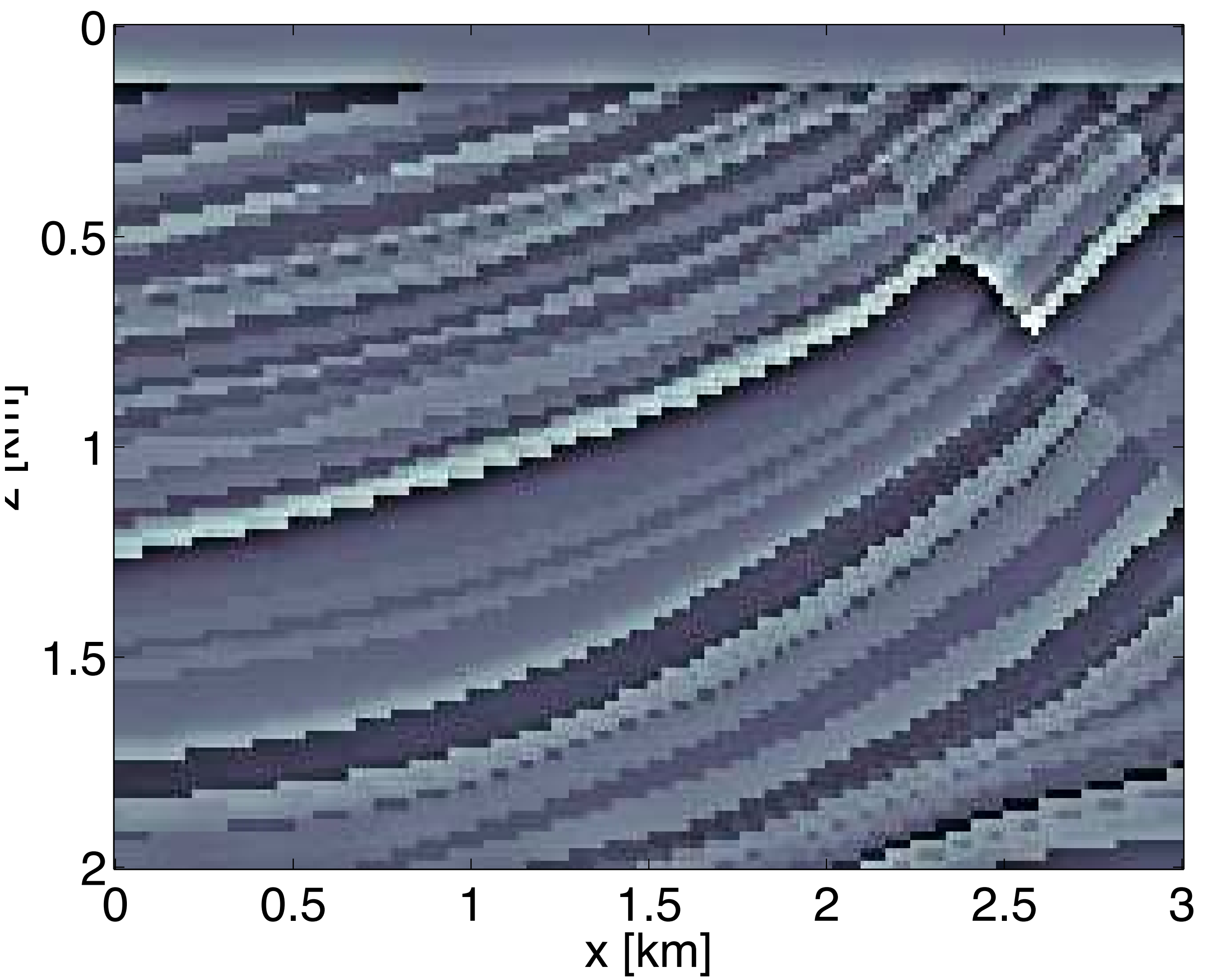
Least Squares:

$$\nabla \boldsymbol{\Phi}(\mathbf{m}) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \nabla \boldsymbol{\mathcal{F}}[\mathbf{m}, \mathbf{q}_{ij}]^T \left(\mathbf{D}_{ij} - \boldsymbol{\mathcal{F}}[\mathbf{m}, \mathbf{q}_{ij}] \right)$$

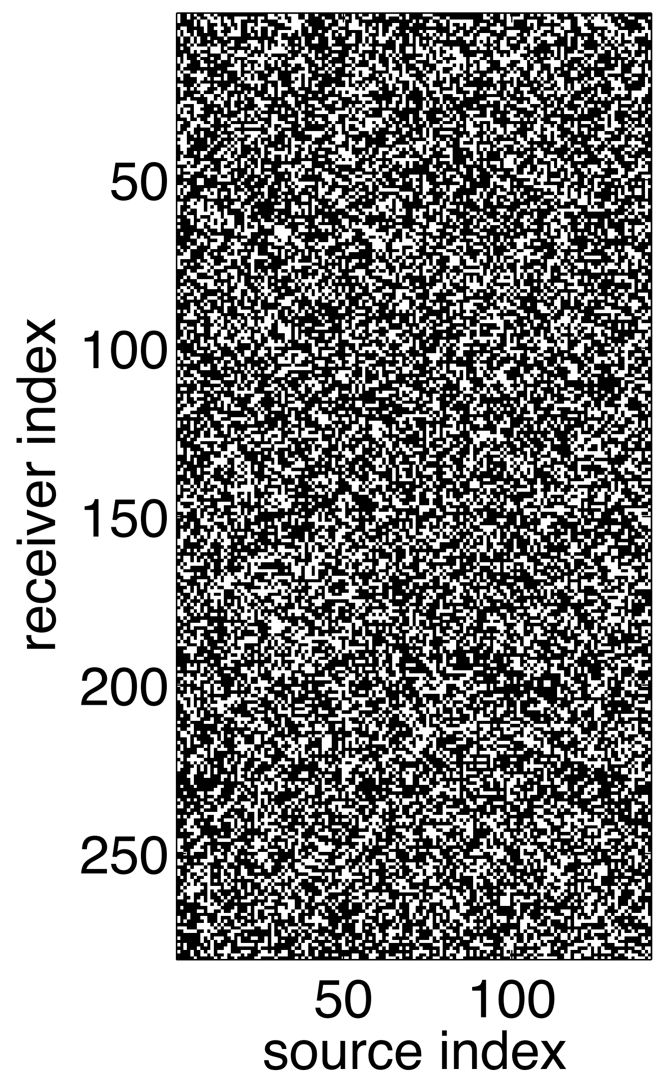
Student's t:

$$\nabla \boldsymbol{\Phi}_{St}(\mathbf{m}) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\nabla \boldsymbol{\mathcal{F}}[\mathbf{m}, \mathbf{q}_{ij}]^T \left(\mathbf{D}_{ij} - \boldsymbol{\mathcal{F}}[\mathbf{m}, \mathbf{q}_{ij}] \right)}{k + (\mathbf{D}_{ij} - \boldsymbol{\mathcal{F}}[\mathbf{m}, \mathbf{q}_{ij}])^2}$$

Marmoussi with 50% data corrupted at random

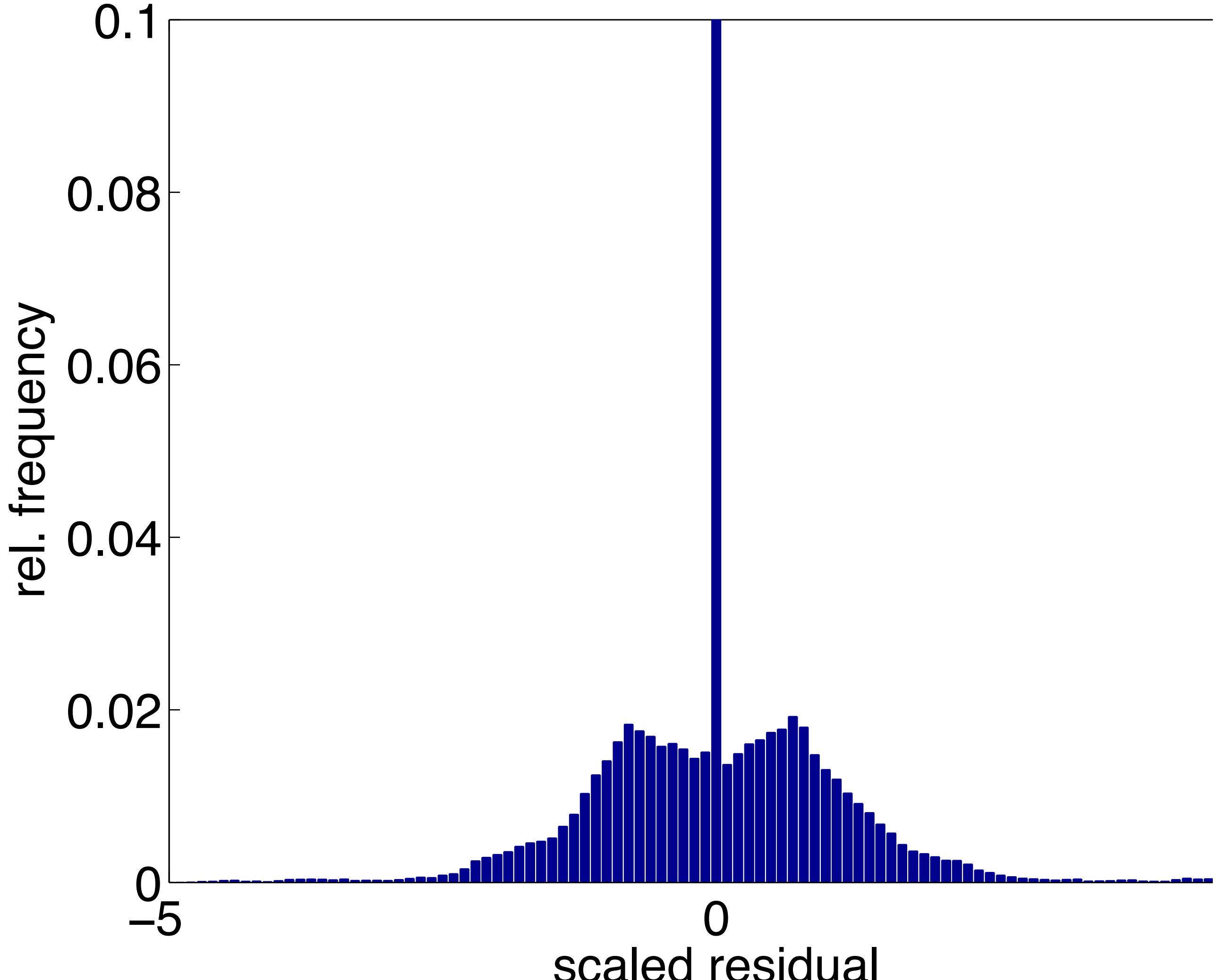


True Reflectivity

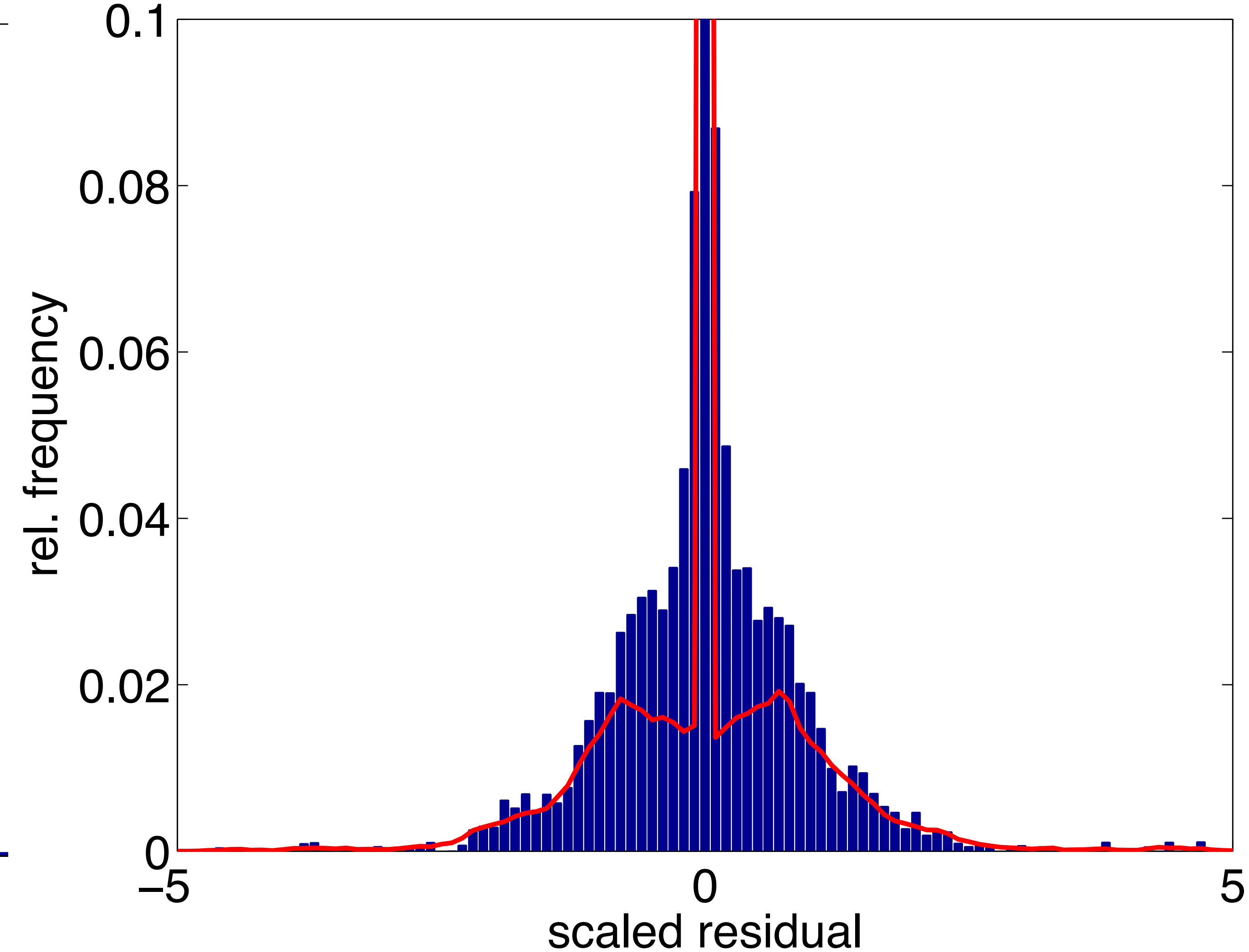


50% missing data

Histograms of residual magnitudes:

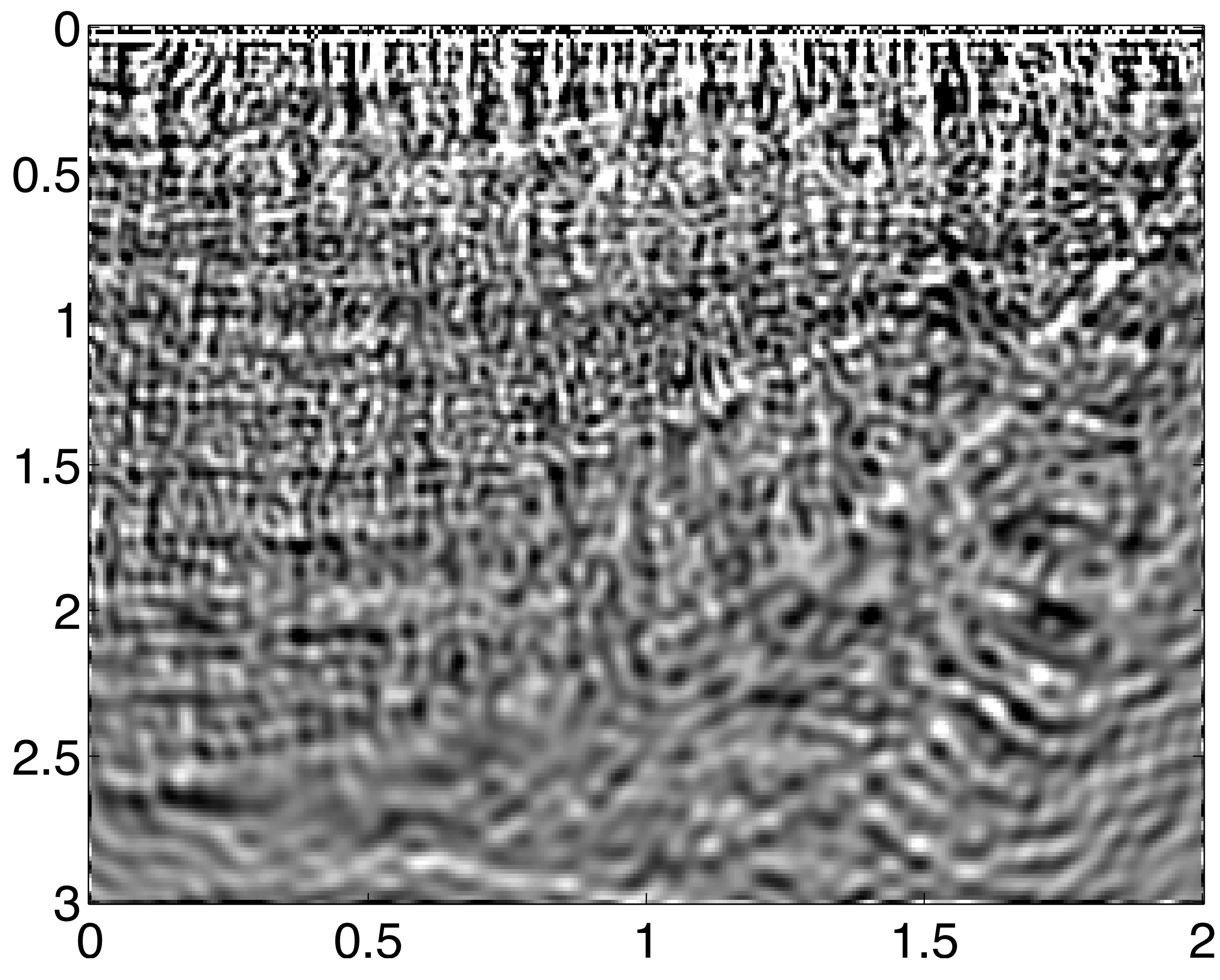


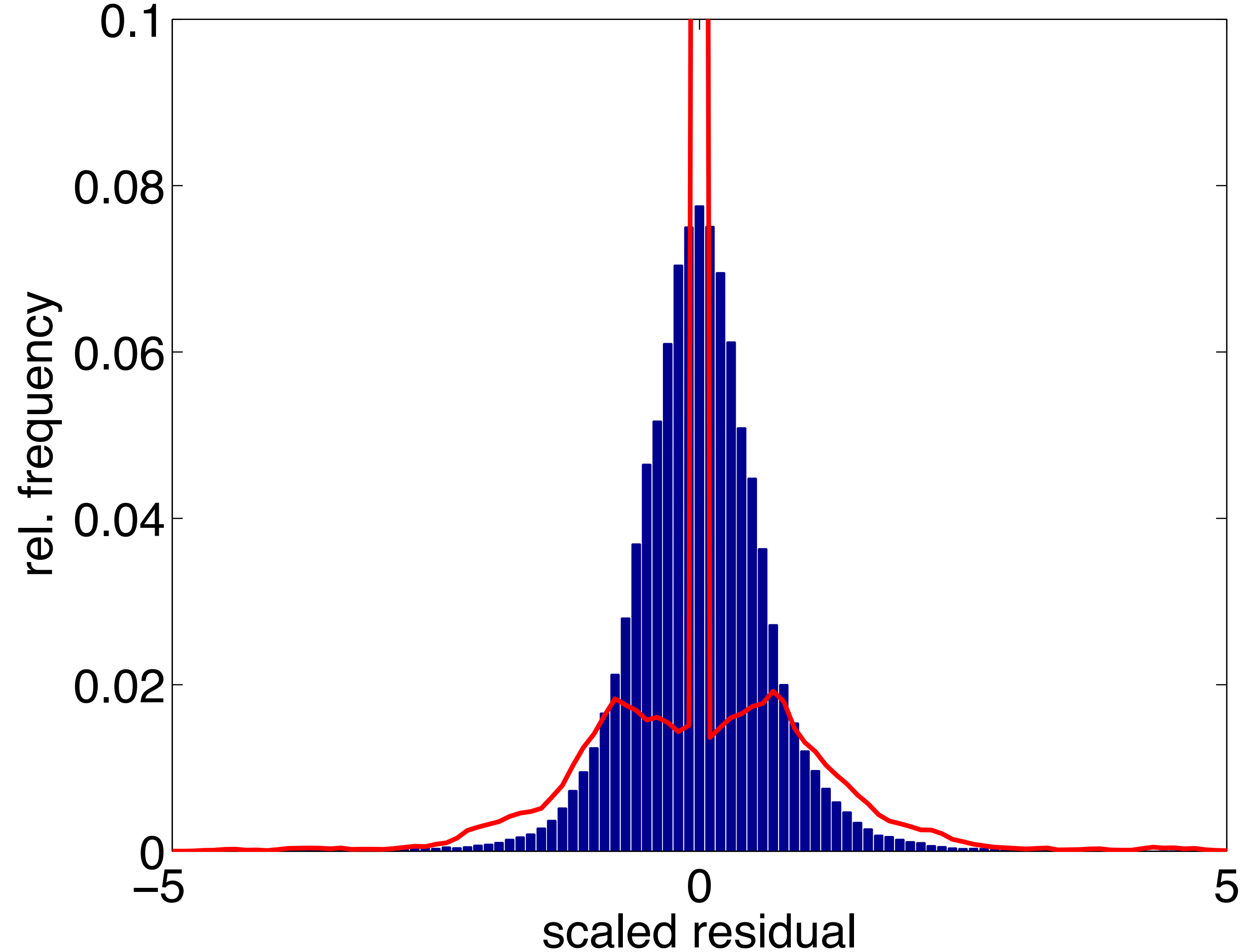
TRUE RESIDUALS



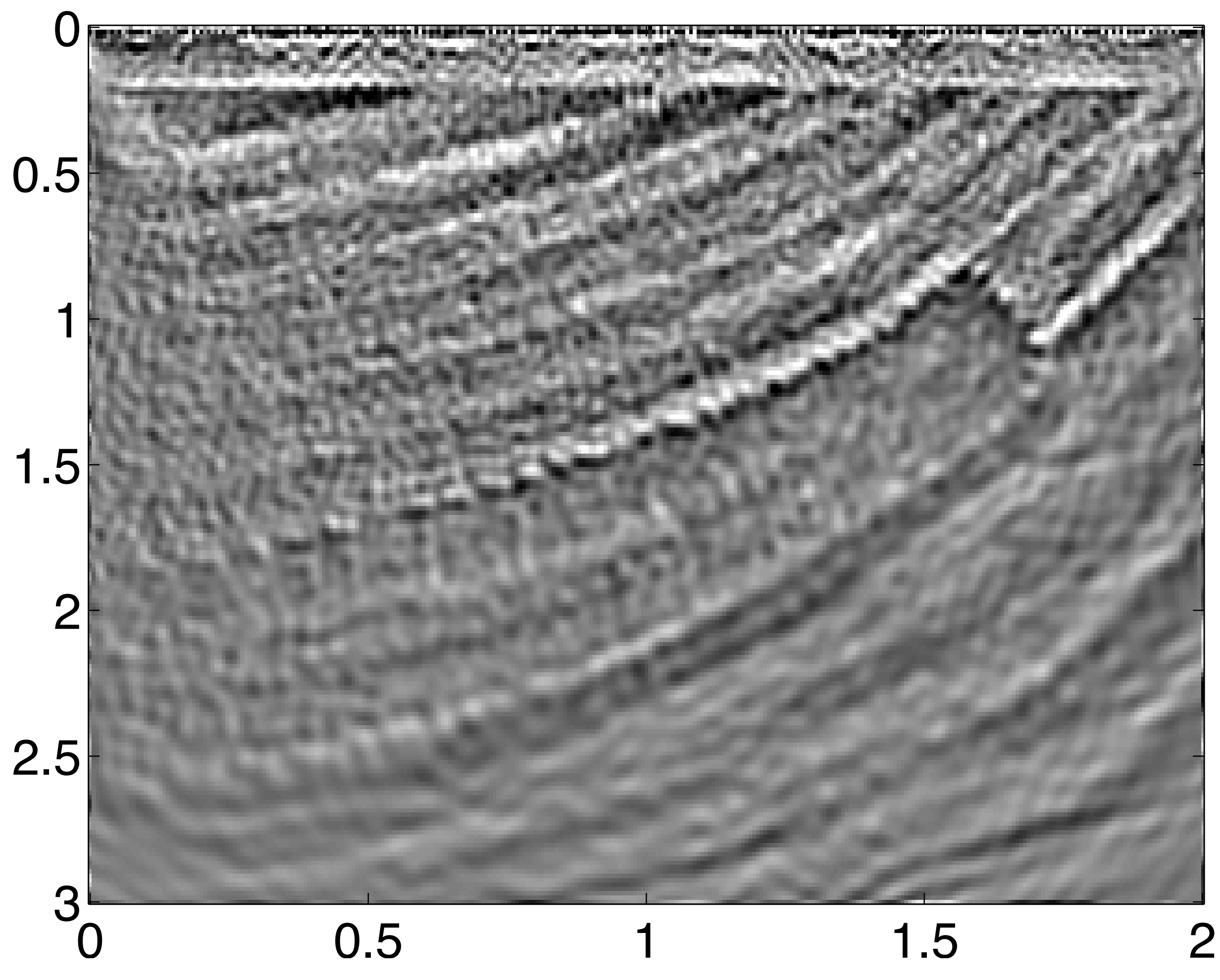
INITIAL RESIDUALS

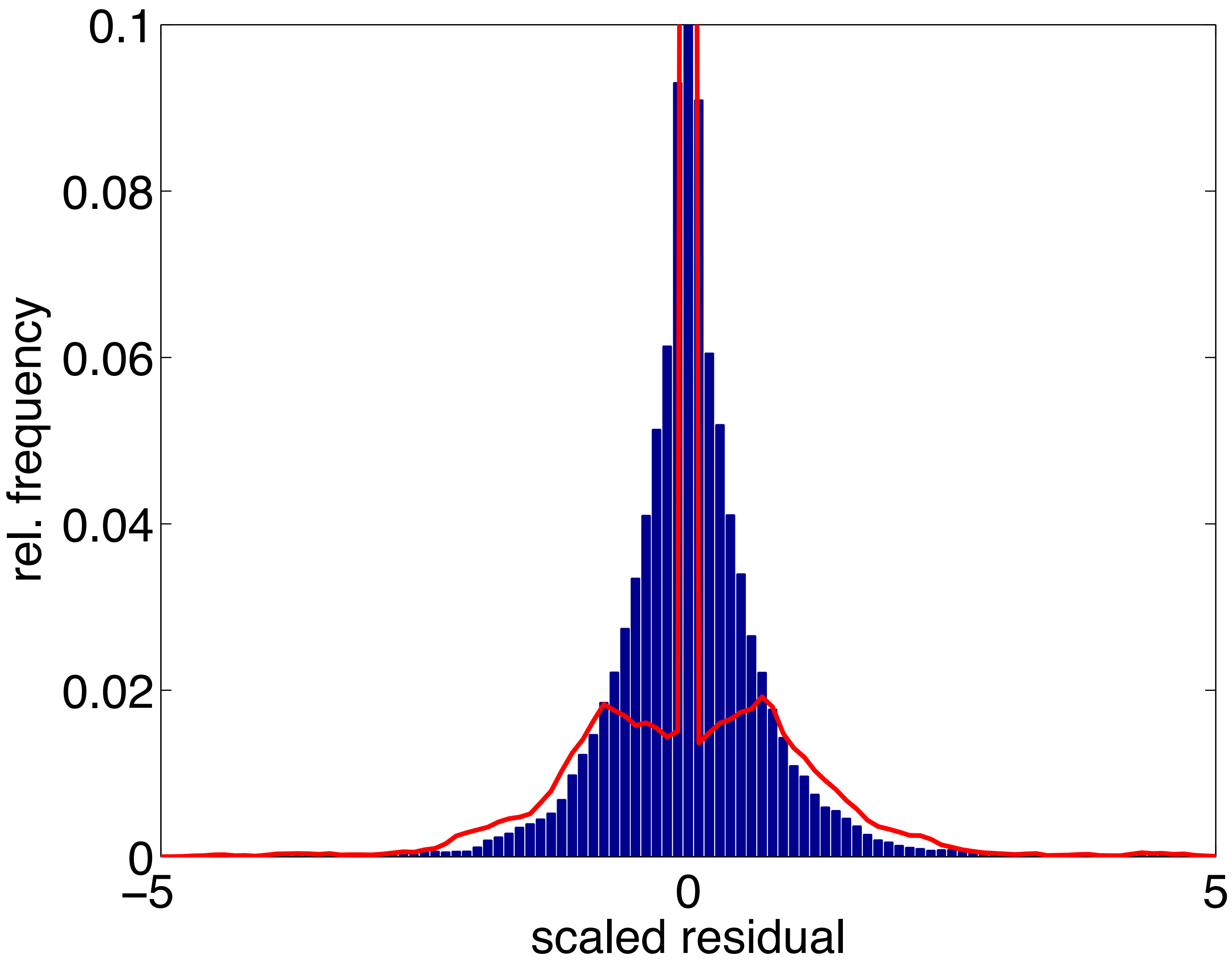
Marmoussi, LS fit, 50% corrupted data



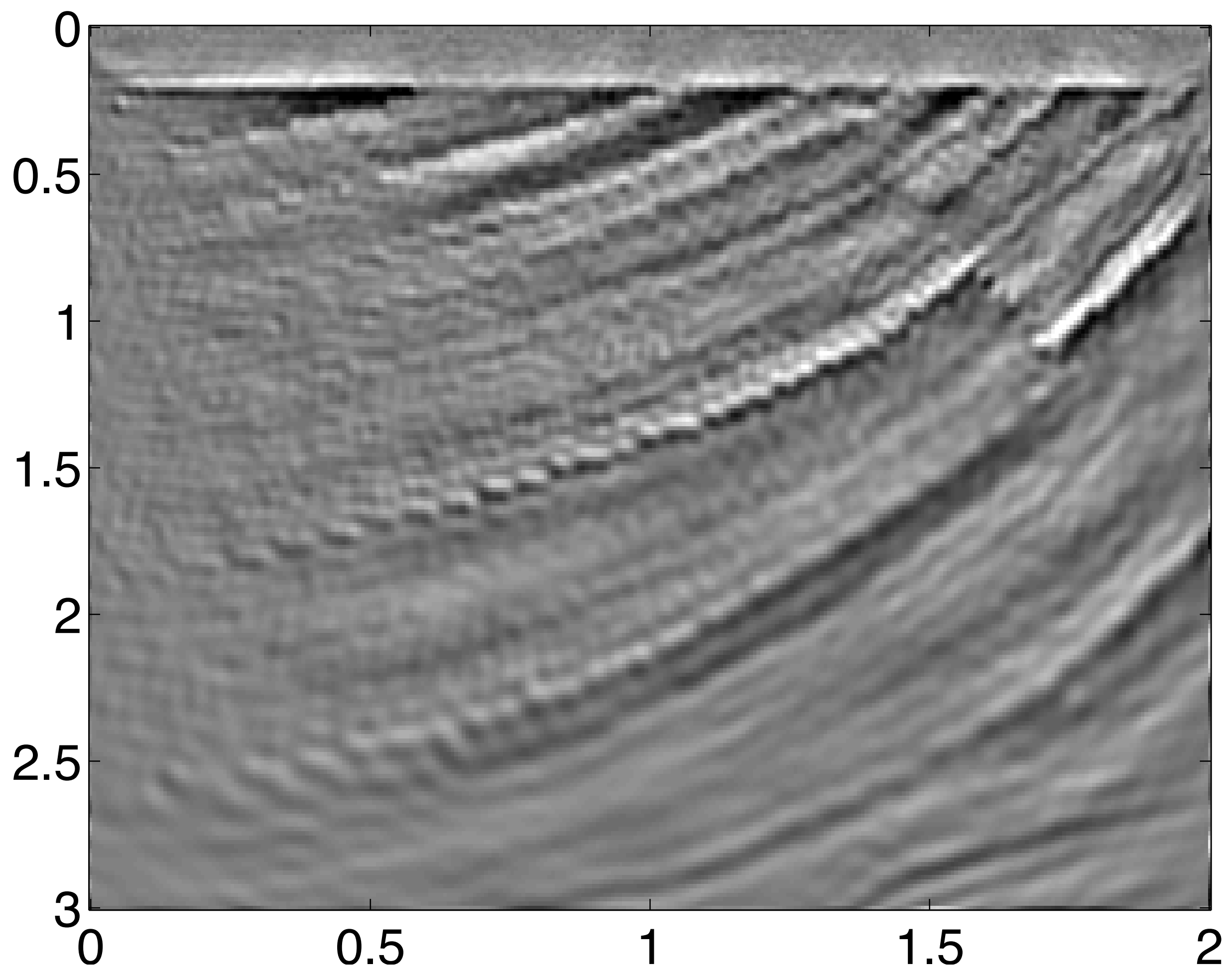


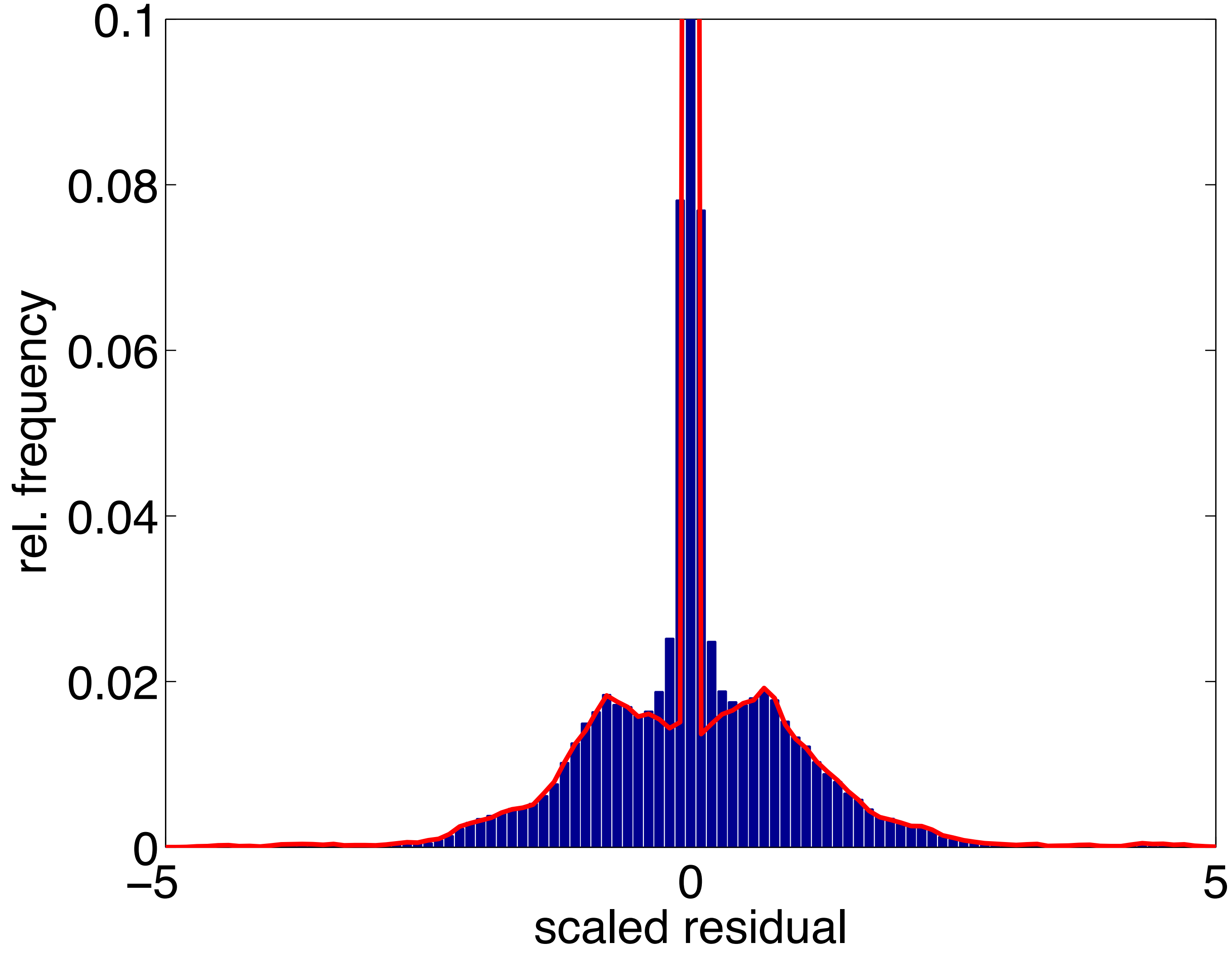
Marmoussi, Huber fit, 50% corrupted data



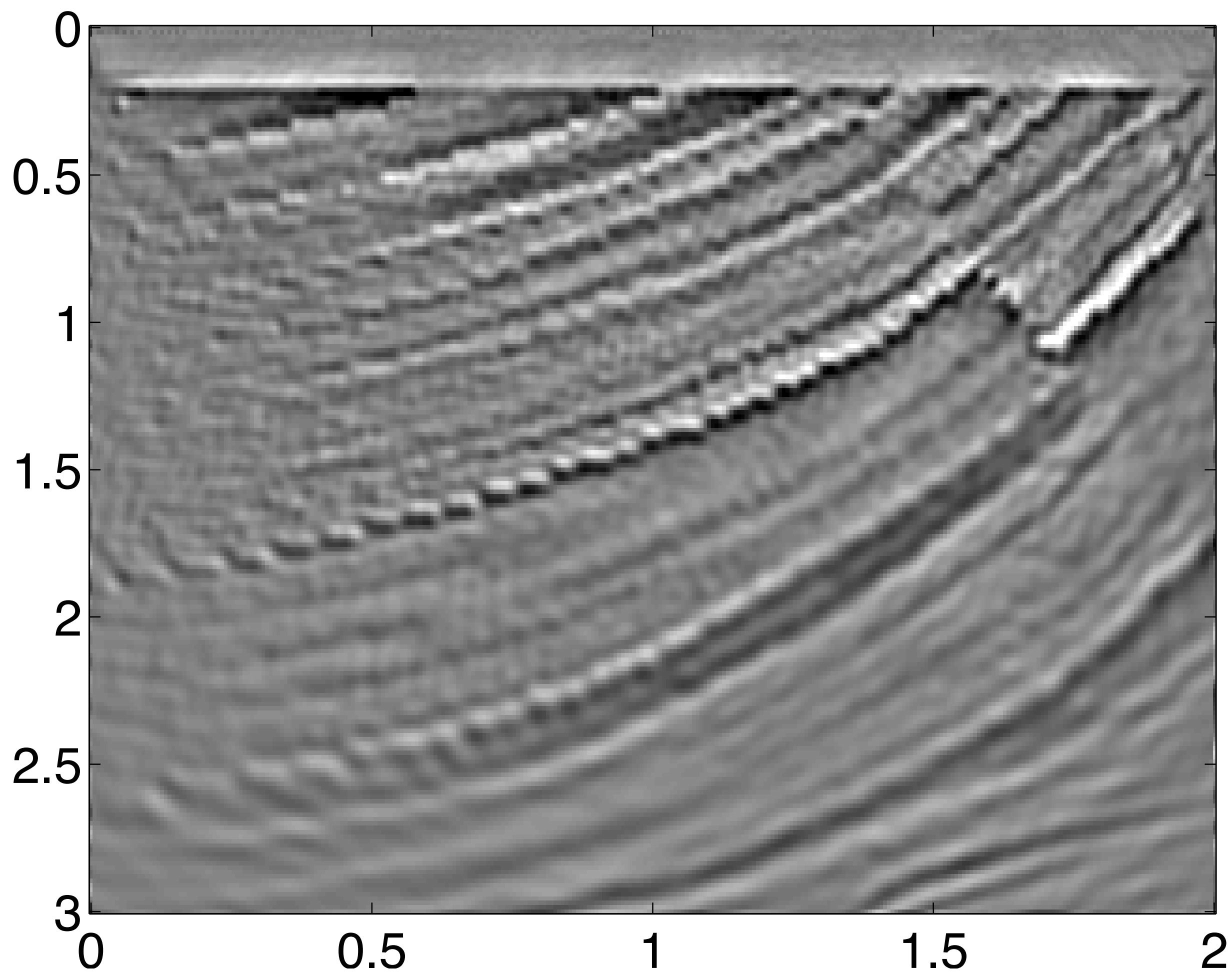


Marmoussi: T fit, 50% corrupted data

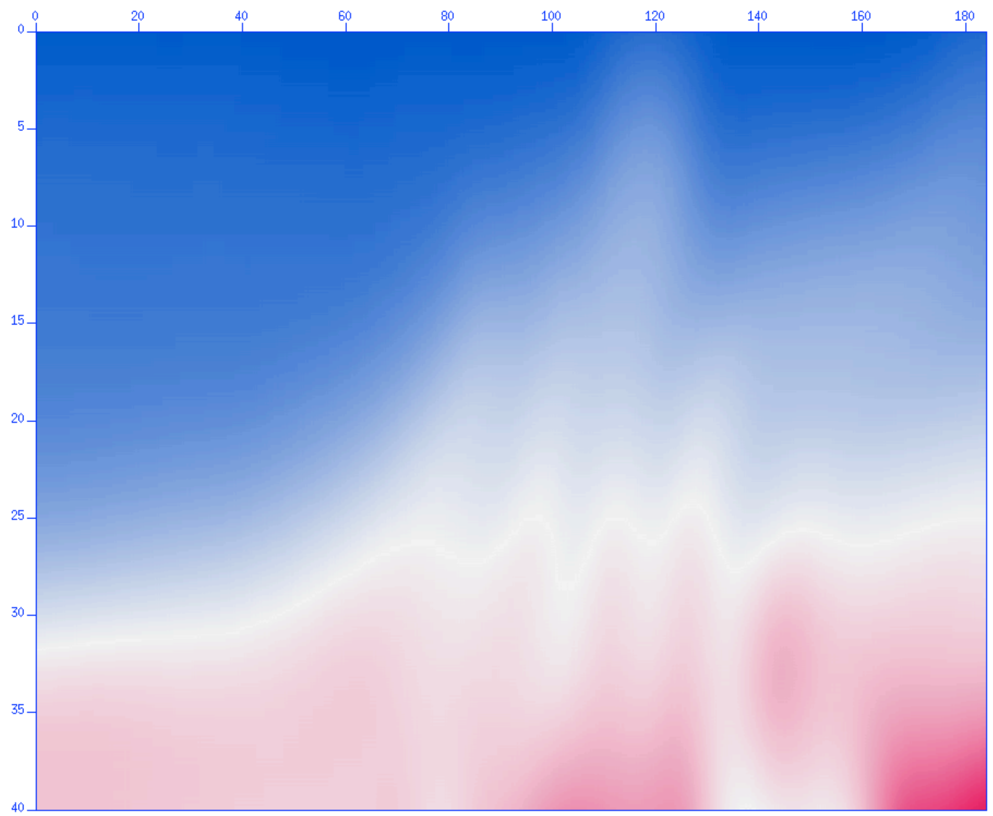




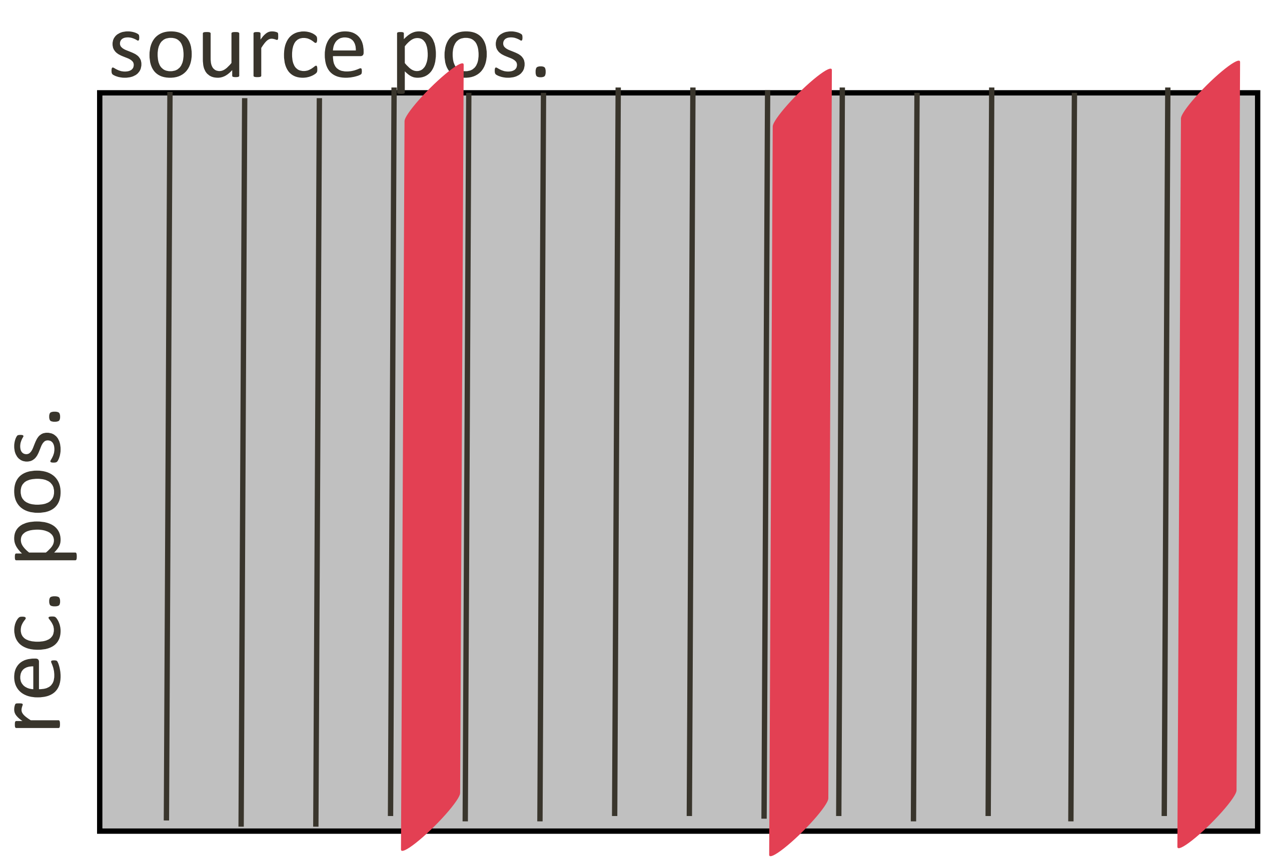
Marmoussi: LS fit, corrupted data ignored



Marmoussi II: Total Implementation

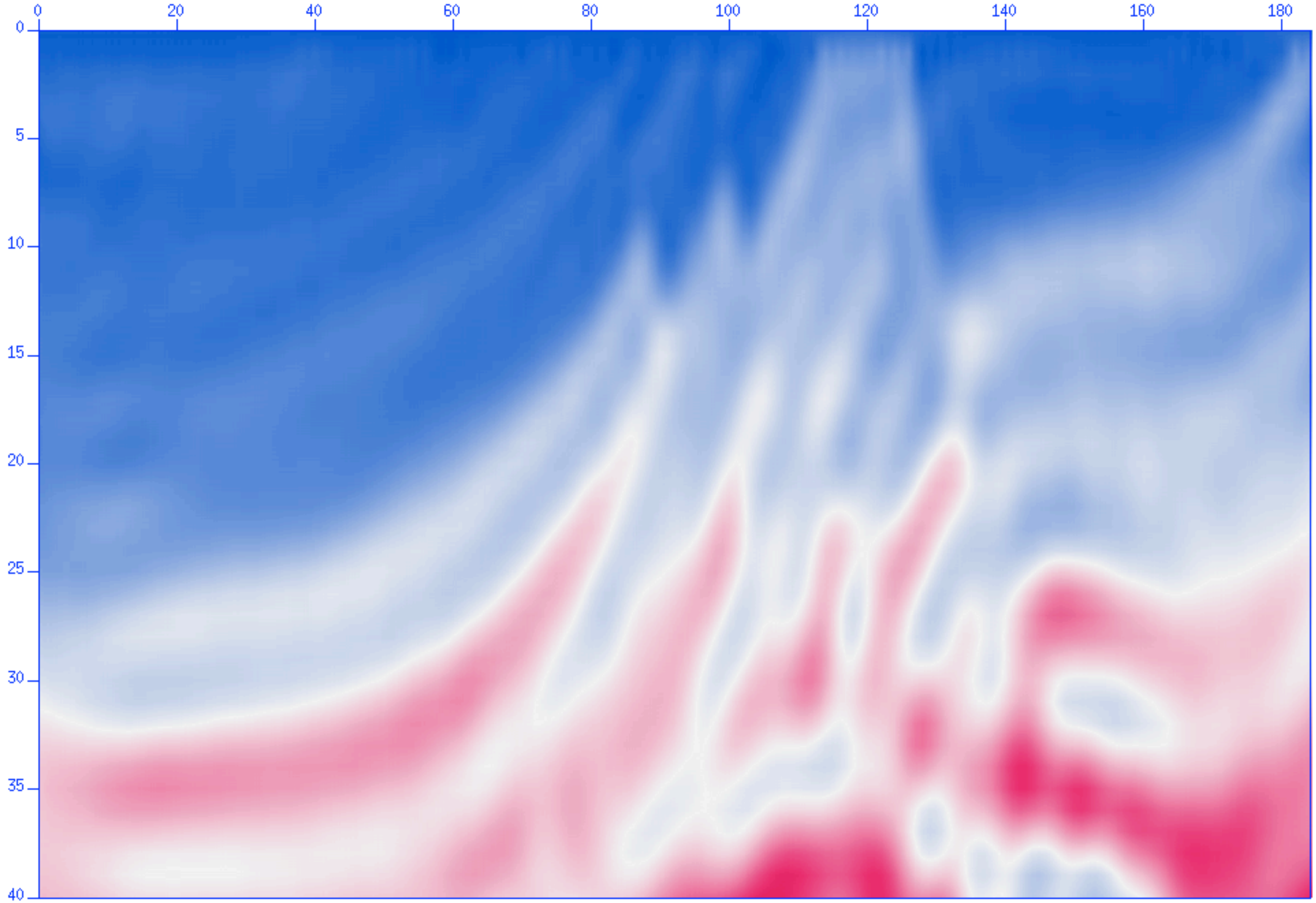


Initial Model, 4 Hz

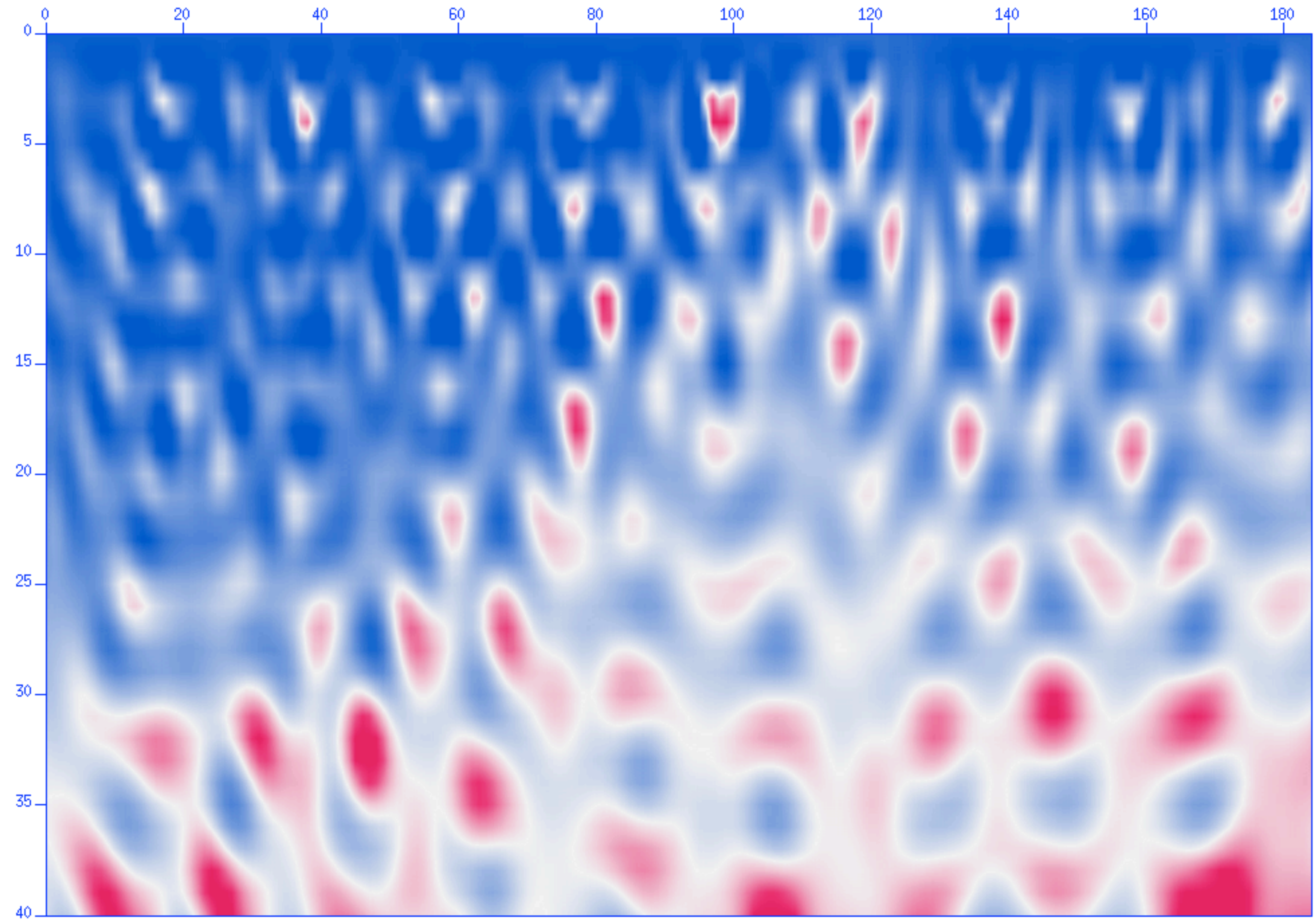


20% Corrupted Data

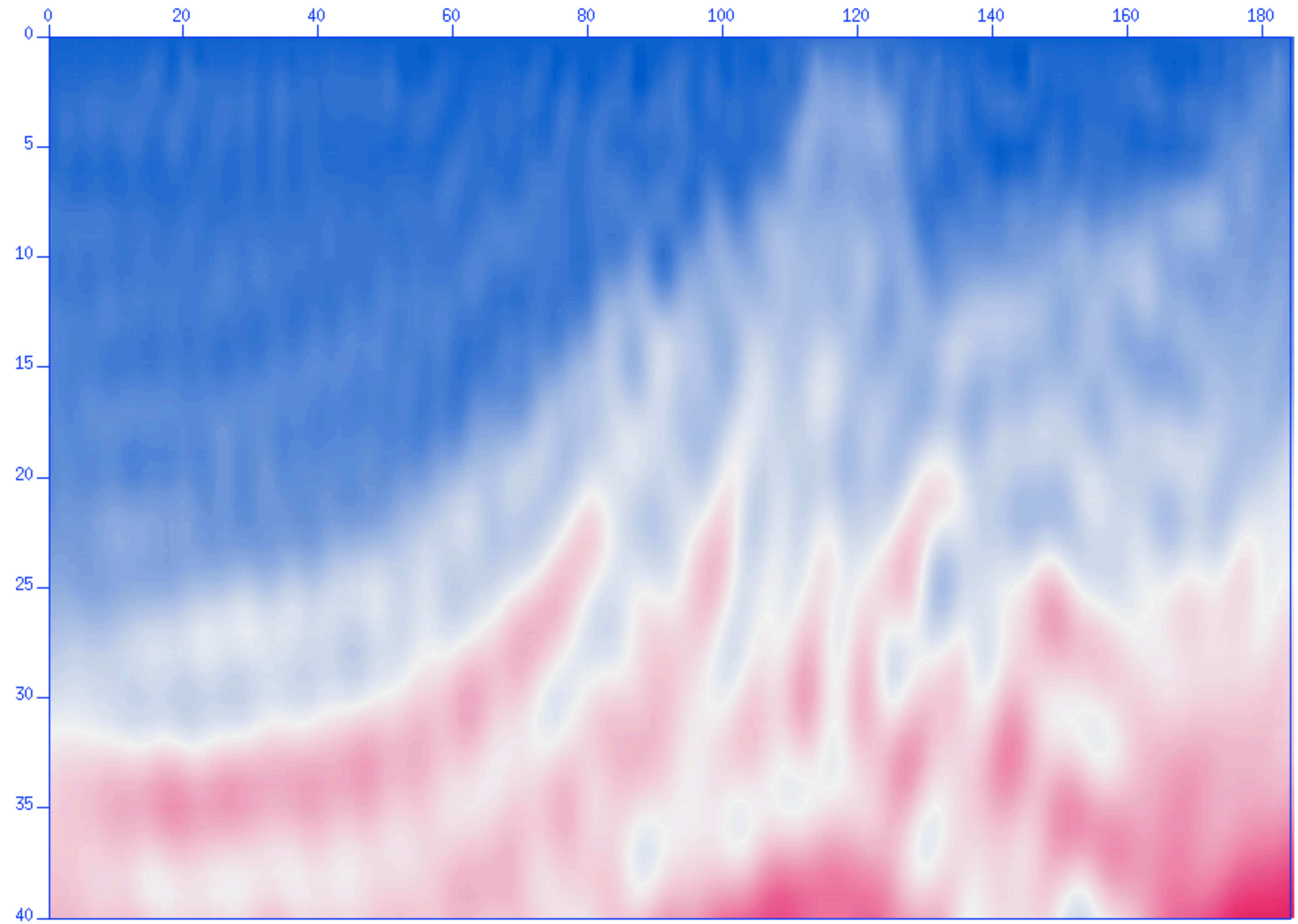
Results: Least Squares with GOOD data, 4Hz



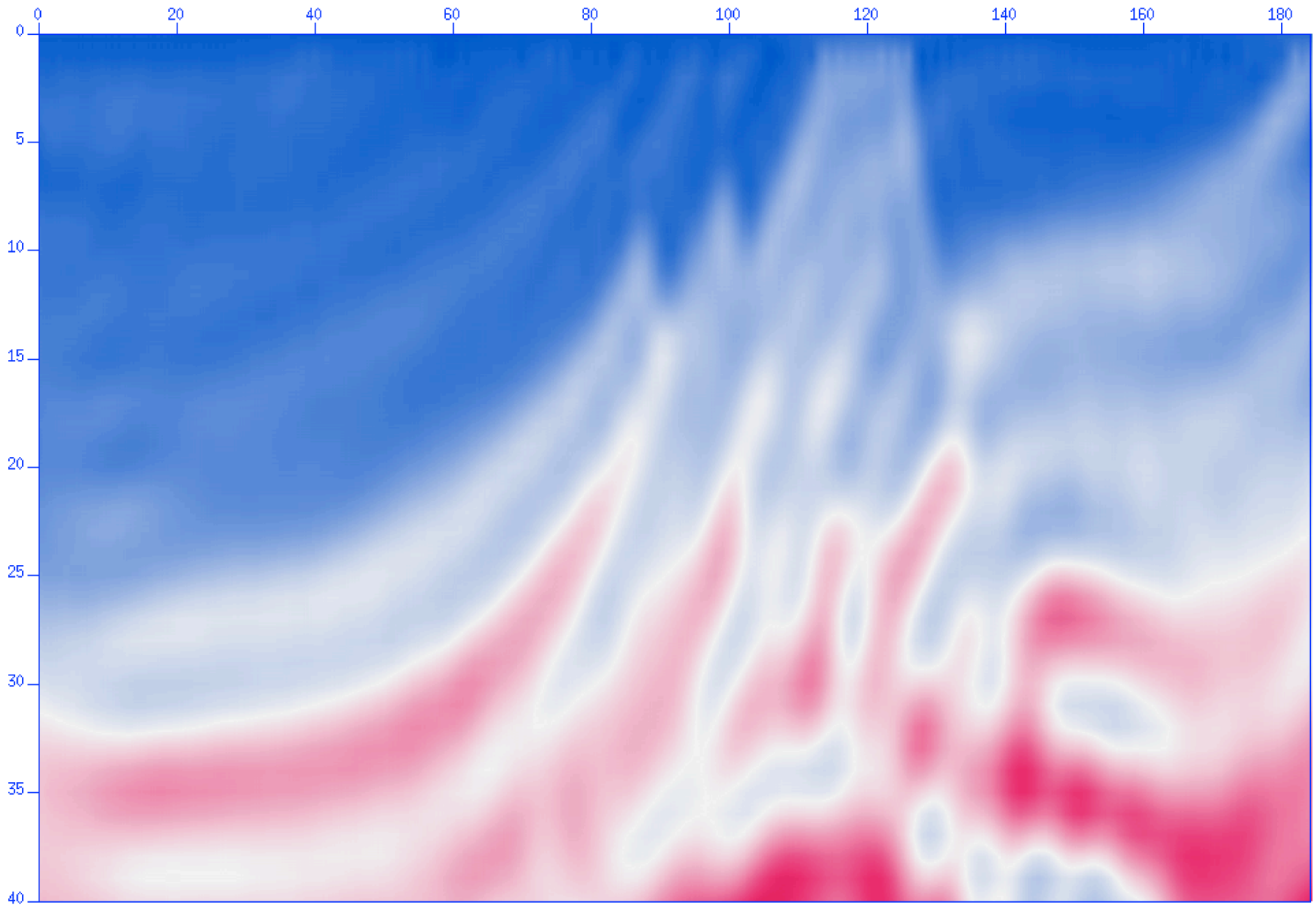
Results: Least Squares with BAD data, 4 Hz



Results: Student's t with BAD data, 10 DF, 4Hz



Results: Least Squares with GOOD data



Source Estimation for Robust Formulations

- We consider general inverse problems of the form

$$\min_{\mathbf{m},\boldsymbol{\alpha}} \Phi(\mathbf{m}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \phi_i \left(\mathbf{r_i}(\mathbf{m},\, \alpha_i)\right) \;,$$

$$\mathbf{r_i}(\mathbf{m},\, \alpha_i) := \mathbf{d_i} - \alpha_i \boldsymbol{\mathcal{F}}_\mathbf{i}(\mathbf{m})\mathbf{q_i}$$

$\mathbf{d_i}$	$n \times 1$ shot record
$\mathbf{q_i}$	$l \times 1$ source
$\mathbf{m}$	parameters to be recovered
$\boldsymbol{\mathcal{F}}_\mathbf{i}(\mathbf{m})$	Forward model (calculated data)
α_i	Unknown source amplitude
ϕ_i	Smooth misfit function (robust)

- We estimate source amplitudes and model parameters **jointly.**

Generalized Variable Projection Approach

- For fixed model parameters, obtain a function of amplitudes only:

$$\rho(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \phi_i(\mathbf{r_i}(\hat{\mathbf{m}}\ , \alpha_i))$$

- Find the optimal amplitudes by minimizing this function:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \phi_i(\mathbf{r_i}(\hat{\mathbf{m}}\ , \alpha_i))$$

- One can easily show that as long as misfit is smooth,

$$\nabla\Phi(\mathbf{m}, \boldsymbol{\alpha}(\mathbf{m})) = \nabla_{\mathbf{m}}\Phi(\mathbf{m}, \hat{\boldsymbol{\alpha}})$$

  - As long as **amplitudes are re-estimated at each step**, we iterate as usual, and still converge to a minimum of the **joint** objective.
  - The key now is to solve the amplitude-only problem **FAST**.

Newton Method for Amplitude-Only Problem

- For each amplitude, implement (scalar) Newton's method:

$$
\begin{aligned}
\alpha_i^{k+1} &= \alpha_i^k + s_i^k d_i^k \\
d_i^k &= -\frac{\nabla \phi_i(\mathbf{r_i}(\hat{\mathbf{m}}, \alpha_i^k))^T \boldsymbol{\mathcal{F}}_i(\hat{\mathbf{m}})\mathbf{q_i}}{\|\boldsymbol{\mathcal{F}}_i(\hat{\mathbf{m}})\mathbf{q_i}\|_H^2}
\end{aligned}
$$

  - When misfit is least squares, this method converges in one iteration and reduces to standard source estimation formula.
  - In general, good solutions are obtained in 5 to 10 iterations, so general source estimation requires only a little more effort than LS source estimation.
  - When the Hessian is not positive definite (Student's t), it is easy to adjust.
  - Updates to amplitudes to not require any forward modeling - note that model is fixed throughout the entire process.

Source Estimation for Student's t-Formulation

- Full algorithm:

$$\begin{aligned}
\alpha_i^{k+1} &= \alpha_i^k - \sum_j \frac{r_{ij}^k f_{ij}}{k + (r_{ij}^k)^2} \Big/ \sum_j \frac{f_{ij}^2}{k + (r_{ij}^k)^2} \\
f_{ij} &= (\boldsymbol{\mathcal{F}}_i(\mathbf{m})\mathbf{q_i})^j \\
r_{ij}^k &= d_{ij} - \alpha_i(\boldsymbol{\mathcal{F}}_i(\mathbf{m})\mathbf{q_i})^j
\end{aligned}$$

- Hessian for Student's t misfit is NOT positive definite, but we use a simple modification to design a nice method.
- In practice, the method converges in just a few iterations, without a line search.
- It took just a few hours to implement Student's t source estimation in a massively parallel FWI code at Total, thanks to Henri Calandra.

Proof of Concept

- We perform a simple experiment:
  - (1) We generate vectors of observed and predicted data randomly.
  - (2) We estimate source amplitude using LS and Student's t methods.
  - (3) We add 'outliers' to a portion of the data
  - (4) We re-estimate the source using LS and Student's t, and compare.

- Results: for vectors of 100,000 measurements, we generated outliers 10000 times larger than actual data points.

	LS Error (%)	T Error (%)
Good data	**0**	$\approx 1\%$
10 outliers	$\approx$ 1000%	$\approx 1\%$
100 outliers	$\approx$ 9000%	$\approx 1\%$

Conclusions

- Robust formulations to FWI that are able to ignore LARGE unexplained artifacts in the data.
- A particularly robust approach is obtained using **heavy tailed** densities, such as the Student's t, and corresponding **non-convex** penalties.
  - It is easy to modify an existing code base to solve the Student's t FWI problem!
- We have also derived a general methodology for robust source estimation using variable projection, and applied it to Robust FWI with Student's t.
- Future work includes strategies for automated estimation of degrees of freedom parameters **k** and uncertainty quantification.

SLIM

Acknowledgements

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, BGP, ConocoPhillips, PGS, Petrobras, Total SA, and WesternGeco.

Marmoussi Example

- We consider a subset of the Marmoussi model
- 151 shots, 301 receivers
- 9 pt. discretization of Helmholtz operator with absorbing boundary; 10 m. spacing on grid
- Sample of Frequencies [5.0, 6.0, 11.5, 14.0, 15.5, 17.5, 23.5] Hz