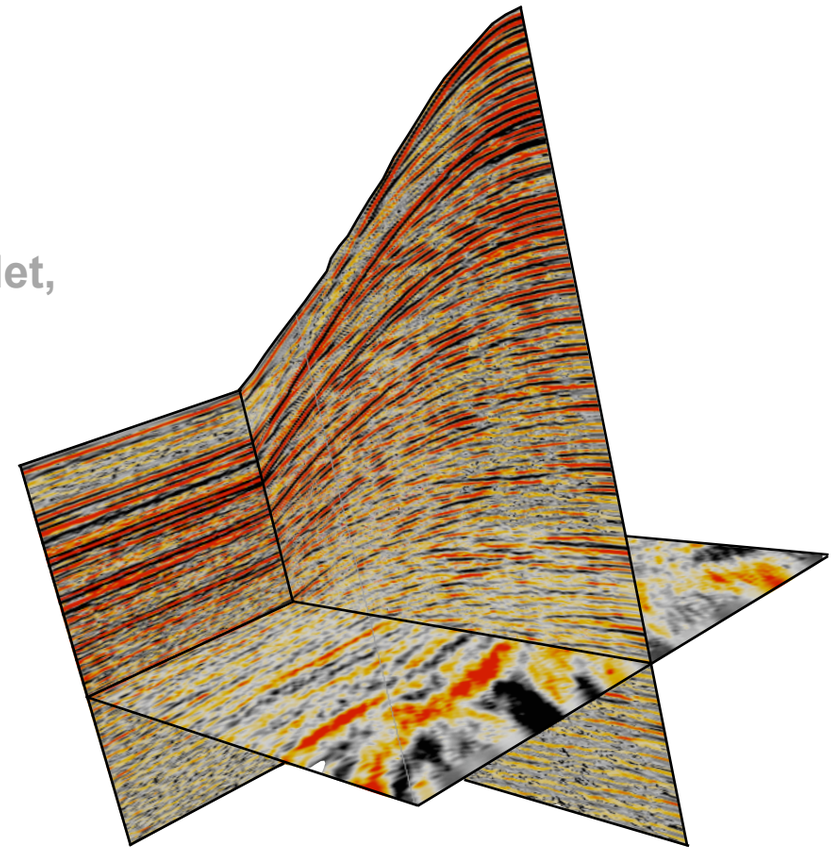
THE UNIVERSITY OF BRITISH COLUMBIA | VANCOUVER

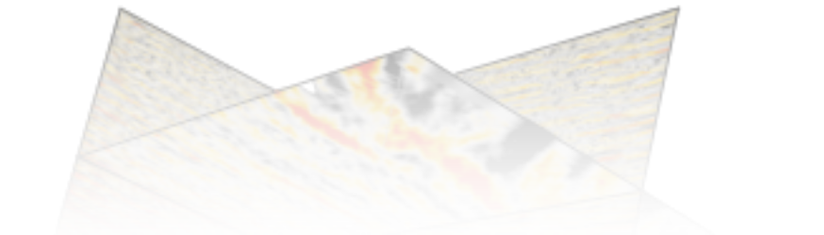
Inverse Problems using Student's t

Aleksandr Aravkin*
saravkin@eos.ubc.ca



Joint work with

Tristan van Leeuwen, Rajiv Kumar, Anais Tamalet, Henri Calandra, Michael P. Friedlander, Felix J. Herrmann.



Outline

- Robust FWI:
  - Motivation and statistical insight
  - Student's t FWI formulations and some results

- Variable Projection For Nuisance Parameters
  - Theoretical Overview
  - Robust source estimation
  - Least Squares noise estimation
  - Student's t parameter estimation

- Sparsity and Student's t
  - Motivating theoretical result
  - Generalized SPGL1 with robust error measure
  - Robust low rank algorithm

Motivation for Robust Formulations

- Errors in measurement, e.g. equipment malfunction
- 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?

Nonlinear Least Squares Formulation

- We consider inverse problems of the form

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

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

- Choice of Gaussian error leads to least squares formulation:

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

Statistical Perspective for Least Squares

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

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

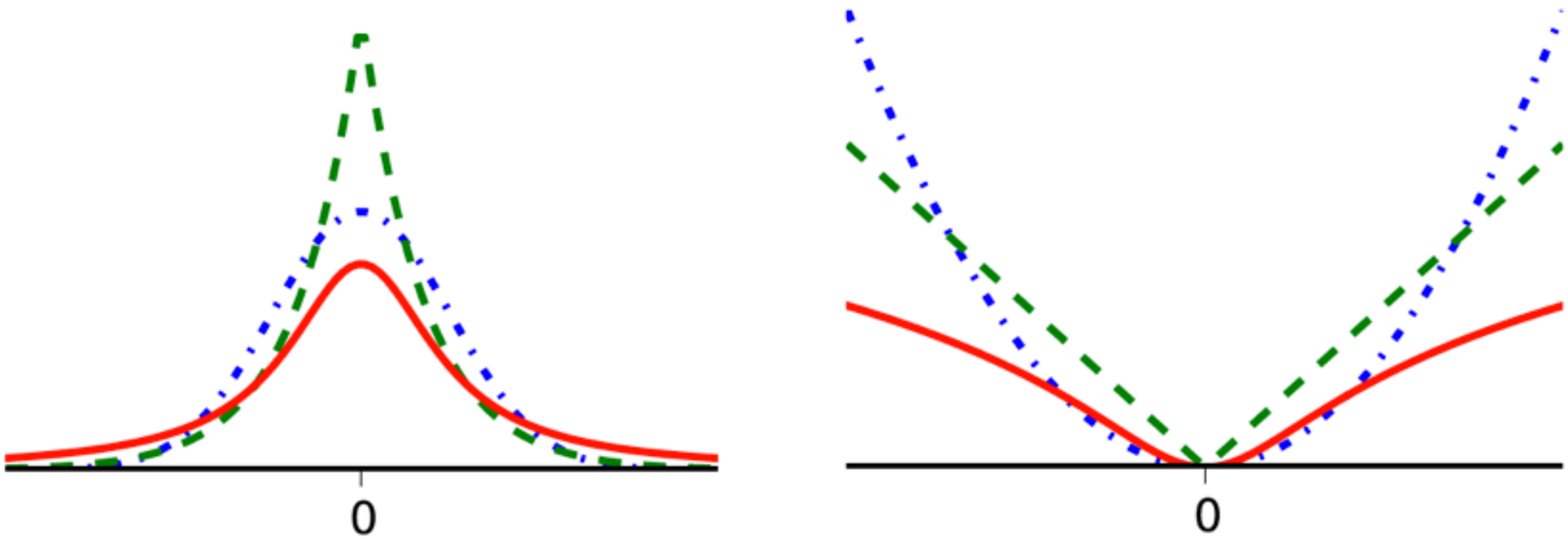
- Equivalence follows from maximum likelihood estimate for model parameters:

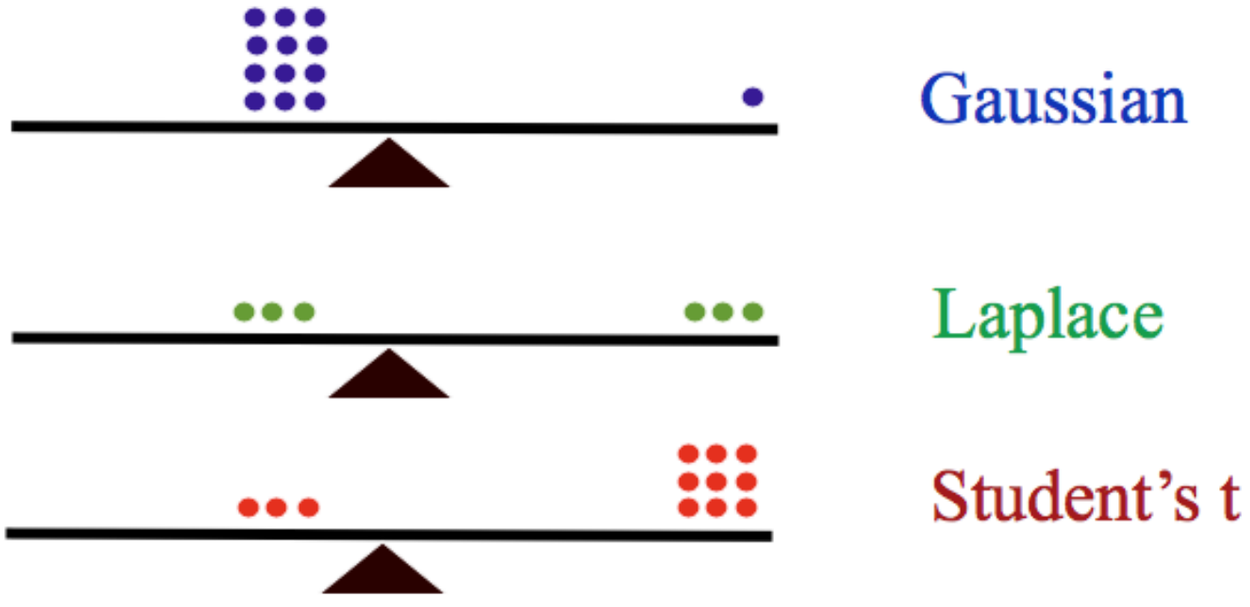
$$\boldsymbol{\mathcal{L}}(\mathbf{m}) \quad \propto \quad \exp\left(-\frac{1}{2}\left\|\mathbf{D} - \boldsymbol{\mathcal{F}}[\mathbf{m};\mathbf{Q}]\right\|_F^2\right)$$

- 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!

Densities, Penalties, and Influence Functions





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.

FWI Using Student's t-distribution

Density:

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

For FWI:

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

Robust Objective:

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

Gradient Comparison

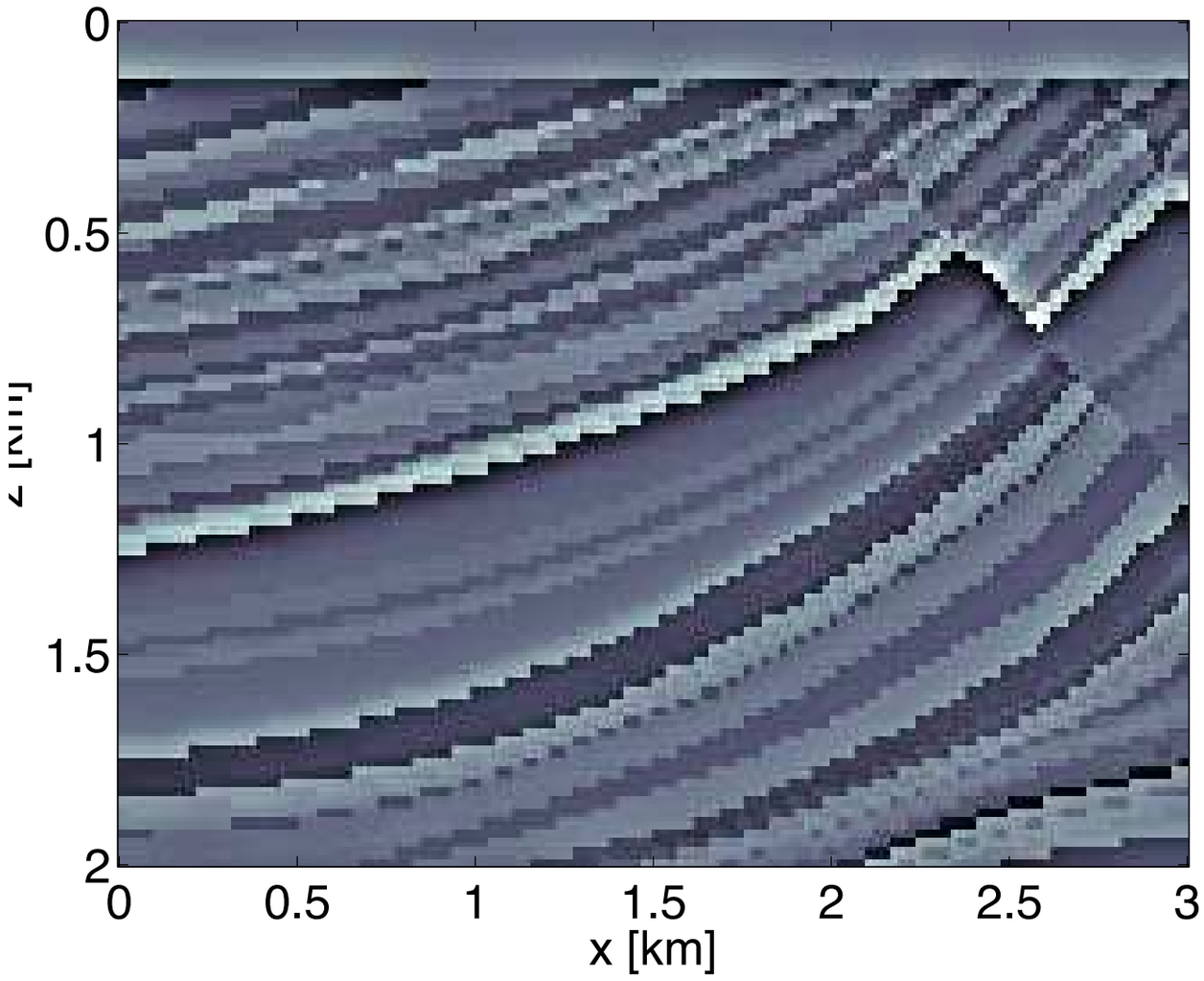
Least Squares:

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

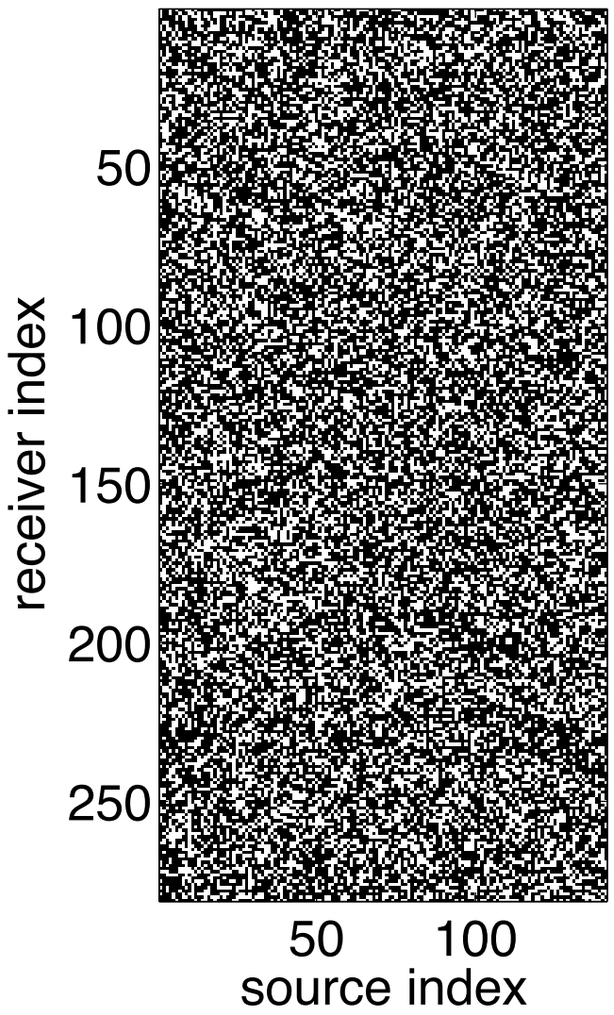
Student's t:

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

Marmoussi with 50% data corrupted at random

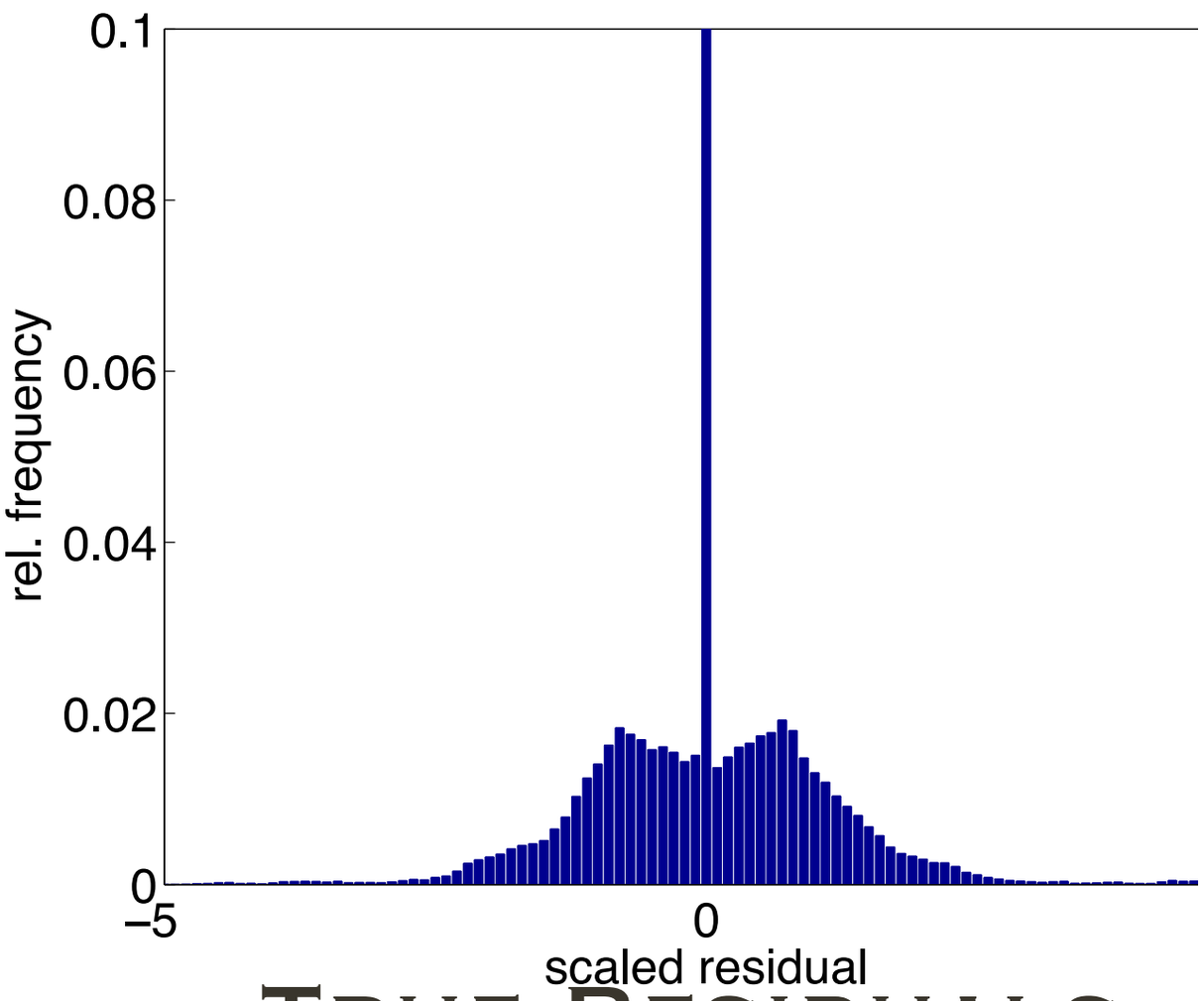


True Reflectivity

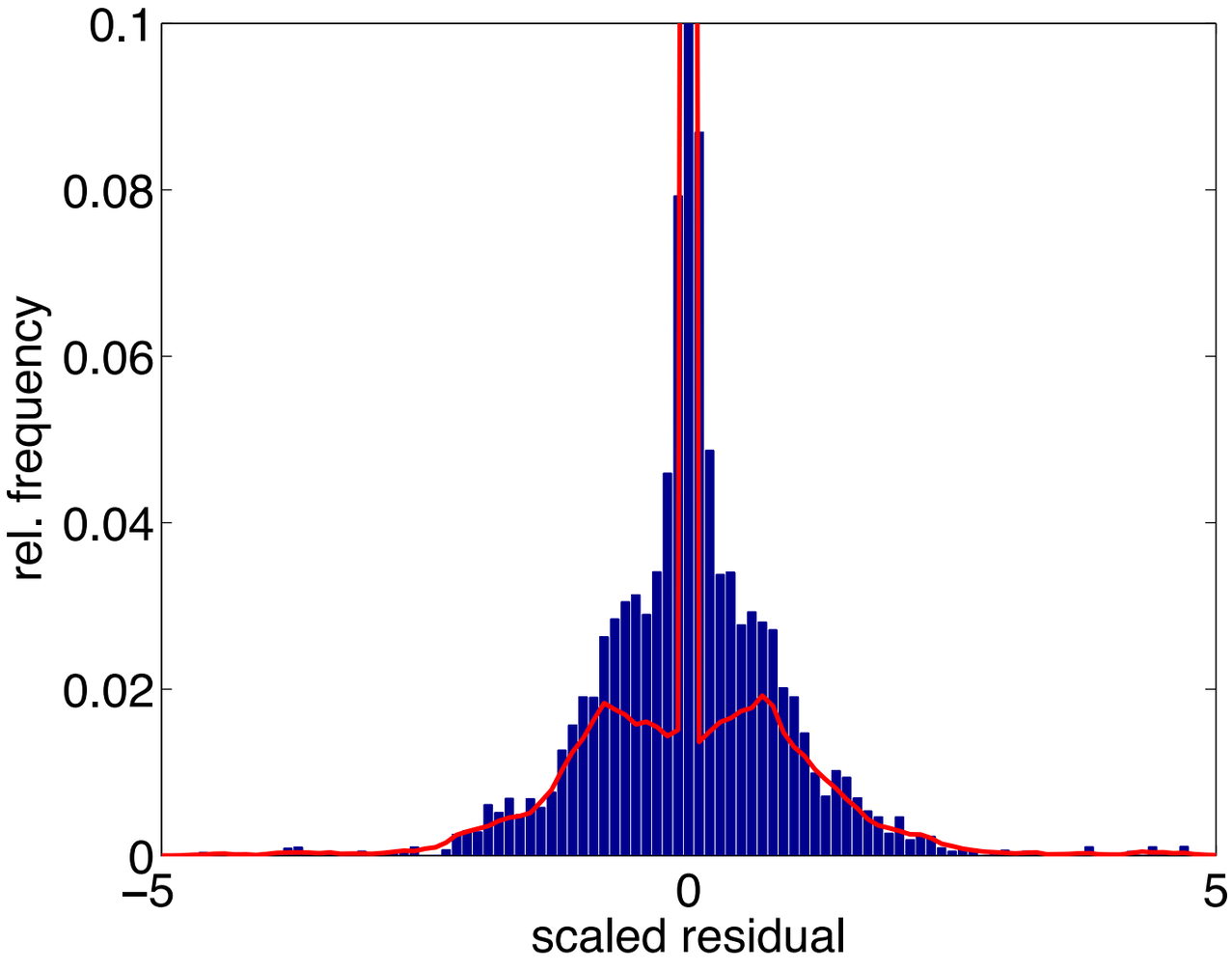


50% missing data

Histograms of residual magnitudes:

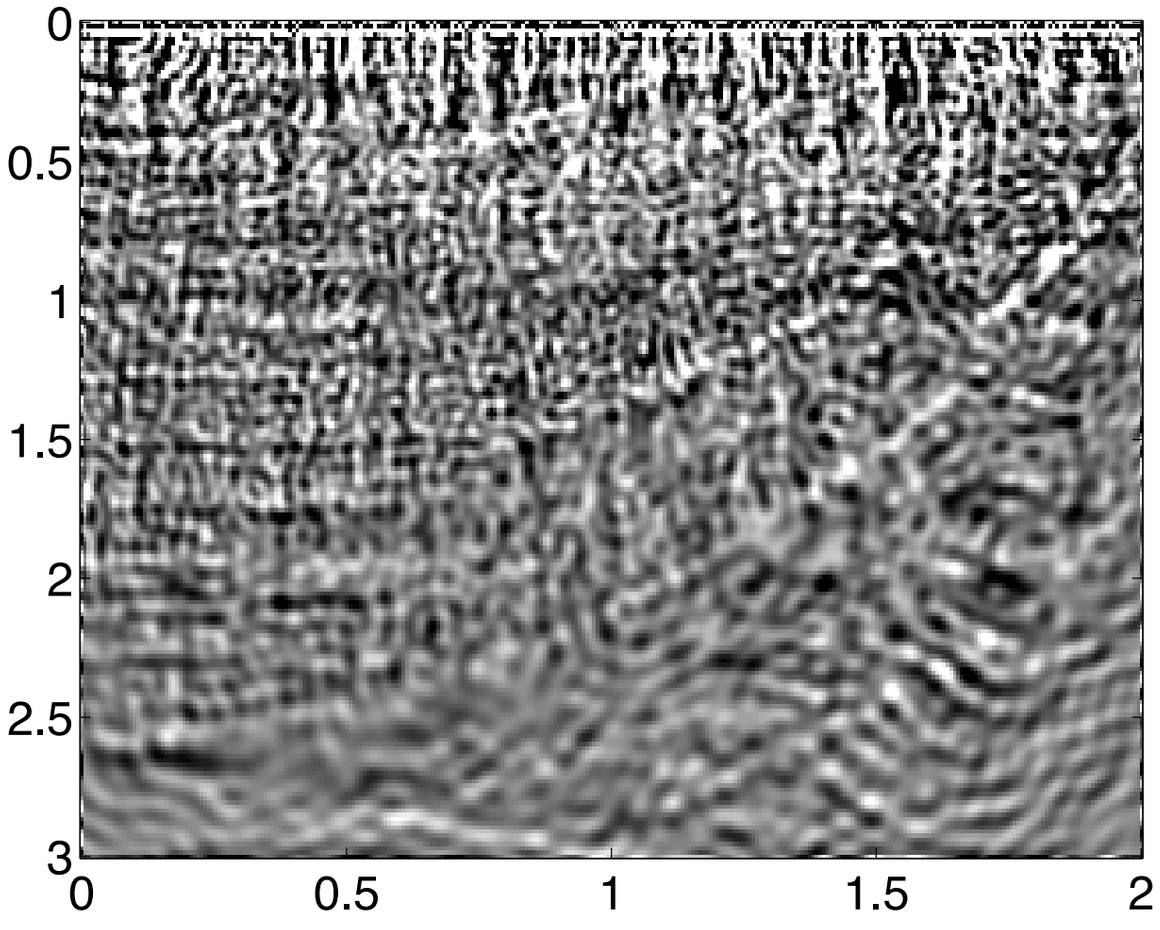


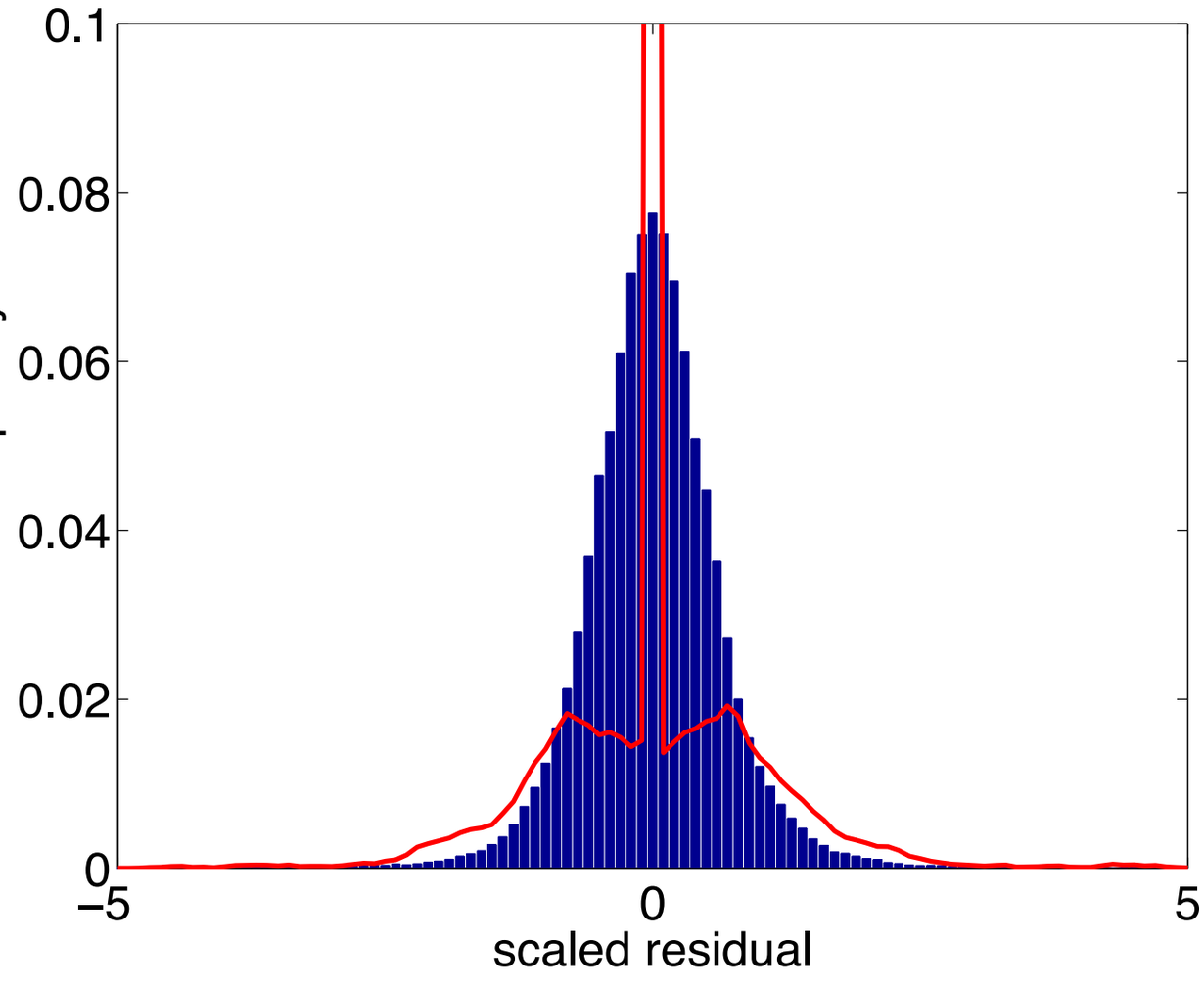
True Residuals



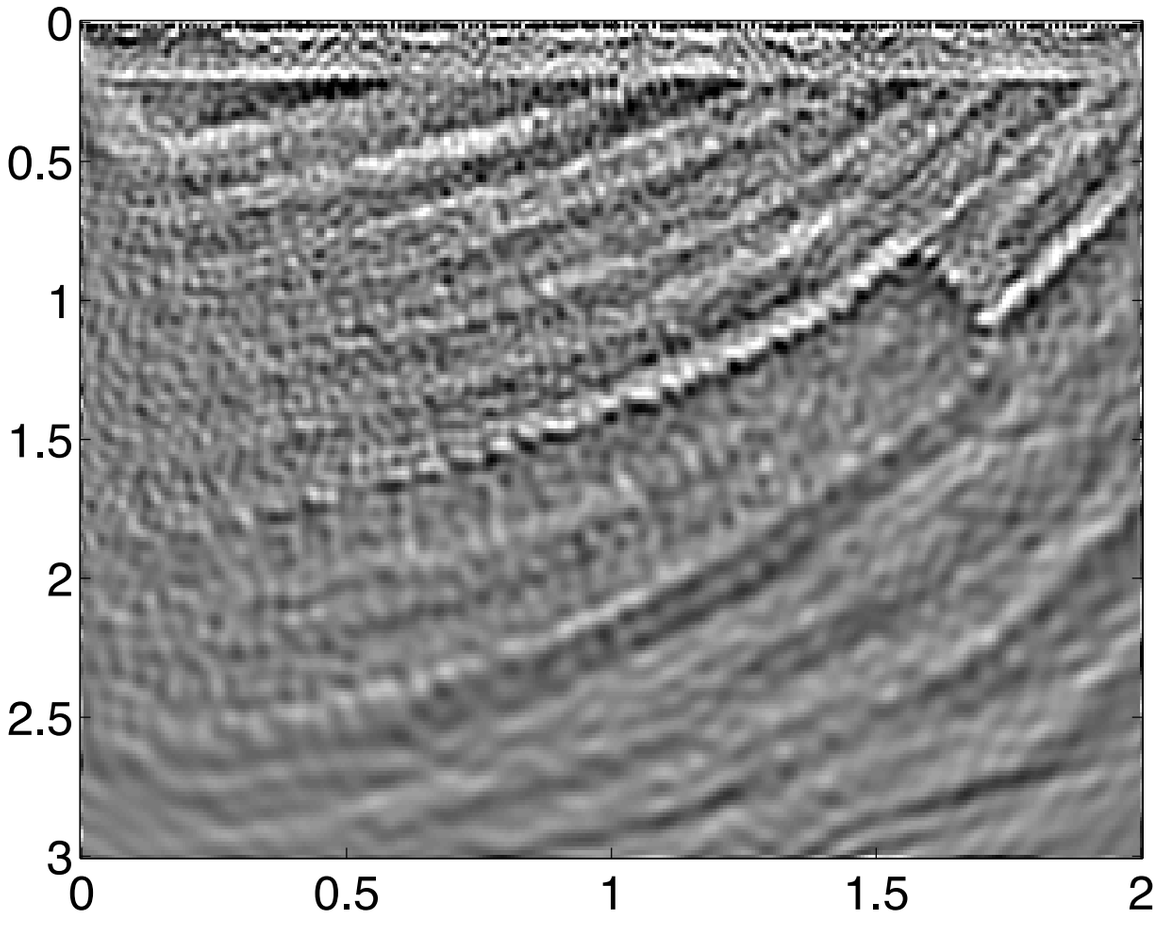
Initial Residuals

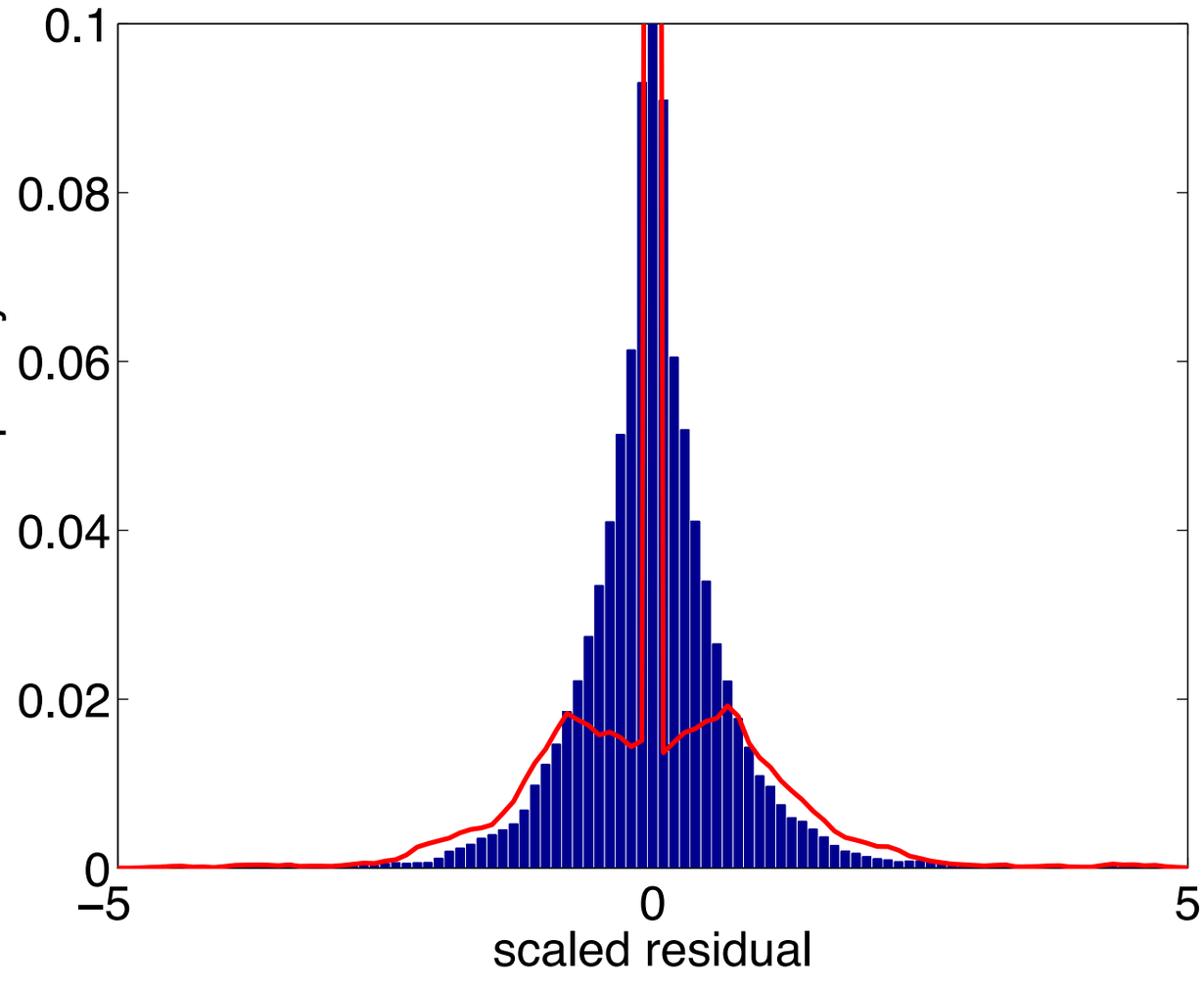
Marmoussi, LS fit, 50% corrupted data



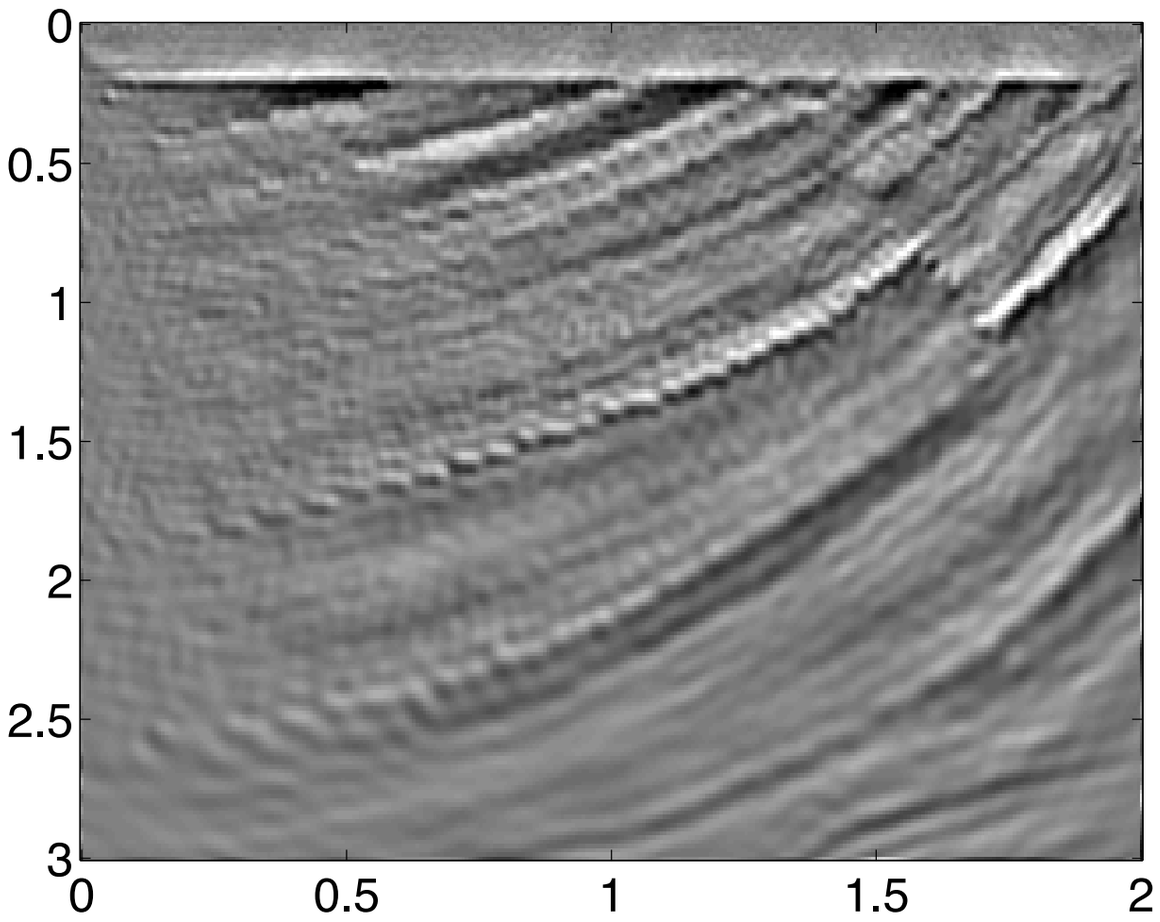


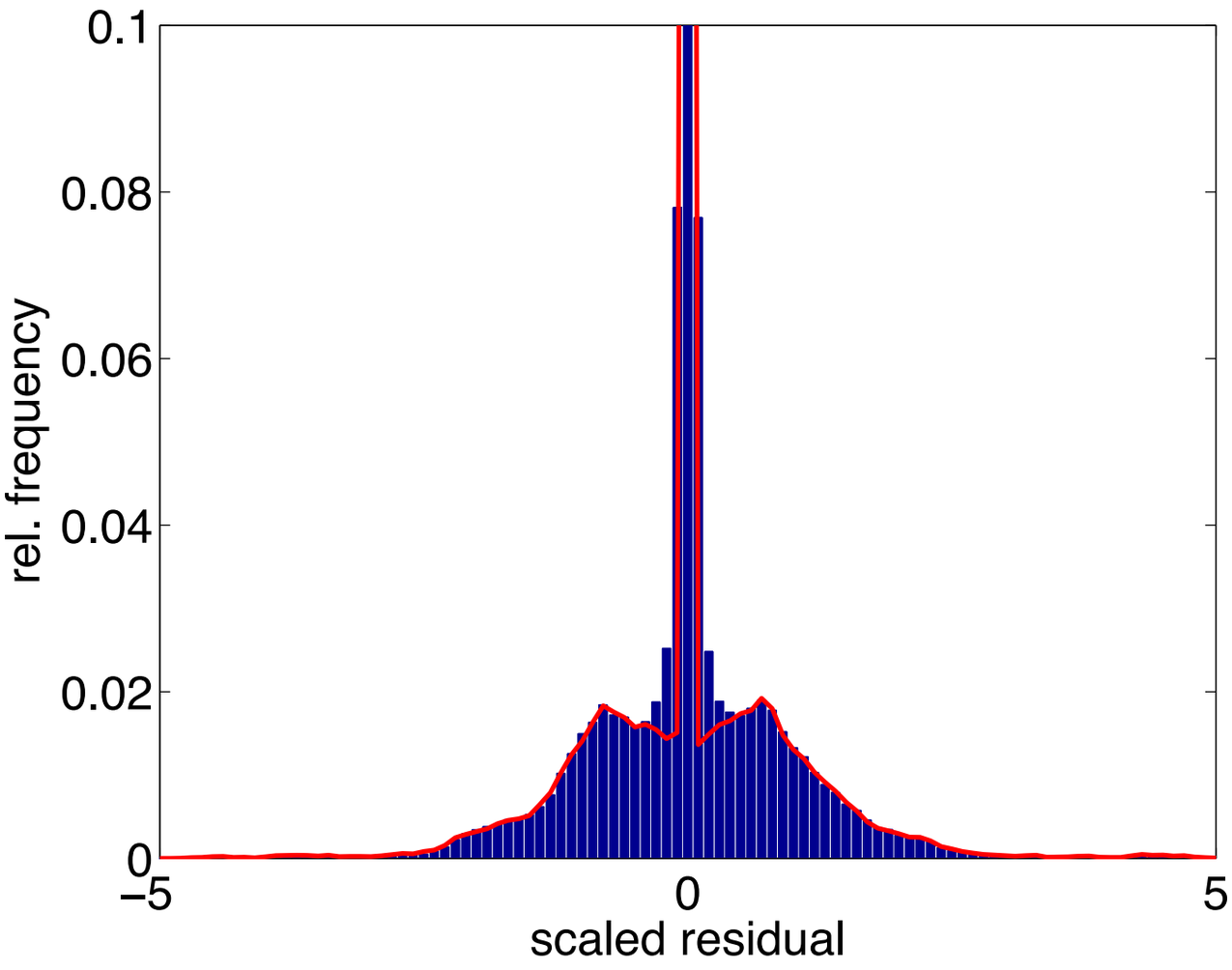
Marmoussi, Huber fit, 50% corrupted data



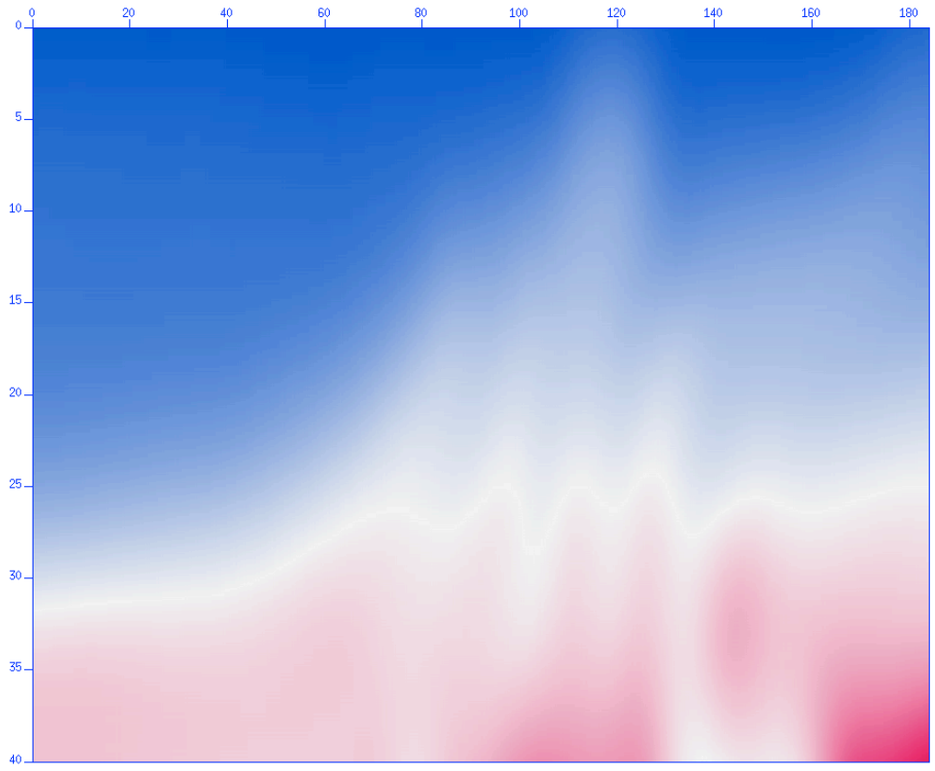


Marmoussi: T fit, 50% corrupted data

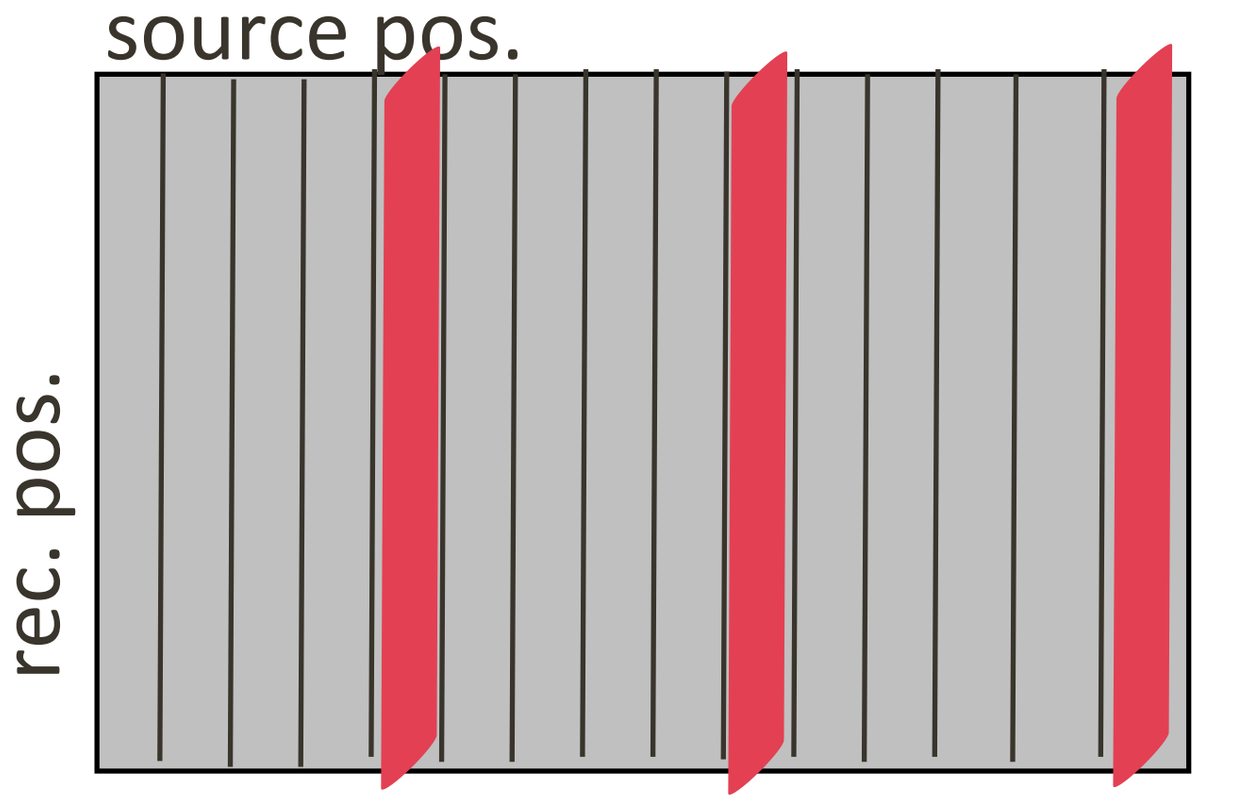




Marmoussi II: Total Implementation

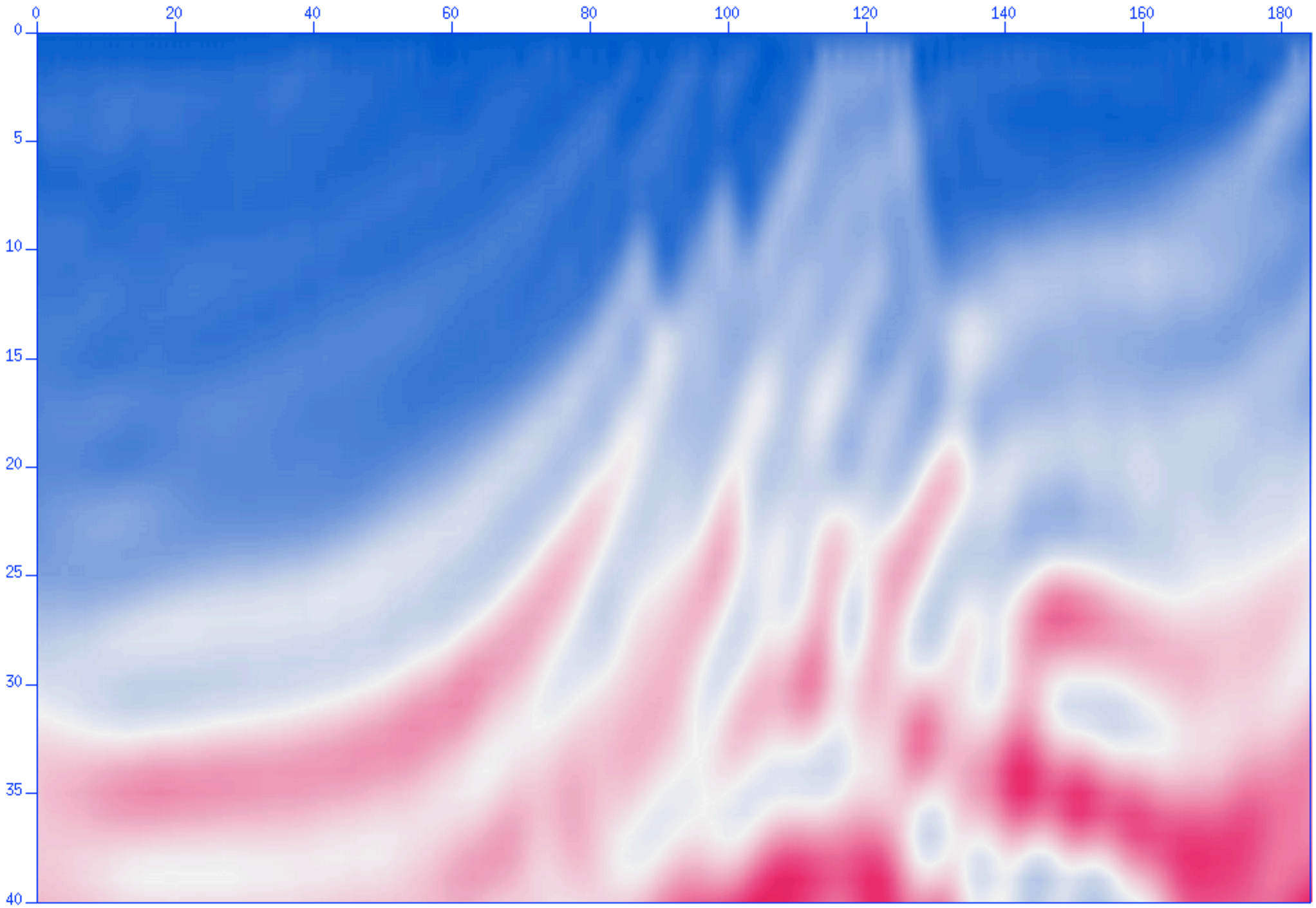


Initial Model, 4 Hz

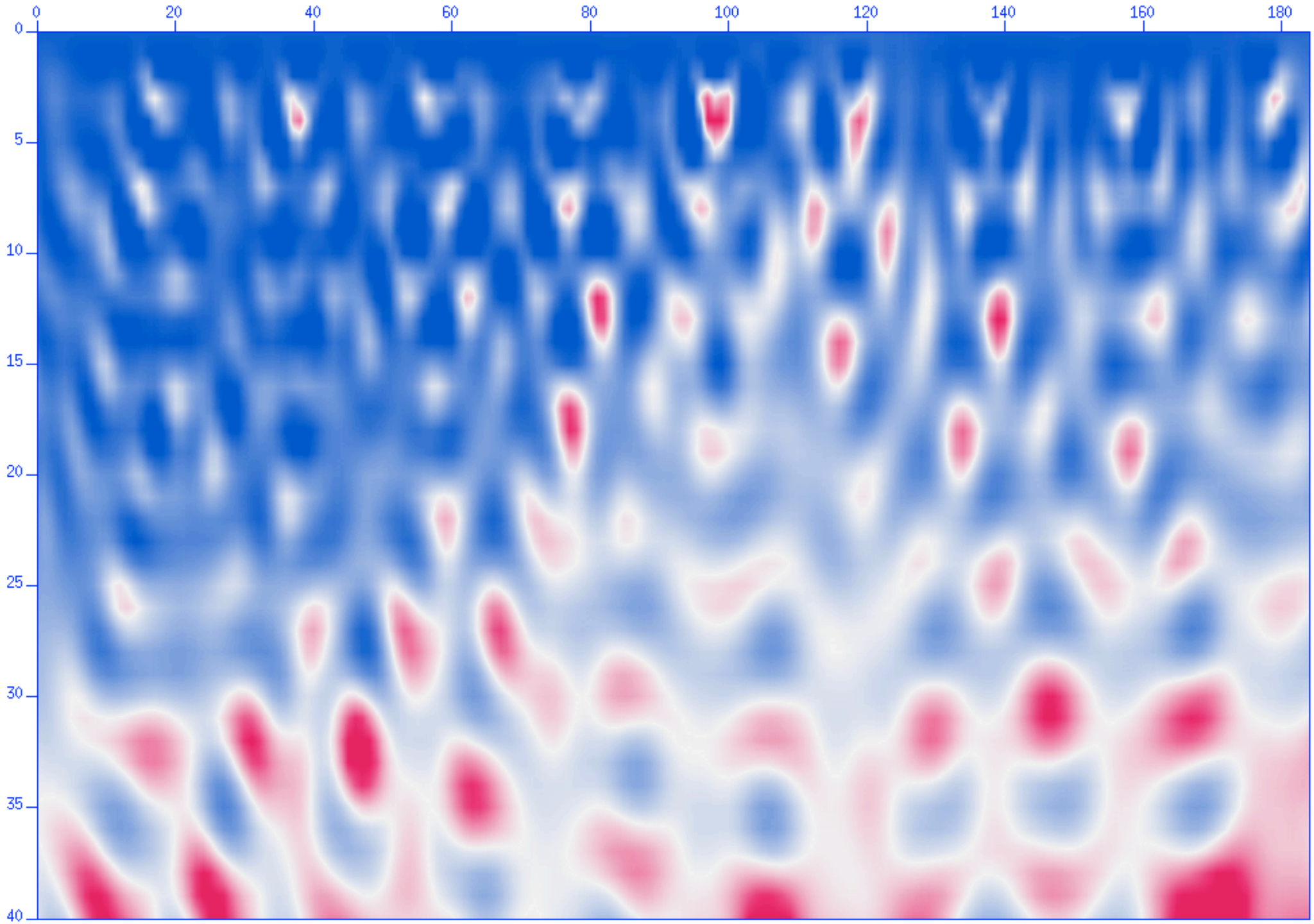


20% Corrupted Data

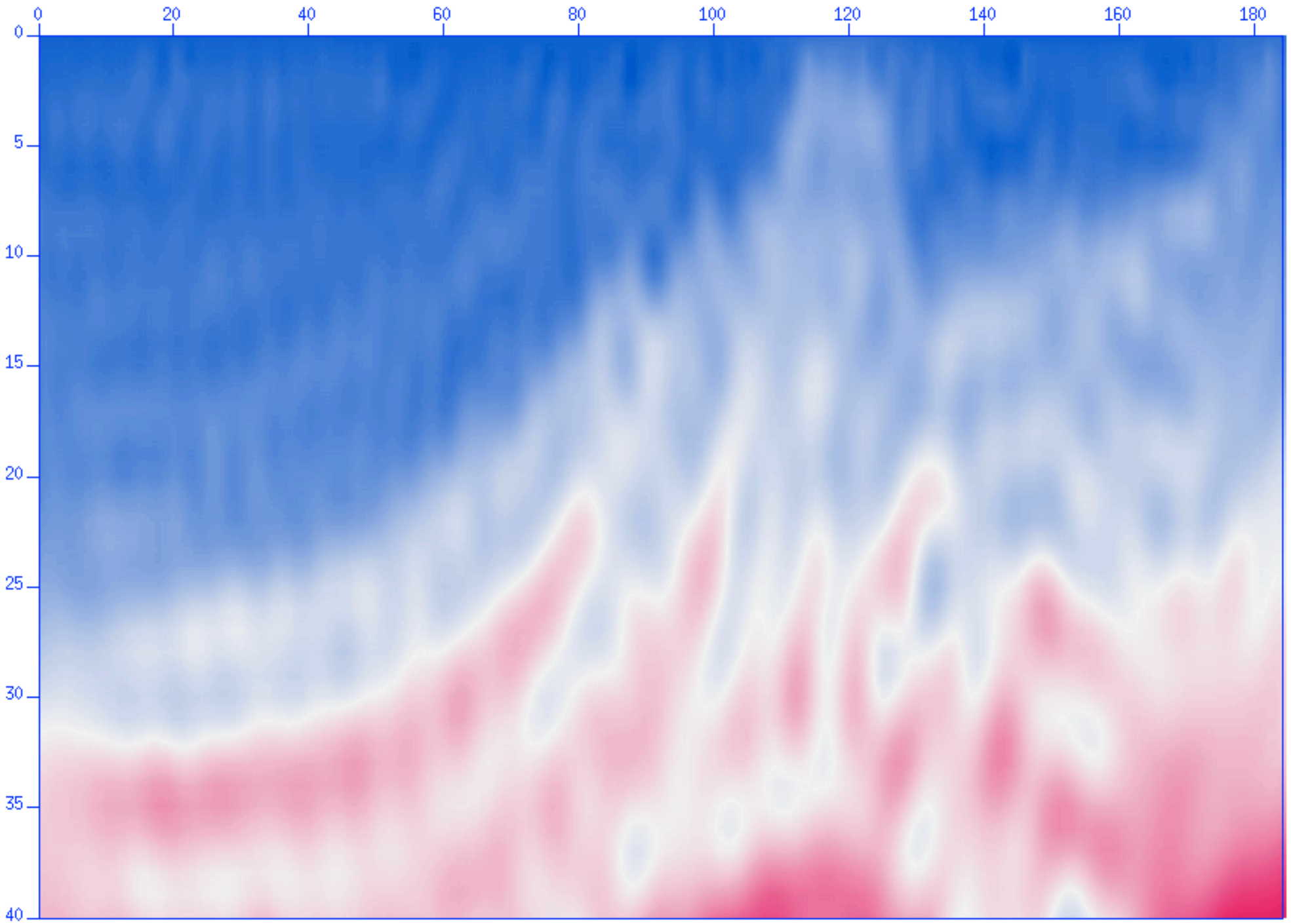
Results: Least Squares with GOOD data, 4Hz



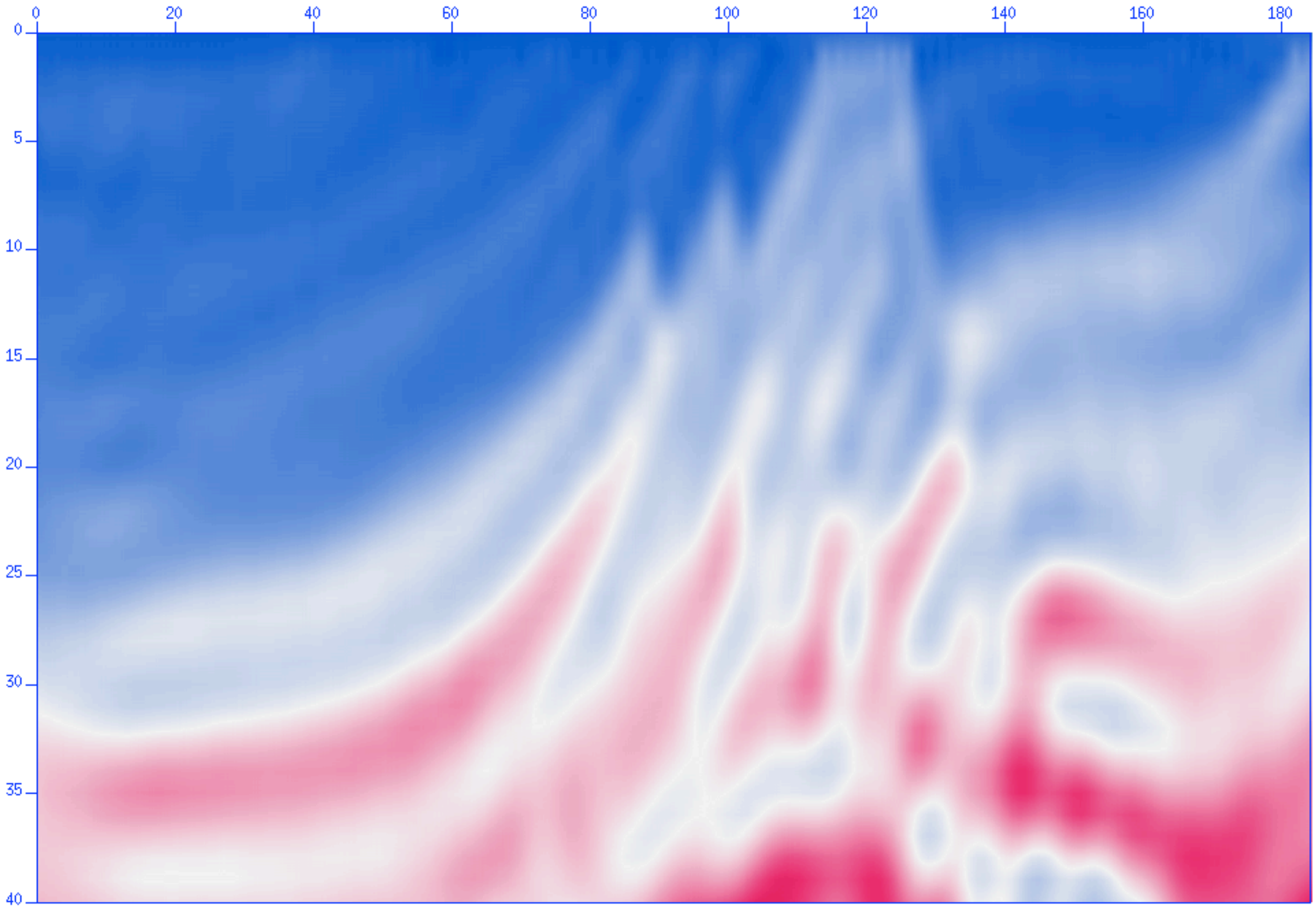
Results: Least Squares with BAD data, 4 Hz



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



Results: Least Squares with GOOD data



Variable Projection for Nuisance Parameters

- Many problems have secondary parameters, which, while not directly of interest, impact inversion for primary parameters:
  - Unknown variance/scale parameters in least squares inversion
  - Unknown source amplitudes
  - Student's t degree of freedom and scale parameters.

- The general problem can be formulated as follows:

$$\min_{\mathbf{m},\mathbf{x}} \mathbf{\Phi}(\mathbf{m},\mathbf{x})$$

- Here, **m** are the primary parameters, e.g. velocity model, and **x** are the (much smaller) nuisance parameters: |**x**|<<|**m**|.

Variable Projection for Nuisance Parameters

- Typically the overall function is solved with an iterative method. At each iteration, we propose updating **x** as follows:

$$\mathbf{x}^{\nu} = \arg\min_{\mathbf{x}} \mathbf{\Phi}(\mathbf{m}^{\nu}, \mathbf{x})$$

- This approach is equivalent to solving the reduced problem

$$\min_{\mathbf{m}} \tilde{\mathbf{\Phi}}(\mathbf{m}) := \mathbf{\Phi}(\mathbf{m}, \mathbf{x}(\mathbf{m}))$$

- Practically, you simply plug in updated **x** parameters into your favorite algorithm for the original problem with fixed **x**.

- Theoretically, you are guaranteed to converge to a local minimum of the full penalty while only working with the reduced objective.

Application I: Variance Estimation for LS

- We extend the FWI formulation to also fit for frequency-specific variances:

$$\min_{\mathbf{m},\boldsymbol{\sigma}} \quad \boldsymbol{\Phi}_{LS}(\mathbf{m},\boldsymbol{\sigma}) := \frac{1}{2}\sum_{j=1}^{n} -\log(\sigma_j) + \frac{\|\mathbf{r_j}(\mathbf{m})\|^2}{\sigma_j^2}$$

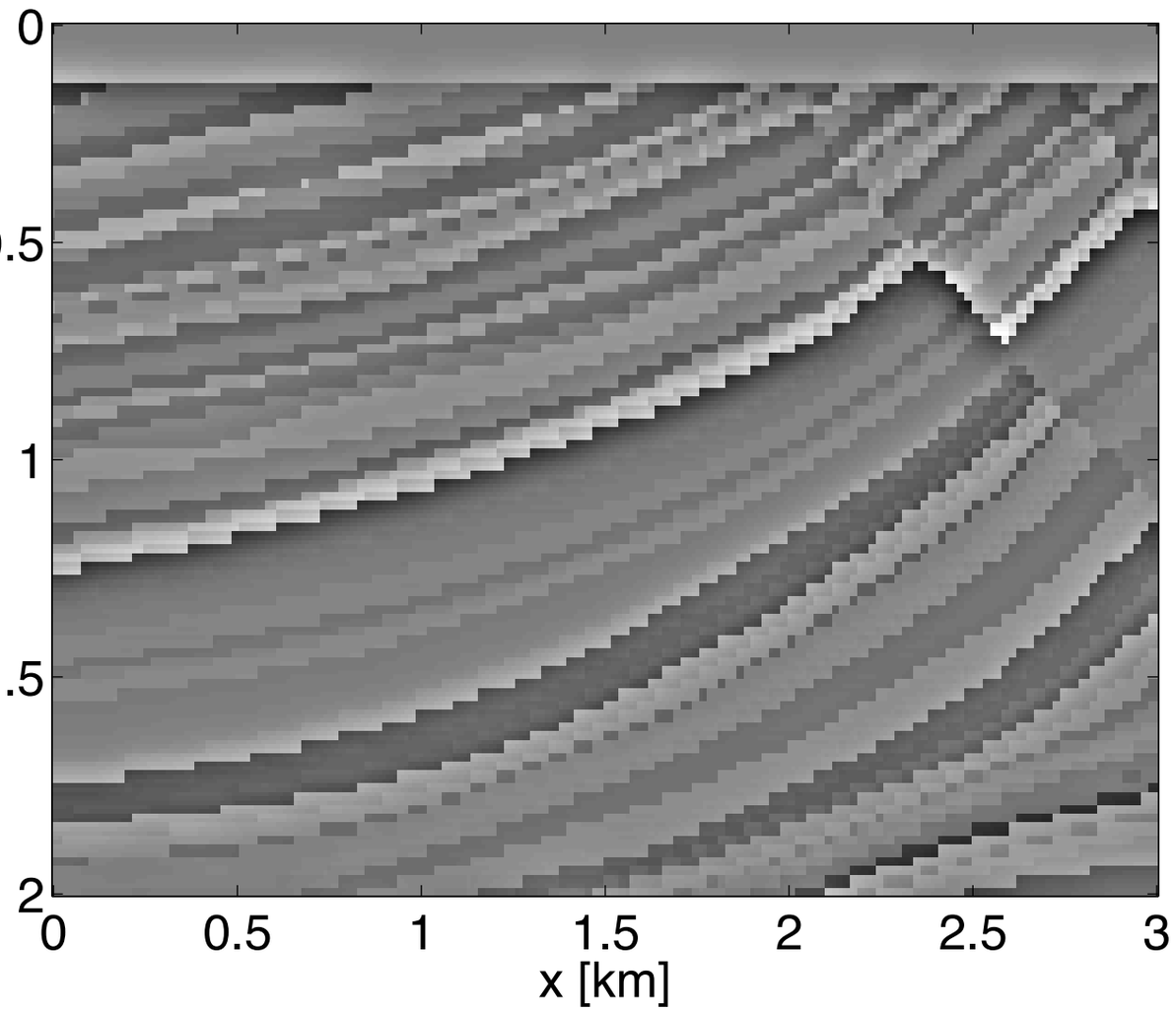
- Our algorithm takes the following form:

$$\mathbf{m}^{\nu+1} = \mathbf{m}^{\nu} - \alpha_{\nu}\sum_{j=1}^{n}\frac{1}{(\sigma_j^{\nu})^2}(\nabla \mathbf{r_j}^{\nu})^T\mathbf{r_j}^{\nu}$$

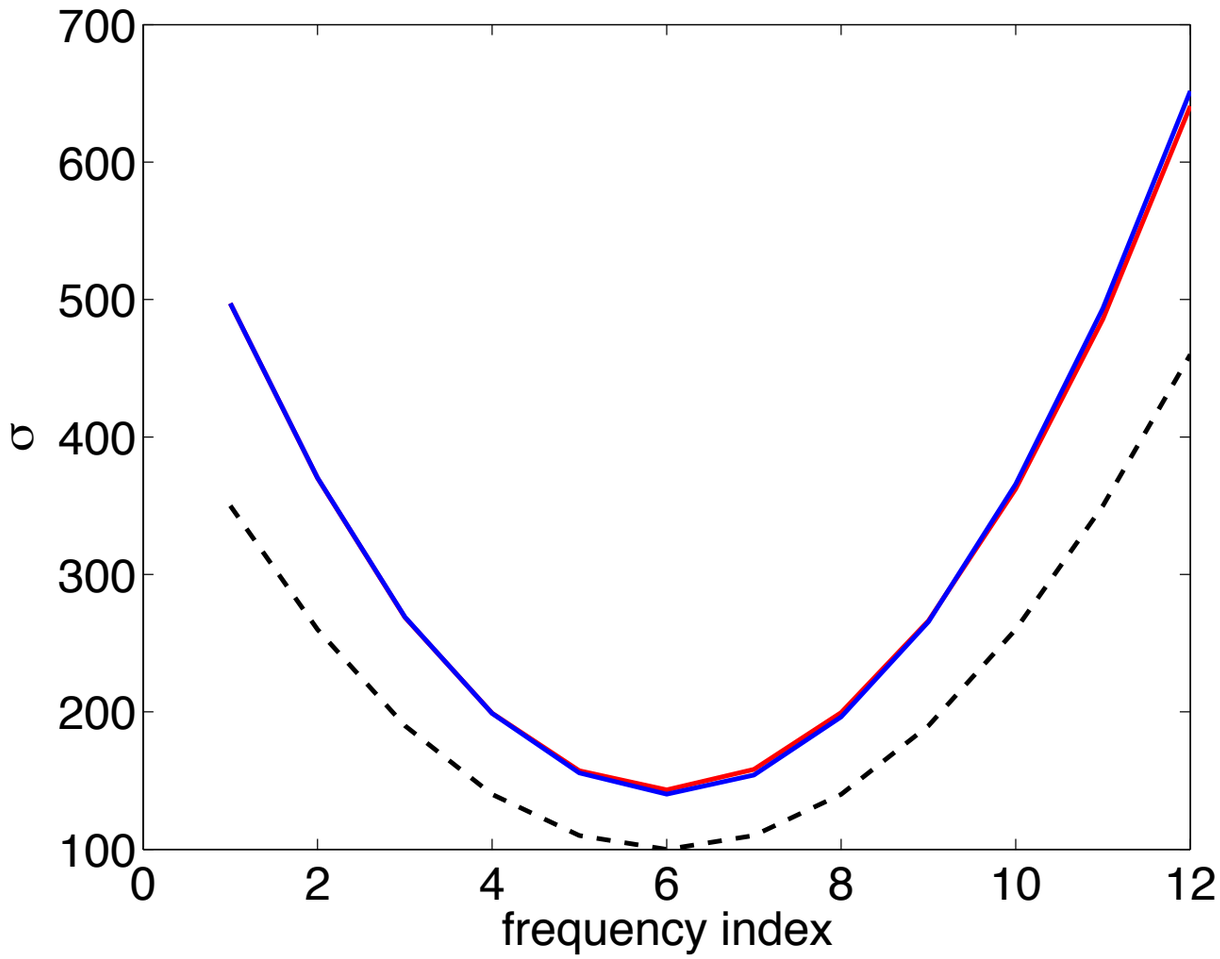
$$\sigma_j^{\nu} = \mathrm{var}(\mathbf{r_j})$$

- The usual workflow is trivially modified, introducing frequency-specific weights that are easily computed.

Experiment: frequency dependent noise

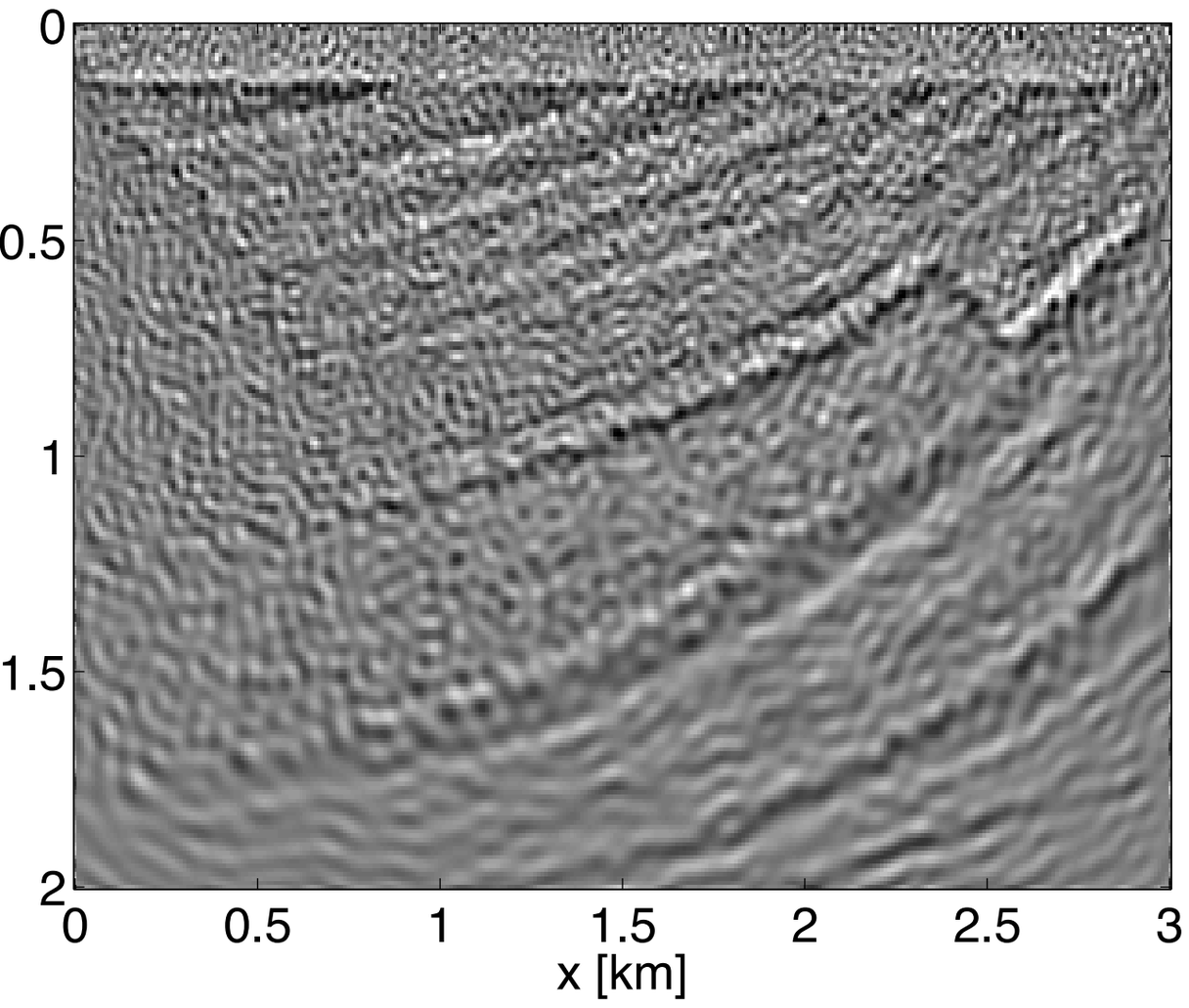


True Perturbation

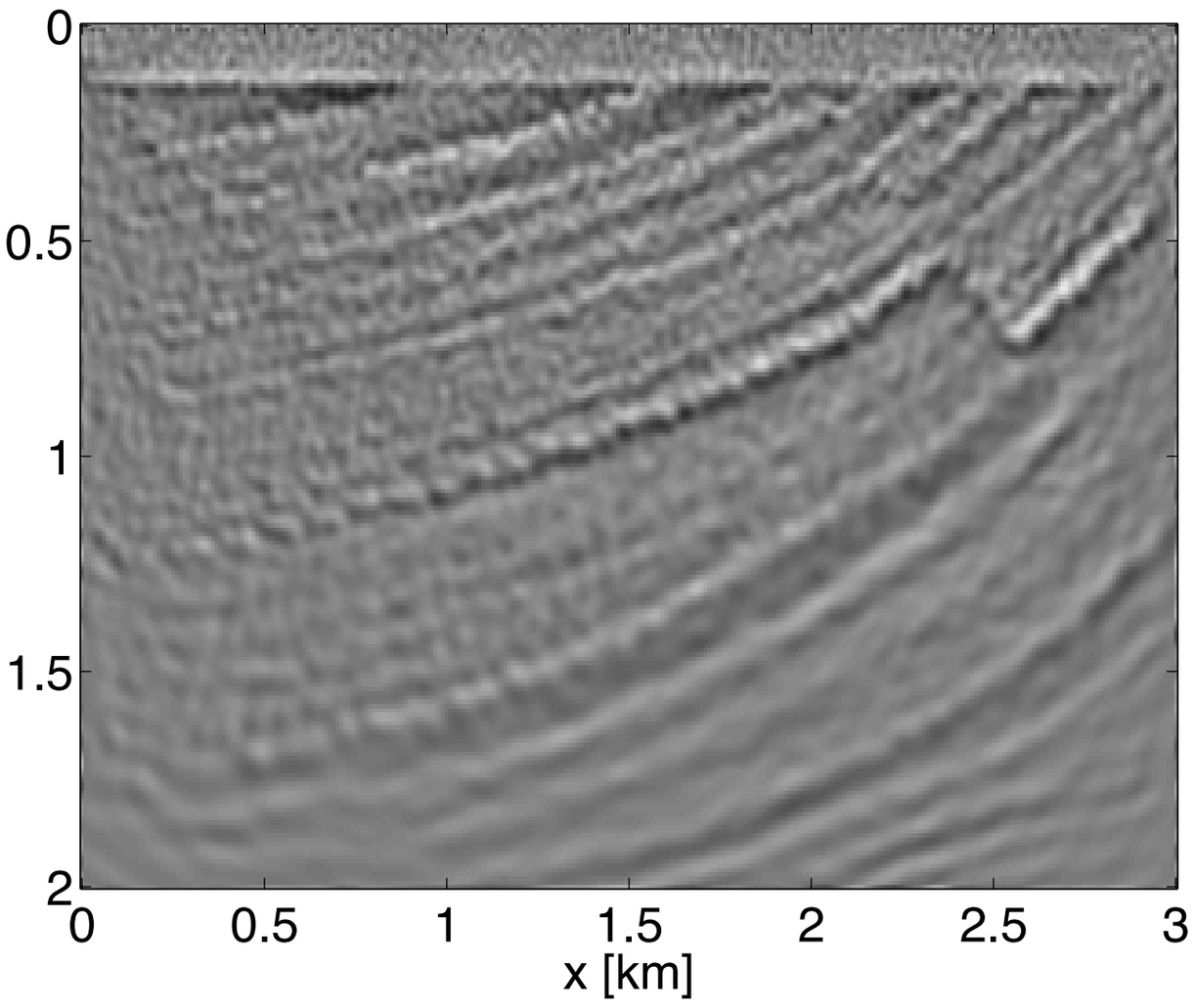


Freq. dependent noise

Results: FWI with/without variance estimation

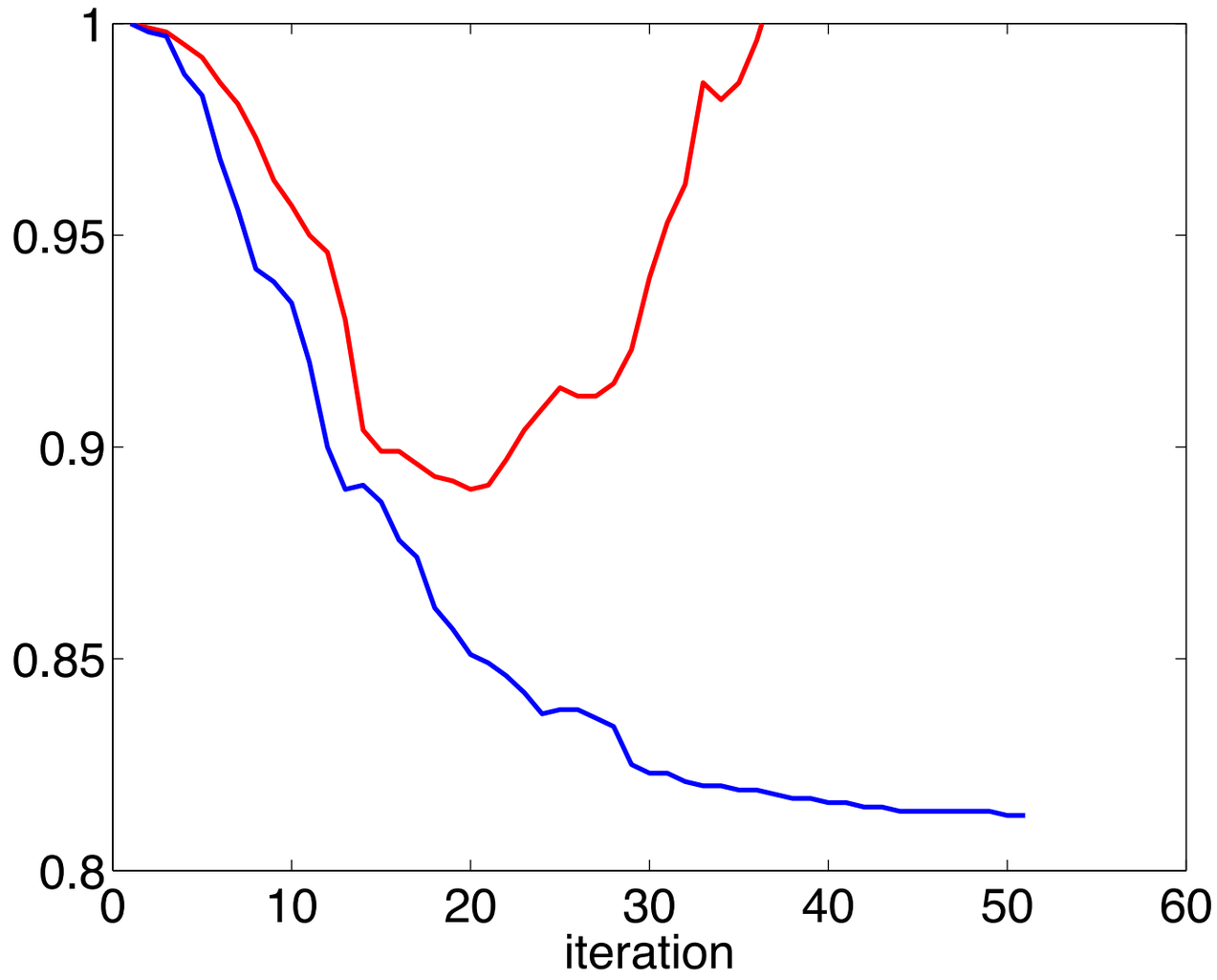


WITHOUT VAR. EST.

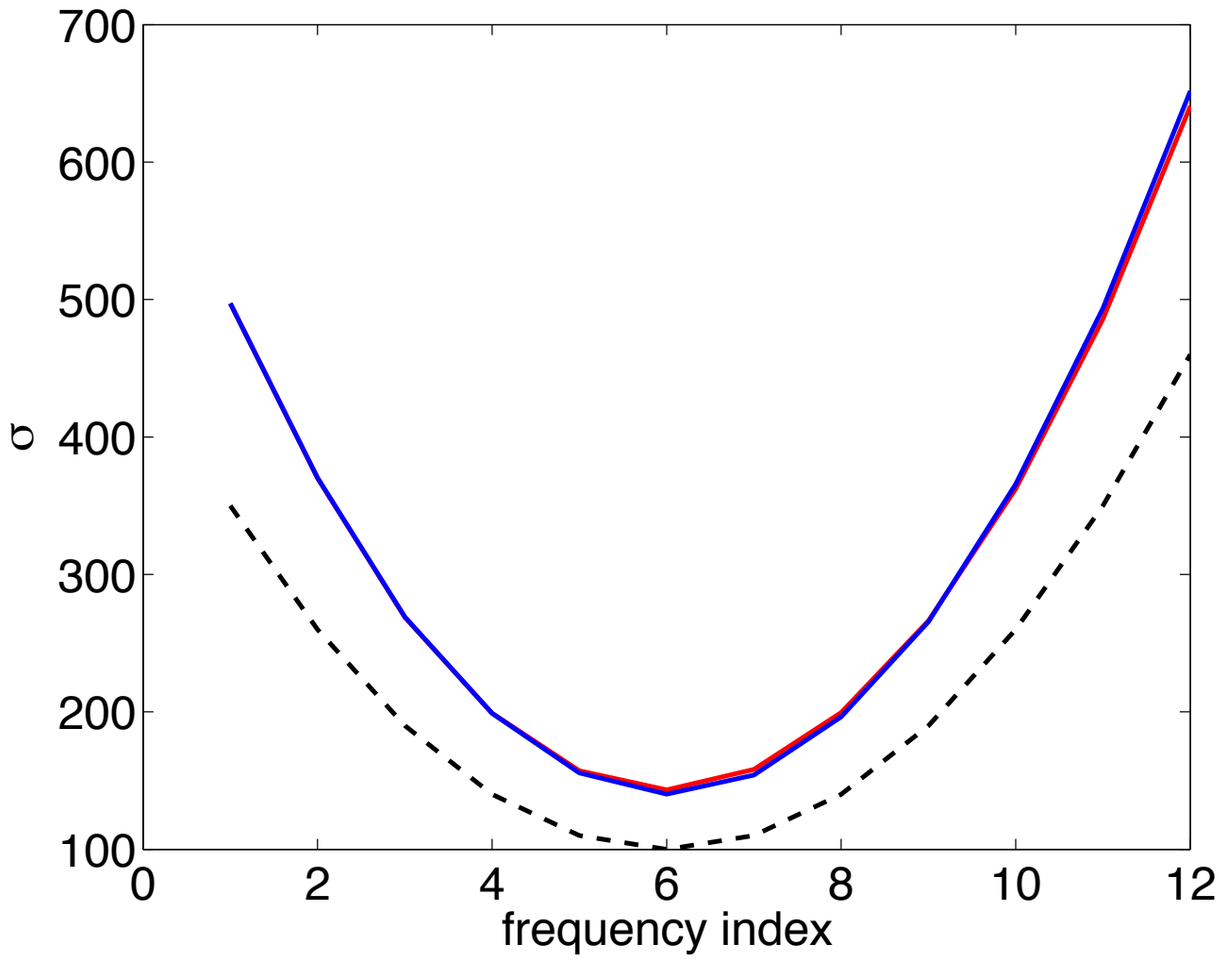


WITH VAR. EST.

Results: FWI with/without variance estimation



Rel. Model Error



Fitted Variances

Application II: Robust Source Estimation

- 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)

- The source amplitudes are the nuisance (**x**) parameters here.

Application of VP to Source Estimation

- Amplitude parameter estimation at each iteration takes the form

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

- These problems separate completely, and for each amplitude we can use Newton's method. We present two cases:
  - Least Squares objective: $\alpha_i = \dfrac{\mathbf{d_i^T}(\boldsymbol{\mathcal{F}}_\mathbf{i}(\mathbf{m})\mathbf{q_i})}{\|\boldsymbol{\mathcal{F}}_\mathbf{i}(\mathbf{m})\mathbf{q_i})\|^2}$
  - Student's t Objective:

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

- Scalar Newton-type method for Student's t source estimation!

Robust Source Estimation: Results

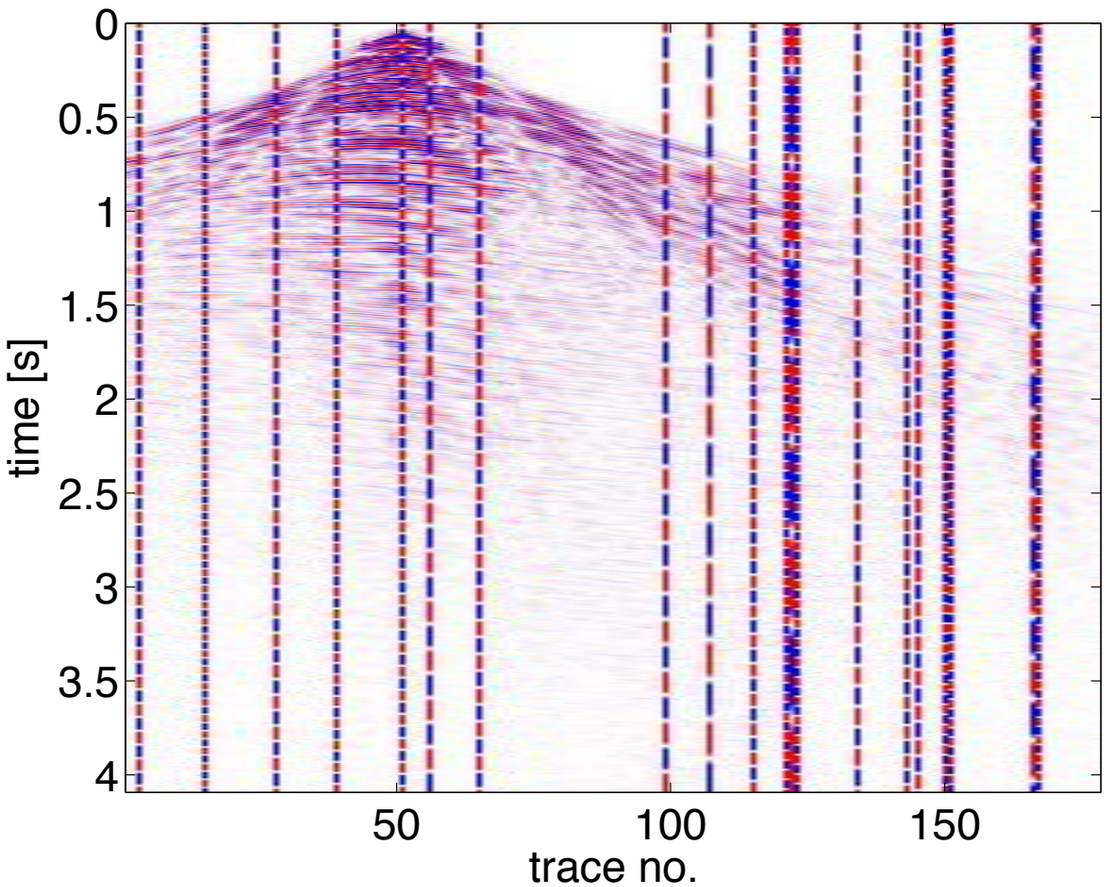


Figure 1 *Data with outliers in the form of bad traces.*

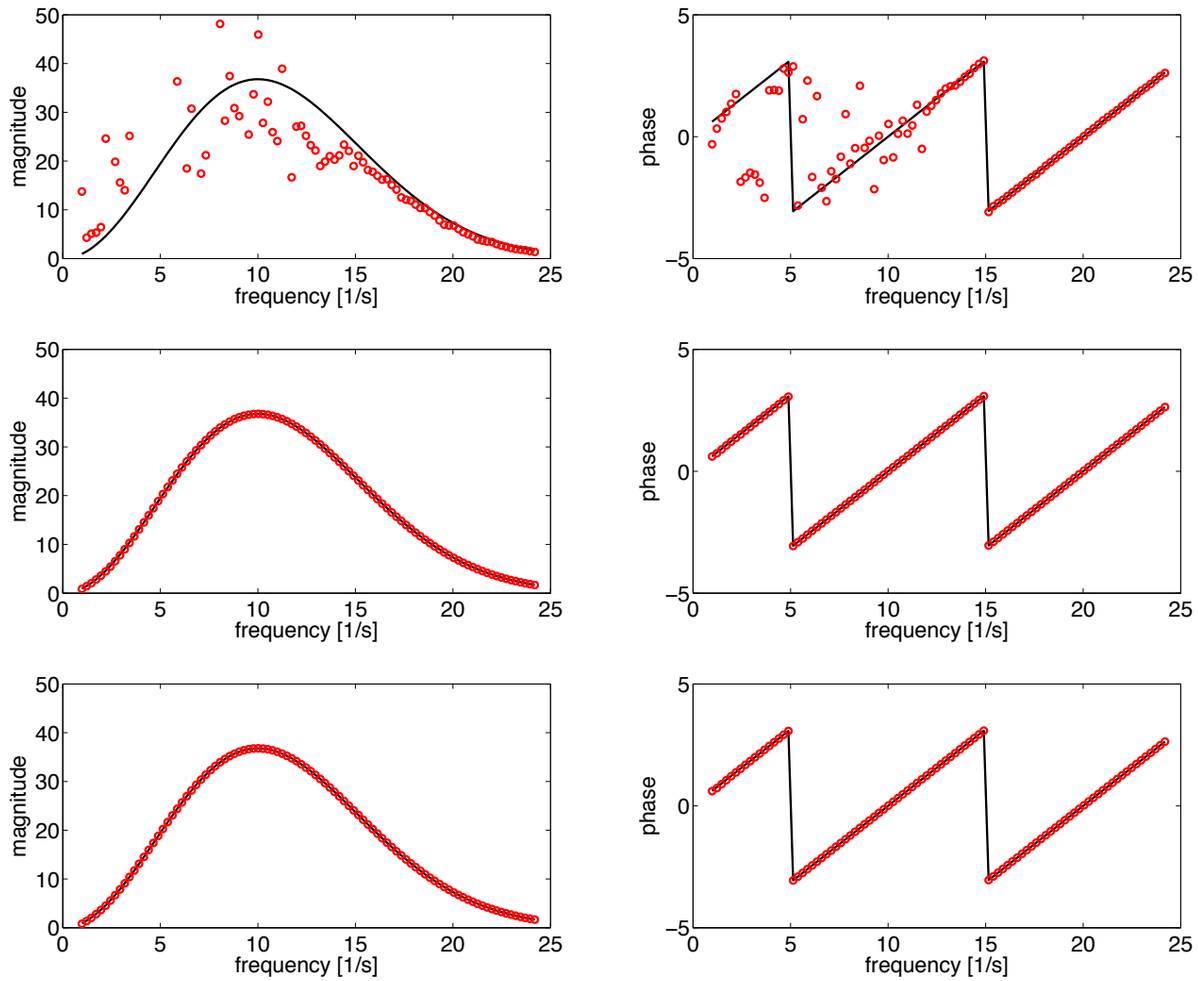
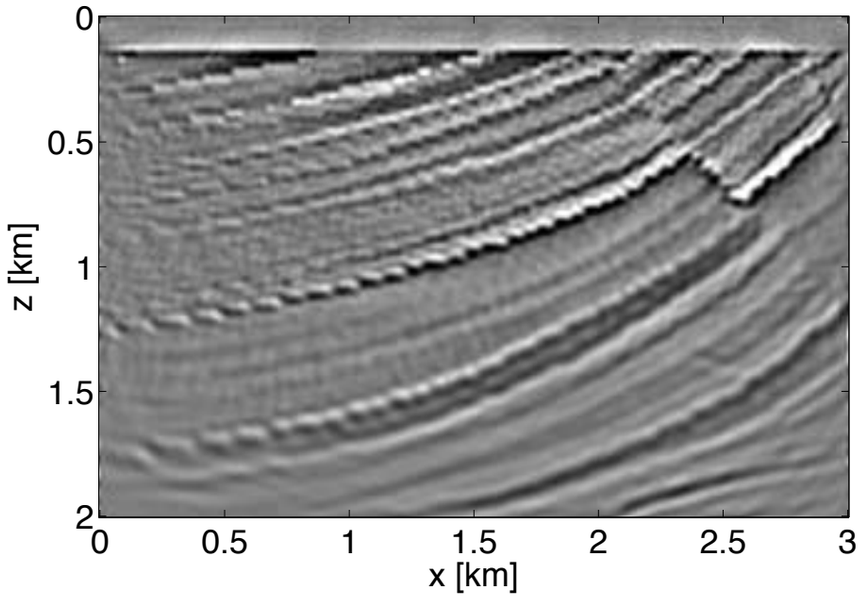
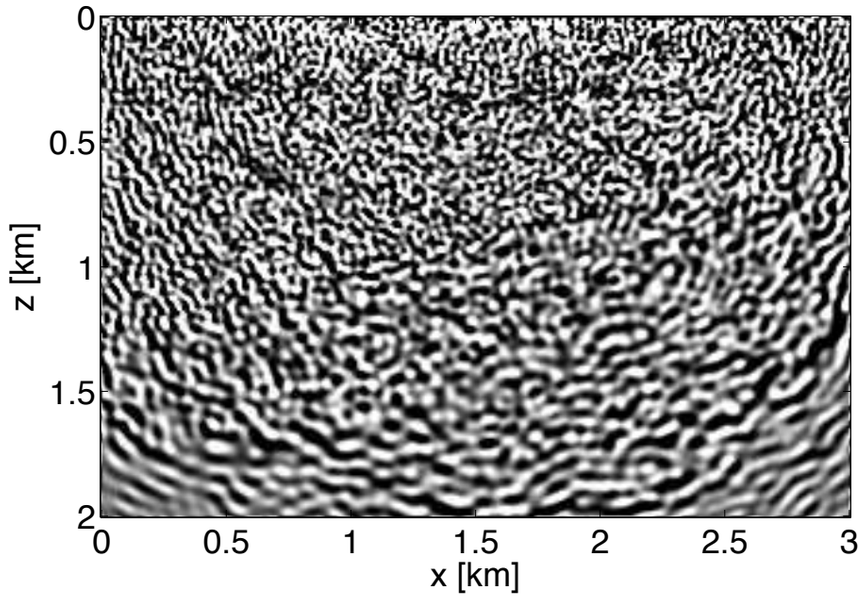


Figure 2 *Estimated source wavelet using Least-Squares (top), Hybrid (middle) and Students t (bottom) approaches.*

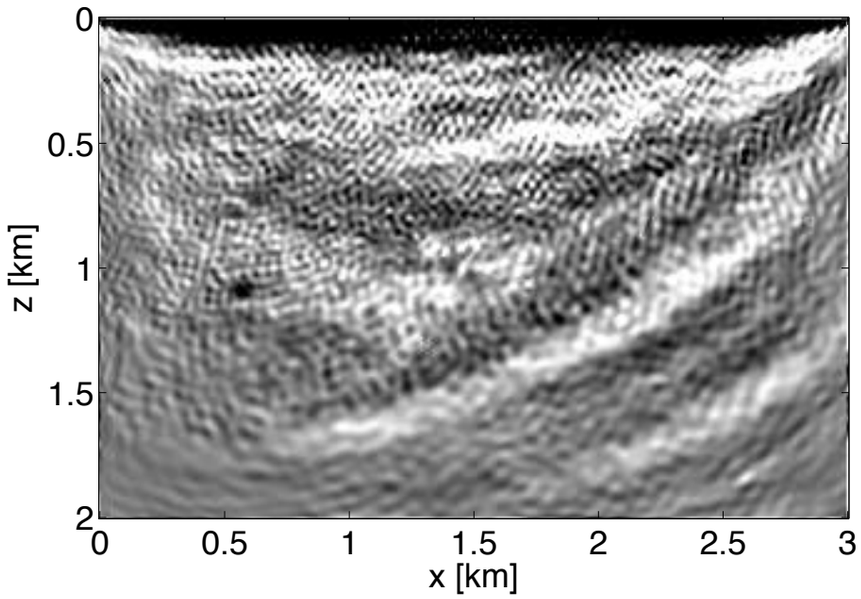
Robust Source Estimation: Results



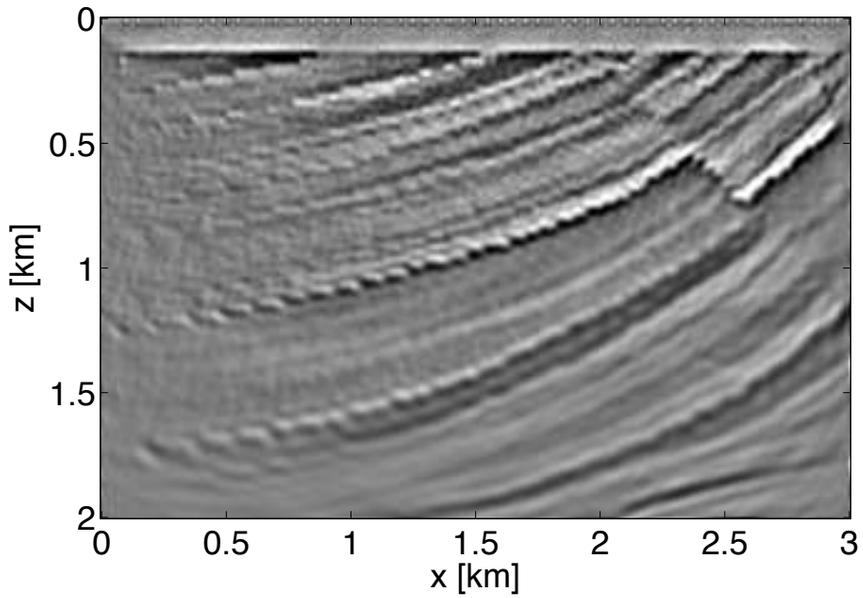
LS-LS w/o noise



LS-LS w noise



ST-LS w noise



ST-ST w noise

Application III: Student's t d.f. estimation

- The performance of Student's t depends on the scale and degrees of freedom parameters.

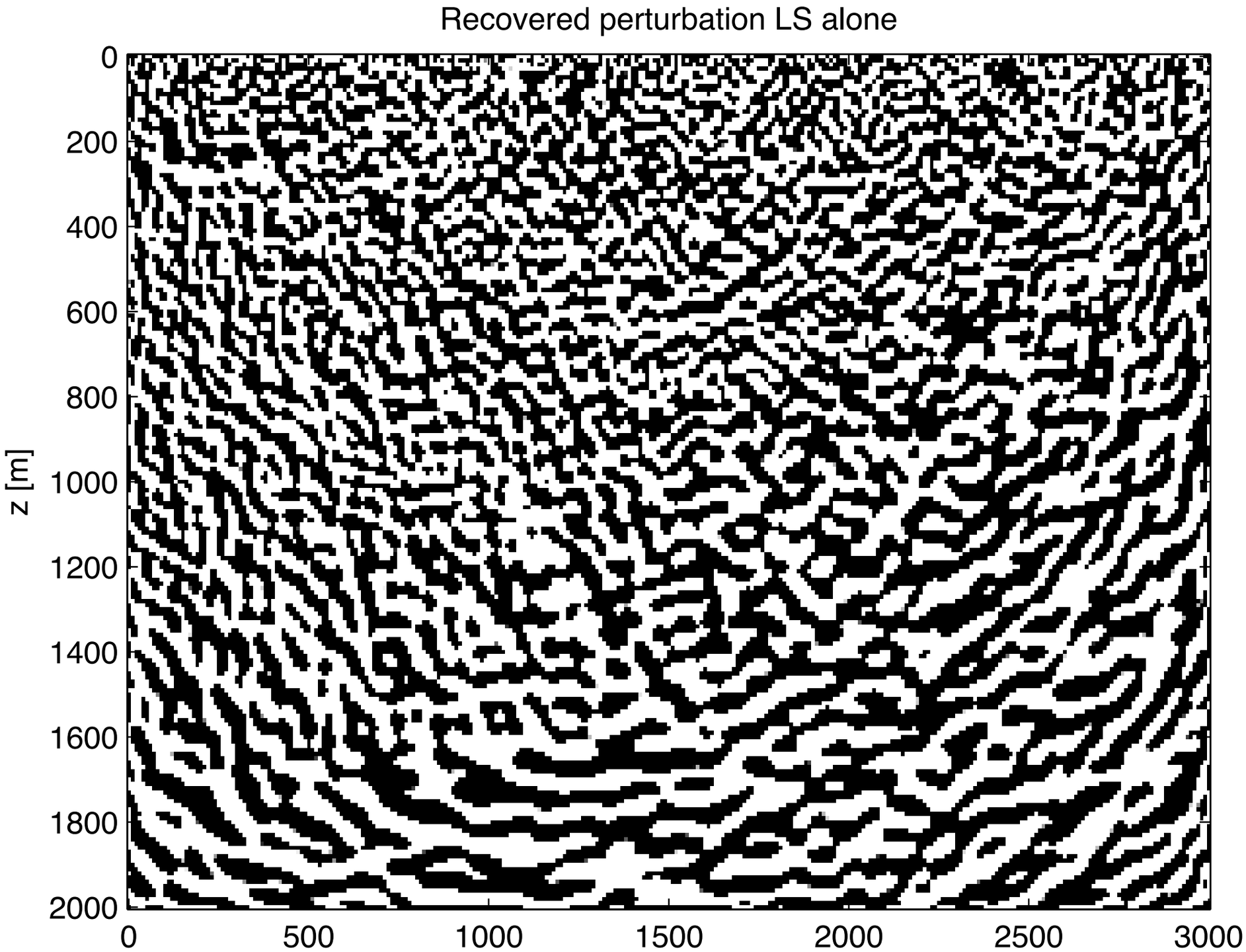
- Thus far, we have been simplifying the problem by using scale = 1 and empirically picking the d.f. parameter.

- To design an automated method, recall first the Student's t density:

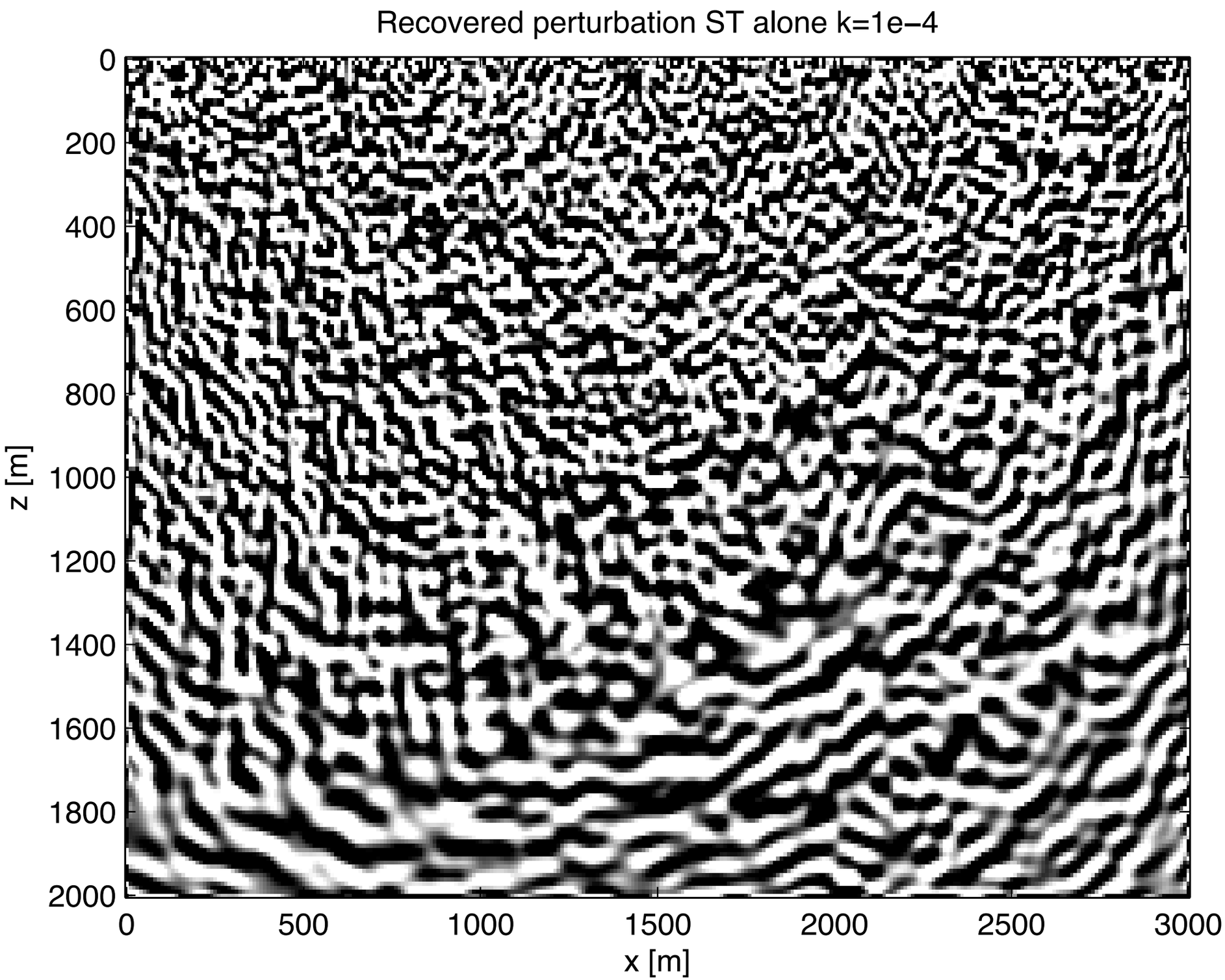
$$\mathbf{p}(\epsilon|\mu,\sigma,k) \quad = \quad \frac{\Gamma(\frac{k+1}{2})}{\sigma\Gamma(\frac{k}{2})\sqrt{\pi k}}\left(1+\frac{(\epsilon-\mu)^2}{k\sigma^2}\right)^{\frac{-(k+1)}{2}}$$

- We begin by showing how the 'effective scale' $k\sigma^2$ affects recovery.

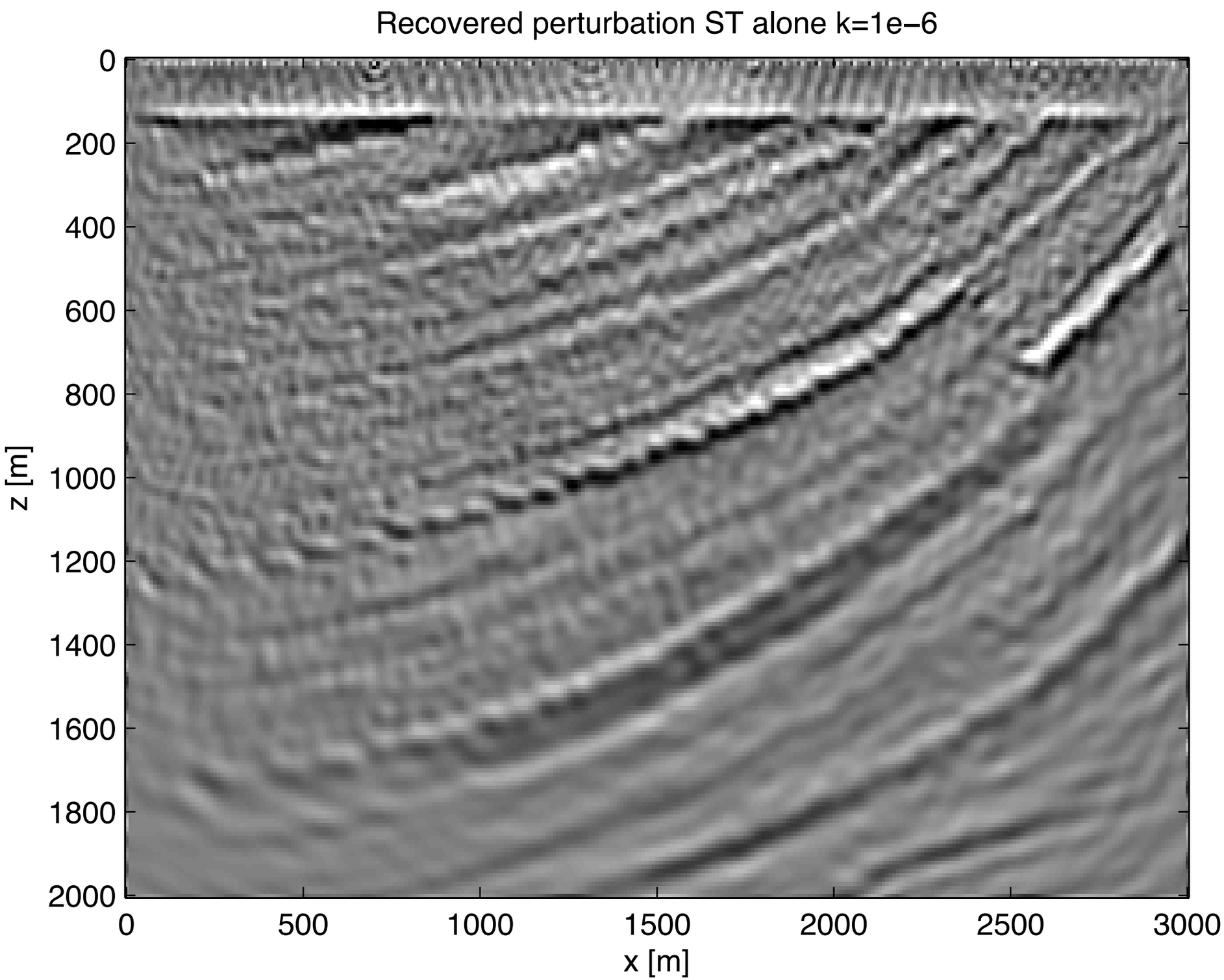
LS for outlier problem:



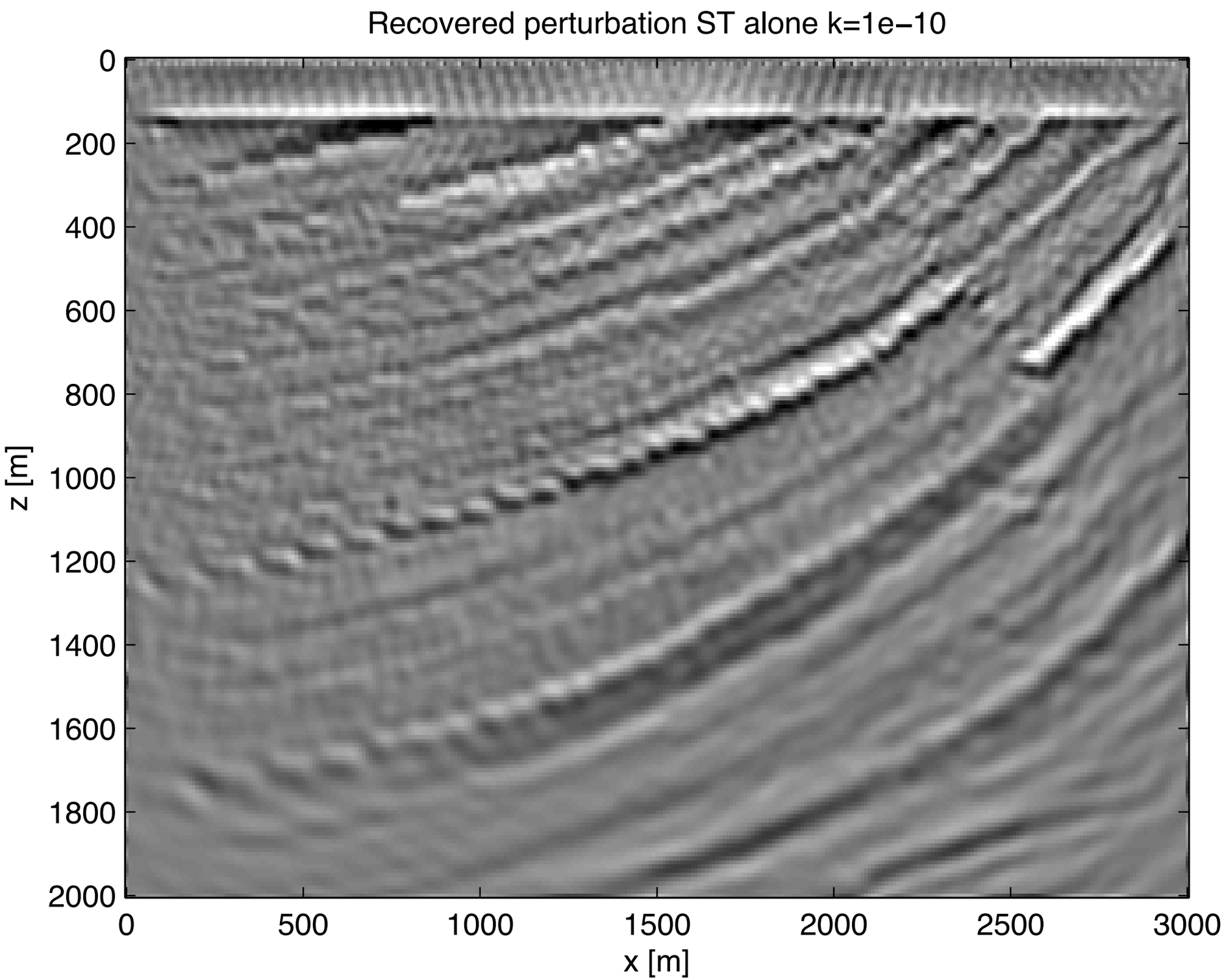
Student’s t with ‘effective scale’ = 1e-4:



Student’s t with ‘effective scale’ = 1e-6:



Student’s t with ‘effective scale’ = 1e-10:



Student's t d.f. estimation

- To develop an automated method, we formulate an extended objective that includes the scale and d.f. as nuisance parameters:

$$\min_{\mathbf{m},k,\sigma^2} -n\log\left(\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\right) + \frac{n}{2}\log(\sigma^2) + \frac{k+1}{2}\sum_{i=1}^{n}\log\left(1+\frac{\mathbf{r_i}^2}{\sigma^2 k}\right)$$

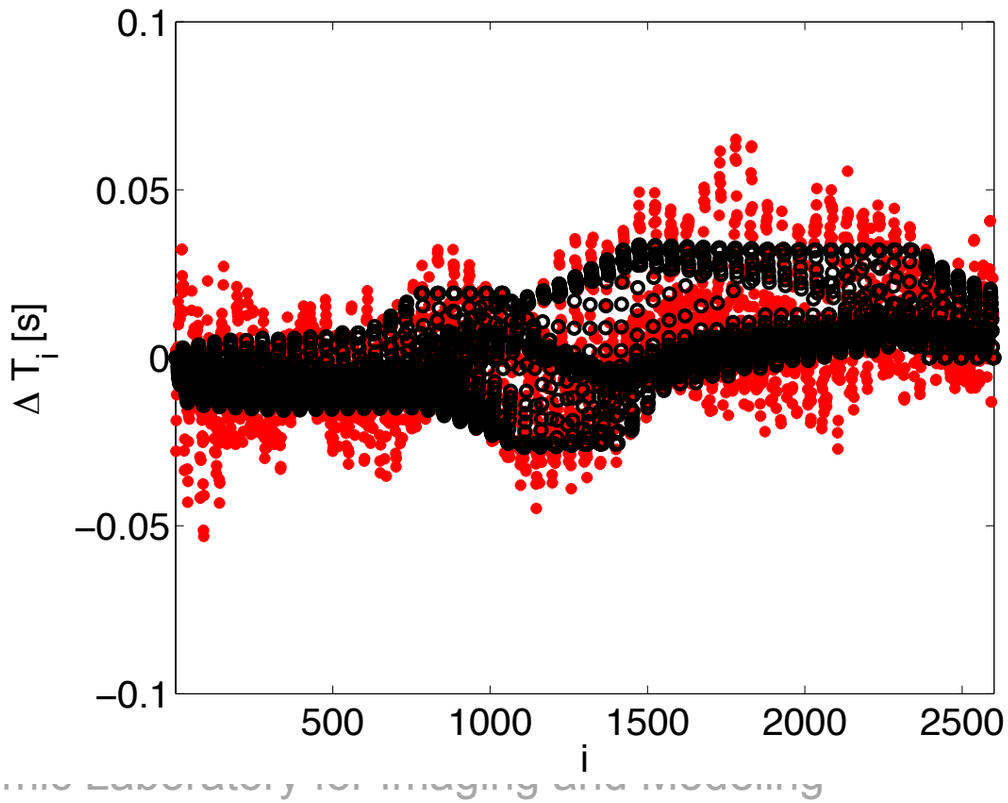
- At each iteration, we minimize over the nuisance parameters with **m** held constant. That's a 2d optimization problem for every frequency!

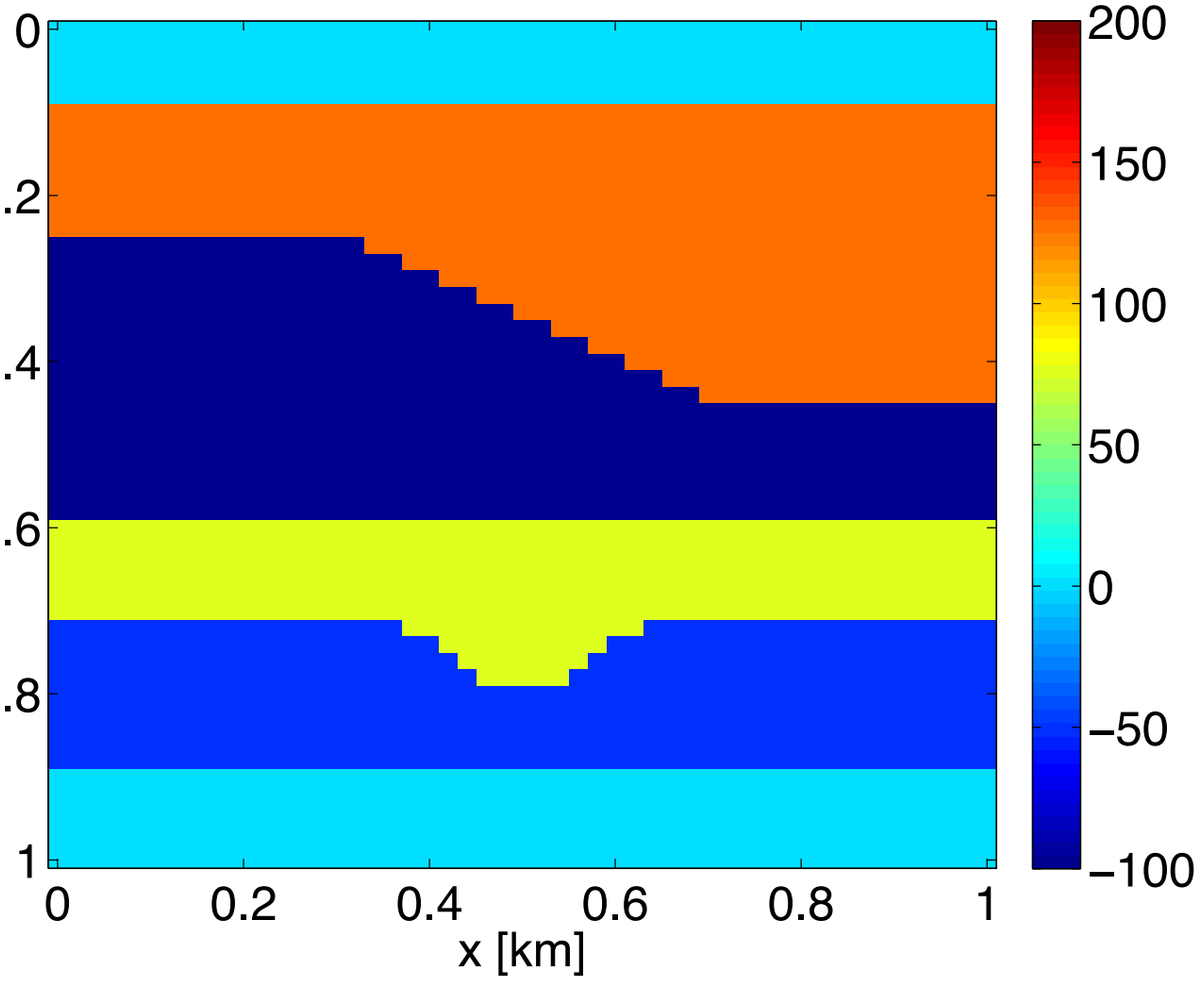
- Estimating scale and degrees of freedom parameters has been addressed in the statistical literature, but this is not how they do it. We think this is better.

Scale and d.f. Estimation Example

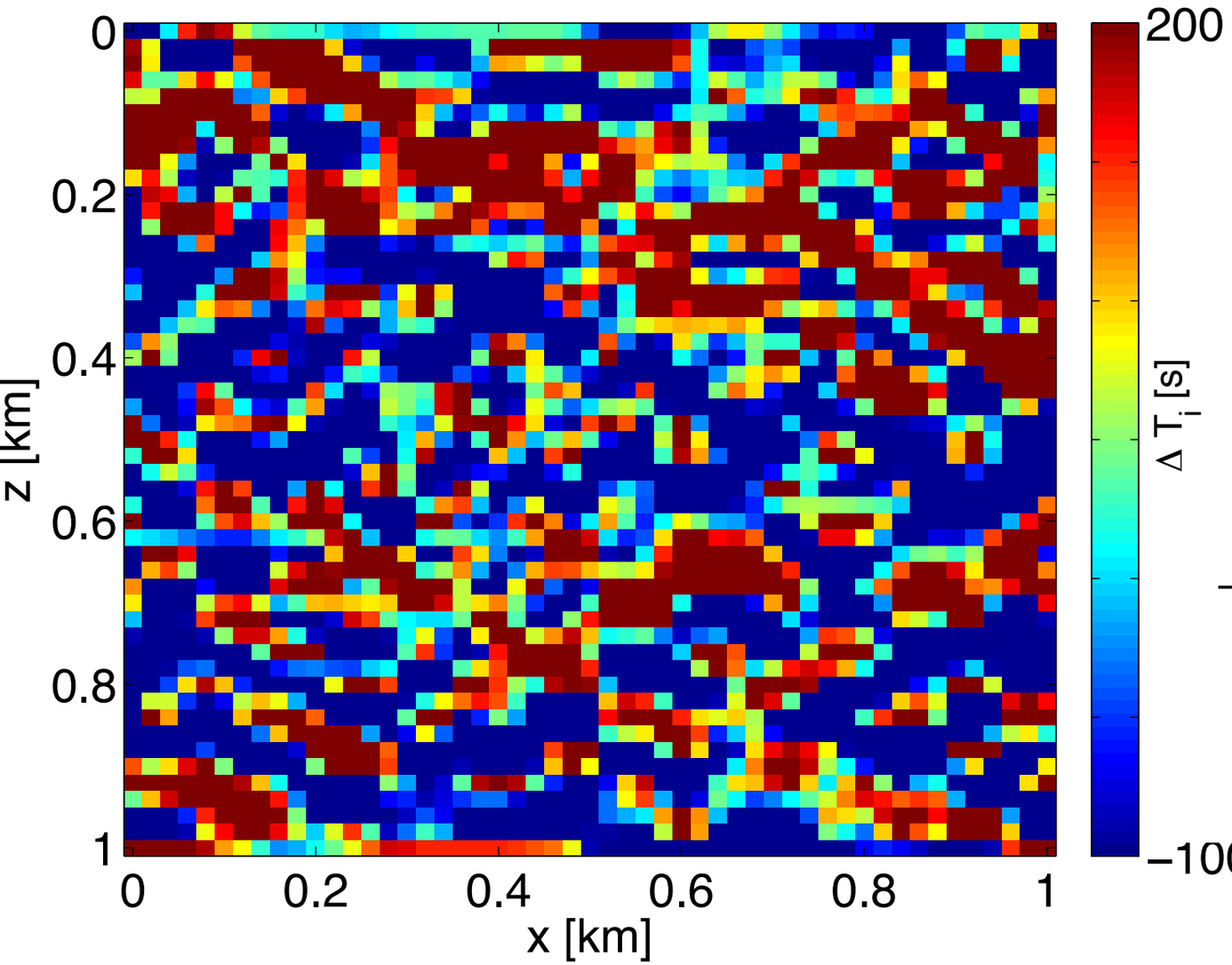
- We set up a traveltime tomography experiment:
  - constant background velocity
  - parameters of interest: perturbation of background velocity
  - Optimization problem:

$$\min_{\mathbf{m},k,\sigma^2} \rho_{\mathbf{m},k}(\Delta T - A\mathbf{m})$$

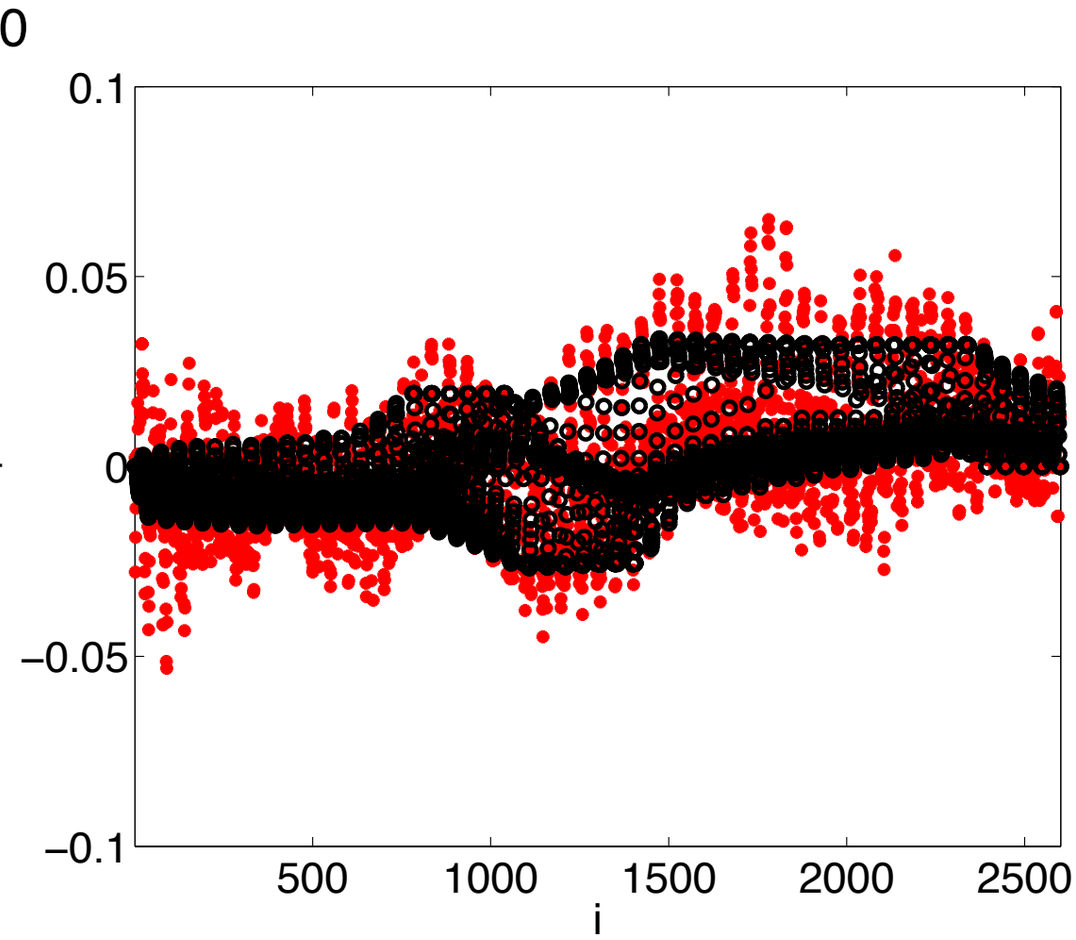




Results: LS Estimation

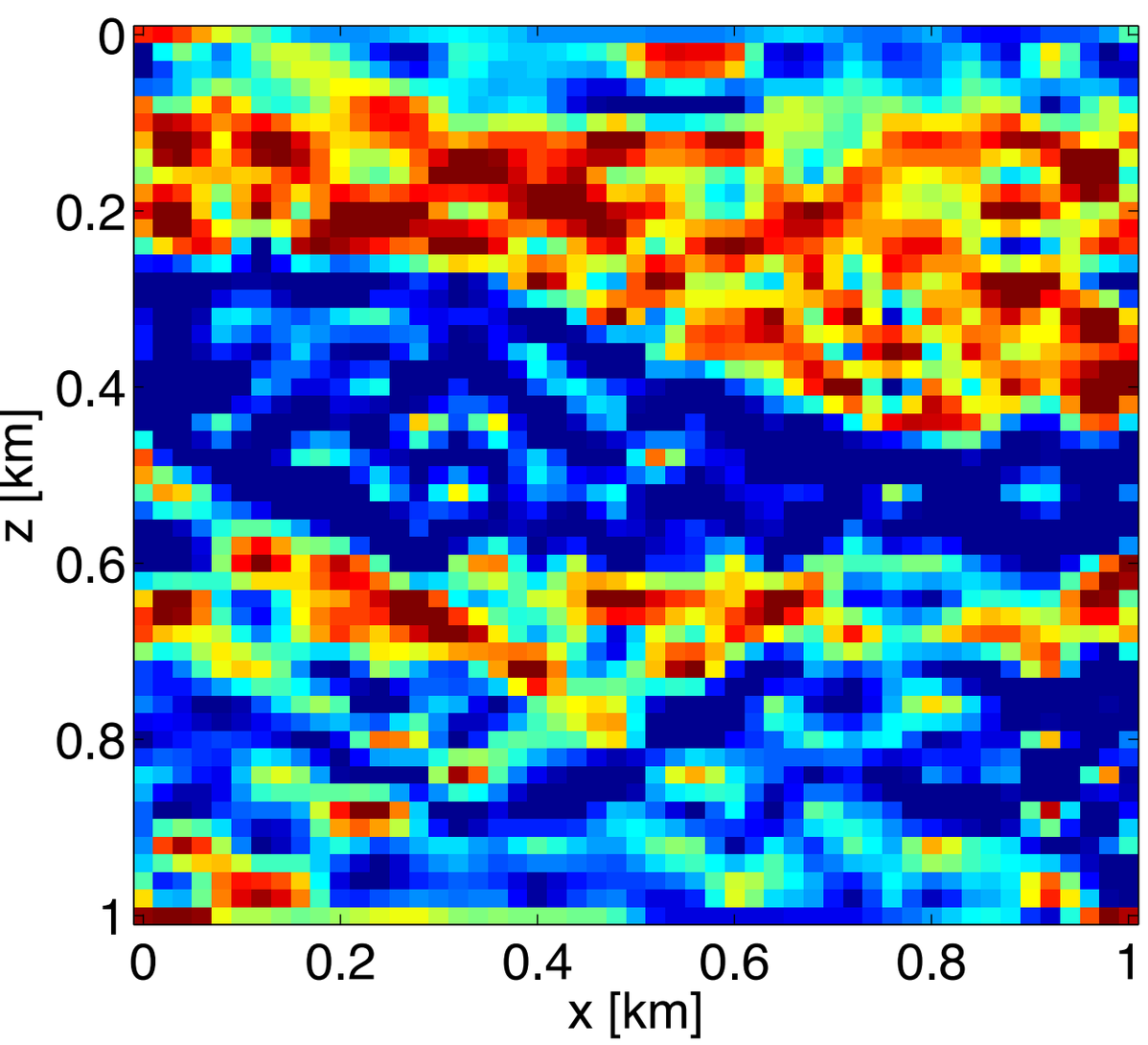


Recovery

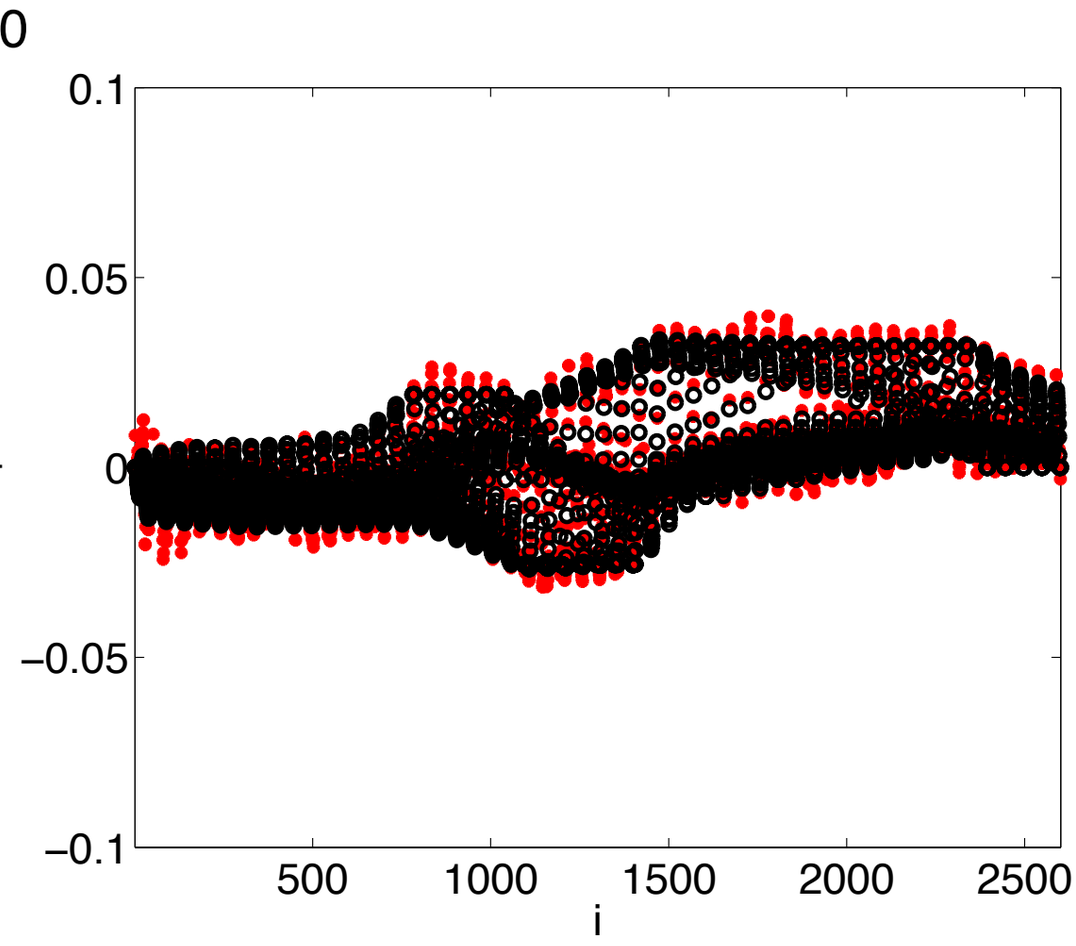


True and Predicted Data

Results: ST estimation, fixed DF and scale

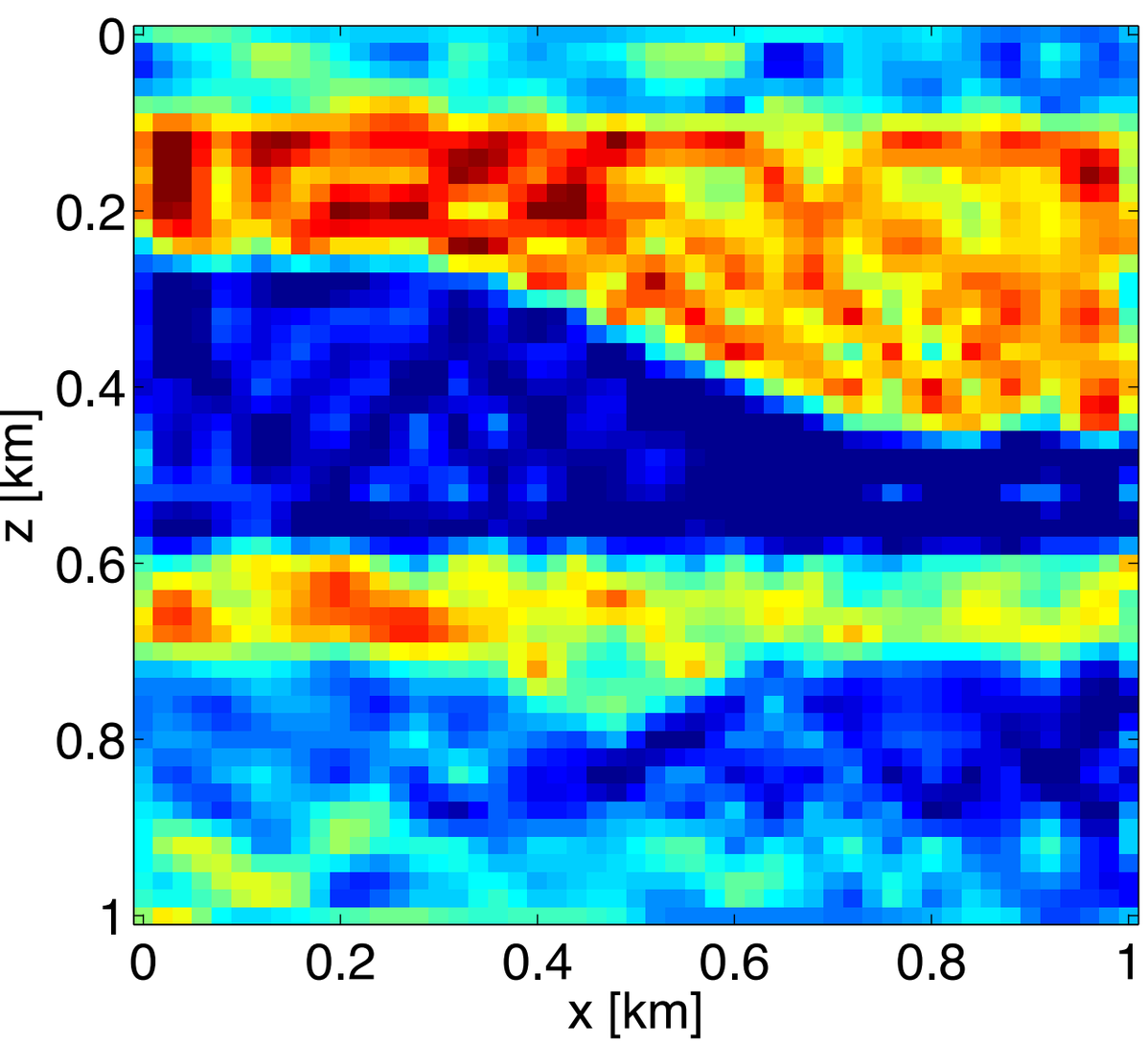


Recovery

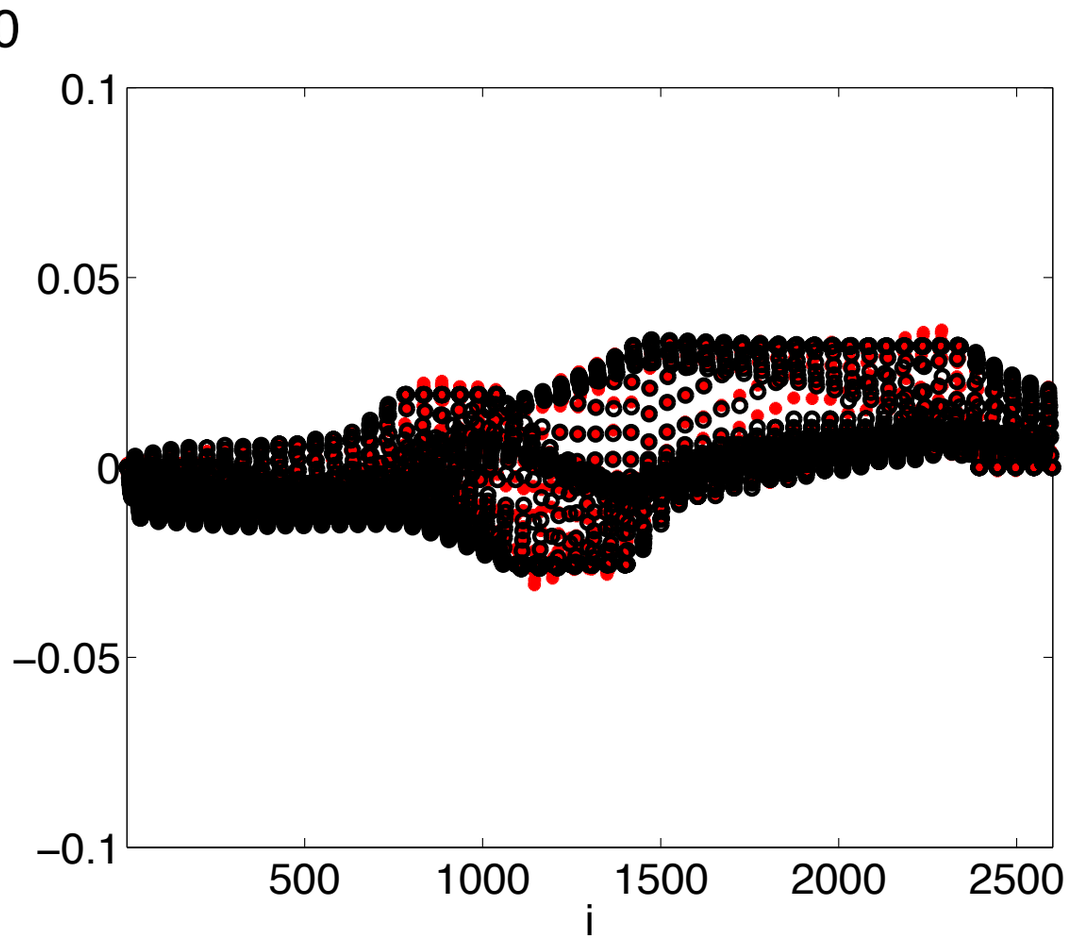


True and Predicted Data

Results: ST estimation, estimated DF and scale



Recovery



True and Predicted Data

Robustness and Sparsity: Review of SPGL1

- SPGL1 is typically known for solving the problem

$$\min_x \|x\|_1 \text{ s.t. } \|Ax - b\|_2 \leq \sigma$$

- The crucial pieces that make it work are
  - Fast projection onto

$$\mathbb{B}_1^\tau = \{x : \|x\|_1 \leq \tau\}$$

  - Simple dual norm:

$$\|\cdot\|_1^* = \|\cdot\|_\infty$$

  - Smooth residual function.

- SPGL1 uses a Newton root finding scheme for the Lasso value function

$$v(\tau) = \min_x \|Ax - b\|_2 \text{ s.t. } \|x\|_1 \leq \tau$$

Root Finding: $v(\tau) = \sigma$

Approximately solve

minimize $\quad \frac{1}{2}\|Ax - b\|_2^2$

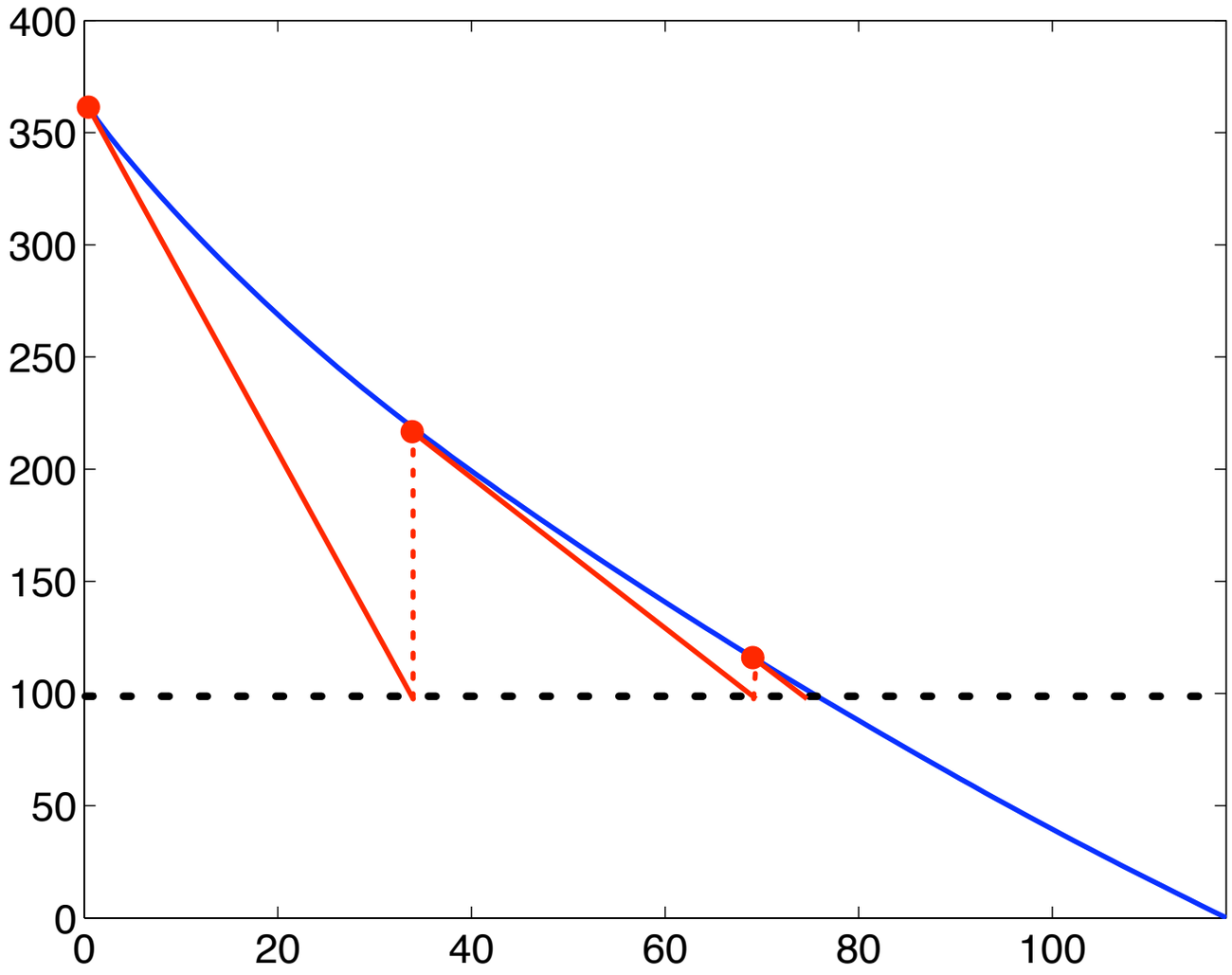
subj to $\quad \|x\|_1 \leq \tau_k$

Newton update

$\tau_{k+1} \leftarrow \tau_k - (v_k - \sigma)/v_k'$

Early termination

monitor duality gap



Generalized SPGL1:

$$\min_x \|x\| \text{ s.t. } h(b - F(x)) \leq \sigma$$

- The crucial pieces that make it work are
  - Fast projection onto $\mathbb{B}^\tau = \{x : \|x\| \leq \tau\}$
  - Simple dual norm: $\|\cdot\|^*$
  - Smooth residual function $h(\cdot)$

- Theorem (A.A, Burke, Friedlander): if h is convex and smooth, F is linear, then

$$v'(\tau) = -\|F^T \nabla h(b - F\bar{x})\|^*$$

- If h is NOT convex, or F not linear, we will just use the formula anyway.

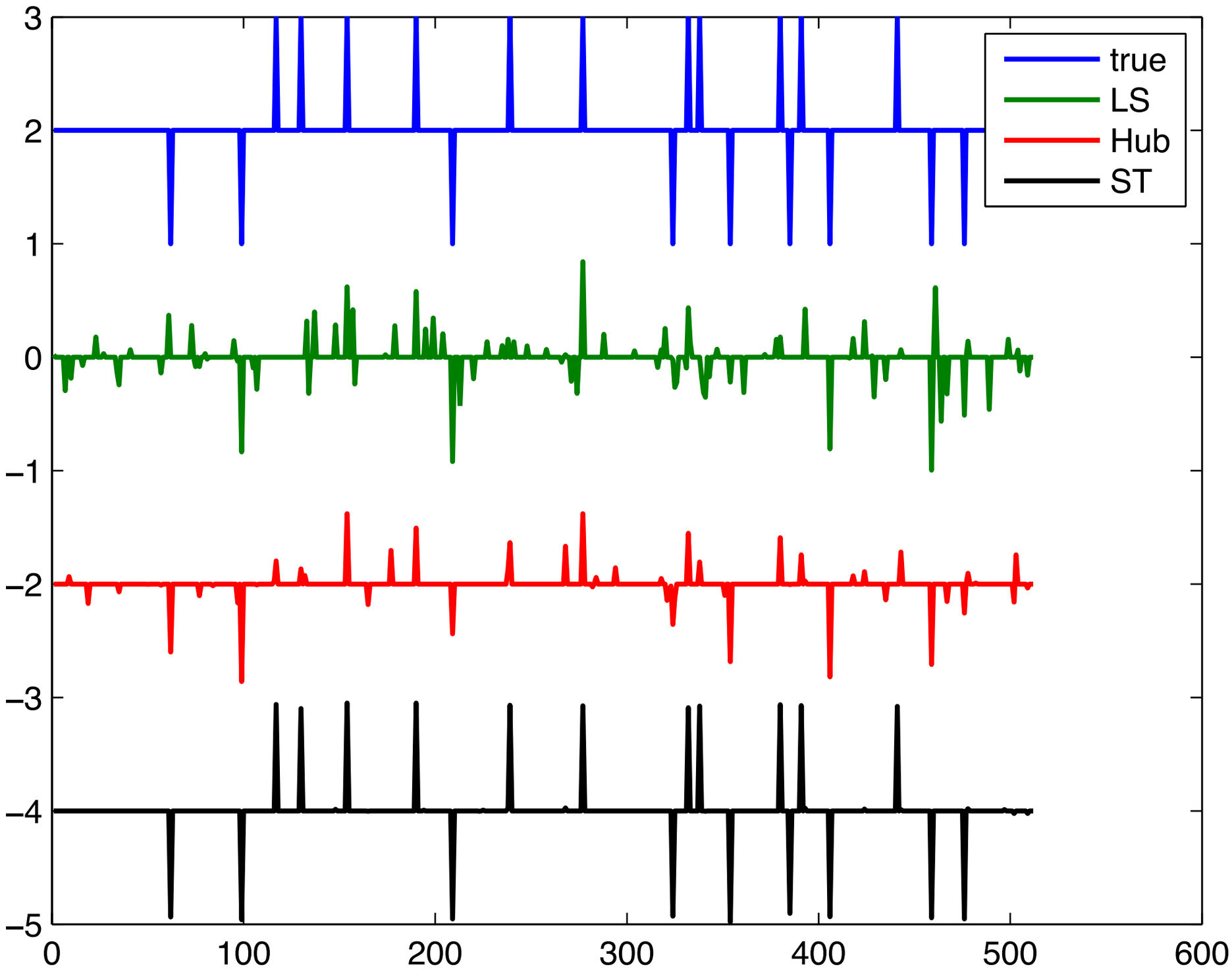
Example 1: Sparse + Robust Recovery

- We play with a Compressive Sensing example, contaminating residuals with outliers.

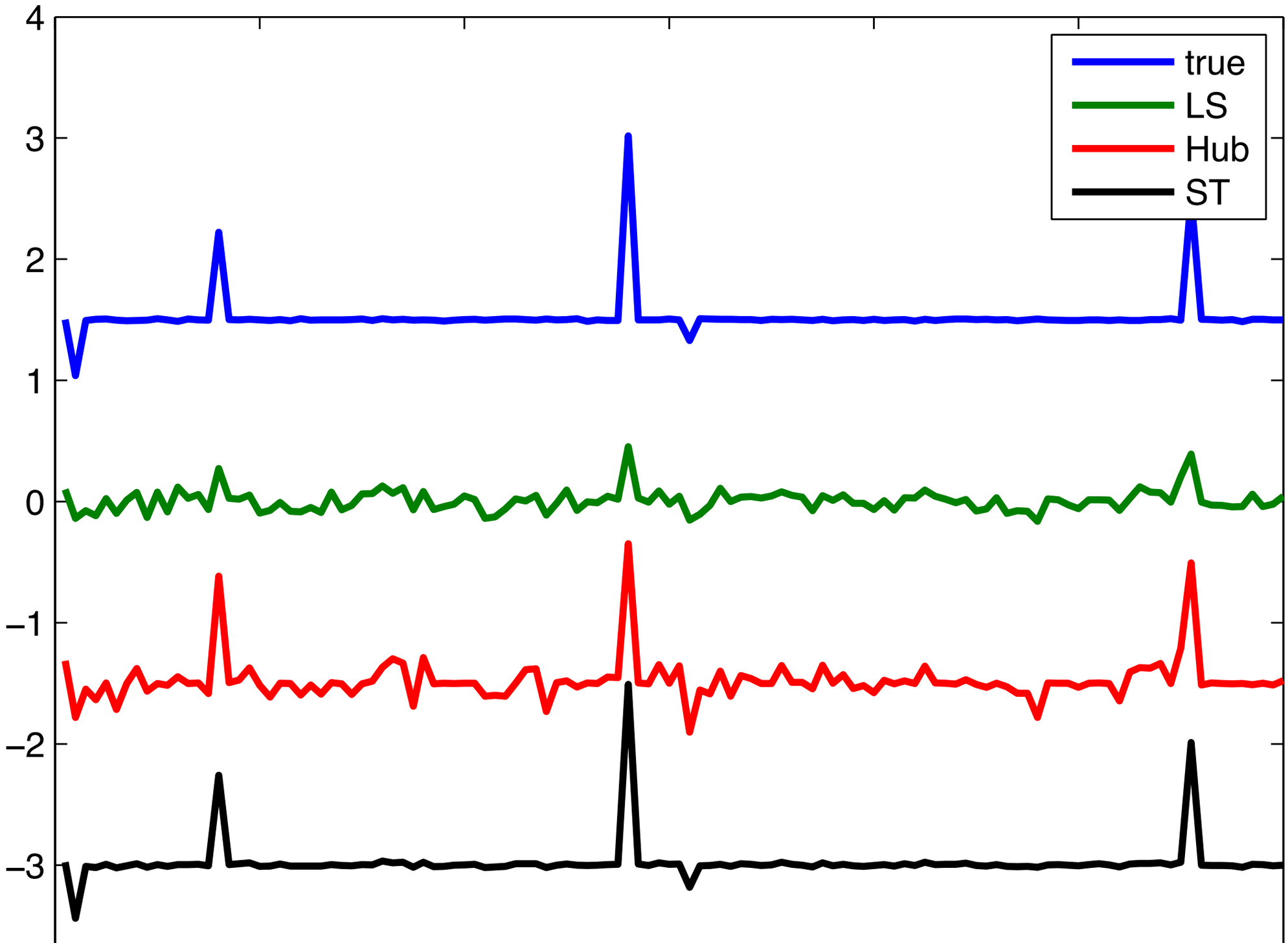
- In this case, the penalty norm is 1-norm, and the operator F is a compressive sensing matrix.

- We try Huber and Student's t penalties (both smooth!)
  - Huber is convex, so Theorem holds.
  - Student's t works just fine
    - Root finder finds the root (which we can tell by the output)
    - Solutions are doing exactly what we want.

Sparse + Robust Results: Signal Recovery



Sparse + Robust Results: Residual Recovery



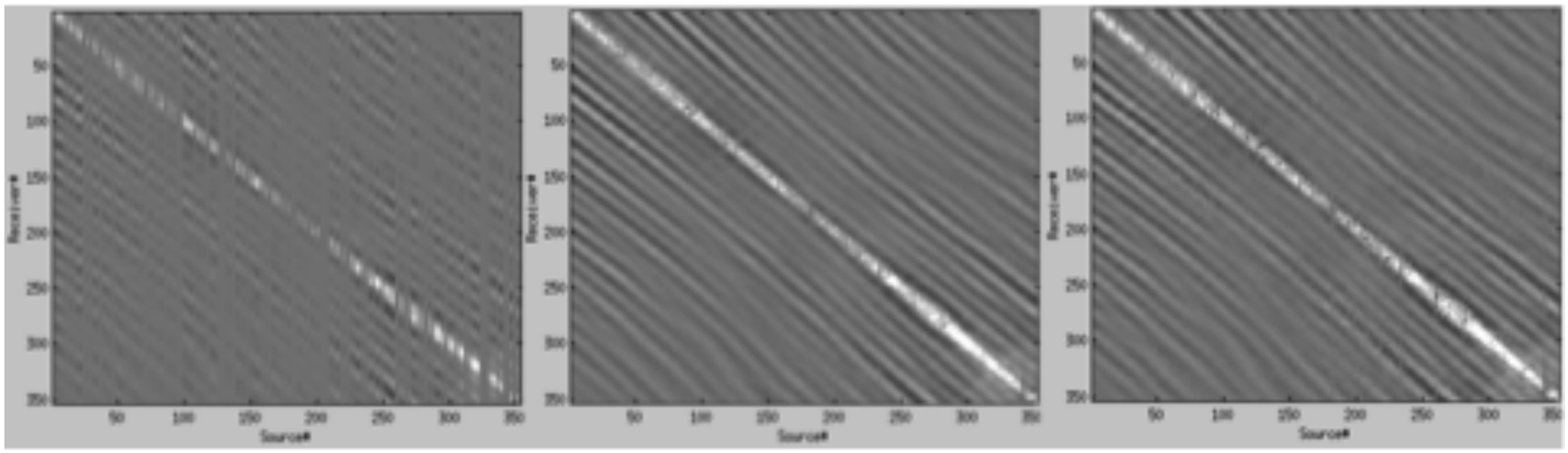
Example 2: Low Rank + Robust Recovery

$$\min_X \|X\|_* \text{ s.t. } h(b - \mathcal{F}(X)) \leq \sigma$$

- We want to solve the Nuclear Norm problem, but we work with L, R factors instead (as discussed in earlier talk).
- We use 50% missing data, and THEN contaminate remaining data with large noise.
- The misfit function is the Student's t penalty.
- The forward model simply picks out the nonzero elements of X

Gulf of Suez: Least square and Student's T+Low Rank

Frequency : 20 Hz, Rank 20



50% missing
+10% outlier

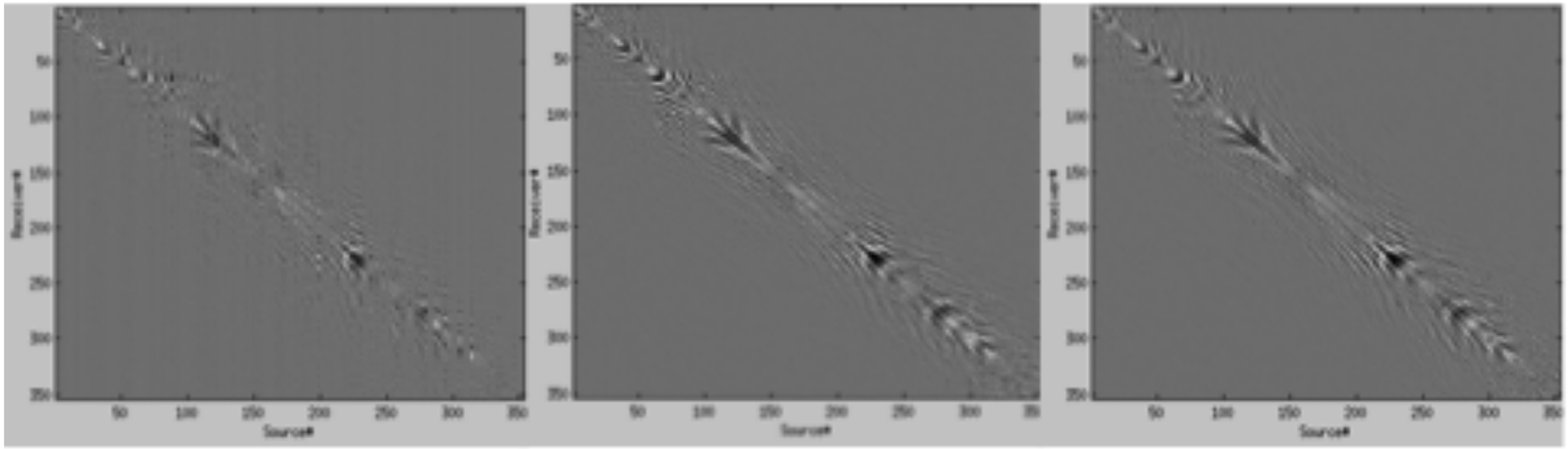
Least Square,
SNR = 24.5 db

Student's t,
SNR = 19.3 db

- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 50x average data value

Gulf of Suez: Least square and Student's T+Low Rank

Frequency : 70 Hz, Rank 20



50% missing
+10% outlier

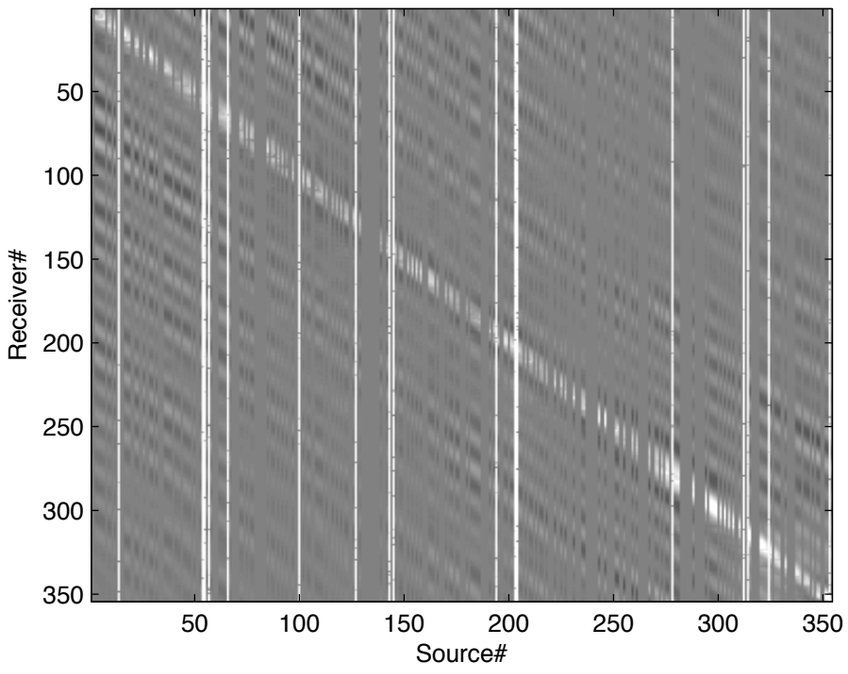
Least Square,
SNR = 15.6 db

Student's t,
SNR = 17.2 db

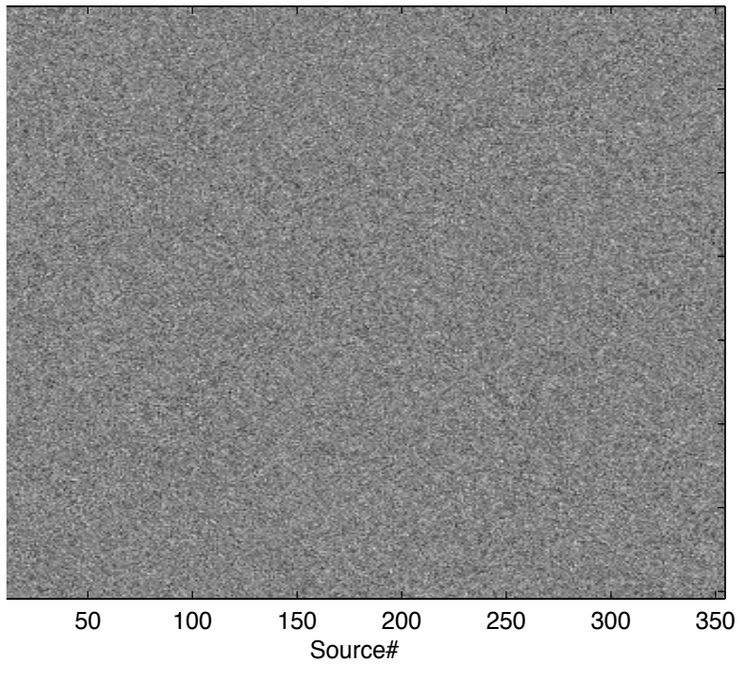
- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 50x average data value

Gulf of Suez: Least square and Student's T+Low Rank

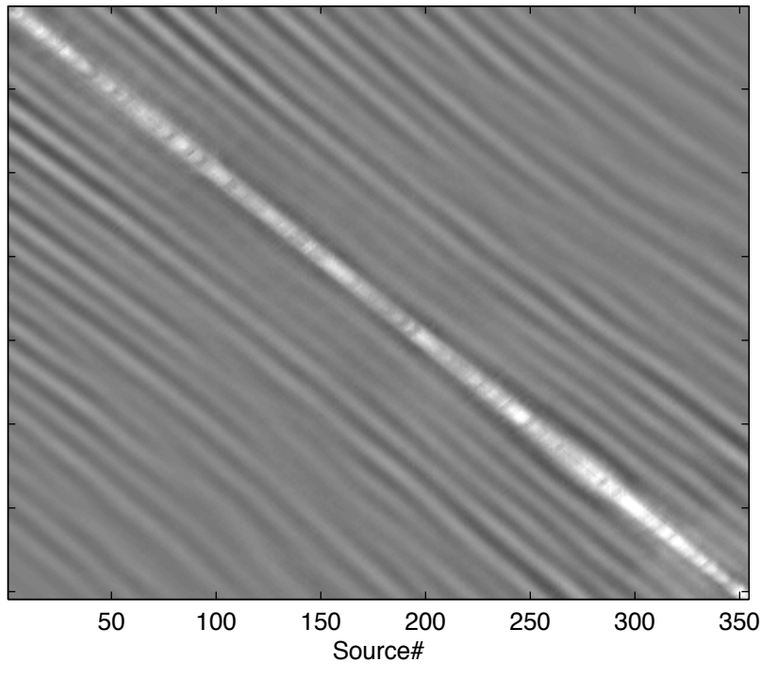
Frequency Slice : 20 Hz, Rank : 20



50 % Missing+10 % corrupted data before interpolation



Recovered data, Least Square
SNR = 4e-7 db

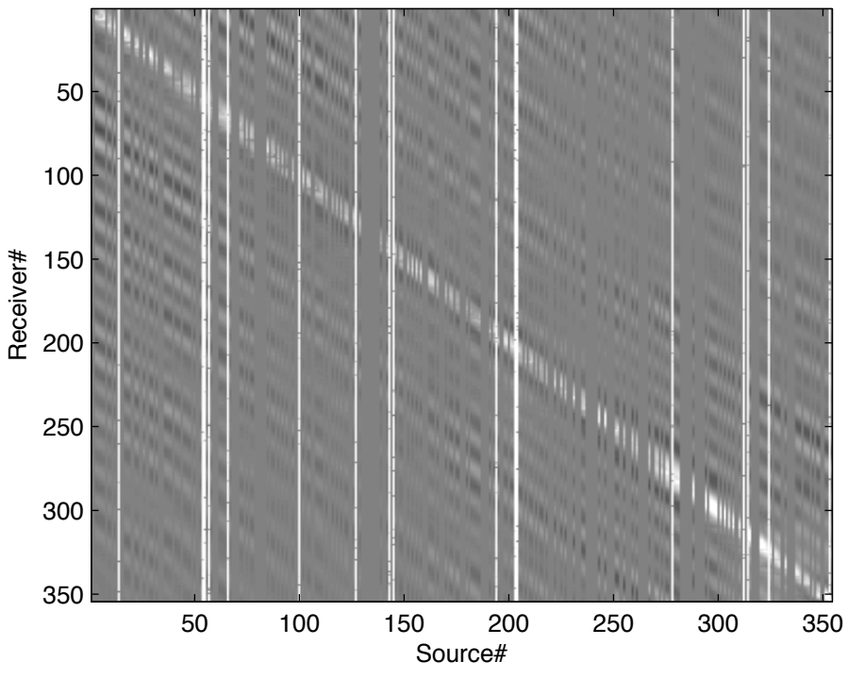


Recovered data, Student's t
SNR = 32.8 db

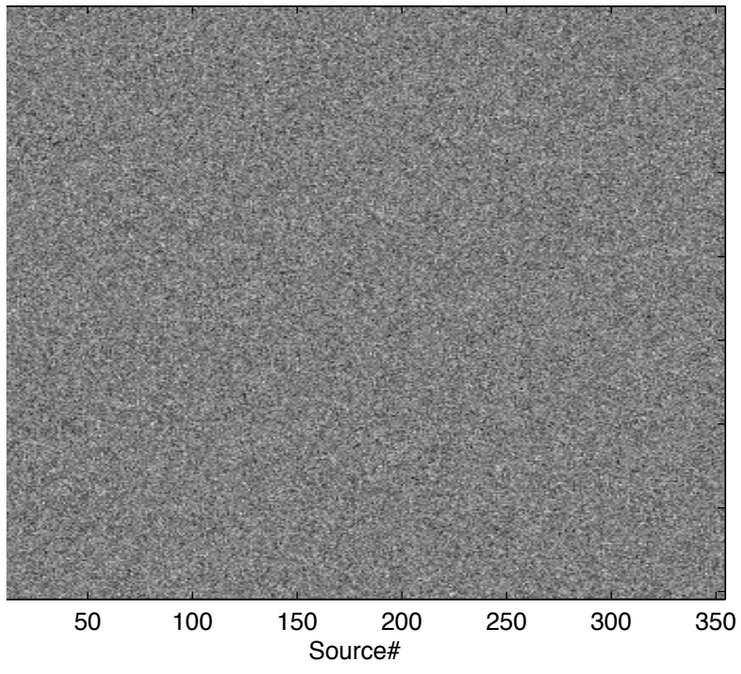
- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 10000x average data value

Gulf of Suez: Least square and Student's T+Low Rank

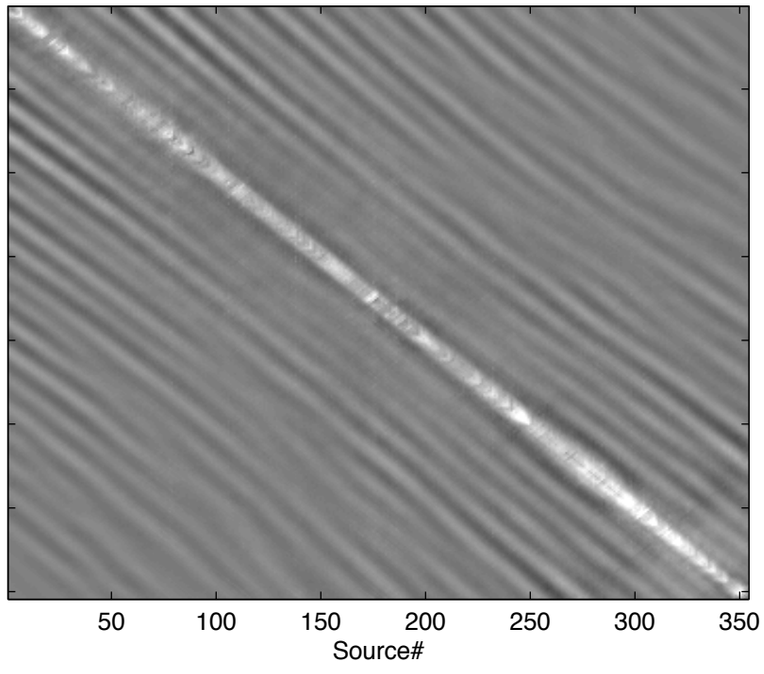
Frequency Slice : 20 Hz, Rank : 40



50 % Missing+10 % corrupted data before interpolation



Recovered data, Least Square
SNR = 5e-7 db

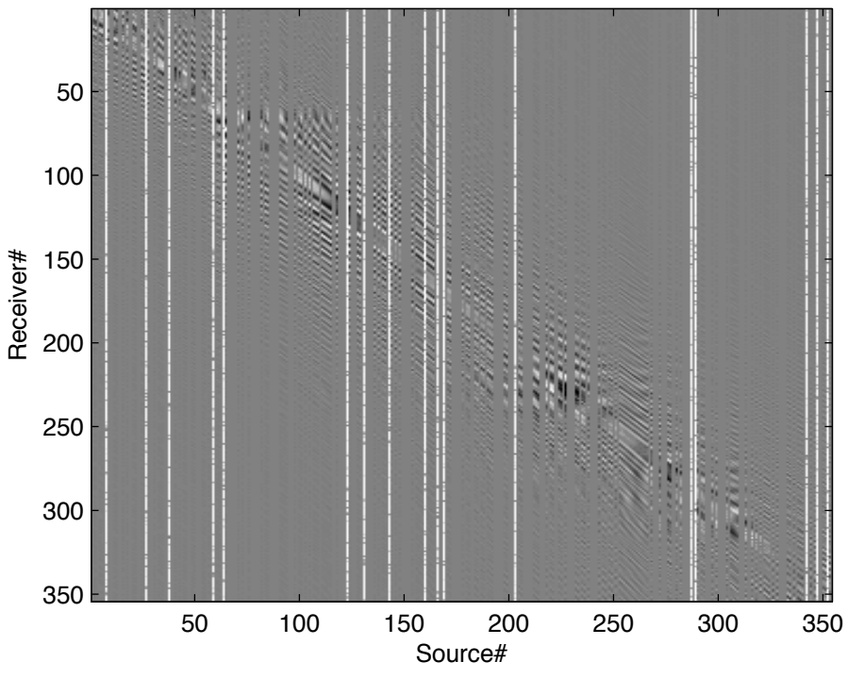


Recovered data, Student's t
SNR = 33.3 db

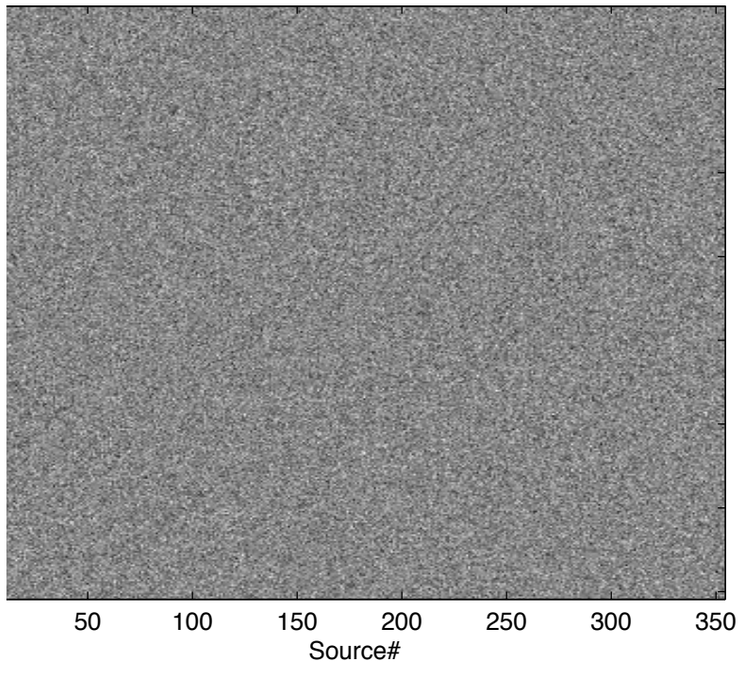
- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 10000x average data value

Gulf of Suez: Least square and Student's T+Low Rank

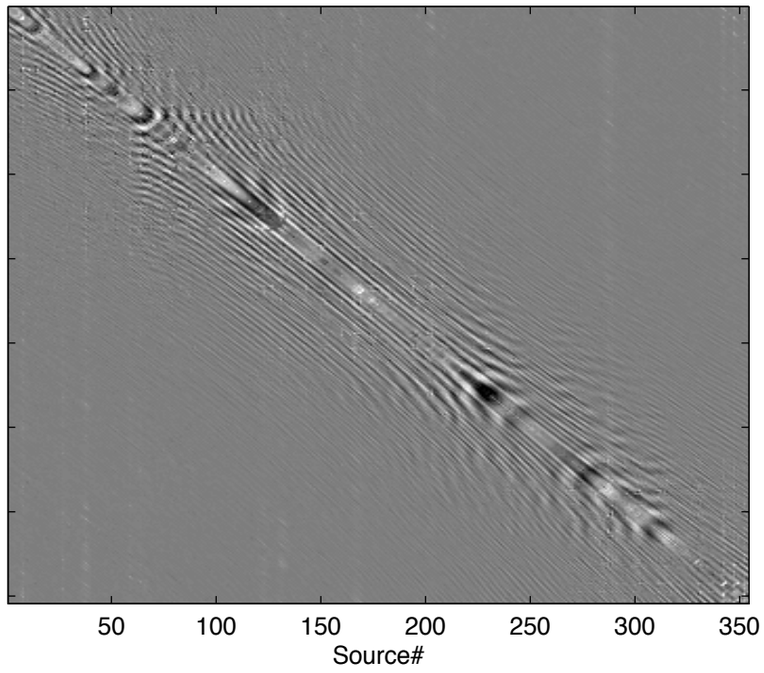
Frequency Slice : 70 Hz, Rank : 20



50 % Missing+10 % corrupted data before interpolation



Recovered data, Least Square
SNR = 1e-8 db

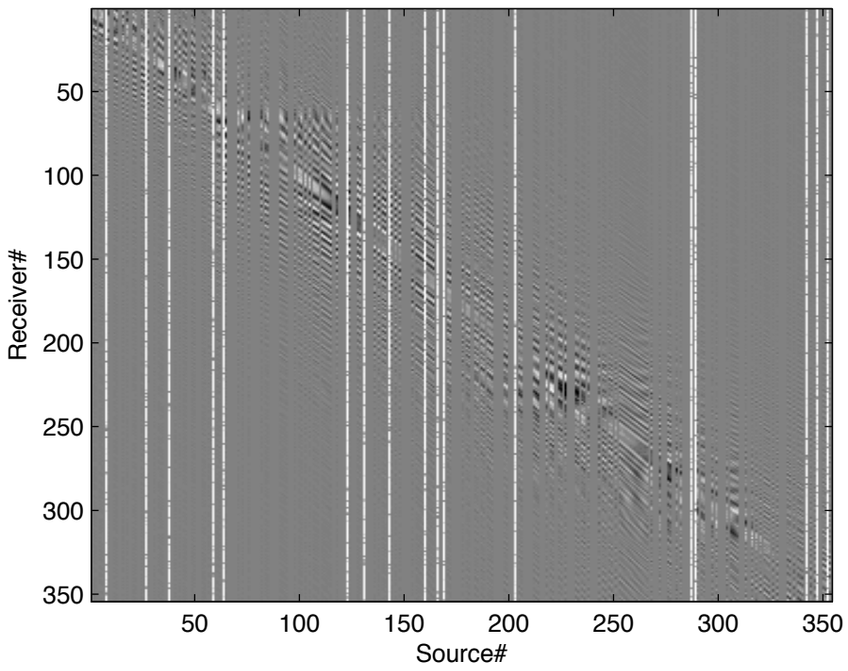


Recovered data, Student's t
SNR = 20.4 db

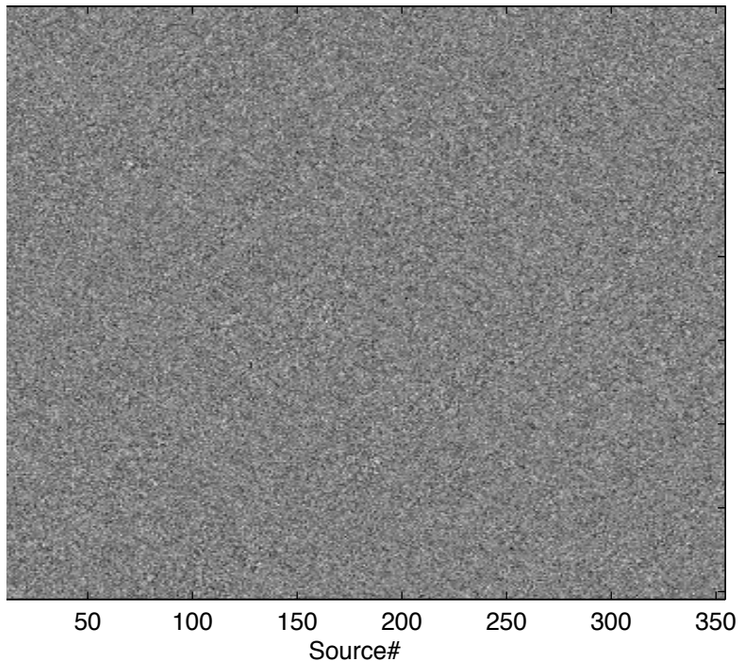
- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 10000x average data value

Gulf of Suez: Least square and Student's T+Low Rank

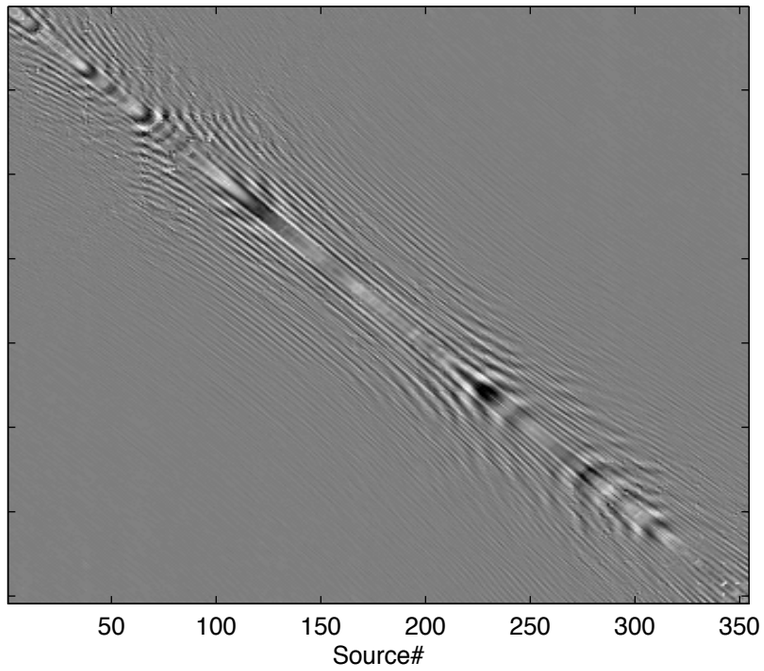
Frequency Slice : 70 Hz, Rank : 40



50 % Missing+10 % corrupted data before interpolation



Recovered data, Least Square
SNR = 1e-7 db



Recovered data, Student's t
SNR = 20.5 db

- 150 SPGL1 iterations; sigma = 10, d.f. = 5e4
- Corrupting noise is 10000x average data value

Conclusions

- Robust formulations to FWI that are able to ignore LARGE unexplained artifacts in the data.

- Scale and degrees of freedom parameters provide tuning knobs that can be tuned using automated data-driven methods.

- Robustness can be combined with sparsity and low rank promotion using the generalized SPGL1 framework.

- Thank you!

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