Extended Robust Formulations for Large Scale Inverse Problems

Aleksandr Y. Aravkin
Computer Science & Earth and Ocean Sciences
University of British Columbia
saravkin@eos.ubc.ca

SLIM Group at UBC
Tristan van Leeuwen
Michael P. Friedlander
Felix J. Herrmann

Total Research (Pau)
Henri Calandra
Raphael Lancretot
Anais Tamalet
Fuchun Gao

SLIM Consortium, December 4, 2012

Outline

- Seismic Inverse Problems

Outline

- Seismic Inverse Problems
- Robust Methods for Inverse Problems
  - Modeling error using parametric densities
  - Theoretical advantages of non-convex penalties/heavy tailed densities
  - Student's t FWI formulation and results

Outline

- Seismic Inverse Problems
- Robust Methods for Inverse Problems
  - Modeling error using parametric densities
  - Theoretical advantages of non-convex penalties/heavy tailed densities
  - Student's t FWI formulation and results
- Variable projection method for nuisance parameters
  - Theory and approach
  - Application 1: Robust source estimation (calibration)
  - Application 2: Self-tuning student's t formulation

Seismic Inverse Problems

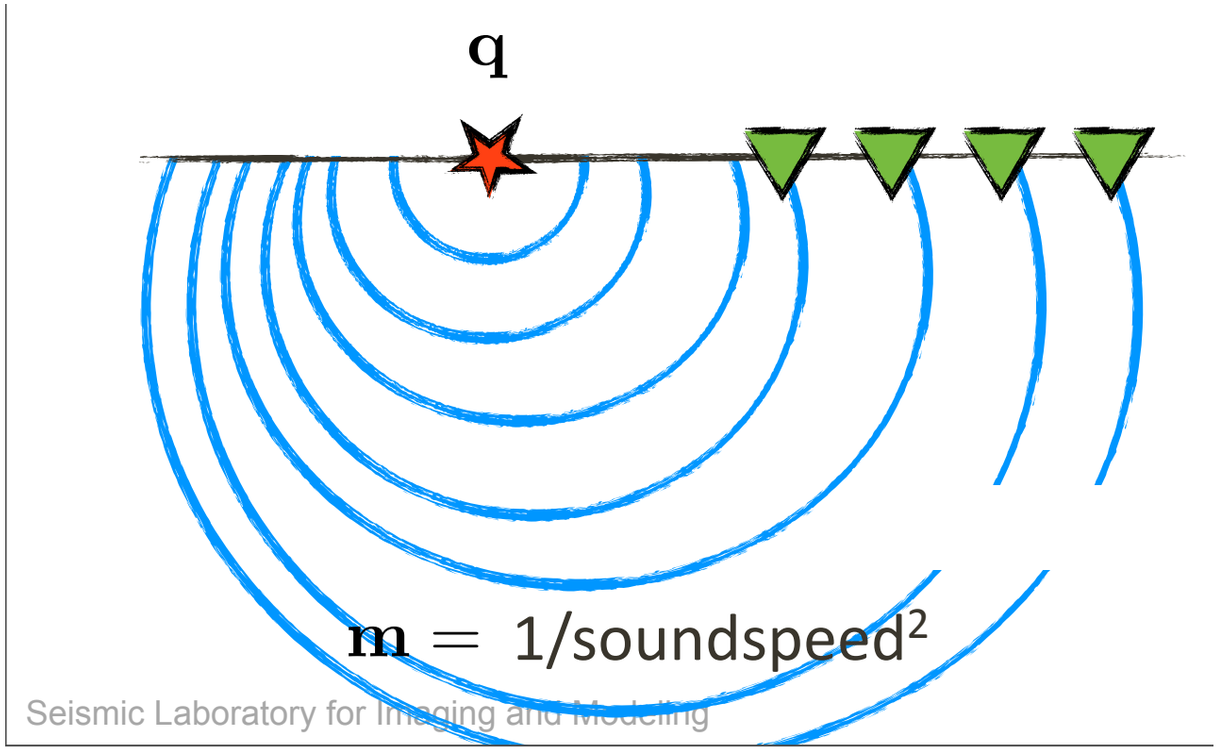
Forward Problem

- The *forward problem* is to predict data given velocity parameters.
- When doing acoustic inversion in the frequency domain, to predict the data from one explosive source $\mathbf{q}$, we solve

$$(\omega^2\mathbf{m} + \nabla^2)\mathbf{u} = \mathbf{q} \ ,$$

and then restrict the solution to the surface with operator P, so that in the discrete setting, the forward model is

$$h(\mathbf{m}) = P \underbrace{(\omega^2\mathbf{m} + \nabla^2)^{-1}\mathbf{q}}_{\mathbf{u}\text{, computed on the fly}} \quad .$$



Inverse problem

- Let $\mathbf{z}$ denote the observed data. We model the relationship between $\mathbf{m}$ and $\mathbf{z}$ using the forward model h and *error model* v:

$$\mathbf{z} = h(\mathbf{m}) + v \ .$$

Inverse problem

- Let $\mathbf{z}$ denote the observed data. We model the relationship between $\mathbf{m}$ and $\mathbf{z}$ using the forward model h and *error model* v:

$$\mathbf{z} = h(\mathbf{m}) + v \ .$$

- We can estimate $\mathbf{m}$ by maximizing the likelihood of $\mathbf{z}$ given a parametric density for errors v. For example, v is i.i.d. Gaussian gives rise to

$$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{z} - h(\mathbf{m})\|^2\right) \quad \longleftrightarrow \quad \overline{\mathbf{m}} = \arg\min_{\mathbf{m}} \|\mathbf{z} - h(\mathbf{m})\|_2^2$$

Inverse problem

- Let $\mathbf{z}$ denote the observed data. We model the relationship between $\mathbf{m}$ and $\mathbf{z}$ using the forward model h and *error model* v:

$$\mathbf{z} = h(\mathbf{m}) + v \ .$$

- We can estimate $\mathbf{m}$ by maximizing the likelihood of $\mathbf{z}$ given a parametric density for errors v. For example, v is i.i.d. Gaussian gives rise to

$$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{z} - h(\mathbf{m})\|^2\right) \quad \longleftrightarrow \quad \overline{\mathbf{m}} = \arg\min_{\mathbf{m}} \|\mathbf{z} - h(\mathbf{m})\|_2^2$$

- In general, if v has negative log $g = -\log(p)$, the maximum likelihood problem is equivalent to

$$\min_{\mathbf{m}} f(\mathbf{m}) := g(\mathbf{z} - h(\mathbf{m})) \ .$$

Inverse problem

- Let $\mathbf{z}$ denote the observed data. We model the relationship between $\mathbf{m}$ and $\mathbf{z}$ using the forward model h and *error model* v:

$$\mathbf{z} = h(\mathbf{m}) + v \ .$$

- We can estimate $\mathbf{m}$ by maximizing the likelihood of $\mathbf{z}$ given a parametric density for errors v. For example, v is i.i.d. Gaussian gives rise to

$$\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{z} - h(\mathbf{m})\|^2\right) \quad \longleftrightarrow \quad \overline{\mathbf{m}} = \arg\min_{\mathbf{m}} \|\mathbf{z} - h(\mathbf{m})\|_2^2$$

- In general, if v has negative log $g = -\log(p)$, the maximum likelihood problem is equivalent to

$$\min_{\mathbf{m}} f(\mathbf{m}) := g(\mathbf{z} - h(\mathbf{m})) \ .$$

- For FWI, $h(\mathbf{m}) = PH^{-1}[\mathbf{m}]q$, where $H[\mathbf{m}] = (\omega^2\mathbf{m} + \nabla^2)$.

Formulation summary

- Given a forward model $h(\mathbf{m})$, we can formulate an inverse problem using model

$$\mathbf{z} = h(\mathbf{m}) + v \ .$$

- The corresponding optimization problem is

$$\min_{\mathbf{m}} f(\mathbf{m}) = g(\mathbf{z} - h(\mathbf{m})) \ ,$$

- The *gradient* is the key computational unit, and for any differentiable g, we can compute the gradient on the fly by solving PDEs.

- We focus on robust and extended modeling formulations, which correspond to choosing a good g and solving the resulting problem.

Robust Methods Via Error Modeling

Why robust methods?

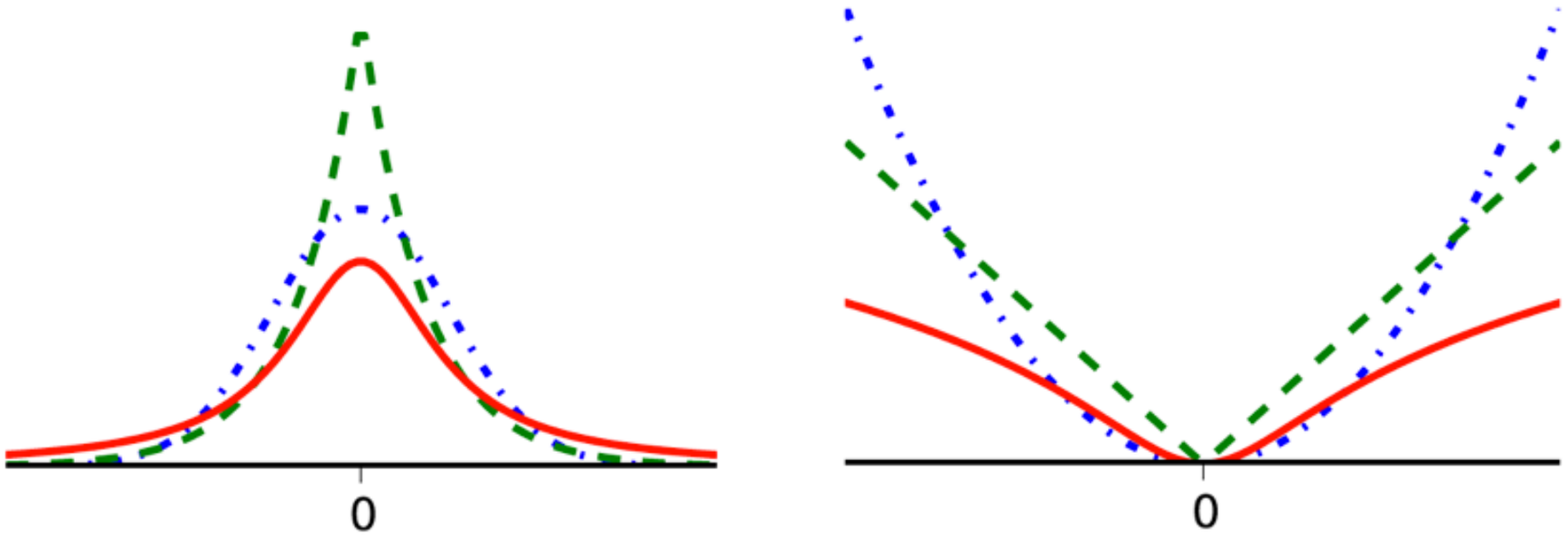
- Unexplained artifacts in the data are a major problem for current inverse problem formulations. In exploration geophysics, data must be cleaned and pre-processed before FWI is done.

- Cleaning data is costly and time consuming — it can take *years*, while solving the inverse problem itself is a matter of months. Moreover, 'cleaning' is not guaranteed to succeed, and you can lose information.

- We main idea of robust methods is to obtain good results **despite** unexplained artifacts or errors in the data.

Densities, and Penalties

- We can design robust methods by changing the *statistical model* for v, which gives a corresponding *error penalty*.



- Least squares: $\min_{\mathbf{m}} \|\mathbf{z} - h(\mathbf{m})\|_2^2 = \sum (z_i - h(\mathbf{m})_i)^2$
- L1 or Huber: $\min_{\mathbf{m}} \|\mathbf{z} - h(\mathbf{m})\|_1 = \sum |z_i - h(\mathbf{m})_i|$
- Student's t: $\min_{\mathbf{m}} \sum \log\left(\nu + (z_i - h(\mathbf{m})_i)^2\right)$

Advantages of Heavy-Tailed Densities

- A convex penalty corresponds to a log-concave density, while a non-convex penalty corresponds to a heavy-tailed density.
- The class of log-concave densities is fundamentally limited from a modeling point of view.

Theorem: A.A., Friedlander, van Leeuwen, Herrmann, Math Prog. 2012

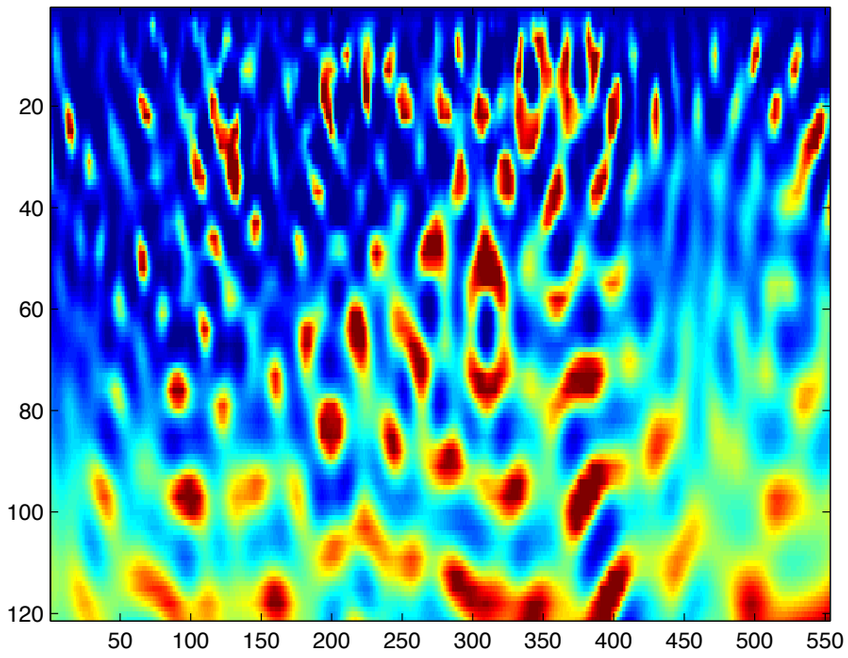
Let $\mathbf{p}$ be a scalar density proportional to $\exp(-g(y))$, with g convex. Then we have

$$P\left(|y| > t + \Delta t \quad \text{given} \quad |y| > t\right) \leq C \exp(-\alpha \Delta t) \ .$$

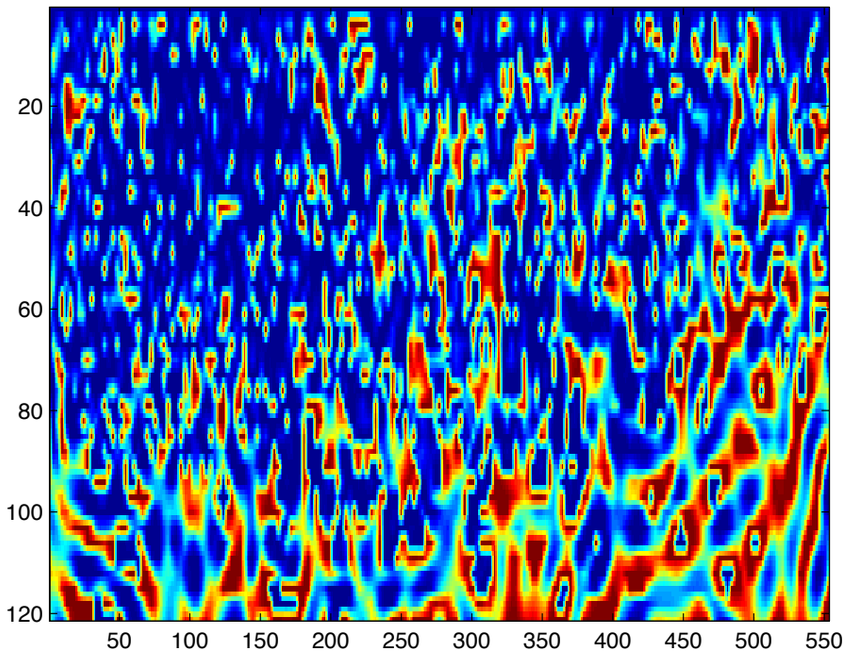
- **Heavy-tailed** models are *less conservative* with regard to outlier distributions.
  - Student's t, Pareto, Cauchy are all heavy-tailed.
  - For Cauchy, $P\left(|y| > 2d \middle| |y| > d\right) \approx 0.5$ for large d!

Total Results: 4% bad data, LS vs. Student's t

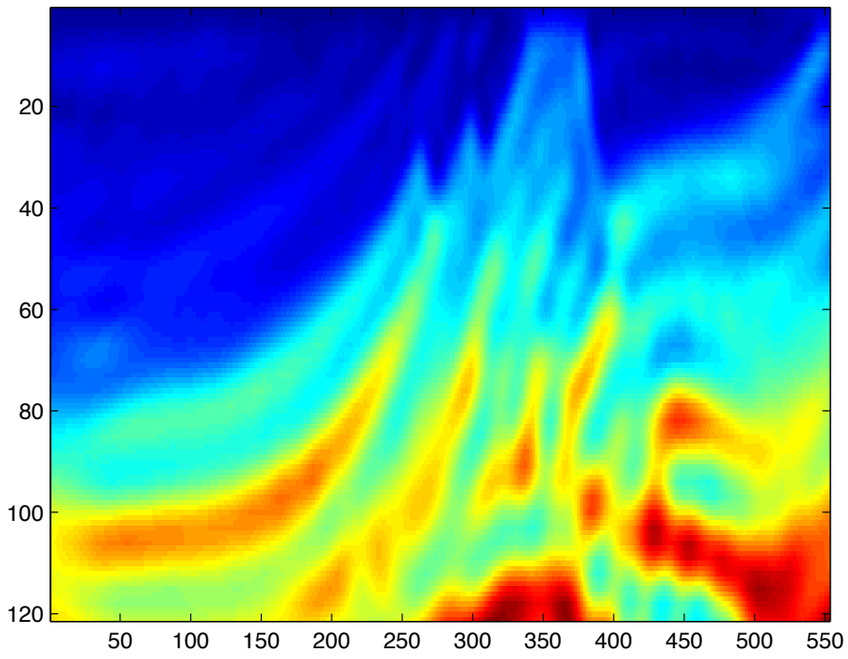
4Hz, LS, 4% bad data



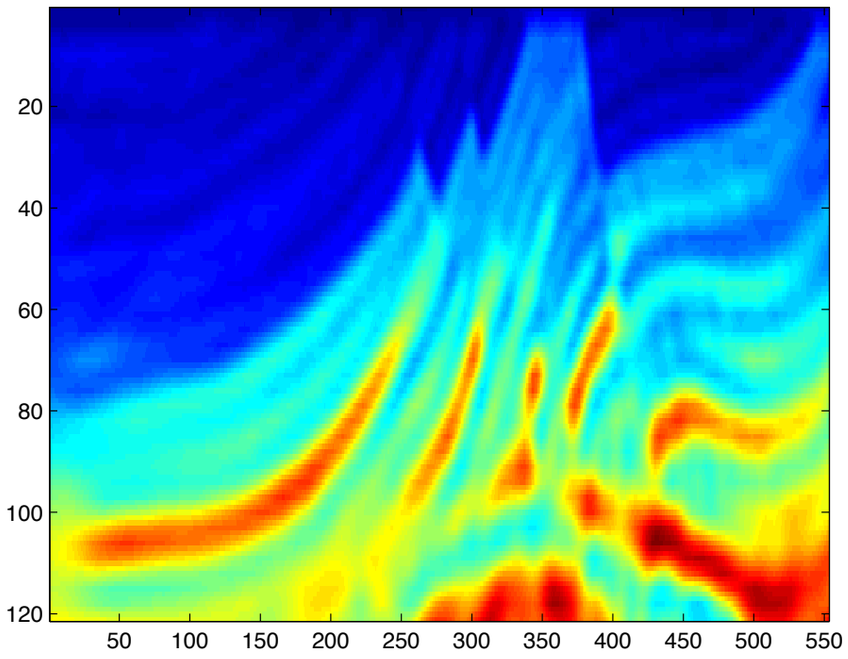
6 Hz, LS, 4% bad data



4Hz, St: $\nu = 100$, 4% bad data

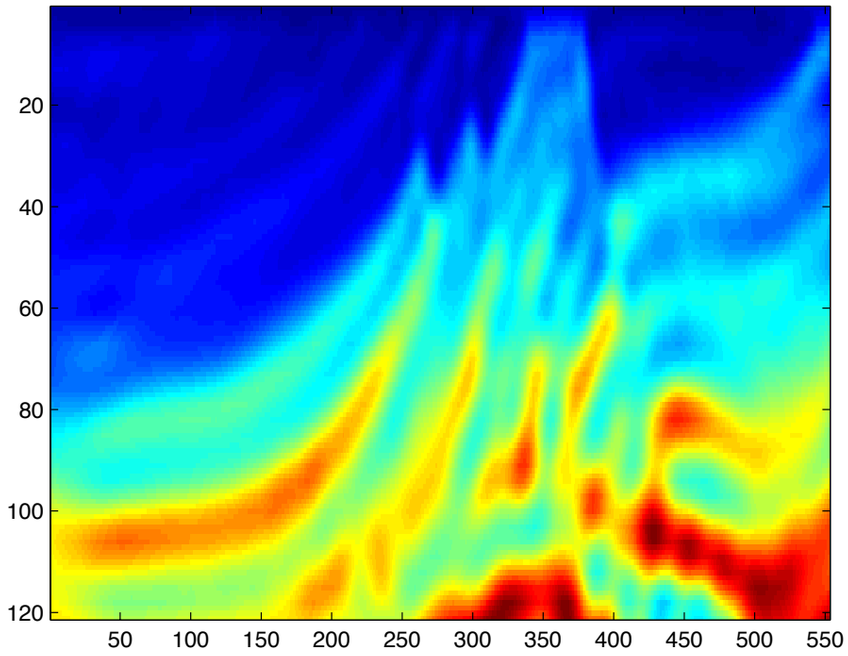


6 Hz, St: $\nu = 100$, 4% bad data

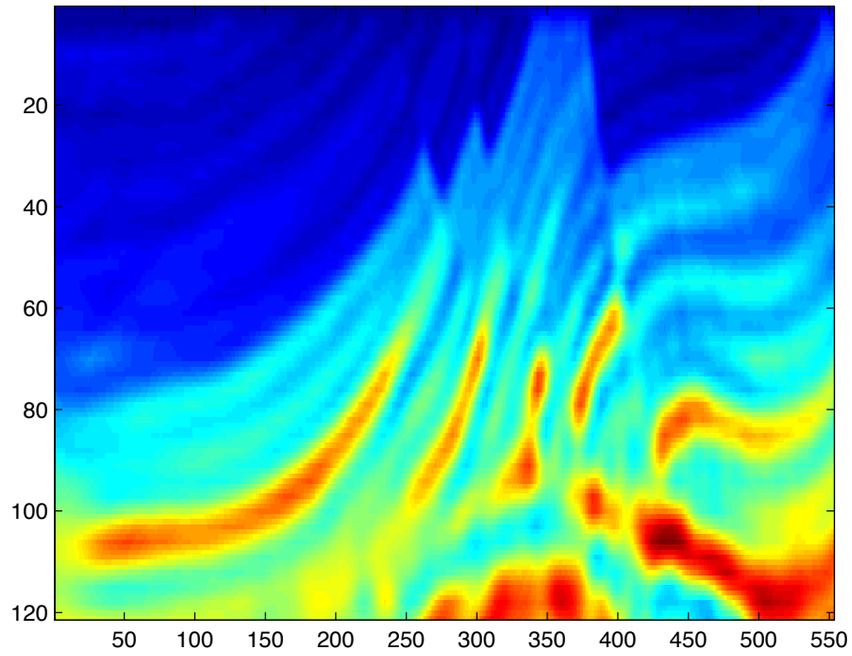


Total Results: St 30% bad data vs. LS good data

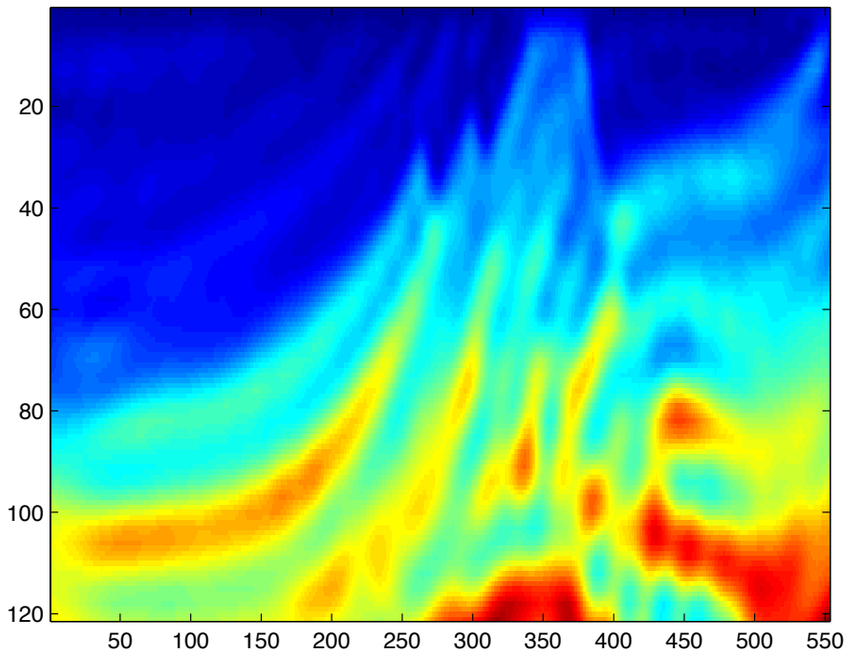
4Hz, LS, good data



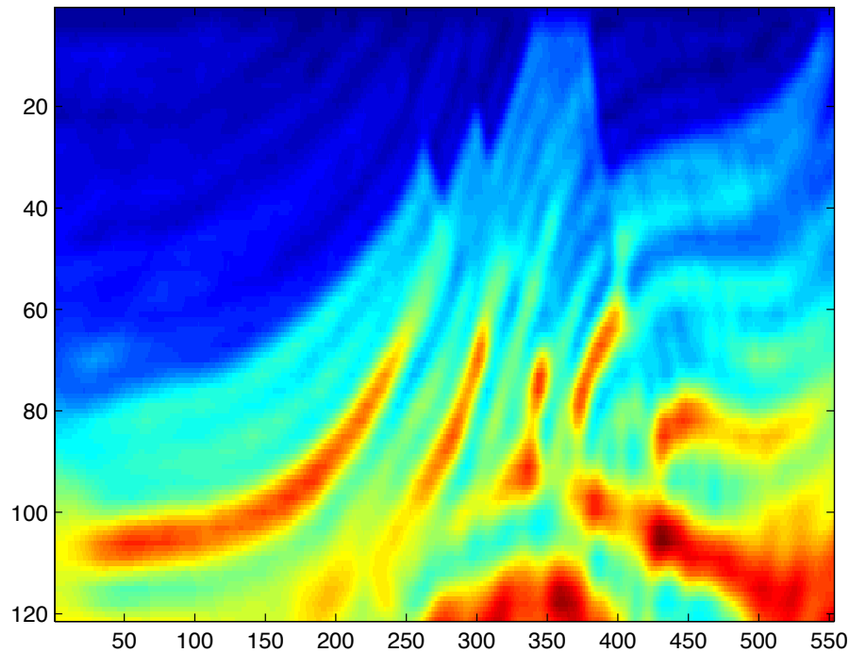
6 Hz, LS, good data



4Hz, St: $\nu = 1$, 30% bad data



6 Hz, St: $\nu = 1$, 30% bad data



Nuisance Parameter Estimation

Nuisance parameters in inverse problems

- Many inverse problems contain *nuisance parameters* which, while not of primary interest, impact our ability to invert for primary parameters.
- Motivating applications:
  - Variance parameters, especially when data comes from several data sets (different shot experiments, different frequencies)
  - Automatic calibration, e.g. *source estimation* in seismic inverse problems.
  - Meta-parameter estimation, e.g. Student's t degrees of freedom.
- For a large class of problems (including all motivating applications) we can *jointly* optimize extended formulations in *both* primary and nuisance parameters.
- In many cases, this can be easily implemented within existing architectures that currently ignore nuisance parameters.

Formulations for Motivating Applications

- If data comes from K sources with Gaussian errors, we can formulate the following (joint) inverse problem: (Anais's talk)

$$\min_{\mathbf{m},\sigma^2} \sum_{i=1}^{K} N_i \log(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2} \|\mathbf{z}_i - h(\mathbf{m})_i\|^2$$

- *Source estimation* means solving for *unknown amplitude parameters*:

$$\min_{\mathbf{m},\alpha} \sum_{i=1}^{M} g(\mathbf{z}_i - \alpha_i h(\mathbf{m})_i) \ .$$

- *Student's t* d.f. parameters may be estimated by minimizing the true negative log likelihood:

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

General Function Class

- All of the motivating formulations take the form

$$\min_{\mathbf{m},\theta} g(\mathbf{m},\theta) \ .$$

- In all of the applications, it is *easy* to compute

$$\overline{\theta}(\mathbf{m}) = \arg\min_{\theta} g(\mathbf{m},\theta) \ .$$

- These facts motivate the definition of a *reduced* objective

$$\tilde{g}(\mathbf{m}) = g(\mathbf{m},\overline{\theta}(\mathbf{m})) \ .$$

- For special types of NLLS problems, this technique is known as *variable projection* (Golub & Pereyra '73, '03, Osborne '07).

Projection Approach

Define $\tilde{g}(\mathbf{m}) = g(\mathbf{m}, \bar{\theta}(\mathbf{m}))$, with $\bar{\theta}(\mathbf{m}) = \arg\min_\theta g(\mathbf{m}, \theta)$.

Theorem: Bell & Burke, '08.

Let $\overline{\mathbf{m}} \in \mathcal{U}$ and $\bar{\theta} \in \mathcal{V}$ be such that $\nabla_\theta g(\overline{\mathbf{m}}, \bar{\theta}) = 0$ and $\nabla^2_\theta g(\overline{\mathbf{m}}, \bar{\theta})$ is positive definite. Then $\tilde{g}(\mathbf{m})$ is twice continuously differentiable, with

$$\nabla_{\mathbf{m}} \tilde{g}(\overline{\mathbf{m}}) = \nabla_{\mathbf{m}} g(\overline{\mathbf{m}}, \bar{\theta}(\overline{\mathbf{m}})) \quad (1)$$

$$\nabla^2_{\mathbf{m}} \tilde{g}(\overline{\mathbf{m}}) = \nabla^2_{\mathbf{m}} g(\overline{\mathbf{m}}, \bar{\theta}(\overline{\mathbf{m}})) + \nabla^2_{\mathbf{m},\theta} g(\overline{\mathbf{m}}, \bar{\theta}(\overline{\mathbf{m}})) \nabla_{\mathbf{m}} \bar{\theta}(\overline{\mathbf{m}}) \, . \quad (2)$$

- We can now design first and second order methods to optimize $\tilde{g}$.
- First order: recompute $\bar{\theta}(\mathbf{m})$ whenever $\mathbf{m}$ is updated, and then use $\bar{\theta}(\mathbf{m})$ (as values!) wherever θ appears in $\nabla_{\mathbf{m}} g(\mathbf{m}, \theta)$. From (1), we have the true gradient of $\tilde{g}$.
- Second order: Use a truncated version of (2), again with $\bar{\theta}(\mathbf{m})$ as values, but this time with your favorite solver, e.g. Gauss-Newton, LM.

A.A. & van Leeuwen, *Estimating Nuisance Parameters in Inverse Problems*, Inverse Problems, October 2012.

Application I: Robust source estimation

- Consider $f(\mathbf{m}, \alpha) = \sum_i g_i(\mathbf{z}_i - \alpha_i h(\mathbf{m})_i)$.
- When $g_i = \|\cdot\|^2$, then $\bar{\alpha}_i = \frac{h(\mathbf{m})_i^\top \mathbf{z}_i}{\|h(\mathbf{m})_i\|^2}$.
- For general g_i, we can very quickly solve for $\bar{\alpha}$ using Newton's method.
- Then, reduced objective is

$$\tilde{f}(\mathbf{m}) = \sum_i g_i(\mathbf{z}_i - \bar{\alpha}_i(\mathbf{m}) h(\mathbf{m})_i),$$

with

$$\nabla \tilde{f}(\mathbf{m}) = \sum_i \nabla h(\mathbf{m})_i^\top \nabla g_i(\mathbf{z}_i - \bar{\alpha}_i(\mathbf{m}) h(\mathbf{m})_i) \ .$$

Application I: Robust FWI with Robust Source estimation

Source Estimation

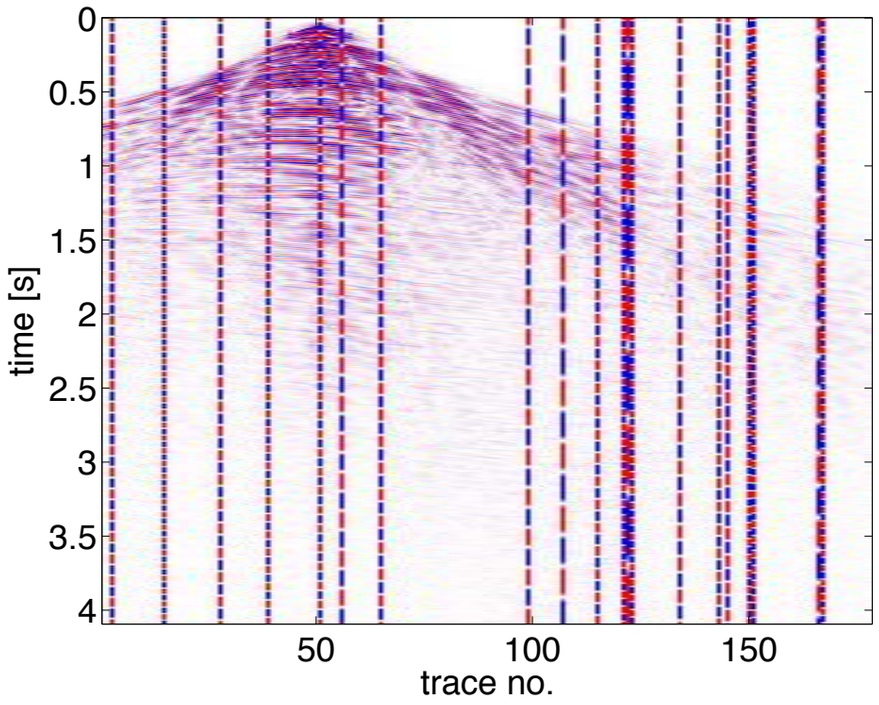


Figure 1 *Data with outliers in the form of bad traces.*

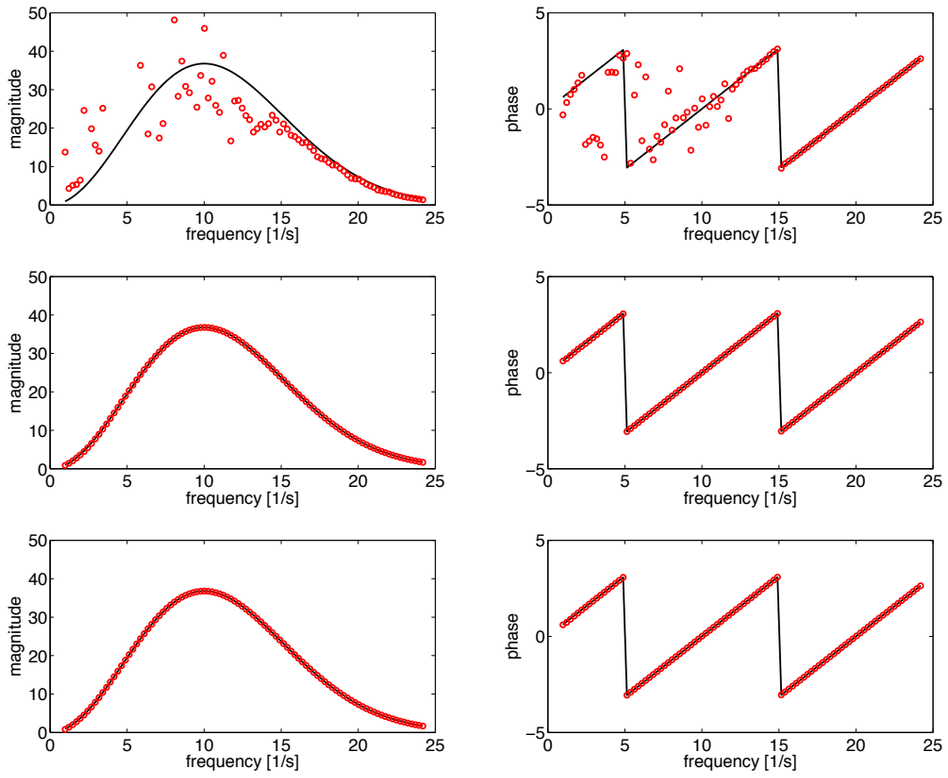
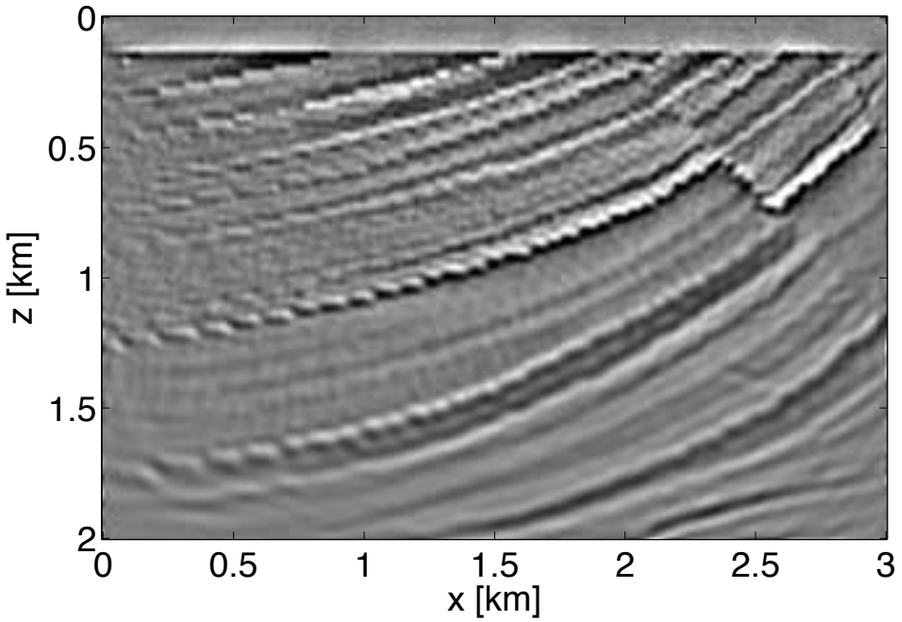


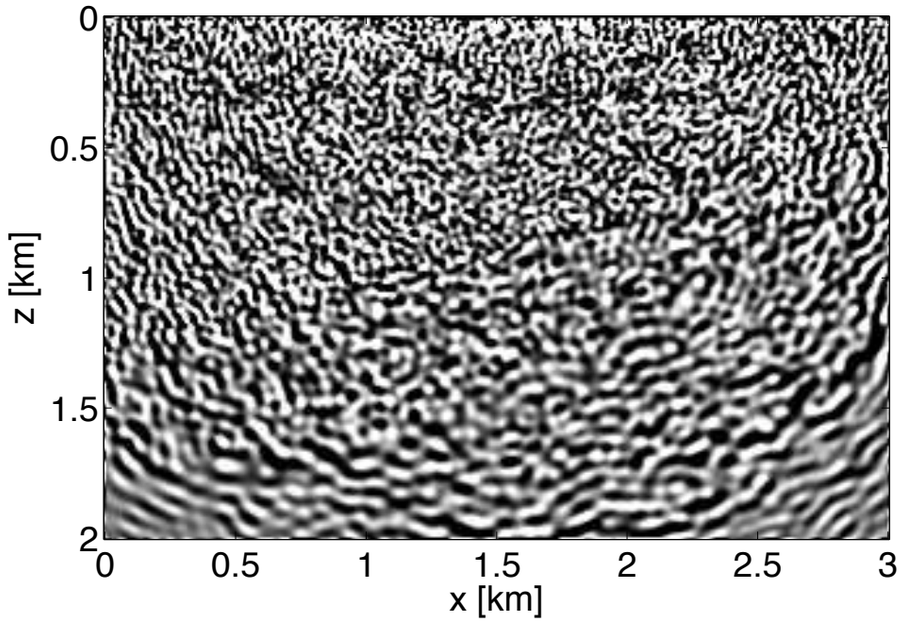
Figure 2 *Estimated source wavelet using Least-Squares (top), Hybrid (middle) and Students t (bottom) approaches.*

Application I: Robust FWI with Robust Source estimation

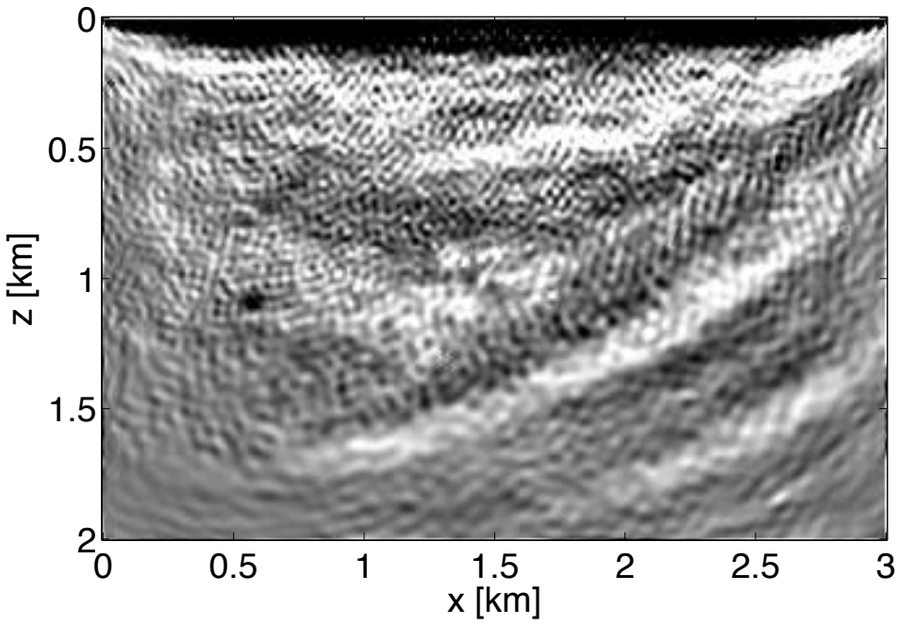
FWI Results



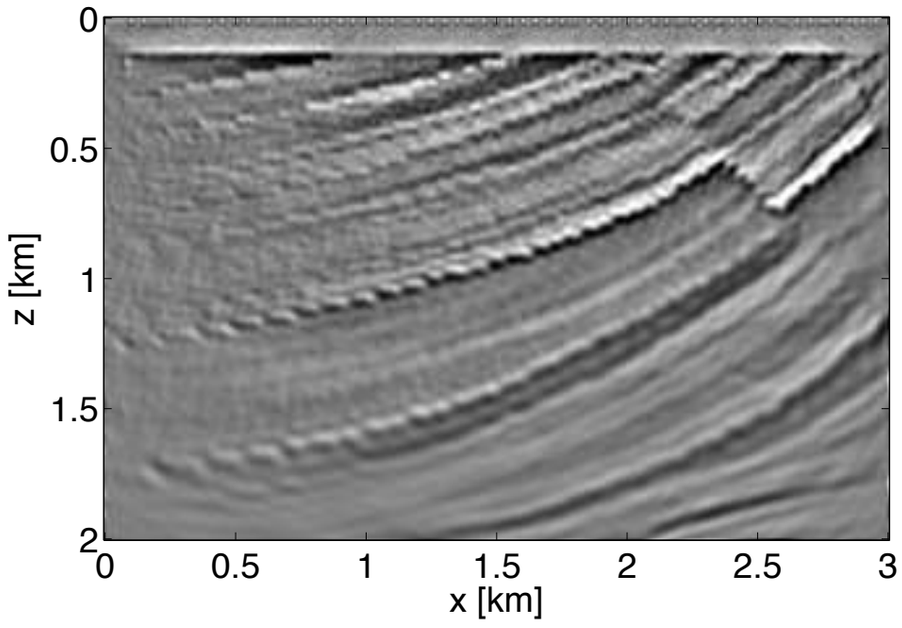
LS-LS w/o noise



LS-LS w noise



ST-LS w noise



ST-ST w noise

Application II: Self-tuning Student's t

- Take the *full log likelihood*:

$$g(\mathbf{m}, \sigma^2, \nu) = -n \log \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \sigma^2}} \right) + \frac{\nu+1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{\mathbf{r}(\mathbf{m})_i^2}{\sigma^2 \nu} \right)$$

Application II: Self-tuning Student's t

- Take the *full log likelihood*:

$$g(\mathbf{m}, \sigma^2, \nu) = -n \log \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \sigma^2}} \right) + \frac{\nu+1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{\mathbf{r}(\mathbf{m})_i^2}{\sigma^2 \nu} \right)$$

- Using change of variables: $\eta = \frac{\nu+1}{2}$, $\varrho = \sigma^2 \nu$, we find

$$(\overline{\varrho}, \overline{\eta}) = \underset{\varrho, \eta}{\arg\min} -n \log \left(\frac{\Gamma(\eta)}{\Gamma\left(\eta - \frac{1}{2}\right)} \right) + \frac{n}{2} \log(\varrho) + \eta \sum_{i=1}^{n} \log \left(1 + \frac{r_i(x)^2}{\varrho} \right) .$$

Application II: Self-tuning Student's t

- Take the *full log likelihood*:

$$g(\mathbf{m}, \sigma^2, \nu) = -n \log \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \sigma^2}} \right) + \frac{\nu+1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{\mathbf{r}(\mathbf{m})_i^2}{\sigma^2 \nu} \right)$$

- Using change of variables: $\eta = \frac{\nu+1}{2}$, $\varrho = \sigma^2 \nu$, we find

$$(\overline{\varrho}, \overline{\eta}) = \underset{\varrho,\eta}{\arg\min} -n \log \left(\frac{\Gamma(\eta)}{\Gamma\left(\eta - \frac{1}{2}\right)} \right) + \frac{n}{2} \log(\varrho) + \eta \sum_{i=1}^{n} \log \left(1 + \frac{r_i(x)^2}{\varrho} \right) .$$

- Taking the derivative with respect to ϱ, we find

$$0 = \frac{n}{2\varrho} - \eta \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2} \implies \eta = \frac{n}{2 \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2}}$$

Application II: Self-tuning Student's t

- Take the *full log likelihood*:

$$g(\mathbf{m}, \sigma^2, \nu) = -n \log \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \sigma^2}} \right) + \frac{\nu+1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{\mathbf{r}(\mathbf{m})_i^2}{\sigma^2 \nu} \right)$$

- Using change of variables: $\eta = \frac{\nu+1}{2}$, $\varrho = \sigma^2 \nu$, we find

$$(\overline{\varrho}, \overline{\eta}) = \operatorname*{arg\,min}_{\varrho, \eta} -n \log \left(\frac{\Gamma(\eta)}{\Gamma\left(\eta - \frac{1}{2}\right)} \right) + \frac{n}{2} \log(\varrho) + \eta \sum_{i=1}^{n} \log \left(1 + \frac{r_i(x)^2}{\varrho} \right) .$$

- Taking the derivative with respect to ϱ, we find

$$0 = \frac{n}{2\varrho} - \eta \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2} \implies \eta = \frac{n}{2 \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2}}$$

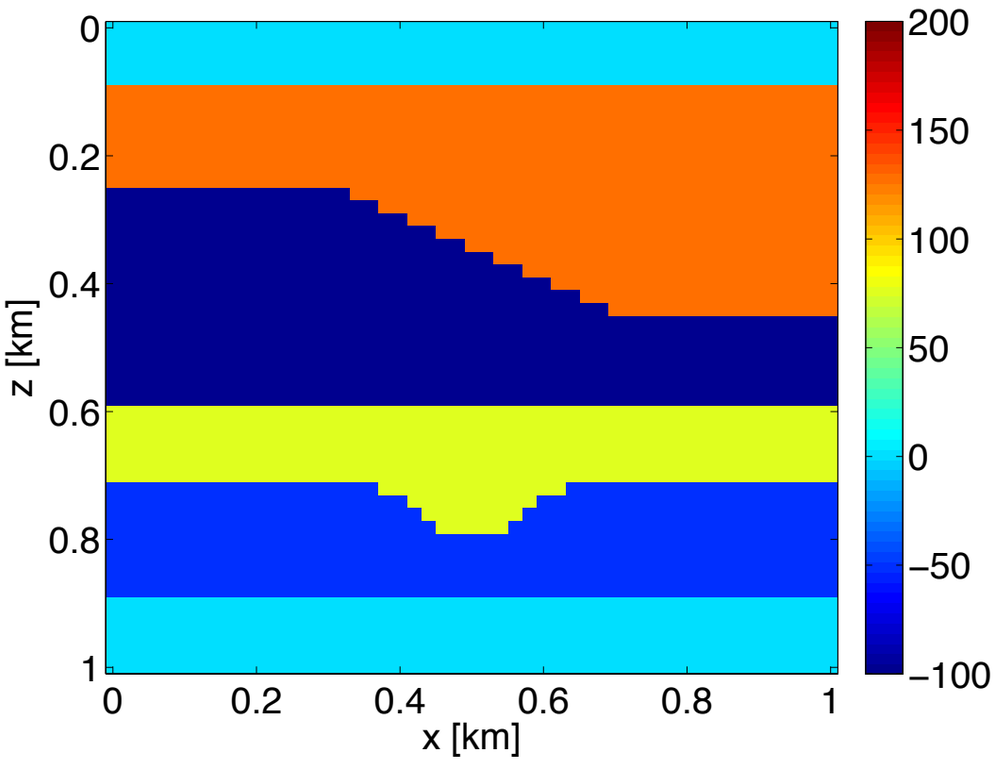
- Plugging back in, we optimize the *scalar* problem

$$\min_{\varrho} -n \log \left(\frac{\Gamma\left[n \Big/ \left(2 \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2} \right) \right]}{\Gamma\left[n \Big/ \left(2 \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2} \right) - \frac{1}{2} \right]} \right) + \frac{n}{2} \log(\varrho) + \frac{n \sum_{i=1}^{n} \log \left(1 + \frac{r_i^2}{\varrho} \right)}{2 \sum_{i=1}^{n} \frac{r_i^2}{\varrho + r_i^2}}$$

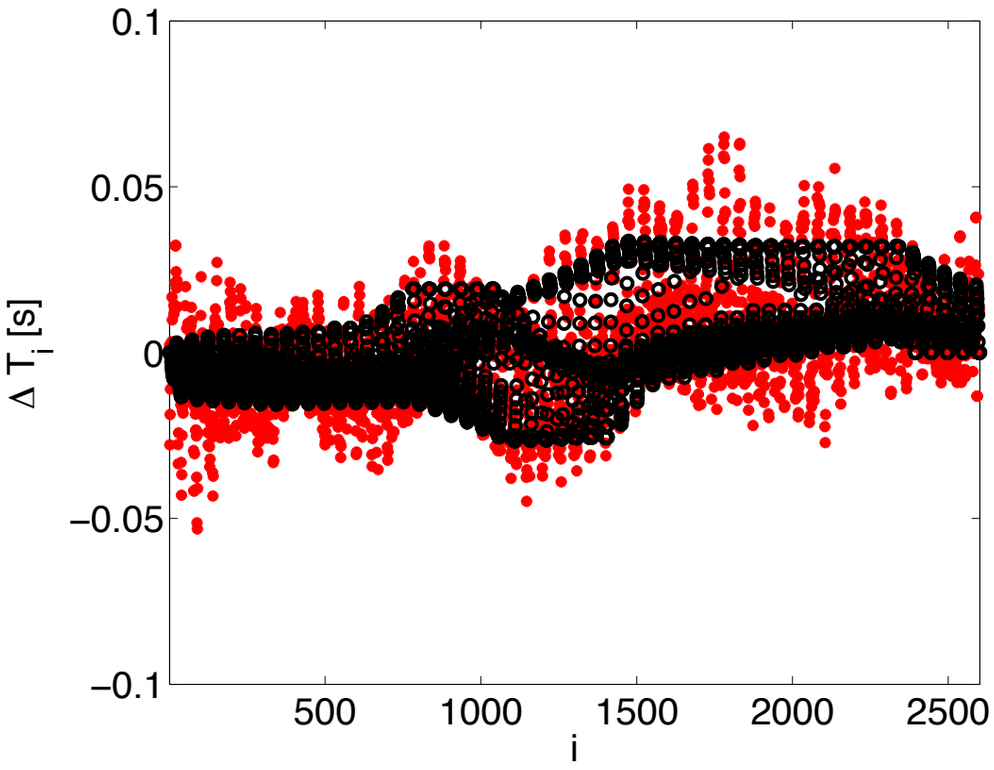
Application II: Results for traveltime tomography

- We set up a traveltime tomography experiment: starting from constant background, we want to invert for a velocity perturbation.
- Linear forward model: $h(\mathbf{m}) = A\mathbf{m}$. Data: $\mathbf{z} = \Delta T$, arrival times.

True perturbation

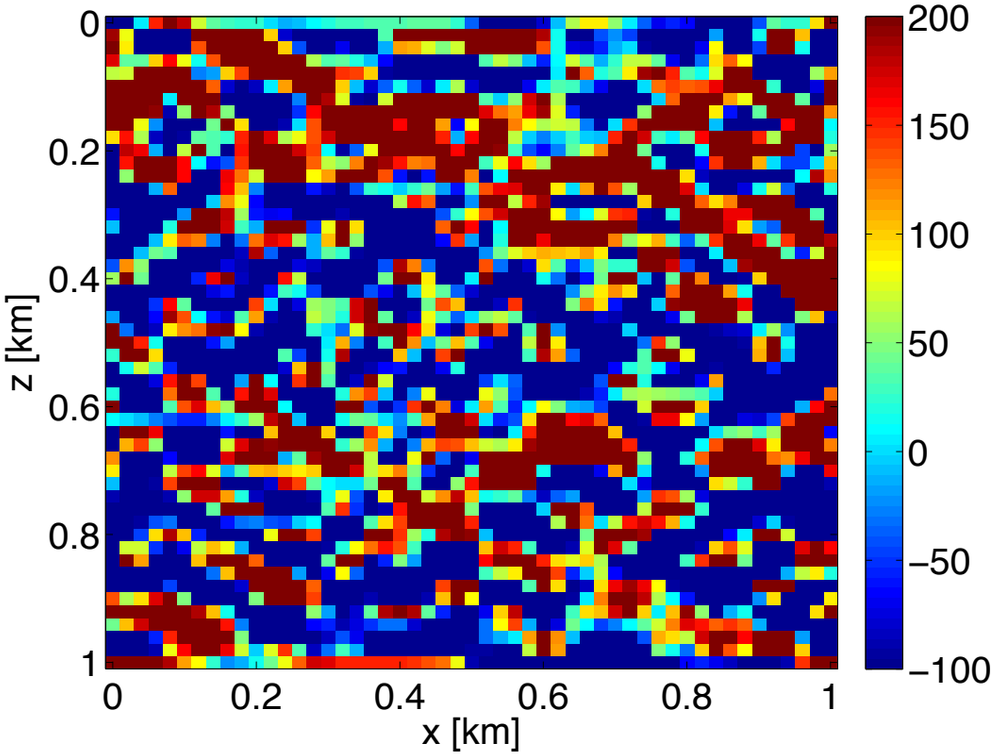


outliers in the data

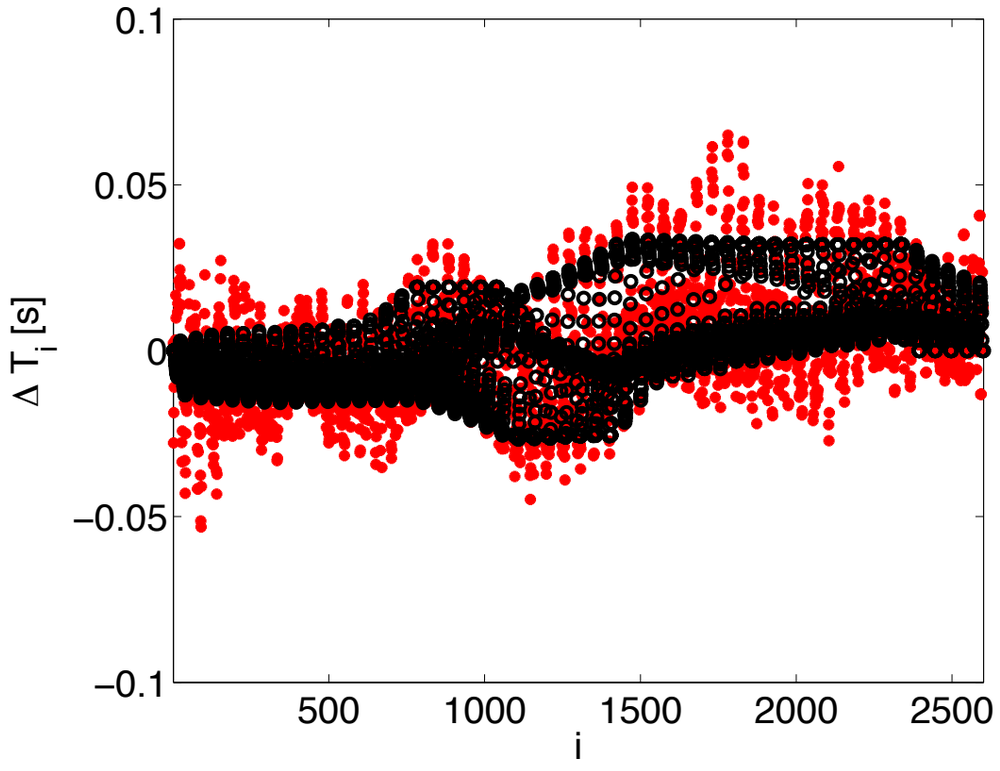


Application II: Results for traveltime tomography

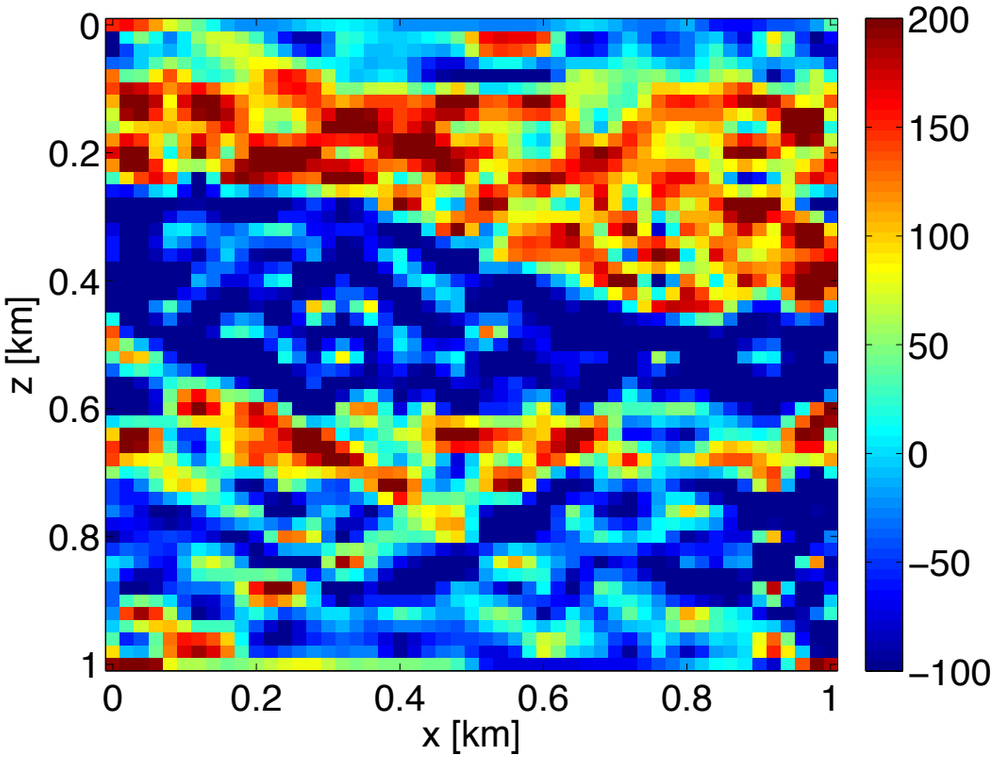
LS results



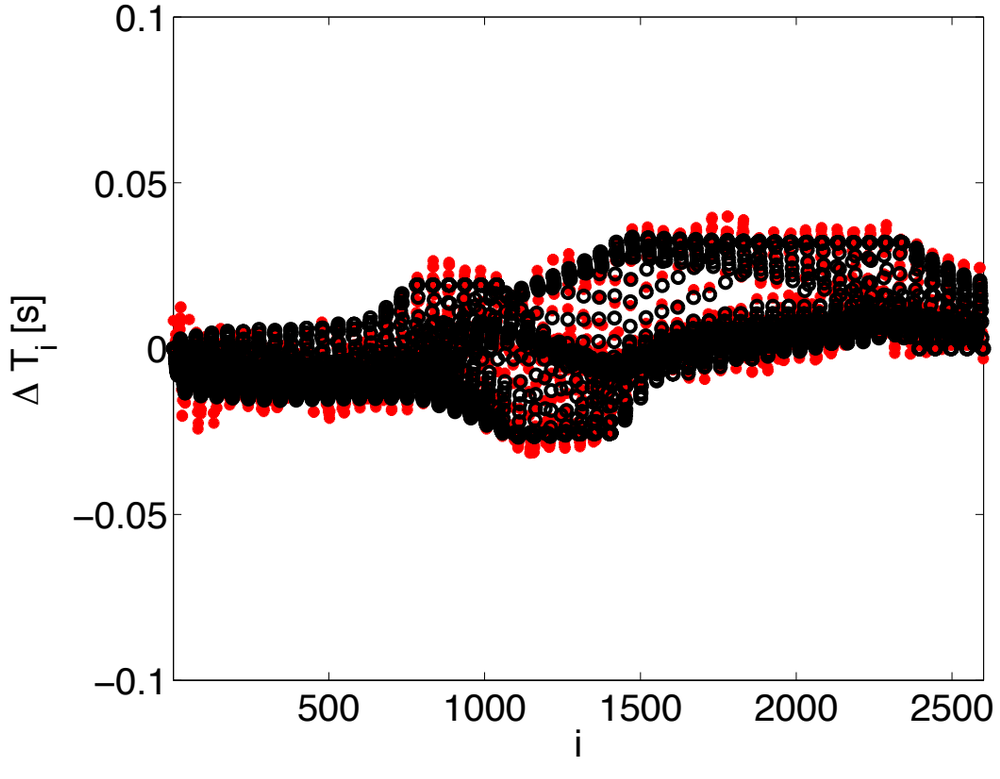
true and predicted data



ST results, ν, σ^2 from initial residual

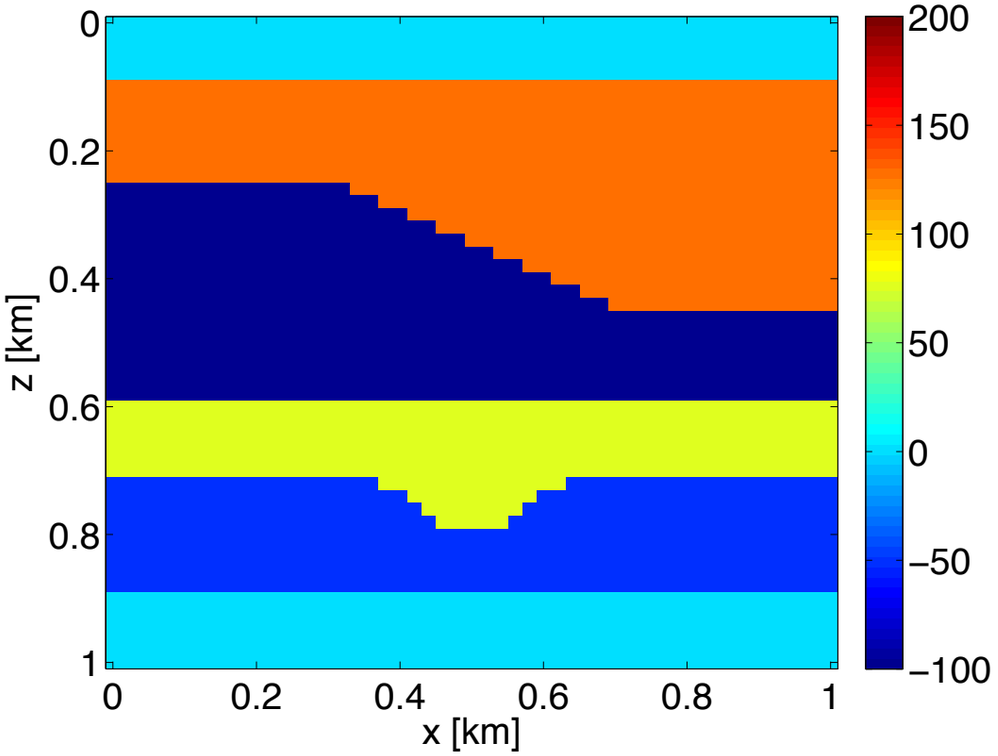


true and predicted data

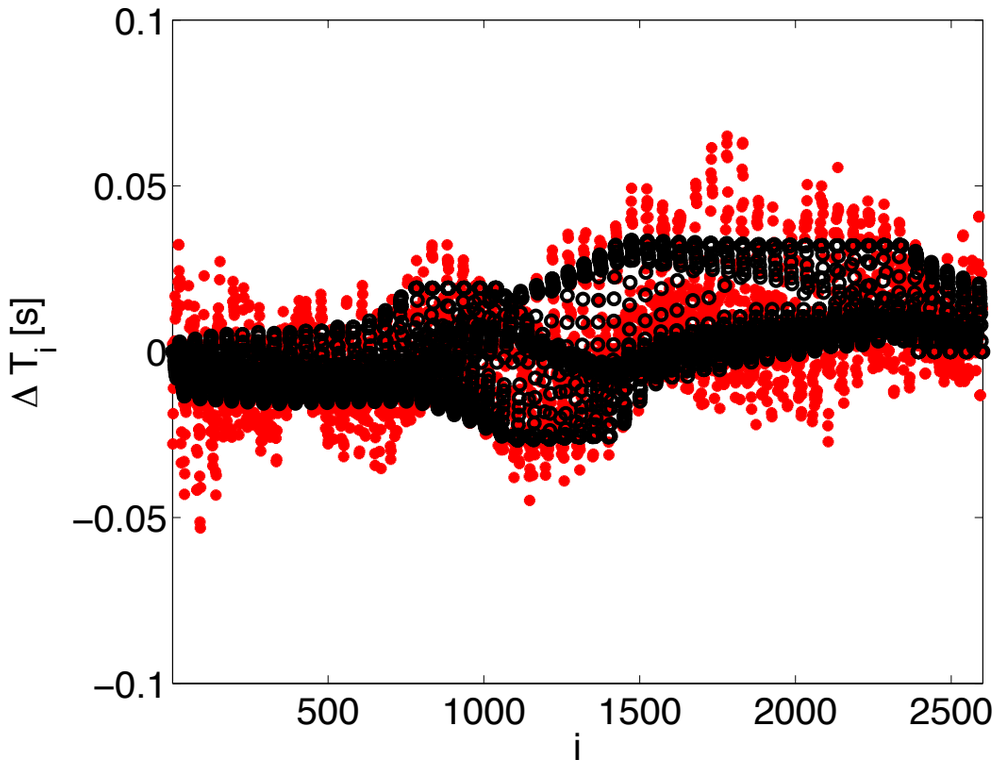


Application II: Results for traveltime tomography

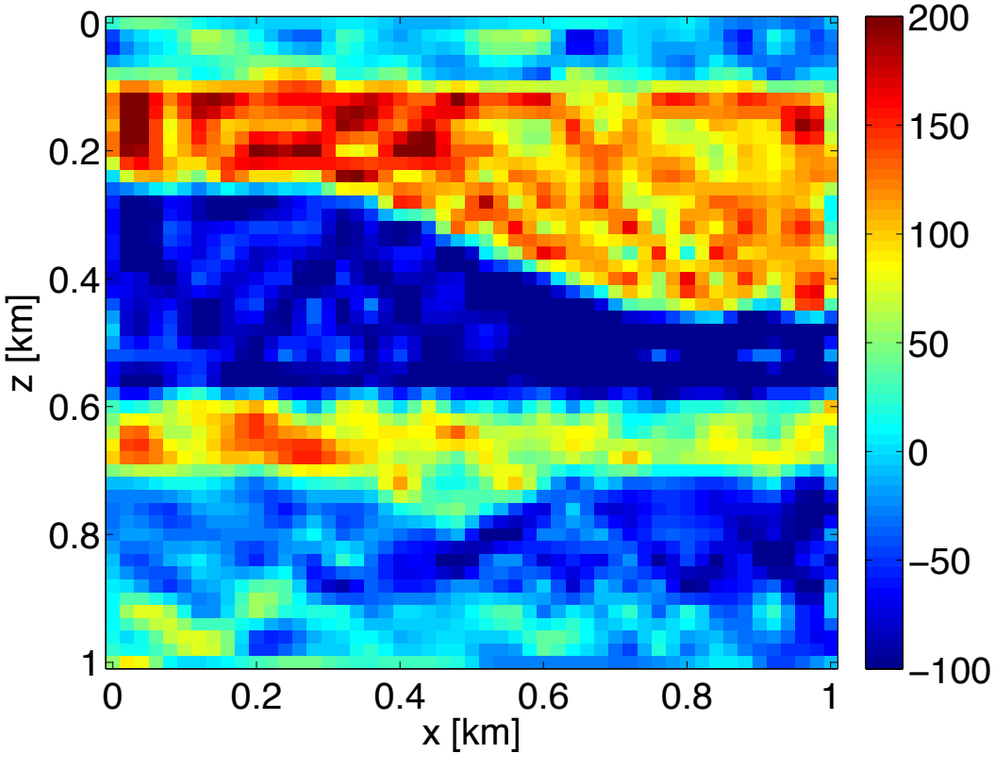
True perturbation



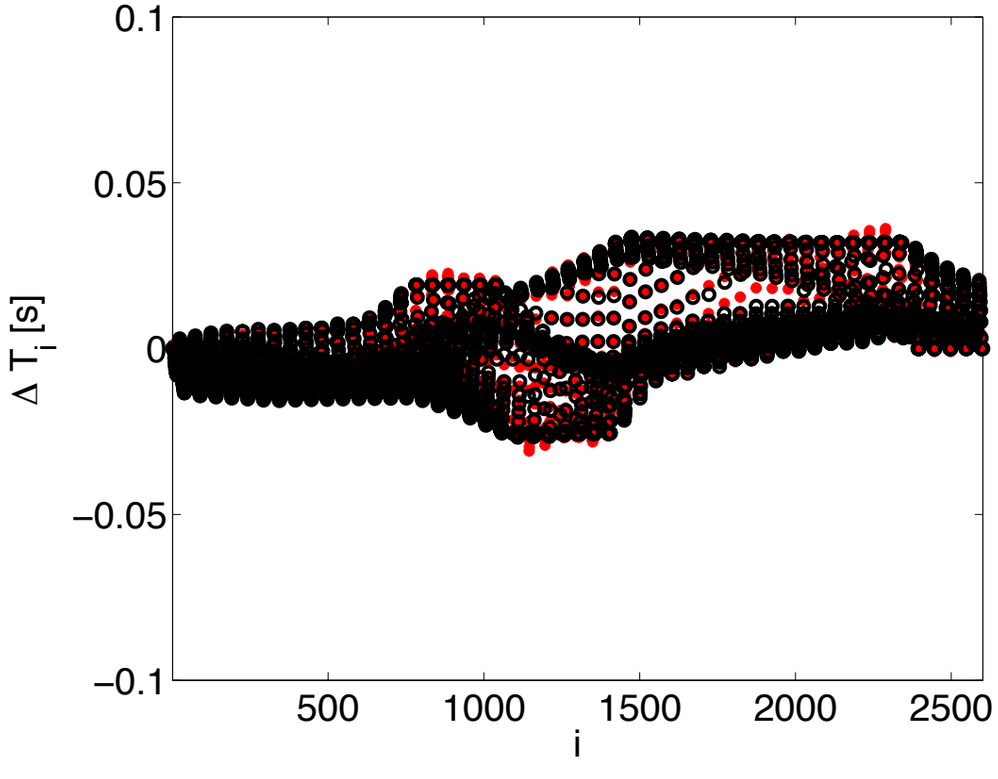
outliers in the data



ST results, ν, σ^2 estimated



true and predicted data



Thank you!

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