Tutorial: Randomized Algorithms in Exploration Seismology

Felix J. Herrmann

SLIM
University of British Columbia

Tutorial: Randomized batching in FWI

Tristan van Leeuwen, Sasha Aravkin, and Michael Friedlander



SLIM
University of British Columbia

Today's agenda

Goals are to

- introduce of *fast* simulation & optimization framework
- borrow ideas from theoretical computer science (machine learning & stochastic optimization)
- to limit IO, which is bound to become the bottle neck
- come up with approaches that are agnostic to type of wave-equation based inversions
- *integrate* these into a *versatile* computational FWI framework that *spends your computational resources only when needed…*

Disclaimer

This tutorial

- does not address ill-conditioning issues of full-waveform inversion
- is not about compressive sensing
- describes work in progress that remains to be tested on field data in 3D
- codes are not yet optimized
- based on time-harmonic formulation but does not rely on it
- statements only valid to 8Hz or so

Lots of room for improvements…

Seismic Laboratory for Imaging and Modeling |

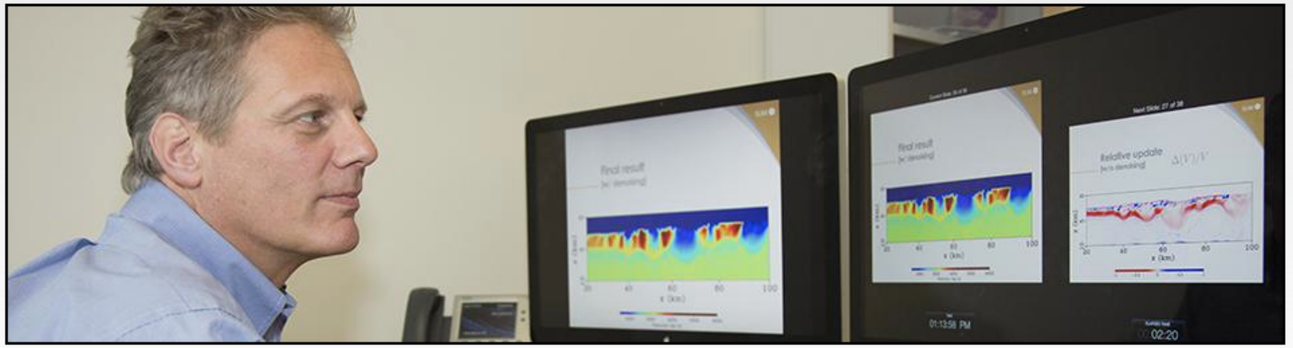
https www.slim.eos.ubc.ca

Reader

News Apple Reviewing Funding Professional SLIM EOS Fun Software Research Hobbies UBC Teaching Travel Research Agenda Your Telecom

Seismic Laboratory for Imaging and Modeling

ABOUT US | PROJECTS | EVENTS | PUBLICATIONS | RESEARCH | SOFTWARE | CONSORTIUM | INTERNAL



SEARCH

Upcoming events

Mon, Aug 31st, 2015
Inaugural Full-Waveform Inversion Workshop, Brazil

Wed, Sep 9th, 2015
Hansruedi Maurer, ETH Zurich "The curse of dimensionality in exploring the subsurface" 4:00 PM, ESB 5104 - 2207 Main Mall, UBC Campus

more

SINBAD Consortium Meeting Fall 2015

New Publications

- **Affordable full subsurface image volume—an application to WEMVA** *Conference (EAGE Workshop on Wave Equation based Migration Velocity Analysis, Madrid)*
- **Irregular grid tensor completion** *Conference (Workshop on Low-rank Optimization and Applications, University of Bonn, Germany)*
- **Wavefield-denoising and source encoding** *Conference (SIAM Conference on Mathematical and Computational Issues in the Geosciences, Stanford University, California)*
- **Sparsity promoting seismic imaging and full-waveform inversion** *Thesis (PhD)*
- **Total variation regularization strategies in full waveform inversion for improving robustness to noise, limited data and poor initializations** *Tech Report*
- **Sparse least-squares seismic imaging with source estimation utilizing multiples** *Conference (PIMS Workshop on Advances in Seismic Imaging and Inversion, University of Alberta, Edmonton)*
- **A new take on compressive time-lapse seismic acquisition, imaging and inversion** *Conference (PIMS Workshop on Advances in Seismic Imaging and Inversion, University of Alberta, Edmonton)*
- **Compressive time-lapse seismic data processing using shared information** *Conference (CSEG,*

Biblio | Seismic Laboratory for Imaging and Modeling

https www.slim.eos.ubc.ca/biblio

Reader

News Apple Reviewing Funding Professional SLIM EOS Fun Software Research Hobbies UBC Teaching Travel Research Agenda Your Telecom

Seismic Laboratory for Imaging and Modeling

ABOUT US PROJECTS EVENTS PUBLICATIONS RESEARCH SOFTWARE CONSORTIUM INTERNAL

Biblio

SEARCH

List Filter

SEARCH PUBLICATIONS

Sort by: Author Keyword Title Type [Year]

Export 64 results: BibTex

Search results for *fwi* [Reset Search]

2015

Philipp Witte, Mathias Louboutin, and Felix J. Herrmann, "**Overview on anisotropic modeling and inversion**". 2015. Abstract BibTex

Felix J. Herrmann and Bas Peters, "**Pros and cons of full- and reduced-space methods for Wavefield Reconstruction Inversion**", in *SIAM Conference on Mathematical and Computational Issues in the Geosciences*, 2015. Abstract BibTex

Brendan Smithyman, Bas Peters, and Felix J. Herrmann, "**Constrained waveform inversion of colocated VSP and surface seismic data**", in *EAGE Annual Conference Proceedings*, 2015. Abstract BibTex

Zhilong Fang, Chia Ying Lee, Curt Da Silva, Felix J. Herrmann, and Rachel Kuske, "**Uncertainty quantification for Wavefield Reconstruction Inversion**", in *EAGE Annual Conference Proceedings*, 2015. Abstract BibTex

Felix Oghenekohwo, Rajiv Kumar, Ernie Esser, and Felix J. Herrmann, "**Using common information in compressive time-lapse full-waveform inversion**", in *EAGE Annual Conference Proceedings*, 2015. Abstract BibTex

Felix J. Herrmann, "**Randomized algorithms in exploration seismology**", in *ASEG Annual Conference Proceedings*, 2015. Abstract BibTex

Mathias Louboutin and Felix J. Herrmann, "**Time compressively sampled full-waveform inversion with stochastic optimization**". 2015. Abstract BibTex

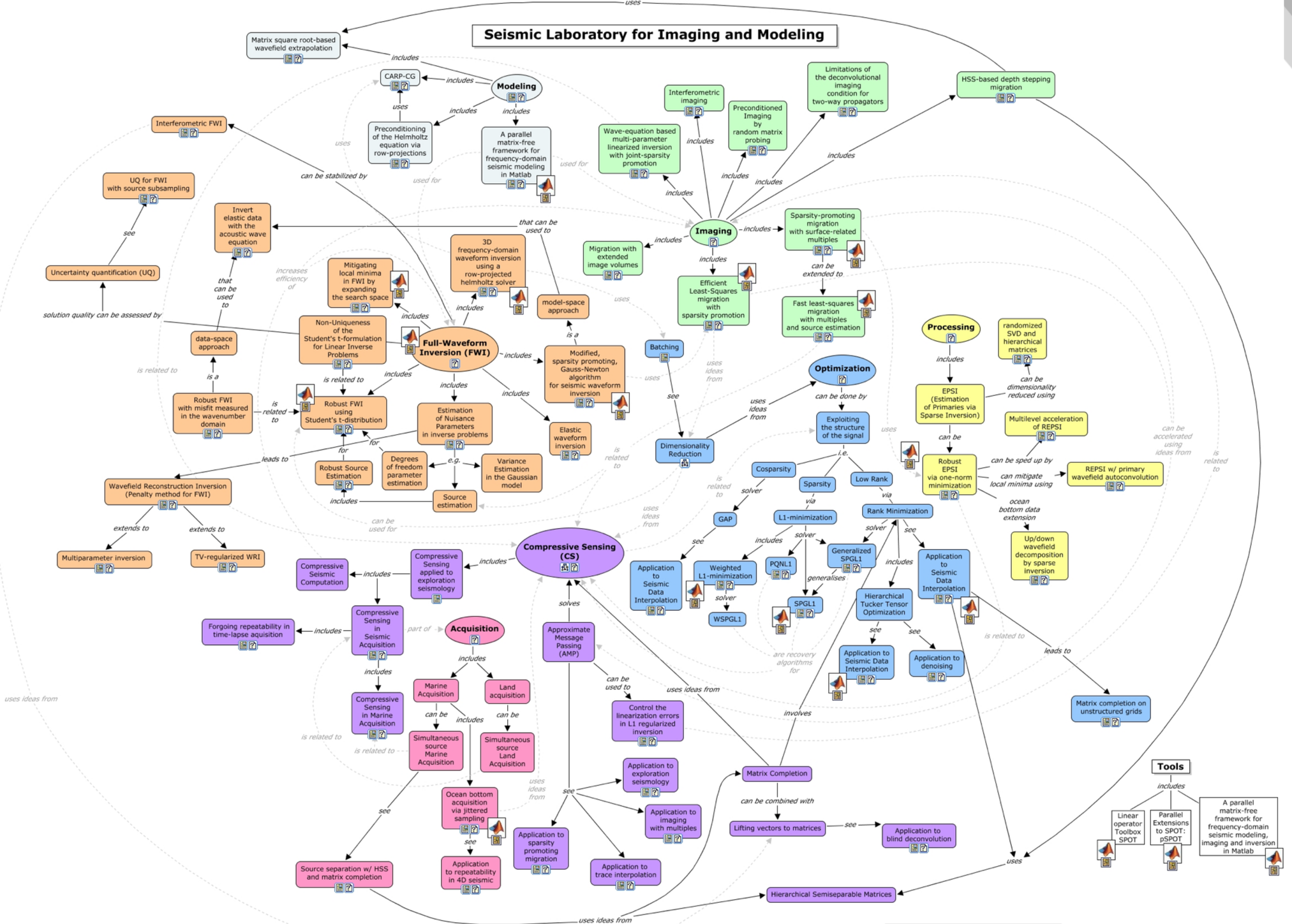
2014

Felix J. Herrmann, Ernie Esser, Tristan van Leeuwen, and Bas Peters, "**Wavefield Reconstruction Inversion (WRI) – a new take on wave-equation based inversion**", in *SEG Workshop on Full Waveform Inversion - Elastic Approaches and Issues with Anisotropy, Nonshallow Inversion, Poor Starting Model; Denver*, 2014. BibTex

Rafael Lago, Art Petrenko, Zhilong Fang, and Felix J. Herrmann, "**Fast solution of time-harmonic wave-equation for full-waveform inversion**", in *EAGE Annual Conference Proceedings*, 2014. Abstract BibTex

Zhilong Fang, Curt Da Silva, and Felix J. Herrmann, "**Fast uncertainty quantification for 2D full-waveform inversion with randomized source subsampling**", in *EAGE Annual Conference Proceedings*, 2014. Abstract BibTex

Seismic Laboratory for Imaging and Modeling



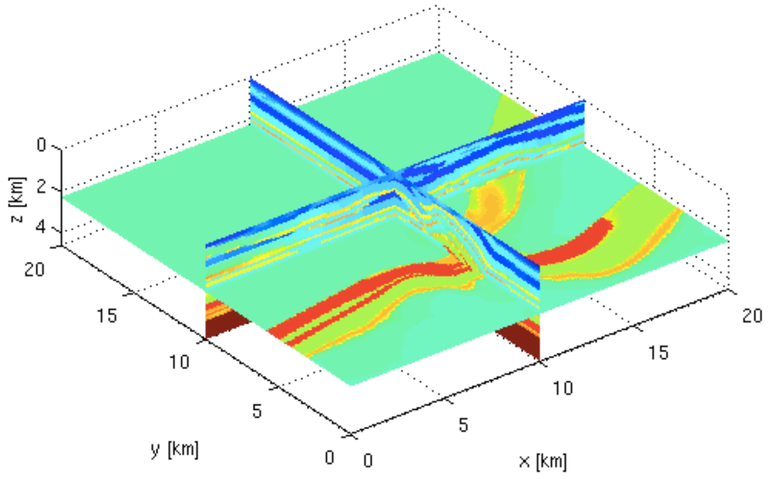
SLIM

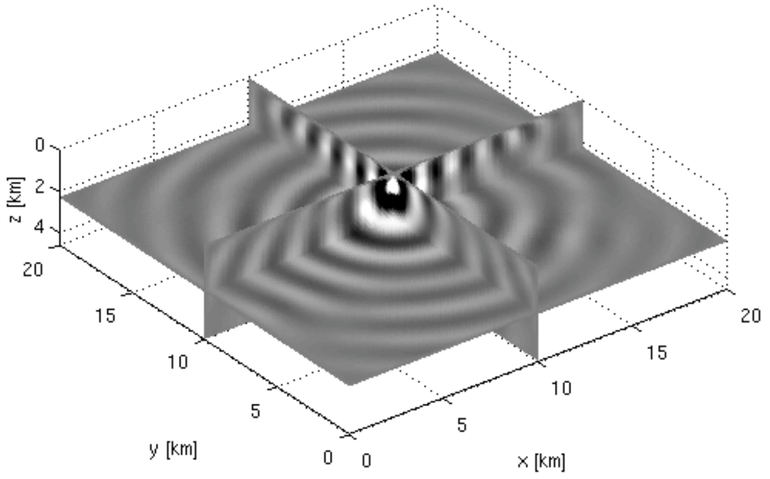
3D Frequency-domain FWI with batching: results

Contents

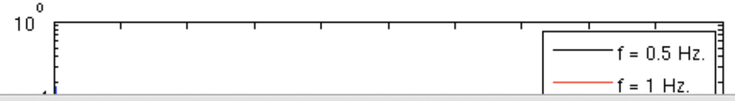
CARP-CG

Here we present some results of the Helmholtz solver on the overthrust model. The model and a wavefield for 2 Hz are shown below.





We compute the wavefield for various frequencies with a fixed number of gridpoints per wavelength. The convergence histories are shown below



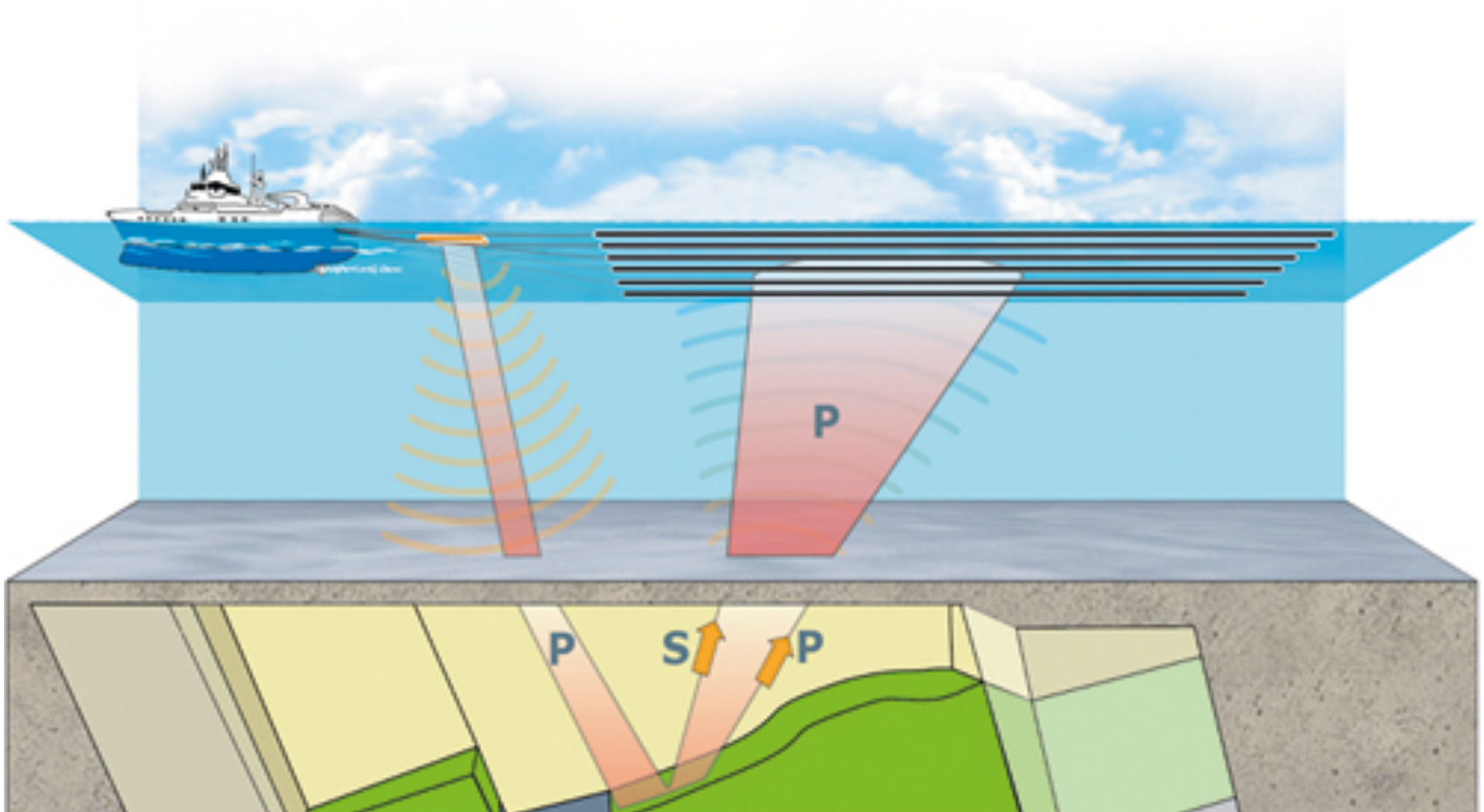
Introduction to full-waveform inversion

SLIM
University of British Columbia

Tristan van Leeuwen and Felix J. Herrmann, "**Fast waveform inversion without source encoding**", *Geophysical Prospecting*, vol. 61, p. 10-19, 2013.
Tristan van Leeuwen and Felix J. Herrmann, "**3D frequency-domain seismic inversion with controlled sloppiness**", *SIAM Journal on Scientific Computing*, vol. 36, p. S192-S217, 2014.

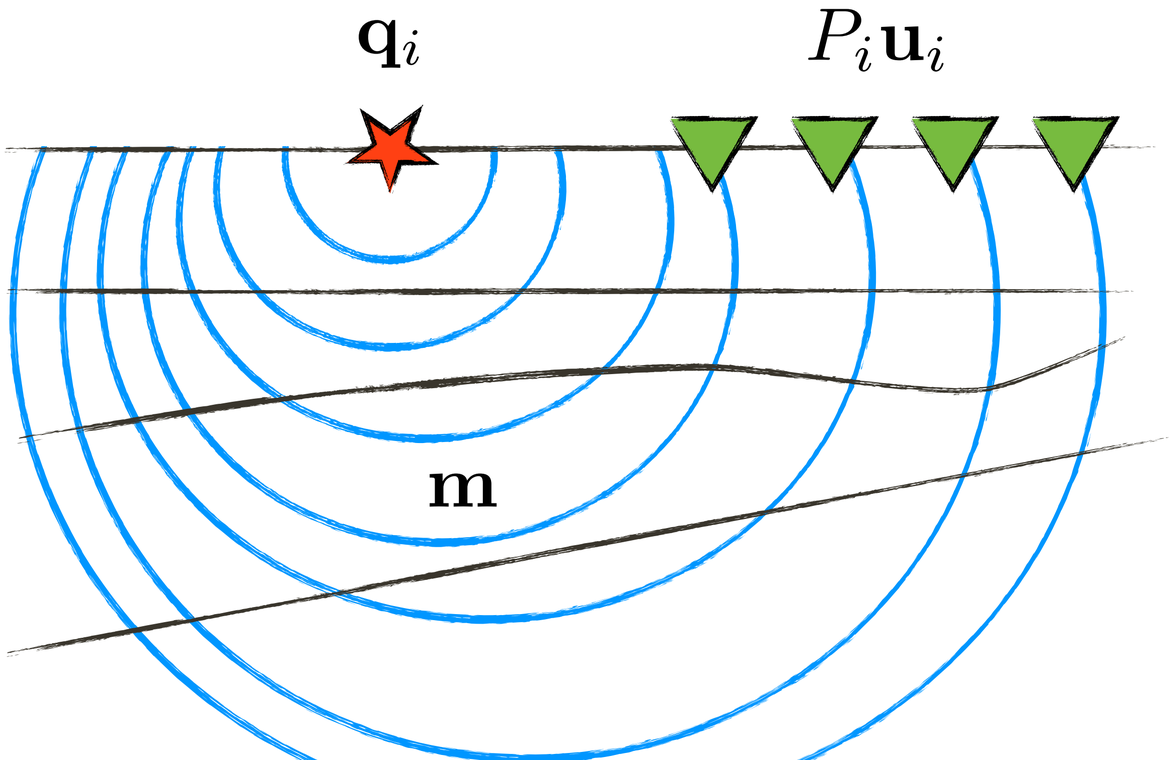
Infer 3D velocity model from *multi-experiment* data:

- $\mathcal{O}(10^9)$ unknowns
- $\mathcal{O}(10^{15})$ datapoints
- propagate $\mathcal{O}(10^2)$ wavelengths



Waveform inversion

Retrieve the medium parameters from partial measurements of the solution of the wave-equation: $A(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$



Discretize then optimize

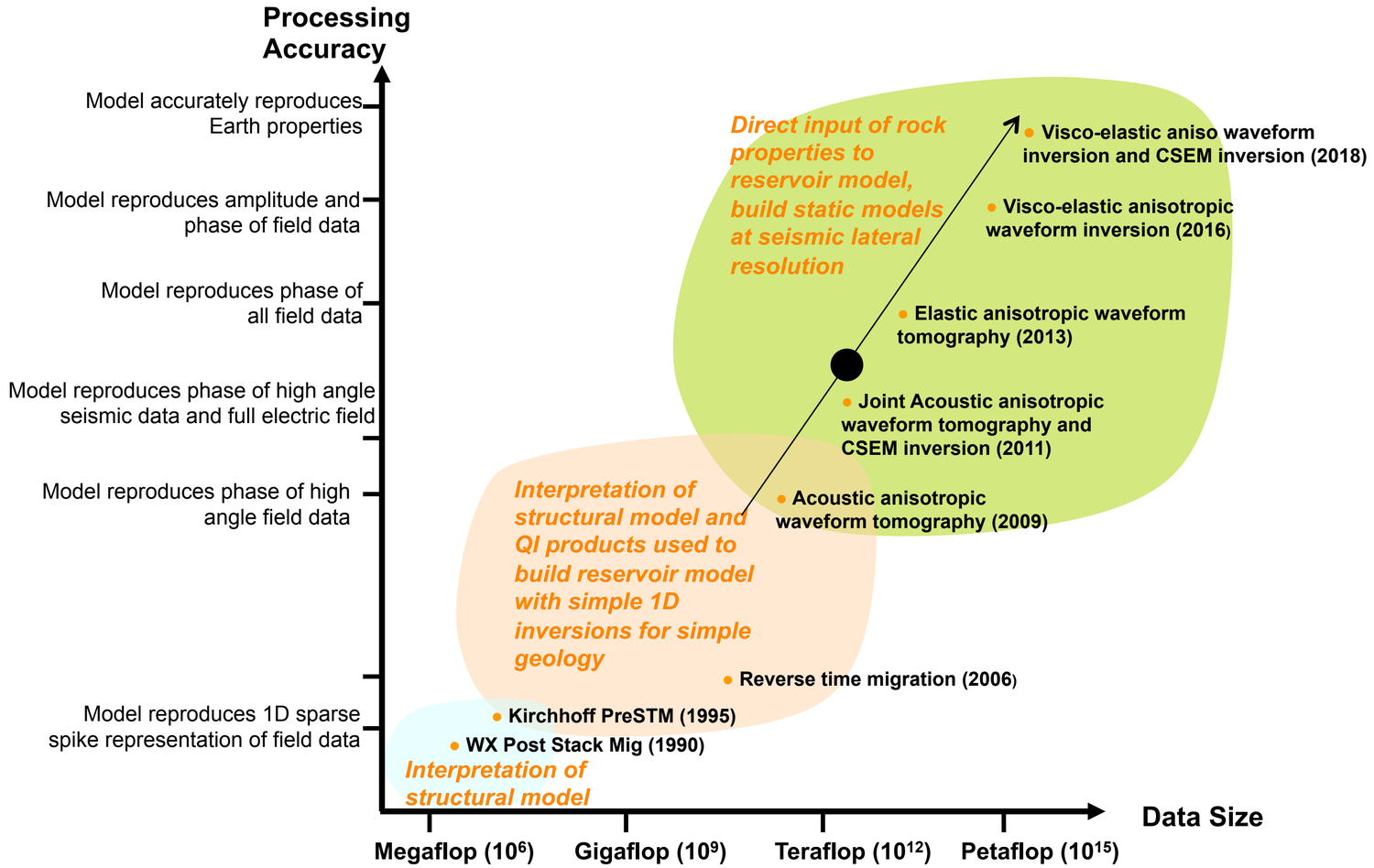
Discretize time-harmonic wave equation

$$\left[\frac{\omega^2}{c(x)^2\rho(x)} + \nabla\cdot\frac{1}{\rho(x)}\nabla\right] u(\omega,x) = q(\omega,x) \;+\; \text{b.c.}'\text{s},$$

into system of equations

$$A\mathbf{u} = \mathbf{q}$$

Computational costs



courtesy. BG Group

Challenges

Computational costs increase

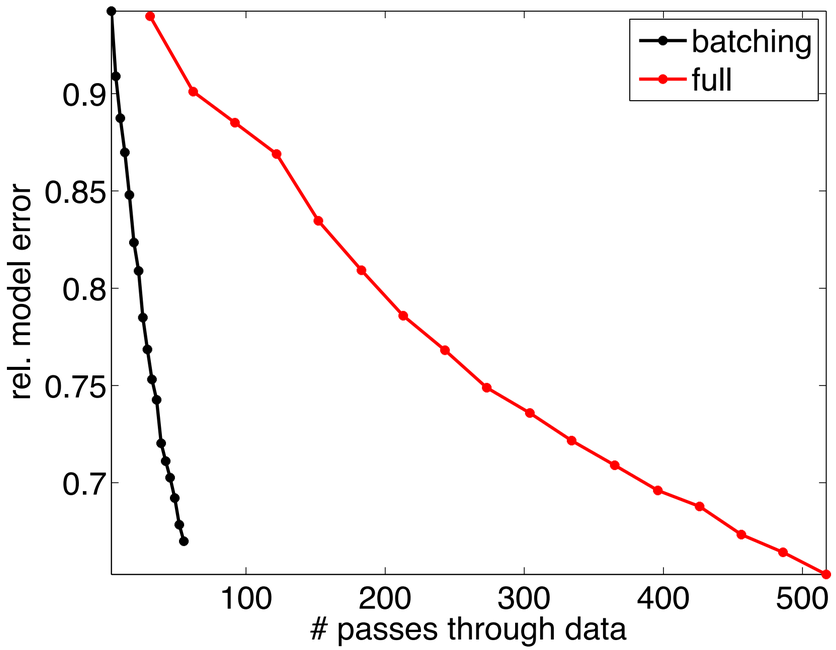
- linearly w/ # of sources
- exponentially with sample density, frequency & survey area

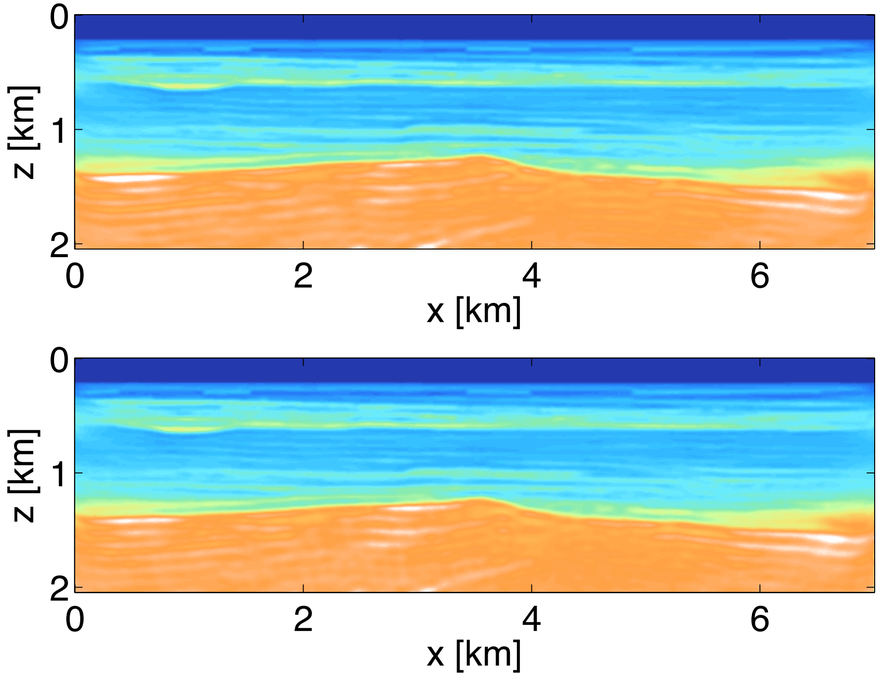
Move to 3D elastic

- sky rocketing costs (X 1000)
- can no longer be met by Moore's law…

Fast randomized optimization

10 x speedup





[van Leeuwen & FJH '11]

Batched randomized computations by industry

thanks to Denes Vigh & Nick Moldoveanu

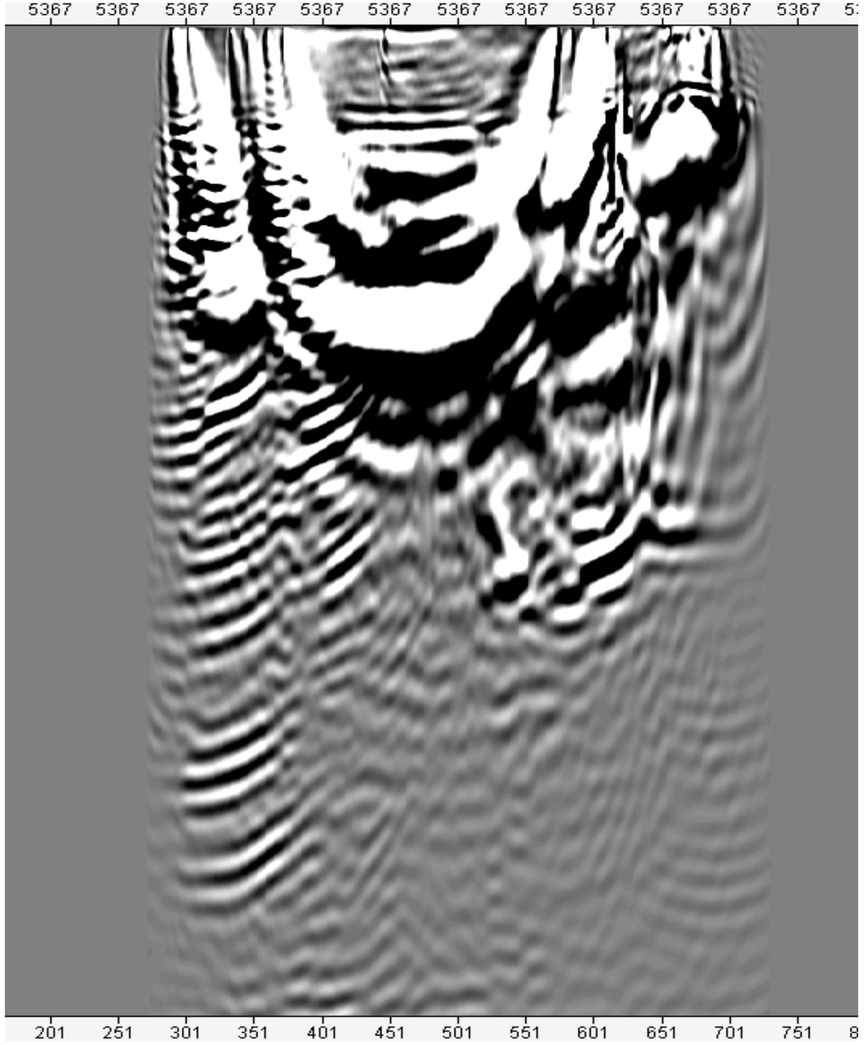
SLIM
University of British Columbia

All Shots

Total NO. Shots = 1749

All shots selected

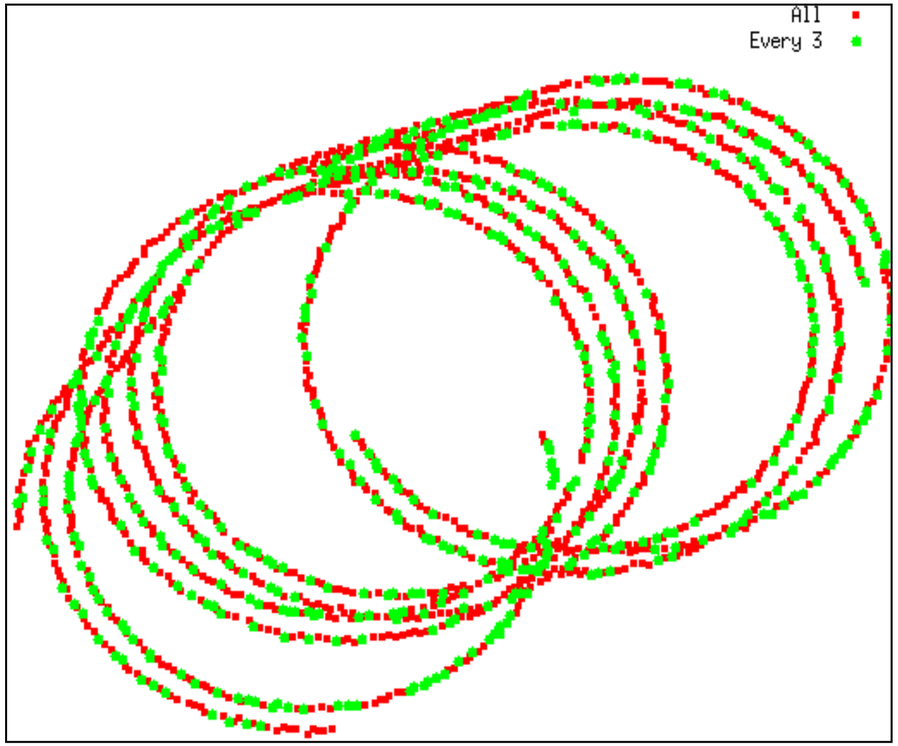
Gradient



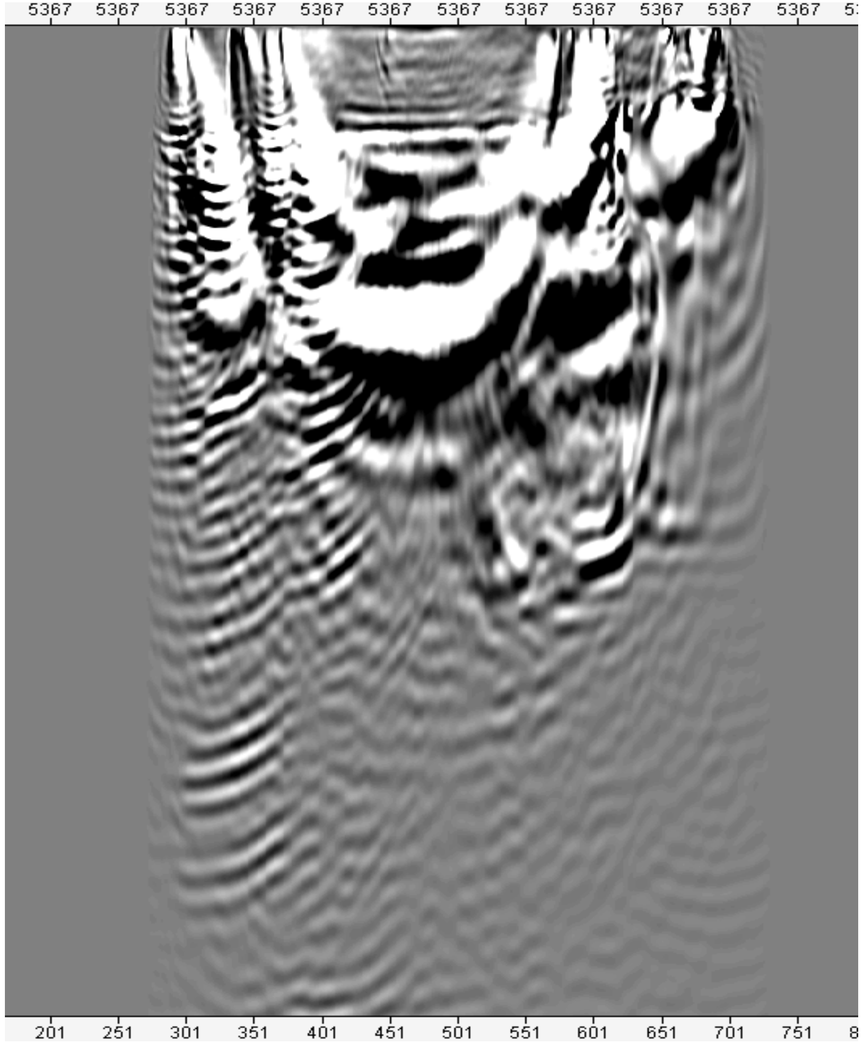
Fixed-increment sampling

Total NO. Shots = 1749

Periodic w/ inc 3 NO. Shots =584



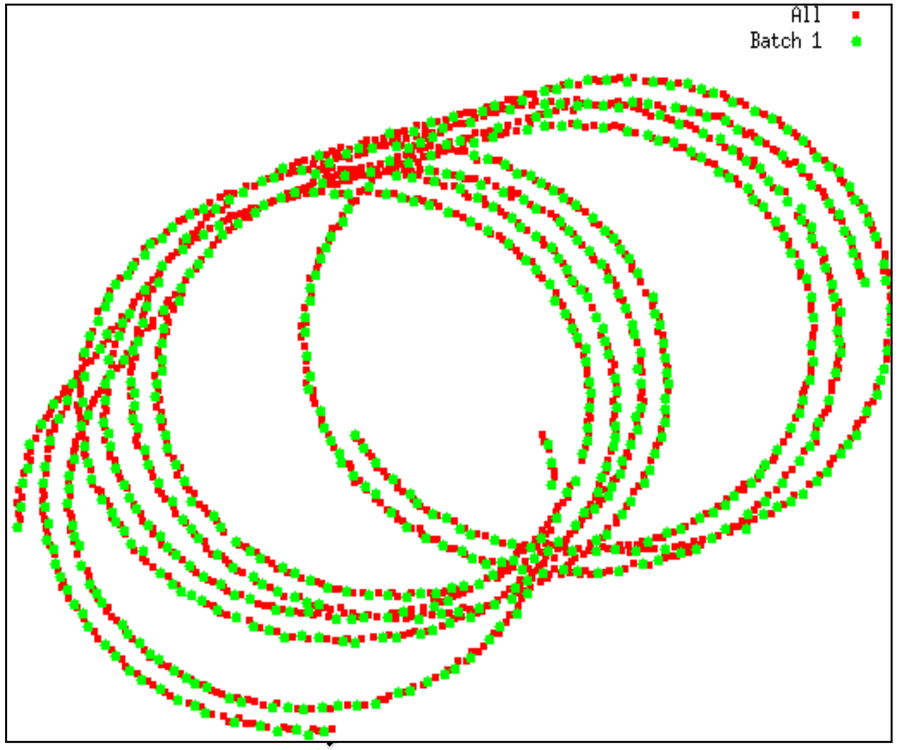
Gradient



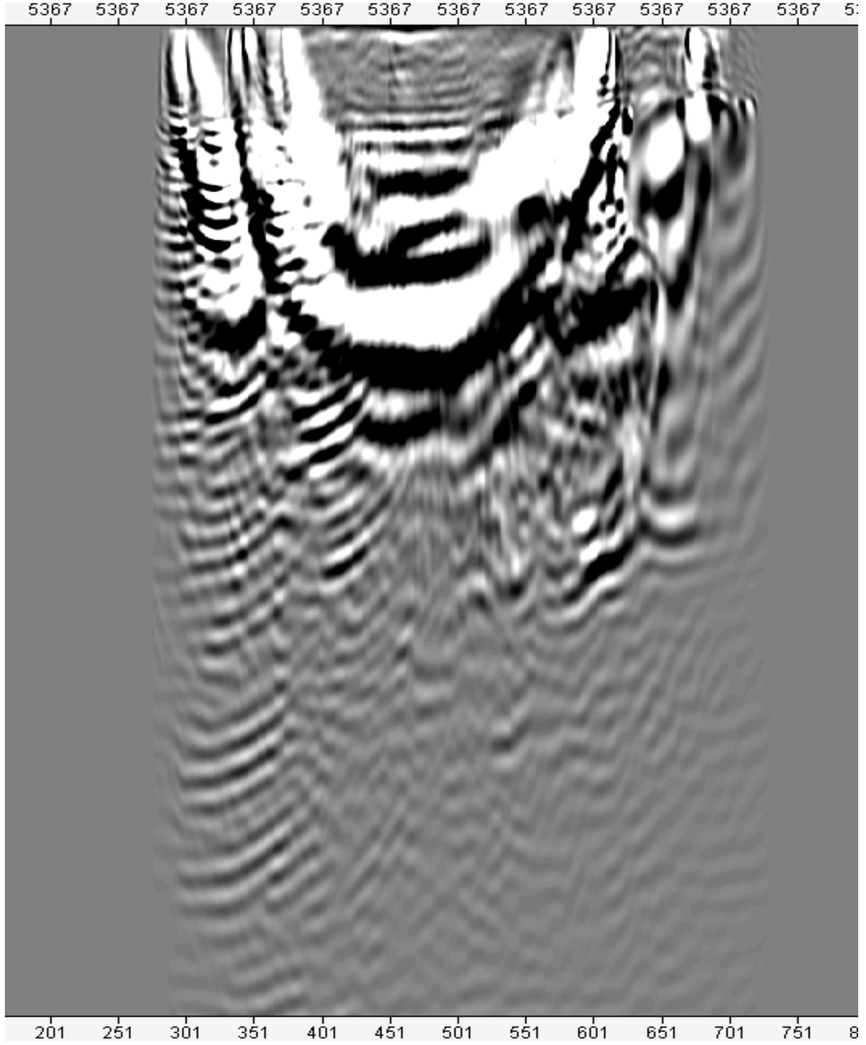
Random-batch sampling 1

Total NO. Shots = 1749

Batch 1 NO. Shots =612



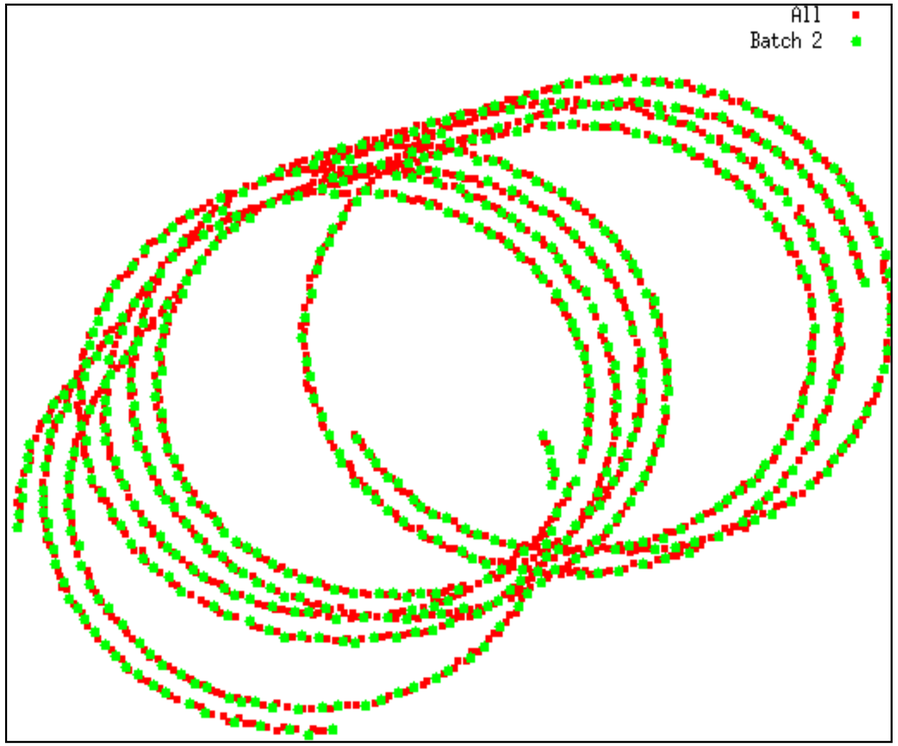
Gradient



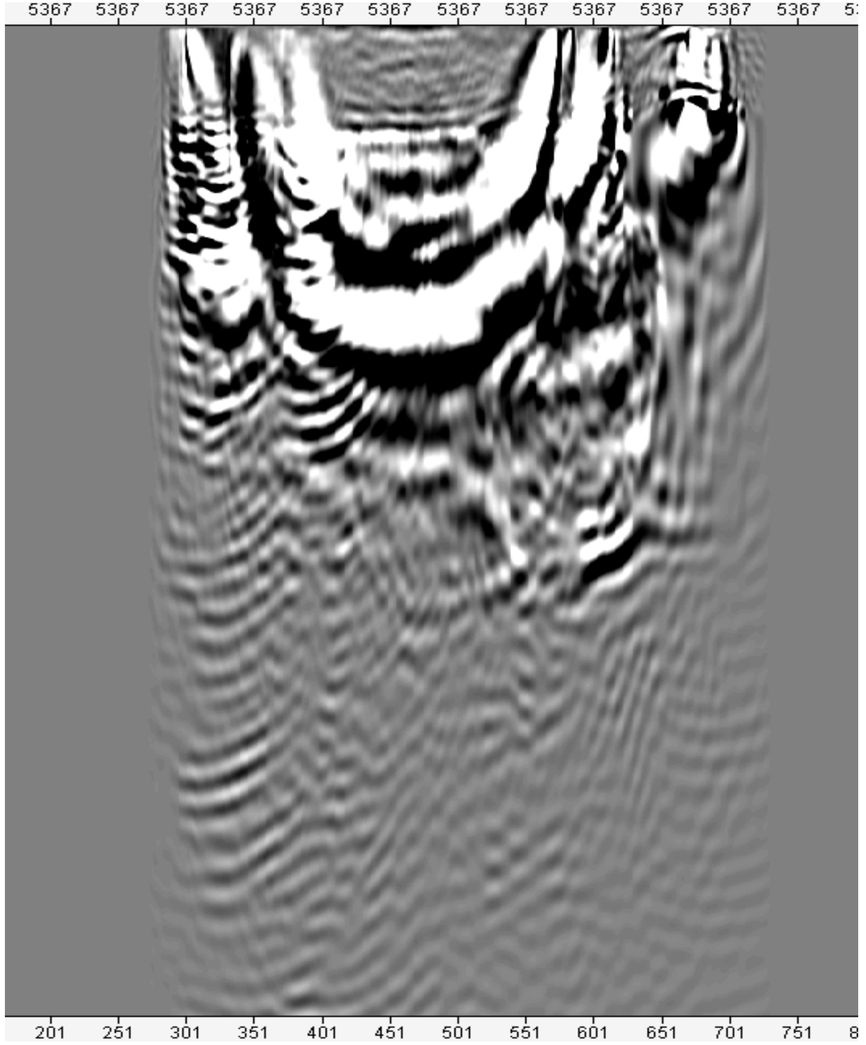
Random-batch sampling 2

Total NO. Shots = 1749

Batch 2 NO. Shots =604



Gradient



Industry uptake

Has resulted in 4X–7X in crease computational efficiency

Is making the difference between rendering a service w/ or w/o a profit

[Heinkenschloss, '98 , Haber, '00]

PDE-constrained optimization
all-at-once full-space approach

simulated data

simulated wavefield

$$\min_{\mathbf{m},\mathbf{u}} \sum_{i=1}^{M} \|P_i\mathbf{u}_i - \mathbf{d}_i\|_2^2 \quad \text{s.t.} \quad A_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$$

observed data

Helmholtz equation

source

- avoids having to solve the PDE explicitly
- sparse (GN) Hessian
- requires storing all variables (**m,u**)
- does **not** scale to industry-scale seismic problems

Lagrangian

misfit functional

Solve $$\min_{\mathbf{m},\mathbf{w},\mathbf{u}} \quad \sum_{i=1}^{M} \rho\big(w_i P_i \mathbf{u}_i - \mathbf{d}_i\big) \quad \text{s.t.} \quad A(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$$

by forming $$L(\mathbf{m},\mathbf{w},\mathbf{u},\mathbf{v}) = \sum_{i=1}^{M} \rho\big(w_i P_i \mathbf{u}_i - \mathbf{d}_i\big) + \mathbf{v}_i^* \big(A(\mathbf{m})\mathbf{u}_i - \mathbf{q}_i\big)$$

and elimination of the wavelet and state variables from

$$\nabla_{\mathbf{u}_i} L = P_i^* \nabla \rho\big(w_i P_i \mathbf{u}_i - \mathbf{d}_i\big) + A(\mathbf{m})^* \mathbf{v}_i,$$

$$\nabla_{\mathbf{v}_i} L = A(\mathbf{m})\mathbf{u}_i - \mathbf{q}_i,$$

$$\nabla_{w_i} L = \big(P_i \mathbf{u}_i\big)^* \nabla \rho\big(w_i P_i \mathbf{u}_i - \mathbf{d}_i\big)$$

$$\nabla_{\mathbf{m}} L = \sum_{i=1}^{M} G(\mathbf{m},\mathbf{u}_i)^* \mathbf{v}_i,$$

Unconstrained reduced formulation

by solving

$$A(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i$$

$$A(\mathbf{m})^H\mathbf{v}_i = P_i^T(\mathbf{d}_i - P_i\mathbf{u}_i)$$

$$w_i = \underset{w}{\operatorname{argmin}}\, \rho\big(wP_i\mathbf{u}_i - \mathbf{d}_i\big).$$

yielding the reduced objective and its gradient

$$\phi(\mathbf{m}) = \sum_{i=1}^{M} \phi_i(\mathbf{m}), \qquad \phi_i(\mathbf{m}) = \rho\big(w_iP_i\mathbf{u}_i - \mathbf{d}_i\big)$$

$$\nabla\phi(\mathbf{m}) = \sum_{i=1}^{M} \nabla\phi_i(\mathbf{m}), \qquad \nabla\phi_i(\mathbf{m}) = \omega^2\mathrm{diag}(\mathbf{u}_i)^*\mathbf{v}_i$$

[Tarantola ’84; Pratt, ’98; Plessix, ’06]

Adjoint-state/reduced-space formulation

Eliminated the constraint

$$\min_{\mathbf{m}} \phi_{\text{red}}(\mathbf{m}) = \sum_{i=1}^{M} \| P_i A_i(\mathbf{m})^{-1} \mathbf{q}_i - \mathbf{d}_i \|_2^2$$

- no need to store all wavefields (block-elimination)
- suitable for black-box optimization (e.g., l-BFGS)
- need to solve forward & adjoint PDEs
- very non-linear dependence on earth model (**m**)
- dense (GN) Hessian, involves additional PDE solves
- **reliance on accurate starting models to avoid cycle skipping**

Frugal FWI

Felix J. Herrmann and Tristan van Leeuwen

Felix J. Herrmann, Andrew J. Calvert, Ian Hanlon, Mostafa Javanmehri, Rajiv Kumar, Tristan van Leeuwen, Xiang Li, Brendan Smithyman, Eric Takam Takougang, and Haneet Wason, "**Frugal full-waveform inversion: from theory to a practical algorithm**", *The Leading Edge*, vol. 32, p. 1082-1092, 2013.

Tristan van Leeuwen and Felix J. Herrmann, "**3D frequency-domain seismic inversion with controlled sloppiness**", *SIAM Journal on Scientific Computing*, vol. 36, p. S192-S217, 2014.

SLIM

University of British Columbia

Frugal optimization

Challenge: Computational costs of FWI scale w/ # of sources

Strategy:

- *reduce* costs by working w/ random *subsets* of sources
- allow for *inaccurate* physics (e.g., PDE solves)
- *convergence* guarantees via *dynamic* accuracy *control*
- *dynamic* increase *size* subsets & *accuracy* PDE solves

Outcome:

- computationally *affordable* scheme for FWI

Framework

Iterative Helmholtz solver:

- w/ low memory imprint & computational overhead, e.g. setup costs
- converges w/o model-dependent tuning

Practical stopping criterion:

- avoid accurate solutions when when model iterate far from true solution

Stochastic optimization strategy:

- exploit separable structure by working w/ small subsets of RHS's
- adaptively grows sample-size} as the optimization proceeds, and

Exploit parallelism:

- model-space via domain decomposition
- data-space parallelism via loops over frequency and/or RHS's

Frugal FWI

$$\min_{\mathbf{m}} \rho(F(\mathbf{m}) - \mathbf{d})$$

robust
formulation

$$A(\mathbf{m})\mathbf{u} = \mathbf{q}$$

versatile
modelling

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \mathbf{s}_k$$

fast optimization strategies

computational framework

Versatile modelling

Challenge:

- FWI modelling problems are large ($\sim 10^9$) gridpoints
- time-harmonic systems are increasingly indefinite for high frequencies

Strategy:

- avoid large *setup, memory* costs & *tuning* parameters
- offer *control* on *precision* wave simulations
- by *increasing* number of *iterations* indirect Krylov solvers

Outcomes:

- *scalable* parallel wave simulations w/ prescribed *tolerance*
- simple *preconditioner* that works for different WE's

fast, complicated,..

VS.

simple, robust, ...

Kaczmarz

The Kaczmarz method solves a system $A\mathbf{x} = \mathbf{b}$ by successive row projections

$$\mathbf{x} := \mathbf{x} + \frac{\lambda_i}{||\mathbf{a}_i||_2^2} \left(b_i - \mathbf{a}_i^T \mathbf{x}\right) \mathbf{a}_i,$$

with relaxation parameter $0 < \lambda_i < 2$

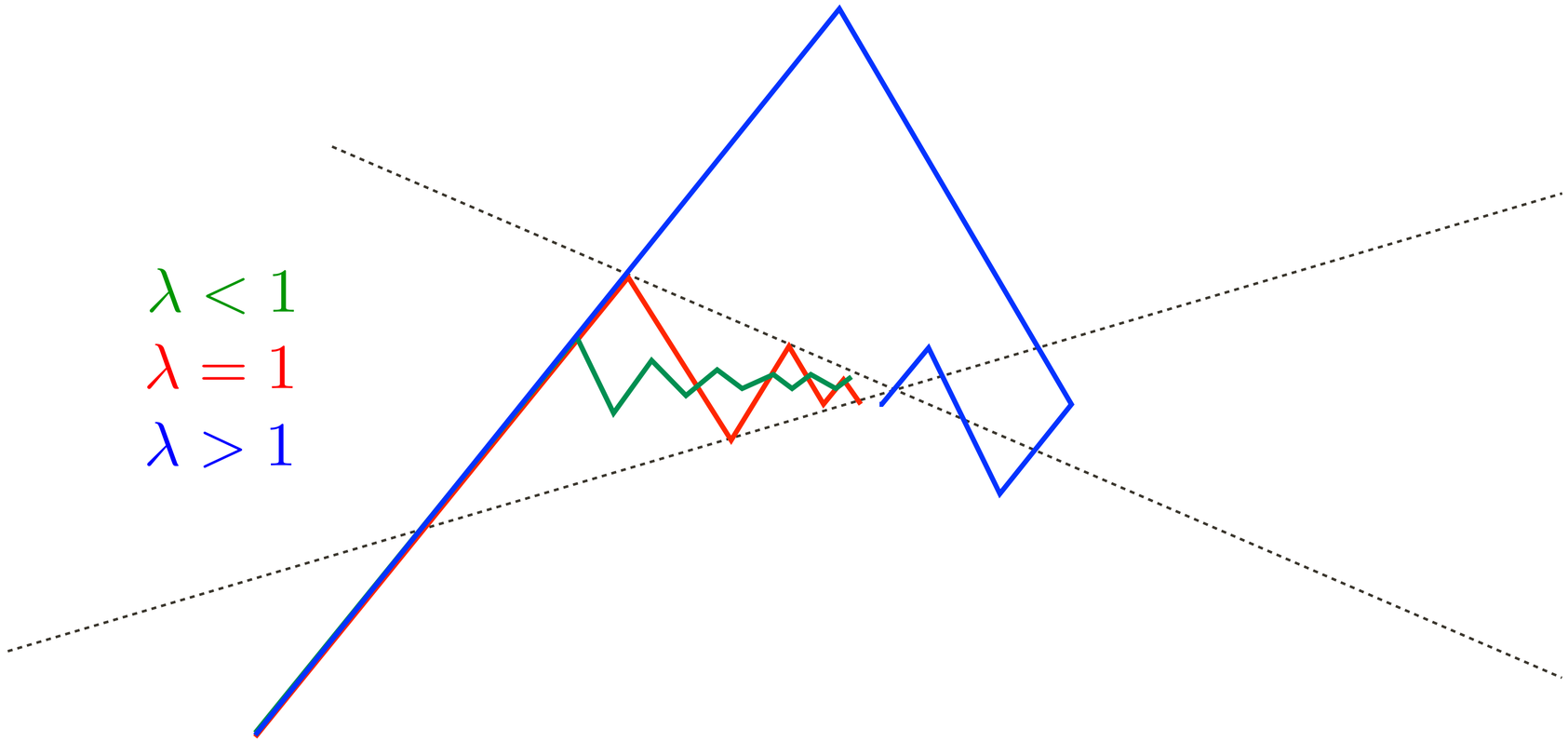
[Kaczmarz, ’37]

Kaczmarz

$\lambda < 1$

$\lambda = 1$

$\lambda > 1$



Kaczmarz

rewrite:

$$\mathbf{x} := \underbrace{\left(I - \frac{\lambda_i}{||\mathbf{a}_i||_2^2} \mathbf{a}_i \mathbf{a}_i^T \right)}_{Q_i} \mathbf{x} + \frac{\lambda_i}{||\mathbf{a}_i||_2^2} b_i \mathbf{a}_i$$

a double sweep yields

$$\mathbf{x} := \underbrace{(Q_1 Q_2 \dots Q_n Q_n \dots Q_1)}_{Q} \mathbf{x} + \underbrace{(\dots)}_{R} \mathbf{b}$$

Kaczmarz

Find a fixed point by solving

$$(I - Q)\mathbf{x} = R\mathbf{b}$$

where $I - Q$ is symmetric and positive semidefinite, so we can use CG (CGMN).

[Bjork & Elfving, ’79]

Gordon & Gordon '11-'12
van Leeuwen et. al. '12

CGMN & CARP-BCG

CGMN: use *simple* Kaczmarz *row* projections

$$\mathbf{u} := \mathbf{u} + \gamma \left(q_i - \mathbf{a}_i{}^* \mathbf{u}\right) \mathbf{a}_i / ||\mathbf{a}_i||_2^2, \quad i = 1 \ldots N,$$

or

$$\mathbf{u} := Q_i \mathbf{u} + \gamma q_i \mathbf{a}_i / ||\mathbf{a}_i||_2^2$$

to form a *preconditioner* deriving from

$$\mathbf{u} := Q\mathbf{u} + R\mathbf{q} \text{ where } Q = Q_1 Q_2 \ldots Q_N Q_N \ldots Q_1$$

with *double* sweeps that

- deals with *multiple* right-hand-sides *simultaneously*
- is *parallelizible* by *projecting* row blocks *independently*
- can be *accelerated* w/ CG

Preconditioned system

Use CG to invert positive semi-definite system of equations

$$(I - Q)\mathbf{u} = R\mathbf{q}$$

equivalent to SSOR on the normal equation AA^*

$$Q = I - A^* H A$$

$$R = A^* H$$

with

$$H = \gamma(2-\gamma)\left(D + \gamma L^*\right)^{-1} D \left(D + \gamma L\right)^{-1}$$

D and L contain the diagonal and lower triangular elements of AA^*

Double sweeps <=> $\mathbf{u} := Q\mathbf{u} + R\mathbf{s}$

Algorithm 1 DKSWP$(A, \mathbf{u}, \mathbf{s}, \gamma)$ Performs a forward and backward Kaczmarz sweep on the matrix A

// forward sweep
for $i = 1$ to N **do**
  $\mathbf{u} := \mathbf{u} + \gamma(q_i - \mathbf{a}_i{}^*\mathbf{u})\mathbf{a}_i/||\mathbf{a}_i||_2^2$
end for
// backward sweep
for $i = N$ to 1 **do**
  $\mathbf{u} := \mathbf{u} + \gamma(q_i - \mathbf{a}_i{}^*\mathbf{u})\mathbf{a}_i/||\mathbf{a}_i||_2^2$
end for
return u

Matrix-free formulation

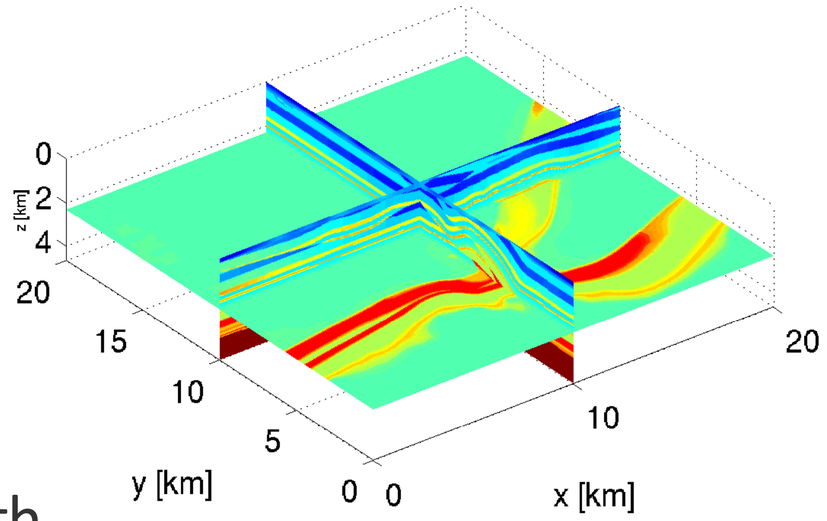
Simple algorithm

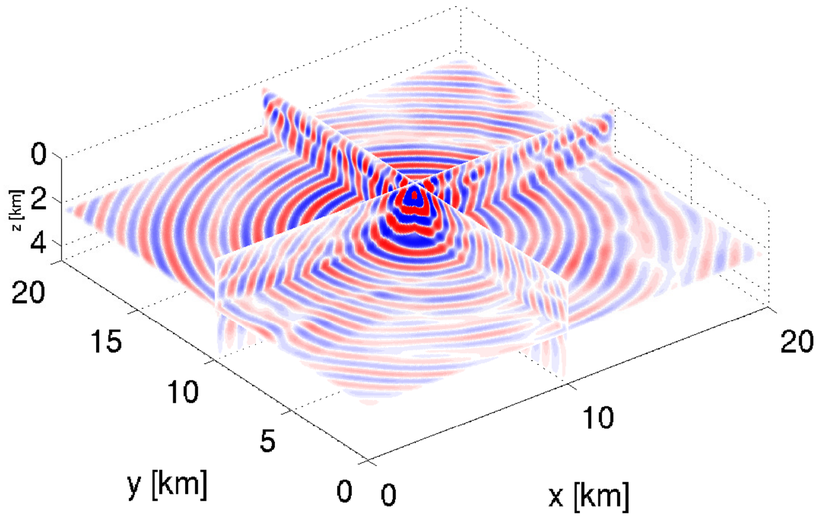
- low setup costs
- matrix-free implementation
- extends to different physics (e.g. Virieux uses this for elastic FWI)
- different discretizations
- on the fly generation of stencils

Extensions

- block
- CRMN
- multilevel

Overtrust model





27 point stencil

10 pts per wavelength

PML

5km X 5km X 2.5Km

[2]These experiments were done on a cluster with 36 IBM x3550 nodes, each with 2 quad-core 2.6 GHz. Intel CPUs and 16 GB memory, connected through a Voltaire Infiniband network. Whenever possible we used a maximum of 2 cores per node to avoid cache conflicts. Timings for more than 64 CPUs may therefore be suboptimal.

CGMN

0.5, 1, 2, 4.5 Hz
non-optimized

comparison Bigstab, GMRES(5), CGMN

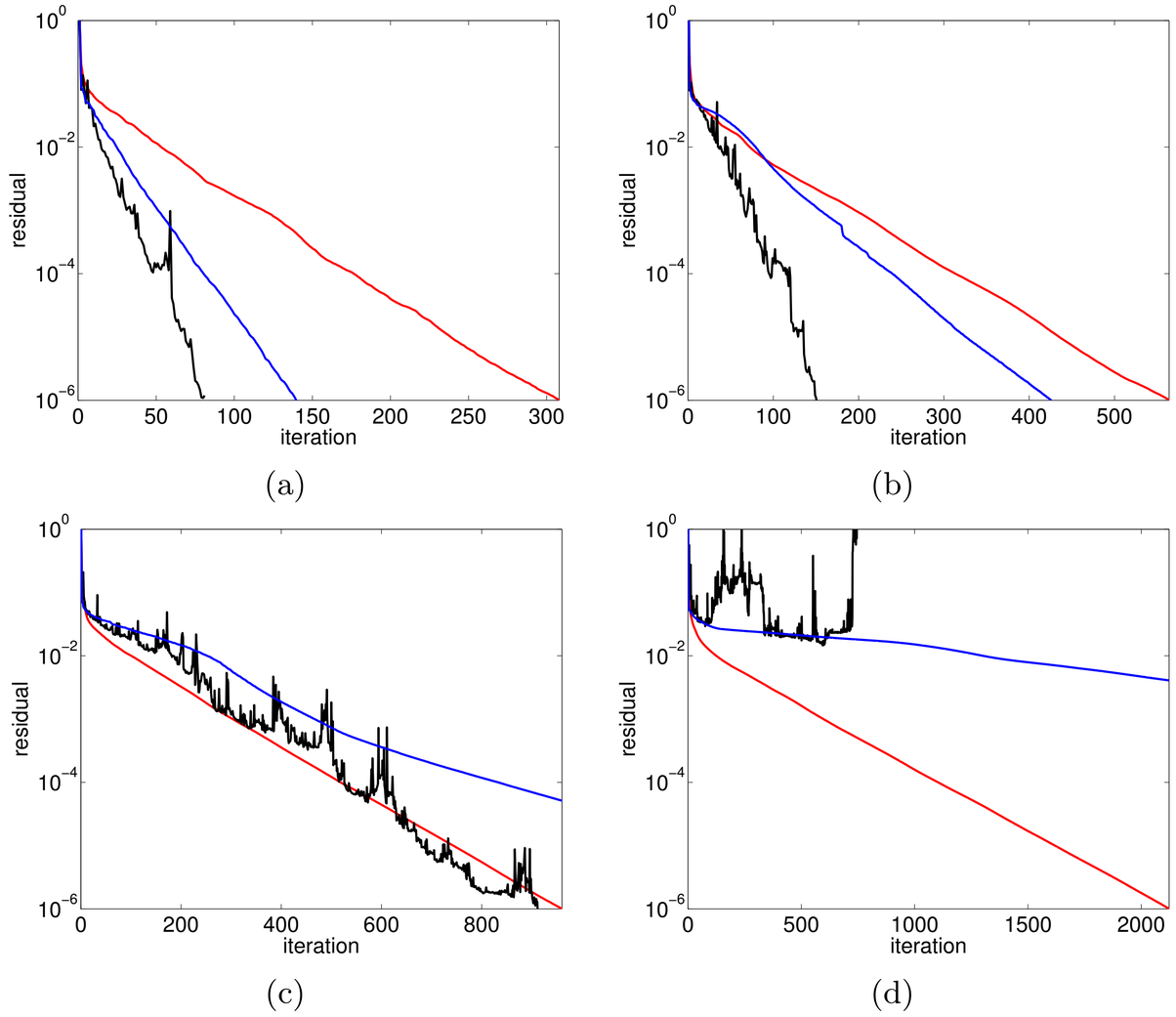


FIG. 4.2. *Convergence histories for BICGstab (black), GMRES(5) (blue) and CGMN (red) for various frequencies: (a) .5 Hz, (b) 1 Hz, (c) 2 Hz and (d) 4 Hz. These plots clearly illustrate the convergence behaviour of the different methods; BiCGstab converges very irregulary, and both GMRES and CGMN decrease the residual monotonically.*

CGMN

0.5, 1, 2, 4 Hz

non-optimized

f [Hz]	N	CGMN	BiCGs	GMRES(5)
0.5	23276.0	308.0	81.0	112.0
1.0	186208.0	564.0	150.0	425.0
2.0	1455808.0	960.0	911.0	963.0
4.0	11646464.0	2123.0	*	*

TABLE 4.1

*Iteration counts for CGMN, BiCGstab and GMRES for different frequencies (using a constant number of gridpoints per wavelength). Here, N denotes the total number of unknowns and * indicates that the method did not converge to the desired tolerance of $\epsilon = 10^{-6}$ within 5000 iterations.*

Block CGMN– w/o deflation

Algorithm 1 BCGMN$(A, U_0, S, \gamma, \epsilon)$ Block-CG algorithm on system $(I-Q)U = RS$, using DKSWP to perform the matrix-vector multiplications

$P_0 = R_0 = \text{DKSWP}(A, U_0, S, \gamma) - U_0$
while $||R_k||_F > \epsilon||B||_F$ **do**
  $Q_k = P_k - \text{DKSWP}(A, P_k, 0, \gamma)$
  $\alpha_k = ({P_k}^* Q_k)^{-1}({R_k}^* R_k)$
  $U_{k+1} = U_k + P_k \alpha_k$
  $R_{k+1} = R_k - Q_k \alpha_k$
  $\beta_k = ({R_k}^* R_k)^{-1}({R_{k+1}}^* R_{k+1})$
  $P_{k+1} = R_k + P_k \beta_k$
  $k = k + 1$
end while

Block CG

0.5,1,2 Hz

sources selected *randomly*

multiple right-hand-sides

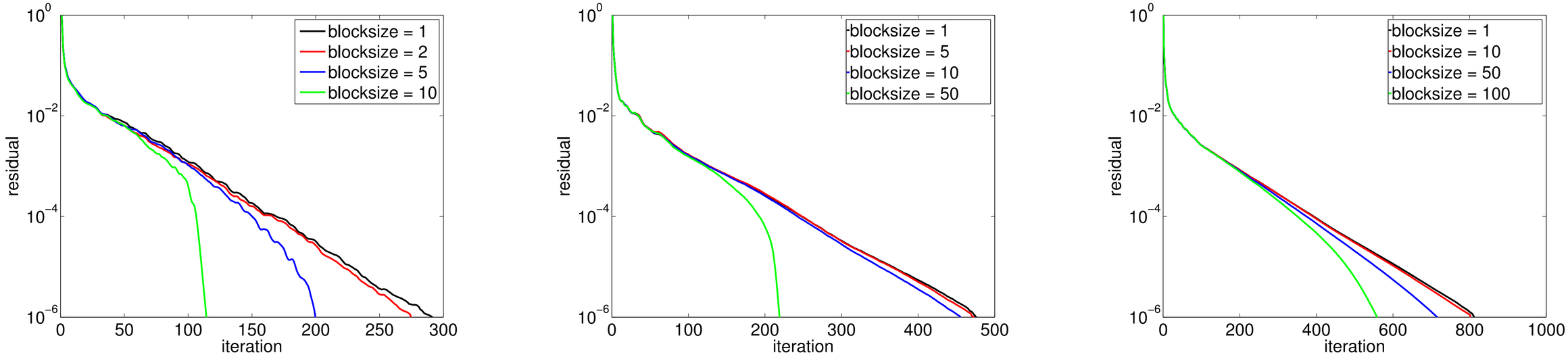


FIG. 4.3. *Convergence histories for block-CGMN with various block-sizes for frequencies (a) .5 Hz, (b) 1 Hz and (c) 2 Hz. Interestingly, the convergence is not sped up uniformly; for the first 100 - 200 iterations, all block-sizes yield the same result. Especially the largest block-sizes show super-linear convergence after a certain number of iterations.*

Block CG

0.5,1, 2 Hz

sources selected randomly

f [Hz]	N	blocksize	iter	time [s]
0.5	23276	1	291	35.9
		2	278	43.3
		5	200	29.7
		10	115	15.2
1.0	186208	1	484	2859.9
		5	477	2419.8
		10	456	2279.7
		50	220	1067.7
2.0	1455808	1	828	125358.2
		10	811	122732.7
		50	716	109424.7
		100	559	82938.2

CARP-CG

parallel over blocks of rows
averaging guarantees convergence

multiple cores

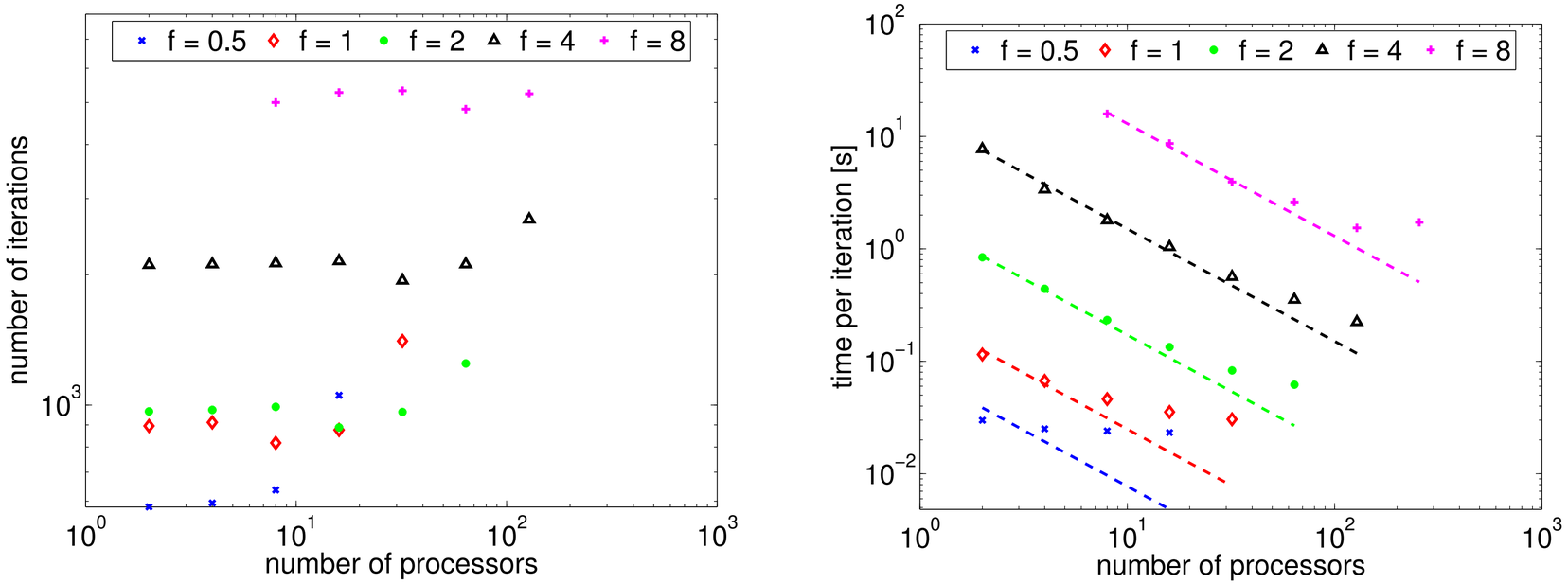


FIG. 4.4. *(a) Number of iterations as a function of the number of processors for CARP-CG. Ideally, the number of iterations should be independent of the number of domains (processors), but the method becomes slightly less effective when the domains become very small. (b) CPU time per iteration as a function of the number of processors for CARP-CG for various frequencies. The dashed line indicates the theoretical CPU time in case of linear speedup.*

Rafael Lago, Art Petrenko, Zhilong Fang, and Felix J. Herrmann, "**Fast solution of time-harmonic wave-equation for full-waveform inversion**", in *EAGE Annual Conference Proceedings*, 2014.

CRMN as a smoother

Algorithm 2: CRMN

Input $\mathbf{x}_0, \mathbf{b} \in \mathbb{C}^n$, $\mathbf{A} \in \mathbb{C}^{n \times n}$
$\mathbf{r}_0 = \text{DKSWP}(\mathbf{A}, \mathbf{b}, \mathbf{x}_0) - \mathbf{x}_0$ and $\mathbf{p}_0 = \mathbf{r}_0$
$\mathbf{A}\mathbf{r}_0 = \mathbf{r}_0$ - $\text{DKSWP}(\mathbf{A}, \mathbf{0}, \mathbf{r}_0)$ and $\mathbf{A}\mathbf{p}_0 = \mathbf{A}\mathbf{r}_0$
for $j = 0..$ until convergence **do**

$$\alpha_j := \frac{\mathbf{r}_j^H \mathbf{A}\mathbf{r}_j}{\mathbf{A}\mathbf{p}_j^H \mathbf{A}\mathbf{p}_j}$$

$$\mathbf{x}_{j+1} := \mathbf{x}_j + \alpha_j \mathbf{p}_j$$

$$\mathbf{r}_{j+1} := \mathbf{r}_j - \alpha_j \mathbf{A}\mathbf{p}_j$$

$$\mathbf{A}\mathbf{r}_{j+1} = \mathbf{r}_{j+1} \text{ - } \text{DKSWP}(\mathbf{A}, \mathbf{0}, \mathbf{r}_{j+1})$$

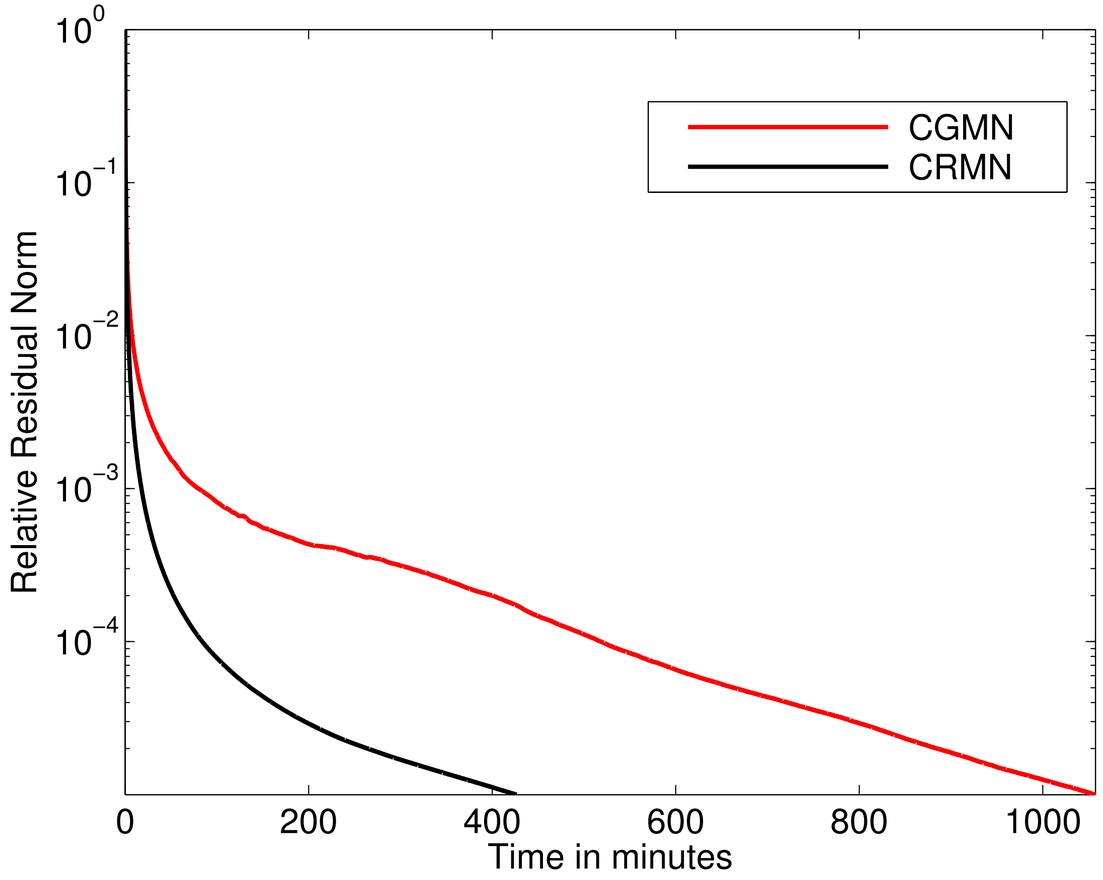
$$\beta_j := \frac{\mathbf{r}_{j+1}^H \mathbf{A}\mathbf{r}_{j+1}}{\mathbf{r}_j^H \mathbf{A}\mathbf{r}_j}$$

$$\mathbf{p}_{j+1} := \mathbf{r}_{j+1} - \beta_j \mathbf{p}_j$$

$$\mathbf{A}\mathbf{p}_{j+1} := \mathbf{A}\mathbf{r}_{j+1} - \beta_j \mathbf{A}\mathbf{p}_j$$

end for

SEG/EAGE overthrust @8Hz



H. Calandra, S. Gratton, R. Lago, X. Pinel, and X. Vasseur. Two-level preconditioned Krylov subspace methods for the solution of three-dimensional heterogeneous Helmholtz problems in seismics. *Numerical Analysis and Applications*, 5:175–181, 2012.

Henri Calandra, Serge Gratton, Xavier Pinel, and Xavier Vasseur. An improved two-grid preconditioner for the solution of three-dimensional Helmholtz problems in heterogeneous media. *Numerical Linear Algebra with Applications*, 2013. to appear.

Three-grid preconditioner

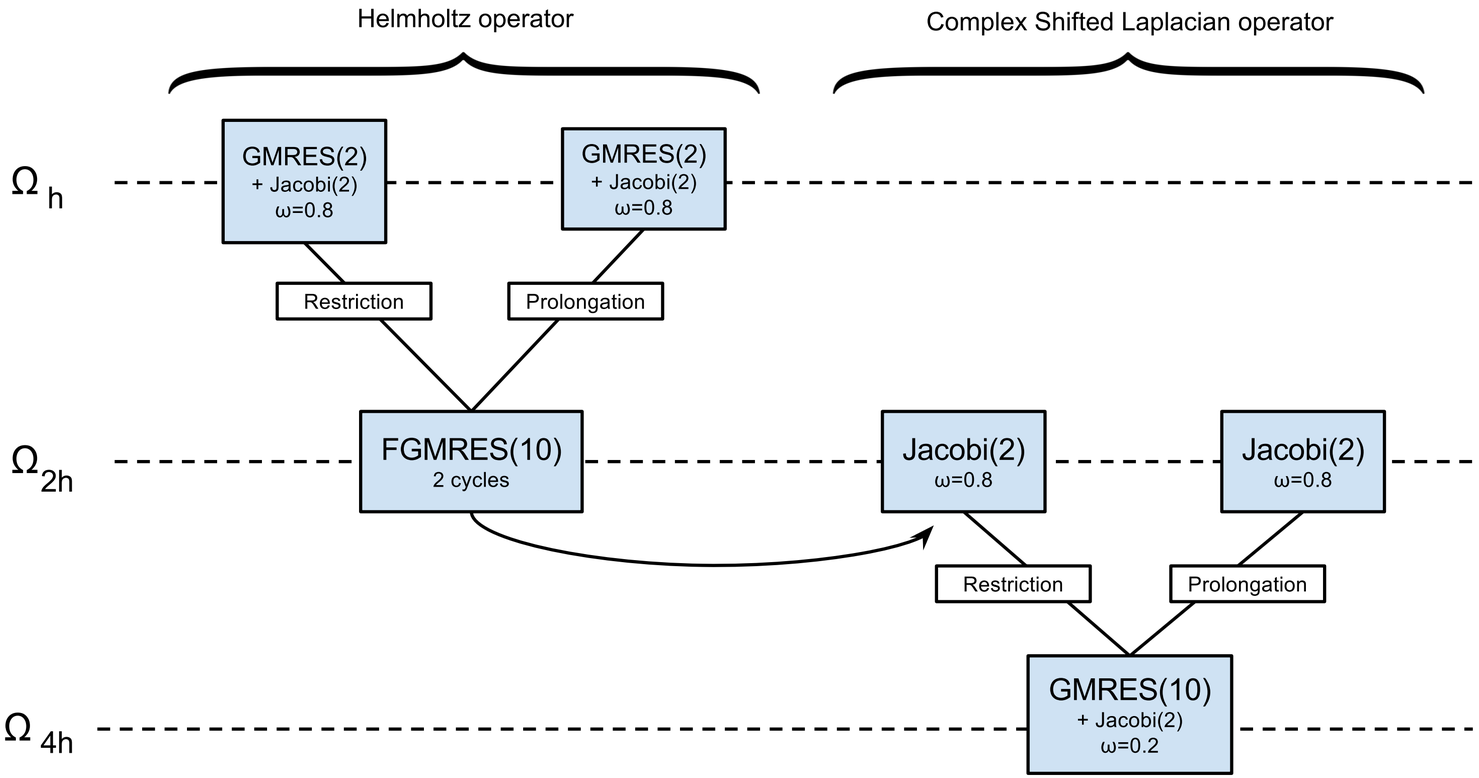


Figure 1: *Representation of the two nested multigrid V-cycle scheme of the $\mathcal{T}_{2V}$ preconditioner.*

Three-grid preconditioner

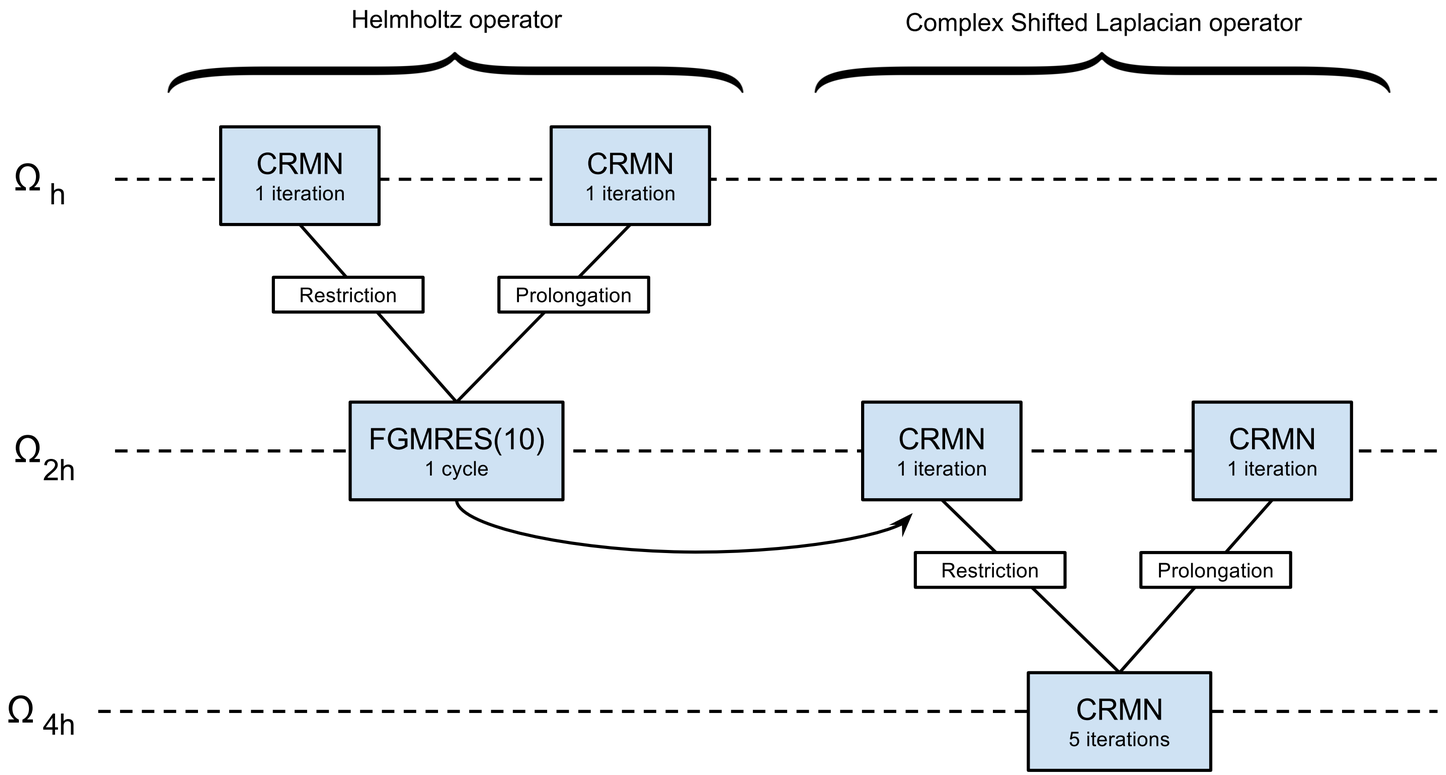


Figure 2: *Representation of the two nested multigrid V-cycle scheme of the ML-CRMN preconditioner.*

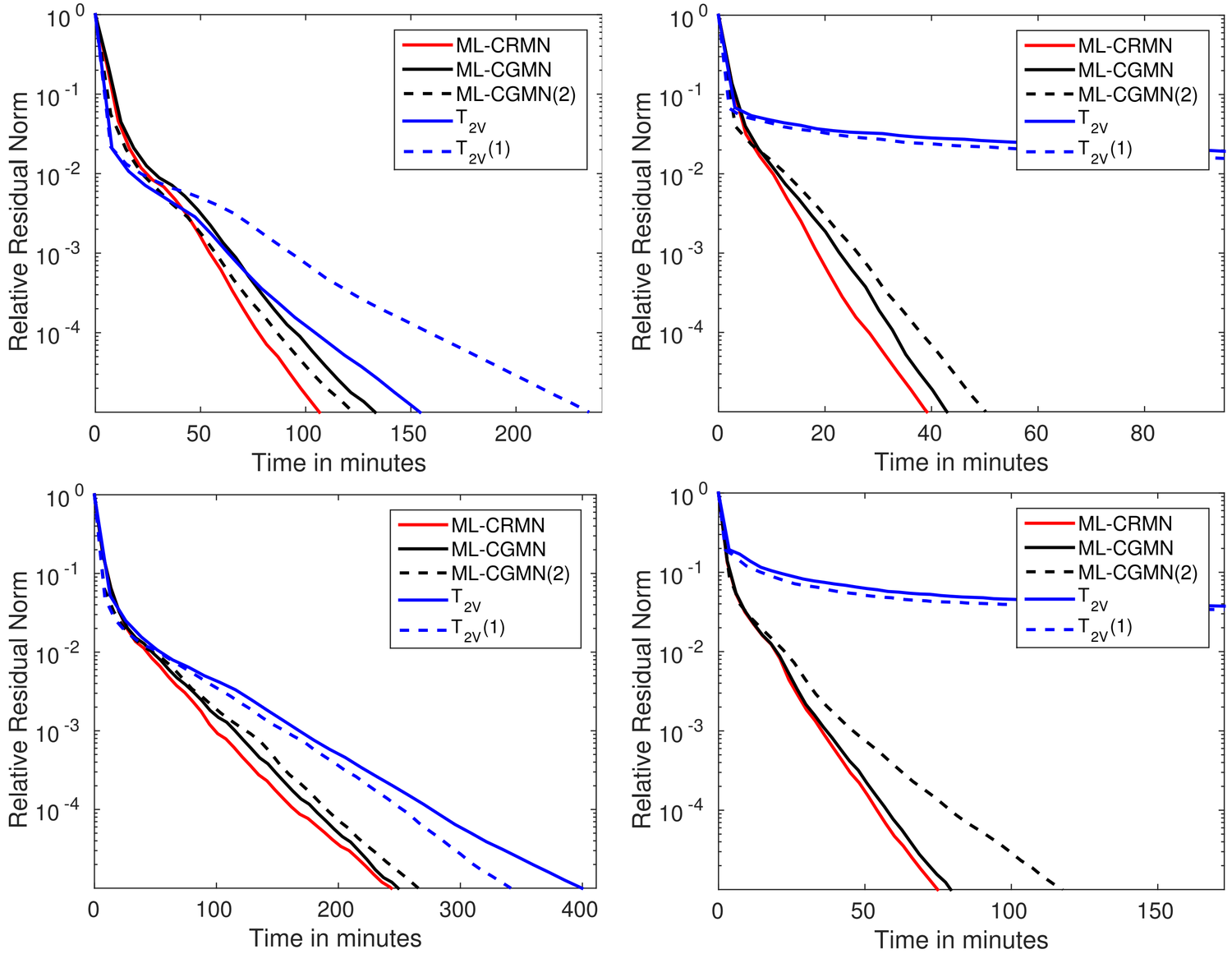


Figure 4: *Relative residual for $\mathcal{T}_{2V}$, ML-CRMN and their variants. Test performed in MATLAB for SEG/EAGE Overthrust (top images) and Salt Dome (bottom images) velocity model, solved at 8Hz discretized. Discrete operators obtained with a 7 (left) and 27 (right) points stencil.*

		SEG/EAGE Overthrust			SEG/EAGE Salt Dome		
		It	T(m)	M(GiB)	It	T(m)	M(GiB)
7pts	$\mathcal{T}_{2V}$	20	158.0	30.09	33	411.7	36.0
	$\mathcal{T}_{2V}(1)$	33	241.0	30.09	41	343.3	36.0
	ML-CRMN	20	108.5	36.50	37	248.4	43.7
	ML-CGMN	22	133.6	34.34	38	251.0	41.2
	ML-CGMN(2)	17	123.9	34.34	35	271.7	41.2
27pts	$\mathcal{T}_{2V}$	-	-	6.57	-	-	7.9
	$\mathcal{T}_{2V}(1)$	-	-	6.57	-	-	7.9
	ML-CRMN	16	41.3	7.91	26	75.2	9.5
	ML-CGMN	18	45.2	7.44	28	80.5	8.9
	ML-CGMN(2)	16	50.2	7.44	33	118.2	8.9

Table 2: *Comparison between the different preconditioners for SEG/EAGE Overthrust (left) and Salt Dome (right) velocity models at 8Hz for both 7 and 27 points stencil. Time is given in minutes and memory is given in GiB. The system matrix for the Overthrust model (all levels) requires 2.39 and 13.98 GiB respectively for the 7 (diagonal only) and 27 points stencil. For Salt Dome, the system matrix requires 2.86 and 16.73 GiB for 7 and 27 points stencils respectively.*

Versatile modelling

Framework:

- preconditioners based on Kaczmarz sweeps are flexible w.r.t.
  - underlying physics
  - discretization
- produce *smooth* errors as a function of # of *iterations*
  - allows for *dynamic* precision *control*
- can handle multiple right-hand-sides & easily parallelizable
  - scales to 3D FWI

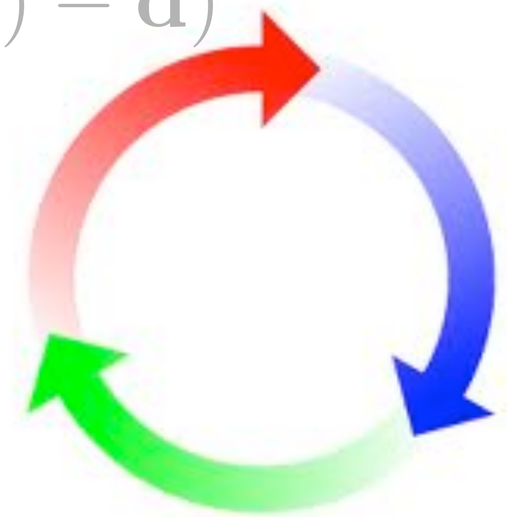
Challenge:

- *translate* into *practice* when *errors* & *convergence* unknown

Frugal FWI

$$\min_{\mathbf{m}} \rho(F(\mathbf{m}) - \mathbf{d})$$

robust
formulation



$$A(\mathbf{m})\mathbf{u} = \mathbf{q}$$

versatile
modelling

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \mathbf{s}_k$$

fast **optimization strategies**

computational framework

Dimensionality reduction

Tristan van Leeuwen and Felix J. Herrmann, "**3D frequency-domain seismic inversion with controlled sloppiness**", *SIAM Journal on Scientific Computing*, vol. 36, p. S192-S217, 2014.

Tristan van Leeuwen and Felix J. Herrmann, "**Fast waveform inversion without source encoding**", *Geophysical Prospecting*, vol. 61, p. 10-19, 2013.

Aleksandr Y. Aravkin, Michael P. Friedlander, Felix J. Herrmann, and Tristan van Leeuwen, "**Robust inversion, dimensionality reduction, and randomized sampling**", *Mathematical Programming*, vol. 134, p. 101-125, 2012.

Eldad Haber, Matthias Chung, and Felix J. Herrmann, "**An effective method for parameter estimation with PDE constraints with multiple right hand sides**", *SIAM Journal on Optimization*, vol. 22, 2012.

Tristan van Leeuwen, Aleksandr Y. Aravkin, and Felix J. Herrmann, "**Seismic waveform inversion by stochastic optimization**", *International Journal of Geophysics*, vol. 2011, 2011.

Leverage structure

Objective function:

$$\phi(\mathbf{m}) = \sum_{i=1}^{M} \phi_i(\mathbf{m}),$$

misfit functional

residual

$$\phi_i(\mathbf{m}) = \rho\left(\overbrace{\mathcal{F}_i(\mathbf{m})\mathbf{q}_i - \mathbf{d}_i}^{\mathbf{r}_i}\right)$$

Reduced costs via

- **batching –** selecting subsets of sources $\mathcal{F}_i(\mathbf{m}) = P_i A^{-1}$ and $\rho(\cdot)$ arbitarry
- **source encoding –** forming source aggregates $\mathcal{F}_i(\mathbf{m}) = P A^{-1}$ and $\rho(\cdot) = \|\cdot\|_2^2$

Tristan van Leeuwen, Aleksandr Y. Aravkin, and Felix J. Herrmann, "**Seismic waveform inversion by stochastic optimization**", *International Journal of Geophysics*, vol. 2011, 2011.

Batching

Approximate the sum–i.e., compute sample average

$$\phi \approx \widetilde{\phi} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \phi_i \quad \text{with} \quad \mathcal{I} \subseteq \{1, 2, \ldots, M\} \quad \text{and} \quad K = |\mathcal{I}| \ll M$$

- leads to errors that decay as batch/sample size *K* increases
- when sampled uniformly

$$E\left(\widetilde{\phi}(\mathbf{m})\right) = \phi(\mathbf{m}) \quad \text{and} \quad E\left(\nabla\widetilde{\phi}(\mathbf{m})\right) = \nabla\phi(\mathbf{m})$$

- can lead to dimensionality reduction

Batching

Difference between sampling **with** or **without** replacement determines

- how the variance of the sample average decreases as K increases
- both cases depend on the sample variance

$$\sigma_g := \frac{1}{M-1}\sum_{i=1}^{M} \|\nabla\phi_i - \nabla\phi\|_2^2$$

but lead to different decay for variance of the ***sample*** averaged gradients

- expectation error **without** replacement

$$E(\||\mathbf{e}_n\|_2^2) = \frac{1}{K}\left(1-\frac{K}{M}\right)\sigma_g$$

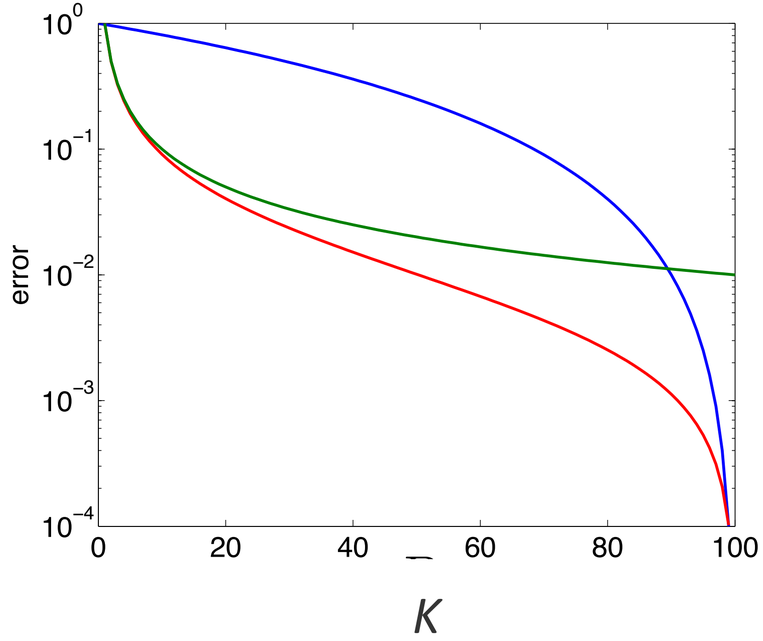
- or **with** replacement

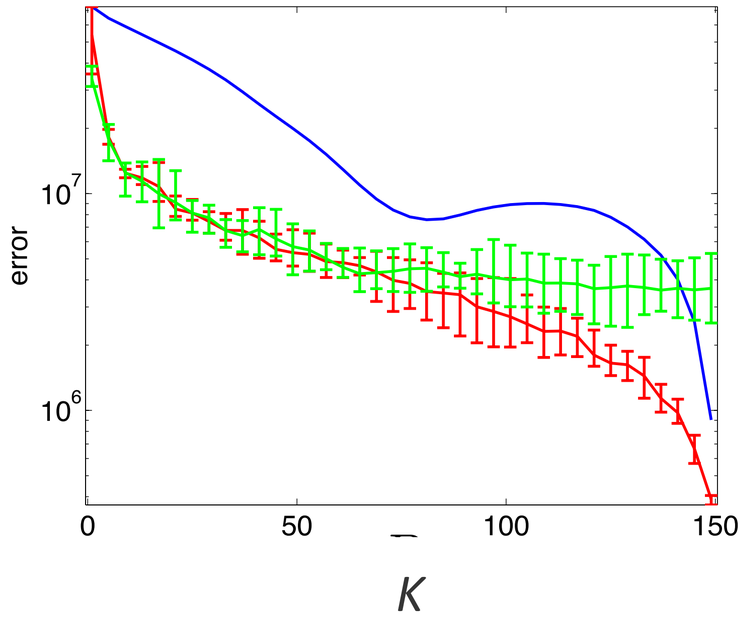
$$E(\||\mathbf{e}_r\|_2^2) = \frac{1}{K}\sigma_g$$

Batching
– *increasing* the sample size

Select sources
- in a pre-scribed order / worst case
- random *without* replacement
- random *with* replacement / *amplitude* source encoding

$$E(\|\|\mathbf{e}_n\|_2^2) = \left(\frac{M-K}{M(M-1)}\right)\frac{M}{K}\max_i(\nabla\phi_i - \phi)$$





Source encoding

When all sources see the same receivers and LS misfit

$$\mathcal{F}_i(\mathbf{m}) = PA^{-1} \text{ and } \rho(\cdot) = \|\cdot\|_2^2$$

we can form source aggregates or supershots w/ simultaneous sources

$$\widetilde{\mathbf{q}} = \sum_{i=1}^{M} w_i \mathbf{q}_i = Q\mathbf{w} \quad \text{with} \quad Q = \left[\mathbf{q}_1 \cdots \mathbf{q}_M\right]$$

yielding

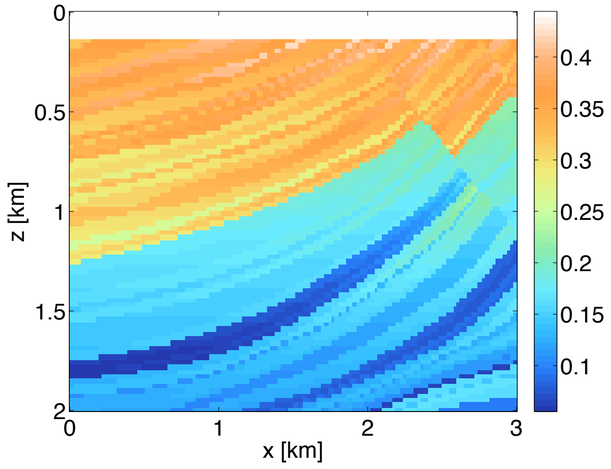
$$\phi = E\left(\widetilde{\phi}(\mathbf{m})\right), \quad \widetilde{\phi}(\mathbf{m}) = PA^{-1}(\mathbf{m})\widetilde{\mathbf{q}} - \widetilde{\mathbf{d}}$$

when covariance

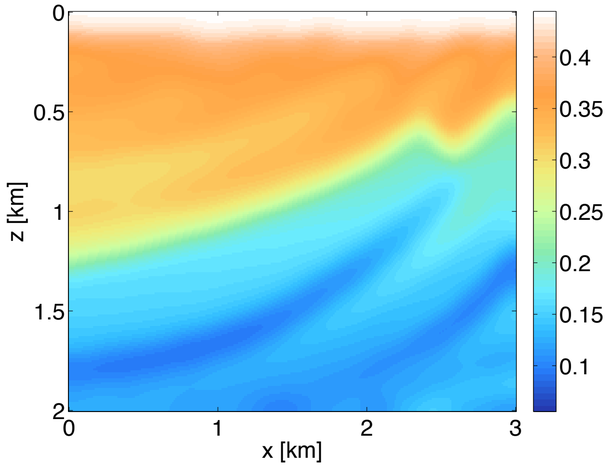
$$E(\mathbf{w}\mathbf{w}^T) = \frac{1}{M} I$$

Stylized example

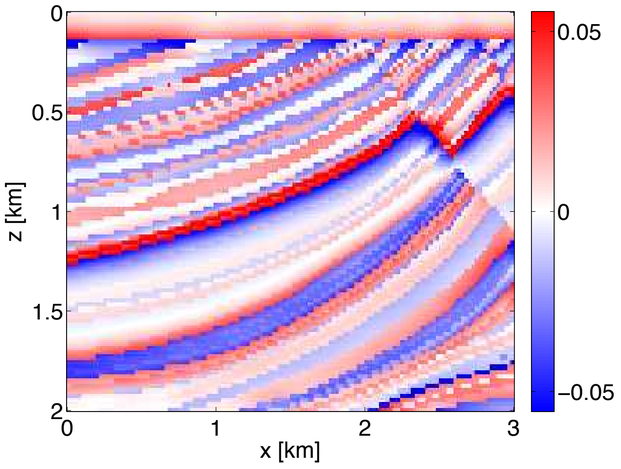
true model



starting model



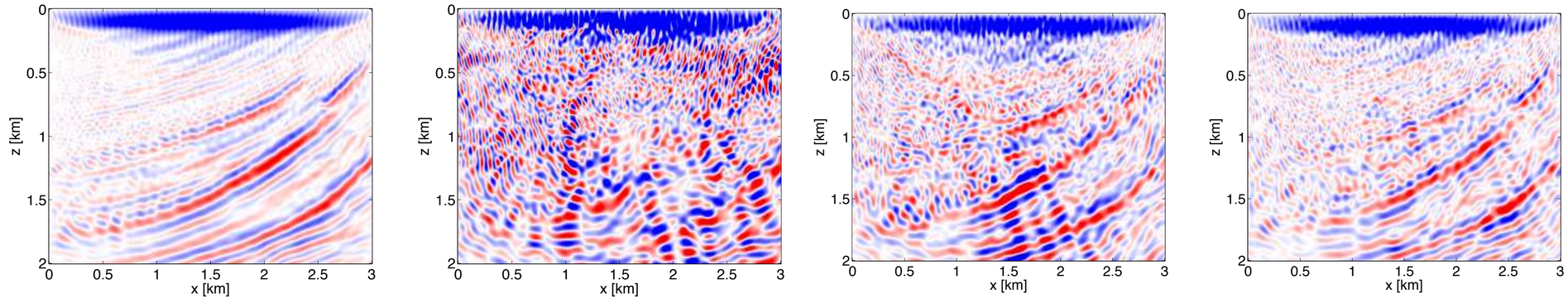
'reflectivity'



Gradients

Search direction for *increasing* batch size K:

$$\mathbf{g}_K \approx \frac{1}{K}\sum_{j=1}^{K} \nabla \mathcal{F}^*[\mathbf{m};\widetilde{\mathbf{q}}_j]\widetilde{\delta\mathbf{d}}_j$$

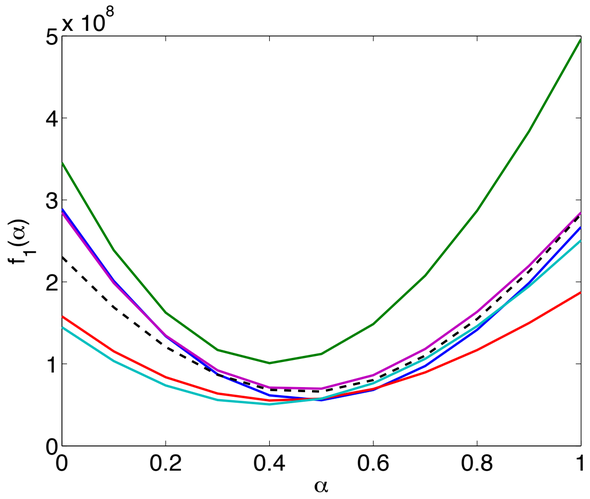


full K=1 K=5 K=10

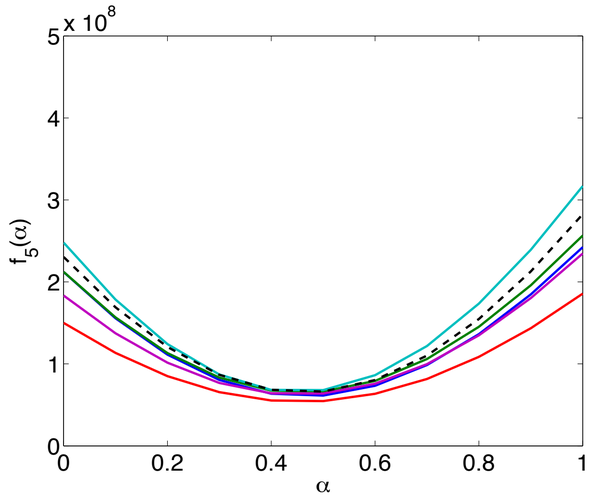
Objective

$$\phi_K(\mathbf{g}_K) = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{2}\|\widetilde{\mathbf{d}}_j - \mathcal{F}[\mathbf{m}+\alpha\mathbf{g}_K;\widetilde{\mathbf{q}}_j]\|_2^2$$

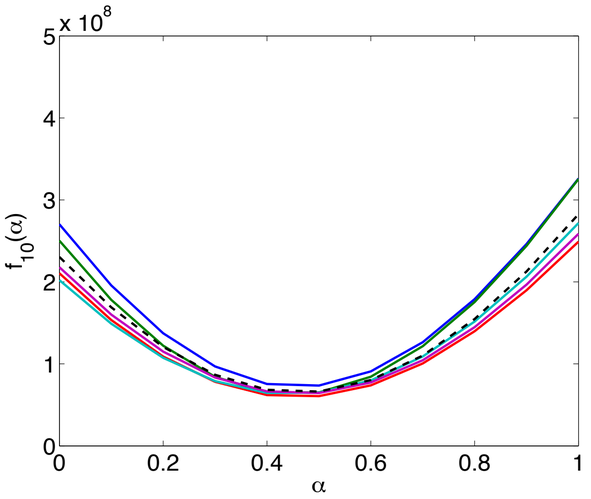
$$|\phi - \phi_K| \sim \mathcal{O}(\frac{1}{\sqrt{K}})$$



K=1



K=5

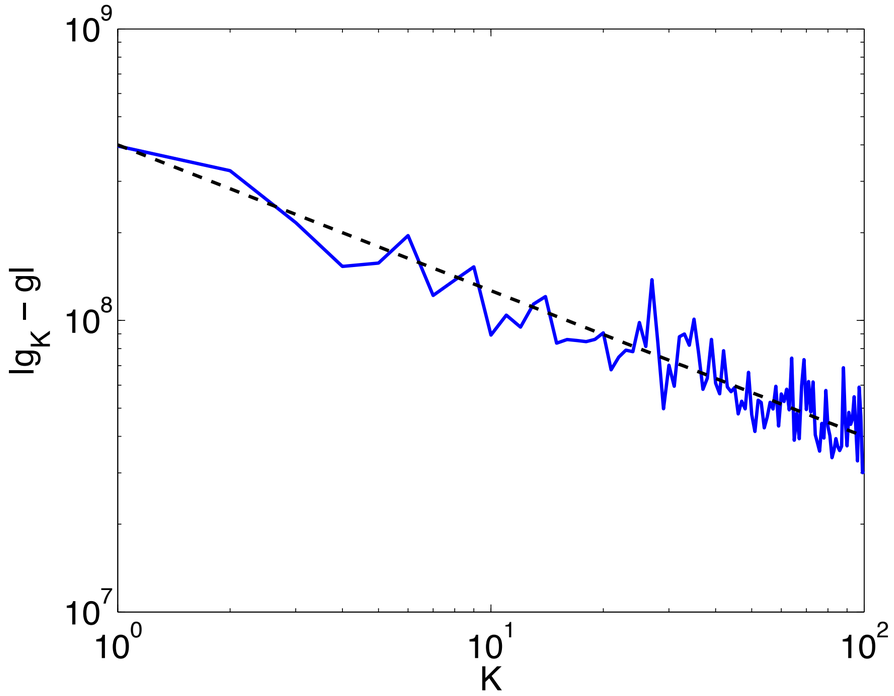


K=10

[Haber, Chung, and FJH, '10; van Leeuwen, Aravkin, FJH, '10]

Decay

$$|\phi - \phi_K| \sim \mathcal{O}(\frac{1}{\sqrt{K}})$$



error between full and sampled gradient

Avron, Haim, and Sivan Toledo. "Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix." *Journal of the ACM (JACM)* 58.2 (2011):
F. Roosta-Khorasani and U. Ascher, J. Found. of Comp. Math. (2014), DOI: 10.1007/s10208-014-9220-1 arXiv1308.2475: Improved bounds on sample size for implicit matrix trace estimators

Random-trace estimation

$$\begin{aligned}\text{trace}\left(R^T R\right) = E\left(\mathbf{w}^T R^T R\mathbf{w}\right) &\approx \frac{1}{K}\sum_{i=1}^{K}\|R\mathbf{w}_i\|_2^2 \\ &\approx \|PA^{-1}QW - DW\|_F^2\end{aligned}$$

- valid for arbitrary normalized *W*
- accuracy estimates exist

$$\Pr\left(\frac{|T_K - T|}{|T|} \leq \epsilon\right) \geq 1 - \delta.$$

Random-trace estimation

TABLE 1: Summary of bounds, adapted from Avron and Toledo [14].

Estimator	Distribution of w	Variance of one sample	Bound on K for (ϵ, δ) bound
Hutchinson $H_K = (1/K)\sum_{j=1}^{K} w_j^\top A w_j$	$\Pr(w_j = \pm 1) = 1/2$	$2(\|A\|_F^2 - \sum_{i=1}^{N} A_{ii}^2)$	$6\epsilon^{-2}\ln(2\,\mathrm{rank}(A)/\delta)$
Gaussian $G_K = (1/K)\sum_{j=1}^{K} w_j^\top A w_j$	$w_j \in N(0,1)$	$2\|A\|_F^2$	$20\epsilon^{-2}\ln(2/\delta)$
Phase encoded $L_K = (N/K)\sum_{j=1}^{K} w_j^T \mathcal{F} A \mathcal{F}^T w_j$	w_j drawn uniformly from $\{e_1, \ldots, e_N\}$	n/a	$2\epsilon^{-2}\ln(4n^2/\delta)\ln(4/\delta)$

Samplings

– Gaussian vs unit vectors

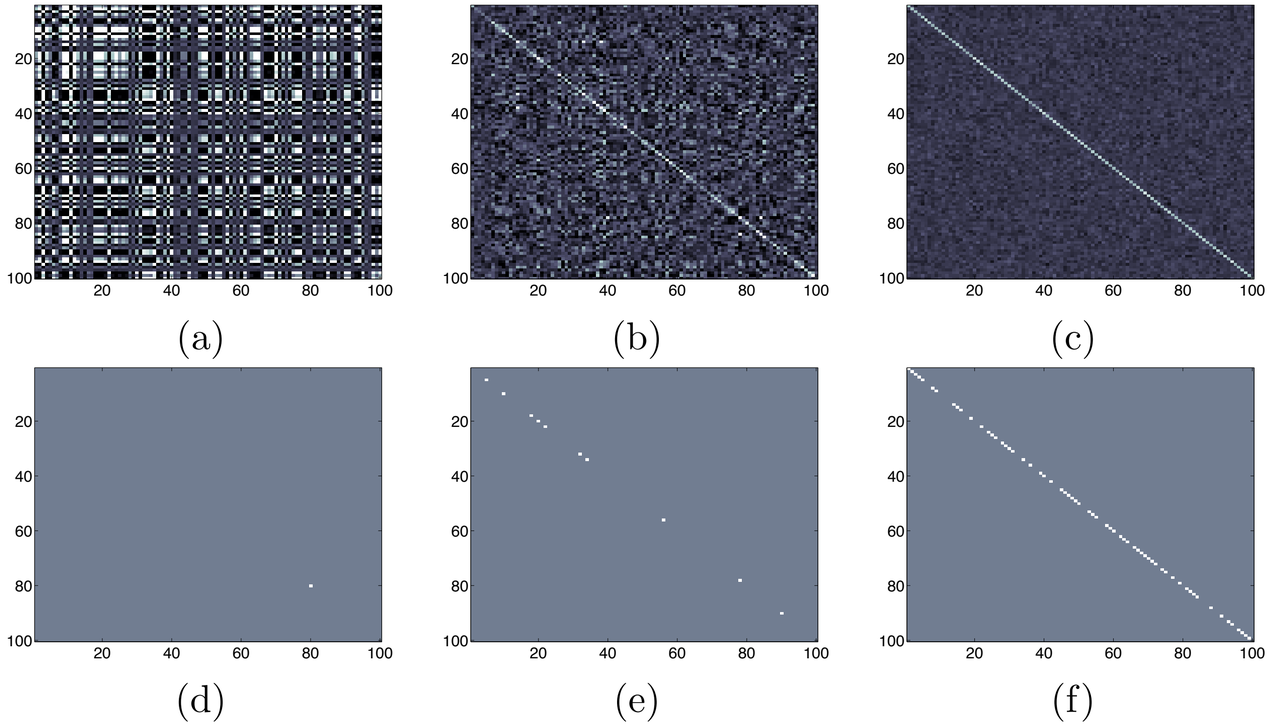


Figure 2: Covariance matrix: $K^{-1}\sum_{i=1}^{K}\mathbf{w}_i{\mathbf{w}_i}^T$ for random Gaussian vectors (top) and random unit vectors (bottom) for $K = \{1, 10, 100\}$ (left to right)

Samplings

– Gaussian vs unit vectors

Gaussian vectors:

- reduce errors in random-trace estimates
- lead to noisy crosstalk
- corresponds to sampling with replacement

Unit vectors:

- establishes link w/ batching
- can be done with and without replacement
- applies to marine sampling & different misfit functionals

In both cases errors are created that may effect optimization…

Stochastic optimization

Tristan van Leeuwen, Aleksandr Y. Aravkin, and Felix J. Herrmann, "**Seismic waveform inversion by stochastic optimization**", *International Journal of Geophysics*, vol. 2011, 2011.
Eldad Haber, Matthias Chung, and Felix J. Herrmann, "**An effective method for parameter estimation with PDE constraints with multiple right hand sides**", *SIAM Journal on Optimization*, vol. 22, 2012.
Aleksandr Y. Aravkin, Michael P. Friedlander, Felix J. Herrmann, and Tristan van Leeuwen, "**Robust inversion, dimensionality reduction, and randomized sampling**", *Mathematical Programming*, vol. 134, p. 101-125, 2012.

SAA

Stochastic-average approximation (SAA):

Solves

$$\min_{\mathbf{x}}\{f(\mathbf{x}) = E_w\left(F(\mathbf{x},\mathbf{w})\right)\}$$

via

$$\min_{\mathbf{x}}\{\widetilde{f}(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} F(\mathbf{x},\mathbf{w}_i)\}$$

- keeps “batch” fixed
- converges w/ probability 1
- select *K* large enough to control error but suffers from bias

SA

Stochastic approximation (SA):

- remove bias by drawing independent source weights at every iteration
- can be done for ***w***'s w/ >1 encodings
- convergence proofs are technical & mostly limited to first-order methods
  - gradient descent
  - no Newton or quasi-Newton
- needs specialized step lengths or averaging that lead to loss of fast convergence
- under certain conditions show to have sublinear error decay after *k* iterations in expectation of

$$|\phi - \phi_k| \sim \mathcal{O}(\frac{1}{\sqrt{k}})$$

Stochastic approximation (SA)

Algorithm 1: Stochastic gradient descent

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; // initial model

while not converged **do**

$\{\widetilde{\mathbf{d}}^k, \widetilde{\mathbf{q}}^k\} \longleftarrow \{\mathbf{D}\mathbf{w}^k, \mathbf{Q}\mathbf{w}^k\}$ with $\mathbf{w}^k \in N(0,1)$; // draw sim. exp.

$\mathbf{g}^k \longleftarrow \nabla \boldsymbol{\mathcal{F}}^*[\mathbf{m}^{k-1}, \widetilde{\mathbf{q}}^k](\widetilde{\mathbf{d}}^k - \boldsymbol{\mathcal{F}}[\mathbf{m}^{k-1}, \widetilde{\mathbf{q}}^k])$; // gradient

$\underline{\mathbf{m}}^{k+1} \longleftarrow \mathbf{m}^k - \gamma^k \mathbf{g}^k$; // update

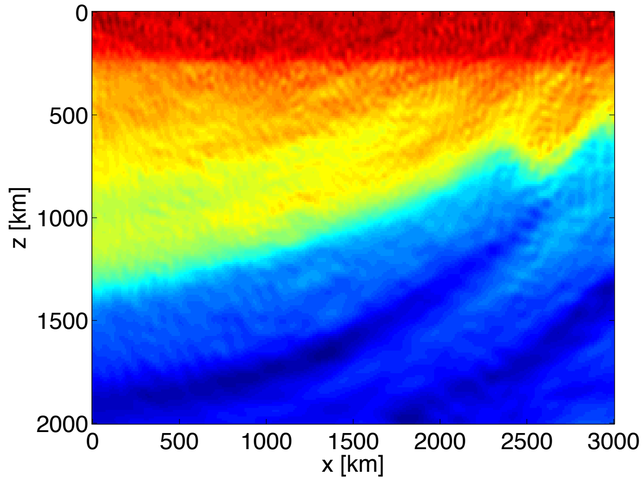
$\mathbf{m}^{k+1} = \frac{1}{k+1}\left(\sum_{i=1}^{k} \mathbf{m}^i + \underline{\mathbf{m}}^{k+1}\right)$; // average

$k \longleftarrow k+1$;

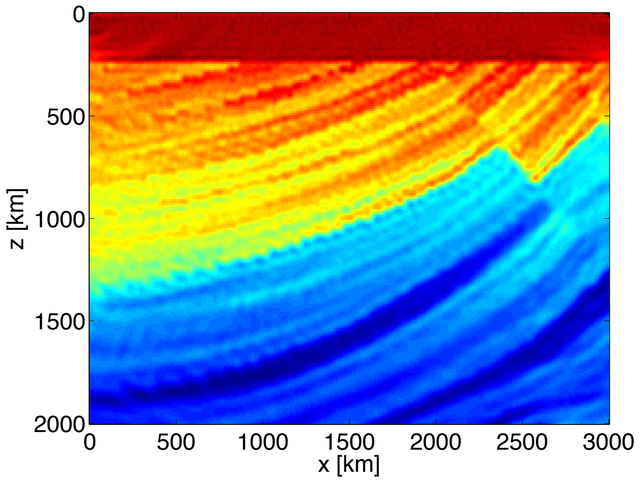
end

[Bertsekas, '96; Haber, Chung, and FJH, '10]

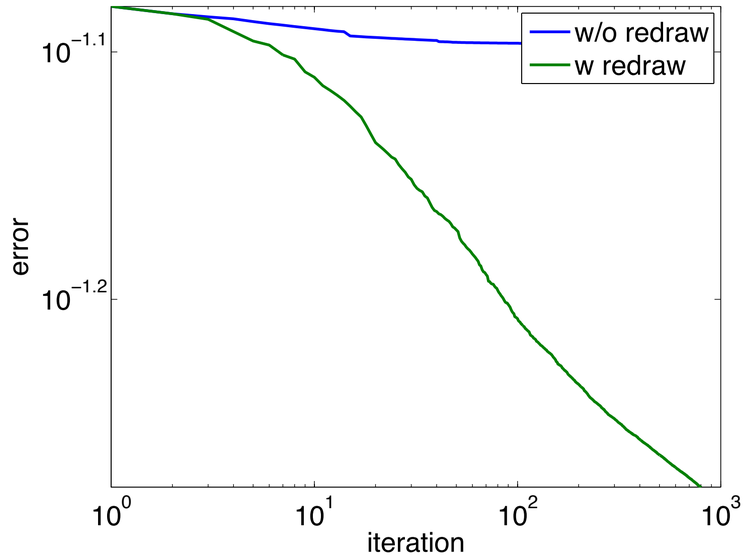
K=1 w/ and w/o redraw [noise-free case]



w/o redraw



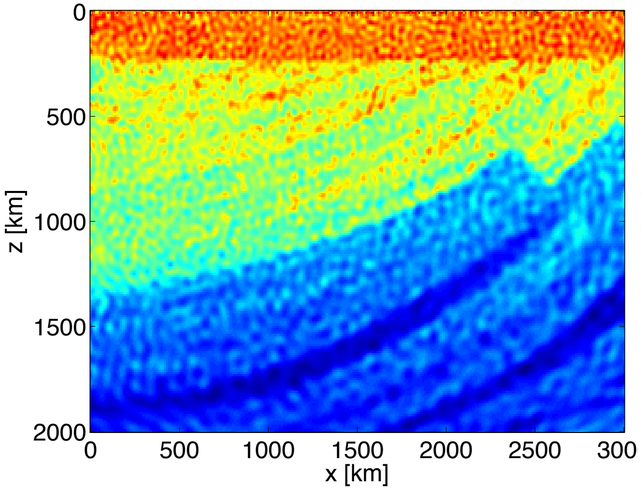
w redraw



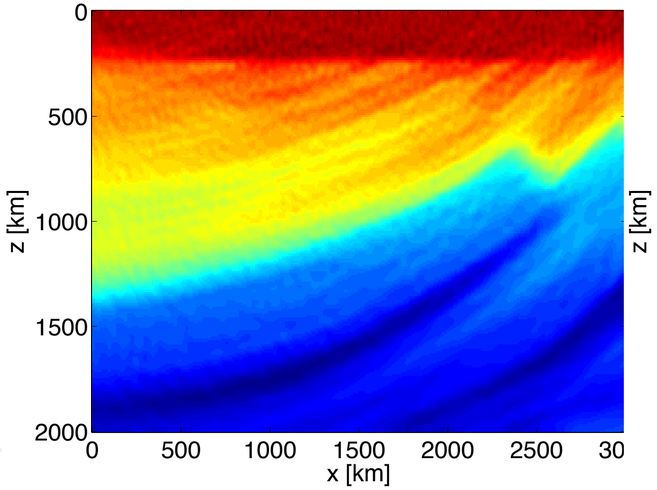
model error K=1

K=1
[noisy case]

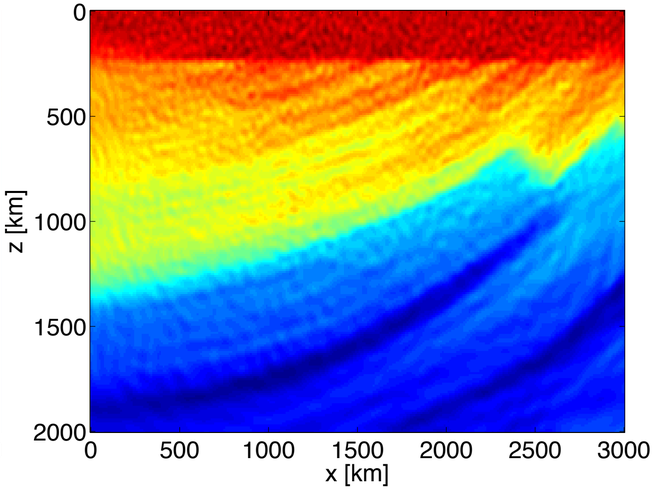
w/o averaging

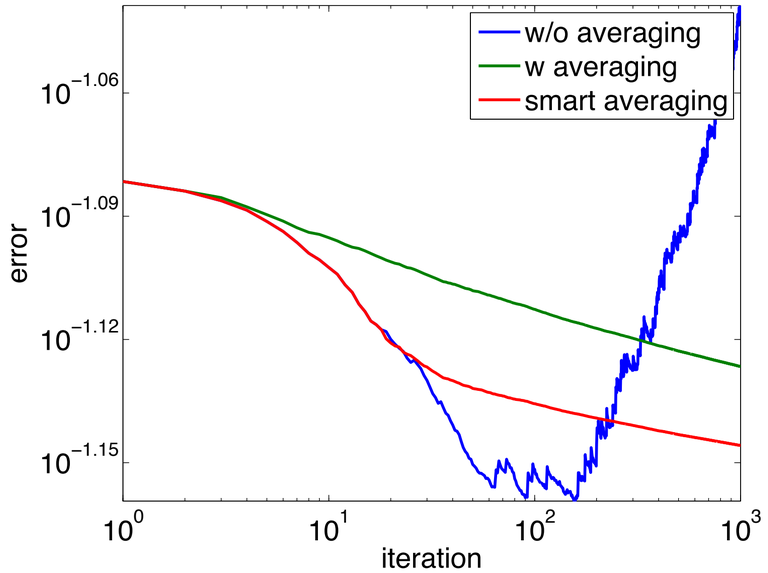


w averaging



smart averaging





Observations

Stochastic-average approximation (SAA)'s error

- decays slowly as a function of K
- is fixed throughout the iterations

Stochastic approximation (SA)

- converges slowly (sub linearly) as a function of k
- may become unstable when data is noisy

Combine deterministic full-gradient with stochastic techniques to get best of both worlds

- fast convergence in the beginning
- control over error by increasing the batch/sample size K guarantee convergence

Hybrid optimization

Michael P. Friedlander and Mark Schmidt, "**Hybrid deterministic-stochastic methods for data fitting**", *SIAM Journal on Scientific Computing*, vol. 34, p. A1380-A1405, 2012.
Tristan van Leeuwen and Felix J. Herrmann, "**Fast waveform inversion without source encoding**", *Geophysical Prospecting*, vol. 61, p. 10-19, 2013.
Aleksandr Y. Aravkin, Michael P. Friedlander, Felix J. Herrmann, and Tristan van Leeuwen, "**Robust inversion, dimensionality reduction, and randomized sampling**", *Mathematical Programming*, vol. 134, p. 101-125, 2012.

Errors & convergence

Consider steepest descent–i.e,

search direction

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \gamma_k \mathbf{s}_k \text{ where } \mathbf{s}_k = -\nabla\phi(\mathbf{m}_k)$$

has linear convergence rate

$$\|\phi(\mathbf{m}_k) - \phi(\mathbf{m}_*)\|_2^2 = \mathcal{O}(c^k), \quad 0 \leq c < 1$$

When error in search directions $\mathbf{s}_k = -\nabla\phi(\mathbf{m}_k) + \mathbf{e}_k$ we have

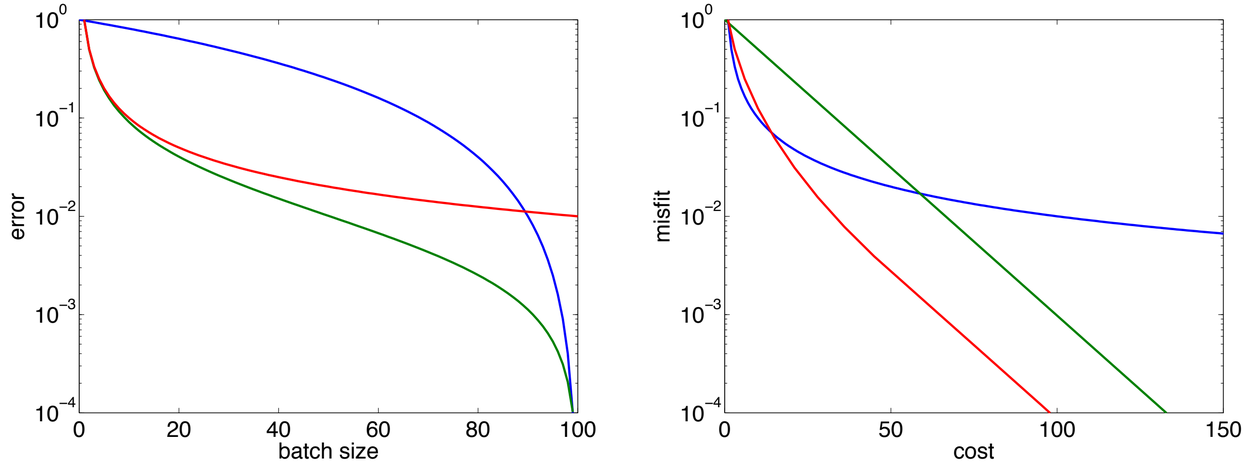
$$\|\phi(\mathbf{m}_k) - \phi(\mathbf{m}_*)\|_2^2 = \mathcal{O}(\max\{c^k, \|\mathbf{e}_k\|_2^2\}), \quad 0 \leq c < 1$$

Have fast (= linear) convergence as long we control the error…

Hybrid optimization

Linear converge rate when we choose a batching strategy so that

$$E(\|\mathbf{e}_k\|_2^2) = \mathcal{O}(c^k), \quad 0 \leq c < 1$$



(a) (b)

Figure 1: (a) Asymptotic behaviour of the error between the approximate and true gradient: worst-case batching (blue), average batching (green), source encoding (red). (b) Asymptotic convergence rate for different optimization strategies: conventional (green), stochastic (blue) and hybrid (red).

Observations

Scheme that interpolates between stochastic gradient & full gradient.

If we want to design optimization scheme w/ linear convergence rate

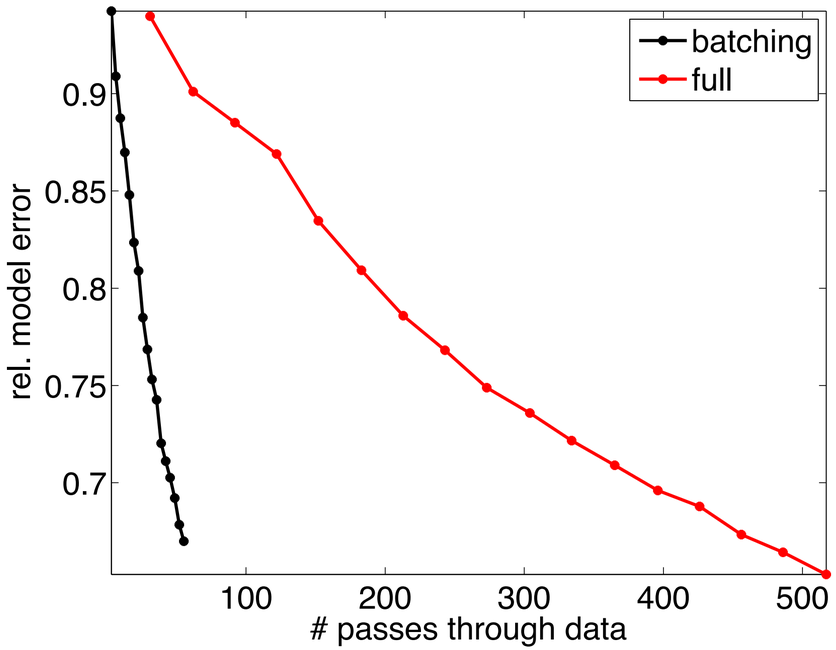
- batch size needs to grow such that $E(\|\mathbf{e}_k\|_2^2)$ decreases with at least that rate
- no benefit to increase batch size at a rate faster then the rate of the original problem–i.e., when problem is ill conditioned so *L* (Lipschitz constant) is small

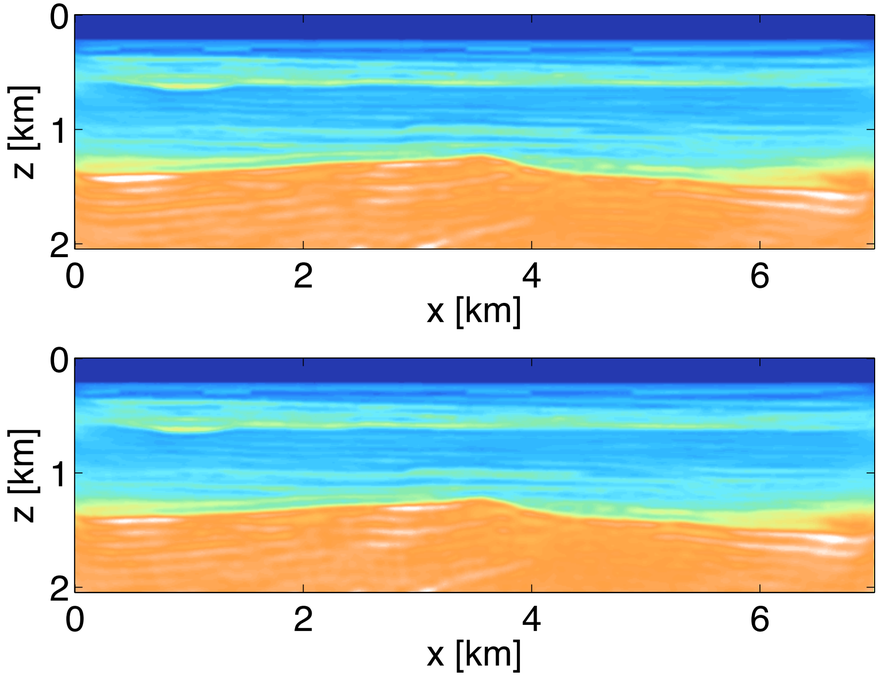
$$\|\nabla\phi(x) - \nabla\phi(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x \text{ and } y$$

Since this rate is unknown in practice **"How to choose the rate of increase of the batch-size in practice is an open problem..."**

Fast optimization

10 x speedup





[van Leeuwen & FJH '11]

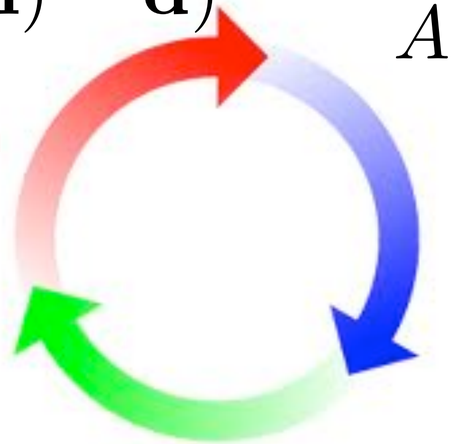
FWI
w/ controlled *sloppiness*

$$\min_{\mathbf{m}} \rho(F(\mathbf{m}) - \mathbf{d})$$

robust
formulation

$$A(\mathbf{m})\mathbf{u} = \mathbf{q}$$

versatile
modelling



$$\mathbf{m}_{k+1} = \mathbf{m}_k + \alpha_k \mathbf{s}_k$$

fast optimization strategies

computational framework

K. van den Doel and U. Ascher, SIAM J. Scient. Comput. (2012), DOI: 10.1137/110826692:
Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements.

Frugal FWI

Van der Doel proposes to use cross validation

- introduces too much overhead
- exponential increase un batch size

Come up w/ a more practical heuristic approach…

Frugal misfit
w/ approximate PDE solves

Heuristic based on *behavior* of the *misfit* as a function of ϵ

$$\phi_i(\mathbf{m}, \epsilon) = \rho(P_i \mathbf{u}_i(\epsilon) - \mathbf{d}_i)$$

by solving PDEs to *tolerance* ϵ.

Ideally find ϵ by guaranteeing

$$|\phi_i(\mathbf{m}, \epsilon) - \phi_i(\mathbf{m}, 0)| \leq \eta \phi_i(\mathbf{m}, 0)$$

true solution

for some fraction η.

[van Leeuwen & FJH '13]

Frugal misfit
w/ approximate PDE solves

Instead find *k* such that

$$|\phi_i(\mathbf{m}, \alpha^k \epsilon) - \phi_i(\mathbf{m}, \alpha^{k+1} \epsilon)| \leq \eta \phi_i(\mathbf{m}, \alpha^{k+1} \epsilon) \quad 0 < \alpha < 1$$

by increasing the precision, i.e., $\epsilon \mapsto \alpha\epsilon$, if this *inequality* does ***not*** hold.

Frugal misfit

Algorithm 1 $\{f, \mathbf{g}\} = \mathsf{misfit}(\mathbf{m}, \mathcal{I}, \eta)$

$\epsilon = 10^{-2}$, $\alpha = 0.5$// Initialization
for $i \in \mathcal{I}$ **do**
  for $k = 0 \rightarrow 10$ **do**
    solve $A(\mathbf{m})\mathbf{u} = \mathbf{q}_i$ up to ϵ // solve forward equation
    $r_k = \rho(P_i\mathbf{u} - \mathbf{d}_i)$ // compute residual
    if $|r_k - r_{k-1}| \leq \eta r_k$ **then**
      break
    else
      $\epsilon = \alpha\epsilon$
    end if
  end for
  solve $A(\mathbf{m})^*\mathbf{v} = P_i{}^*\nabla\rho(P_i\mathbf{u} - \mathbf{d}_i)$ up to ϵ
  $f = f + |\mathcal{I}|^{-1}\rho(P_i\mathbf{u} - \mathbf{d}_i)$ // misfit
  $\mathbf{g} = \mathbf{g} + |\mathcal{I}|^{-1}G(\mathbf{m}, \mathbf{u})^*\mathbf{v}$ // gradient
end for

Stochastic Quasi-Newton

Final algorithm has the following key ingredients:

- draws *independent* random *subsets* for *each* misfit & gradient calculation
- decreases *fraction* $\eta \mapsto \eta/2$ when Wolfe *linesearch* fails
- increases *sample* size when *average* objective does *not* decrease–i.e, if

$$(f_{k+1} + f'_{k+1}) \geq (f_k + f'_k)$$

- Quasi-Newton Hessian w/ l-BFGS requires a single *extra* gradient calculation ensuring the same sample

Stochastic Quasi-Newton

Algorithm 1 Stochastic L-BFGS method

$\eta = 0.1$, $b = 1$, $\beta = 1$, $b_{\max} = M$ // Initialize
choose $\mathcal{I}_0 \subseteq \{1, 2, \ldots, M\}$ s.t. $|\mathcal{I}_0| = b$
$\{f_0, \mathbf{g}_0\} = \mathsf{misfit}(\mathbf{m}_0, \mathcal{I}_0, \eta)$ // *frugal* misfit & gradient at initial guess
while not converged **do**
  $\delta\mathbf{m}_k = \mathsf{lbfgs}(-\mathbf{g}_k, \{\mathbf{t}_l\}_{l=k-m}^{k}, \{\mathbf{y}_l\}_{l=k-m}^{k})$ // low-rank inverse Hessian
  $\{\mathbf{m}_{k+1}, f_{k+1}, \mathbf{g}_{k+1}\} = \mathsf{linesearch}(f_k, \mathbf{g}_k, \delta\mathbf{m}_k)$
  if linesearch successfull **then**
    $\mathbf{t}_{k+1} = \mathbf{m}_{k+1} - \mathbf{m}_k$, $\mathbf{y}_{k+1} = \mathbf{g}_{k+1} - \mathbf{g}_k$ // update L-BFGS vectors
    choose $\mathcal{I}_{k+1} \subseteq \{1, 2, \ldots, M\}$ s.t. $|\mathcal{I}_{k+1}| = b$ // draw new sample
    $\{f'_{k+1}, \mathbf{g}'_{k+1}\} = \mathsf{misfit}(\mathbf{m}_{k+1}, \mathcal{I}_{k+1}, \eta)$ // misfit & gradient new sample
    if $(f_{k+1} + f'_{k+1}) \geq (f_k + f'_k)$ **then**
      $b = \min(b + \beta, b_{\max})$ // increase batch
    end if
    $f_{k+1} = f'_{k+1}$, $\mathbf{g}_{k+1} = \mathbf{g}'_{k+1}$, $k = k + 1$ // Use new misfit & gradient
  else
    $\eta = \eta/2$ // narrow tolerenance
  end if
end while

Edam model

Modelling:

- Spherical anomaly w/ constant velocity of 2500 m/s in constant background of 2000 m/s.
- The model is 1 km in each direction and is discretized with 20 m gridspacing and 10 points are added on each side for the PML layer, leading to a total gridsize of 71 X 71 X 71.
- ”Observed” data are generated by solving the Helmholtz equation up to $\epsilon = 10^{-6}$ for 9 sources (y=0m), 2601 receivers (y=1000m plane) and 3 frequencies 5, 10 and 15 Hz.

Inversion:

- Use all sources & frequencies for $\eta = \{0.1, 0.05, 0.01\}$

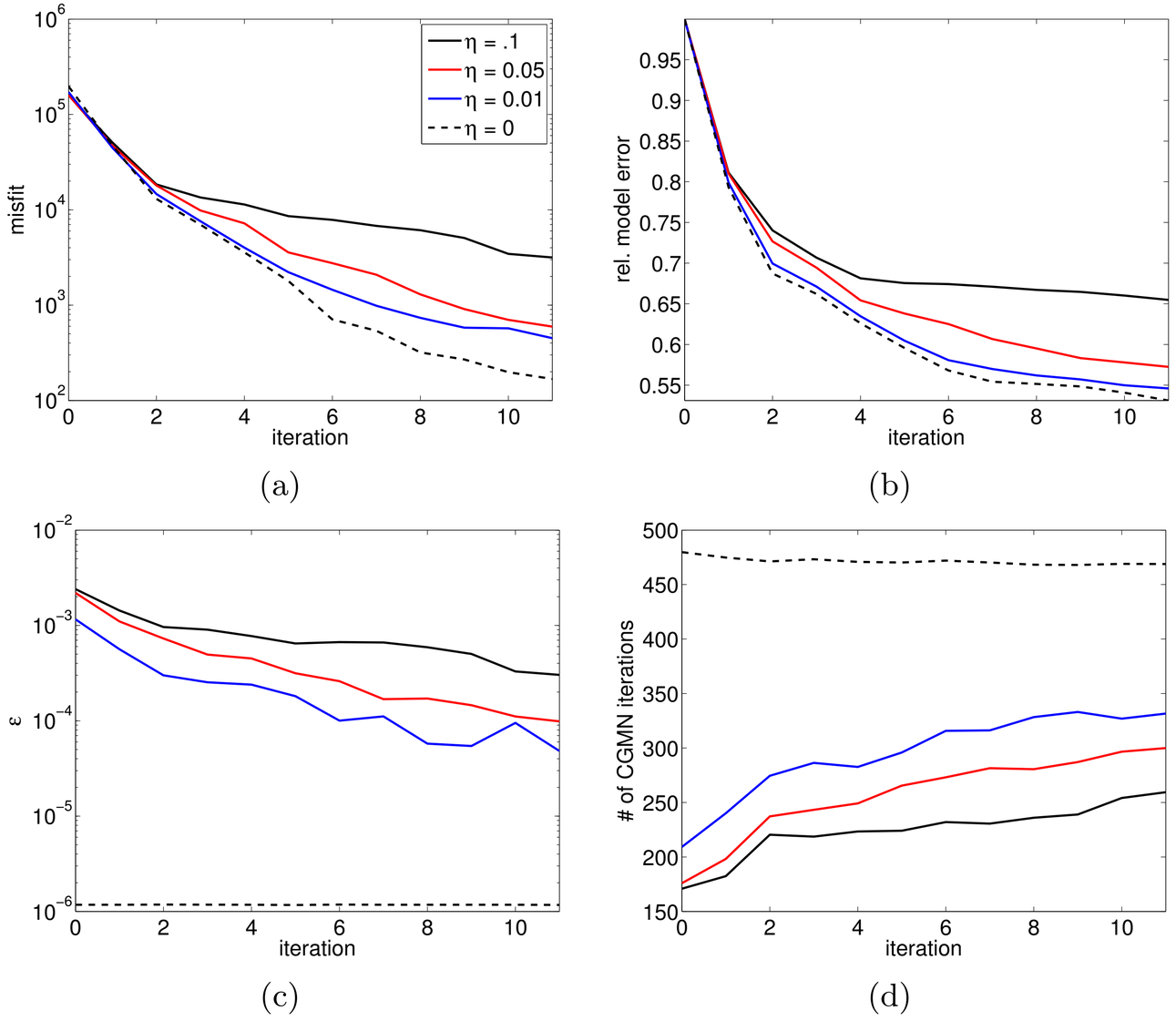


FIG. 4.5. *Convergence in terms of the (a) misfit and (b) rel. model error for the Edam model. The average (over all sources and frequencies) tolerance used by CGMN for each outer iteration is shown in (c) while (d) shows the corresponding number of CGMN iterations required. Using more accurate PDE solves (smaller η) yields results closer to the baseline result $\eta = 0$ as can be seen from (a) and (b). The tolerance used for the PDE solves (c) automatically decreases but stays well above the baseline tolerance of 10^{-6}. Consequently, the number of CGMN iterations required is much lower as can be seen in (d).*

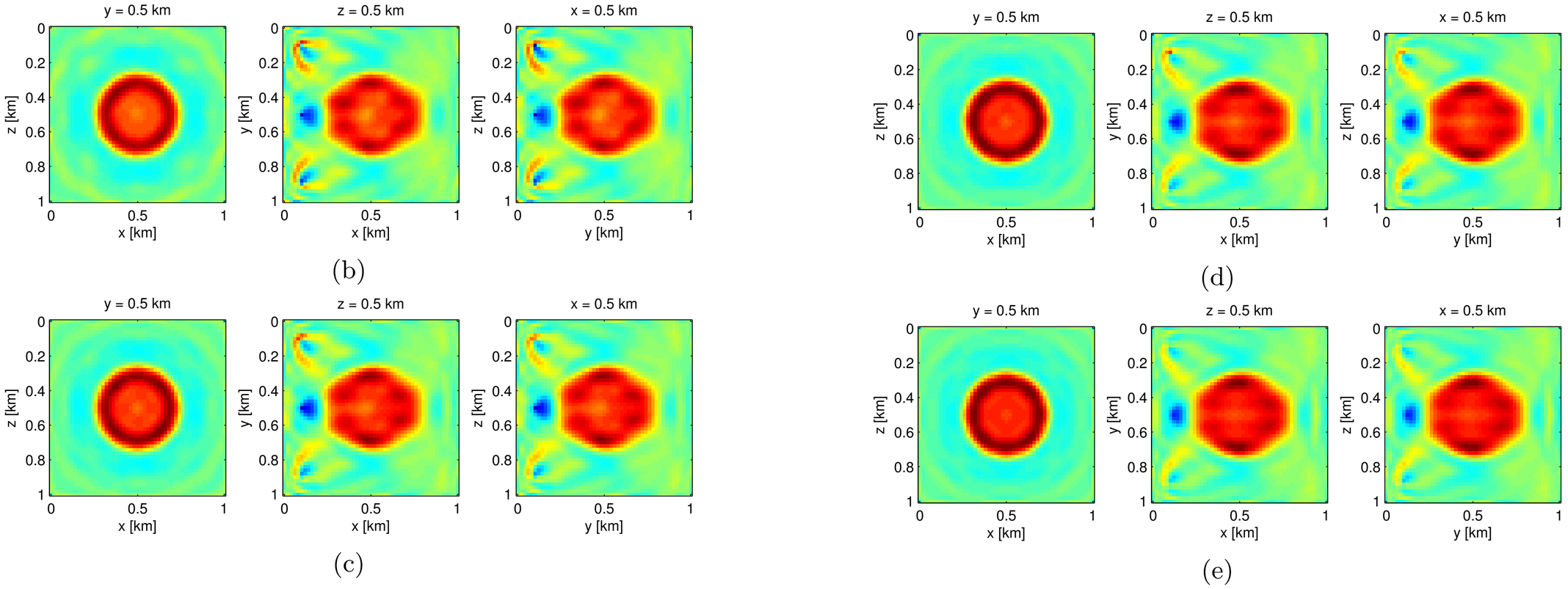


FIG. 4.8. *(The true velocity model used for the Edam experiment is shown in (a). The Reconstructed models using (b) $\eta = 0.1$, (c) $\eta = 0.05$, (d) $\eta = 0.01$ and (e) $\eta = 0$ are also shown on the same colorscale. All the reconstructions are very reasonable when compared to the baseline result (e). Using more accurate solves $\eta = 0.01$ (d) yields less artifacts and is almost identical to the baseline result, however, the computational cost of the baseline was roughly twice as high.*

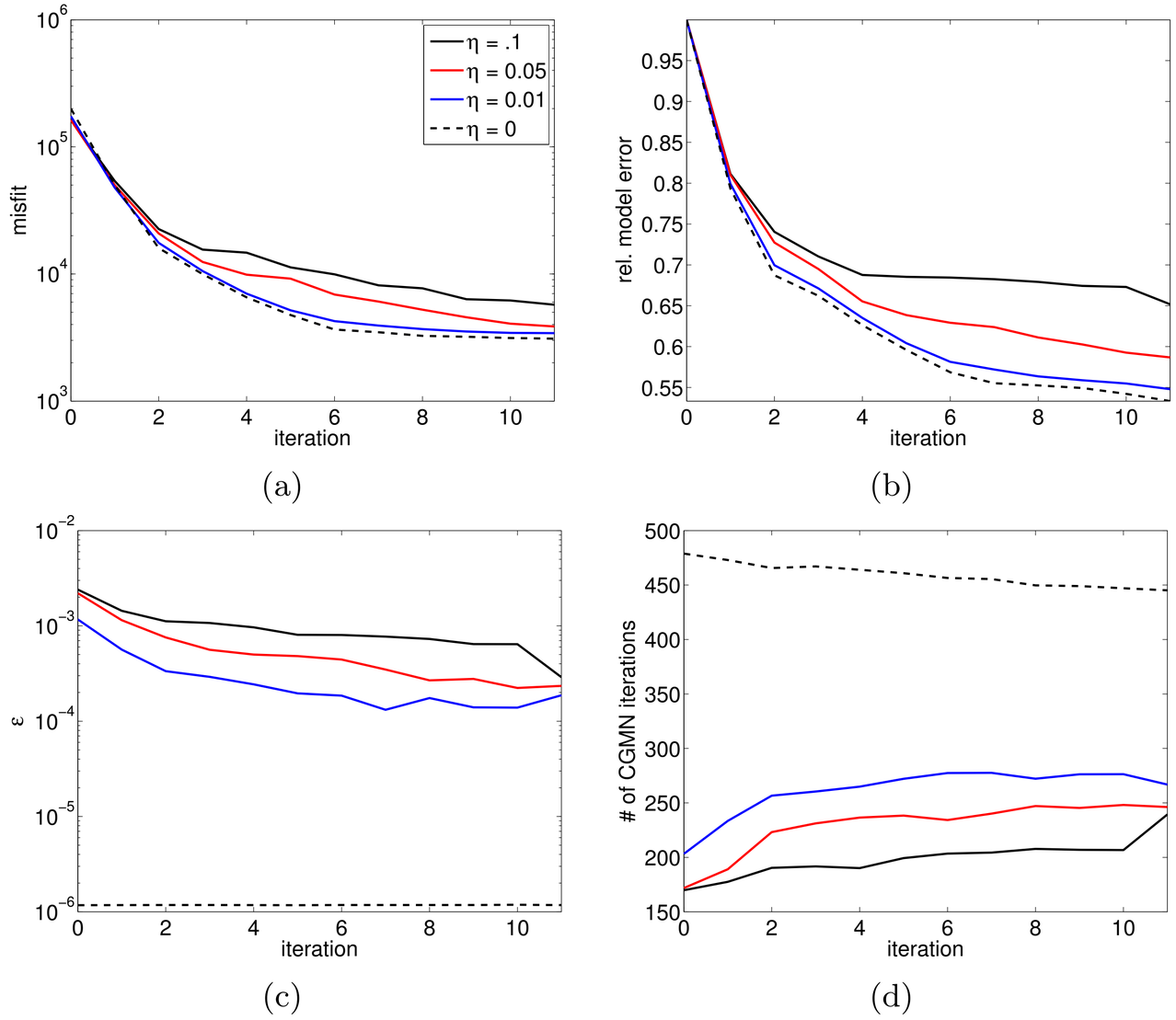


FIG. 4.6. *Convergence in terms of the (a) misfit and (b) rel. model error for the Edam model using noisy data (5 % noise). The average (over all sources and frequencies) tolerance used by CGMN for each outer iteration is shown in (c) while (d) shows the corresponding number of CGMN iterations required. Using more accurate PDE solves (smaller η) yields results closer to the baseline result $\eta = 0$ as can be seen from (a) and (b). The tolerance used for the PDE solves (c) automatically decreases but stays well above the baseline tolerance of 10^{-6}. Consequently, the number of CGMN iterations required is much lower as can be seen in (d).*

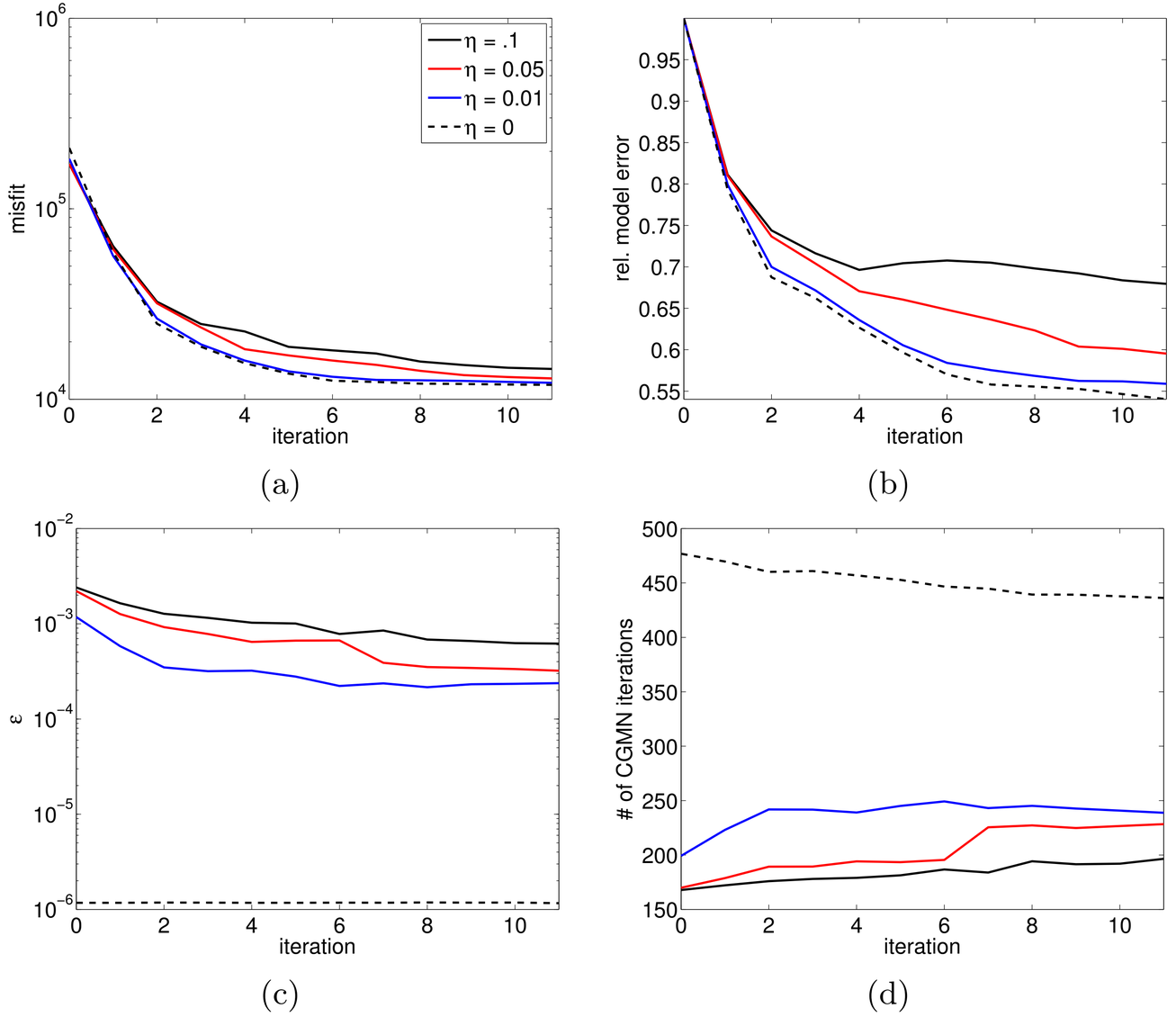


FIG. 4.7. *Convergence in terms of the (a) misfit and (b) rel. model error for the Edam model using noisy data (10 % noise). The average (over all sources and frequencies) tolerance used by CGMN for each outer iteration is shown in (c) while (d) shows the corresponding number of CGMN iterations required. Using more accurate PDE solves (smaller η) yields results closer to the baseline result $\eta = 0$ as can be seen from (a) and (b). The tolerance used for the PDE solves (c) automatically decreases but stays well above the baseline tolerance of 10^{-6}. Consequently, the number of CGMN iterations required is much lower as can be seen in (d).*

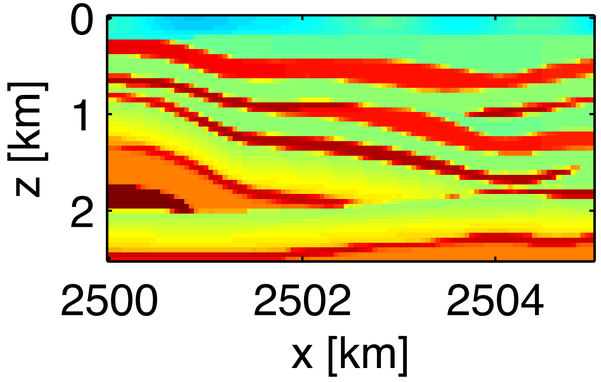
Overthrust model

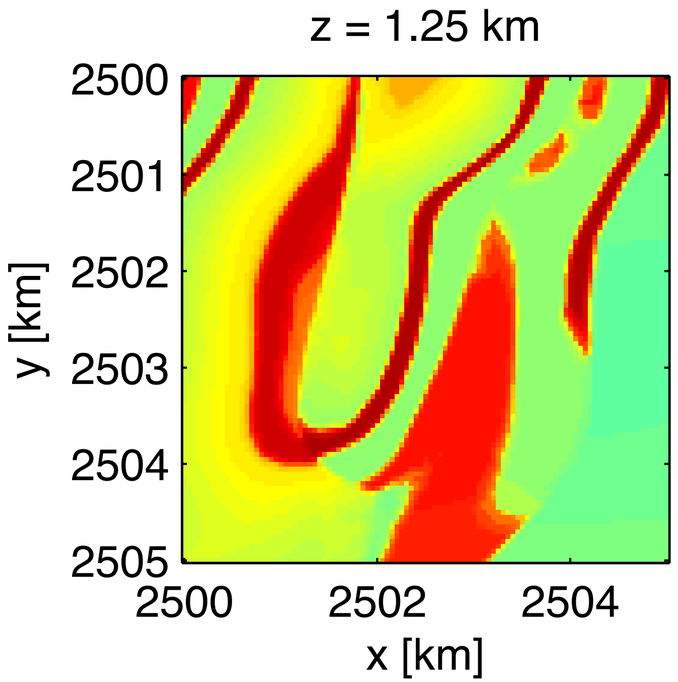
Modellling:

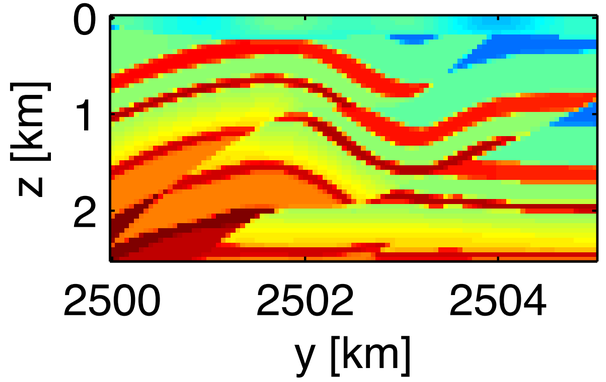
- We generate data using iWAVE, an open-source time-domain finite-difference code
- for a 5 km x 5 km central part of a well-known this benchmark model at 50m gridspacing
- A total of 121 sources and 2601 receivers (both regularly spaced) cover the top of the model

Overthrust model

true model
5km X 5km X 2.5Km
121 sources & 2601 receivers

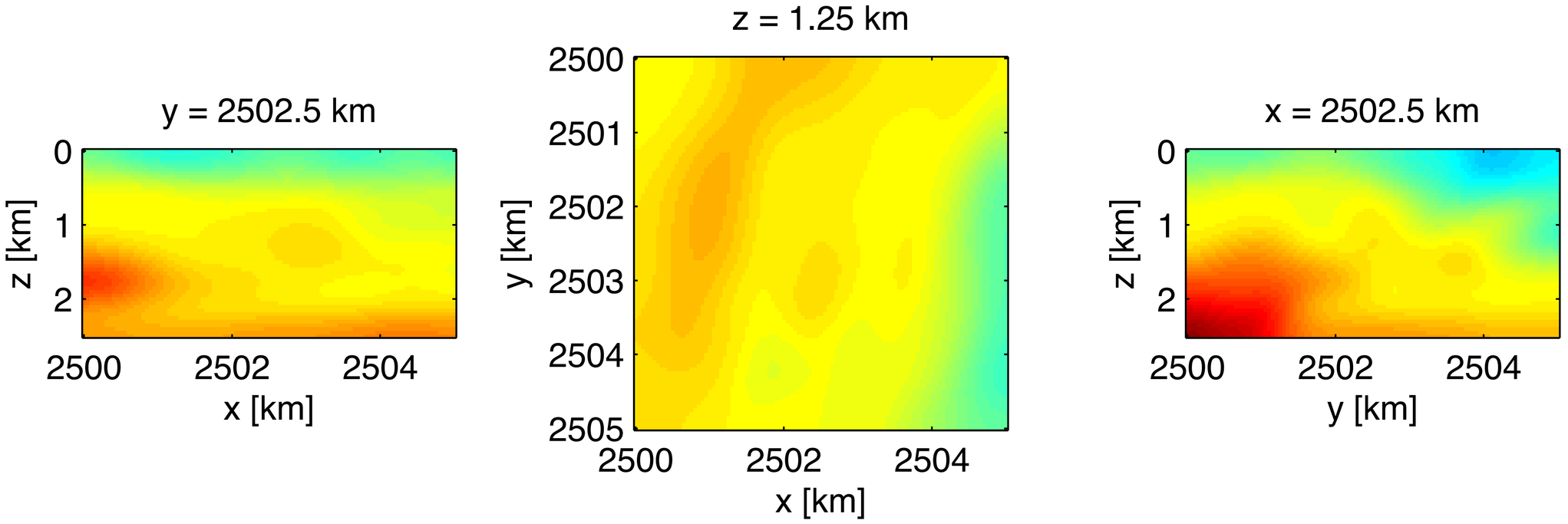






Overthrust model

initial model



Overthrust model

Experiment I:

- invert single frequency of 4Hz
- 100m grid size leading to 36 X 61 X 61
- batch sizes are $b = \{1, 20, 50, 121\}$

Experiment II:

- multiscale inversion $f = \{4, 6, 8\}$ Hz. consecutively using a gridspacing of
- 100 m, 66.67 m and 50 m, respectively
- gridsizes of 36 X 61 X 61, 48 X 86 X 86, and 61 X 111 X 111
- We use either fixed number of b=1 and b=121 sources or an increasing number of sources, starting from b=1.
- For all cases we perform 2 passes through the data for each frequency.

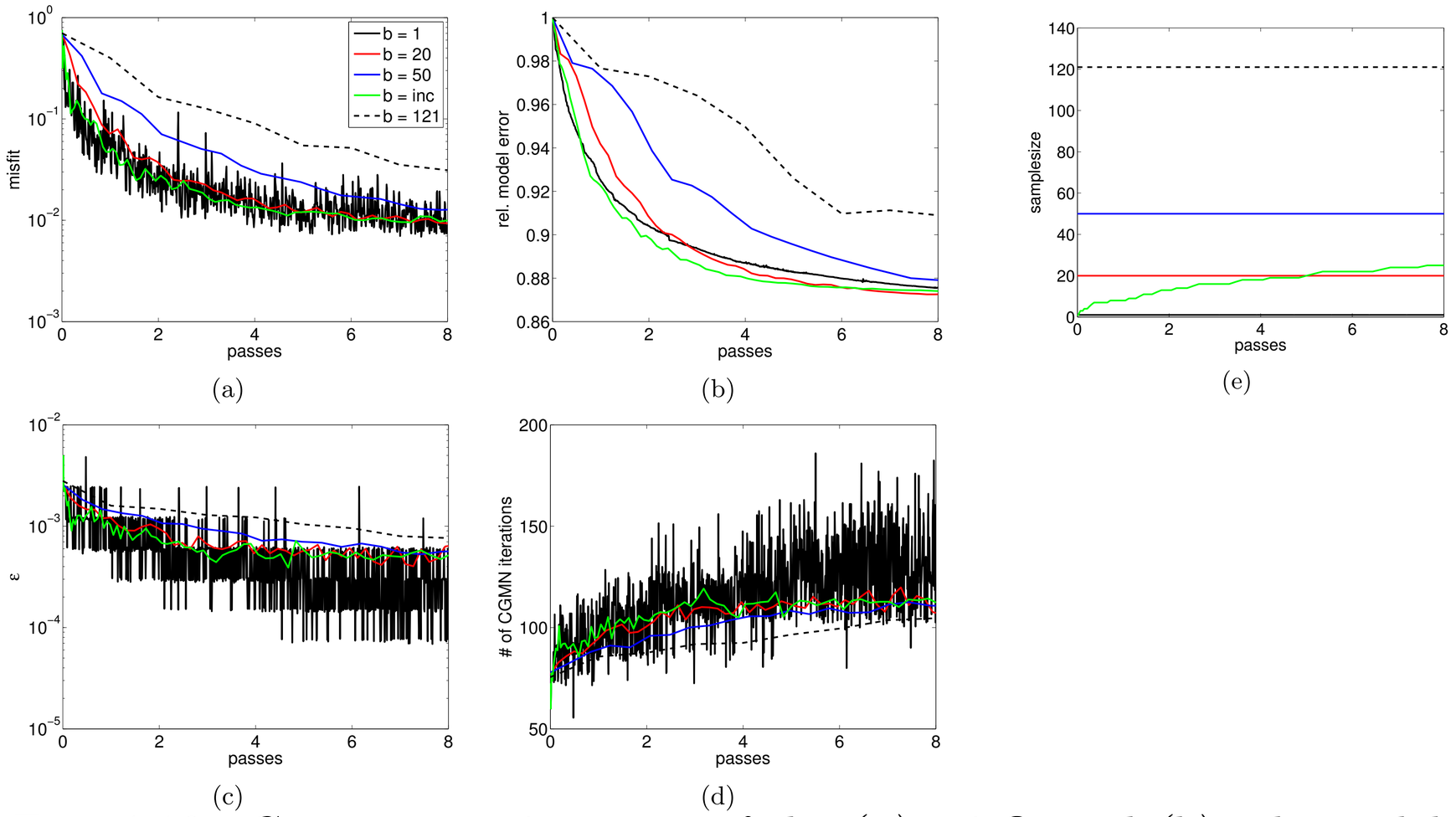
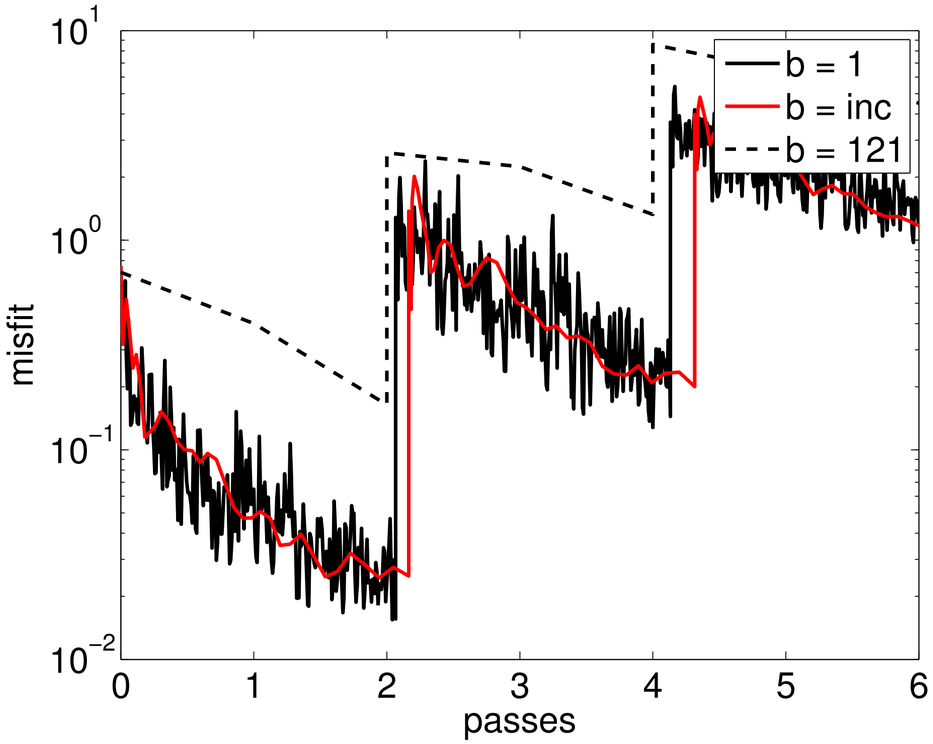


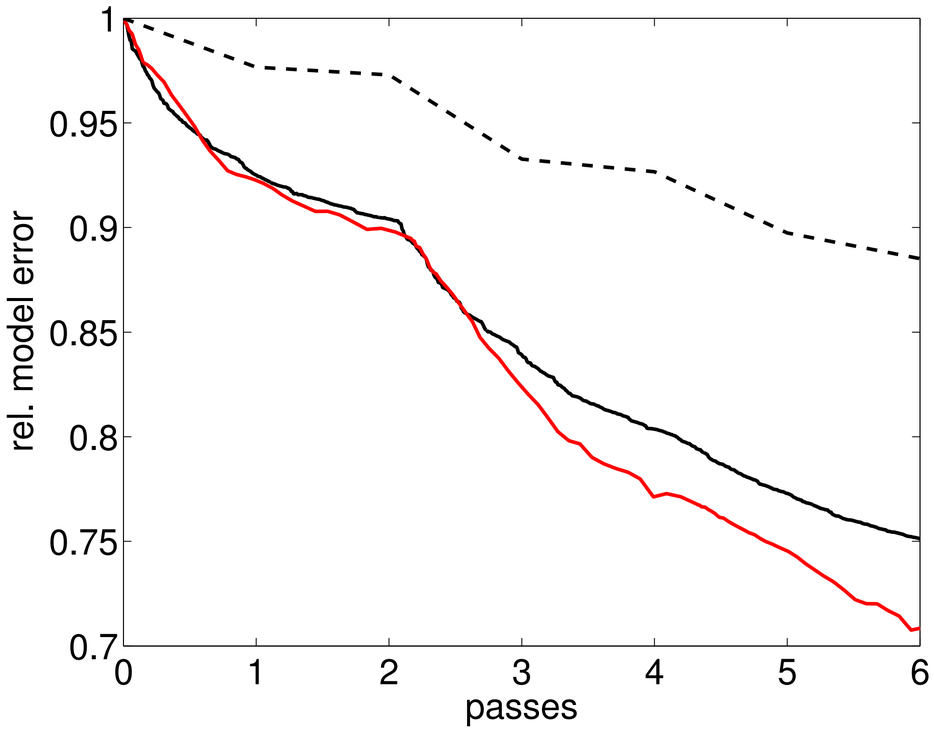
FIG. 4.10. *Convergence in terms of the (a) misfit and (b) rel. model error for the first Overthrust experiment. The average (over all sources) tolerance used by CGMN for each outer iteration is shown in (c) while (d) shows the corresponding number of CGMN iterations required. All results are shown as a function of the effective number of complete passes through the data, and hence have the same computational cost. The sample-size is shown in (e).*

Performance
misfit & relative model error

misfit

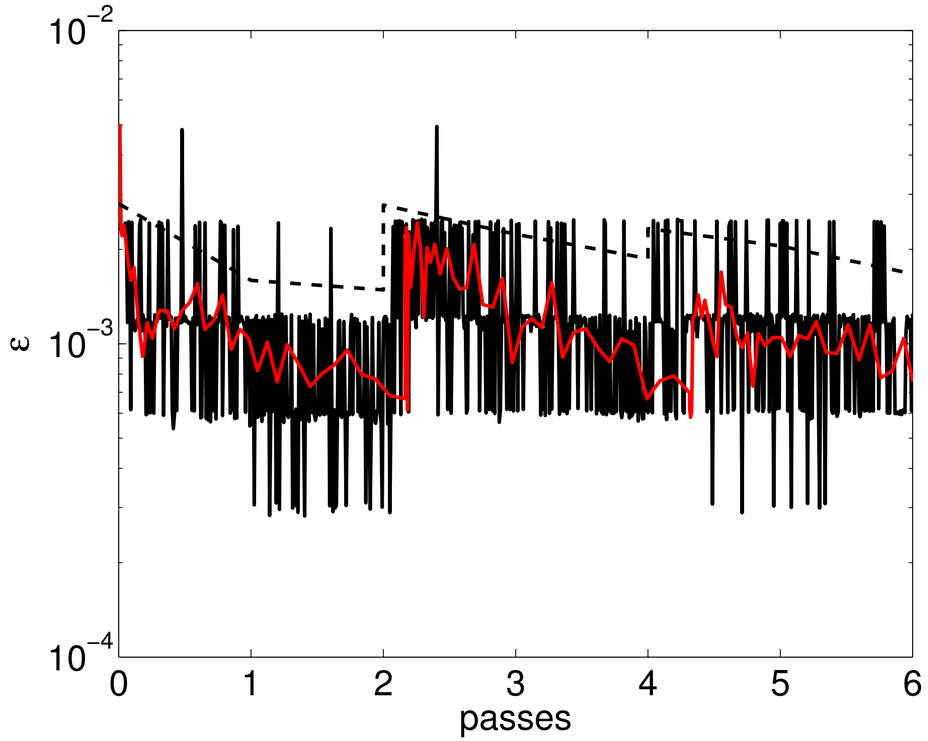


model error

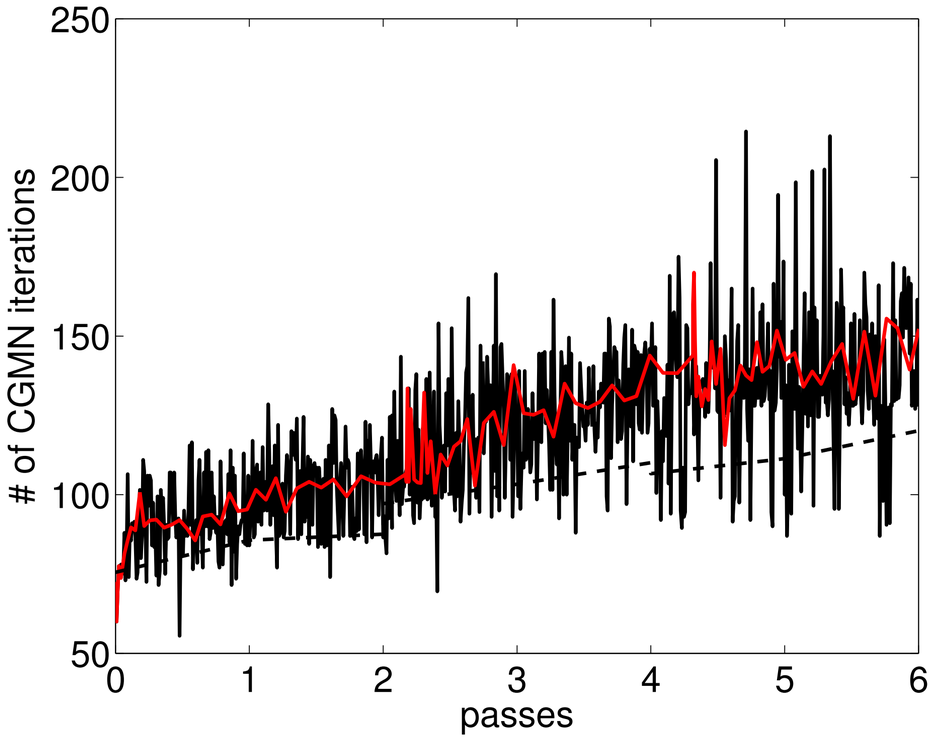


Performance
tolerance & # CARP-CG iterations

accuracy

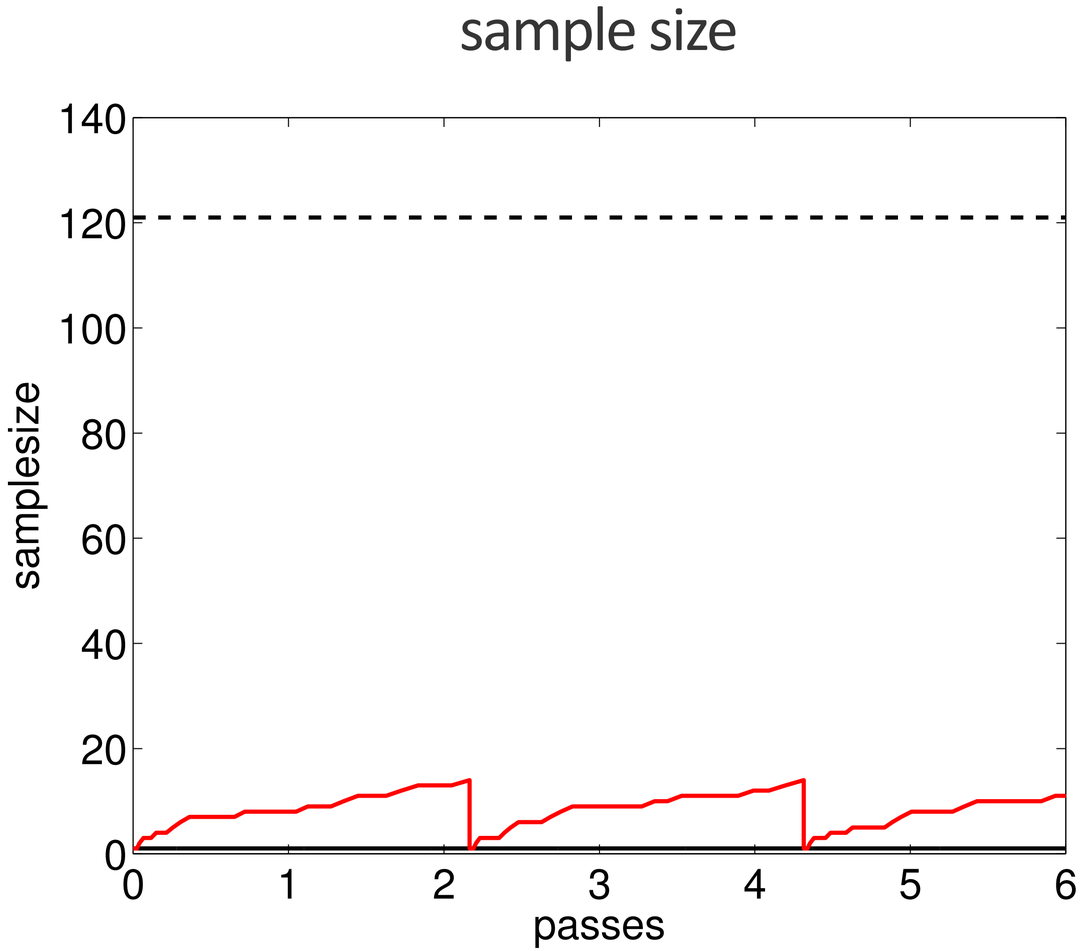


# of iterations



Performance

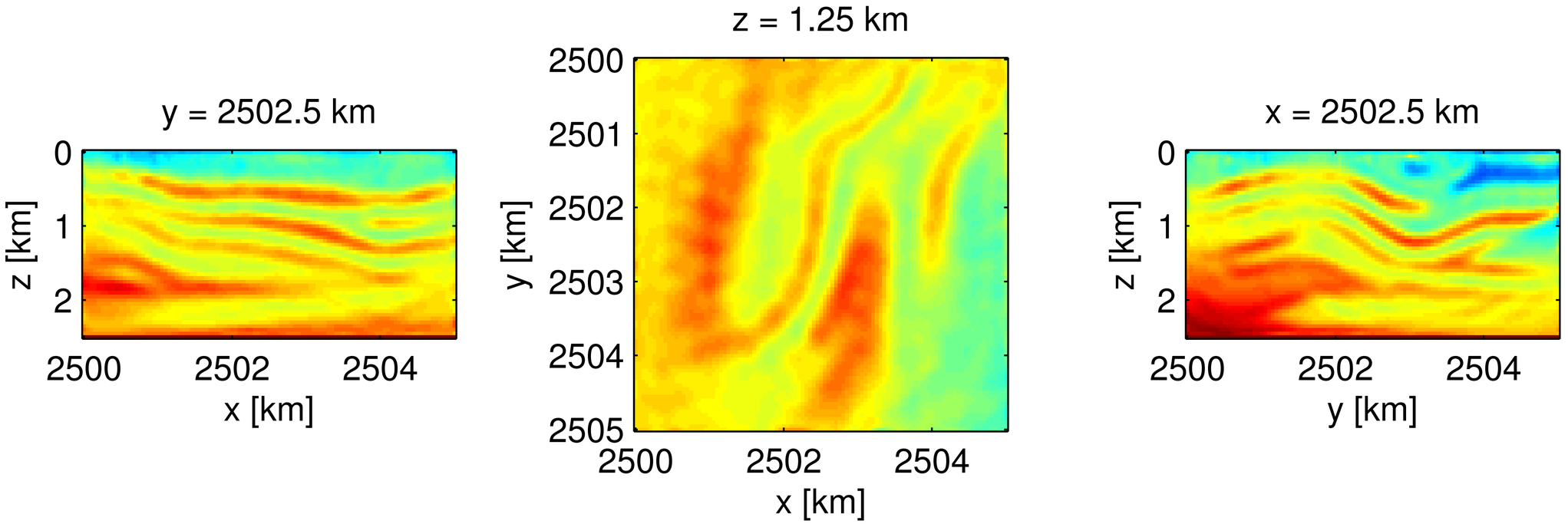
sample size



Overthrust model

recovered model *w/ b=1*

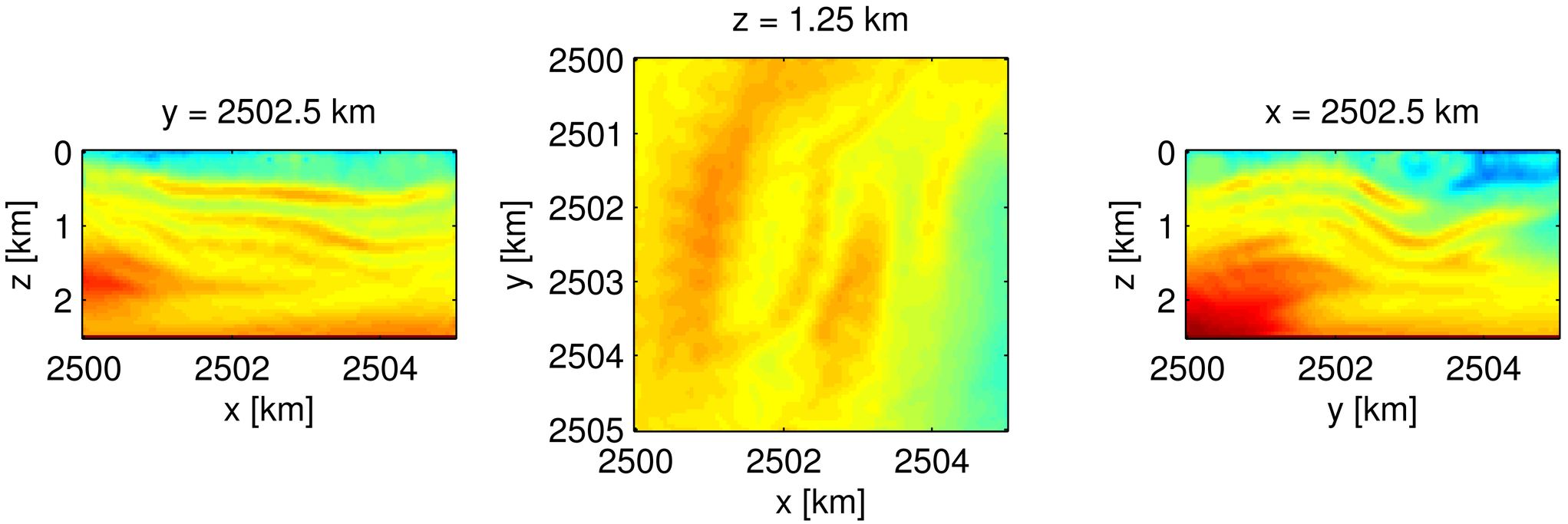
2 passes through *data* for each *(4,6,8) Hz*



Overthrust model

recovered model *w/ b=121*

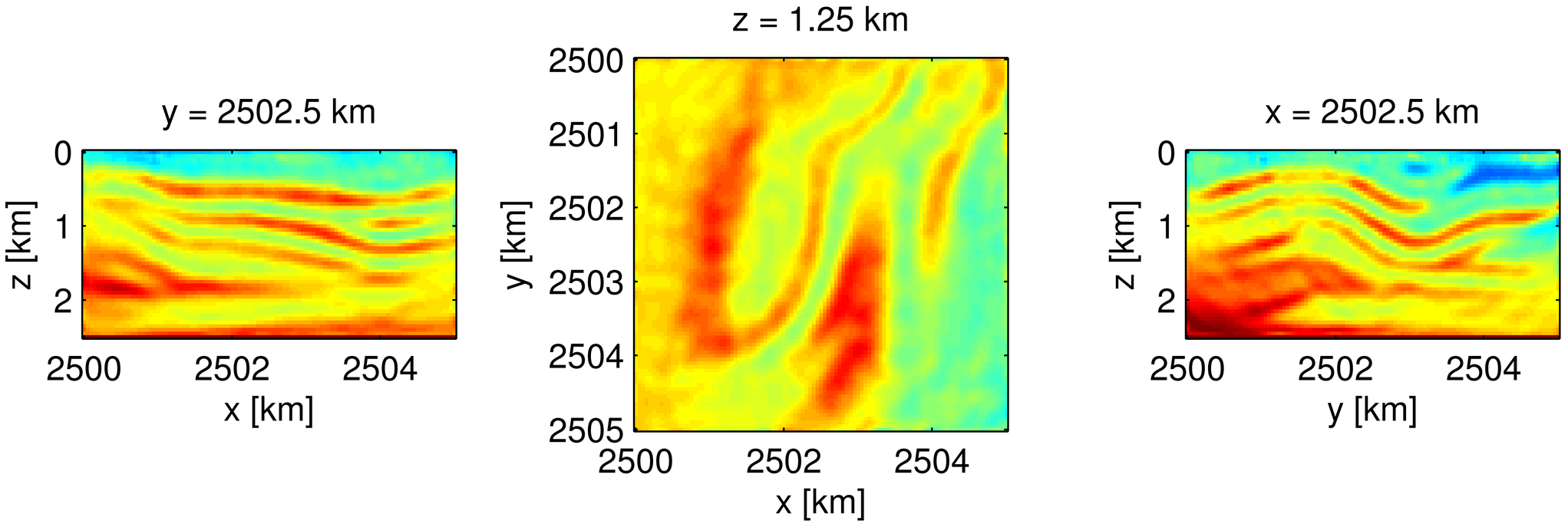
2 passes through *data* for each *(4,6,8) Hz*



Overthrust model

growing sample size

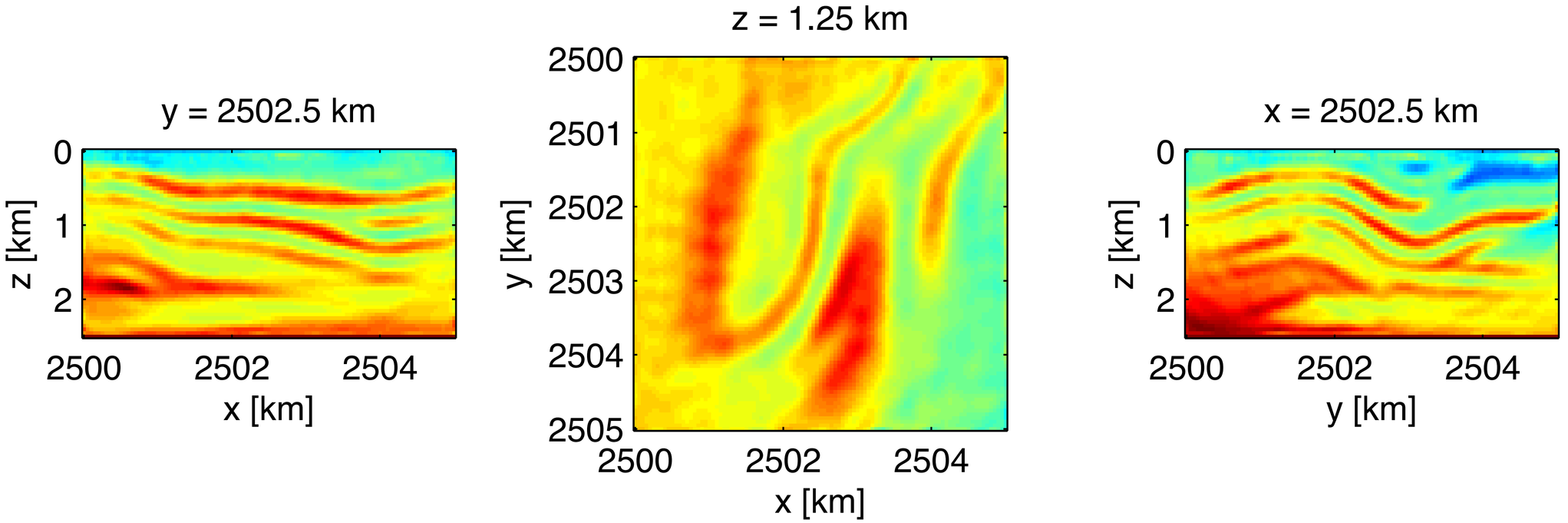
2 passes through *data* for each *(4,6,8) Hz*



Overthrust model

growing sample size

10 passes through *data* for each *(4,6,8) Hz*



Observations

Able to carry out 3-D FWI with *dynamic*

- growth of *sample* size
- *tolerance* PDE solves

Model error decays much *faster* compared to *working* with *all* data

Summary

Main *ingredients* for a *scalable* approach to 3D FWI:

- *iterative* Helmholtz solver w/ *little* memory imprint, computational overhead, and model-dependent tuning
- practical *stopping* criterion for wave simulator
- (stochastic) optimization technique that exploits the *separable* structure of FWI by working w/ *small* subsets
- *strategy* to *increase* sample size and accuracy as needed

Future plans

Use the same *heuristic*

- FWI w/ *penalty* method
- WEMVA w/ random *probing*

Incorporate composite shots from sim. marine

Build in adaptive (stratified) sampling

Carry home message

Insisting on working w/

- *all* data
- *full* accuracy

can be *detrimental* to *FWI* because you can not get there.

When *ill*-conditioned use *less* rather than *more* data & *accuracy.*

Better to *call* for *more* data & *accuracy* only when *strictly* needed.

Less can be really more…

Robust FWI

Aleksandr Y. Aravkin, Michael P. Friedlander, Felix J. Herrmann, and Tristan van Leeuwen, "**Robust inversion, dimensionality reduction, and randomized sampling**", *Mathematical Programming*, vol. 134, p. 101-125, 2012.
Aleksandr Y. Aravkin, Tristan van Leeuwen, Henri Calandra, and Felix J. Herrmann, "**Source estimation for frequency-domain FWI with robust penalties**", in *EAGE Annual Conference Proceedings*, 2012.

MAP estimation

measurement model:

$$\mathbf{d}_i = F_i(\mathbf{m}) + \mathbf{n}_i$$

posterior likelihood:

$$\pi_{\text{post}}(\mathbf{m}) \sim \prod_{i=1}^{K} \pi_{\text{noise}}(F_i(\mathbf{m}) - \mathbf{d}_i)\pi_{\text{prior}}(\mathbf{m})$$

MAP estimation

Maximization of the likelihood

$$\max_{\mathbf{m}} \pi_{\text{post}}(\mathbf{m})$$

is equivalent to

$$\min_{\mathbf{m}} -\log\left(\pi_{\text{post}}(\mathbf{m})\right)$$

MAP estimation

For Gaussian noise we have

$$\pi_{\text{noise}}(\mathbf{r}) \sim \exp\left(-||\mathbf{r}||_2^2\right)$$

which leads to the usual least-squares formulation

$$\min_{\mathbf{m}} \sum_i ||F_i(\mathbf{m}) - \mathbf{d}_i||_2^2$$

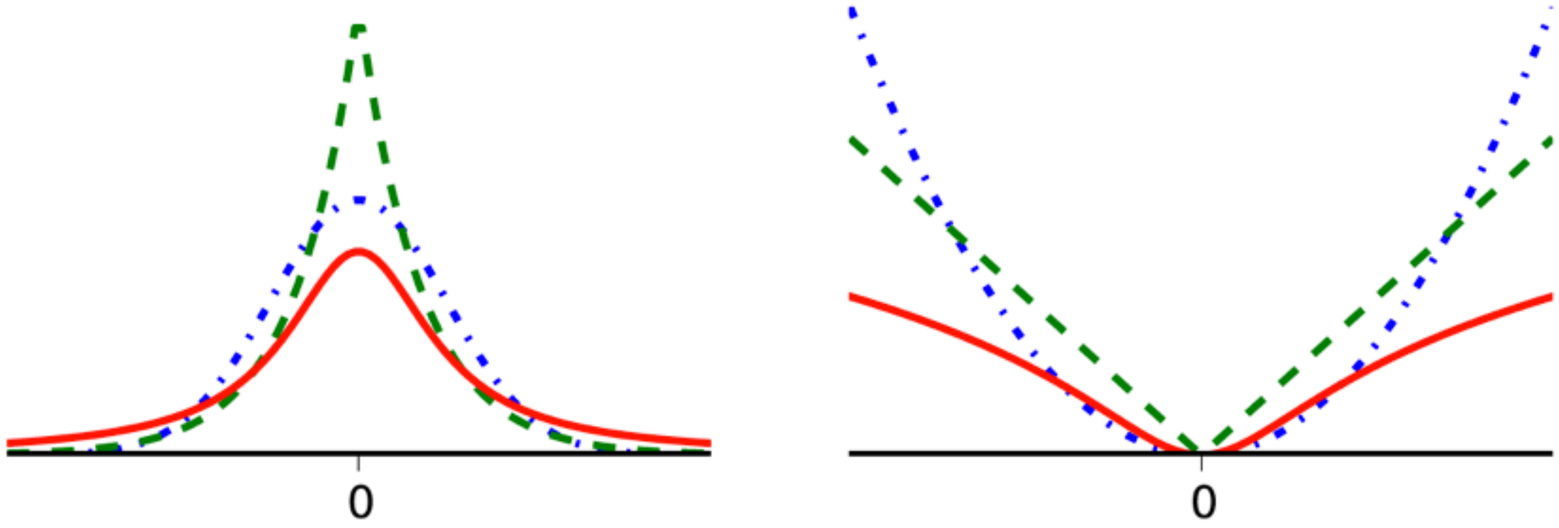
MAP estimation

The use of alternative penalties can be interpreted as using a different noise model

$$\min_{\mathbf{m}} \sum_i \rho\left(F_i(\mathbf{m}) - \mathbf{d}_i\right)$$

MAP estimation

densities & penalties

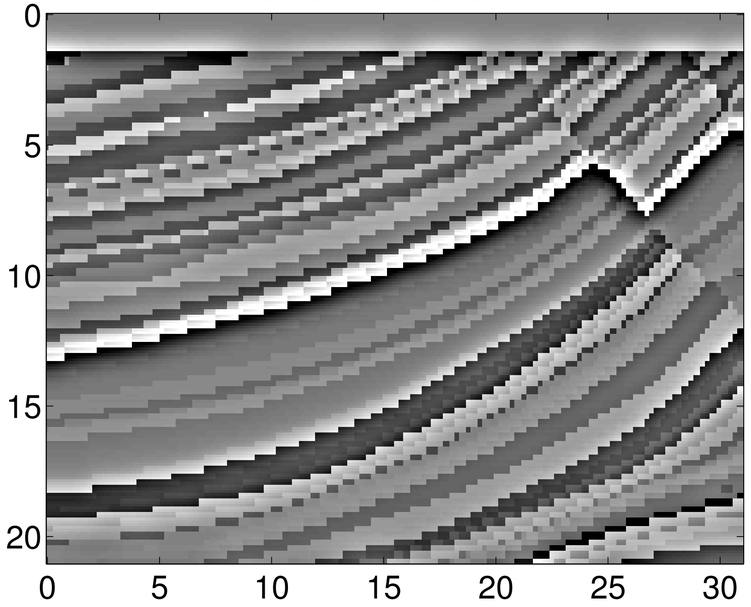


Gaussian, Laplace and Students T

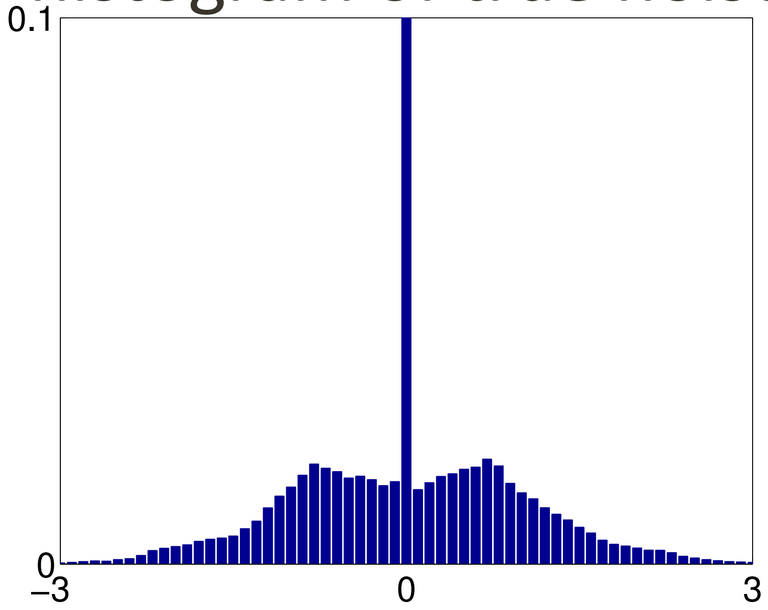
MAP estimation

data with 50% “bad traces”

true model



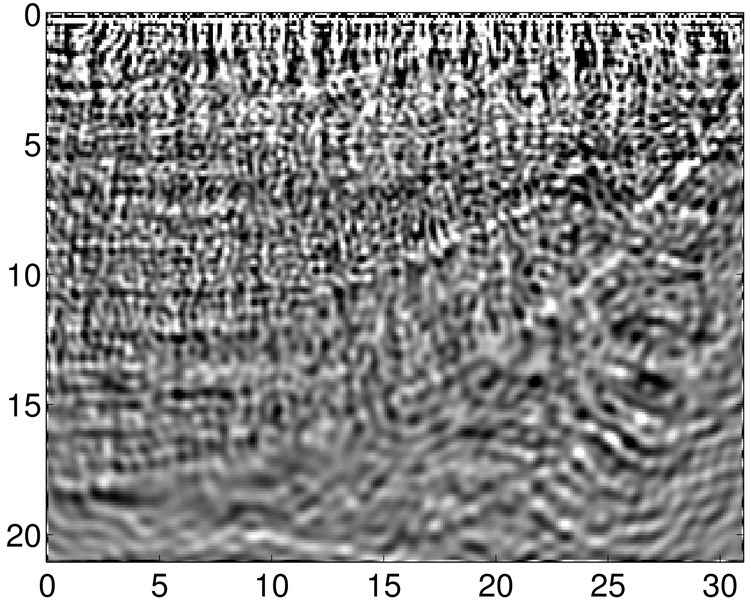
histogram of true noise



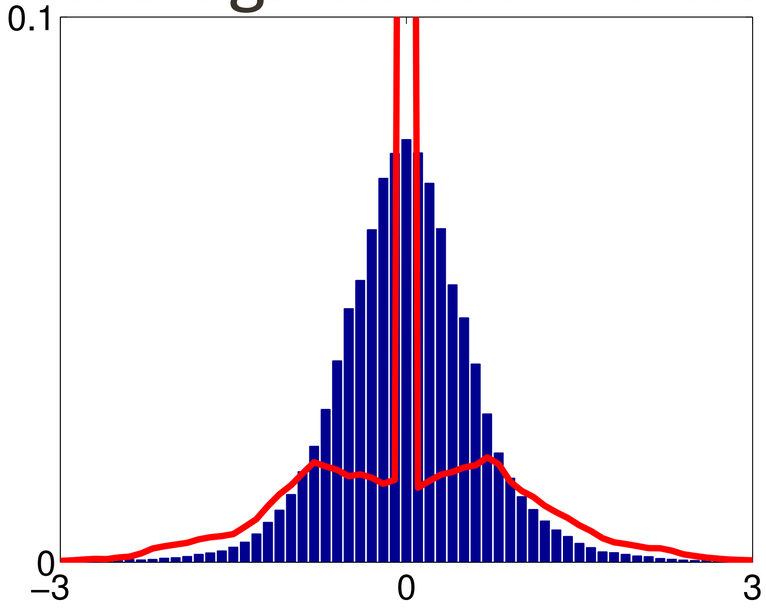
MAP estimation

least-squares penalty

recovered model



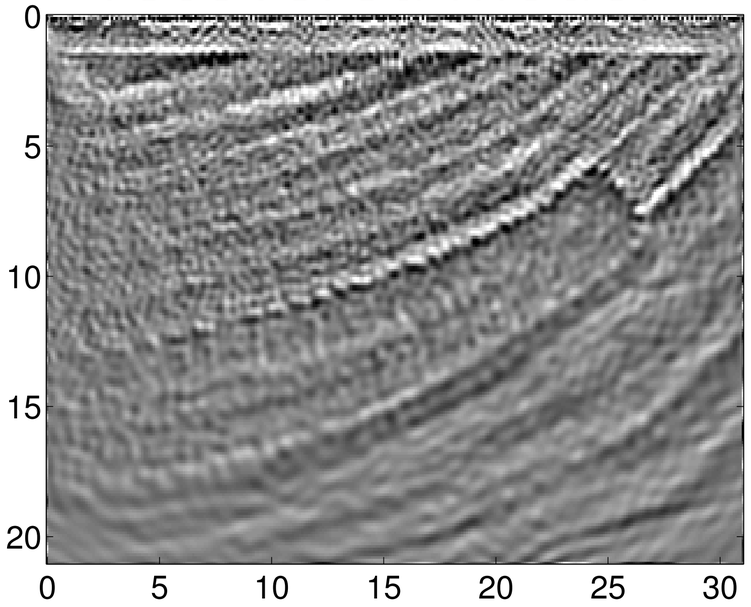
histogram of residual



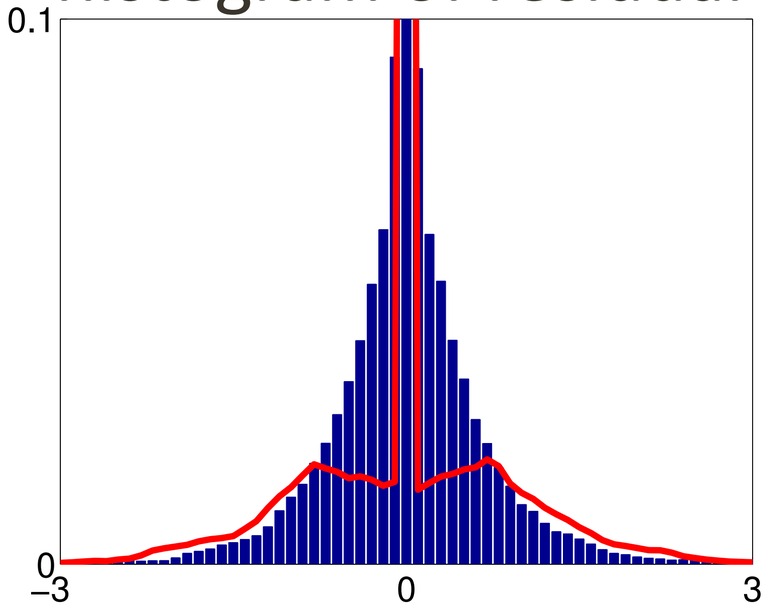
MAP estimation

Huber penalty

recovered model



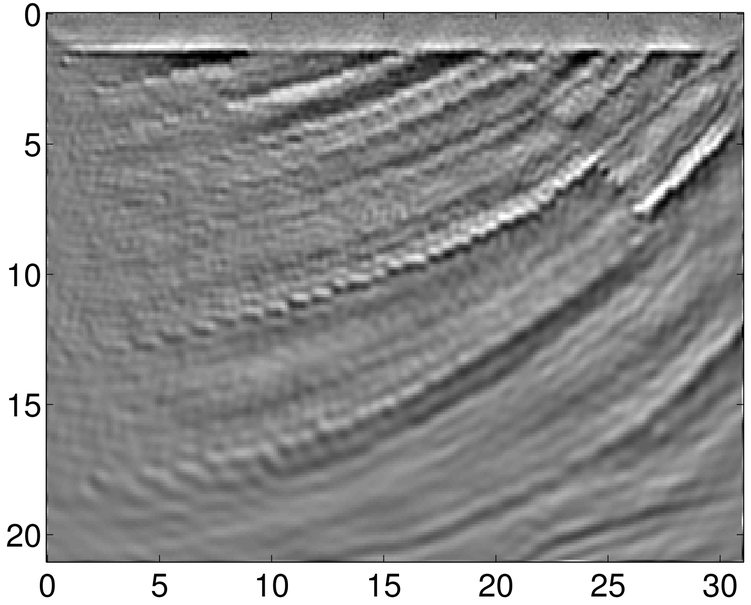
histogram of residual



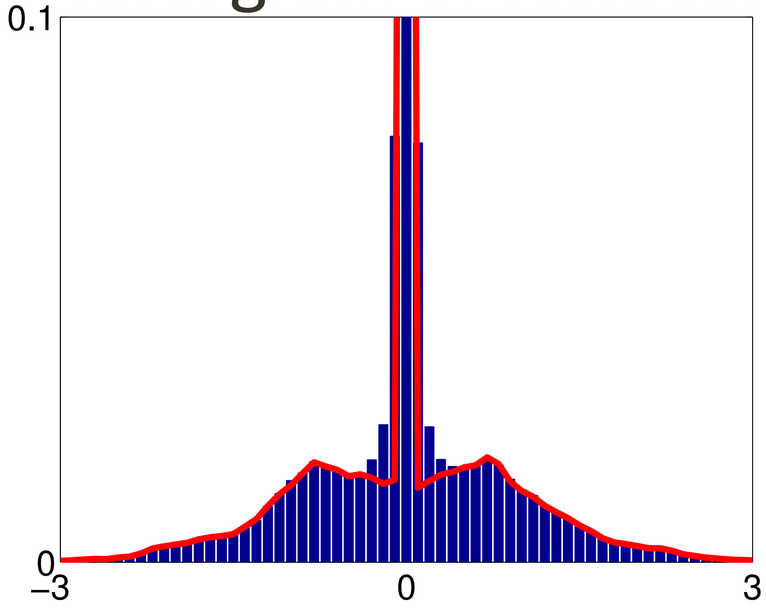
MAP estimation

Students T penalty

recovered model



histogram of residual

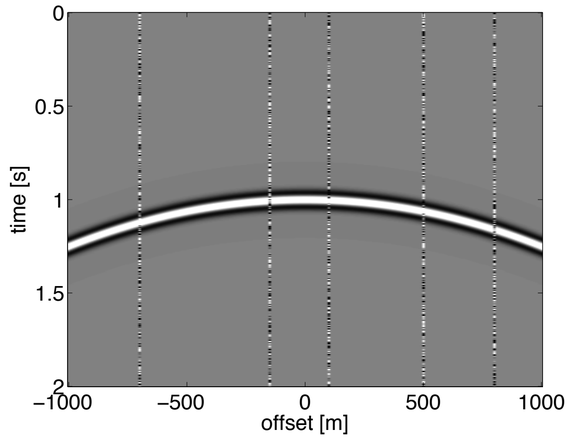


MAP estimation

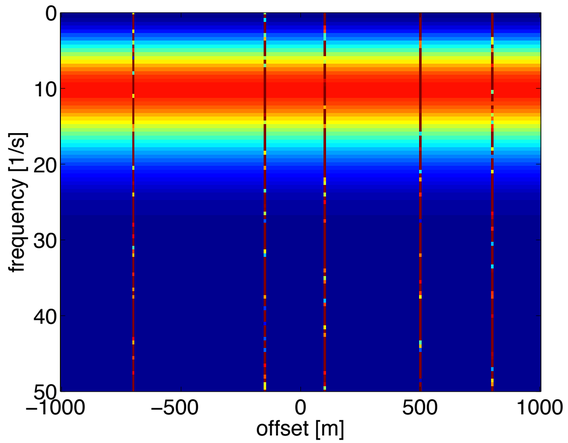
- Noise does *not* come from Students T distribution
- Use of Students T penalty may still be beneficial
- Noise has to be *spiky*

Outliers

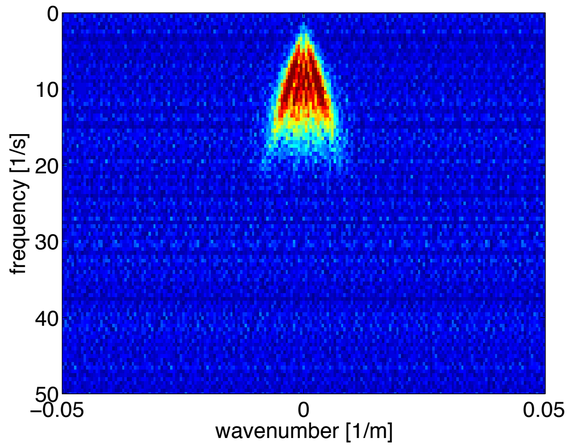
What *is* an outlier?



t,x ✓



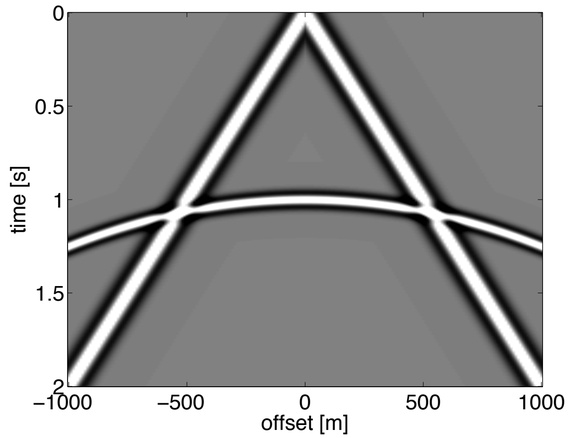
f,x ✓



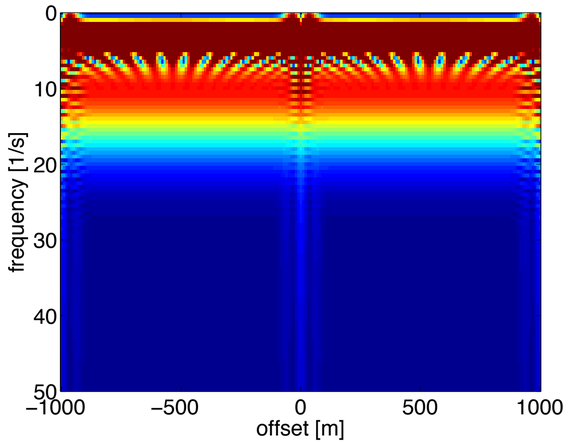
~~f~~,k ✗

Outliers

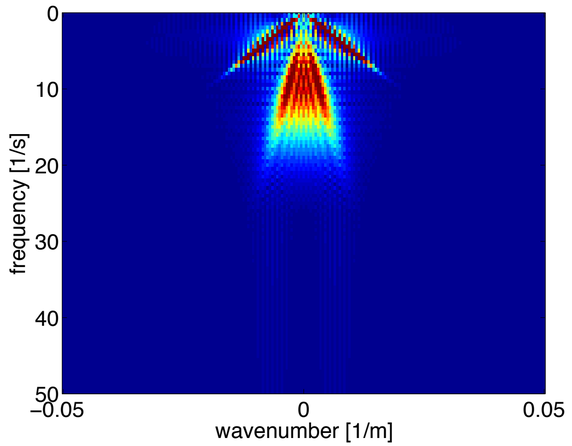
What *is* an outlier?



t,x ✓



f,x ✗



✓ f,k

Outliers

Measure the misfit in a domain that *sparsifies* the noise

$$\min_{\mathbf{m}} \sum_i \rho\left(B\left(F_i(\mathbf{m}) - \mathbf{d}_i\right)\right)$$

e.g., Fourier, Radon, Curvelets,…

Students T

The penalty is given by

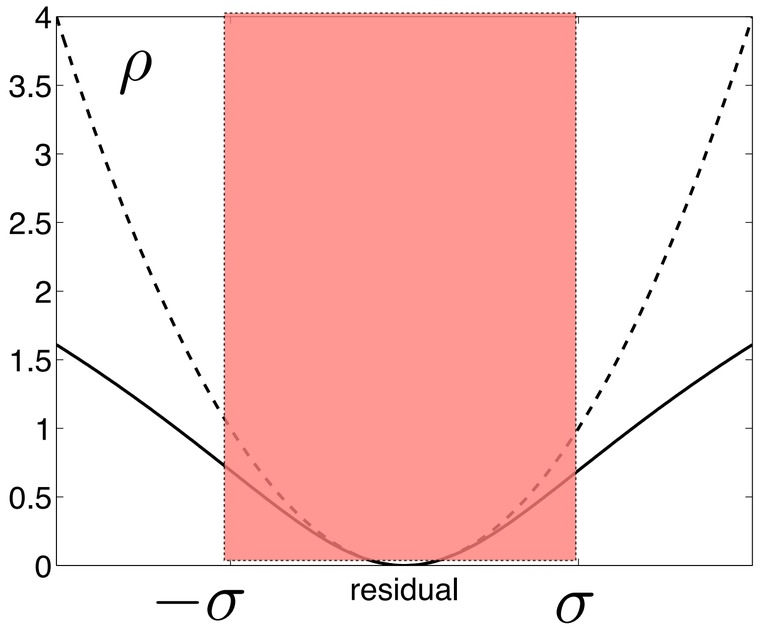
$$\rho(\mathbf{r}) = \sum_j \log(1 + |r_j|^2/\sigma^2)$$

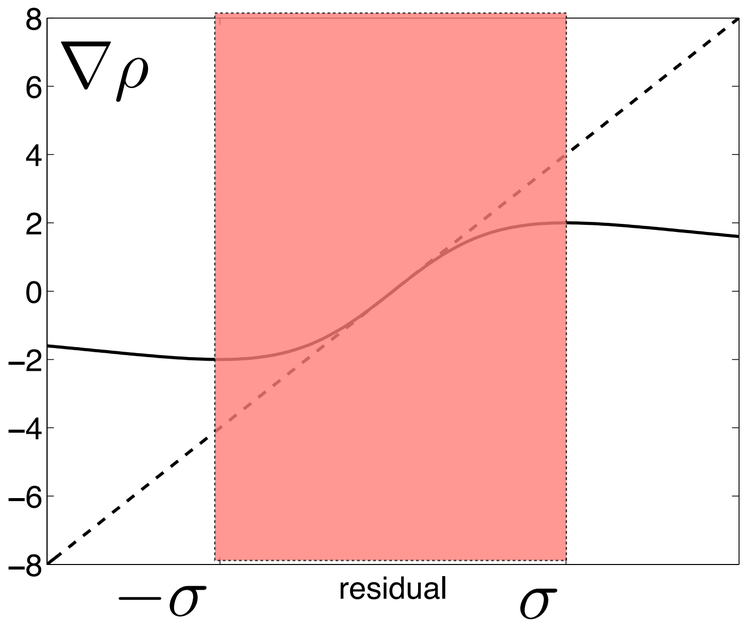
where σ is a scale parameter. The corresponding adjoint source is given by

$$(\nabla\rho)_j = \frac{2r_j}{|r_j|^2 + \sigma^2}$$

Students T

Scale parameter is used to separate outliers from good data





Students T

- *scale* parameter controls which residuals are ignored
- similar to a *weighted* least-squares approach
- how should we *choose* σ ?
- what about *source* estimation?

Source estimation

Use *variable* projection approach on

$$\min_{\mathbf{m},\mathbf{w}} \sum_i \rho\left(B\left(w_i F_i(\mathbf{m}) - \mathbf{d}_i\right)\right)$$

solve *source*-weights as

$$\min_{w_i} \rho\left(B\left(w_i F_i(\mathbf{m}) - \mathbf{d}_i\right)\right)$$

Auto-tuning

Extended Students T penalty:

$$\rho_\sigma(\mathbf{r}) = -N \log\left(\frac{\Gamma(\frac{\sigma^2+1}{2})}{\Gamma(\frac{\sigma^2}{2})\sqrt{\pi\sigma^2}}\right) + \frac{\sigma^2+1}{2}\sum_{j=1}^{N}\log(1 + r_j^2/\sigma^2)$$

find *optimal* σ for a given *residual* by solving

$$\min_\sigma \rho_\sigma(\mathbf{r})$$

Workflow

1. Forward modeling $\mathbf{d}_i^{\mathrm{pred}} = F_i(\mathbf{m}_k)$
2. Estimate source weight (scalar optimization)
3. Compute residual $\mathbf{r}_i = w_i \mathbf{d}_i^{\mathrm{pred}} - \mathbf{d}_i$

4. Estimate scale (scalar optimization)
5. Compute adjoint source $\widetilde{\mathbf{r}}_i = B^* w_i^* \nabla \rho(B\mathbf{r}_i)$

7. Compute gradient $\mathbf{g} = \sum_i \nabla F_i(\mathbf{m}_k)^* \widetilde{\mathbf{r}}_i$

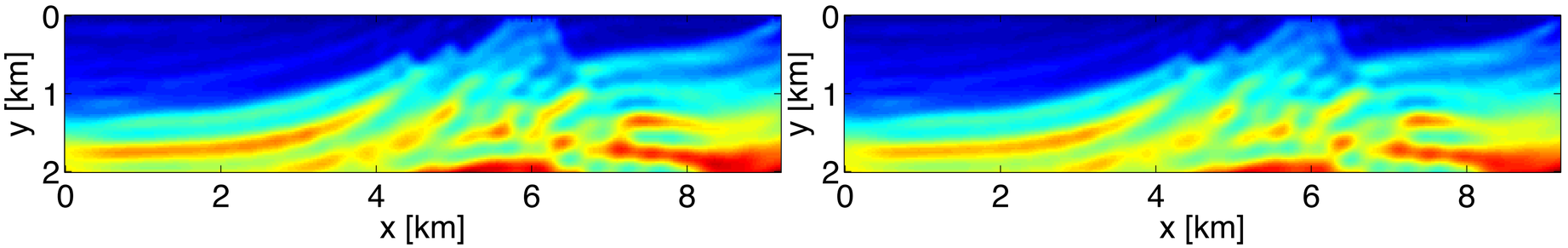
9. update $\mathbf{m}_{k+1} = \mathbf{m}_k - \lambda \mathbf{g}$

Results 1

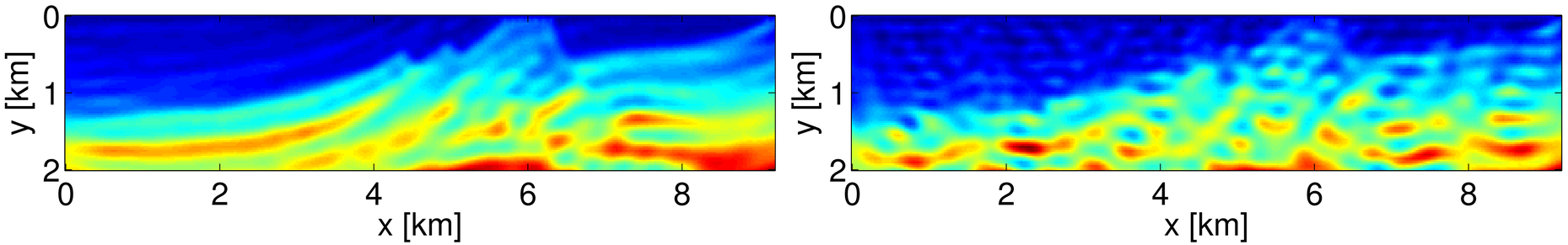
- Marmousi model with *periodic* noise.
- inversion of *single* frequency (4 Hz) with 20 iterations
- Misfit measured in *(f,x)* or *(f,k).*

Results 1

no noise:



periodic noise:

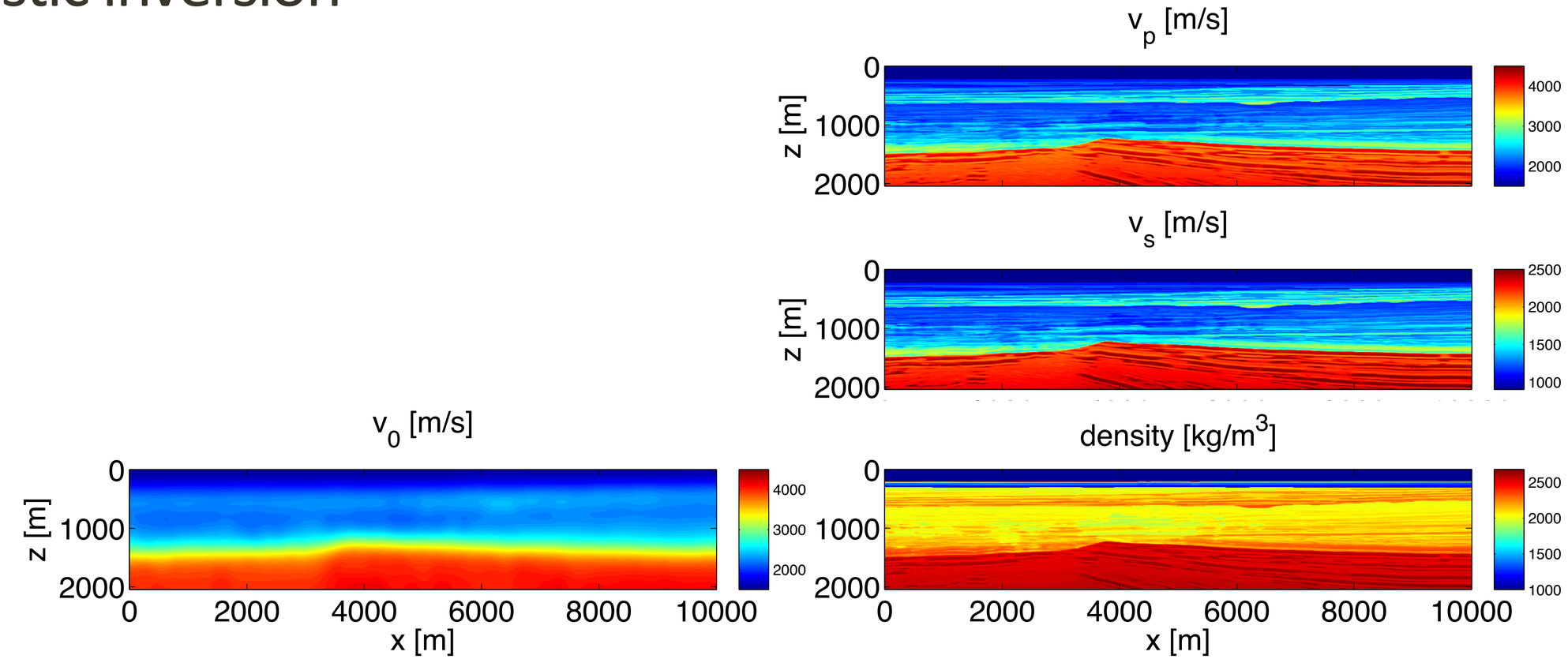


f-k

f-x

Results 2

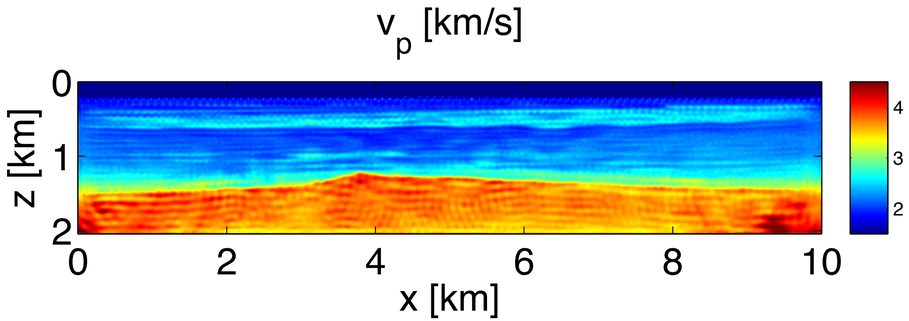
Acoustic inversion



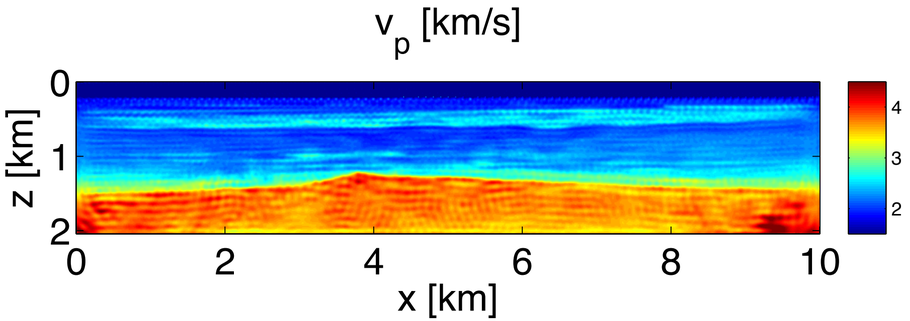
Results 2

Variable density data, no noise

least-squares



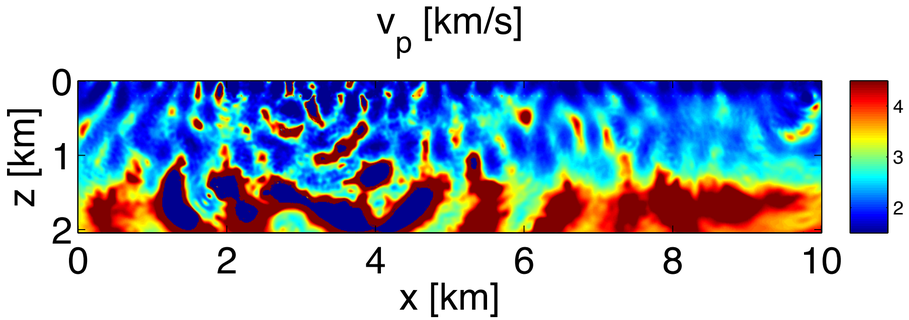
Students T (f,x)



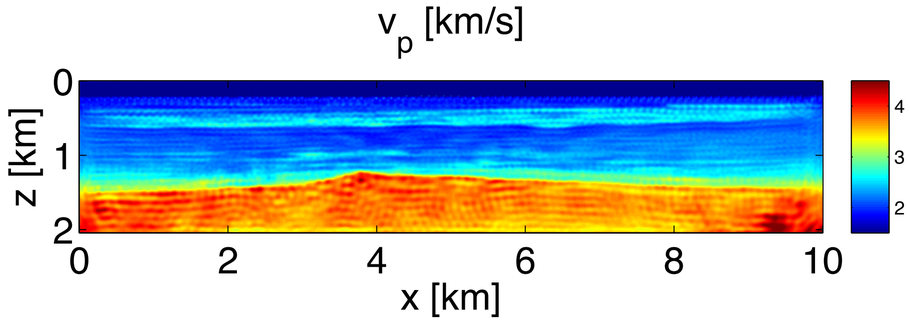
Results 2

Data with bad traces

least-squares



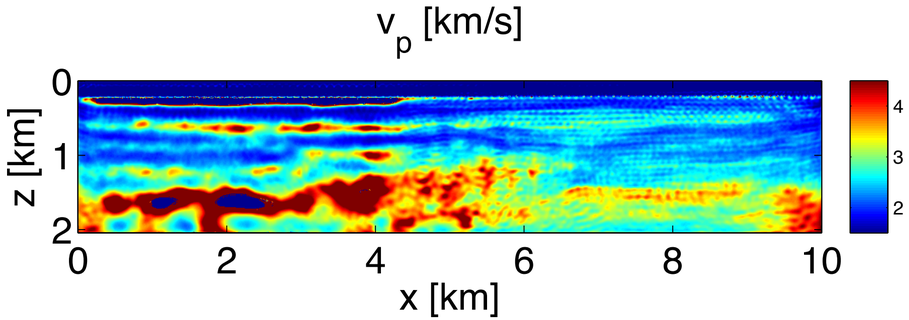
Students T (f,x)



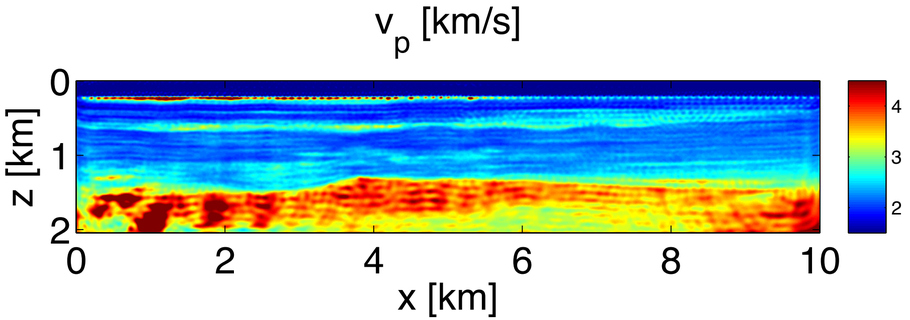
Results 2

Elastic data

least-squares



Students T (f,k)



Acknowledgements

Thank you for your attention !

https://www.slim.eos.ubc.ca/

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, CGG, Chevron, ConocoPhillips, ION, Petrobras, PGS, Statoil, Total SA, WesternGeco, and Woodside.

Questions

Questions

1. Comparing with the CRMN, how much speed up the ML-CRMN can achieve ?
2. Why when the bolck-sizes is bigger, better convergence rate that block CG can achieve?
3. Except uniform sampling and Gaussian sampling, what kind of other sampling strategies can we use ?
4. How much speed up can we obtain using the frugal FWI compared to the conventional FWI?
5. What does the smart averaging means?
6. Does students t only works for the data that has bad traces?