Supercool(ed) least-squares imaging: latest insights in sparsity-promoting migration

Felix J. Herrmann

thanks to Xiang Li

SLIM
Seismic Laboratory for Imaging and Modeling
the University of British Columbia

Big data

http://www.newschool.edu/uploadedImages/events/lang/Data%20Deluge%20compressed(2).jpg

"We are drowning in data but starving for understanding" USGS director Marcia McNutt

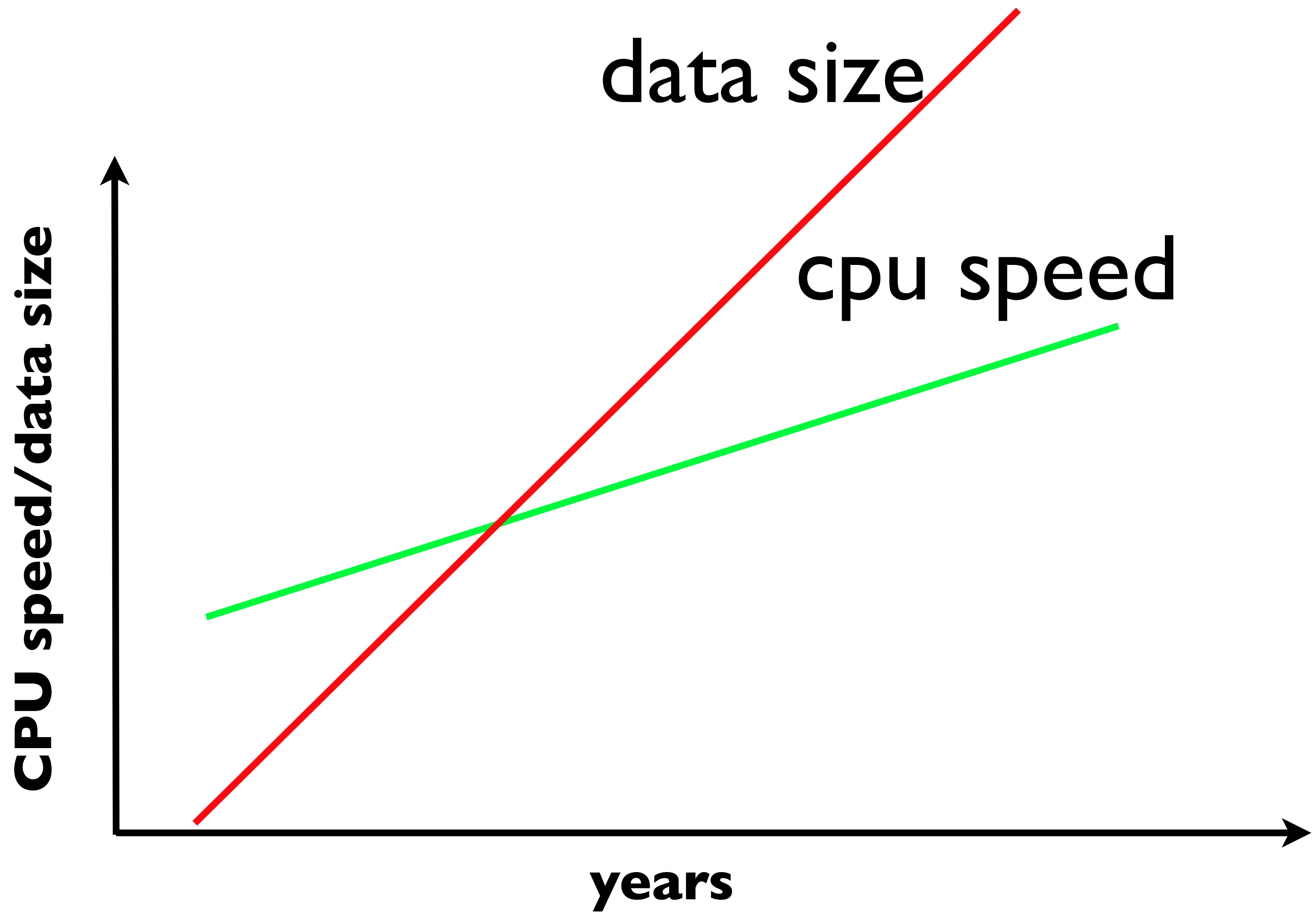
"Got data now what" Carlsson & Ghrist SIAM

http://bigdatablog.emc.com/wp-content/uploads/2012/03/gotbigdata.png

SLIM

Problem

"Data explosion is bigger than Moore's law"



Goals

Replace a '*sluggish*' processing *paradigm* that

- relies on *touching* **all** *data*

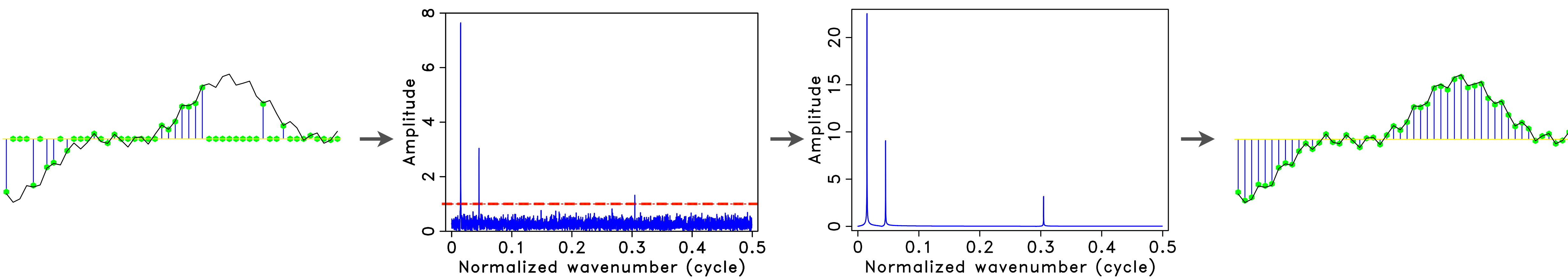
by an *agile* optimization *paradigm* that works on

- **small** *randomized* subsets of data *iteratively*

Confront "*data explosion*" by

- *reducing* acquisition *costs*
- *removing* IO & PDEs-solve *bottlenecks*

Compressive sensing



randomized sampling

detection +
data-consistent
amplitude recovery

recovered signal

$$\min_{\mathbf{x}} \quad \overbrace{\|\mathbf{x}\|_1}^{\text{detection}} \quad \text{subject to} \quad \overbrace{\mathbf{b} = \mathbf{A}\mathbf{x}}^{\text{data-consistent amplitude recovery}}$$

restriction operator

$$\underset{\text{sensing matrix}}{\mathbf{A}} := \overset{\text{restriction operator}}{\mathbf{R}}\underset{\text{inverse Fourier transform}}{\mathbf{F}^H} \qquad \mathbf{A} \in \mathbb{C}^{n \times N} \text{ with } n \ll N$$

sensing matrix

inverse Fourier transform

[Daubechies et. al, ’04; Hennenfent et. al.,’08, Mallat, ’09, Donoho et. al, ’09]

[Montanari, ’12]

SLIM

Convex optimization

Sparse recovery involves *iterations* of the type

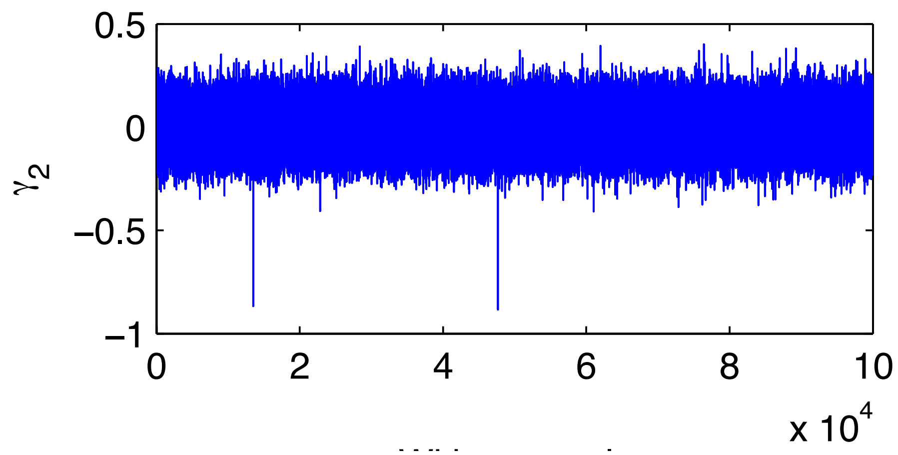
$$\mathbf{x}^{t+1} = \underset{\text{soft threshold}}{\eta_t}\left(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t\right)$$
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t$$

Corresponds to *vanilla* denoising if **A** is a Gaussian matrix.

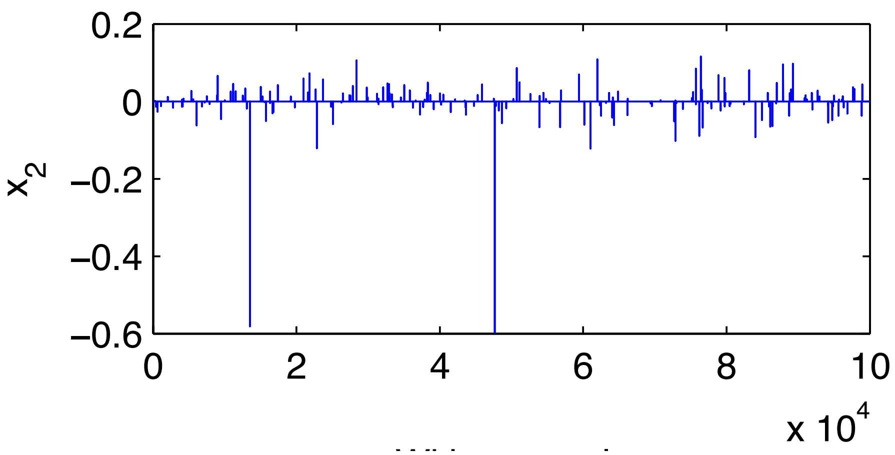
But does the *same* hold for later (t>1) *iterations*...?

Iteration t=1

$$\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t$$

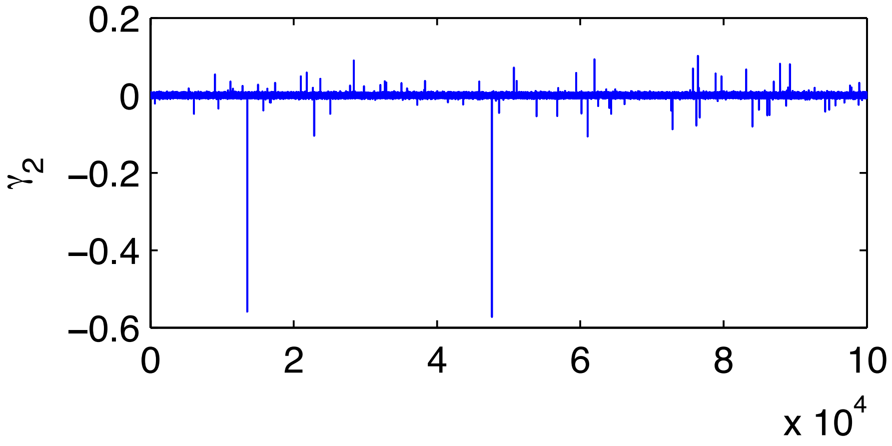


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

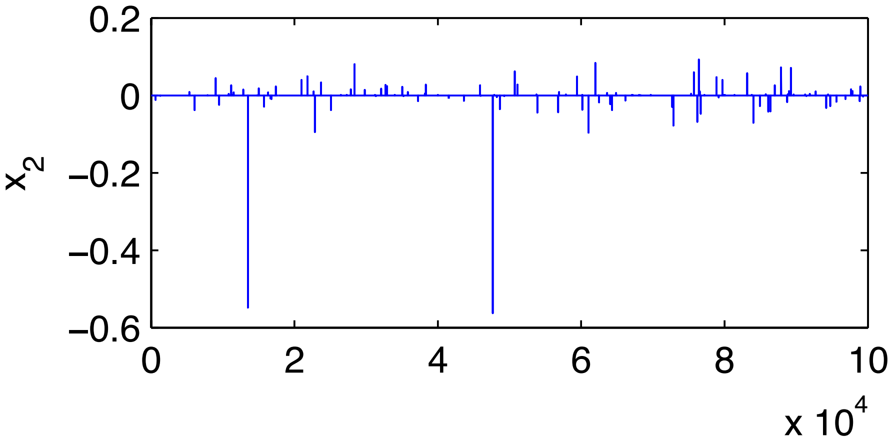


Iteration t=2

$$\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t$$

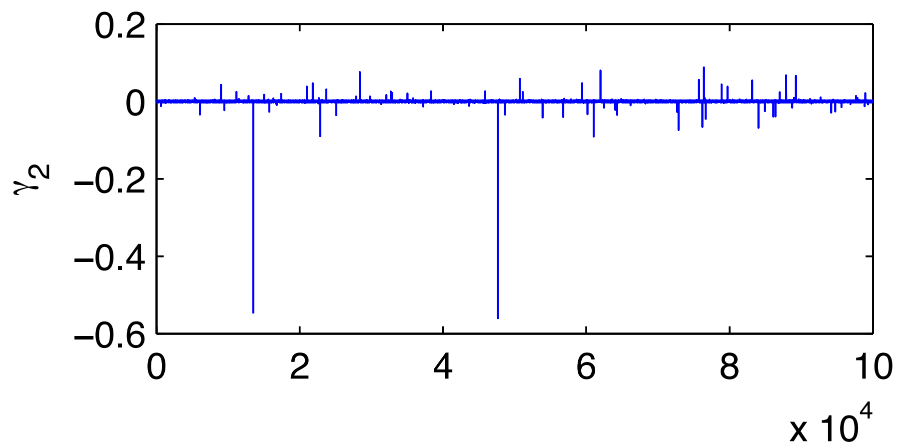


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

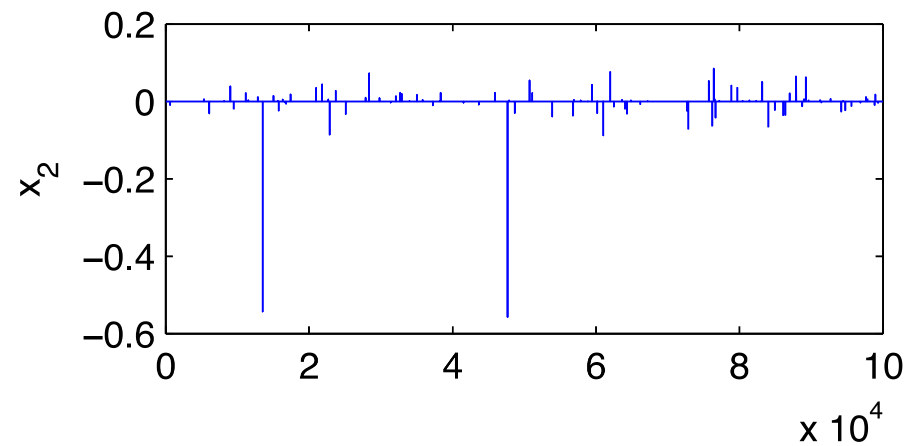


Iteration t=3

$$\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t$$

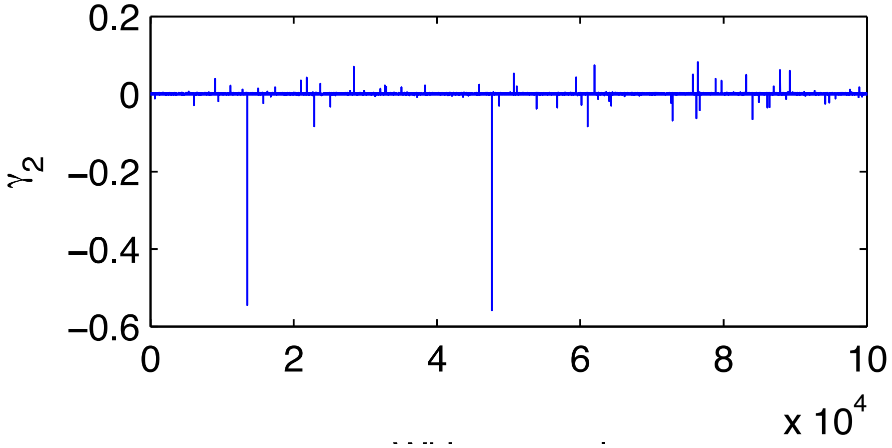


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

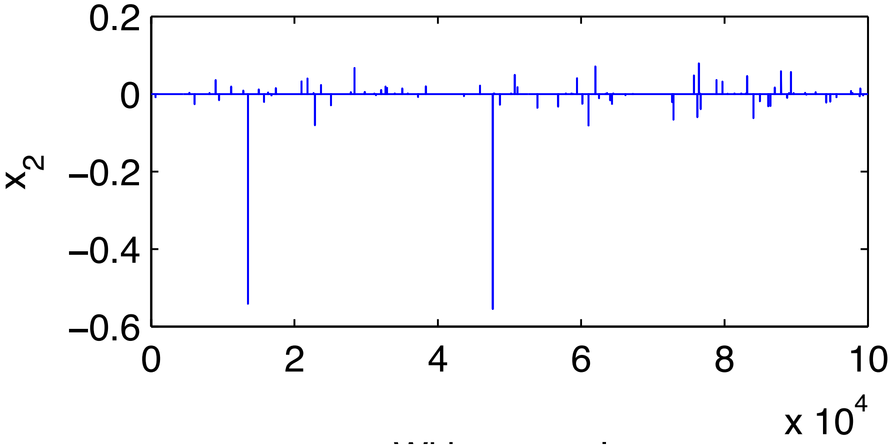


Iteration t=4

$$\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t$$



$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



Problem

After *first* iteration the *interferences* become ‘spiky’ because of *correlations* between model *iterate* $\mathbf{x}^t$ & the *matrix* $\mathbf{A}$

- *assumption* spiky vs Gaussian noise *no* longer *holds*
- renders soft *thresholding* less *effective*

Leads to *slow* convergence of recovery *algorithms*...

Approximate message passing

Add a *term* to *iterative* soft *thresholding*, i.e.,

$$\boxed{\begin{aligned} \mathbf{x}^{t+1} &= \eta_t\left(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t\right) \\ \mathbf{r}^t &= \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \end{aligned}}$$

Holds for

- *normalized* Gaussian matrices $\mathbf{A}_{ij} \in n^{-1/2} N(0,1)$
- large-scale limit and for specific thresholding *strategy*

[Montanari, '12]

SLIM

Approximate message passing

Statistically equivalent to

$$\boxed{\begin{aligned}\mathbf{x}^{t+1} &= \eta_t\left(\mathbf{A}_t^*\mathbf{r}^t + \mathbf{x}^t\right)\\ \mathbf{r}^t &= \mathbf{b}_t - \mathbf{A}_t\mathbf{x}^t\end{aligned}}$$

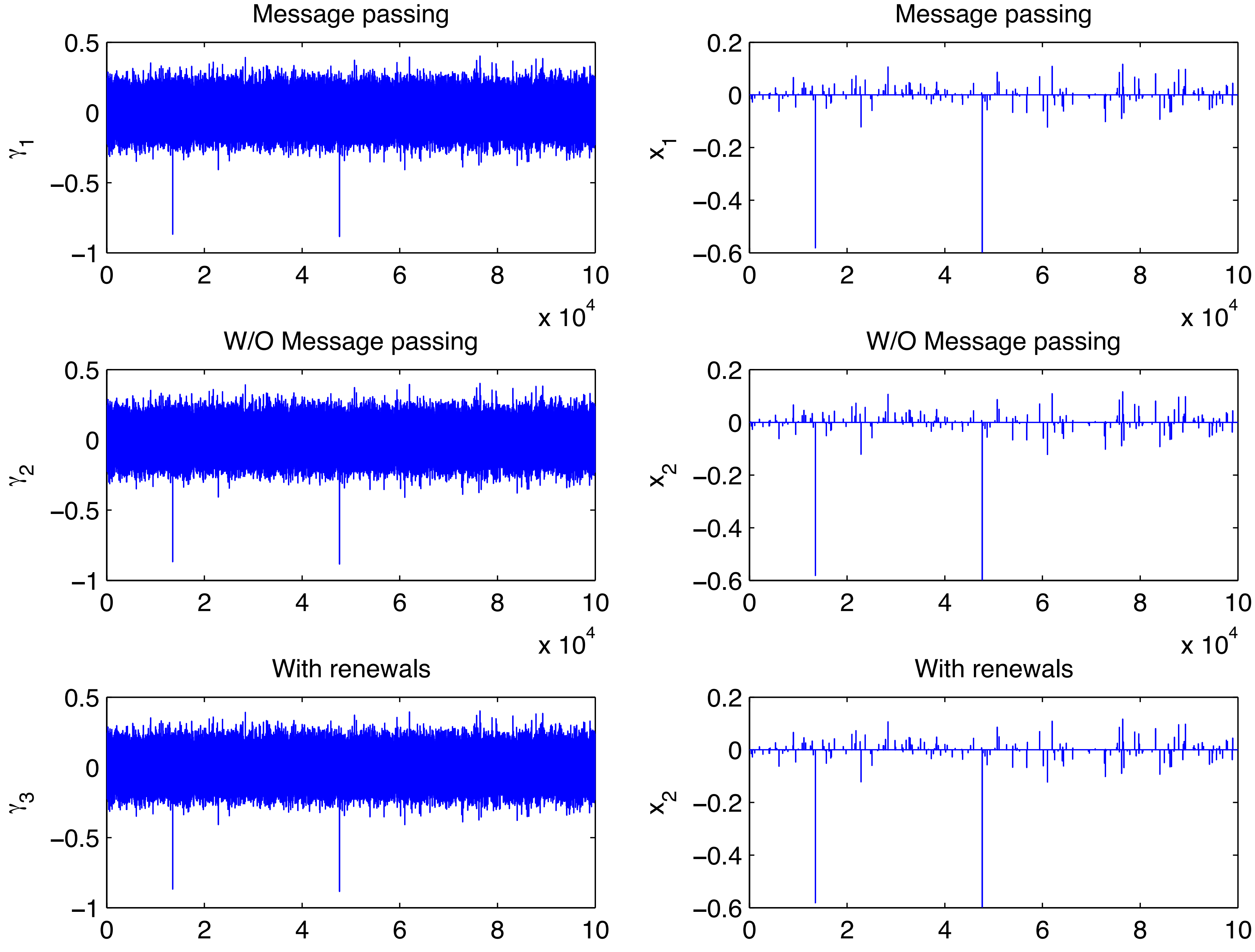
by drawing *new independent* pairs $\{\mathbf{b}_t, \mathbf{A}_t\}$ for *each* iteration

Changes the story *completely*

- breaks *correlation* buildup
- *faster* convergence

Iteration t=1

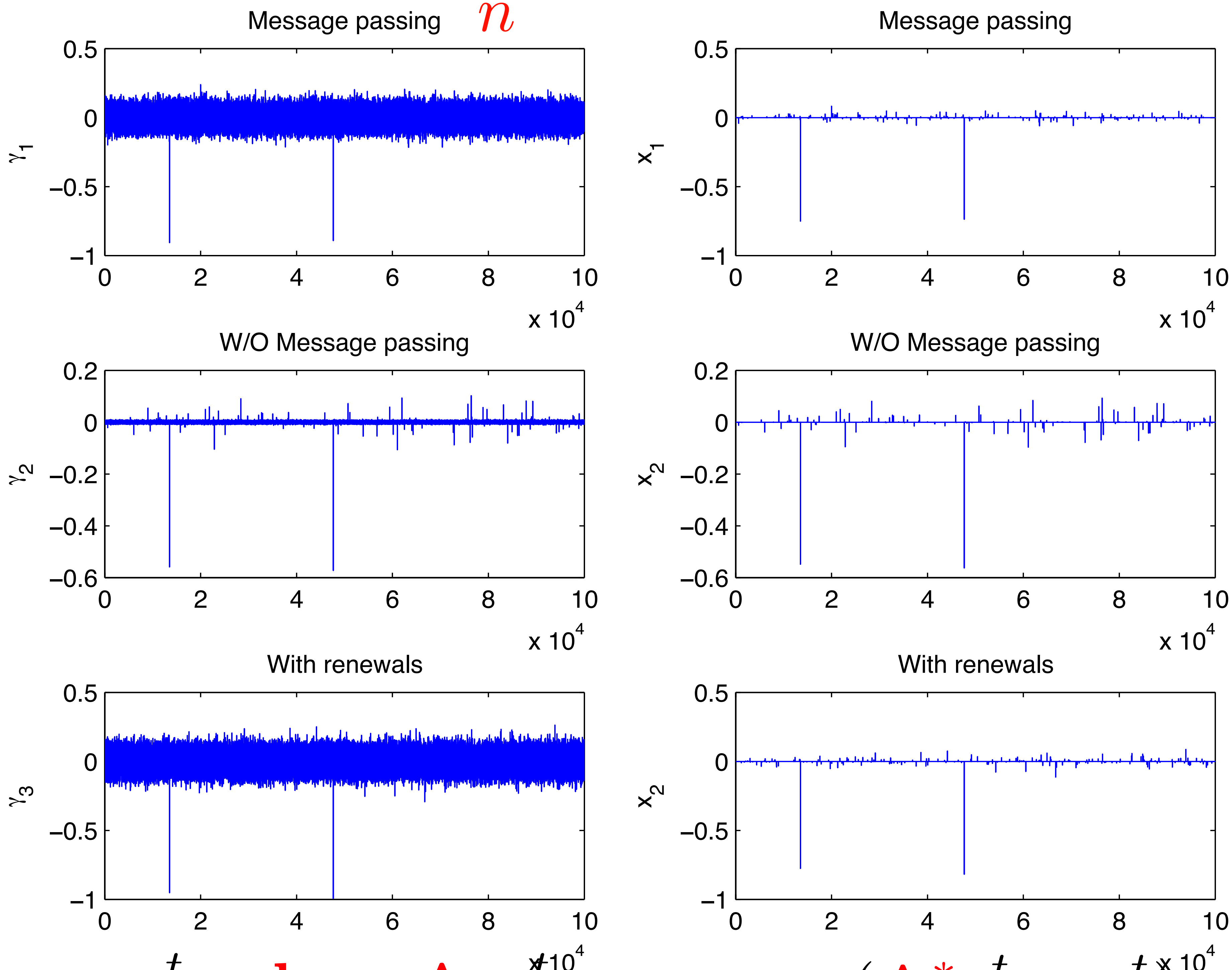
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



$$\mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t\mathbf{x}^t \qquad \eta_t(\mathbf{A}_t^*\mathbf{r}^t + \mathbf{x}^t)$$

Iteration t=2

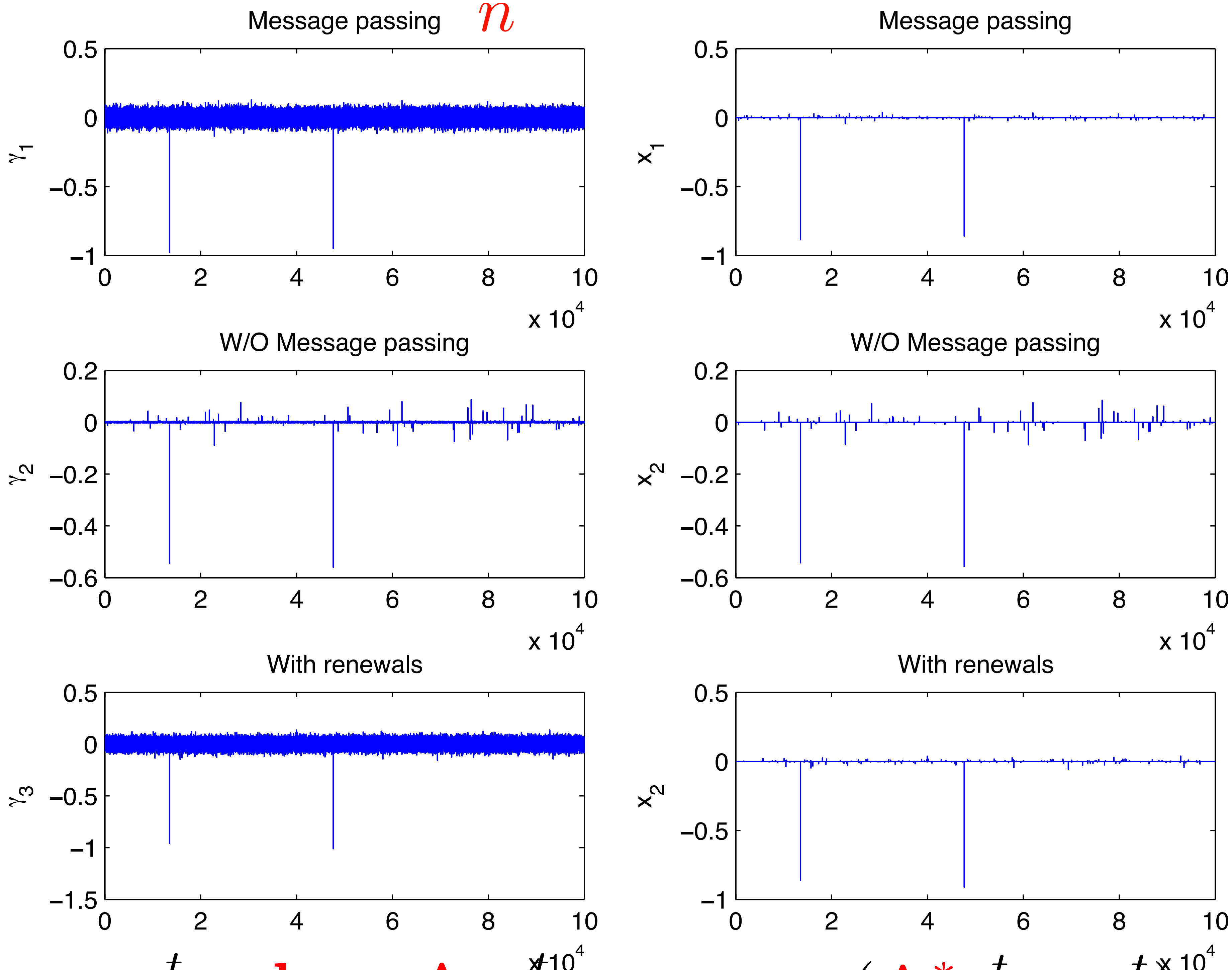
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



$$\mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t\mathbf{x}^t \qquad \eta_t(\mathbf{A}_t^*\mathbf{r}^t + \mathbf{x}^t)$$

Iteration t=3

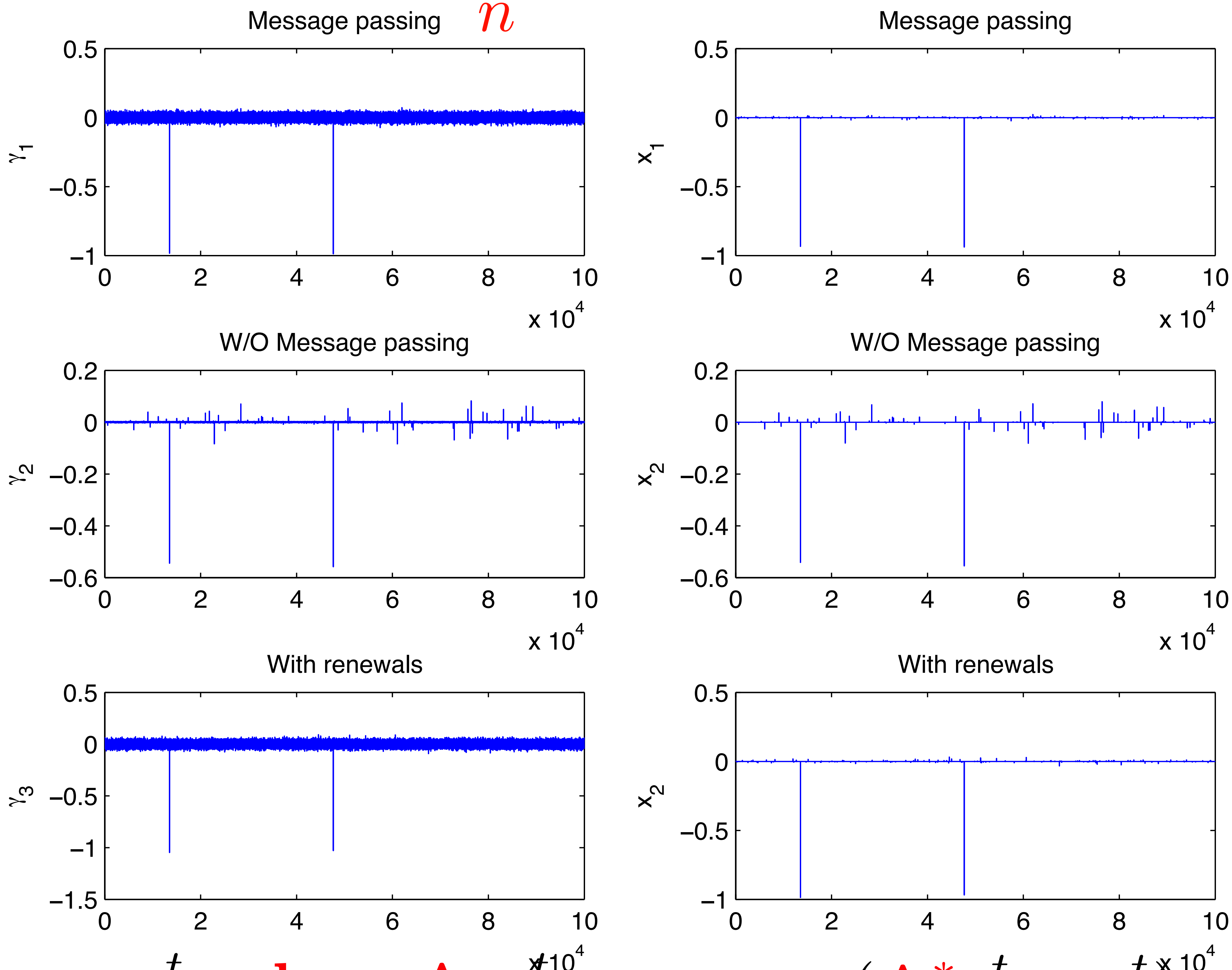
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



$$\mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t\mathbf{x}^t \qquad \eta_t(\mathbf{A}_t^*\mathbf{r}^t + \mathbf{x}^t)$$

Iteration t=4

$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



$$\mathbf{r}^t = \mathbf{b}_t - \mathbf{A}_t\mathbf{x}^t \qquad \eta_t(\mathbf{A}_t^*\mathbf{r}^t + \mathbf{x}^t)$$

Decoupling principle

In large-scale *limit* ($N \to \infty$), the system *decouples* for *each iteration*—i.e.,

$$\left(\mathbf{x}^t + \mathbf{A}^H \mathbf{r}^t\right)_i = (\mathbf{x} + \tilde{\mathbf{w}})_i \text{ for } i = 1 \cdots N$$

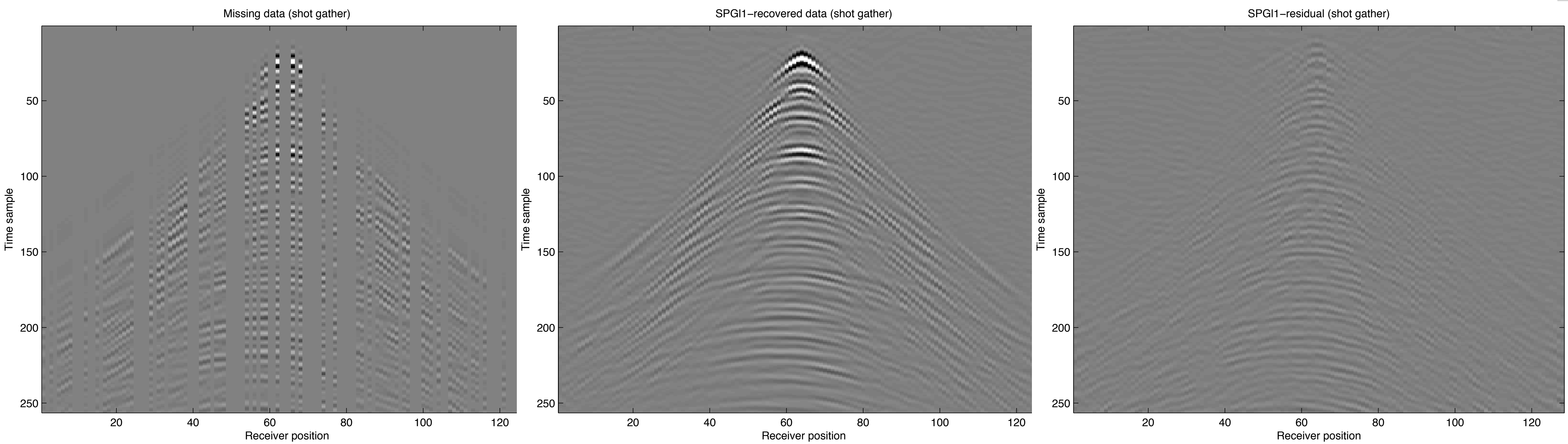
with $\{\tilde{w}_i\}_{i=1 \cdots N}$ *asymptotically* Gaussian

- *each* entry can be treated *separately*
- *estimate* each *entry* by *elementwise soft thresholding* with *carefully* selected threshold *levels*

Missing-trace interpolation [SPGl1]

Recovery with 3D curvelets (N=1.12 X 10^9)

$7.75\,\mathrm{dB}$



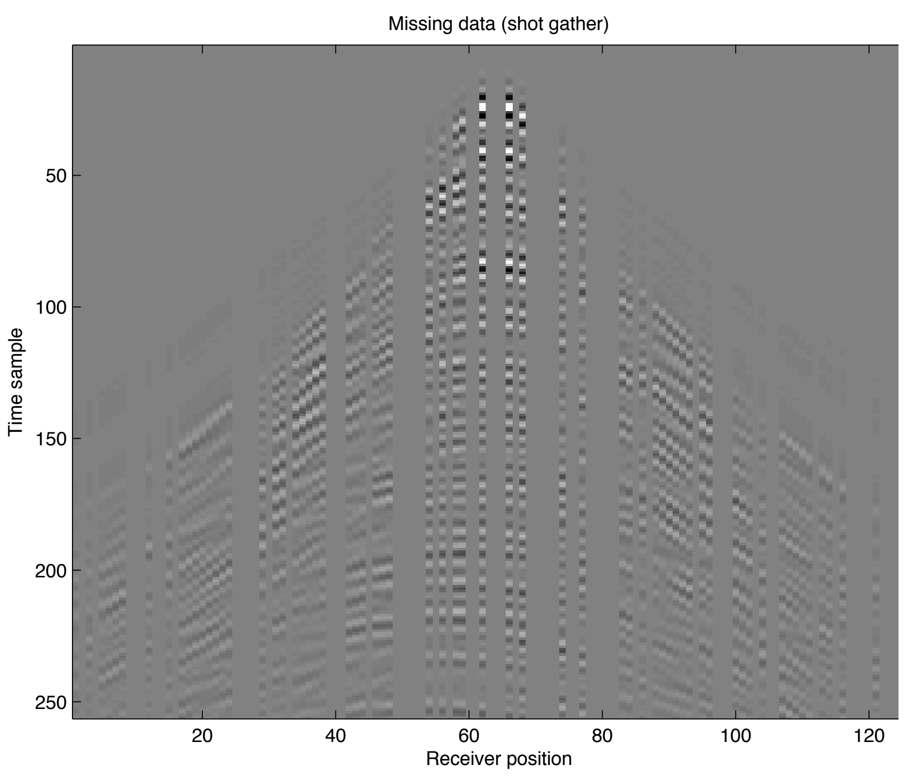
50 % *missing* data

recovery
50 iterations

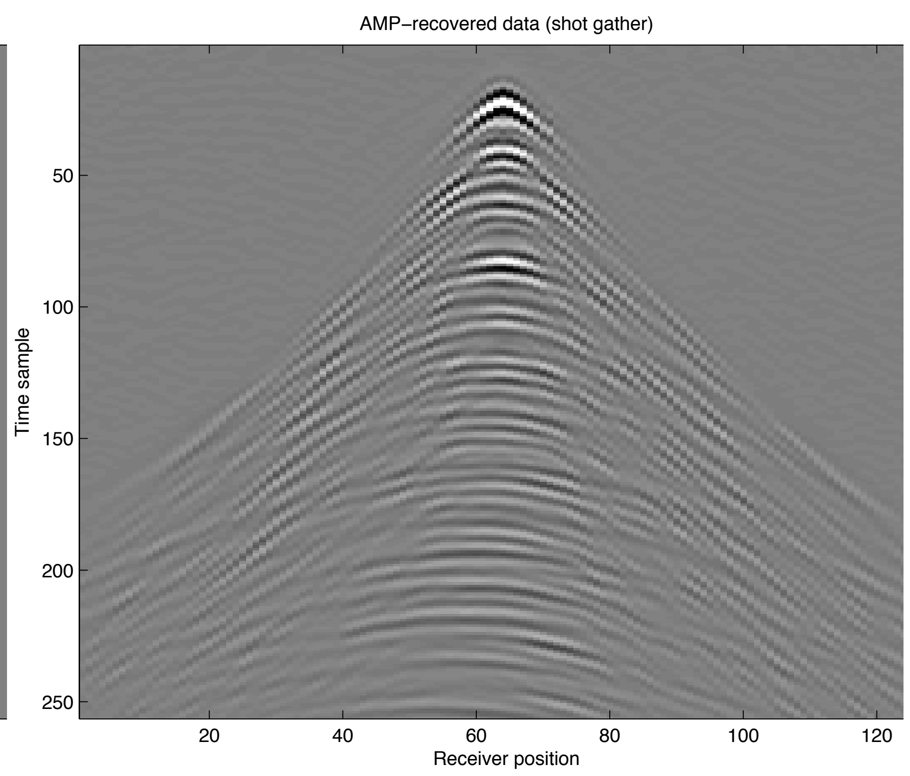
difference

Missing trace interpolation [AMP]

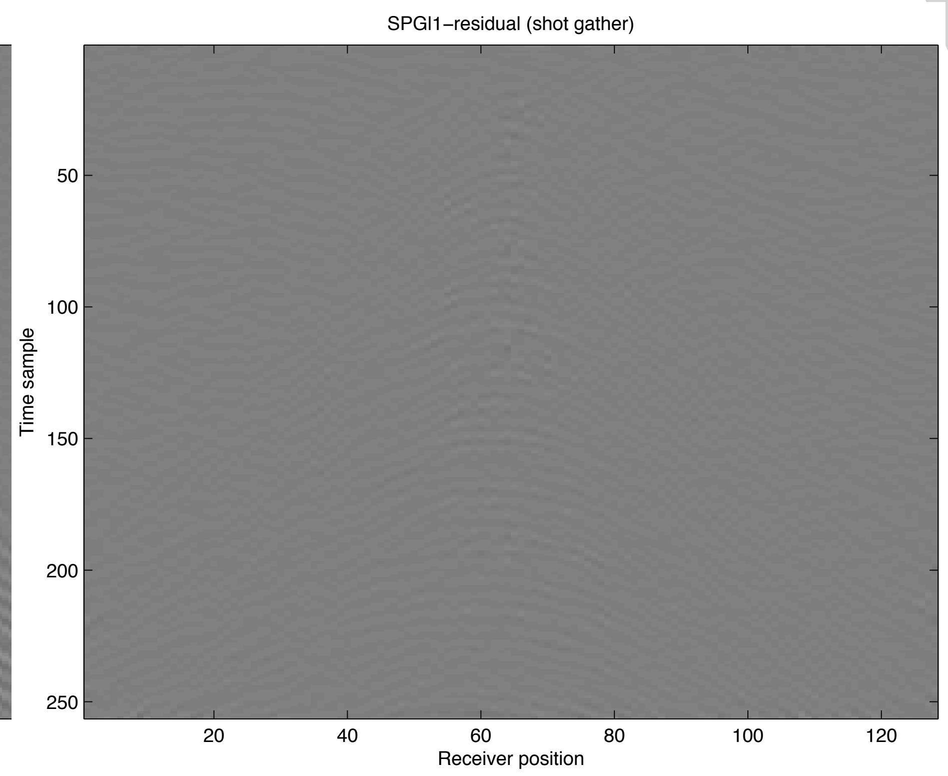
Recovery with 3D curvelets (N=1.12 X 10^9) $9.75\,\mathrm{dB}$



50 % *missing* data



recovery
50 iterations



difference

Observations

Message-pass term has the same *effect* as drawing *independent* experiments $\{\mathbf{b}_t, \mathbf{A}_t\}$

- '*Gaussian*' matrices
- *delicate* normalization and *thresholding* strategy
- *renders* proposed method *impractical*
- can lead to *dramatically* improved convergence

How can we still reap *benefits* from *message* passing in *realistic* less-than-ideal *geophysical* settings?

Problems

In large-scale limit one-norm solvers suffer from:

- *first-order* spectral-gradient methods need many *iterations*
- *second-order* quasi-Newton need to store *multiple* model vectors
- *correlation* buildup that slows down *convergence*

Can *insights* from *AMP* be used to *accelerate* current state-of-the art *one-norm* solvers?

Continuation methods

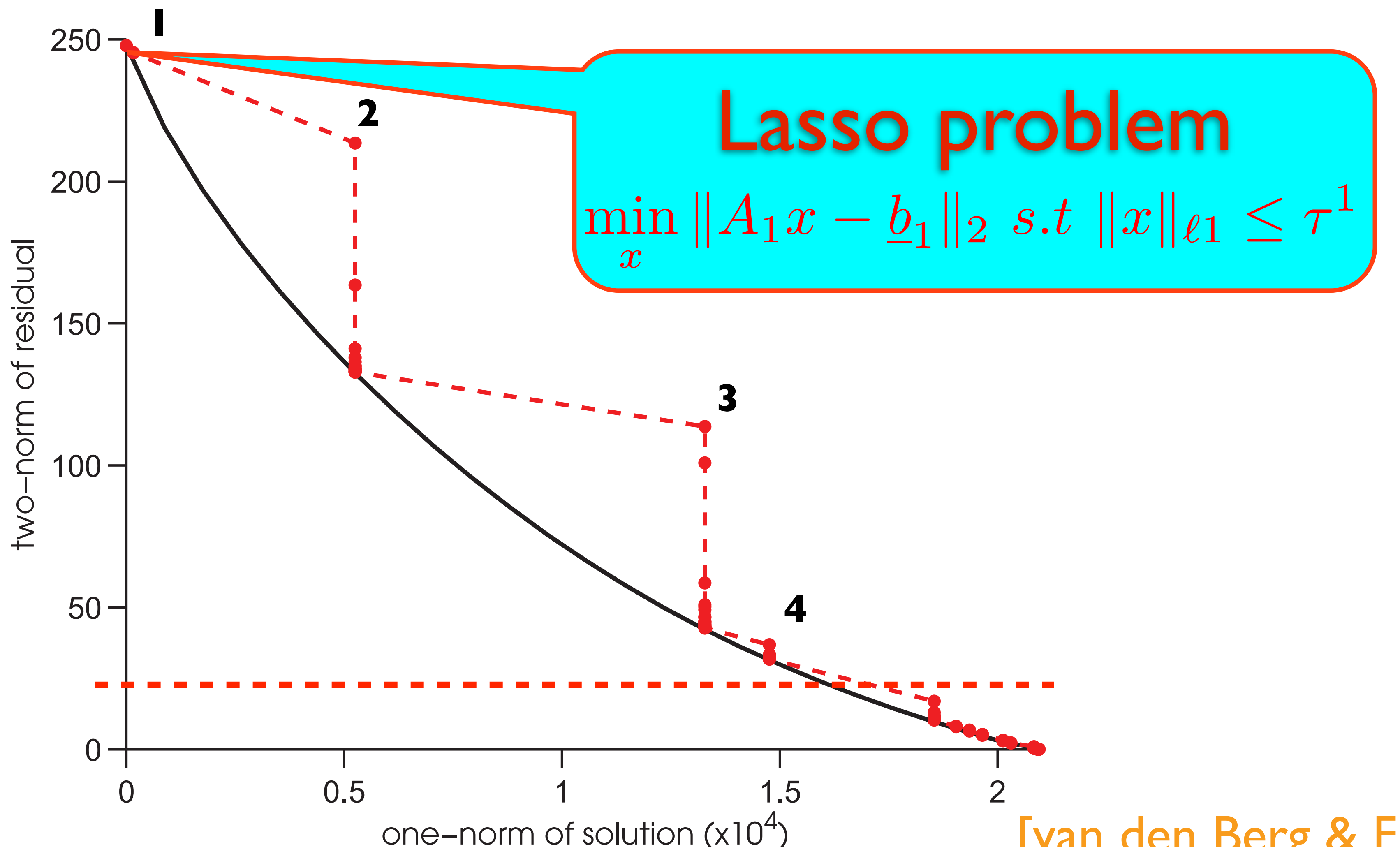
Versatile large-scale *sparsity*-promoting solvers *limit* the number of *matrix-vector* multiplies by *cooling*, which

- slowly allows *components* to *enter* into the *solution*
- solves an *intelligent* series of LASSO *subproblems* for *decreasing* sparsity levels
- uses *convexity* & *smoothness* of Pareto curves with Newton rootfinding

SLIM

Supercooled

spectral-projected gradients



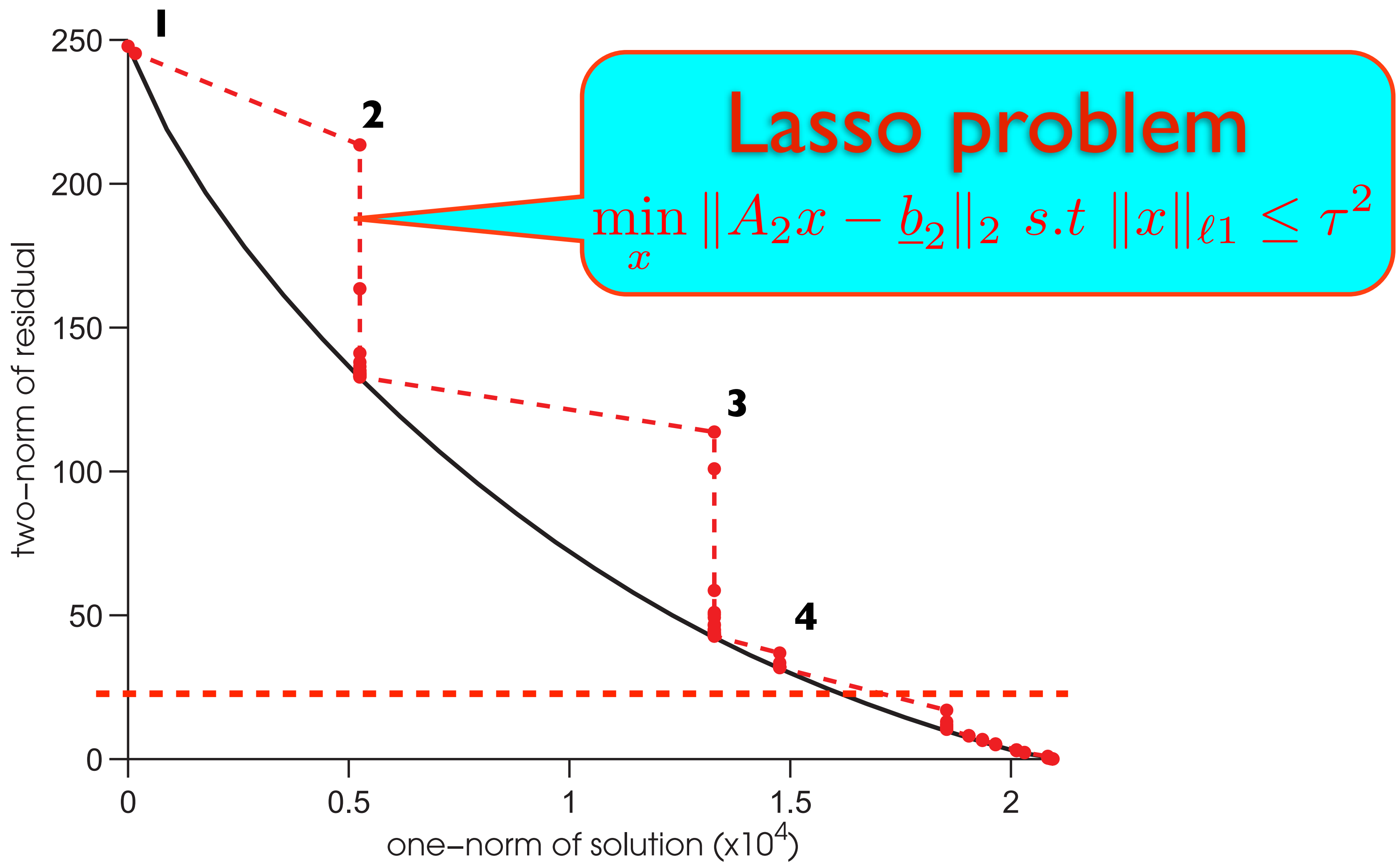
[van den Berg & Friedlander, ’08]

[Hennefent et. al., ’08]

[Lin & FJH, ’09-]

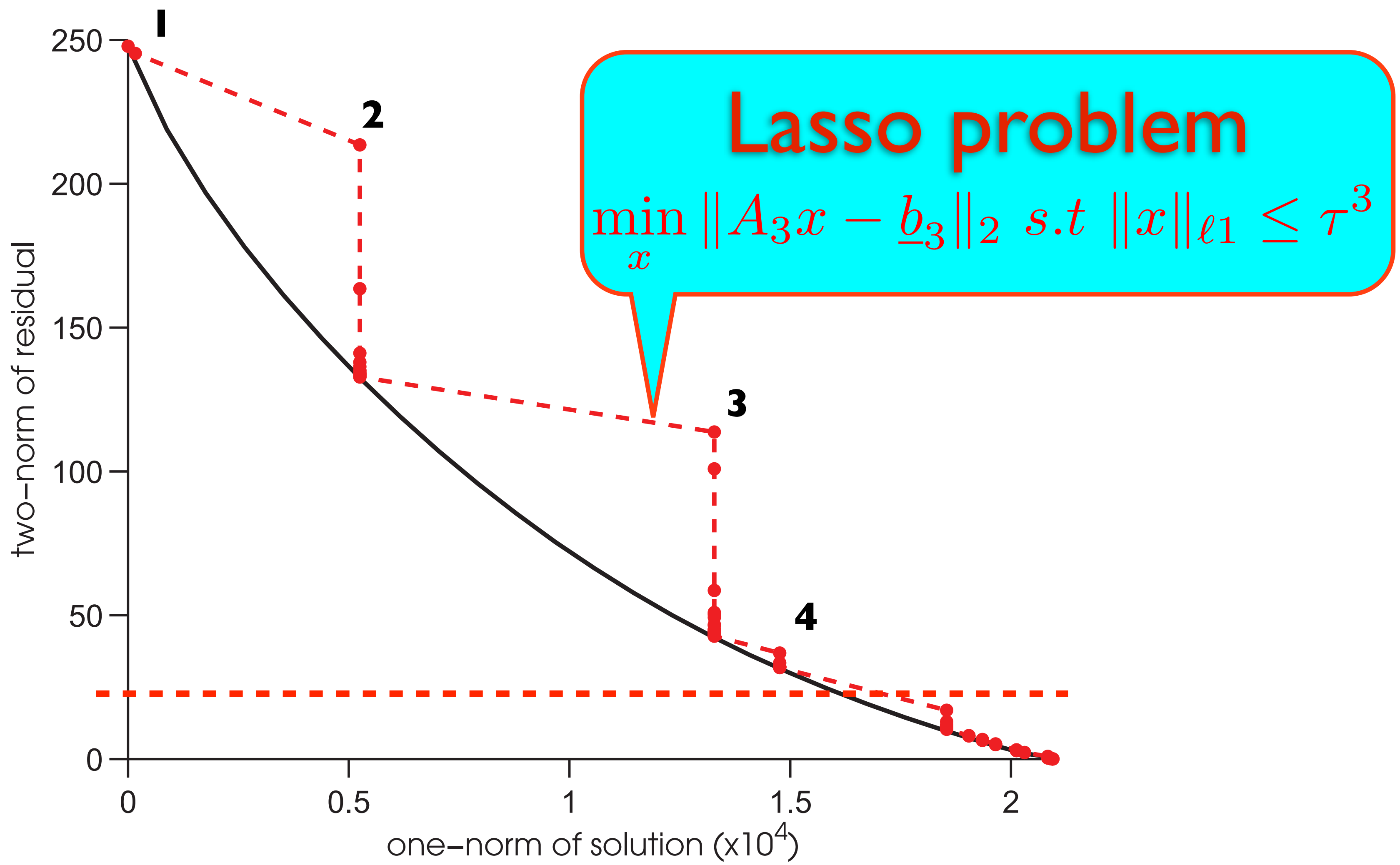
Supercooled

spectral-projected gradients



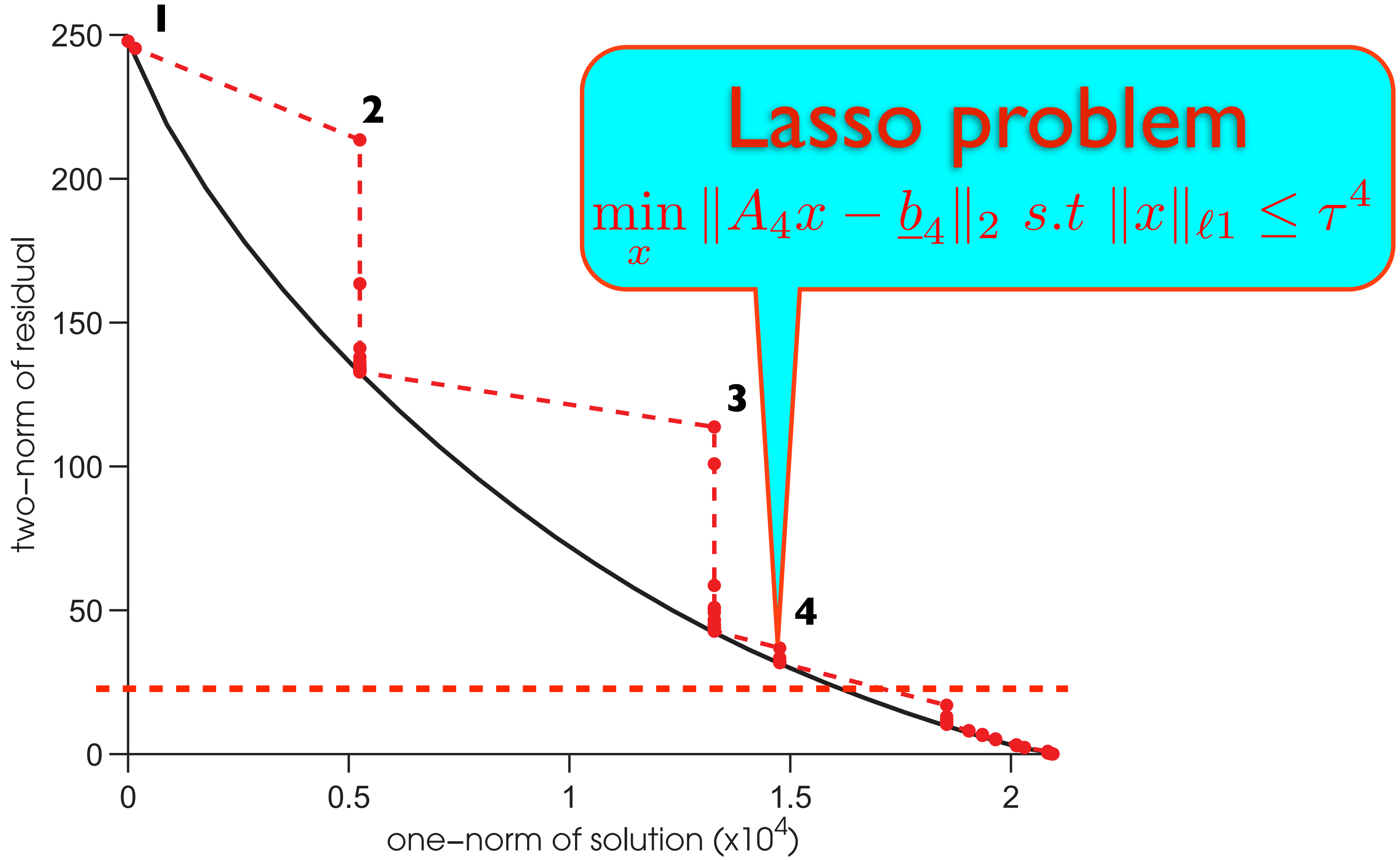
Supercooled

spectral-projected gradients



Supercooled

spectral-projected gradients



Supercooling

Break *correlations* between the model *iterate* and matrix **A** by *rerandomization*

- draw new *independent* $\{\mathbf{b}_t, \mathbf{A}_t\}$ after each LASSO subproblem is solved
- brings in "*extra*" information *without* growing the *system*
- ***minimal*** *extra* computational & memory cost

Supercooled

spectral-projected gradients

Result: Estimate for the model $\mathbf{x}^{t+1}$

$\mathbf{x}^0, \widetilde{\mathbf{x}} \longleftarrow \mathbf{0}$ and $t, \tau^0 \longleftarrow 0$; `// Initialize`

while $t \leq T$ **do**

  $\mathbf{A} \longleftarrow \mathbf{A}_{ij} \sim N(0, 1/\sqrt{n})$; `// Draw new sensing matrix`

  $\mathbf{b} \longleftarrow \mathbf{A}\mathbf{x}$; `// Collect new data`

  $\mathbf{x}^{t+1} \longleftarrow \mathrm{spgl1}(\mathbf{A}, \mathbf{b}, \tau^t, \sigma = 0, \mathbf{x}^t)$; `// Reach Pareto`

  $\tau^t \longleftarrow \|\mathbf{x}^{t+1}\|_1$; `// New initial` τ `value`

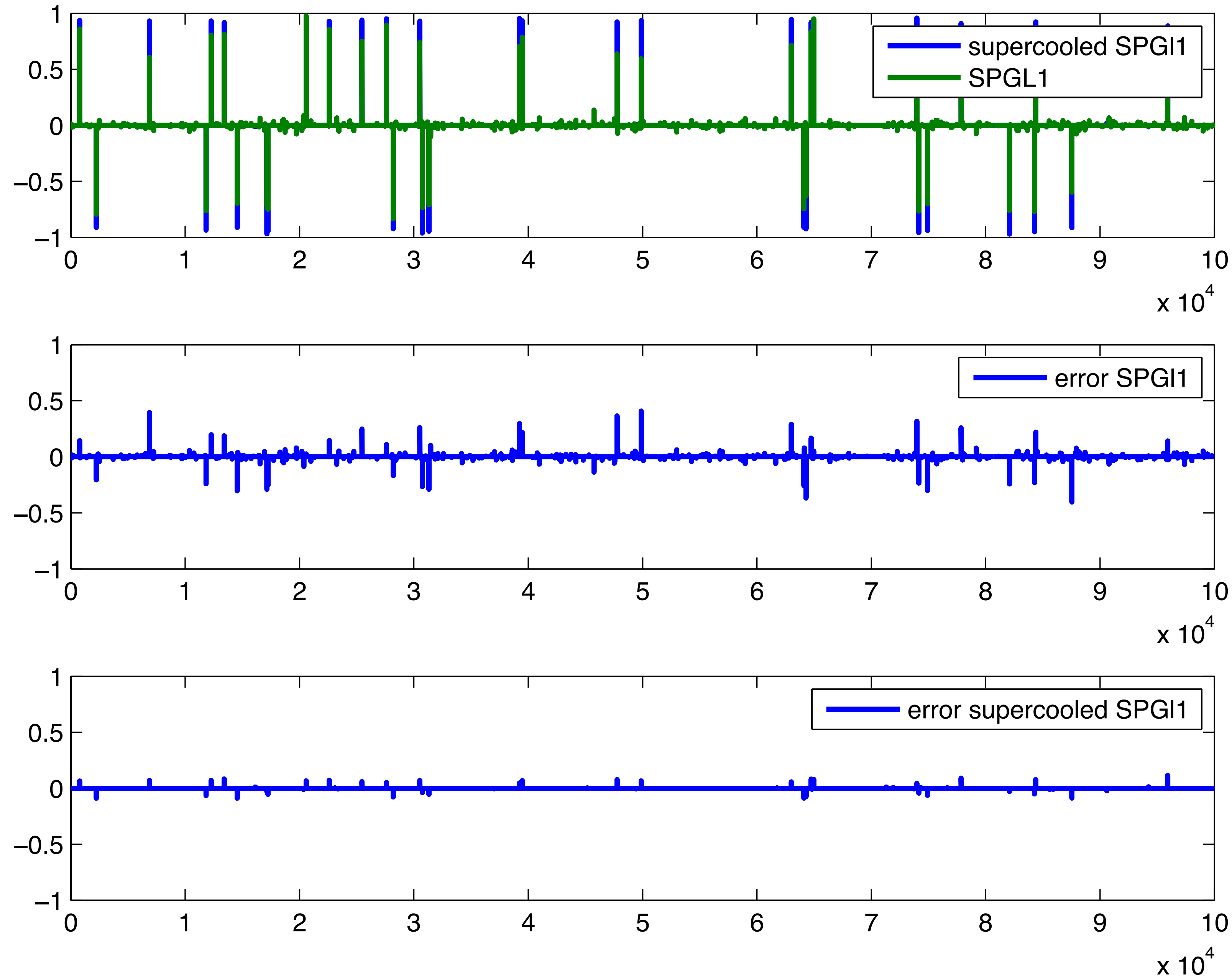
  $t \longleftarrow t + \Delta T$; ; `// Add # of iterations of spgl1`

end

Algorithm 1: Supercooled SPGℓ_1 with message passing.

Sparse example

[n=500; N=10000; k=35; T=50]



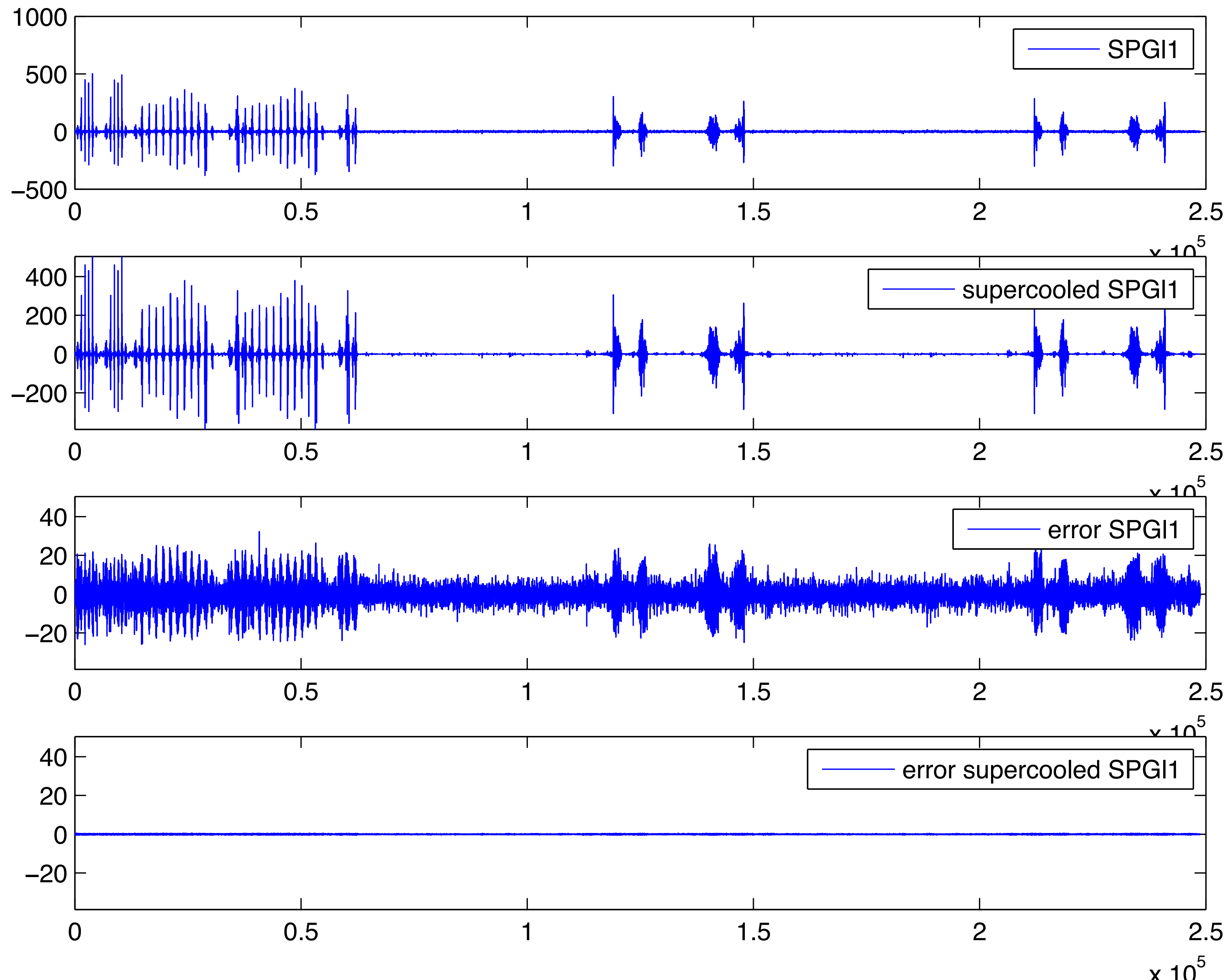
SLIM

Ideal 'Seismic' example

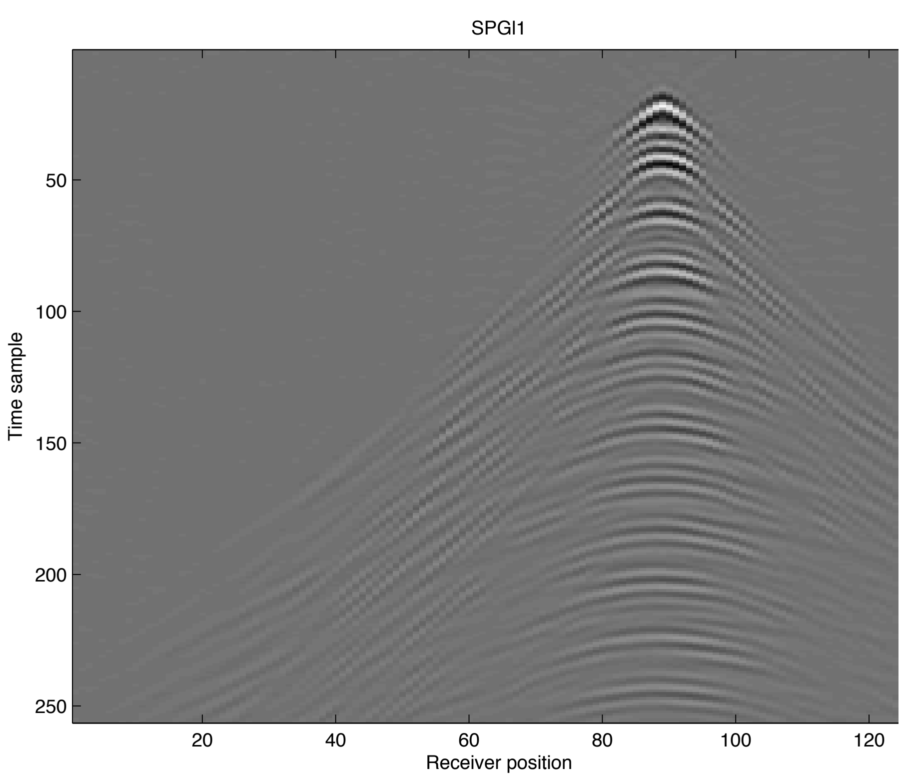
[n/N=0.13;N=248759;T=500]

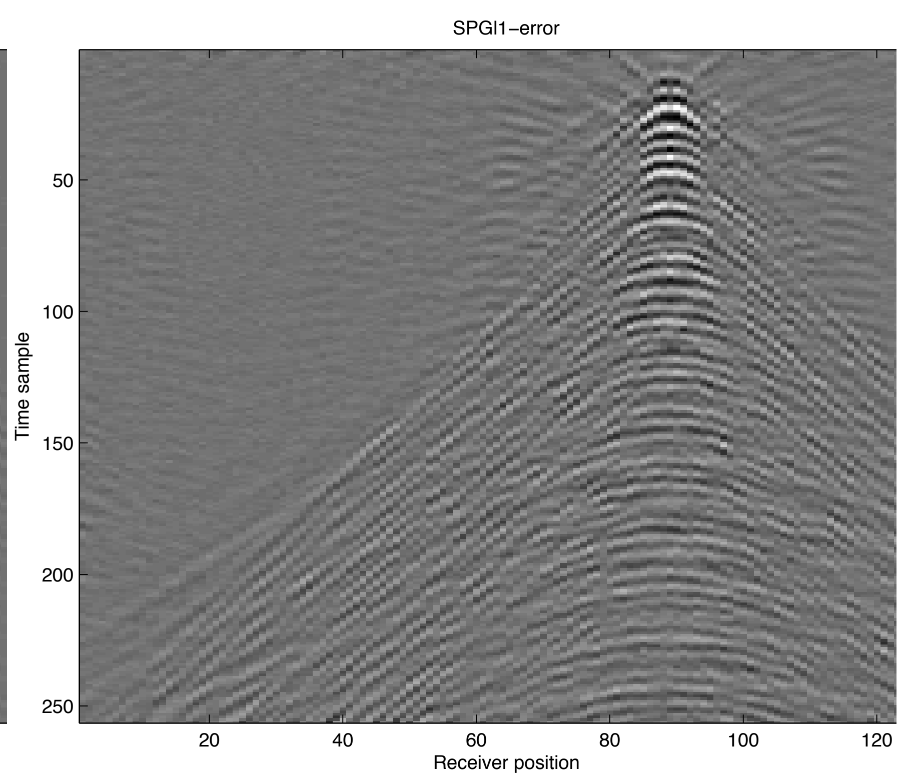
10 X

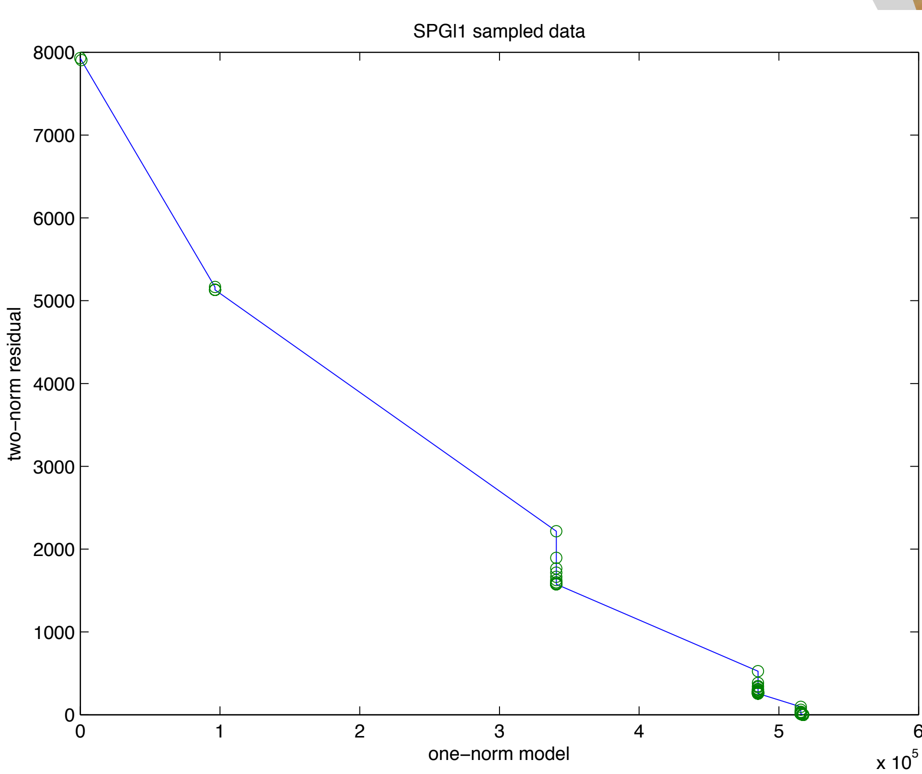
10 X



Ideal 'Seismic' example [n/N=0.13;N=248759;T=500]







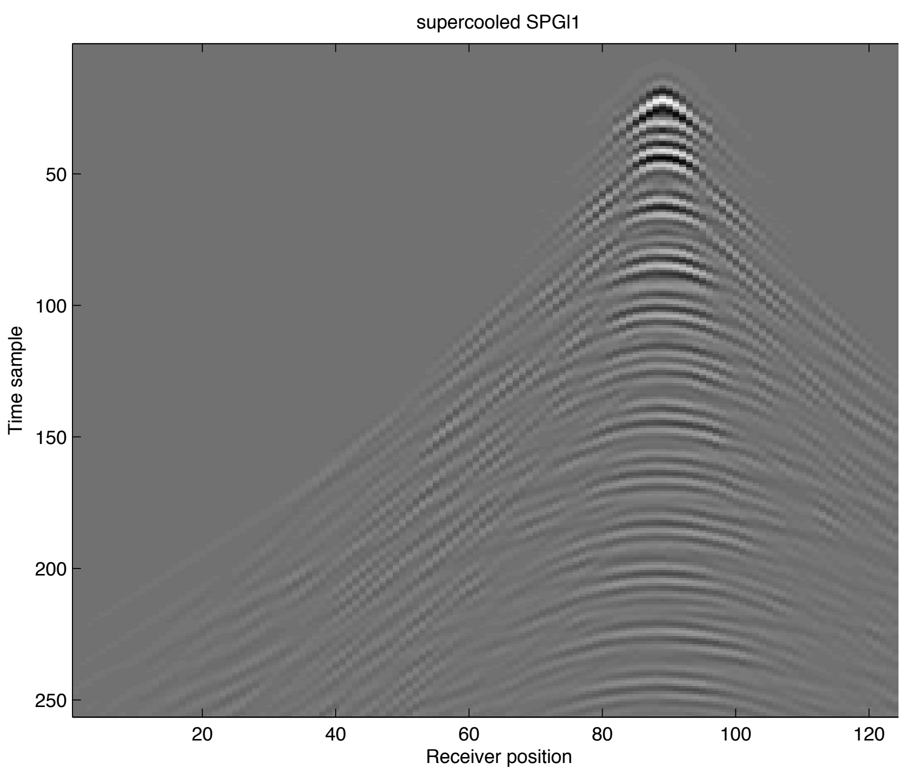
recovery

error

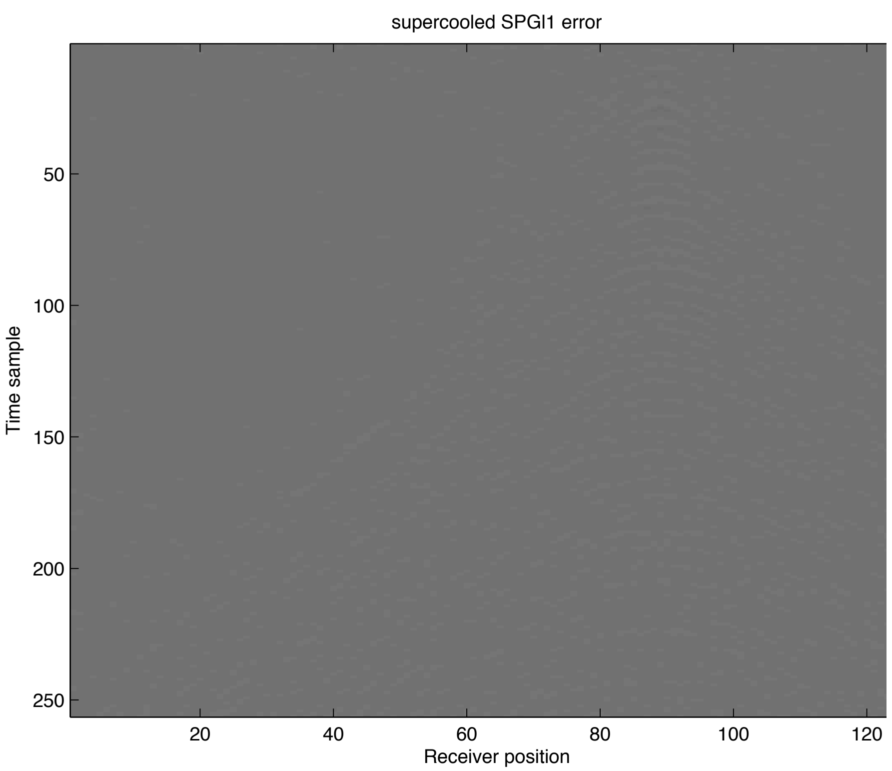
solution *path*

Cooled

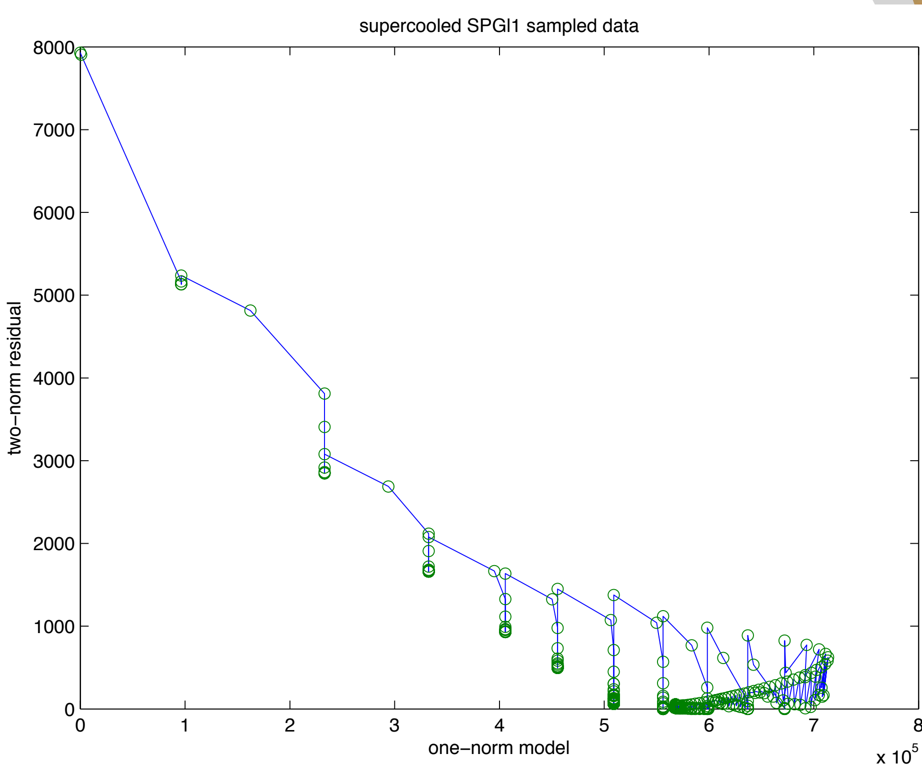
Ideal 'Seismic' example [n/N=0.13;N=248759;T=500]



recovery



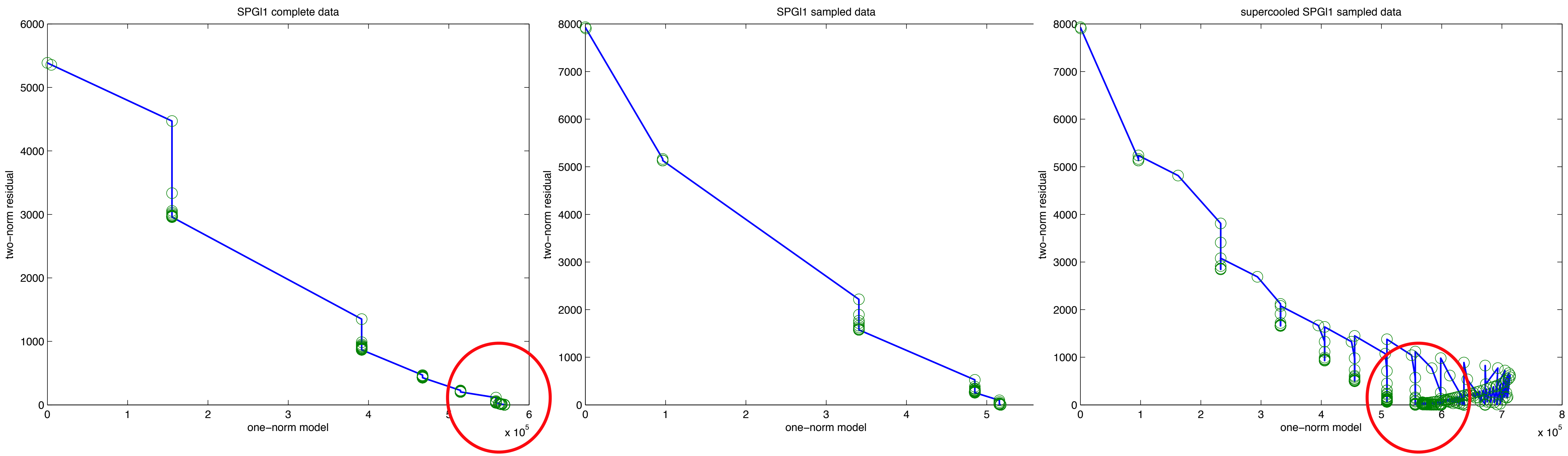
error



solution *path*

Supercooled

Solution paths



Independent redraws of $\{\mathbf{b}_t, \mathbf{A}_t\}$ lead to improved *recovery*

[Romero et. al., 2000;]

[Montanari, 2012]

[Herrmann & Li, 2012]

SLIM

Observations

Independent redraws of $\{\mathbf{b}_t, \mathbf{A}_t\}$ get rid of *small* difficult to remove *interferences*

- working *only* with *subsets* of the *data*

But, aren't we *fooling* ourselves since *proposed* method

- *defeats* the *premise* of *compressive* sampling

Or, are there *data-rich* applications for this method?

- e.g. *efficient* imaging with *random* source *encoding*

Random source-encoded imaging

Replace migration with *all* data (overdetermined system)

$$\widetilde{\mathbf{x}}_{\mathrm{mig}} = \mathbf{A}^*\mathbf{b} \quad \text{approximating} \quad \underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2K}\sum_{i=1}^{K}\|\mathbf{b}_i - \mathbf{A}_i\mathbf{x}\|_2^2$$

with *K* large by *sparsity*-promoting *migration* (underdetermined)

$$\boxed{\underset{\mathbf{x}}{\text{minimize}} \ \|\mathbf{x}\|_1 \quad \text{subject to} \quad \underline{\mathbf{b}}_i = \underline{\mathbf{A}}_i\mathbf{x}, \quad i = 1\cdots K'}$$

with $K' \ll K$ and $\{\underline{\mathbf{b}}_i, \underline{\mathbf{A}}_i\}$ *supershots* & *demigration* operators

Compressive imaging
[*with* message passing]

Select *independent* random source *encodings* after *each* LASSO *subproblem* is solved

- calculate corresponding *supershots*
- *redefine* demigration operator (and its *adjoint*)
(select *independent* simultaneous sources & supershots)

Promote *sparsity* in the *curvelet* domain

Compressive imaging
[*with* message passing]

Result: Estimate for the model $\mathbf{x}^{t+1}$

$\mathbf{x}^0, \widetilde{\mathbf{x}} \longleftarrow \mathbf{0}$ and $t, \tau^0 \longleftarrow 0$; `// Initialize`

while $t \leq T$ **do**

  $\mathbf{W} \longleftarrow \mathbf{W} \in \mathbb{R}^{K \times K'}$ with $W_{ij} \sim N(0, 1/\sqrt{K'})$; `// Random encoding`

  $\{\underline{\mathbf{b}}, \underline{\mathbf{q}}\} \longleftarrow \{\mathbf{DW}, \mathbf{QW}\}$; `// Draw sim sources and data`

  $\underline{\mathbf{A}} \longleftarrow \nabla \boldsymbol{\mathcal{F}}[\mathbf{m}_0; \underline{\mathbf{q}}]$; `// New demigration operator`

  $\mathbf{x}^{t+1} \longleftarrow \mathrm{spgl1}(\underline{\mathbf{A}}, \underline{\mathbf{b}}, \tau^t, \sigma = 0, \mathbf{x}^t)$; `// Reach Pareto`

  $\tau^t \longleftarrow \|\mathbf{x}^{t+1}\|_1$; `// New initial` τ `value`

  $t \longleftarrow t + \Delta T$; `// Add # of iterations of spgl1`

end

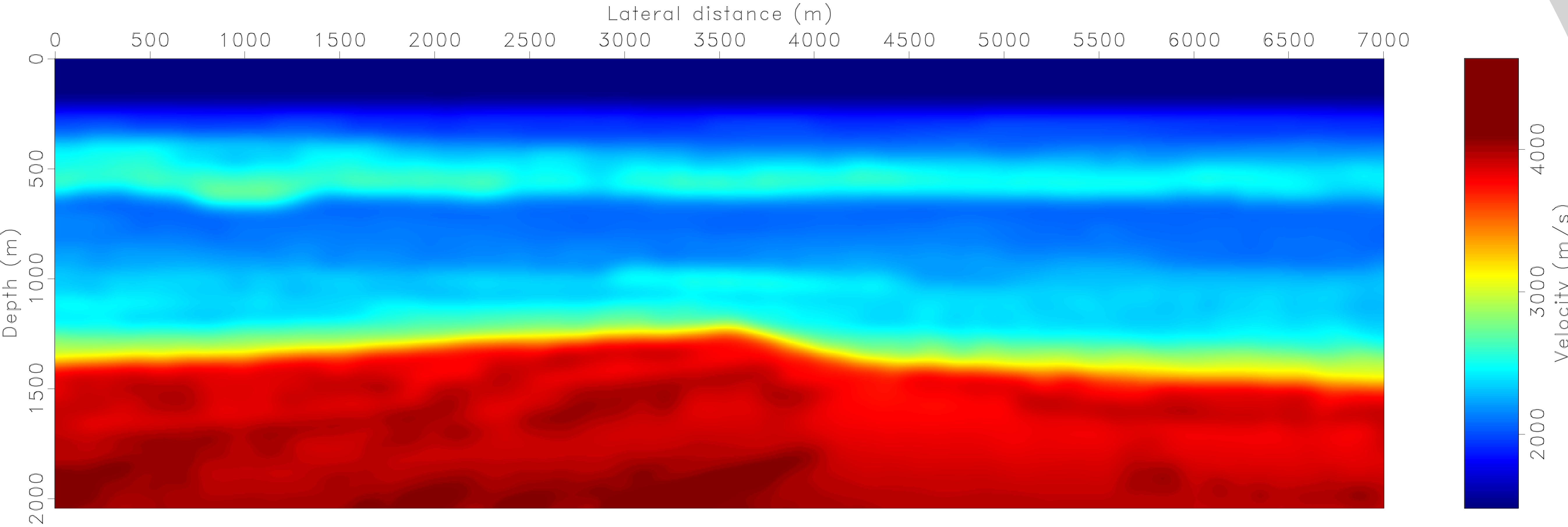
Algorithm 1: Supercooled sparsity-promoting migration.

Imaging results

Time-harmonic Helmholtz:

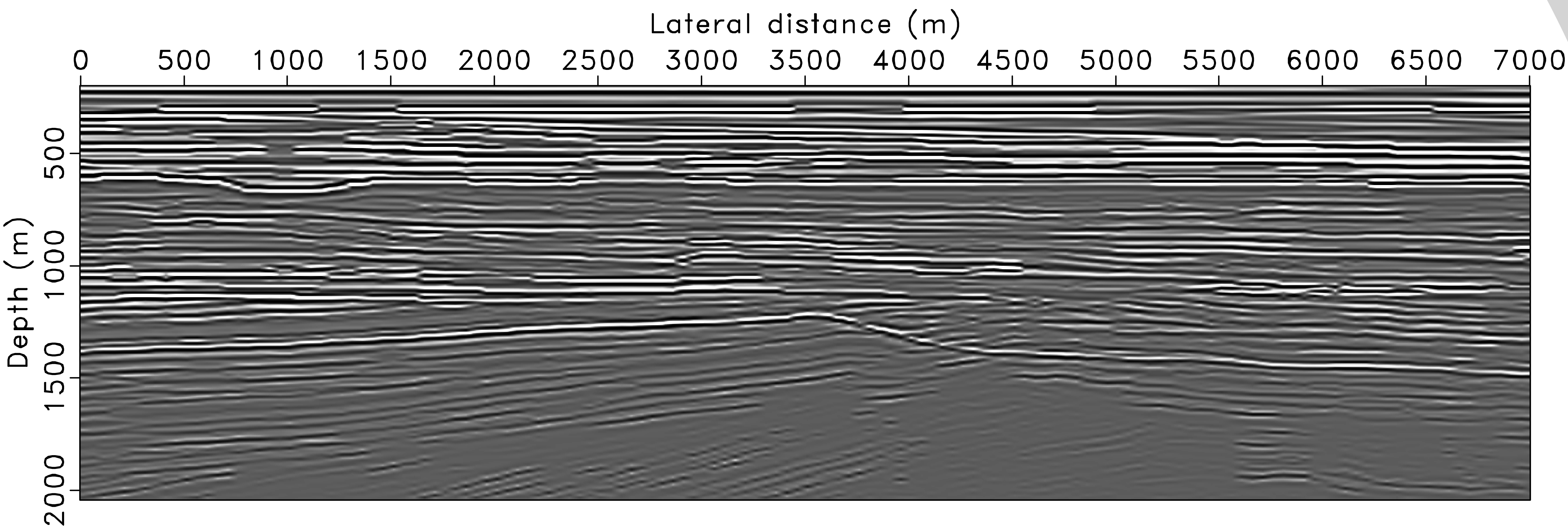
- 409 X 1401 with mesh size of 5m
- 9 point stencil [C. Jo et. al., '96]
- absorbing boundary condition with damping layer with thickness proportional to wavelength
- solve wavefields on the fly with direct solver

Imaging results
[background model]



Migration results

[*true* perturbation]

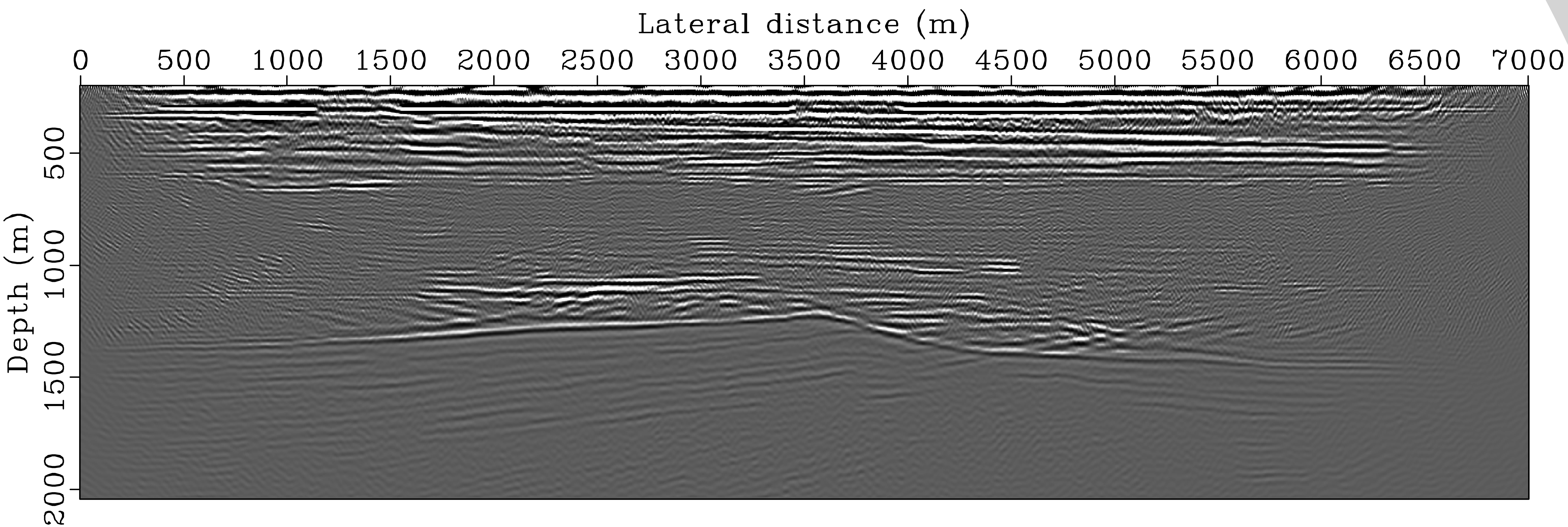


Imaging results

Split-spread surface-free 'land' acquisition:

- 350 sources with sampling interval 20m
- 701 receivers with sampling interval 10m
- maximal offset 7km (3.5 X depth of model)
- Ricker wavelet with central frequency of 30Hz
- recording time for each shot is 3.6s

Migration results
[migration with *all* data]



SLIM

Imaging results

Reduced setup:

- 10 *random* frequencies *(versus 300 frequencies)* (20Hz-50Hz)
- 3 random *simultaneous* shots (*versus 350 sequential shots*)

Significant dimensionality reduction of

$$\boxed{\frac{K'}{K} = 0.0003}$$

[Herrmann & Li, 2011]

Imaging results

Least-squares migration with *randomized supershots:*

$$\delta\widetilde{\mathbf{m}} = \mathbf{S}^* \underset{\delta\mathbf{x}}{\arg\min} \|\delta\mathbf{x}\|_{\ell_2} \quad \text{subject to} \quad \|\delta\underline{\mathbf{d}} - \overbrace{\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_0;\underline{\mathbf{Q}}]}^{\text{demigration}}\mathbf{S}^*\delta\mathbf{x}\|_2 \leq \sigma$$

$$\begin{aligned}
\delta\mathbf{x} &= \text{Sparse curvelet-coefficient vector} \\
\mathbf{S}^* &= \text{Curvelet synthesis} \\
\underline{\mathbf{Q}} &= \text{Simultaneous sources} \\
\delta\underline{\mathbf{d}} &= \text{Super shots}
\end{aligned}$$

Imaging results

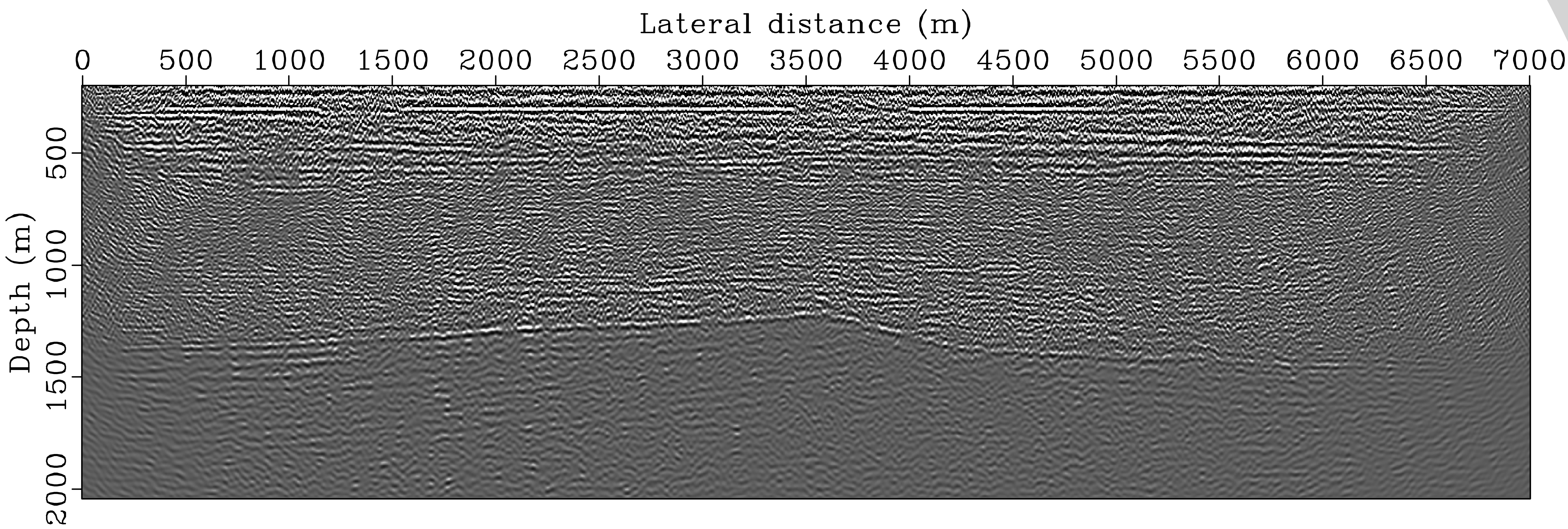
Sparsity-promoting migration with *randomized supershots:*

$$\delta\widetilde{\mathbf{m}} = \mathbf{S}^* \mathop{\arg\min}_{\delta\mathbf{x}} \|\delta\mathbf{x}\|_{\ell_1} \quad \text{subject to} \quad \|\delta\underline{\mathbf{d}} - \overbrace{\nabla\boldsymbol{\mathcal{F}}[\mathbf{m}_0;\underline{\mathbf{Q}}]}^{\text{demigration}}\mathbf{S}^*\delta\mathbf{x}\|_2 \leq \sigma$$

$$\begin{aligned} \delta\mathbf{x} &= \text{Sparse curvelet-coefficient vector} \\ \mathbf{S}^* &= \text{Curvelet synthesis} \\ \underline{\mathbf{Q}} &= \text{Simultaneous sources} \\ \delta\underline{\mathbf{d}} &= \text{Super shots} \end{aligned}$$

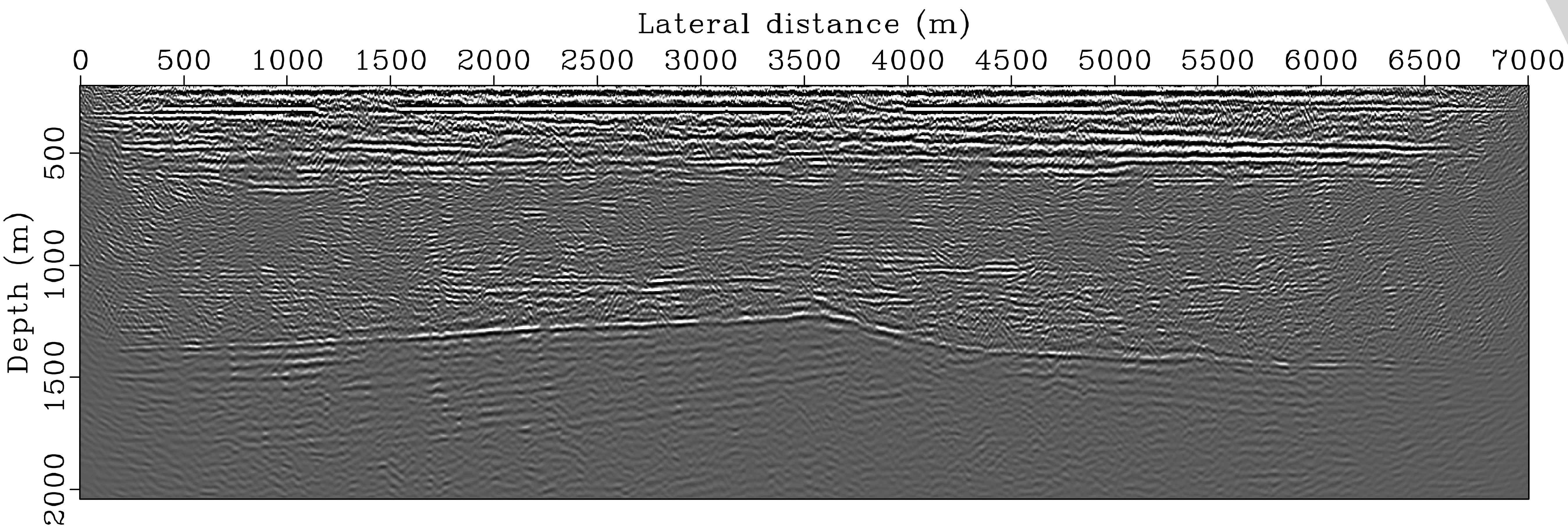
Migration results

[ℓ_2 *without* renewals]



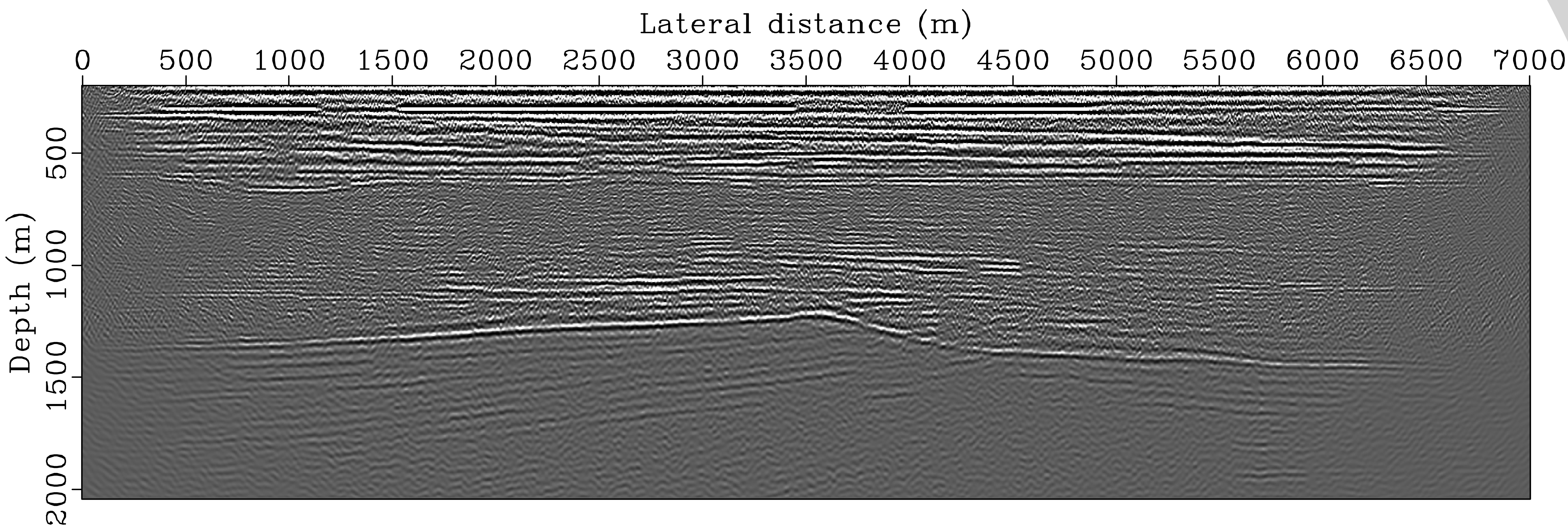
Imaging results

[ℓ_1 *without* renewals]



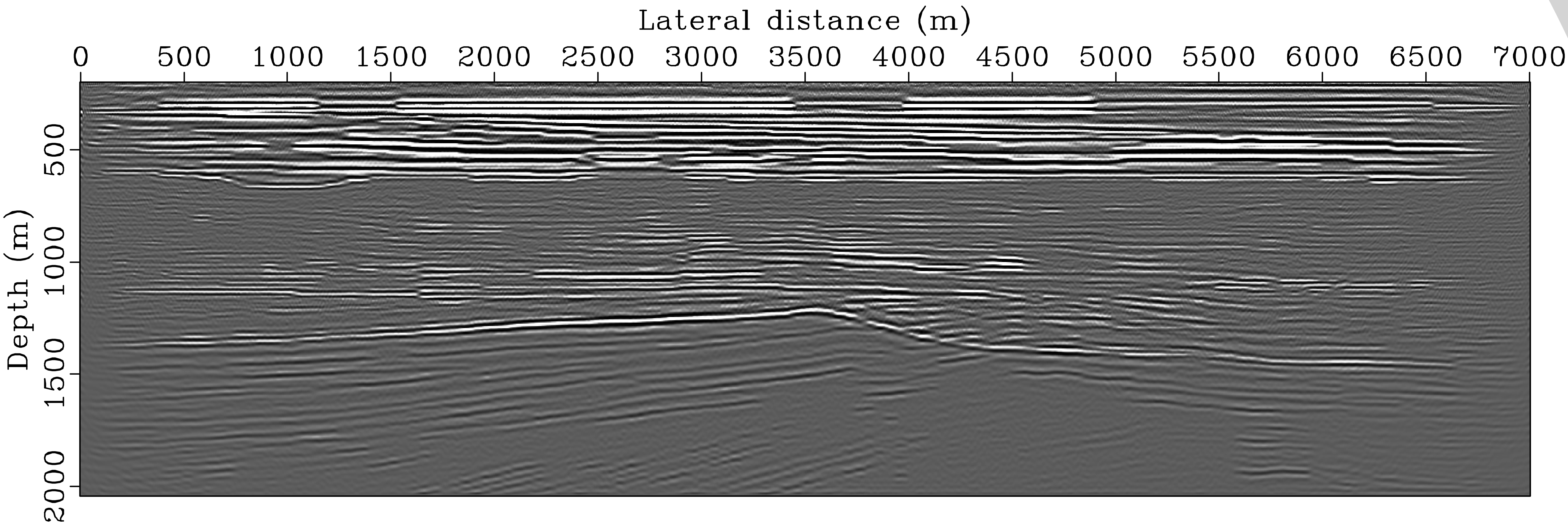
Migration results

[ℓ_2 *with* renewals]



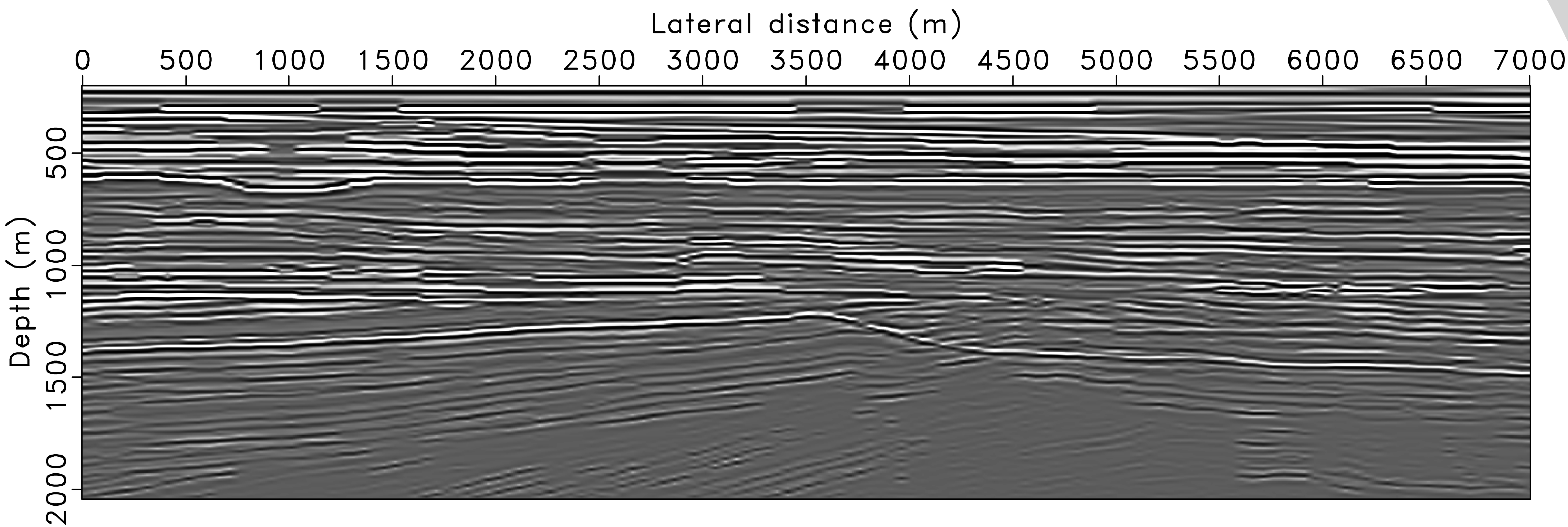
Migration results

[ℓ_1 *with* renewals]



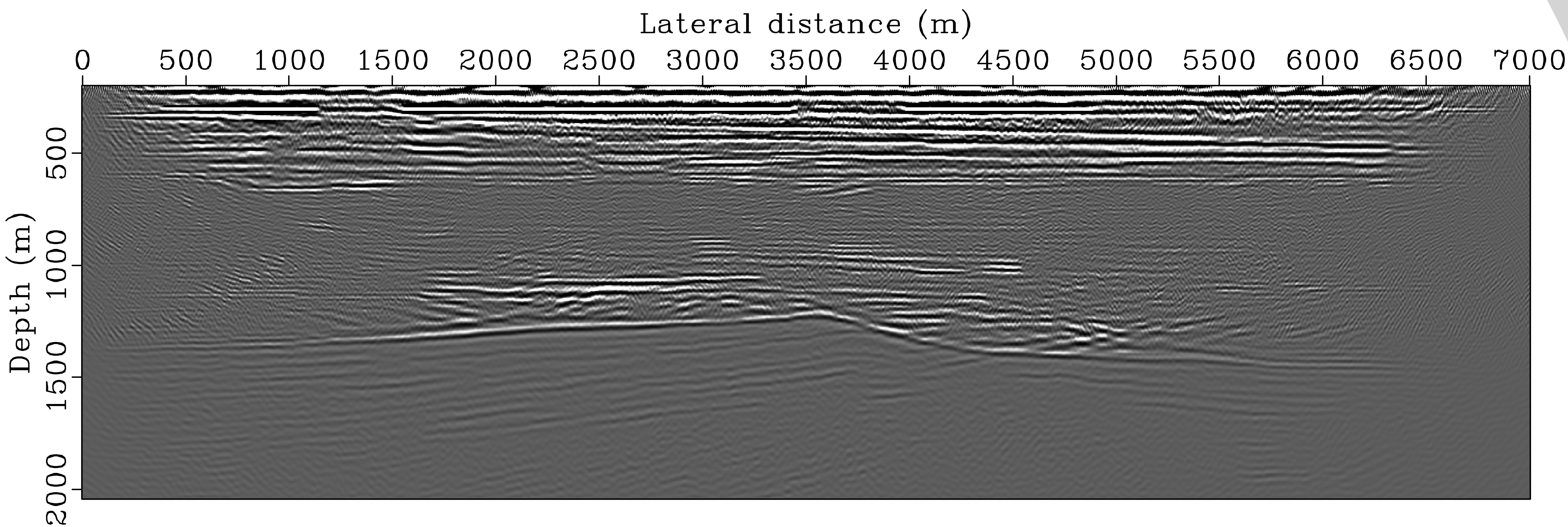
Migration results

[*true* perturbation]



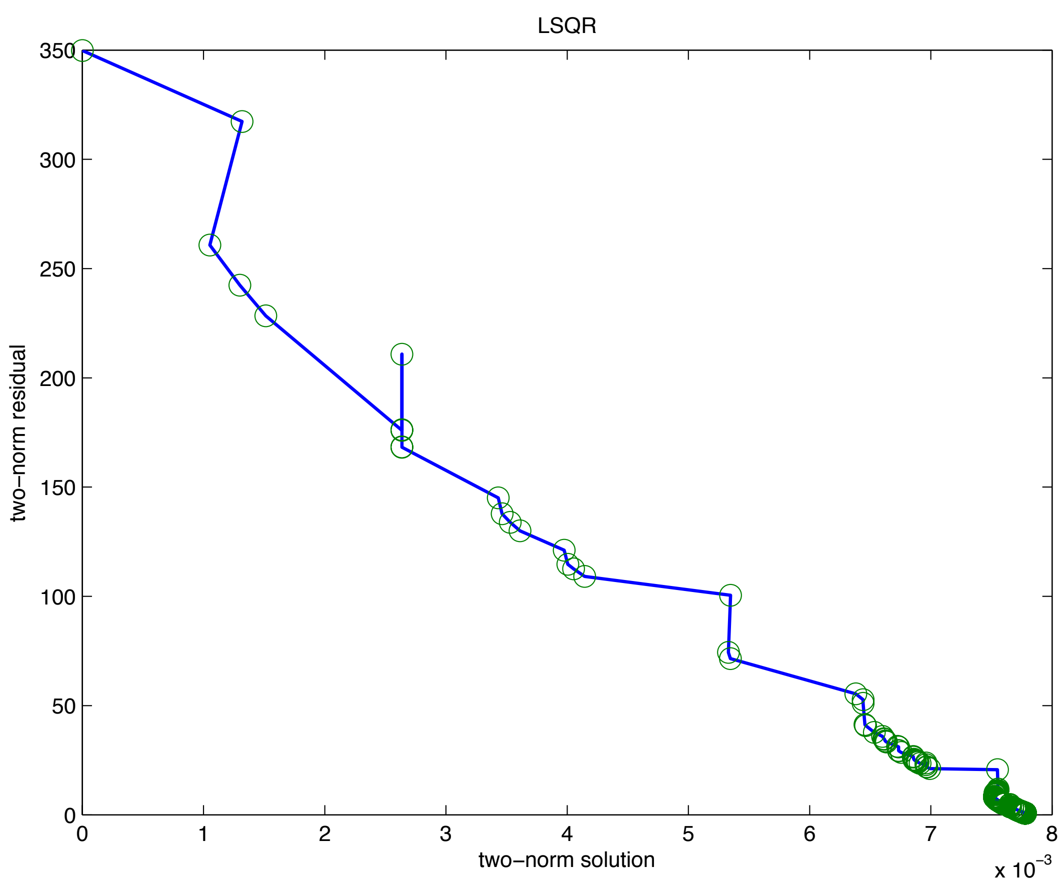
Migration results

[migration with *all* data]

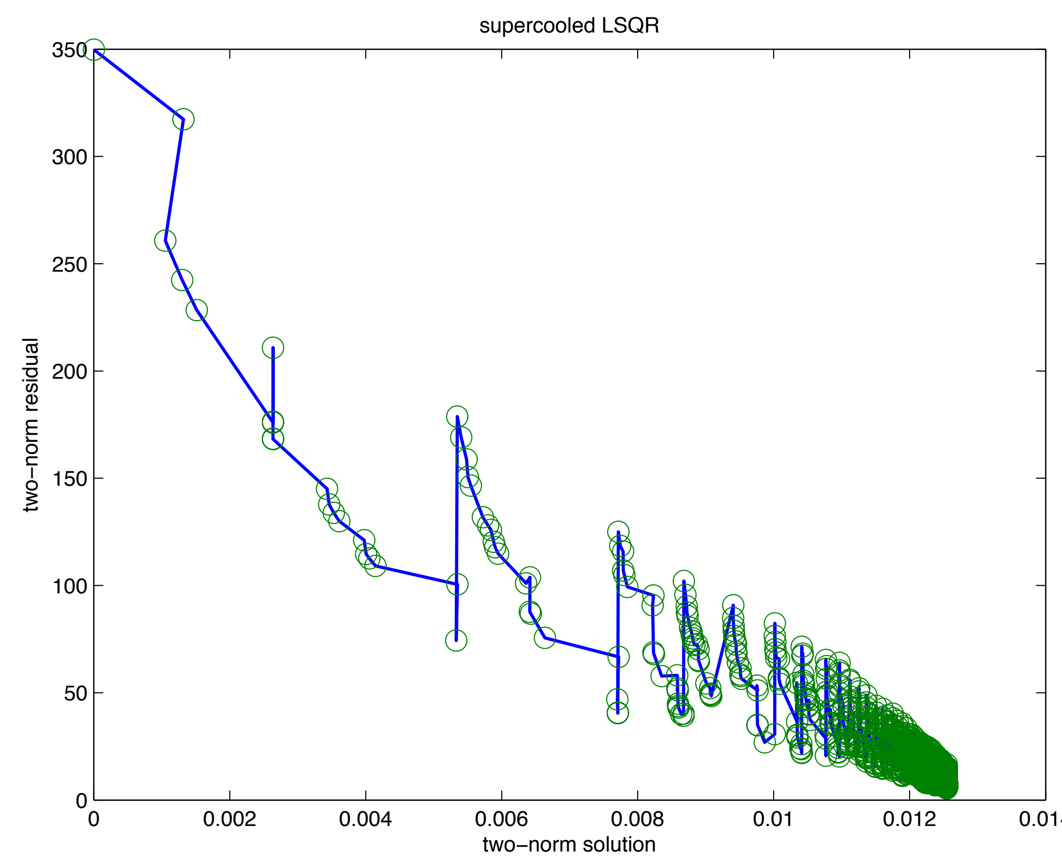


Migration results

[solution paths ℓ_2]



without renewals

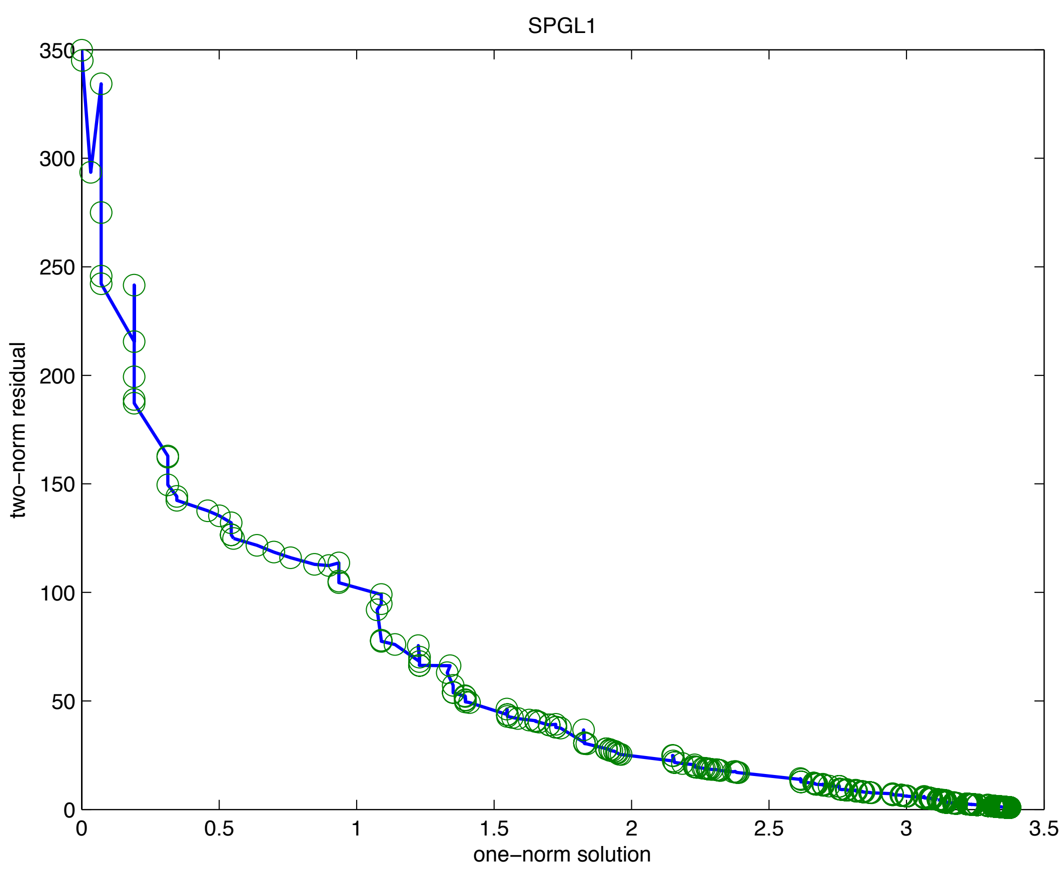


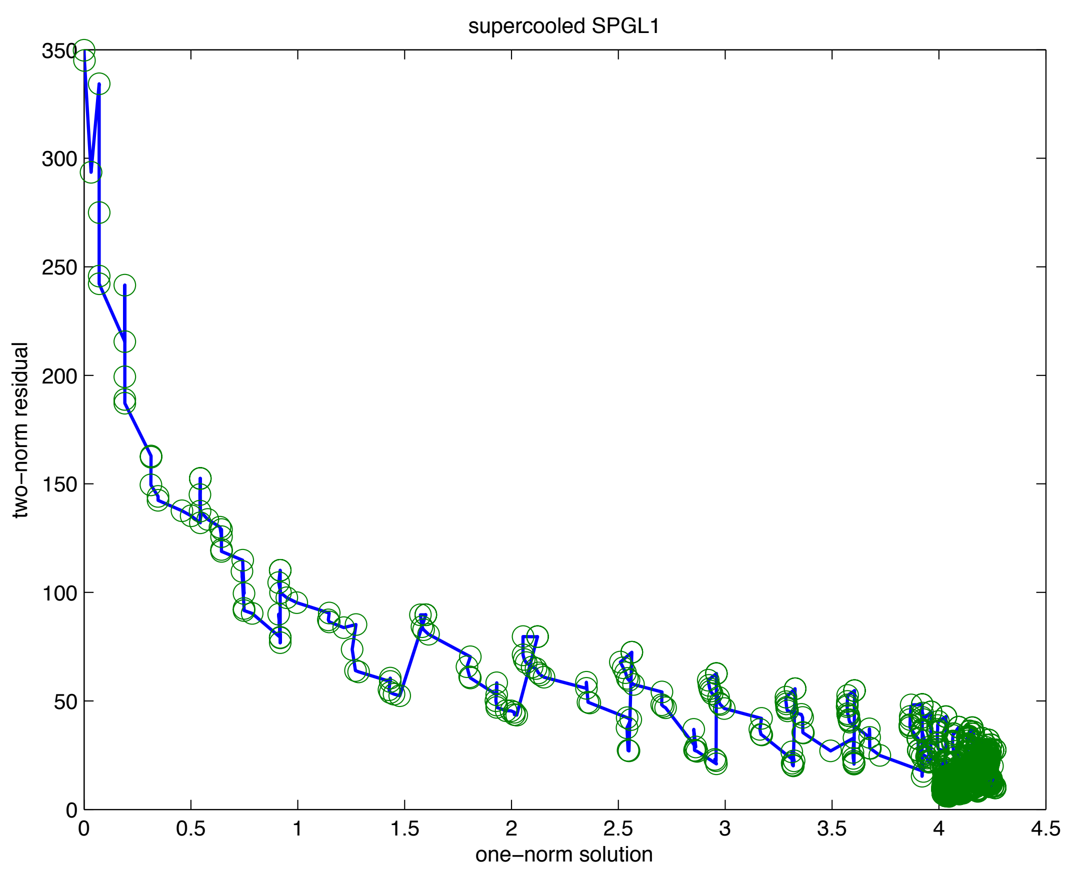
with renewals

SLIM

Migration results

[solution paths ℓ_1]

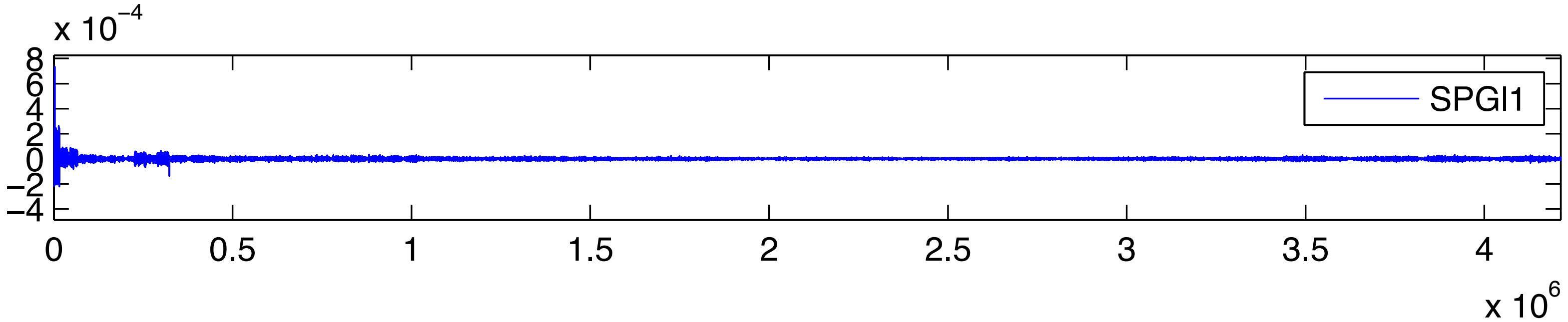


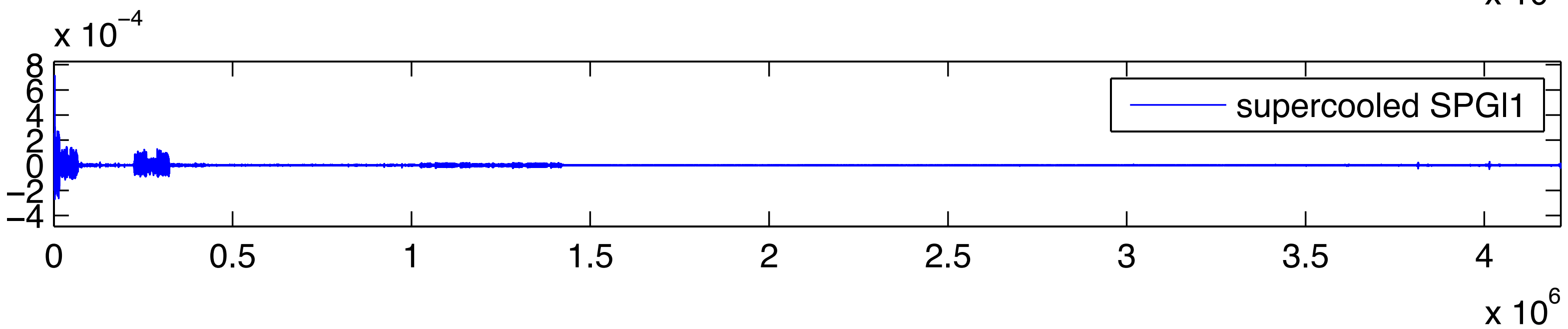


without renewals

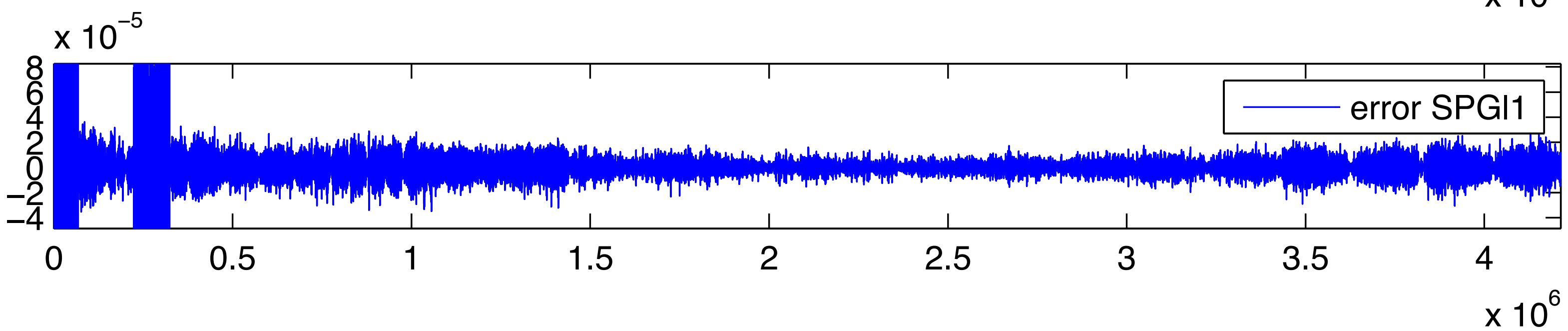
with renewals

Imaging results

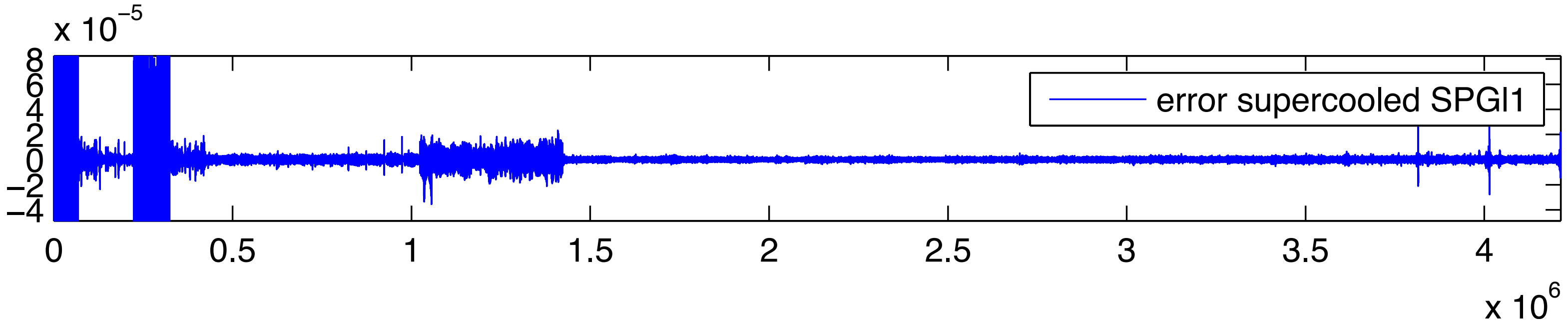




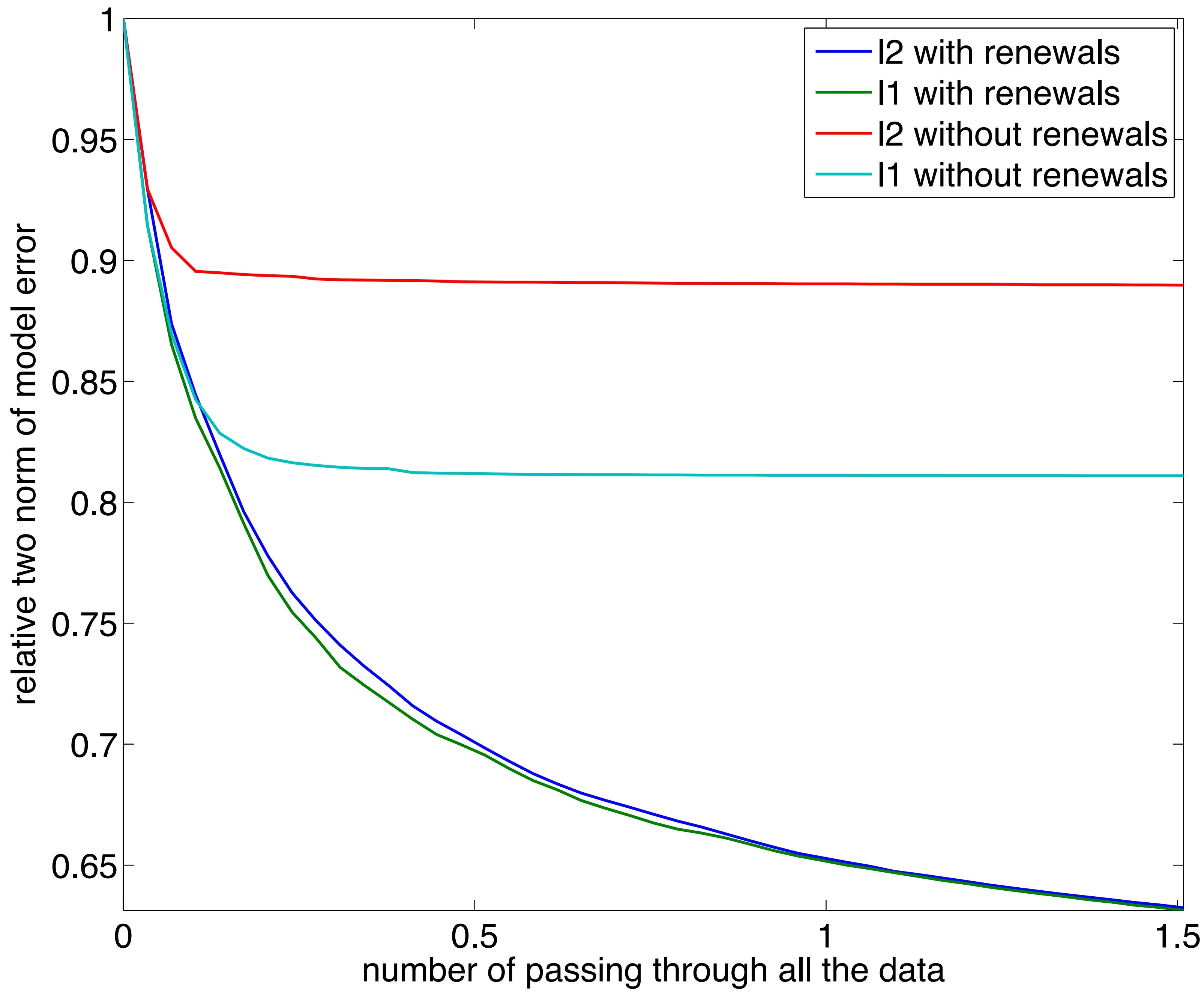
10 X



10 X



Migration results
[model errors]



Conclusions

Message passing improves *image* quality

- computationally *feasible* one-norm *regularization*

Message *passing* via *rerandomization*

- *small* system *size* with *small* IO and memory *imprints*

Possibility to exploit *new* computer *architectures* that employ *model* space *parallelism* to *speed up* wavefield *simulations*...

FWI results

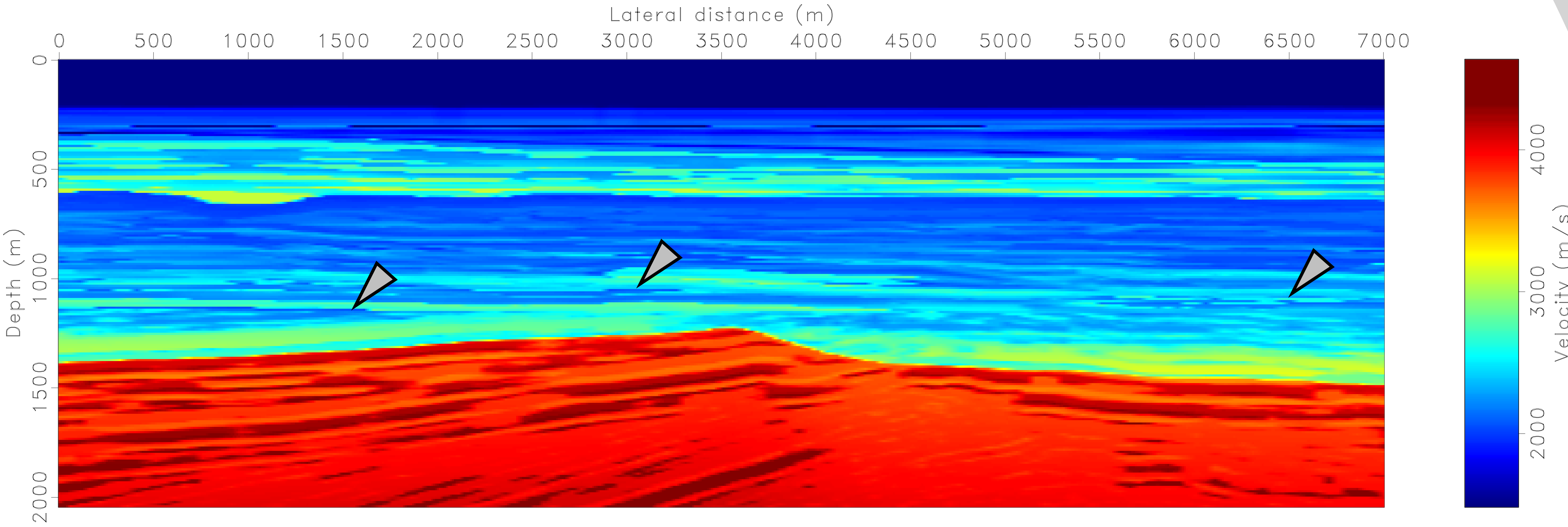
FWI:

- 10 overlapping frequency bands with 10 frequencies (2.9Hz-25Hz)
- 10 Gauss-Newton steps for each frequency band (solved with max 20 spectral-projected gradient iterations)

[Bunks ‘95; Pratt ’98]

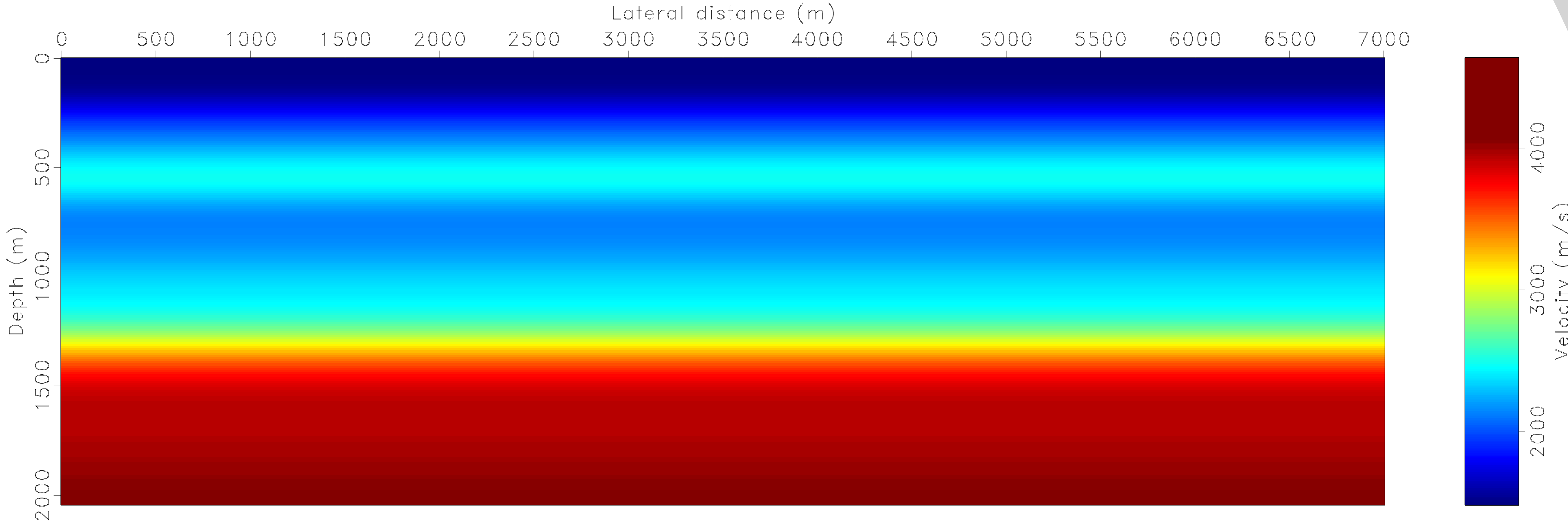
Results GN-FWI

True model



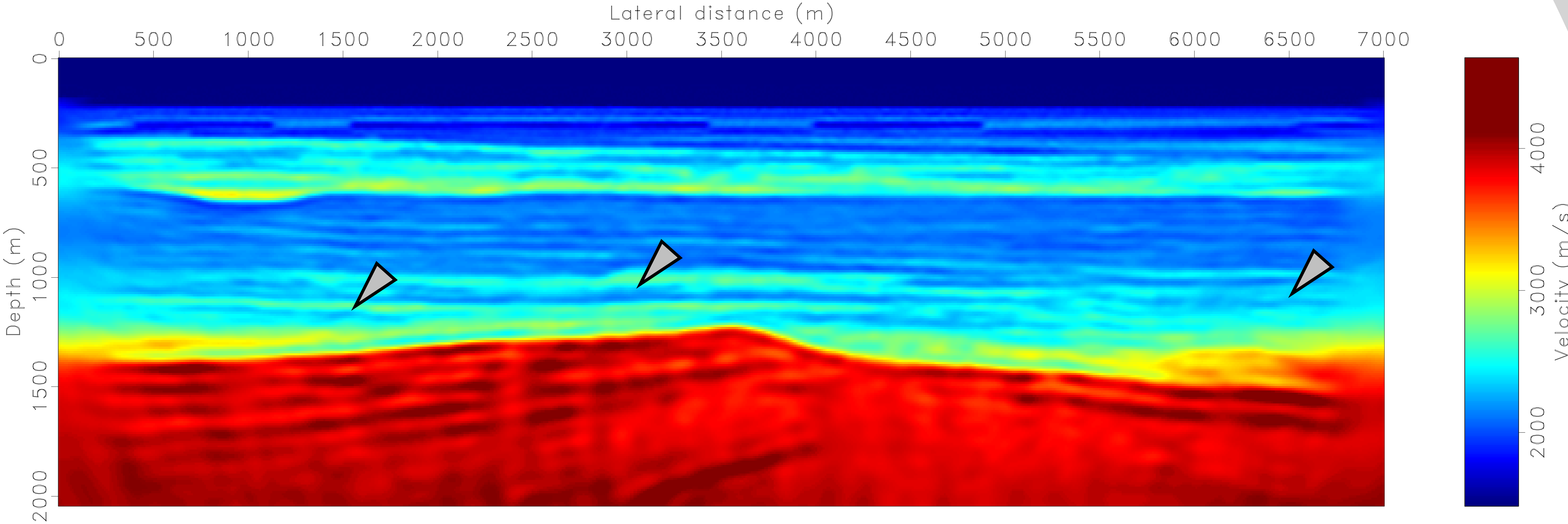
Results GN-FWI

Initial model



Results GN-FWI

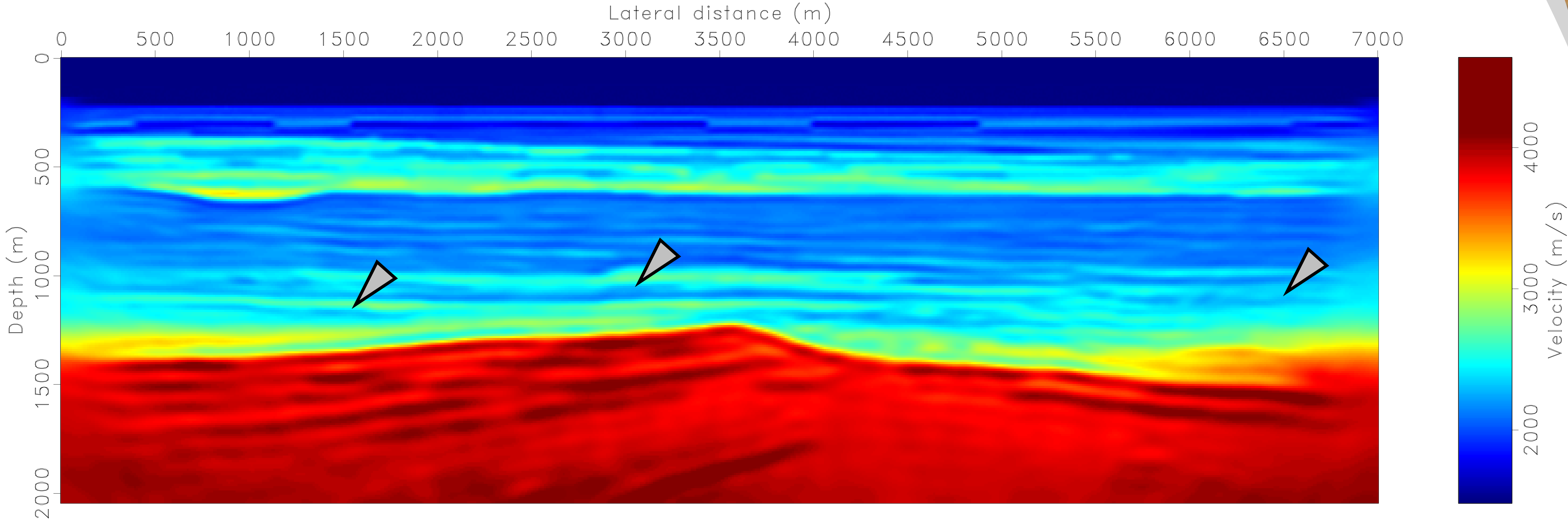
Modified GN 7 sim. shots *without renewals*



25 times speedup compared to full GN

Results GN-FWI

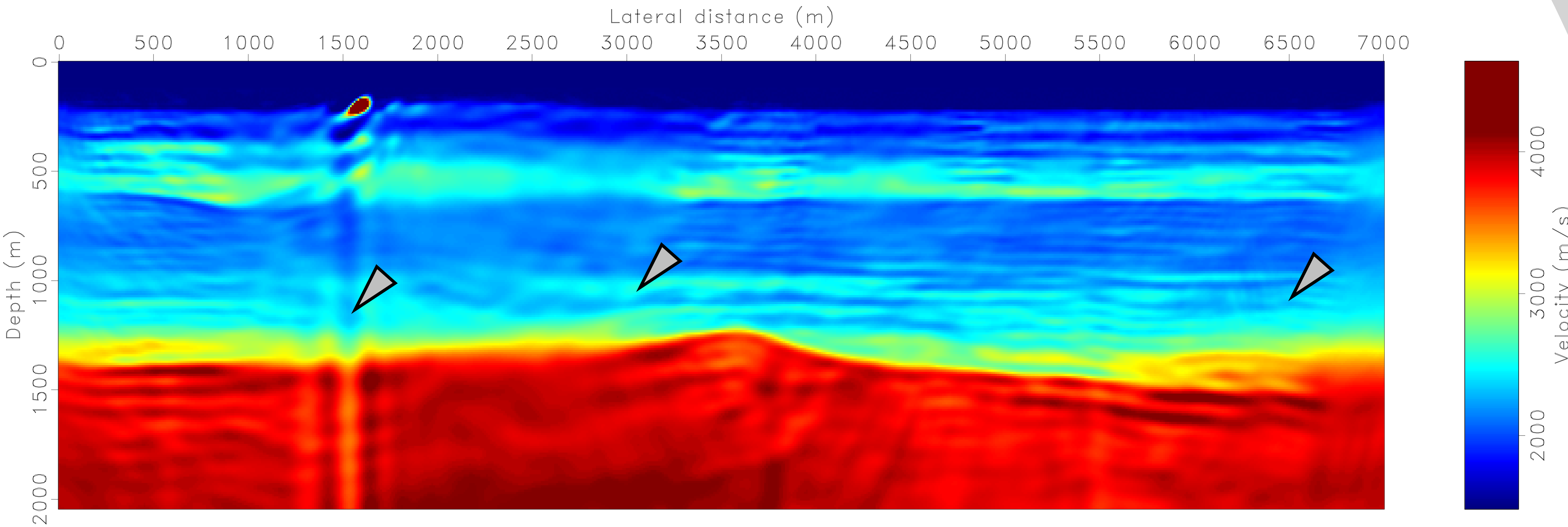
Modified GN 7 sim. shots *with renewals*



25 times speedup compared to full GN

Results GN-FWI

Modified GN 7 seq. shots *without renewals*

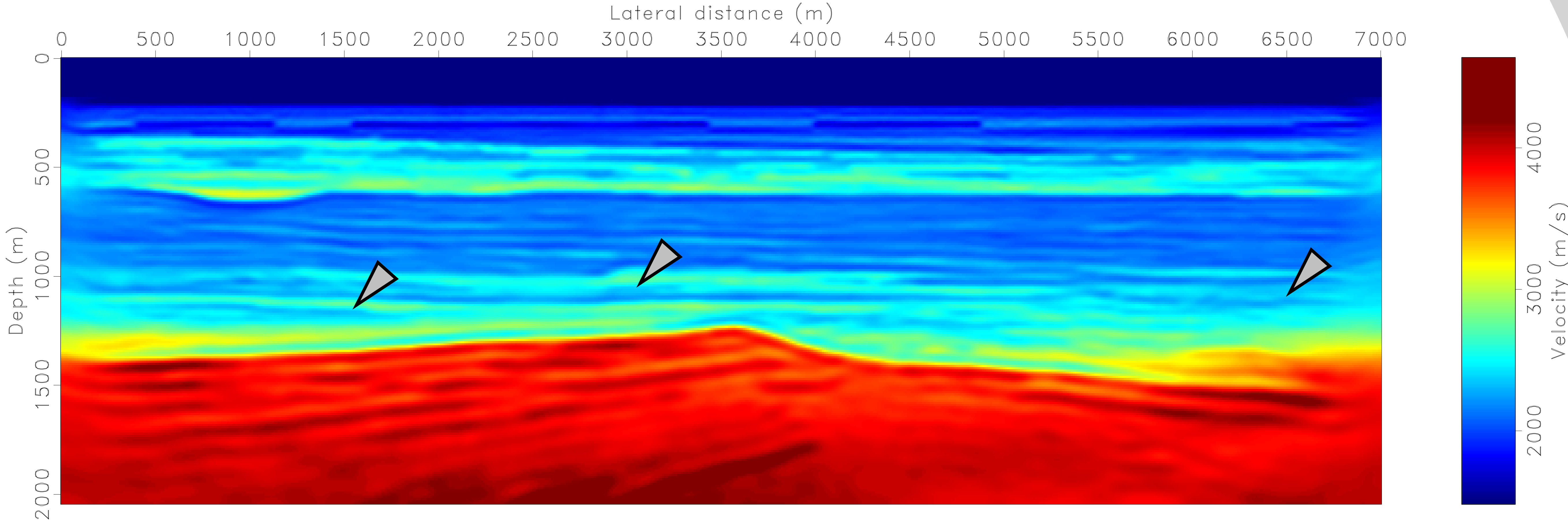


25 times speedup compared to full GN

SLIM

Results GN-FWI

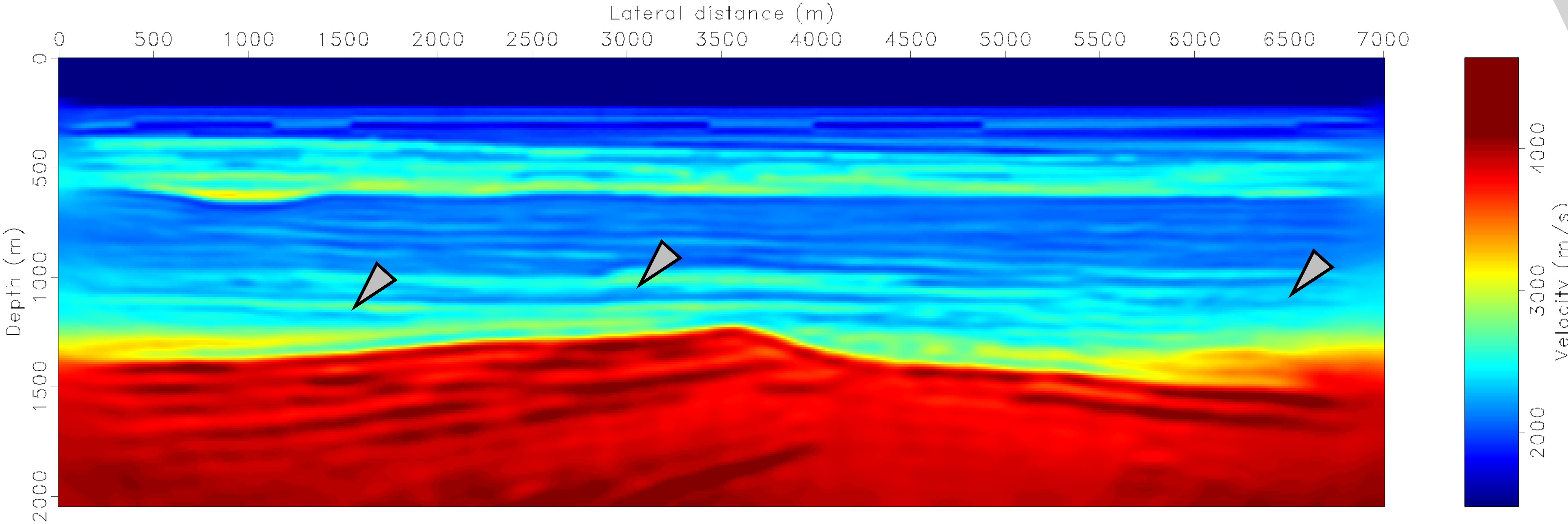
Modified GN 7 seq. shots *with renewals*



25 times speedup compared to full GN

Results GN-FWI

Modified GN 7 sim. shots *with renewals*



25 times speedup compared to full GN