Pass on the message: recent insights in large-scale sparse recovery

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

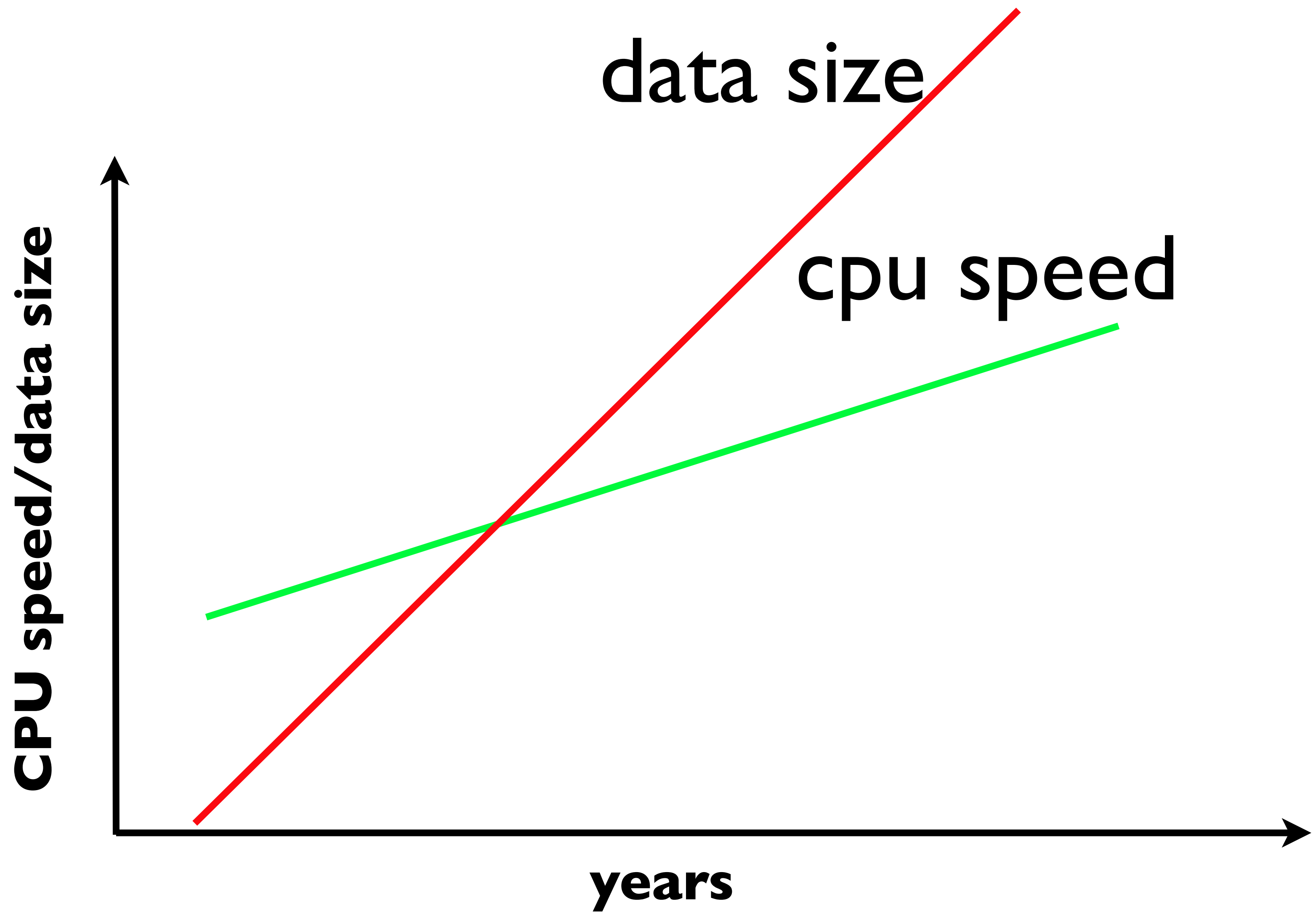
"We are drowning in data but starving for understanding" USGS director Marcia McNutt

"Got data now what" Carlsson & Ghrist SIAM

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

Problem

"Data explosion is bigger than Moore's law"



Goals

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

- relies on *touching* **all** *data* all the *time*

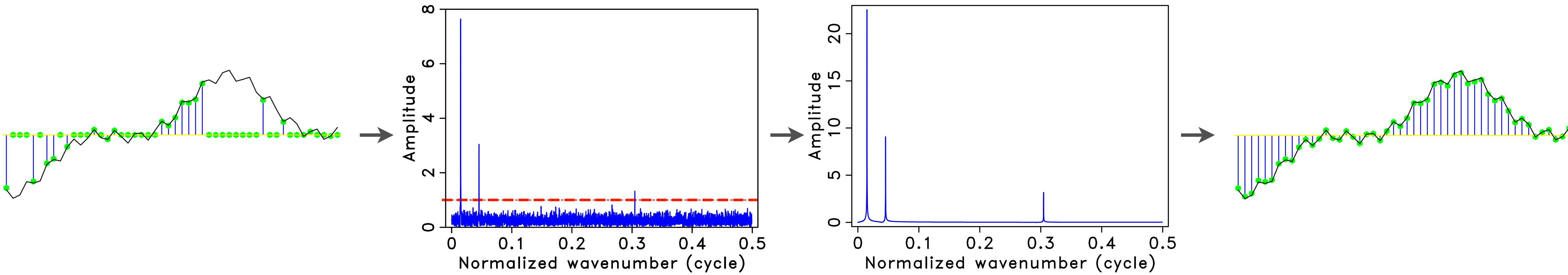
by an *agile* optimization *paradigm* that works on

- **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

$$\mathbf{A} := \mathbf{R}\mathbf{F}^H$$

sensing matrix

inverse Fourier transform

$$\mathbf{A} \in \mathbb{C}^{n \times N} \text{ with } n \ll N$$

[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

soft threshold ↓

$$\mathbf{x}^{t+1} = \eta_t \left(\mathbf{A}^* \mathbf{r}^t + \mathbf{x}^t \right)$$

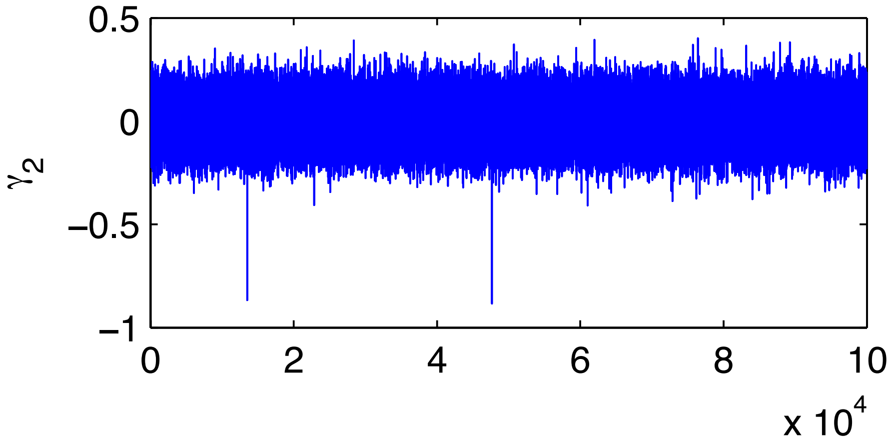
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t$$

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

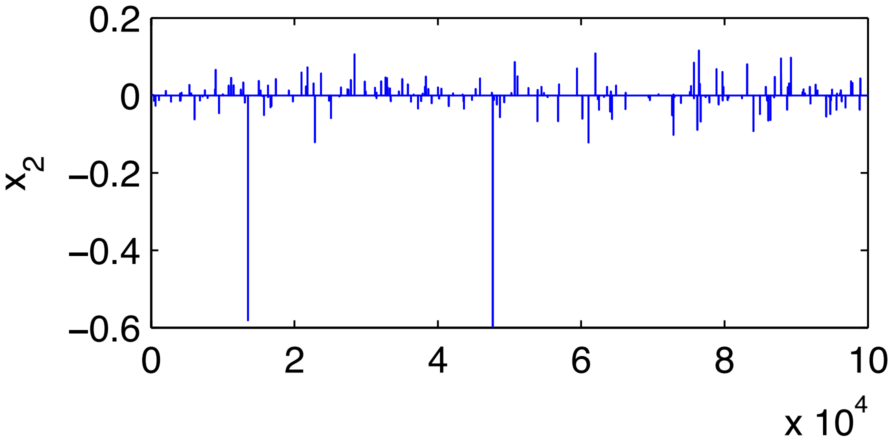
But does the *same* hold for later (t>1) *iterations*...?

Iteration t=1

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

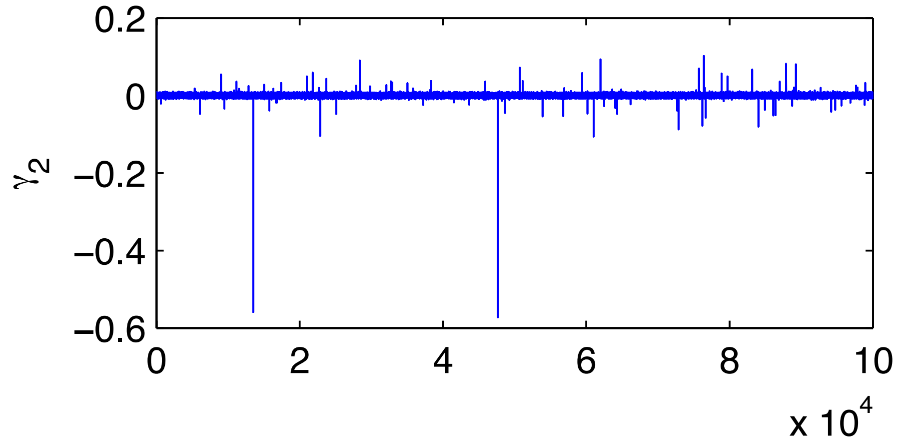


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

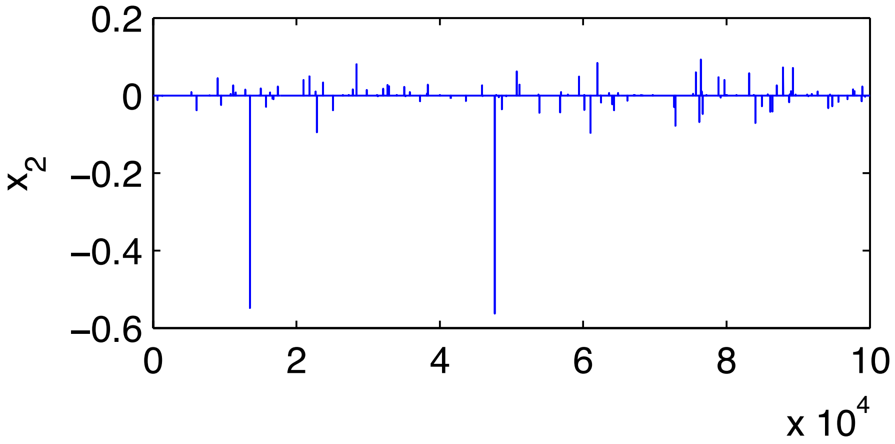


Iteration t=2

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

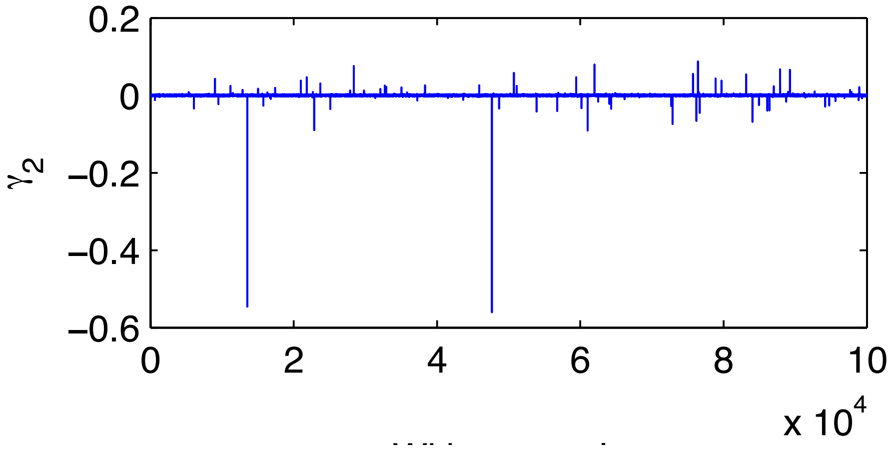


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

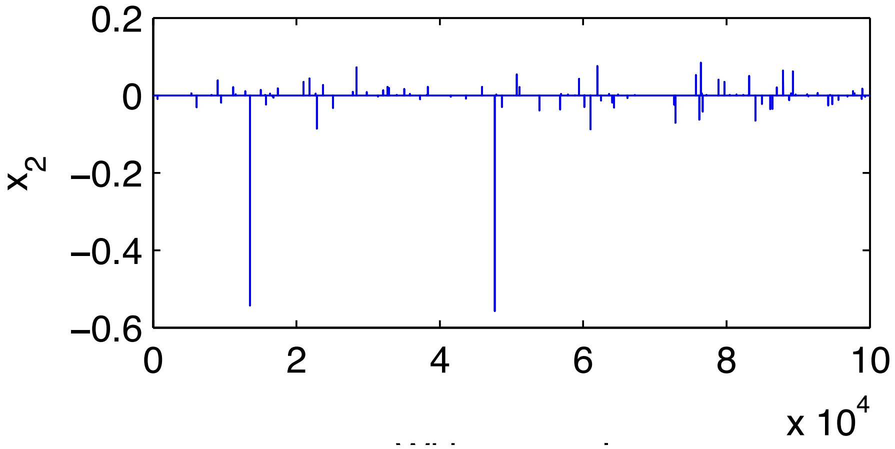


Iteration t=3

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

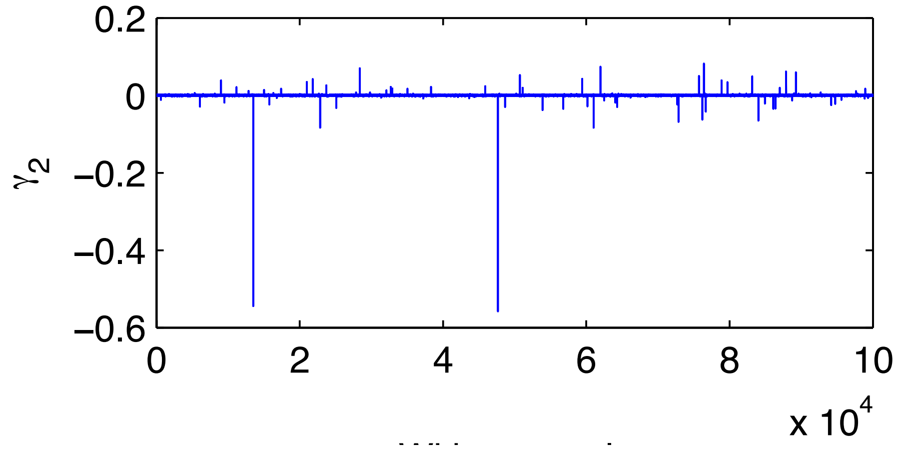


$$\eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$

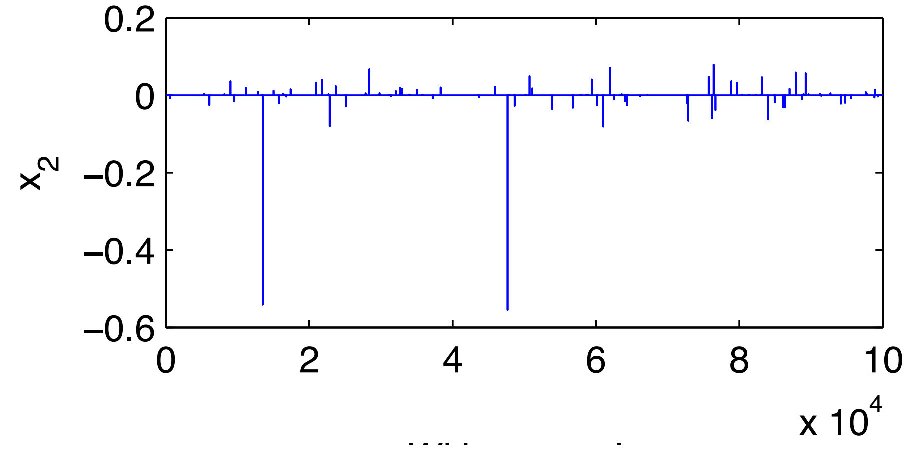


Iteration t=4

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



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



Approximate message passing

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

$$\mathbf{x}^{t+1} = \eta_t\left(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t\right)$$

$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1}$$

← *"message"*

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

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

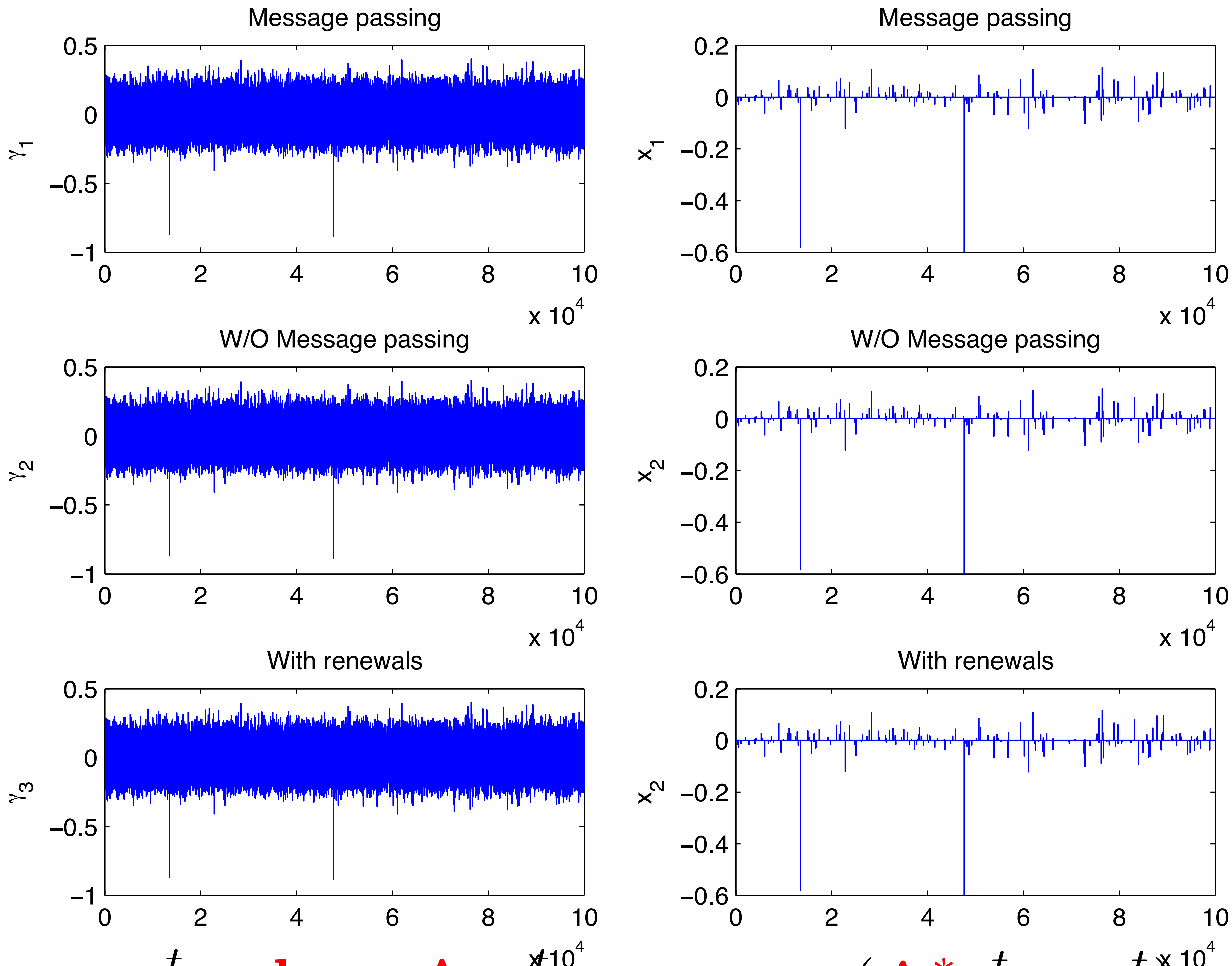
by drawing *new independent* pairs $\{\mathbf{b}_t, \mathbf{A}_t\}$ for *each* iteration

Changes the story *completely*

- breaks *correlation* buildup between model iterate $\mathbf{x}^t$ & the matrix $\mathbf{A}$
- *faster* convergence

Iteration t=1

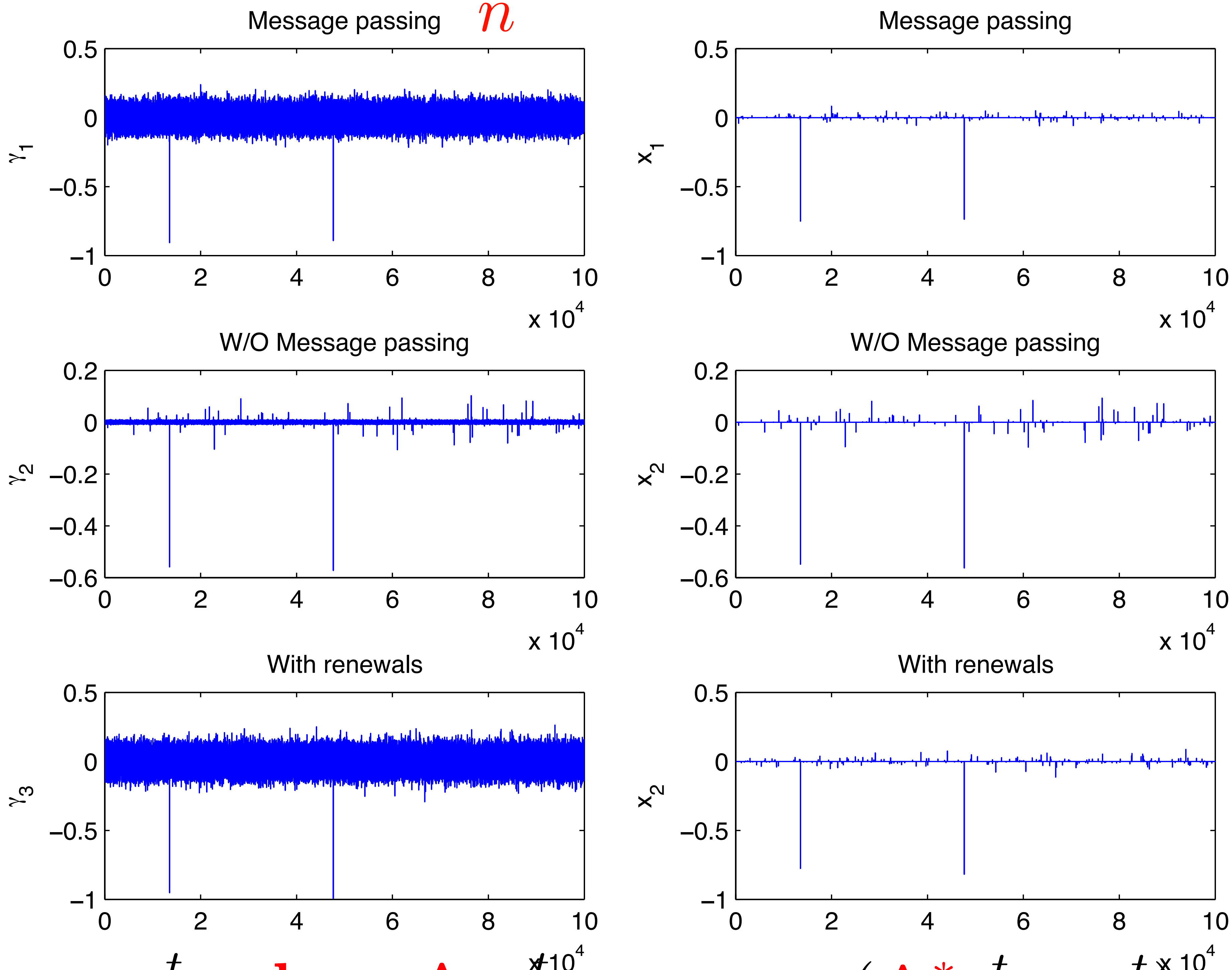
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



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

Iteration t=2

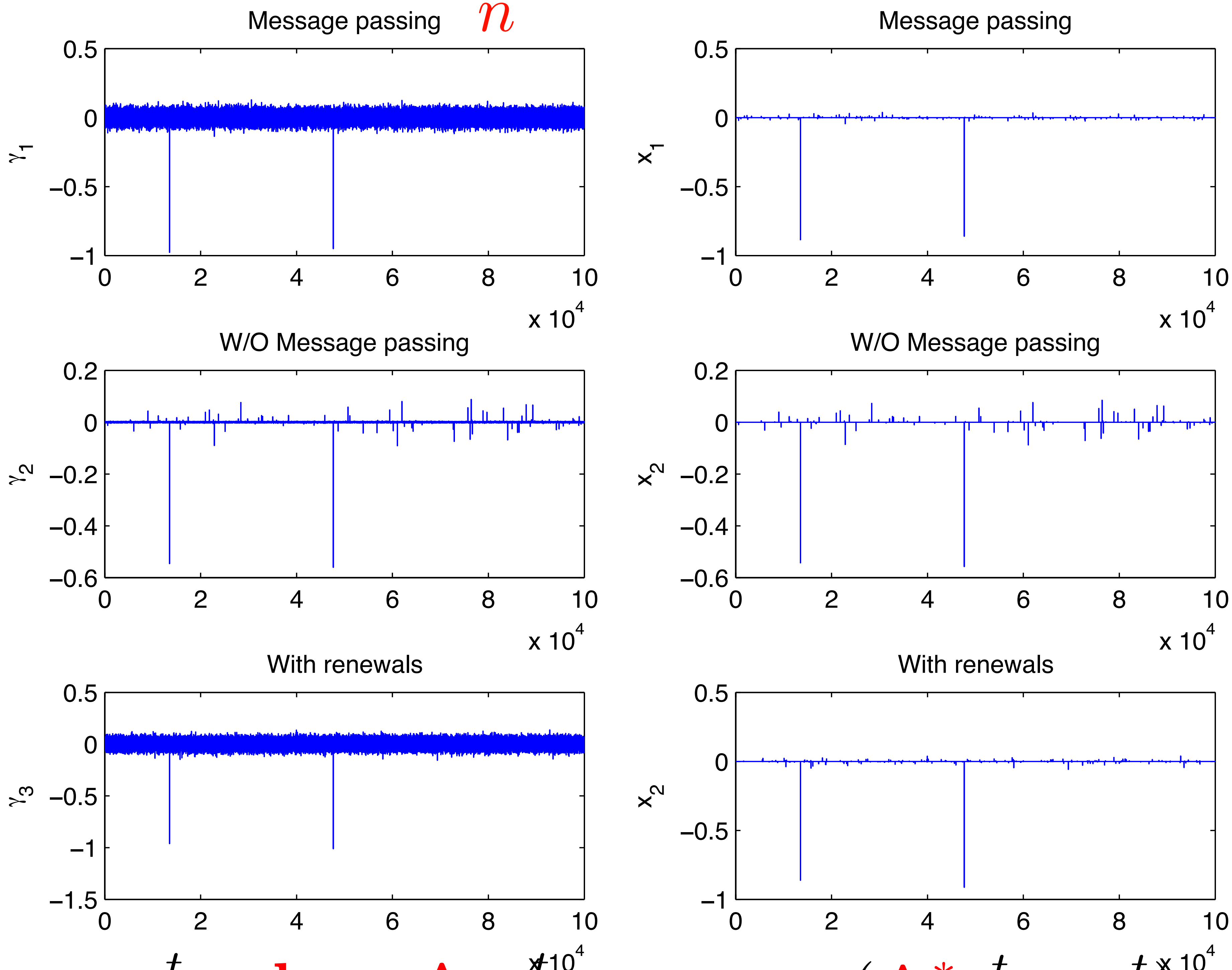
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



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

Iteration t=3

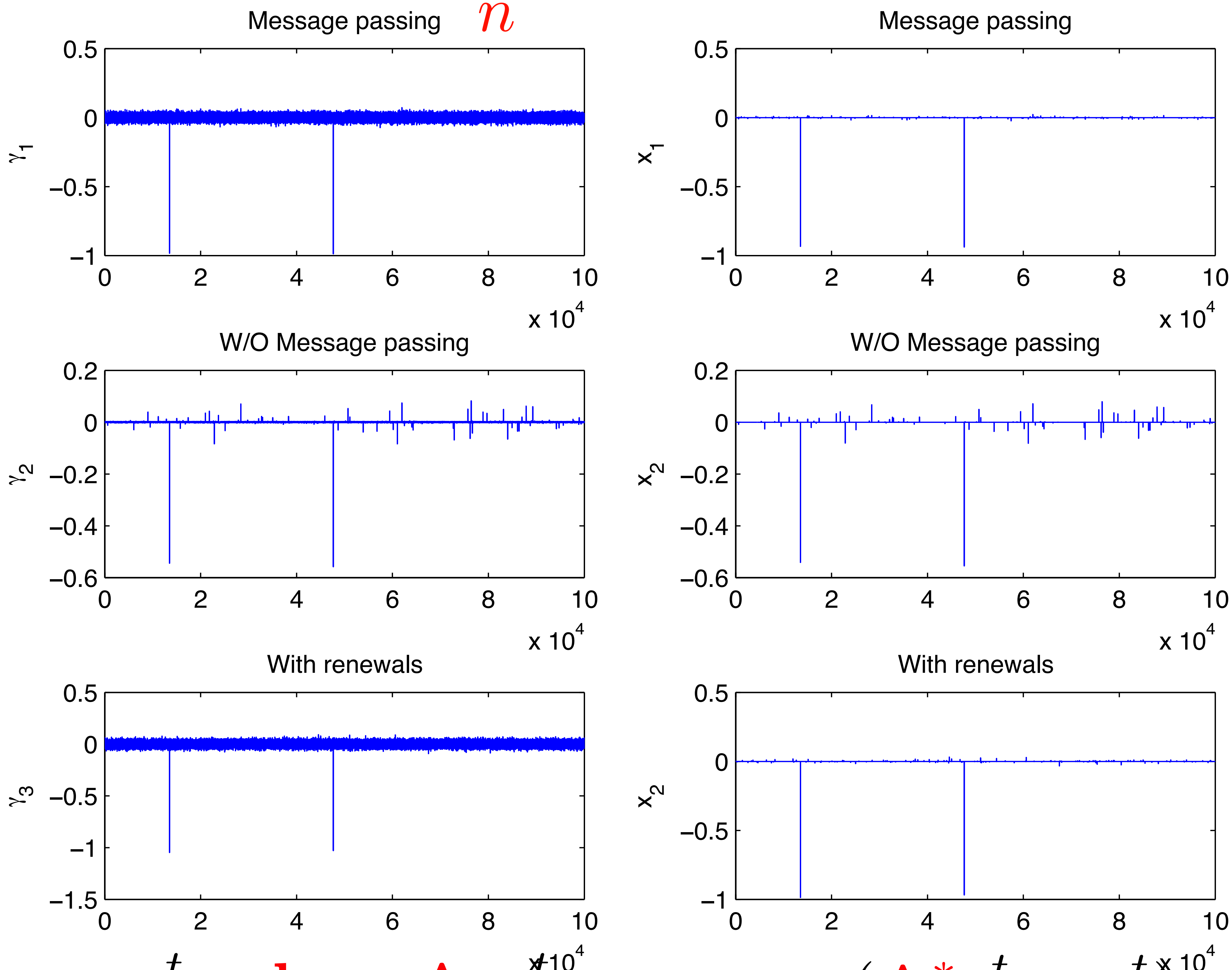
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



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

Iteration t=4

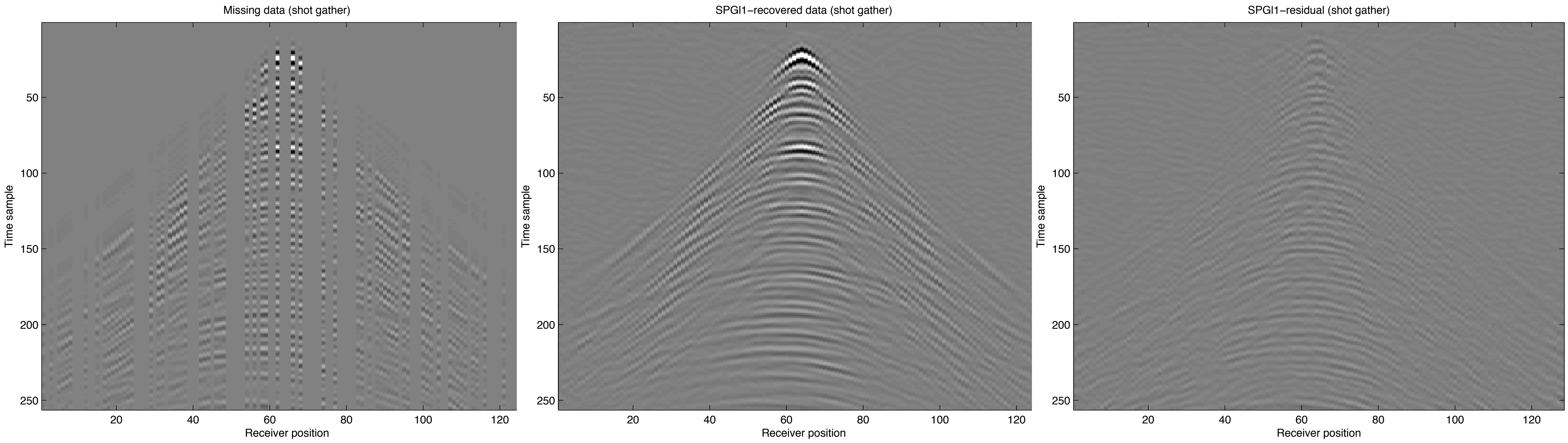
$$\mathbf{r}^t = \mathbf{b} - \mathbf{A}\mathbf{x}^t + \frac{\|\mathbf{x}^{t+1}\|_0}{n}\mathbf{r}^{t-1} \qquad \eta_t(\mathbf{A}^*\mathbf{r}^t + \mathbf{x}^t)$$



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

Missing-trace interpolation [SPGl1]

Recovery with 3D curvelets (N=1.12 X 10^9) 7.75 dB



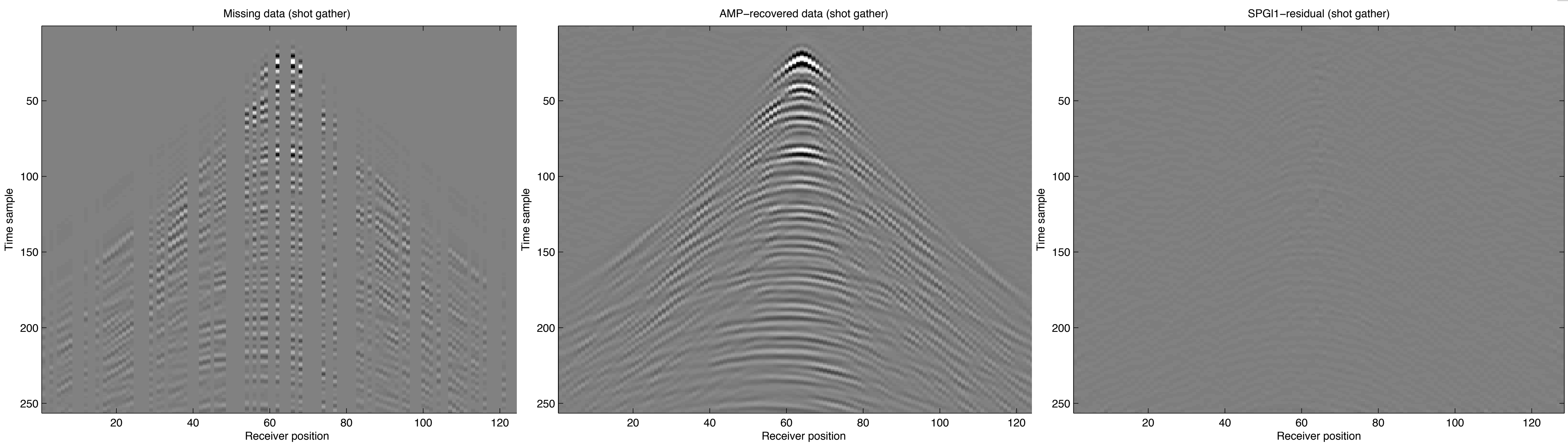
50 % *missing* data

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

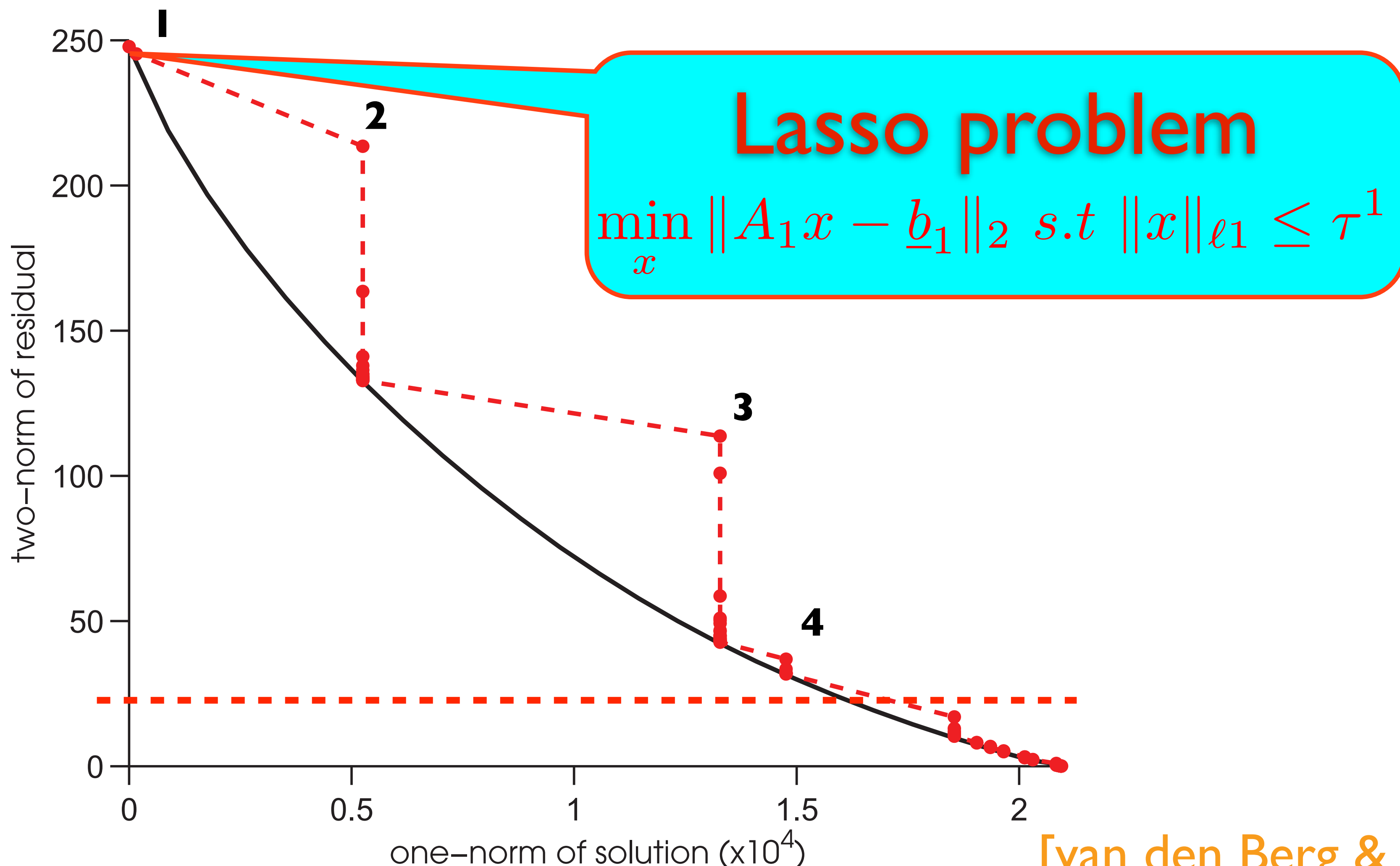
Message-pass term has the same *effect* as drawing *independent* experiments $\{\mathbf{b}_t, \mathbf{A}_t\}$

- "*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?

Supercooled

spectral-projected gradients



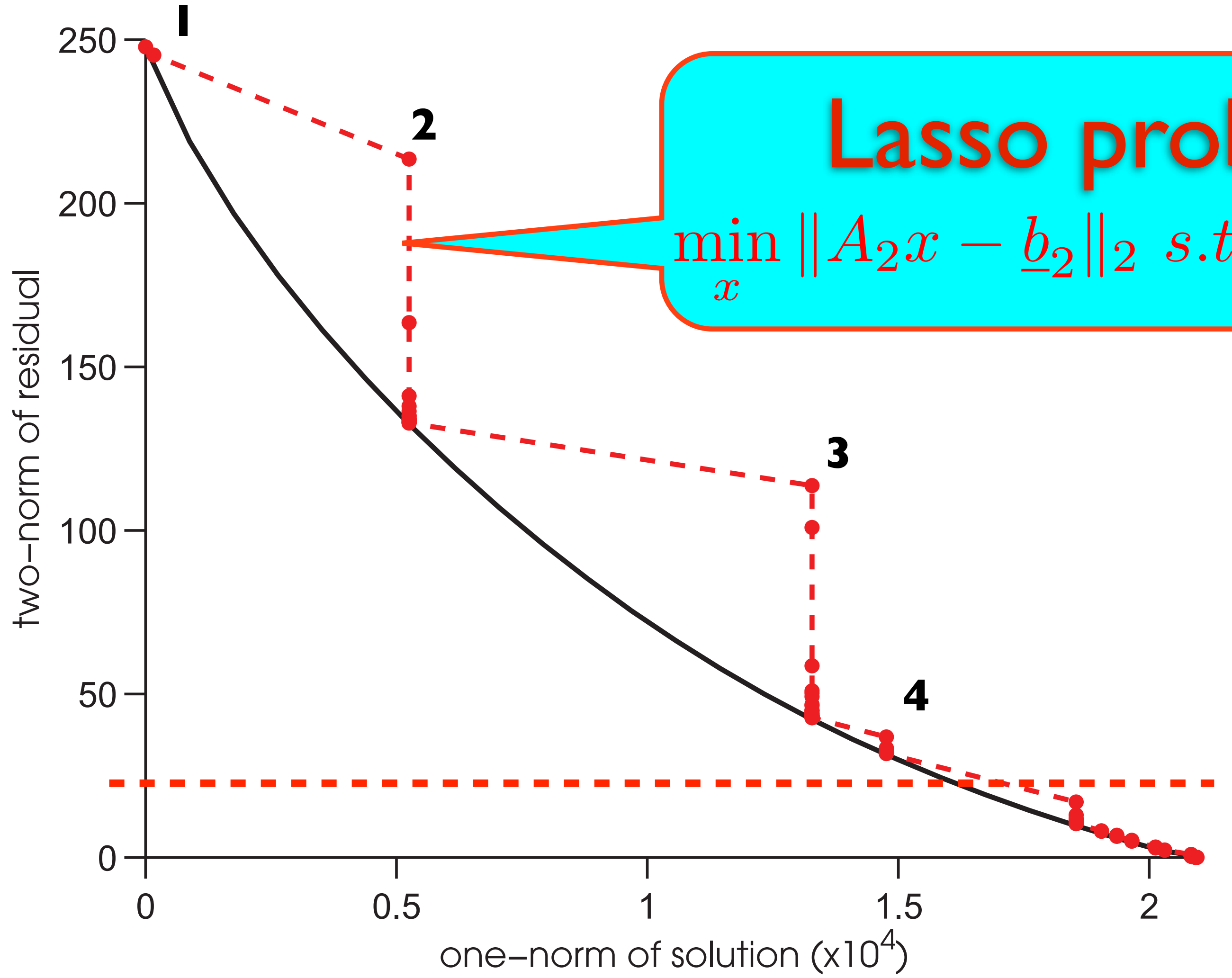
[van den Berg & Friedlander, '08]

[Hennefent et. al., '08]

[Lin & FJH, '09-]

Supercooled

spectral-projected gradients

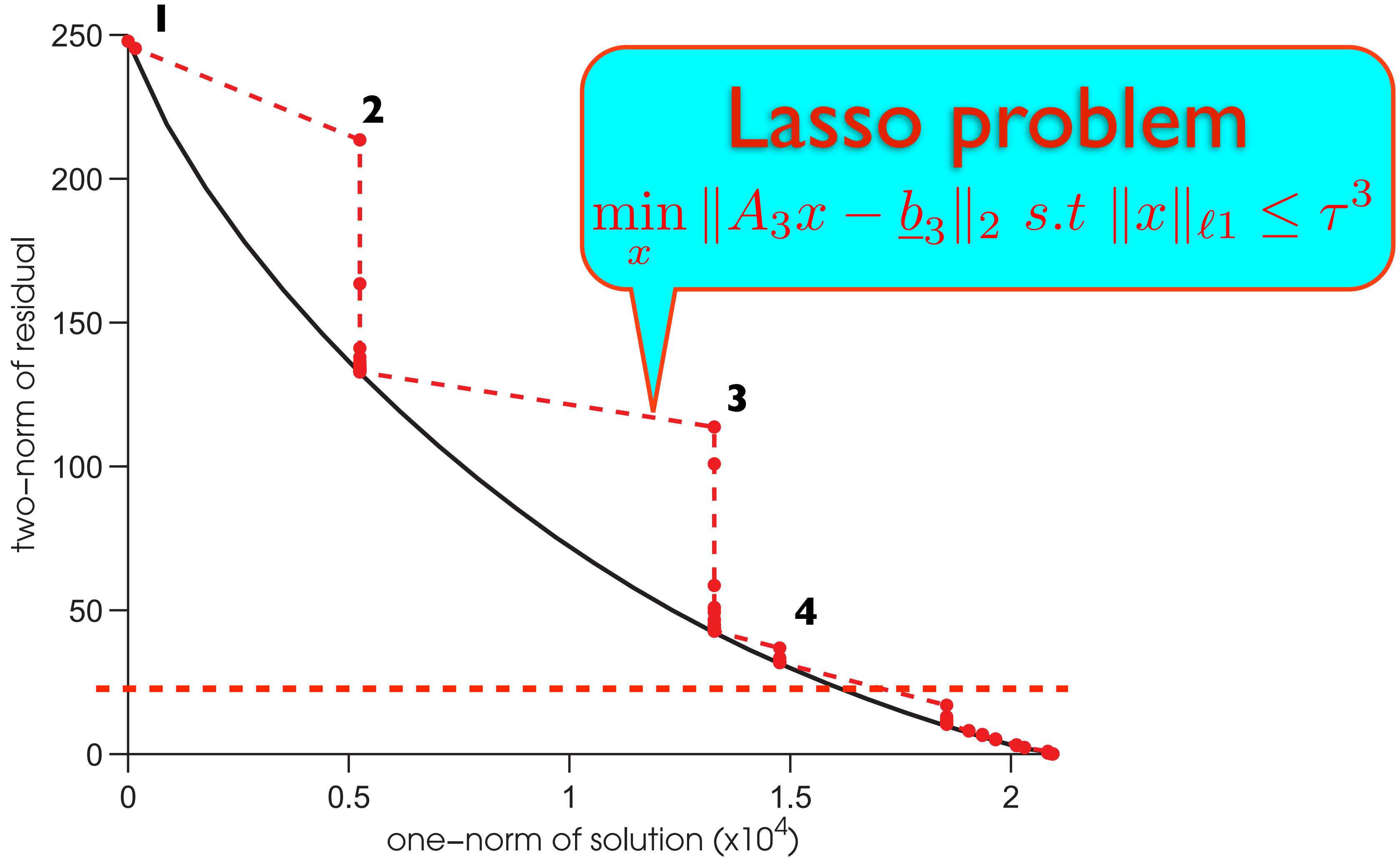


Lasso problem

$$\min_x \|A_2 x - \underline{b}_2\|_2 \ s.t \ \|x\|_{\ell 1} \leq \tau^2$$

Supercooled

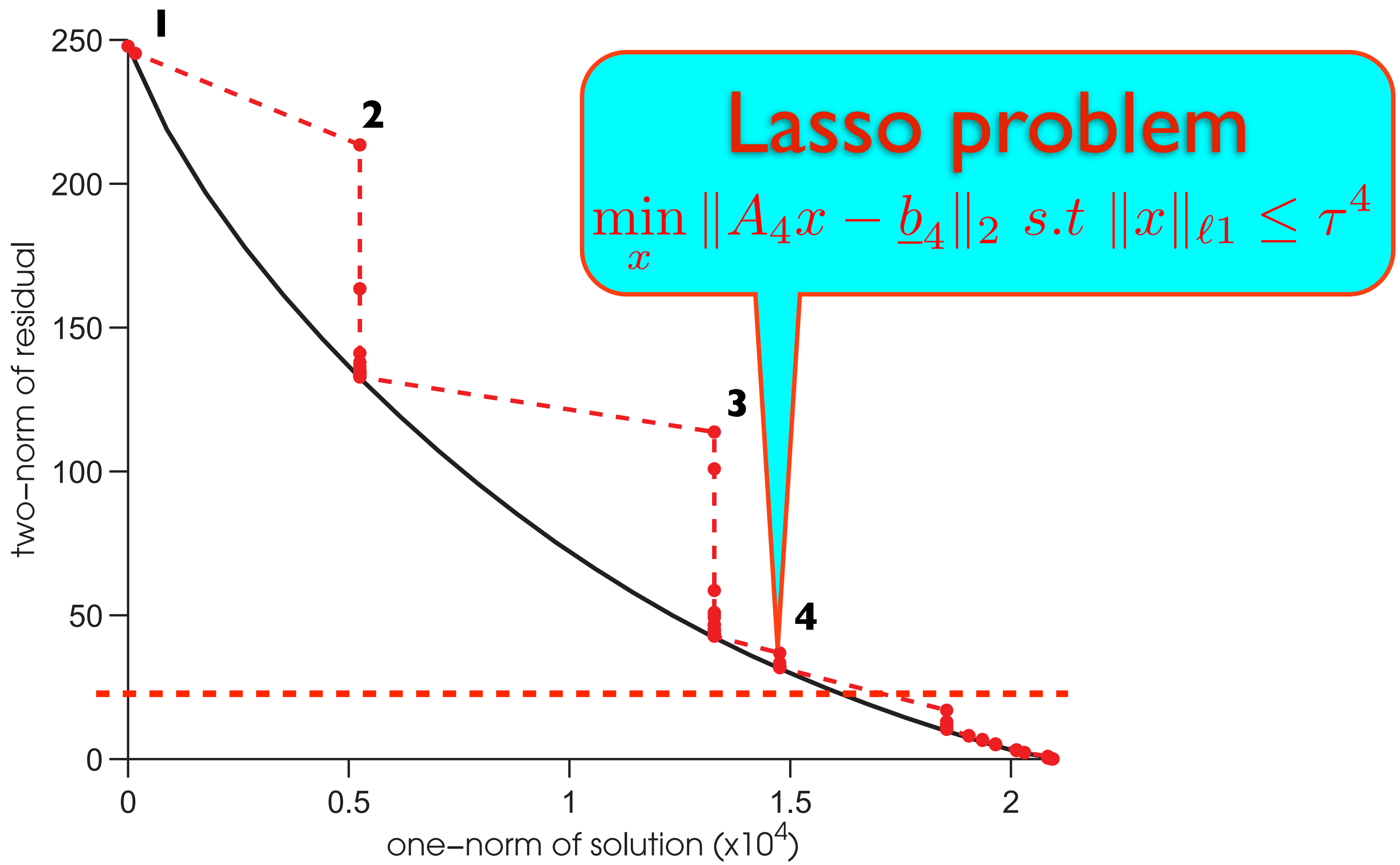
spectral-projected gradients



SLIM

Supercooled

spectral-projected gradients



Supercooling

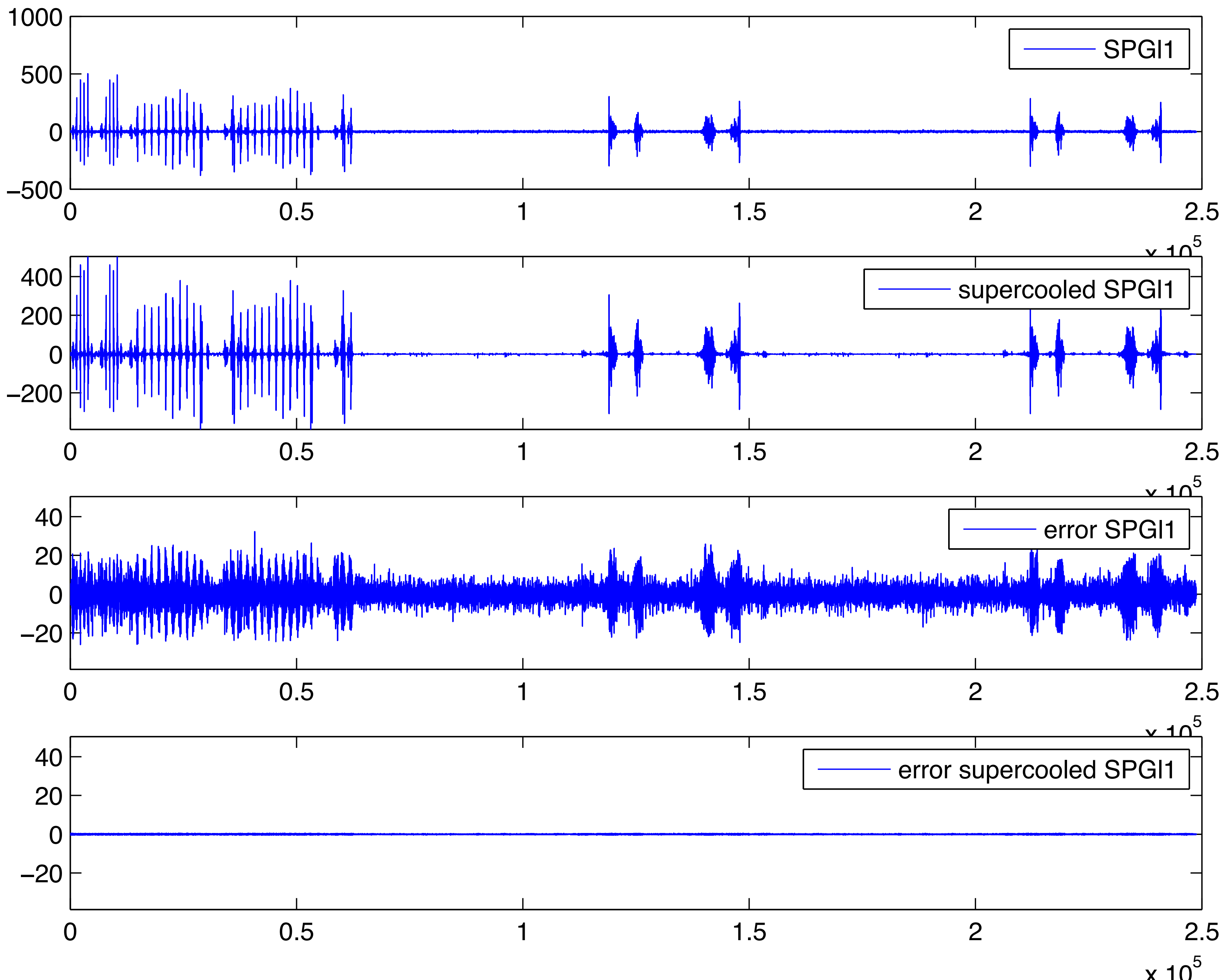
Break *correlations* between the model *iterate* and matrix $\mathbf{A}$ by *rerandomization*

- 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

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

10 X

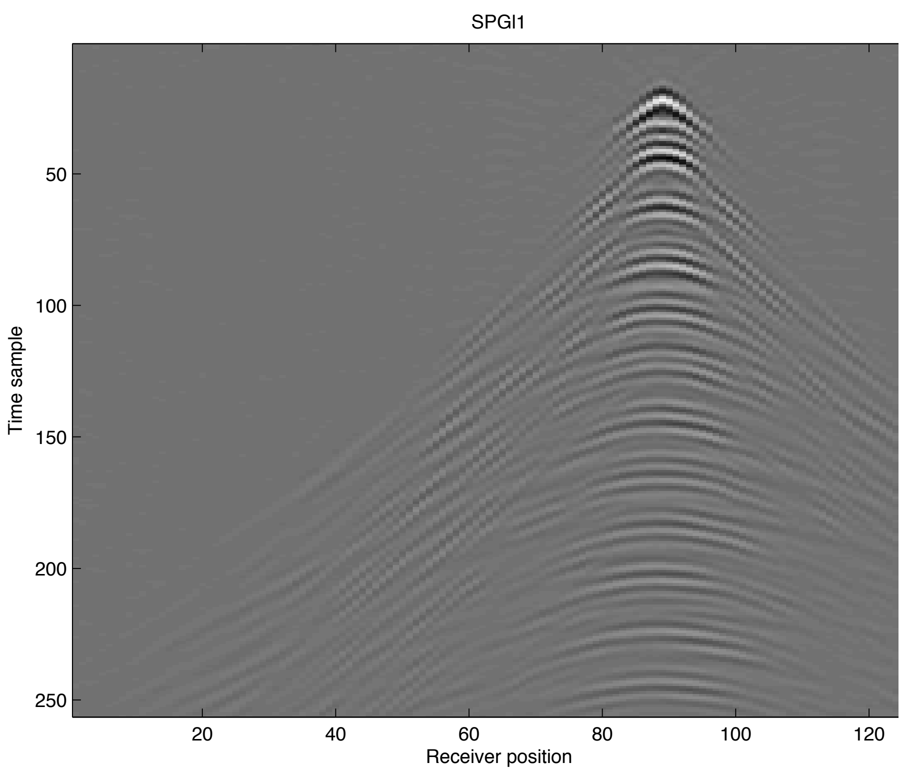
10 X



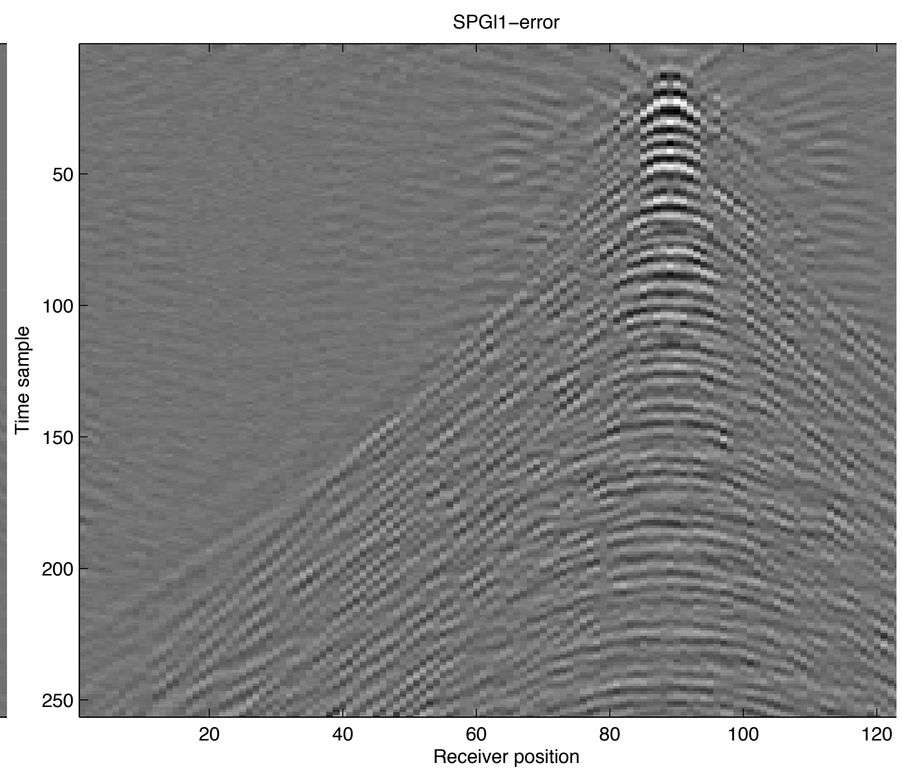
Ideal ‘Seismic’ example

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

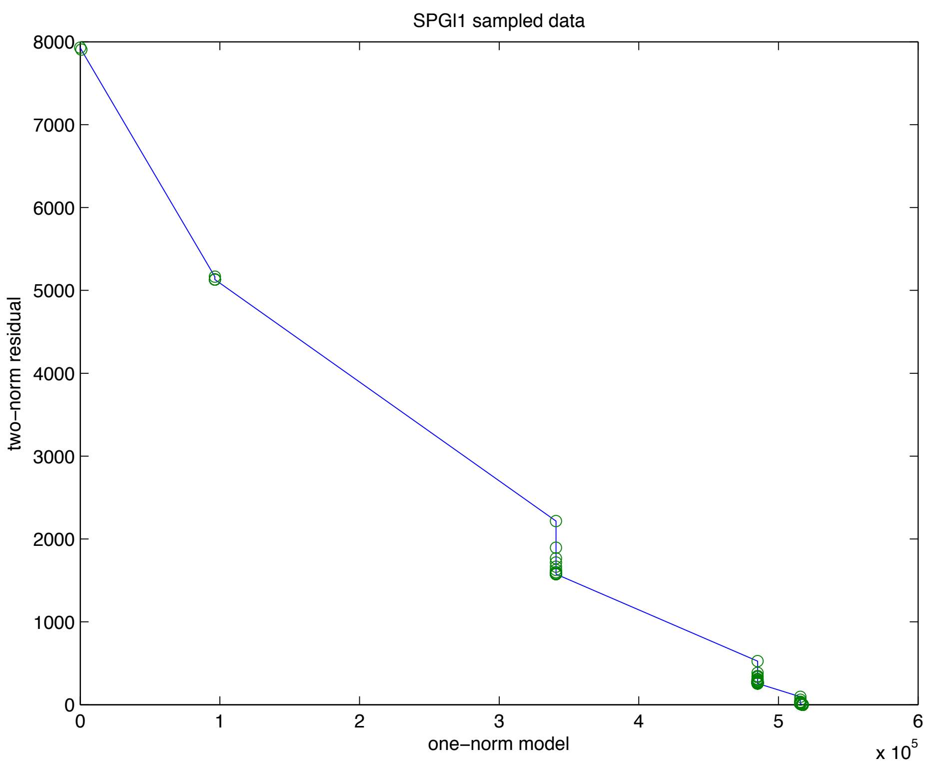
10 X



recovery



error

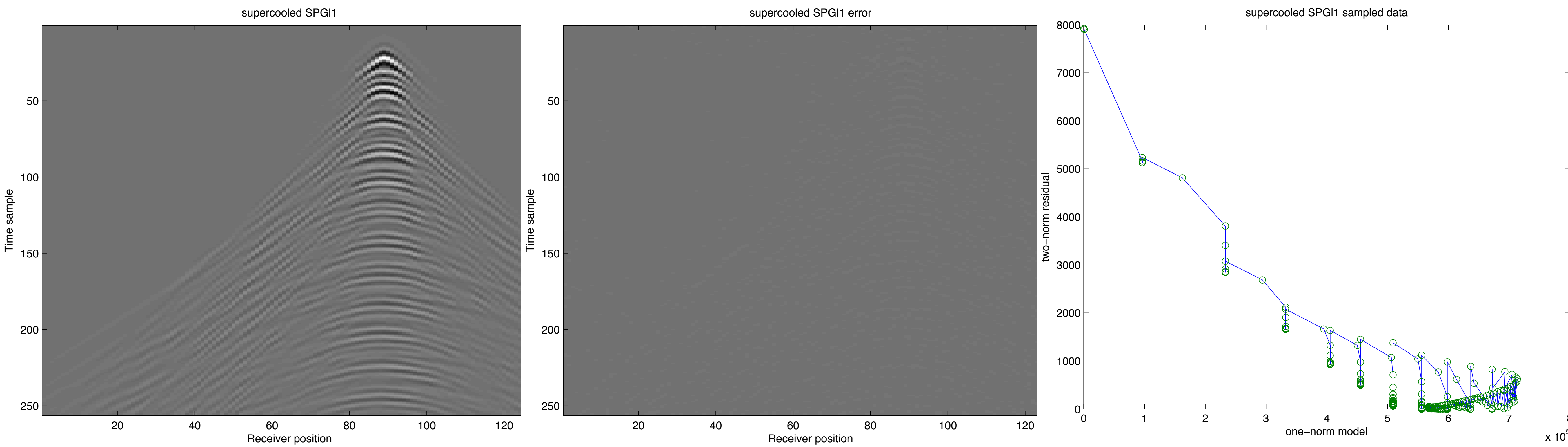


solution *path*

Cooled

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

10 X



recovery *error* solution *path*

Supercooled

[Romero et. al., 2000;]

[Montanari, 2012]

[Herrmann & Li, 2012]

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*

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)

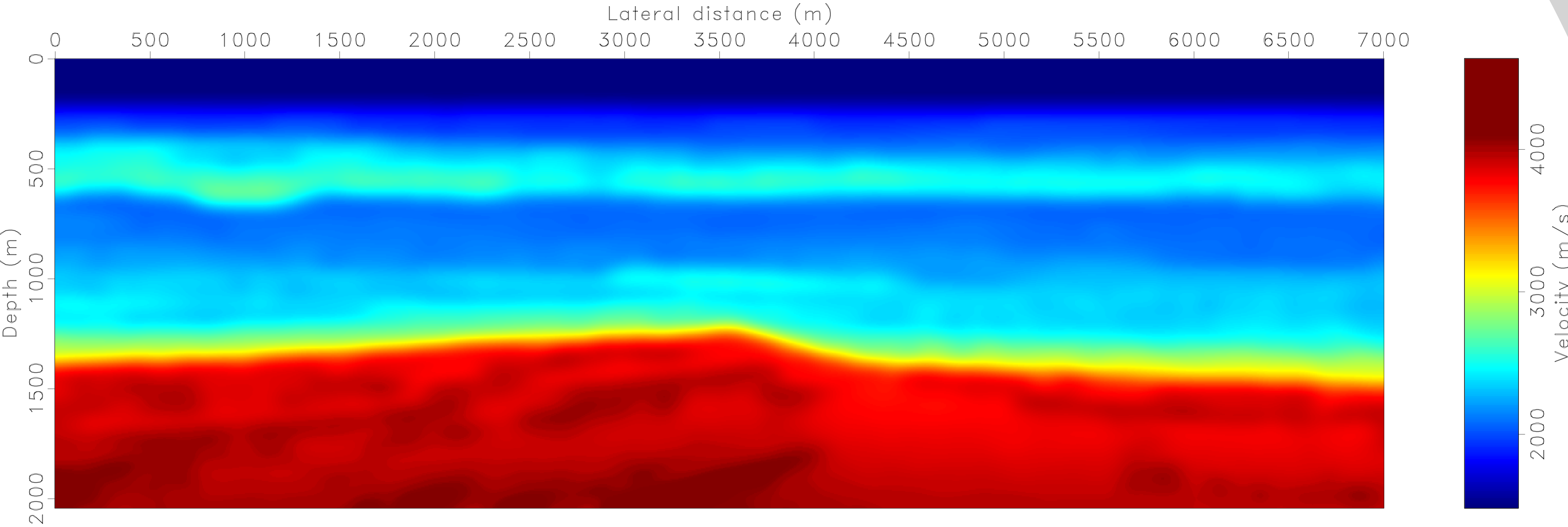
Promote *sparsity* in the *curvelet* domain

Imaging results

Time-harmonic Helmholtz:

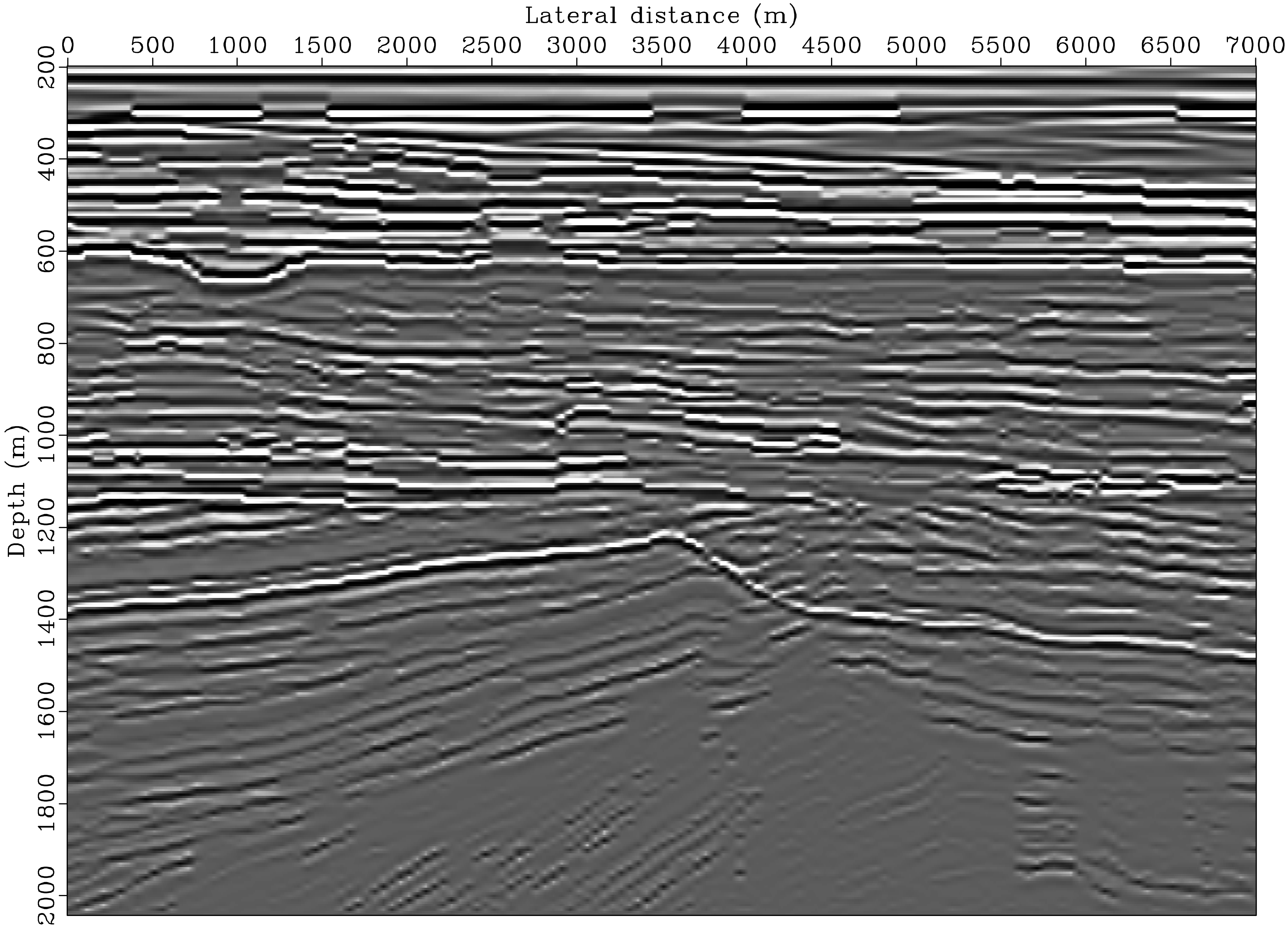
- 409 X 1401 with mesh size of 5m
- 9 point stencil [C. Jo et. al., '96]
- absorbing boundary condition with damping layer with thickness proportional to wavelength
- solve wavefields on the fly with direct solver

Imaging results
[background model]



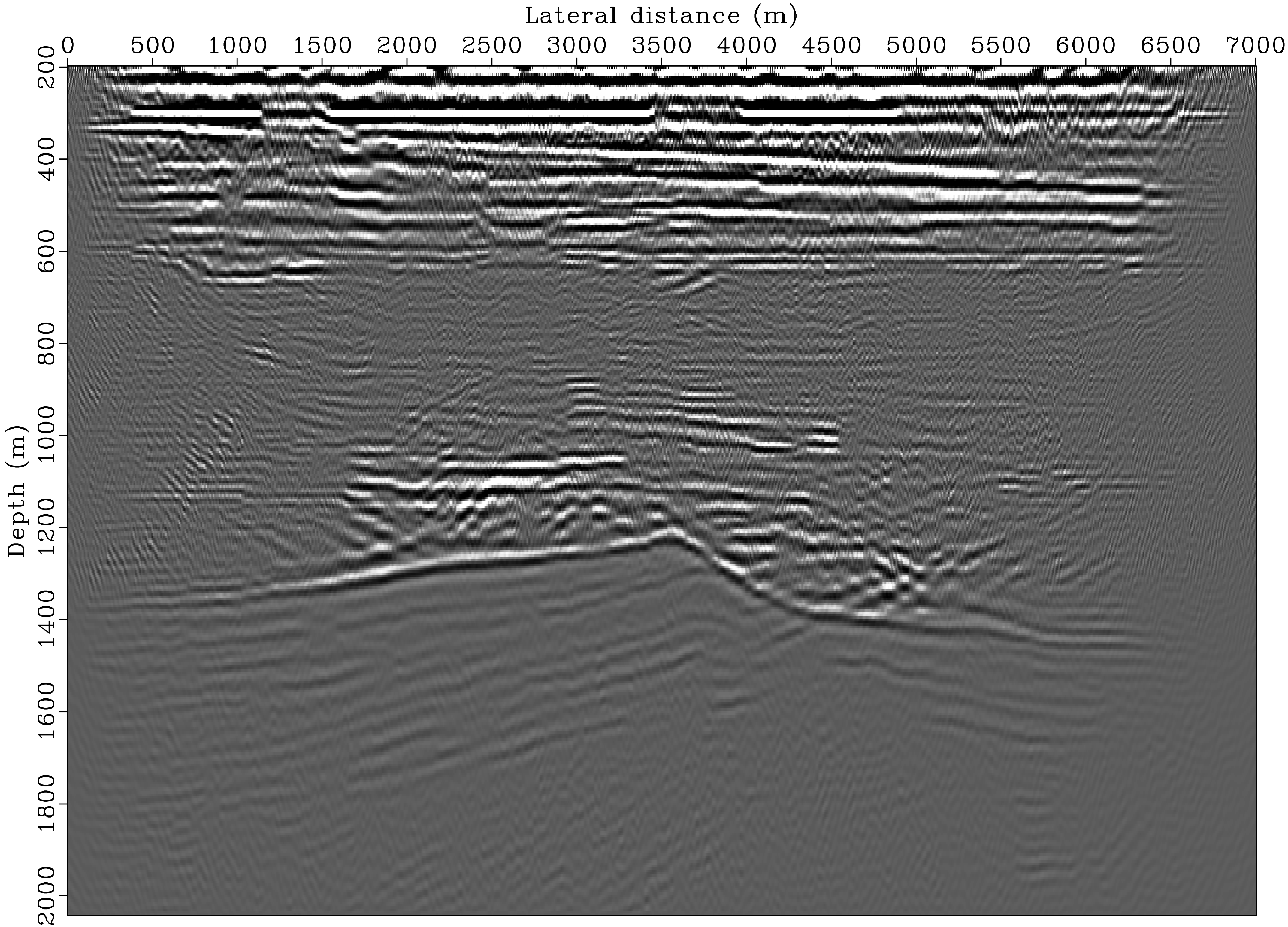
Migration results

[*true* perturbation]



Migration results

[migration with "*all*" data]



Imaging results

***Reduced* setup:**

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

Significant dimensionality reduction of

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

[Herrmann & Li, 2011]

Imaging results

Least-squares migration with *randomized supershots:*

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

$$\delta\mathbf{x} = \text{Sparse curvelet-coefficient vector}$$

$$\mathbf{S}^* = \text{Curvelet synthesis}$$

$$\underline{\mathbf{Q}} = \text{Simultaneous sources}$$

$$\delta\underline{\mathbf{d}} = \text{Super shots}$$

Imaging results

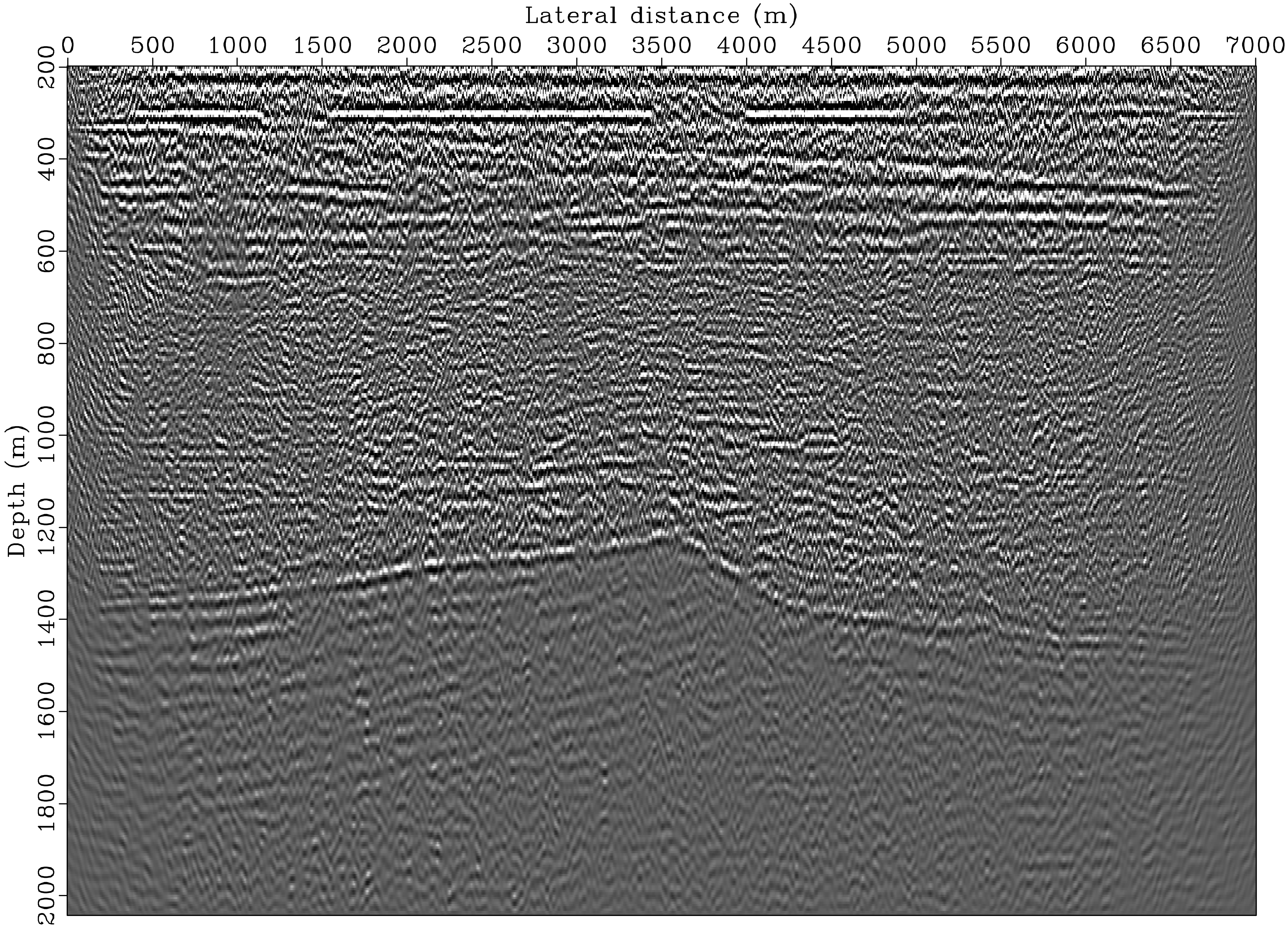
Sparsity-promoting migration with *randomized supershots:*

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

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

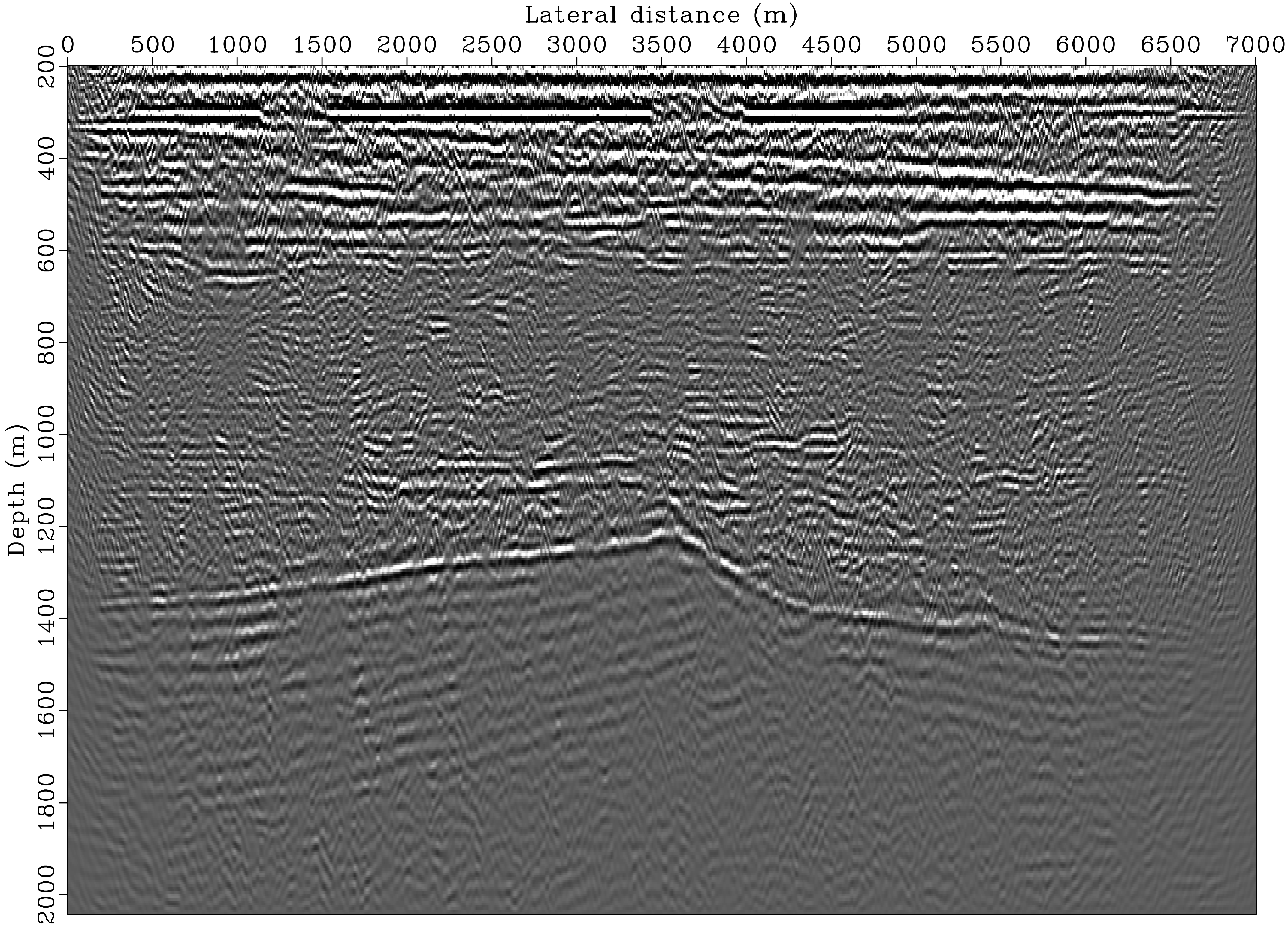
Migration results

[ℓ_2 *without* renewals]



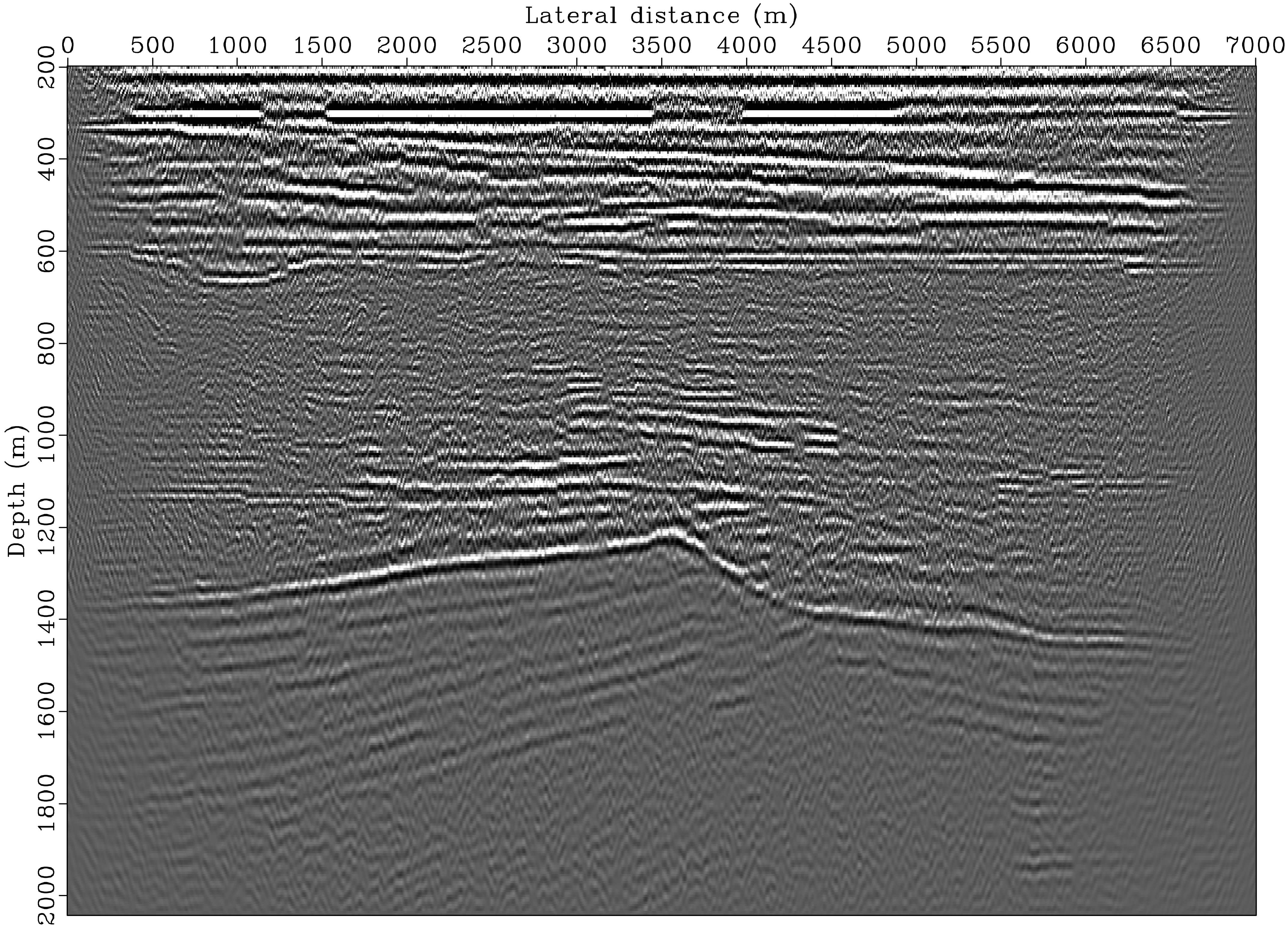
Imaging results

[ℓ_1 *without* renewals]



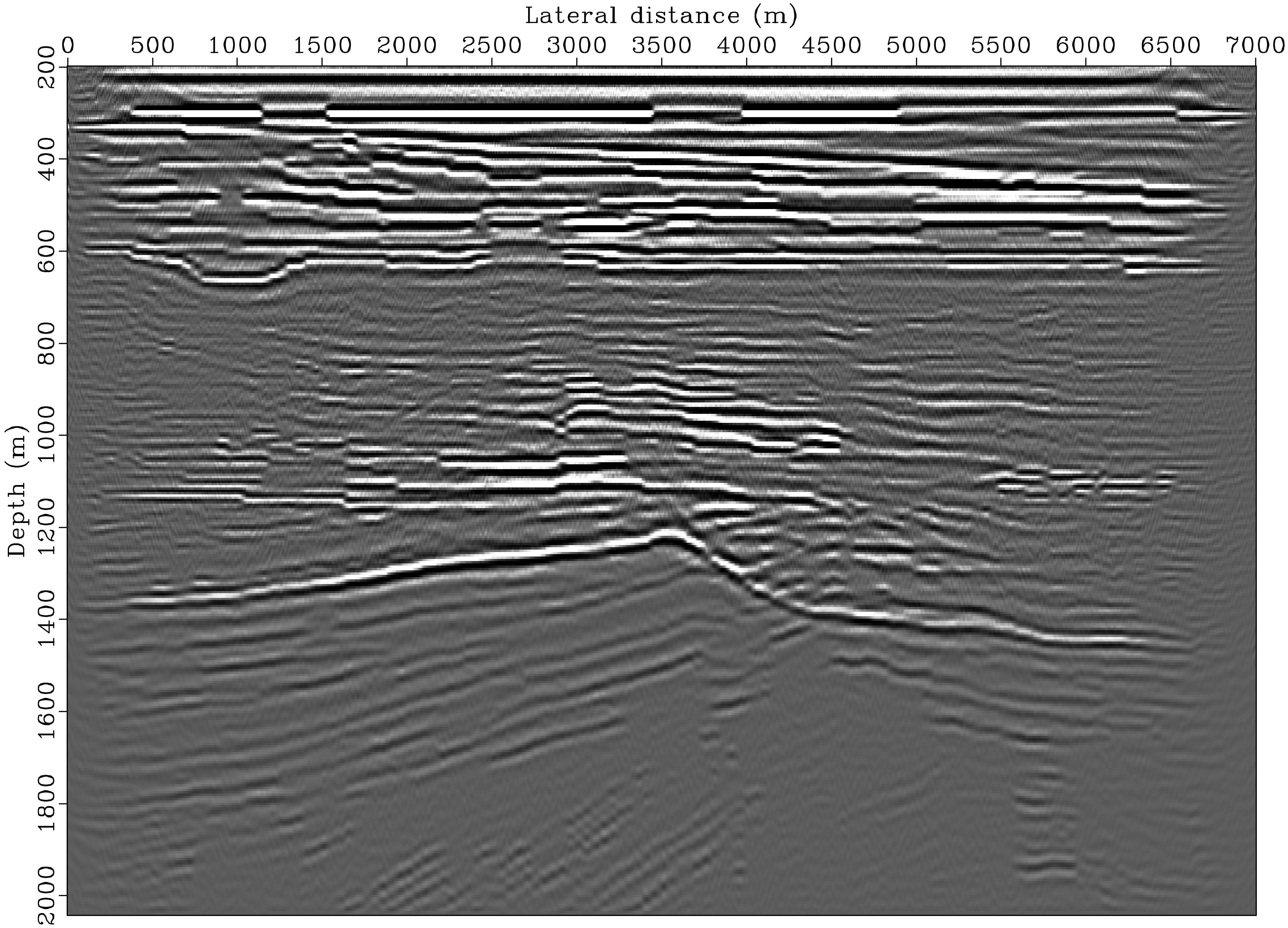
Migration results

[ℓ_2 *with* renewals]

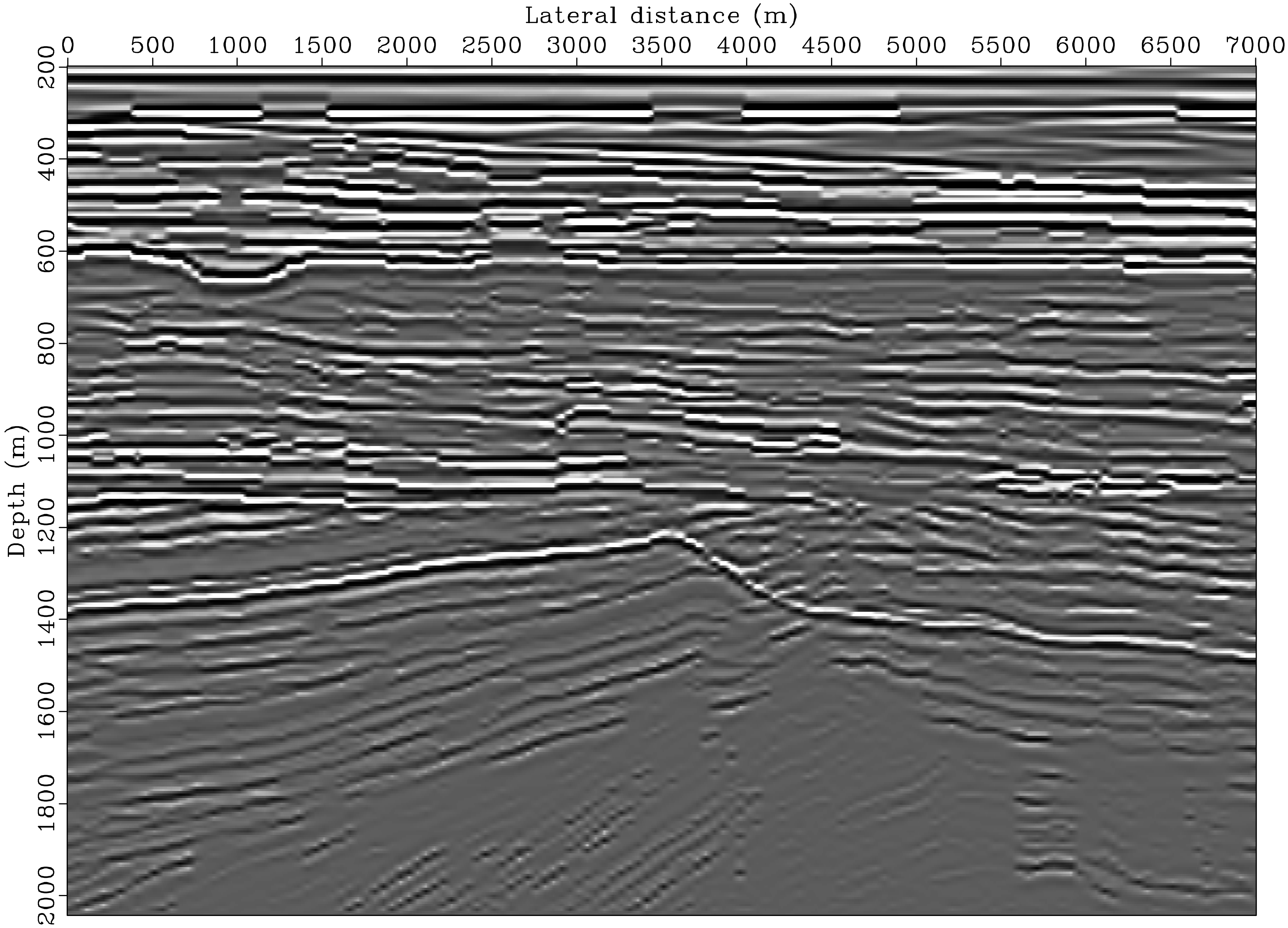


Migration results

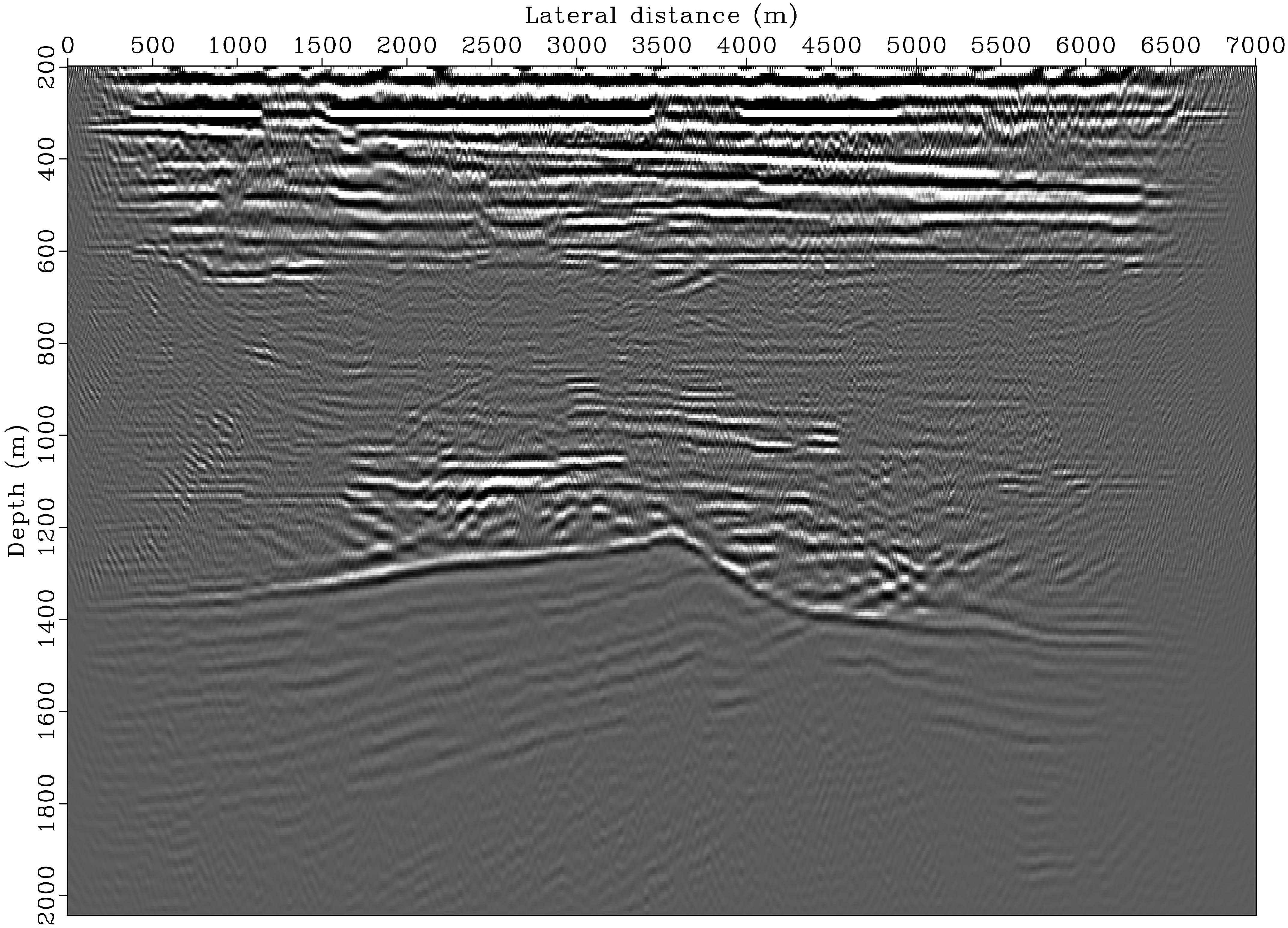
[ℓ_1 *with* renewals]



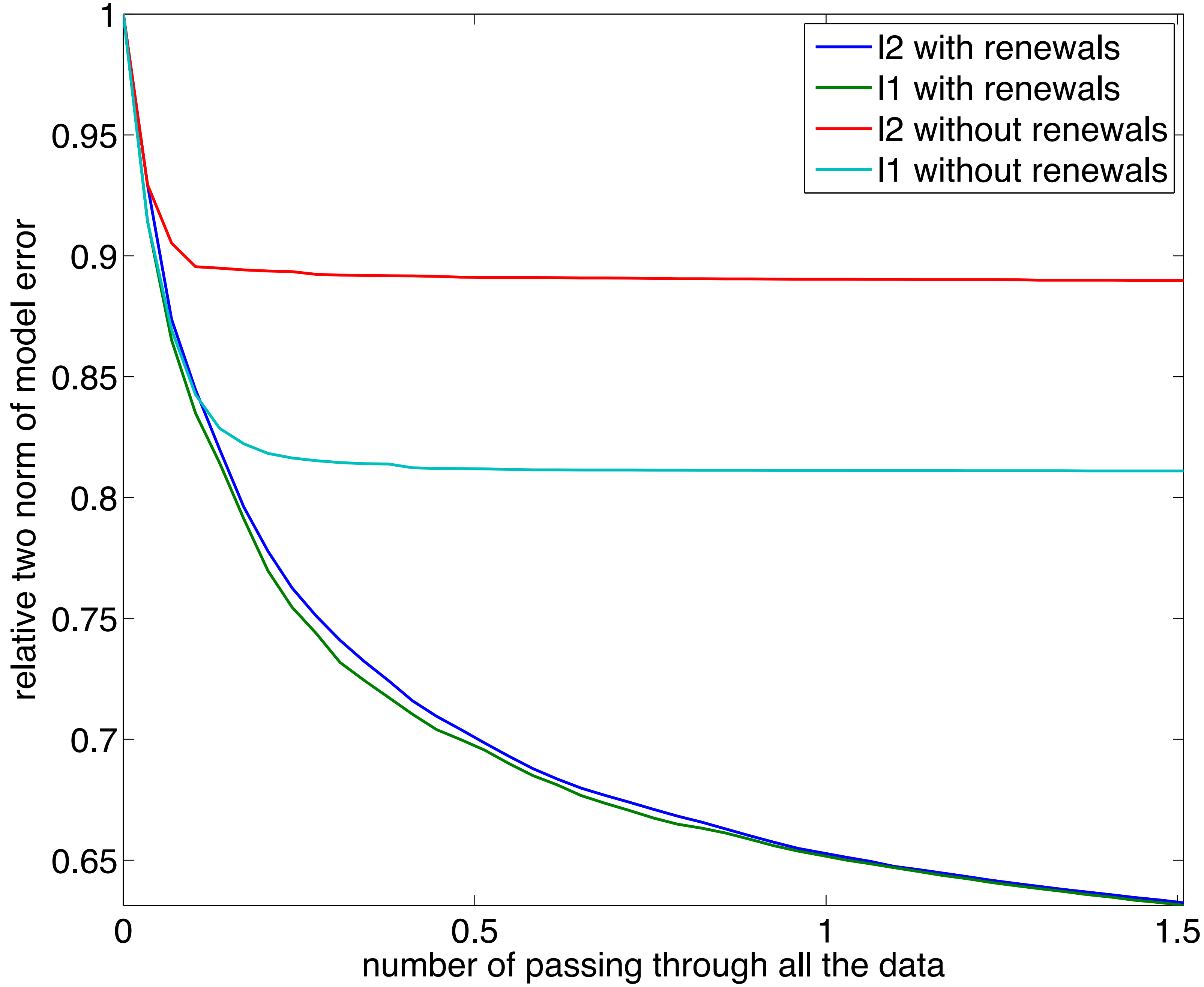
Migration results
[*true* perturbation]



Migration results
[migration with "*all*" data]



Migration results
[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*...

Acknowledgments

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development Grant DNOISE II (375142-08).

We also would like to thank the authors of CurveLab.

This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, ConocoPhillips, Petrobras, PGS, Total SA, and WesternGeco.

NSERC
CRSNG

Further reading

Simultaneous & continuous acquisition:

- *A new look at simultaneous sources* by Beasley et. al., '98.
- *Changing the mindset in seismic data acquisition* by Berkhout '08.

Simultaneous simulations, imaging, and full-wave inversion:

- *Faster shot-record depth migrations using phase encoding* by Morton & Ober, '98.
- *Phase encoding of shot records in prestack migration* by Romero et. al., '00.
- *Efficient Seismic Forward Modeling using Simultaneous Random Sources and Sparsity* by N. Neelamani et. al., '08.
- Compressive simultaneous full-waveform simulation by FJH et. al., '09.
- Randomized dimensionality reduction for full-waveform inversion by FJH & X. Li, '10
- Fast full-wavefield seismic inversion using encoded sources by Krebs et. al., '09
- *An effective method for parameter estimation with PDE constraints with multiple right hand sides. by* Eldad Haber, Matthias Chung, and Felix J. Herrmann. '10
- Seismic waveform inversion by stochastic optimization. Tristan van Leeuwen, Aleksandr Aravkin and FJH, 2010.
- Efficient least-squares imaging with sparsity promotion and compressive sensing by FJH & Li, '12
- Fast randomized full-waveform inversion with compressive sensing by Xiang Li et. al., '12

Further reading

Compressive sensing & sparse solvers

- Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information by Candes, 06.
- Compressed Sensing by D. Donoho, ’06
- Probing the Pareto frontier for basis pursuit solutions by E. van den Berg and M. Friedlander, ’08

Machine learning & message passing

- Message passing algorithms for compressed sensing by David Donoho et. al., 2009
- Graphical Models Concepts in Compressed Sensing by Andrea Montanari, ‘2012

Thank you

www.slim.eos.ubc.ca