Multifidelity conditional normalizing flows for physics-guided Bayesian inference

Ali Siahkoohi [1] Rafael Orozco [1] Gabrio Rizzuti [2] Philipp A. Witte [3]
Mathias Louboutin [1] Felix J. Herrmann [1]

[1]Georgia Institute of Technology

[2]Utrecht University

[3]Microsoft

SLIM
Georgia Institute of Technology

Problem setup

Find $\mathbf{x} \in \mathcal{X}$ such that

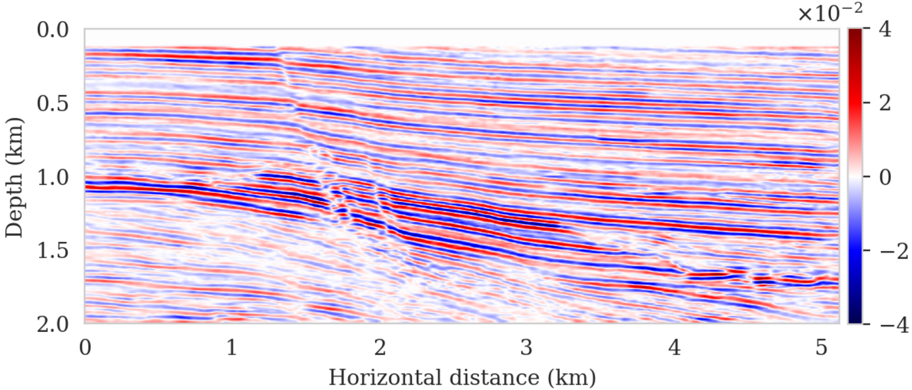
$$\mathbf{y} = F(\mathbf{x}) + \boldsymbol{\epsilon}$$

- expensive (nonlinear) forward operator F
- unknown quantity $\mathbf{x}$
- observed data $\mathbf{y} \in \mathcal{Y}$
- (non-Gaussian) noise $\boldsymbol{\epsilon}$

Bayesian inference

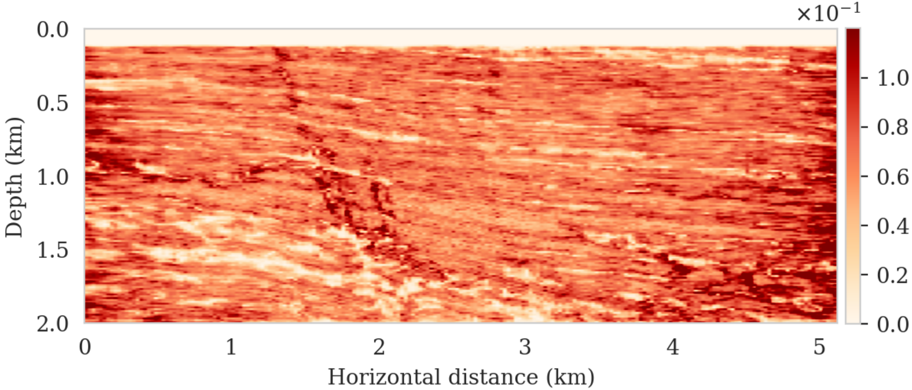
$$p_{\text{post}}\left(\mathbf{x} \mid \mathbf{y}\right) \propto p_{\text{like}}(\mathbf{y} \mid \mathbf{x})\, p_{\text{prior}}(\mathbf{x})$$

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005. ISBN: 978-0-89871-572-9. DOI: 10.1137/1.9780898717921.



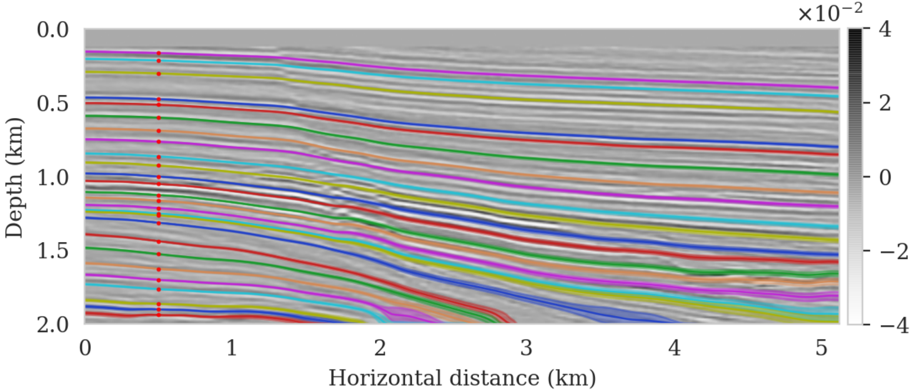
Conditional mean estimate

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. DOI: 10.1190/segam2020-3417560.1.



Pointwise standard deviation, normalized by the envelope of conditional mean

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. DOI: 10.1190/segam2020-3417560.1.



Uncertainty in horizon tracking due to uncertainties in imaging

Xinming Wu and Sergey Fomel. "Least-squares horizons with local slopes and multi-grid correlations". In: *GEOPHYSICS* 83 (4 2018), pp. IM29–IM40. DOI: 10.1190/geo2017-0830.1. URL: https://doi.org/10.1190/geo2017-0830.1.

Xinming Wu and Sergey Fomel. *xinwucwp/mhe*. 2018. URL: https://github.com/xinwucwp/mhe.

Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach". In: *SEG Technical Program Expanded Abstracts 2020*. Sept. 2020. DOI: 10.1190/segam2020-3417560.1.

Markov chain Monte Carlo (MCMC) sampling

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\alpha_k}{2} \nabla_{\mathbf{x}} \Big[\log p_{\text{like}}(\mathbf{y} \mid \mathbf{x}_k) + \log p_{\text{prior}}(\mathbf{x}_k) \Big] + \boldsymbol{\xi}_k, \quad \boldsymbol{\xi}_k \sim \mathrm{N}(\mathbf{0}, \alpha_k \mathbf{I})$$

- asymptotically exact samples

- high-dimensional integration/sampling
- costs of forward operator
- requires choosing a prior distribution

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2004. DOI: 10.1007/978-1-4757-4145-2.

Max Welling and Yee Whye Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. 2011, pp. 681–688. DOI: 10.5555/3104482.3104568.

Prior art: MCMC

Alberto Malinverno and Victoria A Briggs. “Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes”. In: *GEOPHYSICS* 69.4 (2004), pp. 1005–1016. DOI: 10.1190/1.1778243.

Alberto Malinverno and Robert L Parker. “Two ways to quantify uncertainty in geophysical inverse problems”. In: *GEOPHYSICS* 71.3 (2006), W15–W27. DOI: 10.1190/1.2194516.

James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. “A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion”. In: *SIAM Journal on Scientific Computing* 34.3 (2012), A1460–A1487. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598. URL: http://epubs.siam.org/doi/abs/10.1137/110845598.

Anandaroop Ray, Sam Kaplan, John Washbourne, and Uwe Albertin. “Low frequency full waveform seismic inversion within a tree based Bayesian framework”. In: *Geophysical Journal International* 212.1 (Oct. 2017), pp. 522–542. ISSN: 0956-540X. DOI: 10.1093/gji/ggx428. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf. URL: https://doi.org/10.1093/gji/ggx428.

Zhilong Fang, Curt Da Silva, Rachel Kuske, and Felix J. Herrmann. “Uncertainty quantification for inverse problems with weak partial-differential-equation constraints”. In: *GEOPHYSICS* 83.6 (2018), R629–R647. DOI: 10.1190/geo2017-0824.1.

Georgia K Stuart, Susan E Minkoff, and Felipe Pereira. “A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification”. In: *GEOPHYSICS* 84.6 (Nov. 2019), R1003–R1020. DOI: 10.1190/geo2018-0893.1.

Zeyu Zhao and Mrinal K Sen. “A gradient based MCMC method for FWI and uncertainty analysis”. In: *SEG Technical Program Expanded Abstracts 2019*. Aug. 2019, pp. 1465–1469. DOI: 10.1190/segam2019-3216560.1.

M Kotsi, A Malcolm, and G Ely. “Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach”. In: *Geophysical Journal International* 222.2 (May 2020), pp. 1245–1263. ISSN: 0956-540X. DOI: 10.1093/gji/ggaa245.

Andrew Curtis and Anthony Lomax. “Prior information, sampling distributions, and the curse of dimensionality”. In: *Geophysics* 66.2 (2001), pp. 372–378.

Felix J. Herrmann, Ali Siahkoohi, and Gabrio Rizzuti. “Learned imaging with constraints and uncertainty quantification”. In: *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*. Dec. 2019. URL: https://arxiv.org/pdf/1909.06473.pdf.

Variational Inference

Approximate target density $p_{\text{post}}\left(\mathbf{x} \mid \mathbf{y}\right)$ via parametric density $p_{\theta}\left(\mathbf{x} \mid \mathbf{y}\right)$

$$\forall\, \mathbf{x}, \mathbf{y} : \quad p_{\theta}\left(\mathbf{x} \mid \mathbf{y}\right) \approx p_{\text{post}}\left(\mathbf{x} \mid \mathbf{y}\right)$$

- admits stochastic optimization
- cheap sampling after optimization
- known to scale better than MCMC

- requires access to $\mathbf{x} \sim p_{\text{prior}}(\mathbf{x})$ or $p_{\text{prior}}(\mathbf{x})$

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Posterior inference with Normalizing Flows (NFs)

sampling the posterior distribution directly via NFs

Purely data-driven approach

$$\min_{\theta} \mathbb{E}_{\mathbf{y},\mathbf{x}\sim p(\mathbf{y},\mathbf{x})} \left[\frac{1}{2} \left\| G_\theta(\mathbf{y},\mathbf{x}) \right\|^2 - \log \left| \det \nabla_{y,x} G_\theta(\mathbf{y},\mathbf{x}) \right| \right]$$

$$G_\theta(\mathbf{y},\mathbf{x}) = \begin{bmatrix} G_{\theta_y}(\mathbf{y}) \\ G_{\theta_x}(\mathbf{y},\mathbf{x}) \end{bmatrix}, \ \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_y \\ \boldsymbol{\theta}_x \end{bmatrix}$$

conditional NF, $G_\theta : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x$

expectation estimated via training pairs, $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^n \sim p_{y,x}(\mathbf{y},\mathbf{x})$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Ali Siahkoohi and Felix J. Herrmann. "Learning by example: fast reliability-aware seismic imaging with normalizing flows". Apr. 2021. URL: https://arxiv.org/pdf/2104.06255.pdf.

Inference with the data-driven approach

Sampling from $p(\mathbf{x} \mid \mathbf{y})$

$$G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z}) \sim p\left(\mathbf{x} \mid \mathbf{y}\right), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

Posterior density estimation

$$p_\theta(\mathbf{x} \mid \mathbf{y}) = p_{\mathbf{z}}\big(G_{\theta_x}(\mathbf{y}, \mathbf{x})\big) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right|$$

Training dataset for the data-driven approach

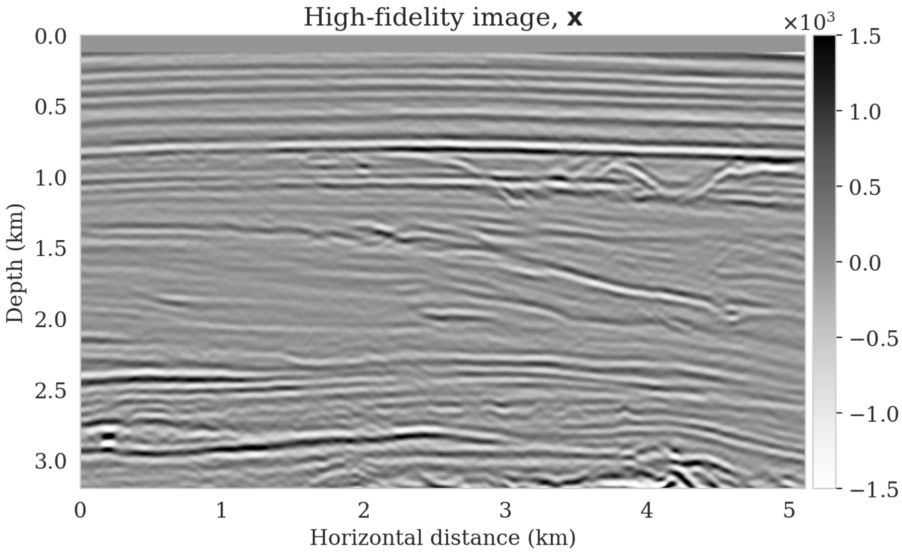
training pairs

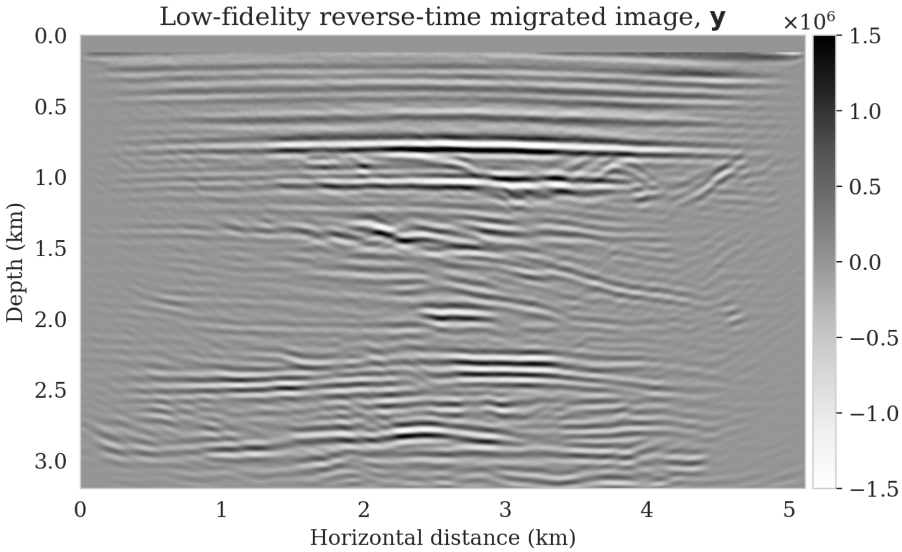
$$\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim p(\mathbf{y}, \mathbf{x})$$

where

$$(\mathbf{y}_i, \mathbf{x}_i) = \left(\mathbf{J}^\top (\mathbf{J}\delta\mathbf{m}_i + \boldsymbol{\epsilon}_i), \delta\mathbf{m}_i\right)$$

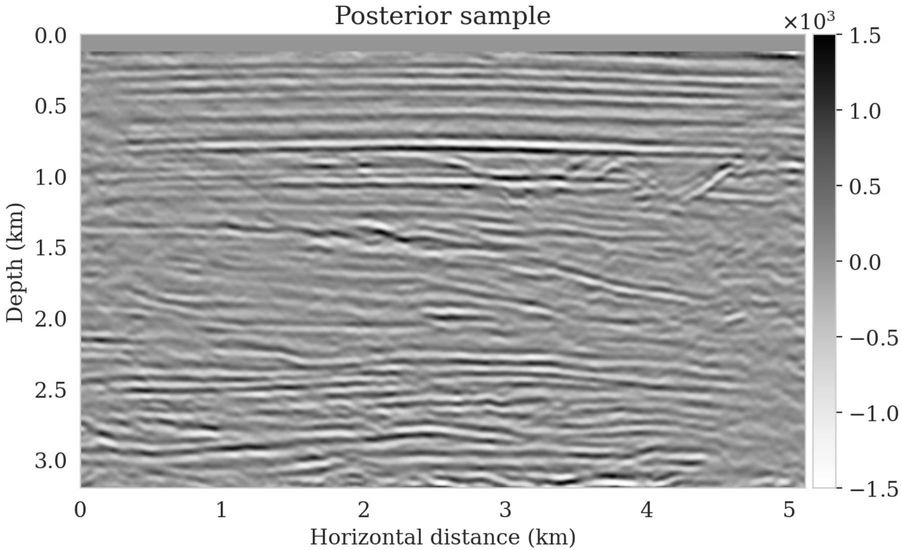
WesternGeco. *Parihaka 3D PSTM Final Processing Report*. Tech. rep. New Zealand Petroleum Report 4582. 2012. URL: `https://wiki.seg.org/wiki/Parihaka-3D`.

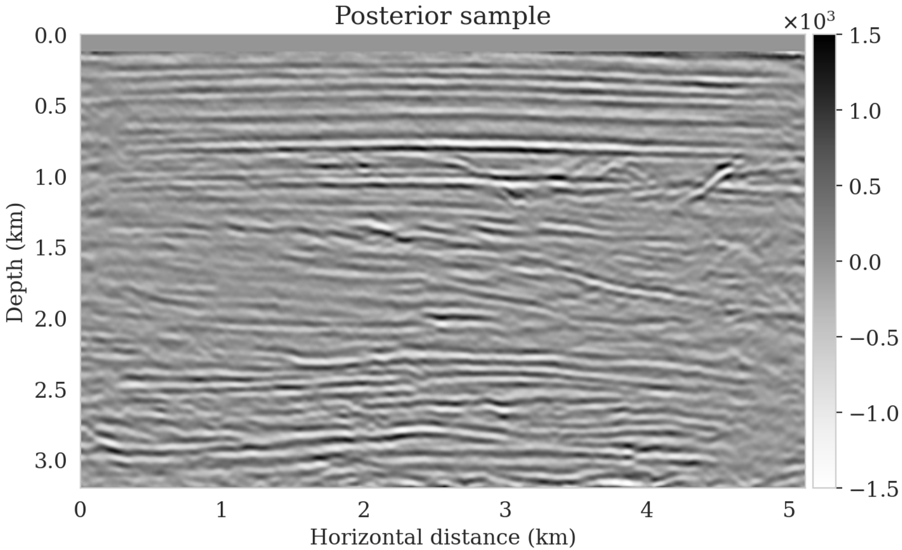

High-fidelity image, **x**
Depth (km)
Horizontal distance (km)
×10³
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

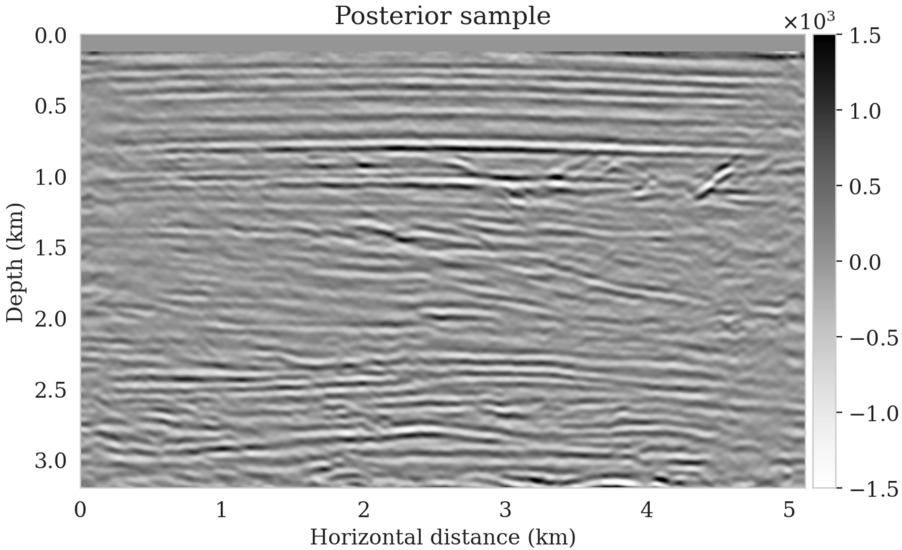

Low-fidelity reverse-time migrated image, **y**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^6
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

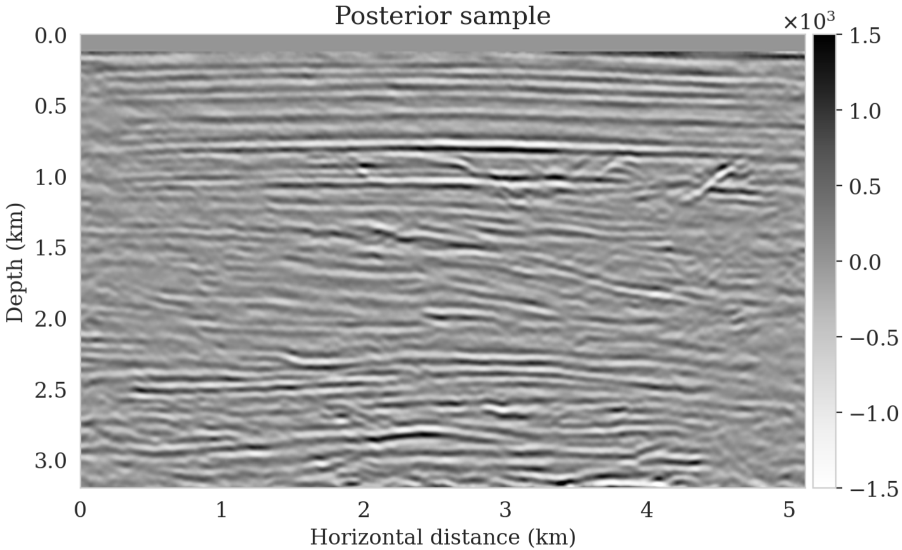
Samples from the posterior

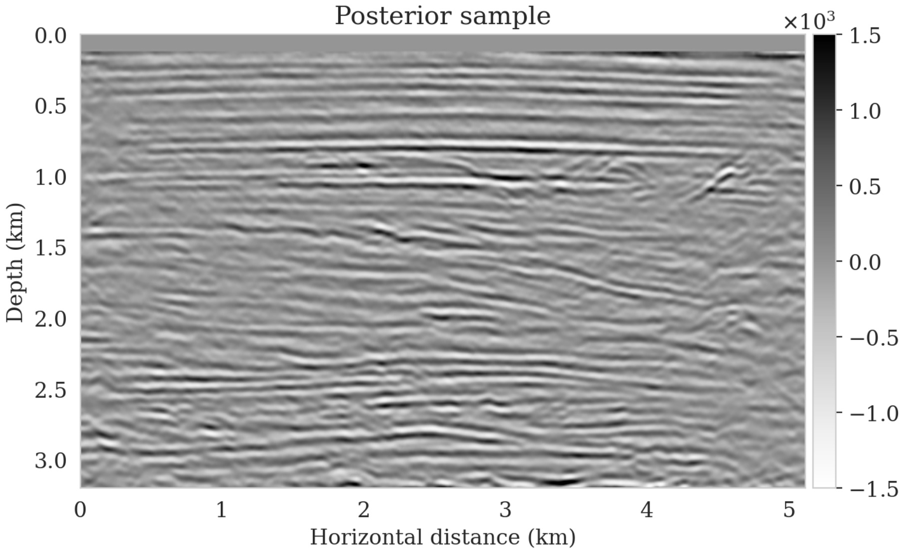
$$G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z}) \sim p\left(\mathbf{x} \mid \mathbf{y}\right), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$


Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

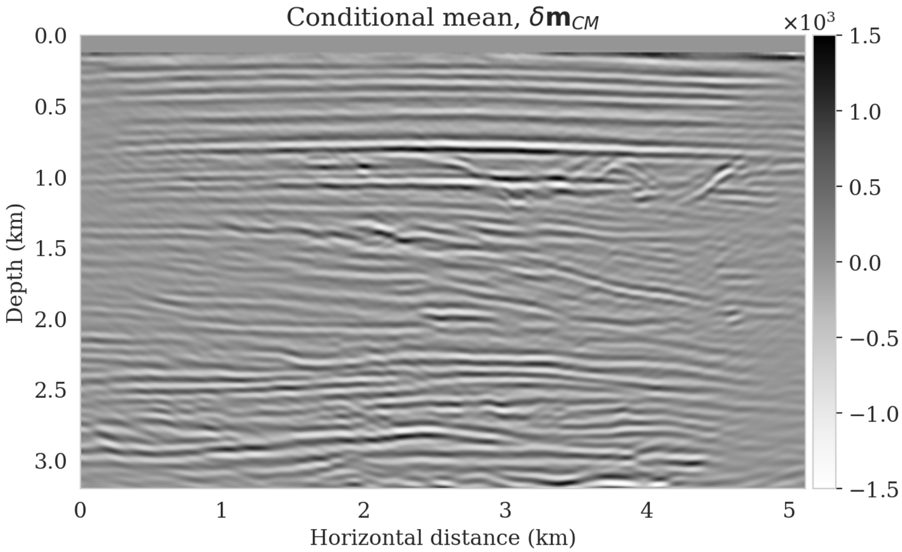

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5


Posterior sample
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

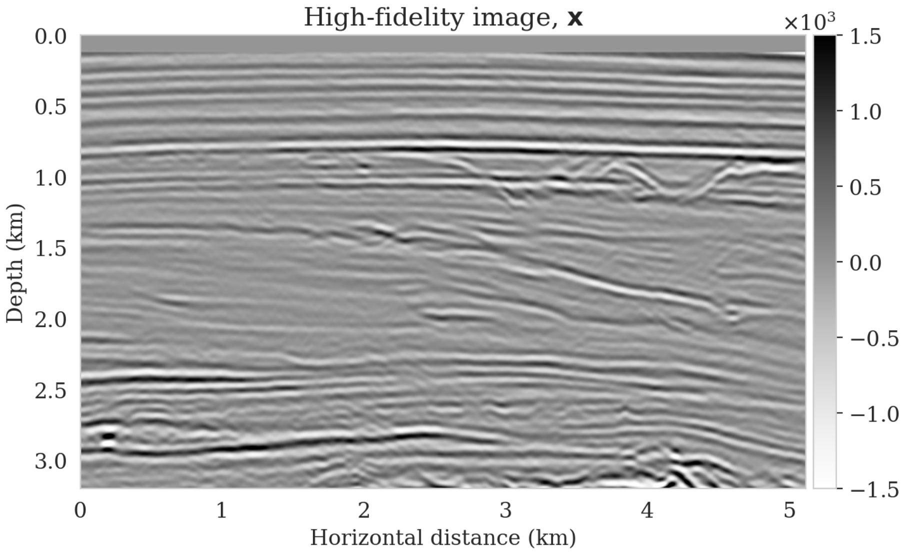

Posterior sample
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

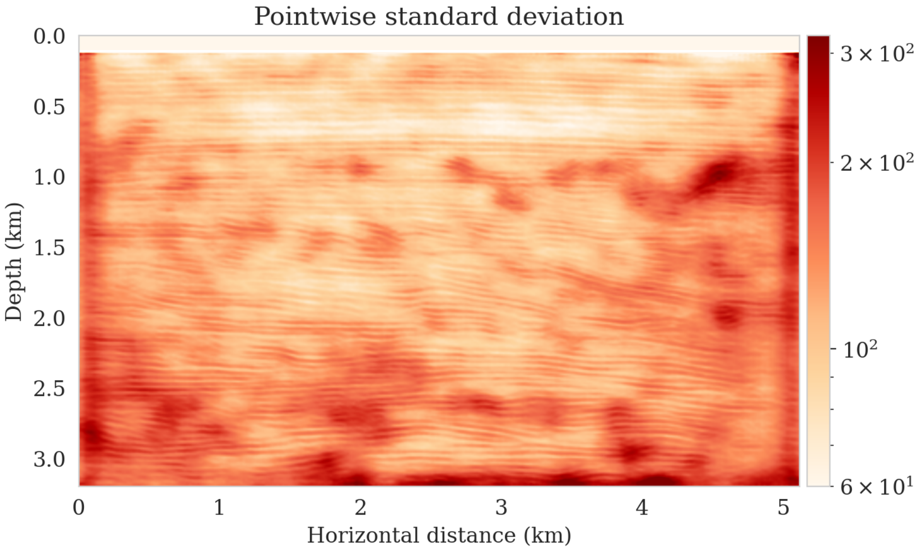
Conditional mean estimate

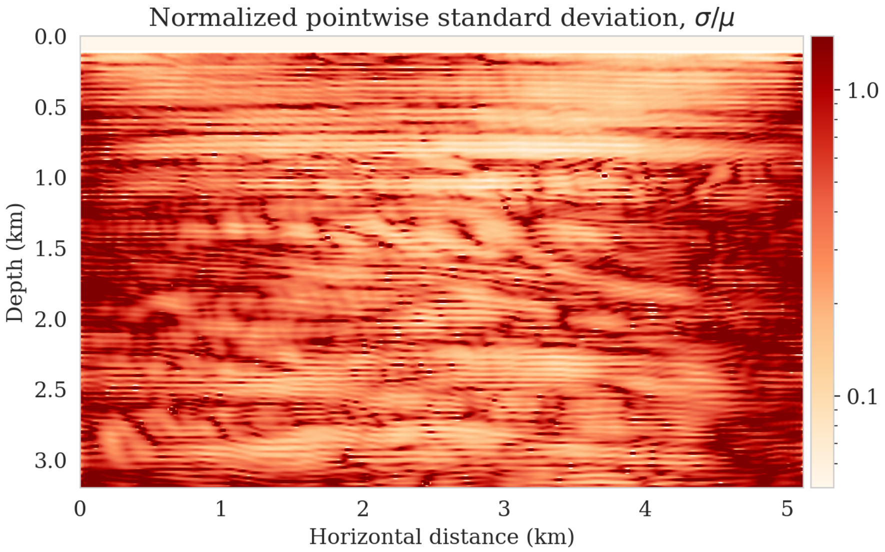
$$\mathbb{E}[\mathbf{x} \mid \mathbf{y}] = \int \mathbf{x}\, p(\mathbf{x} \mid \mathbf{y})\, \mathrm{d}\mathbf{x}$$



Conditional mean, $\delta \mathbf{m}_{CM}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)


High-fidelity image, **x**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Pointwise standard deviation
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
3×10^2
2×10^2
10^2
6×10^1


Normalized pointwise standard deviation, σ/μ
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.0
0.1

Purely data-driven approach

learns the prior distribution from training data

samples from the posterior virtually for free in test time

not specific to one observation $\mathbf{y}$

needs supervised pairs of model and data

heavily relies on the training data to generalize

not directly tied to physics/data

Purely physics-based approach

$$\min_{\phi} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})}\left[\frac{1}{2\sigma^2}\left\|F\big(T_\phi(\mathbf{z})\big)-\mathbf{y}\right\|_2^2-\log p\big(T_\phi(\mathbf{z})\big)-\log\left|\det\nabla_z T_\phi(\mathbf{z})\right|\right]$$

NF, $T_\phi : \mathcal{Z} \to \mathcal{X}$

prior distribution, $p(\mathbf{x})$

specific to one observation, $\mathbf{y}$

Gabrio Rizzuti, Ali Siahkoohi, Philipp A. Witte, and Felix J. Herrmann. "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization". In: *SEG Technical Program Expanded Abstracts*. Sept. 2020, pp. 1541–1545. DOI: 10.1190/segam2020-3428150.1.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. "Bayesian Seismic Tomography using Normalizing Flows". In: (2020). DOI: 10.31223/X53K6G. URL: https://eartharxiv.org/repository/view/1940/.

Inference with the physics-based approach

sampling from $p(\mathbf{x} \mid \mathbf{y})$

$$T_\phi(\mathbf{z}) \sim p\left(\mathbf{x} \mid \mathbf{y}\right), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}),$$

posterior density estimation,

$$p_\phi(\mathbf{x} \mid \mathbf{y}) = p_{\mathbf{z}}\left(T_\phi^{-1}(\mathbf{x})\right) \left| \det \nabla_x T_\phi^{-1}(\mathbf{x}) \right|$$

Purely physics-based approach

tied to the physics and data

no training data needed

requires choosing a prior distribution $p(\mathbf{x})$

repeated evaluations of F and $\nabla F^{\top}$

specific to one observation, $\mathbf{y}$

Multi-fidelity preconditioned scheme

exploit the information in the pretrained conditional NF (data-driven approach)

tie the results to physics and data

Multi-fidelity preconditioned scheme

Use the density encoded by the pretrained conditional NF as a prior

$$p_{\text{prior}}(\mathbf{x}) := p_{\mathbf{z}}\big(G_{\theta_x}(\mathbf{y}, \mathbf{x})\big) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right|$$

allows for using a data-driven prior

removes the bias introduced by hand-crafted priors

can be used as a prior density in inverse problems

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Multi-fidelity preconditioned scheme

Use transfer learning and initialize $T : \mathcal{Z} \to \mathcal{X}$ by

$$T_{\theta_x}(\mathbf{z}) := G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z})$$

can significantly reduce ($5\times$) the cost of the purely physics-based approach

can be used as an implicit prior by reparameterizing the unknown with T

Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. "Invertible generative models for inverse problems: mitigating representation error and dataset bias". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 399–409. URL: http://proceedings.mlr.press/v119/asim20a.html.

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

Preconditioned MAP estimation

$$\min_{\mathbf{z}} \frac{1}{2\sigma^2} \left\| F\left(T_{\theta_x}(\mathbf{z})\right) - \mathbf{y} \right\|_2^2 + \frac{1}{2} \left\| \mathbf{z} \right\|_2^2$$

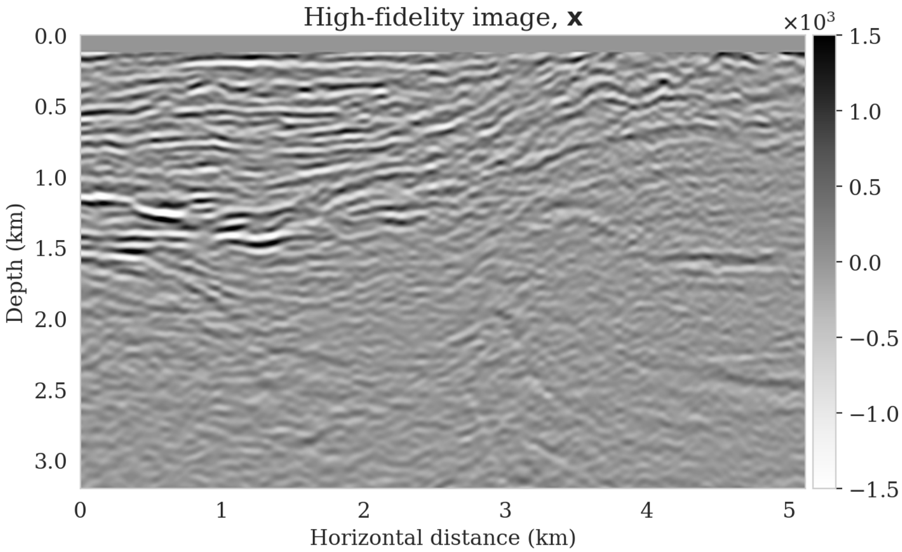
initializing $\mathbf{z} = \mathbf{0}$ acts as an implicit prior

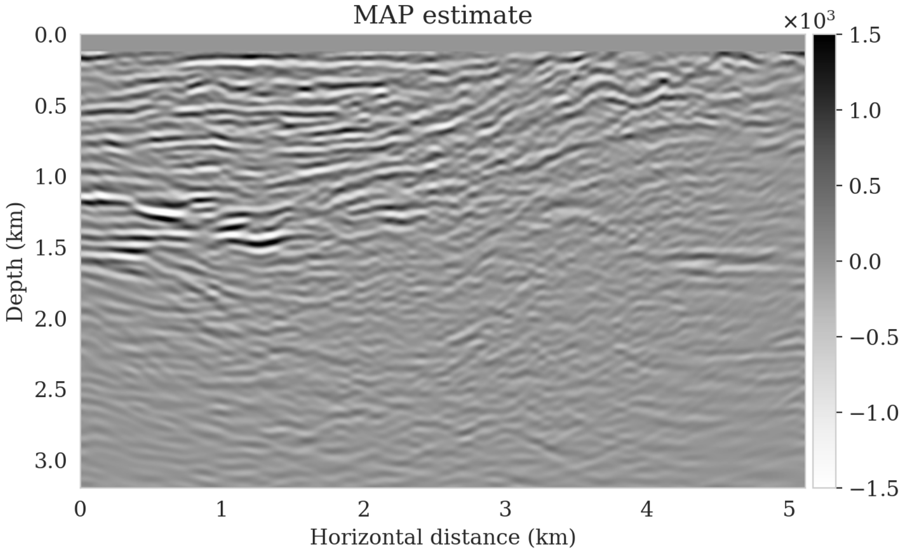
since T is invertible, in principle it can represent any unknown $\mathbf{x} \in \mathcal{X}$

can be used if $\mathbf{y}$ is out-of-distribution

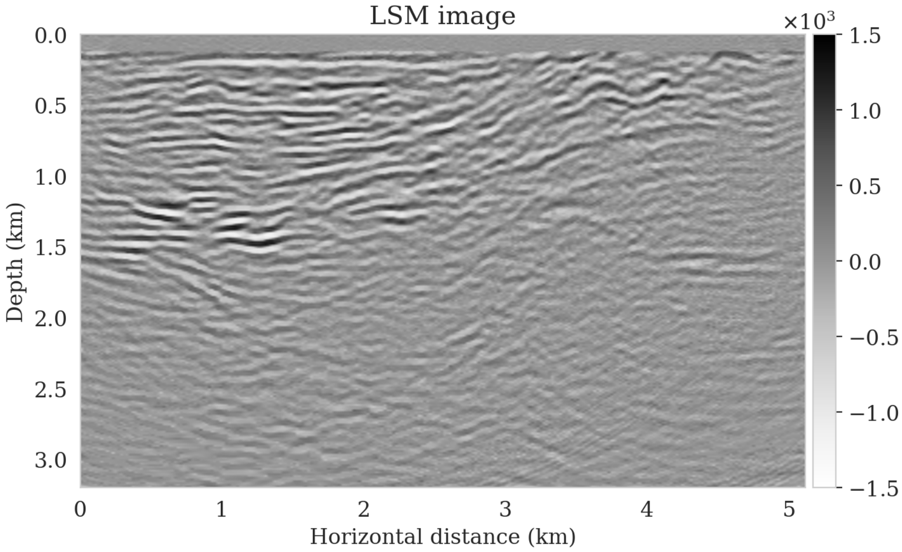
final estimate is tied to the physics/data

Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. "Invertible generative models for inverse problems: mitigating representation error and dataset bias". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 399–409. URL: http://proceedings.mlr.press/v119/asim20a.html.


High-fidelity image, **x**
$\times 10^3$
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



MAP estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


LSM image
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Signal-to-noise ratio comparison

	SNR (dB)
RTM	-54.83
MLE	7.40
Preconditioned MAP	11.29
Data-driven CM	4.34

SNR comparison

Preconditioned physics-based approach

$$\min_{\theta_x} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma^2} \left\| F\big(T_{\theta_x}(\mathbf{z})\big) - \mathbf{y} \right\|_2^2 - \log p_{\text{prior}}\big(T_{\theta_x}(\mathbf{z})\big) - \log \left| \det \nabla_z T_{\theta_x}(\mathbf{z}) \right| \right]$$

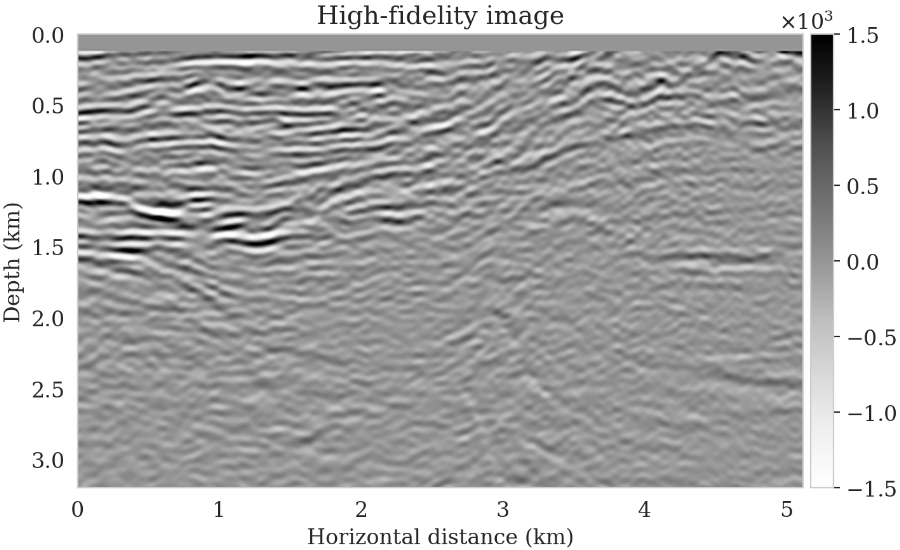
transfer learning $T_{\theta_x}(\mathbf{z}) := G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z})$

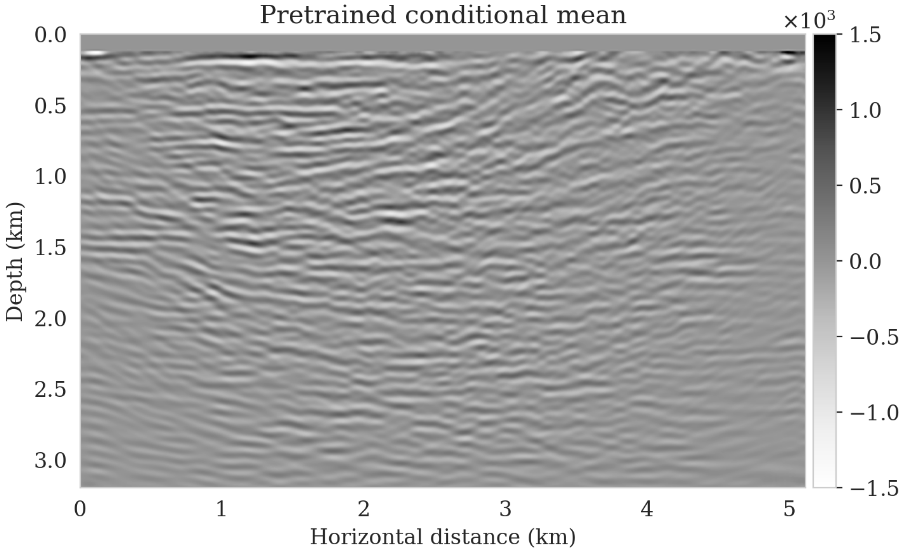
- corrects for errors due to out-of-distribution data
- less evaluations of F and $\nabla F^\top$

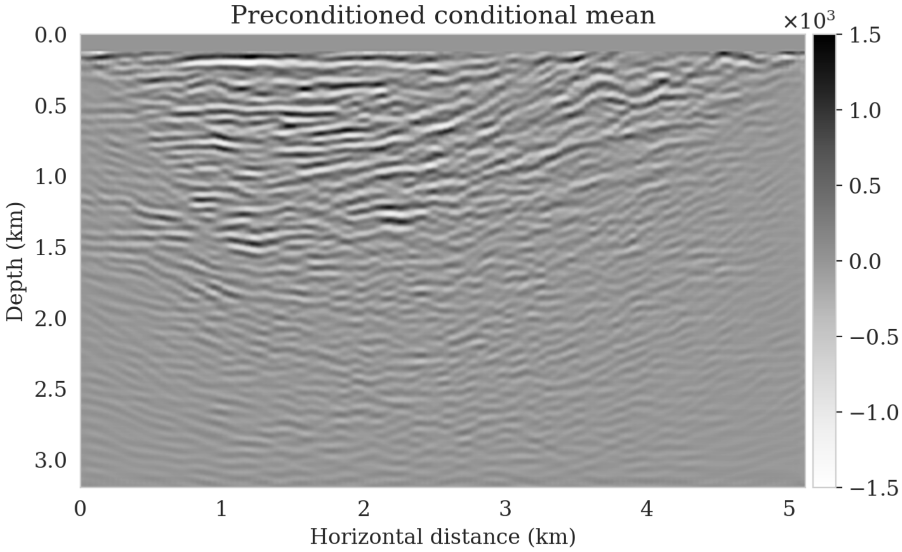
learned prior density, $p_{\text{prior}}(\mathbf{x}) := p_{\mathbf{z}}\big(G_{\theta_x}(\mathbf{y}, \mathbf{x})\big) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right|$

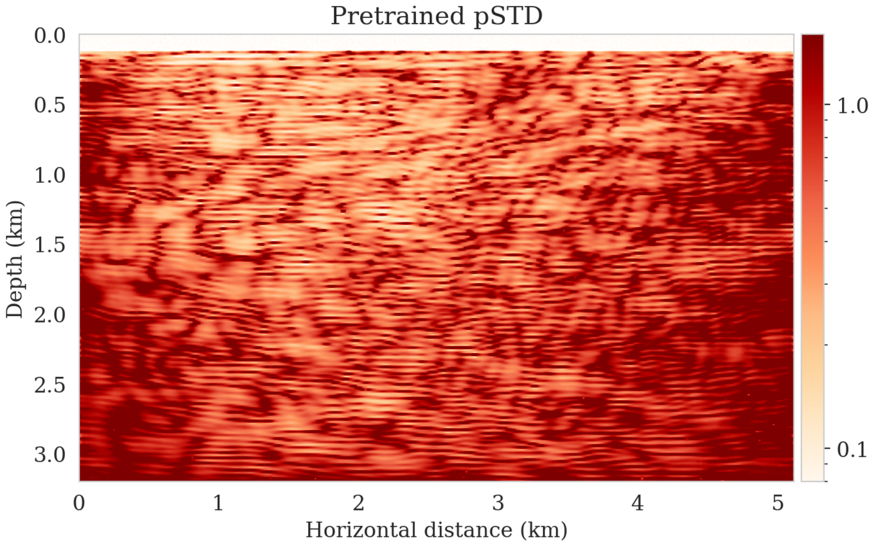
fast to adapt to new observation

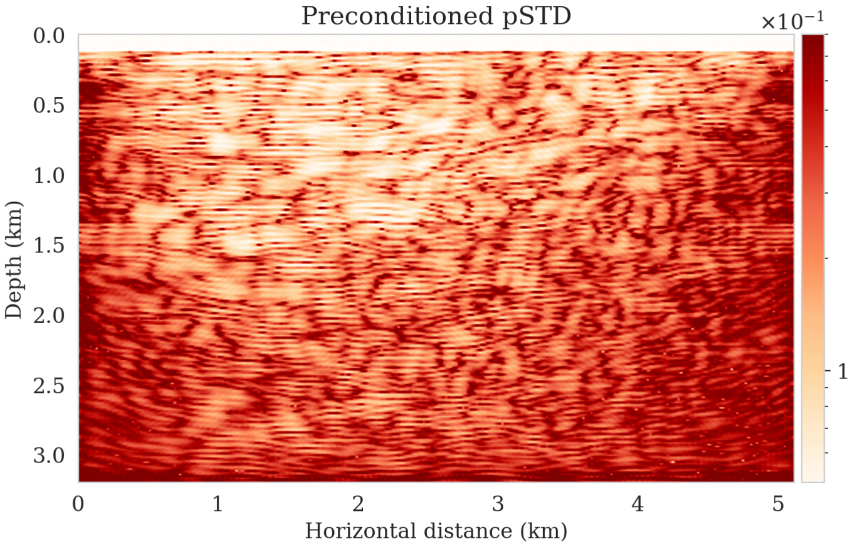
Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.


High-fidelity image
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Pretrained conditional mean
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Preconditioned conditional mean
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Pretrained pSTD
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.0
0.1



Preconditioned pSTD
×10⁻¹
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1

Signal-to-noise ratio comparison

	SNR (dB)
Data-driven CM	4.34
Preconditioned CM	4.76

SNR comparison

Number of PDE solves

	Number of PDE solves	
	F	$\nabla F^\top$
RTM	102	102
MLE	204	204
Preconditioned MAP	510	510
Preconditioned VI approach	$2,000$	$2,000$
MCMC w/ deep prior	$10,000$	$10,000$

Comparison of number of PDE solves in the seismic imaging example

Related work

Purely physics-based approach NF-based Bayesian inference

show orders of magnitude speed up compared to traditional MCMC methods

need to train the NF from scratch for a new observation $\mathbf{y}$

use handcrated priors, often negatively bias the inversion

Gabrio Rizzuti, Ali Siahkoohi, Philipp A. Witte, and Felix J. Herrmann. "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization". In: *SEG Technical Program Expanded Abstracts*. Sept. 2020, pp. 1541–1545. DOI: 10.1190/segam2020-3428150.1.

Xin Zhang and Andrew Curtis. "Seismic tomography using variational inference methods". In: *Journal of Geophysical Research: Solid Earth* 125.4 (2020), e2019JB018589.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. "Bayesian Seismic Tomography using Normalizing Flows". In: (2020). DOI: 10.31223/X53K6G. URL: https://eartharxiv.org/repository/view/1940/.

Related work

Data-driven approaches for directly sampling from the posterior distribution

fast Bayesian inference given new observation $\mathbf{y}$

sample the posterior virtually for free

not directly tied to the physics/data

not reliable when applied to out-of-distribution data

not trivial to scale GAN-based approaches to large-scale problems

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Jonas Adler and Ozan Öktem. "Deep Bayesian Inversion". In: *arXiv preprint arXiv:1811.05910* (2018).

Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. *Conditional Sampling With Monotone GANs*. 2021. arXiv: 2006.06755 [stat.ML].

Jan-Hinrich Nölke et al. "Invertible neural networks for uncertainty quantification in photoacoustic imaging". In: *arXiv preprint arXiv:2011.05110* (2020).

Related work

Injective network for inverse problems and uncertainty quantification

use an injective map to map data to a low-dimensional space

can be used as a prior (projection operator) in inverse problems

heavily relies on availability of training data to generalize

unclear how the learned projection operator behaves when applied to out-of-distribution data

for new observation $\mathbf{y}$, Bayesian inference requires training an additional NF, involving the forward operator

Konik Kothari, AmirEhsan Khorashadizadeh, Maarten de Hoop, and Ivan Dokmanić. *Trumpets: Injective Flows for Inference and Inverse Problems*. 2021. arXiv: 2102.10461 [cs.LG].

Contributions

Take full advantage of existing training data to provide

low-fidelity but fast conditional mean estimate

a first assessment of the image's reliability

preconditioned physics-based high-fidelity MAP estimate via the learned prior

preconditioned, scalable, and physics-based Bayesian inference

Conclusions

Obtaining UQ information is rendered impractical when

- the forward operators are expensive to evaluate
- the problem is high dimensional

There are strong indications that

- Bayesian inference with normalizing flows can lead to orders of magnitude computational improvements compared to MCMC methods
- preconditioning with a pretrained conditional normalizing flow can lead to another order of magnitude speed up

Future work

- characterize uncertainties due to modeling errors

Code

https://github.com/slimgroup/InvertibleNetworks.jl

https://github.com/slimgroup/FastApproximateInference.jl

https://github.com/slimgroup/Software.SEG2021

Acknowledgment

This research was carried out with the support of

Georgia Research Alliance and partners of the ML4Seismic Center

Variational Inference

Approximate distribution p by minimizing the Kullback-Leibner (KL) divergence

$$\mathbb{D}_{\mathsf{KL}}\left(p \mid\mid p_\theta\right) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{p_\theta(\mathbf{x})}\, \mathbf{dx}$$

$$= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\big[- \log p_\theta(\mathbf{x}) + \log p(\mathbf{x})\big],$$

parametric distribution p_θ

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Variational Inference

Objective of variational inference is to solve

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{D}_{\mathsf{KL}}\left(p \,||\, p_\theta\right)$$

p_θ designed to be easily sampled

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (1999), pp. 183–233. DOI: 10.1023/A:1007665907178.

Generative models

probabilistic models to characterize an unknown distribution

Generative models

$$p_\theta(\mathbf{x}) \approx p(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}$$

high dimensional unknown density, $p(\mathbf{x})$

generative model, $p_\theta(\mathbf{x})$, parameterized by $\boldsymbol{\theta}$

$\boldsymbol{\theta}$ estimated via available samples, $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N} \sim p(\mathbf{x})$

Variational Auto Encoders (VAEs)

$$\underset{\theta,\phi}{\arg\max} \prod_{i=1}^{N} p_\theta(\mathbf{x}^{(i)}) = \underset{\theta,\phi}{\arg\max} \prod_{i=1}^{N} \mathrm{ELBO}_{p,q}\left(\mathbf{x}^{(i)}\right)$$

$$= \underset{\theta,\phi}{\arg\max} \prod_{i=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\right],$$

$$p_\theta(\mathbf{x}\mid\mathbf{z}) = \mathrm{N}\left(\mathbf{x}\mid \mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})\right), \quad p_z(\mathbf{z}) = \mathrm{N}\left(\mathbf{z}\mid \mathbf{0}, \mathbf{I}\right)$$

variational approximation with a encoder network, $q_\phi(\mathbf{z}\mid\mathbf{x}) \approx p_\theta(\mathbf{z}\mid\mathbf{x})$

Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR*. 2014. URL: http://arxiv.org/abs/1312.6114.

VAEs—pros vs cons

fast joint and conditional sampling

intractable density evaluation, $p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z})$

high-variance training gradients due to approximating $\mathrm{ELBO}_{p,q}(\mathbf{x})$

Generative adversarial networks (GANs)

$$\min_{\theta} \max_{\phi} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \Big[\log D_{\phi}(\mathbf{x}) \Big] + \mathbb{E}_{\mathbf{z}\sim p_Z(\mathbf{z})} \Big[\log \left(1 - D_{\phi}\left(G_{\theta}(\mathbf{z})\right)\right) \Big]$$

$\mathbf{z} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\right), \quad G_{\theta}(\mathbf{z}) \sim p(\mathbf{x})$

discriminator network, $D_{\phi} : \mathcal{X} \to [0, 1]$

$D_{\phi}(\mathbf{x})$ classifies input based on its distribution—i.e. $\mathbf{x} \sim p(\mathbf{x})$ or $\mathbf{x} \sim p_{\theta}(\mathbf{x})$

Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

GANs—pros vs cons

fast joint and conditional sampling

do not provide a density model

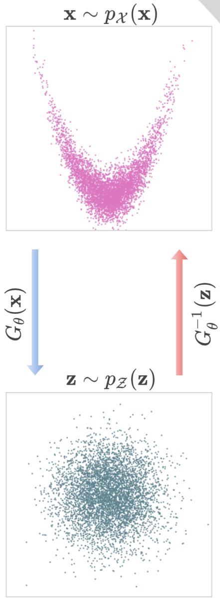
often unstable training due to min-max optimization

Normalizing flows (NFs)

Invertible neural network $G_\theta : \mathcal{X} \rightarrow \mathcal{Z},\ \mathcal{X}, \mathcal{Z} \in \mathbb{R}^d$

- flexible function for variational inference
- Gaussian latent space $p_z(\mathbf{z}) = (2\pi)^{-\frac{d}{2}} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|_2^2}$
- closed-from and exact inverse (up to numerical precision)
- memory-efficient training due to invertibility
- closed-form expression for $p_\theta(\mathbf{x})$

$$p_\theta(\mathbf{x}) = p_z(G_\theta(\mathbf{x})) \left| \det \nabla_x G_\theta(\mathbf{x}) \right| \approx p_{\mathcal{X}}(\mathbf{x})$$



Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1530–1538. URL: http://proceedings.mlr.press/v37/rezende15.html.

SLIM

Training NFs

$$\underset{\boldsymbol{\theta}}{\arg\min}\ \mathbb{D}_{\mathsf{KL}}\left(p_{\mathcal{X}} \mid\mid p_{\theta}\right) = \underset{\boldsymbol{\theta}}{\arg\min}\ \mathbb{E}_{\mathbf{x}\sim p_{\mathcal{X}}(\mathbf{x})}\Big[-\log p_{\theta}(\mathbf{x}) + \log p_{\mathcal{X}}(\mathbf{x})\Big]$$

$$\approx \underset{\theta}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\left[\ \underbrace{\frac{1}{2}\left\|G_{\theta}(\mathbf{x}_i)\right\|_2^2}_{\text{Gaussianizes the input}} - \underbrace{\log\left|\det \nabla_x G_{\theta}(\mathbf{x}_i)\right|}_{\substack{\text{regularizes entropy of } p_{\theta}(\mathbf{x}) \\ \text{e.g., avoids } G_{\theta}(\mathbf{x}) \equiv \mathbf{0}}}\ \right]$$

- Kullback-Leibner (KL) divergence $\mathbb{D}_{\mathsf{KL}}$
- training samples $\{\mathbf{x}^{(i)}\}_{i=1}^{n} \sim p_{\mathcal{X}}(\mathbf{x})$
- $\det \nabla_x G_{\theta}(\mathbf{x})$ comes for free with *affine coupling layers*

Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1530–1538. URL: `http://proceedings.mlr.press/v37/rezende15.html`.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: `http://arxiv.org/abs/1605.08803`.

Inference with NFs

Sampling from $p(\mathbf{x})$,

$$G_{\theta^*}^{-1}(\mathbf{z}) \sim p(\mathbf{x}), \quad \mathbf{z} \sim p_z(\mathbf{z}),$$

and fast density estimation,

$$p(\mathbf{x}) = p_\theta(G_\theta(\mathbf{x})) \left| \det \nabla_{\mathrm{x}} G_\theta(\mathbf{x}) \right| \approx p_z(G_\theta(\mathbf{x})) \left| \det \nabla_{\mathrm{x}} G_\theta(\mathbf{x}) \right|$$

NFs—pros vs cons

fast joint and conditional sampling

exact and tractable density (likelihood) estimation

allows for memory efficient gradient computation

often have less representation power

unable to alter dimensionality

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. "Why Normalizing Flows Fail to Detect Out-of-Distribution Data". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 20578–20589. URL: https://proceedings.neurips.cc/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

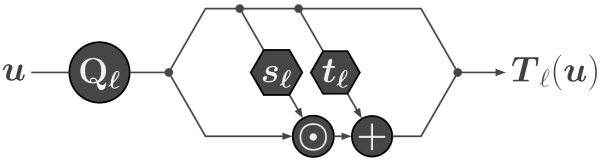
Robert Cornish, Anthony L. Caterini, George Deligiannidis, and Arnaud Doucet. "Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows". In: *ICML*. 2020, pp. 2133–2143. URL: http://proceedings.mlr.press/v119/cornish20a.html.

Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. "SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12685–12696. URL: https://proceedings.neurips.cc/paper/2020/file/9578a63fbe545bd82cc5bbe749636af1-Paper.pdf.

Affine coupling layer

a basic "layer" with an analytic inverse

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.



- an invertible layer with input $\boldsymbol{u}$
- orthogonal Householder reflections $\boldsymbol{Q}_\ell$
- arbitrary neural networks $\boldsymbol{s}_\ell$, $\boldsymbol{t}_\ell$
- output of ℓ^{th} layer $\boldsymbol{T}_\ell(\boldsymbol{u})$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: `http://arxiv.org/abs/1605.08803`.

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: `https://arxiv.org/pdf/1905.10687.pdf`.

Affine coupling layer—forward

$$\boldsymbol{T}_\ell(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \odot \sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right) + \boldsymbol{t}_\ell\left(\tilde{\boldsymbol{u}}_1\right) \end{bmatrix},$$

$$\text{with } \tilde{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix} = \boldsymbol{Q}_\ell \boldsymbol{u}$$

sigmoid function $\sigma\left(\cdot\right)$

elementwise multiplication $\odot$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.

Affine coupling layer—inverse

$$\tilde{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \left(\boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) - \boldsymbol{t}_\ell\left(\boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right)\right)\right) \oslash \sigma\left(\boldsymbol{s}_\ell\left(\boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right)\right)\right) \end{bmatrix},$$

$$\boldsymbol{u} = \boldsymbol{Q}_\ell^{-1} \begin{bmatrix} \tilde{\boldsymbol{u}}_1 \\ \tilde{\boldsymbol{u}}_2 \end{bmatrix}$$

orthogonal learnable matrix $\boldsymbol{Q}_\ell^{-1} = \boldsymbol{Q}_\ell^\top$

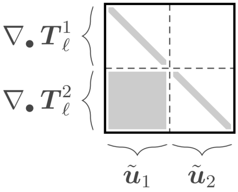
elementwise division $\oslash$

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density estimation using Real NVP". In: *International Conference on Learning Representations, ICLR*. 2016. URL: http://arxiv.org/abs/1605.08803.

Determinant of the Jacobian

$$\nabla_{\boldsymbol{u}} \boldsymbol{T}_\ell(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \begin{bmatrix} \boldsymbol{T}_\ell^1\left(\tilde{\boldsymbol{u}}_1\right) \\ \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \nabla_{\boldsymbol{u}_1} \boldsymbol{T}_\ell^2\left(\tilde{\boldsymbol{u}}_1, \tilde{\boldsymbol{u}}_2\right) & \operatorname{diag}\Big(\sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right)\Big) \end{bmatrix}$$

$$\Rightarrow \log\left|\det \nabla_{\boldsymbol{u}} \boldsymbol{T}_\ell(\boldsymbol{u})\right| = \mathrm{sum}\left(\log \sigma\left(\boldsymbol{s}_\ell\left(\tilde{\boldsymbol{u}}_1\right)\right)\right)$$



sparsity pattern of Jacobian

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Challenges of affine coupling layers

lower representation power due

- sparse triangular Jacobian

requires many layer compositions and orthogonal transforms

- to capture all pixel-wise dependencies

may become numerically non-invertible

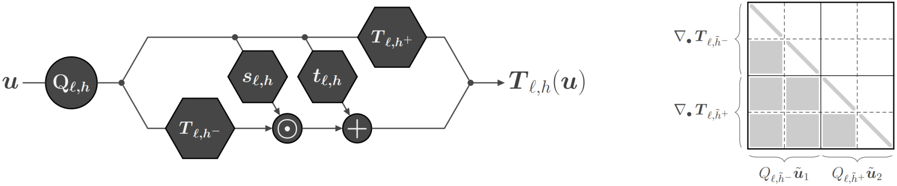
- e.g. due to very small Jacobian singular values

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. “Why Normalizing Flows Fail to Detect Out-of-Distribution Data”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 20578–20589. URL: https://proceedings.neurips.cc/paper/2020/file/ecb9fe2fbb99c31f567e9823e884dbec-Paper.pdf.

Jens Behrmann, Paul Vicol, Kuan-Chieh Wang, Roger Grosse, and Jörn-Henrik Jacobsen. “Understanding and mitigating exploding inverses in invertible neural networks”. In: *arXiv preprint arXiv:2006.09347* (2020).

Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. “SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 12685–12696. URL: https://proceedings.neurips.cc/paper/2020/file/9578a63fbe545bd82cc5bbe749636af1-Paper.pdf.

Hierarchical coupling layers



dense triangular Jacobian

hierarchical architecture

- recursive affine coupling layers
- until input can not be split any further

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Posterior inference with NFs

Train a NF to directly characterize the posterior density

$$p_\theta\left(\mathbf{x}\right) \approx p_{\mathrm{post}}\left(\mathbf{x} \mid \mathbf{y}\right) \propto p_{\mathrm{like}}\left(\mathbf{y} \mid \mathbf{x}\right)\, p_{\mathrm{prior}}\left(\mathbf{x}\right),$$

i.e.,

being able to samples from the posterior

estimate the posterior density

Training a NF for posterior inference

$$
\begin{aligned}
&\arg\min_{\theta} \mathbb{D}_{\mathsf{KL}} \left(p_\theta \mid\mid p_{\mathsf{post}}(\, \cdot \mid \mathbf{y})\right) \\
&= \arg\min_{\theta} \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x}\mid\mathbf{y})} \Big[-\log p_{\mathsf{post}}(\mathbf{x} \mid \mathbf{y}) + \log p_\theta(\mathbf{x}) \Big].
\end{aligned}
$$

KL divergence, $\mathbb{D}_{\mathsf{KL}}$, a distance between two distributions

distribution encoded by the NF, $p_\theta(\mathbf{x}) = p_z(\mathbf{z}) \left| \det \nabla_x T_\theta^{-1}(\mathbf{x}) \right|, \quad \mathbf{z} = T_\theta^{-1}(\mathbf{x})$

SLIM

Posterior inference with NFs

purely physics-based approach

- tied to the physics
- no training data needed
- requires handcrafted priors

purely data-driven approach

- needs supervised pairs of model and data
- heavily relies on the training data to generalize
- ignores the physics

Physics-based variational inference

$$
\begin{aligned}
&\underset{\theta}{\arg\min}\; \mathbb{D}_{\mathrm{KL}}\left(p_\theta \,||\, p_{\mathrm{post}}(\,\cdot \mid \mathbf{y})\right) \\
&= \underset{\theta}{\arg\min}\; \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x})}\left[-\log p_{\mathrm{post}}(\mathbf{x} \mid \mathbf{y}) + \log p_\theta(\mathbf{x})\right] \\
&= \underset{\theta}{\arg\min}\; \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}\left[-\log p_{\mathrm{post}}(T_\theta(\mathbf{z}) \mid \mathbf{y}) + \log p_\theta(T_\theta(\mathbf{z}))\right] \\
&= \underset{\theta}{\arg\min}\; \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}\left[-\log p_{\mathrm{post}}(T_\theta(\mathbf{z}) \mid \mathbf{y}) + \log p_z(\mathbf{z}) - \log \left|\det \nabla_z T_\theta(\mathbf{z})\right|\right] \\
&= \underset{\theta}{\arg\min}\; \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}\left[-\log p_{\mathrm{post}}(T_\theta(\mathbf{z}) \mid \mathbf{y}) - \log \left|\det \nabla_z T_\theta(\mathbf{z})\right|\right].
\end{aligned}
$$

Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 2378–2386. URL: https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Physics-based variational inference

$$\min_{\theta} \mathbb{E}_{\mathbf{z}\sim p_{z}(\mathbf{z})} \left[\frac{1}{2\sigma^2} \left\| F\left(T_\theta(\mathbf{z})\right) - \mathbf{y} \right\|_2^2 - \log p_{\text{prior}}\left(T_\theta(\mathbf{z})\right) - \log \left| \det \nabla_z T_\theta(\mathbf{z}) \right| \right]$$

no training data required

cheap posterior sampling: $T_\theta(\mathbf{z}) \sim p_\theta(\mathbf{x} \mid \mathbf{y}) \approx p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$

requires a prior density, $p_{\text{prior}}(\mathbf{x})$

repeated evaluations of F and $\nabla F^\top$

specific to one observation, $\mathbf{y}$

Qiang Liu and Dilin Wang. "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm". In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016, pp. 2378–2386. URL: https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Xuebin Zhao, Andrew Curtis, and Xin Zhang. "Bayesian Seismic Tomography using Normalizing Flows". In: (2020). DOI: 10.31223/X53K6G. URL: https://eartharxiv.org/repository/view/1940/.

Data-driven variational inference

achieved via a block-triangular transport map, $G_\theta(\mathbf{y}, \mathbf{x})$

training over model and data pairs, $(\mathbf{y}, \mathbf{x}) \sim \widehat{p}_{y,x}(\mathbf{y}, \mathbf{x})$

capable of sampling the posterior for all $\mathbf{y} \sim p_y(\mathbf{y})$

provides an estimate to the posterior density

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Data-driven variational inference

$$\arg\max_{\phi} \mathbb{E}_{\mathbf{y},\mathbf{x}\sim\widehat{p}_{y,x}(\mathbf{y},\mathbf{x})} \left[\log p_{\theta}(\mathbf{y},\mathbf{x})\right]$$

$$= \arg\min_{\phi} \mathbb{E}_{\mathbf{y},\mathbf{x}\sim\widehat{p}_{y,x}(\mathbf{y},\mathbf{x})} \left[\frac{1}{2}\left\|G_{\theta}(\mathbf{y},\mathbf{x})\right\|^2 - \log\left|\det \nabla_{y,x} G_{\theta}(\mathbf{y},\mathbf{x})\right|\right]$$

$$\approx \arg\min_{\phi} \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{2}\left\|G_{\theta}(\mathbf{y}_i,\mathbf{x}_i)\right\|^2 - \log\left|\det \nabla_{y,x} G_{\theta}(\mathbf{y}_i,\mathbf{x}_i)\right|\right]$$

NF, $G_{\theta} : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x$

maximum likelihood estimate for $\boldsymbol{\theta}$ given data pairs, $\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim \widehat{p}_{y,x}(\mathbf{y},\mathbf{x})$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

Block-triangular map

$$G_\theta(\mathbf{y}, \mathbf{x}) = \begin{bmatrix} G_{\theta_y}(\mathbf{y}) \\ G_{\theta_x}(\mathbf{y}, \mathbf{x}) \end{bmatrix}, \quad \phi = \begin{bmatrix} \phi_y \\ \phi_x \end{bmatrix}$$

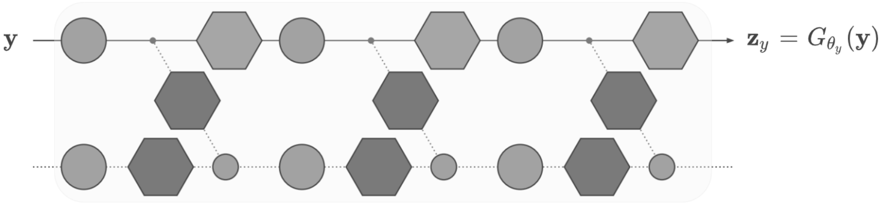
conditional sampling: $G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z}) \sim p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I})$

posterior density estimation: $p_G(\mathbf{x} \mid \mathbf{y}) = p_{\mathbf{z}}\left(G_{\theta_x}(\mathbf{y}, \mathbf{x})\right) \left| \det \nabla_x G_{\theta_x}(\mathbf{y}, \mathbf{x}) \right|.$

Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. *An adaptive transport framework for joint and conditional density estimation*. 2020. arXiv: 2009.10303 [stat.ML].

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: https://arxiv.org/pdf/1905.10687.pdf.

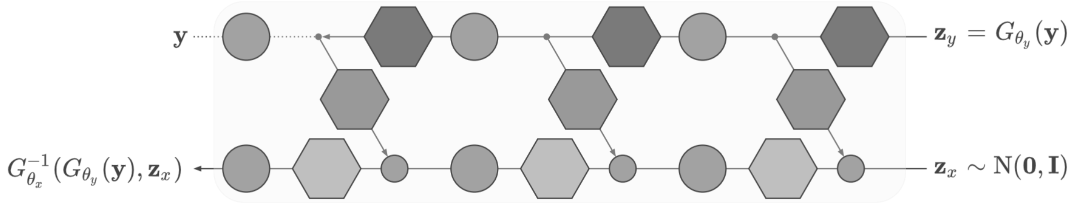
Posterior inference with conditional NFs



$$\mathbf{z}_y = G_{\theta_y}(\mathbf{y})$$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: `https://arxiv.org/pdf/1905.10687.pdf`.

Posterior inference with conditional NFs



$$G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z}_x) \sim p_{\text{post}}\left(\mathbf{x} \mid \mathbf{y}\right), \quad \mathbf{z}_x \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference". In: *Proceedings of AAAI-2021* (2021). URL: `https://arxiv.org/pdf/1905.10687.pdf`.

Data-driven variational inference

$$\min_{\phi} \mathbb{E}_{\mathbf{y},\mathbf{x}\sim p_{y,x}(\mathbf{y},\mathbf{x})} \left[\frac{1}{2}\left\|G_\theta(\mathbf{y},\mathbf{x})\right\|^2 - \log\left|\det \nabla_{y,x} G_\theta(\mathbf{y},\mathbf{x})\right|\right]$$

cheap posterior sampling for all $\mathbf{y}$: $G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}),\mathbf{z}) \sim p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$ for $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I})$

does not involve the expensive forward operator, F

does not require a prior density, $p_{\text{prior}}(\mathbf{x})$

heavily relies on access to training pairs, $\widehat{p}_{y,x} \neq p_{y,x}$

Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. *An adaptive transport framework for joint and conditional density estimation*. 2020. arXiv: 2009.10303 [stat.ML].

Multi-fidelity preconditioned scheme

initialize $T_{\phi_x}(\mathbf{z}) := G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z})$, followed by transfer learning

use $p_G(\mathbf{x})$—density encoded by $G_{\theta_x}^{-1}$—as a (conditional) prior

Multifidelity preconditioned formulations

- insert the conditional NF in physics-based formulations and use transfer learning

$$T_{\theta_x}(\mathbf{z}) := G_{\theta_x}^{-1}(G_{\theta_y}(\mathbf{y}), \mathbf{z})$$

- preconditioned maximum a posteriori (MAP) estimation

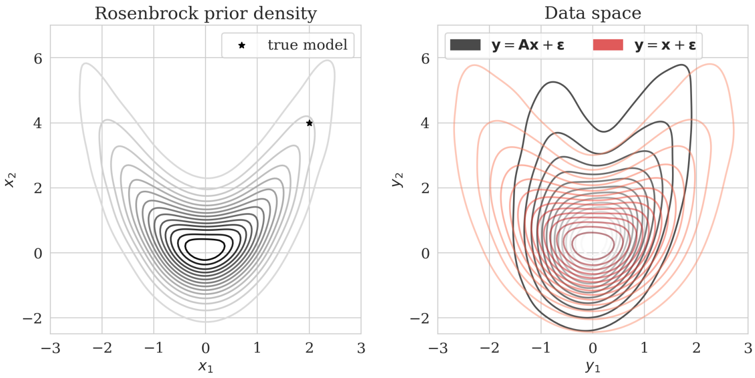
$$\min_{\mathbf{z}} \frac{1}{2\sigma^2} \left\| F\left(T_{\theta_x}(\mathbf{z})\right) - \mathbf{y} \right\|_2^2 + \frac{1}{2} \|\mathbf{z}\|_2^2$$

- preconditioned physics-based posterior learning

$$\min_{\theta_x} \mathbb{E}_{\mathbf{z}\sim \mathrm{N}(\mathbf{0},\mathbf{I})} \left[\frac{1}{2\sigma^2} \left\| F\left(T_{\theta_x}(\mathbf{z})\right) - \mathbf{y} \right\|_2^2 - \log p_{\text{prior}}\left(T_{\theta_x}(\mathbf{z})\right) - \log \left| \det \nabla_z T_{\theta_x}(\mathbf{z}) \right| \right]$$

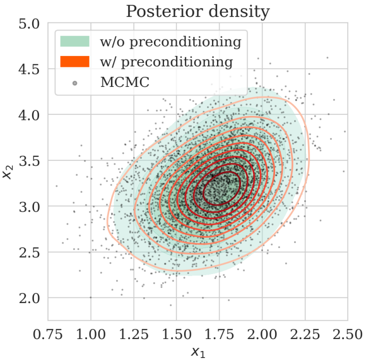
Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

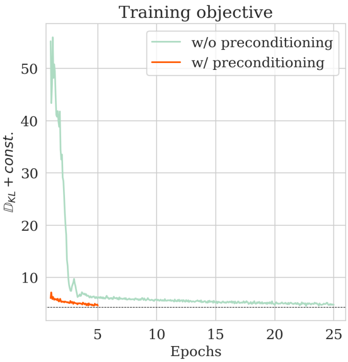
2D Rosenbrock toy example



(left) prior, (right) low- and high-fidelity data

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.



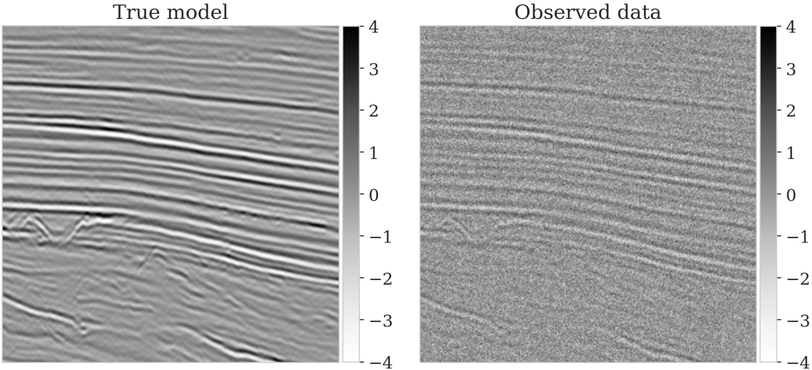


(left) without vs with preconditioning vs MCMC, (right) loss with and without preconditioning

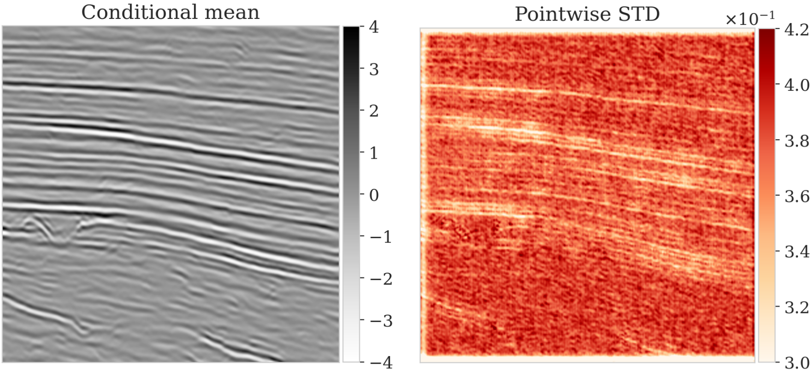
---

Ali Siahkoohi, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. "Preconditioned training of normalizing flows for variational inference in inverse problems". In: *3rd Symposium on Advances in Approximate Bayesian Inference*. Jan. 2021. URL: https://openreview.net/pdf?id=P9m1sMaNQ8T.

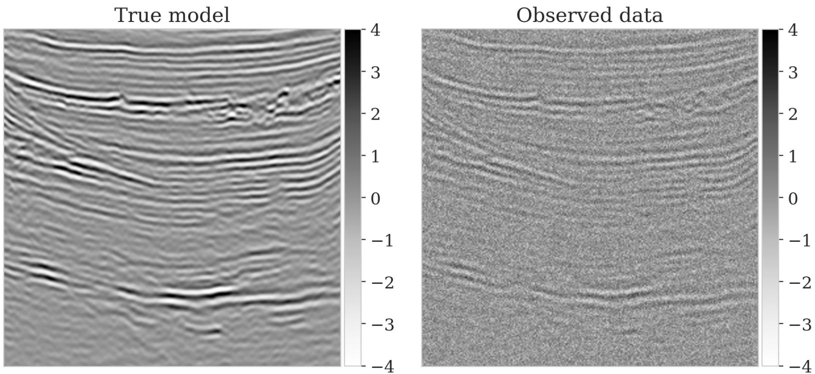
Seismic compressed sensing example



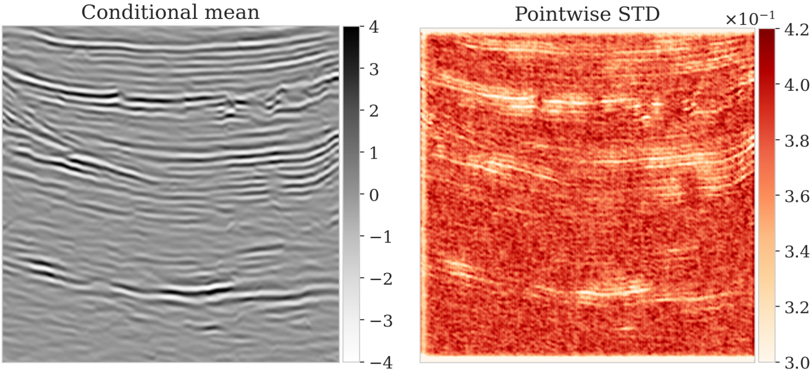
(left) true model, (right) “observed” data



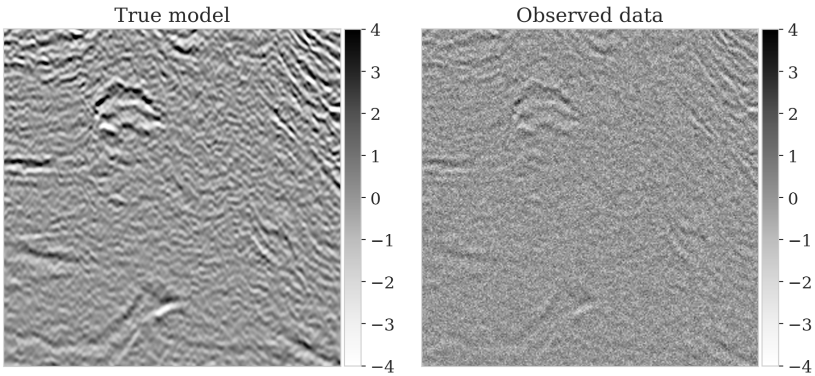
conditional mean (left) and pointwise standard deviation (right)



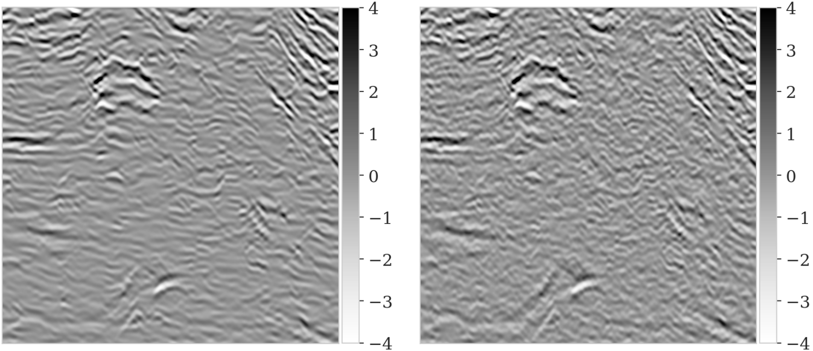
(left) true model, (right) “observed” data



conditional mean (left) and pointwise standard deviation (right)



(left) true model, (right) observed data



conditional mean before (left) and after (right) transfer learning

Seismic imaging examples

Seismic imaging forward model

$$\delta \mathbf{d} = \mathbf{J}[\mathbf{m}_0, \mathbf{q}]\delta \mathbf{m} + \boldsymbol{\eta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

linearized observed data $\delta \mathbf{d}$

source signature $\mathbf{q}$

smooth background model $\mathbf{m}_0$

unknown perturbation model $\delta \mathbf{m}$

linearization error and noise $\boldsymbol{\eta}, \boldsymbol{\epsilon}$

Linearized Born modeling operator

$$\mathbf{J}[\mathbf{m}_0, \mathbf{q}] = \nabla_{\mathbf{m}} \mathbf{P}\mathbf{A}^{-1}[\mathbf{m}]\mathbf{q}\Big|_{\mathbf{m}=\mathbf{m}_0}$$

$$= -\mathbf{P}\underbrace{\mathbf{A}^{-1}[\mathbf{m}_0]\,\mathrm{diag}\left(\nabla\mathbf{A}[(\mathbf{m}_0]\overbrace{\mathbf{A}^{-1}[\mathbf{m}_0]\mathbf{q}}^{\text{PDE solve}}\right)}_{\text{PDE solve}}$$

discretized wave equation $\mathbf{A}[\mathbf{m}]$

restriction operator $\mathbf{P}$

Training dataset for the data-driven approach

Creating perturbation models $\delta\mathbf{m}$

$3075\,\mathrm{m} \times 5120\,\mathrm{m}$ sections from the Parihaka dataset

selected from shallow sections

mainly consists of strong horizontal reflectors

artificial $125\,\mathrm{m}$ water layer to limit the near source imaging artifacts

WesternGeco. *Parihaka 3D PSTM Final Processing Report*. Tech. rep. New Zealand Petroleum Report 4582. 2012. URL: `https://wiki.seg.org/wiki/Parihaka-3D`.

Training dataset for the data-driven approach

For each reflectivity model $\delta\mathbf{m}$ define

$$\mathbf{x} := \delta\mathbf{m}$$

and

$$\mathbf{y} := \mathbf{J}^{\top}(\mathbf{J}\delta\mathbf{m} + \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Training dataset for the data-driven approach

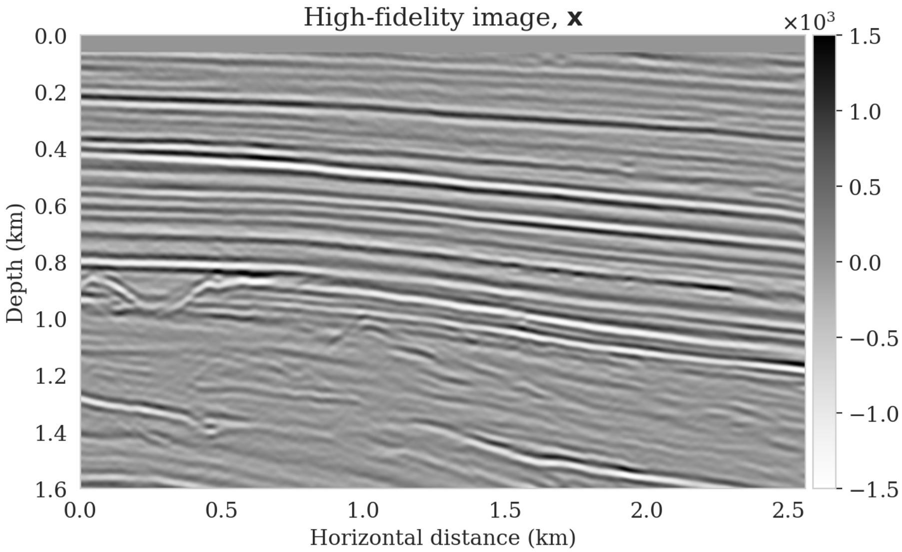
Training pairs:

$$\{\mathbf{y}_i, \mathbf{x}_i\}_{i=1}^{n} \sim p(\mathbf{y}, \mathbf{x})$$

where

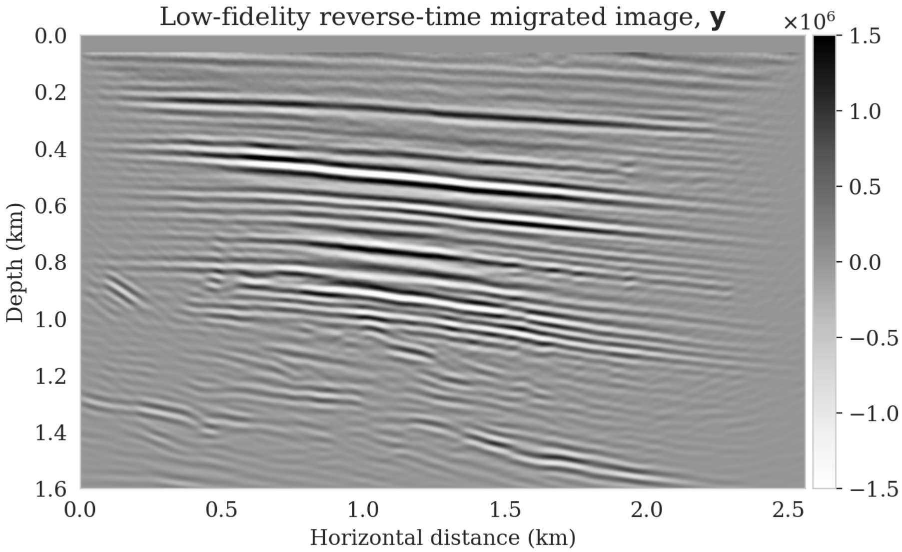
$$(\mathbf{y}_i, \mathbf{x}_i) = \left(\mathbf{J}^\top(\mathbf{J}\delta\mathbf{m}_i + \boldsymbol{\epsilon}_i), \delta\mathbf{m}_i\right)$$

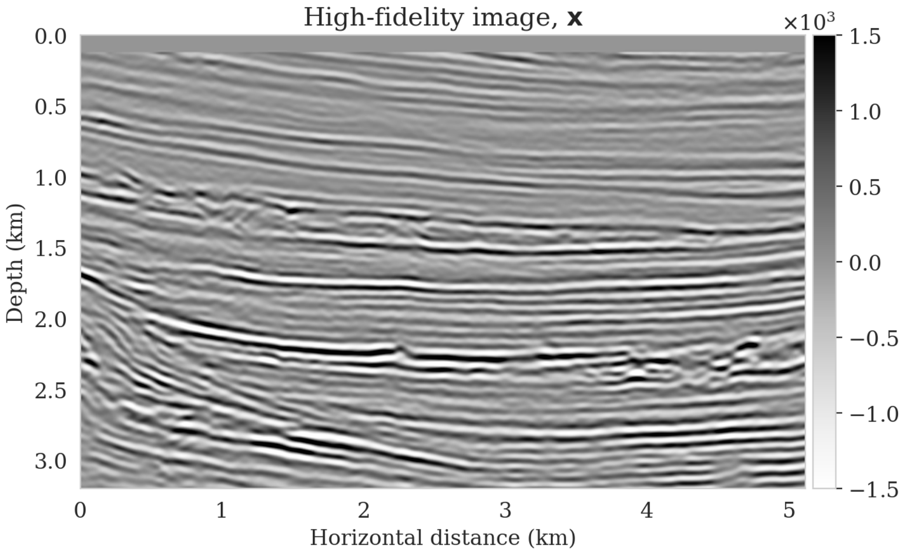
Some data pairs from the training set

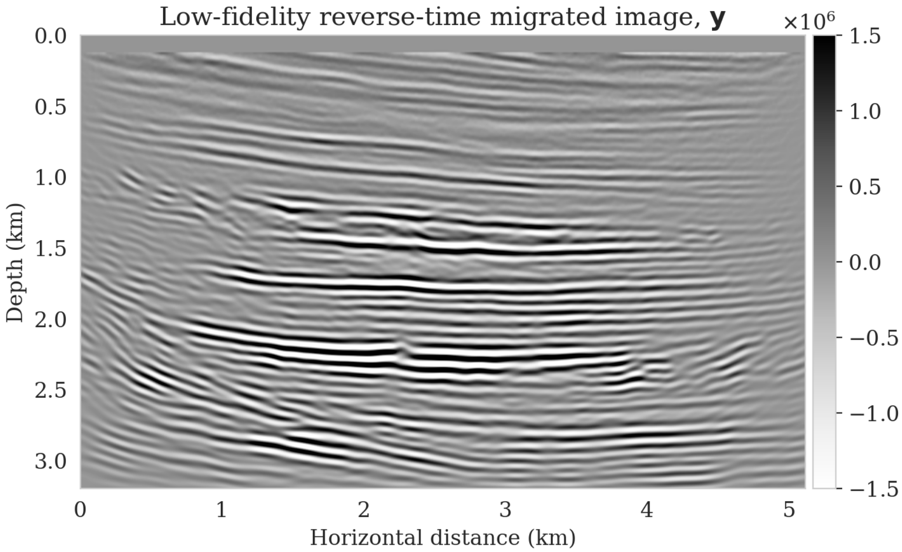


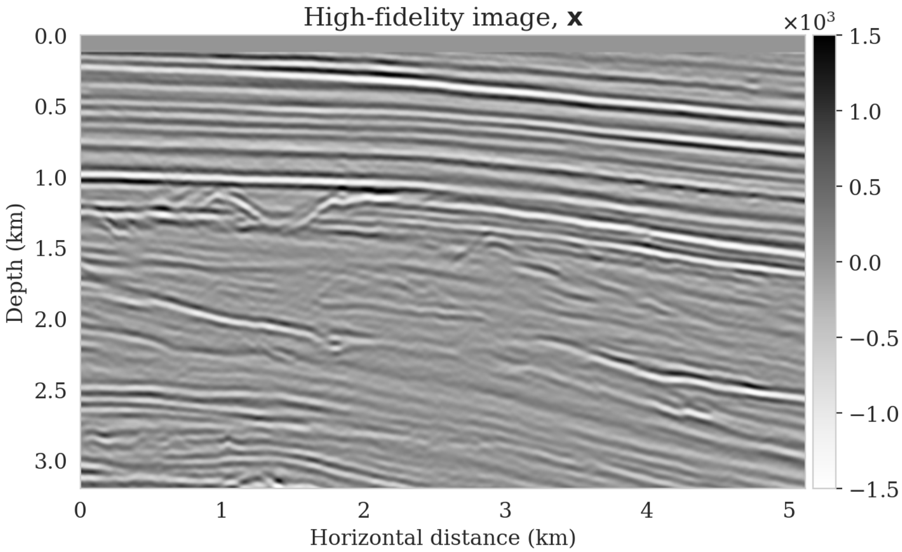
High-fidelity image, **x**
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

SLIM


Low-fidelity reverse-time migrated image, y
Depth (km)
Horizontal distance (km)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0.0
0.5
1.0
1.5
2.0
2.5
×10^6
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

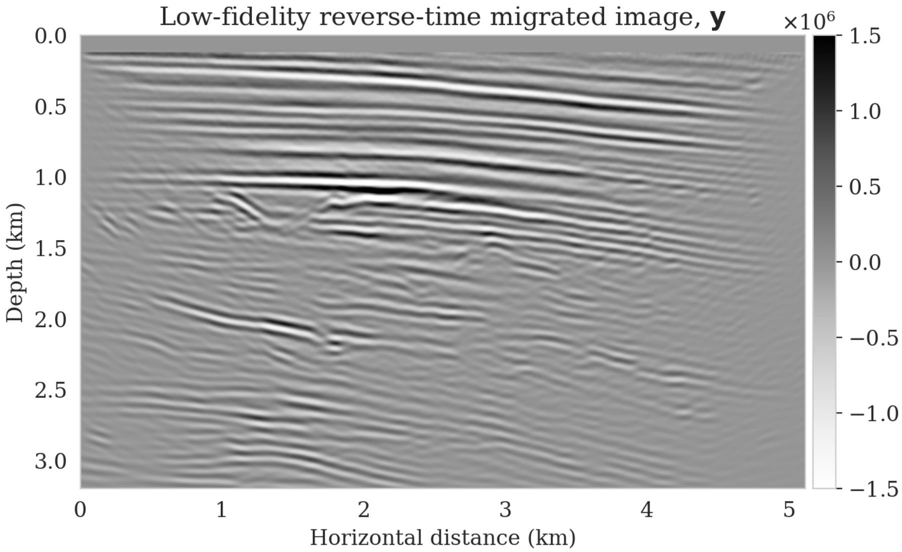

High-fidelity image, **x**
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)


Low-fidelity reverse-time migrated image, **y**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^6
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



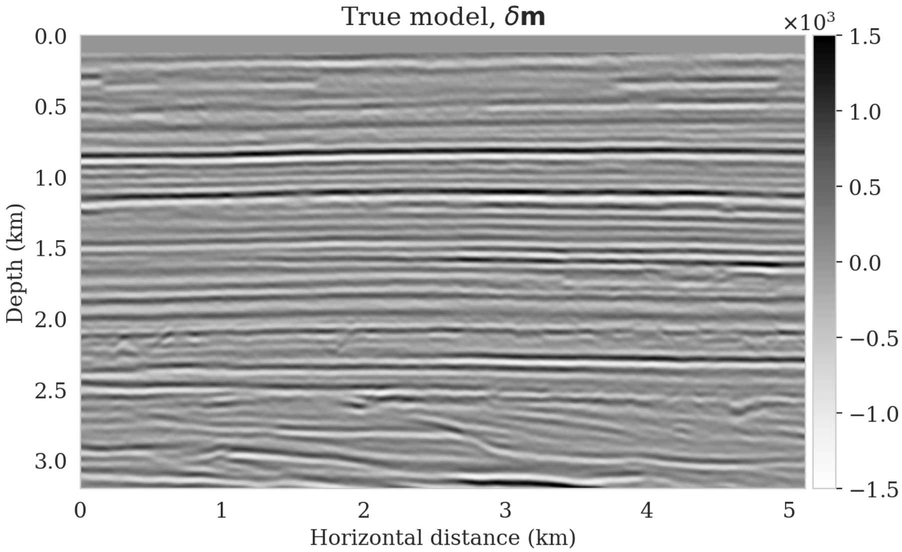
High-fidelity image, **x**
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

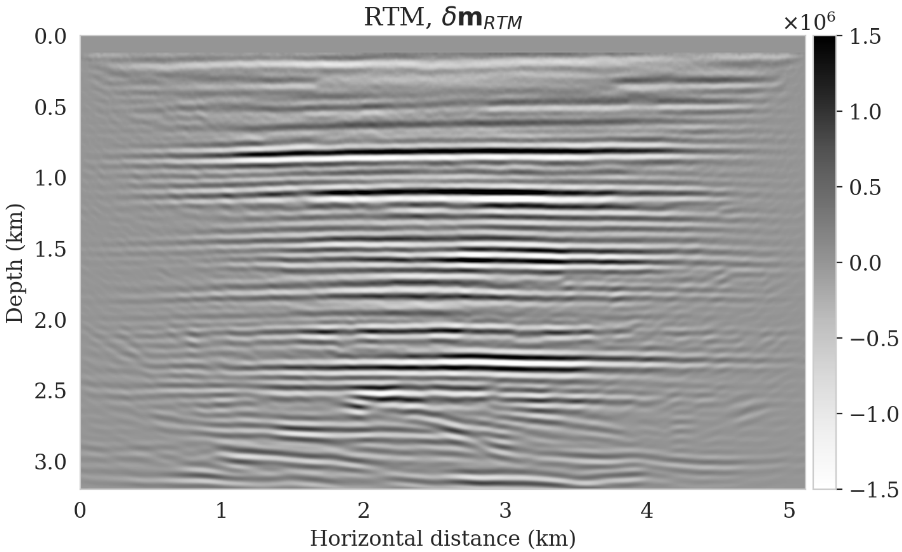
SLIM

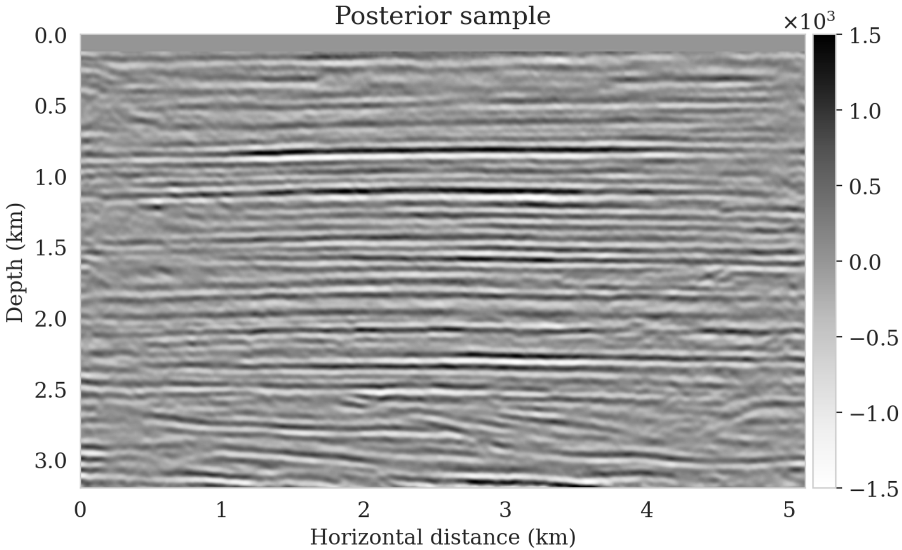

Low-fidelity reverse-time migrated image, **y**
×10^6
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

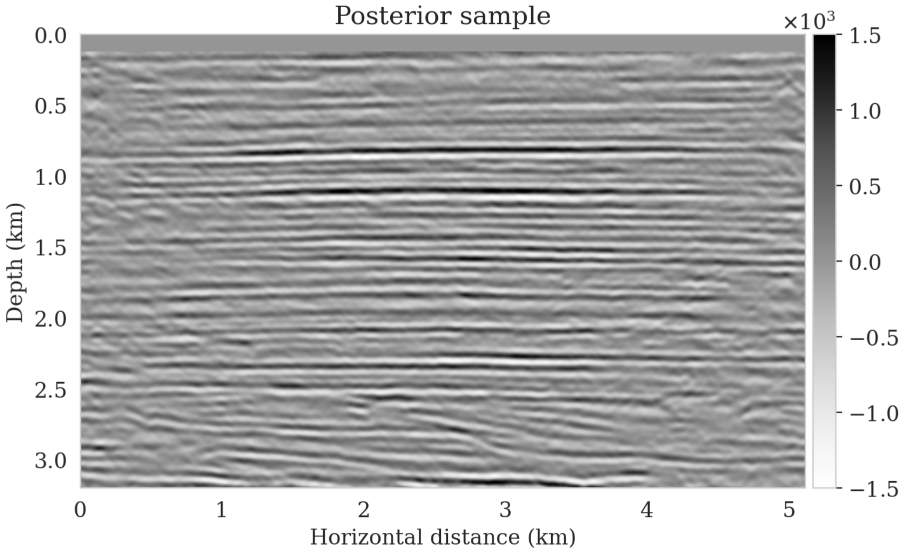
SLIM

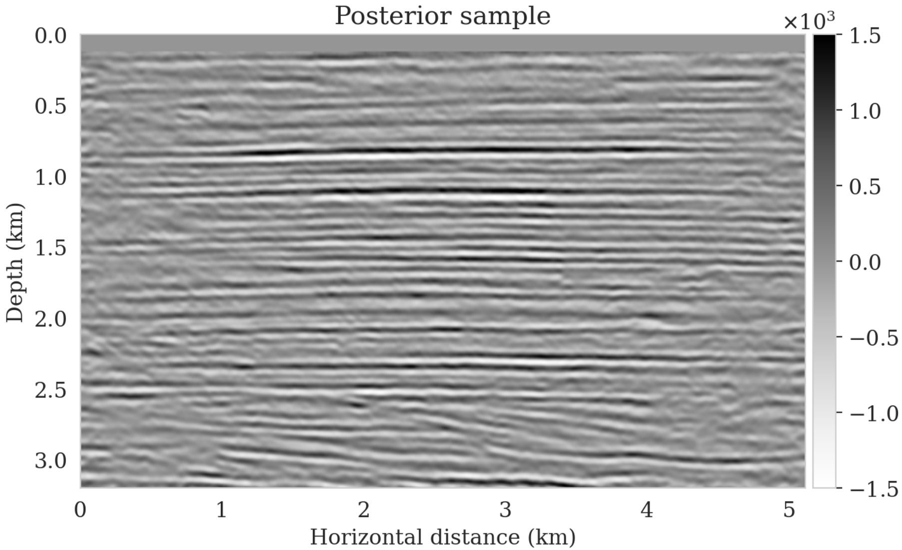
Results

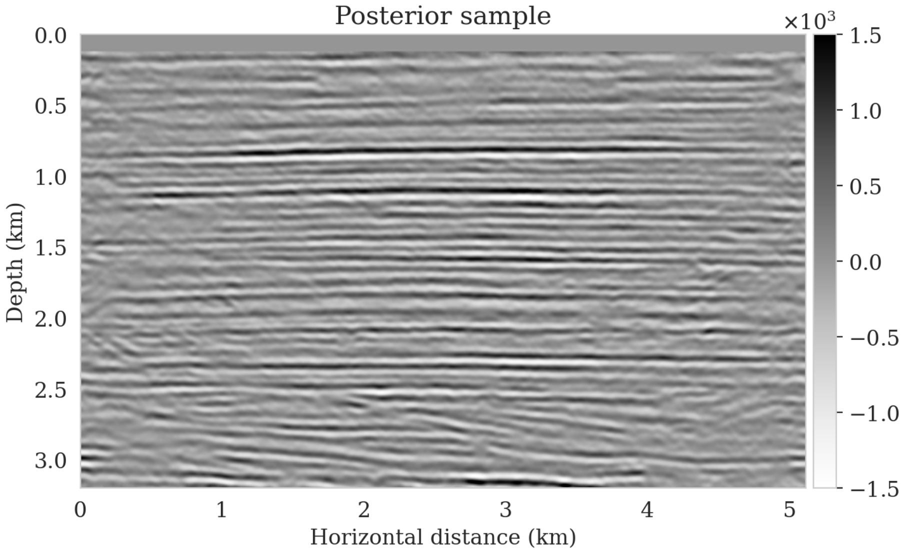

True model, δm
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

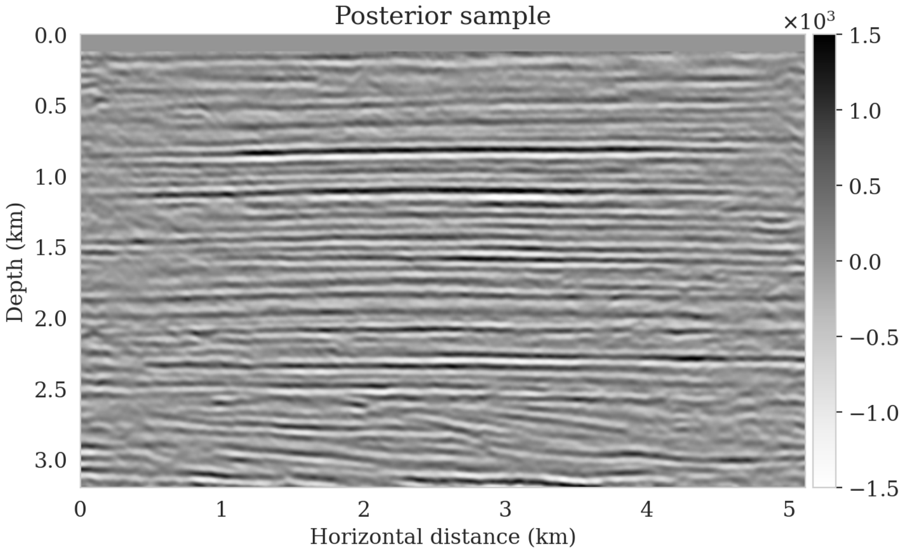

RTM, $\delta \mathbf{m}_{RTM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^6$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

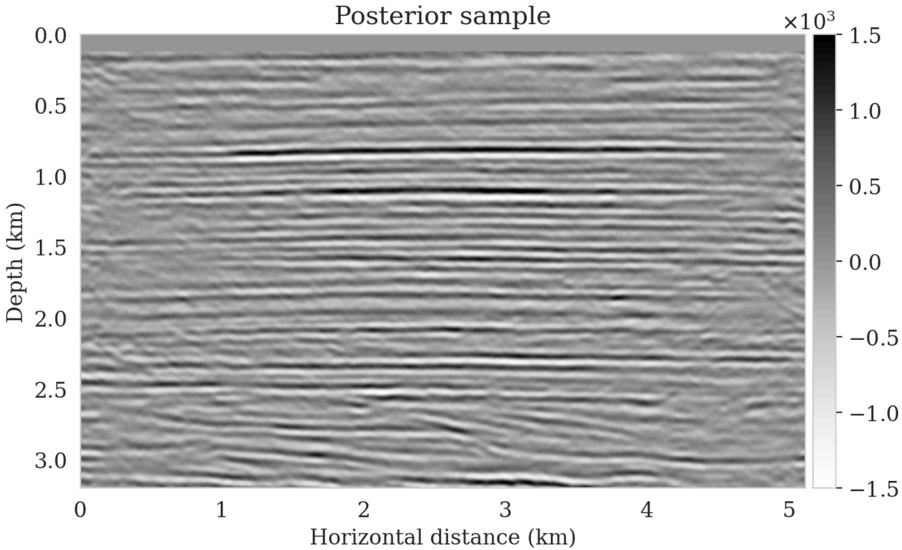

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

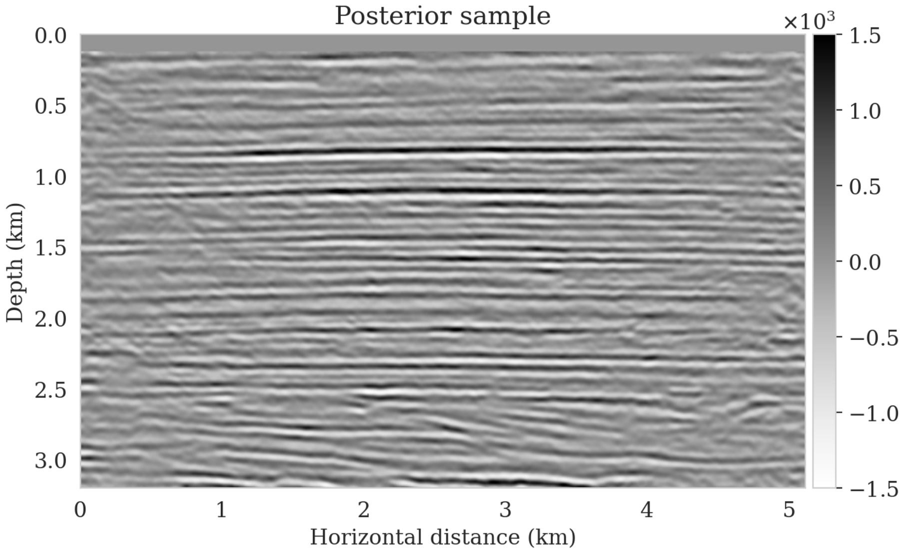

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

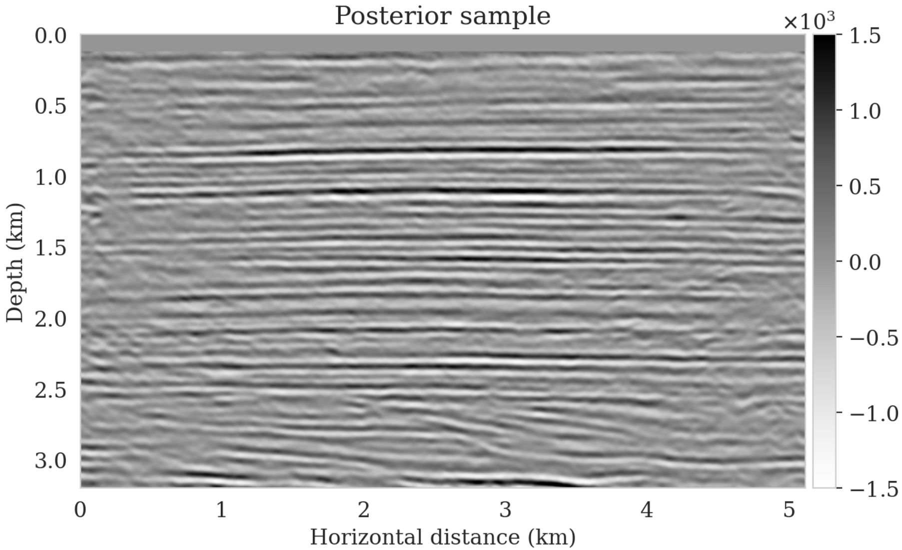

Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)


Posterior sample
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)

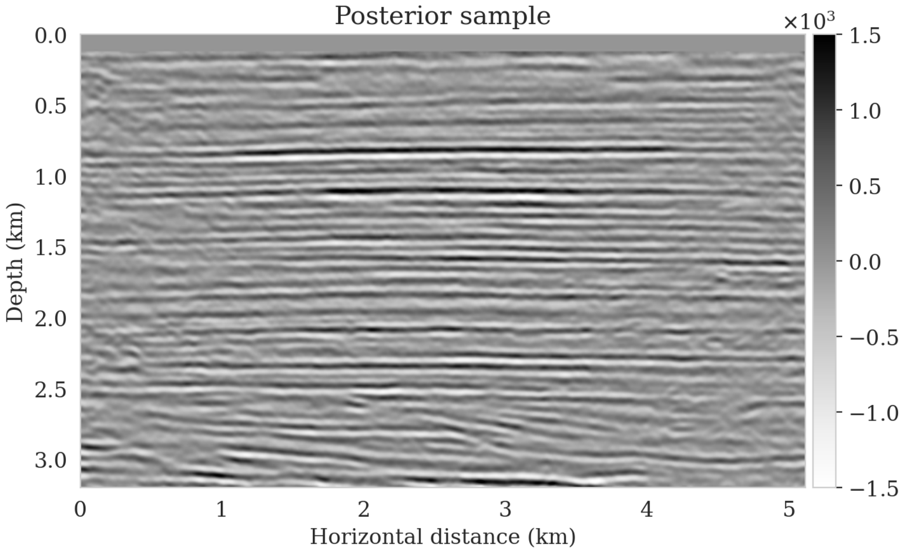
SLIM

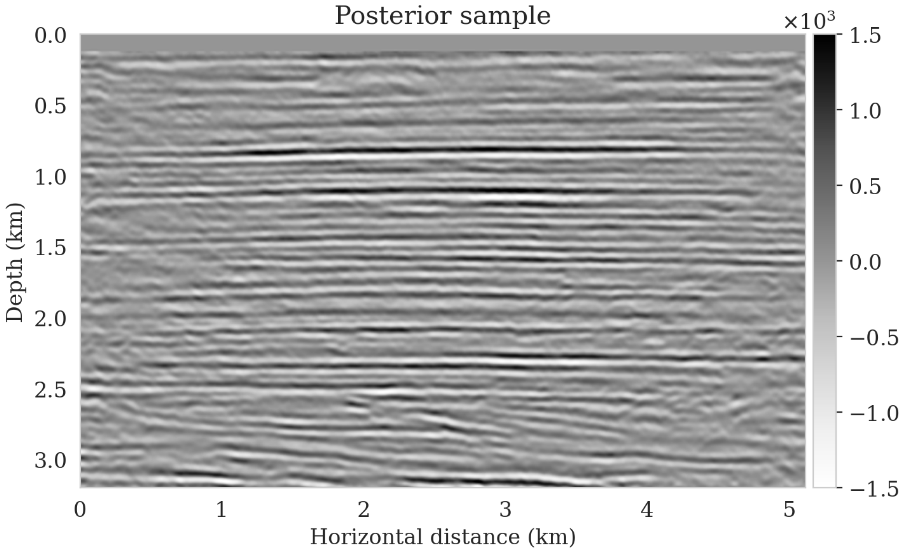

Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

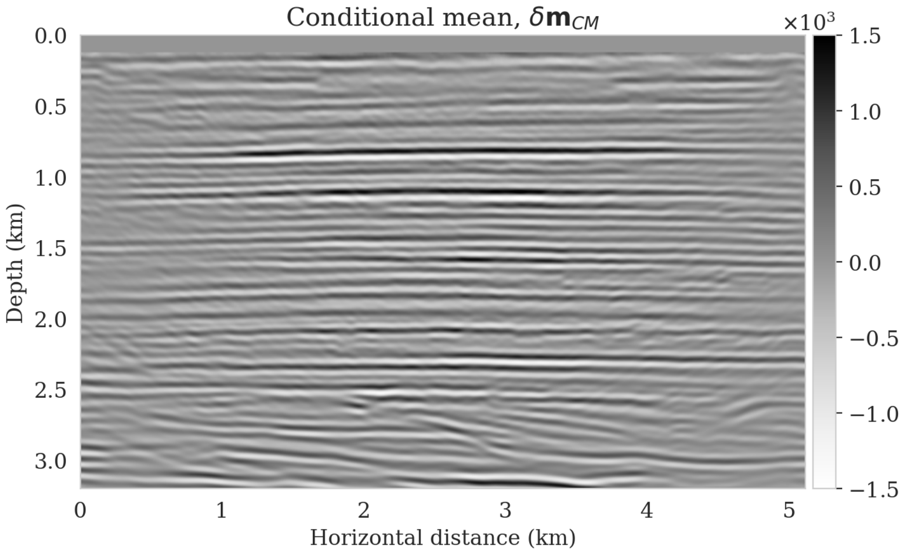

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

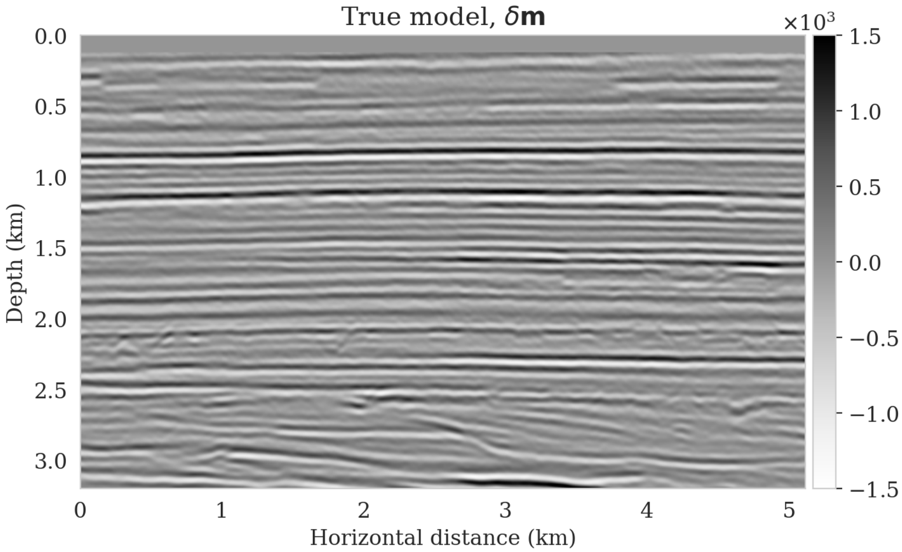


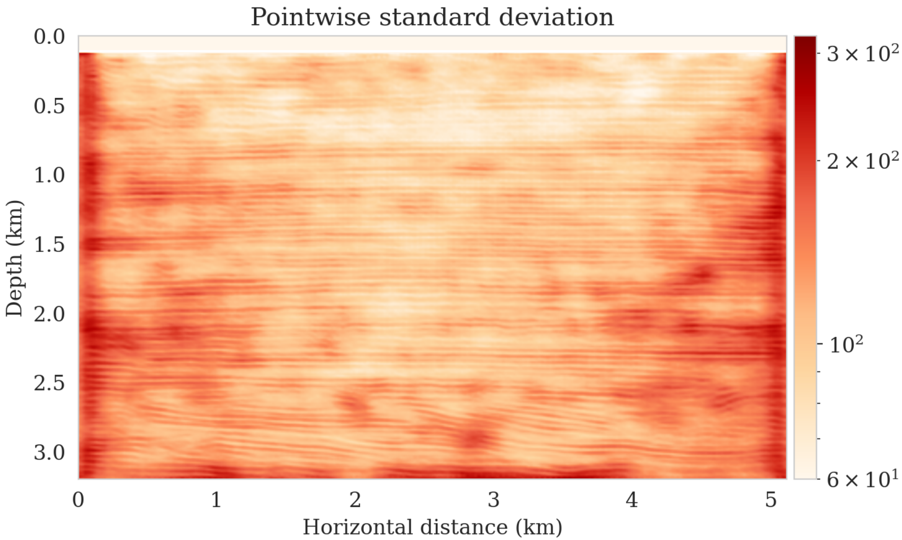
Posterior sample
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

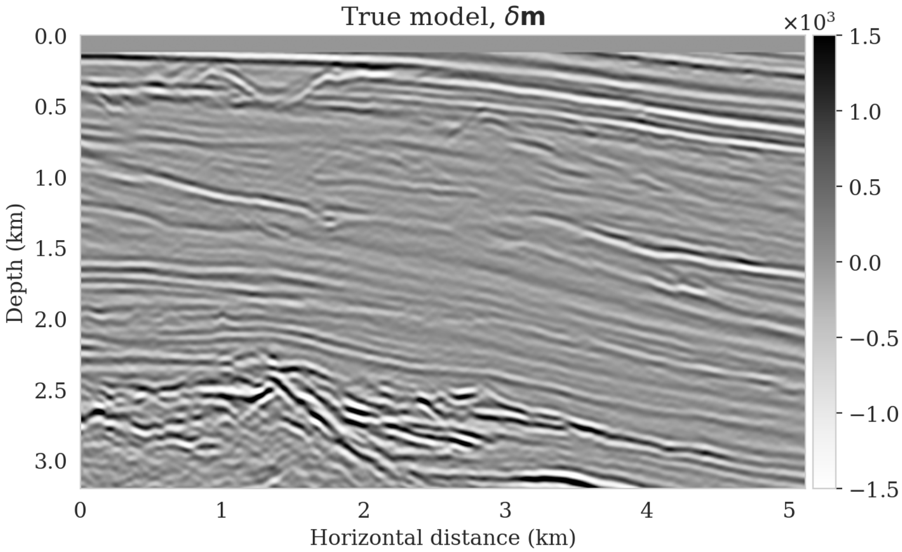

Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Conditional mean, $\delta\mathbf{m}_{CM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

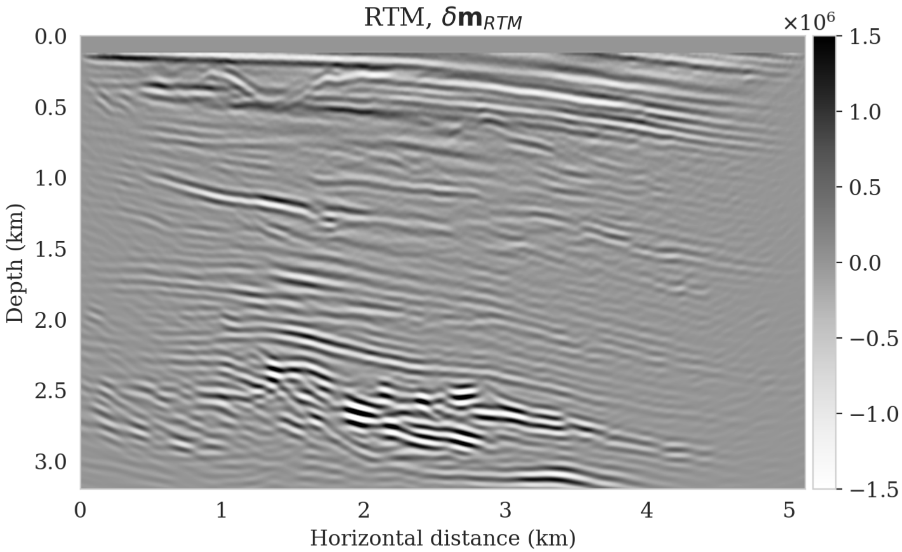

True model, $\delta\mathbf{m}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

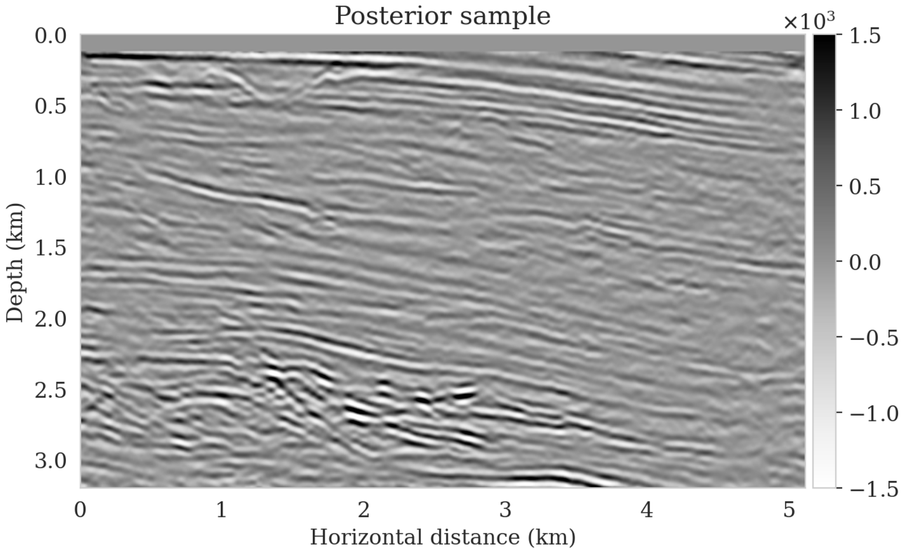


Pointwise standard deviation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
3×10^2
2×10^2
10^2
6×10^1

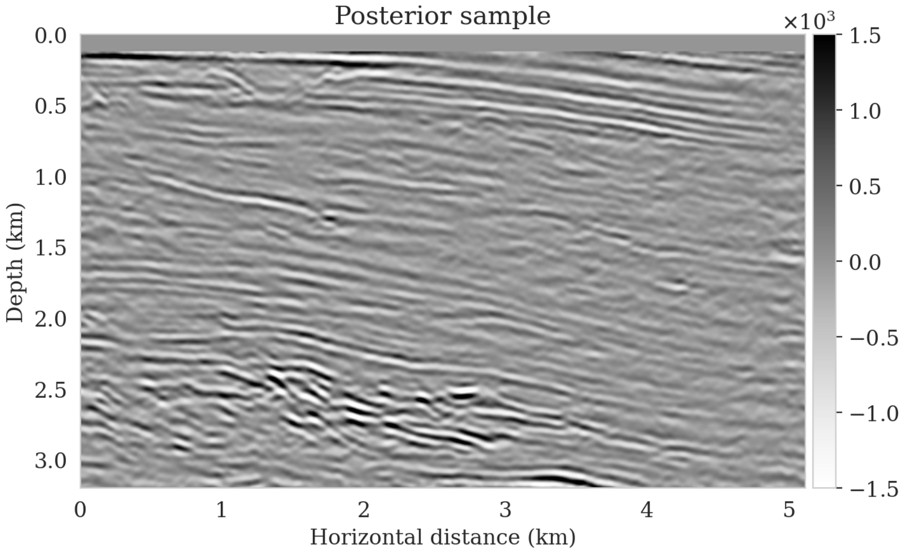

True model, $\boldsymbol{\delta m}$
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

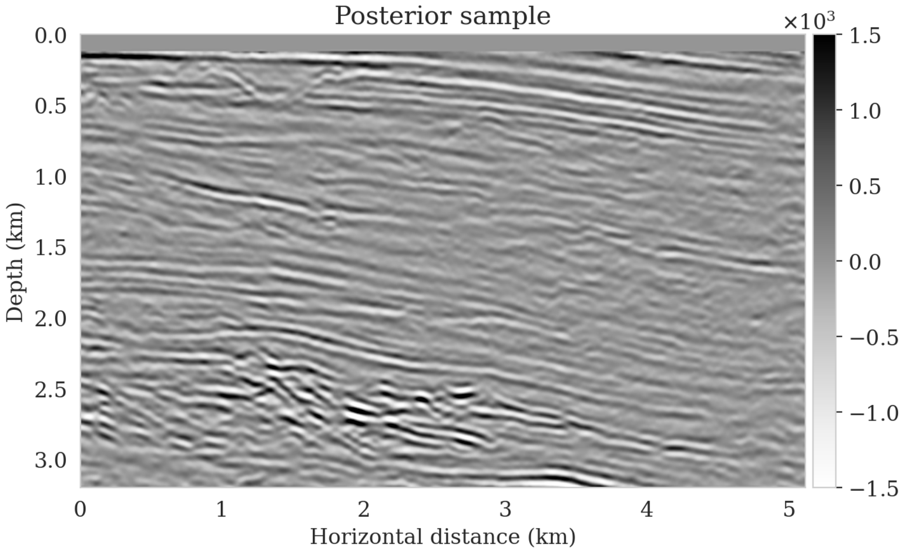
SLIM


RTM, $\delta \mathbf{m}_{RTM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^6$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

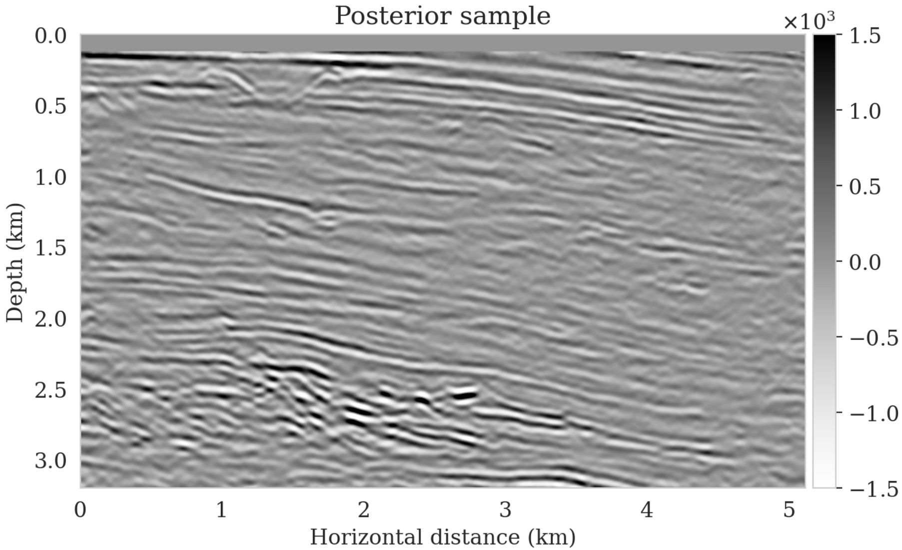


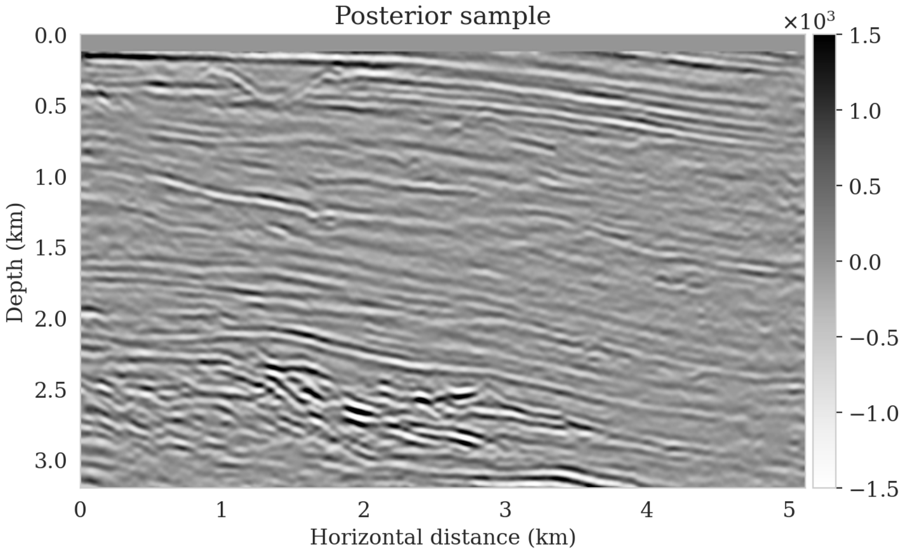
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

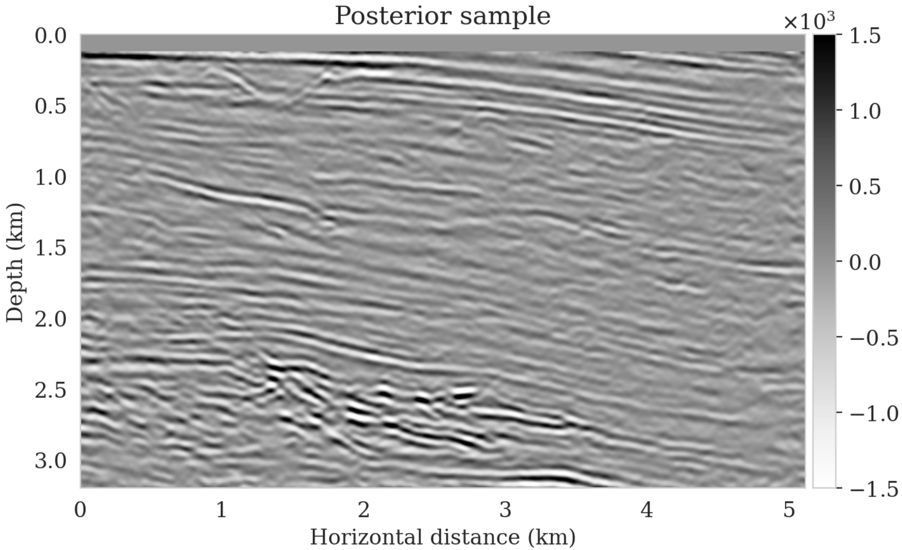


Posterior sample
Depth (km)
Horizontal distance (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

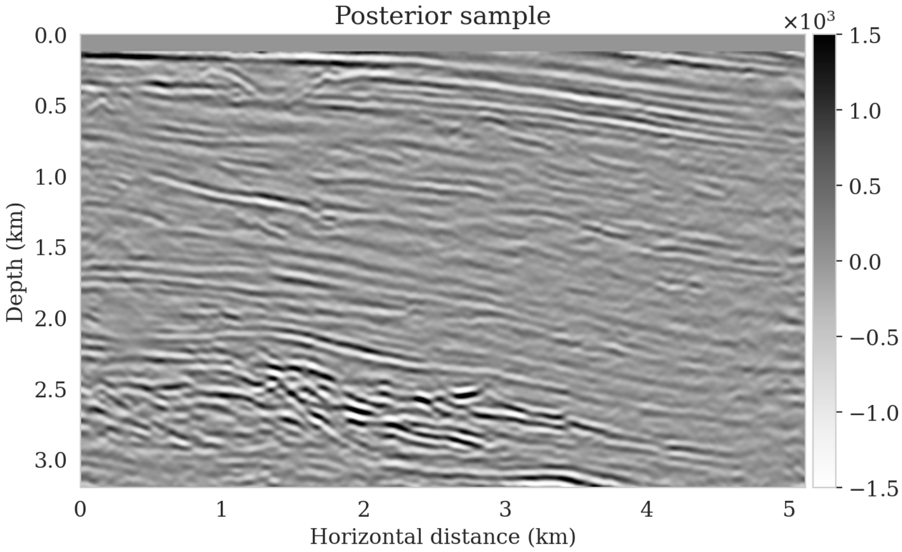

Posterior sample
×10³
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

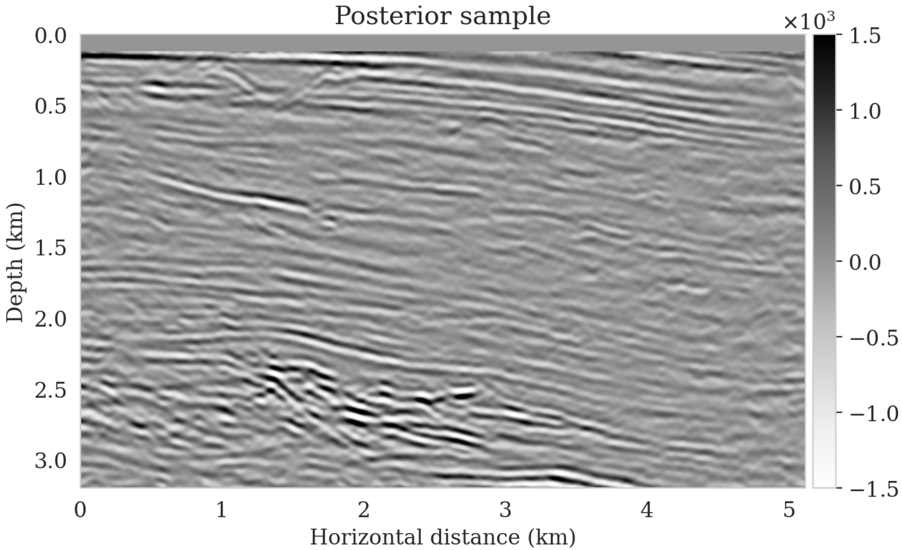


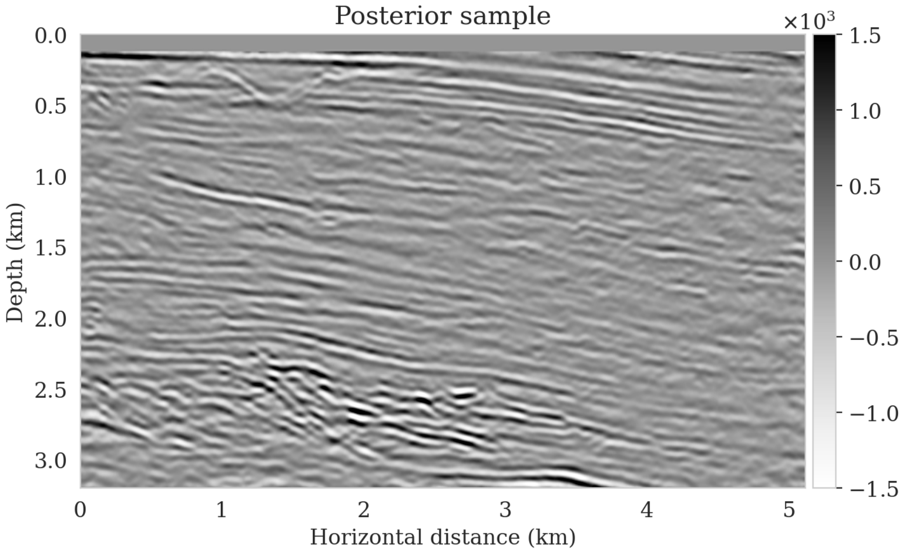
Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

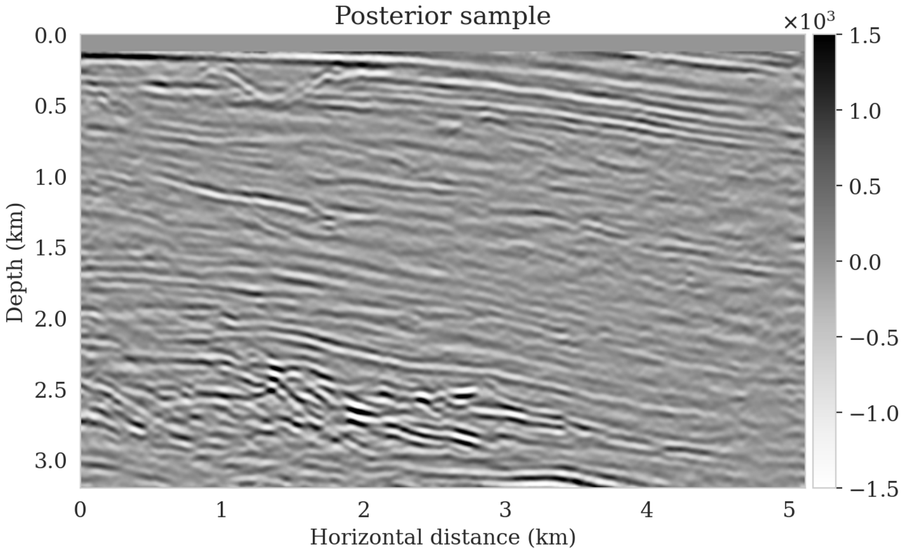

Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

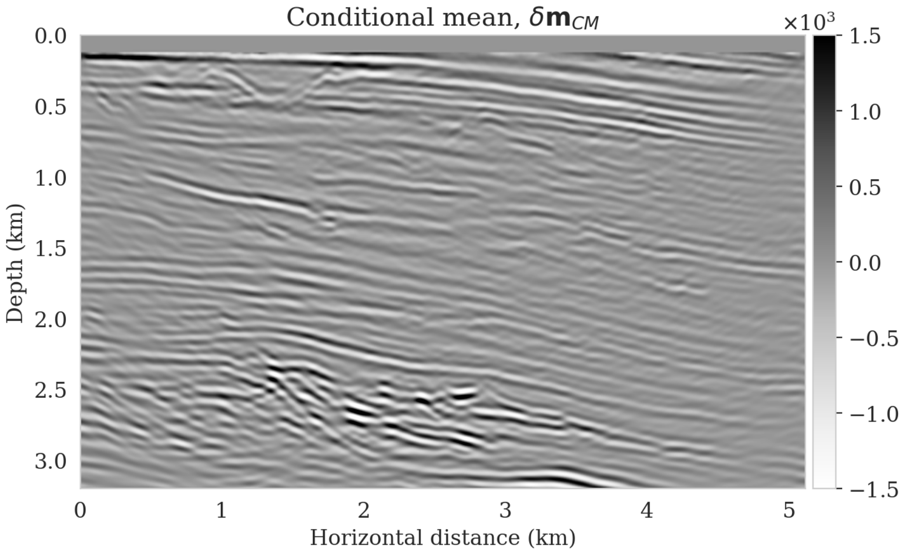
SLIM

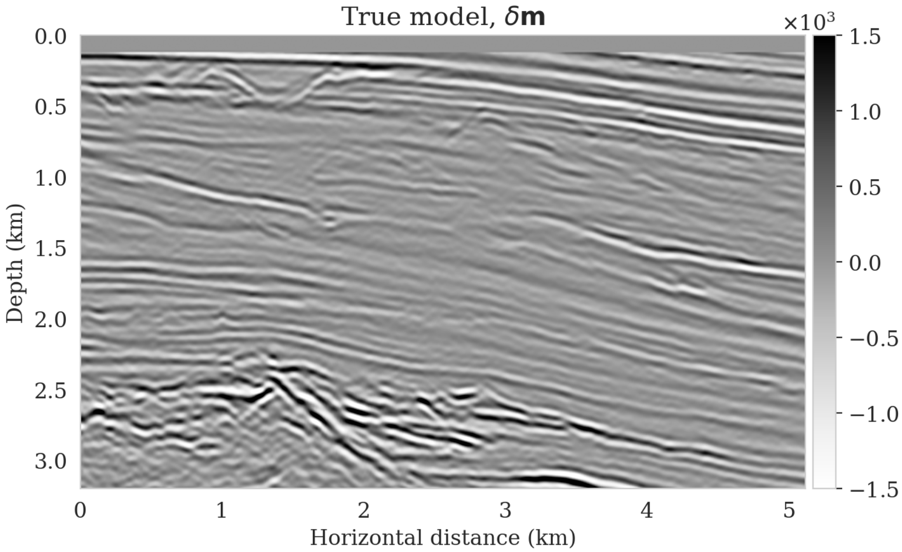

Posterior sample
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)

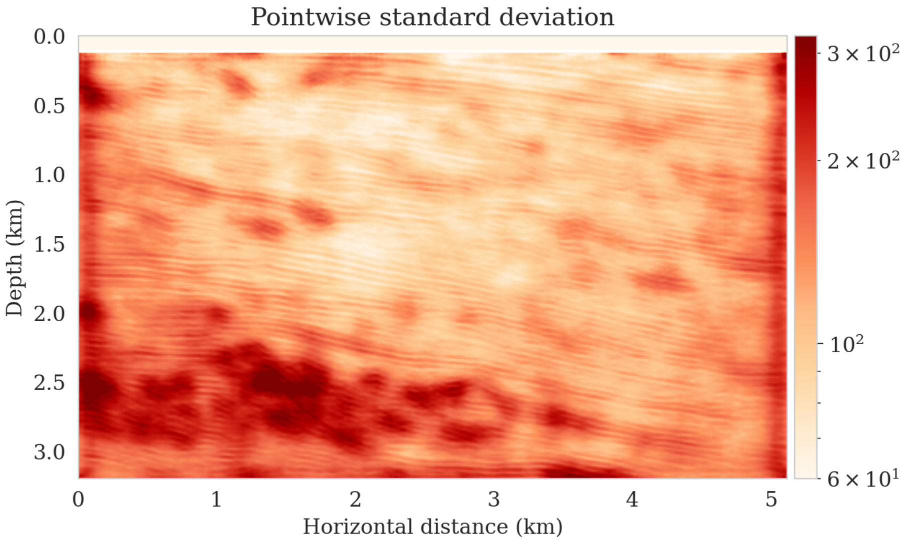

Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
×10^3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Posterior sample
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Depth (km)
0
1
2
3
4
5
Horizontal distance (km)
×10³
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Conditional mean, $\delta\mathbf{m}_{CM}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


True model, $\delta \mathbf{m}$
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Horizontal distance (km)
0
1
2
3
4
5
$\times 10^3$
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5



Pointwise standard deviation
Depth (km)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Horizontal distance (km)
3×10^2
2×10^2
10^2
6×10^1