Variational Inference

Andrew Curtis
Xin Zhang
Atif Nawaz

University of Edinburgh
United Kingdom

Use **Optimisation** to estimate Probability Distributions

Uncertainty = **Family** of all plausible Earth models

- **Variational Inference**
- Travel-Time Tomography
- Application to Grane array data

J Geophys Res, vol.125
Zhang & Curtis, 2020a

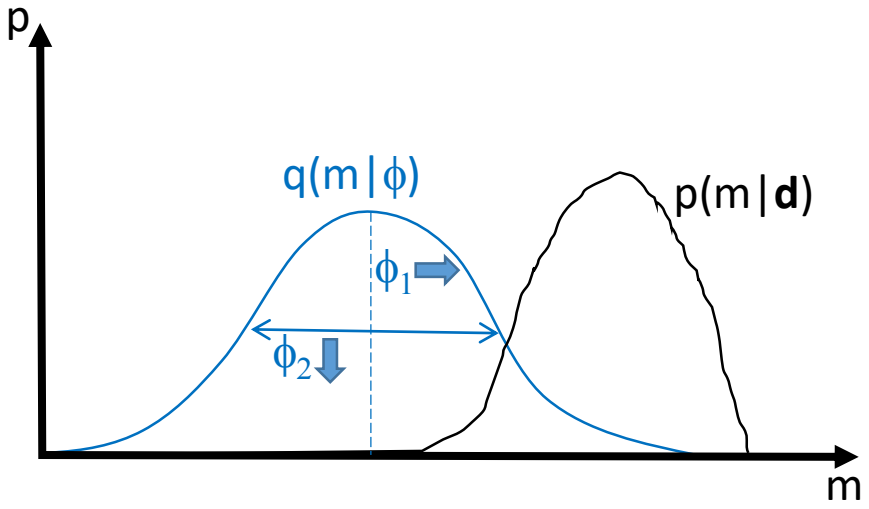
- Variational Full Waveform Inversion

Geophys J Int, vol.222
Zhang & Curtis, 2020b

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



d=data, **m**=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

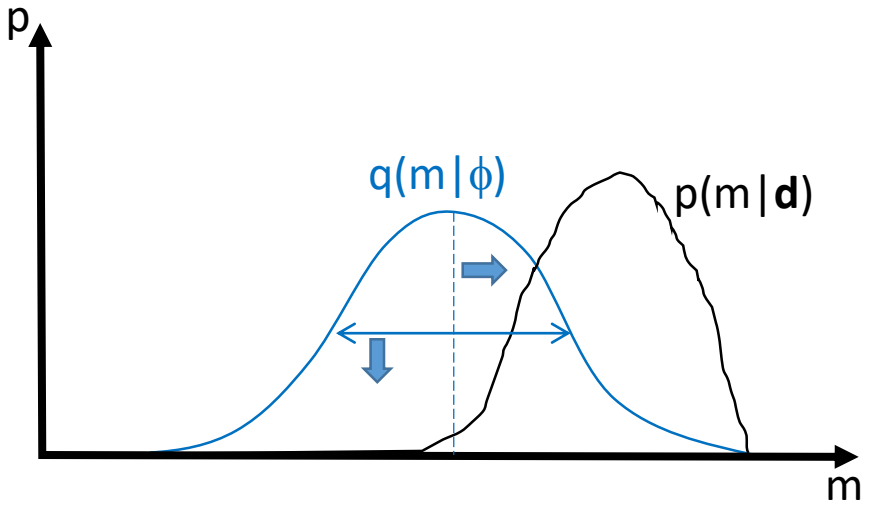
- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$

Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



$\mathbf{d}$=data, $\mathbf{m}$=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

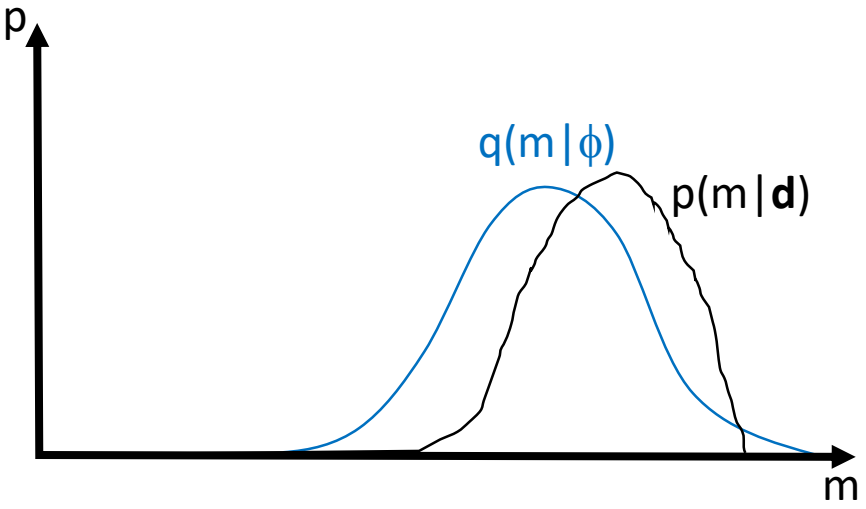
- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$

Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



$\mathbf{d}$=data, $\mathbf{m}$=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

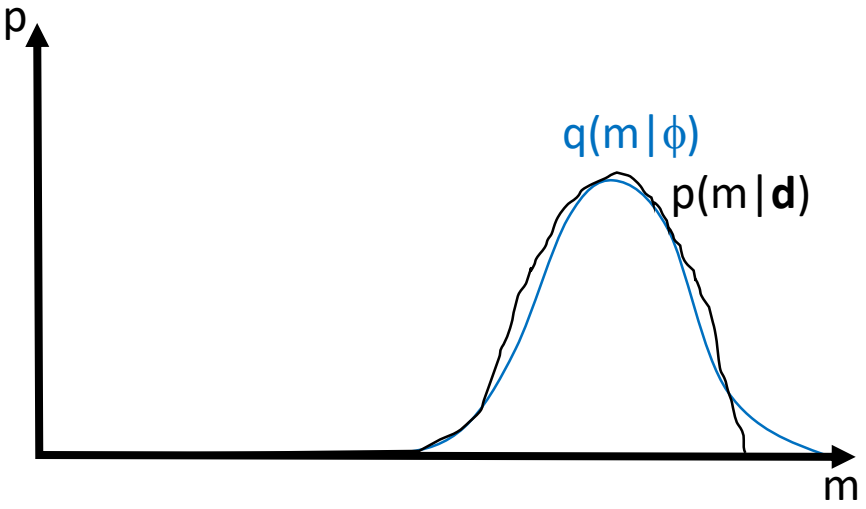
- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$

Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



$\mathbf{d}$=data, $\mathbf{m}$=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

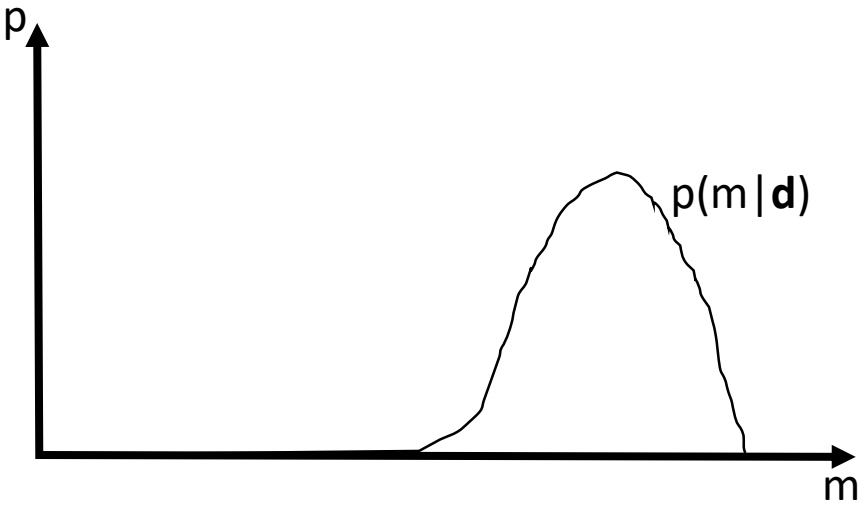
- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



d=data, **m**=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

- Strategy 2: generate a set of **samples** of $p(\mathbf{m}|\mathbf{d})$ by **optimisation**

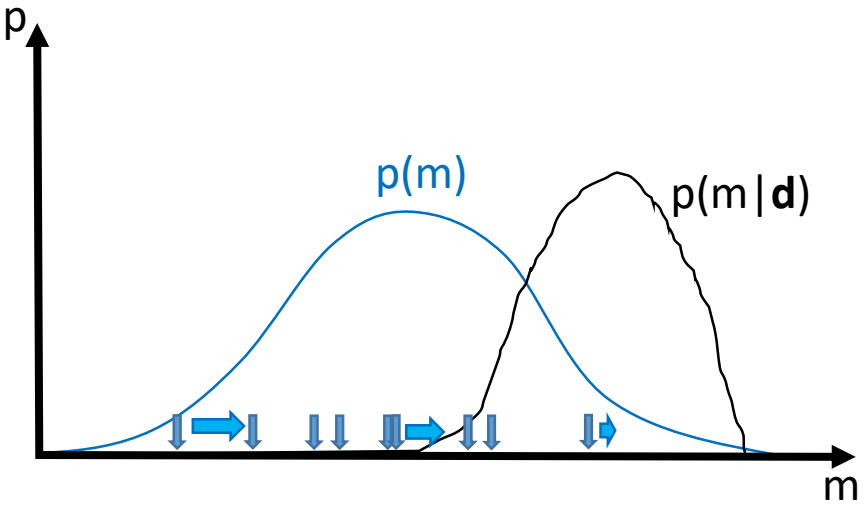
  $\sim q$

  Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



d=data, **m**=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

- Strategy 2: generate a set of **samples** of $p(\mathbf{m}|\mathbf{d})$ by **optimisation**

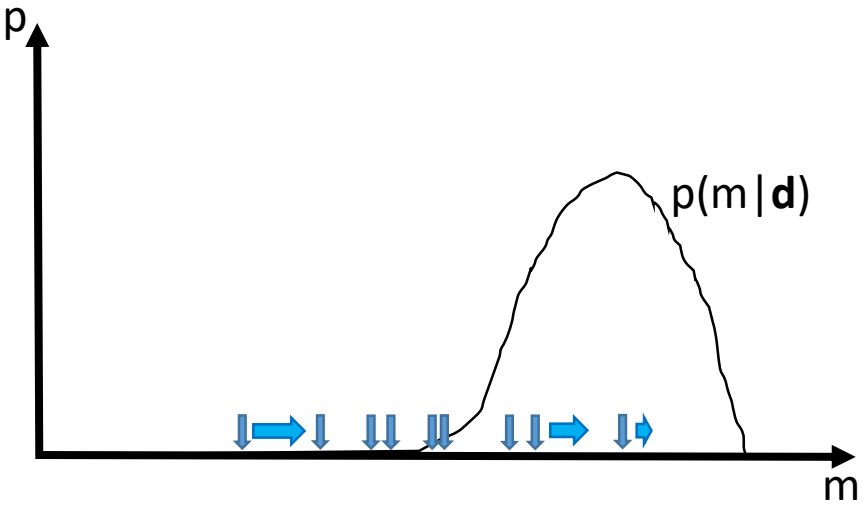
  $\sim q$

  Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



d=data, **m**=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

- Strategy 2: generate a set of **samples** of $p(\mathbf{m}|\mathbf{d})$ by **optimisation**

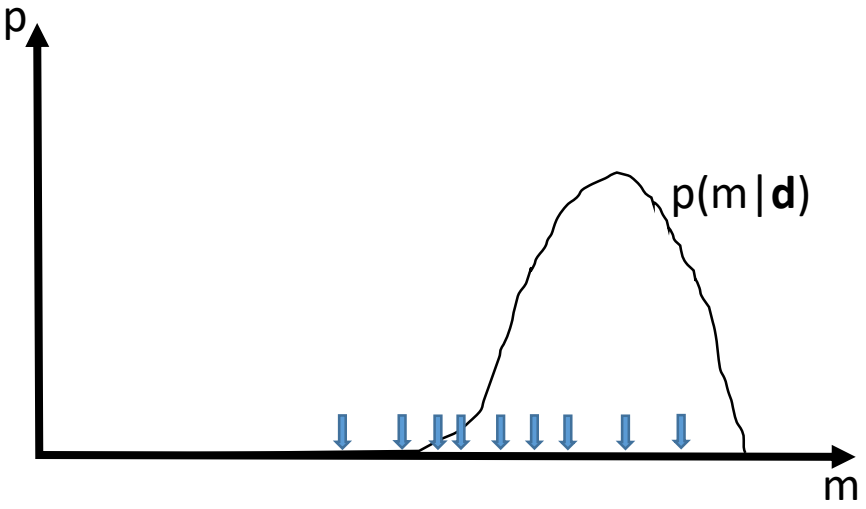
  $\sim q$

  Zhang & Curtis, 2020a

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



d=data, **m**=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

- Strategy 2: generate a set of **samples** of $p(\mathbf{m}|\mathbf{d})$ by **optimisation**

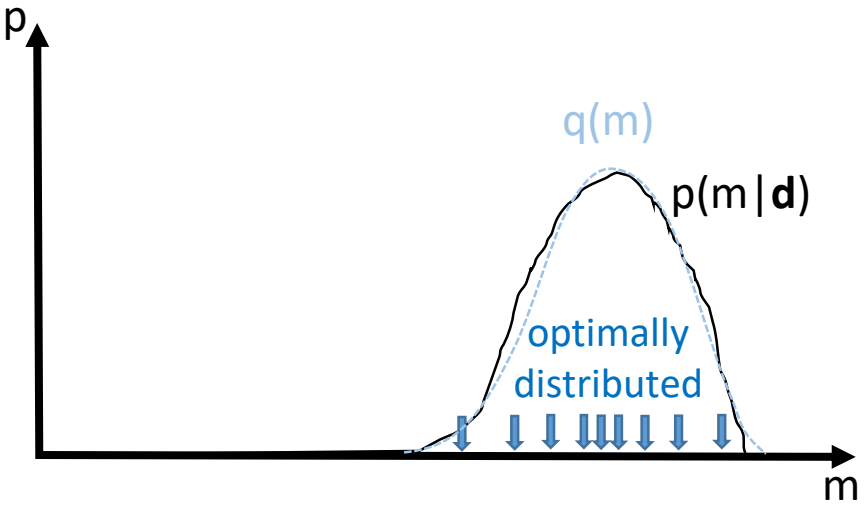
  Zhang & Curtis, 2020a

$$\sim q$$

Variational Inference

- Bayesian solution

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$



$\mathbf{d}$=data, $\mathbf{m}$=parameters

- Variational methods replace **stochastic sampling** with **optimisation** of functions

- Strategy 1: **fit** semi-analytic functions to $p(\mathbf{m}|\mathbf{d})$
  - Choose family of functions $q(\mathbf{m}, \varphi)$, φ=parameters [c.f. φ=Gaussian mean & covar]
  - Optimise φ s.t. $q(\mathbf{m}|\varphi) \cong p(\mathbf{m}|\mathbf{d})$
- Define a measure of difference between q and p; then minimize it.

- Strategy 2: generate a set of **samples** of $p(\mathbf{m}|\mathbf{d})$ by **optimisation**

$\sim q$

Zhang & Curtis, 2020a

Variational Inference

- **Kullback-Liebler divergence** measures difference between q and p:

log(evidence) : intractable

$$KL[q||p] = E_{q(\mathbf{m})}[\log q(\mathbf{m}|\varphi)] - E_{q(\mathbf{m})}[\log p(\mathbf{m}|\mathbf{d})] + \log p(\mathbf{d})$$

- KL ≥ 0 and KL $= 0$ when $q = p$. Rearrange…

log(evidence) : constant w.r.t. q

$$KL[q||p] + E_{q(\mathbf{m})}[\log p(\mathbf{m}|\mathbf{d})] - E_{q(\mathbf{m})}[\log q(\mathbf{m}|\varphi)] = \log p(\mathbf{d})$$

- Evidence lower bound (ELBO)

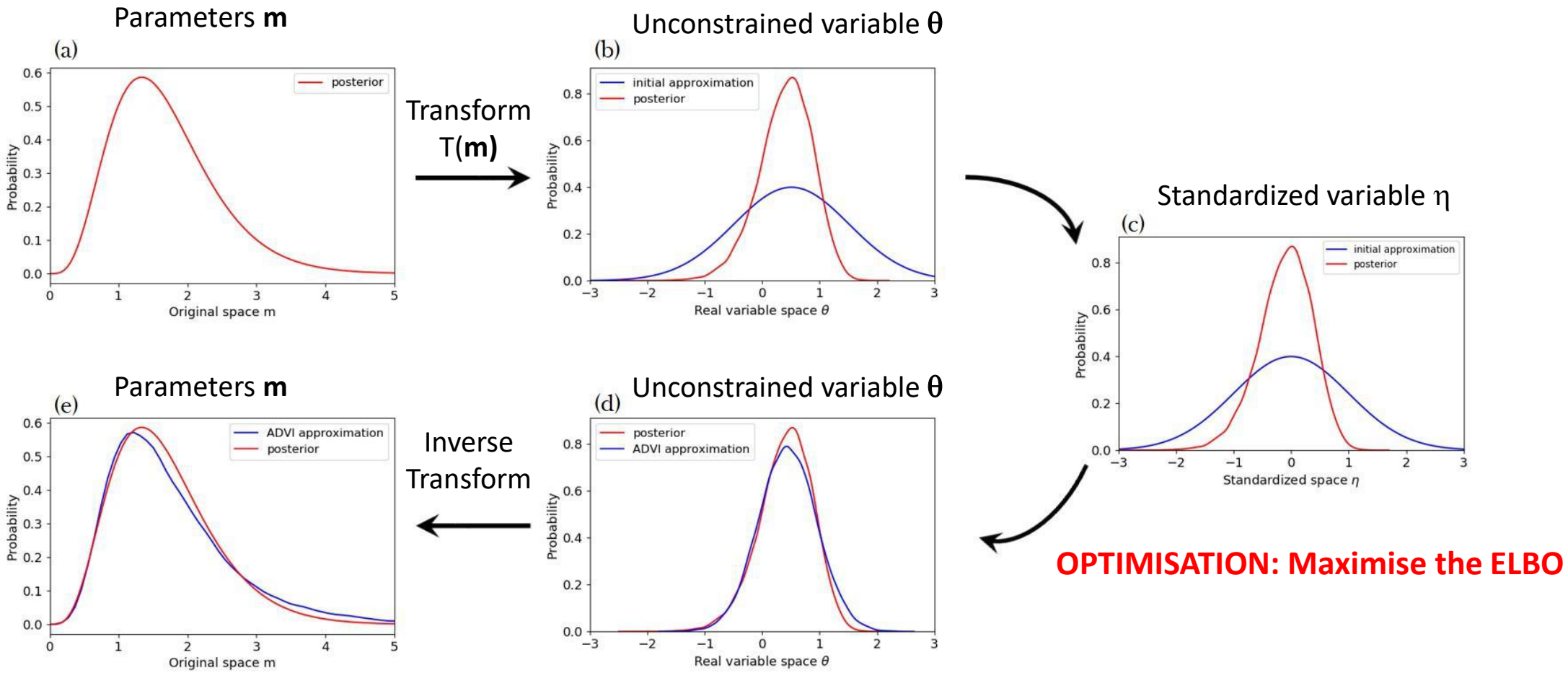
Bayes Rule, prior, likelihood
→ Efficient, case-specific analytical methods (Nawaz & Curtis 2018/19/20)

$$ELBO(q) = E_{q(\mathbf{m})}[\log p(\mathbf{m}|\mathbf{d})] - E_{q(\mathbf{m})}[\log q(\mathbf{m}|\varphi)]$$

Expectations w.r.t. q which we choose

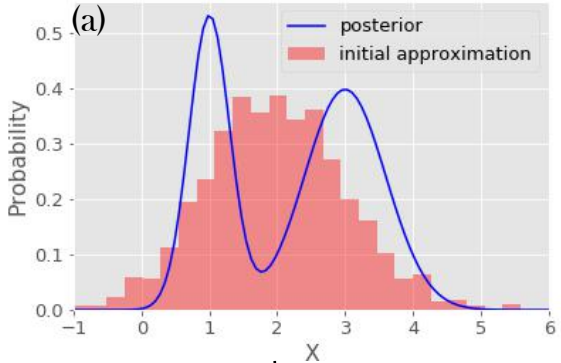
Maximise ELBO ➔ Minimise KL divergence (difference) between q and p.

Automatic Differential Variational Inference (ADVI)



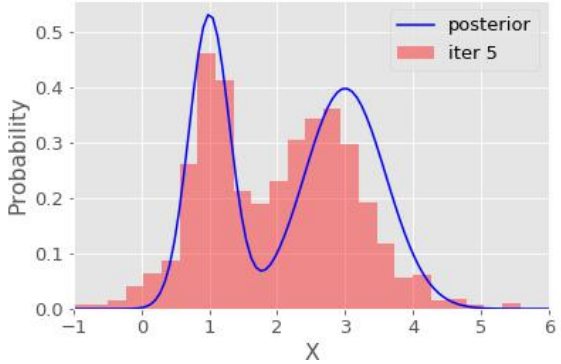
Stein Variational Gradient Descent(SVGD)

- Target probability
- Histogram of particles



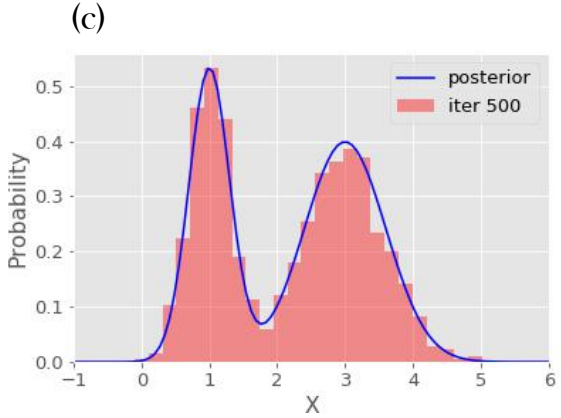
$$T(x) = x + \varepsilon\phi^*(x)$$

OPTIMISATION: Maximise the ELBO

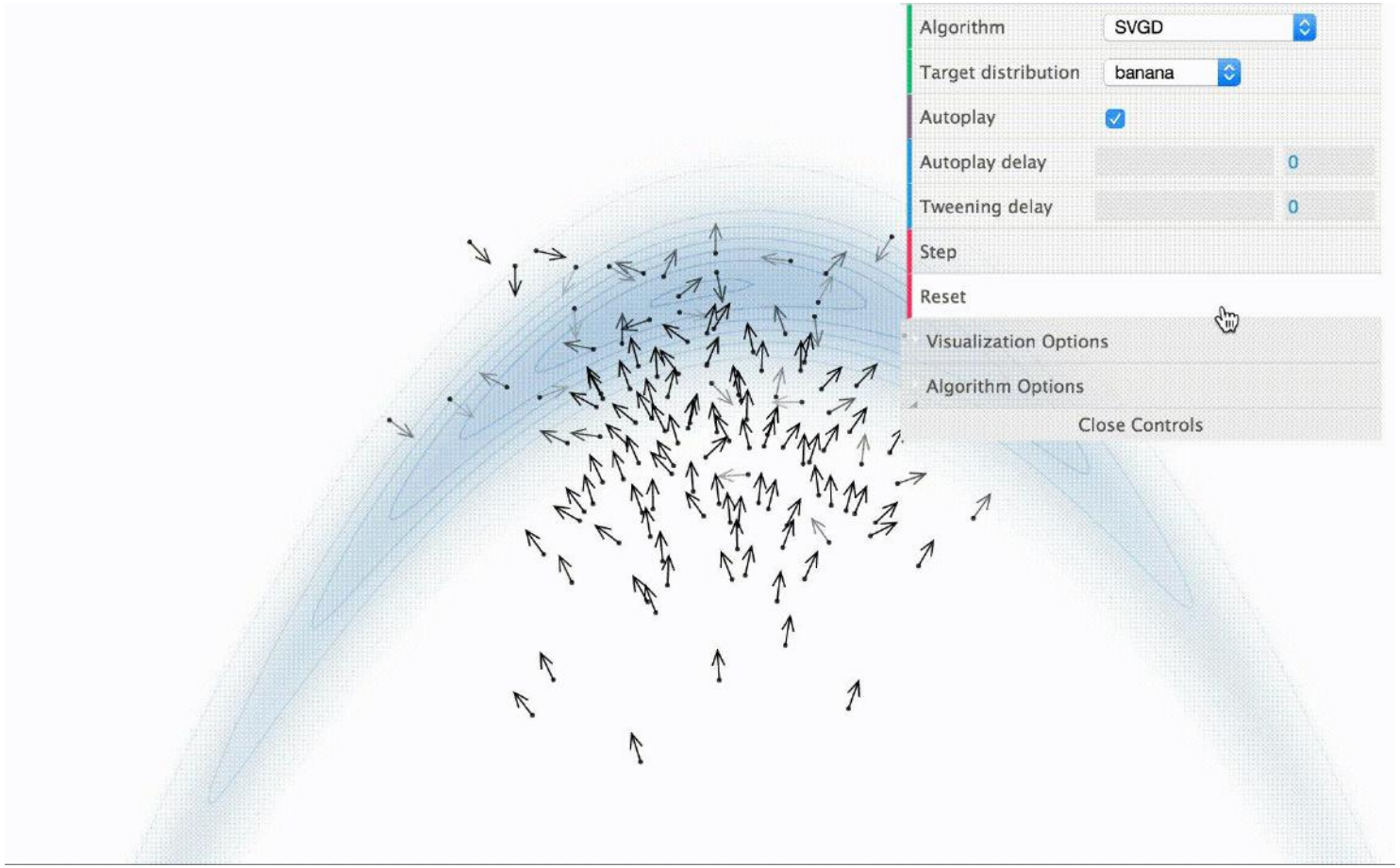


$$T(x) = x + \varepsilon\phi^*(x)$$

OPTIMISATION: Maximise the ELBO

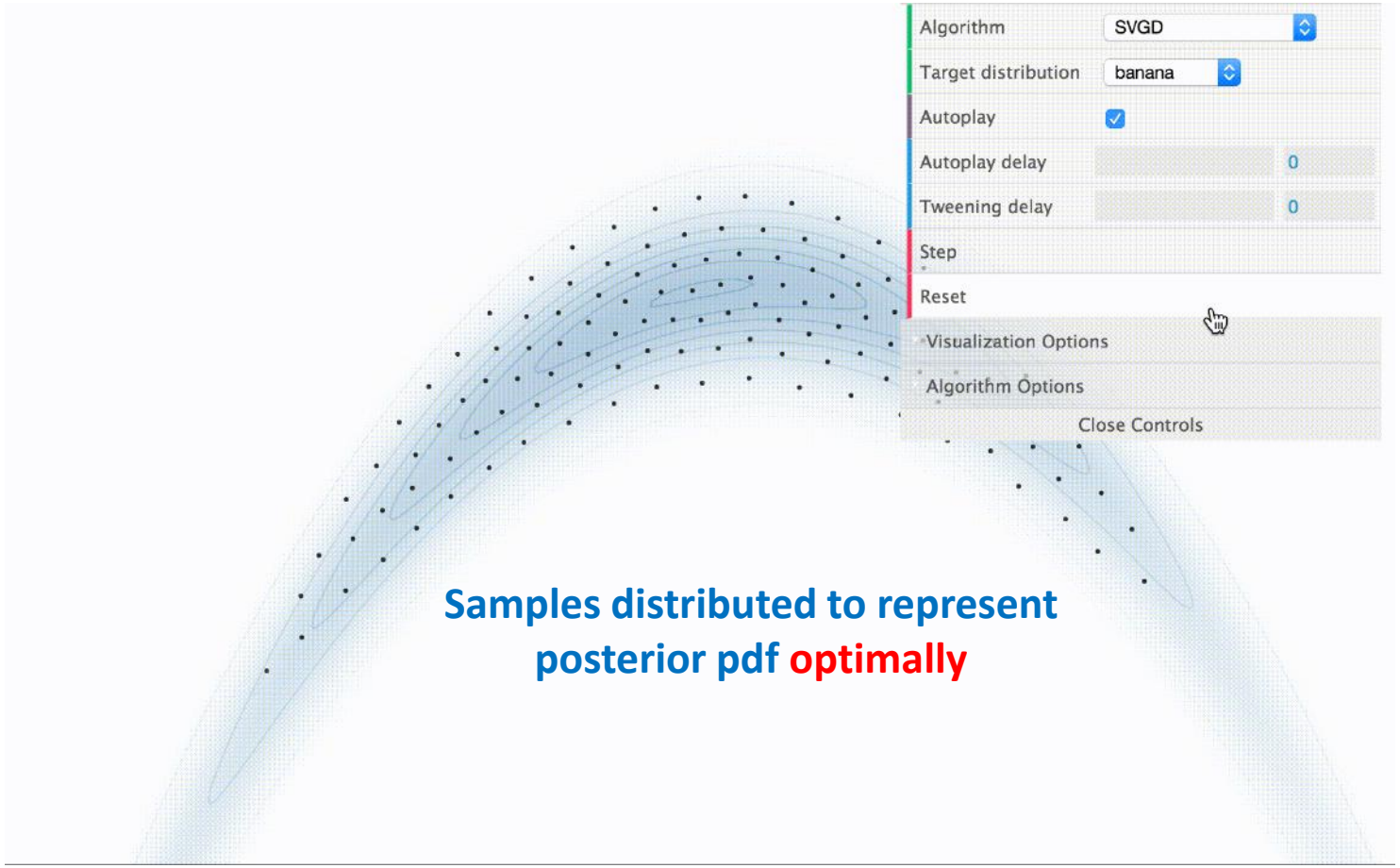


SVGD



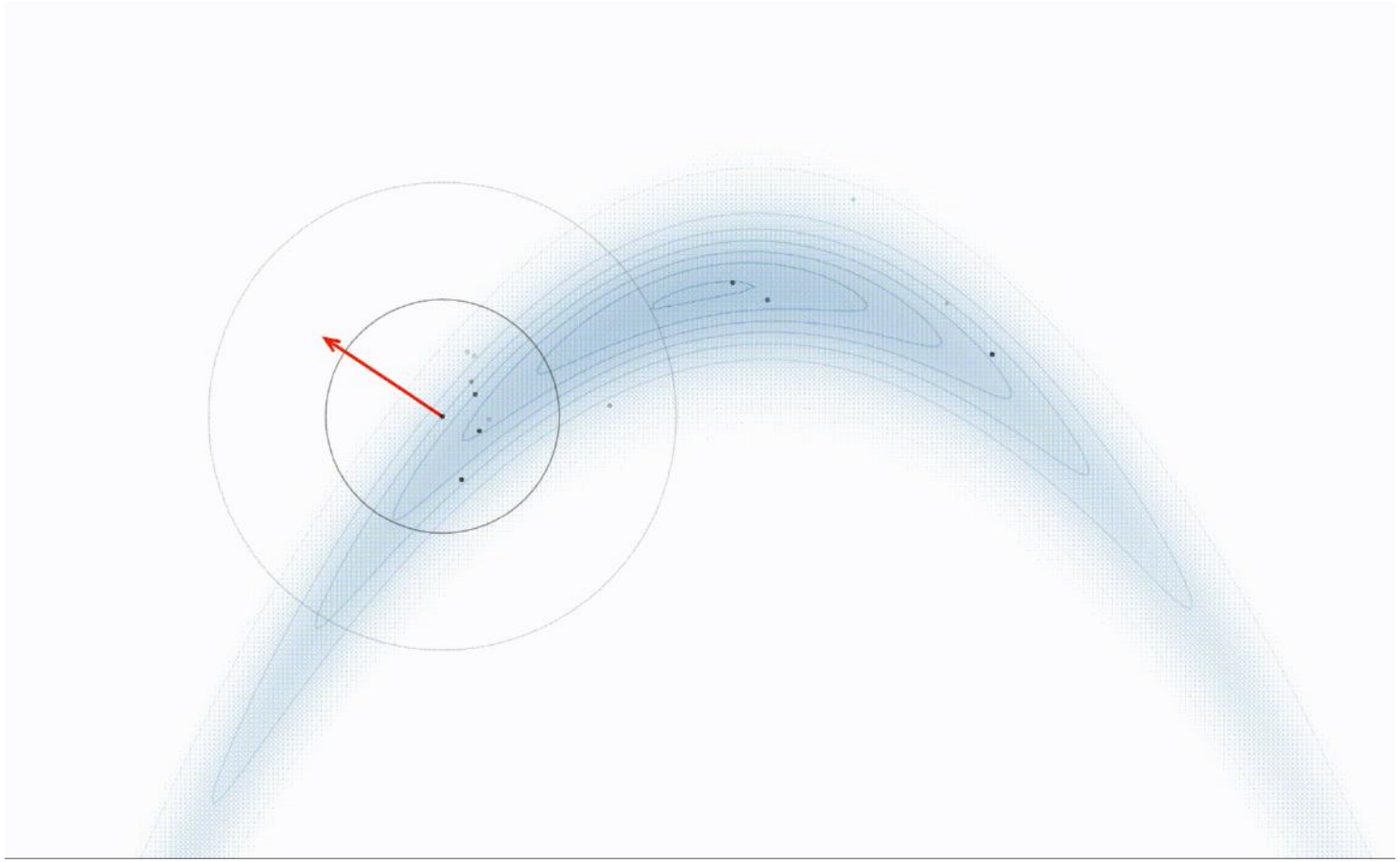
https://chi-feng.github.io/mcmc-demo/app.html

SVGD



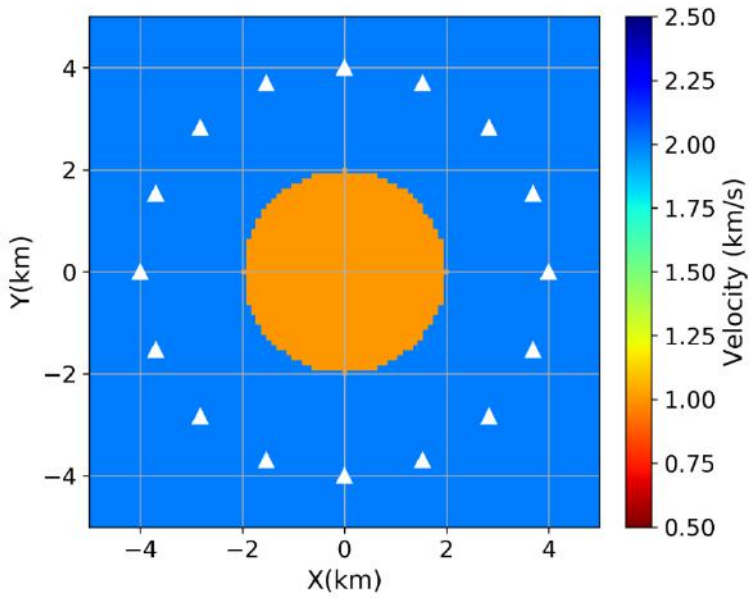
https://chi-feng.github.io/mcmc-demo/app.html

Metropolis-Hastings Monte Carlo



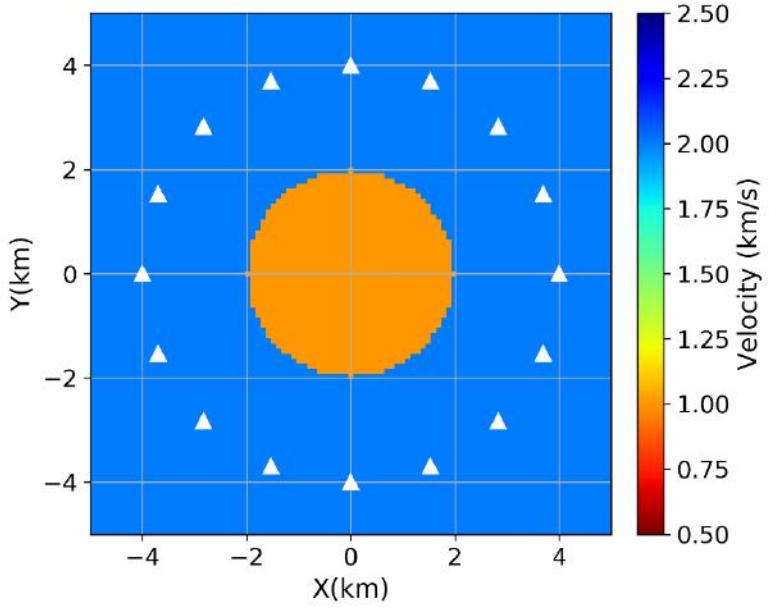
https://chi-feng.github.io/mcmc-demo/app.html

Synthetic tests

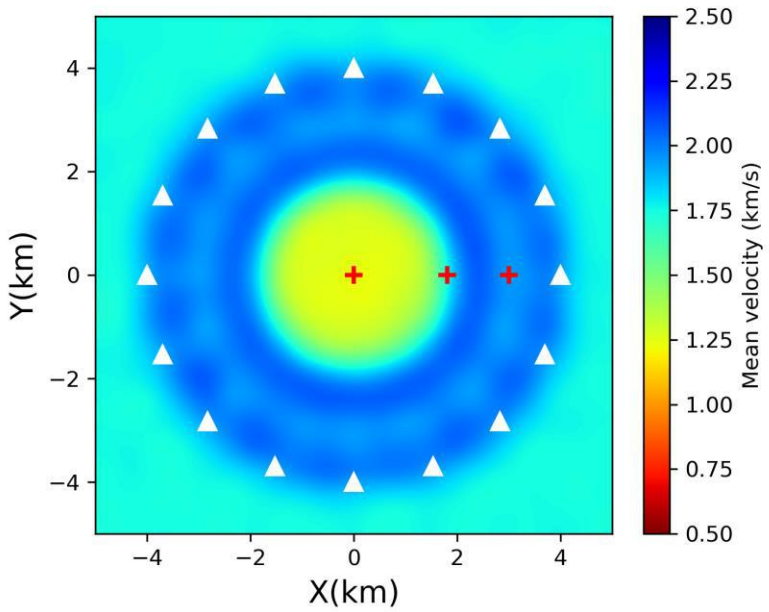


ADVI results

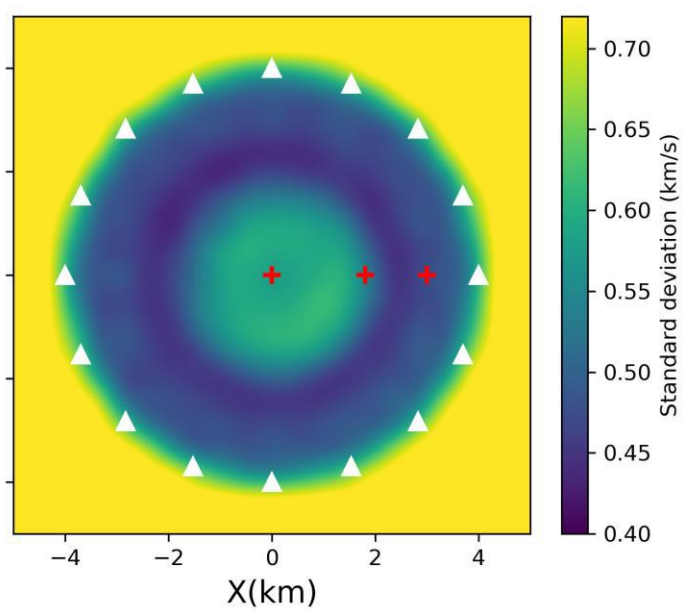
True model



Mean



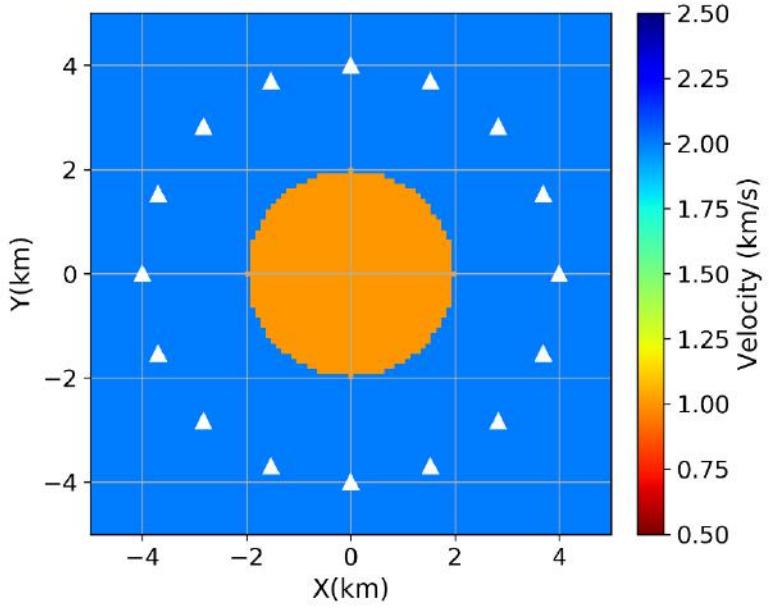
Stdev



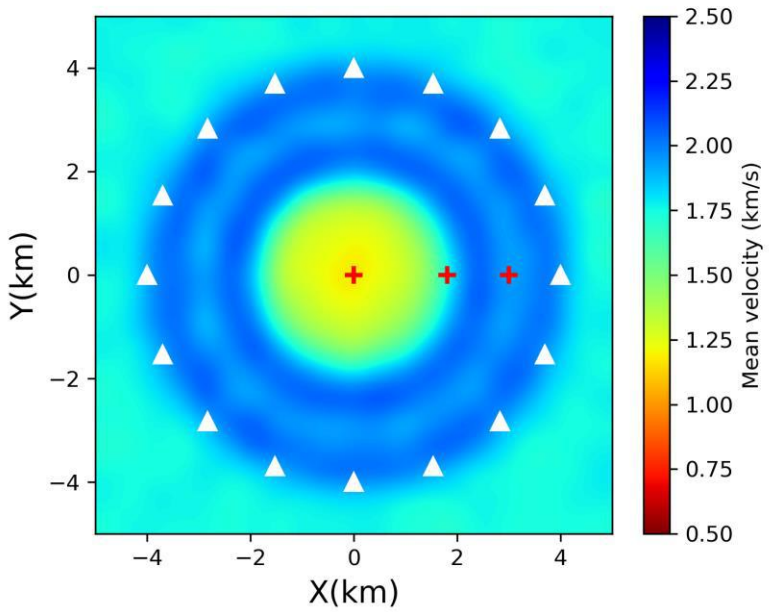
Parameterized by a 21*21 grid

SVGD results

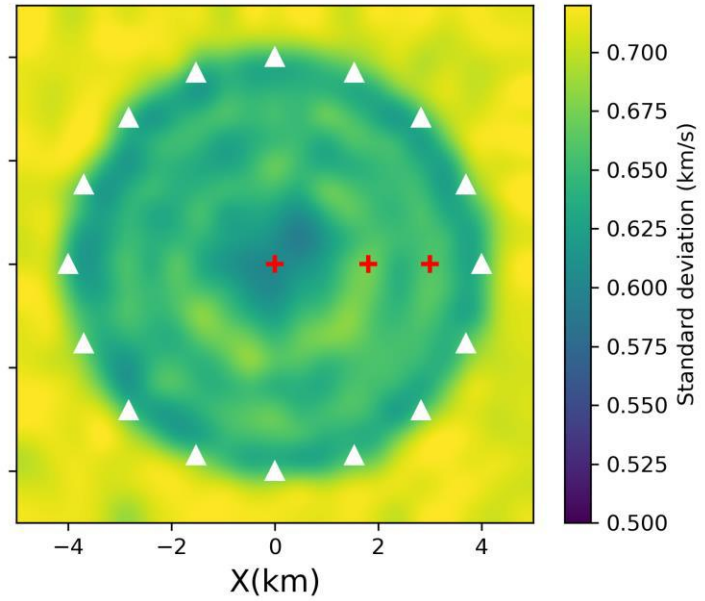
True model



Mean



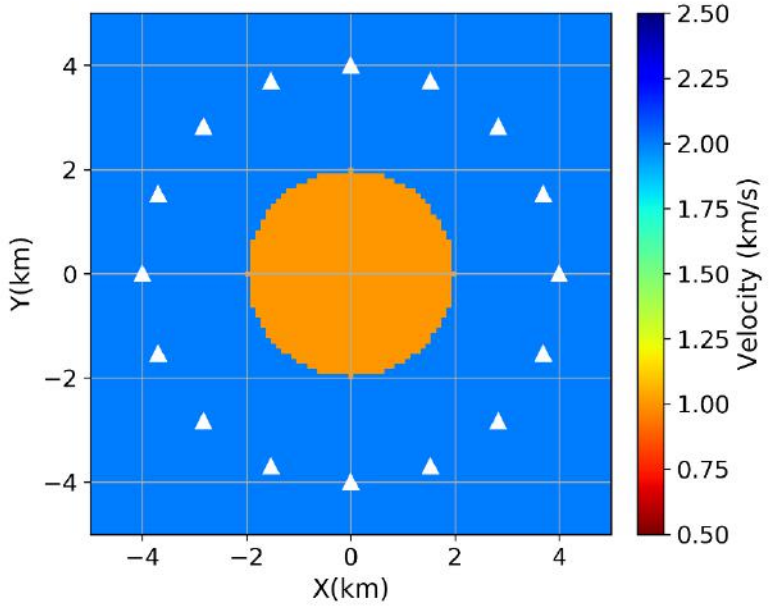
Stdev



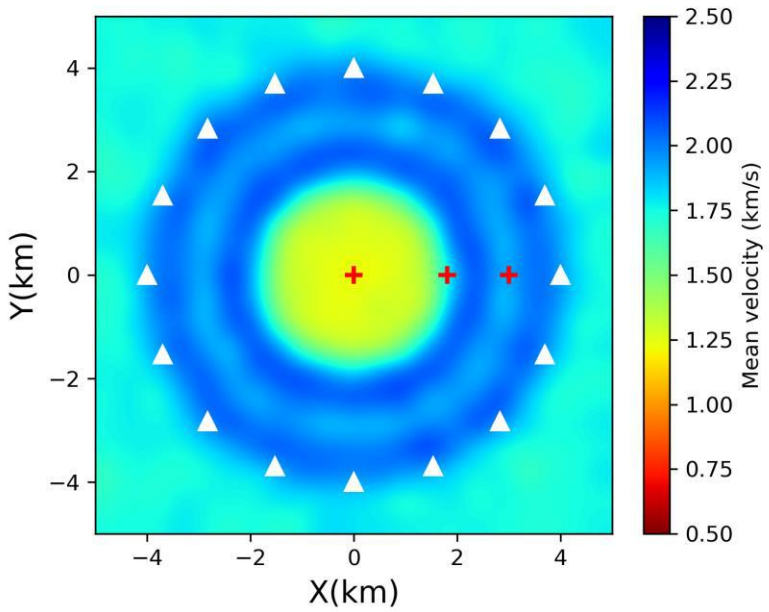
Parameterized by a 21*21 grid

Metropolis-Hastings McMC

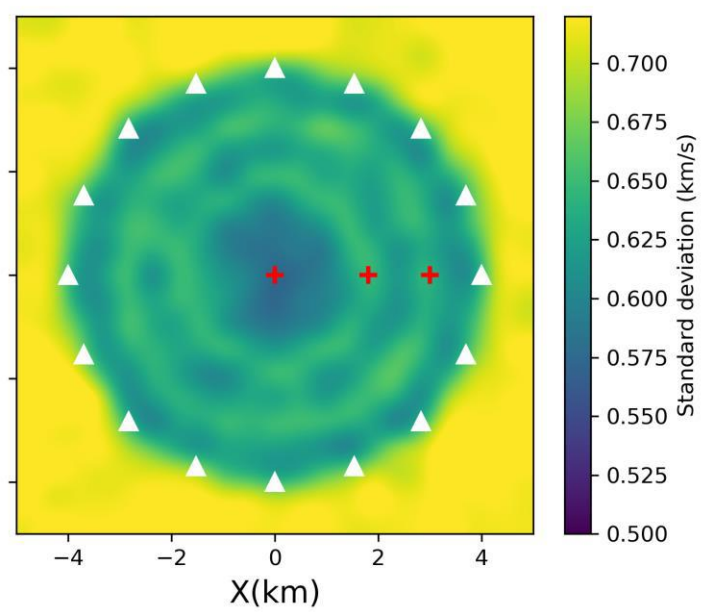
True model



Mean



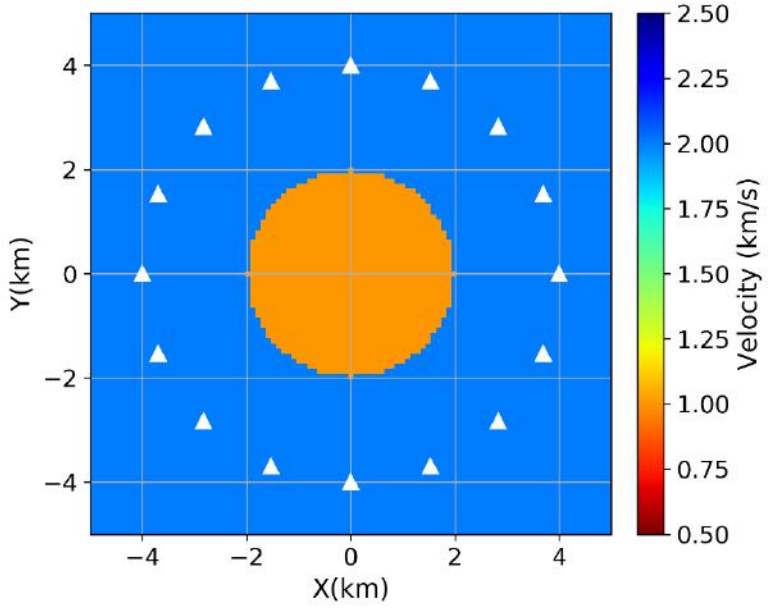
Stdev



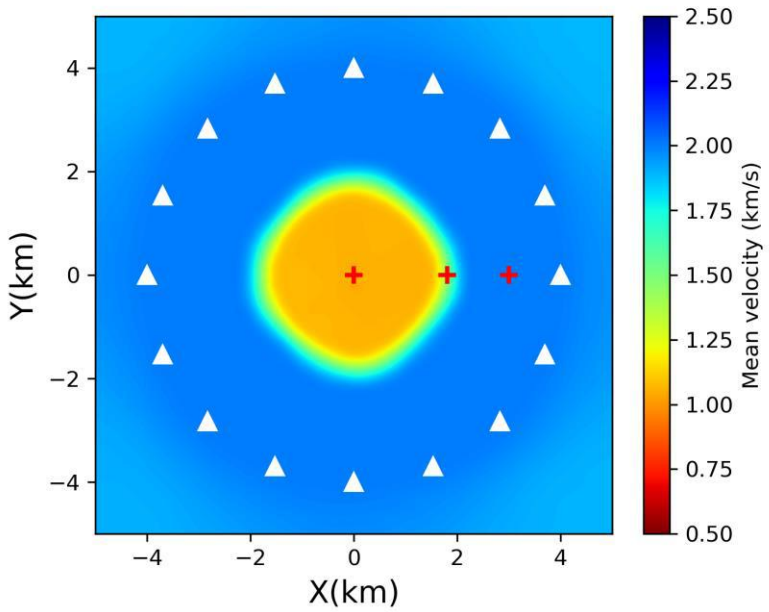
Parameterized by a 21*21 grid

Reversible jump McMC

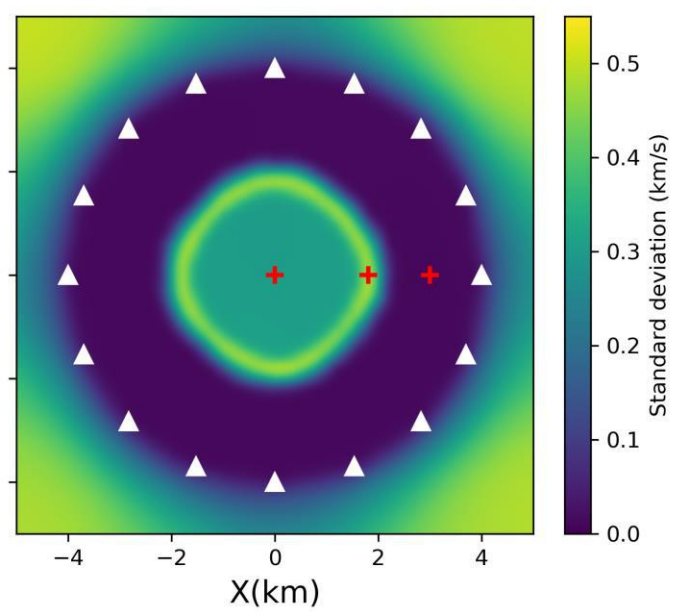
True model



Mean



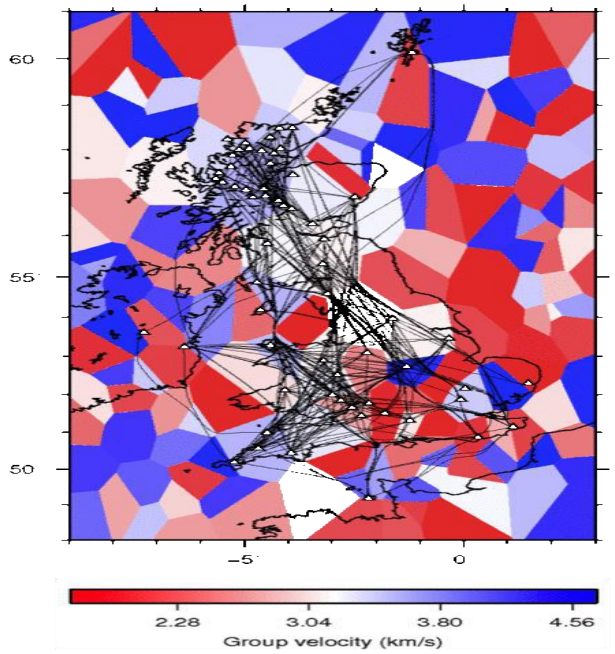
Stdev



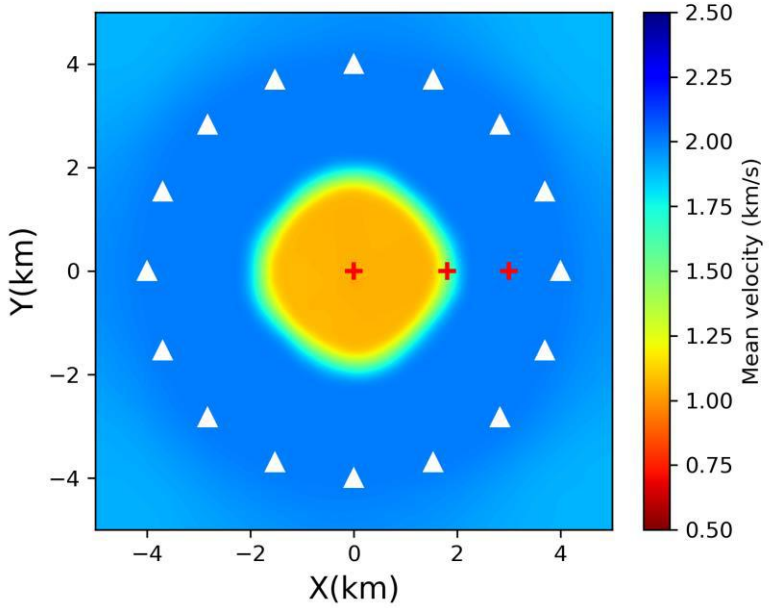
Parameterized by Voronoi cells

Reversible jump McMC

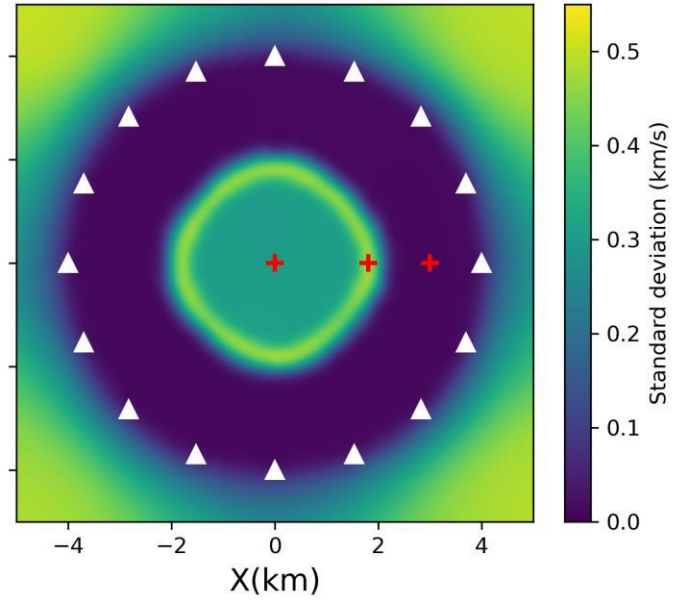
True model



Mean



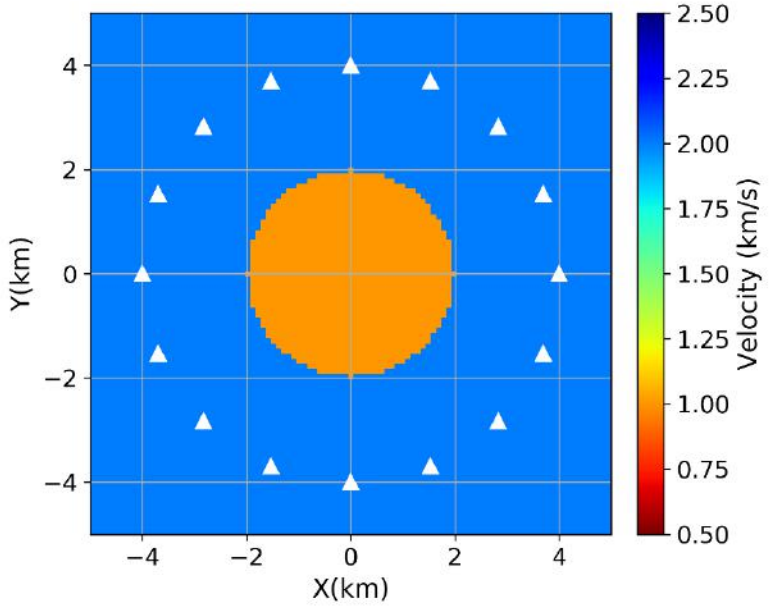
Stdev



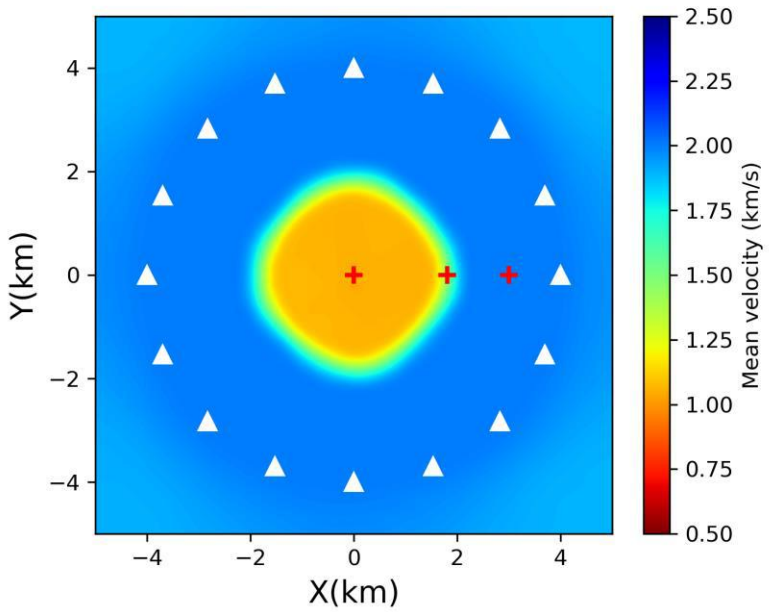
Parameterized by Voronoi cells

Reversible jump McMC

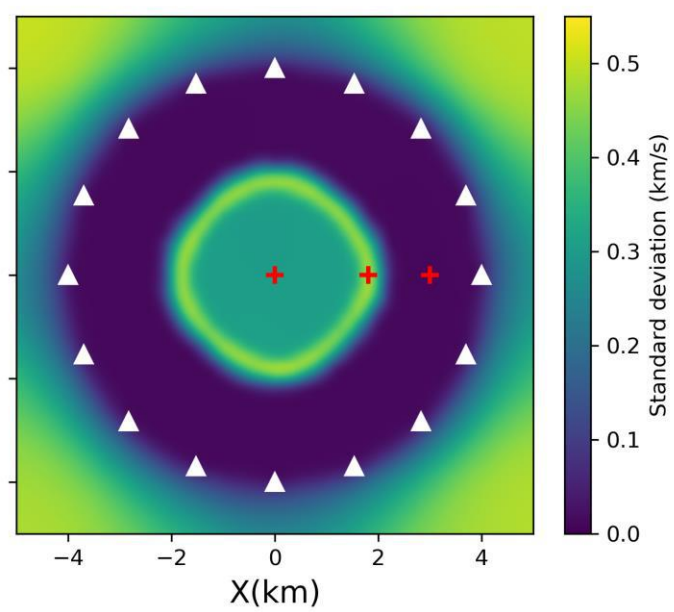
True model



Mean



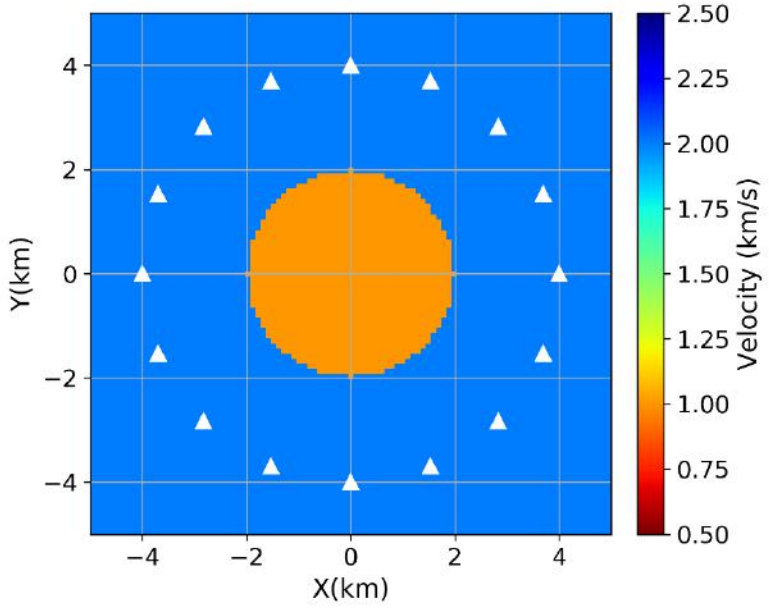
Stdev



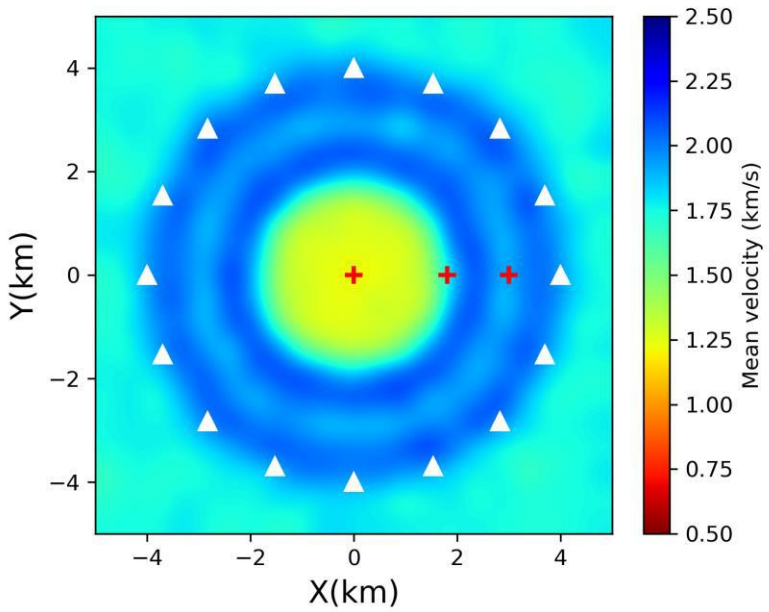
Parameterized by Voronoi cells

Metropolis-Hastings McMC

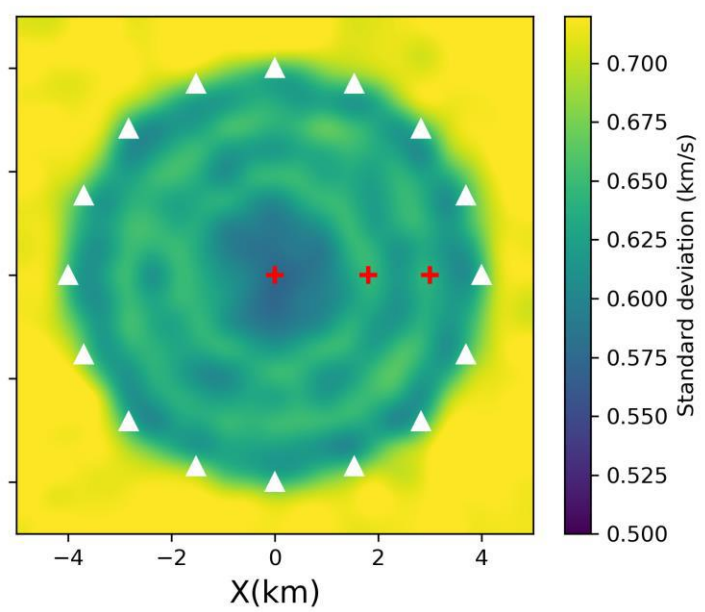
True model



Mean



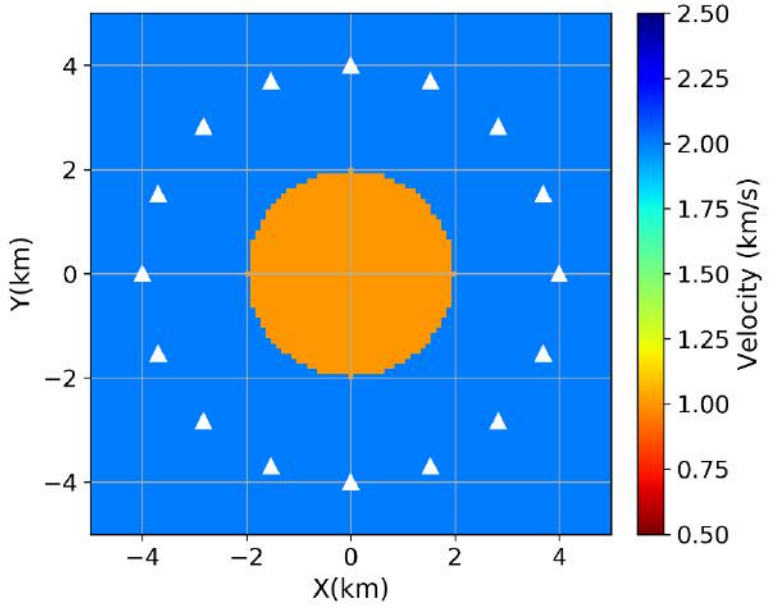
Stdev



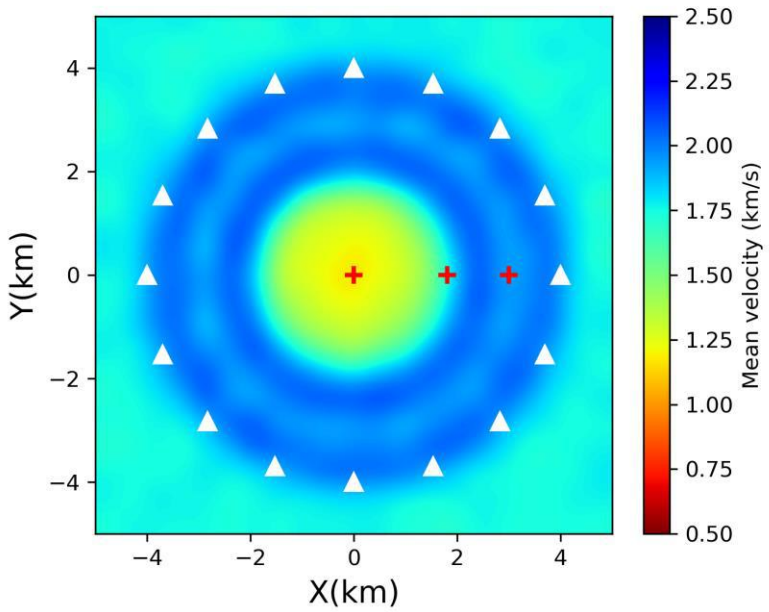
Parameterized by a 21*21 grid

SVGD results

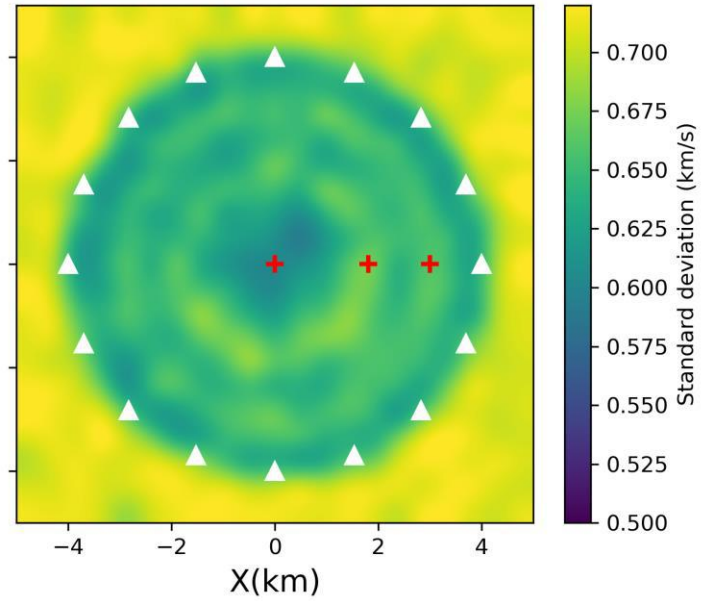
True model



Mean



Stdev



Parameterized by a 21*21 grid

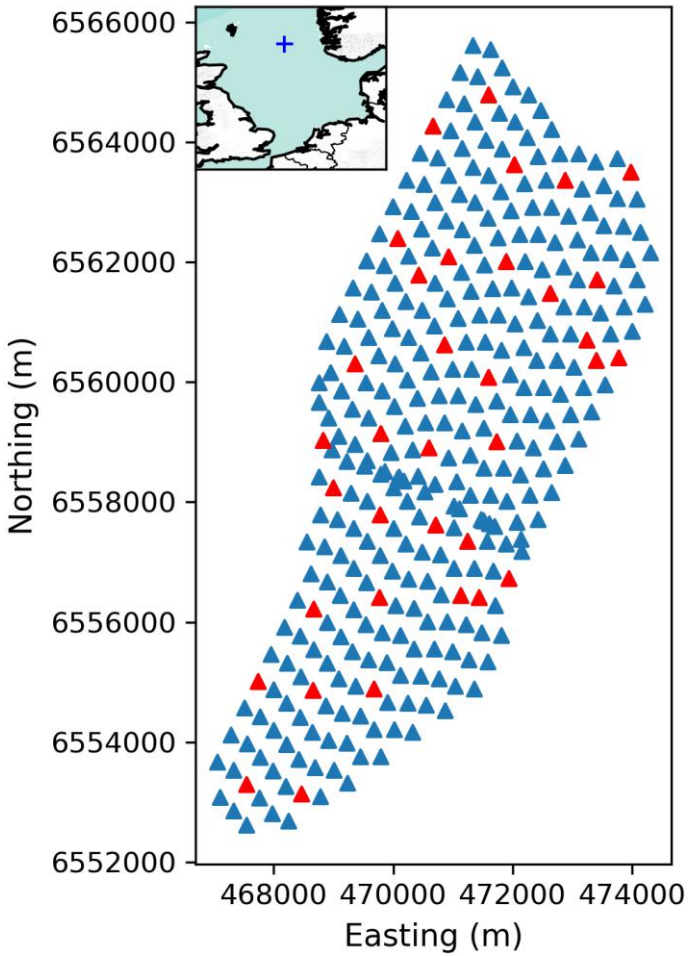
Computational cost

Methods	Number of simulations	CPU hours	Real time (hours)
ADVI	10,000	0.45	0.45
SVGD	400,000	8.53	0.97
MH-McMC	12,000,000	410.3	68.4
Rj-McMC	3,000,000	102.6	17.1

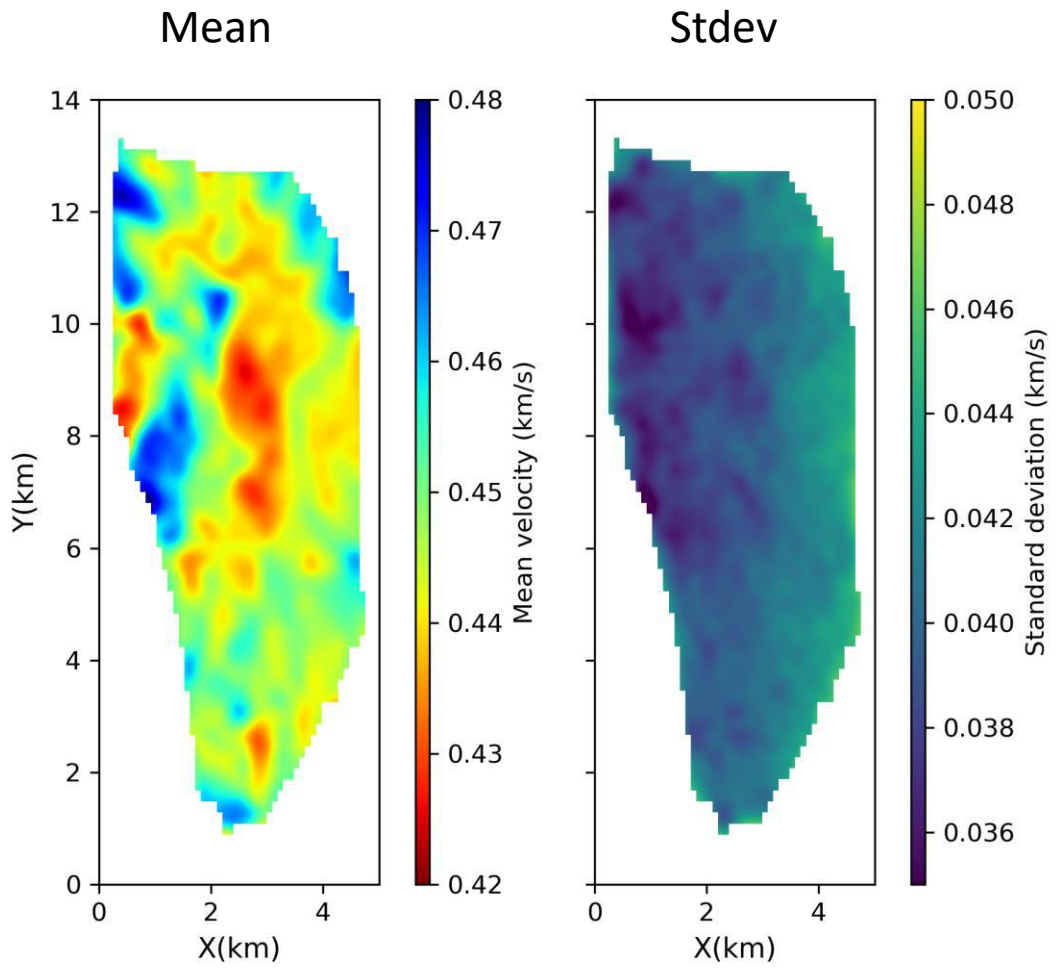
Application to Grane field

346 receivers

35 virtual sources

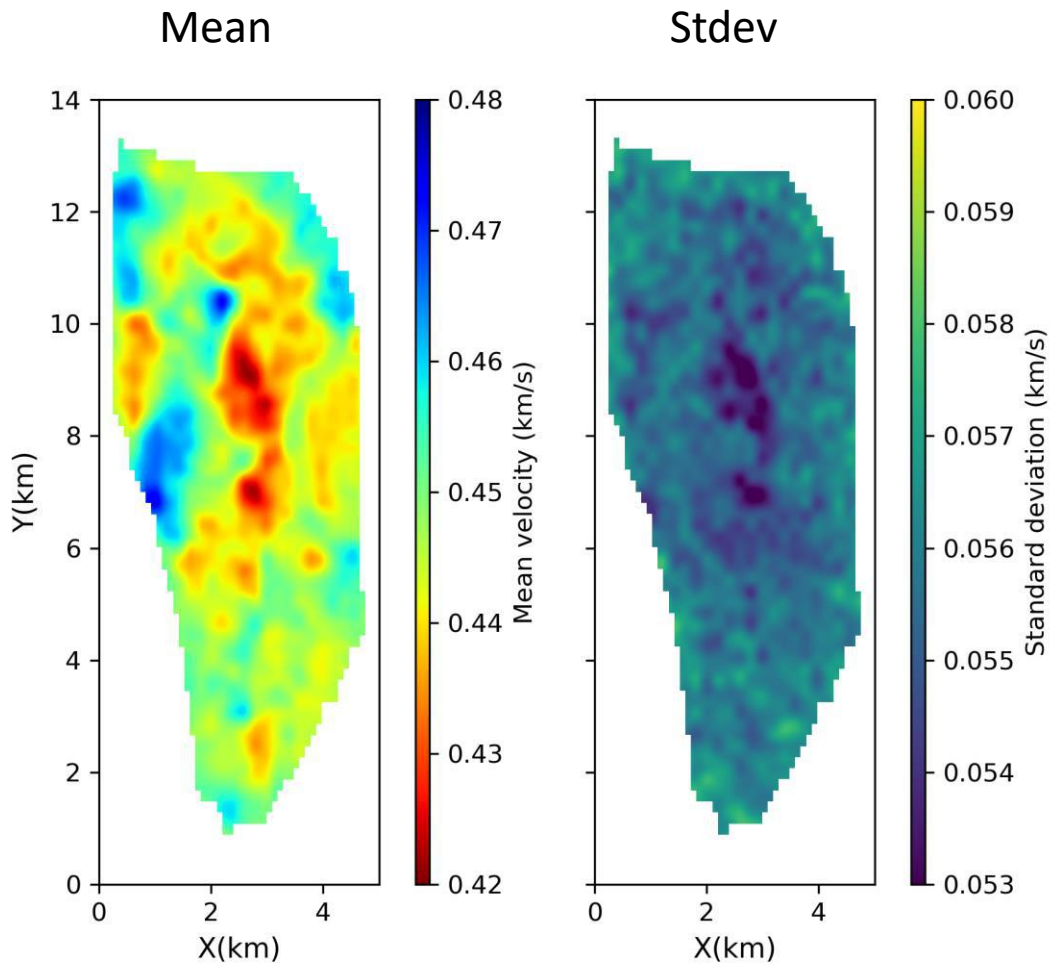


ADVI



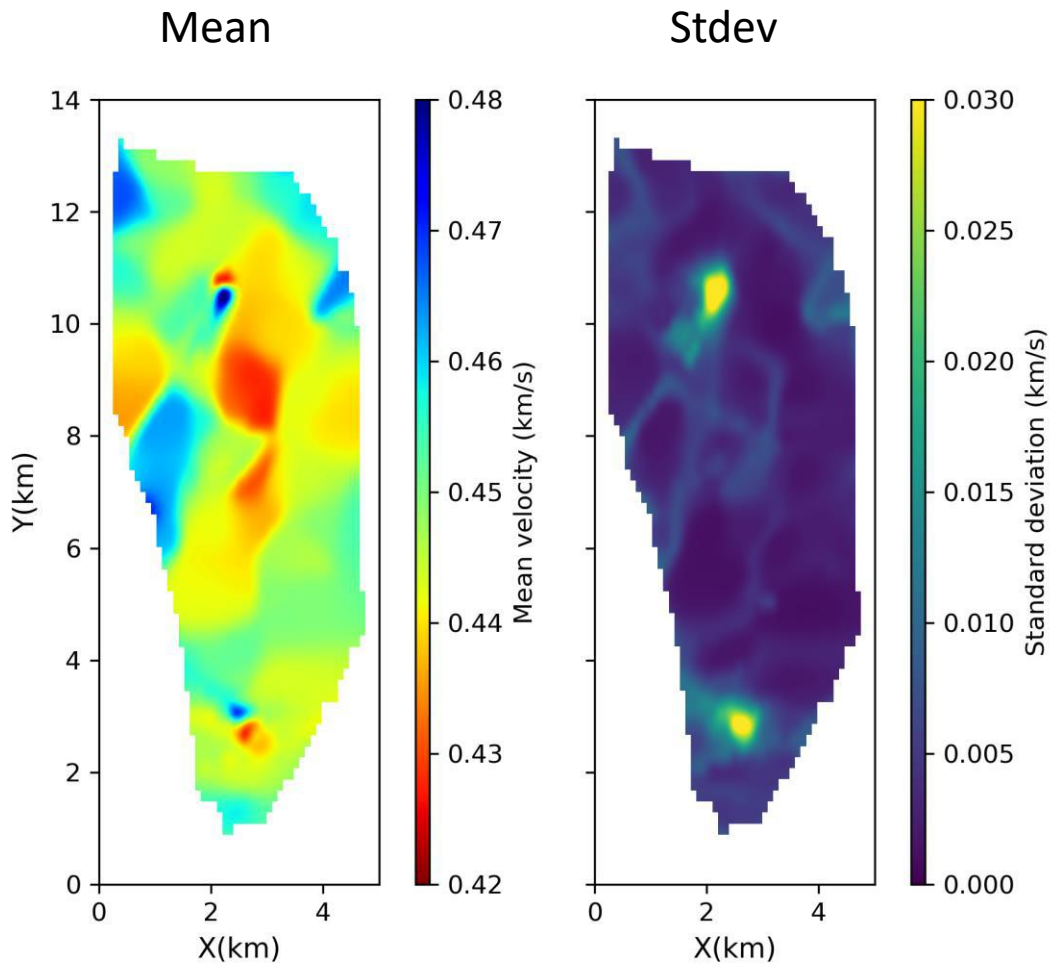
10,000 forward models, 5.1 hours

SVGD



500,000 forward models, 12.1 hours parallelized using 12 cores

rj-McMC

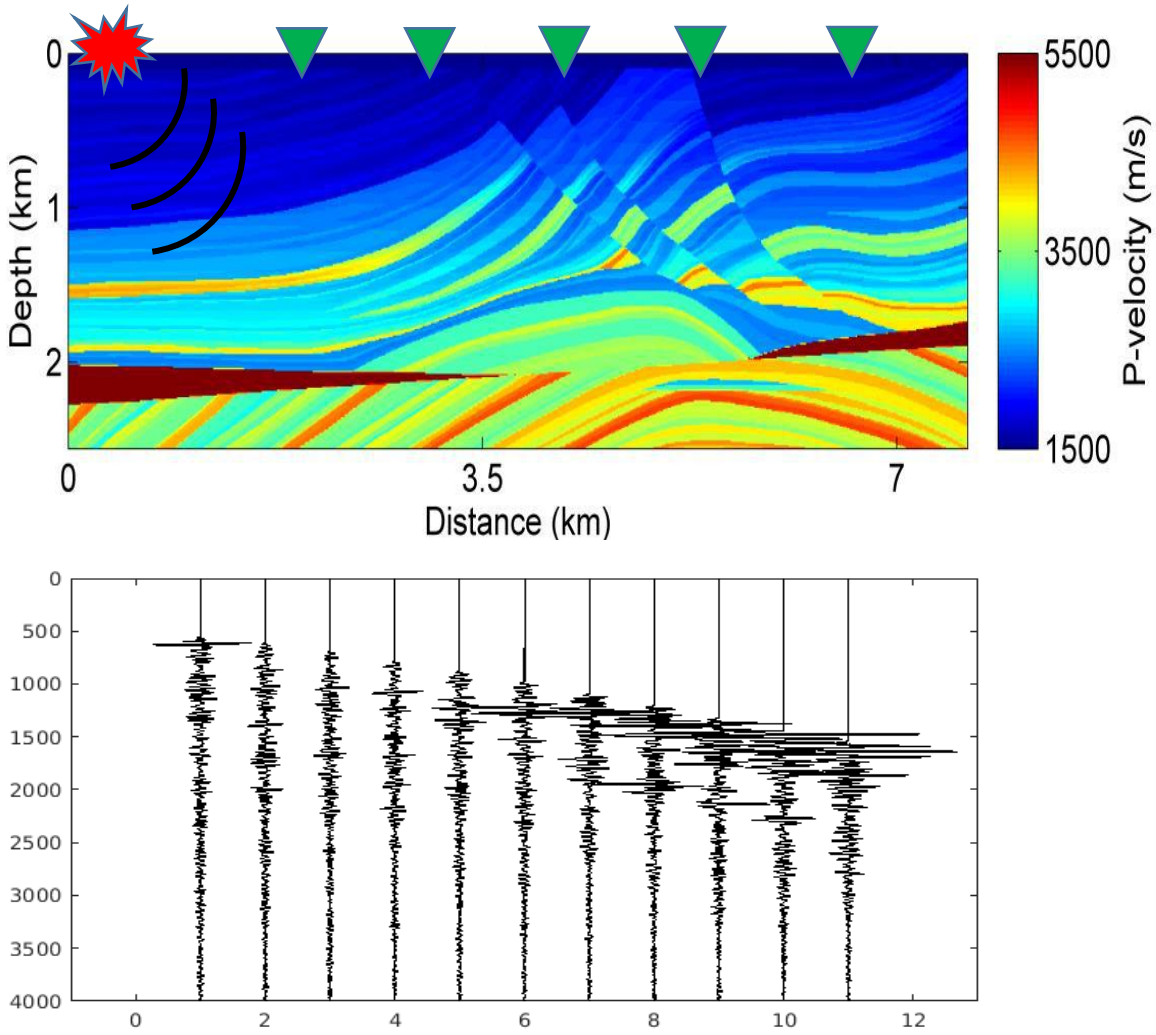


12,800,000 forward models, 5 days running on 16 cores

Variational Full-Waveform Inversion

Full-waveform Inversion

Forward



Inversion

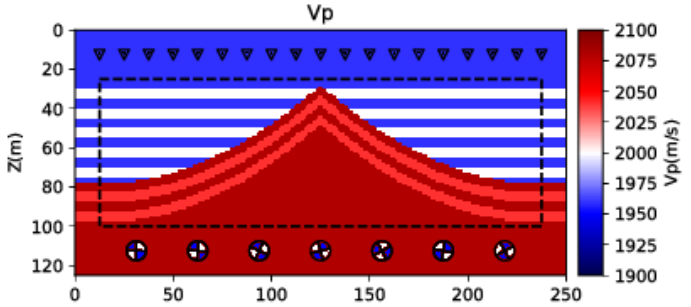
Choose $\mathbf{m}$ to minimize: $\|F(\mathbf{m}) - \mathbf{d}\|$

Synthetic examples

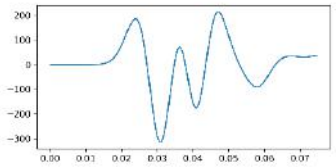
- Fully **elastic** model
- Two component data (x,z)
  7 earthquake sources, 19 receivers
- 180*60*3 free parameters (Vp, Vs, density)
- Uniform priors
  - Vp: 2000 $\pm 100\ m/s$
  - Vs: 800 $\pm 50\ m/s$
  - Density: 1500 $\pm 100\ kg/m^3$

Test case chosen for comparison with Hamiltonian-MC study (*Gebraad et al., 2019*)

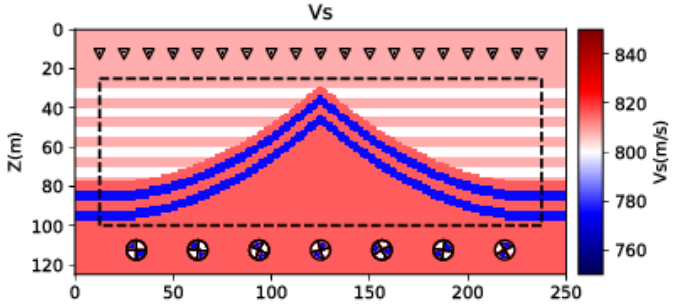
Vp



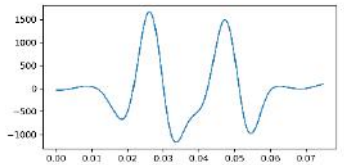
50 Hz



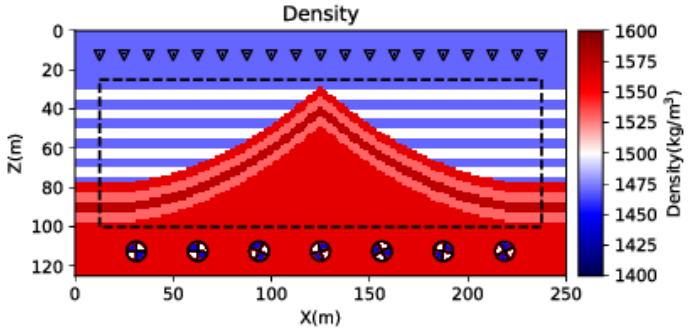
Vs



50 Hz



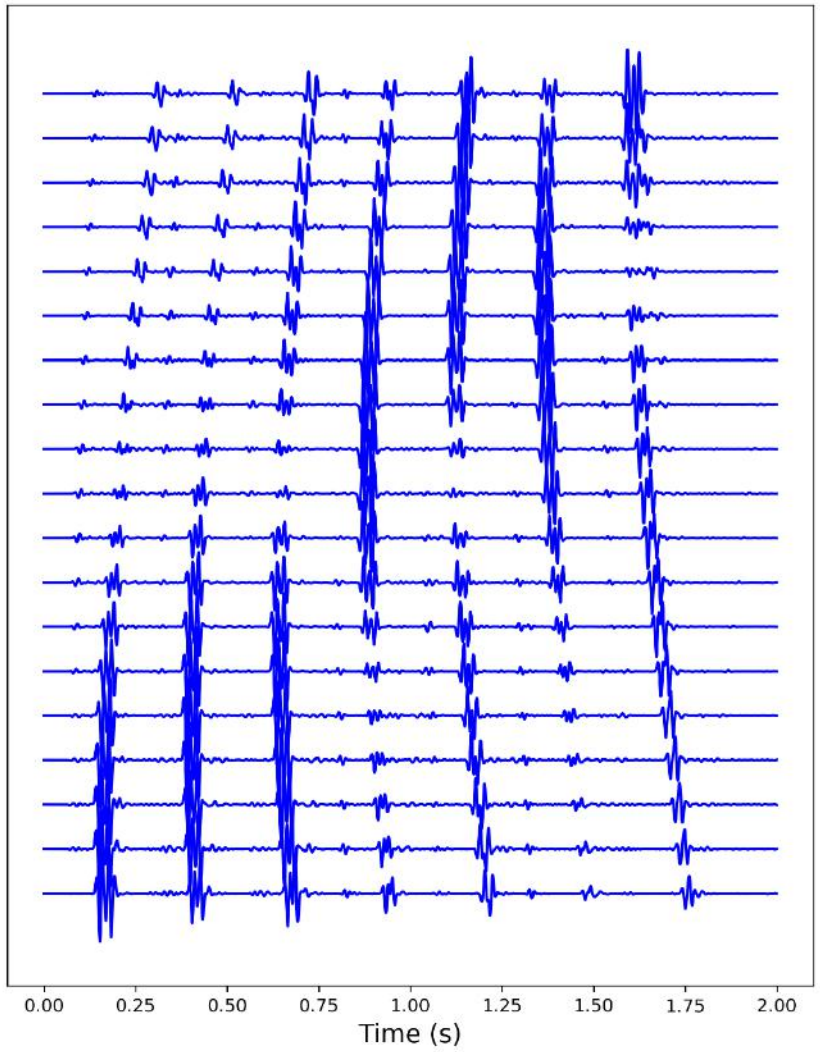
Density



Gebraad et al., 2019 EarthArXiv

Synthetic examples

X-component Data



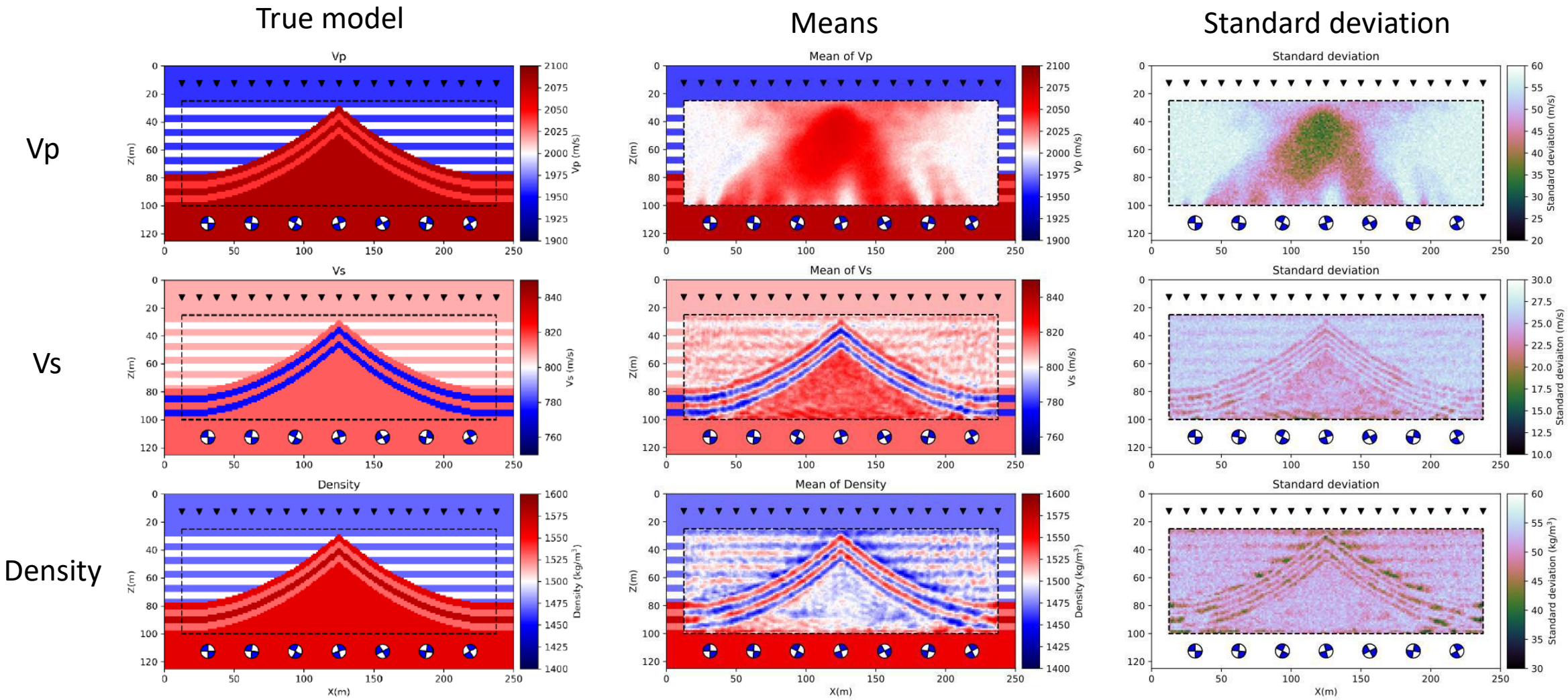
Likelihood: (L2 norm in time domain)

$$p(\mathbf{d}_{obs}|\mathbf{m}) \propto \boldsymbol{e}^{-\boldsymbol{\varphi}(\mathbf{m})}$$

$$\varphi(\mathbf{m}) = \frac{1}{2}\sum_{i,j}\left(\frac{d_{ij,obs} - d_{ij}(\mathbf{m})}{\sigma_{ij}}\right)^2$$

where *i* is receiver number and *j* denotes time samples and $\sigma_{ij} = 1$.

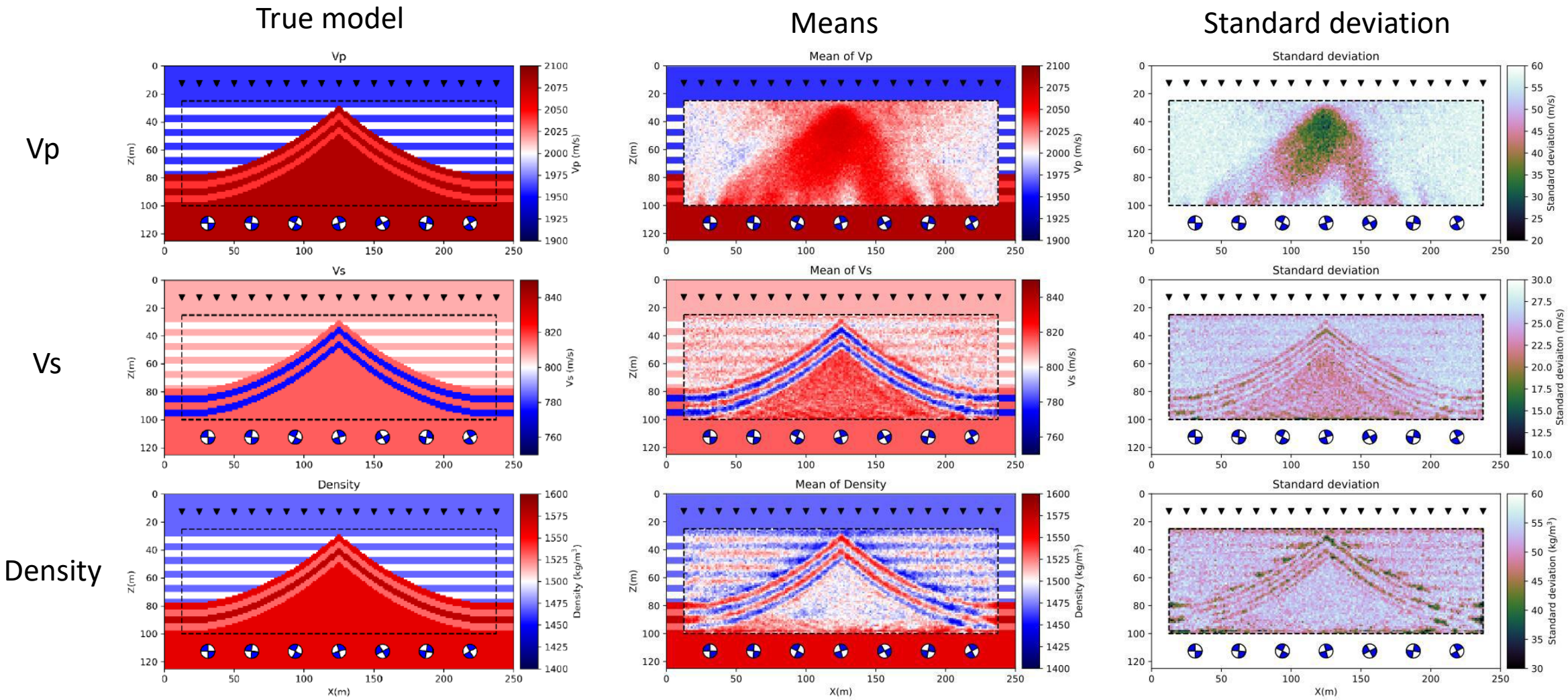
Results of SVGD



600 particles, 600 iterations, parallelized using 16 cores

Zhang & Curtis, 2020b

Results of Hamiltonian Monte Carlo

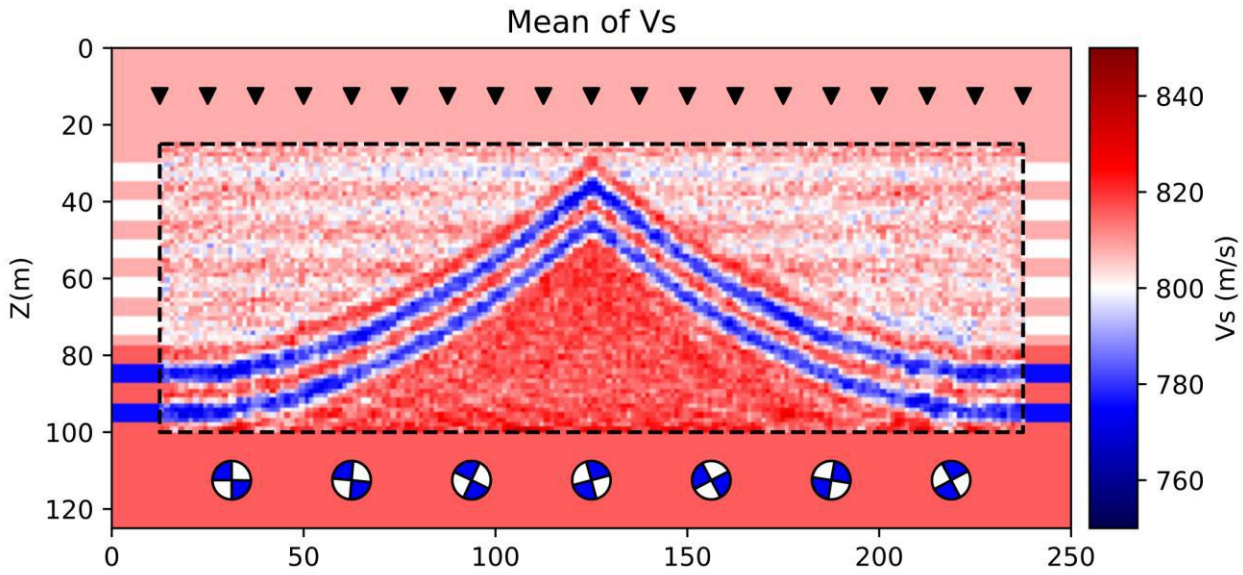


1 chain, about 100,000 simulations, not parallelized

Gebraad et al., 2019 EarthArXiv

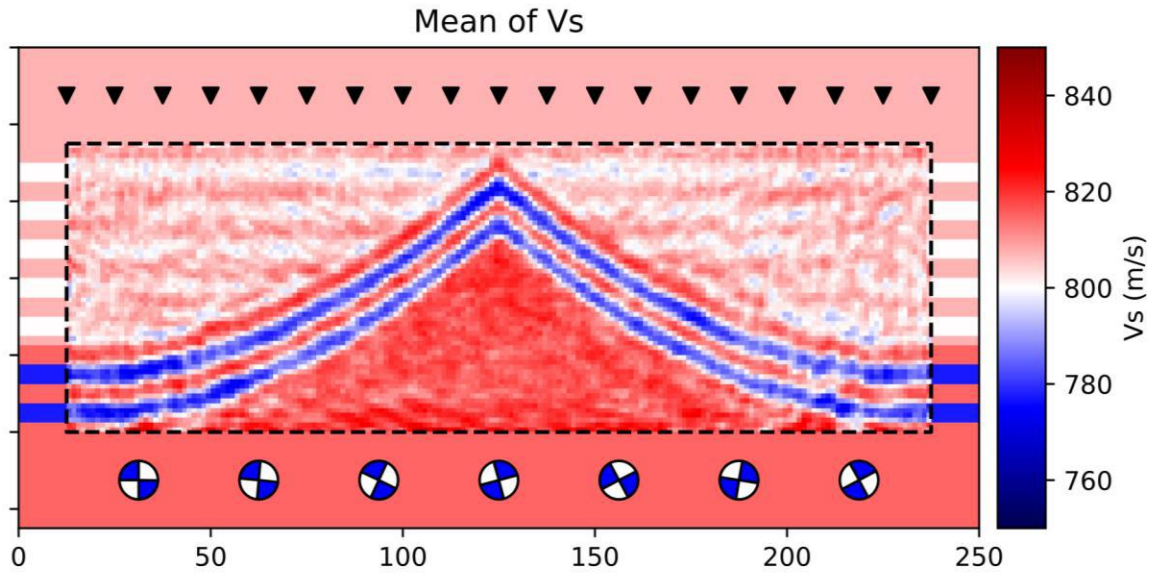
Comparison

Mean Vs from **HMC**



10,000 samples

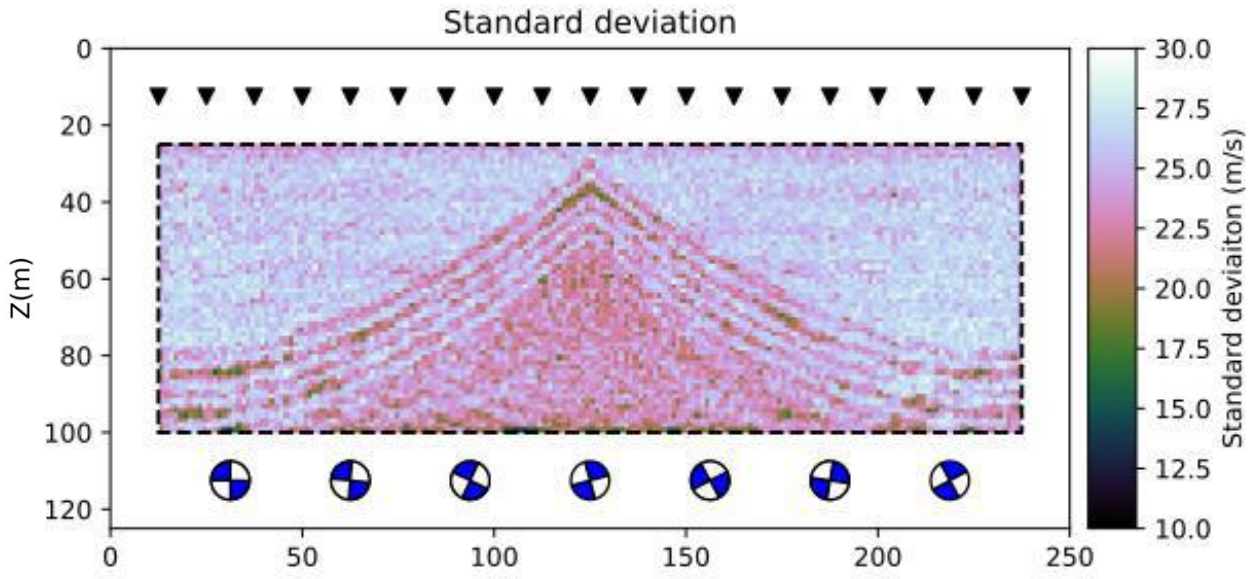
Mean Vs from **SVGD**



600 samples

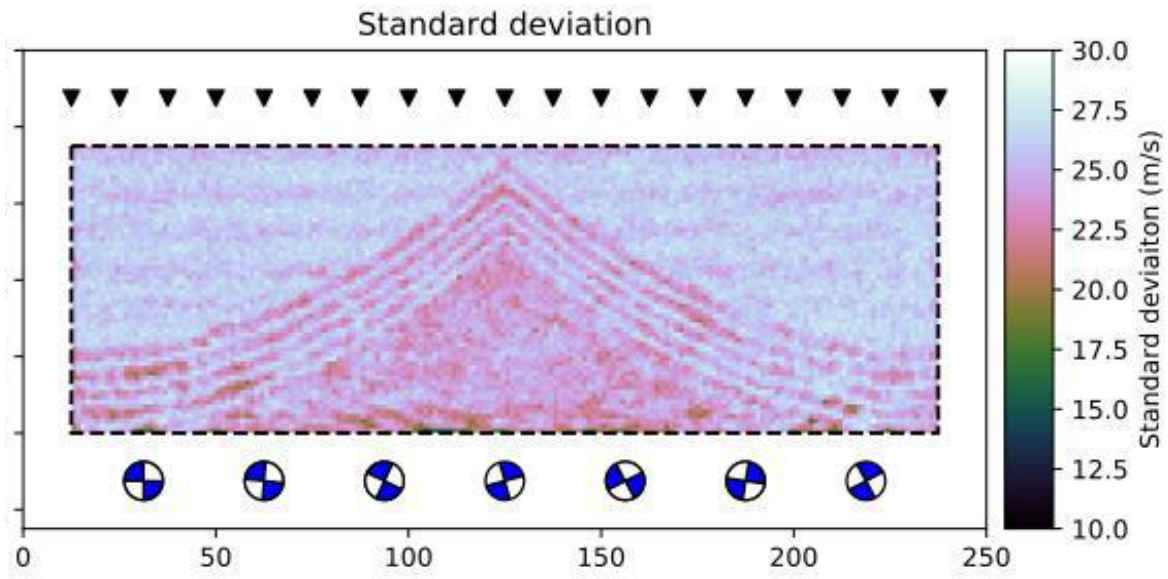
Comparison

Stdev Vs from **HMC**



10,000 samples

Stdev Vs from **SVGD**



600 samples

Optimally Selected to Represent Posterior pdf

Summary

- Introduced two variational inference methods, but there are more!
  - Stochastic Stein Variational Gradient Descent
  - Mixture density neural networks
  - Normalizing Flows
- Compared with Metropolis-Hastings and trans-Dimensional Monte Carlo
  - Variational methods provide efficient alternatives to McMC

- Variational methods under-explored in Seismology – **Try them!**

➔**FOR PAPERS:** *https://blogs.ed.ac.uk/curtis/*

➔**Or email:** **Andrew.Curtis@ed.ac.uk**

Zhang & Curtis (2020a,b)
Nawaz & Curtis (2018,2019,2020)