Facing the Curse of Dimensionality

Klaus Mosegaard

Niels Bohr Institute, University of Copenhagen

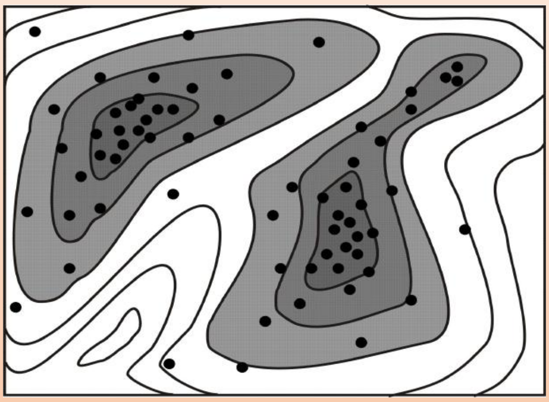
Presentation 21 March 2023 at the SPIN short course, Pitlochry, Scotland

1

Ideal Sampling Solutions to the Non-linear Probabilistic Inverse Problem

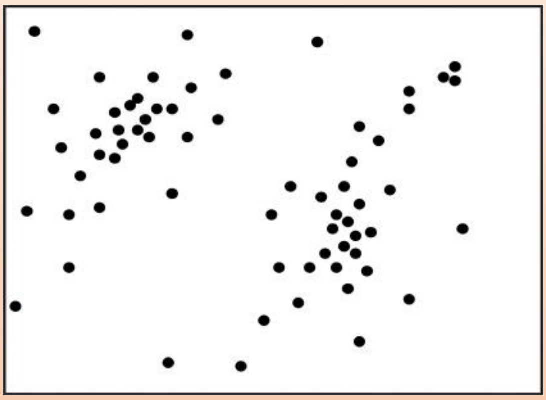
2

Solution: Sampling the Posterior PDF



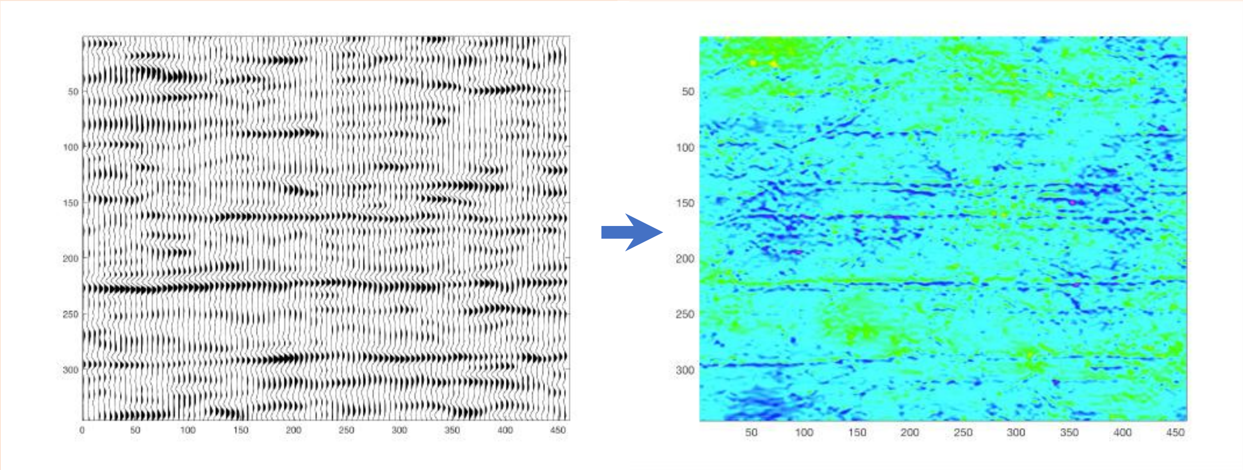
3

Solution: Sampling the Posterior PDF



4

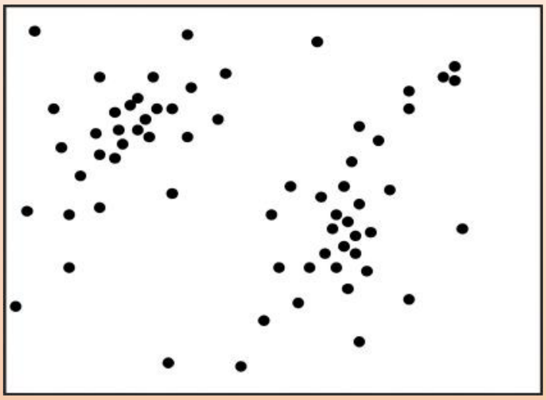
Solution: Sampling the Posterior PDF



(Fernandes and Mosegaard, Geophysical Prospecting 2022)

5

Solution: Sampling the Posterior PDF



...but how difficult is it to obtain such a sample?

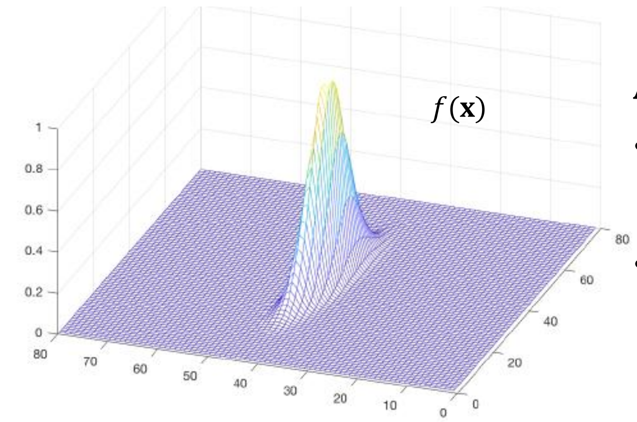
6

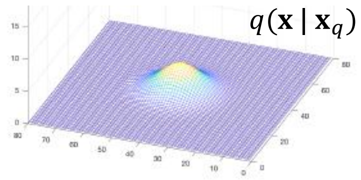
Monte Carlo Algorithms in Spaces of High Dimension

Why pre-knowledge about the distribution is decisive!

7

Blind MCMC Sampling of a Gaussian: A Hard Problem!





Assumptions:

- $\mathbf{x}$ is Gaussian:

$$f(\mathbf{x}) = \mathcal{N}_{\mathbf{x}}\,(\mathbf{x}_0, \mathbf{C}).$$

- Proposal distribution is isotropic Gaussian:

$$q(\mathbf{x} \mid \mathbf{x}_q) \;= \mathcal{N}_{\mathbf{x}}\,(\mathbf{x}_q, \mathbf{C}_q).$$

- Start sampling at f's maximum point $\mathbf{x}_0$.

8

Sampling a Gaussian **without knowing it is a Gaussian**

Examples: Let us consider the case where $\sigma_q^2 = 1$, and $\sigma_n^2 = \frac{1}{n}$:

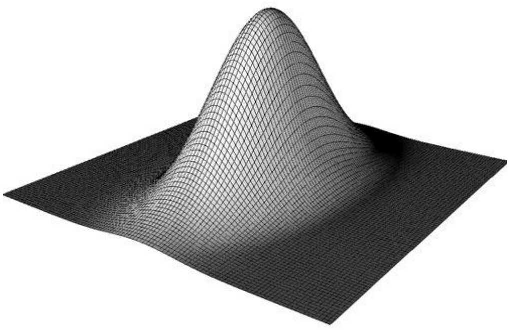
1. $N = 2$:
   Expected acceptance probability: 0.4082
   Mean waiting time between accepted moves: $0.4082^{-1} \approx$ 2.5 iterations

2. $N = 10$:
   Expected acceptance probability: $1.5828 \cdot 10^{-4}$
   Mean waiting time between accepted moves: $\approx$ 6318 iterations.

3. $N = 100$:
   Expected acceptance probability: $1.03 \cdot 10^{-80}$
   Mean waiting time between accepted moves: $\approx 10^{80}$ iterations.

9

Sampling a Gaussian, **knowing that it is Gaussian**: Easy!



Characterized by:

- N components of its mean vector
- $N\,(N+1)/2$ components of its covariance matrix.

The family of Gaussians over an N-dimensional space is a manifold of dimension

$$N + N\,(N+1)/2$$

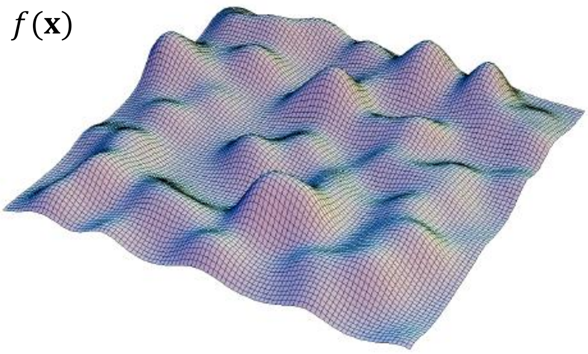
10

- At least $N + N(N+1)/2$ function evaluations are required to characterize ("reconstruct") an N-dimensional Gaussian.
- Consequently, the best conceivable algorithm needs $\sim N + N(N+1)/2$ function evaluations to produce one exact sample of an N-dimensional Gaussian!

Sampling a Gaussian is **not** a hard problem, if you know it is Gaussian

11

Blind Sampling of a Complex Distribution (Hard)



Assume:

- f can be expanded in terms of basis functions:

$$f(\mathbf{x}) = \sum_{j=1}^{J} u_j \varphi_j(\mathbf{x})$$

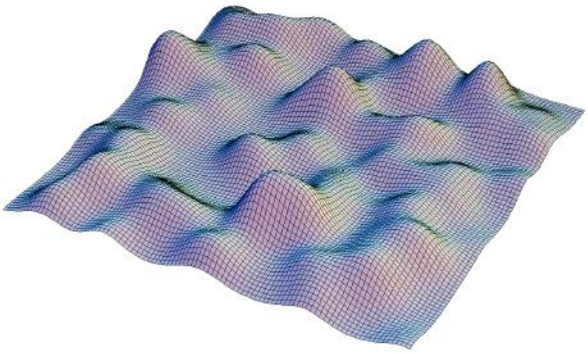
- We have K samples $\mathbf{x}_1, \cdots, \mathbf{x}_K$ and sample values:

$$s_k = f(\mathbf{x}_k) = \sum_{j=1}^{J} u_j \varphi_j(\mathbf{x}_k)$$

Hence, $\mathbf{s} = \mathbf{Fu}$ where $\mathbf{s} = (s_1, \cdots, s_K)$, $\mathbf{u} = (u_1, \cdots, u_J)$, and $F_{kj} = \varphi_j(\mathbf{x}_k)$.

12

Blind Sampling of a Complex Distribution (Hard)



- We have K samples $\mathbf{x}_1, \cdots, \mathbf{x}_K$ and sample values:

$$s_k = f(\mathbf{x}_k) = \sum_{j=1}^{J} u_j \varphi_j(\mathbf{x}_k)$$

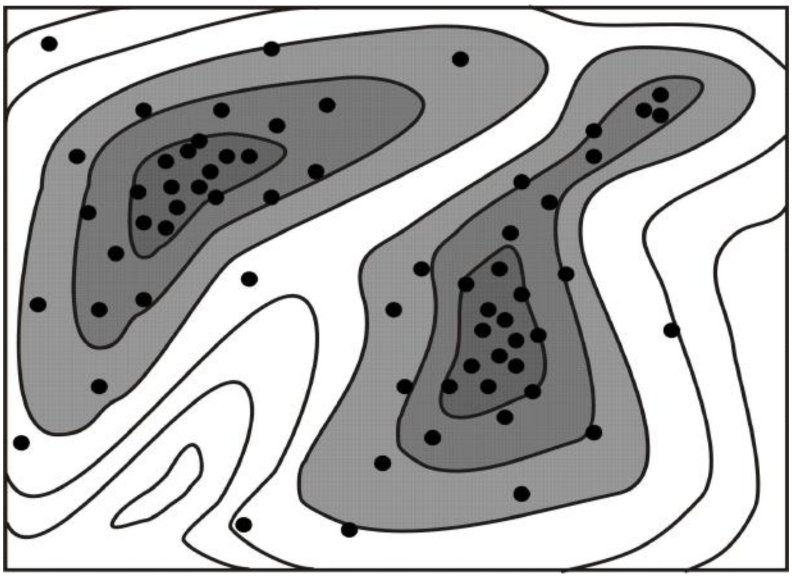
- $\mathbf{s} = \mathbf{F}\mathbf{u}$

$\mathbf{F}^T\mathbf{F}$ singular (e.g., # samples $< J$) $\Longrightarrow$ Incomplete knowledge/sampling
$\Longrightarrow$ Potentially missing "peaks"

If # required base functions grows exponentially with dimension, the problem is **Hard!**

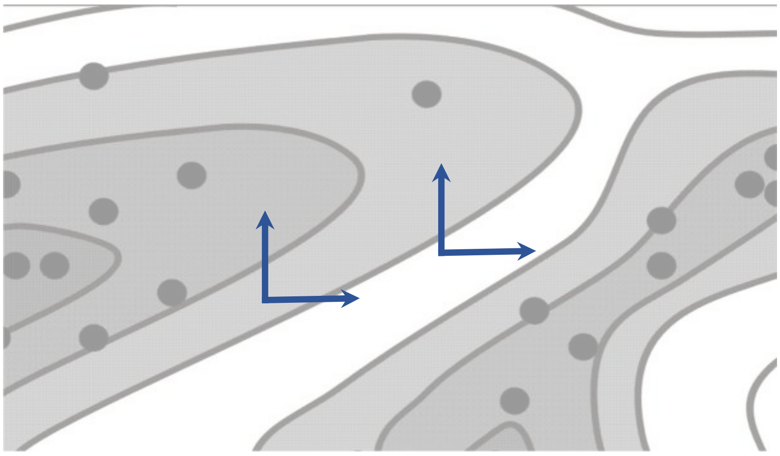
13

Blind Sampling a Complex Distribution (Hard)



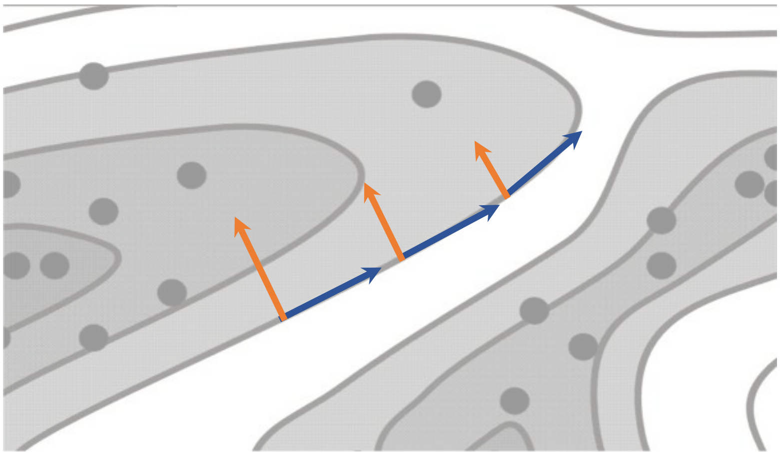
14

Blind Sampling of a Complex Distribution (Hard)



15

Sampling of a Complex Distribution, **having gradients** (Easier)

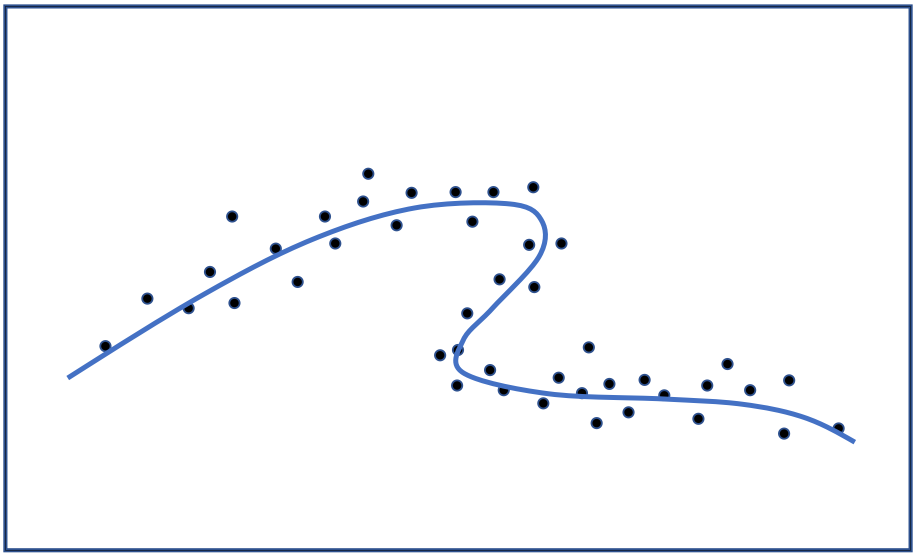


16

When Solutions are Essentially Located in a Lower-Dimensional Subspace

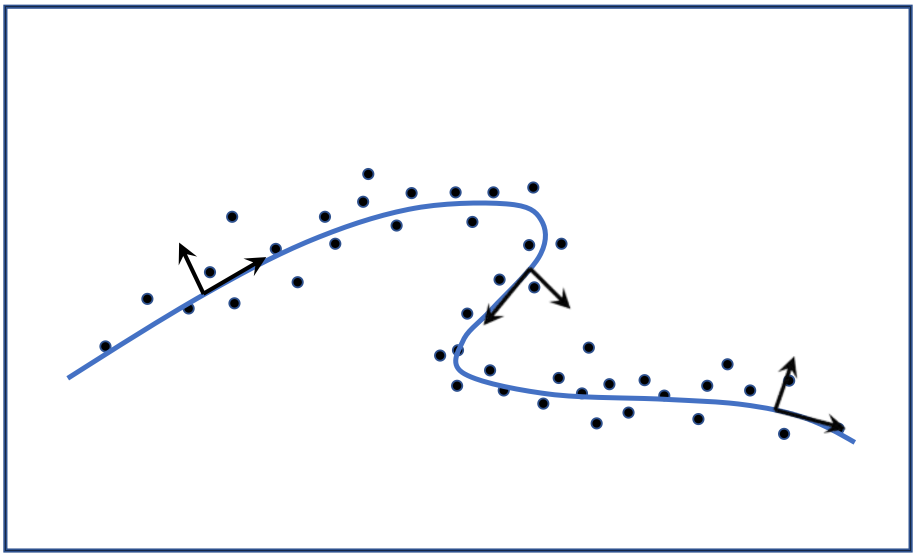
17

Sometimes solutions are essentially located in a lower-dimensional manifold..



18

Sometimes solutions are essentially located in a lower-dimensional manifold..

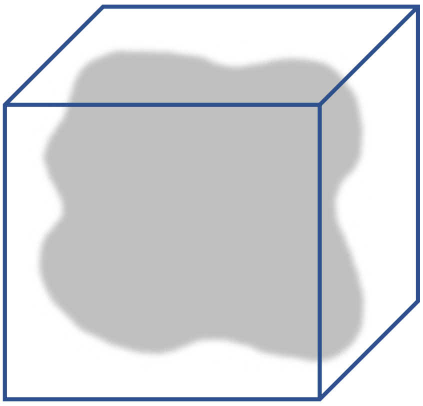


19

Highly Nonlinear Inverse Problems: Dimensionality and Degrees of Freedom

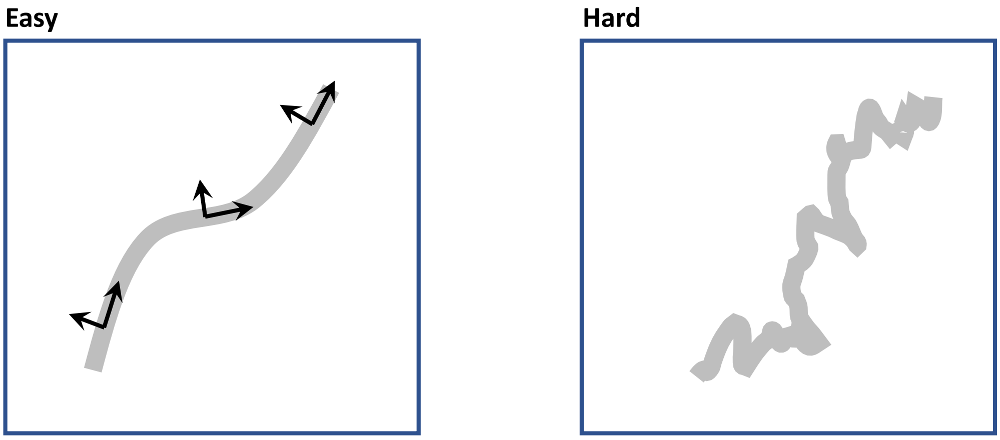
Easy to find acceptable models, but hard to sample due to the high dimension

- Space-filling Distribution
- N degrees of freedom
- Embedded in ND



20

Highly Nonlinear Inverse Problems: Dimensionality and Degrees of Freedom

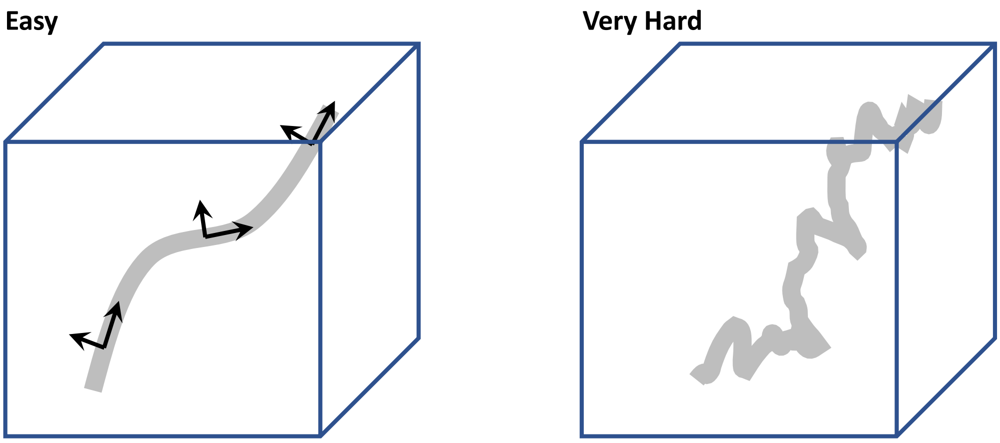


- Parametric distribution
- 1 degree of freedom
- Embedded in 2D

- Non-parametric distribution
- 1 degree of freedom
- Embedded in 2D

21

Highly Nonlinear Inverse Problems: Dimensionality and Degrees of Freedom



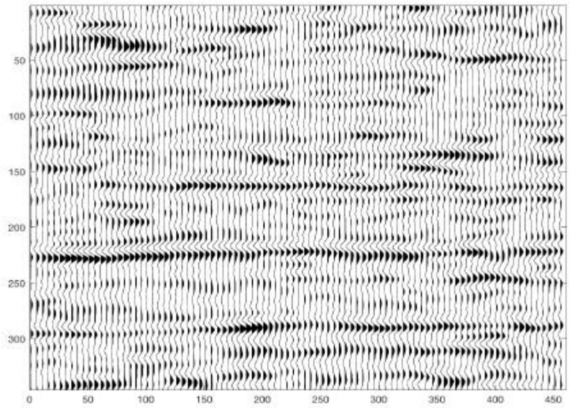
- Parametric distribution
- 1 degree of freedom
- Embedded in ND

- Non-parametric distribution
- 1 degree of freedom
- Embedded in ND

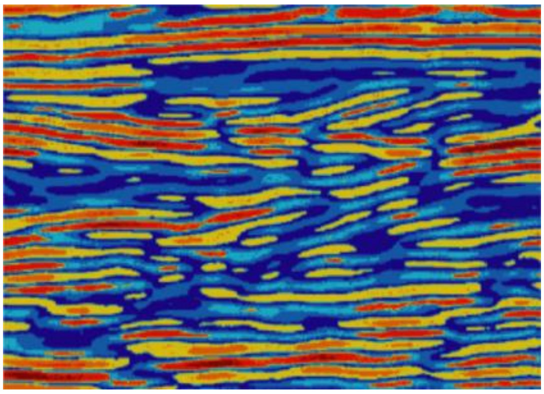
22

A Non-Parametric Posterior from Inversion of Seismic Data with a Multiple-Point Geostatistical Prior

Seismic Data



Model realization from a Multiple-Point Geostatistical Prior



GAIA LAB: https://wp.unil.ch/gaia/mps/ds/

23

Preliminary Conclusion

- A Posterior that is only nonzero close to a **subspace described by (few) local coordinates** is **easy** to sample.
- A Posterior that is only nonzero close to a **subspace without local coordinates** is **difficult** to sample.
- The latter case gets worse when the dimension of the embedding space grows!

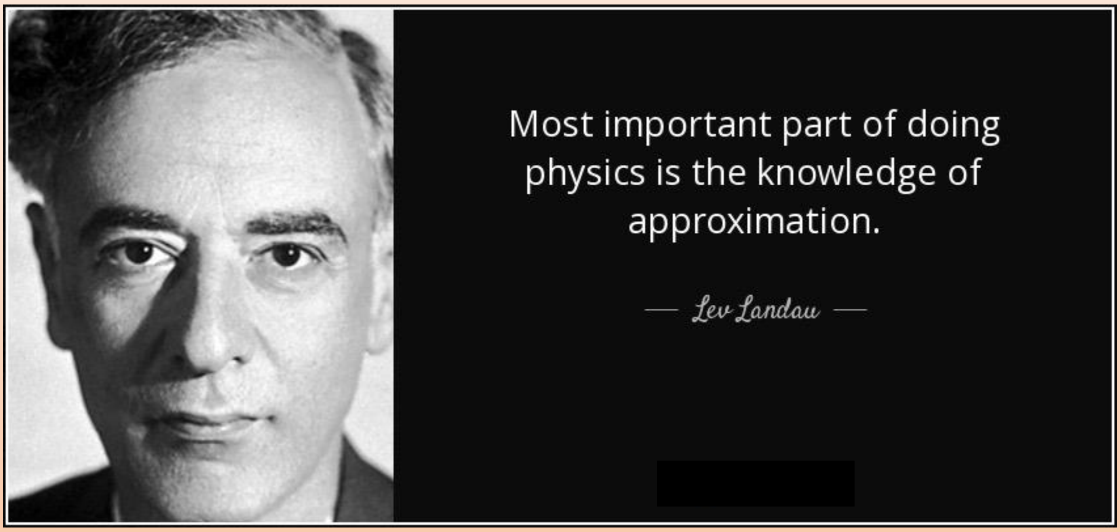
24

MCMC Algorithms with Informed Proposals

Strategies guided by

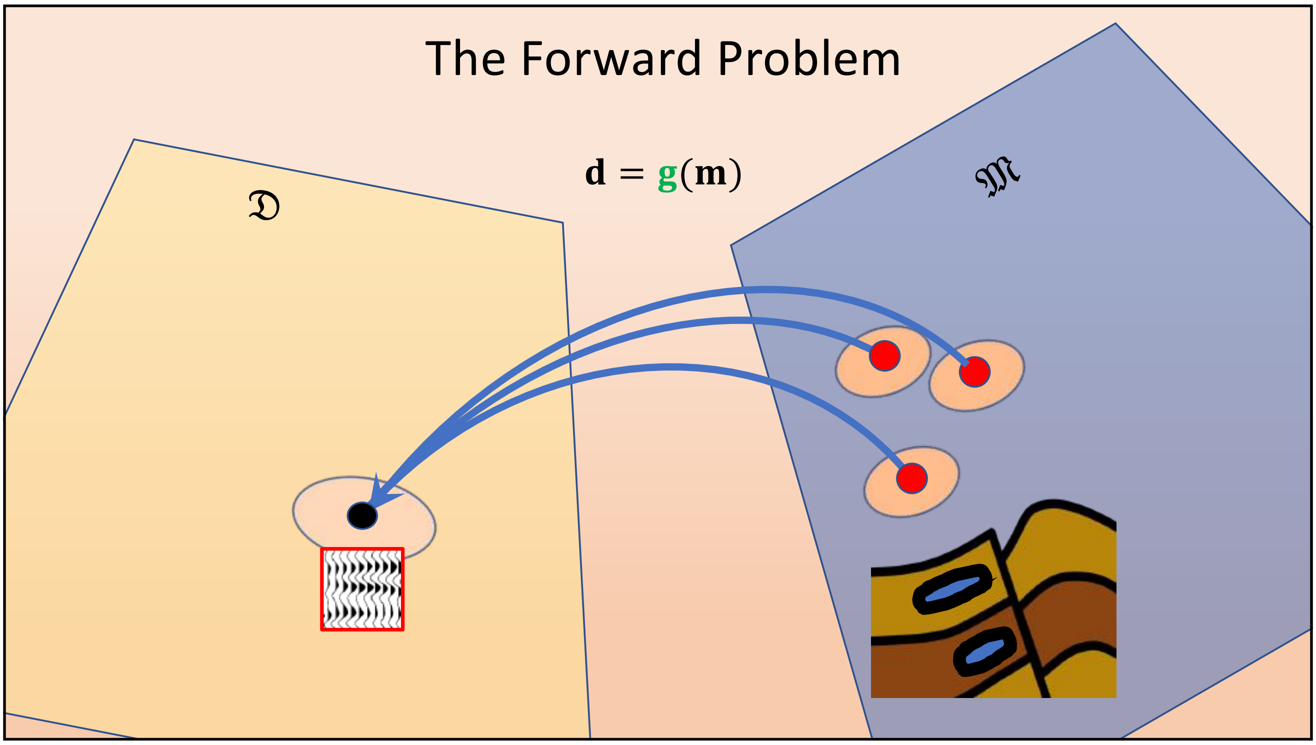
the physics of the problem

25



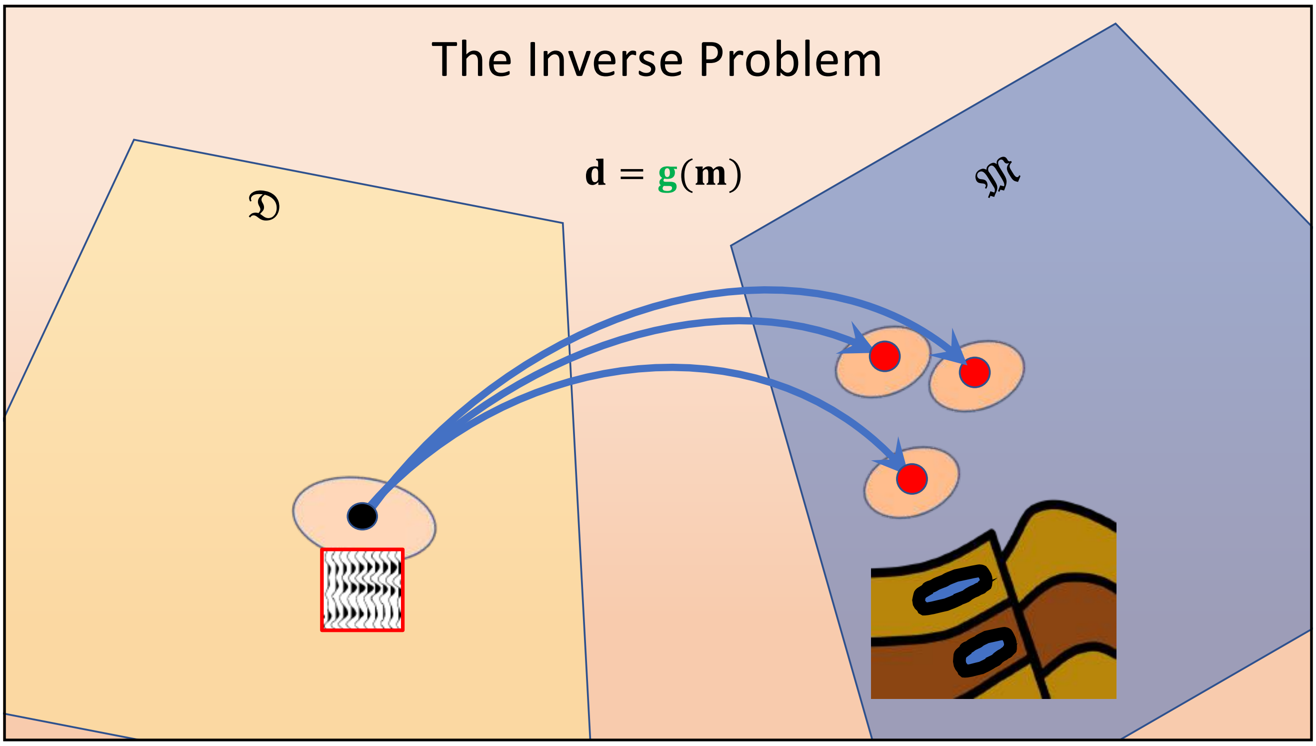
Lev Landau (1908-1968)

26



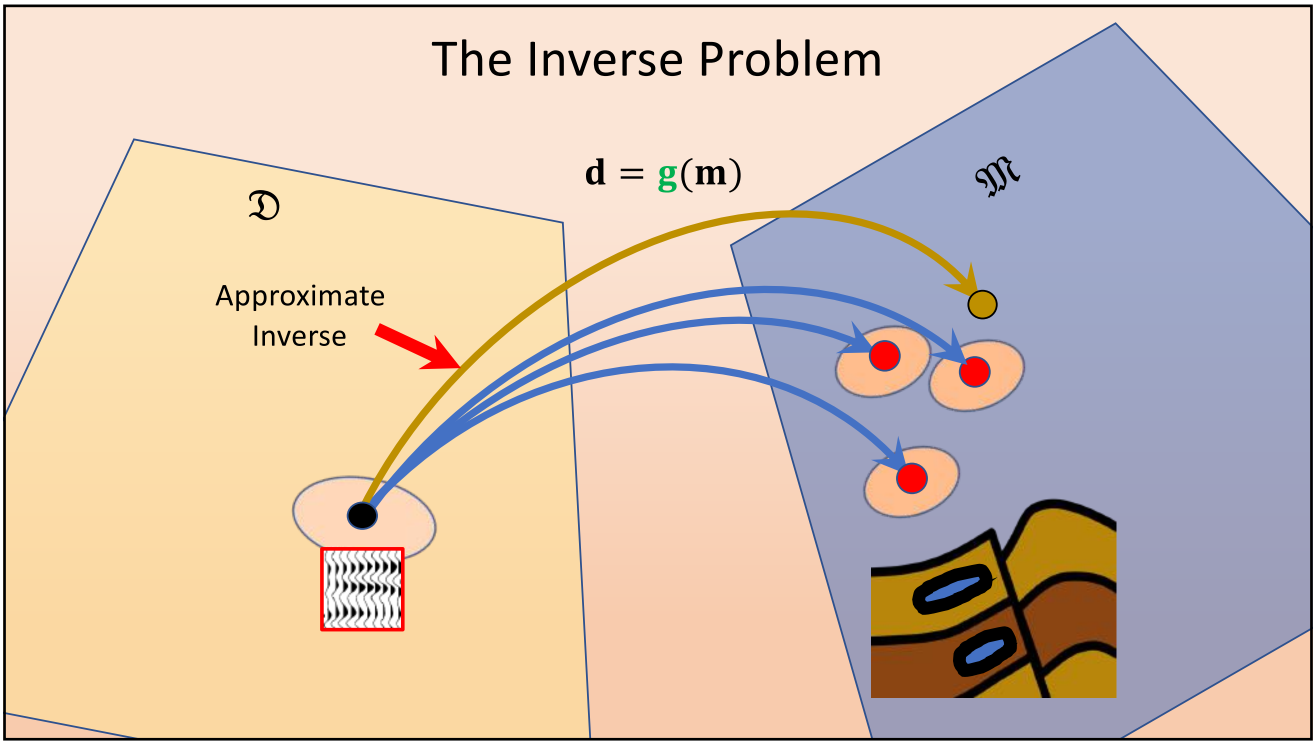
The Forward Problem
$\mathbf{d} = \mathbf{g}(\mathbf{m})$
𝔇
𝔐

27

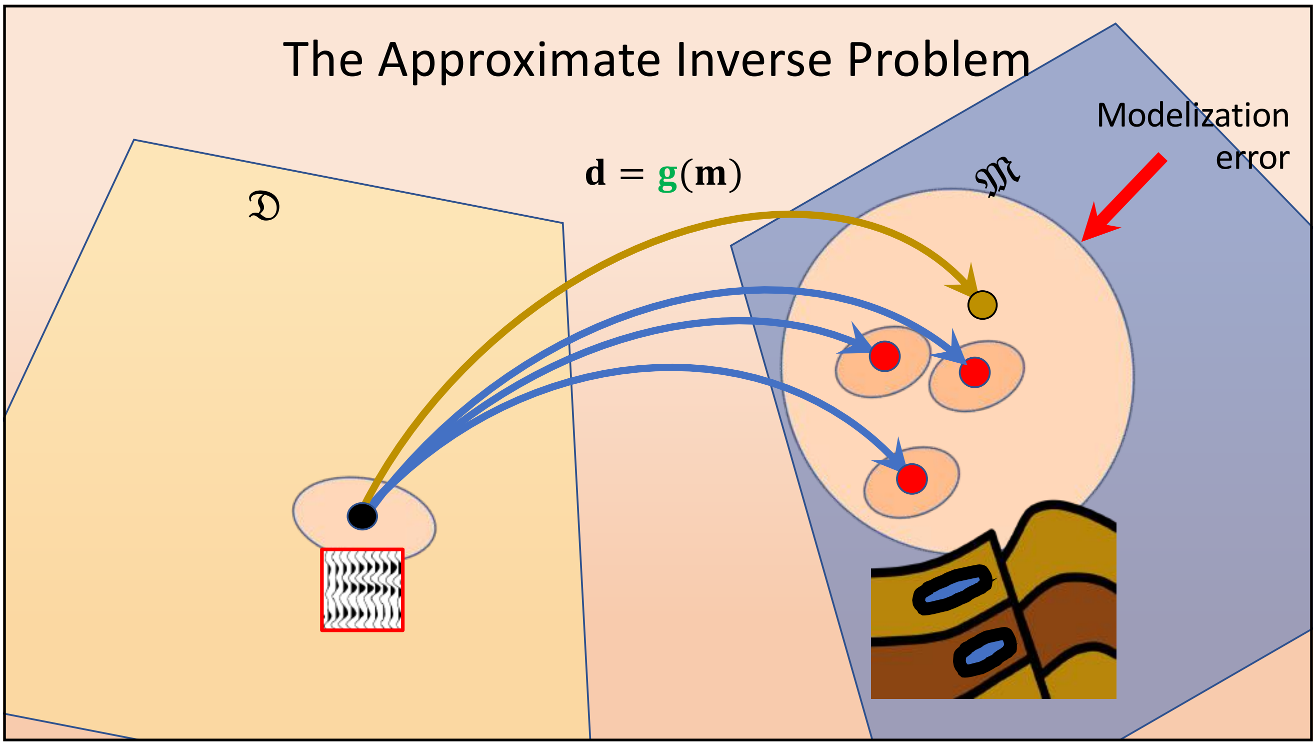


The Inverse Problem
$\mathbf{d} = \mathbf{g}(\mathbf{m})$
𝔇
𝔐

28


The Inverse Problem
$\mathbf{d} = \mathbf{g}(\mathbf{m})$
𝔇
𝔐
Approximate
Inverse

29

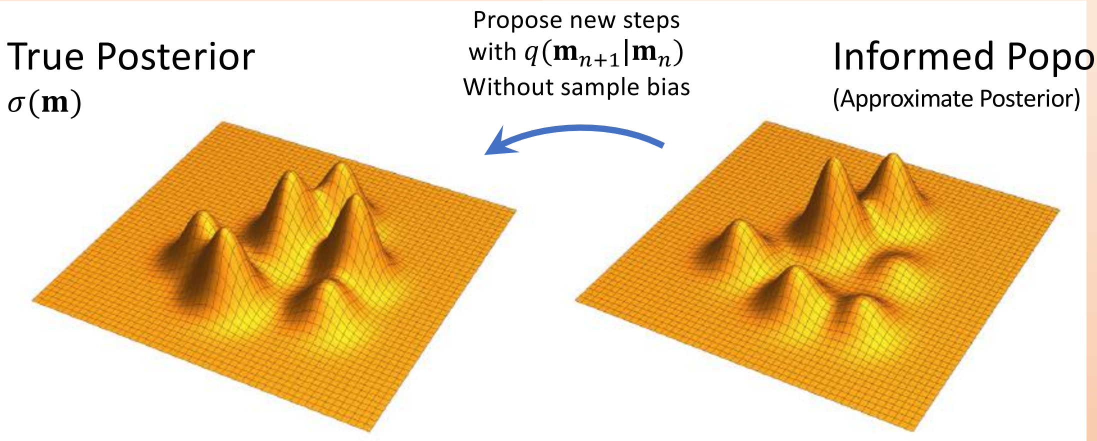

The Approximate Inverse Problem
$\mathbf{d} = \mathbf{g}(\mathbf{m})$
𝔇
𝔐
Modelization
error

30

Building Approximate Physics into MCMC Without an (Asympthotic) Bias

31

MCMC with Informed Proposals: The Idea

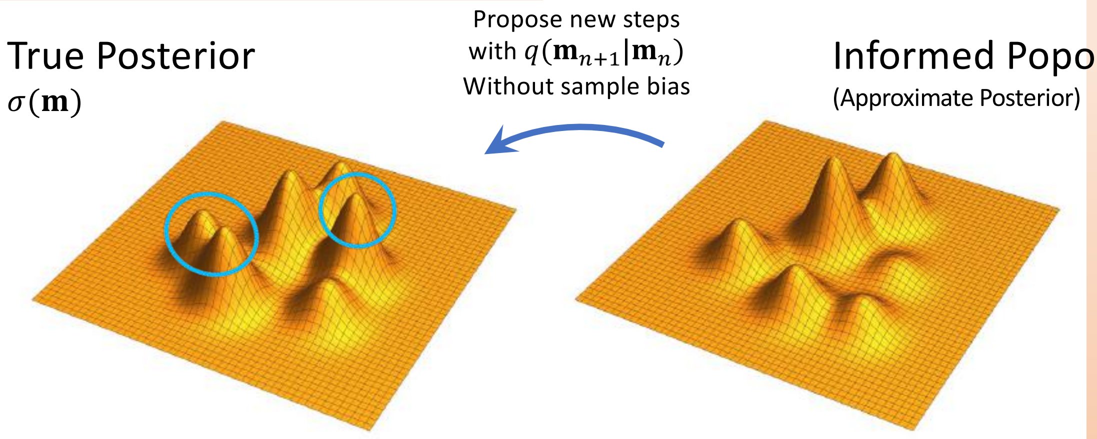


Propose with $q(\mathbf{m}_{n+1}|\mathbf{m}_n)$ Accept with probability:

$$P_{acc} = \frac{\sigma(\mathbf{m}_{n+1})}{\sigma(\mathbf{m}_n)} \frac{q(\mathbf{m}_n|\mathbf{m}_{n+1})}{q(\mathbf{m}_{n+1}|\mathbf{m}_n)}$$

32

MCMC with Informed Proposals: The Idea

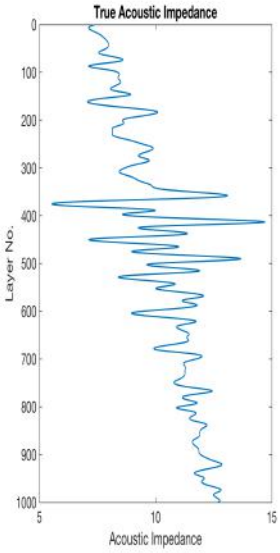


Propose with $q(\mathbf{m}_{n+1}|\mathbf{m}_n)$ Accept with probability:

$$P_{acc} = \frac{\sigma(\mathbf{m}_{n+1})}{\sigma(\mathbf{m}_n)} \frac{q(\mathbf{m}_n|\mathbf{m}_{n+1})}{q(\mathbf{m}_{n+1}|\mathbf{m}_n)}$$

33

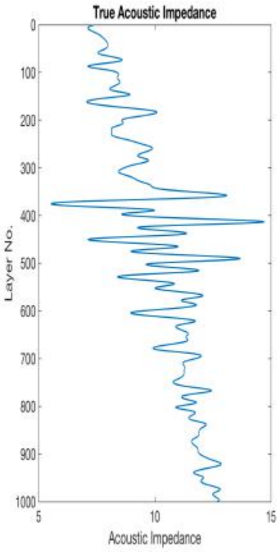
A 1-D Inverse Scattering Problem with 1000-parameters

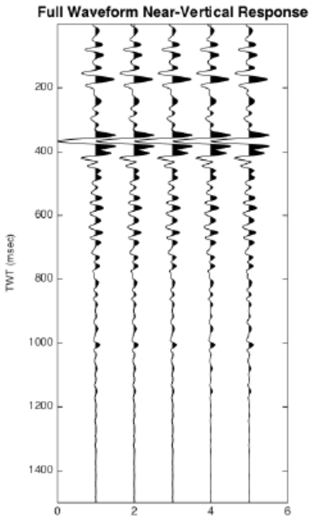


Khoshkholgh, Zunino and Mosegaard, 2021: Informed Proposal Monte Carlo. Geophys. Journ. Int.

34

A 1-D Inverse Scattering Problem with 1000-parameters

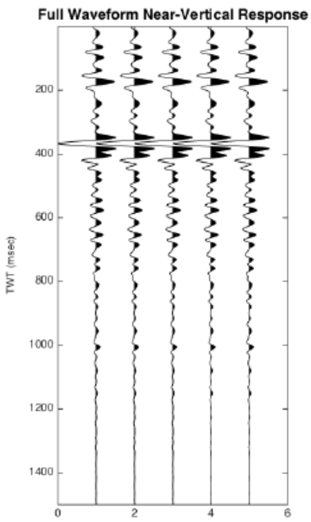


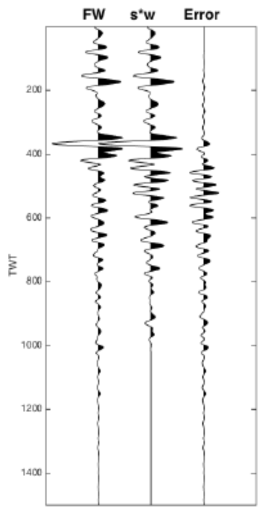


Khoshkholgh, Zunino and Mosegaard, 2021: Informed Proposal Monte Carlo. Geophys. Journ. Int.

35

A 1-D Inverse Scattering Problem with 1000-parameters

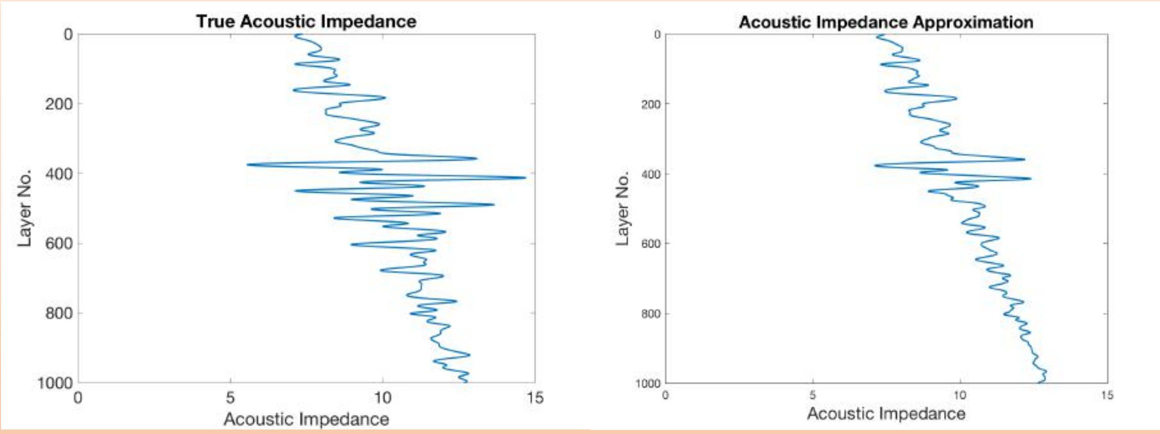




Khoshkholgh, Zunino and Mosegaard, 2021: Informed Proposal Monte Carlo. Geophys. Journ. Int.

36

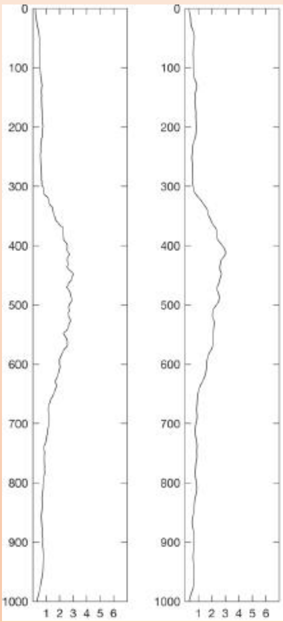
Linear Inversion: An Approximate Solution



37

Approximate (Linear) Inversion: Modelization Error

Envelope of true error

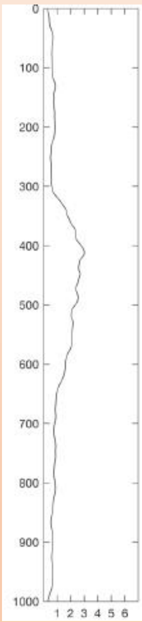


Envelope of 2. order error

1. Assume the approximate model is the true model
2. Simulate fully nonlinear data from this model
3. Find (2. order) approximate solution
4. Compute modelization error

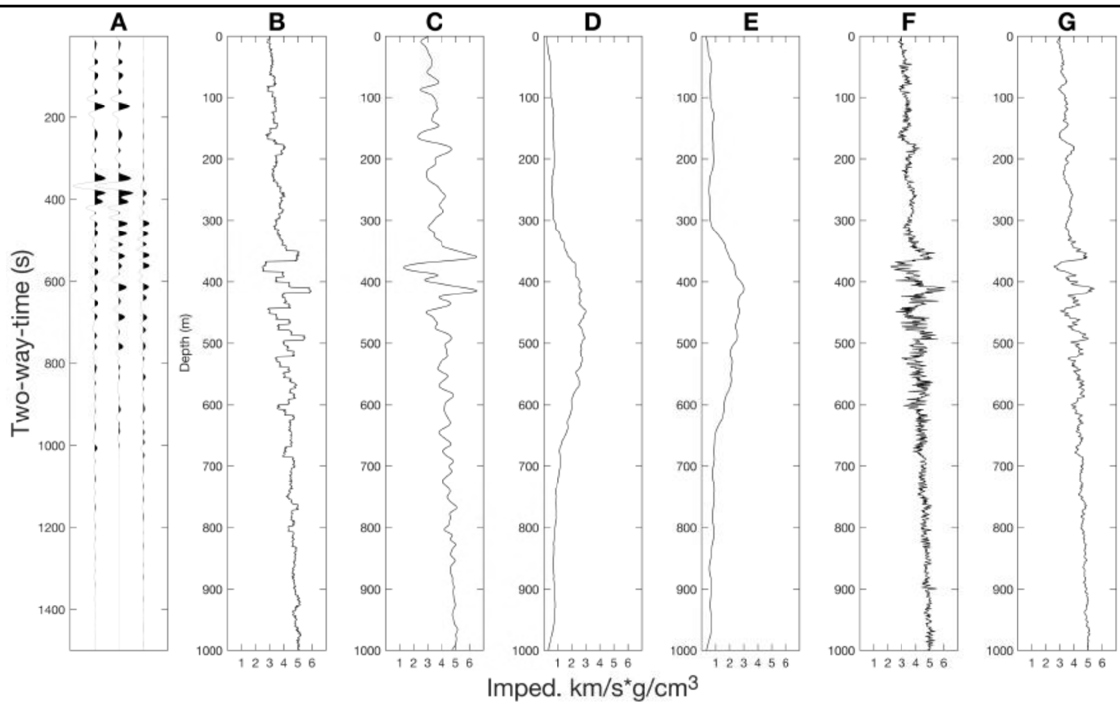
38

Defining the Informed Proposal Distribution



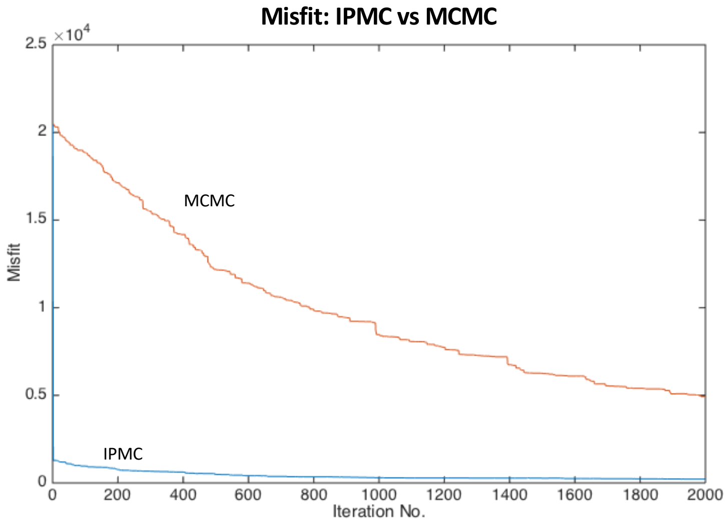
1. Define proposal distribution as a Gaussian centered at the approx. model
2. Use modelization errors at each depth/TWT as standard dev. in the proposal

39



40

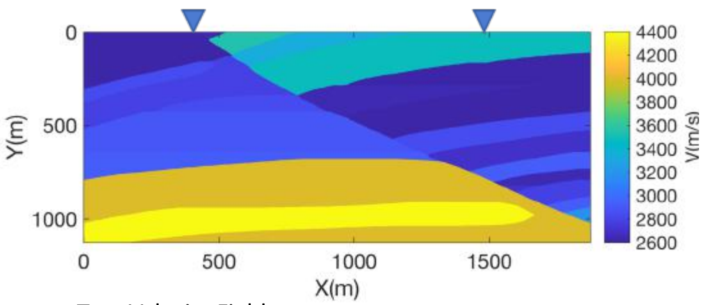
Convergence: Informed-Proposal Monte Carlo



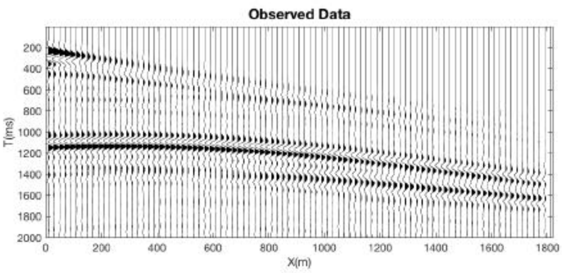
In this example: IPMC equilibrates 10^3 - 10^4 times faster

41

A ~940000-parameter Full-Waveform Acoustic Problem



True Velocity Field

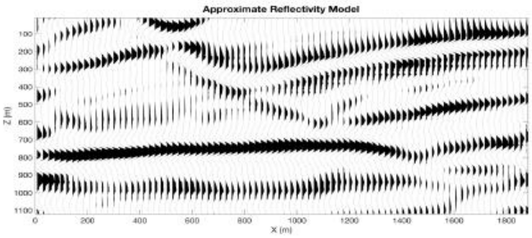


Acoustic wave field recorded at the surface

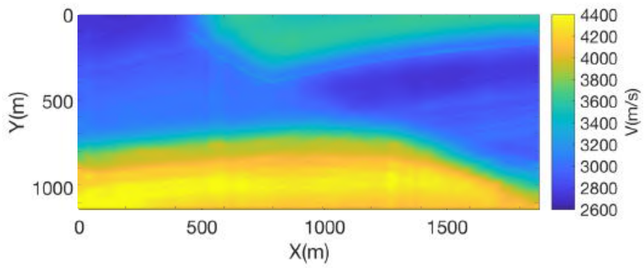
Khoshkholgh, Mosegaard and Zunino (2022)

42

An Approximate Solution



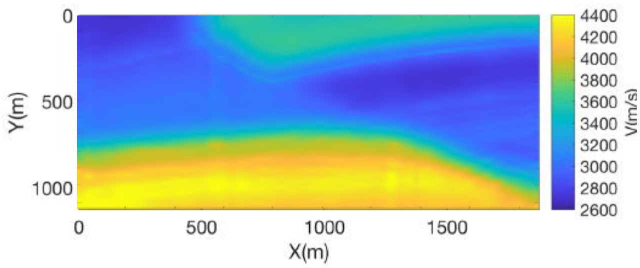
Approximate Reflectivity Model



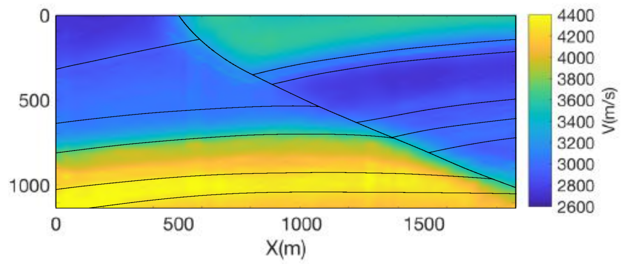
Estimated Acoustic Velocity Field

43

An Approximate Solution from Classical Processing



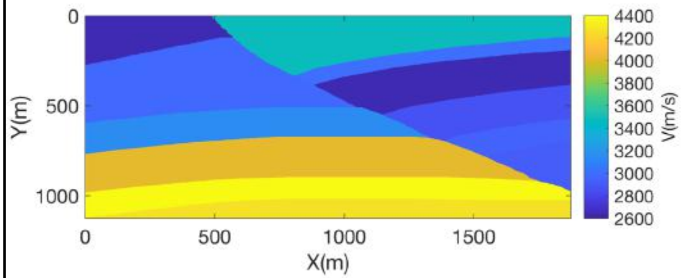
Estimated Acoustic Velocity Field



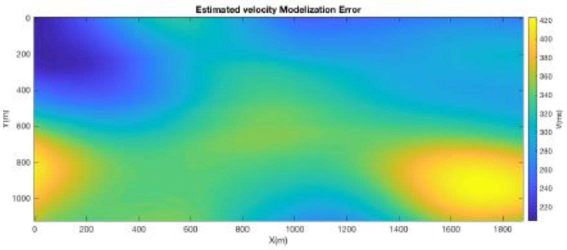
Approximate Velocity Model

44

Creating the Modelization Error Distribution



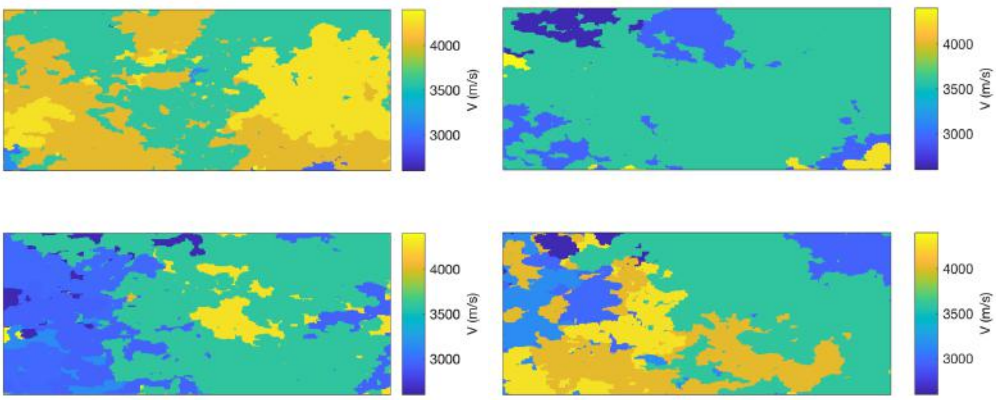
Final Approximate Velocity Model



Envelope $\left(\mathbf{m}_{approx} - \boldsymbol{H}_{approx}\left(\boldsymbol{g}_{precise}(\mathbf{m}_{approx})\right)\right)$

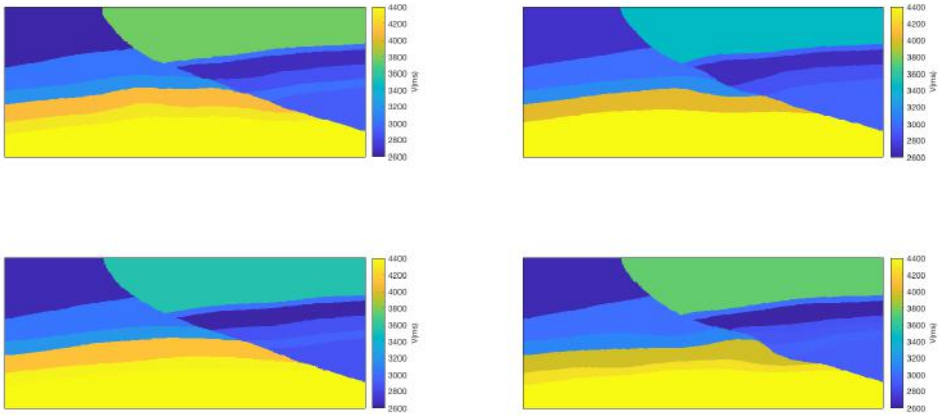
45

Samples from the combined Prior and Modelization Error Distributions



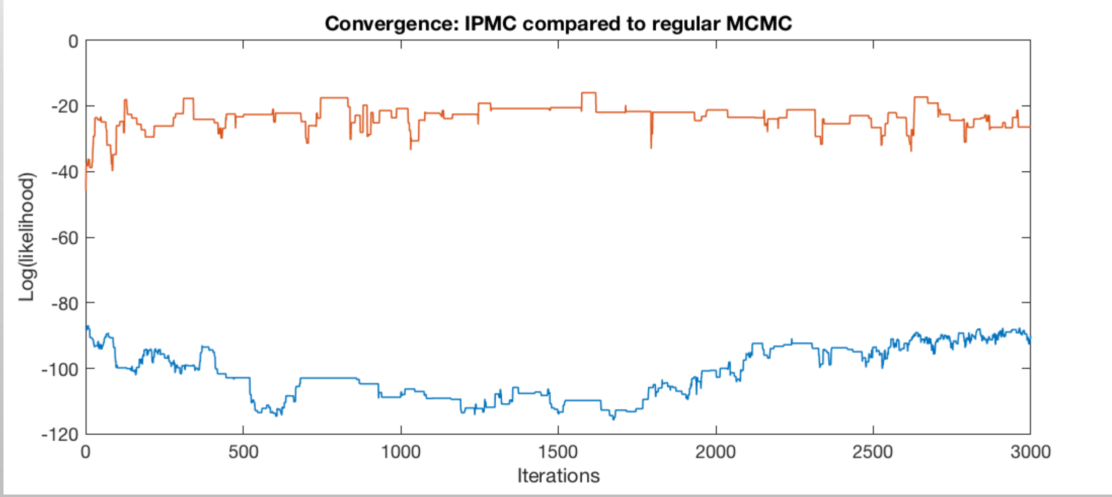
46

Samples from the Posterior Distribution



47

Convergence: Informed Proposal Monte Carlo



In this example: IPMC equilibrates in ~300 iterations

48