Computational Bayesian Inversion

Background, methods and problems

Klaus Mosegaard

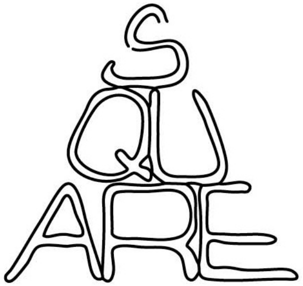
University of Copenhagen

What is probabilistic/Bayesian data analysis?

Probabilistic/Bayesian data analysis (inversion) attempts to weigh each piece of processed information objectively.

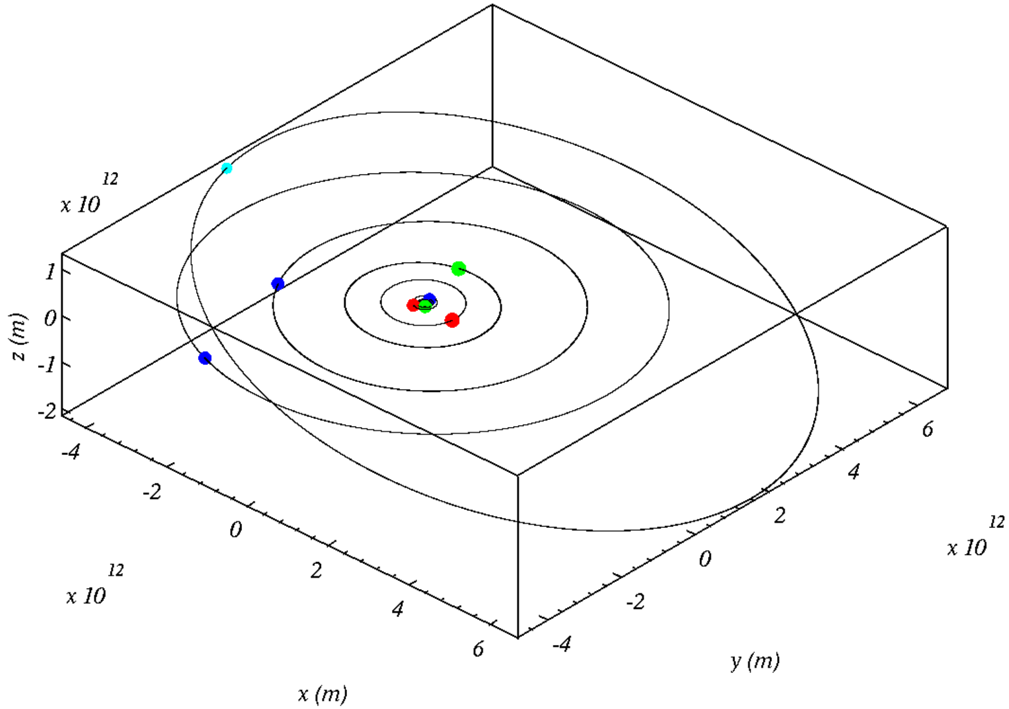
Probabilities are used as weights.

Consistency - a principal theme



- Inconsistency means contradictory results. If results are contradictory, at least one of them is wrong!
- Every step in the computational procedure must, in principle, be documented and agreed upon between analysts, thereby securing a high degree of objectivity.
- Quantification alone is not sufficient to avoid contradictions and therefore potentially meaningless results.
- In the following, we shall focus on *consistency* as a principal theme.

Conditions satisfied by physical laws



- Unique solution (for given initial/boundary conditions)
- Predictions must be independent of the reference frame

PARAMETERIZATION

Model Parameters, observable parameters and their relation

- Model parameters:

$$\boldsymbol{m} = (m_1, m_2, \dots, m_M)$$

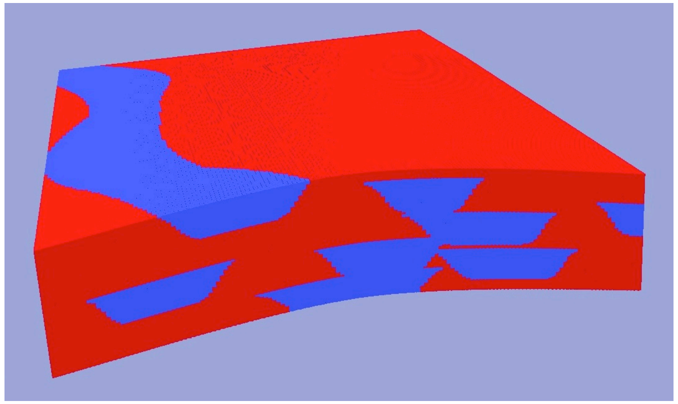
- Data:

$$\boldsymbol{d} = (d_1, d_2, \dots, d_N)$$

- Physical relation (law):

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m})$$

A plan for data analysis



- Parameterize the unknown structure m: $\boldsymbol{m} = \boldsymbol{f}(m)$ to obtain *model parameters* $\boldsymbol{m}$.
- Solve an inverse problem $\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m})$ to infer information about $\boldsymbol{m}$ from data $\boldsymbol{d}$.
- Draw conclusions about the structure: $\boldsymbol{m} \rightarrow m$.

The parameterization process

- An infinite set of orthonormal basis functions $\varphi_1(x,y,z), \varphi_2(x,y,z), \varphi_3(x,y,z), \ldots$
- Parameters $m_1, m_2, \ldots$

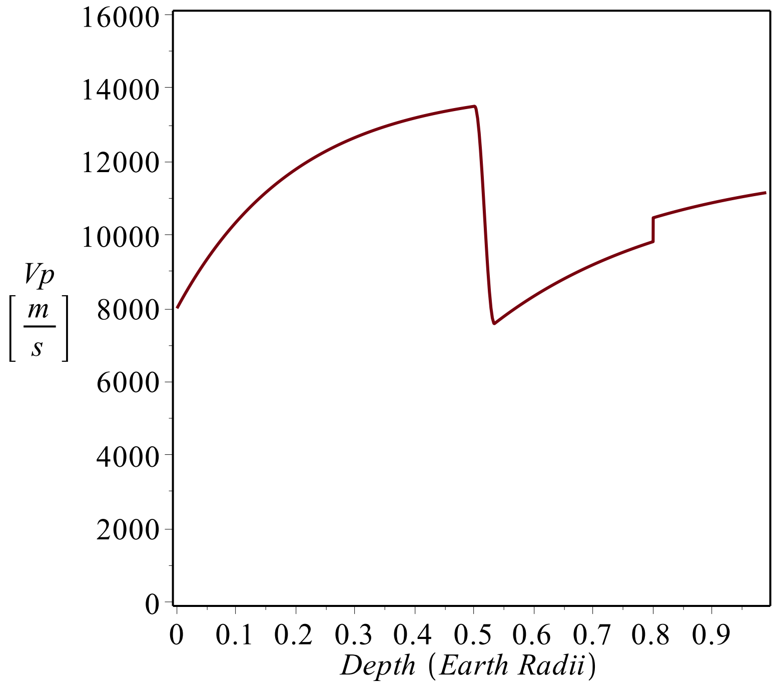
$$m(x,y,z) = \sum_{n=1}^{\infty} m_n \varphi_n(x,y,z)$$

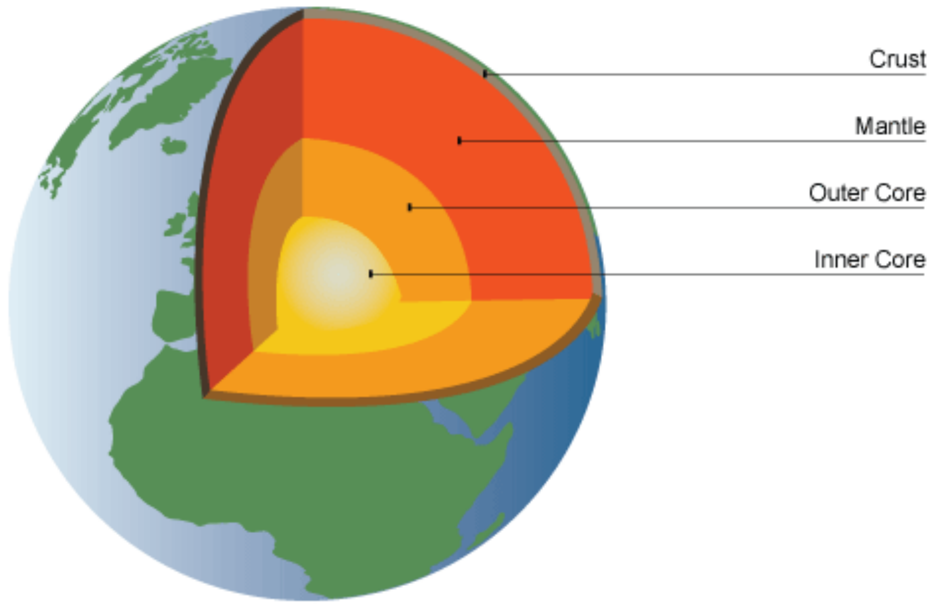
The parameterization process

- Truncate the expansion if necessary
- Keep many parameters to ensure an accurate representation

$$m(x, y, z) = \sum_{n=1}^{N} m_n \varphi_n(x, y, z)$$

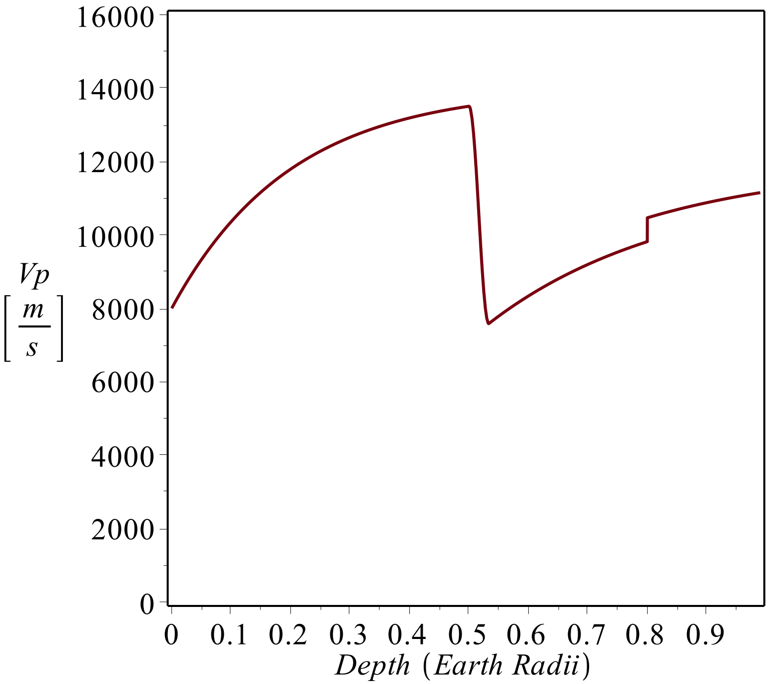
Example: A seismic model of the Earth



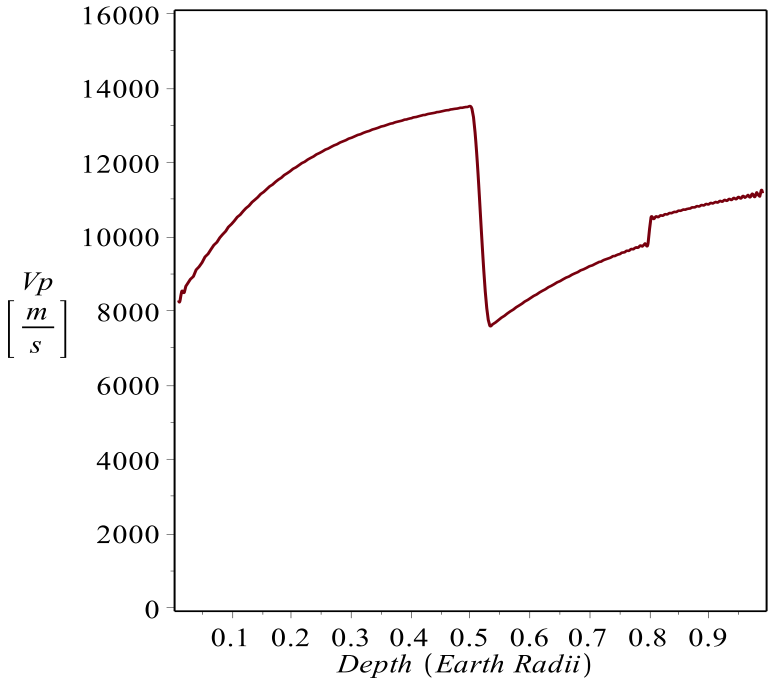


A model not unlike the P-wave velocity in the Earth's interior

Example: A seismic model of the Earth



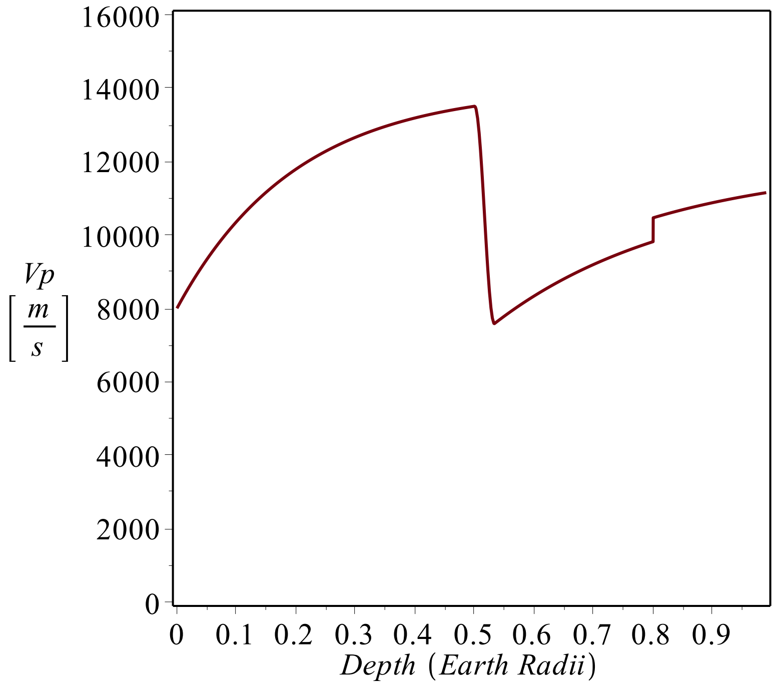
True model



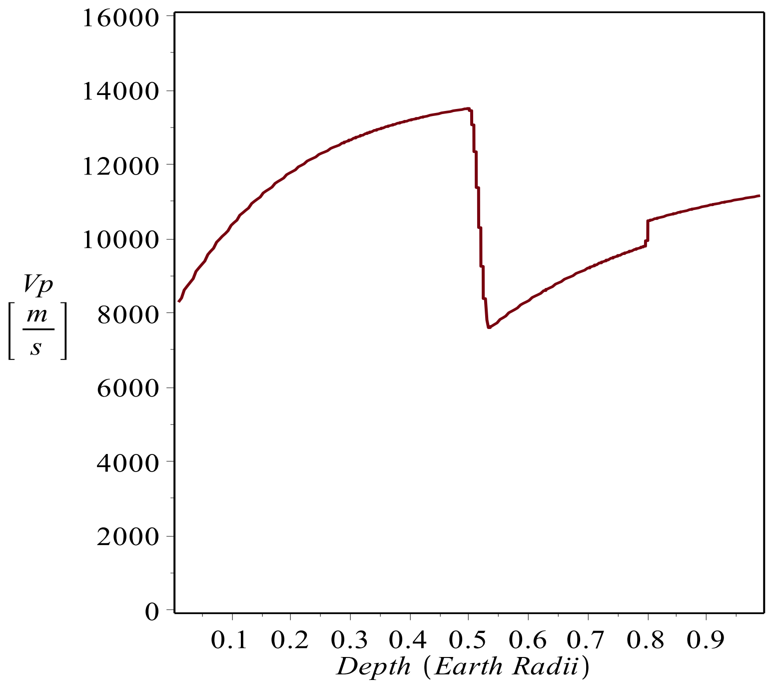
Approximated model

Representation through 128 **Fourier** (sin/cos)-basis functions

Example: A seismic model of the Earth



True model



Approximated model

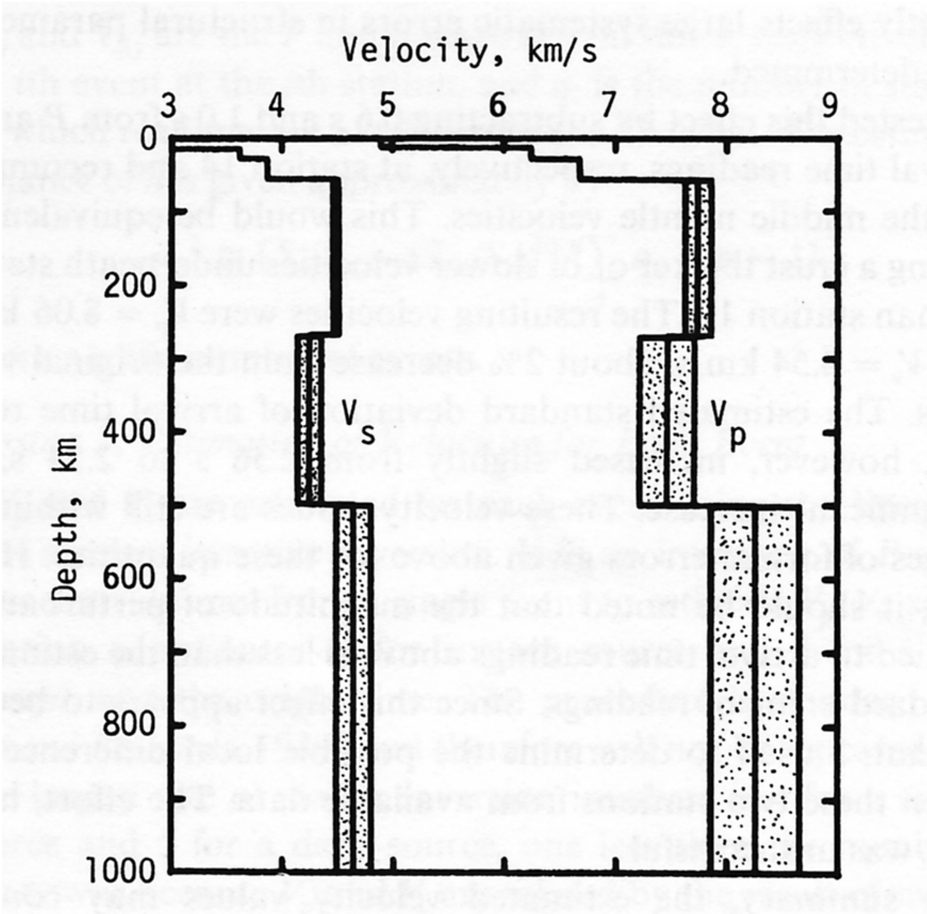
Representation through 256 **Haar**-basis functions

Note! We have invariant results from different bases!

- Even when two analysts choose different set of base functions, they will obtain (almost) the same model.
- The result is **invariant** under a change of base functions.
- The method is **consistent**: There is agreement between the results from different analysts.

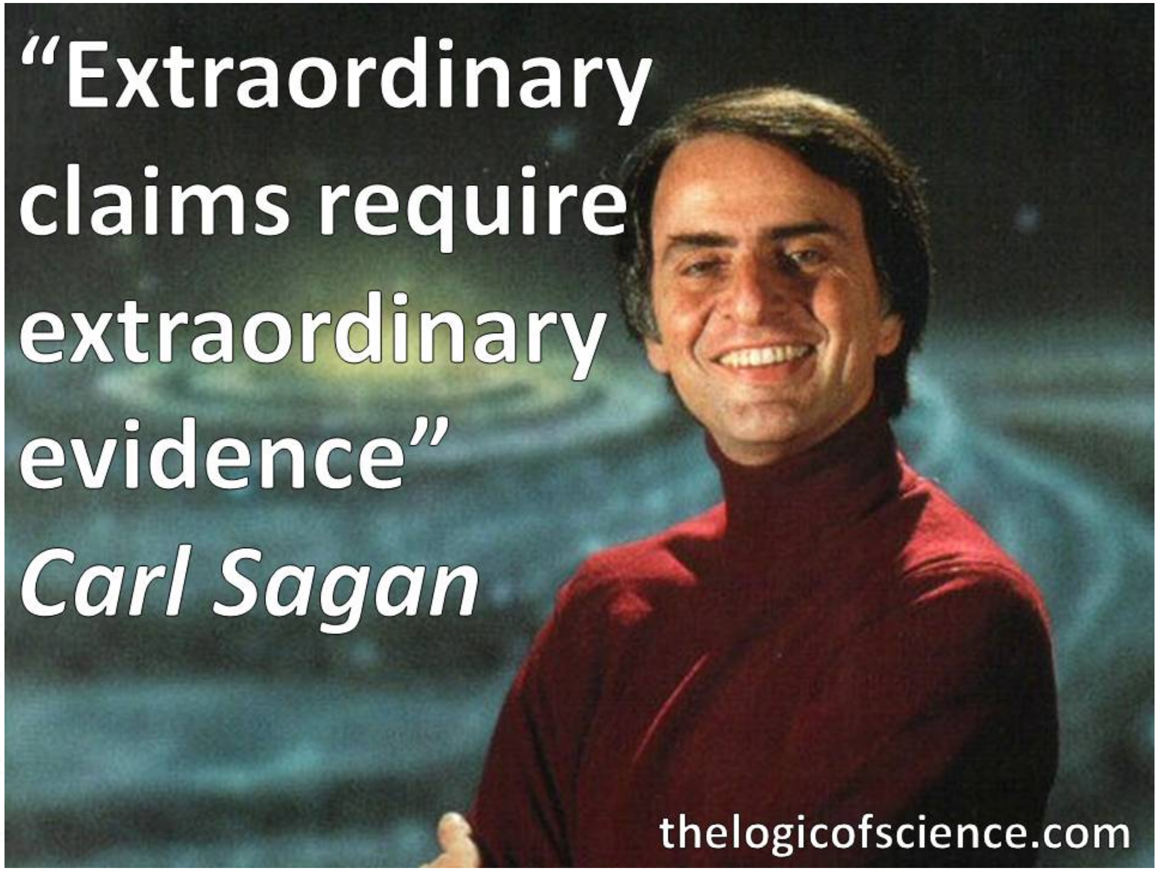
Models with few model parameters

Lunar seismic velocity profile to 1000 km depth



[From Nakamura, *JGR* **88**, 677-686, 1983]

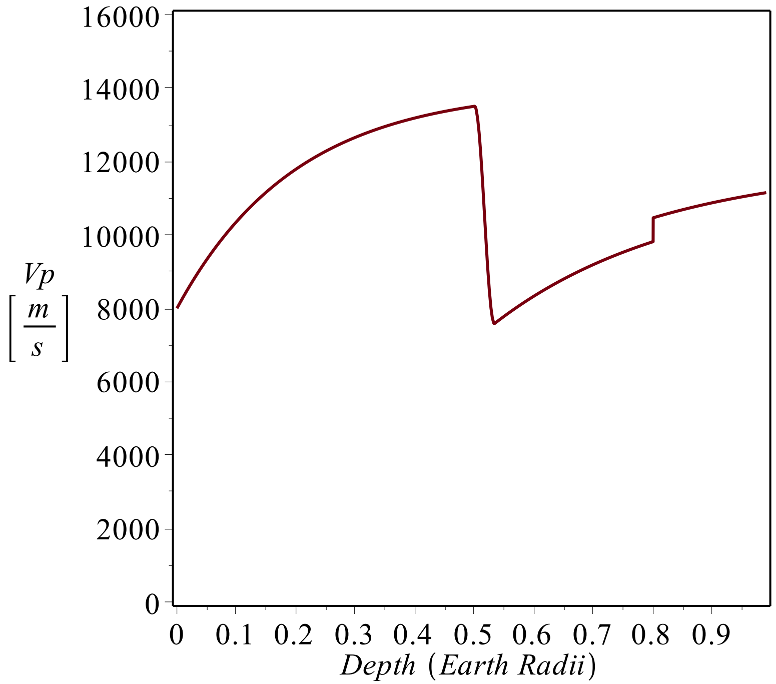
Sparse Models



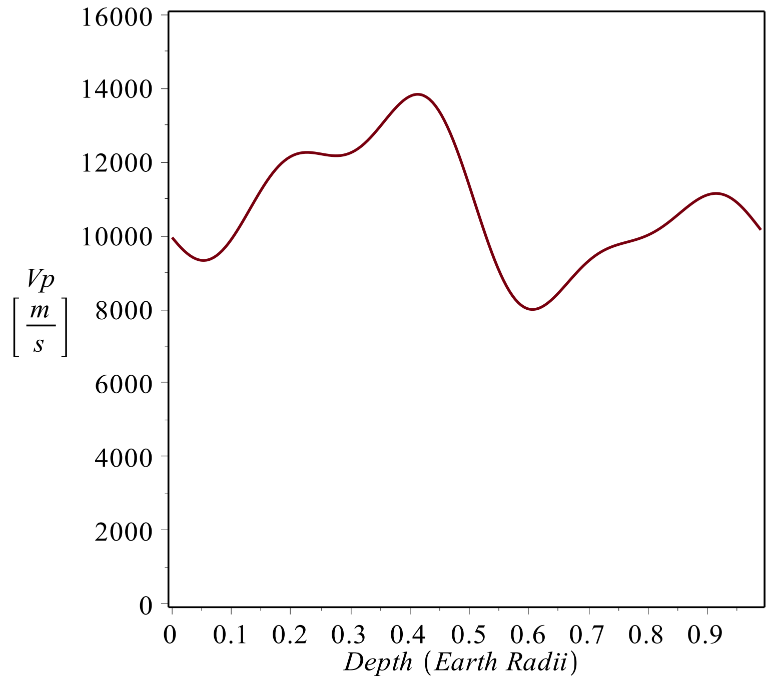
Reasons for sparsity

- To make the problem computationally feasible
- To build-in prior knowledge about structure
- To avoid unnecessary detail (Occam's Razor)

Example: Different sparse models of the Earth with the same misfit: The Fourier basis



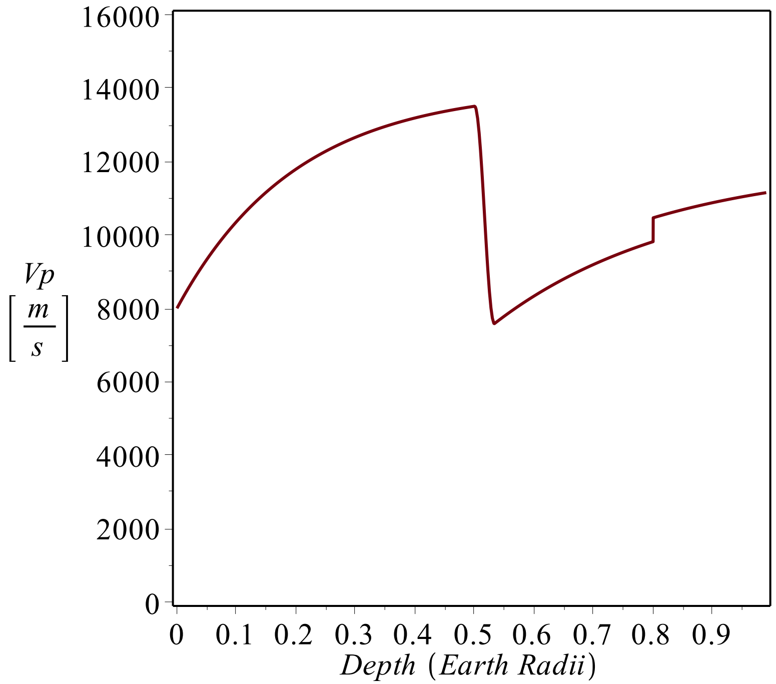
True model



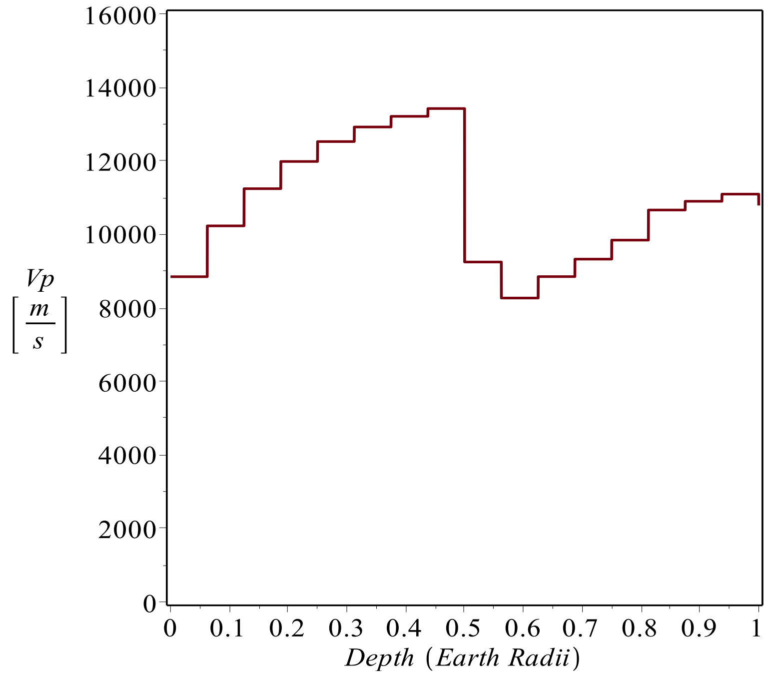
Approximated model

Representation through 4 **Fourier** (sin/cos)-basis functions

Example: Different sparse models of the Earth with the same misfit: The Haar basis



True model



Approximated model

Representation through 16 **Haar**-basis functions

Probabilities

Where do they come from?

What do they mean?

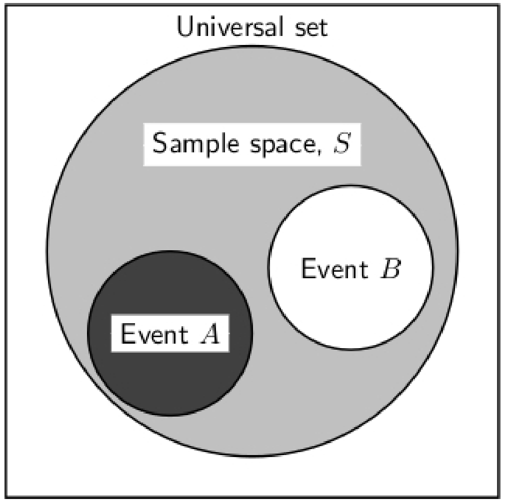
How can they be substantiated?

How can they be challenged?

The basic mathematics

The rules of probability: Kolmogorov (1933)

1. Probabilities lie between 0 and 1
2. The total probability of all possible outcomes is 1
3. The probability $P(A \text{ or } B)$ for two non-overlapping events A and B is equal to $P(A) + P(B)$

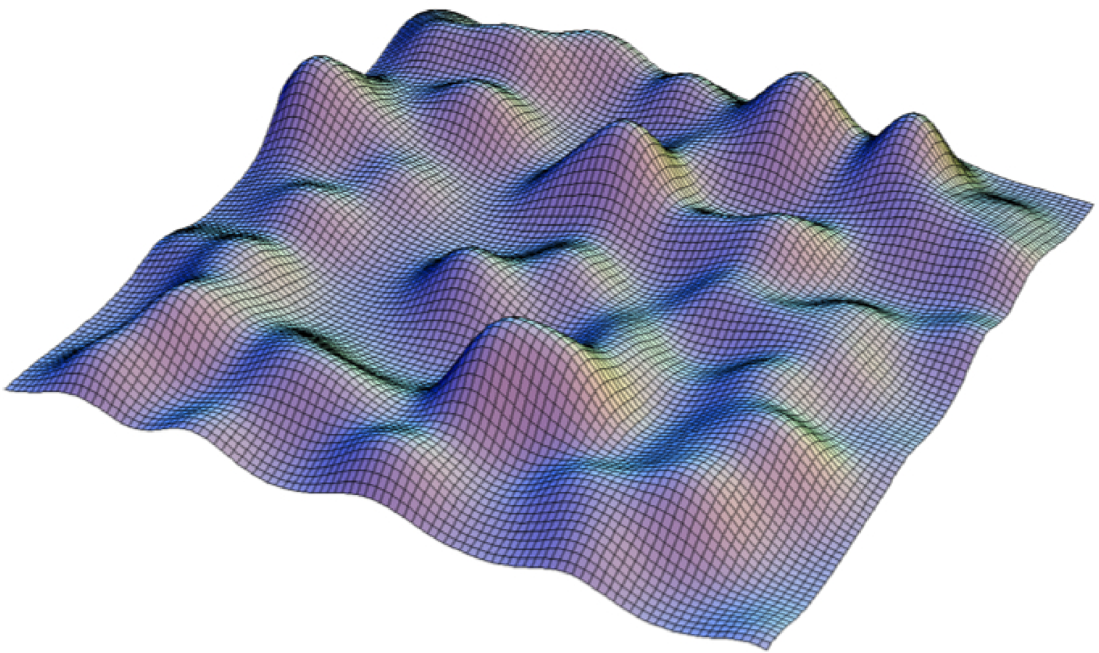


Andrei Nikolaevich Kolomogorov
(1903 – 1987)

$P(\mathcal{A})$ expressed through a probability density

A probability density function $p(\mathbf{x})$ over the parameter space $\mathcal{X}$, is defined such that

$$P(\mathcal{X}) = 1 \ , \quad \text{and} \quad P(\mathcal{A}) = \int_{\mathcal{A}} p(\mathbf{x}) d\mathbf{x} \ \text{ for } \ \mathcal{A} \subseteq \mathcal{X}. \qquad (1)$$

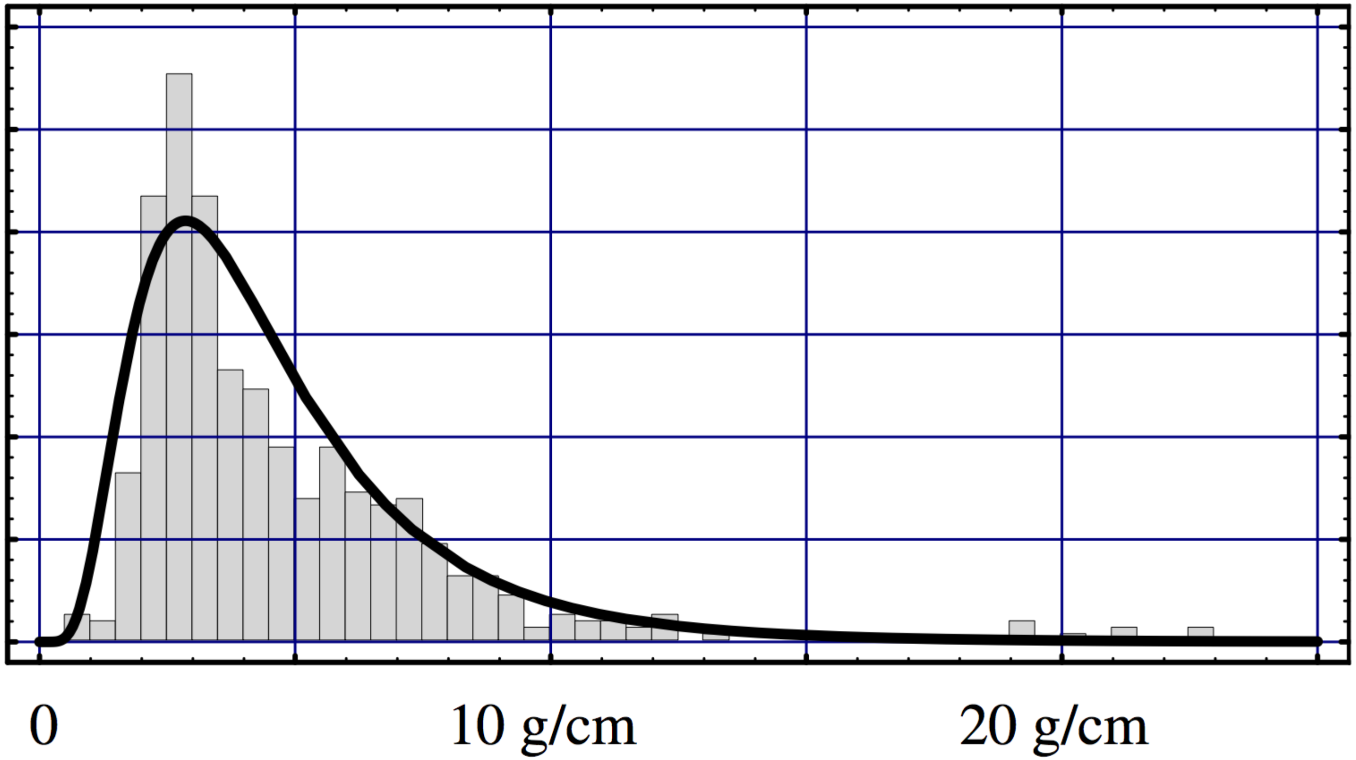


A multi-modal probability density function over a 2-D space

Types and Sources of Uncertainty

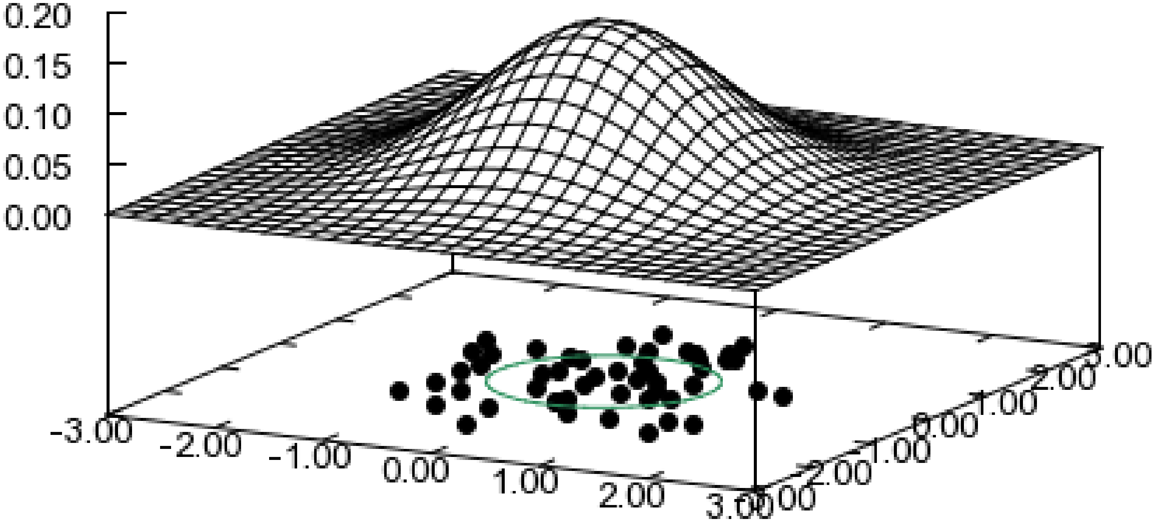
- Probability densities as limits of sampling densities
- Probability from symmetries
- Probability from subjective belief
- Beware of transformed probabilities!

Probability densities as limit sampling densities



Histogram of mass densities of 571 different known minerals in the Earth's crust (Johnson and Olhoeft, 1984)

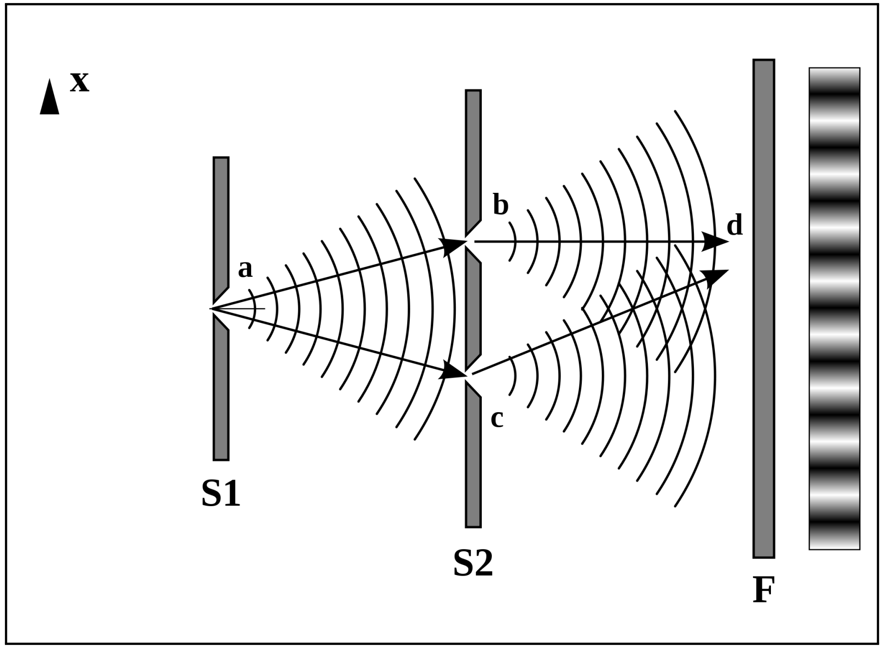
Probability densities as limit sampling densities

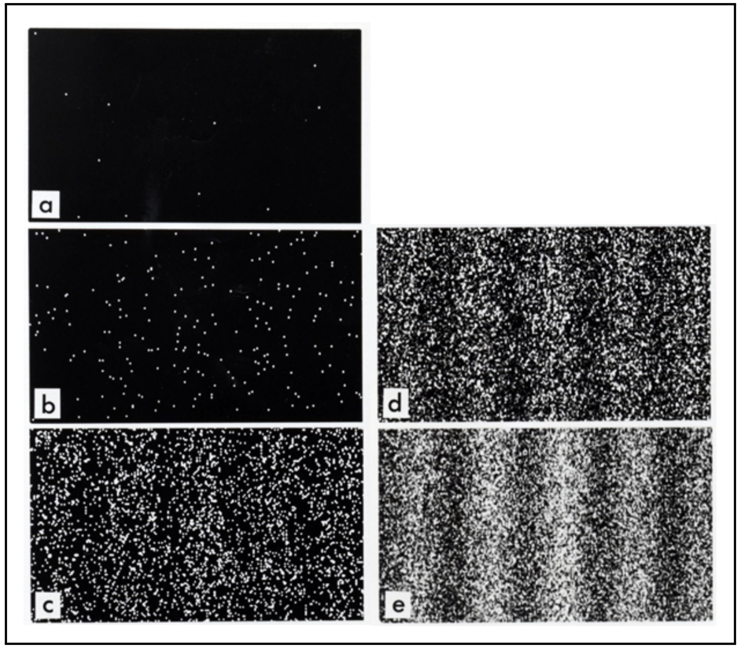


2D probability density as the limit of a 2D sampling density

In high-dimensional spaces, we can also view a probability density as a limit sample density. This is the case when we use Monte Carlo methods to sample the probability density of the solution to an inverse problem (the posterior probability density).

Probability densities as limit sampling densities

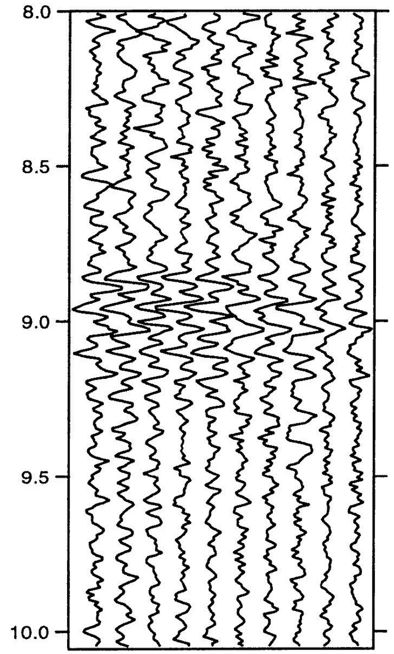




Young's experiment and the quantum dualism between particles and waves

The assumtion of stationarity in time and space: A way of getting more samples

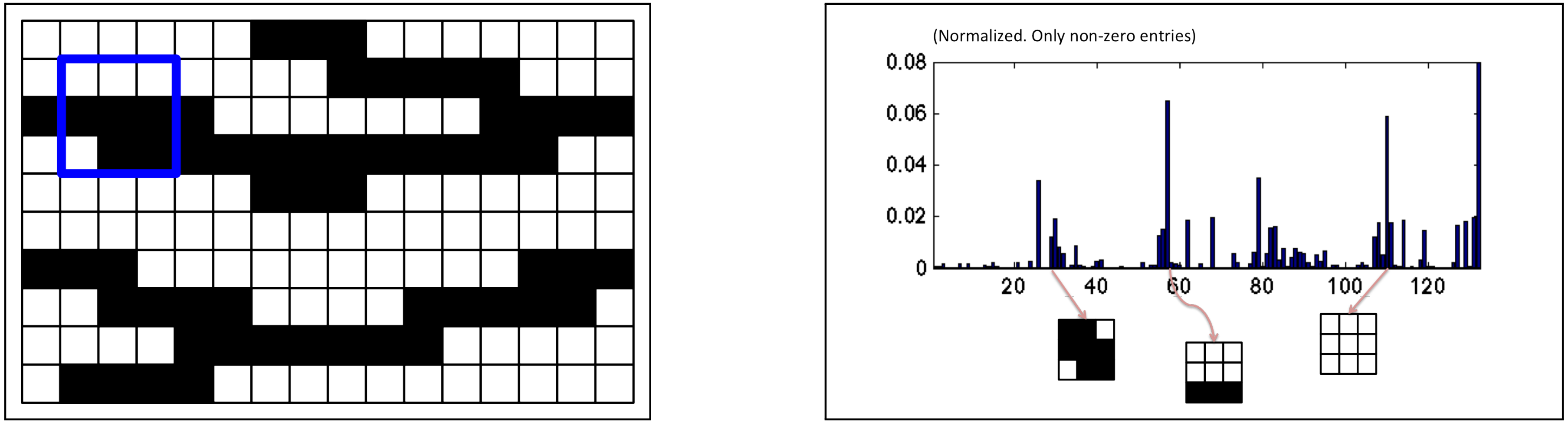
1. In order to obtain many samples it is, in practice, required to use samples from a wide range of points in space/time.
2. It is therefore necessary to assume that these samples satisfy *stationarity*: that their probability distribution is independent of space /time.



Seismic recordings with noise. The noise distribution may be found by assuming that the noise is stationary in time- and space.

Example of the use of stationarity in space: Sequential simulation

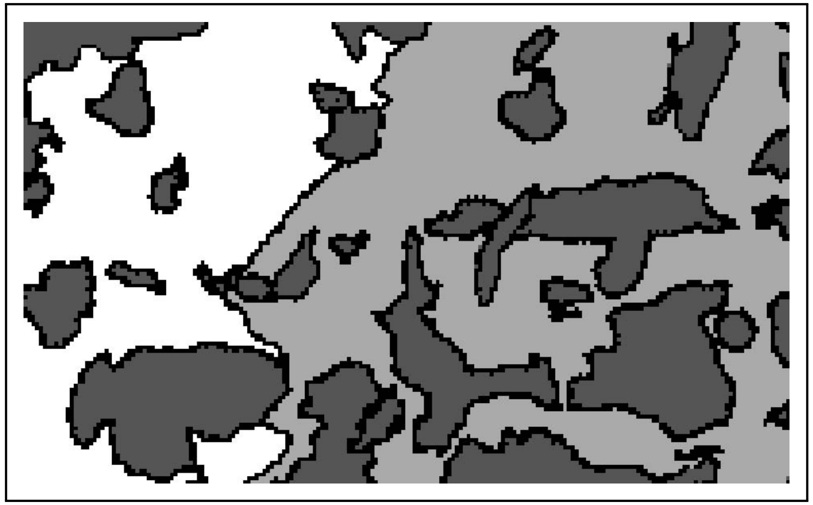
- Pattern frequency distributions obtained from a *training image*.
- The method assumes that the training image is *stationary*, to ensure that patterns from the same distribution can be sampled at different locations in the image.

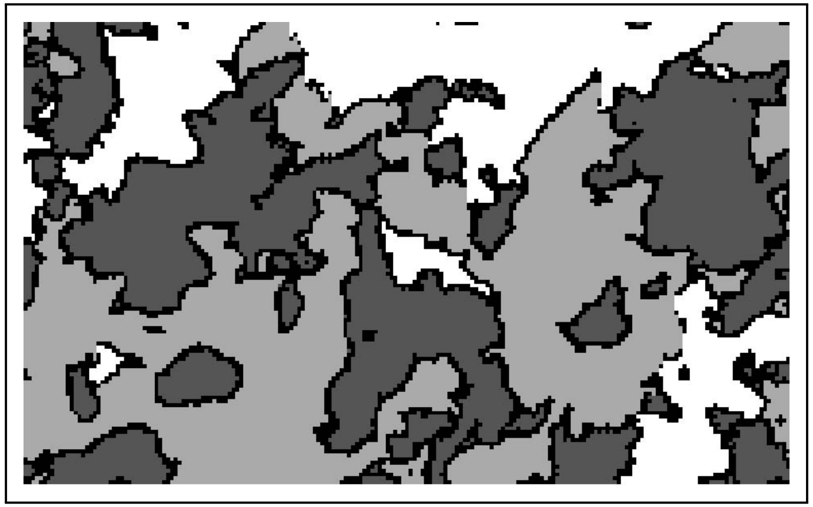


A training image (left), and its pattern histogram.

Example of the use of stationarity in space: Sequential simulation

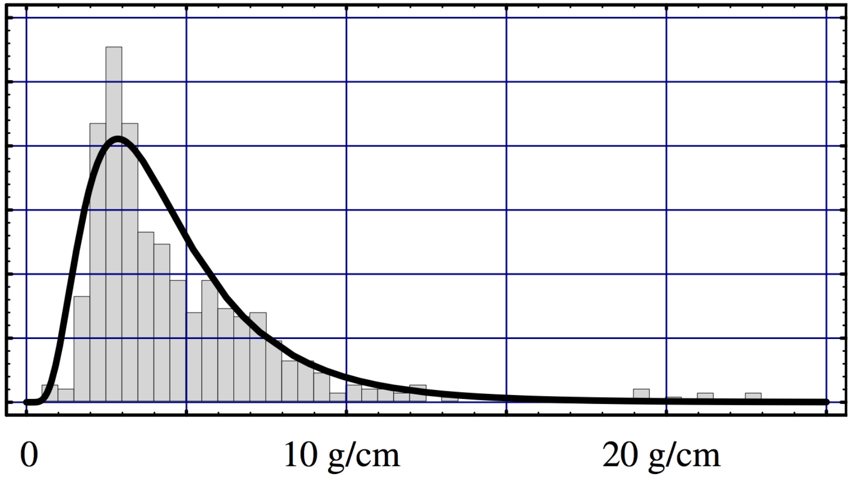
- If the pattern histogram obtained from the training image is used as a pattern *probability density*, we can generate new patterns from this density, and create new images.





A training image (left), and a new realization generated from the pattern histogram (right).

Challenging the probability distribution: Is the histogram a likely outcome of a sampling experiment?



- For the probability distribution to be an acceptable prediction, our histogram must be a likely result of the sampling process

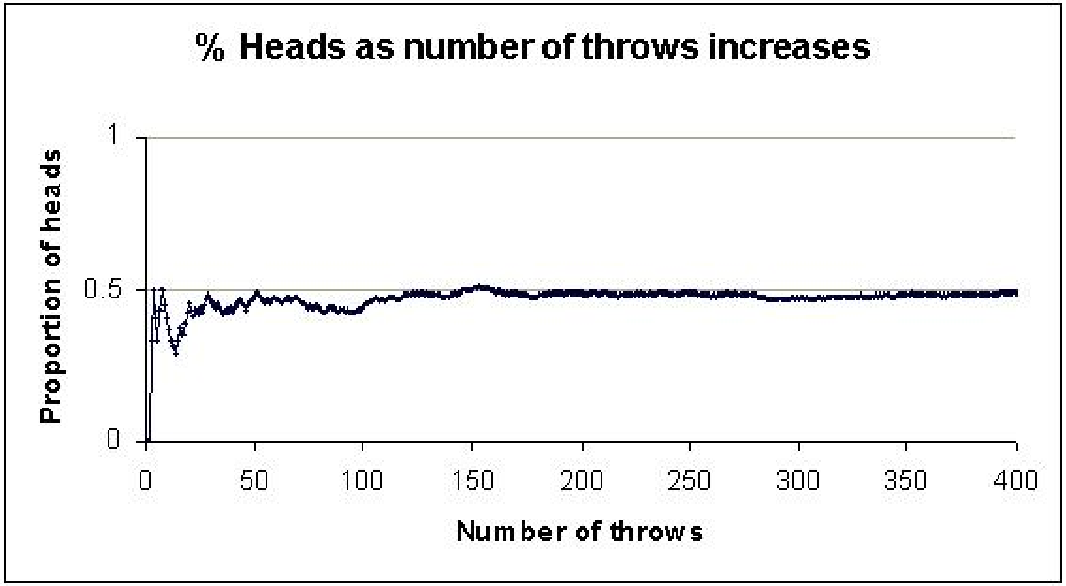
The probability of getting the histogram $\pi_1, \ldots, \pi_K$ with N counts, is given by the *Multinomial Distribution* $p_1, \ldots, p_K$:

$$P(\pi_1, \ldots, \pi_K) = \frac{N!}{\pi_1! \ldots \pi_K!} \; {p_1}^{\pi_1} \ldots {p_K}^{\pi_K} \; .$$

However, the fact that a histogram has a high probability does **not** guarantee that the probability distribution is correct!

Challenging the probability distribution: Is the histogram a likely outcome of a sampling experiment?

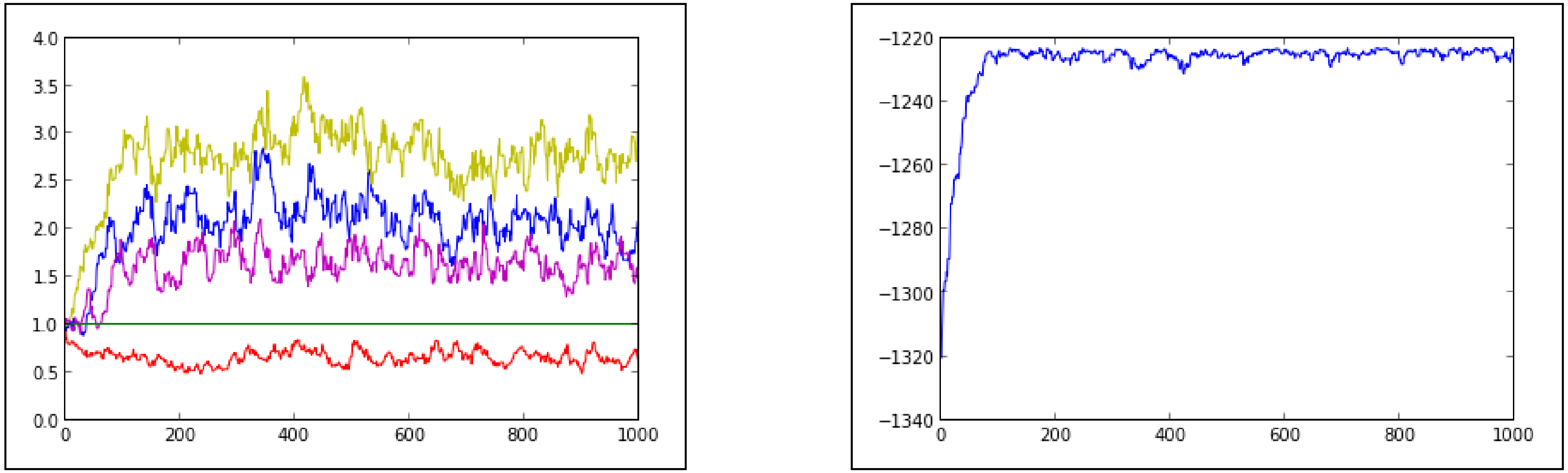
In principle, building a probability density from a histogram (or a 'cloud of sample points') should be done by observing how the sampling density/heights of the histogram columns evolves during the sampling.



Percentage of heads and tails for an increasing number of tosses of a fair coin

Challenging the probability distribution: Is the histogram a likely outcome of a sampling experiment?

Monitoring the evolution of the sampling density or histogram heights (or any function hereof) during Monte Carlo sampling will reveal if the sampling density is not yet close to the probability density.

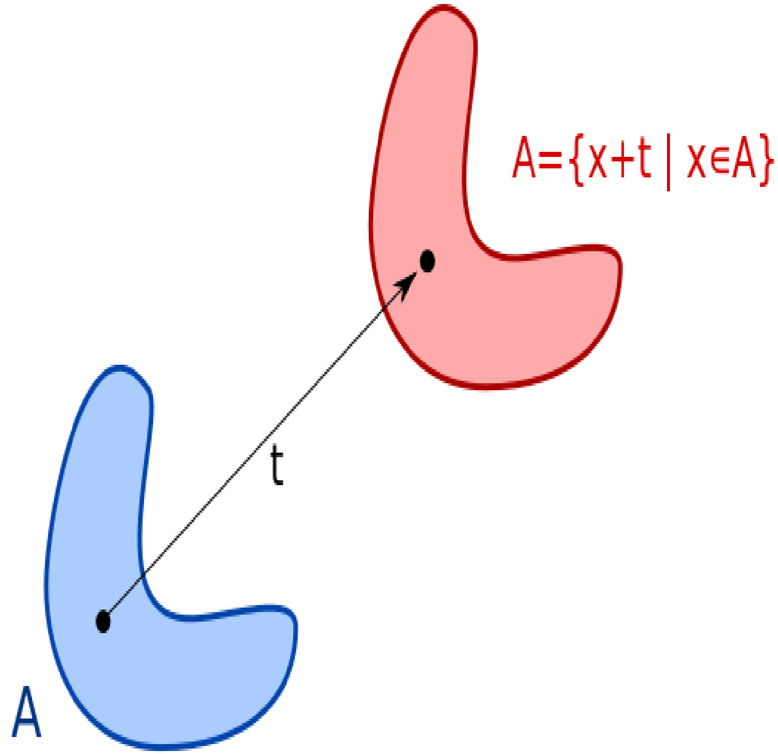


Evolution of parameter values (left) and data misfit (right) during a Monte Carlo sampling of solutions to an inverse problem (Univ. Texas at Austin, 2013)

Probability from symmetries: Rotation invariance

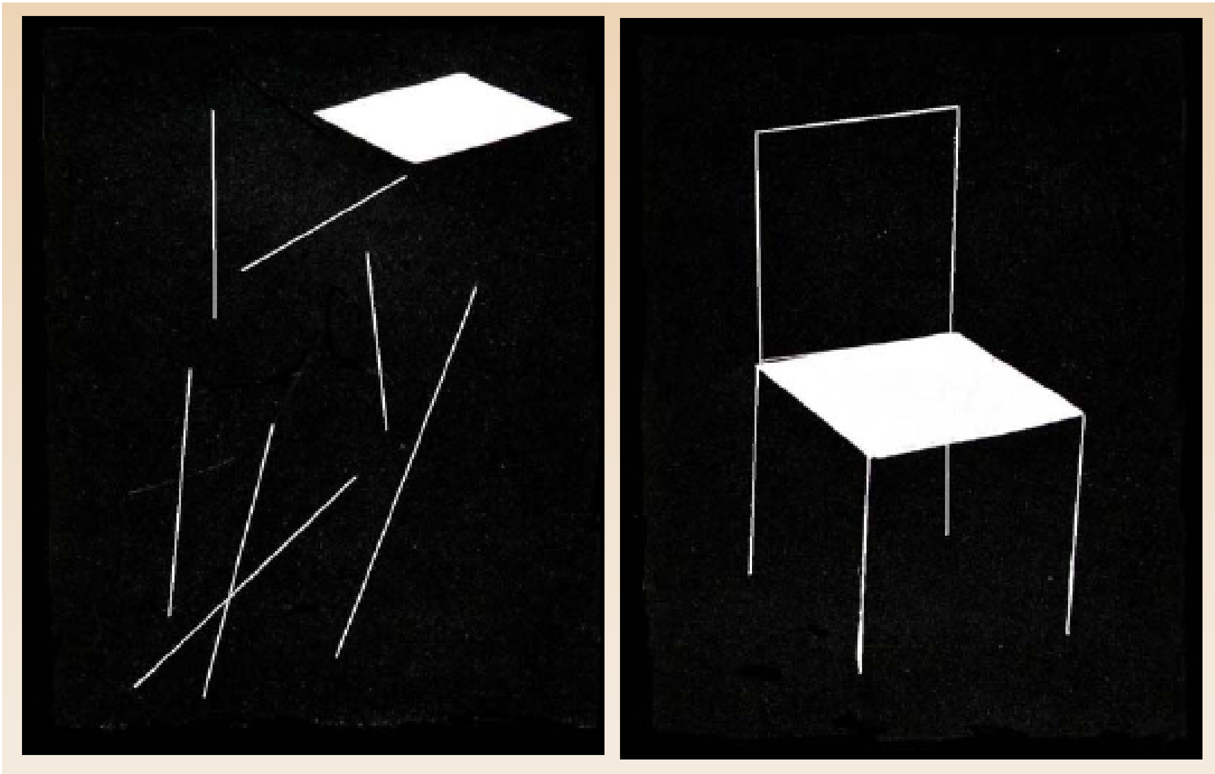
Random generators based on rotation invariance

Probability from symmetries: Translation invariance



Translation invariance means that two similar, translated volumes have the same probability

Probability from subjective belief

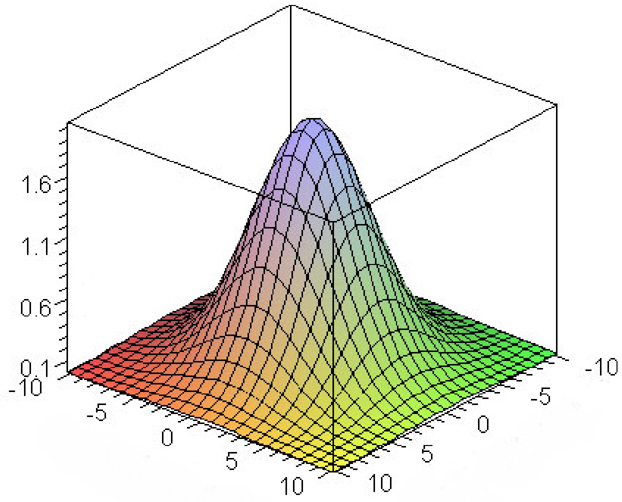


Ames (1950)

Best practice:

Avoid probabilities based on purely subjective belief!

Probability from subjective belief



Gaussian distribution centered at the 'best guess' and with a dispersion expressing how unsure the analyst is.

- Subjective probabilities are probabilities without an empirical or theoretical basis.
- Subjective probabilities are personal and therefore *inconsistent*.

Best practice:

Avoid probabilities based on purely subjective belief!

The most common application of subjective probabilities

Consider a linear inverse problem

$$\mathbf{d} = \mathbf{G}\mathbf{m} \tag{2}$$

Two possible solutions to this problem are:

1. Bayesian (Stochastic) inversion:

$$\mathbf{m} = (\mathbf{G}^T\mathbf{C}_D^{-1}\mathbf{G} + \mathbf{C}_M^{-1})^{-1}\mathbf{G}^T\mathbf{C}_D^{-1}\mathbf{d}_{obs}$$

   where $\mathbf{C}_M$ defines a Gaussian probability distribution often chosen from subjective belief.

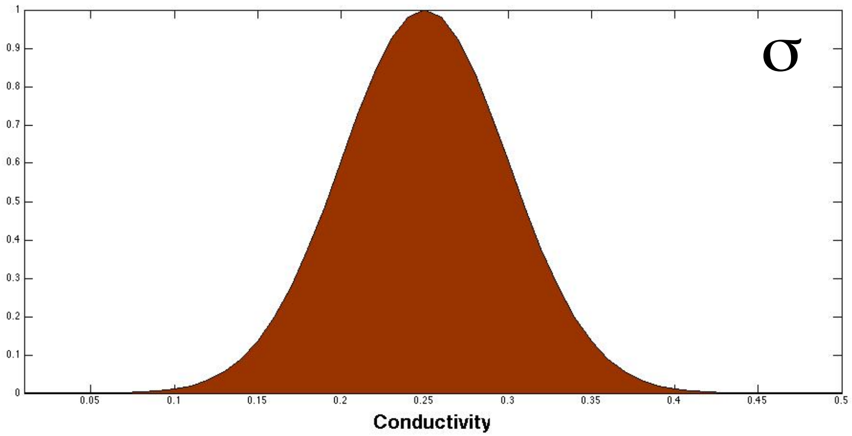
2. If we put $\mathbf{C}_D = \mathbf{I}$ and $\mathbf{C}_M = \frac{1}{\epsilon^2}\mathbf{I}$, this expression is equal to the expression used in inversion through Tikhonov Regularization:

$$\mathbf{m} = (\mathbf{G}^T\mathbf{G} + \epsilon^2\mathbf{I})^{-1}\mathbf{G}^T\mathbf{d}_{obs}$$

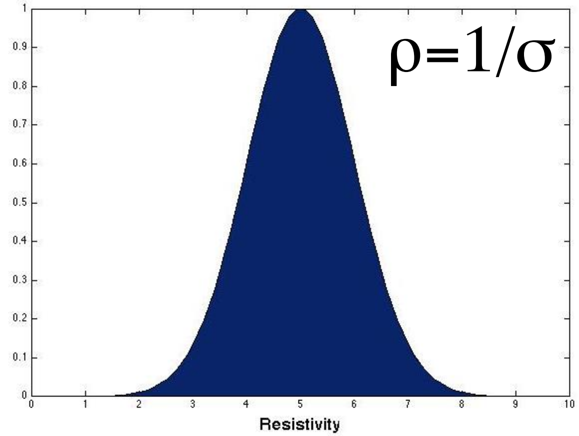
   where ϵ is an undefined, arbitrary parameter to be determined from external considerations.

Conflict between subjective probabilities

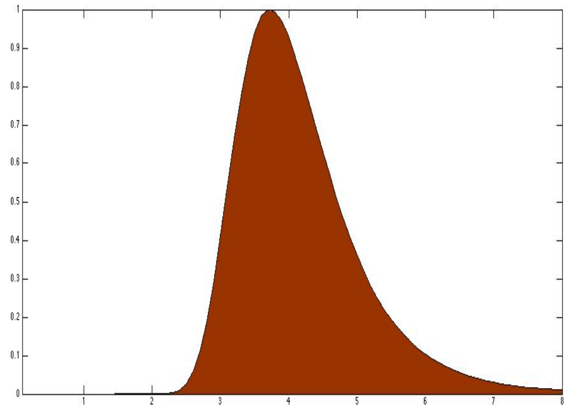
Bob's conductivity distribution



Alice's resistivity distribution:



Bob's computed resistivity distribution:



$$f_A(\rho) = \frac{1}{s_\rho\sqrt{2\pi}} \exp\left(-\frac{(\rho - \rho_{obs})^2}{2{s_\rho}^2}\right)$$

$$g_B(\sigma) = \frac{1}{s_\sigma\sqrt{2\pi}} \exp\left(-\frac{(\sigma - \sigma_{obs})^2}{2{s_\sigma}^2}\right)$$

$$f_B(\rho) = \left|\frac{d\sigma}{d\rho}\right| g_B(\sigma) = \frac{1}{\rho^2}\frac{1}{s_\sigma\sqrt{2\pi}} \exp\left(-\frac{(1/\rho - \sigma_{obs})^2}{2{s_\sigma}^2}\right)$$

PROBABILISTIC INVERSION

Bayes Theorem

From the definition of conditional probability density

$$f(\mathbf{x} \mid \mathbf{y}) \equiv \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{y})}$$

we get:

$$f(\mathbf{m} \mid \mathbf{d}) = \frac{f(\mathbf{d} \mid \mathbf{m}) f(\mathbf{m})}{f(\mathbf{d})}$$

Bayes Theorem

From the definition of conditional probability density

$$f(\mathbf{x}\,|\,\mathbf{y}) \equiv \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{y})}$$

we get:

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

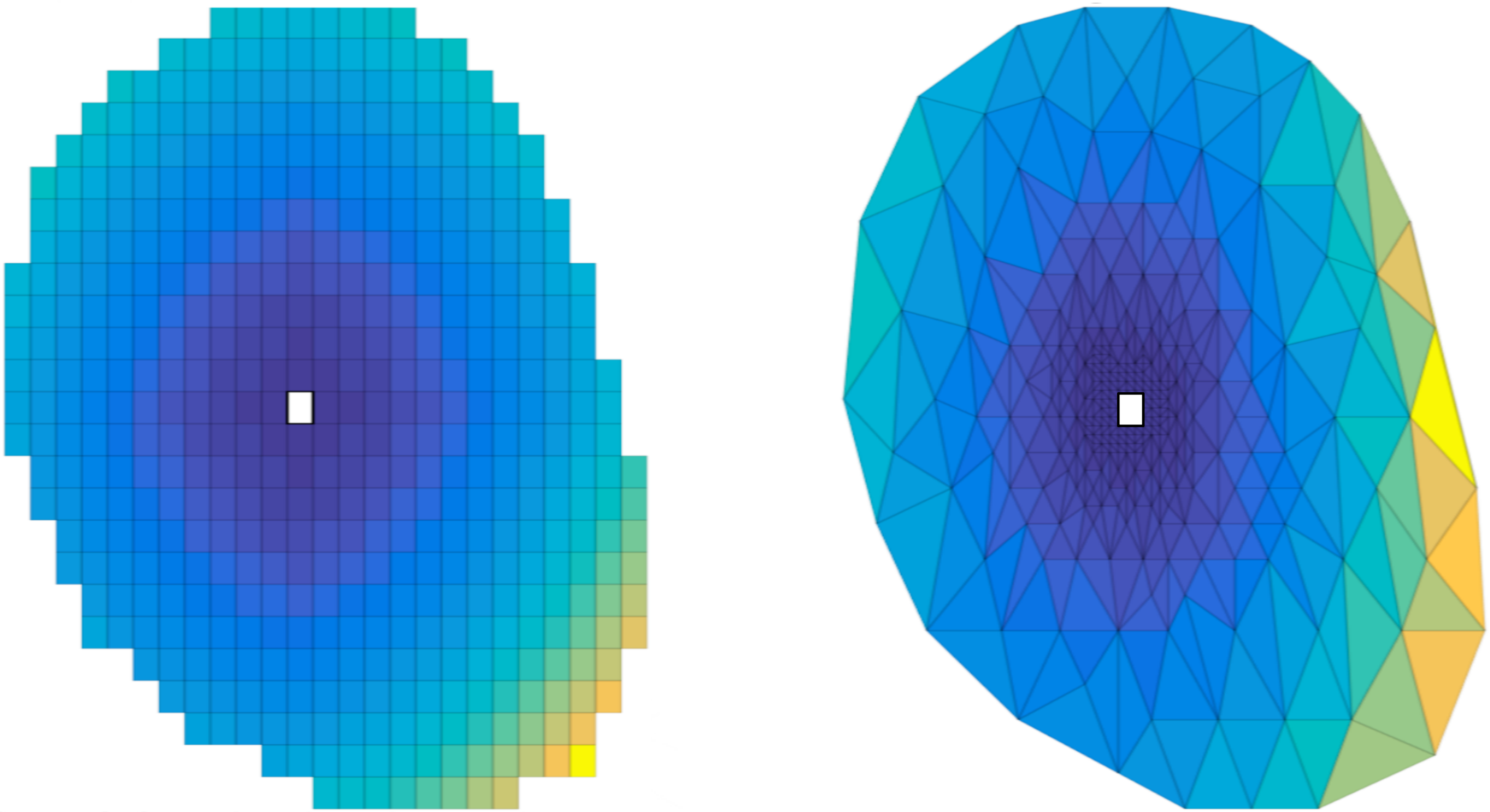
posterior

likelihood

prior

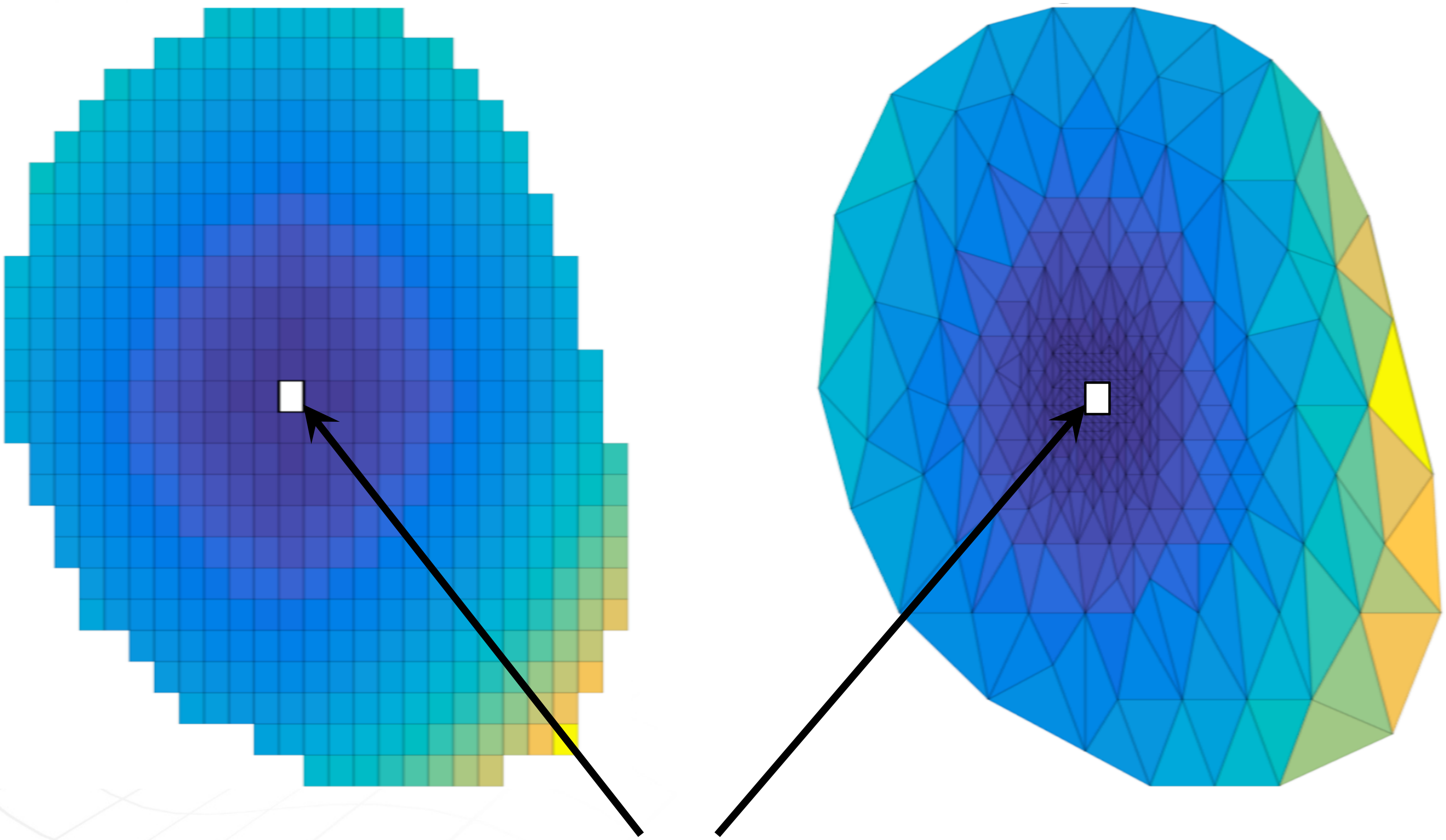
CONDITIONAL PROBABILITIES AND THEIR INHERENT INCONSISTENCY

Computing a conditional probability density



K.-A. Lie et al. (2012)

Computing a conditional probability density

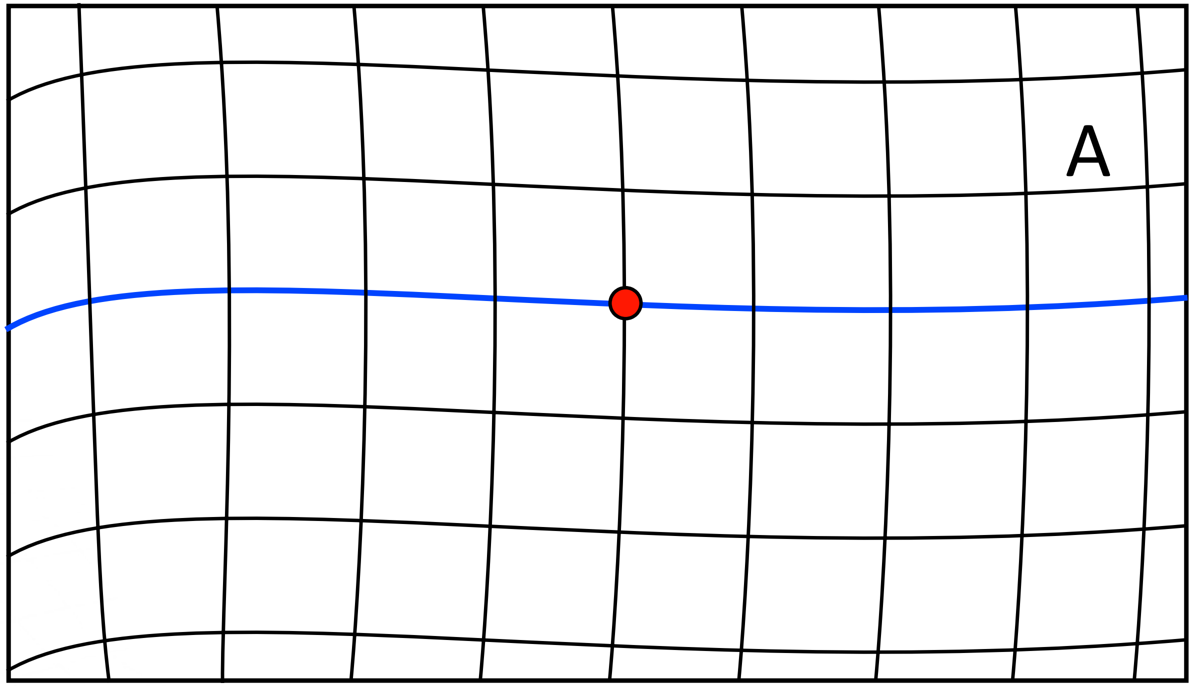


Same conditional PDF ?

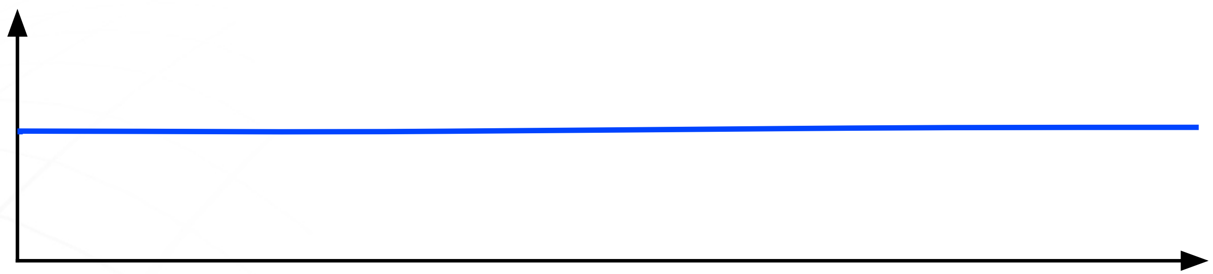
K.-A. Lie et al. (2012)

The Borel Paradox

- Near-Cartesian reference frame
- Equal volumes have equal probabilities

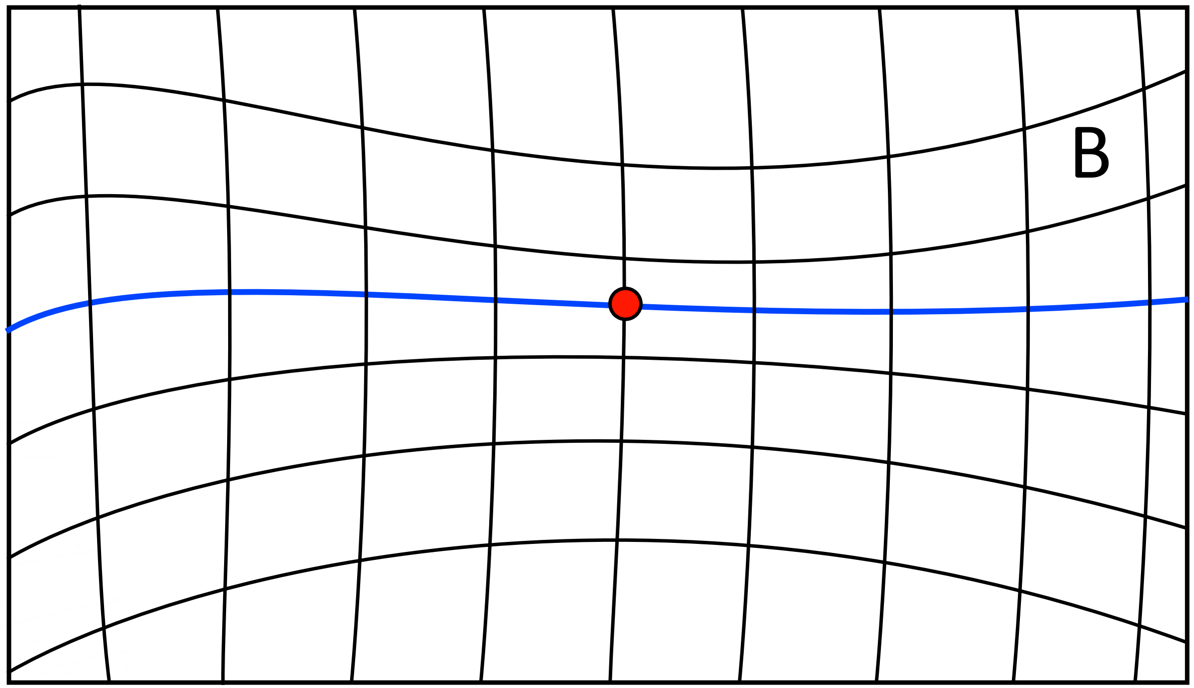


- Conditional probability density is constant

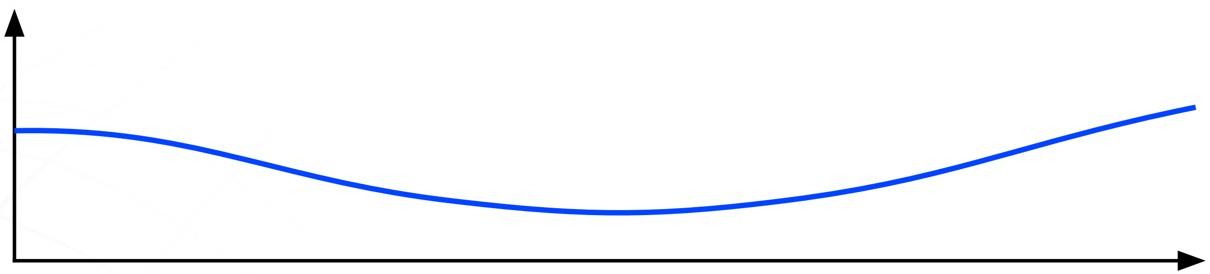


The Borel Paradox

- Non-Cartesian reference frame
- Equal volumes have equal probabilities

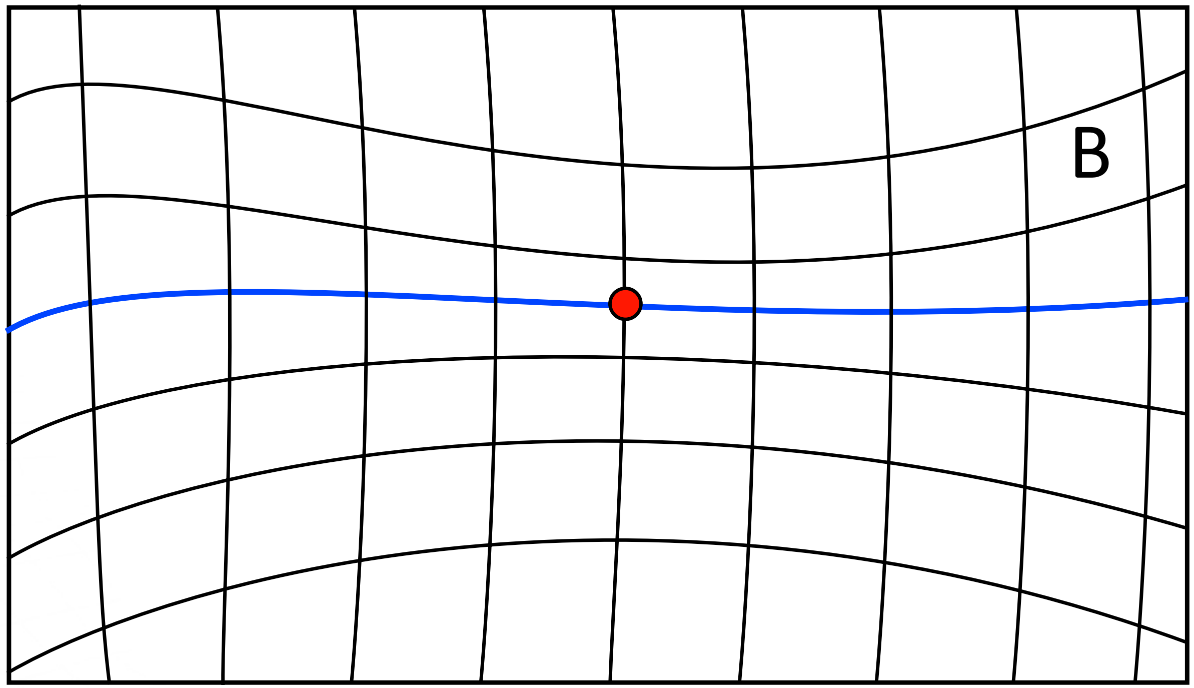


- Conditional probability density is **not** constant

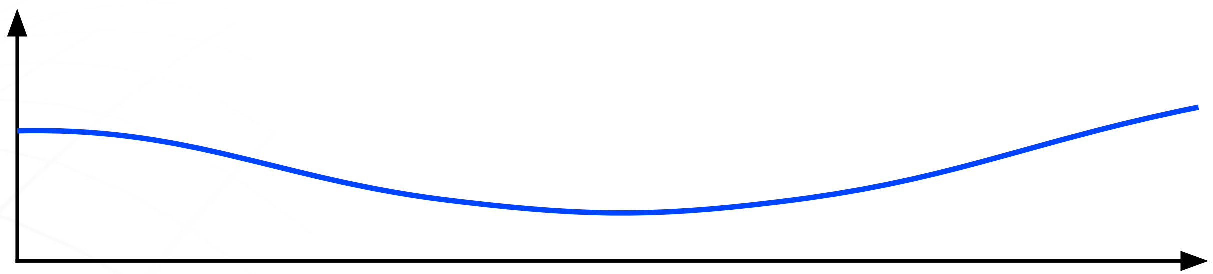


The Borel Paradox

- The metric of the blue subspace **is unchanged**.

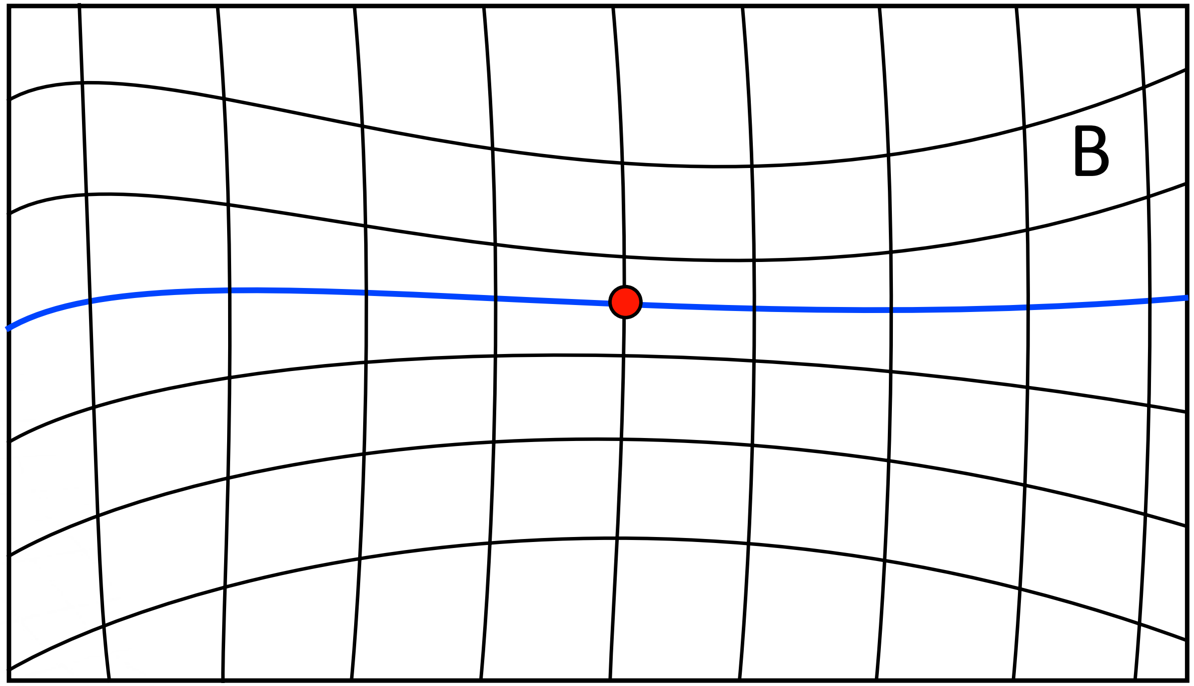


- Conditional probability density **has changed**

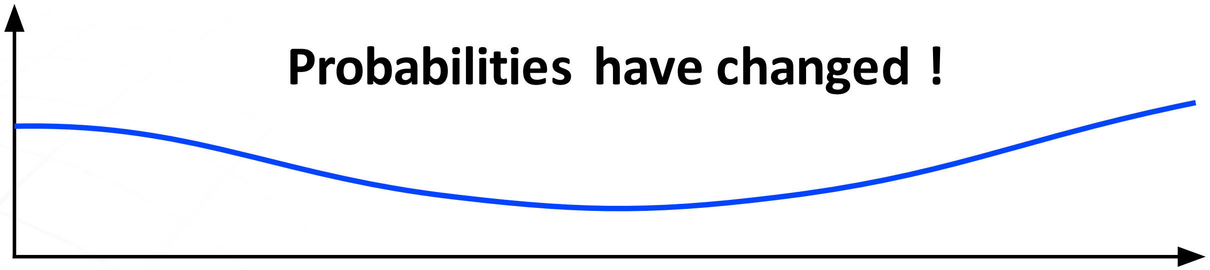


The Borel Paradox

- The metric of the blue subspace **is unchanged**.



- Conditional probability density **has changed**



Conclusion on conditional probability densities

- Conditional probability densities are **inconsistent**, because different analysts may arrive at different (conflicting) results.

Borel's paradox disappears if $f(\mathbf{x})$ is replaced with

$$g(\mathbf{x}) = f(\mathbf{x}) / \mu(\mathbf{x})$$

where $\mu(\mathbf{x})$ is a nonzero *volume density*

(Mosegaard and Tarantola, 2002)

Constructing a probabilistic solution

Constructing a probabilistic solution

From the laws of physics:

$$\boldsymbol{d}_{\boldsymbol{pred}} = g(\boldsymbol{m})$$

Noise $\boldsymbol{n}$ contaminating the observed data:

$$\boldsymbol{d} = \boldsymbol{d}_{\boldsymbol{pred}} + \boldsymbol{n}$$

Assuming that $f_{\boldsymbol{n}}(\boldsymbol{n})$ is the density of the noise, we have the *likelihood function*:

$$f(\boldsymbol{d}|\boldsymbol{m}) = f_{\boldsymbol{n}}\big(\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m})\big)$$

Constructing a probabilistic solution

Likelihood function:

$$f(\boldsymbol{d}|\boldsymbol{m}) = f_{\boldsymbol{n}}\big(\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m})\big)$$

Assuming a normal (Gaussian) distribution with zero mean and variance σ^2:

$$f_{\boldsymbol{n}}(\boldsymbol{n}) = \exp\left(-\frac{\boldsymbol{n}^2}{2\sigma^2}\right)$$

giving

$$f(\boldsymbol{d}|\boldsymbol{m}) = \mathbf{exp}\left(-\frac{\big(\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m})\big)^{\mathbf{2}}}{2\sigma^2}\right)$$

Constructing a probabilistic solution

If we have the marginal probability density $f(\boldsymbol{m})$, often termed the *prior density* $f(\boldsymbol{m})$, we have

$$f(\boldsymbol{m}, \boldsymbol{d}) = f(\boldsymbol{d}|\boldsymbol{m}) f(\boldsymbol{m})$$

If we have observed a concrete realization $\boldsymbol{d_{obs}}$ of $\boldsymbol{d}$, we can compute

$$p(\boldsymbol{m}) = f(\boldsymbol{m}, \boldsymbol{d_{obs}})$$

known as the *posterior distribution* of $\boldsymbol{m}$.

The linear Gaussian problem

The likelihood function:

$$f(\boldsymbol{d}|\boldsymbol{m}) = exp\left(-\frac{1}{2}(\boldsymbol{d} - \boldsymbol{Gm})^T \boldsymbol{C}_{\boldsymbol{n}}^{-1}(\boldsymbol{d} - \boldsymbol{Gm})\right)$$

where $\boldsymbol{C}_{\boldsymbol{n}}$ is the noise covariance matrix, and

$$f(\boldsymbol{m}) = exp\left(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_0)^T \boldsymbol{C}_{\boldsymbol{m}}^{-1}(\boldsymbol{m} - \boldsymbol{m}_0)\right)$$

where $\boldsymbol{m}_0$ is the center (mean) of the prior density, and $\boldsymbol{C}_{\boldsymbol{m}}$ is the prior model covariance matrix.

From the above expressions we get:

$$p(\boldsymbol{m}) = \mathbf{exp}(-\boldsymbol{S}(\boldsymbol{m}))$$

where

$$\boldsymbol{S}(\boldsymbol{m}) = \frac{1}{2}\left[(\boldsymbol{d} - \boldsymbol{Gm})^T \boldsymbol{C}_{\boldsymbol{n}}^{-\mathbf{1}}(\boldsymbol{d} - \boldsymbol{Gm}) + (\boldsymbol{m} - \boldsymbol{m}_0)^T \boldsymbol{C}_{\boldsymbol{m}}^{-\mathbf{1}}(\boldsymbol{m} - \boldsymbol{m}_0)\right].$$

The linear Gaussian problem

It can be shown that the posterior

$$p(\boldsymbol{m}) = \mathbf{exp}(-\boldsymbol{S}(\boldsymbol{m}))$$

with

$$\boldsymbol{S}(\boldsymbol{m}) = \frac{1}{2}\left[(\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m})^T \boldsymbol{C}_{\boldsymbol{n}}^{-1}(\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m}) + (\boldsymbol{m} - \boldsymbol{m}_0)^T \boldsymbol{C}_{\boldsymbol{m}}^{-1}(\boldsymbol{m} - \boldsymbol{m}_0)\right]$$

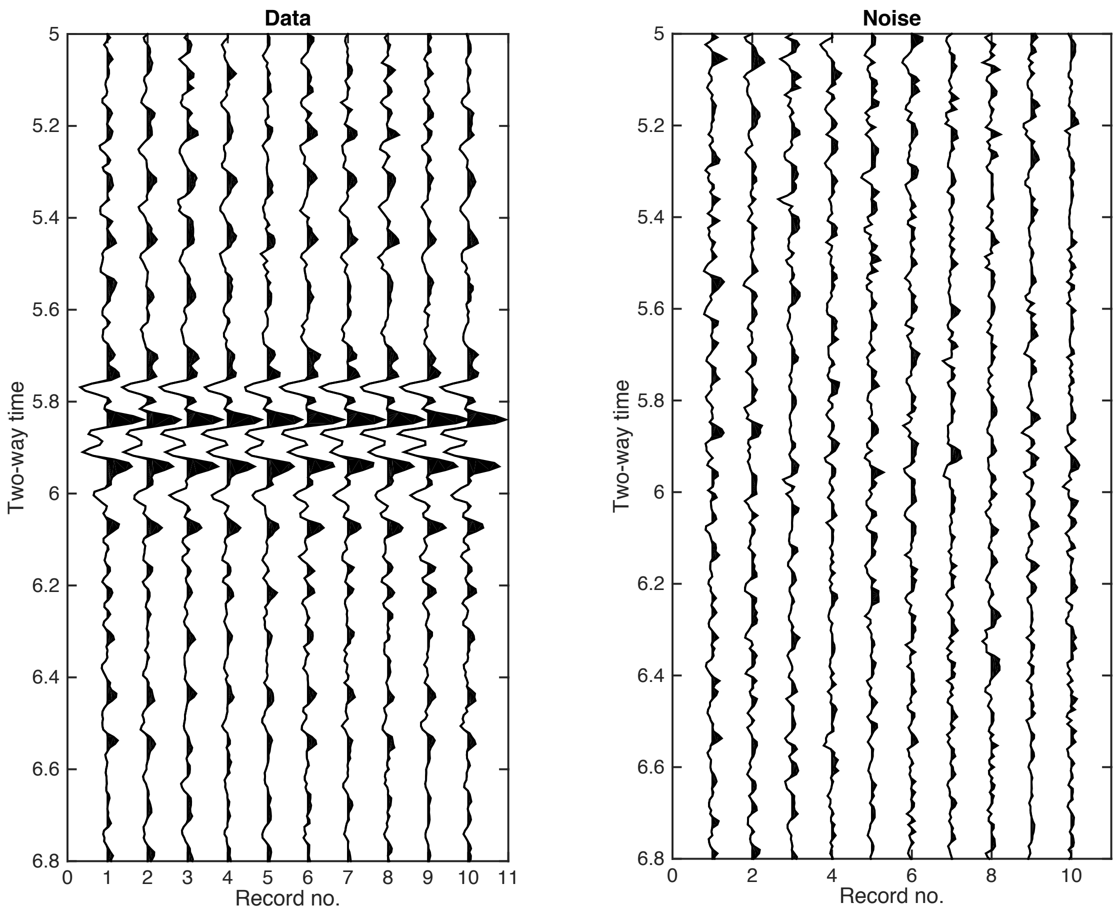
has mean

$$\boldsymbol{m}_{post} = \boldsymbol{m}_0 + \left(\boldsymbol{G}^T \boldsymbol{C}_n^{-1} \boldsymbol{G} + \boldsymbol{C}_{\boldsymbol{m}}^{-1}\right)^{-1} \boldsymbol{G}^T \boldsymbol{C}_{\boldsymbol{n}}^{-1}(\boldsymbol{d} - \boldsymbol{G}\boldsymbol{m}_0)$$

and covariance

$$\boldsymbol{C}_{post} = \left(\boldsymbol{G}^T \boldsymbol{C}_n^{-1} \boldsymbol{G} + \boldsymbol{C}_{\boldsymbol{m}}^{-1}\right)^{-1}.$$

Example: Inversion of seismic data



A selection of deep seismic reflections from the Earth's lithosphere (DRUM profile, BIRPS, 1984). Right: The noise found by assuming horizontal stratification and temporal and spatial stationarity of noise in the data.

Example: Inversion of seismic data

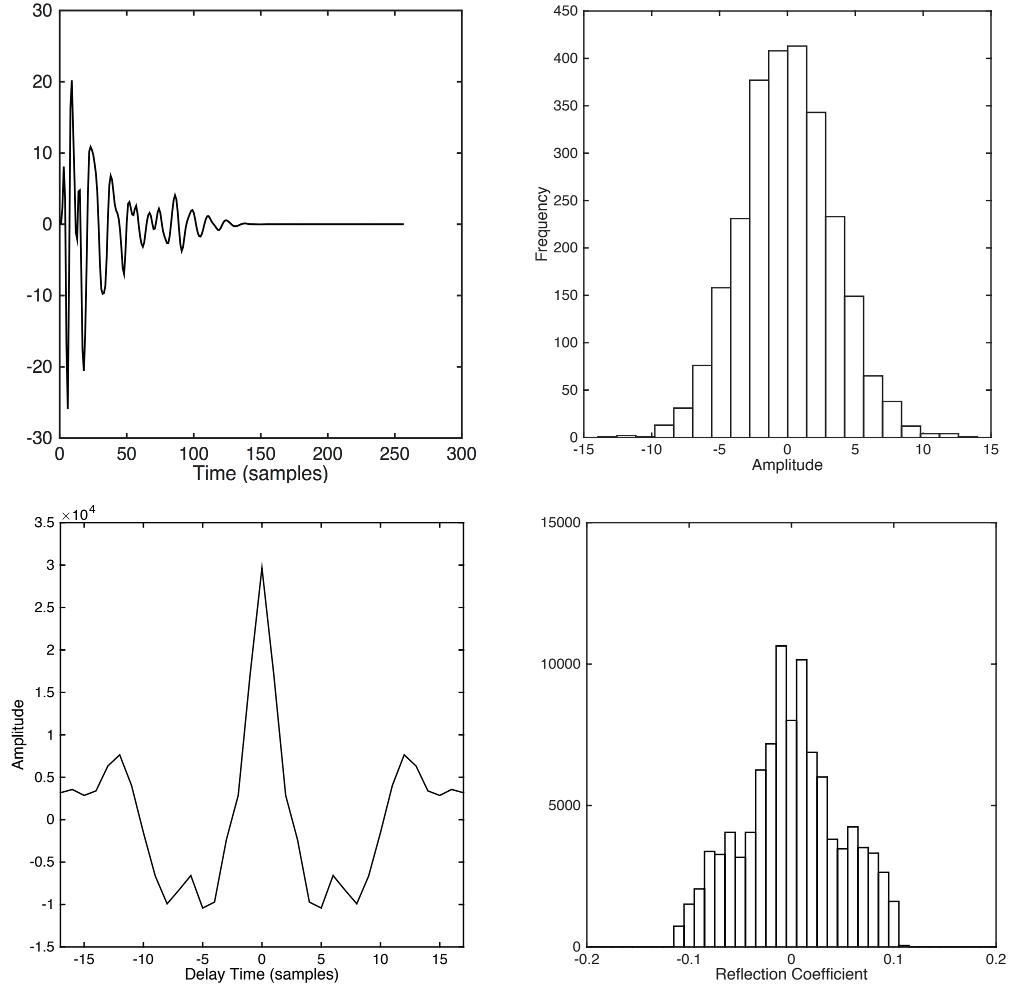
At the ith surface point we measure the seismogram

$$\boldsymbol{d_i} = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{iN} \end{pmatrix} = \boldsymbol{G}\boldsymbol{m_i}$$

where

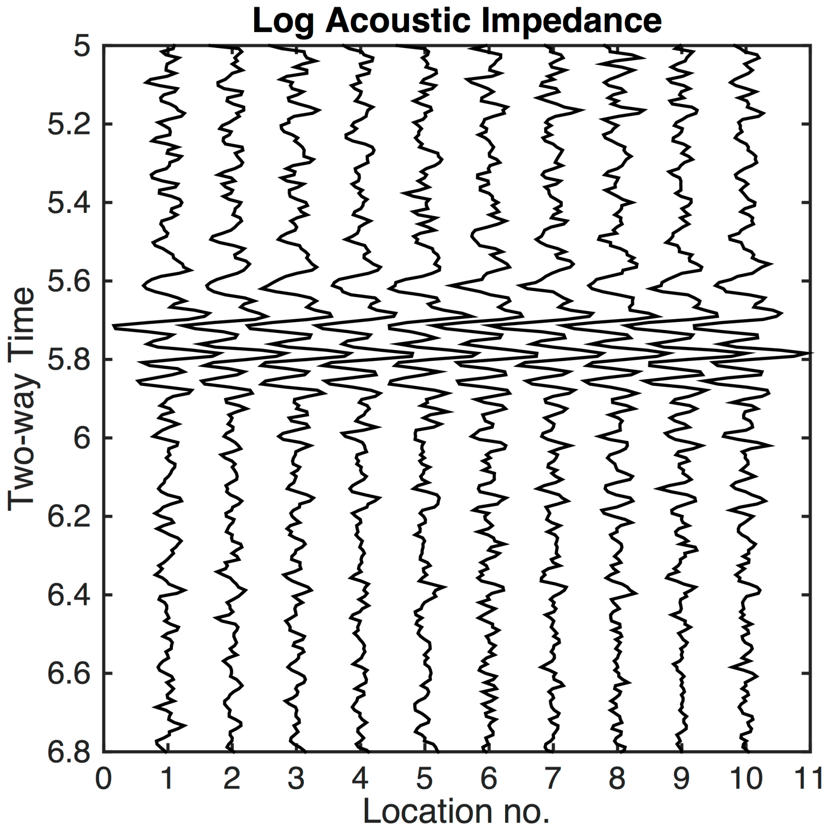
- $\boldsymbol{m_i}$ is the ith column in the matrix $\boldsymbol{M} = \{m_{ij}\}$ containing the unknown acoustic impedances at points (i, j) in the subsurface.
- $\boldsymbol{G}$ is given by $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{D}$ where $\boldsymbol{D}$ performs a differentiation of $\boldsymbol{m_i}$ to obtain an approximate reflectivity
- $\boldsymbol{W}$ is a matrix that convolves the reflectivity with the source signal (wavelet)

Example: Inversion of seismic data



Top left: Recorded wavelet. Top right: Histogram of noise values estimated from data. Bottom left: Estimate of the (temporal) covariance function of the noise. Bottom right: Histogram of reflection coefficients derived from field measurements of rock properties.

Example: Inversion of seismic data



A posteriori mean model obtained from a linear, Gaussian inversion of the data shown in Figure 6 (left). The figure shows a plot of $\log(\boldsymbol{I}/\boldsymbol{I_0})$ where $\boldsymbol{I}$ is the acoustic impedance and $\boldsymbol{I_0}$ is a (here arbitrary) reference value of $\boldsymbol{I}$.

Gaussian, mildly non-linear problems

A mildly non-linear inverse problem:

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m})$$

can be solved iteratively by local approximation to a linear inverse problem:

$$\boldsymbol{m}_{k+1} = \boldsymbol{m}_k + \boldsymbol{\epsilon}_{\boldsymbol{k}}\left(\boldsymbol{G}_{\boldsymbol{k}}^{\boldsymbol{T}}\boldsymbol{C}_n^{-1}\boldsymbol{G}_{\boldsymbol{k}} + \boldsymbol{C}_{\boldsymbol{m}}^{-1}\right)^{-1}\left(\boldsymbol{G}_{\boldsymbol{k}}^{\boldsymbol{T}}\boldsymbol{C}_{\boldsymbol{n}}^{-1}(\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m}_k)) - \boldsymbol{C}_{\boldsymbol{m}}^{-1}(\boldsymbol{m}_k - \boldsymbol{m}_0)\right)$$

where

$$\boldsymbol{G}_{\boldsymbol{k}} = \left(\frac{\partial g_i}{\partial m_j}\right)_{\boldsymbol{m}=\boldsymbol{m}_{\boldsymbol{k}}}.$$

In the limit $k \to \infty$ we obtain the local posterior covariance (of the tangent Gaussian centered at $\boldsymbol{m}_\infty$):

$$\boldsymbol{C}_{post} \approx (\boldsymbol{G}_\infty^T \boldsymbol{C}_n^{-\mathbf{1}} \boldsymbol{G}_\infty + \boldsymbol{C}_{\boldsymbol{m}}^{-1})^{-1}.$$

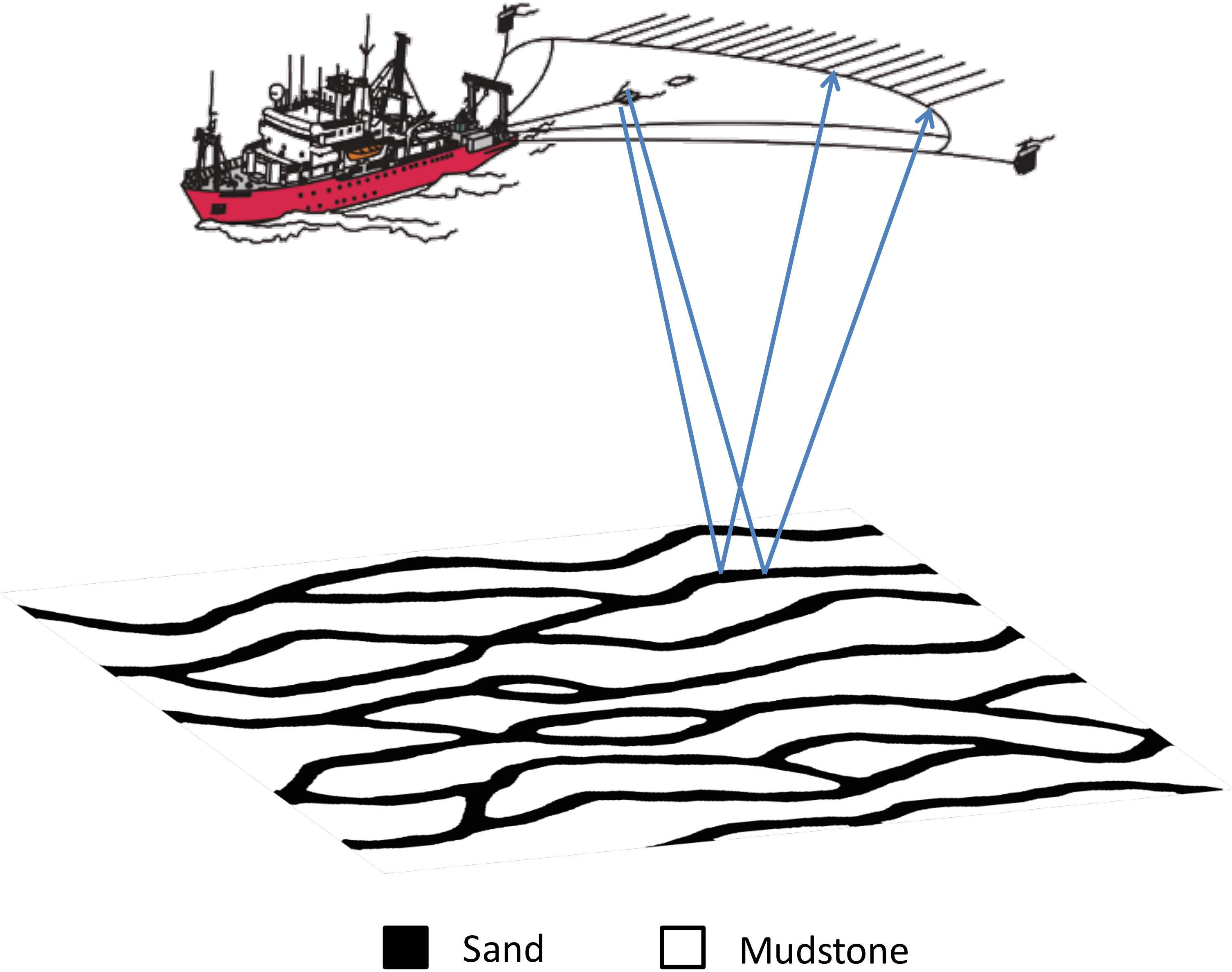
Probabilistic Solutions with Geostatistical Constraints

Example: The Braided River Model

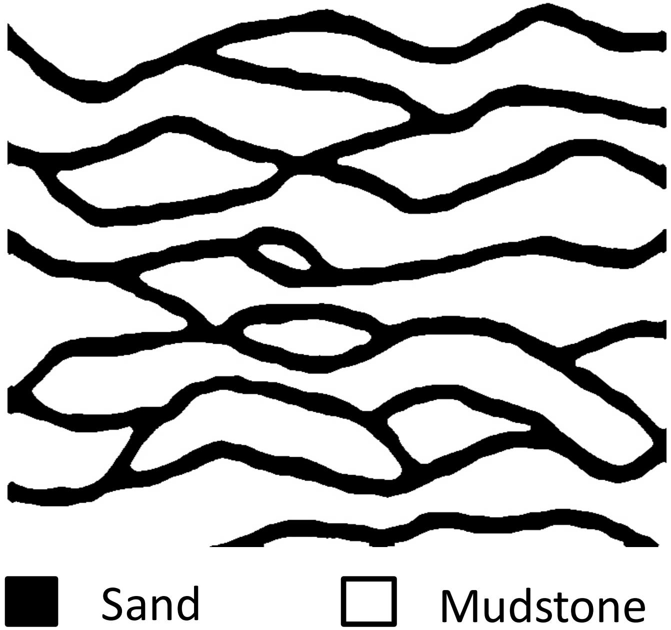
Example: Rakaia-River, New Zealand

Example: Congo-River

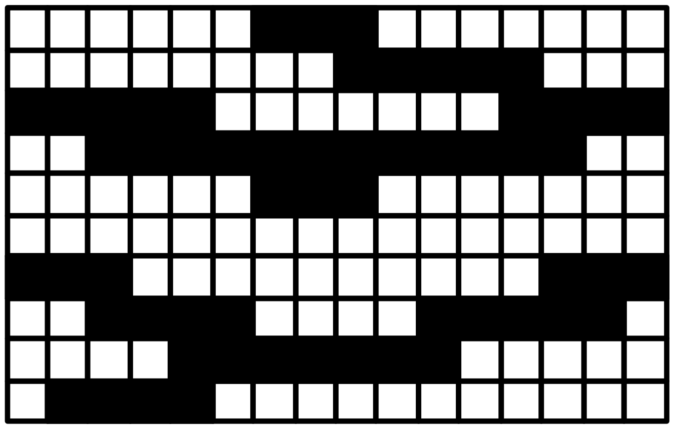
Examples of geo-information: Braided rivers



Examples of geo-information: Braided rivers

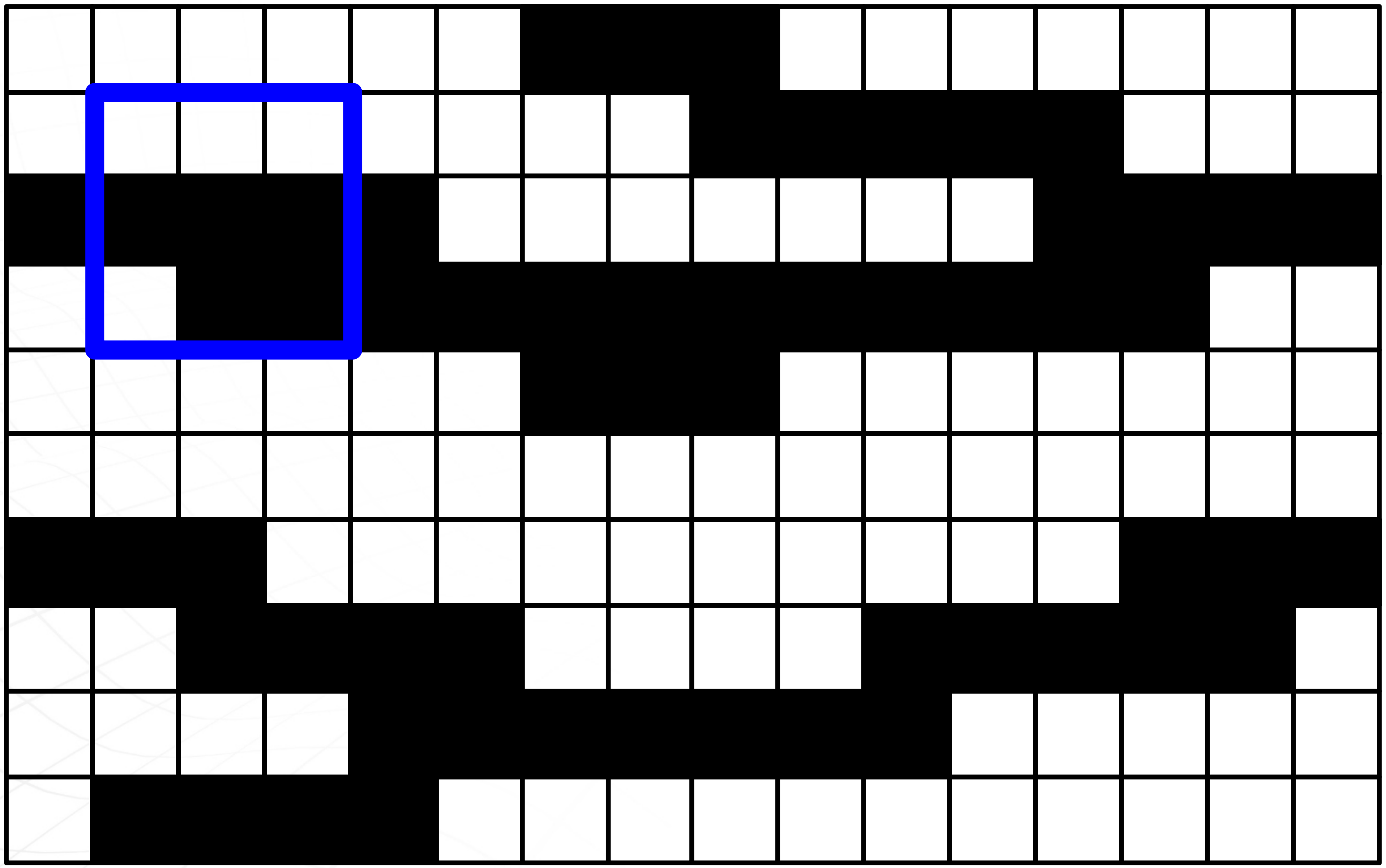


A simple model of a braided river (Strebelle, 2002)



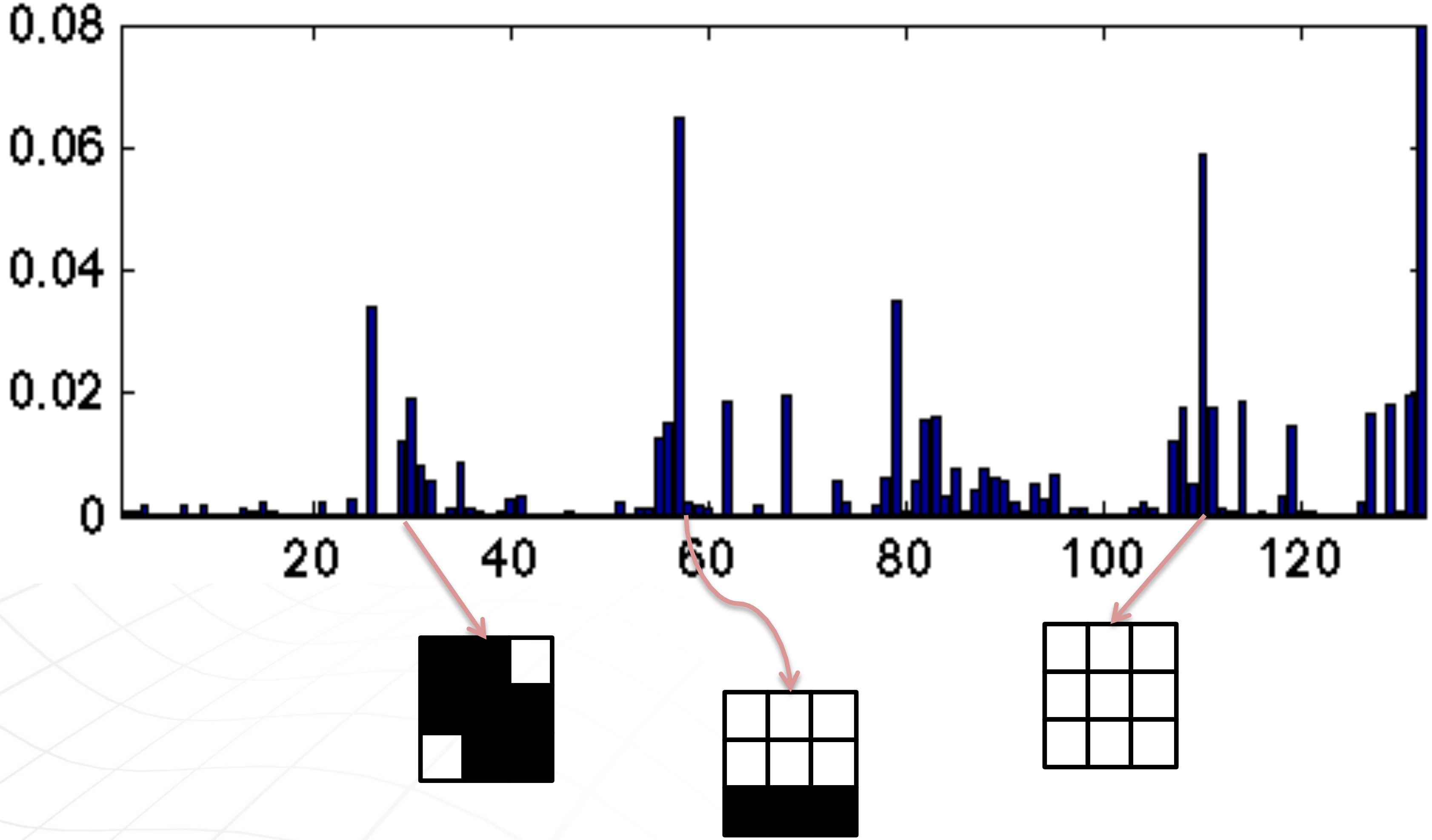
A close-up of part of the pixeled model

Pattern statistics from a geological model

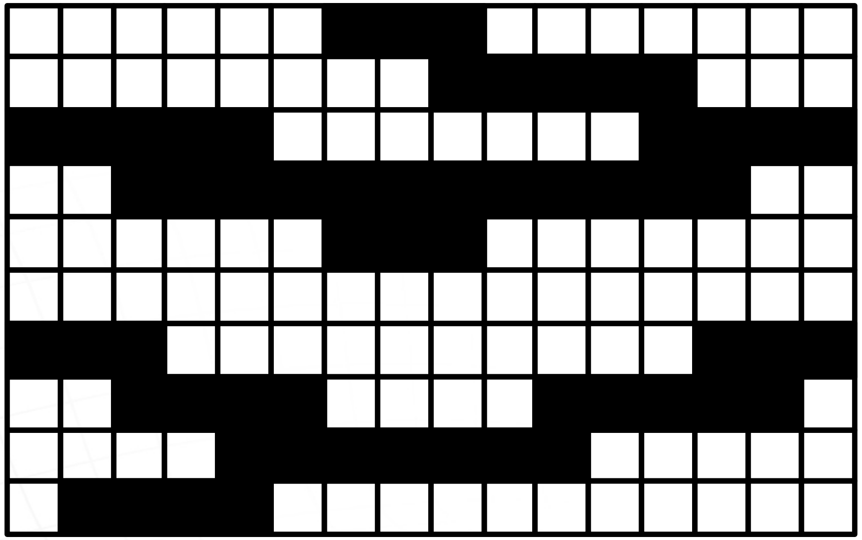


The frequency distribution

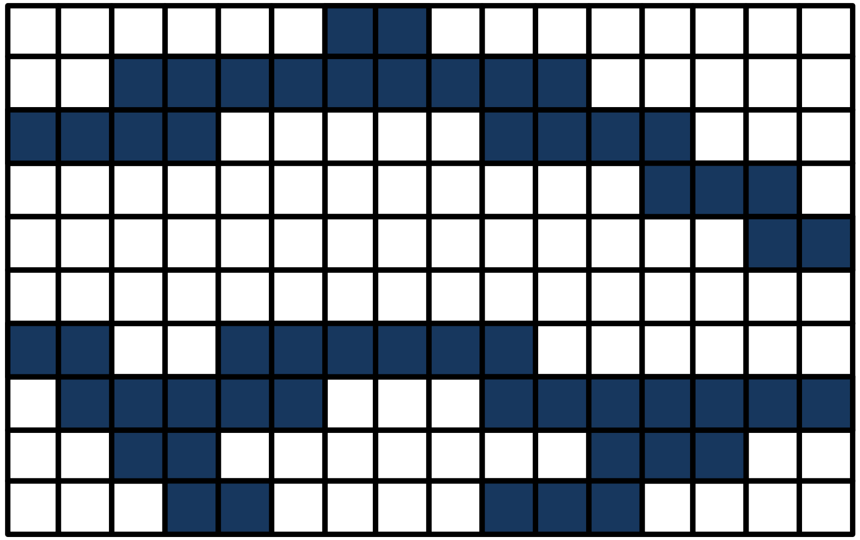
(Normalized. Only non-zero entries)



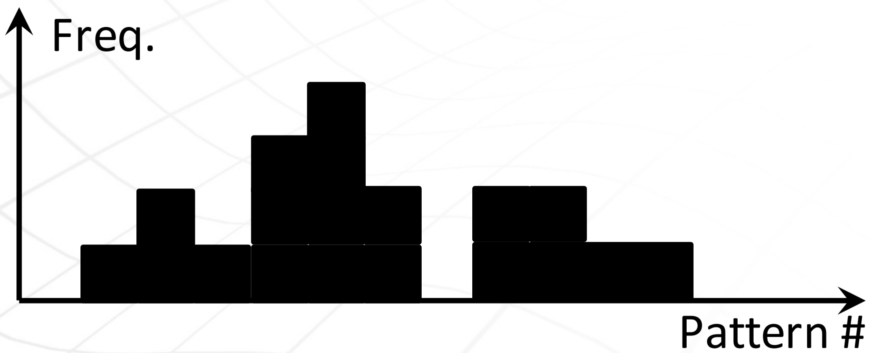
Computing Prior Probability from a Reference Model



Reference Model

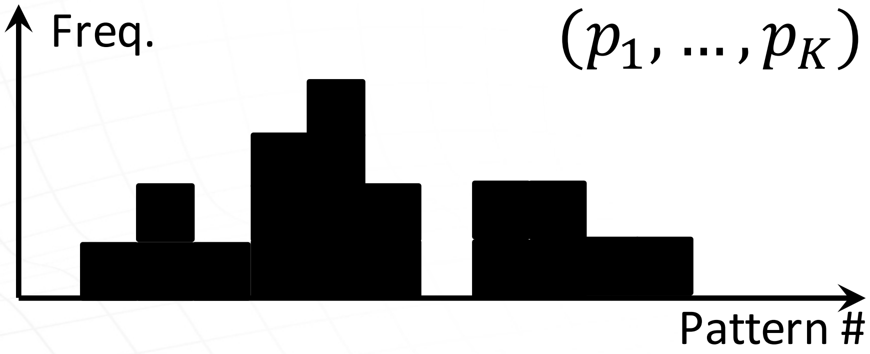


What is the prior probability of this model $\boldsymbol{m}$?

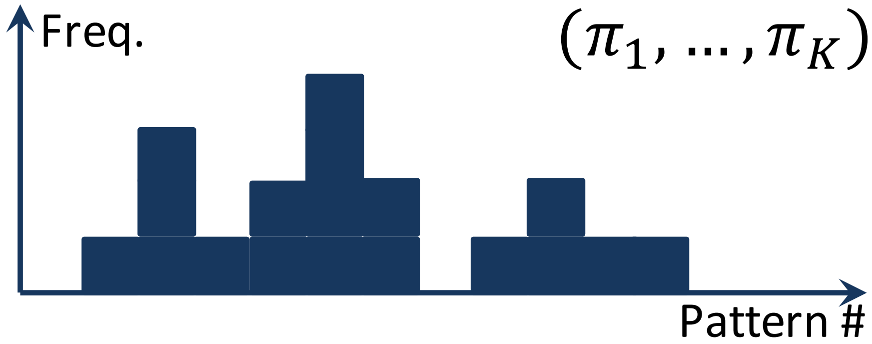


Computing Prior Probability from a Reference Model

Pattern histogram of the reference model



Pattern histogram of the test model $\boldsymbol{m}$



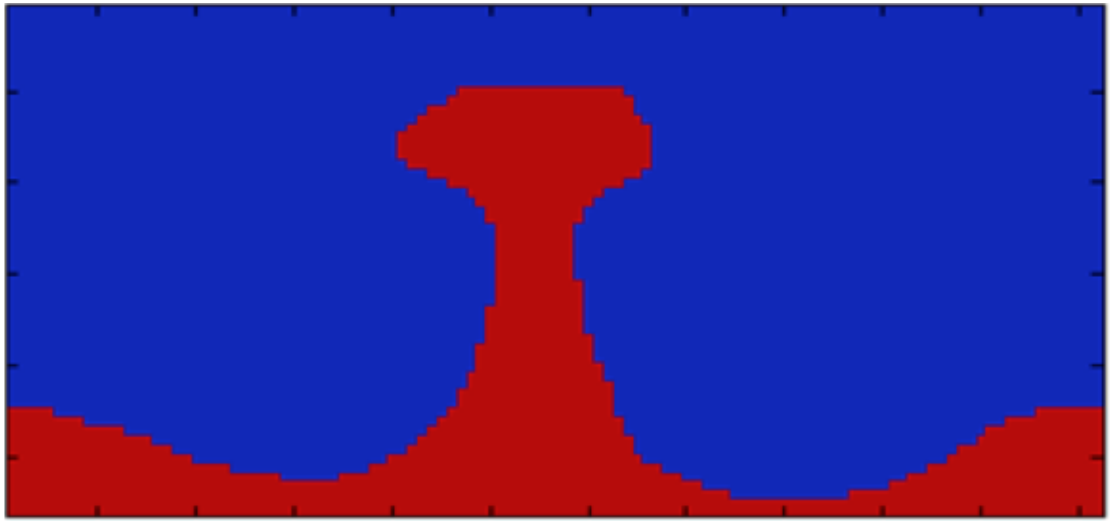
Define the prior probability

$$f(\boldsymbol{m}) \equiv P(\pi_1, \dots, \pi_K) = \frac{N!}{\pi_1! \dots \pi_K!} p_1^{\pi_1} \dots p_K^{\pi_K}$$

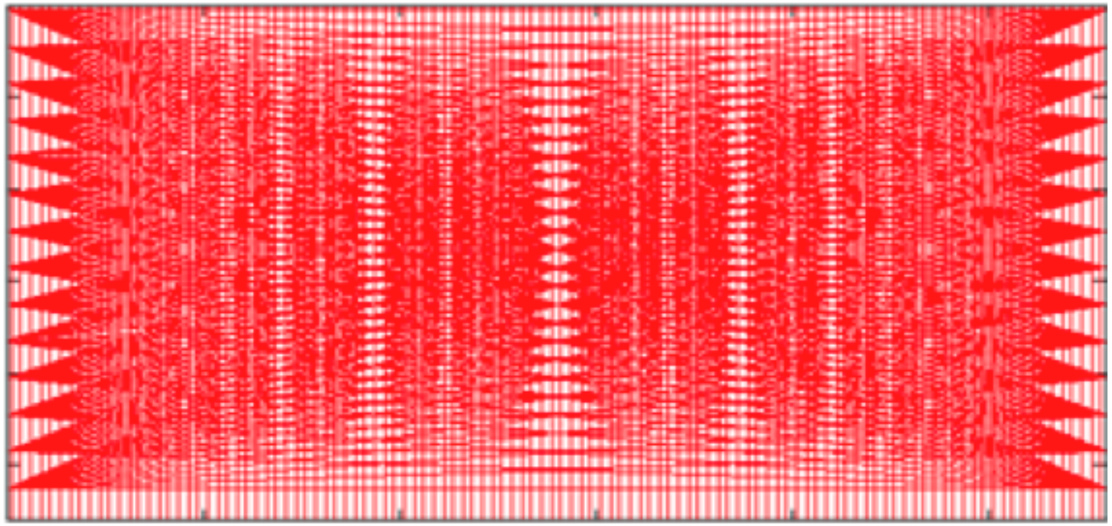
Adding physics to ensure invariant models?

Global Physical Constraints

A salt structure model



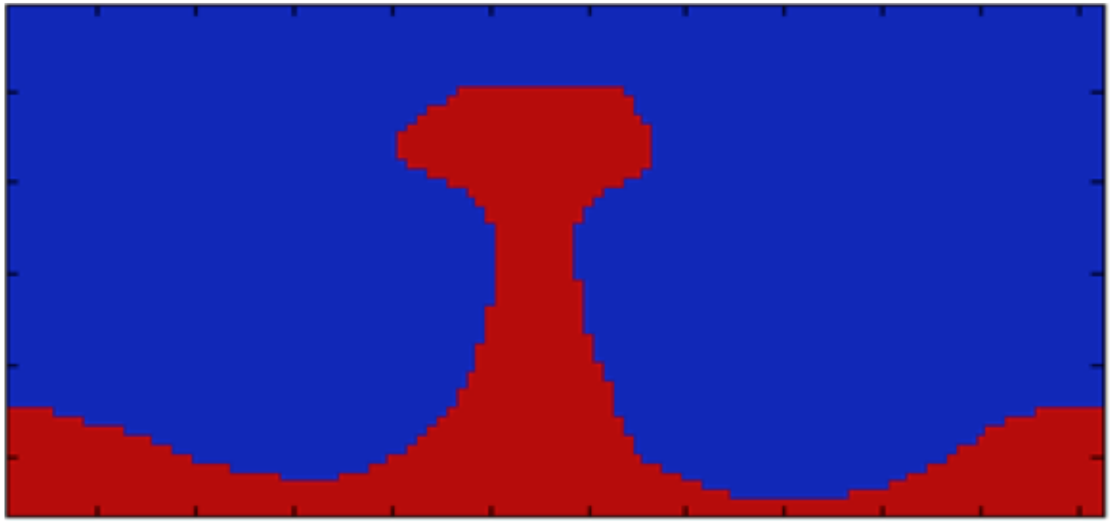
Tomography with vertical and lateral rays



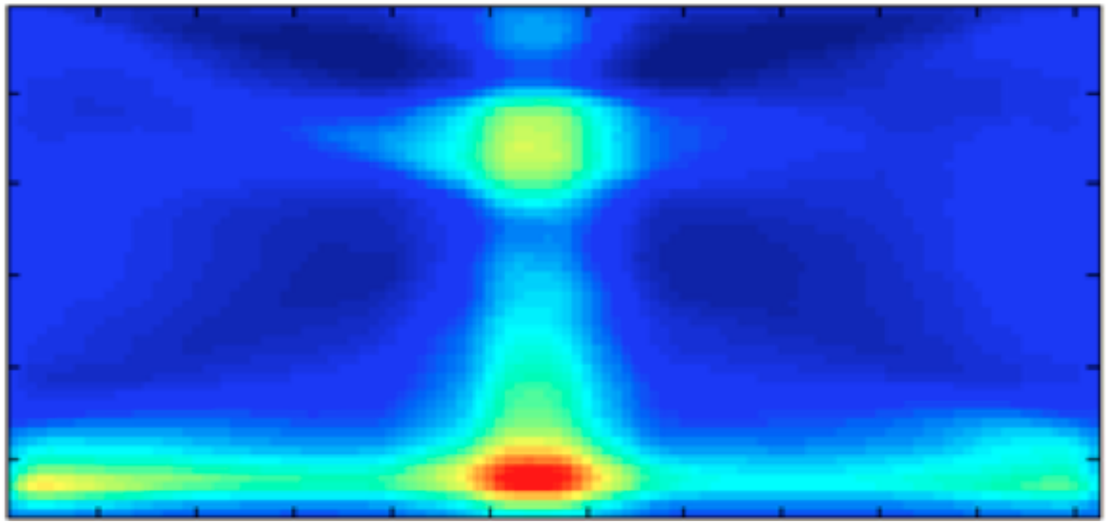
Cordua & Mosegaard, 2014

Global Physical Constraints

A salt structure model



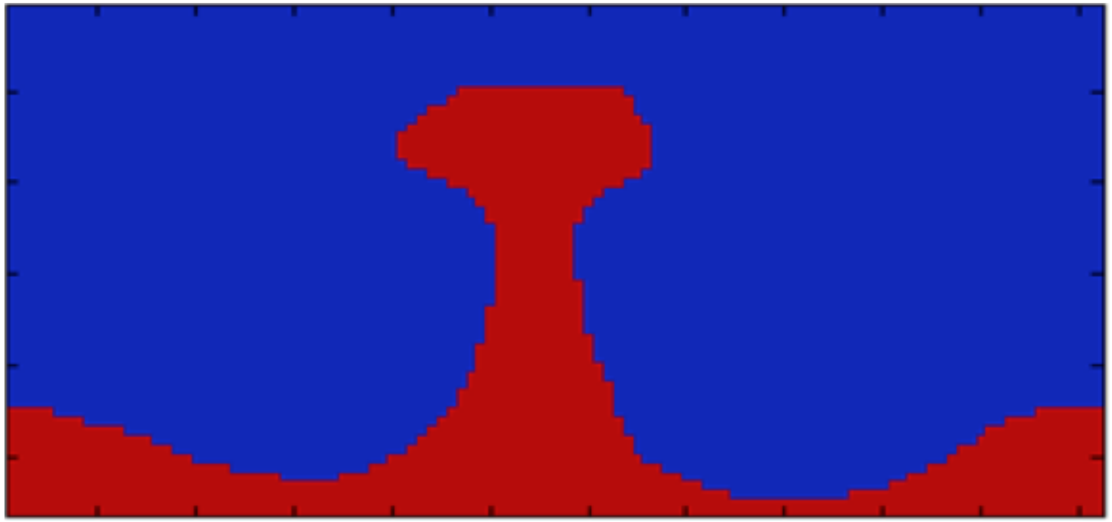
Least Squares inversion with simple Gaussian prior



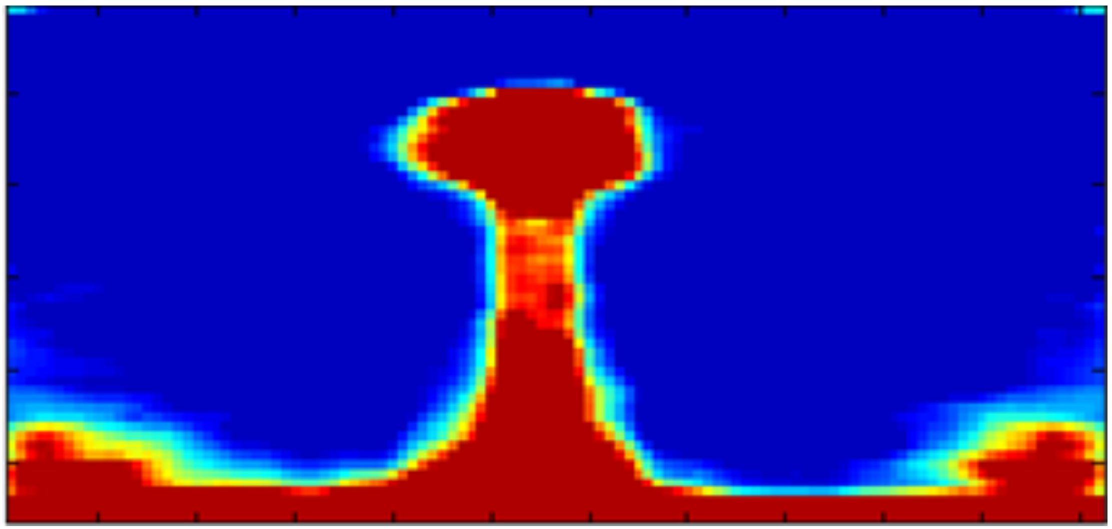
Cordua & Mosegaard, 2014

Global Physical Constraints

A salt structure model



Adding **least-gravitational energy** and **volume preservation** to the prior



Cordua & Mosegaard, 2014