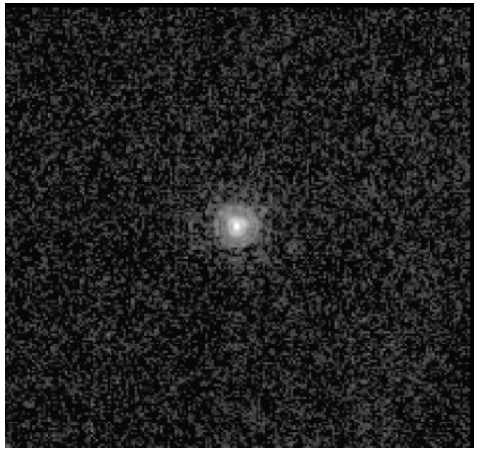
Lecture 1

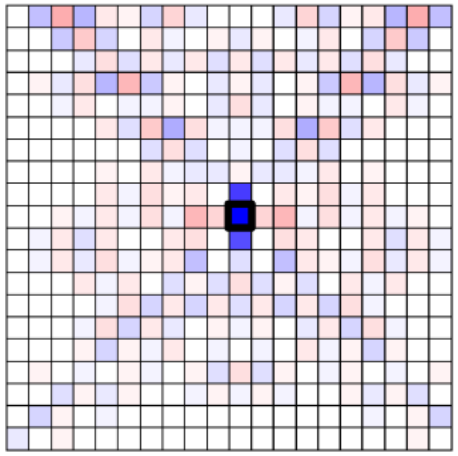
Back to the roots: Least-squares inversion

Andreas Fichtner

ETH Zurich - Seismology and Wave Physics



Optical point-spread function of the Hubble space telescope



Tomographic point-spread function in a 2D traveltime tomography

MOTIVATION

- We often tend to follow fashions. [Right now in the middle of the machine learning hype.]
- We tend to forget the basics. [What we can already do.]

- Lack of truly profound understanding. [Because the foundations are missing.]
- Inability to see broader context and appreciate which new developments are actually useful.
- Tendency to find complicated solutions. [Implement a new technology in order to implement a new technology.]

MOTIVATION

- We often tend to follow fashions. [Right now in the middle of the machine learning hype.]
- We tend to forget the basics. [What we can already do.]

- ➢ Lack of truly profound understanding. [Because the foundations are missing.]
- ➢ Inability to see broader context and appreciate which new developments are actually useful.
- ➢ Tendency to find complicated solutions. [Implement a new technology in order to implement a new technology.]

GOALS OF THIS LECTURE

- Introduce the very basics of linear inverse theory.
- Cover topics that everybody absolutely has to know.
- Foundation for later, more advanced lectures.

OUTLINE

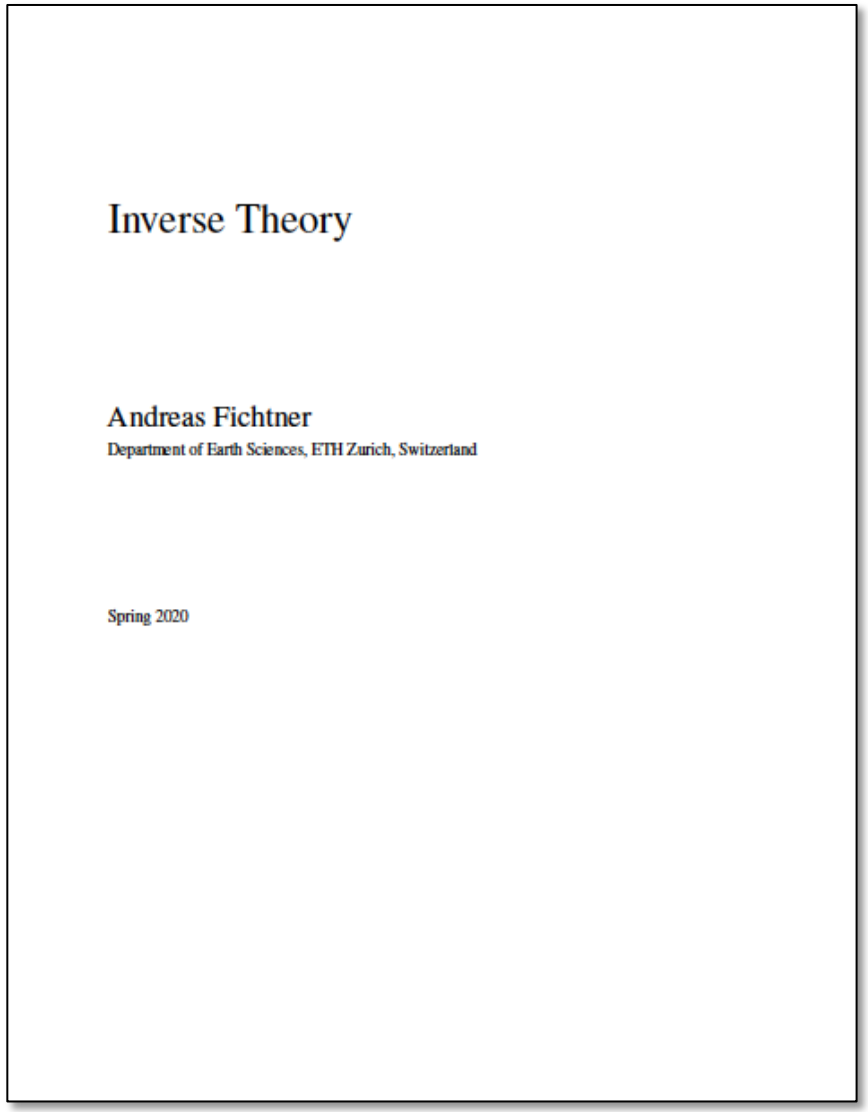
PART I: The basic basics

1. Data, models and physics
2. The nature of inverse problems
3. Fitting lines in *n* dimensions

PART II: Least-squares inversion

1. The least-squares solution
2. Introduction to the model problem: Linear traveltime tomography
3. Prior knowledge vs. regularisation
4. Resolution, point-spread functions and averaging kernels
5. The nullspace

PART III: Introduction to the Jupyter notebook


Inverse Theory

Andreas Fichtner

Department of Earth Sciences, ETH Zurich, Switzerland

Spring 2020

This and much more can be found in the *Inverse Theory* lecture notes.

PART I

The basic basics

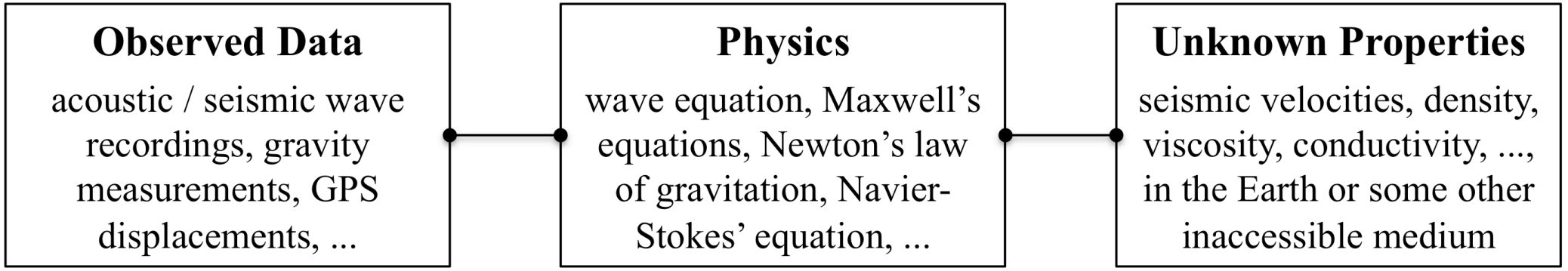
1. Data, models and physics

Observed Data

acoustic / seismic wave recordings, gravity measurements, GPS displacements, ...

Unknown Properties

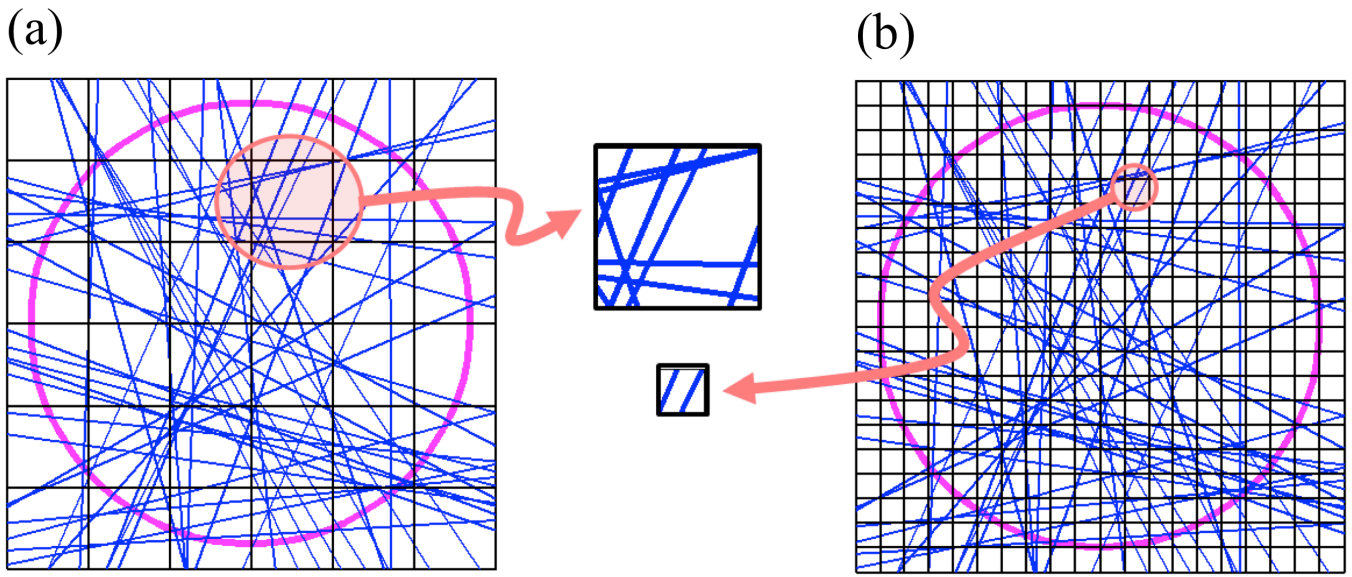
seismic velocities, density, viscosity, conductivity, ..., in the Earth or some other inaccessible medium


Observed Data
acoustic / seismic wave recordings, gravity measurements, GPS displacements, ...
Physics
wave equation, Maxwell's equations, Newton's law of gravitation, Navier-Stokes' equation, ...
Unknown Properties
seismic velocities, density, viscosity, conductivity, ..., in the Earth or some other inaccessible medium

Data And Data Space

- In practice, all data are discrete.
- We collect them into a finite-dimensional vector $\mathbf{d}^{obs}$.
- The ensemble of all conceivable measurements is the data space D.
- Data are a function of control parameters **c**, e.g., measurement location and time, ...
- Data have errors:
  - Random errors [often easy to detect but difficult to describe precisely]
  - Systematic errors [easy to repair but difficult to detect]

- In practice, all models are also discrete.
- We collect them into a finite-dimensional model vector **m**.
- The ensemble of all conceivable models is the model space M.
- Some models are naturally discrete [hypocentral coordinates] others must be discretised [velocity distribution in the Earth]
- Discretisation is largely subjective, and it affects resolution:



good formal resolution
poor spatial resolution

poor formal resolution
good spatial resolution

- In practice, all data are discrete.
- We collect them into a finite-dimensional vector $\mathbf{d}^{obs}$.
- The ensemble of all conceivable measurements is the data space D.
- Data are a function of control parameters **c**, e.g., measurement location and time, ...
- Data have errors:
  - Random errors [often easy to detect but difficult to describe precisely]
  - Systematic errors [easy to repair but difficult to detect]

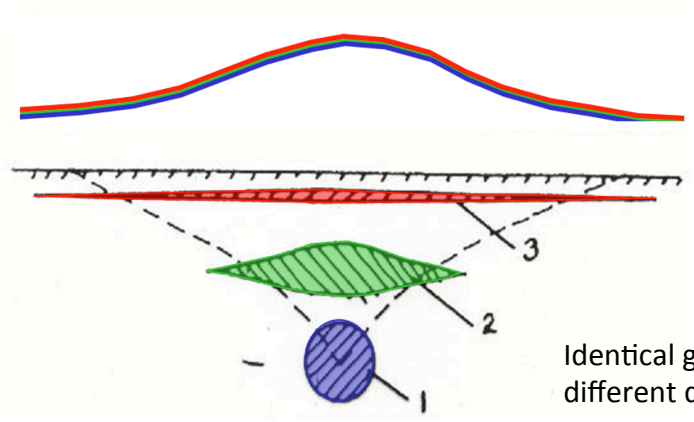
Forward And Inverse Modelling

- Forward modelling: predict synthetic data **d** using a physical theory: $\mathbf{d}=\mathbf{G}(\mathbf{m})$
- Inverse modelling: estimate model on the basis of observed data: $\mathbf{m}^{est}=\mathbf{G}^{-1}(\mathbf{d}^{obs})$
- THE actual inverse problem: $\mathbf{G}^{-1}$ does usually not exist.

2. The nature of inverse problems

Non-Uniqueness

- Non-uniqueness: More than one model **m** describes the observations acceptably well.
- There are different kinds/origins of non-uniqueness:
  1. Physical or inherent non-uniqueness: The nature of physics, nothing we can do.



Identical gravity anomalies caused by very different density anomalies below the surface.

  2. Insufficient data: Lack of coverage in space, time, frequency, ...
  3. Errors in the data.

- Non-uniqueness has many facets, one of them is determinedness.

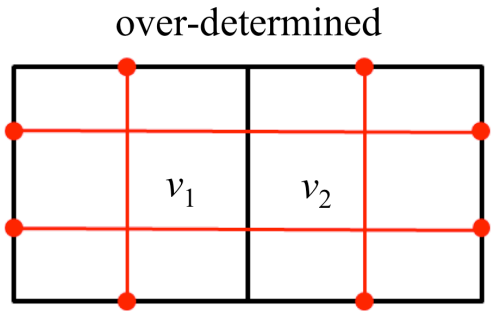


Fig: Illustration of determinedness in a tomographic toy problem.

- Non-uniqueness has many facets, one of them is determinedness.

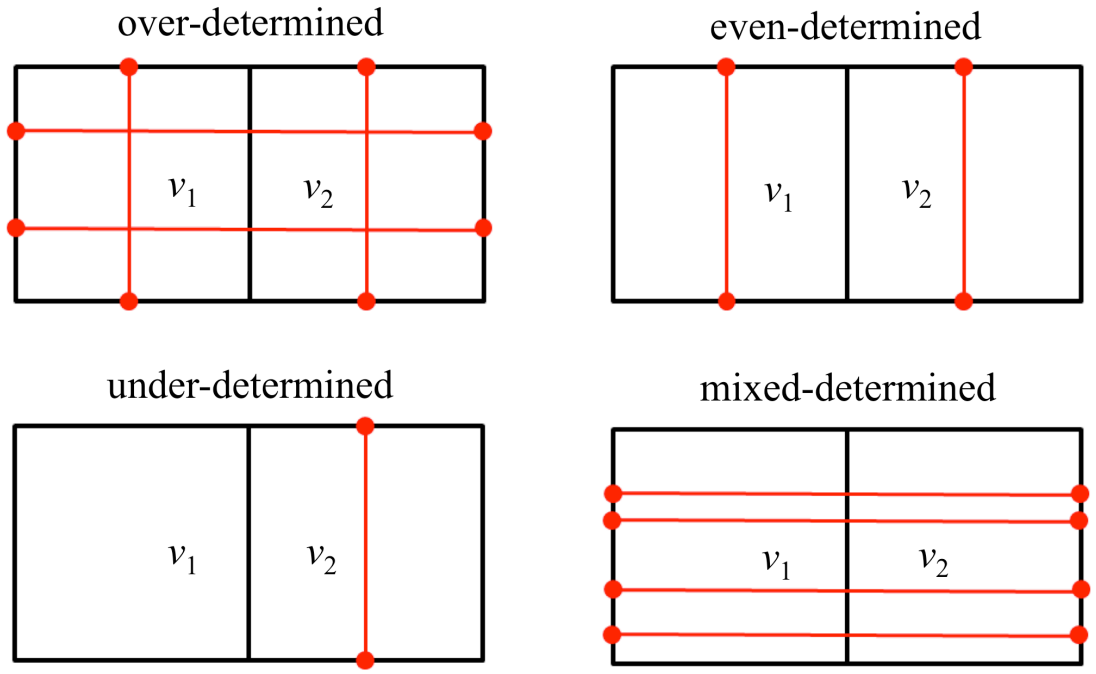


Fig: Illustration of determinedness in a tomographic toy problem.

- Non-uniqueness has many facets, one of them is determinedness.

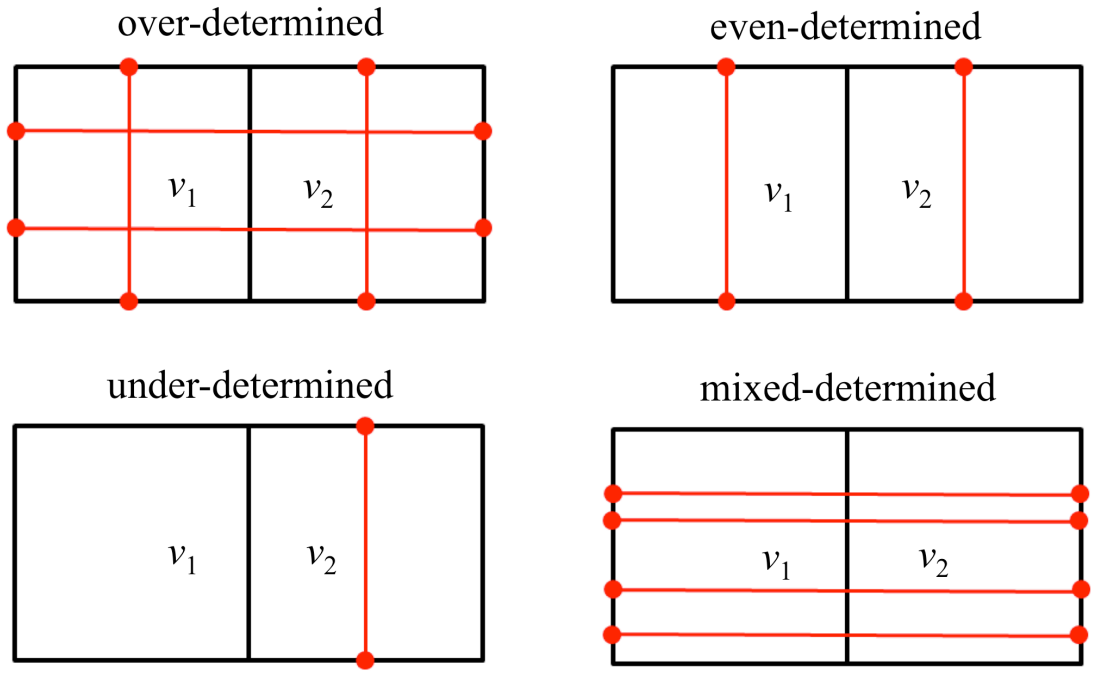


Fig: Illustration of determinedness in a tomographic toy problem.

- Determinedness depends on data but also on model parameterisation!

3. Fitting lines in *n* dimensions

- We consider a collection of observed data points:
- We think we know that the forward modelling equations are linear:

$$d_i = m_1 c_i + m_2$$

- We would like to infer m_1 and m_2

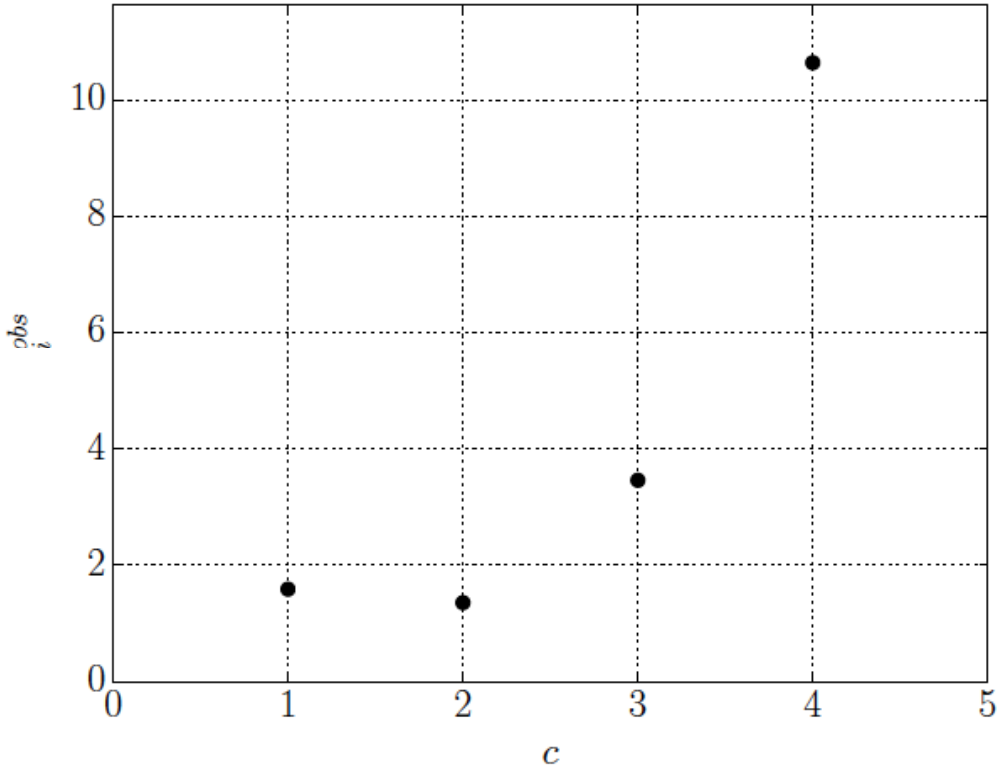


Fig: Observations as function of some control parameters.

- We consider a collection of observed data points:
- We think we know that the forward modelling equations are linear:

$$d_i = m_1 c_i + m_2$$

- We would like to infer m_1 and m_2, but we have some problems:
  1. More measurements than unknowns.
  2. Measurements obviously do not fall onto a line.
- This problem re-occurs frequently also for higher-dimensional (linear) problems **d**=**Gm**.
- This brings us to the concept of:

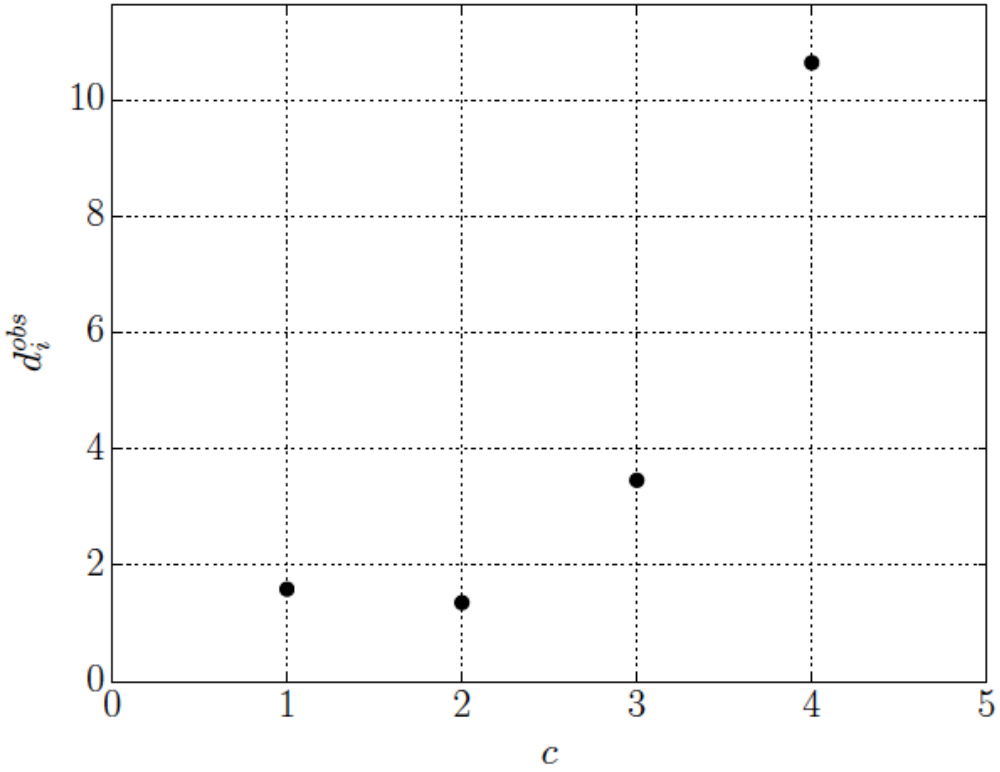


Fig: Observations as function of some control parameters.

- There is obviously no perfect solution to the problem.
- But we may try to find a solution at least such that our observations are matched as closely as possible in a least-squares sense:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)^T \left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)$$

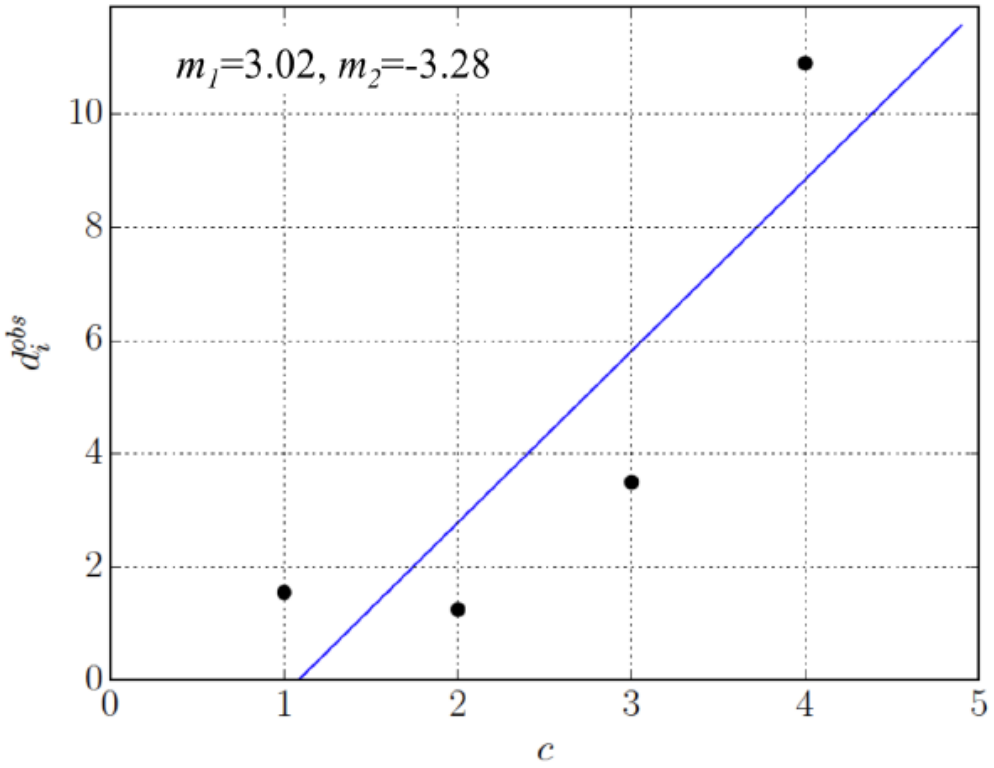


Fig: Simple least-squares fit.

- There is obviously no perfect solution to the problem.
- But we may try to find a solution at least such that our observations are matched as closely as possible in a least-squares sense:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right)^T \left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right)$$

Least-squares misfit functional

- Forcing the first derivative to zero, we can find an optimal model that minimises the misfit:

$$\mathbf{m}^{\text{est}} = \left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{d}^{\text{obs}}$$

- The matrix $(\mathbf{G}^T\mathbf{G})^{-1}$ is called the Moore-Penrose inverse. It may actually not exist.

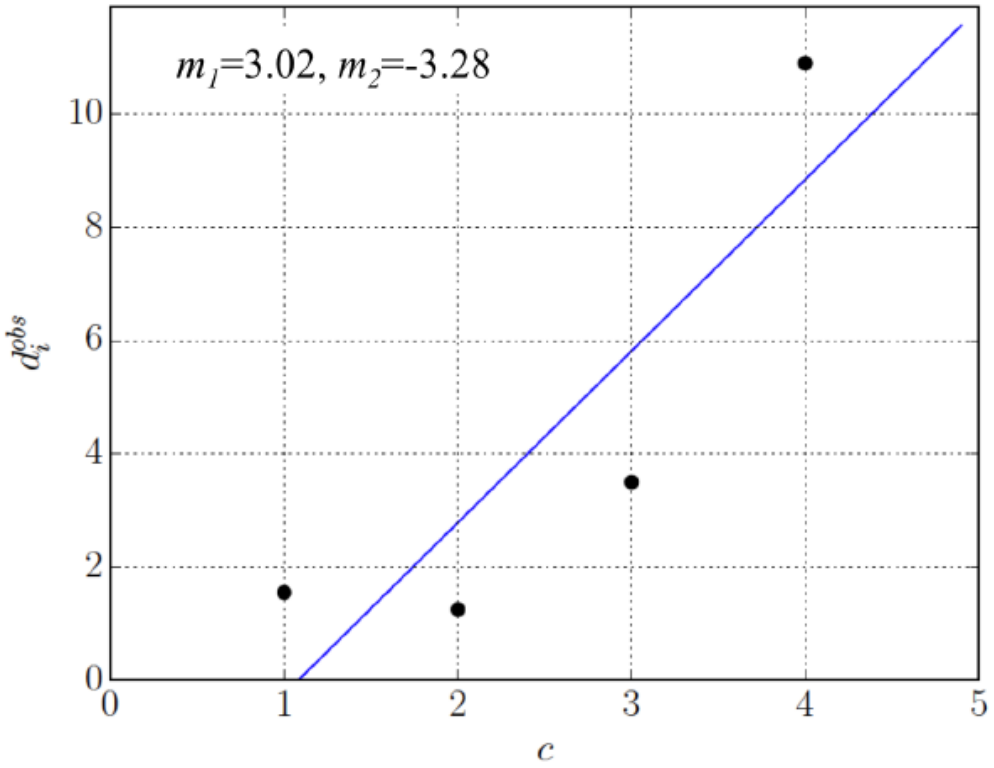


Fig: Simple least-squares fit.

- So far we ignored that data may have errors. [The observation at c=4 may be an outlier, receiving too much weight.]
- Data should be weighted according to their errors:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)^T \mathbf{C}_D^{-1}\left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)$$

- The matrix $\mathbf{C}_D$ is the data covariance matrix:

$$\mathbf{C}_D = \begin{bmatrix} \varepsilon_1^2 & 0 & \dots & 0 & 0 \\ 0 & \varepsilon_2^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \varepsilon_{N-1}^2 & 0 \\ 0 & 0 & \dots & 0 & \varepsilon_N^2 \end{bmatrix}$$

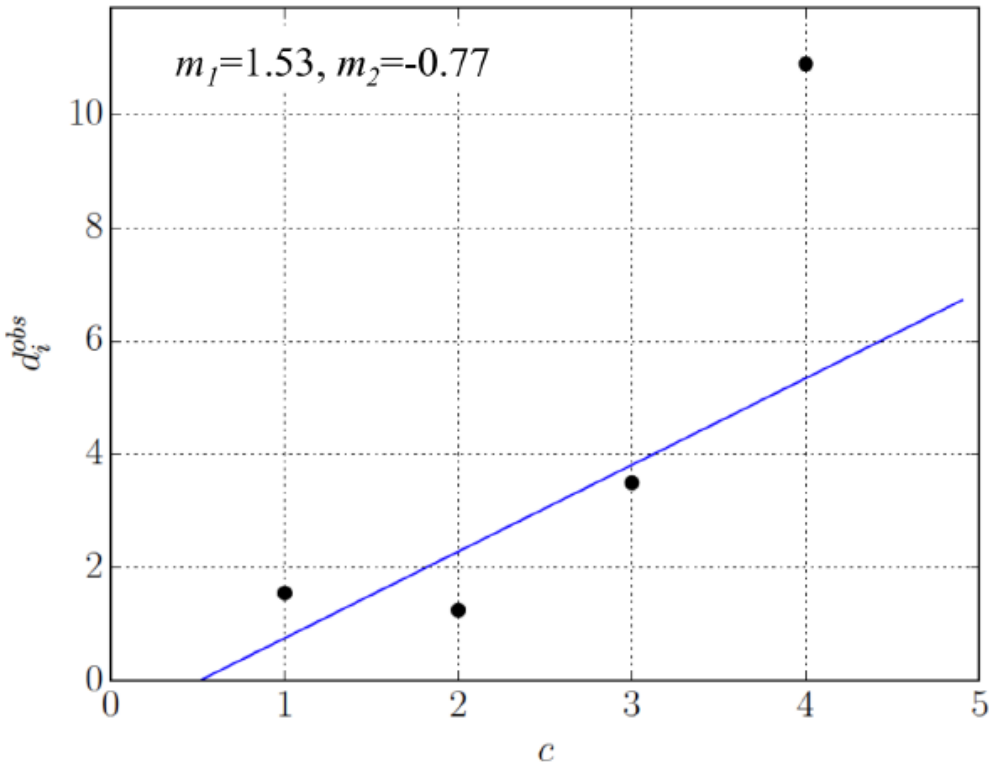


Fig: Least-squares fit with measurement at c=4 receiving less weight.

- So far we ignored that data may have errors. [The observation at c=4 may be an outlier, receiving too much weight.]
- Data should be weighted according to their errors:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)^T \mathbf{C}_D^{-1}\left(\mathbf{d}^{\text{obs}} - \mathbf{Gm}\right)$$

- The matrix $\mathbf{C}_D$ is the data covariance matrix:

$$\mathbf{C}_D = \begin{bmatrix} \varepsilon_1^2 & 0 & \dots & 0 & 0 \\ 0 & \varepsilon_2^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \varepsilon_{N-1}^2 & 0 \\ 0 & 0 & \dots & 0 & \varepsilon_N^2 \end{bmatrix}$$

- Again setting the misfit to zero, we find:

$$\mathbf{m}^{\text{est}} = \left(\mathbf{G}^T \mathbf{C}_D^{-1} \mathbf{G}\right)^{-1} \mathbf{G}^T \mathbf{C}_D^{-1} \mathbf{d}^{\text{obs}}$$

- This still works when $\boldsymbol{C}_D$ is not diagonal.

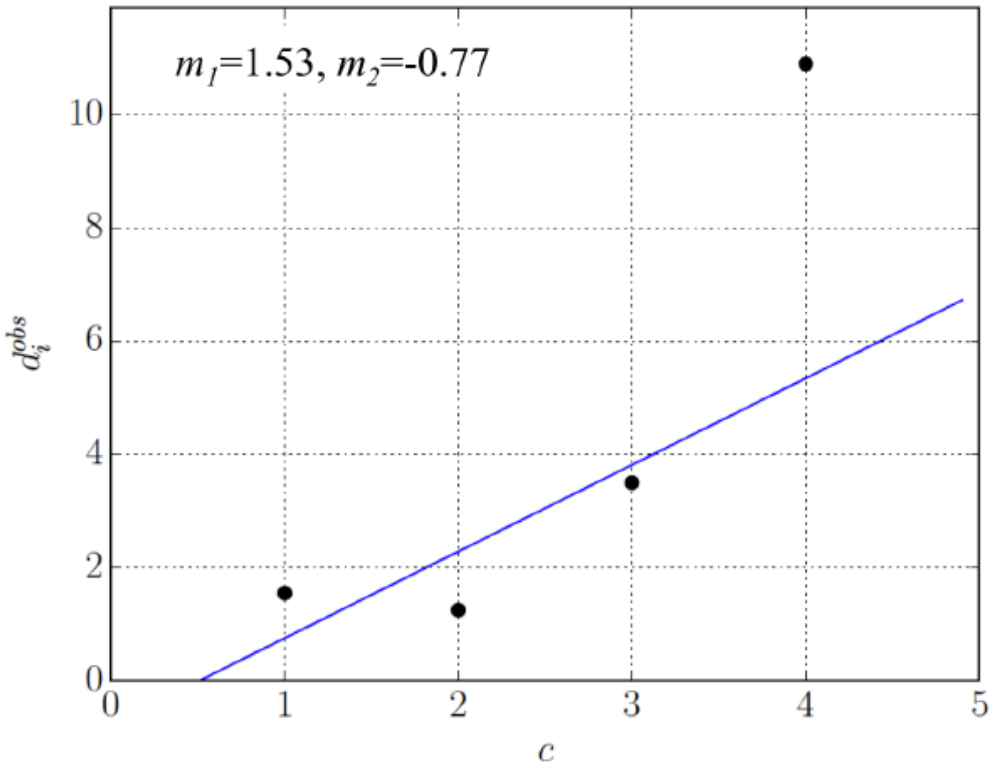


Fig: Least-squares fit with measurement at c=4 receiving less weight.

- We may have some prior information on what meaningful model parameters should be.
- This prior information may be captured in the definition of the least-squares misfit:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right)^T \mathbf{C}_D^{-1}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right) + \frac{1}{2}(\mathbf{m} - \mathbf{m}^{\text{prior}})^T \mathbf{C}_M^{-1}(\mathbf{m} - \mathbf{m}^{\text{prior}})$$

- The matrix $\mathbf{C}_M$ is the prior model covariance matrix.
- The model $\mathbf{m}^{\text{prior}}$ is the a priori most likely model.

- We may have some prior information on what meaningful model parameters should be.
- This prior information may be captured in the definition of the least-squares misfit:

$$\chi(\mathbf{m}) = \frac{1}{2}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right)^T \mathbf{C}_D^{-1}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\mathbf{m}\right) + \frac{1}{2}(\mathbf{m} - \mathbf{m}^{\text{prior}})^T \mathbf{C}_M^{-1}(\mathbf{m} - \mathbf{m}^{\text{prior}})$$

- The matrix $\mathbf{C}_M$ is the prior model covariance matrix.
- The model $\mathbf{m}^{\text{prior}}$ is the a priori most likely model.
- Setting the first derivative to zero, we find a new least-squares solution:

$$\mathbf{m}^{\text{est}} = \left(\mathbf{G}^T \mathbf{C}_D^{-1} \mathbf{G} + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{G}^T \mathbf{C}_D^{-1} \mathbf{d}^{\text{obs}} + \mathbf{C}_M^{-1} \mathbf{m}^{\text{prior}}\right)$$

- Using the estimated model, we compute estimated data:

$$\mathbf{d}^{\text{est}} = \mathbf{G}\mathbf{m}^{\text{est}}$$

- This can be used to compute an estimated misfit:

$$\chi^{\text{est}} = \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\varepsilon_i^2}\left(d_i^{\text{obs}} - d_i^{\text{est}}\right)^2$$

- Ideally, $\chi^{\text{est}} \approx N/2$. Otherwise:
  1. $\chi^{\text{est}} > N/2$: Underfitting. Errors over-estimated and/or too simple forward modelling theory.
  2. $\chi^{\text{est}} < N/2$: Overfitting. Errors under-estimated and/or too complex forward modelling theory.

Remaining Issues

- Definition of misfit is totally arbitrary. Can it be better justified?
- Least squares are very impressed by outliers. Not robust. Can this be improved?
- We only found one solution? Are there more? How can we find them?
- What happens if $\mathbf{G}^T\mathbf{G}$ is actually not invertible?
- What happens if our forward problem is not linear?

Most of these questions will be answered during this MESS.

PART II

Least-squares inversion

1. The least-squares solution

- We return to the least-squares misfit functional:

$$\chi(\mathbf{m}) = \frac{1}{2}(\mathbf{m}-\mathbf{m}^{\text{prior}})^T \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}^{\text{prior}}) + \frac{1}{2}(\mathbf{Gm}-\mathbf{d}^{\text{obs}})^T \mathbf{C}_D^{-1}(\mathbf{Gm}-\mathbf{d}^{\text{obs}})$$

- We return to the least-squares misfit functional:

$$\chi(\mathbf{m}) = \frac{1}{2}(\mathbf{m}-\mathbf{m}^{\text{prior}})^T \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}^{\text{prior}}) + \frac{1}{2}(\mathbf{G}\mathbf{m}-\mathbf{d}^{\text{obs}})^T \mathbf{C}_D^{-1}(\mathbf{G}\mathbf{m}-\mathbf{d}^{\text{obs}})$$

- Interestingly, this can be written in a much simpler form:

$$\chi(\mathbf{m}) = \frac{1}{2}(\mathbf{m}-\tilde{\mathbf{m}})^T \tilde{\mathbf{C}}_M^{-1}(\mathbf{m}-\tilde{\mathbf{m}}) + c$$

- The posterior mean model

$$\tilde{\mathbf{m}} = \left(\mathbf{G}^T \mathbf{C}_D^{-1}\mathbf{G} + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{G}^T \mathbf{C}_D^{-1}\mathbf{d}^{\text{obs}} + \mathbf{C}_M^{-1}\mathbf{m}^{\text{prior}}\right)$$

minimises the least-squares misfit and is often [somewhat incorrectly] regarded as *the* least-squares solution.

- We return to the least-squares misfit functional:

$$\chi(\mathbf{m}) = \frac{1}{2}(\mathbf{m}-\mathbf{m}^{\text{prior}})^T \mathbf{C}_M^{-1}(\mathbf{m}-\mathbf{m}^{\text{prior}}) + \frac{1}{2}(\mathbf{G}\mathbf{m}-\mathbf{d}^{\text{obs}})^T \mathbf{C}_D^{-1}(\mathbf{G}\mathbf{m}-\mathbf{d}^{\text{obs}})$$

- Interestingly, this can be written in a much simpler form:

$$\chi(\mathbf{m}) = \frac{1}{2}(\mathbf{m}-\tilde{\mathbf{m}})^T \tilde{\mathbf{C}}_M^{-1}(\mathbf{m}-\tilde{\mathbf{m}}) + c$$

- The posterior mean model

$$\tilde{\mathbf{m}} = \left(\mathbf{G}^T\mathbf{C}_D^{-1}\mathbf{G} + \mathbf{C}_M^{-1}\right)^{-1}\left(\mathbf{G}^T\mathbf{C}_D^{-1}\mathbf{d}^{\text{obs}} + \mathbf{C}_M^{-1}\mathbf{m}^{\text{prior}}\right)$$

minimises the least-squares misfit and is often [somewhat incorrectly] regarded as *the* least-squares solution.

- The posterior model covariance

$$\tilde{\mathbf{C}}_M = \left(\mathbf{G}^T\mathbf{C}_D^{-1}\mathbf{G} + \mathbf{C}_M^{-1}\right)^{-1}$$

describes how well the solution is constrained.

- A large posterior covariance means that the solution is poorly constrained [quite different models lead to only slightly higher misfit] and vice versa.

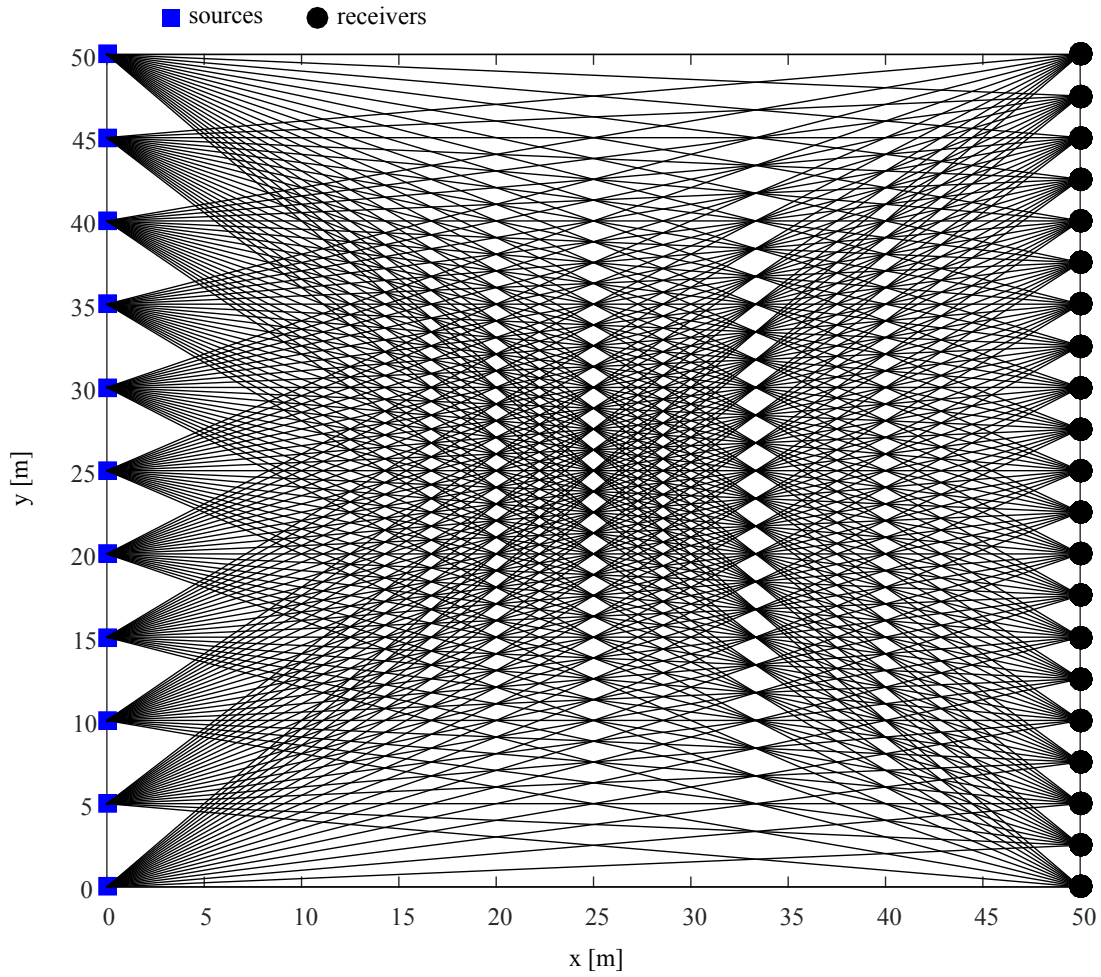
Some Comments

- The solution to the least-squares problem is fully described the posterior mean model and the posterior covariance.
- Both can in principle be computed explicitly. This makes least-squares so attractive.
- All this rests on the assumption that the forward problem is linear.
- One may need to invert a large matrix.
- The matrix may actually not be invertible.

2. Introduction to the model problem: Linear traveltime tomography

Model Problem For Illustration

- To better illustrate the least-squares concept, we consider a simple toy problem:
- 2-D straight-ray traveltime tomography,
  - 11 sources and 21 receivers connected by 231 straight rays,
  - 21 × 21 = 441 blocks of constant slowness used to parameterise the model.



Model Problem For Illustration

- The forward problem is exactly linear: **d**=**Gm**.
  - The model vector **m** contains the slowness values in the blocks.
  - The matrix **G** contains the lengths of the ray segments in each block.
- The problem is under-determined because we have more unknowns (441) than observations (231).
- Though coverage looks good, **G** is extremely sparse, and only around 200 of the 441 eigenvalues of $\mathbf{G}^T\mathbf{G}$ are non-zero.

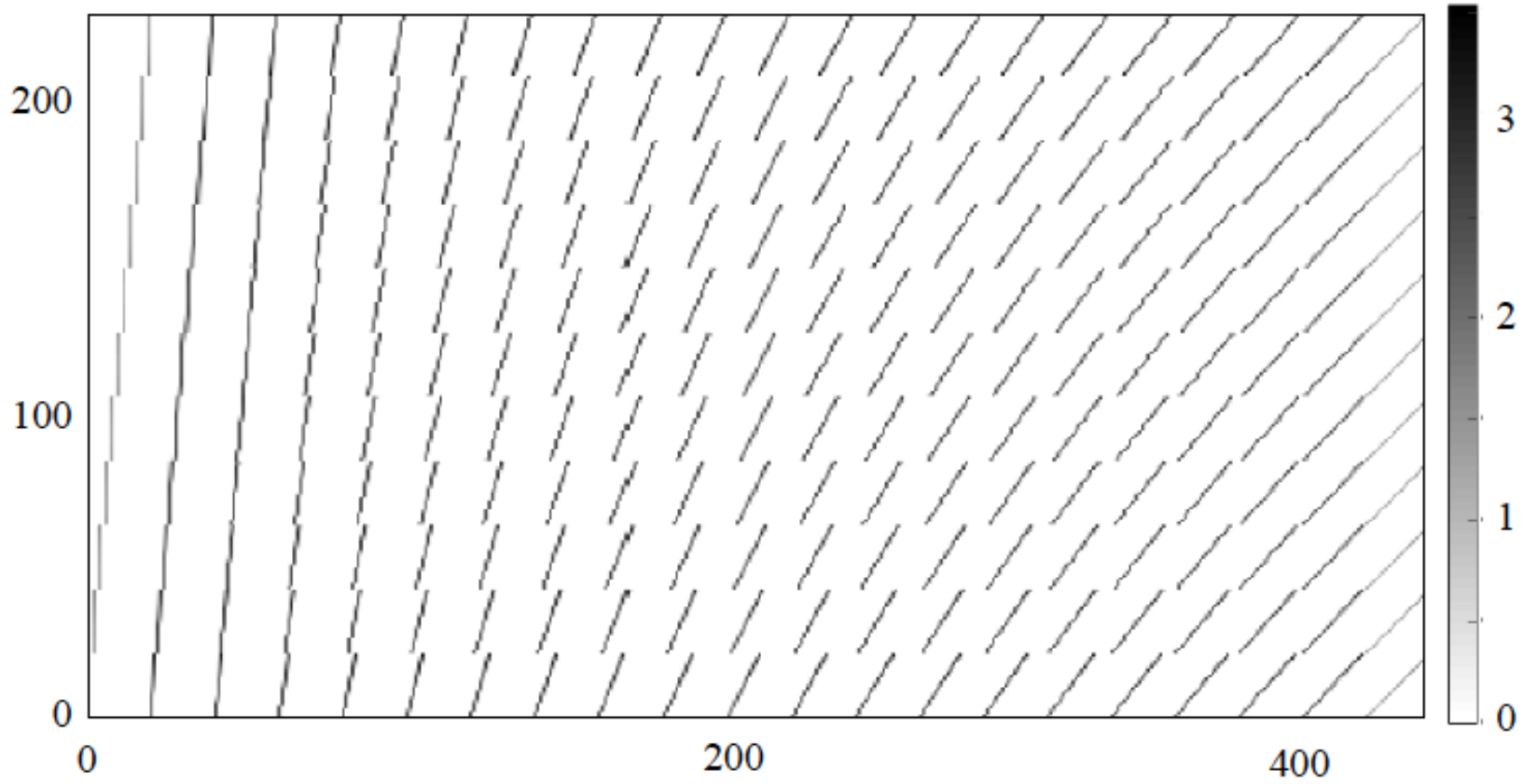


Fig: Visualisation of **G**

- In order to have some ground-truth solution, we commit an inverse crime, i.e., we compute artificial data for some input slowness model:

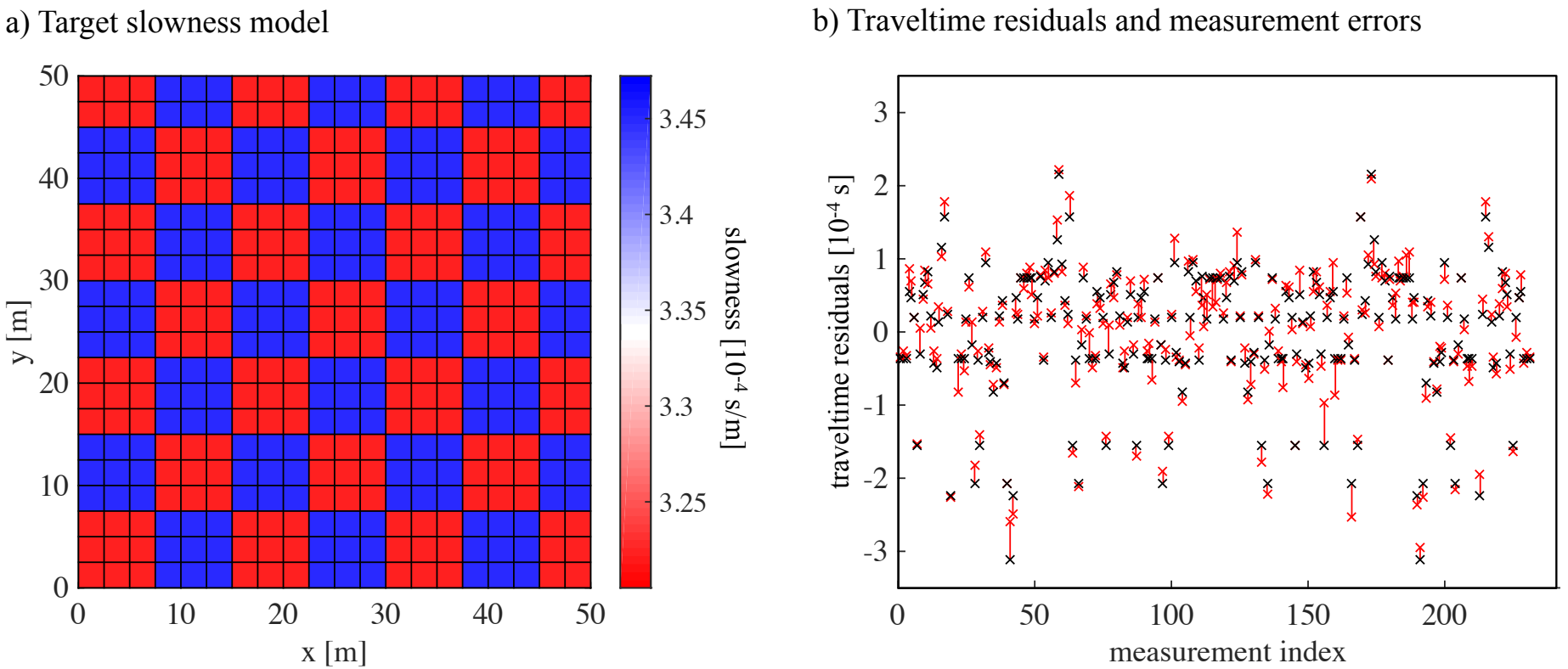


Fig: Summary of the inverse crime setup. (a) Input slowness model. (b) Error-free traveltime residuals with respect to a homogeneous slowness model (black) and traveltime residuals that are artificially polluted by Gaussian errors (red).

3. Prior knowledge vs. regularisation

A Crime Within The Crime

- Obviously, the least-squares problem cannot be solved unless we inject prior knowledge.
- But what prior knowledge do we honestly have? What if this is still not enough?

A Crime Within The Crime

- Obviously, the least-squares problem cannot be solved unless we inject prior knowledge.
- But what prior knowledge do we honestly have? What if this is still not enough?
- Regularisation = injection of artificial prior knowledge with the pragmatic goal to make a matrix invertible [well conditioned].
- It is inherently an act of subjectivity. There is no universally good regularisation.

A Crime Within The Crime

- Obviously, the least-squares problem cannot be solved unless we inject prior knowledge.
- But what prior knowledge do we honestly have? What if this is still not enough?
- Regularisation = injection of artificial prior knowledge with the pragmatic goal to make a matrix invertible [well conditioned].
- It is inherently an act of subjectivity. There is no universally good regularisation.
- In tomography, regularisation is often implemented via a designed prior model covariance:

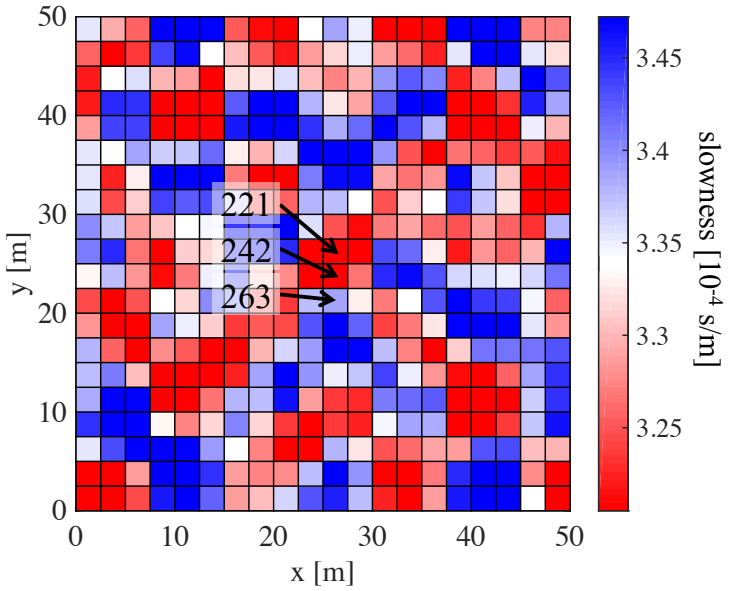
$$(\mathbf{C}_M)_{ij} = \sigma_M^2 \, e^{-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\lambda^2}}$$

- where σ_M describes how well we know individual parameters, and λ describes correlations between neighbouring parameters.
- This causes two different kinds of regularisation:
  1. Damping (via σ_M)
  2. Smoothing (via λ)

A First Solution

- We try σ_M=2e-5 s/m and λ=7.5 m.

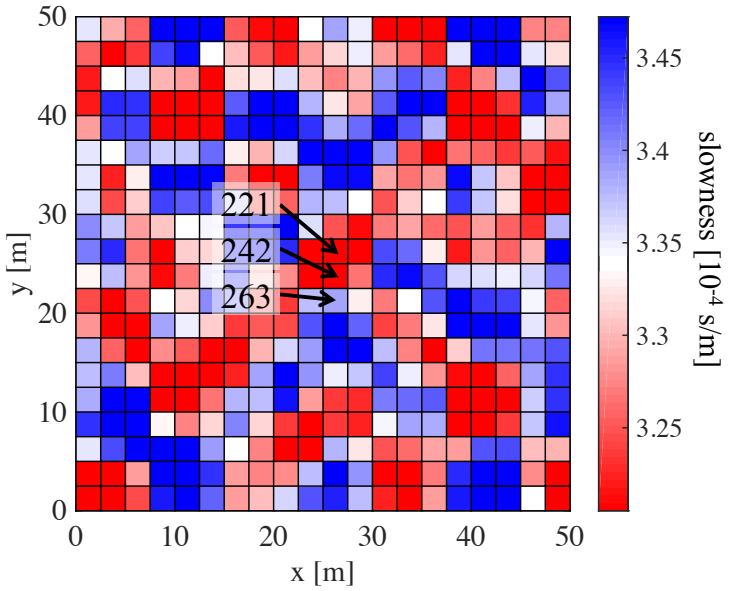
a) Reconstructed slowness distribution



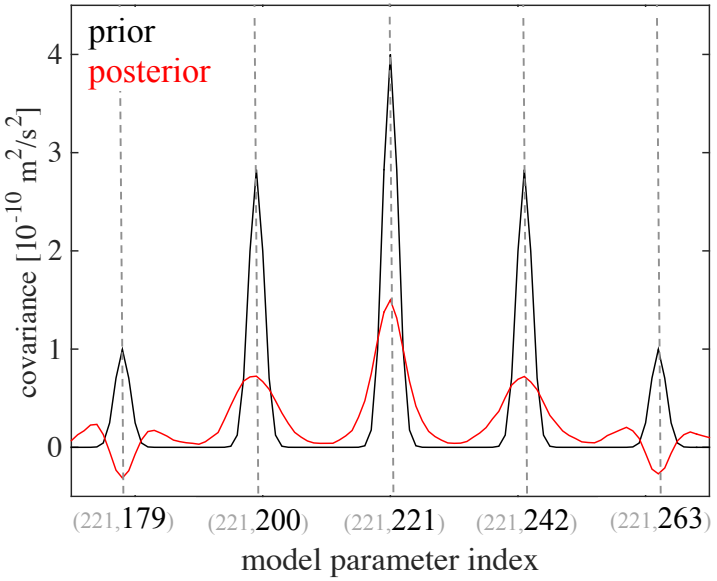
- The input pattern is roughly recovered.

- We try σ_M=2e-5 s/m and λ=7.5 m.

a) Reconstructed slowness distribution



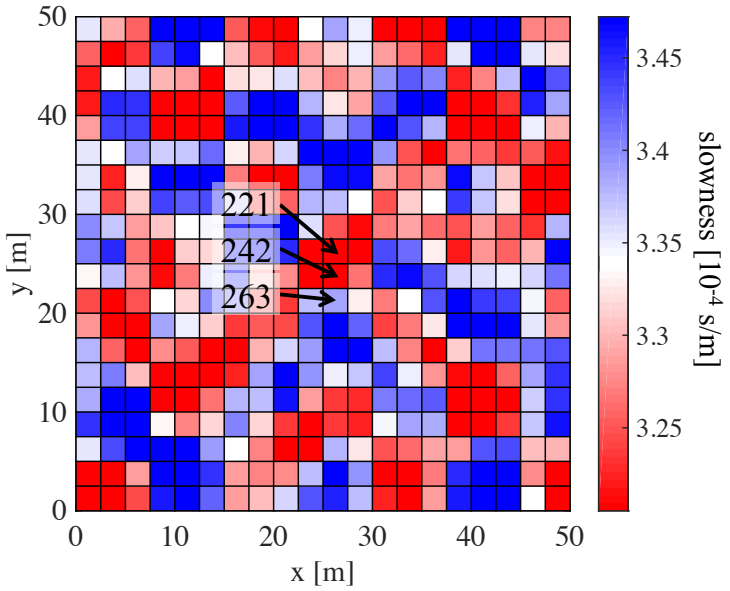
b) Prior and posterior covariance



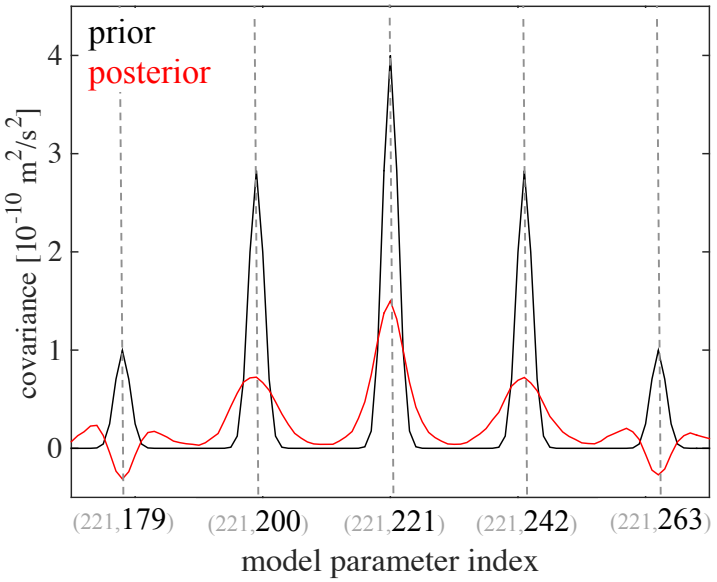
- The input pattern is roughly recovered.
- The posterior covariance is mostly smaller than the prior covariance.

- We try σ_M=2e-5 s/m and λ=7.5 m.

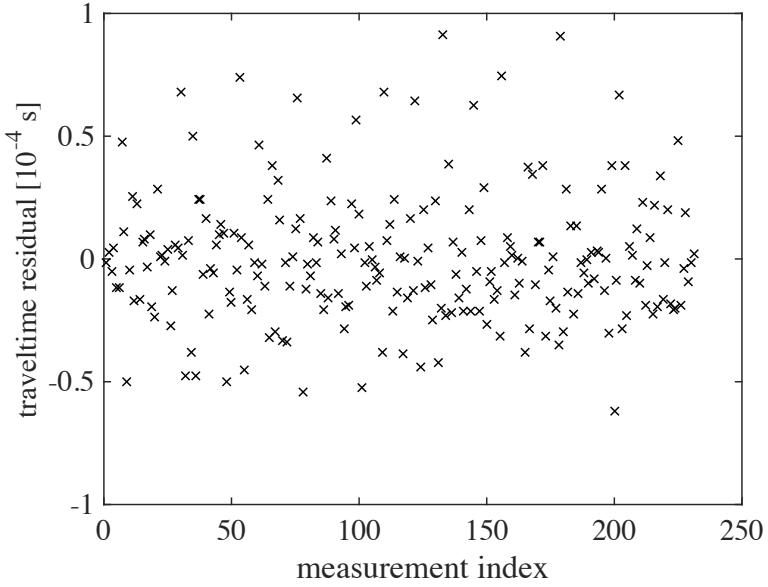
a) Reconstructed slowness distribution



b) Prior and posterior covariance



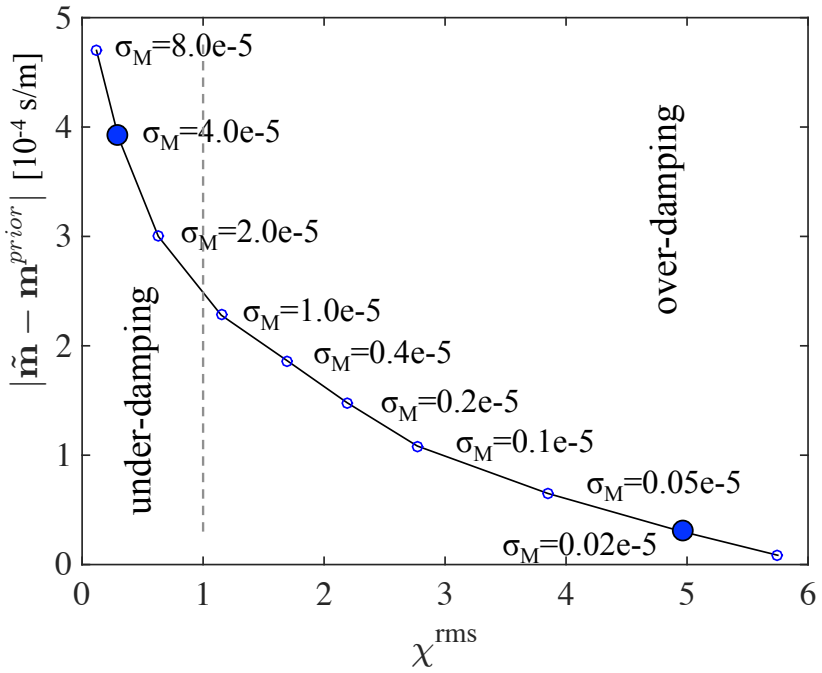
c) Traveltime residuals after reconstruction



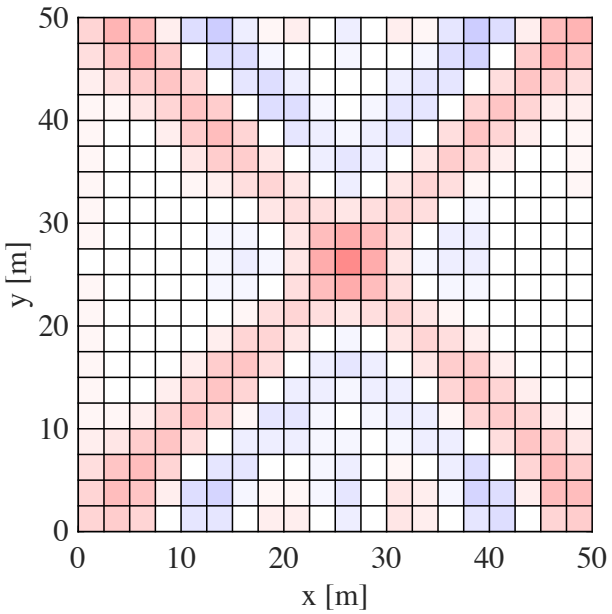
- The input pattern is roughly recovered.
- The posterior covariance is mostly smaller than the prior covariance.
- The traveltime residuals are smaller than for the homogeneous slowness model [mostly below 50 µs compared to 200 µs for the homogeneous model].

- There is no perfect regularisation.
- But we can choose a reasonable regularisation by looking at its influence on data fit and model complexity.

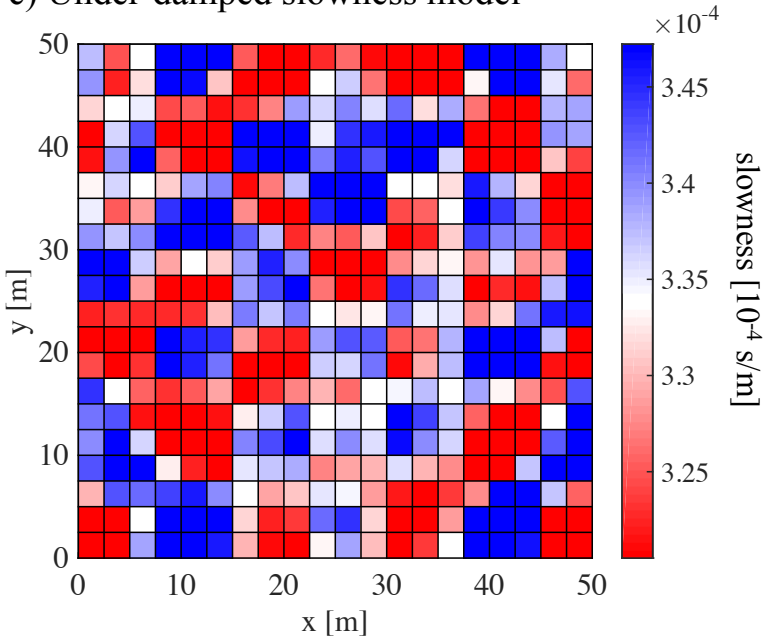
a) L-curve for damping



b) Over-damped slowness model



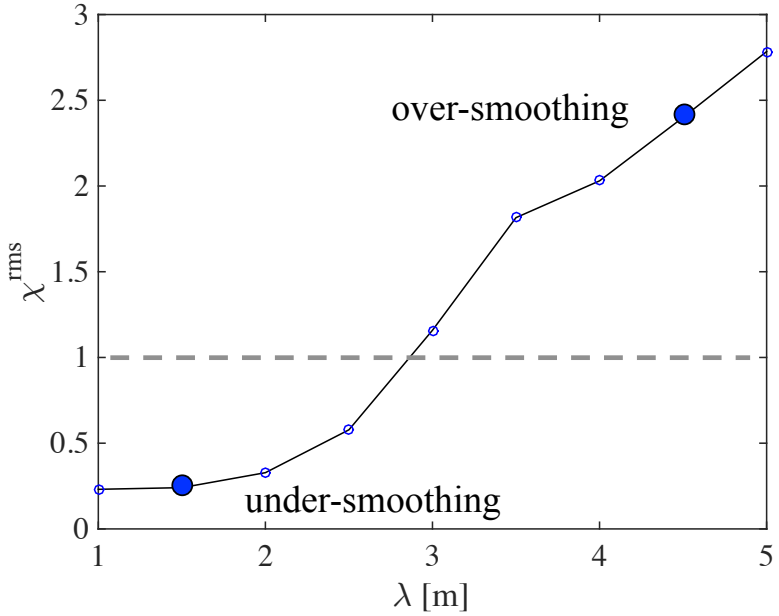
c) Under-damped slowness model



damping [σ_M=0.02•10^{-5} s/m ← σ_M=4.0•10^{-5} s/m]

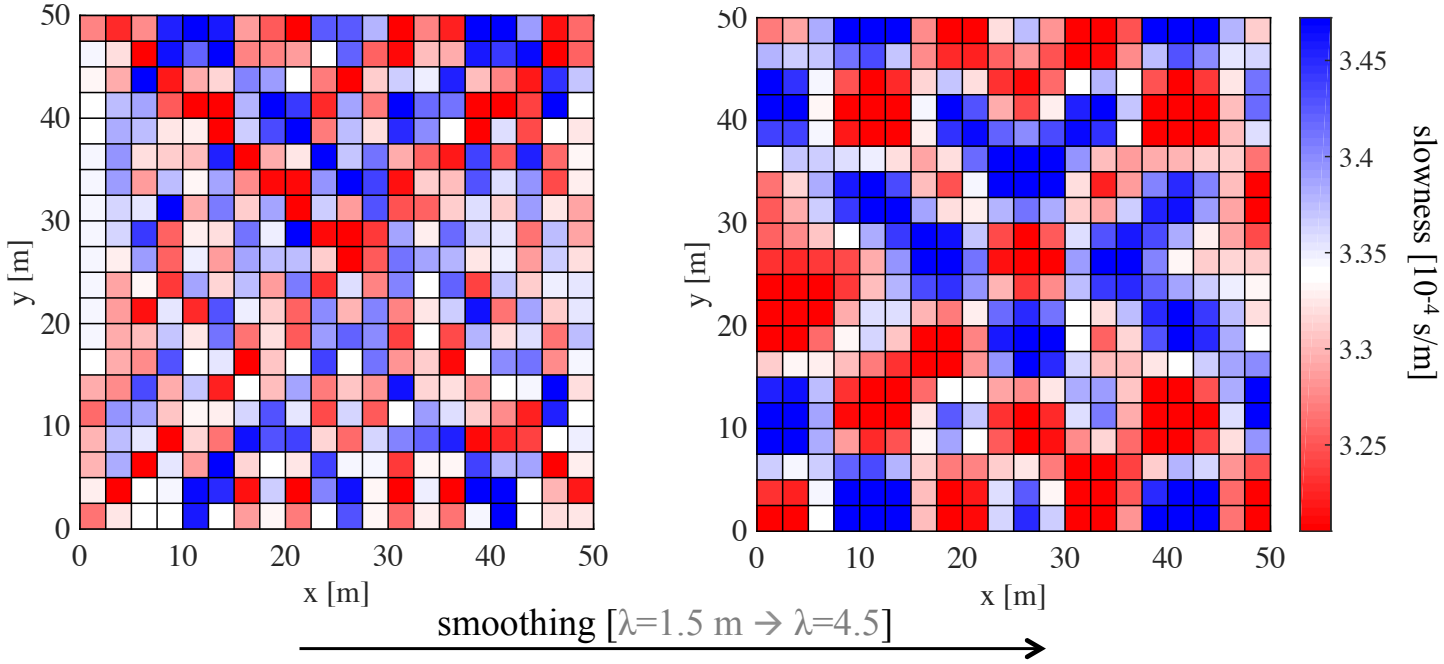
- There is no perfect regularisation.
- But we can choose a reasonable regularisation by looking at its influence on data fit and model complexity.

a) χ^{rms} as a function of smoothing



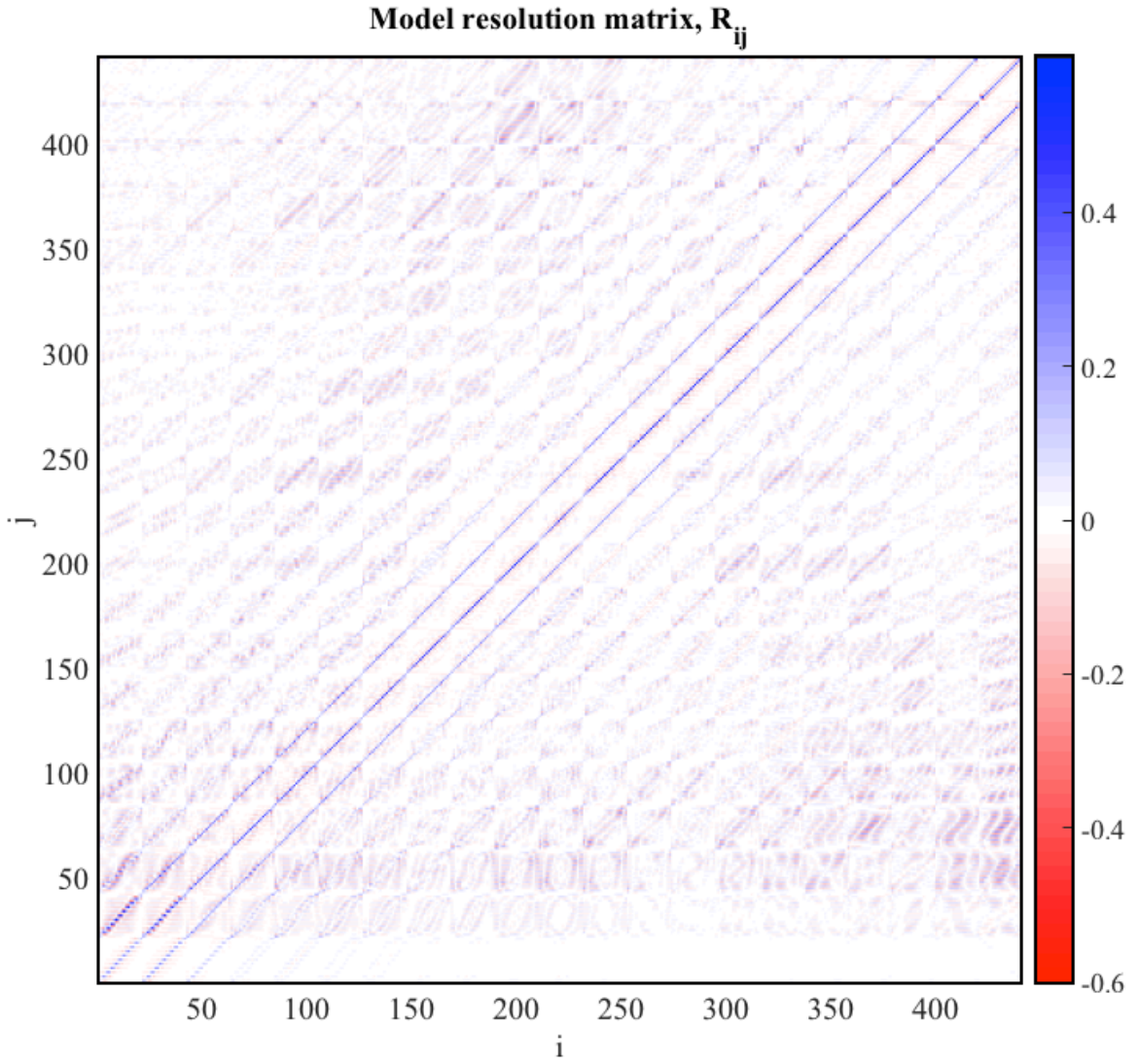
b) Under-smoothed slowness model

c) Over-smoothed slowness model



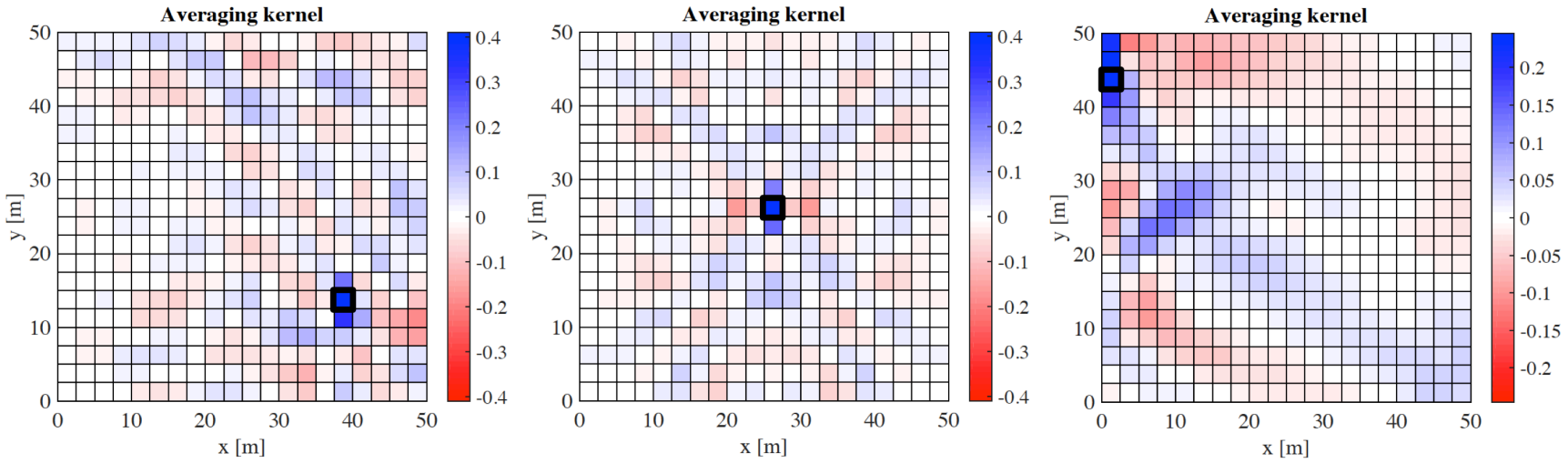
4. Resolution, point-spread functions and averaging kernels

- Within the context of the inverse crime: $\mathbf{d}^{obs}=\mathbf{G}\mathbf{m}^{target}$.
- Using the least-squares concept, we apply some generalised (Moore-Penrose) inverse **L** to $\mathbf{d}^{obs}$ in order to obtain an estimated model $\mathbf{m}^{est}=\mathbf{L}\mathbf{d}^{obs}=\mathbf{L}\mathbf{G}\mathbf{m}^{target}$.
- This gives a linear relation between the 'truth' that we would like to see, $\mathbf{m}^{target}$, and the image $\mathbf{m}^{est}$ that we are able to see with our limited and imperfect data.
- The matrix **R**=**LG** is the model resolution matrix.



Point-Spread Functions And Averaging Kernels

- Rows of the resolution matrix: averaging kernels
- Columns of the resolution matrix: point-spread functions



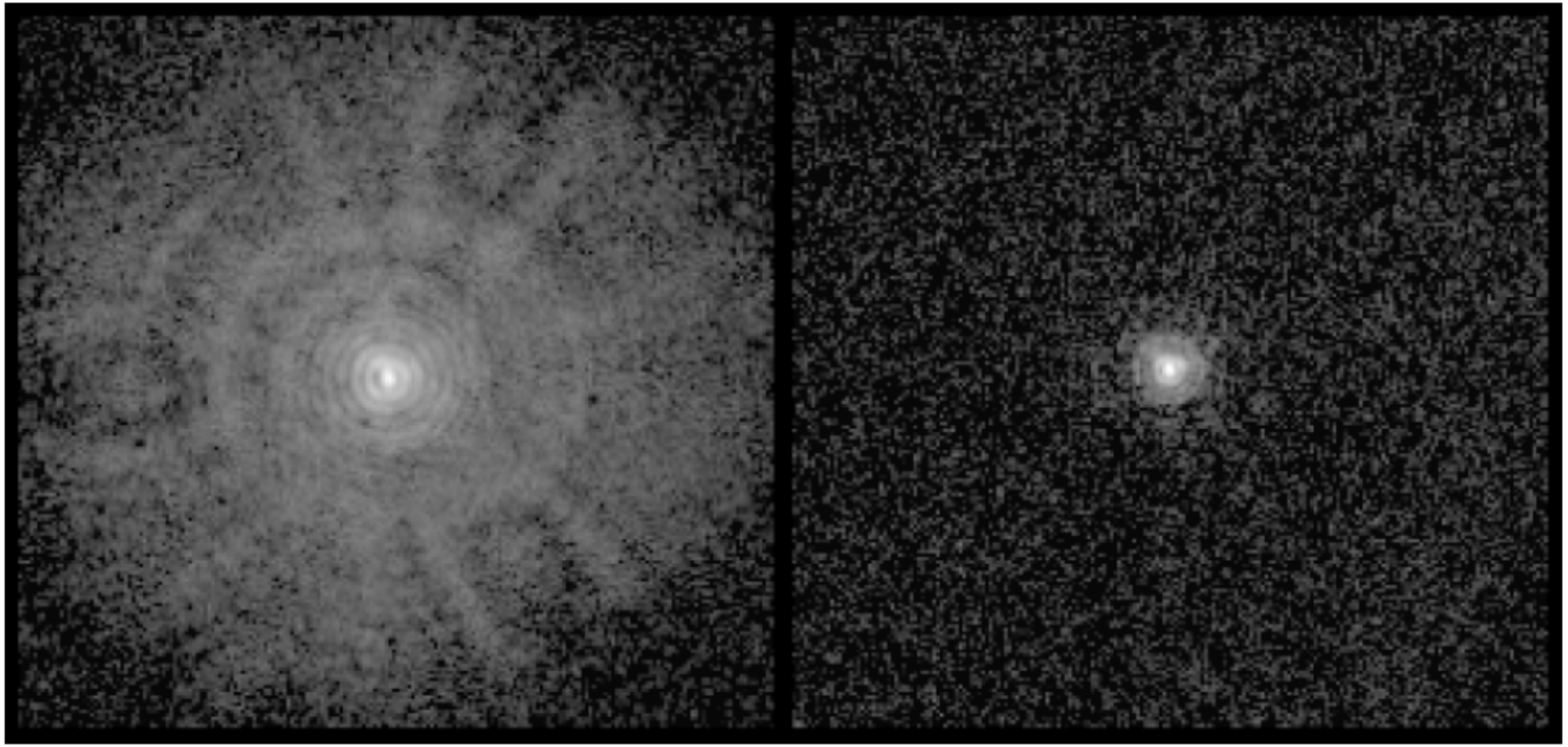


Fig. 6.7 Point-spread functions of the Faint Object Camera (FOC) of the Hubble Space Telescope. **Left**: FOC image of a star using Hubble's flawed mirror, which was around 2 μm too flat. **Right**: FOC image after the COSTAR repair mission that corrected Hubble's optics by providing it with the most expensive glasses ever made. (Copyright: Patrick P Murphy, National Radio Astronomy Observatory)

5. The nullspace