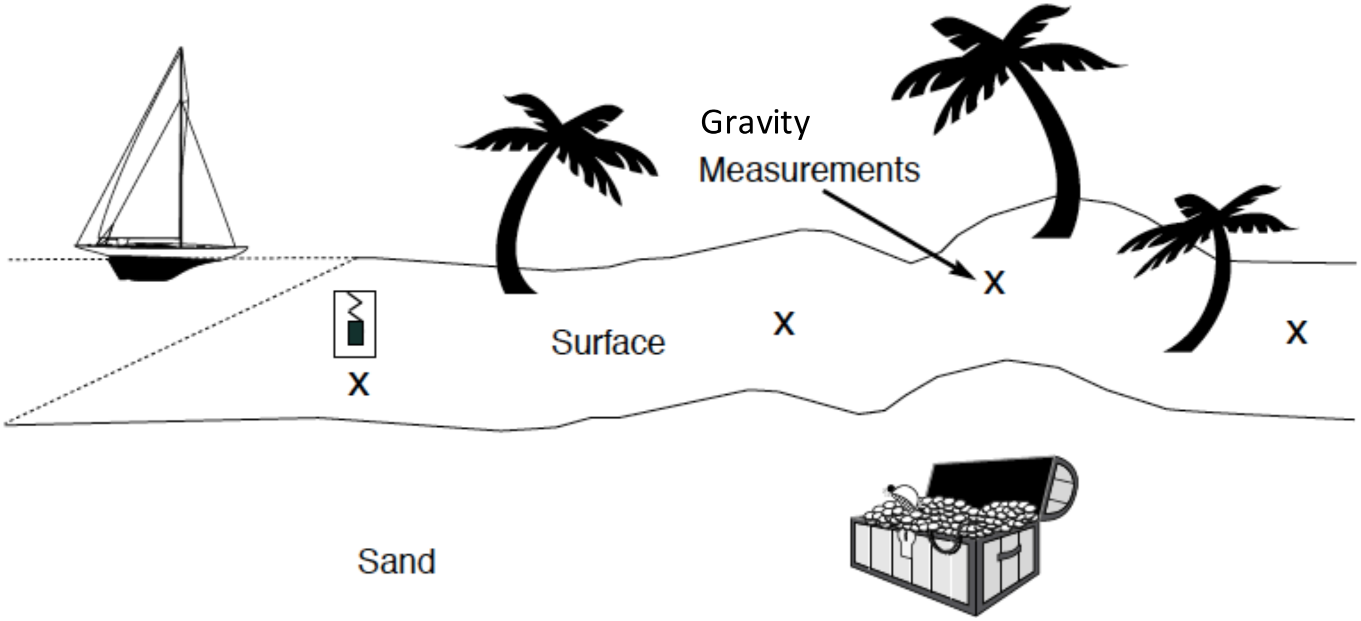
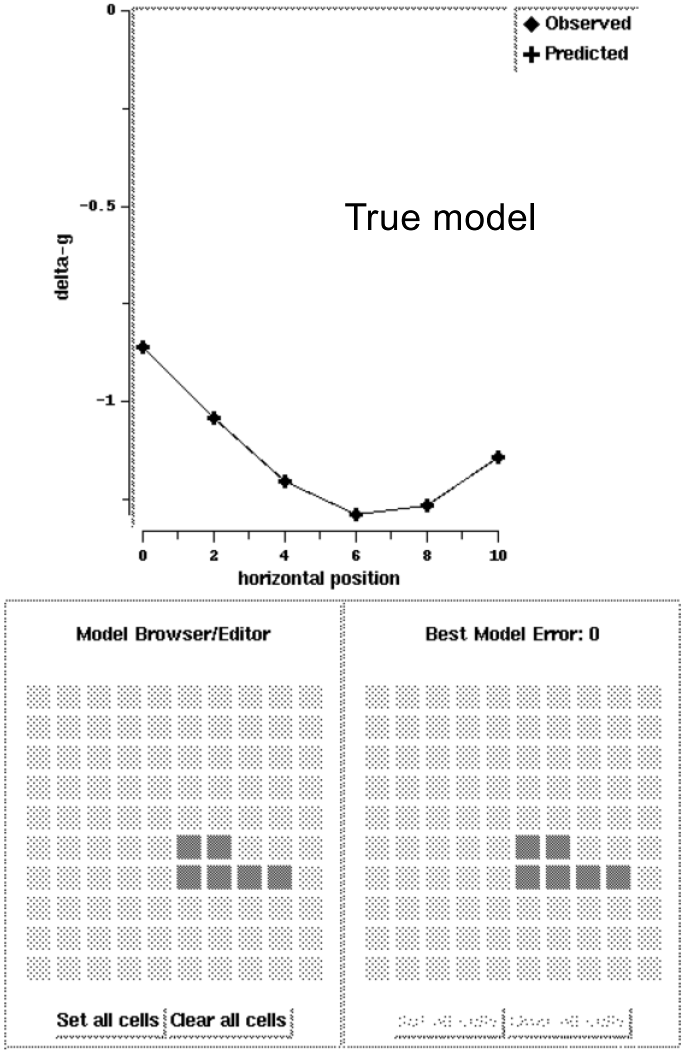
Introduction to geophysical inverse theory

Outline

- What is an inverse problem?
    - Examples
- Linear case
    - Existence/uniqueness/stability of solution
    - Overdetermined/underdetermined systems
    - Weighted, damped least squares
    - Trade-offs ; L- curve
    - Model evaluation
        - Model error
        - Resolution matrix
    - Maximum likelihood approach (if time permits)
- Non-linear case
    - Weakly non-linear optimization
        - Model evaluation
    - Parameter search methods

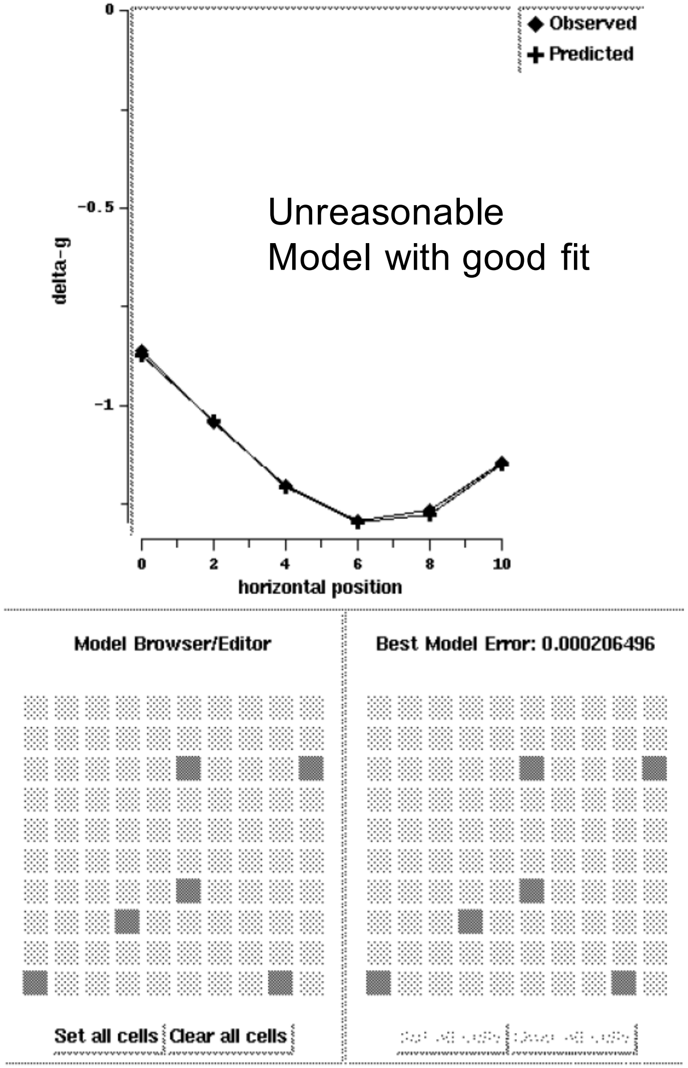
Gold buried in beach example





True model

Best Model Error: 0



Unreasonable
Model with good fit

Best Model Error: 0.000206496

Geophysical Inverse Problem

- Infer some properties of the Earth
  - -> “MODEL” : **m**

- From a set of observations
  - -> “DATA” : **d**

- Assuming a specific method which relates the model parameters to the data
  - -> “THEORY” : **F**

Forward Problem

- Predict a set of observations **d**
- $\mathbf{m} \xrightarrow{\mathbf{F}} \mathbf{d}$
- **F** is a functional that may be:
  - Explicit or implicit
  - Linear or non-linear
- F may contain theoretical assumptions/approximations

F contains the physics

Inverse problem

- Given a set of data d, estimate model parameters.

$$\mathbf{d} \xrightarrow{\mathbf{F}^{-1}} \mathbf{m}$$

- Problem can be stated implicitely:
  - $\mathbf{F}(\mathbf{d}, \mathbf{m}) = 0$
- Or explicitely
  - $\mathbf{d} = \mathbf{F}(\mathbf{m})$

Forward/inverse problem

If it is linear, this relationship can take several forms:

- Discrete:

$$d_i = \sum_{j=1}^{M} G_{ij} m_j$$

- Continuous:

$$d_i = \int G_i(x) m(x) dx$$

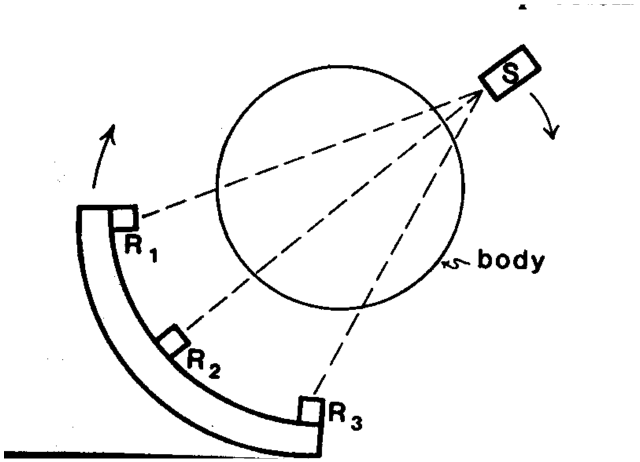
G is called the “kernel” for the problem

- Integral equation:

$$d(y) = \int G(x, y) m(x) dx$$

- Can always be reduced to discrete problem.

Example : catscan



$$\boxed{dI / ds = -c(x, y)I}$$

$$I_i = I_0 \exp\left(-\int_{\text{beam } i} c(x,y)\, ds\right)$$

Non-linear

$$\ln I_0 - \ln I_i = \int_{\text{beam } i} c(x,y)\, ds$$

linearized

Another way to linearize:

$$I_i = I_0 \exp\left(-\int_{\text{beam } i} c(x,y)\, ds\right)$$

Assume net absorption of x rays is small –
Replace exp(-x) by 1-x

Then discretize into small square boxes with constant absorption coefficient c_j

The integral becomes

$I_i = I_0 \left(1 - \sum_{j=1}^{j=m} c_j\, \Delta s_{ij}\right)$ i = 1,N

data

$$\boxed{\Delta I_i = \frac{I_i - I_0}{I_0} = \sum_{j=1}^{M} \Delta s_{ij} c_j}$$

model

→ matrix equation d=Gm

Seismic travel time tomography

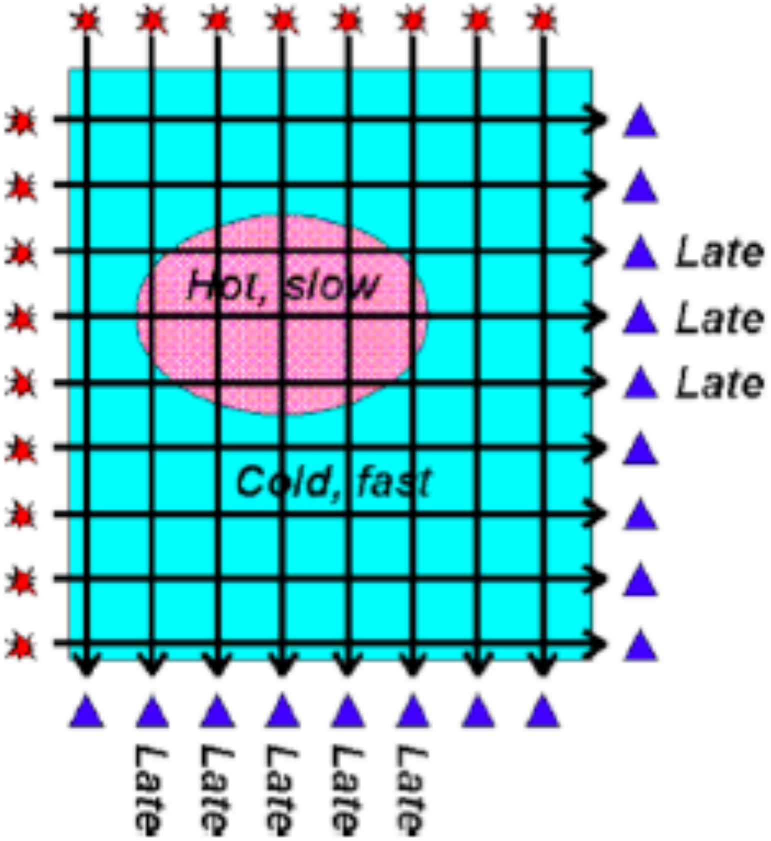
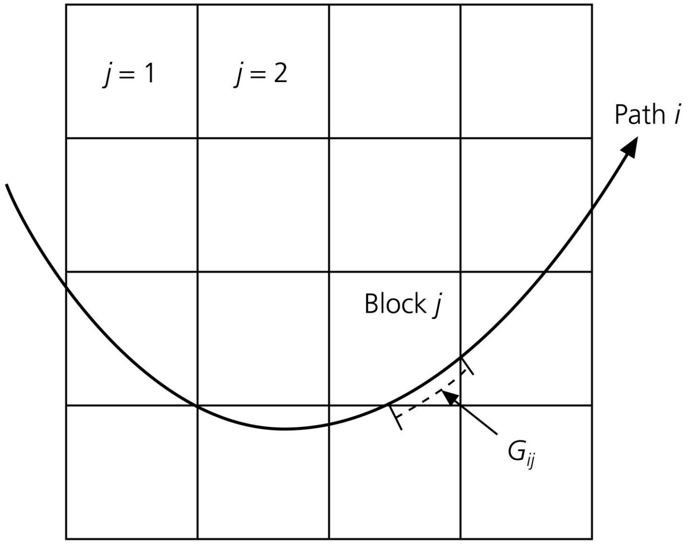


Figure 7.3-1: Geometry of a region being studied using travel time tomography.



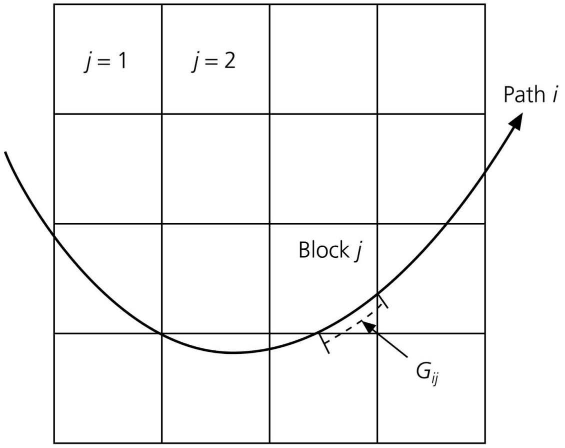
Principles of travel time tomography

1) In the background, “reference” model: Travel time T along a ray γ:

$$T = \int_{\gamma} \frac{1}{v_0(s)} ds = \int_{\gamma} u_0(s) ds$$

$v_0(s)$ velocity at point s on the ray
u= 1/v is the “slowness”

Figure 7.3-1: Geometry of a region being studied using travel time tomography.



The ray path γ is determined by the velocity structure using Snell’s law. Ray theory.

2) Suppose the slowness u is perturbed by an amount δu small enough that the ray path γ is not changed.

The travel time is changed by:

$$\delta T = \int_{\gamma} \delta u\, ds = -\int_{\gamma} \frac{1}{v_0^2} \delta v ds = -\int_{\gamma} \frac{1}{v_0} \frac{\delta v}{v_0} ds$$

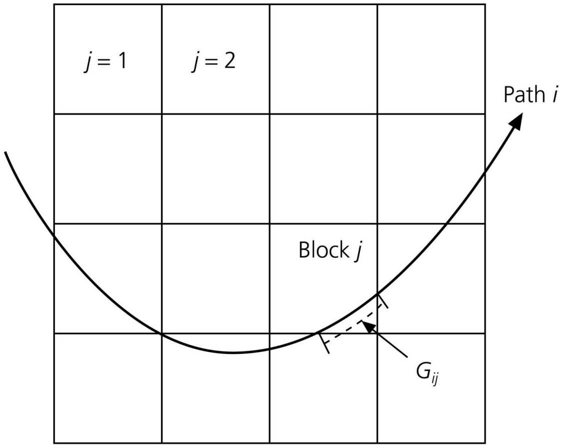
$$\delta T_i = -\int_\gamma \frac{1}{v_0(s)} \frac{\delta v}{v_0}(s) ds = \sum_{j=1}^{j=M} G_{ij} \left(\frac{\delta v}{v_0} \right)_j$$

where:

$$\mathrm{G}_{ij} = -\frac{l_{ij}}{v_0^j}$$

l_{ij} is the distance travelled by ray i in block j
v_0^j is the reference velocity ("starting model") in block j

Figure 7.3-1: Geometry of a region being studied using travel time tomography.



Solving the problem: "Given a set of travel time perturbations δT_i on an ensemble of rays {i=1…N}, determine the perturbations $(dv/v_0)j$ in a 3D model parametrized in blocks (j=1…M}" is solving an inverse problem of the form:

$$\delta\vec{d} = G\delta\vec{m}$$

d= data vector= travel time pertubations δT
m= model vector = perturbations in velocity

or

$$\delta d_i = \sum_{j=1}^{M} G_{ij} \delta m_j \qquad i = 1, N$$

$$\delta\vec{d} = G\delta\vec{m}$$

G has dimensions M x N

$$or$$

Usually N (number of rays) > M (number of blocks):
"over determined system"

$$\delta d_i = \sum_{j=1}^{M} G_{ij}\delta m_j \quad i = 1, N$$

We write: $$G^T\delta\vec{d} = G^T G\delta\vec{m}$$

G^TG is a square matrix of dimensions MxM
If it is invertible, we can write the solution as:

$$\delta\hat{m} = (G^T G)^{-1} G^T \delta d$$

where $(G^TG)^{-1}$ is the inverse of G^TG
In the sense that $(G^TG)^{-1}(G^TG) = I$, I= identity matrix

"<u>least squares solution</u>" – equivalent to minimizing $||d-Gm||_2$

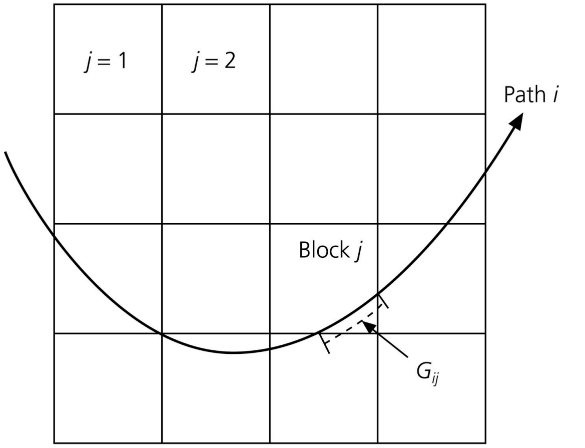
$$\delta\hat{m} = (G^T G)^{-1} G^T \delta d$$

“least squares solution”
Minimizes the sum of squared residuals:

$$\Phi = \sum_{i=1}^{i=N} \left(d_i - \sum_{j=1}^{M} G_{ij}\, m_j \right)^2$$

- G contains assumptions/choices:
  - Theory of wave propagation (ray theory)
  - Parametrization (i.e. blocks of some size)

Figure 7.3-1: Geometry of a region being studied using travel time tomography.



In practice, things are more complicated because G^TG, in general, is singular:

$$\delta d_i = \sum_{j=1}^{M} G_{ij} \delta m_j \quad i = 1, N$$

Some G_{ij} are null (l_{ij}=0)->
infinite elements in the inverse matrix

Inverse problems

- Existence of solution?
- Uniqueness of solution?
  - Null space
- Stability of the solution:
  - Many inverse problems are “ill-posed” - extreme sensitivity to initial conditions:
    - m1 and m2 may not be “close” but the corresponding data elements d1 and d2 can be very “close”

How to choose a solution?

- Special solution that maximizes or minimizes some desireable property through a norm

- For example:
  - Model with the smallest size (norm): $m^T m = ||m||_2 = (m_1{}^2 + m_2{}^2 + m_3{}^2 + \ldots m_M{}^2)^{1/2}$
  - Closest possible solution to a preconceived model $\langle m \rangle$: minimize $||m - \langle m \rangle||_2$

➔ regularization

- <u>Overdetermined system</u>
  - e.g. fitting a straight line through a set of points
  - Typically more data than unknowns (N>M)

- <u>Underdetermined system</u>
  - More unknowns than data (M>N), but equations are consistent (provide independent constraints)

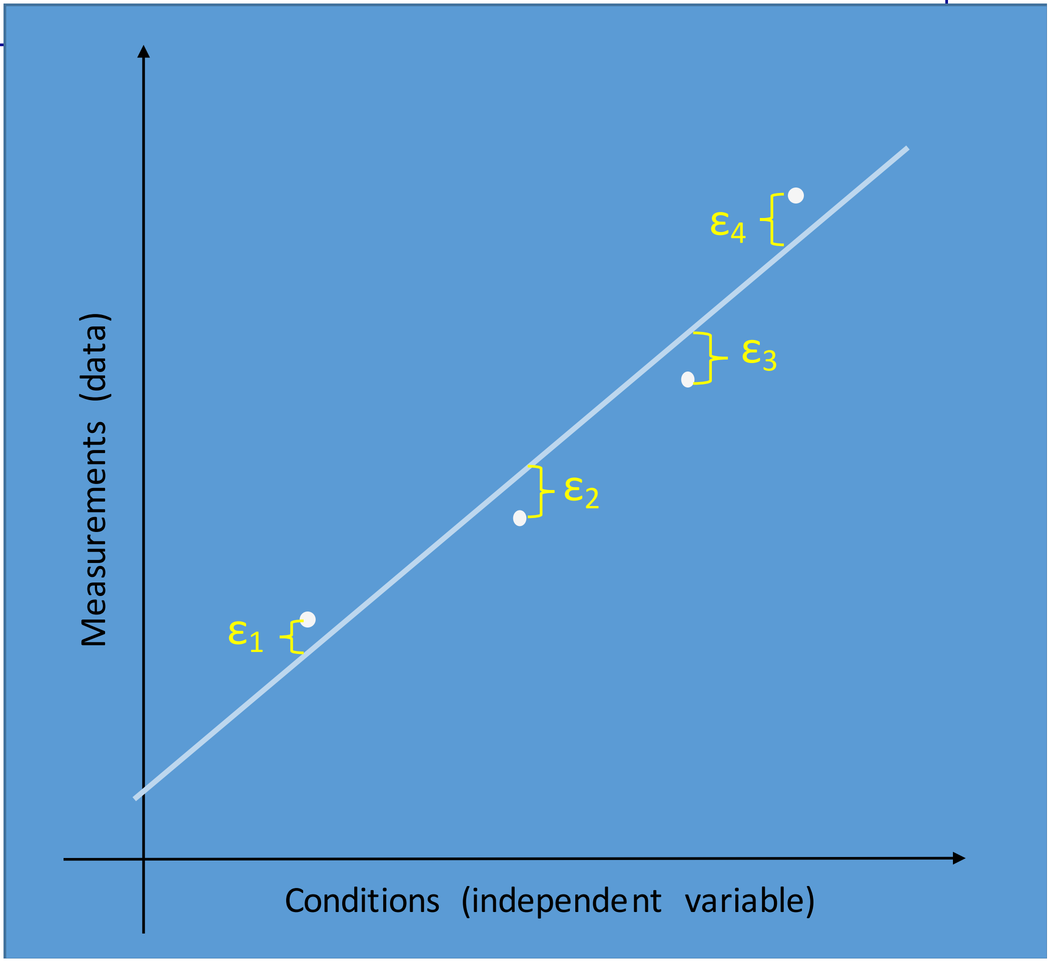
- <u>Mixed determined systems</u>
  - The case for most problems

Least squares solution for overdetermined system (e.g. straight line fitting)

- The equation to solve is:

$$d_i = m_1 + m_2 x_i, \qquad i = 1, N$$
$$N > 2$$

- Minimize the misfit at each point:
- $\epsilon_i = d_i\text{-}(m_1 + m_2 x_i)$
- That is, minimize:
- $\Phi = \varepsilon^T \epsilon = \sum_i (d_i - m_1 - m_2 x_i)^2$



- Set derivatives of E to zero:

$$\frac{\partial E}{\partial m_1} = 2Nm_1 + 2m_2 \sum x_i - 2\sum d_i = 0$$
$$\frac{\partial E}{\partial m_2} = 2m_1 \sum x_i + 2m_2 \sum x_i^2 - 2\sum d_i x_i = 0$$

$\longrightarrow$ m1,m2

- More generally:

$$E = e^T e = (d - Gm)^T (d - Gm)$$
$$= \sum_{i=1}^{N} \left[d_i - \sum_{j=1}^{M} G_{ij} m_j \right] \left[d_i - \sum_{k=1}^{M} G_{ik} m_k \right]$$
$$\frac{\partial E}{\partial m_q} = 0 = 2\sum_{k=1}^{M} m_k \sum_{i=1}^{N} G_{iq} G_{ik} - 2\sum_{i=1}^{N} G_{iq} d_i$$
$$G^T G m - G^T d = 0$$

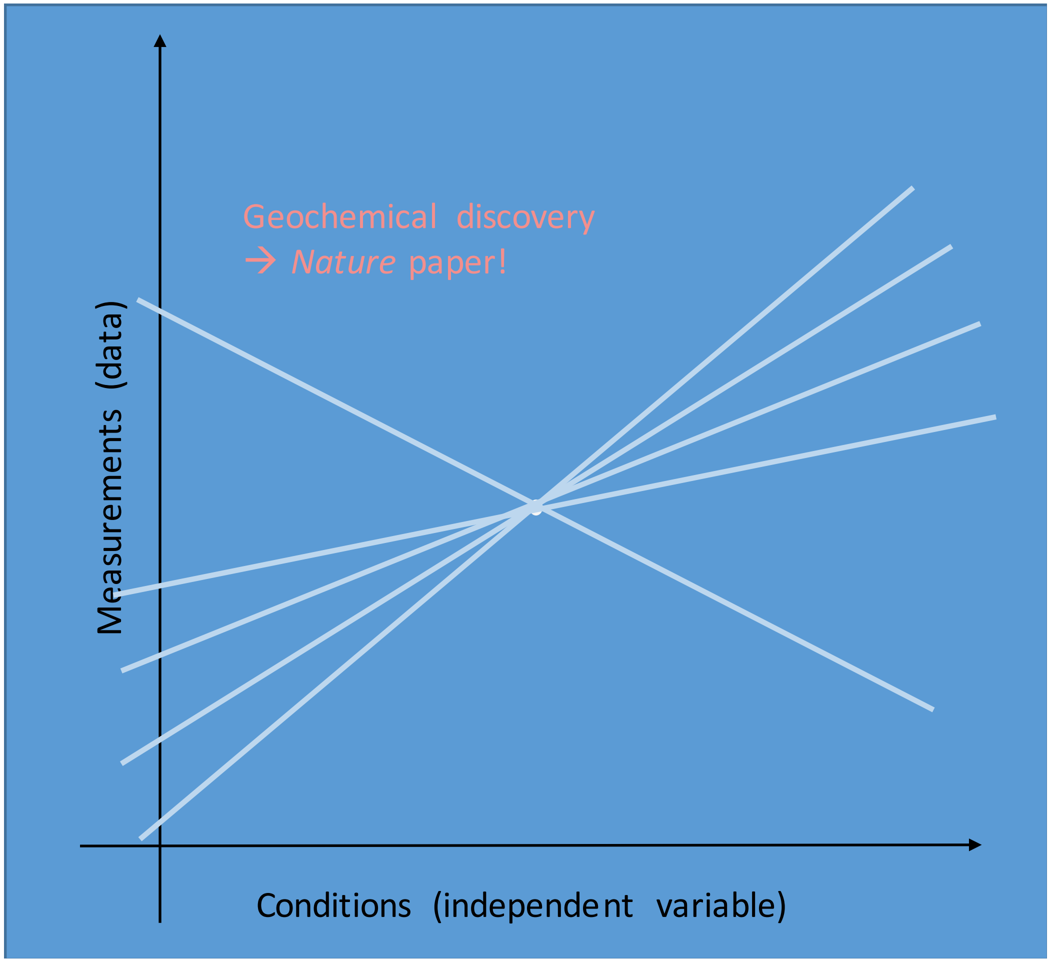
Least squares solution:

$$\longrightarrow \quad \hat{m} = \left[G^T G \right]^{-1} G^T d$$

- Solution doesn't always exist:
  - E.g straight line problem: If we have only one data point . In this case:

$$\left[G^T G\right] = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ x_1 & x_1^2 \end{bmatrix}$$

  - We need to add other constraints



Purely underdetermined problems

- Fewer equations than unknowns (N<M)
- Assume equations independent
- More than one solution – how do we choose the "right one"?
  - Add "a priori information"
    - For example: "line passes through the origin"
    - Parameters have given sign, or lie in given range (e.g. density inside the earth should be between 1 and 50 g/cm3) or its profile close to Adams-Williamson.
- Often, we choose to minimize the norm of the solution (given an appropriate norm)

- Example: minimize the size of the solution as measured by its Euclidian norm:

$$\| m \| = \left(\sum m_i^2 \right)^{1/2}$$

- Solve the following problem:
  - Find m which minimizes L=||m||^{2,} subject to the constraint that d-Gm=0
  - Use “Lagrange multipliers”, minimize:

$$\Phi(m) = \sum_{j=1}^{M} m_j^2 \;\; + \sum_{i=1}^{N} \lambda_i \left[d_i - \sum_{j=1}^{M} G_{ij} m_j \right]$$

$$\frac{\partial \Phi}{\partial m_q} = 2m_q - \sum_{i=1}^{N} \boldsymbol{\lambda}_i G_{iq} = 0$$

$$2\vec{m} = G^T \vec{\boldsymbol{\lambda}}$$

Substitute m in Gm=d:

$$d = G\left\lfloor G^T \lambda / 2 \right\rfloor$$

This gives λ and then m

$$\hat{m} = G^T \left[G G^T \right]^{-1} d$$

If equations independent, GG^T is non-singular

Mixed-determined problems

- 2 – Minimize some combination of the misfit and the solution size:

$$\Phi(m) = e^T e + \varepsilon^2 m^T m$$

e=d-Gm

- Then the solution is the "damped least squares solution":

$$\hat{m} = \left[G^T G + \varepsilon^2 I\right]^{-1} G^T d$$

Tikhonov regularization

Summary

Overdetermined:
Minimize error
“Least squares”

$$\hat{m} = \left[G^T G\right]^{-1} G^T d$$

Underdetermined:
Minimize model size
“Minimum length”

$$\hat{m} = G^T \left[G G^T\right]^{-1} d$$

Mixed-determined:
Minimize both
“Damped least squares”

$$\hat{m} = \left[G^T G + \varepsilon^2 I\right]^{-1} G^T d$$

Concept of 'Generalized Inverse'

- Generalized inverse (G^{-g}) is the matrix in the linear inverse problem that multiplies the data to provide an estimate of the model parameters;

$$\hat{m} = G^{-g} d$$

  - For Least Squares $$G^{-g} = \left[G^T G\right]^{-1} G^T$$
  - For Damped Least Squares $$G^{-g} = \left[G^T G + \varepsilon^2 I\right]^{-1} G^T$$
  - Note : Generally $G^{-g} \neq G^{-1}$

- Generalize to the case where we want to find the solution closest to some particular model <m>, called the "a priori model":
  - Replace m by m-<m>

- Generalize to other norms:
  - Example: minimize roughness, i.e. difference between adjacent model parameters.
  - Consider ||Dm|| instead of ||m|| and minimize:

$$\left[Dm\right]^T\left[Dm\right] = m^T D^T Dm = m^T W_m m$$

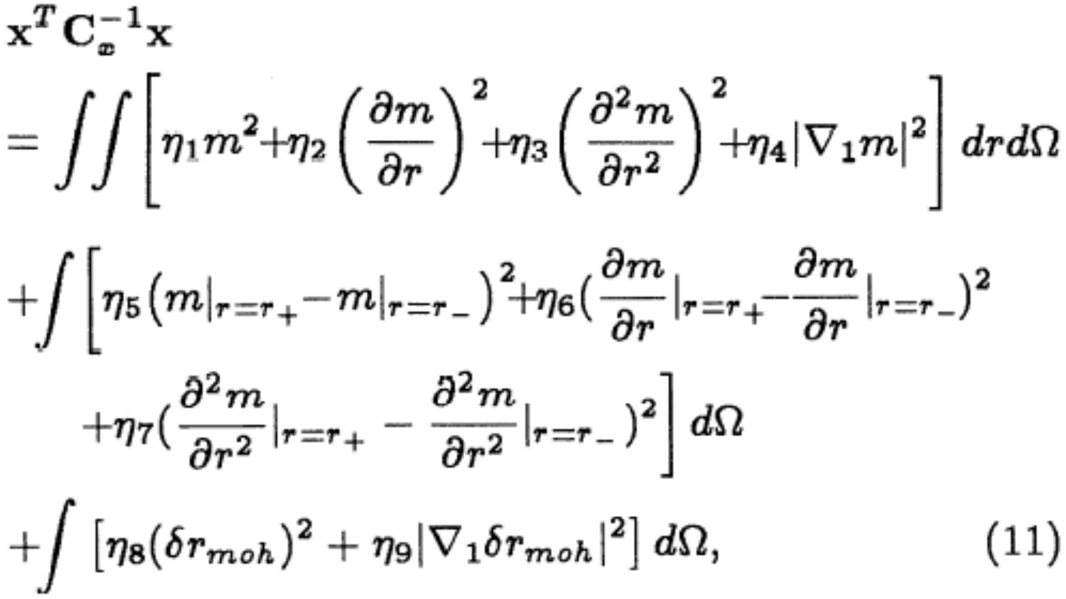
$$Dm = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \\ m_M \end{bmatrix}$$

  - More generally, minimize:

L = $(m - < m >)^T W_m (m - < m >)$

$$m^{est} = < m > + W_m G^T [G W_m G^T]^{-1} [d - G < m >]$$

Example of model covariance definition

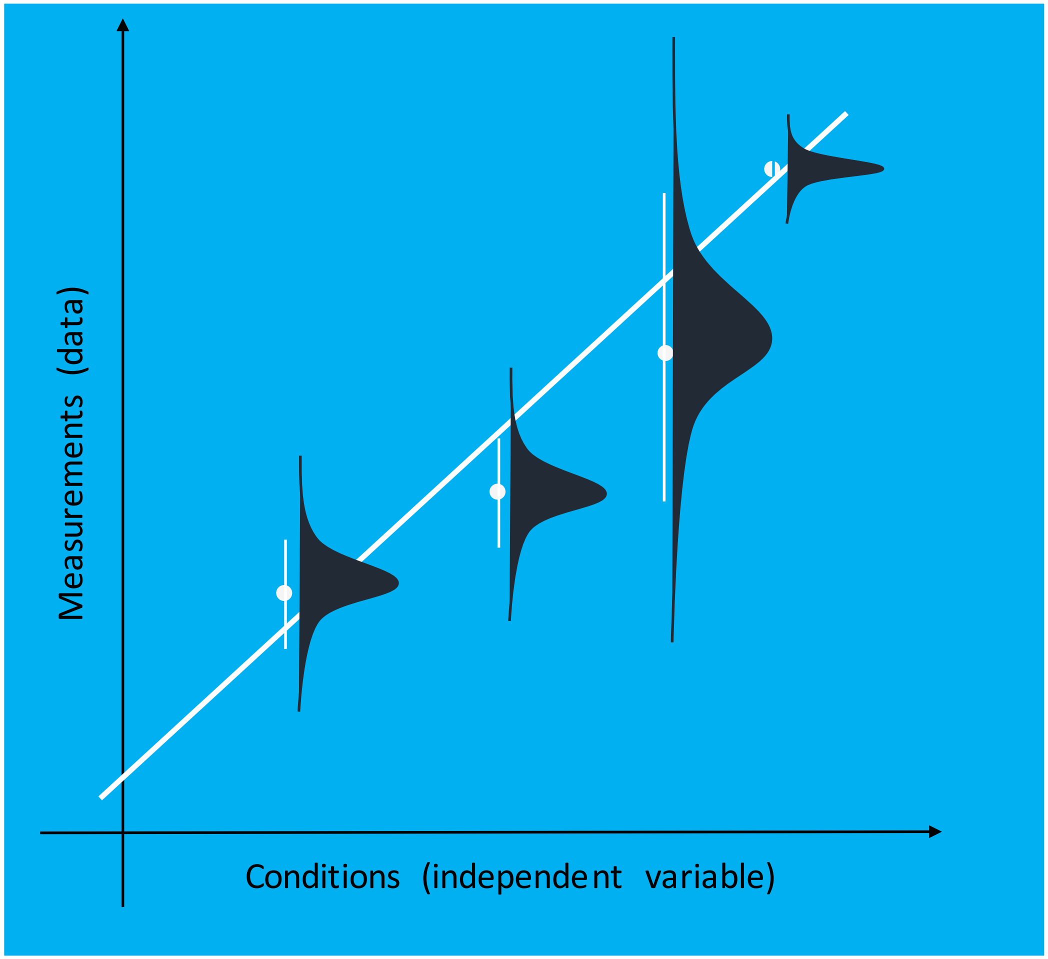


$$
\begin{aligned}
&\mathbf{x}^T \mathbf{C}_x^{-1} \mathbf{x} \\
&= \iint \left[\eta_1 m^2 + \eta_2 \left(\frac{\partial m}{\partial r} \right)^2 + \eta_3 \left(\frac{\partial^2 m}{\partial r^2} \right)^2 + \eta_4 |\nabla_1 m|^2 \right] dr d\Omega \\
&+ \int \Bigg[\eta_5 \left(m|_{r=r_+} - m|_{r=r_-} \right)^2 + \eta_6 \left(\frac{\partial m}{\partial r}|_{r=r_+} - \frac{\partial m}{\partial r}|_{r=r_-} \right)^2 \\
&\quad + \eta_7 \left(\frac{\partial^2 m}{\partial r^2}|_{r=r_+} - \frac{\partial^2 m}{\partial r^2}|_{r=r_-} \right)^2 \Bigg] d\Omega \\
&+ \int \left[\eta_8 (\delta r_{moh})^2 + \eta_9 |\nabla_1 \delta r_{moh}|^2 \right] d\Omega, \qquad (11)
\end{aligned}
$$

Li and Romanowicz, JGR, 1996 – SAW12D

- We may also want to weigh the misfits in data space (some observations more accurate than others). Instead of $\varepsilon^T\varepsilon$, minimize:
    - $\Phi=\varepsilon^T W_e \varepsilon$

$$\Phi(m) = \sum_{j=1}^{N_d} \frac{\left(d_j - g_j(m)\right)^2}{{\sigma_j}^2}$$



- Completely *overdetermined problem*:
    - Minimize Φ, then solution is:

$$m^{est} = [G^T W_e G]^{-1} G^T W_e d$$

Weighted damped least squares

- In general we will want to minimize:
  - $\Phi + \varepsilon^2 L$
- The solution then has the form:

$$m^{est} = <m> + [G^T W_e G + \varepsilon^2 W_m]^{-1} G^T W_e [d - G <m>]$$

$$or, \quad equivalently:$$

$$m^{est} = <m> + W_m^{-1} G^T [G W_m^{-1} G^T + \varepsilon^2 W_e^{-1}]^{-1} [d - G <m>]$$

For more rigorous and complete treatment (incl. non-linear):
See Tarantola (1985) Inverse problem theory
Tarantola and Valette (1982) RevGeo

Weighted damped least squares

Data weighting

Model conditioning

$$m^{est} = < m > + \underbrace{[G^T W_e G + \varepsilon^2 W_m]^{-1} G^T W_e \overbrace{[d - G < m >]}^{\text{Misfit of reference model}}}_{\text{Perturbation to reference model}}$$

Other types of a priori information

- Linear equality constraints:
  - Fm=h
  - Example: mean of the model must equal some value h_1:

$$Fm = \frac{1}{M}[11....1]\begin{bmatrix} m_1 \\ m_2 \\ \\ \\ m_M \end{bmatrix} = h_1$$

We obtain the augmented matrix equation:

$$\begin{bmatrix} G^T G & F^T \\ F & 0 \end{bmatrix}\begin{bmatrix} m \\ \lambda \end{bmatrix} = \begin{bmatrix} G^T d \\ h \end{bmatrix}$$

- Use Lagrange multipliers…minimize $\Phi=\varepsilon^T\varepsilon$ subject to the constraint Fm-h=0

Constrained fitting of a straight line

- d =m_1+m_2x, where line must pass through the point (x_0, d_0)
- Constraint:

$$Fm = \begin{bmatrix} 1 & x_0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = [d_0]$$

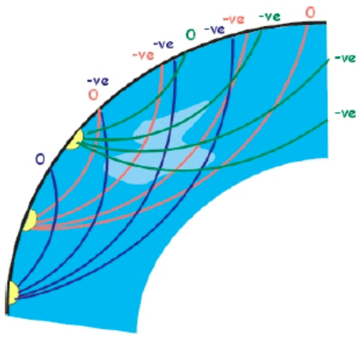
- The problem to solve is then:

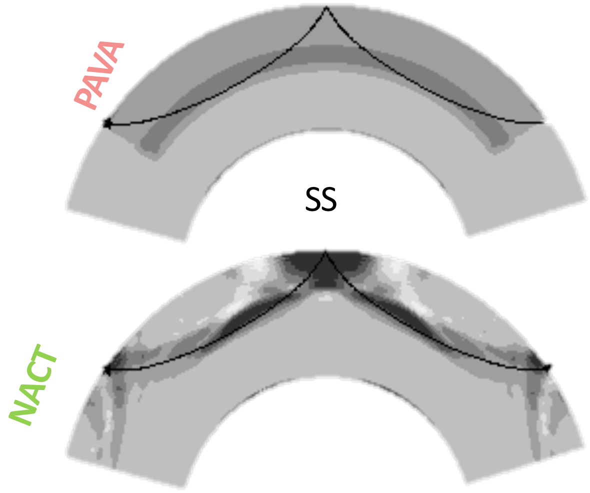
$$\begin{bmatrix} \sum d_i \\ \sum d_i x \\ d_0 \end{bmatrix} = \begin{bmatrix} N & \sum x_i & 1 \\ \sum x_i & \sum x_i^2 & x_0 \\ 1 & x_0 & 0 \end{bmatrix} \begin{bmatrix} \hat{m}_1 \\ \hat{m}_2 \\ \lambda_1 \end{bmatrix}$$

Ingredients of an inversion:

- *Importance of sampling/coverage*
  - mixture of data types
- *Parametrization*
  - Physical (Vs, Vp, ρ, anisotropy, attenuation)
  - Geometry (local versus global functions, size of blocks)
- *Theory of wave propagation*
  - e.g. for travel times: finite frequency kernels/ray theory

Whole Earth tomography





Summary points..

- In order to get more reliable and robust answers, we need to weigh the data appropriately to make sure we focus on fitting the most reliable data

- We also need to specify a priori characteristics of the model through model weighting or regularization

- These are often not necessarily constrained well by the data, and so are “tuneable” parameters in our inversions

Once a solution is found…

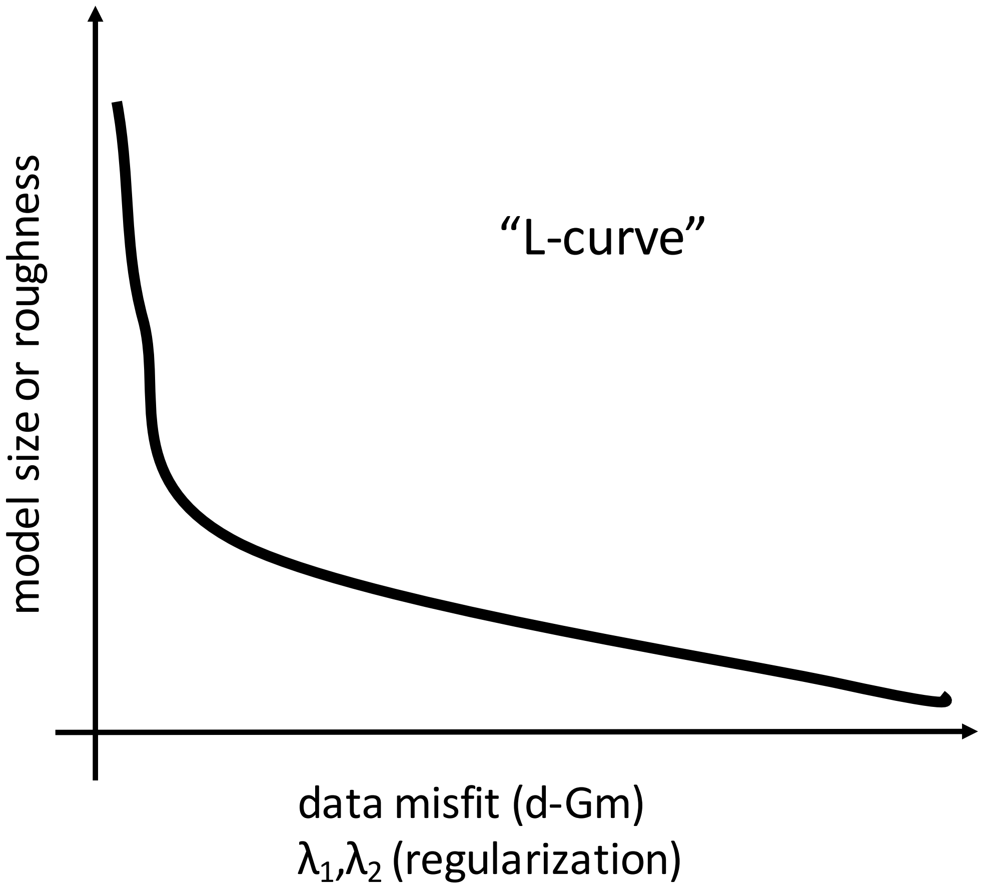
- How certain are we in our results?

- How well is the dataset able to resolve the chosen model parameterization?

- Are there model parameters or combinations of model parameters that we can’t resolve?

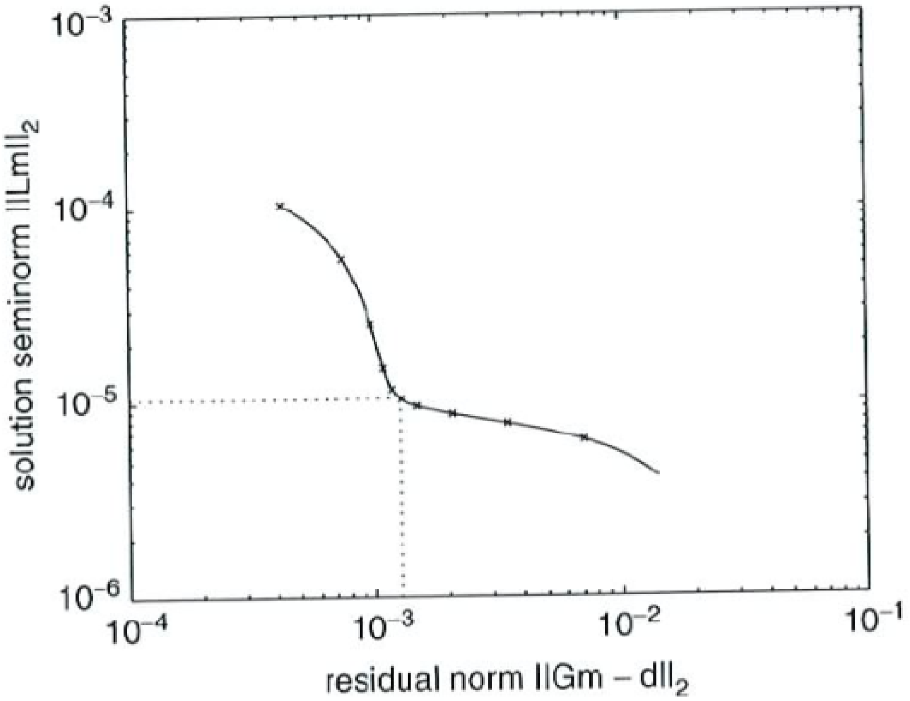
Trade offs: example of damped least squares

- As you increase the damping parameter θ, more priority is given to model-norm part of functional.
  - Increases Prediction Error
  - Decreases model structure
  - Model will be biased toward smooth solution
- How to choose θ so that model is not overly biased?
- Leads to idea of trade-off analysis.



Implementation

- Run inversion for a number of different regularization values.
- Plot data residual versus model norm for different inversions.
- Choose inversion result at knee of 'L' curve.



Model evaluation

- Model resolution:
  - - Given the geometry of data collection and the choices of model parameterization and regularization, how well are we able to image target structures?

- Model error:
  - - Given the errors in our measurements and the a priori model constraints (regularization), what is the uncertainty of the resolved model?

Model Resolution Matrix (linear case!)

- How accurately is the value of an inversion parameter recovered?
- How small of an object can be imaged ?

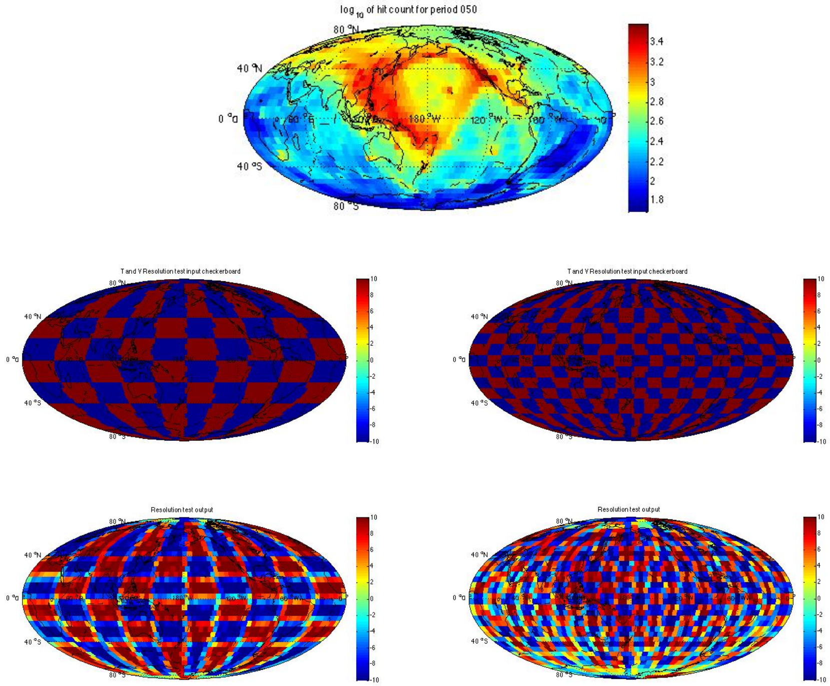
- Model resolution matrix R:

$$\hat{m} = G^{-g} d^{obs} = G^{-g} G m_{true} = R m_{true}$$

- R can be thought of as a spatial filter that is applied to the true model to produce the estimated values.
  - Often just main diagonal analyzed to determine how spatial resolution changes with position in the image.
  - Off-diagonal elements provide the 'filter functions' for every parameter.

The resolution matrix

- Think of it as a filter that runs a target model through the data geometry and regularization to see how your inversion can see different kinds of structure
- Does *not* account for errors in theory or noise in data



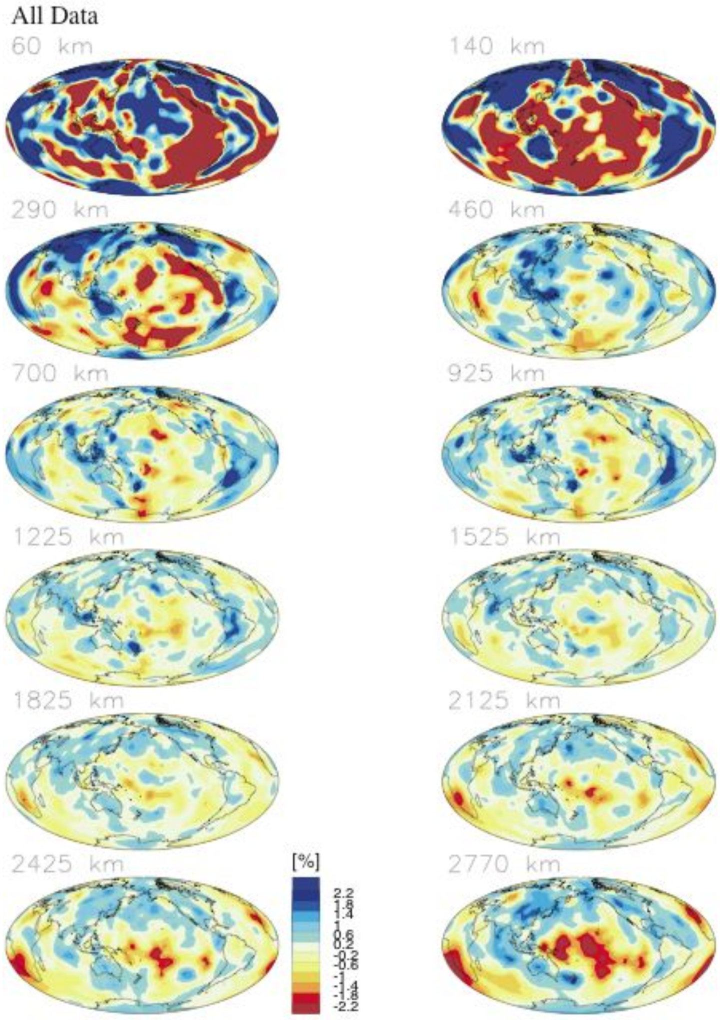
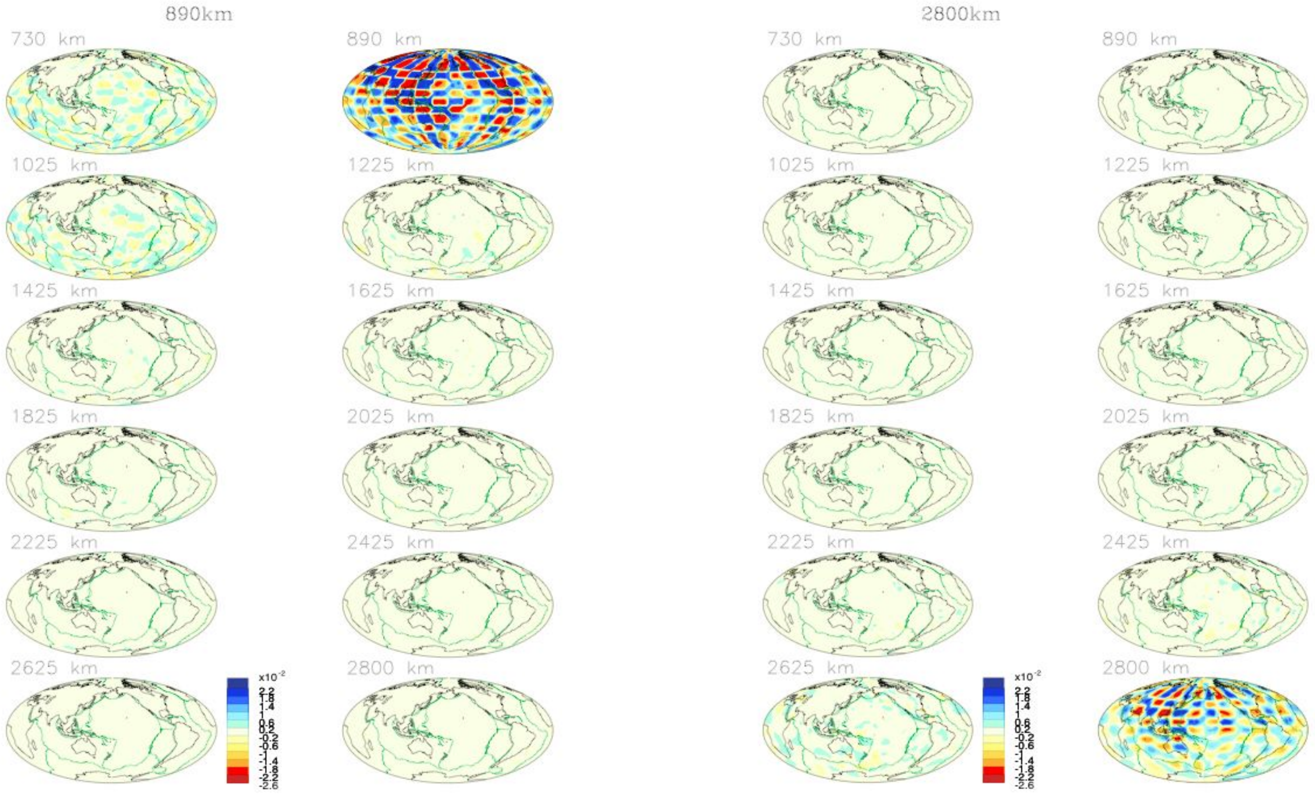


Plate 2. Cross-sections of the Earth at various depths showing perturbations of shear velocity (in percent). This high-resolution model was constructed using surface-wave and free-oscillation data as well as body wave data. Compare with plate 1. Note that the upper mantle is completely different but the anomalies in the lower mantle appear to be remarkably robust.

Masters, CIDER 2010

Checkerboard test



$$\hat{m} = Rm_{true}$$

$$R = G^{-g}G$$

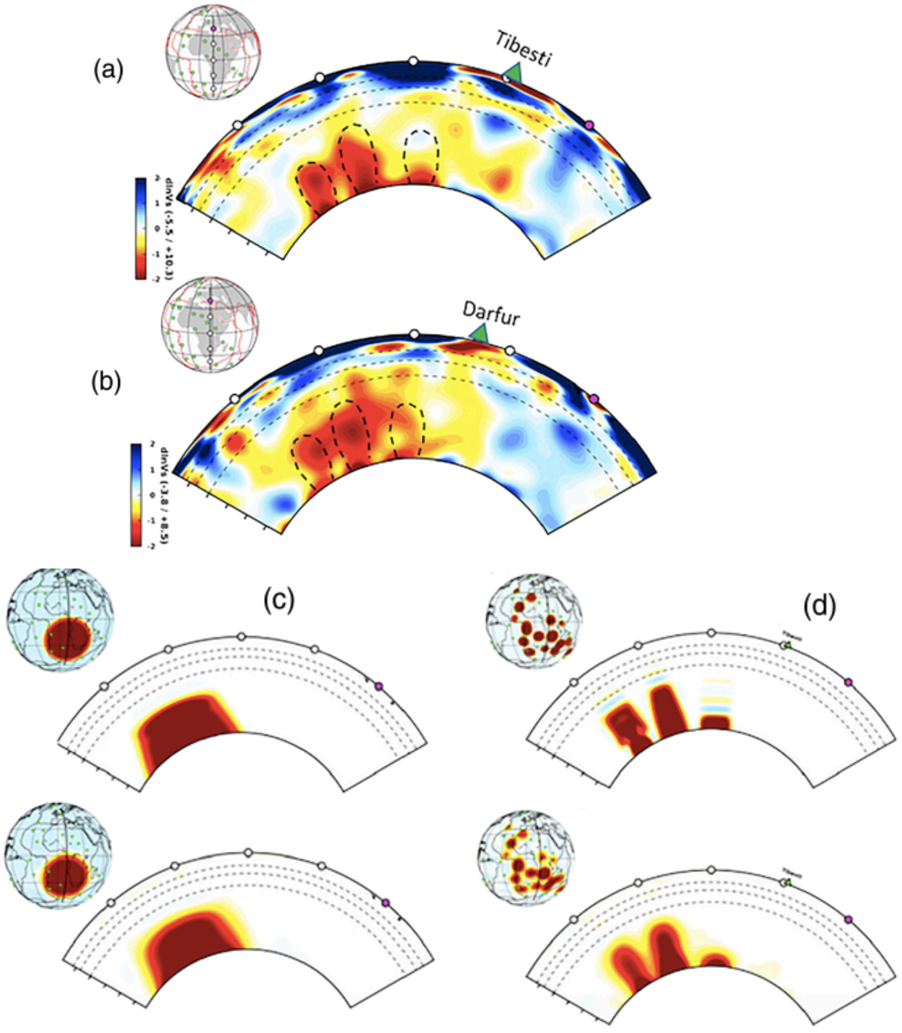
R contains theoretical assumptions
on wave propagation, parametrization
And assumes the problem is linear

After Masters, CIDER 2010

Beware the checkerboard!

- Checkerboard tests really only reveal how well the experiment can resolve checkerboards of various length scales

- For example, if the study is interpreting vertically or laterally continuous features, it might make more sense to use input models which test the ability of the inversion to resolve continuous or separated features



Davaille and Romanowicz, 2020, Tectonics

Model error

- Resolution matrix tests ignore effects of data error

- Very good apparent resolution can often be obtained by decreasing damping/regularization

- If we assume a linear problem with Gaussian errors, we can propagate the data errors directly to model error

- Simple covariance: assume data are uncorrelated with equal variance σ_d^2:

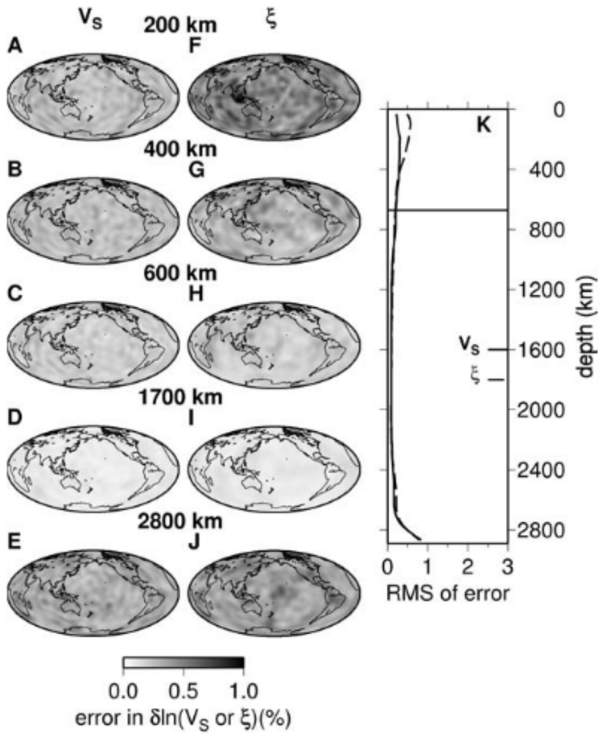
$$\hat{m} = Md + v$$

$$[\operatorname{cov} m] = M[\operatorname{cov} d]M^T$$

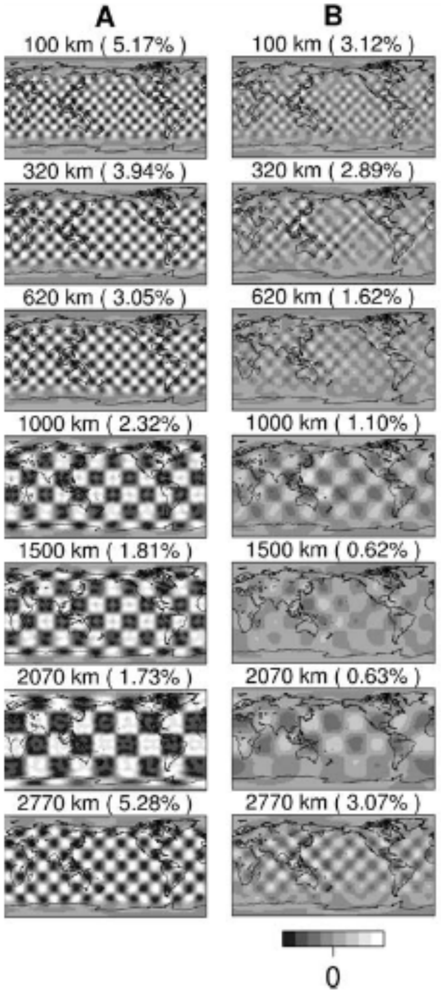
- For the least squares solution:

$$[\operatorname{cov} m] = [[G^TG]^{-1}G^T]\sigma_d^2[[G^TG]^{-1}G^T]^T = \sigma_d^2[G^TG]^{-1}$$

Linear approaches: resolution/error tradeoff



Bootstrap error map (Panning and Romanowicz, 2006)



Checkerboard resolution map

Takeaway :

- In order to understand a model produced by an inversion, we need to consider resolution and error

- Both of these are affected by the choices of regularization
  - More highly constrained models will have lower error, but also poorer resolution, as well as being biased towards the reference model

- Ideally, one should explore a wide range of possible regularization parameters

Damped weighted least squares

Data weighting

Model weighting

$$m^{est} = <m> + \underbrace{[G^T W_e G + \varepsilon^2 W_m]^{-1} G^T W_e \overbrace{[d - G<m>]}^{\text{Misfit of reference model}}}_{\text{Perturbation to reference model}}$$

Maximum likelihood estimation

I- Consider data space

- Assume that observations are independent.

- Given a model m, we have a probability function $f_i(d_i|m)$ for each observation I

- Likelihood function: joint probability density for a vector of independent observations d:

$$L(m\,|\,d) = f_1(d_1\,|\,m) f_2(d_2\,|\,m)..f_n(d_n\,|\,m)$$

- Choose the model that MAXIMIZES L(m/d)

Data with normally distributed errors

- If the linear inverse problem is discrete, the Max. Likelihood solution is then the least squares solution.
- To show this:
  - Assume that data have independent random errors that are normally distributed
- Probability density for d_i, with standard deviation σ_i and expected value zero:

$$f_i(d_i \mid m) = \frac{1}{(2\pi)^{1/2}\sigma_i} e^{-(d_i-(Gm)_i)^2/2\sigma_i^2}$$

- Likelihood function is :

$$L(m \mid d) = \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^{n}\sigma_i} \Pi_{i=1}^{n} e^{-(d_i-(Gm)_i)^2/2\sigma_i^2}$$

$$L(m \mid d) = \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^{n}\sigma_i}\Pi_{i=1}^{n}e^{-(d_i-(Gm)_i)^2/2\sigma_i^2} = A\exp(-\sum_{i=1}^{n}(d_i-(Gm)_i)^2/2\sigma_i^2)$$

$$= A\exp(-\frac{1}{2}(d-Gm)^t C_D^{-1}(d-Gm)\ \ )$$

- L is a monotonically increasing function:
  - Maximize L <-> maximize log L. The problem becomes a minimization problem:

$$\min\sum_{i=1}^{n}\frac{(d_i-(Gm)_i)^2}{\sigma_i^2} = \min\left[(d-Gm)^t C_D^{-1}(d-Gm)\right]$$

i.e. least squares

  - This amounts to scaling the system of equations by a diagonal weighting matrix:

$$W = C_D^{-1} = diag(1/\sigma_1, 1/\sigma_2 .. 1/\sigma_n)$$

II- Now consider Model space

- A priori information on the model m:
  - Assume m is a sample of a known Gaussian distribution whose mean is <m> (the "prior" – what we previously called the "reference model") and covariance matrix C_M. The corresponding probability density function is, in model space:

$$\rho_M(m) = A\exp(-\frac{1}{2}(m-<m>)^t C_M^{-1}(m-<m>)$$

  - where A is a constant

- Assuming that data and model are independent, the joint likelihood function will have the form:

$$L(m \mid d) = \frac{1}{(2\pi)^{n/2}\Pi_{i=1}^{n}\sigma_i}\Pi_{i=1}^{n} e^{-(d_i-(Gm)_i)^2/2\sigma_i^2} \times \rho_M(m)$$

$$= A\exp\left[-\sum_{i=1}^{n}(d_i-(Gm)_i)^2/2\sigma_i^2 - \frac{1}{2}(m-<m>)^t C_M^{-1}(m-<m>)\right]$$

$$= A\exp\left[-\frac{1}{2}(d-Gm)^t C_D^{-1}(d-Gm) - \frac{1}{2}(m-<m>)^t C_M^{-1}(m-<m>)\right]$$

Maximizing L(m/d) is then equivalent to minimizing:

$$2S(m) = (d-Gm)^t C_D^{-1}(d-Gm) + (m-<m>)^t C_M^{-1}(m-<m>)$$

Which results in the already established solution in the linear case:

Damped weighted least squares (linear case) - > Maximum likelihood solution

Data weighting

Model weighting

$$m^{est} =< m > +[G^T W_e G + \varepsilon^2 W_m]^{-1} G^T W_e \overbrace{[d - G < m >]}^{\text{Misfit of reference model}}$$

Or: $$m^{est} = m_{prior} \underbrace{+ (G^t C_D^{-1} G + C_M^{-1})^{-1} G^t C_D^{-1} (d_{obs} - G m_{prior})}_{\text{Perturbation to reference model}}$$

$W_e = C_D^{-1}$

$W_m = C_M^{-1}$

With <m> = “m prior”

A posteriori covariance: $\tilde{C}_M = (G^t C_D^{-1} G + C_M^{-1})^{-1}$

What about non-linear problems?

General non-linear inverse problem:
Tarantola and Valette formalism

Tarantola A. and B. Valette (1982) Rev. Geophys. 20, 219-232.

Non-linear case

- In this case, we have: d = g(m), where g is non-linear, and we cannot replace g by the matrix G. We have to write:

$$L(m\,|\,d) =$$

$$= A\exp\left[-\frac{1}{2}(d-g(m))^t C_D^{-1}(d-g(m)) - \frac{1}{2}(m-<m>)^t C_M^{-1}(m-<m>)\right]$$

- And minimize:

$$2S(m) = (d-g(m))^t C_D^{-1}(d-g(m)) + \frac{1}{2}(m-<m>)^t C_M^{-1}(m-<m>)$$

- Case I: weak non-linearity

  - g(m) can be linearized around m_{prior}=<m>

$$g(m) \sim g(m_{prior}) + G(m - m_{prior}) \qquad G_{ij} = \frac{\partial g^i}{\partial m_j}(m_{prior})$$

  - Then minimization problem leads to the solution:

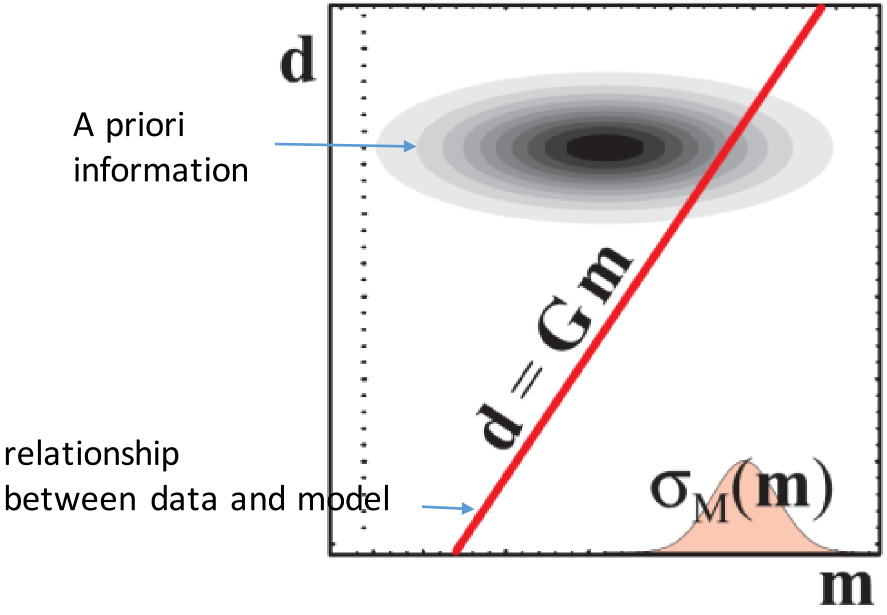
$$m_{est} = m_{prior} + (G^t C_D^{-1} G + C_M^{-1})^{-1} G^t C_D^{-1}(d_{obs} - g(m_{prior}))$$

  - The a posteriori covariance is still:

$$\tilde{C}_M = (G^t C_D^{-1} G + C_M^{-1})^{-1}$$

Linear/ Weakly non-linear

Linear – least squares works and $\tilde{C}_M$ is accurate

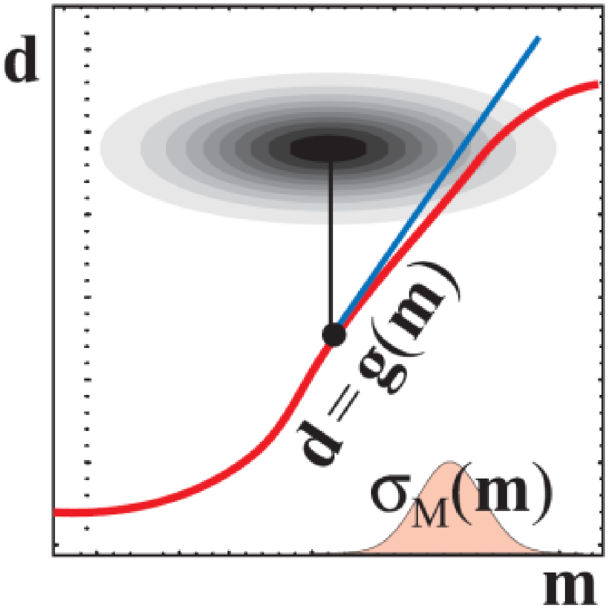


A priori information

relationship between data and model

Linearizable around starting model

$\tilde{C}_M$ is probably OK



Tarantola, 2005

- In the case when non-linearity is too strong, but we can assume g(m) is linear in a neighborhood containing significant model probability, then we need to use an iterative method in order to approach the solution correctly.

  - Starting if possible as close as possible to the global maximum.
  - Often one chooses $m_0 = m_{prior}$

  - Eg. Iterative quasi Newton method will write:

$$m_{n+1} = m_n + (G_n^t C_D^{-1} G_n + C_M^{-1})^{-1} \left[G_n^t C_D^{-1} (d_{obs} - d_n) + C_M^{-1} (m_{prior} - m_n) \right]$$

  - With $d_n = g(m_n)$ and G_n calculated at m_n.

Importance of theory (i.e. "g"): forward versus inverse problem

- It is *incorrect* to replace g(m) by Gm in the term d-g(m) that appears at the n'th iteration of the inversion.
  - *This is only correct if the problem is linear*

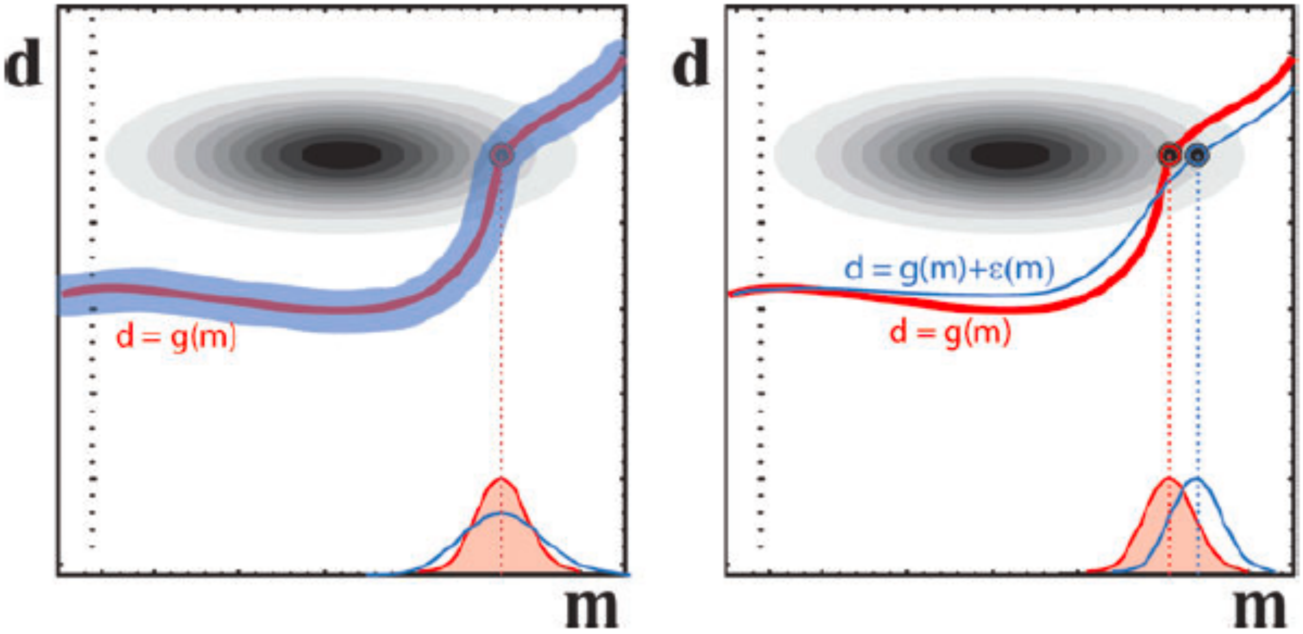
- Likewise, it is important to have as accurate a theoretical representation of g(m) in this residual term, as possible.
  - *d-g(m) defines the (non-flat) misfit surface and this allows us to determine the location of the true minimum*

- On the other hand, one can be more tolerant of approximations in the computations of the matrix G at each step
  - *In the limit, all one needs to know is an approximate direction of the gradient. If it is approximately right, but not rigorously "true", the next iteration will help correct it*

- In the appendix of Lekic and Romanowicz (2011), we introduce errors in both g and G and show that the bias in G has a *second order* effect on the error in the solution**, whereas the bias in g has a *first order* effect.

Case 1:
Error on g has a gaussian distribution around the true relationship
➾ Model distribution is wider but still Gaussian

Case 2:
Non-Gaussian bias in the theory “g”
-=>The resulting model is biased



Lekic and Romanowicz, 2011 GJI – after Tarantola (2005) Fig 3.2

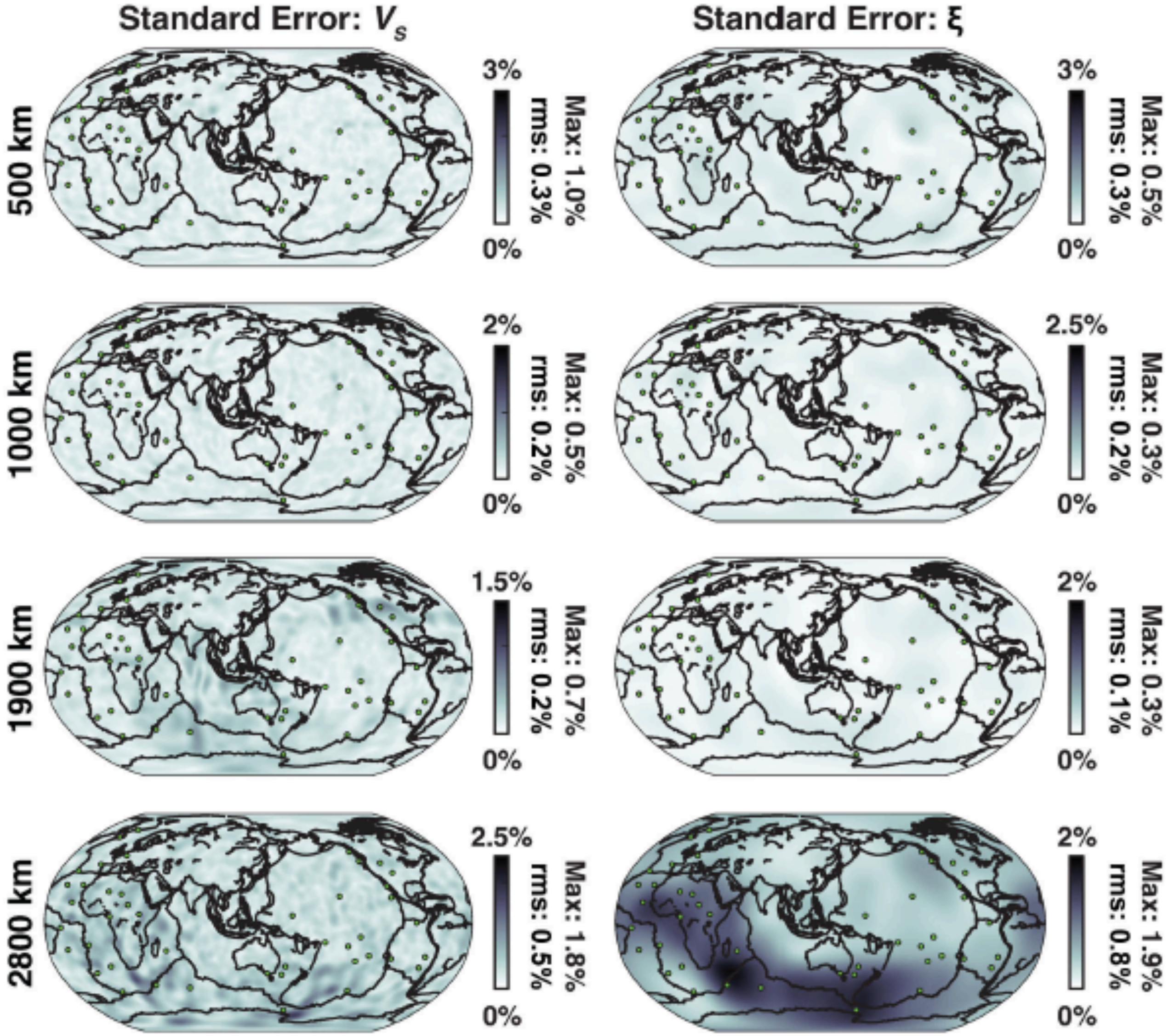
=> choice of using SEM for forward modeling part of full waveform inversion
and less heavy computationally NACT (asymptotic mode coupling) for the inverse part
This also allows us to use the faster converging Gauss-Newton inversion method (as opposed to conjugate-gradients as in inversions based on “adjoints”

Standard error by Jackknife approach

$$\hat{\sigma}_{\text{jackknife}} = \sqrt{\frac{n-d}{d \cdot C} \sum_{i=1}^{C} \left[\theta_{(i)} - \hat{\theta}\right]^2},$$

$$\text{where } \hat{\theta} = \frac{1}{C} \sum_{i=1}^{C} \theta_{(i)} \text{ and } C = \frac{n!}{d!(n-d)!},$$

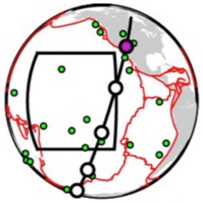
n = number of datasets =12
d = number of deletions per Dataset realization



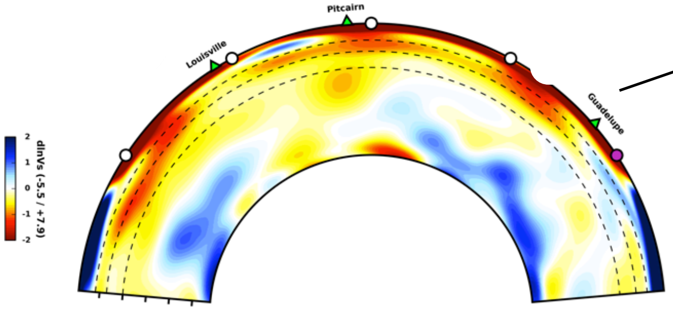
Model SEMUCB_WM1
– French and Romanowicz (2015)

Inversion with different starting model

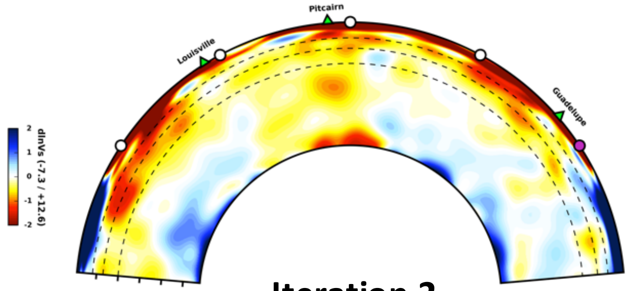
PITCAIRN



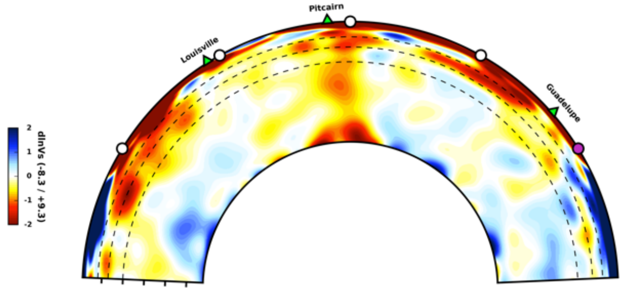
Starting model:S362ANI



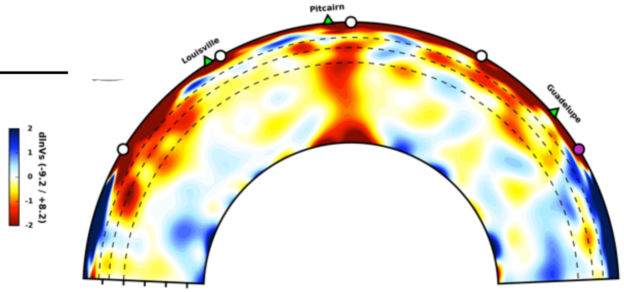
Iteration 1



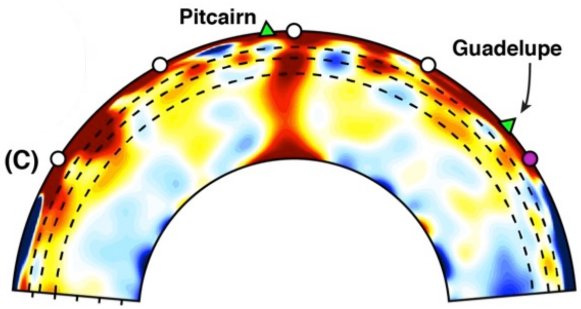
Iteration 2



Iteration 3

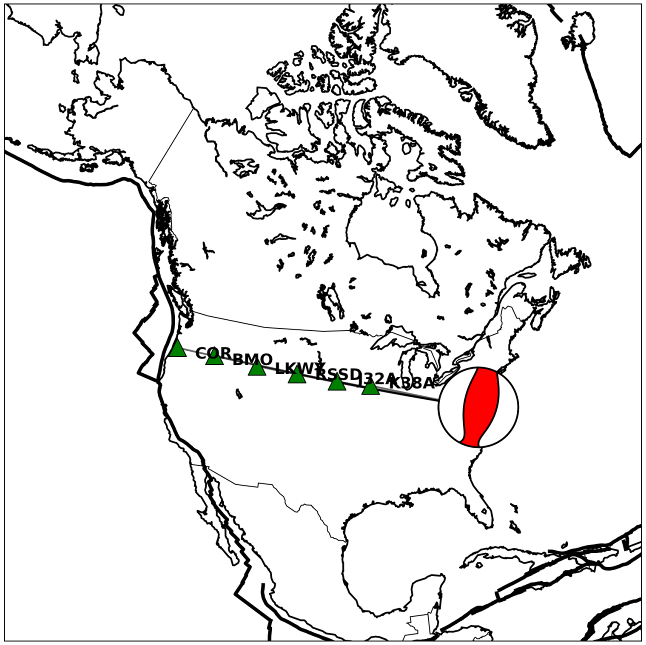


Target model: SEMUCB_WM1

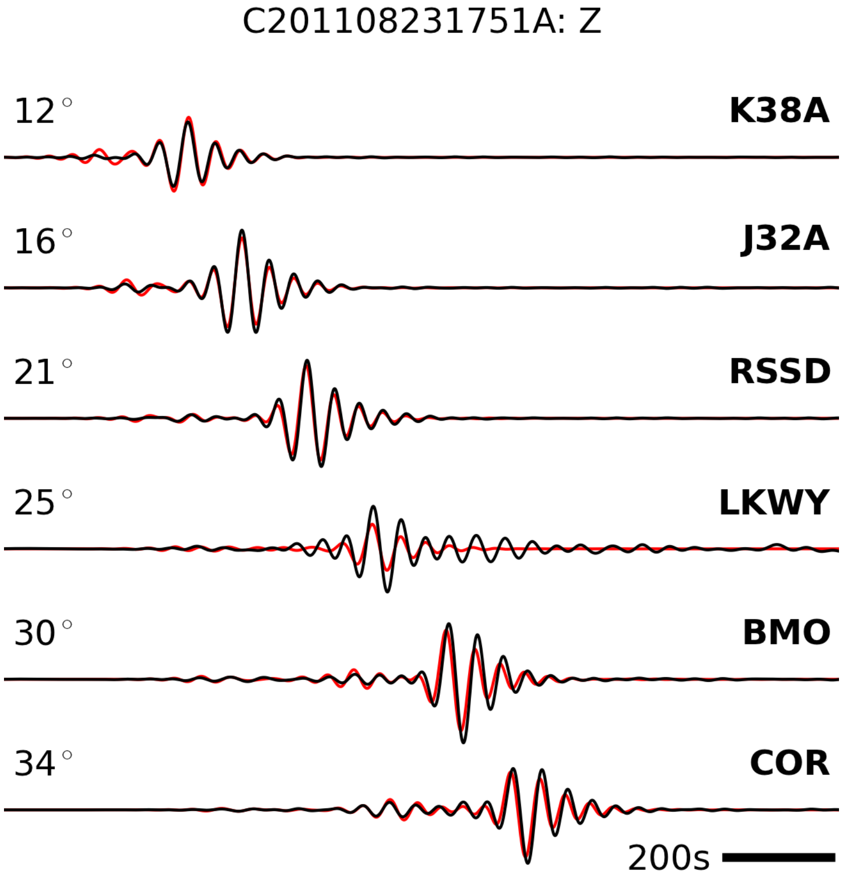


Romanowicz et al., unpublished

SEMum2 validation using RegSEM



08/23/2011 Virginia eq.Mw 57

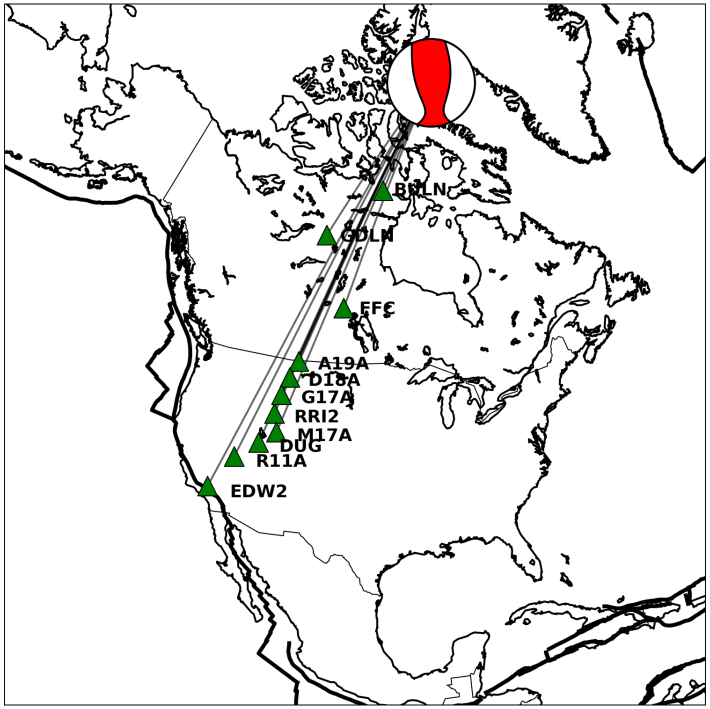


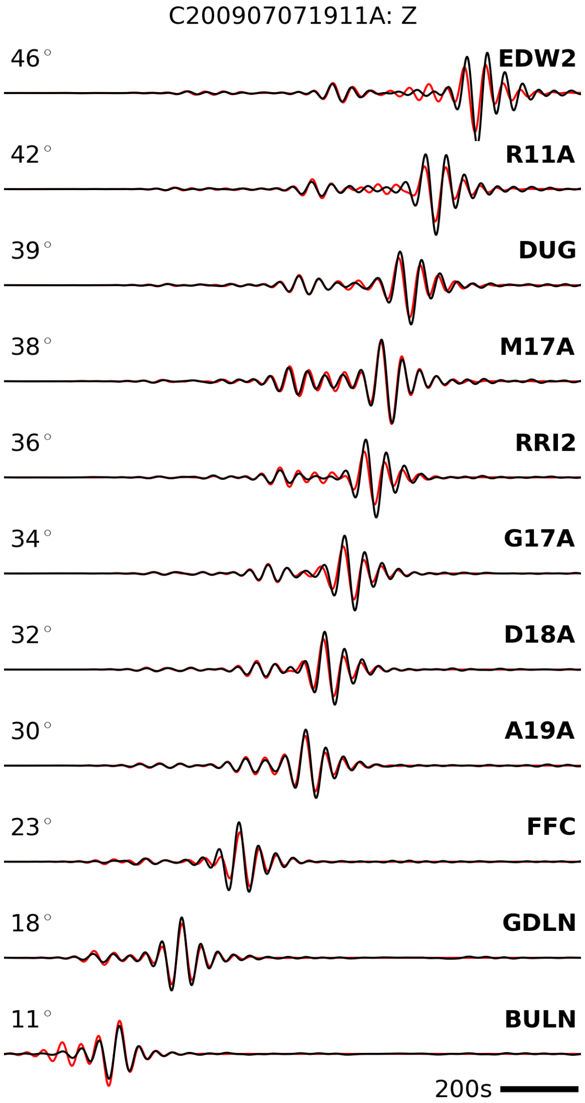
Data

Synthetics SEMum2

SEMum2 constructed with data at Δ>15°,T>60 s;
Here: comparison is shown down to 40 s and regional distances
Event not used in the inversion

French et al., 2013



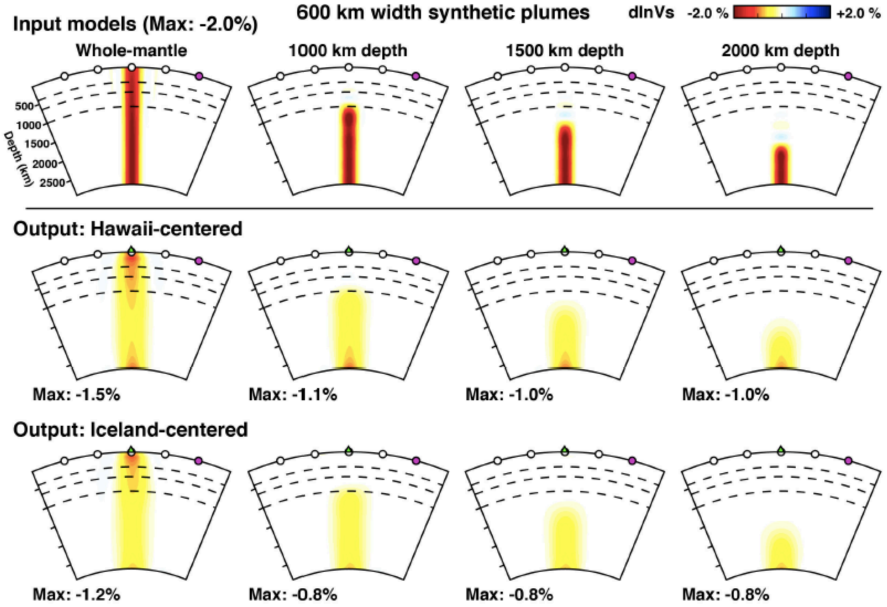


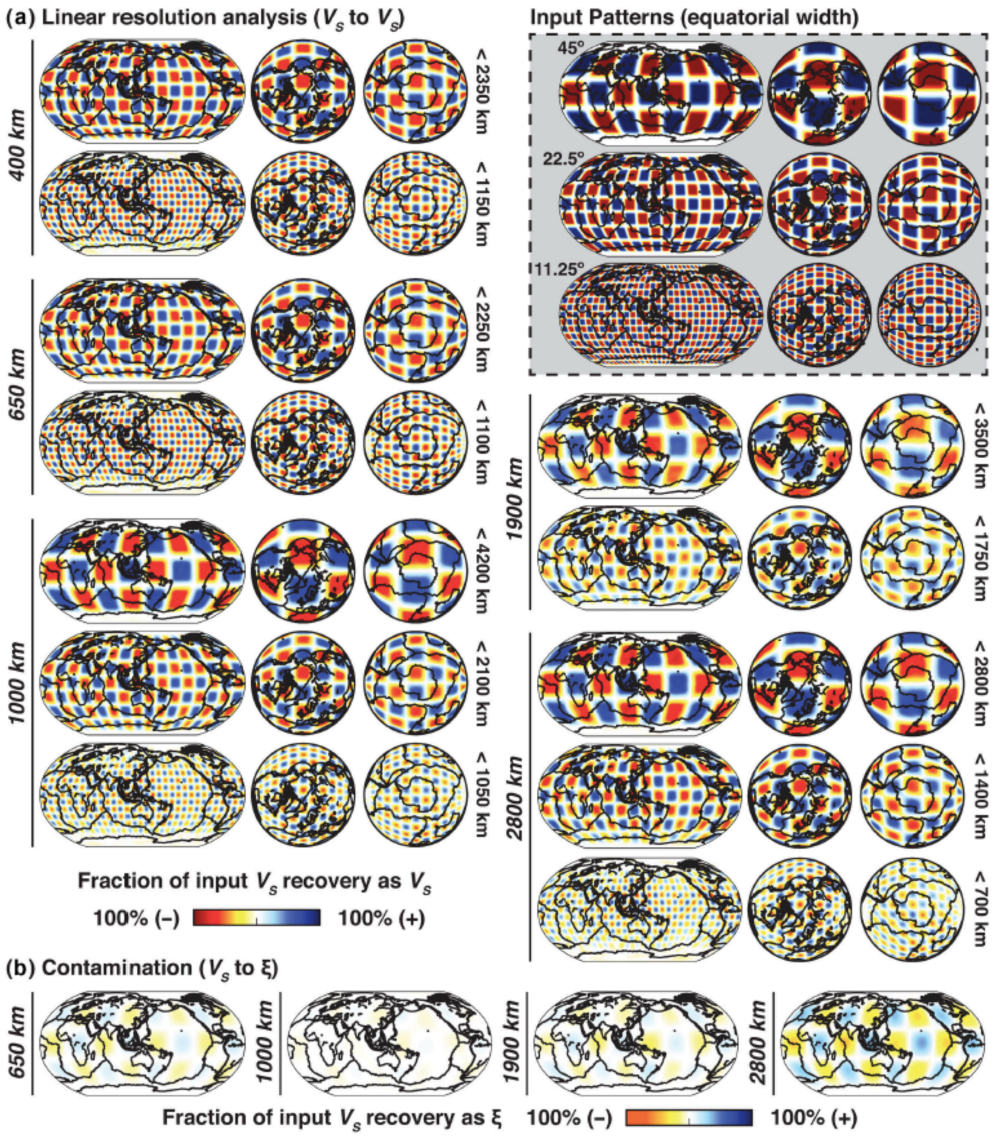
Data

Synthetics SEMum2

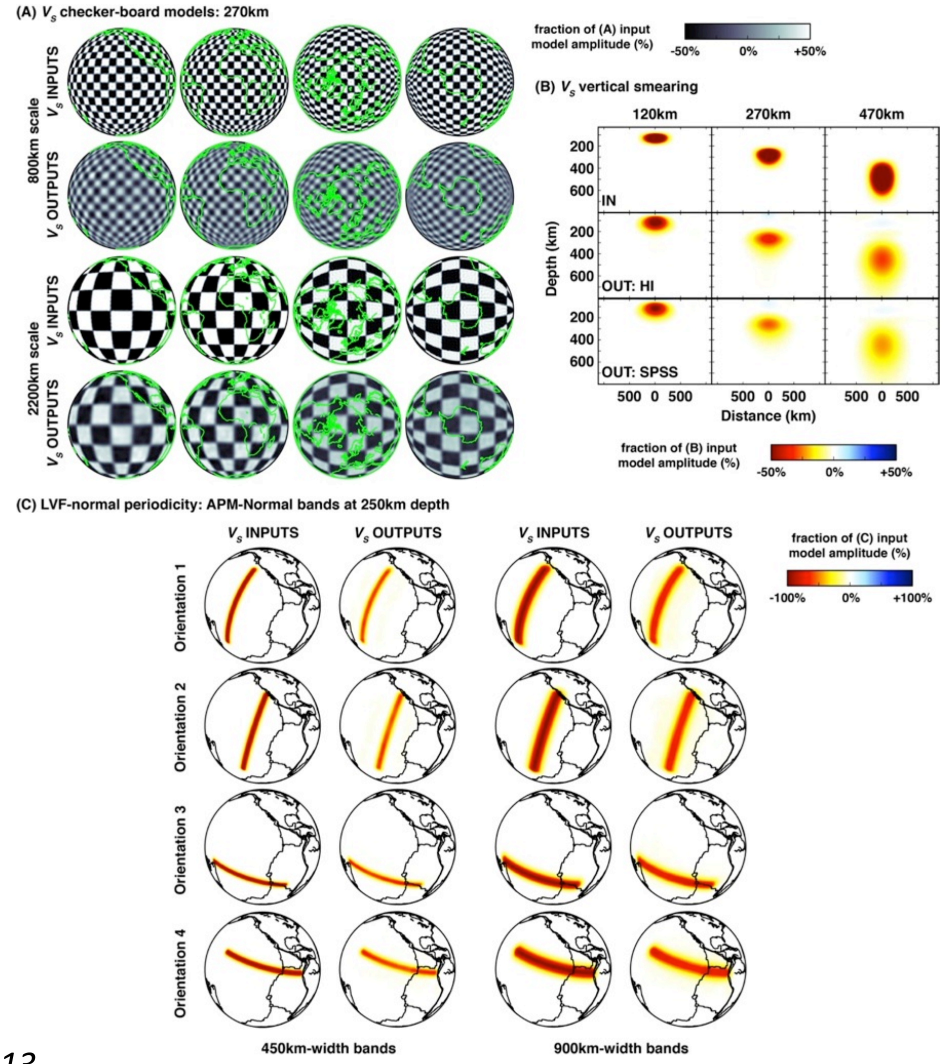
SEMum2 validation using RegSEM

Linear resolution analysis




(a) Linear resolution analysis (V_s to V_s)
Input Patterns (equatorial width)
45°
22.5°
11.25°
400 km
< 2350 km
< 1150 km
650 km
< 2250 km
< 1100 km
1000 km
< 4200 km
< 2100 km
< 1050 km
1900 km
< 3500 km
< 1750 km
2800 km
< 2800 km
< 1400 km
< 700 km
Fraction of input V_s recovery as V_s
100% (–) 100% (+)
(b) Contamination (V_s to ξ)
650 km
1000 km
1900 km
2800 km
Fraction of input V_s recovery as ξ 100% (–) 100% (+)

French and Romanowicz (2014)



SEMum2:
French, Lekic, Romanowicz, 2013

Linearizable around most likely model

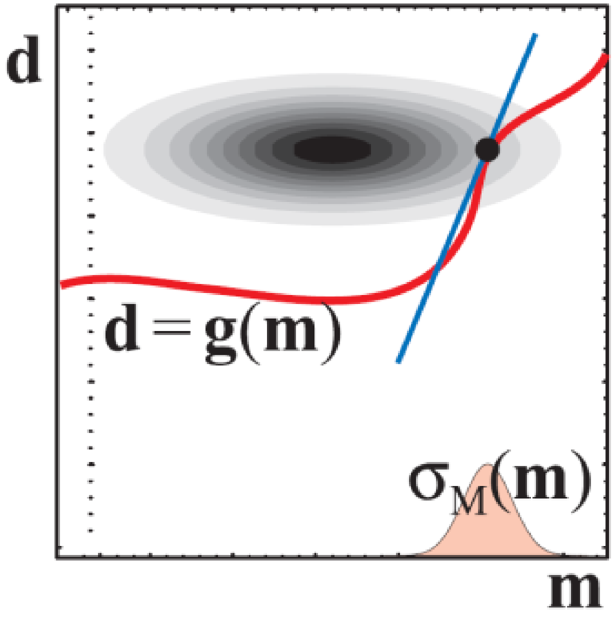
Must iterate!

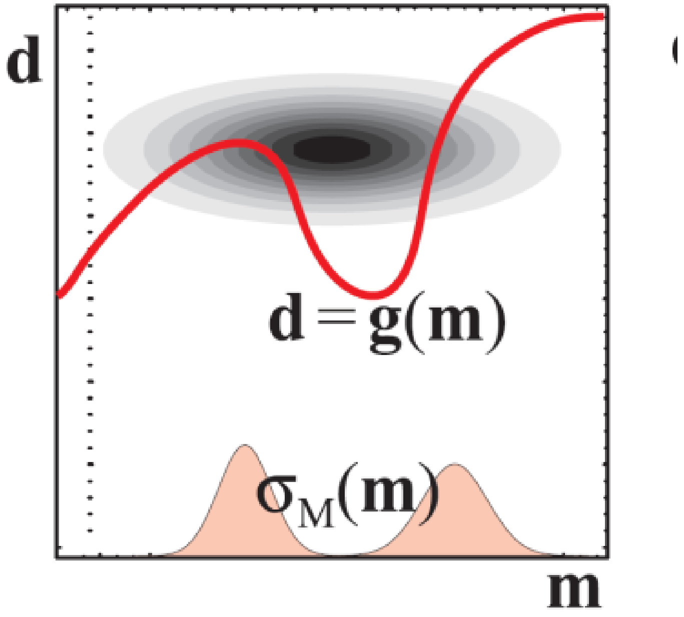
$\tilde{C}_M$ might be OK

Non-linear → multi-modal posterior on **m,** $\tilde{C}_M$ is woefully invalid!

Don’t use least-squares!

rjMcMC ☺





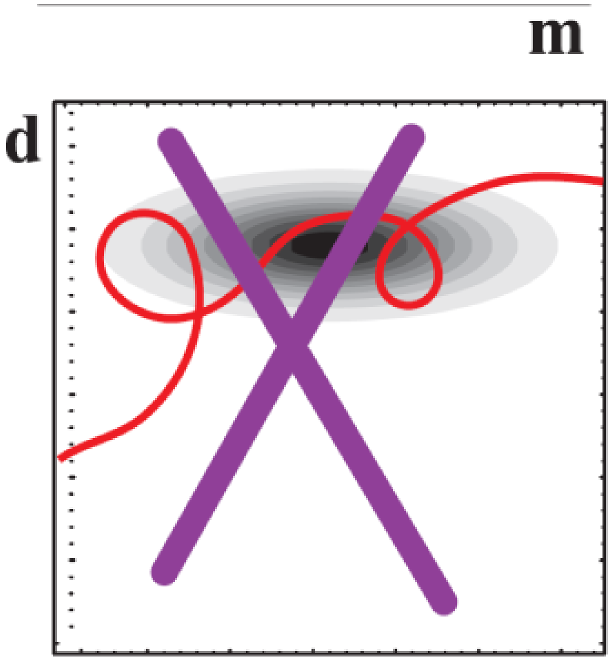
Tarantola, 2005

Trumpian / Brexistential

nothing will work! *Good luck*!

Give up and get a drink

(quote after Ved Lekic, 2018)



Tarantola, 2005

Model space search approaches

- When the relationship between data and model – i.e. g(m) – is non-linear, linear approaches can be inadequate, i.e. stuck in local minima and underestimating model error.

- Many current approaches focus on exploration of the model space
  Eliminate need for
  - → less biased estimates of model parameters

  - Some have flexible parameterization: "transdimensional"
  - Some estimate data uncertainty: "hierarchical"
  - Yield ensemble of models that can be analyzed to map uncertainty and non-uniqueness

Exploit vs. explore?

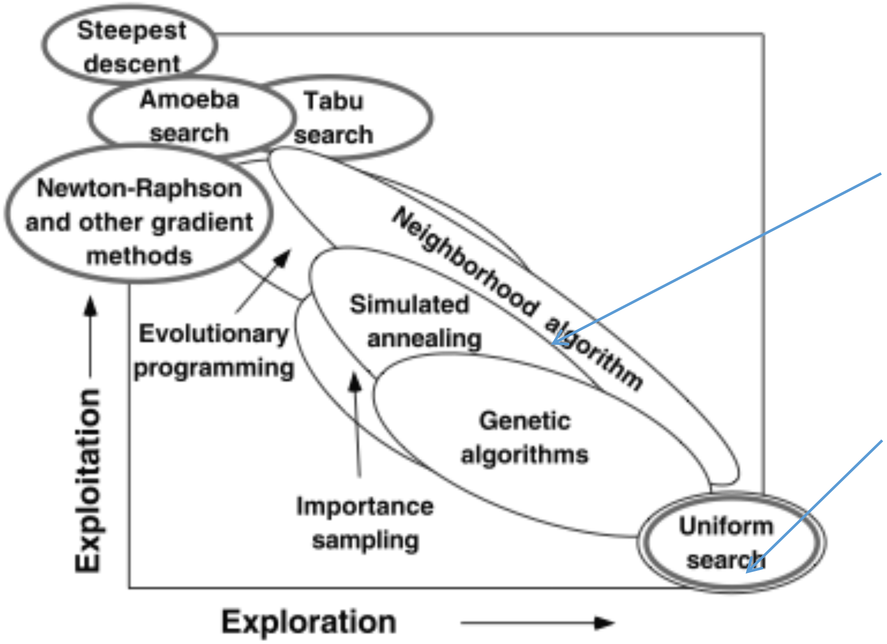


Figure 6. A schematic representation of various search/optimization algorithms in terms of the degrees to which they explore the parameter space and exploit information. Shaded borders indicate a deterministic (non–Monte Carlo) method. Uniform search includes the deterministic grid search.

Markov Chain Monte Carlo and various Bayesian approaches

Grid search, Monte Carlo search

From Sambridge, 2002

Press, 1968 Monte Carlo inversion

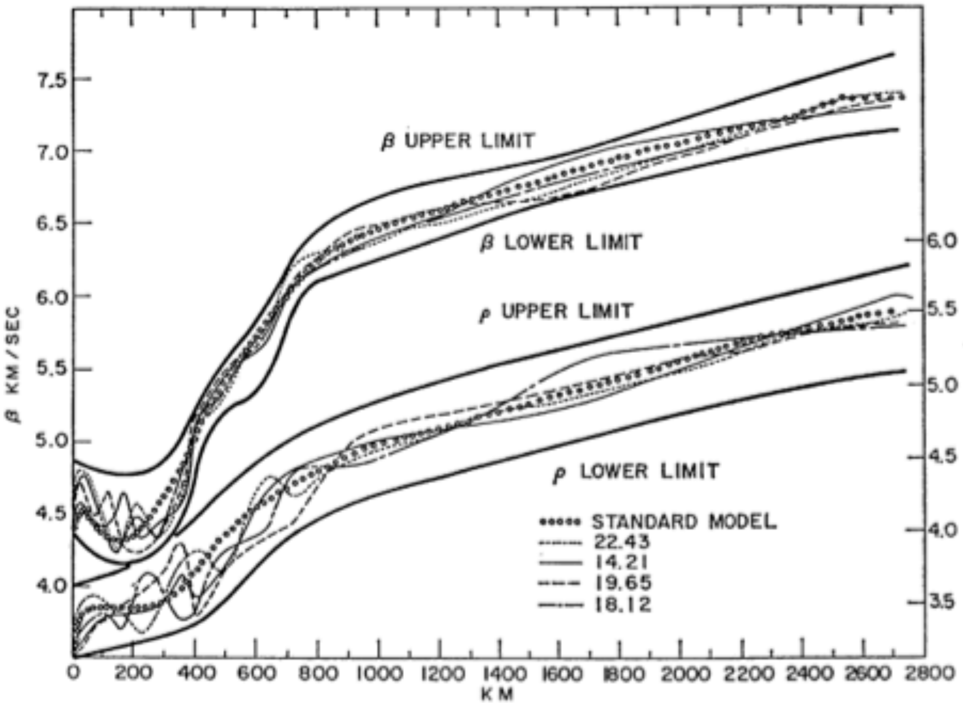


Figure 3. The six seismic and density Earth models that passed all tests shown in Figure 2 from the 5 million generated (from *Press* [1968]).

sample inverse problem

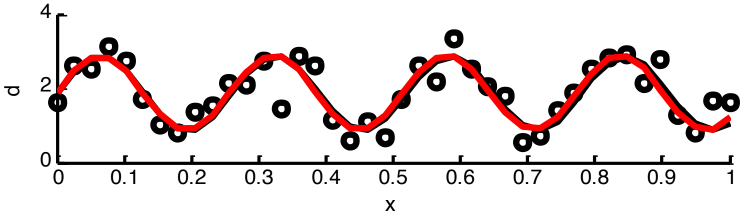
$$d_i(x_i) = sin(\omega_0 m_1 x_i) + m_1 m_2$$

with $\omega_0 = 20$

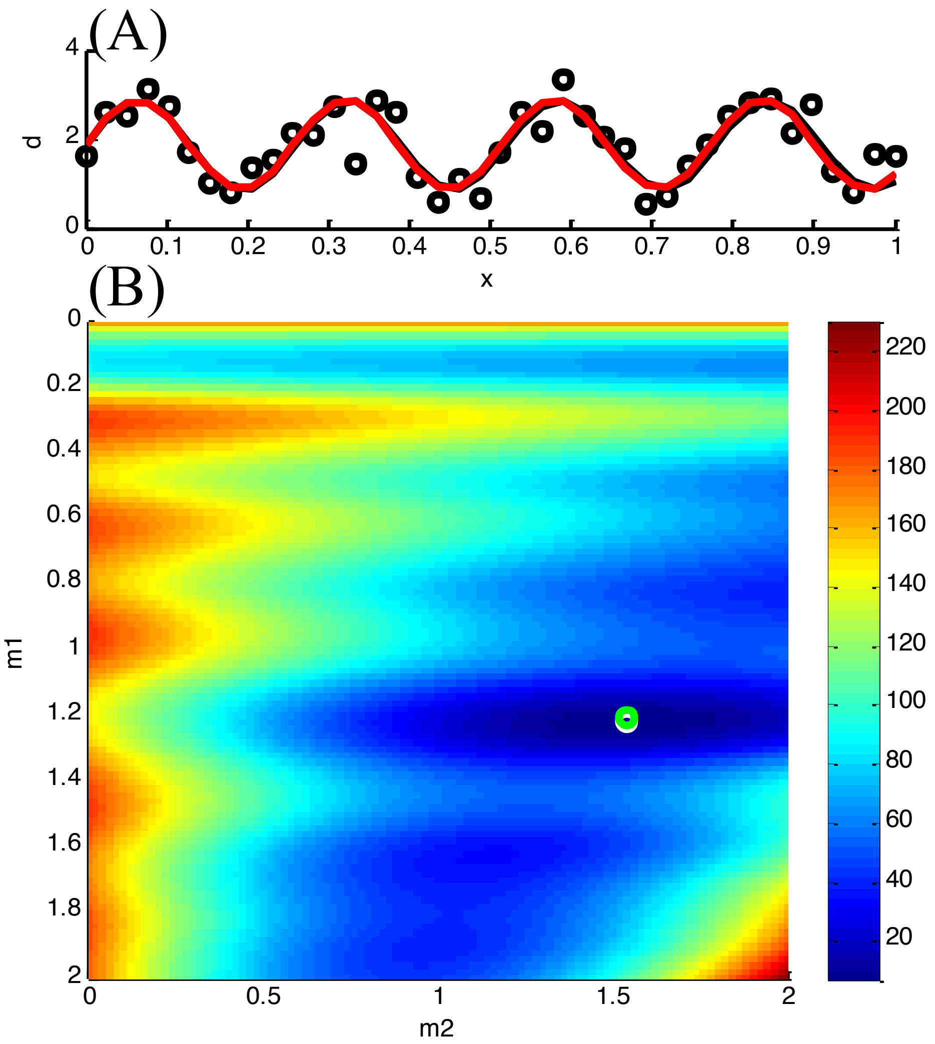
true solution

$m_1 = 1.21$, $m_2 = 1.54$

$N = 40$ noisy data



Grid search



Example from
Menke, 2012

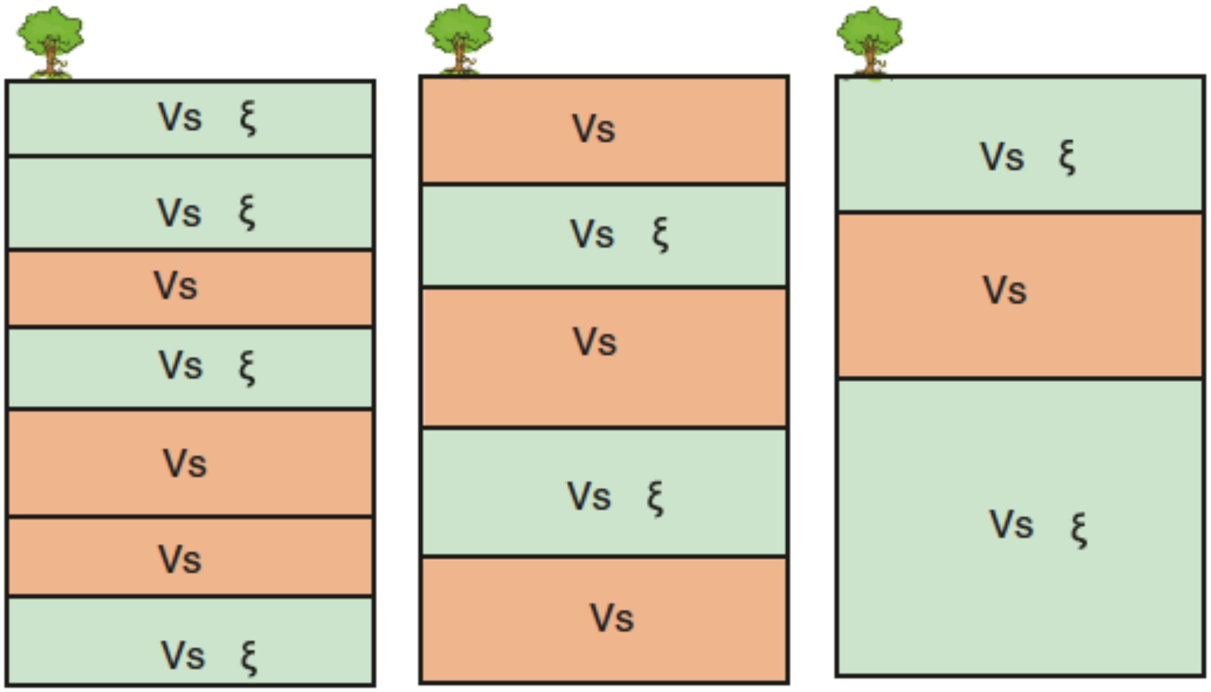
Markov Chain Monte Carlo (and other Bayesian approaches)

- Many derived from Metropolis-Hastings algorithm which uses randomly sampled models that are accepted or rejected based on the relative change in misfit from previous model

- End result is many (often millions) of models with sample density proportional to the probability of the various models

Trans-dimensional inversion:

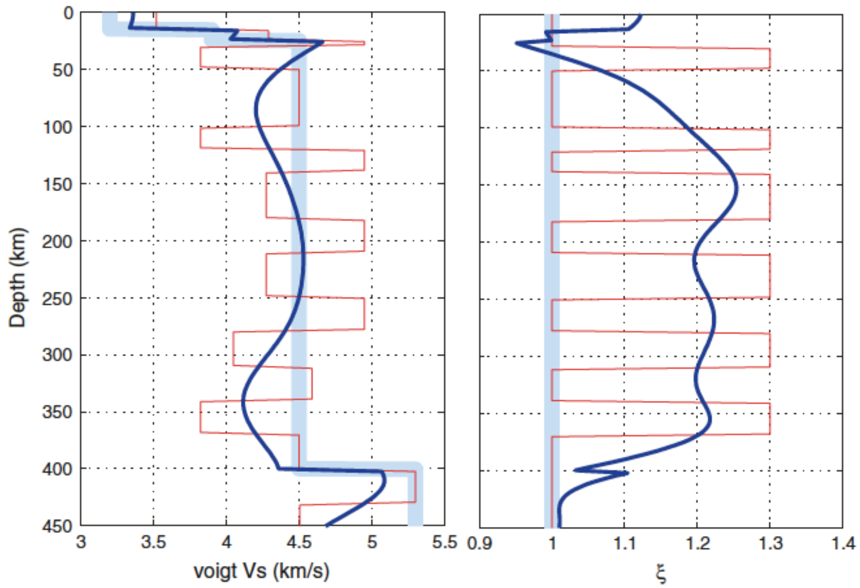
Number of parameters (here layers) is itself considered as an unknown



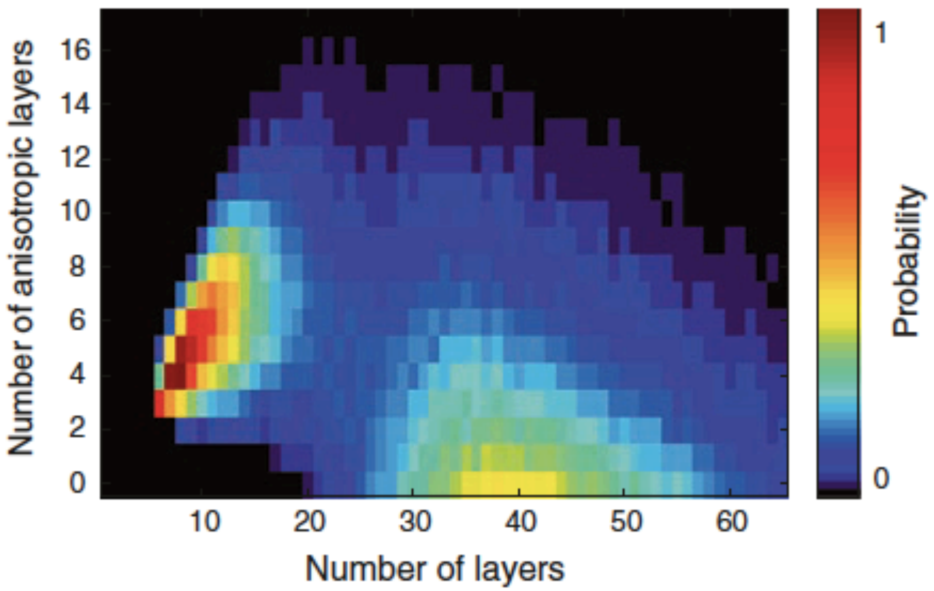
Bodin et al., 2014

Trans-dimensional inversion:

Residual Homogeneization



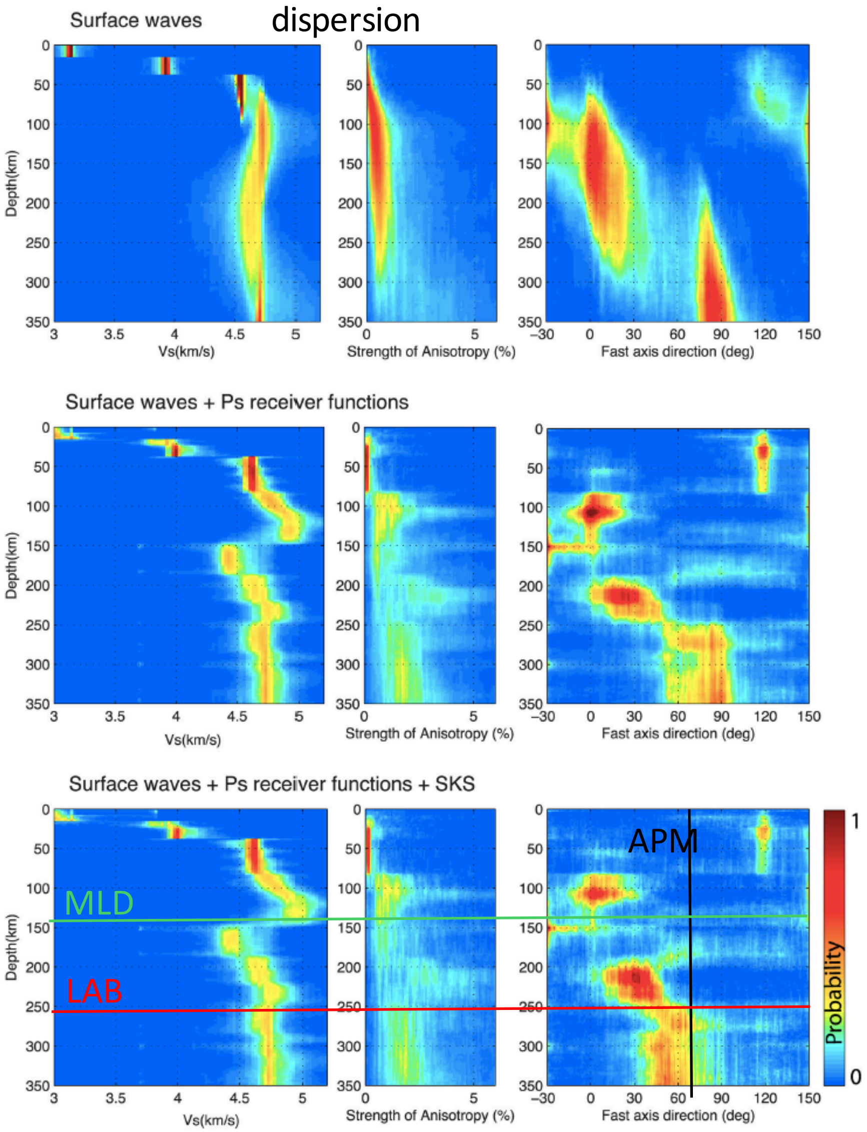
Trade-off between heterogeneity and radial anisotropy



Bodin et al., 2014

Transdimensional inversion
for Vs and azimuthal anisotropy

Station FFC (Canada)



Bodin et al., 2016, GJI

Evaluating an inverse model paper

- How well does the data sample the region being modeled? Is the data any good to begin with?
- Is the problem linear or not? Can it be linearized? Should it?
- What kind of theory are they using for the forward problem?
- What inverse technique are they using? Does it make sense for the problem?
- What's the model resolution and error? Did they explain what regularization choices they made and what effect it has on the model?

For further reference

- Textbooks
  - Gubbins, “Time Series Analysis and Inverse Theory for Geophysicists”, 2004
  - Menke, “Geophysical Data Analysis: Discrete Inverse Theory” 3rd ed., 2012
  - Parker, “Geophysical Inverse Theory”, 1994
  - Scales, Smith, and Treitel, “Introductory Geophysical Inverse Theory”, 2001
  - Tarantola, “Inverse Problem Theory and Methods for Model Parameter Estimation”, 2005