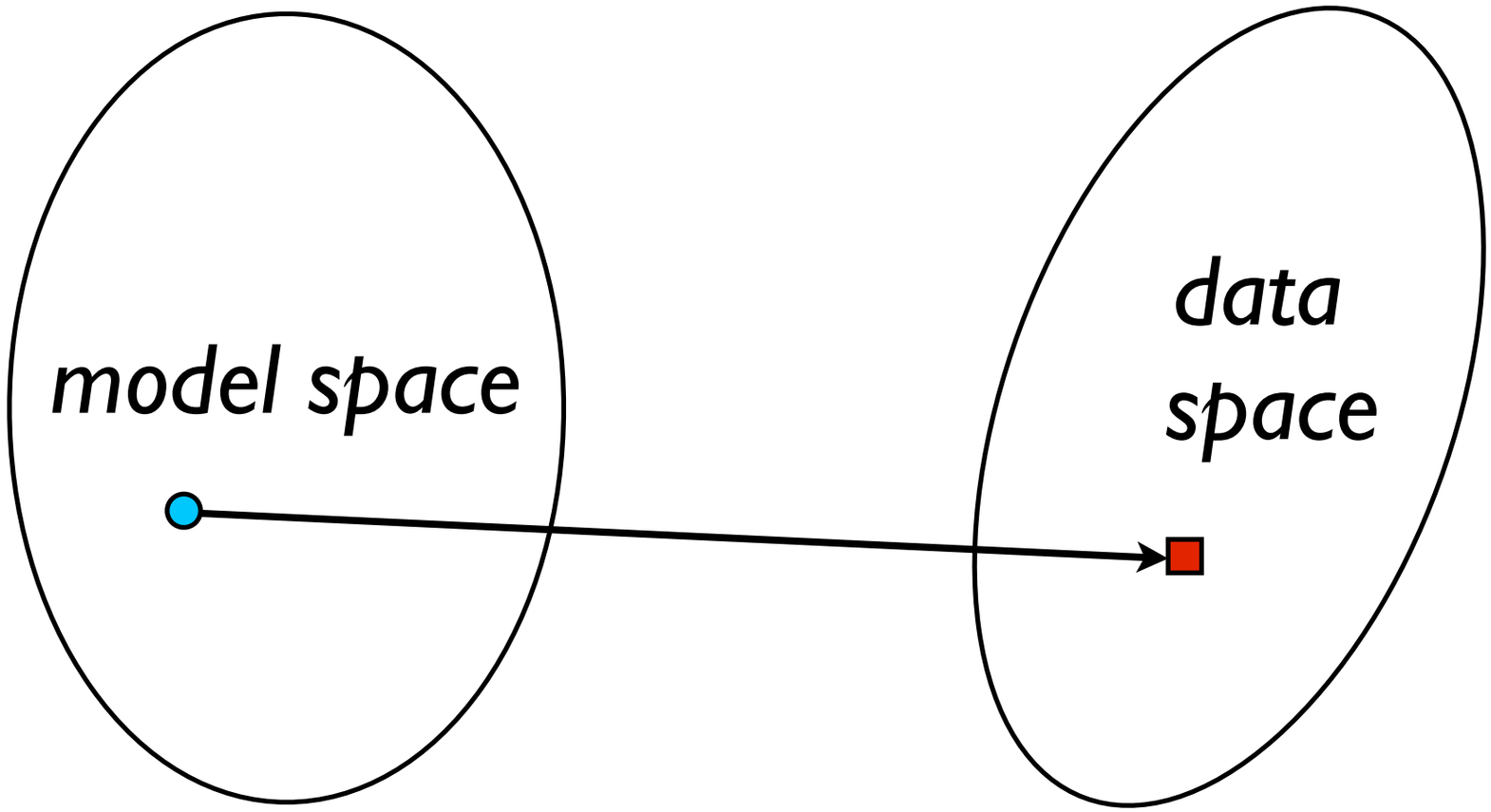
SIOG223A
A Very Brief Introduction to Geophysical Modeling and Inverse Theory.

Transitioning from parameter estimation to inverse theory, dealing with nonlinear problems and optimization.

Background reading see Bob Parker's optimization notes.
For much more on this topic take SIOG230.

Forward modeling:



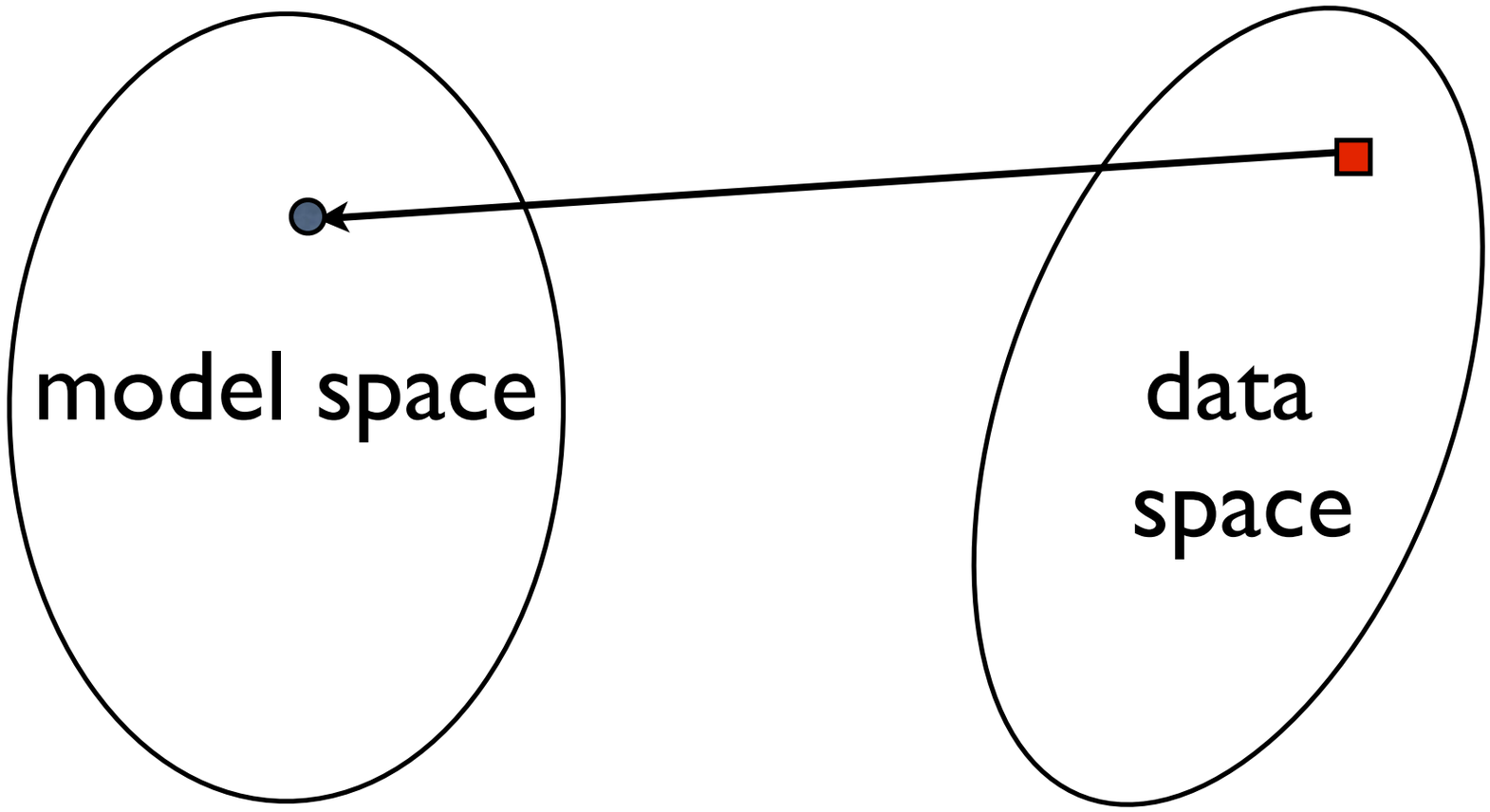
$$\hat{\mathbf{d}} = f(\mathbf{x}, \mathbf{m})$$ Some forward functional *f*

$$\mathbf{m} = (m_1, m_2, ..., m_N)$$ Model parameters (layers, blocks, ...)

$$\mathbf{x} = (x_1, x_2, x_3, ..., x_{kM})$$ Independent variables (freqs., locations, ...)

$$\hat{\mathbf{d}} = (\hat{d}_1, \hat{d}_2, \hat{d}_3, ..., \hat{d}_M)$$ Predicted data (gravity, magnetic, electric, ...)

Inverse modeling:



Given real (observed) data $\mathbf{d} = (d_1, d_2, d_3, ..., d_M)$

with errors $\sigma = (\sigma_1, \sigma_2, ..., \sigma_M)$

find an $\mathbf{m}$

There are several approaches to inversion:

Stochastic

Monte Carlo, Markov Chains
Genetic Algorithms
Simulated annealing, etc.
(Bayesian Searches)

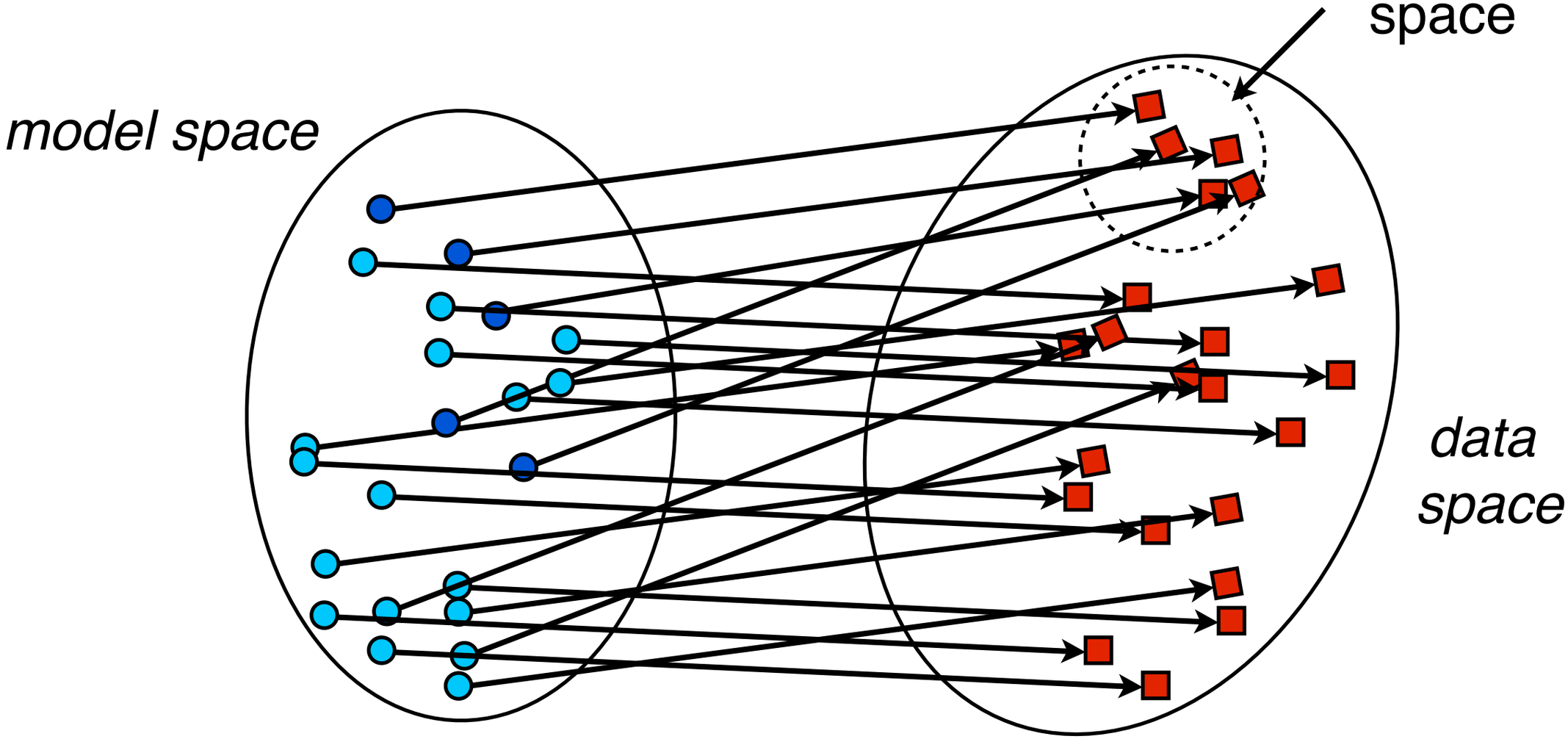
Deterministic

Newton Algorithms
Steepest descent
Conjugate Gradients
Quadratic (and Linear) Programming, etc.

Analytical

D+ (1D MT)
Bilayer (1D resistivity)
Ideal body theory in gravity and magnetism

Stochastic methods:

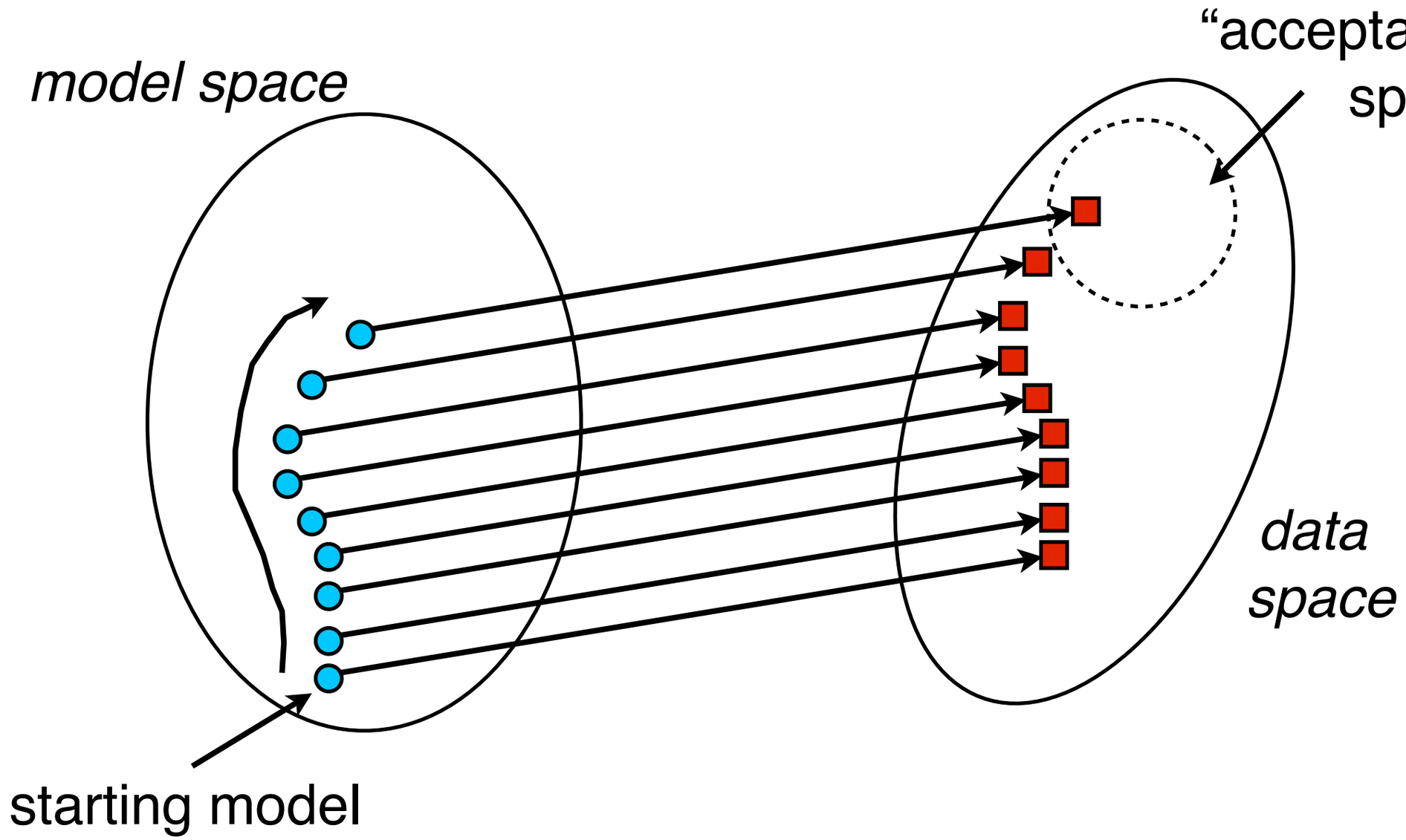


A useful approach, largely restricted to simple problems (because millions of models required), with most of the subtlety in model generation methods.

The advantages are that (i) only forward calculations are made and (ii) some probabilities can be obtained on model parameters. Best for sparsely parameterized models. One needs to be careful that bounds on explored model space don't unduly influence the outcome.

Deterministic

Newton Algorithms
Steepest descent
Conjugate Gradients



Analytical

e.g. D+ (1D MT)

and across the insulating interval $z_k < z < z_{k+1}$ we find

$$E_{k+1} = E_k + (z_{k+1} - z_k) D_k^+ \tag{52}$$

$$= E_k + (z_{k+1} - z_k) D_{k+1}^- \tag{53}$$

Define the admittance just above the k-th conductor in the usual way

$$C_k = -E_k / D_k^- \,. \tag{54}$$

Then by means of equations (50), (51) and (53) we can eliminate the E_k and $D_k^\pm$ as we did for uniform layers (although C_k is not continuous):

$$C_k = \frac{E_k}{-D_k^-} = \frac{E_k}{\mathrm{i}\omega\mu_0\tau_k E_k - D_k^+} = \frac{1}{\mathrm{i}\omega\mu_0\tau_k - D_k^+/E_k} \tag{55}$$

$$= \frac{1}{\mathrm{i}\omega\mu_0\tau_k - D_{k+1}^-/E_k} = \frac{1}{\mathrm{i}\omega\mu_0\tau_k - \dfrac{D_{k+1}^-}{E_{k+1} - (z_{k+1} - z_k) D_{k+1}^-}} \,. \tag{56}$$

Finally, dividing by D_{k+1}^- in the bottom tier we find the connection between the admittance at one level to the one above:

$$C_k = \frac{1}{\mathrm{i}\omega\mu_0\tau_k + \dfrac{1}{z_{k+1} - z_k + C_{k+1}}} \,. \tag{57}$$

We could solve (48) by recurring upwards in the familiar way, starting with $E(H) = 0 = C_{K+1}$, to get the value of $E(0)$ and hence of $C_1 = c(\omega)$. But now we do something different: we substitute repeatedly from the top, and we get a magnificent **continued fraction** for the admittance:

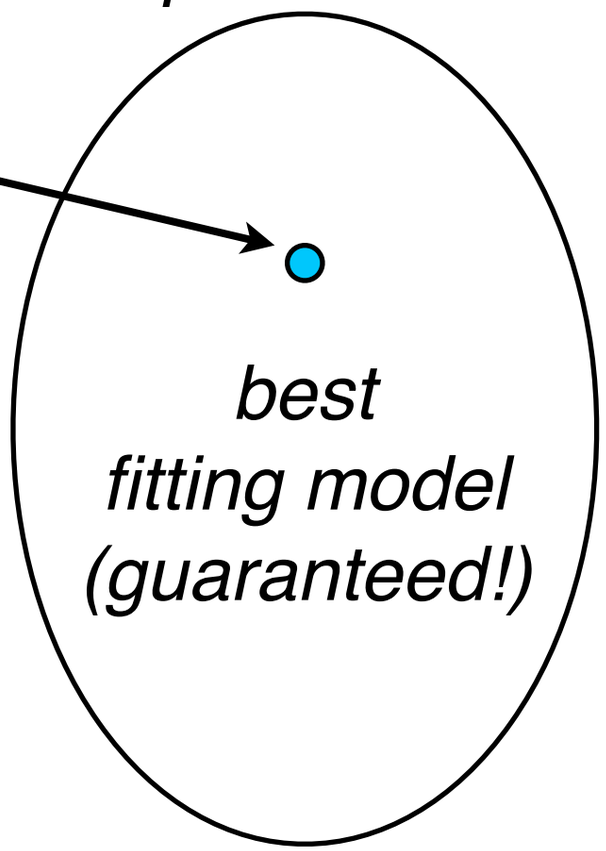
$$c(\omega) = z_1 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_1 + \cfrac{1}{z_2 - z_1 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_2 + \cfrac{1}{z_3 - z_2 + \cfrac{1}{\mathrm{i}\omega\mu_0\tau_3 + \cdots \cfrac{1}{H - z_K}}}}}} \,. \tag{58}$$

The initial z_1 allows us to put an insulator at $z = 0$, rather than a conducting sheet at the surface. While not exactly the same as the continued fractions described in the introduction, (58) can be rearranged by similar elementary algebra to be a *finite* partial fraction expansion:

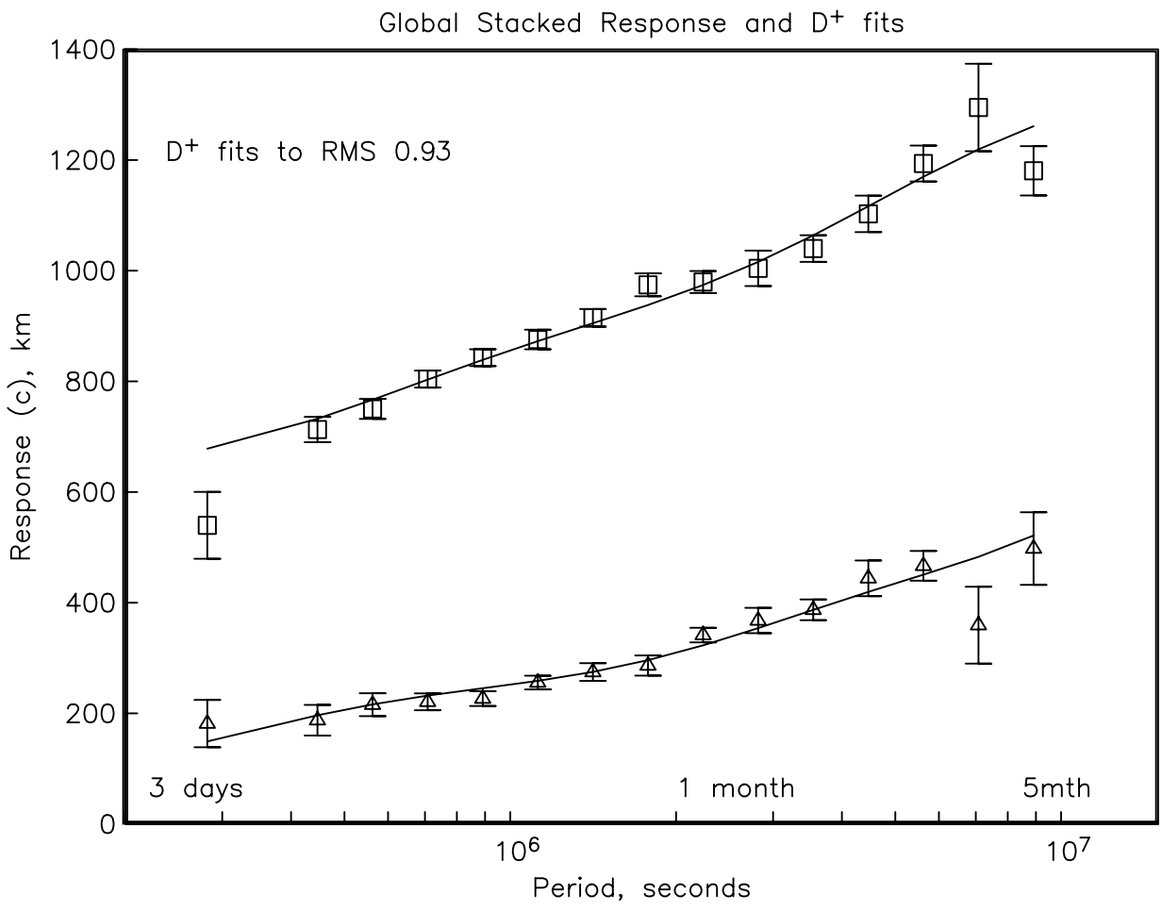
$$c(\omega) = z_1 + \sum_{k=1}^{K} \frac{\alpha_k}{\lambda_k + \mathrm{i}\omega} \,. \tag{59}$$

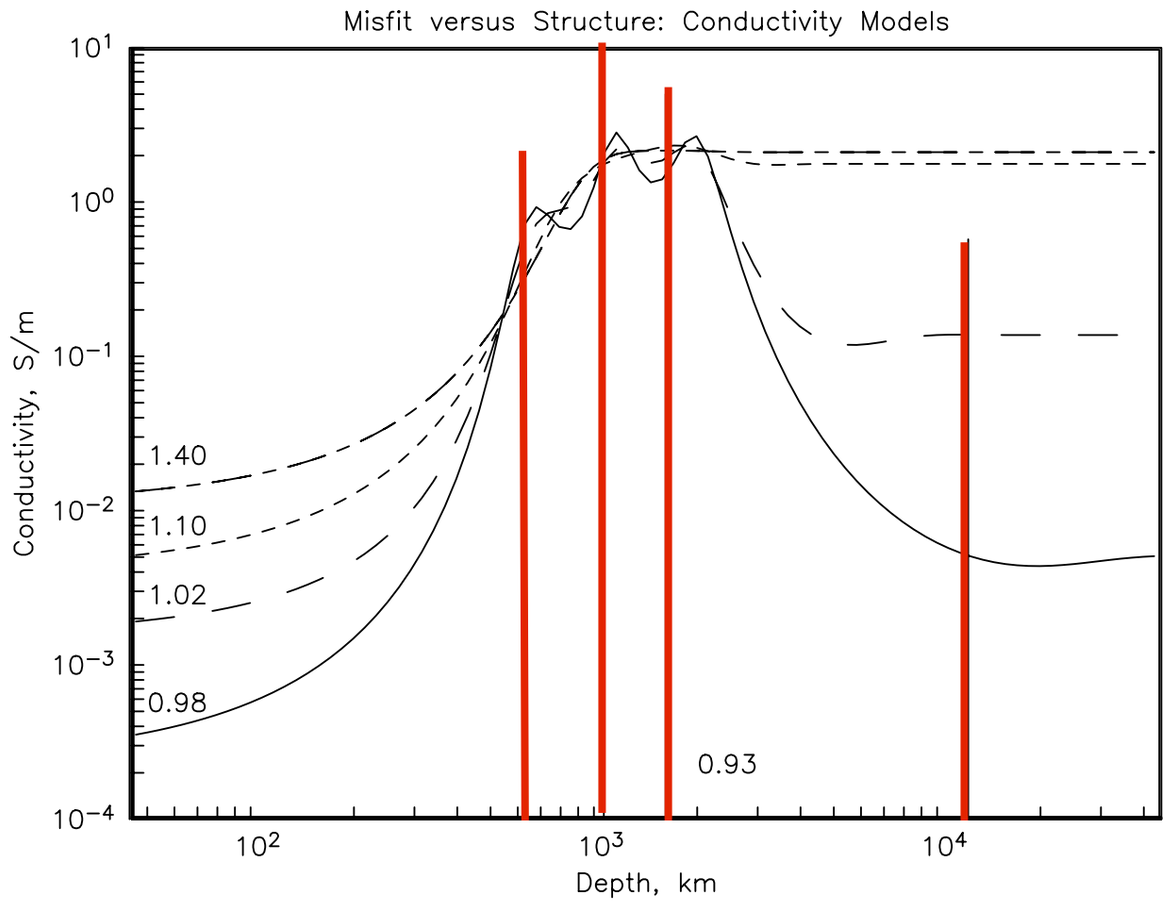
my data

model space



These solutions are guaranteed best fitting but usually pathological. For example, in 1D MT or global induction the least squares solutions are delta functions in electrical conductivity.





Existence and Uniqueness: Is there a solution to the inverse problem? Is there only one solution?

Finite noisy data for a linear problem (say, gravity)

An infinite number of solutions fit the data

Finite noisy data for a nonlinear problem

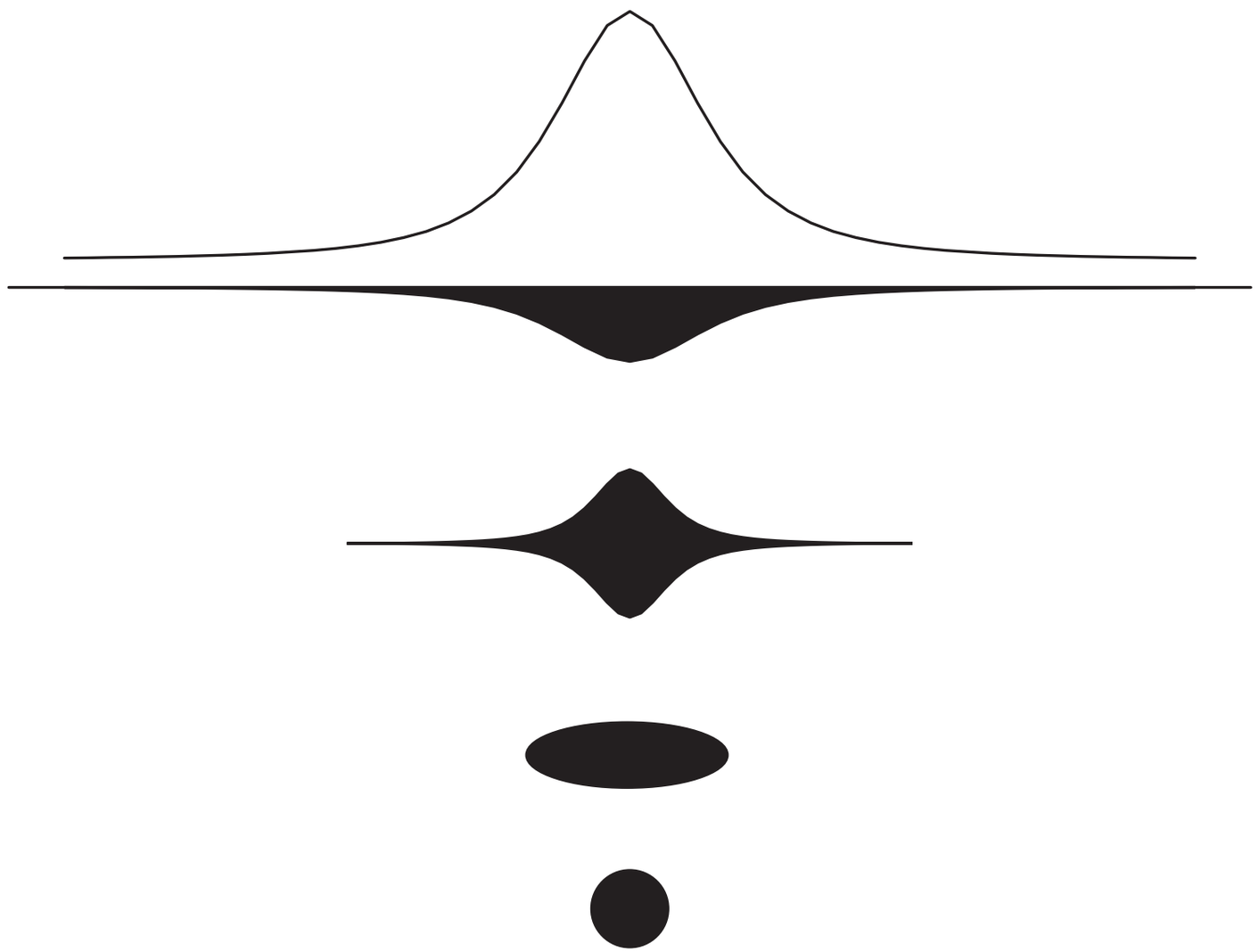
Either zero or an infinite number of solutions fit the data

Infinite exact data

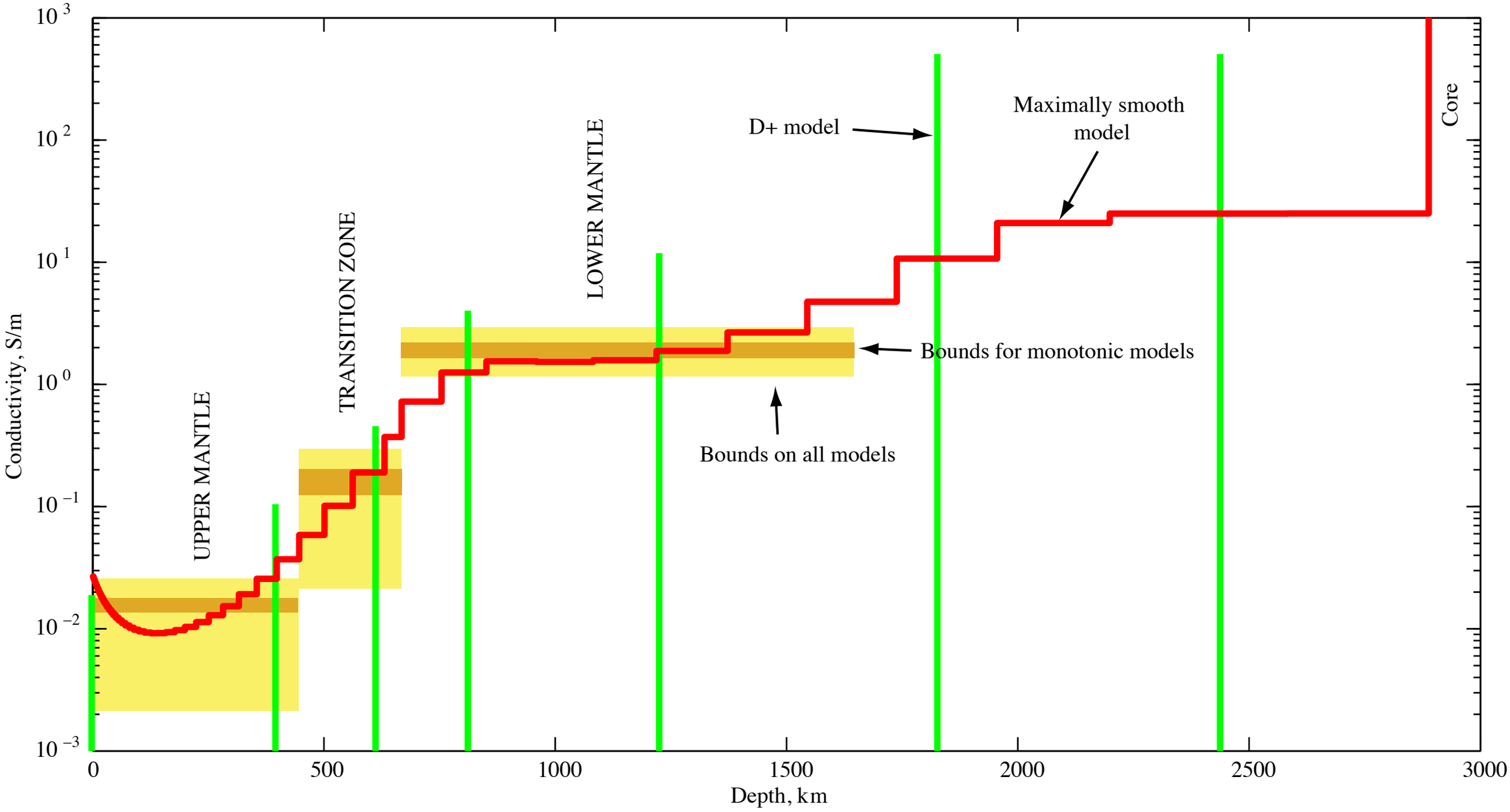
A unique solution has been shown to exist for a few cases.

As someone put it, there is no such thing as being a little bit non-unique.

What we have talked about so far is **model construction**. For a great many geophysicists this is what they think of when inversion is mentioned. More rigorous approaches try to obtain bounds on model properties - something that is true of all models. The classic example is total mass from gravity:

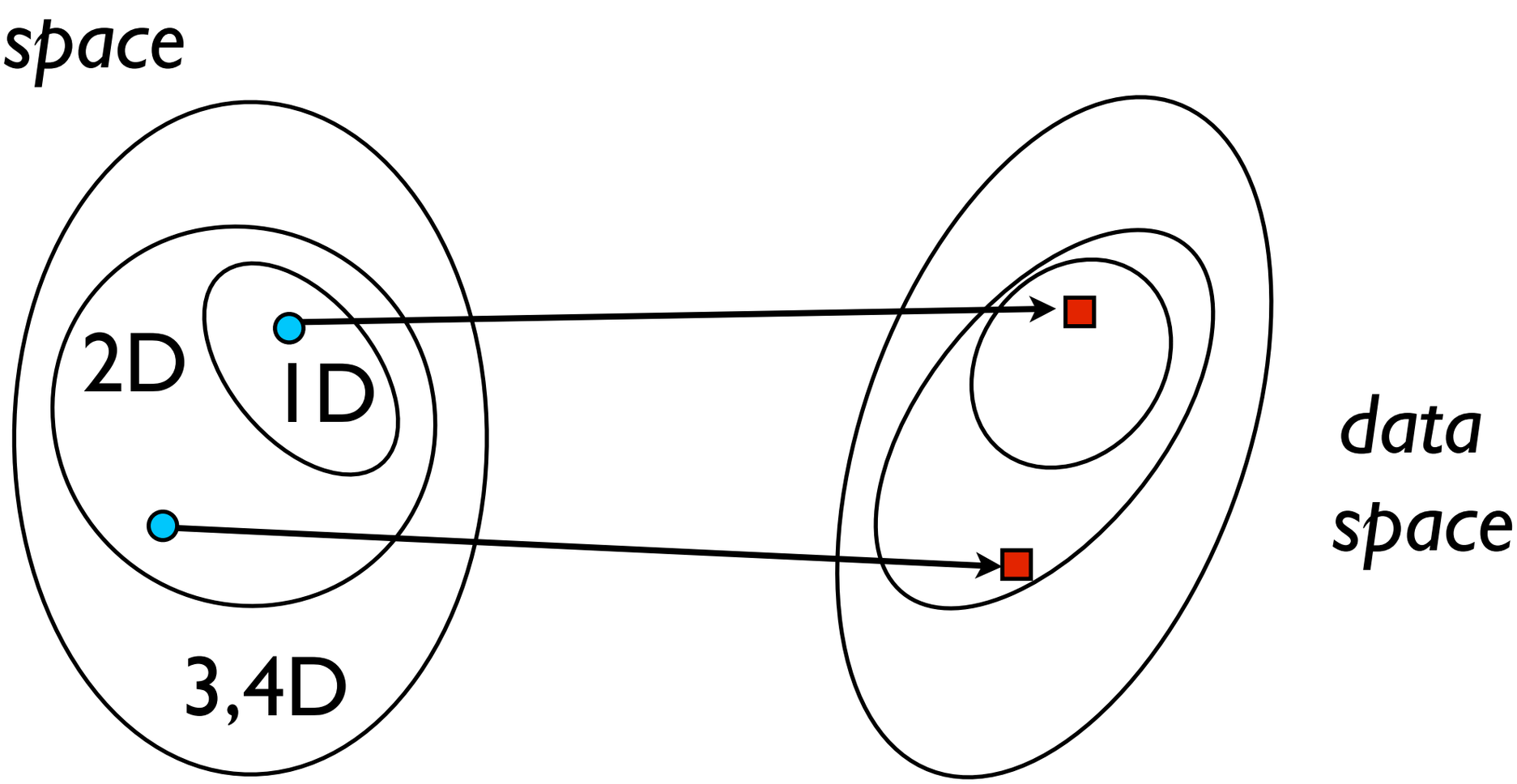


An example from global conductivity studies:



Where in model space you are is determined by your parameterization - this also determines where in data space you can be.

In non-linear geophysical problems, even forward modeling can involve a challenging computational effort.



One could ask the question: "*Can our mathematics ever completely describe nature?*".

The trite answer, of course, is "*No*". However, it is more useful to understand the nature of the limitations:

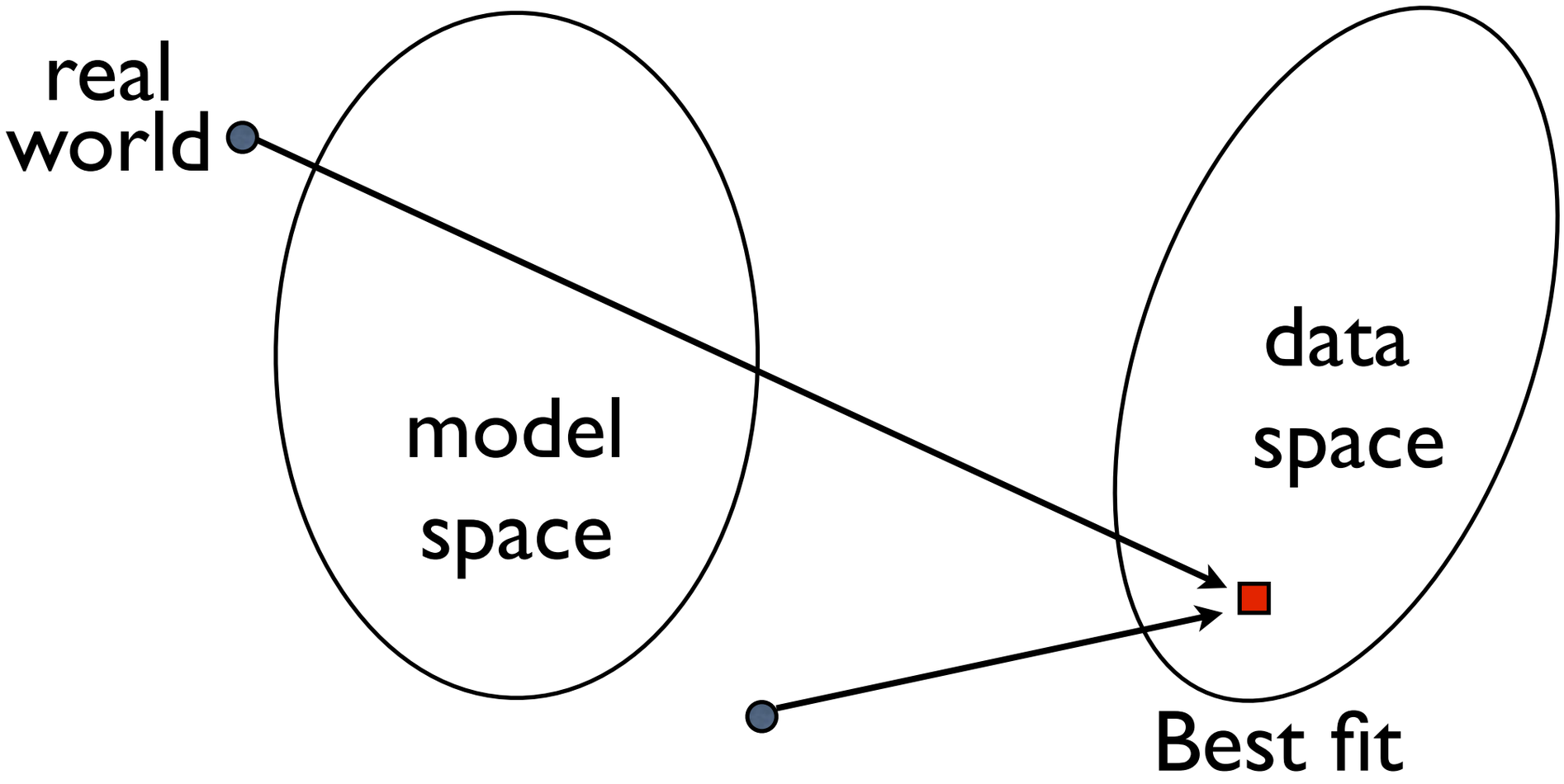
Are the physics sufficient (e.g. scalar versus anisotropic models)?

Is the forward computational machinery accurate?

Is the dimensionality of model space large enough? (1D, 2D, 3D, 4D)

Is the discretization fine enough and the model size big enough?

One can rarely afford to blindly ensure these are all achieved, so intelligence and understanding must be applied, perhaps by trial and error.



Even with your best efforts, the real world is unlikely to be captured by your model parameterization, and the best fitting model almost certainly won't be either. Understanding this can be important.

So what constitutes an
“adequate” fit to the data?

For noisy data (read: **all** data), we need a measure of how well a given model fits. Sum of squares is the venerable way:

$$\chi^2 = \sum_{i=1}^{M} \frac{1}{\sigma_i^2} \left[d_i - f(x_i, \mathbf{m}) \right]^2$$

or

$$\chi^2 = ||\mathbf{W}\mathbf{d} - \mathbf{W}\hat{\mathbf{d}}||^2$$

where $\mathbf{W}$ is a diagonal of reciprocal data errors

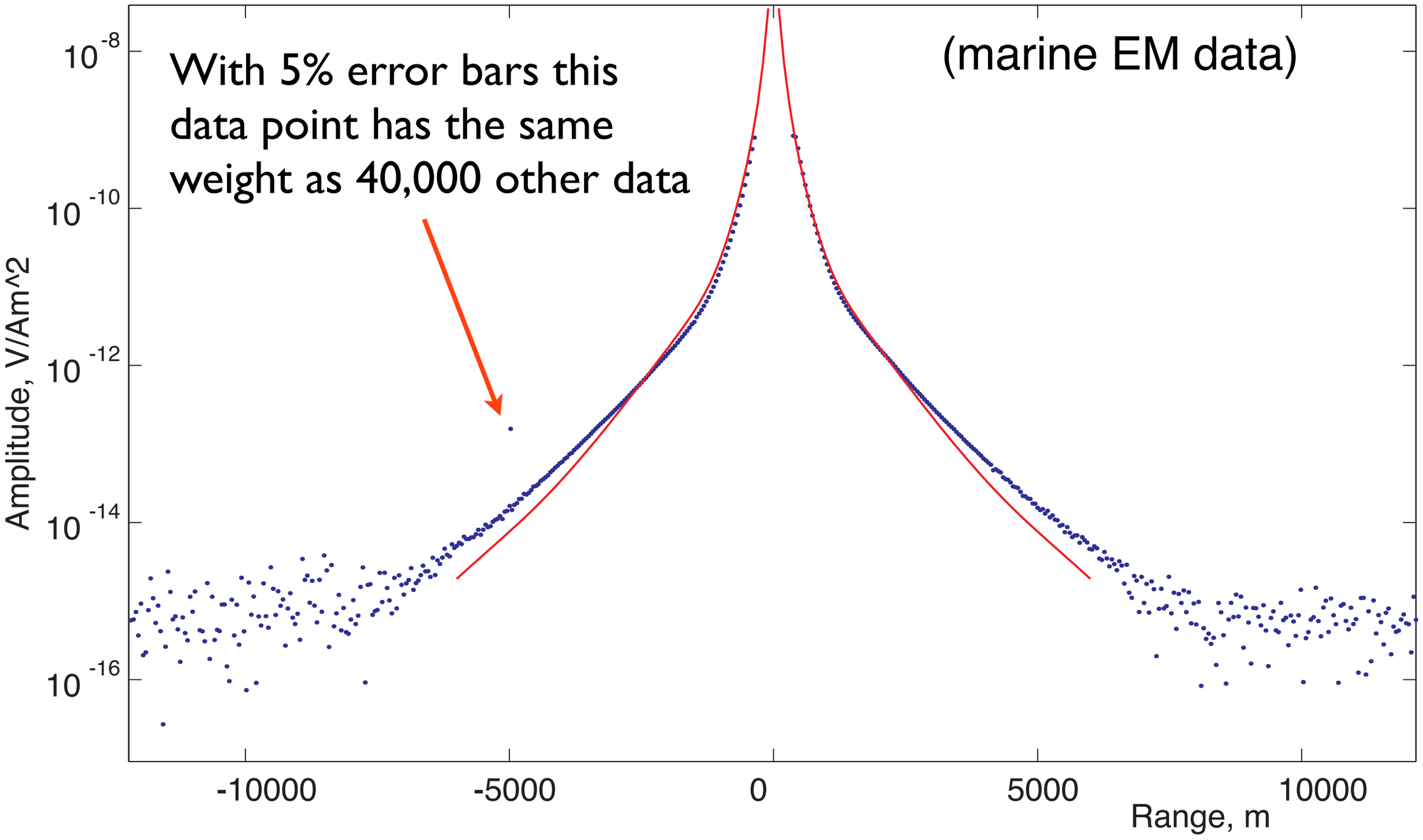
$$\mathbf{W} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, ..., 1/\sigma_M) \quad .$$

I like to remove the dependence on data number and use RMS:

$$\mathrm{RMS} = \sqrt{\chi^2/M} \quad .$$

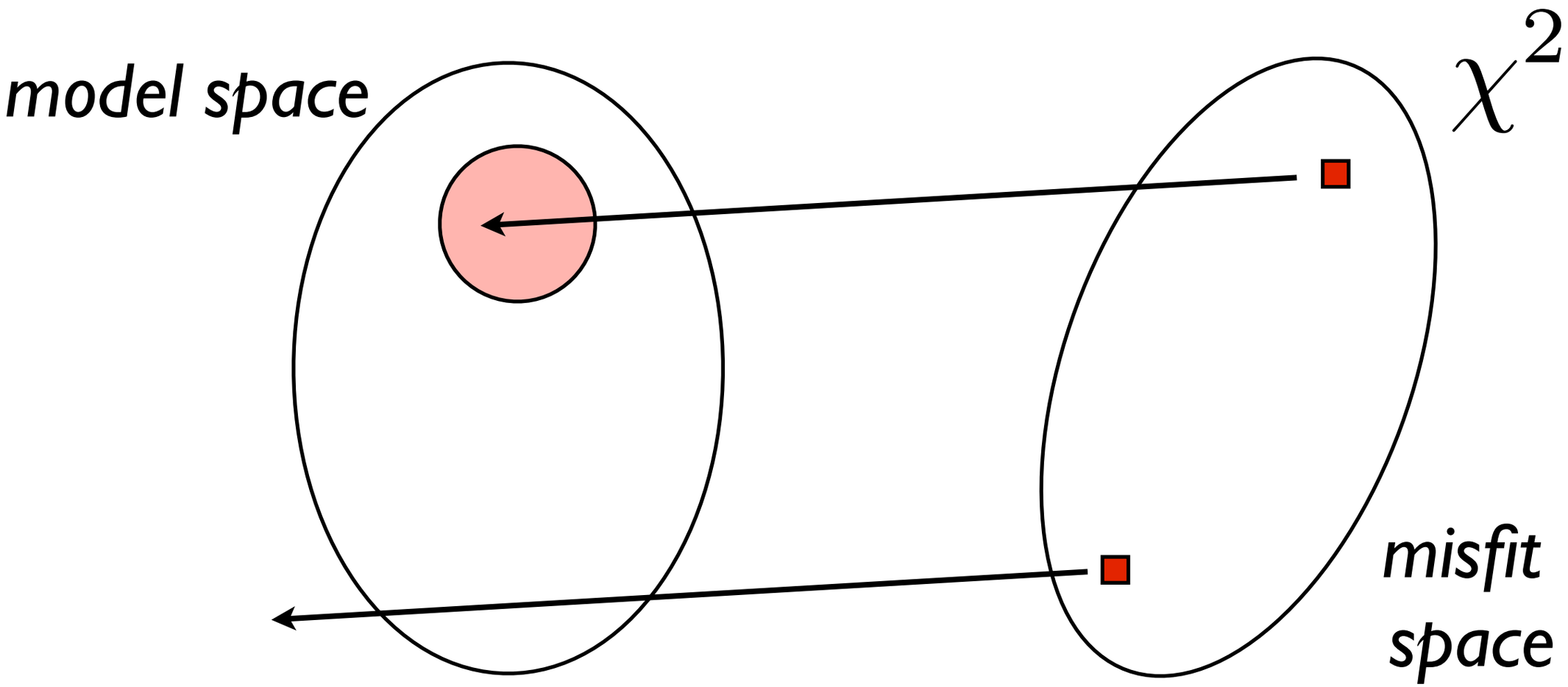
We could use other measures of fit, but the quadratic measure works with the mathematics of minimization, and for Gaussian errors LS is a maximum likelihood, minimum variance, unbiased solution. But...

... sum-squared misfit measures are unforgiving of outliers:



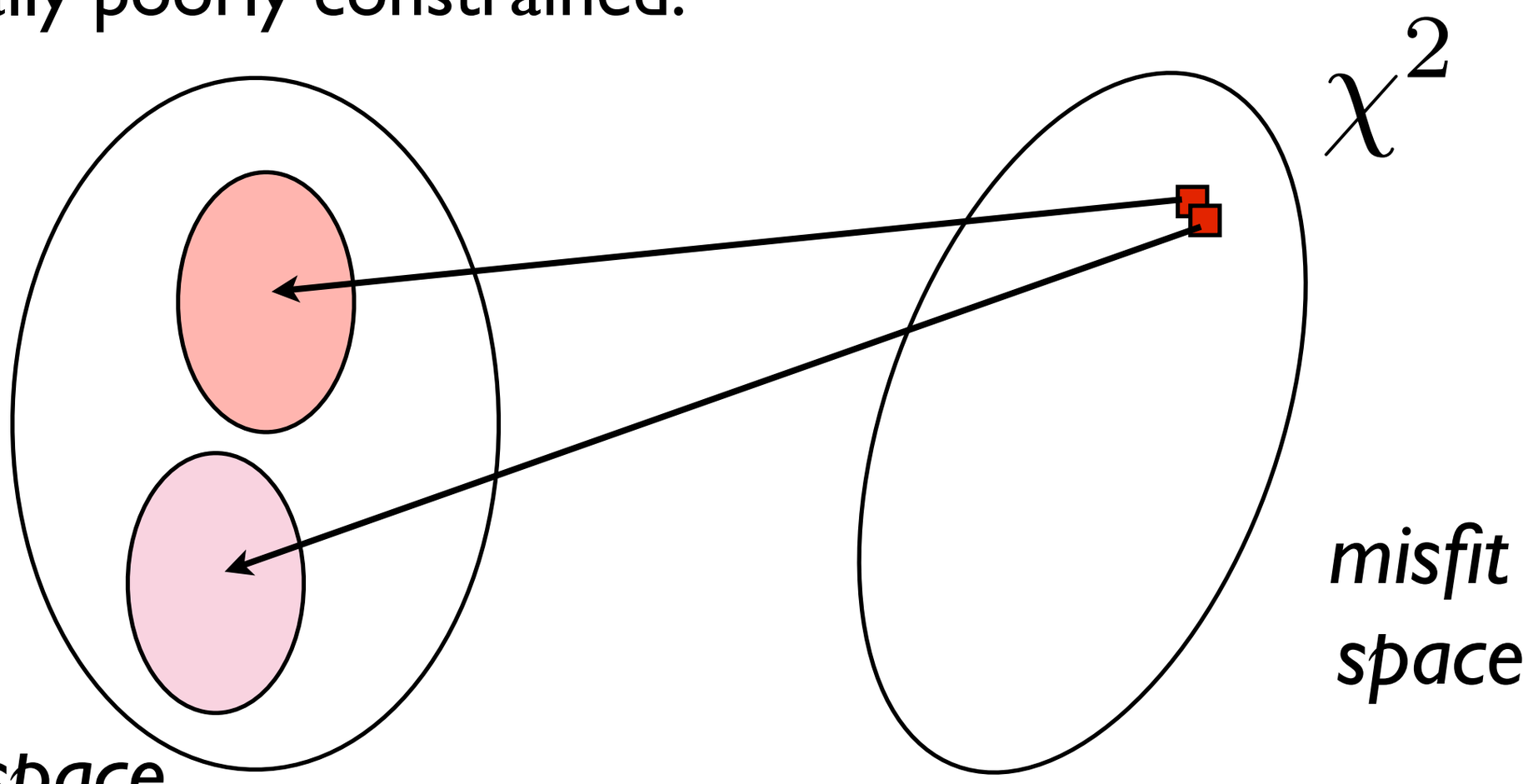
All through the inversion process you should monitor weighted residuals to ensure that there are no bad guys out there.

Geophysical inversion is non-unique:



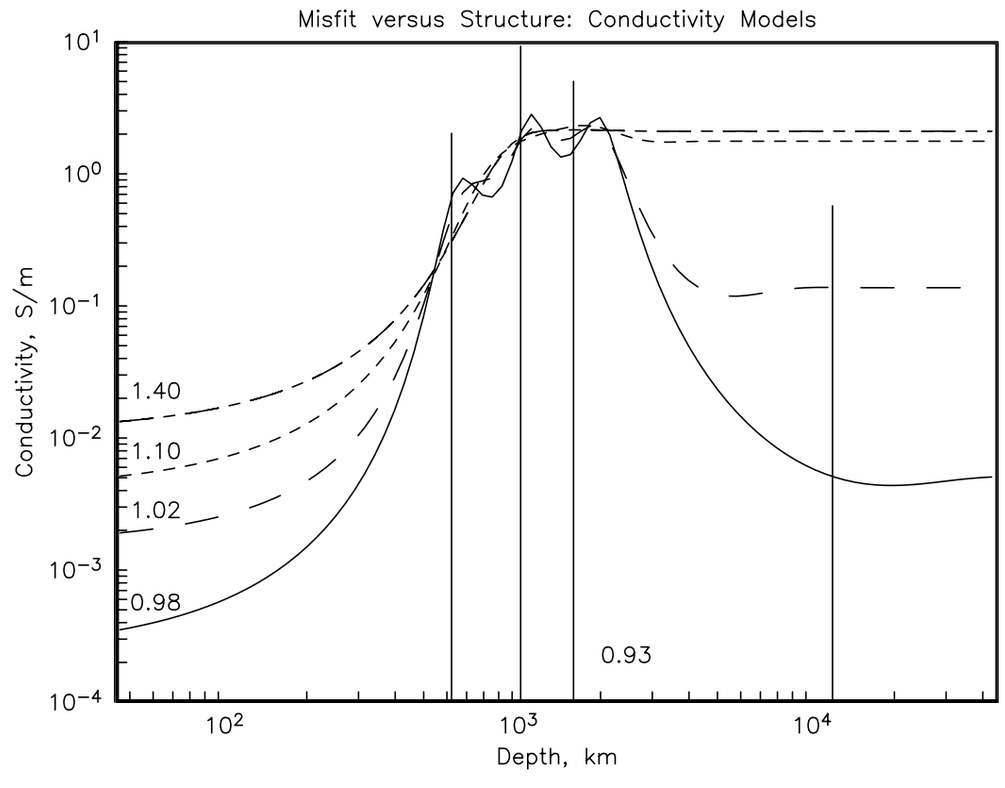
A single misfit will map into an infinite number of models (or none at all!).

and usually poorly constrained:



A small distance in χ^2 corresponds to a large distance in $\mathbf{m}$

(And don't forget: the minimum χ^2 is likely outside your model parameterization).



Speaking of model space parameterization:

$$\hat{\mathbf{d}} = f(\mathbf{x}, \mathbf{m}) \qquad \text{Some forward functional } f$$

$$\mathbf{m} = (m_1, m_2, ..., m_N) \qquad \text{Model parameters}$$

In the real world, N (model size) is infinite (even in 1D). How we proceed from here depends on whether N is small, moderately large, or infinite.

Small (sparse) parameterizations can be handled with parameterized inversions (e.g. Marquardt) or stochastic inversions. The concept of least squares fitting works because sparse models don't have the freedom to mimic the pathological true least squares solutions.

Infinite N requires real inverse theory.

A lot of geophysical model construction deals with moderately large N.

To invert non-linear forward problems we often linearize around a starting model:

$$\hat{\mathbf{d}} = f(\mathbf{m}_1) = f(\mathbf{m}_0 + \Delta\mathbf{m}) \approx f(\mathbf{m}_0) + \mathbf{J}\Delta\mathbf{m}$$

using a matrix of derivatives

$$J_{ij} = \frac{\partial f(x_i, \mathbf{m}_0)}{\partial m_j}$$

and a model perturbation

$$\Delta\mathbf{m} = \mathbf{m}_1 - \mathbf{m}_0 = (\delta m_1, \delta m_2, ..., \delta m_N)$$

Now our expression for χ^2

$$\chi^2 = ||\mathbf{W}\mathbf{d} - \mathbf{W}\hat{\mathbf{d}}||^2$$

Is then

$$\chi^2 \approx ||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m}_0) + \mathbf{W}\mathbf{J}\Delta\mathbf{m}||^2$$

For a least squares solution we solve in the usual way by differentiating and setting to zero to get a linear system:

$$\beta = \alpha \Delta \mathbf{m}$$

where

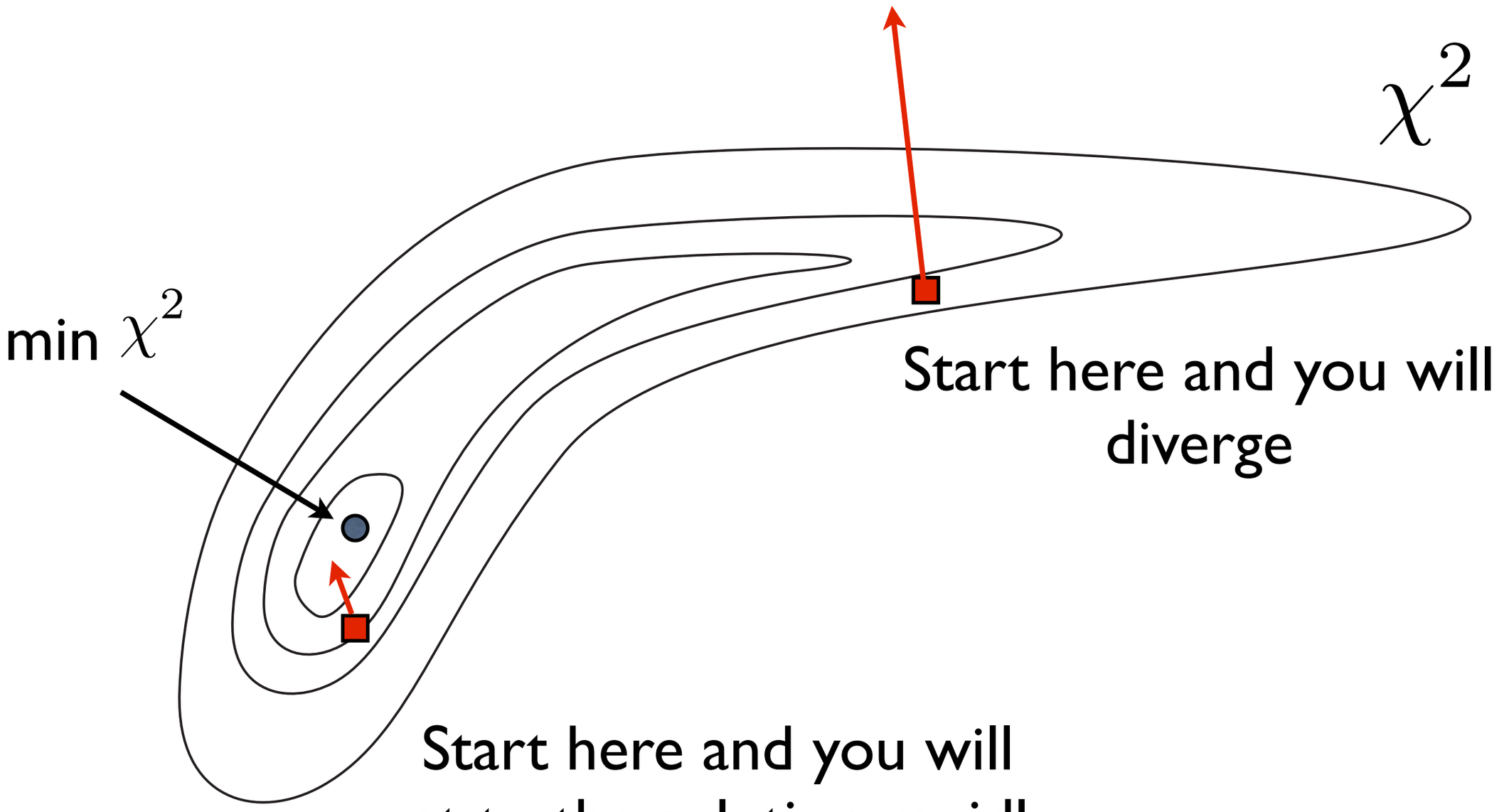
$$\beta = (\mathbf{WJ})^T \mathbf{W}(\mathbf{d} - f(\mathbf{m_0}))$$
$$\alpha = (\mathbf{WJ})^T \mathbf{WJ} \quad .$$

So, given a starting model $\mathbf{m}_0$ we can find an update $\Delta \mathbf{m}$ and iterate until we converge. (This is Gauss-Newton.)

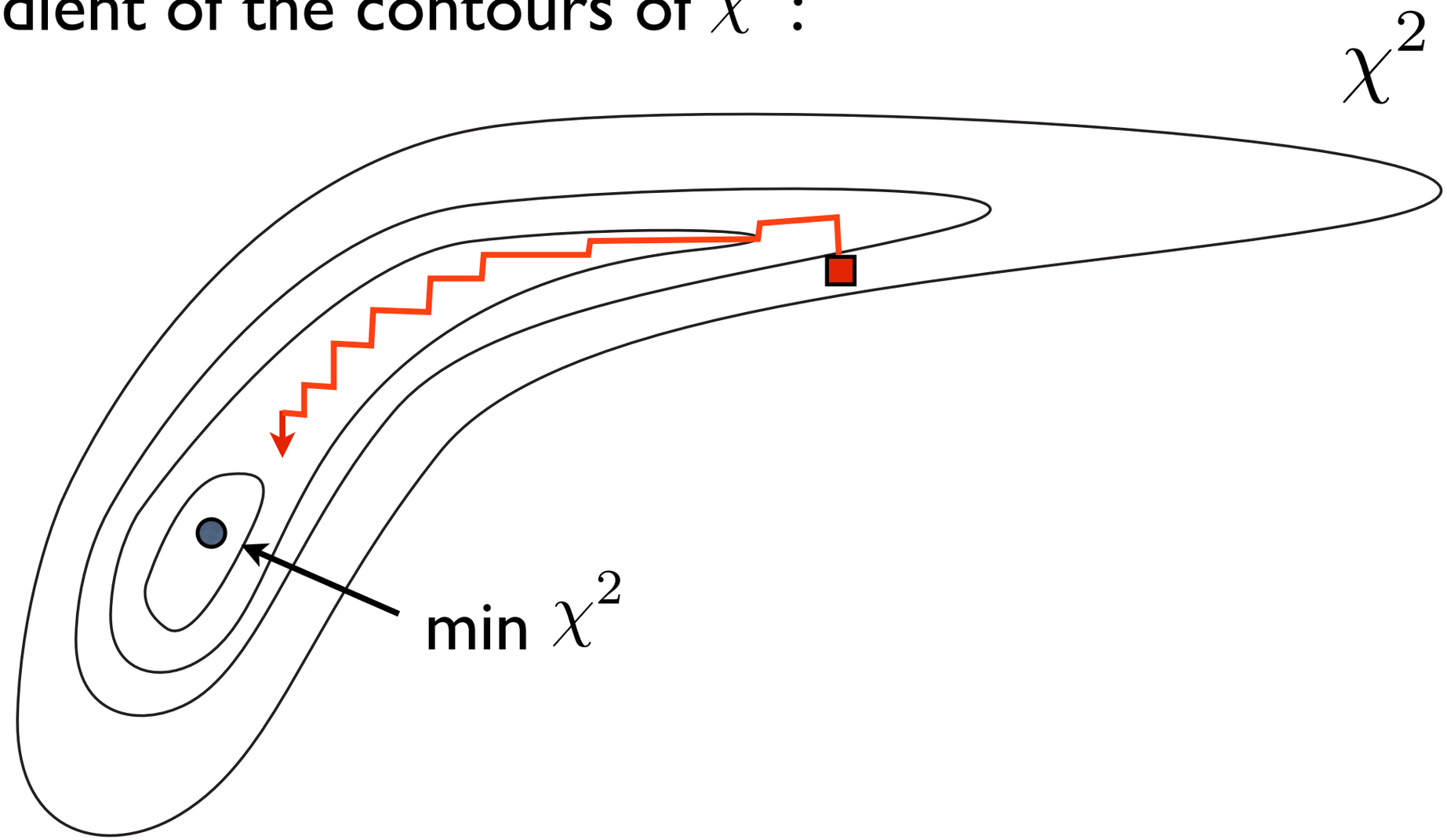
$$\Delta \mathbf{m} = \alpha^{-1} \beta$$

This can only work for small N (it isn't even defined for $N > M$). Even for very sparse parameterizations, it rarely works without modification.

You need to start "linearly close" to the solution for this to work. Long, thin, valleys in χ^2 space are common and present problems for Gauss-Newton methods.



Another approach is "steepest descent", in which you go down the gradient of the contours of χ^2:



These solutions are of the form

$$\Delta \mathbf{m} = \mu(\mathbf{WJ})^T[\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}))]$$

but the steps are always orthogonal and the step size is proportional to the slope, so this method stalls near the solution

The Marquardt method combines the steepest descent method and Newton method in one algorithm by modifying the curvature matrix:

$$\alpha_{jk} = \alpha_{jk}(1+\lambda) \quad \text{for } j = k$$

$$\alpha_{jk} = \alpha_{jk} \qquad \text{for } j \neq k \quad .$$

When λ is large, this reduces to steepest descent. When λ is small, it reduces to the Newton method. Starting with large λ and then reducing it close to the solution works very well.

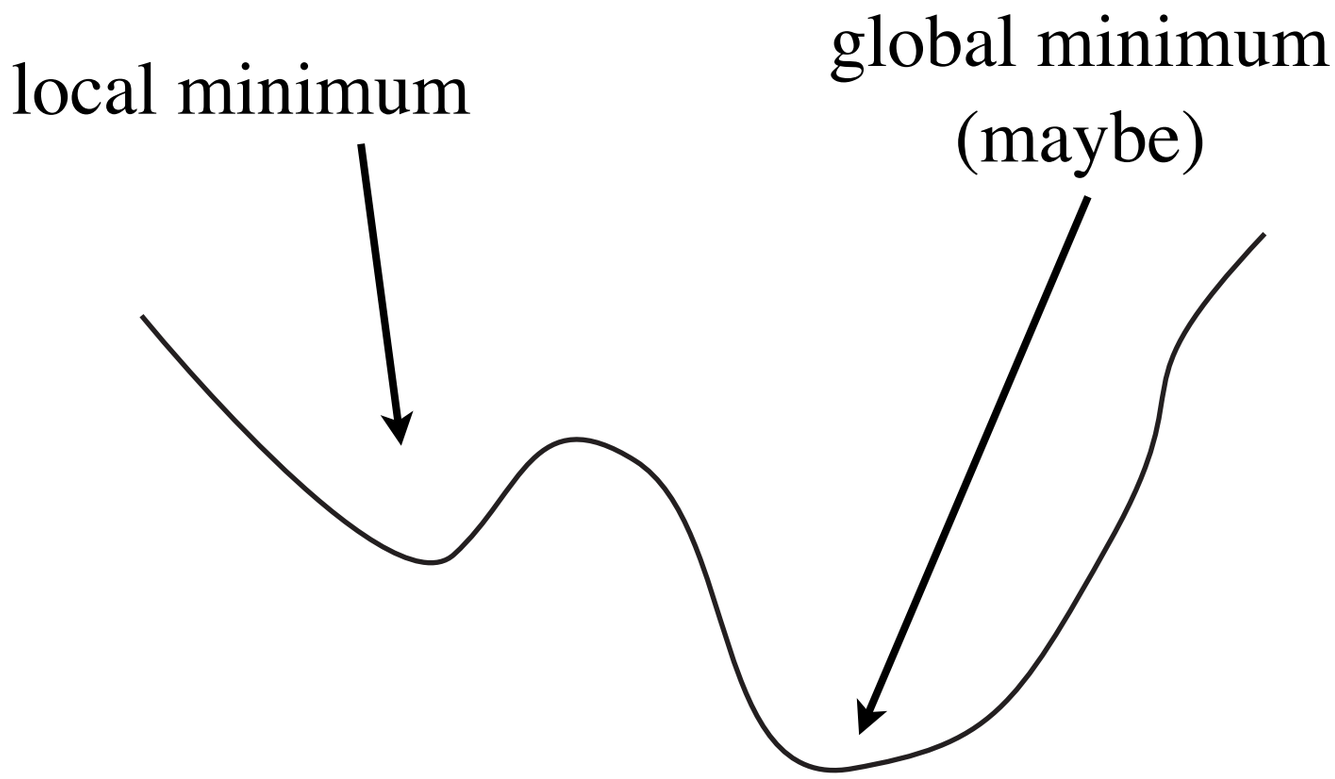
For problems that are naturally discretely parameterized, Marquardt is hard to beat. For sparse parameterizations of infinite dimensional models, the parameterization (e.g. number of layers chosen) has a big influence on the outcome.

Global versus local minima:

For nonlinear problems, there are no guarantees that Gauss-Newton will converge.

There are no guarantees that if it does converge the solution is a global one.

The solution might well depend on the starting model.



If you increase N too much, even with the Marquardt approach the solution goes unstable.

If N is big then the solutions become unstable, oscillatory, and generally useless (they are probably trying to converge to D+ type solutions).

Almost all inversion today incorporates some type of regularization, which minimizes some aspect of the model as well as fit to data:

$$U = ||\mathbf{R}\mathbf{m}_1||^2 + \mu^{-1}\left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m}_1)||^2\right)$$

where $\mathbf{Rm}$ is some measure of the model and μ is a trade-off parameter or Lagrange multiplier.

$$U = ||\mathbf{R}\mathbf{m}_1||^2 + \mu^{-1}\left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m}_1)||^2\right)$$

In 1D a typical $\mathbf{R}$ might be:

$$\mathbf{R}_1 = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ & & \ddots & & & \ddots & \\ & & & & & -1 & 1 \end{pmatrix}$$

m_1	-1						
m_2	+1	-1					
m_3		+1	-1				
m_4			+1	-1			
m_5				+1	-1		
m_6					+1	-1	
m_7						+1	-1
m_8							+1

which extracts a measure of slope. This stabilizes the inversion, creates a unique solution, and manufactures models with useful properties.

This is easily extended to 2D and 3D modelling.

$$U = ||\mathbf{R}\mathbf{m}_1||^2 + \mu^{-1}\left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m}_1)||^2\right)$$

When μ is small, model roughness is ignored and we try to fit the data. When μ is large, we smooth the model at the expense of data fit.

One approach is to choose μ and minimize U by least squares. There are various sets of machinery to do this (Newton, quasi-Newton, conjugate gradients, etc.). With many of these methods μ must be chosen by trial and error, increasing the computational burden and introducing some subjectivity.

Picking μ *a priori* is simply choosing how rough your model is compared to the data misfit. But, we've no idea how rough our model should be. However, we ought to have a decent idea of how well our data can be fit.

The Occam approach is to introduce some acceptable fit to the data (χ^2_*) and minimize:

$$U = ||\mathbf{R}\mathbf{m}_1||^2 + \mu^{-1}\left(||\mathbf{W}\mathbf{d} - \mathbf{W}f(\mathbf{m}_1)||^2 - \chi^2_*\right)$$

$$U = ||\mathbf{Rm}_1||^2 + \mu^{-1}\left(||\mathbf{Wd} - \mathbf{W}f(\mathbf{m}_1)||^2 - \chi_*^2\right)$$

Linearizing:

$$U = ||\mathbf{Rm}_1||^2 + \mu^{-1}\left(||\mathbf{Wd} - \mathbf{W}\left(f(\mathbf{m}_0) + \mathbf{J}(\mathbf{m}_1 - \mathbf{m}_0)\right)||^2 - \chi_*^2\right)$$

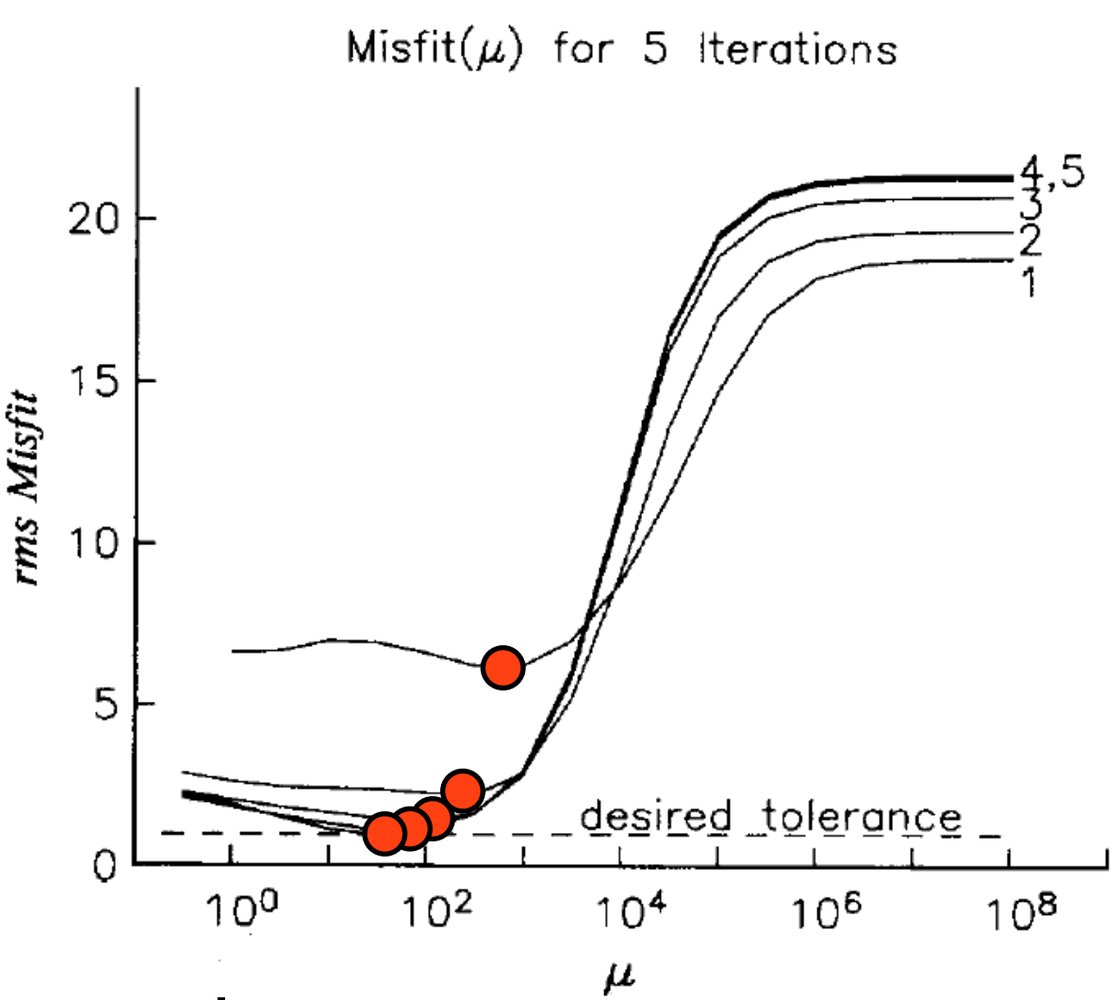
After differentiation and setting to zero we get

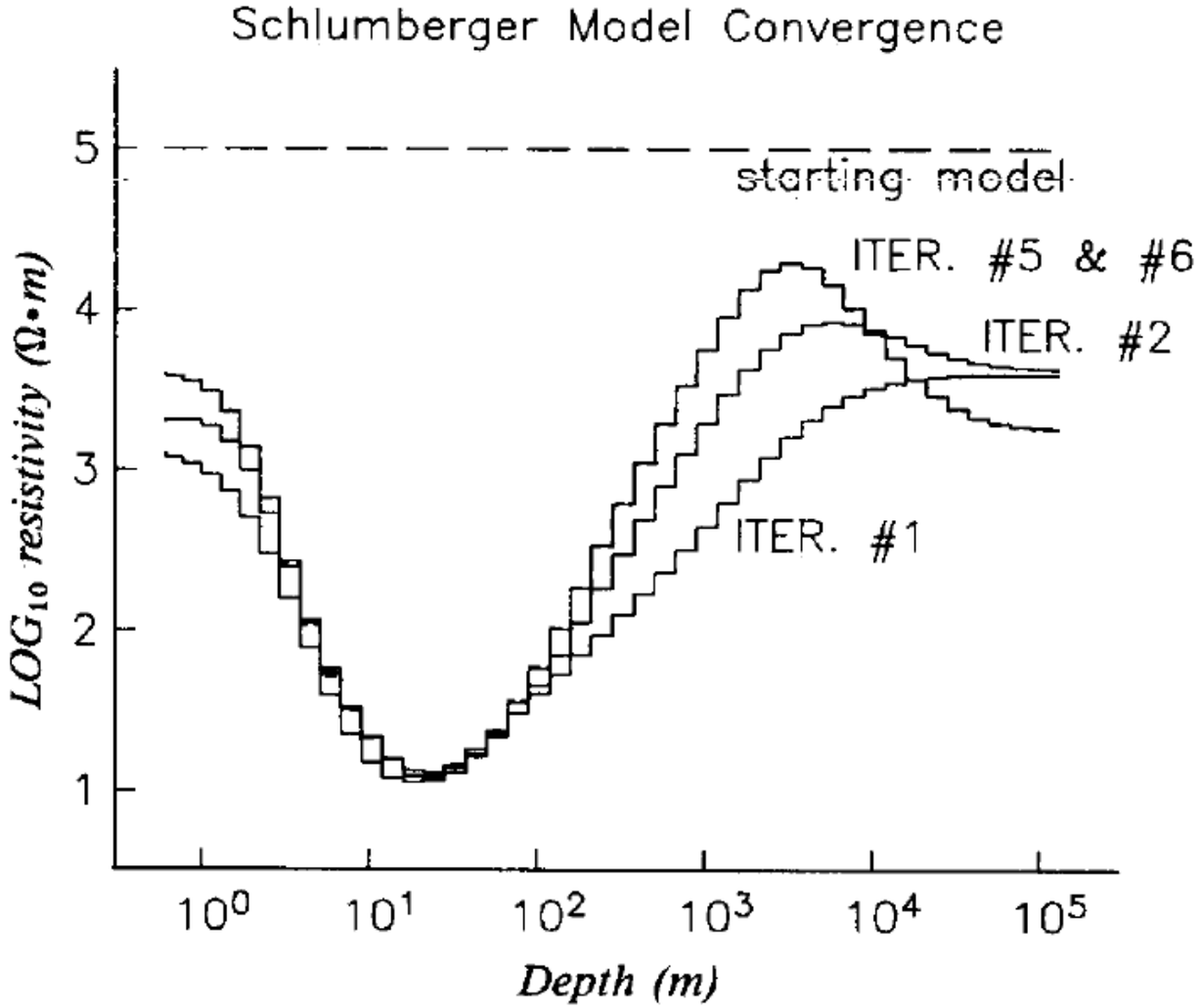
$$\mathbf{m}_1 = \left[\mu\mathbf{R}^T\mathbf{R} + (\mathbf{WJ})^T\mathbf{WJ}\right]^{-1}(\mathbf{WJ})^T\mathbf{W}(\mathbf{d} - f(\mathbf{m_0}) + \mathbf{Jm}_0)$$

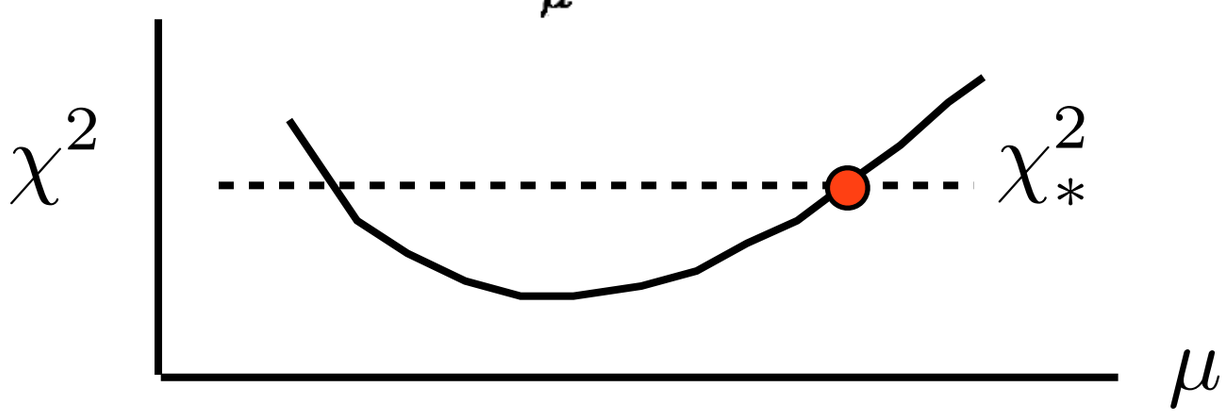
The only thing we need is to find the right value for μ.

(Note we are solving for the next model $\mathbf{m}_1$ directly instead of $\Delta\mathbf{m}$. Bob Parker calls these “leaping” and “creeping” algorithms.)

Occam finds μ by carrying out a line search to find the ideal value. Before χ^2_* is reached, we minimize χ^2. After χ^2_* is reached we choose the μ which gives us exactly χ^2_*.





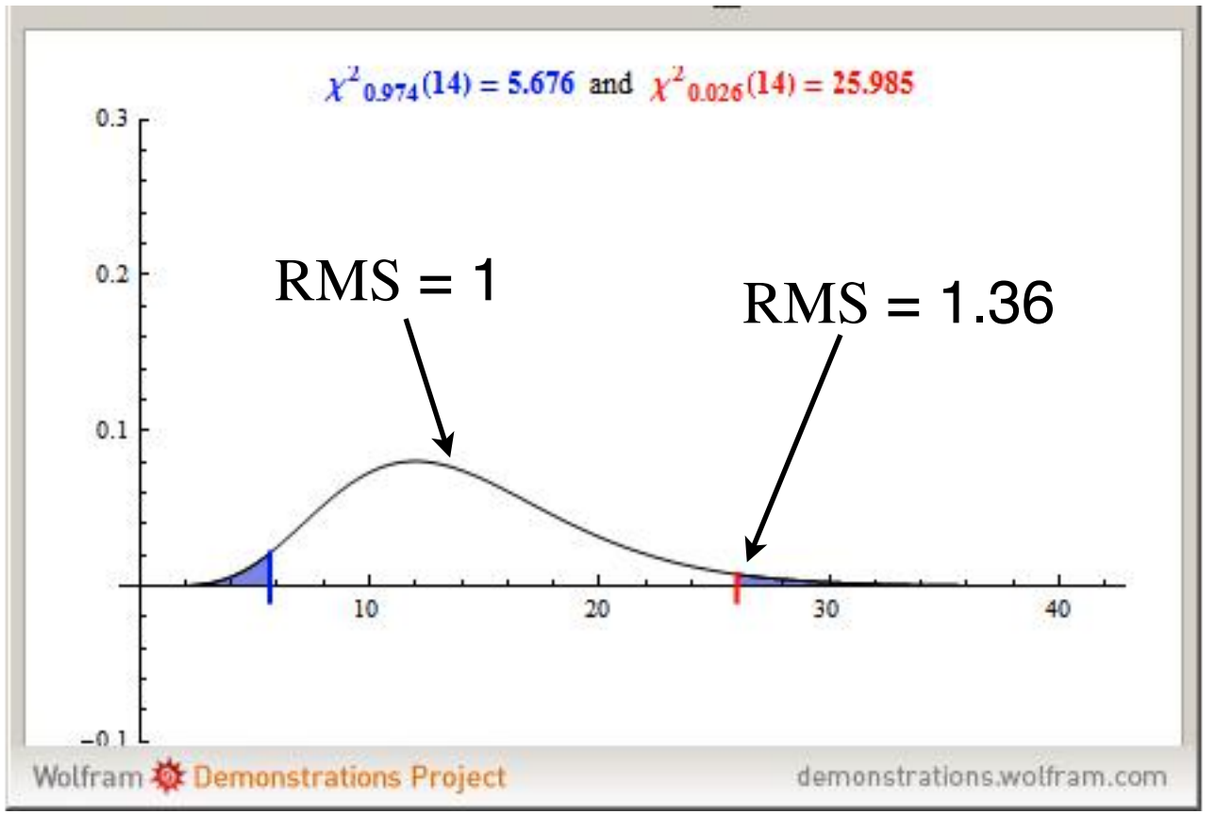


How to choose χ^2_* ?

For zero-mean, Gaussian, independent errors, the sum-square misfit

$$\chi^2 = ||\mathbf{W}\mathbf{d} - \mathbf{W}\hat{\mathbf{d}}||^2$$

is chi-squared distributed with M degrees of freedom. The expectation value is just M, which corresponds to an RMS of one, and so this could be a reasonable target misfit. Or, one could look up the 95% (or other) confidence interval for chi-squared M.



So, if our errors are well estimated and well behaved, this provides a statistical guideline for choosing χ^2_* .

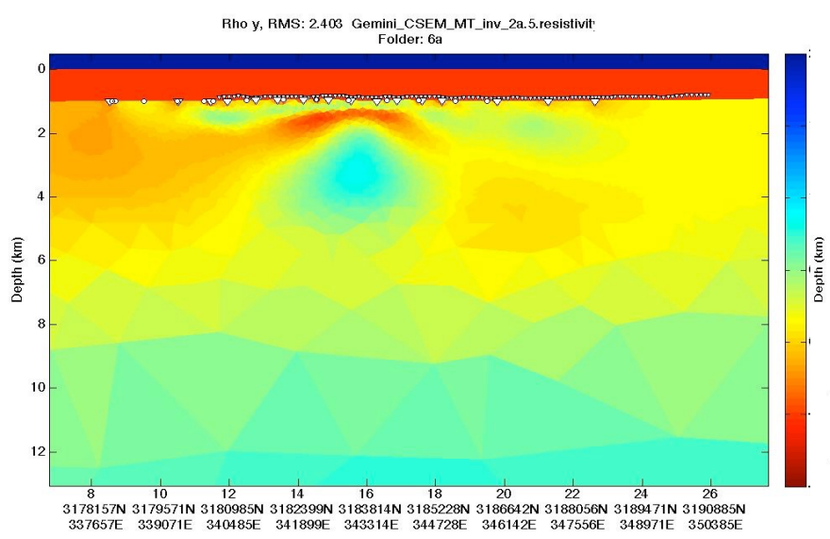
Errors come from

- statistical processing errors
- systematic errors such as instrument calibrations, and
- “geological or geophysical noise” (our inability to parameterize fine details of geology or extraneous physical processes).

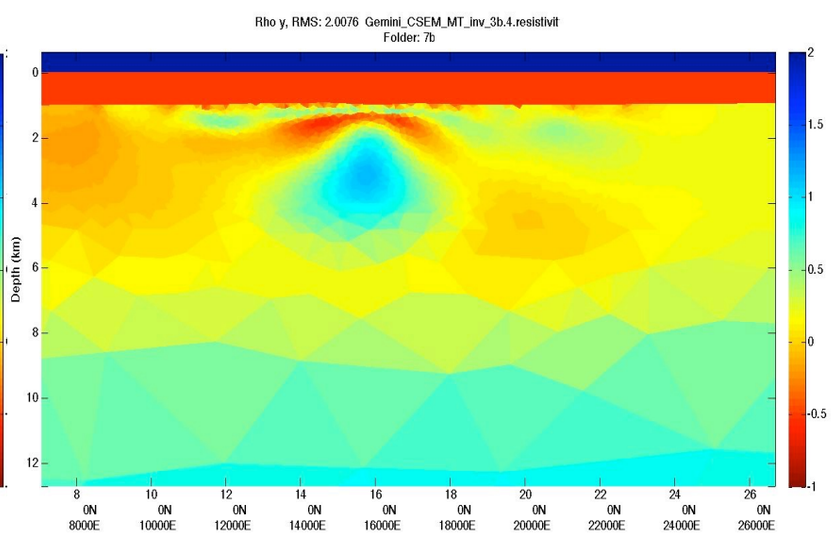
Instrument noise should be captured by processing errors, but some error models assume stationarity (i.e. noise statistics don’t vary with time). In practice, we only have a good handle on processing errors - everything else is lumped into a noise floor.

Even with well-estimated errors, choice of misfit can still be somewhat subjective.

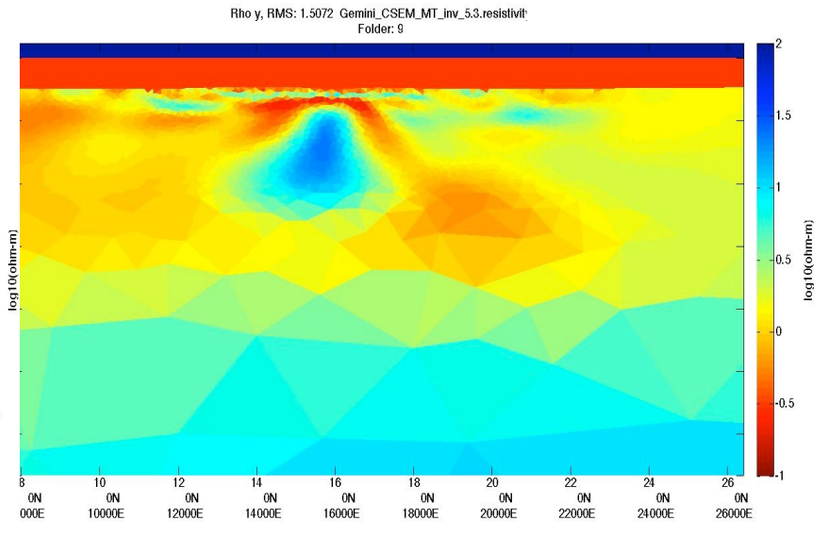
Joint 2D inversion of marine CSEM (3 frequencies, no phase) and MT (Gemini salt prospect, Gulf of Mexico):



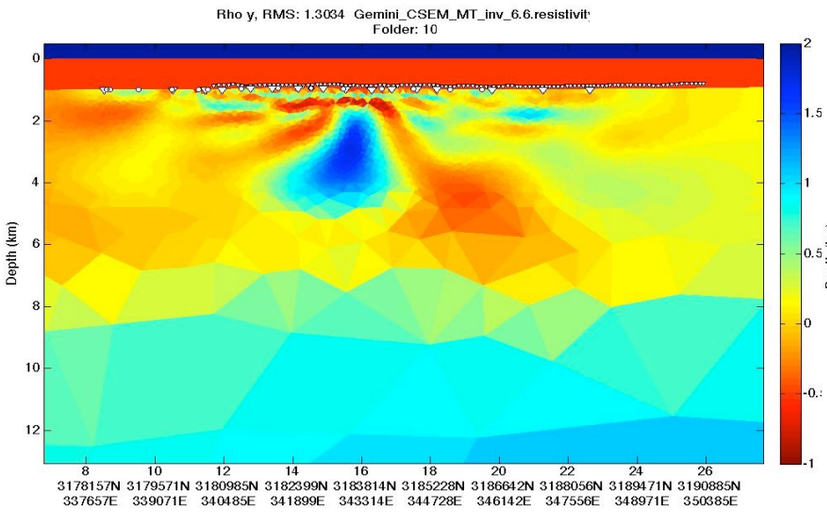
Target misfit 2.4



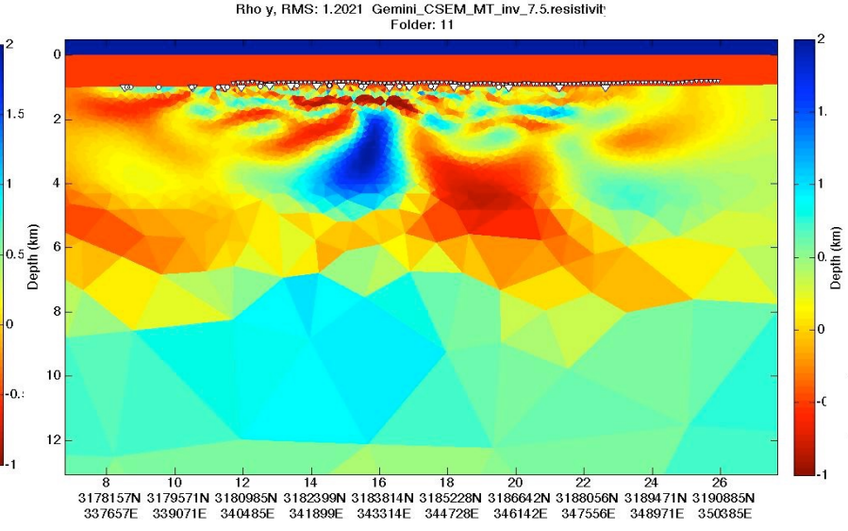
Target misfit 2.0



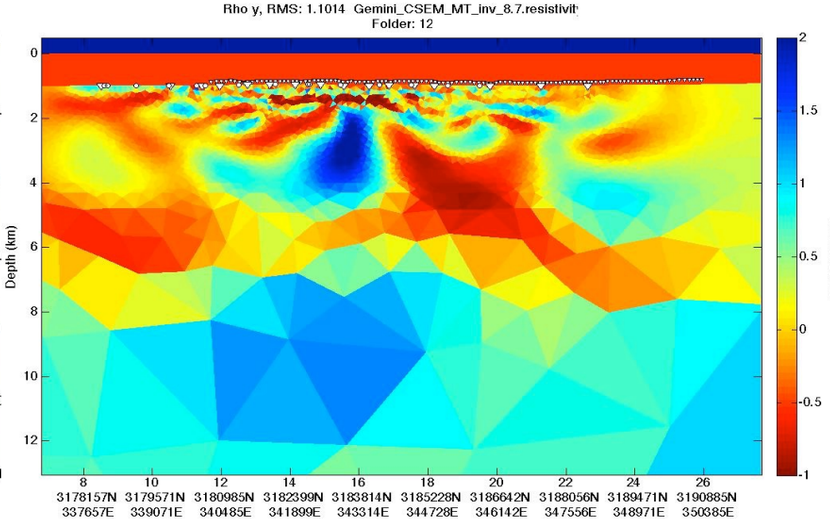
Target misfit 1.5



Target misfit 1.3

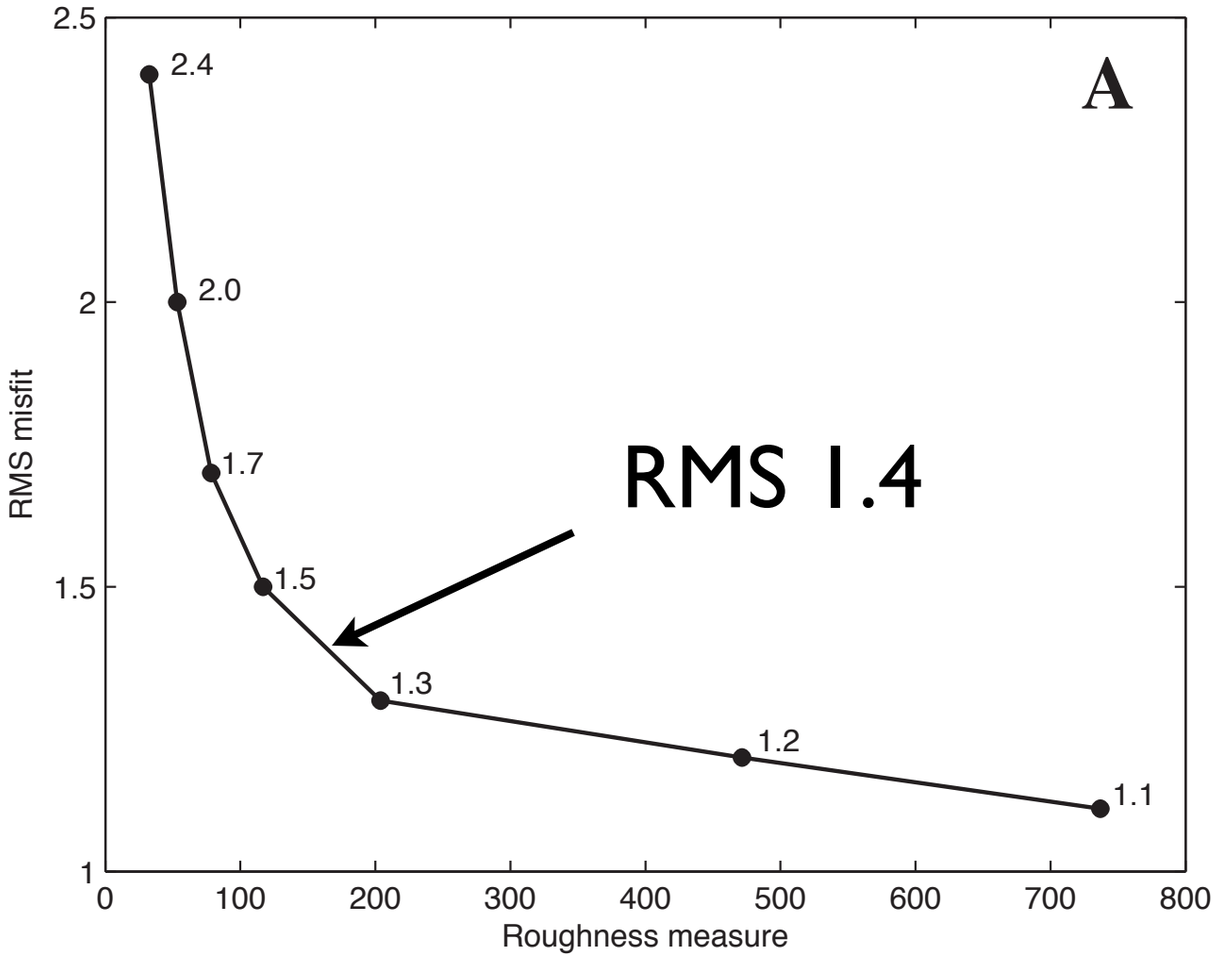


Target misfit 1.2

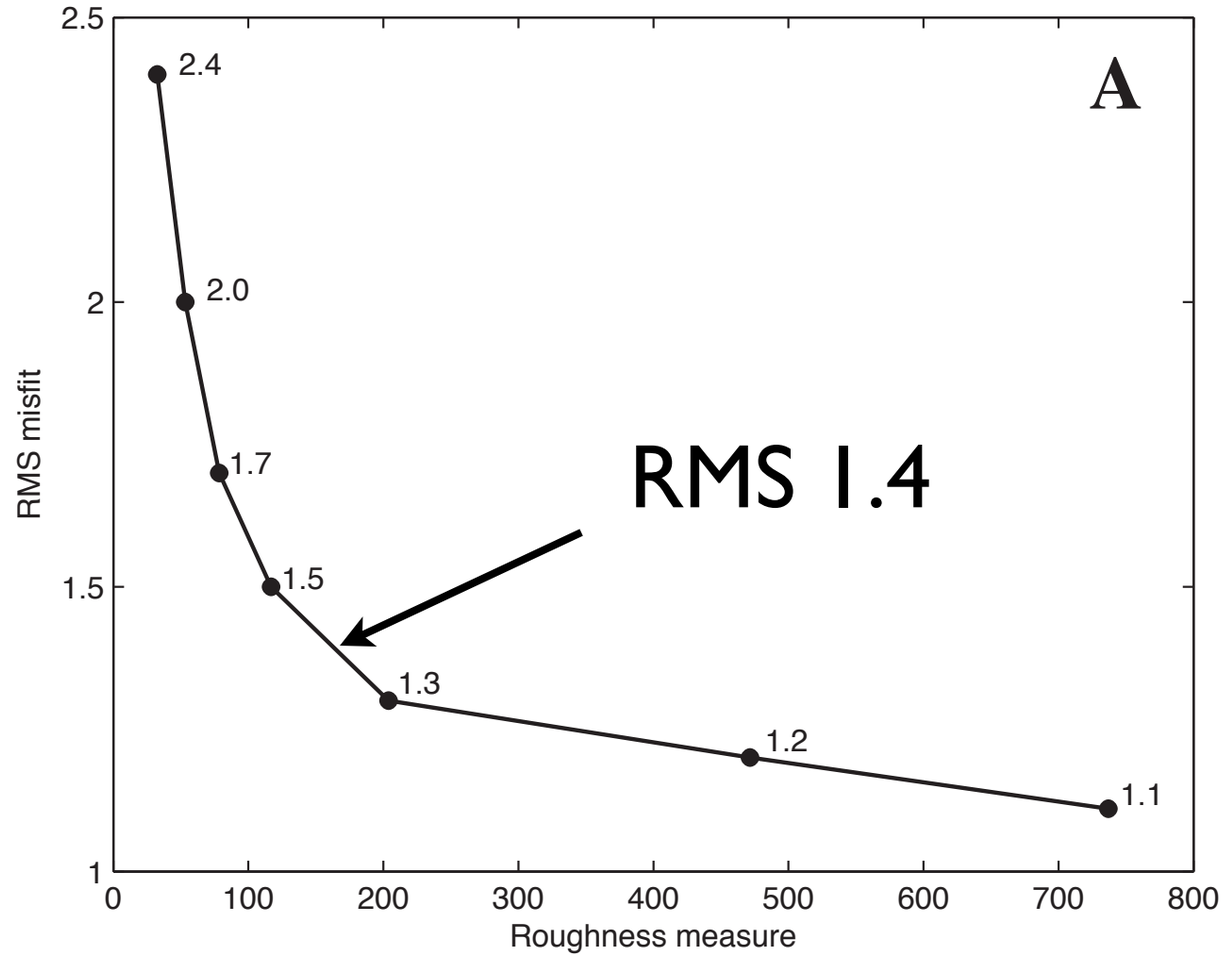


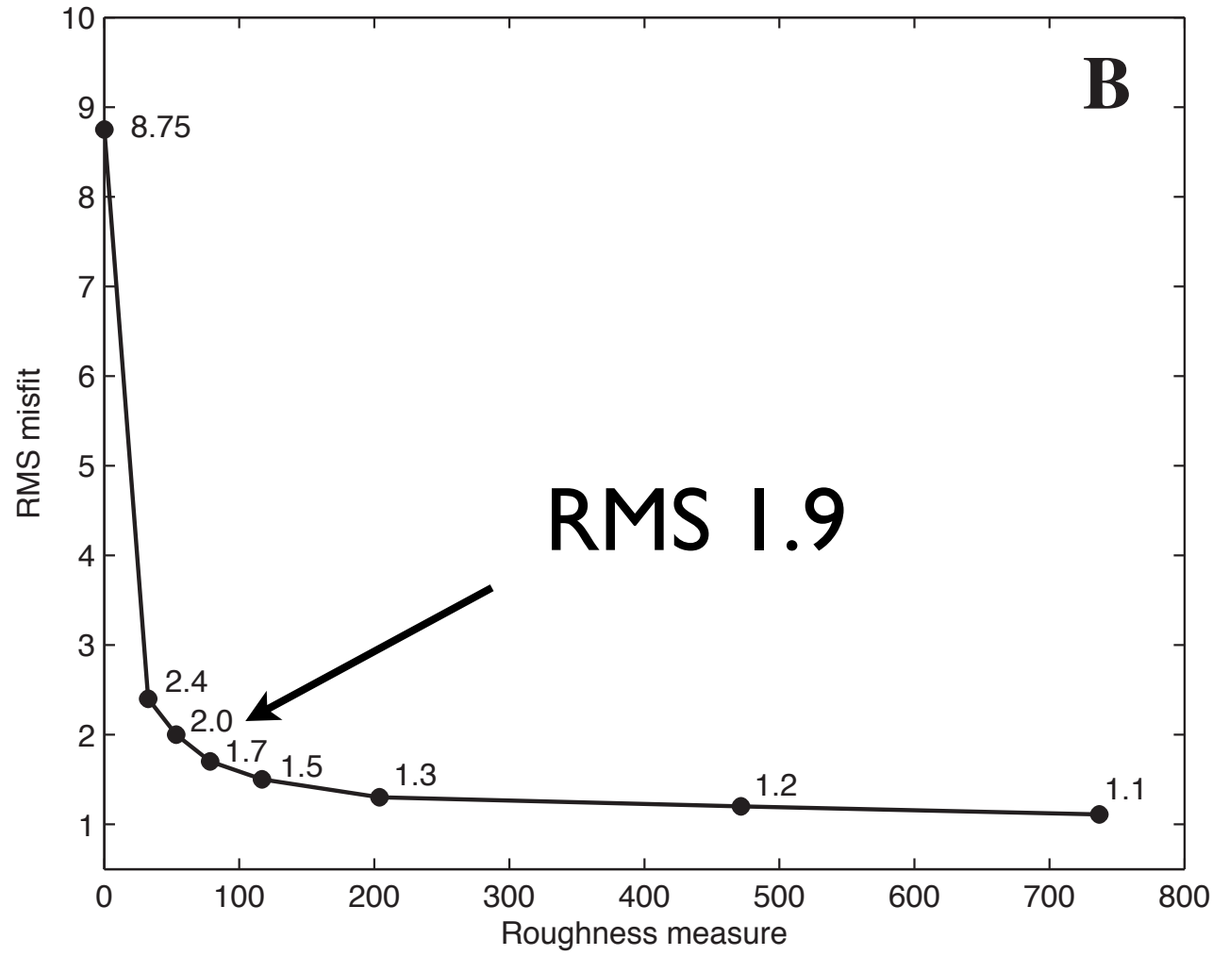
Target misfit 1.1

Beware of trade-off (“L”) curves:



Beware of trade-off (“L”) curves:



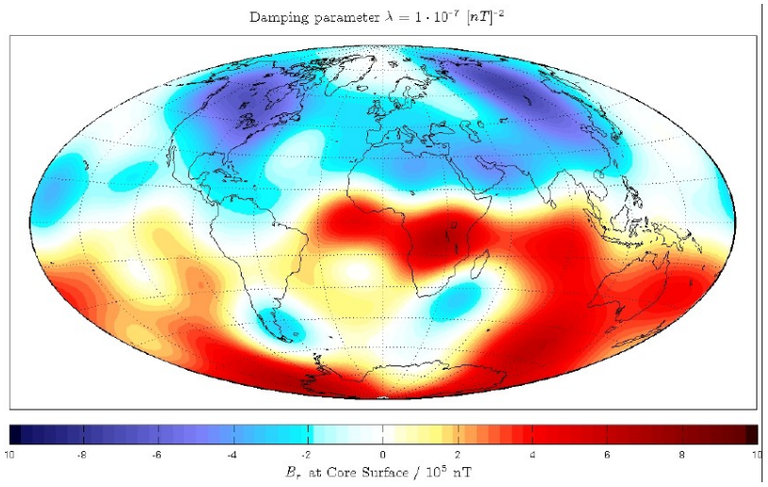


(they are not as objective as proponents say...)

Choice of Misfit Again

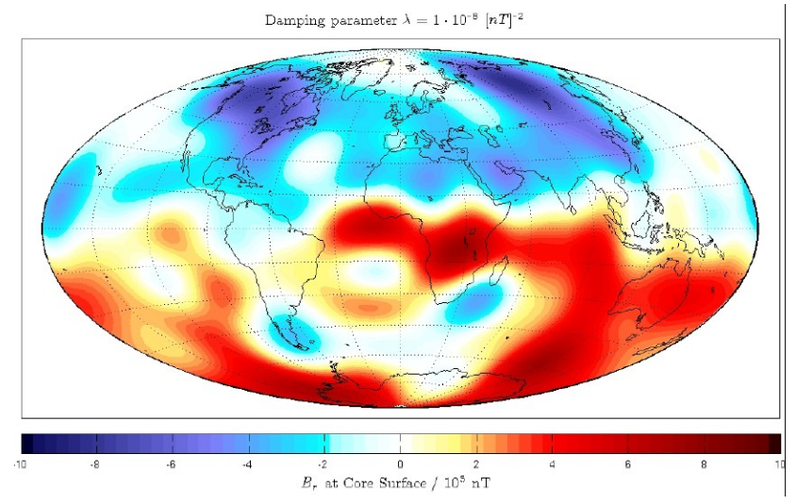
Regularization

Solution on the trade-off curve



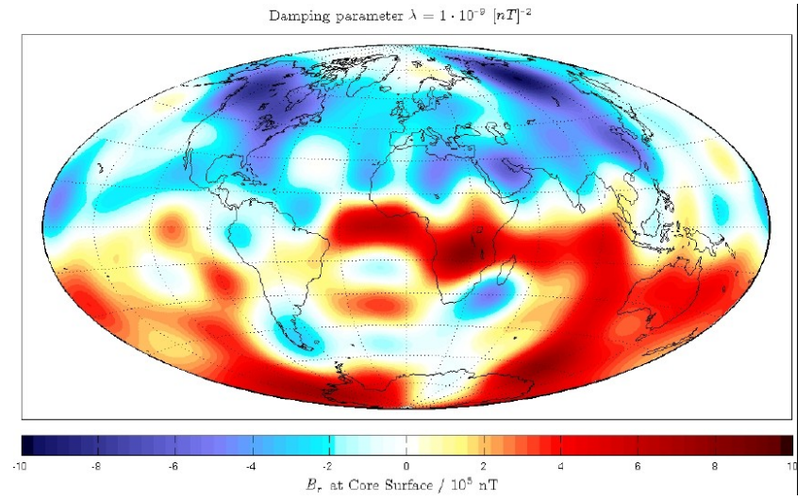
Regularization

Solution on the trade-off curve



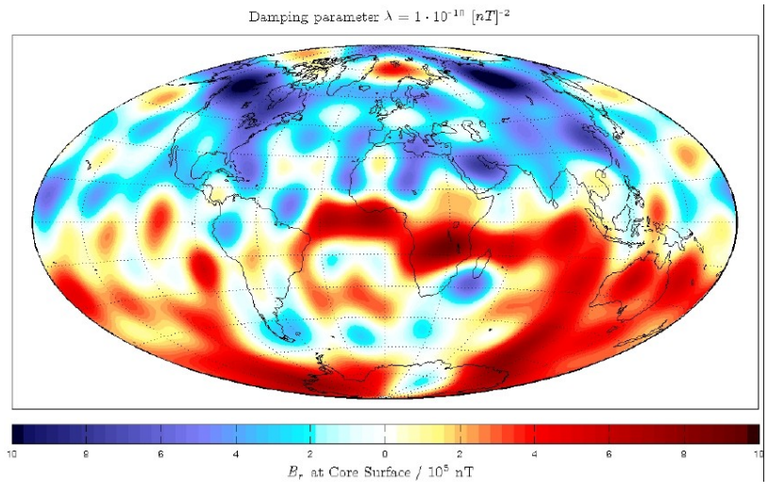
Regularization

Solution on the trade-off curve



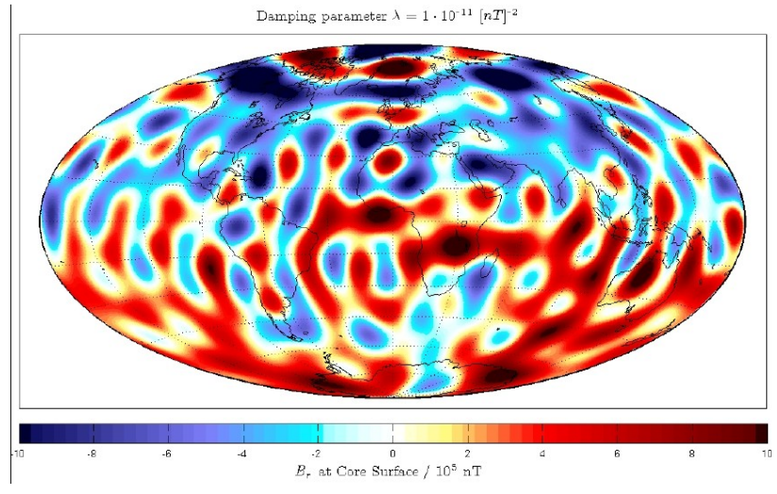
Regularization

Solution on the trade-off curve



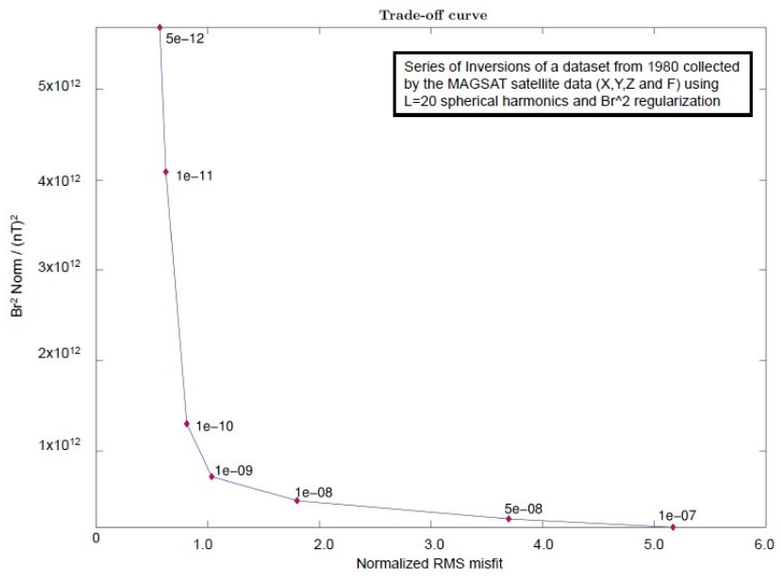
Regularization

Solution on the trade-off curve

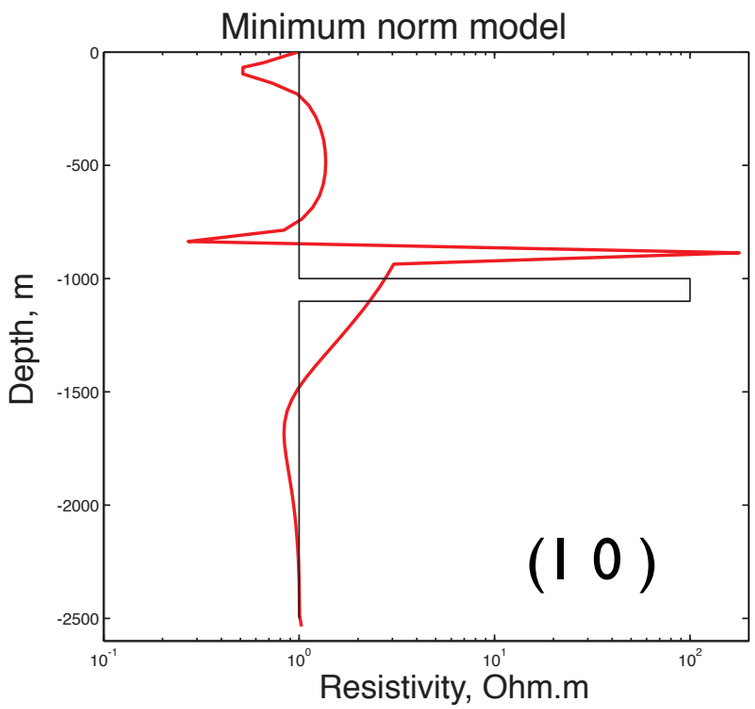


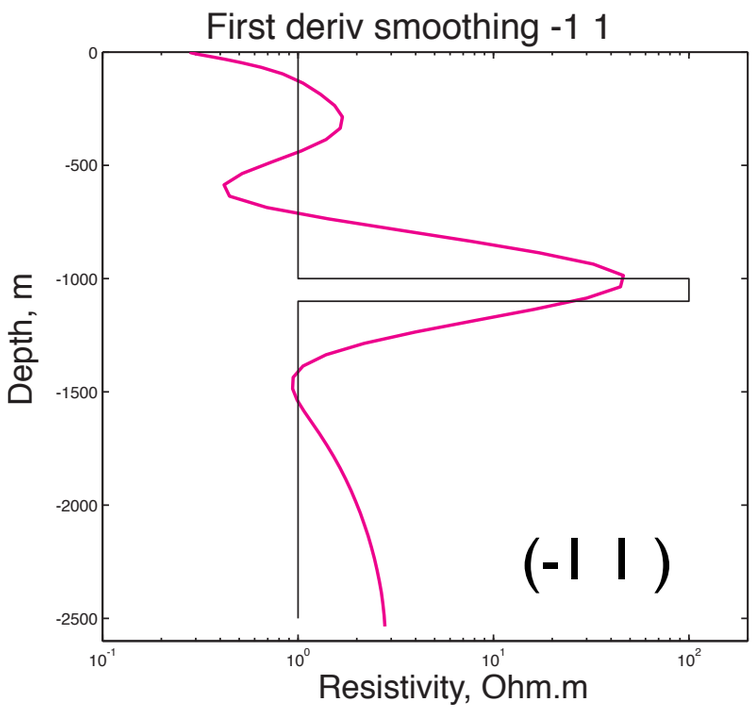
Regularization

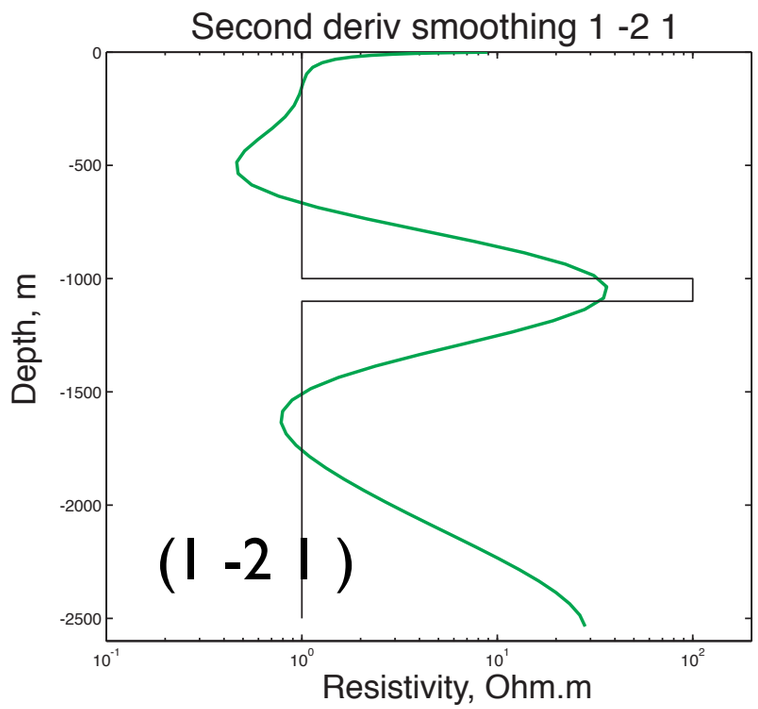
Example of core surface field trade-off curve

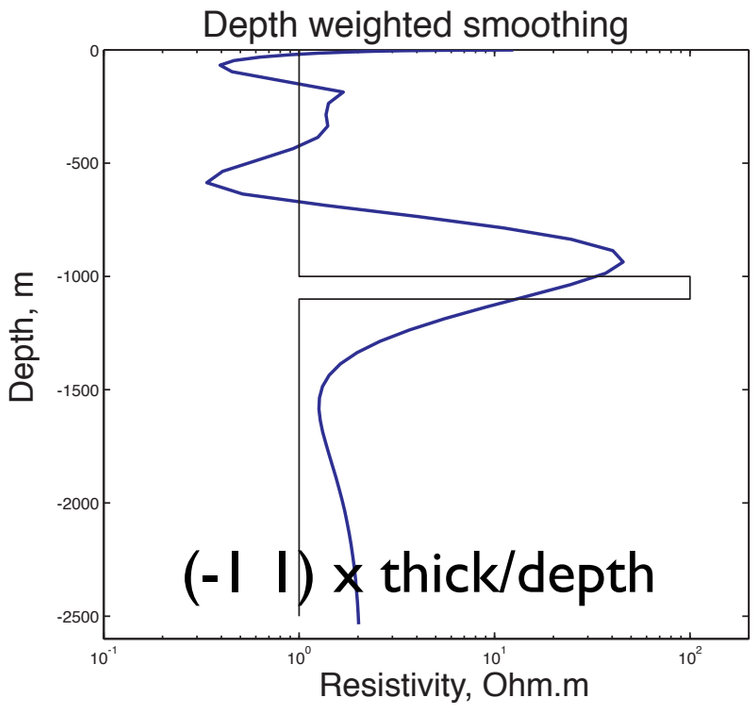


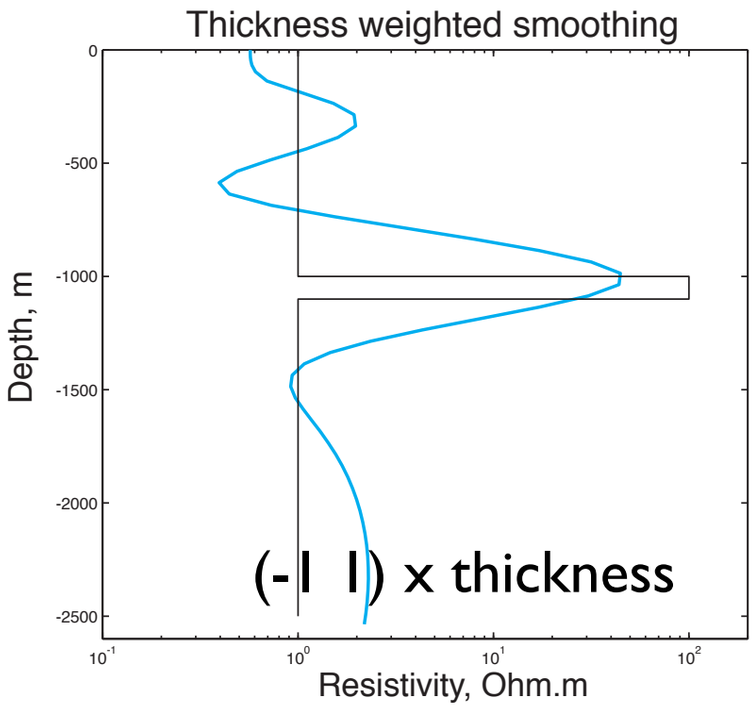
For a given misfit, the model now depends on ($\mathbf{R}$)











Finally, remember that geophysical inversion for model construction depends on much more than the data alone: