Inverse Problems in Geophysics
Part 10: Probability and Likelihood

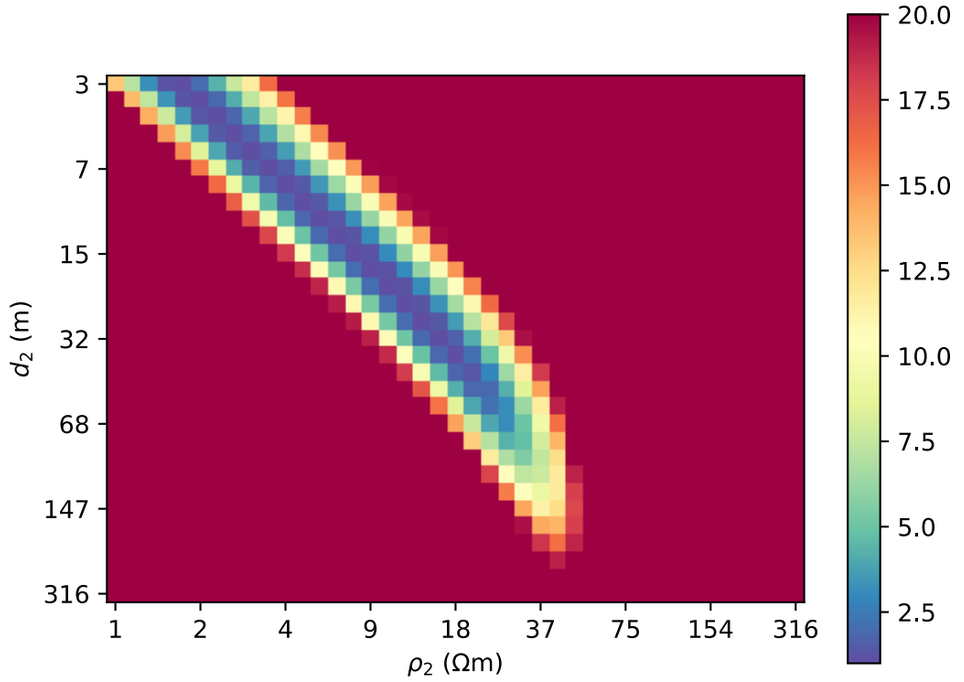
2. MGPY+MGIN

Thomas Günther
thomas.guenther@geophysik.tu-freiberg.de

TU BERGAKADEMIE FREIBERG
TUBAF
Die Ressourcenuniversität.
Seit 1765.

Recap

- linear problems: least-squares solution of (regularized) problem
- non-linear problems: linearization of $f(m) \Rightarrow$ linear problem for $\Delta\mathbf{m}$ and $\Delta\mathbf{d} = \mathbf{d} - \mathbf{f}(\mathbf{m})$
- **grid search**: systematic search through model space



Objective function for VES

Alternatives to grid search

Monte Carlo search

draw random samples and accept them if the error is improved

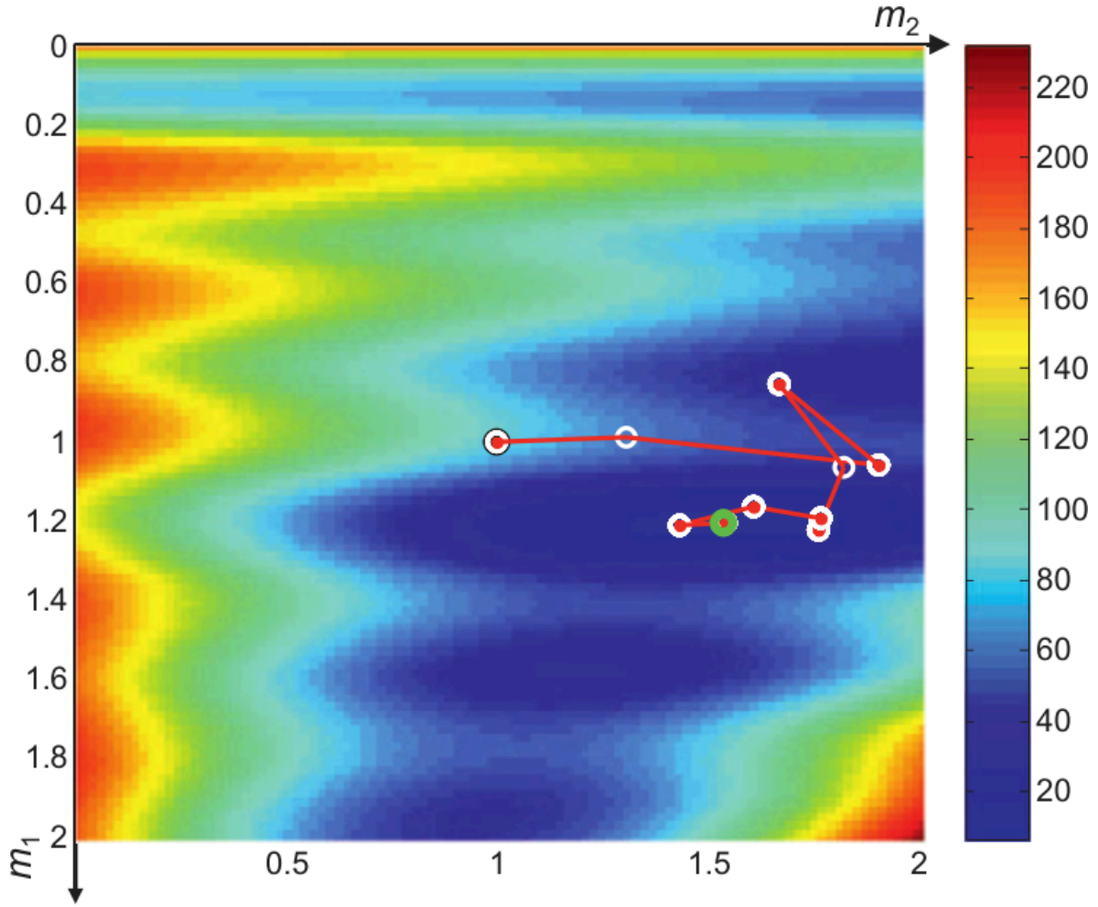
undirected search (Newtons method is directed)

Simulated annealing

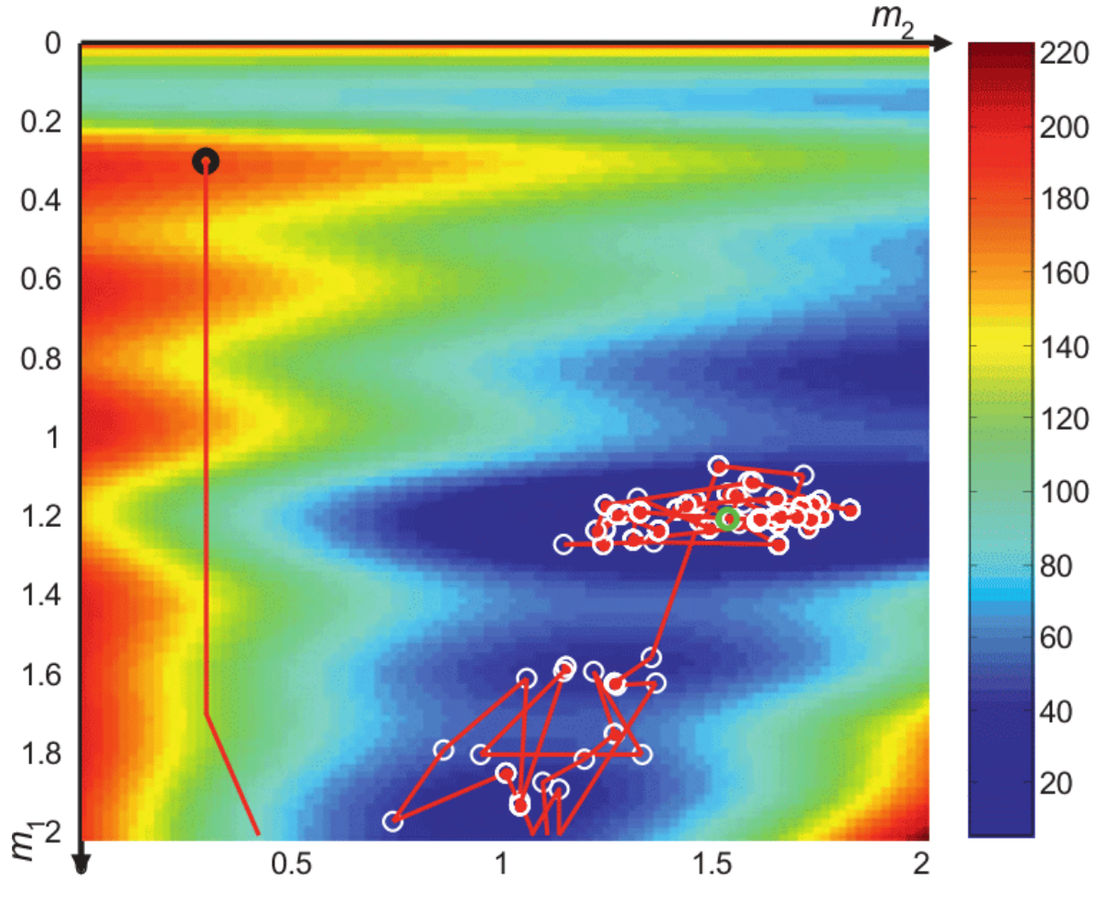
decrease temperature controlling particle movements:

high T: undirected, low T: search in vicinity of current model

Monte Carlo vs. Simulated Annealing

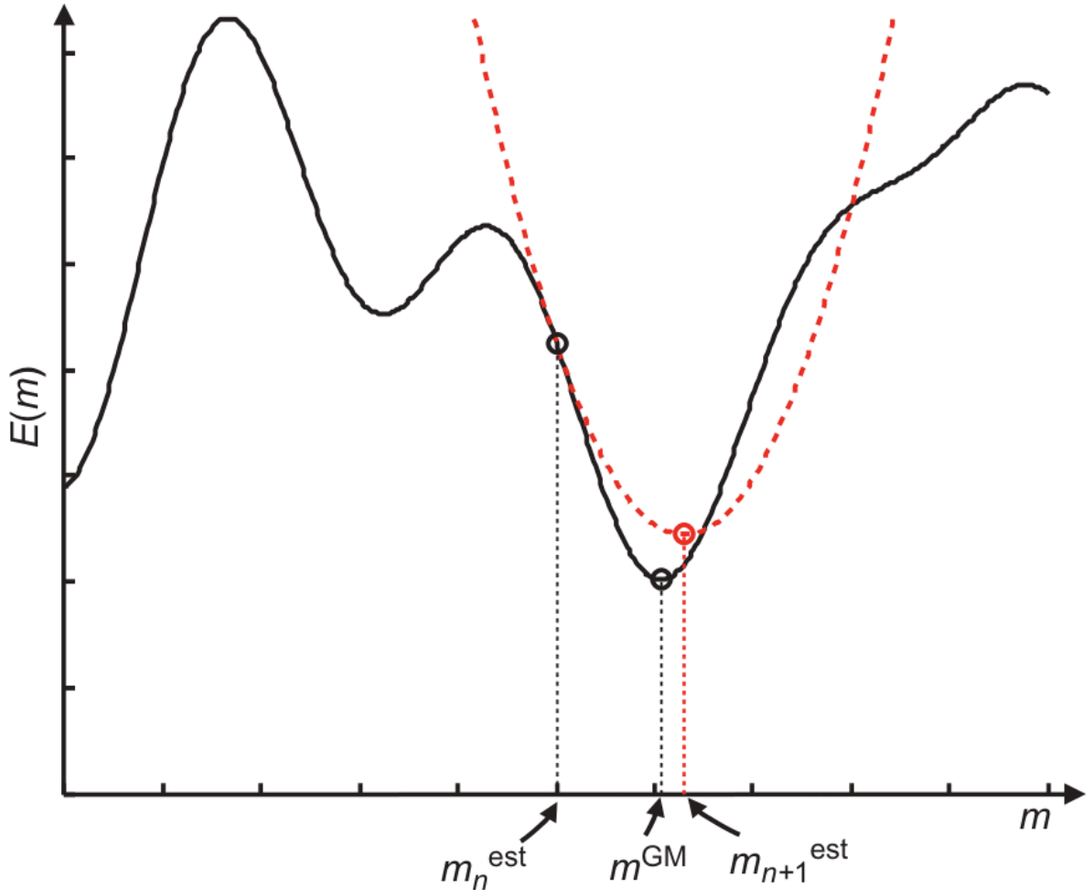


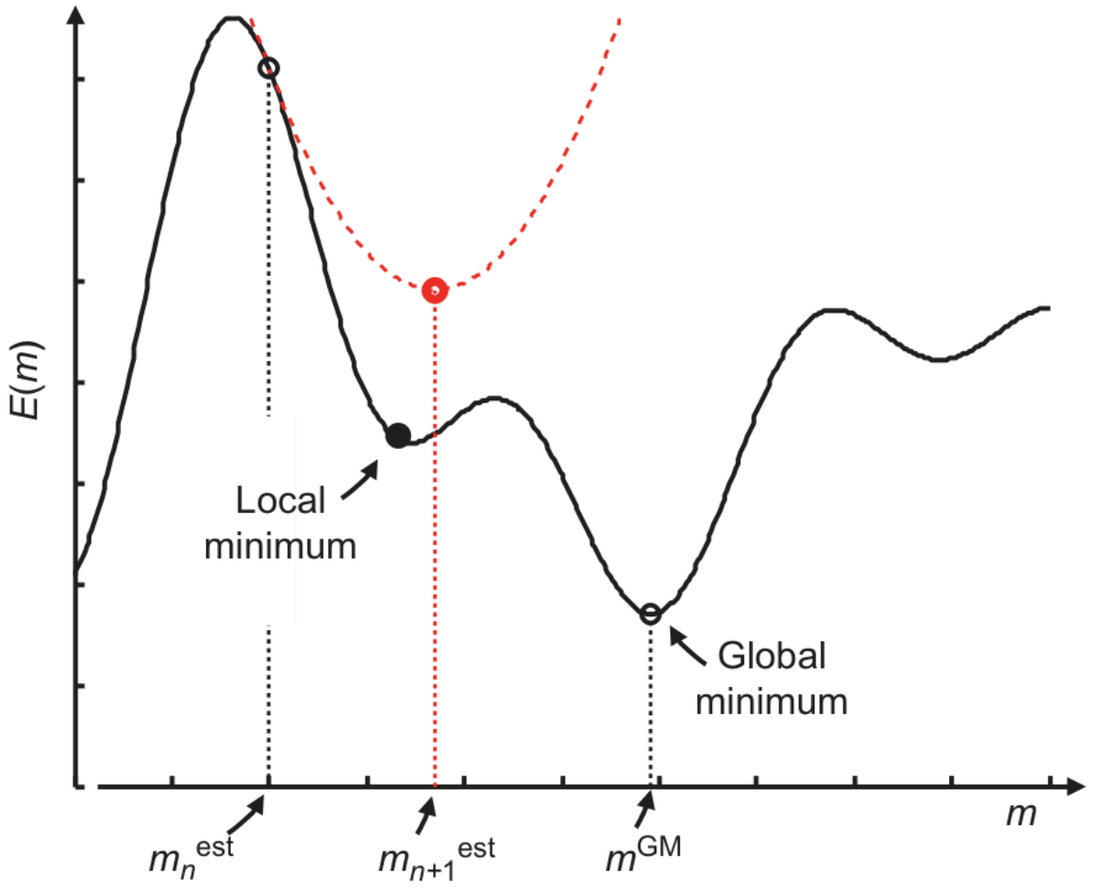
Monte Carlo method



Simulated Annealing

Newtons method (Menke, 2012)





linearize with value, slope and curvature of Φ_d

Gauss-Newton minimization

1. Choose starting model $\mathbf{m}^0$ and set n=0
2. Compute model response $\mathbf{f}(\mathbf{m}^n)$
3. Compute sensitivity matrix $\mathbf{S}^n$
4. Solve linearized subproblem $\mathbf{S}^n \Delta\mathbf{m}^n = \Delta\mathbf{d} = (\mathbf{d} - \mathbf{f}(\mathbf{m}^n))$
5. Optimize line search parameter τ^n
6. Update model by $\mathbf{m}^{n+1} = \mathbf{m}^n + \tau^n \Delta\mathbf{m}^n$
7. If convergence quit, otherwise $n \leftarrow n + 1$ & proceed with 2.

Computation of the sensitivity matrix

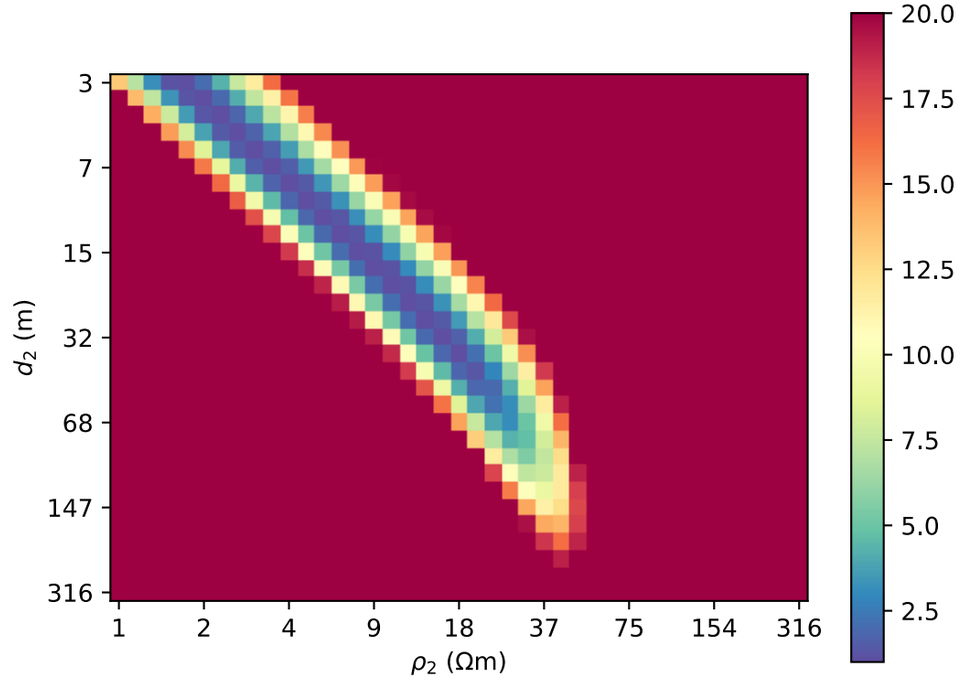
- analytically (derivation of the forward operator)
- transforming the PDE and its numerical solution
- perturbation method (brute force)

$$\frac{\partial f_i(\mathbf{m})}{\partial m_j} \approx \frac{f_i(\mathbf{m} + \delta_j \Delta m) - f_i(\mathbf{m})}{\Delta m}$$

with the Dirac vector $\delta_j = [0, \ldots, 0, 1, 0, \ldots, 0]^T$

$\Rightarrow$ one full forward computation for every model parameter

Model transformation



Objective function for VES

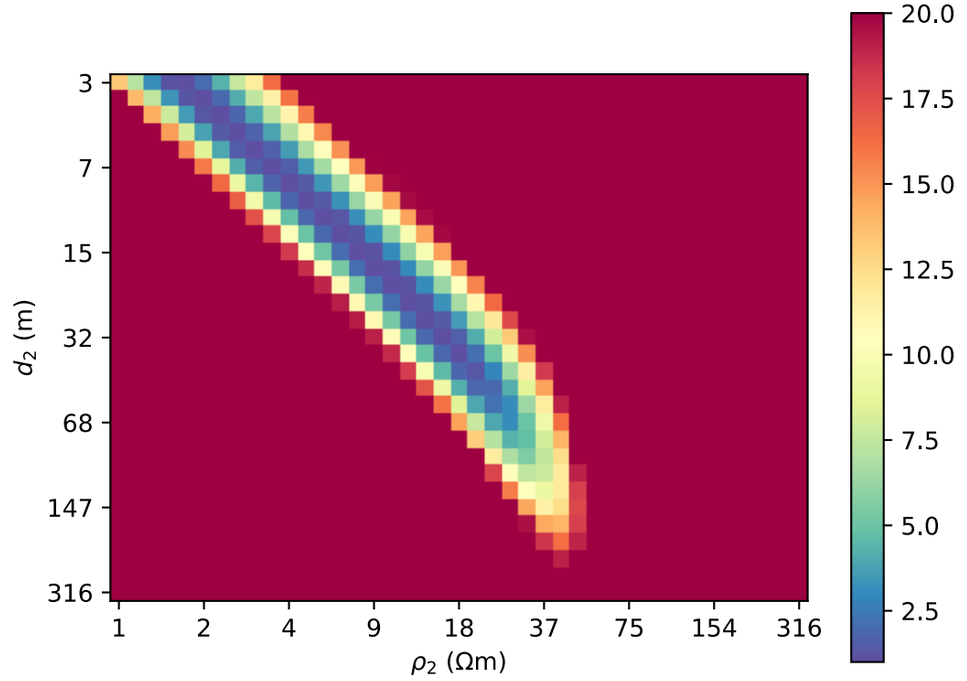
To keep the parameters positive, we often invert for the logarithms.

If we invert for $\hat{m}$ instead of m, we use the chain rule

$$\frac{\partial f}{\partial \hat{m}} = \frac{\partial f}{\partial m} \cdot \frac{\partial m}{\partial \hat{m}} = \frac{\partial f}{\partial m} \Big/ \frac{\partial \hat{m}}{\partial m}$$

E.g. $\partial \log \rho / \partial \rho = 1/\rho$

Data transformation



Objective function for VES

Often, measured data show a wide range so that we use the logarithm

If we invert $\hat{d}$ instead of d, we use the chain rule

$$\frac{\partial \hat{f}}{\partial m} = \frac{\partial f}{\partial m} \cdot \frac{\partial \hat{f}}{\partial f}$$

E.g. $\partial \log \rho^a / \partial \rho^a = 1/\rho_a$

Combined model and data transformation

Sensitivity $S_{ij} = \frac{\partial \rho_i^a}{\partial \rho_j}$

Data $d_i = \log \rho_i^a$, model parameter $m_i = \log \rho_j$

$\Rightarrow$ Jacobian matrix $\frac{\partial \hat{f}}{\partial \hat{m}} = \frac{\partial f}{\partial m} \cdot \frac{\partial \hat{f}}{\partial f} \Big/ \frac{\partial \hat{m}}{\partial m}$

$$J_{ij} = \frac{\partial \log \rho_i^a}{\partial \log \rho_j} = \frac{\partial \rho_i^a}{\partial \rho_j} \cdot \frac{\rho_j}{\rho_i^a}$$

Regularization

$$\Phi = (\mathbf{d} - \mathbf{f}(\mathbf{m}))^T(\mathbf{d} - \mathbf{f}(\mathbf{m})) + \lambda(\mathbf{c} - \mathbf{C}\mathbf{m})^T(\mathbf{c} - \mathbf{C}\mathbf{m})$$

$$b_i = \frac{\partial \Phi}{\partial m_i} = -2\mathbf{S}^T(\mathbf{d} - \mathbf{f}(\mathbf{m})) - 2\lambda\mathbf{C}^T(\mathbf{c} - \mathbf{C}\mathbf{m})$$

$$B_{ij} = \frac{\partial^2 \Phi}{\partial m_i \partial m_j} = \frac{\partial b_i}{\partial m_j} \approx 2\mathbf{S}^T\mathbf{S} + 2\lambda\mathbf{C}^T\mathbf{C}$$

$$\Rightarrow (\mathbf{S}^T\mathbf{S} + \lambda\mathbf{C}^T\mathbf{C})\Delta\mathbf{m} = \mathbf{S}^T(\mathbf{d} - \mathbf{f}(\mathbf{m}^n)) + \lambda\mathbf{C}^T(\mathbf{c} - \mathbf{C}\mathbf{m})$$

Resolution matrices for non-linear problems

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \mathbf{S}^\dagger(\mathbf{d} - \mathbf{f}(\mathbf{m})) + \mathbf{C}^\dagger\mathbf{C}(\mathbf{m}^0 - \mathbf{m}^k)$$

with

$$\mathbf{S}^\dagger = (\mathbf{S}^T\mathbf{S} + \lambda\mathbf{C}^T\mathbf{C})^{-1}\mathbf{S}^T$$

$$\mathbf{C}^\dagger = (\mathbf{S}^T\mathbf{S} + \lambda\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$$

and their combination $\mathbf{S}^\dagger\mathbf{S} + \mathbf{C}^\dagger\mathbf{C} = \mathbf{I}$

Resolution matrices for non-linear problems

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \mathbf{S}^\dagger(\mathbf{d} - \mathbf{f}(\mathbf{m})) + \mathbf{C}^\dagger\mathbf{C}(\mathbf{m}^0 - \mathbf{m}^k)$$

assuming we “come close”, i.e. $\mathbf{m}^{est} = \mathbf{m}^{k+1}$ and $\Delta\mathbf{m}^k$ is small. With

$$\mathbf{d} = \mathbf{f}(\mathbf{m}^{true}) + \mathbf{n} = \mathbf{f}(\mathbf{m}^k) + \mathbf{S}(\mathbf{m}^{true} - \mathbf{m}^k) + \mathbf{n}$$

$$\Rightarrow \mathbf{m}^{est} = \mathbf{m}^k + \mathbf{S}^\dagger\mathbf{S}(\mathbf{m}^{true} - \mathbf{m}^k) + \mathbf{C}^\dagger\mathbf{C}(\mathbf{m}^0 - \mathbf{m}^k) + \mathbf{S}^\dagger\mathbf{n}$$

$$\Rightarrow \mathbf{m}^{est} = \mathbf{m}^k + \mathbf{S}^\dagger\mathbf{S}\mathbf{m}^{true} - (\mathbf{S}^\dagger\mathbf{S} + \mathbf{C}^\dagger\mathbf{C})\mathbf{m}^k + \mathbf{C}^\dagger\mathbf{C}\mathbf{m}^0 + \mathbf{S}^\dagger\mathbf{n}$$

$$\Rightarrow \mathbf{m}^{est} = \mathbf{R}^M\mathbf{m}^{true} + (\mathbf{I} - \mathbf{R}^M)\mathbf{m}^0 + \mathbf{S}^\dagger\mathbf{n}$$

Data resolution for non-linear problems

$$\mathbf{f}(\mathbf{m}^{k+1}) = \mathbf{S}(\mathbf{S}^{\dagger}(\mathbf{f}(\mathbf{m}^{true}) + \mathbf{d} - \mathbf{f}(\mathbf{m}^{k}))) + \mathbf{C}^{\dagger}\mathbf{C}(\mathbf{m}^{0} - \mathbf{m}^{k})$$

$$\mathbf{f}(\mathbf{m}^{est}) = \mathbf{R}^{D}\mathbf{f}(\mathbf{m}^{true}) + \mathbf{R}^{D}\mathbf{n} + (\mathbf{I} - \mathbf{R}^{D})\mathbf{f}(\mathbf{m}^{0})$$

$$\mathbf{f}(\mathbf{m}^{est}) = \mathbf{R}^{D}(\mathbf{f}(\mathbf{m}^{true}) + \mathbf{n}) + (\mathbf{I} - \mathbf{R}^{D})\mathbf{f}(\mathbf{m}^{0})$$

$$\mathbf{f}(\mathbf{m}^{est}) = \mathbf{R}^{D}\mathbf{d} + (\mathbf{I} - \mathbf{R}^{D})\mathbf{f}(\mathbf{m}^{0})$$

Response through data space & starting response through null space

Gradient methods

Sometimes, computing the Frechet derivatives is too expensive, but the misfit Φ_d and its derivatives $\partial \Phi_d / \partial m_i$ are easy to compute

Steepest descent method

Go into the direction where Φ_d decreases:

$$\nu = -\frac{\nabla \Phi_d}{|\nabla \Phi_d|}$$

Line search using Armijo rule

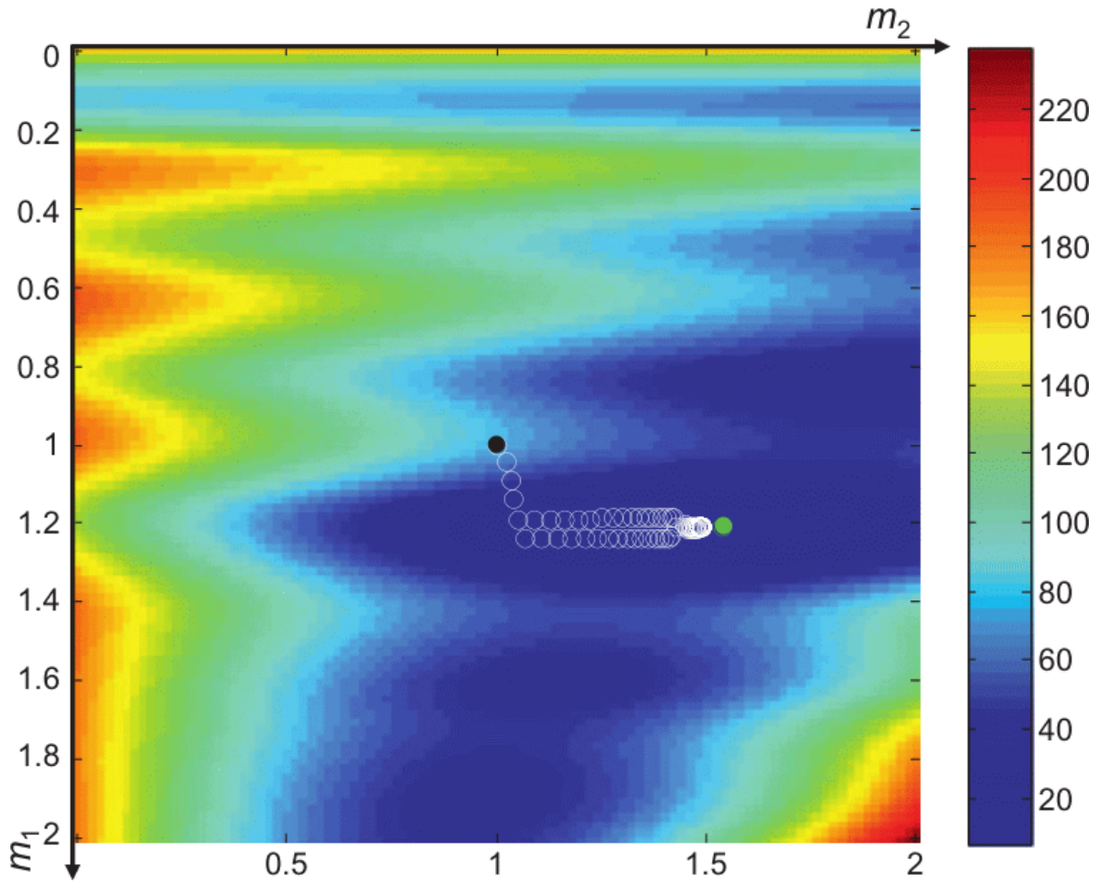
$$\mathbf{m}^{k+1} = \mathbf{m}^{k} + \alpha\nu$$

Armijo's rule to determine α

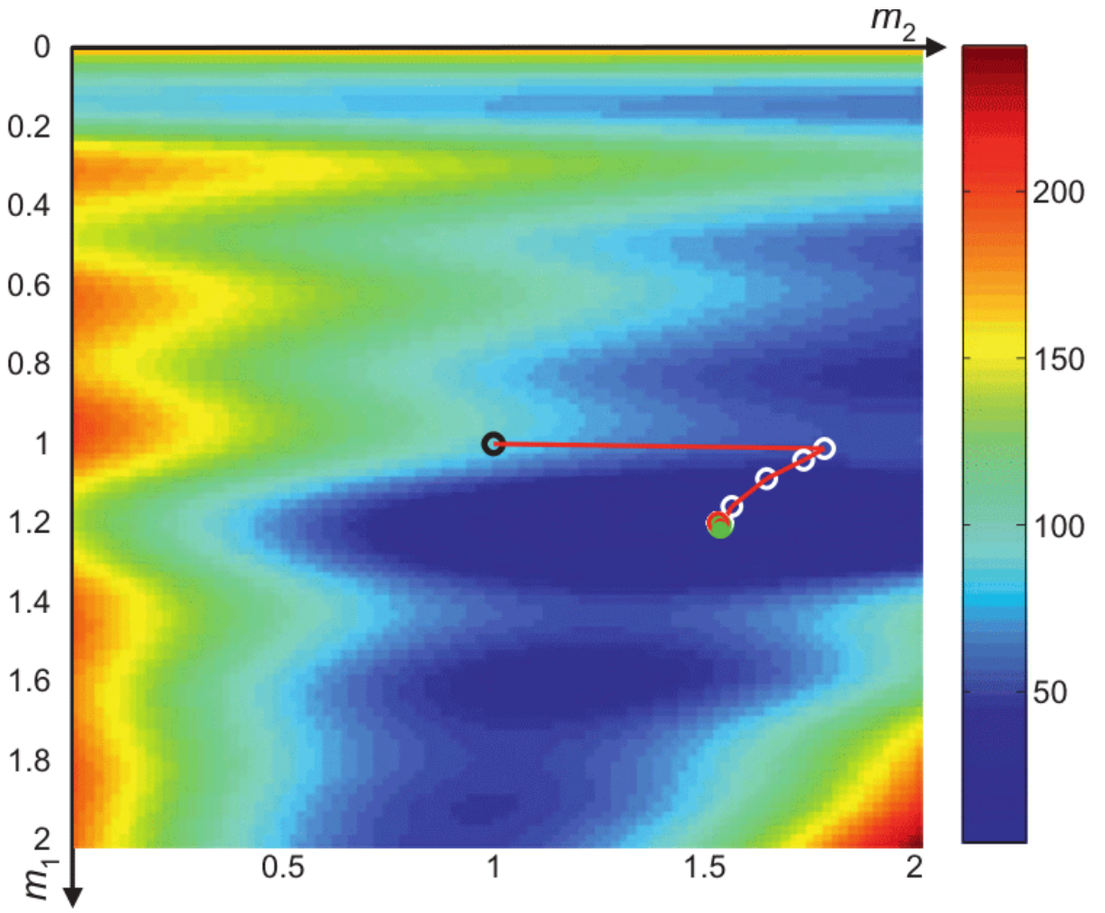
$$\Phi_d(\mathbf{m}^{k+1}) \leq \Phi_d(\mathbf{m}^{k}) + c\alpha\nu^T\boldsymbol{\nabla}\Phi_d$$

- c-empirical constant, typically 1e-4
- start with large α and decrease until rule is fulfilled

Gradient vs. Newton

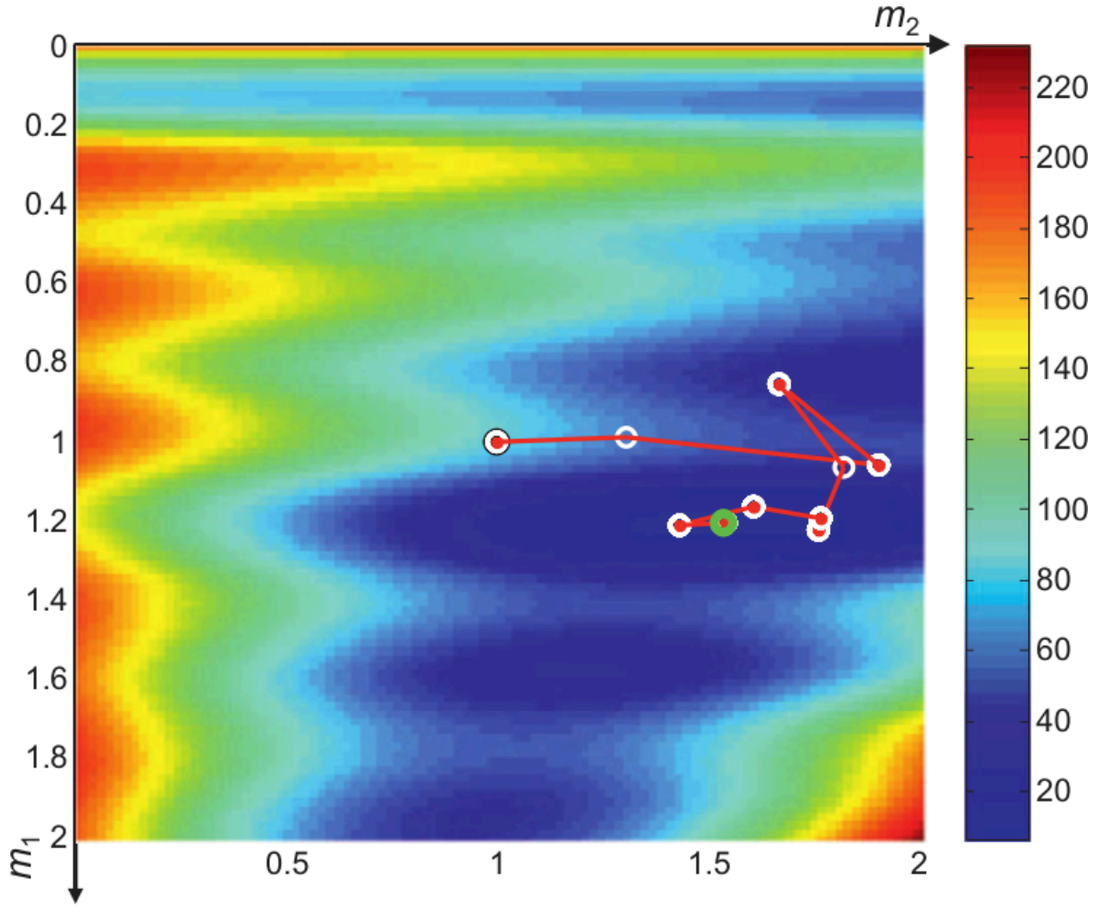


Gradient method

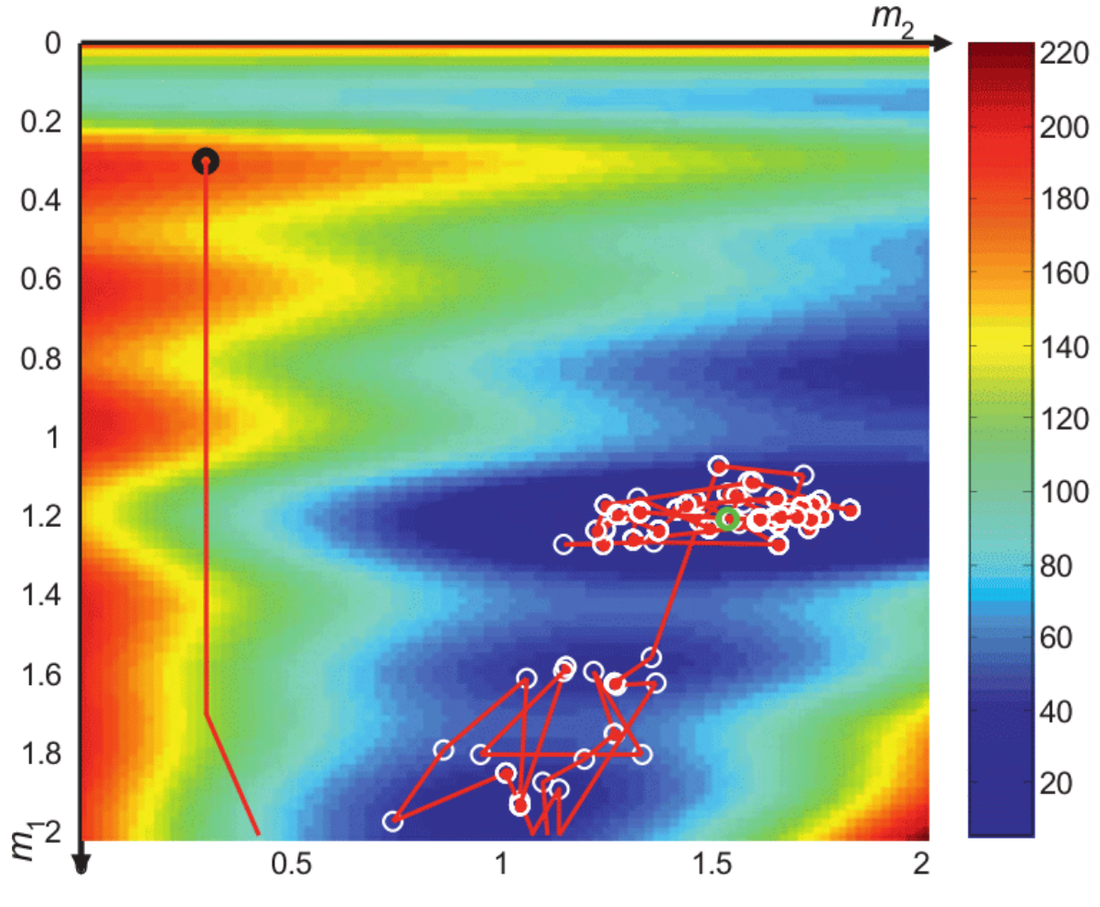


Newtons method

Monte Carlo vs. Simulated Annealing



Monte Carlo method



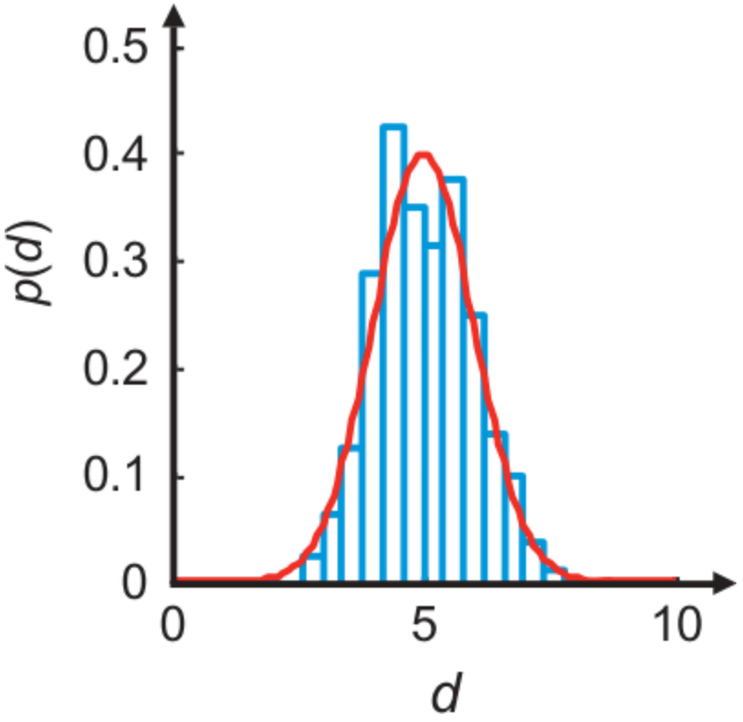
Simulated Annealing

Probability and likelihood

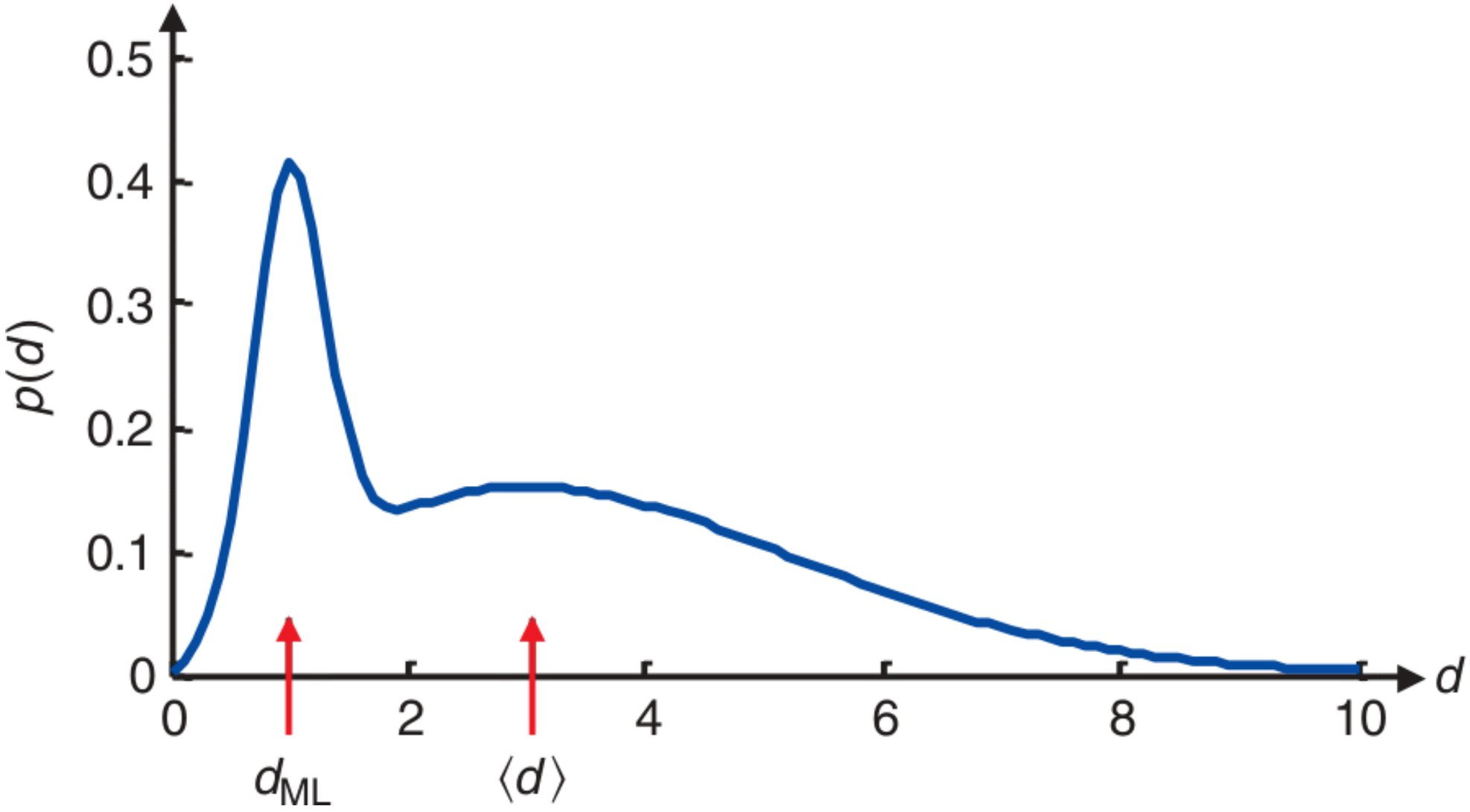
random variables: probability through many repetitions

maximum $p(d)$ is most likely

expectation: $\langle d \rangle = \int d \cdot p(d)\, \mathrm{d}d$



Probability density function



Variance

$$\sigma^2 = \int (d - \langle d \rangle)^2 p(d)\, \mathrm{d}d$$

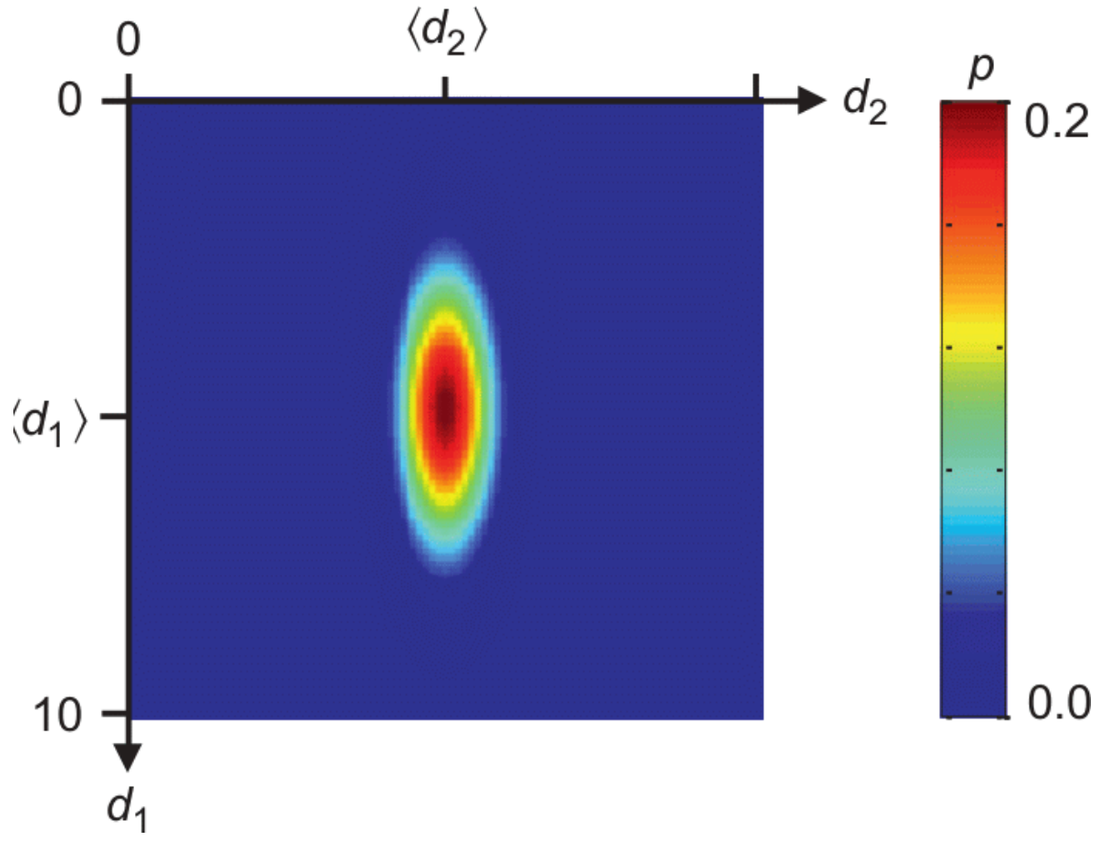
σ is a measure of the width of the distribution

related to standard deviation and mean of sampling

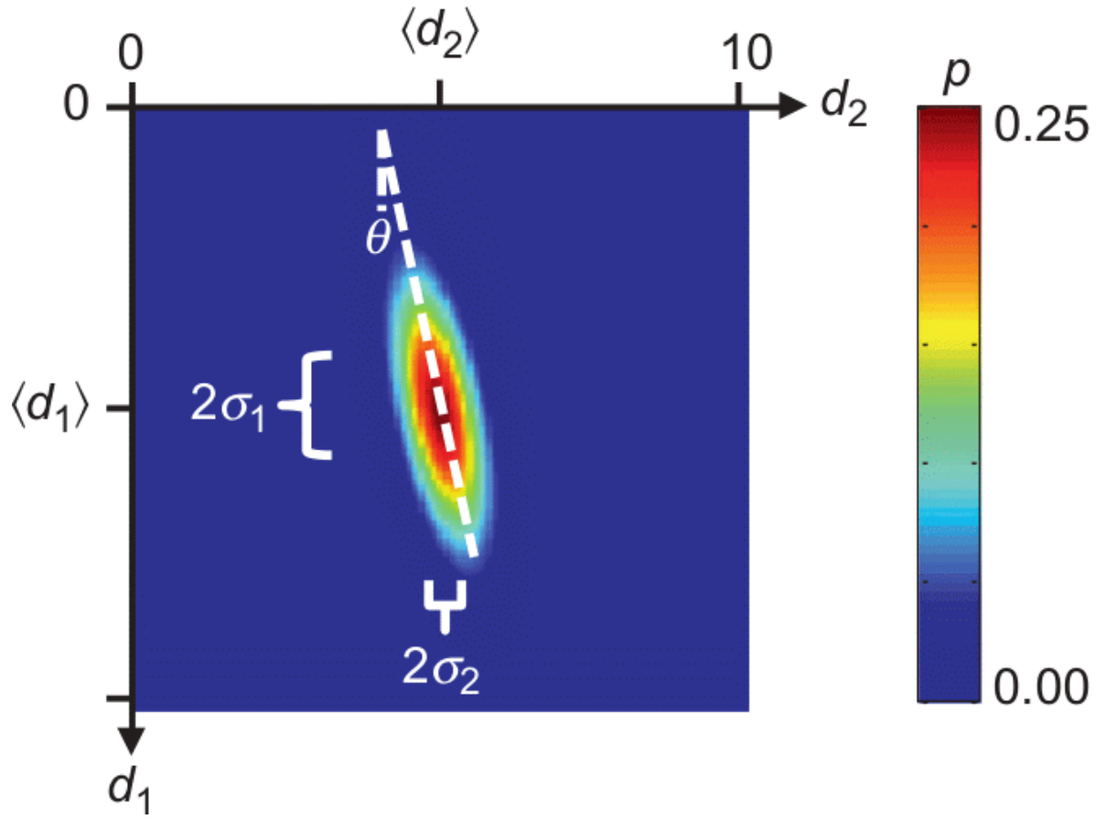
$$\sigma_{est}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (d_i - \langle d \rangle)^2 \quad \text{with} \quad \langle d \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i$$

Correlated data

independent: $p(\mathbf{d}) = p(d_1)p(d_2)\ldots p(d_N)$



uncorrelated data (Menke, 2012)



correlated data (Menke, 2012)

Covariance

(measure of correlation between data)

$$\mathrm{cov}(d_1, d_2) = \int\int (d_1 - \langle d_1 \rangle)(d_2 - \langle d_2 \rangle) p(d_1, d_2)\, \mathrm{d}d_1\, \mathrm{d}d_2$$

$$\langle d_i \rangle = \int \ldots \int d_i p(\mathbf{d})\, \mathrm{d}d_1 \ldots \mathrm{d}d_N$$

Covariance propagation

Linear problem $\mathbf{m} = \mathbf{M}\mathbf{d}$, e.g. $\mathbf{m} = \mathbf{G}^{\dagger}\mathbf{d}$

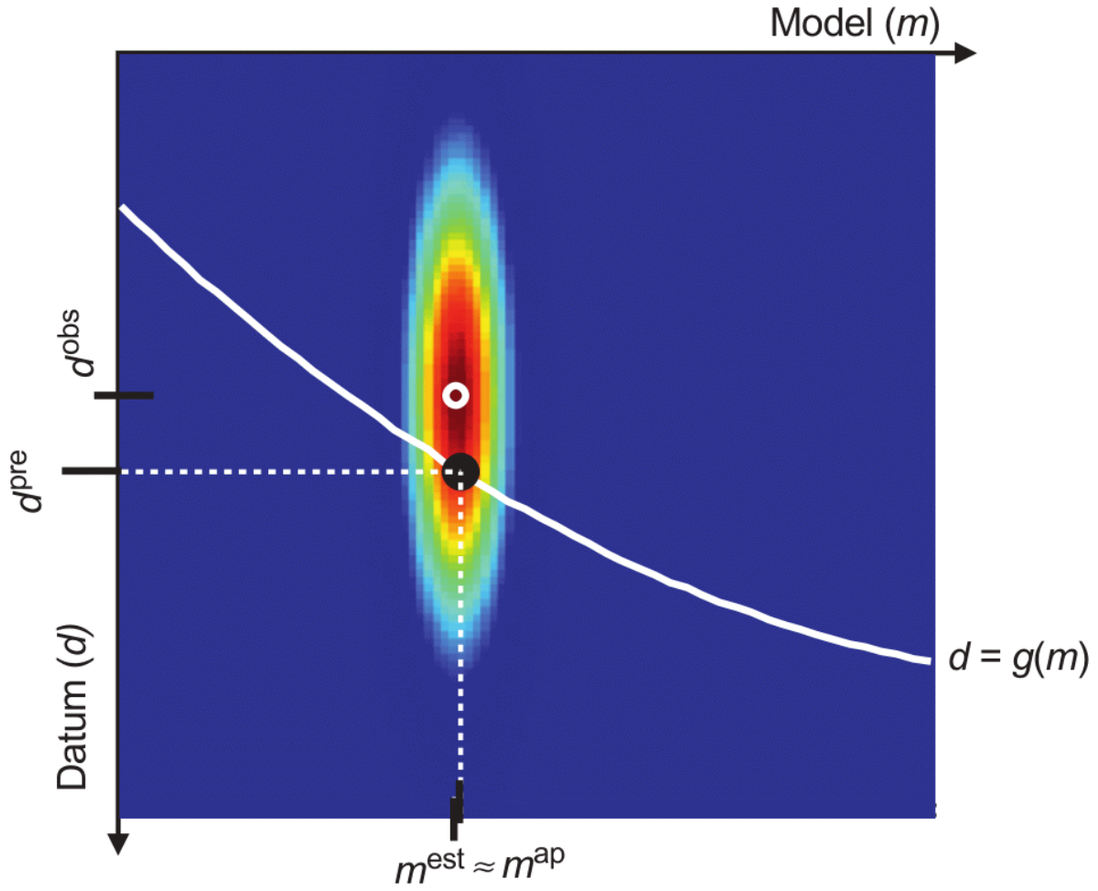
Mean value $\langle \mathbf{m} \rangle = \mathbf{M} \langle \mathbf{d} \rangle + \mathbf{n}$ and covariance

$$\mathrm{cov}(\mathbf{m}) = \mathbf{M}\mathrm{cov}(\mathbf{d})\mathbf{M}^T$$

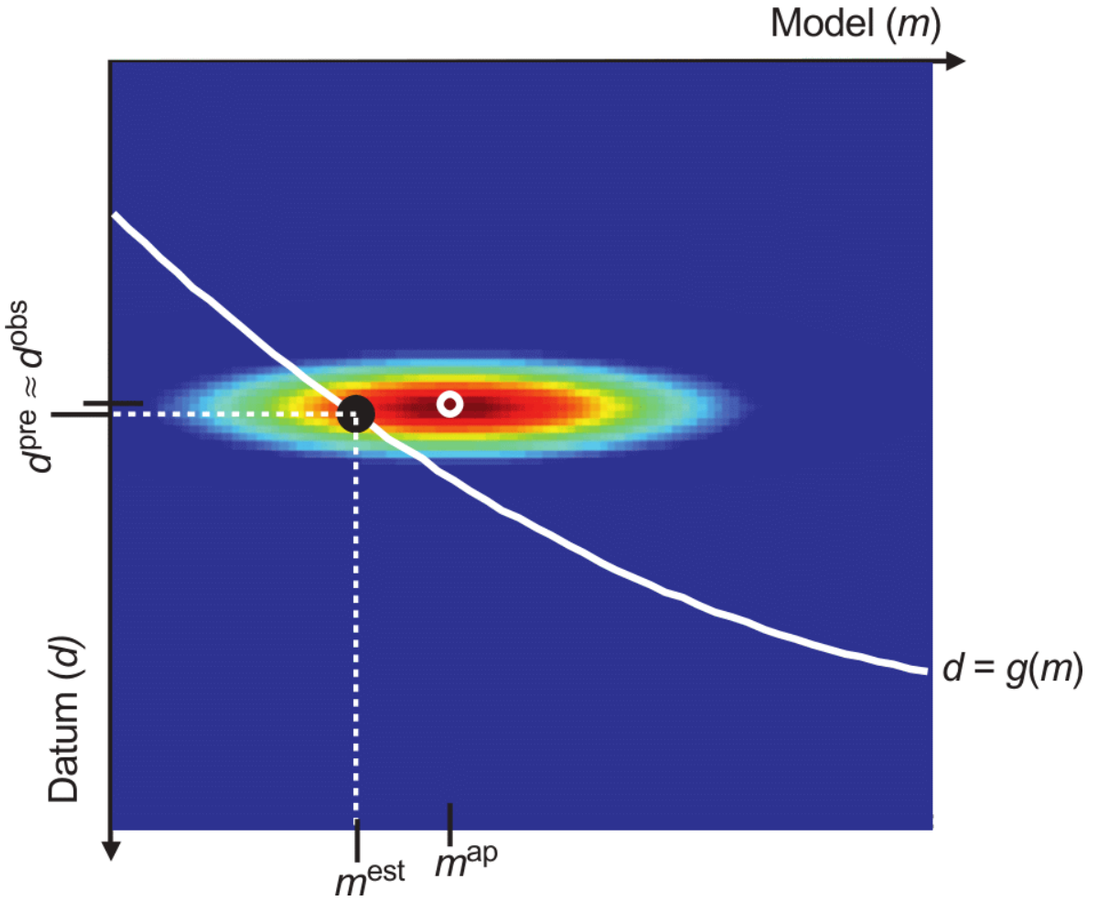
Least-squares: $\mathbf{M} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$, uncorrelated data: $\mathrm{cov}(\mathbf{d}) = \sigma_d^2\mathbf{I}$

$$\Rightarrow \mathrm{cov}(\mathbf{m}) = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\sigma_d^2\mathbf{I}((\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T)^T = \sigma_d^2(\mathbf{G}^T\mathbf{G})^{-1}$$

A priori knowledge



accurate prior model (Menke, 2012)



accurate data (Menke, 2012)

Bayes' theorem

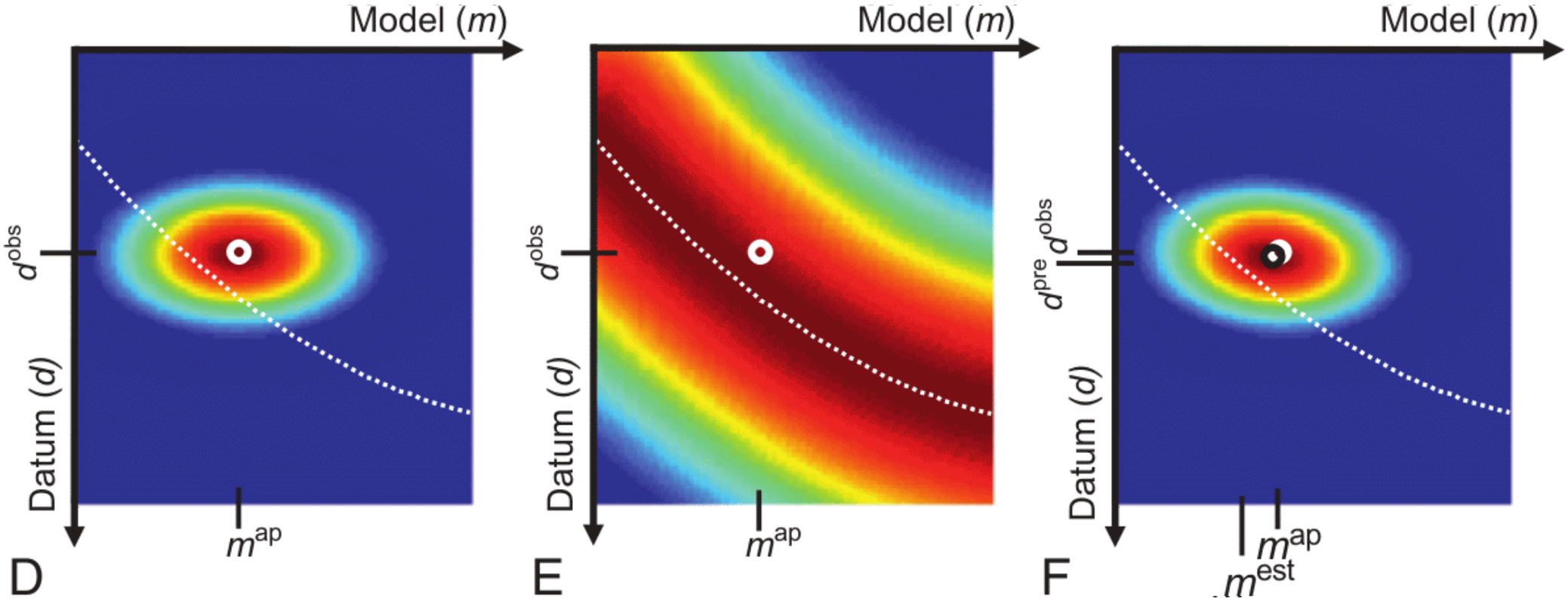
Conditional probability

$$p(d_1|d_2) = \frac{p(d_1,d_2)}{p(d_2)}$$

$$p(\mathbf{m}|\mathbf{d})p(\mathbf{d}) = p(\mathbf{d}|\mathbf{m})p(\mathbf{m})$$

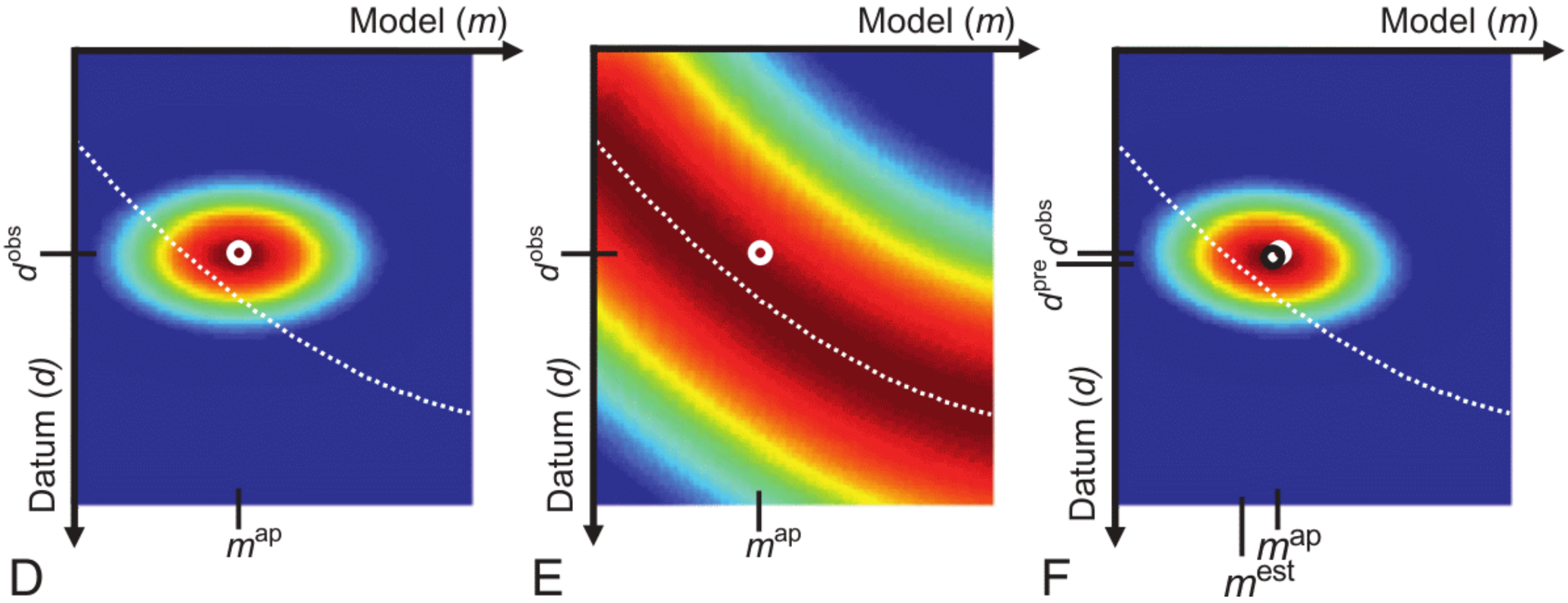
$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}$$

A priori (Menke, 2012)



A: *a priori pdf* $p_a(\mathbf{m}, \mathbf{d})$, B: *conditional pdf* $p_g(\mathbf{m}, \mathbf{g})$, C: product $p_t(\mathbf{m}, \mathbf{d}) = p_a(\mathbf{m}, \mathbf{d}) p_g(\mathbf{m}, \mathbf{d})$, white

A priori and


Model (m)
Datum (d)
d^{obs}
m^{ap}
D
Model (m)
Datum (d)
d^{obs}
m^{ap}
E
Model (m)
Datum (d)
d^{pre} d^{obs}
m^{ap}
m^{est}
F

Monte Carlo methods

Monte Carlos search: randomly draw solutions from grid

Markow-Chain-Monte-Carlo

Metropolis-Hastings