Lecture 8

The Principle of Maximum Likelihood

Syllabus

Lecture 01	Describing Inverse Problems
Lecture 02	Probability and Measurement Error, Part 1
Lecture 03	Probability and Measurement Error, Part 2
Lecture 04	The L_2 Norm and Simple Least Squares
Lecture 05	A Priori Information and Weighted Least Squared
Lecture 06	Resolution and Generalized Inverses
Lecture 07	Backus-Gilbert Inverse and the Trade Off of Resolution and Variance
Lecture 08	**The Principle of Maximum Likelihood**
Lecture 09	Inexact Theories
Lecture 10	Nonuniqueness and Localized Averages
Lecture 11	Vector Spaces and Singular Value Decomposition
Lecture 12	Equality and Inequality Constraints
Lecture 13	L_1 , L_∞ Norm Problems and Linear Programming
Lecture 14	Nonlinear Problems: Grid and Monte Carlo Searches
Lecture 15	Nonlinear Problems: Newton's Method
Lecture 16	Nonlinear Problems: Simulated Annealing and Bootstrap Confidence Intervals
Lecture 17	Factor Analysis
Lecture 18	Varimax Factors, Empircal Orthogonal Functions
Lecture 19	Backus-Gilbert Theory for Continuous Problems; Radon's Problem
Lecture 20	Linear Operators and Their Adjoints
Lecture 21	Fréchet Derivatives
Lecture 22	Exemplary Inverse Problems, incl. Filter Design
Lecture 23	Exemplary Inverse Problems, incl. Earthquake Location
Lecture 24	Exemplary Inverse Problems, incl. Vibrational Problems

Purpose of the Lecture

Introduce the spaces of all possible data,
all possible models and the idea of likelihood

Use maximization of likelihood as a guiding principle for solving inverse problems

Part 1

The spaces of all possible data, all possible models and the idea of *likelihood*

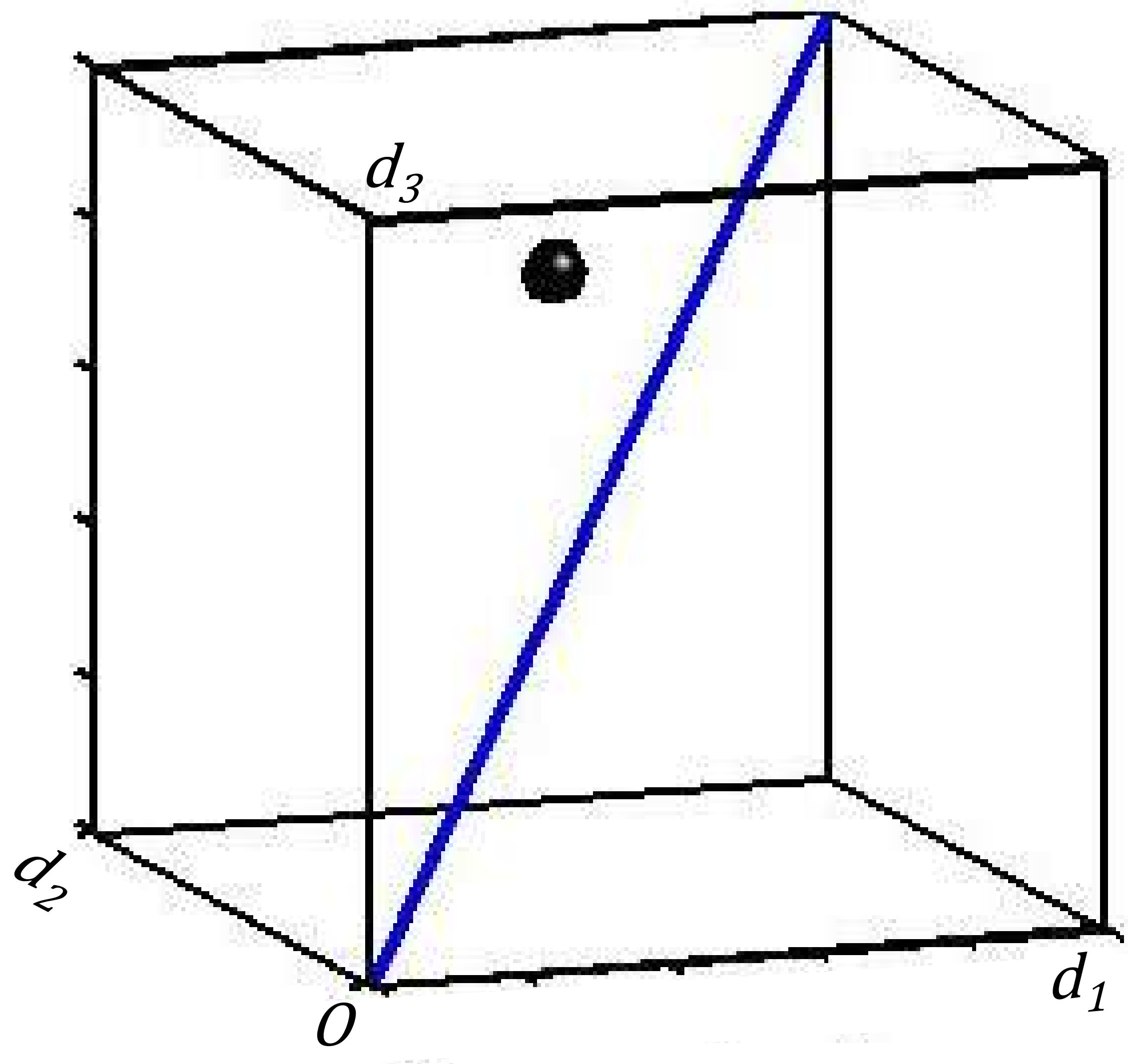
viewpoint

the observed data is one point in the space of all possible observations

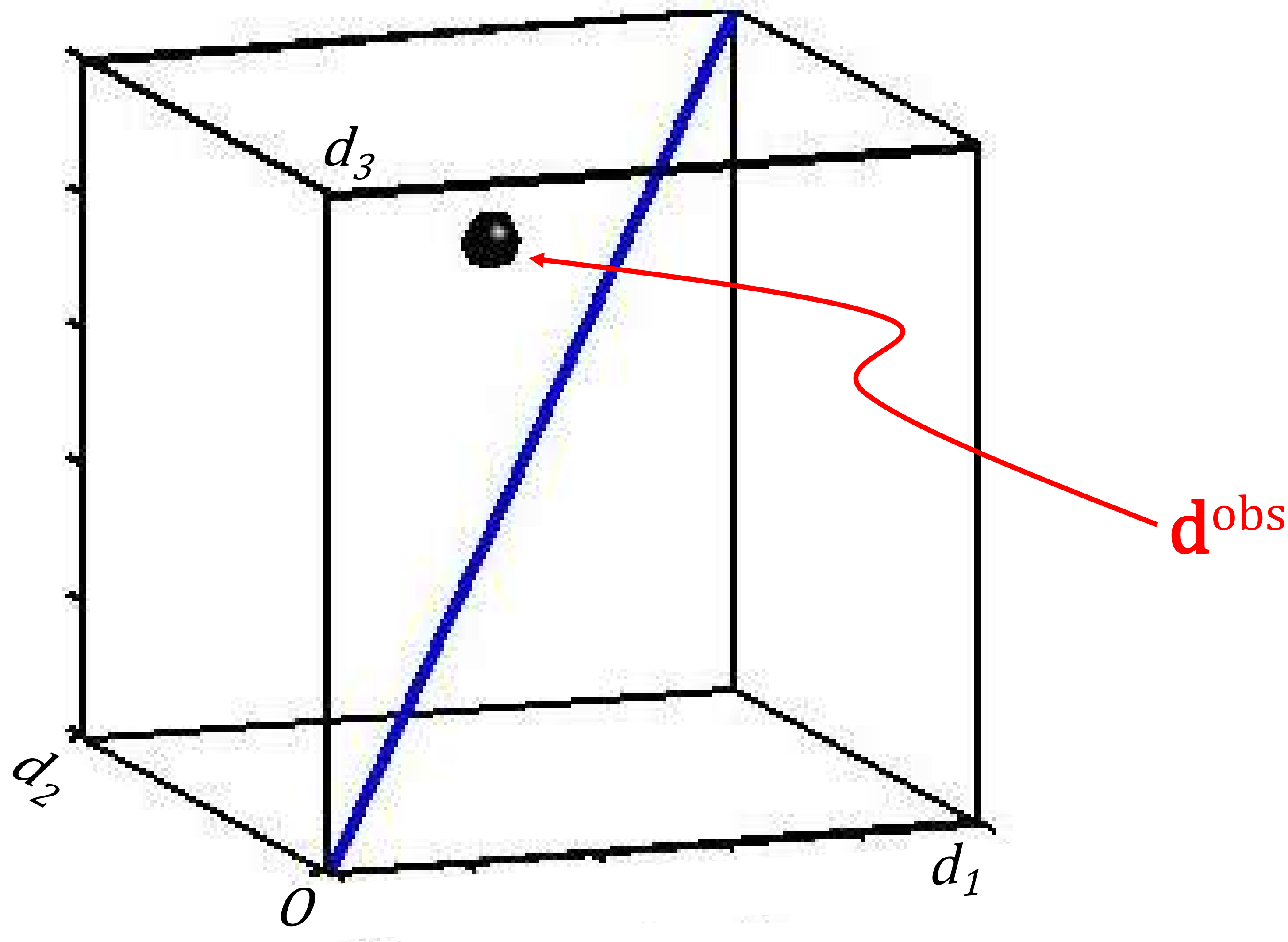
or

$\mathbf{d}^{obs}$ is a point in $S(\mathbf{d})$

plot of $\mathbf{d}^{obs}$



plot of $\mathbf{d}^{obs}$



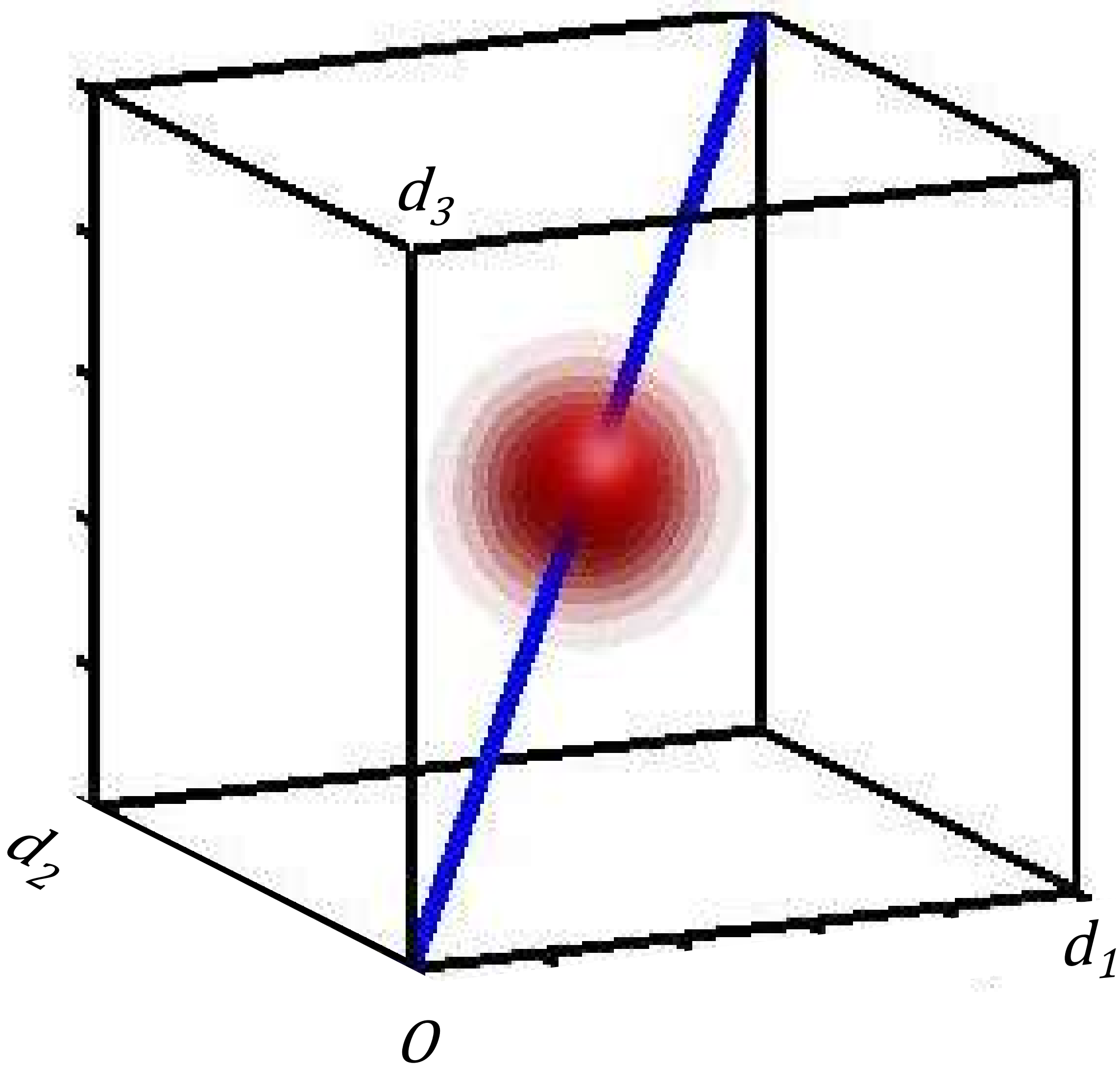
now suppose ...

the data are independent
each is drawn from a Gaussian distribution
with the same mean m_1 and variance σ^2

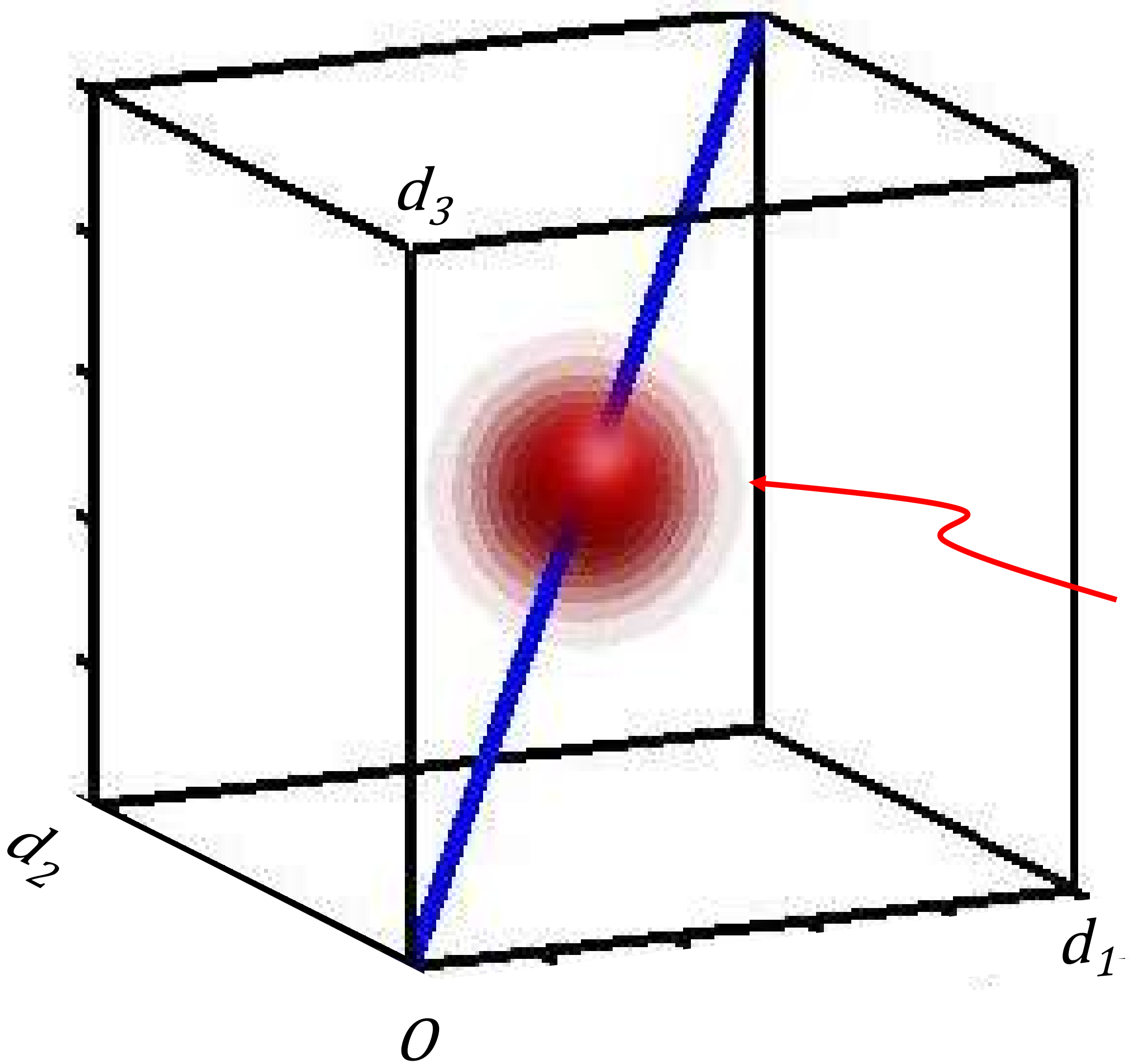
(but m_1 and σ unknown)

$$p(\mathbf{d}) = \sigma^{-N}(2\pi)^{-N/2}\exp\left[-\tfrac{1}{2}\sigma^{-2}\sum_{i=1}^{N}[d_i - m_1]^2\right]$$

plot of $p(\mathbf{d})$



plot of $p(\mathbf{d})$



cloud centered on $\mathbf{d_1=d_2=d_3}$ **with radius proportional to σ**

now interpret …

$$p(\mathbf{d}^{\text{obs}})$$

as the probability that the observed data was in fact observed

$$L = \log p(\mathbf{d}^{\text{obs}})$$

called the *likelihood*

find parameters in the distribution

maximize

p($\mathbf{d}^{obs}$)

with respect to m_1 and σ

maximize the probability that the observed data were in fact observed

the

Principle of Maximum Likelihood

Example

$$p(\mathbf{d}) = \sigma^{-N}(2\pi)^{-N/2}\exp\left[-½\sigma^{-2}\sum_{i=1}^{N}[d_i - m_1]^2\right]$$

$$L = \log(p(\mathbf{d}^{obs})) = -Nlog(\sigma) - ½\sigma^{-2}\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)^2$$

$$\frac{\partial L}{\partial m_1} = 0 = -½\sigma^{-2}2m_1\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)$$

$$\frac{\partial L}{\partial \sigma} = 0 = -\frac{N}{\sigma} + \sigma^{-3}\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)^2$$

solving the two equations

$$m_1^{est} = \frac{1}{N}\sum_{i=1}^{N} d_i^{obs} \quad \text{and} \quad \sigma^{est} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)^2\right]^{½}$$

solving the two equations

$$m_1^{est} = \frac{1}{N}\sum_{i=1}^{N} d_i^{obs} \quad \text{and} \quad \sigma^{est} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)^2\right]^{½}$$

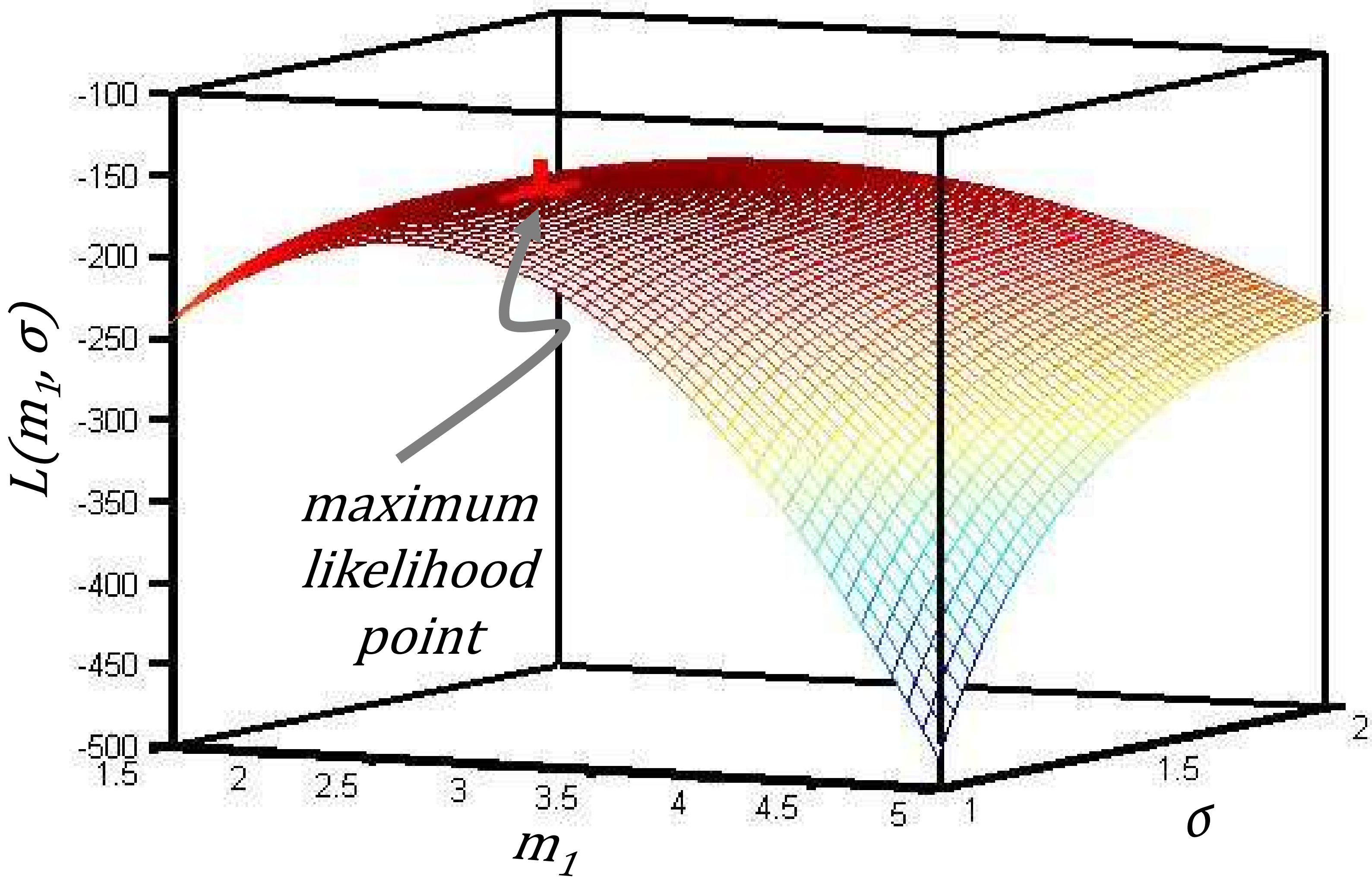
usual formula for the sample mean

almost the usual formula for the sample standard deviation these two estimates linked to the assumption of the data being Gaussian-distributed

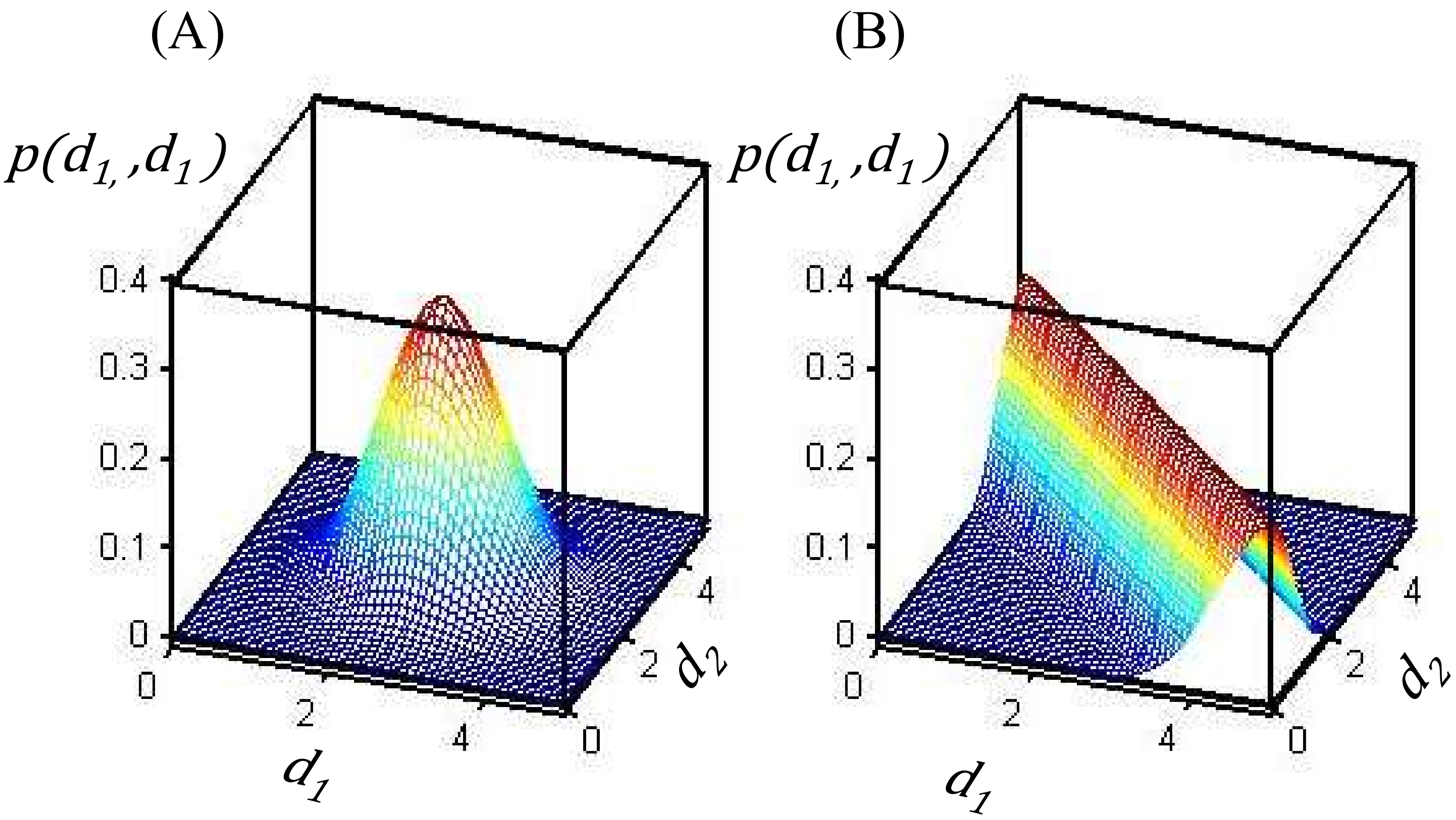
$$m_1^{est} = \frac{1}{N}\sum_{i=1}^{N} d_i^{obs} \quad \text{and} \quad \sigma^{est} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{obs} - m_1\right)^2\right]^{½}$$

might get a different formula for a different p.d.f.

example of a likelihood surface



likelihood maximization process will fail if p.d.f. has no well-defined peak



Part 2

Using the maximization of likelihood as a guiding principle for solving inverse problems

linear inverse problem for with Gaussian-distibuted data with known covariance [cov $\mathbf{d}$]

assume

$\mathbf{Gm}=\mathbf{d}$

gives the mean $\mathbf{d}$

$$p(\mathbf{d}) \propto \exp[-\tfrac{1}{2}(\mathbf{d}-\mathbf{Gm})^{\mathrm{T}}[\mathrm{cov}\,\mathbf{d}]^{-1}(\mathbf{d}-\mathbf{Gm})]$$

principle of maximum likelihood

maximize $L = \log p(\mathbf{d}^{\text{obs}})$

minimize

$$(\mathbf{d}^{\text{obs}} - \mathbf{Gm})^{\text{T}}[\text{cov}\,\mathbf{d}]^{-1}(\mathbf{d}^{\text{obs}} - \mathbf{Gm})$$

with respect to $\mathbf{m}$

principle of maximum likelihood

maximize $L = \log p(\mathbf{d}^{\text{obs}})$

minimize

$$\mathrm{E} = (\mathbf{d}^{\text{obs}} - \mathbf{Gm})^{\mathrm{T}}[\operatorname{cov}\mathbf{d}]^{-1}(\mathbf{d}^{\text{obs}} - \mathbf{Gm})$$

This is just weighted least squares

principle of maximum likelihood

when data Gaussian-distributed
solve **Gm=d** with weighted least squares

with weighting of

$$[\mathrm{cov}\,\mathbf{d}]^{-1}$$

special case of uncorrelated data
each datum with a different variance

$$[\mathrm{cov}\,\mathbf{d}]_{\mathrm{ii}} = \sigma_{di}^{\;2}$$

minimize

$$E = \sum_{i=1}^{N} \sigma_{di}^{-2} e_i^2$$

special case of uncorrelated data
each datum with a different variance

$$[\text{cov}\,\mathbf{d}]_{ii} = \sigma_{di}^{2}$$

minimize

$$E = \sum_{i=1}^{N} \sigma_{di}^{-2} e_i^2$$

errors
weighted by
their *certainty*

but what about a priori information?

probabilistic representation of a priori information

probability that the model parameters are near **m**
given by p.d.f.

$$p_A(\mathbf{m})$$

probabilistic representation of a priori information

probability that the model parameters are near $\mathbf{m}$

given by p.d.f.

$$p_A(\mathbf{m})$$

centered at a priori value $\langle \mathbf{m} \rangle$

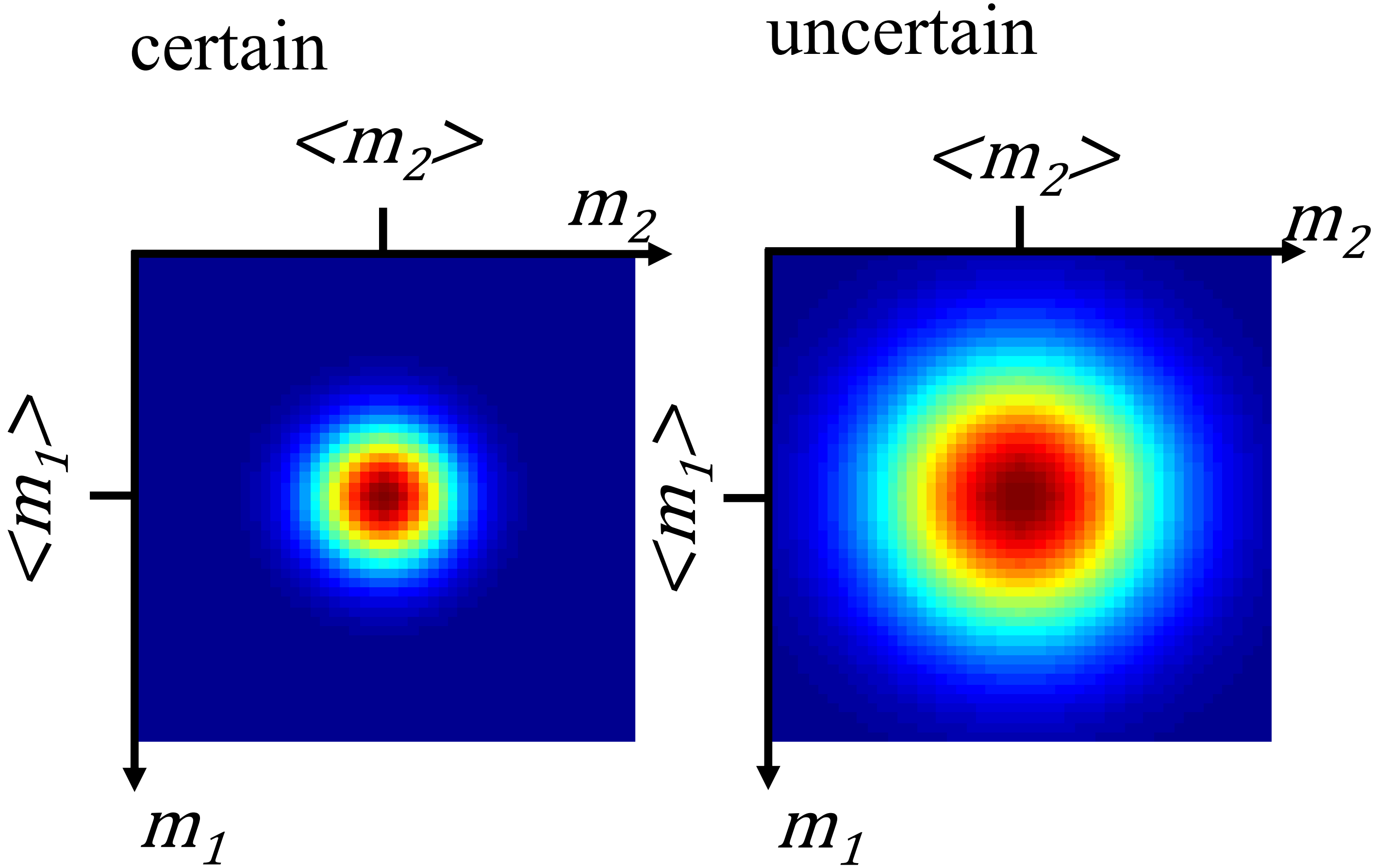
probabilistic representation of a priori information

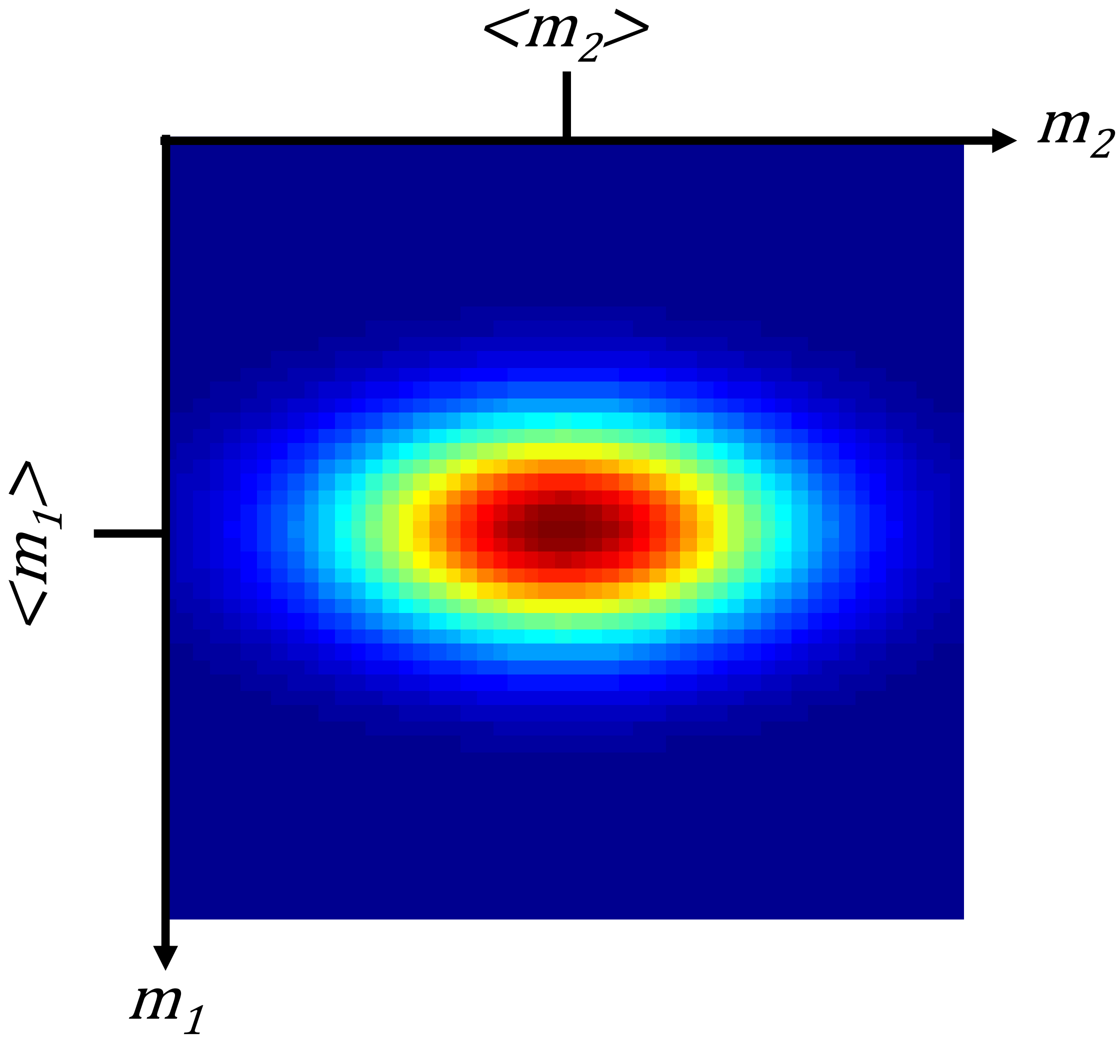
probability that the model parameters are near $\mathbf{m}$

given by p.d.f.

$$p_A(\mathbf{m})$$

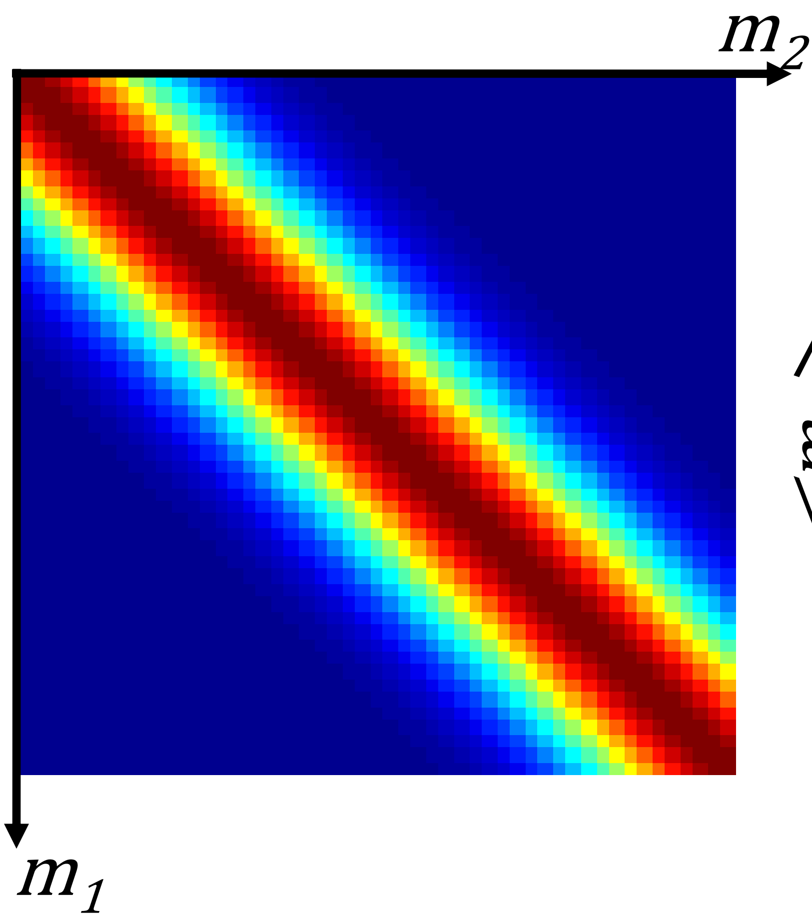
variance reflects uncertainty in a priori information


certain
uncertain
$<m_2>$
m_2
$<m_1>$
m_1
$<m_2>$
m_2
$<m_1>$
m_1

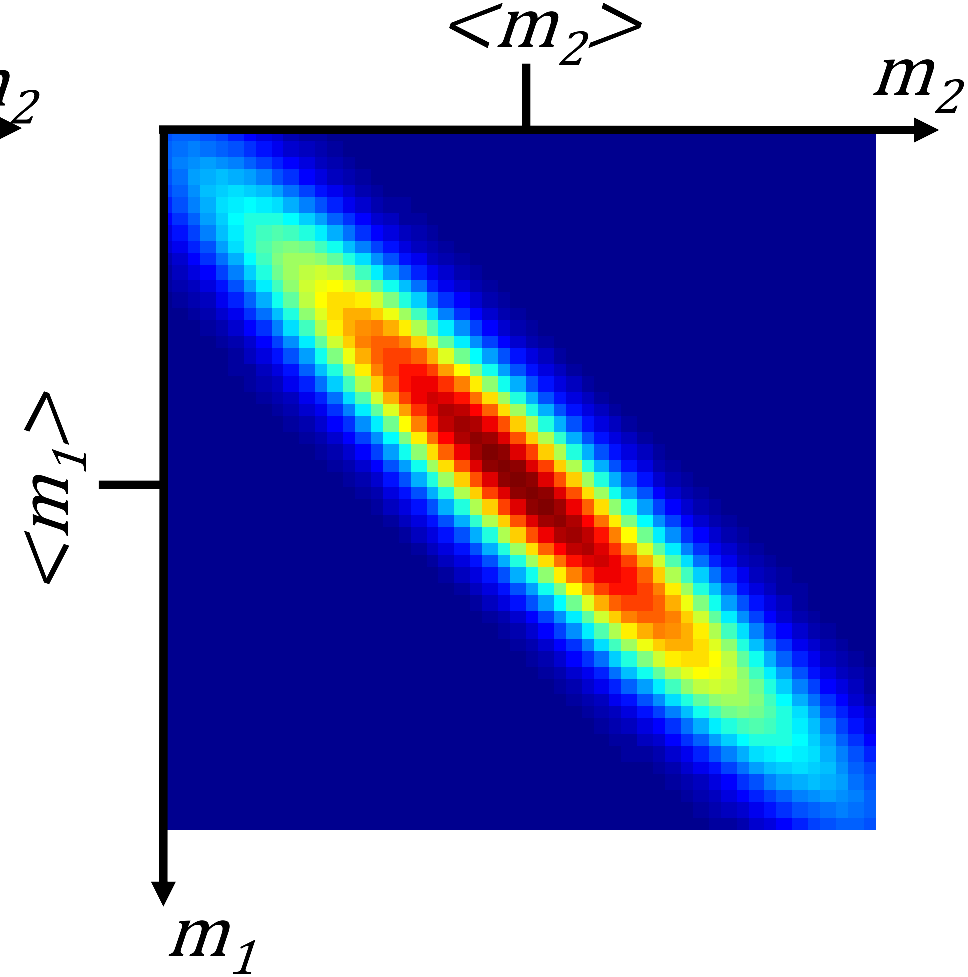


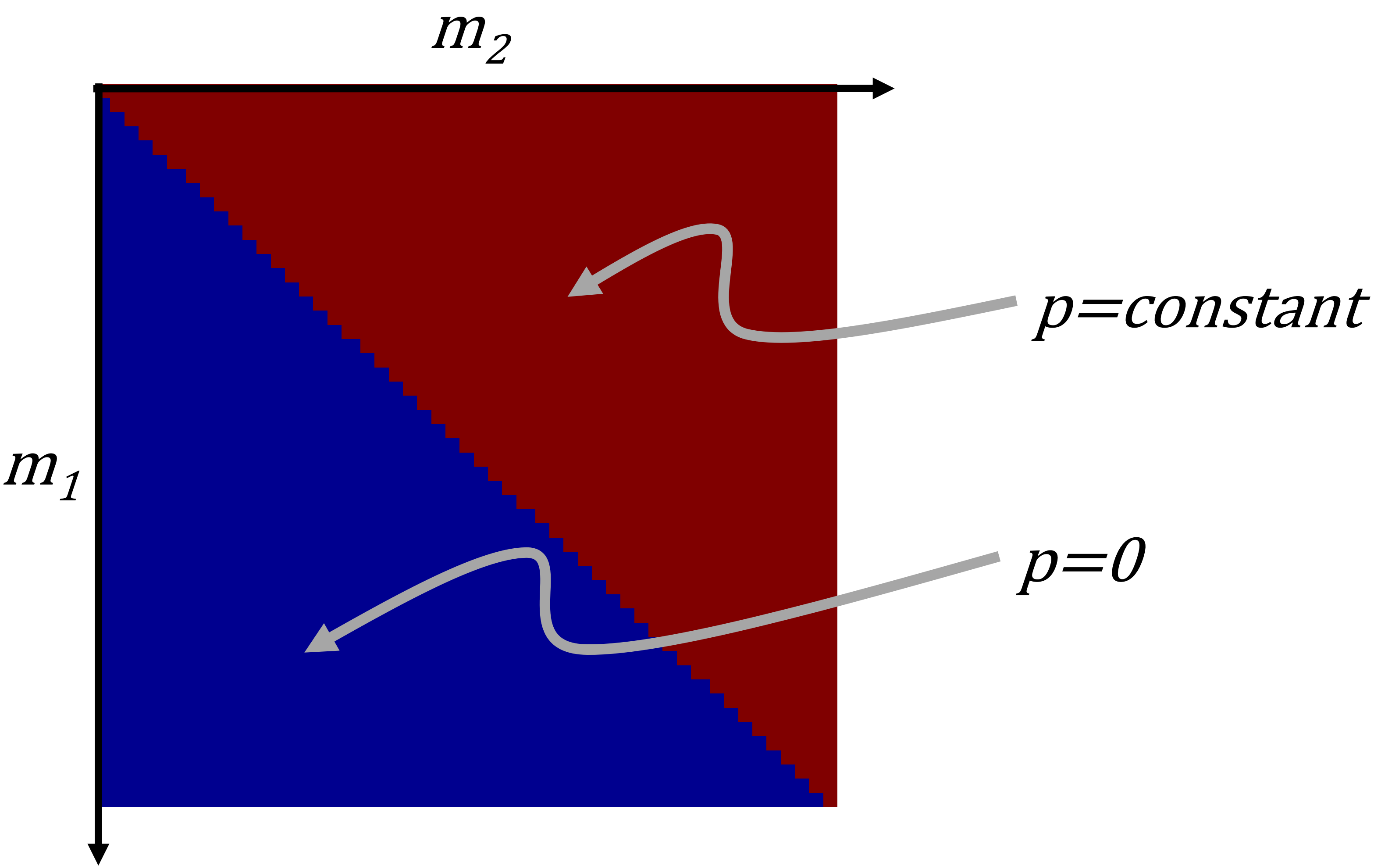
$\langle m_2 \rangle$
m_2
$\langle m_1 \rangle$
m_1

linear relationship



approximation with Gaussian





m_2
m_1
p=constant
p=0

assessing the information content in $p_A(\mathbf{m})$

Do we know a little about $\mathbf{m}$

or

a lot about $\mathbf{m}$?

Information Gain, **S**

$$S[p_A(\mathbf{m})] = \int p_A(\mathbf{m}) \log\left[\frac{p_A(\mathbf{m})}{p_N(\mathbf{m})}\right] \mathrm{d}^M \mathbf{m}$$

-*S* called Relative Entropy,

Relative Entropy, S
also called Information Gain

$$S[p_A(\mathbf{m})] = \int p_A(\mathbf{m}) \log\left[\frac{p_A(\mathbf{m})}{p_N(\mathbf{m})}\right] \mathrm{d}^M\mathbf{m}$$

null p.d.f.
state of no knowledge

Relative Entropy, S
also called Information Gain

$$S[p_A(\mathbf{m})] = \int p_A(\mathbf{m}) \log\left[\frac{p_A(\mathbf{m})}{p_N(\mathbf{m})}\right] \mathrm{d}^{\mathrm{M}}\mathbf{m}$$

uniform p.d.f. might
work for this

probabilistic representation of data

probability that the data are

near **d**

given by p.d.f.

$$p_A(\mathbf{d})$$

probabilistic representation of data

probability that the data are
near $\mathbf{d}$
given by p.d.f.

$$p(\mathbf{d})$$

centered at
observed data
d^{obs}

probabilistic representation of data

probability that the data are near **d** given by p.d.f.

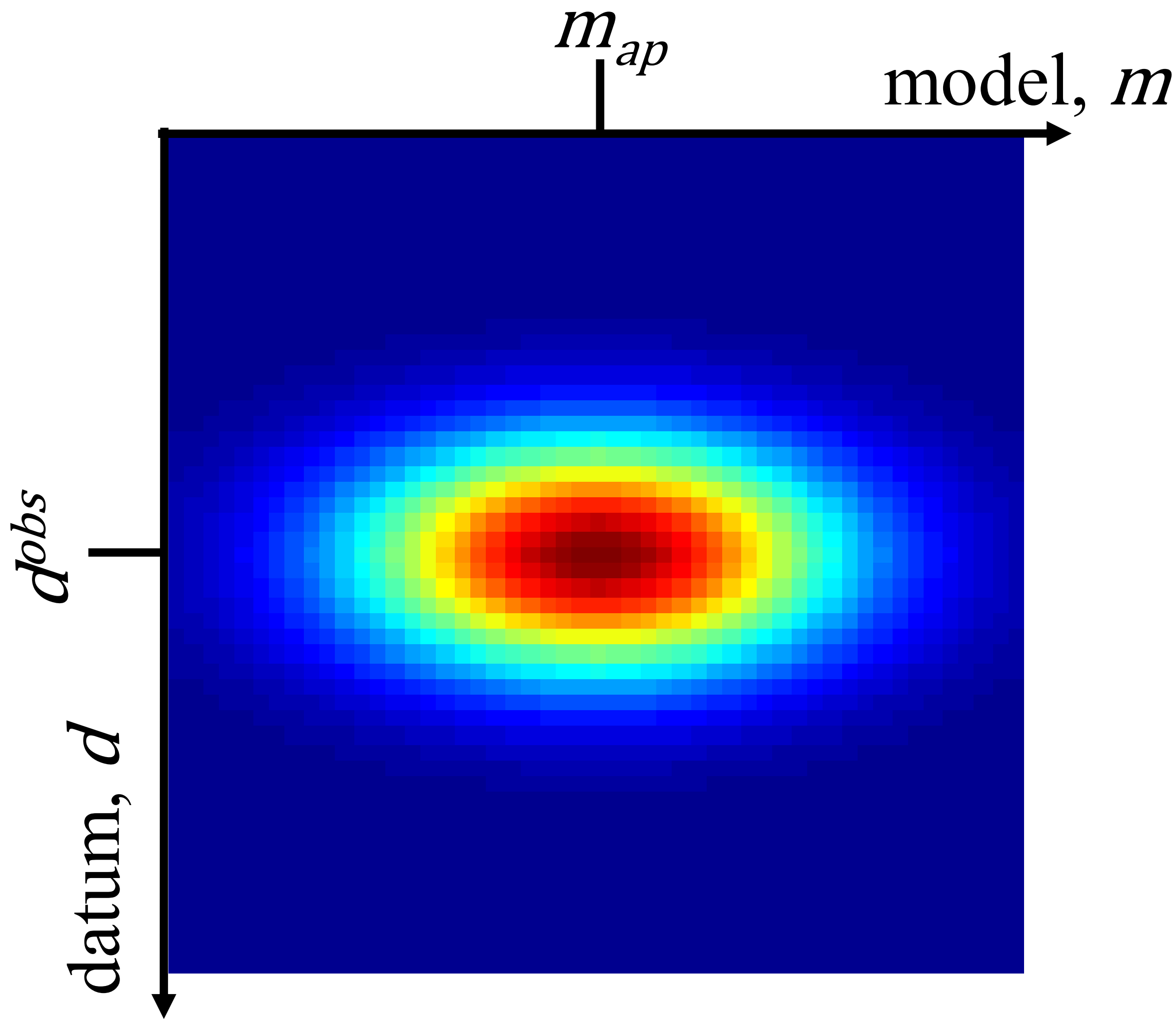
$$p(\mathbf{d})$$

variance reflects uncertainty in measurements probabilistic representation of both prior information and observed data

assume observations and a priori information are uncorrelated

$$p_A(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m})p_A(\mathbf{d})$$

Example of $p_A(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m})p_A(\mathbf{d})$



the theory

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

is a surface in the combined space of data and model parameters

on which the estimated model parameters and predicted data must lie the theory

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

is a surface in the combined space of data and model parameters

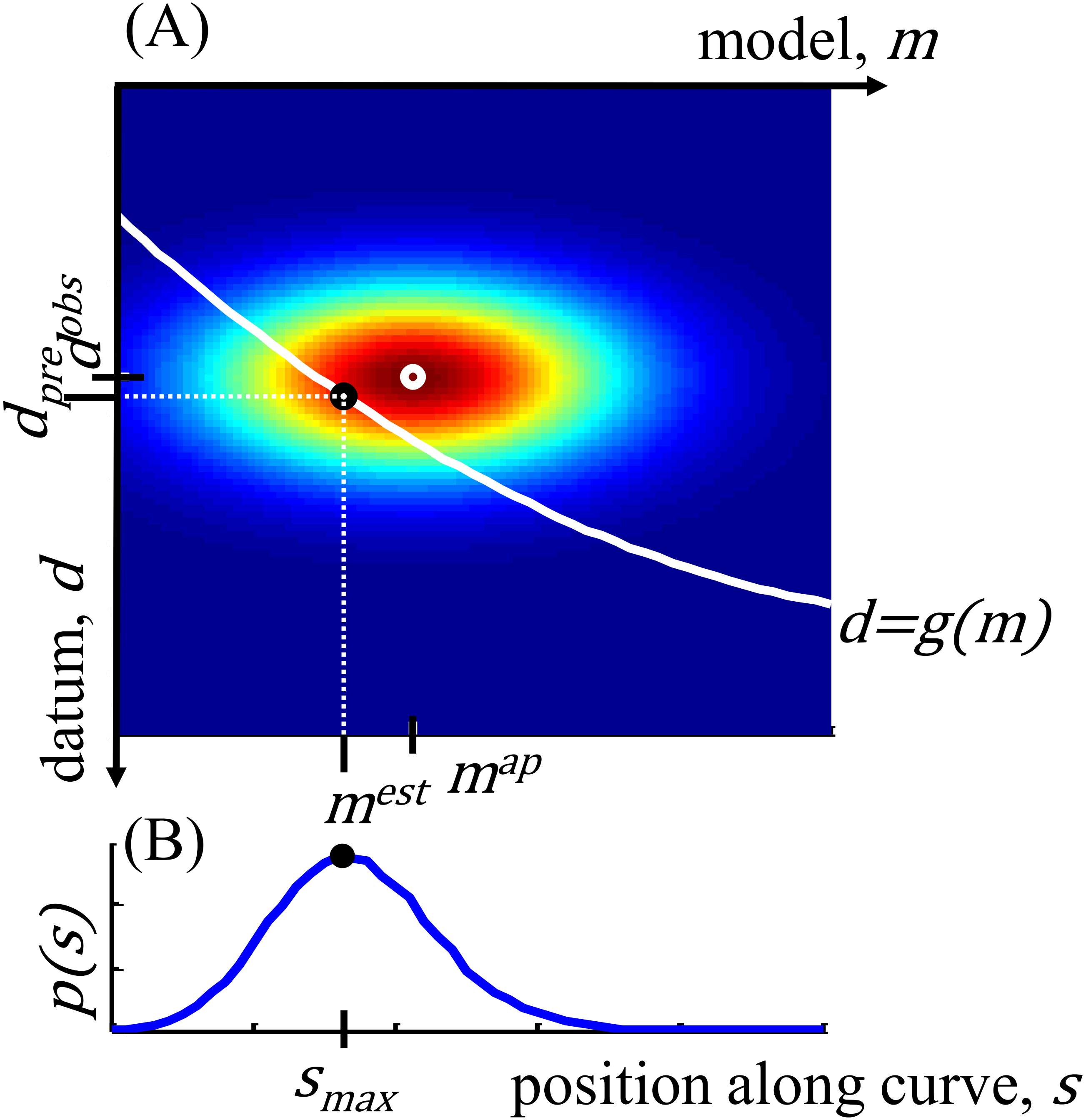
on which the estimated model parameters and predicted data must lie

for a linear theory
the surface is planar

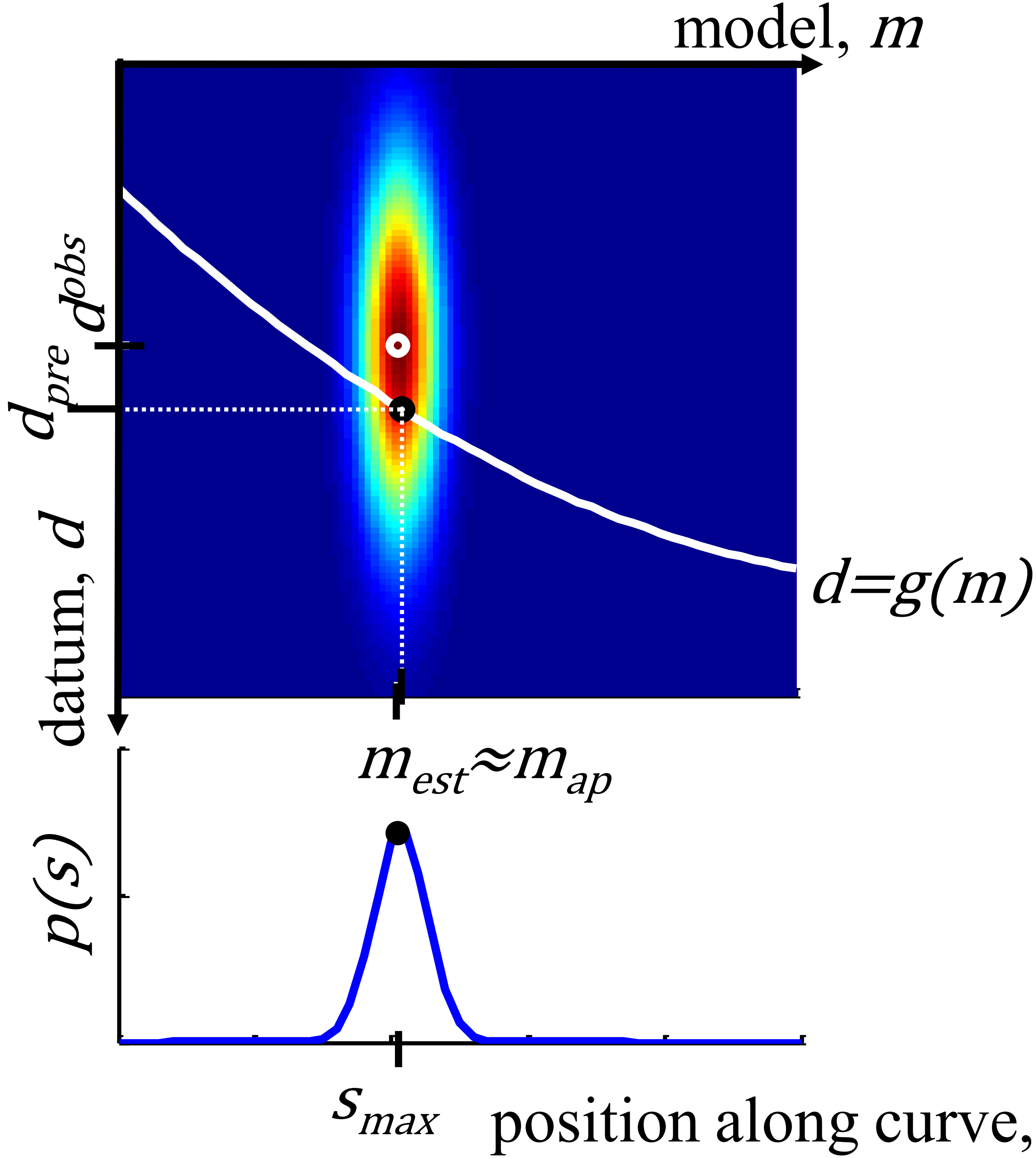
the principle of maximum likelihood says

maximize $p_A(\mathbf{m}, \mathbf{d}) = p_A(\mathbf{m})p_A(\mathbf{d})$

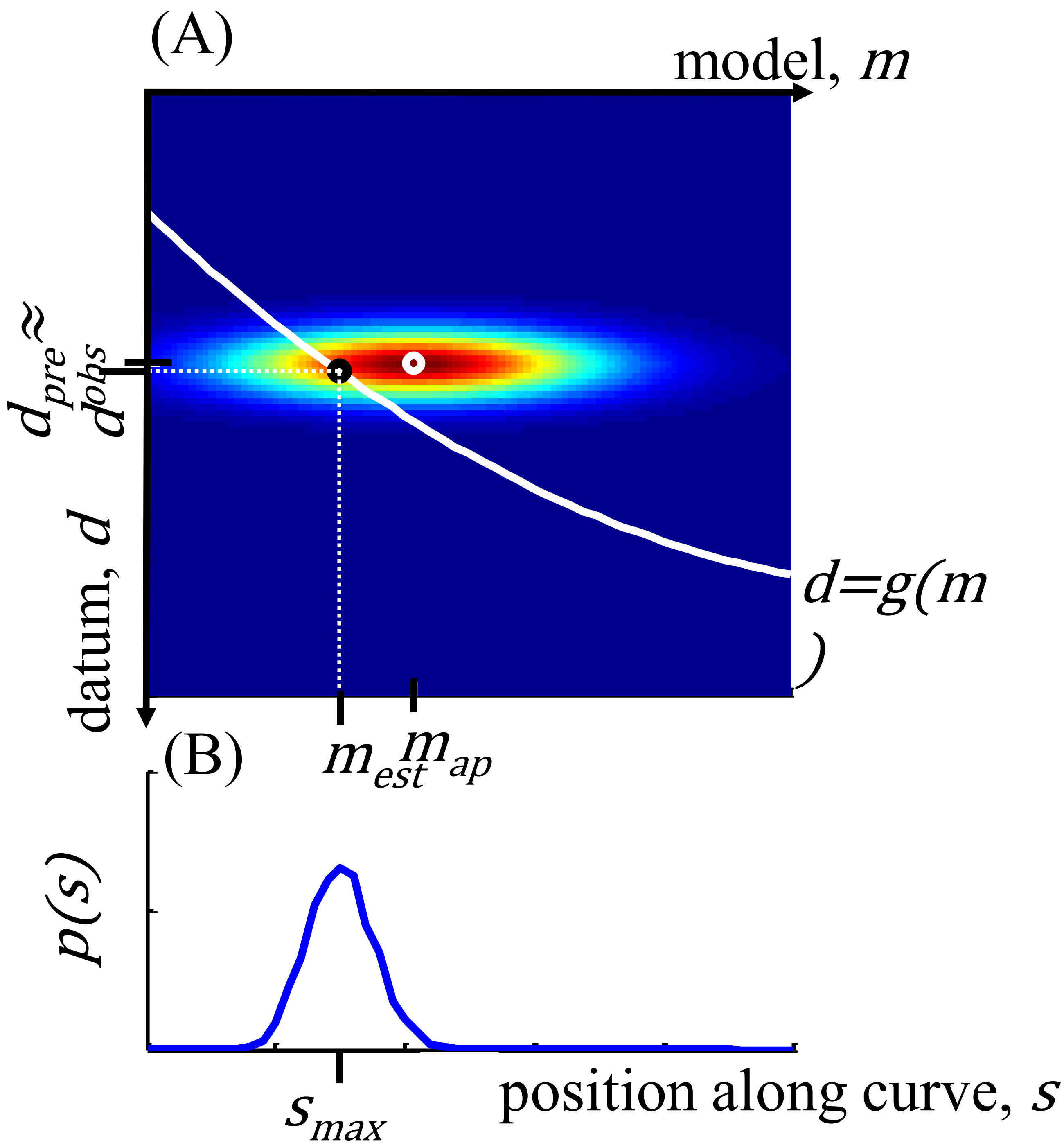
on the surface $\mathbf{d}=\mathbf{g}(\mathbf{m})$



(A)
model, m
d_{pre}
d^{obs}
datum, d
$d=g(m)$
m^{est}
m^{ap}
(B)
$p(s)$
s_{max}
position along curve, s



model, m
d^{obs}
d_{pre}
datum, d
$d=g(m)$
$m_{est} \approx m_{ap}$
$p(s)$
s_{max}
position along curve, s



(A)
model, m
d_pre ≈ d_obs
datum, d
d=g(m)
m_est
m_ap
(B)
p(s)
s_max
position along curve, s

principle of maximum likelihood
with
Gaussian-distributed data
Gaussian-distributed a priori information

minimize $\Phi(\mathbf{m}) = L(\mathbf{m}) + \mathrm{E}(\mathbf{m})$ with respect to $\mathbf{m}$ with

$$L(\mathbf{m}) = (\mathbf{m} - \langle\mathbf{m}\rangle)^{\mathrm{T}}[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}}^{-1}(\mathbf{m} - \langle\mathbf{m}\rangle)$$

$$E(\mathbf{m}) = \left(\mathbf{Gm} - \mathbf{d}^{\mathrm{obs}}\right)^{\mathrm{T}}[\mathrm{cov}\,\mathbf{d}]^{-1}\left(\mathbf{Gm} - \mathbf{d}^{\mathrm{obs}}\right)$$

this is just weighted least squares with

$$\varepsilon^2 \mathbf{W}_m = [\mathrm{cov}\,\mathbf{m}]^{-1} \quad \text{and} \quad \mathbf{W}_e = [\mathrm{cov}\,\mathbf{d}]^{-1}$$

so we already know the solution

solve **Fm=f** with simple least squares

$$\mathbf{F} = \begin{bmatrix} [\operatorname{cov}\mathbf{d}]^{-1/2}\mathbf{G} \\ [\operatorname{cov}\mathbf{m}]_A^{-1/2}\mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} [\operatorname{cov}\mathbf{d}]^{-1/2}\mathbf{d}^{\mathbf{obs}} \\ [\operatorname{cov}\mathbf{m}]_A^{-1/2}\langle\mathbf{m}\rangle \end{bmatrix}$$

when $[\text{cov}\,\mathbf{d}]=\sigma_d^2\mathbf{I}$ and $[\text{cov}\,\mathbf{m}]=\sigma_m^2\mathbf{I}$

$$\mathbf{F} = \begin{bmatrix} [\text{cov}\,\mathbf{d}]^{-1/2}\mathbf{G} \\ [\text{cov}\,\mathbf{m}]_A^{-1/2}\mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} [\text{cov}\,\mathbf{d}]^{-1/2}\mathbf{d}^{\mathbf{obs}} \\ [\text{cov}\,\mathbf{m}]_A^{-1/2}\langle\mathbf{m}\rangle \end{bmatrix}$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{G} \\ \varepsilon\mathbf{I} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} \mathbf{d}^{\mathbf{obs}} \\ \varepsilon\langle\mathbf{m}\rangle \end{bmatrix} \quad \text{with} \quad \varepsilon^2 = \frac{\sigma_d^2}{\sigma_m^2}$$

this provides and answer to the question

What should be the value of ε^2 in damped least squares?

The answer

$$\varepsilon^2 = \frac{\sigma_d^2}{\sigma_m^2}$$

it should be set to the ratio of variances of the data and the a priori model parameters if the a priori information is

$$\mathbf{Hm}=\mathbf{h}$$

with covariance $[\text{cov}\,\mathbf{h}]_A$

then the $\mathbf{Fm}=\mathbf{f}$ becomes

$$\mathbf{F}=\begin{bmatrix}[\text{cov}\,\mathbf{d}]^{-1/2}\mathbf{G}\\ [\text{cov}\,\mathbf{h}]_A^{-1/2}\mathbf{H}\end{bmatrix} \quad \text{and} \quad \mathbf{f}=\begin{bmatrix}[\text{cov}\,\mathbf{d}]^{-1/2}\mathbf{d}^{\text{obs}}\\ [\text{cov}\,\mathbf{h}]_A^{-1/2}\mathbf{h}\end{bmatrix}$$

the most useful formula in inverse theory

$\mathbf{Gm}=\mathbf{d}^{\text{obs}}$ with covariance $[\text{cov}\ \mathbf{d}]$

$\mathbf{Hm}=\mathbf{h}$ with covariance $[\text{cov}\ \mathbf{h}]_A$

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

with

$$\mathbf{F} = \begin{bmatrix} [\text{cov}\ \mathbf{d}]^{-\frac{1}{2}}\mathbf{G} \\ [\text{cov}\ \mathbf{h}]_A^{-\frac{1}{2}}\mathbf{H} \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} [\text{cov}\ \mathbf{d}]^{-\frac{1}{2}}\mathbf{d}^{\text{obs}} \\ [\text{cov}\ \mathbf{h}]_A^{-\frac{1}{2}}\mathbf{h} \end{bmatrix}$$