Lecture 9

Inexact Theories

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, Empirical 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

Discuss how an inexact theory can be represented

Solve the inexact, linear Gaussian inverse problem

Use maximization of relative entropy as a guiding principle for solving inverse problems

Introduce F-test as way to determine whether one solution is "better" than another

Part 1

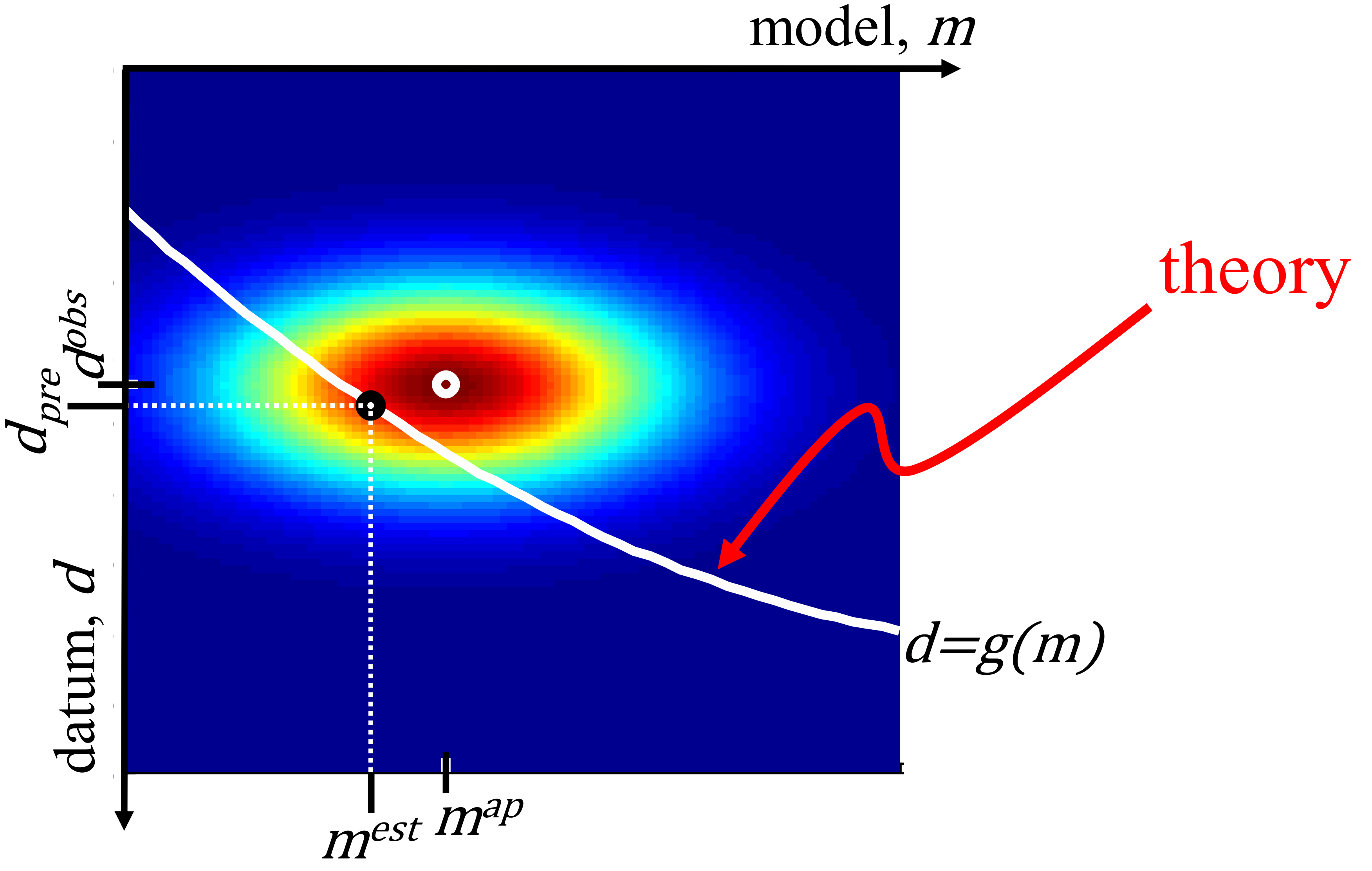
How Inexact Theories can be Represented

How do we generalize the case of

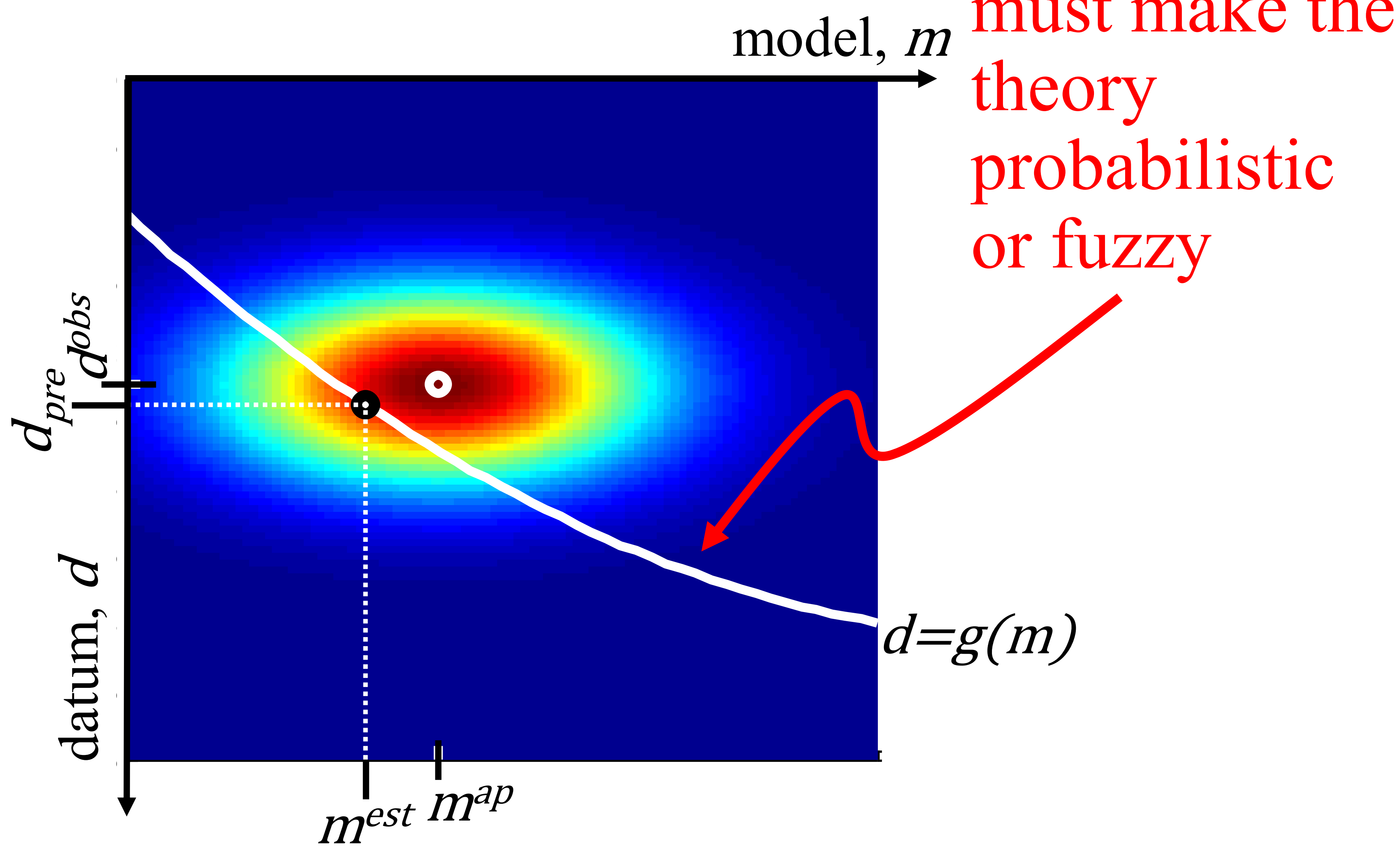
an exact theory

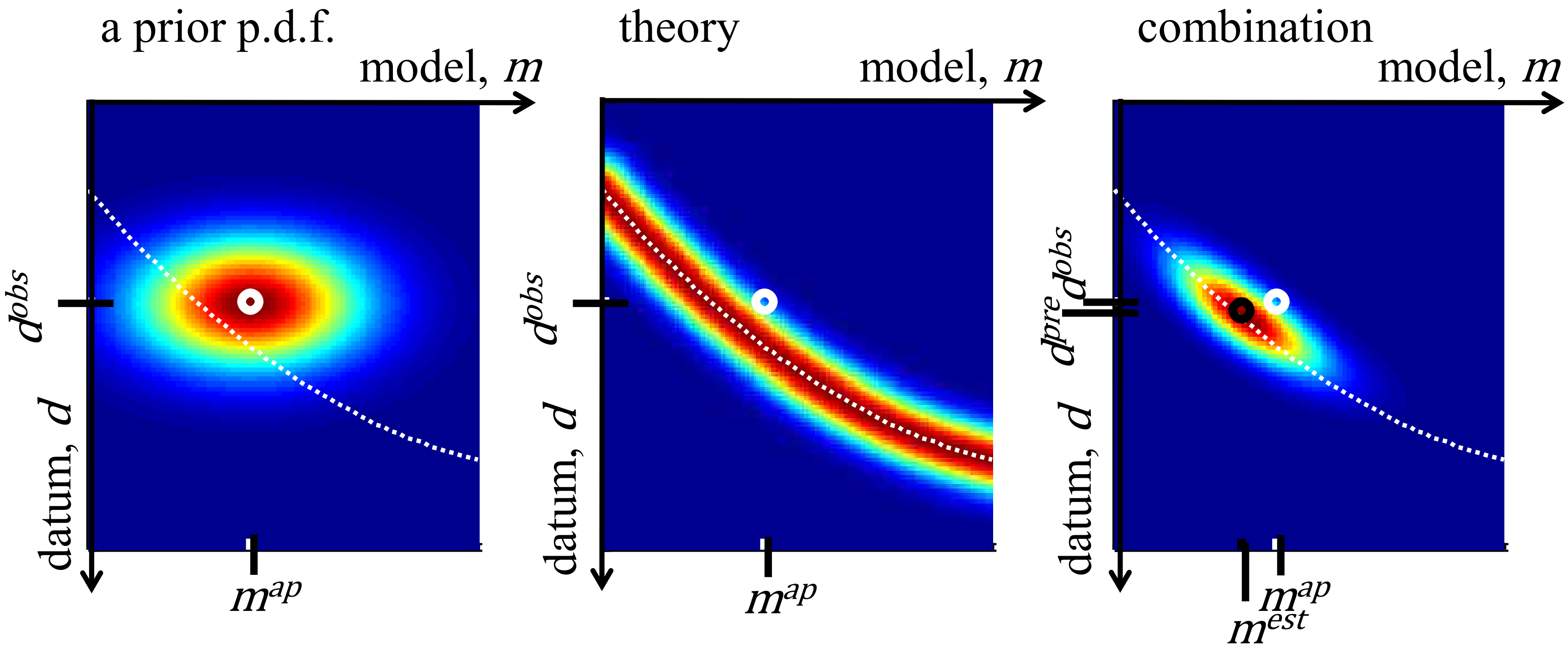
to one that is inexact?

exact theory case



to make theory inexact ...




a prior p.d.f.
model, m
d^obs
datum, d
m^ap
theory
model, m
d^obs
datum, d
m^ap
combination
model, m
d^pre
d^obs
datum, d
m^ap
m^est

how do you

combine

two probability density functions ?

how do you
combine
two probability density functions ?

so that the information in them is combined ...

desirable properties

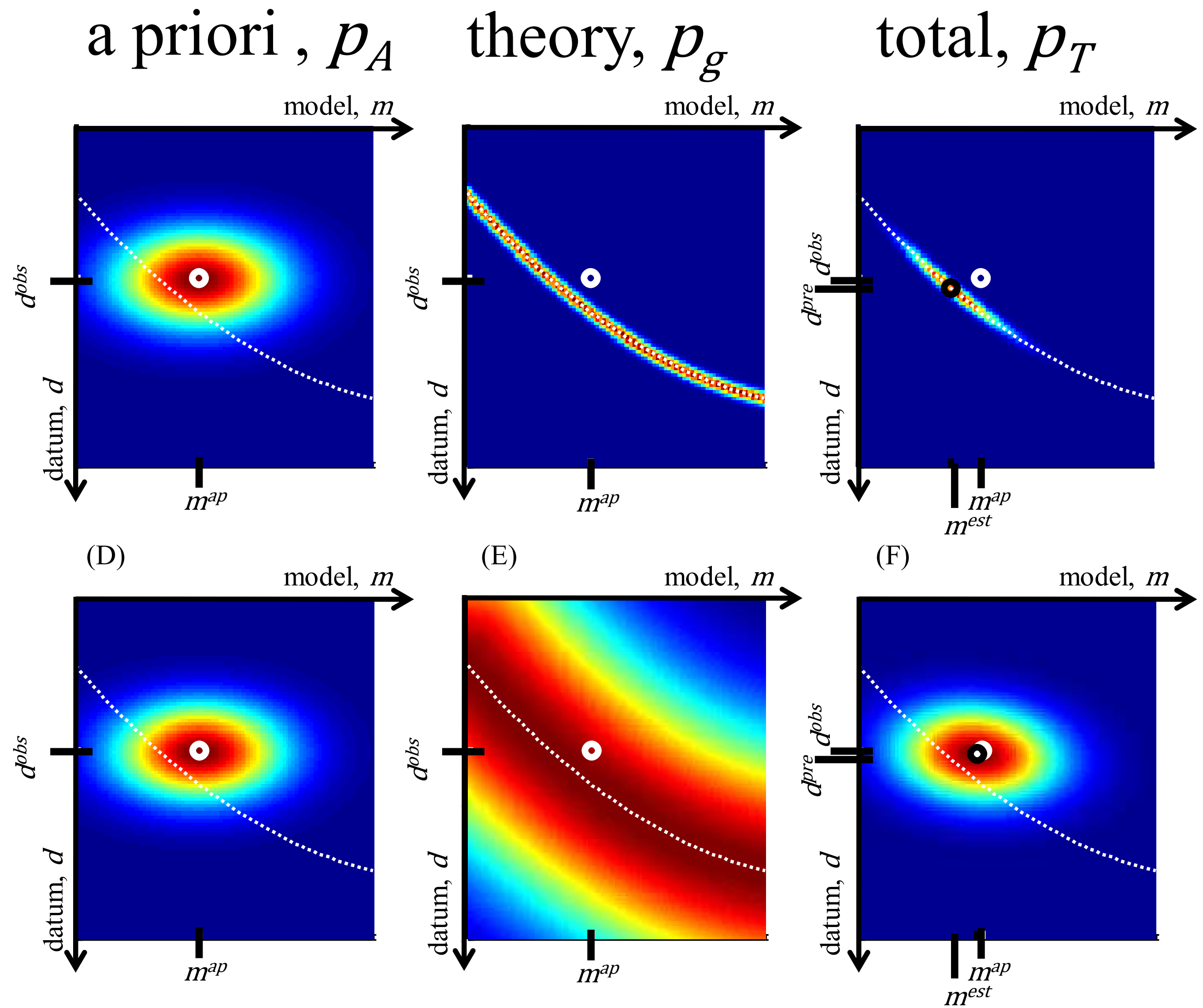
order shouldn’t matter

combining something with the null distribution should leave it unchanged

combination should be invariant under change of variables

Answer

$$p_3 = \frac{p_1 p_2}{p_N}$$



a priori , p_A
theory, p_g
total, p_T
model, m
model, m
model, m
d^{obs}
d^{obs}
d^{obs}
d^{pre}
datum, d
datum, d
datum, d
m^{ap}
m^{ap}
m^{ap}
m^{est}
(D)
(E)
(F)
model, m
model, m
model, m
d^{obs}
d^{obs}
d^{obs}
d^{pre}
datum, d
datum, d
datum, d
m^{ap}
m^{ap}
m^{ap}
m^{est}

“solution to inverse problem”
maximum likelihood point of

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

(with $p_N \propto constant$)

simultaneously gives
$\mathbf{m}^{est}$ and $\mathbf{d}^{pre}$

$$p_T(\mathbf{m}, \mathbf{d})$$

probability that the estimated model parameters are near **m** and the predicted data are near **d**

$$p_p(\mathbf{m}) = \int p_T(\mathbf{m}, \mathbf{d})\ \mathrm{d}^{\mathrm{N}}\mathbf{d}$$

probability that the estimated model parameters are near **m** irrespective of the value of the predicted data

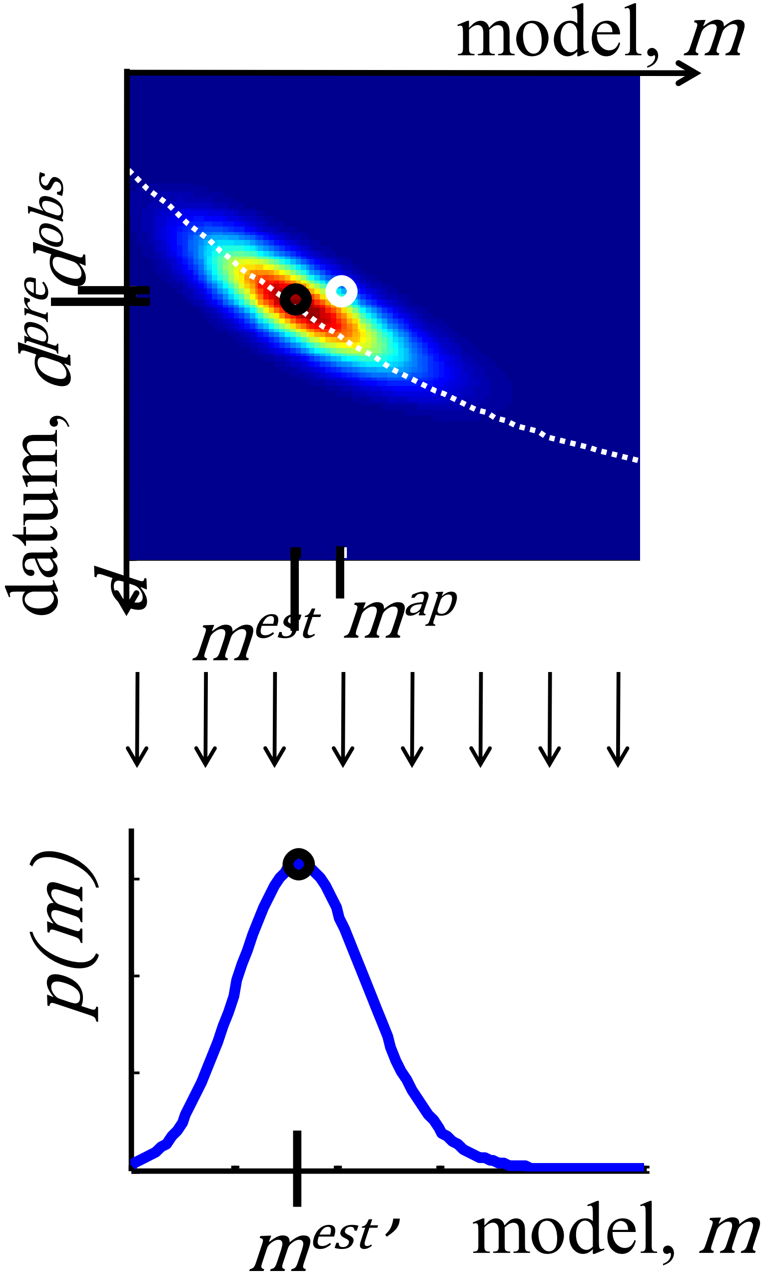
conceptual problem

$$p_T(\mathbf{m}, \mathbf{d})$$

and

$$p_p(\mathbf{m}) = \int p_T(\mathbf{m}, \mathbf{d})\, \mathrm{d}^N \mathbf{d}$$

do not necessarily have maximum likelihood points at the same value of **m**



model, m
datum, d
d^pre
d^obs
m^est
m^ap
p(m)
m^est'
model, m

illustrates the problem in defining a
definitive solution
to an inverse problem

illustrates the problem in defining a **definitive solution** to an inverse problem

fortunately
if all distributions are Gaussian
the two points are the same

Part 2

Solution of the inexact linear Gaussian inverse problem

Gaussian a priori information

$$p_A(\mathbf{m}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{m} - \langle\mathbf{m}\rangle)^{\mathrm{T}}[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}}^{-1}(\mathbf{m} - \langle\mathbf{m}\rangle)\right]$$

Gaussian a priori information

$$p_A(\mathbf{m}) \propto \exp\left[-\tfrac{1}{2}(\mathbf{m} - \langle\mathbf{m}\rangle)^{\mathrm{T}}[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}}^{-1}(\mathbf{m} - \langle\mathbf{m}\rangle)\right]$$

a priori values of model parameters

their uncertainty

Gaussian observations

$$p_A(\mathbf{d}) \propto \exp\left[-\tfrac{1}{2}\left(\mathbf{d}-\mathbf{d}^{\mathrm{obs}}\right)^{\mathrm{T}}[\operatorname{cov}\mathbf{d}]^{-1}\left(\mathbf{d}-\mathbf{d}^{\mathrm{obs}}\right)\right]$$

Gaussian observations

$$p_A(\mathbf{d}) \propto \exp\left[-\tfrac{1}{2}\left(\mathbf{d}-\mathbf{d}^{\text{obs}}\right)^{T}[\text{cov}\,\mathbf{d}]^{-1}\left(\mathbf{d}-\mathbf{d}^{\text{obs}}\right)\right]$$

observed data

measurement error

Gaussian theory

$$p_g(\mathbf{m}, \mathbf{d}) \propto \exp[-½(\mathbf{d} - \mathbf{Gm})^{\mathrm{T}}[\mathrm{cov}\,\mathbf{g}]^{-1}(\mathbf{d} - \mathbf{Gm})]$$

Gaussian theory

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

linear theory

uncertainty in theory

mathematical statement of problem

find $(\mathbf{m},\mathbf{d})$ that maximizes

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

and, along the way, work out the form of $p_T(\mathbf{m},\mathbf{d})$

notational simplification

group **m** and **d** into single vector $\mathbf{x} = [\mathbf{d}^T, \mathbf{m}^T]^T$

group $[\text{cov}\ \mathbf{m}]_A$ and $[\text{cov}\ \mathbf{d}]_A$ into single matrix

$$[\text{cov}\ \mathbf{x}] = \begin{bmatrix} [\text{cov}\ \mathbf{d}] & 0 \\ 0 & [\text{cov}\ \mathbf{m}]_A \end{bmatrix}$$

write **d-Gm=0** as **Fx=0** with **F=[I, –G]**

after much algebra, we find
$p_T(\mathbf{x})$ is a Gaussian distribution

with mean

$$\mathbf{x}^* = \left\{\mathbf{I} - [\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\left[[\mathrm{cov}\,\mathbf{g}] + \mathbf{F}[\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\right]^{-1}\mathbf{F}\right\}\langle\mathbf{x}\rangle$$

and variance

$$[\mathrm{cov}\,\mathbf{x}^*] = \left\{\mathbf{I} - [\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\left[[\mathrm{cov}\,\mathbf{g}] + \mathbf{F}[\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\right]^{-1}\mathbf{F}\right\}[\mathrm{cov}\,\mathbf{x}]$$

after much algebra, we find
$p_T(\mathbf{x})$ is a Gaussian distribution

with mean

$$\mathbf{x}^* = \left\{\mathbf{I} - [\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\left[[\mathrm{cov}\,\mathbf{g}] + \mathbf{F}[\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\right]^{-1}\mathbf{F}\right\}\langle\mathbf{x}\rangle$$

solution to inverse problem

and variance

$$[\mathrm{cov}\,\mathbf{x}^*] = \left\{\mathbf{I} - [\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\left[[\mathrm{cov}\,\mathbf{g}] + \mathbf{F}[\mathrm{cov}\,\mathbf{x}]\mathbf{F}^T\right]^{-1}\mathbf{F}\right\}[\mathrm{cov}\,\mathbf{x}]$$

after pulling $\mathbf{m}^{\text{est}}$ out of $\mathbf{x}^*$

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle \mathbf{m} \rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle \mathbf{m} \rangle$$

$$\text{with} \quad \mathbf{G}^{-g} = [\text{cov}\,\mathbf{m}]_A \mathbf{G}^T\{[\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] + \mathbf{G}[\text{cov}\,\mathbf{m}]_A \mathbf{G}^T\}^{-1}$$

after pulling $\mathbf{m}^{est}$ out of $\mathbf{x}^*$

$$\mathbf{m}^{est} = \langle \mathbf{m} \rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{obs} - \mathbf{G}\langle \mathbf{m} \rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{obs} + [\mathbf{I} - \mathbf{R}]\langle \mathbf{m} \rangle$$

with $\mathbf{G}^{-g} = [\mathrm{cov}\,\mathbf{m}]_A \mathbf{G}^T \{[\mathrm{cov}\,\mathbf{d}] + [\mathrm{cov}\,\mathbf{g}] + \mathbf{G}[\mathrm{cov}\,\mathbf{m}]_A \mathbf{G}^T\}^{-1}$

reminiscent of $\mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}$
minimum length solution

after pulling $\mathbf{m}^{\text{est}}$ out of $\mathbf{x}^*$

$$\mathbf{m}^{\text{est}} = \langle\mathbf{m}\rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle\mathbf{m}\rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle\mathbf{m}\rangle$$

$$\text{with} \quad \mathbf{G}^{-g} = [\text{cov}\,\mathbf{m}]_A \mathbf{G}^T \{[\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] + \mathbf{G}[\text{cov}\,\mathbf{m}]_A \mathbf{G}^T\}^{-1}$$

error in theory adds to error in data

after pulling $\mathbf{m}^{\text{est}}$ out of $\mathbf{x}^*$

$$\mathbf{m}^{\text{est}} = \langle\mathbf{m}\rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle\mathbf{m}\rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle\mathbf{m}\rangle$$

with $\mathbf{G}^{-g} = [\text{cov}\,\mathbf{m}]_A\mathbf{G}^T\{[\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] + \mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^T\}^{-1}$

solution depends on the values of the prior information only to the extent that the model resolution matrix is different from an identity matrix

and after algebraic manipulation

$$\mathbf{m}^{\text{est}} = \langle\mathbf{m}\rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle\mathbf{m}\rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle\mathbf{m}\rangle$$

with $\mathbf{G}^{-g} = [\text{cov}\,\mathbf{m}]_A\mathbf{G}^T\{[\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}] + \mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^T\}^{-1}$

which also equals

$$\mathbf{G}^{-g} = \{\mathbf{G}^T([\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}])^{-1}\mathbf{G} + [\text{cov}\,\mathbf{m}]_A^{-1}\}^{-1}\mathbf{G}^T([\text{cov}\,\mathbf{d}] + [\text{cov}\,\mathbf{g}])^{-1}$$

reminiscent of $(\mathbf{G}^T\mathbf{G})^{-1}\,\mathbf{G}^T$
least squares solution

interesting aside

weighted least squares solution

is equal to the

weighted minimum length solution

what did we learn?

for linear Gaussian inverse problem

inexactness of theory
just adds to
inexactness of data

Part 3

Use maximization of relative entropy as a guiding principle for solving inverse problems

from last lecture

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

Do we know a little about **m**

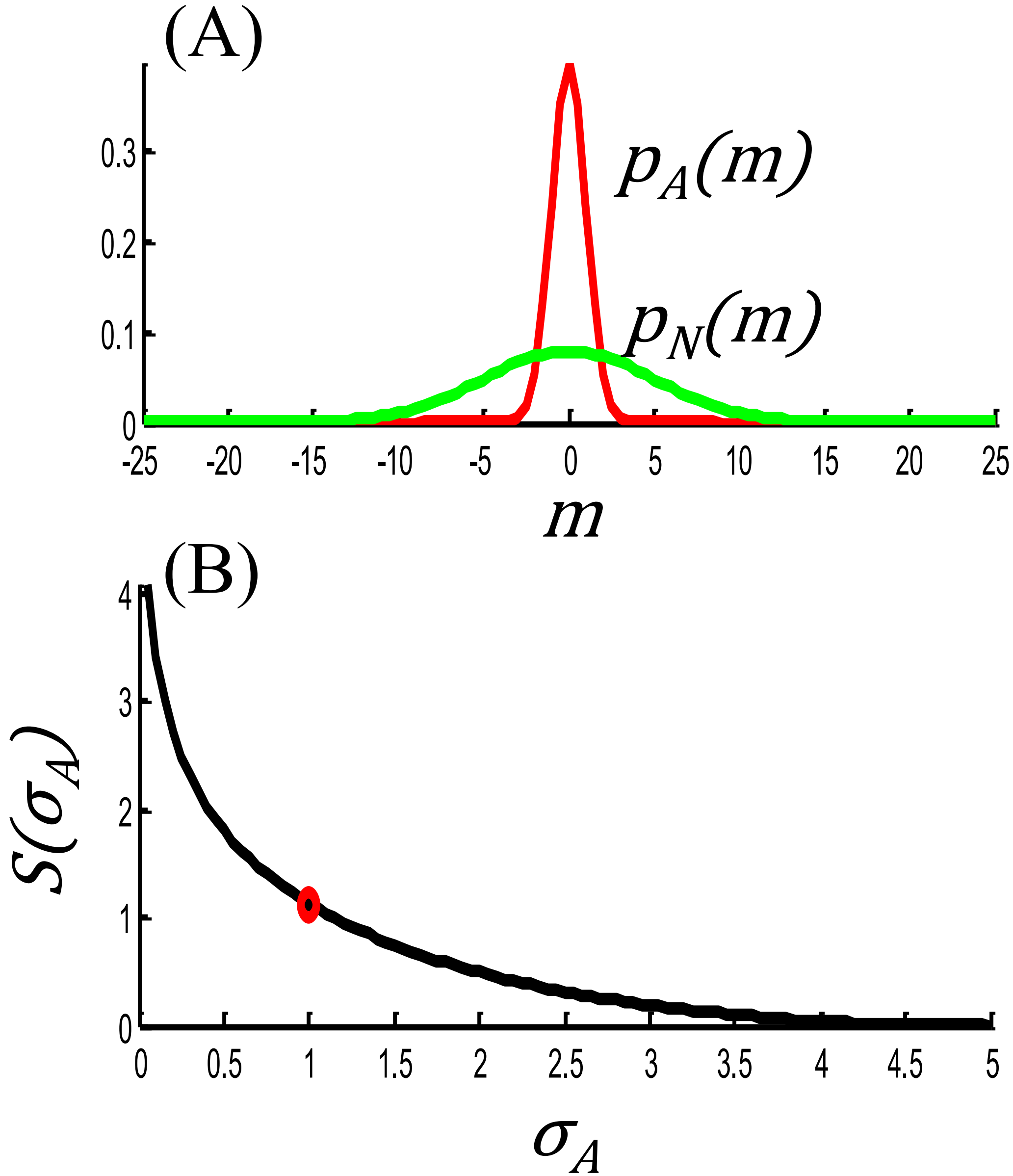
or

a lot about **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



(A)
0.3
0.2
0.1
0
-25
-20
-15
-10
-5
0
5
10
15
20
25
$p_A(m)$
$p_N(m)$
m
(B)
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
$S(\sigma_A)$
σ_A

Principle of
Maximum Relative Entropy

or if you prefer

Principle of
Minimum Information Gain

find solution p.d.f. $p_T(\mathbf{m})$ that has the largest relative entropy as compared to a priori p.d.f. $p_A(\mathbf{m})$

or if you prefer

find solution p.d.f. $p_T(\mathbf{m})$ that has smallest possible new information as compared to a priori p.d.f. $p_A(\mathbf{m})$

$$\text{minimize } S = \int p_T(\mathbf{m}) \log\left(\frac{p_T(\mathbf{m})}{p_A(\mathbf{m})}\right) \mathrm{d}^{\mathrm{M}} m \quad \text{with constraints}$$

$$\int p_T(\mathbf{m}) \mathrm{d}^{\mathrm{M}} m = 1 \quad \text{and} \quad \int p_T(\mathbf{m})(\mathbf{d} - \mathbf{G}\mathbf{m}) \, \mathrm{d}^{\mathrm{M}} m = 0$$

$$\text{minimize } S = \int p_T(\mathbf{m}) \log\left(\frac{p_T(\mathbf{m})}{p_A(\mathbf{m})}\right) \mathrm{d}^{\mathrm{M}} m \quad \text{with constraints}$$

$$\int p_T(\mathbf{m}) \mathrm{d}^{\mathrm{M}} m = 1 \quad \text{and} \quad \int p_T(\mathbf{m})(\mathbf{d} - \mathbf{G}\mathbf{m})\, \mathrm{d}^{\mathrm{M}} m = 0$$

properly
normalized
p.d.f.

data is satisfied in the mean
or
expected value of error is zero

After minimization using Lagrange Multipliers process

$p_T(\mathbf{m})$ is Gaussian with maximum likelihood point $\mathbf{m}^{est}$ satisfying

$$\mathbf{m}^{est} - \langle\mathbf{m}\rangle = [\mathrm{cov}\,\mathbf{m}]_A \mathbf{G}^T \{\mathbf{G}[\mathrm{cov}\,\mathbf{m}]_A \mathbf{G}^T\}^{-1} \{\mathbf{d} - \mathbf{G}\langle\mathbf{m}\rangle\}$$

After minimization using Lagrane Multipliers process

$p_T(\mathbf{m})$ is Gaussian with maximum likelihood point $\mathbf{m}^{est}$ satisfying

$$\mathbf{m}^{est} - \langle\mathbf{m}\rangle = [\text{cov}\,\mathbf{m}]_A \mathbf{G}^T\{\mathbf{G}[\text{cov}\,\mathbf{m}]_A\mathbf{G}^T\}^{-1}\{\mathbf{d} - \mathbf{G}\langle\mathbf{m}\rangle\}$$

just the weighted minimum length solution

What did we learn?

Only that the
Principle of Maximum Entropy
is yet another way of deriving
the inverse problem solutions
we are already familiar with

Part 4

F-test

as way to determine whether one solution is "better" than another

Common Scenario

two different theories

solution m^{est}_A
M_A model parameters
prediction error E_A

solution m^{est}_B
M_B model parameters
prediction error E_B

Suppose $E_B < E_A$

Is B really better than A ?

What if B has many more model parameters than A

$$M_B >> M_A$$

Is B fitting better any surprise?

Need to against Null Hypothesis

The difference in error is due to random variation suppose error **e** has a Gaussian p.d.f.

uncorrelated

uniform variance σ_d

estimate variance

$$(\sigma_d^{est})^2 = \frac{1}{\nu}\sum_{i=1}^{N} e_i^2 = \frac{E}{\nu} \quad \text{with} \quad \nu = N - M$$

want to known the probability density function of

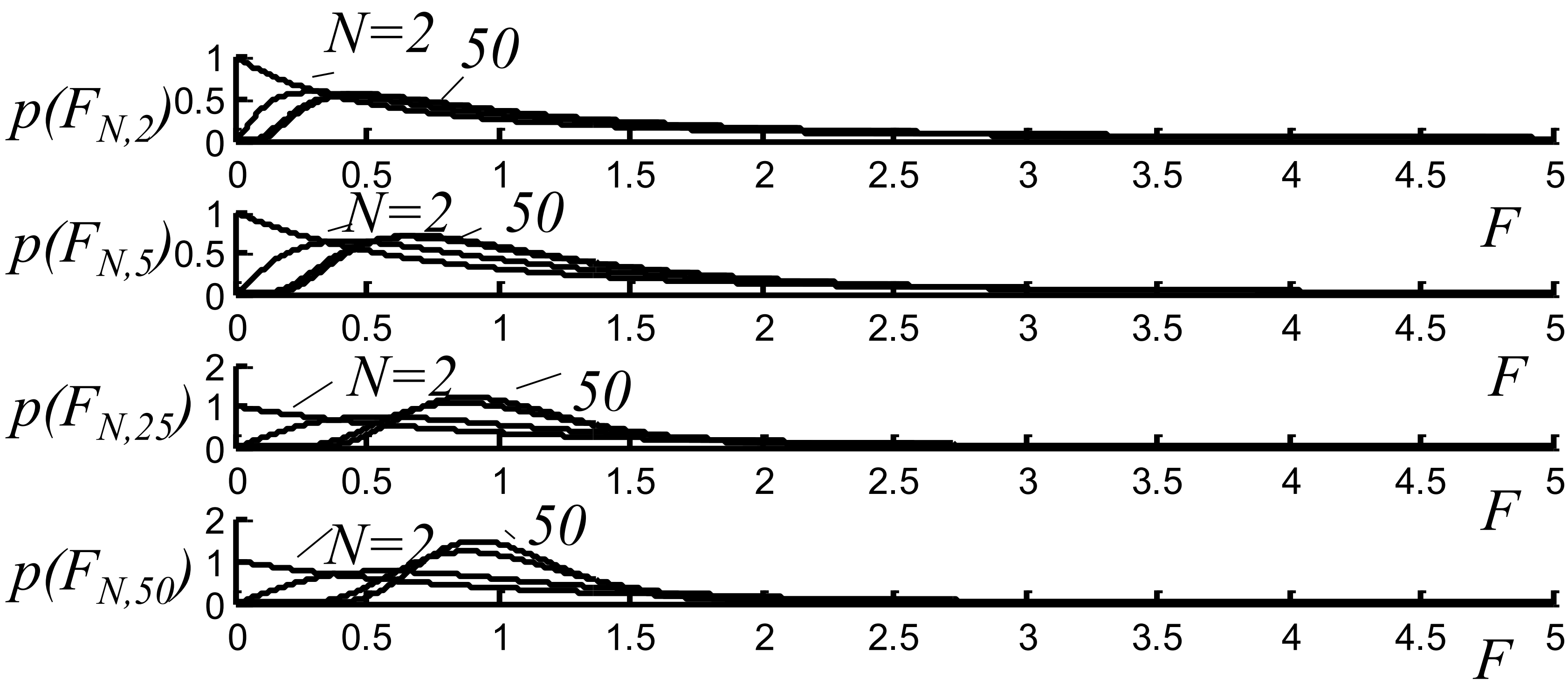
$$(\sigma_{dA}^{est})^2 / (\sigma_{dB}^{est})^2$$

actually, we'll use the quantity

$$F(\nu_A, \nu_B) = \frac{(\sigma_{dA}^{est})^2 / (\sigma_{dA}^{true})^2}{(\sigma_{dB}^{est})^2 / (\sigma_{dA}^{true})^2}$$

which is the same, as long as the two theories that we're testing is applied to the same data

p.d.f. of *F* is known



as is its mean and variance

$$\langle F \rangle = \frac{\nu_B}{\nu_B - 2} \qquad \sigma_F^2 = \frac{2\nu_B^2(\nu_A + \nu_B - 2)}{\nu_A(\nu_B - 2)^2(\nu_B - 4)}$$

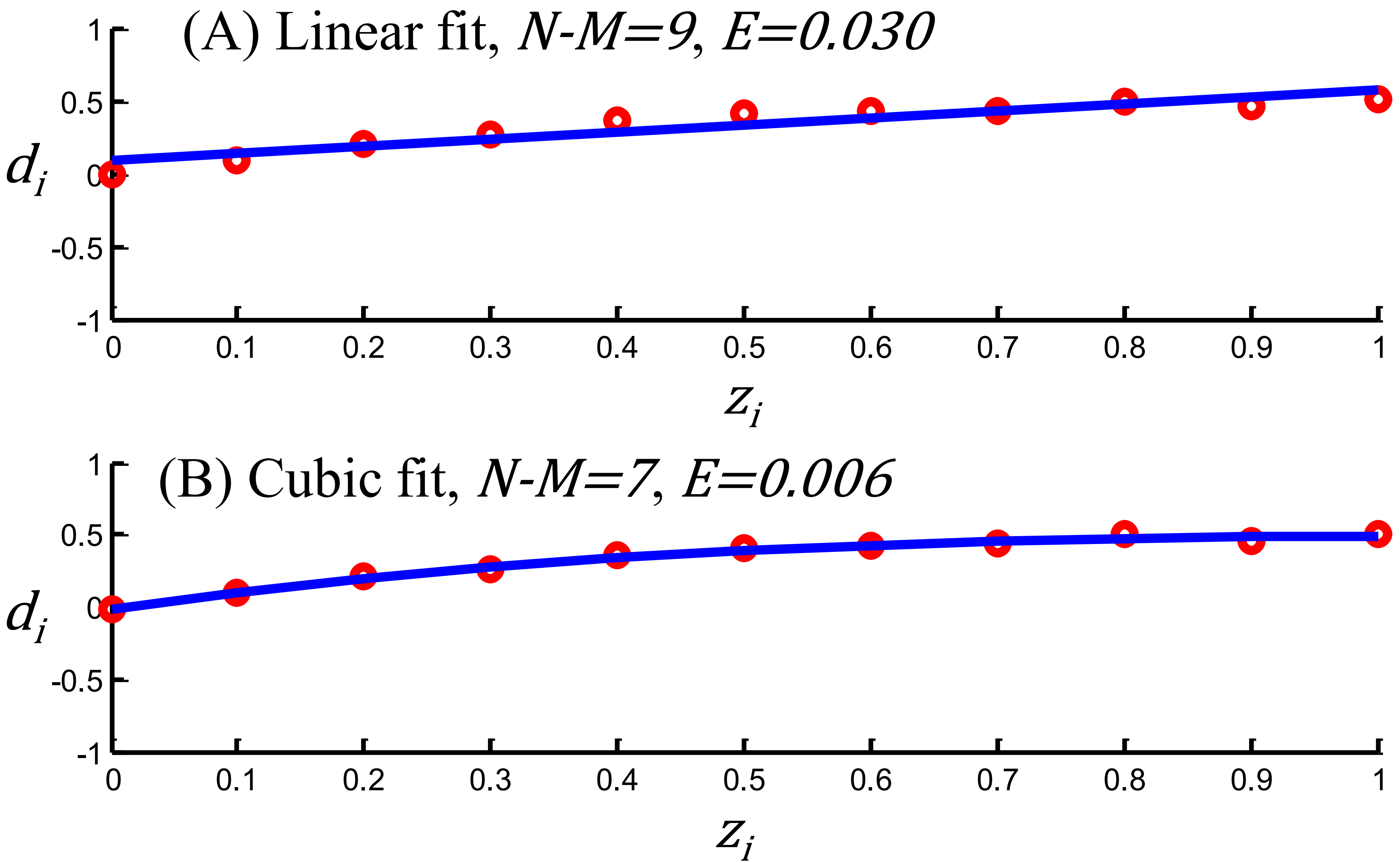
example

same dataset fit with

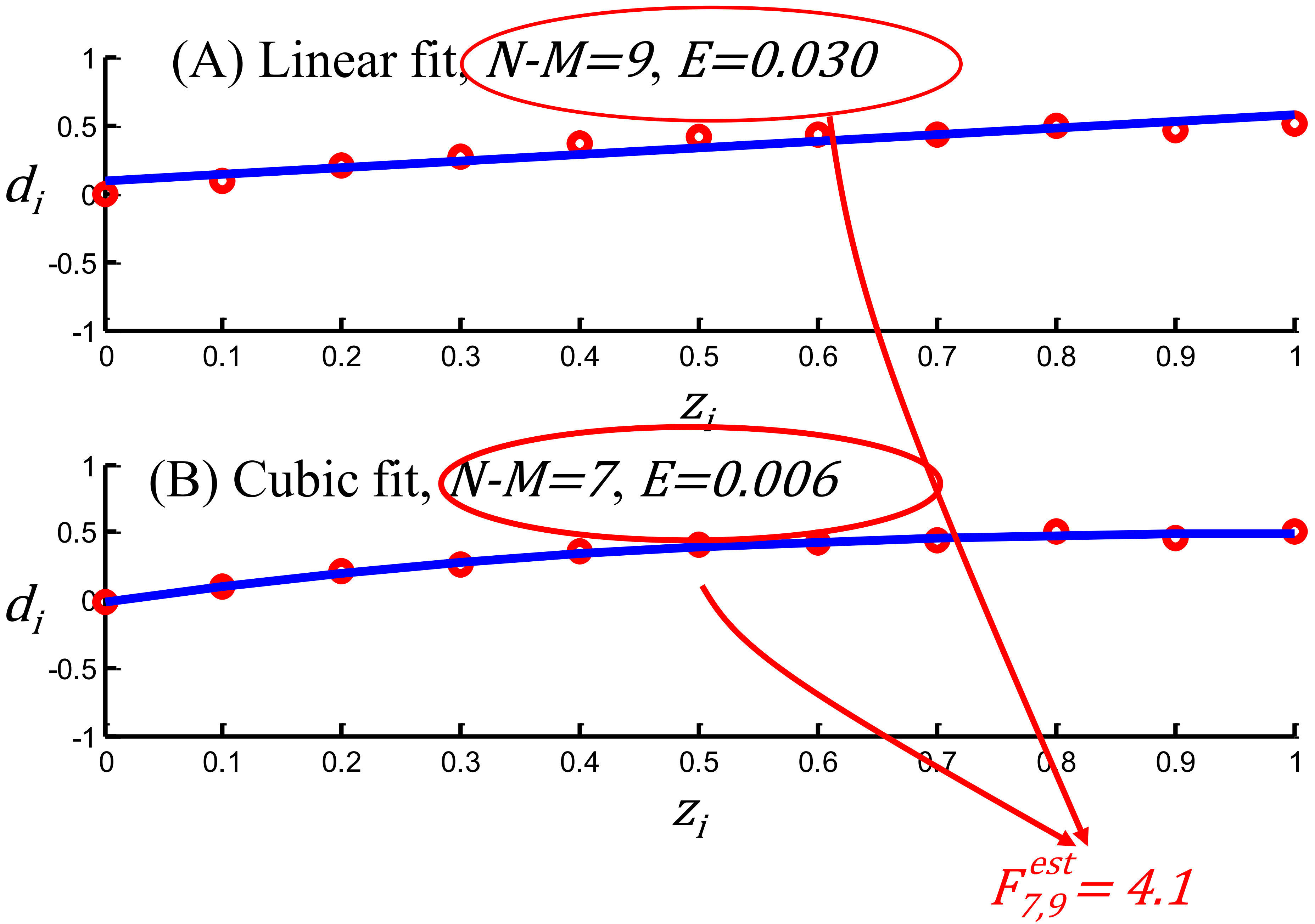
a straight line

and

a cubic polynomial



(A) Linear fit, N-M=9, E=0.030
1
0.5
0
-0.5
-1
d_i
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
z_i
(B) Cubic fit, N-M=7, E=0.006
1
0.5
0
-0.5
-1
d_i
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
z_i



(A) Linear fit, N-M=9, E=0.030
1
0.5
0
-0.5
-1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
d_i
z_i
(B) Cubic fit, N-M=7, E=0.006
1
0.5
0
-0.5
-1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
d_i
z_i
$F^{est}_{7,9} = 4.1$

$$P(F < 1/F^{est} \text{ or } F > F^{est})$$

probability that

$$F > F^{est}$$

(cubic fit seems better than linear fit)

by random chance alone

or

$$F < 1/F^{est}$$

(linear fit seems better than cubic fit)

by random chance alone

in *MatLab*

```
P = 1 - (fcdf(Fobs,vA,vB)-fcdf(1/Fobs,vA,vB));
```

answer: 6%

The Null Hypothesis

that the difference is due to random variation

cannot be rejected to 95% confidence