Lecture 4

The L_2 Norm and Simple Least Squares

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 concept of prediction error and the norms that quantify it

Develop the Least Squares Solution

Develop the Minimum Length Solution

Determine the covariance of these solutions

Part 1

prediction error and norms

The Linear Inverse Problem

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

The Linear Inverse Problem

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

data kernel

model parameters

data

an estimate of the model parameters can be used to predict the data

$$\mathbf{G}\mathbf{m}^{\text{est}} = \mathbf{d}^{\text{pre}}$$

but the prediction may not match the observed data
(e.g. due to observational error)

$$\mathbf{d}^{\text{pre}} \neq \mathbf{d}^{\text{obs}}$$

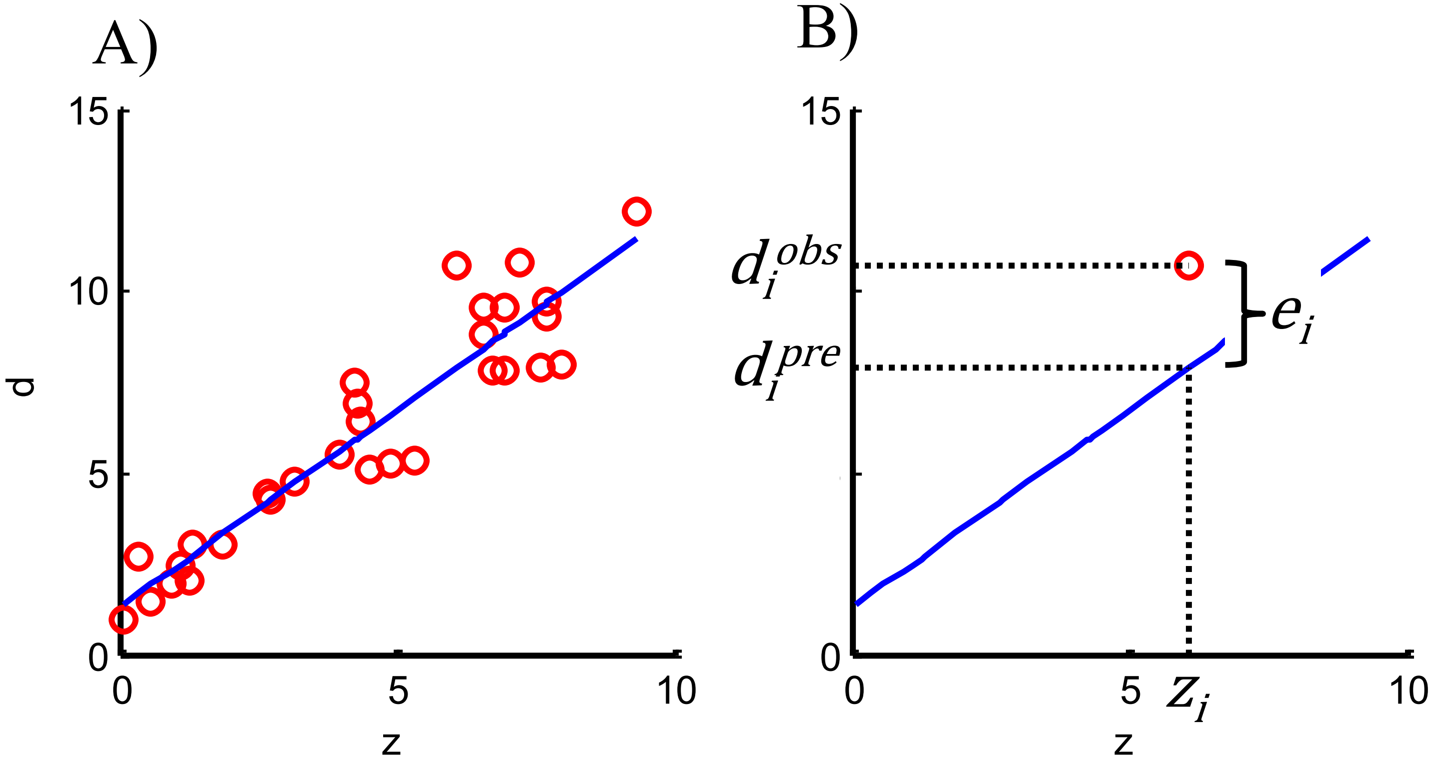
this mismatch leads us to define the prediction error

$$\mathbf{e} = \mathbf{d}^{obs} - \mathbf{d}^{pre}$$

$$\mathbf{e} = 0$$

when the model parameters exactly predict the data

example of prediction error for line fit to data



“norm”
rule for quantifying the overall size of the error vector **e**

lot’s of possible ways to do it

L_n family of norms

$$L_1 \text{ norm:} \quad \|\mathbf{e}\|_1 = \left[\sum_i |e_i|^1\right]$$

$$L_2 \text{ norm:} \quad \|\mathbf{e}\|_2 = \left[\sum_i |e_i|^2\right]^{1/2}$$

$$L_n \text{ norm:} \quad \|\mathbf{e}\|_n = \left[\sum_i |e_i|^n\right]^{1/n}$$

L_n family of norms

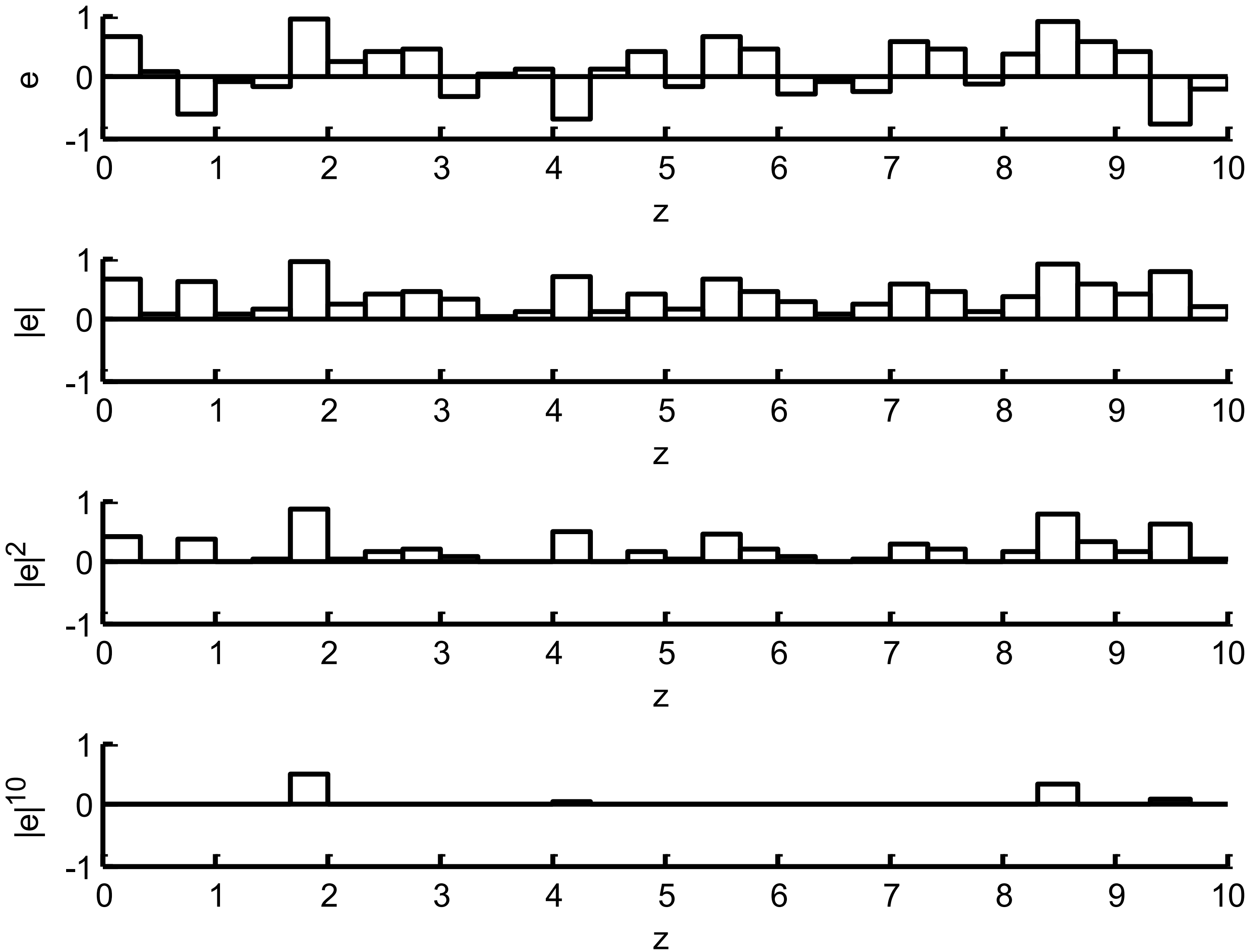
$$L_1 \text{ norm:} \quad \|\mathbf{e}\|_1 = \left[\sum_i |e_i|^1\right]$$

$$L_2 \text{ norm:} \quad \|\mathbf{e}\|_2 = \left[\sum_i |e_i|^2\right]^{½}$$

Euclidian length

$$L_n \text{ norm:} \quad \|\mathbf{e}\|_n = \left[\sum_i |e_i|^n\right]^{1/n}$$

higher norms give increaing weight to largest element of **e**



limiting case

$$L_\infty \text{ norm:} \quad \|\mathbf{e}\|_\infty = \max_i |e_i|$$

guiding principle for solving an inverse problem

find the $\mathbf{m}^{est}$

that minimizes $E=||\mathbf{e}||$

with

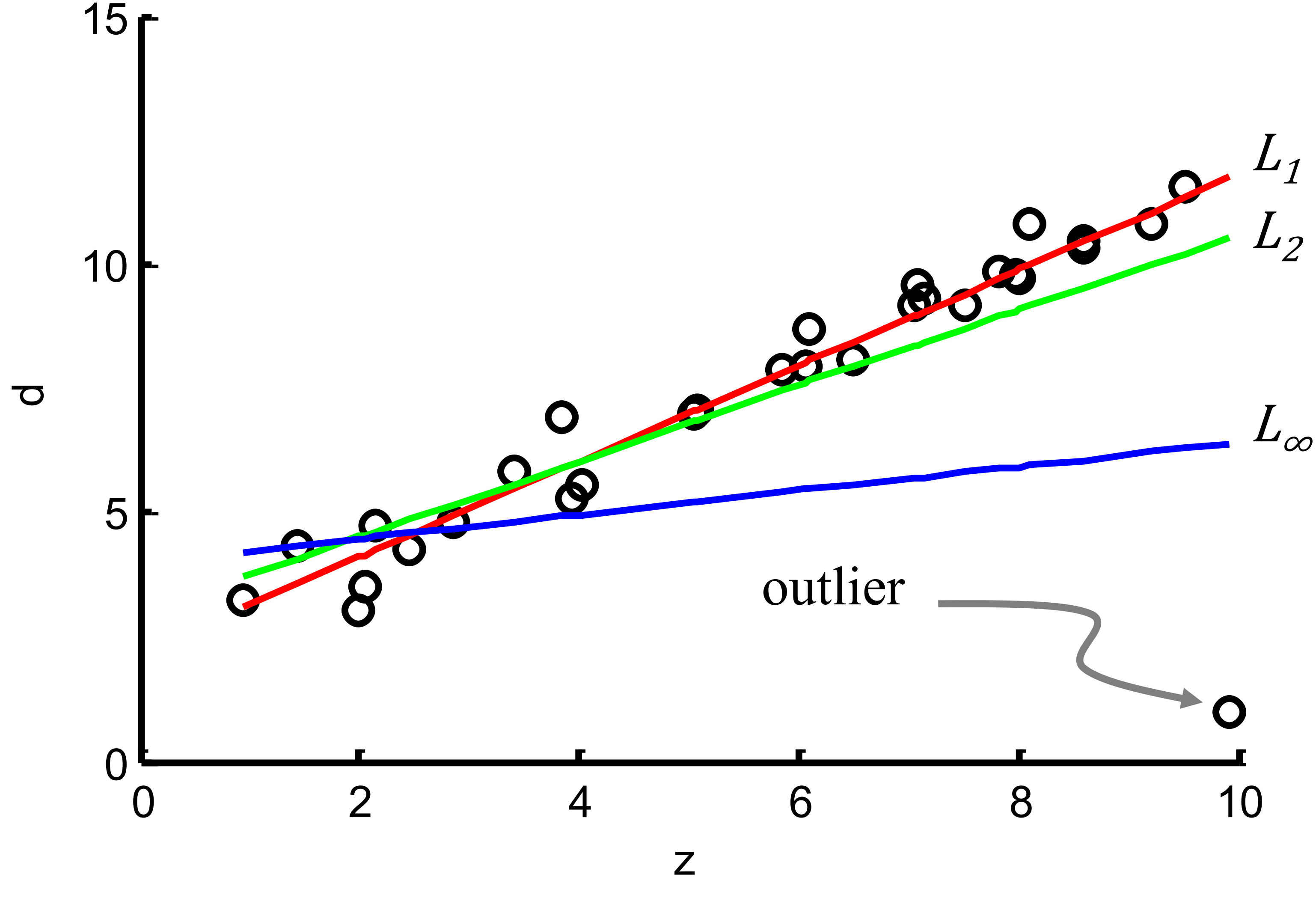
$$\mathbf{e} = \mathbf{d}^{obs} - \mathbf{d}^{pre}$$

and

$$\mathbf{d}^{pre} = \mathbf{G}\mathbf{m}^{est}$$

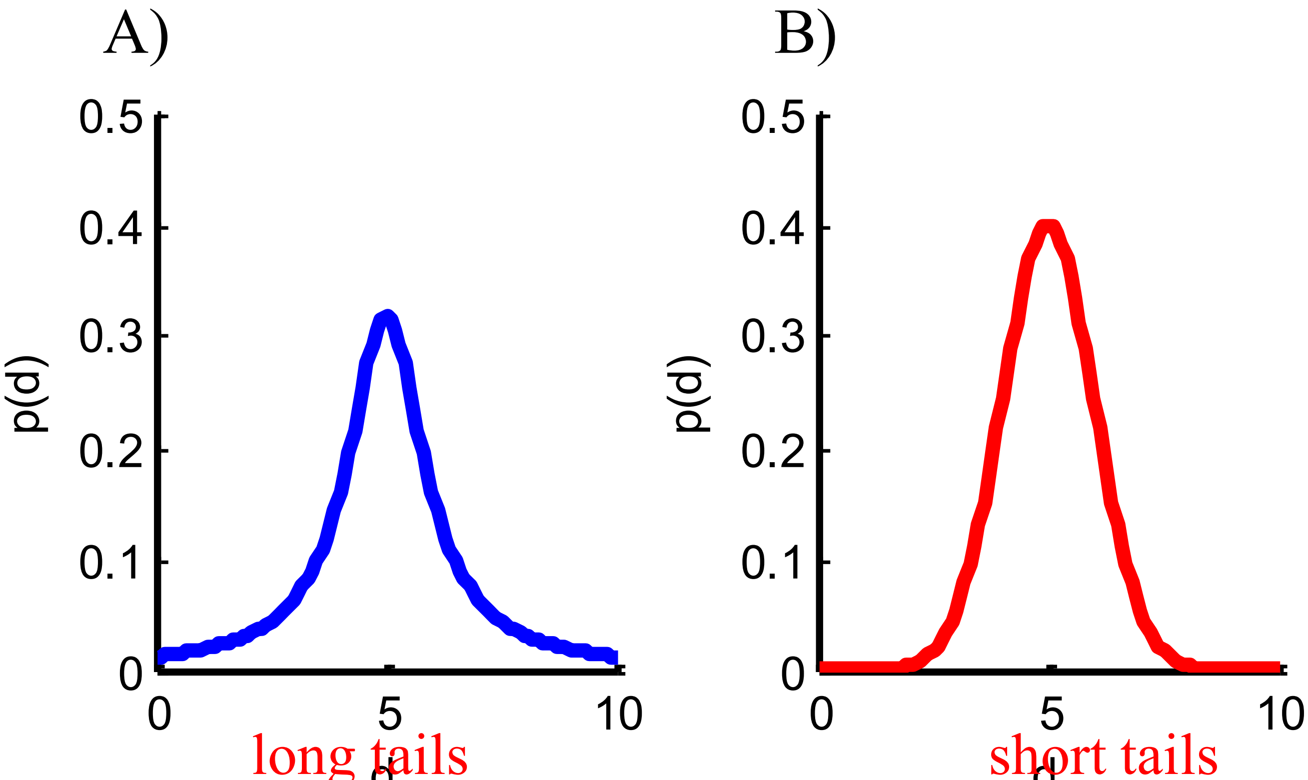
but which norm to use?

it makes a difference!



15
10
5
0
d
0
2
4
6
8
10
z
L1
L2
L∞
outlier

Answer is related to the distribution of the error. Are outliers common or rare?



A)	B)
long tails	short tails
outliers common	outliers uncommon
outliers unimportant	outliers important
use low norm	use high norm
gives low weight to outliers	gives high weight to outliers

as we will show later in the class ...

use L_2 norm
when data has
Gaussian-distributed error

Part 2

Least Squares Solution to $\mathbf{Gm}=\mathbf{d}$

L_2 norm of error is its Euclidian length

$$E = \sum_{i=1}^{N} e_i^2 \quad = \mathbf{e}^{\mathrm{T}}\mathbf{e}$$

so E is the square of the Euclidean length

mimimize E

Principle of Least Squares

Least Squares Solution to **Gm=d**

$$E = \mathbf{e}^T\mathbf{e} = (\mathbf{d} - \mathbf{Gm})^T(\mathbf{d} - \mathbf{Gm}) = \sum_{i=1}^{N}\left[d_i - \sum_{j=1}^{M} G_{ij}m_j\right]\left[d_i - \sum_{k=1}^{M} G_{ik}m_k\right]$$

minimize E with respect to m_q

$$\partial E/\partial m_q = 0$$

$$E = \mathbf{e}^T\mathbf{e} = (\mathbf{d} - \mathbf{Gm})^T(\mathbf{d} - \mathbf{Gm}) = \sum_{i=1}^{N}\left[d_i - \sum_{j=1}^{M} G_{ij}m_j\right]\left[d_i - \sum_{k=1}^{M} G_{ik}m_k\right]$$

so, multiply out

$$E = \sum_{j=1}^{M}\sum_{k=1}^{M} m_j m_k \sum_{i=1}^{N} G_{ij}G_{ik} - 2\sum_{j=1}^{M} m_j \sum_{i=1}^{N} G_{ij}d_i + \sum_{i=1}^{N} d_i d_i$$

first term

$$\frac{\partial}{\partial m_q}\left[\sum_{j=1}^{M}\sum_{k=1}^{M} m_j m_k \sum_{i=1}^{N} G_{ij} G_{ik}\right] = \sum_{j=1}^{M}\sum_{k=1}^{M}[\delta_{jq} m_k + m_j \delta_{kq}] \sum_{i=1}^{N} G_{ij} G_{ik}$$

$$= 2\sum_{k=1}^{M} m_k \sum_{i=1}^{N} G_{iq} G_{ik}$$

first term

$$\frac{\partial}{\partial m_q}\left[\sum_{j=1}^{M}\sum_{k=1}^{M} m_j m_k \sum_{i=1}^{N} G_{ij}G_{ik}\right] = \sum_{j=1}^{M}\sum_{k=1}^{M}\left[\delta_{jq} m_k + m_j \delta_{kq}\right]\sum_{i=1}^{N} G_{ij}G_{ik}$$

$$= 2\sum_{k=1}^{M} m_k \sum_{i=1}^{N} G_{iq}G_{ik}$$

$\partial m_j / \partial m_q = \delta_{jq}$

since m_j and m_q are independent variables

Kronecker delta
(elements of identity matrix)

$$[\mathbf{I}]_{ij} = \delta_{ij}$$

$$\mathbf{a} = \mathbf{I}\mathbf{b} = \mathbf{b}$$

$$a_i = \Sigma_j \, \delta_{ij} \, b_j = b_i$$

$$a_i = \Sigma_j \, \delta_{ij} \, b_j = b_i$$

i

second term

$$-2\frac{\partial}{\partial m_q}\left[\sum_{j=1}^{M} m_j \sum_{i=1}^{N} G_{ij} d_i\right] = -2\sum_{j=1}^{M} \delta_{jq} \sum_{i=1}^{N} G_{ij} d_i = -2\sum_{i=1}^{N} G_{iq} d_i$$

third term

$$\frac{\partial}{\partial m_q}\left[\sum_{i=1}^{N} d_i d_i\right] = 0$$

putting it all together

$$\frac{\partial E}{\partial m_q} = 0 = 2\sum_{k=1}^{M} m_k \sum_{i=1}^{N} G_{iq}G_{ik} - \; -2\sum_{i=1}^{N} G_{iq}d_i$$

or

$$\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{m} - \mathbf{G}^{\mathrm{T}}\mathbf{d} = 0$$

presuming $[\mathbf{G}^{\mathrm{T}}\mathbf{G}]$ has an inverse

Least Square Solution

$$\mathbf{m}^{\mathrm{est}} = [\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}$$

presuming $[\mathbf{G}^T\mathbf{G}]$ has an inverse

Least Square Solution

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

memorize

example
straight line problem

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

$$\begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} d \\ d_2 \\ \vdots \\ d_N \end{bmatrix}$$

$$\mathbf{G}^{\mathrm{T}}\mathbf{G} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \end{bmatrix} \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_N \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{bmatrix}$$

$$\mathbf{G}^{\mathrm{T}}\mathbf{d} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_N \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} d_i \\ \sum_{i=1}^{N} d_i z_i \end{bmatrix}$$

$$\mathbf{m}^{\text{est}} = [\mathbf{G}^{\text{T}}\mathbf{G}]^{-1}\mathbf{G}^{\text{T}}\mathbf{d} = \begin{bmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{N} d_i \\ \sum_{i=1}^{N} d_i z_i \end{bmatrix}$$

in practice,
no need to multiply matrices analytically

just use *MatLab*

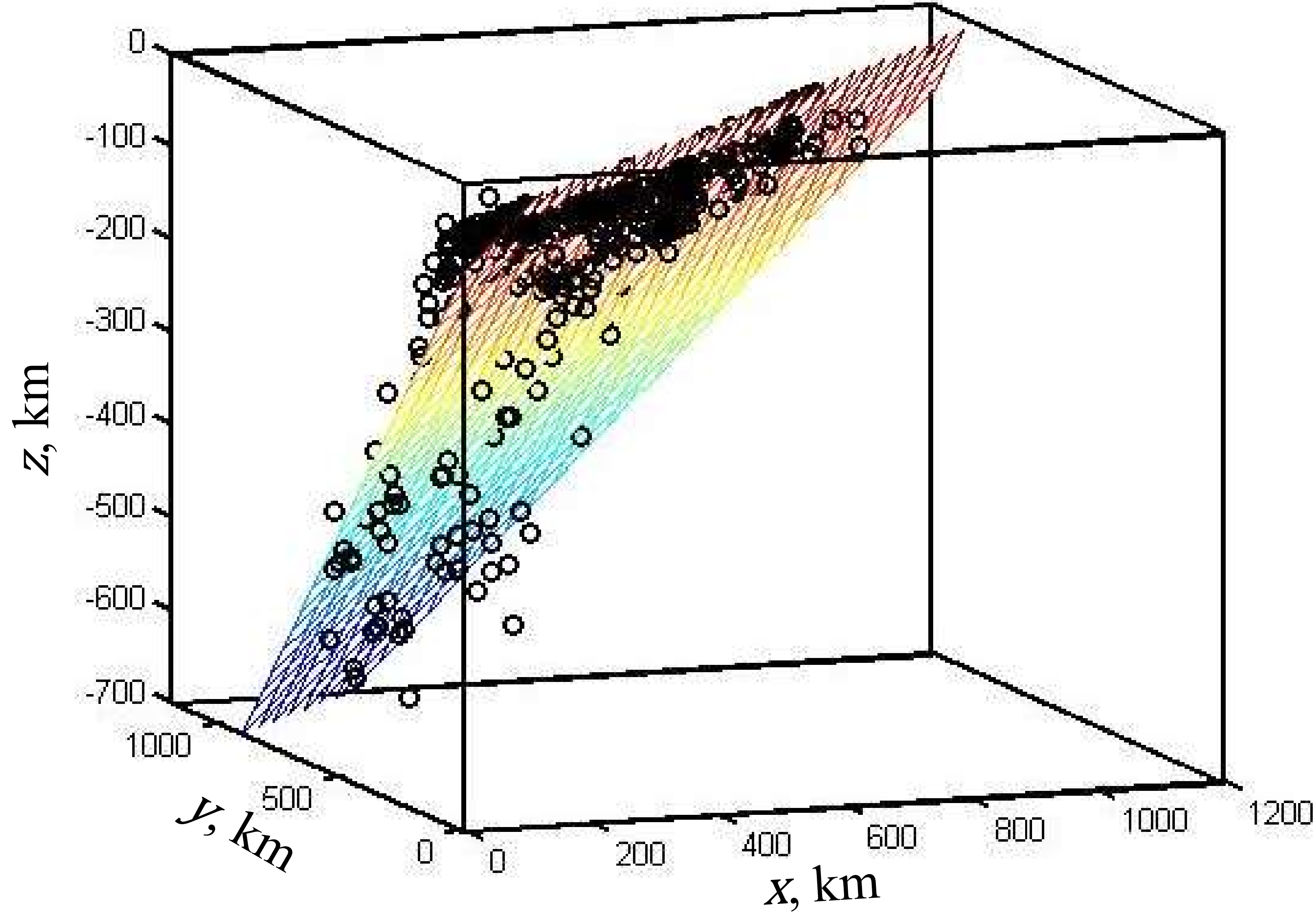
```
mest = (G'*G)\(G'*d);
```

another example
fitting a plane surface

$$d_i = m_1 + m_2 x_i + m_3 y_i$$

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

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & y_N \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} = \begin{bmatrix} d \\ d_2 \\ \vdots \\ d_N \end{bmatrix}$$



0
-100
-200
-300
-400
-500
-600
-700
z, km
1000
500
0
y, km
0
200
400
600
800
1000
1200
x, km

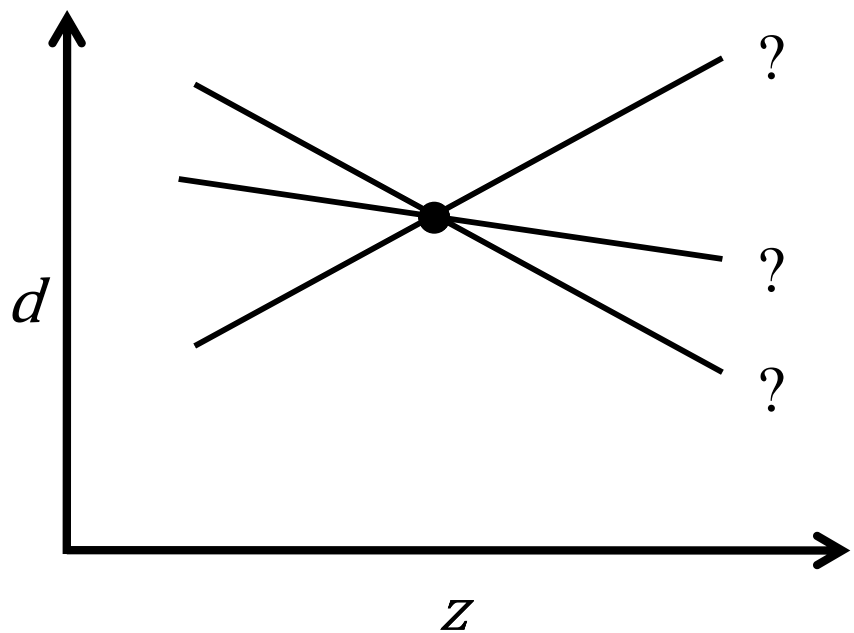
Part 3

Minimum Length Solution

but Least Squares will fail

when $[\mathbf{G}^T\mathbf{G}]$ has no inverse

example
fitting line to a single point



$$[\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1} = \begin{bmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{bmatrix}^{-1} \rightarrow \begin{bmatrix} 1 & z_1 \\ z_1 & z_1^2 \end{bmatrix}^{-1}$$

$$[\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1} = \begin{bmatrix} N & \sum_{i=1}^{N} z_i \\ \sum_{i=1}^{N} z_i & \sum_{i=1}^{N} z_i^2 \end{bmatrix}^{-1} \rightarrow \begin{bmatrix} 1 & z_1 \\ z_1 & z_1^2 \end{bmatrix}^{-1}$$

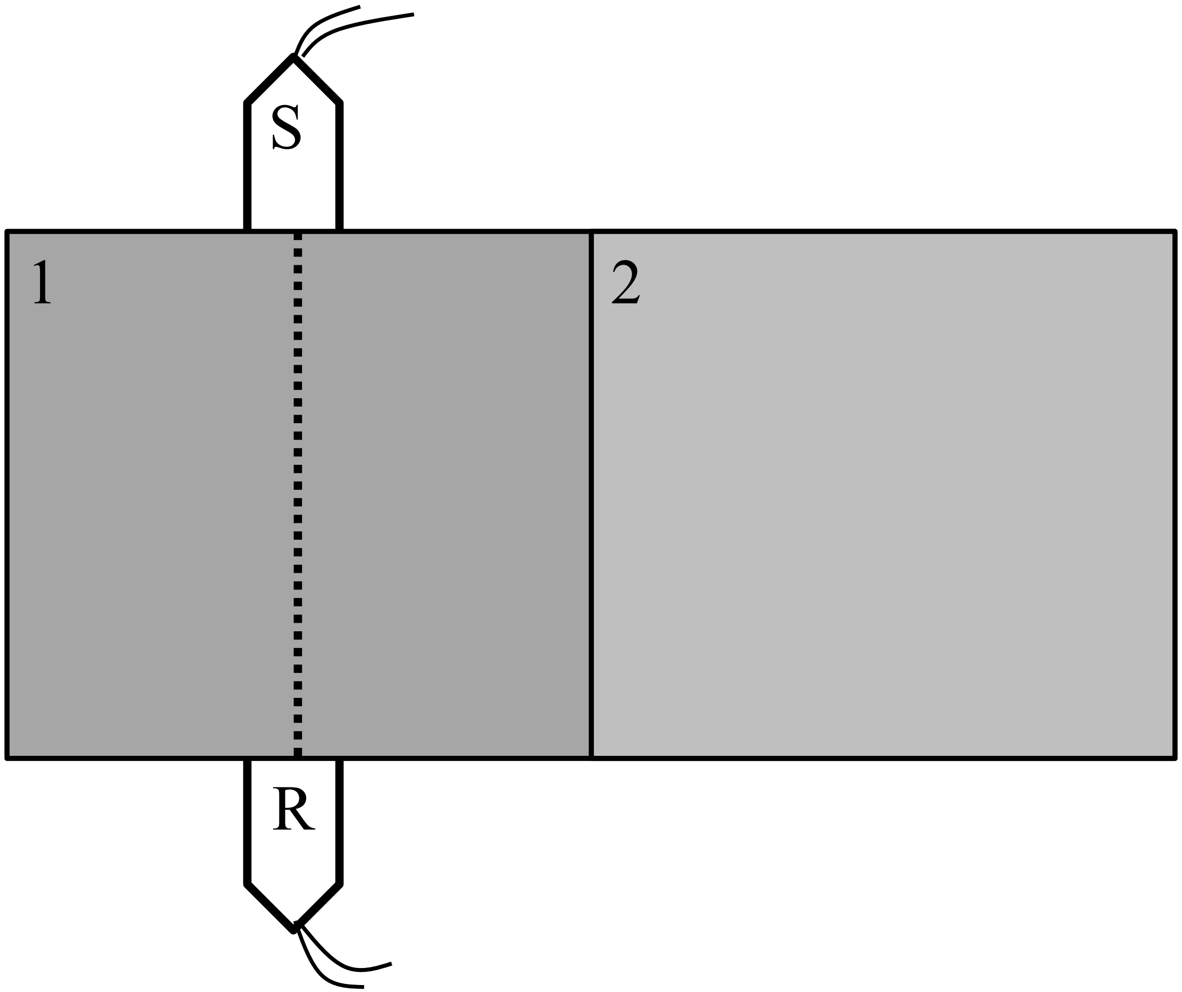
zero determinant
hence no inverse

Least Squares will fail

when more than one solution minimizes the error

the inverse problem is "underdetermined"

simple example of an underdetermined problem



What to do?

use another guiding principle

"a priori" information about the solution in the case

choose a solution that is small

minimize $\|\mathbf{m}\|_2$

simplest case
“purely underdetermined”

more than one solution has zero error

minimize $L = ||\mathbf{m}||_2^2$

with the constraint that $\mathbf{e} = 0$

Method of Lagrange Multipliers

minimize L with constraints

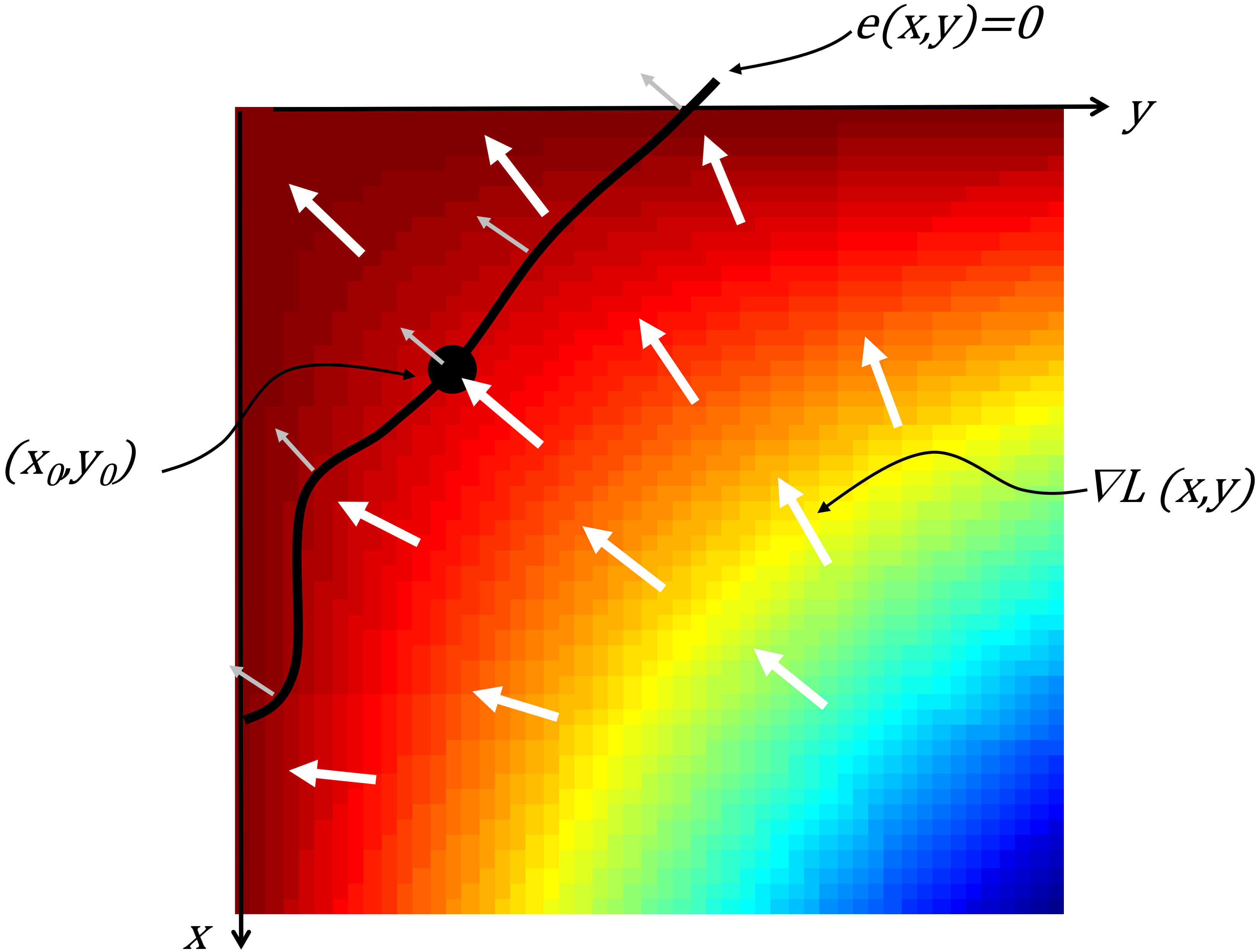
$C_1=0,\ C_2=0,\ \ldots$

equivalent to

minimize $\Phi = L + \lambda_1 C_1 + \lambda_2 C_2 + \ldots$

with no constraints

λs called "Lagrange Multipliers"



e(x,y)=0
y
(x0,y0)
∇L (x,y)
x

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

$$\frac{\partial \Phi}{\partial m_q} = \sum_{i=1}^{M} 2\frac{\partial m_i}{\partial m_q} m_i - \sum_{i=1}^{N} \lambda_i \sum_{j=1}^{M} G_{ij} \frac{\partial m_j}{\partial m_q} = 2m_q - \sum_{i=1}^{N} \lambda_i G_{iq}$$

$$2\mathbf{m} = \mathbf{G}^T \boldsymbol{\lambda} \quad \text{and} \quad \mathbf{Gm} = \mathbf{d}$$

$$\tfrac{1}{2}\mathbf{G}\mathbf{G}^T \boldsymbol{\lambda} = \mathbf{d}$$

$$\boldsymbol{\lambda} = 2[\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}$$

$$\mathbf{m} = \mathbf{G}^T [\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{d}$$

presuming $[\mathbf{G}\mathbf{G}^T]$ has an inverse

Minimum Length Solution

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

presuming $[\mathbf{G}\mathbf{G}^T]$ has an inverse

Minimum Length Solution

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

memorize

Part 4

Covariance

Least Squares Solution

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

Minimum Length Solution

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

both have the linear form

$$\mathbf{m} = \mathbf{M}\mathbf{d}$$

but if

$$\mathbf{m} = \mathbf{M}\mathbf{d}$$

then

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

when data are uncorrelated with uniform variance $\sigma_d^{\,2}$

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

so

Least Squares Solution

$$[\mathrm{cov}\,\mathbf{m}] = [\mathbf{G}^{\mathrm{T}}\mathbf{G}\,]^{-1}\mathbf{G}^{\mathrm{T}}\sigma_d^{\,2}\,\mathbf{G}[\mathbf{G}^{\mathrm{T}}\mathbf{G}\,]^{-1}$$

$$[\mathrm{cov}\,\mathbf{m}] = \sigma_d^{\,2}\,[\mathbf{G}^{\mathrm{T}}\mathbf{G}\,]^{-1}$$

Minimum Length Solution

$$[\mathrm{cov}\,\mathbf{m}] = \mathbf{G}^{\mathrm{T}}\,[\mathbf{G}\mathbf{G}^{\mathrm{T}}\,]^{-1}\,\sigma_d^{\,2}\,[\mathbf{G}\mathbf{G}^{\mathrm{T}}\,]^{-1}\mathbf{G}$$

$$[\mathrm{cov}\,\mathbf{m}] = \sigma_d^{\,2}\,\mathbf{G}^{\mathrm{T}}\,[\mathbf{G}\mathbf{G}^{\mathrm{T}}\,]^{-2}\mathbf{G}$$

Least Squares Solution

$$[\text{cov}\ \mathbf{m}] = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T \sigma_d^2\, \mathbf{G}[\mathbf{G}^T\mathbf{G}]^{-1}$$

$$[\text{cov}\ \mathbf{m}] = \sigma_d^2\, [\mathbf{G}^T\mathbf{G}]^{-1}$$

memorize

Minimum Length Solution

$$[\text{cov}\ \mathbf{m}] = \mathbf{G}^T\, [\mathbf{G}\mathbf{G}^T]^{-1}\, \sigma_d^2\, [\mathbf{G}\mathbf{G}^T]^{-1}\mathbf{G}$$

$$[\text{cov}\ \mathbf{m}] = \sigma_d^2\, \mathbf{G}^T\, [\mathbf{G}\mathbf{G}^T]^{-2}\mathbf{G}$$

where to obtain the value of σ_d^2

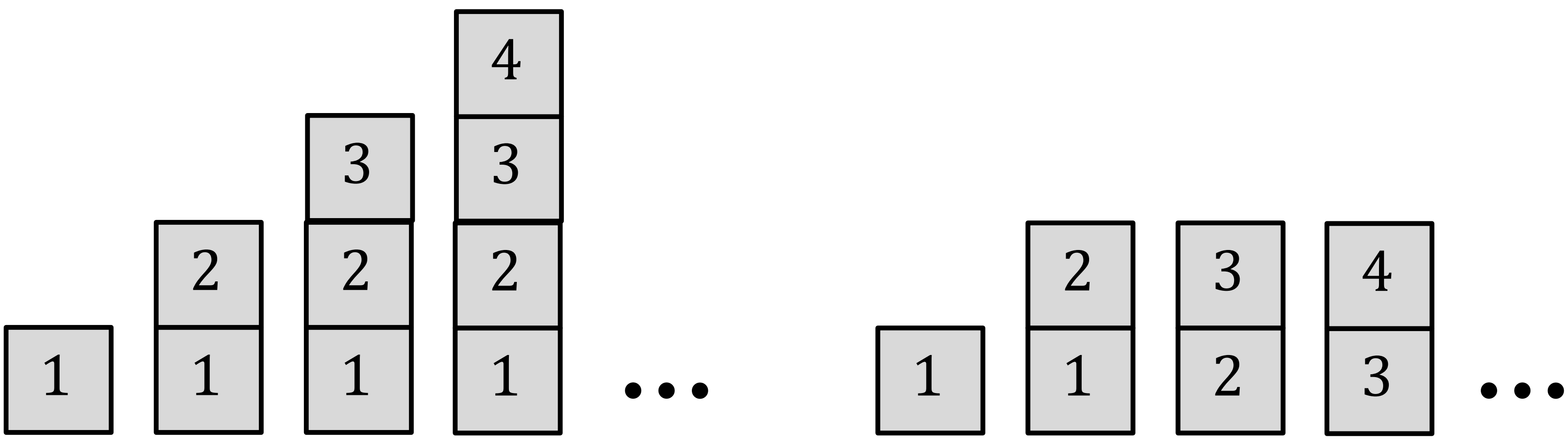
a priori value – based on knowledge of accuracy of measurement technique

my ruler has 1 mm divisions, so $\sigma_d \approx ½ mm$

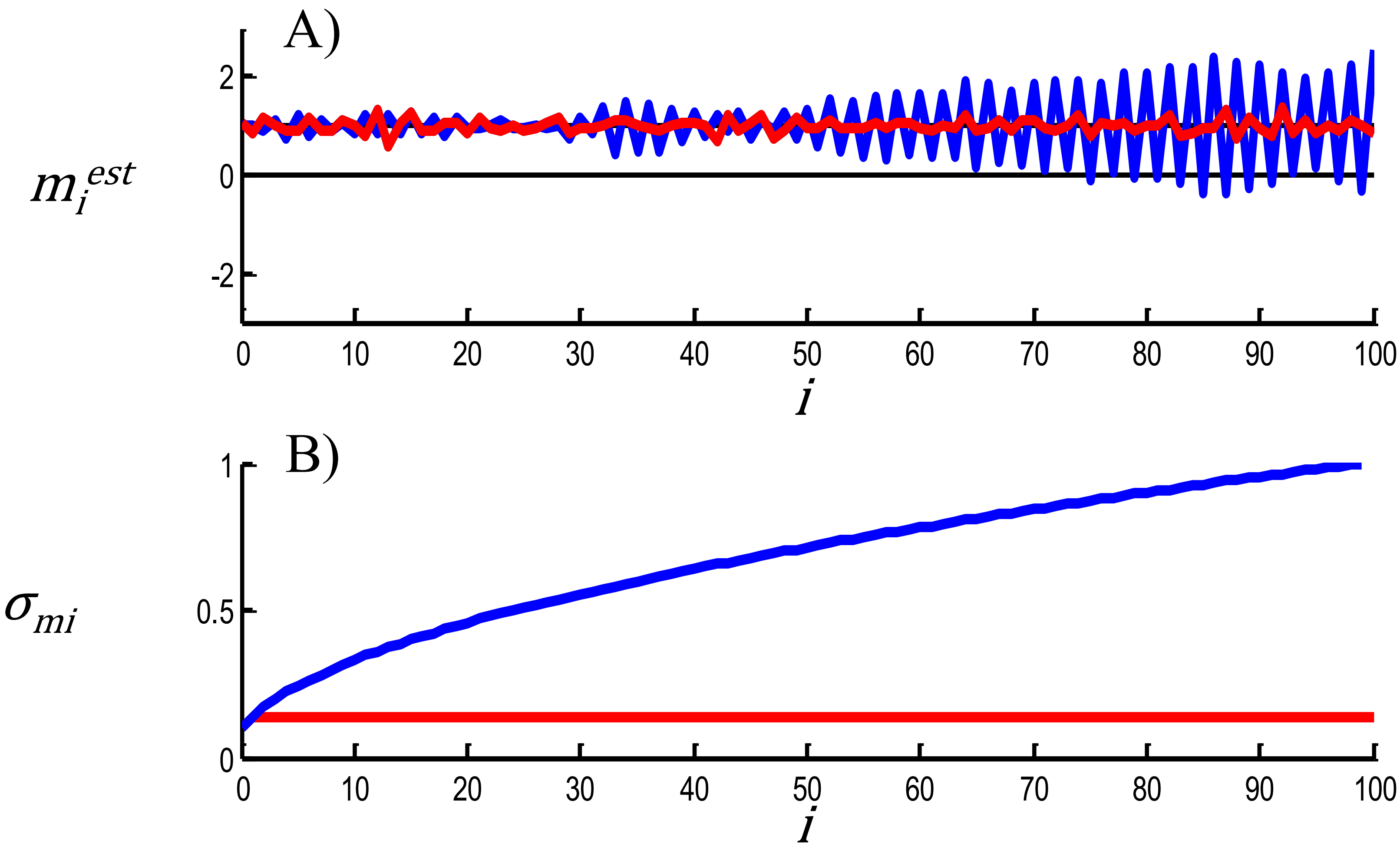
a posteriori value – based on prediction error

$$\sigma_d^2 \approx \frac{1}{N-M} \sum_{i=1}^{N} e_i^2$$

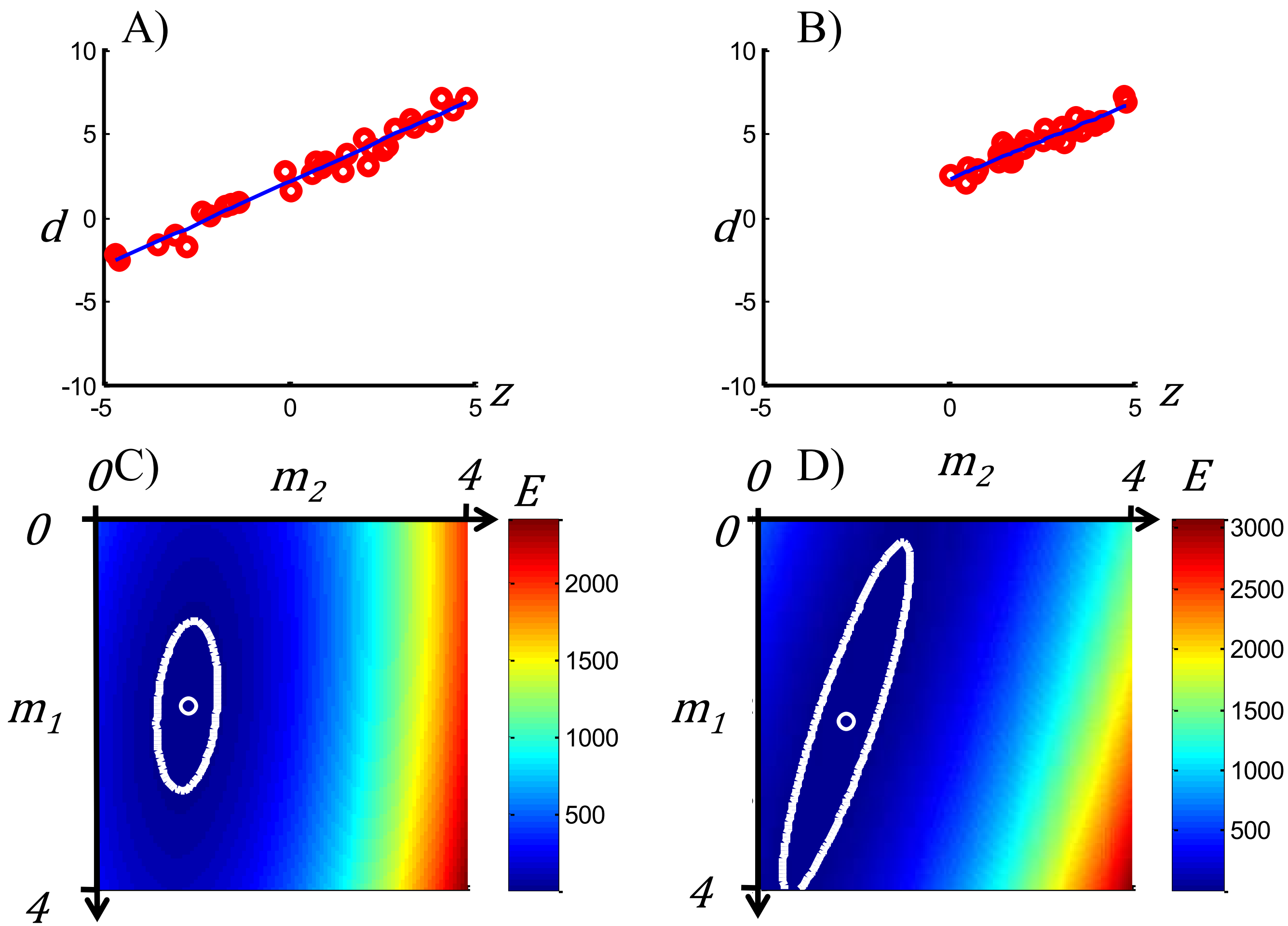
variance critically dependent on experiment design (structure of $\mathbf{G}$)



which is the better way to weigh a set of boxes ?


A)
2
0
-2
m_i^{est}
0
10
20
30
40
50
60
70
80
90
100
i
B)
1
0.5
0
σ_{mi}
0
10
20
30
40
50
60
70
80
90
100
i

Relationship between [cov **m**] and Error Surface



Taylor Series expansion of the error about its minimum

$$\Delta E = E(\mathbf{m}) - E(\boldsymbol{m}^{est}) = [\mathbf{m} - \mathbf{m}^{est}]^{\mathrm{T}} \left[\frac{1}{2} \frac{\partial^2 E}{\partial \mathbf{m}^2} \right]_{\mathbf{m}=\mathbf{m}^{est}} [\mathbf{m} - \mathbf{m}^{est}]$$

Taylor Series expansion of the error about its minimum

$$\Delta E = E(\mathbf{m}) - E(\boldsymbol{m}^{est}) = [\mathbf{m} - \mathbf{m}^{est}]^T \left[\frac{1}{2}\frac{\partial^2 E}{\partial \mathbf{m}^2}\right]_{\mathbf{m}=\mathbf{m}^{est}} [\mathbf{m} - \mathbf{m}^{est}]$$

curvature matrix
with elements
$\partial^2 E / \partial m_i \partial m_j$

for a linear problem
curvature is related to $\mathbf{G}^T\mathbf{G}$

$$E = (\mathbf{Gm}-\mathbf{d})^T(\mathbf{Gm}-\mathbf{d}) =$$

$$\mathbf{m}^T[\mathbf{G}^T\mathbf{G}]\mathbf{m}-\mathbf{d}^T\mathbf{Gm}-\mathbf{m}^T\mathbf{G}^T\mathbf{d}+\mathbf{d}^T\mathbf{d}$$

so

$$\partial^2 E/\, \partial m_i \partial m_j = [\mathbf{G}^T\mathbf{G}]_{ij}$$

and since

$$[\text{cov}\,\mathbf{m}] = \sigma_d^2\,[\mathbf{G}^T\mathbf{G}]^{-1}$$

we have

$$[\text{cov}\,\mathbf{m}] = \sigma_d^2\,[\mathbf{G}^T\mathbf{G}]^{-1} = \sigma_d^2 \left[\frac{1}{2}\frac{\partial^2 E}{\partial \mathbf{m}^2}\right]^{-1}_{\mathbf{m}=\mathbf{m}^{est}}$$

$$[\text{cov}\,\mathbf{m}] = \sigma_d^2\,[\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1} = \sigma_d^2 \left[\frac{1}{2}\frac{\partial^2 E}{\partial \mathbf{m}^2}\right]^{-1}_{\mathbf{m}=\mathbf{m}^{est}}$$

the sharper the minimum
the higher the curvature
the smaller the covariance