Technical Computing for the Earth Sciences, Lecture 2:

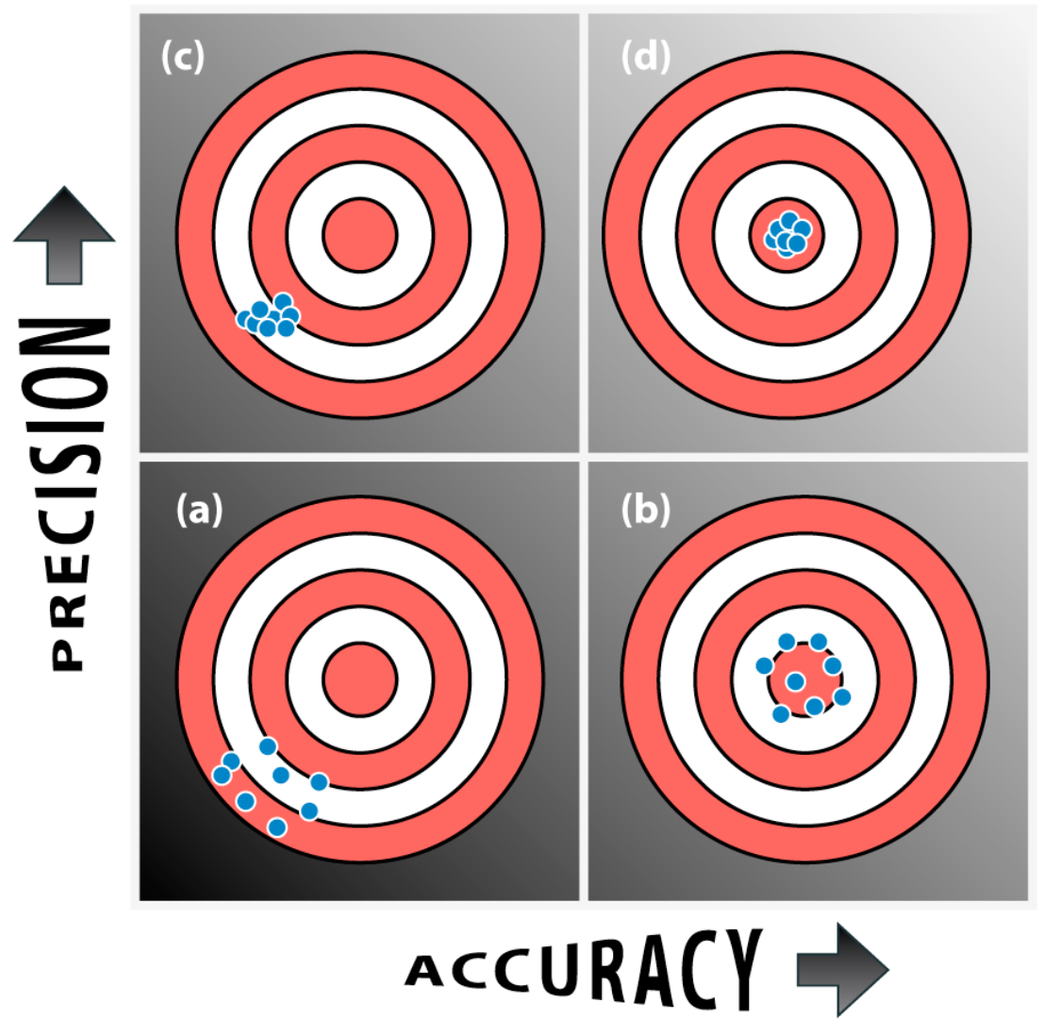
Data and uncertainty
Datasets as vectors and matrices

EARS 80.03

Uncertainty

What is uncertainty?

An estimate of “distance from some true or central value”?

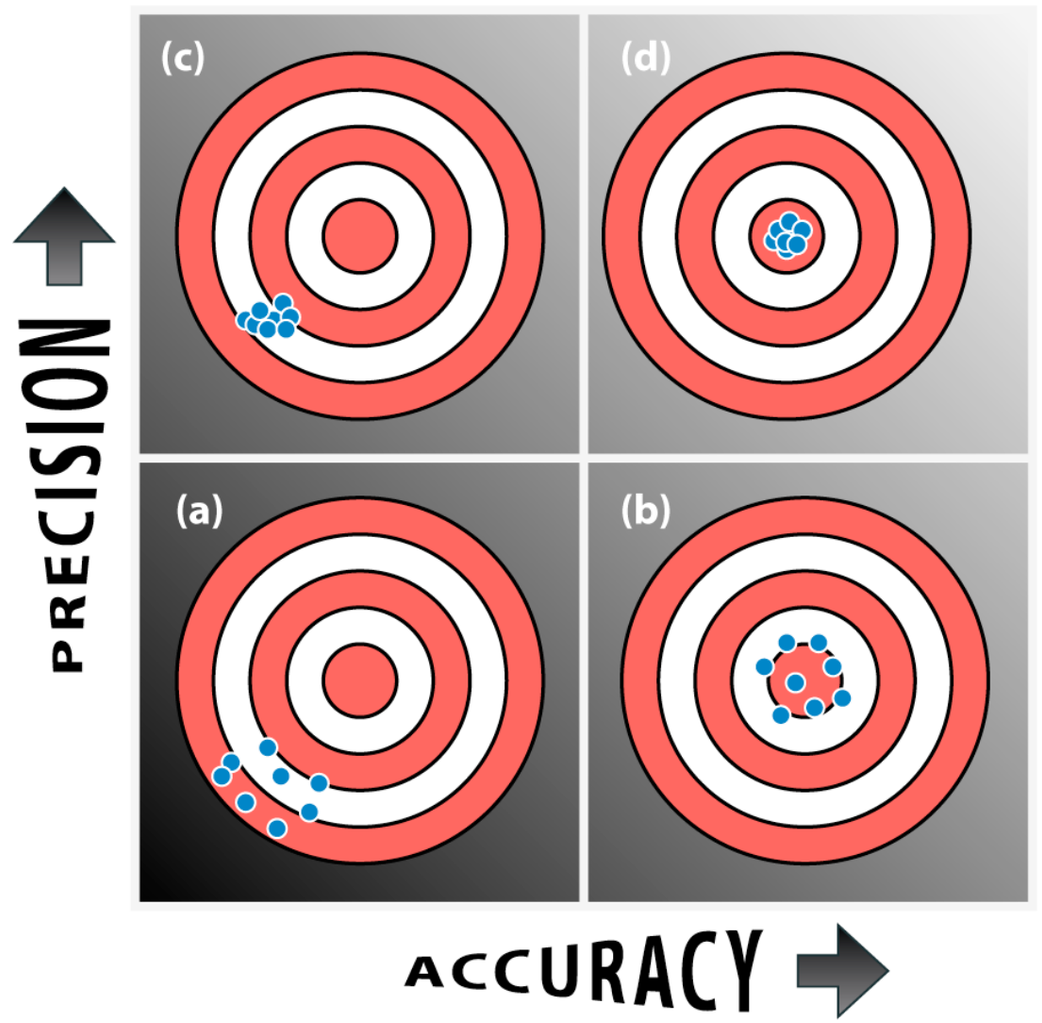


Uncertainty

What is uncertainty?

An estimate of “distance from some true or central value”?

- Accuracy: distance from true value
- Precision: distance from central value (scatter)
- Can’t always determine accuracy, but can always determine precision



First, some terminology around vectors and matrices

Vectors

A vector is a one-dimensional array of numbers

(or as B & V put it, “an ordered finite list of numbers”)

e.g.:

$$\begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix}$$

- A vector of length *n* can be thought of as a point in *n*-dimensional space

Matrices

A matrix is a two-dimensional array of numbers

e.g.:

$$\begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 0 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix}$$

- Have rows (horizontal) and columns (vertical)
- Matrices have various interpretations. Can be thought of:
  - as just a bunch of vectors
  - as systems of linear equations
  - as functions that act on vectors
  - etc.

Vectors and matrices

Can be multiplied

e.g.:

$$\begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 0 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix} * \begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix}$$

Vectors and matrices

Can be multiplied

e.g.:

$$\begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 0 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix} * \begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix}$$

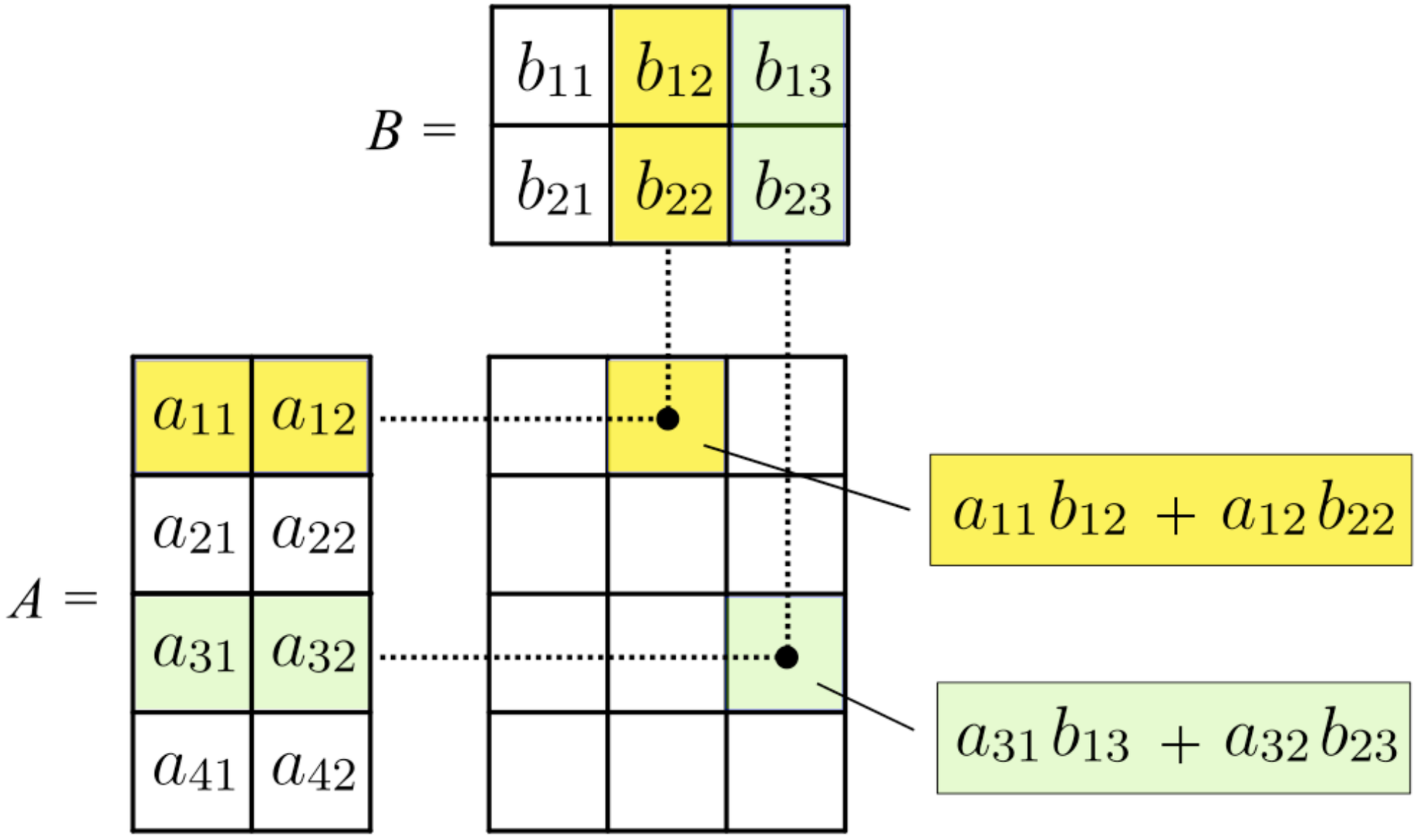
$0 \times -1.1 + 1 \times 0 + -2.3 \times 3.6 + 0.1 \times -7.2$

etc.

Vectors and matrices

Can be multiplied

e.g.:



Vectors and matrices

Can be multiplied

e.g.:

$$\begin{bmatrix} 0 & 1 & -2.3 & 0.1 \\ 1.3 & 4 & -0.1 & 0 \\ 4.1 & -1 & 0 & 1.7 \end{bmatrix} * \begin{bmatrix} -1.1 \\ 0.0 \\ 3.6 \\ -7.2 \end{bmatrix}$$

-9

-1.7

-16.75

- Matrix multiplication isn’t commutative
- Have to be right size (# columns in left matrix have to equal # rows in right matrix)
- Try it out in Julia to see what does and doesn’t work

Uncertainty

How do we measure uncertainty?

“Distance from some true or central value”?

Two problems:

A. How do we measure distance?

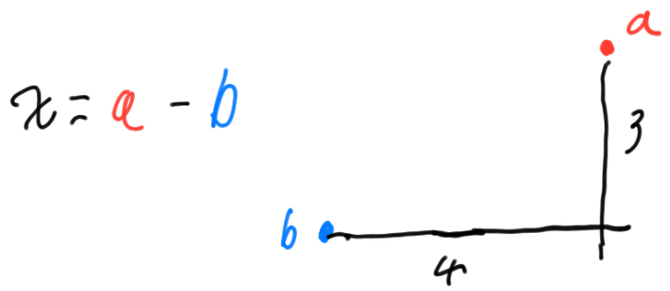
B. What should we use as the true or central value?

A: How do we measure distance

Norms

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Examples:



p = 0

- Technically not valid
- Sometimes said to be # of non-zero dimensions of **x** (i.e., # of dimensions in which **a** and **b** differ)

p = 1

- “Manhattan norm” or “taxicab norm”

p = 2

- “Euclidian norm” or “Euclidian distance”

lim p -> ∞

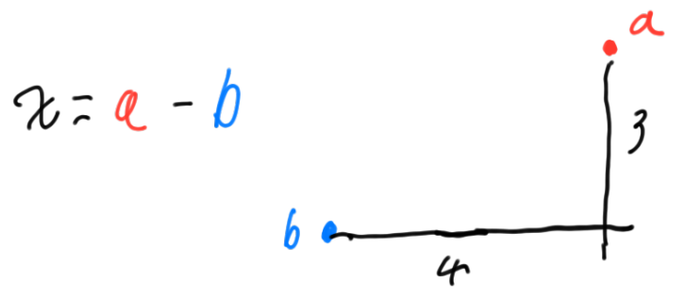
- Largest element of **x**

A: How do we measure distance

Norms

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Examples:



p = 0

- Technically not valid
- Sometimes said to be # of non-zero dimensions of **x** (i.e., # of dimensions in which **a** and **b** differ)

$= 1+1$
$= 2$

p = 1

- “Manhattan norm” or “taxicab norm”

$= 4+3$
$= 7$

p = 2

- “Euclidian norm” or “Euclidian distance”

$= \sqrt{4^2+3^2}$
$= \sqrt{16+9}$
$= \sqrt{25}$
$= 5$

lim p -> ∞

- Largest element of **x**

$= 4$

Let's take off the outer ^(1/p) just to make this a bit simpler

- *Discrepancy*

Say we have some dataset ***D*** containing a whole bunch of data points
*Let's assume these points (vectors) are one-dimensional (i.e, just numbers) to make this easier to think about

What is the value **m** that minimizes the discrepancy of **D** - **m**, i.e.
$\Sigma |\mathbf{D}_i - \mathbf{m}|^p$?

p = 0

Want a point **m** that…

- Minimizes the number of times that one of the points in **D** is not equal to **m**

*If we're willing to take the slightly non-standard position that 0^0 = 0

p = 1

Want a point **m** that…

- Minimizes the average distance between **m** and each point in **D**

p = 2

Want a point **m** that…

- Minimizes the average *squared* distance between **m** and each point in **D**

*Link for more detail

Let's take off the outer ^(1/p) just to make this a bit simpler

- *Discrepancy*

Say we have some dataset ***D*** containing a whole bunch of data points
*Let's assume these points (vectors) are one-dimensional (i.e, just numbers) to make this easier to think about

What is the value **m** that minimizes the discrepancy of **D** - **m**, i.e.
$\Sigma |\mathbf{D}_i - \mathbf{m}|^p$?

Say **D** = [1,1,2,3,8]

p = 0

Want a point **m** that…

- Minimizes the number of times that one of the points in **D** is not equal to **m**

*If we're willing to take the slightly non-standard position that 0^0 = 0

p = 1

Want a point **m** that…

- Minimizes the average distance between **m** and each point in **D**

p = 2

Want a point **m** that…

- Minimizes the average *squared* distance between **m** and each point in **D**

*Link for more detail

Let's take off the outer ^(1/p) just to make this a bit simpler

- *Discrepancy*

Say we have some dataset ***D*** containing a whole bunch of data points
*Let's assume these points (vectors) are one-dimensional (i.e, just numbers) to make this easier to think about

What is the value **m** that minimizes the discrepancy of **D** - **m**, i.e.
$\Sigma |\mathbf{D}_i - \mathbf{m}|^p$?

Say **D** = [1,1,2,3,8]

p = 0

Want a point **m** that…

- Minimizes the number of times that one of the points in **D** is not equal to **m**

*If we're willing to take the slightly non-standard position that 0^0 = 0

= 1

p = 1

Want a point **m** that…

- Minimizes the average distance between **m** and each point in **D**

= 2

p = 2

Want a point **m** that…

- Minimizes the average *squared* distance between **m** and each point in **D**

= 3

*Link for more detail

Let's take off the outer ^(1/p) just to make this a bit simpler

- *Discrepancy*

Say we have some dataset ***D*** containing a whole bunch of data points
*Let's assume these points (vectors) are one-dimensional (i.e, just numbers) to make this easier to think about

What is the value **m** that minimizes the discrepancy of **D** - **m**, i.e.
$\Sigma |\mathbf{D}_i - \mathbf{m}|^p$?

Say **D** = [1,1,2,3,8]

p = 0

Want a point **m** that…

- Minimizes the number of times that one of the points in **D** is not equal to **m**

*If we're willing to take the slightly non-standard position that 0^0 = 0

= 1

MODE

p = 1

Want a point **m** that…

- Minimizes the average distance between **m** and each point in **D**

= 2

MEDIAN

p = 2

Want a point **m** that…

- Minimizes the average *squared* distance between **m** and each point in **D**

= 3

MEAN

*Link for more detail

What should we use as the metric for distance and for central value when measuring scatter?

- p=1 norm, p=1 central value (median)?

  -> *median absolute deviation*

  = median(abs(x .- median(x)))

- p=2 norm, p=2 central value (mean)?

  -> *standard deviation*

$$= \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2 \right)^{1/2}$$

almost the same as sqrt(mean(x .- mean(x))^2), if it weren't for the -1 in N-1 (that's to do with sample vs population statistics)

Some useful functions for working with data and uncertainty

```
julia> rand(dims) # Uniform random numbers between 0 and 1
julia> randn(dims) # Gaussian random numbers with μ=0, σ=1
```

In stdlibs:

```
julia> using Statistics, StatsBase
julia> mean(data)
julia> std(data) # Standard deviation
julia> median(data)
julia> percentile(data)
```