Bayesian Scientific Computing
Day 1

Daniela Calvetti, Erkki Somersalo

Case Western Reserve University
Department of Mathematics, Applied Mathematics and Statistics

Lyngby, December 2019

Introduction

Mathematical model building has often a *natural* direction (causality, locality):

Forward problem:

$$\text{Cause} \longrightarrow \text{Effect}$$

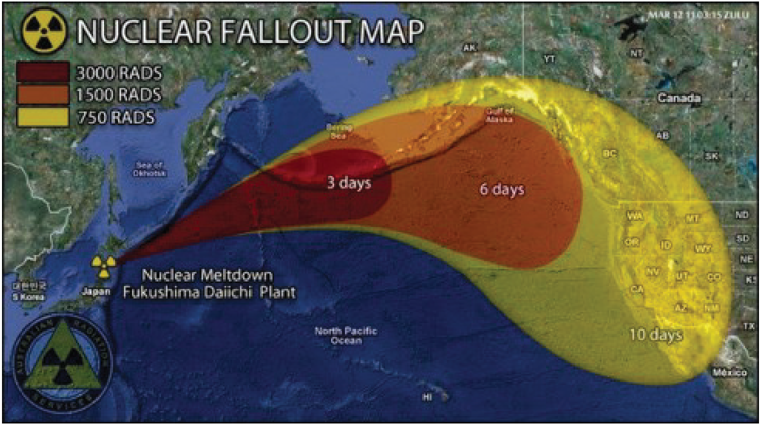
Conceptually tractable, mathematically often the easier (but not necessarily easy) direction.

Inverse problem:

$$\text{Effect} \longrightarrow \text{Cause}$$

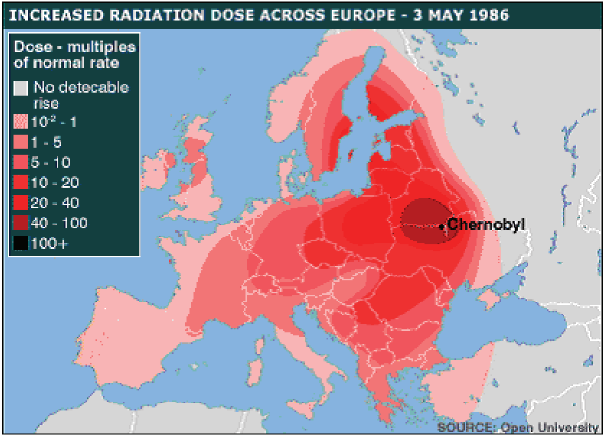
Conceptually and technically challenging (may not even have a solution).

Introduction

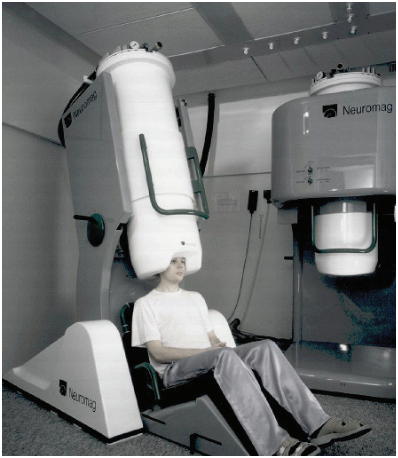


Forward problem: Predict the nuclear fallout
Inverse problem: Find the source of the leak

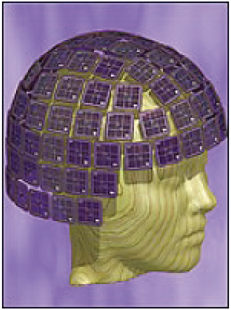
Example: Non-cooperative Target

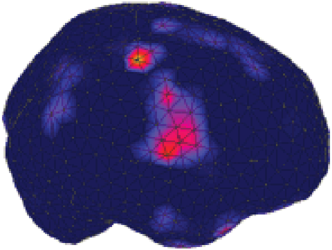


Example: Magnetoencephalography (MEG)



Example: Magnetoencephalography (MEG)





Biot-Savart law for a current dipole at $\vec{r} = \vec{r}_0$:

$$b(\vec{r}) = \vec{n}(\vec{r}) \cdot \vec{B}(\vec{r}) = \frac{\mu_0}{4\pi} \frac{\vec{n}(\vec{r}) \cdot \vec{q} \times (\vec{r} - \vec{r}_0)}{|\vec{r} - \vec{r}_0|^3}.$$

Example: Magnetoencephalography (MEG)

Many dipoles, many measurements:

$$b_j = \sum_{n=1}^{N} \sum_{k=1}^{3} a_{jn}^k q_k^n + \mathrm{noise},$$

or in matrix form,

$$b = \mathsf{A}x + e.$$

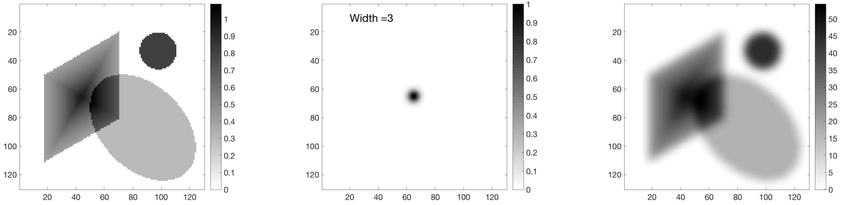
where

$$e = \text{noise}.$$

Inverse problems and ill-posedness

Blurring by convolution,

$$g(r) = \int K(r - r')f(r')dr' + \varepsilon \xrightarrow{\text{discretize}} b = Ax + \varepsilon.$$



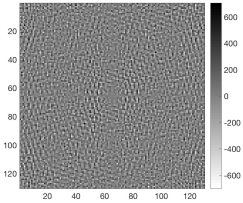
Solving with or without regularization

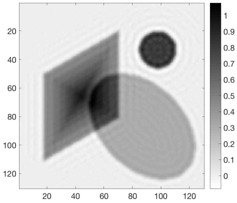
Naive inversion,

$$x_0 = \mathsf{A}^\dagger b,$$

versus Tikhonov regularized inversion,

$$x_\alpha = \mathrm{argmin}\{\|b - \mathsf{A}x\|^2 + \alpha^2\|x\|^2\}.$$





even with data with 5 digits of accuracy.

Causality and Loss of Information

- Second Law of Thermodynamics: "The entropy of a closed system can never decrease: $\Delta S \geq 0$ ".
- Interpretation: The arrow of time points in larger entropy $\Rightarrow$ In causal processes, information is lost (Entropy = lack of information)
- Solving inverse problems is a fight against the Second Law of Thermodynamics.
- To be successful, lost information needs to be replaced by extra information.

Bayesian methods are a systematic way
to merge information from different sources.

Example from Thermodynamics

Consider a thin rod of unit length, identified with the interval $[0,1]$.

$$u(x,t) = \text{temperature at } x \in [0,1] \text{ at time } t \geq 0.$$

End points held at fixed temperature,

$$u(0,t) = u(1,t) = a = \text{known constant}.$$

The temperature satisfies the heat equation,

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2},$$

where D is the heat conductivity of the material.

Inverse Problem: Against Causality

Inverse Problem:*Given the final temperature distribution $u(x, T)$, is it possible to recover $u(x, 0)$?*

$$u(x,0) \stackrel{\text{forward}}{\longrightarrow} u(x,T) \stackrel{\text{inverse}}{\longrightarrow} u(x,0).$$

Question: *What if we simply reverse time: $t \to -t$? Define $v(x,t) = u(x,T-t)$,*

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}, \quad \longrightarrow \quad -\frac{\partial v}{\partial t} = D\frac{\partial^2 v}{\partial x^2},$$

with initial value

$$v(x,0) = u(x,T).$$

Obviously $v(x,T) = u(x,0)$. Can we solve it numerically? Let's try (Excercise)!

Inverse Problems and Bayesian Inference

- In inverse problems, the goal is to estimate a quantity that is not directly observable, by using indirect observations
- In statistical inference, the goal is to estimate a probability distribution based on random draws from it

What is the connection?

- Noisy measurements are random draws from a distribution
- The distribution depends on the unknown quantity that we are interested in

Bayesian approach: *Develop a formalism to express what one can* **believe** *about the values of the unknown, given the measured data.*

Probability

Two classical ways of *understanding* probability:

- *Frequentist's definition*: Probability can be understood in terms of frequencies in repeated experiments
- *Bayesian definition:* Probability is a subject's expression of degree of belief.

Bayesian probability = subjective probability

Bayesian Subjective Probability

Example: Predict the continuation of a sequence:

26535	89793	23846	26433	83279	50288	41971
69399	37510	58209	74944	59230	78164	...

Subject 1:

- The sequence looks random
- Can be modeled as a realization of a random process
- The distribution is estimated by counting the frequencies of the numbers $0 \cdots 9$

Bayesian Subjective Probability

Example: Predict the continuation of a sequence:

$$\begin{array}{lllllllll} 3,14159 & 26535 & 89793 & 23846 & 26433 & 83279 & 50288 & 41971 \\ & 69399 & 37510 & 58209 & 74944 & 59230 & 78164 & \ldots \end{array}$$

Subject 2:

- This looks like the decimal approximation of π
- Prediction from a reference source
- Minimal or no uncertainty (although Subject 2 may be wrong!).

Bayesian Subjective Probability

Bayesian probability:

- Expresses a subject's level of uncertainty (lack of information).
- Asserts that randomness is not the object's but the *subject's* property.
- May be subjective, but needs to be defendable.

Note: Subjective is *not* the same as arbitrary.

Deterministic Quantities as Random Variables

"Why should we interpret a fully deterministic quantity as a random variable?"

Consider the following two "experiments":

1. Predict the outcome of a coin toss I'm going to make.
2. Guess the outcome of a coin toss I already made without showing you the outcome.

- In the first experiment, the outcome is "naturally" a random variable (result of a random process).
- In the second one, the outcome is fully deterministic; however, you don't know the outcome.

The same concepts of probability apply to both experiments.

Bayes' Formula

Reverend Thomas Bayes (1701–1761), philosopher, theologist.

Bayes' formula appeared in his paper

- "An Essay towards solving a Problem in the Doctrine of Chances", read posthumously in the Royal Society in 1763.
- Bayes's motivation: Studied the proofs for existence of God.

Bayes' Formula

Pierre-Simon Laplace (1749 - 1827), French mathematician and natural scientist, one of the developers of modern probability and statistics.

- Gambling
- Astronomy

- " Mémoires sur la probabilité des causes par les évènemens" (1774): *Inverse probability*.
- "Memoir on comets" (1813) Contains Bayes' formula in its present form, and a scientific application.

Inverse Probability

Inverse probability (through an example)

- Given an urn with a known number of black and white balls, we can compute the probability of drawing (say) a white ball ("forward probability").
- The inverse probability is to figure out the ratio between the numbers of white and black balls, given observed draws from the urn.

Subjective Probability

BRUNO DE FINETTI 1906–1985

"Probability does not exist!"
(B. de Finetti: *Theory of Probability*)

Subjective Probability

Probability = uncertainty, or subject's lack of information.

> *"The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence."*

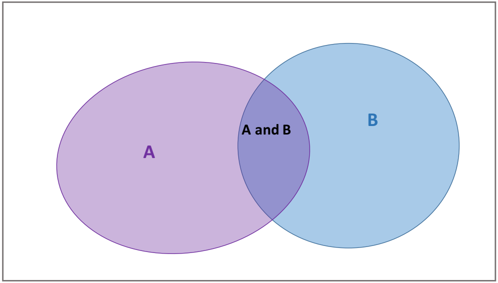
– Bruno de Finetti

Elementary Version of Bayes' Formula

$$\boxed{\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\mathrm{P}(A)}{\mathrm{P}(B)}}$$

- A, B are events. The even B is observed.
- $\mathrm{P}(A) =$ the probability of the event A.
- $\mathrm{P}(B) =$ the probability of the event B.
- $\mathrm{P}(B \mid A) =$ probability of B assuming that A happens.
- $\mathrm{P}(A \mid B) =$ probability of A assuming that B happens.

Visual Justification of Bayes' Formula



$$
\begin{aligned}
\mathrm{P}(A,B) &= \mathrm{P}(A \mid B)\mathrm{P}(B) \\
&= \mathrm{P}(B \mid A)\mathrm{P}(A),
\end{aligned}
$$

$$
\mathrm{P}(A) = \mathrm{P}(A \mid B)\mathrm{P}(B) + \mathrm{P}(A \mid \neg B)\mathrm{P}(\neg B).
$$

Example

A woman detects a lump in her breast and gets a mammography. Later, she is recalled for further diagnostics due to a positive mammography result. What is her probability of having breast cancer?

- Event $B =$ positive finding that makes the doctor suspect cancer.
- Event $A =$ the patient has breast cancer.

Example

Assume the following statistical data, based on long-term patient records:

	Malignant tumor	Benign tumor
Positive (recall)	0.8	0.6
Negative (no recall)	0.2	0.4

Interpretation in terms of conditional probabilities:

$$
\begin{aligned}
\mathrm{P}(B \mid A) &= 0.8, \quad \mathrm{P}(B \mid \neg A) = 0.6 \\
\mathrm{P}(\neg B \mid A) &= 0.2, \quad \mathrm{P}(\neg B \mid \neg A) = 0.4
\end{aligned}
$$

Here, $\neg A =$ "not A."

Example

More information: According medical records, for one out of four patients reporting a lump in the breast, the tumor is a malignant lesion.
Interpretation:

$$\mathrm{P}(A) = 0.25, \quad P(\neg A) = 0.75.$$

We still need $\mathrm{P}(B)$. This is obtained as follows.

$$\mathrm{P}(B) = \mathrm{P}(B \mid A)\mathrm{P}(A) + \mathrm{P}(B \mid \neg A)\mathrm{P}(\neg A),$$

that is, "B may happen with A or with $\neg A$." In numbers:

$$\mathrm{P}(B) = 0.8 \times 0.25 + 0.6 \times 0.75 = 0.65.$$

Example

Use Bayes' formula:

$$\begin{aligned}
\mathrm{P}(A \mid B) &= \frac{\mathrm{P}(B \mid A)\mathrm{P}(A)}{\mathrm{P}(B)} \\
&= \frac{0.8 \times 0.25}{0.65} \approx 0.31.
\end{aligned}$$

We see that the spontaneous reaction "*The patient has cancer with probability 0.8*" is way too pessimistic!

Prosecutor's Fallacy

A newborn dies without a visible reason (SIDS). The probability of such event is very low; should one suspect that the parents have killed the baby?

Sally Clark case, UK 1998: Two newborns died of SIDS, and the mother was accused of murder.

Argument: "1/73 million chance that SIDS happens independently twice in the same family."

A faulty reasoning leads to those suspicions, and a probabilistic one cleared the mother.

- Event B = Newborn dies with no visible reason.
- Event A = The parents have killed the baby.

Then

$$\mathrm{P}(B) = \mathrm{P}(B \mid A)\mathrm{P}(A) + \mathrm{P}(B \mid \neg A)\mathrm{P}(\neg A),$$

How big must $\mathrm{P}(A)$ to support the murder conclusion?

Random Variables, Continuum State Space

Probability distribution of a real valued random variable X,

$$\mu_X\big((a,b)\big) = \mathrm{P}\big\{a < X < b\big\}.$$

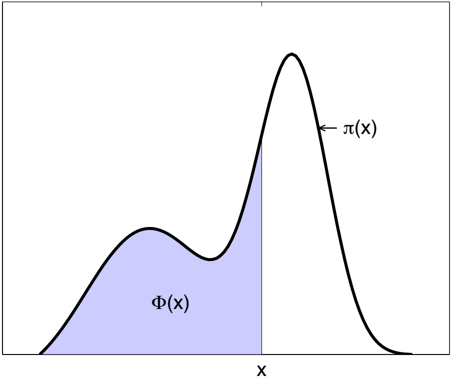
If a function $\pi_X(x) \geq 0$ exists such that

$$\mu_X(B) = \int_B \pi_X(x)dx, \quad B \subset \mathbb{R},$$

we refer to π_X as the *probability density of* X.
Normalization:

$$\int_{\mathbb{R}} \pi_X(x)dx = \mathrm{P}\{X \in \mathbb{R}\} = 1.$$

Random Variables, Continuum State Space



Cumulative distribution

$$\Phi_X(x) = \mathrm{P}\{X < x\} = \int_{-\infty}^{x} \pi_X(x')dx'.$$

Joint Probability Density

For two random variables X and Y, define

$$\mu_{XY}(A \times B) = \mathrm{P}\big\{X \in A, Y \in B\big\},$$

where μ_{XY} is the joint probability distribution.
Assume the existence of of the joint probability density π_{XY},

$$\mu_{XY}(A \times B) = \int_A \int_B \pi_{XY}(x, y) dydx.$$

Again,

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \pi_{XY}(x, y) dxdy = 1.$$

Multivariate Random Variables

Real valued random variables $X_1, X_2, \ldots, X_n$.
Define

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \in \mathbb{R}^n.$$

Joint probability density,

$$\pi_X(x_1, x_2, \ldots, x_n) \geq 0,$$

defines the probability density in $\mathbb{R}^n$: For $B \in \mathbb{R}^n$,

$$\mathrm{P}\{X \in B\} = \int_B \pi_X(x)dx = \int \cdots \int_B \pi_X(x_1, \ldots, x_n)dx_1 \ldots dx_n.$$

Marginal Densities

Define the marginal densities,

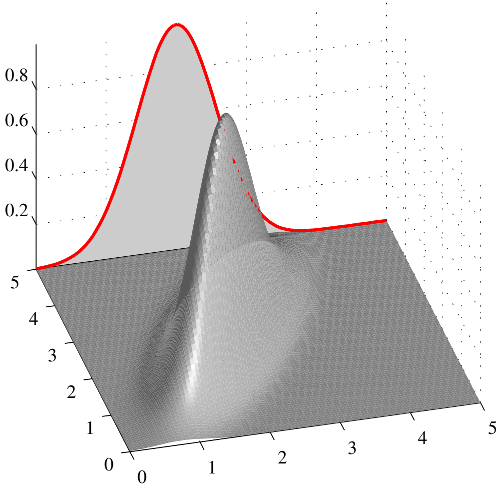
$$\pi_X(x) = \int_{\mathbb{R}} \pi_{XY}(x,y)dy, \quad \pi_Y(y) = \int_{\mathbb{R}} \pi_{XY}(x,y)dx,$$

which satisfy

$$\mathrm{P}\big\{X \in A\big\} = \int_A \pi_X(x)dx = \int_A \int_{\mathbb{R}} \pi_{XY}(x,y)dydx = \mathrm{P}\big\{X \in A, Y \in \mathbb{R}\big\},$$

$$\mathrm{P}\big\{Y \in B\big\} = \int_B \pi_Y(y)dy = \int_B \int_{\mathbb{R}} \pi_{XY}(x,y)dxdy = \mathrm{P}\big\{Y \in B, X \in \mathbb{R}\big\}.$$

Marginal Density Visualized



Conditional Densities

Since

$$\pi_Y(y) = \int_{\mathbb{R}} \pi_{XY}(x,y)dx,$$

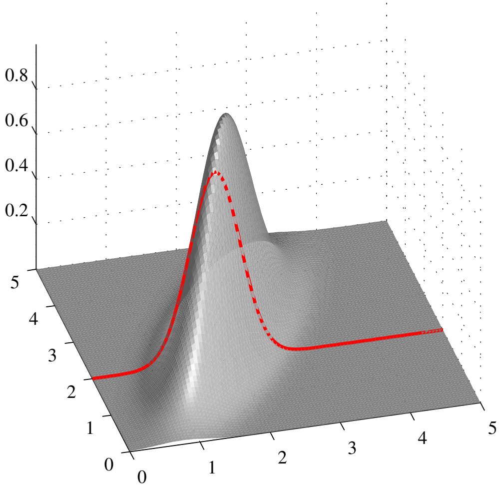
so if $\pi_Y(y) \neq 0$, we have

$$\frac{\pi_{XY}(x,y)}{\pi_Y(y)} \geq 0, \quad \int_{\mathbb{R}} \frac{\pi_{XY}(x,y)}{\pi_Y(y)} dx = 1.$$

Define a probability density $\pi_{X|Y}(x \mid y)$ through the formula

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{XY}(x,y)}{\pi_Y(y)}.$$

Conditional Density Visualized



Independency

Two random variables are **independent if and only if** the joint probability density can be factored as

$$\boxed{\pi_{XY}(x, y) = \pi_X(x)\pi_Y(y).}$$

Observe: For independent variables

$$\pi_{X|Y}(x \mid y) = \frac{\pi_{XY}(x, y)}{\pi_Y(y)} = \frac{\pi_X(x)\pi_Y(y)}{\pi_Y(y)} = \pi_X(x),$$

- By observing Y we learn nothing about X.

Converse: If

$$\pi_{X|Y}(x \mid y) = \pi_X(x),$$

the joint probability density factors and independency follows.

Bayes' Formula

Interpretation: The conditional density is the probability density of X, assuming that the variable Y takes on value $Y = y$.
Symmetrically,

$$\pi_{Y|X}(y \mid x) = \frac{\pi_{XY}(x,y)}{\pi_X(x)},$$

assuming that $\pi_X(x) \neq 0$.
Solving for the joint density leads to

$$\pi_{XY}(x,y) = \pi_{X|Y}(x \mid y)\pi_Y(y) = \pi_{Y|X}(y \mid x)\pi_X(x),$$

Bayes' Formula

The equation

$$\pi_{XY}(x,y) = \pi_{X\mid Y}(x \mid y)\pi_Y(y) = \pi_{Y\mid X}(y \mid x)\pi_X(x),$$

implies

Bayes' formula for probability densities:

$$\boxed{\pi_{X\mid Y}(x \mid y) = \frac{\pi_{Y\mid X}(y \mid x)\pi_X(x)}{\pi_Y(y)}.}$$

This is the key formula in Bayesian scientific computing

$$\begin{aligned} \pi_X &= \text{prior density},\\ \pi_{Y\mid X} &= \text{likelihood density},\\ \pi_{X\mid Y} &= \text{posterior density}. \end{aligned}$$

$$\boxed{\text{posterior} \propto \text{prior} \times \text{likelihood}.}$$

Expectation, Variance

Given a $\mathbb{R}$-valued random variable X, the *expectation* is the center of mass of the probability distribution,

$$\mathsf{E}\{X\} = \int_{\mathbb{R}} x\pi_X(x)dx = \overline{x}.$$

The *variance* is the expectation of the squared deviation from the expectation,

$$\mathrm{var}(X) = \mathsf{E}\{(X - \overline{x})^2\} = \int_{\mathbb{R}} (x - \overline{x})^2 \pi_X(x)dx,$$

assuming that the integrals converge.

Example: Gaussian Distributions

A random variable $X \in \mathbb{R}$ is normally distributed, or Gaussian,

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

if

$$\mathrm{P}\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx.$$

Probability density:

$$\pi_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Expectation, Variance

Example: Gaussian distribution:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = \mu,$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = \sigma^2,$$

Expectation, Variance

Example: Cauchy distribution:

$$\pi_X(x) = \frac{\alpha}{\pi}\frac{1}{1+\alpha^2 x^2},$$

$$\int_{-\infty}^{\infty} |x|\pi_X(x)dx = \infty,$$

and therefore, the integral defining the expectation is non-convergent.

Multivariate vs. Univariate

Define

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

with each component X_i being an $\mathbb{R}$-valued variable.
Probability density of X = joint probability density $\pi = \pi_X : \mathbb{R}^n \to \mathbb{R}_+$ of its components.
Expectation is

$$\overline{x} = \int_{\mathbb{R}^n} x\pi_X(x)dx \in \mathbb{R}^n,$$

or, componentwise,

$$\overline{x}_i = \int_{\mathbb{R}^n} x_i\pi_X(x)dx \in \mathbb{R}, \quad 1 \leq i \leq n.$$

Multivariate vs. Univariate

Observe:

$$\int_{\mathbb{R}^n} x_1 \pi_X(x) dx = \int_{\mathbb{R}} x_1 \underbrace{\int_{\mathbb{R}^{n-1}} \pi_X(x_1, x_2, \ldots, x_n) dx_2 \ldots dx_n}_{\text{marginal density of } x_1}$$

$$= \int_R x_1 \pi_{X_1}(x_1) dx_1.$$

Similarly the other components.

Multivariate vs. Univariate

The *covariance matrix* is defined as

$$\mathrm{cov}(X) = \int_{\mathbb{R}^n} (x - \overline{x})(x - \overline{x})^\mathsf{T} \pi_X(x) dx \in \mathbb{R}^{n \times n},$$

or, componentwise,

$$\mathrm{cov}(X)_{ij} = \int_{\mathbb{R}^n} (x_i - \overline{x}_i)(x_j - \overline{x}_j) \pi_X(x) dx \in \mathbb{R}^{n \times n}, \quad 1 \leq i, j \leq n.$$

Multivariate vs. Univariate

Recall: A matrix $\mathsf{C} \in \mathbb{R}^{m\times m}$ is *symmetric, positive definite* (SPD), if

1. $$\mathsf{C}^{\mathsf{T}} = \mathsf{C}, \quad \text{(symmetry)}$$

2. $$v^{\mathsf{T}}\mathsf{C}v > 0 \text{ for all } v \neq 0. \quad \text{(positive definiteness)}$$

Equivalent conditions: :

(2a.) C admits a Cholesky factorization,

$$\mathsf{C} = \mathsf{R}^{\mathsf{T}}\mathsf{R}, \qquad \mathsf{R} \text{ is upper triangular, } \mathsf{R}_{jj} > 0.$$

(2b.) C admits an eigenvalue decomposition

$$\mathsf{C} = \mathsf{V}\mathsf{D}\mathsf{V}^{\mathsf{T}}, \quad \mathsf{V} \text{ orthogonal, } \mathsf{D} \text{ diagonal, } \mathsf{D}_{jj} > 0.$$

Multivariate vs. Univariate

Covariance matrix is *symmetric* and *positive semi-definite*: For any $v \in \mathbb{R}^n$, $v \neq 0$,

$$\begin{aligned} v^{\mathsf{T}}\mathrm{cov}(X)v &= \int_{\mathbb{R}^n} \left[v^{\mathsf{T}}(x-\overline{x})\right]\left[(x-\overline{x})^{\mathsf{T}}v\right]\pi_X(x)dx \qquad (1) \\ &= \int_{\mathbb{R}^n} \left(v^{\mathsf{T}}(x-\overline{x})\right)^2\pi_X(x)dx \geq 0. \end{aligned}$$

$$v^{\mathsf{T}}\mathrm{cov}(X)v = \text{ variance of } X \text{ into the direction } v.$$

Usually, it is assumed that X has non-vanishing variance in all directions, and the covariance is SPD.

Multivariate vs. Univariate

Denote by $x_i' \in \mathbb{R}^{n-1}$ the vector x with the ith component deleted:

$$
\begin{aligned}
\mathrm{cov}(X)_{ii} &= \int_{\mathbb{R}^n} (x_i - \overline{x}_i)^2 \pi_X(x) dx \\
&= \int_{\mathbb{R}} (x_i - \overline{x}_i)^2 \underbrace{\left(\int_{\mathbb{R}^{n-1}} \pi_X(x_i, x_i') dx_i' \right)}_{=\pi_{X_i}(x_i)} dx_i \\
&= \int_{\mathbb{R}} (x_i - \overline{x}_i)^2 \pi_{X_i}(x_i) dx_i = \mathrm{var}(X_i).
\end{aligned}
$$

Conclusion: Diagonal of the covariance matrix gives the variances of the individual components.

Multivariate vs. Univariate

Consider two random variables, X_1 and X_2, and assume that they are independent and Gaussian:

$$X_j \sim \mathcal{N}(0, \sigma_j^2), \quad j = 1, 2.$$

$$\begin{aligned}
\mathrm{P}\big\{a_1 < x_1 < b_1, a_2 < x_2 < b_2\big\} &= \mathrm{P}\big\{a_1 < x_1 < b_1\big\}\mathrm{P}\big\{a_2 < x_2 < b_2\big\} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} \int_{a_1}^{b_1} \exp\left(-\frac{1}{2\sigma_1^2}x_1^2\right) dx_1 \times \frac{1}{\sqrt{2\pi\sigma_2^2}} \int_{a_2}^{b_2} \exp\left(-\frac{1}{2\sigma_2^2}x_2^2\right) dx_2 \\
&= \int\int_Q \pi_{X_1 X_2}(x_1, x_2) dx_1, dx_2,
\end{aligned}$$

Multivariate vs. Univariate

where

$$Q = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2,$$

and $\pi_{X_1 X_2}(x_1, x_2) = \pi_X(x)$ is a two-dimensional Gaussian density,

$$\pi_X(x) = \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right)\right)$$

Multivariate vs. Univariate

Defining

$$\mathsf{C} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix},$$

the two-dimensional Gaussian density can be written as

$$\begin{aligned} \pi_X(x) &= \frac{1}{\sqrt{(2\pi)^2\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2}\right)\right) \\ &= \frac{1}{\sqrt{(2\pi)^2|\mathsf{C}|}} \exp\left(-\frac{1}{2}x^{\mathsf{T}}\mathsf{C}^{-1}x\right), \end{aligned}$$

where

$$|\mathsf{C}| = |\det(\mathsf{C})| = \sigma_1^2\sigma_2^2.$$

Multivariate vs. Univariate

This leads to the definition of a general *multivariate Gaussian probability density with independent components*:

Let

$$\mathsf{C} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}.$$

Define

$$\pi_X(x) = \frac{1}{\sqrt{(2\pi)^m |\mathsf{C}|}} \exp\left(-\frac{1}{2} x^\mathsf{T} \mathsf{C}^{-1} x\right),$$

where $|\mathsf{C}| = |\det(\mathsf{C})| = \sigma_1^2 \sigma_2^2 \cdots \sigma_n^2$.

Multivariate vs. Univariate

Geometric intuition in two dimensions: The curve

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = \text{constant}$$

represents an ellipse with principal axes along the coordinate axes.
Add a rotation: Define

$$\mathsf{U} = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right],$$

and rotated random variable X',

$$X' = \mathsf{U}^{\mathsf{T}} X.$$

Assume that the components X_1' and X_2' are independent Gaussian,

$$X_j' \sim \mathcal{N}(0, \sigma_j^2), \quad j = 1, 2.$$

Multivariate vs. Univariate

If

$$\mathsf{D} = \begin{bmatrix} \sigma_1^2 & \\ & \sigma_2^2 \end{bmatrix},$$

then the probability density of X' is

$$\pi_{X'}(x') = \frac{1}{\sqrt{(2\pi)^m |\mathsf{D}|}} \exp\left(-\frac{1}{2}(x')^\mathsf{T}\mathsf{D}^{-1}x'\right),$$

Probability density of $X = \mathsf{U}x'$: Jacobian determinant of a rotation $=1$, so

$$x = \mathsf{U}x' \Rightarrow dx = dx'.$$

Multivariate vs. Univariate

$$
\begin{aligned}
\pi_X(x) &= \frac{1}{\sqrt{(2\pi)^m|\mathsf{D}|}}\exp\left(-\frac{1}{2}(\mathsf{U}^\mathsf{T}x)^\mathsf{T}\mathsf{D}^{-1}\mathsf{U}^\mathsf{T}x\right) \\
&= \frac{1}{\sqrt{(2\pi)^m|\mathsf{D}|}}\exp\left(-\frac{1}{2}x^\mathsf{T}(\mathsf{U}\mathsf{D}^{-1}\mathsf{U}^\mathsf{T})x\right) \\
&= \frac{1}{\sqrt{(2\pi)^m|\mathsf{C}|}}\exp\left(-\frac{1}{2}x^\mathsf{T}\mathsf{C}^{-1}x\right),
\end{aligned}
$$

where

$$
\mathsf{C} = \mathsf{U}\mathsf{D}\mathsf{U}^\mathsf{T}, \quad |\mathsf{C}| = |\mathsf{D}|.
$$

Multivariate vs. Univariate

Geometric interpretation: The curve

$$x^{\mathsf{T}} \mathsf{C}^{-1} x = \text{constant}, \quad \mathsf{C} \in \mathbb{R}^{2\times 2}$$

is a rotated ellipse.
More generally, in $\mathbb{R}^n$,

$$x^{\mathsf{T}} \mathsf{C}^{-1} x = \text{constant}, \quad \mathsf{C} \in \mathbb{R}^{n\times n}$$

is an ellipsoidal hypersurface.

Multivariate Gaussian

Multivariate extension of Gaussian densities in $\mathbb{R}$: $X \in \mathbb{R}^n$ is Gaussian, if its probability density is

$$\pi_X(x) = \left(\frac{1}{(2\pi)^n|\mathsf{C}|}\right)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\mathsf{C}^{-1}(x-\mu)\right),$$

where $\mu \in \mathbb{R}^n$, $\mathsf{C} \in \mathbb{R}^{n\times n}$ is symmetric positive definite, that is,

$$\mathsf{C}^\mathsf{T} = \mathsf{C}, \quad v^\mathsf{T}\mathsf{C}v > 0 \quad \text{for all } v \neq 0.$$

$$\mathrm{P}\{X \in B\} = \int_B \pi_X(x)dx.$$

Notation,

$$X \sim \mathcal{N}(\mu, \mathsf{C}).$$

Multivariate Gaussian

A straightforward computation shows: If $X \sim \mathcal{N}(\mu, \mathsf{C})$, then

$$\mathsf{E}\{X\} = \mu,$$

$$\mathrm{cov}(X) = \mathsf{C}.$$

Inverse problems are ill-posed

A problem is called *well-posed* if

- It has a solution (existence)
- The solution is unique (uniqueness)
- Small perturbations in the problem setting lead to small perturbation in the solution (stability)

Problems that are not well-posed are called *ill-posed*

Inverse problems are practically always ill-posed; they fail to satisfy at least one, but often several of the conditions defining well-posedness.

Linear algebra: Basic concepts

Inner product: If $x, y \in \mathbb{R}^n$,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

the inner product is defined as

$$x^{\mathsf{T}} y = \sum_{j=1}^{n} x_j y_j = y^{\mathsf{T}} x.$$

The vectors are said to be *orthogonal* if and only if

$$x^{\mathsf{T}} y = 0.$$

Linear algebra: Basic concepts

Matrix-vector product: If $\mathsf{A} \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, then

$$b = \mathsf{A}x \in \mathbb{R}^m, \quad b_k = \sum_{j=1}^{n} A_{kj}x_j, \quad 1 \leq k \leq m.$$

Denoting

$$\mathsf{A} = \left[\begin{array}{cccc} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{array}\right], \quad a_j \in \mathbb{R}^m,$$

we have

$$\mathsf{A}x = \sum_{j=1}^{n} x_j a_j = \text{ linear combination of } a_1, \ldots, a_n.$$

Linear algebra: Basic concepts

Matrix-matrix product: Let $\mathsf{A} \in \mathbb{R}^{m \times n}$, $\mathsf{B} \in \mathbb{R}^{n \times k}$. Then

$$\mathsf{C} = \mathsf{A}\mathsf{B} \in \mathbb{R}^{m \times k}, \quad C_{ij} = \sum_{\ell=1}^{n} A_{i\ell} B_{\ell j}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq k.$$

Denoting

$$\mathsf{B} = \left[\begin{array}{cccc} b_1 & b_2 & \cdots & b_k \end{array} \right],$$

we have

$$\mathsf{A}\mathsf{B} = \left[\begin{array}{cccc} \mathsf{A}b_1 & \mathsf{A}b_2 & \cdots & \mathsf{A}b_k \end{array} \right],$$

Recall:

$$(\mathsf{A}\mathsf{B})^{\mathsf{T}} = \mathsf{B}^{\mathsf{T}}\mathsf{A}^{\mathsf{T}}.$$

Linear algebra: Basic concepts

Given vectors $a_1, a_2, \ldots, a_k \in \mathbb{R}^n$, we say that the vectors are *linearly independent* if

$$c_1 a_1 + c_2 a_2 + \ldots + c_k a_k = 0 \Leftrightarrow c_1 = c_2 = \ldots = c_k = 0.$$

In matrix notation: Denoting

$$\mathsf{A} = \left[\begin{array}{cccc} a_1 & a_2 & \cdots & a_k \end{array}\right] \in \mathbb{R}^{n\times k}, \quad c = \left[\begin{array}{c} c_1 \\ \vdots \\ c_k \end{array}\right],$$

the condition reads

$$\mathsf{A}c = 0_n \Leftrightarrow c = 0_k.$$

Here 0_n (0_k) is the null vector in $\mathbb{R}^n$ $(\mathbb{R}^k)$.

Linear algebra: Basic concepts

Given k vectors $a_1, \ldots a_k$, the subspace spanned by these vectors is defined as

$$H = \mathrm{span}\{a_1, \ldots, a_k\} = \{x \in \mathbb{R}^n \quad | \quad x = c_1 a_1 + \ldots c_k a_k \\ \text{for some } c_1, \ldots, c_k \in \mathbb{R}\}.$$

If the vectors $a_1, \ldots, a_k$ are linearly independent, they form a *basis* of H. The *dimension* of a subspace $H \subset \mathbb{R}^n$ is the maximum number of independent vectors that span the subspace H.

Linear algebra: Basic concepts

Given a matrix $\mathsf{A} \in \mathbb{R}^{m\times n}$,

- The *null space* of A, denoted as $\mathcal{N}(\mathsf{A})$ is defined as

$$\mathcal{N}(\mathsf{A}) = \big\{x \in \mathbb{R}^n \mid \mathsf{A}x = 0\big\}.$$

- The *range* of A, denoted as $\mathcal{R}(\mathsf{A})$ is defined as

$$\mathcal{R}(\mathsf{A}) = \big\{y \in \mathbb{R}^m \mid y = \mathsf{A}x \text{ for some } x \in \mathbb{R}^n\big\}.$$

Linear algebra: Basic concepts

Another characterization for $\mathcal{R}(A)$: Denote the column vectors of A as $a_1, \ldots, a_n$,

$$\mathsf{A} = \left[\begin{array}{cccc} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{array} \right], \quad a_j \in \mathbb{R}^m.$$

We have

$$\mathsf{A}x = \sum_{j=1}^{n} x_j a_j = \text{ linear combination of } a_1, \ldots, a_n.$$

Therefore,

$$\mathcal{R}(\mathsf{A}) = \mathrm{span}\{a_1, \ldots, a_n\} \quad = \quad \begin{array}{l} \text{the linear space of all} \\ \text{linear combinations of columns of A.} \end{array}$$

Linear algebra: Basic concepts

Consider the linear equation

$$\mathsf{A}x = b, \quad \mathsf{A} \in \mathbb{R}^{m\times n}. \tag{2}$$

- If $\mathcal{N}(\mathsf{A}) \neq \{0\}$, the solution of (**??**), if it exists, is non-unique: If $x^* \in \mathbb{R}^n$ is a solution and $0 \neq x_0 \in \mathcal{N}(\mathsf{A})$, then

$$\mathsf{A}(x^* + x_0) = \mathsf{A}x^* + \mathsf{A}x_0 = \mathsf{A}x^* = b,$$

  and so $x^* + x_0$ is also a solution.
- If $\mathcal{R}(\mathsf{A}) \neq \mathbb{R}^m$, the problem (**??**) does not always have a solution: If $b \notin \mathcal{R}(\mathsf{A})$, then, by definition, there is no $x \in \mathbb{R}^n$ such that (**??**) holds.

Linear algebra: Basic concepts

The *rank* of a matrix $A \in \mathbb{R}^{m\times n}$, denoted as $\mathrm{rank}(A)$, has two *equivalent* definitions:

1. $\mathrm{rank}(A) =$ max. number of linearly independent columns of A,
2. $\mathrm{rank}(A) =$ max. number of linearly independent rows of A.

Linear algebra: Basic concepts

Observations:

1. $\mathrm{rank}(\mathsf{A}) \leq \min\{m, n\}$.
2. We have

$$\dim\big(\mathcal{R}(\mathsf{A})\big) = \dim\big(\mathrm{span}\{a_1, \ldots, a_n\}\big) = \mathrm{rank}(\mathsf{A}).$$

Therefore, if $\mathrm{rank}(\mathsf{A}) < m$, the problem $\mathsf{A}x = b$ may not have a solution.

Linear algebra: Basic concepts

Closer look at the null space: Let $x \in \mathcal{N}(\mathsf{A})$, that is $\mathsf{A}x = 0$. Then, for any $y \in \mathbb{R}^m$, we have

$$0 = y^\mathsf{T}\mathsf{A}x = \left(y^\mathsf{T}\mathsf{A}\right)x = \left(\mathsf{A}^\mathsf{T}y\right)^\mathsf{T}x \text{ for all } y \in \mathbb{R}^m.$$

The vectors $\mathsf{A}^\mathsf{T}y$ span the subspace $\mathcal{R}(\mathsf{A}^\mathsf{T}) \subset \mathbb{R}^n$. Conclusion:

$$\mathcal{N}(\mathsf{A}) \perp \mathcal{R}(\mathsf{A}^\mathsf{T}).$$

In fact, we can conclude that

$$\begin{aligned} \mathcal{N}(\mathsf{A}) &= \mathcal{R}(\mathsf{A}^\mathsf{T})^\perp = \left\{x \in \mathbb{R}^n \mid x \perp z \text{ for every } z \in \mathcal{R}(\mathsf{A}^\mathsf{T})\right\} \\ &= \text{orthocomplement of } \mathcal{R}(\mathsf{A}^\mathsf{T}). \end{aligned}$$

Linear algebra: Basic concepts

Denote

$$\mathsf{A} = \begin{bmatrix} \alpha_1^\mathsf{T} \\ \alpha_2^\mathsf{T} \\ \vdots \\ \alpha_m^\mathsf{T} \end{bmatrix}, \quad \alpha_j \in \mathbb{R}^n,$$

or, equivalently,

$$\mathsf{A}^\mathsf{T} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_m \end{bmatrix} \in \mathbb{R}^{n\times m}.$$

Since

$$\mathcal{R}(\mathsf{A}^\mathsf{T}) = \mathrm{span}\{\alpha_1, \ldots, \alpha_m\},$$

we conclude that

$$\dim\big(\mathcal{R}(\mathsf{A}^\mathsf{T})\big) = \mathrm{rank}(\mathsf{A}).$$

Linear algebra: Basic concepts

Conclusion: If $\mathrm{rank}(\mathsf{A}) < n$, then $\mathcal{R}(\mathsf{A}^\mathsf{T}) \neq \mathbb{R}^n$, and its orthocomplement $\mathcal{N}(\mathsf{A})$ is *at least* one-dimensional. In fact,

$$\dim\big(\mathcal{N}(\mathsf{A})\big) = n - \dim\big(\mathcal{R}(\mathsf{A}^\mathsf{T})\big) = n - \mathrm{rank}\big(\mathsf{A}\big).$$

We add a third observation to the previous ones:

Linear algebra: Basic concepts

Observations:

1. $\mathrm{rank}(\mathsf{A}) \leq \min\{m, n\}$.
2. We have
$$\dim\big(\mathcal{R}(\mathsf{A})\big) = \dim\big(\mathrm{span}\{a_1, \ldots, a_n\}\big) = \mathrm{rank}(\mathsf{A}).$$
Therefore, if $\mathrm{rank}(\mathsf{A}) < m$, the problem $\mathsf{A}x = b$ may not have a solution.
3. We have
$$\dim\big(\mathcal{N}(\mathsf{A})\big) = n - \dim\big(\mathcal{R}(\mathsf{A}^\mathsf{T})\big) = n - \mathrm{rank}\big(\mathsf{A}\big),$$
Therefore, if $\mathrm{rank}(\mathsf{A}) < n$, the problem $\mathsf{A}x = b$ cannot have a unique solution.

Linear algebra: Conclusions

1. To guarantee the existence of the solution $\mathsf{A}x = b$, $\mathsf{A} \in \mathbb{R}^{m\times n}$, we need to have
$$\mathrm{rank}(\mathsf{A}) = m.$$
2. To guarantee uniqueness of the solution, we need to have
$$\mathrm{rank}(\mathsf{A}) = n.$$
3. The problem cannot be well-posed unless $m = n$ and the matrix A is full rank,
$$\det(\mathsf{A}) \neq 0,$$
that is, the matrix is square and invertible.

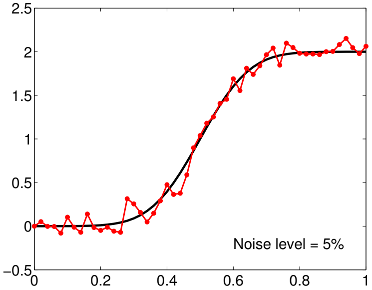
Therefore, *almost every linear problem you encounter is ill-posed!*

Example: Numerical Differentiation

Let $f : [0,1] \to \mathbb{R}$ be a differentiable function, $f(0) = 0$.

$$\text{Data } = f(t_j) + \text{ noise}, \quad t_j = \frac{j}{n}, \quad j = 1, 2, \ldots n.$$

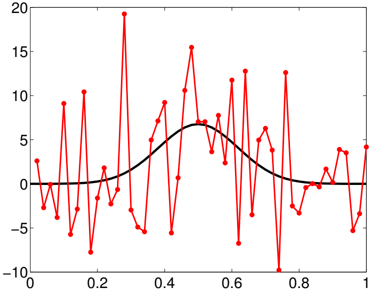
Problem: Estimate $f'(t_j)$.



Naive Solution: Finite Difference Approximation

Approximate

$$f'(t_j) \approx \frac{f(t_j) - f(t_{j-1})}{h}, \quad h = \frac{1}{n}.$$



Naive Solution: Finite Difference Approximation

Where is the problem?

$$b_j = f(t_j) + e_j,$$

$$x_j = \frac{b_j - b_{j-1}}{h} = \underbrace{n\left(f(t_j) - f(t_{j-1})\right)}_{\approx f'(t_j)} + n\left(e_j - e_{j-1}\right).$$

The noise is amplified by a factor of the order $\sim n = 50$.

Formulation as an Inverse Problem

Denote $g(t) = f'(t)$. Then,

$$f(t) = \int_0^t g(\tau) d\tau.$$

Linear model:

$$\text{Data} = b_j = f(t_j) + e_j = \int_0^{t_j} g(\tau) d\tau + e_j,$$

where e_j is the noise.

Discretization

Write

$$\int_0^{t_j} g(\tau)d\tau \approx \frac{1}{n}\sum_{k=1}^{j} g(t_k).$$

By denoting $g(t_k) = x_k$,

$$b = \mathsf{A}x + e,$$

where

$$\mathsf{A} = \frac{1}{n}\begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Properties of the Matrix

The matrix A is invertible: For a triangular matrix,

$$\det(\mathsf{A}) = a_{11}a_{22}\cdots a_{nn} = \left(\frac{1}{n}\right)^n \neq 0.$$

Therefore,

$$\operatorname{rank}(\mathsf{A}) = m = n,$$

and the problem has a unique solution. One can check that

$$\mathsf{A}^{-1} = n\begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

corresponding to the finite difference method.

Properties of the Matrix

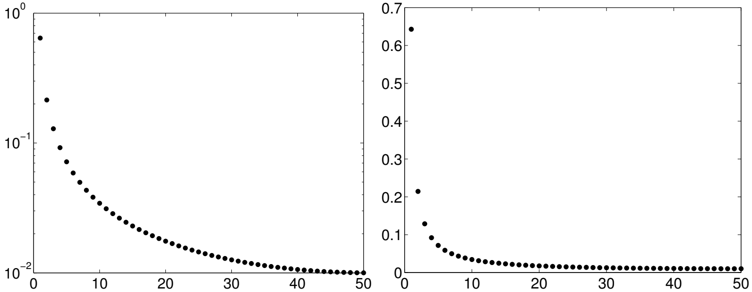
Singular values (revisited soon):

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^{\mathsf{T}},$$

where $\mathsf{U}, \mathsf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices,

$$\mathsf{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix}, \quad d_1 \geq d_2 \geq \ldots d_n > 0.$$

Singular Value Analysis



Ratio of smallest and largest singular values (= condition number of A):

$$r(\mathsf{A}) = \frac{d_1}{d_n} = 64.27$$

In inverse problems, $r(\mathsf{A})$ is often *much* larger than this (10^6–10^8)!

SVD Revisited

Fundamental result: Any rectangular real matrix $\mathsf{A} \in \mathbb{R}^{m\times n}$ can be decomposed as

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^\mathsf{T},$$

where $\mathsf{U} \in \mathbb{R}^{m\times m}$, $\mathsf{V} \in \mathbb{R}^{n\times n}$ are orthogonal matrices, that is,

$$\mathsf{U}\mathsf{U}^\mathsf{T} = \mathsf{U}^\mathsf{T}\mathsf{U} = \mathsf{I}_m, \quad \mathsf{V}\mathsf{V}^\mathsf{T} = \mathsf{V}^\mathsf{T}\mathsf{V} = \mathsf{I}_n,$$

and $\mathsf{D} \in \mathbb{R}^{m\times n}$ is a diagonal matrix. The diagonal entires of D are non-negative, and are usually ordered in a non-decreasing order,

$$d_1 \geq d_2 \geq \ldots \geq d_{\min(m,n)} \geq 0.$$

SVD Revisited

Remark 1: Denote

$$\mathsf{V} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}, \quad v_j \in \mathbb{R}^n.$$

Orthogonality of V is *equivalent* to saying that the column vectors v_j form an orthonormal basis of $\mathbb{R}^n$.
In particular, for every $x \in \mathbb{R}^n$,

$$x = \sum_{j=1}^{n} \left(v_j^\mathsf{T} x\right) v_j,$$

where

$$v_j^\mathsf{T} x \quad = \quad \text{orthogonal projection of } x \text{ onto the subspace spanned by } v_j.$$

SVD Revisited

Remark 2: Orthogonality of a matrix V means that the action $x \mapsto \mathsf{V}x$ is

- a rotation, or
- a reflection, or
- a permutation.

that is, a conformal (= angle-preserving) linear transformation.

$$(\mathsf{V}x)^\mathsf{T}\mathsf{V}y = x^\mathsf{T}\underbrace{\mathsf{V}^\mathsf{T}\mathsf{V}}_{=\mathsf{I}_n}y = x^\mathsf{T}y,$$

that is,

$$\cos\angle(\mathsf{V}x, \mathsf{V}y) = \cos\angle(x, y), \quad \|\mathsf{V}x\| = \|u\|.$$

SVD Revisited

Remark 3: A rectangular (=non-square) diagonal matrix looks like

$$\mathsf{D} = \begin{bmatrix} d_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & d_m & 0 & \cdots & 0 \end{bmatrix}, \quad m < n,$$

and

$$\mathsf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad m > n.$$

SVD in Action

Given the SVD of a matrix $A \in \mathbb{R}^{m\times n}$,

$$Ax = \sum_{j=1}^{\max(m,n)} u_j d_j \left(v_j^{\mathsf{T}} x\right) = \sum_{j=1}^{r} u_j d_j \left(v_j^{\mathsf{T}} x\right),$$

where r is the number of non-zero singular values,

$$d_1 \geq d_2 \geq \ldots \geq d_r > d_{r+1} = \ldots d_{\min(m,n)} = 0.$$

$$r = \mathrm{rank}(A).$$

SVD in Action

Null space of a matrix:

$$\mathcal{N}(\mathsf{A}) = \left\{x \in \mathbb{R}^n \mid \mathsf{A}x = 0\right\} = ?$$

$$\mathsf{A}x = \sum_{j=1}^{r} u_j d_j \left(v_j^{\mathsf{T}} x\right) = 0,$$

if and only if

$$x \perp v_1, v_2, \ldots, v_r,$$

that is,

$$x = \sum_{j=r+1}^{n} x_j v_j \in \mathrm{span}\left\{v_{r+1}, \ldots, v_n\right\}.$$

SVD in Action

Range of a matrix

$$\mathcal{R}(\mathsf{A}) = \big\{ y \in \mathbb{R}^m \mid y = \mathsf{A}x \text{ for some } x \in \mathbb{R}^n \big\} = ?$$

$$b = \mathsf{A}x = \sum_{j=1}^{r} u_j d_j \big(v_j^{\mathsf{T}} x \big)$$

for some x if and only if

$$b = \sum_{j=1}^{r} (u_j^{\mathsf{T}} b) u_j \in \mathrm{span}\big\{ u_1, u_2, \ldots, u_r \big\}.$$

SVD in Action

Solving a linear equations in terms of the SVD:

$$b = \mathsf{A}x,$$

or in terms of SVD,

$$\sum_{j=1}^{m} (u_j^\mathsf{T} b) u_j = \sum_{j=1}^{r} d_j \big(v_j^\mathsf{T} x\big) u_j.$$

SVD in Action

- Solution exists if and only if $b \in \mathcal{R}(\mathsf{A})$, that is,

$$b \perp u_{r+1}, \dots, u_m.$$

- For the solution to exist, we must have

$$d_j \big(v_j^\mathsf{T} x\big) = (u_j^\mathsf{T} b), \quad 1 \leq j \leq r.$$

- For any $x_0 \in \mathcal{N}(\mathsf{A})$, the vector

$$x = \sum_{j=1}^{r} \frac{(u_j^\mathsf{T} b)}{d_j} v_j + x_0 = \sum_{j=1}^{r} \frac{(u_j^\mathsf{T} b)}{d_j} v_j + \sum_{j=r+1}^{n} x_j v_j$$

is a solution.

Four Fundamental Subspaces

Given a matrix $\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^\mathsf{T} \in \mathbb{R}^{m\times n}$,

$$\mathrm{Rank}(\mathsf{A}) = r = \#\text{ of non-zero singular values} \leq \min(m,n),$$

$$\mathsf{U} = \left[\begin{array}{c|c} \leftarrow r \rightarrow & \leftarrow (m-r) \rightarrow \\ \mathcal{R}(\mathsf{A}) & \mathcal{N}(\mathsf{A}^\mathsf{T}) \\ & \end{array}\right] \in \mathbb{R}^{m\times m},$$

$$\mathsf{V} = \left[\begin{array}{c|c} \leftarrow r \rightarrow & \leftarrow (n-r) \rightarrow \\ \mathcal{R}(\mathsf{A}^\mathsf{T}) & \mathcal{N}(\mathsf{A}) \\ & \end{array}\right] \in \mathbb{R}^{n\times n},$$

Backwards Heat Equation and Linear Algebra

Heat equation:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}, \quad u(0,t) = u(1,t) = a, \quad u(x,0) = u_0(x).$$

Look for the solution of the form

$$u(x,t) = a + \sum_{j=1}^{\infty} u_j(t) \sin \pi j x,$$

and initial value as

$$u_0(x) = a + \sum_{j=1}^{\infty} \alpha_j \sin \pi j x,$$

Backwards Heat Equation and Linear Algebra

Substitute into the heat equation,

$$\sum_{j=1}^{\infty} u_j'(t) \sin \pi j x = -\sum_{j=1}^{\infty} D(\pi j)^2 u_j(t) \sin \pi j x,$$

which is satisfied if

$$u_j'(t) = -D(\pi j)^2 u_j(t), \quad u_j(0) = \alpha_j.$$

Solution:

$$u_j(t) = \alpha_j e^{-D(\pi j)^2 t}.$$

Backwards Heat Equation and Linear Algebra

Forward map:

$$\mathsf{A} : u(x,0) \mapsto u(x,T),$$

$$a + \sum_{j=1}^{\infty} \alpha_j \sin \pi j x \mapsto a + \sum_{j=1}^{\infty} \alpha_j e^{-D(\pi j)^2 T} \sin \pi j x.$$

Inverse map (formally)

$$\mathsf{A}^{-1} : u(x,T) \mapsto u(x,0),$$

$$a + \sum_{j=1}^{\infty} \beta_j \sin \pi j x \mapsto a + \sum_{j=1}^{\infty} \beta_j e^{D(\pi j)^2 T} \sin \pi j x.$$

In Matrix Language

Finite approximation (set $a = 0$): Discretize,

$$x_k = \frac{k}{n}, \quad 1 \leq k \leq n-1, \quad \text{(interior points)}$$

then truncate the trigonometric series,

$$u(x_k, 0) = \sqrt{\frac{2}{n}} \sum_{j=1}^{n-1} \alpha_j \sin \pi \frac{jk}{n},$$

or

$$U_0 = \left[\begin{array}{c} u(x_1, 0) \\ \vdots \\ u(x_{n-1}, 0) \end{array} \right] = \mathsf{S} \left[\begin{array}{c} \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{array} \right],$$

where $\mathsf{S} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the matrix with entries

$$\mathsf{S}_{jk} = \sqrt{\frac{2}{n}} \sin \pi \frac{jk}{n}, \quad 1 \leq j, k \leq n-1.$$

Similarly, we write

$$u(x_k, T) = \sqrt{\frac{2}{n}} \sum_{j=1}^{n-1} \beta_j \sin \pi \frac{jk}{n},$$

or, in matrix form,

$$U_T = \mathsf{S}\beta.$$

From the spectral analysis, we know that

$$\beta_j = e^{-D(\pi j)^2 T} \alpha_j = \lambda_j \alpha_j,$$

or, in matrix form,

$$\beta = \Lambda \alpha,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{n-1} \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

You can verify (using trigonometry or numerically) that

$$\mathsf{S}^{-1} = \mathsf{S}^{\mathsf{T}} = \mathsf{S}.$$

Therefore:

$$U_T = \mathsf{S}\beta = \mathsf{S}\Lambda\alpha = \mathsf{S}\Lambda\mathsf{S}^{\mathsf{T}} U_0.$$

We have found a propagation matrix

$$\mathsf{A} = \mathsf{S}\Lambda\mathsf{S}^{\mathsf{T}} : U_0 \mapsto U_T.$$

It is invertible (in theory),

$$\mathsf{A}^{-1} = S\Lambda^{-1} S^{\mathsf{T}},$$

but this formula is useless in practice.

Summary

- The ill-posedness of a linear problem depends on the dimensions of the matrix and the rank of it
- SVD gives the means to determine the rank of the matrix, the basis of the null space and the basis of the range
- Sensitivity to noise depends on the distribution of the singular values (condition number)
- SVD is a useful tool for analyzing linear problems, but may be impractical to compute