Machine learning in geoscience notes

Tariq Alkhalifah and Omar Saad

February 1, 2025

Chapter 1

The basics of Linear Algebra for Machine Learning

Linear algebra is fundamental to machine learning (ML), particularly in geosciences, where data can be high-dimensional and complex. In ML, we often solve large systems of equations, and that requires knowledge of Linear Algebra. As we will see later, many components of the ML model, as well as the data, will be represented in terms of vectors and matrices, and even tensors. This chapter covers core concepts in linear algebra, including focusing on singular value decomposition (SVD) and Principal Component Analysis (PCA) for dimensionality reduction and matrix inversion, both of which are essential for transforming and interpreting data in geoscientific applications.

The father of linear Algebra is Al-Khwarizmi, a Persian mathematician, astronomer, and scholar of the Islamic Golden Age. His writing in "The book of Algebra and balancing", written in 820 CE, provided foundational contributions to the field. In his seminal work, he introduced the principles of algebra as a distinct mathematical discipline. He developed methods for solving linear and quadratic equations, emphasizing the use of systematic procedures (algorithms) to find solutions. Al-Khwarizmi's work not only laid the groundwork for modern algebra but also introduced the concept of balancing equations, which remains central to mathematics today. His contributions were later translated into Latin and influenced European mathematicians during the Renaissance, playing a pivotal role in the development of mathematics worldwide, including machine learning.

1.1 Vectors, Matrices, and Basic Operations

Vectors and matrices are essential structures in linear algebra that represent data and transformations. In geosciences, vectors might represent spatial data points or subsurface model grid values, while matrices can represent grids of spatial measurements, but also operators essential to relate models to data.

1.1.1 Vectors

A vector is an ordered list of numbers:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{1.1}$$

1.1.2 Matrices

A matrix is a two-dimensional array of numbers organized in rows and columns. For example, an $m \times n$ matrix $\mathbf{A}$ has m rows and n columns, as follows:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{1.2}$$

where $\ldots$ imply the rest of the elements of the matrix. A square matrix ($m = n$) is a symmetric if $\mathbf{A}^T = \mathbf{A}$, where T is the transpose operation, in which

$$\mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{1.3}$$

1.2 Matrix Operations

1.2.1 Matrix Addition and Scalar Multiplication

Matrix addition and scalar multiplication are element-wise operations:
- Addition: $\mathbf{A} + \mathbf{B}$ adds matrices of the same dimensions element-by-element.
- Scalar Multiplication: $\alpha\mathbf{A}$, where each element in $\mathbf{A}$ is multiplied by scalar α.

1.2.2 Matrix Multiplication

Matrix multiplication, on the other hand, combines two matrices $\mathbf{A}$ (of dimensions $m \times n$) and $\mathbf{B}$ (of dimensions $n \times p$) to produce a matrix $\mathbf{C}$ of dimensions

$m \times p$:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \tag{1.4}$$

where a_{ij}, b_{ij}, and c_{ij} are the elements of matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively.

1.3 Matrix Inversion

Matrix inversion is critical for solving systems of linear equations and in applications such as linear regression and inversion.

1.3.1 Definition of Matrix Inversion

For a square matrix $\mathbf{A}$, the inverse $\mathbf{A}^{-1}$ exists if:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{1.5}$$

where $\mathbf{I}$ is the identity matrix, given by

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{1.6}$$

1.3.2 Calculating the Inverse, and the determinant

Calculating the inverse for large matrices is both complicated and costly. To provide an idea of how the inverse works and to appreciate the complexity, we consider the smallest matrix. For a 2×2 matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the inverse is given by:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \tag{1.7}$$

where the determinant $\det(\mathbf{A})$ is:

$$\det(\mathbf{A}) = ad - bc \tag{1.8}$$

For larger matrices, more complex methods such as Gaussian elimination or decompositions (e.g., LU decomposition) are used.

1.3.3 Applications of Matrix Inversion in Geosciences

Matrix inversion is essential in solving linear systems for tasks like: - **Linear Regression:** Solving for the coefficients that best fit a linear model to data.
- **Data Reconstruction:** Inverting transformation matrices to return transformed data to their original space.

- Linear traveltime tomography: Inverting measured wave traveltimes for the velocity model through which the wave propagated.

In many geoscientific applications, the inversion of large matrices is required, such as in spatial data interpolation using kriging or in estimating covariance structures in geospatial modeling.

1.4 Eigenvalues, Eigenvectors, and Principal Component Analysis (PCA)

PCA is a powerful technique for analyzing data and reducing dimensionality, which transforms high-dimensional data into a lower-dimensional form while preserving as much variance as possible. In geosciences, PCA can be used to reduce the complexity of spatial datasets, making machine learning models more efficient and interpretable.

1.4.1 Eigenvalues and Eigenvectors

For a square matrix $\mathbf{A}$, an eigenvector $\mathbf{v}$ and corresponding eigenvalue λ satisfy:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}. \tag{1.9}$$

Eigenvalues represent the scaling factor along the direction of their corresponding eigenvectors, which remain unchanged in direction under the transformation represented by $\mathbf{A}$, which is a special direction characterizing the transformation.

1.4.2 Singular value decomposition (SVD) and Pseudoinverses

SVD utilizes the concept of eigenvalues and eigenvectors to perform a form matrix decomposition that allows us to compute matrix inverses even for non-invertible matrices. Specifically, it has the form:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1.10}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{\Sigma}$ is a diagonal matrix of singular values. If $\mathbf{A}$ is a square symmetric matrix, then $\mathbf{V} = \mathbf{U}$, with columns given by the eigenvectors of the matrix, with the corresponding eigenvalues in $\mathbf{\Sigma}$. In this case, the inverse of $\mathbf{A}$ is simply given by

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^T, \tag{1.11}$$

where the inverse of a diagonal matrix is simply given by the inverse of its elements along the diagonal.

Another important character of a matrix is its rank. The rank of $\mathbf{A}$, rank($\mathbf{A}$), is given by the number of nonzero eigenvalues for a matrix. Note for a matrix with rank($\mathbf{A}$)$< n$, we will have zero eigenvalues, and computing the inverse using

SVD will be problematic. The pseudoinverse is an approximation of the inverse in which we ignore zero and sometimes small eigenvalues by setting their inverse to zero. Ignoring eigenvalues, especially small ones, is a form of dimensionality reduction, as each nonzero eignvalue corresponds to an orthogonal eigenvector, representing a dimension. However, for small eigenvalues that representation is small and could possibly be dropped.

1.4.3 Principal Component Analysis (PCA)

PCA identifies the principal components of a dataset by finding eigenvectors of the covariance matrix. These components are the directions in which the data varies most, allowing us to project high-dimensional data into a lower-dimensional space, corresponding to the highest eigenvalues.

Steps in PCA:

1. Data Centering: Subtract the mean of each feature from the dataset to create a zero-centered dataset.
2. Covariance Matrix Calculation: Calculate the covariance matrix of the centered data:
$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}, \tag{1.12}$$
where $\mathbf{X}$ is the data matrix with rows representing data points and columns representing features.
3. Eigenvalue and Eigenvector Computation: Calculate the eigenvalues and eigenvectors of the covariance matrix $\mathbf{C}$.
4. Selection of Principal Components: Select the top k eigenvectors corresponding to the largest k eigenvalues.
5. Data Projection: Project the data onto the selected eigenvectors (principal components) to reduce dimensionality:
$$\mathbf{X}_{\text{reduced}} = \mathbf{X}\mathbf{W}, \tag{1.13}$$
where $\mathbf{W}$ is the matrix of selected eigenvectors.

1.4.4 Applying PCA to Geoscience Data

PCA is highly applicable to geosciences, where large, multi-dimensional datasets (e.g., seismic and satellite imagery or climate model outputs) are common. PCA helps reduce dimensionality, enabling more efficient data storage, faster processing, and often improved model accuracy by removing noise and redundant information. It also provides a lot of insight into the data and their distribution.

Example application: PCA can be used in the analysis of seismic images. By applying PCA to seismic images, we can identify the distribution of images along the principal patterns across, like the horizontal layering, providing insights into the distribution of the main features within the seismic image.

1.4.5 Interpreting Principal Components

In geoscientific datasets, principal components often correspond to meaningful physical patterns. For instance, in seismic images, the first principal component might capture the reflectivity, while subsequent components could capture faults, and other features less prominent.

1.5 Summary

This chapter covered essential linear algebra concepts, with an emphasis on matrix inversion and PCA, both of which are crucial for machine learning in geosciences. Matrix inversion is essential for solving linear systems and understanding transformations, while PCA is a widely used technique for reducing dimensionality and extracting important features from complex geospatial data. We will visit PCA more than once in this course.

In the following chapters, we will dive deeper into machine learning techniques that build upon these linear algebra foundations, applying them to solve real-world problems in geosciences.

References

1. Strang, G. (2006). *Linear Algebra and Its Applications*. Brooks Cole.
2. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Chapter 2

Probability, Bayes' Theorem, and Statistical Inference in Machine Learning

At the heart of machine learning (ML) are the statistics involved in training the neural network model, and those involve the statistics of the training dataset as well as the resulting predictions. As a result, a trained ML model that makes predictions is referred to as a hypothesis. This term is derived from the statistical nature of the ML process. In fact, Bayesian inference has become central to many machine learning approaches, especially in fields like geosciences, where data often come with uncertainties and noise. This chapter covers probability fundamentals with a focus on Bayes' Theorem, laying the groundwork for Bayesian methods that are critical in handling and interpreting geospatial data.

Abu Yosuf al-Kindī's "Manuscript on Deciphering Cryptographic Messages" (9th century) laid an early foundation for probability and statistics by introducing **frequency analysis**, a fundamental statistical method. In this work, al-Kindī systematically analyzed the frequency of letters in Arabic texts to break encrypted messages, demonstrating that different letters appear with varying probabilities. This concept, which relies on counting occurrences and estimating likelihoods, became a precursor to modern statistical inference.

More recently, probability and statistics have often been attributed to the work of Blaise Pascal (1623–1662) and Pierre de Fermat (1607–1665), who laid the foundation of probability theory through their correspondence on gambling problems in the 17th century. Their discussions led to the "concept of expected value", which we will cover here, as well as the "binomial probability distribution", which helped solve problems involving fair bets and risk assessment. Around the same time, Christiaan Huygens (1629–1695) formalized these ideas

in his book De Ratiociniis in Ludo Aleae (1657), making probability a structured mathematical discipline. In statistics, Carl Friedrich Gauss (1777–1855) contributed significantly with the "normal distribution" and the "method of least squares", which are fundamental to modern statistical analysis.

Here we will review in summary some of the main components of this topic, but for more details, look into the references listed below.

2.1 Basics of Probability

Among the many roles probability plays, its most important component is that it allows us to quantify and reason about uncertainty. In geosciences, predicting events like the location of natural resources, earthquake sources, or floods, requires probabilistic models to capture and interpret uncertainty in observations.

2.1.1 Definitions and Concepts

- **Experiment:** A repeatable process that produces a well-defined outcome. In geosciences, an experiment could be the measurement of a specific weather parameter or recording seismic activity.
- **Sample Space (Ω):** The set of all possible outcomes of an experiment, e.g., all possible earthquake magnitudes. It could be finite or infinite, and it could be countable or uncountable. An example of an infinite uncountable set is the set of natural numbers.
- **Event** (E): A subset of the sample space that represents an outcome or set of outcomes of interest, like a single measurement.

The probability of an event E is defined as:

$$P(E) = \frac{\text{Number of E outcomes}}{\text{Total number of outcomes}}, \tag{2.1}$$

For example, in a toss of an unbiased coin, the sample space equals 2 (heads or tails), and the probability of sampling heads (H) with one toss is $P(H) = \frac{1}{2} = 0.5$.

2.2 Random Variables and Probability Distributions

A *random variable* is a function that assigns a numerical value to each outcome of an experiment. Random variables are classified as "discrete" or "continuous".

Discrete Random Variable

A *discrete random variable* X takes on a finite or countably infinite set of values. For example, the number of earthquake occurrences in a given year is a discrete random variable.

- **Probability Mass Function (PMF):** $p(x) = P(X = x)$, the probability that X takes on a particular value x. The values of p are always positive and

between zero and one, and the sum of p over all possible outcomes $x \in X$ equals 1.

Continuous Random Variable

A *continuous random variable* Y can take on any value within a certain range. For example, seismic data measurements are often modeled as continuous data.

- **Probability Density Function (PDF):** $f(y)$, which represents the density of Y around y. It satisfies the following relation:

$$P(a \leq Y \leq b) = \int_a^b f(y)\, dy. \tag{2.2}$$

f can take on values larger than 1. However, P is always positive and f is positive, and the maximum of P is 1 when a and b bound the entire space.

2.3 Joint, Marginal, and Conditional Distributions

In probability theory, the concepts of **joint distribution**, **marginal distribution**, and **conditional distribution** are fundamental to understanding the relationships between multiple random variables, such as model and data, or predictions and neural network parameters.

2.3.1 Joint Distribution

The **joint distribution** describes the probability of two or more random variables taking specific values simultaneously. For two discrete random variables X and Y, the joint probability mass function (PMF) is represented as $P(X = x, Y = y)$, where x and y are specific values of X and Y, respectively.

For continuous random variables, the joint probability density function (PDF) is represented by $f_{X,Y}(x, y)$, where it represents the density of the joint distribution at the point (x, y).

2.3.2 Marginal Distribution

The **marginal distribution** of a random variable is obtained by summing (or integrating) the joint distribution over the other variable(s). Focusing on the continuous random variables, the marginal PDF of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy, \tag{2.3}$$

and the marginal PDF of Y is:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx. \tag{2.4}$$

2.3.3 Conditional Distribution

The **conditional distribution** describes the probability distribution of one random variable given the value of another. Again focusing on continuous random variables, the conditional PDF of X given $Y = y$ is:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \tag{2.5}$$

provided $f_Y(y) > 0$. In fact, since the same holds the other way around (replacing X and Y), a more general form is given by

$$f_{X,Y}(x, y) = f_{X|Y}(x \mid y) f_Y(y) = f_{Y|X}(y \mid x) f_X(x). \tag{2.6}$$

We can actually derive the Bayes relation from this equation by writing it in probability distribution form:

$$P(x, y) = P(x \mid y)P(x) = P(y \mid x)P(x). \tag{2.7}$$

2.3.4 A discrete example

Consider two random variables X and Y with the following joint PMF, representing for example two biased coins with heads labeled as 1 and tails as 0:

$$P(X = x, Y = y) = \begin{cases} 0.2 & \text{if } (x, y) = (0, 0), \\ 0.3 & \text{if } (x, y) = (0, 1), \\ 0.1 & \text{if } (x, y) = (1, 0), \\ 0.4 & \text{if } (x, y) = (1, 1). \end{cases} \tag{2.8}$$

The marginal PMF of X is:

$$P(X = 0) = 0.2 + 0.3 = 0.5, \quad P(X = 1) = 0.1 + 0.4 = 0.5. \tag{2.9}$$

The marginal PMF of Y is:

$$P(Y = 0) = 0.2 + 0.1 = 0.3, \quad P(Y = 1) = 0.3 + 0.4 = 0.7. \tag{2.10}$$

The conditional PMF of X given $Y = 1$ is:

$$P(X = 0 \mid Y = 1) = \frac{P(X = 0, Y = 1)}{P(Y = 1)} = \frac{0.3}{0.7} \approx 0.428, \tag{2.11}$$

$$P(X = 1 \mid Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{0.4}{0.7} \approx 0.571. \tag{2.12}$$

The total should round up to 1.

2.4 Bayes' Theorem and inference

Bayes' Theorem is a foundational concept in probability that allows us to update the probability of a hypothesis based on new evidence (like data). We start with a probability for the results we expect, like the subsurface model, or events occurring, like earthquakes, and that probability is updated thanks to new measurements. Thus, it is widely used in geosciences to revise beliefs about uncertain events when additional data is observed.

Statement of Bayes' Theorem

For two events x and y, where $P(x) > 0$:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}, \tag{2.13}$$

where:
- $P(x|y)$ is the *posterior probability*: the probability of event x (hypothesis) given that y (evidence) has occurred.
- $P(y|x)$ is the *likelihood*: the probability of observing y given that x is true.
- $P(x)$ is the *prior probability*: the initial probability of x before observing y.
- $P(y)$ is the *marginal probability* of y: the total probability of y occurring.

This theorem allows us to calculate the probability of a hypothesis being true after new data is considered. In machine learning and geosciences, Bayes' Theorem is used to build models that update predictions based on incoming data.

Thus, Bayesian inference is a statistical inference method that combines prior information with evidence to form a posterior probability. This is especially useful in geosciences, where models can incorporate prior knowledge of geological conditions.

Following the above definition, we repeat the main components as follows:

1. Prior ($P(x)$): Initial belief before observing any evidence.
2. Likelihood ($P(y|x)$): Probability of the observed data given the hypothesis.
3. Posterior ($P(x|y)$): Updated belief after observing the evidence.

Thus,

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}. \tag{2.14}$$

2.5 Bayesian Modeling in Geosciences

In geoscientific machine learning, Bayesian models are commonly applied to handle the uncertainties in measurements and predictions. Some key areas include:
- **Seismic Hazard Analysis:** By combining historical data and recent seismic observations, Bayesian models update predictions on the likelihood of future

earthquakes in a region.
- **Geospatial Prediction:** Bayesian methods are used to update models on climate or mineral distributions as new observational data becomes available.

Example: Estimating the Probability of a Oil reservoir

Let:
- H: Presence of an oil reservoir in the field.
- D: The observed seismic data in which the image contains a bright spot indicating the presence of a reservoir.

To estimate the posterior probability that an oild reservoir exists given bright spots in the image, we can use Bayes' Theorem:

$$P(\text{Reservoir}|\text{Data}) = \frac{P(\text{Data}|\text{Reservoir})\ \ P(\text{Reservoir})}{P(\text{Data})}, \tag{2.15}$$

where:
- P(Data|Reservoir): Likelihood of observing the data if a reservoir exists. - P(Reservoir): Prior probability of a reservoir based on prior surveys. - P(Data): Marginal probability of the observed data.

2.5.1 2.6: Advantages of Bayesian Methods in Machine Learning

Bayesian approaches in machine learning provide several advantages for geosciences:
- **Incorporation of Prior Knowledge:** Bayesian models allow the integration of historical or expert knowledge.
- **Uncertainty Quantification:** Bayesian models provide probabilistic outputs that quantify the level of uncertainty.
- **Adaptive or active learning:** Bayesian inference updates models dynamically as new data arrive, which is essential in fields with evolving data.

2.6 Sampling the Posterior Distribution

In Bayesian inference, and considering ML problems where θ represents the neural network (NN) model parameters (weights and biases), and D are the training data, possibly including inputs and labels, we often wish to sample from the posterior distribution $p(\theta|D)$ given by Bayes' theorem:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{2.16}$$

However, computing the normalization constant $p(D) = \int p(D|\theta)p(\theta)d\theta$ is often intractable. Thus, we resort to numerical methods to sample from the posterior distribution. In the following, we discuss some of the commonly used methods.

2.6.1 Markov Chain Monte Carlo (MCMC)

MCMC methods construct a Markov chain whose stationary distribution is the target posterior $p(\theta|D)$. The two most common MCMC algorithms are:

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm generates a sequence of samples using a proposal distribution $q(\theta'|\theta)$. Given the current state θ_t at step t, a new sample θ' is proposed, and it is accepted with probability:

$$A(\theta'|\theta_t) = \min\left(1, \frac{p(D|\theta')p(\theta')q(\theta_t|\theta')}{p(D|\theta_t)p(\theta_t)q(\theta'|\theta_t)}\right). \tag{2.17}$$

If accepted, we set $\theta_{t+1} = \theta'$, otherwise we retain $\theta_{t+1} = \theta_t$.

Gibbs Sampling

Gibbs sampling is a special case of MCMC where we sample each parameter sequentially from its conditional posterior:

$$\theta_i^{(t+1)} \sim p(\theta_i|\theta_{-i}^{(t)}, D). \tag{2.18}$$

This method is particularly useful when the full conditional distributions are easy to sample from.

2.6.2 Hamiltonian Monte Carlo (HMC)

HMC improves on MCMC by using gradient information to explore the posterior more efficiently. It introduces auxiliary momentum variables r and defines a Hamiltonian function:

$$H(\theta, r) = U(\theta) + K(r), \tag{2.19}$$

where $U(\theta) = -\log p(\theta|D)$ is the potential energy and $K(r) = \frac{1}{2}r^T M^{-1} r$ is the kinetic energy. The system evolves according to Hamiltonian dynamics, and proposals are generated using numerical integration methods such as leapfrog integration.

2.6.3 Variational Inference (VI)

Instead of sampling, variational inference approximates the posterior with a simpler distribution $q(\theta; \lambda)$, where λ are variational parameters. The goal is to minimize the Kullback-Leibler (KL) divergence:

$$\lambda^* = \arg\min_{\lambda} \mathrm{KL}(q(\theta; \lambda)||p(\theta|D)). \tag{2.20}$$

This optimization is often performed using stochastic gradient descent.

Each of these methods has its own advantages and is used depending on the complexity of the posterior and computational constraints. For more information, please visit the references.

References

1. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

2. Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.

3. Smith, R. C. (2013). *Uncertainty Quantification: Theory, Implementation, and Applications*. SIAM.