

Discussion 2

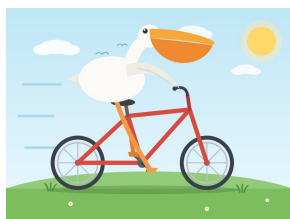


fig: a pelican on a bicycle
drawn by Qwen 3.8 27B

This week's resources:

- Matrix differentiation guide (written by Randal Barnes) link
- Mathematics for Machine Learning link
- 3Blue1Brown: Essence of Linear Algebra playlist
- Simon Willison's pelican on a bicycle benchmark for LLMs link

1 Proof of K-means Convergence

Consider the K-means algorithm applied to data points $\{x_n\}_{n=1}^N \subset \mathbb{R}^D$ with K clusters. Define binary indicator variables $r_{nk} \in \{0, 1\}$ with $\sum_{k=1}^K r_{nk} = 1$ for each n , and cluster centers $\{\mu_k\}_{k=1}^K$. The K-means objective is

$$J(\{r_{nk}\}, \{\mu_k\}) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2.$$

The K-means algorithm alternates between

1. **Update assignments:** Set $r_{nk} = 1$ for the k that minimizes $\|x_n - \mu_k\|^2$ and $r_{nj} = 0$ for all $j \neq k$.
2. **Update means:** Set the cluster centers to

$$\mu_k = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}} \quad (\text{for clusters with } \sum_n r_{nk} > 0).$$

Prove that the K-means algorithm converges after a finite number of iterations. (Hint: It is sufficient to show that the objective is monotonically decreasing, and that there are a finite number of iterations).

2 PC-yay!

- (a) A warehouse robot records the following four two-dimensional displacement vectors:

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} -1 \\ -2 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

Each coordinate measures displacement in meters. We will use PCA to find the direction in which the robot's motion varies the most.

- (i) Compute the sample mean

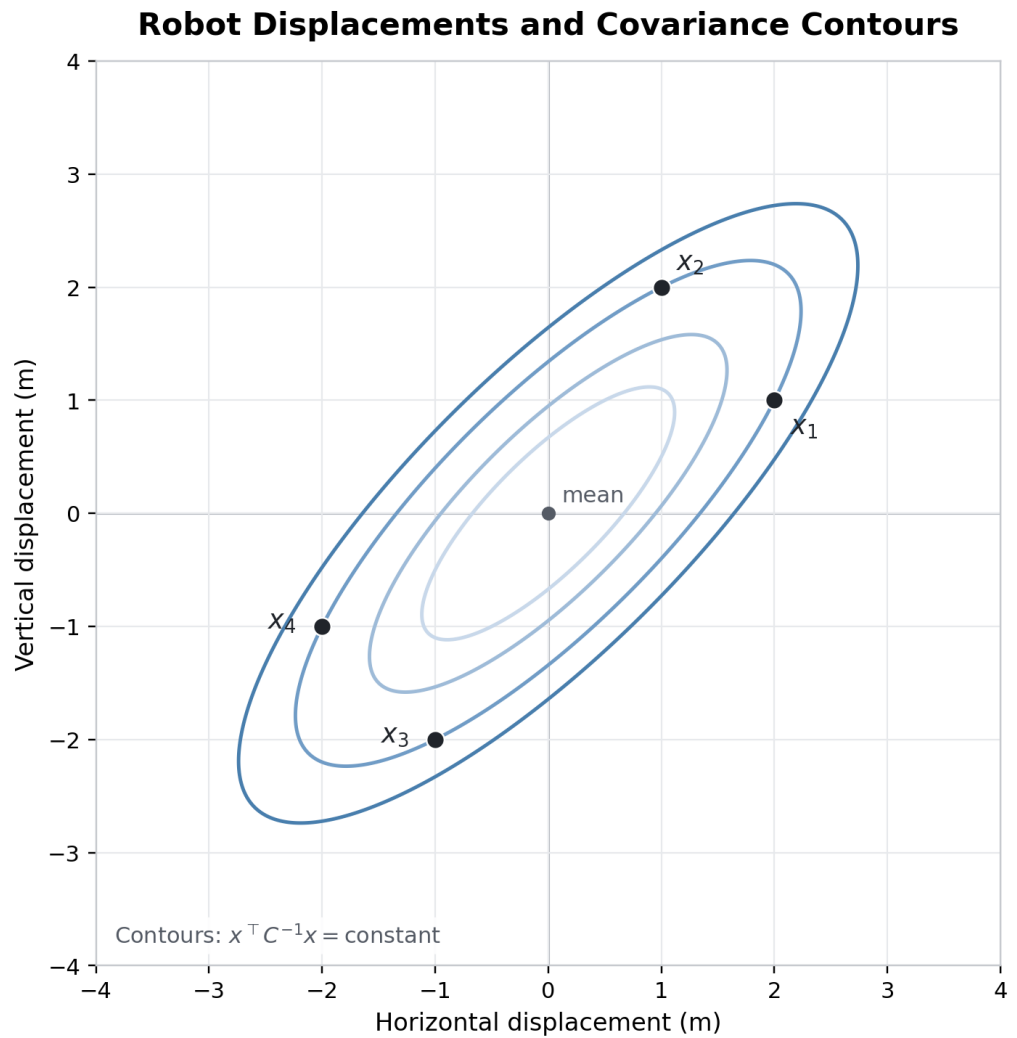
$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i.$$

- (ii) Using the normalization $1/N$, compute the empirical covariance matrix

$$C = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top.$$

- (iii) Find the eigenvalues of C and a corresponding unit eigenvector for each eigenvalue.

- (iv) The plot below shows the four robot displacements and several contours of equal covariance-scaled distance from the mean. Starting at the mean, draw and label the two eigenvector directions $\mathbf{v}_1$ and $\mathbf{v}_2$. Make the length of each arrow proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, respectively, using the same scale factor for both arrows. Which eigenvector is the first principal component? Describe its physical meaning for the robot.



- (v) Project each centered displacement onto the first principal component by computing $z_i = \mathbf{v}_1^T (\mathbf{x}_i - \boldsymbol{\mu})$.
- (vi) Suppose the robot stores only the one-dimensional values z_i . What information about its motion is lost? Give a specific example from the four recorded displacements.

(vii) The robot later records a new displacement

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Project this displacement onto the first principal component, then reconstruct its two-dimensional approximation using

$$\hat{\mathbf{x}}_{\text{new}} = \boldsymbol{\mu} + \left(\mathbf{v}_1^\top (\mathbf{x}_{\text{new}} - \boldsymbol{\mu}) \right) \mathbf{v}_1.$$

In a safety-critical robot, why might replacing $\mathbf{x}_{\text{new}}$ with this approximation be problematic?

3 Calculus

(a) Consider the function

$$f(x, y) = \sigma(ax + by),$$

where $a, b \in \mathbb{R}$ and $\sigma(t) = \frac{1}{1 + e^{-t}}$ for $t \in \mathbb{R}$.

(i) Show that $\frac{d\sigma}{dt} = \sigma(t) (1 - \sigma(t))$.

(ii) Using the result you showed in part (i) and the chain rule, compute $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$.

(b) For $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, define

$$g(\mathbf{x}) = \sum_{j=1}^n x_j^2.$$

Compute the partial derivative $\frac{\partial g}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$.

(c) Let $\mathbf{w} \in \mathbb{R}^n$ be a constant vector and define the scalar function

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{j=1}^n w_j x_j.$$

Compute $\frac{\partial h}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$.

(d) Let $\mathbf{w} \in \mathbb{R}^n$ be a constant vector and define the scalar function

$$\ell(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^n x_j^2 - \mathbf{w}^\top \mathbf{x} = \frac{1}{2} g(\mathbf{x}) - h(\mathbf{x}).$$

In single-variable calculus, we can minimize/maximize a function $f(x)$ using its derivatives. Here ℓ takes a whole vector $\mathbf{x}$ as input instead of a single number, but we can reuse that same idea one coordinate at a time.

(i) Using your answers to parts (b) and (c), compute $\frac{\partial \ell}{\partial x_i}$ for a generic coordinate $i \in \{1, \dots, n\}$. Stacking these n partial derivatives into a vector gives the *gradient*

$$\nabla \ell(\mathbf{x}) = \begin{bmatrix} \frac{\partial \ell}{\partial x_1} \\ \dots \\ \frac{\partial \ell}{\partial x_n} \end{bmatrix}.$$

- (ii) Recall from single-variable calculus that we find values of x that minimize/maximize $f(x)$ by solving $\frac{df}{dx} = 0$; the solutions are called *critical points*. In vector-calculus, a critical point of ℓ is a vector $\mathbf{x}$ where the partials $\frac{\partial \ell}{\partial x_i} = 0$ for all i components of $\mathbf{x}$ (equivalently $\nabla \ell(\mathbf{x}) = \mathbf{0}$).

Set all n partial derivatives equal to zero and solve for the minimizer $\mathbf{x}^*$. (*Hint: the i -th component of $\mathbf{x}^*$ is the value of x_i that makes $\frac{\partial \ell}{\partial x_i} = 0$; solve for one coordinate and then stack each component into the vector $\mathbf{x}^*$*)

- (iii) In single-variable calculus, once we have a critical point we check the sign of the second derivative: $f''(x) > 0$ means the function f is concave up, so the point is a minimum. Use this same second-derivative idea to argue that $\mathbf{x}^*$ is a minimum, and not a maximum or a saddle point.