Foundations of Machine Learning
African Masters in Machine Intelligence

Summary Statistics

Marc Deisenroth

Quantum Leap Africa
African Institute for Mathematical Sciences, Rwanda

Department of Computing
Imperial College London

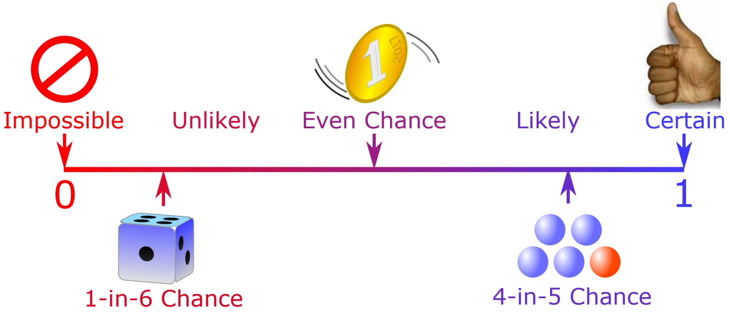
@mpd37
mdeisenroth@aimsammi.org

September 27, 2018

Some Learning Material

- Book: `https://mml-book.com` (Chapter 6)
- MOOC: `https://www.coursera.org/learn/pca-machine-learning` (Week 1)

Probabilities



- Describe a frequency ratio of events
- Express uncertainty when making predictions
- Capture a degree of belief about a hypothesis

▶▶ Probabilities are sufficient for reasoning under uncertainty

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

- Here: We are imprecise and call $p(\cdot)$ a probability distribution

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

- Here: We are imprecise and call $p(\cdot)$ a probability distribution
- If $\boldsymbol{x}$ is continuous $p(\boldsymbol{x})$ is not the probability of event $\boldsymbol{x}$ happening

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x})d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

- Here: We are imprecise and call $p(\cdot)$ a probability distribution
- If $\boldsymbol{x}$ is continuous $p(\boldsymbol{x})$ is not the probability of event $\boldsymbol{x}$ happening
- Cumulative distribution function

$$p(x \leqslant t) = \int_{-\infty}^{t} p(x)dx \in [0, 1]$$

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x})d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

- Here: We are imprecise and call $p(\cdot)$ a probability distribution
- If $\boldsymbol{x}$ is continuous $p(\boldsymbol{x})$ is not the probability of event $\boldsymbol{x}$ happening
- Cumulative distribution function

$$p(x \leqslant t) = \int_{-\infty}^{t} p(x)dx \in [0, 1]$$

- Joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ of two random variables $\boldsymbol{x}, \boldsymbol{y}$

Probability Distributions

- Probability density function (continuous $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \int p(\boldsymbol{x}) d\boldsymbol{x} = 1$$

- Probability mass function (discrete $\boldsymbol{x}$)

$$p(\boldsymbol{x}) \geqslant 0\,, \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

- Here: We are imprecise and call $p(\cdot)$ a probability distribution
- If $\boldsymbol{x}$ is continuous $p(\boldsymbol{x})$ is not the probability of event $\boldsymbol{x}$ happening
- Cumulative distribution function

$$p(x \leqslant t) = \int_{-\infty}^{t} p(x) dx \in [0, 1]$$

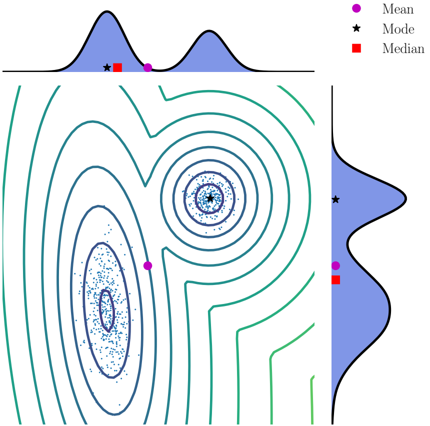
- Joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ of two random variables $\boldsymbol{x}, \boldsymbol{y}$
- Conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$ of $\boldsymbol{x}$ given $\boldsymbol{y}$

Summary Statistics

$$\mathcal{D} = \{-1, 1, 2\}$$

- Summarize datasets or random variables by describing some of their properties
- Examples:
  - Mean/Expected value (average): 2/3
  - Variance (related to spread of the data around the mean): 1.56
  - Median (data point "in the middle", i.e., value so that another data point is equally likely to be greater or smaller) $\neq$ mean: 1

Mean, Mode, Median (Continuous Distributions)



Mean (Expected Value)

- "Average"
- Does not have to be part of the dataset or a plausible realization of a random variable (▸▸ dice)

Mean (Expected Value)

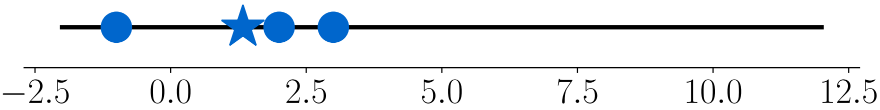
- "Average"
- Does not have to be part of the dataset or a plausible realization of a random variable (⏩ dice)

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}_{\boldsymbol{x}} \in \mathbb{R}^D \qquad \text{if } \boldsymbol{x} \in \mathbb{R}^D \text{ is continuous}$$

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n =: \boldsymbol{\mu}_{\boldsymbol{x}} \in \mathbb{R}^D \qquad \text{if } \boldsymbol{x} \in \mathbb{R}^D \text{ is discrete}$$

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{x_1}[x_1] \\ \vdots \\ \mathbb{E}_{x_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$$

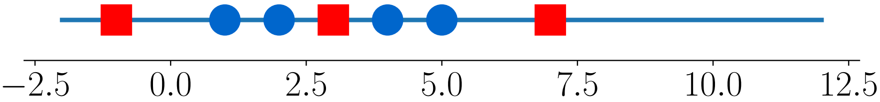
Empirical Mean



- Random variable $\boldsymbol{x} \in \mathbb{R}^D$
- N concrete realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, $\boldsymbol{x}_n \in \mathbb{R}^D$
- Empirical mean (estimate of the true mean:

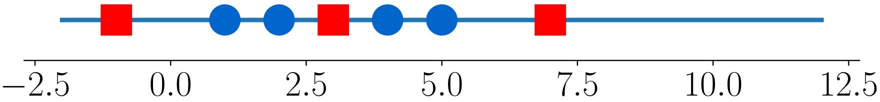
$$\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \in \mathbb{R}^D$$

Variance



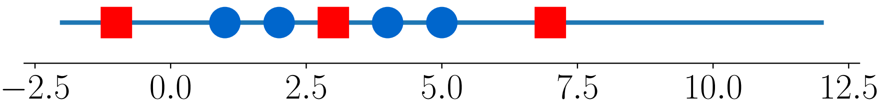
- Both datasets have the same (empirical) mean

Variance



- Both datasets have the same (empirical) mean
- Need a different quantity to describe "spread" of the data around the mean ▶▶ Variance
- Variance: Expected (squared) distance of data from the mean

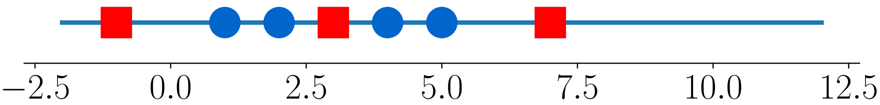
Variance



- Both datasets have the same (empirical) mean
- Need a different quantity to describe "spread" of the data around the mean ⏩ Variance
- Variance: Expected (squared) distance of data from the mean

$$\mathbb{V}[\boldsymbol{x}] := \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top]$$

Variance



- Both datasets have the same (empirical) mean
- Need a different quantity to describe "spread" of the data around the mean ▶▶ Variance
- Variance: Expected (squared) distance of data from the mean

$$\mathbb{V}[\boldsymbol{x}] := \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top]$$

$$\mathbb{V}[\boldsymbol{x}] := \int (\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top p(\boldsymbol{x}) d\boldsymbol{x} \in \mathbb{R}^{D\times D} \qquad \text{if } \boldsymbol{x} \in \mathbb{R}^D \text{ is continuous}$$

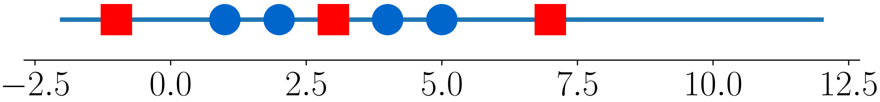
$$\mathbb{V}[\boldsymbol{x}] := \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_x)(\boldsymbol{x}_n - \boldsymbol{\mu}_x)^\top \in \mathbb{R}^{D\times D} \qquad \text{if } \boldsymbol{x} \in \mathbb{R}^D \text{ is discrete}$$

Empirical Variance

- Random variable $\boldsymbol{x} \in \mathbb{R}^D$
- N concrete realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, $\boldsymbol{x}_n \in \mathbb{R}^D$
- Empirical variance (estimate of the true variance):

$$\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \in \mathbb{R}^{D\times D}$$

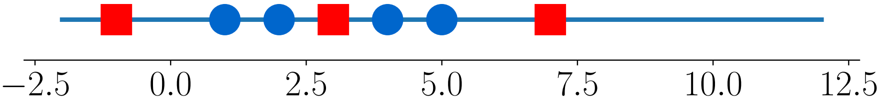
Empirical Variance



$$\mathcal{D}_1 = \{1, 2, 4, 5\}, \qquad \mathcal{D}_2 = \{-1, 3, 7\}$$

Compute the empirical variances for both datasets

Empirical Variance

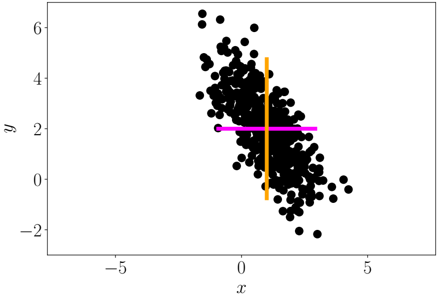


$$\mathcal{D}_1 = \{1, 2, 4, 5\}, \qquad \mathcal{D}_2 = \{-1, 3, 7\}$$

Compute the empirical variances for both datasets

- $\mathbb{V}[\mathcal{D}_1] = 2.5$
- $\mathbb{V}[\mathcal{D}_2] = 10.66$ ▶▶ $\mathcal{D}_2$ is more spread (around the mean) than $\mathcal{D}_1$
- Standard deviation $\sqrt{\mathbb{V}[\cdot]}$ describes the spread more naturally and possesses the same units as the mean

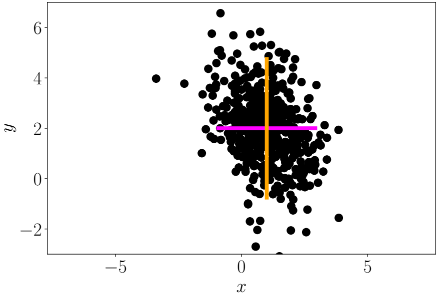
Covariance and Cross-Covariance



- Variances along each axis remain constant, but properties of the dataset change
- Variances insufficient to characterize the relationship/correlation of two random variables

  ▶▶ Cross-covariance

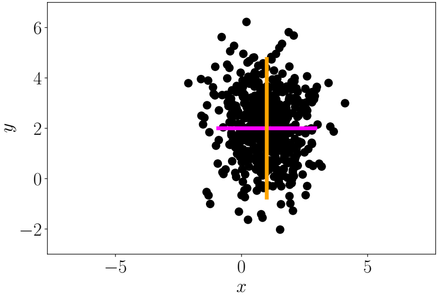
Covariance and Cross-Covariance



- Variances along each axis remain constant, but properties of the dataset change
- Variances insufficient to characterize the relationship/correlation of two random variables

▸▸ Cross-covariance

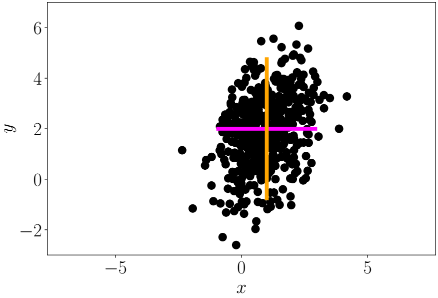
Covariance and Cross-Covariance



- Variances along each axis remain constant, but properties of the dataset change
- Variances insufficient to characterize the relationship/correlation of two random variables

  ▸▸ Cross-covariance

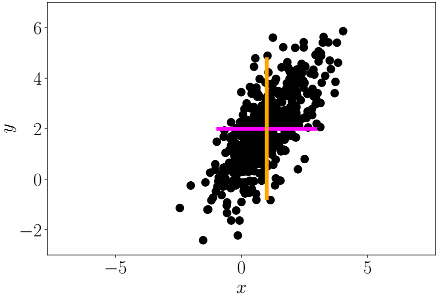
Covariance and Cross-Covariance



- Variances along each axis remain constant, but properties of the dataset change
- Variances insufficient to characterize the relationship/correlation of two random variables

  ▸▸ Cross-covariance

Covariance and Cross-Covariance



- Variances along each axis remain constant, but properties of the dataset change
- Variances insufficient to characterize the relationship/correlation of two random variables

  ▸▸ Cross-covariance

Cross-Covariance (2)

$\boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{y} \in \mathbb{R}^E$. Then

- Cross-covariance:

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top]$$

Cross-Covariance (2)

$\boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{y} \in \mathbb{R}^E$. Then

- Cross-covariance:

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top]$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \iint (\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top p(\boldsymbol{x}) p(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \in \mathbb{R}^{D \times E}$$

if $\boldsymbol{x}, \boldsymbol{y}$ are continuous

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_x)(\boldsymbol{y}_n - \boldsymbol{\mu}_y)^\top \in \mathbb{R}^{D \times E}$$

if $\boldsymbol{x}, \boldsymbol{y}$ are discrete

Cross-Covariance (2)

$\boldsymbol{x} \in \mathbb{R}^D, \boldsymbol{y} \in \mathbb{R}^E$. Then

- Cross-covariance:

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top]$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \iint (\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{y} - \boldsymbol{\mu}_y)^\top p(\boldsymbol{x})p(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \in \mathbb{R}^{D\times E}$$

if $\boldsymbol{x}, \boldsymbol{y}$ are continuous

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] := \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_x)(\boldsymbol{y}_n - \boldsymbol{\mu}_y)^\top \in \mathbb{R}^{D\times E}$$

if $\boldsymbol{x}, \boldsymbol{y}$ are discrete

- $\mathbb{V}[\boldsymbol{x}] = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}] \in \mathbb{R}^{D\times D}$
- $\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]^\top \in \mathbb{R}^{D\times E}$

Covariance Matrix

- Random variable

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix} \in \mathbb{R}^D$$

- Variance of this D-dimensional random variable is given by a covariance matrix

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \begin{bmatrix} \mathbb{V}[x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathbb{V}[x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \cdots & & \mathbb{V}[x_D] \end{bmatrix} \in \mathbb{R}^{D \times D}$$

- Covariance matrix is symmetric, positive (semi-)definite

Mean and (Co)Variance

Mean and (co)variance are often useful to describe properties of data distributions (expected values and spread).

Summary

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu} \qquad \left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)$$

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]^\top$$

Mean and (Co)Variance

Mean and (co)variance are often useful to describe properties of data distributions (expected values and spread).

Summary

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu} \qquad \left(\frac{1}{N}\sum_{n=1}^{N} x_n\right)$$

$$\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}]\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]^\top$$

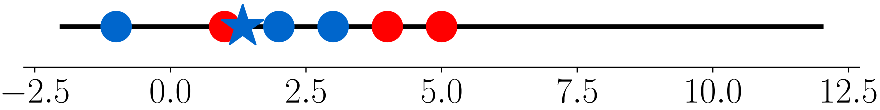
Compute the mean and (co)variance of the following datasets

$$\mathcal{D}_1 := \{-2, -1, 2\} \qquad \mathcal{D}_2 := \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 4 \end{bmatrix} \right\}$$

Translation: Effect on the Mean

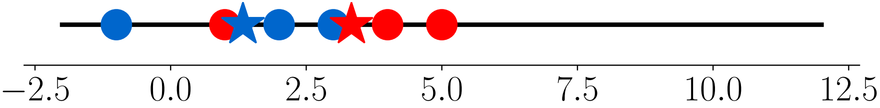
- What happens to the mean of a dataset if we shift/translate it by 2?

Translation: Effect on the Mean



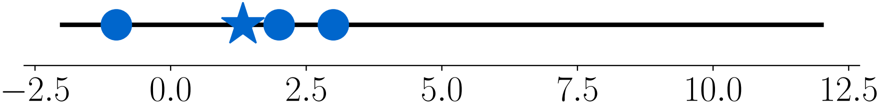
- What happens to the mean of a dataset if we shift/translate it by 2?

Translation: Effect on the Mean



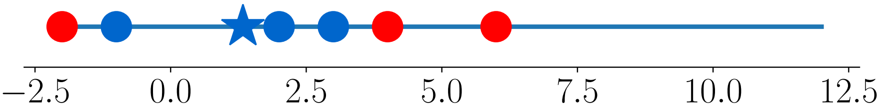
- What happens to the mean of a dataset if we shift/translate it by 2?

Scaling: Effect on the Mean



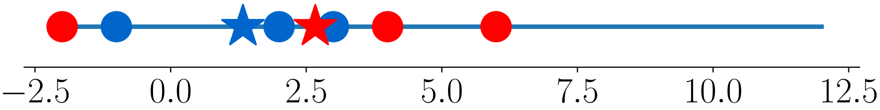
- What happens to the mean of a dataset if we scale it by 2?

Scaling: Effect on the Mean



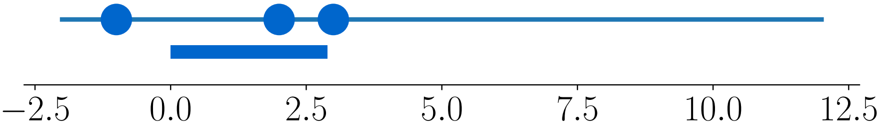
- What happens to the mean of a dataset if we scale it by 2?

Scaling: Effect on the Mean



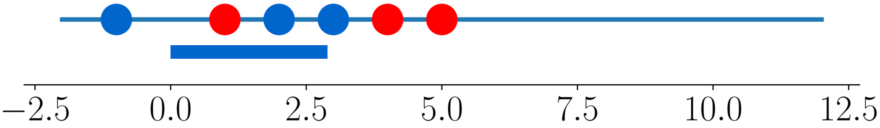
- What happens to the mean of a dataset if we scale it by 2?

Translation: Effect on the Variance



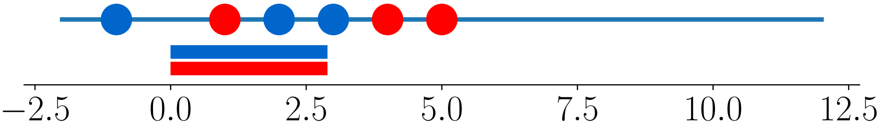
- What happens to the variance of a dataset if we shift it by 2?

Translation: Effect on the Variance



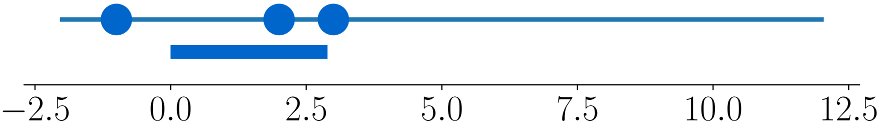
- What happens to the variance of a dataset if we shift it by 2?

Translation: Effect on the Variance



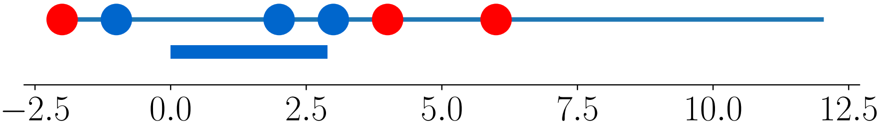
- What happens to the variance of a dataset if we shift it by 2?

Scaling: Effect on the Variance



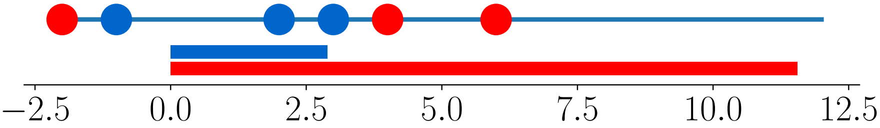
- What happens to the variance of a dataset if we scale it by 2?

Scaling: Effect on the Variance



- What happens to the variance of a dataset if we scale it by 2?

Scaling: Effect on the Variance



- What happens to the variance of a dataset if we scale it by 2?

Linear/Affine Transformations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\mu},\ \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] =$$

$$\mathbb{V}[\boldsymbol{y}] =$$

Linear/Affine Transformations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\mu},\ \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] =$$

Linear/Affine Transformations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\mu},\ \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] =$$

Linear/Affine Transformations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\mu},\ \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathbb{V}_{\boldsymbol{x}}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_{\boldsymbol{x}}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$$

Sum of Random Variables

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$ random variables. Then

$$\mathbb{E}[\boldsymbol{x} \pm \boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] \pm \mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]$$
$$\mathbb{V}[\boldsymbol{x} \pm \boldsymbol{y}] = \mathbb{V}[\boldsymbol{x}] + \mathbb{V}[\boldsymbol{y}] \pm \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] \pm \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{x}]$$