Motivation: Parameter Estimation

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

October 3, 2022

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes
- Things in the world are influenced by many (hidden) factors

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes
- Things in the world are influenced by many (hidden) factors
- Things seem random to us, until we understand them

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes
- Things in the world are influenced by many (hidden) factors
- Things seem random to us, until we understand them
- In ML, we use data to improve understanding

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes
- Things in the world are influenced by many (hidden) factors
- Things seem random to us, until we understand them
- In ML, we use data to improve understanding
- To express our understanding, we use **probability**

Machine Learning

- ML's goals: **predict** the world, and **optimise** outcomes
- Things in the world are influenced by many (hidden) factors
- Things seem random to us, until we understand them
- In ML, we use data to improve understanding
- To express our understanding, we use **probability**

Machine Learning and Statistics
are **almost the same**

Overview

Probability

Statistical Modelling

Assumed: Probabilities and Densities

I assume that you know about:

- Probability spaces and random variables
- Probability densities, e.g. $P(0.4 \leqslant X \leqslant 0.5) = \int_{0.4}^{0.5} p_X(x)\mathrm{d}x$
- Joint random variables, e.g. $P(X = 3, Y = 4)$
- Joint random densities, e.g.

$$P(0.4 \leqslant X \leqslant 0.5, 0.8 \leqslant Y \leqslant 0.85) = \int_{0.4}^{0.5}\int_{0.8}^{0.85} p_{XY}(x, y)\mathrm{d}y\mathrm{d}x \quad (1)$$

See exercise sheet for recap of notation, exercises, and pointers to revision material.

Probabilities and vectors

- Statistical modelling deals with many random variables at a time.

Probabilities and vectors

- Statistical modelling deals with many random variables at a time.
- Collect them in **vectors** for easier notation.

Probabilities and vectors

- Statistical modelling deals with many random variables at a time.
- Collect them in **vectors** for easier notation.
- Not conceptually different from joint distributions on scalars.

Probabilities and vectors

- Statistical modelling deals with many random variables at a time.
- Collect them in **vectors** for easier notation.
- Not conceptually different from joint distributions on scalars.

Skill: Probabilities and vectors

- For multiple joint random variables $X_1, X_2, X_3 \in \mathbb{R}$ with density

$$p_{X_1,X_2,X_3}(x_1, x_2, x_3)\,, \tag{2}$$

  we can interchangably denote this using a vector RV $X \in \mathbb{R}^3$ with density $p_X(\mathbf{x})$.

Skill: Probabilities and vectors

- For multiple joint random variables $X_1, X_2, X_3 \in \mathbb{R}$ with density

$$p_{X_1,X_2,X_3}(x_1, x_2, x_3)\,, \tag{2}$$

  we can interchangably denote this using a vector RV $X \in \mathbb{R}^3$ with density $p_X(\mathbf{x})$.
- This can shorten notation when specifying densities, e.g. for $0 \leqslant x_n \leqslant 1$, we can have

$$p(x_1, x_2, x_3) = \frac{1}{C}(x_1^2 + x_2^2 + x_3^2) = \frac{1}{C}||\mathbf{x}||^2 \tag{3}$$

Skill: Probabilities and vectors

- For multiple joint random variables $X_1, X_2, X_3 \in \mathbb{R}$ with density

$$p_{X_1,X_2,X_3}(x_1, x_2, x_3)\,, \tag{2}$$

  we can interchangably denote this using a vector RV $X \in \mathbb{R}^3$ with density $p_X(\mathbf{x})$.

- This can shorten notation when specifying densities, e.g. for $0 \leqslant x_n \leqslant 1$, we can have

$$p(x_1, x_2, x_3) = \frac{1}{C}(x_1^2 + x_2^2 + x_3^2) = \frac{1}{C}||\mathbf{x}||^2 \tag{3}$$

- See exercise sheet for some practice.

Overview

Probability

Statistical Modelling

Example: The shaking desk

Figure: xkcd #228

- Why is it shaking?
- How can I stop it?
- How much will it shake?

Statistical View on the World

Data Generating Process

We assume that the data we observe is the outcome of some random process. Each observation is one random variable. In this course, probabilities of the data generating process are denoted with $\mathbb{P}(\cdot)$, and which has distribution $\pi(\cdot)$.

Statistical View on the World

Data Generating Process

We assume that the data we observe is the outcome of some random process. Each observation is one random variable. In this course, probabilities of the data generating process are denoted with $\mathbb{P}(\cdot)$, and which has distribution $\pi(\cdot)$.

Example:

- We observe a dataset of 3 values $\{x_n\}_{n=1}^3$.
- This has density $\pi_{X_1,X_2,X_3}(x_1, x_2, x_3)$.

Independent Identically Distributed

Independent Identically Distributed (iid) Assumption

Often, we assume that random variables in a dataset are independent and identically distributed, which means that each random variable has the same distribution. Groups of RVs can also be iid.

Independent Identically Distributed

Independent Identically Distributed (iid) Assumption

Often, we assume that random variables in a dataset are independent and identically distributed, which means that each random variable has the same distribution. Groups of RVs can also be iid.

Examples:

$$\pi_{X_1,X_2,X_3}(x_1, x_2, x_3) = \prod_{n=1}^{3} \pi(x_n)$$

$$\pi_{X_1,Y_1,X_2,Y_2,\ldots}(x_1, y_1, x_2, y_2, \ldots) = \pi_{X,Y}(\mathbf{x}, \mathbf{y}) \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^N$$

$$= \prod_{n=1}^{N} \pi(x_n, y_n)$$

Statistical Model

Statistical Model

A statistical model is a random process used by a statistician (i.e. “us”) to explain data coming from the data generating process. In this course, probabilities of the model are denoted with $P(\cdot)$, which has the distribution $p(\cdot)$.

Statistical Model

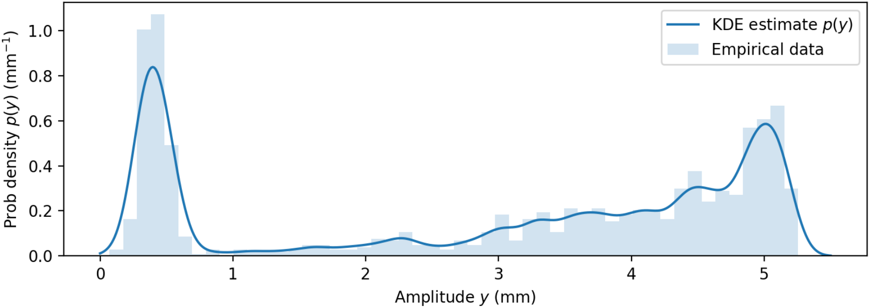
Statistical Model

A statistical model is a random process used by a statistician (i.e. "us") to explain data coming from the data generating process. In this course, probabilities of the model are denoted with $P(\cdot)$, which has the distribution $p(\cdot)$.

- The data generating process $\pi(\cdot)$ is unknown, and therefore usually different to our statistical model $p(\cdot)$.
- Our goal is to make them similar!
- A model often depends on some parameters $\boldsymbol{\theta}$, denoted as $p(\cdot|\boldsymbol{\theta})$, which are used to adjust the statistical model to fit the data.

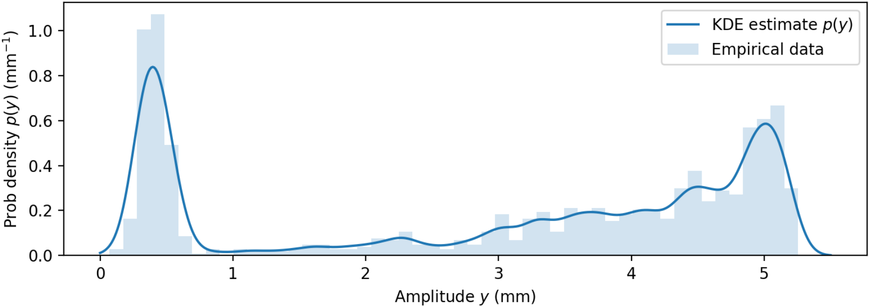
The shaking desk: Gathering data

Measure amplitude at various fixed points during the day.



The shaking desk: Gathering data

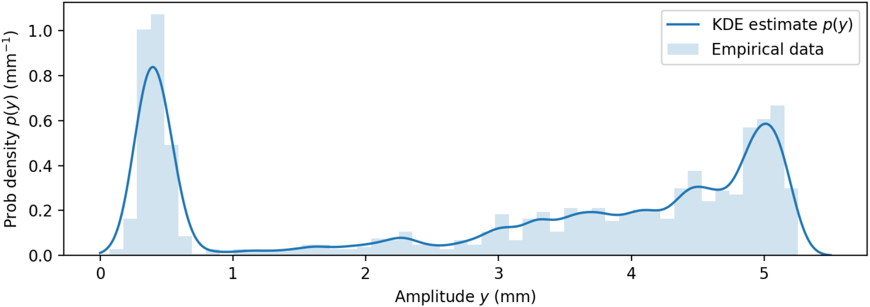
Measure amplitude at various fixed points during the day.



- We observe data as an iid sample from $\pi(y)$.
- Our model for unknown phenomenon: $p(y|\boldsymbol{\theta})$.

The shaking desk: Gathering data

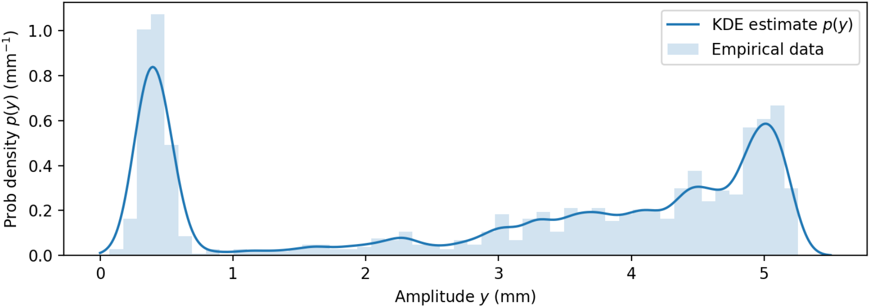
Measure amplitude at various fixed points during the day.



- We observe data as an iid sample from $\pi(y)$.
- Our model for unknown phenomenon: $p(y|\boldsymbol{\theta})$.
- Parameters $\boldsymbol{\theta}$ adjust shape of $p(y|\boldsymbol{\theta})$

The shaking desk: Gathering data

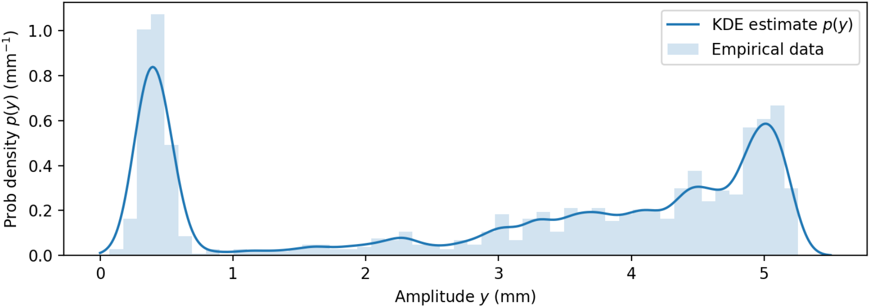
Measure amplitude at various fixed points during the day.



- We observe data as an iid sample from $\pi(y)$.
- Our model for unknown phenomenon: $p(y|\boldsymbol{\theta})$.
- Parameters $\boldsymbol{\theta}$ adjust shape of $p(y|\boldsymbol{\theta})$ ⏩ Find from data.
- Can we predict whether the shaking will happen?

The shaking desk: Gathering data

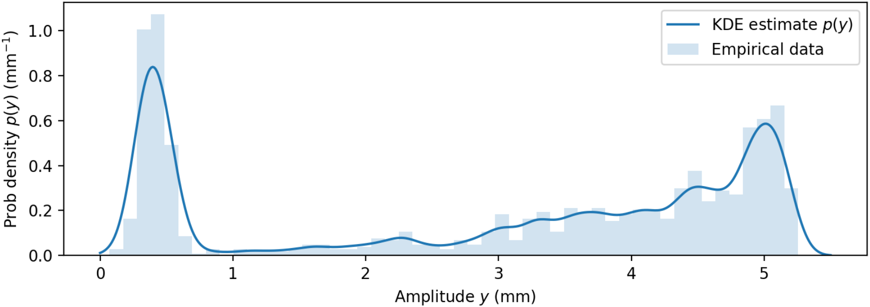
Measure amplitude at various fixed points during the day.



- We observe data as an iid sample from $\pi(y)$.
- Our model for unknown phenomenon: $p(y|\boldsymbol{\theta})$.
- Parameters $\boldsymbol{\theta}$ adjust shape of $p(y|\boldsymbol{\theta})$ ⏩ Find from data.
- Can we predict whether the shaking will happen? ⏩ No.

The shaking desk: Gathering data

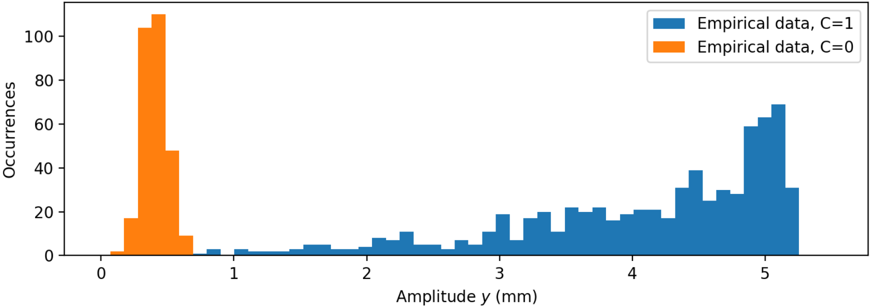
Measure amplitude at various fixed points during the day.



- We observe data as an iid sample from $\pi(y)$.
- Our model for unknown phenomenon: $p(y|\boldsymbol{\theta})$.
- Parameters $\boldsymbol{\theta}$ adjust shape of $p(y|\boldsymbol{\theta})$ ⏩ Find from data.
- Can we predict whether the shaking will happen? ⏩ No.
- We can at least predict: Unlikely to be small amplitude.

The shaking desk: Gathering data

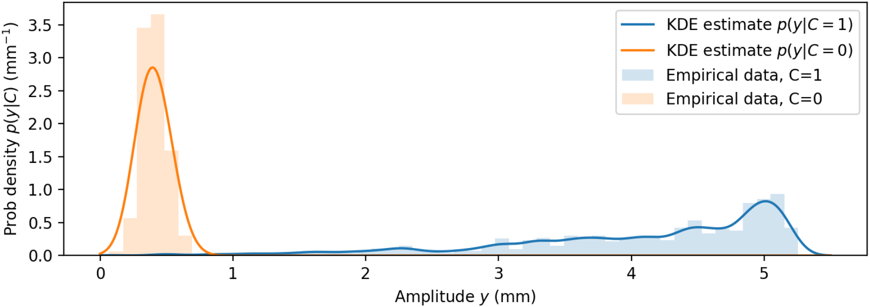
More data: Also measure whether your colleague is present (C).



- Now we observe pairs c_n, y_n from a generating process $\pi(c, y)$
- We now find two models $p(y|C = 0, \boldsymbol{\theta})$ and $p(y|C = 1, \boldsymbol{\theta})$
- It seems that $C = 1$ indicates larger shaking
- Given C, we can now predict with more certainty!

The shaking desk: Gathering data

More data: Also measure whether your colleague is present (C).



- Now we observe pairs c_n, y_n from a generating process $\pi(c, y)$
- We now find two models $p(y|C = 0, \boldsymbol{\theta})$ and $p(y|C = 1, \boldsymbol{\theta})$
- It seems that $C = 1$ indicates larger shaking
- Given C, we can now predict with more certainty!

Example: The shaking desk

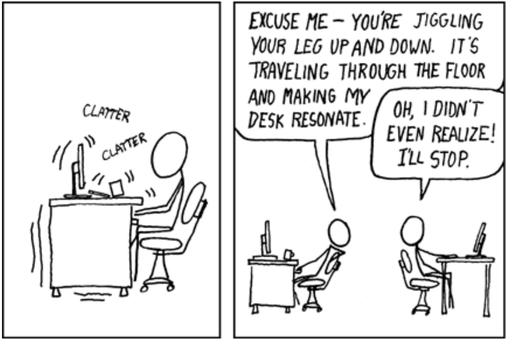
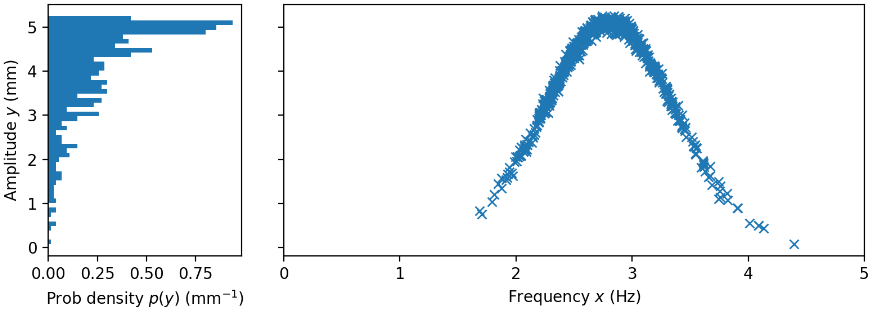


Figure: xkcd #228

- Why is it shaking?
- How can I stop it?
- How much will it shake?

The shaking desk: Gathering more data

Data: For colleague present ($C = 1$), measure jiggling frequency x.



- Could we estimate on model per x? I.e. $p(y|x, \boldsymbol{\theta})$.
- We should be able to predict the amplitude very accurately!
- Uncertainty reduces, predictions improve

Example: The shaking desk

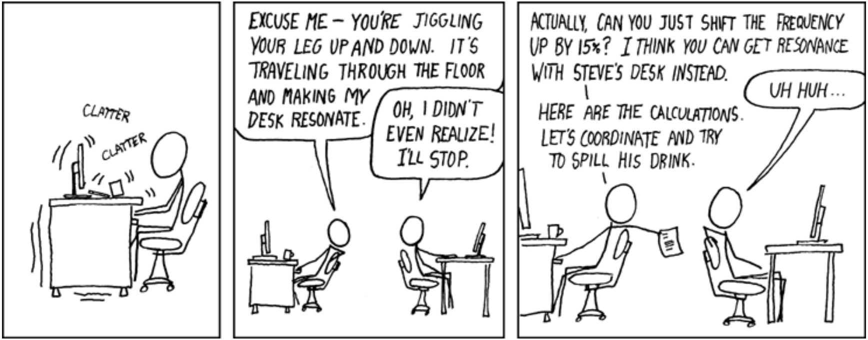
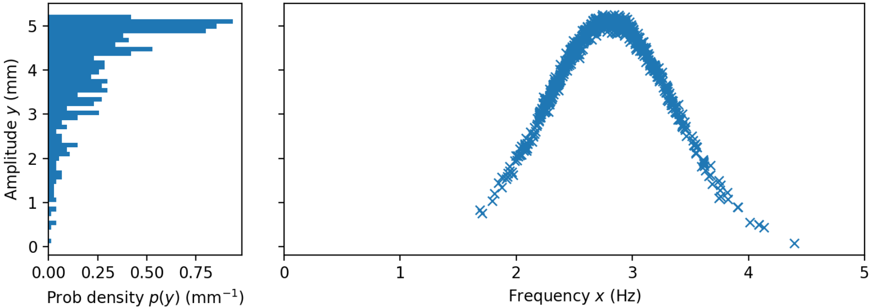


Figure: xkcd #228

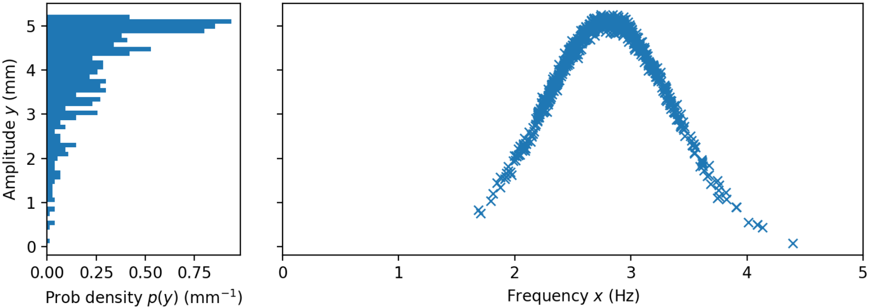
- Why is it shaking?
- How can I stop it?
- How much will it shake?

Curve Fitting



Curve fitting problem!

Curve Fitting



Curve fitting problem!

Regression