Investigating accelerometry data as a potential tool to help understand rare diseases

Niels Heijnekamp

Supervised by Dr Martin Bootsma and Dr Peter van Hasselt

June 19, 2025

Data can help in understanding diseases

Data can help in understanding diseases

	time	X	Y	Z
900088	2024-10-10 14:20:02	-0.036910392	-0.2834109	0.0427539228
900089	2024-10-10 14:20:02	0.033089608	-0.3454109	0.0007539228
900090	2024-10-10 14:20:02	0.147089608	-0.5254109	0.0747539228
900091	2024-10-10 14:20:03	0.197089608	-0.8184109	0.0977539228

Caption: Example of the available raw data.

Data can help in understanding diseases

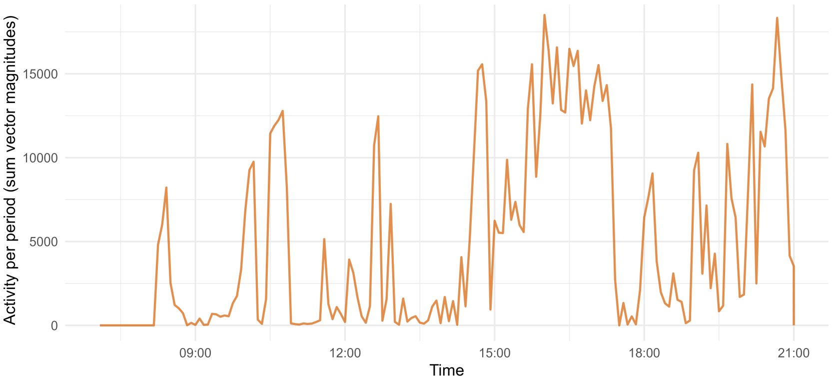
	time	X	Y	Z
900088	2024-10-10 14:20:02	-0.036910392	-0.2834109	0.0427539228
900089	2024-10-10 14:20:02	0.033089608	-0.3454109	0.0007539228
900090	2024-10-10 14:20:02	0.147089608	-0.5254109	0.0747539228
900091	2024-10-10 14:20:03	0.197089608	-0.8184109	0.0977539228

Caption: Example of the available raw data.

Advantages accelerometry

- Potentially abundant
- Little effort
- Minimal disturbance

Current efforts focus on a long time scale



Caption: Estimated activity levels from accelerometry.

Table of contents

Measurements are comparable across individuals

Data details

- Worn on left hip
- Uninterrupted sampling
- 30 Hz

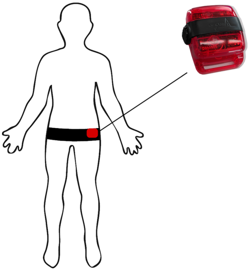
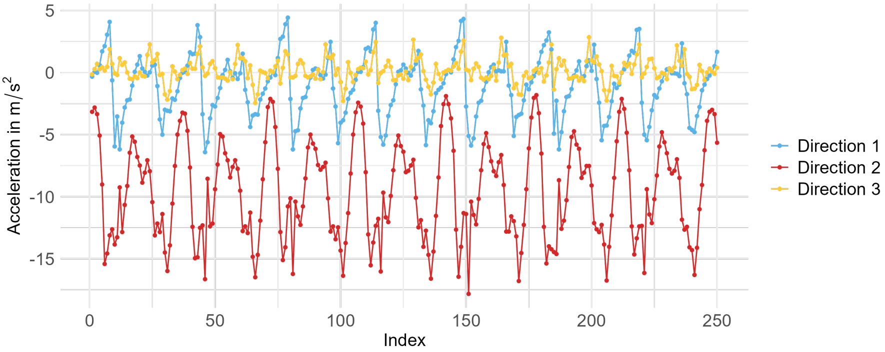


Figure reference: "Using unsupervised machine learning to quantify physical activity from accelerometry in a diverse and rapidly changing population". Available from: `https://www.researchgate.net/figure/Shows-the-accelerometer-the-ActiGraph-GT3X_and-how-it-was-instructed-to-be-worn_fig1_369824532` [accessed 8 Jun 2025].

Measurements are made on three orthogonal axes



Caption: Acceleration as a three-dimensional time series.

Measurements are with respect to a free fall

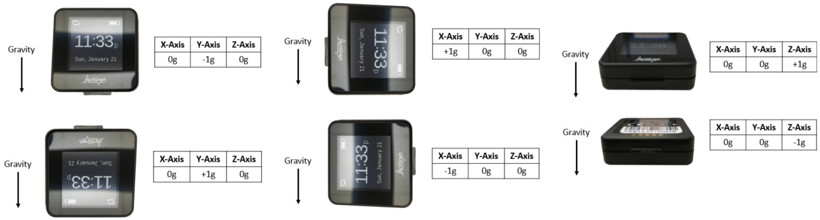


Figure reference: ActiGraph. (2020). ActiGraph Link device orientation (Serial numbers starting with TAS). Available from `https://actigraphcorp.my.site.com/support/s/article/ActiGraph-Link-device-orientation-Serial-numbers-starting-with-TAS` [accessed 28 Jan 2025]

Measurements are relative to the accelerometer

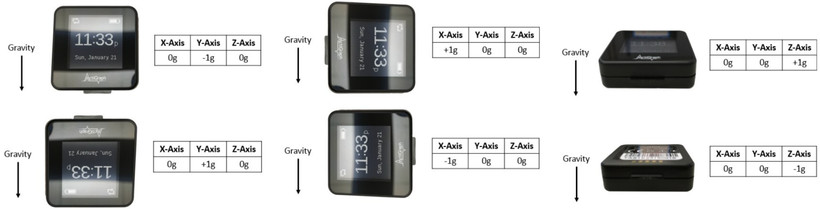


Figure reference: ActiGraph. (2020). ActiGraph Link device orientation (Serial numbers starting with TAS). Available from `https://actigraphcorp.my.site.com/support/s/article/ActiGraph-Link-device-orientation-Serial-numbers-starting-with-TAS` [accessed 28 Jan 2025]

Data of two diseases has been collected

Details Hurler group

- Metabolic disease
- Progressive deterioration, terminal
- Damage to skeletal system
- Nine individuals ages 9-17, unsupervised

Data of two diseases has been collected

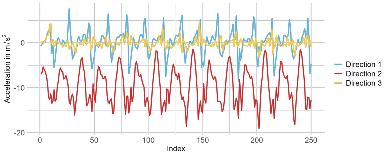
Details Hurler group

- Metabolic disease
- Progressive deterioration, terminal
- Damage to skeletal system
- Nine individuals ages 9-17, unsupervised

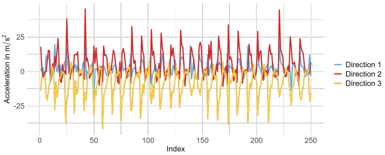
Details Batten group

- Neurodegenerative disease
- Progressive deterioration, terminal
- Vision and walking impaired
- Ten individuals ages 8-19, walking tests including videos

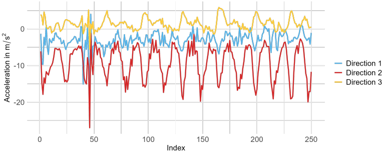
Accelerometer output shares certain properties



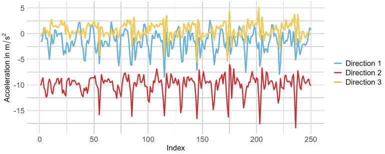
(a) Healthy adult (walking).



(b) Healthy child (running).



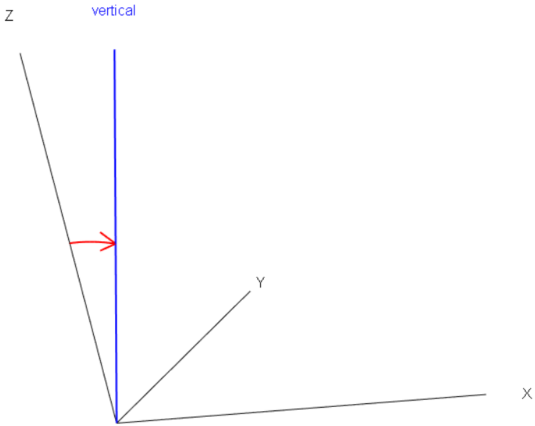
(c) Child with Hurler (walking).



(d) Child with Batten (walking).

Caption: Examples of walking.

Vertical acceleration can be estimated from gravity...



Caption: Example rotation of the accelerometer relative to Earth's surface.

...although there is no such thing as vertical acceleration

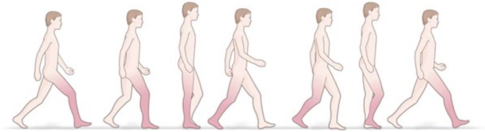


Figure reference: Musculoskeletal Key. (n.d.). Gait and posture analysis. In Musculoskeletal Key. Retrieved June 13, 2025, from `https://musculoskeletalkey.com/gait-and-posture-analysis/`. Modified here.

...although there is no such thing as vertical acceleration

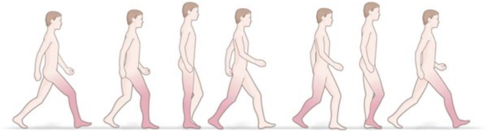
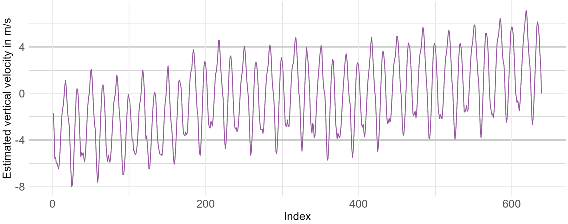
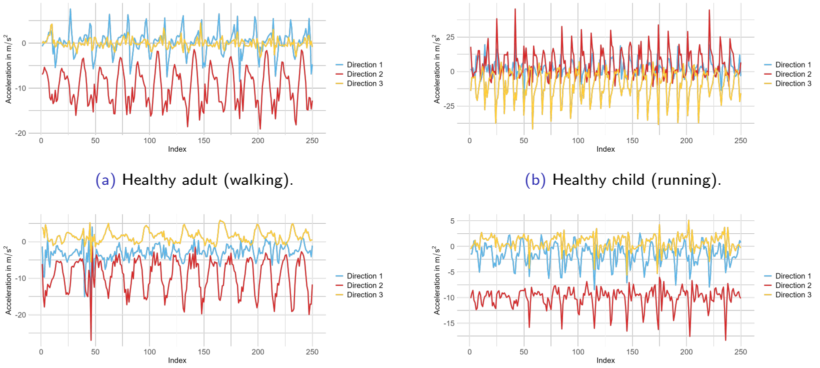


Figure reference: Musculoskeletal Key. (n.d.). Gait and posture analysis. In Musculoskeletal Key. Retrieved June 13, 2025, from `https://musculoskeletalkey.com/gait-and-posture-analysis/`. Modified here.



Caption: Drifting effect in estimated velocity.

Degree of periodicity is hard to make precise



(a) Healthy adult (walking).

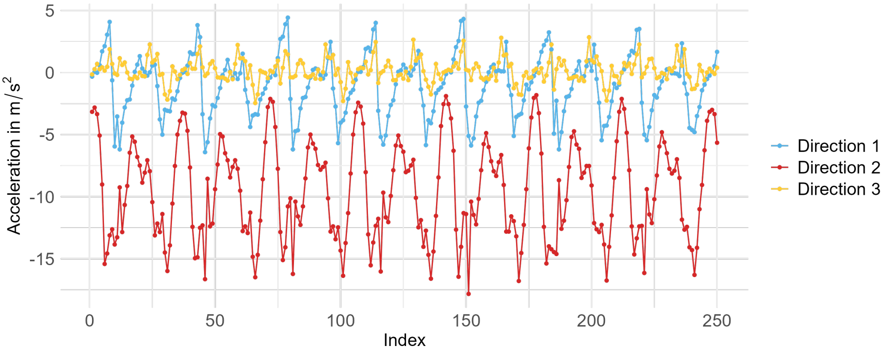
(b) Healthy child (running).

(c) Child with Hurler (walking).

(d) Child with Batten (walking).

Caption: Examples of walking.

We are interested in strength of periodicity of time series



Caption: Example of a three-dimensional time series.

Several ideas exist for determining strength of periodicity

Ways to measure periodicity

Several ideas exist for determining strength of periodicity

Ways to measure periodicity

- Correlation at period lag

Several ideas exist for determining strength of periodicity

Ways to measure periodicity

- Correlation at period lag
- Similarity with periodic function

Several ideas exist for determining strength of periodicity

Ways to measure periodicity

- Correlation at period lag
- Similarity with periodic function
- Discrete Fourier transform

Several ideas exist for determining strength of periodicity

Ways to measure periodicity

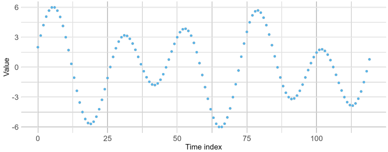
- Correlation at period lag
- Similarity with periodic function
- Discrete Fourier transform

Discrete Fourier transform (DFT)

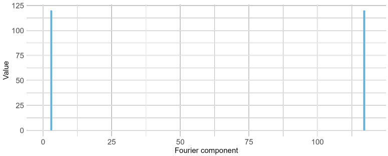
Let $(x_t)_{t \in T}$ be a univariate time series with $T = \{0, 1, ..., n-1\}$. The discrete Fourier transform DFT is given by $\mathrm{DFT}\left((x_t)_t\right) = (y_t)_{t \in T}$, where

$$y_t = \sum_{k=0}^{n-1} x_k \cdot \exp\left(-2i\pi \frac{tk}{n}\right).$$

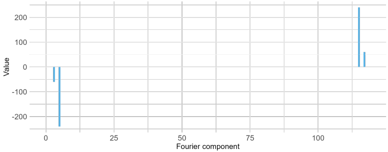
Fourier transform measures “strength” of frequencies



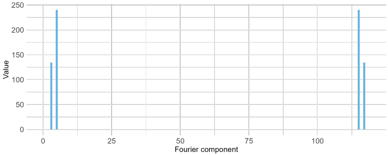
(a) The function
$f : x \mapsto \sin(\frac{\pi}{20}x) + 2\cos(\frac{\pi}{20}x) + 4\sin(\frac{\pi}{12}x)$.



(b) Real part of the DFT.



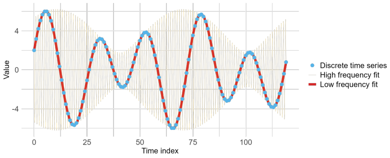
(c) Imaginary part of the DFT.



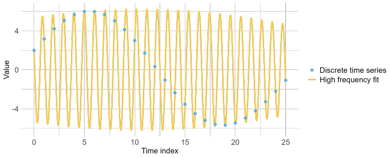
(d) Modulus of the DFT.

Caption: The DFT applied to a sum of sinusoids.

Only part of the DFT's outcome is of interest



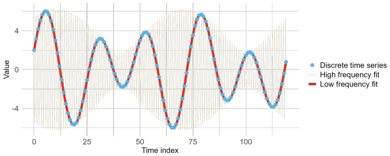
(a) Two sinusoid reconstructions of f.



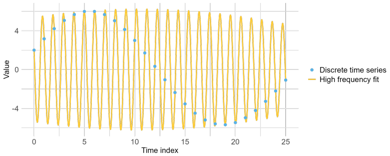
(b) The high frequency reconstruction.

Caption: Different sinusoid reconstructions.

Only part of the DFT's outcome is of interest



(a) Two sinusoid reconstructions of f.



(b) The high frequency reconstruction.

Caption: Different sinusoid reconstructions.

Recap of details

- Periodicity is "concentratedness" frequencies
- Consider low frequencies and modulus
- Rescale to capture strength?

Fourier entries should be squared to capture energy

Parseval's Theorem

Let $(x_t)_t$ be a time series of length n and $(y_t)_t$ the sequence generated by applying the DFT to $(x_t)_t$. Then,

$$\sum_{k=0}^{n-1} |x_k|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |y_k|^2.$$

Fourier entries should be squared to capture energy

Parseval's Theorem

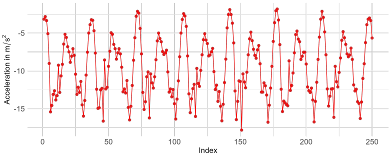
Let $(x_t)_t$ be a time series of length n and $(y_t)_t$ the sequence generated by applying the DFT to $(x_t)_t$. Then,

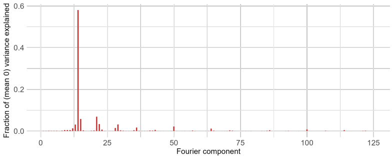
$$\sum_{k=0}^{n-1} |x_k|^2 = \frac{1}{n} \sum_{k=0}^{n-1} |y_k|^2.$$

Connection with sample variance

$$\sigma^2((x_k)_k) = \frac{1}{n} \sum_{k=0}^{n-1} |x_k|^2 = \frac{1}{n^2} \sum_{k=0}^{n-1} |y_k|^2.$$

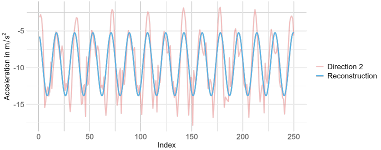
Energy spectral density is a potential tool to find periodicities



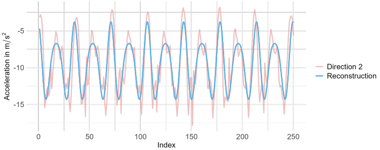


Caption: A time series and its energy spectral density (low frequencies, mean 0).

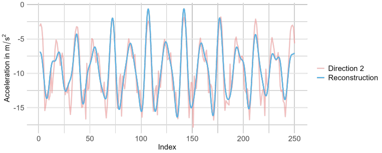
Energy spectral density is a potential tool to find periodicities



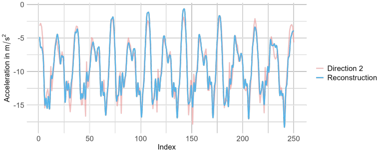
(a) Mean plus one component. Explains 0.580.



(b) Mean plus two components. Explains 0.649.



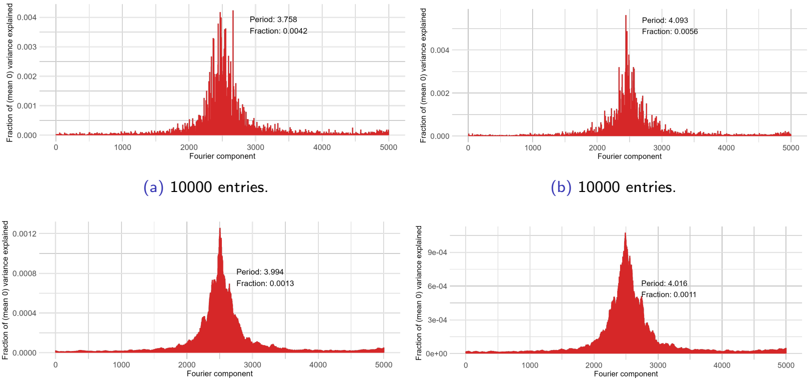
(c) Mean plus five components. Explains 0.773.



(d) Mean plus twenty components. Explains 0.930.

Caption: Reconstructions using sinusoids and the corresponding fraction of variance explained.

Variance can be reduced by smoothing energy



(a) 10000 entries.

(b) 10000 entries.

(c) Moving average over 50 entries applied to (a).

(d) Moving average over 50 entries applied to (b).

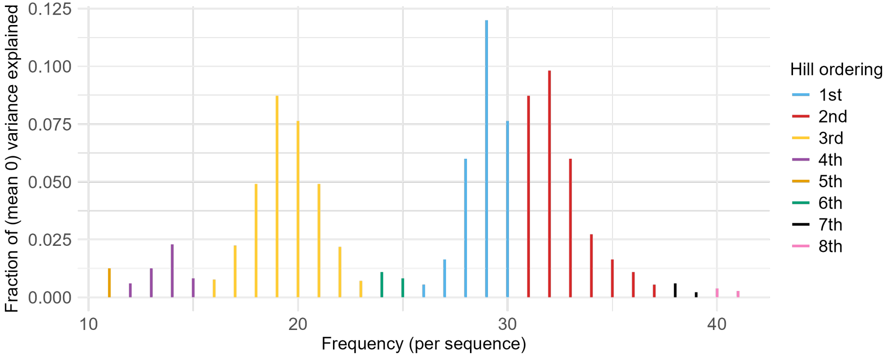
Caption: Fraction of variance explained compared to smoothed fraction of variance explained.

The amount of smoothing comes with a trade-off

Increasing bandwidth

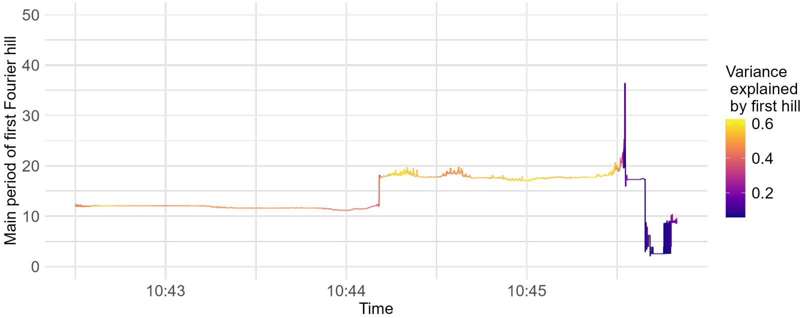
+ Lowers variance
+ Places peaks near centre
– Lowers resolution
– Introduces bias

We should capture "hills" when we know the number of underlying frequencies captured



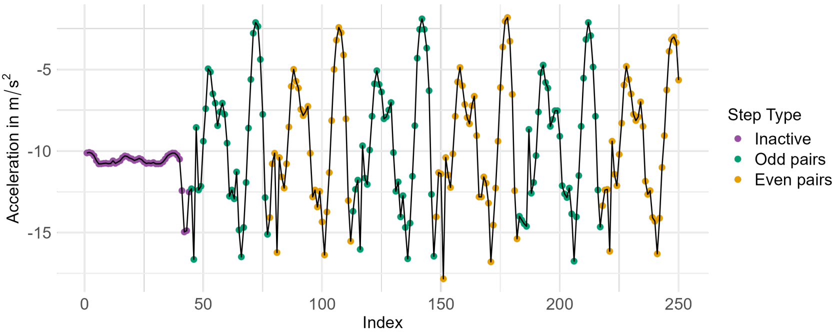
Caption: An example of Fourier hills.

We need local measures denoting step potential



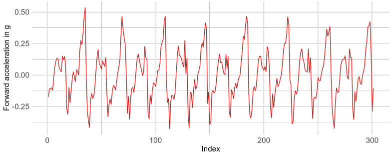
Caption: Results of the preparatory extraction phase.

We find steps through their period and correlation

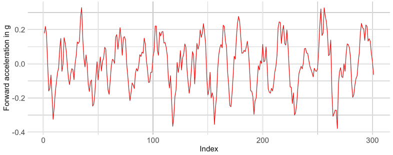


Caption: Vertical acceleration of a step sequence.

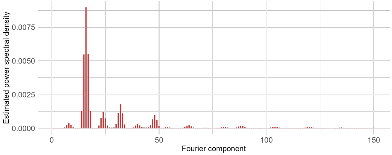
Collected sequences need a processing phase



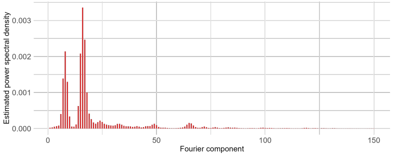
(a) Forward acceleration walking.



(b) Forward acceleration cycling.



(c) Estimated fraction explained in (a).



(d) Estimated fraction explained in (b).

Caption: Comparison of walking and cycling.

We can verify results whenever logs are available

Video

Different approaches to analysing our step data are possible

Approaches

Different approaches to analysing our step data are possible

Approaches

- Define features

Different approaches to analysing our step data are possible

Approaches

- Define features
- Cluster analysis

Different approaches to analysing our step data are possible

Approaches

- Define features
- Cluster analysis
- Noise analysis

We rescale steps before comparing them

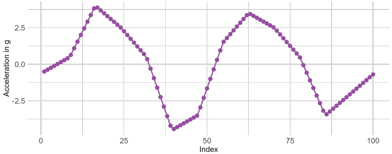
(a) Example step.

(b) Example step.

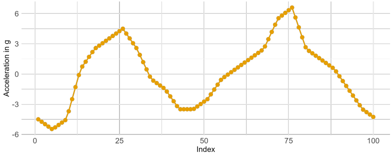
We rescale steps before comparing them

(a) Example step.

(b) Example step.



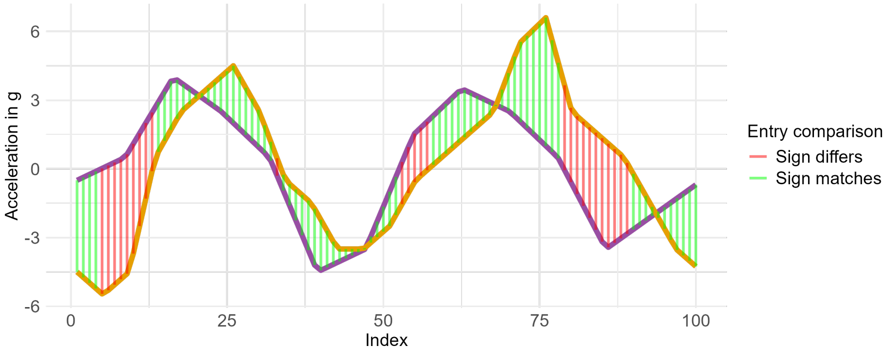
(c) Rescaled version of (a).



(d) Rescaled version of (b).

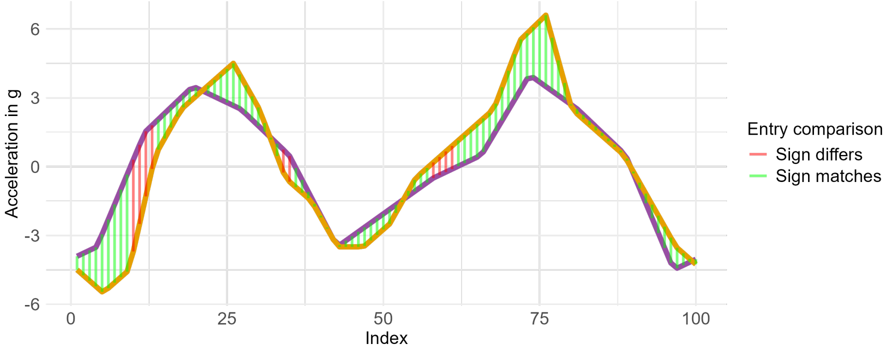
Caption: Examples of rescaling.

We consider several distances and dissimilarities



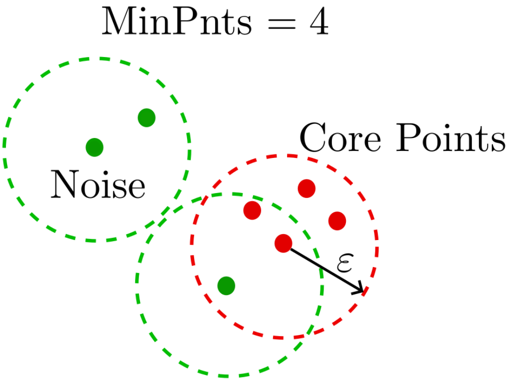
Caption: Step comparison without realignment.

We realign steps before comparing them



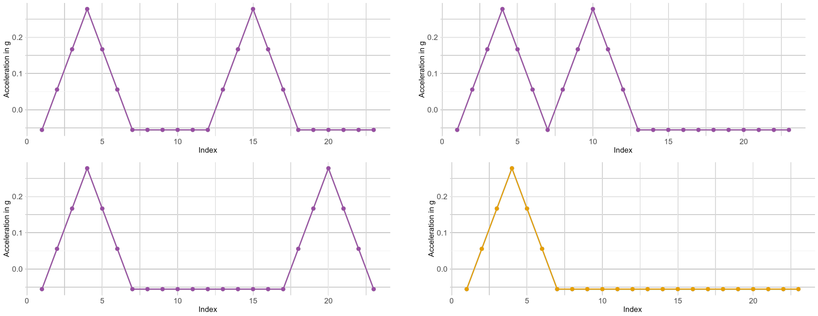
Caption: Step comparison with realignment.

Density based methods are suitable for outlier detection



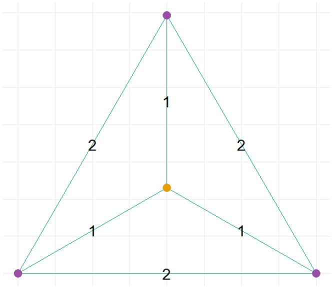
Caption: Noise and core points DBSCAN*.
Figure reference: Made by Sam Lindauer.

We want a method that works on dissimilarities



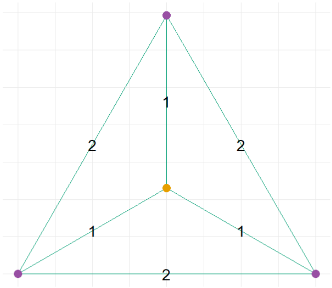
Caption: Different example steps.

We want a method that works on dissimilarities

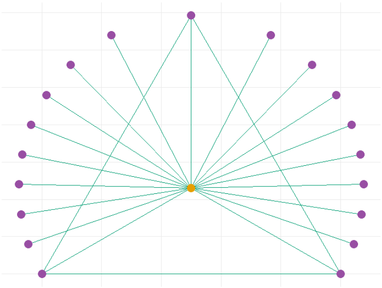


Caption: Scaled projection of example steps.

We want a method that works on dissimilarities

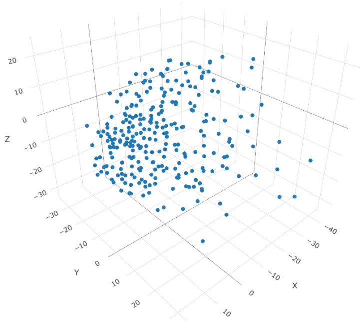


Caption: Scaled projection of example steps.



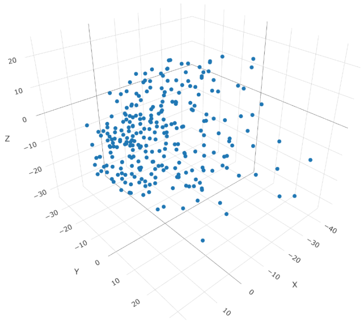
Caption: Scaled projection of more example steps.

We show density based methods for example data

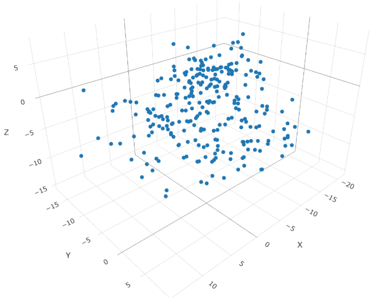


Caption: Absolute sum dissimilarities.

We show density based methods for example data

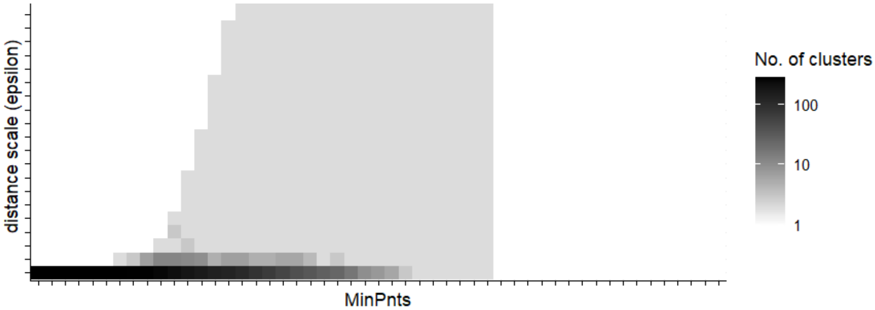


Caption: Absolute sum dissimilarities.



Caption: Sign difference dissimilarities.

Choice of parameters can be unclear



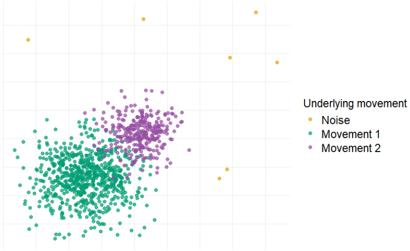
Caption: Clusters for different parameters (“local DBSCAN”).

Different dissimilarities give similar results

No.	Sum noise
5	0.215
2	0.116
6	0.114
10	0.103
8	0.086
4	0.074
9	0.054
3	0.028
1	0.018
7	0.012

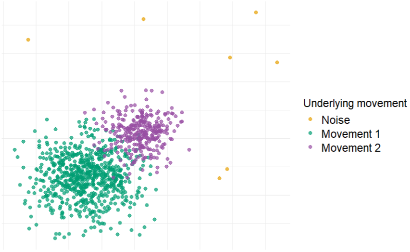
No.	Sign noise
5	0.270
2	0.098
6	0.097
8	0.082
10	0.078
3	0.064
4	0.064
9	0.055
7	0.043
1	0.029

Results allow for various interpretations

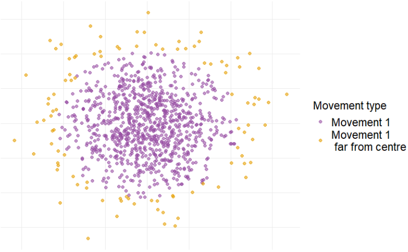


Caption: Noise as a separate category.

Results allow for various interpretations

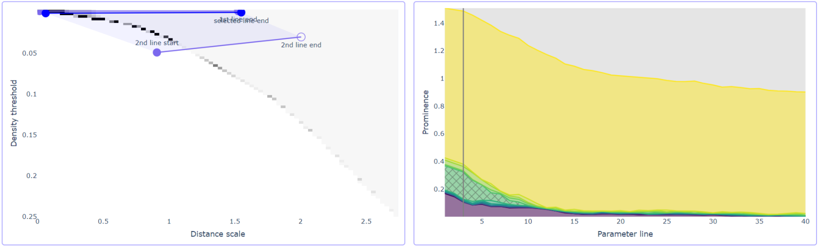


Caption: Noise as a separate category.



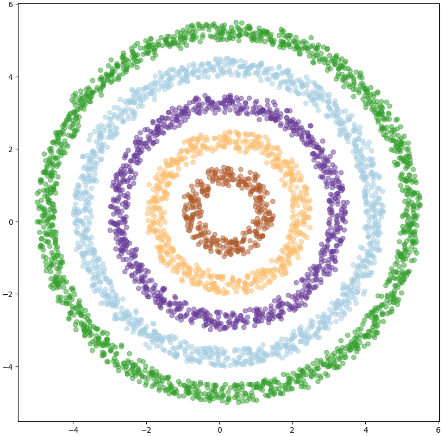
Caption: Noise as a definitional necessity.

Persistable finds a “persistent” parameter choice through DBSCAN*



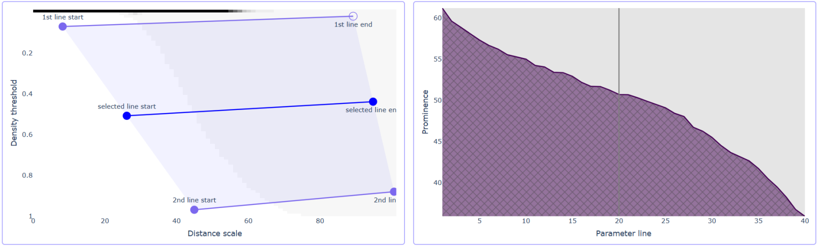
Caption: Persistable applied to example data.

Persistable finds a “persistent” parameter choice



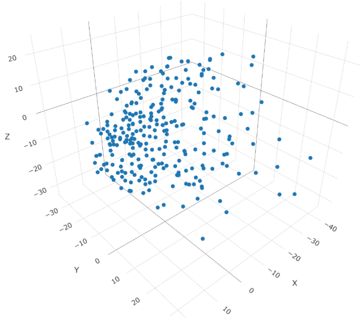
Caption: Example data.

Persistable indicates one cluster for the absolute sum data

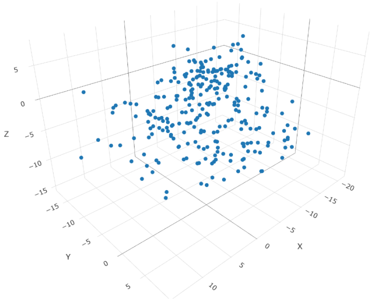


Caption: Persistable applied to the absolute sum data.

Persistable finds no noise on the example data



Caption: Robust Persistable clustering absolute sum data.



Caption: Robust Persistable clustering sign data.

We can compare connectedness to walking quality

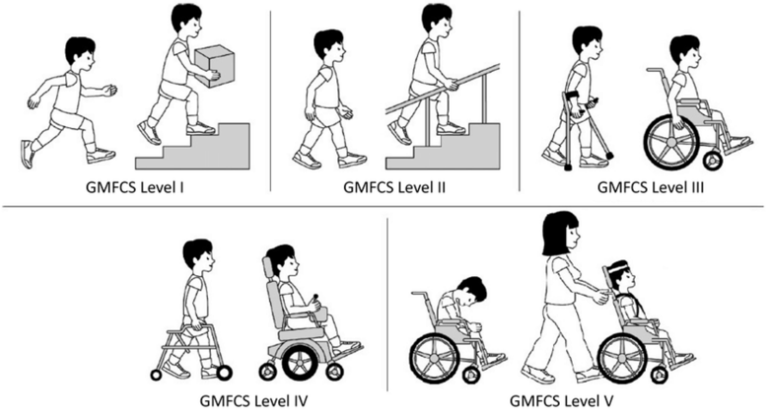
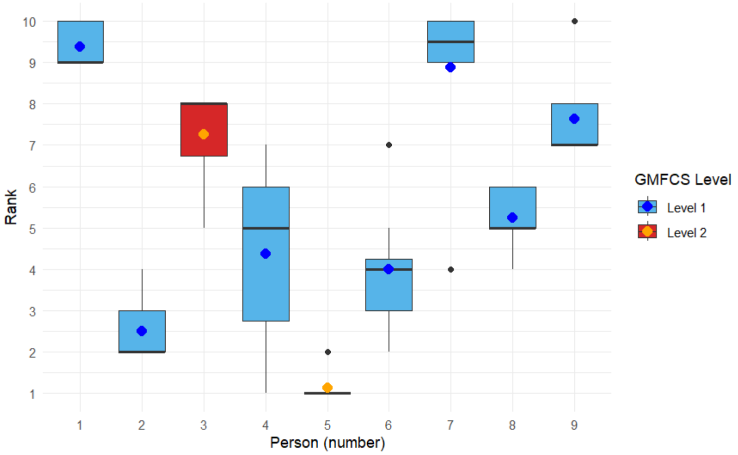


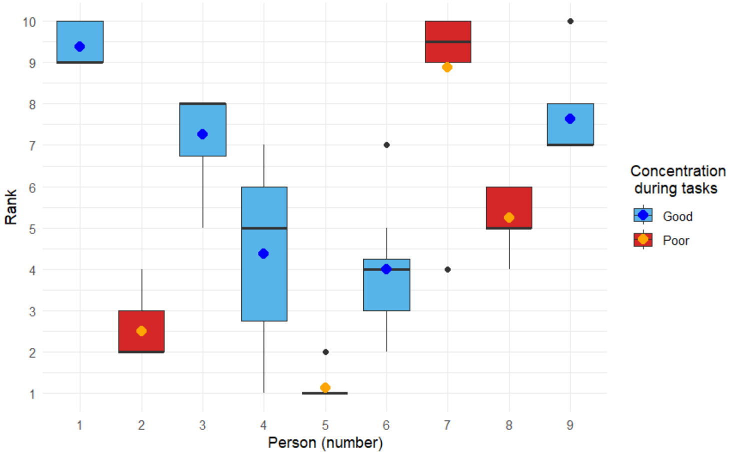
Figure reference: About Wheelchair. (n.d.). levels-of-gross-motor-function-classification-system-gmfcs. Retrieved June 10, 2025, from
`https://about-wheelchair.com/levels-of-gross-motor-function-classification-system-gmfcs/`.

Connectedness potentially correlates with walking quality



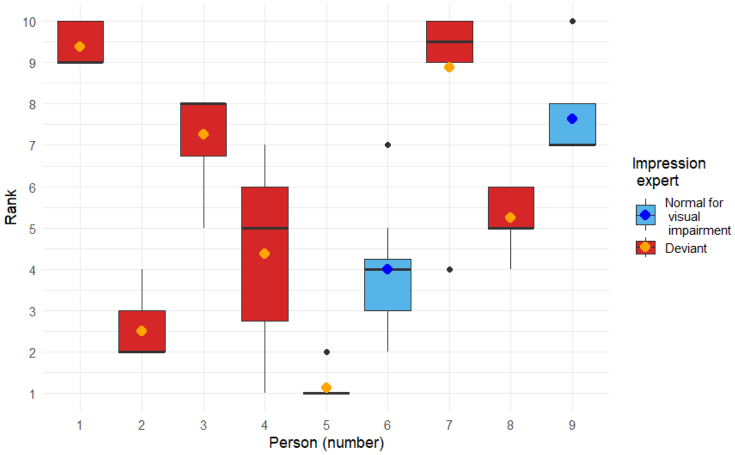
Caption: Connectedness compared to GMFCS level.

Connectedness potentially correlates with concentration



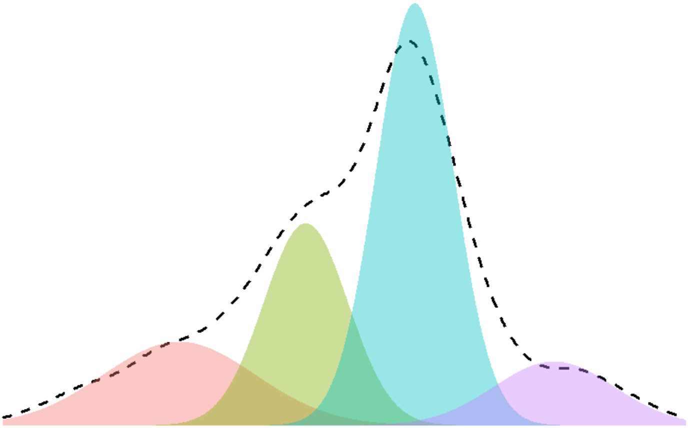
Caption: Connectedness compared to concentration.

Connectedness potentially correlates with anomalies



Caption: Connectedness compared to expert impression.

Our results point to using model based clustering



Caption: An example of data generated through Gaussians.

Future research can focus on several directions

Next steps

Future research can focus on several directions

Next steps

- Improve on data

Future research can focus on several directions

Next steps

- Improve on data
- Improve on assumptions

Future research can focus on several directions

Next steps

- Improve on data
- Improve on assumptions
- Improve on methods

Too early to say if results are promising

Results

- Can extract steps mathematically
- Improved understanding of nature of noise and data
- *Potentially* found a way to assess walking quality

Step extraction can be improved in several ways

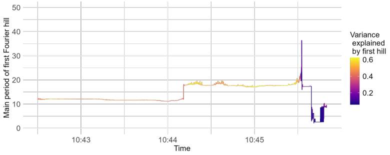
Assumptions

- May improve through orientation, consistent positioning, more accelerometers, more measurements.
- Step processing and angle assumptions are simplified.
- Nature of data is shown indirectly.

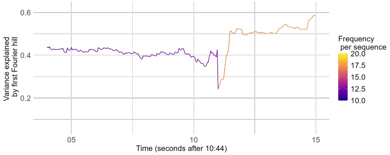
Research indicates promising directions

Next steps

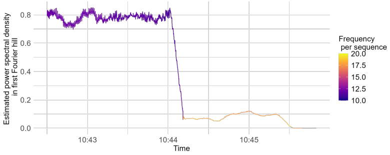
- Use more of the data
- Improving on extraction and processing opens many doors
- Move on to model based clustering



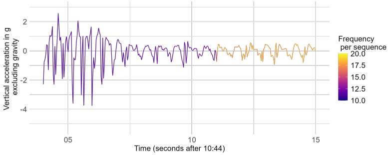
(a) Period over time.



(b) Periodicity strength at switch



(c) Power spectral density over time.



(d) Raw data at switch.

Caption: Results of the preparatory extraction phase.

Walking is modelled as a stochastic process

Interested in walking periods

Univariate stochastic process

A univariate stochastic process is a collection $(X_t)_{t \in T}$, where X_t is a random variable (over some probability space) and T is a sequence denoting time.

The walking process is wide-sense stationary

Let $(X_t)_t$ be a stochastic process and let

- $\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t)\,\mathrm{d}t$,
- $\gamma_X(s,t) = \operatorname{cov}(X_s, X_t) = \mathbb{E}[(X_s - \mathbb{E}[X_s])(X_t - \mathbb{E}[X_t])]$.

Wide-sense stationary stochastic process

Then $(X_t)_t$ is wide-sense stationary if

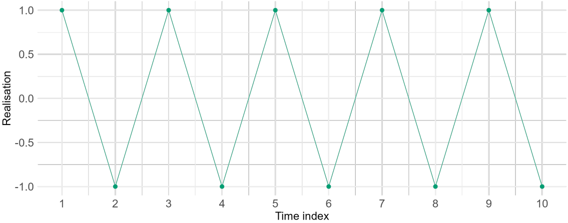
- $\gamma_X(t,t) < \infty$ for $t \in T$,
- $\mathbb{E}[X_s] = \mathbb{E}[X_t]$ for all $s, t \in T$,
- $\gamma_X(s+t, s) = \gamma_X(t, 0)$ for $s, t, s+t \in T$.

Wide-sense stationary models are common

Consider $(X_t)_{t\in T}$, $T = \{0, 1, 2, \ldots\}$ such that

$$\begin{cases} \mathbb{P}\left[X_0 = -1\right] = \frac{1}{2}, \\ \mathbb{P}\left[X_0 = 1\right] = \frac{1}{2}. \end{cases}$$

and $X_t = -X_{t-1}$ for $t \geq 1$.

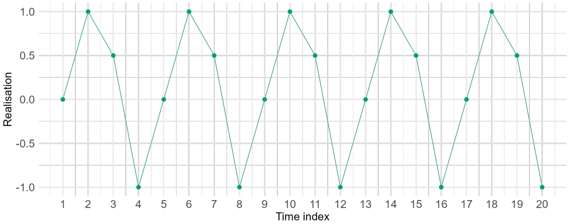


Wide-sense stationary models are common

Consider $(X_t)_{t \in T}$, $T = \{0, 1, 2, \dots\}$ such that

$$\mathbb{P}[X_0 = -1] = \mathbb{P}[X_0 = 0] = \mathbb{P}[X_0 = 1] = \mathbb{P}\left[X_0 = \frac{1}{2}\right] = \frac{1}{4},$$

and X_t is the value that comes after X_{t-1} in $(0, 1, 0.5, -1, 0, 1, \dots)$.

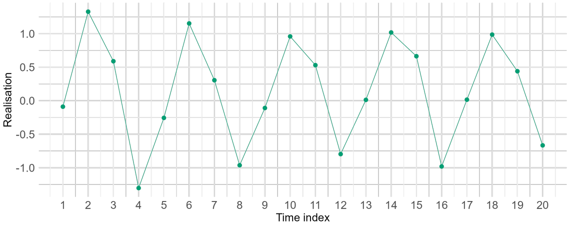


Wide-sense stationary models are common

Consider $(Y_t)_{t \in T}$, $T = \{0, 1, 2, ...\}$ such that

$$Y_t = X_t + \mathcal{N}(0, 0.04),$$

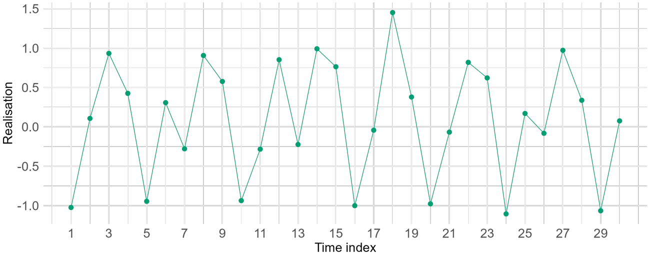
with $(X_t)_t$ as before and $\mathcal{N}(\mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2.



Wide-sense stationary models are common

Consider $(W_t)_t = (Z_t)_t + \mathcal{N}(0, 0.04)$, where for all t

$$\begin{cases} Z_t = X_t \text{ with probability } 0.9, \\ Z_t = 0 \text{ with probability } 0. \end{cases}$$



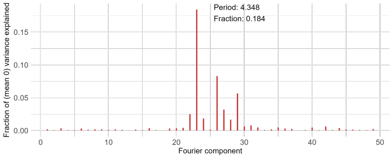
Wide-sense stationary processes has power spectral density

Power spectral density (Wiener-Khinchin-Einstein Theorem)

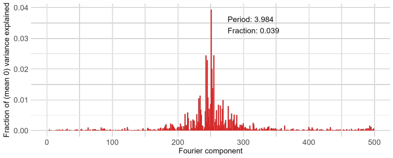
Consider a wide-sense stationary stochastic process $(X_t)_{t\in T}$. Then under light assumptions, e.g. $T = \{0, 1, ..., n-1\}$, the power spectral density at index $k \in \{0, 1, ..., n-1\}$ is given by

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\left|\sum_{t=0}^{n-1} X_t \cdot \exp\left(-2\pi i \frac{tk}{n}\right)\right|^2\right].$$

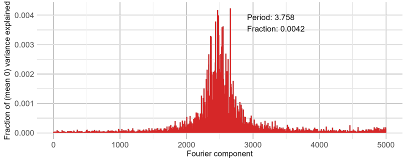
Energy spectral density is an estimator of power spectral density with large variance



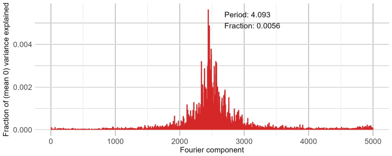
(a) 100 entries.



(b) 1000 entries.



(c) 10000 entries.



(d) 10000 entries.

Caption: Energy spectral density of the stochastic process $(W_t)_t$.

Different clustering techniques exist

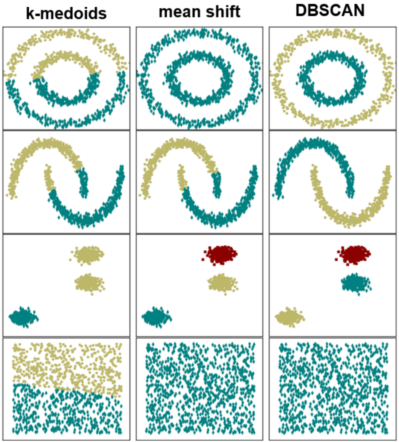


Figure reference: "Scikit-learn: Machine learning in Python" by F. Pedregosa et al. Journal of Machine Learning Research vol. 12. Modified here.

We can verify the nature of noise through k-medoids.

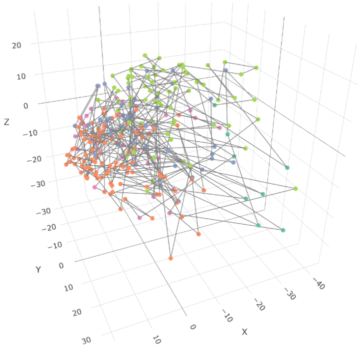
k-medoids

We let $M = (x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ be k cluster centres from n observations $(x_j)_j$, such that the cost C is minimised, where

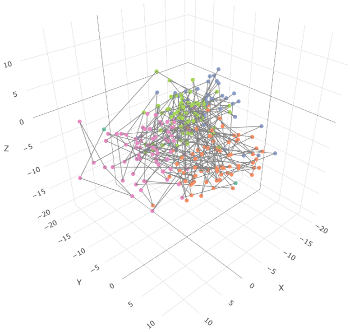
$$C(M) = \sum_{i=1}^{n} \min_{m \in M} \left(D(m, x_i) \right).$$

A point x_i is then assigned to $\mathrm{argmin}_{m \in M} \left(D(m, x_i) \right)$, arbitrarily breaking ties.

k-medoids affirms noise is not a separate category

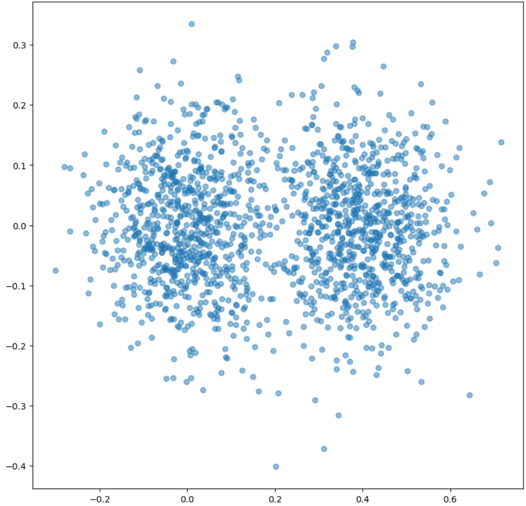


Caption: k-medoids for absolute sum dissimilarity.

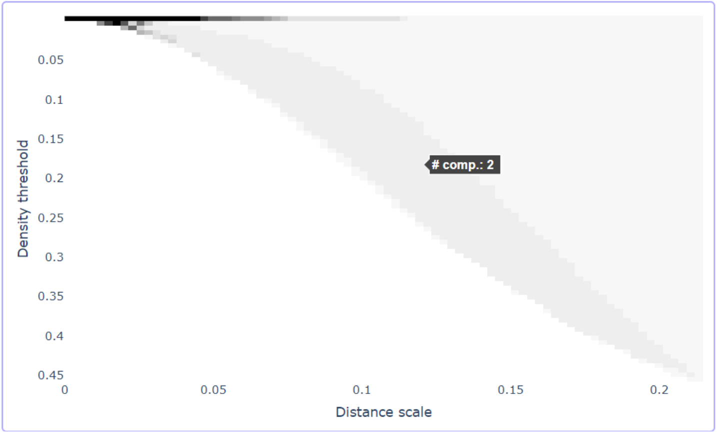


Caption: k-medoids for sign dissimilarity.

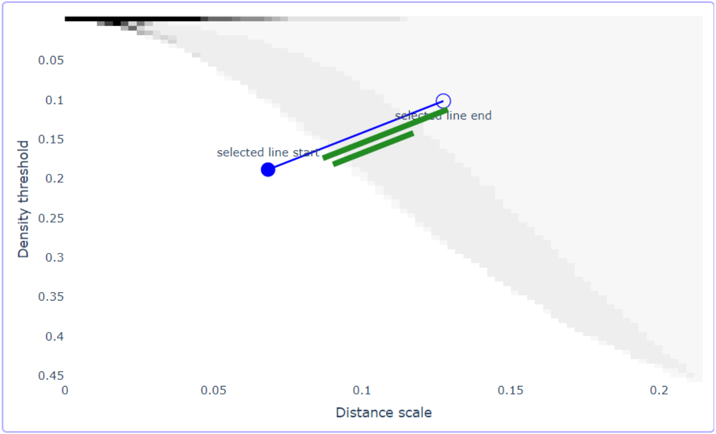
Example data



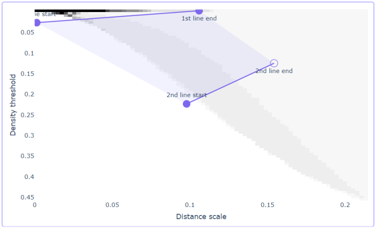
Component counting plane

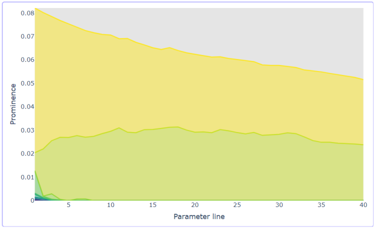


Flattening algorithm

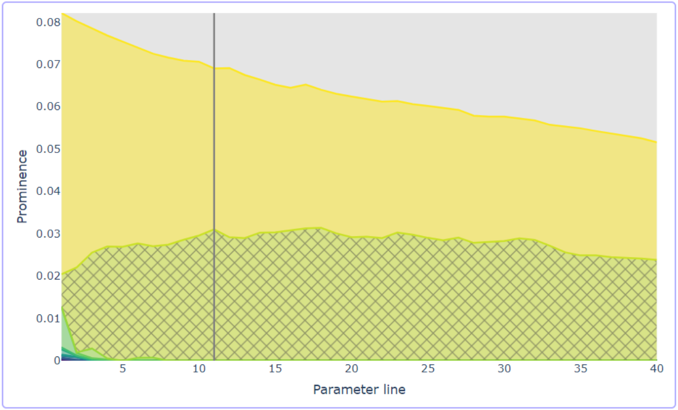


The vineyard





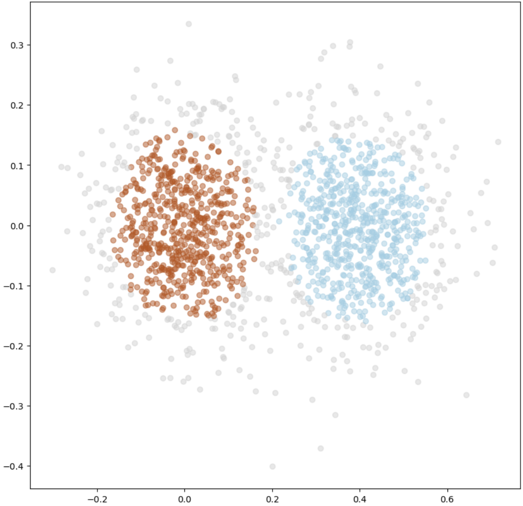
Parameter selection



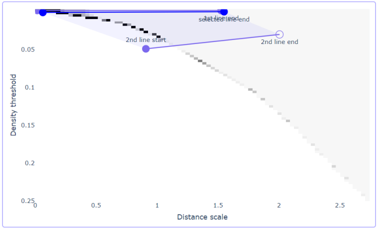
Parameter selection On / Off

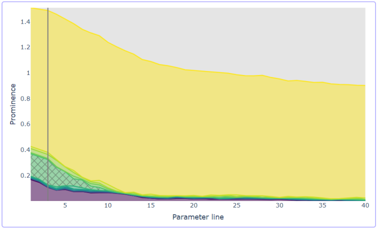
Line number 11 Gap number 2 Choose parameter

Result

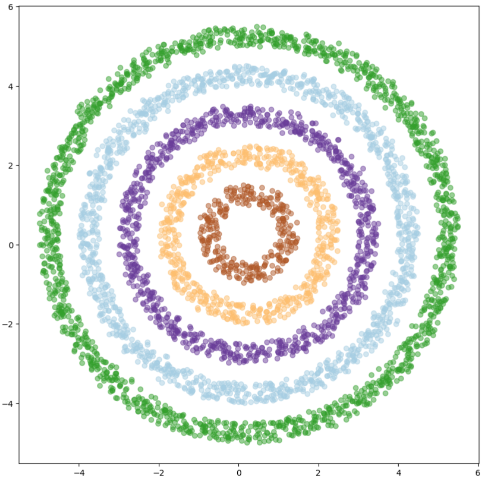


Applying Persistable

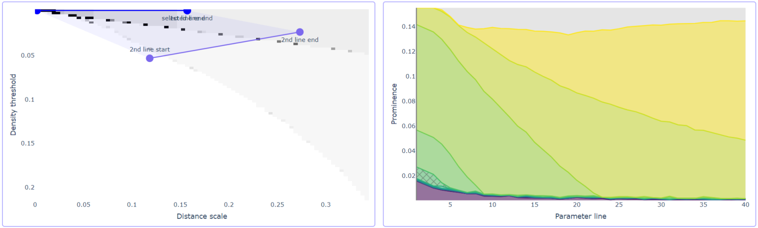




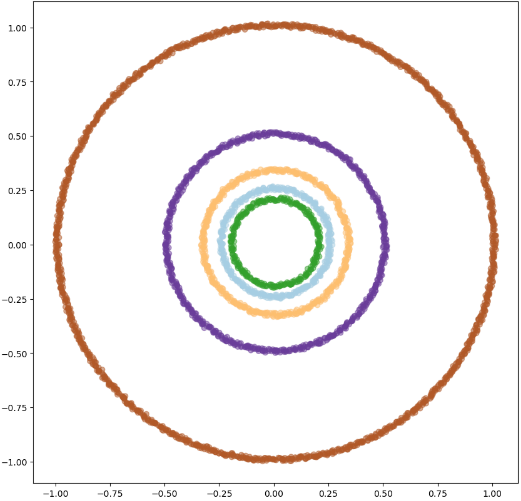
Applying Persistable



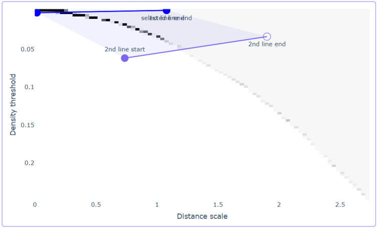
Applying Persistable

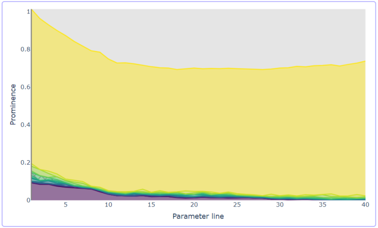


Applying Persistable

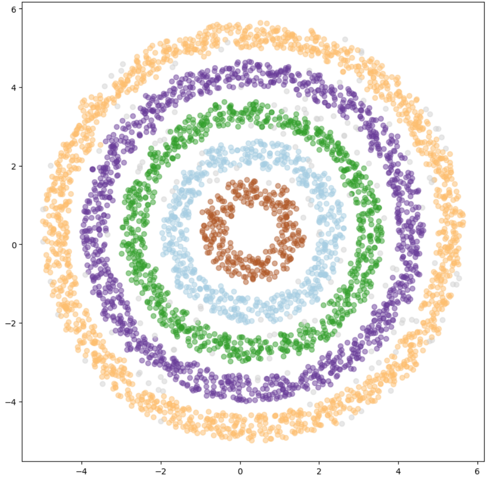


Applying Persistable

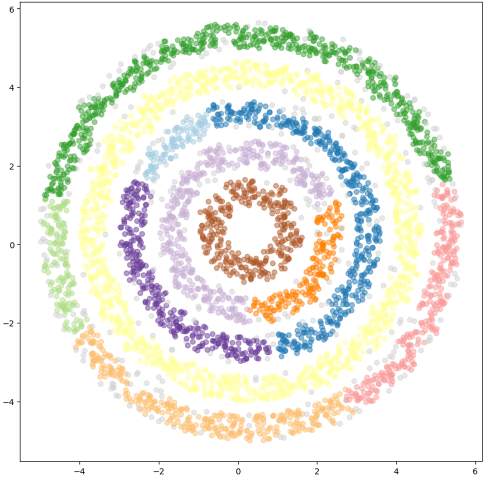




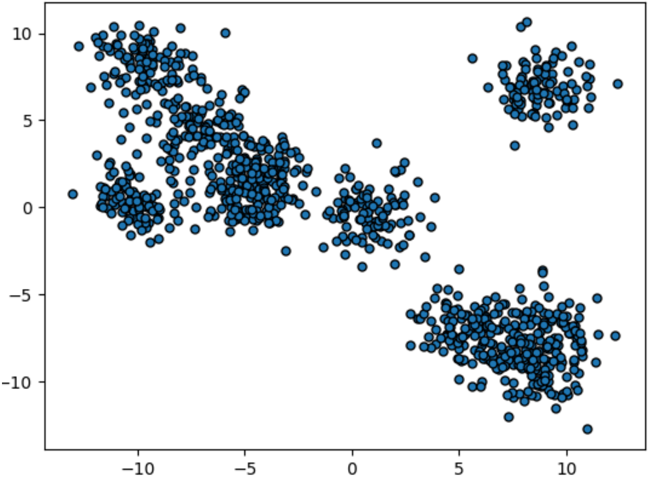
Applying Persistable



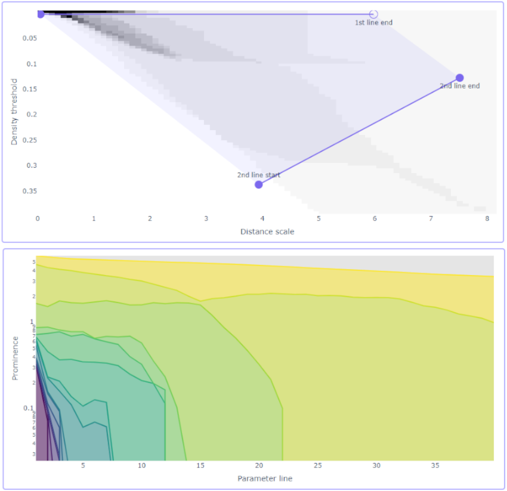
Applying Persistable



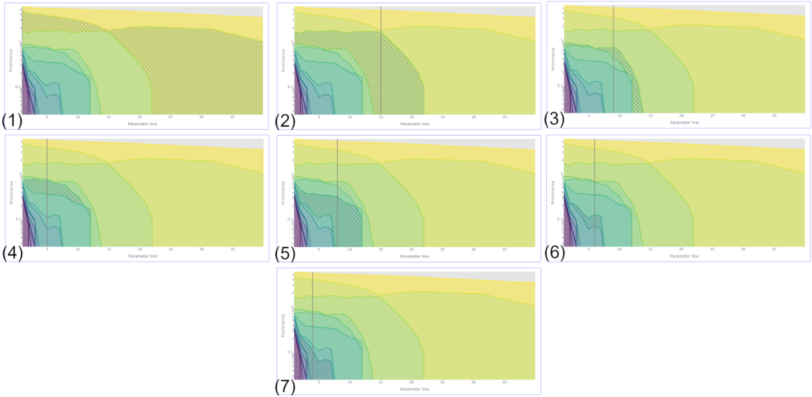
Applying Persistable



Applying Persistable



Applying Persistable



Applying Persistable