Sparsity 101: Statistical estimators Central location and dispersion

Laurent Duval

IFP Energies nouvelles

2013

Introduction

What is the trend? Where is the outlier?

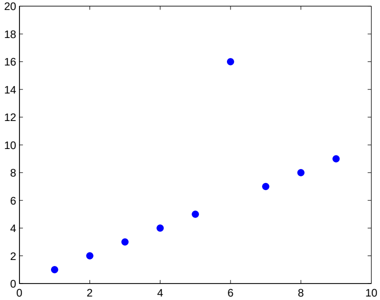

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

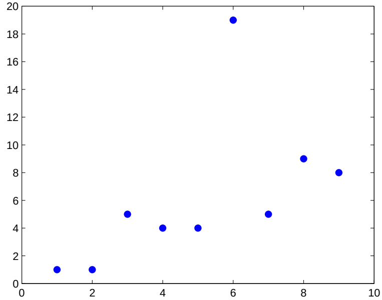

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

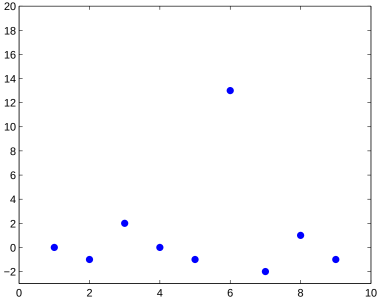

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

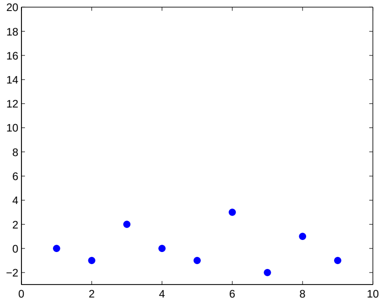

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

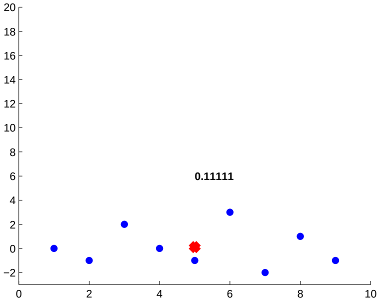

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

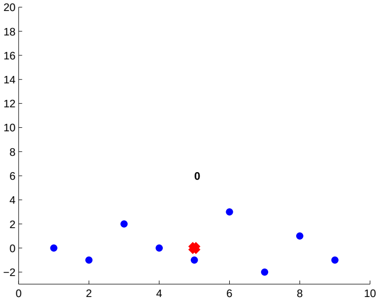

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

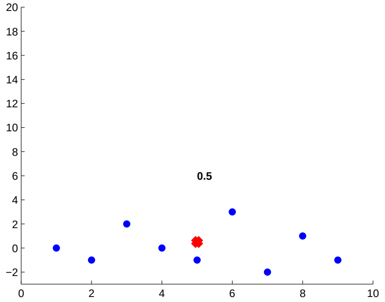

Figure : Toy noise problem

Introduction

What is the trend? Where is the outlier?

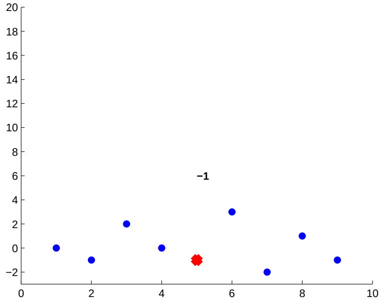

Figure : Toy noise problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N} \sum x_i$

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N} \sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$ ($\frac{N+1}{2}$ position)

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$ (and anything else in between?)

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum (x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum w_i(x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum (x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m}) \leftarrow \arg\min \sum |x_i - m|$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum (x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m}) \leftarrow \arg\min \sum w_i|x_i - m|$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i)$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum (x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m}) \leftarrow \arg\min \sum |x_i - m|$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i) \leftarrow \arg\min \max |x_i - m|$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i))$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \leftarrow \arg\min \sum (x_i - m)^2$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m}) \leftarrow \arg\min \sum |x_i - m|$
- *mid-range*: $\frac{1}{2}(\min x_i + \max x_i) \leftarrow \arg\min \max |x_i - m|$
- *mid-hinge*: $\frac{1}{2}(Q1(x_i) + Q3(x_i)) \leftarrow \arg\min \mathrm{med}|x_i - m|$
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem

Standard estimators

Compute a representative central location $\overline{m}$?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*
- *mid-hinge*
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*
- *mid-hinge*
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \;\; \to \sigma = \sqrt{\frac{1}{N}\sum(x_i - m)^2}$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$
- *mid-range*
- *mid-hinge*
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$ $\rightarrow \sigma = \sqrt{\frac{1}{N}\sum(x_i - m)^2}$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$ $\rightarrow \text{MAD} = \text{med}|x_i - m|$
- *mid-range*
- *mid-hinge*
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$ $\rightarrow \sigma = \sqrt{\frac{1}{N}\sum(x_i - m)^2}$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$ $\rightarrow \text{MAD} = \text{med}|x_i - m|$
- *mid-range* $\rightarrow \frac{1}{2}(\max x_i - \min x_i)$ (half-range, mid-span)
- *mid-hinge*
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i$ $\rightarrow \sigma = \sqrt{\frac{1}{N}\sum(x_i - m)^2}$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m})$ $\rightarrow \mathrm{MAD} = \mathrm{med}|x_i - m|$
- *mid-range* $\rightarrow \frac{1}{2}(\max x_i - \min x_i)$ (half-range, mid-span)
- *mid-hinge* $\rightarrow \frac{1}{2}(Q3(x_i) - Q1(x_i))$ (IQR/mid-spread)
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Standard estimators

Compute a representative dispersion/spread ?

- mean: $\overline{m} = \frac{1}{N}\sum x_i \quad \rightarrow \sigma = \sqrt{\frac{1}{N}\sum(x_i - m)^2}$
- median: $\#(x_i < \overline{m}) = \#(x_i > \overline{m}) \quad \rightarrow \mathrm{MAD} = \mathrm{med}|x_i - m|$
- *mid-range* $\rightarrow \frac{1}{2}(\max x_i - \min x_i)$ (half-range, mid-span)
- *mid-hinge* $\rightarrow \frac{1}{2}(Q3(x_i) - Q1(x_i))$ (IQR/mid-spread)
- mode: $\arg\max p(x)$
- even: $\min x_i$, $\max x_i$
- arbitrary algorithmic choice?
- no, answer to an optimization problem
- with a natural spread

Bonuses: add shape factor (skewness, kurtosis); $\frac{Q1+2\times\mathrm{med}+Q3}{4}\ldots$

Standard estimators: ℓ_2

Least-squares estimator of weighted central location:

$$f(m) = \sum w_i (x_i - m)^2$$

$$\frac{df}{dm} = \sum -2 w_i (x_i - m)$$

$$\frac{df}{dm} = 0 \Leftrightarrow \sum -2 w_i x_i = \sum -2 w_i m$$

$$\sum w_i x_i = \sum w_i m$$

$$\overline{m} = \frac{\sum w_i x_i}{\sum w_i}$$

Weighted sum location → all linear filters; Gaussian noise

Standard estimators: ℓ_1

Least-magnitude estimator of (weighted) central location:

$$f(m) = \sum w_i |x_i - m|$$

$$\frac{df}{dm} \approx \sum -w_i \mathrm{sign}(x_i - m)$$

$$\frac{df}{dm} = 0 \Leftrightarrow \#_{w_i}(x_i < \overline{m}) = \#_{w_i}(x_i > \overline{m})$$

Equality reached when $\overline{m}$ stands "in between"

$$\overline{m} = \mathrm{median}_{w_i} x_i$$

Sorted location $\rightarrow$ gen. weighted median filters; Laplace noise

Other (less standard) estimators

Importance of measurement units

- *harmonic*, geometric, arithmetico-geometric means
  100 km at 150 km h^{-1}, 100 km at 100 km h^{-1} $\rightarrow$ 120 km h^{-1}
- M-estimators, L-estimators, robust statistics
- no natural dispersion in general
- time vs individuals; 1D/2D; representative scale; transforms

A robust Gaussian noise dispersion estimator (details)

$$\hat{\sigma} \simeq \frac{\mathrm{median}|c_i|}{0.6745}$$

Use with wavelet shrinkage operators (soft , hard, garrote, etc.)

$$\hat{c}_i = S(c_i, \Lambda(\sigma))$$

Noise level estimation and wavelets

Figure : Noise estimation with standard 1D wavelets

Noise level estimation and wavelets

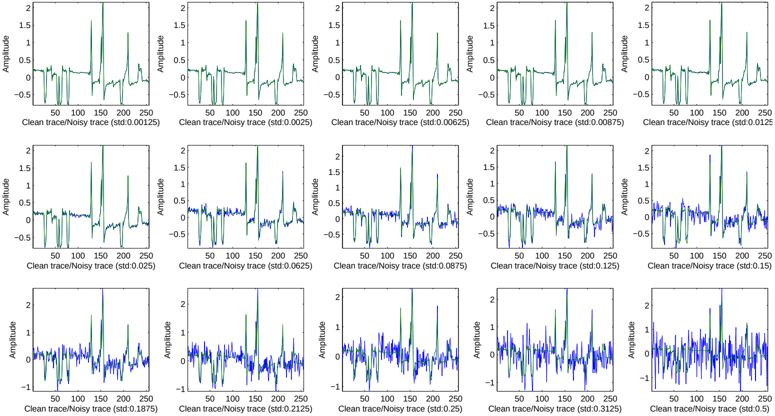

Figure : Noise estimation with standard 1D wavelets

Noise level estimation and wavelets

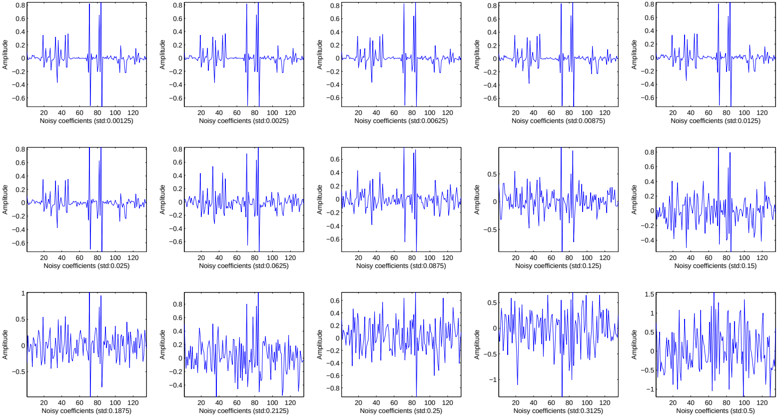

Figure : Noise estimation with standard 1D wavelets

Noise level estimation and wavelets

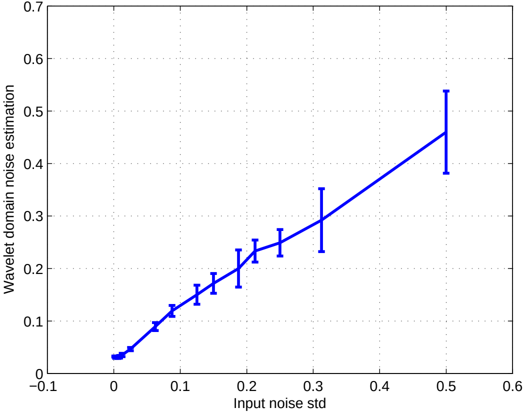

Figure : Noise estimation with standard 1D wavelets