Ranking Function Based Parameter Estimation

Benedek Bálint Novák

Supervisor:
Balázs Csanád Csáji

Eötvös Loránd Tudományegyetem

2025 Január 9.

The framework

Given:

- A family of probability distributions $\{\mathbb{P}_\vartheta | \vartheta \in \Theta\}$
  (Θ is a metric space)
- $(x_1, ..., x_n)$ i.i.d. sample from $\mathbb{P}_{\vartheta^*}$
- Black box B that can generate new sample given parameter ϑ

Black box types:

- 1. Generates an i.i.d. sample from $\mathbb{P}_\vartheta$
- 2. Generates a sample given $(\vartheta, q) \in \Theta \times [0, 1]^d$ which has distribution $\mathbb{P}_\vartheta$ if q is drawn from a uniform distribution, i.e. it is a function of ϑ and q that has distribution $\mathbb{P}_\vartheta$. (e.g. inverse of CDF)

Goal:

Approximate ϑ^*

The framework

The Resampling framework

- 1. Generate $m-1$ alternative samples $S^{(1)}, \ldots S^{(m-1)}$ (we denote the original sample with $S^{(0)}$) from $\mathbb{P}_{\vartheta}$
- 2. Assign a real number to each sample based on ϑ and its values called *reference variable*: $Z^{(i)}(\vartheta) := T(S^{(i)}(\vartheta), \vartheta)$
- 3. Rank the samples based on the reference variables:
- 4. Denote the *rank* of the original sample with $\mathcal{R}(\vartheta) \in \{1, \ldots, m\}$

Theorem

$\mathbb{P}(\vartheta^* \in \{\vartheta \in \Theta | \mathcal{R}(\vartheta) \leq q\}) = \frac{q}{m}$ if there is a strict ordering a.s.

Remark

If the reference variables are constructed in such a way that a lower value corresponds to a better fit, then $\operatorname*{argmin}_{\vartheta \in \Theta} \mathcal{R}(\vartheta)$ is a good approximation for ϑ^*.

Motivation

Advantages

- The framework can be used even if we don't know the density functions explicitly
- It also constructs a confidence region for the estimate
- The reference variables are customisable, they can even be black boxes

Reference Variable

Examples of Reference Variables

- ML based reference variable: $Z^{(i)}(\vartheta) = ||\nabla_\vartheta \mathcal{L}(\vartheta, S^{(i)})||^2$
- MMD based reference variable:
  $Z^{(i)}(\vartheta) = \widehat{\mathrm{MMD}}^2[S^{(i)}(\vartheta), S^{(m)}(\vartheta)]$

where $S^{(m)}$ denotes an extra sample and $\widehat{\mathrm{MMD}}^2$ is an unbiased estimator for the Maximum Mean Discrepancy of the two probability distributions.

Remark

The MMD is a customisable similarity measure of probability distributions. Note that MMD based reference variable doesn't require any knowledge about the distributions besides the samples.

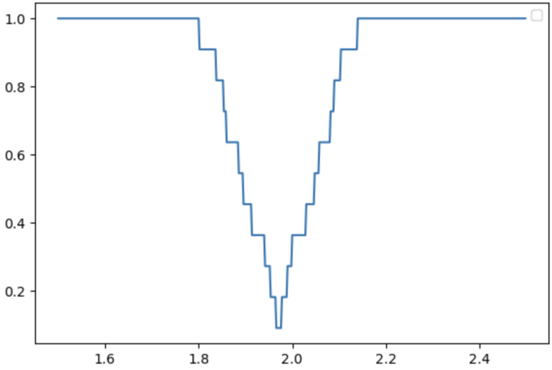
Parameter Estimation

Idea:

$$\widehat{\vartheta} \in \operatorname*{argmin}_{\vartheta \in \Theta} \mathcal{R}(\vartheta)$$

Problem:

Hard to optimize



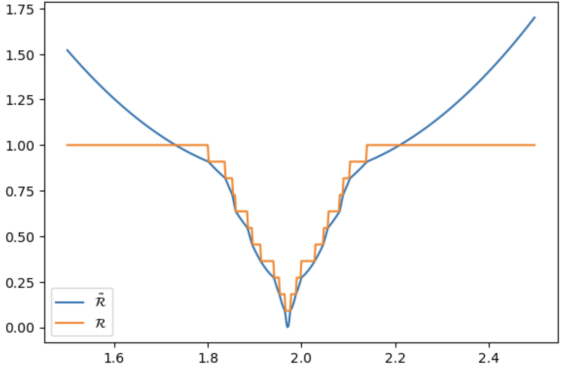
Parameter Estimation

Idea:

$$\widehat{\vartheta} \in \underset{\vartheta \in \Theta}{\operatorname{argmin}} \; \mathcal{R}(\vartheta)$$

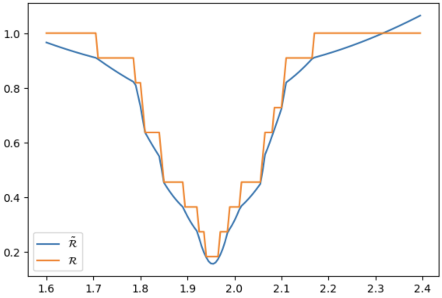
Solution:

Smoothed rank



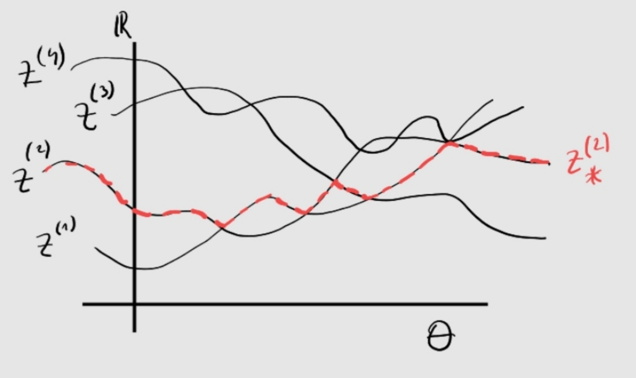
Smoothed rank

$$\tilde{\mathcal{R}}(\vartheta) = \begin{cases} \dfrac{Z^{(0)}}{Z_*^{(1)}} & \text{if } Z^{(0)} < Z_*^{(1)} \\ k + \dfrac{Z^{(0)} - Z_*^{(k)}}{Z_*^{(k+1)} - Z_*^{(k)}} & \text{if } Z_*^{(k)} \leq Z^{(0)} < Z_*^{(k+1)} \\ m - 1 + \dfrac{Z^{(0)}}{Z_*^{(m-1)}} & \text{if } Z_*^{(m-1)} \leq Z^{(0)} \end{cases}$$



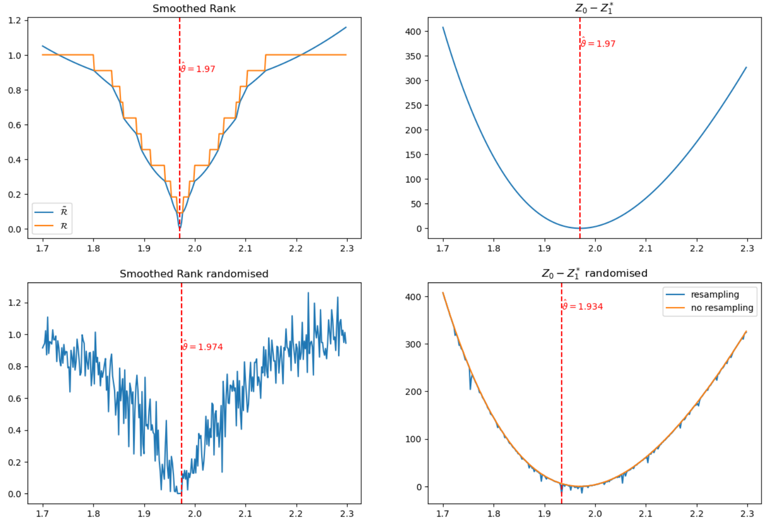
Continuity of the Smoothed rank

The continuity of the Smoothed rank follows from the continuity of $Z^{(k)}(\vartheta)$, because their pointwise ordered version $(Z_*^{(k)}(\vartheta))$ is also continuous.
Illustration for the proof idea of the continuity of $Z_*^{(k)}(\vartheta)$:



Other solutions

$n = 250, m = 10$



Asymptotic behaviour

Proposition

$\lim_{m\to\infty} \mathcal{R}(\vartheta) = F_{Z(\vartheta)}(Z^{(0)}(\vartheta))$ where $F_{Z(\vartheta)}$ denotes the CDF of $Z^{(i)}(\vartheta)$ for every $i \neq 0$

Proposition

If $\inf\left\{Z_*^{(1)}(\vartheta)\right\} = 0$, then

$$\lim_{m\to\infty}\left(Z^{(0)}(\vartheta) - Z_*^{(1)}(\vartheta)\right) = Z^{(0)}(\vartheta) - \inf\left\{Z_*^{(1)}(\vartheta)\right\} = Z^{(0)}(\vartheta)$$

Asymptotic behaviour

$n = 250, m = 1000$