Hypothesis test based estimation

Ádám Jung

Supervisor : Balázs Csanád Csáji

2024

ELTE | FACULTY OF SCIENCE

Introduction

Goal : Estimating the generating distribution of a sample within a parametric family.

Based on : Non-asymptotic hypothesis tests with exact type I. error probability.

Solution : Optimizing for parameters defining distributions least distinguishable from the true distribution.

Reproducing Kernel Hilbert Spaces (RKHS)

- Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$.
- Positive definite kernel $k(\cdot,\ \cdot)$: symmetric, satisfies

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0 \qquad \forall x_i \in \mathcal{X}, \forall a_i \in \mathbb{R}$$

- Maps points $x \in \mathcal{X}$ to $\mathcal{H}$ via $x \mapsto k(\cdot, x)$.
- Reproducing property :

$$\langle h, k(\cdot, x) \rangle_{\mathcal{H}} = h(x) \qquad \forall h \in \mathcal{H}, \forall x \in \mathcal{X}$$

Kernel Mean Embedding (KME)

- KME of distribution Q : $\mu_Q = \int_{\mathcal{X}} k(\cdot, x) Q(dx)$.
- As a consequence of the reproducing property

$$\mathbb{E}[h(X)] = \langle h, \mu_Q \rangle_{\mathcal{H}} \quad \forall h \in \mathcal{H}; \quad \text{where } X \sim Q$$

- Maximum Mean Discrepancy (MMD) : $\mathrm{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$.
- By the reproducing property :

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}[k(X, X')] - 2\mathbb{E}[k(X, Y)] + \mathbb{E}[k(Y, Y')]$$

KME with Riesz kernel

- Consider $\mathcal{X} = \mathbb{R}^d$ and the *Riesz* (or energy) kernel :

$$k(u, v) = -\|u - v\|^r \quad \text{with} \quad r \in (0, 2).$$

- In the case of $d = r = 1$ there are closed form solutions for $\mathrm{MMD}^2(P_\theta, \hat{Q})$, where
  - $\hat{Q} = \frac{1}{m} \sum_{i=1}^{m} \delta_{y_i}$
  - P_θ : a commonly used parametric distr. family (e.g., normal, mixture of normal, etc.)

Hypothesis Tests

- Parametric family $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$.
- i.i.d. sample $\{y_1, \dots, y_n\}$ from $Q := P_{\theta^*}$
- Construct hypothesis tests for

$$H_0 : P_\theta = Q$$
$$H_1 : P_\theta \neq Q$$

Resampling framework

- Let $\mathcal{S}^{(0)}$ denote the original sample $\{y_1, \dots, y_n\}$.
- Generate $m-1$ alternative set of samples $\mathcal{S}^{(j)}(\theta)$ from P_θ :

$$\mathcal{S}^{(j)}(\theta) = \{y_1^{(j)}, \dots, y_n^{(j)}\} \qquad j = 1, \dots, m-1$$

- Observe that $\theta = \theta^* \implies \mathcal{S}^{(0)}, \dots, \mathcal{S}^{(m-1)}$ are exchangeable.
- Let $\mathcal{R}(\theta)$ denote a *ranking function*, defining the rank of $\mathcal{S}^{(0)}$ among $\mathcal{S}^{(1)}(\theta), \dots, \mathcal{S}^{(m-1)}(\theta)$.

Theorem

For any ranking function $\mathcal{R}$ and parameters p, q, m, we have

$$\mathbb{P}(\theta^* \in \tilde{\Theta}) = \frac{q-p+1}{m}, \text{ where } \tilde{\Theta} = \{\theta \in \Theta \mid p \le \mathcal{R}(\theta) \le q\}.$$

Let $\mathcal{R}(\theta) = 1 + \sum_{i=1}^{m-1} \mathbb{1}\left\{\|\nabla_\theta \ell(\theta, \mathcal{S}^{(i)})\|^2 < \|\nabla_\theta \ell(\theta, \mathcal{S}^{(0)})\|^2\right\}$.

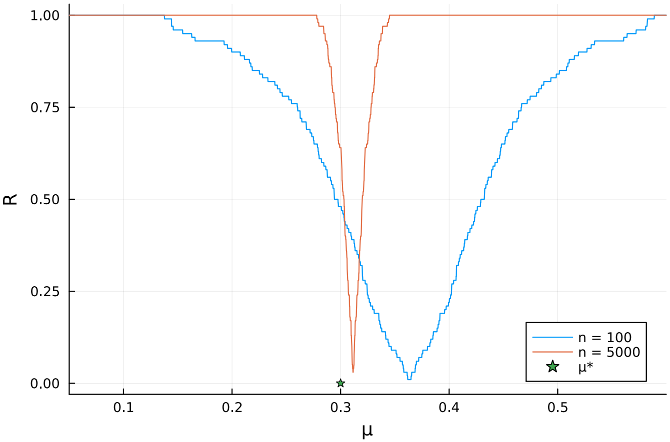

Figure – Normalized rank statistics for $P_\theta = \mathcal{N}(\mu, 1)$, $(m = 100)$.

Let $\mathcal{R}(\theta) = 1 + \sum_{i=1}^{m-1} \mathbb{1}\left\{\mathrm{MMD}^2(P_\theta, \mathcal{S}^{(i)}) < \mathrm{MMD}^2(P_\theta, \mathcal{S}^{(0)})\right\}$

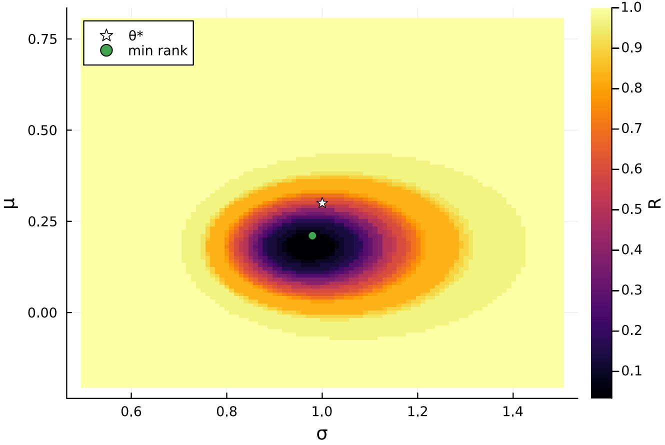

Figure – Normalized rank statistics for $P_\theta = \mathcal{N}(\mu, \sigma^2)$, $(m = 30)$.

Optimization

- Point estimate : $\widehat{\theta} \in \arg\min_{\theta \in \Theta} \mathcal{R}(\theta)$.
- Difficulties :
  i) $\mathcal{R}$ is a piece-wise constant function
  ii) Completely flattens out for parameter values far from θ^*
- Possible solutions :
  i) Use gradient-free optimization methods (e.g., Nelder-Mead)
  ii) Start with a small subset of the observations, and find $\widehat{\theta}$ iteratively for larger sample sizes

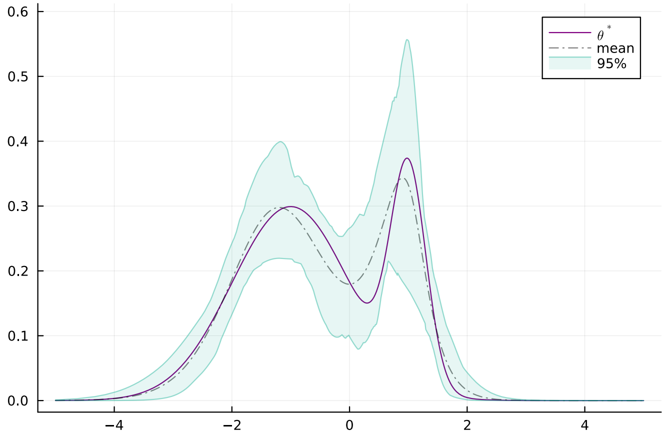

Figure – The result of 50 repeated estimations of a mixture $0.25 \cdot \mathcal{N}(1, 0.3) + 0.75 \cdot \mathcal{N}(-1, 1)$ with $\mathcal{R}$ being the MMD based construction. ($n = 60$, $m = 3000$).

Thank you for listening