Hypothesis test based estimation

Ádám Jung

Supervisor: Balázs Csanád Csáji

May 2024

1 Introduction

In this document we will present a novel method for estimating the generating distribution of a sample.

Candidate distributions are proposed from a parametric family of distributions and then using non-asymptotic hypothesis tests with exact type I. error for distribution fitting, we aim to optimise for parameters which define distributions least distinguishable from the true distribution of the sample.

2 Reproducing Kernel Hilbert Spaces

Let $\mathcal{X}$ be an arbitrary set and let $\mathcal{H}$ be a Hilbert space of $\mathcal{X} \to \mathbb{R}$ type functions with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. If for all $x \in \mathcal{X}$ the $\delta_x : \mathcal{H} \to \mathbb{R}$ evaluation functional (mapping $h \in \mathcal{H}$ to $h(x)$) is continuous, then $\mathcal{H}$ is called a *Reproducing Kernel Hilbert Space* (RKHS).

A $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ function is called a *positive definite kernel*, if it is symmetric in its arguments and for all $n \in \mathbb{N}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 \quad \forall a_i \in \mathbb{R}, x_i \in \mathcal{X}, i \in [n] \tag{1}$$

holds. If $\mathcal{H}$ is an RKHS then it has the so called *reproducing property* i.e. there exists a unique pd kernel k, satisfying

$$\langle h, k(\cdot, x) \rangle_{\mathcal{H}} = h(x) \quad \forall x \in \mathcal{X};\ \forall h \in \mathcal{H}. \tag{2}$$

k is called the *reproduciong kernel* of $\mathcal{H}$ and in fact $k(u, v)$ is nothing else but the Riesz representer of δ_x evaluated at u.

By the Moore-Arnoszjan theorem there is a one to one correspondence between positive definite kernels and RKSHs, meaning that given a pd kernel one can uniquely construct an RKHS $\mathcal{H}_k$ with k being its reproducing kernel.

2.1 Kernel Mean Embedding

Given $\mathcal{X}$ and k it is possible to map $x \in \mathcal{X}$ points to the (possibly) infinite dimensional function space $\mathcal{H}_k$ by the map $x \mapsto k(\cdot, x)$.

There is a similar way to define a map from probability distributions defined on $\mathcal{X}$ to $\mathcal{H}$ by taking the expected value of the mapped values. More formally if $M_+^1(\mathcal{X})$ denotes the probability distributions defined on $\mathcal{X}$, then the so called *kernel mean embedding* (KME) of $Q \in M_+^1(\mathcal{X})$ is defined as

$$\mu_Q := \int_{\mathcal{X}} k(\cdot, x) Q(dx). \tag{3}$$

If $\mathbb{E}[\sqrt{k(X, X)}] < \infty$ holds for $X \sim Q$, then we have[3]

$$\langle h, \mu_Q \rangle = \mathbb{E}[h(X)] \quad \forall h \in \mathcal{H}, \tag{4}$$

which can be seen as a reproducing property of the expectation operation.

If the $\mu : M_+^1 \to \mathcal{H}$ kernel mean map is injective, then the associated kernel is called *characteristic*. In this case

$$||\mu_P - \mu_Q||_{\mathcal{H}}^2 = 0 \iff P = Q \tag{5}$$

and therefore the so called *Maximum Mean Discrepany* $\text{MMD}^2(P, Q) := ||\mu_P - \mu_Q||_{\mathcal{H}}^2$ can be used to compare distributions for equality.

Using the reproducing property (4) the MMD distance of distributions P and Q can be expressed as

$$\text{MMD}^2(P, Q) = \mathbb{E}[k(X, X')] - 2\mathbb{E}[k(X, Y)] + \mathbb{E}[k(Y, Y')], \tag{6}$$

where X', Y' are independent copies of X, Y with distributions P, Q respectively.

If we only have an empirical estimate of Q based on i.i.d. observations $x_1, \ldots, x_n$; $x_i \sim Q$ then the empirical counterpart of (3) is

$$\hat{\mu}_Q = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, x_i). \tag{7}$$

For the MMD of two empirical distributions based on i.i.d. samples $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^m$ from P and Q respectively, an unbiased estimate for (6) is

$$\widehat{\mathrm{MMD}}^2(P,Q) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i,x_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i,y_j). \quad (8)$$

Depending on the choice of kernel, if $P = P_\theta$ is a parametric distribution and $\hat{Q}$ is an empirical distribution based on samples, the expectations defining $\mathrm{MMD}^2(P_\theta, \hat{Q})$ might be expressed in closed form, as a function of the parameters.

Riesz kernel Lets consider the case, when $\mathcal{X} = \mathbb{R}^d$ and we use a special case of distance based kernels[4] the so called *Riesz* (aka. energy) kernel :

$$k(u,v) = -||u-v||^r \quad r \in (0,2). \quad (9)$$

Since k is only *conditionally* positive definite (ie. (1) only holds with the additional assumption $\sum_i a_i = 0$), it is needed to define a center point $x_0 \in \mathbb{R}^d$, and only the the modified version

$$\tilde{k}(u,v) := -||u-v||^r + ||u-x_0||^r + ||v-x_0||^r$$

defines a positive definite kernel. However since we are only interested in computing the MMD distance of distributions and the terms in which x_0 is present are cancelled out in (6), it is justified to work with k instead of $\tilde{k}$ as they provide the exact same results.

In the case of $d = r = 1$ there are closed form solutions for $\mathrm{MMD}^2(P_\theta, \hat{Q})$ for many commonly used parametric distributions.[2] For example if $P_\theta = \mathcal{N}(\mu, \sigma^2)$, and $\hat{Q} = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}$ (where δ_y denotes the measure of a point mass at y) then

$$\mathrm{MMD}^2(P_\theta, \hat{Q}) = \frac{4}{m}\sum_{i=1}^{m}(y_i-\mu)\left[\varphi\left(\frac{y-\mu}{\sigma}\right)-1\right] + 2\sigma^2\Phi\left(\frac{y-\mu}{\sigma}\right) - \frac{2\sigma}{\sqrt{\pi}} - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|y_i-y_j|, \quad (10)$$

where φ and Φ are the density and cumulative distribution function of $\mathcal{N}(0,1)$ respectively.

Remark 1 *Equation (10) is derived based on literature about* proper scoring rules *(PSR), which is a closely related concept to MMD.*[5] *In the PSR literature there are also available closed form expressions for other parametric distribution families (e.g. mixture of normal).*

3 Hypothesis Tests

Based on the resampling framework described in[1] for binary classification, we will present a similar method for constructing exact hypothesis tests for distribution fitting.

Suppose we have a parametric family of probability distributions

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}, \quad (11)$$

and an i.i.d. sample $\{y_1, \dots, y_n\}$ from $Q_{\theta^*} \in \mathcal{P}$. In this section our goal is to construct hypothesis tests for

$$H_0 : P_\theta = Q_{\theta^*}$$
$$H_1 : P_\theta \neq Q_{\theta^*},$$

with non-asymptotic guarantees for exact probability of the type I. error.

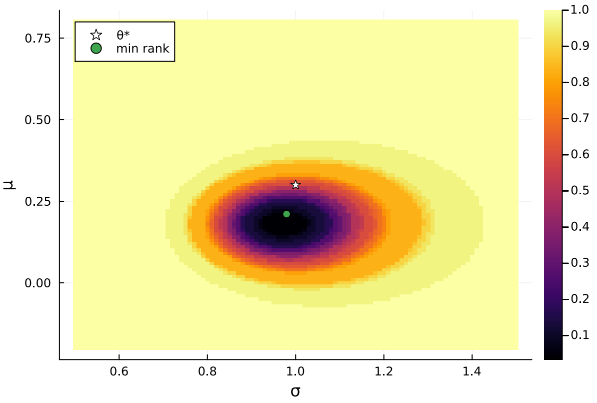

Figure 1: Normalized rank statistics for a two parameter normal distibution $\mathcal{N}(\mu, \sigma^2)$. The ranking function was the second construction of section 5, with $m = 30$ generated samples, each with size $n = 100$.

The algorithm is based on generating $m-1$ alternative set of i.i.d. samples given a candidate parameter θ, which are denoted as

$$\mathcal{S}^{(j)}(\theta) := \{y_1^{(j)}, \dots, y_n^{(j)}\}; \qquad y_1^{(j)} \sim P_\theta \quad (12)$$

for $j = 1, \dots, m-1$. For brevity we denote the original sample $\{y_1, \dots, y_n\}$ with $\mathcal{S}^{(0)} := \{y_1^{(0)}, \dots, y_n^{(0)}\}$.

A function $\psi : \mathbb{A}^m \to [m]$ is called a *ranking function*, if for all $a_1, \dots, a_m \in \mathbb{A}$ it is is invariant for reordering its last $m-1$ arguments, and for all $i \neq j$ we have

$$\psi(a_i, \{a_k\}_{k\neq i}) \neq \psi(a_j, \{a_k\}_{k\neq j}).$$

From *Lemma 1* of[1] if $A_1, \dots, A_m$ are exchangeable (a.s.) pairwise different random elements from $\mathbb{A}$, then $\psi(A_1, \dots, A_m)$ has uniform distribution on $\{1, \dots, m\}$.

Given a ranking function ψ and the $\mathcal{S}^{(j)}(\theta)$, $j \in [m]$ samples, lets define a *confidence region* for θ^* as

$$\tilde{\Theta} := \{\theta \in \Theta \mid p \leq \psi(\mathcal{S}^{(0)}, \{\mathcal{S}^{(k)}(\theta)\}_{k\neq 0}) \leq q\}. \quad (13)$$

By *Theorem 1* of[1] for all ψ ranking functions we have

$$\mathbb{P}(\theta^* \in \tilde{\theta}) = \frac{q-p+1}{m}, \quad (14)$$

meaning that with an appropriate choice of p, q and m the type I. error of the test can be exactly controlled.

A method is called *consistent*, if for all $\theta \neq \theta^*$

$$\mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \{\theta \in \tilde{\Theta}_n\}\right) = 0 \quad (15)$$

holds, $\tilde{\Theta}_n$ denoting the confidence region based on a sample of size n.

4 Optimisation

In this section we propose a general method for constructing point estimates from the hypothesis tests of section 3.

Let $\mathcal{R}(\theta)$ denote the *rank* of the original sample among the generated samples, i.e.

$$\mathcal{R}(\theta) := \psi(\mathcal{S}^{(0)}, \{\mathcal{S}^{(k)}(\theta)\}_{k\neq 0}). \quad (16)$$

As we will see in section 5 the proposed ranking functions are defined in a way, that for a false $\theta \neq \theta^*$ parameter the rank of $\mathcal{S}^{(0)}$ tends to be the largest. Therefore we set $p = 1$ and we are aiming to find parameters, for which $\mathcal{R}(\theta)$ is small, since these are the ones we can only reject with high probability of type I. error.

Lets define the point estimate as a parameter that minimizes the rank statistic

$$\hat{\theta} := \min_{\theta\in\Theta} \mathcal{R}(\theta). \quad (17)$$

Solving this minimisation can be generally a hard problem, since $\mathcal{R}(\theta)$ is a piece-wise constant function which completely flattens out for parameter values that are fare from θ^* (see figure 2, and 1).

In general (17) can be solved with gradient-free optimisation methods such as the *Nelder–Mead method*, but it is necessary to have a good θ_0 initial guess to successfully start the minimization algorithm.

Remark 2 *Quite counter-intuitively for smaller sample sizes it is easier to start the optimization (17), since the bigger uncertainty about the parameter results in much larger confidence regions, therefore it is more easy to find a sufficient initial θ_0. (see fig. 2)*

A possible approach for finding θ_0 for large sample sizes is to start with only a subset of the observations, find $\hat{\theta}$ and then use it as θ_0 for a larger subset of the observations, solving (17) iteratively.

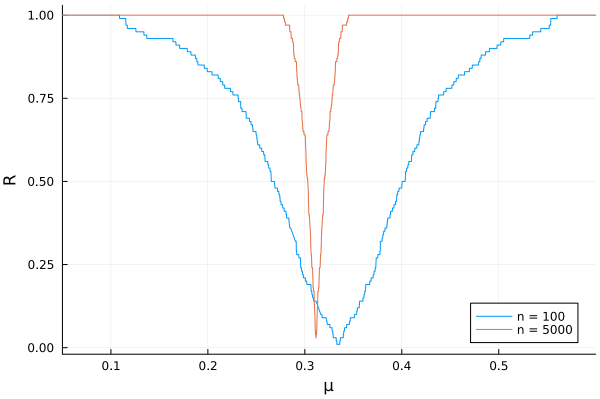

Figure 2: Normalized rank statistics for estimating the expected value $\mu \in \mathbb{R}$ of a normal distribution $\mathcal{N}(\mu, 1)$. There was $m = 100$ generated samples, and the ranking function was the first construction of section 5.

5 Experiments

In this section we present two constructions for the ranking function ψ.

i) The first approach is based on the maximum likelihood equation, i.e. let $l(\theta; y_1, \dots, y_n)$ be the log-likelihood function of P_θ, and lets define *reference variables* $Z^{(i)}(\theta)$ for $i = 0, \dots, m-1$ as

$$Z^{(i)}(\theta) := ||\nabla_\theta l(\theta, y_1^{(i)}, \dots, y_n^{(i)})||^2, \quad (18)$$

and let

$$\mathcal{R}(\theta) = 1 + \sum_{i=1}^{m-1} \mathbb{I}(Z^{(i)}(\theta) < Z^{(0)}). \quad (19)$$

ii) The second construction is based on the MMD distance of P_θ and $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}$ using the Riesz kernel. The reference variables are defined as

$$Z^{(i)}(\theta) = \mathrm{MMD}^2(P_\theta, \hat{Q}_0), \tag{20}$$

and $\mathcal{R}(\theta)$ is constructed the same way as in (19).

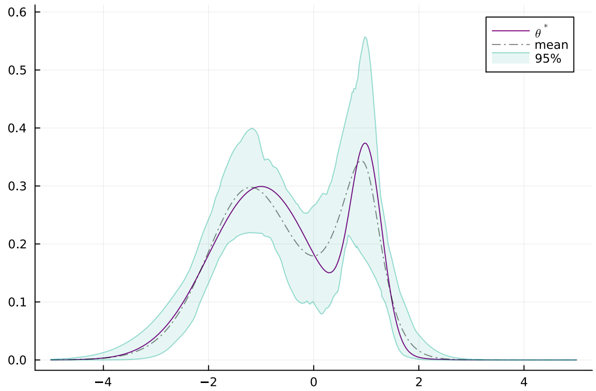

Figure 3: The result of 50 repeated estimations of a mixture model of $0.25 \cdot \mathcal{N}(1, 0.3) + 0.75 \cdot \mathcal{N}(-1, 1)$ with the second construction of section 5, ($n = 60$, $m = 3000$).

References

[1] Balázs Csanád Csáji and Ambrus Tamás. Semiparametric uncertainty bounds for binary classification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 4427–4432, 2019.

[2] Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules, 2018.

[3] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017.

[4] Dino Sejdinovic, Arthur Gretton, Bharath Sriperumbudur, and Kenji Fukumizu. Hypothesis testing using pairwise distances and associated kernels (with appendix), 2012.

[5] Erik Zawadzki and Sebastien Lahaie. Nonparametric scoring rules. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Mar. 2015.