KSD Aggregated Goodness-of-fit Test

Antonin Schrab, Benjamin Guedj, Arthur Gretton

University College London
Centre for Artificial Intelligence
Gatsby Computational Neuroscience Unit
Inria London

Goodness-of-fit problem

Problem: Having access to

- the score function $\nabla \log p(\cdot)$ of a model with density p on $\mathbb{R}^d$,
- some samples $\mathbb{Z}_N \coloneqq (Z_1, \dots, Z_N)$ where $Z_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$,

can we decide whether the densities p and q are the same?

Hypothesis testing:

$$\mathcal{H}_0 \colon p = q \qquad \text{against} \qquad \mathcal{H}_a \colon p \neq q$$

Type I error: controlled by α by design

$$\mathbb{P}(\text{reject } \mathcal{H}_0 \mid \mathcal{H}_0 \text{ is true}) \leq \alpha$$

Type II error: can be controlled by β for $\|p - q\|_2$ *'large enough'*

$$\mathbb{P}(\text{fail to reject } \mathcal{H}_0 \mid \mathcal{H}_0 \text{ is false}) \leq \beta$$

Kernel Stein Discrepancy Test for fixed kernel

Chwialkowski et al. **A kernel test of goodness of fit**. ICML 2016.

Liu et al. **A kernelized Stein discrepancy for goodness-of-fit tests**. ICML 2016.

Stein kernel: $h_{p,k}(x,y)$ depending on score $\nabla \log p(\cdot)$ and kernel k

Stein identity: $\mathbb{E}_p[h_{p,k}(Z,\cdot)] = 0$

KSD: $\mathrm{KSD}^2_{p,k}(q) := \mathbb{E}_{q,q}[h_{p,k}(Z,Z')]$

Estimator: $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) := \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} h_{p,k}(Z_i, Z_j)$

Quantile: $\widehat{q}^{\,k}_{1-\alpha}$ estimated with either parametric or wild bootstrap

Test: reject $\mathcal{H}_0 : p = q$ if $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) > \widehat{q}^{\,k}_{1-\alpha}$

Time complexity: $\mathcal{O}\left(BN^2\right)$

KSDAgg: KSD Aggregated test

Multiple testing: finite collection K of kernels

KSDAgg (KSD Aggregated test):

reject $\mathcal{H}_0\colon p = q$ if $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) > \widehat{q}^{\,k}_{1-u_\alpha w_k}$ for some $k \in K$

Weights: $(w_k)_{k\in K}$ satisfying $\sum_{k\in K} w_k \leq 1$
e.g. uniform weights: $w_k = 1/|K|$

Correction: u_α defined to control type I error and maximize power

$$\sup\left\{ u > 0 : \mathbb{P}_{\mathcal{H}_0}\left(\max_{k\in K} \left(\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) - \widehat{q}^{\,k}_{1-u w_k} \right) > 0 \right) \leq \alpha \right\}$$

- more powerful than Bonferroni correction as $u_\alpha \geq \alpha$
- power does not necessarily decrease to 0 as $|K|$ increases

Time complexity: $\mathcal{O}\big(|K|\,(B_1 + B_2)\,N^2\big)$

KSDAgg: power guarantees

Theorem (no regularity assumption)

KSDAgg type II error $\mathbb{P}(\textit{fail to reject } \mathcal{H}_0 \mid \mathcal{H}_0 \textit{ is false}) \leq \beta$ *if*

$$\|p-q\|_2^2 \;\geq\; \min_{k\in K}\left(\left\|(p-q) - T_{h_{p,k}}(p-q)\right\|_2^2 + C\log\left(\frac{1}{\alpha w_k}\right)\frac{\sqrt{C_k}}{\beta N}\right)$$

- **Integral transform:** $(T_\kappa f)(y) \coloneqq \int_{\mathbb{R}^d} \kappa(x,y) f(x)\,\mathrm{d}x$
- **Assumption:** $C_k \coloneqq \mathbb{E}_{q,q}\left[h_{p,k}(Z,Z')^2\right] < \infty$

Theorem (restricted Sobolev regularity assumption)

Under restrictive assumptions, if $p-q$ *lies in a restricted Sobolev ball, then the type II error for KSDAgg is controlled by* β *when*

$$\|p-q\|_2^2 \;\geq\; C\left(\frac{N}{\log\log N}\right)^{-4s/(4s+d)}$$

KSDAgg: KSD Aggregated Goodness-of-fit Test

- **Kernel selection:** adaptive test which aggregates over a collection of kernels using a powerful multiple testing correction

- **Non-asymptotic level and power guarantees**

- **Experiments:** KSDAgg outperforms KSD test with median bandwidth & KSD test with bandwidth selected on split data. KSDAgg matches the power of the KSD test with bandwidth selected using extra data.

- **Linear-time variant of KSDAgg:**
  Schrab A., Kim I., Guedj B., Gretton A. **Efficient Aggregated Kernel Tests using Incomplete U-statitics**. NeurIPS 2022.