KSD Aggregated Goodness-of-fit Tests
KSDAgg & KSDAggInc

Antonin Schrab

University College London

Centre for Artificial Intelligence

Gatsby Computational Neuroscience Unit

Inria London

a.schrab@ucl.ac.uk antoninschrab.github.io

KSD Aggregated Goodness-of-fit Test

Antonin Schrab
† ‡ §

Benjamin Guedj
† §

Arthur Gretton
‡

† Centre for Artificial Intelligence, UCL
‡ Gatsby Computational Neuroscience Unit, UCL
§ Inria London

Efficient Aggregated Kernel Tests using Incomplete *U*-statistics

Antonin Schrab[† ‡ §]

Ilmun Kim[*]

Benjamin Guedj[†§]

Arthur Gretton[‡]

† Centre for Artificial Intelligence, UCL

‡ Gatsby Computational Neuroscience Unit, UCL

§ Inria London

* Department of Statistics & Data Science, Yonsei University

Kernel Stein Discrepancy

Goodness-of-fit problem:

- model with probability density p or score function $\nabla \log p(\cdot)$ on $\mathbb{R}^d$
- samples $\mathbb{Z}_N := (Z_1, \dots, Z_N)$, $Z_i \overset{\text{iid}}{\sim} q$ in $\mathbb{R}^d$

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

Stein kernel: $h_{p,k}(x, y)$ defined in terms of $\nabla \log p(\cdot)$ and kernel k

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

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

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

Incomplete *U*-statistic: $\overline{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) := \frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} h_{p,k}(Z_i, Z_j)$

KSD Goodness-of-fit Tests

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

- quantile $\widehat{q}^{\,k}_{1-\alpha}$ estimated using wild/parametric bootstrap
- complexity $\mathcal{O}\big(B\,N^2\big)$

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

- positive weights $(w_k)_{k\in\mathcal{K}}$ summing to 1
- correction u_α maximizes power while retaining well-calibrate level α
- complexity $\mathcal{O}\big(|\mathcal{K}|\,B\,N^2\big)$

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

- complexity $\mathcal{O}\big(|\mathcal{K}|\,B\,|\mathcal{D}|\big)$
- linear-time test for the choice $|\mathcal{D}| = cN$ for some small $c \in \mathbb{N}$

Come check out my poster for details and discussions!

- **Power guarantees:** upper bound on uniform separation rates
- **Trade-off:** computational efficiency versus rate of convergence

KSD Aggregated Goodness-of-fit Tests

Antonin Schrab I. Kim B. Guedj A. Gretton

UCL CENTRE FOR ARTIFICIAL INTELLIGENCE GATSBY

KSDAgg: KSD Aggregated Goodness-of-fit Test

KSDAggInc: Efficient Aggregated Kernel Tests using Incomplete U-statistics

Contributions

- Aggregate KSD tests with different kernels or bandwidths
- Quantiles estimated via wild or parametric bootstraps
- No data splitting (known to result in a loss in power)
- Uniform separation rate upper bound for general kernels
- Propose efficient tests based on incomplete U-statistics
- Quantify trade-off efficiency versus rate of convergence

Goodness-of-fit problem

Are samples drawn from the model?

- model density p (or score function $\nabla \log p(z)$)
- samples $\mathbb{Z}_N := (Z_1, \dots, Z_N)$ drawn $Z_i \overset{\text{iid}}{\sim} q$

Hypothesis testing:

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

Kernel Stein Discrepancy

Stein kernel: $h_{p,k}(x, y)$ in terms of $\nabla \log p(z)$ with kernel k

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

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

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

KSD test for fixed kernel k

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

Quantile: $\widehat{q}^{\,k}_{1-\alpha}$ is $\lceil B(1-\alpha) \rceil$-th largest bootstrapped value

Wild bootstrap: $\dfrac{1}{N(N-1)} \displaystyle\sum_{1 \leq i \neq j \leq N} \varepsilon_i \varepsilon_j h_{p,k}(Z_i, Z_j)$, $\varepsilon_i \overset{\text{iid}}{\sim} \{\pm 1\}$

Parametric bootstrap: $\dfrac{1}{N(N-1)} \displaystyle\sum_{1 \leq i \neq j \leq N} h_{p,k}(\widetilde{Z}_i, \widetilde{Z}_j)$, $\widetilde{Z}_i \overset{\text{iid}}{\sim} p$

KSDAgg for collection of kernels $\mathcal{K}$

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

Weights (prior): $(w_k)_{k \in \mathcal{K}}$ satisfying $\sum_{k \in \mathcal{K}} w_k \leq 1$

Correction: u_α maximum value such that the level estimated via Monte-Carlo is well-calibrated at α

More powerful than conservative Bonferroni correction

KSDAgg Uniform separation rate

Integral transform: $(T_\kappa f)(y) := \displaystyle\int_{\mathbb{R}^d} \kappa(x, y) f(x) \,\mathrm{d}x$

Kernel assumption: $A_k := \mathbb{E}_{q,q}[h_{p,k}(Z, Z')^2] < \infty$

If $\|p - q\|_2^2$ is greater than

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

then **KSDAgg** has power at least $1 - \beta$.

Incomplete U-statistic

Estimator: $\overline{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N) := \dfrac{1}{N(N-1)} \displaystyle\sum_{(i,j) \in \mathcal{D}_N} h_{p,k}(Z_i, Z_j)$

Design: $\mathcal{D}_N$ random / deterministic subset of $\{(i,j)\}_{1 \leq i \neq j \leq N}$

Linear time: $|\mathcal{D}_N| = cN$ for some fixed integer $c \in \mathbb{N}$

KSDAggInc Uniform separation rate

KSDAggInc: use $\overline{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N)$ instead of $\widehat{\mathrm{KSD}}^2_{p,k}(\mathbb{Z}_N)$

Uniform separation rate: same condition as for **KSDAgg** with N multiplied by an extra cost factor $|\mathcal{D}_N| / N^2$

- $|\mathcal{D}_N| \asymp N^2$: recover **KSDAgg** rate
- $N \lesssim |\mathcal{D}_N| \lesssim N^2$: cost $|\mathcal{D}_N| / N^2$ incurred in **KSDAgg** rate
  Trade-off: computational efficiency / rate convergence
- $|\mathcal{D}_N| \lesssim N$: no guarantee that rate converges to 0

Experiments

Gaussian-Bernoulli Restricted Boltzmann Machine: graphical model with binary hidden variable $h \in \{\pm 1\}^{d_h}$ & continuous observable variable $x \in \mathbb{R}^{d_x}$ with joint density

$$p(x, h) = \frac{1}{Z} \exp\left(\frac{1}{2} x^\top B h + b^\top x + c^\top h - \frac{1}{2} \|x\|_2^2 \right)$$

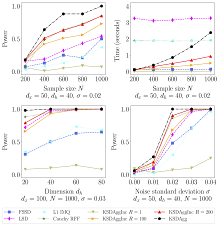
- model: GBRBM with $B \in \{\pm 1\}^{d_x \times d_h}$, $b \in \mathbb{R}^{d_x}$, $c \in \mathbb{R}^{d_h}$
- samples: GBRBM with noise $\mathcal{N}(0, \sigma)$ injected into B

Collection: Gaussian kernels with scaled median bandwidth

Parameter R: number of subdiagonals of the kernel matrix

FSSD: Jitkrittum et al. 2017 **LSD:** Grathwohl et al. 2020

L1 IQM & Cauchy RFF: Huggins and Mackey 2018



KSDAgg paper code

AggInc paper code