Aggregated Kernel Tests

These works were supported by UK Research and Innovation [grant number EP/S021566/1].

MMD Aggregated Two-sample Test
KSD Aggregated Goodness-of-fit Test
Efficient Aggregated Kernel Tests

Antonin Schrab

Engineering and Physical Sciences Research Council

Two-sample testing

Are two samples identically distributed?

🐧🐦🐥 🐠🐟🐡

Applications

- Clinical trial
- Change point detection
- Combining datasets
- Quality evaluation of generated samples
- Causal discovery using conditional GANs
- Domain adaptation: train/test datasets

Independence testing

Are paired samples dependent or independent?

(🇬🇧,☕) (🇫🇷,🥖) (🇮🇹,🍕)

Applications

- *Medicine: drug / recovery*
- *Neuroscience: stimulus / brain activity*
- *Genomics: gene selection*
- *Finance: stock market returns dependence*
- *Econometrics: economic independence hypothesis*
- *Machine Learning: feature selection*

Goodness-of-fit testing

Are samples coming from a given model?

🖥️ 📈📉📊

Applications

- *Fitting models to data verification*
- *Sample generation verification*
- *Sampling methods verification*
- *Model change point detection*
- *Model selection*
- *Composite testing: generalise to family of models*

Kernel: measure of similarity

k(🐥, 🐦) = 💯 k(🐥,🐠) = 0️⃣

Advantages

- **Generality:** Allows for any type of data (numbers, images, graphs, text, audio)
- **Kernel trick:** Work efficiently with infinite number of dimensions

Kernel-based measures

- **MMD:** Maximum Mean Discrepancy
- **HSIC:** Hilbert Schmidt Independence Criterion
- **KSD:** Kernel Stein Discrepancy

Expressive measures depending on the choice of kernel and kernel bandwidth

How to choose the kernel or kernel bandwidth?

Aggregated kernel tests

- **Problem:** Importance of testing on different length scales 🌍🇪🇺🇫🇷🏙️📍
- **Solution:** Aggregate tests with multiple kernel bandwidths
- **Theory:** Minimax optimality and adaptivity over Sobolev balls
- **Practice:** Outperform state-of-the-art adaptive kernel tests in terms of power

🚀 Computationally efficient aggregated tests 🚀

- **`Big data' problem:** Access to millions of data points (long compute times)
- **Solution:** Linear-time variants of the three quadratic-time aggregated tests
- **Method:** Subsampling entries of the kernel matrix
- **Trade-off:** Between computational time and cost in minimax rate