A Linear-Time Kernel Goodness-of-Fit Test

Wittawat Jitkrittum[1] Wenkai Xu[1] Zoltán Szabó[2] Kenji Fukumizu[3] Arthur Gretton[1]

[1]Gatsby Unit, University College London [2]CMAP, École Polytechnique [3]The Institute of Statistical Mathematics

Summary

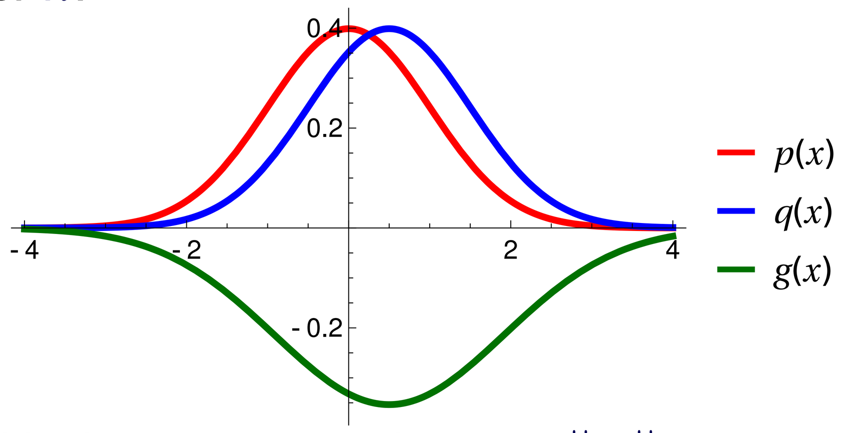
- **Given**: $\{\mathbf{x}_i\}_{i=1}^n \sim q$ (unknown), and a density p.
- **Goal**: Test $H_0: p = q$ vs $H_1: p \neq q$ quickly.
- **New multivariate goodness-of-fit test (FSSD)**:

1. Nonparametric: arbitrary, unnormalized p. $\mathbf{x} \in \mathbb{R}^d$.
2. Linear-time: $\mathcal{O}(n)$ runtime complexity. Fast.
3. Interpretable: tell where p does not fit the data.

Previous: Kernel Stein Discrepancy (KSD)

- Let $\xi(\mathbf{x},\mathbf{v}) := \frac{1}{p(\mathbf{x})}\nabla_{\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.

Stein witness function: $\mathbf{g}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\xi(\mathbf{x},\mathbf{v})]$ where $\mathbf{g} = (g_1,\ldots,g_d)$ and each $g_i \in \mathcal{F}$, an RKHS associated with kernel k.



Known: Under some conditions, $\|\mathbf{g}\|_{\mathcal{F}^d} = 0 \iff p = q$.

[Chwialkowski et al., 2016, Liu et al., 2016]

Statistic: $\mathrm{KSD}^2 = \|\mathbf{g}\|^2_{\mathcal{F}^d} = \overbrace{\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{y}\sim q}}^{\text{double sums}} h_p(\mathbf{x},\mathbf{y}) \approx \frac{2}{n(n-1)}\sum_{i<j} h_p(\mathbf{x}_i,\mathbf{x}_j)$. where

$$h_p(\mathbf{x},\mathbf{y}) := [\nabla_{\mathbf{x}}\log p(\mathbf{x})]\,k(\mathbf{x},\mathbf{y})\,[\nabla_{\mathbf{y}}\log p(\mathbf{y})] + \nabla_{\mathbf{x}}\nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y}) + [\nabla_{\mathbf{y}}\log p(\mathbf{y})]\,\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{y}) + [\nabla_{\mathbf{x}}\log p(\mathbf{x})]\,\nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y}).$$

Characteristics of KSD:

✓ Nonparametric. Applicable to a wide range of p.
✓ Do not need the normalizer of p.
✗ Runtime: $\mathcal{O}(n^2)$. Computationally expensive. ☹

Linear-Time KSD (LKS) Test: [Liu et al., 2016]

$$\|\mathbf{g}\|^2_{\mathcal{F}^d} \approx \frac{2}{n}\sum_{i=1}^{n/2} h_p(\mathbf{x}_{2i-1},\mathbf{x}_{2i}).$$

✓ Runtime: $\mathcal{O}(n)$. ✗ High variance. Low test power. ☹

The Finite Set Stein Discrepancy (FSSD)

Idea: Evaluate witness $\mathbf{g}$ at J locations $\{\mathbf{v}_1,\ldots,\mathbf{v}_J\}$. Fast.

$$\mathrm{FSSD}^2 = \frac{1}{dJ}\sum_{j=1}^{J}\|\mathbf{g}(\mathbf{v}_j)\|_2^2.$$

Proposition (FSSD is a discrepancy measure). *Main conditions:*

1. *(Nice kernel) Kernel k is C_0-universal, and* **real analytic** *(Taylor series at any point converges) e.g., Gaussian kernel.*
2. *(Vanishing boundary)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.
3. *(Avoid "blind spots") Locations $\{\mathbf{v}_1,\ldots,\mathbf{v}_J\}$ are drawn from a distribution η which has a density.*

Then, for any $J \geq 1$, η-a.s. $\mathrm{FSSD}^2 = 0 \iff p = q$.

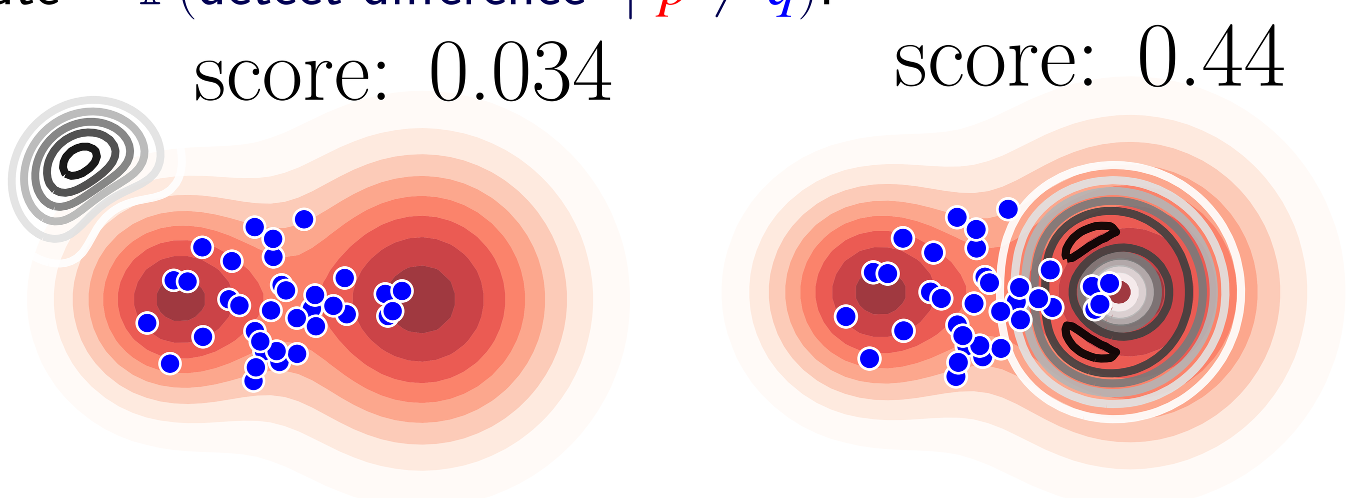
Characteristics of FSSD:

✓ Nonparametric. ✓ Do not need the normalizer of p.
✓ Runtime: $\mathcal{O}(n)$. ✓ Higher test power than LKS. ☺ ☺

Model Criticism with FSSD

Proposal: Optimize locations $\{\mathbf{v}_1,\ldots,\mathbf{v}_J\}$ and kernel bandwidth by $\arg\max$ score $= \mathrm{FSSD}^2/\sigma_{H_1}$ (runtime: $\mathcal{O}(n)$).

Proposition: This procedure maximizes the true positive rate $= \mathbb{P}(\text{detect difference} \mid p \neq q)$.

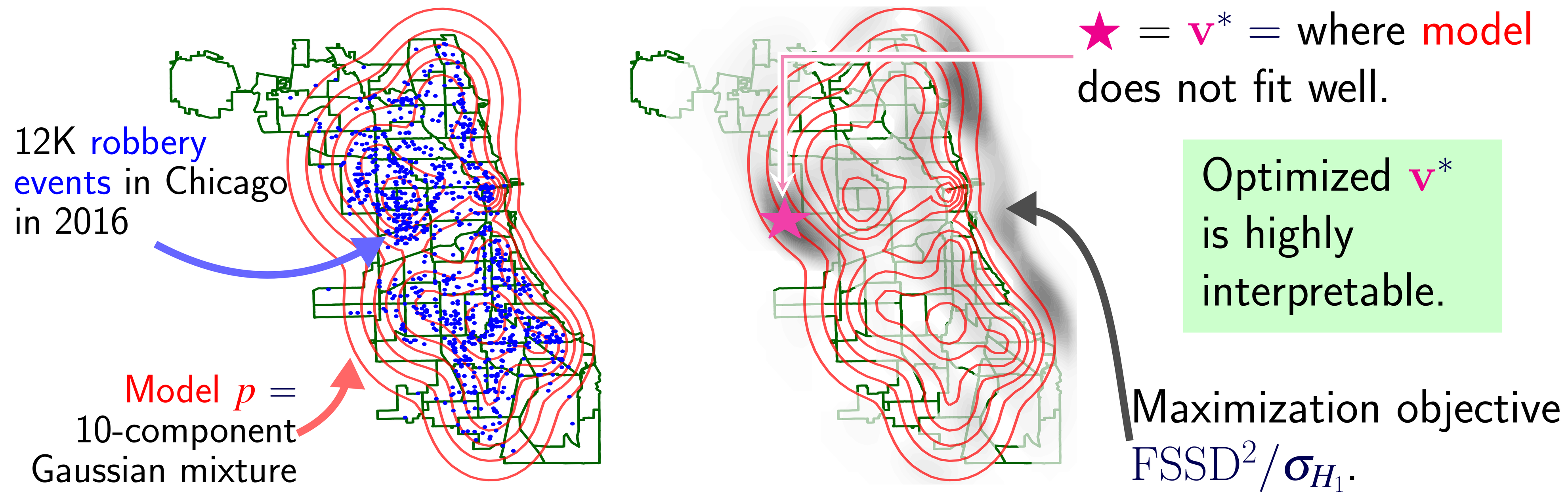


WJ, WX, and AG thank the Gatsby Charitable Foundation for the financial support. ZSz was financially supported by the Data Science Initiative. KF has been supported by KAKENHI Innovative Areas 25120012.

Contact: `wittawat@gatsby.ucl.ac.uk`

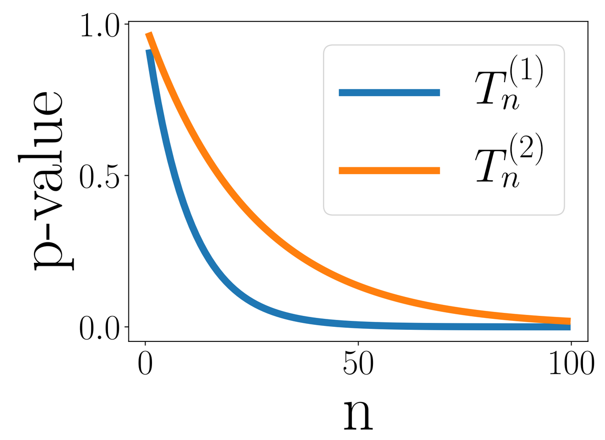
Code: `github.com/wittawatj/kernel-gof`

Interpretable Features for Model Criticism



Bahadur Slope and Bahadur Efficiency

- Bahadur slope ≈ rate of p-value → 0 of statistic T_n under H_1. High = good.
- Bahadur efficiency = ratio $\frac{\text{slope}^{(1)}}{\text{slope}^{(2)}}$ of slopes of two tests. > 1 means test$^{(1)}$ better.
- **Results**: Slopes of FSSD and LKS tests when $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$.



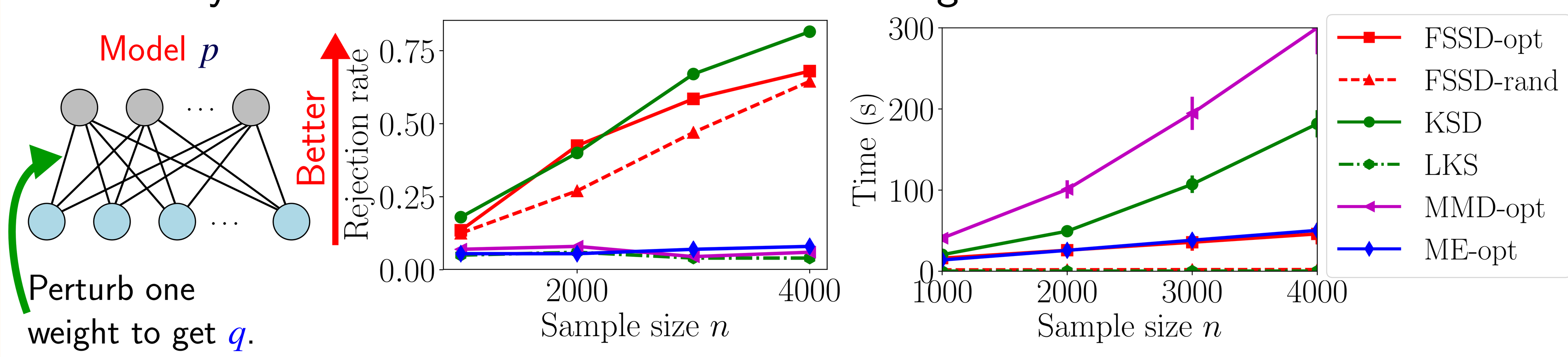
Proposition. *Let σ_k^2, κ^2 be kernel bandwidths of FSSD and LKS. Fix $\sigma_k^2 = 1$. Then, $\forall \mu_q \neq 0$, $\exists \mathbf{v} \in \mathbb{R}$, $\forall \kappa^2 > 0$, the Bahadur efficiency*

$$\frac{\mathrm{slope}^{(\mathrm{FSSD})}(\mu_q, \mathbf{v}, \sigma_k^2)}{\mathrm{slope}^{(\mathrm{LKS})}(\mu_q, \kappa^2)} > 2.$$

FSSD is statistically more efficient than LKS.

Experiment: Restricted Boltzmann Machine

- 40 binary hidden units. $d = 50$ visible units. Significance level $\alpha = 0.05$.



- FSSD-opt, (FSSD-rand) = Proposed tests. $J = 5$ optimized, (random) locations.
- MMD-opt [Gretton et al., 2012] = State-of-the-art two-sample test (quadratic-time).
- ME-opt [Jitkrittum et al., 2016] = Linear-time two-sample test with optimized locations.
- **Key**: FSSD ($\mathcal{O}(n)$), KSD ($\mathcal{O}(n^2)$) have comparable power. FSSD is much faster.