Kernel Stein Tests for Multiple Model Comparison

Jen Ning Lim[1] Makoto Yamada[2,3] Bernhard Schölkopf[1] Wittawat Jitkrittum[1]

[1]Max Planck Institute for Intelligent Systems [2]Kyoto University [3]RIKEN AIP

Summary

- **Given**: l candidate models $\mathcal{M} = \{P_1, \dots, P_l\}$ and n samples $\{\mathbf{y}_i\}_{i=1}^n \sim R$ from unknown data distribution R. In $\mathbb{R}^d$.
- **Goal**: For each model $P_i \in \mathcal{M}$, determine if it (the model) is worse that the best candidate model P_J in $\mathcal{M}$.
- **Contribution**:
  - Proposed two methods for constructing non-parametric tests for *multiple model comparison*.
    1. Mult: The dataset is split into two independent portions for selection (i.e., determining $P_{\hat{J}}$) and testing.
    2. PSI: The same dataset is used for selection and testing.
  - When $l = 2$, PSI has provably higher true positive rate (TPR) than Mult.
  - Both methods control false positive rate (FPR).

Multiple Model Comparison

- Test $H_{0,i} : D(P_i, R) \leq D(P_J, R)$ v.s. $H_{1,i} : D(P_i, R) > D(P_J, R)$ where P_J is the best candidate model in $\mathcal{M}$.
- For each model P_i, we either have samples $\{\mathbf{x}_j\}_{j=1}^n \sim P_i$ (D = MMD) or have the log density $\log p_i(x)$ (D = KSD).
- **Problem**: P_J is unknown.
- **Idea**: Estimate $\hat{J}$ from data. We choose $P_{\hat{J}} \in \mathrm{argmin}_{P_i \in \mathcal{M}} \hat{D}(P_i, R)$.
- **Proposal**: Consider hypotheses conditioned on the selection
  - $H_{0,i}^{\hat{J}} : D(P_i, R) \leq D(P_{\hat{J}}, R) \mid P_{\hat{J}}$ is the best,
  - $H_{1,i}^{\hat{J}} : D(P_i, R) > D(P_{\hat{J}}, R) \mid P_{\hat{J}}$ is the best.
- *Conditional* null hypothesis $\implies$ *valid* post selection tests that account for selection bias.

False and True Positive Rates

- We use false positive rate (FPR) and true positive rate (TPR) to measure the performance of our algorithm.
- True positive rate (TPR) is the proportion of models correctly designated as worse than the best P_J.
- False positive rate (FPR) is the proportion of models incorrectly designated as worse than the best P_J.
- A good test has high TPR and low FPR.

Test Statistic

Our test statistic is $\hat{S} := \sqrt{n}[\hat{D}(P_i, R) - \hat{D}(P_{\hat{J}}, R)]$ where $\hat{D}$ is an unbiased estimator of KSD or MMD.

- **Before selection**, $\hat{S}$ is normally distributed as $n \to \infty$.
- **After selection**, $\hat{S}$ under the *conditional null* is asymptotically:
  - **Normal** if the data are split into two independent portions: one for selection and the other for inference.
  - **Truncated normal** if the full dataset is used for both selection and testing, by the polyhedral lemma of Lee et. al. 2016.

- For all $i = 1, \dots, l$, reject $H_{0,i}$ if $\hat{S} > (1-\alpha)$-quantile of the distribution of $\hat{S}$ **after** selection.
- Reject $H_{0,i} \implies$ Model P_i is worse than the best $P_{\hat{J}}$.
- This asymptotically controls the type-I error(and FPR) at α. ✓

Theoretical Result

Theorem: Let P_1, P_2 be two candidate models, and R be a data generating distribution. Assume that P_1, P_2 and R are distinct. Given $\alpha \in [0, \frac{1}{2}]$ and split proportion $\rho \in (0,1)$ for Mult so that $(1-\rho)n$ samples are used for selecting $P_{\hat{J}}$ and ρn samples for testing, for all $n \gg N = \left(\frac{\sigma \Phi^{-1}(1-\frac{\alpha}{2})}{\mu(1-\sqrt{\rho})}\right)^2$, we have

$$\mathrm{TPR}_{PSI} \gtrapprox \mathrm{TPR}_{Mult}.$$

✓PSI can yield higher TPR than Mult.
✓Holds for both $D = MMD$ and $D = KSD$.

Lemma: Define the selective type-I error for the i^{th} model to be

$$s(i, \hat{J}) := \mathbb{P}(\text{reject } H_{0,i}^{\hat{J}} \mid H_{0,i}^{\hat{J}} \text{ is true}, P_{\hat{J}} \text{ is selected}).$$

If $s(i, \hat{J}) \leq \alpha$ for all $i, \hat{J} \in \{1, \dots, l\}$, then

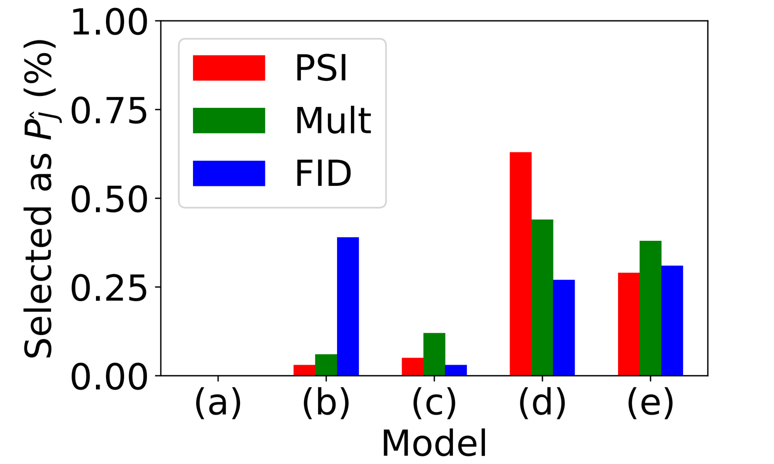
$$\mathrm{FPR} \leq \alpha.$$

✓Both of our test controls provably controls FPR.

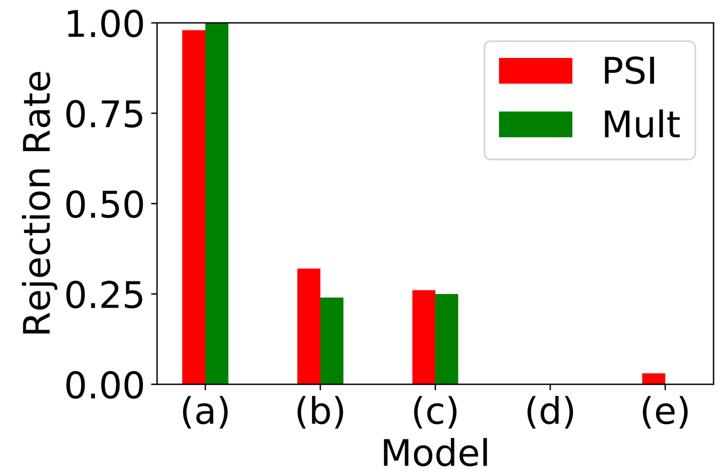
Mixture of CelebA (D = MMD)

Task: If the true distribution is composed of 50% smile and 50% non-smile (images), which of the following models are the closest?

(a) GAN: Smile 50%, No-smile 50%.
(b) Real: Smile 60%, No-smile 40%.
(c) Real: Smile 40%, No-smile 60%.
(d) Real: Smile 51%, No-smile 49%.
(e) Real: Smile 52%, No-smile 48%.

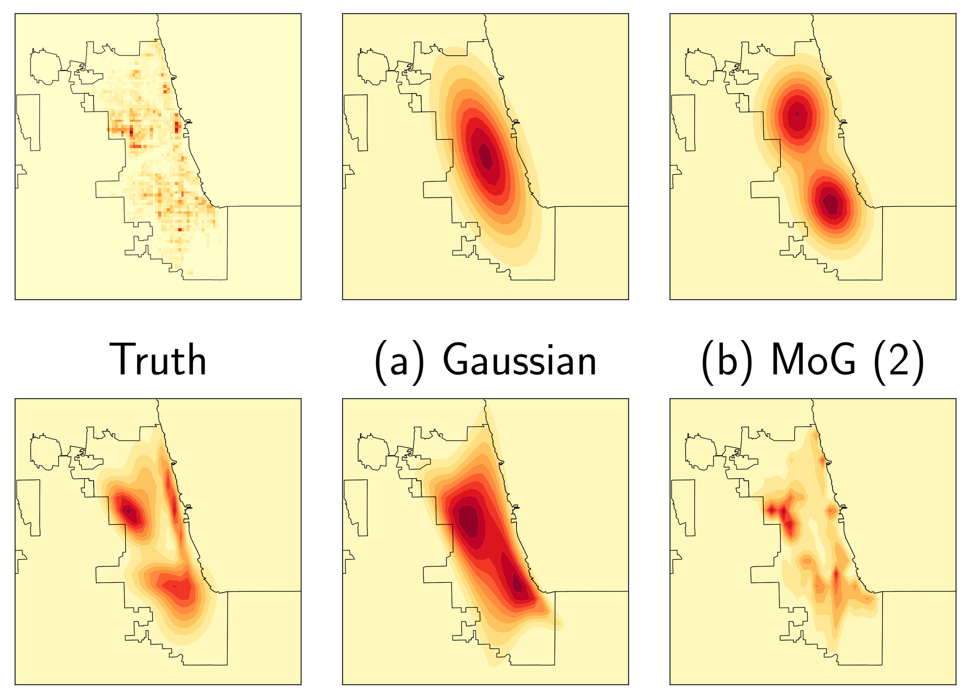


Results:
- Ranking by FID $\implies$ noisy selection.
- Testing indicates that (d) and (e) are the best.
- Performance of PSI and Mult similar.



Chicago Crime (D = KSD)

Task: Best model for representing the crime activity in Chicago?



Truth (a) Gaussian (b) MoG (2)
(c) MoG (5) (d) MADE (e) MAF (5)

Results:
- Ranking by KSD $\implies$ (c) and (e) are selected.
- Negative Log Likelihood (NLL) favors the most complex model (e).
- PSI has higher rejection rate than Mult.