Informative Features for Comparing Distributions

Wittawat Jitkrittum

Max Planck Institute for Intelligent Systems

wittawat.com

DALI 2019, San Sebastian, Spain

4 September 2019

They Play a Big Part in My PhD Journey

- **Arthur Gretton** (Gatsby Unit, UCL)
- Zoltán Szabó (École Polytechnique)
- Massimiliano Pontil (Istituto Italiano di Tecnologia & UCL)
- Nando de Freitas (University of Oxford & DeepMind)
- Peter Dayan (Max Planck Institute for Biological Cybernetics)
- Members of Gatsby Unit, UCL
- Kenji Fukumizu (Institute of Statistical Mathematics)
- Mijung Park (Max Planck Institute for Intelligent Systems)
- Dino Sejdinovic (University of Oxford)
- Nicolas Heess (DeepMind)
- Ali Eslami (DeepMind)
- Balaji Lakshminarayanan (DeepMind)
- Maneesh Sahani (Gatsby Unit, UCL)
- Kacper Chwialkowski (Voleon)
- Wenkai Xu (Gatsby Unit, UCL)
- My family and friends
- $\vdots$

My PhD Thesis

- At Gatsby Unit, University College London.
  - Supervisor: Arthur Gretton.
- Thesis: Kernel-Based Distribution Features for Statistical Tests and Bayesian Inference
  - Study algorithms to extract **interpretable** "features" from distributions
- Focus: scalable algorithms $\mathcal{O}(n)$ + theoretical justification

Problems tackled:

1.
2.
3. Dependence measure
4. Amortized message passing with expectation propagation

My PhD Thesis

- At Gatsby Unit, University College London.
  - Supervisor: Arthur Gretton.
- Thesis: Kernel-Based Distribution Features for Statistical Tests and Bayesian Inference
  - Study algorithms to extract **interpretable** "features" from distributions
- Focus: scalable algorithms $\mathcal{O}(n)$ + theoretical justification

Problems tackled:

1
2
3 Dependence measure
4 Amortized message passing with expectation propagation

My PhD Thesis

- At Gatsby Unit, University College London.
  - Supervisor: Arthur Gretton.
- Thesis: Kernel-Based Distribution Features for Statistical Tests and Bayesian Inference
  - Study algorithms to extract **interpretable** "features" from distributions
- Focus: scalable algorithms $\mathcal{O}(n)$ + theoretical justification

Problems tackled:

1. Two-sample testing
2. Model criticism
3. Dependence measure
4. Amortized message passing with expectation propagation

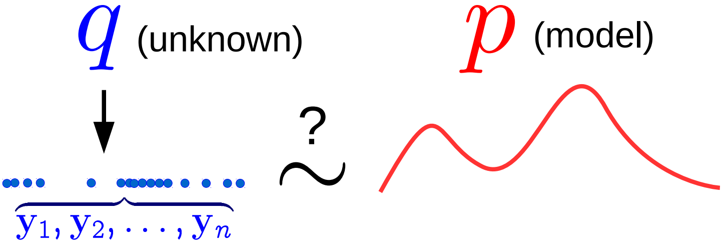
My PhD Thesis

- At Gatsby Unit, University College London.
  - Supervisor: Arthur Gretton.
- Thesis: Kernel-Based Distribution Features for Statistical Tests and Bayesian Inference
  - Study algorithms to extract **interpretable** "features" from distributions
- Focus: scalable algorithms $\mathcal{O}(n)$ + theoretical justification

Problems tackled:

1. **Two-sample testing** $\leftarrow$ **(this talk)**
2. **Goodness-of-fit testing** $\leftarrow$ **(this talk)**
3. Dependence measure
4. Amortized message passing with expectation propagation

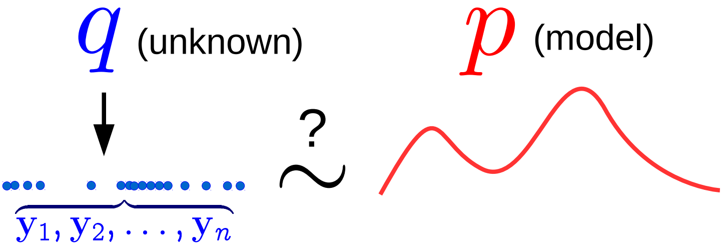
Problem Setting: Distribution Comparison



Test goal: Do data follow the model p?

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Tell where the model is wrong. ★.

Problem Setting: Distribution Comparison

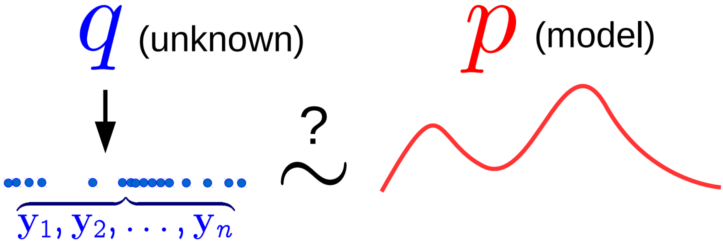


Test goal: Do data follow the model p?

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Tell where the model is wrong. ★.

Problem Setting: Distribution Comparison

Goodness-of-fit testing

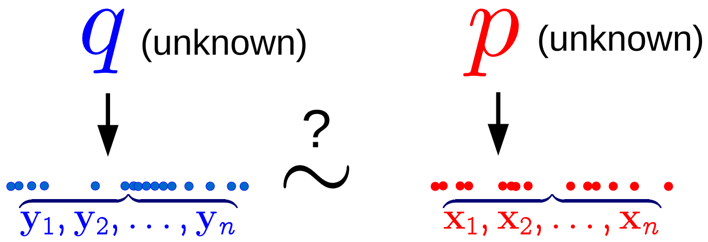


Test goal: Do data follow the model p?

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. **Interpretable.** Tell where the model is wrong. ★.

Problem Setting: Distribution Comparison

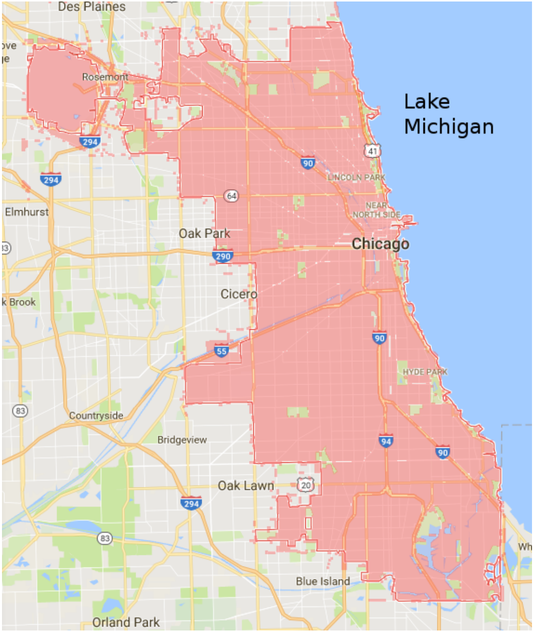
Two-sample testing



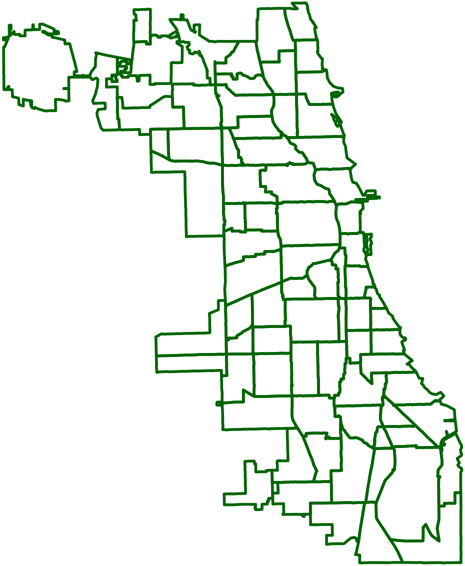
Test goal: Do data follow the model p?

1. Nonparametric.
2. Linear-time. Runtime is $\mathcal{O}(n)$. Fast.
3. Interpretable. Tell where the model is wrong. .

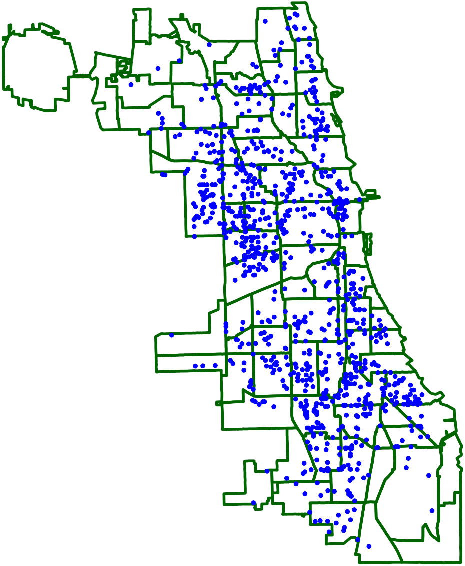
Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)



Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

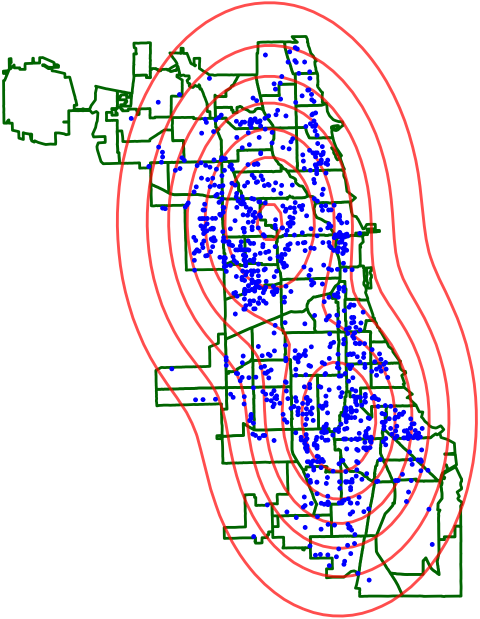


Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)



- Robbery event coordinates (samples from q).
- **Goal**: Model spatial density.

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

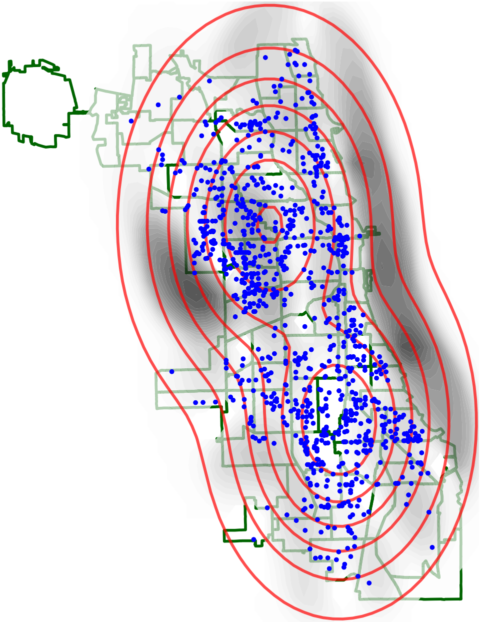


A candidate model

p = Mixture of 2 Gaussians.

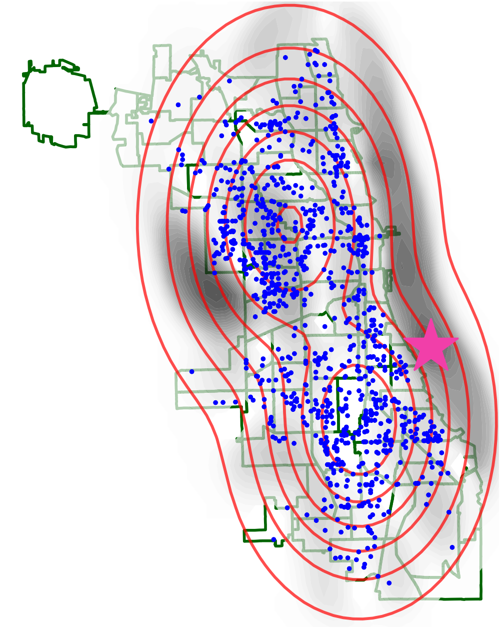
Is p a good model?

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)



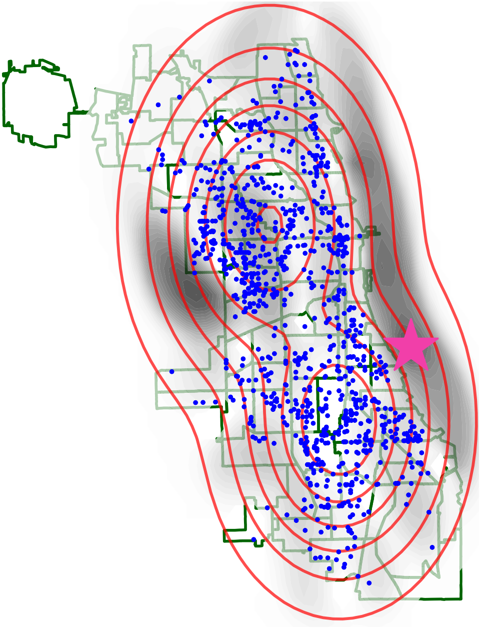
Score surface
(black = large mismatch)

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)



★ = optimized $\mathbf{v}$.

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)



★ = optimized **v**.

No robbery in Lake Michigan.

Sharp data boundary. Not follow Gaussian tails.

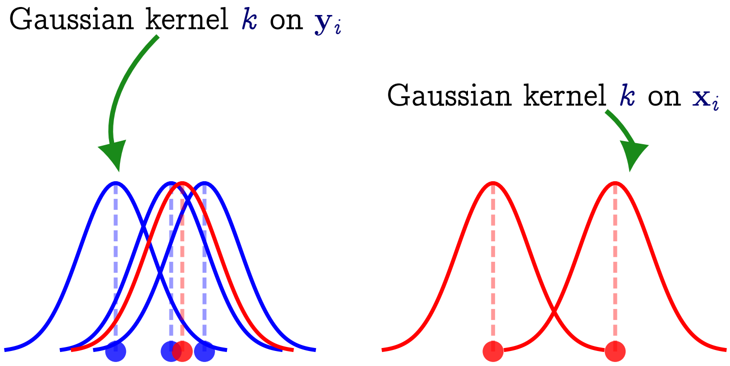
The Witness Function [Gretton et al., 2012]

The Witness Function [Gretton et al., 2012]

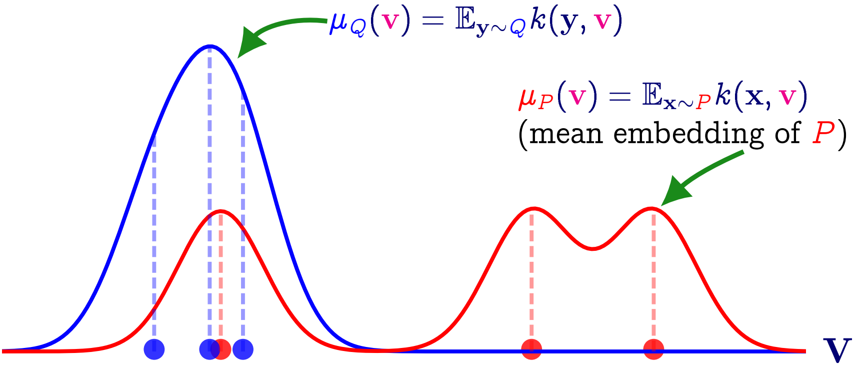
Observe $Y_n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim Q$

Observe $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim P$

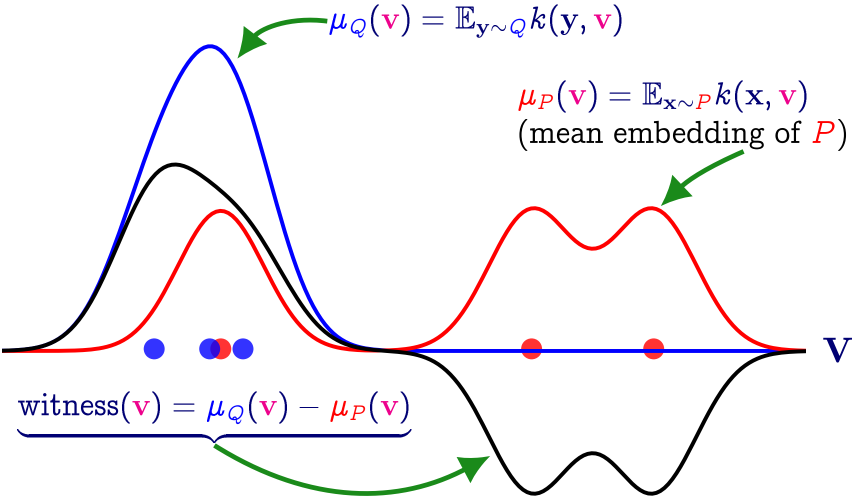
The Witness Function [Gretton et al., 2012]



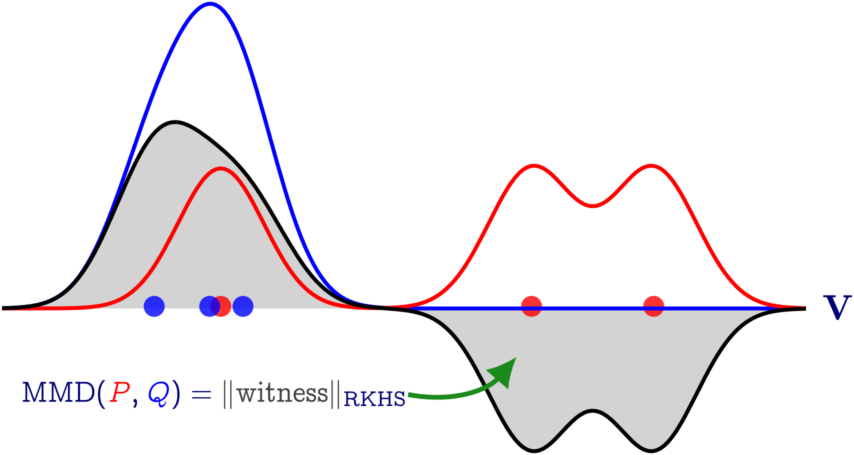
The Witness Function [Gretton et al., 2012]



The Witness Function [Gretton et al., 2012]

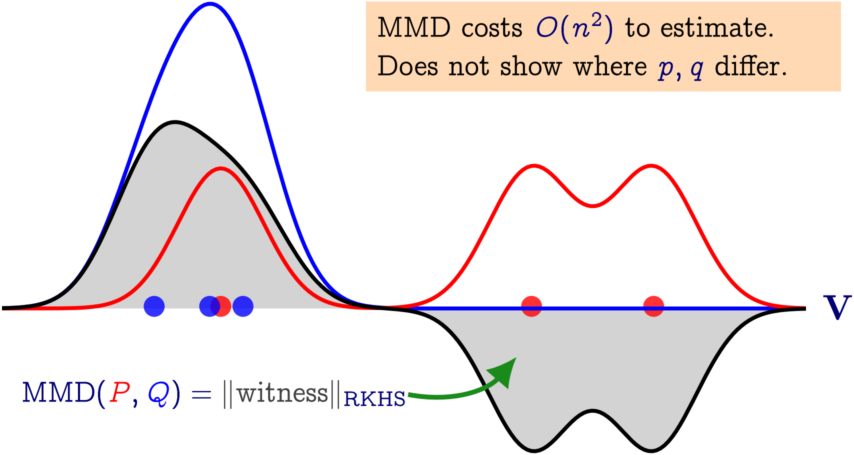


The Witness Function [Gretton et al., 2012]



The Witness Function [Gretton et al., 2012]

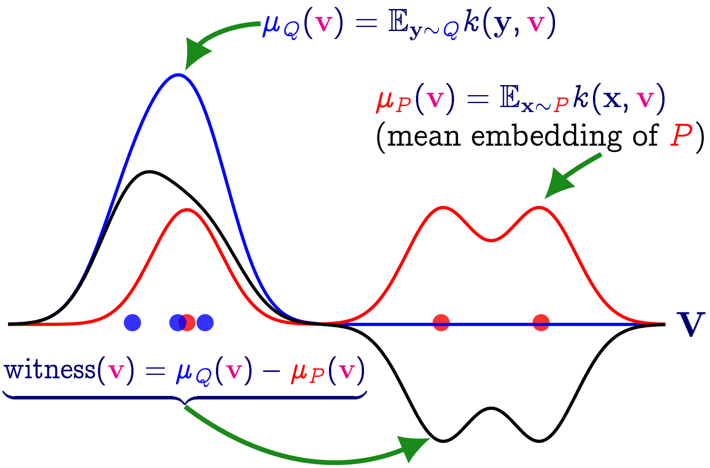
MMD costs $O(n^2)$ to estimate.
Does not show where p, q differ.



$\text{MMD}(P, Q) = \|\text{witness}\|_{\text{RKHS}}$

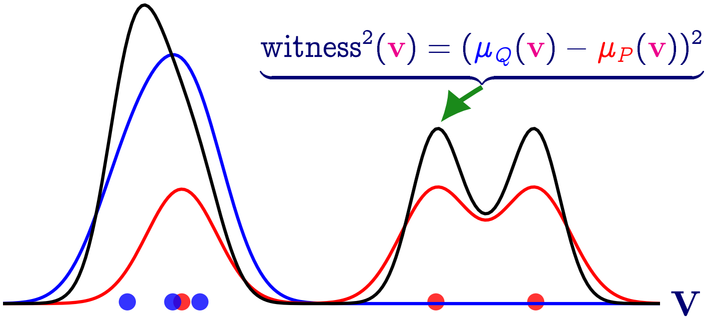
Proposal: The Unnormalized Mean Embeddings Statistic

[Chwialkowski et al., 2015, Jitkrittum et al., 2016]



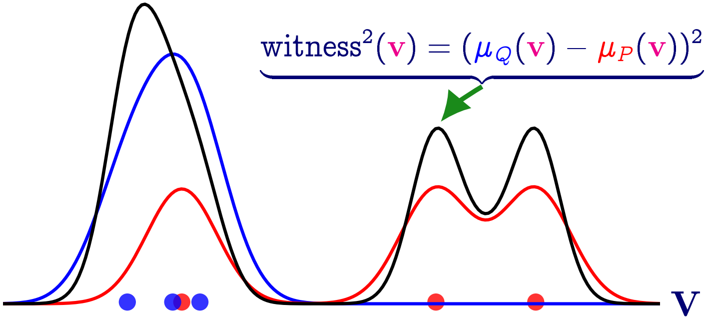
Proposal: The Unnormalized Mean Embeddings Statistic

[Chwialkowski et al., 2015, Jitkrittum et al., 2016]



Proposal: The Unnormalized Mean Embeddings Statistic

[Chwialkowski et al., 2015, Jitkrittum et al., 2016]



- Given J **optimized** test locations $V := \{\mathbf{v}_j\}_{j=1}^J = \{ ★, \ldots, ★ \}$,

$$\mathrm{UME}^2(P, Q) = \frac{1}{J} \sum_{j=1}^{J} \text{witness}^2(\mathbf{v}_j).$$

- Can be estimated in $\mathcal{O}(Jn)$.

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \mathrm{test\,power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

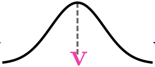
Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

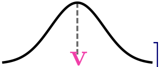
- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \mathrm{test\ power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

$$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad] - \mathbb{E}_{\mathbf{x}\sim p}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad]$$

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \mathrm{test\,power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

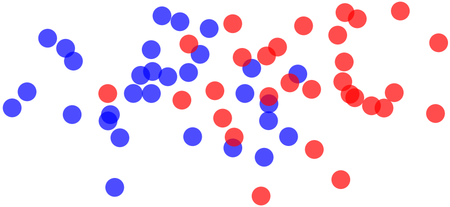
$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[$ 

$] - \mathbb{E}_{\mathbf{x}\sim p}[$ 

$]$

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \text{test power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

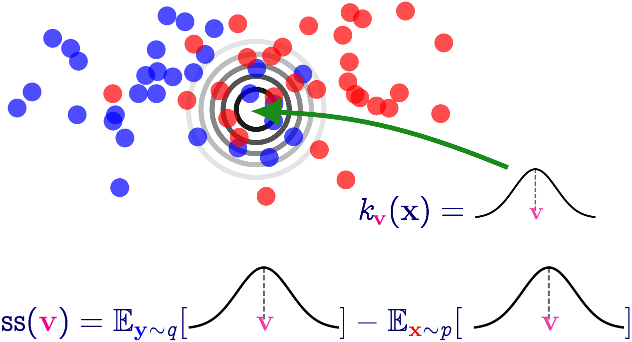


$$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad \mathbf{v} \quad] - \mathbb{E}_{\mathbf{x}\sim p}[\quad \mathbf{v} \quad]$$

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \text{test power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

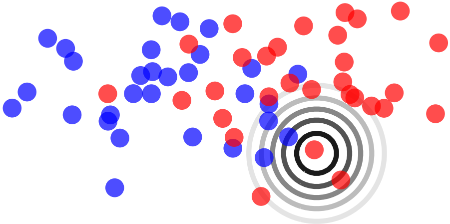
score: 0.008

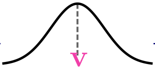
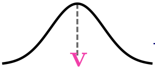


Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \text{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \text{test power}$.
- $\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}$

score: 1.6

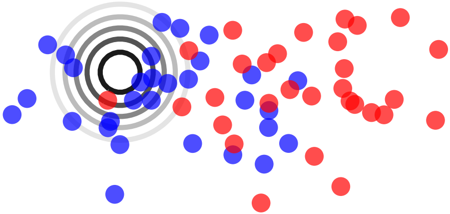


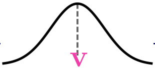
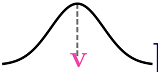
$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[$  $] - \mathbb{E}_{\mathbf{x}\sim p}[$  $]$

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \text{test power}.$
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

score: 13

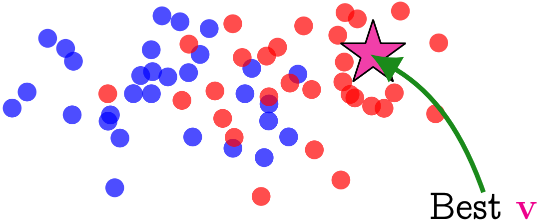


$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[$  $] - \mathbb{E}_{\mathbf{x}\sim p}[$  $]$

Interpretable Two-Sample Test with UME (NeurIPS 2016, oral)

- **Propose**: Find test location(s) $\mathbf{v}$ which maximize the probability of detecting differences (test power) between q and p.
- Show that $\arg\max_{\mathbf{v}} \mathrm{score}(\mathbf{v}) \implies \arg\max_{\mathbf{v}} \text{test power}$.
- $\mathrm{score}(\mathbf{v}) = \frac{\mathrm{witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}$

score: 25



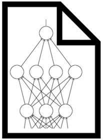
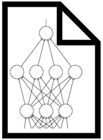
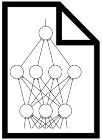
$$\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\ \mathbf{v}\ ] - \mathbb{E}_{\mathbf{x}\sim p}[\ \mathbf{v}\ ]$$

Bayesian Inference Vs. Deep Learning Papers

Papers on Bayesian inference

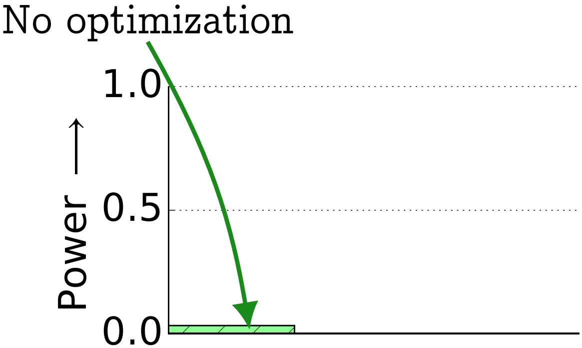
$\mathsf{X} = \{$ , ,  $, \ldots \} \sim p$

Papers on deep learning

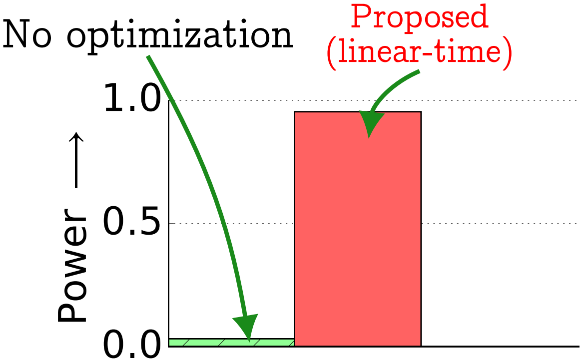
$\mathsf{Y} = \{$ , ,  $, \ldots \} \sim q$

- NeurIPS papers (1988-2015)
- Sample size $n = 216$.
- Random 2000 nouns (dimensions). TF-IDF representation.

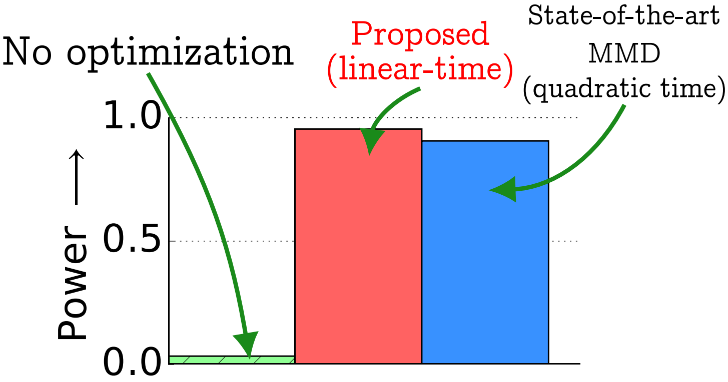
Bayesian Inference Vs. Deep Learning Papers



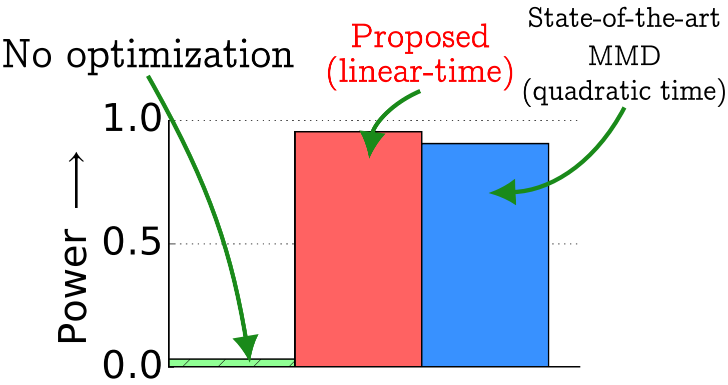
Bayesian Inference Vs. Deep Learning Papers



Bayesian Inference Vs. Deep Learning Papers



Bayesian Inference Vs. Deep Learning Papers



Learned test location ★ (a new document):

infer, Bayes, Monte Carlo, adaptor, motif,
haplotype, ECG, covariance, Boltzmann

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

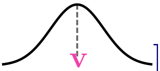
Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

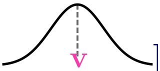
Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

$$(\text{Stein})\ \text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad] - \mathbb{E}_{\mathbf{x}\sim p}[\quad T_p k_{\mathbf{v}}(\mathbf{x}) \quad]$$

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

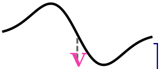
(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{y}\sim q}[\, T_p$ 

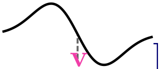
$] - \mathbb{E}_{\mathbf{x}\sim p}[\, T_p$ 

$]$

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

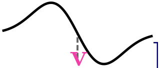
$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}\Big[$$ 

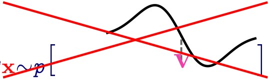
$$\Big] - \mathbb{E}_{\mathbf{x}\sim p}\Big[$$ 

$$\Big]$$

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{y}\sim q}[$ 

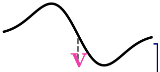
$] - \mathbb{E}_{\mathbf{x}\sim p}[$ 

$]$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{y}\sim q}[$ 

$]$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad]$$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad]$$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad]$$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

signal-to-noise ratio

Interpretable Goodness-of-Fit Test (NeurIPS 2017 Best Paper)

Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{x}\sim p}[k_{\mathbf{v}}(\mathbf{x})]$.

$$\text{(Stein) witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}[\quad T_p k_{\mathbf{v}}(\mathbf{y}) \quad]$$

Idea: Define T_p such that $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$, for any $\mathbf{v}$.

Proposal: Good $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

signal-to-noise ratio

- $\text{score}(\mathbf{v})$ can be estimated in **linear-time**.

What is $T_p k_{\mathbf{v}}$?

Recall Stein witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}(T_p k_{\mathbf{v}})(\mathbf{y}) - \cancel{\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x})}$

What is $T_p k_{\mathbf{v}}$?

Recall Stein witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}(T_p k_{\mathbf{v}})(\mathbf{y}) - \cancel{\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x})}$

$$(T_p k_{\mathbf{v}})(\mathbf{x}) = \frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})].$$

Then, $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0.$

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

What is $T_p k_{\mathbf{v}}$?

Recall Stein witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}(T_p k_{\mathbf{v}})(\mathbf{y}) - \cancel{\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x})}$

$$(T_p k_{\mathbf{v}})(\mathbf{x}) = \frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0$.

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

What is $T_p k_{\mathbf{v}}$?

Recall Stein witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{y}\sim q}(T_p k_{\mathbf{v}})(\mathbf{y}) - \cancel{\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x})}$

$$(T_p k_{\mathbf{v}})(\mathbf{x}) = \frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k(\mathbf{x},\mathbf{v})p(\mathbf{x})].$$

Normalizer cancels

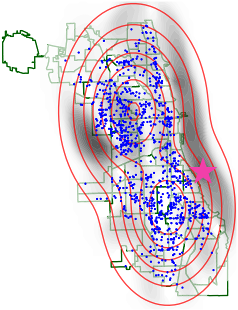
Then, $\mathbb{E}_{\mathbf{x}\sim p}(T_p k_{\mathbf{v}})(\mathbf{x}) = 0.$

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

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{x})\right] &= \int_{-\infty}^{\infty}\left[\frac{1}{\cancel{p(\mathbf{x})}}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]\cancel{p(\mathbf{x})}\,d\mathbf{x}\\
&= \int_{-\infty}^{\infty}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\,d\mathbf{x}\\
&= [k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]_{\mathbf{x}=-\infty}^{\mathbf{x}=\infty}\\
&= 0
\end{aligned}$$

(assume $\lim_{|\mathbf{x}|\to\infty} k(\mathbf{v},\mathbf{x})p(\mathbf{x})$)

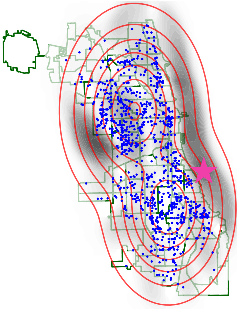
Conclusions



Proposed new tests for two-sample and goodness-of-fit testing:

1. Nonparametric
2. Linear-time
3. **Interpretable with** ★

Conclusions



Proposed new tests for two-sample and goodness-of-fit testing:

1. Nonparametric
2. Linear-time
3. **Interpretable with** ★

NeurIPS 2019 Tutorial
Interpretable Comparison of Distributions and Models
Wittawat Jitkrittum, Dougal Sutherland, Arthur Gretton

Questions?

Thank you

The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j))^2.$$

Proposition 1 (Chwialkowski et al., 2015, Jitkrittum et al., 2016).

Assume

1. *Kernel k is real analytic, integrable, and characteristic,*
2. *V is drawn from η, a distribution with a density e.g., standard normal.*

Then, for any $J > 0$, any P, Q, $\mathrm{UME}^2(P, Q) = 0$ iff $P = Q$, η-almost surely.

- **Key**: Evaluating witness^2 is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. J is small e.g., 10.

The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j))^2.$$

Proposition 1 (Chwialkowski et al., 2015, Jitkrittum et al., 2016).

Assume

1. *Kernel k is real analytic, integrable, and characteristic,*
2. *V is drawn from η, a distribution with a density e.g., standard normal.*

Then, for any $J > 0$, any P, Q, $\mathrm{UME}^2(P, Q) = 0$ iff $P = Q$, η-almost surely.

- **Key**: Evaluating witness^2 is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. J is small e.g., 10.

The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J}\sum_{j=1}^{J}(\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j))^2.$$

Proposition 1 (Chwialkowski et al., 2015, Jitkrittum et al., 2016).

Assume

1. *Kernel k is real analytic, integrable, and characteristic,*
2. *V is drawn from η, a distribution with a density e.g., standard normal.*

Then, for any $J > 0$, any P, Q, $\mathrm{UME}^2(P, Q) = 0$ iff $P = Q$, η-almost surely.

- Key: Evaluating witness^2 is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. J is small e.g., 10.

The Unnormalized Mean Embeddings (UME) Statistic

$$\mathrm{UME}^2(P, Q) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(\mathbf{v}_j) - \mu_Q(\mathbf{v}_j))^2.$$

Proposition 1 (Chwialkowski et al., 2015, Jitkrittum et al., 2016).

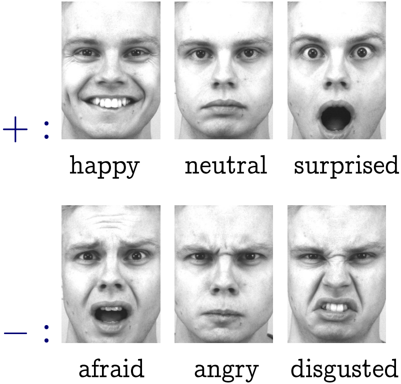
Assume

1. *Kernel k is real analytic, integrable, and characteristic,*
2. *V is drawn from η, a distribution with a density e.g., standard normal.*

Then, for any $J > 0$, any P, Q, $\mathrm{UME}^2(P, Q) = 0$ iff $P = Q$, η-almost surely.

- **Key**: Evaluating witness^2 is enough to detect the difference (in theory).
- Runtime complexity: $\mathcal{O}(Jn)$. J is small e.g., 10.

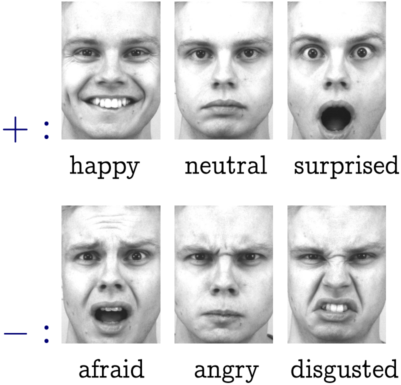
Distinguishing Positive/Negative Emotions

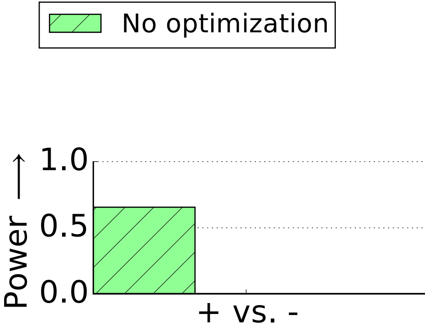


- 35 females and 35 males (Lundqvist et al., 1998).
- $48 \times 34 = 1632$ dimensions. Pixel features.
- $n = 201$.

- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

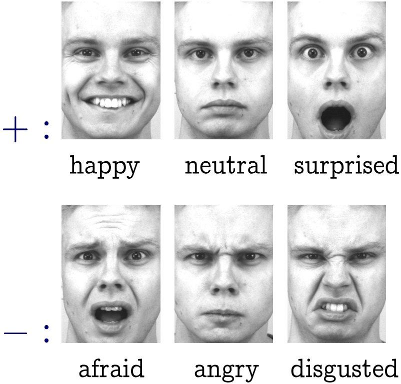
Distinguishing Positive/Negative Emotions

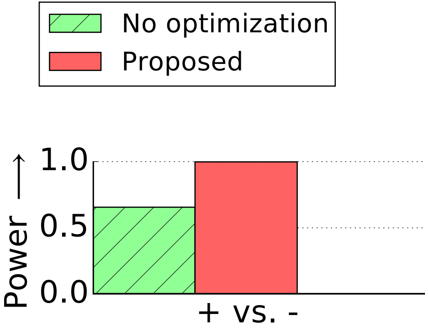




- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

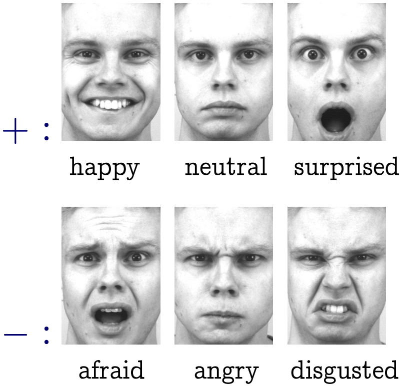
Distinguishing Positive/Negative Emotions

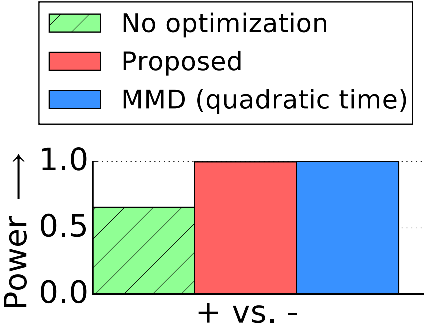




- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

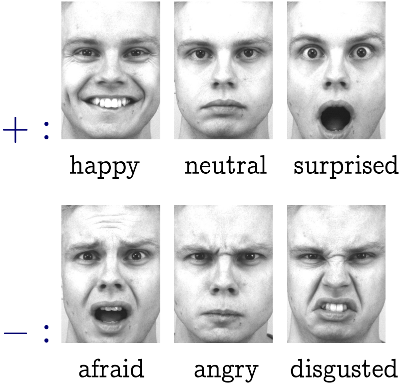
Distinguishing Positive/Negative Emotions

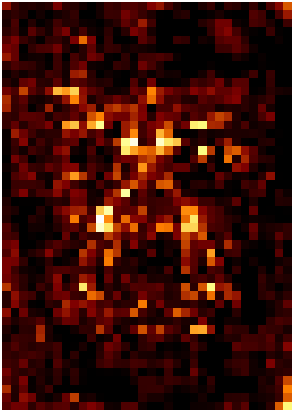




- Test power comparable to the state-of-the-art MMD test.
- Informative features: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions

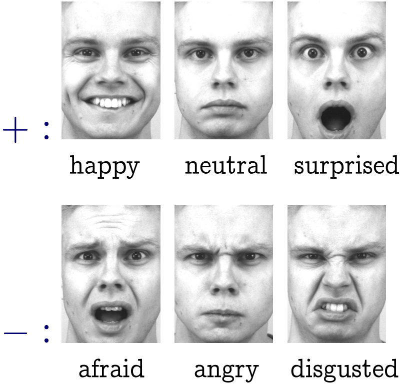




Learned ★

- Test power comparable to the state-of-the-art MMD test.
- **Informative features**: differences at the nose, and smile lines.

Distinguishing Positive/Negative Emotions



Learned ★

- Test power comparable to the state-of-the-art MMD test.
- **Informative features:** differences at the nose, and smile lines.

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y}) \sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

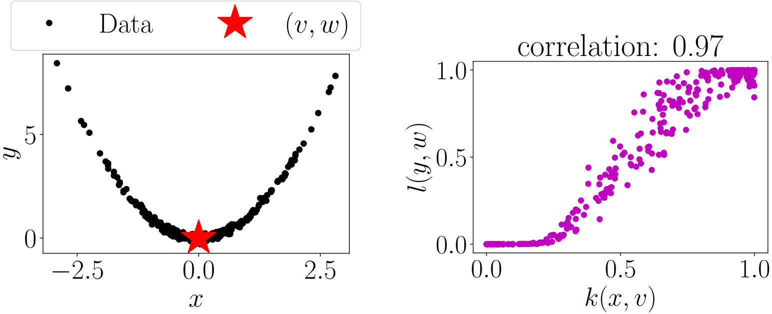
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



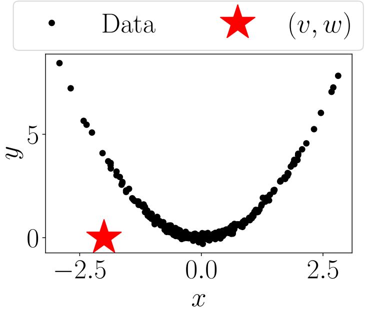
Proposal: The Finite-Set Independence Criterion (FSIC)

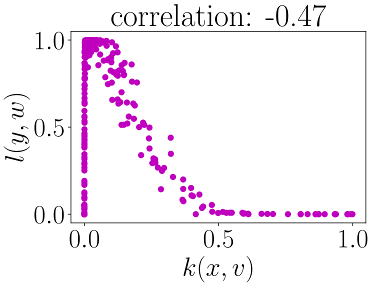
1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$





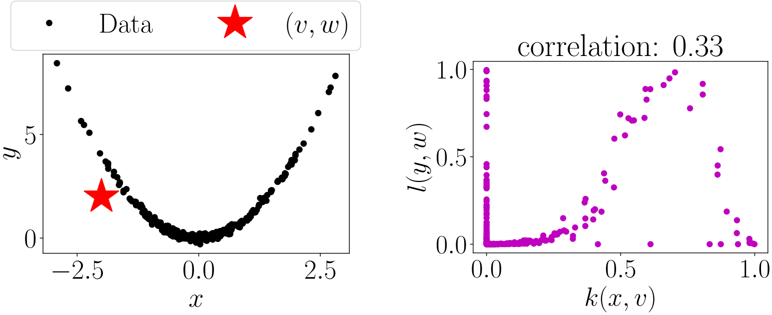
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



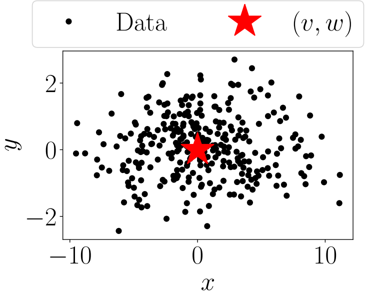
Proposal: The Finite-Set Independence Criterion (FSIC)

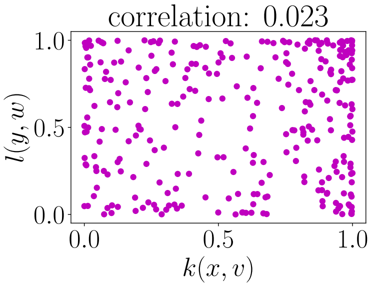
1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$





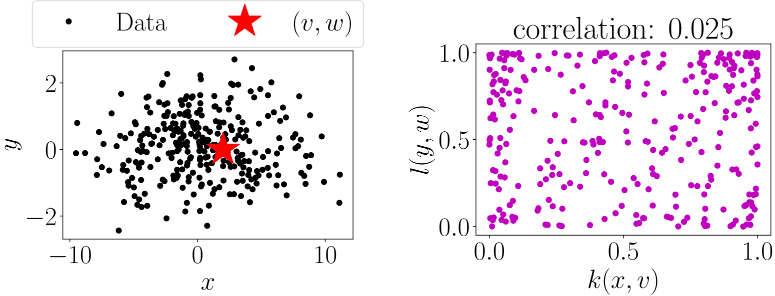
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y}) \sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



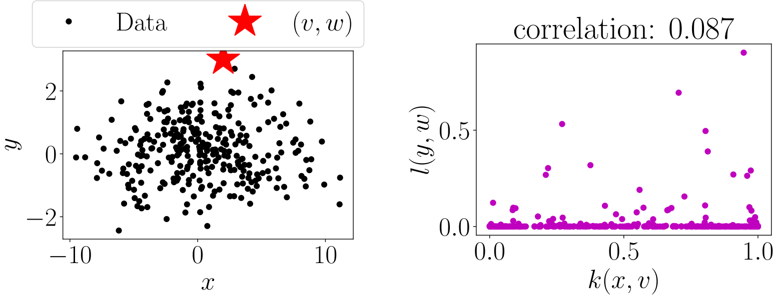
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **test location** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

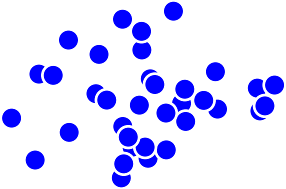
3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

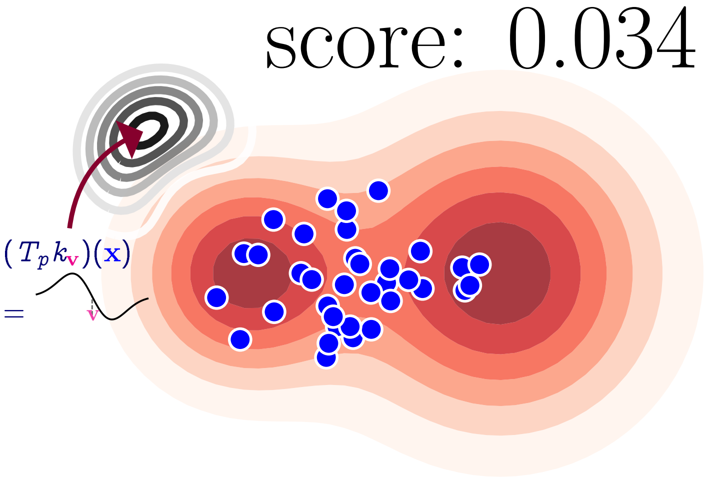


Proposal: Model Criticism with the Stein Witness



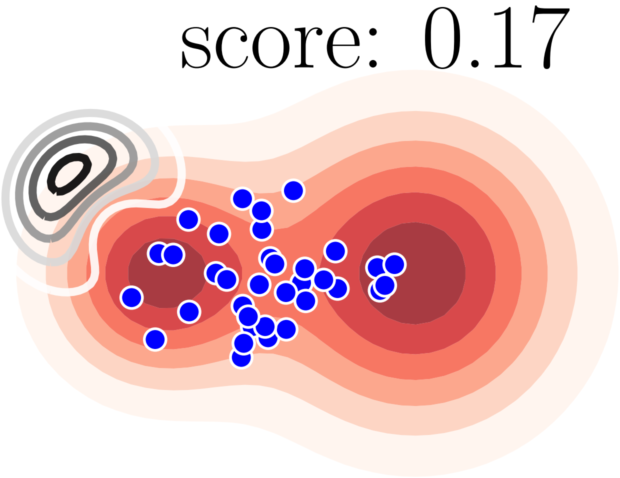
$$\mathrm{score}(\mathbf{v}) = \frac{\mathrm{Stein\ Witness}^2(\mathbf{v})}{\mathrm{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



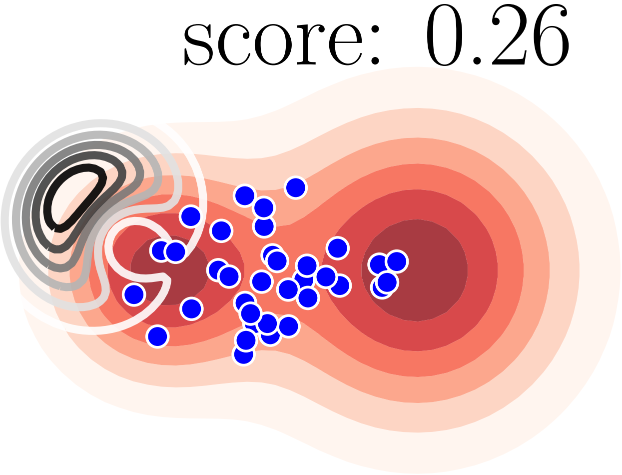
$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



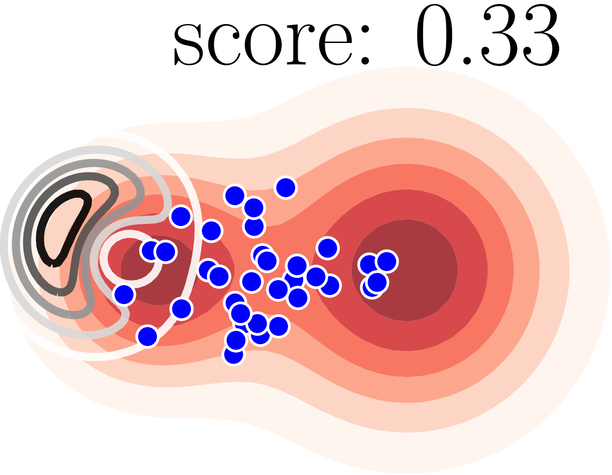
$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



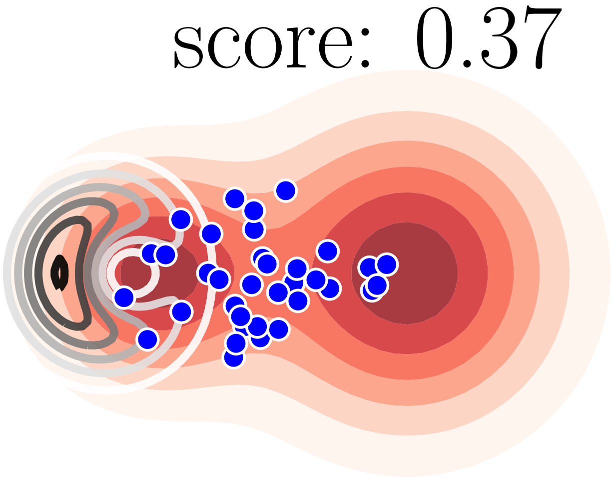
$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



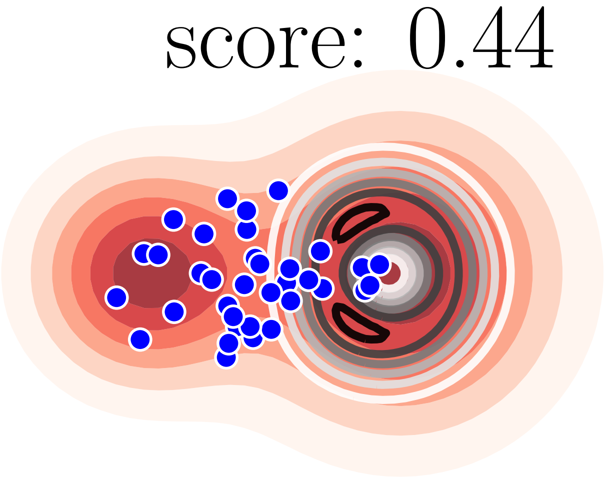
$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



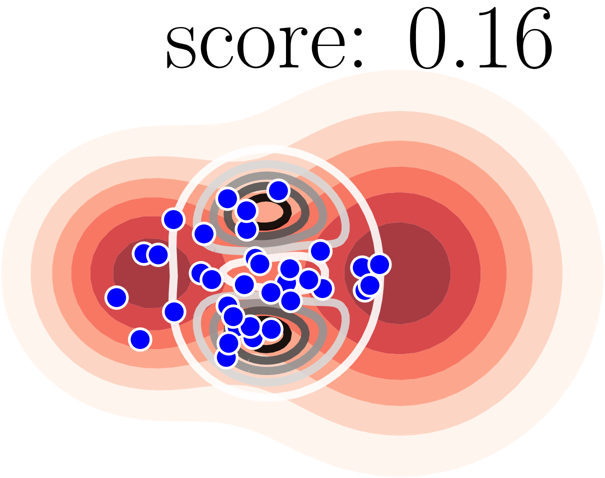
$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

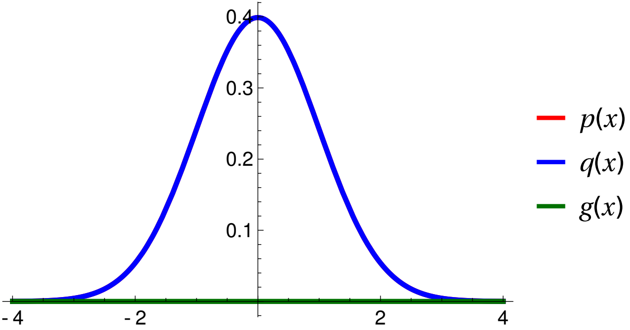
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{Stein Witness}^2(\mathbf{v})}{\text{uncertainty}(\mathbf{v})}.$$

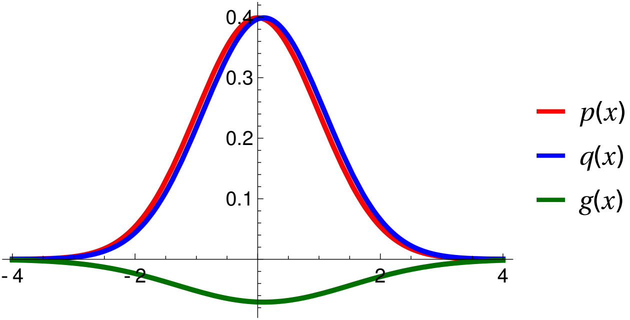
Proposal: The Finite Set Stein Discrepancy (FSSD)

- Stein witness function: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



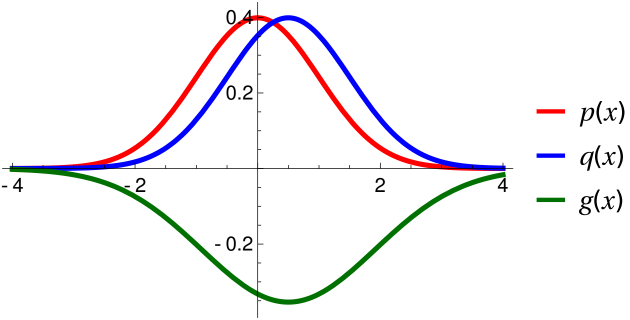
Proposal: The Finite Set Stein Discrepancy (FSSD)

- Stein witness function: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



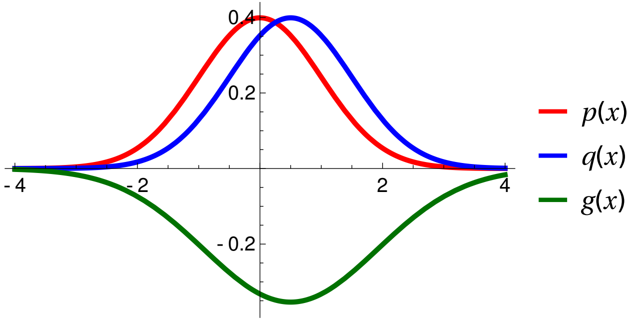
Proposal: The Finite Set Stein Discrepancy (FSSD)

- Stein witness function: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right].$



Proposal: The Finite Set Stein Discrepancy (FSSD)

- Stein witness function: $\mathbf{g}(\mathbf{v}) := \mathbb{E}_{\mathbf{x}\sim q}\left[\frac{1}{p(\mathbf{x})}\frac{d}{d\mathbf{x}}[k_{\mathbf{v}}(\mathbf{x})p(\mathbf{x})]\right]$.



- **FSSD statistic**: Evaluate g^2 at J test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$.

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

FSSD is a Discrepancy Measure

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

Theorem 1 (FSSD is a discrepancy measure).

Main conditions:

1. *(Nice kernel) Kernel k is C_0-universal, and real analytic e.g., Gaussian kernel.*
2. *(Vanishing boundary)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})k_{\mathbf{v}}(\mathbf{x}) = \mathbf{0}$.
3. *(Avoid "blind spots") Locations $\mathbf{v}_1, \ldots, \mathbf{v}_J \sim \eta$ which has a density.*

Then, for any $J \geq 1$, η-almost surely,

$$\boxed{\text{FSSD}^2 = 0 \iff p = q.}$$

Summary: Evaluating the witness at random locations is sufficient to detect the discrepancy between p, q.

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

What is $T_p k_{\mathbf{v}}$?

Recall $\mathrm{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right]$$

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})} \frac{d}{d\mathbf{y}} [k(\mathbf{y}, \mathbf{v}) p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] = \int_{-\infty}^{\infty} \left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] p(\mathbf{y}) d\mathbf{y}$$

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] = \int_{-\infty}^{\infty}\left[\frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\right]p(\mathbf{y})\,\mathrm{d}\mathbf{y}$$

What is $T_p k_{\mathbf{v}}$?

Recall $\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})} \frac{d}{d\mathbf{y}} [k(\mathbf{y}, \mathbf{v}) p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0$.

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] = \int_{-\infty}^{\infty} \left[\frac{1}{\cancel{p(\mathbf{y})}} \frac{d}{d\mathbf{y}} [k_{\mathbf{v}}(\mathbf{y}) p(\mathbf{y})]\right] \cancel{p(\mathbf{y})}\, d\mathbf{y}$$

What is $T_p k_{\mathbf{v}}$?

Recall witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0$.

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

Proof:

$$\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] = \int_{-\infty}^{\infty}\left[\frac{1}{\cancel{p(\mathbf{y})}}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\right]\cancel{p(\mathbf{y})}\,d\mathbf{y}$$
$$= \int_{-\infty}^{\infty}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\,d\mathbf{y}$$

What is $T_p k_{\mathbf{v}}$?

Recall witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0.$

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

Proof:

$$\begin{aligned}\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] &= \int_{-\infty}^{\infty}\left[\frac{1}{\cancel{p(\mathbf{y})}}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\right]\cancel{p(\mathbf{y})}\,\mathrm{d}\mathbf{y} \\ &= \int_{-\infty}^{\infty}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\,\mathrm{d}\mathbf{y} \\ &= [k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]_{\mathbf{y}=-\infty}^{\mathbf{y}=\infty}\end{aligned}$$

What is $T_p k_{\mathbf{v}}$?

Recall witness$(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}(T_p k_{\mathbf{v}})(\mathbf{x}) - \cancel{\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y})}$

$$(T_p k_{\mathbf{v}})(\mathbf{y}) = \frac{1}{p(\mathbf{y})}\frac{d}{d\mathbf{y}}[k(\mathbf{y},\mathbf{v})p(\mathbf{y})].$$

Normalizer cancels

Then, $\mathbb{E}_{\mathbf{y}\sim p}(T_p k_{\mathbf{v}})(\mathbf{y}) = 0$.

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

Proof:

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim p}\left[(T_p k_{\mathbf{v}})(\mathbf{y})\right] &= \int_{-\infty}^{\infty}\left[\frac{1}{\cancel{p(\mathbf{y})}}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\right]\cancel{p(\mathbf{y})}\,d\mathbf{y} \\
&= \int_{-\infty}^{\infty}\frac{d}{d\mathbf{y}}[k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]\,d\mathbf{y} \\
&= [k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})]_{\mathbf{y}=-\infty}^{\mathbf{y}=\infty} \\
&= 0
\end{aligned}$$

(assume $\lim_{|\mathbf{y}|\to\infty} k_{\mathbf{v}}(\mathbf{y})p(\mathbf{y})$)

Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approxeq$ rate of p-value $\to 0$ under H_1 as $n \to \infty$.
- Measure a test's sensitivity to the departure from H_0.

$$H_0 \colon \theta = \mathbf{0},$$
$$H_1 \colon \theta \neq \mathbf{0}.$$

- Typically $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2}c(\theta)n\right)$ where $c(\theta) > 0$ under H_1, and $c(\mathbf{0}) = 0$ [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

$$c(\theta) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

where $F(t) =$ CDF of T_n under H_0.

- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approxeq$ rate of p-value $\to 0$ under H_1 as $n \to \infty$.
- Measure a test's sensitivity to the departure from H_0.

$$H_0 \colon \boldsymbol{\theta} = \mathbf{0},$$
$$H_1 \colon \boldsymbol{\theta} \neq \mathbf{0}.$$

- Typically $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2} c(\boldsymbol{\theta}) n\right)$ where $c(\boldsymbol{\theta}) > 0$ under H_1, and $c(\mathbf{0}) = 0$ [Bahadur, 1960].
- $c(\boldsymbol{\theta})$ higher $\implies$ more sensitive. Good.

Bahadur slope

$$c(\boldsymbol{\theta}) := -2 \operatorname*{plim}_{n \to \infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

where $F(t) = $ CDF of T_n under H_0.

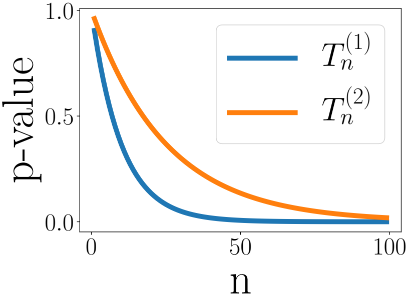
- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approxeq$ rate of p-value $\to 0$ under H_1 as $n \to \infty$.
- Measure a test's sensitivity to the departure from H_0.

$$H_0\colon \theta = \mathbf{0},$$
$$H_1\colon \theta \neq \mathbf{0}.$$

- Typically $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2} c(\theta) n\right)$ where $c(\theta) > 0$ under H_1, and $c(\mathbf{0}) = 0$ [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.



Bahadur slope

$$c(\theta) := -2 \operatorname*{plim}_{n\to\infty} \frac{\log\left(1 - F(T_n)\right)}{n},$$

where $F(t) = $ CDF of T_n under H_0.

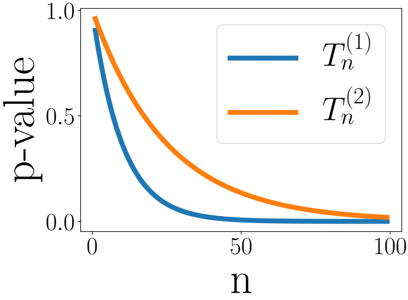
- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slope and Bahadur Efficiency

- Bahadur slope $\approxeq$ rate of p-value $\to 0$ under H_1 as $n \to \infty$.
- Measure a test's sensitivity to the departure from H_0.

$$H_0 \colon \boldsymbol{\theta} = \mathbf{0},$$
$$H_1 \colon \boldsymbol{\theta} \neq \mathbf{0}.$$

- Typically $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2} c(\boldsymbol{\theta}) n\right)$ where $c(\boldsymbol{\theta}) > 0$ under H_1, and $c(\mathbf{0}) = 0$ [Bahadur, 1960].
- $c(\boldsymbol{\theta})$ higher $\implies$ more sensitive. Good.



Bahadur slope

$$c(\boldsymbol{\theta}) := -2 \operatorname*{plim}_{n \to \infty} \frac{\log(1 - F(T_n))}{n},$$

where $F(t)$ = CDF of T_n under H_0.

- Bahadur efficiency = ratio of slopes of two tests.

Gaussian Mean Shift Problem

Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$.

- Assume $J = 1$ location for $n\widehat{\mathrm{FSSD}^2}$. Gaussian kernel (bandwidth $= \sigma_k^2$)

$$c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2) = \frac{\sigma_k^2\left(\sigma_k^2+2\right)^3 \mu_q^2 e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+1}}}{\sqrt{\frac{2}{\sigma_k^2}+1}\left(\sigma_k^2+1\right)\left(\sigma_k^6+4\sigma_k^4+\left(v^2+5\right)\sigma_k^2+2\right)}.$$

- For LKS, Gaussian kernel (bandwidth $= \kappa^2$).

$$c^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \frac{\left(\kappa^2\right)^{5/2}\left(\kappa^2+4\right)^{5/2} \mu_q^4}{2\left(\kappa^2+2\right)\left(\kappa^8+8\kappa^6+21\kappa^4+20\kappa^2+12\right)}.$$

Theorem 2 (FSSD is at least two times more efficient).

Fix $\sigma_k^2 = 1$ for $n\widehat{\mathrm{FSSD}^2}$. Then, $\forall \mu_q \neq 0$, $\exists v \in \mathbb{R}$, $\forall \kappa^2 > 0$, we have Bahadur efficiency

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

Gaussian Mean Shift Problem

Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$.

- Assume $J = 1$ location for $n\widehat{\mathrm{FSSD}^2}$. Gaussian kernel (bandwidth $= \sigma_k^2$)

$$c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2) = \frac{\sigma_k^2 \left(\sigma_k^2 + 2\right)^3 \mu_q^2 e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+1}}}{\sqrt{\frac{2}{\sigma_k^2} + 1}\left(\sigma_k^2 + 1\right)\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2 + 5\right)\sigma_k^2 + 2\right)}.$$

- For LKS, Gaussian kernel (bandwidth $= \kappa^2$).

$$c^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \frac{\left(\kappa^2\right)^{5/2}\left(\kappa^2 + 4\right)^{5/2} \mu_q^4}{2\left(\kappa^2 + 2\right)\left(\kappa^8 + 8\kappa^6 + 21\kappa^4 + 20\kappa^2 + 12\right)}.$$

Theorem 2 (FSSD is at least two times more efficient).

Fix $\sigma_k^2 = 1$ for $n\widehat{\mathrm{FSSD}^2}$. Then, $\forall \mu_q \neq 0$, $\exists v \in \mathbb{R}$, $\forall \kappa^2 > 0$, we have Bahadur efficiency

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

Gaussian Mean Shift Problem

Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, 1)$.

- Assume $J = 1$ location for $n\widehat{\mathrm{FSSD}^2}$. Gaussian kernel (bandwidth $= \sigma_k^2$)

$$c^{(\mathrm{FSSD})}(\mu_q, v, \sigma_k^2) = \frac{\sigma_k^2 \left(\sigma_k^2 + 2\right)^3 \mu_q^2 e^{\frac{v^2}{\sigma_k^2+2} - \frac{(v-\mu_q)^2}{\sigma_k^2+1}}}{\sqrt{\frac{2}{\sigma_k^2} + 1}\left(\sigma_k^2 + 1\right)\left(\sigma_k^6 + 4\sigma_k^4 + \left(v^2 + 5\right)\sigma_k^2 + 2\right)}.$$

- For LKS, Gaussian kernel (bandwidth $= \kappa^2$).

$$c^{(\mathrm{LKS})}(\mu_q, \kappa^2) = \frac{\left(\kappa^2\right)^{5/2}\left(\kappa^2 + 4\right)^{5/2} \mu_q^4}{2\left(\kappa^2 + 2\right)\left(\kappa^8 + 8\kappa^6 + 21\kappa^4 + 20\kappa^2 + 12\right)}.$$

Theorem 2 (FSSD is at least two times more efficient).

Fix $\sigma_k^2 = 1$ *for* $n\widehat{\mathrm{FSSD}^2}$. *Then,* $\forall \mu_q \neq 0$, $\exists v \in \mathbb{R}$, $\forall \kappa^2 > 0$, *we have Bahadur efficiency*

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

Bahadur Slopes of FSSD and LKS

Theorem 3.

The Bahadur slope of $n\widehat{\mathrm{FSSD}^2}$ *is*

$$c^{(\mathrm{FSSD})} := \mathrm{FSSD}^2/\omega_1,$$

where ω_1 *is the maximum eigenvalue of* $\Sigma_p := \mathrm{cov}_{\mathbf{x}\sim p}[\boldsymbol{\tau}(\mathbf{x})]$*.*
The Bahadur slope of the linear-time kernel Stein (LKS) statistic $\sqrt{n}\widehat{S_l^2}$ *is*

$$c^{(\mathrm{LKS})} = \frac{1}{2}\frac{\left[\mathbb{E}_q h_p(\mathbf{x}, \mathbf{x}')\right]^2}{\mathbb{E}_p\left[h_p^2(\mathbf{x}, \mathbf{x}')\right]},$$

where h_p *is the U-statistic kernel of the KSD statistic.*

Illustration: Optimization Objective

- Consider $J = 1$ location.
- Training objective $\frac{\widehat{\mathrm{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, ★ = best $\mathbf{v}$.

$$p = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \text{ vs. } q = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

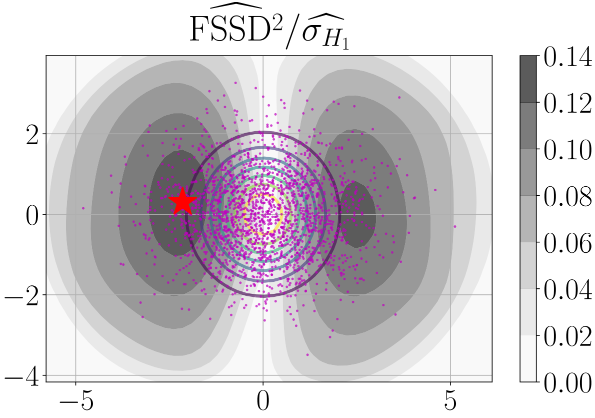
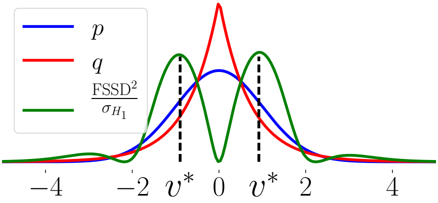
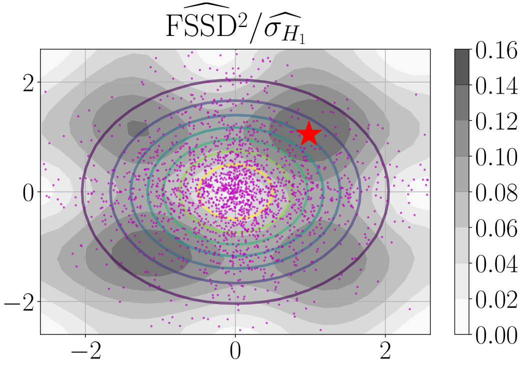


Illustration: Optimization Objective

- Consider $J = 1$ location.
- Training objective $\frac{\widehat{\mathrm{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, ★ = best $\mathbf{v}$.

$$p = \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ vs. } q = \text{Laplace with same mean \& variance.}$$





References I

Bahadur, R. R. (1960).
Stochastic comparison of tests.
The Annals of Mathematical Statistics, 31(2):276–295.