A Linear-Time Kernel Goodness-of-Fit Test

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





wittawat@gatsby.ucl.ac.uk

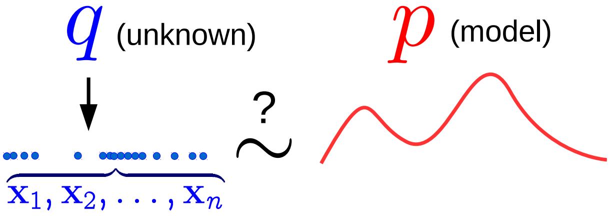
[1]Gatsby Unit, University College London
[2]CMAP, École Polytechnique
[3]The Institute of Statistical Mathematics, Tokyo

MLTrain Workshop: Learn How to Code a Paper
9 December 2017

Problem Setting: Goodness-of-Fit Test

$H_0: p = q$

$H_1: p \neq q$



The developed test:

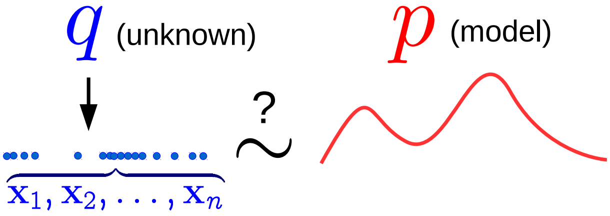
1. (Testing) Outputs "reject H_0" or "fail to reject H_0", and p-value.
2. If "reject H_0", shows a location $\mathbf{v}$ where the model does not fit well. Interpretable.

Runtime complexity is $\mathcal{O}(n)$. Fast.

Problem Setting: Goodness-of-Fit Test

$H_0: p = q$

$H_1: p \neq q$



The developed test:

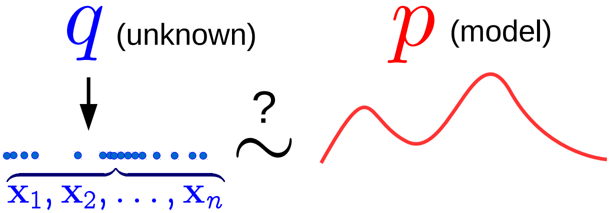
1. (Testing) Outputs "reject H_0" or "fail to reject H_0", and p-value.
2. If "reject H_0", shows a location $\mathbf{v}$ where the model does not fit well. Interpretable.

Runtime complexity is $\mathcal{O}(n)$. Fast.

Problem Setting: Goodness-of-Fit Test

$H_0: p = q$

$H_1: p \neq q$



The developed test:

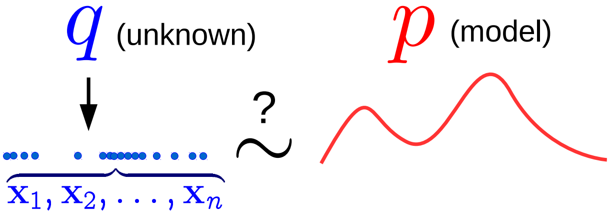
1. <u>(Testing)</u> Outputs "reject H_0" or "fail to reject H_0", and p-value.
2. <u>(Model criticism)</u> If "reject H_0", shows a location $\mathbf{v}$ where the model does not fit well. Interpretable.

Runtime complexity is $\mathcal{O}(n)$. <u>Fast</u>.

Problem Setting: Goodness-of-Fit Test

$H_0: p = q$

$H_1: p \neq q$



The developed test:

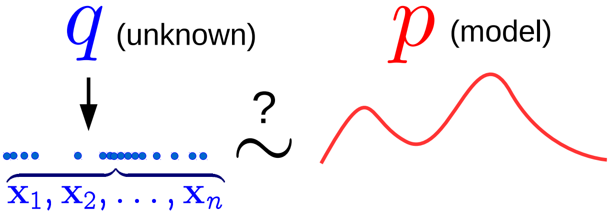
1. (Testing) Outputs "reject H_0" or "fail to reject H_0", and p-value.
2. (Model criticism) If "reject H_0", shows a location $\mathbf{v}$ where the model does not fit well. Interpretable.

Runtime complexity is $\mathcal{O}(n)$. **Fast**.

Problem Setting: Goodness-of-Fit Test

$H_0 : p = q$

$H_1 : p \neq q$

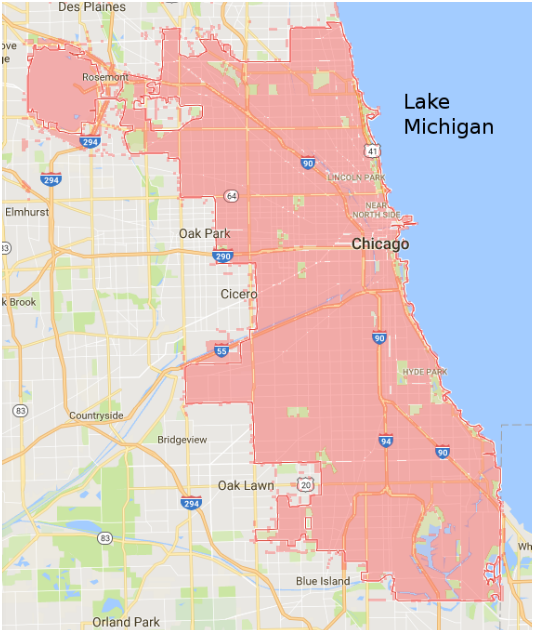


The developed test:

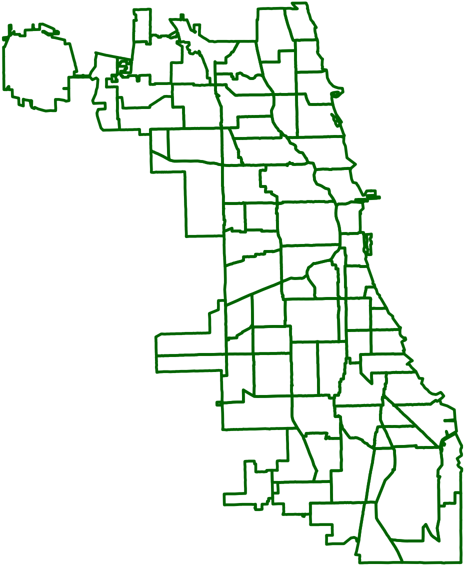
1. (Testing) Outputs "reject H_0" or "fail to reject H_0", and p-value.
2. **(Model criticism)** If "reject H_0", shows a location $\mathbf{v}$ where the model does not fit well. Interpretable.

Runtime complexity is $\mathcal{O}(n)$. **Fast**.

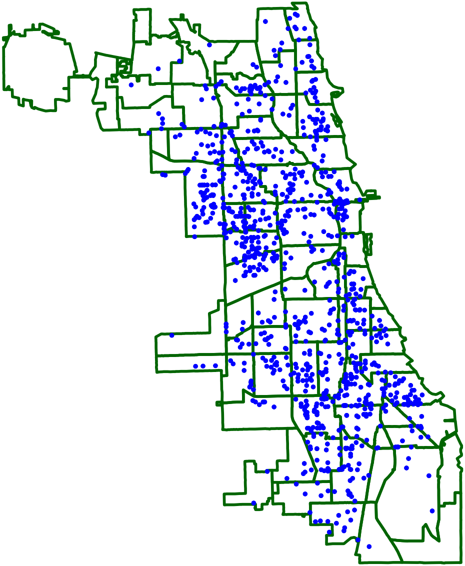
Interpretable Features: Chicago Crime



Interpretable Features: Chicago Crime

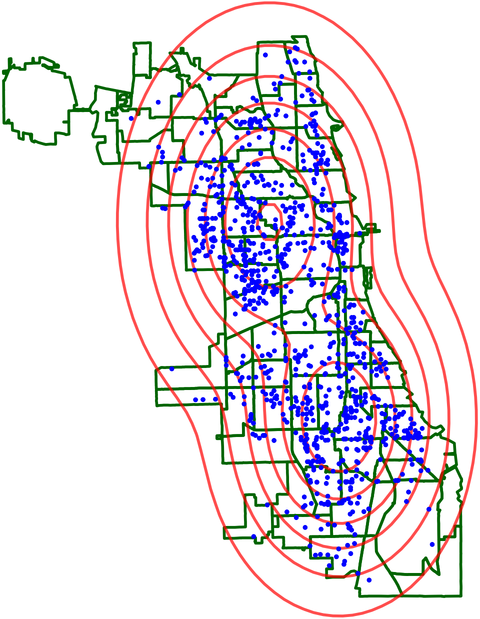


Interpretable Features: Chicago Crime



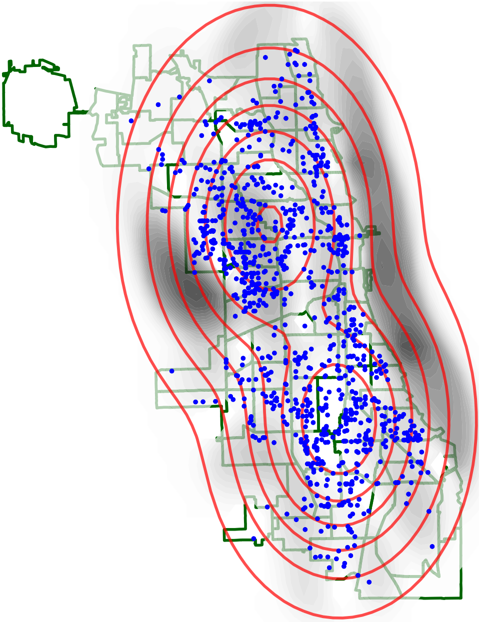
- $n = 11957$ robbery events in Chicago in 2016.
  - lat/long coordinates $=$ sample from q.
- Model spatial density with Gaussian mixtures.

Interpretable Features: Chicago Crime



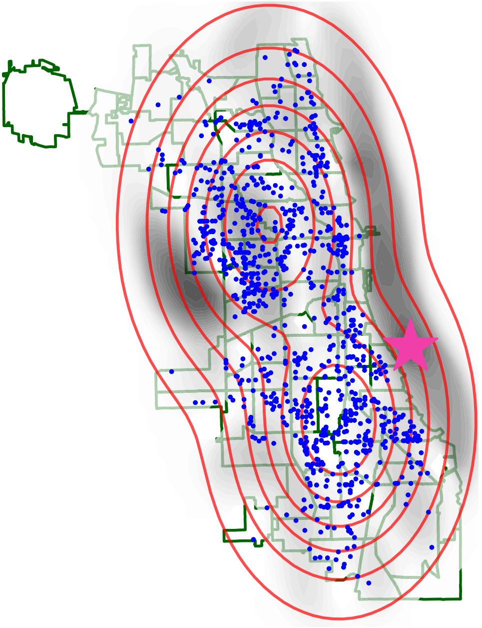
Model p = 2-component Gaussian mixture.

Interpretable Features: Chicago Crime



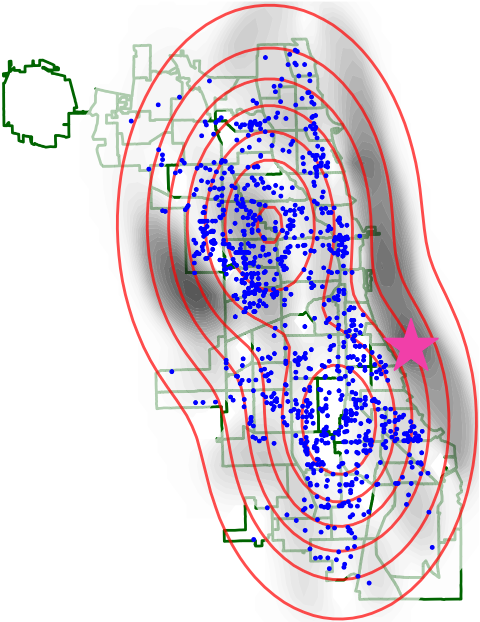
Score surface

Interpretable Features: Chicago Crime



★ = optimized $\mathbf{v}$.

Interpretable Features: Chicago Crime



★ = optimized $\mathbf{v}$.

No robbery in Lake Michigan.

Score Function for Model Criticism

Proposal: A good location $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

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

Goodness-of-fit test:

- Find $\mathbf{v}^* = \arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$.
- Use $\text{signal}^2(\mathbf{v}^*)$ as the test statistic.
- General form: $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$.

Score Function for Model Criticism

Proposal: A good location $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

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

Goodness-of-fit test:

- Find $\mathbf{v}^* = \arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$.
- Use $\text{signal}^2(\mathbf{v}^*)$ as the test statistic.
- General form: $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$.

Score Function for Model Criticism

Proposal: A good location $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

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

Goodness-of-fit test:

- Find $\mathbf{v}^* = \arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$.
- Use $\text{signal}^2(\mathbf{v}^*)$ as the test statistic.
- General form: $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$.

Score Function for Model Criticism

Proposal: A good location $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

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

Goodness-of-fit test:

- Find $\mathbf{v}^* = \arg\max_{\mathbf{v}} \text{score}(\mathbf{v})$.
- Use $\text{signal}^2(\mathbf{v}^*)$ as the test statistic.
- General form: $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$.

Demo

Use Jupyter notebook.

signal($\mathbf{v}$) and noise($\mathbf{v}$)

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})} = \frac{|\mathbb{E}_{\mathbf{x}\sim q}[\,T_p k_{\mathbf{v}}(\mathbf{x})]|}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[\,T_p k_{\mathbf{v}}(\mathbf{x})]}}.$$

where

$$T_p k_{\mathbf{v}}(\mathbf{x}) := k_{\mathbf{v}}(\mathbf{x}) \frac{d}{d\mathbf{x}} \log p(\mathbf{x}) + \frac{d}{d\mathbf{x}} k_{\mathbf{v}}(\mathbf{x}).$$

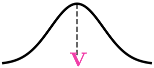
- $\frac{d}{d\mathbf{x}} \log p(\mathbf{x})$ does not depend on the normalizer.
-

signal($\mathbf{v}$) and noise($\mathbf{v}$)

$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})} = \frac{|\mathbb{E}_{\mathbf{x}\sim q}[T_p k_{\mathbf{v}}(\mathbf{x})]|}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[T_p k_{\mathbf{v}}(\mathbf{x})]}}.$$

where

$$T_p k_{\mathbf{v}}(\mathbf{x}) := k_{\mathbf{v}}(\mathbf{x}) \frac{d}{d\mathbf{x}} \log p(\mathbf{x}) + \frac{d}{d\mathbf{x}} k_{\mathbf{v}}(\mathbf{x}).$$

- $\frac{d}{d\mathbf{x}} \log p(\mathbf{x})$ does not depend on the normalizer.
- $k_{\mathbf{v}}(\mathbf{x}) =$  $=$ a kernel (e.g., Gaussian) centered at $\mathbf{v}$.

Model $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right).$$

$$\log p(\mathrm{x}) = -\frac{\|\mathrm{x}\|^2}{2} - \frac{d}{2}\log 2\pi.$$

$$\frac{d}{d\mathrm{x}} \log p(\mathrm{x}) = -\mathrm{x}.$$

- In the implementation, only need to specify $\tilde{p}(\mathrm{x}) = \exp\left(-\frac{\|\mathrm{x}-\mu\|^2}{2}\right)$.
- autograd automatically computes $\frac{d}{d\mathrm{x}} \log p(\mathrm{x})$.

Model $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right).$$

$$\log p(\mathbf{x}) = -\frac{\|\mathbf{x}\|^2}{2} - \frac{d}{2}\log 2\pi.$$

$$\frac{d}{d\mathrm{x}} \log p(\mathrm{x}) = -\mathrm{x}.$$

- In the implementation, only need to specify $\tilde{p}(\mathrm{x}) = \exp\left(-\frac{\|\mathrm{x}-\mu\|^2}{2}\right)$.
- autograd automatically computes $\frac{d}{d\mathrm{x}} \log p(\mathrm{x})$.

Model $p = \mathcal{N}(\mathbf{0}, \mathbf{I})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right).$$

$$\log p(\mathbf{x}) = -\frac{\|\mathbf{x}\|^2}{2} - \frac{d}{2}\log 2\pi.$$

$$\frac{d}{d\mathbf{x}} \log p(\mathbf{x}) = -\mathbf{x}.$$

- In the implementation, only need to specify $\tilde{p}(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}-\mu\|^2}{2}\right)$.
- autograd automatically computes $\frac{d}{d\mathbf{x}} \log p(\mathbf{x})$.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
   - Two locations can collapse to the same point.
   - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
   - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
   - Two locations can collapse to the same point.
   - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
   - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\mathrm{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
    - Two locations can collapse to the same point.
    - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
    - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
   - Two locations can collapse to the same point.
   - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
   - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
   - Two locations can collapse to the same point.
   - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
   - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Limitations and Technical Conditions

Some limitations (that can be fixed in future work).

1. $\text{score}(\mathbf{v}_1, \ldots, \mathbf{v}_J)$ does not penalize locations that are too close to each other.
   - Two locations can collapse to the same point.
   - **Solution**: Use a normalized statistic [Jitkrittum et al., 2016]. Explicit penalty.
2. (Vanishing boundary condition) Require $\lim_{\|\mathbf{x}\| \to \infty} k(\mathbf{x}, \mathbf{v}) p(\mathbf{x}) = 0$ for any $\mathbf{v}$.
   - Require the domain to be full $\mathbb{R}^d$ in many cases.
3. Optimizing $\{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ jointly by gradient ascent may not be the best way.

Conclusions

- A new discrepancy measure between a density p and a dataset.

Proposed a new goodness-of-fit test.

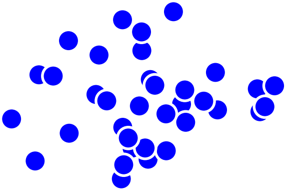
1. Can be applied to a wide range of models p.
2. Linear-time. Fast.
3. Interpretable.

Python code: https://github.com/wittawatj/kernel-gof

Questions?

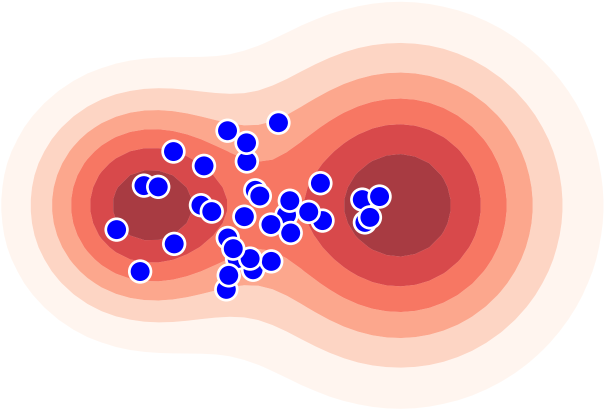
Thank you

Proposal: Model Criticism with the Score



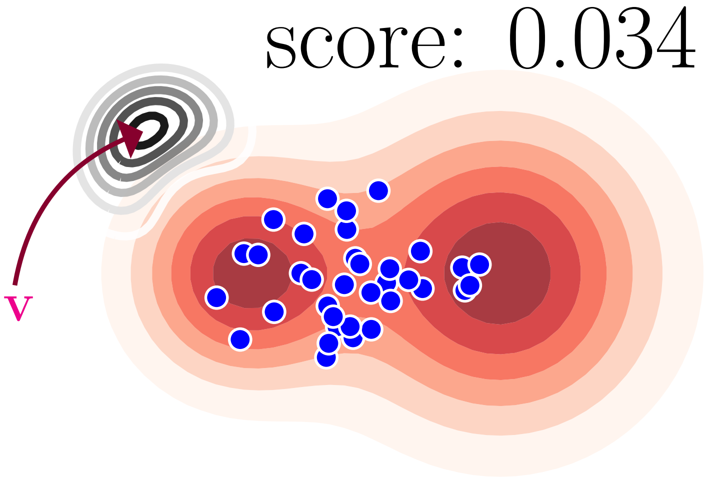
$$\mathrm{score}(\mathbf{v}) = \frac{|\mathrm{signal}(\mathbf{v})|}{\mathrm{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



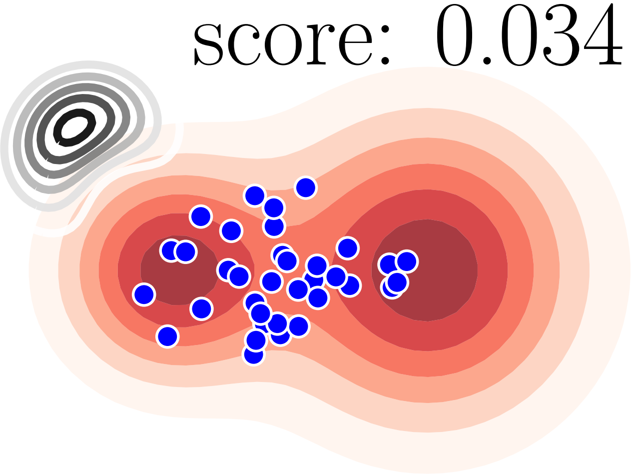
$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



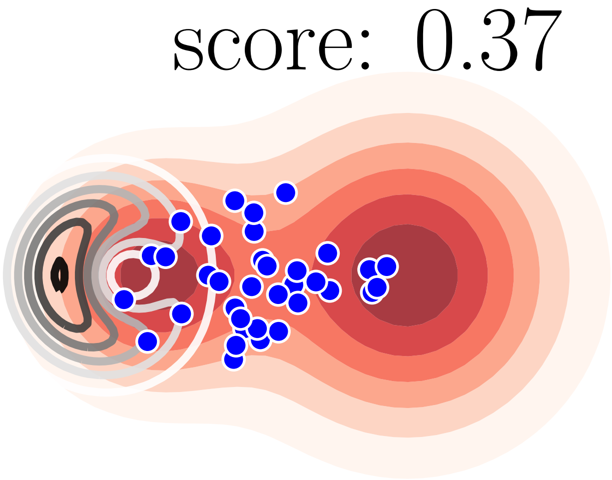
$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



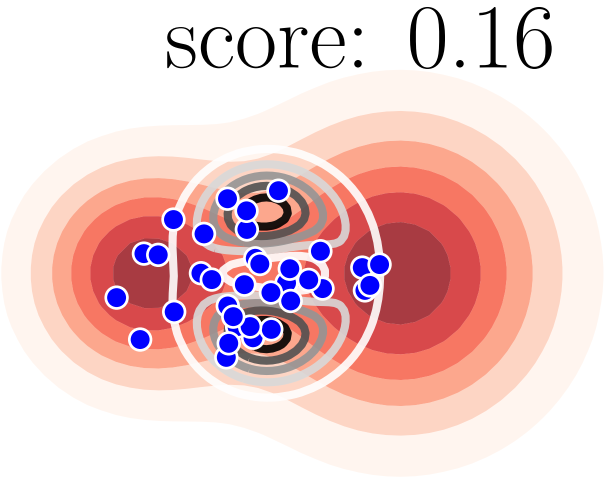
$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



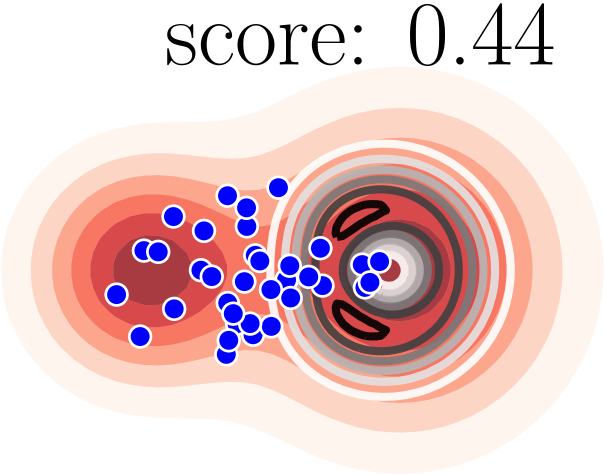
$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Score



$$\text{score}(\mathbf{v}) = \frac{|\text{signal}(\mathbf{v})|}{\text{noise}(\mathbf{v})}.$$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

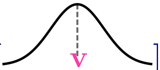
Model Criticism by Maximum Mean Discrepancy [?]

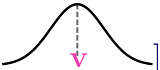
- Find a location $\mathbf{v}$ at which q and p differ most [?].

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

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

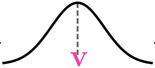
$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[$ 

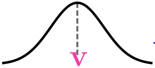
$] - \mathbb{E}_{\mathbf{y}\sim p}[$ 

$]$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

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

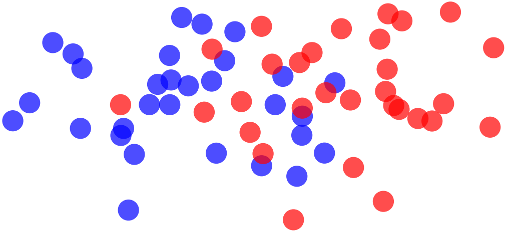
$\Big] - \mathbb{E}_{\mathbf{y}\sim p}\Big[$ 

$\Big]$

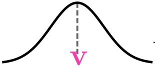
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].



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

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

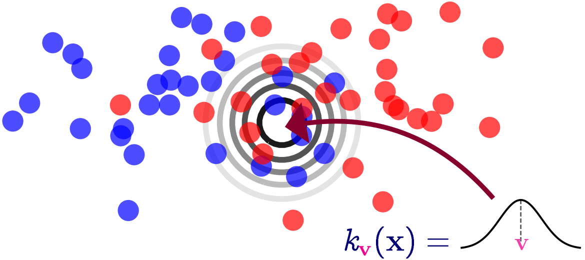
$]$

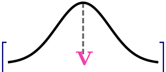
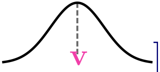
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

score: 0.008



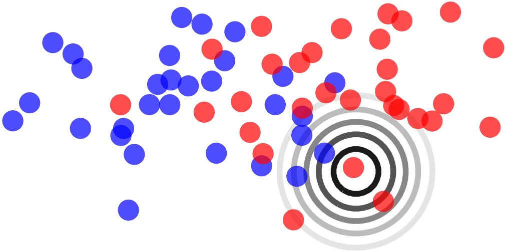
$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}\Big[$  $\Big] - \mathbb{E}_{\mathbf{y}\sim p}\Big[$  $\Big]$

$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

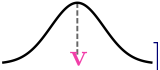
Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

score: 1.6



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

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

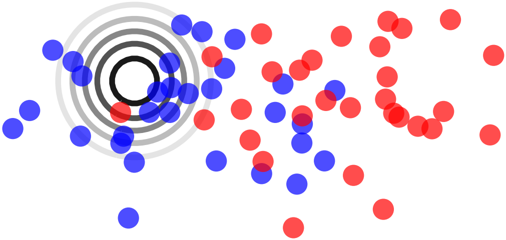
$]$

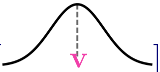
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

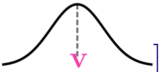
Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

score: 13



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

] − $\mathbb{E}_{\mathbf{y}\sim p}$[

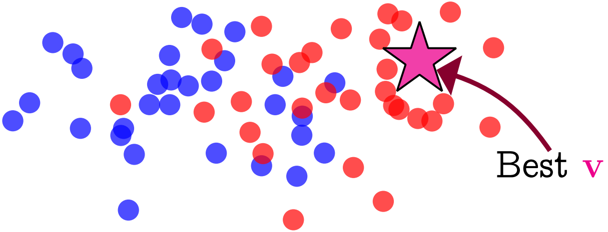
]

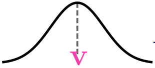
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

score: 25



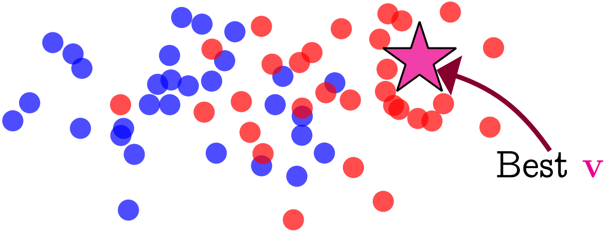
$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[$  $] - \mathbb{E}_{\mathbf{y}\sim p}[$  $]$

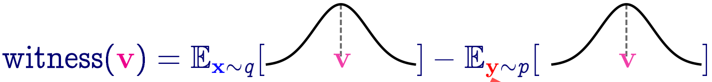
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Model Criticism by Maximum Mean Discrepancy [?]

- Find a location $\mathbf{v}$ at which q and p differ most [?].

score: 25





$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

No sample from p.
Difficult to generate.

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

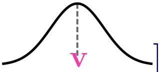
The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

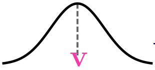
Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]$.

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

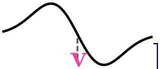
(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{x}\sim q}[\, T_p$ 

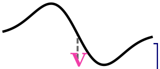
$] - \mathbb{E}_{\mathbf{y}\sim p}[\, T_p$ 

$]$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

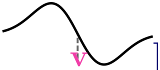
(Stein) witness($\mathbf{v}$) $= \mathbb{E}_{\mathbf{x}\sim q}[$ 

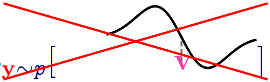
$] - \mathbb{E}_{\mathbf{y}\sim p}[$ 

$]$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

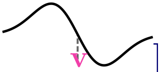
$] - \mathbb{E}_{\mathbf{y}\sim p}[$ 

$]$

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

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

$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

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

$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

signal-to-noise ratio

The Stein Witness Function [Liu et al., 2016, Chwialkowski et al., 2016]

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

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

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

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

$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

signal-to-noise ratio

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

FSSD is a Discrepancy Measure

Theorem 1.

Let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$ *be drawn i.i.d. from a distribution* η *which has a density. Let* $\mathcal{X}$ *be a connected open set in* $\mathbb{R}^d$*. Assume*

1. *(Nice RKHS) Kernel* $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is* C_0*-universal, and real analytic.*
2. *(Stein witness not too rough)* $\|g\|_{\mathcal{F}}^2 < \infty$*.*
3. *(Finite Fisher divergence)* $\mathbb{E}_{\mathbf{x}\sim q}\|\nabla_{\mathbf{x}} \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\|^2 < \infty$ *.*
4. *(Vanishing boundary)* $\lim_{\|\mathbf{x}\|\to\infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$*.*

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

$$\mathrm{FSSD}^2 = 0 \textit{ if and only if } p = q.$$

- Gaussian kernel $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$ works.
- In practice, $J = 1$ or $J = 5$.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ where $\sigma_{H_1}^2 := 4\mu^\top\Sigma_q\mu$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

But, how to estimate Σ_p? No sample from p!

- **Theorem**: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, and $\{\omega_i\}_{i=1}^{dJ}$ be the eigenvalues of Σ_p.

1. *Under $H_0 : p = q$, asymptotically $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{H_1})$ where $\sigma^2_{H_1} := 4\mu^\top\Sigma_q\mu$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

But, how to estimate Σ_p? No sample from p!

- **Theorem**: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})}\partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})p(\mathbf{x})] \in \mathbb{R}^d$.
- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, *and* $\{\omega_i\}_{i=1}^{dJ}$ *be the eigenvalues of* Σ_p.

1. *Under* $H_0 : p = q$, *asymptotically* $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of* n.
2. *Under* $H_1 : p \neq q$, *we have* $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{H_1})$ *where* $\sigma^2_{H_1} := 4\boldsymbol{\mu}^\top\Sigma_q\boldsymbol{\mu}$. *Implies* $\mathbb{P}$(*reject* H_0) $\to 1$ *as* $n \to \infty$.

But, how to estimate Σ_p? No sample from p!

- **Theorem**: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

Asymptotic Distributions of $\widehat{\mathrm{FSSD}^2}$

- Recall $\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})} \partial_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v}) p(\mathbf{x})] \in \mathbb{R}^d$.
- $\tau(\mathbf{x}) :=$ vertically stack $\xi(\mathbf{x}, \mathbf{v}_1), \ldots \xi(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\mu := \mathbb{E}_{\mathbf{x} \sim q}[\tau(\mathbf{x})]$.
- $\Sigma_r := \mathrm{cov}_{\mathbf{x} \sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ \times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, *and* $\{\omega_i\}_{i=1}^{dJ}$ *be the eigenvalues of* Σ_p.

1. *Under* $H_0 : p = q$, *asymptotically* $n\widehat{\mathrm{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of* n.
2. *Under* $H_1 : p \neq q$, *we have* $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ *where* $\sigma_{H_1}^2 := 4\mu^\top \Sigma_q \mu$. *Implies* $\mathbb{P}(\textit{reject } H_0) \to 1$ *as* $n \to \infty$.

But, how to estimate Σ_p? No sample from p!

- **Theorem**: Using $\hat{\Sigma}_q$ (computed with $\{\mathbf{x}_i\}_{i=1}^n \sim q$) still leads to a consistent test.

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 $\mathbf{v}$.

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

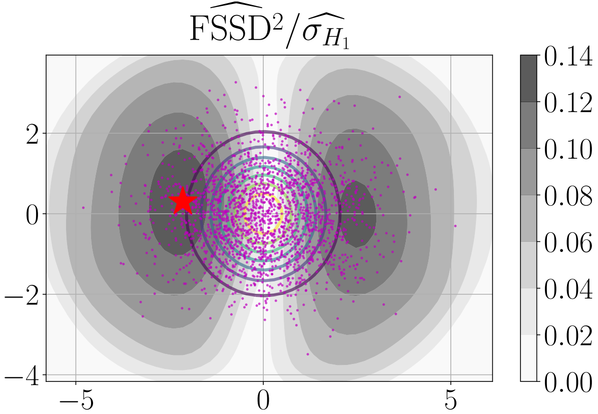
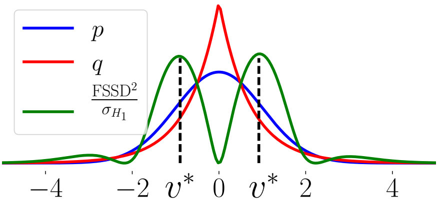
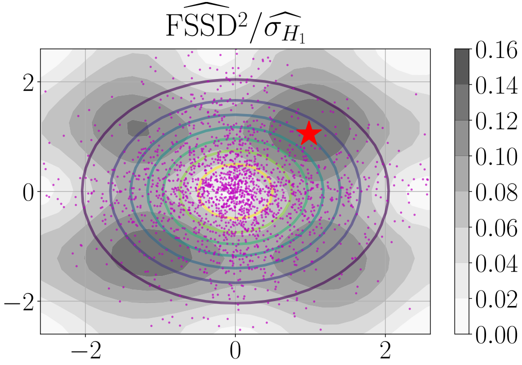


Illustration: Optimization Objective

- 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