Recent Advances in Kernel Methods for Model Criticism

Wittawat Jitkrittum

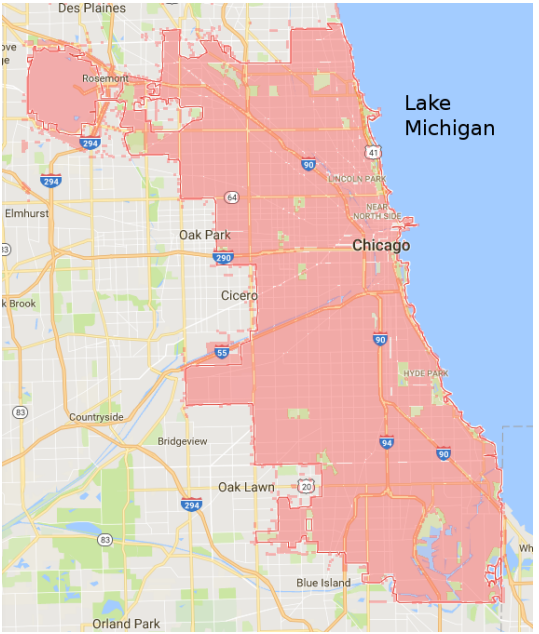
wittawat.com

Max Planck Institute for Intelligent Systems

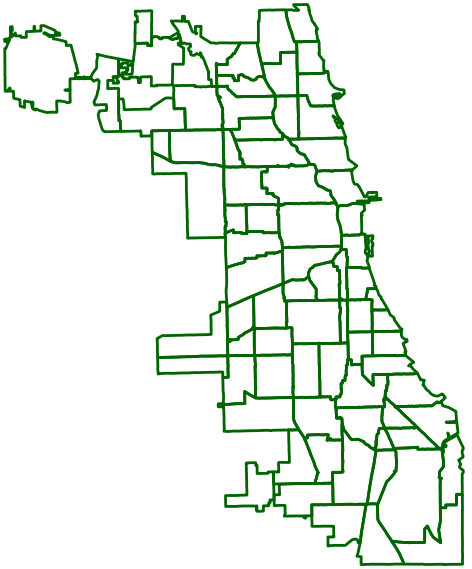
Symposium on Frontier Research in Information Science and Technology
At VISTEC

20 December 2018

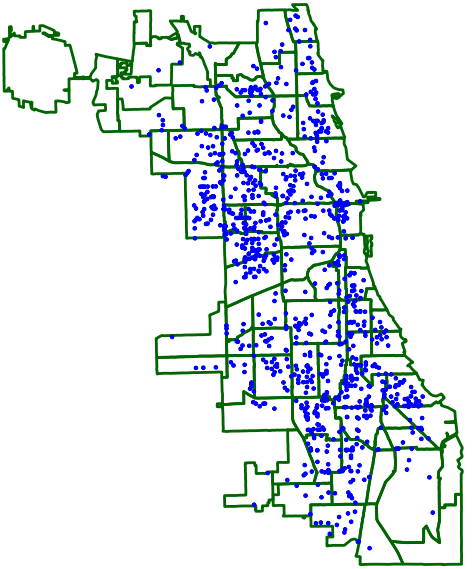
Example of Model Criticism



Example of Model Criticism

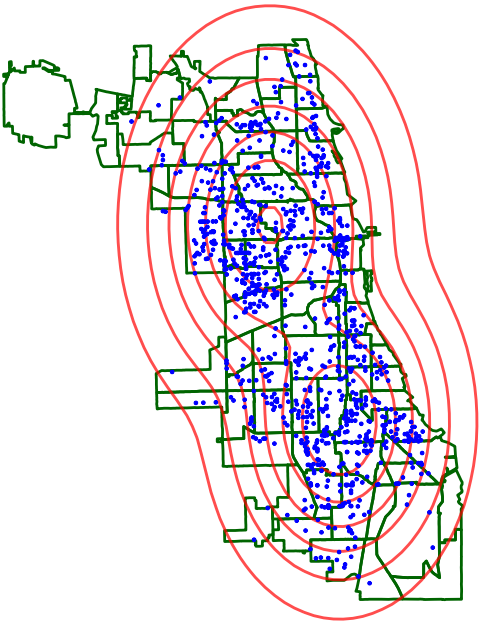


Example of Model Criticism



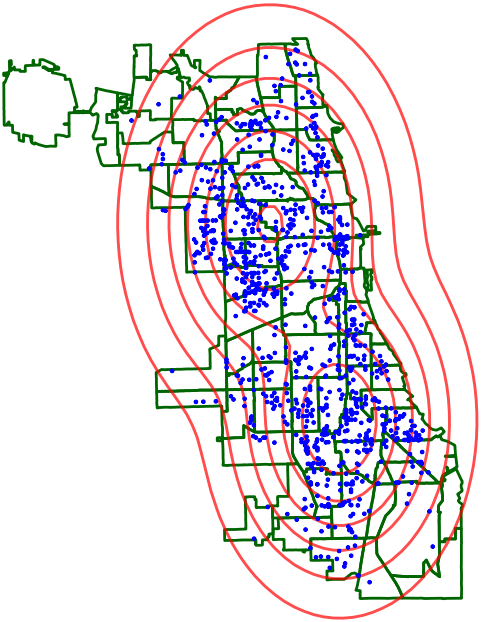
Data = robbery events in Chicago in 2016.

Example of Model Criticism



Is this a good model?

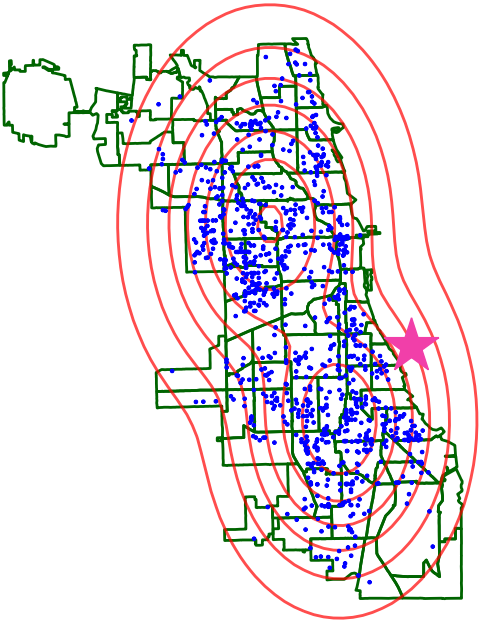
Example of Model Criticism



Goals:

1. Determine if a (complicated) model fits the data.
2. If it does not, show a location where it fails.

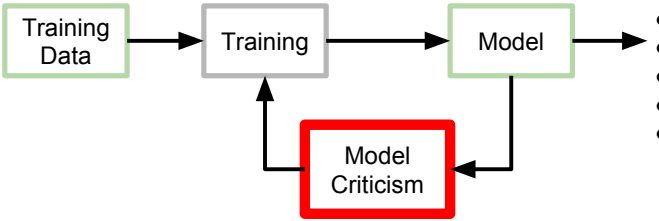
Example of Model Criticism



Goals:

1. Determine if a (complicated) model fits the data.
2. If it does not, show a location where it fails.

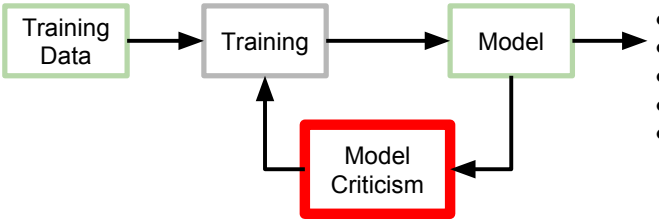
Machine Learning Pipeline



- Language model
- Image generator
- Denoiser
- Crime rate model
- etc.

- Arguably all machine learning tasks build models of some sort.
- Model criticism is of obvious value.
- Kernel-based techniques make nonparametric model criticism possible.

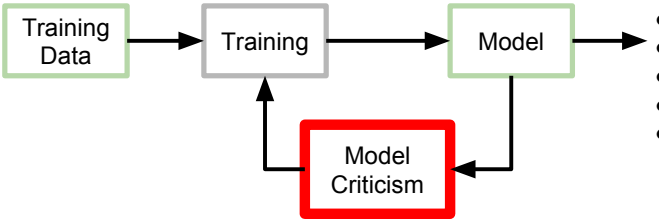
Machine Learning Pipeline



- Language model
- Image generator
- Denoiser
- Crime rate model
- etc.

- Arguably all machine learning tasks build models of some sort.
- Model criticism is of obvious value.
- Kernel-based techniques make nonparametric model criticism possible.

Machine Learning Pipeline

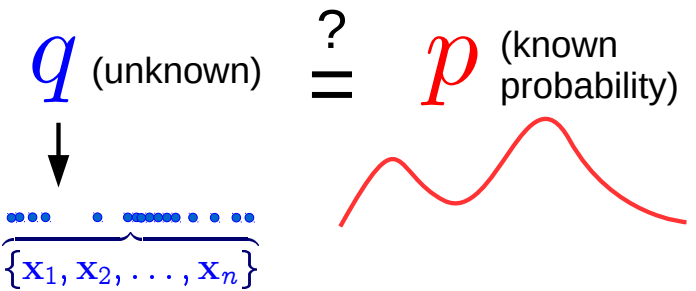


- Arguably all machine learning tasks build models of some sort.
- Model criticism is of obvious value.
- Kernel-based techniques make nonparametric model criticism possible.

Formulation for Model Criticism (Setting #1)

Given:

1. A sample $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$ (unknown) in $\mathbb{R}^d$,
2. Unnormalized density p (known model).

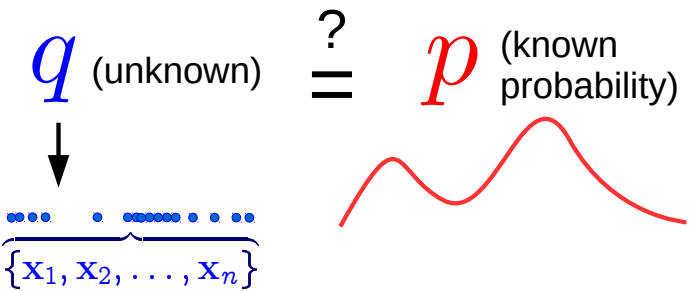


- Check whether $p = q$. Also known as goodness-of-fit testing.
- Various applications in many domains e.g., images, text, tabular data.

Formulation for Model Criticism (Setting #1)

Given:

1. A sample $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$ (unknown) in $\mathbb{R}^d$,
2. Unnormalized density p (known model).

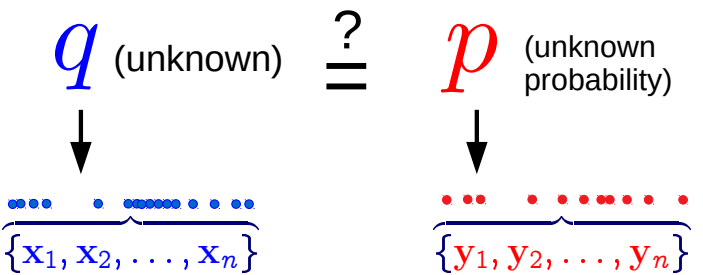


- Check whether $p = q$. Also known as goodness-of-fit testing.
- Various applications in many domains e.g., images, text, tabular data.

Model Criticism by Samples (Setting #2)

Given:

1. A sample $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} q$ (unknown) on $\mathbb{R}^d$,
2. A sample $\{\mathbf{y}_i\}_{i=1}^n \overset{i.i.d.}{\sim} p$ (implicit) on $\mathbb{R}^d$.



- Do not know exact form of model p. Can only sample.
- Also known as two-sample testing.

Application 1: Diagnose Generative Adversarial Nets (GANs)

Real images (CelebA HQ)

Generated from Progressive GAN p

[Karras et al., 2018]

GAN:

$z \sim p_0(z)$ (latent code)

$y = g(z)$ (generate an image)

where g is a deep net. Density p not available.

- Are real and generated images drawn from the same distribution?

Application 2: Learn Distinguishing Features

- **Have:** Two samples X, Y from unknown distributions q and p.

Positive emotions

$X = \{$  ,  ,  $, \ldots\} \sim q$

Negative emotions

$Y = \{$  ,  ,  $, \ldots\} \sim p$

- Learn distinguishing location (face) that indicates where q and p differ.

Application 2: Learn Distinguishing Features

- **Have:** Two samples X, Y from unknown distributions q and p.

Positive emotions

$X = \{$    $, \ldots\} \sim q$

Negative emotions

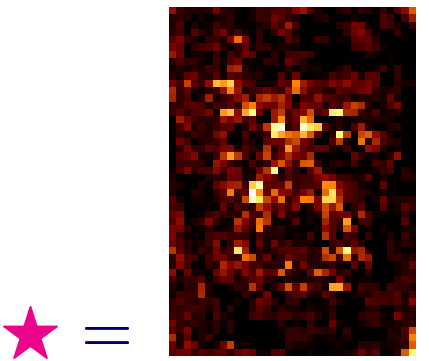
$Y = \{$    $, \ldots\} \sim p$

- Learn distinguishing location (face) that indicates where q and p differ.

Application 2: Learn Distinguishing Features

From the two collections

produce a new point indicating where to look for the differences

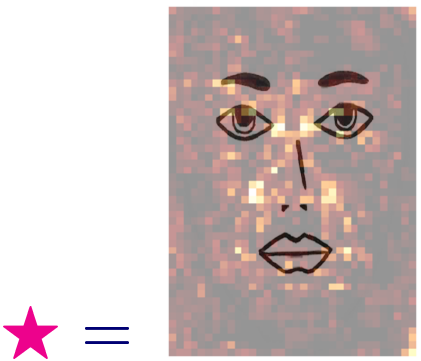


Application 2: Learn Distinguishing Features

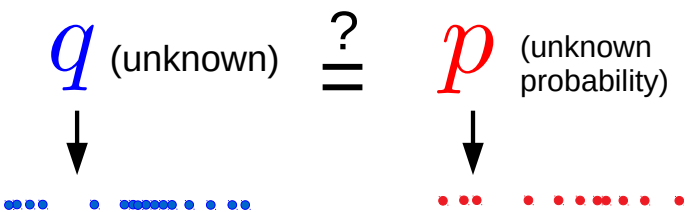
From the two collections

and

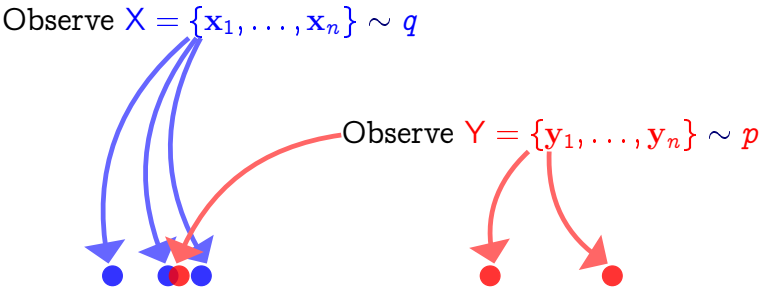

produce a new point indicating where to look for the differences



Model Criticism for Two Samples (Setting #2)



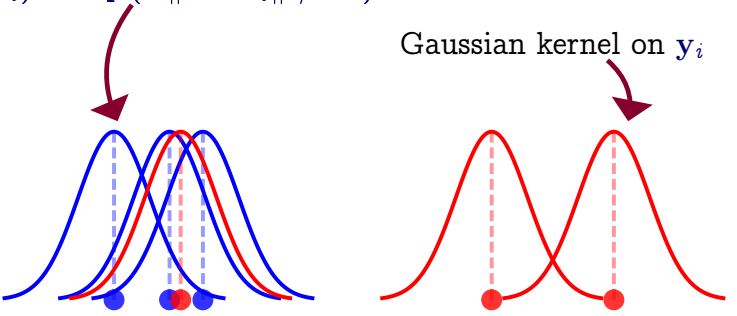
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)


Observe $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim q$
Observe $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim p$

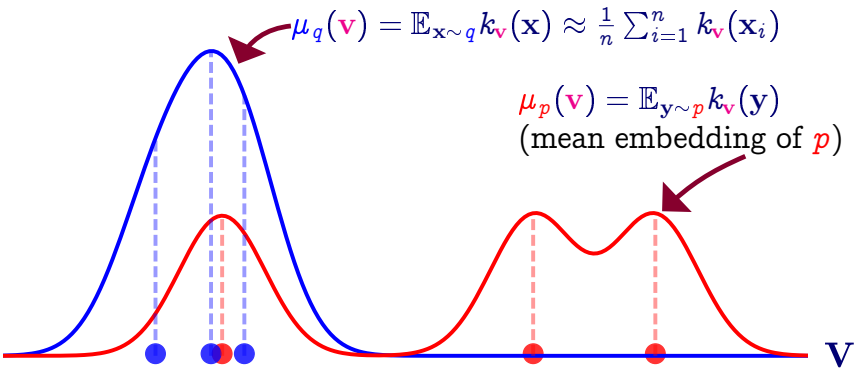
Gaussian kernel on $\mathbf{x}_i$.

$k_{\mathbf{v}}(\mathbf{x}_i) = \exp\left(-\|\mathbf{v} - \mathbf{x}_i\|^2/2\sigma^2\right)$

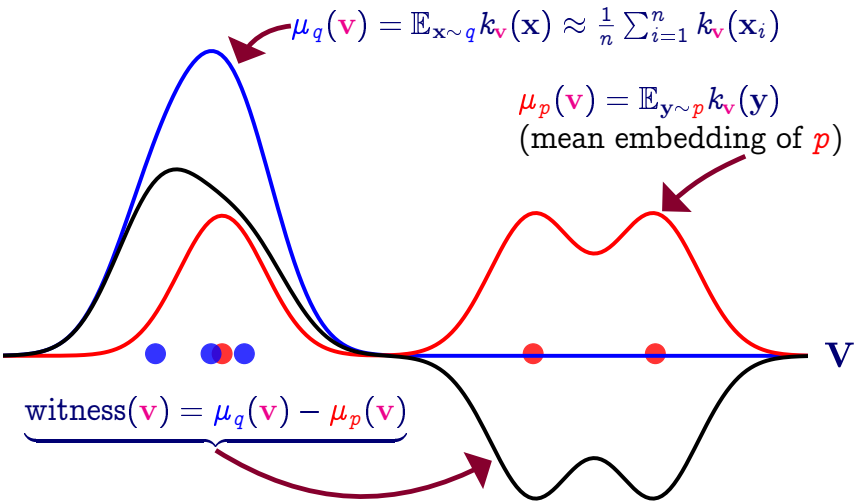
Gaussian kernel on $\mathbf{y}_i$



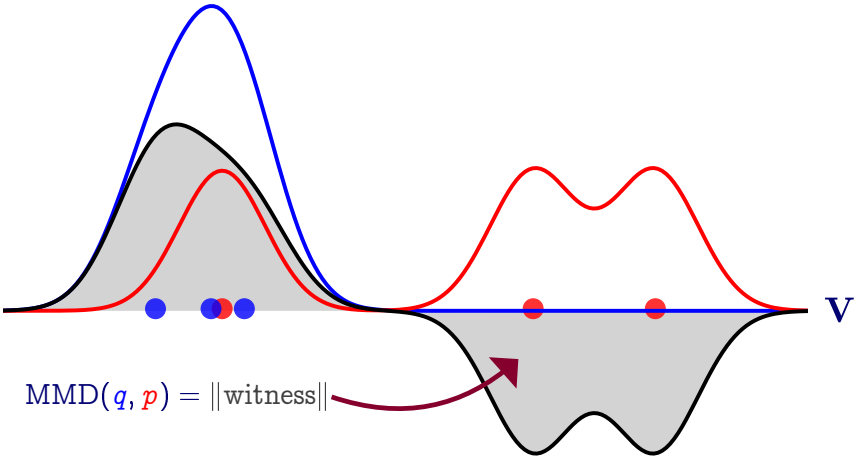
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



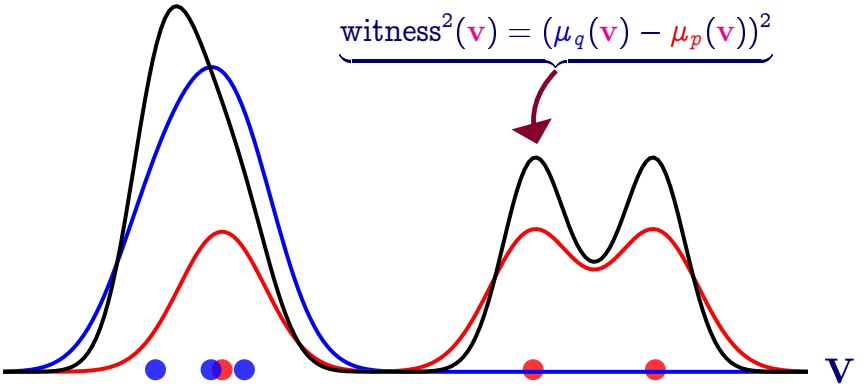
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



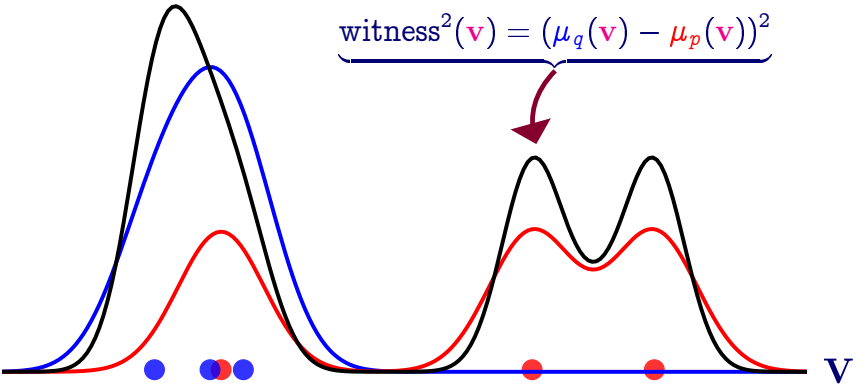
Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



Maximum Mean Discrepancy (MMD) Witness Function (Gretton et al., 2012)



- witness$^2(\mathbf{v})$ can be used to find a good test location $\mathbf{v}^* =$ ★ .

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].

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

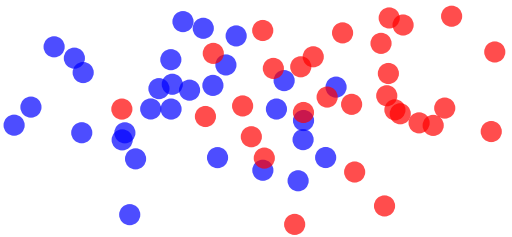
Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].

$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

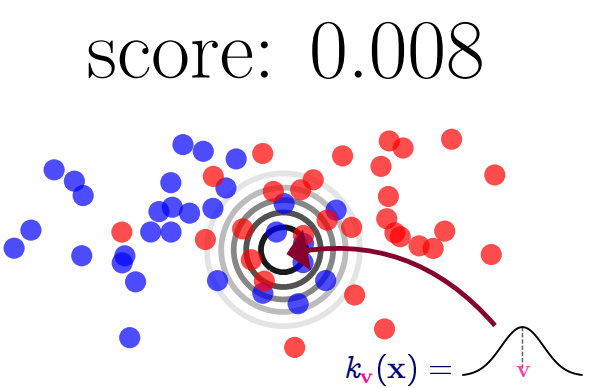
- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].



$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].

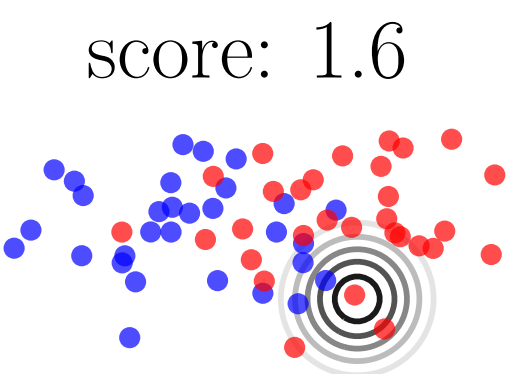


$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

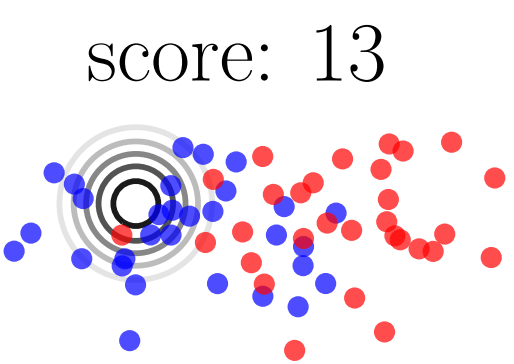
- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].



$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

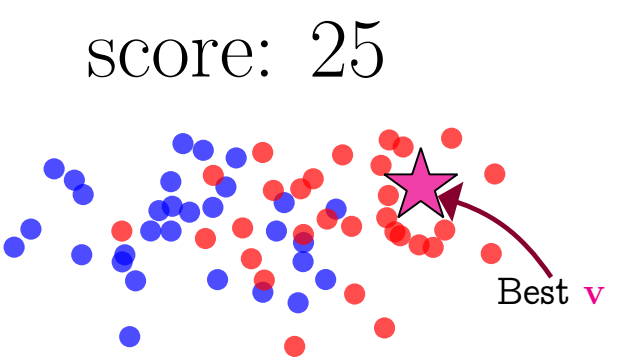
- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].



$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Model Criticism by the MMD Witness

- Find a location $\mathbf{v}$ at which q and p differ most [Jitkrittum et al., 2016].



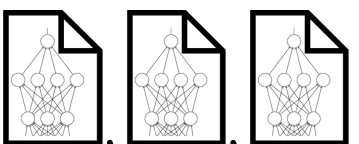
$$\text{witness}(\mathbf{v}) = \mathbb{E}_{\mathbf{x}\sim q}[\quad k_{\mathbf{v}}(\mathbf{x}) \quad] - \mathbb{E}_{\mathbf{y}\sim p}[\quad k_{\mathbf{v}}(\mathbf{y}) \quad],$$
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})} = \frac{\text{witness}^2(\mathbf{v})}{\sqrt{\mathbb{V}_{\mathbf{x}\sim q}[k_{\mathbf{v}}(\mathbf{x})] + \mathbb{V}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]}}.$$

Bayesian Inference Vs. Deep Learning Papers

Papers on Bayesian inference

$X = \{$  $, \ldots\} \sim q$

Papers on deep learning

$Y = \{$  $, \ldots\} \sim p$

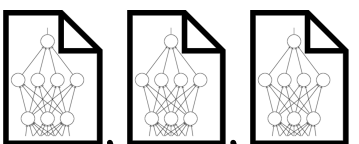
- NIPS papers (1988-2015). Sample size $n = 216$.
- Random 2000 nouns (dimensions). TF-IDF representation.
- **Learned ★:** infer, Bayes, Monte Carlo, adaptor, motif, haplotype, ECG

Bayesian Inference Vs. Deep Learning Papers

Papers on Bayesian inference

$X = \{$  $, \ldots\} \sim q$

Papers on deep learning

$Y = \{$  $, \ldots\} \sim p$

- NIPS papers (1988-2015). Sample size $n = 216$.
- Random 2000 nouns (dimensions). TF-IDF representation.
- **Learned ★:** infer, Bayes, Monte Carlo, adaptor, motif, haplotype, ECG

Distinguishing NIPS Articles

Paper categories

1. Bayesian inference
2. Deep learning
3. Learning theory
4. Neuroscience

Learned ★ (bags of words):

Bayes vs Bayes: collabor, traffic, bay, permut, net, central, occlus, mask.
Bayes vs Neuro: spike, markov, cortex, dropout, recurr, iii, gibb.
Learn vs Neuro: polici, interconnect, hardwar, decay, histolog, edg, period.
Bayes vs Learn: infer, Markov, graphic, segment, bandit, boundary, favor
Learn vs Deep: deep, forward, delay, subgroup, bandit, receptor, invariance

Distinguishing NIPS Articles

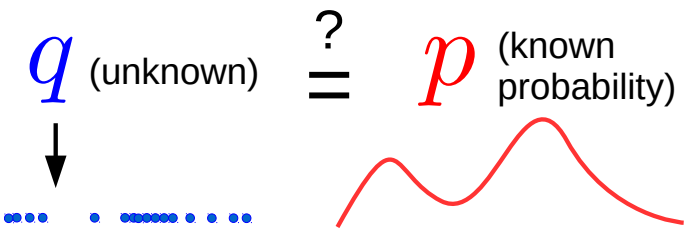
Paper categories

1. Bayesian inference
2. Deep learning
3. Learning theory
4. Neuroscience

Learned ★ (bags of words):

Bayes vs Bayes: collabor, traffic, bay, permut, net, central, occlus, mask.
Bayes vs Neuro: spike, markov, cortex, dropout, recurr, iii, gibb.
Learn vs Neuro: polici, interconnect, hardwar, decay, histolog, edg, period.
Bayes vs Learn: infer, Markov, graphic, segment, bandit, boundary, favor
Learn vs Deep: deep, forward, delay, subgroup, bandit, receptor, invariance

Model Criticism for one Sample and a Model (Setting #1)



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]

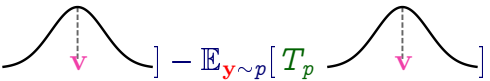
Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]$.

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

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

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

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



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

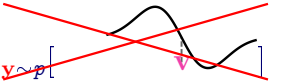
Problem: No sample from p. Cannot estimate $\mathbb{E}_{\mathbf{y}\sim p}[k_{\mathbf{v}}(\mathbf{y})]$.

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

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



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

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

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

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

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: Informative $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

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

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

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

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: Informative $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

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: Informative $\mathbf{v}$ should have high

$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

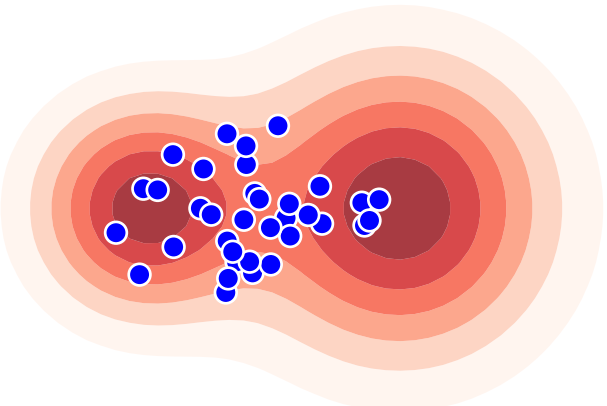
signal-to-noise ratio

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

Proposal: Model Criticism with the Stein Witness

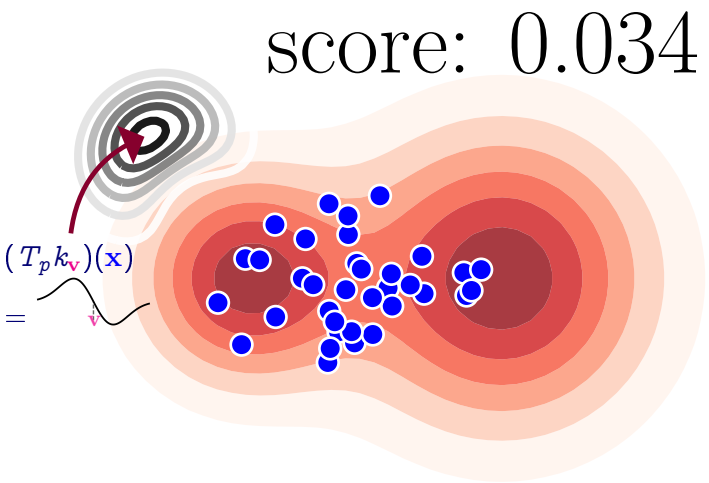
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



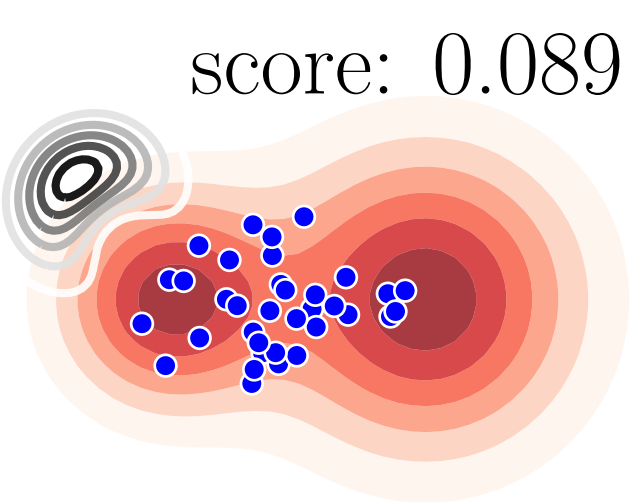
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

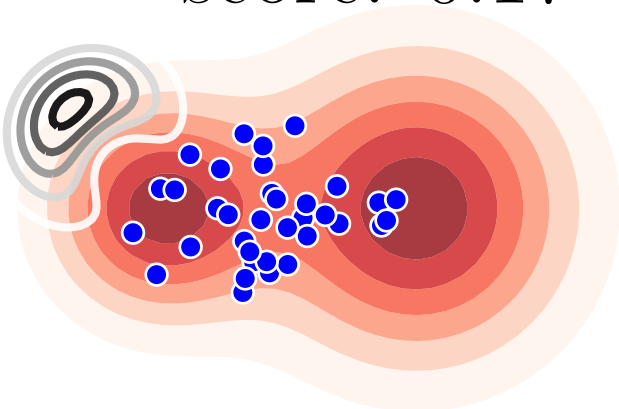
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

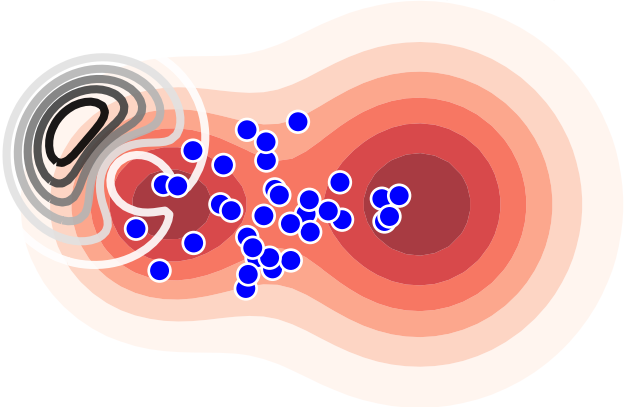
score: 0.17



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

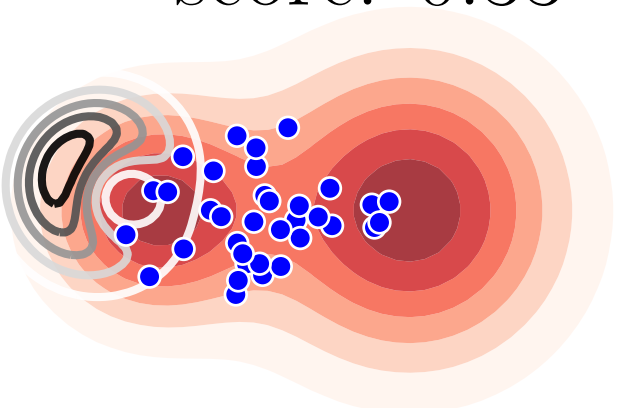
score: 0.26



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

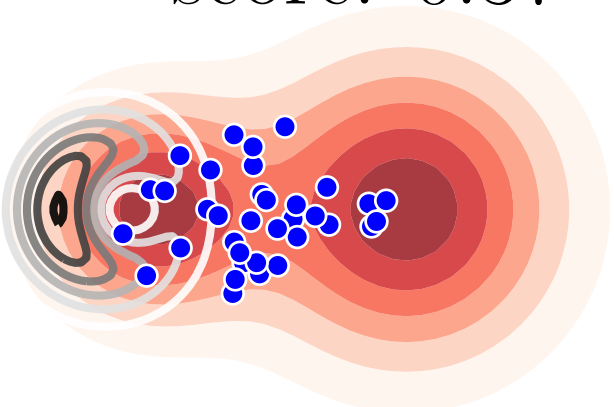
score: 0.33



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

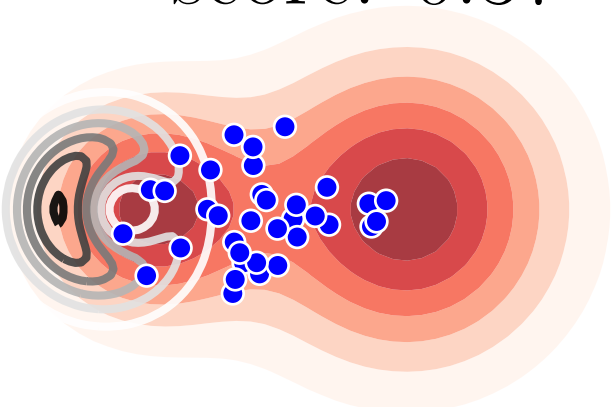
score: 0.37



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

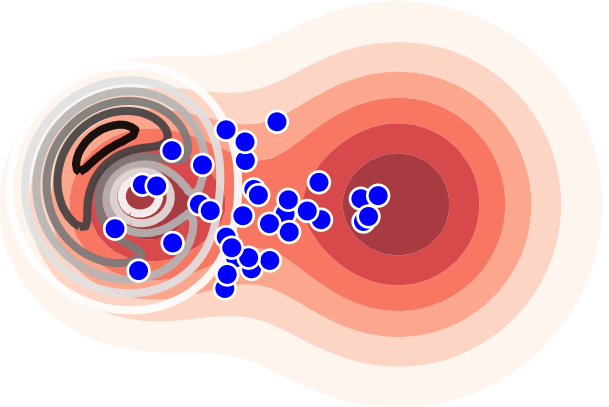
score: 0.37



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

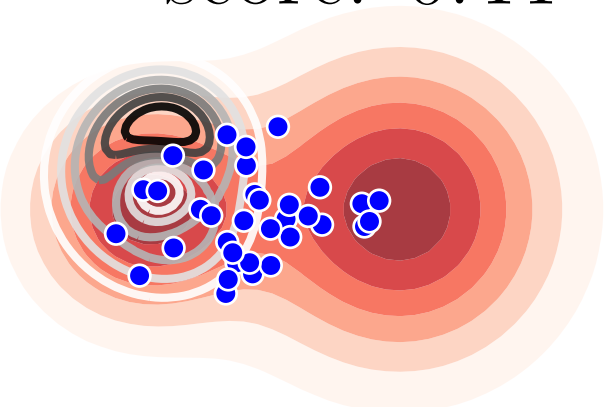
score: 0.45



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

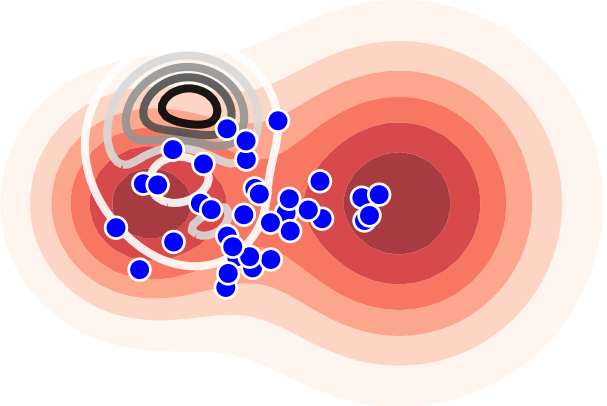
score: 0.44



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

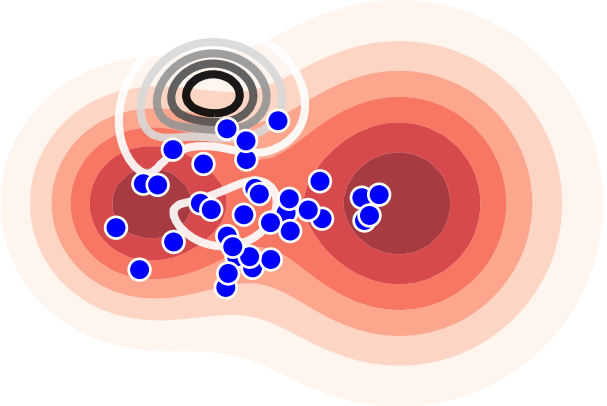
score: 0.39



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

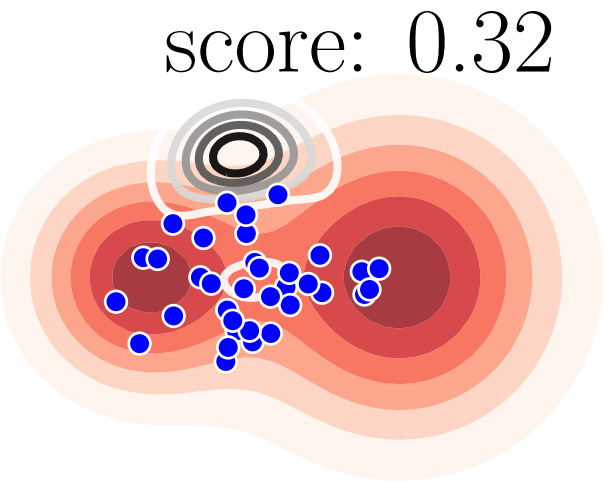
Proposal: Model Criticism with the Stein Witness

score: 0.31



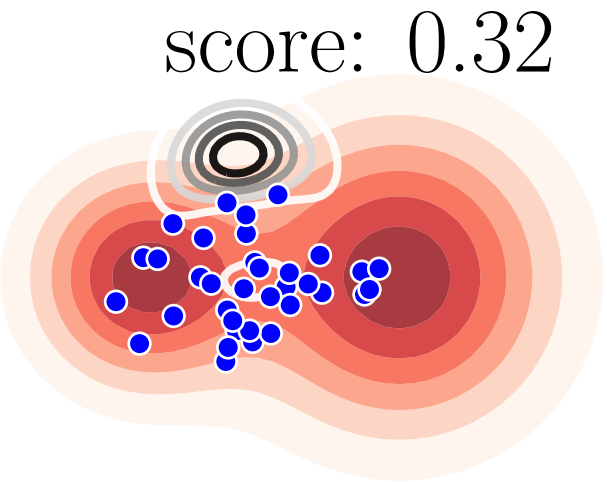
$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

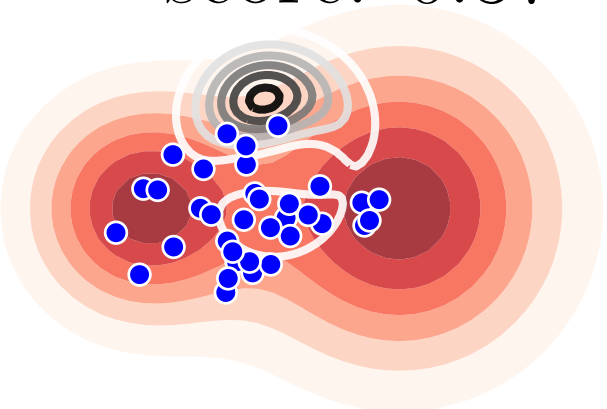
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

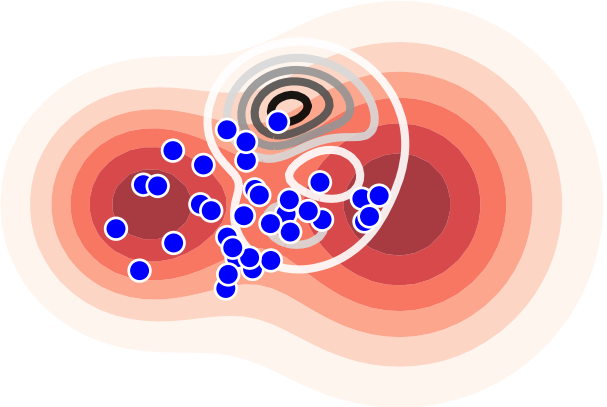
score: 0.37



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

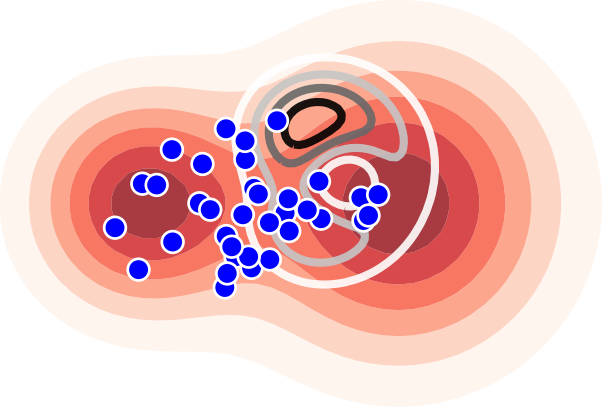
score: 0.48



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

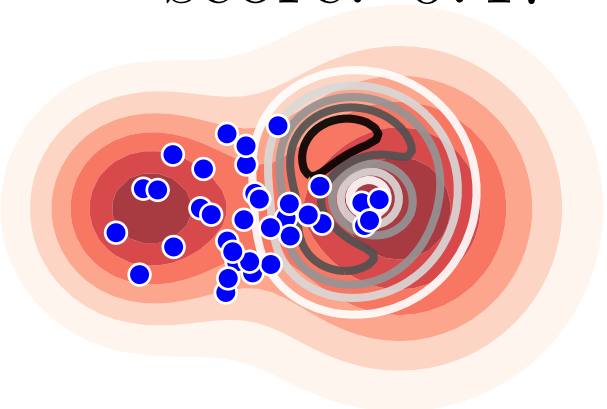
score: 0.49



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

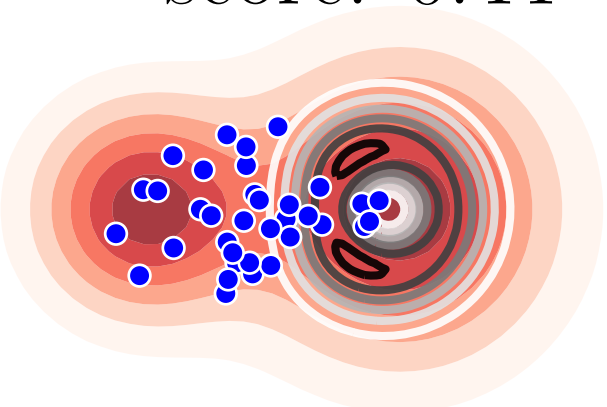
score: 0.47



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

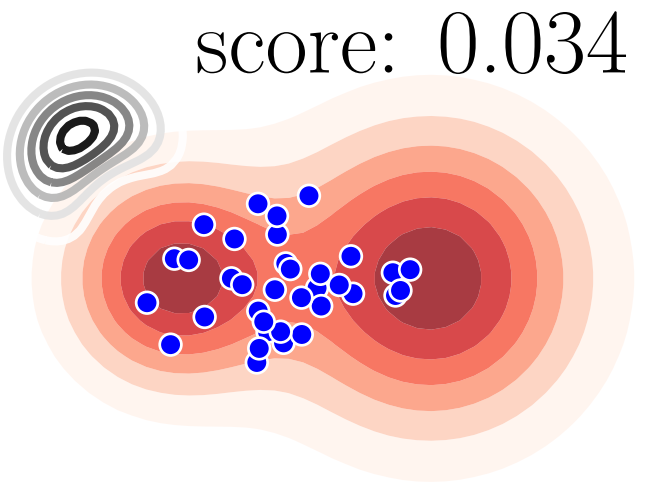
Proposal: Model Criticism with the Stein Witness

score: 0.44



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

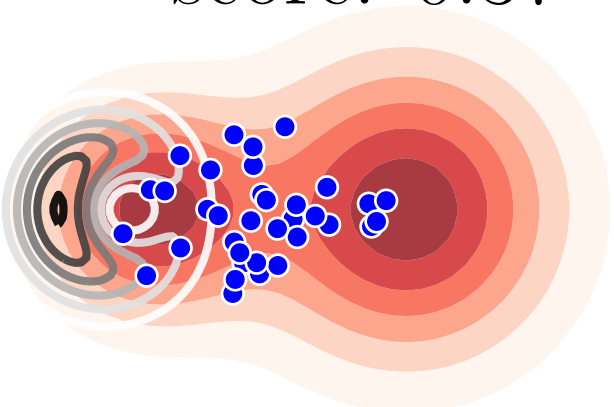
Proposal: Model Criticism with the Stein Witness



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

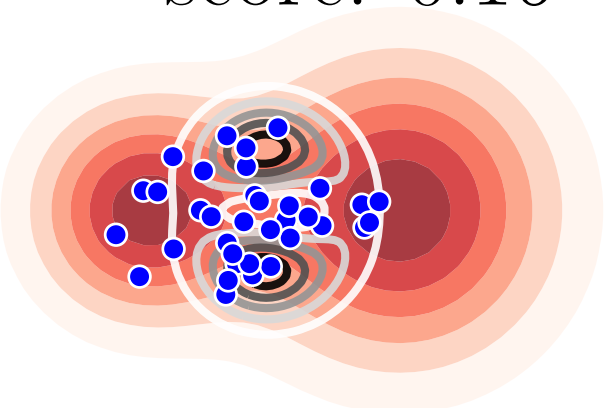
score: 0.37



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness

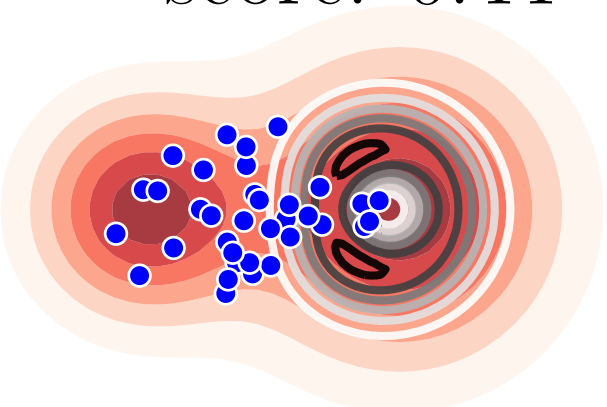
score: 0.16



$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

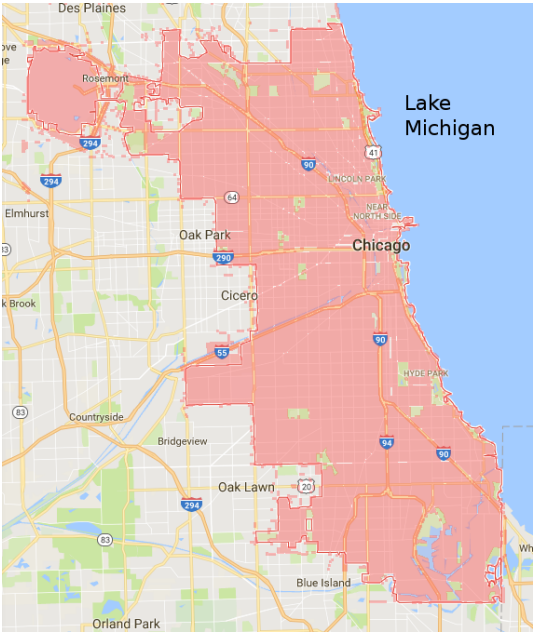
Proposal: Model Criticism with the Stein Witness

score: 0.44

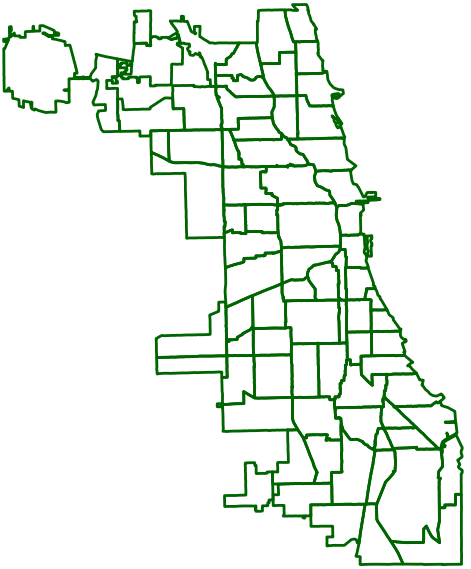


$$\text{score}(\mathbf{v}) = \frac{\text{witness}^2(\mathbf{v})}{\text{noise}(\mathbf{v})}.$$

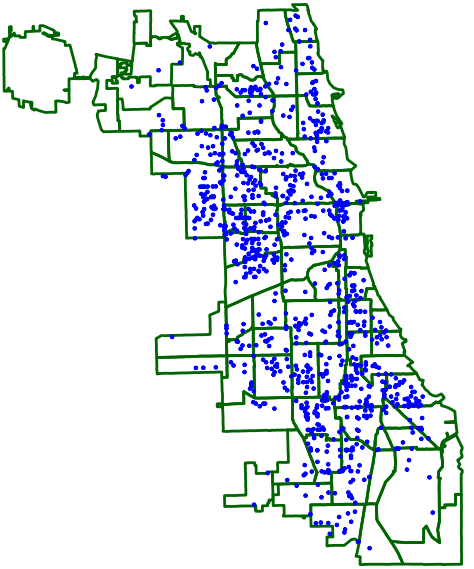
Interpretable Test Locations: Chicago Crime



Interpretable Test Locations: Chicago Crime

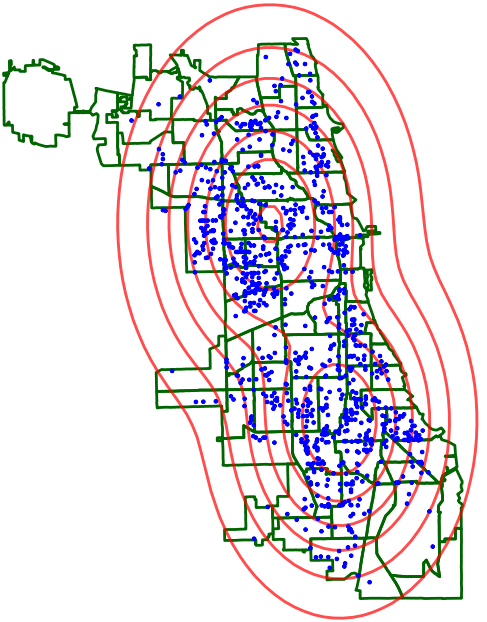


Interpretable Test Locations: Chicago Crime



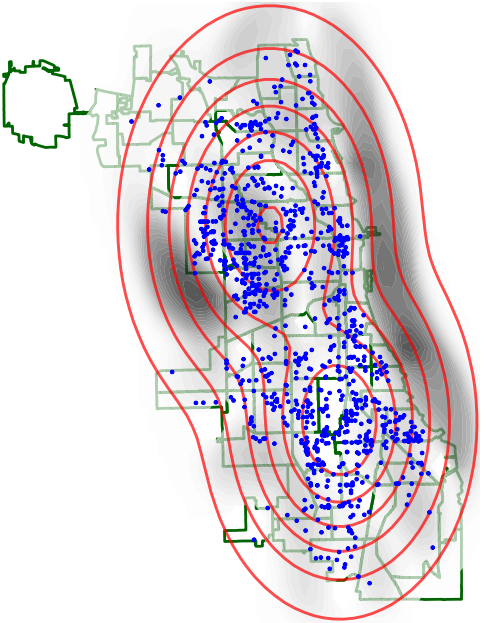
- $n = 11957$ robbery events in Chicago in 2016.
  - lat/long coordinates $=$ sample from q.
- Model spatial density with Gaussian mixtures.

Interpretable Test Locations: Chicago Crime



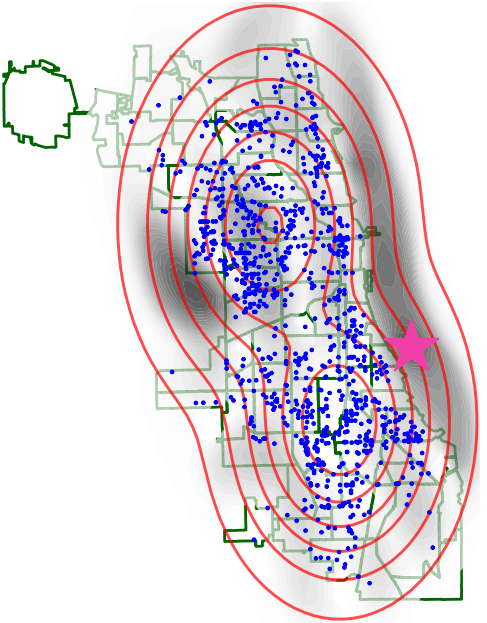
Model p = 2-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



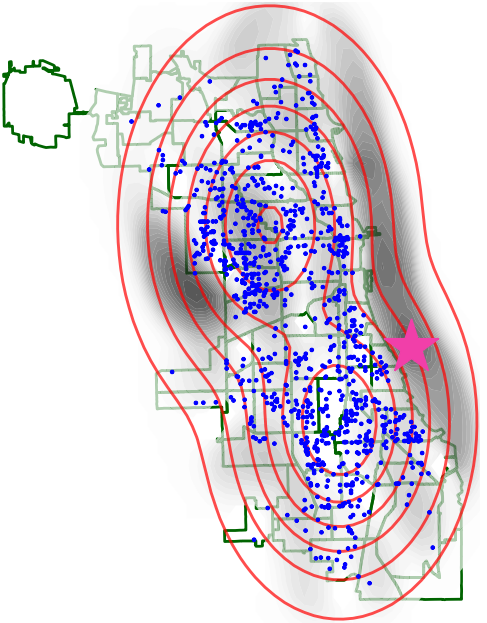
Score surface

Interpretable Test Locations: Chicago Crime



★ = optimized $\mathbf{v}$.

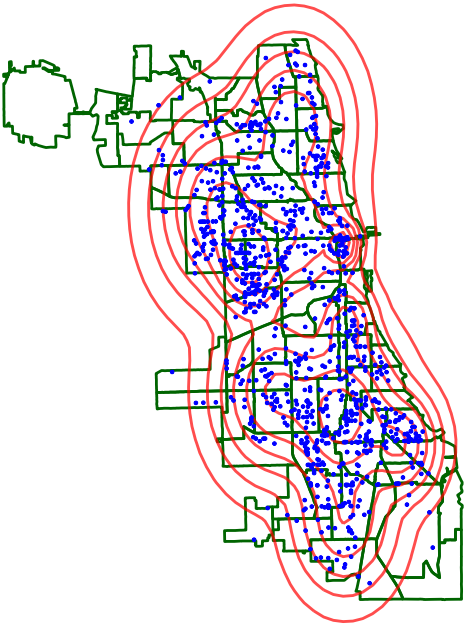
Interpretable Test Locations: Chicago Crime



★ = optimized $\mathbf{v}$.

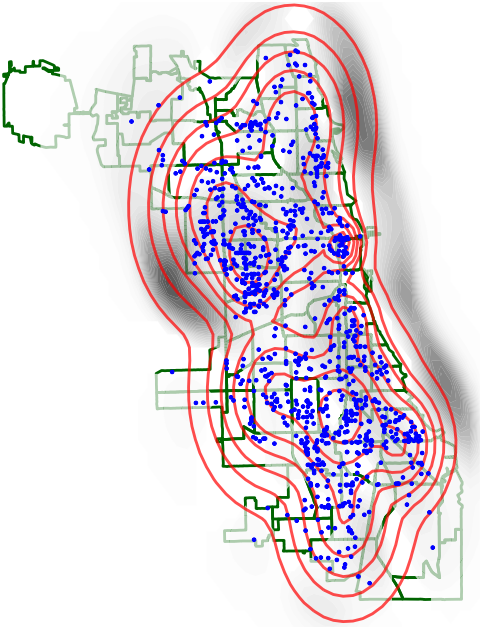
No robbery in Lake Michigan.

Interpretable Test Locations: Chicago Crime



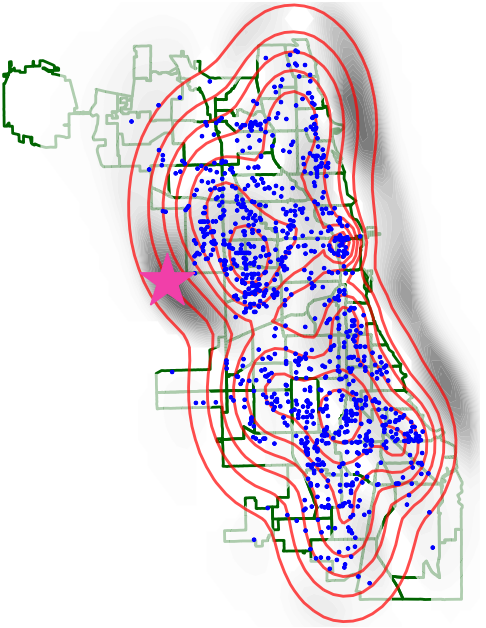
Model p = 10-component Gaussian mixture.

Interpretable Test Locations: Chicago Crime



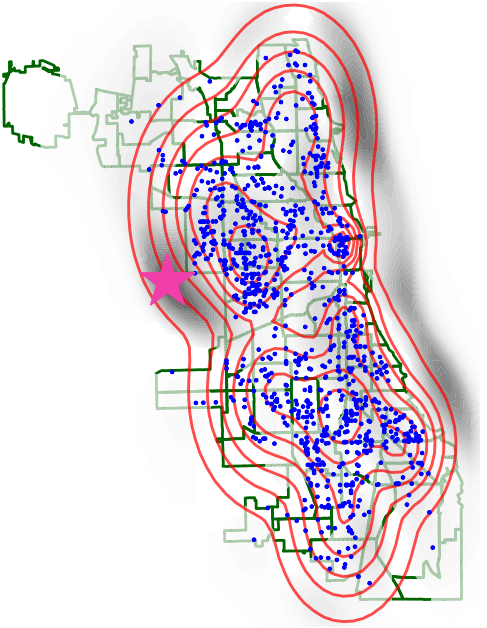
Capture the right tail better.

Interpretable Test Locations: Chicago Crime



Still, does not capture the left tail.

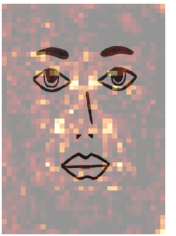
Interpretable Test Locations: Chicago Crime



Still, does not capture the left tail.

Learned test locations are interpretable.

Conclusion



Learned ★

Fast methods to learn key point(s)
for model criticism (or distinguishing two distributions)

- Based on a kernel function.
- Interpretable results.
- Theoretical guarantees.

Relevant References

- **Interpretable Distribution Features with Maximum Testing Power**
  Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, Arthur Gretton
  Python code: https://github.com/wittawatj/interpretable-test
  NIPS 2016
- **A Linear-Time Kernel Goodness-of-Fit Test**
  Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, Arthur Gretton
  Python code: https://github.com/wittawatj/kernel-gof
  NIPS 2017 (Best Paper)
- **Informative Features for Model Comparison**
  Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, Arthur Gretton
  Python code: https://github.com/wittawatj/kernel-mod
  NIPS 2018

Questions?

Thank you

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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:

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

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

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

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

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:

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

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

Application: Flow-based Generative Model [Rezende & Mohamed, 2015]

Samples from GLOW
[Kingma & Dhariwal, 2018]

$$z \sim p_0(z) \text{ (latent code)}$$
$$x = g(z) \text{ (produce image)}$$

where g is invertible.

- Model density $p(x)$ can be derived.

$$p(x) = p_0(g^{-1}(x)) \prod_{i=1}^{L} \left| \det \left(\frac{dh_i}{dh_{i-1}} \right) \right|$$

- Given real images $\{x_1, \ldots, x_n\}$, criticize the flow-based model p.
- How good is the model p compared to real images?

Application: Flow-based Generative Model [Rezende & Mohamed, 2015]

Samples from GLOW
[Kingma & Dhariwal, 2018]

$$z \sim p_0(z) \text{ (latent code)}$$
$$x = g(z) \text{ (produce image)}$$

where g is invertible.

- Model density $p(x)$ can be derived.

$$p(x) = p_0(g^{-1}(x)) \prod_{i=1}^{L} \left| \det \left(\frac{dh_i}{dh_{i-1}} \right) \right|$$

- Given real images $\{x_1, \ldots, x_n\}$, criticize the flow-based model p.
- How good is the model p compared to real images?

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* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *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*

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

- 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 $J = 5$.

References I

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016). Interpretable Distribution Features with Maximum Testing Power. In *NIPS*, pages 181–189.