Kernel tests of goodness-of-fit using Stein's method



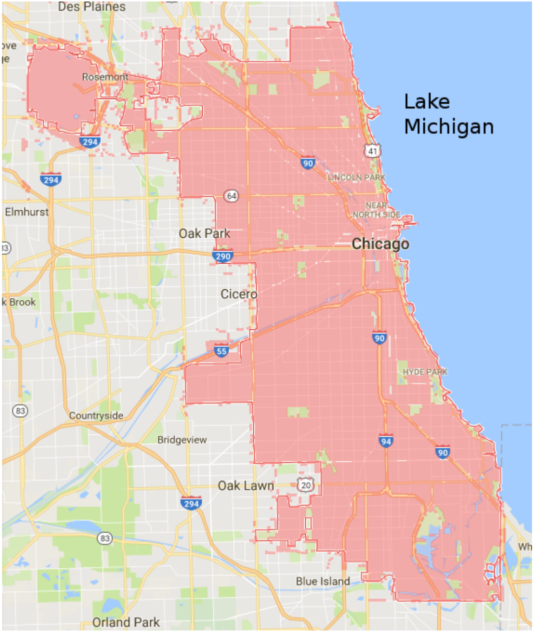



Arthur Gretton
gretton@gatsby.ucl.ac.uk

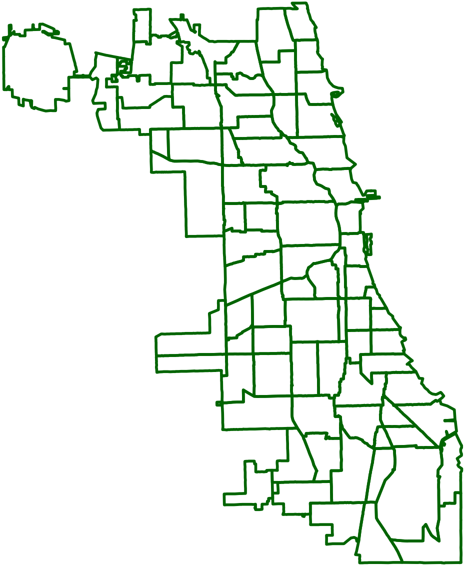
Gatsby Unit, University College London

Duke University 2019

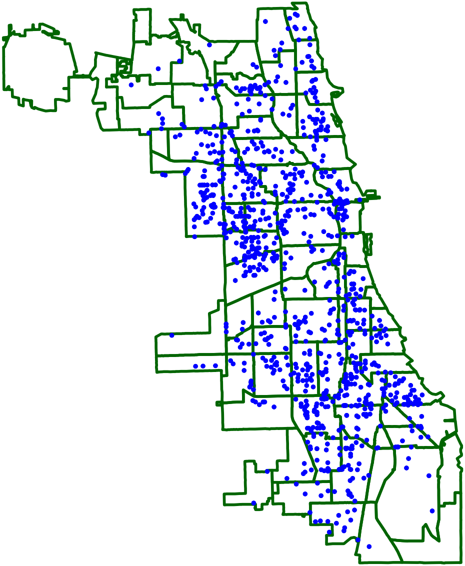
Model Criticism



Model Criticism

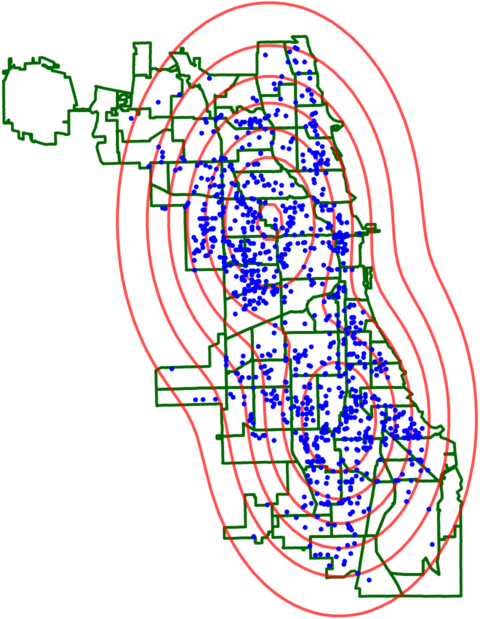


Model Criticism



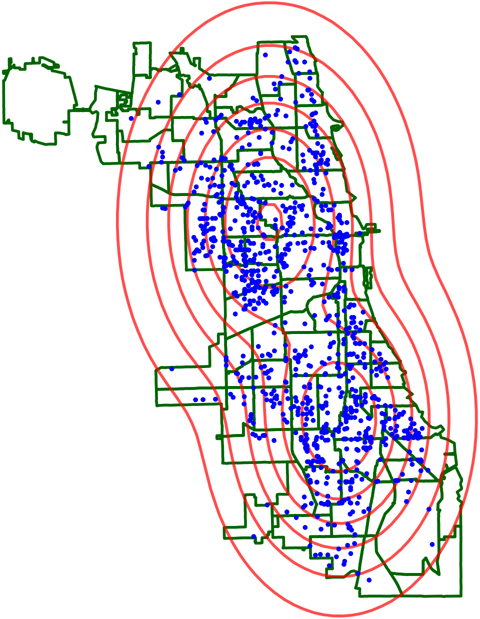
Data = robbery events in Chicago in 2016.

Model Criticism



Is this a good model?

Model Criticism



Goals: Test if a (complicated) model fits the data.

Outline

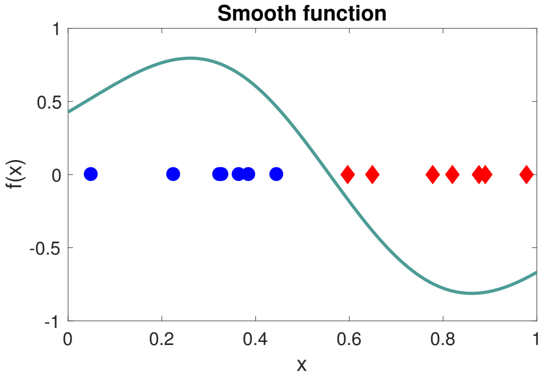
- The kernel Stein discrepancy Chwialkowski, Strathmann, G. ICML 2016
  - Comparing two models via samples: MMD and the witness function.
  - Comparing a sample and a model: Stein modification of the witness class
- A Linear-Time Kernel Goodness-of-Fit Test
  Jitkrittum, Xu, Szabo, Fukumizu, G. NeurIPS 2017
  - Features learned to maximise (estimate of) test power
  - Better asymptotic relative efficiency vs a “naive” linear time test
- Relative hypothesis tests with latent variables
  Kanagawa, Jitkrittum, Mackey, Fukumizu, G. 2019

Integral probability metrics

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

$$\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)$$

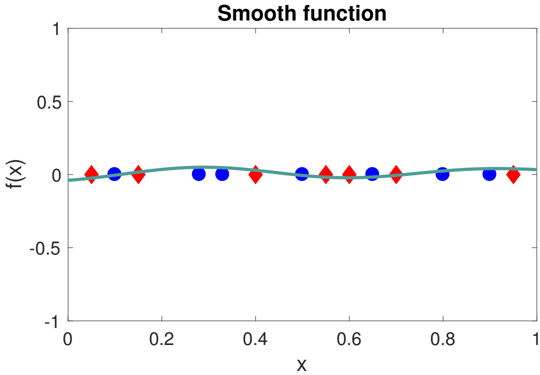


Integral probability metrics

Integral probability metric:

Find a "well behaved function" $f(x)$ to maximize

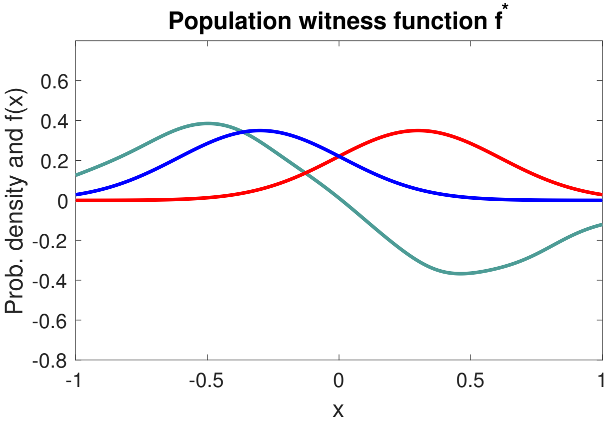
$$\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)$$



The MMD: an integral probability metric

Maximum mean discrepancy: RKHS function for P vs Q

$$MMD(P, Q; \mathcal{F}) := \sup_{\|f\|_{\mathcal{F}} \le 1} \left[\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)\right]$$



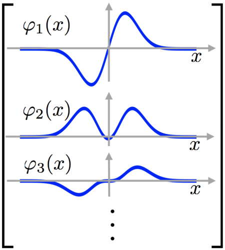
The MMD: an integral probability metric

Maximum mean discrepancy: RKHS function for P vs Q

$$MMD(P, Q; \mathcal{F}) := \sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X) \right]$$

Functions are linear combinations of features:

$$f(x) = \langle f, \varphi(x) \rangle_{\mathcal{F}} = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x) = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{bmatrix}^\top$$



$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^2 \leq 1$$

The MMD: an integral probability metric

Maximum mean discrepancy: RKHS function for P vs Q

$$MMD(P, Q; \mathcal{F}) := \sup_{\|f\|_{\mathcal{F}} \leq 1} [\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)]$$

For characteristic RKHS $\mathcal{F}$, $MMD(P, Q; F) = 0$ iff $P = Q$

Other choices for witness function class:

- Bounded continuous [Dudley, 2002]
- Bounded varation 1 (Kolmogorov metric) [Müller, 1997]
- Lipschitz (Wasserstein distances) [Dudley, 2002]

The MMD: an integral probability metric

Maximum mean discrepancy: RKHS function for P vs Q

$$MMD(P, Q; \mathcal{F}) := \sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)\right]$$

Expectations of functions are linear combinations of expected features

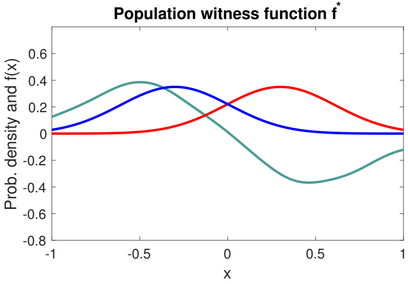
$$\mathbf{E}_P(f(X)) = \mathbf{E}_P \langle f, \varphi(X)\rangle_{\mathcal{F}} = \langle f, \mathbf{E}_P \varphi(X)\rangle_{\mathcal{F}} = \langle f, \mu_P\rangle_{\mathcal{F}}$$

(if feature map φ Bochner integrable; always true if kernel is bounded)

Integral prob. metric vs feature mean difference

The MMD:

$$MMD(P, Q; \mathcal{F}) = \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right]$$



Integral prob. metric vs feature mean difference

The MMD:

$$MMD(P, Q; \mathcal{F})$$
$$= \sup_{\|f\|\leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)\right]$$
$$= \sup_{\|f\|\leq 1} \left\langle f, \mu_P - \mu_Q \right\rangle_{\mathcal{F}}$$

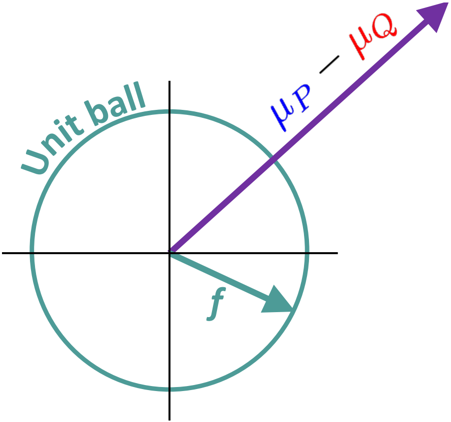
use

$$\mathbf{E}_P f(X) = \mathbf{E}_P \left\langle \varphi(X), f \right\rangle_{\mathcal{F}}$$
$$= \left\langle \mu_P, f \right\rangle_{\mathcal{F}}$$

Integral prob. metric vs feature mean difference

The MMD:

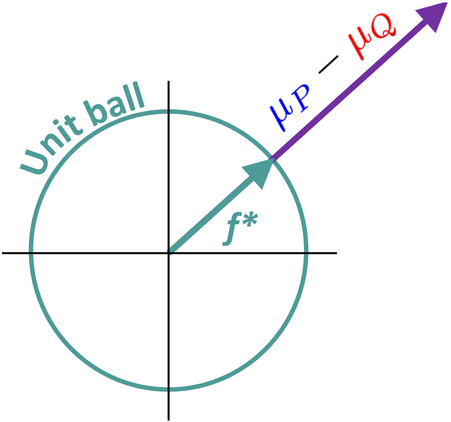
$$MMD(P, Q; \mathcal{F})$$
$$= \sup_{\|f\| \leq 1} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



Integral prob. metric vs feature mean difference

The MMD:

$$MMD(P, Q; \mathcal{F})$$
$$= \sup_{\|f\|\leq 1} [\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y)]$$
$$= \sup_{\|f\|\leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}}$$



$$f^* = \frac{\mu_P - \mu_Q}{\|\mu_P - \mu_Q\|}$$

Integral prob. metric vs feature mean difference

The MMD:

$$
\begin{aligned}
& MMD(P, Q; \mathcal{F}) \\
&= \sup_{\|f\| \leq 1} \left[\mathbf{E}_P f(X) - \mathbf{E}_Q f(Y) \right] \\
&= \sup_{\|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \|\mu_P - \mu_Q\|
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
f^*(v) &= \langle f, \varphi(v) \rangle_{\mathcal{F}} \\
&\propto \langle \mu_P - \mu_Q, \varphi(v) \rangle_{\mathcal{F}} \\
&= \langle \mathbf{E}_P \varphi(X) - \mathbf{E}_Q \varphi(Y), \varphi(v) \rangle_{\mathcal{F}} \\
&= \mathbf{E}_P k(X, v) - \mathbf{E}_Q k(Y, v)
\end{aligned}
$$

The maximum mean discrepancy

The maximum mean discrepancy in terms of **expected kernels:**

$$
\begin{aligned}
MMD^2(P, Q; \mathcal{F}) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\
&= \underbrace{\mathbf{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(y, y')}_{\text{(a)}} - 2\underbrace{\mathbf{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

(a)= within distrib. similarity, (b)= cross-distrib. similarity.

The maximum mean discrepancy

The maximum mean discrepancy in terms of **expected kernels:**

$$
\begin{aligned}
MMD^2(P, Q; \mathcal{F}) &= \|\mu_P - \mu_Q\|^2_{\mathcal{F}} \\
&= \underbrace{\mathbf{E}_P k(x, x')}_{\text{(a)}} + \underbrace{\mathbf{E}_Q k(y, y')}_{\text{(a)}} - \underbrace{2\mathbf{E}_{P,Q} k(x, y)}_{\text{(b)}}
\end{aligned}
$$

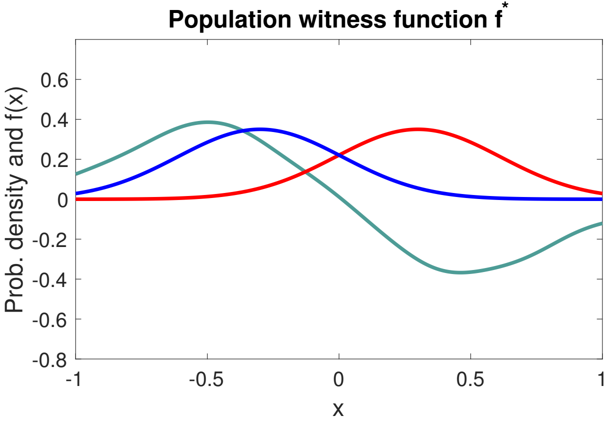
(a)= within distrib. similarity, (b)= cross-distrib. similarity.

Proof:

$$
\begin{aligned}
\|\mu_P - \mu_Q\|^2_{\mathcal{F}} &= \langle \mu_P - \mu_Q, \mu_P - \mu_Q \rangle_{\mathcal{F}} \\
&= \langle \mu_P, \mu_P \rangle_{\mathcal{F}} + \langle \mu_Q, \mu_Q \rangle_{\mathcal{F}} - 2 \langle \mu_P, \mu_Q \rangle_{\mathcal{F}} .
\end{aligned}
$$

Model criticism

$$\mathrm{MMD}(P, Q; \mathcal{F}) = \sup_{\|f\|_{\mathcal{F}} \le 1} [\mathbf{E}_q f - \mathbf{E}_p f]$$



Can we compute MMD with samples from Q and a **model** P?

Problem: usualy can't compute $\mathbf{E}_p f$ in closed form.

Stein idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \le 1} [\mathbf{E}_q f - \mathbf{E}_p f]$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Then

$$\mathbf{E}_p T_p f = 0$$

subject to appropriate boundary conditions.

Proof:

$$E_p [T_p f]$$

$$\int \left[\frac{d}{dx} (f(x) p(x)) \right] dx$$

$$= [f(x)p(x)]_{-\infty}^{\infty} = 0$$

Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)

Stein idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} \left[\mathbf{E}_q f - \mathbf{E}_p f\right]$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Then

$$\mathbf{E}_p T_p f = 0$$

subject to appropriate boundary conditions.

Proof:

$$E_p [T_p f] = \int \left[\frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))\right] p(x) dx$$

$$\int \left[\frac{d}{dx} (f(x) p(x))\right] dx$$

$$= [f(x) p(x)]_{-\infty}^{\infty} = 0$$

Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)

Stein idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} [\mathbf{E}_q f - \mathbf{E}_p f]$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x))$$

Then

$$\mathbf{E}_p T_p f = 0$$

subject to appropriate boundary conditions.

Proof:

$$E_p [T_p f] \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x)p(x)) \right] \cancel{p(x)} dx$$

$$\int \left[\frac{d}{dx} (f(x)p(x)) \right] dx$$

$$= [f(x)p(x)]_{-\infty}^{\infty} = 0$$

Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)

Stein idea

To get rid of $\mathbf{E}_p f$ in

$$\sup_{\|f\|_{\mathcal{F}} \leq 1} \left[\mathbf{E}_q f - \mathbf{E}_p f\right]$$

we define the (1-D) **Stein operator**

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Then

$$\mathbf{E}_p T_p f = 0$$

subject to appropriate boundary conditions.

Proof:

$$E_p [T_p f] \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} (f(x) p(x)) \right] \cancel{p(x)} dx$$

$$\int \left[\frac{d}{dx} (f(x) p(x)) \right] dx$$

$$= [f(x) p(x)]_{-\infty}^{\infty} = 0$$

Gorham and Mackey (NeurIPS 15), Oates, Girolami, Chopin (JRSS B 2016)

Kernel Stein Discrepancy

Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx} (f(x) p(x))$$

Kernel Stein Discrepancy (KSD)

$$\mathrm{KSD}_p(Q) = \sup_{\|g\|_{\mathcal{F}} \le 1} \mathbf{E}_q T_p g - \mathbf{E}_p T_p g$$

Kernel Stein Discrepancy

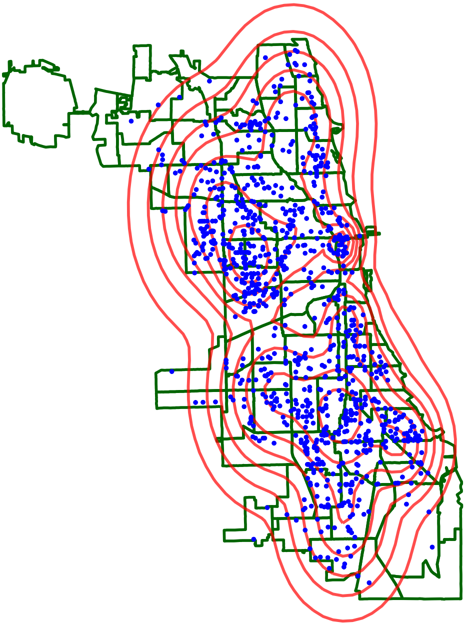
Stein operator

$$T_p f = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x) \right)$$

Kernel Stein Discrepancy (KSD)

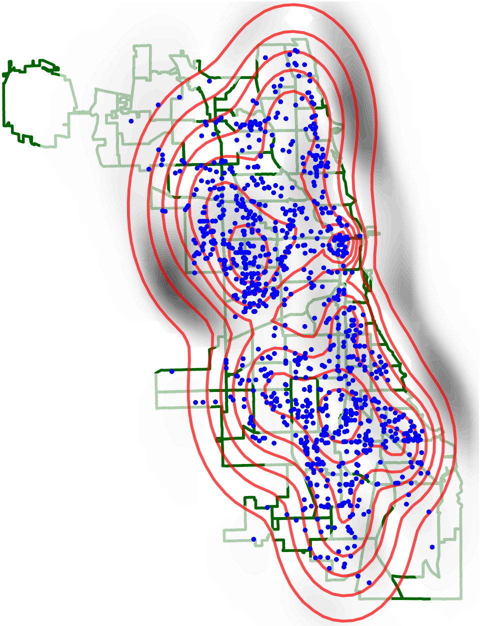
$$\mathrm{KSD}_p(Q) = \sup_{\|g\|_{\mathcal{F}} \le 1} \mathbf{E}_q T_p g - \cancel{\mathbf{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \le 1} \mathbf{E}_q T_p g$$

The witness function: Chicago Crime



Model p = 10-component Gaussian mixture.

The witness function: Chicago Crime



Witness function g shows mismatch

Simple expression using kernels

Re-write stein operator as:

$$\left[T_p f\right](x) = \frac{1}{p(x)} \frac{d}{dx}\left(f(x)p(x)\right)$$
$$= f(x)\frac{d}{dx}\log p(x) + \frac{d}{dx}f(x)$$

Can we define "Stein features" in $\mathcal{F}$?

$$\left[T_p f\right](x) = \left(\frac{d}{dx}\log p(x)\right) f(x) + \frac{d}{dx}f(x)$$
$$=: \langle f, \underbrace{\xi(x)}_{\text{stein features}} \rangle_{\mathcal{F}}$$

where $\mathbf{E}_{x\sim p}\xi(x) = 0$.

Simple expression using kernels

Re-write stein operator as:

$$[T_p f](x) = \frac{1}{p(x)} \frac{d}{dx}\left(f(x)p(x)\right) = f(x)\frac{d}{dx}\log p(x) + \frac{d}{dx}f(x)$$

Can we define "Stein features" in $\mathcal{F}$?

$$[T_p f](x) = \left(\frac{d}{dx}\log p(x)\right) f(x) + \frac{d}{dx}f(x) =: \langle f, \underbrace{\xi(x)}_{\text{stein features}} \rangle_{\mathcal{F}}$$

where $\mathbf{E}_{x\sim p}\xi(x) = 0$.

The kernel trick for derivatives

Reproducing property for the derivative: for differentiable $k(x, x')$,

$$\frac{d}{dx}f(x) = \left\langle f, \frac{d}{dx}\varphi(x) \right\rangle_{\mathcal{F}}$$

The kernel trick for derivatives

Reproducing property for the derivative: for differentiable $k(x, x')$,

$$\frac{d}{dx}f(x) = \left\langle f, \frac{d}{dx}\varphi(x) \right\rangle_{\mathcal{F}}$$

Using kernel derivative trick in (a),

$$\begin{aligned}
[T_p f](x) &= \left(\frac{d}{dx}\log p(x)\right) f(x) + \frac{d}{dx}f(x) \\
&= \left\langle f, \left(\frac{d}{dx}\log p(x)\right)\varphi(x) + \underbrace{\frac{d}{dx}\varphi(x)}_{(a)} \right\rangle_{\mathcal{F}} \\
&=: \langle f, \xi(x)\rangle_{\mathcal{F}}.
\end{aligned}$$

Kernel stein discrepancy: derivation

Closed-form expression for KSD:

$$
\begin{aligned}
\mathrm{KSD}_p(Q) &= \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathbf{E}_{x\sim q}\left([T_p g](x)\right) \\
&= \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathbf{E}_{x\sim q} \left\langle g, \xi_x \right\rangle_{\mathcal{F}} \\
&\overset{(a)}{=} \sup_{\|g\|_{\mathcal{F}} \leq 1} \left\langle g, \mathbf{E}_{x\sim q}\xi_x \right\rangle_{\mathcal{F}} = \left\|\mathbf{E}_{x\sim q}\xi_x\right\|_{\mathcal{F}}
\end{aligned}
$$

Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

Kernel stein discrepancy: derivation

Closed-form expression for KSD:

$$
\begin{aligned}
\mathrm{KSD}_p(Q) &= \sup_{\|g\|_{\mathcal{F}} \le 1} \mathbf{E}_{x\sim q}\left([T_p g](x)\right) \\
&= \sup_{\|g\|_{\mathcal{F}} \le 1} \mathbf{E}_{x\sim q} \langle g, \xi_x \rangle_{\mathcal{F}} \\
&\overset{(a)}{=} \sup_{\|g\|_{\mathcal{F}} \le 1} \langle g, \mathbf{E}_{x\sim q}\xi_x \rangle_{\mathcal{F}} = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}
\end{aligned}
$$

Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

Kernel stein discrepancy: derivation

Closed-form expression for KSD:

$$\begin{aligned}\mathrm{KSD}_p(Q) &= \sup_{\|g\|_{\mathcal{F}}\le 1} \mathbf{E}_{x\sim q}\left([T_p g](x)\right)\\ &= \sup_{\|g\|_{\mathcal{F}}\le 1} \mathbf{E}_{x\sim q}\left\langle g, \xi_x\right\rangle_{\mathcal{F}}\\ &\underset{(a)}{=} \sup_{\|g\|_{\mathcal{F}}\le 1} \left\langle g, \mathbf{E}_{x\sim q}\xi_x\right\rangle_{\mathcal{F}} = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}\end{aligned}$$

Caution: (a) requires a condition for the Riesz theorem to hold,

$$\mathbf{E}_{x\sim q}\left(\frac{d}{dx}\log p(x)\right)^2 < \infty.$$

Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

Does the Riesz condition matter?

Consider the standard normal,

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

Then

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

If q is a Cauchy distribution, then the integral

$$\mathbf{E}_{x\sim q}\left(\frac{d}{dx}\log p(x)\right)^2 = \int_{-\infty}^{\infty} x^2 q(x)\, dx$$

is undefined.

Does the Riesz condition matter?

Consider the standard normal,

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

Then

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

If q is a Cauchy distribution, then the integral

$$\mathbf{E}_{x\sim q} \left(\frac{d}{dx} \log p(x)\right)^2 = \int_{-\infty}^{\infty} x^2 q(x) dx$$

is undefined.

Kernel stein discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $x, x' \sim q$,

$$\mathrm{KSD}_p^2(Q) = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}^2 = \mathbf{E}_{x,x'\sim q} h_p(x, x')$$

where

$$h_p(x,x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x,x') + \mathbf{s}_p(x)^\top k_2(x,x') + \mathbf{s}_p(x')^\top k_1(x,x') + \mathrm{tr}\left[k_{12}(x,x')\right]$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \frac{\nabla p(x)}{p(x)}$
- $k_1(a,b) := \nabla_x k(x,x')|_{x=a,x'=b} \in \mathbb{R}^d$,
  $k_2(a,b) := \nabla_{x'} k(x,x')|_{x=a,x'=b} \in \mathbb{R}^d$,
- $k_{12}(a,b) := \nabla_x \nabla_{x'} k(x,x')|_{x=a,x'=b} \in \mathbb{R}^{d\times d}$

Kernel stein discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $x, x' \sim q$,

$$\mathrm{KSD}_p^2(Q) = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}^2 = \mathbf{E}_{x,x'\sim q} h_p(x, x')$$

where

$$\begin{aligned} h_p(x, x') &= \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^\top k_2(x, x') \\ &\quad + \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right] \end{aligned}$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \frac{\nabla p(x)}{p(x)}$
- $k_1(a, b) := \nabla_x k(x, x')|_{x=a,x'=b} \in \mathbb{R}^d$,
  $k_2(a, b) := \nabla_{x'} k(x, x')|_{x=a,x'=b} \in \mathbb{R}^d$,
- $k_{12}(a, b) := \nabla_x \nabla_{x'} k(x, x')|_{x=a,x'=b} \in \mathbb{R}^{d\times d}$

Kernel stein discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $x, x' \sim q$,

$$\mathrm{KSD}_p^2(Q) = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}^2 = \mathbf{E}_{x,x'\sim q} h_p(x, x')$$

where

$$\begin{aligned} h_p(x, x') = {} & \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^\top k_2(x, x') \\ & + \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right] \end{aligned}$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \frac{\nabla p(x)}{p(x)}$
- $k_1(a, b) := \nabla_x k(x, x')|_{x=a, x'=b} \in \mathbb{R}^d$,
  $k_2(a, b) := \nabla_{x'} k(x, x')|_{x=a, x'=b} \in \mathbb{R}^d$,
- $k_{12}(a, b) := \nabla_x \nabla_{x'} k(x, x')|_{x=a, x'=b} \in \mathbb{R}^{d \times d}$

Do not need to normalize p, or sample from it.

Kernel stein discrepancy: population expression

Test statistic when $x \in \mathbb{R}^d$, given *independent* $x, x' \sim q$,

$$\mathrm{KSD}_p^2(Q) = \|\mathbf{E}_{x\sim q}\xi_x\|_{\mathcal{F}}^2 = \mathbf{E}_{x,x'\sim q} h_p(x, x')$$

where

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') + \mathbf{s}_p(x)^\top k_2(x, x') + \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right]$$

- $\mathbf{s}_p(x) \in \mathbb{R}^d = \frac{\nabla p(x)}{p(x)}$
- $k_1(a, b) := \nabla_x k(x, x')|_{x=a, x'=b} \in \mathbb{R}^d$,
  $k_2(a, b) := \nabla_{x'} k(x, x')|_{x=a, x'=b} \in \mathbb{R}^d$,
- $k_{12}(a, b) := \nabla_x \nabla_{x'} k(x, x')|_{x=a, x'=b} \in \mathbb{R}^{d\times d}$

If kernel is C_0-universal and Q satisfies $\mathbf{E}_{x\sim q}\left\|\nabla\left(\log\frac{p(x)}{q(x)}\right)\right\|^2 < \infty$, then $\mathrm{KSD}_p^2(Q) = 0$ iff $P = Q$.

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \ldots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') \\ - \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right]$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x,
$\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \ldots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') - \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right]$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x,
$\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

A discrete kernel: $k(x, x') = \exp\left(-d_H(x, x')\right)$, where
$d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x_d \neq x'_d)$.

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

KSD for discrete-valued variables

Discrete domains: $\mathcal{X} = \{1, \ldots, L\}^D$ with $L \in \mathbb{N}$.

The population KSD (discrete):

$$\mathrm{KSD}_p^2(Q) = \mathbf{E}_{x,x' \sim q} h_p(x, x')$$

where

$$h_p(x, x') = \mathbf{s}_p(x)^\top \mathbf{s}_p(x') k(x, x') - \mathbf{s}_p(x)^\top k_2(x, x') \\ - \mathbf{s}_p(x')^\top k_1(x, x') + \mathrm{tr}\left[k_{12}(x, x')\right]$$

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, Δ_x^{-1} is cyclic backwards difference on x, $\mathbf{s}_p(x) = \frac{\Delta p(x)}{p(x)}$

A discrete kernel: $k(x, x') = \exp\left(-d_H(x, x')\right)$, where $d_H(x, x') = D^{-1} \sum_{d=1}^{D} \mathbb{I}(x_d \neq x'_d)$.

$\mathrm{KSD}_p^2(Q) = 0$ iff $P = Q$ if

- Gram matrix over all the configurations in $\mathcal{X}$ is strictly positive definite,
- $P > 0$ and $Q > 0$.

Ranganath et al. (NeurIPS 2016), Yang et al. (ICML 2018)

Empirical statistic and asymptotics

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

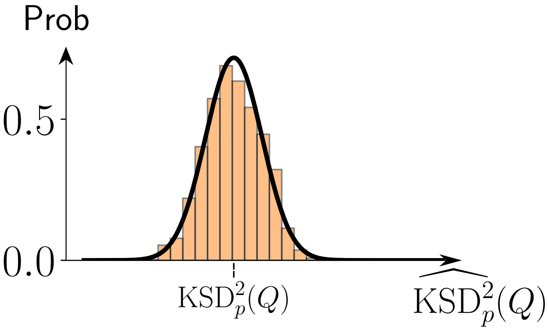
Empirical statistic and asymptotics

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

Asymptotic distribution when $P \neq Q$:

$$\sqrt{n}\left(\widehat{\mathrm{KSD}_p^2}(Q) - \mathrm{KSD}_p^2(Q)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_{h_p}^2) \qquad \sigma_{h_p}^2 = 4\mathrm{Var}_x[\mathbf{E}_{x'}[h_p(x, x')]].$$



Empirical statistic and asymptotics

The empirical statistic:

$$\widehat{\mathrm{KSD}_p^2}(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

Asymptotic distribution when $P = Q$:

$$n\widehat{\mathrm{KSD}_p^2}(Q) \sim \sum_{\ell=1}^{\infty} \lambda_\ell Z_\ell^2$$

$$\lambda_i \psi_i(x') = \int_{\mathcal{X}} h_p(x, x') \psi_i(x) \, dp(x)$$

$$Z_\ell \sim \mathcal{N}(0, 1) \quad \text{i.i.d.}$$

Test threshold via wild bootstrap.

A naive linear time statistic

A running average:

$$\widehat{LKS_p^2}(Q) := \frac{2}{n} \sum_{i=1}^{n/2} h_p(x_{2i-1}, x_{2i}).$$

Asymptotically normal when $P \neq Q$ *and* when $P = Q$.

A naive linear time statistic

A running average:

$$\widehat{LKS_p^2}(Q) := \frac{2}{n}\sum_{i=1}^{n/2} h_p(x_{2i-1}, x_{2i}).$$

Asymptotically normal when $P \neq Q$ *and* when $P = Q$.

Can we do better? Wishlist:

1. still linear-time
2. adaptive (parameters automatically tuned)
3. more interpretable

Linear-time, interpretable Goodness-of-fit Test

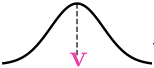
Stein Witness Function at a Single Location

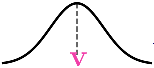
Idea:

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

Stein Witness Function at a Single Location

Idea:

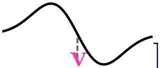
$(\text{Stein})\ \text{witness}(\mathbf{v}) = \mathbf{E}_{\mathbf{x}\sim q}[T_p$ 

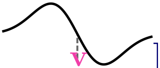
$] - \mathbf{E}_{\mathbf{y}\sim p}[T_p$ 

$]$

Stein Witness Function at a Single Location

Idea:

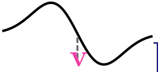
$\text{(Stein) witness}(\mathbf{v}) = \mathbf{E}_{\mathbf{x}\sim q}[$ 

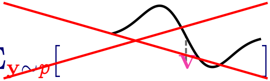
$] - \mathbf{E}_{\mathbf{y}\sim p}[$ 

$]$

Stein Witness Function at a Single Location

Idea:

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

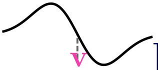
$] - \mathbf{E}_{\mathbf{y}\sim p}[$ 

$]$

Stein Witness Function at a Single Location

Idea:

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



Stein Witness Function at a Single Location

Idea:

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

Stein Witness Function at a Single Location

Idea:

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

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

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

Stein Witness Function at a Single Location

Idea:

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

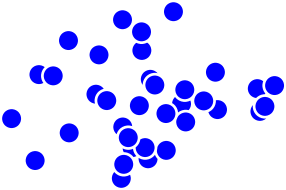
Proposal: Good $\mathbf{v}$ should have high

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

signal-to-noise ratio

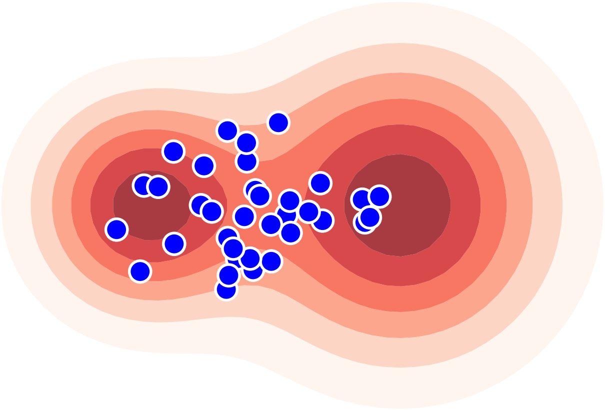
- witness($\mathbf{v}$) and standard deviation($\mathbf{v}$) can be estimated in linear time.

Proposal: Model Criticism with the Stein Witness



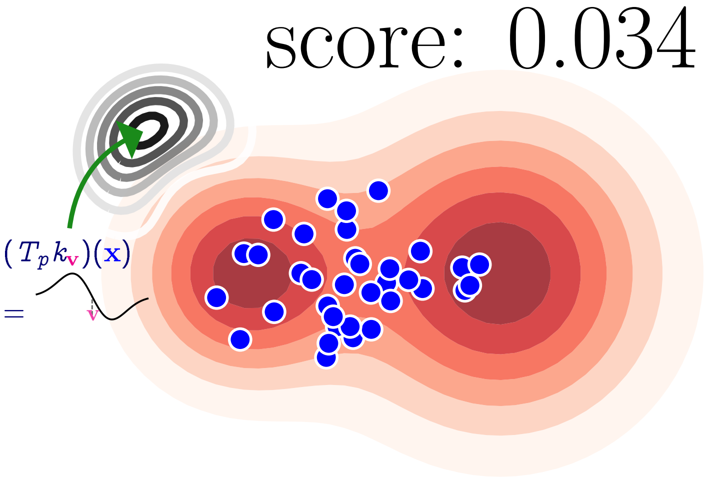
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



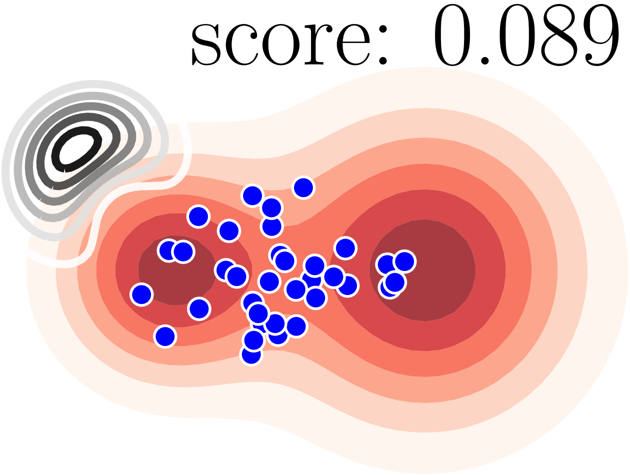
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



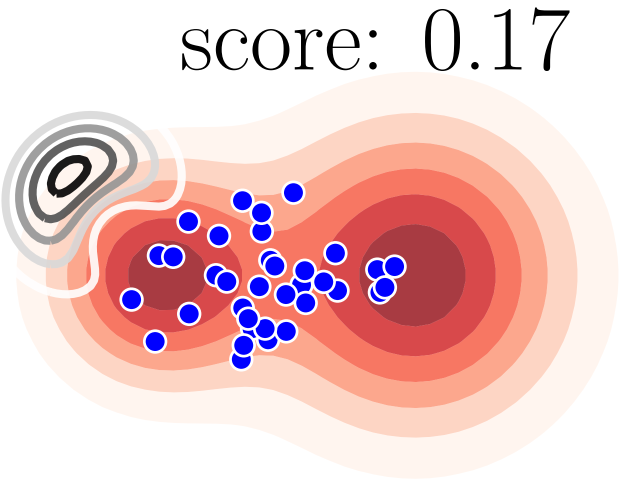
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



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

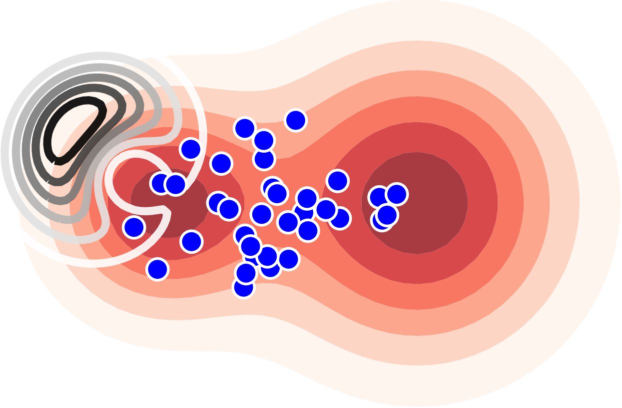
Proposal: Model Criticism with the Stein Witness



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

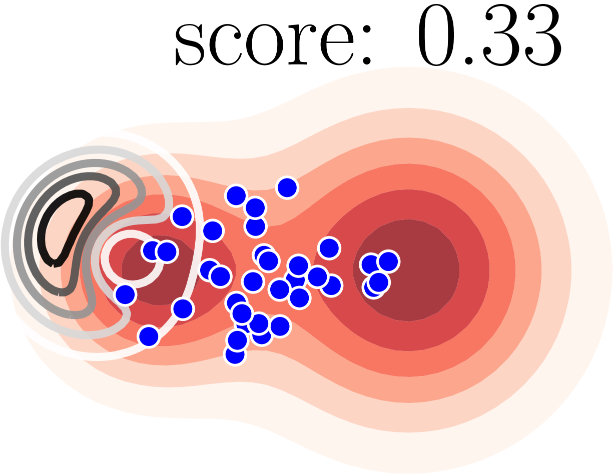
Proposal: Model Criticism with the Stein Witness

score: 0.26



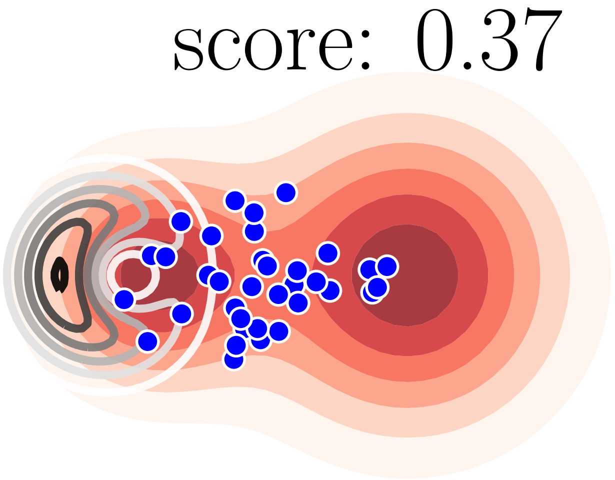
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



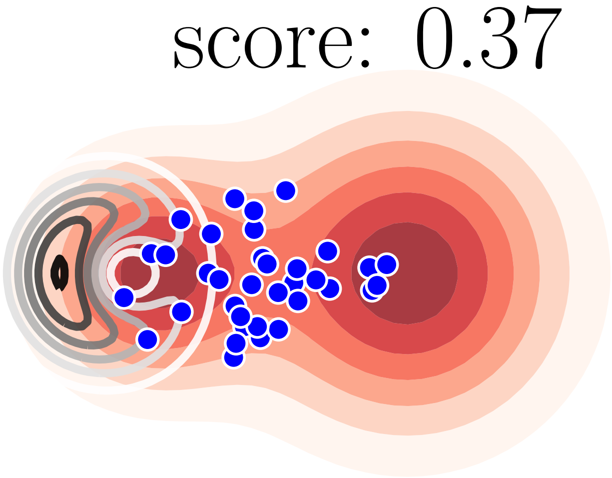
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



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

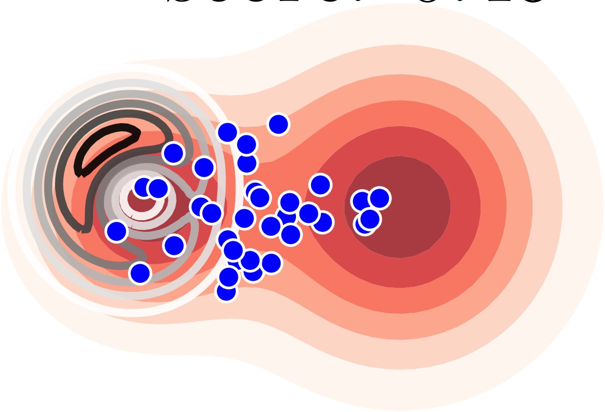
Proposal: Model Criticism with the Stein Witness



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

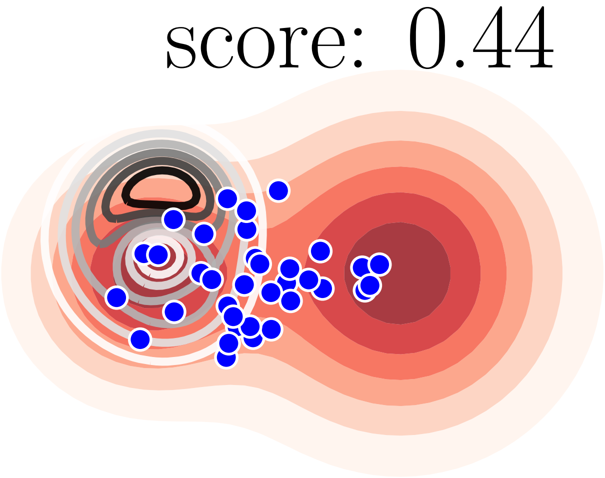
Proposal: Model Criticism with the Stein Witness

score: 0.45



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

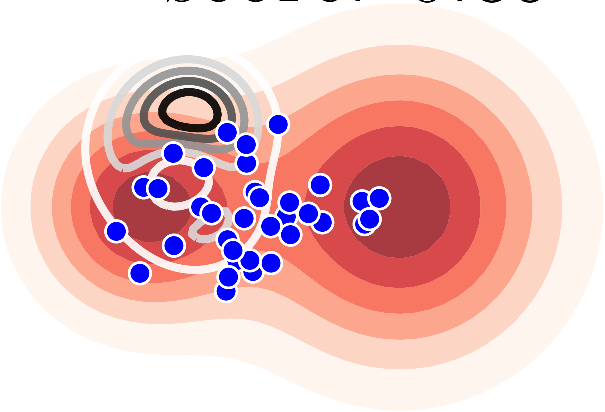
Proposal: Model Criticism with the Stein Witness



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

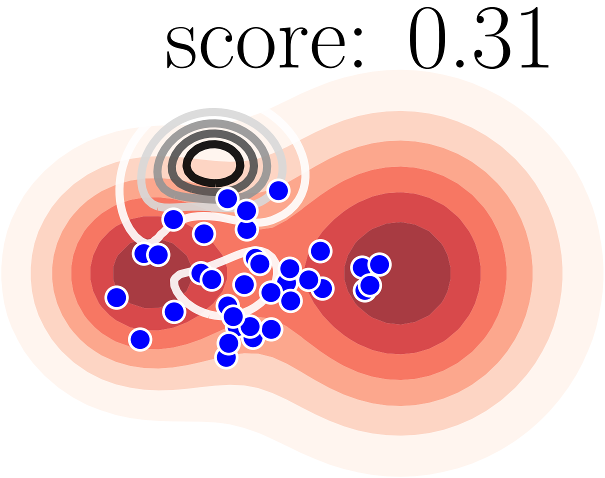
Proposal: Model Criticism with the Stein Witness

score: 0.39



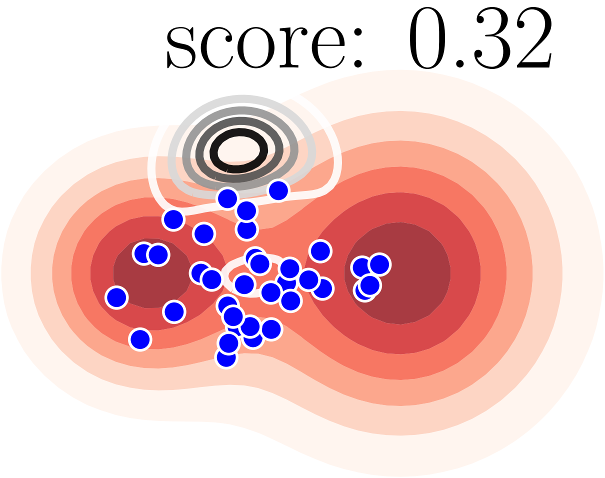
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



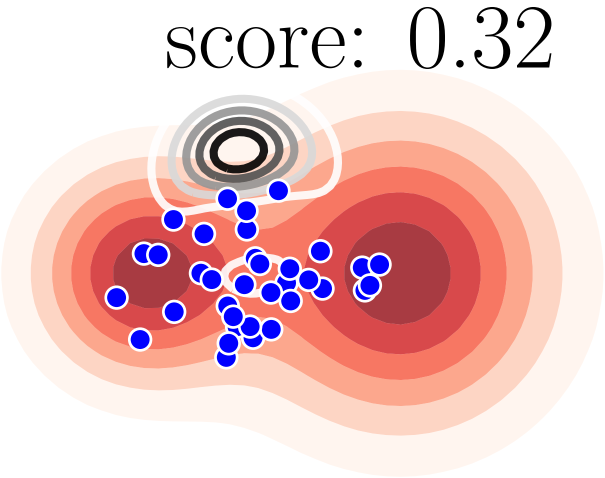
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



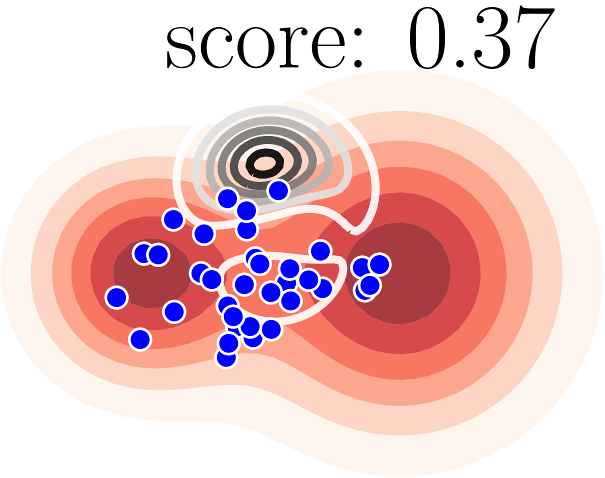
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



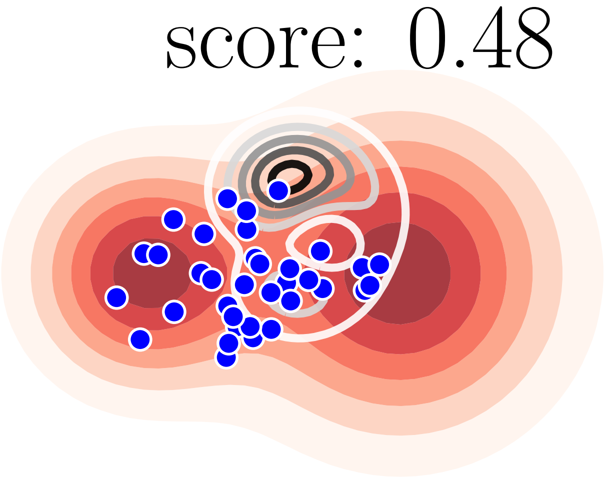
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



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

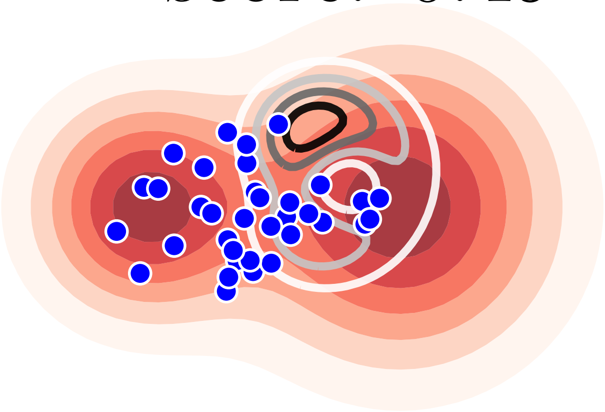
Proposal: Model Criticism with the Stein Witness



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

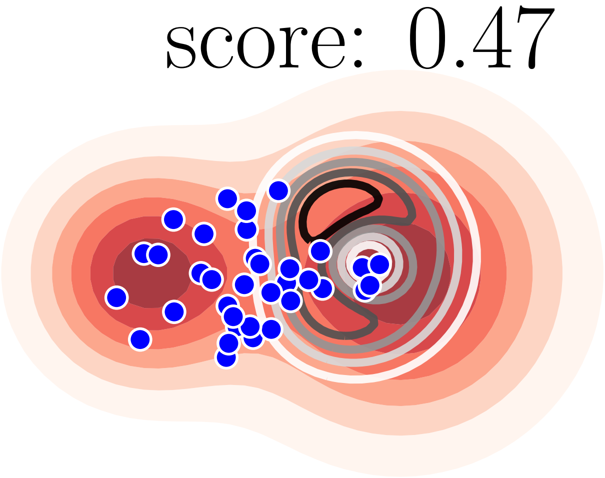
Proposal: Model Criticism with the Stein Witness

score: 0.49



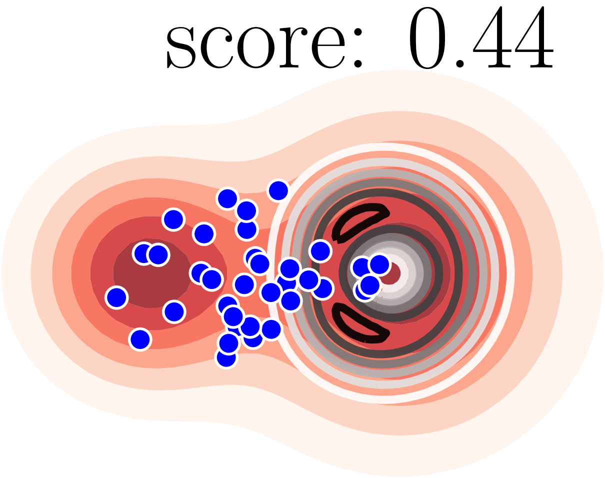
$$\text{score}(\mathbf{v}) = \frac{|\text{witness}(\mathbf{v})|}{\text{standard deviation}(\mathbf{v})}.$$

Proposal: Model Criticism with the Stein Witness



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

Proposal: Model Criticism with the Stein Witness



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

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. *(Riesz condition holds) $||g||^2_{\mathcal{F}} < \infty$.*
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 condition) $\lim_{||\mathbf{x}|| \to \infty} p(\mathbf{x})\mathbf{g}(\mathbf{x}) = \mathbf{0}$.*

Then, η-almost surely

$$\mathrm{FSSD}^2 = 0 \text{ if and only if } p = q, \text{ for any } J \geq 1.$$

- 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 $J = 5$.

More on FSSD^2

- When $d > 1$, the Stein witness $\mathbf{g}$ has d outputs.
- Define
$$\xi(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})} \nabla_{\mathbf{x}} [p(\mathbf{x}) k(\mathbf{x}, \mathbf{v})] \in \mathbb{R}^d.$$
- d-output Stein witness
$$\mathbf{g}(\mathbf{v}) = \mathbb{E}_{\mathbf{x} \sim q} \xi(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^d.$$
- General form:
$$\text{FSSD}^2 = \frac{1}{dJ} \sum_{j=1}^{J} \|\mathbf{g}(\mathbf{v}_j)\|_2^2,$$

where unbiased estimator $\widehat{\text{FSSD}^2}$ computable in $\mathcal{O}(d^2 J n)$.

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

- $\tau(\mathbf{x}) :=$ 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: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- Equivalently, $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\mu\|_2^2$ (mean feature).
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \stackrel{i.i.d.}{\sim} \mathcal{N}(0,1)$, 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 asymptotically consistent test.

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

- $\tau(\mathbf{x}) :=$ 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: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- Equivalently, $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\mu\|_2^2$ (mean feature).
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ 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* $\mathbb{P}$(*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 asymptotically consistent test.

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ 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: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x} \sim q}[\boldsymbol{\tau}(\mathbf{x})]$.
- Equivalently, $\mathrm{FSSD}^2 = \frac{1}{dJ} \|\boldsymbol{\mu}\|_2^2$ (mean feature).
- $\Sigma_r := \mathrm{cov}_{\mathbf{x} \sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ \times dJ}$ 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 $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ 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 asymptotically consistent test.

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

- $\tau(\mathbf{x}) :=$ 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: $\mu := \mathbb{E}_{\mathbf{x}\sim q}[\tau(\mathbf{x})]$.
- Equivalently, $\text{FSSD}^2 = \frac{1}{dJ}\|\mu\|_2^2$ (mean feature).
- $\Sigma_r := \text{cov}_{\mathbf{x}\sim r}[\tau(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ 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 $n\widehat{\text{FSSD}^2} \xrightarrow{d} \sum_{i=1}^{dJ}(Z_i^2 - 1)\omega_i$.*
   - *Easy to simulate to get p-value.*
   - *Simulation cost independent of n.*
2. *Under $H_1 : p \neq q$, we have $\sqrt{n}(\widehat{\text{FSSD}^2} - \text{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ 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 asymptotically consistent test.

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ 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: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$.
- Equivalently, $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$ (mean feature).
- $\Sigma_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ 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 $\sqrt{n}(\widehat{\mathrm{FSSD}^2} - \mathrm{FSSD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_{H_1}^2)$ 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 asymptotically consistent test.

Parameter Tuning

- Jointly optimise locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ for more test power

Proposition 2 (Approx. power for large n).

Under H_1, for large n and fixed threshold r, the test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

$$\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}\right),$$

where $\Phi =$ CDF of $\mathcal{N}(0,1)$.

- For large n, second term dominates. So

$$\arg\max_{V,\sigma_k^2} (\text{power}) \approx \arg\max_{V,\sigma_k^2} \frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}.$$

- Split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize V on tr. Test on te.

Parameter Tuning

- Jointly optimise locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ for more test power

Proposition 2 (Approx. power for large n).

Under H_1, for large n and fixed threshold r, the test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$

$$\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}\right),$$

where $\Phi =$ CDF of $\mathcal{N}(0, 1)$.

- For large n, second term dominates. So

$$\arg\max_{V, \sigma_k^2} (\text{power}) \approx \arg\max_{V, \sigma_k^2} \frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}.$$

- Split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize V on tr. Test on te.

Parameter Tuning

- Jointly optimise locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ for more test power

Proposition 2 (Approx. power for large n).

Under H_1, for large n and fixed threshold r, the test power $\mathbb{P}(\textit{reject } H_0 \mid H_1 \textit{ true})$

$$\mathbb{P}_{H_1}(n\widehat{\mathrm{FSSD}^2} > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{FSSD}^2}{\sigma_{H_1}}\right),$$

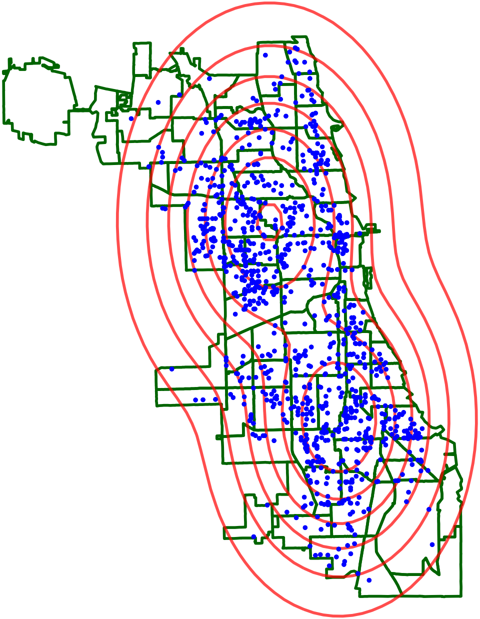
where $\Phi =$ CDF of $\mathcal{N}(0,1)$.

- For large n, second term dominates. So

$$\arg\max_{V,\sigma_k^2} \text{(power)} \approx \arg\max_{V,\sigma_k^2} \frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}.$$

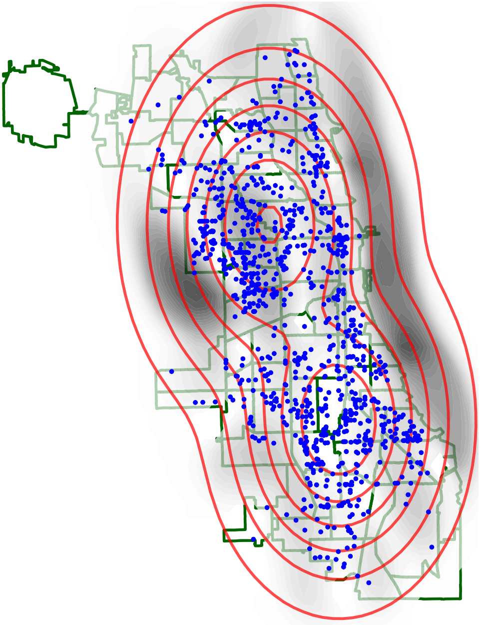
- Split $\{\mathbf{x}_i\}_{i=1}^n$ into independent training/test sets. Optimize V on tr. Test on te.

Interpretable Features: Chicago Crime



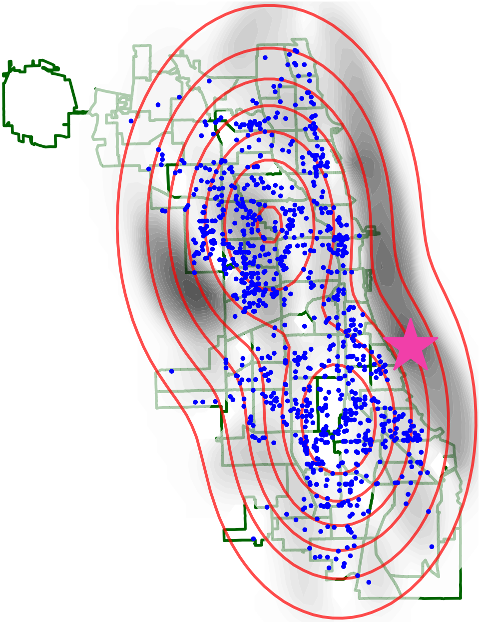
Model p = 2-component Gaussian mixture.

Interpretable Features: Chicago Crime



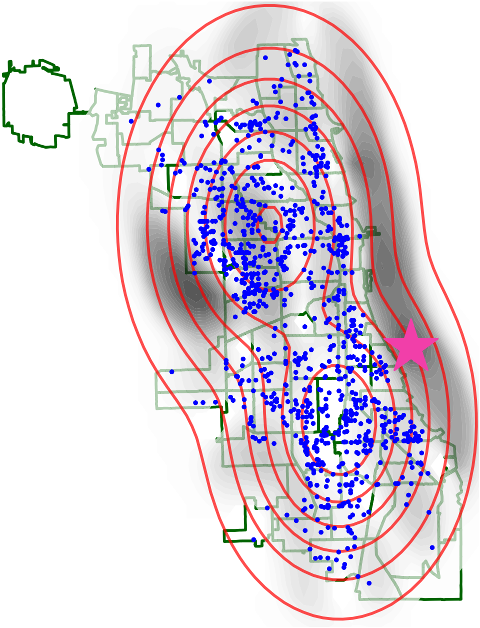
Score surface

Interpretable Features: Chicago Crime



★ = optimized $\mathbf{v}$.

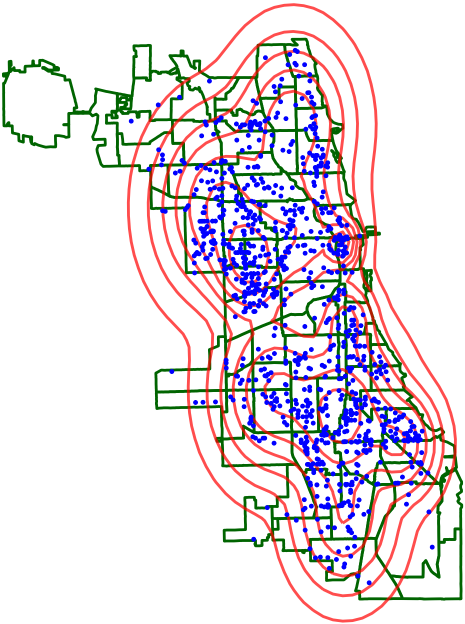
Interpretable Features: Chicago Crime



★ = optimized **v**.

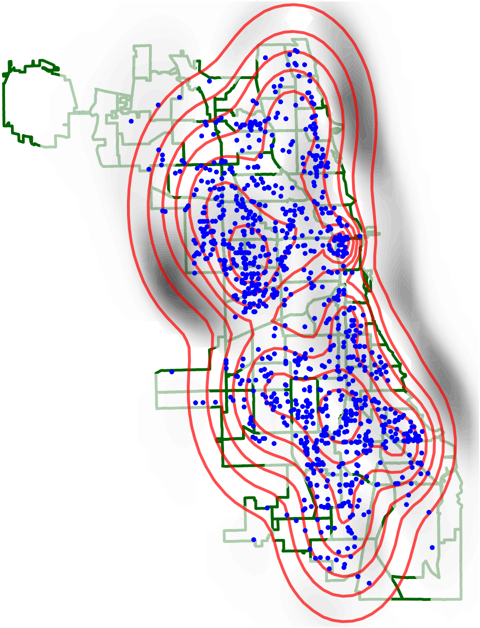
No robbery in Lake Michigan.

Interpretable Features: Chicago Crime



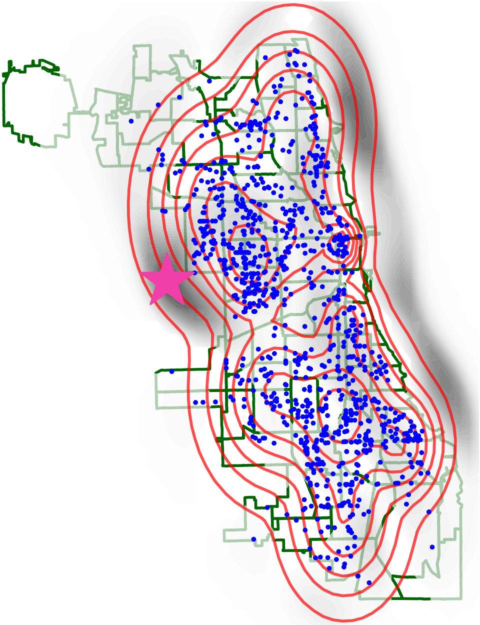
Model p = 10-component Gaussian mixture.

Interpretable Features: Chicago Crime



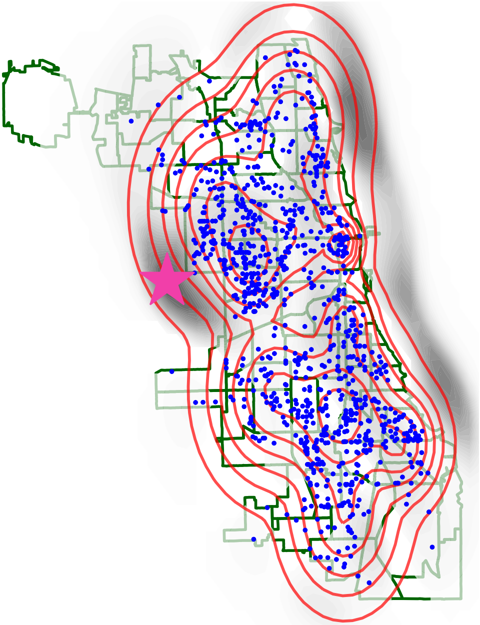
Capture the right tail better.

Interpretable Features: Chicago Crime



Still, does not capture the left tail.

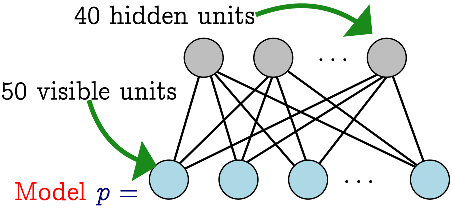
Interpretable Features: Chicago Crime



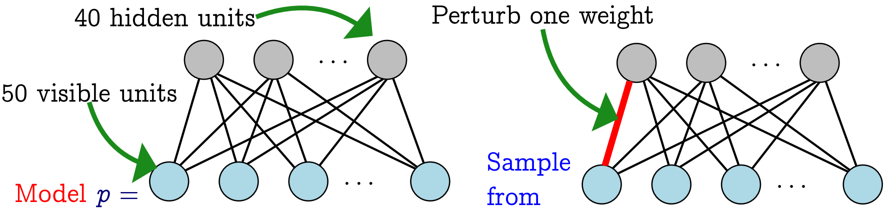
Still, does not capture the left tail.

Learned test locations are interpretable.

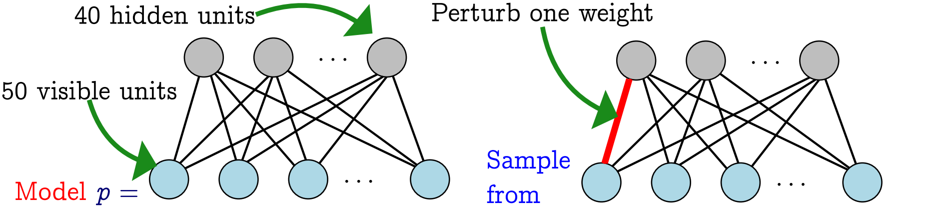
Experiment: Restricted Boltzmann Machine (RBM)

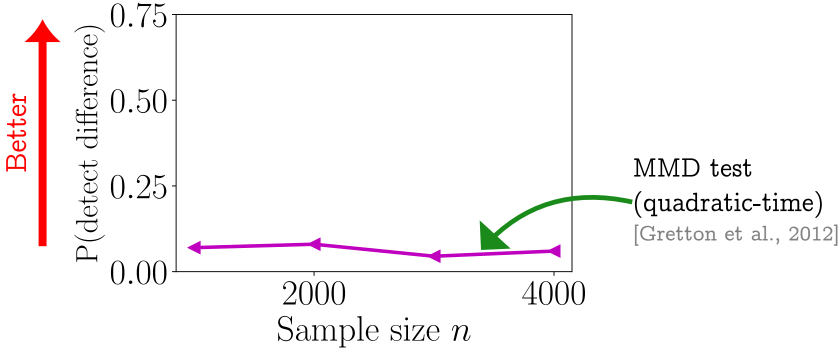


Experiment: Restricted Boltzmann Machine (RBM)

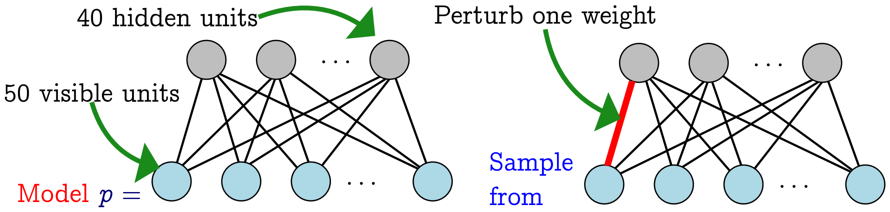


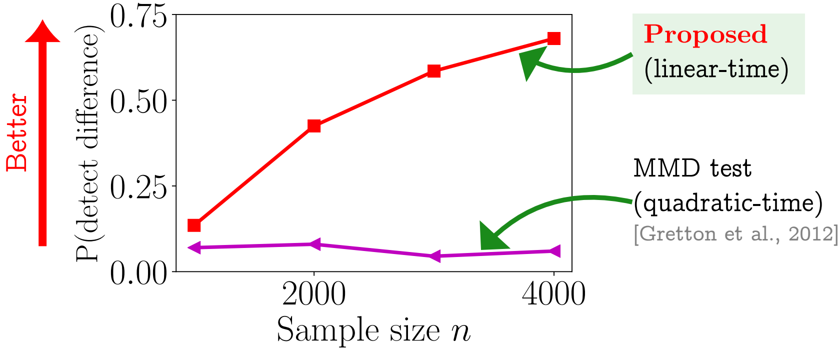
Experiment: Restricted Boltzmann Machine (RBM)





Experiment: Restricted Boltzmann Machine (RBM)





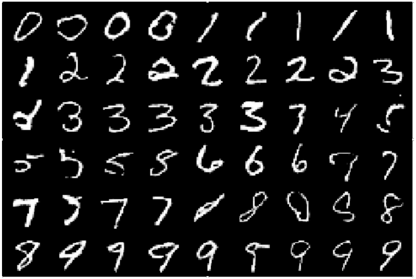
Model Criticism

"All models are wrong."

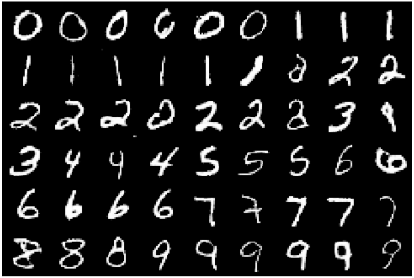
G. Box (1976)

Relative model comparison

- **Have:** two candidate models P and Q, and samples $\{x_i\}_{i=1}^n$ from reference distribution R
- **Goal:** which of P and Q is better?



Samples from GAN, Goodfellow et al. (2014)

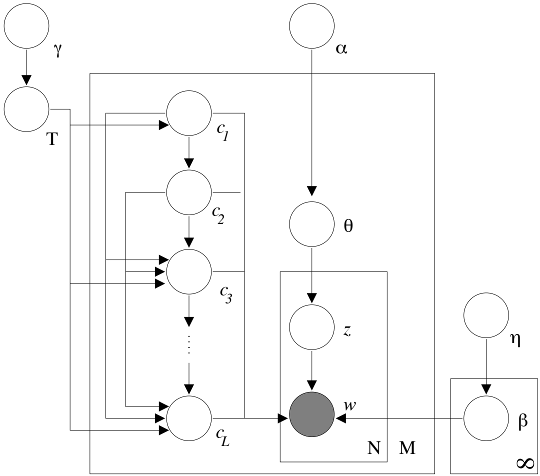


Samples from LSGAN, Mao et al. (2017)

Which model is better?

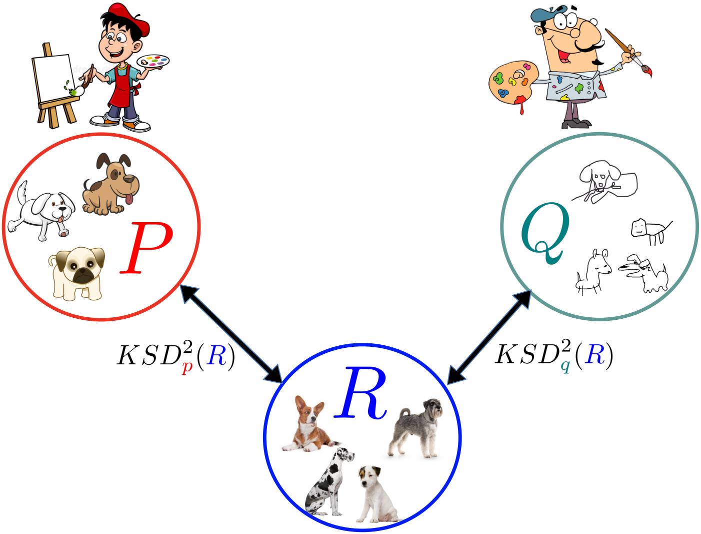
Most interesting models have latent structure

Graphical model representation of hierarchical LDA with a nested CRP prior, Blei et al. (2003)



Relative goodness-of-fit testing

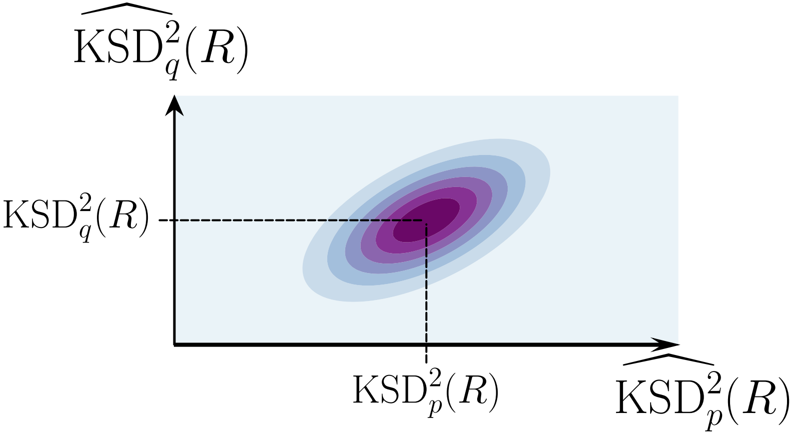
- Two generative models P and Q, data $\{x_i\}_{i=1}^{n} \sim R$.
- Neither model gives a perfect fit ($P \neq R$ and $Q \neq R$).



Joint asymptotic normality

Joint asymptotic normality when $P \neq R$ and $Q \neq R$

$$\sqrt{n}\begin{bmatrix} \widehat{\mathrm{KSD}^2_p}(R) - \mathrm{KSD}^2_p(R) \\ \widehat{\mathrm{KSD}^2_q}(R) - \mathrm{KSD}^2_q(R) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{h_p} & \sigma_{h_p h_q} \\ \sigma_{h_p h_q} & \sigma^2_{h_q} \end{bmatrix}\right)$$



Joint asymptotic normality

Joint asymptotic normality when $P \neq R$ and $Q \neq R$

$$\sqrt{n}\left[\begin{array}{c}\widehat{\mathrm{KSD}_p^2}(R) - \mathrm{KSD}_p^2(R) \\ \widehat{\mathrm{KSD}_q^2}(R) - \mathrm{KSD}_q^2(R)\end{array}\right] \xrightarrow{d} \mathcal{N}\left(\left[\begin{array}{c}0 \\ 0\end{array}\right], \left[\begin{array}{cc}\sigma_{h_p}^2 & \sigma_{h_p h_q} \\ \sigma_{h_p h_q} & \sigma_{h_q}^2\end{array}\right]\right)$$

Difference in statistics is asymptotically normal:

$$\sqrt{n}\left[\widehat{\mathrm{KSD}_p^2}(R) - \widehat{\mathrm{KSD}_q^2}(R) - \left(\mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R)\right)\right]$$
$$\xrightarrow{d} \mathcal{N}\left(0, \sigma_{h_p}^2 + \sigma_{h_q}^2 - 2\sigma_{h_p h_q}\right)$$

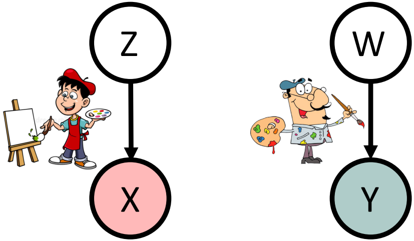
$\implies$ a statistical test with null hypothesis $\mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R) \leq 0$ is straightforward.

Latent variable models

Can we compare latent variable models with KSD?

$$p(x) = \int p(x|z)p(z)dz$$

$$q(y) = \int q(y|w)p(w)dw$$



Recall multi-dimensional Stein operator:

$$[T_p f](x) = \left\langle \underbrace{\frac{\nabla p(x)}{p(x)}}_{(a)}, f(x) \right\rangle + \langle \nabla, f(x) \rangle .$$

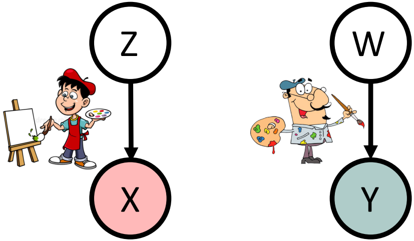
Expression (a) requires marginal $p(x)$, often intractable...

Latent variable models

Can we compare latent variable models with KSD?

$$p(x) = \int p(x|z)p(z)\,dz$$

$$q(y) = \int q(y|w)p(w)\,dw$$



Recall multi-dimensional Stein operator:

$$[T_p f](x) = \left\langle \underbrace{\frac{\nabla p(x)}{p(x)}}_{(a)}, f(x) \right\rangle + \langle \nabla, f(x) \rangle .$$

Expression (a) requires marginal $p(x)$, often intractable...

...but sampling can be straightforward!

Monte Carlo approximation

Approximate the integral using $\{z_j\}_{j=1}^m \sim p(z)$:

$$p(x) = \int p(x|z)p(z)\,dz$$
$$\approx p_m(x) = \frac{1}{m}\sum_{j=1}^{m} p(x|z_j)$$

Estimate KSDs with approxiomate densities:

$$\widehat{\mathrm{KSD}_p^2}(R) - \widehat{\mathrm{KSD}_q^2}(R) \approx \widehat{\mathrm{KSD}_{p_m}^2}(R) - \widehat{\mathrm{KSD}_{q_m}^2}(R)$$

Monte Carlo approximation

Approximate the integral using $\{z_j\}_{j=1}^m \sim p(z)$:

$$p(x) = \int p(x|z)p(z)\,dz$$
$$\approx p_m(x) = \frac{1}{m}\sum_{j=1}^m p(x|z_j)$$

Estimate KSDs with approxiomate densities:

$$\widehat{\mathrm{KSD}^2_p}(R) - \widehat{\mathrm{KSD}^2_q}(R) \approx \widehat{\mathrm{KSD}^2_{p_m}}(R) - \widehat{\mathrm{KSD}^2_{q_m}}(R)$$

Recall

$$\sqrt{n}\left[\widehat{\mathrm{KSD}^2_p}(R) - \widehat{\mathrm{KSD}^2_q}(R) - \left(\mathrm{KSD}^2_p(R) - \mathrm{KSD}^2_q(R)\right)\right]$$
$$\xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{h_p} + \sigma^2_{h_q} - 2\sigma_{h_p h_q}\right)$$

$\rightarrow$ if m is large, can we simply substitute p_m and q_m ?

Simple proof of concept

Check $\widehat{\mathrm{KSD}}_p^2(R) \approx \widehat{\mathrm{KSD}}_{p_m}^2(R)$ with a toy model:

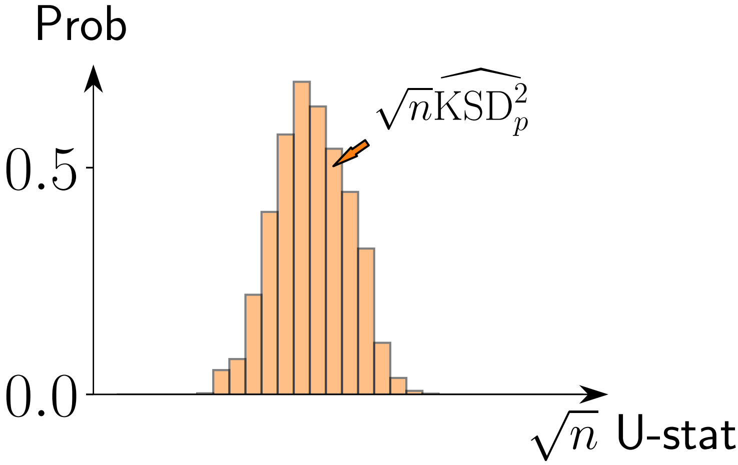
- Model: Beta-Binomial $\mathrm{BetaBinom}(\alpha, \beta)$

$$p(x|z) = \binom{N}{x} z^x (1-z)^{n-x}, \;\; p(z) = \mathrm{Beta}(a, b)$$

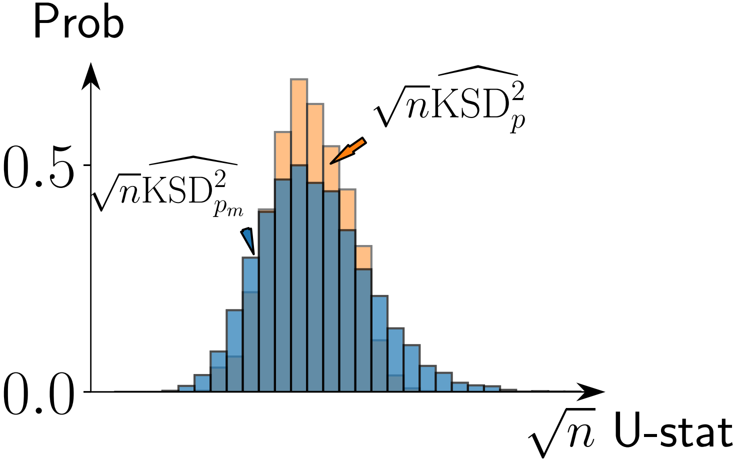
  - Latent $z \in (0, 1)$: success probability for binomial likelihood
  - Marginal $p(x)$: tractable (given by the beta function)

- Generate $\sqrt{n}\widehat{\mathrm{KSD}}_p^2(R)$ and $\sqrt{n}\widehat{\mathrm{KSD}}_{p_m}^2(R)$
  $\rightarrow$ what do their distribution look like?

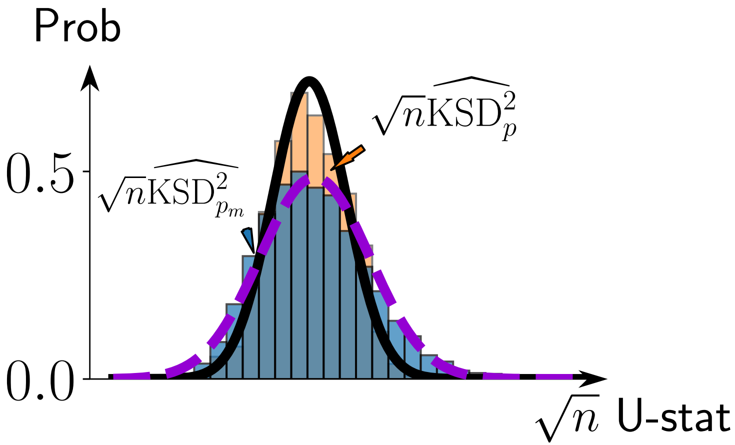
Effect of sampling the latents (Beta-binomial)



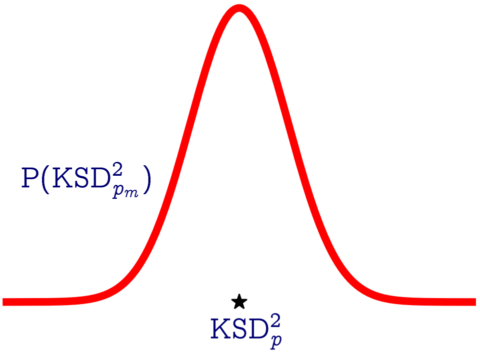
Effect of sampling the latents (Beta-binomial)



Effect of sampling the latents (Beta-binomial)

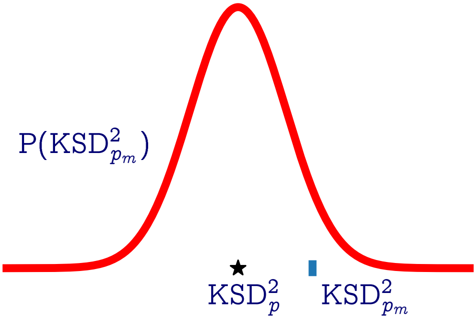


Why this happens



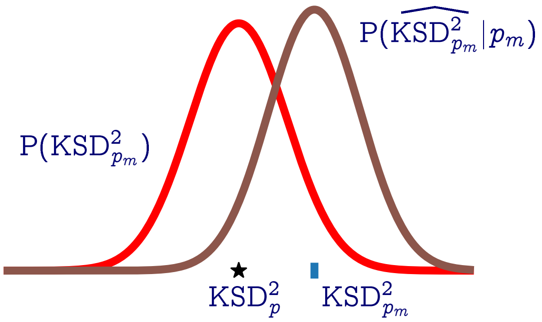
$KSD^2_{p_m}(R)$ is normally distributed around $KSD^2_p(R)$
(approximation error)

Why this happens



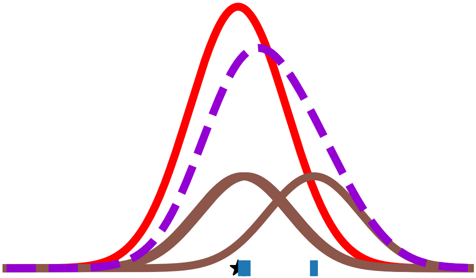
Approximation p_m gives a random draw $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



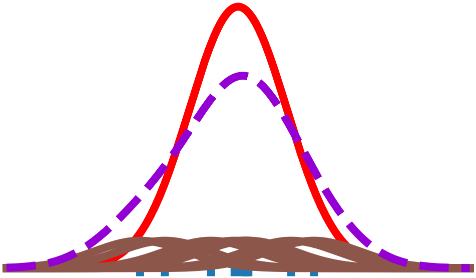
$\widehat{\mathrm{KSD}}^2_{p_m}(R)$ is normally distributed around $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



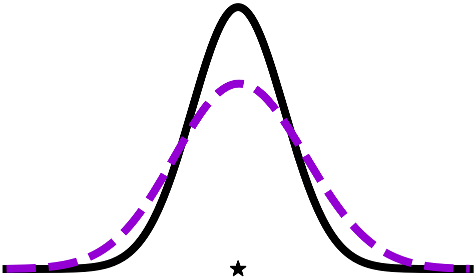
Distribution of $\widehat{\mathrm{KSD}^2_{p_m}}(R)$ is
averaged over random draws of $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



Distribution of $\widehat{\mathrm{KSD}^2_{p_m}}(R)$ is
averaged over random draws of $\mathrm{KSD}^2_{p_m}(R)$

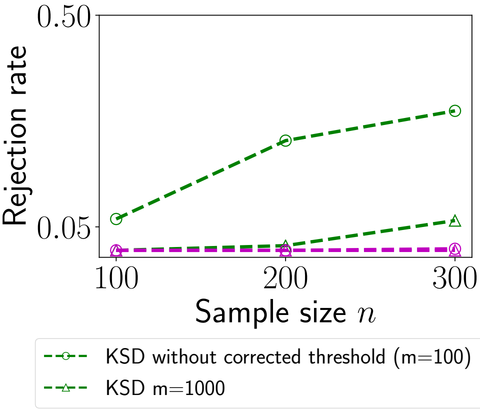
Why this happens



$\widehat{\mathrm{KSD}^2_{p_m}}(R)$ has a higher variance than $\widehat{\mathrm{KSD}^2_{p}}(R)$

Correction for this effect

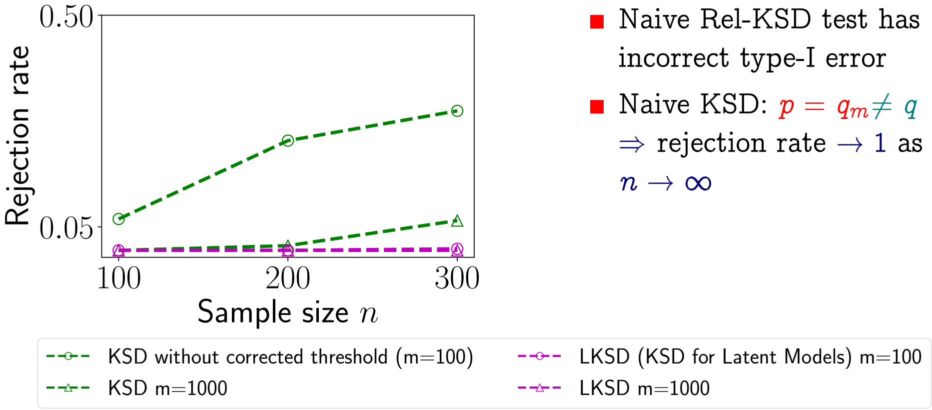
- BetaBinomial models with $p = q_m$ vs q
  $\rightarrow$numerical vs closed-form marginalisation.
- With correction for increased $\widehat{\mathrm{KSD}}^2_{q_m}(R)$ variance,
  null accepted w.p. $1 - \alpha$.



- $Q = \mathrm{BetaBinom}(5 + a, 1 + b)$
- $P = q_m$
- $R = \mathrm{BetaBinom}(a, b)$
- $k(x, x') = \exp(-\mathbb{I}(x \neq x'))$
- $\alpha = 0.05$

Correction for this effect

- BetaBinomial models with $p = q_m$ vs q
  $\rightarrow$numerical vs closed-form marginalisation.
- With correction for increased $\widehat{\mathrm{KSD}^2_{q_m}}(R)$ variance,
  null accepted w.p. $1-\alpha$.



- Naive Rel-KSD test has incorrect type-I error
- Naive KSD: $p = q_m \neq q$ $\Rightarrow$ rejection rate $\rightarrow 1$ as $n \rightarrow \infty$

Asymptotics for approximate KSD

We have asymptotic normality for $\mathrm{KSD}^2_{p_m}(R)$,

$$\sqrt{m}(\mathrm{KSD}^2_{p_m}(R) - \mathrm{KSD}^2_{p}(R)) \xrightarrow{d} \mathcal{N}(0, \gamma_p^2)$$

The fine print:

- $\inf_x p(x) > 0$
- $\sup_x \left| \frac{dp(x)}{dx} \right| < \infty$
- (Uniform CLT) Likelihoods $\{p(x|\cdot) | x \in \mathcal{X}\}$ and derivatives $\{\frac{d}{dx} p(x|\cdot) | x \in \mathcal{X}\}$ are $p(z)$ - Donsker class

Asymptotic distribution for relative KSD test

Asymptotic distribution of approximate KSD estimate
$(n, m) \to \infty$, $\frac{n}{m} \to r \in [0, \infty)$:

$$\sqrt{n}\left[\left(\widehat{\mathrm{KSD}^2_{p_m}}(R) - \widehat{\mathrm{KSD}^2_{q_m}}(R)\right) - \left(\mathrm{KSD}^2_p(R) - \mathrm{KSD}^2_q(R)\right)\right] \xrightarrow{d} \mathcal{N}(0, c^2)$$

where

- $c = \sigma_{pq}\sqrt{1 + r(\gamma_{pq}/\sigma_{pq})^2}$
- $\gamma^2_{pq} = \lim_{m\to\infty} m \cdot \mathrm{Var}\left[\mathbf{E}_{x,x'} h_{p_m}(x, x') - \mathbf{E}_{x,x'} h_{q_m}(x, x')\right]$
- $\sigma^2_{pq} = \lim_{n\to\infty} n \cdot \mathrm{Var}\left[\widehat{\mathrm{KSD}^2_p}(R) - \widehat{\mathrm{KSD}^2_q}(R)\right]$

Fine print:

- $h_p(x, x') - h_q(x, x')$ has a finite third moment
- Additional technical conditions

Relative test, further detail

Theorem (Asymptotic distribution of random kernel U-statistic).

- *Let*
  - $U_{n,m}$ *: a U-statistic defined by a random U-statistic kernel* H_m
  - U_n *: a U-statistic defined by a fixed U-statistic kernel* h
- *Assume that*
  - $\sigma^2_{H_m} \to \sigma^2_h$ *in probability*
  - $\nu_3(H_m) \to \nu_3(h) < \infty$ *in probability*
    where $\nu_3(H_m) = \mathbf{E}_{x,x'} \left| H_m(x,x') - \mathbf{E}_{x,x'} H_m(x,x') \right|^3$
  - $Y_m := \sqrt{m}\Big(\mathbf{E}_n[U_{n,m}|H_m] - \mathbf{E}_n[U_n]\Big) \xrightarrow{d} Y$
- *Then, with* $n/m \to r \in [0,\infty)$,

$$\lim_{n,m\to\infty} \Pr\left[\sqrt{n}(U_{n,m} - \mathbf{E}_n U_n) < t\right] = \mathbf{E}_Y \left[\Phi\left(\frac{t - \sqrt{r}\,Y}{\sigma_h}\right)\right]$$

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network $\mathrm{SBN}(W)$:

$$R(x|z) = \mathrm{sigmoid}(Wz),\ \ R(z) = \mathcal{N}(0, I),\ \ W \in \mathbb{R}^{30\times 10}$$

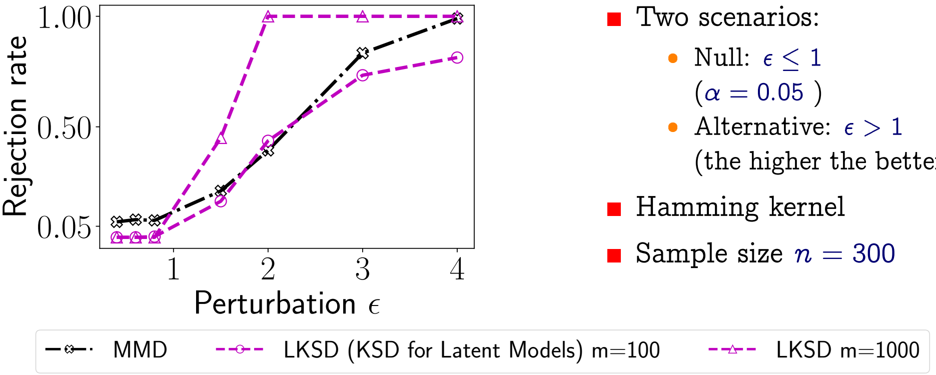
- Models: $P = \mathrm{SBN}(W + \epsilon[\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$, $Q = \mathrm{SBN}(W + [\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$
- Only the first column of weight W is perturbed by ϵ

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network $\mathrm{SBN}(W)$:

$$R(x|z) = \mathrm{sigmoid}(Wz),\ \ R(z) = \mathcal{N}(0, I),\ \ W \in \mathbb{R}^{30\times 10}$$

- Models: $P = \mathrm{SBN}(W + \epsilon[\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$, $Q = \mathrm{SBN}(W + [\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$
- Only the first column of weight W is perturbed by ϵ



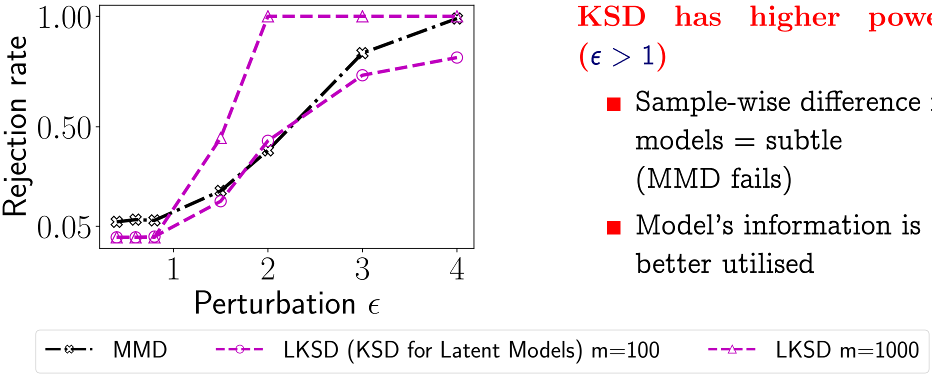
- Two scenarios:
  - Null: $\epsilon \leq 1$ ($\alpha = 0.05$)
  - Alternative: $\epsilon > 1$ (the higher the better)
- Hamming kernel
- Sample size $n = 300$

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network $\text{SBN}(W)$:

$$R(x|z) = \text{sigmoid}(Wz), \;\; R(z) = \mathcal{N}(0, I), \;\; W \in \mathbb{R}^{30\times 10}$$

- Models: $P = \text{SBN}(W + \epsilon[\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$, $Q = \text{SBN}(W + [\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$
- Only the first column of weight W is perturbed by ϵ



KSD has higher power ($\epsilon > 1$)

- Sample-wise difference in models = subtle (MMD fails)
- Model's information is better utilised

Papers referenced

A Linear-Time Kernel Goodness-of-Fit Test.
Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, Arthur Gretton
https://arxiv.org/abs/1705.07673

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

A Kernel Stein Test for Comparing Latent Variable Models
Heishiro Kanagawa, Wittawat Jitkrittum, Lester Mackey, Kenji Fukumizu, Arthur Gretton
https://arxiv.org/abs/1907.00586

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Efficiency comparison, linear-time 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.

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

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

Bahadur slope

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

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

- Bahadur efficiency = ratio of slopes of two tests.

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.

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

- Typically $\text{pval}_n \approx \exp\left(-\frac{1}{2} c(\boldsymbol{\theta}) n\right)$ where $c(\boldsymbol{\theta}) > 0$ under H_1, and $c(\mathbf{0}) = 0$. [?].
- $c(\boldsymbol{\theta})$ 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.

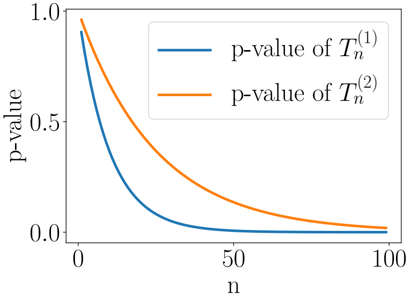
- Bahadur efficiency = ratio of slopes of two tests.

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.

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

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



Bahadur slope

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

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

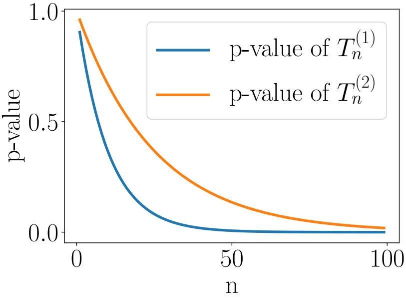
- Bahadur efficiency = ratio of slopes of two tests.

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.

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

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



Bahadur slope

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

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

- Bahadur efficiency = ratio of slopes of two tests.

Bahadur Slopes of FSSD and LKS

Theorem 2.

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* $\Sigma_p := \mathrm{cov}_{\mathbf{x}\sim p}[\tau(\mathbf{x})]$.

Theorem 3.

The Bahadur slope of the linear-time kernel Stein (LKS) statistic $\sqrt{n}\widehat{S_l^2}$ *is*

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

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

- Let's consider a specific case ...

Bahadur Slopes of FSSD and LKS

Theorem 2.

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* $\Sigma_p := \mathrm{cov}_{\mathbf{x}\sim p}[\tau(\mathbf{x})]$.

Theorem 3.

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.*

- Let's consider a specific case ...

Gaussian Mean Shift Problem

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

- Assume $J = 1$ feature 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 $= \kappa^2$).

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

Gaussian Mean Shift Problem

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

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

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

Gaussian Mean Shift Problem

Theorem 4 (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*

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