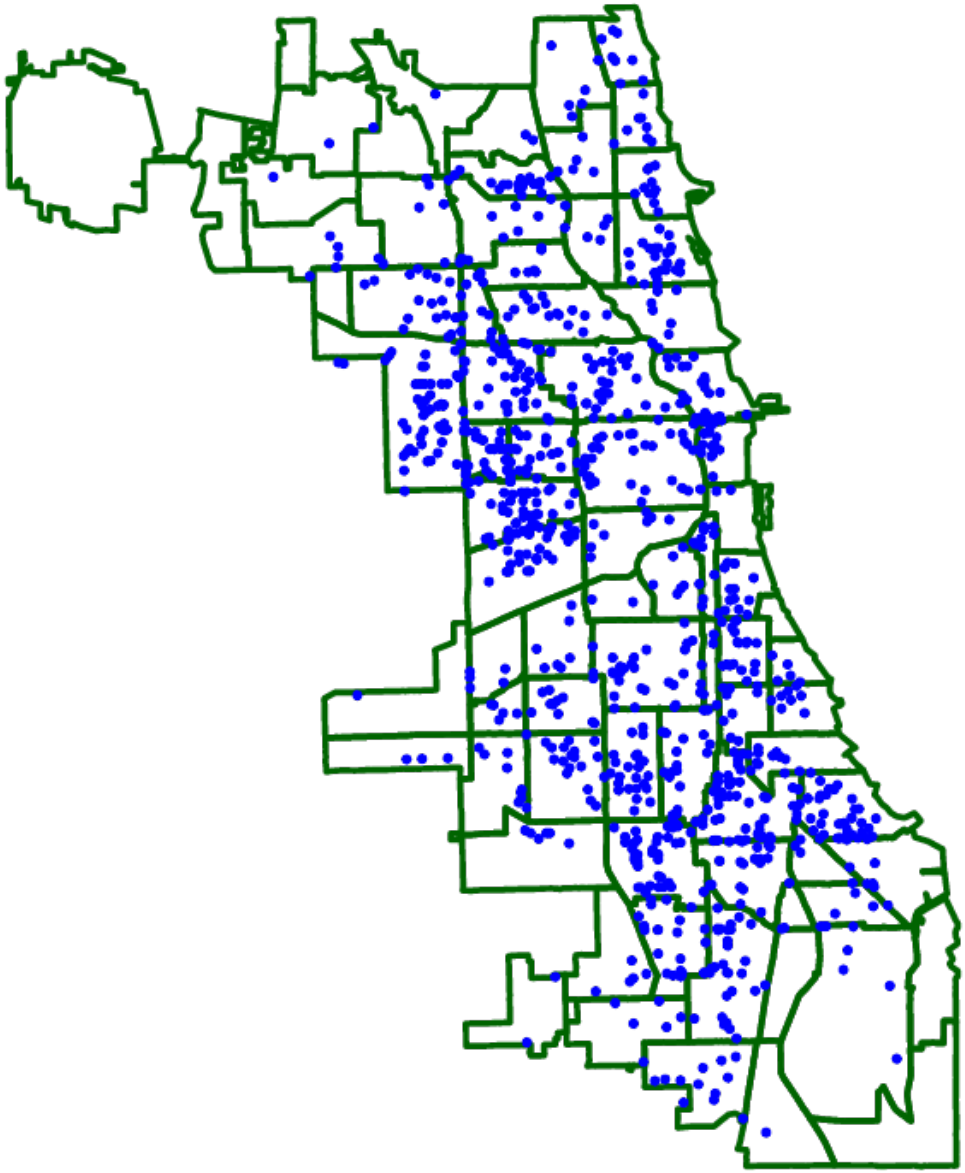
# RKHS in ML: Comparing a Sample and a Model

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

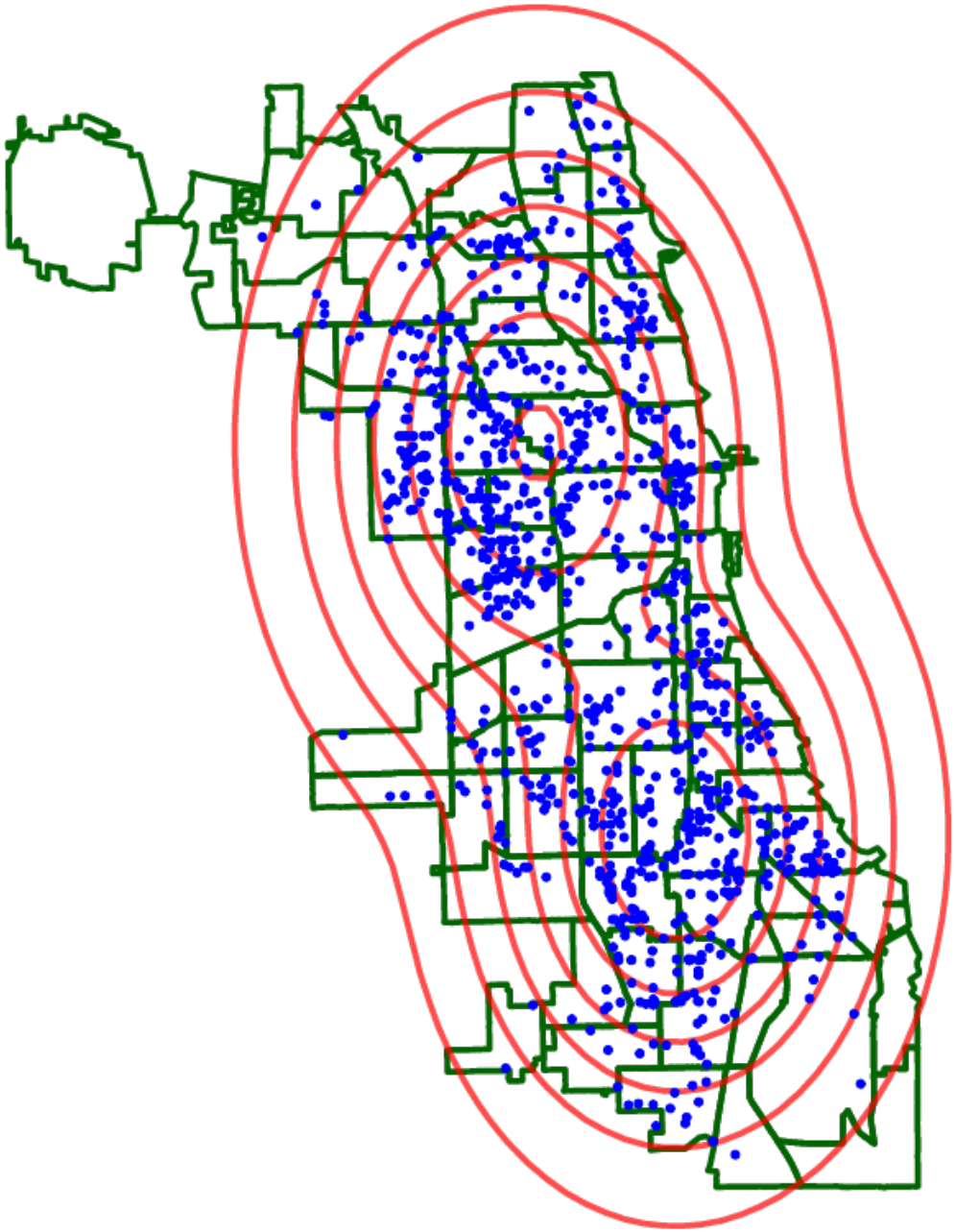
November 27, 2024

# Model Criticism



Data = robbery events in Chicago in 2016.

# Model Criticism



Is this a good model?

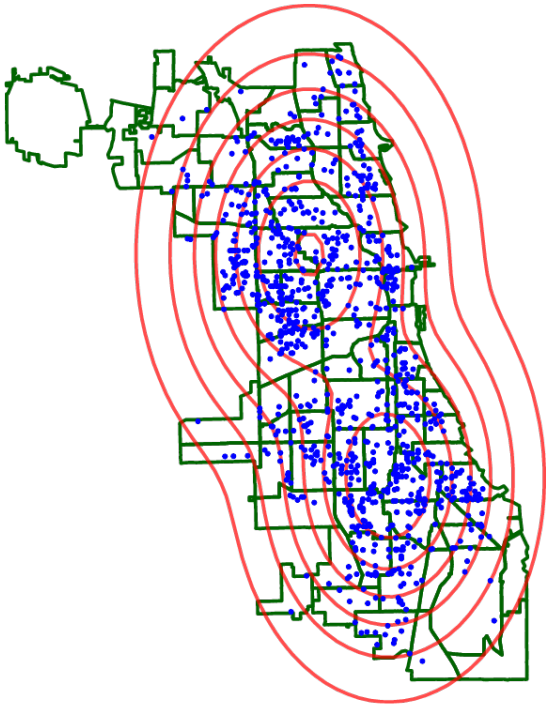
# Model Criticism

"All models are wrong."

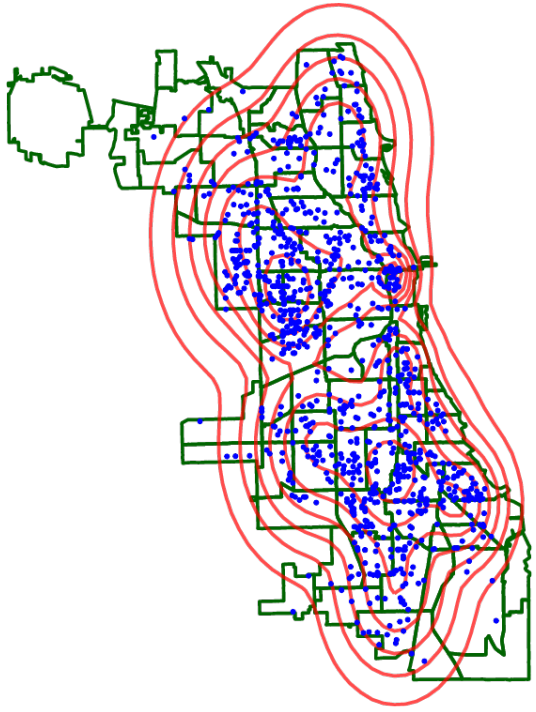
G. Box (1976)

# 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?



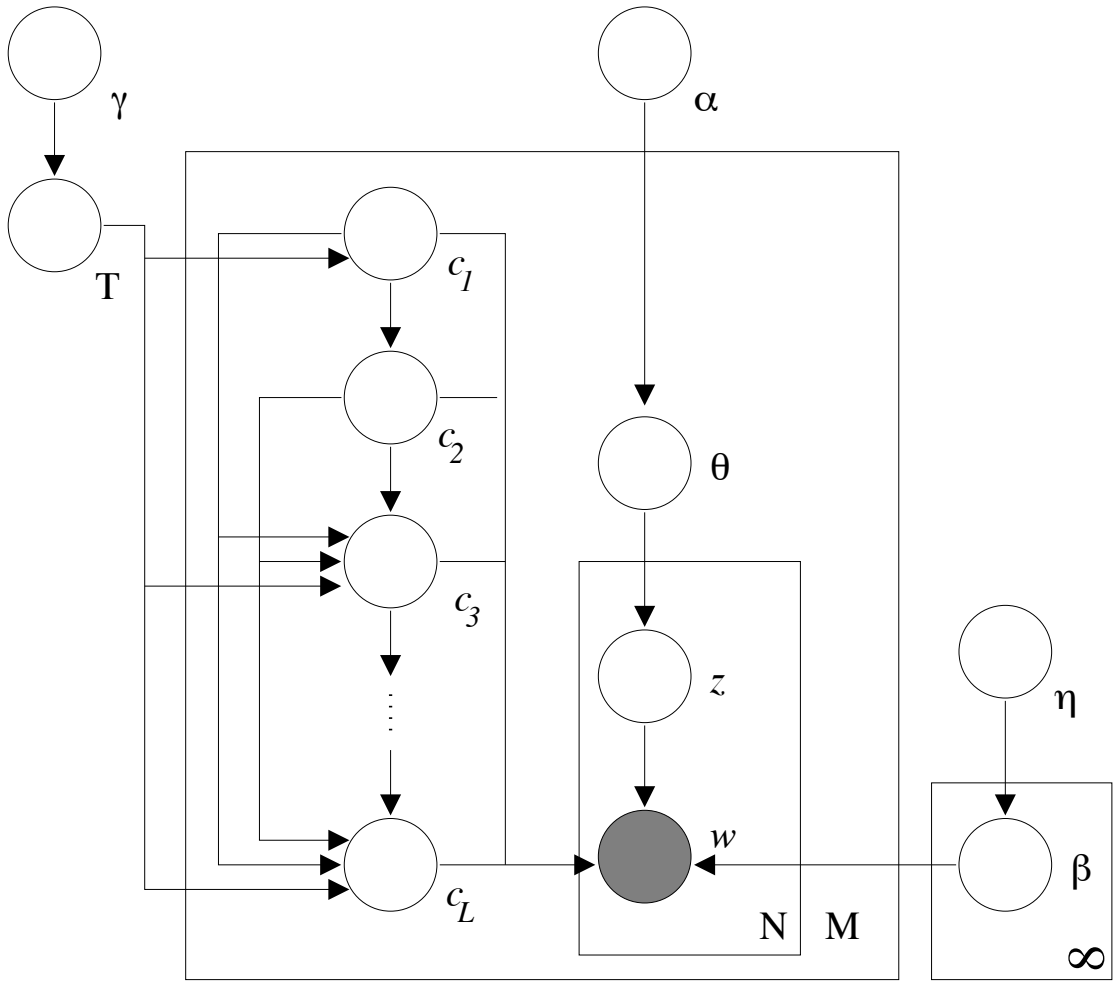
$P$ : two components



$Q$ : ten components

# Most interesting models have latent structure

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



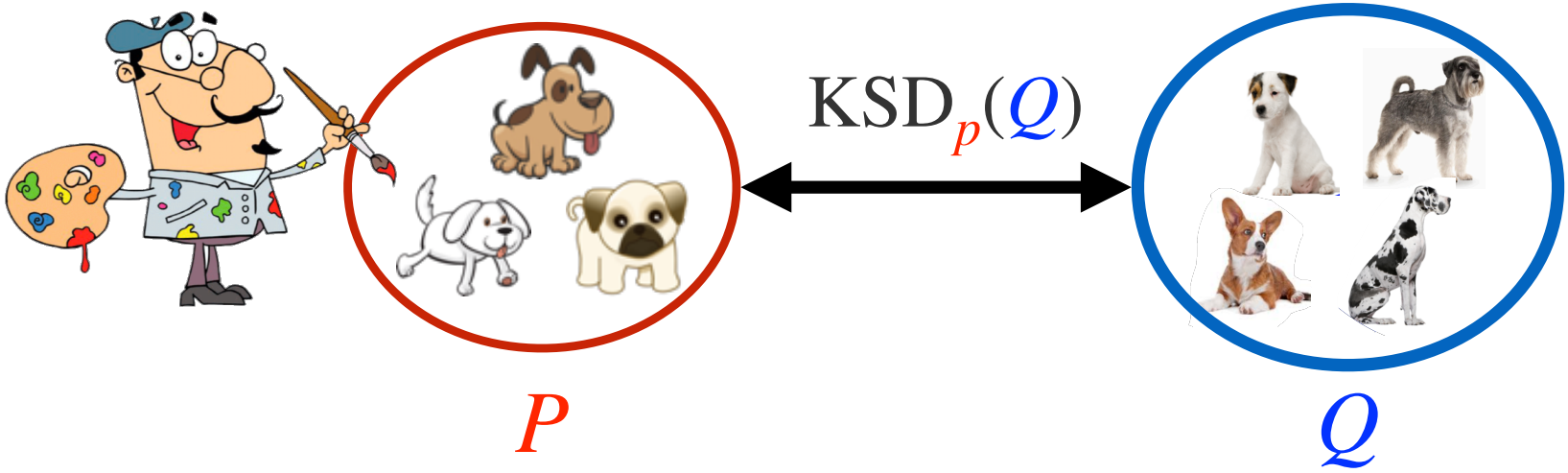
# Outline

## Relative goodness-of-fit tests for Models with Latent Variables

- The Maximum Mean Discrepancy: an integral probability metric
  - maximize difference in expectations using an RKHS witness class
- The kernel Stein discrepancy
  - Comparing a sample and a model: Stein modification of the witness class
- Constructing a relative hypothesis test using the KSD
- Relative hypothesis tests with latent variables

# Kernel Stein Discrepancy

- Model $P$, data $\{x_i\}_{i=1}^n \sim Q$.
- "All models are wrong" ($P \neq Q$).

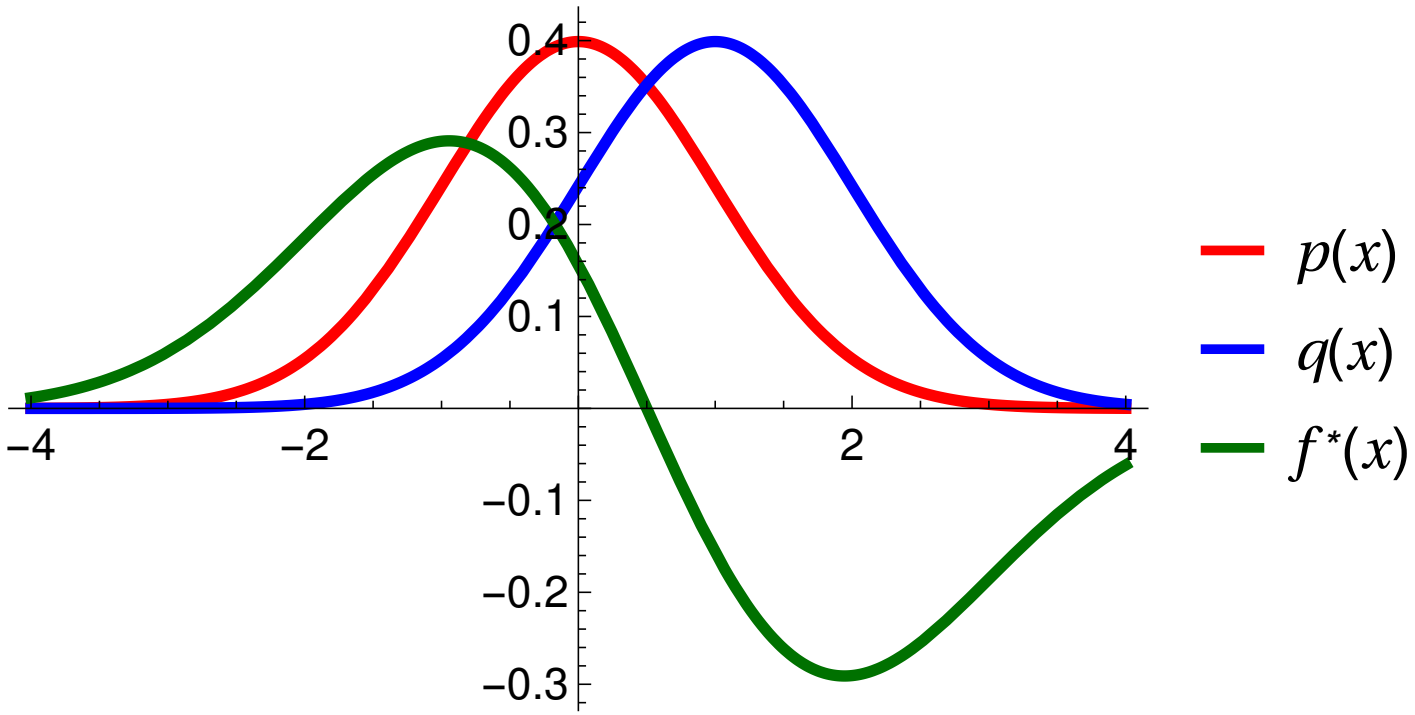


# Comparing a sample and model

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

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

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



# Stein idea

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

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

we use the (1-D) Langevin Stein operator

$$\left[\mathcal{A}_p f\right](x) = \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x)\right)$$

Then

$$\mathrm{E}_p \mathcal{A}_p f = 0$$

subject to appropriate boundary conditions.

$$\mathrm{E}_p \left[\mathcal{A}_p f\right] = \int \left[\frac{1}{\cancel{p(x)}} \frac{d}{dx} \left(f(x) p(x)\right)\right] \cancel{p(x)} dx = \left[f(x) p(x)\right]_{-\infty}^{\infty}$$

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

# Stein idea

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

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

we use the (1-D) Langevin Stein operator

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

Then

$$\mathrm{E}_p \mathcal{A}_p f = 0$$

subject to appropriate boundary conditions.

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

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

# Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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

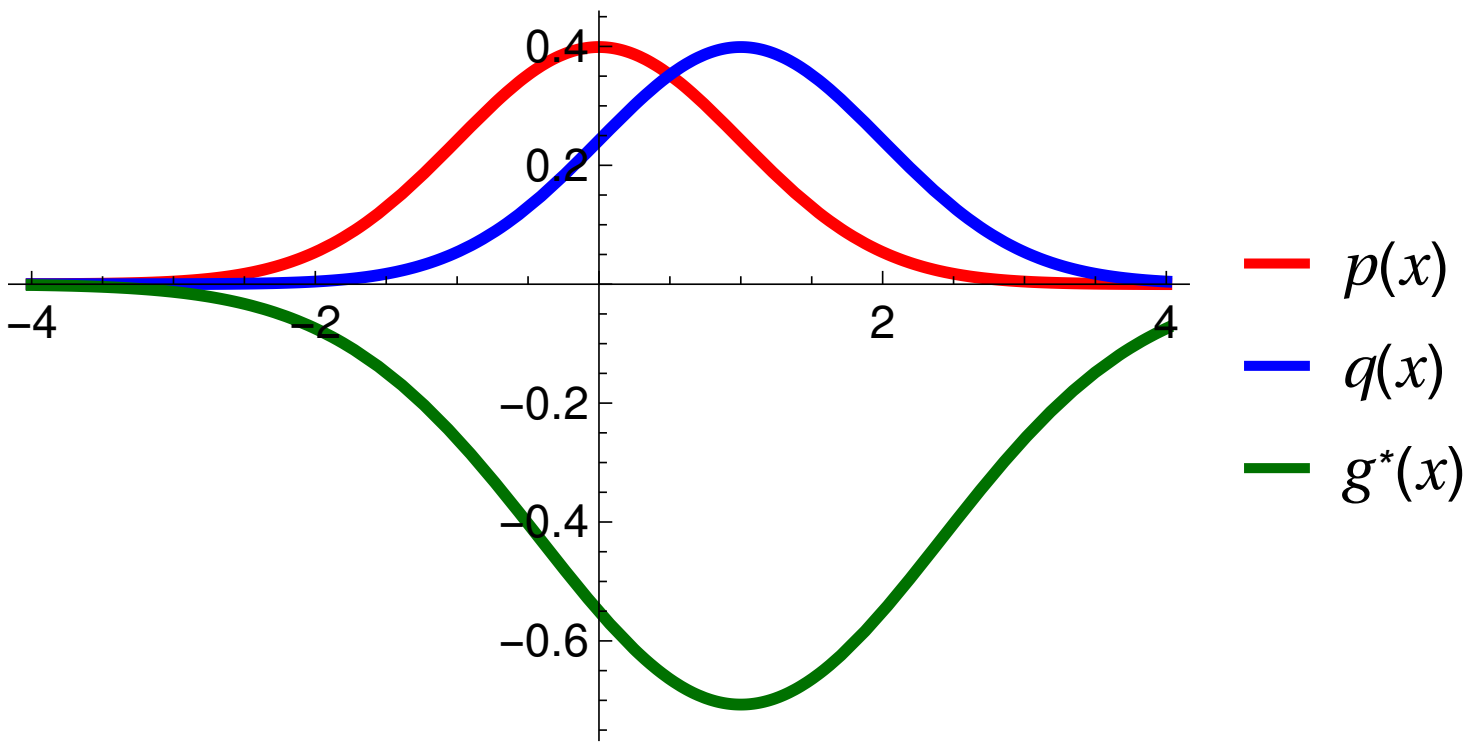
# Kernel Stein Discrepancy

Stein operator

$$\mathcal{A}_p f = \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x))$$

Kernel Stein Discrepancy (KSD)

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



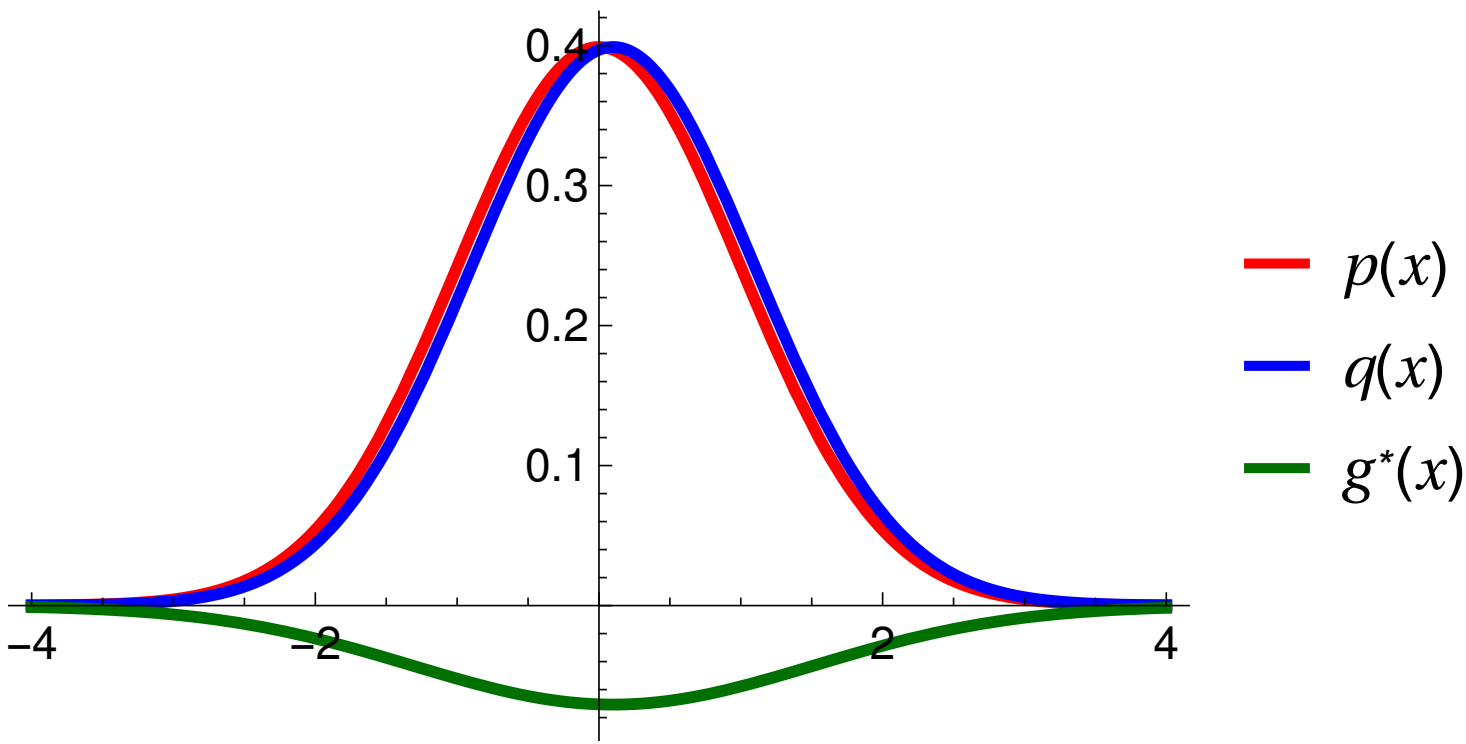
# Kernel Stein Discrepancy

Stein operator

$$\mathcal{A}_p f = \frac{1}{p(x)} \frac{d}{dx} (f(x)p(x))$$

Kernel Stein Discrepancy (KSD)

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



# Computing the kernel Stein discrepancy

How do we get the KSD in closed form (with kernels)?

Can we define "Stein features"?

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

# Computing the kernel Stein discrepancy

How do we get the KSD in closed form (with kernels)?

Can we define "Stein features"?

$$
\begin{aligned}
\left[\mathcal{A}_p f\right](x) &= \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x)\right) \\
&= \frac{d}{dx} f(x) + f(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\
&= f(x) \frac{d}{dx} \log p(x) + \frac{d}{dx} f(x)
\end{aligned}
$$

# Computing the kernel Stein discrepancy

How do we get the KSD in closed form (with kernels)?

Can we define "Stein features"?

$$
\begin{aligned}
\left[\mathcal{A}_p f\right](x) &= \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x)\right) \\
&= \frac{d}{dx} f(x) + f(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\
&= f(x) \frac{d}{dx} \log p(x) + \frac{d}{dx} f(x) \\
&\stackrel{?}{=} \langle f, \underbrace{\xi(x)}_{\text{stein features}} \rangle_{\mathcal{F}}
\end{aligned}
$$

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

# Computing the kernel Stein discrepancy

How do we get the KSD in closed form (with kernels)?

Can we define "Stein features"?

$$
\begin{aligned}
\left[\mathcal{A}_p f\right](x) &= \frac{1}{p(x)} \frac{d}{dx} \left(f(x) p(x)\right) \\
&= \frac{d}{dx} f(x) + f(x) \frac{1}{p(x)} \frac{d}{dx} p(x) \\
&= f(x) \frac{d}{dx} \log p(x) + \frac{d}{dx} f(x) \\
&\stackrel{?}{=} \langle f, \underbrace{\xi(x)}_{\text{stein features}} \rangle_{\mathcal{F}}
\end{aligned}
$$

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

Intended destination:

$$
\mathrm{KSD}(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} \langle g, E_{z \sim q} \xi_z \rangle_{\mathcal{F}} = \| E_{z \sim q} \xi_z \|_{\mathcal{F}}
$$

# Stein RKHS features

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

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

Steinwart, Christmann, Support Vector Machines (2008), Lemma 4.3.4

# Stein RKHS features

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

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

Using kernel derivative trick in $(a)$,

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

Steinwart, Christmann, Support Vector Machines (2008), Lemma 4.3.4

# Proof: kernel derivative trick (on $[-\pi, \pi]$)

Proof: differentiable translation invariant $k(x, x')$, $\mathcal{X} := [-\pi, \pi]$, periodic boundary

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

# Proof: kernel derivative trick (on $[-\pi, \pi]$)

Proof: differentiable translation invariant $k(x, x')$, $\mathcal{X} := [-\pi, \pi]$, periodic boundary

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

Fourier series representation:

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \qquad \hat{f}_\ell = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \exp(-\imath \ell x)\, dx.$$

Fourier series representation of derivative:

$$\frac{d}{dx} f(x) \xrightarrow{F.S.} \left\{ (\imath \ell) \hat{f}_\ell \right\}_{\ell=-\infty}^{\infty}$$

# Proof: kernel derivative trick (on $[-\pi, \pi]$)

Proof: differentiable translation invariant $k(x, x')$, $\mathcal{X} := [-\pi, \pi]$, periodic boundary

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

Fourier series representation:

$$f(x) = \sum_{\ell=-\infty}^{\infty} \hat{f}_\ell \exp(\imath \ell x), \qquad \hat{f}_\ell = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \exp(-\imath \ell x)\, dx.$$

Fourier series representation of derivative:

$$\frac{d}{dx} f(x) \xrightarrow{F.S.} \left\{ (\imath \ell) \hat{f}_\ell \right\}_{\ell=-\infty}^{\infty} \qquad \frac{d}{dx} k(x, \cdot) = \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \exp(\imath \ell (x - \cdot))$$

# Proof: kernel derivative trick (on $[-\pi, \pi]$)

From previous slide,

$$\frac{d}{dx} f(x) \xrightarrow{F.S.} \left\{ (\imath \ell) \hat{f}_\ell \right\}_{\ell=-\infty}^{\infty} \qquad \frac{d}{dx} k(x, \cdot) = \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \exp\left(\imath \ell (x - \cdot)\right)$$

We can write

$$\left\langle f, \frac{d}{dx} k(x, \cdot) \right\rangle_{\mathcal{F}} = \sum_{\ell=-\infty}^{\infty} \frac{\left(\hat{f}_\ell\right) \left( \overline{-\imath \ell \hat{k}_\ell \exp(-\imath \ell x)} \right)}{\hat{k}_\ell}$$

$$= \sum_{\ell=-\infty}^{\infty} (\imath \ell) \left(\hat{f}_\ell\right) \left(\exp(\imath \ell x)\right) = \frac{d}{dx} f(x).$$

# Does the mean stein embedding exist?

The KSD is written:

$$
\begin{aligned}
\left[T_p f\right](z) &= \left(\frac{d}{dz}\log p(z)\right) f(z) + \frac{d}{dz} f(z) \\
&= \left\langle f, \left(\frac{d}{dz}\log p(z)\right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot) \right\rangle_{\mathcal{F}} \\
&=: \langle f, \xi_z \rangle_{\mathcal{F}} .
\end{aligned}
$$

# Does the mean stein embedding exist?

The KSD is written:

$$
\begin{aligned}
[T_p f](z) &= \left(\frac{d}{dz}\log p(z)\right) f(z) + \frac{d}{dz} f(z) \\
&= \left\langle f, \left(\frac{d}{dz}\log p(z)\right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot) \right\rangle_{\mathcal{F}} \\
&=: \langle f, \xi_z \rangle_{\mathcal{F}} .
\end{aligned}
$$

Next step: show that

$$
E_{z\sim q}\left[T_p f\right] = E_{z\sim q}\left\langle f, \xi_z \right\rangle_{\mathcal{F}} = \left\langle f, E_{z\sim q}\xi_z \right\rangle_{\mathcal{F}} .
$$

# Does the mean stein embedding exist?

The KSD is written:

$$
\begin{aligned}
\left[T_p f\right](z) &= \left(\frac{d}{dz}\log p(z)\right) f(z) + \frac{d}{dz} f(z) \\
&= \left\langle f, \left(\frac{d}{dz}\log p(z)\right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot) \right\rangle_{\mathcal{F}} \\
&=: \langle f, \xi_z \rangle_{\mathcal{F}} .
\end{aligned}
$$

Next step: show that

$$
E_{z\sim q}\left[T_p f\right] = E_{z\sim q}\left\langle f, \xi_z \right\rangle_{\mathcal{F}} = \left\langle f, E_{z\sim q}\xi_z \right\rangle_{\mathcal{F}} .
$$

Riesz theorem!

# Next step: taking expectations

Riesz theorem: need boundedness,

$$|E_{z\sim q} \langle f, \xi_z \rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \lambda$$

for some $\lambda \in \mathbb{R}$.

By Jensen and Cauchy-Schwarz,

$$\begin{aligned} |E_{z\sim q} \langle f, \xi_z \rangle_{\mathcal{F}}| &\leq E_{z\sim q} |\langle f, \xi_z \rangle_{\mathcal{F}}| \\ &\leq \|f\|_{\mathcal{F}} \underbrace{E_{z\sim q} \|\xi_z\|_{\mathcal{F}}}_{\text{bounded?}}. \end{aligned}$$

# Next step: taking expectations

Riesz theorem: need boundedness,

$$|E_{z\sim q}\langle f, \xi_z\rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}}\,\lambda$$

for some $\lambda \in \mathbb{R}$.

By Jensen and Cauchy-Schwarz,

$$\begin{aligned}|E_{z\sim q}\langle f, \xi_z\rangle_{\mathcal{F}}| &\leq E_{z\sim q}|\langle f, \xi_z\rangle_{\mathcal{F}}| \\ &\leq \|f\|_{\mathcal{F}} \underbrace{E_{z\sim q}\|\xi_z\|_{\mathcal{F}}}_{\text{bounded?}}.\end{aligned}$$

# Next step: taking expectations

Compute the squared norm:

$$\|\xi_z\|_{\mathcal{F}}^2 = \langle \xi_z, \xi_z \rangle_{\mathcal{F}}$$

$$= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot), \dots \right\rangle_{\mathcal{F}}$$

$$= \underbrace{\left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot), \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) \right\rangle_{\mathcal{F}}}_{(A)}$$

$$+ \underbrace{\left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(B) = \frac{d}{dx}\frac{d}{dx'} k(x-x') \big|_{x=x'=z}}$$

$$+ 2 \underbrace{\left\langle \left( \frac{d}{dx} \log p(x) \right) k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(C)}$$

# Next step: taking expectations

Compute the squared norm:

$$\|\xi_z\|_{\mathcal{F}}^2 = \langle \xi_z, \xi_z \rangle_{\mathcal{F}}$$

$$= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot), \dots \right\rangle_{\mathcal{F}}$$

$$= \underbrace{\left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot), \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) \right\rangle_{\mathcal{F}}}_{(A)}$$

$$+ \underbrace{\left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(B) = \frac{d}{dx}\frac{d}{dx'} k(x-x') \big|_{x=x'=z}}$$

$$+ 2 \underbrace{\left\langle \left( \frac{d}{dx} \log p(x) \right) k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(C)}$$

# Next step: taking expectations

Compute the squared norm:

$$
\begin{aligned}
\|\xi_z\|_{\mathcal{F}}^2 &= \langle \xi_z, \xi_z \rangle_{\mathcal{F}} \\
&= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot), \dots \right\rangle_{\mathcal{F}} \\
&= \underbrace{\left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot), \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) \right\rangle_{\mathcal{F}}}_{(A)} \\
&+ \underbrace{\left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(B) = \frac{d}{dx}\frac{d}{dx'} k(x-x') \big|_{x=x'=z}} \\
&+ 2 \underbrace{\left\langle \left( \frac{d}{dx} \log p(x) \right) k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(C)}
\end{aligned}
$$

# Next step: taking expectations

Compute the squared norm:

$$
\begin{aligned}
\|\xi_z\|_{\mathcal{F}}^2 &= \langle \xi_z, \xi_z \rangle_{\mathcal{F}} \\
&= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot), \dots \right\rangle_{\mathcal{F}} \\
&= \underbrace{\left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot), \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) \right\rangle_{\mathcal{F}}}_{(A)} \\
&+ \underbrace{\left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(B) = \frac{d}{dx}\frac{d}{dx'} k(x-x') \big|_{x=x'=z}} \\
&+ 2 \underbrace{\left\langle \left( \frac{d}{dx} \log p(x) \right) k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(C)}
\end{aligned}
$$

# Next step: taking expectations

Compute the squared norm:

$$
\begin{aligned}
\|\xi_z\|_{\mathcal{F}}^2 &= \langle \xi_z, \xi_z \rangle_{\mathcal{F}} \\
&= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) + \frac{d}{dz} k(z,\cdot), \dots \right\rangle_{\mathcal{F}} \\
&= \underbrace{\left\langle \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot), \left( \frac{d}{dz} \log p(z) \right) k(z,\cdot) \right\rangle_{\mathcal{F}}}_{(A)} \\
&+ \underbrace{\left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(B) = \frac{d}{dx}\frac{d}{dx'} k(x-x') \big|_{x=x'=z}} \\
&+ 2 \underbrace{\left\langle \left( \frac{d}{dx} \log p(x) \right) k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}}_{(C)}
\end{aligned}
$$

# First two (easy) terms

First term (A):

$$
\begin{aligned}
(A) &= \left\langle \left( \frac{d}{dz} \log p(z) \right) k(z, \cdot), \left( \frac{d}{dz} \log p(z) \right) k(z, \cdot) \right\rangle_{\mathcal{F}} \\
&= \left[ \left( \frac{d}{dz} \log p(z) \right)^2 \underbrace{k(z, z)}_{=c} \right]
\end{aligned}
$$

# First two (easy) terms

## Second term (B):

$$(B) = \left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} \frac{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x)\right] \overline{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x')\right]}}{\hat{k}_\ell} \Big|_{x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} -(\imath\ell)^2 \hat{k}_\ell \underbrace{\exp\left(\imath\ell(x'-x)\right)}_{=1 \text{ when } x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} \ell^2 \hat{k}_\ell =: C > 0$$

# First two (easy) terms

## Second term (B):

$$(B) = \left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} \frac{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x)\right] \overline{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x')\right]}}{\hat{k}_\ell} \Big|_{x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} -(\imath\ell)^2 \hat{k}_\ell \underbrace{\exp\left(\imath\ell(x'-x)\right)}_{=1 \text{ when } x=x'=z}$$

$$= \sum_{\ell=-\infty}^{\infty} \ell^2 \hat{k}_\ell =: C > 0$$

# First two (easy) terms

## Second term (B):

$$
\begin{aligned}
(B) &= \left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} \frac{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x)\right] \overline{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x')\right]}}{\hat{k}_\ell} \Big|_{x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} -(\imath\ell)^2 \hat{k}_\ell \underbrace{\exp\left(\imath\ell(x'-x)\right)}_{=1 \text{ when } x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} \ell^2 \hat{k}_\ell =: C > 0
\end{aligned}
$$

# First two (easy) terms

## Second term (B):

$$
\begin{aligned}
(B) &= \left\langle \frac{d}{dx} k(x,\cdot), \frac{d}{dx'} k(x',\cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} \frac{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x)\right] \overline{\left[-\imath\ell \hat{k}_\ell \exp(-\imath\ell x')\right]}}{\hat{k}_\ell} \Big|_{x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} -(\imath\ell)^2 \hat{k}_\ell \underbrace{\exp\left(\imath\ell(x'-x)\right)}_{=1 \text{ when } x=x'=z} \\
&= \sum_{\ell=-\infty}^{\infty} \ell^2 \hat{k}_\ell =: C > 0
\end{aligned}
$$

# Third term

## Third term (C):

$$(C) = \left\langle \left( \frac{d}{dx} \log p(x) \right) k(x, \cdot), \frac{d}{dx'} k(x', \cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z}$$

$$= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} \frac{\left[ \hat{k}_\ell \exp(-\imath \ell x) \right] \left[ (-\imath \ell) \hat{k}_\ell \exp(-\imath \ell x') \right]}{\hat{k}_\ell} \Big|_{x=x'=z}$$

$$= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \underbrace{\exp\left( \imath \ell (x' - x) \right)}_{=1 \text{ when } x=x'=z}$$

$$= 0.$$

# Third term

## Third term (C):

$$
\begin{aligned}
(C) &= \left\langle \left( \frac{d}{dx} \log p(x) \right) k(x, \cdot), \frac{d}{dx'} k(x', \cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} \frac{\left[ \hat{k}_\ell \exp(-\imath \ell x) \right] \overline{\left[ (-\imath \ell) \hat{k}_\ell \exp(-\imath \ell x') \right]}}{\hat{k}_\ell} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \underbrace{\exp\left( \imath \ell (x' - x) \right)}_{=1 \text{ when } x=x'=z} \\
&= 0.
\end{aligned}
$$

# Third term

## Third term (C):

$$
\begin{aligned}
(C) &= \left\langle \left( \frac{d}{dx} \log p(x) \right) k(x, \cdot), \frac{d}{dx'} k(x', \cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} \frac{\left[ \hat{k}_\ell \exp(-\imath \ell x) \right] \overline{\left[ (-\imath \ell) \hat{k}_\ell \exp(-\imath \ell x') \right]}}{\hat{k}_\ell} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \underbrace{\exp\left( \imath \ell (x' - x) \right)}_{=1 \text{ when } x=x'=z} \\
&= 0.
\end{aligned}
$$

# Third term

## Third term (C):

$$
\begin{aligned}
(C) &= \left\langle \left( \frac{d}{dx} \log p(x) \right) k(x, \cdot), \frac{d}{dx'} k(x', \cdot) \right\rangle_{\mathcal{F}} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} \frac{\left[ \hat{k}_\ell \exp(-\imath \ell x) \right] \overline{\left[ (-\imath \ell) \hat{k}_\ell \exp(-\imath \ell x') \right]}}{\hat{k}_\ell} \Big|_{x=x'=z} \\
&= \left( \frac{d}{dz} \log p(z) \right) \sum_{\ell=-\infty}^{\infty} (\imath \ell) \hat{k}_\ell \underbrace{\exp\left( \imath \ell (x' - x) \right)}_{=1 \text{ when } x=x'=z} \\
&= 0.
\end{aligned}
$$

# Putting it all together

We found:

$$\|\xi_z\|_{\mathcal{F}}^2 = C + \left(\frac{d}{dz}\log p(z)\right)^2 c,$$

Thus for boundedness, we have the condition:

$$E_{z\sim q}\|\xi_z\|_{\mathcal{F}} = E_{z\sim q}\sqrt{C + \left(\frac{d}{dx}\log p(x)\right)^2 c}$$

$$\leq \sqrt{E_{z\sim q}\left[C + \left(\frac{d}{dz}\log p(z)\right)^2 c\right]},$$

So Riesz holds when $E_{z\sim q}\left(\frac{d}{dz}\log p(z)\right)^2 < \infty$

# Putting it all together

We found:

$$\|\xi_z\|_{\mathcal{F}}^2 = C + \left(\frac{d}{dz}\log p(z)\right)^2 c,$$

Thus for boundedness, we have the condition:

$$\begin{aligned} E_{z\sim q}\,\|\xi_z\|_{\mathcal{F}} &= E_{z\sim q}\sqrt{C + \left(\frac{d}{dx}\log p(x)\right)^2 c} \\ &\leq \sqrt{E_{z\sim q}\left[C + \left(\frac{d}{dz}\log p(z)\right)^2 c\right]}, \end{aligned}$$

So Riesz holds when $E_{z\sim q}\left(\frac{d}{dz}\log p(z)\right)^2 < \infty$

# 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

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

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

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

is undefined.

# Kernel Stein discrepancy: population expression

Multi-dimensional Stein operator:

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

# Kernel Stein discrepancy: population expression

Population kernel Stein discrepancy (in $\mathbb{R}^D$):

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

where

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

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

# Kernel Stein discrepancy: population expression

Population kernel Stein discrepancy (in $\mathbb{R}^D$):

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

where

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

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

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

# Constructing threshold for a statistical test

Given samples $\{z_i\}_{i=1}^n \sim q$, empirical KSD (test statistic) is:

$$\widehat{\mathrm{KSD}}(p, q, \mathcal{F}) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_p(z_i, z_j).$$

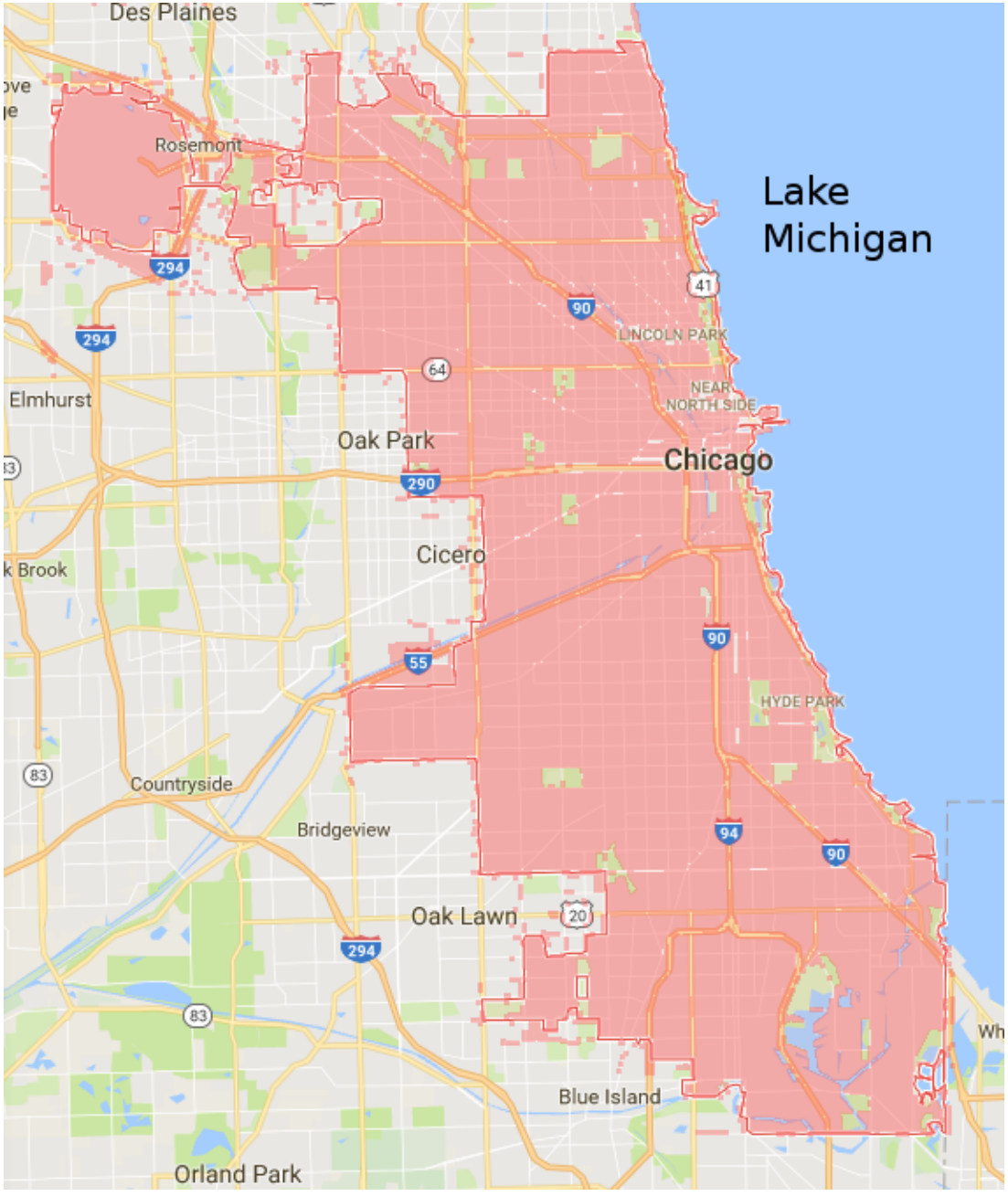
When $q = p$, U-statistic is degenerate. Estimate of null distribution with wild bootstrap:

$$\widetilde{KSD}(p, q, \mathcal{F}) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sigma_i \sigma_j h_p(z_i, z_j).$$

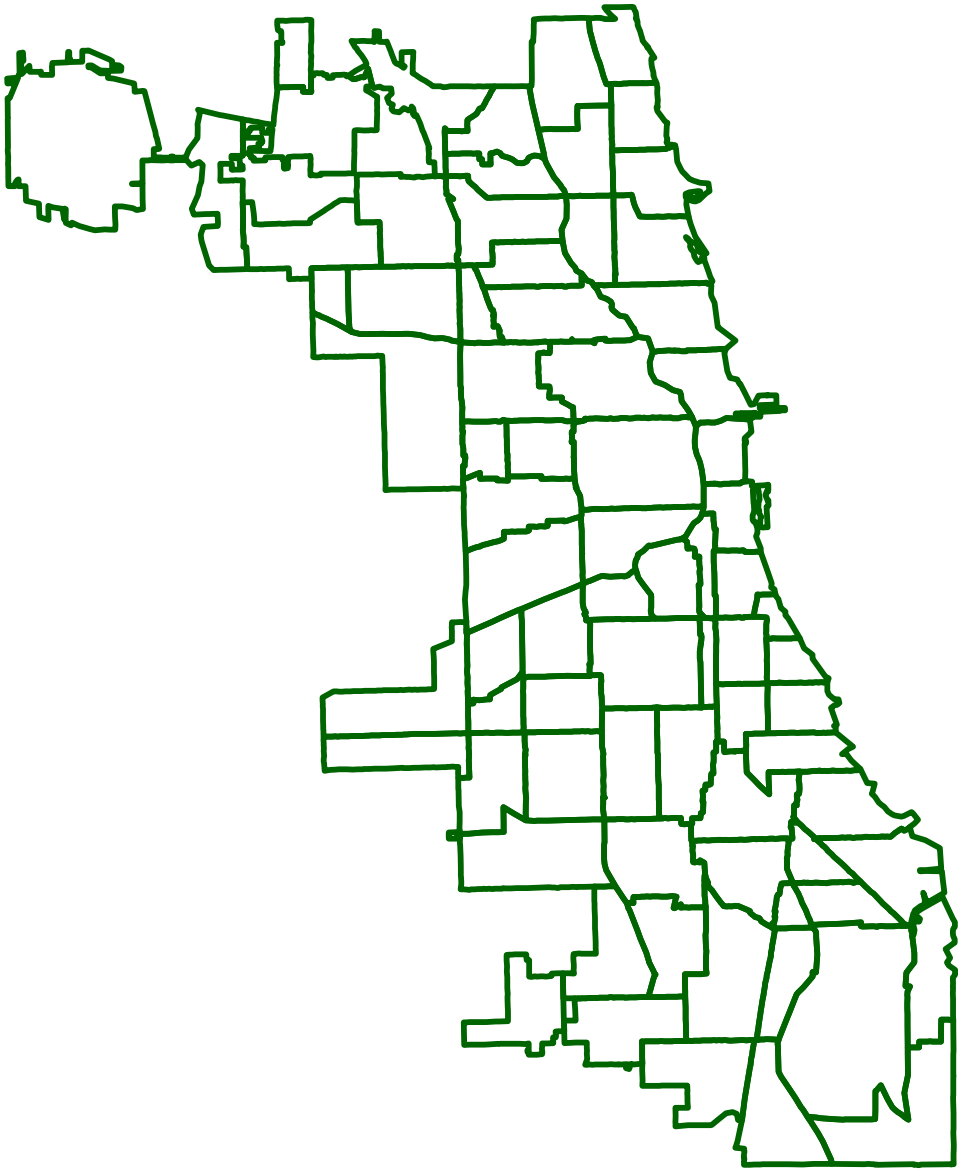
where $\{\sigma_i\}_{i=1}^n$ i.i.d, $E(\sigma_i) = 0$, and $E(\sigma_i^2) = 1$

- Consistent estimate of the null distribtion when $q = p$
- Consistent test (Type II error goes to zero) under a rich class of alternatives Chwialkowski, Strathmann, G., ICML 2016

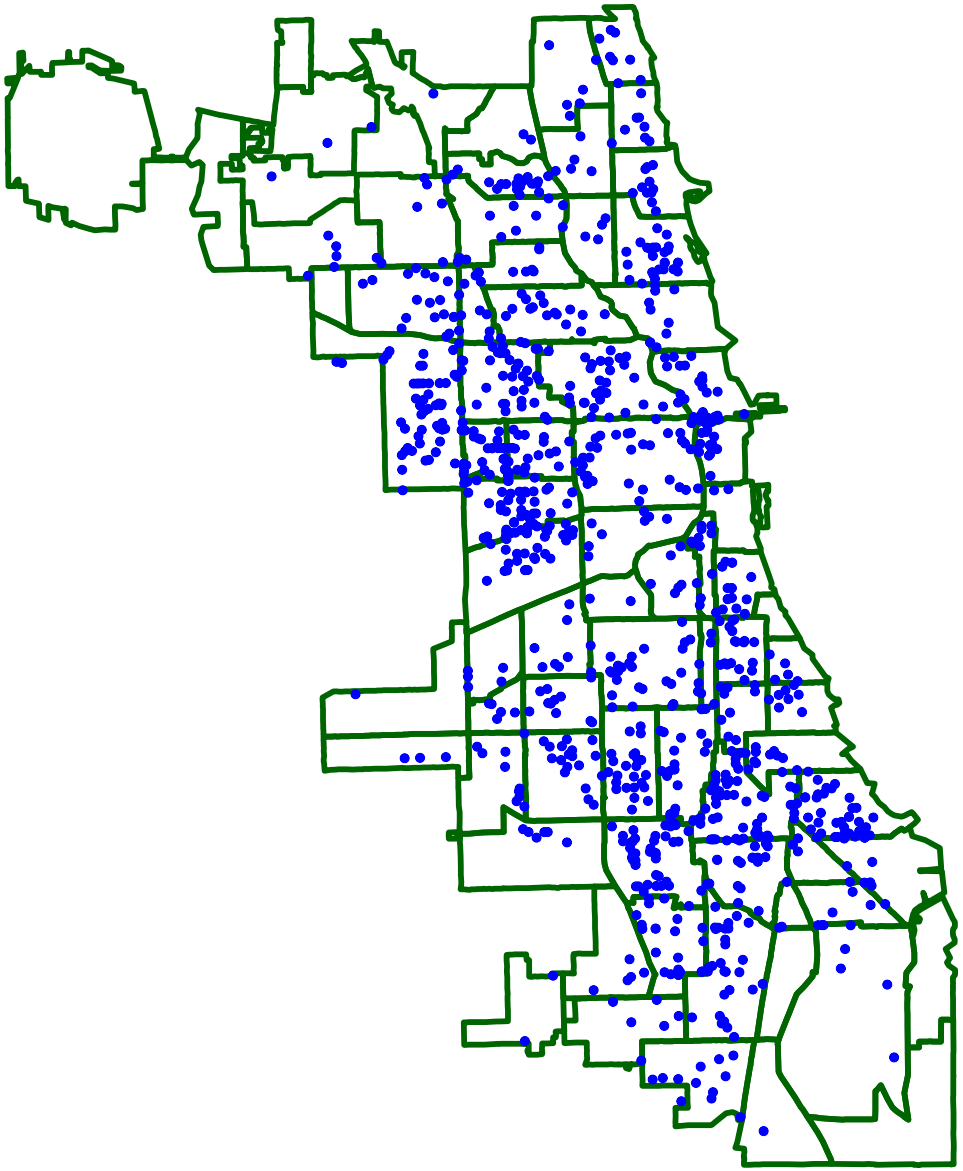
# Model Criticism



# Model Criticism

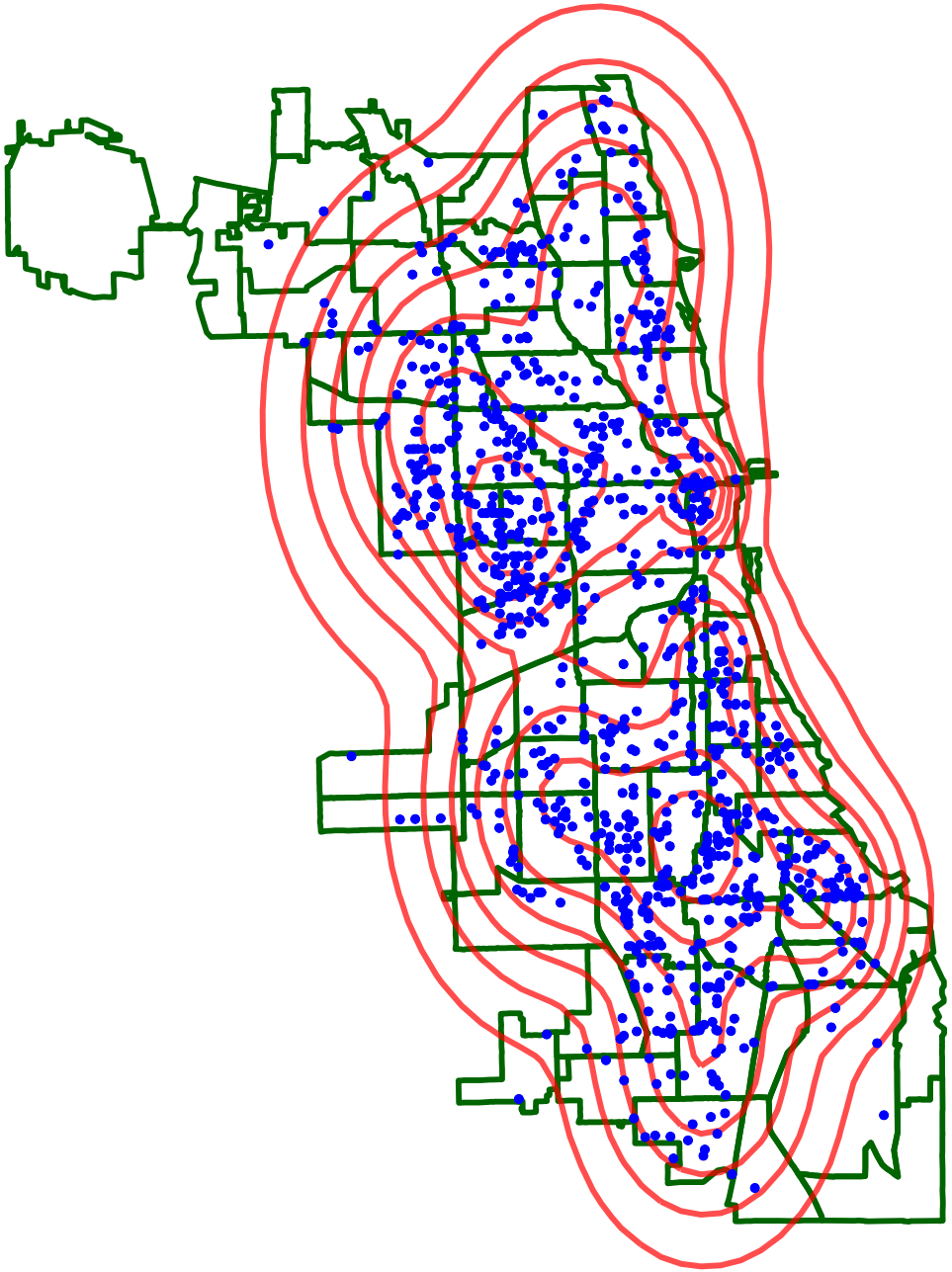


# Model Criticism



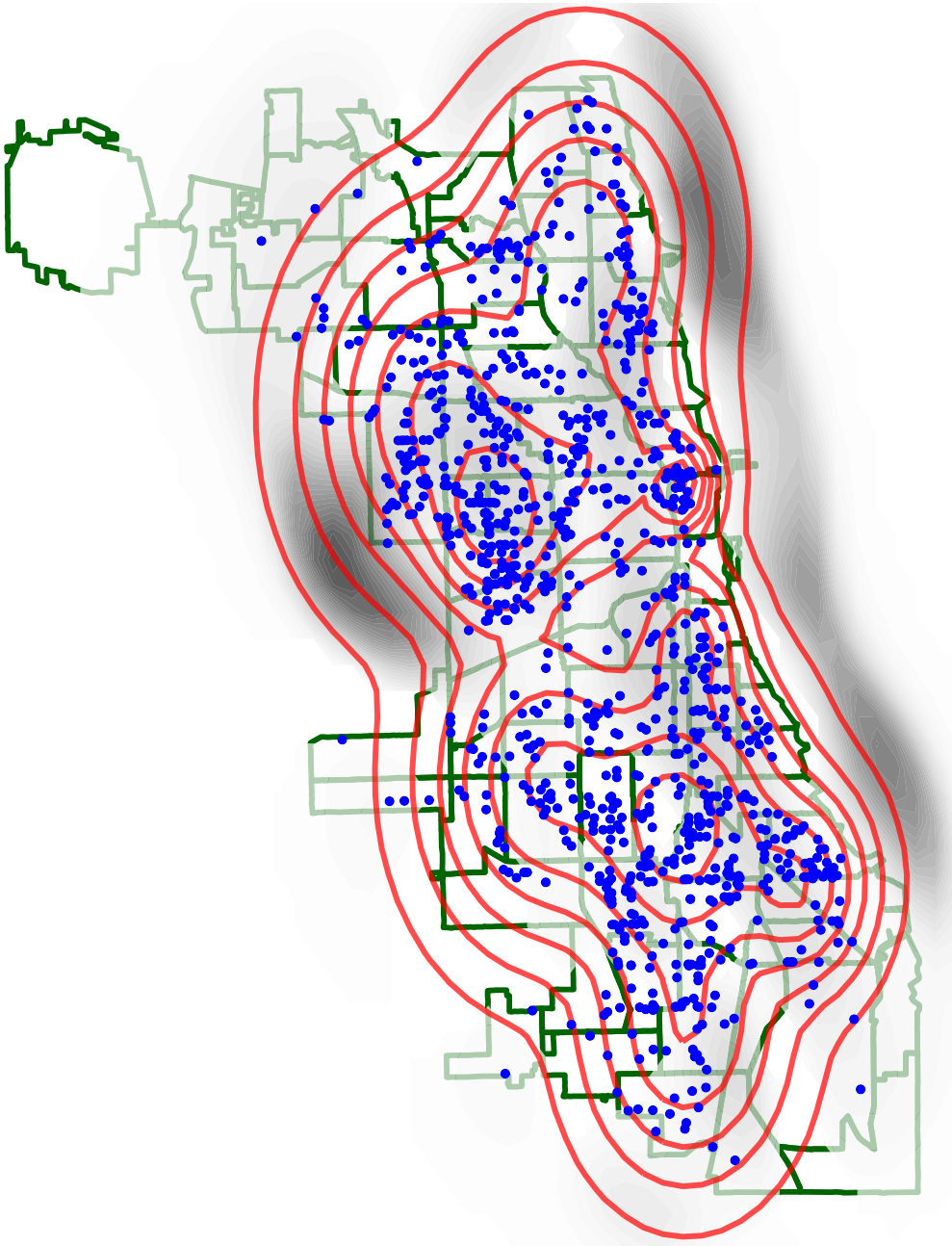
Data = robbery events in Chicago in 2016.

# The witness function: Chicago Crime



Model $p$ = 10-component Gaussian mixture.

# The witness function: Chicago Crime



Witness function $g$ shows mismatch

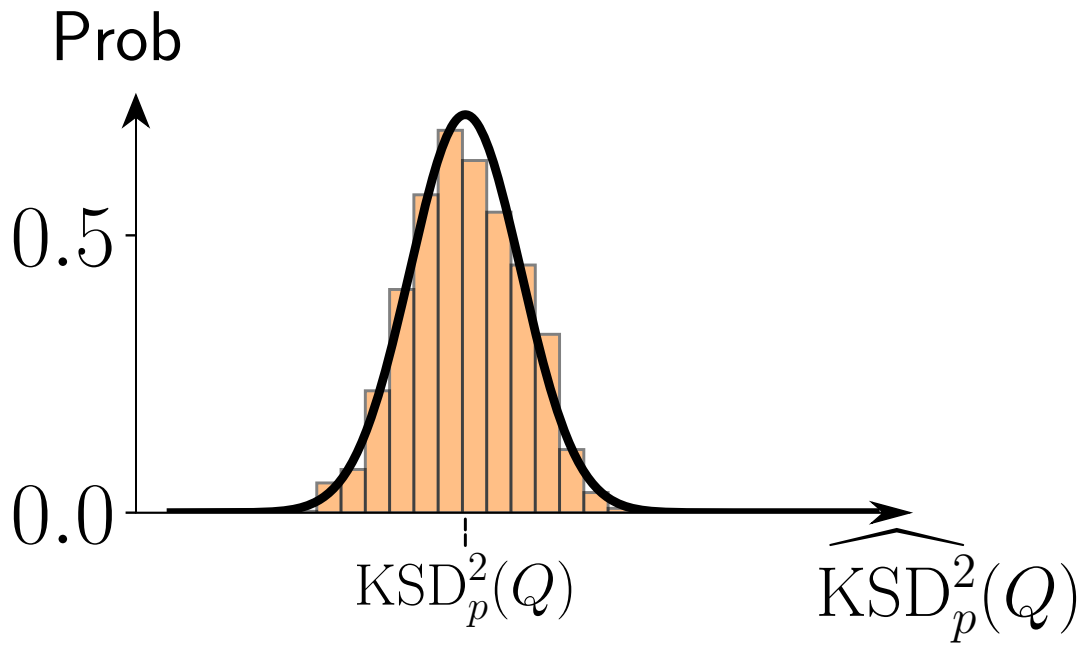
# Empirical statistic, asymptotic normality for $P \neq Q$

The empirical statistic:

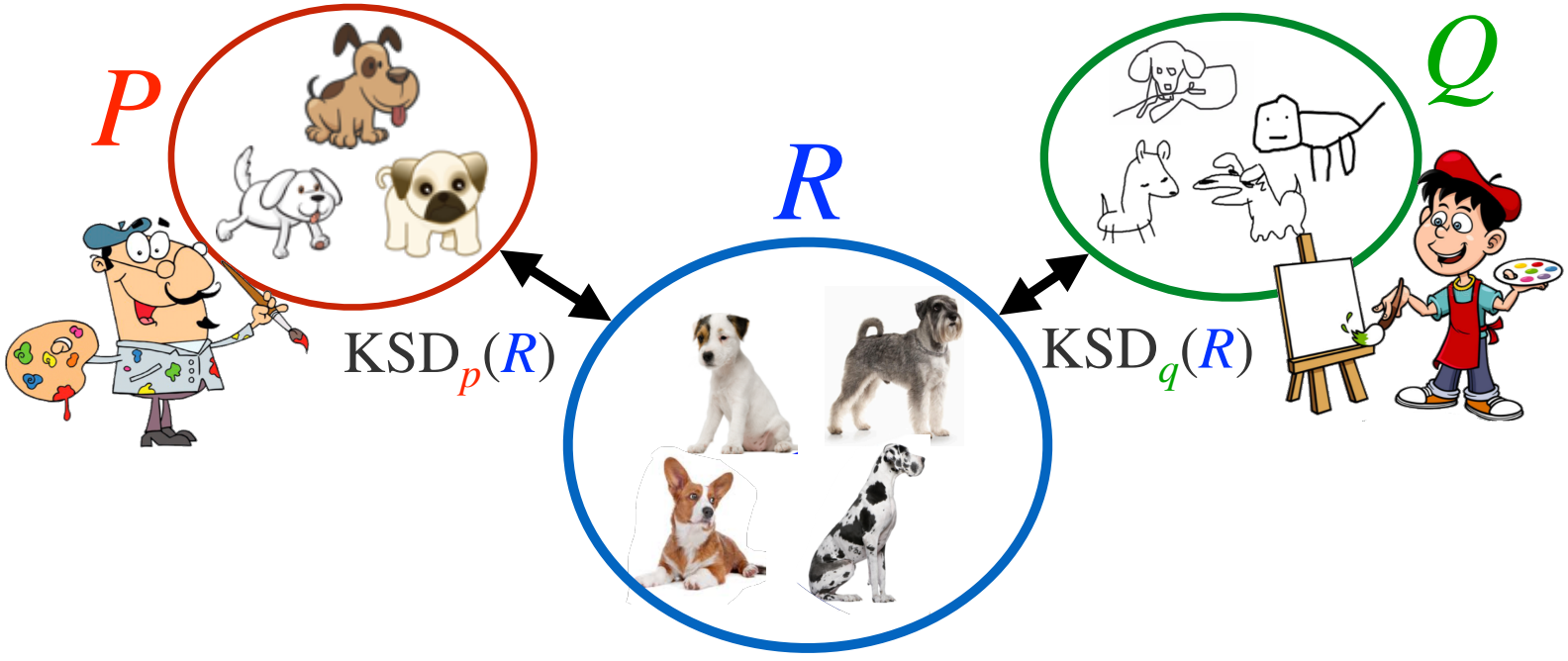
$$\widehat{\mathrm{KSD}}_p^2(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

Asymptotic distribution when $q \neq p$:

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



# Relative goodness-of-fit testing




- Two latent variable models $P$ and $Q$, data $\{x_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} R$.
- Distinct models $p \neq q$

Hypotheses:

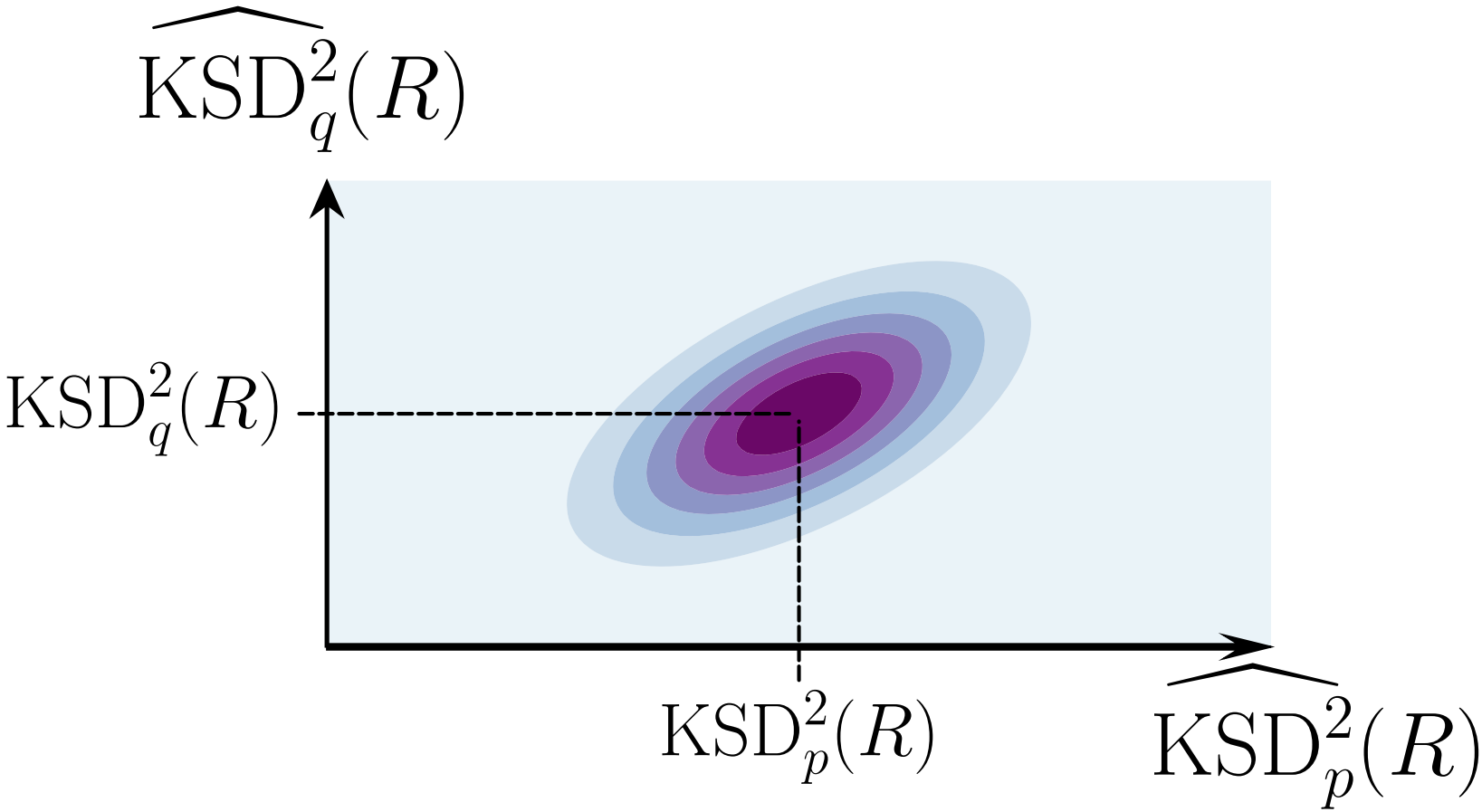
$$H_0 : \mathrm{KSD}_p(R) \leq \mathrm{KSD}_q(R) \text{ vs. } H_1 : \mathrm{KSD}_p(R) > \mathrm{KSD}_q(R)$$

($H_0$ : '$P$ is as good as $Q$, or better' vs. $H_1$ : '$Q$ is better' )

# Relative GOF testing: joint asymptotic normality

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

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



# Relative GOF testing: joint asymptotic normality

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

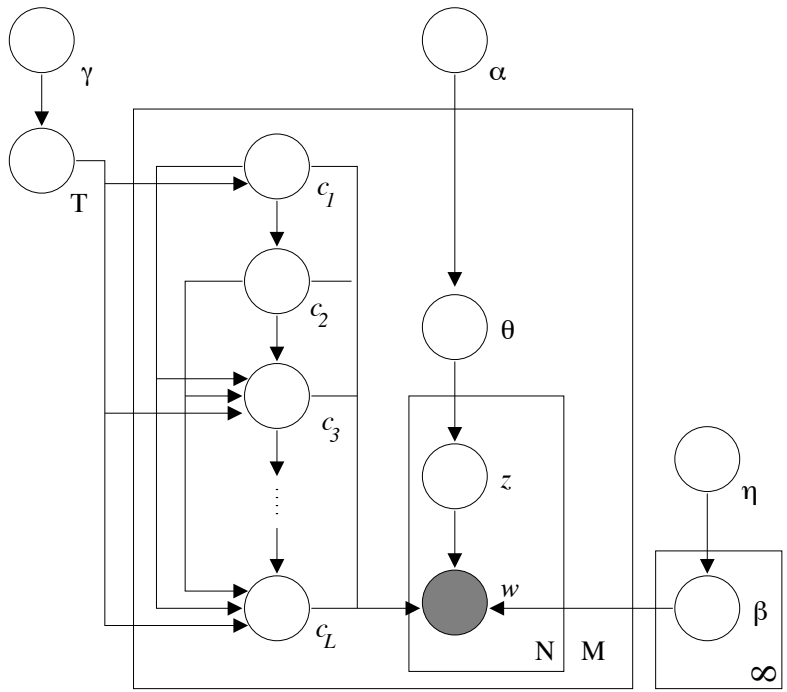
$$\sqrt{n}\begin{bmatrix} \widehat{\mathrm{KSD}_p^2}(R) - \mathrm{KSD}_p(R) \\ \widehat{\mathrm{KSD}_q^2}(R) - \mathrm{KSD}_q(R) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{h_p}^2 & \sigma_{h_p h_q} \\ \sigma_{h_p h_q} & \sigma_{h_q}^2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

Difference in statistics is asymptotically normal:

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

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

# Latent variable models

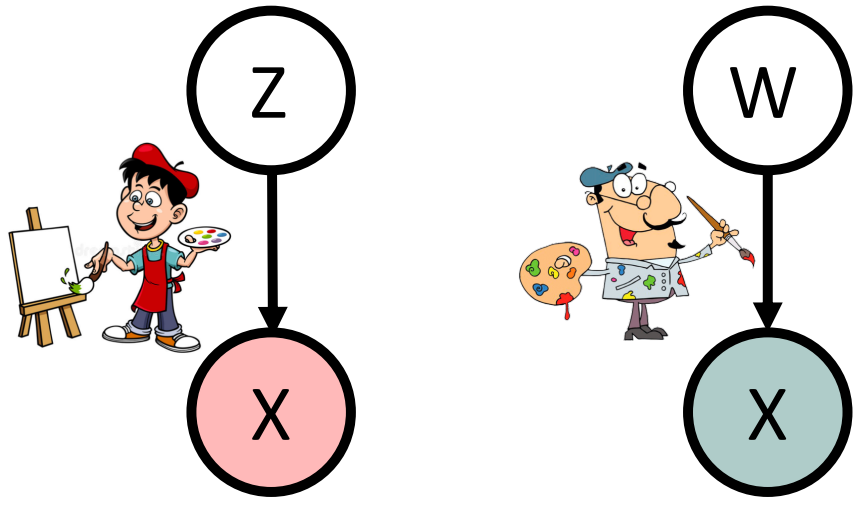


# Latent variable models

Can we compare latent variable models with KSD?

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

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




Multi-dimensional Stein operator:

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

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

# What not to do

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 KSD with approximate density:

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

# What not to do

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 KSD with approximate density:

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

Problem: $\widehat{\mathrm{KSD}^2_{p_m}}(R)$ asymptotically normal but slow bias decay.

# MCMC approximation of score function

Result we use:

$$\mathrm{s}_p(x) = \mathbb{E}_{z|x}[\mathrm{s}_p(x|z)]$$

Proof:

$$\begin{aligned}\mathrm{s}_p(x) = \frac{\nabla p(x)}{p(x)} &= \frac{1}{p(x)} \int \nabla p(x|z) \mathrm{d}p(z) \\ &= \int \frac{\nabla p(x|z)}{p(x|z)} \cdot \frac{p(x|z)\, dp(z)}{p(x)} = \mathbb{E}_{z|x}[\mathrm{s}_p(x|z)],\end{aligned}$$

Friel, N., Mira, A. and Oates, C. J. (2016) Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. Bayesian Analysis, 11, 215–245.

# MCMC approximation of score function

Result we use:

$$\mathrm{s}_p(x) = \mathbb{E}_{z|x}[\mathrm{s}_p(x|z)]$$

Proof:

$$\begin{aligned}\mathrm{s}_p(x) = \frac{\nabla p(x)}{p(x)} &= \frac{1}{p(x)} \int \nabla p(x|z) \mathrm{d}p(z) \\ &= \int \frac{\nabla p(x|z)}{p(x|z)} \cdot \frac{p(x|z)\,dp(z)}{p(x)} = \mathbb{E}_{z|x}[\mathrm{s}_p(x|z)],\end{aligned}$$

Approximate intractable posterior $\mathbb{E}_{z|x_i}[\mathrm{s}_p(x_i|z)]$

$$\bar{\mathrm{s}}_p(x_i; z_i^{(t)}) := \frac{1}{m} \sum_{j=1}^{m} \mathrm{s}_p(x_i|z_{i,j}^{(t)}) \approx \mathrm{s}_p(x_i)$$

with $z_i^{(t)} = (z_{i,1}^{(t)}, \ldots, z_{i,m}^{(t)})$ via MCMC (after $t$ burn-in steps)

Friel, N., Mira, A. and Oates, C. J. (2016) Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods. Bayesian Analysis, 11, 215–245.

# KSD for latent variable models

Recall earlier KSD estimate:

$$U_n(P) = \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j) \ (\approx \mathrm{KSD}_p^2(R))$$

# KSD for latent variable models

Recall earlier KSD estimate:

$$U_n(P) = \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j) \ (\approx \mathrm{KSD}_p^2(R))$$

KSD estimate for latent variable models:

$$U_n^{(t)}(P) := \frac{1}{n(n-1)} \sum_{i \neq j} \bar{H}_p\left[(x_i, z_i^{(t)}), (x_j, z_j^{(t)})\right] \ (\approx \mathrm{KSD}_p^2(R))$$

where $\bar{H}_p$ is the Stein kernel $h_p$ with $\mathrm{s}_p(x_i)$ replaced with $\bar{\mathrm{s}}_p(x_i; z_i^{(t)})$.

# Return to relative GOF test, latent variable models

Hypotheses:

$$H_0 : \mathrm{KSD}_p(R) \leq \mathrm{KSD}_q(R) \text{ vs. } H_1 : \mathrm{KSD}_p(R) > \mathrm{KSD}_q(R)$$

($H_0$ : '$P$ is as good as $Q$, or better' vs. $H_1$ : '$Q$ is better' )

# Return to relative GOF test, latent variable models

Hypotheses:

$$H_0 : \mathrm{KSD}_p(R) \leq \mathrm{KSD}_q(R) \text{ vs. } H_1 : \mathrm{KSD}_p(R) > \mathrm{KSD}_q(R)$$

($H_0$ : '$P$ is as good as $Q$, or better' vs. $H_1$ : '$Q$ is better' )

Strategy:

- Estimate the difference $\mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R)$ by

$$D_n^{(t)}(P, Q) = U_n^{(t)}(P) - U_n^{(t)}(Q).$$

- If $D_n^{(t)}(P, Q)$ is sufficiently large, reject $H_0$.
  - "Sufficient": control type-I error (falsely rejecting $H_0$)
  - Requires the (asymptotic) behaviour of $D_n^{(t)}(P, Q)$

# Asymptotic distribution for relative KSD test

Asymptotic distribution of approximate KSD estimate $n, t \to \infty$:

$$\sqrt{n}\left[D_n^{(t)}(P, Q) - \mu_{PQ}\right] \xrightarrow{d} \mathcal{N}(0, \sigma_{PQ}^2)$$

where

$$\mu_{PQ} = \mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R),$$
$$\sigma_{PQ}^2 = \lim_{n,t\to\infty} n \cdot \mathrm{Var}\left[D_n^{(t)}(P, Q)\right].$$

Fine print:

- The double limit requires fast bias decay
  $\sqrt{n}\left[\mathbb{E}\{D_n^{(t)}(P, Q)\} - \mu_{PQ}\right] \to 0$
- The fourth moment of $\bar{H}_p^{(t)} - \bar{H}_q^{(t)}$ has finite limit sup. $(t \to \infty)$.

# Asymptotic distribution for relative KSD test

Asymptotic distribution of approximate KSD estimate $n, t \to \infty$:

$$\sqrt{n}\left[D_n^{(t)}(P, Q) - \mu_{PQ}\right] \xrightarrow{d} \mathcal{N}(0, \sigma_{PQ}^2)$$

where

$$\mu_{PQ} = \mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R),$$
$$\sigma_{PQ}^2 = \lim_{n,t\to\infty} n \cdot \mathrm{Var}\left[D_n^{(t)}(P, Q)\right].$$

Level-$\alpha$ test:

$$\text{Reject } H_0 \text{ if } D_n^{(t)}(P, Q) \geq \frac{\hat{\sigma}_{PQ}}{\sqrt{n}} c_{1-\alpha}$$

- $c_{1-\alpha}$ is $(1-\alpha)$-quantile of $\mathcal{N}(0, 1)$.
- $\hat{\sigma}_{PQ}$ estimated via jackknife

# Experiments

# Experiment 1: sensitivity to model difference

- Data $R$ : Probabilistic Principal Component Analysis PPCA($A$):

$$x_i \in \mathbb{R}^{100} \sim \mathcal{N}(Az_i, I), \;\; z_i \in \mathbb{R}^{10} \sim \mathcal{N}(0, I_z)$$

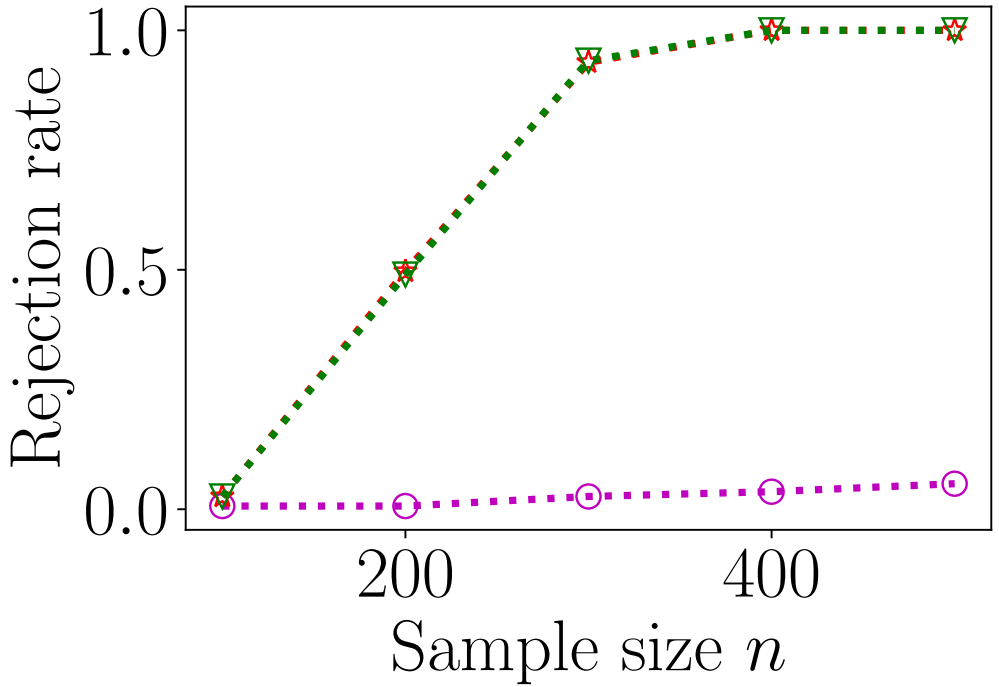
- Generate $P$, $Q$ : perturb $(1,1)$-entry : $A_\delta = A + \delta E_{1,1}$

# Experiment: sensitivity to model difference

- Data $R$ : Probabilistic Principal Component Analysis PPCA($A$):

$$x_i \in \mathbb{R}^{100} \sim \mathcal{N}(Az_i, I), \; z_i \in \mathbb{R}^{10} \sim \mathcal{N}(0, I_z)$$

- Generate $P$, $Q$ : perturb $(1,1)$-entry : $A_\delta = A + \delta E_{1,1}$




- Alt. $H_1$ ($Q$ is better):
  - $P$'s perturbation $\delta_P = 2$
  - $Q$'s perturbation $\delta_Q = 1$
- IMQ kernel: $k(x, x') = \left(1 + \|x - x'\|_2^2 / \sigma_{\text{med}}^2\right)^{-1/2}$
- NUTS-HMC with sample size $m = 500$ (after $t = 200$ steps).

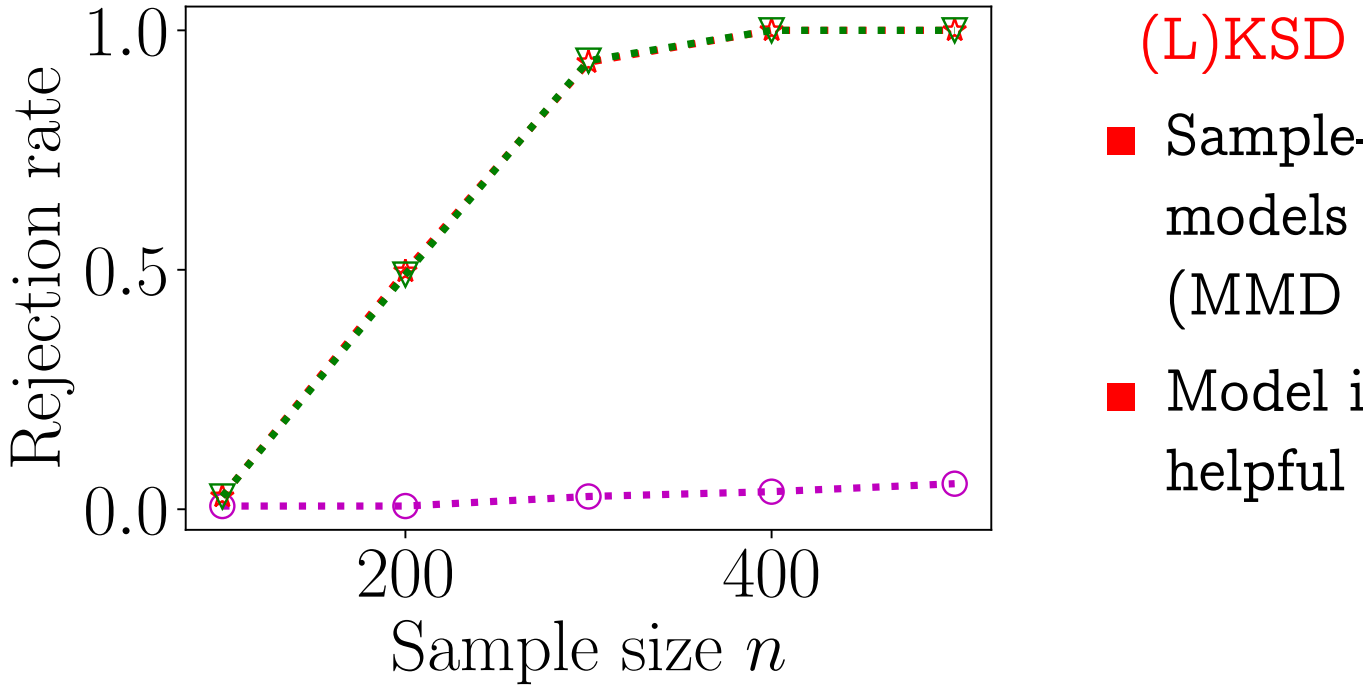
Hoffman and Gelman (JMLR 2014)

# Experiment: sensitivity to model difference

- Data $R$ : Probabilistic Principal Component Analysis PPCA($A$):

$$x_i \in \mathbb{R}^{100} \sim \mathcal{N}(Az_i, I), \; z_i \in \mathbb{R}^{10} \sim \mathcal{N}(0, I_z)$$

- Generate $P$, $Q$ : perturb $(1,1)$-entry : $A_\delta = A + \delta E_{1,1}$



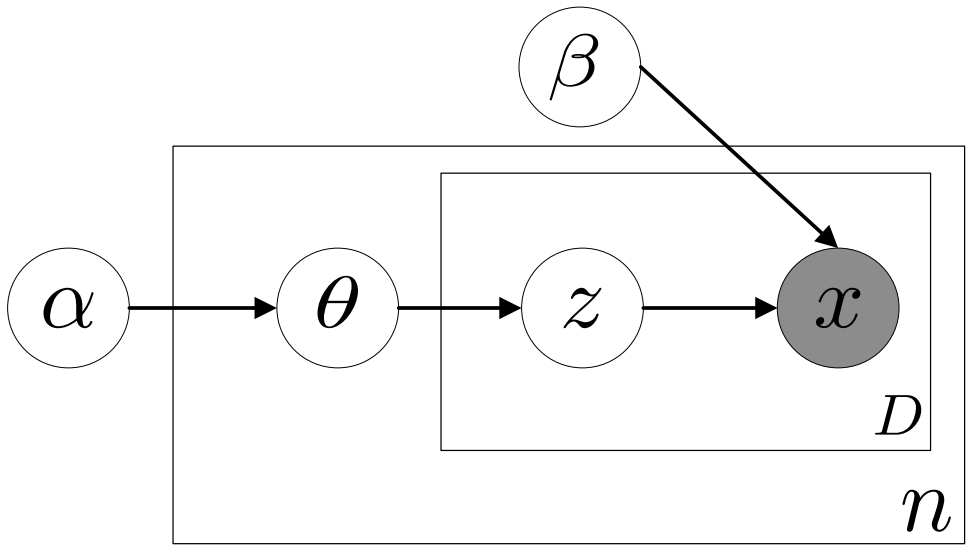

(L)KSD = higher power

- Sample-wise difference in models = subtle (MMD fails)
- Model information is helpful

Hoffman and Gelman (JMLR 2014)

# Experiment 2: topic models for arXiv articles

- Data $R$ : arXiv articles from category stat.TH (stat theory) :
- Models $P$, $Q$ : LDAs trained on articles from different categories
  - $P$ : math.PR (math probability theory)
  - $Q$ : stat.ME (stat methodology). $H_1$: $Q$ is better

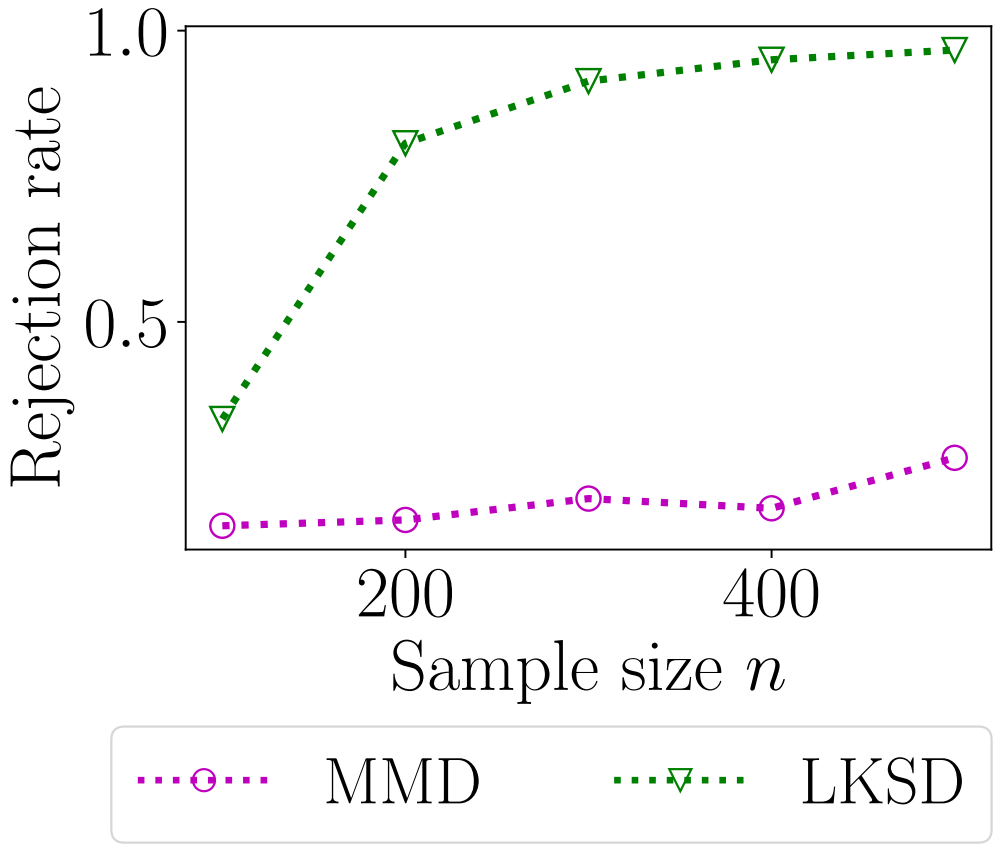



Graphical model of LDA

Blei, Ng, Jordan (JMLR 2003)

# Experiment 2: topic models for arXiv articles

- Data $R$ : arXiv articles from category stat.TH (stat theory) :
- Models $P$, $Q$ : LDAs trained on articles from different categories (100 topics)
  - $P$ : math.PR (math probability theory)
  - $Q$ : stat.ME (stat methodology). $H_1$: $Q$ is better




- $\mathcal{X} = \{1, \ldots, L\}^D$, $D = 100$, $L = 126,190$.
- IMQ kernel in BoW rep.: $k(x, x') = (1 + \|B(x) - B(x')\|_2^2)^{-1/2}$
- MCMC size $m = 5000$ (after $t = 500$ steps).

# Conclusion

## Relative goodness-of-fit tests for Models with Latent Variables

- The kernel Stein discrepancy
  - Comparing two models via samples: MMD and the witness function.
  - Comparing a sample and a model: Stein modification of the witness class
- Constructing a relative hypothesis test using the KSD
- Relative hypothesis tests with latent variables

# References

A Kernel Test of Goodness of Fit
Kacper Chwialkowski, Heiko Strathmann, Arthur Gretton
https://arxiv.org/abs/1602.02964

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