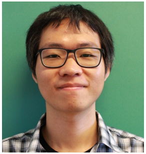
Relative Goodness-of-Fit Tests for Models with Latent Variables

Arthur Gretton

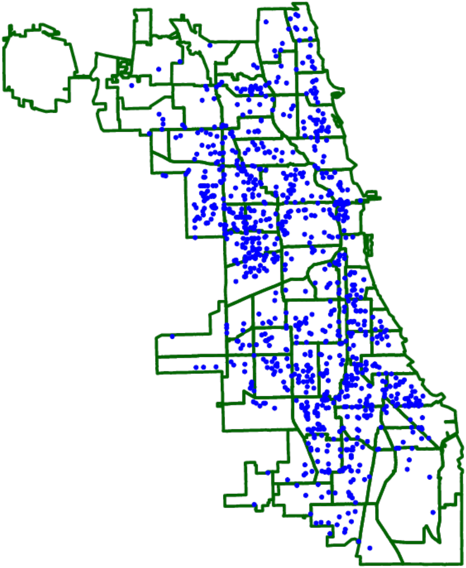





Gatsby Computational Neuroscience Unit,
Deepmind

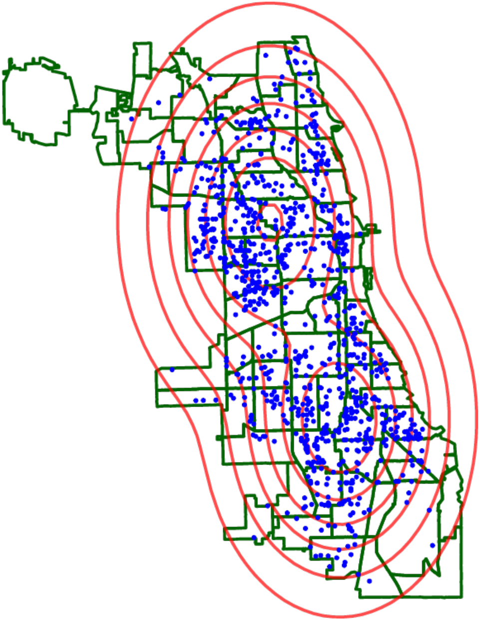
Department of Statistics, Columbia, 2023

Model Criticism



Data = robbery events in Chicago in 2016.

Model Criticism



Is this a good model?

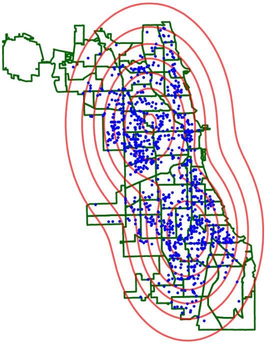
Model Criticism

"All models are wrong."

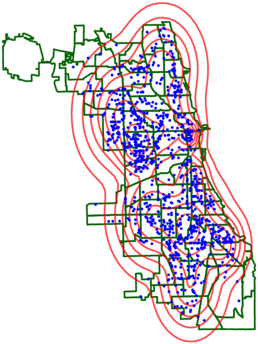
G. Box (1976)

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?



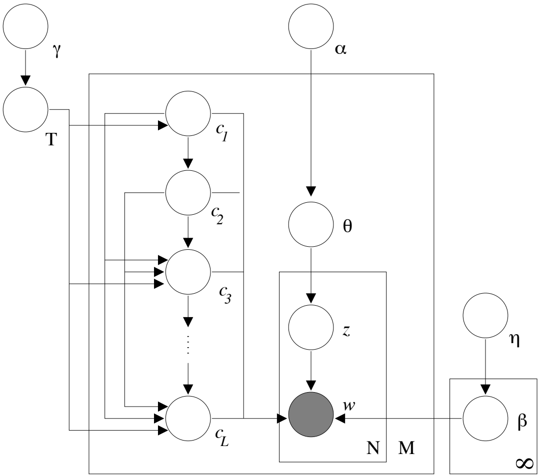
P : two components



Q : ten components

Most interesting models have latent structure

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



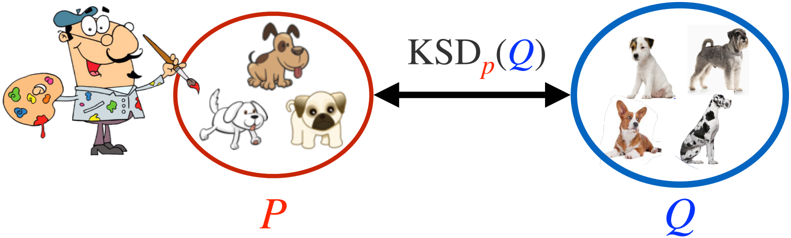
Outline

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" ($P \neq Q$).

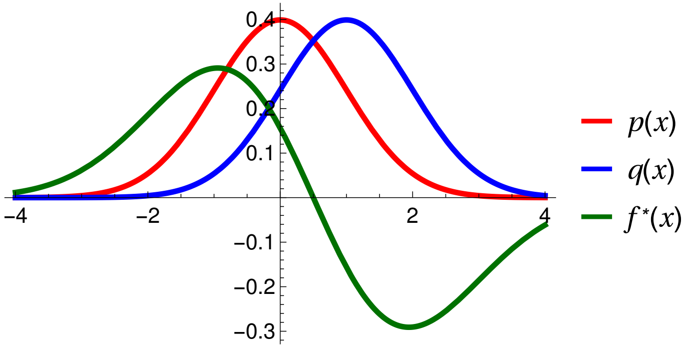


Comparing a sample and model

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

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

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.

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

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

Stein idea

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

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

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.

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)

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

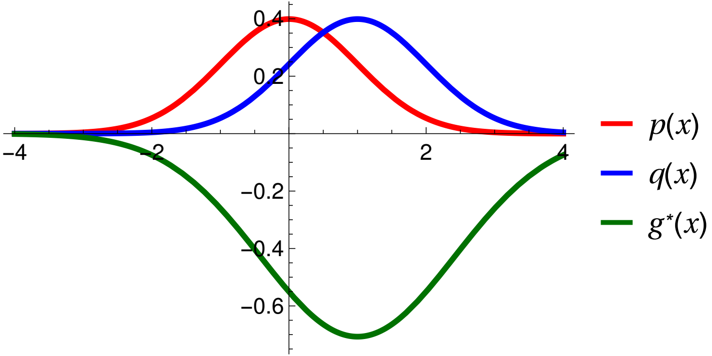
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



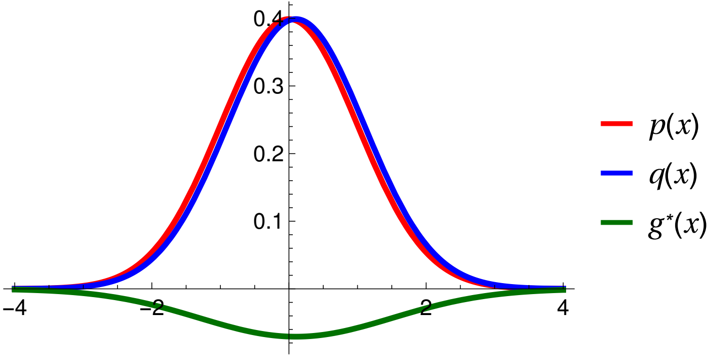
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



Computing the kernel Stein discrepancy

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

Can we define "Stein features"?

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

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"?

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

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

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

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

Kernel Stein discrepancy: derivation

Closed-form expression for KSD: given independent $x, x' \sim Q$, then

$$\mathrm{KSD}_p(Q) = \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathrm{E}_{x\sim q}\left([\mathcal{A}_p g](x)\right)$$

$$= \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathrm{E}_{x\sim q} \langle g, \xi_x \rangle_{\mathcal{F}}$$

$$\underset{(a)}{=} \sup_{\|g\|_{\mathcal{F}} \leq 1} \langle g, \mathrm{E}_{x\sim q}\xi_x \rangle_{\mathcal{F}} = \|\mathrm{E}_{x\sim q}\xi_x\|_{\mathcal{F}}$$

Kernel Stein discrepancy: derivation

Closed-form expression for KSD: given independent $x, x' \sim Q$, then

$$\mathrm{KSD}_p(Q) = \sup_{\|g\|_{\mathcal{F}} \le 1} \mathrm{E}_{x \sim q}\left([\mathcal{A}_p g](x)\right)$$

$$= \sup_{\|g\|_{\mathcal{F}} \le 1} \mathrm{E}_{x \sim q} \langle g, \xi_x \rangle_{\mathcal{F}}$$

$$\overset{(a)}{=} \sup_{\|g\|_{\mathcal{F}} \le 1} \langle g, \mathrm{E}_{x \sim q} \xi_x \rangle_{\mathcal{F}} = \|\mathrm{E}_{x \sim q} \xi_x\|_{\mathcal{F}}$$

Kernel Stein discrepancy: derivation

Closed-form expression for KSD: given independent $x, x' \sim Q$, then

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

Caution: (a) requires boundedness (Riesz),

$$\left|E_{z\sim q}\left\langle f, \xi_z\right\rangle_{\mathcal{F}}\right| \le \|f\|_{\mathcal{F}} \underbrace{E_{z\sim q}\|\xi_z\|_{\mathcal{F}}}_{\text{bounded?}}$$

Kernel Stein discrepancy: derivation

Closed-form expression for KSD: given independent $x, x' \sim Q$, then

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

Caution: (a) requires boundedness (Riesz),

$$\left| E_{z\sim q} \langle f, \xi_z \rangle_{\mathcal{F}} \right| \leq \|f\|_{\mathcal{F}} \underbrace{E_{z\sim q} \|\xi_z\|_{\mathcal{F}}}_{\text{bounded?}}$$

Leading term

$$\|\xi_z\|^2_{\mathcal{F}} = \left\langle \left(\frac{d}{dz} \log p(z)\right) k(z, \cdot), \left(\frac{d}{dz} \log p(z)\right) k(z, \cdot) \right\rangle_{\mathcal{F}} + \dots$$

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

Kernel Stein discrepancy: derivation

Closed-form expression for KSD: given independent $x, x' \sim Q$, then

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

Kernel expression in $\mathbb{R}$:

$$\begin{aligned}
&\| \mathrm{E}_{x \sim q} \xi_x \|_{\mathcal{F}}^2 \\
&= \left\| \mathrm{E}_{x \sim q} \left(\varphi(x) \frac{d}{dx} \log p(x) + \frac{d}{dx} \varphi(x) \right) \right\|_{\mathcal{F}}^2 \\
&= \mathrm{E}_{x, x' \sim Q} \left(k(x, x') \frac{\partial p(x)}{p(x)} \frac{\partial p(x')}{p(x')} + \partial_1 k(x, x') \frac{\partial p(x')}{p(x')} \right. \\
&\qquad \left. + \partial_2 k(x, x') \frac{\partial p(x)}{p(x)} + \partial_{12} k(x, x') \right)
\end{aligned}$$

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

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

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

where

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

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

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

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

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

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 $P = Q$.

KSD for discrete-valued variables

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

The population KSD (discrete):

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

where

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

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

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):

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

where

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

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

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):

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

where

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

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

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)

Constructing threshold for a statistical test

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

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

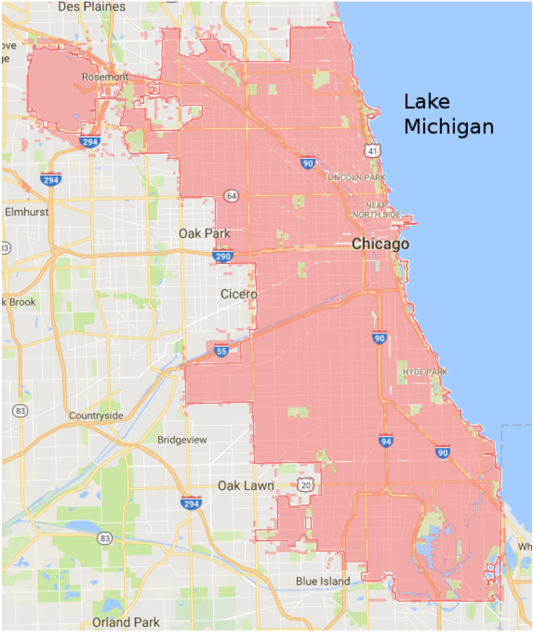
When $q = p$, U-statistic is degenerate. Estimate of null distribution with wild bootstrap:

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

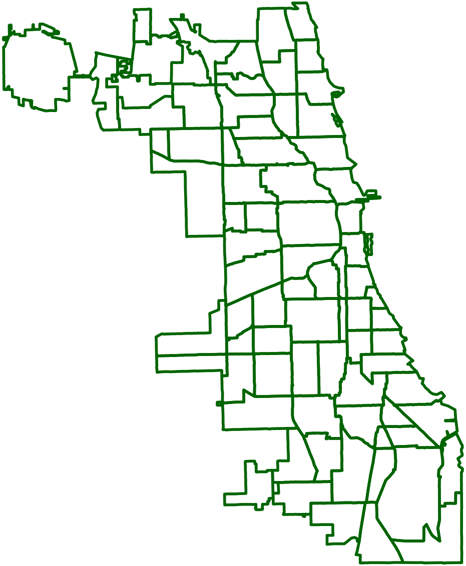
where $\{\sigma_i\}_{i=1}^n$ i.i.d, $E(\sigma_i) = 0$, and $E(\sigma_i^2) = 1$

- 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

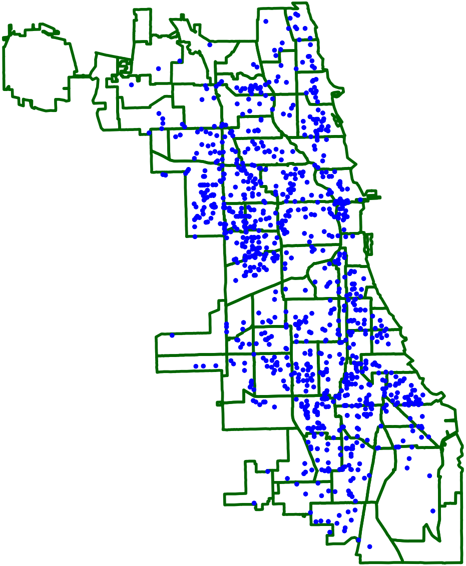
Model Criticism



Model Criticism

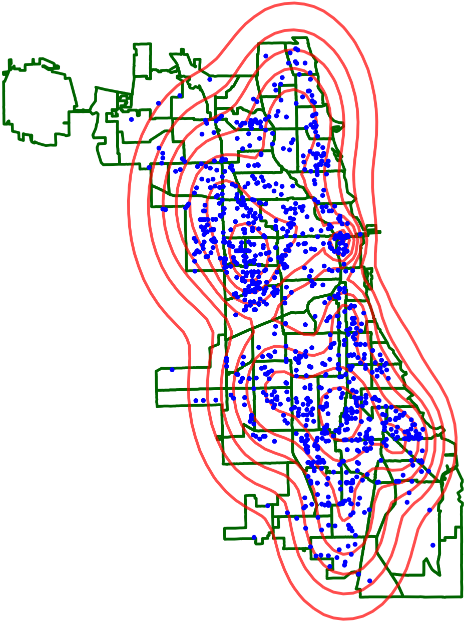


Model Criticism



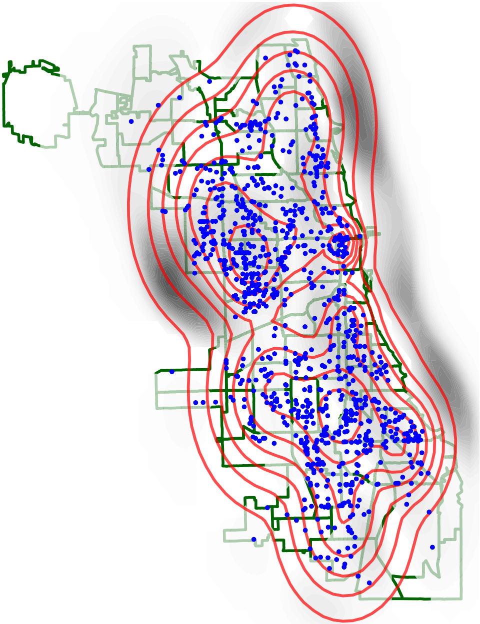
Data = robbery events in Chicago in 2016.

The witness function: Chicago Crime



Model p = 10-component Gaussian mixture.

The witness function: Chicago Crime



Witness function g shows mismatch

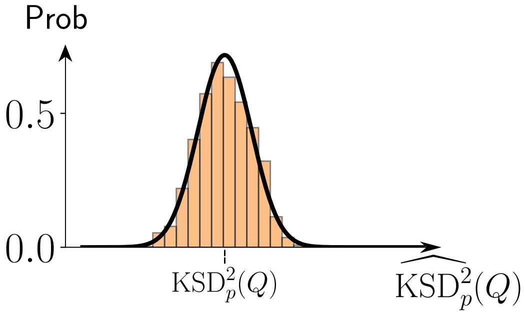
Empirical statistic, asymptotic normality for $P \neq Q$

The empirical statistic:

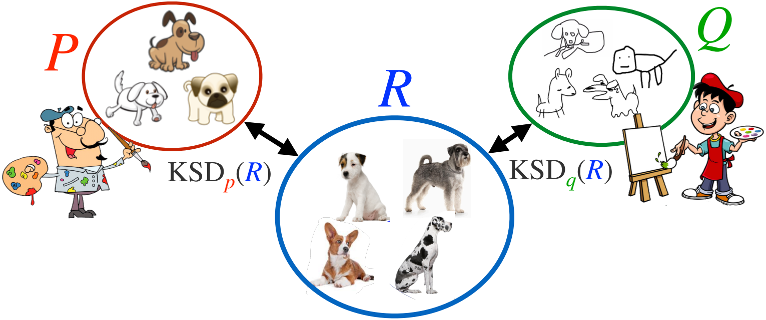
$$\widehat{\mathrm{KSD}^2_p}(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

Asymptotic distribution when $q \neq p$:

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



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:

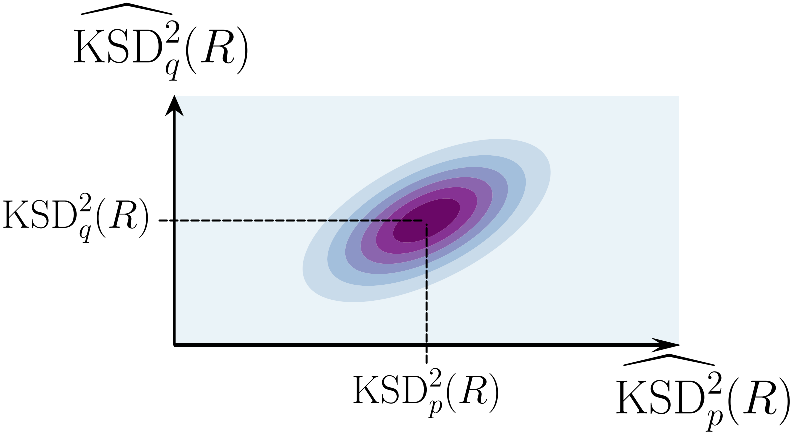
$$H_0 : \mathrm{KSD}_p(R) \leq \mathrm{KSD}_q(R) \text{ vs. } H_1 : \mathrm{KSD}_p(R) > \mathrm{KSD}_q(R)$$

$$(H_0 : \text{`}P \text{ is as good as } Q\text{, or better' vs. } H_1 : \text{`}Q \text{ is better'})$$

Relative GOF testing: joint asymptotic normality

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

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



Relative GOF testing: joint asymptotic normality

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

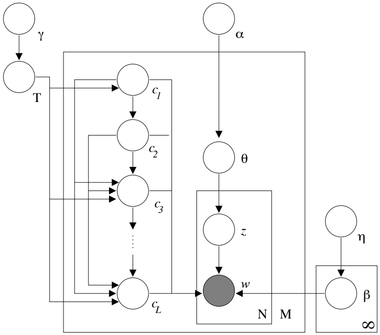
$$\sqrt{n} \begin{bmatrix} \widehat{\mathrm{KSD}^2_p}(R) - \mathrm{KSD}_p(R) \\ \widehat{\mathrm{KSD}^2_q}(R) - \mathrm{KSD}_q(R) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{h_p} & \sigma_{h_p h_q} \\ \sigma_{h_p h_q} & \sigma^2_{h_q}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

Difference in statistics is asymptotically normal:

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

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

Latent variable models

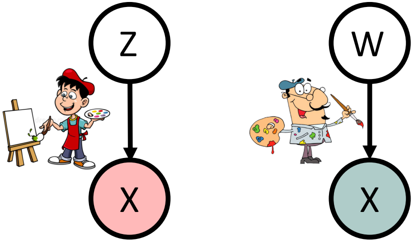


Latent variable models

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:

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

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:

$$s_p(x) = \mathbb{E}_{z|x}[s_p(x|z)]$$

Proof:

$$s_p(x) = \frac{\nabla p(x)}{p(x)} = \frac{1}{p(x)} \int \nabla p(x|z) \mathrm{d}p(z)$$

$$= \int \frac{\nabla p(x|z)}{p(x|z)} \cdot \frac{p(x|z)\, dp(z)}{p(x)} = \mathbb{E}_{z|x}[s_p(x|z)],$$

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:

$$\mathrm{s}_p(x) = \frac{\nabla p(x)}{p(x)} = \frac{1}{p(x)} \int \nabla p(x|z) \mathrm{d}p(z)$$

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

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:

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

$$(H_0 : \text{`}P \text{ is as good as } Q, \text{ or better' vs. } H_1 : \text{`}Q \text{ is better'})$$

Return to relative GOF test, latent variable models

Hypotheses:

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

$$(H_0 : \text{`}P \text{ is as good as } Q, \text{ or better' vs. } H_1 : \text{`}Q \text{ is better' })$$

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 $n, t \to \infty$:

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

where

$$\mu_{PQ} = \mathrm{KSD}_p^2(R) - \mathrm{KSD}_q^2(R),$$

$$\sigma^2_{PQ} = \lim_{n,t\to\infty} n \cdot \mathrm{Var}\left[D_n^{(t)}(P, Q)\right].$$

Fine print:

- The double limit requires fast bias decay $\sqrt{n}[\mathbb{E}\{D_n^{(t)}(P, Q)\} - \mu_{PQ}] \to 0$
- 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 $n, t \to \infty$:

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

where

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

Level-α 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):

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

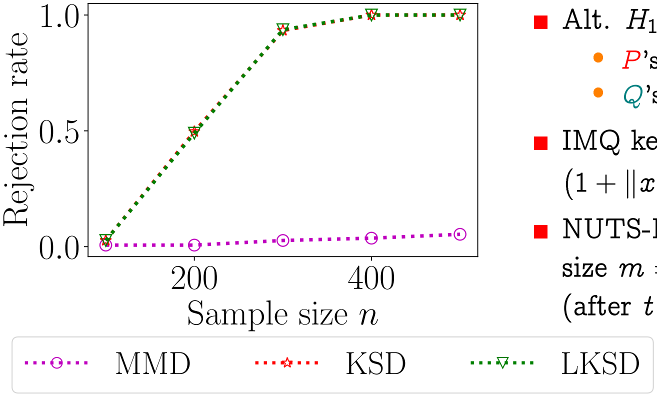
- Generate P, Q : perturb $(1,1)$-entry : $A_\delta = A + \delta E_{1,1}$

Experiment 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 : $A_\delta = A + \delta\, E_{1,1}$



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

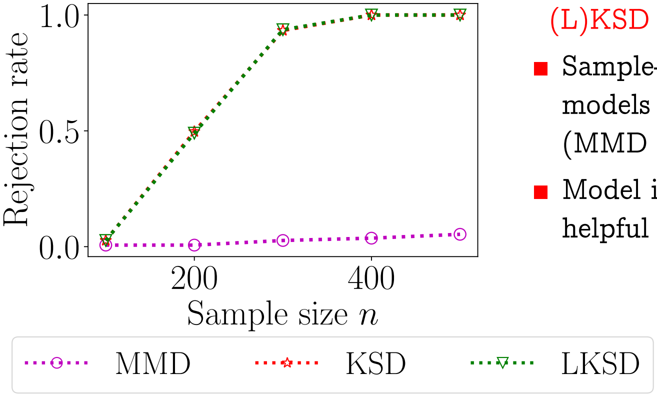
Hoffman and Gelman (JMLR 2014)

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 : $A_\delta = A + \delta E_{1,1}$



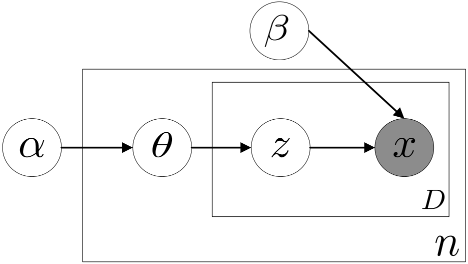
(L)KSD = higher power

- 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

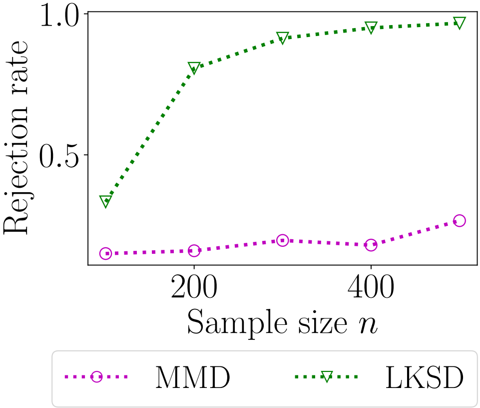


Graphical model of LDA

Blei, Ng, Jordan (JMLR 2003)

Experiment 2: topic models for arXiv articles

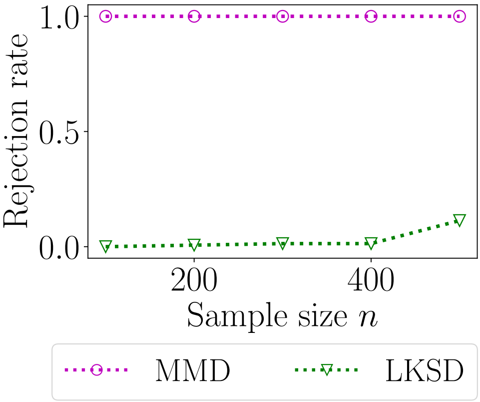
- Data R : arXiv articles from category stat.TH (stat theory) :
- Models P, Q : LDAs trained on articles from different categories (100 topics)
  - P : math.PR (math probability theory)
  - Q : stat.ME (stat methodology). H_1: Q is better



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

A failure mode

- Data R : arXiv articles from category stat.TH (stat theory) :
- Models P, Q : LDAs trained on articles from different categories (100 topics)
  - P : cs.LG (CS machine learning)
  - Q : stat.ME (stat methodology). H_1: Q is better



- $\mathcal{X} = \{1, \ldots, L\}^D$, $D = 100$, $L = 208,671$.
- IMQ kernel in BoW rep.: $k(x, x') = \left(1 + \|B(x) - B(x')\|_2^2\right)^{-1/2}$
- MCMC size $m = 5000$ (after $t = 500$ steps).

What went wrong?

Recall (one-dimension, informally)

$$\mathrm{s}_p(x) = \frac{p(x+1)}{p(x)} - 1$$

Numerical instability arises when

- Observed word x has low probability
- Word next to x in vocabulary has non-negligible probability

Zanella-Barker Stein operator

Zanella-Barker Stein operator (1-D):

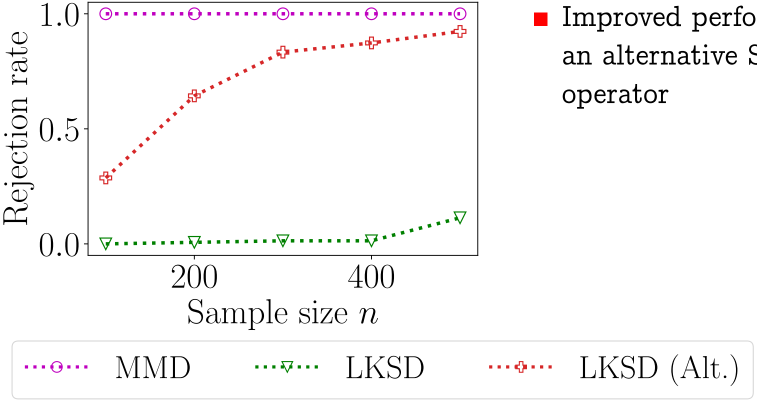
$$\mathcal{A}_p^{\mathrm{ZB}} f(x) = \sum_{\tilde{x} \in \{x+1, x-1\}} \frac{p(\tilde{x})}{p(\tilde{x}) + p(x)} \cdot \{f(\tilde{x}) - f(x)\}$$

- More stable: the ratio $p(\tilde{x})/\{p(\tilde{x}) + p(x)\}$ is always between 0 and 1.
- Similarly applies to latent variable models.

Hodgkinson, Salomone, and Roosta (2020); Shi, Zhou, Hwang, Titsias, and Mackey. (2022)

A resolution to the failure mode

- Data R : arXiv articles from category stat.TH (stat theory) :
- Models P, Q : LDAs trained on articles from different categories (100 topics)
  - P : cs.LG (CS machine learning)
  - Q : stat.ME (stat methodology). H_1: Q is better



- Improved performance by an alternative Stein operator

Can sampler influence test power?

How important is the quality of $\frac{1}{m}\sum_{j=1}^{m} \mathrm{s}_{p}(x|z_j^{(t)})$?

Experiment with PPCA:

- P : MALA with a bad step size (poor sampler)
- Q : NUTS-HMC (good sampler)

Expectation:

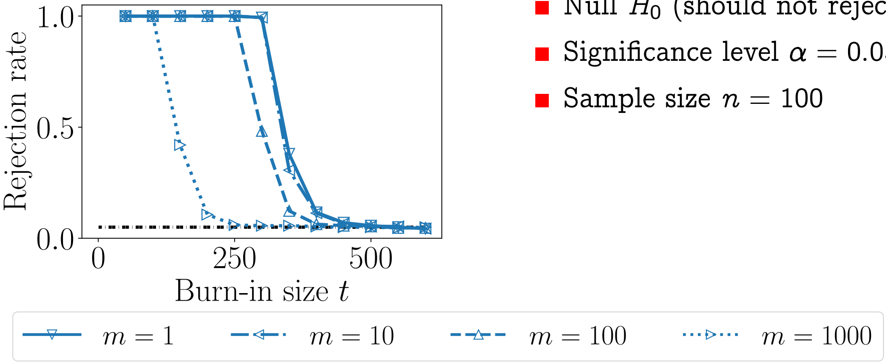
If poor, the test would reject even if P and Q are equally good

Can sampler influence test power?

How important is the quality of $\frac{1}{m}\sum_{j=1}^{m} \mathrm{s}_p(x|z_j^{(t)})$?

Experiment with PPCA:

- P : MALA with a bad step size (poor sampler)
- Q : NUTS-HMC (good sampler)



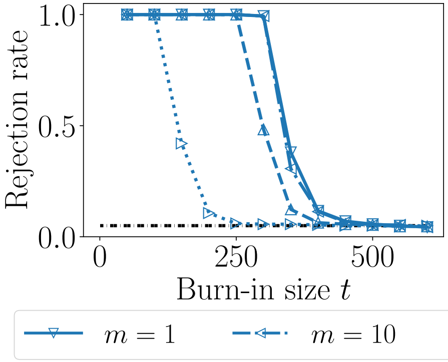
- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 100$

Can sampler influence test power?

How important is the quality of $\frac{1}{m}\sum_{j=1}^{m} \mathrm{s}_{p}(x|z_j^{(t)})$?

Experiment with PPCA:

- P : MALA with a bad step size (poor sampler)
- Q : NUTS-HMC (good sampler)



- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 100$

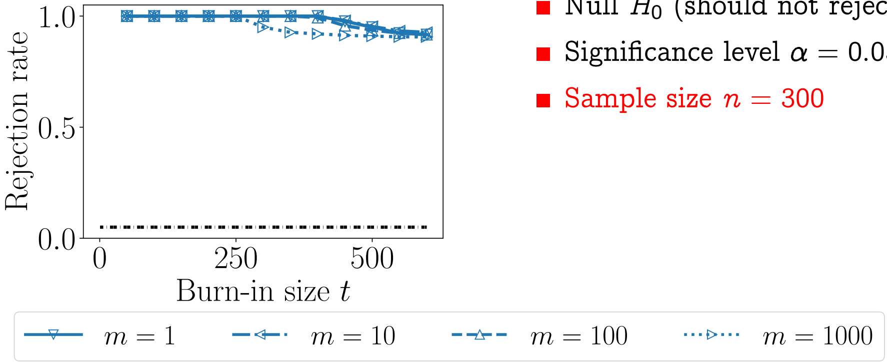
Sufficient burn-in
$\rightarrow$ correct type-I error

Can sampler influence test power?

How important is the quality of $\frac{1}{m}\sum_{j=1}^{m} \mathrm{s}_p(x|z_j^{(t)})$?

Experiment with PPCA:

- P : MALA with a bad step size (poor sampler)
- Q : NUTS-HMC (good sampler)



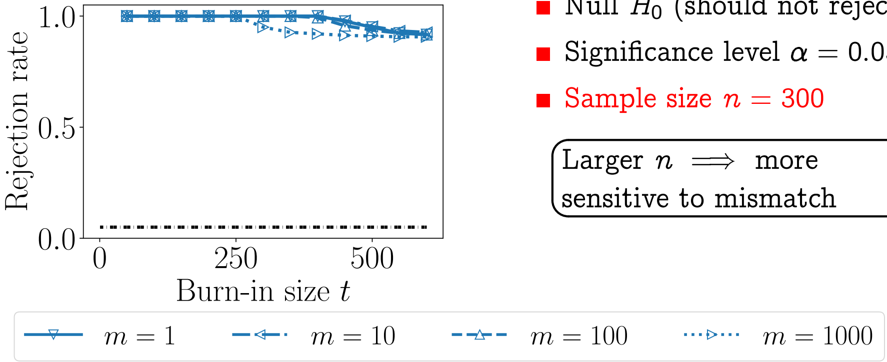
- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 300$

Can sampler influence test power?

How important is the quality of $\frac{1}{m}\sum_{j=1}^{m} \mathrm{s}_p(x|z_j^{(t)})$?

Experiment with PPCA:

- P : MALA with a bad step size (poor sampler)
- Q : NUTS-HMC (good sampler)



- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 300$

Larger $n \implies$ more sensitive to mismatch

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

Research support

Work supported by:

The Gatsby Charitable Foundation

Deepmind

Questions?

KSD Riesz condition proof (detailed)

The KSD is written:

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

KSD Riesz condition proof (detailed)

The KSD is written:

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

Step 2: show that

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

KSD Riesz condition proof (detailed)

The KSD is written:

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

Step 2: show that

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

Riesz theorem!

Next step: taking expectations

Riesz theorem: need boundedness,

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

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

By Jensen and Cauchy-Schwarz,

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

Next step: taking expectations

Riesz theorem: need boundedness,

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

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

By Jensen and Cauchy-Schwarz,

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

Next step: taking expectations

Compute the squared norm:

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

$$= \left\langle \left(\frac{d}{dz} \log p(z) \right) k(z, \cdot) + \|\xi_z\|^2_{\mathcal{F}} \, \|\xi_z\|^2_{\mathcal{F}} \, \|\xi_z\|^2_{\mathcal{F}} \, \frac{d}{dz} k(z, \cdot), \ldots \right\rangle_{\mathcal{F}}$$

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

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

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

Next step: taking expectations

Compute the squared norm:

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

Next step: taking expectations

Compute the squared norm:

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

Next step: taking expectations

Compute the squared norm:

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

Next step: taking expectations

Compute the squared norm:

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

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):

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

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

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

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

First two (easy) terms

Second term (B):

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

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

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

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

First two (easy) terms

Second term (B):

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

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

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

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

First two (easy) terms

Second term (B):

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

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

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

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

Third term

Third term (C):

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

Third term

Third term (C):

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

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

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

$$= 0.$$

Third term

Third term (C):

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

Third term

Third term (C):

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

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