A Fast Goodness-of-Fit Test with Analytic Kernel Embeddings

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

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

Greek Stochastics (Milos, Greece)

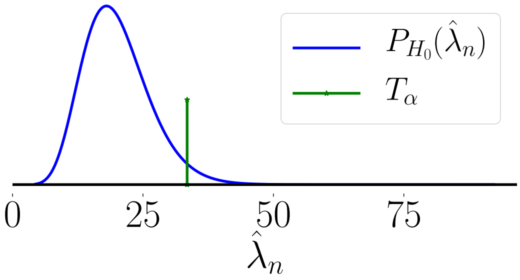
14 July 2017

What Is Goodness-of-fit Testing?

- Given a known density p (model), and sample $\{\mathbf{x}_i\}_{i=1}^n \sim q$ (unknown) defined on $\mathcal{X} \subseteq \mathbb{R}^d$, test

$$\begin{aligned} H_0 &: p = q, \\ \text{vs. } H_1 &: p \neq q, \end{aligned}$$

  $\equiv$ test whether $\{\mathbf{x}_i\}_{i=1}^n \sim p$.
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n > T_\alpha$ (threshold).
- $T_\alpha = (1 - \alpha)$-quantile of the null distribution.



Settings & Motivations

- Many classic tests assume a family for p (e.g., Gaussian), or are for univariate variables.
- **Want** a multivariate nonparametric test.

Recent kernel Stein discrepancy (KSD) test [Liu et al., 2016, Chwialkowski et al., 2016]:

- ✓ Nonparametric i.e., mild assumption on p, q. Kernel-based.
- ✗ Slow. Runtime: $\mathcal{O}(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernel.

Propose the Finite-Set Stein Discrepancy (FSSD).

1. Nonparametric.
2. Linear-time. Runtime complexity: $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., well-defined criterion for parameter tuning.
4. Interpretable. Tells where the model does not fit the data.

Settings & Motivations

- Many classic tests assume a family for p (e.g., Gaussian), or are for univariate variables.
- **Want** a multivariate nonparametric test.

Recent kernel Stein discrepancy (KSD) test
[Liu et al., 2016, Chwialkowski et al., 2016]:

- ✓ Nonparametric i.e., mild assumption on p, q. Kernel-based.
- ✗ Slow. Runtime: $\mathcal{O}(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernel.

Propose the Finite-Set Stein Discrepancy (FSSD).

1. Nonparametric.
2. Linear-time. Runtime complexity: $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., well-defined criterion for parameter tuning.
4. Interpretable. Tells where the model does not fit the data.

Settings & Motivations

- Many classic tests assume a family for p (e.g., Gaussian), or are for univariate variables.
- **Want** a multivariate nonparametric test.

Recent kernel Stein discrepancy (KSD) test [Liu et al., 2016, Chwialkowski et al., 2016]:

- ✓ Nonparametric i.e., mild assumption on p, q. Kernel-based.
- ✗ Slow. Runtime: $\mathcal{O}(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernel.

Propose the Finite-Set Stein Discrepancy (FSSD).

1. Nonparametric.
2. Linear-time. Runtime complexity: $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., well-defined criterion for parameter tuning.
4. Interpretable. Tells where the model does not fit the data.

Stein Idea in Kernel Stein Discrepancy (KSD)

- Consider $d = 1$.
- Define a **Stein operator** of p as

$$(T_p f)(x) = \frac{\partial_x [f(x)p(x)]}{p(x)},$$

for some real-valued function f.

- Assume $\lim_{|x|\to\infty} f(x)p(x) = 0$. Then,

$$\boxed{\mathbb{E}_{x\sim q}(T_p f)(x) = 0 \iff p = q.}$$

- Proof of $\Leftarrow$

$$\begin{aligned}\mathbb{E}_{x\sim p}(T_p f)(x) &= \int_{-\infty}^{\infty} \frac{\partial_x [f(x)p(x)]}{\cancel{p(x)}} \cancel{p(x)} \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \partial_x [f(x)p(x)] \, \mathrm{d}x = [f(x)p(x)]_{x=-\infty}^{x=\infty} = 0.\end{aligned}$$

- Only certain f makes $\Rightarrow$ true.

Stein Idea in Kernel Stein Discrepancy (KSD)

- Consider $d = 1$.
- Define a **Stein operator** of p as

$$(T_p f)(x) = \frac{\partial_x[f(x)p(x)]}{p(x)},$$

for some real-valued function f.
- Assume $\lim_{|x|\to\infty} f(x)p(x) = 0$. Then,

$$\boxed{\mathbb{E}_{x\sim q}(T_p f)(x) = 0 \iff p = q.}$$

- Proof of $\Leftarrow$

$$\mathbb{E}_{x\sim p}(T_p f)(x) = \int_{-\infty}^{\infty} \frac{\partial_x[f(x)p(x)]}{\cancel{p(x)}} \cancel{p(x)}\, \mathrm{d}x$$
$$= \int_{-\infty}^{\infty} \partial_x[f(x)p(x)]\, \mathrm{d}x = [f(x)p(x)]_{x=-\infty}^{x=\infty} = 0.$$

- Only certain f makes $\Rightarrow$ true.

Stein Idea in Kernel Stein Discrepancy (KSD)

- Consider $d = 1$.
- Define a **Stein operator** of p as

$$(T_p f)(x) = \frac{\partial_x [f(x)p(x)]}{p(x)},$$

for some real-valued function f.

- Assume $\lim_{|x|\to\infty} f(x)p(x) = 0$. Then,

$$\boxed{\mathbb{E}_{x\sim q}(T_p f)(x) = 0 \iff p = q.}$$

- Proof of $\Leftarrow$

$$\begin{aligned}\mathbb{E}_{x\sim p}(T_p f)(x) &= \int_{-\infty}^{\infty} \frac{\partial_x [f(x)p(x)]}{\cancel{p(x)}} \cancel{p(x)} \,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \partial_x [f(x)p(x)] \,\mathrm{d}x = [f(x)p(x)]_{x=-\infty}^{x=\infty} = 0.\end{aligned}$$

- Only certain f makes $\Rightarrow$ true.

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \langle f, k'(\cdot, x) \rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = (\langle a, b \rangle + \sigma)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a - b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \left\langle f, k'(\cdot, x)\right\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = \left(\langle a, b\rangle + \sigma\right)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a-b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \langle f, k'(\cdot, x)\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = \left(\langle a, b\rangle + \sigma\right)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a-b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \langle f, k'(\cdot, x)\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = (\langle a, b\rangle + \sigma)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a-b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \langle f, k'(\cdot, x)\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = (\langle a, b\rangle + \sigma)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a - b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \left\langle f, k'(\cdot, x)\right\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = \left(\langle a, b\rangle + \sigma\right)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a - b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- RKHS: computational tractability.
  - $\mathcal{F} = \{\mathcal{X} \to \mathbb{R}$ functions$\}$ Hilbert space with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ repr. kernel if
    1. for all $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{F}$ (generators),
    2. $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{F}}$ (reproducing property).
  - $\exists \phi : \mathcal{X} \to \mathcal{F}$ Hilbert such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.
- Similarly for derivatives

$$f'(x) = \left\langle f, k'(\cdot, x) \right\rangle_{\mathcal{F}}.$$

- Examples:

$$k_G(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad k_p(a, b) = \left(\langle a, b \rangle + \sigma\right)^p,$$

$$k_{M,\frac{3}{2}}(a, b) = \left(1 + \frac{\sqrt{3}\,\|a - b\|_2}{\sigma}\right) e^{-\frac{\sqrt{3}\|a-b\|_2}{\sigma}}.$$

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- KSD = square of

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \sup_{\|f\|_{\mathcal{F}}\leq 1} \left\langle f, \underbrace{\mathbb{E}_{x\sim q}\left\{k(\cdot,x)\partial_x \log p(x) + \partial_x k(\cdot,x)\right\}}_{=:g} \right\rangle_{\mathcal{F}}$$

$$= \|g\|_{\mathcal{F}},$$

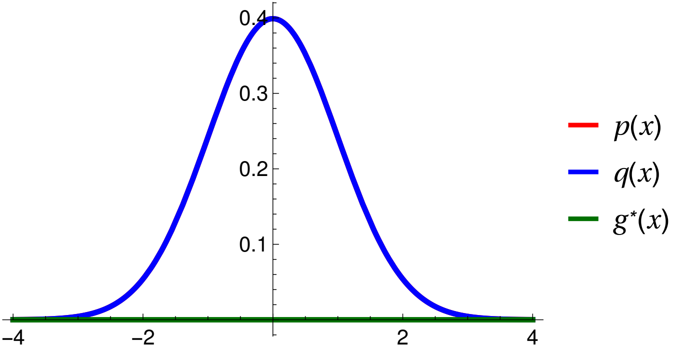
- Take the RKHS norm of **Stein witness** function $g = g^*$.

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- KSD = square of

$$\sup_{\|f\|_{\mathcal{F}}\leq 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \sup_{\|f\|_{\mathcal{F}}\leq 1} \left\langle f, \underbrace{\mathbb{E}_{x\sim q}\left\{k(\cdot,x)\partial_x \log p(x) + \partial_x k(\cdot,x)\right\}}_{=:g} \right\rangle_{\mathcal{F}}$$

$$= \|g\|_{\mathcal{F}},$$



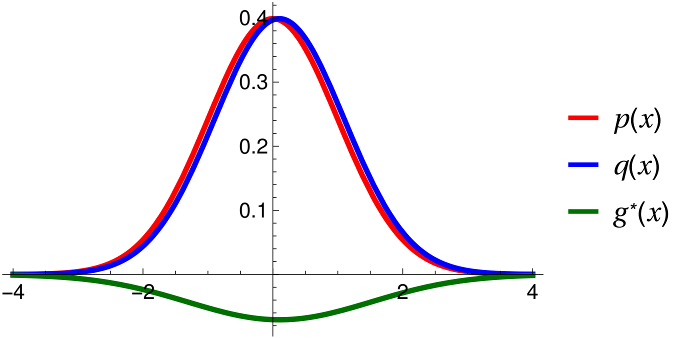
- Take the RKHS norm of **Stein witness** function $g = g^*$.

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- KSD = square of

$$\sup_{\|f\|_{\mathcal{F}} \le 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \sup_{\|f\|_{\mathcal{F}} \le 1} \left\langle f, \underbrace{\mathbb{E}_{x\sim q}\left\{k(\cdot,x)\partial_x \log p(x) + \partial_x k(\cdot,x)\right\}}_{=:g} \right\rangle_{\mathcal{F}}$$

$$= \|g\|_{\mathcal{F}},$$



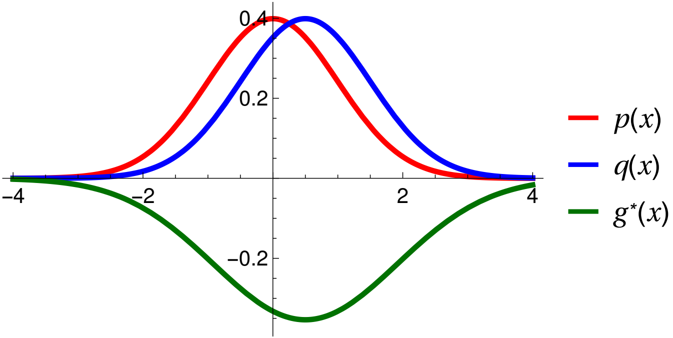
- Take the RKHS norm of **Stein witness function** $g = g^*$.

Kernel Stein Discrepancy (KSD)

- If considering all $f \in$ unit ball in an RKHS $\mathcal{F}$, then $\Rightarrow$ holds.
- KSD = square of

$$\sup_{\|f\|_{\mathcal{F}} \le 1} \mathbb{E}_{x\sim q}(T_p f)(x) = \sup_{\|f\|_{\mathcal{F}} \le 1} \left\langle f, \underbrace{\mathbb{E}_{x\sim q}\left\{k(\cdot,x)\partial_x \log p(x) + \partial_x k(\cdot,x)\right\}}_{=:g} \right\rangle_{\mathcal{F}}$$

$$= \|g\|_{\mathcal{F}},$$



- Take the RKHS norm of **Stein witness** function $g = g^*$.

Kernel Stein Discrepancy (KSD)

Closed-form expression for KSD: given $x, x' \sim q$, then
[Liu et al., 2016, Chwialkowski et al., 2016]

$$S^2 = \|g\|_{\mathcal{F}}^2 = \overbrace{\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}}^{\text{double sum}} h_p(x, x')$$

where

$$\begin{aligned} h_p(x, y) &:= [\partial_x \log p(x)]\, k(x, y)\, [\partial_x \log p(y)] \\ &+ [\partial_y \log p(y)]\, \partial_x k(x, y) \\ &+ [\partial_x \log p(x)]\, \partial_y k(x, y) \\ &+ \partial_x \partial_y k(x, y) \end{aligned}$$

and k is RKHS kernel for $\mathcal{F}$.

- ✓ Only depends on kernel k and $\partial_x \log p(x)$.
- ✓ Do not need to normalize p, or sample from it.
- ✗ The "double sum" makes it $\mathcal{O}(d^2 n^2)$. Slow.

Kernel Stein Discrepancy (KSD)

Closed-form expression for KSD: given $x, x' \sim q$, then
[Liu et al., 2016, Chwialkowski et al., 2016]

$$S^2 = \|g\|_{\mathcal{F}}^2 = \overbrace{\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}}^{\text{double sum}} h_p(x, x')$$

where

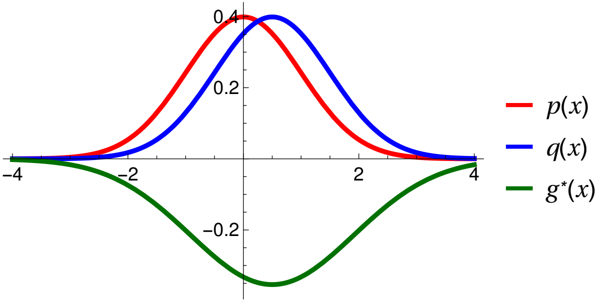
$$\begin{aligned} h_p(x, y) := & \left[\partial_x \log p(x)\right] k(x, y) \left[\partial_x \log p(y)\right] \\ & + \left[\partial_y \log p(y)\right] \partial_x k(x, y) \\ & + \left[\partial_x \log p(x)\right] \partial_y k(x, y) \\ & + \partial_x \partial_y k(x, y) \end{aligned}$$

and k is RKHS kernel for $\mathcal{F}$.

- ✓ Only depends on kernel k and $\partial_x \log p(x)$.
- ✓ Do not need to normalize p, or sample from it.
- ✗ The "double sum" makes it $\mathcal{O}(d^2 n^2)$. Slow.

Proposal: the Finite Set Stein Discrepancy (FSSD)

Take g (Stein witness function), and evaluate g^2 at finitely many locations.



- Test locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\} \subset \mathbb{R}^d$.
- Population FSSD (when $d = 1$)

$$\mathrm{FSSD}^2 := \frac{1}{J} \sum_{j=1}^{J} g^2(\mathbf{v}_j).$$

- g can be computed in $\mathcal{O}(d^2 n)$.

FSSD is a Discrepancy Measure

Theorem 1.

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

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

Then, η-almost surely

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

- 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

$$\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}) := \frac{1}{p(\mathbf{x})} \partial_{\mathbf{x}}[p(\mathbf{x}) k(\mathbf{x}, \mathbf{v})] \in \mathbb{R}^d.$$

- d-output Stein witness

$$\mathbf{g}(\mathbf{v}) = \mathbb{E}_{\mathbf{x} \sim q} \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^d.$$

- General form:

$$\mathrm{FSSD}^2 = \frac{1}{dJ} \sum_{j=1}^{J} \|\mathbf{g}(\mathbf{v}_j)\|_2^2,$$

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

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_1), \ldots \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$; $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$.
- $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$.

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\{\omega_i\}_{i=1}^{dJ}$ the eigenvalues of $\boldsymbol{\Sigma}_p$, $\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}\left\|\boldsymbol{\tau}(x)^T\boldsymbol{\tau}(x')\right\|_2^2 < \infty$.

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 $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu}$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

But, how to estimate $\boldsymbol{\Sigma}_p$? No sample from p!

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

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_1), \ldots \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$; $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$.
- $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$.

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\{\omega_i\}_{i=1}^{dJ}$ *the eigenvalues of* $\boldsymbol{\Sigma}_p$, $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}\left\|\boldsymbol{\tau}(\mathbf{x})^T\boldsymbol{\tau}(\mathbf{x}')\right\|_2^2 < \infty$.

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* $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu}$. *Implies* $\mathbb{P}($*reject* $H_0) \to 1$ *as* $n \to \infty$.

But, how to estimate $\boldsymbol{\Sigma}_p$? No sample from p!

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

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_1), \ldots \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$; $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$.
- $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$.

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\{\omega_i\}_{i=1}^{dJ}$ the eigenvalues of $\boldsymbol{\Sigma}_p$, $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}\left\|\boldsymbol{\tau}(\mathbf{x})^T\boldsymbol{\tau}(\mathbf{x}')\right\|_2^2 < \infty$.

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 $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu}$. Implies $\mathbb{P}$(reject H_0) $\to 1$ as $n \to \infty$.*

But, how to estimate $\boldsymbol{\Sigma}_p$? No sample from p!

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

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_1), \ldots \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$; $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$.
- $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$.

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\{\omega_i\}_{i=1}^{dJ}$ the eigenvalues of $\boldsymbol{\Sigma}_p$, $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}\left\|\boldsymbol{\tau}(\mathbf{x})^T\boldsymbol{\tau}(\mathbf{x}')\right\|_2^2 < \infty$.

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 $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu}$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

But, how to estimate $\boldsymbol{\Sigma}_p$? No sample from p!

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

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

- $\boldsymbol{\tau}(\mathbf{x}) :=$ vertically stack $\boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_1), \ldots \boldsymbol{\xi}(\mathbf{x}, \mathbf{v}_J) \in \mathbb{R}^{dJ}$. Feature vector of $\mathbf{x}$.
- Mean feature: $\boldsymbol{\mu} := \mathbb{E}_{\mathbf{x}\sim q}[\boldsymbol{\tau}(\mathbf{x})]$; $\mathrm{FSSD}^2 = \frac{1}{dJ}\|\boldsymbol{\mu}\|_2^2$.
- $\boldsymbol{\Sigma}_r := \mathrm{cov}_{\mathbf{x}\sim r}[\boldsymbol{\tau}(\mathbf{x})] \in \mathbb{R}^{dJ\times dJ}$ for $r \in \{p, q\}$.

Proposition 1 (Asymptotic distributions).

Let $Z_1, \ldots, Z_{dJ} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\{\omega_i\}_{i=1}^{dJ}$ the eigenvalues of $\boldsymbol{\Sigma}_p$, $\mathbb{E}_{\mathbf{x}\sim q}\mathbb{E}_{\mathbf{x}'\sim q}\left\|\boldsymbol{\tau}(\mathbf{x})^T\boldsymbol{\tau}(\mathbf{x}')\right\|_2^2 < \infty$.

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 $\sigma_{H_1}^2 := 4\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_q\boldsymbol{\mu}$. Implies $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

But, how to estimate $\boldsymbol{\Sigma}_p$? No sample from p!

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

Parameter Tuning

- Any random locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ work when $n \to 0$. But, for finite n, tuning will increase the performance.
- Test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$.

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 on tr. Goodness-of-fit test on te.

Parameter Tuning

- Any random locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ work when $n \to 0$. But, for finite n, tuning will increase the performance.
- Test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$.

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 on tr. Goodness-of-fit test on te.

Parameter Tuning

- Any random locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ work when $n \to 0$. But, for finite n, tuning will increase the performance.
- Test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$.

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 on tr. Goodness-of-fit test on te.

Parameter Tuning

- Any random locations $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_J\}$ work when $n \to 0$. But, for finite n, tuning will increase the performance.
- Test power $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$.

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 $\mathcal{N}(0,1)$.

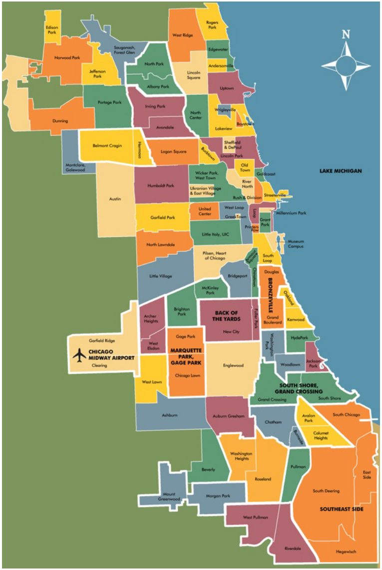
- For large n, second term dominates. So

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

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

Interpretable Features: Chicago Crime

- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



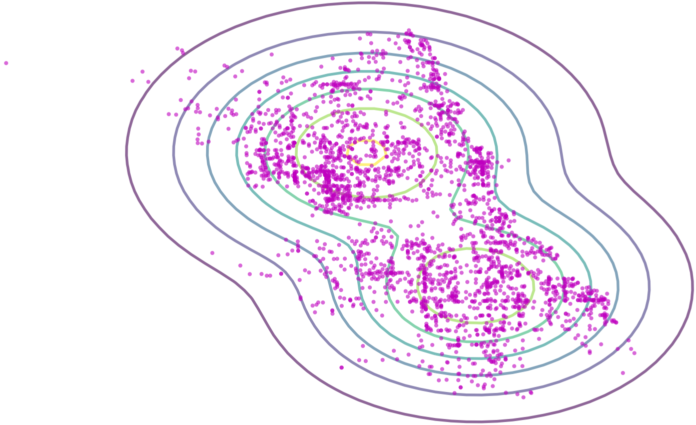
Interpretable Features: Chicago Crime

- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.

Robbery events = data from q.

Interpretable Features: Chicago Crime

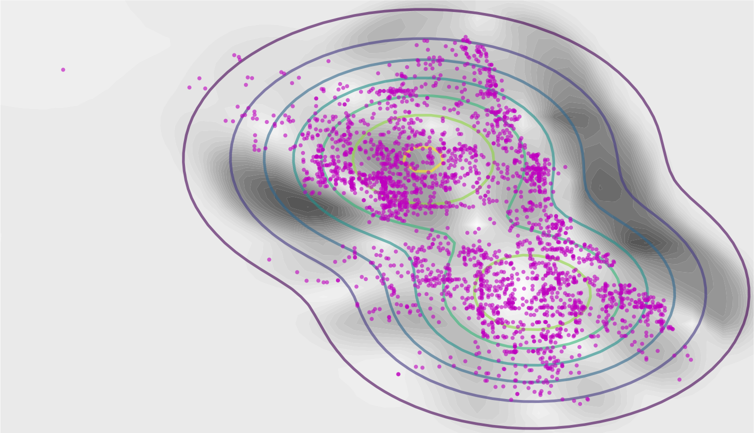
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Fit a 2-component Gaussian mixture $\rightarrow p$.

Interpretable Features: Chicago Crime

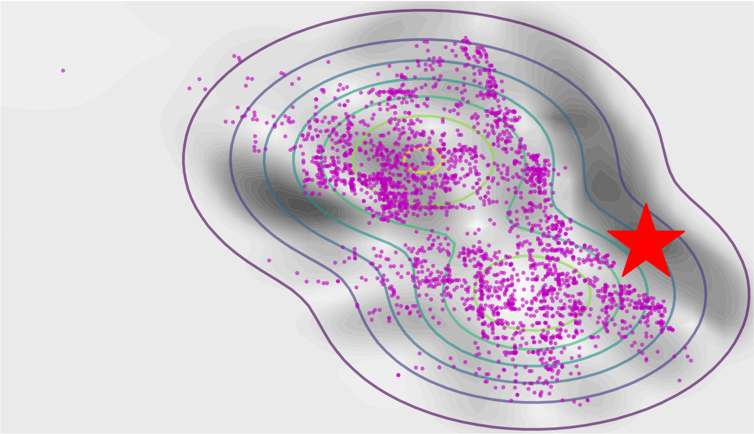
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Optimization objective $\frac{\widehat{\mathrm{FSSD}^2}}{\widehat{\sigma_{H_1}}}$.

Interpretable Features: Chicago Crime

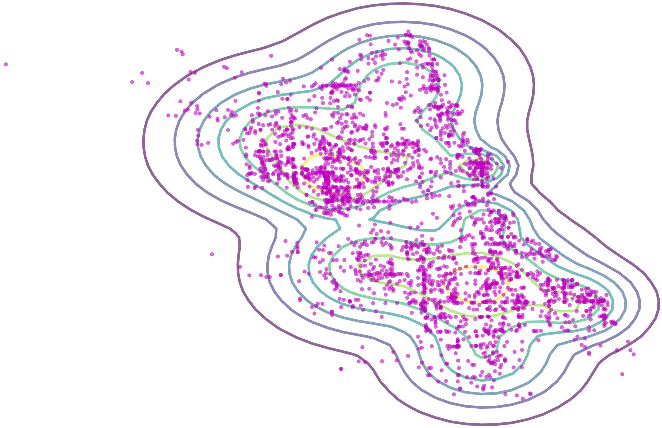
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



No Gaussian tail on the right. Lake Michigan, sharp data boundary.

Interpretable Features: Chicago Crime

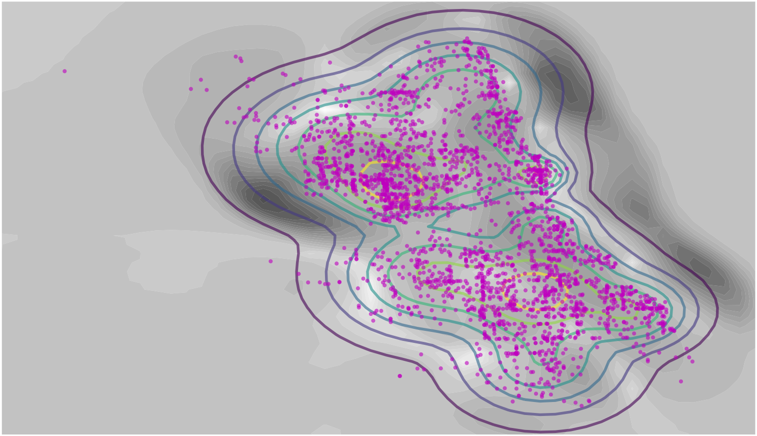
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Fit a 10-component Gaussian mixture $\rightarrow p$.

Interpretable Features: Chicago Crime

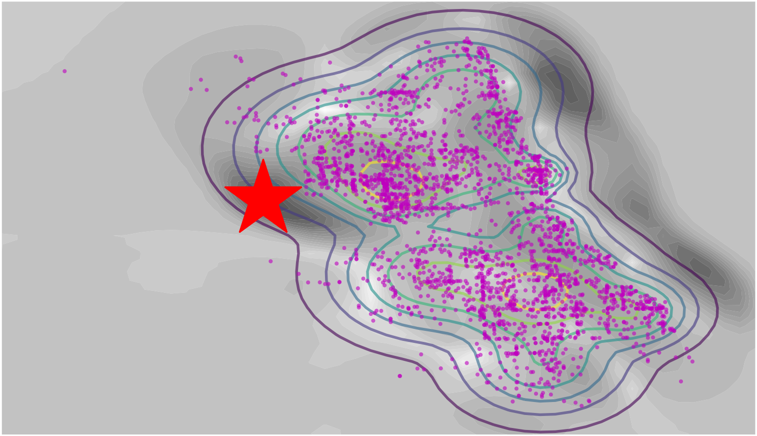
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Capture the right tail better.

Interpretable Features: Chicago Crime

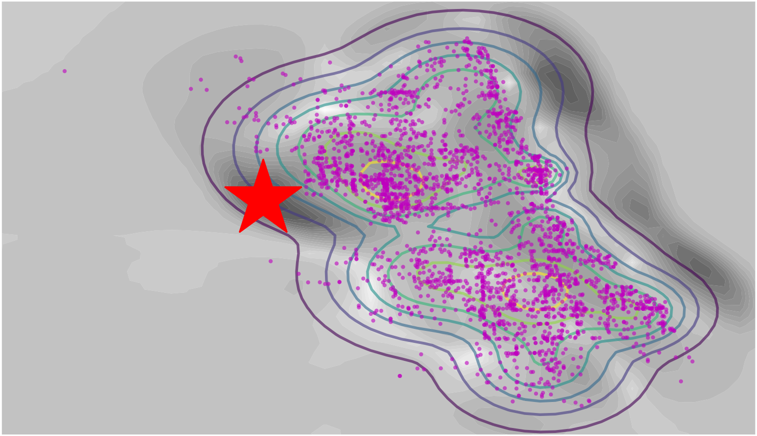
- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Still does not capture the left tail.

Interpretable Features: Chicago Crime

- $n = 11957$ robbery events in Chicago in 2016.
- Model spatial density with Gaussian mixtures.



Still does not capture the left tail.

FSSD features (test locations) are interpretable.

Simulation Settings

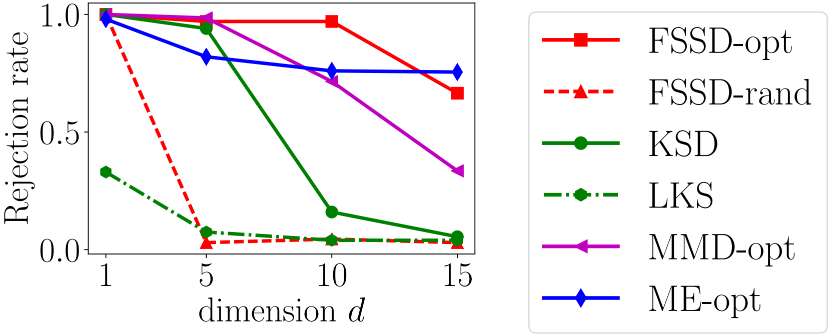
- Gaussian kernels $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|_2^2}{2\sigma_k^2}\right)$

	Method	Description
1	**FSSD-opt**	Proposed. With optimization. $\mathcal{O}(n)$.
2	FSSD-rand	Proposed. Random test locations.
3	KSD	Quadratic-time kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]
4	LKS	Linear-time running average version of KSD.
5	MMD-opt	MMD two-sample test [Gretton et al., 2012]. With optimization.
6	ME-test	Mean Embeddings two-sample test [Jitkrittum et al., 2016]. With optimization.

- FSSD tests use $J = 5$ locations.
- Two-sample tests need to draw sample from p.
- Tests with optimization use 20% of the data.
- $\alpha = 0.05$. 200 trials.

Gaussian vs. Laplace

- p = Gaussian. q = Laplace. Same mean and variance. High-order moments differ.
- Sample size $n = 1000$.



- Optimization increases the power.
- Two-sample tests can perform well in this case (p, q clearly differ).

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$

  where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{h} \in \{\pm 1\}^{d_h}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$. $d = 50$, $d_h = 40$.

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$

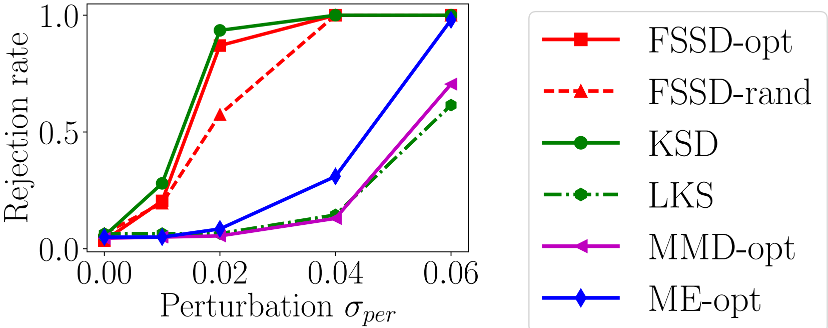
  where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{h} \in \{\pm 1\}^{d_h}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$. $d = 50$, $d_h = 40$.

Gaussian-Bernoulli Restricted Boltzmann Machine (RBM)

- $p(\mathbf{x})$ is the marginal of

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left(\mathbf{x}^\top \mathbf{B}\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2\right),$$

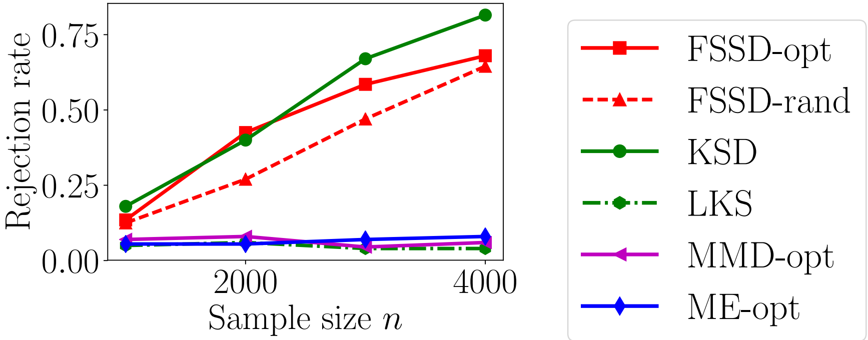
  where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{h} \in \{\pm 1\}^{d_h}$ is latent. Randomly pick $\mathbf{B}, \mathbf{b}, \mathbf{c}$.
- $q(\mathbf{x}) = p(\mathbf{x})$ with i.i.d. $\mathcal{N}(0, \sigma_{per})$ noise added to all entries of $\mathbf{B}$.
- Sample size $n = 1000$. $d = 50$, $d_h = 40$.



KSD, FSSD-opt comparable. LKS has low power.

Harder RBM Problem

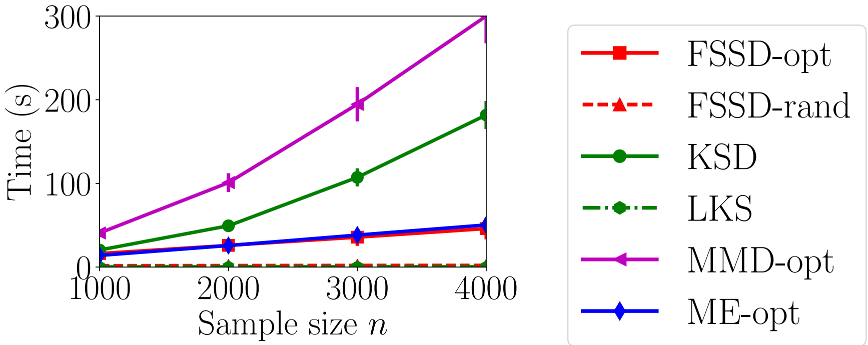
- Now, perturb only one entry of $\mathbf{B} \in \mathbb{R}^{50\times 40}$.
- $B_{1,1} \leftarrow B_{1,1} + \mathcal{N}(0, \sigma^2_{per} = 0.1^2)$. Entries of $\mathbf{B}$ are random $\{-1, 1\}$.



- Two-sample tests fail. Samples from p, q look roughly the same.
- FSSD-opt is comparable to KSD at low n. One order of magnitude faster.

Harder RBM Problem

- Now, perturb only one entry of $\mathbf{B} \in \mathbb{R}^{50\times 40}$.
- $B_{1,1} \leftarrow B_{1,1} + \mathcal{N}(0, \sigma_{per}^2 = 0.1^2)$. Entries of $\mathbf{B}$ are random $\{-1, 1\}$.



- Two-sample tests fail. Samples from p, q look roughly the same.
- FSSD-opt is comparable to KSD at low n. One order of magnitude faster.

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 $c(\mathbf{0}) = 0$. [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

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

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 $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2}c(\theta)n\right)$ where $c(\theta) > 0$ under H_1, and $c(\mathbf{0}) = 0$. [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.

Bahadur slope

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

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

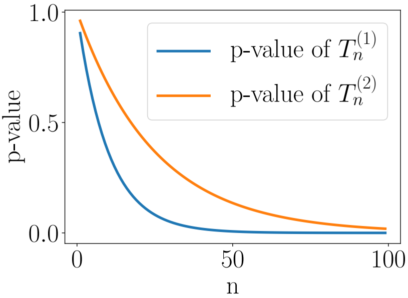
- Bahadur efficiency = ratio of slopes of two tests.

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.

$$
\begin{aligned}
H_0 &: \theta = \mathbf{0}, \\
H_1 &: \theta \neq \mathbf{0}.
\end{aligned}
$$

- Typically $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2}c(\theta)n\right)$ where $c(\theta) > 0$ under H_1, and $c(\mathbf{0}) = 0$. [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.



Bahadur slope

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

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

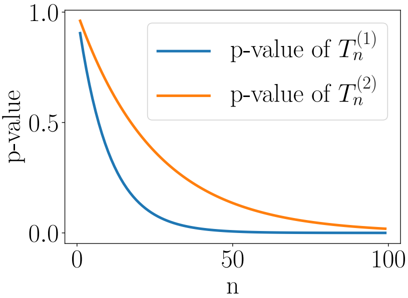
- Bahadur efficiency = ratio of slopes of two tests.

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 $\mathrm{pval}_n \approx \exp\left(-\frac{1}{2}c(\theta)n\right)$ where $c(\theta) > 0$ under H_1, and $c(\mathbf{0}) = 0$. [Bahadur, 1960].
- $c(\theta)$ higher $\implies$ more sensitive. Good.



Bahadur slope

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

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

- Bahadur efficiency = ratio of slopes of two tests.

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

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

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 $= \kappa^2$).

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

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

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

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 $= \kappa^2$).

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

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

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

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 $= \kappa^2$).

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

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

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

Conclusions

- Proposed The Finite Set Stein Discrepancy (**FSSD**).
- Goodness-of-fit based on FSSD is
  1. nonparametric,
  2. linear-time,
  3. adaptive (parameters automatically tuned),
  4. interpretable.
- When $p = \mathcal{N}(0, 1), q = \mathcal{N}(\mu_q, 1)$, FSSD is theoretically at least two times more efficient (Bahadur efficiency) than LKS.

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`

Questions?

Thank you

Linear-Time Kernel Stein Discrepancy (LKS)

- [Liu et al., 2016] also proposed a linear version of KSD.
- For $\{\mathbf{x}_i\}_{i=1}^n \sim q$, KSD test statistic is

$$\widehat{S^2} = \frac{2}{n(n-1)} \sum_{i<j} h_p(\mathbf{x}_i, \mathbf{x}_j).$$

- LKS test statistic is a "running average"

$$\widehat{S_l^2} = \frac{2}{n} \sum_{i=1}^{n/2} h_p(\mathbf{x}_{2i-1}, \mathbf{x}_{2i}).$$

- Both unbiased. LKS has $\mathcal{O}(d^2 n)$ runtime.
- ✗ LKS has high variance. Poor test power.
  - We will show this empirically and theoretically.

FSSD and KSD in 1D Gaussian Case

Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$.

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

$$\mathrm{FSSD}^2 = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}} \left(\left(\sigma_k^2+1\right)\mu_q + v\left(\sigma_q^2-1\right)\right)^2}{\left(\sigma_k^2+\sigma_q^2\right)^3}.$$

- If $\mu_q \neq 0, \sigma_q^2 \neq 1$, and $v = -\frac{(\sigma_k^2+1)\mu_q}{(\sigma_q^2-1)}$, then $\mathrm{FSSD}^2 = 0$!
  - This is why v should be drawn from a distribution with a density.
- For KSD, Gaussian kernel (bandwidth $= \kappa^2$).

$$S^2 = \frac{\mu_q^2\left(\kappa^2+2\sigma_q^2\right) + \left(\sigma_q^2-1\right)^2}{\left(\kappa^2+2\sigma_q^2\right)\sqrt{\frac{2\sigma_q^2}{\kappa^2}+1}}.$$

FSSD and KSD in 1D Gaussian Case

Consider $p = \mathcal{N}(0,1)$ and $q = \mathcal{N}(\mu_q, \sigma_q^2)$.

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

$$\mathrm{FSSD}^2 = \frac{\sigma_k^2 e^{-\frac{(v-\mu_q)^2}{\sigma_k^2+\sigma_q^2}} \left(\left(\sigma_k^2 + 1\right) \mu_q + v \left(\sigma_q^2 - 1\right) \right)^2}{\left(\sigma_k^2 + \sigma_q^2\right)^3}.$$

- If $\mu_q \neq 0, \sigma_q^2 \neq 1$, and $v = -\frac{(\sigma_k^2+1)\mu_q}{(\sigma_q^2-1)}$, then $\mathrm{FSSD}^2 = 0$!
  - This is why v should be drawn from a distribution with a density.
- For KSD, Gaussian kernel (bandwidth $= \kappa^2$).

$$S^2 = \frac{\mu_q^2 \left(\kappa^2 + 2\sigma_q^2\right) + \left(\sigma_q^2 - 1\right)^2}{\left(\kappa^2 + 2\sigma_q^2\right) \sqrt{\frac{2\sigma_q^2}{\kappa^2} + 1}}.$$

Illustration: Optimization Objective

- Consider $J = 1$ location. In $\mathbb{R}^2$.
- Training objective $\frac{\widehat{\mathrm{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, $\bigstar$ = best $\mathbf{v}$.

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

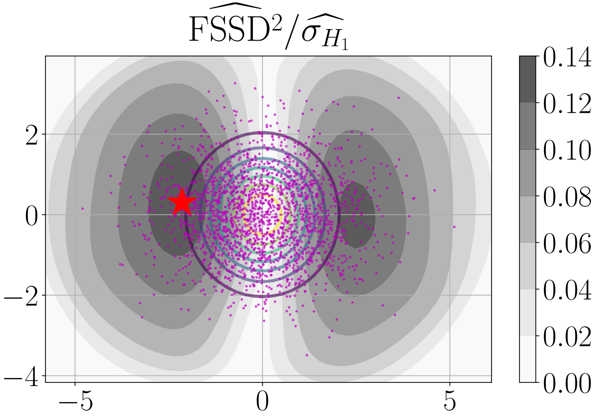
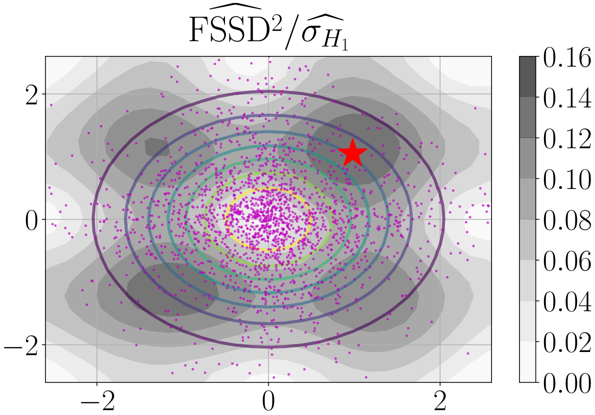


Illustration: Optimization Objective

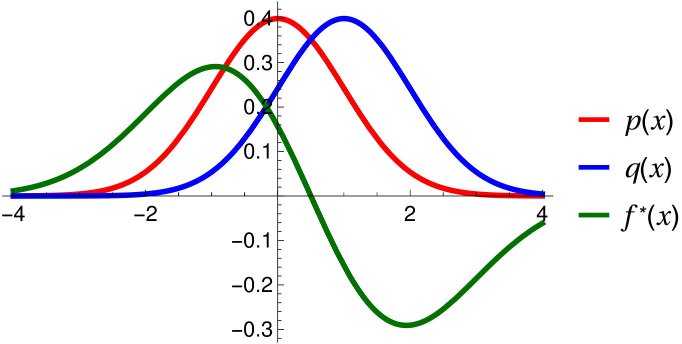
- Consider $J = 1$ location. In $\mathbb{R}^2$.
- Training objective $\frac{\widehat{\mathrm{FSSD}^2}(\mathbf{v})}{\widehat{\sigma_{H_1}}(\mathbf{v})}$ (gray), p in wireframe, $\{\mathbf{x}_i\}_{i=1}^n \sim q$ in purple, ★ = best $\mathbf{v}$.

$p = \mathcal{N}(\mathbf{0}, \mathbf{I})$ vs. $q = \mathrm{Laplace}$ with same mean & variance.



Statistical Model Criticism with MMD

$$MMD(p, q) = \|f^*\|^2 = \sup_{\|f\|_{\mathcal{F}} \leq 1} [E_p f - E_p f]$$



$f^*(x)$ is the witness function

Can we compute MMD with samples from q and a **model** p?

Problem: usually can't compute $E_p f$ in closed form.

Proof of Stein Idea

Consider the class

$$G = \{\partial_x f + f(\partial_x \log p) | f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned}\mathbb{E}_p g(X) &= \mathbb{E}_p \left[\partial_x f(X) + f(X)\partial_x \log p(X)\right] \\ &= \int \partial_x f(x)p(x) + f(x)\partial_x p(x)dx \\ &= \int_{-\infty}^{\infty} (f(x)p(x))dx \\ &= [f(x)p(x)]_{x=-\infty}^{x=\infty} \\ &= 0\end{aligned}$$

Proof of Stein Idea

Consider the class

$$G = \{\partial_x f + f(\partial_x \log p) | f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned}
\mathbb{E}_p g(X) &= \mathbb{E}_p \left[\partial_x f(X) + f(X) \partial_x \log p(X)\right] \\
&= \int \partial_x f(x) p(x) + f(x) \partial_x p(x) dx \\
&= \int_{-\infty}^{\infty} (f(x) p(x)) dx \\
&= [f(x) p(x)]_{x=-\infty}^{x=\infty} \\
&= 0
\end{aligned}$$

Proof of Stein Idea

Consider the class

$$G = \{\partial_x f + f(\partial_x \log p) | f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned}
\mathbb{E}_p g(X) &= \mathbb{E}_p \left[\partial_x f(X) + f(X)\partial_x \log p(X)\right] \\
&= \int \partial_x f(x)p(x) + f(x)\partial_x p(x) dx \\
&= \int_{-\infty}^{\infty} (f(x)p(x)) dx \\
&= [f(x)p(x)]_{x=-\infty}^{x=\infty} \\
&= 0
\end{aligned}$$

Proof of Stein Idea

Consider the class

$$G = \{\partial_x f + f(\partial_x \log p) | f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned}
\mathbb{E}_p g(X) &= \mathbb{E}_p \left[\partial_x f(X) + f(X)\partial_x \log p(X)\right] \\
&= \int \partial_x f(x)p(x) + f(x)\partial_x p(x)dx \\
&= \int_{-\infty}^{\infty} (f(x)p(x))dx \\
&= [f(x)p(x)]_{x=-\infty}^{x=\infty} \\
&= 0
\end{aligned}$$

Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - E_p T_p g$$

Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$

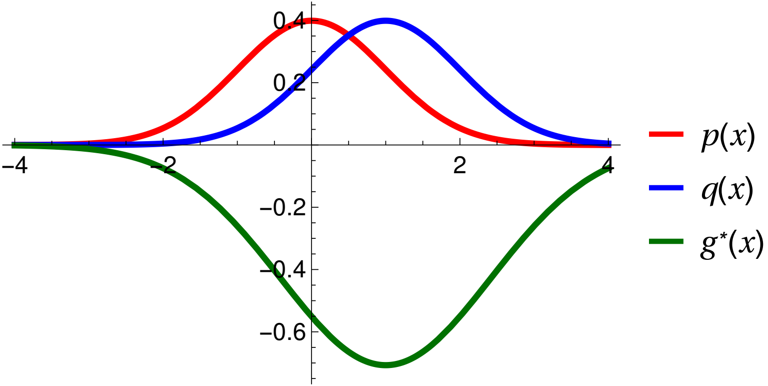
Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



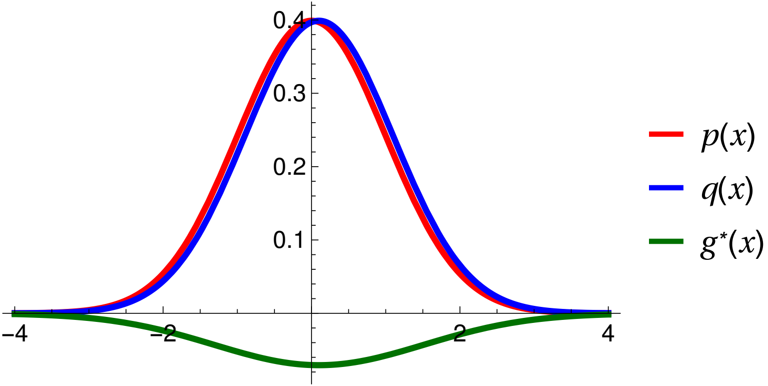
Kernel Stein Discrepancy

Stein operator

$$T_p f = \partial_x f + f \partial_x (\log p)$$

Kernel Stein Discrepancy (KSD)

$$MSD(p, q, \mathcal{F}) = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g - \cancel{E_p T_p g} = \sup_{\|g\|_{\mathcal{F}} \leq 1} E_q T_p g$$



References I

Bahadur, R. R. (1960).
Stochastic comparison of tests.
The Annals of Mathematical Statistics, 31(2):276–295.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *ICML*, pages 2606–2615.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A Kernel Two-Sample Test.
JMLR, 13:723–773.

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

References II

Liu, Q., Lee, J., and Jordan, M. (2016).
A Kernelized Stein Discrepancy for Goodness-of-fit Tests.
In *ICML*, pages 276–284.