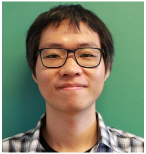
Relative Goodness-of-Fit Tests for Models with Latent Variables

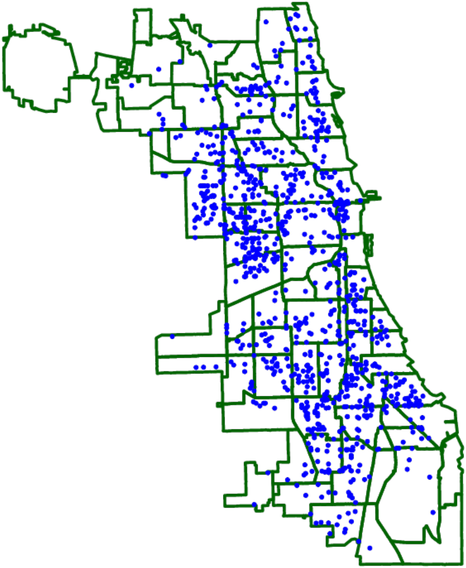
Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

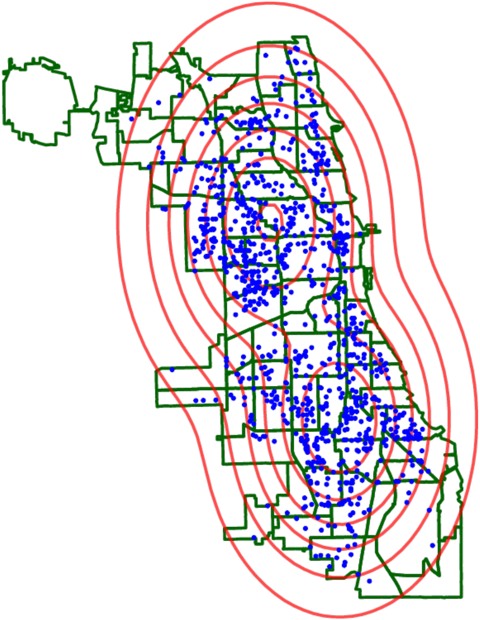
CIRM 2022

Model Criticism



Data = robbery events in Chicago in 2016.

Model Criticism



Is this a good model?

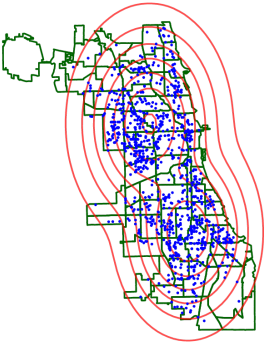
Model Criticism

"All models are wrong."

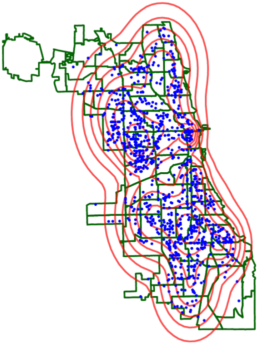
G. Box (1976)

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?



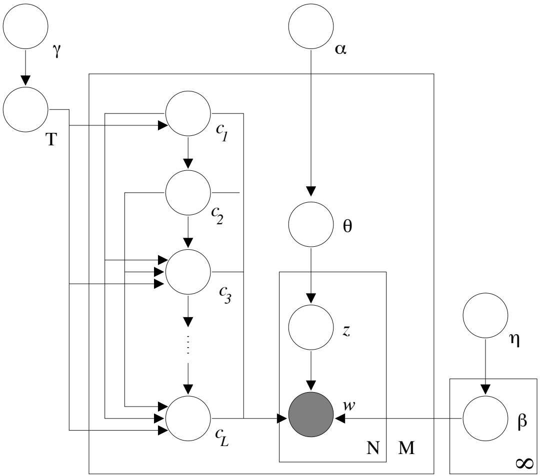
P : two components



Q : ten components

Most interesting models have latent structure

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



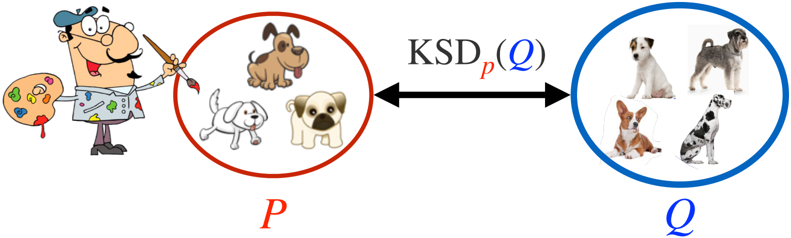
Outline

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

- The 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

Kernel Stein Discrepancy

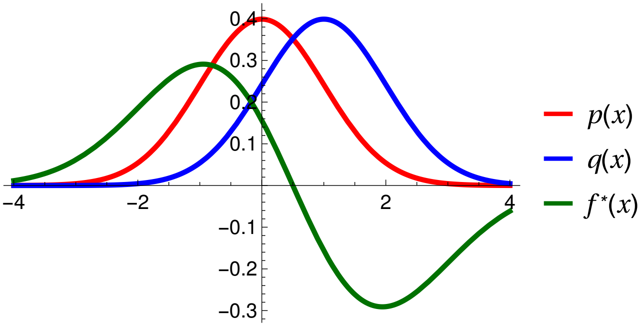
- Model P, data $\{x_i\}_{i=1}^n \sim Q$.
- "All models are wrong" ($P \neq Q$).



MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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



All of kernel methods

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\ldots \varphi_\ell(x) \ldots] \in \ell_2$$

For positive definite k,

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Features are solutions to kernel eigenvalue equation

$$\lambda_\ell e_\ell(x) = \int k(x, x') e_\ell(x') \, dp(x') \, dx'$$

$$\varphi_\ell(x) = \sqrt{\lambda_\ell} e_\ell(x)$$

where $p(x)$ finite Borel measure satisfying Mercer (e.g. supported on $\mathcal{X}$ where $\mathcal{X}$ compact).

All of kernel methods

Kernels: dot products of features

Feature map $\varphi(x) \in \mathcal{F}$,

$$\varphi(x) = [\dots \varphi_\ell(x) \dots] \in \ell_2$$

For positive definite k,

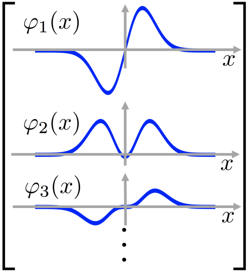
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Infinitely many features $\varphi(x)$, dot product in closed form!

Exponentiated quadratic kernel

$$k(x, x') = \exp\left(-\gamma \left\| x - x' \right\|^2\right)$$

$$p(x) = \mathcal{N}(0, 1)$$

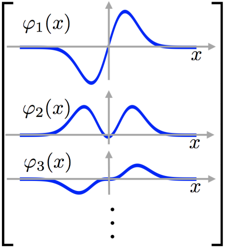
$\varphi(x) =$ 

Features: Gaussian Processes for Machine learning, Rasmussen and Williams, Ch. 4.

All of kernel methods

Functions are linear combinations of features:

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



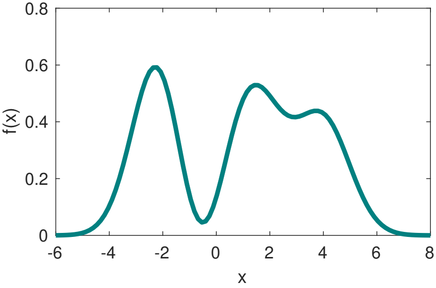
$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^{\,2}$$

All of kernel methods

"The kernel trick"

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x)$$

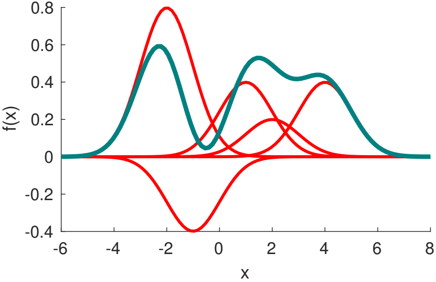
$$= \sum_{i=1}^{m} \alpha_i \underbrace{k(x_i, x)}_{\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{F}}}$$



All of kernel methods

"The kernel trick"

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x)$$
$$= \sum_{i=1}^{m} \alpha_i \underbrace{k(x_i, x)}_{\langle \varphi(x_i), \varphi(x) \rangle_{\mathcal{F}}}$$



$$f_\ell := \sum_{i=1}^{m} \alpha_i \varphi_\ell(x_i)$$

Function of infinitely many features expressed using m coefficients.

MMD: an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

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

Other choices for witness function class:

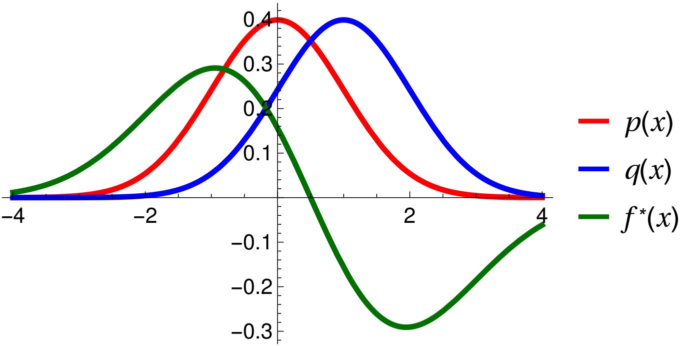
- Bounded continuous [Dudley, 2002]
- Bounded variation 1 (Kolmogorov metric) [Müller, 1997]
- 1-Lipschitz (Wasserstein distances) [Dudley, 2002]

Statistical model criticism: toy example

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

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

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)

Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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

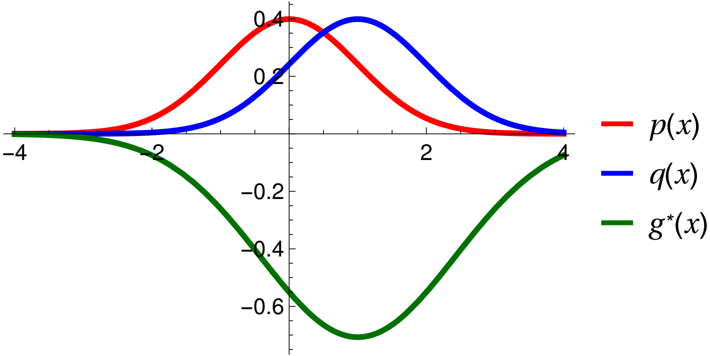
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



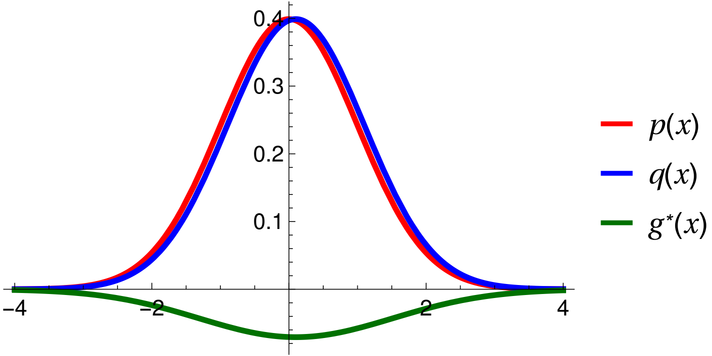
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



Computing the kernel Stein discrepancy

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

Can we define "Stein features"?

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

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

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),

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

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

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

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

Kernel Stein discrepancy: derivation

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

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

Caution: (a) requires a condition for Bochner integrability,

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

Kernel Stein discrepancy: derivation

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

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

Kernel expression:

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

Does the Bochner 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 Bochner 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

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

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

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

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

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

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

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

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

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

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

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

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, 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: $\mathcal{X} = \{1, \dots, L\}^D$ 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

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

$k_1(x, x') = \Delta_x^{-1} k(x, x')$, 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)

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

The empirical statistic:

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

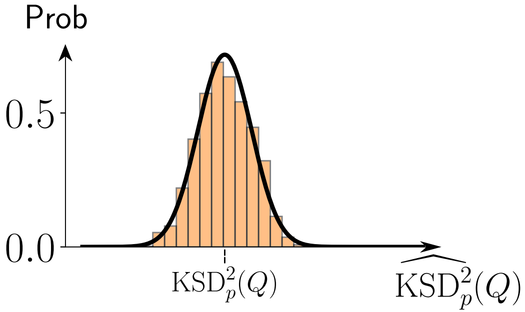
Empirical statistic, asymptotic normality for $P \neq Q$

The empirical statistic:

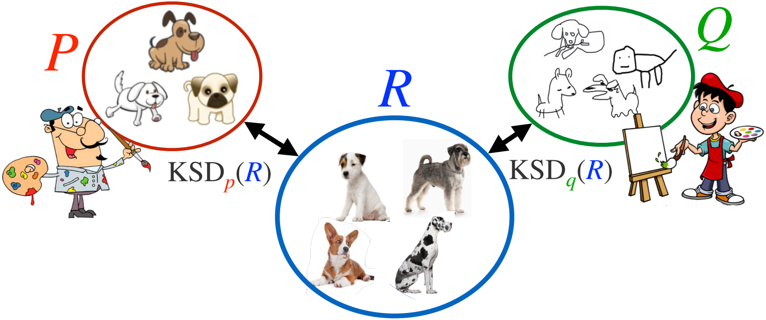
$$\widehat{\mathrm{KSD}^2_p}(Q) := \frac{1}{n(n-1)} \sum_{i \neq j} h_p(x_i, x_j).$$

Asymptotic distribution when $P \neq Q$:

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



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:

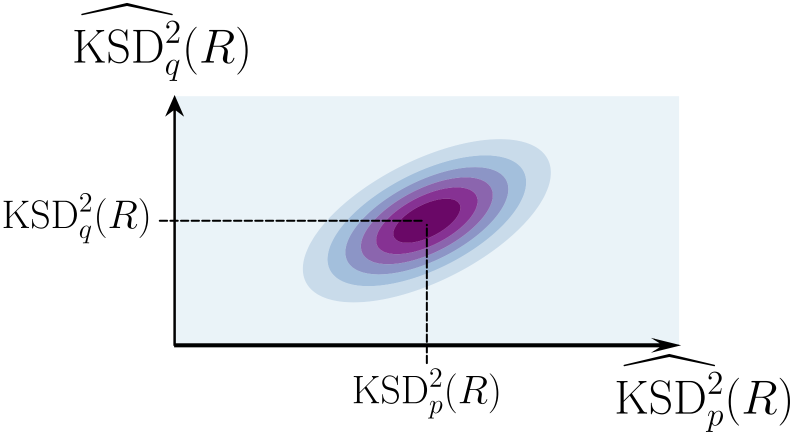
$$H_0 : \mathrm{KSD}_p(R) \leq \mathrm{KSD}_q(R) \text{ vs. } H_1 : \mathrm{KSD}_p(R) > \mathrm{KSD}_q(R)$$

$$\left(H_0 : \text{‘}P \text{ is as good as } Q\text{, or better’ vs. } H_1 : \text{‘}Q \text{ is better’} \right)$$

Relative GOF testing: joint asymptotic normality

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

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



Relative GOF testing: joint asymptotic normality

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

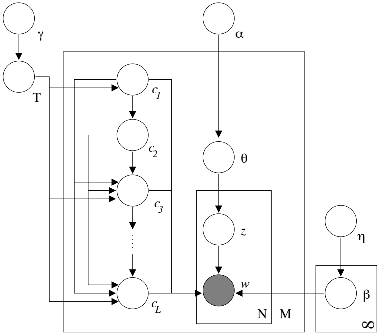
$$\sqrt{n}\begin{bmatrix} \widehat{\mathrm{KSD}^2_p}(R) - \mathrm{KSD}_p(R) \\ \widehat{\mathrm{KSD}^2_q}(R) - \mathrm{KSD}_q(R) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{h_p} & \sigma_{h_p h_q} \\ \sigma_{h_p h_q} & \sigma^2_{h_q}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

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

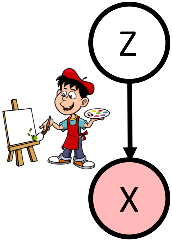


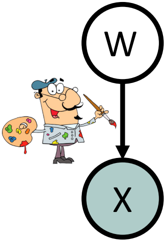
Latent variable models

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

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

Estimate KSD with approximate density:

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

What not to do

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

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

Estimate KSD with approximate density:

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

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

MCMC approximation of score function

Result we use:

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

Proof:

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

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

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:

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

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

Strategy:

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

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

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

Asymptotic distribution for relative KSD test

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

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

where

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

Fine print:

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

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

Fine print:

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

Asymptotic distribution for relative KSD test

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

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

where

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

Level-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):

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

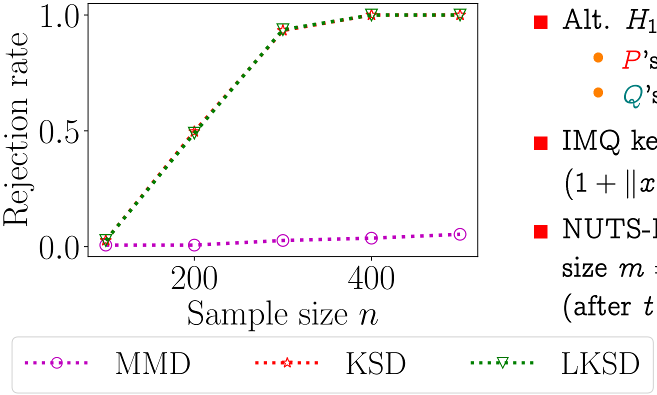
- Generate P, Q : perturb $(1,1)$-entry : $A_\delta = A + \delta E_{1,1}$

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: $k(x, x') = \left(1 + \|x - x'\|_2^2/\sigma_{\text{med}}^2\right)^{-1/2}$
- NUTS-HMC with sample size $m = 500$ (after $t = 200$ steps).

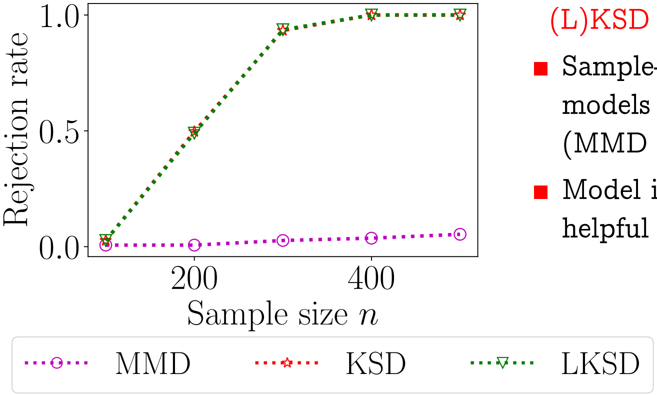
Hoffman and Gelman (JMLR 2014)

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



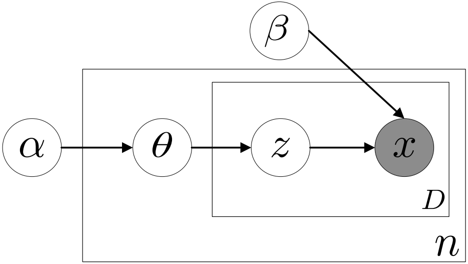
(L)KSD = higher power

- 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)

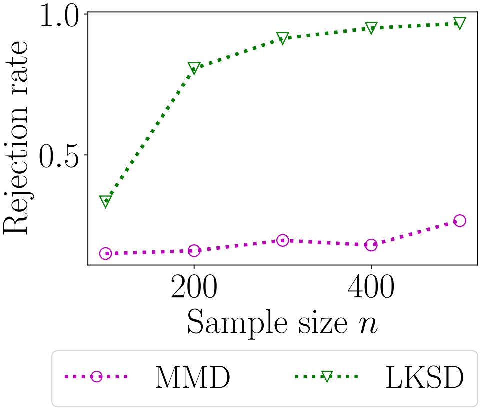


Graphical model of LDA

Blei, Ng, Jordan (JMLR 2003)

Experiment 2: topic models for arXiv articles

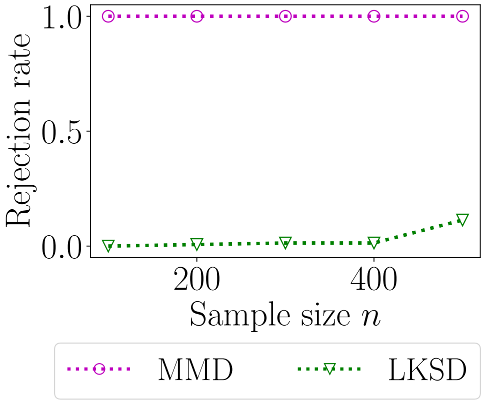
- Data R : arXiv articles from category stat.TH (stat theory) :
- Models P, Q : LDAs trained on articles from different categories (100 topics)
  - P : math.PR (math probability theory)
  - Q : stat.ME (stat methodology)



- $\mathcal{X} = \{1, \ldots, L\}^D$, $D = 100$, $L = 126, 190$.
- 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)



- $\mathcal{X} = \{1, \ldots, L\}^D$, $D = 100$, $L = 208,671$.
- 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):

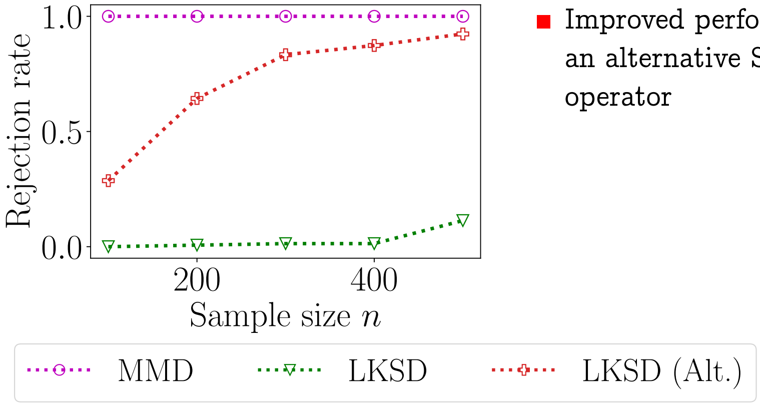
$$\mathcal{A}_p^{\mathrm{ZB}} f(x) = \sum_{\tilde{x} \in \{x+1, x-1\}} \frac{p(\tilde{x})}{p(\tilde{x}) + p(x)} \cdot \{f(\tilde{x}) - f(x)\}$$

- 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)



- Improved performance by an alternative Stein operator

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

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT

Poor MCMC hurts the test

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:

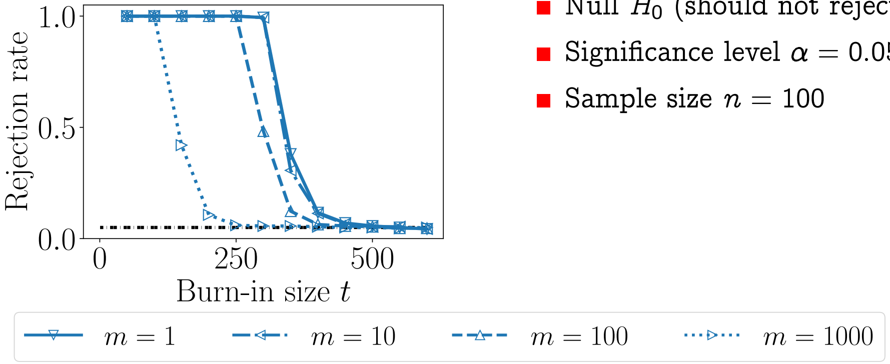
If poor, the test would reject even if P and Q are equally good

Poor MCMC hurts the test

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)



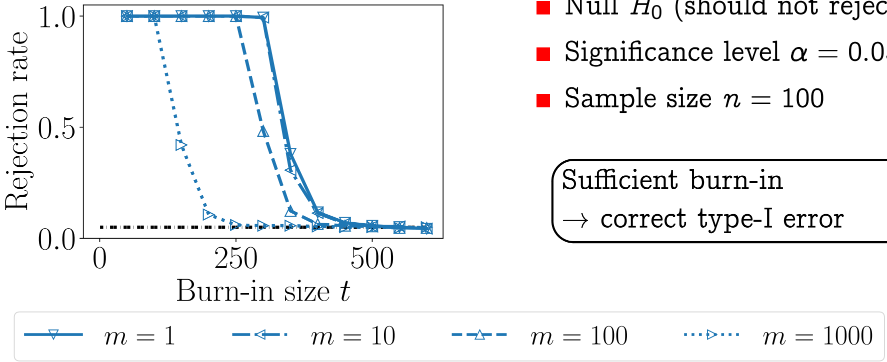
- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 100$

Poor MCMC hurts the test

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 $n = 100$

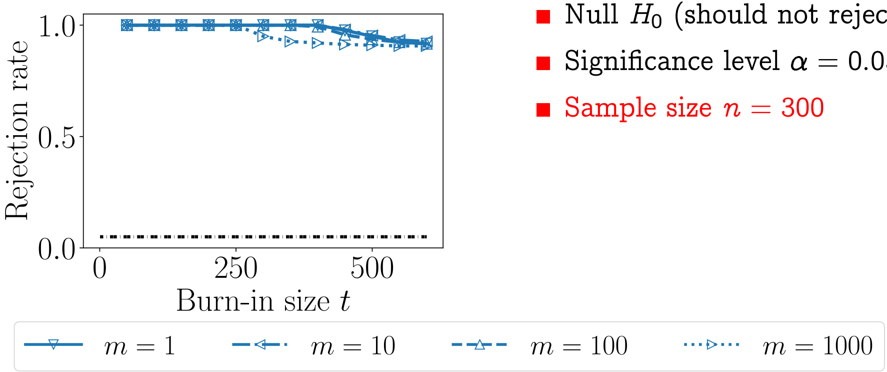
Sufficient burn-in
→ correct type-I error

Poor MCMC hurts the test

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)



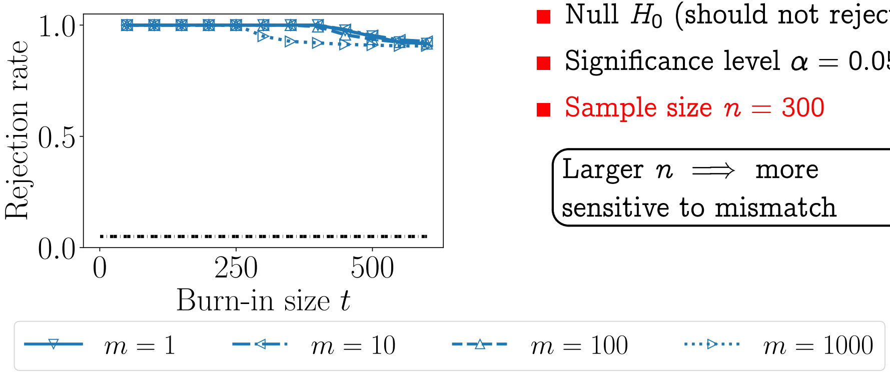
- Null H_0 (should not reject)
- Significance level $\alpha = 0.05$
- Sample size $n = 300$

Poor MCMC hurts the test

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