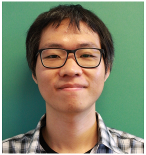
Relative Goodness-of-Fit Tests for Models with Latent Variables

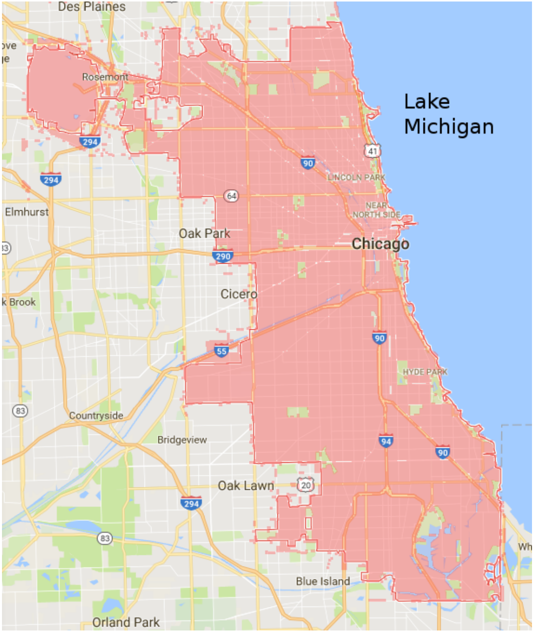
Arthur Gretton



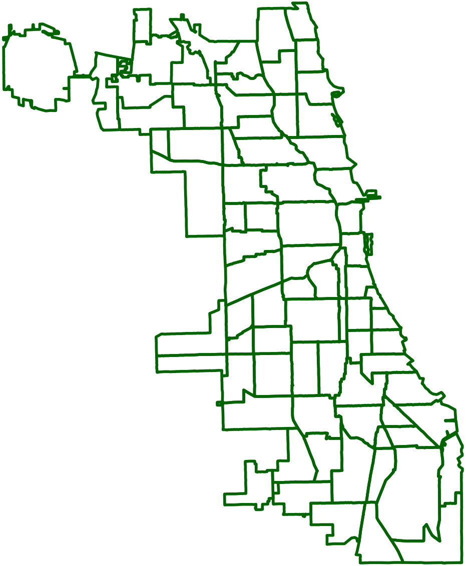
Gatsby Computational Neuroscience Unit,
University College London

June 15, 2019

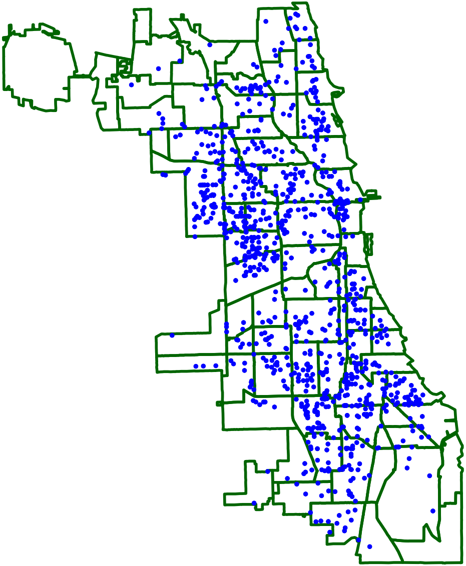
Model Criticism



Model Criticism

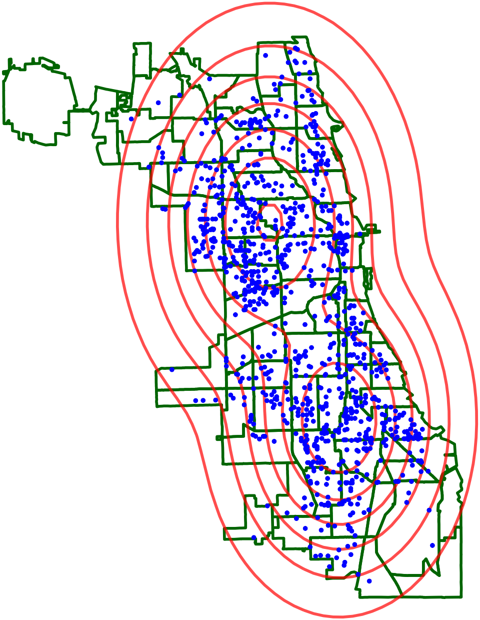


Model Criticism



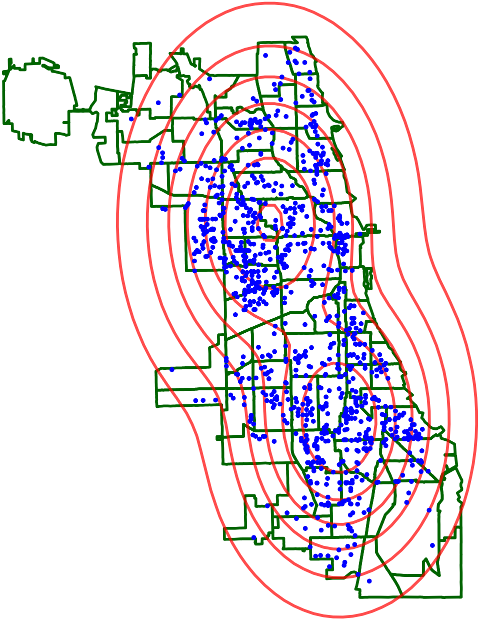
Data = robbery events in Chicago in 2016.

Model Criticism



Is this a good model?

Model Criticism



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

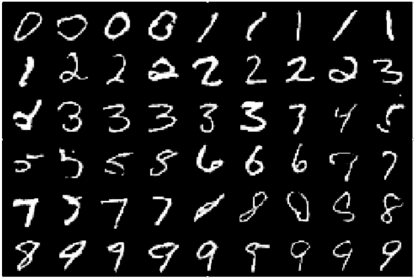
Model Criticism

"All models are wrong."

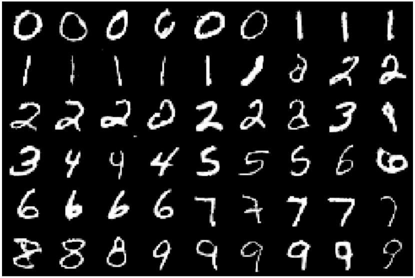
G. Box (1976)

Relative model comparison

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



Samples from GAN,
Goodfellow et al. (2014)

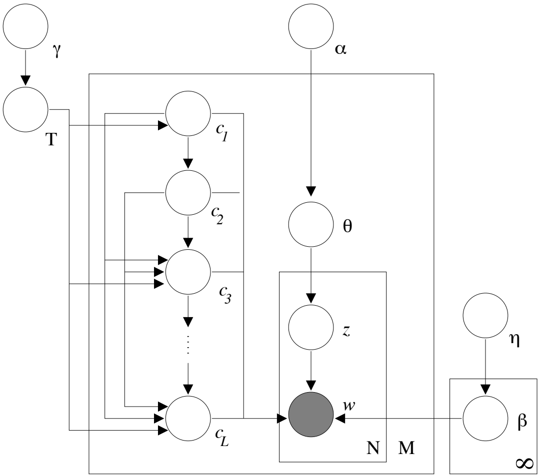


Samples from LSGAN,
Mao et al. (2017)

Which model is better?

Most interesting models have latent structure

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



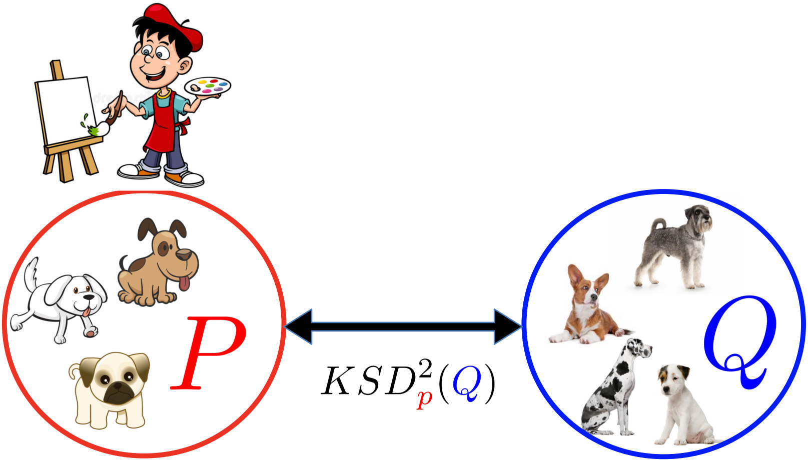
Outline

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 (new, unpublished)

Kernel Stein Discrepancy

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

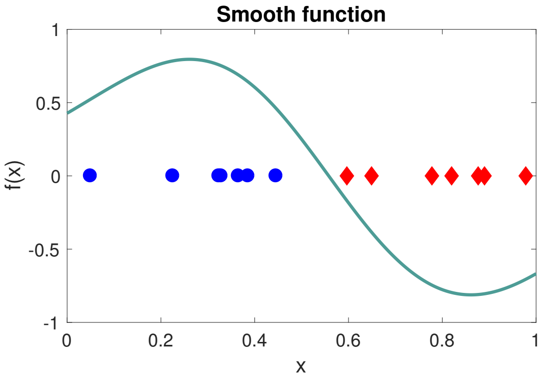


Integral probability metrics

Integral probability metric:

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

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

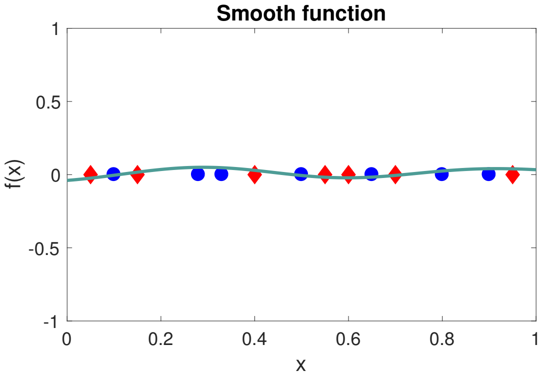


Integral probability metrics

Integral probability metric:

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

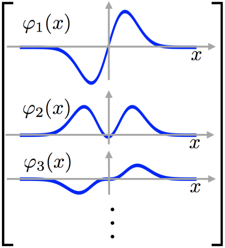
$$\mathbf{E}_Q f(Y) - \mathbf{E}_P f(X)$$



All of kernel methods

Functions are linear combinations of features:

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



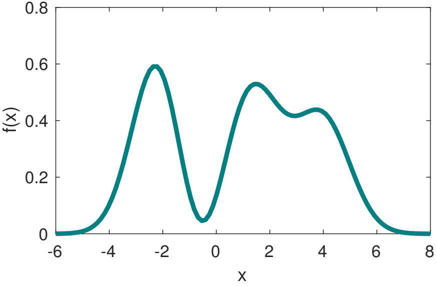
$$\|f\|_{\mathcal{F}}^2 := \sum_{i=1}^{\infty} f_i^{\,2}$$

All of kernel methods

"The kernel trick"

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

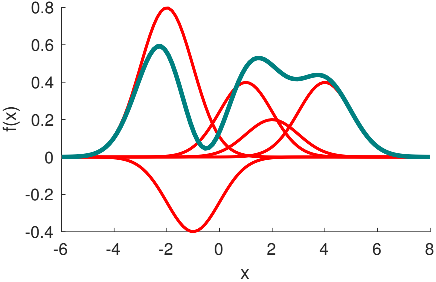
$$= \sum_{i=1}^{m} \alpha_i k(x_i, x)$$



All of kernel methods

"The kernel trick"

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x)$$
$$= \sum_{i=1}^{m} \alpha_i k(x_i, x)$$



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

Function of infinitely many features expressed using m coefficients.

MMD as an integral probability metric

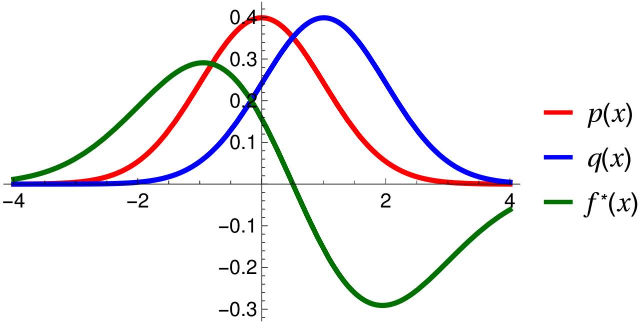
Maximum mean discrepancy: smooth function for P vs Q

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

MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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



MMD as an integral probability metric

Maximum mean discrepancy: smooth function for P vs Q

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

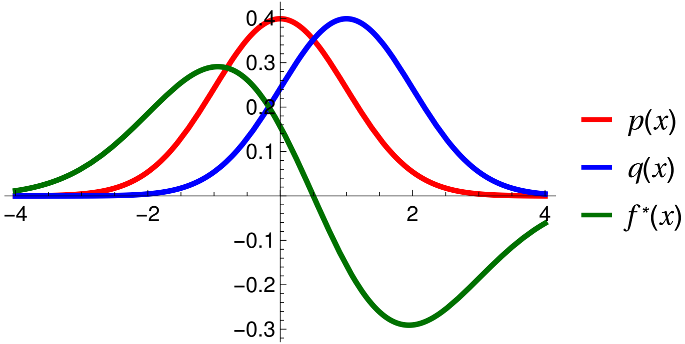
For characteristic RKHS $\mathcal{F}$, $\mathrm{MMD}(P, Q; \mathcal{F}) = 0$ iff $P = Q$

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

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



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

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

Stein idea

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

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

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

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

Then

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

subject to appropriate boundary conditions.

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)

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

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)

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

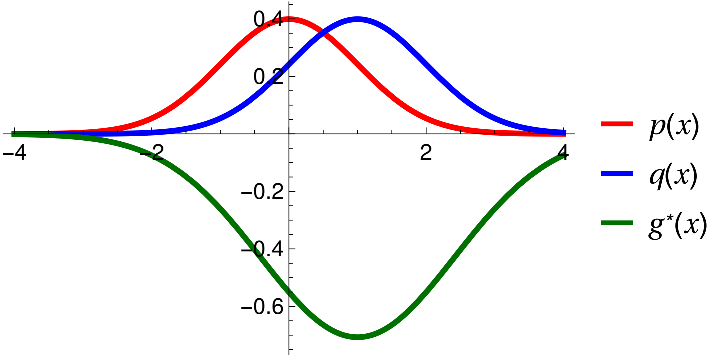
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



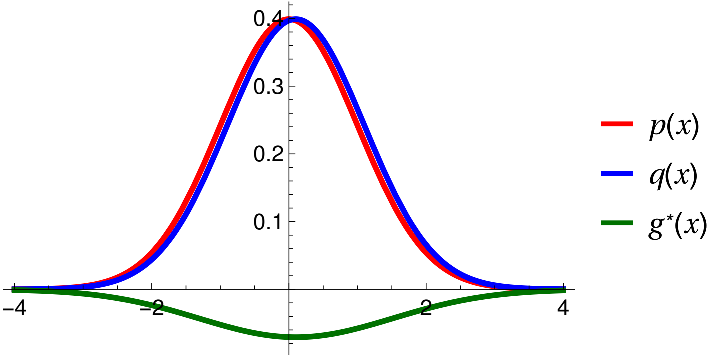
Kernel Stein Discrepancy

Stein operator

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

Kernel Stein Discrepancy (KSD)

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



Simple expression using kernels

Re-write stein operator as:

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

Can we define "Stein features"?

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

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

Simple expression using kernels

Re-write stein operator as:

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

Can we define "Stein features"?

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

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

The kernel trick for derivatives

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

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

The kernel trick for derivatives

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

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

Using kernel derivative trick in (a),

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

Kernel stein discrepancy: derivation

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

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

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

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

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

Kernel stein discrepancy: derivation

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

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

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

Kernel stein discrepancy: derivation

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

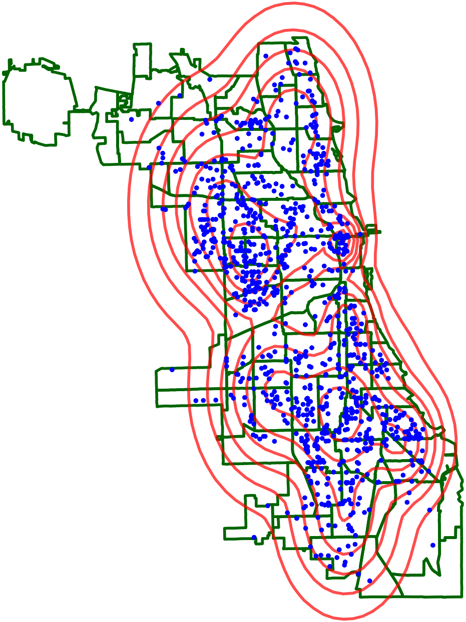
$$\begin{aligned}
\mathrm{KSD}_p(Q) &= \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathbf{E}_{x\sim q}\left([\mathcal{A}_p g](x)\right) \\
&= \sup_{\|g\|_{\mathcal{F}} \leq 1} \mathbf{E}_{x\sim q} \langle g, \xi_x \rangle_{\mathcal{F}} \\
&\underset{(a)}{=} \sup_{\|g\|_{\mathcal{F}} \leq 1} \langle g, \mathbf{E}_{x\sim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

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

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

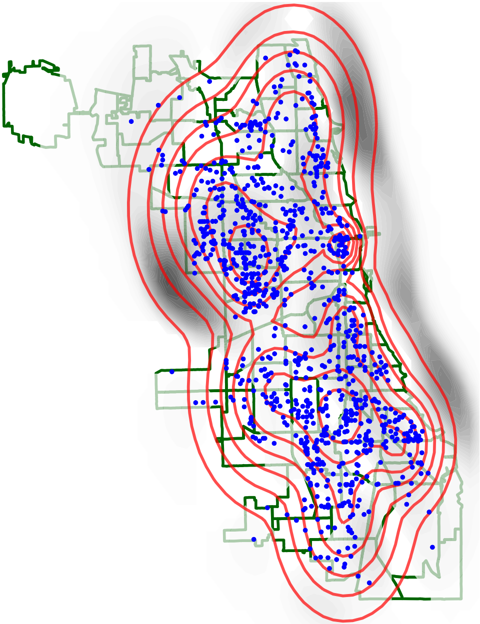
Chwialkowski, Strathmann, G., (ICML 2016) Liu, Lee, Jordan (ICML 2016)

The witness function: Chicago Crime



Model p = 10-component Gaussian mixture.

The witness function: Chicago Crime



Witness function g shows mismatch

Does the Riesz condition matter?

Consider the standard normal,

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

Then

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

If q is a Cauchy distribution, then the integral

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

is undefined.

Does the Riesz condition matter?

Consider the standard normal,

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

Then

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

If q is a Cauchy distribution, then the integral

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

is undefined.

Kernel stein discrepancy: population expression

Test statistic:

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

where

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

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

Kernel stein discrepancy: population expression

Test statistic:

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

where

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

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

Kernel stein discrepancy: population expression

Test statistic:

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

where

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

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

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

Kernel stein discrepancy: population expression

Test statistic:

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

where

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

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

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

KSD for discrete-valued variables

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

The population KSD (discrete):

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

where

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

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

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

KSD for discrete-valued variables

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

The population KSD (discrete):

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

where

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

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

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

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

KSD for discrete-valued variables

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

The population KSD (discrete):

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

where

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

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

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

The empirical statistic:

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

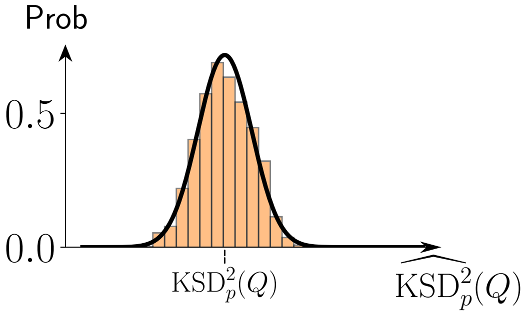
Empirical statistic, asymptotic normality for $P \neq Q$

The empirical statistic:

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

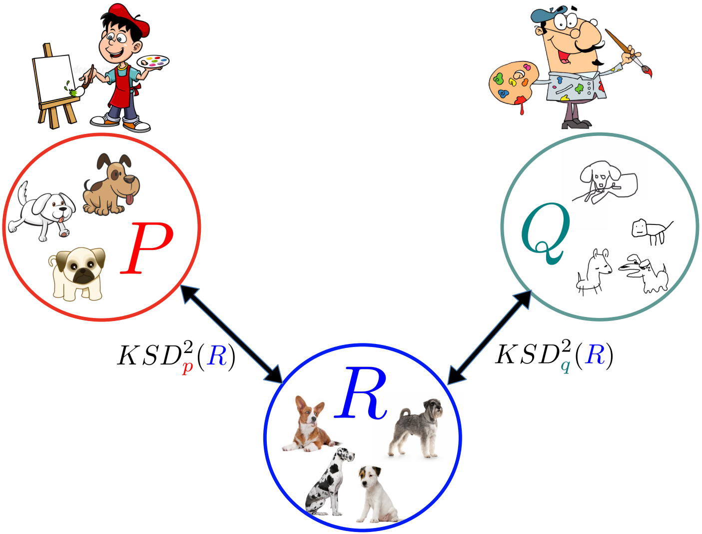
Asymptotic distribution when $P \neq Q$:

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



Relative goodness-of-fit testing

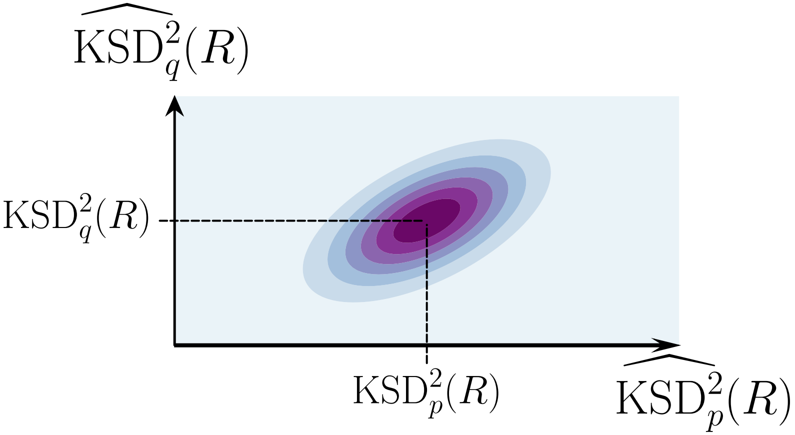
- Two generative models P and Q, data $\{x_i\}_{i=1}^n \sim R$.
- Neither model gives a perfect fit ($P \neq R$ and $Q \neq R$).



Joint asymptotic normality

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

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



Joint asymptotic normality

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

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

Difference in statistics is asymptotically normal:

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

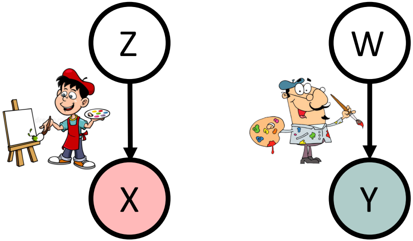
$\Longrightarrow$ a statistical test with null hypothesis $\mathrm{KSD}^2_p(R) - \mathrm{KSD}^2_q(R) \leq 0$ is straightforward.

Latent variable models

Can we compare latent variable models with KSD?

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

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



Recall multi-dimensional Stein operator:

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

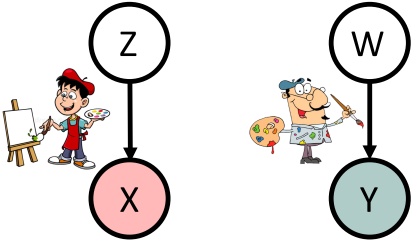
Expression (a) requires marginal $p(x)$, often intractable...

Latent variable models

Can we compare latent variable models with KSD?

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

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



Recall multi-dimensional Stein operator:

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

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

...but sampling can be straightforward!

Monte Carlo approximation

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

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

Estimate KSDs with approxiomate densities:

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

Monte Carlo approximation

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

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

Estimate KSDs with approxiomate densities:

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

Recall

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

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

Simple proof of concept

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

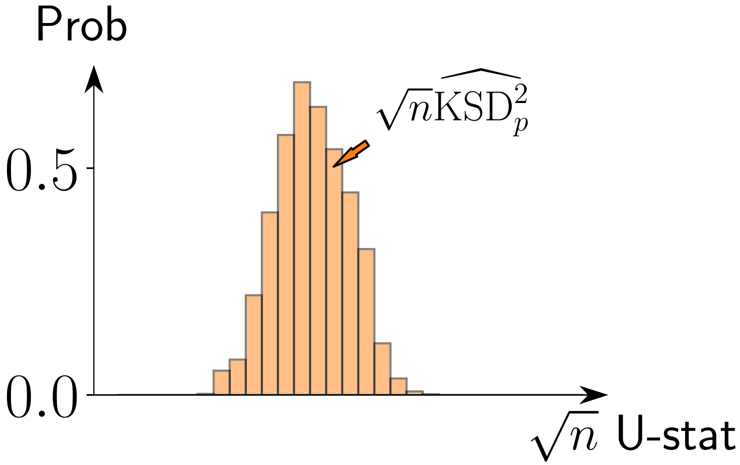
- Model: Beta-Binomial $\mathrm{BetaBinom}(\alpha, \beta)$

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

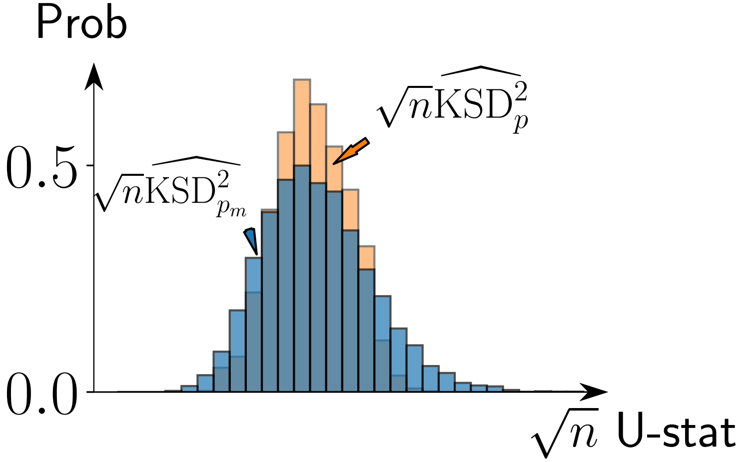
  - Latent $z \in (0, 1)$: success probability for binomial likelihood
  - Marginal $p(x)$: tractable (given by the beta function)

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

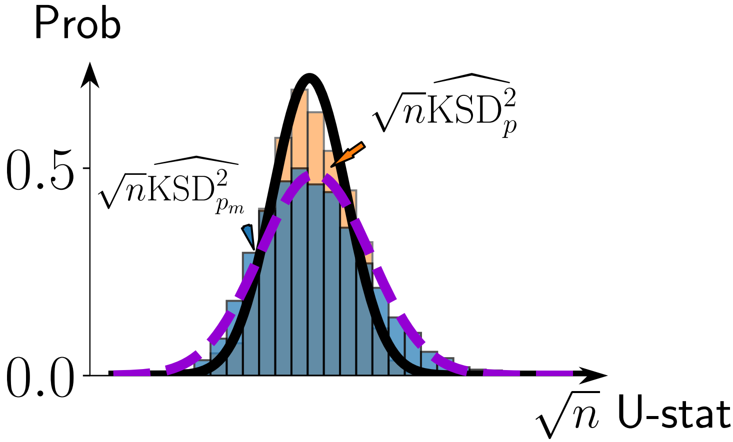
Effect of sampling the latents (Beta-binomial)



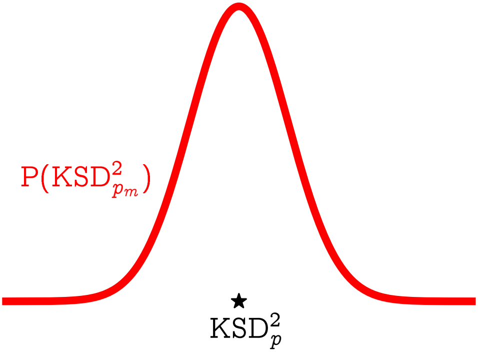
Effect of sampling the latents (Beta-binomial)



Effect of sampling the latents (Beta-binomial)

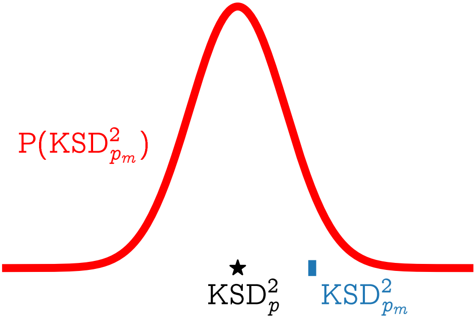


Why this happens



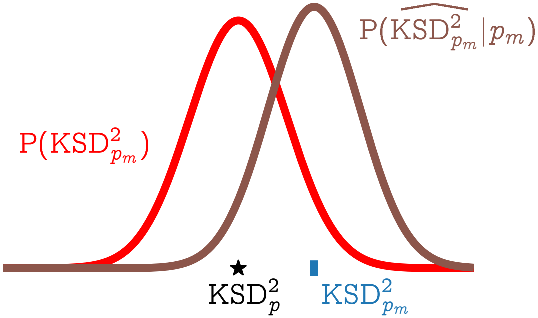
$\mathrm{KSD}^2_{p_m}(R)$ is normally distributed around $\mathrm{KSD}^2_p(R)$
(approximation error)

Why this happens



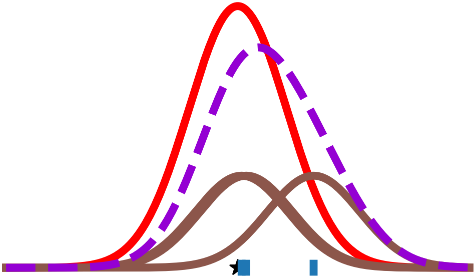
Approximation p_m gives a random draw $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



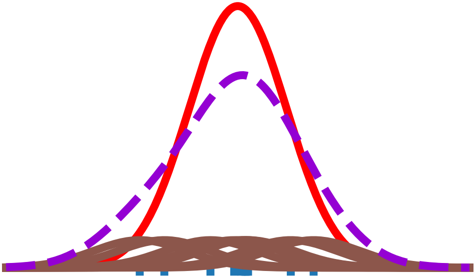
$\widehat{\mathrm{KSD}}^2_{p_m}(R)$ is normally distributed around $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



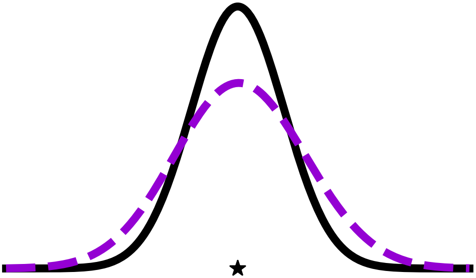
Distribution of $\widehat{\mathrm{KSD}}^2_{p_m}(R)$ is
averaged over random draws of $\mathrm{KSD}^2_{p_m}(R)$

Why this happens



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

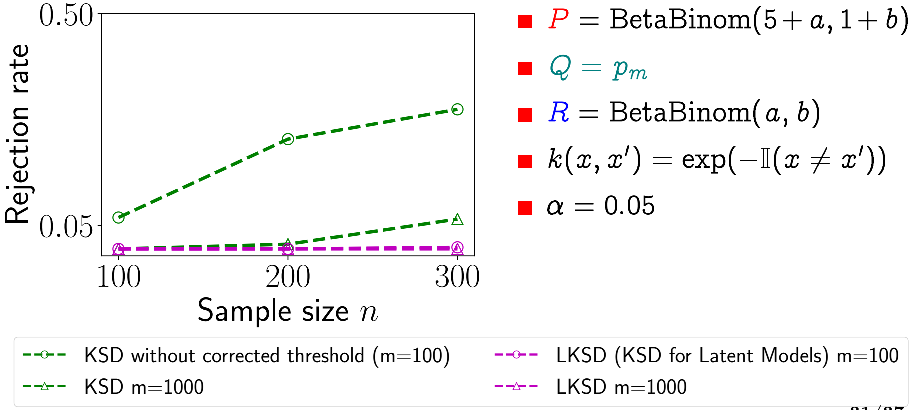
Why this happens



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

Correction for this effect

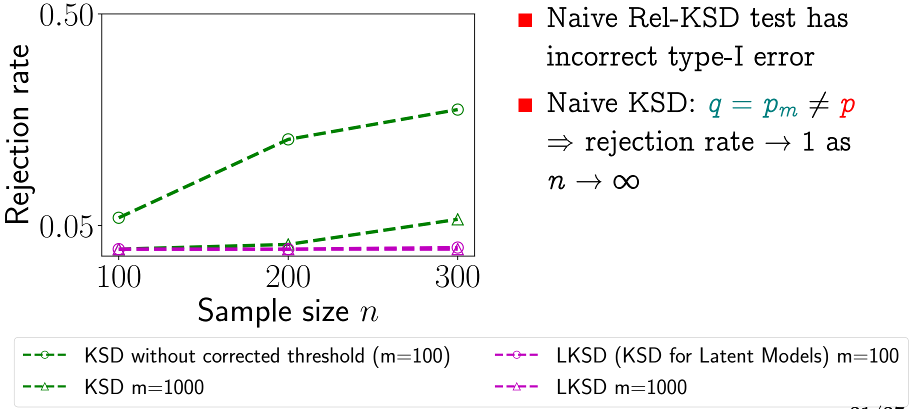
- BetaBinomial models with p vs $q = p_m$: numerical vs closed-form marginalisation.
- With correction for increased $\widehat{\mathrm{KSD}^2_{p_m}}(R)$ variance, null accepted w.p. $1-\alpha$.



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

Correction for this effect

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



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

Asymptotics for approximate KSD

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

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

The fine print:

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

Asymptotic distribution for relative KSD test

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

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

where

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

Fine print:

- $h_p(x, x') - h_q(x, x')$ has a finite third moment
- An additional technical condition (next slide)

Main theorem

Theorem (Asymptotic distribution of random kernel U-statistic)

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

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

Main theorem

Theorem (Asymptotic distribution of random kernel U-statistic)

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

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

Main theorem

Theorem (Asymptotic distribution of random kernel U-statistic)

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

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

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network SBN(W):

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

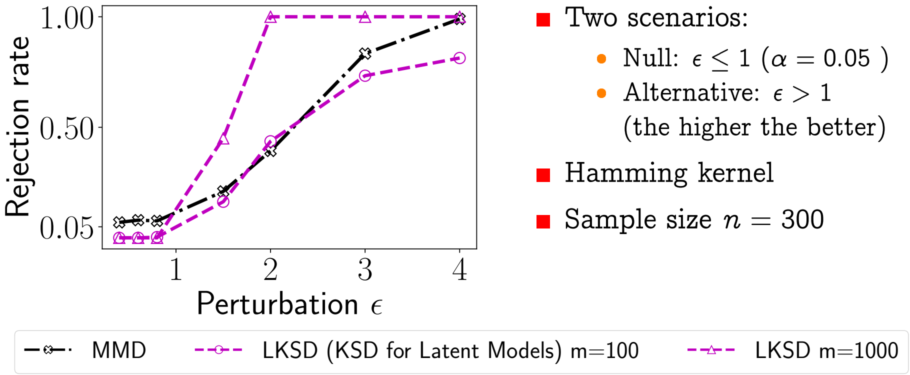
- Models: $P = \text{SBN}(W + \epsilon[\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$, $Q = \text{SBN}(W + [\mathbf{1}, \mathbf{0}, \ldots, \mathbf{0}])$
- Only the first column of weight W is perturbed by ϵ

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network SBN(W):

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

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



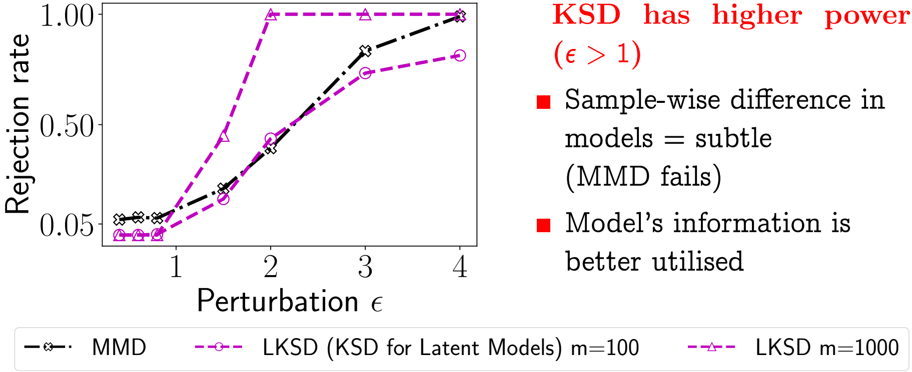
- Two scenarios:
  - Null: $\epsilon \leq 1$ ($\alpha = 0.05$)
  - Alternative: $\epsilon > 1$ (the higher the better)
- Hamming kernel
- Sample size $n = 300$

Experiment: sensitivity to model difference

- Data R = Sigmoid Belief Network SBN(W):

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

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



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

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

Questions?


THE KERNEL
BREWERY LONDON
BALTIC PORTER
7.3% ABV
THE KERNEL
EXPORT STOUT
LONDON