Interpretable Comparison of Generative Models

Wittawat Jitkrittum
Max Planck Institute for Intelligent Systems
Google Research
wittawat.com

Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, Arthur Gretton

EURECOM, Data Science Seminar
5 November 2020
Model Comparison

Which model is better? P or Q?

- Both models $P, Q$ can be wrong.
- **Goal**: pick the better one.
Outline

1 Problem setting
2 Motivations for the proposed test
3 Hypothesis testing 101
4 The Unnormalized Mean Embeddings (UME) statistic (3-sample test)
   1 Asymptotic distributions
   2 Interpretability
5 Experiments
6 The Finite Set Stein Discrepancy (FSSD) statistic (2 density models and 1 set of samples)
Problem Setting

- \( P, Q \): candidate generative models that can be sampled e.g., GANs.
- \( R \): data generating distribution (unknown).

Observe \( X_n \overset{i.i.d.}{\sim} P, Y_n \overset{i.i.d.}{\sim} Q, \) and \( Z_n \overset{i.i.d.}{\sim} R \) be three sets of samples, each of size \( n \).

\( H_0: P \) and \( Q \) model \( R \) equally well

\( H_1: Q \) models \( R \) better.

Formulate as

\[
H_0: D(P, R) - D(Q, R) = 0
\]
\[
H_1: D(P, R) - D(Q, R) > 0,
\]

for some distance \( D \).

Relative goodness-of-fit testing.

Statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \). Large, positive \( \implies Q \) is better.
Problem Setting

- \( P, Q \): candidate generative models that can be sampled e.g., GANs.
- \( R \): data generating distribution (unknown).
- Observe \( X_n \overset{i.i.d.}{\sim} P \), \( Y_n \overset{i.i.d.}{\sim} Q \), and \( Z_n \overset{i.i.d.}{\sim} R \) be three sets of samples, each of size \( n \).

\[
H_0: P \text{ and } Q \text{ model } R \text{ equally well}
\]
\[
H_1: Q \text{ models } R \text{ better.}
\]

- Formulate as

\[
H_0: D(P, R) - D(Q, R) = 0
\]
\[
H_1: D(P, R) - D(Q, R) > 0,
\]

for some distance \( D \).

- Relative goodness-of-fit testing.

- Statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \). Large, positive \( \implies Q \) is better.
Problem Setting

- $P, Q$: candidate generative models that can be sampled e.g., GANs.
- $R$: data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$H_0$: $P$ and $Q$ model $R$ equally well

$H_1$: $Q$ models $R$ better.

- Formulate as

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0,$$

for some distance $D$.

- Relative goodness-of-fit testing.

- Statistic: $\hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R)$. Large, positive $\implies Q$ is better.
Problem Setting

- $P, Q$: candidate generative models that can be sampled e.g., GANs.
- $R$: data generating distribution (unknown).
- Observe $X_n \overset{i.i.d.}{\sim} P$, $Y_n \overset{i.i.d.}{\sim} Q$, and $Z_n \overset{i.i.d.}{\sim} R$ be three sets of samples, each of size $n$.

$$H_0: P \text{ and } Q \text{ model } R \text{ equally well}$$
$$H_1: Q \text{ models } R \text{ better.}$$

- Formulate as

$$H_0: D(P, R) - D(Q, R) = 0$$
$$H_1: D(P, R) - D(Q, R) > 0,$$

for some distance $D$.

- Relative goodness-of-fit testing.
- Statistic: $\hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R)$. Large, positive $\implies Q$ is better.
Problem Setting

- \( P, Q \): candidate generative models that can be sampled e.g., GANs.
- \( R \): data generating distribution (unknown).
- Observe \( X_n \overset{i.i.d.}{\sim} P \), \( Y_n \overset{i.i.d.}{\sim} Q \), and \( Z_n \overset{i.i.d.}{\sim} R \) be three sets of samples, each of size \( n \).

\[
H_0: \text{P and Q model R equally well} \\
H_1: \text{Q models R better.}
\]

- Formulate as

\[
H_0: D(P, R) - D(Q, R) = 0 \\
H_1: D(P, R) - D(Q, R) > 0,
\]

for some distance \( D \).
- Relative goodness-of-fit testing.
- Statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \). Large, positive \( \implies \) \( Q \) is better.
Motivations

A common approach:
Compare \( \hat{D}(P, R) \) and \( \hat{D}(Q, R) \) estimated from samples (e.g., FID). If \( \hat{D}(Q, R) < \hat{D}(P, R) \), conclude that \( Q \) is better than \( P \).

Problems:

1. Noisy decision. \( \hat{D} \) is random. \( \rightarrow \) Statistical testing accounts for this.
2. Not interpretable. A scalar \( \hat{D} \) is not informative enough.

\[
Q = \text{LSGAN} \quad [\text{Mao et al., 2017}] \quad P = \text{GAN} \quad [\text{Goodfellow et al., 2014}]
\]

- 1’s from \( Q \) are better. But 3’s from \( P \) are better.
- Our interpretable test can output this information.
Motivations

A common approach:
Compare \( \hat{D}(P, R) \) and \( \hat{D}(Q, R) \) estimated from samples (e.g., FID).
If \( \hat{D}(Q, R) < \hat{D}(P, R) \), conclude that \( Q \) is better than \( P \).

Problems:

1. Noisy decision. \( \hat{D} \) is random. → Statistical testing accounts for this.
2. Not interpretable. A scalar \( \hat{D} \) is not informative enough.

\[ Q = \text{LSGAN} \; [\text{Mao et al., 2017}] \quad P = \text{GAN} \; [\text{Goodfellow et al., 2014}] \]

- 1’s from \( Q \) are better. But 3’s from \( P \) are better.
- Our interpretable test can output this information.
Motivations

A common approach:
Compare $\hat{D}(P, R)$ and $\hat{D}(Q, R)$ estimated from samples (e.g., FID).
If $\hat{D}(Q, R) < \hat{D}(P, R)$, conclude that $Q$ is better than $P$.

Problems:

1. Noisy decision. $\hat{D}$ is random. → Statistical testing accounts for this.
2. Not interpretable. A scalar $\hat{D}$ is not informative enough.

$Q = \text{LSGAN}$ [Mao et al., 2017] $P = \text{GAN}$ [Goodfellow et al., 2014]

- 1’s from $Q$ are better. But 3’s from $P$ are better.
- Our interpretable test can output this information.
Motivations

**A common approach:**
Compare \( \hat{D}(P, R) \) and \( \hat{D}(Q, R) \) estimated from samples (e.g., FID).
If \( \hat{D}(Q, R) < \hat{D}(P, R) \), conclude that \( Q \) is better than \( P \).

**Problems:**

1. Noisy decision. \( \hat{D} \) is random. → Statistical testing accounts for this.
2. Not interpretable. A scalar \( \hat{D} \) is not informative enough.

\[
\begin{align*}
Q &= \text{LSGAN} \ [\text{Mao et al., 2017}] \\
P &= \text{GAN} \ [\text{Goodfellow et al., 2014}]
\end{align*}
\]

- 1’s from \( Q \) are better. But 3’s from \( P \) are better.
- Our interpretable test can output this information.
Review: Hypothesis Testing

\[ H_0 : D(P, R) - D(Q, R) = 0 \]
\[ H_1 : D(P, R) - D(Q, R) > 0. \]

Test statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \)
Review: Hypothesis Testing

\[ H_0: D(P, R) - D(Q, R) = 0 \]
\[ H_1: D(P, R) - D(Q, R) > 0. \]

Test statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \)

\[ p_{H_0} = \text{distribution of } \hat{S}_n \text{ when } H_0 \text{ is true.} \]
Review: Hypothesis Testing

\[ H_0: D(P, R) - D(Q, R) = 0 \]
\[ H_1: D(P, R) - D(Q, R) > 0. \]

Test statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \)

Null distribution \( p_{H_0}(\hat{S}_n) = \) distribution of \( \hat{S}_n \) when \( H_0 \) is true.
Review: Hypothesis Testing

\[ H_0: D(P, R) - D(Q, R) = 0 \]
\[ H_1: D(P, R) - D(Q, R) > 0. \]

Test statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \)

- Null distribution \( p_{H_0}(\hat{S}_n) \) = distribution of \( \hat{S}_n \) when \( H_0 \) is true.
- \( T_\alpha = (1 - \alpha) \)-quantile of \( p_{H_0} \). Need to know \( p_{H_0} \).
Review: Hypothesis Testing

\[ H_0: D(P, R) - D(Q, R) = 0 \]
\[ H_1: D(P, R) - D(Q, R) > 0. \]

Test statistic: \[ \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \]

- Null distribution \( p_{H_0}(\hat{S}_n) = \) distribution of \( \hat{S}_n \) when \( H_0 \) is true.
- \( T_\alpha = (1 - \alpha) \)-quantile of \( p_{H_0} \). Need to know \( p_{H_0} \).
Review: Hypothesis Testing

$H_0: D(P, R) - D(Q, R) = 0$

$H_1: D(P, R) - D(Q, R) > 0.$

Test statistic: $\hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R)$

- Null distribution $p_{H_0}(\hat{S}_n) = \text{distribution of } \hat{S}_n \text{ when } H_0 \text{ is true.}$
- $T_\alpha = (1 - \alpha)$-quantile of $p_{H_0}$. Need to know $p_{H_0}$.
- Test: Reject $H_0$ when $\hat{S}_n > T_\alpha$. False rejection rate of $H_0$ is $\alpha$. 
Review: Hypothesis Testing

\( H_0: \ D(P, R) - D(Q, R) = 0 \)

\( H_1: \ D(P, R) - D(Q, R) > 0. \)

Test statistic: \( \hat{S}_n = \hat{D}(P, R) - \hat{D}(Q, R) \)

- **Null distribution** \( p_{H_0}(\hat{S}_n) = \text{distribution of } \hat{S}_n \text{ when } H_0 \text{ is true}. \)
- \( T_\alpha = (1 - \alpha)\)-quantile of \( p_{H_0}. \) Need to know \( p_{H_0}. \)
- Test: Reject \( H_0 \) when \( \hat{S}_n > T_\alpha. \) False rejection rate of \( H_0 \) is \( \alpha. \)
The Witness Function (Gretton et al., 2012)
Observe $Z_n = \{z_1, \ldots, z_n\} \sim R$

Observe $X_n = \{x_1, \ldots, x_n\} \sim P$
The Witness Function (Gretton et al., 2012)

Gaussian kernel $k$ on $z_i$

Gaussian kernel $k$ on $x_i$
The Witness Function (Gretton et al., 2012)

\[ \mu_R(v) = \mathbb{E}_{z \sim R} k(z, v) \]

\[ \mu_P(v) = \mathbb{E}_{x \sim P} k(x, v) \]

(mean embedding of \( P \))
The Witness Function (Gretton et al., 2012)

\[ \mu_R(v) = \mathbb{E}_{z \sim R} k(z, v) \]

\[ \mu_P(v) = \mathbb{E}_{x \sim P} k(x, v) \]

(mean embedding of \( P \))

\[ \text{witness}(v) = \mu_R(v) - \mu_P(v) \]
The Witness Function (Gretton et al., 2012)

\[
\text{MMD}(P, R) = \|\text{witness}\|_{\text{RKHS}}
\]
The Unnormalized Mean Embeddings Statistic (Chwialkowski et al., 2015)

$$\mu_R(v) = \mathbb{E}_{z \sim R} k(y, v)$$

$$\mu_P(v) = \mathbb{E}_{x \sim P} k(x, v)$$

(mean embedding of $P$)

$$\text{witness}(v) = \mu_R(v) - \mu_P(v)$$
The Unnormalized Mean Embeddings Statistic (Chwialkowski et al., 2015)

\[ \text{witness}^2(v) = (\mu_R(v) - \mu_P(v))^2 \]
The Unnormalized Mean Embeddings Statistic (Chwialkowski et al., 2015)

\[ \text{witness}^2(v) = (\mu_R(v) - \mu_P(v))^2 \]

Given \( J \) test locations \( V := \{v_j\}_{j=1}^J \) (\( V \) gives interpretability later),

\[ \text{UME}^2_V(P, R) = \frac{1}{J} \sum_{j=1}^{J} \text{witness}^2(v_j) = U^2_P. \]

\( \text{UME}^2_V \) will be \( D \) for model comparison.
The Unnormalized Mean Embeddings (UME) Statistic

\[ \text{UME}_V^2(P, R) = U_P^2 = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2. \]

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

Assume

1. Kernel \( k \) is real analytic, integrable, and characteristic;
2. \( V \) is drawn from \( \eta \), a distribution with a density.

Then, for any \( J > 0 \), any \( P \) and \( R \),

\[ \text{UME}_V^2(P, R) = 0 \text{ iff } P = R, \]

\( \eta \)-almost surely.

- **Key**: Evaluating \( \text{witness}^2(v) \) is enough to detect the difference (in theory).
- Runtime complexity: \( \mathcal{O}(Jn) \). \( J \) is small.
The Unnormalized Mean Embeddings (UME) Statistic

\[
UME^2_V(P, R) = U^2_P = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2.
\]

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

Assume

1. Kernel \( k \) is real analytic, integrable, and characteristic;
2. \( V \) is drawn from \( \eta \), a distribution with a density.

Then, for any \( J > 0 \), any \( P \) and \( R \),

\[
UME^2_V(P, R) = 0 \text{ iff } P = R,
\]

\( \eta \)-almost surely.

- **Key**: Evaluating \( \text{witness}^2(v) \) is enough to detect the difference (in theory).
- Runtime complexity: \( \mathcal{O}(Jn) \). \( J \) is small.
The Unnormalized Mean Embeddings (UME) Statistic

\[ \text{UME}_V^2(P, R) = U_P^2 = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2. \]

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

Assume

1. Kernel \( k \) is real analytic, integrable, and characteristic;
2. \( V \) is drawn from \( \eta \), a distribution with a density.

Then, for any \( J > 0 \), any \( P \) and \( R \),

\[ \text{UME}_V^2(P, R) = 0 \text{ iff } P = R, \]

\( \eta \)-almost surely.

- **Key**: Evaluating witness\(^2(\nu)\) is enough to detect the difference (in theory).
- Runtime complexity: \( \mathcal{O}(Jn) \). \( J \) is small.
The Unnormalized Mean Embeddings (UME) Statistic

\[
UME_V^2(P, R) = U_P^2 = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2.
\]

Proposition (Chwialkowski et al., 2015, Jitkrittum et al., 2016)

Assume

1. Kernel \( k \) is real analytic, integrable, and characteristic;
2. \( V \) is drawn from \( \eta \), a distribution with a density.

Then, for any \( J > 0 \), any \( P \) and \( R \),

\[
UME_V^2(P, R) = 0 \text{ iff } P = R,
\]

\( \eta \)-almost surely.

- **Key**: Evaluating \( \text{witness}^2(v) \) is enough to detect the difference (in theory).
- Runtime complexity: \( \mathcal{O}(Jn) \). \( J \) is small.
Asymptotic Distribution of $\text{UME}_V^2(P, R) = \hat{U}_P^2$

Proposition (Asymptotic distribution of $\hat{U}_P^2$)

If $P \neq R$, for any $V$, as $n \to \infty$

$$\sqrt{n} \left[ \text{UME}_V^2(P, R) - \text{UME}_V^2(P, R) \right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

where $\zeta_P^2 := (\psi_P - \psi^R)^\top (C_P + C^R)(\psi_P - \psi^R) > 0$.

- Let $\psi_P := \mathbb{E}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^J$. ~ Mean of the features.
- Let $C_P := \text{cov}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^{J \times J}$. ~Covariance of the features.
- Define $\psi_V(y) := \frac{1}{\sqrt{J}} (k(y, v_1), \ldots, k(y, v_J))^\top \in \mathbb{R}^J$.

Main point: When $P \neq R$, $\text{UME}_V^2(P, R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\hat{S}_n = \text{UME}_V^2(P, R) - \text{UME}_V^2(Q, R)$ which is ...?
Asymptotic Distribution of $\hat{\text{UME}}_V^2(P, R) = \hat{U}_P^2$

Proposition (Asymptotic distribution of $\hat{U}_P^2$)

If $P \neq R$, for any $V$, as $n \to \infty$

$$\sqrt{n} \left[ \text{UME}_V^2(P, R) - \text{UME}_V^2(P, R) \right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

where $\zeta_P^2 := (\psi^P - \psi^R)^\top(C^P + C^R)(\psi^P - \psi^R) > 0$.

- Let $\psi^P := \mathbb{E}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^J$. ~ Mean of the features.
- Let $C^P := \text{cov}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^{J \times J}$. ~Covariance of the features.
- Define $\psi_V(y) := \frac{1}{\sqrt{J}}(k(y, v_1), \ldots, k(y, v_J))^\top \in \mathbb{R}^J$.

Main point: When $P \neq R$, $\hat{\text{UME}}_V^2(P, R)$ is asymptotically normally distributed. Simple.

But we will need the distribution of $\hat{S}_n = \hat{\text{UME}}_V^2(P, R) - \hat{\text{UME}}_V^2(Q, R)$ which is ... ?
Asymptotic Distribution of $\widehat{\text{UME}}_V^2(P, R) = \widehat{U}_P^2$

Proposition (Asymptotic distribution of $\widehat{U}_P^2$)

If $P \neq R$, for any $V$, as $n \to \infty$

$$\sqrt{n} \left[ \widehat{\text{UME}}_V^2(P, R) - \text{UME}_V^2(P, R) \right] \xrightarrow{d} \mathcal{N}(0, 4\zeta_P^2),$$

where $\zeta_P^2 := (\psi^P - \psi^R)^\top (C^P + C^R)(\psi^P - \psi^R) > 0$.

- Let $\psi^P := \mathbb{E}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^J$. ~Mean of the features.
- Let $C^P := \text{cov}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^{J \times J}$. ~Covariance of the features.
- Define $\psi_V(y) := \frac{1}{\sqrt{J}} (k(y, v_1), \ldots, k(y, v_J))^\top \in \mathbb{R}^J$.

**Main point:** When $P \neq R$, $\widehat{\text{UME}}_V^2(P, R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\widehat{S}_n = \widehat{\text{UME}}_V^2(P, R) - \widehat{\text{UME}}_V^2(Q, R)$ which is ...?
Asymptotic Distribution of $\hat{\text{UME}}_V^2(P, R) = \hat{U}_P^2$

Proposition (Asymptotic distribution of $\hat{U}_P^2$)

If $P \neq R$, for any $V$, as $n \to \infty$

$$\sqrt{n} \left[ \hat{\text{UME}}_V^2(P, R) - \text{UME}_V^2(P, R) \right] \overset{d}{\to} \mathcal{N}(0, 4\zeta_P^2),$$

where $\zeta_P^2 := (\psi^P - \psi^R)^\top (C^P + C^R)(\psi^P - \psi^R) > 0$.

- Let $\psi^P := \mathbb{E}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^J$. ~ Mean of the features.
- Let $C^P := \text{cov}_{x \sim P}[\psi_V(x)] \in \mathbb{R}^{J \times J}$. ~Covariance of the features.
- Define $\psi_V(y) := \frac{1}{\sqrt{J}} (k(y, v_1), \ldots, k(y, v_J))^\top \in \mathbb{R}^J$.

**Main point:** When $P \neq R$, $\hat{\text{UME}}_V^2(P, R)$ is asymptotically normally distributed. Simple.

- But we will need the distribution of $\hat{S}_n = \hat{\text{UME}}_V^2(P, R) - \hat{\text{UME}}_V^2(Q, R)$ which is ... ?
UME²(P, R) and UME²(Q, R) are Correlated

- Write \( U_P^2 = \text{UME}^2(P, R) \) and \( U_Q^2 = \text{UME}^2(Q, R) \).
- Let \( S := U_P^2 - U_Q^2 \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).

Proposition (Joint distribution of \( \widehat{U}_P^2 \) and \( \widehat{U}_Q^2 \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions, for any \( V \),

1. \( \sqrt{n} \left( \left( \frac{U_P^2}{U_Q^2} \right) - \left( \frac{U_P^2}{U_Q^2} \right) \right) \xrightarrow{d} \mathcal{N} \left( 0, 4 \left( \frac{\zeta_P^2}{\zeta_P^2} \right) \right) \).

2. \( \sqrt{n} \left( \hat{S}_n - S \right) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta_P^2 - 2\zeta_P^2 + \zeta_Q^2) \right) \).

So, the asymptotic null distribution is normal. Easy to get \( T_\alpha \).

- [1] → use theory of multivariate U-statistics
UME\textsubscript{V}(P, R) and UME\textsubscript{V}(Q, R) are Correlated

- Write \( U_P^2 = UME\textsubscript{V}(P, R) \) and \( U_Q^2 = UME\textsubscript{V}(Q, R) \).
- Let \( S := U_P^2 - U_Q^2 \). So \( H_0: S = 0 \) and \( H_1: S > 0 \).

Proposition (Joint distribution of \( \widehat{U}_P^2 \) and \( \widehat{U}_Q^2 \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions, for any \( V \),

\[
\sqrt{n} \left( \left( \frac{U_P^2}{U_Q^2} \right) - \left( \frac{U_P^2}{U_Q^2} \right) \right) \overset{d}{\to} \mathcal{N} \left( 0, 4 \begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} \right).
\]

\[
\sqrt{n} \left( \hat{S}_n - S \right) \overset{d}{\to} \mathcal{N} \left( 0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2) \right).
\]

So, the asymptotic null distribution is normal. Easy to get \( T_\alpha \).

- [1] \( \rightarrow \) use theory of multivariate U-statistics
- [2] \( \rightarrow \) continuous mapping theorem. Follows from [1].
UME\(^2\)_V(P, R) and UME\(^2\)_V(Q, R) are Correlated

- Write \( U^2_P = \text{UME}^2(P, R) \) and \( U^2_Q = \text{UME}^2(Q, R) \).
- Let \( S := U^2_P - U^2_Q \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).

Proposition (Joint distribution of \( \hat{U}^2_P \) and \( \hat{U}^2_Q \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions, for any \( V \),

1. \( \sqrt{n} \left( \begin{pmatrix} \hat{U}^2_P \\ \hat{U}^2_Q \end{pmatrix} - \begin{pmatrix} U^2_P \\ U^2_Q \end{pmatrix} \right) \xrightarrow{d} \mathcal{N} \left( 0, 4 \begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} \right) \).

2. \( \sqrt{n} \left( \hat{S}_n - S \right) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta^2_P - 2\zeta_{PQ} + \zeta^2_Q) \right) \).

So, the asymptotic null distribution is normal. Easy to get \( T_\alpha \).

- [1] → use theory of multivariate U-statistics
UME^2_V(P, R) and UME^2_V(Q, R) are Correlated

- Write \( U_P^2 = UME^2(P, R) \) and \( U_Q^2 = UME^2(Q, R) \).
- Let \( S := U_P^2 - U_Q^2 \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).

Proposition (Joint distribution of \( \widehat{U_P^2} \) and \( \widehat{U_Q^2} \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions, for any \( V \),

1. \( \sqrt{n} \left( \left( \frac{\widehat{U_P^2}}{U_P^2} \right) - \left( \frac{\widehat{U_Q^2}}{U_Q^2} \right) \right) \xrightarrow{d} \mathcal{N} \left( 0, 4 \begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} \right) \).

2. \( \sqrt{n} \left( \hat{S}_n - S \right) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2) \right) \).

So, the asymptotic null distribution is normal. Easy to get \( T_\alpha \).

- [1] → use theory of multivariate U-statistics
UME_2^2(P, R) and UME_2^2(Q, R) are Correlated

- Write \( U_P^2 = UME_2^2(P, R) \) and \( U_Q^2 = UME_2^2(Q, R) \).
- Let \( S := U_P^2 - U_Q^2 \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).

Proposition (Joint distribution of \( \hat{U}_P^2 \) and \( \hat{U}_Q^2 \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions, for any \( V \),

1. \( \sqrt{n} \left( \left( \begin{array}{c} \hat{U}_P^2 \\ \hat{U}_Q^2 \end{array} \right) - \left( \begin{array}{c} U_P^2 \\ U_Q^2 \end{array} \right) \right) \xrightarrow{d} \mathcal{N} \left( \begin{array}{cc} 0 & \zeta_P^2 \\ \zeta_P Q & \zeta_Q^2 \end{array} \right) \).

2. \( \sqrt{n} \left( \hat{S}_n - S \right) \xrightarrow{d} \mathcal{N} \left( 0, 4 \left( \zeta_P^2 - 2 \zeta_P Q + \zeta_Q^2 \right) \right) \).

So, the asymptotic null distribution is normal. Easy to get \( T_\alpha \).

- [1] \( \rightarrow \) use theory of multivariate U-statistics
- [2] \( \rightarrow \) continuous mapping theorem. Follows from [1].
Choose Test Locations \( V = \{v_j\}_{j=1}^J \) in Practice

- Pick \( V \) so as to maximize the test power.
- \( H_0 : U_P^2 - U_Q^2 = 0 \) vs. \( H_1 : U_P^2 - U_Q^2 > 0 \) (i.e., \( Q \) is better).
Choose Test Locations $V = \{v_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0: U_P^2 - U_Q^2 = 0$ vs. $H_1: U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

Under $H_0: U_P^2 - U_Q^2 = 0$,
Choose Test Locations $V = \{v_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

Under $H_0 : U_P^2 - U_Q^2 = 0$,
Choose Test Locations $V = \{ v_j \}_{j=1}^{J}$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

Under $H_1 : U_P^2 - U_Q^2 > 0$,
Choose Test Locations $V = \{v_j\}_{j=1}^{J}$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_p^2 - U_q^2 = 0$ vs. $H_1 : U_p^2 - U_q^2 > 0$ (i.e., $Q$ is better).

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$

Test statistic $p_{H_0}$

$T_\alpha$

$p_{H_1}$

Test statistic

-2 0 2 4

Test statistic

-2 0 2 4

Test statistic
Choose Test Locations $V = \{v_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

**Test power** = $P(\text{reject } H_0 \mid H_1 \text{ true}) = P(\text{Decide } Q \text{ better} \mid Q \text{ better})$

- Split the data into tr and te. Optimize $V$ on tr. Test on te.
Choose Test Locations $V = \{v_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0: U_P^2 - U_Q^2 = 0$ vs. $H_1: U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better} \mid Q \text{ better})$

- Split the data into $tr$ and $te$. Optimize $V$ on $tr$. Test on $te$.
- Optimized $V$ show where $Q$ is better than $P$. 

Test statistic $p_{H_0}$ $T_\alpha$ $p_{H_1}$
Choose Test Locations $V = \{v_j\}_{j=1}^J$ in Practice

- Pick $V$ so as to maximize the test power.
- $H_0 : U_P^2 - U_Q^2 = 0$ vs. $H_1 : U_P^2 - U_Q^2 > 0$ (i.e., $Q$ is better).

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\text{Decide } Q \text{ better } \mid Q \text{ better})$

Split the data into $tr$ and $te$. Optimize $V$ on $tr$. Test on $te$.

- Optimized $V$ show where $Q$ is better than $P$.
- For large $n$, $\arg \max_V \text{ power} = \arg \max_V f(V)$ where $f = \frac{\text{mean of } p_{H_1}}{\text{std of } p_{H_1}}$.

Call $f$ the power criterion.
Recall the witness function between $P$ and $R$:

$$witness_{P,R}(v) = \mathbb{E}_{x \sim P} k(x, v) - \mathbb{E}_{z \sim R} k(z, v)$$

for some positive definite kernel $k(x, v)$.

Assume only one test location $v$. Recall

$$UME_{V}^{2}(P, R) = witness_{P,R}^{2}(v) = (\mu_{P}(v) - \mu_{R}(v))^{2}$$
Rel-UME: Difference of Two Witness Functions
Power criterion($\mathbf{v}$) = $f(\mathbf{v})$ is a function such that maximizing it corresponds to maximizing the test power.

$$f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})}$$

- $f(\mathbf{v}) > 0 \implies Q$ is better in the region around $\mathbf{v}$
- $f(\mathbf{v}) < 0 \implies P$ is better in the region around $\mathbf{v}$
Rel-UME: Difference of Two Witness Functions

Power criterion \( f(\mathbf{v}) = f(\mathbf{v}) \) is a function such that maximizing it corresponds to maximizing the test power.

\[
f(\mathbf{v}) = \frac{\text{witness}^2_{P,R}(\mathbf{v}) - \text{witness}^2_{Q,R}(\mathbf{v})}{\text{standard deviation}_{P,Q,R}(\mathbf{v})}
\]

- \( f(\mathbf{v}) > 0 \implies Q \text{ is better in the region around } \mathbf{v} \)
- \( f(\mathbf{v}) < 0 \implies P \text{ is better in the region around } \mathbf{v} \)
Rel-UME: Difference of Two Witness Functions

Power criterion($v$) = $f(v)$ is a function such that maximizing it corresponds to maximizing the test power.

$$f(v) = \frac{\text{witness}_{P,R}^2(v) - \text{witness}_{Q,R}^2(v)}{\text{standard deviation}_{P,Q,R}(v)}$$

- $f(v) > 0 \implies Q$ is better in the region around $v$
- $f(v) < 0 \implies P$ is better in the region around $v$
Where Does Each GAN Do Better?

\[ Q = \text{LSGAN} \quad \text{[Mao et al., 2017]} \]

\[ P = \text{GAN} \quad \text{[Goodfellow et al., 2014]} \]

- Set \( V = 40 \) (real) images of digit \( i = 0, \ldots, 9 \).
- Evaluate power criterion with \( n = 2000 \).
- \( Q \) is better at “1” and “5”. \( P \) is slightly better at “3”. *Interpretable.*
Where Does Each GAN Do Better?

\[ Q = \text{LSGAN} \quad \text{[Mao et al., 2017]} \]

\[ P = \text{GAN} \quad \text{[Goodfellow et al., 2014]} \]

- Set \( V = 40 \) (real) images of digit \( i = 0, \ldots, 9 \).
- Evaluate power criterion with \( n = 2000 \).
- \( Q \) is better at “1” and “5”. \( P \) is slightly better at “3”. Interpretable.
Where Does Each GAN Do Better?

\( Q = \) LSGAN [Mao et al., 2017]

\( P = \) GAN [Goodfellow et al., 2014]

- Set \( V = 40 \) (real) images of digit \( i = 0, \ldots, 9 \).
- Evaluate power criterion with \( n = 2000 \).
- \( Q \) is better at “1” and “5”. \( P \) is slightly better at “3”. Interpretable.
Where Does Each GAN Do Better?

\( Q = \text{LSGAN} \) [Mao et al., 2017]

\( P = \text{GAN} \) [Goodfellow et al., 2014]

- Set \( V = 40 \) (real) images of digit \( i = 0, \ldots, 9 \).
- Evaluate power criterion with \( n = 2000 \).
- \( Q \) is better at “1” and “5”. \( P \) is slightly better at “3”. Interpretable.
Where Does Each GAN Do Better?

\[ Q = \text{LSGAN} \quad [\text{Mao et al., 2017}] \]

\[ P = \text{GAN} \quad [\text{Goodfellow et al., 2014}] \]

- Set \( V = 40 \) (real) images of digit \( i = 0, \ldots, 9 \).
- Evaluate power criterion with \( n = 2000 \).
- \( Q \) is better at “1” and “5”. \( P \) is slightly better at “3”. \textit{Interpretable}.

(Gaussian kernel on top of features from a CNN classifier.)
We thank all the reviewers for constructive comments. We will revise the paper accordingly. Recall: $J$ = number of test locations with low distinguishing power are intrinsically noisier and harder to define. Our intention was to use the same image features in comparing their captions are correct. There are typos in lines 312-314. Lines 24-25 note that the test statistic is compared to a threshold $H_0$ which cannot be provided by Rel-MMD, KID, and FID. We note that there is no easy way to control false rejection rate of $H_0$.

In null hypothesis statistical testing, the test statistic is compared to a threshold $H_0$ which cannot be provided by Rel-MMD, KID, and FID. We note that there is no easy way to control false rejection rate of $H_0$.

As noted by rev 2, a key advantage of our new linear-time tests is its ability to determine the better model even when the $q$-quantiles are very similar (Fig 4d, perturbation only slightly above 0.3).

As noted by rev 2, a key advantage of our new linear-time tests is its ability to determine the better model even when the $q$-quantiles are very similar (Fig 4d, perturbation only slightly above 0.3).

Revs 1, 3: Advantages/disadvantages.
Revs 1, 2: GAN comparison.

Figure 2c shows the top 15 test locations as sorted (true) $R$ = \{automobile, cat\}

- $P$ = \{airplane, cat\}, $Q$ = \{automobile, cat\}
- $P$, $Q$, $R$ = \{automobile, cat\}
- $P$, $Q$, $R$ = \{automobile, cat\}

- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

\textbf{Experiment on CIFAR10}\n
- $P$ = \{airplane, cat\}, $Q$ = \{automobile, cat\}
- (true) $R$ = \{automobile, cat\}

- $P$, $Q$, $R$ = \{automobile, cat\}

- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.
Experiment on CIFAR10

- \( P = \{\text{airplane, cat}\} \),
  \( Q = \{\text{automobile, cat}\} \)
- (true) \( R = \{\text{automobile, cat}\} \)
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.
**Experiment on CIFAR10**

- $P = \{\text{airplane, cat}\}$, $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

Histogram of power criterion values $f(v)$ evaluated at $v = \{\text{airplane, automobile, cat}\}$.

- All non-negative. $\implies Q$ is equally good or better than $P$ everywhere.


We thank all the reviewers for constructive comments. We will revise the paper accordingly. Recall: $J$ = number of test locations $\mathcal{L}$ with low distinguishing power are intrinsically noisier and harder to define. Our intention was to use the same image which have power criterion values close to 0, meaning that these images can be generated equally well by both models work with asymptotic relative efficiency. Sub-figures in Fig 2 and 312-314 should be:

Figure 2c shows the top 15 test locations as sorted descendingly by the criterion.

Images $\mathbf{v}$ with the lowest values of $f(\mathbf{v}) \approx 0$. $\implies P, Q$ perform equally well in these regions.

---

### Experiment on CIFAR10

- $P = \{\text{airplane, cat}\}$,
  $Q = \{\text{automobile, cat}\}$
- (true) $R = \{\text{automobile, cat}\}$
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.
Experiment on CIFAR10

- \( P = \{\text{airplane, cat}\}, \)  
  \( Q = \{\text{automobile, cat}\} \)
- (true) \( R = \{\text{automobile, cat}\} \)
- Gaussian kernel on 2048 features extracted by the Inception-v3 network at the pool3 layer.

Images \( \mathbf{v} \) with the highest values of \( f(\mathbf{v}) > 0. \) \( \implies Q \) is better than \( P \) in these regions.
Problem Setting 2

- \( p, q \): probability density functions up to the normalizer
- \( r \): unknown data generating density (unknown).
- Observe \( Z_n \sim R \) and have explicit \( p, q \).

\[ H_0: \quad p \text{ and } q \text{ model } r \text{ equally well} \]
\[ H_1: \quad q \text{ models } r \text{ better.} \]

- Formulate as

\[ H_0: \quad D(p, r) - D(q, r) = 0 \]
\[ H_1: \quad D(p, r) - D(q, r) > 0, \]

for some distance \( D \).

- Statistic: \( \hat{S}_n = \hat{D}(p, r) - \hat{D}(q, r) \). Large, positive \( \Rightarrow \) \( Q \) is better.

- Same as before except \( p, q \) are now explicit density functions. No samples.
Problem Setting 2

- \( p, q \): probability density functions up to the normalizer
- \( r \): unknown data generating density (unknown).
- Observe \( Z_n \stackrel{i.i.d.}{\sim} R \) and have explicit \( p, q \).

\[
H_0: \ p \text{ and } q \text{ model } r \text{ equally well}
\]
\[
H_1: \ q \text{ models } r \text{ better.}
\]

- Formulate as

\[
H_0: \ D(p, r) - D(q, r) = 0
\]
\[
H_1: \ D(p, r) - D(q, r) > 0,
\]

for some distance \( D \).

- Statistic: \( \hat{S}_n = \hat{D}(p, r) - \hat{D}(q, r) \). Large, positive \( \implies \) \( Q \) is better.
- Same as before except \( p, q \) are now explicit density functions. No samples.
Problem Setting 2

- $p, q$: probability density functions up to the normalizer
- $r$: unknown data generating density (unknown).
- Observe $Z_n \sim R$ and have explicit $p, q$.

\[ H_0: \quad p \text{ and } q \text{ model } r \text{ equally well} \]
\[ H_1: \quad q \text{ models } r \text{ better.} \]

- Formulate as

\[ H_0: \quad D(p, r) - D(q, r) = 0 \]
\[ H_1: \quad D(p, r) - D(q, r) > 0, \]

for some distance $D$.

- Statistic: $\hat{S}_n = \hat{D}(p, r) - \hat{D}(q, r)$. Large, positive $\Rightarrow Q$ is better.
- Same as before except $p, q$ are now explicit density functions. No samples.
Problem Setting 2

- $p, q$: probability density functions up to the normalizer
- $r$: unknown data generating density (unknown).
- Observe $Z_n \overset{i.i.d.}{\sim} R$ and have explicit $p, q$.

\[
H_0: \text{ } p \text{ and } q \text{ model } r \text{ equally well} \\
H_1: q \text{ models } r \text{ better.}
\]

- Formulate as

\[
H_0: D(p, r) - D(q, r) = 0 \\
H_1: D(p, r) - D(q, r) > 0,
\]

for some distance $D$.

- Statistic: $\hat{S}_n = \hat{D}(p, r) - \hat{D}(q, r)$. Large, positive $\implies Q$ is better.
- Same as before except $p, q$ are now explicit density functions. No samples.
Problem Setting 2

- \( p, q \): probability density functions up to the normalizer
- \( r \): unknown data generating density (unknown).
- Observe \( Z_n \overset{i.i.d.}{\sim} R \) and have explicit \( p, q \).

\[
H_0: \text{ } p \text{ and } q \text{ model } r \text{ equally well}
\]
\[
H_1: q \text{ models } r \text{ better.}
\]

- Formulate as

\[
H_0: D(p, r) - D(q, r) = 0
\]
\[
H_1: D(p, r) - D(q, r) > 0,
\]

for some distance \( D \).

- Statistic: \( \hat{S}_n = \hat{D}(p, r) - \hat{D}(q, r) \). Large, positive \( \implies Q \) is better.

- Same as before except \( p, q \) are now explicit density functions. No samples.
The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall witness $v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)]$

**Problem**: No sample from $p$. Cannot estimate $\mathbb{E}_{x \sim p}[k_v(x)]$ easily.
The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall witness \( v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)] \)

**Problem:** No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_v(x)] \) easily.

\[
(\text{Stein) witness}(v) = \mathbb{E}_{z \sim r}[T_p k_v(z)] - \mathbb{E}_{x \sim p}[T_p k_v(x)]
\]
The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall witness \( v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)] \)

**Problem**: No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_v(x)] \) easily.

(Stein) witness \( v = \mathbb{E}_{z \sim r}[T_p(z)] - \mathbb{E}_{x \sim p}[T_p(x)] \)
Recall witness \( \mathbf{v} = \mathbb{E}_{z \sim r}[k_{\mathbf{v}}(z)] - \mathbb{E}_{x \sim p}[k_{\mathbf{v}}(x)] \)

**Problem:** No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_{\mathbf{v}}(x)] \) easily.
Recall witness \( \mathbf{v} = \mathbb{E}_{z \sim r}[k_{\mathbf{v}}(z)] - \mathbb{E}_{x \sim p}[k_{\mathbf{v}}(x)] \)

**Problem**: No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_{\mathbf{v}}(x)] \) easily.

(Stein) witness \( \mathbf{v} = \mathbb{E}_{z \sim r}[k_{\mathbf{v}}(z)] - \mathbb{E}_{x \sim p}[k_{\mathbf{v}}(x)] \)

**Idea**: Define \( T_p \) such that \( \mathbb{E}_{x \sim p}(T_p k_{\mathbf{v}})(x) = 0 \), for any \( \mathbf{v} \).
Recall witness \( v \) = \( \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)] \)

**Problem**: No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_v(x)] \) easily.

(Stein) witness \( v \) = \( \mathbb{E}_{z \sim r}[ v ] \)

**Idea**: Define \( T_p \) such that \( \mathbb{E}_{x \sim p}(T_p k_v)(x) = 0 \), for any \( v \).
Recall witness ($v$) = $\mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)]$

**Problem**: No sample from $p$. Cannot estimate $\mathbb{E}_{x \sim p}[k_v(x)]$ easily.

(Stein) witness ($v$) = $\mathbb{E}_{z \sim r}[T_p k_v(z)]$

**Idea**: Define $T_p$ such that $\mathbb{E}_{x \sim p}(T_p k_v)(x) = 0$, for any $v$. 
The Finite Set Stein Discrepancy (FSSD) (NeurIPS 2017 Best Paper)

Recall witness \( v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)] \)

**Problem**: No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_v(x)] \) easily.

(Stein) witness \( (v) = \mathbb{E}_{z \sim r}[T_p k_v(z)] \)

**Idea**: Define \( T_p \) such that \( \mathbb{E}_{x \sim p}(T_p k_v)(x) = 0 \), for any \( v \).

- UME defined with this new Stein witness function is called the Finite-Set Stein Discrepancy (Jitkrittum et al., 2017).
The Finite Set Stein Discrepancy (FSSD) \(\text{(NeurIPS 2017 Best Paper)}\)

Recall witness \(v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)]\)

**Problem**: No sample from \(p\). Cannot estimate \(\mathbb{E}_{x \sim p}[k_v(x)]\) easily.

\[
(\text{Stein}) \text{ witness}(v) = \mathbb{E}_{z \sim r}[T_p k_v(z)]
\]

**Idea**: Define \(T_p\) such that \(\mathbb{E}_{x \sim p}(T_p k_v)(x) = 0\), for any \(v\).

- UME defined with this new Stein witness function is called the **Finite-Set Stein Discrepancy** (Jitkrittum et al., 2017).
- \(T_p\) is called a **Stein operator**.

\[
(T_p k_v)(z) = \frac{1}{p(z)} \frac{d}{dz} [k_v(z) p(z)],
\]

which is independent of the normalizer of \(p\).
Recall witness \( v = \mathbb{E}_{z \sim r}[k_v(z)] - \mathbb{E}_{x \sim p}[k_v(x)] \)

**Problem:** No sample from \( p \). Cannot estimate \( \mathbb{E}_{x \sim p}[k_v(x)] \) easily.

\[
(\text{Stein}) \text{ witness}(v) = \mathbb{E}_{z \sim r}[T_p k_v(z)]
\]

**Idea:** Define \( T_p \) such that \( \mathbb{E}_{x \sim p}(T_p k_v)(x) = 0 \), for any \( v \).

- UME defined with this new Stein witness function is called the **Finite-Set Stein Discrepancy** (Jitkrittum et al., 2017).
- \( T_p \) is called a **Stein operator**.

\[
(T_p k_v)(z) = \frac{1}{p(z)} \frac{d}{dz}[k_v(z) p(z)],
\]

which is independent of the normalizer of \( p \).

- Can construct **Rel-FSSD** test similarly: optimize \( V \) to show where \( Q \) is better, asymptotic normality, etc.
FSSD is a Proper Discrepancy Measure

- \( \text{FSSD}^2(p, r) = \frac{1}{dj} \sum_{j=1}^{J} \| g_{p,r}(v_j) \|^2 \) where
  \[ g_{p,r}(v) = \mathbb{E}_{z \sim r} \left[ \frac{1}{p(z)} \frac{d}{dz} [k_v(z)p(z)] \right] \] (Stein witness).

Theorem (FSSD is a discrepancy measure (Jitkrittum et al., 2017))

Main conditions:

1. **(Nice kernel)** Kernel \( k \) is \( C^0 \)-universal, and real analytic e.g., Gaussian kernel.
2. **(Vanishing boundary)** \( \lim_{\|x\| \to \infty} p(x)k_v(x) = 0 \).
3. **(Avoid “blind spots”)** Locations \( v_1, \ldots, v_J \sim \eta \) which has a density.

Then, for any \( J \geq 1, \eta \)-almost surely,

\[ \text{FSSD}^2 = 0 \iff p = r. \]

**Summary:** Evaluating the witness at random locations is sufficient to detect the discrepancy between \( p, r \).
FSSD is a Proper Discrepancy Measure

\[ \text{FSSD}^2(p, r) = \frac{1}{d^J} \sum_{j=1}^{J} \|g_{p,r}(v_j)\|^2_{2} \]

where

\[ g_{p,r}(v) = \mathbb{E}_{z \sim r} \left[ \frac{1}{p(z)} \frac{d}{dz} [k_v(z)p(z)] \right] \] (Stein witness).

Theorem (FSSD is a discrepancy measure (Jitkrittum et al., 2017))

Main conditions:

1. **(Nice kernel)** Kernel \( k \) is \( C_0 \)-universal, and **real analytic** e.g., Gaussian kernel.

2. **(Vanishing boundary)** \( \lim_{\|x\| \to \infty} p(x)k_v(x) = 0 \).

3. **(Avoid “blind spots”)** Locations \( v_1, \ldots, v_J \sim \eta \) which has a density.

Then, for any \( J \geq 1 \), \( \eta \)-almost surely,

\[ \text{FSSD}^2 = 0 \iff p = r. \]

**Summary:** Evaluating the witness at random locations is sufficient to detect the discrepancy between \( p, r \).
Unlike UME which cares about probability mass, FSSD cares about shape of density functions.

In FSSD, $p, q$ are represented by $\nabla_x \log p(x)$ and $\nabla_y \log q(y)$ (instead of samples).
Unlike UME which cares about probability mass, FSSD cares about shape of density functions.

In FSSD, $p, q$ are represented by $\nabla_x \log p(x)$ and $\nabla_y \log q(y)$ (instead of samples).
**Summary**

Propose a model comparison test: Relative UME:

- **Statistical testing**: account for randomness of the distance
- **Linear-time**: runtime complexity = $O(n)$
- **Interpretable**: tells where $Q$ is better $P$ (vice versa)

Another variant: Relative FSSD: $P, Q$ are explicit (unnormalized) density functions. No need to sample.

---

**Main reference:**

- Informative Features for Model Comparison
  W. Jitkrittum, H. Kanagawa, P. Sangkloy, J. Hays, B. Schölkopf, A. Gretton
  NeurIPS 2018
  Python code: https://github.com/wittawatj/kernel-mod

**Extension**: relative test for comparing latent-variable models.

- A Kernel Stein Test for Comparing Latent Variable Models
  H. Kanagawa, W. Jitkrittum, L. Mackey, K. Fukumizu, A. Gretton
  https://arxiv.org/abs/1907.00586
Summary

Propose a model comparison test Relative UME:

- **Statistical testing**: account for randomness of the distance
- **Linear-time**: runtime complexity = $O(n)$
- **Interpretable**: tells where $Q$ is better $P$ (vice versa)

Another variant Relative FSSD: $P, Q$ are explicit (unnormalized) density functions. No need to sample.

Main reference:

- Informative Features for Model Comparison
  W. Jitkrittum, H. Kanagawa, P. Sangkloy, J. Hays, B. Schölkopf, A. Gretton
  NeurIPS 2018

  **Python code**: [https://github.com/wittawatj/kernel-mod](https://github.com/wittawatj/kernel-mod)

Extension: relative test for comparing latent-variable models.

- A Kernel Stein Test for Comparing Latent Variable Models
  H. Kanagawa, W. Jitkrittum, L. Mackey, K. Fukumizu, A. Gretton
Questions?

Thank you
Experiment on CelebA

- Two datasets for training two models.
- Center-cropped CelebA images to $64 \times 64$ pixels.
Experiment on CelebA

Model for smiling faces (S)  Model for non-smiling faces (N)

- Trained with DCGAN. Get two models.
Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\text{FID}(P, R) > \text{FID}(Q, R)$, claim $Q$ is better.

- $RS =$ real smiling images. $RN =$ real non-smiling images.
- $RM =$ mixture of RS and RN

<table>
<thead>
<tr>
<th>Case</th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
<th>Truth</th>
<th>Rel-UME J10</th>
<th>Rel-UME J40</th>
<th>Rel-UME MMD</th>
<th>FID</th>
<th>FID diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>S</td>
<td>S</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.53</td>
<td>-0.045 ± 0.52</td>
</tr>
<tr>
<td>2.</td>
<td>RS</td>
<td>RS</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0.7</td>
<td>0.04 ± 0.19</td>
</tr>
<tr>
<td>3.</td>
<td>S</td>
<td>N</td>
<td>RN</td>
<td>Rej</td>
<td>0.57</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>5.25 ± 0.75</td>
</tr>
<tr>
<td>4.</td>
<td>S</td>
<td>N</td>
<td>RM</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-4.55 ± 0.82</td>
</tr>
</tbody>
</table>

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.
Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\text{FID}(P, R) > \text{FID}(Q, R)$, claim $Q$ is better.
- $\text{RS} = \text{real smiling images}$. $\text{RN} = \text{real non-smiling images}$.
- $\text{RM} = \text{mixture of RS and RN}$

<table>
<thead>
<tr>
<th>Case</th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
<th>Truth</th>
<th>Rel-UME</th>
<th>Rel-UME</th>
<th>FID</th>
<th>FID diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>J10</td>
<td>J40</td>
<td>MMD</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>S</td>
<td>S</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.53</td>
</tr>
<tr>
<td>2.</td>
<td>RS</td>
<td>RS</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0.7</td>
</tr>
<tr>
<td>3.</td>
<td>S</td>
<td>N</td>
<td>RN</td>
<td>Rej</td>
<td>0.57</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4.</td>
<td>S</td>
<td>N</td>
<td>RM</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.
## Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming \( Q \) is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If \( \text{FID}(P, R) > \text{FID}(Q, R) \), claim \( Q \) is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

<table>
<thead>
<tr>
<th>Case</th>
<th>( P )</th>
<th>( Q )</th>
<th>( R )</th>
<th>Truth</th>
<th>Rel-UME</th>
<th>Rel-MMD</th>
<th>FID</th>
<th>FID diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>S</td>
<td>S</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.53</td>
</tr>
<tr>
<td>2.</td>
<td>RS</td>
<td>RS</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0.7</td>
</tr>
<tr>
<td>3.</td>
<td>S</td>
<td>N</td>
<td>RN</td>
<td>Rej</td>
<td>0.57</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4.</td>
<td>S</td>
<td>N</td>
<td>RM</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- FID claims \( Q \) is better when the two models are equally good. Not account for uncertainty.
- All have high test power when \( Q \) is indeed better.
## Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming $Q$ is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If $\text{FID}(P, R) > \text{FID}(Q, R)$, claim $Q$ is better.
- $\text{RS} =$ real smiling images. $\text{RN} =$ real non-smiling images.
- $\text{RM} =$ mixture of RS and RN

<table>
<thead>
<tr>
<th>Case</th>
<th>$P$</th>
<th>$Q$</th>
<th>$R$</th>
<th>Truth</th>
<th>Rel-UME J10</th>
<th>Rel-UME J40</th>
<th>Rel-MMD</th>
<th>FID</th>
<th>FID diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>S</td>
<td>S</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.53</td>
<td>-0.045 ± 0.52</td>
</tr>
<tr>
<td>2.</td>
<td>RS</td>
<td>RS</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0.7</td>
<td>0.04 ± 0.19</td>
</tr>
<tr>
<td>3.</td>
<td>S</td>
<td>N</td>
<td>RN</td>
<td>Rej</td>
<td>0.57</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>5.25 ± 0.75</td>
</tr>
<tr>
<td>4.</td>
<td>S</td>
<td>N</td>
<td>RM</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-4.55 ± 0.82</td>
</tr>
</tbody>
</table>

- FID claims $Q$ is better when the two models are equally good. Not account for uncertainty.
- All have high test power when $Q$ is indeed better.
Experiment on CelebA

- Report avg rejection rate (e.g., rate of claiming Q is better).
- Fréchet Inception Distance (FID) (Heusel et al., 2017). Not a test. If \( \text{FID}(P, R) > \text{FID}(Q, R) \), claim Q is better.
- **RS** = real smiling images. **RN** = real non-smiling images.
- **RM** = mixture of RS and RN

<table>
<thead>
<tr>
<th>Case</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>Truth</th>
<th>Rel-UME</th>
<th>Rel-MMD</th>
<th>FID</th>
<th>FID diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>J10</td>
<td>J40</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MMD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>S</td>
<td>S</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.53</td>
</tr>
<tr>
<td>2.</td>
<td>RS</td>
<td>RS</td>
<td>RS</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.03</td>
<td>0.7</td>
</tr>
<tr>
<td>3.</td>
<td>S</td>
<td>N</td>
<td>RN</td>
<td>Rej</td>
<td>0.57</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>4.</td>
<td>S</td>
<td>N</td>
<td>RM</td>
<td>Not rej</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- FID claims Q is better when the two models are equally good. Not account for uncertainty.
- All have high test power when Q is indeed better.
Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.

$q$ is closer to $r$. So, $H_1$ is true.
Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.

$q$ is closer to $r$. So, $H_1$ is true.

Rel-MMD (Bounliphone et al., 2014) suffers from a wrong choice of Gaussian bandwidth.

Proposed Rel-UME, Rel-FSSD can optimize their parameters (maximizing test power).
Experiment: 2d Blobs

- Problem in $\mathbb{R}^2$. Difference in small scale relative to the global structure.
- $q$ is closer to $r$. So, $H_1$ is true.

Rel-MMD (Bounliphone et al., 2014) suffers from a wrong choice of Gaussian bandwidth.

Proposed Rel-UME, Rel-FSSD can optimize their parameters (maximizing test power).
Rewriting UME

- \( V := \{v_1, \ldots, v_J\} = J \) test locations

\[
\text{UME}_V^2(P, R) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2
\]
Rewriting UME

- \( V := \{v_1, \ldots, v_J\} = J \) test locations

\[
UME^2_V(P, R) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2
\]

\[
= \frac{1}{J} \left\| \begin{pmatrix} \mu_P(v_1) \\ \vdots \\ \mu_P(v_J) \end{pmatrix} - \begin{pmatrix} \mu_R(v_1) \\ \vdots \\ \mu_R(v_J) \end{pmatrix} \right\|_2^2
\]
Rewriting UME

- \( V := \{v_1, \ldots, v_J\} = J \) test locations

\[
\text{UME}^2_V(P, R) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2
\]

\[
= \frac{1}{J} \left\| \begin{pmatrix} \mu_P(v_1) \\ \\
\vdots \\ \\
\mu_P(v_J) \end{pmatrix} - \begin{pmatrix} \mu_R(v_1) \\ \\
\vdots \\ \\
\mu_R(v_J) \end{pmatrix} \right\|_2^2
\]

\[
= \frac{1}{J} \left\| \mathbb{E}_{x \sim P} \begin{pmatrix} k(x, v_1) \\ \\
\vdots \\ \\
k(x, v_J) \end{pmatrix} - \mathbb{E}_{z \sim R} \begin{pmatrix} k(z, v_1) \\ \\
\vdots \\ \\
k(z, v_J) \end{pmatrix} \right\|_2^2
\]
 Rewriting UME

- \( V := \{v_1, \ldots, v_J\} = J \) test locations

\[
\text{UME}^2_V(P, R) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2
\]

\[
= \frac{1}{J} \left\| \begin{pmatrix}
\mu_P(v_1) \\
\vdots \\
\mu_P(v_J)
\end{pmatrix} - \begin{pmatrix}
\mu_R(v_1) \\
\vdots \\
\mu_R(v_J)
\end{pmatrix} \right\|^2_2
\]

\[
= \frac{1}{J} \left\| \mathbb{E}_{x \sim P} \begin{pmatrix}
k(x, v_1) \\
\vdots \\
k(x, v_J)
\end{pmatrix} - \mathbb{E}_{z \sim R} \begin{pmatrix}
k(z, v_1) \\
\vdots \\
k(z, v_J)
\end{pmatrix} \right\|^2_2
\]

Let \( \psi_V(x) := \frac{1}{\sqrt{J}} (k(x, v_1), \ldots, k(x, v_J))^\top \in \mathbb{R}^J \). Equivalently,

\[
\text{UME}^2_V(P, R) = \left\| \mathbb{E}_{x \sim P} [\psi_V(x)] - \mathbb{E}_{z \sim R} [\psi_V(z)] \right\|^2_2.
\]
Rewriting UME

- \( V := \{v_1, \ldots, v_J\} = J \) test locations

\[
UME^2_V(P, R) = \frac{1}{J} \sum_{j=1}^{J} (\mu_P(v_j) - \mu_R(v_j))^2
\]

\[
= \frac{1}{J} \left\| \begin{pmatrix} \mu_P(v_1) \\ \vdots \\ \mu_P(v_J) \end{pmatrix} - \begin{pmatrix} \mu_R(v_1) \\ \vdots \\ \mu_R(v_J) \end{pmatrix} \right\|_2^2
\]

\[
= \frac{1}{J} \left\| \mathbb{E}_{x \sim P} \begin{pmatrix} k(x, v_1) \\ \vdots \\ k(x, v_J) \end{pmatrix} - \mathbb{E}_{z \sim R} \begin{pmatrix} k(z, v_1) \\ \vdots \\ k(z, v_J) \end{pmatrix} \right\|_2^2
\]

Let \( \psi_V(x) := \frac{1}{\sqrt{J}} (k(x, v_1), \ldots, k(x, v_J))^\top \in \mathbb{R}^J \). Equivalently,

\[
UME^2_V(P, R) = \left\| \mathbb{E}_{x \sim P}[\psi_V(x)] - \mathbb{E}_{z \sim R}[\psi_V(z)] \right\|_2^2.
\]

- Empirical \( \hat{UME}^2(P, R) = \) replace \( \mathbb{E} 's \) above with \( \frac{1}{n} \sum_{i=1}^{n} \).
UME\textsuperscript{2}(P, R) and UME\textsuperscript{2}(Q, R) are Correlated

- Write \( U_P^2 = UME\textsuperscript{2}(P, R) \) and \( U_Q^2 = UME\textsuperscript{2}(Q, R) \).
- Let \( S := U_P^2 - U_Q^2 \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).
- Let \( C_S \) := \text{cov}_{y \sim S} [\psi_V(y)] \) where \( S \in \{P, Q, R\} \).
- Let \( M := \begin{pmatrix} \psi^P_V - \psi^R_V & 0 \\ 0 & \psi^Q_W - \psi^R_W \end{pmatrix} \).
- Let \( \begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} := M^\top \begin{pmatrix} C^P_V + C^R_V & C^R_V \\ (C^R_V)^\top & C^Q_W + C^R_W \end{pmatrix} M \).

Proposition (Joint distribution of \( \hat{U}_P^2 \) and \( \hat{U}_Q^2 \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions,

1. \( \sqrt{n} \begin{pmatrix} \hat{U}_P^2 \\ \hat{U}_Q^2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, 4 \begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} \right) \);

2. \( \sqrt{n} (\hat{S}_n - S) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta^2_P - 2\zeta_{PQ} + \zeta^2_Q) \right) \).

So, asymptotic null distribution is normal. Easy to get \( T_\alpha \).
UME^2_V(P, R) and UME^2_V(Q, R) are Correlated

- Write $U_P^2 = UME^2(P, R)$ and $U_Q^2 = UME^2(Q, R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C_V^S := \text{cov}_{y \sim S}[\psi_V(y)]$ where $S \in \{P, Q, R\}$.
- Let $M := \begin{pmatrix} \psi^P_V - \psi^R_V & 0 \\ 0 & \psi^Q_W - \psi^R_W \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} := M^\top \begin{pmatrix} C_V^P + C_W^R & C_V^R \\ (C_V^R)^\top & C_W^Q + C_W^R \end{pmatrix} M$

Proposition (Joint distribution of $\hat{U}_P^2$ and $\hat{U}_Q^2$)

Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,

1. $\sqrt{n} \begin{pmatrix} \hat{U}_P^2 \\ \hat{U}_Q^2 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, 4 \begin{pmatrix} \zeta^2_P & \zeta_{PQ} \\ \zeta_{PQ} & \zeta^2_Q \end{pmatrix} \right)$;

2. $\sqrt{n} (\hat{S}_n - S) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta^2_P - 2\zeta_{PQ} + \zeta^2_Q) \right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$. 
UME^2_V(P, R) and UME^2_V(Q, R) are Correlated

- Write \( U^2_P = UME^2(P, R) \) and \( U^2_Q = UME^2(Q, R) \).
- Let \( S := U^2_P - U^2_Q \). So \( H_0 : S = 0 \) and \( H_1 : S > 0 \).
- Let \( C^S_V := \text{cov}_{y \sim S}[\psi_V(y)] \) where \( S \in \{P, Q, R\} \).
- Let \( M := \begin{pmatrix} \psi^P_V - \psi^R_V & 0 \\ 0 & \psi^Q_W - \psi^R_W \end{pmatrix} \).
- Let \( \left( \begin{array}{c} \zeta^2_P \\ \zeta^2_Q \\ \zeta^2_{PQ} \end{array} \right) := M^\top \left( \begin{array}{cc} C^P_V + C^R_V & C^R_V \\ (C^R_V)^\top & C^Q_W + C^R_W \end{array} \right) M \).

Proposition (Joint distribution of \( \hat{U}^2_P \) and \( \hat{U}^2_Q \))

Assume that \( P, Q \) and \( R \) are all distinct. Under mild conditions,

1. \( \sqrt{n} \left( \begin{array}{c} \hat{U}^2_P \\ \hat{U}^2_Q \end{array} \right) \xrightarrow{d} \mathcal{N} \left( \begin{array}{c} 0 \\ 0 \end{array} \right, \begin{pmatrix} \zeta^2_P & \zeta^2_{PQ} \\ \zeta^2_{PQ} & \zeta^2_Q \end{pmatrix} \) ;

2. \( \sqrt{n} \left( \hat{S}_n - S \right) \xrightarrow{d} \mathcal{N} \left( 0, 4(\zeta^2_P - 2\zeta_{PQ} + \zeta^2_Q) \right) \).

So, asymptotic null distribution is normal. Easy to get \( T_\alpha \).
UME^2_V(P, R) and UME^2_V(Q, R) are Correlated

- Write $U_P^2 = UME^2(P, R)$ and $U_Q^2 = UME^2(Q, R)$.
- Let $S := U_P^2 - U_Q^2$. So $H_0 : S = 0$ and $H_1 : S > 0$.
- Let $C_S^V := \text{cov}_{y \sim S}[\psi_V(y)]$ where $S \in \{P, Q, R\}$.
- Let $M := \begin{pmatrix} \psi_V^P - \psi_V^R & 0 \\ 0 & \psi_W^Q - \psi_W^R \end{pmatrix}$.
- Let $\begin{pmatrix} \zeta_P^2 \\ \zeta_{PQ} \\ \zeta_Q^2 \end{pmatrix} := M^\top \begin{pmatrix} C_V^P + C_R^R \\ (C_V^R)^\top \\ C_W^Q + C_W^R \end{pmatrix} M$

Proposition (Joint distribution of $\widehat{U}_P^2$ and $\widehat{U}_Q^2$)

Assume that $P, Q$ and $R$ are all distinct. Under mild conditions,

1. $\sqrt{n} \left( \begin{pmatrix} \widehat{U}_P^2 \\ \widehat{U}_Q^2 \end{pmatrix} - \begin{pmatrix} U_P^2 \\ U_Q^2 \end{pmatrix} \right) \overset{d}{\to} \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix} , 4 \begin{pmatrix} \zeta_P^2 & \zeta_{PQ} \\ \zeta_{PQ} & \zeta_Q^2 \end{pmatrix} \right)$;

2. $\sqrt{n} \left( \widehat{S}_n - S \right) \overset{d}{\to} \mathcal{N} \left( 0, 4(\zeta_P^2 - 2\zeta_{PQ} + \zeta_Q^2) \right)$.

So, asymptotic null distribution is normal. Easy to get $T_\alpha$. 

28/21
Experiment: Mean Shift

- Model 1: $p = \mathcal{N}([0.5, 0, \ldots, 0], I)$. Model 2: $q = \mathcal{N}([1, 0, \ldots, 0], I)$
- Data distribution $r = \mathcal{N}(0, I)$. Defined on $\mathbb{R}^{50}$.
- Set $\alpha = 0.05$. Should not reject $H_0$. 
Experiment: Mean Shift

- Model 1: $p = \mathcal{N}([0.5, 0, \ldots, 0], I)$. Model 2: $q = \mathcal{N}([1, 0, \ldots 0], I)$
- Data distribution $r = \mathcal{N}(0, I)$. Defined on $\mathbb{R}^{50}$.
- Set $\alpha = 0.05$. Should not reject $H_0$.

![Graph showing rejection rate vs. sample size for different tests (Rel-UME J1, Rel-UME J5, Rel-FSSD J1, Rel-FSSD J5, Rel-MMD).]
Experiment: Mean Shift

- Model 1: \( p = \mathcal{N}([0.5, 0, \ldots, 0], I) \). Model 2: \( q = \mathcal{N}([1, 0, \ldots, 0], I) \)
- Data distribution \( r = \mathcal{N}(0, I) \). Defined on \( \mathbb{R}^{50} \).
- Set \( \alpha = 0.05 \). Should not reject \( H_0 \).

![Graph showing time vs sample size](image)

- MMD runs in \( O(n^2) \) time.
- Proposed Rel-UME and Rel-FSSD run in \( O(n) \).
Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- \( p, q, r \) are all RBM models. \( d = 20 \) dimensions. \( n = 2000 \).
- \( g_{B,b,c}(x) := \frac{1}{Z} \sum_h \exp \left( x^\top Bh + b^\top x + c^\top h - \frac{1}{2}||x||^2 \right) \) where \( h \in \{-1, 1\}^5 \).
- Define \( r(x) := g_{B,b,c}(x) \) for some randomly drawn \( B, b, c \).
- Let \( p(x) := g_{B^p,b,c}(x) \), and \( q(x) := g_{B^q,b,c}(x) \).
- \( B^p = B \) but with \( \epsilon \) added to its first entry \( B_{1,1} \)
- \( B^q = B \) but with 0.3 added to its first entry \( B_{1,1} \)
- If \( \epsilon > 0.3 \), \( q \) is better. Should reject \( H_0 \).
Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- \( p, q, r \) are all RBM models. \( d = 20 \) dimensions. \( n = 2000 \).
- \( g_{B,b,c}(x) := \frac{1}{Z} \sum_h \exp \left( x^\top Bh + b^\top x + c^\top h - \frac{1}{2} ||x||^2 \right) \) where \( h \in \{-1, 1\}^5 \).
- Define \( r(x) := g_{B,b,c}(x) \) for some randomly drawn \( B, b, c \).
- Let \( p(x) := g_{B^p,b,c}(x) \), and \( q(x) := g_{B^q,b,c}(x) \).
- \( B^p = B \) but with \( \epsilon \) added to its first entry \( B_{1,1} \)
- \( B^q = B \) but with 0.3 added to its first entry \( B_{1,1} \)
- If \( \epsilon > 0.3 \), \( q \) is better. Should reject \( H_0 \).
Experiment: Gaussian-Bernoulli Restricted Boltzmann Machine

- $p, q, r$ are all RBM models. $d = 20$ dimensions. $n = 2000$.
- $g_{B,b,c}(x) := \frac{1}{Z} \sum_h \exp \left( x^\top B h + b^\top x + c^\top h - \frac{1}{2} ||x||^2 \right) \text{ where } h \in \{-1, 1\}^5$.
- Define $r(x) := g_{B,b,c}(x)$ for some randomly drawn $B, b, c$.
- Let $p(x) := g_{B^p,b,c}(x)$, and $q(x) := g_{B^q,b,c}(x)$.
- $B^p = B$ but with $\epsilon$ added to its first entry $B_{1,1}$
- $B^q = B$ but with 0.3 added to its first entry $B_{1,1}$
- If $\epsilon > 0.3$, $q$ is better. Should reject $H_0$. 
Models and true distribution are very close. Difficult.

FSSD has access to the density. Higher power than UME, MMD (rely on samples).
What is $T_p k_v$?

Recall Stein witness($v$) = $E_{y \sim q}(T_p k_v)(y)$ - $E_{x \sim p}(T_p k_v)(x)$
What is $T_p k_v$?

Recall Stein witness $(\mathbf{v}) = \mathbb{E}_{y \sim q}(T_p k_v)(y) - \mathbb{E}_{x \sim p}(T_p k_v)(x)$

$$(T_p k_v)(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \frac{d}{d\mathbf{x}} [k(\mathbf{x}, \mathbf{v}) p(\mathbf{x})].$$

Then, $\mathbb{E}_{x \sim p}(T_p k_v)(\mathbf{x}) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]
What is $T_p k_v$?

Recall Stein witness $v = \mathbb{E}_{y \sim q}(T_p k_v)(y) - \mathbb{E}_{x \sim p}(T_p k_v)(x)$

$$(T_p k_v)(x) = \frac{1}{p(x)} \frac{d}{dx} [k(x, v)p(x)].$$

Then, $\mathbb{E}_{x \sim p}(T_p k_v)(x) = 0$.

[Liu et al., 2016, Chwialkowski et al., 2016]
What is $T_p k_v$?

Recall Stein witness \( v \) = \( \mathbb{E}_{y \sim q}(T_p k_v)(y) - \mathbb{E}_{x \sim p}(T_p k_v)(x) \)

\[
(T_p k_v)(x) = \frac{1}{p(x)} \frac{d}{dx} [k(x, v)p(x)].
\]

Then, \( \mathbb{E}_{x \sim p}(T_p k_v)(x) = 0. \)

[Liu et al., 2016, Chwialkowski et al., 2016]

\[
\mathbb{E}_{x \sim p} [(T_p k_v)(x)] = \int_{-\infty}^{\infty} \left[ \frac{1}{p(x)} \frac{d}{dx} [k_v(x)p(x)] \right] p(x) \, dx
\]

\[
= \int_{-\infty}^{\infty} \frac{d}{dx} [k_v(x)p(x)] \, dx
\]

\[
= [k_v(x)p(x)]_{x=-\infty}^{x=\infty}
\]

\[
= 0
\]

(assume \( \lim_{|x| \to \infty} k(v, x)p(x) \))
Interpretable Distribution Features with Maximum Testing Power
Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, Arthur Gretton
NIPS 2016 (oral)
**Paper/code:** https://github.com/wittawatj/interpretable-test

A Linear-Time Kernel Goodness-of-Fit Test
Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, Arthur Gretton
NIPS 2017 (oral, best paper)
**Paper/code:** https://github.com/wittawatj/kernel-gof