# 'Which model is best?' Composite Relative goodness-of-fit testing with kernels

VMFS3

Supervisor : Dr François-Xavier Briol

University College London (UCL) - Department of Statistical Science

September 19, 2022

# Part1. Introduction

# Use Generative Model as intro

1. **What is it?** Generative Model v.s. Preditive Model
2. **When we need?** Data Augmentation, Image/vedio generation
3. **How to build?**
    - Parametric generative model: **Graphical Model** (Boltzmann machine), mixture gaussian etc...
    - Non-parametric generative models: GAN, VAE, diffusion models etc...

# Challenges of Generative Model

Two challenges of generative model:

1. **Model selection**
   - All models are wrong, but some are useful. – George E.P.Box
   - The model fits the data?
   - Which model fits better?
   - How to evaluate?
2. **Computationally expensive**
   - Large amount of data are needed.
   - Normalization constant term $Z$: $p(x) = f(x)/Z$

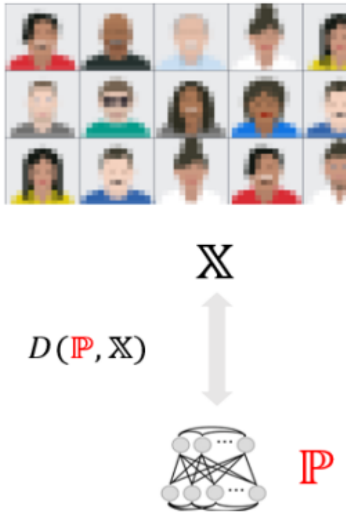
# Use hypothesis testing for model selection

Two types of hypothesis testing:

- Goodness-of-fit testing (GOF)
  - Composite Goodness-of-fit testing (C-GOF)
  - Relative Goodness-of-fit testing (R-GOF)
  - **Composite relative goodness-of-fit testing** (CR-GOF).
- Two sample test (TST)

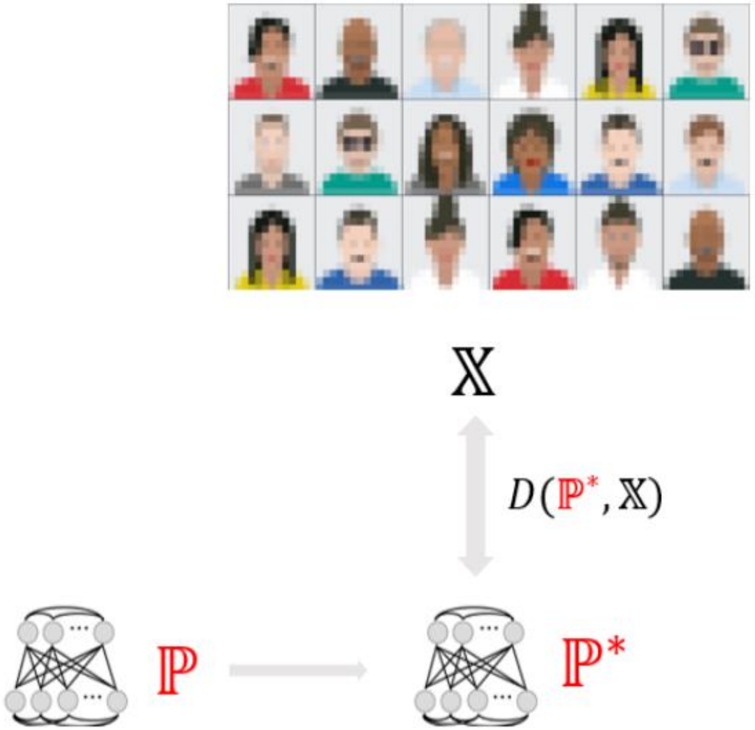
# Goodness-of-fit testing (GOF)

- Suppose $D(\mathbb{P}, \mathbb{Q})$ is a statistical discrepancy. $D(\mathbb{P}, \mathbb{Q}) \geq 0$ and $D(\mathbb{P}, \mathbb{Q}) = 0$ iff $\mathbb{P}, \mathbb{Q}$.
- Given sample $\{X_i\}_{i=1}^{n}$, $\mathbb{X}$ is its population. $\mathbb{P}_\theta$ is parametric model where $\theta \in \Theta$.
- **GOF:** determine if the model $\mathbb{P}_\theta$ fits the data $\mathbb{X}$
- $H_0 : \mathbb{P} \in \mathbb{X}, H_1 : \mathbb{P} \notin \mathbb{X}$



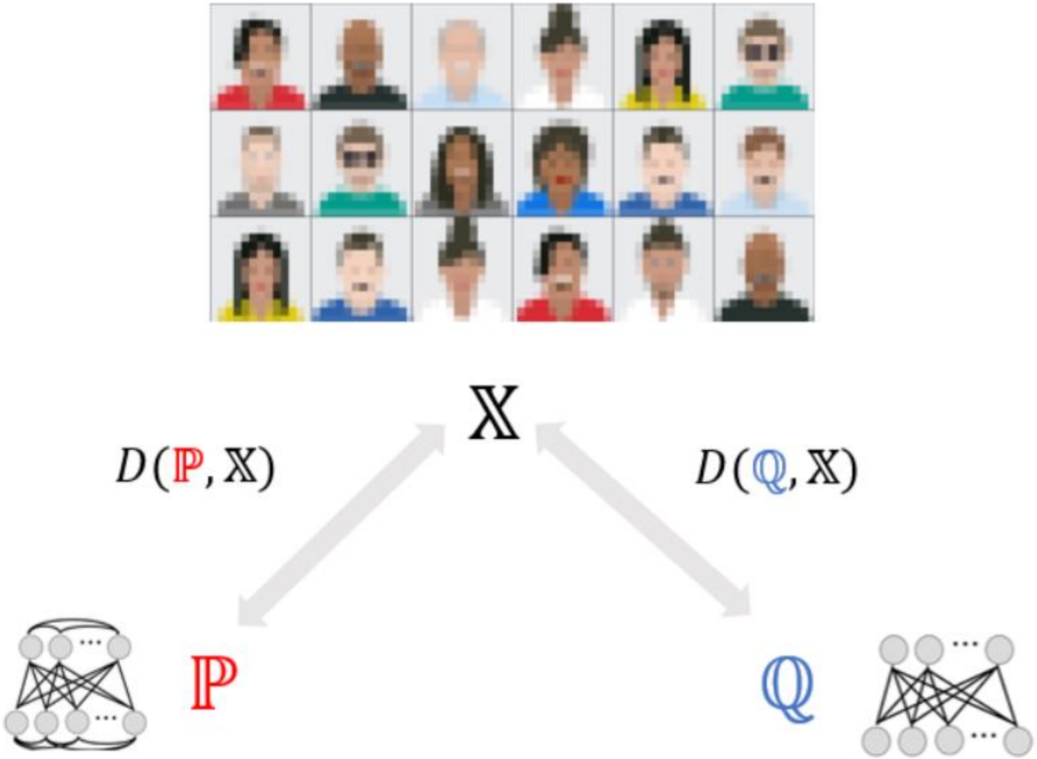
# Composite goodness-of-fit testing (C-GOF)

- **C-GOF:** determine if the model set $\{\mathbb{P}\}_{\theta\in\Theta}$ fits the data $\mathbb{X}$
- $H_0 : \{\mathbb{P}\}_{\theta\in\Theta} \in \mathbb{X}$, $H_1 : \{\mathbb{P}\}_{\theta\in\Theta} \notin \mathbb{X}$



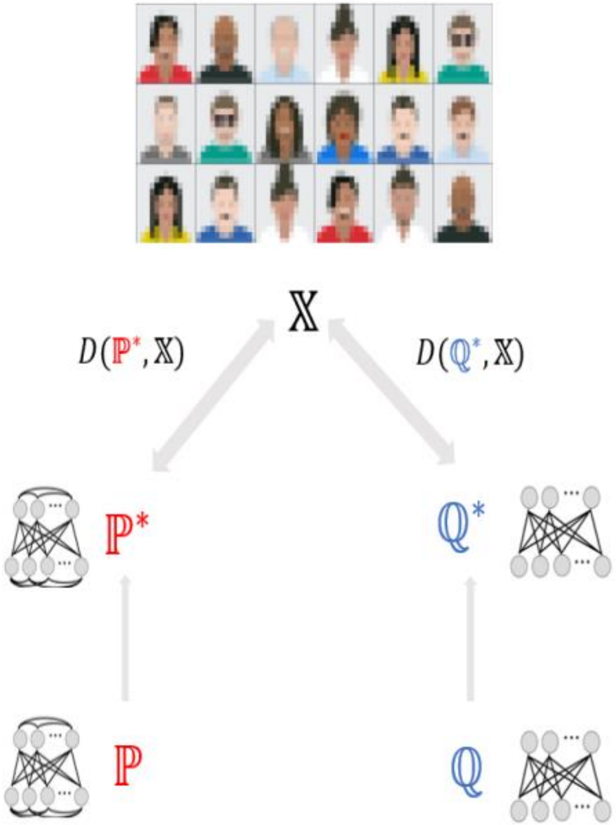
# Relative goodness-of-fit testing (R-GOF)

- **R-GOF:** determine whether model $\mathbb{P}_\theta$ or $\mathbb{Q}_\phi$ fits the data $\mathbb{X}$ better.
- $H_0 : D(\mathbb{P}_\theta, \mathbb{X}) - D(\mathbb{Q}_\phi, \mathbb{X}) \leq 0$, $H_1 : D(\mathbb{P}_\theta, \mathbb{X}) - D(\mathbb{Q}_\phi, \mathbb{X}) > 0$



# Composite Relative goodness-of-fit testing

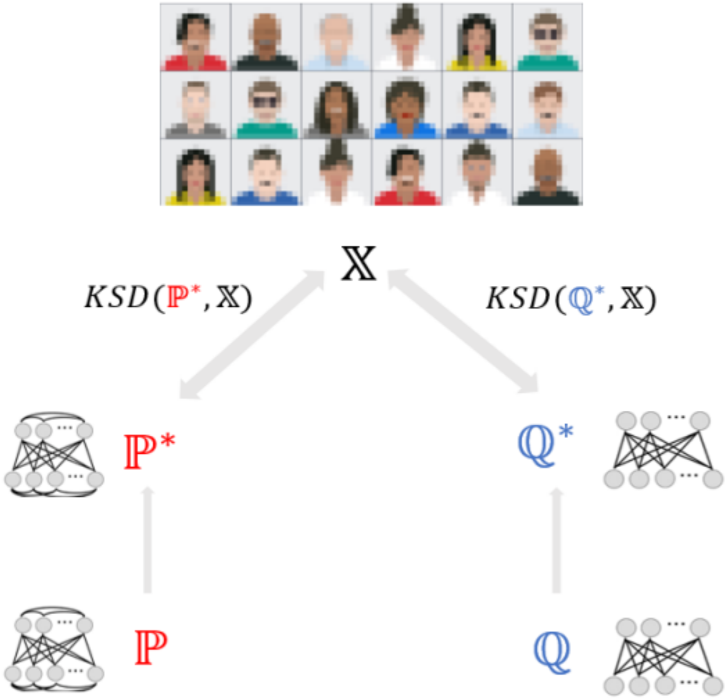
- **CR-GOF:** determine whether model set $\{\mathbb{P}\}_{\theta\in\Theta}$ or $\{\mathbb{Q}\}_{\phi\in\Phi}$ fits the data $\mathbb{X}$ better. Select $\mathbb{P}_{\theta^*}, \mathbb{Q}_{\phi^*}$ from $\{\mathbb{P}\}_{\theta\in\Theta}$ and $\{\mathbb{Q}\}_{\phi\in\Phi}$
- $H_0 : D(\mathbb{P}_{\theta^*}, \mathbb{X}) - D(\mathbb{Q}_{\phi^*}, \mathbb{X}) \leq 0$, $H_1 : D(\mathbb{P}_{\theta^*}, \mathbb{X}) - D(\mathbb{Q}_{\phi^*}, \mathbb{X}) > 0$



# How to choose a statistical discrepancy $D$?

- Different $D$ result in different hypothesis testing.
- Classical GOF methods (Chi-square test, K-S test, C-V test) cannot applied for the models that only known up to a normalization constant term.
- Use Kernelized Stein discrepancy (KSD) as $D$
- **Kernel-based composite relative goodness-of-fit testing (KCR-GOF)**

# KCR-GOF




- KSD bypass the expensive computation of $Z$.
- What is KSD?

# Part2. Kernelized Stein discrepancy (KSD)

# Kernelized Stein discrepancy (KSD)

- Two kinds of statistical discrepancy: IPM and $\phi$ discrepancy.

**Definition (Integral Pseudo-probability Metrics)**

$$\mathcal{D}_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mathbb{P}}[f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)] \right|$$

- Different choice of $f$ result in different IPMs.
- For KSD, $f$ is choosed from a unit-ball in RKHS. Why?
  1. Reproducing property
  2. Stein discrepancy

# Foundation of RKHS: Kernel Method

## Definition (kernel function)

Given a Hilbert space $\mathcal{H}$ and a non-empty set $\mathcal{X}$ as well as a map function $\phi(x) : \mathcal{X} \to \mathcal{H}$. Suppose $\mathbf{x}$, $\mathbf{x}' \in \mathcal{X}$, then a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function that

$$k(\mathbf{x},\mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}} \tag{1}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$.

# Foundation of RKHS: Kernel Method

## Definition (Positive definite function)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, ..., a_n) \in \mathbb{R}^n, \forall (x_1, ..., x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 \tag{2}$$

The function $k(\cdot, \cdot)$ is strictly positive definite if for mutually distinct $x_i$, the equality holds only when all the $a_i$ are zero.

## Lemma

Let $\mathcal{H}$ be any Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then a kernel function $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function. And the reverse direction also holds.

# Reproducing Kernel Hilbert spaces

- RKHS is a Hilbert space which contain function with special property call *reproducing property*.

$$\phi(\mathbf{x}) = (x_1, x_2, \sqrt{2}x_1x_2)^T \in \mathbb{R}^3, \text{where } \mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2 \quad (3)$$

- Let's define a function $f : \mathbf{x} \in \mathbb{R}^2 \to f(\mathbf{x}) \in \mathbb{R}^1$ with the feature $(x_1, x_2, \sqrt{2}x_1x_2)^T \in \mathbb{R}^3$ where $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$ as below:

$$f(x) = ax_1 + bx_2 + c\sqrt{2}x_1x_2$$

- Equivalent representation for $f$ just using its coefficients:

$$f(\cdot) = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

- $\langle f(\cdot), k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$, where $k(\cdot, x) = \phi(x)$

# Reproducing Kernel Hilbert spaces (RKHS)

**Definition (RKHS)**

A Hilbert space $\mathcal{H}$ of functions is a reproducing kernel Hilbert space (RKHS) if

1. $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$
2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

# Stein's method

## Definition (Score function)

Assume that $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and $p(x)$ is a smooth density of $\mathcal{X}$, the (Stein) Score function of $p$ is defined as

$$s_p = \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)} \tag{4}$$

## Definition (Stein class)

A function $f : \mathcal{X} \to \mathbb{R}$ is in the Stein class $\mathcal{F}$ of $p$ if $f$ is smooth and satisfies

$$\int_{x \in \mathcal{X}} \nabla_x (f(x)p(x))dx = 0 \tag{5}$$

# Stein's method

## Definition (Stein operator)

Given a target probability distribution $\mathbb{P}$ on some set $\mathcal{X}$ and its stein class $\mathcal{F}$, suppose $\mathcal{A}_p$ is a linear opeartor acting on any function $f \in \mathcal{F}$. We call $\mathcal{A}_p$ is a stein operator of $P$ if the below euqation is hold:

$$\mathbb{E}_{X\sim p}[\mathcal{A}_p f(X)] = 0 \tag{6}$$

The most popular choice of Stein operator is the Langevin Stein operator.

## Definition (Langevin Stein operator)

$$\mathcal{A}_p f(x) = s_p(x) f(x) + \bigtriangledown_x f(x) = \frac{1}{p(x)} \frac{d}{dx}(f(x)p(x)) \tag{7}$$

# Kernelized Stein discrepancy (KSD)

## Definition (Stein discrepancy)

$\mathbb{P}$ is a target distribution support on a non-empty set $\mathcal{X}$, suppose $\mathcal{F}$ is the stein class of it and $\mathcal{A}_p$ is the stein operator acting on the stein class of $\mathbb{P}$. Then the stein discrepancy between $\mathbb{P}$ and another distribution $\mathbb{Q}$ can is given below, with appropriate norm $||\cdot||^*$:

$$\begin{aligned} \mathbb{S}(\mathbb{P},\mathbb{Q}) &= \sup_{f\in\mathcal{F}} ||(\mathbb{E}_{X\sim Q}[\mathcal{A}_p f(X)] - \mathbb{E}_{X\sim Q}[\mathcal{A}_p f(X)])||^* \\ &= \sup_{f\in\mathcal{F}} ||(\mathbb{E}_{X\sim Q}[\mathcal{A}_p f(X)])||^* \end{aligned} \tag{8}$$

# Kernelized Stein discrepancy (KSD)

**Definition 2.5.1**(Langevin Kernel Stein Discrepancy (KSD))

- KSD is a kind of Stein discrepancy:

$$
\begin{aligned}
\mathbb{S}(\mathbb{P}, \mathbb{Q}) &= \sup_{f \in \mathcal{F}} ||(\mathbb{E}_{X \sim Q}[\mathcal{A}_p f(X)] - \mathbb{E}_{X \sim Q}[\mathcal{A}_p f(X)])||^* \\
&= \sup_{f \in \mathcal{F}} ||(\mathbb{E}_{X \sim Q}[\mathcal{A}_p f(X)])||^*
\end{aligned} \tag{9}
$$

- Stein operator: Langevin Stein operator
- norm: $L - 2$ norm
- choose $f$ from a unit-ball of RKHS

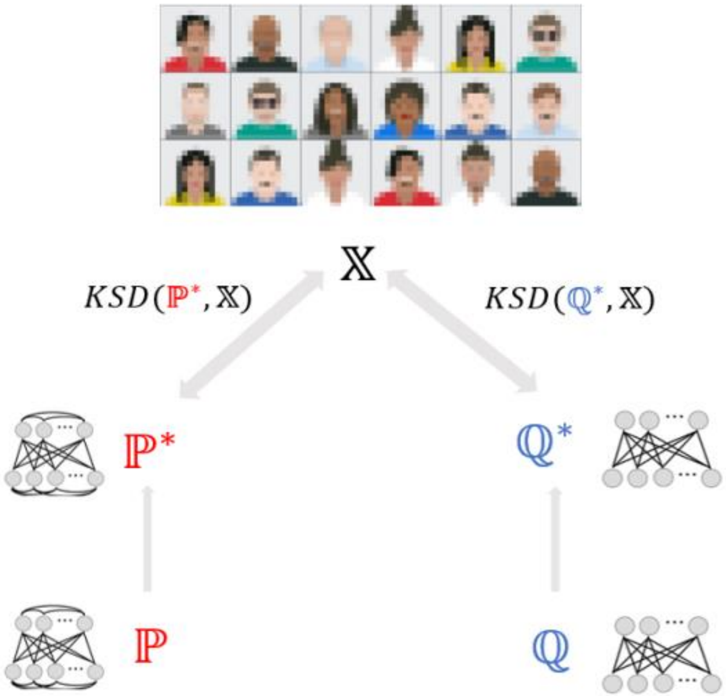
# Kernelized Stein discrepancy (KSD)

Given the i.i.d sample $\{x_i\}$ drawn from an unknown $p$ and the score function $s_q(x)$, we can estimate $\mathbb{S}(p, q)$ by

$$\hat{\mathbb{S}}_u(p, q) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_q(x_i, x_j) \tag{10}$$

# Part3. Methodology

# KCR-GOF




- KSD bypass the expensive computation of $Z$.

# Null hypothesis of KCR-GOF

$$H_0 : \mathbb{S}(\mathbb{P}_{\theta^*}, \mathbb{X}) \leq \mathbb{S}(\mathbb{Q}_{\phi^*}, \mathbb{X})$$
$$H_1 : \mathbb{S}(\mathbb{P}_{\theta^*}, \mathbb{X}) > \mathbb{S}(\mathbb{Q}_{\phi^*}, \mathbb{X}) \tag{11}$$

If the distribution family $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ fits the distribution $\mathbb{X}$ better than $\{\mathbb{Q}_\phi\}_{\phi\in\Phi}$, then there exist at least one model $\mathbb{P}_{\theta^*} \in \{\mathbb{P}_\theta\}_{\theta\in\Theta}$ that $\mathbb{P}_{\theta^*}$ fits $\mathbb{X}$ better than all the models $\{\mathbb{Q}_\phi\}_{\phi\in\Phi}$. Formally,

$$\exists \mathbb{P}_{\theta^*} \in \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \forall \mathbb{Q}_\phi \in \{\mathbb{Q}_\phi\}_{\phi\in\Phi} : \quad \mathbb{S}(\mathbb{P}_{\theta^*}, \mathbb{X}) \leq \mathbb{S}(\mathbb{Q}_\phi, \mathbb{X}) \tag{12}$$

$$\mathbb{P}_{\hat{\theta}} \in \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \quad \text{where } \hat{\theta} = \arg\min_{\theta\in\Theta} \hat{\mathbb{S}}_u(\mathbb{P}_\theta, X_n) \tag{13}$$

# Test statistic

As before, $\mathbb{P}_\theta$ is a parametric model with parameter $\theta$ and $\mathbb{Q}_\phi$ is a parametric model with parameter $\phi$. $\mathbb{X}$ is the population distribution (or data generating process) of a given sample $X_n = \{x_1, x_2, ..., x_n\}$. Then the null hypothesis can be equally rewrited in form of the difference of squared KSDs:

$$\mathbb{S}(\mathbb{P}_\theta, \mathbb{X}) - \mathbb{S}(\mathbb{Q}_\phi, \mathbb{X}) \leq 0 \tag{14}$$

The above equation motivates us to design a test statistic to estimate the above difference of squared KSDs.

Recall that $\mathbb{S}(\mathbb{P}_\theta, \mathbb{X}) = E_{x_i,x_j}[h_{p_\theta}(x_i, x_j)]$ where $x_i, x_j \in \mathbb{X}$. Given $X_n = \{x_1, x_2, ..., x_n\}$, it can be estimated by a U-statistic:

$$\hat{\mathbb{S}}_u(\mathbb{P}_\theta, X_n) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{p_\theta}(x_i, x_j) \tag{15}$$

where

$$\begin{aligned} h_{p_\theta}(x, y) = & \bigtriangledown \log \mathbb{P}_\theta(x)^T \bigtriangledown \log \mathbb{P}_\theta(y) k(x, y) \\ & + \bigtriangledown \log \mathbb{P}_\theta(y)^T \bigtriangledown_x k(x, y) \\ & + \bigtriangledown \log \mathbb{P}_\theta(x)^T \bigtriangledown_y k(x, y) \\ & + \langle \bigtriangledown_x k(x, \cdot), \bigtriangledown_y k(\cdot, y) \rangle_{\mathcal{H}} \end{aligned} \tag{16}$$

## Difference of KSDs

$$\mathbb{S}(\mathbb{P}_\theta, \mathbb{Q}_\phi) = \mathbb{S}(\mathbb{P}_\theta) - \mathbb{S}(\mathbb{Q}_\phi) = E_{x_i, x_j}[h_{p_\theta, q_\phi}(x_i, x_j)] \tag{17}$$

where $h_{p_\theta, q_\phi}(x_i, x_j) = h_{p_\theta}(x_i, x_j) - h_{q_\phi}(x_i, x_j)$ for $x_i, x_j \in \mathbb{X}$.
When two assumptions is statisfied ($h$ is a symmetric matrix and $E[h(x_i, x_j)] < \infty$, then we can consturct a U-statistic. The difference of two U-statistic is also a U-statistic. Similar to the previous definition, let's define their difference as below:
**Test statistic**

$$\hat{\mathbb{S}}_u(\mathbb{P}_\theta, \mathbb{Q}_\phi) = \hat{\mathbb{S}}(\mathbb{P}_\theta, \mathbb{X}) - \hat{\mathbb{S}}_u(\mathbb{Q}_\phi, \mathbb{X}) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h_{p_\theta, q_\phi}(x_i, x_j) \tag{18}$$

where $h_{p_\theta, q_\phi}(x_i, x_j) = h_{p_\theta}(x_i, x_j) - h_{q_\phi}(x_i, x_j)$ for $x_i, x_j \in X_n$.

# Test procedure

Given two sets of candidate models $\{\mathbb{P}_\theta\}_{\theta\in\Theta}, \{\mathbb{Q}_\phi\}_{\phi\in\Phi}$ (usually only known up to normalization term) and a sample $X_n = \{x_1, x_2, \ldots, x_n\} \in \mathbb{X}$. The test procedure of our novel composite relative goodness-of-fit test is a two-stages testing as below:

**Stage 1 (Estimation):**

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \hat{\mathbb{S}}_u(\mathbb{P}_\theta, X_n), \qquad \hat{\phi} = \arg\min_{\phi\in\Phi} \hat{\mathbb{S}}_u(\mathbb{Q}_\phi, X_n)$$

**Stage 2 (Testing):** reject $H_0$ if $\hat{\mathbb{S}}_u(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}_{\hat{\phi}}) \geq c_\alpha$

**Algorithm 1:** Wild bootstrap test

**Input:** $X_n, \mathbb{P}_\theta, \mathbb{Q}_\phi, \alpha, b$

$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{S}(\mathbb{P}_\theta, X_n)$;

$\hat{\phi} = \arg\min_{\phi \in \Phi} \mathbb{S}(\mathbb{Q}_\phi, X_n)$;

**for** $k \in \{1, \ldots, b\}$ **do**

  $w^{(k)} = (w_1, \ldots, w_n)$;

  $\Delta^{(k)} = \frac{1}{n} \sum_{i,j=1}^{n} w_i^{(k)} w_j^{(k)} h_{p_{\hat{\theta}}, q_{\hat{\phi}}}(x_i, x_j)$;

**end**

$c_\alpha = \text{quantile}(\{\Delta^{(1)}, \ldots, \Delta^{(b)}\}, 1 - \alpha)$;

**if** $\Delta = \hat{\mathbb{S}}_u(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}_{\hat{\phi}}) \geq c_\alpha$ **then**

  reject the null;

**else**

  Do not reject;

**end**

**Algorithm 2:** Parametric bootstrap test

**Input:** $X_n, \mathbb{P}_\theta, \mathbb{Q}_\phi, \alpha, b$

$\hat{\theta} = \arg\min_{\theta\in\Theta} \mathbb{S}(\mathbb{P}_\theta, X_n)$;

$\hat{\phi} = \arg\min_{\phi\in\Phi} \mathbb{S}(\mathbb{Q}_\phi, X_n)$;

**for** $k \in \{1, ..., b\}$ **do**

$\quad X_n^{(k)} \sim \mathbb{P}_{\hat{\theta}}($ for type I error$), \mathbb{Q}_{\hat{\phi}}($ for power$)$; $\hat{\theta}^{(k)} = \arg\min_{\theta\in\Theta} \mathbb{S}(\mathbb{P}_\theta, X_n^{(k)})$;

$\quad \hat{\phi}^{(k)} = \arg\min_{\phi\in\Phi} \mathbb{S}(\mathbb{Q}_\phi, X_n^{(k)})$; $\Delta^{(k)} = \hat{\mathbb{S}}_u(\mathbb{P}_{\hat{\theta}^{(k)}}, \mathbb{Q}_{\hat{\phi}^{(k)}})$;

**end**

$c_\alpha = \text{quantile}(\{\Delta^{(1)}, ..., \Delta^{(b)}\}, 1-\alpha)$;

**if** $\Delta = \hat{\mathbb{S}}_u(\mathbb{P}_{\hat{\theta}}, \mathbb{Q}_{\hat{\phi}}) \geq c_\alpha$ **then**

$\quad$ reject the null;

**else**

$\quad$ Do not reject;

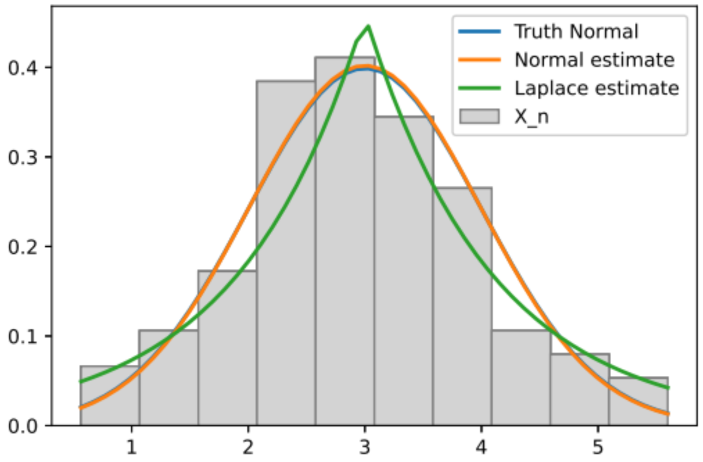
**end**

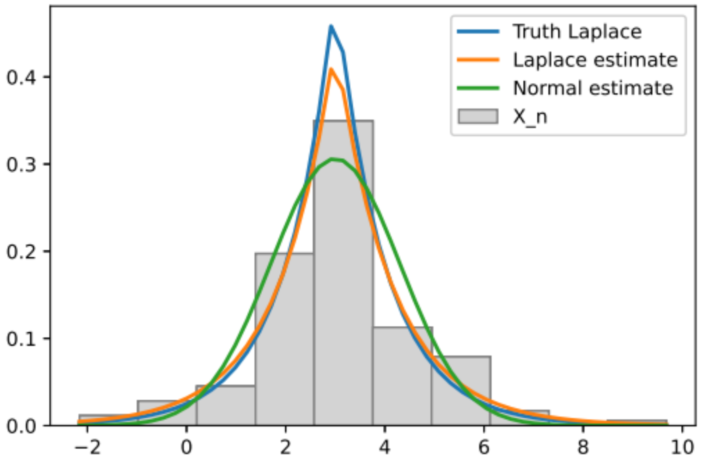
# Part4. Experiment

# Gaussian v.s. Laplace

**Experiemernt 1:** We set the first set of candidate models $\{\mathbb{P}\}_{\theta\in\Theta}$ to be the Gaussian model, and the second set of candiate models $\{\mathbb{Q}\}_{\phi\in\Phi}$ to be the Laplace model. Under $H_0$, we generate sample $X_n = \{x_i\}_{i=1}^n \sim \mathcal{N}(3,1)$ with different sample size $n$. Under $H_1$, we generate sample $X_n = \{x_i\}_{i=1}^n \sim \text{Laplace}(3,1)$.

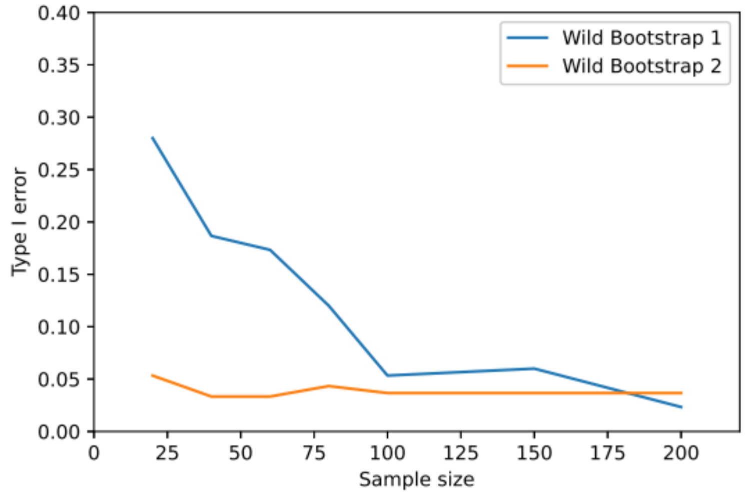
# Gaussian v.s. Laplace

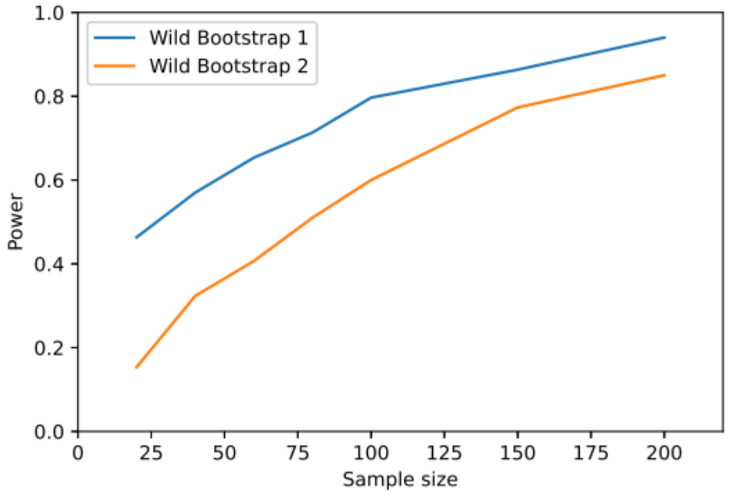


# Gaussian v.s. Laplace

Type I error and power:



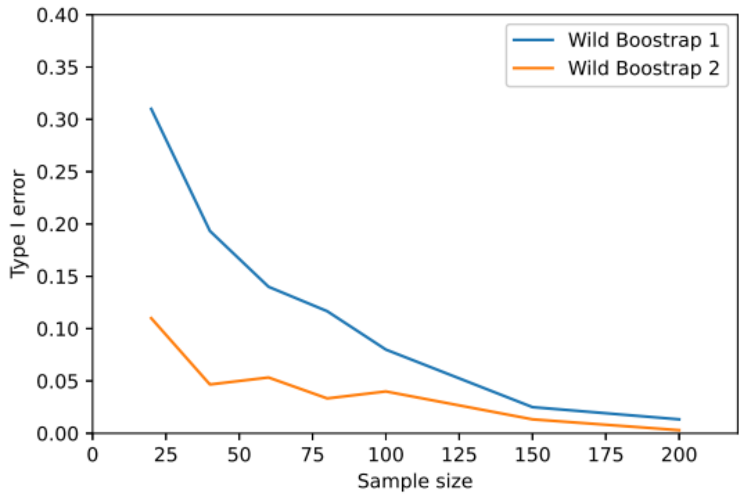



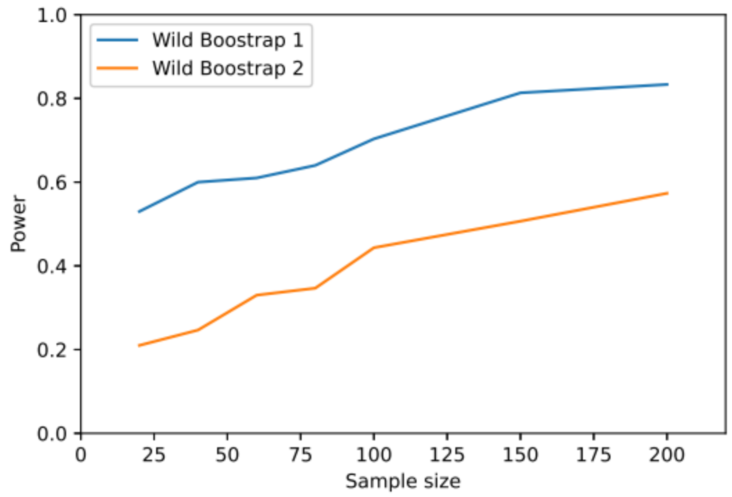
# Gaussian v.s. Laplace

**Experiemernt 2:** We reverse the null hypothesis by setting $\{\mathbb{P}\}_{\theta\in\Theta}$ to be the Laplace model, and set $\{\mathbb{Q}\}_{\phi\in\Phi}$ to be the Gaussian model. Under $H_0$, we generate sample $X_n = \{x_i\}_{i=1}^n \sim \mathcal{N}(3,1)$ with different sample size $n$. Under $H_1$, we generate sample $X_n = \{x_i\}_{i=1}^n \sim \text{Laplace}(3,1)$.

# Gaussian v.s. Laplace

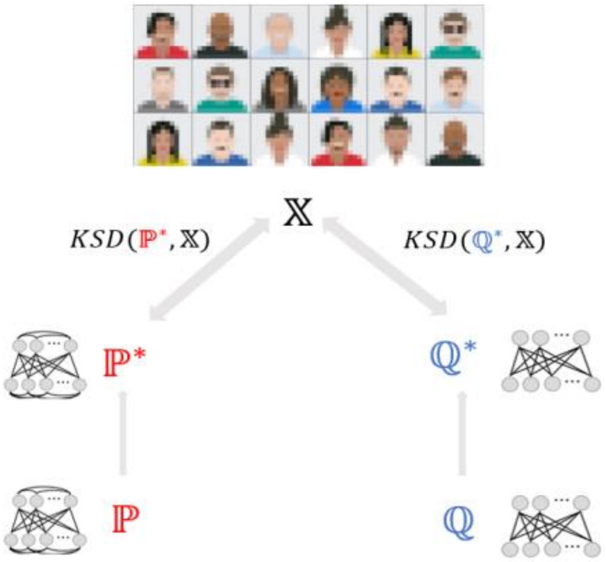
reverse the null hypothesis:



# Conclusion

- Kernel-based Composite Relative Goodness-of-fit testing (KCR-GOF)




- KSD bypass the expensive computation of $Z$.
- Experiment: Gaussian Model v.s. Laplace Model
- Fuure work: Kernelized exponential Model and kernel choice.