Forefront of the Two Sample Problem

From classical to state-of-the-art methods

Yuchi Matsuoka

What is the Two Sample Problem?

$$X_1, \dots, X_l \sim P \; i.i.d$$
$$Y_1, \dots, Y_n \sim Q \; i.i.d$$

- Two Sample Problem

$$P = Q \; ?$$

- Example:
  - Is there a difference in blood glucose level between two groups?
  - Is there a difference in test score between two schools?

Parametric Approach

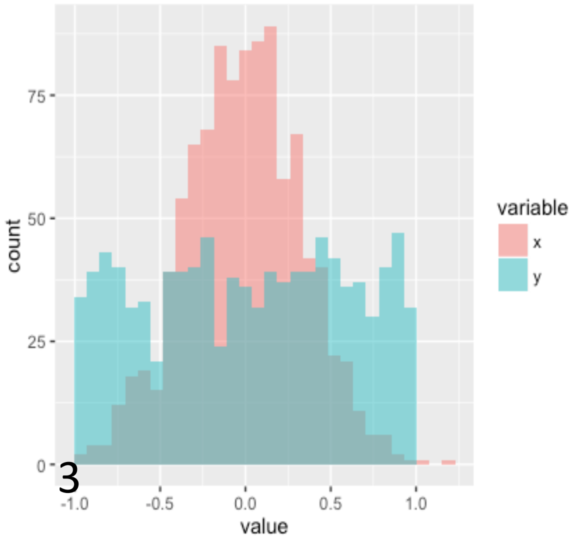
- Assume the type of distribution and compare their empirical moments.
  - Two sample t tests. (1st order moment)
  - F tests. (2nd order moment)

- If the parametric assumptions are not satisfied, they cannot work well.
  - Ex:

$$X \sim N(0, 1/3), \quad Y \sim U(-1, 1)$$

$$E[X] = E[Y] = 0$$

$$V[X] = V[Y] = 1/3$$



Ex: Welch's t-test and F test

In [1]:

```
x <- rnorm(500,mean=0,sd=sqrt(1/3))
y <- runif(500,-1,1)
t.test(x,y)
```

```
        Welch Two Sample t-test

data:  x and y
t = 0.14157, df = 997.72, p-val
ue = 0.8874
alternative hypothesis: true di
fference in means is not equal
 to 0
95 percent confidence interval:
 -0.06953235  0.08034515
sample estimates:
   mean of x    mean of y
 0.001976782 -0.003429617
```

In [2]:

```
var.test(x, y)
```

```
        F test to compare two v
ariances

data:  x and y
F = 1.034, num df = 499, denom
 df = 499, p-value = 0.7089
alternative hypothesis: true ra
tio of variances is not equal t
o 1
95 percent confidence interval:
 0.8674189 1.2326030
sample estimates:
ratio of variances
          1.034013
```

Apparently, they cannot take into account 3rd or higher order moments.

Nonparametric Approach (Kolmogorov Smirnov test)

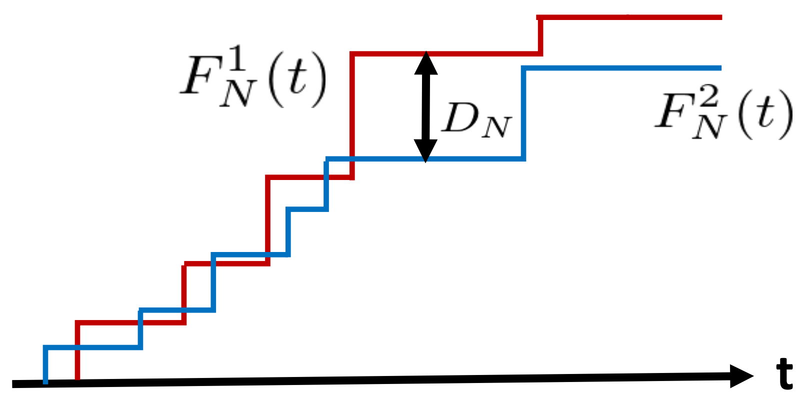
- One dimensional variables
  - Empirical Distribution

$$F_N(t) = \frac{1}{N}\sum_{i=1}^{N} I(X_i \leq t)$$

- KS test statistics

$$D_N = \sup_{t\in\mathbb{R}} \left| F_N^1(t) - F_N^2(t) \right|$$

Image:



Nonparametric Approach (Mann–Whitney *U* test)

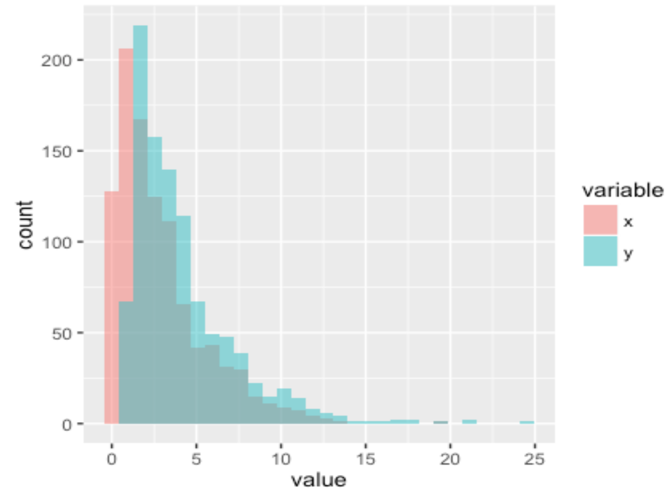
- Define kernel

$$h(x,y) = \mathbf{1}\{X \leq Y\}$$

And corresponding U-statistic is

$$U_{\ell,n} = \frac{1}{\ell n}\sum_{i,j}\mathbf{1}\{X_i \leq Y_j\}$$

Test statistic: $\ell n U_{\ell,n}$



- Mann-Whitney U statistic is average number of $X \leq Y$ for possible combinations.

Asymptotic properties

✓For more general kernel, see Van der Vaart(2000).

- Consider kernel $h(x_i, y_j)$, and U-statistics

$$U_{\ell,n} = \frac{1}{\ell n} \sum_{i,j} h(X_i, Y_j)$$

- Assume, $\frac{\ell}{N} \to \gamma, \;\; \frac{n}{N} \to 1-\gamma$. Then,

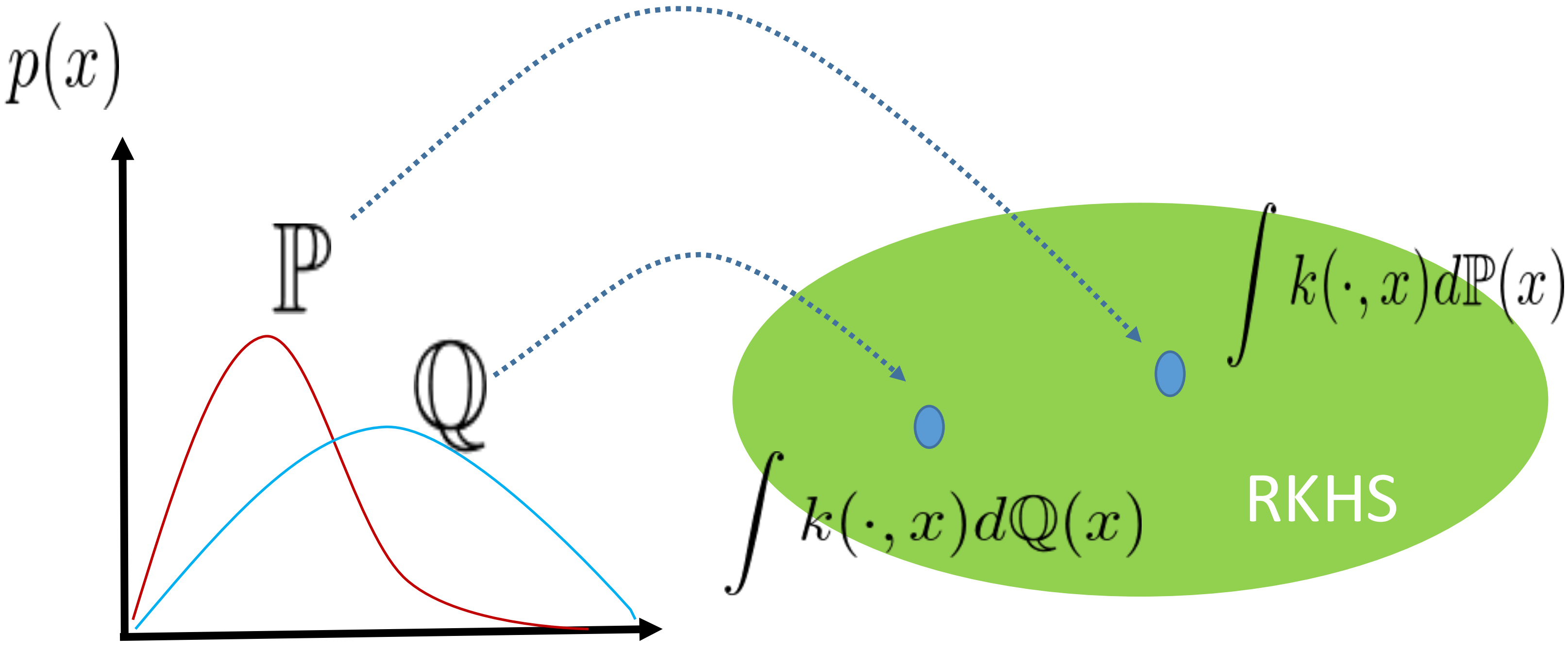
$$\sqrt{N}(U - \theta) \xrightarrow{\mathrm{d}} N(0, \zeta_{1,0}/\gamma + \zeta_{0,1}/(1-\gamma))$$

where, $\theta = E[h(X_1, Y_1)]$,

$$\zeta_{1,0} = \mathrm{Cov}[h(X_1, Y_1), h(X_1, Y_1')], \;\; \zeta_{0,1} = \mathrm{Cov}[h(X_1, Y_1), h(X_1', Y_1)]$$

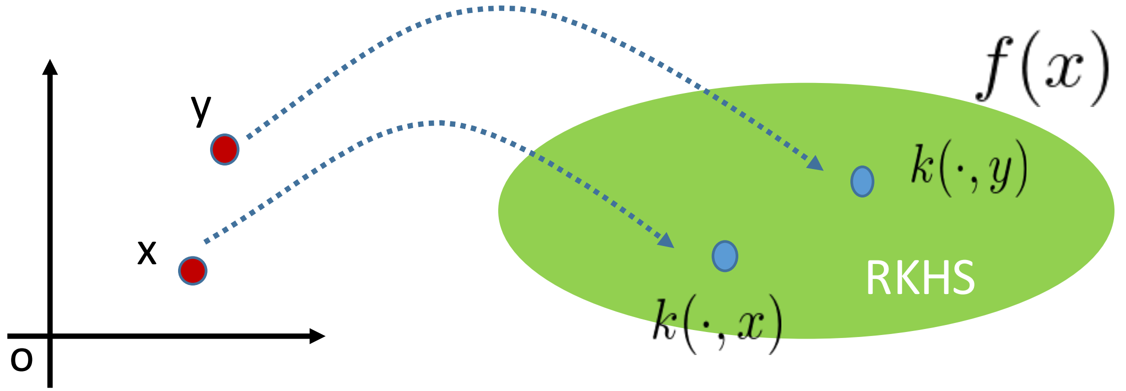
Extension to kernel methods

- Idea: RKHS embedding of distributions.



Relationship with usual Kernel Method

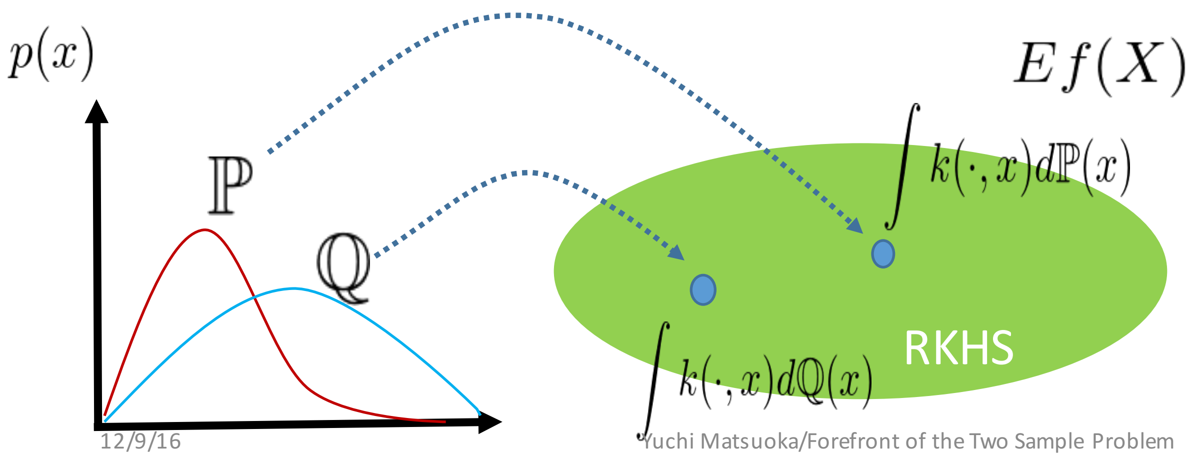
Feature map



Reproducing property

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

Measure map



Reproducing property

$$Ef(X) = \langle f, \int k(\cdot, x) d\mathbb{P}(x) \rangle_{\mathcal{H}_k}$$

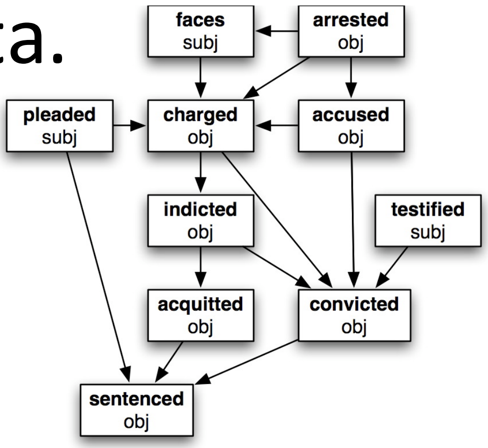
We write

$$m_{\mathbb{P}}^k := \int k(\cdot, x) d\mathbb{P}(x)$$

Why Kernel?

- Characteristic kernels hold all the information of the moment.
  - This can be checked easily. Consider the taylor expansion of characteristic kernels, e.g. gaussian kernel, and take expectations by any distribution.

- They can be defined for any data.
  - Of course, they can be used for multi-dimensional data. Furthermore, for structured data such as strings.



Applications of Kernel Method

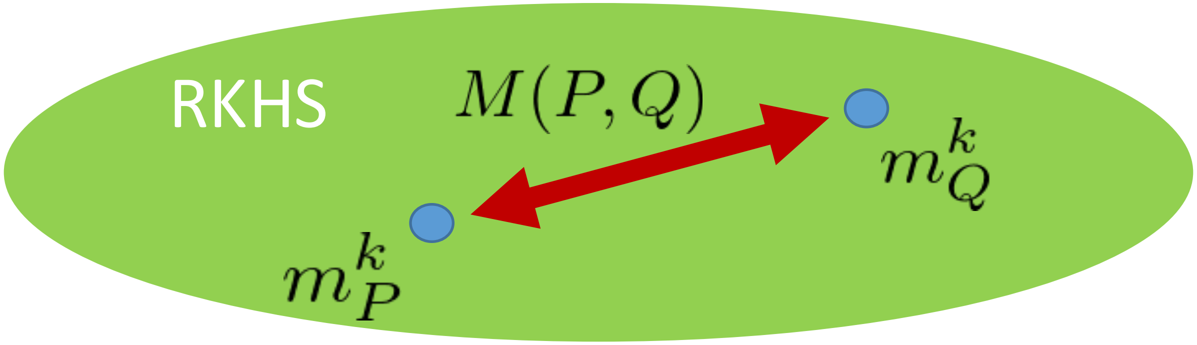
- Feature map perspective
  - SVM, Kernel PCA,
    Kernel CCA, Kernel FDA,
    Kernel Ridge Regression, SVR, etc…
- Measure map perspective
  - Kernel Two Sample Test, HSIC,
    Kernel Dimensionality Reduction
  - Kernel Bayes Rule,
    Kernel Monte Carlo Filter,
    Kernel Spectral Algorithm for HMM,
    Support Measure Machines, etc…

Measure distance of distributions in RKHS

- MMD: maximum mean discrepancy
  - To conduct test, we define the distance of distributions as MMD;

$$M^2(P,Q) \equiv ||m_P^k - m_Q^k||^2_{\mathcal{H}_k}$$

- If $P = Q$, MMD becomes 0.



Empirical Estimator of MMD

- By replacing kernel mean with its empirical estimator,

$$\hat{M}_{\ell,n} = ||\hat{m}_P - \hat{m}_Q||^2_{\mathcal{H}}$$

$$= \frac{1}{\ell^2}\sum_{a,b=1}^{\ell} k(X_a, X_b) + \frac{1}{n^2}\sum_{c,d=1}^{n} k(Y_c, Y_d) - \frac{2}{\ell n}\sum_{a=1}^{\ell}\sum_{c=1}^{n} k(X_a, Y_c).$$

- We use its unbiased version;

$$U_{\ell,n} = \frac{1}{\ell(\ell-1)}\sum_{a\neq b} k(X_a, X_b) + \frac{1}{n(n-1)}\sum_{c\neq d} k(Y_c, Y_d)$$

$$-\frac{2}{\ell n}\sum_{a=1}^{\ell}\sum_{c=1}^{n} k(X_a, Y_c).$$

Relationship with U statistics

- $U_{\ell,n}$ is an 2 sample U-statistics using kernel

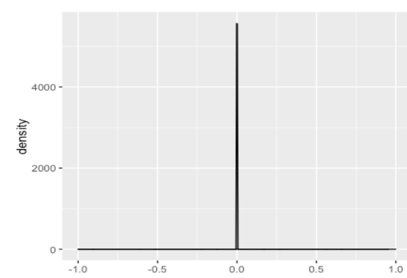
$$h(x_1, x_2; , y_1, y_2) = k(x_1, x_2) + k(y_1, y_2) \\ -\frac{1}{2}\{k(x_1, y_1) + k(x_1, y_2) + k(x_2, y_1) + k(x_2, y_2)\}$$

- We can get its asymptotic null distribution applying theory of U-statistics!!

Difficulty

- Reminded the claim of theorem of asymptotic normality of U-statistics, it calculate a quantity, $\zeta_{1,0}, \zeta_{0,1}$.
- This case, it becomes 0.
- This implies that we have to multiply quantities bigger than $\sqrt{N}$ in order not to asymptotic distribution degenerate.

$$\sqrt{N}(U_{\ell,m} - \theta)$$



Key Theorem

Let total sample size is N, and N = l + n.

Assume

$$\frac{\ell}{N} \to \gamma, \quad \frac{n}{N} \to 1 - \gamma$$

Then, Under the null hypothesis $P = Q$,

$$NU_{\ell,n} \overset{\mathrm{d}}{\to} \sum_{i=1}^{\infty} \lambda_i \left(Z_i^2 - \frac{1}{\gamma(1-\gamma)} \right)$$

$$\text{where } Z_i \overset{\text{i.i.d}}{\sim} N\left(0, \frac{1}{\gamma(1-\gamma)}\right),$$

- (Proof) See 福水(2010).

Who are $\{\lambda_i\}_{i=1}^{\infty}$ s?

- A non-zero eigenvalues of integral operator over $L^2(P)$ that has kernel

$$\tilde{k}(x,y) = k(x,y) - E[k(x,X)] - E[k(X,y)] + E[k(X,\tilde{X})]$$

$$\tilde{X}, X \overset{\text{i.i.d}}{\sim} P.$$

i.e. non-negative real value that satisfies

$$\int \tilde{k}\phi_i(x,y)dP(y) = \lambda_i \phi_i(x).$$

Proposed method to get a critical value

- We have to find $\{\lambda_i\}_{i=1}^{\infty}$. This can be estimated consistently by eigenvalues of the gram matrix defined by,

$$\tilde{K}_{ij} = k(X_i, X_j) - \frac{1}{N}\sum_{b=1}^{N} k(X_i, X_b) - \frac{1}{N}\sum_{a=1}^{N} k(X_a, X_j) + \frac{1}{N^2}\sum_{a,b=1}^{N} k(X_a, X_b) = (Q_N K Q_N)_{ij}.$$

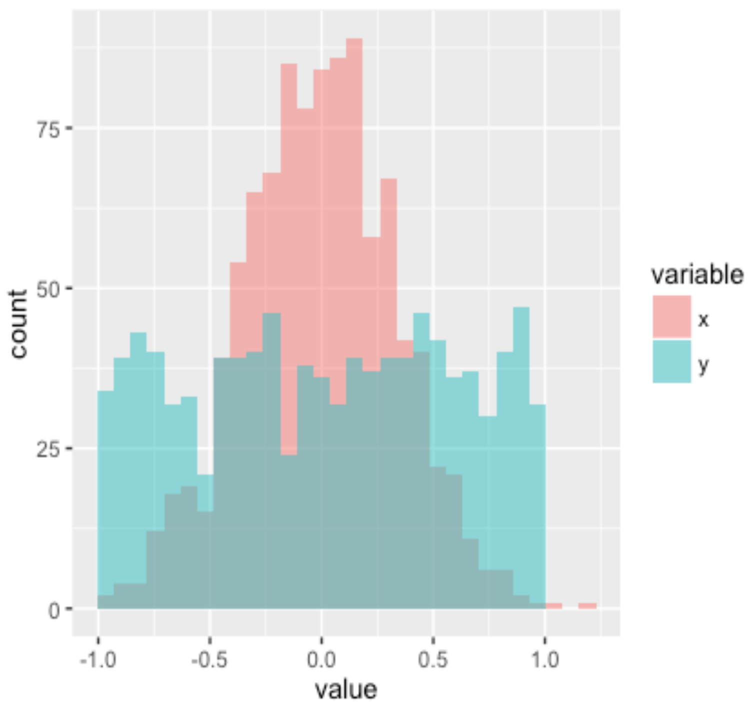
$$\text{where.} \qquad Q_N = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$$

(called centered gram matrix.)

Power analysis

- Synthetic Data

$$X \sim N(0, 1/3), \quad Y \sim U(-1, 1)$$



What times is null hypothesis rejected on 100 trials.

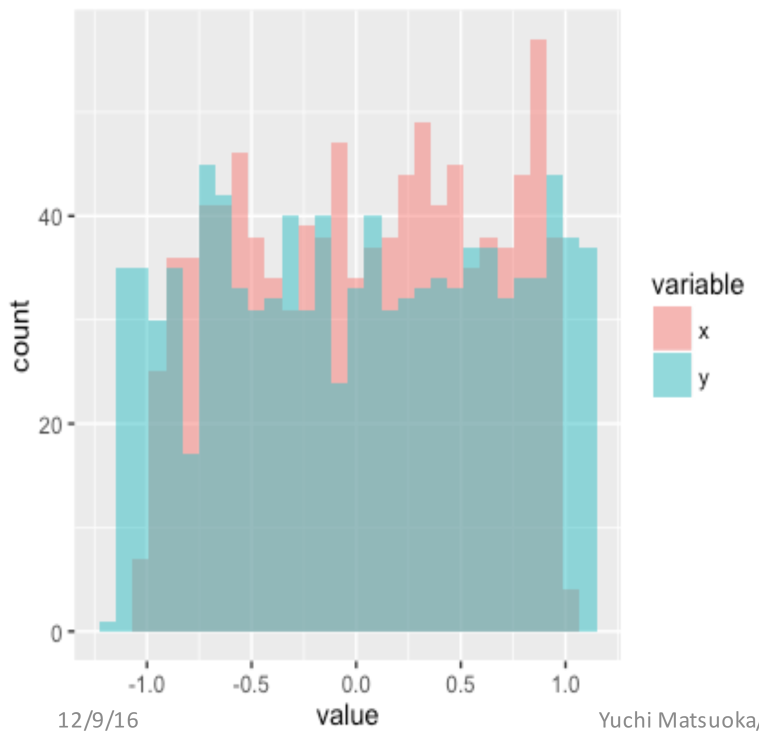
EachSampleSize	KolmogorovSmirnov	Mann_Whitney	Kernel
100	92	5	100
500	100	7	100

Kernel Two Sample Test sperilor to Other tests in terms of power.

Power analysis

- Synthetic Data

$$X \sim U(-1,1), \quad Y \sim U(-1.15, 1.15)$$



What times is null hypothesis rejected on 100 trials.

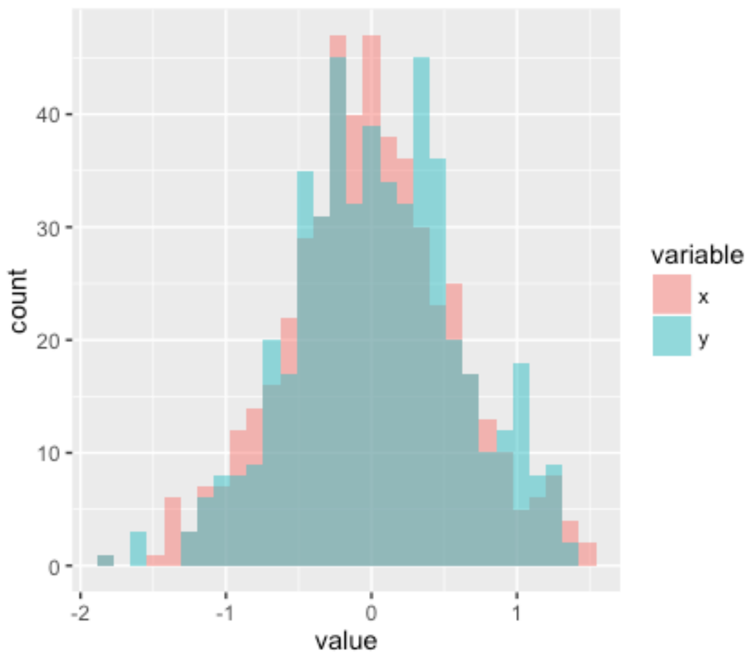
EachSampleSize	KolmogorovSmirnov	Mann_Whitney	Kernel
100	7	2	23
500	45	6	77
1000	97	10	100

Kernel Two Sample Test sperilor to Other tests in terms of power.

Significance analysis

- Synthetic data

$$X \sim N(0, 1/3), \quad Y \sim N(0, 1/3)$$



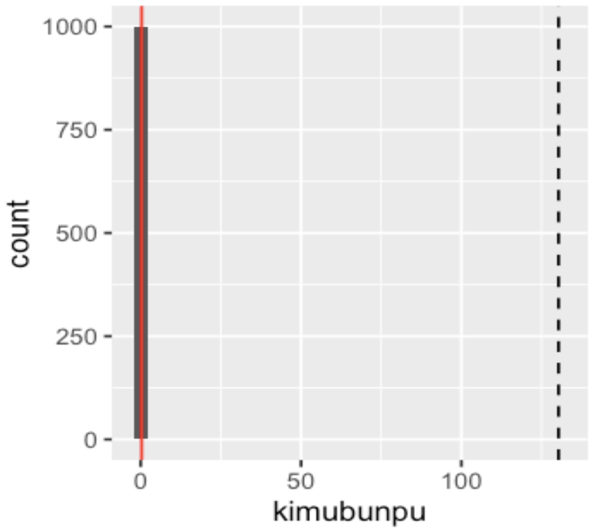
Type 1 error rate on 5000 trials.

EachSampleSize	KolmogorovSmirnov	Mann_Whitney	Kernel
100	0.038	0.048	0.067
500	0.050	0.061	0.048
1000	0.050	0.045	0.057

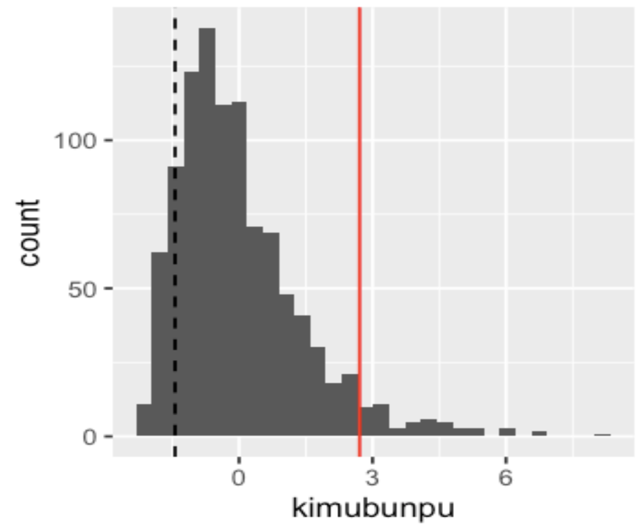
All tests output the expected values.

Multidimensional case: iris data(setosa and virginica)

- Labeled 5-dim data (m=50, l=50, N=100)
  - Is there a difference between these features?



- Setosa and Setosa
  - Null hypothesis should not be rejected.



Histogram: estimated null distribution. Red line: critical value. Dash Line: test statistic.

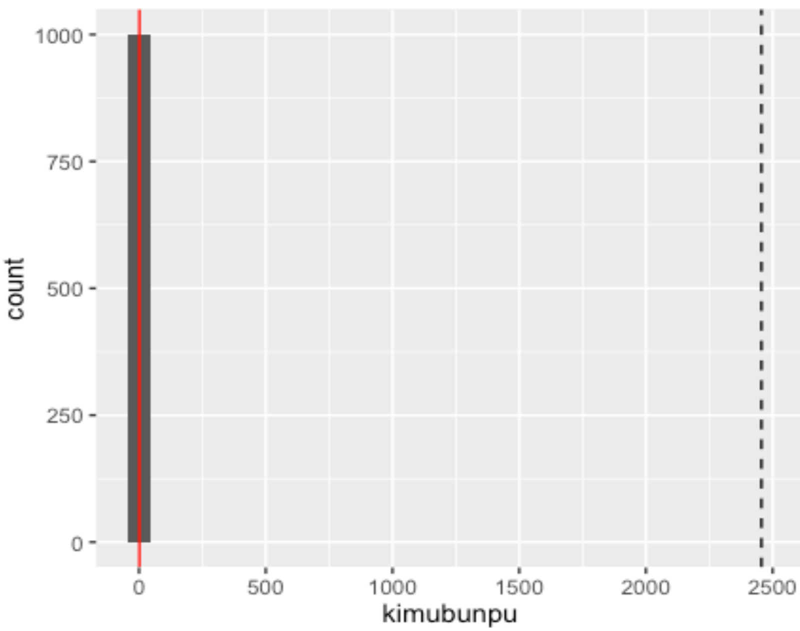
High dimensional case: MNIST data(hand-written digit data)

- Labeled(1~9) data with 784 features.
  - Each feature represents 0 or 1 of the pixel.

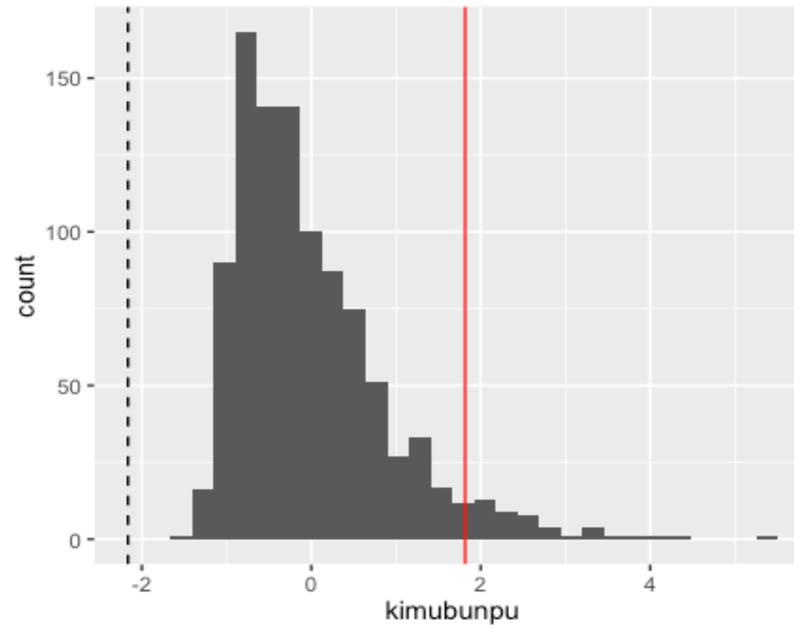
- Is it possible to classify these numbers based on the distribution?
- Can Kernel two sample test overcome the curse of dimensionality?

Digit “1” vs Digit “2”

- Compare group of “1” and “2”.
  - Each group is about 4000 sample.



- Divide “1” into two groups.



Histogram: estimated null distribution.
Red line: critical value.
Dash Line: test statistic.

Reference

- A. Gretton, et al. A fast, consistent kernel two-sample test. *Advances in neural information processing systems (2009).*
- A. Gretton, et al. A kernel two-sample test. *Journal of Machine Learning Research* 13: 723-773. (2012).
- 福水健次. カーネル法入門—正定値カーネルによるデータ解析. *朝倉 書店,* (2010).
- K. Muandet, et al. Kernel Mean Embedding of Distributions: A Review and Beyonds. *arXiv:1605.09522* (2016).
- A. W. Van der Vaart. Asymptotic statistics. *Cambridge series in statistical and probabilistic mathematics.* (2000).