Two-Stage U-Statistics for Hypothesis Testing

G. Shieh, H.-C. Ho

Arthur Gretton's notes

February 6, 2014

U-Statistics

Given a sample drawn i.i.d. according to a probability measure P,

$$\{x_i\}_{i=1}^{n}$$

we want the minimum variance unbiased estimator of

$$E_P h(X_1, \dots X_K),$$

where h is assumed symmetric, and the $X_i \sim P$ are independent random variables. The estimator is

$$U_n = C(n,k)^{-1} \sum_{C(n,k)} h(x_{i_1}, \dots, x_{i_k}),$$

where $\sum_{C(n,k)}$ is the sum over the combinations $C(n,k)$.

The U-statistic is degnerate of order $j < k$ if

$$E_P h(x_1, \ldots, x_j, X_{j+1}, \ldots X_k) = 0.$$

Note that the degeneracy of a U-statistic might be indeterminate. An example is MMD:

$$h(Z, Z') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y)$$

where $Z := (X, Y)$ and $X \sim P$, $Y \sim Q$.

Degenerate U-Statistics

The U-statistic is degnerate of order $j < k$ if

$$E_P h(x_1, \ldots, x_j, X_{j+1}, \ldots X_k) = 0.$$

Note that the degeneracy of a U-statistic might be indeterminate. An example is MMD:

$$h(Z, Z') = k(X, X') + k(Y, Y') - k(X, Y') - k(X', Y)$$

where $Z := (X, Y)$ and $X \sim P$, $Y \sim Q$.
Take expectation wrt one variable:

$$E_Z h(Z, z) = E_X k(X, x') + E_Y k(Y, y') - E_X k(X, y') - E_Y k(x', Y)$$

This is zero when $P = Q$, but may not be zero when $P \neq Q$ (depends on kernel).

Degenerate U-statistics are a problem

When a U-statistic is degenerate, the asymptotic distribution is complicated.
Again, example of MMD (assume equal number of samples from P and Q):

$$n\mathrm{MMD}^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Q_i^2 - 2),$$

where

$$Q_i \sim \mathcal{N}(0,2) \text{ i.i.d.}, \qquad \int_{\mathcal{X}} \underbrace{\tilde{k}(x,x')}_{\text{centred}} \psi_i(x) d\Pr(x) = \lambda_i \psi_i(x')$$

Degenerate U-statistics are a problem

When a U-statistic is degenerate, the asymptotic distribution is complicated.
Again, example of MMD (assume equal number of samples from P and Q):

$$n\mathrm{MMD}^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i (Q_i^2 - 2),$$

where

$$Q_i \sim \mathcal{N}(0,2) \text{ i.i.d.}, \qquad \int_{\mathcal{X}} \underbrace{\tilde{k}(x,x')}_{\text{centred}} \psi_i(x) d\Pr(x) = \lambda_i \psi_i(x')$$

A solution that doesn't work too well: the incomplete U-statistic.
Make N independent draws of $h(x_1, \ldots, x_k)$, where $N/n \to 0$.
By central limit theorem, this is asymptotically normal.

A better solution

1. Divide the data into m blocks of size l.
2. Compute a U-statistic on jth block,

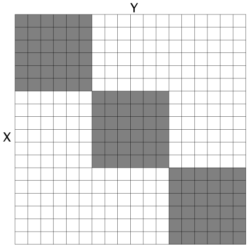
$$I_j := l^{-1/2} \sum_{\mathbf{i} \sim C(l_j, k)} h(x_{i_1}, \ldots, x_{i_k}),$$

where $\mathbf{i} \sim C(l_j, k)$ denotes sampling l times *with replacement* from $C(l, k)$ in jth block (slightly odd)

3. Take an average:

$$T_{m,1} = m^{-1} \sum_{j=1}^{m} I_j.$$

This converges in distribution to a Gaussian (central limit theorem). Variance can be computed e.g. by bootstrap.



Provably more powerful test than incomplete U-statistic

An even better solution

Define a new U-statistic over the blocks. Need to know the degeneracy d.
(MMD has $d = 1$ under $\mathcal{H}_0$)
Define the centered block U-statistic

$$\tilde{I}_j := l^{(d+1)/2} \left(C(l,k)^{-1} \sum_{C(l,k)} h(X_{i_1}, \ldots, X_{i_k}) - \hat{\theta} \right)$$

where $\hat{\theta}$ is the U-statistic computed on the whole sample (for centering).

An even better solution

Define a new U-statistic over the blocks. Need to know the degeneracy d.
(MMD has $d = 1$ under $\mathcal{H}_0$)
Define the centered block U-statistic

$$\tilde{I}_j := l^{(d+1)/2} \left(C(l,k)^{-1} \sum_{C(l,k)} h(X_{i_1}, \ldots, X_{i_k}) - \hat{\theta} \right)$$

where $\hat{\theta}$ is the U-statistic computed on the whole sample (for centering).
Define a test statistic:

$$T_{m,t} = C(m,t)^{-1} \sum_{C(m,t)} \left[\tilde{I}_{j_1} \times \ldots \times \tilde{I}_{j_t} + l^{t(d+1)/2} \left(\hat{\theta}\right)^t \right],$$

where t is the order of the U-statistic ($t \geq 2$ improves over previous method)

Asymptotics of $T_{m,t}$ are tractable

Under $\mathcal{H}_0$, given $d \geq 1$,

$$m^{t/2} T_{m,t} \xrightarrow{D} v_{d+1}^t H_t(Z),$$

where

- $Z \sim \mathcal{N}(0, 1)$,
- H_k is the kth Hermite polynomial, $H_k(x) = (-1)^k e^{x^2/2} \left(\frac{d^k e^{-x^2/2}}{dx^k} \right)$.
    - $H_2(Z) = Z^2 - 1$
- $v_{d+1} = \sigma_{d+1} \sqrt{(d+1)!} C(k, d+1)$, σ_{d+1} is leading non-zero term in U-statistic variance expansion.

A surprising result

TU statistics can yield more powerful tests than the full U-statistics.
Proved when:

- t=2
- The U-statistic is degenerate under $\mathcal{H}_0$, non-degenerate under $\mathcal{H}_1$.

Idea:
The probability of correctly rejecting the null hypothesis for the U-statistic:

$$1 - \Phi\left[\left(C_U(\alpha)n^{-d/2} - \theta\sqrt{n}\right)/\sigma\right],$$

Φ is standard normal CDF
The probabilty of correctly rejecting the null hypothesis for the TU-statistic:

$$1 - G\left[\left(C_{TU}(\alpha) - n\theta^d l^d\right)/v_{d+1}^2\right],$$

where G is CDF of $\chi_1^2 - 1$. Recall $l = \left[n^\lambda\right]$ for $0 < \lambda < 1$.