Topics in U-statistics and Risk Estimation

Qing Wang and Bruce G. Lindsay

March 17, 2011
Outline

■ Introduction of U-statistics
■ Proposed Unbiased Variance Estimator
■ Implementation in Risk Estimation
■ Future Work
■ Reference
Suppose $F$ is a $p$-variate distribution function ($p \in \mathbb{N}^+$), denoted as $F(x) = F(x^{(1)}, ..., x^{(p)})$. We are considering a parameter of interest $\theta$ which can be written as a functional of $F$ with the following form:

$$\theta(F) = \int \cdots \int K(x_1, x_2, ..., x_m) dF(x_1) dF(x_2) \cdots dF(x_m)$$

where $x_1, ..., x_m$ are all $p$-variate and $K$ is a symmetric function of $m$ arguments.

- Given a sample of size $n$ ($n \geq m$, $m =$ size of the kernel) from $F$, $K(X_1, X_2, ..., X_m)$ is an unbiased estimate of the parameter $\theta$.

- However, intuition reminds us that there should be some better estimators, since $K(X_1, X_2, ..., X_m)$ does not use up the entire data set.
Suppose $F$ is a $p$-variate distribution function ($p \in \mathbb{N}^+$), denoted as $F(x) = F(x^{(1)},...,x^{(p)})$. We are considering a parameter of interest $\theta$ which can be written as a functional of $F$ with the following form:

$$\theta(F) = \int \cdots \int K(x_1, x_2, ..., x_m) dF(x_1)dF(x_2)\cdots dF(x_m)$$

where $x_1, ..., x_m$ are all $p$-variate and $K$ is a symmetric function of $m$ arguments.

- Given a sample of size $n$ ($n \geq m$, $m =$ size of the kernel) from $F$, $K(X_1, X_2, ..., X_m)$ is an unbiased estimate of the parameter $\theta$.

- However, intuition reminds us that there should be some better estimators, since $K(X_1, X_2, ..., X_m)$ does not use up the entire data set.
Definition (Hoeffding (1948))

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables (vectors) and $K(x_1, ..., x_m)$ be a symmetric real-valued function of $m$ arguments, then a U-statistic is defined as:

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < ... < i_m \leq n} K(X_{i_1}, ..., X_{i_m})$$

- The unbiasedness of $U_n$ follows from the unbiasedness of $K$.
- It can be seen that $U_n$ is a function of order statistics (which is a set of sufficient statistics).
- When we are doing nonparametric inference, the set of order statistics is a complete sufficient statistic if the underlying distribution family is large enough (Fraser (1954)). Therefore, a U-statistic is the best unbiased estimator in this context by Rao-Blackwell Theorem.
Definition (Hoeffding (1948))

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables (vectors) and $K(x_1, ..., x_m)$ be a symmetric real-valued function of $m$ arguments, then a U-statistic is defined as:

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < ... < i_m \leq n} K(X_{i_1}, ..., X_{i_m})$$

- The unbiasedness of $U_n$ follows from the unbiasedness of $K$.
- It can be seen that $U_n$ is a function of order statistics (which is a set of sufficient statistics).
- When we are doing nonparametric inference, the set of order statistics is a complete sufficient statistic if the underlying distribution family is large enough (Fraser (1954)). Therefore, a U-statistic is the best unbiased estimator in this context by Rao-Blackwell Theorem.
As a motivating problem, we consider risk estimation in the context of nonparametric kernel density estimation.

Consider a probability density function of a continuous random variable $X$, denoted as $f(x)$.

Consider the most visible and used density estimation method, the nonparametric kernel density estimator which is defined for a kernel $K$ as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where $x \in \mathcal{R}$, $h > 0$, and $K_h(t) = \frac{1}{h}K(t/h)$. 

Example in Risk Estimation
As a motivating problem, we consider risk estimation in the context of nonparametric kernel density estimation.

- Consider a probability density function of a continuous random variable $X$, denoted as $f(x)$.
- Consider the most visible and used density estimation method, the nonparametric kernel density estimator which is defined for a kernel $K$ as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i),$$

where $x \in \mathcal{R}$, $h > 0$, and $K_h(t) = \frac{1}{h} K(t/h)$. 
One option to select the “optimal” bandwidth is to compute a risk function which measures the “average distance” between $\hat{f}_h(x)$ and $f(x)$ in a certain fashion, and the best bandwidth $h^*$ is considered as the one that yields the smallest risk score.

In practice, given a dataset one estimates the risk function and uses $\hat{h}^*$.

As a U-statistic is the best unbiased estimator for nonparametric inferences, we would like to construct a U-statistic form estimator for the risk that arises from $L^2$ loss function.
U-statistic Form $L^2$ Risk Estimator

- $L^2$ loss based on a Gaussian kernel $K_t(x, x_0) = \frac{1}{t\sqrt{2\pi}} e^{-\frac{(x-x_0)^2}{2t^2}}$

The risk of $\hat{f}_h(\cdot)$ based on $L^2$ loss is defined as:

$$\text{Risk}_{L^2, n} = E[X_n] \left[ \int (f(x) - \hat{f}_h(x))^2 dx \right]$$

$$\propto \frac{n-1}{n} E[K_{\sqrt{2}h}(X_1, X_2)] - 2E[K_h(X_1, X_2)].$$

Therefore, a U-statistic estimate for the above relative risk can be constructed as

$$U_{L^2} = \frac{1}{\binom{n}{2}} \sum_{i<j} K_{L^2,h}(X_i, X_j), \text{ where}$$

$$K_{L^2,h}(x_1, x_2) = \frac{n-1}{n} K_{\sqrt{2}h}(x_1, x_2) - 2K_h(x_1, x_2).$$

It can be shown that the above risk estimator $U_{L^2}$ is identical to UCV estimator.

Qing Wang and Bruce G. Lindsay

Department of Statistics, Pennsylvania State University

Topics in U-statistics and Risk Estimation
U-statistic Form $L^2$ Risk Estimator

- $L^2$ loss based on a Gaussian kernel $K_t(x, x_0) = \frac{1}{t\sqrt{2\pi}} e^{-\frac{(x-x_0)^2}{2t^2}}$

The risk of $\hat{f}_h(\cdot)$ based on $L^2$ loss is defined as:

$$Risk_{L^2,n} = E[X_n \left[ \int (f(x) - \hat{f}_h(x))^2 dx \right]]$$

$$\propto \frac{n-1}{n} E[K_{\sqrt{2}h}(X_1, X_2)] - 2E[K_h(X_1, X_2)].$$

Therefore, a U-statistic estimate for the above relative risk can be constructed as

$$U_{L^2} = \frac{1}{\binom{n}{2}} \sum_{i<j} K_{L^2,h}(X_i, X_j), \text{ where}$$

$$K_{L^2,h}(x_1, x_2) = \frac{n-1}{n} K_{\sqrt{2}h}(x_1, x_2) - 2K_h(x_1, x_2).$$

It can be shown that the above risk estimator $U_{L^2}$ is identical to UCV estimator.
As a U-statistic is an unbiased estimator of the parameter of interest, exploring its variance to evaluate the parameter estimation is always crucial and of interest.

In the case of risk estimation, we may want to know how precise we estimate the risk function by a U-statistic.
Established Results

- **Theorem 1:** For $m$ fixed, $n \to \infty$.
  Suppose the kernel function $K$ is twice integrable. And let
  \[ \sigma_1^2 = \text{Var}\{E(K(X_1, \ldots, X_m)|X_1)\} \]
  with $0 < \sigma_1^2 < \infty$, then
  \[ \sqrt{n}(U_n - \theta) \to N(0, m^2 \sigma_1^2). \]

- **Theorem 2:** Suppose
  \[ \phi_c = E(K(X_1, \ldots, X_m)|(X_1, \ldots, X_c)), 1 \leq c \leq m \]
  and
  \[ \sigma_c^2 = \text{Var}(\phi_c), 1 \leq c \leq m. \]
  We have
  \[ \text{Var}(U_n) = \frac{1}{\binom{n}{m}} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2. \]
Established Results

- **Theorem 1**: For $m$ fixed, $n \to \infty$.
  
  Suppose the kernel function $K$ is twice integrable. And let $\sigma_1^2 = \text{Var}\{E(K(X_1, \ldots, X_m)|X_1)\}$ with $0 < \sigma_1^2 < \infty$, then
  
  $$\sqrt{n}(U_n - \theta) \to N(0, m^2 \sigma_1^2).$$

- **Theorem 2**: Suppose
  
  $$\phi_c = E(K(X_1, \ldots, X_m)|(X_1, \ldots, X_c)), 1 \leq c \leq m$$
  and
  
  $$\sigma_c^2 = \text{Var}(\phi_c), 1 \leq c \leq m.$$ We have
  
  $$\text{Var}(U_n) = \frac{1}{n \choose m} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.$$
Although the established results give us the asymptotic distribution of U-statistics and their asymptotic variance, they are not reliable whenever $n$ is not large or the kernel size $m$ is not negligible compared with the sample size $n$.

On the other hand, the closed form variance of a U-statistic is complicated in form, especially when both $m$ and $n$ are large.

We propose an unbiased variance estimator of a general U-statistic which is applicable even for the cases that $m/n$ is a fixed fraction.
Although the established results give us the asymptotic distribution of U-statistics and their asymptotic variance, they are not reliable whenever $n$ is not large or the kernel size $m$ is not negligible compared with the sample size $n$.

On the other hand, the closed form variance of a U-statistic is complicated in form, especially when both $m$ and $n$ are large.

We propose an unbiased variance estimator of a general U-statistic which is applicable even for the cases that $m/n$ is a fixed fraction.
Consider a U-statistic defined as

\[ U_n = \frac{1}{\binom{n}{m}} \sum_{i} K(S_i), \text{ where } S_i \text{ is a size-}m \text{ sample out of } i.i.d. X_1, ..., X_n. \]

Define

- Let \( Q(0) = \frac{1}{N_0} \sum_{P_0} K(S_1)K(S_2), \) where \( P_0 \) contains all pairs of size-\( m \) samples with no overlaps and \( N_0 \) is the number of pairs in \( P_0 \). Then, we have \( E[Q(0)] = [E(U_n)]^2. \)

- Let \( Q(m) = \frac{1}{N_m} \sum_{P_m} K(S_1)K(S_2), \) where \( P_m \) contains all pairs of size-\( m \) samples and \( N_0 \) is the number of pairs in \( P_m \). Then, we have \( Q(m) = U_n^2. \)

Therefore, \( Q(m) - Q(0) \) is an unbiased estimate of \( \text{Var}(U_n). \)
Construction of the Unbiased Variance Estimator

Consider a U-statistic defined as

\[ U_n = \frac{1}{\binom{n}{m}} \sum_i K(S_i), \]

where \( S_i \) is a size-\( m \) sample out of \( i.i.d. X_1, \ldots, X_n \).

Define

- Let \( Q(0) = \frac{1}{N_0} \sum_{P_0} K(S_1)K(S_2) \), where \( P_0 \) contains all pairs of size-\( m \) samples with no overlaps and \( N_0 \) is the number of pairs in \( P_0 \). Then, we have \( E[Q(0)] = [E(U_n)]^2 \).

- Let \( Q(m) = \frac{1}{N_m} \sum_{P_m} K(S_1)K(S_2) \), where \( P_m \) contains all pairs of size-\( m \) samples and \( N_0 \) is the number of pairs in \( P_m \). Then, we have \( Q(m) = U_n^2 \).

Therefore, \( Q(m) - Q(0) \) is an unbiased estimate of \( Var(U_n) \).
**Theorem 3:** Suppose $U_n$ is a U-statistic with a kernel $K$ of size $m$, $m \leq n/2$. Denote

$$\hat{V}_u = Q(m) - Q(0)$$

where $Q(m) - Q(0)$ are defined before. Then, $\hat{V}_u$ is an unbiased estimator of $\text{Var}(U_n)$.

- $\hat{V}_u$ is a function of the order statistics and so is the best unbiased estimator of $\text{Var}(U_n)$. (New result.)
- $\hat{V}_u$ itself can be written as a U-statistic with a kernel function of order $2m$. Therefore, it has asymptotic normality under certain regularity conditions in the fixed $m$ case. When $m/n$ is a fixed fraction, lower and upper bounds for $\text{Var}(\hat{V}_u)$ can be obtained.
Theorem 3: Suppose $U_n$ is a U-statistic with a kernel $K$ of size $m$, $m \leq n/2$. Denote

$$\hat{V}_u = Q(m) - Q(0)$$

where $Q(m) - Q(0)$ are defined before. Then, $\hat{V}_u$ is an unbiased estimator of $\text{Var}(U_n)$.

- $\hat{V}_u$ is a function of the order statistics and so is the best unbiased estimator of $\text{Var}(U_n)$. (New result.)

- $\hat{V}_u$ itself can be written as a U-statistic with a kernel function of order $2m$. Therefore, it has asymptotic normality under certain regularity conditions in the fixed $m$ case. When $m/n$ is a fixed fraction, lower and upper bounds for $\text{Var}(\hat{V}_u)$ can be obtained.
Recall the constructed U-statistic estimation for $L^2$ loss. As the U-statistic risk estimate is constructed based on a kernel of size two, it is feasible to calculate the complete U-statistic and accomplish the proposed unbiased variance estimator $\hat{V}_u = Q(m) - Q(0)$.

A simulation comparison with nonparametric and smoothed bootstrap estimators were conducted and the results are shown in the following table.
Recall the constructed U-statistic estimation for $L^2$ loss.

As the U-statistic risk estimate is constructed based on a kernel of size two, it is feasible to calculate the complete U-statistic and accomplish the proposed unbiased variance estimator $\hat{V}_u = Q(m) - Q(0)$.

A simulation comparison with nonparametric and smoothed bootstrap estimators were conducted and the results are shown in the following table.
Unbiased Variance Estimator of $U_{L^2}$

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>Unbiased</th>
<th>Nonparametric</th>
<th>Smoothed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave. $\hat{\text{Var}}(U_{L^2})$</td>
<td>0.000464</td>
<td>0.000467</td>
<td>0.000499</td>
<td>0.000493</td>
</tr>
<tr>
<td>$SD{\hat{\text{Var}}(U_{L^2})}$</td>
<td>1.525e-4</td>
<td>1.417e-4</td>
<td>1.307e-4</td>
<td></td>
</tr>
<tr>
<td>Bias/SD${\hat{\text{Var}}(U_{L^2})}$</td>
<td>0.01967 $^1$</td>
<td>0.2470</td>
<td>0.2219</td>
<td></td>
</tr>
<tr>
<td>Computation</td>
<td>40.76 hr</td>
<td>2.88 hr</td>
<td>4.47 hr</td>
<td></td>
</tr>
</tbody>
</table>

Table: Risk Based on $L^2$ Distance.

$^1$True value is zero.

$^2$R = 200 size-100 samples were drawn independently from standard normal distribution. For each bootstrap algorithm, 1,000 resamples were considered. Bandwidth $h$ was taken to be the minimizer of $U_{L^2}$ with subsample size $n/2$. 
Notice that the unbiased variance estimator based on formula \( \hat{V}_u = Q(m) - Q(0) \) can be considered as a subsampling estimation. Therefore, the comparison between the unbiased variance estimator and bootstrap variance estimators is then equivalent to the comparison between subsampling and bootstrapping.

According to Table 1, it can be seen that subsampling estimate (i.e. the unbiased variance estimate) will have negligible bias but larger standard deviation compared with its bootstrap counterparts.

We are expecting that there should exist a compromise variance estimator between the subsampling and bootstrapping ones that has smaller variation but without large bias. Such an estimator should be a good trade-off solution for the subsampling and bootstrapping algorithms.
Reference


Qing Wang and Bruce G. Lindsay

Department of Statistics, Pennsylvania State University

Topics in U-statistics and Risk Estimation
Thank You