The Pennsylvania State University
The Graduate School

TOPICS IN U-STATISTICS AND RISK ESTIMATION

A Thesis in
Statistics
by
Qing Wang

© 2010 Qing Wang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2010

The thesis of Qing Wang was reviewed and approved* by the following:

Bruce G. Lindsay
Willaman Professor of Statistics and Department Head
Thesis Advisor

Naomi S. Altman
Associate Professor of Statistics

David R. Hunter
Associate Professor of Statistics

Runze Li
Professor of Statistics and Chair of Graduate Program

*Signatures are on file in the Graduate School.

Abstract

We consider estimating the variance of a general U-statistic when it is used as an unbiased estimator of the parameter of interest $\theta = E(K)$ where K is the kernel function. Long established results demonstrate the asymptotic normality of U-statistics and their asymptotic variance under regularity conditions. However, these asymptotic results are based on the assumption that the sample size n goes to infinity; they are not so reliable when n is not large or the kernel size m is not negligible compared with n. In addition, it can be seen that the asymptotic variance is always optimistic. On the other hand, the exact finite sample variance of a U-statistic is complicated in form when m is large. We consider an alternative approach to estimate its variance which has a relatively simple form. This variance estimator is the best unbiased estimator and therefore is applicable even for the cases that m/n is a fixed fraction. We also consider methods to estimate the U-statistic and its variance by "m out of n"resampling. Especially, two resampling schemes have been developed, both of which provide an unbiased realization of the unbiased variance estimator.

In order to further investigate the proposed method, we apply it in risk estimation under the context of nonparametric density estimation. We constructed U-statistic form risk estimators based on L^2 loss and Kullback-Leibler loss respectively, the former of which is comparable to the bagged CV score introduced in [19]. To evaluate the proposed variance estimator in risk estimation, we have carried out a simulation comparison with some bootstrap variance estimators.

Table of Contents

List of Figures

List of Tables

List of Symbols

$O(S_1, S_2)$	Sample overlap between two size-m samples, i.e. the number of elements in common between S_1 and S_2, p. 19
P_k	The set containing all pairs of size-m samples with overlap less than or equal to k, p. 19
$N(k)$	The number of pairs in P_k, p. 19
$Q(k)$	Average of products of paired kernel functions taking values from P_k, p. 19
$\hat{V}_u$	The proposed unbiased variance estimator, p. 20
$\hat{V}_{u1}$	The first proposed non-negative variance estimator, p. 28
$\hat{V}_{u2}$	The second proposed non-negative variance estimator, p. 28
$\hat{V}_{u3}$	The third proposed non-negative unbiased variance estimator, p. 29
$K^{V_u}(S_{2m})$	The kernel function for the U-statistic representation of the proposed unbiased variance estimator, where S_{2m} is a sample of size $2m$ (assume that the corresponding U-statistic is based on a kernel of size m), p. 22
$\hat{f}_h(x)$	The nonparametric kernel density estimator at point x, p. 53
ISE	Integrated squared error, p. 55
$MISE$	Mean-integrated squared error, p. 55
$AMISE$	Asymptotic mean integrated squared error, p. 57
U_{L^2}	U-statistic form risk estimator based on L^2 distance, p. 67

U_{KL} U-statistic form risk estimator based on Kullback-Leibler distance, p. 72

Acknowledgments

This master thesis would not have been completed without the help of my advisor, Doctor Bruce G. Lindsay. I am heartily grateful for his guidance, encouragement and supervision during the past year. I would also like to thank Doctor Naomi Altman and Doctor David Hunter, both of whom provided me with insightful suggestions for my research during the oral defence of my comprehensive exam.

Lastly, I owe my gratitude to all of those who provided me helpful suggestions and supported me in any aspect during the completion of this master thesis.

Qing Wang

Chapter 1

Motivation

Since Hoeffding (1948) [21] introduced the definition of U-statistics, this class of statistics has been widely used in both theoretical and applied statistical problems. Given any unbiased estimator $\hat{\theta} = K(X_1, ..., X_m)$ for the parameter of interest θ, a U-statistic can be represented as a conditional expectation of the unbiased estimator conditional on the order statistics $X_{(1)}, ..., X_{(n)}$. Notice that when we are doing nonparametric inference, the set of order statistics is the complete sufficient statistic if the underlying distribution family is large enough (see Fraser (1954) [14]). Therefore, U-statistic is the best unbiased estimator based on Rao-Blackwell Theorem. Long established results verify the asymptotic normality of U-statistics, the formula for the asymptotic variance, and also provide exact finite sample results, such as the closed form variance under regularity conditions. However, the asymptotic results are based on the assumption that the sample size n goes to infinity with the kernel fixed; they are not so reliable when n is not large or the kernel size m is not negligible compared with n. On the other hand, the exact finite sample variance of a U-statistic is complicated in form and difficult to estimate when the kernel size m is large. In order to create an asymptotic setting for these problems, we will later suppose that m grows with n. Of course, this will mean that the kernel K itself changes with sample size n.

The starting point of this study is to consider a general U-statistic with large kernel size m. We will develop an alternative approach to estimate its variance which has a relatively simple form. More specifically, the first primary goal of this thesis is to describe and investigate the best unbiased estimator of the U-

statistic variance which is applicable regardless of the magnitude of m relative to n, except that m must be $\leq n/2$. In addition, since the construction of the unbiased variance estimator has a very general form, it can be applied to the cases of degenerate U-statistics, where the standard asymptotic formula is invalid, and multivariate k-sample U-statistics. As a further step toward making the unbiased estimator more practical for large m, two resampling schemes have been developed, both of which provide unbiased realizations of the U-statistic estimator and for the unbiased estimator of its variance.

In this chapter, I will introduce the definition of a complete univariate / multivariate U-statistic and the generalized k-sample U-statistic. Then, some well-known and long established results concerning their asymptotic behaviors and exact closed form variance for finite samples will be discussed in the following.

1.1 Introduction to U-Statistics

1.1.1 One-Sample U-Statistics

Suppose F is a p-variate distribution function ($p \in N^+$)i.e.$F(x) = F(x^{(1)}, ..., x^{(p)})$, and we are considering a parameter of interest θ which can be written as a functional of the distribution function F. If furthermore, we assume that the functional θ has the form of

$$\theta(F) = \int ... \int K(x_1, x_2, ..., x_m) dF(x_1) dF(x_2) ... dF(x_m) \tag{1.1.1}$$

where $x_1, ..., x_m$ are all p-variate and K is a kernel function of m symmetric arguments ("symmetric arguments" means that the value of the kernel function does not depend on the order of its arguments), then, it can be seen that given a sample of size n ($n \geq m$) i.e. $X_1, X_2, ..., X_n$ i.i.d. from F, $K(X_1, ..., X_m)$ is an unbiased estimate of the parameter θ. In other words, $E[K(X_1, ..., X_m)] = \theta$.

However, intuition reminds us that there should be some better estimators, since $K(X_1, ..., X_m)$ does not use up the entire dataset. Based on Rao-Blackwell Theorem, conditional on the order statistics (which is a set of sufficient statistics), the conditional expectation of $K(X_1, ..., X_m)$ is the best unbiased estimator with

the form:

$$E[K(X_1, ..., X_m)|X_{(1)}, ..., X_{(n)}] = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m \le n} K(X_{i_1}, ..., X_{i_m}).$$

Back to 1948, Hoeffding [21] defined a group of statistics with the form shown above and named them as U-statistics.

Definition 1.1: Let $X_1, X_2, ..., X_n$ be a sample of 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:

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m \le n} K(X_{i_1}, ..., X_{i_m}) \tag{1.1.2}$$

In fact, the requirement that the kernel function is symmetric in its m arguments is not necessary. For instance, suppose $\tilde{K}(x_1, ..., x_m)$ is an asymmetric kernel function. Let

$$K(x_1, ..., x_m) = \frac{1}{m!} \sum_{(i_1,...,i_m) \in Perm(1,...,m)} \tilde{K}(x_{i_1}, ..., x_{i_m})$$

where the summation is taken over all the possible permutations of the m arguments. Then, K is a symmetric kernel function, and the definition of U-statistics in (1.1.2) is equivalent to the one mentioned in [21]:

$$U_n = \frac{1}{n(n-1)...(n-m+1)} \sum_{(i_1,...,i_m) \in Perm(1,...,m)} \tilde{K}(x_{i_1}, ..., x_{i_m}) \tag{1.1.3}$$

1.1.2 k-Sample U-Statistics

Now, let us consider k i.i.d. samples of random variables (vectors) from k independent distribution functions $F_1, ..., F_k$. Namely,

$$\begin{aligned} X_{1,1}, ..., X_{1,n_1} i.i.d. &\sim F_1; \\ X_{2,1}, ..., X_{2,n_2} i.i.d. &\sim F_2; \\ &... \\ X_{k,1}, ..., X_{k,n_k} i.i.d. &\sim F_k. \end{aligned}$$

In addition, assume K is a kernel function of $m = \sum_{i=1}^{k} m_i$ arguments with m_i arguments from sample i and is symmetric in terms of each m_i arguments ($1 \leq i \leq k$). Then, the generalized definition of a k-sample U-statistic is given by:

$$U_n = [\binom{n_1}{m_1}\binom{n_2}{m_2}...\binom{n_k}{m_k}]^{-1} \sum ... \sum K(x_{1,i_1}, ..., x_{1,i_{m_1}}; ...; x_{k,i_1}, ..., x_{k,i_{m_k}}) \tag{1.1.4}$$

1.1.3 Toy Examples

In the following, we are going to discuss two simple examples of U-statistics, one of which is under the one-sample framework, and the other is within the two-sample U-statistic framework.

Example 1.1 (sample variance) Let $X_1, ..., X_n$ be a i.i.d. sample from some distribution with mean μ and variance σ^2. Consider the symmetric kernel function

$$K(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2.$$

According to Definition 1.1, the corresponding U-statistic is:

$$U_n = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \frac{1}{2}(x_i - x_j)^2$$

By elementary calculation, we can simplify it in the following way:

$$\begin{aligned} U_n &= \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{1}{2}(x_i - x_j)^2 = \frac{1}{n(n-1)}(\sum_{i<j} x_i^2 - 2\sum_{i<j} x_i x_j + \sum_{i<j} x_j^2) \\ &= \frac{1}{n(n-1)}(\sum_j \sum_i x_i^2 - \sum_j \sum_i x_i x_j) \\ &= \frac{1}{n(n-1)}(n\sum_i x_i^2 - n^2 \bar{x}^2) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

which is actually equal to the sample variance S_n^2.

Note:

- The unbiased variance estimator, sample variance S_n^2, has the form as a U-statistic.

- Based on the defined U-statistic, the unbiasedness of S_n^2 can be shown in the following way:

$$E(U_n) = E[K(X_1, X_2)] = \frac{1}{2}E[(X_1 - X_2)^2] = \sigma^2$$

  which seems more straightforward than the approach via using the form of S_n^2.

- The U-statistic expression of the sample variance considers S_n^2 as a function of between-subject variation instead of the deviations from the mean as indicated in the traditional expression.

Example 1.2 (compare two sample variances) Let $X_1, ..., X_{n_1}; Y_1, ..., Y_{n_2}$ be two i.i.d. samples from two continuous distributions with mean μ_k and variance $\sigma_k^2 (k = 1, 2)$. Define a symmetric kernel function as follows

$$K(x_1, x_2; y_1, y_2) = \frac{1}{2}(x_1 - x_2)^2 - \frac{1}{2}(y_1 - y_2)^2.$$

Then, we have

$$\theta = E[K(X_1, X_2; Y_1, Y_2)] = \sigma_1^2 - \sigma_2^2$$

Thus, the corresponding two-sample U-statistic is

$$U = [\binom{n_1}{2}\binom{n_2}{2}]^{-1} \sum_{1 \le i < j \le n_1} \sum_{1 \le k < l \le n_2} K(x_i, x_j; y_k, y_l)$$

If $\theta = 0$ (namely, $\sigma_1^2 = \sigma_2^2$), then the two samples have the same variance. Therefore, the defined 2-sample U-statistic can be used to conduct the hypothesis testing with the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.

More examples of k-sample U-statistics can be found in Kowalski and Tu (2008)[23].

1.2 Established Results for U-Statistics

Recall the example of sample variance introduced in Example 1.1. It is true that under normality assumption i.e. $X_1, X_2, ..., X_n$ are i.i.d. $Normal(\mu, \sigma^2)$, the exact distribution of $(n-1)\frac{S_n^2}{\sigma^2}$ is Chi-Square with degrees of freedom $n-1$. Then, the mean and variance of S_n^2 follow immediately from the Chi-Square distribution. However, if we want to relax the restriction of normality and turn to an arbitrary distribution, it may become a lot more challenging to obtain or estimate the variance of S_n^2 in a traditional way.

As a U-statistic is an unbiased estimator of the parameter of interest (in the case of S_n^2, the parameter of interest is the population variance), exploring its variance to measure the parameter estimation is always crucial and of interest. One of the desirable properties of U-statistics is its asymptotic normality which will be stated shortly below.

Theorem 1.1: Suppose the square of the kernel function, K^2, is integrable, and let $\sigma_1^2 = Var[E(K(X_1, ..., X_m))|X_1]$ with $0 < \sigma_1^2 < \infty$, then

$$\sqrt{n}(U_n - \theta) \to N(0, m^2\sigma_1^2) \tag{1.2.1}$$

Besides, we also have the following results in Hoeffding (1948) [21] :

Theorem 1.2: Let $\phi_c(x_1, ..., x_c) = E[K(X_1, ..., X_m)|X_1 = x_1, ..., X_c = x_c]$, and $\sigma_c^2 = Var[\phi_c(X_1, ..., X_c)]; 1 \le c \le m$. Then, we have

$$Var(U_n) = \frac{1}{\binom{n}{m}} \sum_{c=1}^{m} \binom{m}{c}\binom{n-m}{m-c}\sigma_c^2 \tag{1.2.2}$$

Actually, by decomposing U_n into orthogonal terms, we can get

$$U_n = \theta + \sum_{c=1}^{m} [\frac{\binom{m}{c}}{\binom{n}{c}} \sum_{(n,c)} h^{(c)}(x_{i_1}, ..., x_{i_c})] \tag{1.2.3}$$

where

$$h^{(c)}(x_1, ..., x_c) = \phi_c(x_1, ..., x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, ..., x_{i_j}) - \theta; 1 \le c \le m. \tag{1.2.4}$$

It can be shown that $h^{(c)}(X_1, ..., X_c)(1 \leq c \leq m)$ have mean zero and are uncorrelated with each other. The form in (1.2.3) is usually referred as Hoeffding-decomposition, and the components defined in (1.2.4) are called the orthogonal terms in Hoeffding-decomposition.

If furthermore, we assume $Var[h^{(c)}(X_{i_1}, ..., X_{i_c})] = \delta_c^2$, then, $Var(U_n)$ can be written as

$$Var(U_n) = \sum_{c=1}^{m} \binom{m}{c}^2 \binom{n}{c}^{-1} \delta_c^2. \tag{1.2.5}$$

This is an alternative closed form representation of $Var(U_n)$ as compared with (1.2.2).

Theorem 1.3: The quantities $\sigma_1^2, ..., \sigma_m^2$ as defined in Theorem 1.2 satisfy the inequalities

$$0 \leq \frac{\sigma_c^2}{c} \leq \frac{\sigma_d^2}{d} \text{ if } 1 \leq c < d \leq m. \tag{1.2.6}$$

Proof:

Let

$$\gamma_c = \sigma_c^2 - \binom{c}{1}\sigma_{c-1}^2 + \binom{c}{2}\sigma_{c-2}^2 + ... + (-1)^{c-1}\binom{c}{c-1}\sigma_1^2; 1 \leq c \leq m.$$

Then, we have

$$\gamma_c \geq 0,$$

and

$$\sigma_c^2 = \gamma_c + \binom{c}{1}\gamma_{c-1} + ... + \binom{c}{c-1}\gamma_1.$$

Based on above equation, we have for $1 \leq c < d \leq m$

$$\begin{aligned} c\sigma_d^2 - d\sigma_c^2 &= c\sum_{i=1}^{d}\binom{d}{i}\gamma_i - d\sum_{i=1}^{c}\binom{c}{i}\gamma_i \\ &= \sum_{i=1}^{c}[c\binom{d}{i} - d\binom{c}{i}]\gamma_i + c\sum_{i=c+1}^{d}\binom{d}{i}\gamma_i \end{aligned}$$

Since $\gamma_i \geq 0$, and $c\binom{d}{i} - d\binom{c}{i} \geq 0$ if $1 \leq i \leq c \leq d$, every term in the two sums

of the above equation is nonnegative, which yields that

$$c\sigma_d^2 - d\sigma_c^2 \geq 0.$$

□

Theorem 1.4: The variance $Var(U_n)$ of a U-statistic $U_n = U(X_1, ..., X_n)$ (1.1.2), where $X_1, ..., X_n$ are independent and identically distributed, satisfies the inequalities

$$\frac{m^2}{n}\sigma_1^2 \leq Var(U_n) \leq \frac{m}{n}\sigma_m^2 \tag{1.2.7}$$

$nVar(U_n)$ is a decreasing function of n,

$$(n+1)Var(U_{n+1}) \leq nVar(U_n) \tag{1.2.8}$$

which takes on its upper bound $m\sigma_m^2$ for $n = m$ and tends to its lower bound $m^2\sigma_1^2$ as n increases:

$$Var(U_m) = \sigma_m^2 \tag{1.2.9}$$

$$lim_{n\to\infty} nVar(U_n) = m^2\sigma_1^2 \tag{1.2.10}$$

Although we have the asymptotic normality of U-statistics, the asymptotic results are based on the assumption that the sample size n goes to infinity with m fixed, or equivalently, the kernel size m is negligible compared with n. However, we may encounter the cases that the kernel size m is not small compared with n. One way to express this asymptotically is to say that m/n is a fixed fraction, say γ, as n goes to infinity. Of course, in such cases, the kernel $K(x_1, ..., x_m)$ must also be modelled as a function of m increasing to infinity. Note that in this case, the U-statistic could fail to be consistent. The lower bound on variance (1.2.7) can be expressed as $n \cdot \gamma^2 \cdot \sigma_1^2$. Hence, second-moment inconsistency will occur unless $\sigma_1^2 \cdot n$ goes to zero. Indeed, the usual convergence rate of $1/n$ cannot hold unless $\sigma_1^2 \cdot n^2$ converges to a constant. On the other hand, the upper bound will be $\gamma \cdot \sigma_m^2$. If the kernels are bounded as m grows, then this gives a finite variance asymptotically. Furthermore, (1.2.7) reveals that by using the asymptotic variance to estimate the U-statistic variance, we are always optimistic.

If m/n is a fixed fraction, the standard asymptotic results are no longer reliable, and the asymptotic variance in (1.2.10) may not be a good estimate of the variance of a U-statistic. On the other hand, the exact form of the variance is computationally intensive, especially when the kernel size m is large.

Chapter 2

Extensions of U-Statistics

2.1 Fixed m Incomplete U-Statistics

In practice, when the sample size n and the kernel size m are both large, it is not efficient to compute the complete U-statistic (1.1.2), since the number of exhaustive combinations of size-m samples out of n (i.e. $\binom{n}{m}$) is enormous. As a result, researchers seek to use a subset of the size-m samples to compute a corresponding statistic with a similar form to a U-statistic which also harbours good properties.

2.1.1 General Fixed m Incomplete U-Statistics

Blom (1976) [1] pointed out that due to the strong dependency of many of the size-m samples used to construct U_n (1.1.2), it seems reasonable to consider less than $N = \binom{n}{m}$ terms without losing too much information. Furthermore, he defined a general class of surrogates of complete U-statistics, called "**incomplete U-statistics**".

$$U_{inc} = \frac{1}{B}\sum_{i=1}^{B} K(X_{i_1}, ..., X_{i_m}) \tag{2.1.1}$$

In this definition, B can be any positive integer less than or equal to N, and it is also allowed to take repetitions of size-m samples out of n.

Now, suppose $n = m \cdot k$. If we randomly divide the n observations into k non-overlapped samples of size m, the k subsamples are independent. Accordingly, we can define an incomplete U-statistic of k independent terms.With this construction,

we have

$$Var(U_{inc}) = \frac{1}{k}Var[K(X_1, ..., X_m)] = \frac{1}{k}\sigma_m^2 \text{ where } k = n/m.$$

Recall that for the complete U-statistic, we have $\frac{m^2}{n}\sigma_1^2 \leq Var(U_n) \leq \frac{m}{n}\sigma_m^2$ (1.2.7). Therefore, an incomplete U-statistic may be as competitive as its complete counterpart, as is seen by comparing their variances. Furthermore, the asymptotic efficiency of U_{inc} compared with U_n is $m\sigma_1^2/\sigma_m^2$. This ratio may be close to 1 and will be exactly equal to 1 if the kernel function K is an arithmetic mean of a size-1 kernel i.e. $K(x_1, ..., x_m) = \frac{1}{m}\sum_{i=1}^{m}\phi(x_i)$, in which case U_n and U_{inc} are identical for each n.

2.1.2 A Special Case–Reduced U-Statistics

Brown and Kildea (1978) [3] put forward a definition of a **reduced U-statistic** which averages the symmetric kernel functions K under certain restrictions. Actually, a reduced U-statistic is just a special case of Blom's incomplete U-statistic which was referred as a balanced U-statistic in [1]. As a general result for balanced U-statistics, Blom [1] showed that the upper and lower bounds for variance of U_n (1.2.7) also hold for the balanced U-statistics. In addition, a reduced U-statistic is asymptotically normally distributed as its complete counterpart.

2.1.2.1 Reduced U-Statistics of Order 2

For a simple illustration, let us first consider the case that the kernel size m equals to 2.

Let C_t be a set of pairs (i, j) with $1 \leq i < j \leq n$ such that each positive integer less than or equal to n appears in exactly $2t$ pairs in C_t, and let

$$P_t = \{(X_i, X_j) : (i, j) \in C_t\}.$$

Then, a reduced U-statistic for $m = 2$ case is defined as

$$U_{reduced} = \frac{1}{nt}\sum_{P_t} K(X_i, X_j). \tag{2.1.2}$$

It is easily seen that P_t only contains nt size-2 samples in total (because each integer $\leq n$ appears $2t$ times, then there are $2t \cdot n$ integers in total which is equivalent to nt pairs). Therefore, as $n \to \infty$, the computational effort required to calculate $U_{reduced}$ is negligible compared with that needed to compute U_n, the complete U-statistic. Furthermore, notice that each pair in P_t can be seen as a size-2 sample drawn from $X_1, ..., X_n$ independently from others. Therefore, $U_{reduced}$ is also an unbiased estimator of the parameter of interest θ.

In their paper, they also showed that for fixed m, the reduced U-statistics have asymptotic efficiency compared with the corresponding U-statistics under certain conditions. In addition, the reduced U-statistics are also asymptotically normal. On this point, the reduced U-statistics shed some light on computational efficiency by preserving the property of asymptotic normality.

The following notations and results are based on Brown and Kildea (1978) [3].

Notation:

In some applications, the kernel function $K(\cdot)$ may depend on sample size n. In order to include this case, we use $K_n(\cdot)$ to represent the kernel function which is possibly depending on n, and let

$$\begin{aligned} W_n &= \sum_{C_t} K_n(X_i, X_j) && (2.1.3) \\ \theta_n &= E[K_n(X_1, X_2)] && (2.1.4) \\ \sigma_n^2 &= Var[K_n(X_1, X_2)] && (2.1.5) \\ \text{and } \rho_n\sigma_n^2 &= Cov[K_n(X_1, X_2), K_n(X_1, X_3)] && (2.1.6) \end{aligned}$$

It can be shown that

$$Var(W_n) = nt\sigma_n^2[1 + 2(2t-1)\rho_n]$$

Theorem 2.1 If the finite limits $\sigma^2 = lim_{n\to\infty}\sigma_n^2$ and $\rho\sigma^2 = \lim_{n\to\infty} \rho_n\sigma_n^2$ both exist, if $\sigma^2 > 0$, and if

$$\{K_n(X_1, X_2) - \theta_n, n \geq 1\} \text{is uniformly square integrable}$$

then $(nt)^{-\frac{1}{2}}(W_n - nt\theta_n)$ converges in distribution as $n \to \infty$ to a normal law with mean zero and variance $\sigma^2[1 + 2(2t-1)\rho]$.

Corollary 2.1: When $\rho\sigma^2 > 0$, the estimator $\{(nt)^{-1}W_n, n \geq 1\}$ of $\{\theta_n, n \geq 1\}$ has asymptotic efficiency $2t\rho\{\frac{1}{2} + (2t-1)\rho\}^{-1}$, relative to the corresponding complete U-statistic estimators, as $n \to \infty$.

It can be seen from Corollary 2.1 that for large enough t the asymptotic efficiency of the reduced U-statistic may be arbitrarily close to 1. From another aspect, the best case for the reduced U statistics is when ρ is large enough. We will later show that ρ is no greater than $1/2$ (2.1.8), but the efficiency is one at this value; the worst case may happen when $\rho = 0$, which results in a zero asymptotic efficiency.

Theorem 2.2: Under the conditions and notation of Theorem 2.1, now let $W_n^{(1)}, \cdots, W_n^{(p)}$ be reduced U-statistics corresponding to the sets of pairs ${C_{t_1}}^{(1)}, \cdots, {C_{t_p}}^{(p)}$. Then $\{W_n^{(1)}, \cdots, W_n^{(p)}\}$, when suitably normalized, converges in distribution as $n \to \infty$ to a multivariate normal distribution.

Corollary 2.2: Let $\{{C_{t_\alpha}}^{(\alpha)}, 1 \leq \alpha \leq p\}$ be disjoint. Then, for $\alpha \neq \beta$,

$$Cov(W_n^{(\alpha)}, W_n^{(\beta)}) = 4nt_\alpha t_\beta \rho_n \sigma_n^2$$

and the covariance structure of the limit distribution in Theorem 2.1 is determined.

2.1.2.2 Representation of $E[K_n(S_1)K_n(S_2)|\#$ overlaps]

In fact, based on Hoeffding-decomposition technique, we can express

$$E[K_n(S_1)K_n(S_2)|\#\text{ overlaps between } S_1 \text{ and } S_2]$$

in terms of the variances of the orthogonal terms in Hoeffding-decomposition through the following way.

Denote

$$\phi_c(x_1, ..., x_c) = E[K_n(X_1, ..., X_m)|X_1 = x_1, ..., X_c = x_c]$$

where $1 \leq c \leq m$.

Let

$$\begin{aligned} h^{(1)}(x_1) &= \phi_1(x_1) \\ h^{(2)}(x_1, x_2) &= \phi_2(x_1, x_2) - \phi_1(x_1) - \phi_1(x_2) - \theta_n \\ &... \\ h^{(c)}(x_1, ..., x_c) &= \phi_c(x_1, ..., x_c) - \sum_{j=1}^{c-1} \sum_{(c,j)} h^{(j)}(x_{i_1}, ..., x_{i_j}) - \theta_n \end{aligned}$$

Therefore, we have

$$\phi_c(x_1, ..., x_c) = \sum_{j=1}^{c} \sum_{(c,j)} h^{(j)}(x_{i_1}, ..., x_{i_j}) + \theta_n; 1 \leq c \leq m$$

And it can be shown that (Lee (1990) [24])

$$\begin{aligned} E[h^{(j)}(X_{i_1}, ..., X_{i_j})] &= 0 \\ Cov[h^{(j)}(X_{i_1}, ..., X_{i_j}), h^{(j\prime)}(X_{i_1}, ..., X_{i_{j\prime}})] &= 0 \text{ if } j \neq j\prime \\ Cov[h^{(j)}(X_{i_1}, ..., X_{i_j}), h^{(j)}(X_{l_1}, ..., X_{l_j})] &= 0 \text{ if } (X_{i_1}, ..., X_{i_j}) \cap (X_{l_1}, ..., X_{l_j}) \neq \emptyset \end{aligned}$$

That is, $h^{(j)}(x_{i_1}, ..., x_{i_j}); 1 \leq j \leq m$ have mean zero and are uncorrelated with each other. Here we call them the orthogonal terms in Hoeffding-decomposition as mentioned earlier.

Lemma 2.1: If we denote $Var[h^{(j)}(X_{i_1}, ..., X_{i_j})] = \delta_j^2$ where $1 \leq j \leq k \leq m$, k is the number of overlaps between two size-m samples (S_1, S_2), then we have

$$E[K_n(S_1)K_n(S_2)|\# \text{ overlaps } = k] = \sum_{j=1}^{k} \binom{k}{j} \delta_j^2 + \theta_n^2 \tag{2.1.7}$$

Proof:

$$
\begin{aligned}
& E[K_n(S_1)K_n(S_2)|\#\text{ overlaps between } S_1 \text{ and } S_2 = k] \\
= \; & E\{E[K_n(S_1)K_n(S_2)|\text{overlaps} = (X_1, ..., X_k)]\} \\
= \; & E\{E[K_n(S_1)|X_1, ..., X_k]E[K_n(S_2)|X_1, ..., X_k]\} \\
= \; & E[\phi_k(X_1, ..., X_k)\phi_k(X_1, ..., X_k)] \\
= \; & E[(\sum_{j=1}^{k}\sum_{(k,j)} h^{(j)}(x_{i_1}, ..., x_{i_k}) + \theta_n)(\sum_{l=1}^{k}\sum_{(k,l)} h^{(l)}(x_{i_1}, ..., x_{i_l}) + \theta_n)] \\
= \; & \sum_{j=1}^{k}\sum_{l=1}^{k}\sum_{(k,j)}\sum_{(k,l)} E[h^{(j)}(x_{i_1}, ..., x_{i_j})h^{(l)}(x_{i_1}, ..., x_{i_l})] \\
+ \; & 2\theta_n \sum_{j=1}^{k}\sum_{l=1}^{k} E[h^{(j)}(x_{i_1}, ..., x_{i_j})] + \theta_n^2 \\
= \; & \sum_{j=1}^{k}\sum_{(k,j)} E[(h^{(j)}(x_{i_1}, ..., x_{i_j}))^2] + \theta_n^2 \\
= \; & \sum_{j=1}^{k}\sum_{(k,j)} Var[h^{(j)}(x_{i_1}, ..., x_{i_j})] + \theta_n^2 \\
= \; & \sum_{j=1}^{k} \binom{k}{j} \delta_j^2 + \theta_n^2
\end{aligned}
$$

□

For the case of $m = 2$,

$$
\begin{aligned}
& Cov[K_n(X_1, X_2), K_n(X_1, X_3)] \\
= \; & E[K_n(X_1, X_2)K_n(X_1, X_3)] - \{E[K_n(X_1, X_2)]\}^2 \\
= \; & \delta_1^2 + \theta_n^2 - \theta_n^2 \\
= \; & \delta_1^2
\end{aligned}
$$

since the number of overlaps between the two size-2 samples equals to 1 in this case.

Besides, based on the result of A.J. Lee (1990) [24], we also have

$$\sigma_c^2 = \sum_{j=1}^{c} \binom{c}{j} \delta_j^2; 1 \leq c \leq m$$

where σ_c^2 is defined to be $Var[\phi_c(X_1, ..., X_c)]$, and $\phi_c(x_1, ..., x_c)$ is defined at the beginning of Section 2.1.2.2 (pp. 13).

Therefore, for $m = 2$, $\sigma_2^2 = 2\delta_1^2 + \delta_2^2$. In addition, formula (2.1.6) yields that

$$0 \leq \rho_n = \frac{\delta_1^2}{2\delta_1^2 + \delta_2^2} \leq \frac{1}{2} \tag{2.1.8}$$

The upper bound occurs when $\delta_2 = 0$ while δ_1^2 is any positive value.

2.1.2.3 Reduced U-Statistics of Order $m > 2$

We next generalize the idea of reduced U-statistics of order 2 to higher order case. Consider a symmetric kernel function K with m arguments. Define

$$C_t = \{(i_1, ..., i_m) : \text{each integer} \leq n \text{ appears in exactly } mt \text{ pairs in } C_t\}$$

And, let P_t be the number of size-m samples in C_t. Since there are nt size-m samples in P_t, the generalized reduced U-statistic can be defined as

$$U_{reduced} = \frac{1}{nt} \sum_{P_t} K_n(X_{i_1}, ..., X_{i_m}) \tag{2.1.9}$$

Definition (2.1.2) can be generalized to $m = M; M \in \mathbf{N}^+$ case naturally, and the results stated in Theorem 2.1, 2.2 also hold for reduced U-statistics with orders $m > 2$.

2.2 Random Subsampling and Incomplete U-Statistics

Although a reduced U-statistic is less computational expensive and also has desirable properties, its construction is still under the restriction of fixed-m framework,

while our goal is to deal with problems with fixed m/n in practice. When m/n is a fixed fraction, $\binom{n}{m}$ will become even more sizeable for large n case. So, finding efficient alternatives to complete U-statistics turns out to be a practical issue under this scenario.

A simple idea is to construct an incomplete U-statistic by random resampling:

$$\tilde{U}_B = \frac{1}{B}\sum_{b=1}^{B} K(\tilde{S}_b) \tag{2.2.1}$$

where $\tilde{S}_1, ..., \tilde{S}_B$ are drawn randomly and independently from $\mathbf{S} = \{(X_{i_1}, ..., X_{i_m}) : 1 \leq i_1 < ... < i_m \leq n\}$. This methodology is often called "subsampling".(See the book by Politis et al. (1999) [29].)

If we consider $\mathbf{S}$ to be the set of all possible samples of size m with $1 \leq i_1 < ... < i_m \leq n$, then, $\tilde{S}_1, ..., \tilde{S}_B$ is just a subsample of size B from $\mathbf{S}$. It is obvious that for i.i.d. case, $\tilde{U}_B$ is a subsampling unbiased estimator of the complete U statistic, and so also an unbiased estimator of θ. And the computation of $\tilde{U}_B$ is negligible compared with the number of steps needed to compute U_n. Therefore, $\tilde{U}_B$ could be a good alternative estimator of θ for the case that m/n is a fixed fraction.

Recall the closed form variance for U-statistic in (1.2.2), i.e.

$$Var(U_n) = \sum_{c=1}^{m} \frac{\binom{m}{c}\binom{n-m}{m-c}}{\binom{n}{m}}\sigma_c^2$$

where $\sigma_c^2 = Var\{E[K(X_1, ..., X_m)|X_1, ..., X_c]\}$,$1 \leq c \leq m$.

Define the overlap variable $X = O(S_1, X_2)$, the number of overlaps between two size-m samples. It follows a hypergeometric distribution with probability mass

$$P(X = k) = \frac{\binom{m}{c}\binom{n-m}{m-c}}{\binom{n}{m}}$$

which is just the weight of σ_c^2 in the closed form variance.

Based on hypergeometric$(n; m, m)$ distribution, we have

$$\begin{aligned} E(X) &= \frac{m^2}{n}, \\ Var(X) &= \frac{m^2(n-m)^2}{n^2(n-1)}. \end{aligned}$$

Now, consider $\frac{X}{m}$, we have

$$\begin{aligned} Var(\frac{X}{m}) &= \frac{1}{m^2}\frac{m^2(n-m)^2}{n^2(n-1)} \\ &= \frac{(n-m)^2}{n^2(n-1)} \\ &= \frac{(1-m/n)^2}{n-1} \\ &\to 0 \\ E(\frac{X}{m}) &= \frac{1}{m}\frac{m^2}{n} \\ &= \frac{m}{n} \\ &= \gamma \end{aligned}$$

Therefore, by $Var(\frac{X}{m}) \to 0, E(\frac{X}{m}) = \gamma$, we have $\frac{X}{m} \to \gamma$ **in probability**. In other words, as m, n go to infinity with $m/n = \gamma$, the overlap distribution tends to be degenerate at the "mean overlap" i.e. $m\gamma$.

As a result, the weight of $\sigma^2_{m\gamma}$ goes to 1 asymptotically, meaning that when $m/n = \gamma$ is a fixed fraction, the asymptotic variance of the U-statistic is going to $\sigma^2_{m\gamma}/n$.

Recall (1.2.6) that $\frac{\sigma_a^2}{a} \leq \frac{\sigma_b^2}{b}, 1 \leq a < b \leq m$, we have $\sigma_a^2 \leq \sigma_b^2, 1 \leq a < b \leq m$. This has implications on sampling: in terms of incomplete U-statistics, we may seek to construct designed samples with the smallest possible overlaps in order to achieve asymptotic efficiency.

So far, we have found surrogates for U-statistics for both fixed m and fixed m/n cases. However, as mentioned above, we are also aiming to find a reliable variance estimator when the variance calculation works poorly. Ideally, we would be able to construct a variance estimator which is applicable even for the case that m/n is a fixed fraction.

Chapter 3

The Unbiased Variance Estimator

3.1 Construction of the Unbiased Variance Estimator

We now demonstrate how one can construct an unbiased estimator of the variance of an arbitrary U-statistic, assuming that $m \leq n/2$.

Consider a U-statistic defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_i K(S_i), \text{ where } S_i \text{ is a size} - m \text{ sample out of i.i.d. } X_1, ..., X_n \tag{3.1.1}$$

Define the sample overlap

$$O(S_1, S_2) = \text{ number of elements in common between } S_1 \text{ and } S_2 \tag{3.1.2}$$

Let

$$P_k = \{(S_1, S_2) : O(S_1, S_2) \leq k\} \tag{3.1.3}$$

$$N_k = \text{ number of pairs in } P_k \tag{3.1.4}$$

Let

$$Q(k) = \frac{1}{N_k} \sum_{P_k} K(S_1)K(S_2) \tag{3.1.5}$$

Note that

$$E(Q(0)) = E(U_n)^2, Q(m) = U_n^2.$$

Therefore,

$$E[Q(m) - Q(0)] = E(U_n^2) - E(U_n)^2 = Var(U_n).$$

That is, Q(m)-Q(0) is an unbiased estimate of $Var(U_n)$.

Theorem 3.1: Suppose U_n is a U-statistic with a kernel K of size m, $m \leq n/2$. Denote

$$\hat{V}_u = Q(m) - Q(0) \tag{3.1.6}$$

where $Q(m)$ and $Q(0)$ are defined in (3.1.5). Then, $\hat{V}_u$ is an unbiased estimator of $Var(U_n)$. Furthermore, it is a function of the order statistics and so is the best unbiased estimator of $Var(U_n)$.

Example 3.1: Consider a data set $x_1, ..., x_n$ from some distribution with mean μ and variance σ^2. Assume the parameter of interest is $\theta = \mu$. Let $K(x) = x$ be the kernel function, which results in a U-statistic

$$U_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Based on the defined unbiased variance estimate (3.1.6), we have

$$\begin{aligned} Q(m) &= U_n^2 = \bar{X}^2 \\ Q(0) &= \frac{1}{\binom{n}{2}} \sum_{i<j} X_i X_j \end{aligned}$$

It can be shown that

$$\hat{V}_u = Q(m) - Q(0) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n} S_n^2,$$

and,

$$E(\hat{V}_u) = \frac{1}{n} E(S_n^2) = \frac{\sigma^2}{n}.$$

Note:

- To our knowledge, the proposed unbiased variance estimator (3.1.6) is a new method for variance estimation. The thesis of Folsom (1986) [13] from University of North Carolina at Chapel Hill had this result in a more complex form, but further results seem to be absent.

- Because the construction of the unbiased variance estimate is only based on sample overlaps, its validity does not depend on whether the U-statistic is degenerate or not. Therefore, the form of the unbiased variance estimator $\hat{V}_u$ can be applied to both degenerate and non-degenerate cases.

Proposition 3.1: The unbiased variance estimator defined in (3.1.6) can be generalized to the k-sample U-statistic case (1.1.4).

First of all, let us consider a 2-sample U-statistic based on two independent populations with distribution functions F and G. Denote the two samples as $X_1, ..., X_{n_1}; Y_1, ..., Y_{n_2}$ where $X_1, ..., X_{n_1}$ i.i.d. $\sim F$ and $Y_1, ..., Y_{n_2}$ i.i.d. $\sim G$. The 2-sample U-statistics defined by the kernel function K has the form:

$$U = \frac{1}{\binom{n_1}{m_1}} \frac{1}{\binom{n_2}{m_2}} \sum_{1 \le i_1 < ... < i_{m_1} \le n_1} \sum_{1 \le j_1 < ... < j_{m_2} \le n_2} K(X_{i_1}, ..., X_{i_{m_1}}; Y_{j_1}, ..., Y_{j_{m_2}})$$

Assume $m = m_1 + m_2$, and S_1, S_2 are two samples of size m with m_1 components from distribution F and m_2 components from distribution G. Then, in this case the sample overlap can be defined in a similar fashion such that $\{O(S_1, S_2) \le k\}$ contains all pairs of size-m samples (S_1, S_2) where the number of common elements between S_1 and S_2 is at most k. Compared with the one-sample case, the only difference here is that the overlaps can result from two groups of components corresponding to the two distributions. In addition, denote P_k as the set containing all the pairs of size-m samples with overlaps at most k; $N(k)$ is the number of pairs in P_k.

$$Q(0) = \frac{1}{N_0} \sum_{P_0} K(S_1) K(S_2) \tag{3.1.7}$$

$$Q(m) = \frac{1}{N_m} \sum_{P_m} K(S_1) K(S_2) \tag{3.1.8}$$

Similar to the one-sample U-statistics case, we have

$$E[Q(0)] = E[K(S_1)K(S_2)] = E[K(S_1)]E[K(S_2)] = U^2, Q(m) = U^2.$$

Therefore, $\hat{Vu} = Q(m) - Q(0)$ is an unbiased estimator of $Var(U_n)$ for the 2-sample U-statistic case.

The generalization to 2-sample case can also be extended to k-sample U-statistics ($k \geq 2$) in the same fashion.

3.2 The U-Statistic Form of $\hat{V}_u$ and Its Asymptotic Behavior

In fact, the unbiased variance estimate, $\hat{V}_u = Q(m) - Q(0)$, can be re-expressed as a complete U-statistic with a kernel K^{V_u} of size $2m$.

Proposition 3.2: Consider samples of size $2m$, S_{2m}. Let S_a, S_b be two subsamples of size m out of S_{2m}. Define

$$K^{V_u}(S_{2m}) = K_m^{V_u}(S_{2m}) - K_0^{V_u}(S_{2m}) \tag{3.2.1}$$

where

$$K_0^{V_u}(S_{2m}) = \frac{\binom{n}{2m}}{\binom{n}{m}\binom{n-m}{m}} \sum_{S_a, S_b \subset S_{2m}} K(S_a)K(S_b)I\{S_a \cap S_b = \emptyset\}, \tag{3.2.2}$$

$$K_m^{V_u}(S_{2m}) = \frac{\binom{n}{2m}}{\binom{n}{m}^2} \sum_{S_a, S_b \subset S_{2m}} K(S_a)K(S_b)\omega(a,b), \tag{3.2.3}$$

$\omega(a,b) = 1/n(a,b)$, and $n(a,b) = \binom{n-(2m-k)}{k}$ where $k = O(S_a, S_b)$ i.e. $n(a,b)$ is the number of different size-$2m$ samples S_{2m} in which S_a, S_b are subsets. Then,

$$\hat{V}_u = Q(m) - Q(0) = \frac{1}{\binom{n}{2m}} \sum_{S_{2m}} K^{V_u}(S_{2m}) \tag{3.2.4}$$

Proof:

Firstly, we are going to write $Q(m)$ as a U-statistic with kernel size $2m$. In other words, we want to find the weights, $\omega(a,b)$, such that the following formula

$$\binom{n}{2m}^{-1}\sum_{S_{2m}}[\sum_{S_a,S_b\subset S_{2m}} K(S_a)K(S_b)w(a,b)]$$

is proportional to

$$\frac{1}{\binom{n}{m}^2}[\sum_{S_a,S_b\subset S_n} K(S_a)K(S_b)].$$

In the second formula, each pair of S_a, S_b appears once inside the bracketed sum. In the first formula, each pair of S_a, S_b appears once or zero times inside the bracket and will appear once in the outer sum for each size-$2m$ sample that contains that pair. Now, denote $n(a,b)$ as the number of different size-$2m$ samples S_{2m} in which S_a, S_b are subsets. Then, we can set $\omega(a,b) = 1/n(a,b)$, and then, with adjusting the initial constants, the two formulas will be equal. That is, let

$$K_m^{V_u}(S_{2m}) = \frac{\binom{n}{2m}}{\binom{n}{m}^2}\sum_{S_a,S_b\subset S_{2m}} K(S_a)K(S_b)\omega(a,b)$$

we have

$$Q(m) = \frac{1}{\binom{n}{2m}}\sum_{S_{2m}} K_m^{V_u}(S_{2m})$$

When it comes to $Q(0)$ i.e. $\frac{1}{\sum I\{S_a\cap S_b=\emptyset\}}\sum_{S_a,S_b\subset S_{2m}} K(S_a)K(S_b)I\{S_a\cap S_b=\emptyset\}$, for each non-overlapped pair S_a, S_b there is only one S_{2m} that contains them. So, $n(a,b) = 1$. Similarly, we can define

$$K_0^{V_u}(S_{2m}) = \frac{\binom{n}{2m}}{\sum I\{S_a\cap S_b=\emptyset\}}\sum_{S_a,S_b\subset S_{2m}} K(S_a)K(S_b)I\{S_a\cap S_b=\emptyset\},$$

where $\sum I\{S_a\cap S_b=\emptyset\} = \binom{n}{m}\binom{n-m}{m}$.

Then, we have

$$Q(0) = \frac{1}{\binom{n}{2m}}\sum_{S_{2m}} K_0^{V_u}(S_{2m})$$

If we denote $K^{V_u}(S_{2m}) = K_m^{V_u}(S_{2m}) - K_0^{V_u}(S_{2m})$, then the unbiased variance estimate $\hat{V}_u$ has the form as a U-statistic:

$$\hat{V}_u = Q(m) - Q(0) = \frac{1}{\binom{n}{2m}} \sum_{S_{2m}} K^{V_u}(S_{2m})$$

□

Example 3.2: In the previous example (Example 3.1), we have that $n(a,b) = 1$, if $a \neq b$; $n(a,b) = n-1$, if $a = b$. Therefore,

$$\begin{aligned} K_m^{V_u}(x_1, x_2) &= \frac{\binom{n}{2}}{n^2}\left(\frac{x_1^2 + x_2^2}{n-1} + 2x_1x_2\right), \\ K_0^{V_u}(x_1, x_2) &= \frac{\binom{n}{2}}{n(n-1)} 2x_1x_2, \end{aligned}$$

and

$$\begin{aligned} K^{V_u}(x_1, x_2) &= K_m^{V_u}(x_1, x_2) - K_0^{V_u}(x_1, x_2), \\ &= \binom{n}{2}\left[\frac{1}{n^2}\left(\frac{x_1^2 + x_2^2}{n-1} + 2x_1x_2\right) - \frac{1}{n(n-1)} 2x_1x_2\right] \\ &= \binom{n}{2} \frac{1}{n^2(n-1)} (x_1 - x_2)^2 \\ &= \frac{1}{n} \frac{(x_1 - x_2)^2}{2}. \end{aligned}$$

As a result,

$$\begin{aligned} Q(m) - Q(0) &= \frac{1}{\binom{n}{2}} \sum_{i<j} K^{V_u}(X_i, X_j) \\ &= \frac{1}{n} \cdot \frac{1}{\binom{n}{2}} \sum_{i<j} \frac{(X_i - X_j)^2}{2} \\ &= \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \frac{1}{n} S_n^2 \end{aligned}$$

which is the same as the result in Example 3.1 based on formula (3.1.6).

As a U-statistic, $\hat{V}_u$ has asymptotic normality under certain regularity conditions in the fixed m case (Theorem 1.1). When m/n is a fixed fraction, lower and upper bounds of $Var(\hat{V}_u)$ can be obtained based on (1.2.7) accordingly.

3.3 Negative Values of $\hat{V}_u$ and Proposed Fix-ups

It can be seen that although $E(Q(k)-Q(0)) \geq 0$ (because the fact that pairs with some overlaps should be more positively correlated than pairs with no overlaps), it is numerically possible that $Q(k)-Q(0) \leq 0$,$k > 1$. A simple example illustrating this phenomenon will be shown below.

Example 3.3: Consider $m = 2$ and $n = 4$, i.e. the original data set is $x_1, ..., x_4$, and K is a kernel function of order 2. Suppose that $K(x_1, x_2) = K(x_3, x_4) = 1$, $K(x_1, x_3) = K(x_1, x_4) = K(x_2, x_3) = K(x_2, x_4) = 0$. Then,

$$\hat{V}_u = Q(m) - Q(0) = U_n^2 - Q(0) = (\frac{1}{3})^2 - \frac{1}{3} = -\frac{2}{9}.$$

The potential negative problem results from larger $Q(0)$ value compared with $Q(m)$. That is, it is possible to obtain negative estimate of $Var(U_n)$ by $\hat{V}_u$, which is unreasonable. Therefore, some adjustments to $Q(m)-Q(0)$ should be considered. The following two lemmas lead to a natural adjustment to the unbiased variance estimate.

3.3.1 The First Proposal

Lemma 3.1:

$$E[Q(m) - Q(k)] \leq E[Q(m) - Q(k-1)] \leq Var(U_n) \text{ for all } k \tag{3.3.1}$$

Proof:

Let A_k denote the set of all pairs of size-m samples with overlaps exactly equal to k, and let n_k be the number of pairs in A_k, $0 \leq k \leq m$. It is easily seen that $P_k = \cup_{l=0}^{k} A_k$, and $N_k = \sum_{l=0}^{k} n_k$, where P_k, N_k are defined in (3.1.3), (3.1.4).

Given $1 \leq k \leq m$, let's first consider $Q(k) - Q(k-1)$. By definition in (3.1.5), we have

$$\begin{aligned} Q(k) - Q(k-1) &= \frac{1}{N_k}[\sum_{A_0} K(S_i)K(S_j) + ... + \sum_{A_k} K(S_i)K(S_j)] \\ &- \frac{1}{N_{k-1}}[\sum_{A_0} K(S_i)K(S_j) + ... + \sum_{A_{k-1}} K(S_i)K(S_j)] \\ &= \frac{-n_k}{N_k \cdot N_{k-1}}[\sum_{A_0} K(S_i)K(S_j) + ... + \sum_{A_{k-1}} K(S_i)K(S_j)] \\ &+ \frac{1}{N_k}\sum_{A_k} K(S_i)K(S_j) \end{aligned}$$

Recall the conclusion in Lemma 2.1, that is,

$$E[K(S_i)K(S_j)|\#\text{ overlaps} = k] = \sum_{j=1}^{k} \binom{k}{j}\delta_j^2 + \theta^2$$

where $\delta_j^2 = Var[h^{(j)}(X_{i_1}, ..., X_{i_j})]$.

We have

$$\begin{aligned} E[Q(k) - Q(k-1)] &= \frac{-n_k}{N_k N_{k-1}}[n_0\theta^2 + n_1(\delta_1^2 + \theta^2) + ... + n_{k-1}(\sum_{j=1}^{k-1}\binom{k-1}{j}\delta_j^2 + \theta^2)] \\ &+ \frac{n_k}{N_k}(\sum_{j=1}^{k}\binom{k}{j}\delta_j^2 + \theta^2) \\ &= \frac{n_k}{N_k}\frac{\sum_{l=1}^{k-1} n_l}{N_{k-1}}\sum_{j=1}^{k}\binom{k}{j}\delta_j^2 - \frac{n_k}{N_k N_{k-1}}\sum_{l=1}^{k-1}[n_l\sum_{j=1}^{l}\binom{l}{j}\delta_j^2] \\ &+ \frac{-n_k}{N_k N_{k-1}}\sum_{j=1}^{k-1} n_j \cdot \theta^2 + \frac{n_k}{N_k}\theta^2 \\ &= \frac{n_k}{N_k N_{k-1}}\sum_{l=1}^{k-1} n_l[\sum_{j=1}^{k}\binom{k}{j}\delta_j^2 - \sum_{j=1}^{l}\binom{l}{j}\delta_j^2] \end{aligned}$$

Notice that each term $\sum_{j=1}^{k}\binom{k}{j}\delta_j^2 - \sum_{j=1}^{l}\binom{l}{j}\delta_j^2 \geq 0$ for $1 \leq l \leq k-1; 1 \leq k \leq m$. Therefore, we have

$$E[Q(k) - Q(k-1)] \geq 0$$

The fact that

$$E[Q(m) - Q(k-1)] - E[Q(m) - Q(k)] = E[Q(k) - Q(k-1)],$$

yields the result in Lemma 3.1 immediately.

□

Lemma 3.2:

$$Q(m) - Q(m-1) = \frac{1}{N(N-1)} \sum [K(S) - U_n]^2 := S_U^2 \tag{3.3.2}$$

where $N = \binom{n}{m}$.

Proof:

Since

$$\begin{aligned} Q(m-1) &= \frac{N_m}{N_{m-1}}[Q(m) - \frac{1}{N_m} \sum_{A_m} K(S_i)(S_j)] \\ &= Q(m) + \frac{N_m - N_{m-1}}{N_{m-1}} Q(m) - \frac{1}{N_{m-1}} \sum_{i=1}^{N} K^2(S_i), \end{aligned}$$

we have

$$\begin{aligned} Q(m) - Q(m-1) &= \frac{1}{N_{m-1}} \sum_{i=1}^{N} K^2(S_i) - \frac{N_m - N_{m-1}}{N_{m-1}} Q(m) \\ &= \frac{1}{N_{m-1}} \sum_{i=1}^{N} K^2(S_i) - \frac{n_m}{N_{m-1}} Q(m) \\ &= \frac{1}{N_{m-1}} [\sum_{i=1}^{N} K^2(S_i) - n_m U_n^2]. \end{aligned}$$

Notice that $n_m = \binom{n}{m} = N$, $N_{m-1} = N(N-1)$, and $U_n = \frac{1}{N} \sum_{i=1}^{N} K(S_i)$. Therefore,

$$Q(m) - Q(m-1) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (K(S_i) - U_n)^2.$$

□

It is interesting to notice that S_U^2, as an estimator of $Var(U_n)$, is biased downwards from the first lemma and is nonnegative from the second lemma. (Remember that $Q(m) - Q(0)$ as an unbiased estimator may result in negative estimate.) Moreover, it is strictly positive unless all $K(S)$ are degenerate at U_n, i.e. $K(S) = U_n$ for any S. This suggests a simple fix-up to the unbiased estimator $\hat{V}_u$:

$$\hat{V}_{u1} = max\{S_U^2, \hat{V}_u\} \tag{3.3.3}$$

It is true that $\hat{V}_{u1}$ is nonnegative and must have positive bias, since $\hat{V}_{u1} \geq \hat{V}_u$ and $\hat{V}_u$ is unbiased. Furthermore, $\hat{V}_{u1}$ is forced to be larger than the estimator one would use in the case when the $K(S)$ values are highly independent. (When the $K(S)$'s are highly independent, $Var(U_n) \approx \frac{1}{\binom{n}{m}} Var[K(S)] \leq S_U^2 \leq \hat{V}_{u1}$.)

Example 3.4: In the previous example (Example 3.3), we find that

$$\hat{V}_{u1} = max\{\frac{1}{6 \times 5}[2 \times (1 - \frac{1}{3})^2 + 4 \times (0 - \frac{1}{3})^2], -\frac{2}{9}\} = \frac{2}{45}$$

which is nonnegative.

3.3.2 The Second Proposal

One might reasonably argue that $\hat{V}_{u1}$ is not subtle enough. Consider that pairs of samples with relatively small overlap could be used together with no overlap cases in order to create a better estimator than S_U^2. One possibility is to use:

$$\hat{V}_{u2} = max_k\{Q(m) - Q(k) : 0 \leq k \leq m\} \tag{3.3.4}$$

It is clear that $\hat{V}_{u2} \geq \hat{V}_{u1} \geq \hat{V}_u$. That is, by preserving nonnegativity, we potentially increase the positive bias. Notice that each of the terms $Q(m) - Q(k)$ $(k = 1, ..., m)$ has a negative bias (Lemma 3.2). So, we might expect that $\hat{V}_{u2}$ does not overall have a large positive bias.

3.3.3 The Third Proposal

Stepping forward, after realizing that $\hat{V}_{u2}$ could bring a rather substantial computational burden, we are hoping to consider further strategies that will not only save computation but also strengthen the $\hat{V}_u$ estimator.

Now, consider the distribution of the sample overlap, denoted as $X = O(S_1, S_2)$ (3.1.2). It can be seen that X has the probability mass:

$$Prob(X = k) = \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}}; k = 0, 1, ..., m \tag{3.3.5}$$

$$\Rightarrow \text{ Hypergeometric(n;m,m).}$$

We might pick a value $k^\star$ such that the fractional overlap among its samples is relatively low among all overlaps, but do so that we still have enough pairs (S_1, S_2) to be averaged over. In other words, we could choose $k^\star$ as the αth percentile of the overlap distribution, i.e. *Hypergeometric(*$n; m, m$*).* And then, one could use

$$\hat{V}_{u3} = max\{S_U^2, Q(m) - Q(k^\star), \hat{V}_u\}. \tag{3.3.6}$$

Recall that the mean and variance of hypergeometric distribution with parameters (n, m, m) are $\frac{m^2}{n}$ and $\frac{m^2(n-m)^2}{n^2(n-1)}$ respectively, and hypergeometric distribution can be approximated by binomial distribution, $Binomial(n, p)$, with $p = \frac{m}{n}$ under certain conditions. Given the relationship between binomial distribution and normal distribution, it is reasonable to connect hypergeometric distribution with normal distribution. As a result, the specification of the $k^\star$ value can be obtained based on normal approximation to hypergeometric distribution (Feller's Lemma).

Lemma 3.3 (Feller's Lemma [10])

If $n \to \infty, m \to \infty$ so that $m/n \to t \in (0, 1)$ and $x_k := (k - mp)/\sqrt{mpq} \to x$, then:

$$P(k; m, n) \sim \frac{e^{-ax^2/2}}{\sqrt{2\pi mpq(1-t)}}; a := \frac{1}{1-t} \tag{3.3.7}$$

where $P(k; m, n)$ is the point mass of *Hypergeometric(*$n; m, m$*)* at k, and $p = m/n$ in our case.

Proof:

Denote:

$$\begin{aligned} P(k;m,n) &= p^k q^{m-k}\binom{m}{k}\cdot R(k;m,n), \text{ where} \\ R(k;m,n) &= \frac{\Pi_{j=1}^{k-1}(1-\frac{j}{m})\Pi_{j=1}^{m-k-1}(1-\frac{j}{n-m})}{\Pi_{j=1}^{m-1}(1-\frac{j}{n})}, \text{ and} \\ q &= 1-p. \end{aligned}$$

By Stirling's formula, i.e.

$$n! = \sqrt{2\pi n}(\frac{n}{e})^n,$$

together with the assumptions that $\frac{m}{n} \to t$ and

$$k/m \sim t + x\sqrt{qt/m},$$

$$k/(n-m) \sim t + x\sqrt{pt/(n-m)},$$

we have:

$$\Pi_{j=1}^{m-1}(1-\frac{j}{n}) \sim \frac{e^{-nt}}{(1-t)^{n(1-t)+1/2}}(1+O(\frac{1}{n}))$$

$$\Pi_{j=1}^{k-1}(1-\frac{j}{m}) \sim \frac{e^{np(t+x\sqrt{qt/m})}}{(1-t-x\sqrt{qt/m})^{m(1-t-x\sqrt{qt/m})}}(1+O(\frac{1}{n}))$$

$$\Pi_{j=1}^{m-k-1}(1-\frac{j}{n-m}) \sim \frac{e^{-(n-m)(t-x\sqrt{pt/(n-m)})}}{(1-t+x\sqrt{pt/(n-m)})^{(n-m)(1-t+x\sqrt{pt/(n-m)})}}(1+O(\frac{1}{n}))$$

Then, we have:

$$\begin{aligned} R(k;m,n) &\approx \frac{(1-t)^{n(1-t)+1/2}}{e^{-nt}} \\ &\cdot \frac{e^{-m(t+x\sqrt{qt/m})}}{(1-t-x\sqrt{qt/m})^{m(1-t-x\sqrt{qt/m})+1/2}} \\ &\cdot \frac{e^{-(n-m)(t-x\sqrt{\frac{pt}{n-m}})}}{(1-t+x\sqrt{\frac{pt}{n-m}})^{(n-m)(1-t+x\sqrt{\frac{pt}{n-m}})+1/2}} \end{aligned}$$

Hence, we have:

$$\begin{aligned} logR(k;m,n) &\approx [n(1-t)+\frac{1}{2}]log(1-t) - m(t+x\sqrt{qt/m}) \\ &- (n-m)(t-x\sqrt{\frac{pt}{n-m}}+nt) \\ &- [m(1-t-x\sqrt{qt/m}+\frac{1}{2}]log(1-t-x\sqrt{qt/m}) \\ &- [(n-m)(1-t+x\sqrt{\frac{pt}{n-m}})+\frac{1}{2}]log(1-t+x\sqrt{\frac{pt}{n-m}}) \end{aligned}$$

which can be simplified as

$$logR(k;m,n) \approx I + A + B + C$$

where

$$\begin{aligned} I &= -mt-(n-m)t+nt-mx\sqrt{\frac{qt}{m}}+(n-m)x\sqrt{\frac{pt}{n-m}} \\ &= -mx\sqrt{\frac{qt}{m}}+(n-m)x\sqrt{\frac{pt}{n-m}} \\ &\to x\sqrt{(1-c)cnt}-x\sqrt{(1-c)cnt} \\ &= 0 \text{ as } n\to\infty; \end{aligned}$$

and

$$\begin{aligned} A &= n(1-t)log(1-t)-m(1-t)log(1-t-x\sqrt{\frac{qt}{m}}) \\ &- (n-m)(1-t)log(1-t+x\sqrt{\frac{pt}{n-m}}) \\ B &= \frac{1}{2}log(1-t)-\frac{1}{2}log(1-t-x\sqrt{\frac{qt}{m}})-\frac{1}{2}log(1-t+x\sqrt{\frac{pt}{n-m}}) \\ C &= mx\sqrt{\frac{qt}{m}}[log(1-t)-x\sqrt{\frac{qt}{m}}\frac{1}{1-t}+o(\frac{1}{\sqrt{m}})] \\ &- (n-m)x\sqrt{\frac{pt}{n}-m}[log(1-t)+x\sqrt{\frac{pt}{n-m}}\frac{1}{1-t}+o(\frac{1}{\sqrt{n-m}})] \end{aligned}$$

Furthermore,

$$\begin{aligned}
A &= (n-m)(1-t)log(1-t) - m(1-t)[-x\sqrt{\frac{qt}{m}}\frac{1}{1-t} - \frac{x^2qt}{2m}\frac{1}{(1-t)^2} + o(\frac{1}{m})] \\
&- (n-m)(1-t)[log(1-t) + x\frac{pt}{n-m} - \frac{x^2pt}{2(n-m)}\frac{1}{(1-t)^2} + o(\frac{1}{n-m})] \\
&= n(1-t)log(1-t) - m(1-t)log(1-t) + (m-n)(1-t)log(1-t) \\
&+ x\sqrt{mqt} + \frac{x^2qt}{2(1-t)} + o(1) - x\sqrt{(n-m)pt} + \frac{x^2pt}{2(1-t)} + o(1) \\
&= x\sqrt{mqt} - x\sqrt{(n-m)pt} + \frac{tx^2}{2(1-t)} + o(1) \\
&\to \frac{tx^2}{2(1-t)} \\
B &= \frac{1}{2}log(1-t) - \frac{1}{2}log(1-t-x\sqrt{\frac{qt}{m}}) - \frac{1}{2}log(1-t+x\sqrt{\frac{pt}{n-m}}) \\
&\to -\frac{1}{2}log(1-t) \\
C &= x\sqrt{mqt}log(1-t) - x\sqrt{(n-m)pt}log(1-t) - \frac{qtx^2}{1-t} - \frac{ptx^2}{1-t} + o(1) \\
&= x\sqrt{nc(1-c)t}log(1-t) - x\sqrt{nc(1-c)t}log(1-t) - \frac{tx^2}{1-t} + o(1) \\
&\to -\frac{tx^2}{1-t}
\end{aligned}$$

as $n \to \infty$ and $\frac{m}{n} = \gamma = t$.

Therefore, we have

- $A \to \frac{tx^2}{2(1-t)}$ as $n \to \infty (m \to \infty)$
- $B \to -\frac{1}{2}log(1-t)$ as $n \to \infty (m \to \infty)$
- $C \to -\frac{tx^2}{1-t}$ as $n \to \infty (m \to \infty)$

In sum,

$$R(k;m,n) \sim \frac{1}{\sqrt{1-t}} e^{-\frac{tx^2}{2(1-t)}}.$$

That is,

$$P(k;m,n) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi mpq}} \frac{1}{\sqrt{1-t}} e^{-\frac{tx^2}{2(1-t)}}.$$

Therefore, we have

$$P(k;m,n) \sim \frac{1}{\sqrt{2\pi mpq(1-t)}} \cdot e^{-ax^2/2}; a := \frac{1}{1-t}.$$

□

According to Feller's Lemma, we could obtain the αth percentile of the overlap distribution based on normal approximation to hypergeometric distribution. One possible choice would be to use the 25th percentile of the overlap distribution.

In our example, we can easily verify the conditions in Lemma 3.3:

- $n \to \infty, m \to \infty$ such that $\frac{m}{n} \to t$

  In our case, $\frac{m}{n} = \gamma$ fixed. Take $t = \gamma$, the first condition is satisfied.

- $x_k = \frac{k-mp}{\sqrt{mpq}} \to x$

  Based on the assumption that $\frac{k}{m} \sim t + x\sqrt{\frac{qt}{m}} \Leftrightarrow k \sim mt + x\sqrt{mqt}$ and $p = t = \gamma, q = 1-\gamma$, we have

  $$\frac{k-mp}{\sqrt{mpq}} \sim \frac{mp + x\sqrt{mqt} - mp}{\sqrt{mpq}} = x.$$

  That is, the second condition is also satisfied.

If we want to get the αth percentile of the distribution of $O(S_1, S_2)$, denoted as $k^\star$, then the normal approximated value can be solved via:

$$\frac{\sqrt{1/(1-\gamma)}\frac{k^\star - m\gamma}{\sqrt{m\gamma(1-\gamma)}}}{\sqrt{m\gamma(1-\gamma)^2}} \approx z_\alpha,$$

where z_α is the αth percentile of standard normal distribution.

That is,

$$k^\star \approx z_\alpha m\gamma(1-\gamma)^2 + m\gamma = m\gamma[1 + z_\alpha(1-\gamma)^2].$$

3.4 Comparison with Some Bootstrap Variance Estimators

In the following, I will compare the unbiased variance estimator (i.e.$\hat{V}_u$) with some bootstrap variance estimators. I am going to use the kernel function in Example 1.1 which corresponds to a U-statistic equivalent to the sample variance S_n^2, i.e.

$$K(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2.$$

Furthermore, suppose $X_1, ..., X_n i.i.d. \sim Normal(0, 1)$.

Under standard normal assumption,

$$(n-1)S_n^2 \sim X^2(n-1).$$

Therefore, the true variance of the U-statistic (which is S_n^2) is just $\frac{2}{n-1}$.I will measure the estimation accuracy of the unbiased formula method and several bootstrap approaches (nonparametric, balanced, smooth, and parametric bootstrap methods). In addition, I am also going to report the computation time required for each method to obtain the same number of estimations.

Our goal is to estimate $\theta = Var(U_n)$ which can be written as a functional of a population distribution function F (where F is the standard normal distribution in this case), denoted as $\theta(F)$. The basic principal of bootstrap is to use the same functional of a estimated distribution function $\hat{F}$ to estimate $Var(U_n)$. In the sense of nonparametric bootstrap, $\hat{F}$ is commonly chosen to be the empirical distribution; while in the sense of parametric bootstrap, $\hat{F}$ is usually chosen to be a normal distribution with mean equal to the sample mean and variance equal to the sample variance, if the original distribution F is assumed to be normal. Balanced and smooth bootstrap methods add further restriction or modifications on how to draw the bootstrap samples.

By resampling $B = 1000$ size-n samples conditional on each sample of $X_1, ..., X_n$ i.i.d. $\sim N(0, 1)$, there is one estimate of $\theta(F) = Var(U_n)$ by the unbiased variance formula (3.1.6) and each of the bootstrap methods. In order to evaluate each method and measure the accuracy of these several different variance estimators, I

chose $R = 1000$ size-n samples drawn from $N(0, 1)$ and resampled $B = 1000$ times for the bootstrap procedure given each of the size-n samples.

The average of mean estimations (i.e. $Ave\{E(\hat{U}_n)\}$), the average of variance estimations (i.e. $Ave\{\hat{Var}(U_n)\}$), standard deviation of variance estimations (i.e. $SD\{\hat{Var}(U_n)\}$), and computing time for calculating $R = 1000$ estimates are summarized in the following three tables.

Table 3.1. Comparison with Bootstrap Variance Estimators ($n = 6, m = 2$)

$n = 6$	True	Formula	Nonpar.	Balanced	Smooth	Par.
$Ave\{E(\hat{U}_n)\}$	1	0.9897	0.8244	0.8251	0.9887	0.9909
$Ave\{\hat{Var}(U_n)\}$	0.4	0.3780	0.2403	0.2401	0.4065	0.5490
$SD\{\hat{Var}(U_n)\}$		0.6051	0.3208	0.3180	0.5180	0.6993
$Time$		0.02sec	1.74sec	1.66sec	5.60sec	5.57sec

Table 3.2. Comparison with Bootstrap Variance Estimators ($n = 8, m = 2$)

$n = 8$	True	Formula	Nonpar.	Balanced	Smooth	Par.
$Ave\{E(\hat{U}_n)\}$	1	0.9972	0.8729	0.8726	0.9963	0.9961
$Ave\{\hat{Var}(U_n)\}$	0.2857	0.2834	0.1978	0.1980	0.2833	0.3643
$SD\{\hat{Var}(U_n)\}$		0.4285	0.2561	0.2575	0.3373	0.4065
$Time$		0.14sec	2.77sec	3.02sec	8.40sec	7.46sec

Table 3.3. Comparison with Bootstrap Variance Estimators ($n = 25, m = 2$)

$n = 25$	True	Formula	Nonpar.	Balanced	Smooth	Par.
$Ave\{E(\hat{U}_n)\}$	1	0.9940	0.9539	0.9544	0.9940	0.9936
$Ave\{\hat{Var}(U_n)\}$	0.0833	0.0836	0.0743	0.0741	0.0815	0.0892
$SD\{\hat{Var}(U_n)\}$		0.0617	0.0535	0.0530	0.0571	0.0536
$Time$		10.62sec	24.75sec	25.88sec	42.28sec	38.32sec

Conclusion:

Within our expectation, compared with other bootstrap counterparts, $\hat{V}_u$ provided us with the least-bias estimates for both the mean and variance of the U-statistic. In terms of computing time, the unbiased formula approach was about 100 times more efficient for small n cases and was about 3 times more efficient for the $n = 25$ case.

However, for this specific example, nonparametric bootstrap methods tended to result in estimators with much less standard deviation than the unbiased formula approach.

From another point of view, the comparison between unbiased variance estimator with bootstrap variance estimators is actually equivalent to the comparison between subsampling and bootstrapping. The subsampling approach is superior in terms of unbiasedness and computation efficiency. The standard deviation of subsampling becomes closer to that of bootstrapping as sample size n gets larger. We will investigate these relationships further later in the thesis.

Chapter 4

Two Unbiased Resampling Schemes, Their Comparisons and Properties

4.1 Two Resampling Schemes

For problems with large n and m values, it is computationally expensive to compute U-statistics and the corresponding $Q(m), Q(0)$ values. In Chapter 2, we have discussed some approaches to realize efficient unbiased estimation by reduced U-statistics or general incomplete U-statistics. In this chapter, I will introduce two resampling schemes to realize our unbiased variance estimator $\hat{V}_u$ efficiently.

To reduce computing time and alleviate the calculation complexity, I consider estimating $Q(m)$ and $Q(0)$ based on B bootstrapped size-m samples. It can be shown that the way I define $\hat{Q}(m)^{(k)}$ and $\hat{Q}(0)^{(k)}$ (k=1,2) will result in resampling unbiased estimates of $Q(m)$ and $Q(0)$ given the number of non-overlapping pairs of size-m samples. Here, resampling unbiasedness will mean that for a fixed data set, the mean over repeated resampling equals $Q(m)$ and $Q(0)$ respectively.

4.1.1 Type 1 Resampling Scheme

Let $x_1, ..., x_n$ be the original data from a distribution F. If we randomly draw B data subsets of size m with replacement, we get B elements of $\{S_r : r = 1, ..., N; N = \binom{n}{m}\}$ where we use S_r to represent data subsets of size m and use r to represent the labels on the possible data subsets. Equivalently, we are drawing

labels $L_1 = r_1, L_2 = r_2, ..., L_B = r_B$ from $\{1 \leq i_1 < ... < i_m \leq n\}$.

Let the **set-label space** be all vectors of length B with coordinates from set $\{\{1, ..., n\} - \text{choose } m\}$. Denote this label space as LSS. This is, in a basic sense, the sample space for the subsampling experiment, and we will say that we are drawing a set of B "set labels" with replacement.

By drawing B set labels $r_1, ..., r_B$, we use $S_{r_1}, ..., S_{r_B}$ to estimate the complete U-statistic U_n by

$$\tilde{U}_B = \frac{1}{B}\sum_{i=1}^{B} K(S_{r_i})$$

which is actually equivalent to the incomplete U-statistic defined in (2.2.1).

Since the set labels were drawn with replacement from LSS, we have

$$\begin{aligned} E[\frac{1}{B}\sum_{i=1}^{B} K(S_{r_i})] &= E[K(S_{r_1})] \\ &= \frac{1}{N}\sum_{r=1}^{N} K(S_r) \\ &= U_n \end{aligned}$$

That is, $\tilde{U}_B$ is a subsampling unbiased estimate of the corresponding complete U-statistic.

Then, a natural extension is whether we could use the same subsampling experiment to construct a subsampling unbiased realization of the unbiased variance estimate of the complete U-statistic.

First consider

$$\hat{Q}(m)^{(1)} = \frac{1}{B(B-1)}\sum_{i \neq j} K(S_{r_i})K(S_{r_j}) \quad (4.1.1)$$

$$= \frac{1}{B(B-1)}[(K(S_{r_1}) + ... + K(S_{r_B}))^2 - \sum_{i=1}^{B} K^2(S_{r_i})] \quad (4.1.2)$$

Notice that the term on the first line (4.1.1) involves $\binom{B}{2}$ calculations, but the second line (4.1.2) is much shorter to calculate. Recall the incomplete U-statistic defined in (2.2.1), in fact, $K(S_{r_1}) + ... + K(S_{r_B}) = B\tilde{U}_B$. So, the formula can be

further simplified as:

$$\hat{Q}(m)^{(1)} = \frac{1}{B(B-1)}[B^2\tilde{U}_B^2 - \sum_{i=1}^{B} K^2(S_{r_i})] \tag{4.1.3}$$

Proposition 3.1: $\hat{Q}(m)^{(1)}$ is subsampling unbiased.

Proof:

Since $\hat{Q}(m)^{(1)}$ only counts for distinct pairs of (S_{r_i}, S_{r_j}) based on resampling WITH replacement, we have

$$\begin{aligned} E[K(S_{r_i})K(S_{r_j})] &= \frac{1}{N^2}\sum_{l=1}^{N}\sum_{k=1}^{N} K(S_l)K(S_k) \\ &= U_n^2 \\ &= Q(m), \end{aligned}$$

where $N = \binom{n}{m}$, and $i \neq j$.

That is, $\hat{Q}(m)^{(1)}$ is an unbiased realization of $Q(m)$ over the resampling distribution, given the size-n data set $x_1, ..., x_n$.

□

Denote

$$\tilde{I}(i,j) = I\{S_i \cap S_j = \emptyset, \{i,j\} \subseteq \{L_1, ..., L_B\}\} \tag{4.1.4}$$

$$\hat{Q}(0)^{(1)} = \frac{1}{\sum_{i\neq j;i,j=1,...,N} \tilde{I}(i,j)} \sum_{i\neq j;i,j=1,...,N} K(S_i)K(S_j)\tilde{I}(i,j) \tag{4.1.5}$$

Notice that in (4.1.5) the resampling randomness is only in $\tilde{I}(i,j)$.

Proposition 3.2: Denote

$$\sum_{i\neq j;i,j=1,...,N} \tilde{I}(i,j) = C > 0,$$

$\hat{Q}(0)^{(1)}$ is a conditional subsampling unbiased realization of $Q(0)$ given the value of C.

Namely, we have

$$E[\hat{Q}(0)^{(1)}|\sum_{i\neq j;i,j=1,...,N}\tilde{I}(i,j)=C]=Q(0) \tag{4.1.6}$$

where $C>0$.

Proof:

Let the **constrained set-label space** be the subset of set-label space LSS such that the number of non-overlapped pairs generated by the set-label sample is exactly C. Denote the constrained set label space as $LSS_C=\{(L_1,...,L_B)\in LSS|\sum_{i\neq j;i,j=1,...,N}\tilde{I}(i,j)=C\}$.

Claims:

1. Every set-label sample of size B is equally likely. That is, every $(L_1,...,L_B)\in LSS$ is equally likely.
   Proof:
   $$P\{(L_1,...,L_B)=(l_1,...,l_B)\}=(\frac{1}{N})^B$$
   for any B subset labels of size m, by the property of resampling with replacement.

2. With the constrained label space (i.e. given the constant C), every set-label sample is equally likely. Therefore, the conditional distribution of set-label samples given the constraint is uniform.

3. Fix C, assume there is at least one non-overlapping pair of subsets, say S_{r_1} and S_{r_2}, with labels $\{r_1,r_2\}$. Now consider any other pair that is non-overlapped, say S_{r_i} and S_{r_j}, with labels $\{r_i,r_j\}$. We claim that the labels $\{r_1,r_2\}$ are found in exactly the same number of constrained label sets as are $\{r_i,r_j\}$, and so all such non-overlapping pairs are equally likely. (This is intuitive, as the process should not favour any one pair over another.)

4. Also notice that a permutation of indices does not affect the overlap counts.

Proof:

Let (r_1, r_2) and (r_i, r_j) be two pairs of non-overlapped indices.

Without loss of generality, assume the indices for r_1 are $(1, 2, ..., m)$; indices for r_2 are $(m+1, ..., 2m)$. Similarly, assume the indices for r_i are $(a_1, ..., a_m)$; indices for r_j are $(a_{m+1}, ..., a_{2m})$. And neither r_i nor r_j has indices among $a_{2m+1} < ... < a_n$.

Construct a permutation map by setting $perm(i) = a_i$. That is, we now have a mapping of indices: $1, 2, ..., n \rightarrow^{\star} a_1, ..., a_m, a_{m+1}, ..., a_{2m}, ..., a_n$. This is invertible. Now apply this transformation, $\star$, to all the possible sets r in LSS or LSS_C. Each r generates a new set $r^{\star}$ containing the transformed labels. This induces a mapping on the set of labels, so that r gets mapped into $r^{\star}$, the label for $S_{r^{\star}}$. A little thought tells us that this mapping is also $1-1$ and onto, from the integers $\{\{1, ..., n\} - \text{choose m}\}$ onto the same integers. (Given any set-label sample $t \in \{\{1, ..., n\} - \text{choose } m\}$, we need to find an r such that $r^{\star} = t$. We can do so by applying the inverse permutation to the indices in t to generate r.)

Now this mapping takes an element $L_1, ..., L_B$ of LSS, and maps into another element of LSS, denoted as $L_1^{\star}, ..., L_B^{\star}$. This mapping is also $1-1$ and onto.

Note that a permutation map on the indices cannot change the overlap counts, so $\star$ is also a map from LSS_C into LSS_C. It is also $1-1$ and onto LSS_C because of this. (Let M be the number of elements of LSS_C. Since the $\star$ map is one-to-one, the image set must have M elements. But the image set is contained in LSS_C, of size M, so the image set must equal LSS_C).

Finally, since all the non-overlapping pairs are equally likely to appear, the expectation of $\hat{Q}(0)^{(1)}$ given the constraint is just the average value times the constant C, as needed.

That is,

$$\begin{aligned}
E(\hat{Q}(0)^{(1)} | \sum_{i \neq j} \tilde{I}(i,j) = C) &= E[\frac{1}{C} \sum_{a \neq b} K(S_{L_a}) K(S_{L_b}) \tilde{I}(L_a, L_b) | \sum_{i \neq j} \tilde{I}(i,j) = C] \\
&= \frac{1}{C} E[\sum_{a \neq b} K(S_{L_a}) K(S_{L_b}) I(L_a, L_b) | \sum_{i \neq j} \tilde{I}(i,j) = C] \\
&= \frac{1}{C} C \frac{\sum_{r \neq t} K(S_r) K(S_t) I\{S_r \cap S_t = \emptyset\}}{\sum_{r \neq t} I\{S_r \cap S_t = \emptyset\}} \\
&= Q(0)
\end{aligned}$$

Thus, we have proved the unbiasedness of $\hat{Q}(0)^{(1)}$ given the number of non-

overlapping pairs C when $C > 0$.
□

As a result, the Type 1 resampling realization of the unbiased variance estimator, $\hat{V}_u$, can be defined as

$$\hat{V}_u^{(1)} = \hat{Q}(m)^{(1)} - \hat{Q}(0)^{(1)} \tag{4.1.7}$$

One concern we might have for the Type 1 resampling scheme is that for fixed B, the number of sampled pairs without common elements might be rare or even non-existent, especially when B is small. As a result, the estimate of Q(0) using $\hat{Q}(0)^{(1)}$ may not be good. In order to overcome this drawback of Type 1 resampling scheme, we consider another approach–the Type 2 resampling scheme.

4.1.2 Type 2 Resampling Scheme

Let $S_{1,2m}, S_{2,2m}, ..., S_{M,2m}, M = \binom{n}{2m}$ be an enumeration of all possible size-$2m$ samples drawn from the original data $x_1, ..., x_n$.

Now, consider randomly selecting $B/2$ samples with replacement out of the full set with probability $\frac{1}{M}$ for each of the $S_{i,2m}, 1 \leq i \leq M$ in the full set. Equivalently, we are drawing $B/2$ labels of size $2m$ from $\{\{1, ..., n\} - \text{choose } 2m\}$, denoted as $L_{1,2m} = r_{1,2m}, L_{2,2m} = r_{2,2m}, ..., L_{B/2,2m} = r_{B/2,2m}$. The subscripts m and $2m$ represent the number of elements in the set-label sample. (Without loss of generality, here we assume $B/2$ is an integer.)

Then, we randomly split each size-$2m$ set-label sample into two size-m labels. That is, we have $r_{1,2m} = r_{1,m} \cup \bar{r}_{1,m}, ..., r_{B/2,2m} = r_{B/2,m} \cup \bar{r}_{B/2,m}$. Namely,

$$S_{r_{i,2m}} = (S_{r_{i,m}}, S_{\bar{r}_{i,m}}); 1 \leq i \leq B/2. \tag{4.1.8}$$

Therefore, in Type 2 resampling scheme, we have at least $B/2$ pairs of size-m data subsets $(S_{r_i,m}, S_{\bar{r}_i,m})$ without common elements.

Similarly as discussed in Section 4.1.1, here we can use the B split size-m data subsets, $S_{r_{1,m}}, ..., S_{r_{B/2,m}}, ..., S_{\bar{r}_{B/2,m}}$, to estimate the corresponding complete

U-statistic, i.e.

$$\tilde{U}_B = \frac{1}{B}\sum_{i=1}^{B/2}[K(S_{r_{i,m}}) + K(S_{\bar{r}_{i,m}})] \tag{4.1.9}$$

which is an unbiased estimate of U_n. (Note, we could split the size-$2m$ sample up into all its possible pairs of nonoverlapping size-m samples. But this might be a huge number–it seems that it is likely to be more statistically efficient to just generate a new size-$2m$ sample and split it randomly.)

Similarly, define

$$\begin{aligned}
& \hat{Q}(m)^{(2)} \\
= & \frac{\sum_{i\neq j}[K(S_{r_i,m})K(S_{r_j,m}) + K(S_{\bar{r}_i,m})K(S_{\bar{r}_j,m}) + K(S_{r_i,m})K(S_{\bar{r}_j,m})]}{4\cdot(B/2)\cdot(B/2-1)} \\
= & \frac{1}{2B(B/2-1)}\{[K(S_{r_{1,m}}) + ... + K(S_{r_{B/2,m}}) + ... + K(S_{\bar{r}_{B/2,m}})]^2 \\
- & \sum_{i=1}^{B/2}[K^2(S_{r_{i,m}}) + K^2(S_{\bar{r}_{i,m}})] - 2\sum_{i=1}^{B/2}K(S_{r_{i,m}})K(S_{\bar{r}_{i,m}})\}
\end{aligned}$$

Again, denote

$$\tilde{I}(i,j) = I\{S_i \cap S_j = \emptyset, \{i,j\} \subseteq \{r_{1,m}, ..., r_{B/2,m}, \bar{r}_{1,m}, ..., \bar{r}_{B/2,m}\}\}$$

$$\hat{Q}(0)^{(2)} = \frac{1}{\sum_{i\neq j}\tilde{I}(i,j)}\sum_{i\neq j}K(S_i)K(S_j)\tilde{I}(i,j)$$

The variance estimation by Type 2 resampling scheme is

$$\hat{V}_u^{(2)} = \hat{Q}(m)^{(2)} - \hat{Q}(0)^{(2)} \tag{4.1.10}$$

The unbiasedness of $\hat{Q}(m)^{(2)}$ and $\hat{Q}(0)^{(2)}$ can be shown in a similar fashion as discussed in section 4.1.1.

4.2 Properties of the Two Resampling Schemes

In this section we will study how the two sampling schemes affect the number of non-overlapping pairs available to estimate $Q(0)$.

Type 1 Scheme:
$X_1, ..., X_n, \binom{n}{m}$ size-m samples. Draw B size-m samples out of $\binom{n}{m}$ with replacement.

$$P(non-overlaps) = \frac{\binom{n-m}{m}}{\binom{n}{m}} \tag{4.2.1}$$

$$E_1(non-overlaps) = \frac{\binom{n-m}{n}\binom{B}{2}}{\binom{n}{m}} \tag{4.2.2}$$

Type 2 Scheme:
$X_1, ..., X_n, \binom{n}{2m}$ size-$2m$ samples. Draw $B/2$ size-2m samples out of $\binom{n}{2m}$ with replacement and split each one into two size-m samples.

$$E_2(non-overlaps) = \frac{B}{2} + \frac{\binom{n-m}{n}(\binom{2\cdot\frac{B}{2}}{2} - \frac{B}{2})}{\binom{n}{m}} = \frac{B}{2} + \frac{\binom{n-m}{m}(\frac{B(B-2)}{2})}{\binom{n}{m}} \tag{4.2.3}$$

$$E_2 - E_1 = \frac{B}{2} + \frac{\binom{n-m}{m}}{\binom{n}{m}}[\frac{B(B-2)}{2} - \frac{B(B-1)}{2}] = \frac{B}{2}[1 - \frac{\binom{n-m}{m}}{\binom{n}{m}}] \tag{4.2.4}$$

Relative improvement:

$$\frac{E_2 - E_1}{E_1} = \frac{\frac{B}{2}[1 - \frac{\binom{n-m}{m}}{\binom{n}{m}}]}{\binom{B}{2}\frac{\binom{n-m}{m}}{\binom{n}{m}}} = \frac{1 - \binom{n-m}{m}/\binom{n}{m}}{(B-1)\binom{n-m}{m}/\binom{n}{m}} \tag{4.2.5}$$

Denote $p(n,m) = \frac{\binom{n-m}{m}}{\binom{n}{m}}$,then

$$\text{relative improvement} = \frac{1 - p(n,m)}{(B-1)p(n,m)} \tag{4.2.6}$$

It can be seen that when either B or $p(m,n) = \frac{\binom{n-m}{m}}{\binom{n}{m}}$ is large enough, the relative improvement will become slight or negligible.

The gain of Type 2 resampling of having more non-overlapped pairs is a doubling or better if $\frac{1-p(n,m)}{(B-1)p(n,m)} \geq 1$. Namely,

$$\frac{1 - p(n,m)}{p(n,m)} \geq B - 1 \tag{4.2.7}$$

From another aspect, by elementary calculation we have:

$$E_1(non-overlaps) = \frac{\binom{n-m}{m}}{2\binom{n}{m}} B(B-2) + \frac{\binom{n-m}{m}}{\binom{n}{m}} \frac{B}{2}$$

$$E_2(non-overlaps) = \frac{\binom{n-m}{m}}{2\binom{n}{m}} B(B-2) + \frac{B}{2}$$

The difference of non-overlaps between the two resampling schemes is $\frac{B}{2}(1 - \frac{\binom{n-m}{m}}{\binom{n}{m}})$. Notice that

$$\frac{\binom{n-m}{m}}{\binom{n}{m}} = \frac{n-m}{n} \frac{n-1-m}{n-1} \frac{n-2-m}{n-2} ... \frac{n-m+1-m}{n-m+1},$$

then, we can conclude that:

- When m is close to 1(i.e. the kernel size m is small compared with sample size n), the gain of Type 2 resampling scheme is slight.
- When m is close to n(i.e. the kernel size m is relatively large compared with sample size n), the gain of Type 2 resampling scheme is substantial.

Let's revisit the kernel function used in Example 1.1, and we will continue with the assumption that $n = 8, m = 2$ and $X_1, ..., X_8 i.i.d. \sim Normal(0, 1)$. Based on the simulation result in Table 3.2, we can also count the average of non-overlapping pairs in each of the resampling schemes.

Table 4.1. Non-overlapping Pairs of the Two Resampling Schemes

Non-overlaps	$B = 8$	$B = 12$	$B = 16$	$B = 20$	$B = 24$	$B = 28$
Type 1 Scheme	15.12	35.39	64.07	102.07	147.99	202.35
Type 2 Scheme	17.02	38.56	68.50	107.07	154.71	210.87
Rel.Improvement	0.1421	0.0806	0.0662	0.05436	0.0460	0.0383

The exact expected non-overlaps for the two resampling schemes can be calculated based on formulas (4.2.2) and (4.2.3). For instance, when $B = 8$, the theoretical values are $E_1 = 15$ and $E_2 = 16.86$, very close to our simulation results. For other B values, the expected non-overlaps based on the simulation can also be shown to be close to the theoretical values.

Moreover, the simulation result also confirmed that the relative improvement became smaller as resample size B increased, which can be visually seen from the following plot.

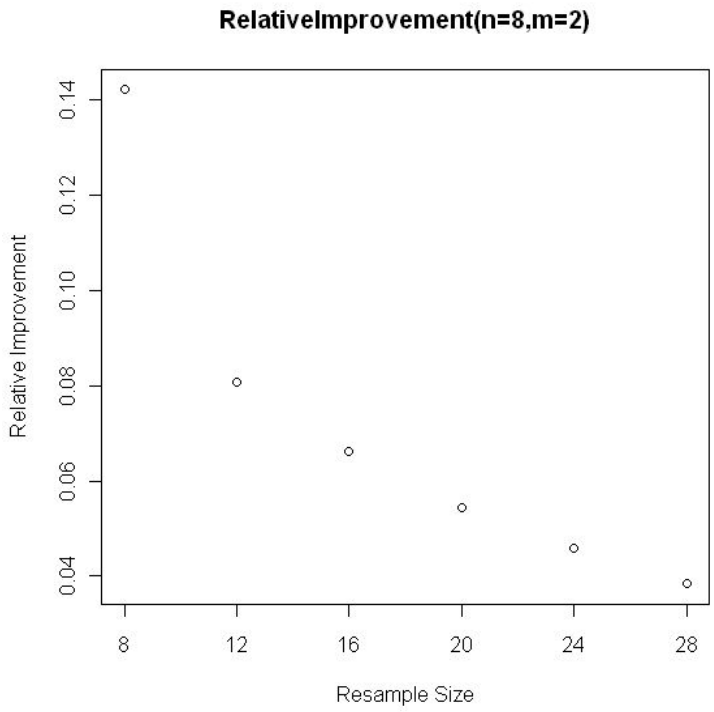

Figure 4.1. Relative Improvement

Chapter 5

Nonparametric Density Estimation

In this chapter, we will be discussing a practical implementation of the unbiased variance estimator of a U-statistic in the context of nonparametric kernel density estimation. First of all, let us have a brief review of the existing approaches to accomplish probability density estimation.

5.1 Introduction

The **probability density function** of a continuous random variable X, say $f(\cdot)$, is usually defined to be a nonnegative real-valued function with property

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

In statistical practice, we sometimes know little about the underlying probability density of the random variable. Therefore, in order to understand how the random variable is distributed or to compute the probability of the random variable taking values in a certain interval, we need to construct an estimate of the true probability density in some fashion based on a data set. The idea of nonparametric density estimation was first proposed by Fix and Hodges (1951) [12] with applications in discriminant analysis. It has since been widely used in many other areas in statistics. The importance of density estimation in exploring and presenting data can be found in Silverman (1986) [33].

There are two basic approaches to estimating a probability density function,

parametric or nonparametric. The parametric approach assumes that the underlying density f comes from a known parametric family of density functions. In this case, our main task is to estimate the unknown, fixed parameters in the parametric form based on the data set and replace the unknown true parameters by their estimates in order to obtain a density estimator. For instance, suppose that the probability density belongs to a normal family with mean μ and variance σ^2. Then, we can simply use the sample mean and sample variance to estimate μ and σ^2 respectively, denoted as $(\hat{\mu}, \hat{\sigma}^2)$, and consider the density of $Normal(\hat{\mu}, \hat{\sigma}^2)$ as the estimator. In contrast, nonparametric approach frees one from the restriction of the fixed parametric form of the underlying density so that the data can speak for themselves in the estimation of f.

Throughout this chapter, we will mainly discuss nonparametric density estimation methods. In the following, I will introduce several existing nonparametric density estimation tools but will focus on kernel density estimator. Later on, the unbiased U-statistic risk estimators based on L^2 distance and Kullback-Leibler distance will be proposed. A simulation study will be conducted to demonstrate how to select the optimal bandwidth in kernel estimation based on U-statistic risk estimators, followed by the implementation of the unbiased variance estimator for these U-statistic risk estimators.

5.2 Histogram

The histogram is the oldest and most easily interpreted tool for density estimation. For simplicity, we will only consider the univariate case throughout this chapter, i.e. the continuous random variable $X \in \mathcal{R}$.

Given a origin x_0 and a bin width h, we can partition the support of the underlying density f, say $[a, b)$, by a grid of bins with the form

$$[x_0 + (k-1) \cdot h, x_0 + k \cdot h)$$

with positive values of k. Without losing generality, assume $1 \leq k \leq L, k \in \mathcal{Z}$, and $a = x_0 + 0 \cdot h < x_0 + 1 \cdot h < ... < x_0 + L \cdot h = b$.

Suppose the size of the observed data is n. That is, we denote the data set as

$X_1, X_2, ..., X_n$. Then, the density estimation at point x can be expressed as

$$\hat{f}(x) = \frac{1}{nh}(\text{number of observations in the same bin as } x) \tag{5.2.1}$$

If we let n_k be the number of observed data in the kth bin, then we have

$$\hat{f}(x) = \frac{n_k}{nh}, x \in [x_0 + (k-1) \cdot h, x_0 + k \cdot h), \tag{5.2.2}$$

where $k = 1, ..., L$.

As seen in formula (5.2.2), the bin width h is directly related to the density estimation at each point. A simple illustration of the effect of bin width can be seen from the following figure, where *Sepal.Length* in the *iris data* of Fisher [11] was used to construct the histograms. The left panel used bin width $h = 0.5$, while the right panel used bin width $h = 0.2$. It can be clearly seen that the histogram tends to be smoother with larger bin width. Especially in this example, the spiky features in the right panel do not appear with bin width $h = 0.5$.

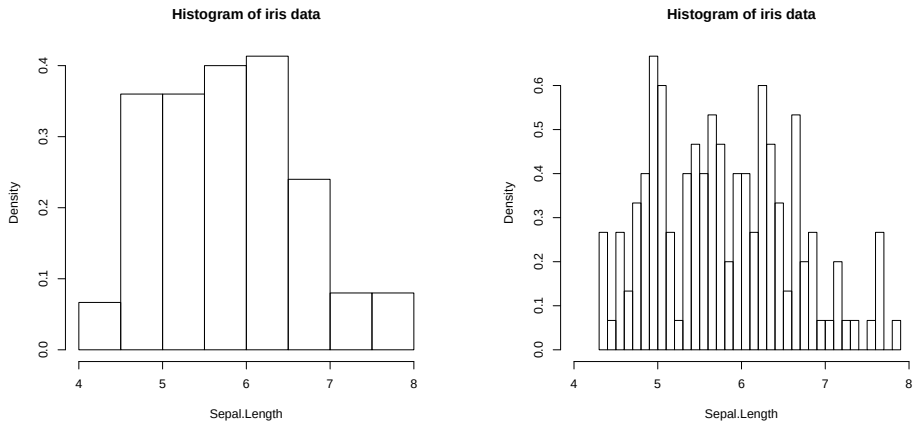

Figure 5.1. Histogram of iris data (Sepal.Length) with bin width $h = 0.5$ (left panel) and $h = 0.2$ (right panel)

Actually, the bin width can be changed from cell to cell. In other words, we can use distinct bin widths for different bins in constructing the histogram, say that h_k is the bin width for the kth bin ($k = 1, ..., L$). Then, the bins partitioning

the support of f becomes

$$[x_0 + (k-1) \cdot h_k, x_0 + k \cdot h_k); k = 1, 2, ..., L.$$

The density estimation at point x can be written as

$$\hat{f}(x) = \frac{n_k}{nh_k}, x \in [x_0 + (k-1) \cdot h_k, x_0 + k \cdot h_k)$$

where n_k is the number of observations in the kth bin, and $k = 1, 2, ..., L$.

Although the histogram gives us a representation of the empirical distribution that is easy to construct and visualize, it is far from satisfactory. Some of its defects include the discontinuities at the bin boundaries, zero value outside the range of the grid of bins, and the major effect of bin width choice, as was shown in Figure 5.1. Moreover, if we alter the origin of the histogram, the shape of the histogram may also change dramatically, which leads to inconsistent performance by using different origins.

The following two histograms are still based on the data set of *Sepal.Length* from the *iris data*. In the left panel, the origin $x_0 = 3.5$; in the right panel, the origin $x_0 = 4$. One of the major differences between these two graphs is that the left panel is visually right skewed while the right one is roughly symmetric.

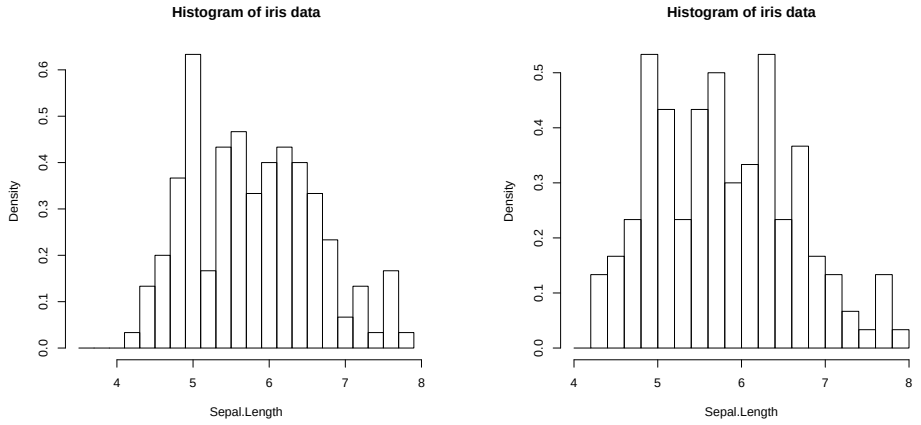

Figure 5.2. Histogram of iris data (Sepal.Length) with origin $x_0 = 3.5$ (left panel) and $x_0 = 4$ (right panel)

In sum, the histogram method suffers from several major drawbacks. It usually cannot be used as more than an exploratory tool for the underlying empirical distribution.

5.3 Orthogonal Series Estimation

The orthogonal series estimator is based upon Fourier expansion and aims to estimate the coefficients in the Fourier series for the probability density function.

Define the basis of the Fourier expansion as:

$$\phi_0(x) = 1 \tag{5.3.1}$$
$$\phi_{2k-1}(x) = cos(kx) \tag{5.3.2}$$
$$\phi_{2k}(x) = sin(kx) \tag{5.3.3}$$

where $k = 1, 2, ...$

The probability function, f, can be represented as

$$f(x) = \sum_{t=0}^{\infty} c_t \phi_t(x), \tag{5.3.4}$$

where c_t's are the coefficients.

Recall the property of orthogonality between *sine* and *cosine* functions, i.e.

$$\begin{aligned}
\int_{-\pi}^{\pi} sin(mx)cos(nx)dx &= 0, \\
\int_{-\pi}^{\pi} sin(mx)sin(nx)dx &= \pi\delta(m-n), \\
\int_{-\pi}^{\pi} cos(mx)cos(nx)dx &= \pi\delta(m-n),
\end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function, i.e. $\delta(m-n) = 1$ if and only if $m = n$ and 0 otherwise.

Therefore, by simple mathematical calculation, we have

$$c_t = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\phi_t(x)dx, t \geq 0.$$

It is easily seen that $c_t = \frac{1}{\pi} E_f[\phi_t(x)]$ and therefore $\hat{c}_t = \frac{1}{n\pi} \sum_{i=1}^{n} \phi_t(x_i)$ is an unbiased estimate of c_t $(t \geq 0)$. However, Silverman [33] has pointed out that $\sum_{t=0}^{\infty} \hat{c}_t \phi_t(x)$ is not a good estimate of f because the series converge to a linear form of Dirac delta functions, i.e.

$$\omega(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - X_i).$$

Notice that

$$\hat{c}_t = \frac{1}{\pi} \int_{-\pi}^{\pi} \omega(x) \phi_t(x) dx, t \geq 0.$$

Therefore, $\hat{c}_t$ are the exact Fourier coefficients for the function ω but not for f.

Because ω is the discrete, empirical density, in order to obtain a useful estimate for the continuous density f, a simple solution introduced in [33] is to smooth ω by truncating the expansion $\sum_{t=0}^{\infty} \hat{c}_t \phi_t(x)$ at a certain point. That is, consider

$$\hat{f}_(x) = \sum_{t=0}^{T} \hat{c}_t \phi_t(x)$$

as the estimate for f. The choice of T determines the amount of smoothing.

What is more, Silverman [33] also suggested another approach by imposing weights to the Fourier series. That is, consider the weighted Fourier series

$$\hat{f}(x) = \sum_{t=0}^{\infty} \lambda_t \hat{c}_t \phi_t(x),$$

where the weights λ_t satisfy $\lambda_t \to 0$ as $t \to \infty$. In this case, the convergence rate of λ_t to zero determines the amount of smoothing.

Besides of the two nonparametric density estimators introduced in Section 5.2 and 5.3, other density estimation approaches include the nearest neighbour method, the variable kernel method, the general weighted function estimator, the maximum penalized likelihood estimator, the pseudo-likelihood estimator, the kernel density estimator, and more. In the following, we will focus on the kernel density estimator, which is the most visible and used density estimation method. Details of other methods can be found in [33] and [34].

5.4 Kernel Density Estimator

5.4.1 Introduction

Nowadays, the most popular density estimation method is the kernel density estimator. Suppose $X_1, X_2, ..., X_n$ is an i.i.d. sample from some probability density f. Then, the kernel density estimator at point x is defined for a kernel K as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}), x \in \mathcal{R}, h > 0. \tag{5.4.1}$$

If we denote $K_h(x) = \frac{1}{h} K(\frac{x}{h})$, then the kernel density estimator can be rewritten as

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \tag{5.4.2}$$

The function K is called the kernel. It is a kernel of order k if it satisfies the following assumptions (Turlach (1993) [34]):

1. K is symmetric, i.e. $K(u) = K(-u)$.
2. $\int_R K(u)du = 1$.
3. $\int_R u^j K(u)du = 0$ for $j = 1, ..., k-1$.
4. $\int_R u^k K(u)du \neq 0$.

It can be seen that the symmetry of K implies that k must be an even number. For $k = 2$, K is non-negative and is itself a probability density. For the order k to be 4 or larger, it is necessary for K to take negative values and thus may result in negative density estimations. Although K with higher order ($k \geq 4$) harbours better smoothness properties, negative values for density estimates are not desirable in practice. Furthermore, the simulation study in Marron and Wand (1992) [27] revealed that for some difficult-to-estimate underlying densities the sample size n needs to be in millions in order for higher order kernel functions to predominate over kernel functions of order 2. Consequently, kernels of order 2 are generally preferable and will be considered as the conventional choice throughout this section.

Some examples of univariate kernel functions are listed below:

Table 5.1. Examples of Univariate Kernel Functions

Kernel Function	$K(x)$
Rectangular	$\frac{1}{2}I_{\{\lvert x\rvert\leq 1\}}$
Triangular	$(1-\lvert x\rvert)I_{\{\lvert x\rvert\leq 1\}}$
Bartlett-Epanechnikov	$\frac{3}{4}(1-x^2)I_{\{\lvert x\rvert\leq 1\}}$
Biweight	$\frac{15}{16}(1-x^2)^2I_{\{\lvert x\rvert\leq 1\}}$
Triweight	$\frac{35}{32}(1-x^2)^3I_{\{\lvert x\rvert\leq 1\}}$
cos	$\frac{\pi}{4}cos(\frac{\pi}{2}x)I_{\{\lvert x\rvert\leq 1\}}$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$

A graphical demonstration of the compact support kernel functions are displayed in Figure 5.3:

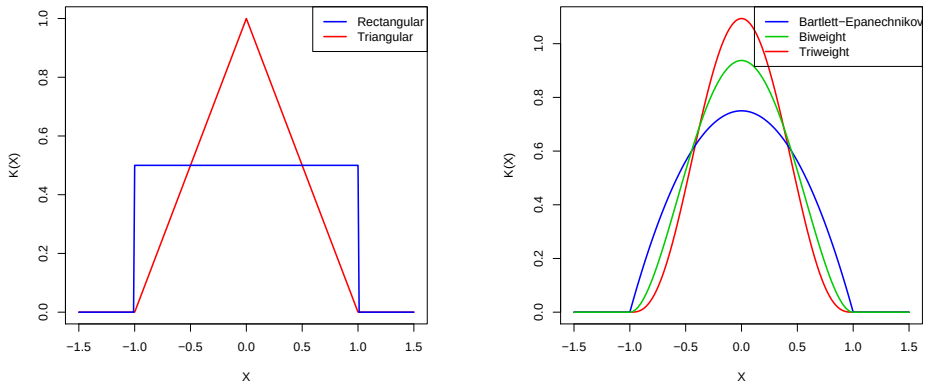

Figure 5.3. Compact Support Kernel Functions. Left panel: Rectangular and Triangular kernels. Right panel: Bartlett-Epanechnikov, Biweight, and Triweight kernels.

Although the choice of kernel function more or less affects the accuracy of density estimation, the bandwidth choice h in kernel density estimation is much more crucial (Hardle et. al. (1994) pp.57-61 [15]). In short, bandwidth h controls the roughness of the fitted density curve and plays a similar role as bin width does in the histogram and smoothing parameter does in the smoothing spline method. As a result, two important practical issues that address many researchers' attention

are how to select the optimal bandwidth given the data set and what criterion should be used in bandwidth selection.

5.4.2 Assessing the Kernel Density Estimator

In this subsection, we will review several measurements that are currently used to evaluate the bandwidth-dependent kernel density estimator.

5.4.2.1 L^2 Distance-based Assessment

One of the most popular measurements to evaluate how closely $\hat{f}$ approximates f for a given data set is the **integrated squared error** (also called L^2 loss), which is defined as

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2 dx, \tag{5.4.3}$$

where $\hat{f}_h(\cdot)$ is the kernel density estimator defined in formula (5.4.1).

If furthermore, we are interested in the performance of $\hat{f}$ averaging over all possible data sets, then we can take expectation of $ISE(h)$ over the data set which yields another measurement, the **mean integrated squared error** (also called L^2 risk).

$$MISE(h) = IMSE(h) = E\{\int (\hat{f}_h(x) - f(x))^2 dx\} \tag{5.4.4}$$

One thing to notice here is that by Fubini's Theorem, we have

$$\begin{aligned} MISE(h) &= E\{\int (\hat{f}_h(x) - f(x))^2 dx\} \\ &= \int E\{\hat{f}_h(x) - f(x)\}^2 dx \end{aligned}$$

In other words, the integrated mean-squared error ($IMSE$) is equal to the mean integrated squared error ($MISE$).

In addition, $MISE(h)$ has the following presentation ([34]):

$$\begin{aligned} MISE(h) &= \int E_f\{\hat{f}_h(x) - f(x)\}^2 dx \\ &= \int E\{\hat{f}_h(x) - E[\hat{f}_h(x)] + E[\hat{f}_h(x)] - f(x)\}^2 dx \\ &= \int E\{\hat{f}_h(x) - E[\hat{f}_h(x)]\}^2 + \{E[\hat{f}_h(x)] - f(x)\}^2 dx \\ &= \int Var[\hat{f}_h(x)]dx + \int bias^2[\hat{f}_h(x)]dx \end{aligned}$$

If we denote the integrated variance of $\hat{f}_h(x)$ as

$$IV(h) = \int Var[\hat{f}_h(x)]dx,$$

and denote the integrated squared bias of $\hat{f}_h(x)$ as

$$IB(h) = \int bias^2[\hat{f}_h(x)]dx.$$

Then,

$$MISE(h) = IV(h) + IB(h).$$

It can be shown that:

$$\begin{aligned} IV(h) &= \frac{R(K)}{nh} - \frac{1}{n}\int (K_h * f)^2(x)dx \\ IB(h) &= \int (K_h * f - f)^2(x)dx \\ &= \int (K_h * f)^2(x)dx - 2\int (K_h * f)(x)f(x)dx + \int f^2(x)dx \end{aligned}$$

where $R(K) = \int K^2(x)dx$, and $*$ denotes the convolution of two functions with $K * L(x) = \int K(x-u)L(u)du = \int K(u)L(x-u)du$.

With the assumptions that f has at least $k+2$ derivatives and K is a kernel

function of order k, $IV(h)$ and $IB(h)$ can be simplified asymptotically to

$$\begin{aligned} IV(h) &= \frac{R(K)}{nh} + \frac{1}{n}R(f) + O(n^{-1}h^k) \\ IB(h) &= \frac{h^{2k}}{(k!)^2}\mu_k^2(K)R(f^{(k)}) + O(h^{2k+4}) \end{aligned}$$

where $\mu_k(K) = \int x^k K(x)dx$, and $f^{(k)}$ is the kth derivative of f.

Now, if we assume $h \to 0$ such that $n \cdot h \to \infty$ as $n \to \infty$, then it follows the **asymptotic mean integrated squared error** $(AMISE)$:

$$AMISE(h) = (nh)^{-1}R(K) + h^{2k}(\mu_k(K)/k!)^2 R(f^{(k)}) \tag{5.4.5}$$

Adapting the notations used in [34], denote

$$\hat{h}_0 = argmin_h ISE(h) \tag{5.4.6}$$

$$h_0 = argmin_h MISE(h) \tag{5.4.7}$$

$$h_\infty = argmin_h AMISE(h) \tag{5.4.8}$$

Note that $\hat{h}_0$ depends on the dataset because $ISE(h)$ does.

Based on (5.4.5), h_∞ has the following closed form:

$$h_\infty = (\frac{R(K)(k!)^2}{2k\mu_k^2(K)R(f^{(k)})})^{\frac{1}{2k+1}} \cdot n^{-\frac{1}{2k+1}}. \tag{5.4.9}$$

For the case of $k = 2$, we have

$$h_\infty = (\frac{R(K)}{\mu_2^2(K)R(f^{(2)})})^{1/5} \cdot n^{-1/5} \tag{5.4.10}$$

Turlach (1993) [34] stated that h_∞ is usually a poor approximation to h_0 unless the sample size n is in the millions. In other words, h_∞ is rarely considered as a satisfying practical surrogate to h_0.

One thing to notice is that all of the above criteria (formulas (5.4.3), (5.4.4), (5.4.5)) involve the unknown underlying density f. In other words, it is not feasible to compute their minimizers. As a result, we need to estimate the measure of fit based on observations somehow in order to select the "optimal" bandwidth $\hat{h}$. Hall

and Marron (1991) [16] showed that for any data-driven bandwidth selector $\hat{h}$, *in the best case* its relative rate of convergence to $\hat{h}_0$ is $n^{-1/10}$, while its relative rate of convergence to h_0 is $n^{-1/2}$.

$$\frac{\hat{h}}{\hat{h}_0} = 1 + O_p(n^{-1/10}) \tag{5.4.11}$$

$$\frac{\hat{h}}{h_0} = 1 + O_p(n^{-1/2}) \tag{5.4.12}$$

This shows that a bandwidth selection method $\hat{h}$ that aims to target $\hat{h}_0$ is likely to be highly variable, just as $\hat{h}_0$ itself is. In comparison, the $\hat{h}$ targeting h_0 may adapt less well to the sample at hand.

5.4.2.2 Kullback-Leibler Distance-based Assessment

We can also evaluate the kernel density estimator based on Kullback-Leibler loss.

Kullback-Leibler distance (loss) for measuring the closeness between two density functions is defined as

$$d(f, \hat{f}_h) = \int log \frac{f(x)}{\hat{f}_h(x, \mathcal{X}_n)} f(x) dx \tag{5.4.13}$$

where $f(x)$ is the true underlying density, and $\mathcal{X}_n$ is the data set of size n.

Then, the corresponding risk dependent on h is

$$Risk = E_{\mathcal{X}_n}[d(f, \hat{f}_h(\mathcal{X}_n))] \tag{5.4.14}$$

More investigation for risk based on Kullback-Leibler distance will be discussed in a later subsection.

5.5 Applications of U-Statistics in Risk Estimation

5.5.1 Established Results

In order to obtain the minimizer over h of the L^2 distance-based criteria in Section 5.4.2.1, the terms involving f need to be estimated, or we need to otherwise approximate the measure of fit based on the data set. Several well-known approaches have been developed and thoroughly studied, some of which will be summarized briefly in the following.

1. The "Quick and Dirty" Method:

   Here, consider the $AMISE$ criterion. As shown in formula (5.4.5),

   $$AMISE(h) = (nh)^{-1}R(K) + h^{2k}(\mu_k(K)/k!)^2 R(f^{(k)}).$$

   Since $R(f^{(k)})$ is unknown, the exact minimizer of (5.4.5) is unattainable. One possible solution is to choose a "reference density" for f and calculate $R(f^{(k)})$ by substituting f with its reference. For instance, if we assume f is a normal density with mean 0 and variance σ^2 and K is a Gaussian kernel, then we can work out the exact form of $R(f^{(k)})$ with $k = 2$ easily. In this case, the optimal bandwidth, usually called *rule-of-thumb* bandwidth, is

   $$h_{rot} = 1.06\hat{\sigma}n^{-1/5} \tag{5.5.1}$$

   where $\hat{\sigma}$ is an estimate of σ, such as the sample standard deviation.

   A more robust version is to consider interquartile range $\hat{R}$ instead of $\hat{\sigma}$. As an alternative to $\hat{\sigma}$, we can take the minimum between $\hat{\sigma}$ and $\hat{R}/1.34$. Silverman (1986) [33] showed that for Gaussian kernel, $\hat{R} \approx 1.34\hat{\sigma}$. If we use the more robust estimator of σ, we have

   $$\hat{h}_{rot} = 1.06min\{\hat{\sigma}, \frac{\hat{R}}{1.34}\}n^{-1/5}. \tag{5.5.2}$$

   Another possibility is to use the lower bound for $R(f^{(2)})$ $(k = 2)$ as a substi-

tute. This is called the "minimal smoothing principle". By plugging in the lower bound of $R(f^{(2)})$ into $AMISE(h)$, the minimizer becomes

$$\hat{h}_{MSP} = 3 \cdot (35)^{-1/5} \cdot \hat{\sigma} \cdot (R(K)/\mu_2^2(K))^{1/5} \cdot n^{-1/5}. \tag{5.5.3}$$

Note: Turlach [34] stated that $\hat{h}_{rot}$ and $\hat{h}_{MSP}$ usually perform pretty well when f is a uni-modal probability density. However, for multi-modal case, the selected bandwidth tends to oversmooth the data and conceals the detailed features of the underlying density.

2. Cross-Validation Methods

   - *Leave-One-Out Cross-Validation (Unbiased CV)*:
     Define the cross-validation objective function to be

     $$UCV(h) = R(\hat{f}_h) - 2\sum_{i=1}^{n} \hat{f}_{h,-i}(\mathbf{X}_{<-\mathbf{i}>})$$

     where $\mathbf{X}_{<-\mathbf{i}>}$ is the set of data except the ith observation.

     It can be easily shown that

     $$E[UCV(h)] = MISE(h) - R(f).$$

     That is, $UCV(h)$ is an unbiased estimator of $MISE(h)$ ignoring the term $R(f)$ which is independent of h. Moreover, $UCV(h)$ can also be seen as an estimate of $ISE(h)$ except for a constant term.

     Therefore, $UCV(h)$ can be viewed as either an $MISE$-based criterion or an ISE-based criterion. Rudemo (1982) [31] first proposed $UCV(h)$ with the aim of seeking the minimizer of an estimate of $MISE(h)$, while Bowman (1984) [2] independently found the same bandwidth selector by trying to approximate $ISE(h)$.

     Note:

     – The convergence rate of $\hat{h}_{UCV}$ to both h_0 and $\hat{h}_0$ is $n^{-1/10}$. Based on formula (5.4.12) and (5.4.12), it does not accomplish the best

convergence rate as a bandwidth selector in aim of reaching h_0 but is optimal in approximating $\hat{h}_0$. Since the convergence rate of $\hat{h}_{UCV}$ is very slow, it suffers from the problem of sample variation.

- Another practical problem for unbiased CV is that $UCV(h)$ often has more than one local minima. It is recommended that we choose the rightmost value of h where a local minima occurs (see [34]).
- Although strong criticisms of unbiased CV can be found in many articles, Loader (1999) [26] was in favour of this classical approach. She also verified its soundness via some simulation studies and real data examples.

- *Biased Cross-Validation*:

  Biased cross-validation was first appeared in Scott and Terrell (1987) [32]. Recall formula (5.4.5), i.e.

$$AMISE(h) = (nh)^{-1}R(K) + h^{2k}(\mu_k(K)/k!)^2 R(f^{(k)}); k = 2.$$

  The "Quick-and-Dirty" method is based on a "reference density" of f and to minimize the resulting score function. Instead, [32] estimated $R(f^{(2)})$ by $R(\hat{f}_h^{(2)})$ to obtain a objective function $BCV(h)$ based on which minimization is realized with respect to h.

  It can be shown that (see [32]):

$$\begin{aligned} R(\hat{f}_h^{(2)}) &= \frac{1}{n}K_h^{(2)} * K_h^{(2)}(0) + \frac{1}{n^2}\sum_{i\neq j} K_h^{(2)} * K_h^{(2)}(X_i - X_j) \\ &= \frac{1}{nh^5}K^{(2)} * K_h^{(2)}(0) + \frac{1}{n^2h^5}\sum_{i\neq j} K^{(2)} * K^{(2)}(X_i - X_j) \\ &= \frac{1}{nh^5}R(K^{(2)}) + \frac{1}{n^2h^5}K^{(2)} * K^{(2)}(X_i - X_j) \end{aligned}$$

  Then, the biased cross-validation criterion can be expressed as

$$BCV(h) = \frac{1}{nh}R(K) + h^4\frac{\mu_2^2(K)}{4n^2}\sum_{i\neq j} K_h^{(2)} * K_h^{(2)}(X_i - X_j). \quad (5.5.4)$$

  Note:

  - The minimizer of $BCV(h)$, denoted as $\hat{h}_{BCV}$, is the optimal bandwidth proposed by [32]. Although $\hat{h}_{BCV}$ has the same relative convergence rate as $\hat{h}_{UCV}$, the difference compared with $\hat{h_0}$ or h_0 was shown to be always smaller for $\hat{h}_{BCV}$.
  - In the case of more than one minimizer of the $BCV(h)$ function, it was shown that the best performance is often obtained by picking the smallest value of h for which a local minimum occurs.

- *Smoothed Cross-Validation*
  (Hall, Marron, and Park (1992) [17])
  Recall the representation

$$MISE(h) = IV(h) + IB(h)$$

  where $IV(h)$ and $IB(h)$ are defined in section 5.4.2.

  In Marron and Wand (1992) [27] it was shown that $\frac{1}{nh}R(K)$ is a good estimator of $IV(h)$. When it comes to the second term $IB(h)$, [17] suggested to use $\hat{f}_g$, where $\hat{f}_g$ is another kernel density estimator with a possibly different bandwidth g and kernel function L. After simplification, it can be shown that

$$\hat{IB}(h) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_h * K_h - 2K_h + K_0) * L_g * L_g(X_i - X_j)$$

  where K_0 denotes the Dirac delta function.

  Furthermore, by using the approximation $n \approx n - 1$, we have

$$SCV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)}\sum_{i\neq j}(K_h * K_h - 2K_h + K_0) * L_g * L_g(X_i - X - j) \tag{5.5.5}$$

  Note:

  - The name "Smoothed Cross-Validation" comes from the fact that

by using $n \approx n-1$, $UCV(h)$ can be written as

$$UCV(h) = \frac{1}{nh}R(K) + \frac{1}{n(n-1)}\sum_{i \neq j}(K_h * K_h - 2K_h + K_0)(X_i - X_j),$$

assuming there are no duplicates in the data set. That is, $SCV(h)$ can be viewed as a representation of $UCV(h)$ where the differences $X_i - X_j$ are pre-smoothed in some fashion.

– Hall et. al. [17] proposed to use the minimizer of $SCV(h)$ as the suggested bandwidth, called $\hat{h}_{SCV}$. They showed that g and L can be properly chosen such that the convergence rate of $\hat{h}_{SCV}$ to h_0 can reach the best case, i.e.$O_p(n^{-1/2})$. However, in order to achieve this best convergence rate, the kernel function L has to be at least of order 6.

- *Bandwidth Factorized Smoothed Cross-Validation*
(Jones, Marron, and Park (1991) [22])
Using the same score function as $SCV(h)$ but allowing g to be related to h, Jones et. al [22] modified the Smooth Cross-Validation approach and denoted the target function as $JMP(h)$.

$$JMP(h) = \frac{R(K)}{nh} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_h \times K_h - 2K_h + K_0) \times L_g \times L_g(X_i - X_j) \tag{5.5.6}$$

One important result for $JMP(h)$ is that by choosing $g \sim n^{-23/45}h^{-2}$, the minimizer of $JMP(h)$, called $\hat{h}_{JMP}$, can achieve root-n convergence rate even if K and L are both of order 2.

5.5.2 Unbiased Estimate for L^2 Risk

In this subsection, we will propose an unbiased estimator of L^2 risk with the form of a U-statistic.

Define

$$Risk_{L^2,n}(h) = E_{\mathcal{X}_n}[\int (f(x) - \hat{f}_h(x))^2 dx]$$

where $f(x)$ is the true underlying density, $\mathcal{X}_n$ is the set of data/observations of size

n based on which the nonparametric kernel estimation at x is taken, and

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

It can be seen that the defined $Risk_{L^2,n}(h)$ is actually $MISE$ (mean-integrated squared error). By Fubini's theorem, $MISE = IMSE$. Then, we can decompose the risk function in the following way:

$$Risk_{L^2,n}(h) = \int E_{\mathcal{X}_n}[f(x)^2 - 2f(x)\hat{f}_h(x) + \hat{f}_h(x)^2]dx.$$

Notice that $E_{\mathcal{X}_n}[f(x)^2]$ is not related to h. So, we only need to focus on

$$\begin{aligned} Risk_{L^2,n}(h) &\propto -2\int E_{\mathcal{X}_n}[f(x)\hat{f}_h(x)]dx + \int E_{\mathcal{X}_n}[\hat{f}_h(x)^2]dx \\ &= -2E_{\mathcal{X}_n}[\int f(x)\hat{f}_h(x)dx] + E_{\mathcal{X}_n}[\int \hat{f}_h(x)^2 dx] \end{aligned}$$

Denote

$$\begin{aligned} Term1 &= E_{\mathcal{X}_n}[\int f(x)\hat{f}_h(x)dx] \\ &= E_{\mathcal{X}_n}[\int f(x)\frac{1}{n}\sum_{i=1}^{n} K_h(X_i, x)dx] \\ &= E_{\mathcal{X}_2}[K_h(X_1, X_2)] \end{aligned}$$

Moreover, let

$$\begin{aligned} Term2 &= E_{\mathcal{X}_n}[\int \hat{f}_h(x)^2 dx] \\ &= E_{\mathcal{X}_n}[\frac{1}{n^2}\int \sum_i \sum_j K_h(X_i, x)K_h(X_j, x)dx] \\ &= E_{\mathcal{X}_n}[\frac{1}{n^2}\sum_i \sum_j K_{\sqrt{2}h}(X_i, X_j)] \\ &= \frac{1}{n^2}\{nE_{\mathcal{X}_2}[K_{\sqrt{2}h}(X_1, X_1)] + n(n-1)E_{\mathcal{X}_2}[K_{\sqrt{2}h}(X_1, X_2)]\} \\ &= E[K_{\sqrt{2}h}(X_1, X_2)] + \frac{1}{n}E[K_{\sqrt{2}h}(X_1, X_1) - K_{\sqrt{2}h}(X_1, X_2)] \end{aligned}$$

where $K_h(\cdot)$ stands for a Gaussian kernel with mean 0 and variance h^2; $K_{\sqrt{2}h}$ is a Gaussian kernel with mean 0 and variance $2 \cdot h^2$. The third equality in the above derivation is based on the fact that for Gaussian kernels, we have

$$\int K_h(X_i, x)K_h(X_j, x)dx = K_{\sqrt{2}h}(X_i, X_j).$$

In sum,

$$\begin{aligned} Risk_{L^2,n}(h) &\propto E[K_{\sqrt{2}h}(X_1, X_2)] - 2E[K_h(X_1, X_2)] \\ &+ \frac{1}{n}E[K_{\sqrt{2}h}(X_1, X_1) - K_{\sqrt{2}h}(X_1, X_2)]. \end{aligned}$$

Following the footsteps of Ray and Lindsay (2008) [30] and Lindsay and Liu (2009) [25], we propose to estimate the risk $Risk_{L^2,m}(h)$ at sample sizes m that are not equal to n. There are several reasons that one might wish to consider such an approach. One particular motivation here is that Lindsay and Liu found in their problem (not density estimation) that the risk at $m = n$ was very difficult to estimate. There is some reason to believe this holds for density estimation as well. Moreover, as Ray and Lindsay pointed out, the well known BIC model selection criterion corresponds to using minimum risk estimation with $m = n/\{log(n) - 1\}$. (Note: Throughout this Chapter, we use m to represent the subsample size.)

We also know that the cross-validation criterion for bandwidth selection tends to overfit the density (Loader (1999)[26]). So, we consider $Risk_{L^2,m}(h)$, i.e. using $\frac{1}{m}$ to replace $\frac{1}{n}$ in the second line above with $m < n$, as an alternative measurement to assess the density estimation. Later in Figure 5.4 it will be seen that the minimizer h_{L^2} of $Risk_{L^2,m}(h)$ is decreasing in m. That is, by taking $m < n$ in $Risk_{L^2,m}(h)$, the optimal bandwidth selected based on minimization of $Risk_{L^2,m}(h)$ is larger than the one obtained in minimizing $Risk_{L^2,n}(h)$. This therefore potentially adjusts the undersmoothing problem of cross-validation algorithm.

In other words, instead of using $Risk_{L^2,n}(h)$ we can refer to

$$\begin{aligned} Risk_{L^2,m}(h) &\propto E[K_{\sqrt{2}h}(X_1, X_2)] - 2E[K_h(X_1, X_2)] \\ &+ \frac{1}{m}E[K_{\sqrt{2}h}(X_1, X_1) - K_{\sqrt{2}h}(X_1, X_2)] \end{aligned}$$

as the measurement to evalulate the density estimation. The suggested bandwidth is the one that minimizes $Risk_{L^2,m}(h)$.

When K is a Gaussian kernel, by taking derivative of $Risk_{L^2,m}(h)$ with respect to h and setting it equal to zero, we can solve for the value of m for which a particular h would be optimal:

$$m_h = \frac{\frac{1}{2h^2} + \frac{(1+2h^2)^2-4h^2(1+2h^2)-1}{[(1+2h^2)^2-1]^{3/2}}}{\frac{(1+2h^2)^2-4h^2(1+2h^2)-1}{[(1+2h^2)^2-1]^{3/2}} - \frac{\sqrt{2}[(1+h^2)^2-2h^2(1+h^2)-1]}{[(1+h^2)^2-1]^{3/2}}}. \tag{5.5.7}$$

Inverting this formula gives us the **exact** formula to calculate h_{L^2} given a value of m.

A figure illustrating the relationship between m_h and h_{L^2} together with a comparison with h_{rot} (5.5.1) under the assumption of the Gaussian kernel is shown below:

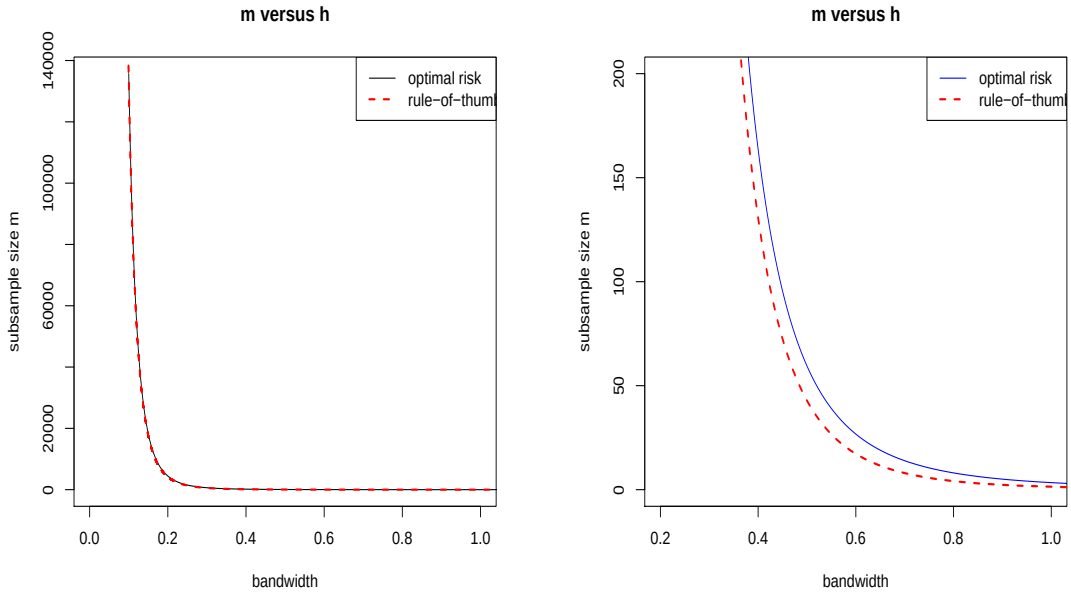

Figure 5.4. Relationship between m and h, shown at two different scales

Figure 5.4 indicates that for a given value of subsample size m, the optimal bandwidth selected by (5.5.7) is slightly larger than the rule of thumb h_{rot} for the case that K is a Gaussian kernel.

As mentioned earlier, instead of using the exact form of $Risk_{L^2,m}(h)$, we can construct an unbiased estimator of $Risk_{L^2,m}(h)$ based on a U-statistic and select the optimal bandwidth for kernel density estimation by minimizing the estimated risk, denoted as U_{L^2}:

$$U_{L^2} := \hat{Risk}_{L^2,m}(h) = \frac{1}{\binom{n}{2}} \sum_{i<j} K_h^\star(x_i, x_j) \tag{5.5.8}$$

where

$$K_h^\star(x_1, x_2) = K_{\sqrt{2}h}(x_1, x_2) - 2K_h(x_1, x_2) + \frac{1}{m}[\frac{1}{2h\sqrt{\pi}} - K_{\sqrt{2}h}(x_1, x_2)]. \tag{5.5.9}$$

Note that U_{L^2} is equivalent to the unbiased cross-validation formula (also called leave-one-out cross-validation) when $m = n-1$. That is, both of them are unbiased estimates for relative $MISE$ and functions of the order statistics (modulo terms that do not depend on h). In addition, the bagged bandwidth selector (bagging on CV) proposed by Hall and Robinson (2009) [19] is actually nothing more than the bandwidth selector based on $\hat{Risk}_m(h)$ if one makes m equal to their bootstrap size.

If we minimize (5.5.8) over h, we have the inverse relationship
$\hat{m}_h =$

$$\frac{\frac{1}{\binom{n}{2}} \sum_{i<j}[\frac{1}{2h^2\sqrt{\pi}} - \frac{1}{2h^2\sqrt{\pi}} e^{-(x_i-x_j)^2/(4h^2)} + \frac{(x_i-x_j)^2}{4h^4\sqrt{\pi}} e^{-(x_i-x_j)^2/(4h^2)}]}{\frac{1}{\binom{n}{2}} \sum_{i<j}[-\frac{e^{-\frac{(x_i-x_j)^2}{4h^2}}}{2h^2\sqrt{\pi}} + \frac{(x_i-x_j)^2}{4h^4\sqrt{\pi}} e^{-\frac{(x_i-x_j)^2}{4h^2}} + \frac{2 \cdot e^{-\frac{(x_i-x_j)^2}{2h^2}}}{\sqrt{2\pi}h^2} - \frac{2(x_i-x_j)^2}{\sqrt{2\pi}h^4} e^{-\frac{(x_i-x_j)^2}{2h^2}}]}. \tag{5.5.10}$$

This function describes the dependence structure between m and h in determining the least estimated L^2 risk.

5.5.2.1 A Simulation Study

Consider the objective function

$$\begin{aligned} f_{L^2,m}(h) &= \frac{1}{\binom{n}{2}}\sum_{i<j}[\frac{1}{2h^2\sqrt{\pi}} - \frac{1}{2h^2\sqrt{\pi}}e^{-(x_i-x_j)^2/(4h^2)} + \frac{(x_i-x_j)^2}{4h^4\sqrt{\pi}}e^{-(x_i-x_j)^2/(4h^2)}] \\ &- m\cdot\frac{1}{\binom{n}{2}}\sum_{i<j}[-\frac{1}{2h^2\sqrt{\pi}}e^{-\frac{(x_i-x_j)^2}{4h^2}} + \frac{(x_i-x_j)^2}{4h^4\sqrt{\pi}}e^{-\frac{(x_i-x_j)^2}{4h^2}} + \frac{2}{\sqrt{2\pi}h^2}e^{-\frac{(x_i-x_j)^2}{2h^2}} \\ &- \frac{2(x_i-x_j)^2}{\sqrt{2\pi}h^4}e^{-\frac{(x_i-x_j)^2}{2h^2}}], \end{aligned}$$

which can be obtained based on rearranging (5.5.10) and setting one side to zero.

For a fixed value of m, the bandwidth selector can be obtained by seeking the root(s) of $f_{L^2,m}(h)$. If $f_{L^2,m}(h)$ has a unique solution, the optimal bandwidth can be easily attained via bisection algorithm; otherwise, we need to take the rightmost root of $f_{L^2,m}(h)$ as suggested in [31].

In the following simulation study, I drew $R = 1000$ samples of size 100 independently from $Normal(0,1)$ and considered $m = 30, 40, ..., 100$. The average minimizer of $\hat{Risk}_{L^2,m}(h)$ and the standard deviation of the $R = 1000$ bandwidth selectors for each m value can be found in the following table.

Table 5.2. Relationship between m and $\hat{h}_{L^2}$ by Minimizing U_{L^2}

m	30	40	50	60	70	80	90	100
$E[\hat{h}_{L^2}]$	0.5902	0.5508	0.5221	0.4994	0.4809	0.4663	0.4544	0.4426
$SD[\hat{h}_{L^2}]$	0.0890	0.0958	0.1019	0.1081	0.1125	0.1160	0.1183	0.1222

Furthermore, we can compute the simulation mean of exact ISE for the density estimator $\hat{f}_{\hat{h}_{L^2}}(x)$ based on formula (5.4.3).

Table 5.3. $ISE(\hat{h}_{L2})$

m	30	40	50	60
$E[ISE_{\hat{h}_{L2}}]$	0.0078	*0.0075*	*0.0075*	0.0076
Relative Improvement over $m = n$	8.2%	*11.8%*	*11.8%*	10.6%
$SD[ISE_{\hat{h}_{L2}}]$	0.0054	*0.0056*	*0.0056*	0.0065
m	70	80	90	100
$E[ISE_{\hat{h}_{L2}}]$	0.0078	0.0080	0.0082	0.0085
Relative Improvement over $m = n$	8.2%	5.9%	3.5%	0%
$SD[ISE_{\hat{h}_{L2}}]$	0.0070	0.0077	0.0082	0.0087

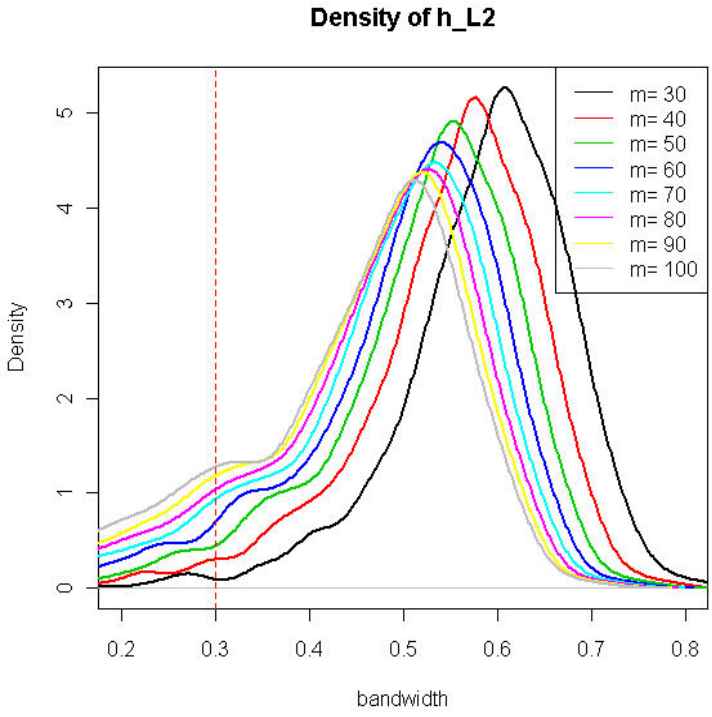

Figure 5.5. Density Estimates for $\hat{h}_{L^2}$

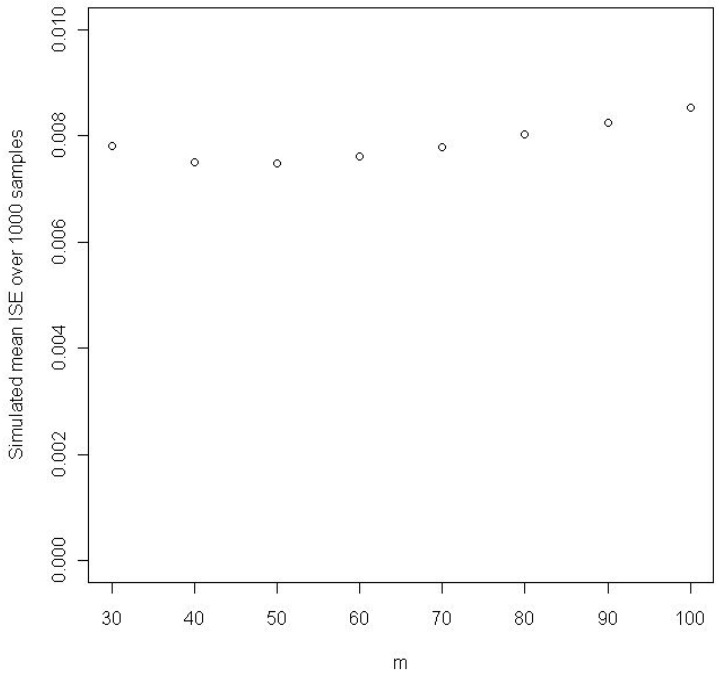

Figure 5.6. Mean $ISE(\hat{h}_{L^2})$ over the Samples

Conclusion:

Table 5.2 indicates that as m decreased the "optimal" bandwidth selected became larger. The ratio of means was about 1.20 when one took m as half of n. Furthermore, the average ISE was always smaller when one used the bandwidth selector $\hat{h}_{L^2}$ based on minimizing $\hat{Risk}_m(h), m < n$. What is more, by taking $m = n/2$, the average ISE was the smallest in the simulation study. From this aspect, the bandwidth selector based on the estimated L^2 risk with $m < n$ not only gave us a larger bandwidth that helps to alleviate undersmoothing problem of the kernel density estimator, improving the accuracy of the bandwidth selector, it also yielded smaller average integrated squared errors. Moreover, the variation in ISE was considerably reduced when $m < n$.

Figure 5.5 displays density estimates for distribution of $\hat{h}_{L^2}$ given different values of m. It indicates that the mean of $\hat{h}_{L^2}$ shifted to the right (i.e. becomes larger) as m got smaller. In addition, the density curve narrowed for smaller value of m. The differences in left-hand tails were striking: there were rare probabilities to

obtain "optimal" bandwidth less than 0.3 by minimizing $Risk_{L^2,m}(h)$ with respect to h when $m \leq n/2$, and this range of bandwidth choices usually result in the overfitting problem.

Figure 5.6 illustrates comparison of the mean $ISE(\hat{h}_{L^2})$ over samples given different values of m. It is clearly seen that when $m = n/2$ we had the smallest integrated squared errors in average over samples, with an improvement of about 11.8% over $m = n$. This discovery may lead us to an interesting investigation for future work which is to compare this approach with Hall's rescaled bagging methods introduced in [19].

5.5.3 Unbiased Estimate for Kullback-Leibler Risk

In this subsection, we are going to propose a U-statistic for the unbiased estimation for the risk based on Kullback-Leibler distance.

Recall the formulas in (5.4.13) and (5.4.14):

$$d(f, \hat{f}_h) = \int log \frac{f(x)}{\hat{f}_h(x, \mathcal{X}_n)} f(x)dx$$

where $f(x)$ is the true underlying density, and $\mathcal{X}_n$ is a data set of size n.

Also, remember that

$$\begin{aligned} Risk_{KL,n}(h) &= E_{\mathcal{X}_n}[d(f, \hat{f}_h(\mathcal{X}_n))] \\ &= E_{\mathcal{X}_n}\{\int [log f(x)] f(x)dx\} - E_{\mathcal{X}_n}\{\int [log \hat{f}_h(x, \mathcal{X}_n)] f(x)dx\}. \end{aligned}$$

Notice that minimizing the risk is equivalent to maximizing

$$\begin{aligned} \text{Negative Relative Risk}_{KL,n}(h) &= \int E_{\mathcal{X}_n}[log \hat{f}_h(x, \mathcal{X}_n)] f(x)dx \\ &= E_{\mathcal{X}_{n+1}}[log \hat{f}_h(X, \mathcal{X}_n)] \end{aligned}$$

Actually, instead of using the full data set $\mathcal{X}_n$, we can define a corresponding negative relative risk based on a subsample of size m, $\mathcal{X}_m$. That is, define

$$\text{Negative Relative Risk}_{KL,m}(h) = E_{\mathcal{X}_{m+1}}[log \hat{f}_h(X, \mathcal{X}_m)] \tag{5.5.11}$$

Now, consider a symmetric kernel function of order $m+1, m \leq n-1$

$$K_h^{\star\star}(\mathcal{X}_{m+1}) = \frac{1}{m+1}\sum_{i=1}^{m+1} log\hat{f}_h(X_i, \mathcal{X}_{<-i>}), \tag{5.5.12}$$

where $\mathcal{X}_{m+1}$ is a subset of size $m+1$ out of $\mathcal{X}_n = (X_1, ..., X_n)$, X_i is the ith observation in $\mathcal{X}_{m+1}$, and $\mathcal{X}_{<-i>}$ is the m observations in $\mathcal{X}_{m+1}$ except X_i.

Based on the above kernel function (5.5.12), we can construct a U-statistic which is an unbiased estimate of the "Negative Relative $Risk_{KL,m}(h)$", denoted as U_{KL}.

$$U_{KL} := \textbf{Negative Relative } \hat{\text{Risk}}_{KL,m}(h) = \frac{1}{\binom{n}{m+1}} \sum_{(n,m+1)} K_h^{\star\star}(X_{i_1}, ..., X_{i_{m+1}}) \tag{5.5.13}$$

In practice, due to the enormous number of possible subsamples of size $m+1$ when n is large, i.e.$\binom{n}{m+1}$, we can use an incomplete U-statistic to estimate the Negative Relative $\text{Risk}_{KL,m}(h)$. That is,

$$\tilde{U}_{KL,B} = \frac{1}{B}\sum_{b=1}^{B} K_h^{\star\star}(S_b), \tag{5.5.14}$$

where K_h is the kernel function defined in (5.5.12), and S_b is a sample of size $m+1$ out of $\mathcal{X}_n$.

By maximizing $\tilde{U}_{KL,B}$ (5.5.14), we can obtain the optimal bandwidth $\hat{h}_{KL}(m)$ which minimizes the estimated KL risk.

The full expression for $\tilde{U}_{KL,B}$ is:

$$\begin{aligned}
\tilde{U}_{KL,B} &= \frac{1}{B}\sum_{b=1}^{B} K_h^{\star\star}(S_b) = \frac{1}{B}\sum_{b=1}^{B}(\frac{1}{m+1}\sum_{i=1}^{m+1} log\hat{f}_h(S_{b(i)}, S_{b(-i)})) \\
&= \frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1} log(\frac{1}{\sqrt{2\pi}mh}\sum_{j\neq i} e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}) \\
&= \frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1} log(e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}) - log(\sqrt{2\pi}mh)
\end{aligned}$$

where $S_{b(i)}$ represents the ith component of the size-$m+1$ sample S_b, and $S_{b(-i)}$

represents the components in S_b except the ith one.

By taking the derivatives of $\tilde{U}_{KL,B}$ with respect to h and setting it to be 0, we have:

$$\begin{aligned}
0 &= \frac{d}{dh}\tilde{U}_{KL,B} \\
&= \frac{1}{h^3}\frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1}\frac{\sum_{j\neq i}(S_{b(i)}-S_{b(j)})^2 e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}}{\sum_{j\neq i} e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}} - \frac{1}{h} \\
&\Rightarrow \\
\frac{1}{h} &= \frac{1}{h^3}\frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1}\frac{\sum_{j\neq i}(S_{b(i)}-S_{b(j)})^2 e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}}{\sum_{j\neq i} e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}} \\
h^2 &= \frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1}\frac{\sum_{j\neq i}(S_{b(i)}-S_{b(j)})^2 e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}}{\sum_{j\neq i} e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}}
\end{aligned}$$

Notice that this is not an explicit representation of h. Consider the following target function/objective function

$$f_{KL,m}(h) = \frac{1}{B(m+1)}\sum_{b=1}^{B}\sum_{i=1}^{m+1}\frac{\sum_{j\neq i}(S_{b(i)}-S_{b(j)})^2 e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}}{\sum_{j\neq i} e^{-\frac{(S_{b(i)}-S_{b(j)})^2}{2h^2}}} - h^2 \qquad (5.5.15)$$

Given a value of m, the optimal bandwidth $\hat{h}_{KL}$ can be obtained by seeking the positive root of the above objective function (5.5.15).

5.5.3.1 A Simulation Study

In the following simulation study, I considered $R = 100$ samples of size 100 from $Normal(0, 1)$ and took $B = 1000$ in the target function $f_{KL,m}$. (i.e. I used 1000 i.i.d. samples of size $m+1$ ($m+1 \leq n = 100$) to construct the incomplete U-statistic (formula (5.5.14)) which is an unbiased estimator of the negative relative KL risk.) Furthermore, I considered $m = 30, 40, 50, ..., 100$. Given each m value, the root of the target function $f_{KL,m}(h)$ gave us the optimal bandwidth with respect to the m that minimizes the estimated KL risk.

Table 5.4. Relationship between m and $\hat{h}_{KL}$ by Minimizing Est KL Risk

$m+1$	30	40	50	60	70	80	90	100
$E[\hat{h}_{KL}]$	0.5658	0.5247	0.4951	0.4713	0.4502	0.4336	0.4169	0.4027
$SD[\hat{h}_{KL}]$	0.0656	0.0723	0.0775	0.0816	0.0883	0.0913	0.0950	0.0983

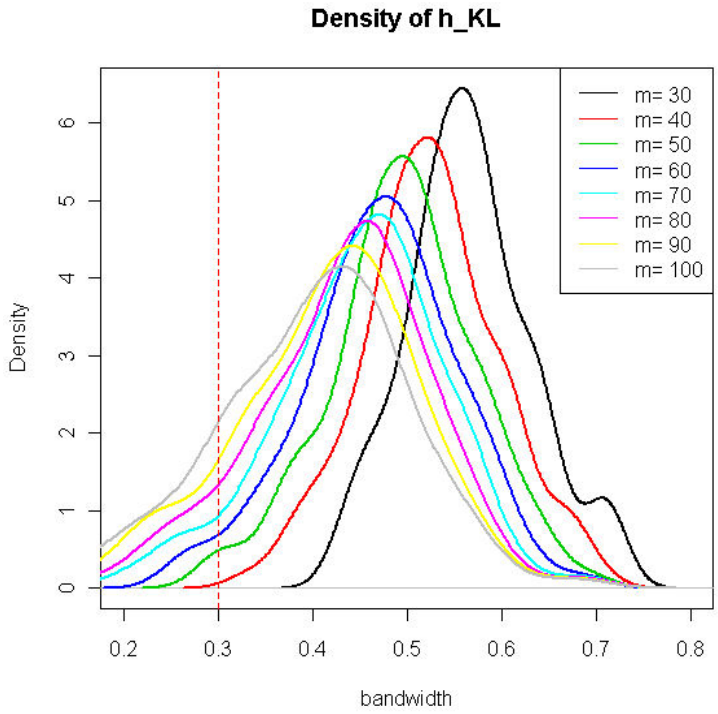

Figure 5.7. Density Estimates for $\hat{h}_{KL}$

Analogous to Table 5.3, here we can also compute the simulated mean Kullback-Leibler loss by using bandwidth $\hat{h}_{KL,m}$ dependent on the value of m. Recall the definition of Kullback-Leibler distance (5.4.13):

$$\begin{aligned} d(f, \hat{f}_h) &= \int log \frac{f(x)}{\hat{f}_h(x, \mathcal{X}_n)} f(x) dx \\ &= \int log[f(x)] f(x) dx - \int log[\hat{f}_h(x, \mathcal{X}_n)] f(x) dx. \end{aligned}$$

Notice that with normal assumption for f, the first term in the above expression can be solved explicitly and is about -1.419. Therefore, the Kullback-Leibler

distance between $\hat{f}_h$ and f then equals to

$$-1.419 - \int log[\hat{f}_h(x, \mathcal{X}_n)]f(x)dx.$$

By substituting h with the m-dependent bandwidth choice $\hat{h}_{KL,m}$ into the above expression, we can evaluate the performance of the kernel density estimator $\hat{f}_{\hat{h}_{KL,m}}(x, \mathcal{X}_n)$ at a given data set $\mathcal{X}_n$. In other words, we are interested in the measurement

$$-1.419 - \int log[\hat{f}_{\hat{h}_{KL,m}}(x, \mathcal{X}_n)]f(x)dx,$$

where the integral can be estimated by Monte Carlo Method. That is, we can draw independent samples $X_1, X_2, ..., X_B$ from the true underlying distribution F with density f and use

$$-1.419 - \frac{1}{B}\sum_{i=1}^{B} log[\hat{f}_{\hat{h}_{KL,m}}(X_i, \mathcal{X}_n)]$$

to approximate $d(\hat{f}_{\hat{h}_{KL,m}}, f)$.

Furthermore, by averaging over different samples, we can obtain the simulated mean Kullback-Leibler loss and compare the kernel density estimators that use different m-dependent bandwidths. Table 5.5 and Figure 5.8 below display the relative results.

Table 5.5. Relative Kullback-Leibler Loss ($\hat{h}_{KL}$)

$m+1$	30	40	50	60
$E[Relative\ Loss(\hat{h}_{KL})]$	0.0383	0.0354	0.0339	*0.0330*
Relative Improvement over $m+1=n$	-12.9%	-4.4%	0.0%	*2.7%*
$SD[Relative\ Loss(\hat{h}_{KL})]$	0.0059	0.0050	0.0042	*0.0038*
$m+1$	70	80	90	100
$E[Relative\ Loss(\hat{h}_{KL})]$	0.0330	0.0330	0.0332	0.0339
Relative Improvement over $m+1=n$	2.7%	2.7%	2.1%	0.0%
$SD[Relative\ Loss(\hat{h}_{KL})]$	0.0043	0.0050	0.0060	0.0081

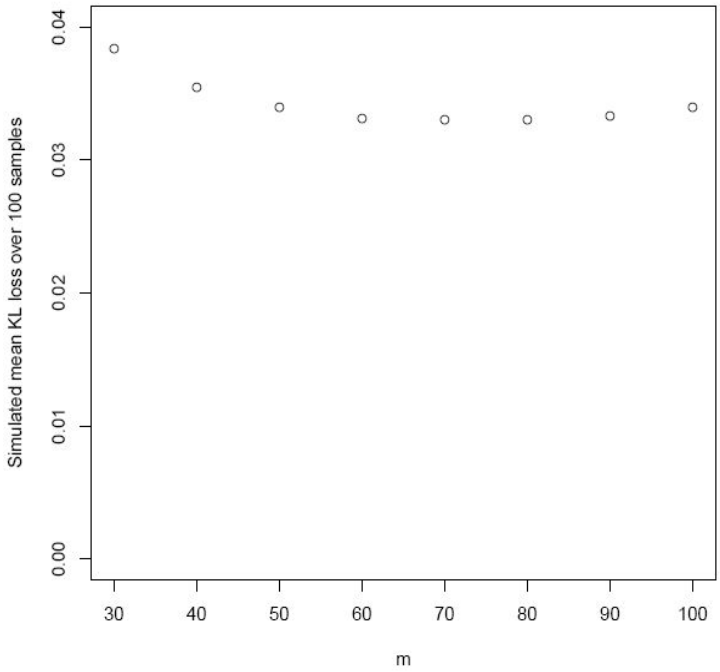

Figure 5.8. Mean Relative Kullback-Leibler Loss over the Samples

Conclusions:

The results are in Table 5.4, 5.5 and Figure 5.7, 5.8.

As m became smaller, the selected bandwidth increased with its standard deviation decreasing. That is, by using $m < n$ samples rather than the full data set to construct the unbiased KL risk estimator, we potentially ameliorate the problems of both sample fluctuation and density overfitting. Table 5.5 revealed that we can always improve the mean KL loss by using kernel size $m + 1$ bigger than $n/2$ and less than n. Especially, when $m + 1 = (3/5) \cdot n$, the simulated mean KL loss was the smallest, although its superiority compared with the case that $m + 1 = n/2$ was not dramatic. What is more, compared with the case discussed in Section 5.5.2, the gain of reducing mean loss by using subsample size $m < n$ to estimate the risk was not as very obvious here when use Kullback-Leibler distance.

Figure 5.7 displays the density estimates for $\hat{h}_{KL}$. The mean of $\hat{h}_{KL}$ shifted to the right as m became smaller. Meanwhile, the variance of $\hat{h}_{KL}$ reduced as m got larger. Most importantly, we can see the differences in the left tails. For $m \leq n/2$

there were rare probabilities to select bandwidths less than 0.3, and this range of values usually cause the undersmoothing problem.

When it comes to the comparison with the bandwidth selector based on L^2 loss, in general $\hat{h}_{KL}$ yields slightly smaller optimal bandwidths than the L^2 loss approach. Moreover, the former method also gives us bandwidth selector with slightly smaller standard deviation given the same value of m.

5.5.4 Implementing the Unbiased Variance Estimator

In order to investigate the performance of our proposed unbiased variance estimator under the context of risk estimation, we carried out simulation study by comparing it with some bootstrap variance estimators.

5.5.4.1 Variance of U_{L^2}

For the unbiased risk estimator dependent on L^2 loss, the U-statistic is constructed based on kernel $K_h^{\star}(x_1, x_2)$ (5.5.9). Because the kernel size is small, it is feasible to calculate the complete U-statistic and accomplish the unbiased variance estimation by (3.16). The simulation results are shown in Table 5.6.

	True (simulated)	Unbiased	Nonparametric	Balanced	Smoothed	Parametric
$\hat{E}(U_{L^2})$	-0.272916	-0.273623	-0.281437	-0.28136	-0.27955	-0.27552
$\hat{Var}(U_{L^2})$	0.000464	0.000467	0.000499	0.000502	0.000493	0.000478
$SD\{\hat{Var}(U_{L^2})\}$		1.525e-4	1.417e-4	1.416e-4	1.307e-4	5.850e-5
CompTime		40.76 hr	2.88 hr	3.02 hr	4.47 hr	2.33 hr

Table 5.6. Risk Based on L^2 Distance: $R = 200$ size-n samples ($n = 100$) are drawn independently from standard normal distribution. For each bootstrap algorithm, 1,000 resamples of size-n are considered. The simulated true values are based on 5,000 random samples on the basis of (5.5.15). The standard deviation for the Gaussian kernel (h) is taken to be the selected bandwidth by minimizing U_{L^2} when $m = n/2$.

As seen in Table 5.6, when kernel size is relative small compared with sample size n, the complete unbiased variance estimator does not have much gain in

terms of computation. However, its unbiasedness is obvious while compared with bootstrap variance estimators.

What is more, by referring to Hoeffding-decomposition, the variance of U_{L^2} has the following closed form

$$Var(U_{L^2}) = \frac{4}{\binom{n}{1}}\delta_1^2 + \frac{1}{\binom{n}{2}}\delta_2^2, \tag{5.5.16}$$

where $\delta_j^2 = Var[h^j(X_1, ..., X_j)](j = 1, 2)$, and $h^{(j)}(x_1, ..., x_j) = \phi_j(x_1, ..., x_j) - \sum_{c=1}^{j-1}\sum_{(j,c)} h^{(c)}(x_{\nu_1}, ..., x_{\nu_c}) - \theta$ are the orthogonal terms in H-decomposition .

In particular, with the form of $K_h^\star(x_1, x_2)$ in (5.5.9) determined by a Gaussian kernel, we have

$$\begin{aligned}\phi_1(x) &= \frac{1}{2hm\sqrt{\pi}} + \frac{m-1}{m}\frac{1}{\sqrt{2\pi(2h^2+1)}}e^{-x^2/(2(2h^2+1))} \\ &- \sqrt{\frac{2}{\pi(h^2+1)}}e^{-x^2/(2(h^2+1)}\end{aligned}$$

in this case.

Since here the kernel size is negligible compared with the sample size n, using the asymptotic variance as an estimate seems justifiable. The asymptotic variance for U_{L^2} is the first term in (5.5.16) and is about 0.000447 by simulation. As stated in Chapter 1, the asymptotic variance is always biased downwards, even though it is a reasonable alternative when the fraction of kernel size over sample size is 1/50. In addition, the asymptotic variance estimator is also smaller than the unbiased variance estimator on average.

5.5.4.2 Variance of U_{KL}

For the U-statistic risk estimator based on Kullback-Leibler loss, the kernel function is of order $n/2$. In this case, using the asymptotic variance estimator is clearly inappropriate. In addition, when sample size n is large, computing the closed form variance also becomes impractical. Therefore, it is reasonable for us to consider alternative methods, such as the proposed unbiased variance estimator.

However, one problem here is that the number of possible subsamples of size $n/2$ is truly huge for large n. Consequently, we will use the incomplete U-statistic

and apply the Type 2 resampling scheme (4.1.10) to realize the unbiased variance estimator. In Table 5.7, two different numbers of subsamples used to construct the incomplete U-statistic are considered and compared.

$B =$ 1000 subsamples	True (simulated)	Unbiased	Nonparametric	Balanced	Smoothed	Parametric
$\hat{E}(U_{KL})$	-1.44660	-1.44654	-1.42464	-1.42458	-1.43036	-1.46915
$\hat{Var}(U_{KL})$	0.005114	0.004919	0.004306	0.004327	0.004364	0.005336
$SD\{\hat{Var}(U_{KL})\}$		0.002258	0.001105	0.001126	0.001104	0.000249
CompTime		0.018 hr	8.40 hr	8.39 hr	8.58 hr	8.52 hr
$B =$ 5000 subsamples	True (simulated)	Unbiased	Nonparametric	Balanced	Smoothed	Parametric
$\hat{E}(U_{KL})$	-1.44660	-1.44661	-1.42463	-1.42599	-1.43023	-1.44853
$\hat{Var}(U_{KL})$	0.005114	0.004885	0.004317	0.004327	0.004330	0.005182
$SD\{\hat{Var}(U_{KL})\}$		0.002176	0.001086	0.001093	0.001080	0.000311
CompTime		0.089 hr	42.58 hr	42.33 hr	42.21 hr	42.69 hr

Table 5.7. Risk Based on Kullback-Leibler Distance: $R = 200$ size-n samples ($n = 100$) are drawn independently from standard normal distribution for each method. For each bootstrap algorithm, I resampled 1,000 times for each size-n sample. The standard deviation for the Gaussian kernel (h) is taken to be the selected bandwidth by minimizing U_{KL} when $m = n/2$.

When m/n is a relatively large fraction, the unbiased variance estimator is much less computationally intensive compared with bootstrap methods. Moreover, it provides us with an estimate with the highest accuracy, although bootstrap variance estimators tend to be more precise. Besides, we can also notice that the number of subsamples B used to construct the incomplete U-statistic does not seem to be crucial in improving the estimation.

Chapter 6

Future Work

6.1 Comparing Bootstrapping and Subsampling

First of all, notice that the unbiased variance estimator based on formula (3.1.6) 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 3.1 to 3.3 (and also Table 5.6 and 5.7), it can be seen that subsampling estimate (i.e. the unbiased variance estimate) can have smaller bias but larger standard deviation compared with its bootstrap counterpart. A similar conclusion can be made from Table 4 in Lindsay and Liu (2009) [25]. The difference between these two estimates become negligible as the effective sample size (i.e. n/m, where n is the sample size, and m represents the subsample size) gets large enough. We are expecting that there should exist a variance estimator in between bootstrap and subsampling that has smaller variation but has some small bias. Such an estimator should be a good trade-off solution for the subsampling and bootstrapping algorithms. In order to further understand the merits of these two approaches, we plan to compare their forms on the basis of Hoeffding-decomposition structure and investigate whether we can modify the dominating bootstrap-version decomposed variance term(s) somehow in order to attain an improved variance estimator.

6.2 Estimating Variance with $\hat{V}_u$

When it comes to the two resampling schemes, both of them are aimed to provide an unbiased realization of the unbiased variance estimator, $\hat{V}_u$. In this aspect, they should result in non-significantly different realizations. However, from another aspect, Type 2 resampling scheme tends to use more nonoverlapping pairs of size-m samples to estimate $Q(0)$. Especially when the resample size B is not large, the difference of the number of nonoverlapping pairs used in Type 2 scheme should be substantially larger than that used in Type 1 resampling scheme. The gain of Type 2 resampling scheme in terms of realizing $\hat{V}_u$ and estimating the variance of a U-statistic need to be thoroughly studied and stated via some simulation examples.

6.3 Forcing Positive Variance Estimations

The possible negative estimations based on the unbiased variance estimate formula (3.1.6) leads to three potential adjustments proposed in Chapter 3. However, there remains a question about how to realize these adjustments efficiently in practice. One natural thought is whether we could involve resampling idea into the realization of the proposed adjustments. Moreover, comparing the proposed improvements on $\hat{V}_u$ also deserves more efforts. For instance, some simulation examples should be developed to compare the performance of the complete unbiased variance estimator $\hat{V}_u$ with these adjusted non-negative variance estimators.

Bibliography

[1] G. Blom, Some Properties of Incomplete U-statistics, *Biometrica (1976)*, Vol.63, No.3, pp.573-580.

[2] A. Bowman, An Alternative Method of Cross-Validation for the Smoothing of Density Estimates, *Biometrika, 71 (1984)*, pp. 353-360.

[3] B.M. Brown, and D.G. Kildea, Reduced U-Statistics and the Hodges-Lehmann Estimator, *Institiute of Methematical Statistics (1978)*, Vol.6, No.4, pp. 828-835.

[4] M.R. Chernick, Bootstrap Methods: A Guide for Practitioners and Researchers, *Wiley Series in Probability and Statistics*, 2008.

[5] A. DasGupta, Asymptotic Theory of Statistics and Probability, *Springer Texts in Statistics*, 2008.

[6] A.C. Davison, D.V. Hinkley, and E. Schechtman, Efficint Bootstrap Simulation, *Biometrika (1986)*, Vol.73, No.3, pp.555-566.

[7] B. Efron, Bootstrap Methods: Another Look at The Jackknife, *The Annals of Statistics (1979)*, Vol.7, No.1, pp.1-26.

[8] B. Efron, and C. Stein, The Jackknife Estimate of The Variance, *The Annals of Statistics (1981)*, Vol.9, No.3, pp.586-596.

[9] B. Efron, The Introduction to Bootstrap, *Chapman and Hall*, 1982.

[10] W. Feller, An Introduction to Probability Theory and It Applications, *Wiley Series in Probability*, 1967, pp.194.

[11] R.A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics (1936)*, Vol. 7, No. 2., pp. 179-188.

[12] E. Fix and J.L. Hodges, Discriminatory Analysis. Nonparametric Discrimination; Consistency Properties. *Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.* (Reprinted as pp. 261-279 of Agrawala, 1977.)

[13] R.E. Folsom, Probability Sample U-Statistics: Theory and Applications for Complex Sample Designs, *Presented at the American Statistical Association Meeting, Section on Survey Research Methods*, 1986.

[14] D.A.S. Fraser, Completeness of Order Statistics, *Canadian Journal of Mathematics 6*, pp. 42-45.

[15] W. Hardel, M. Muller, S. Sperlich and A. Werwatz, Nonparametric and Semiparametric Models, *Springer*, 1994.

[16] P. Hall and J.S. Marron, Lower Bounds for Bandwidth Selection in Density Estimation, *Probability Theory and Related Fields, 90 (1991a)*, pp. 149-173.

[17] P. Hall, J.S. Marron, and B.U. Park, Smoothed Cross-Validation, *Probability Theory and Related Fields*, 1992.

[18] P. Hall, The Bootstrap and Edgeworth Expansion, *Springer*, 1999.

[19] P. Hall, and A.P. Robinson, Reducing Variability of Crossvalidation for Smoothing-parameter Choice, *Biometrika (2009)*, Vol.96, No.1, pp.175-186.

[20] P.R. Halmos, The Theory of Unbiased Estimation, *The Annals of Mathmatical Statistics (1946)*, Vol.17, No.1, pp.34-43.

[21] W. Hoeffding, A Class of Statistics with Asymptotically Normal Distribution, *Institute of Mathematical Statistics (1948)*, Vol.19, No.3, pp.293-325.

[22] M.C. Jones, J.S. Marron, and B.U. Park (1991), A Simple Root n Bandwidth Selector, *Annals of Statistics, 19 (1991)*, pp. 1919-1932.

[23] J. Kowalski, and X.M. Tu, Modern Applied U-Statistics, *Wiley*, 2008.

[24] A.J. Lee, U-Statitics–Theory and Practice, *Dekker*, 1990.

[25] B.G. Lindsay and J. Liu, Model Assessment Tools for a Model False World, *Statistical Science, 24*, pp. 303-318.

[26] C.R. Loader, Bandwidth Selection: Classical or Plug-in?, *Annals of Statistics, Vol.27 No.2, 1999*, pp.415-438.

[27] J.S. Marron and M.P. Wand, Exact Mean Integrated Squared Error, *Annals of Statistics, 20 (1992)*, pp. 712-736.

[28] A.M. Polansky, Upper Bounds on the True Coverage of Bootstrap Percentile Type Confidence Interval, *The American Statistician 53*, pp.362-369.

[29] D.N. Politis, J.P. Romano, and M. Wolf, Subsampling, *Springer*, 1999.

[30] S. Ray, and B.G. Lindsay, Model Selection in High-Dimensions: A Quadratic-risk Based Approach, *Journal of the Royal Statistical Society-Series B (2008)*, Vol. 70, Part 1, pp.95-118.

[31] M. Rudemo, Empirical Choice of Histograms and Kernel Density Estimators, *Scandanavian Journal of Statistics, 9 (1982)*, pp. 65-78.

[32] D.W. Scott and G.R. Terrell, Biased and Unbiased Cross-Validation in Density Estimation, *Journal of American Statistical Association 82 (1987)*, pp. 1131-1146.

[33] B.W. Silverman, Density Estimation for Statistics and Data Analysis, *Monographs on Statistics and Applied Probability, London: Chapman and Hall (1986).*

[34] B.A. Turlach, *C.O.R.E. and Institut de Statistique (1993)*, pp. 23-493.

[35] J. Wu, Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis, *The Annals of Statistics (1986)*, Vol.14, No.4, pp.1261-1295.