Extrapolation Techniques in U-statistic Variance Estimation

Qing Wang*
Department of Mathematical Sciences, Bentley University, Waltham, MA 02452

Abstract
This paper considers the problem of variance estimation of a U-statistic. Extrapolation techniques are proposed to overcome the drawback of having negative values when using the unbiased U-statistic variance estimator (Wang and Lindsay, 2014). Following the proposal of the linearly extrapolated variance estimator (Wang and Chen, 2015), we consider nonlinear extrapolation method and devise a variance estimator that is nearly second-order unbiased. Simulation studies indicate that the second-order extrapolated variance estimator has smaller mean squared error compared to the unbiased variance estimator and the jackknife variance estimator across a wide selection of distributions. We also discuss the advantage of the proposal compared to its jackknife counterpart in regression analysis and model selection.

Key Words: Hoeffding-decomposition, linear extrapolation, nonlinear extrapolation, variance estimation, unbiased, U-statistic

1. Introduction

Let $X_1, \ldots, X_n$ be an independent, identically distributed sample from some distribution. A U-statistic with kernel function ϕ of k components is defined as (Hoeffding, 1948)

$$U_n = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} \phi(X_{i_1}, \ldots, X_{i_k}).$$

Without loss of generality, let ϕ be a symmetric function that is permutation invariant in its k components. Since function ϕ is often scalar-valued in applications, we focus on the case that $\phi \in \mathcal{R}$. We call k the kernel size of U_n; it is the smallest integer such that $E\{\phi(X_1, \ldots, X_k)\} = \theta$, where θ is the parameter of interest. The kernel size k is also referred to as the degree of U_n. U-statistic is an unbiased estimator for parameter θ. In the context of nonparametric inference where the set of order statistics $(X_{(1)}, \ldots, X_{(n)})$ is the complete sufficient statistic (Fraser, 1954), U_n is the minimum-variance unbiased estimator. Because most unbiased estimators in common use have a U-statistic representation, obtaining a reliable estimator for the variance of U_n is crucial in statistical inference and practical applications.

Hoeffding (1948) gives the closed-form expression for the variance of U_n:

$$\text{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{k-c} \sigma_c^2, \tag{1.1}$$

where $\sigma_c^2 = \text{Var}\{\phi_c(X_1, \ldots, X_c)\}$, and $\phi_c(x_1, \ldots, x_c) = E\{\phi(X_1, \ldots, X_k) X_1 = x_1, \ldots, X_c = x_c\}$ for $1 \le c \le k$. It was also shown that U_n admits an asymptotic normal distribution with asymptotic variance $k^2\sigma_1^2/n$, provided that ϕ is twice integrable and $0 < \sigma_1^2 < \infty$. However, the exact U-statistic variance (1.1) is complicated in form, and the asymptotic variance of U_n is not necessarily reliable when the kernel size k is not small

*Part of the research was done when the author was an Assistant Professor of Statistics at Williams College in Williamstown, Massachusetts.

compared to the sample size n. Wang and Lindsay (2014) propose an unbiased variance estimator of U_n, denoted as $\hat{V}_u$, that is of a simple quadratic form and is applicable as long as the ratio $k/n \leq 1/2$.

Let N_c be the number of pairs of subsamples of size k that have at most c overlaps $(0 \leq c \leq k)$, and let $O(S_a, S_b)$ be the number of overlaps between subsets S_a and S_b. The unbiased variance estimator (Wang and Lindsay, 2014) is defined as

$$\hat{V}_u = Q(k) - Q(0), \tag{1.2}$$

where $Q(c) = N_c^{-1} \sum_{O(S_a,S_b)\leq c} \phi(S_a)\phi(S_b)$ for $0 \leq c \leq k$. The investigation of an unbiased variance estimator of U_n has occurred in some previous literature, such as Folsom (1984) and Maesono (1998). It can be shown that the various proposals of the unbiased variance estimator of U_n are equivalent. However, the definition of $\hat{V}_u$ in Wang and Lindsay (2014) is of a much simpler form.

Although $\hat{V}_u$ is unbiased and easy to compute with the help of partition resampling scheme (Wang and Lindsay, 2014), it is possible for $\hat{V}_u$ to yield negative values (see Example 2 in Wang and Lindsay (2014)). To overcome this drawback, Wang and Chen (2015) propose a general class of linearly extrapolated variance estimators that are non-negative. The class of linearly extrapolated variance estimators (Wang and Chen, 2015) can be viewed as generalization of the leave-one-out jackknife variance estimator (Efron and Stein, 1981), but they are more computationally efficient than the jackknife estimator in half-sampling cross-validation problems. Let U_m denote a U-statistic computed based on a subsample of size m. Building upon the linear extrapolation method, we extend the approximate relationship between Var(U_n) and Var(U_m) from linear to a nonlinear form. We will show that with the help of second-order extrapolation technique the resulting U-statistic variance estimator is nearly second-order unbiased. Thus, it is more accurate than the linearly extrapolated variance estimator.

The rest of the paper is organized as follows: We first introduce two extrapolated variance estimators of a U-statistic in Section 2, using first-order and second-order extrapolation techniques respectively. In Section 3 we present a simulation study to confirm that the extrapolated variance estimators are always non-negative, while the unbiased variance estimator in Wang and Lindsay (2014) might yield negative values. In addition, we compare the performance of the extrapolated variance estimators with the unbiased variance estimator and the conventional jackknife variance estimator in terms of bias, variance, and mean squared error in a study of assessing the variance of the unbiased sample variance, where the data are generated from a wide selection of distributions. In Section 4 we discuss the advantage of the proposed extrapolated variance estimators in comparison to its jackknife counterpart in the context of regression analysis and model selection. We will conclude this paper with some discussions in Section 5.

2. Extrapolation Techniques in Variance Estimation

In this section we consider practical solutions for the problem of having possibly negative values when using the unbiased variance estimator $\hat{V}_u$ (1.2). We propose to first estimate the variance of a U-statistic at a subsample size m, also referred to as a *fictional sample size*, that may be smaller than the original sample size n, and then extrapolate the variance estimator from m to n to remove the bias incurred in the subsampling stage. The extrapolated variance estimator is always non-negative. Moreover, we anticipate that this subsampling plus extrapolation methodology can help to reduce the variation of the variance estimator, similar as what has been seen in the context of kernel density bandwidth selection in Marron (1987), Hall and Robinson (2009), and Wang and Lindsay (2015).

2.1 First-order Extrapolation Technique

Let $U_{n-1}^{(-i)}$ be a U-statistic defined on a data subset of size $n-1$, without the ith observation. The conventional jackknife estimator in the context of U-statistic variance estimation is defined by

$$\hat{V}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(U_{n-1}^{(-i)} - \frac{1}{n}\sum_{j=1}^{n}U_{n-1}^{(-j)}\right)^2.$$

Efron and Stein (1981) consider $\hat{V}_J$ as a linearly extrapolated variance estimator. The term $\sum_{i=1}^{n}\left(U_{n-1}^{(-i)} - \frac{1}{n}\sum_{j=1}^{n}U_{n-1}^{(-j)}\right)^2$ is viewed as a estimator for Var(U_{n-1}), and $(n-1)/n$ is a constant multiplier for adjusting the difference between Var(U_n) and Var(U_{n-1}). It is well known that the jackknife variance estimator is often less computationally expensive than the bootstrap methods, but it is always biased upwards.

Following the footsteps of Efron and Stein (1981), Wang and Chen (2015) study a general class of linearly extrapolated variance estimators. In the context of U-statistic variance estimation, one first constructs an unbiased variance estimator for a U-statistic at subsample size m ($m \leq n/2$), denoted as $\hat{V}_m$, and then extrapolates it from m to n based on a linear approximate relationship. Denote the U-statistic computed based on a subsample S of size m as $U_m(S)$. The first-order extrapolated variance estimator can be expressed as

$$\hat{V}_{\text{ex1}} = \frac{m}{n}\left\{\binom{n}{m}\binom{n-m}{m}\right\}^{-1}\sum_{O(S_a,S_b)=0}\frac{\{U_m(S_a)-U_m(S_b)\}^2}{2}, \tag{2.1}$$

where S_a and S_b are subsamples of size m, and $O(S_a, S_b)$ is the number of overlapping elements between these two data subsets. Note that

$$\hat{V}_m := \left\{\binom{n}{m}\binom{n-m}{m}\right\}^{-1}\sum_{O(S_a,S_b)=0}\frac{\{U_m(S_a)-U_m(S_b)\}^2}{2} \tag{2.2}$$

is an unbiased estimator for Var(U_m). The condition of $m \leq n/2$ is crucial for the construction of an unbiased variance estimator for a U-statistic at fictional size m. When n and m are both large, one can approximate $\hat{V}_{\text{ex1}}$ by independently drawing B disjoint pairs of subsamples $(S_{b,1}, S_{b,2})$ $(1 \leq b \leq B)$. Then, $\hat{V}_{\text{ex1}}$ can be approximated by

$$\hat{V}_{\text{ex1}}^B = \frac{m}{n}\frac{1}{B}\sum_{b=1}^{B}\frac{(U_m(S_{b,1})-U_m(S_{b,2}))^2}{2}. \tag{2.3}$$

Remark 1. *By construction the linearly extrapolated variance estimator, $\hat{V}_{ex1}$ or $\hat{V}_{ex1}^B$, is an average of square differences, and therefore is always non-negative.*

Let h_c $(1 \leq c \leq k)$ be the cth orthogonal term in Hoeffding decomposition (Hoeffding, 1948; Lee, 1990), defined by

$$h_c(x_1,\ldots,x_c) = \phi_c(x_1,\ldots,x_c) - \sum_{l=1}^{c-1}\sum_{(c,l)}h_l(x_{i_1},\ldots,x_{i_l}) - \theta,$$

and $h_1(x_1) = \phi_1(x_1) - \theta$. The closed-form expression of Var(U_n) in (1.1) can be equivalently written as

$$\text{Var}(U_n) = \sum_{c=1}^{k}\binom{n}{c}^{-1}\binom{k}{c}^2\delta_c^2, \tag{2.4}$$

where $\delta_c^2 = \text{Var}(h_c)$ $(1 \le c \le k)$. Thus, for any $m \le n$

$$\text{Var}(U_m) = \sum_{c=1}^{k} \binom{m}{c}^{-1} \binom{k}{c}^2 \delta_c^2.$$

Because

$$\binom{n}{c}^{-1} \binom{k}{c}^2 < (m/n) \binom{m}{c}^{-1} \binom{k}{c}^2$$

for $m < n$ and $c \ge 2$. We have $\text{Var}(U_n) < (m/n)\text{Var}(U_m)$. Thus, we introduce some positive bias by extrapolating the variance from m to n based on an approximate linear relationship.

Remark 2. *Wang and Chen (2015) express the expectation of $\hat{V}_{ex1}$ in terms of the variance of the orthogonal terms in Hoeffding decomposition:*

$$E(\hat{V}_{ex1}) = \frac{k^2}{n}\delta_1^2 + \frac{m}{n}\sum_{c=2}^{k} \binom{k}{c}^2 \binom{m}{c}^{-1} \delta_c^2.$$

Thus, $\hat{V}_{ex1}$ is first-order unbiased. Its second-order bias can be written as

$$\textit{second-order bias}(\hat{V}_{ex1}) = \binom{k}{2}^2 \left\{ \frac{2}{n(m-1)} - \frac{2}{n(n-1)} \right\} \delta_2^2.$$

The positive bias becomes larger as m gets smaller for a fixed sample size n, or when the sample size n gets smaller for a fixed ratio of m/n.

The construction of $\hat{V}_{\text{ex1}}$ is based on an approximate linear relationship between $\text{Var}(U_m)$ and $\text{Var}(U_n)$, which may not be accurate for relatively small value of n. In the following we will consider a nonlinear approximate relationship that aims to reduce the bias in $\hat{V}_{\text{ex1}}$, especially for small sample size n.

2.2 Second-order Extrapolation Technique

Below we will show how to construct a second-order extrapolated estimator for $\text{Var}(U_n)$ by referring to the closed-form expression of the U-statistic variance. Without loss of generality, we demonstrate the method by extrapolating from two fictional sizes, $m = n/2$ and $m = n/4$.

We first rewrite the U-statistic variance as follows:

$$\text{Var}(U_n) = \sum_{c=1}^{k} \binom{n}{c}^{-1} \binom{k}{c}^2 \delta_c^2 = \frac{1}{n}\sigma^2 \left\{ 1 + \frac{a}{n-1} + O(1/n^2) \right\},$$

where $\sigma^2 = k^2\delta_1^2$, and $a = (k-1)^2\delta_2^2/(2\delta_1^2)$. One can approximate $\text{Var}(U_n)$ by $(1/n)\sigma^2 e^{a/(n-1)}$. For simplicity of estimating the unknown parameter a, we write the approximate variance at size n as

$$V_n := (1/n)\sigma^2 e^{a/n}.$$

At fictional sizes $m = n/4$ and $m = n/2$, we have

$$V_{n/4} = \frac{4}{n}\sigma^2 e^{4a/n} \text{ and } V_{n/2} = \frac{2}{n}\sigma^2 e^{2a/n}.$$

Then, $V_{n/2}/V_{n/4} = (1/2)e^{-2a/n}$ and $V_n/V_{n/2} = (1/2)e^{-a/n}$. Thus,

$$V_n = \frac{1}{2}V_{n/2}e^{-a/n} = \frac{1}{2}V_{n/2}\sqrt{2V_{n/2}/V_{n/4}}. \tag{2.5}$$

One can view $\sqrt{2V_{n/2}/V_{n/4}}$ as a shrinkage factor; it equals 1 if the ratio of $V_{n/2}/V_{n/4}$ is exactly 1/2. Denote the second-order extrapolated variance estimator of Var(U_n) as $\hat{V}_{\text{ex2}}$. It is defined by

$$\hat{V}_{\text{ex2}} = (1/\sqrt{2})\hat{V}_{n/2}^{3/2}\hat{V}_{n/4}^{-1/2} = \hat{V}_{\text{ex1}}\sqrt{2\hat{V}_{n/2}/\hat{V}_{n/4}}, \tag{2.6}$$

where $\hat{V}_{n/2}$ and $\hat{V}_{n/4}$ are unbiased estimators for Var$(U_{n/2})$ and Var$(U_{n/4})$, as defined in equation (2.2) when setting $m = n/2$ and $m = n/4$ respectively. As the definition of $\hat{V}_m$ $(m \leq n/2)$ is straightforward, the realization of $\hat{V}_{\text{ex2}}$ is simple in structure and does not involve higher-order computational cost compared to $\hat{V}_{\text{ex1}}$.

Remark 3. *The second-order extrapolated variance estimator can be expressed as* $\hat{V}_{ex2} = \hat{V}_{ex1}e^{-\hat{a}/n}$*, where* $e^{-\hat{a}/n} = \sqrt{2\hat{V}_{n/2}/\hat{V}_{n/4}}$*. Because* $0 < e^{-\hat{a}/n} < 1$*,* $\hat{V}_{ex2}$ *is smaller than* $\hat{V}_{ex1}$ *and therefore corrects the positive bias in* $\hat{V}_{ex1}$*.*

Theorem 1. *Let* $\hat{V}_{ex2}$ *be the second-order extrapolated variance estimator as defined in equation* (2.6)*. Under weak regularity conditions so that* $\hat{V}_m$ (2.2) *has finite variance,* $\hat{V}_{ex2}$ *is nearly second-order unbiased. Thus, it has smaller bias compared to the linearly extrapolated variance estimator* $\hat{V}_{ex1}$*.*

For proof, please see Appendix.

Remark 4. *Hinkley (1977) and Wu (1986)) show that the jackknife variance estimator yields poor performance with large positive bias in regression analysis. The numerical study in Wang and Chen (2015) reveals that the first-order extrapolated variance estimator has much smaller bias than its jackknife counterpart in assessing the variance of a U-statistic risk estimate for a parametric model. In addition, the linearly extrapolated variance estimator has significant computational advantage than the jackknife method in half-sampling cross-validation problems. The computational cost of* $\hat{V}_{ex2}$ *is of the same order as that for* $\hat{V}_{ex1}$*. Thus, the second-order extrapolated variance estimator is superior to the conventional jackknife method in the regression context. We will demonstrate a simulation comparison in regression risk estimation in Section 4.*

3. Simulation Study

In this section we study the numerical performance of the proposed second-order extrapolated variance estimator in comparison to the linearly extrapolated variance estimator $\hat{V}_{\text{ex1}}$, the unbiased variance estimator $\hat{V}_u$, and the jackknife variance estimator $\hat{V}_J$. We consider a simple but practical scenario where the the parameter of interest θ is the variance of the underlying distribution. The U-statistic estimate for the variance is the unbiased sample variance, i.e. $U_n = S^2 = \{1/(n-1)\}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Our goal is to evaluate the variance of the unbiased sample variance, denoted as Var(U_n).

Table 1: A list of distributions under consideration

Name	Description
Standard Normal	Normal with mean 0 and standard deviation 1
Mixture 1	Binomal normal mixture: $0.5N(-1.5,1)+0.5N(1.5,1)$
Mixture 2	Normal mixture with an outlying mode: $0.5N(0,1)+0.5N(0,0.1)$
Gamma(5,2)	Gamma with shape parameter 5 and scale parameter 2
t(10)	t distribution with 10 degrees of freedom

We randomly generate $R = 500$ samples of size n ($n = 10, 30, 50, 100$) from some distribution. To investigate whether the proposed variance estimator is robust to outlying, bimodal, skewed, or heavy-tailed features of the data, a list of five different distributions are considered with descriptions shown in Table 1. Here we choose relatively small sample size n, as the difference between various variance estimators diminishes as n gets large. For each given sample of size n, we compute the proposed first-order and second-order extrapolated variance estimators. In the realization of $\hat{V}_{\text{ex1}}$ (2.3), we set $B = 1000$ and consider subsample sizes $m = n/2$ and $m = n/4$. The calculation of $\hat{V}_{\text{ex2}}$ is based on formula (2.6). We also compute the unbiased variance estimator $\hat{V}_u$ proposed in Wang and Lindsay (2014). The unbiased variance estimator $\hat{V}_u$ is realized based on equation (1.2) for $n = 10$ and $n = 30$. When $n = 50$ or 100, the calculation of $Q(0)$ in equation (1.2) involves an average of $\binom{n}{2}\binom{n-2}{2}$ terms, which is computational expensive to realize. Therefore, for $n = 50$ and 100 we approximate $Q(0)$ using C randomly generated pairs of disjoint subsets of size two. We consider four different values of C, i.e. 1000, 10000, 100000, and 1000000, in the following simulation comparison.

3.1 Possible negative values in $\hat{V}_u$

Table 2: Number of negative values produced by $\hat{V}_u$ out of 500 samples

Sample size	Normal	Mixture 1	Mixture 2	Gamma(5,2)	t(10)
$n = 10$	0	1	0	0	0
$n = 30$	0	0	0	0	0
$C = 1000$					
$n = 50$	0	0	0	1	12
$n = 100$	2	23	0	33	79
$C = 10000$					
$n = 50$	38	89	7	18	25
$n = 100$	130	166	67	97	94
$C = 100000$					
$n = 50$	0	4	0	0	0
$n = 100$	13	36	1	7	3
$C = 1000000$					
$n = 50$	0	0	0	0	0
$n = 100$	0	0	0	0	0

We first look at the number of negative values produced by $\hat{V}_u$ out of the 500 replications. Table 2 shows that it is possible for the unbiased variance estimator to yield negative values. Even when the exact formula (1.2) is used for small sample size $n = 10$, one sample generated from the bimodal normal mixture distribution yields a negative value of $\hat{V}_u$. When

$n = 50$ or 100, the number of negative values of $\hat{V}_u$ increases significantly as $Q(0)$ is approximated by a smaller number of disjoint random subsets. Only when C increases to one million do all the values of $\hat{V}_u$ become positive. Although Wang and Lindsay (2014) discuss possible fix-ups for the issue of having negative values of $\hat{V}_u$, the forced-positive variance estimator is quite liberal with very small value. This may lead to undesirable large probability of committing Type I error when used to construct a test statistic in a statistical hypothesis test setting. Thus, whenever the unbiased variance estimator gives negative estimates, the extrapolated variance estimators may be considered. Furthermore, it will be seen later in Table 3 and Table 4 that even when the unbiased variance estimator produces positive variance estimates, the second-order extrapolated variance estimator yields comparable or even better performance in achieving a smaller mean squared error.

3.2 Comparison between Extrapolated Variance Estimators and Unbiased Variance Estimator

We now compare the performance between the extrapolated variance estimators and the unbiased variance estimators in terms of mean, standard deviation, and mean squared error. We summarize in Table 3 and Table 4 the simulation results. Since the jackknife variance estimator can be viewed as a linearly extrapolated variance estimator, we also include the results of $\hat{V}_J$ in the tables below. When $n = 50$ or 100 we approximate $Q(0)$ in $\hat{V}_u$ using $C = 1000000$ disjoint subsamples. From Table 2 we know that with $C = 1000000$ the unbiased variance estimator $\hat{V}_u$ yields positive values for all distributions under consideration. However, we will see later that even with positive values $\hat{V}_u$ does not outperform the second-order extrapolated variance estimator.

When considering the linearly extrapolated variance estimator $\hat{V}_{\text{ex1}}$ at different fictional sizes, using $m = n/2$ leads to the best result. This agrees with the fact that $m = n/2$ leads to the smallest bias in $\hat{V}_{\text{ex1}}$, as shown in Wang and Chen (2015). The second-order extrapolated variance estimator seems to correct the positive bias in $\hat{V}_{\text{ex1}}$; its superiority in terms of bias is significant for small sample size n. The unbiased variance estimator tends to have larger variation compared to the second-order extrapolated variance estimator. Wang and Lindsay (2014) show that $\hat{V}_u$ has a U-statistic expression itself. Thus, $\hat{V}_u$ is the minimum-variance unbiased variance estimator in the context of nonparametric inference. However, our numerical results indicate that a smaller variance and a smaller mean squared error could be achieved by relaxing the unbiasedness condition. Moreover, besides the well-known positive bias of the jackknife variance estimator, the jackknife method seems more variable than the extrapolated variance estimators. Overall, the second-order extrapolated variance estimator is a clear winner in achieving a smaller standard deviation and mean squared error across different sample sizes and distributions. The advantage of using second-order extrapolation is particularly obvious when the sample size n is small. The performance of these variance estimators become more and more similar as the sample size n increases.

Table 3: Comparison of different variance estimators. The smallest mean squared error for each distribution and sample size is highlighted in bold.

Normal						
Sample size n		$\hat{V}_u$	$\hat{V}_{\text{ex1}}$		$\hat{V}_{\text{ex2}}$	$\hat{V}_J$
			$(m = n/2)$	$(m = n/4)$		
10	Mean	0.2025	0.2291	0.4615	0.1650	0.2286
	SD	0.2350	0.2530	0.4476	0.1973	0.2514
(Truth: 0.2222)	MSE	0.0556	0.0641	0.2576	**0.0422**	0.0633
30	Mean	0.0653	0.0677	0.0791	0.0628	0.0677
	SD	0.0455	0.0457	0.0522	0.0430	0.0464
(Truth: 0.0690)	MSE	0.0021	0.0021	0.0028	**0.0019**	0.0022
50	Mean	0.0395	0.0405	0.0441	0.0388	0.0405
	SD	0.0209	0.0212	0.0225	0.0207	0.0211
(Truth: 0.0408)	MSE	0.0004	0.0004	0.0005	**0.0004**	0.0004
100	Mean	0.0197	0.0199	0.0203	0.0198	0.0200
	SD	0.0078	0.0075	0.0075	0.0075	0.0074
(Truth: 0.0202)	MSE	0.0001	0.0001	0.0001	**0.0001**	0.0001
Mixture 1						
10	Mean	1.3863	1.6941	4.1513	1.1193	1.6881
	SD	1.3992	1.5177	2.9510	1.1538	1.5027
(Truth: 1.2872)	MSE	1.9674	2.4689	16.9120	**1.3594**	2.4188
30	Mean	0.3876	0.4134	0.5168	0.3706	0.4138
	SD	0.1908	0.1980	0.2293	0.1862	0.1968
(Truth: 0.3673)	MSE	0.0368	0.0413	0.0749	**0.0347**	0.0409
50	Mean	0.2303	0.2393	0.2708	0.2252	0.2391
	SD	0.0922	0.0906	0.0985	0.0881	0.0907
(Truth: 0.2286)	MSE	0.0085	0.0083	0.0115	**0.0078**	0.0083
100	Mean	0.1125	0.1134	0.1181	0.1113	0.1114
	SD	0.0385	0.0289	0.0301	0.0289	0.0287
(Truth: 0.1113)	MSE	0.0015	0.0008	0.0010	**0.0008**	0.0008

Table 4: Comparison of different variance estimators. The smallest mean squared error for each distribution and sample size is highlighted in bold.

Sample size n		$\hat{V}_u$	$\hat{V}_{\text{ex1}}$		$\hat{V}_{\text{ex2}}$	$\hat{V}_J$
			$(m = n/2)$	$(m = n/4)$		
Mixture 2						
10	Mean	0.1580	0.1649	0.2535	0.1343	0.1651
	SD	0.3470	0.3525	0.4764	0.3062	0.3519
(Truth: 0.1504)	MSE	0.1205	0.1244	0.2376	**0.0940**	0.1241
30	Mean	0.0451	0.0458	0.0507	0.0437	0.0457
	SD	0.0486	0.0492	0.0531	0.0475	0.0489
(Truth: 0.0421)	MSE	0.0024	0.0024	0.0029	**0.0023**	0.0024
50	Mean	0.0266	0.0267	0.0284	0.0260	0.0267
	SD	0.0216	0.0214	0.0224	0.0210	0.0214
(Truth: 0.0250)	MSE	0.0005	0.0005	0.0005	**0.0004**	0.0005
100	Mean	0.0127	0.0128	0.0128	0.0127	0.0128
	SD	0.0066	0.0065	0.0066	0.0066	0.0065
(Truth: 0.0125)	MSE	0.0000	0.0000	0.0000	**0.0000**	0.0000
Gamma(5,2)						
10	Mean	0.5325	0.5288	0.9429	0.4058	0.5766
	SD	1.2445	1.0993	1.6048	0.9296	1.2845
(Truth: 0.5345)	MSE	1.5488	1.2084	2.7423	**0.8807**	1.6517
30	Mean	0.1592	0.1607	0.1831	0.1509	0.1630
	SD	0.2363	0.2176	0.2415	0.2072	0.2379
(Truth: 0.1704)	MSE	0.0559	0.0474	0.0585	**0.0433**	0.0566
50	Mean	0.0917	0.0933	0.1007	0.0901	0.0928
	SD	0.1004	0.0974	0.1034	0.0949	0.1003
(Truth: 0.1014)	MSE	0.0102	0.0096	0.0107	**0.0091**	0.0101
100	Mean	0.0483	0.0491	0.0497	0.0489	0.0489
	SD	0.0416	0.0385	0.0388	0.0386	0.0412
(Truth: 0.0504)	MSE	0.0017	0.0015	0.0015	**0.0015**	0.0017
t(10)						
10	Mean	0.5518	0.5964	1.0462	0.4585	0.5962
	SD	2.0676	2.1075	2.8910	1.8100	2.1020
(Truth: 0.5049)	MSE	4.2770	4.4501	8.6508	**3.2783**	4.4269
30	Mean	0.1569	0.1606	0.1824	0.1507	0.1607
	SD	0.2439	0.2461	0.2677	0.2360	0.2456
(Truth:0.1595)	MSE	0.0595	0.0606	0.0722	**0.0558**	0.0603
50	Mean	0.0970	0.0985	0.1059	0.0951	0.0986
	SD	0.1089	0.1076	0.1136	0.1047	0.1075
(Truth: 0.0950)	MSE	0.0119	0.0116	0.0130	**0.0110**	0.0116
100	Mean	0.0486	0.0489	0.0495	0.0486	0.0489
	SD	0.0455	0.0447	0.0447	0.0447	0.0447
(Truth: 0.0473)	MSE	0.0021	0.0020	0.0020	**0.0020**	0.0020

4. Application to Regression Analysis and Model Selection

In regression analysis one often wants to find the most parsimonious model with sufficient goodness of fit. Many existing model selection criteria, such as the AIC (Akaike, 1974) and BIC (Schwarz, 1978) model selection tools, are constructed by estimating the Kullback-Leibler risk of a fitted model with a certain training sample size (Wang and Lindsay, 2014). However, every risk estimate suffers sampling variation. Without evaluating the variance of a risk estimator, one cannot know for sure whether the model with the smallest risk score is truly the optimal one or not.

The proposed extrapolated variance estimators have great practical value in regression analysis and model selection. It was discussed in Hindley (1977), Wu (1986), and also noted in Wang and Chen (2015) that the jackknife variance estimator yields large positive bias in unbalanced regression situations. When used to assess the variation of a risk estimate for a fitted model, the large positive bias in the jackknife variance estimator is likely to impact inferential decisions. For instance, under the widely used "one-standard-error" rule (Hastie et al., 2009), where one selects the most parsimonious model whose risk is within one standard error of the optimal risk score, using the jackknife variance estimator may result in choosing an over-parsimonious model.

We consider the same simulation scenario as discussed in Wang and Chen (2015). A multiple linear regression model is considered whose true relationship is defined as follows:

$$Y_i = 1 + 8X_{i,1} + 5X_{i,2} + 3X_{i,3} + 1X_{i,4} + 0.1X_{i,5} + \epsilon_i \ (1 \le i \le 100).$$

We simulate $R = 500$ data sets of size $n = 100$. For each data set the x-variables are independently generated from Uniform(0,1), and the random errors are simulated from Normal distribution with mean 0 and standard deviation 0.1. The parameter of interest is the true Kullback-Leibler risk of the Least Square fit of the above linear regression model. The unbiased risk estimate based on Kullback-Leibler distance is of a U-statistic form. For instance, if the kernel size of the U risk estimator is $k = n/2$, then the cross-validation score can be written as

$$\begin{aligned} U_n &= \binom{n}{n/2}^{-1} \sum_{(n,n/2)} \phi(S_{n/2}), \\ \phi(S_{n/2}) &= \frac{1}{n/2} \sum_{(\mathbf{x}_i, y_i) \in S_{n/2}} \log f_{\hat{\beta}(S^{-i}_{n/2-1})}(y_i \mid \mathbf{x}_i), \end{aligned}$$

where the notation $(n, n/2)$ means the summation is taken over all subsamples $S_{n/2}$ of size $n/2$. Here $S^{-i}_{n/2-1}$ is a delete-i data subset of size $n/2 - 1$, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i_5})^T$, $\beta = (\beta_0, \ldots, \beta_5)^T$, and $f_{\hat{\beta}}$ is the estimated density function for the response Y.

We consider estimating the variance of the Kullback-Leibler risk score using different variance estimators, including the linearly extrapolated variance estimator $\hat{V}_{\text{ex1}}$ with $m = n/2$, the second-order extrapolated variance estimator $\hat{V}_{\text{ex2}}$ with $m = n/2$ and $n/4$, and the jackknife variance estimator. The true variance of the risk score is approximated based on 10,000 random data sets. The following table summarizes the average variance estimate, the standard deviation , and the mean squared error of each variance estimator.

Table 5: Comparison of different variance estimators in risk analysis

Estimator	Mean	SD	MSE
Truth	0.00552		
$\hat{V}_{\text{ex1}}$	0.00595	0.00229	5.43×10^{-6}
$\hat{V}_{\text{ex2}}$	0.00511	0.00217	4.92×10^{-6}
$\hat{V}_J$	0.00895	0.00331	2.27×10^{-5}

As noticed in Table 5, the jackknife variance estimator shows non-negligible positive bias that is over 60% of the true variance in the example under consideration. In addition, the mean squared error of the jackknife estimator is about four times larger than that of the extrapolated variance estimators. In comparison, both extrapolation methods yield variance

estimates that are close to the truth on average and with smaller standard deviation and mean squared error. The second-order extrapolation seems to provide improvement over the linear extrapolation technique. The extrapolated variance estimators clearly outperform the conventional jackknife method in the regression context.

5. Discussion

In this paper we consider using extrapolation techniques in U-statistic variance estimation. In particular, a second-order extrapolated variance estimator is proposed. The second-order extrapolation technique corrects the positive bias in the linearly extrapolated variance estimator and leads to a variance estimator that is nearly second-order unbiased.

The construction of the extrapolated variance estimators, as well as the unbiased variance estimator $\hat{V}_u$, requires the kernel size $k \leq n/2$ (or $k \leq n/4$ for second-order extrapolation). This condition limits the applications of these variance estimators in K-fold cross-validation problems ($K \geq 2$). Wang and Lindsay (2014) show that the kernel size k of an unbiased risk estimator in K-fold cross-validation is $1 + n(K-1)/K$, bigger than $n/2$. Bengio and Grandvalet (2004) point out that there is no unbiased variance estimator for K-fold cross-validation. That is, neither the unbiased variance estimator nor the extrapolated variance estimators apply to the commonly used ten-fold or leave-one-out cross-validation scenarios. In a future project we will study how to find a general variance estimator for a U-statistic in the context of K-fold cross-validation.

References

[1] H. Akaike. A new look at the statistical identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[3] B. Efron and C. Stein. The jackknife estimation of variance. *The Annals of Statistics*, 9(3):586–596, 1981.

[4] R.E. Folsom. *Probability sample U-statistics: theory and applications for complex sample designs*. PhD thesis, University of North Carolina, Chapel Hill, 1984.

[5] D.A.S. Fraser. Completeness of order statistics. *Canadian Journal of Mathematics*, 6:42–45, 1954.

[6] P. Hall and A.P. Robinson. Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika*, 96(1):175–186, 2009.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[8] D.V. Hinkley. Jackknifing in unbalanced situations. *Technometrics*, 19:285–292, 1977.

[9] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

[10] A.J. Lee. *U-statistics: theory and practice*. New York : M. Dekker, 1990.

[11] Y. Maesono. Asymptotic comparisons of several variance estimators and their effects for studentizations. *Annals of the Institute of Statistical Mathematics*, 50(3):451–470, 1998.

[12] J.S. Marron. Partitioned cross-validation. *Econometric Reviews*, 6:271–283, 1987.

[13] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[14] Q. Wang and S. Chen. A general class of linearly extrapolated variance estimators. *Statistics&Probability Letters*, 98:29–38, 2015.

[15] Q. Wang and B.G. Lindsay. Variance estimation of a general u-statistic with application to cross-validation. *Statistica Sinica*, 24(3):1117–1141, 2014.

[16] Q. Wang and B.G. Lindsay. Improving cross-validated bandwidth selector using subsampling-extraplation techniques. *Computational Statistics and Data Analysis*, 89:51–71, 2015.

[17] C.F.J Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14:1261–1295, 1986.

Appendix A: Proof for Theorem 1

Proof. Because $\hat{V}_m$ $(m \leq n/2)$ is unbiased for $\text{Var}(U_m)$ and has finite variance, by Chebyshev's inequality $\hat{V}_m$ converges to $\text{Var}(U_m)$ in probability. That is,

$$\hat{V}_{n/2} \xrightarrow{P} \text{Var}(U_{n/2}) \text{ and } \hat{V}_{n/4} \xrightarrow{P} \text{Var}(U_{n/4}).$$

Because both $\text{Var}(U_{n/2})$ and $\text{Var}(U_{n/4})$ are non-zero constants, by Slutsky's Theorem we have

$$2\hat{V}_{n/2}/\hat{V}_{n/4} \xrightarrow{P} 2\text{Var}(U_{n/2})/\text{Var}(U_{n/4}).$$

Therefore,

$$\sqrt{2\hat{V}_{n/2}/\hat{V}_{n/4}} \xrightarrow{P} \sqrt{2\text{Var}(U_{n/2})/\text{Var}(U_{n/4})}.$$

We write $\sqrt{2\hat{V}_{n/2}/\hat{V}_{n/4}} = \sqrt{2\text{Var}(U_{n/2})/\text{Var}(U_{n/4})} + o_P(1)$, and thus

$$\hat{V}_{\text{ex2}} = \hat{V}_{\text{ex1}} \left(\sqrt{2\text{Var}(U_{n/2})/\text{Var}(U_{n/4})} + o_P(1) \right).$$

By the closed-form expression of the U-statistic variance in equation (2.4), we have

$$\begin{aligned}
\frac{2\text{Var}(U_{n/2})}{\text{Var}(U_{n/4})} &= \frac{2\sum_{j=1}^{k}\binom{k}{j}^2\binom{n/2}{k}^{-1}\delta_j^2}{\sum_{j=1}^{k}\binom{k}{j}^2\binom{n/4}{k}^{-1}\delta_j^2} \\
&= \frac{\frac{4k^2}{n}\delta_1^2+\frac{4k^2(k-1)^2}{n(n-2)}\delta_2^2+o(1/n^2)}{\frac{4k^2}{n}\delta_1^2+\frac{8k^2(k-1)^2}{n(n-4)}+o(1/n^2)} \\
&= \frac{1+\frac{(k-1)^2}{n-2}\frac{\delta_2^2}{\delta_1^2}+o(1/n)}{1+\frac{2(k-1)^2}{n-4}\frac{\delta_2^2}{\delta_1^2}+o(1/n)} \\
&= \frac{1+\frac{(k-1)^2}{n-2}\frac{\delta_2^2}{\delta_1^2}+o(1/n)}{1+2(k-1)^2\left(\frac{1}{n-2}+\frac{2}{(n-2)(n-4)}\right)\frac{\delta_2^2}{\delta_1^2}+o(1/n)} \\
&= \frac{1+\frac{(k-1)^2}{n-2}\frac{\delta_2^2}{\delta_1^2}+o(1/n)}{1+\frac{2(k-1)^2}{n-2}\frac{\delta_2^2}{\delta_1^2}+o(1/n)}
\end{aligned}$$

Let $s=(k-1)^2\delta_2^2/\{(n-2)\delta_1^2\}$. The expectation of $\hat{V}_{\text{ex2}}$ can be expressed as

$$\begin{aligned}
E(\hat{V}_{\text{ex2}}) &= \left\{\frac{k^2}{n}\delta_1^2+\frac{k^2(k-1)^2}{n(n-2)}\delta_2^2+o(1/n^2)\right\}\left\{\sqrt{\frac{1+s+o(1/n)}{1+2s+o(1/n)}}+o_P(1)\right\} \\
&= \left\{\frac{k^2}{n}\delta_1^2+\frac{k^2(k-1)^2}{n(n-2)}\delta_2^2+o(1/n^2)\right\}\left\{\sqrt{1-\frac{s}{1+2s}+o(1/n)}+o_P(1)\right\}
\end{aligned}$$

Denote $t=s/(1+2s)+o(1/n)$, and consider $\sqrt{1-t}$. Because $\mid t \mid < 1$, we can apply Taylor series and expand $\sqrt{1-t}$ around $t=0$ as follows.

$$\sqrt{1-t}=1-\frac{t}{2}+\text{Remainder}.$$

Notice that t is of order $1/n$. Thus, the remainder in the expansion is $o(1/n)$, and

$$\begin{aligned}
E(\hat{V}_{\text{ex2}}) &= \left\{\frac{k^2}{n}\delta_1^2+\frac{k^2(k-1)^2}{n(n-2)}\delta_2^2+o(1/n^2)\right\}\left\{1-\frac{(k-1)^2\delta_2^2}{2(n-2)\delta_1^2}+o(1/n)+o_P(1)\right\} \\
&= \frac{k^2}{n}\delta_1^2+\left\{\frac{k^2(k-1)^2}{n(n-2)}-\frac{k^2(k-1)^2}{2n(n-2)}\right\}\delta_2^2+o(1/n^2)+(1/n)o_P(1)
\end{aligned}$$

Thus,

$$n\left\{E(\hat{V}_{\text{ex2}})-\left(\frac{k^2}{n}\delta_1^2+\frac{k^2(k-1)^2}{2n(n-2)}\delta_2^2+o(1/n^2)\right)\right\}\xrightarrow{P}0.$$

Stochastically, the second-order bias in $\hat{V}_{\text{ex2}}$ is

$$\binom{k}{2}^2\left\{\frac{2}{n(n-2)}-\frac{2}{n(n-1)}\right\}\delta_2^2\approx 0.$$

In comparison, the second-order positive bias in $\hat{V}_{\text{ex1}}$ is

$$\binom{k}{2}^2\left\{\frac{4}{n(n-2)}-\frac{2}{n(n-1)}\right\}\delta_2^2.$$

In addition, from Theorem 2 in Wang and Chen (2015) the jackknife variance estimator has second-order bias

$$\binom{k}{2}^2 \left\{ \frac{4n(n-2)}{(n-1)^4} - \frac{2}{n(n-1)} \right\} \delta_2^2.$$

Hence, the second-order extrapolation technique successfully corrects the positive bias in $\hat{V}_{\text{ex1}}$ and $\hat{V}_J$. □