Statistica Sinica **24** (2014), 1117-1141
doi:http://dx.doi.org/10.5705/ss.2012.215

VARIANCE ESTIMATION OF A GENERAL U-STATISTIC WITH APPLICATION TO CROSS-VALIDATION

Qing Wang and Bruce Lindsay

Williams College and The Pennsylvania State University

Abstract: This paper addresses the problem of variance estimation for a general U-statistic. U-statistics form a class of unbiased estimators for those parameters of interest that can be written as $E\{\phi(X_1,\ldots,X_k)\}$, where ϕ is a symmetric kernel function with k arguments. Although estimating the variance of a U-statistic is clearly of interest, asymptotic results for a general U-statistic are not necessarily reliable when the kernel size k is not negligible compared with the sample size n. Such situations arise in cross-validation and other nonparametric risk estimation problems. On the other hand, the exact closed form variance is complicated in form, especially when both k and n are large. We have devised an unbiased variance estimator for a general U-statistic. It can be written as a quadratic form of the kernel function ϕ and is applicable as long as $k \leq n/2$. In addition, it can be represented in a familiar analysis of variance form as a contrast of between-class and within-class variation. As a further step to make the proposed variance estimator more practical, we developed a partition resampling scheme that can be used to realize the U-statistic and its variance estimator simultaneously with high computational efficiency. A data example in the context of model selection is provided. To study our estimator, we construct a U-statistic cross-validation tool, akin to the BIC criterion for model selection. With our variance estimator we can test which model has the smallest risk.

Key words and phrases: Best unbiased estimator, cross-validation, likelihood risk, model selection, partition resampling, U-statistic, variance.

1. Introduction

Suppose we are considering a parameter of interest θ that can be written as a functional of F with the form $\theta(F) = \int\cdots\int \phi(x_1,\ldots,x_k)dF(x_1)\cdots dF(x_k)$, where F may be either univariate or multivariate, and ϕ is a symmetric function with k arguments. Here "symmetry" means that the value of ϕ does not change by rearranging its k components. Given a random sample of size n, $\mathcal{X}_n = (X_1,\ldots,X_n)$, statistic $\phi(X_1,\ldots,X_k)$ can be used as an unbiased estimator for the parameter θ if $n \geq k$. Hoeffding (1948) introduced a class of statistics, called U-statistics, which average over all kernel functions with subsamples of size k taken out of $\mathcal{X}_n$. It is generally defined in the form

$$U_n = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} \phi(X_{i_1}, \ldots, X_{i_k}). \tag{1.1}$$

Although the function ϕ is often scalar-valued in applications, most of the theory can easily be transferred to the vector-valued case.

The main achievement of this paper is the identification and practical development of an unbiased estimator of the variance of a U-statistic. This estimator exists provided $k \le n/2$. We also develop fast subsampling method that leads simultaneously to an efficient incomplete U-statistic along with its variance estimate. We think that this estimator and its spin-offs will be valuable in many problems where the size of the U-statistic kernel is large.

We use our variance estimator to make new contributions to the model selection literature. Standard cross-validation tools often lead to U-statistics with large kernel sizes. Our concern with these methods is that many rely almost entirely on point estimates of risk, even though these cross-validation risk estimates can be highly variable. It seems to us that a conservative approach would be to avoid selecting larger models when we are unsure whether they actually improve risk. To do this, we need to measure the variability in the differences between model risk estimates. We do so here by estimating their variance of the differences between paired risk estimates and constructing t-type statistics. We show our assessment is quite different from the "1-SE rule" (Breiman et al. (1984)).

In these model selection settings standard nonparametric variance estimation tools can be highly biased and computationally expensive. We think the unbiased estimator, which we show is faster to compute, could play an important role. We demonstrate its use here in two contexts. In one problem, a nonparametric estimation of model power, we compare by simulation our method with a variety of other nonparametric variance estimators. This is done in Section 5. In a second problem, in Section 6, we devise a likelihood cross-validation (LCV) tool that is akin to a BIC estimator of risk (Schwarz (1978)), and apply it to the selection of a logistic model. In this example we can show the close relationship of the LCV and BIC estimates, and use our variance estimation to develop a conservative method of minimizing risk.

2. Unbiased Variance Estimation

In this section we present some background on U-statistics, then introduce the unbiased variance estimator $\hat{V}_u$ (2.3). We consider the estimation of variance for incomplete U-statistics, as well as a partition representation of $\hat{V}_u$ that leads to a new subsampling scheme.

2.1. Standard U-statistic results

The statistic U_n found in (1.1) is clearly an unbiased estimator of parameter θ. In addition, it can be written as a function of the order statistics. As the set of order statistics is the complete sufficient statistic if the underlying distribution family is large enough (Fraser (1954)), U_n is the best unbiased estimator of θ in the context of nonparametric inference.

Theoretical results concerning a general U-statistic include its asymptotic normality under certain regularity conditions and its exact closed form variance. These results first appeared in Hoeffding (1948). The asymptotic U-statistic variance, $k^2\sigma_1^2/n$, is constructed according to Theorem 5.2 in Hoeffding (1948):

$$\lim_{n\to\infty} n\mathrm{Var}\,(U_n) = k^2\sigma_1^2,$$

where $\sigma_1^2 = \mathrm{Var}\,[E\{\phi(X_1,\ldots,X_n) \mid X_1\}]$. However, the asymptotic results are based on the assumption that the sample size n goes to infinity with the kernel ϕ, and hence k, fixed. Equivalently, the kernel size k is negligible compared with n. In addition, Theorem 5.2 in Hoeffding (1948) reveals that by using the asymptotic variance to estimate the U-statistic variance, we are always optimistic. The exact form of the variance for a general U-statistic

$$\mathrm{Var}\,(U_n) = \binom{n}{k}^{-1}\sum_{c=1}^{k}\binom{k}{c}\binom{n-k}{k-c}\sigma_c^2, \tag{2.1}$$

where $\phi_c(x_1,\ldots,x_c) = E\{\phi(X_1,\ldots,X_n) \mid X_1 = x_1,\ldots,X_c = x_c\}$ and $\sigma_c^2 = \mathrm{Var}\,(\phi_c)$ for $1 \le c \le k$, is complicated in form and appears computationally intensive, especially when both k and n are large. As a result, our first goal is to derive a general estimator for the variance of a U-statistic that is of a relatively simple form. Its realization does not depend on the computation of the conditional variances σ_c^2's. In addition, the proposed variance estimator exists as long as the kernel size $k \le n/2$.

2.2. The construction of the unbiased variance estimator $\hat{V}_u$

We demonstrate how one can construct an unbiased estimator of the variance of an arbitrary U-statistic, assuming that $k \le n/2$. Consider the complete U-statistic denoted as $U_n = \mathbb{N}^{-1}\sum_{i=1}^{\mathbb{N}}\phi(S_i)$, where $\mathbb{N}$ is the number of distinct size-k samples, say S_i, taken out of $X_1,\ldots,X_n$. Define the sample overlap $O(S_1,S_2)$ as the number of elements in common between S_i and S_j. Take $P_c = \{(S_i,S_j) \mid O(S_i,S_j) \le c\}$, and let N_c be the number of pairs in P_c. Define

$$Q(c) = N_c^{-1}\sum_{P_c}\phi(S_i)\phi(S_j) \quad (0 \le c \le k). \tag{2.2}$$

Theorem 1. *If U_n is a U-statistic with a kernel ϕ of size k, $k \le n/2$, and*

$$\hat{V}_u = Q(k) - Q(0), \tag{2.3}$$

where $Q(k)$ and $Q(0)$ are defined in (2.2)*, then $\hat{V}_u$ is an unbiased estimator of* $\mathrm{Var}\,(U_n)$*. Furthermore, it is a function of the order statistics and so is the best unbiased estimator of* $\mathrm{Var}\,(U_n)$*.*

Proof. With a little algebra one can show that $E\{Q(0)\} = \{E(U_n)\}^2$, and $Q(k) = U_n^2$. Therefore, $E\{Q(k) - Q(0)\} = E(U_n^2) - \{E(U_n)\}^2 = \mathrm{Var}\,(U_n)$. That is, $Q(k) - Q(0)$ is an unbiased estimate of $\mathrm{Var}\,(U_n)$. This yields the result in Theorem 1.

The theory can be extended to the estimation of vector-valued $\phi(S)$ by defining matrix-valued versions of $\mathbf{Q}(c) = \sum_{P_c} \phi(S_i)\phi(S_j)^T/N_c \quad (0 \le c \le k)$.

Example 1. Consider an independent and identically distributed sample $X_1, \ldots, X_n$ from some distribution with mean μ and finite variance σ^2. Take the parameter of interest to be $\theta = \mu$. Let $\phi(x) = x$ be the kernel function, which results in the U-statistic $U_n = \sum_{i=1}^n X_i/n = \bar{X}$. Based on (2.3), we have

$$Q(k) = U_n^2 = \bar{X}^2, \quad Q(0) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} X_i X_j.$$

It can be shown that $\hat{V}_u = Q(k) - Q(0) = \sum_{i=1}^n \left(X_i - \bar{X}\right)^2 / \{n(n-1)\}$, and $E(\hat{V}_u) = \sigma^2/n = \mathrm{Var}\,(U_n)$.

Remark 1. To our knowledge, (2.3) is a new method for variance estimation. Folsom has this result in a more complex form (Folsom (1984, p.79)), with the estimator expressed as

$$\left\{2\binom{n}{k}\binom{n-k}{k}\right\}^{-1} \sum_{S_a \cap S_b = \emptyset} \{\phi(S_a) - \phi(S_b)\}^2 - \binom{n}{k}^{-1} \sum_a \{\phi(S_a) - U_n\}^2.$$

His result focuses on probability sample U-statistics in survey statistics, and further results seem to be absent.

Remark 2. There is some related work in the area of variance estimation in cross-validation. Nadeau and Bengio (2003) considered training samples of $n/2$ or less, but their estimator was unbiased for the U-statistic with sample size $n/2$, not n, and so it showed considerable positive bias. Markatou et al. (2005) considered larger training samples, and corrected for bias by using moment approximations based on Taylor series.

Remark 3. As the proposed variance estimator $\hat{V}_u$ is a function of the order statistics, it can be represented as a complete U-statistic with a new kernel function ψ of size $2k$ (see Appendix for proof). This kernel depends on n, as is reflective of the fact that the variance of U_n depends on n. Hence, even if k is fixed, further work is needed to establish the asymptotic properties of the estimator.

Remark 4. Although we do not pursue it further here, it can be shown that the unbiased estimator for σ_1^2 has the form $(N_1/n_1)\{Q(1) - Q(0)\}$, where n_1 is the number of pairs of size-k subsamples that have exactly one common element. From this it follows that the unbiased estimator for σ_1^2 also has the form of a complete U-statistic with kernel size $2k$, making it no easier to estimate unbiasedly than $\mathrm{Var}\,(U_n)$.

2.3. Estimation of variance under subsampling

To reduce the expense in computing a complete U-statistic, Blom (1976) proposed to consider an incomplete U-statistic. It is generally defined as $U^{inc} = C^{-1}\sum \phi(S_i)$, where S_i is a size-k subsample from $\mathcal{X}_n$ $(1 \le i \le C, C \in \mathcal{N}^+)$. A special version is a realization by random subsampling (Politis, Romano, and Wolf (1999)). We denote the sampled incomplete U-statistic with C random subsamples as $\tilde{U}_C$. It can be written as $\tilde{U}_C = C^{-1}\sum_{i=1}^{C}\phi(\tilde{S}_i)$, where each $\tilde{S}_i$ is a random subsample of size k taken out of $\mathcal{X}_n$. We show here how one can estimate the variance of $\tilde{U}_C$ unbiasedly.

The variance of $\tilde{U}_C$ can be decomposed as

$$\mathrm{Var}\,(\tilde{U}_C) = \mathrm{Var}\left[E\Big\{C^{-1}\sum_{i=1}^{C}\phi(\tilde{S}_i)\Big|\mathcal{X}_n\Big\}\right] + E\left[\mathrm{Var}\left\{C^{-1}\sum_{i=1}^{C}\phi(\tilde{S}_i)\Big|\mathcal{X}_n\right\}\right].$$

Because $\phi(\tilde{S}_1), \ldots, \phi(\tilde{S}_C)$ are sampled independently given the data $\mathcal{X}_n$, we then have $\mathrm{Var}\,(\tilde{U}_C) = \mathrm{Var}\,(U_n) + E[\{C(C-1)\}^{-1}\sum_i\{\phi(\tilde{S}_i) - \tilde{U}_C\}^2]$.

Denote the unbiased estimator for $\mathrm{Var}\,(\tilde{U}_C)$ as

$$\hat{V}_u^{inc} = \hat{V}_u + \{C(C-1)\}^{-1}\sum_{i=1}^{C}\left(\phi(\tilde{S}_i) - \tilde{U}_C\right)^2.$$

As $\hat{V}_u$ is the best unbiased estimate for $\mathrm{Var}\,(U_n)$, $\hat{V}_u^{inc}$ is the best unbiased estimate for $\mathrm{Var}\,(\tilde{U}_C)$. The second term on the right represents the extra variance due to using an incomplete statistic. This is a new formula for variance estimation in the subsampling case. However, we see later that approximation of $\hat{V}_u$ is challenging under simple subsampling.

2.4. The analysis of variance representation of $\hat{V}_u$

There is an alternative representation of $\hat{V}_u$ that is quite useful for constructing a better subsampling scheme for incomplete U-statistics. Equations (2.2) and (2.3) show that the proposed unbiased variance estimator $\hat{V}_u$ can be written as a quadratic form involving the kernel function $\phi(S)$. This leads us to write the statistic using matrix notation to better understand its structure. We show here how it can be represented in a familiar analysis of variance form as a contrast of between-class and within-class variation.

For any given kernel size k, we partition the sample space into a maximal number of subsamples of size k, say $S_1, \ldots, S_m$, where $mk \leq n$. The resulting $\phi(S_1), \ldots, \phi(S_m)$ are independent random variables. To simplify our notation, assume $n = mk$. We let $\mathbb{B}$ be the number of different ways one can partition the data set. For $a = 1, \ldots, \mathbb{B}$, let the ath partition be a unique sequence of nonoverlapping size-k samples $S_{a,1}, \ldots, S_{a,m}$. We represent the variance estimator based on partitions and claim that it is equal to the best unbiased estimator.

Define the complete variance estimator based on partitions to be

$$\hat{V}_{\text{partition}} = \frac{1}{\mathbb{B}} \sum_{a=1}^{\mathbb{B}} \left[\frac{1}{m} \sum_{j=1}^{m} \left\{ \frac{1}{m-1} \left(\phi(S_{a,j}) - \bar{\phi}_a \right)^2 - \left(\bar{\phi}_a - \bar{\phi} \right)^2 \right\} \right], \tag{2.4}$$

where $\bar{\phi}_a = \sum_{j=1}^{m} \phi(S_{a,j})/m$, and $\bar{\phi} = \sum_{a=1}^{\mathbb{B}} \bar{\phi}_a / \mathbb{B} = U_n$.

Proposition 1. *The complete variance estimator based on partitions is equal to the unbiased variance estimator, $\hat{V}_{partition} = \hat{V}_u$.*

Note that $\hat{V}_{\text{partition}}$ can be re-expressed as

$$\hat{V}_{\text{partition}} = \frac{1}{m} \sigma^2_{\text{WP}} - \sigma^2_{\text{BP}}, \tag{2.5}$$

where $\sigma^2_{\text{WP}} = \sum_{a=1}^{\mathbb{B}} \sum_{j=1}^{m} \left(\phi(S_{a,j}) - \bar{\phi}_a \right)^2 / \{\mathbb{B}(m-1)\}$ is the within-partition variance, and $\sigma^2_{\text{BP}} = \sum_{a=1}^{\mathbb{B}} \left(\bar{\phi}_a - \bar{\phi} \right)^2 / \mathbb{B}$ is the between-partition variance.

One interesting question here is how the formula in (2.4) relates to variance estimation when one uses an incomplete U-statistic based on a single partition, say $S_1, \ldots, S_m$. As $\phi(S_1), \ldots, \phi(S_m)$ are independent, the true variance for the corresponding incomplete U-statistic U^{inc} is Var $\{\phi(S)\}/m$. The best unbiased estimator of Var (U^{inc}) would be $\sum_{j=1}^{m} \left(\phi(S_j) - \bar{\phi} \right)^2 / \{m(m-1)\}$. This same term is estimated in $\hat{V}_{\text{partition}}$ by σ^2_{WP}/m. The term $1 - (m\sigma^2_{\text{BP}}/\sigma^2_{\text{WP}})$ therefore represents the relative decrease in variance arising from using the complete U-statistic instead of a single partition.

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

In this section we show $\hat{V}_u$ has a weakness that has a simple repair. We start by showing that it is numerically possible for $\hat{V}_u = Q(k) - Q(0)$ to be negative when $k \geq 2$.

Example 2. Consider $k = 2$ and $n = 4$. Denote the original data set as $x_1, \ldots, x_4$, and let ϕ be a kernel function of order 2. Suppose that $\phi(x_1, x_2) = \phi(x_3, x_4) = 1$, and $\phi(x_1, x_3) = \phi(x_1, x_4) = \phi(x_2, x_3) = \phi(x_2, x_4) = 0$. Then,

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

The estimator corresponds to a quadratic form. But, the matrix has negative eigenvalues, as is clear from (2.4). Negative estimates of $\mathrm{Var}\,(U_n)$ by $\hat{V}_u$ are clearly undesirable. And, we propose an easy-to-compute adjustment based on two lemmas.

Lemma 1. *$E\{Q(k) - Q(c)\} \leq E\{Q(k) - Q(c-1)\} \leq \mathrm{Var}\,(U_n)$ for all $1 \leq c \leq k$.*

Lemma 2. *If $S_U^2 = Q(k) - Q(k-1)$,*

$$S_U^2 = \frac{1}{\mathbb{N}(\mathbb{N}-1)} \sum_{i=1}^{\mathbb{N}} \{\phi(S_i) - U_n\}^2, \quad \mathbb{N} = \binom{n}{k}.$$

Thus S_U^2, viewed as an estimator of $\mathrm{Var}\,(U_n)$, is both biased downwards from Lemma 1, and nonnegative from Lemma 2. Moreover, it is strictly positive unless $\phi(S_i) = U_n$ for any size-k sample S_i. Since $\hat{V}_u - S_U^2$ estimates a positive quantity but can take negative values, there is a simple fix to the unbiased estimator $\hat{V}_u$, denoted as $\hat{V}_u^+$,

$$\hat{V}_u^+ = \max\{\hat{V}_u, S_U^2\}. \tag{3.1}$$

As S_U^2 is consistent for small values, using $\hat{V}_u^+$ guarantees that the adjusted variance estimator is no smaller than S_U^2. In addition, this proposed fix does not have the discontinuity (as a function of the data) present in another intuitive adjustment, $\tilde{V} = \hat{V}_u \mathbb{I}\{\hat{V}_u > 0\} + S_U^2 \mathbb{I}\{\hat{V}_u \leq 0\}$. In our simulation study it will be seen that the proposed adjustment has similar performance as $\hat{V}_u$ as to means, standard deviations, and mean squared errors.

Example 3. Continuing with Example 2, we find that

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

It is clear that $\hat{V}_u^+$ has a positive bias. We will examine the differences between $\hat{V}_u^+$ and $\hat{V}_u$ in our numerical study of Section 5.3.

4. Partition and Other Resampling Schemes

For problems with large n and k, the number of possible subsamples of size k is enormous, and so it is challenging to compute U_n. It is even more challenging to compute $\hat{V}_u$, which requires summation over all the pairs of nonoverlapped subsamples. We introduce a partition resampling scheme that can be used to realize U_n and its variance estimator $\hat{V}_u$ (2.3) simultaneously with high statistical and computational efficiency.

4.1. Partition resampling scheme

A single partition $(S_1, \ldots, S_m)$ of the sample creates the most efficient incomplete U-statistic possible using just m subsamples. Thus it seems a natural building block for creating a sampling scheme to compute the incomplete U-statistics.

For $b = 1, \ldots, B$, let $P_b = (\tilde{S}_{b,1}, \ldots, \tilde{S}_{b,m})$ be the bth random partition of the size-n sample into m disjoint subsets of size k. Without loss of generality, assume that $n = mk$. We propose to sample with replacement B times from the set of all partitions. We call this the *partition resampling scheme*. Thus for kernel size 2, we first sample a partition consisting $m = n/2$ subsets of size 2; each such partition generates a term $\bar{\phi}_b = \sum_{j=1}^m \phi(\tilde{S}_{b,j})/m$. We then construct B such partitions and average over partitions to get an incomplete U-statistic.

One can estimate $\mathrm{Var}\,(U_n)$ by random partition resampling based on a formula analogous to (2.5). Let $W(P_b) = \sum_{j=1}^m (\phi(\tilde{S}_{b,j}) - \bar{\phi}_b)^2/(m-1)$ be the sums of squares within the bth partition, $b = 1, \ldots, B$. We define the random partition variance estimator as

$$\hat{V}^{inc}_{\mathrm{partition},B} = \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{W(P_b)}{m} - (\bar{\phi}_b - \bar{\phi}^{inc})^2\right\} := \frac{1}{B}\sum_{b=1}^{B}\gamma_b \tag{4.1}$$

where $\bar{\phi}^{inc} = \sum_{b=1}^B \bar{\phi}_b/B$. It can also be represented as a contrast of between-partition and within-partition variation as in (2.5). $\hat{V}^{inc}_{\mathrm{partition},B} = \tilde{\sigma}^2_{\mathrm{WP}}/m - \tilde{\sigma}^2_{\mathrm{BP}}$, where $\tilde{\sigma}^2_{\mathrm{WP}} = \sum_{b=1}^B \sum_{j=1}^m \left\{\phi(\tilde{S}_{b,j}) - \bar{\phi}_b\right\}^2 / \{B(m-1)\}$ is the sampled within-partition variance, and $\tilde{\sigma}^2_{\mathrm{BP}} = \sum_{b=1}^B \left(\bar{\phi}_b - \bar{\phi}^{inc}\right)^2/B$ is the sampled between-partition variance. It can be shown that $\hat{V}^{inc}_{\mathrm{partition},B}$ is subsampling unbiased for $\mathrm{Var}\,(U^{inc})$.

Moreover,

$$\hat{V}^{inc}_{\text{partition},B} = (\bar{\phi}^{inc})^2 - B^{-1}\sum_{b=1}^{B}\tilde{Q}_b(0), \quad \tilde{Q}_b(0) = \binom{m}{2}^{-1}\sum_{1\le i<j\le m}\phi(\tilde{S}_{b,i})\phi(\tilde{S}_{b,j}).$$

That is, the random partition resampling only considers the nonoverlapped pairs within each partition in the estimation of $Q(0)$.

4.2. Determining the partition size B

The selection of the number of partitions B is important for the properties of the estimated variance. The good news is that one can estimate adequacy of B from the subsampling data. The first property we consider is the efficiency of the incomplete estimator relative to the complete estimator. The variance of the partition variance estimator (4.1) can be written as $\text{Var}\,(\hat{V}^{inc}_{\text{partition},B}) = \text{Var}\,\{E(\hat{V}^{inc}_{\text{partition},B} \mid \mathcal{X}_n)\} + E\{\text{Var}\,(\hat{V}^{inc}_{\text{partition},B} \mid \mathcal{X}_n)\}$, so

$$\text{Var}\,(\hat{V}^{inc}_{\text{partition},B}) = \text{Var}\,(\hat{V}_u) + \text{mean partition resampling variance}. \tag{4.2}$$

In practice, we want the mean partition resampling variance small so as to have a more accurate estimator.

From the perspective of subsampling partitions P_b, the statistic $\hat{V}^{inc}_{\text{partition},B}$ can be written as a U-statistic of order 2:

$$\hat{V}^{inc}_{\text{partition},B} = \frac{1}{\binom{B}{2}}\sum_{1\le i<j\le B}\Big\{\frac{W(P_{b_i}) + W(P_{b_j})}{2m} - \frac{B-1}{B}\frac{(\bar{\phi}_{b_i} - \bar{\phi}_{b_j})^2}{2}\Big\}.$$

We can estimate its variance unbiasedly using the methods in this paper. But, as a simpler method, (4.1) has $\hat{V}^{inc}_{\text{partition},B}$ close to being an average over partitions that are generated randomly. Thus, we can also evaluate the accuracy of the incomplete realization by examining its standard error based on $SD(\gamma_b)/\sqrt{B}$, and thereby ensure that the deviation from $\hat{V}_u$ is minimal.

In a simulation study one can estimate $\text{Var}\,(\hat{V}^{inc}_{\text{partition},B})$ and the mean partition subsamping variance, and so use (4.2) to estimate the complete estimator variance, corresponding to $B = \infty$. In a data analysis, it is useful to interpret the simulation error in terms of how much our inference could change if B were infinite. Suppose that when we estimate $\hat{V}_u$ with $\hat{V}^{inc}_{\text{partition},B}$, there is an estimated resampling variance, say τ_B^2, due to B not being infinite. We could then infer that $\hat{V}_u$ would be very likely to be in the range $\hat{V}^{inc}_{\text{partition},B} \pm 2\tau_B$ if B were to increase without limit. Therefore, if we consider the change in the t-statistic that would be due to letting B become infinite, we get the ratio

$$\frac{U_n/\sqrt{\hat{V}^{inc}_{\text{partition},B}}}{U_n/\sqrt{\hat{V}_u}} = \sqrt{\frac{\hat{V}_u}{\hat{V}^{inc}_{\text{partition},B}}}$$

which is highly likely to lie in the range $\sqrt{1 \pm 2\tau_B / \hat{V}^{inc}_{\text{partition},B}}$. In our data example, we will report these values as a measure of the adequacy of our subsampling efforts.

4.3. Estimating $\text{Var}\,(U_n)$ with simple random sampling

The incomplete U-statistic based on partition resampling is statistically efficient compared with the randomly sampled incomplete U-statistic when both use the same number of size-k subsamples in their constructions. Thus, if we let U^{inc}_{simp} and U^{inc}_{part} be the incomplete U-statistics computed from simple random subsampling and partition resampling respectively, $\text{Var}\,(U^{inc}_{\text{part}}) \leq \text{Var}\,(U^{inc}_{\text{simp}})$. For proofs of these results, the 2012 Pennsylvania State University Ph.D. thesis by Q. Wang is available electronically from Penn State University Library.

A further drawback of simple random sampling arises because one needs a good method to estimate $Q(0)$ in $\hat{V}_u$, and the simple random sampling method can fail to generate enough non-overlapped pairs as needed to calculate $Q(0)$ accurately.

5. Studying the U-Estimator

We designed a simulation study to compare various nonparametric methods in estimating the variance of a general U-statistic. We considered a U-statistic with a large kernel size k so that the standard asymptotic U-statistic variance would likely be highly biased and the closed form U variance is difficult to compute. In this context, we compared the proposed unbiased variance estimator with several standard resampling estimators. We present speed comparisons and discuss the bias problems in the other methods. We also studied the possible negative values of the unbiased variance estimator along with the proposed simple fix-up.

5.1. A simulation study

Liu and Lindsay (2009) assessed the quality of the fit of a model to a data set through goodness-of-fit testing used in an inverse fashion. They assumed the alternative was true, and estimated the statistical power of detection of the alternative at sample sizes less than n, constructing a U-statistic for this estimation. Their results showed that this to be a difficult estimation problem.

Here we apply this methodology to the problem of evaluating the estimated power of a one-sample Kolmogorov-Smirnov test for normality. We compare the performance of the unbiased variance estimator $\hat{V}_u$ with some bootstrap and jackknife variance estimators. It is seen that $\hat{V}_u$ provides an estimator that is competitive with its bootstrap and jackknife counterparts on features such as

bias, variance, and computational speed. We later use the same setting to investigate the performance of the nonnegative variance estimator $\hat{V}_u^+$ in comparison with $\hat{V}_u$.

Consider a sample of size n taken independently from a logistic distribution with location parameter 0 and scale parameter 1. We wish to test whether these data are normal. We take subsamples of size k from the data set and test them for normality. The power of a one-sample Kolmogorov-Smirnov test for normality at significance level 0.05 and sample size k is

$$\text{power}_k = P_k(\text{test statistic} \geq \text{threshold}) = P_k(p\text{-value} \leq 0.05).$$

In this case the parameter of interest $\theta_k = E(\mathbb{I}\{p\text{-value} \leq 0.05 \mid \mathcal{X}_k\})$ can be estimated unbiasedly by a U-statistic with kernel function $\phi(\mathcal{X}_k) = \mathbb{I}\{p\text{-value} \leq 0.05 \mid \mathcal{X}_k\}$ of size k, where k is the training sample size. Our proposed variance estimator $\hat{V}_u$ applies as long as $k \leq n/2$.

5.2 Simulation results

The simulation setting was as follows: $R = 500$ samples of size $n = 100$ were drawn independently from a logistic distribution. For each sample of size n, we tested for normality using a one-sample Kolmogorov-Smirnov test based on subsets of size $k = pn$ where $p = 0.1, 0.25$, and 0.5. The logistic density is difficult to distinguish from the normal at this sample size. The mean estimated power at the three sample sizes were 0.158, 0.283, and 0.523, respectively. The columns in Table 1 to Table 3 correspond to the unbiased estimator realized by partition resampling, the naive nonparametric bootstrap estimator, the smooth bootstrap estimator, and the jackknife estimator. For more on bootstrap and jackknife methods, see Efron (1987).

The partition resampling scheme was implemented in estimating the U estimator and its variance, simultaneously, with mB randomly partitioned subsamples of size k. For $mB = 1,000$, this corresponds to $B = 100$ random partitions for $k = 0.1n$, $B = 250$ random partitions for $k = 0.25n$, and $B = 500$ random partitions for $k = 0.5n$. To standardize our comparisons, we used the same number of subsamples of size k when we constructed the bootstrap and jackknife estimates (the results in the following tables were based on 2,000 subsamples for each estimator). We have summarized in Tables 1 to 3 the following: the average of the variance estimates, the standard deviation of the variance estimates, and the ratio of absolute bias over standard deviation. The latter measure reflects the shift in the sampling distribution of the statistic away from the true value, but in a scale free way. In addition, we also report the average computation time *in minutes per estimator*.

Table 1. Comparison with bootstrap and jackknife estimators ($n = 100, k = 0.1n$).

	$\hat{V}_u$ ($mB = \infty$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 1,000$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 2,000$)	Nonpara. Bootstrap	Smooth Bootstrap	Jackknife
Ave$\{\widehat{\text{Var}}\,(U_n)\}$	0.0025	0.0027	0.0026	0.0034	1.7×10^{-5}	0.0067
SD$\{\widehat{\text{Var}}\,(U_n)\}$	0.0021	0.0027	0.0024	0.0018	2.9×10^{-5}	0.0039
$\mid$ Bias $\mid$ /SD	0.000	0.084	0.034	0.513	85.890	1.097
Computation		0.042	0.082	5.682	4.407	0.893

Table 2. Comparison with bootstrap and jackknife estimators ($n = 100, k = 0.25n$).

	$\hat{V}_u$ ($mB = \infty$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 1,000$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 2,000$)	Nonpara. Bootstrap	Smooth Bootstrap	Jackknife
Ave$\{\widehat{\text{Var}}\,(U_n)\}$	0.0162	0.0161	0.0161	0.0200	8.5×10^{-5}	0.0259
SD$\{\widehat{\text{Var}}\,(U_n)\}$	0.0125	0.0131	0.0128	0.0066	1.2×10^{-4}	0.0130
$\mid$ Bias $\mid$ /SD	0.000	0.005	0.008	0.575	133.161	0.748
Computation		0.020	0.037	8.017	5.142	1.816

Table 3. Comparison with bootstrap and jackknife estimators ($n = 100, k = 0.5n$).

	$\hat{V}_u$ ($mB = \infty$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 1,000$)	$\hat{V}^{inc}_{\text{partition},B}$ ($mB = 2,000$)	Nonpara. Bootstrap	Smooth Bootstrap	Jackknife
Ave$\{\widehat{\text{Var}}\,(U_n)\}$	0.0578	0.0569	0.0568	0.0469	1.8×10^{-4}	0.0822
SD$\{\widehat{\text{Var}}\,(U_n)\}$	0.0571	0.0578	0.0573	0.0147	2.7×10^{-4}	0.0397
$\mid$ Bias $\mid$ /SD	0.000	0.016	0.017	0.740	210.912	0.0616
Computation		0.014	0.051	8.039	8.213	1.540

The standard deviation of $\hat{V}_u$ in column one was estimated based on (4.2). With $mB = 2,000$, the variance of the incomplete variance estimator was close to $\text{Var}\,(\hat{V}_u)$. The bias of the bootstrap and jackknife estimators were significantly larger than that of $\hat{V}_u$. The ordinary jackknife can cause significant positive bias, as seen in Table 1−3. For other evidence, see Table 1 in Wu (1986). The smoothing that is used in the smooth bootstrap estimator creates normality in the data and thus results in significant bias. With partition resampling scheme, the unbiased variance estimator is efficient to compute compared to its bootstrap and jackknife counterparts. This advantage is more obvious for half-sampling, as each size-n sample is partitioned into two disjoint half-samples which can be used to estimate $Q(0)$ directly.

As the unbiased variance estimator can be written as a complete U-statistic itself (Remark 3), it can be thought of as a subsampling unbiased estimator, where we average over a sampling-without-replacement scheme. In comparison, bootstrap methods are based on sampling with replacement. The subsampling approach is superior in terms of unbiasedness.

5.3. A numerical study of the negative estimation problem

We investigated the performance of the simple nonnegative variance estimator $\hat{V}_u^+$ based on a simulation study with the same setting as in the previous section ($mB = 2,000$).

We have $\hat{V}_u^+ = \max\{\hat{V}_u, S_U^2\}$, where $S_U^2 = Q(k) - Q(k-1)$. In this section we use the complete data notation for simplicity, but the simulation calculations were based on incomplete estimators. Figure 1 displays the relationship between $\hat{V}_u$ and S_U^2 based on the simulation as in Section 5.2. The rightmost vertical dashed line in each plot marks the level of the true value for the variance. The filled squares represent values of $\hat{V}_u^+ = S_U^2$ when $\hat{V}_u < 0$, the negative values we wish to eliminate. The filled triangles represent values of $\hat{V}_u^+ = S_U^2$ when $S_U^2 > \hat{V}_u > 0$, the cases where we use a downward biased estimate to substitute for $\hat{V}_u$.

Remark 5. In Figure 1, the scale of the vertical axis is different from the horizontal axis. Notice the structural relationship between S_U^2 and $\hat{V}_U$ in panel (c). We think it is interesting mathematical conundrum rather than a statistically useful structure.

The simulation generated negative values for $\hat{V}_u$, at 1.9% when $k = 0.1n$, 1.3% when $k = 0.25n$, and 3.5% when $k = 0.5n$. The number of negative values decreased as the partition sampling size B was increased. Figure 1 shows that it is rare for $S_U^2 > \hat{V}_u > 0$. Indeed, S_U^2 is so small in magnitude and so relatively tightly distributed that most cases occurred when $\hat{V}_u$ was negative. In addition, when $\hat{V}_u^+ > \hat{V}_u$, it was often closer to $\mathrm{Var}\,(U_n)$. This helped to reduce the mean squared error of the positive estimator as seen in Table 4. The variance estimator worked well in this study; we discuss future work needed on the distribution of $\hat{V}_u$ and $\hat{V}_u^+$ in the discussion section.

We can compare S_U^2, the forced positive estimator $\hat{V}_u^+$, and $\hat{V}_u$ in terms of their mean, standard deviation, and mean squared error. The results based on the same simulation are summarized in Table 4.

Table 4 indicates that S_U^2 was consistently much smaller than $\hat{V}_u$ in expectation. The positive variance estimator $\hat{V}_u^+$ led to an estimate with negligible bias as compared with the unbiased variance estimator $\hat{V}_u$. In addition, its variance was slightly smaller than the variance of $\hat{V}_u$. We conclude that the forced positive estimator is an inexpensive but effective amendment to variance estimation.

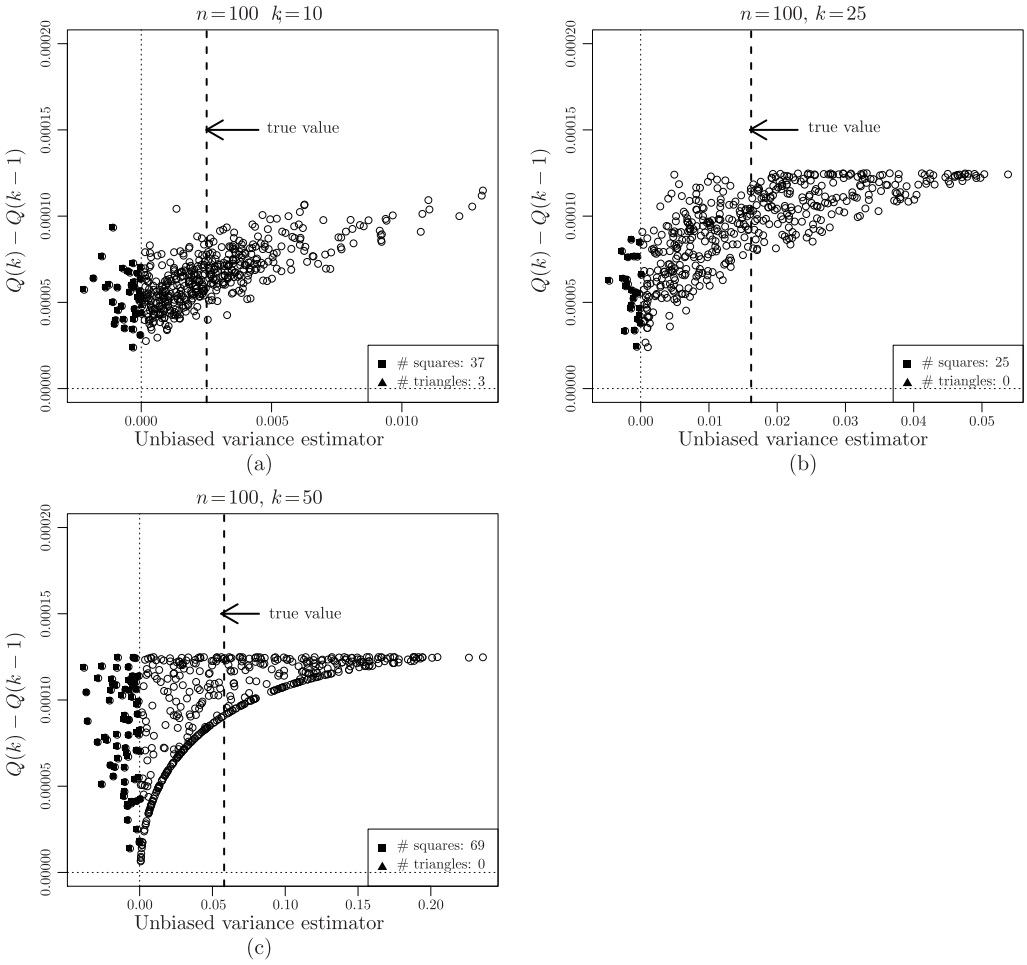

Figure 1. Plots of S_U^2 versus $\hat{V}_u$.

Table 4. Comparing $S_U^2, \hat{V}_u$, and $\hat{V}_u^+$.

	$n=100, k=10$			$n=100, k=25$			$n=100, k=50$		
	S_U^2	$\hat{V}_u$	$\hat{V}_u^+$	S_U^2	$\hat{V}_u$	$\hat{V}_u^+$	S_U^2	$\hat{V}_u$	$\hat{V}_u^+$
True		0.0025			0.0162			0.0578	
Mean	0.00007	0.0026	0.0026	0.00009	0.0161	0.0161	0.00010	0.0568	0.0583
SD	0.00002	0.0024	0.0023	0.00002	0.0128	0.0128	0.00003	0.0573	0.0555
MSE	5.93e-6	5.98e-6	5.68e-6	2.59e-4	1.66e-4	1.64e-4	3.33e-3	3.33e-3	3.31e-3

6. Model Selection with Real Data Example

Although unconventional, we treat cross-validation as a U-statistic methodology. In this section, we demonstrate our method by considering a likelihood

cross-validation problem (Van der Laan, Dudoit, and Keles (2004)).

In this methodology, the kernel size k is $\tilde{n}+1$, where $\tilde{n}$ ($\tilde{n} < n$) is the size of the training sample. Thus, if we are to estimate the variability of the risk estimation, we need to use training samples with size $\tilde{n} < n/2$. There are a number of important reasons to use smaller training samples. For one, see Hall and Robinson (2009), where "bagging cross-validation" based on training samples of size $n/2$ was shown to significantly improve mean integrated square error performance in kernel density bandwidth selection. For others, see Shao (1993), Van der Laan, Dudoit, and Keles (2004), and Shah and Samworth (2012).

We present an additional reason to consider smaller training samples in likelihood cross-validation. Standard likelihood cross-validation corresponds to a nonparametric version of AIC-type risk estimation (Akaike (1974)). We argue that likelihood cross-validation with training samples of size $\tilde{n} = n/(\log n - 1)$ is equivalent to the standard BIC risk estimation (Schwarz (1978)), and verify this in a rich data set that has many parametric models to compare. We note that likelihood cross-validation is more generally applicable than AIC and BIC, but we have chosen a model where both exist and can be compared. We then show how to use our U-statistic variance estimation methodology to eliminate models that are not significantly smaller in BIC risk than other, more parsimonious, models.

6.1. Model selection methodology

Let the true distribution have the density function $\tau(x)dx$. For each δ in an index set, say $\delta \in \{1, 2, 3, \ldots\}$, let model $\mathcal{M}_\delta$ be a class of densities $\{m_{\theta_\delta} \mid \theta_\delta \in \Theta_\delta \subseteq \mathcal{R}^{p_\delta}\}$ indexed by the parameter θ_δ. Here p_δ is the dimension of parameter θ_δ. In a later example, each model $\mathcal{M}_\delta$ corresponds to a particular logistic regression model, and our goal is to pick the model $\mathcal{M}_\delta$ with the smallest risk.

The Kullback-Leiber distance between τ and m_{θ_δ} is

$$d(\tau, \delta) = \int \tau(x) \log\{\tau(x)/m_{\theta_\delta}(x)\} dx.$$

Let $\mathcal{X}_n$ be a sample of size n, and let $\hat{\theta}_\delta(\mathcal{X}_n)$ be the parameter estimator for model $\mathcal{M}_\delta$ based on that sample. Then, the relative Kullback-Leibler risk for using fitted model $m_{\hat{\theta}_\delta(\mathcal{X}_n)}$ has the form $R(\delta, n) = -E_{\mathcal{X}_{n+1}}\{\log \hat{m}_{\hat{\theta}_\delta(\mathcal{X}_n)}(X_{n+1})\}$, where $\mathcal{X}_{n+1}$ is the size-$(n+1)$ sample $(\mathcal{X}_n, X_{n+1})$. Risk depends on the sample size n used to estimate θ_δ.

Ray and Lindsay (2008) noted that one can create a more flexible family of risk estimators by defining the relative risk based on a subsample of size $\tilde{n}$. Indeed, much of the risk estimation literature does this implicitly through the

choice of the sample size for the "training set". We call $\tilde{n}$ the training sample size and estimate the parameter θ_δ by $\hat{\theta}_\delta(\mathcal{X}_{\tilde{n}})$, with

$$R(\delta, \tilde{n}) = -E_{\mathcal{X}_{\tilde{n}+1}} \left\{ \log \hat{m}_{\hat{\theta}_\delta(\mathcal{X}_{\tilde{n}})}(X_{\tilde{n}+1}) \right\}. \tag{6.1}$$

We construct a U-statistic estimate for the relative risk $R(\delta, \tilde{n})$ by defining a symmetric kernel function of size $k = \tilde{n} + 1$ ($\tilde{n} \le n - 1$),

$$\phi_{KL,\delta}(\mathcal{X}_{\tilde{n}+1}) = -\frac{1}{\tilde{n}+1} \sum_{i=1}^{\tilde{n}+1} \log \hat{m}_{\hat{\theta}_\delta(\mathcal{X}_{(-i)})}(X_i), \tag{6.2}$$

where $\mathcal{X}_{\tilde{n}+1}$ is a sample of size $\tilde{n}+1$ taken out of $\mathcal{X}_n$, X_i is the ith observation in $\mathcal{X}_{\tilde{n}+1}$, and $\mathcal{X}_{(-i)}$ contains the $\tilde{n}$ observations in $\mathcal{X}_{\tilde{n}+1}$ except X_i. The kernel size k in (6.2) depends on the training sample size $\tilde{n}$ which is usually of order n.

The U-statistic estimate for the relative risk $R(\delta, \tilde{n})$ has the form

$$\text{LCV}(\delta, \tilde{n}) = \frac{1}{\binom{n}{\tilde{n}+1}} \sum_{1 \le i_1 < \cdots < i_{\tilde{n}+1} \le n} \phi_{KL,\delta}(X_{i_1}, \ldots, X_{i_{\tilde{n}+1}}). \tag{6.3}$$

The letters "LCV" stand for likelihood cross-validation. To reduce computational efforts, we can use an incomplete U-statistic to estimate $R(\delta, \tilde{n})$,

$$\text{LCV}_B(\delta, \tilde{n}) = \frac{1}{B} \sum_{i=1}^{B} \phi_{KL,\delta}(\tilde{S}_i) \ (B \in \mathcal{N}^+). \tag{6.4}$$

Here $\phi_{KL,\delta}$ is defined in (6.2), and $\tilde{S}_i$ is a subsample of size $\tilde{n}+1$ taken out of $\mathcal{X}_n$.

An alternative for estimating the risk in $R(\delta, n)$ is the AIC method (Akaike (1974)). If we generalize this method to an arbitrary $\tilde{n}$, we get a formula based on some asymptotic expansions that include regularity assumptions. This generalization of AIC estimates the relative risk $R(\delta, \tilde{n})$ with AIC$/(2n)$ using training sample size $\tilde{n}$ based on, see Lindsay and Liu (2007) and Ray and Lindsay (2008),

$$\text{AIC}(\delta, \tilde{n}) = 2n \left\{ -\frac{l}{n} + \frac{p_\delta}{2n} + \frac{p_\delta}{2\tilde{n}} \right\},$$

where l is the log-likelihood based on the fitted model $m_{\hat{\theta}(\mathcal{X}_{\tilde{n}})}$. The conventional AIC has $\tilde{n} = n$; the conventional BIC (Schwarz (1978) can be generated with $\tilde{n} = n/(\log n - 1)$ (Ray and Lindsay (2008)).

The likelihood cross-validation statistic (6.3) is an alternative to AIC$(\delta, \tilde{n})/2n$ as a method for estimating the same risk value. If one uses $\tilde{n} = n/(\log n - 1)$, it is an analogue of the BIC method.

Generally speaking, $\text{LCV}(\delta, \tilde{n})$ and $\text{AIC}(\delta, \tilde{n})$ are not always asymptotically equivalent, as the derivation of $\text{AIC}(\delta, \tilde{n})$ is based on asymptotic chi-square approximations. These are invalid when τ is not in the model or when standard likelihood asymptotics fail to hold even when the model is true. For a model that is not regular, like a mixture model, it might be preferable to construct a U-statistic risk estimate, knowing that the AIC criterion has weaker foundations.

Remark 6. Following Benjamini and Gavrilov (2009), if one compares two nested parametric models, say $\mathcal{M}_i$ and $\mathcal{M}_j$ with $p_j = p_i + 1$, based on the generalized AIC method, the difference in their AIC scores is $\text{AIC}(\delta = i, \tilde{n}) - \text{AIC}(\delta = j, \tilde{n}) = 2(l_j - l_i) - (\tilde{n}+n)/\tilde{n}$, where l_δ is the log-likelihood based on the fitted model $m_{\hat{\theta}_\delta(\mathcal{X}_{\tilde{n}})}$ $(\delta = i, j)$. When the smaller model is true, under regularity conditions $2(l_j - l_i) \to \mathcal{X}^2(1)$ as $n \to \infty$. Therefore, when considering the hypothesis test $H_0 : \mathcal{M}_i$ versus $H_a : \mathcal{M}_j$, the decision of favoring the larger model $\mathcal{M}_j$ is equivalent to rejecting the null hypothesis with a level of significance α approximately equal to $\text{P}\left\{\mathcal{X}^2(1) > (\tilde{n} + n)/\tilde{n}\right\}$. We call this number the "p-value index". This level is about 0.16 for $\tilde{n} = n$. When $\tilde{n} = n/2$, it corresponds to $\alpha = 0.08$; when $\tilde{n} = n/3$, it yields $\alpha = 0.05$. This provides further motivation for using half or less subsampling in the estimation of θ_δ when conservative modelling is desired.

6.2. Numerical results

We consider the Census Income data, also known as the Adult data. It is available in the R package "arules" under the name AdultUCI. This data set was extracted from the 1994 census database. A task is to determine whether an individual makes over 50 thousand a year given a set of predictors. The AdultUCI data set has records of 15 variables from 48,842 individuals. It has been cited in Kohavi (1996), Caruana and Niculescu-Mizil (2004), and Agrawal, Ikant, and Thomas (2005).

We focused on a set of nine variables, including age (A), workclass (W), education (E), marital-status (M), occupation (O), race (R), sex (S), hours-per-week (H), and income (I). The attributes age and hours-per-week were each discretized into four categories. After some data manipulations, our data set had $n = 26897$ observations and $p = 9$ attributes. More details of data trimming process can be found in the Appendix.

To predict whether an individual makes more than 50 thousand dollars a year, it is natural to consider income (I), a binary variable, as the response. Take the probability that the ith individual with yearly income higher than 50 thousand, coded as 1, is p_i. Then, income is Bernoulli(p_i), where p_i is the probability that the ith observation is 1. We fit logistic regression models

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{linear form of some predictors.}$$

Table 5. Candidate Models.

$\mathcal{M}_\delta$	
Model 1	M
Model 2	M+E
Model 3	M+E+O
Model 4	M+E+O+A
Model 5	M+E+O+A+H
Model 6	M+E+O+A+H+W
Model 7	M+E+O+A+H+W+S
Model 8	M+E+O+A+H+W+S+R

Table 6. Compare Different Model Selection Methods.

Model	AIC$/(2n)$	Rank_{AIC}	BIC$/(2n)$	Rank_{BIC}	LCV	Rank_{LCV}
1	0.4686	8	0.4697	8	0.4721	8
2	0.4183	7	0.4205	7	0.4233	7
3	0.4035	6	0.4076	6	0.4114	6
4	0.3918	5	0.3964	5	0.4003	5
5	0.3846	4	0.3897	4	0.3935	3
6	0.3823	3	0.3884	**1**	0.3921	**1**
7	0.3822	2	0.3884	2	0.3922	2
8	0.3820	**1**	0.3888	3	0.3940	4

After initial model screening using stepwise selection, the best model in the BIC sense was identified as the one with six main effects. We considered eight candidate models with only first-order terms in our model comparison. These models are summarized in Table 5. Here we only considered the best model of each size.

The model selection index for each criterion is summarized in Table 6. We compare the generalized AIC method, the BIC method, and the LCV method with training sample size $\tilde{n} = n/(\log n - 1) = 2,923$. This size of $\tilde{n}$ yields a p-value index of 0.0014 (see Remark 6). The calculation of the LCV scores was based on implementing a partition resampling scheme with $B = 55$ random partitions, where each partition yields 9 disjoint subsets of size $\tilde{n} + 1$. After a partition was created, all models were fit using that same partition (this is necessary if one is to accurately estimate the variance of the difference in risk estimates, see Section 6.3). Table 6 also includes the rankings of the eight candidate models based on optimization of each criterion.

The result in Table 6 demonstrates how numerically similar the BIC method and the LCV method can be in model selection for parametric models. We know that the BIC has some nice properties, like consistency, and we have shown in Table 6 that the BIC/(2n) criterion provides almost identical model rankings and nearly identical risk estimation as the LCV. (Although the rankings for Model

5 and Model 8 are different between the BIC and the LCV methods, it can be shown that the risk in Model 8 is not significantly worse than Model 5. However, as Model 5 is more parsimonious, we would prefer Model 5 to Model 8, as seen in the last column of Table 7.) In comparison, the AIC method tended to favor larger models, as it has a smaller penalty term for model complexity compared with the BIC criterion.

6.3. Pairwise model comparison based on the U-statistic criterion

One fundamental problem with using risk estimation to choose models is that the high variability in the risk estimates can lead one to choose a less parsimonious model than is needed to minimize risk. Note that the LCV score for Model 6 is only slightly smaller than that of Model 5. If the difference in risks is not significant, then one may prefer to select Model 5 even if its LCV score is slightly larger.

Suppose we want to compare two models, say $\mathcal{M}_i$ and $\mathcal{M}_j$, where $\mathcal{M}_i$ represent the more parsimonious model (not necessarily nested in $\mathcal{M}_j$). Let $\text{LCV}_{\text{diff}}(\tilde{n}) = \text{LCV}(\delta = i, \tilde{n}) - \text{LCV}(\delta = j, \tilde{n})$, and take $\theta = E\{\text{LCV}_{\text{diff}}(\tilde{n})\}$. We wish to use the parsimonious model unless there is significant evidence against it, and so consider the hypotheses

$$H_0 : \theta \leq 0; \ H_a : \theta > 0. \tag{6.5}$$

We use an analogue of the paired t-test to test this hypothesis.

Note that $\text{LCV}(\delta = i, \tilde{n})$ and $\text{LCV}(\delta = j, \tilde{n})$ are U-statistics and denote their kernels as ϕ_i and ϕ_j. The difference LCV_{diff} is a U-statistic with kernel $\phi_{\text{diff}} = \phi_i - \phi_j$, and its variance can be estimated unbiasedly with $\hat{V}_{\text{diff}}$ based on (2.3). For subsampling, we use the same partitions on each data set. In this example, both the kernel size k and the sample size n are relatively large ($k = 2{,}924, n = 26{,}897$). We calculated LCV_{diff} and their variance estimates based on partition resampling with $B = 55$. Rejection follows if $\text{LCV}_{\text{diff}}/\sqrt{\hat{V}_{\text{diff}}} > 1.645$. This is fairly conservative relative to the one-standard-error rule discussed in the next section. Since the selected model depends on this cut-off, one could decide in view of models selected at different cut-offs. Ideally it would not make much difference.

Comparing Model 5 and Model 6, for instance, $\text{LCV}_{\text{diff}} = 0.00133$, $\sqrt{\hat{V}_{\text{diff}}} = 0.000438$, and $t = \text{LCV}_{\text{diff}}/\sqrt{\hat{V}_{\text{diff}}} = 3.04$. (The ratio of the t-statistics, $\sqrt{\hat{V}_u}/\sqrt{\hat{V}_{\text{diff}}}$, is predicted to be within the range (0.89,1.10) according to the discussion in Section 4.2. Therefore, a larger partition size B would likely not change the significance of the t-statistic.) We might therefore conclude that the risk of Model 5 is significantly larger than that of Model 6, and we may prefer Model 6 to Model 5.

Table 7. Pairwise Comparison of Risks.

δ	1	2	3	4	5	6	7	8
1	X	2	3	4	5	**6**	7	8
2		X	3	4	5	**6**	7	8
3			X	4	5	**6**	7	8
4				X	5	**6**	7	8
5					X	**6**	7	5
6						X	**6**	**6**
7							X	7
8								X

Table 7 summarizes the results of pairwise tests between all of the eight candidate models of Table 5. Here, the integer value in each cell gives the "winner" of each pair taken at the nominal $\alpha = 0.05$ level, where winner means the larger model if it is significantly better than the smaller one; otherwise, the winner is the smaller model.

If row i has entries only equal to "i", then that model is not significantly worse in risk to any larger model. If column j has entries only equal to "j", then model j is significantly better than any smaller model. In Table 7 we see that Model 6 had both these characteristics.

6.4. Comparison to the one-standard-error (1-SE) rule

A referee has pointed out that our model selection procedure bears similarity to the 1-SE rule of Breiman et al. (1984) that is a commonly used rule of thumb (see, for example, Hastie, Tibshirani, and Friedman (2009)). The 1-SE rule suggests that one uses the most parsimonious model whose risk is no more than one standard error above the risk of the best model. Let $\mathcal{M}_j$ be the model with the smallest LCV score and $\mathcal{M}_i$ be a more parsimonious model. Based on the 1-SE rule, we should prefer Model i unless $(\text{LCV}_i - \text{LCV}_j)/\sqrt{\hat{V}_j} > 1$, where $\hat{V}_j$ is the variance estimate of Model j. We could have carried out such an analysis using our variance estimator to compute the standard error of LCV_j.

The 1-SE rule looks like a t-test rule with a modified test statistic. The test statistic used in our pairwise model comparison was $T_1 = \text{LCV}_{\text{diff}}/\sqrt{\hat{V}_{\text{diff}}}$. This would equal $T_2 = \text{LCV}_{\text{diff}}/\sqrt{2\hat{V}_j}$ if the U-statistics involved had equal variances and if they were statistically independent. Under those assumptions, the 1-SE rule would be roughly equivalent to our pairwise comparison when used at critical value $1/\sqrt{2}$. However, there is a good reason to think that the two U-statistics involved show positive association, and so we anticipate that $\hat{V}_{\text{diff}} < 2\hat{V}_j$. A table demonstrates the comparison between $\hat{V}_j$ and $\hat{V}_{\text{diff}}$ based on our data example.

Table 8. Comparison between individual variance and pairwise variance.

Model	1	2	3	4	5	6	7	8
$\sqrt{\hat{V}_j}$	0.00317	0.00325	0.00329	0.00329	0.00328	0.00332	0.00332	0.00328
$\sqrt{\hat{V}_{\text{diff}}}$	NA	0.00203	0.00112	0.00097	0.00076	0.00044	0.00012	0.00052
$\frac{\sqrt{2\hat{V}_j}}{\sqrt{\hat{V}_{\text{diff}}}}$	NA	2.2668	4.1445	4.8080	6.0819	10.7067	39.7538	8.9612

The value $\hat{V}_{\text{diff}}$ under Model j is the estimated variance of LCV_{diff} when comparing Model $j-1$ with Model j.

The last row shows the t-ratio T_1/T_2. We conclude that T_1 is much larger than T_2. Thus, if we used the same critical value for T_2 as for T_1, with T_2 we would be much less likely to choose the larger model despite the evidence of its superiority. The effect of correlation was not only large, it was quite variable, making it hard to predict the magnitude of this conservative behavior. We think that this is an important finding for cross-validation methodology. It is clear that one can estimate the difference in risk between two models much more accurately than it would appear from one-at-a-time standard errors.

7. Discussion

We have shown that one can successfully build an unbiased estimator for U-statistics of large kernel size k. The main limitation on our method, at least as applied to cross-validation, is that we require $k \leq n/2$. Since the kernel size k is often determined by the size of the training sample size, say $\tilde{n}$, this would seem to limit the method to estimating risk variances when the training sample size does not satisfy $\tilde{n} < n/2$. For example, for likelihood cross-validation we need $k = \tilde{n} + 1 \leq n/2$. As a consequence one could not directly estimate the risk for such commonly used methods as leave-one-out or ten-fold cross-validation.

We believe these limitations can be overcome. For example, the cross-validation method, as applied to the squared error risk in density estimation, has a U-statistic kernel of size 2, independent of the training sample size, and so our method can be applied directly. We have preliminary results showing that it is often more effective to estimate risk with $\tilde{n}$ small, and then extrapolate the answer to create a risk estimate for sample size n. For more on this see the bagging bandwidth selection method of Hall and Robinson (2009). Their conclusion was that half samples were much better than leave-one-out cross-validation. Furthermore, Shao (1993) showed that consistency of cross-validation in model selection problems can require $\tilde{n}/n \to 0$.

We hope to build a better understanding of the asymptotic properties of the unbiased estimator as well as its nonparametric competitors. In addition, we

anticipate the development of methods that would enable one to balance bias and variance in a more flexible way.

Acknowledgement

This research was partially supported by the National Science Foundation (DMS 0714839). The authors would like to thank the associate editor and two referees for their valuable comments and suggestions that lead to considerable improvement of the original manuscript.

Appendix

Proof for the U-statistic representation of $\hat{V}_u$

Let S_a, S_b be two subsamples of size k taken from S_{2k}. We determine weights $\omega(a,b)$ such that $\binom{n}{2k}^{-1}\sum_{S_{2k}\subseteq\mathcal{X}_n}\sum_{S_a,S_b\subset S_{2k}}\phi(S_a)\phi(S_b)w(a,b)$ is proportional to $Q(k)=\binom{n}{k}^{-2}\sum_{S_a,S_b\subset S_n}\phi(S_a)\phi(S_b)$.

In the first of these, each pair (S_a,S_b) appears once or zero times inside the bracket and will appear once in the outer sum for each size-$2k$ sample that contains the pair. In the second, each pair (S_a,S_b) appears once inside the bracketed sum. Let $n(a,b)$ be the number of different size-$2k$ samples S_{2k} in which S_a,S_b are subsets. Then we can set $\omega(a,b)=1/n(a,b)$. After adjusting the initial constants,

$$Q(k)=\binom{n}{2k}^{-1}\sum_{S_{2k}\subseteq\mathcal{X}_n}\psi_k(S_{2k}),\ \psi_k(S_{2k})=\binom{n}{2k}\binom{n}{k}^{-2}\sum_{S_a,S_b\subset S_{2k}}\phi(S_a)\phi(S_b)\omega(a,b).$$

Consider $Q(0)=\sum_{P_0}\phi(S_a)\phi(S_b)/N_0$. It is easily shown that

$$\begin{aligned}Q(0)&=\binom{n}{2k}^{-1}\sum_{S_{2k}\subseteq\mathcal{X}_n}\psi_0(S_{2k}),\ \psi_0(S_{2k})\\&=\binom{n}{2k}N_0^{-1}\sum_{S_a,S_b\subset S_{2k}}\phi(S_a)\phi(S_b)\mathbb{I}\{S_a\cap S_b=\emptyset\}.\end{aligned}$$

If $\psi(S_{2k})=\psi_k(S_{2k})-\psi_0(S_{2k})$, $\hat{V}_u$ has a representation as a complete U-statistic:

$$\hat{V}_u=Q(k)-Q(0)=\frac{1}{\binom{n}{2k}}\sum_{S_{2k}\subseteq\mathcal{X}_n}\psi(S_{2k}).$$

Proof for Proposition 1. Because $\sum_{j=1}^{m}\left\{\phi(S_{a,j})-\bar{\phi}_a\right\}^2=\sum_{j\neq l}\{\phi(S_{a,j})$

$-\phi(S_{a,l})\}^2/(2m)$, we have

$$\sum_{a=1}^{\mathbb{B}}\sum_{j=1}^{m}\{\phi(S_{a,j})-\bar{\phi}_a\}^2=\frac{1}{2m}\sum_{a=1}^{\mathbb{B}}\sum_{j\neq l}\{\phi(S_{a,j})-\phi(S_{a,l})\}^2$$
$$=\sum_{a=1}^{\mathbb{B}}\sum_{j=1}^{m}\phi(S_{a,j})^2-\frac{1}{m}\sum_{a=1}^{\mathbb{B}}\sum_{j,l=1}^{m}\phi(S_{a,j})\phi(S_{a,l}).$$

Note that $\sum_{a=1}^{\mathbb{B}}(\bar{\phi}_a-\bar{\phi})^2=\sum_{a=1}^{\mathbb{B}}\sum_{j,l=1}^{m}\phi(S_{a,j})\phi(S_{a,l})/m^2-\mathbb{B}\bar{\phi}^2$. Among the $\mathbb{B}$ partitions (each with m nonoverlapped subsets of size k) there are $m(m-1)\mathbb{B}$ nonoverlapped within-class pairs in total. Therefore,

$$\frac{1}{m(m-1)\mathbb{B}}\Big\{\frac{1}{2m}\sum_{a=1}^{\mathbb{B}}\sum_{j\neq l}(\phi(S_{a,j})-\phi(S_{a,l}))^2-m(m-1)\sum_{a=1}^{\mathbb{B}}(\bar{\phi}_a-\bar{\phi})^2\Big\}$$
$$=\bar{\phi}^2-\frac{1}{m(m-1)\mathbb{B}}\sum_{a=1}^{\mathbb{B}}\sum_{j\neq l}\phi(S_{a,j})\phi(S_{a,l})=Q(k)-Q(0).$$

Proof for Lemma 1. Let A_c denote the set of all pairs of size-k samples with overlaps exactly equal to c, and let n_c be the number of pairs in A_c $(0\le c\le k)$. It is easily seen that $P_c=\bigcup_{l=0}^{c}A_c$, and $N_c=\sum_{l=0}^{c}n_c$, where P_c, N_c are defined in Section 2.2.

Given c $(1\le c\le k)$, consider $Q(c)-Q(c-1)$. By (2.2), we have $Q(c)=N_c^{-1}\left\{\sum_{A_0}\phi(S_i)\phi(S_j)+\cdots+\sum_{A_c}\phi(S_i)\phi(S_j)\right\}$. Then, it can be shown that

$$Q(c)-Q(c-1)=\frac{-n_c}{N_cN_{c-1}}\Big\{\sum_{A_0}\phi(S_i)\phi(S_j)+\cdots+\sum_{A_{c-1}}\phi(S_i)\phi(S_j)\Big\}$$
$$+\frac{1}{N_c}\sum_{A_c}\phi(S_i)\phi(S_j).$$

Because $E\{\phi(S_i)\phi(S_j)\mid O(S_1,S_2)=j\}=\sigma_c^2+\theta^2$, where $\sigma_c^2=\mathrm{Var}\{\phi_c(X_1,\ldots,X_c)\}$, we have

$$E\{Q(c)-Q(c-1)\}=\frac{n_c}{N_c}(\sigma_c^2+\theta^2)-\frac{n_c}{N_cN_{c-1}}\sum_{j=0}^{c-1}n_j(\sigma_j^2+\theta^2)$$
$$=\frac{n_c}{N_cN_{c-1}}\sum_{j=0}^{c-1}n_j(\sigma_c^2-\sigma_j^2).$$

Since $\sigma_c^2\ge\sigma_j^2$ $(1\le j<c\le k)$ (see Theorem 5.1 in Hoeffding (1948)), each term inside the summation is nonnegative. Then, $E\{Q(c)-Q(c-1)\}\ge 0$. The

fact that $E\{Q(k)-Q(c-1)\}-E\{Q(k)-Q(c)\}=E\{Q(c)-Q(c-1)\}$ yields the result.

Proof for Lemma 2.
Since $Q(k-1)=(N_k/N_{k-1})\left\{Q(k)-(1/N_k)\sum_{A_k}\phi(S_i)\phi(S_j)\right\}$, we have

$$Q(k)-Q(k-1)=\frac{1}{N_{k-1}}\sum_{i=1}^{\mathbb{N}}\phi^2(S_i)-\frac{N_k-N_{k-1}}{N_{k-1}}Q(k)=\frac{1}{N_{k-1}}\left\{\sum_{i=1}^{\mathbb{N}}\phi^2(S_i)-n_kU_n^2\right\}.$$

Notice that $n_k=\mathbb{N}$, $N_{k-1}=\mathbb{N}(\mathbb{N}-1)$, and $U_n=\sum_{i=1}^{\mathbb{N}}\phi(S_i)/\mathbb{N}$. Therefore, $Q(k)-Q(k-1)=\sum_{i=1}^{\mathbb{N}}(\phi(S_i)-U_n)^2/\{\mathbb{N}(\mathbb{N}-1)\}$.

Data Set in Section 6

For the Census Income data used in Section 6, we first removed attributes "fnlwgt", "education-num", "native-country", "relationship", "capital-gain", and "capital-loss", as those variables are either of little interest or redundant given other existing predictors. In addition, we discarded observations from individuals with education level lower than high school, since those people commonly do not have stable income. We discretized variable "age" into the categories Young, Middle-aged, Senior, Old, based on cut-off intervals (0,30], (30,45], (45, 65], (65, 100); variable "hours-per-week" (number of hours worked per week) was discretized into Part-time, Full-time, Over-time, Workaholic, based on cut-off intervals (0, 25], (25, 45], (45, 60], (60, 200). Then, we trimmed all of the observations with missing values. After these adjustments, we ended up with a data frame with $n = 26897$ observations and the following nine categorical variables (the number of levels for each variable is shown inside the parentheses): age (4), workclass (8), education (8), occupation (14), race (5), sex (2), hours-per-week (4), and income (2).

References

Akaike, H. (1974). A new look at the statistical identification. *IEEE Trans. Automat. Control* **19**, 716-723.

Agrawal, R., Ikant, R., and Thomas, D. (2005). Privacy Preserving OLAP. *SIGMOD Conference.*

Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Statist.* **3**, 179-198.

Blom, G. (1976). Some properties of incomplete u-statistics. *Biometrika* **63**, 573-580.

Breiman, L, Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Chapman and Hall/CRC.

Caruana, R. and Niculescu-Mizil, A. (2004). An empirical evaluation of supervised learning for ROC area. *Proceedings of the* 1*st Workshop on ROC Analysis*, 1-8.

Efron, B. (1987). *The jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial Mathematics.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Folsom, R. E. (1984). Probability sample U-statistics: theory and applications for complex sample designs. Thesis, University of North Carolina, Chapel Hill.

Fraser, D. A. S. (1954). Completeness of order statistics. *Canad. J. Math.* **6**, 42-45.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *Elements of Statistical Learning*. Springer, New York.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Inst. Math. Statist.* **19**, 293-325.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 202-207.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker, New York.

Lindsay, B. G. and Liu, J. (2007). Model Assessment Tools for a Model False World, unsubmitted manuscript (available from B.G. Lindsay).

Liu, J. and Lindsay, B. G. (2009). Model assessment tools for a model false world, *Statist. Sci.* **24**, 255-385.

Markatou, M., Tian, H., Biswas, S. and Hripcsak, G. (2005). Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.* **6**, 1127-1168.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning* **52**, 239-281.

Politis, D. N. R., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *J. Roy. Statist. Soc. Ser. B* **70**, 95-118.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shah, R. D. and Samworth, R. J. (2012). Variable selection with error control: another look at stability selection. *J. Roy. Statist. Soc. Ser. B* **74**, 50-80.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Van der Laan, M. J., Dudoit, S. and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statist. Appl. Genet. Mol. Biol.* **3**, Article 4.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261-1295.

Department of Mathematics and Statistics, Williams College, Williamstown MA 01267, USA.

E-mail: qing.w.wang@williams.edu

Department of Statistics, The Pennsylvania State University, University Park PA 16802, USA.

E-mail: bgl@stat.psu.edu

(Received July 2012; accepted July 2013)