Statistica Sinica **27** (2017), 1155-1174
doi:https://doi.org/10.5705/ss.202015.0463

PSEUDO-KERNEL METHOD IN U-STATISTIC VARIANCE ESTIMATION WITH LARGE KERNEL SIZE

Qing Wang and Bruce Lindsay

Wellesley College and Pennsylvania State University

Abstract: This paper addresses the problem of variance estimation of a general U-statistic with large kernel size (degree) k. U-statistics form a class of unbiased estimators. It was first proposed in Hoeffding (1948) and has since been widely used in many statistical applications. Wang and Lindsay (2014) propose an unbiased variance estimator for a general U-statistic; it is applicable provided that the kernel size k is at most half of the sample size n. This condition restricts its application to common K-fold cross-validation problems. We devise a pseudo-kernel variance estimator that can be realized in the same fashion as the unbiased variance estimator, but is defined based on a pseudo-kernel function of degree two. We demonstrate how to construct a pseudo-kernel function and show that the resulting variance estimator is second-order unbiased. Moreover, we develop an efficient realization of the proposal in the context of K-fold cross-validation. The proposed variance estimator shows comparable performance with significantly improved computational efficiency compared to its bootstrap and jackknife counterparts in simulation and data analysis in the context of model selection using the "one-standard-error" rule.

Key words and phrases: K-fold cross-validation, Kullback-Leibler distance, pseudo-kernel, second-order unbiased, U-statistic, variance estimation.

1. Introduction

Variance measures the uncertainty of a random quantity. Therefore, variance estimation is crucial in evaluating the performance of a point estimator or conducting inference for a statistical methodology. In statistical practice, an unbiased estimator is often desired. Because most unbiased estimators in common use can be written in the form of a U-statistic, in this paper we focus on the problem of variance estimation of a U-statistic.

Consider a parameter of interest θ that is defined as the expectation of a symmetric function ϕ with k components.

$$\theta = E[\phi(X_1, \ldots, X_k)], \tag{1.1}$$

where $X_1, \ldots, X_k$ are independent and identically distributed (i.i.d.) random variables, and symmetry means that the function ϕ is permutation invariant

of its k components. We call ϕ the kernel function and k the kernel size. The smallest integer k for (1.1) to hold is also referred to as the degree of the statistical functional θ. Given an i.i.d. sample of size n ($n \geq k$), $\mathcal{X}_n = (X_1, \ldots, X_n)$, from some distribution with probability density f, a U-statistic with a symmetric kernel function ϕ of degree k is defined in Hoeffding (1948) as

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

Although the kernel function ϕ is often scalar-valued, most of the results can be easily generalized to vector-valued cases.

Let $\phi_c(x_1, \ldots, x_c) = E[\phi(X_1, \ldots, X_k) \mid X_1 = x_1, \ldots, X_c = x_c]$. Write $\mathrm{Var}[\phi_c(X_1, \ldots, X_c)] = \sigma_c^2$ ($1 \leq c \leq k$). Hoeffding (1948) gives the closed-form expression of the variance of U_n:

$$\mathrm{Var}(U_n) = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{k-c} \sigma_c^2. \tag{1.2}$$

Under the conditions that f is square-integrable and $0 < \sigma_1^2 < \infty$, U_n admits an asymptotic normal distribution with asymptotic variance $k^2\sigma_1^2/n$. However, the exact U-statistic variance given in (1.2) is complicated in form, and the asymptotic variance of U_n is not necessarily reliable when the kernel size k is not small compared to the sample size n.

Wang and Lindsay (2014) propose an unbiased variance estimator, denoted as $\widehat{V}_u$. It is of an elementary quadratic form and easy to realize with the help of a partition resampling scheme proposed in Wang and Lindsay (2014). Let (S_a, S_b) represent a pair of data subsets, each of size k. Denote $|S_a \cap S_b|$ as the number of overlaps between S_a and S_b. Provided that $k \leq n/2$, the unbiased variance estimator $\widehat{V}_u$ can be written as

$$\widehat{V}_u = U_n^2 - Q(0), \text{ where} \tag{1.3}$$
$$Q(0) = \left[\binom{n}{k} \binom{n-k}{k} \right]^{-1} \sum_{|S_a \cap S_b| = 0} \phi(S_a)\phi(S_b).$$

Since $E[U_n^2 - Q(0)] = E(U_n^2) - [E(U_n)]^2$, the unbiasedness of $\widehat{V}_u$ follows. Maesono (1998) compares several U-statistic variance estimators, including an unbiased estimator. Maesono's method is based on finding unbiased estimates for the σ_c^2 terms in (1.2). Let $\sum_{(n,j)}$ denote a summation taken over all subsets of size j ($1 \leq j \leq n$). Following the notations in Maesono (1998), it can be shown that

$$Q(0) = \binom{n}{2k}^{-1} \sum_{(n,2k)} \zeta_0(X_1, \ldots, X_{2k}),$$

$$U_n^2 = \binom{n}{k}^{-1} \sum_{c=0}^{k} \binom{k}{c} \binom{n-k}{k-c} \widehat{\lambda}_k,$$

where $\widehat{\lambda}_k = \binom{n}{2k-c}^{-1} \sum_{(n,2k-c)} \zeta_k(x_{i_1}, \ldots, x_{i_{2k-c}})$, and

$$\zeta_c(x_1, \ldots, x_{2k-c}) = \binom{2k-c}{c}^{-1} \binom{2k-2c}{k-c}^{-1} \sum_{|S_a \cap S_b| = c} \phi(S_a)\phi(S_b) \text{ for } 0 \le c \le k.$$

As a result, Maesono's unbiased variance estimator is equivalent to $\widehat{V}_u$. However, the proposal of Wang and Lindsay (2014) is of a much simpler form.

The construction of the $Q(0)$ term in $\widehat{V}_u$ requires that $k \le n/2$. This condition restricts its application to common K-fold cross-validation problems. For more on cross-validation, see Picard and Cook (1984), Shao (1993), and Hastie, Tibshirani and Friedman (2009). Ray and Lindsay (2008) and Wang and Lindsay (2014) show that the unbiased estimator for the Kullback-Leibler risk of a parametric model can be written as a U-statistic with kernel size $k = m + 1$, where m is the training sample size. Thus, the U-statistic estimator for the K-fold cross-validated risk has kernel size $k = n(K-1)/K + 1$, bigger than $n/2$. In this case the unbiased variance estimator $\widehat{V}_u$ is no longer applicable. Bengio and Grandvalet (2004) argue that there is no universal unbiased estimator for the variance of K-fold cross-validation. This paper aims to develop reliable and efficient estimators for the variance of a U-statistic with relatively large kernel size k. We will focus the discussion on applications of K-fold cross-validation under Kullback-Leibler loss function.

We first notice that the kernel function of the U-statistic estimator for the Kullback-Leibler risk of a parametric model can be approximated asymptotically by a kernel function of two components. We define a pseudo-kernel function as the "fake" kernel derived from U_n when assuming the degree of U_n is two. Using the degree-two pseudo-kernel function, one can construct a variance estimator of U_n based on equation (1.3). We call it a pseudo-kernel variance estimator. When U_n is non-degenerate, i.e. $0 < \sigma_1^2 < \infty$, the proposed pseudo-kernel variance estimator is second-order unbiased (Theorem 1) and consistent (Theorem 2). We also notice that the well-known delete-one jackknife variance estimator can be expressed in the form of an unbiased variance estimator using a pseudo-kernel of degree one (Remark 3). Efron and Stein (1981) and Wang and Chen (2015)

show that the conventional jackknife variance estimator is first-order unbiased. Therefore, the proposed pseudo-kernel variance estimator is more accurate than the conventional jackknife method.

The rest of the paper is organized as follows: In Section 2 we propose a pseudo-kernel method in variance estimation. We show how one can construct a pseudo-kernel function of degree two based on a given data set and also discuss the construction of pseudo-kernels with higher degree. We prove some theoretical properties of the resulting pseudo-kernel variance estimator. In Section 3 we demonstrate an efficient realization of the proposal in the context of ten-fold cross-validation. Then, we compare the performance of the pseudo-kernel variance estimator with its bootstrap and jackknife counterparts using simulations and a real data set.

2. Pseudo-Kernel Method in U-statistic Variance Estimation

2.1. Asymptotic kernel for the U-statistic Kullback-Leibler risk estimator

Consider a family of parametric models $\mathcal{M} = \{m_\theta \mid \theta \in \Theta\}$ with unknown parameter $\theta \in \Theta \subseteq \mathcal{R}^p$ ($p \in \mathcal{Z}^+$). Denote the Maximum Likelihood (ML) estimator of the parameter as $\widehat{\theta}(\mathcal{X}_m)$, where $\mathcal{X}_m = (X_1, \ldots, X_m)$ is a training sample of size m taken out of $\mathcal{X}_n$. Write the log-likelihood function of the fitted model as $\log f_{\widehat{\theta}(\mathcal{X}_m)}(x)$. Wang and Lindsay (2014) show that the unbiased estimator for the Kullback-Leibler risk in model selection can be written as

$$U_n = \binom{n}{m+1}^{-1} \sum_{(n,m+1)} \phi(\mathcal{X}_{m+1}),$$

$$\phi(\mathcal{X}_{m+1}) = -(m+1)^{-1} \sum_{i=1}^{m+1} \log f_{\widehat{\theta}(\mathcal{X}_m^{(-i)})}(X_i), \tag{2.1}$$

where the summation $\sum_{(n,m+1)}$ is taken over all subsets of size $m+1$ out of $\mathcal{X}_n$, $\mathcal{X}_m^{(-i)}$ represents a training sample of size m without X_i, and $f_{\widehat{\theta}(\mathcal{X}_m^{(-i)})}(X_i)$ is the estimated probability density function evaluated at point X_i. Because the parameter θ is estimated using data subset $\mathcal{X}_m^{(-i)}$ ($1 \le i \le m+1$), the degree of the kernel function $\phi(\mathcal{X}_{m+1})$ (2.1) is $k = m+1$. Therefore, unless the training sample size $m < n/2$ one cannot apply the unbiased variance estimator $\widehat{V}_u$ directly. Below we show that the kernel function $\phi(\mathcal{X}_{m+1})$ (2.1) with possibly large degree has an approximate symmetric kernel function with two components.

Consider an i.i.d. sample of size n, $X_1, \ldots, X_n$, from some parametric distri-

bution f_θ ($\theta \in \Theta \subseteq \mathcal{R}^p,\ p \in \mathcal{Z}^+$). Let θ^* be the true value of the parameter in Θ, and let $\widehat{\theta}(\mathcal{X}_m)$ be the Maximum Likelihood estimator. Assume the regularity conditions C1–C5 (Appendix A1) hold such that $\widehat{\theta}(\mathcal{X}_m)$ is consistent for θ^*. With a Taylor series expansion, the kernel function $\phi(\mathcal{X}_{m+1})$ (2.1) can be approximated asymptotically by a symmetric function of two components. The error of the approximation is of order $o(\|\widehat{\theta}(\mathcal{X}_m) - \theta^*\|)$ where $\|\cdot\|$ represents the Euclidean norm. We call the approximate symmetric kernel function an asymptotic kernel; it is defined as

$$\phi^*(x_i, x_j) = \frac{1}{4}\left[\log f_{\theta^*}(x_i) + \log f_{\theta^*}(x_j)\right] + u(x_j|\theta^*)^T \mathbf{I}(\theta^*)^{-1} u(x_i|\theta^*), \qquad (2.2)$$

where $u(x_i|\theta^*) = \partial \log f_\theta(x_i)/\partial\theta \mid_{\theta=\theta^*}$, and $\mathbf{I}(\theta^*)$ is the Fisher Information matrix

$$\mathbf{I}(\theta^*) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right] \mid_{\theta=\theta^*}.$$

When $\widehat{\theta}(\mathcal{X}_m)$ is $\sqrt{m}$-consistent, the error of the approximation is of order $O_p(1/\sqrt{m})$. (A detailed derivation of the asymptotic kernel in (2.2) can be found in the supplementary materials.)

There exists an alternative approximation to the kernel function $\phi(\mathcal{X}_{m+1})$ (2.1) that leads to an interesting interpretation of the risk estimator, akin to the generalized AIC (Wang and Lindsay (2014)).

Let $u(x_i|\theta) = \partial \log/\partial\theta f_\theta(x_i) = \partial l/\partial\theta(\theta|x_i)$. We have

$$\begin{aligned}\phi(\mathcal{X}_{m+1}) = &\frac{1}{m+1}\sum_{i=1}^{m+1}\log f_{\theta^*}(X_i)\\ &+ \frac{1}{m(m+1)}\sum_{i\neq j} u(X_j|\theta^*)^T \mathbf{I}(\theta^*)^{-1} u(X_i|\theta^*) + o(\|\widehat{\theta}(\mathcal{X}_m) - \theta^*\|).\end{aligned}$$

Given an i.i.d. data subset $\mathcal{X}_{m+1}$ of size $m+1$, denote the ML estimator based on $\mathcal{X}_{m+1}$ as $\widehat{\theta}_{m+1} := \widehat{\theta}(\mathcal{X}_{m+1})$. For large enough m, we have

$$\begin{aligned}\phi(\mathcal{X}_{m+1}) \approx &\frac{1}{m+1}\sum_{i=1}^{m+1}\log f_{\widehat{\theta}_{m+1}}(X_i)\\ &+ \frac{1}{m(m+1)}\sum_{i\neq j} u(X_j|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_i|\widehat{\theta}_{m+1}).\end{aligned}$$

Because $\sum_{i=1}^{m+1} u(X_i|\widehat{\theta}_{m+1}) = \sum_{i=1}^{m+1} l'(\widehat{\theta}_{m+1}|X_i) = 0$, we have

$$\left[\sum_{i=1}^{m+1} u(X_i|\widehat{\theta}_{m+1})^T\right] \mathbf{I}(\widehat{\theta}_{m+1})^{-1} \left[\sum_{j=1}^{m+1} u(X_j|\widehat{\theta}_{m+1})\right]$$

$$= \sum_{i=1}^{m+1} \sum_{j=1}^{m+1} u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_j|\widehat{\theta}_{m+1}) = 0,$$

which yields

$$\sum_{i \neq j} u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_j|\widehat{\theta}_{m+1})$$
$$= -\sum_{i=1}^{m+1} u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_i|\widehat{\theta}_{m+1}).$$

Thus,

$$\phi(\mathcal{X}_{m+1}) \approx \frac{1}{m+1} l(\widehat{\theta}_{m+1}) - \frac{1}{m(m+1)} \sum_{i=1}^{m+1} u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_i|\widehat{\theta}_{m+1}),$$

where $l(\widehat{\theta}_{m+1}) = \sum_{i=1}^{m+1} l(\widehat{\theta}_{m+1}|X_i)$ is the joint log-likelihood.

Because $X_1, \ldots X_{m+1}$ are i.i.d., by the Law of Large Numbers,

$$\frac{1}{m+1} \sum_{i=1}^{m+1} \left[u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_i|\widehat{\theta}_{m+1}) \right]$$
$$\xrightarrow{a.s.} E[u(X_1 \mid \widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_1 \mid \widehat{\theta}_{m+1})]$$

as m goes to infinity. Since

$$E[u(X_1 \mid \widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_1 \mid \widehat{\theta}_{m+1})]$$
$$= E\{\text{trace}[u(X_1 \mid \widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_1 \mid \widehat{\theta}_{m+1})]\}$$
$$= \text{trace}\{E[\mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_1 \mid \widehat{\theta}_{m+1}) u(X_1 \mid \widehat{\theta}_{m+1})^T]\}$$
$$= p,$$

we have $(m+1)^{-1} \sum_{i=1}^{m+1} u(X_i|\widehat{\theta}_{m+1})^T \mathbf{I}(\widehat{\theta}_{m+1})^{-1} u(X_i|\widehat{\theta}_{m+1}) \to p$ as $m \to \infty$. Thus, for large enough m $\phi(\mathcal{X}_{m+1}) \approx c[l(\widehat{\theta}_{m+1}) - p]$, where p is the dimension of parameter θ and c is a fixed constant. Namely, $\phi(\mathcal{X}_{m+1})$ is asymptotically proportional to $-2[l(\widehat{\theta}_{m+1}) - p]$.

Remark 1. The symmetric kernel function $\phi(\mathcal{X}_{m+1})$ in (2.1) is asymptotically equivalent to the generalized AIC evaluated at subsample size $m+1$. Therefore, a U-statistic risk estimator, defined as an average of $\phi(\mathcal{X}_{m+1})$ over n-choose-$(m+1)$ subsamples of size $m+1$, is approximately an average of the AIC scores, each computed based on a subsample of size $m+1$. When comparing AIC at size $m+1$ with AIC at size n, the difference comes from the log-likelihood term, the source of bias incurred by using subsample size $m+1 < n$.

2.2. Construction of pseudo-kernel of degree two

Equation (2.2) gives us an explicit expression of the asymptotic kernel function for the U-statistic estimator of a Kullback-Leibler risk. However, it is not directly applicable without knowing the true parameter θ^*. In the following we give the definition of a pseudo-kernel function, a "fake" kernel function of lower degree. We propose to use the pseudo-kernel function to construct a U-statistic variance estimator, applicable even when the kernel size k of U_n is large.

Definition 1. *Consider a U-statistic U_n, defined on a symmetric kernel function ϕ of degree k ($k \geq 2$). The function $\phi_{PS}(x_i, x_j)$ is a degree-two pseudo-kernel function if*

$$\phi_{PS}(x_i, x_j) = \binom{n}{2} U_n - \binom{n-1}{2} U_{n-1}^{(-i)} - \binom{n-1}{2} U_{n-1}^{(-j)} + \binom{n-2}{2} U_{n-2}^{(-i,-j)}, \quad (2.3)$$

where $U_{n-1}^{(-i)}$ is a U-statistic computed based on a data subset of size $n-1$ without the ith observation, and $U_{n-2}^{(-i,-j)}$ is a U-statistic computed based on a data subset of size $n-2$ excluding both the ith and the jth observations.

The motivation for the definition of $\phi_{\mathrm{PS}}(x_i, x_j)$ is as follows: Assume that U_n has an asymptotic kernel function of degree two, such as in the case of Kullback-Leibler risk estimation shown in Subsection 2.1, and ignore the errors in the asymptotic approximation. Then, for any i, j ($1 \leq i < j \leq n$),

$$\begin{aligned}
\binom{n}{2} U_n &= \sum_{1 \leq k < l \leq n} \phi_{\mathrm{PS}}(x_k, x_l), \\
\binom{n-1}{2} U_{n-1}^{(-i)} &= \sum_{1 \leq k < l \leq n} \phi_{\mathrm{PS}}(x_k, x_l) - \sum_{l \neq i} \phi_{\mathrm{PS}}(x_i, x_l), \\
\binom{n-1}{2} U_{n-1}^{(-j)} &= \sum_{1 \leq k < l \leq n} \phi_{\mathrm{PS}}(x_k, x_l) - \sum_{k \neq j} \phi_{\mathrm{PS}}(x_k, x_j), \\
\binom{n-2}{2} U_{n-2}^{(-i,-j)} &= \sum_{1 \leq k < l \leq n} \phi_{\mathrm{PS}}(x_k, x_l) - \sum_{l \neq i} \phi_{\mathrm{PS}}(x_i, x_l) - \sum_{k \neq j} \phi_{\mathrm{PS}}(x_k, x_j) \\
&\quad + \phi_{\mathrm{PS}}(x_i, x_j).
\end{aligned}$$

It follows that $\phi_{\mathrm{PS}}(x_i, x_j)$ has the form (2.3).

Using the pseudo-kernel function ϕ_{PS} of degree two, one can follow the construction of an unbiased variance estimator in (1.3) to obtain a variance estimator for U_n. We call the resulting estimator a *pseudo-kernel variance estimator*, defined as

$$\widehat{V}_{\mathrm{PS}} = {U_{\mathrm{PS}}}^2 - \left[\binom{n}{2}\binom{n-2}{2}\right]^{-1} \sum \phi_{\mathrm{PS}}(X_{i_1}, X_{i_2})\phi_{\mathrm{PS}}(X_{j_1}, X_{j_2}), \tag{2.4}$$

where the summation is taken over all pairs of non-overlapped subsamples of size two, and $U_{\mathrm{PS}} = \sum_{1\le i<j\le n} \phi_{\mathrm{PS}}(X_i, X_j)/\{n!/[2!(n-2)!]\}$.

Although $\widehat{V}_{\mathrm{PS}}$ is constructed based on ϕ_{PS}, not the true kernel, we can show that, when U_n is non-degenerate, the pseudo-kernel variance estimator $\widehat{V}_{\mathrm{PS}}$ (2.4) is always second-order unbiased. This property makes the proposed estimator preferable to the jackknife variance estimator. For a proof of Theorem 1, see Appendix A2.

Theorem 1. *Consider a non-degenerate U-statistic U_n of degree k ($k \ge 2$). The pseudo-kernel variance estimator $\widehat{V}_{PS}$ defined in* (2.4) *is second-order unbiased.*

Besides the second-order unbiasedness of the devised pseudo-kernel variance estimator, it can be shown that $\widehat{V}_{\mathrm{PS}}$ is also a consistent estimator for the true variance $\mathrm{Var}(U_n)$. The consistency property is stated in Theorem 2 and the proof can be found in the supplementary materials.

Theorem 2. *Let U_n be a non-degenerate U-statistic with degree k ($k \ge 2$). The pseudo-kernel variance estimator $\widehat{V}_{PS}$ defined in* (2.4) *satisfies for any $\epsilon > 0$,*

$$P(\mid \widehat{V}_{PS} - \mathit{Var}(U_n) \mid > \epsilon) \to 0 \ \mathit{as}\ n \to \infty.$$

Remark 2. If the degree of a U-statistic is exactly two ($k = 2$), the pseudo-kernel method yields the unbiased variance estimator $\widehat{V}_u$.

Remark 3. As with the the degree-two pseudo-kernel, one can define a degree-one pseudo-kernel as

$$\phi_{\mathrm{PS}}(x_i) = \binom{n}{2} U_n - \binom{n-1}{2} U_{n-1}^{(-i)}.$$

The conventional leave-one-out jackknife variance estimator can be written in the form of $\widehat{V}_u$ based on a pseudo-kernel function of degree one. (For proof, see the supplementary materials.)

Remark 4. The asymptotic kernel $\phi^*(x_1, x_2)$, (2.2), for the kernel function $\phi(\mathcal{X}_{m+1})$, (2.1), of a U-statistic risk estimator is independent of the training sample size m. Therefore, in the context of K-fold cross-validation the pseudo-kernel function can be realized by deleting an entire block of observations rather than removing one observation at a time. We investigate this idea further in Section 3 to simplify the computational cost of the proposed variance estimator.

2.3. Pseudo-Kernel of higher degree

Following the construction of a pseudo-kernel function of degree two, one can define a pseudo-kernel of higher degree in a similar fashion. For instance, a pseudo-kernel function of degree three can be written as

$$\begin{aligned}\phi_{\mathrm{PS}}(x_{i_1}, x_{i_2}, x_{i_3}) = &\binom{n}{3} U_n - \binom{n-1}{3}\left(U_{n-1}^{(-i_1)} + U_{n-1}^{(-i_2)} + U_{n-1}^{(-i_3)}\right) \\ &+ \binom{n-2}{3}\left(U_{n-2}^{(-i_1,-i_2)} + U_{n-2}^{(-i_1,-i_3)} + U_{n-2}^{(-i_2,-i_3)}\right) \\ &- \binom{n-3}{3} U_{n-3}^{(-i_1,-i_2,-i_3)},\end{aligned} \tag{2.5}$$

where $U_n, U_{n-1}^{(-i_1)}$, and $U_{n-2}^{(-i_1,-i_2)}$ $(1 \le i_1 < i_2 \le n)$ are defined in the same way as in Definition 1, and $U_{n-3}^{(-i_1,-i_2,-i_3)}$ is a U-statistic computed based on a data subset of size $n-3$, excluding the i_1th, the i_2th, and the i_3th observations from the data set.

Using a pseudo-kernel function of degree three, (2.5), the resulting pseudo-kernel variance estimator $\widehat{V}_{\mathrm{PS}}$ is

$$\widehat{V}_{\mathrm{PS}} = U_{\mathrm{PS}}^2 - \left[\binom{n}{3}\binom{n-3}{3}\right]^{-1} \sum \phi_{\mathrm{PS}}(X_{i_1}, X_{i_2}, X_{i,3}) \phi_{\mathrm{PS}}(X_{j_1}, X_{j_2}, X_{j_3}),$$

where the summation is taken over all pairs of non-overlapped data subsets of size three, and $U_{\mathrm{PS}} = \binom{n}{3}^{-1} \sum_{1 \le i_1 < i_2 < i_3 \le n} \phi_{\mathrm{PS}}(x_{i_1}, x_{i_2}, x_{i_3})$.

Using a degree-three pseudo-kernel, the pseudo-kernel variance estimator is third-order unbiased (Theorem 3). The proof for Theorem 3 is straightforward but involves tedious manipulations of the orthogonal terms in Hoeffding decomposition (Hoeffding (1948)). A sketch of the proof can be found in the supplementary materials.

Theorem 3. *Consider a non-degenerate U-statistic U_n of degree k $(k \ge 3)$. Using a degree-three pseudo-kernel function as defined in* (2.5)*, the resulting variance estimator $\widehat{V}_{PS}$ is third-order unbiased.*

Hypothetically one can define a pseudo-kernel function of any degree k^* $(2 \le k^* \le k)$ in an adaptive fashion. As the degree of the pseudo-kernel function increases, the bias of the resulting pseudo-kernel variance estimator decreases. The improvement in bias is of order n^{-k^*} $(k^* \ge 2)$, which may be negligible for large n. Given the marginal gain in accuracy and the incurred computational cost by using $k^* \ge 3$, we do not pursue the implementation of pseudo-kernel functions of higher degree. The degree-two pseudo-kernel function would mostly

yield satisfactory result in practice, as shown later in Section 3, and we focus on the discussion of the pseudo-kernel variance estimator with $k^* = 2$.

3. Pseudo-Kernel Variance Estimator in Ten-Fold Cross-Validation

Ten-fold cross-validation is one of the most widely used algorithms in statistical practice and machine learning (Kohavi (1995)). We now demonstrate how to realize the pseudo-kernel variance estimator $\widehat{V}_{\mathrm{PS}}$ (2.4) with high computational efficiency in the case of ten-fold cross-validation.

3.1. An efficient realization

Without loss of generality, assume n is a multiple of $K = 10$. We partition the data set into ten blocks, each of size $\widetilde{n} = n/10$, say $S_1, \ldots, S_{10}$. Write the observations in subsample S_l as $x_{l,t}$ $(1 \le l \le 10; 1 \le t \le \widetilde{n})$. Let L be a loss function used to evaluate the closeness between an independent observation $x_{l,t}$ and its prediction $\widehat{x}_{l,t}^{(-S_l)}$, where $\widehat{x}_{l,t}^{(-S_l)}$ is obtained by fitting the model excluding the lth block of observations. The conventional ten-fold cross-validated risk estimator can be expressed as

$$U_{\mathrm{CV}} = \frac{1}{10}\sum_{l=1}^{10}\frac{1}{\widetilde{n}}\sum_{x_{l,t}\in S_l} L(\widehat{x}_{l,t}^{(-S_l)}, x_{l,t}),$$

and the remove-one-block and remove-two-block cross-validated risk estimators can be written as

$$U_{\mathrm{CV}}^{(-S_i)} = \frac{1}{9}\sum_{l\neq i}\frac{1}{\widetilde{n}}\sum_{x_{l,t}\in S_l} L(\widehat{x}_{l,t}^{(-S_i,-S_l)}, x_{l,t}),$$

$$U_{\mathrm{CV}}^{(-S_i,-S_j)} = \frac{1}{8}\sum_{l\neq i \text{ and } l\neq j}\frac{1}{\widetilde{n}}\sum_{x_{l,t}\in S_l} L(\widehat{x}_{l,t}^{(-S_i,-S_j,-S_l)}, x_{l,t}).$$

Here $\widehat{x}_{l,t}^{(-S_i,-S_l)}$ represents the predicted value for $x_{l,t} \in S_l$ when the model is fitted without S_i and S_l $(1 \le i \le 10;\ 1 \le l \le 10;\ l \neq i)$, and $\widehat{x}_{l,t}^{(-S_i,-S_j,-S_l)}$ is the predicted value for $x_{l,t} \in S_l$ when the model is fitted without $S_i, S_j,$ and S_l $(1 \le i < j \le 10;\ 1 \le l \le 10;\ l \neq i \text{ or } j)$. Simply put, $U_{\mathrm{CV}}^{(-S_i)}$ is a nine-fold CV risk estimator after removing S_i, and $U_{\mathrm{CV}}^{(-S_i,-S_j)}$ is an eight-fold CV risk estimator after removing both S_i and S_j from the data.

Under Kullback-Leibler distance, the U-statistic risk estimator has an asymptotic kernel function of degree two. Then,

$$U_{\mathrm{CV}} \approx \binom{n}{2}^{-1} \sum_{1\le s<t\le n} \phi_{\mathrm{PS}}(x_s, x_t), \tag{3.1}$$

$$U_{\mathrm{CV}}^{(-S_i)} \approx \binom{0.9n}{2}^{-1} \sum_{x_s, x_t \notin S_i} \phi_{\mathrm{PS}}(x_s, x_t), \tag{3.2}$$

$$U_{\mathrm{CV}}^{(-S_j)} \approx \binom{0.9n}{2}^{-1} \sum_{x_s, x_t \notin S_j} \phi_{\mathrm{PS}}(x_s, x_t), \tag{3.3}$$

$$U_{\mathrm{CV}}^{(-S_i,-S_j)} \approx \binom{0.8n}{2}^{-1} \sum_{x_s, x_t \notin S_i \cup S_j} \phi_{\mathrm{PS}}(x_s, x_t), \tag{3.4}$$

where the summation $\sum_{x_s,x_t\notin S_i}$ is taken over $1 \le s < t \le n$ and $x_s, x_t \notin S_i$ for $1 \le i \le 10$; the summation $\sum_{x_s,x_t\notin S_i\cup S_j}$ is taken over $1 \le s < t \le n$ and $x_s, x_t \notin S_i \cup S_j$ for $1 \le i < j \le 10$.

Take

$$\widetilde{\phi}(i,j) = \frac{1}{\widetilde{n}^2}\left[\binom{n}{2}U_{\mathrm{CV}} - \binom{0.9n}{2}U_{\mathrm{CV}}^{(-S_i)} - \binom{0.9n}{2}U_{\mathrm{CV}}^{(-S_j)} + \binom{0.8n}{2}U_{\mathrm{CV}}^{(-S_i,-S_j)}\right].$$

According to (3.1) to (3.4), we have

$$\widetilde{\phi}(i,j) \approx \frac{1}{\widetilde{n}^2} \sum_{x_{i,s}\in S_i} \sum_{x_{j,t}\in S_j} \phi_{\mathrm{PS}}(x_{i,s}, x_{j,t}).$$

It is easy to see that the average of $\widetilde{\phi}(i,j)$ over $1 \le i < j \le 10$ is an approximation for U_{CV}. In addition, $\widetilde{\phi}(i,j)\widetilde{\phi}(l,k)$ with $(i,j)\cap(l,k) = \emptyset$ is an average of $\phi_{\mathrm{PS}}(x_{i,s}, x_{j,t})\phi_{\mathrm{PS}}(x_{l,s}, x_{k,t})$ terms, where $x_{i,s} \in S_i, x_{j,t} \in S_j, x_{l,s} \in S_l$, and $x_{k,t} \in S_k$ $(1 \le i < j < 10; 1 \le l < k < 10; 1 \le s, t \le \widetilde{n})$. Thus, the average of $\widetilde{\phi}(i,j)\widetilde{\phi}(l,k)$ over $(i,j)\cap(l,k) = \emptyset$ can be used to approximate the $Q(0)$ term in $\widehat{V}_u$ (1.3) directly. Moreover, the pseudo-kernel variance estimator at (2.4) can be approximated by

$$\widetilde{V}_{\mathrm{PS}} := {U_{CV}^*}^2 - \left[\binom{10}{2}\binom{8}{2}\right]^{-1} \sum_{(i,j)\cap(s,t)=\emptyset} \widetilde{\phi}(i,j)\widetilde{\phi}(s,t), \tag{3.5}$$

where $U_{\mathrm{CV}}^* = \binom{10}{2}^{-1}\sum_{1\le i<j\le 10}\widetilde{\phi}(i,j)$.

Fix the number of folds K in a cross-validation algorithm. Although the number of observations in each fold increases with a larger sample size n, the number of terms being averaged over in (3.5) depends only on K. Therefore, the computational cost for realizing $\widetilde{V}_{\mathrm{PS}}$ (3.5) does not grow substantially with greater n, whereas the computational cost of its bootstrap and jackknife counterparts can increase dramatically with larger n.

3.2. A simulation study in ten-fold cross-validation

We consider a simulation study of regression analysis using ten-fold cross-validation. The methodology can be easily generalized to other cross-validation scenarios. We compare the performance of the pseudo-kernel variance estimator $\widetilde{V}_{\mathrm{PS}}$ (3.5) with the nonparametric bootstrap and the leave-one-out jackknife variance estimators. For more on bootstrap and jackknife, see Efron and Stein (1981); Efron and Gong (1983); Efron and Tibshirani (1983), and Efron (1987).

Consider a linear regression model between response Y_i and seven continuous predictors $X_{i,j}$ $(1 \le i \le n; 1 \le j \le 7)$. The true regression relationship is assumed to be

$$Y_i = 1 + 15X_{i,1} + 8X_{i,2} + 5X_{i,3} + 3X_{i,4} + 1X_{i,5} + 0.01X_{i,6} + 0X_{i,7} + \epsilon_i \ (1 \le i \le n).$$

To simulate the data sets, we generated predictors X_i's, each of size n, independently from Uniform(0,1), and standardized each predictor to have zero mean and unit standard deviation. We simulated random errors of size n from Normal(0, $\sqrt{0.1}$), and obtained the values for the response variable Y. We repeated this process $R = 1,000$ times to get 1,000 independent data sets. For each sample of size n, we performed ten-fold cross-validation based on Kullback-Leibler distance to evaluate the ordinary least-square model fit. In this case, the U-statistic estimator of the risk can be written as

$$U_{\mathrm{CV}} = -\frac{1}{10}\sum_{l=1}^{10}\frac{1}{\widetilde{n}}\sum_{(x_t,y_t)\in S_l}\log f_{\widehat{\beta}^{(-S_l)}}(y_t|x_t), \tag{3.6}$$

which is proportional to

$$\frac{1}{10}\sum_{l=1}^{10}\frac{1}{\widetilde{n}}\sum_{(x_t,y_t)\in S_l}(y_t - x_t\widehat{\beta}^{(-S_l)})^2,$$

where $x_t = (1, x_{t,1}, \ldots, x_{t,7})^T$ and $\beta = (\beta_0, \ldots, \beta_7)^T$. Here $f_{\widehat{\beta}^{(-S_l)}}$ is the estimated density function of response Y_t when the model is fitted without the lth block of observations. As shown in equation (3.6), with normal random errors the ten-fold CV estimator for the Kullback-Leibler risk is proportional to the risk estimator based on L^2 distance. The goal here is to assess the variation of U_{CV} at (3.6).

We focused on the full linear regression model, including all seven predictors, and compared the efficient pseudo-kernel variance estimator $\widetilde{V}_{\mathrm{PS}}$ at (3.5), the nonparametric bootstrap variance estimator with 500 bootstrap samples for each data set, and the leave-one-out jackknife variance estimator. Table 1 summarizes the average estimated variance, the simulated standard deviation, and the mean

Table 1. Performance of different variance estimators in ten-fold CV.

$n = 100$	True	Pseudo	Bootstrap	Jackknife
$\widehat{E}\{\mathrm{Var}(U_{\mathrm{CV}})\}$	0.00547	0.00480	0.00648	0.00747
$\widehat{SD}\{\mathrm{Var}(U_{\mathrm{CV}})\}$		0.00357	0.00184	0.00237
MSE$\times 10^3$		0.01319	0.00441	0.00962
Time (hours)		1.67	6.23	2.35
$n = 200$	True	Pseudo	Bootstrap	Jackknife
$\widehat{E}\{\mathrm{Var}(U_{\mathrm{CV}})\}$	0.00261	0.00237	0.00279	0.00302
$\widehat{SD}\{\mathrm{Var}(U_{\mathrm{CV}})\}$		0.00138	0.00053	0.00059
MSE$\times 10^3$		0.00196	0.00031	0.00052
Time (hours)		1.87	6.80	2.79
$n = 500$	True	Pseudo	Bootstrap	Jackknife
$\widehat{E}\{\mathrm{Var}(U_{\mathrm{CV}})\}$	0.00105	0.00103	0.00104	0.00108
$\widehat{SD}\{\mathrm{Var}(U_{\mathrm{CV}})\}$		0.00052	0.00014	0.00012
MSE$\times 10^3$		0.00027	0.00020	0.00015
Time (hours)		3.62	16.59	13.53

squared error for each estimator. The computation time given in Table 1 is the total time in hours spent to compute $R = 1,000$ variance estimates. The true variance was approximated based on a simulation with 50,000 random samples.

Table 1 shows that the pseudo-kernel method is indeed an efficient way to estimate the variation of a cross-validation score, especially for large sample size n. When the sample size n is small, the pseudo-kernel variance estimator is less biased but may be more variable than its bootstrap and jackknife counterparts. In this case, the computational advantage of using $\widetilde{V}_{\mathrm{PS}}$ is not as significant. When the sample size n is relatively large, the proposal yields comparable performance compared to the bootstrap and jackknife variance estimators, but with much less computational cost. The bootstrap and jackknife variance estimators become very expensive to compute for large sample size n. In addition, the jackknife variance estimator always shows large positive bias in regression analysis, as also noted in Wu (1986) and Hinkley (1977).

Next, we considered how different the decision of model selection would be if we used different methods to estimate the variance of U_{CV}. Since the seven x-variables were independently generated from the same distribution, and standardized, we ranked the significance of the predictors by the magnitudes of their corresponding coefficients. We compared seven models, with Model 7 being the full model and Model 1 being the single-predictor model, as shown in Table 2. We only show results for sample size $n = 200$ (results with other sample sizes are very similar).

Table 2. Models under comparison.

Model	Predictors						
Model 1	X_1						
Model 2	X_1	X_2					
Model 3	X_1	X_2	X_3				
Model 4	X_1	X_2	X_3	X_4			
Model 5	X_1	X_2	X_3	X_4	X_5		
Model 6	X_1	X_2	X_3	X_4	X_5	X_6	
Model 7	X_1	X_2	X_3	X_4	X_5	X_6	X_7

If one simply selects the model with the smallest ten-fold CV risk score, then out of 1,000 random samples one would select Model 5 on 742 occasions, Model 6 on 178, and Model 7 on 80. We investigated what would happen if one implemented the "one-standard-error" (1-SE) rule (Hastie, Tibshirani and Friedman (2009)), selecting the most parsimonious model whose CV risk estimate is within one standard error of the minimum CV risk estimate. Here all three methods chose Model 5 all the time. Although Model 6 was the true model, the coefficient of predictor X_6 was very close to 0 and therefore the risk of Model 5 was not significantly larger than that of Model 6. Again, the pseudo-kernel variance estimator is much more efficient to compute.

Remark 5. As noted in Wang and Lindsay (2014), one can conduct pairwise model comparisons and evaluate the variance of the difference between two risk scores. This proposal can be applied to estimating the variance of the difference in risk scores.

Remark 6. The unbiased variance estimator of U_n proposed in Wang and Lindsay (2014) may yield negative values, but we did not encounter any negative variance estimates when computing $\widetilde{V}_{\mathrm{PS}}$ in this example. Should this occur, one can consider the simple fix-ups in Wang and Lindsay (2014), or can implement the extrapolation techniques discussed in Wang and Chen (2015).

3.3. Data analysis

Consider a heart disease data set available from the UCI Machine Learning Repository. We focused on the response of whether or not a patient has heart disease and the thirteen predictor variables. The data set first appeared in Aha and Kibler (1988) and has been analyzed in Chai et al. (2004); Huang et al. (2004), and Zhou and Jiang (2012).

After removing observations with missing values, there are 296 observations.

Table 3. Models under comparison. Predictors marked as "X" are included in the corresponding model.

p	thal	exang	ca	slope	cp	sex	trestbps	thalach	chol	fbs	restecg	oldpeak	age
1	X												
2	X	X											
3	X	X	X										
4	X	X	X	X									
5	X	X	X	X	X								
6	X	X	X	X	X	X							
7	X	X	X	X	X	X	X						
8	X	X	X	X	X	X	X	X					
9	X	X	X	X	X	X	X	X	X				
10	X	X	X	X	X	X	X	X	X	X			
11	X	X	X	X	X	X	X	X	X	X	X		
12	X	X	X	X	X	X	X	X	X	X	X	X	
13	X	X	X	X	X	X	X	X	X	X	X	X	X

For convenience of partitioning the data we focused on a subset of 290 of them, and considered a logistic regression model to determine the classification of each patient. Here the ten-fold CV risk estimator based on Kullback-Leibler distance is

$$U_{\mathrm{CV}} = -\frac{1}{10}\sum_{l=1}^{10}\frac{1}{\widetilde{n}}\sum_{(x_t,y_t)\in S_l}[y_t \log \widehat{p}_t^{(-S_l)} + (1-y_t)\log(1-\widehat{p}_t^{(-S_l)})],$$

where $\widehat{p}_t^{(-S_l)}$ is the predicted probability of heart disease based on a logistic regression model fitted without the lth block of observations. Although there are 2^{13} possible models, we only considered the best model of each size in the BIC sense. Thus, we worked with 13 models for model comparison, as shown explicitly in Table 3.

We computed the ten-fold cross-validation score of each candidate model. Figure 1 indicates that the full model has the smallest estimated Kullback-Leibler risk. If one does not take into account of the variation of the risk estimator, Model 13 would have been selected as the best model. In practice, it is clear that Model 13 is not necessarily the optimal choice, as some more parsimonious models, such as Models 5 to 12, have CV scores very similar to that of Model 13. We implemented the 1-SE rule to see whether a more parsimonious model would be selected instead of the full model. We used the same set of methods to estimate the standard error of the ten-fold CV risk score.

From Figure 2 it can be seen that, based on the 1-SE rule, all three variance estimation methods lead to the same conclusion of favoring Model 4, as its CV

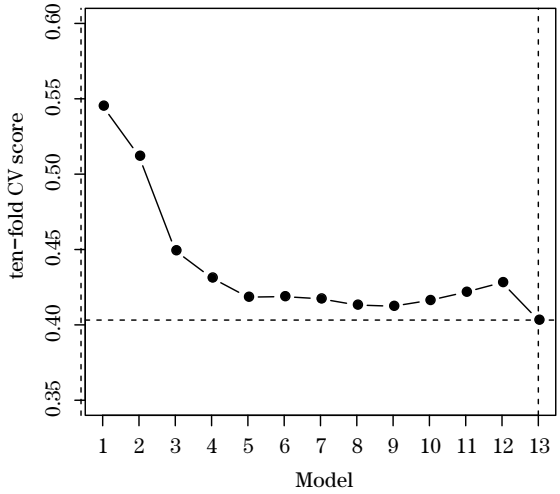

Figure 1. Ten-fold cross-validation scores for the candidate models.

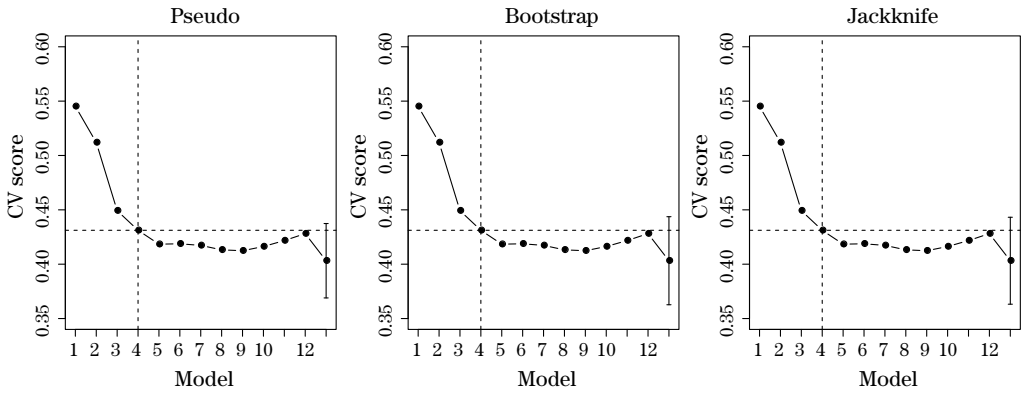

Figure 2. Implementation of 1-SE rule, where the standard error was estimated by different methods.

risk score is within one standard error of the minimum. Again, the pseudo-kernel variance estimator is much cheaper to compute compared to its bootstrap and jackknife counterparts.

Acknowledgment

We would like to thank the Editor, an associate editor, and two referees for their valuable comments and suggestions that lead to a much improved version of the original manuscript. Part of the work of the first author was done when she was at Williams College and when she was at Bentley University.

Appendix

A1. Regularity conditions

Under the following regularity conditions C1–C5 (Lehmann (2004)), the Maximum Likelihood (ML) estimator $\widehat{\theta}$ of θ^* is consistent, i.e. $\widehat{\theta} \xrightarrow{P} \theta^*$ as $m \to \infty$.

C1. The parameter space Θ is open and bounded.

C2. The probability density function f_θ is identifiable, i.e.

$$\theta \neq \theta_0 \text{ implies } f_\theta(x) \neq f_{\theta_0}(x).$$

C3. The probability density function $f_\theta(x)$ is continuous in x.

C4. The set $A = \{x : f_\theta(x) > 0\}$ is independent of θ.

C5. For all x in A, $f_\theta(x)$ is differentiable with respect to θ.

A2. Proof of Theorem 1

Proof. Consider the orthogonal terms in Hoeffding decomposition (Hoeffding (1948))

$$h^{(c)}(x_1, \ldots, x_c) = \phi_c(x_1, \ldots, x_c) - \sum_{j=1}^{c} \sum_{(c,j)} h^{(j)}(x_{i_1}, \ldots, x_{x_j}) - \theta,$$

where $\theta = E\left[\phi(X_1, \ldots, X_k)\right]$ and

$$\phi_c(x_1, \ldots, x_c) = E\left[\phi(X_1, \ldots, X_k) | X_1 = x_1, \ldots, X_c = x_c\right]$$

for $1 \leq c \leq k$.

We have

$$U_n = \theta + \binom{n}{k}^{-1} \sum_{j=1}^{k} \binom{n-j}{k-j} \sum_{(n,j)} h^{(j)}(x_{\nu_1}, \ldots, x_{\nu_j}),$$

$$U_{n-1}^{(-s)} = \theta + \binom{n-1}{k}^{-1} \sum_{j=1}^{k} \binom{n-1-j}{k-j} \sum_{(n-1^{(-s)},j)} h^{(j)}(x_{\nu_1}, \ldots, x_{\nu_j}),$$

$$U_{n-2}^{(-s,-t)} = \theta + \binom{n-2}{k}^{-1} \sum_{j=1}^{k} \binom{n-2-j}{k-j} \sum_{(n-2^{(-s,-t)},j)} h^{(j)}(x_{\nu_1}, \ldots, x_{\nu_j}).$$

The notation $\sum_{(n,j)}$ indicates that the summation is taken over all subsets of size j taken out of $\mathcal{X}_n$, and $\sum_{(n-1^{(-s)},j)}$ sums over all subsets of size j taken out of $\mathcal{X}_{n-1}^{(-s)}$ where $\mathcal{X}_{n-1}^{(-s)}$ is a data subset of size $n-1$ excluding X_s, and $\sum_{(n^{(-s,-t)},j)}$

sums over all subsets of size j taken out of $\mathcal{X}_{n-2}^{(-s,-t)}$ where $\mathcal{X}_{n-2}^{(-s,-t)}$ is data subset of size $n-2$ excluding X_s and X_t.

The pseudo-kernel of degree two can be expressed as

$$\begin{aligned}
\phi_{\mathrm{PS}}(x_s,x_t) &= \binom{n}{2}U_n - \binom{n-1}{2}U_{n-1}^{(-s)} - \binom{n-1}{2}U_{n-1}^{(-t)} + \binom{n-2}{2}U_{n-2}^{(-s,-t)} \\
&= \binom{n}{2}\left(\theta + \binom{n}{k}^{-1}\sum_{j=1}^{k}\binom{n-j}{k-j}\sum_{(n,j)} h^{(j)}(x_{\nu_1},\ldots,x_{\nu_j})\right) \\
&\quad - \binom{n-1}{2}\left(\theta + \binom{n-1}{k}^{-1}\sum_{j=1}^{k}\binom{n-1-j}{k-j}\sum_{(n-1^{(-s)},j)} h^{(j)}(x_{\nu_1},\ldots,x_{\nu_j})\right) \\
&\quad - \binom{n-1}{2}\left(\theta + \binom{n-1}{k}^{-1}\sum_{j=1}^{k}\binom{n-1-j}{k-j}\sum_{(n-1^{(-t)},j)} h^{(j)}(x_{\nu_1},\ldots,x_{\nu_j})\right) \\
&\quad + \binom{n-2}{2}\left(\theta + \binom{n-2}{k}^{-1}\sum_{j=1}^{k}\binom{n-2-j}{k-j}\sum_{(n-2^{(-s,-t)},j)} h^{(j)}(x_{\nu_1},\ldots,x_{\nu_j})\right).
\end{aligned}$$

Here the terms related to θ can be simplified to θ; the terms related to $h^{(1)}$ can be simplified to $(k/2)\left[h^{(1)}(x_s)+h^{(1)}(x_t)\right]$; the terms related to $h^{(2)}$ are

$$\begin{aligned}
&\binom{n}{2}\binom{n}{k}^{-1}\binom{n-2}{k-2}\sum_{1\le i<j\le n} h^{(2)}(x_i,x_j) \\
&- \binom{n-1}{2}\binom{n-1}{k}^{-1}\binom{n-3}{k-2}\left(\sum_{1\le i<j\le n} h^{(2)}(x_i,x_j) - \sum_{1\le i<s} h^{(2)}(x_i,x_s) - \sum_{s<j\le n} h^{(2)}(x_s,x_j)\right) \\
&- \binom{n-1}{2}\binom{n-1}{k}^{-1}\binom{n-3}{k-2}\left(\sum_{1\le i<j\le n} h^{(2)}(x_i,x_j) - \sum_{1\le i<t} h^{(2)}(x_i,x_t) - \sum_{t<j\le n} h^{(2)}(x_t,x_j)\right) \\
&+ \binom{n-2}{2}\binom{n-2}{k}^{-1}\binom{n-4}{k-2}\left(\sum_{1\le i<j\le n} h^{(2)}(x_i,x_j) - \sum_{1\le i<t} h^{(2)}(x_i,x_t) - \sum_{t<j\le n} h^{(2)}(x_t,x_j)\right) \\
&- \binom{n-2}{2}\binom{n-2}{k}^{-1}\binom{n-4}{k-2}\left(\sum_{1\le i<s} h^{(2)}(x_i,x_s) + \sum_{s<j\le n} h^{(2)}(x_s,x_j) - h^{(2)}(x_s,x_t)\right),
\end{aligned}$$

which can be simplified to $[k(k-1)/2]h^{(2)}(x_s,x_t)$.

Therefore, for $1\le s<t\le n$

$$\phi_{\mathrm{PS}}(x_s,x_t) = \theta + (k/2)\left(h^{(1)}(x_s)+h^{(1)}(x_t)\right) + [k(k-1)/2]h^{(2)}(x_s,x_t) + \text{remainder}.$$

The remainder only depends on $h^{(c)}$ with $3\le c\le k$. The leading term in the

remainder is, with $c = 3$,

$$\frac{k(k-1)(k-2)}{2(n-2)} \sum_{i \neq s \text{ and } i \neq t} h^{(3)}(x_s, x_t, x_i) + O\left(\frac{k(k-1)(k-2)}{n^2} \sum_{i \neq s \text{ and } i \neq t} h^{(3)}(x_s, x_t, x_i) \right).$$

The other terms in the remainder are of lower orders and can be expressed in terms of $h^{(j)}$ $(4 \leq j \leq k)$ in a similar way.

As a result,

$$U_{\mathrm{PS}} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \phi_{\mathrm{PS}}(x_i, x_j)$$
$$= \theta + \frac{k}{n} \sum_{i=1}^{n} h^{(1)}(x_i) + \binom{n}{k}^{-1} \binom{n-2}{k-2} \sum_{(n,2)} h^{(2)}(x_i, x_j) + \text{remainder},$$

and it is easy to show that the variance of the remainder terms is of order n^{-3}.

Let $\delta_c^2 = \mathrm{Var}(h^{(c)})$ for $1 \leq c \leq k$. The U-statistic variance can be expressed as (Hoeffding (1948); Lee (1990)):

$$\mathrm{Var}(U_n) = \sum_{c=1}^{k} \binom{k}{c}^2 \binom{n}{c}^{-1} \delta_c^2.$$

If one follows the construction of the unbiased variance estimator (1.3) but changes the original kernel function to the pseudo-kernel function, one would get a second-order unbiased variance estimator. That is, the coefficients of the δ_1^2 and δ_2^2 terms in $E(\widehat{V}_{\mathrm{PS}})$ match those in $\mathrm{Var}(U_n)$. Therefore, $\widehat{V}_{\mathrm{PS}}$ is a second-order unbiased estimator for $\mathrm{Var}(U_n)$.

References

Aha, D. and Kibler, D. (1988). Instance-based prediction of heart-disease presence with the cleveland database. *Technical report, University of California.*

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

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* **5**, 1089-1105.

Chai, X., Deng, L., Yang, Q. and Ling, C. X. (2004). Test-cost sensitive naive bayes classification. *Proceedings of the Fourth IEEE International Conference on Data Mining*, 51-58.

Efron, B. (1987). Bootstrap methods: another look at the jackknife. *Journal of the American Statistical Association* **82**, 171-185.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**, 36-48.

Efron, B. and Stein, C. (1981). The jackknife estimation of variance. *The Annals of Statistics* **9**, 586-596.

Efron, B. and Tibshirani, R. (1983). *An Introduction to Bootstrap*. Chapman&Hall/CRC.

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

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19**, 285-292.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19**, 293-325.

Huang, K., Yang, H., King, I., Lyu, M. R. and Chan, L. (2004). Biased minimax probability machine for medical diagnosis. In the *Eighth International Symposium on Artificial Intelligence and Mathematics*.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence* **2**, 1137-1143.

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

Lehmann, E. L. (2004). *Elements of Large-Sample Theory*. Springer.

Maesono, Y. (1998). Asymptotic comparisons of several variance estimators and their effects for studentizations. *Annals of the Institute of Statistical Mathematics* **50**, 451-470.

Picard, R. and Cook, D. (1984). Cross-validation for regression models. *Journal of the American Statistical Association* **79**, 575-583.

Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *Journal of the Royal Statistical Society-Series B* **70**, 95-118.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-494.

Wang, Q. (2012). Investigation of topics in U-statistics and their applications in risk estimation and cross-validation (doctoral dissertation). *The Pennsylvania State University*.

Wang, Q. and Chen, S. (2015). A general class of linearly extrapolated variance estimators. *Statistics & Probability Letters* **98**, 29-38.

Wang, Q. and Lindsay, B. G. (2014). Variance estimation of a general U-statistic with application to cross-validation. *Statistica Sinica* **24**, 1117-1141.

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

Zhou, Z. and Jiang, Y. (2012). Nec4.5: Neural ensemble based c4.5. *IEEE Transactions on Knowledge and Data Engineering* **16**, 770.

Department of Mathematics, Wellesley College, 106 Central Street, Wellesley, MA 02481, USA

E-mail: qwang@wellesley.edu

Department of Statistics, The Pennsylvania State University, University Park, PA 16801, USA

In memoriam of Dr. Bruce Lindsay.

(Received December 2015; accepted May 2016)