Statistics and Probability Letters 98 (2015) 29–38

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A general class of linearly extrapolated variance estimators

Qing Wang *, Shiwen Chen

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

ARTICLE INFO

Article history:
Received 10 June 2014
Received in revised form 8 December 2014
Accepted 11 December 2014
Available online 18 December 2014

Keywords:
ANOVA decomposition
Jackknife
Hoeffding decomposition
Linear extrapolation
Variance estimation
U-statistic

ABSTRACT

A general class of linearly extrapolated variance estimators was developed as an extension of the conventional leave-one-out jackknife variance estimator. In the context of U-statistic variance estimation, the proposed variance estimator is first-order unbiased. After showing the equivalence between the Hoeffding decomposition (Hoeffding, 1948) and the ANOVA decomposition (Efron and Stein, 1981), we study the bias property of the proposed variance estimator in comparison to the conventional jackknife method. Simulation studies indicate that the proposal has comparable performance to the jackknife method when assessing the variance of the sample variance in various distributions. An application to half-sampling cross-validation indicates that the proposal is more computationally efficient and shows better performance than its jackknife counterpart in the context of regression analysis.

1. Introduction

Variance measures the uncertainty of a random variable. Therefore, variance estimation is crucial in evaluating the performance of a point estimator or a statistical methodology. Nowadays, one of the commonly used variance estimation techniques is the leave-one-out jackknife variance estimator (Quenouille, 1949; Tukey, 1958). Denote the parameter of interest as θ. Given an i.i.d. sample of size n, $X_1, \ldots, X_n$, the jackknife variance estimator for statistic $\hat{\theta} = T(X_1, \ldots, X_n)$ is defined as

$$\hat{V}_J = \frac{n-1}{n} \sum_{i=1}^{n} (T_{n-1}^{-i} - \bar{T}_{n-1})^2, \tag{1.1}$$

where $T_{n-1}^{-i} = T(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ and $\bar{T}_{n-1} = n^{-1} \sum_{i=1}^{n} T_{n-1}^{-i}$.

Efron and Stein (1981) consider the jackknife variance estimator as a linearly extrapolated estimator: one first constructs a variance estimator at subsample size $n-1$ using $\sum_{i=1}^{n} (T_{n-1}^{-i} - \bar{T}_{n-1})^2$, and then extrapolates it from $n-1$ to the original sample size n by multiplying $(n-1)/n$. Following the footsteps of Efron and Stein (1981), we consider an extension of the conventional jackknife methodology. The main contribution of this paper is the proposal of a general class of linearly extrapolated variance estimators and the investigation of its bias property. We also demonstrate an application of the proposed variance estimator in half-sampling cross-validation.

The new methodology can be summarized as follows: we first devise a variance estimator at subsample size m ($m < n$) and then extrapolate it from m to n using linear approximation. In the context of U-statistic variance estimation, an unbiased variance estimator at size m can be obtained as long as the kernel size $k \leq m \leq n/2$ (see Section 2). Then, the bias in the linearly extrapolated variance estimator can be formally evaluated. We prove in Section 3 the equivalence between the Hoeffding decomposition (Hoeffding, 1948) and the ANOVA decomposition (Efron and Stein, 1981), which facilitates the

* Correspondence to: 18 Hoxsey Street, Williamstown, MA 01267, USA. Tel.: +1 413 597 4960.
E-mail address: qww1@williams.edu (Q. Wang).

http://dx.doi.org/10.1016/j.spl.2014.12.011
0167-7152/

comparison between the proposed variance estimator and the leave-one-out jackknife variance estimator. We show in Section 3 that both estimators are first-order unbiased, and their biases can be expressed explicitly. We demonstrate the performance of the proposal in comparison to the jackknife method in two simulation studies in Section 4. The proposed variance estimator shows comparable performance to jackknife method in a simulation study that assesses the variance of the unbiased sample variance. The proposal seems to outperform the jackknife estimator with high computational efficiency in the context of half-sampling cross-validation. It can be seen in Section 4.2 that the flexibility of choosing subsample size m ($m \leq n/2$) in the proposed variance estimator leads to efficient realization of the cross-validation algorithm. In the end, we conclude our paper with some final remarks and discussions.

2. Linearly extrapolated variance estimator

Given an i.i.d. sample of size n, a U-statistic (Hoeffding, 1948) is defined as

$$U_n = \binom{n}{k}^{-1} \sum_{1\leq i_1<\cdots<i_k\leq n} \phi(X_{i_1}, \ldots, X_{i_k}), \tag{2.1}$$

where $\phi(x_1, \ldots, x_k)$ is a symmetric kernel function with k components. It is an unbiased estimator for the parameter $\theta = E\{\phi(X_1, \ldots, X_k)\}$. As most unbiased point estimators can be written as a U-statistic, throughout this paper we will focus on the problem of U-statistic variance estimation. However, the proposed variance estimator can be easily generalized to other statistics that do not have a U-statistic representation.

Hoeffding (1948) derives the closed-form expression of the variance of a U-statistic. However, calculating the exact variance is computational expensive, especially when both n and k are large. Moreover, the asymptotic variance of a general U-statistic (see Theorem 7.1 in Hoeffding, 1948) is not necessarily reliable when the kernel size k is not negligible compared to the sample size n. In this paper, we propose a linearly extrapolated variance estimator that is easy to construct and is applicable as long as $k \leq n/2$. In addition, the proposed variance estimator is first-order unbiased in the context of U-statistic variance estimation, which makes it a valuable competitor of the jackknife variance estimator.

Consider a U-statistic defined in (2.1). For any $k \leq m$, let U_m be the U-statistic computed based on a subsample of size m, say $\mathcal{X}_m = (X_1, \ldots, X_m)$. Denote

$$U_m = U_m(X_1, \ldots, X_m) = \binom{m}{k}^{-1} \sum_{1\leq i_1<\cdots<i_k\leq m} \phi(X_{i_1}, \ldots, X_{i_k}).$$

Theorem 1. *Let U_n be a U-statistic based on a symmetric kernel ϕ of size k, where $k \leq m \leq n/2$. Given an i.i.d. sample of size n, let S_m and S_m^* be mutually exclusive subsamples of size m. Define*

$$\hat{V}_m = \left\{\binom{n}{m}\binom{n-m}{m}\right\}^{-1} \sum_{(S_m, S_m^*)\subseteq \mathcal{X}_n} \frac{\{U_m(S_m) - U_m(S_m^*)\}^2}{2}.$$

Then, $\hat{V}_m$ is an unbiased estimator of $Var(U_m)$. The linearly extrapolated variance estimator of U_n can be expressed as

$$\hat{V}_{ex} = \frac{m}{n}\hat{V}_m \tag{2.2}$$

which is a first-order unbiased estimator for $Var(U_n)$.

Proof. Because S_m and S_m^* are nonoverlapping data subsets of size m,

$$E(\hat{V}_m) = (1/2)E[\{U_m(S_m) - U_m(S_m^*)\}^2] = E(U_m^2) - \{E(U_m)\}^2 = \mathrm{Var}(U_m).$$

The unbiasedness of $\hat{V}_m$ follows.

To show the first-order unbiasedness of $\hat{V}_{\mathrm{ex}}$, notice that

$$E(\hat{V}_{\mathrm{ex}}) = \frac{m}{n}E(\hat{V}_m) = \frac{m}{n}\mathrm{Var}(U_m).$$

Based on the exact formula of a U-statistic variance (Hoeffding, 1948), we have

$$\mathrm{Var}(U_m) = \sum_{j=1}^{k} \binom{k}{j}^2 \binom{m}{j}^{-1} \delta_j^2,$$

where δ_j^2 is the variance of the jth orthogonal term in Hoeffding decomposition (Hoeffding, 1948; Lee, 1990).

$$E(\hat{V}_{\mathrm{ex}}) = \frac{m}{n}\sum_{j=1}^{k} \binom{k}{j}^2 \binom{m}{j}^{-1} \delta_j^2 = \frac{k^2}{n}\delta_1^2 + \frac{m}{n}\sum_{j=2}^{k} \binom{k}{j}^2 \binom{m}{j}^{-1} \delta_j^2.$$

Therefore, $\hat{V}_{\mathrm{ex}}$ is a first-order unbiased estimator for $\mathrm{Var}(U_n)$. □