Computational Statistics and Data Analysis 89 (2015) 51–71

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Improving cross-validated bandwidth selection using subsampling-extrapolation techniques

Qing Wang [a,*], Bruce G. Lindsay [b]

[a] Department of Mathematics and Statistics, Williams College, Williamstown, MA, USA

[b] Department of Statistics, Pennsylvania State University, University Park, PA, USA

HIGHLIGHTS

- A two-stage subsampling-extrapolation bandwidth selection procedure is proposed.
- An automatic nested cross-validation method is developed to select the subsample size.
- The extrapolated bandwidth selectors achieve a smaller mean square error.
- The second-order extrapolated bandwidth selector has a relative convergence rate $n^{-1/4}$.

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form 19 February 2015
Accepted 3 March 2015
Available online 16 March 2015

Keywords:
Bandwidth selection
Cross-validation
Extrapolation
L^2 distance
Nonparametric kernel density estimator
Subsampling

ABSTRACT

Cross-validation methodologies have been widely used as a means of selecting tuning parameters in nonparametric statistical problems. In this paper we focus on a new method for improving the reliability of cross-validation. We implement this method in the context of the kernel density estimator, where one needs to select the bandwidth parameter so as to minimize L^2 risk. This method is a two-stage subsampling-extrapolation bandwidth selection procedure, which is realized by first evaluating the risk at a fictional sample size m ($m \leq$ sample size n) and then extrapolating the optimal bandwidth from m to n. This two-stage method can dramatically reduce the variability of the conventional unbiased cross-validation bandwidth selector. This simple first-order extrapolation estimator is equivalent to the rescaled "bagging-CV" bandwidth selector in Hall and Robinson (2009) if one sets the bootstrap size equal to the fictional sample size. However, our simplified expression for the risk estimator enables us to compute the aggregated risk without any bootstrapping. Furthermore, we developed a second-order extrapolation technique as an extension designed to improve the approximation of the true optimal bandwidth. To select the optimal choice of the fictional size m given a sample of size n, we propose a nested cross-validation methodology. Based on simulation study, the proposed new methods show promising performance across a wide selection of distributions. In addition, we also investigated the asymptotic properties of the proposed bandwidth selectors.

1. Introduction

Cross-validation methodology has long been a popular method for selecting tuning parameters in non and semiparametric models. However, it has also been criticized for its high variability and its corresponding tendency to overfit the data.

* 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.csda.2015.03.005

This paper develops new methods for the improvement of the conventional cross-validation procedures. It is based on a blending of U-statistic estimation and asymptotic theory. These new methods are realized by estimating the cross-validation risk with small training sets, then extrapolating the results to the desired sample size. The extrapolation step requires some asymptotic theory, but only the rate of convergence, not any unknown constants. We will show that such a two-stage procedure can dramatically reduce the high variability and overfitting that is the major liability of the conventional unbiased cross-validation.

We view our results as part of the following paradigm: when one is estimating nonparametrically a statistical property of samples of target size m, such as the risk inherent in using a particular model, then one can do a much more accurate estimation when the target m is much smaller than the actual sample size n. The intuition is that there are many, many more subsamples of size $n/2$, say, than there are subsamples of size n or $n-1$.

To motivate our extrapolation methodology, we will here show how it works when used in risk estimation in the context of nonparametric kernel density estimation. In the process we will also show that for this problem the risk function for arbitrary m is surprisingly simple. In particular, cross-validation estimation at an arbitrary training sample size of m does not require repeated subsampling at size m, thereby greatly speeding up and improving accuracy of the methods we propose. We believe this to be a major new insight in the kernel density estimation literature.

To simplify notation, consider a univariate random variable $X \in \mathcal{R}$. In statistical practice, we often know little about the underlying distribution of X which is crucial in exploratory or inferential analysis (Silverman, 1986). So, our main task is to estimate the unknown density function $f(x)$ based on a set of observations. In this paper, we focus on the nonparametric kernel density estimator (Fix and Hodges, 1951). Given an i.i.d. sample of size n, $\mathcal{X}_n = (X_1, \ldots, X_n)$, the kernel density estimator at x is defined for a kernel K as

$$\hat{f}_h(x \mid \mathcal{X}_n) = n^{-1} \sum_{i=1}^{n} K_h(X_i - x) \quad (x \in \mathcal{R}), \tag{1.1}$$

where $h > 0$ is called the bandwidth parameter. Here $K_h(t) = h^{-1}K(t/h)$ and function K is the kernel function. As the choice of K does not greatly affect the density estimation (Hardle et al., 1994), throughout this paper we consider a commonly used location kernel function, the Gaussian kernel.

$$K_h(x - x_0) = (h\sqrt{2\pi})^{-1} e^{-(x-x_0)^2/(2h^2)} \sim N(x_0, h^2). \tag{1.2}$$

However, our proposed methodologies do not depend on the choice of K, and the theoretical results in this paper will be stated in terms of an arbitrary symmetric kernel function K of order r ($r \geq 2$). For the definition of the order of a kernel function, please see Turlach (1993).

Although one has free choice of the kernel function in a density estimator, the choice of the bandwidth h is generally viewed as much more crucial. In order to select the optimal smoothing parameter h, we need to evaluate how closely $\hat{f}_h$ can approximate f for a given data set. Most bandwidth selectors are based on first choosing a risk function that measures the error made in using a particular bandwidth h. One can then estimate the risk function for a given data set and choose the bandwidth that minimizes the empirical risk. Such bandwidth selectors are referred to as data-driven methods.

The main result of this paper is to propose a two-stage subsampling-extrapolation bandwidth selection procedure. This work is closely related to the rescaled bagging cross-validation method of Hall and Robinson (2009) and the partitioned cross-validation method of Marron (1987). Recent work involving bagging and subsampling in problems other than kernel density estimation includes Meinshausen and Buhlmann (2010) and Shah and Samworth (2012). Unlike the bandwidth selectors discussed in Park and Marron (1990) and Sheather and Jones (1991), which are based on asymptotic theory, our proposed methodology is a hybrid of the cross-validation method and the asymptotic theory. As such it does not require the estimation of $R(f'')$ or a third-stage estimation of $R(f''')$. (By convention, we denote $R(g) = \int g^2(x)dx$ for any given function g.) Hence, it is more straightforward to implement than plug-in estimators. Most importantly, it can be used in a wide variety of problems where plug-in methodology is not available.

We present an extensive simulation study in Section 4.1 to compare the proposed methods with the conventional cross-validation estimator. It will be seen that our bandwidth selectors achieve a smaller expected integrated square error that is much closer to the theoretical optimum than the standard cross-validation. Moreover, a comparison of the proposed methods to indirect cross-validation (Savchuk et al., 2011, 2010; Mammen et al., 2012) can be found in Section 4.2. In addition, we compare our methods to the asymptotic selection of the subsample size m that was described in Marron (1987).

2. U-statistic estimate of L^2 risk

In this section, we will derive a simple U-statistic form estimator for the risk that arises from L^2 distance. It is a new representation for the unbiased risk estimator and enables us to calculate the aggregated risk at subsamples of size m ($m \leq n$) much more efficiently than the repeated bootstrapping done in Hall and Robinson (2009) or the partitioning method used in Marron (1987).