6th St.Petersburg Workshop on Simulation (2009) 1-3

A Note on Data-Adaptive Bandwidth Selection for Sequential Kernel Smoothers

Ansgar Steland[1]

Abstract

Sequential kernel smoothers form a class of procedures covering various known methods for the problem of detecting a change in the mean as special cases. In applications, one often aims at estimation, prediction and detection of changes. We propose to use sequential kernel smoothers and study a sequential cross-validation algorithm to choose the bandwidth parameter assuming that observations arrive sequentially at equidistant time instants. An uniform weak law of large number and a consistency result for the cross-validated bandwidth is discussed.

1 Introduction

Let us assume that observations $Y_n = Y_{Tn}$, $1 \leq n \leq T$, T the maximum sample size, arrive sequentially and satisfy the model equation

$$Y_n = m(n/T) + \epsilon_n, \qquad n = 1, 2, \ldots, T,\ T \geq 1,$$

for some bounded and piecewise continuous function $m : [0, \infty) \to \mathbb{R}$. The errors $\{\epsilon_n : n \in \mathbb{N}\}$ form a sequence of i.i.d. random variables such that

$$E(\epsilon_n) = 0, \qquad E(\epsilon_1^4) < \infty. \tag{1}$$

Consequently, $m(t)$, $t \in [0, 1]$, models the process mean during the relevant time frame $[0, T]$. In practice, an analysis has often to solve three problems. (i) Estimation of the current process mean. (ii) One-step prediction of the process mean. (iii) Signaling when there is evidence that the process mean differs from an assumed (null) model. Usually, different statistics are used for these problems. To ease interpretation and applicability, we will base the detector on the same statistic used for estimation and prediction. Our reasoning is that a method which fits the data well and has convincing prediction properties should also possess reasonable detection properties for a large class of alternatives models.

We confine ourselves to closed end procedures where monitoring stops at a (usually large) time horizon T. The proposed kernel smoother is controlled by a bandwidth parameter which controls the degree of smoothing. As well known,

[1]RWTH Aachen University, E-mail: steland@stochastik.rwth-aachen.de

its selection is crucial, particularly for estimation and prediction accuracy. We propose to select the bandwidth sequentially by minimizing a sequential version of the cross-validation criterion. The topic has been quite extensively studied in the literature assuming the classic regression estimation framework where the data gets dense as the sample size increases. A comprehensive monograph of the general methodology is [7]. For references to the literature on estimation of regression functions that are smooth except some discontinuity (change-) points see the recent work of [2].

Before proceeding, let us discuss our assumptions on m. Often the information about the problem of interest is not sufficient to setup a (semi-) parametric model for the process mean m and the distribution of the error terms, which would allow us to use methods based on, e.g., likelihood ratios. In this paper it is only assumed that

$$m \in \mathrm{Lip}, \quad m(t) > 0, \qquad t > 0, \qquad \text{and } \|m\|_\infty < \infty \ , \tag{2}$$

where Lip denotes the class of Lipschitz continuous functions. Under these general conditions, one should use detectors which avoid (semi-) parametric specifications about the shape of m, and nonparametric smoothers $\widehat{m}_n$ which estimate some monotone functional of the process mean and are sensitive with respect to changes of the mean. For these reasons, we confine our study to detectors of the form

$$S_T = \inf\{\lfloor s_0 T \rfloor \le t \le T : \widehat{m}_t > c\}.$$

Here c is a threshold (control limit), $s_0 \in (0,1)$ determines through $\lfloor T s_0 \rfloor$ the start of monitoring, $\lfloor x \rfloor$ denoting the integer part of x, and $\{\widehat{m}_n : n \in \mathbb{N}\}$ is a sequence of $\sigma(Y_1, \ldots, Y_n)$-measureable statistics. Specifically, we study the following sequential kernel smoother

$$\widetilde{m}_n = \widetilde{m}_{n,h} = \frac{1}{h} \sum_{i=1}^{n} K([i-n]/h) Y_i, \qquad n = 1, 2, \ldots$$

and the associated normed version

$$\widehat{m}_n = \widehat{m}_{n,h} = \widetilde{m}_h / \frac{1}{h} \sum_{j=1}^{n} K([i-n]/h),$$

respectively, which are related to the classic Nadaraya-Watson estimator. It is worth noting that various classic control chart statistics are obtained as special cases. Denoting the target value by μ_0, the CUSUM chart is based on $C_n = \sum_{i=1}^{n} [X_i - (\mu_0 + K)]$ where $\{X_n\}$ denotes the observed process and K is the reference value. This chart corresponds to the choice $K(z) = 1$ if $Y_n = X_n - (\mu_0 + K)$ for all n. The EWMA recursion, $\widehat{m}_n = \lambda Y_n + (1-\lambda)\widehat{m}_{n-1}$ with starting value $\widehat{m}_0 = Y_0$, $\lambda \in (0,1)$ a smoothing parameter, corresponds to the kernel $K(z) = e^{|z|}$ and the bandwidth $h = 1/\log(1-\lambda)$.

Our assumptions on the smoothing kernel are as follows.

$$K \in \mathrm{Lip}(\mathbb{R}; [0,\infty)),\ \|K\|_\infty < \infty,\ \mathrm{supp}(K) \subset [-1,1], \text{ and } K > 0 \text{ on } (0,1). \tag{3}$$

For the bandwidth $h > 0$ we assume that

$$\lim_{T\to\infty} T/h = \xi \tag{4}$$

for some constant $\xi \in (0, \infty)$, which guarantees that in our design the number of observations on which $\widehat{m}_T$ depends converges to ∞, as $T \to \infty$. In practice, one can select ξ and put $h = T/\xi$.

In [4, 5, 6] procedures based on the sequential smoother $\widehat{m}_n$ are studied, which allow us to detect changes in the mean of a stationary or random walk series of observations. The asymptotic theory was studied as well. Specifically, in [5] it is shown that under the assumptions of the present paper the process $\{\sqrt{T}\widehat{m}_{\lfloor Ts\rfloor,h} : s \in [0,1]\}$ satisfies a functional central limit theorem when $m = 0$, i.e.,

$$\sqrt{T}\widehat{m}_{\lfloor Ts\rfloor,h} \Rightarrow \mathbb{M}(s),$$

for some centered Gaussian process $\{\mathbb{M}(s) : s \in [0,1]\}$ which depends on ξ.This result can be used to construct detection procedures with pre-specified statistical properties. E.g., when choosing the control limit such that the type I error rate satisfies $P(S_T \le T) = \alpha$ when $m = 0$ for some given significance level $\alpha \in (0,1)$, the control limit also depends on ξ. The question arises, how one can or should select the bandwidth $h \sim T$ and the parameter ξ, respectively.

In this paper we propose to select the bandwidth $h > 0$ such that the Y_t are well approximated by sequential predictions $\widehat{m}_t$ which are calculated from past data $Y_1, \ldots, Y_{t-1}$. For that purpose we propose a sequential version of the cross-validation criterion based on sequential leave-one-out estimates.

2 Sequential Cross-Validation

The idea of cross-validation is to choose parameters such that the corresponding estimates provide a good fit on average. To achieve this goal, one may consider the average squared distance between observations, Y_i, and predictions as an approximation of the integrated squared distance. To avoid over-fitting and interpolation, the prediction of Y_i is determined using the reduced sample where Y_i is omitted. Since, additionally, we aim at selecting the bandwidth h to obtain a good fit when using the sequential estimate, we consider

$$\widehat{m}_{h,-i} = N_{T,-i}^{-1}\frac{1}{h}\sum_{j=1}^{i-1} K([j-i]/h)Y_j, \qquad i = 2, 3, \ldots$$

with the constant $N_{T,-i} = h^{-1}\sum_{j=1}^{i-1} K([j-i]/h)$. Notice that $\widehat{m}_{h,-i}$ can be regarded as a sequential leave-one-out estimate. The corresponding *sequential leave-one-out cross-validation criterion* is defined as

$$CV_s(h) = \frac{1}{T}\sum_{i=2}^{\lfloor Ts\rfloor}(Y_i - \widehat{m}_{h,-i})^2, \qquad h > 0.$$

The cross-validation bandwidth at time s is now obtained by minimizing $CV_s(h)$ for fixed s. Notice that $CV_s(h)$ is a sequential unweighted version of the criterion studied by [3] in the classic regression function estimation framework. We do not consider a weighted CV sum, since we have in mind that the selected bandwidth is used to obtain a good fit for past and current observations. However, similar results as presented here can be obtained for the weighted criterion $T^{-1}\sum_{i=1}^{n} K([i-n]/h)(Y_i - \widehat{m}_{h,-i})^2$ as well. Notice that due to

$$CV_s(h) = \frac{1}{T}\sum_{i=1}^{\lfloor Ts \rfloor} Y_i^2 - \frac{2}{T}\sum_{i=2}^{\lfloor Ts \rfloor} Y_i \widehat{m}_{h,-i} + \frac{1}{T}\sum_{i=2}^{\lfloor Ts \rfloor} \widehat{m}_{h,-i}^2,$$

minimizing $CV_s(h)$ is equivalent to minimizing

$$C_{T,s}(h) = -\frac{2}{T}\sum_{i=2}^{\lfloor Ts \rfloor} Y_i \widehat{m}_{h,-i} + \frac{1}{T}\sum_{i=2}^{\lfloor Ts \rfloor} \widehat{m}_{h,-i}^2.$$

Thus, we will study $C_{T,s}(h)$ in the sequel. Cross-validation is expensive in terms of computational costs and minimizing $C_{T,s}$ for all s is not feasible in many case. Therefore and to simplify exposition, let us fix a finite number of time points

$$0 < s_1 < \cdots < s_N,$$

$N \in \mathbb{N}$. Later we shall relax this assumption and allow that N is an increasing function of T. At time s_i the cross-validation criterion is minimized to select the bandwidth, $h_i^* = h_i^*(Y_1, \ldots, Y_{s_i})$, and that bandwidth is used during the time interval $[s_i, s_{i+1})$, $i = 1, \ldots, N$.

3 Asymptotic Results

The question arises which function is estimated by $C_{T,s}(h)$. Our first result identifies the limit and shows convergence in mean.

Theorem 1. *We have*

$$E(C_{T,s}(h)) \to C_\xi(s) = -2\frac{\int_0^s \int_0^r \xi K(\xi(u-r)) m(\xi u)\, du dr}{\int_0^s \xi K(\xi(r-s))\, dr} + \frac{\int_0^s \xi^2 \int_0^r \int_0^r K(\xi(u-r)) K(\xi(v-r)) m(u) m(v)\, du\, dv\, dr}{\int_0^s \xi K(\xi(r-s))\, dr}, \tag{5}$$

as $T \to \infty$, uniformly in $s \in [s_0, 1]$.

Before proceeding, let us consider an example where the function $C_\xi(s)$ possesses a well-separated minimum.

Example 1. *Suppose K is given by $K(z) = (1-|z|)1_{[0,1]}(z)$ for $z \in \mathbb{R}$. Further, let us consider the nonlinear function $m(t) = x(x-0.2)(x-0.4)$. Clearly, $C_\xi(s)$ is a polynomial of order 4 with coefficients which depend on s. Figure 1 depicts $C_\xi(s)$ for some values of ξ. The locations of the (real) roots of $\frac{\partial}{\partial \xi} C_\xi(s)$ depend on $s \in [0,1]$ and are shown in Figure 1 as well.*

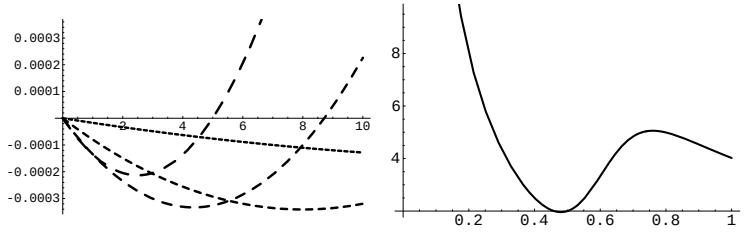

Figure 1: *Left panel:* The function $C_s(\xi)$, $\xi \in (0, 20]$, for $s \in \{0.1, 0.2, 0.3, 0.4\}$. *Right panel:* The optimal values for ξ as a function of $s \in (0, 1]$.

We will now study the uniform mean squared convergence of the random function $C_{T,s}(h)$. Define $\mathcal{S}_N = \{s_i : 1 \leq i \leq N\}$.

Theorem 2. *We have*

$$E \sup_{s \in \mathcal{S}_N} |C_{T,s}(h) - E(C_{T,s}(h))|^2 = O(T^{-1}),$$

as $T \to \infty$.

Current research focuses on the following generalization which allows that the number of time points where cross validation is conducted is a function of the maximum sample size T.

Conjecture 3. *Assume* $N = N_T$ *and*

$$0 < s_0 < s_{N1} < \cdots < s_{NN} \leq 1, \qquad N \geq 1, \tag{6}$$

and put $\mathcal{S}_N = \{s_{Ni} : 1 \leq i \leq N\}$. *Given the assumptions of Theorem 2 there exists some* $\gamma > 0$ *with* $\frac{N_T}{T^\gamma} = o(1)$, *such that*

$$E \sup_{s \in \mathcal{S}_N} |C_{T,s}(h) - E(C_{T,s}(h))|^2 = o(1)$$

Combining the above statements, we obtain

Theorem 4. *Suppose that* (1) *and* (6)

$$E \sup_{s \in \mathcal{S}_N} |C_{T,s}(h) - C_s(\xi))|^2 \to 0,$$

as $T \to \infty$.

We shall now extend the above results to study weak consistency of the cross-validation bandwidth under fairly general and weak assumptions. Having in mind the fact that $h \sim T$, let us simplify the setting by assuming that

$$h = h(\xi) = T/\xi, \qquad \xi \in [1, \Xi],$$

for some fixed $\Xi \in (1, \infty)$. This means, h and ξ are now equivalent parameters for each T. We also restrict the optimization to a compact interval, which is not restrictive for applications. Now $\widehat{m}_{h,-i}$ can be written as

$$\widehat{m}_{h,-i} = \frac{1}{(i-1)h} \sum_{j=1}^{i-1} K(\xi(j-i)/T) Y_j.$$

With some abuse of notation, let us also write

$$C_{T,s}(\xi) = C_{T,s}(T/\xi).$$

Theorem 5. *For any* $s \in [s_0, 1]$

$$\sup_{\xi \in [1,\Xi]} |C_{T,s}(\xi) - EC_{T,s}(\xi)| = o_P(1), \tag{7}$$

and

$$\sup_{\xi \in [1,\Xi]} |C_{T,s}(\xi) - C_{\xi}(s)| = o_P(1), \tag{8}$$

as $T \to \infty$.

We are now in a position to formulate the following conjecture on the asymptotic behavior of the the cross-validated sequential bandwidth selector.

Conjecture 6. *Suppose* $C_\xi(s)$ *possesses a well-separated minimum* $\xi^* \in [1, \Xi]$, *i.e.,*

$$\sup_{\xi \in [1,\Xi]: |\xi - \xi^*| \geq \varepsilon} C_\xi(s) > C_{\xi^*}(s).$$

for every $\varepsilon > 0$. *Then*

$$\text{argmin}_{\xi \in [1,\Xi]} C_{T,s}(\xi) \xrightarrow{P} \xi^*.$$

References

[1] Gijbels I., Goderniaux A.C. (2004) Bandwidth Selection for Changepoint Estimation in Nonparametric Regression. *Technometrics*, **46**, 1, 76-86. Oliver and Boyd, Edinburgh.

[2] Härdle W., Marron J.S. (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465-1481.

[3] Schmid W., Steland A. (2000) Sequential control of non-stationary processes by nonparametric kernel control charts. *AStA Adv. Stat. Anal.*, **84**, 315-336.

[4] Steland A. (2004) Sequential control of time series by functionals of kernel-weighted empirical processes under local alternatives. *Metrika*, **60**, 229-249.

[5] Steland A. (2005) Random walks with drift - A sequential approach. *J. Time Ser. Anal.*, **26** (6), 917-942.

[6] Wand M.P., Jones M.C. (1995) *Kernel Smoothing*. Chapman & Hall, Boca Raton.