Journal of Statistical Planning and Inference 165 (2015) 104–126

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Long-run variance estimation for spatial data under change-point alternatives

Béatrice Bucchia *, Christoph Heuser

Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Köln, Germany

ARTICLE INFO

Article history:
Received 1 August 2014
Received in revised form 8 January 2015
Accepted 4 April 2015
Available online 23 April 2015

MSC:
62H15
62E20
62M99
60G60
62H12

Keywords:
Long-run variance estimation
Change-point estimation
Change-point detection
Random fields

ABSTRACT

In this paper, we consider the problem of estimating the long-run variance (matrix) of an $\mathbb{R}^p$-valued multiparameter stochastic process $\{X_{\mathbf{k}}\}_{\mathbf{k}\in[1,n]^d}$, $(n, p, d \in \mathbb{N}, p, d$ fixed) whose mean-function has an abrupt jump. We consider processes of the form

$$X_{\mathbf{k}} = Y_{\mathbf{k}} + \mu + I_{C_n}(\mathbf{k})\Delta,$$

where I_C is the indicator function for a set C, the change-set $C_n \subset [1, n]^d$ is a finite union of rectangles and $\mu, \Delta \in \mathbb{R}^p$ are unknown parameters. The stochastic process $\{Y_{\mathbf{k}} : \mathbf{k} \in \mathbb{Z}^d\}$ is assumed to fulfill a weak invariance principle. Due to the non-constant mean, kernel-type long-run variance estimators using the arithmetic mean of the observations as a mean estimator have an unbounded error for changes Δ that do not vanish for $n \to \infty$. To reduce this effect, we use a mean estimator which is based on an estimation of the set C_n. In the case where $C_n = (\lfloor n\boldsymbol{\theta}_1^0 \rfloor, \lfloor n\boldsymbol{\theta}_2^0 \rfloor]$ is a rectangle, we introduce an estimator $\hat{C}_n = (\lfloor n\hat{\boldsymbol{\theta}}_1 \rfloor, \lfloor n\hat{\boldsymbol{\theta}}_2 \rfloor]$ and study its convergence rate.

1. Introduction

In this paper, we present and analyze a kernel-type long-run variance matrix (LRV in the following) estimator for a multivariate random field under the assumption of a non-constant mean. Such an estimator is needed e.g. in change-point analysis when one is interested in testing whether a given data-set is stationary or whether there is a jump in the mean, dividing the data into two sets with (different) constant means. In this case, the magnitude of the difference between the arithmetic means over suitable subsets of the data can be used as an indicator of the likelihood of a non-constant mean. The resulting tests are often based on the asymptotic behavior of the test statistic under the null hypothesis. For tests based on the partial sums of observations under suitable weak dependence conditions, a functional central limit theorem can be used to determine the distributional limit of the test statistic as a function of a multiparameter Brownian motion, and appropriate normalization can be used to standardize the limit process, leaving the LRV Σ as the only nuisance parameter. In order to construct asymptotic tests it is therefore important to estimate Σ consistently under the null hypothesis, so that the unknown LRV Σ may be replaced by its estimator for sufficiently large sample sizes. This has already been widely investigated for processes with constant mean functions, amongst others by Newey and West (1986) and Andrews (1991) for multivariate time series and later by Politis and Romano (1996), Robinson (2007) and Lavancier (2008) for univariate random fields. Most of the publications on the subject focus on the (null hypothesis) case of constant means to derive consistency of the LRV estimators. However, since the estimator for Σ is often used as a scaling factor in change-point tests, it is also

* Corresponding author. Tel.: +49 0 221 470 2493; fax: +49 0 221 470 6073.
E-mail addresses: bbucchia@math.uni-koeln.de (B. Bucchia), cheuser@math.uni-koeln.de (C. Heuser).

http://dx.doi.org/10.1016/j.jspi.2015.04.005
0378-3758/

important to have an estimator which remains stable and bounded with respect to a change under the alternative. Otherwise, error in the estimation of Σ might lead to tests which display lower power for bigger changes. For example Vogelsang (1999) and Crainiceanu and Vogelsang (2001) investigate the problem of nonmonotonic power under data-dependent bandwidth choices for a test of mean shift in a univariate time series, noting that this might even lead to tests with no power against "obvious" changes, which could be detected with the naked eye. They conclude that this is due to the fact that the LRV estimator is constructed under the (misspecified) model of a stable mean. Indeed, under alternatives with abrupt changes in the mean, the arithmetic mean displays a bias which causes associated kernel-type LRV estimators to diverge for growing bandwidths. In order to avoid this effect – or at least attenuate it – we consider LRV estimators that use a mean estimator which is more adapted to the change alternative. Depending on the accuracy of the change-set estimation, it is then possible to obtain a consistent estimator. This method has been well studied in the time series literature. For instance, Juhl and Xiao (2009) present an LRV estimator for a univariate time series which remains consistent and bounded under both the null and alternative hypotheses, where the mean function fulfills a Lipschitz condition under the alternative, and Antoch et al. (1997), Kejriwal (2009) and Hušková and Kirch (2010) investigate an At-Most-One-Change location model. The aim of this paper is to extend this methodology to the random field case.

This paper is organized as follows: In Section 2, we present notations, the model and the main assumptions on the considered process. In Section 3, we study the behavior of an LRV estimator constructed without taking the change into account and compare it to a modification which makes use of estimators for the magnitude and location of the change. Section 4 gives an example of a change-set estimator with the associated estimation rate. Finally, Section 5 contains a small simulation study in order to give an impression of the finite sample behavior of the estimators and associated change-point tests, both for simulated data and a real data-set. Technical proofs are relegated to the Appendix.

2. Model and main assumptions

The following notations will be used throughout this paper. Let $\mathbb{R}^d$ ($d \in \mathbb{N}$) be the vector space of real vectors equipped with the usual partial order. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we write $\mathbf{x} \vee \mathbf{y} = (\max\{x_1, y_1\}, \dots, \max\{x_d, y_d\})^T$ and $\mathbf{x} \wedge \mathbf{y} = (\min\{x_1, y_1\}, \dots, \min\{x_d, y_d\})^T$ as well as $\lfloor \mathbf{x} \rfloor = (\lfloor x_1 \rfloor, \dots, \lfloor x_d \rfloor)^T$ for the integer part of $\mathbf{x}$, $|\mathbf{x}| = (|x_1|, \dots, |x_d|)^T$ and $[\mathbf{x}] = x_1 \cdots x_d$. We use the notations $x^{(i)}$ or x_i for the ith entry of a vector and analogously for matrices. The notation $\|\cdot\|$ is used to denote the maximum norm $\|\mathbf{x}\| = \max_{i=1,\dots,d} |x_i|$. Furthermore, for any integer $k \in \mathbb{N}_0$, we denote $(k, \dots, k)' \in \mathbb{N}_0^d$ by $\underline{\mathbf{k}}$. A rectangle in $\mathbb{R}^d$ is a set of the form

$$(\mathbf{x}, \mathbf{y}] = \{\mathbf{z} = (z_1, \dots, z_d)^T : x_i < z_i \le y_i,\ i = 1, \dots, d\}$$

for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ ($(\mathbf{x}, \mathbf{y}] = \emptyset$, if $x_i \ge y_i$ for some $i \in \{1, \dots, d\}$). A rectangle in $\mathbb{Z}^d$ is the intersection of a rectangle in $\mathbb{R}^d$ and the set $\mathbb{Z}^d$. We denote the Lebesgue measure on $\mathbb{R}^d$ by λ. Note that for the union of two disjoint rectangles $(\mathbf{k}_1, \mathbf{m}_1]$ and $(\mathbf{k}_2, \mathbf{m}_2]$ with endpoints $\mathbf{k}_i, \mathbf{m}_i \in \mathbb{Z}^d$ it holds that

$$\lambda((\mathbf{k}_1, \mathbf{m}_1] \cup (\mathbf{k}_2, \mathbf{m}_2]) = \#((\mathbf{k}_1, \mathbf{m}_1] \cap \mathbb{Z}^d) + \#((\mathbf{k}_2, \mathbf{m}_2] \cap \mathbb{Z}^d),$$

where $\#A$ denotes the cardinality of a finite set A. Therefore, we do not always explicitly distinguish between the notations and take $\lambda(C)$ to mean either the Lebesgue measure of a set in $\mathbb{R}^d$ or (for finite sets) its cardinality. To simplify notation we write $\lambda(\mathbf{k}, \mathbf{m}] = \lambda((\mathbf{k}, \mathbf{m}])$ for any rectangle $(\mathbf{k}, \mathbf{m}]$. We denote the symmetric difference of two sets A and B by $A \triangle B$. For a function $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^d$, the increment of f over a rectangle $(\mathbf{s}, \mathbf{t}] \subset D$ takes the form

$$f(\mathbf{s}, \mathbf{t}] = \begin{cases} \sum_{\boldsymbol{\varepsilon} \in \{0,1\}^d} (-1)^{d - \sum_{i=1}^d \varepsilon_i} f(\mathbf{s} + \boldsymbol{\varepsilon}(\mathbf{t} - \mathbf{s})), & \mathbf{s} < \mathbf{t} \\ 0, & \mathbf{s} \not< \mathbf{t}. \end{cases}$$

Unless stated otherwise, we will always denote the complement of a set $R \subseteq (\underline{\mathbf{0}}, \underline{\mathbf{n}}]$ by $R^c = (\underline{\mathbf{0}}, \underline{\mathbf{n}}] \setminus R$ and take sums of the form $\sum_{\mathbf{k} \in R}$ to mean the summation over all $\mathbf{k} \in R \cap \mathbb{Z}^d$. The data-generating process considered here is an $\mathbb{R}^p$-valued random field $\{X_{\mathbf{k}}\}$ with

$$X_{\mathbf{k}} = Y_{\mathbf{k}} + \mu + I_{\mathbf{k} \in C_n} \Delta = Y_{\mathbf{k}} + \mu(\mathbf{k}), \quad \mathbf{k} \in [1, n]^d \cap \mathbb{Z}^d, \tag{1}$$

with a shift Δ that fulfills $\Delta^T \Delta > 0$, a subset $C_n \subset [1, n]^d$ and the mean function $\mu(\mathbf{k}) = EX_{\mathbf{k}} = \mu + I_{\mathbf{k} \in C_n} \Delta$. All the parameters are considered unknown. Since the mean deviates from its value μ on C_n, we call this the change-set. In particular, we have $C_n = (0, k_2^0]$ ($d = 1$) and $C_n = \left(\mathbf{k}_1^0, \mathbf{k}_2^0\right]$ ($d \ge 1$) in mind. For such rectangles C_n the resulting change-set problem is the straightforward generalization to the multiparameter case of a one-dimensional change-point problem with two change-points $0 < k_0 < m_0 < n$. This type of problem is known in the change-point literature as an epidemic change-point. A more detailed description of the epidemic change-point problem and its multiparameter version, as well as some references to further research, can be found in Bucchia (2014). In order to allow slightly more general change-sets for the LRV estimation, we consider the following case: