

# Self-normalization for Time Series: A Review of Recent Developments <sup>1</sup>

February 24, 2015

XIAOFENG SHAO

**Abstract:** This article reviews some recent developments on the inference of time series data using the self-normalized approach. We aim to provide a detailed discussion about the use of self-normalization in different contexts, and highlight distinctive feature associated with each problem and connections among these recent developments. The topics covered include: confidence interval construction for a parameter in a weakly dependent stationary time series setting, change point detection in the mean, robust inference in regression models with weakly dependent errors, inference for nonparametric time series regression, inference for long memory time series, locally stationary time series and near-integrated time series, change point detection and two sample inference for functional time series, as well as the use of self-normalization for spatial data and spatial-temporal data. Some new variations of the self-normalized approach are also introduced with additional simulation results. We also provide a brief review of related inferential methods, such as blockwise empirical likelihood and subsampling, which were recently developed under the fixed- $b$  asymptotic framework. We conclude the paper with a summary of merits and limitations of self-normalization in the time series context and potential topics for future investigation.

**Keywords:** Confidence Interval, Dependence, Functional Data, Inference, Locally Stationary, Long Memory, Resampling, Studentization, Testing.

---

<sup>1</sup>Xiaofeng Shao is Associate Professor, Department of Statistics, University of Illinois, at Urbana-Champaign, Champaign, IL 61820 (e-mail: xshao@illinois.edu). The research is supported in part by NSF grants DMS-0804937, DMS-1104545 and DMS-1407037. I would like to thank Michael Stein and Steve Lalley for insightful comments on my seminar presentation at the university of Chicago and Weibiao Wu for helpful discussions on Bahadur representation of sample quantiles. I thank Yeonwoo Rho, Xianyang Zhang and Zhou Zhou for helpful comments on an early version and Xianyang Zhang for working out the details of Example 4.1. I am also grateful to the two referees for constructive comments that led to a substantial improvement of the paper. All remaining errors are mine.

# 1 Self-normalization for iid data

Self-normalization has a long history in statistics that can be traced back to the work of Student (1908). In this pathbreaking paper, Student (the pseudonym of William Sealy Gosset) introduced the celebrated  $t$  statistic and the  $t$ -distribution, and emphasized the need for developing small sample theory. The paper has been widely commented upon by prominent statisticians, see Lehmann (1999) and Zabell (2008) and the references therein. Among its many contributions to statistics, Student's  $t$ -statistic is the first prototypical example of self-normalization (SN, hereafter), see de la Pena, Lai and Shao (2009) for a recent book-length treatment of this subject. Note that the latter book mainly focuses on probabilistic aspects of self-normalization for iid (independent and identically distributed) random variables and martingales, whereas this paper is concerned with SN-related statistical methodology and theory for time series and dependent data.

We shall start by briefly reviewing the  $t$ -statistic and its generalization in statistics. For a random sample  $(X_i)_{i=1}^n$  from the normal distribution denoted as  $N(\mu, \sigma^2)$  where  $(\mu, \sigma^2)$  is unknown, let  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$  be the sample mean and  $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  be the sample variance. To test  $H_0 : \mu = \mu_0$  versus  $H_a : \mu \neq \mu_0$ , we form the  $t$ -statistic as

$$t_n(\mu_0) = \frac{\bar{X}_n - \mu_0}{S_n / \sqrt{n}}.$$

Under the null hypothesis,  $t_n(\mu_0) =_D t(n-1)$ , where  $t(n-1)$  stands for the  $t$  distribution with  $n-1$  degrees of freedom and “ $=_D$ ” stands for equality in distribution. Before Gosset's work, the standard normal distribution  $N(0, 1)$  was used as an approximation to the finite sample distribution of  $t_n(\mu_0)$  under the null, which was natural as  $\frac{\bar{X}_n - \mu_0}{\sigma / \sqrt{n}} =_D N(0, 1)$  under the null hypothesis and  $S_n \rightarrow_p \sigma$  as  $n \rightarrow \infty$ , where “ $\rightarrow_p$ ” stands for convergence in probability. In small samples, however, the variability of  $S_n$  does matter and the normal approximation, which acts as if  $S_n$  is the constant  $\sigma$ , may result in a substantial error.

In the past century, Gosset's  $t$ -statistic has been generalized to studentized statistic/quantity in a general setting and studentization has become an important principle in conducting statistical inference. Suppose we have  $\theta = T(F) \in \mathbb{R}$ , where  $F$  is the marginal cumulative distribution function (cdf) of the random sample  $(X_i)_{i=1}^n$  and  $T$  is a functional. A natural estimator of  $\theta$  is  $\theta_n = T(F_n)$ , where  $F_n$  is the empirical cdf. Suppose that under suitable regularity conditions we have  $\sqrt{n}(\theta_n - \theta) \rightarrow_D N(0, \sigma_F^2)$ , where  $\sigma_F^2$  is the asymptotic variance and “ $\rightarrow_D$ ” stands for convergence in distribution. To construct a confidence

interval for  $\theta$ , we can rely on the studentized quantity

$$G_n(\theta) = \frac{\sqrt{n}(\theta_n - \theta)}{\hat{\sigma}_F},$$

where  $\hat{\sigma}_F^2$  is a consistent estimator of  $\sigma_F^2$ . It can be a plug-in estimator, i.e.,  $\hat{\sigma}_F^2 = \sigma_{F_n}^2$ . For example, in the mean case,  $\sigma_F^2 = \mathbb{E}_F(X - \mathbb{E}_F(X))^2$  and  $\sigma_{F_n}^2 = \mathbb{E}_{F_n}(X - \mathbb{E}_{F_n}(X))^2 = n^{-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$ . Since  $G_n(\theta) \rightarrow_D N(0, 1)$ , the interval can be constructed based on the approximate pivot  $G_n(\theta)$ . The traditional method described above requires deriving a close-form expression for the asymptotic variance, and then finding a consistent variance estimate by using a direct plug-in approach. Alternatively, consistent estimation of asymptotic variance can be based on jackknife, bootstrap or subsampling methods, so no theoretical formula for asymptotic variance needs to be derived; see Shao and Tu (1995) Chapter 1 for a nice summary of the merits of jackknife and bootstrap methods.

The self-normalized approach to inference for time series was first introduced in Shao (2010a) as an extension of the idea of Kiefer, Vogelsang and Bunzel (2000) and Lobato (2001). Simply put, it is a novel studentization technique which takes advantage of the unidirectional ordering of time series and avoids or reduces the bandwidth choice, the latter of which is common in existing inferential methods for time series. In Section 2, we provide a review of the main idea in Shao (2010a), some additional remarks, variations and simulation comparisons. Section 3 contains extensions of the SN approach to other inference problems for univariate time series. Section 4 is devoted to the inference of functional time series. Section 5 briefly mentions an extension of the SN approach to spatial data and an application to assess the properties of a space-time covariance function. In Section 6, we review some related inferential methods recently developed under the fixed- $b$  asymptotic framework. Section 7 concludes. Lastly, let me apologize in the beginning for the necessarily selective coverage and for putting the emphasis on methods and extensions that are mostly associated with my own work.

## 2 Self-normalization for time series

In time series analysis, dependence is the rule rather than the exception. Suppose we have a strictly stationary time series  $X_t, t \in \mathbb{Z}$ , and we are interested in constructing a confidence interval for its mean  $\mu = \mathbb{E}(X_t)$ . For the natural estimator  $\bar{X}_n$ , we know that

under suitable moment and weak dependence conditions,

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow_D N(0, s^2),$$

where  $s^2 = \sum_{k=-\infty}^{\infty} \gamma(k)$  with  $\gamma(k) = \text{cov}(X_0, X_k)$ . Following the traditional approach, we would try to find a consistent estimator for the asymptotic variance  $s^2$ , which is called long-run variance in econometrics. Since  $s^2 = 2\pi f(0)$ , where  $f(\cdot)$  is the spectral density function of  $X_t$ , a consistent estimator of  $s^2$  is basically a consistent estimator of  $f(\lambda)$  evaluated at zero frequency up to a multiplicative constant. In spectral analysis, spectral density estimation is a well-studied problem and a popular class of estimators is called the lag-window estimator [see Priestley (1981)]. At zero frequency, it admits the following form,

$$s_n^2 = \sum_{k=1-n}^{n-1} K(k/l_n) \hat{\gamma}_n(k), \quad (1)$$

where  $K(\cdot)$  is a kernel function,  $l_n$  is a bandwidth parameter, and  $\hat{\gamma}_n(k) = n^{-1} \sum_{t=|k|+1}^n (X_t - \bar{X}_n)(X_{t-|k|} - \bar{X}_n)$  is sample autocovariance at lag  $k$ . This is also the same as the HAC (heteroscedasticity-autocorrelation consistent) estimator popular in econometrics literature. The bandwidth parameter  $l_n$  needs to satisfy  $1/l_n + l_n/n \rightarrow 0$  as  $n \rightarrow \infty$  to ensure the consistency of  $s_n^2$ . The choice of bandwidth affects the finite sample performance, so it is natural to ask whether there is an optimal choice that can be made data driven. One way to find the expression of  $l_{n,opt}$  is by minimizing the mean squared error of  $s_n^2$ , and this was derived in Andrews (1991) and Newey and West (1994) in the multivariate setting. For the Bartlett kernel, i.e.,  $K(x) = (1 - |x|)\mathbf{1}(|x| \leq 1)$ , which is locally linear around the origin,

$$l_{n,opt} = 1.1447(\alpha(1)n)^{1/3}, \text{ where } \alpha(1) = \{f'(0)/f(0)\}^2,$$

and  $f'(0)$  denotes the first order derivative of  $f$  at zero frequency. The optimal bandwidths for the kernels that are locally quadratic around the origin are of form  $Cn^{1/5}$ , where  $C$  depends on the second order derivative of  $f$  at zero and the kernel; see Equation (6.2) of Andrews (1991). The difficulty lies in the estimation of the derivative of spectral density at the first or second order, which is again a nonparametric estimation problem and requires the choice of another bandwidth parameter. To avoid additional bandwidth choice, one can obtain an estimate of  $\alpha(1)$  or  $C$  above by assuming/approximating the data generating

process with a parametric model. For example, if the (approximated) time series model is AR(1), then  $\alpha(1)$  only depends on the AR(1) parameter  $\rho$  and it can be estimated by a direct plug-in approach, i.e.,

$$\hat{l}_{n,opt} = 1.1447(\hat{\alpha}(1)n)^{1/3}, \text{ where } \hat{\alpha}(1) = \frac{4\hat{\rho}^2}{(1 - \hat{\rho})^2(1 + \hat{\rho})^2}, \quad (2)$$

and  $\hat{\rho} = \sum_{t=2}^n (X_t - \bar{X}_n)(X_{t-1} - \bar{X}_n) / \sum_{t=2}^n (X_{t-1} - \bar{X}_n)^2$  is the AR(1) parameter estimate. Of course, the validity of this formula much depends on how close the AR(1) model is to the true data generating process.

If the goal is to find the optimal bandwidth for the testing problem:  $H_0 : \mu = \mu_0$  versus  $H_a : \mu \neq \mu_0$ , in terms of balancing type I and type II errors, then the optimal bandwidth that minimizes the mean squared error is not necessarily the testing-optimal bandwidth; see Sun, Phillips and Jin (2008) for detailed discussions. In the testing context, for a given loss function, the optimal bandwidth can be derived but it nevertheless depends on the derivative of spectral density at zero frequency, a quantity that is more difficult to estimate than  $f(0)$  in terms of nonparametric convergence rate. It is worth mentioning that another approach is to approximate the data generating process with an AR( $p$ ) model, where  $p$  is allowed to grow to  $\infty$  as  $n \rightarrow \infty$ . Then the resulting estimate of  $f(0)$  or its derivatives is simply the spectral density of the estimated AR( $p$ ) process or its derivative at zero frequency. It is expected that if the underlying process is AR( $\infty$ ) with iid innovations, then this sieve approach [Berk (1974)] also provides a consistent estimate. It nevertheless requires the choice of  $p$ , which is typically selected using model selection criteria (e.g., BIC). For stationary nonlinear time series, it is not known whether this AR sieve approach is still consistent as the theoretical framework in Berk (1974) excludes nonlinear processes.

To avoid choosing the bandwidth parameter, Kiefer, Vogelsang and Bunzel (2000, KVB, hereafter) and Lobato (2001) proposed to use an inconsistent studentizer in the inference. While KVB dealt with the inference of regression parameter in dynamic regression models and Lobato (2001) tackled the testing problem of whether the autocovariances at first  $L$  lags are equal to zero, the formulation of their ideas turn out to be identical in the mean case; see Section 2 of Lobato (2001) for an illuminating discussion. In the mean case, they used the self-normalizer

$$W_n^2 = n^{-2} \sum_{t=1}^n (S_t - t\bar{X}_n)^2, \text{ where } S_t = \sum_{j=1}^t X_j.$$

Let  $D[0, 1]$  denote the space of functions on  $[0, 1]$  which are right continuous and have left limits, endowed with the Skorokhod topology (Billingsley, 1968). Denote by  $\Rightarrow$  weak convergence in  $D[0, 1]$  or more generally in the  $\mathbb{R}^d$ -valued function space  $D^d[0, 1]$ , where  $d \in \mathbb{N}$ . Let  $\lfloor a \rfloor$  denote the integer part of  $a \in \mathbb{R}$ . Assuming that

$$n^{-1/2} \sum_{t=1}^{\lfloor nr \rfloor} (X_t - \mu) \Rightarrow sB(r), \quad (3)$$

i.e., the functional central limit theorem (FCLT) holds, then by the continuous mapping theorem, we can derive

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{W_n^2}} \rightarrow_D Q_1 := \frac{B(1)}{\sqrt{\int_0^1 \{B(u) - uB(1)\}^2 du}}. \quad (4)$$

Unlike the  $t$ -statistic, in which the self-normalizer  $S_n^2$  is a consistent estimator of the variance  $\sigma^2$ ,  $W_n^2$  is not consistent for  $s^2$  but is proportional to  $s^2$ . Since the nuisance parameter  $s$  is proportional to both the denominator and the numerator in (4), it gets canceled out and the distribution of  $Q_1$  becomes pivotal. Similar to the  $t$  distribution, the denominator  $\sqrt{\int_0^1 \{B(u) - uB(1)\}^2 du}$  is independent of the numerator  $B(1)$  and the distribution of  $Q_1$  can be regarded as a mixture of standard normal distribution. The density of  $U_1 = Q_1^2$  has been plotted in Lobato (2001) using Monte-carlo simulations, and its moments and tail behavior have been analytically studied by Abadir and Paruolo (1997). It is worth noting that the distribution of  $U_1$  has heavier tails than that of  $\chi_1^2$ .

**REMARK 2.1.** There is a growing literature on the so-called fixed- $b$  asymptotics for time series inference initiated by Kiefer and Vogelsang (2005). The basic idea is that in deriving the limiting distribution of the studentized quantity, where the lag-window estimator  $s_n^2$  is the studentizer, we hold  $b = l_n/n \in (0, 1]$  as a fixed constant in the asymptotics. This is different from the conventional small- $b$  asymptotics [Newey and West (1987)], under which  $b \rightarrow 0$  and  $s_n^2$  consistently estimates  $s^2$ . Under the fixed- $b$  asymptotics,  $s_n^2$  is no longer consistent but is still proportional to  $s^2$ . In the limit, we have

$$\frac{n(\bar{X}_n - \mu)^2}{s_n^2} \rightarrow_D \mathcal{D}(\{B(r)\}_{r \in [0, 1]}; K(\cdot), b)$$

where the limiting distribution depends on the kernel and  $b$  and its detailed form can be found in Kiefer and Vogelsang (2005). For a given  $b$  and kernel, the limiting distribution

is pivotal and its critical values can be found by Monte carlo simulation. In other words, the limiting distribution captures the choice of the kernel and bandwidth, thus two users who use different kernel functions or bandwidths end up with using different limiting distributions as reference distributions. This turns out to be more advantageous than the traditional approach where the distribution of  $\chi_1^2$  is used for different choices of bandwidths and kernel functions. The superiority of the fixed- $b$  limiting distribution to the small- $b$  counterpart has been demonstrated in simulations and in theory; see Jansson (2004), Sun et al. (2008) and Zhang and Shao (2013) for high order distribution theory. As pointed out in Shao (2010a), the SN approach is a special case of the fixed- $b$  approach for the mean inference in that  $2W_n^2 = s_n^2$  when we take the Bartlett kernel and let  $b = 1$  (i.e.,  $l_n = n$ ); see Kiefer and Vogelsang (2002).

**REMARK 2.2.** Müller (2007) showed that in the inference of the mean for a stationary Gaussian time series, the consistent long run variance estimator can converge to any positive value when there is contamination in the autocorrelations. In particular, he defined two neighborhoods of contaminated models via the deviation of the covariance matrix of the contaminated model from that of the benchmark model, and the contamination neighborhoods include (1) local-to-unity Gaussian AR(1) process [Chan and Wei (1987), Phillips (1987)], (2) Gaussian white noise with a relatively low frequency seasonal component, and (3) Gaussian white noise with a Gaussian outlier at certain date. In other words, the inference based on consistent long run variance estimator is fragile. On the other hand, he showed that the fixed- $b$  approach, which corresponds to inconsistent long run variance estimation, is robust against the contamination. Since the SN approach is a special case of the fixed- $b$  approach in the mean case, this suggests that the SN approach is robust. Compared to the traditional approach, the robustness of the SN approach comes with a price, that is, efficiency, as reflected by the power loss in testing or longer length in confidence interval construction (see Section 2.1). Thus we may view the SN approach as a robust alternative to the existing inference approaches, where consistent variance estimation is involved.

Within the fixed- $b$  approach, there is also a tradeoff between size distortion and power loss, as seen from the simulation results in Kiefer and Vogelsang (2005). With larger  $b$ , the size distortion diminishes but the power loss increases for a given kernel; see Figures 11-15 in Kiefer and Vogelsang (2005). According to Figures 3-10 in Kiefer and Vogelsang (2005), the power is essentially the same across kernels for small  $b$ , but for large  $b$ , Bartlett kernel

corresponds to the highest power, compared to quadratic spectral kernel, Daniell kernel and others. On the other hand, the size distortion for the Bartlett kernel is the largest among all kernels for large  $b$ . Since the SN approach corresponds to the use of the Bartlett kernel and setting  $b = 1$ , it can be viewed as a special member of fixed- $b$  family that has the largest possible  $b$ , so size distortion is the least for a given kernel and achieves the highest local power among kernels compared in Kiefer and Vogelsang (2005) once  $b = 1$  is fixed.

## 2.1 Large sample comparisons of the intervals and tests

To further understand the difference between the traditional inference approach and the SN approach, we offer a detailed comparison of the two intervals and corresponding tests based on large sample approximation in this subsection and some additional simulation comparison in Section 2.3, which complement the results in Lobato (2001) and Shao (2010a). Note that in the mean case, the  $100\alpha\%$  intervals can be constructed as follows:

$$\text{SN-based interval : } [\bar{X}_n - \sqrt{U_{1,\alpha}}W_n/\sqrt{n}, \bar{X}_n + \sqrt{U_{1,\alpha}}W_n/\sqrt{n}], \quad (5)$$

$$\text{Traditional interval : } [\bar{X}_n - \sqrt{\chi_{1,\alpha}^2}s_n/\sqrt{n}, \bar{X}_n + \sqrt{\chi_{1,\alpha}^2}s_n/\sqrt{n}], \quad (6)$$

where  $U_{1,\alpha}$  and  $\chi_{1,\alpha}^2$  are the  $100\alpha\%$  upper quantile of the distribution  $U_1$  and  $\chi_1^2$ , respectively.

For a confidence interval, we care about its interval length and coverage accuracy. In large samples, the coverage accuracy is determined by the magnitude of the approximation error when approximating the finite sample distribution of the studentized quantity with its limiting distribution. For the self-normalized mean, such high order expansion of its finite sample distribution has been derived under the framework of the Gaussian location model. In particular, Jansson (2004) showed that

$$\sup_{x \in \mathbb{R}} \left| P \left( \frac{n(\bar{X}_n - \mu)^2}{W_n^2} \leq x \right) - P(U_1 \leq x) \right| = O(n^{-1} \log(n)), \quad (7)$$

which was further refined by Sun et al. (2008) by dropping the  $\log(n)$  term (also see Zhang and Shao (2013)). By contrast, Velasco and Robinson (2001) showed that

$$\sup_{x \in \mathbb{R}} \left| P \left( \frac{n(\bar{X}_n - \mu)^2}{s_n^2} \leq x \right) - P(\chi_1^2 \leq x) \right| = O(n^{-(1-\delta)}), \text{ for some } \delta > 0, \quad (8)$$



where  $\delta$  depends on the choice of kernel, the bandwidth parameter  $l_n$  in  $s_n^2$  (see (1)) and the smoothness of the spectral density around zero frequency. Therefore the approximation error associated with the self-normalized quantity  $\frac{n(\bar{X}_n - \mu)^2}{W_n^2}$  is smaller, resulting in better coverage accuracy for the SN-based interval. Although the above-mentioned results are only stated for Gaussian time series, we conjecture that the same phenomenon still carries over to the non-Gaussian case, as seen from the simulation results in Lobato (2001) and Shao (2010a). On the other hand, under some mild moment and weak dependence conditions on  $X_t$ , it can be shown that  $2W_n^2$  is an asymptotically unbiased estimator of  $s^2$ , with bias of order  $O(n^{-1})$  and variance of order  $O(1)$ . The asymptotically nonnegligible variance is captured in the first order limiting distribution  $U_1$ . Compared to the consistent studentizer  $s_n^2$  used in HAC-based testing, there is a reduction in the bias, which partially explains more accurate size (coverage) in testing (confidence interval construction); see Simonoff (1993) for a nice discussion of the role of the bias of variance estimator in the inference.

To compare the interval lengths, we examine (i) the probability that SN-based interval is shorter and (ii) ratio of expected interval lengths (say, SN versus traditional). Under the assumptions that  $W_n^2 \rightarrow_D s^2 \int_0^1 \tilde{B}(r)^2 dr$ , where  $\tilde{B}(r) = B(r) - rB(1)$  and  $s_n^2 \rightarrow_p s^2$ , we can derive that

$$P(\sqrt{U_{1,\alpha}} W_n < \sqrt{\chi_{1,\alpha}^2} s_n) \rightarrow P(\alpha) := P\left(U_{1,\alpha} \int_0^1 \tilde{B}(r)^2 dr < \chi_{1,\alpha}^2\right)$$

and with some additional assumptions, the following holds

$$\frac{\sqrt{U_{1,\alpha}} \mathbb{E}(W_n)}{\sqrt{\chi_{1,\alpha}^2} \mathbb{E}(s_n)} \rightarrow R(\alpha) := \frac{\sqrt{U_{1,\alpha}} \mathbb{E}\left(\sqrt{\int_0^1 \tilde{B}(r)^2 dr}\right)}{\sqrt{\chi_{1,\alpha}^2}}$$

Please insert Table 1 here!

Table 1 tabulates the values of  $P(\alpha)$  and  $R(\alpha)$  for  $\alpha = 0.5, 0.75, 0.9, 0.95, 0.99$ . It can be seen that  $P(\alpha)$  decreases and  $R(\alpha)$  increases as  $\alpha$  increases,  $P(\alpha)$  is larger than 50% for  $\alpha > 0.75$ , and  $R(\alpha)$  can be smaller than 1 for small enough  $\alpha$ . For conventional confidence levels  $\alpha = 0.9, 0.95, 0.99$ , the SN-based interval is on average wider than the traditional interval, indicating a loss of power in the testing context.

Please insert Figure 1 here!

To quantify the power loss in the large sample precisely, we can also compare the two tests in a similar manner. Note that an asymptotic level- $\gamma$  (say  $\gamma = 0.05$ ) test for  $H_0 : \mu = \mu_0$  versus  $H_1 : \mu \neq \mu_0$  can be constructed as follows:

$$\begin{aligned} \text{SN-based test function : } & \mathbf{1}(\{\sqrt{n}(\bar{X}_n - \mu_0)\}^2/W_n^2 > U_{1,1-\gamma}), \\ \text{Traditional test function : } & \mathbf{1}(\{\sqrt{n}(\bar{X}_n - \mu_0)\}^2/s_n^2 > \chi_{1,1-\gamma}^2). \end{aligned}$$

Under the local alternative  $\mu_n = \mu_0 + \delta s/\sqrt{n}$ ,  $\delta \neq 0$ , we can derive the limit of local power as

$$\begin{aligned} P(\text{SN-based test rejects } H_0 | \mu = \mu_n) &= P(\{\sqrt{n}(\bar{X}_n - \mu_0)\}^2/W_n^2 > U_{1,1-\gamma} | \mu = \mu_n) \\ &= P\left(\frac{\{\sqrt{n}(\bar{X}_n - \mu_n)\}^2 + n(\mu_n - \mu_0)^2 + 2\sqrt{n}(\bar{X}_n - \mu_n)(\mu_n - \mu_0)}{W_n^2} > U_{1,1-\gamma} | \mu = \mu_n\right) \\ &\rightarrow P\left(\frac{\{B(1) + \delta\}^2}{\int_0^1 \{B(u) - uB(1)\}^2 du} > U_{1,1-\gamma}\right) =: P_{SN}(\gamma, \delta) \end{aligned}$$

and

$$\begin{aligned} P(\text{Traditional test rejects } H_0 | \mu = \mu_n) &= P(\{\sqrt{n}(\bar{X}_n - \mu_0)\}^2/s_n^2 > \chi_{1,1-\gamma}^2 | \mu = \mu_n) \\ &\rightarrow P(\{B(1) + \delta\}^2 > \chi_{1,1-\gamma}^2) =: P_{TR}(\gamma, \delta). \end{aligned}$$

To calculate the limiting local powers, we approximate the standard Brownian motion with the standardized partial sum of 1000 iid standard normal random variables and repeat 50000 times. Figure 1 shows the approximation of two probabilities  $P_{SN}(\gamma, \delta)$  and  $P_{TR}(\gamma, \delta)$  as a function of  $\delta$  for two levels  $\gamma = 5\%$  and  $10\%$ . It can be seen that the power curves appear symmetric with respect to  $\delta$ , there is a power loss with the SN-based test, and the power loss decreases when we increases the level from 0.05 to 0.1. If we look at the relative percentage of power loss, i.e.,  $\frac{P_{TR}(\gamma, \delta) - P_{SN}(\gamma, \delta)}{P_{TR}(\gamma, \delta)}$ , the largest loss is about 28.5% for  $\gamma = 0.05$  and it occurs around  $\delta = \pm 1.9$ , whereas the largest loss is about 20% for  $\gamma = 0.1$  and it occurs around  $\delta = \pm 1.7$ . This is consistent with the finding in Lobato (2001), who showed that there is a loss of power for the SN approach in testing the zeroness of autocovariances at the first  $L$  lags compared to the traditional approach, where consistent estimation of asymptotic variance is involved. Also see Shao (2010a) for numerical evidence of power loss.

## 2.2 General framework

Suppose we are interested in the inference of the median of  $X_t$ , i.e., the parameter of interest is  $\theta = F^{-1}(1/2)$ , where  $F$  is the marginal cdf of  $X_t$ . Then a natural estimator is  $\hat{\theta}_n = F_n^{-1}(1/2)$ , where  $F_n$  is the empirical cdf based on time series observations  $\{X_t\}_{t=1}^n$ . Under suitable weak dependence and moment conditions, we can establish that

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow_D N(0, \sigma^2),$$

where  $\sigma^2 = \{4g(\theta)^2\}^{-1} \sum_{k=-\infty}^{\infty} \text{cov}\{1 - \mathbf{21}(X_0 \leq \theta), 1 - \mathbf{21}(X_k \leq \theta)\}$  with  $g(\cdot)$  being the density function of  $X_1$ ; see Bühlmann (2002). A direct consistent estimation of asymptotic variance is possible, but is rather involved. It requires the estimation of  $g(\theta)$  using a kernel density estimate, and the estimation of  $\sum_{k=-\infty}^{\infty} \text{cov}\{1 - \mathbf{21}(X_0 \leq \theta), 1 - \mathbf{21}(X_k \leq \theta)\}$  using lag-window type estimate, both of which involve bandwidth selections. Alternative approaches, such as blockwise empirical likelihood [Kitamura (1997)], subsampling [Politis and Romano (1994)] and the moving block bootstrap [Künsch (1989), Liu and Singh (1992)], can be used for variance estimation or confidence interval construction, but they also involve a choice of block size. The bandwidth parameter or block size serves as a tuning parameter and different choices may lead to very different finite sample results. Theoretically, it results in a tradeoff between bias and variance in estimation and balances Type I error and Type II error in testing. The data driven bandwidth (block size) choice has been discussed by Hall, Horowitz and Jing (1996), Lahiri (2003, chapter 7), Politis, Romano and Wolf (1999, Chapter 9), among others, but most of the algorithms require the choice of another tuning parameter, and no good guidance seems available in the literature.

In Shao (2010a), the self-normalized approach was formally proposed as a general methodology to constructing confidence intervals for quantities in time series analysis without any bandwidth parameter. It naturally extends the proposal of Lobato (2001) from the mean case to more general setting, including the median as a special case. For a strictly stationary univariate time series  $(X_t)$ , let  $\theta = T(F_m) \in \mathbb{R}^q$  be the quantity of interest, where  $T$  is a functional,  $F_m$  is the marginal cdf for  $Y_t = (X_t, \dots, X_{t+m-1})'$  for a fixed  $m \in \mathbb{N}$ . Given time series observations  $\{X_t\}_{t=1}^n$ , a natural estimator of  $\theta$  is  $\hat{\theta}_N = T(\hat{F}_{m,N})$ , where  $N = n - m + 1$ ,  $\hat{F}_{m,N} = N^{-1} \sum_{j=1}^N \delta_{Y_j}$  with  $\delta_y$  denoting the point mass at  $y \in \mathbb{R}^m$ . The SN approach uses recursive estimates  $\{\hat{\theta}_{1,t}\}_{t=1}^N$ , where  $\hat{\theta}_{1,t} = T(\hat{F}_{m,t})$  is based on  $\{Y_1, \dots, Y_t\}$ , in the construction of its self-normalizer. To see why recursive

estimates are useful, we write

$$\hat{\theta}_{1,t} = \theta + t^{-1} \sum_{j=1}^t IF(Y_j; F_m) + R_{1,t}, \quad t = 1, \dots, N, \quad (9)$$

where  $IF(Y_j; F_m)$  is the influence function of  $T$  [Hampel, Ronchetti, Rousseeuw and Stahel (1986)] and  $\{R_{1,t}\}_{t=1}^N$  are the remainder terms. In the case of the median,  $\theta = F^{-1}(1/2)$ ,  $m = 1$  and  $IF(Y_j; F_m) = \{1 - 2\mathbf{1}(X_j \leq \theta)\}/\{2g(\theta)\}$ . For approximately linear statistics, the remainder term  $R_{1,n}$  is typically asymptotically negligible, i.e.,

$$\sqrt{N}R_{1,N} = o_p(1). \quad (10)$$

For example,  $R_{1,N} = 0$  in the mean case, and  $R_{1,N} = O_{a.s.}(N^{-3/4}(\log N)^{3/2})$  in the median case [see Wu (2005)]. Here a sequence of random variables  $Z_n$  is said to be  $O_{a.s.}(1)$  ( $o_{a.s.}(1)$ ) if  $Z_n$  is almost surely bounded (converges to zero). Furthermore, we assume

$$\mathbb{E}(IF(Y_t; F_m)) = 0 \text{ and } N^{-1/2} \sum_{t=1}^N IF(Y_t; F_m) \rightarrow_D N(0, \Sigma(F_m)), \quad (11)$$

where  $\Sigma(F_m) = \sum_{k=-\infty}^{\infty} \text{cov}(IF(Y_0; F_m), IF(Y_k; F_m))$  is the long-run variance of the stationary process  $\{IF(Y_t; F_m)\}_{t \in \mathbb{Z}}$ . Combining (10) and (11), we can then derive the asymptotic normality for  $\hat{\theta}_N$ , i.e.,  $\sqrt{N}(\hat{\theta}_N - \theta) \rightarrow_D N(0, \Sigma(F_m))$ .

The validity of the SN approach requires stronger assumptions than those needed for the asymptotic normality. In particular, the following two assumptions are stated in Shao (2010a):

Assumption I: Assume that

$$\mathbb{E}\{IF(Y_t; F_m)\} = 0, \text{ and } N^{-1/2} \sum_{t=1}^{\lfloor rN \rfloor} IF(Y_t; F_m) \Rightarrow \Delta B_q(r), \quad (12)$$

where  $\Delta$  is a  $q \times q$  lower triangular matrix with nonnegative diagonal entries and  $B_q(\cdot)$  is a  $q$ -dimensional vector of independent Brownian motions. Further assume that  $\Delta\Delta' = \Sigma(F_m)$  is positive definite.

Assumption II: Assume that  $N^{1/2}R_{1,N} = o_p(1)$  and

$$N^{-2} \sum_{t=1}^N |tR_{1,t}|^2 = o_p(1), \quad (13)$$

where  $|v|$  denotes the Euclidean norm of a vector  $v \in \mathbb{R}^q$ . Thus Assumption I strengthens (11) from CLT to FCLT for the partial sum of influence functions, which can be satisfied by a wide class of time series models; see Shao (2010a) for more discussions. Compared to (10), Assumption II additionally imposes (13), which basically requires the remainder process  $\{R_{1,t}\}_{t=1}^n$  to be uniformly negligible in a weighted  $L^2$  sense. The verification of Assumption II for the smooth function model has been described in Shao (2010a). For sample quantiles, under certain regularity conditions, Wu (2005) asserts that  $R_{1,n} = O_{a.s.}(n^{-3/4}(\log n)^{3/2})$  for weakly dependent linear and nonlinear processes, which implies that  $n^{-2} \sum_{t=1}^n |tR_{1,t}|^2 = o_{a.s.}(n^{-1/2}(\log n)^3) = o_p(1)$ .

Define the self-normalization matrix

$$W_N^2 = N^{-2} \sum_{t=1}^N t^2 (\hat{\theta}_t - \hat{\theta}_N)(\hat{\theta}_t - \hat{\theta}_N)'. \quad (14)$$

In Shao (2010a), it was shown that under Assumptions I and II, the self-normalized quantity

$$N(\hat{\theta}_N - \theta)'(W_N^2)^{-1}(\hat{\theta}_N - \theta) \rightarrow_D U_q, \quad (15)$$

where  $U_q = B_q(1)'V_q^{-1}B_q(1)$  and  $V_q = \int_0^1 \{B_q(r) - rB_q(1)\}\{B_q(r) - rB_q(1)\}'dr$ .

The key idea behind the use of recursive estimates in the formation of self-normalization matrix lies in the following identity:

$$\frac{k}{\sqrt{N}}(\hat{\theta}_k - \hat{\theta}_N) = \frac{1}{\sqrt{N}} \sum_{j=1}^k \left\{ IF(Y_j; F_m) - N^{-1} \sum_{j=1}^N IF(Y_j; F_m) \right\} + \frac{1}{\sqrt{N}} \{kR_k - \frac{k}{N}NR_N\} \quad (16)$$

for  $k = 1, \dots, N$ . To extend Lobato's SN approach from the mean to other quantities, it would be natural to define the self-normalizer as a continuous mapping of the partial sum process of the influence functions, where the mapping used in Lobato (2001) basically corresponds to

$$\mathcal{M} : h \in D^q[0, 1] \rightarrow \mathcal{M}(h) = \int_0^1 h(r)h(r)'dr. \quad (17)$$

However, the influence function may contain unknown nuisance parameters (e.g.,  $g(\theta)$  is unknown in the influence function of sample median). To avoid direct consistent estimation of the unknown parameter, the above identity suggests that we can apply the mapping  $\mathcal{M}$

to the process based on recursive estimates on the left hand side of (16). This strategy is expected to work if the second term on the right hand side of (16) is asymptotically negligible. Under Assumption II, it turns out that our self-normalizer  $W_N^2 \rightarrow_D \Delta V_q \Delta'$  and consequently it follows from the continuous mapping theorem that (15) holds, thus the use of recursive estimates to form a self-normalizer is justified. It is worth noting that verifying Assumption II is quite nontrivial in practice and it requires a tedious case-by-case study. Recently Volgushev and Shao (2014) provided some new theory that allows us to derive the convergence in distribution in (15) by utilizing the weak convergence of sequential empirical process and functional delta method, see Section 3.1 of Volgushev and Shao (2014) for details.

**REMARK 2.3.** The applicability of the SN approach can be considerably widened to include statistics that do not fall into the above-mentioned framework. In Shao (2010a), it was shown that it is also applicable to certain quantities that can be written as  $\theta = T(F^\infty)$ , where  $F^\infty$  stands for the distribution of the whole process  $(X_t)_{t \in \mathbb{Z}}$ . A notable example in this framework is the so-called spectral mean and its ratio version. In addition, we expect the applicability of the SN approach to the  $U$ -statistic given that it has a similar expansion as in (9) and is asymptotically normal with  $\sqrt{n}$  convergence rate in the non-degenerate case.

**REMARK 2.4.** Jackknife [Quenouille (1949), Tukey (1958)] is well known to be a method useful for estimating/reducing the bias and estimating the variance of an estimate without knowing the theoretical formula. For time series data, the traditional leave-one-out jackknife does not work and a blockwise jackknife corresponding to “leave a block out” was proposed by Künsch (1989) and was shown to be consistent. Our self-normalizer is formed in a way similar to blockwise jackknife as a block of consecutive observations (i.e.,  $(Y_{t+1}, \dots, Y_N)$  for  $t = 1, \dots, N$ ) is deleted and the estimate based on the remaining sample is calculated. However, in the blockwise jackknife, each time a block of fixed size is deleted whereas in our SN approach a block of diminishing size is removed. As a result, our self-normalizer is not a consistent estimate of asymptotic variance but is proportional to it, and it leads to a nonstandard but pivotal limiting distribution for our self-normalized quantity.

**REMARK 2.5.** In the mean case, the SN approach is a special case of the fixed- $b$  approach, but in other settings, this relationship does not necessarily hold, as demonstrated in Rho

and Shao (2013a). In the latter paper, the effect of prewhitening was investigated when applied to both SN approach and the KVB approach in time series regression models with dynamic regressors. There the SN approach was found to differ from the KVB approach presented in Kiefer, Vogelsang and Bunzel (2000), which is a special case of the fixed- $b$  approach [Kiefer and Vogelsang (2005)] formulated for dynamic regression models; see Remark 3.1 in Rho and Shao (2013a). Additionally, simulation results suggest that using a prewhitened variance estimator helps to reduce the size distortion in the strongly dependent case, and that the SN-based test and its prewhitened version tend to have more accurate size than the KVB counterparts especially when temporal dependence is strong.

**REMARK 2.6.** The self-normalized approach to time series inference can be interpreted as an extension of the classical self-normalization in  $t$ -statistic to time series setting and it inherits a number of features possessed by the  $t$ -test: (i) The  $t$ -test is based on a pivotal quantity (or statistic), whereas our self-normalized quantity is asymptotically pivotal; (ii) There is no bandwidth parameter involved in both procedures; (iii) the reference distribution is  $t$  distribution in  $t$ -test and a function of Brownian motions in our case, so both are non-normal; (iv) the approximation to the finite sample distribution is more accurate than the existing counterpart. For iid normal data,  $t(n-1)$  provides an error-free approximation to the finite sample distribution of  $t$ -statistic under the null, which was approximated by standard normal distribution before Gosset's invention of  $t$  distribution. Similarly, the approximation of the finite sample distribution of the self-normalized quantity using  $U_1$  is more accurate than that achieved using a consistent long run variance estimator  $s_n^2$  as a studentizer and  $\chi_1^2$  as the limiting distribution [see (7) and (8)]. Compared to the  $t$ -test, a distinctive feature for the SN in the time series context is that we use inconsistent variance estimate, whereas consistent variance estimate  $S_n^2$  is used in  $t$ -statistic. Consistent variance estimation and subsequent studentization have been frequently used in statistics, but for the purpose of inference (e.g., testing and confidence interval construction), consistent variance estimate is not a necessity.

## 2.3 Some variations and simulation results

As mentioned in Shao (2010a), the self-normalizer  $W_N^2$  is not unique and the specific form (14) was influenced by Lobato (2001), since it corresponds to Lobato's mapping  $\mathcal{M}$  (see (17)) and reduces to Lobato's self-normalizer in the mean case. In this subsection, we

describe several variants of the self-normalizer and examine their finite sample performance in terms of coverage rates and average interval lengths. For the convenience of notation, we shall restrict our discussion to a scalar parameter  $\theta = T(F)$ , where  $T$  is a functional and  $F$  is the marginal cdf of a stationary univariate time series  $X_t$ . The generalization to a vector parameter and the multivariate time series setting is straightforward.

Variant 1 (SN<sub>1</sub>): a key feature of Shao's generalization of Lobato's scheme is to use recursive estimates  $\{\hat{\theta}_{1,t}\}_{t=1}^n$ , which are based on a forward scan [McElroy and Politis (2007)]. If we use a backward scan, i.e.,  $\{(x_n), (x_{n-1}, x_n), \dots, (x_1, \dots, x_n)\}$ , the self-normalizer would admit the form  $W_{1n}^2 = n^{-2} \sum_{t=1}^n t^2 (\hat{\theta}_{n-t+1,n} - \hat{\theta}_{1,n})^2$ , where  $\hat{\theta}_{i,j}$  stands for the plug in estimator of  $\theta$  based on the subsample  $(X_i, \dots, X_j)$ ,  $i \leq j$ . Then under Assumption I, (10) and

$$n^{-2} \sum_{t=1}^n (n-t+1)^2 |R_{t,n}|^2 = o_p(1), \quad (18)$$

where  $R_{t,n}$  denotes the remainder term in the expansion of  $\hat{\theta}_{t,n}$  as in (9), it can be shown that

$$\frac{n(\hat{\theta}_n - \theta)^2}{W_{1n}^2} \rightarrow_D U_1.$$

So this also naturally extends Lobato's method in a sense since in the mean case  $W_{1n}^2 = W_n^2$ . If the distributions of  $(X_1, \dots, X_n)$  and  $(X_n, \dots, X_1)$  are identical, i.e., the time series is reversible, then the distribution of  $(R_{1,n}, R_{2,n}, \dots, R_{n-1,n}, R_{n,n})$  is the same as that of  $(R_{1,n}, R_{1,n-1}, \dots, R_{1,2}, R_{1,1})$  and (18) is equivalent to (13).

Variant 2 (SN<sub>2</sub>): we use all recursive subsample estimates, i.e.,  $\hat{\theta}_{i,j}$ ,  $1 \leq i \leq j \leq n$  in the formation of self-normalizer. Let

$$W_{2n}^2 = n^{-3} \sum_{i=1}^n \sum_{j=i}^n (j-i+1)^2 (\hat{\theta}_{i,j} - \hat{\theta}_{1,n})^2.$$

By Proposition 3.1 in Volgushev and Shao (2014), we can derive under some assumptions that

$$\frac{n(\hat{\theta}_n - \theta)^2}{W_{2n}^2} \rightarrow_D \frac{B(1)^2}{\int_0^1 \int_x^1 \{B(y) - B(x) - (y-x)B(1)\}^2 dy dx} =: J_1$$

The 100 $\alpha$ th upper critical values of the limiting distribution  $J_1$ , denoted as  $J_{1,\alpha}$  are tabulated in Table 2. The simulated critical values are obtained by approximating the standard Brownian motion by standardized sum of 5000 standard normal random variables.



Please insert Table 2 here!

Variant 3 (SN<sub>3</sub>): in this variant, the self-normalizer is taken to be  $W_{3n}^2 = \{W_n^2 + W_{1n}^2\}/2$ , i.e., the average of the self-normalizers used in the original SN approach of Shao (2010a) and SN<sub>1</sub>. This idea was suggested by Professor Michael Stein in my seminar at the University of Chicago on May 21, 2012. Since  $W_{3n}^2 = W_n^2$  in the mean case, this variation can also be considered as a natural extension of Lobato's method. It then follows that

$$\frac{n(\hat{\theta}_n - \theta)^2}{W_{3n}^2} \rightarrow_D U_1$$

provided that Assumptions I and II as well as (18) hold. Another interesting point was raised by Professor Steve Lalley after my seminar in Chicago. He pointed out that the original self-normalizer  $W_n^2$  exhibits certain degree of asymmetry in the way the data  $(X_1, \dots, X_n)$  are used. This is in fact true since the  $i$ th data point  $X_i$  is used  $(n - i + 1)$  times in the recursive estimates  $\{\hat{\theta}_{1,1}, \dots, \hat{\theta}_{1,n}\}$  for  $i = 1, \dots, n$ , although different weights are assigned to  $\hat{\theta}_{1,i} - \hat{\theta}_{1,n}$  in  $W_n^2$ . A more general approach would be to incorporate a weighting scheme, say  $w(i/n)$ , where  $w(\cdot)$  is a nonnegative weight function, and form the self-normalizer as  $W_n^2 = n^{-2} \sum_{t=1}^n w(t/n) t^2 (\hat{\theta}_{1,t} - \hat{\theta}_{1,n})^2$  (see Jansson (2004)), but it is not clear how to choose the weighting function. Note that the variants SN<sub>2</sub> and SN<sub>3</sub> fulfill the symmetry requirement, although the computation associated with SN<sub>2</sub> is substantially more than that for SN<sub>3</sub>, which is twice of SN<sub>1</sub> (or SN) in computational cost.

Please insert Table 3 here!

Table 3 shows the empirical coverage rates (in percentage) and ratios of average interval widths for confidence intervals of the marginal median based on 50000 monte carlo replications. The data generating process is a Gaussian AR(1) model with  $\rho = 0.2, 0.5, 0.8, -0.4$ . Samples sizes  $n = 50$  and  $200$ . When  $\rho = -0.4$ , SN and SN<sub>1</sub> exhibit undercoverage, whereas the intervals corresponding to SN<sub>2</sub> and SN<sub>3</sub> overcover. When  $\rho$  is positive, it can be seen that the coverages of SN<sub>2</sub> and SN<sub>3</sub> are closer to the nominal level than SN and SN<sub>1</sub> in most cases, although SN<sub>3</sub> delivers the widest interval among the four. The performance of SN<sub>1</sub> is mostly comparable to that for SN, which may be due to the reversibility of the Gaussian time series. SN<sub>2</sub> not only improves the coverage accuracy of SN for most cases, but also corresponds to slightly shorter interval. The performance of SN<sub>2</sub> will be further examined in the mean case below. We would like to mention that our simulation results are quite limited and it will be a worthwhile effort to compare these four approaches for more data generating processes and other quantities of interest. The coverage accuracy

associated with  $SN_2$  and  $SN_3$  are interesting and further theoretical or simulation evidence are needed to justify the use of these variants.

As a complement to the simulation results presented in Lobato (2001) and Shao (2010a), we provide additional simulation comparison of the following approaches in the confidence interval construction for the mean of a Gaussian AR(1) time series. In particular, we consider (i) SN approach, which is equivalent to  $SN_1$  and  $SN_3$  in the mean case; (ii) traditional approach (denoted as “TR”), where the long run variance is consistently estimated using the lag window estimator (1) with the Bartlett kernel and Andrews’ data driven bandwidth  $\hat{l}_{n,opt}$  (see (2)); (iii)  $SN_2$ ; (iv) Infeasible approach, where normal approximation is used with long run variance  $s^2$  assumed to be known. Note that the finite sample variance  $n\text{var}(\bar{X}_n) \neq s^2$  for a finite  $n$ , so there is still an approximation error with this infeasible approach, leading to a deviation of the empirical coverage from the nominal level.

Please insert Table 4 here!

It can be seen from Table 4 that when  $\rho$  is positive, all methods lead to undercoverage, which corresponds to the negative leading term in the edgeworth expansion of the finite sample distribution of the studentized mean; see Zhang and Shao (2013) for more discussions. SN substantially outperforms the traditional method and even slightly outperforms the infeasible approach in terms of coverage accuracy. When  $\rho = -0.4$ , most methods exhibit overcoverage. The traditional method delivers accurate coverage rate 90.1% at the 90% nominal level when  $n = 50$ , but its empirical coverage deviates away from the nominal level from 90.1% to 91% as  $n$  increases from  $n = 50$  to 200. This is not the case for SN and other methods, for which the coverage rates generally get closer to the nominal level as sample size increases. The performance of  $SN_2$  is comparable or can be noticeably worse than SN (see, e.g., the result for  $\rho = 0.8$ ) in terms of coverage accuracy, although the average interval length of SN is slightly wider than the  $SN_2$  counterpart. Apparently, the SN-based interval is wider than other counterparts for all cases. The simulation result presented here suggests that (i) the normal approximation can perform poorly in the situation of small sample size and strong dependence even with some sensible bandwidth choice; (ii) SN can even outperform infeasible normal approximation (with long run variance assumed to be known) in terms of coverage accuracy for some cases at the price of longer interval length; (iii) the advantage of  $SN_2$  over SN seen in the median case does not carry over to the mean case.

### 3 Extensions to other inference problems for univariate time series

In this section, we present/discuss extensions of the SN approach to several important inference problems in a univariate setting, including change point detection in Section 3.1, confidence interval construction for regression parameter in regression models with fixed regressors and weakly dependent errors in Section 3.2, inference in nonparametric time series regression in Section 3.3, inference for long memory time series in Section 3.4, as well as inference for the mean of a time series with time varying second order properties in Section 3.5 and for the mean of near-integrated time series in Section 3.6.

#### 3.1 Change point detection

Let  $(X_t)_{t=1}^n$  be time series observations. We want to test the null hypothesis

$$H_0 : \mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots = \mathbb{E}(X_n)$$

against the alternative hypothesis

$$H_1 : \mathbb{E}(X_1) = \cdots = \mathbb{E}(X_{k^*}) \neq \mathbb{E}(X_{k^*+1}) = \cdots = \mathbb{E}(X_n),$$

where the location of the change point  $1 \leq k^* \leq n-1$  is unknown. There is a huge literature in econometrics and statistics on change point detection for time series; see Perron (2006) and Aue and Horváth (2013) for recent reviews. A common class of test statistics is based on the so-called CUSUM (cumulative sum) process defined as

$$T_n(\lfloor nr \rfloor) = n^{-1/2} \sum_{t=1}^{\lfloor nr \rfloor} (X_t - \bar{X}_n), \quad r \in [0, 1].$$

The celebrated Kolmogorov-Smirnov test admits the form

$$KS_n = \sup_{r \in [0, 1]} |T_n(\lfloor nr \rfloor)|/s_n = \sup_{k=1, \dots, n} |T_n(k)|/s_n,$$

where  $s_n$  is the square root of the consistent long run variance estimator defined in (1). Under the null hypothesis and assuming (3), we can show that  $KS_n \rightarrow_D \sup_{r \in [0, 1]} |\tilde{B}(r)|$ . Thus the asymptotic null distribution of  $KS_n$  is pivotal. The difficulty is the choice of  $l_n$  in

the variance estimate  $s_n^2$ . Asymptotic theory requires  $l_n/n + 1/l_n = o(1)$  but this does not give any practically useful guidance on the choice of the bandwidth. A possible strategy is to use Andrews' bandwidth selection rule that aims for a minimal mean squared error for  $s_n^2$ .

Please insert Figure 2 here!

Figure 2 shows the power of  $KS_n$  at 5% level for the following alternative:

$$X_t = \begin{cases} u_t, & 1 \leq t \leq n/2, \\ \eta + u_t, & n/2 + 1 \leq t \leq n = 200, \end{cases} \quad (19)$$

where  $u_t = 0.5u_{t-1} + \varepsilon_t$ ,  $\varepsilon_t \sim \text{iid } N(0, 1)$ . Here the bandwidth  $l_n$  is chosen by using the formulae in (2). As we see from Figure 2, the power can decrease to zero as the magnitude of change  $\eta$  gets very large, which is counterintuitive as we expect a test to be more powerful when the alternative gets farther away from the null. In the literature, this phenomenon was called nonmonotonic power [Vogelsang 1999] and its source has been identified through theoretical analysis and simulations by Vogelsang (1999), Deng and Perron (2008) and Juhl and Xiao (2009), among others. Heuristically, the decrease in power accompanied with larger shift is due to the fact that the bandwidth  $l_n$  is severely biased upward under the alternative, which leads to an inflation in the estimate of the scale  $s$ . When the scale estimate  $s_n$  is too large, the  $KS_n$  test statistic tends to be small, resulting in a low rejection rate. Note that the fixed bandwidth (e.g.  $n^{1/3}$ ) is immune to the nonmonotonic power problem but is not adaptive to the magnitude of autocorrelations in the series and it could lead to severe size distortion; see Shao and Zhang (2010) for simulation evidence.

Following the description of the SN idea in the previous section, a seemingly natural extension of the SN approach to the change point testing problem is to replace  $s_n$  in  $KS_n$  by  $W_n$ . In other words, let

$$\widetilde{KS}_n = \sup_{r \in [0, 1]} |T_n(\lfloor nr \rfloor) / W_n| = \sup_{k=1, \dots, n} |T_n(k) / W_n|.$$

Under  $H_0$  and assuming (3),  $\widetilde{KS}_n \rightarrow_D \sup_{r \in [0, 1]} |\widetilde{B}(r)| / [\int_0^1 \widetilde{B}(r)^2 dr]^{1/2}$ , which is still pivotal. However, as shown in Shao and Zhang (2010), the above naive extension fails to attenuate the non-monotonic power problem; see Figure 1 therein.

The major problem with  $\widetilde{KS}_n$  is that it does not take into account the change-point alternative. To circumvent the problem, Shao and Zhang (2010) proposed a new self-

normalization process, i.e., for  $k = 1, \dots, n-1$ ,

$$V_n(k) = n^{-2} \left[ \sum_{t=1}^k \{S_{1,t} - (t/k)S_{1,k}\}^2 + \sum_{t=k+1}^n \{S_{t,n} - (n-t+1)/(n-k)S_{k+1,n}\}^2 \right],$$

where  $S_{i,j} = \sum_{t=i}^j X_t$  for  $i \leq j$ . The SN-based test statistic was defined as

$$\widehat{KS}_n = \sup_{k=1, \dots, n-1} T_n(k)' V_n^{-1}(k) T_n(k). \quad (20)$$

Assuming (3), the limiting null distribution of  $\widehat{KS}_n$  can be derived via the continuous mapping theorem as

$$\mathcal{L}(1) := \sup_{r \in [0,1]} [B(r) - rB(1)]' V(r)^{-1} [B(r) - rB(1)], \text{ where}$$

$$V(r) = \int_0^r \{B(u) - (u/r)B(r)\}^2 du + \int_r^1 [B(1) - B(u) - \frac{1-u}{1-r}\{B(1) - B(r)\}]^2 du. \quad (21)$$

The critical values of  $\mathcal{L}(1)$  have been provided in Shao and Zhang (2010) via simulations. Note that the normalization factor  $V_n(k)$  in our test depends on  $k$ , whereas those in  $KS_n$  and  $\widehat{KS}_n$  stay the same for all  $k$ . This distinction has important implications in their power behaviors.

Through extensive simulations, the SN-based test has been shown to have superior size compared to some recently proposed alternatives by Crainiceanu and Vogelsang (2007) and Juhl and Xiao (2009), which have been developed to alleviate the nonmonotonic power problem. The power of the SN-based test was seen to be monotonic although there has been no rigorous theory that justifies the nonmonotonic power property of the SN-based test. The extension to testing for a change point in other quantities, such as marginal quantiles and spectral distribution function at certain frequencies, have also been developed in Shao and Zhang (2010).

**REMARK 3.1.** A key point we want to make in this part is that the self-normalizer used in confidence interval construction is no longer suitable for the problem of change point testing as it is not able to alleviate the non-monotonic power problem. The new self-normalizer, which takes into account one change point alternative, seems to work well as it delivers good size and respectable power. It is worth noting that our SN-based change point test

is tailored to one abrupt change. To detect gradual change or in the presence of trends, the above test is not expected to work without a nontrivial modification. Therefore, the self-normalizer depends on the problem of interest and different self-normalizers may be needed for different problems.

### 3.2 Time series regression with deterministic regressors

The SN approach in Lobato (2001) and Shao (2010a) was for stationary time series and it has been extended by Zhou and Shao (2013) to the following linear regression problem where the response variable can be nonstationary in mean. Consider the following linear model:

$$Y_t = \mathbf{x}_{t,n}'\beta + e_t, \quad t = 1, 2, \dots, n, \quad (22)$$

where  $\mathbf{x}_{t,n} = (x_{t1,n}, x_{t2,n}, \dots, x_{tp,n})'$ ,  $1 \leq t \leq n$ , are  $p \times 1$  known deterministic design vectors,  $\beta \in \mathbb{R}^p$  is the unknown regression parameter and  $e_t$  is a stationary and weakly dependent time series error process. For example, when  $\mathbf{x}_{t,n} = (1, t/n)'$ , then our interest is on estimation and inference of linear trend with dependent errors, a well-studied topic in time series analysis. In particular, it includes the stationary time series framework in Shao (2010a) as a special case, since if we let  $\mathbf{x}_{t,n} = 1$ , then  $Y_t = \beta + e_t$  are stationary. Zhou and Shao (2013) adopted an M-estimation framework, i.e.,

$$\hat{\beta}_n = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{t=1}^n M(Y_t - \mathbf{x}_{t,n}'\beta),$$

where  $M$  is a convex function. Special cases include least squares regression and quantile regression, among others. Under certain conditions on the weak dependence of  $e_t$ , the design vectors  $\{\mathbf{x}_{t,n}\}_{t=1}^n$ , and the function  $M(\cdot)$ , it can be shown that the regression parameter estimate  $\hat{\beta}_n$  is asymptotically normal. Let  $\psi(x) = M'(x)$ ,  $\varphi(x) = \mathbb{E}[\psi(e_1 + x)]$  and  $\Sigma(r) = \lim_{n \rightarrow \infty} (rn)^{-1} \sum_{j=1}^{\lfloor nr \rfloor} \mathbf{x}_{j,n} \mathbf{x}_{j,n}'$  for  $r \in (0, 1]$ . Theorem 1 in Zhou and Shao (2013) asserts that

$$\sqrt{n}(\hat{\beta}_n - \beta) \rightarrow_D N(0, \sigma_\infty^2 \Sigma^{-1}(1) / [\varphi'(0)]^2), \quad (23)$$

where  $\sigma_\infty^2 = \sum_{t=-\infty}^{\infty} \operatorname{cov}(\psi(e_0), \psi(e_t))$  is the long run variance associated with the stationary process  $\psi(e_t)$ . In the case  $M(x) = |x|$  and  $\mathbf{x}_{t,n} = 1$ ,  $\hat{\beta}_n$  corresponds to sample median

of a stationary time series. As we discussed earlier, consistent estimation of the asymptotic variance of  $\hat{\beta}_n$  in the sample median case is possible but is quite involved. Alternatively, one can use the residual block-based bootstrap method but it seems there is no theoretical/empirical study on the choice of block size in this setting. Our simulation suggests that the inference is sensitive to the block size and the optimal block size that corresponds to most accurate coverage is tied to the magnitude of dependence in the error process and the design matrix.

Naturally we seek an extension of the SN approach to the regression setting by using the idea of recursive estimation. Theorem 2 in Zhou and Shao (2013) establishes a uniform Bahadur representation for  $\hat{\beta}_{[rn]}$ ,  $r \in [\epsilon, 1]$  (i.e., the estimate of  $\beta$  based on  $(\mathbf{x}_{t,n}, Y_t)_{t=1}^{\lfloor rn \rfloor}$ ), where  $\epsilon \in (0, 1)$  is the trimming proportion. Specifically, we have

$$\sup_{\epsilon \leq r \leq 1} \left| \varphi'(0) r \sqrt{n} \Sigma(r) (\hat{\beta}_{[rn]} - \beta) - \sum_{j=1}^{\lfloor rn \rfloor} \psi(e_j) \mathbf{x}_{j,n} / \sqrt{n} \right| = o_{\mathbb{P}}(1). \quad (24)$$

Note that the uniformness is with respect to  $r \in [\epsilon, 1]$ , and for technical reasons, we were unable to prove (24) when  $\epsilon = 0$ . Based on the above result, it is quite straightforward to derive that

$$n(\hat{\beta}_n - \beta)^T \{W_n^2(\epsilon)\}^{-1} (\hat{\beta}_n - \beta) \rightarrow_D \Gamma'(1) \left\{ \int_{r=\epsilon}^1 \{\Gamma(r) - r\Gamma(1)\} \{\Gamma(r) - r\Gamma(1)\}' dr \right\}^{-1} \Gamma(1),$$

where  $W_n^2(\epsilon) = \sum_{t=\lfloor \epsilon n \rfloor}^n t^2 (\hat{\beta}_t - \hat{\beta}_n)(\hat{\beta}_t - \hat{\beta}_n)^T / n^2$  and  $\Gamma(\cdot)$  is a zero-mean Gaussian process with covariance  $\text{cov}(\Gamma(r_1), \Gamma(r_2)) = \{\Sigma(\max(r_1, r_2))\}^{-1}$ . Therefore the limiting distribution of the self-normalized quantity is free of the dependence in the error process, but it is not pivotal as it depends on the design matrix in a nontrivial manner. To perform the inference, a simulation-based approach was proposed in Zhou and Shao (2013) to approximate the nonpivotal limiting distribution. To capture the effect of the design matrix, we generate iid standard normal random variables  $Z_1, Z_2, \dots, Z_n$ . Let  $\hat{\Gamma}_r = (\sum_{j=1}^{\lfloor rn \rfloor} \mathbf{x}_{j,n} \mathbf{x}_{j,n}' / \lfloor nr \rfloor)^{-1} \sum_{j=1}^{\lfloor rn \rfloor} Z_j \mathbf{x}_{j,n} / \sqrt{n}$  and

$$D_n = \hat{\Gamma}_1' \left\{ \int_{r=\epsilon}^1 (\hat{\Gamma}_r - r\hat{\Gamma}_1)(\hat{\Gamma}_r - r\hat{\Gamma}_1)' dr \right\}^{-1} \hat{\Gamma}_1.$$

Then it can be shown that  $D_n \rightarrow_D \Gamma'(1) \left\{ \int_{r=\epsilon}^1 \{\Gamma(r) - r\Gamma(1)\} \{\Gamma(r) - r\Gamma(1)\}' dr \right\}^{-1} \Gamma(1)$ . We can repeat the generation of iid standard normal  $\{Z_t\}_{t=1}^n$  for a sufficiently large number

of times (say  $B$  times) and denote the simulated values as  $\{D_{j,n}\}_{j=1}^B$ . Confidence region or tests for  $\beta$  can be conducted by using the discrete uniform distribution over  $\{D_{j,n}\}_{j=1}^B$  as the reference distribution. Simulation results in Zhou and Shao (2013) suggest that (i) the SN-based interval is wider than the one delivered by the residual block bootstrap, but has better coverage accuracy; (ii) compared to other choices of  $\epsilon s$ ,  $\epsilon = 0.1$  delivers reasonably accurate coverage for both median regression and ordinary least squares methods and for several combinations of error types with varying degree of dependence.

**REMARK 3.2.** It is worth noting that the approach in KVB is similar to SN, but it does not involve any trimming parameter. However, KVB's approach was developed for inference in time series regression models with (approximately stationary) dynamic regressors and it does not seem applicable when the regressors are fixed. Also their framework takes advantage of the analytical form of the ordinary least squares estimate and does not allow for quantile regression. When the regressor is fixed and nonconstant, we see that the pivotalness of the limiting distribution of the SN quantity is lost. This is due to the fact that the process based on the standardized recursive estimates depend on the whole trajectory of the limiting design matrix, i.e.,  $\{\Sigma(r), r \in [\epsilon, 1]\}$ , which can be mimicked by using a simple simulation-based approach. In addition, the introduction of the trimming parameter  $\epsilon$  is necessary for theory to go through but also has an impact on the finite sample performance. We shall leave a more detailed discussion to Section 7.

### 3.3 Nonparametric time series regression

This part is based on the work of Kim, Zhao and Shao (2015). Consider the inference of the conditional mean function  $\mu(x) = \mathbb{E}(Y_i|X_i = x)$ , where  $(X_i, Y_i)_{i=1}^n$  are time series observations from a bivariate stationary process. Let

$$\hat{\mu}_n(x) = \left\{ \sum_{i=1}^n K\left(\frac{x - X_i}{b_n}\right) \right\}^{-1} \sum_{i=1}^n Y_i K\left(\frac{x - X_i}{b_n}\right)$$

be the Nadaraya-Watson kernel smoothing estimate, where  $K(\cdot)$  is a kernel function, and  $b_n > 0$  is the bandwidth parameter. Under suitable regularity conditions, we have

$$\sqrt{nb_n} \frac{\hat{\mu}_n(x) - \mu(x) - b_n^2 \rho(x)}{s(x)} \rightarrow_D N(0, 1), \quad (25)$$



where  $b_n^2 \rho(x)$  is the bias term and  $s^2(x)$  is the asymptotic variance. Write  $e_i = Y_i - \mu(X_i)$ . Then it can be derived that

$$\rho(x) = \left[ \frac{p'_X(x)\mu'(x)}{p_X(x)} + \frac{\mu''(x)}{2} \right] \int_{\mathbb{R}} u^2 K(u) du, \quad s^2(x) = \frac{\sigma^2(x)}{p_X(x)} \int_{\mathbb{R}} K^2(u) du,$$

where  $p_X(\cdot)$  is the density function of  $X_t$  and  $\sigma^2(x) = \mathbb{E}(e_i^2 | X_i = x)$ . To construct a point-wise confidence interval for  $\mu(x)$ , the traditional approach involves consistent estimation of  $s^2(x)$  through an extra smoothing procedure and normal approximation. It boils down to nonparametric estimation of  $\sigma^2(x)$  and  $p_X(x)$  using two bandwidths. The selection of bandwidths are not easy and it may involve another user-chosen number, especially for dependent data. Also normal approximation may be unsatisfactory in small samples leading to poor coverage.

To alleviate the difficulty in the traditional inference procedure, Kim et al. (2015) proposed to extend the SN approach to nonparametric setting. Their extension consists of the following steps:

(1) Bias reduction using high order kernel. For a symmetric kernel that is locally quadratic around zero, the optimal bandwidth  $b_n$  that minimizes the mean squared error is of order  $Cn^{-1/5}$ . If we use  $K^*(u) = 2K(u) - K(u/\sqrt{2})/\sqrt{2}$  and  $b_n = Cn^{-1/5}$  then we can verify  $\int_{\mathbb{R}} u^2 K^*(u) du = 0$  so the bias is asymptotically negligible relative to the variance.

(2) To avoid a direct consistent estimation of  $s^2(x)$ , we again rely upon the recursive estimates of  $\mu(x)$ , i.e.,  $\hat{\mu}_m(x)$  based on  $(X_t, Y_t)_{t=1}^m$ , where  $m = \lfloor \epsilon n \rfloor, \dots, n$ . Similar to Section 3.2,  $\epsilon \in (0, 1)$  is a trimming constant and there is technical difficulty in establishing the following functional central limit theorem with  $\epsilon = 0$ . Letting  $b_{m,n} = b_n(n/m)^{1/5}$  and  $b_n = Cn^{1/5}$  (chosen to be optimal order in practice), it was shown in Theorem 1 of Kim et al. (2015) that

$$\left\{ \frac{\sqrt{nb_n p_X(x)}}{\sigma(x)} t^{4/5} [\hat{\mu}_{\lfloor nt \rfloor}(x) - \mu(x) - b_{\lfloor nt \rfloor, n}^2 \rho(x)] \right\}_{\epsilon \leq t \leq 1} \Rightarrow \{G_t\}_{\epsilon \leq t \leq 1}, \quad (26)$$

where  $\{G_t\}_{\epsilon \leq t \leq 1}$  is a mean zero Gaussian process with covariance function given by

$$C_K(t, t') = \text{cov}(G_t, G_{t'}) = \min(t, t') \int_{\mathbb{R}} K(t^{1/5} u) K(t'^{1/5} u) du. \quad (27)$$

(3) Combining steps (1) and (2), we can use bias reduced recursive estimates to form the self-normalizer  $S_n(x) = n^{-13/10} \left\{ \sum_{m=\lfloor \epsilon n \rfloor}^n m^{8/5} |\hat{\mu}_m^*(x) - \hat{\mu}_n^*(x)|^2 \right\}^{1/2}$ , where  $\hat{\mu}_m^*(x)$  is

the kernel smoothing estimate using the high order kernel  $K^*(\cdot)$  and bandwidth  $b_{m,n}$ . Then it follows by a straightforward argument that

$$\xi_n(x) := \frac{\hat{\mu}_n^*(x) - \mu(x)}{S_n(x)} \rightarrow_D \frac{G_1^*}{\sqrt{\int_{\epsilon}^1 |G_t^* - t^{4/5} G_1^*|^2 dt}} = \xi, \quad (28)$$

where  $\{G_t^*\}_{t=\epsilon}^1$  is a mean zero Gaussian process with covariance function  $C_{K^*}$ . Consequently, an asymptotic  $100\alpha\%$  confidence interval for  $\mu(x)$  is  $\hat{\mu}_n^*(x) \pm q_{\alpha} S_n(x)$ , where  $q_{\alpha}$  is the  $100\alpha\%$  quantile of  $|\xi|$ . Thus for a given kernel function and  $\epsilon$ , the limiting distribution of the self-normalized quantity  $\xi_n(x)$ , i.e.,  $\xi$  is pivotal, and its critical values have been provided via simulations in Kim et al. (2015) for  $\epsilon = 0.1$  and standard Gaussian kernel. Simulations in Kim et al. (2015) showed that the SN approach is vastly superior to the traditional counterpart, where the extra smoothing is applied to consistently estimate the unknown nuisance parameter in asymptotic variance, in terms of coverage accuracy.

**REMARK 3.3.** Compared to the SN approach in Shao (2010a), the main difference in its extension to nonparametric inference problems is twofold: (i) The partial sum process based on recursive estimates in Shao (2010a) has a standard Brownian motion (with stationary and independent increments) limit up to a scaling constant, whereas the limit for the process based on our standardized nonparametric recursive estimates is a Gaussian process with nonstationary and dependent increments; (ii) In Shao (2010a), the SN-based inference method does not require any bandwidth parameter, whereas in nonparametric setting, we need to estimate conditional mean function using the bandwidth parameter  $b_n$ , which seems necessary. What the SN approach does is to avoid consistent variance estimation, which can be difficult because of the involvement of extra bandwidths. The use of recursive estimates of the nonparametric mean function as the self-normalizer again deliver a pivotal limit, although the limit is no longer a function of Brownian motion. Again, a trimming parameter  $\epsilon$  needs to be introduced and its optimal choice is desired but seems difficult.

### 3.4 Long memory time series

The extension of the SN method to long memory time series is not without complications. To illustrate the idea, we first consider the inference of the mean  $\mu$  of a strictly stationary long range dependent (LRD, hereafter) process  $\{X_t\}_{t \in \mathbb{Z}}$ . For a LRD process with finite

second moments,

$$\gamma(k) = \text{cov}(X_0, X_k) = k^{-\kappa} L(k), \quad k \geq 1,$$

where  $\kappa \in (0, 1)$  and  $L(\cdot)$  is a slowly varying function at infinity. Therefore the long run variance  $\sum_{k=-\infty}^{\infty} \gamma(k) = \infty$  and correspondingly, the spectral density function blows up at zero frequency. An equivalent way of characterizing the long memory phenomenon is in terms of the local behavior of the integrable spectral density function  $\{f(\lambda), |\lambda| \leq \pi\}$  around the origin. Typically,

$$f(\lambda) \sim C_d |\lambda|^{-2d}, \quad \text{as } \lambda \rightarrow 0,$$

where  $d = 1/2 - \kappa/2 \in (0, 1/2)$ ,  $C_d > 0$  and the symbol " $\sim$ " means that the ratio of the terms on the two sides converges to one. In the literature, a widely used class of processes that allows for long-range dependence is the so-called linear processes. Write  $X_t = \mu + \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}$ , where  $\varepsilon_k$  are iid random variables with zero mean and finite second moment. For linear processes, it can be shown that

$$n^{1/2-d}(\bar{X}_n - \mu) \rightarrow_D \sigma N(0, 1)$$

for some  $\sigma > 0$ . In other words, the convergence rate of sample mean is slower than  $n^{1/2}$  due to the long range dependence. To perform the inference based on normal approximation, we need to estimate  $d$  and  $\sigma$  consistently. HAC estimator of  $\sigma$  with long memory correction (or the so-called MAC, memory and autocorrelation consistent estimator) has been proposed by Robinson (2005) and further studied in Abadir, Distaso and Giraitis (2009). Let  $\hat{d}$  and  $\hat{\sigma}$  be the estimator of  $d$  and  $\sigma$  respectively. Under the assumption that  $\hat{d} - d = o_p(1/\log(n))$  and some other mild conditions, Abadir et al. (2009) provided second order expansions for both HAC and MAC estimators and showed that  $n^{1/2-\hat{d}}(\bar{X}_n - \mu)/\hat{\sigma} \rightarrow_D N(0, 1)$  for the MAC estimator  $\hat{\sigma}$ . For the HAC estimator, the asymptotic normality holds only when the long memory index  $d$  is below some threshold. It is worth noting that the asymptotic framework adopted in Abadir et al. (2009) is small- $b$  asymptotics.

The extension of the self-normalized approach to the long memory case hinges on the following functional CLT

$$n^{-1/2-d} \sum_{j=1}^{\lfloor nr \rfloor} (X_j - \mu) \rightarrow_D \sigma B(r; d),$$

where  $B(r; d)$  is the fractional Brownian motion. For a linear process, it holds provided that  $a_k \sim k^{d-1}L_1(k)$ , where  $L_1(\cdot)$  is a slowly varying function at infinity [Davidson and de Jong (2000)]. Also see Wu and Shao (2006) and the references therein. By the continuous mapping theorem,

$$H_n(\mu) := \frac{n(\bar{X}_n - \mu)^2}{n^{-2} \sum_{t=1}^n (S_t - t\bar{X}_n)^2} \rightarrow_D \frac{B(1; d)^2}{\int_0^1 \{B(r; d) - rB(1; d)\}^2 dr} \equiv H_d,$$

therefore the limiting distribution  $H_d$  depends on  $d$  and is not pivotal. To perform the inference, one way is to estimate  $d$  by  $\hat{d}$  using log periodogram regression [Geweke and Porter-Hudak (1983)] or local Whittle estimation [Künsch (1987)], and use the simulated critical values corresponding to the distribution of  $H_{\hat{d}}$ . Alternatively, one can use a subsampling approach to approximate the distribution of  $H_d$ ; see Fan (2010) and Jach, McElroy and Politis (2012). Note that Fan (2010) also considered another class of long memory processes, namely, transformation of Gaussian processes, for which studentized sample mean may not be asymptotically normal [Taqqu (1975, 1979), Dobrushin and Major (1979)].

As mentioned in Shao (2010a), the SN method is a special case of the fixed- $b$  approach [Kiefer and Vogelsang (2005)] in the mean case. McElroy and Politis (2012, 2014) have extended the fixed- $b$  approach to the long memory setting for the inference of mean, spectral density and spectral distribution function. Their results show that the limiting distributions of the studentized sample mean as well as spectral density and distribution estimates depend on the kernel, the magnitude of memory, and the taper under the fixed- $b$  asymptotics.

The SN approach has also been extended to inference for the parameter vector in stationary long memory time series models with dependent errors. A prominent example of such models is the so-called FARIMA-GARCH (Fractional AutoRegressive Integrated Moving Average - Generalized Autoregressive Conditional Heteroscedastic) model. In Shao (2012), the inference of the parameter in the FARIMA part was studied while the GARCH error was treated nonparametrically. The popular Whittle likelihood estimator (i.e., frequency domain maximum likelihood estimator) is asymptotically normal but the asymptotical variance admits a complicated form. In particular, it depends on the fourth order cumulant spectrum of the unobserved error process and no consistent estimation has been addressed in the long memory context. The difficulty can be circumvented by adopting the SN approach, where no consistent estimation of asymptotic variance is required. As demonstrated in Shao (2012), the SN-based method is able to produce reasonable coverage

at a large sample size for both short and long time series models with various types of errors in comparison with the residual block bootstrap method. An interesting feature of this work in the long memory context is that no bandwidth parameter is involved in the SN-based inference and the limiting distribution of the SN quantity is still  $U_q$ . This is different from inference for the mean of a LRD process, where the limiting distribution of the SN quantity is non-pivotal and the approximation requires the use of bandwidth-dependent plug-in or resampling approaches.

### 3.5 Locally stationary time series

In Rho and Shao (2013b), we studied the estimation and SN-based inference of locally stationary time series regression models with fixed regressors and locally stationary errors. To illustrate the difference from the use of the SN method for stationary time series, we shall consider the following simple locally stationary time series model,

$$X_{t,n} = \mu + e_{t,n}, \quad t = 1, \dots, n,$$

where the mean  $\mu$  is constant,  $e_{t,n} = G(t/n, \mathcal{F}_t)$  is a mean zero locally stationary series. Here  $\mathcal{F}_t = (\dots, \epsilon_{t-1}, \epsilon_t)$ , where  $\epsilon_t$  are iid random variables,  $G : [0, 1] \times \mathbb{R}^\infty \rightarrow \mathbb{R}$  is a measurable function such that  $G(r, \mathcal{F}_t)$  is a properly defined random variable for all  $r \in [0, 1]$ . This framework was first introduced by Zhou and Wu (2009) to allow both nonlinearity and local stationarity, and it covers a wide range of locally stationary processes.

A natural estimator of  $\mu$  is the sample mean  $\bar{X}_n$ , which is also the ordinary least squares estimator. Under certain regularity conditions, we have that

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow_D N(0, \sigma^2)$$

where  $\sigma^2 = \int_0^1 \sigma^2(r) dr$ , with  $\sigma^2(r) = \sum_{k=-\infty}^{\infty} \text{cov}(G(r, \mathcal{F}_0), G(r, \mathcal{F}_k))$  being the (local) long run variance of  $G(r, \mathcal{F}_t)$ . To adopt the SN method, the key is to derive a functional central limit theorem for the partial sum process. Under suitable conditions, it can be shown that

$$n^{-1/2} \sum_{t=1}^{\lfloor nr \rfloor} (X_t - \mu) \Rightarrow B(r, \sigma(\cdot)) := \int_0^r \sigma(u) dB(u).$$

Let  $S_{t,n} = \sum_{j=1}^t X_{j,n}$ . Then the SN quantity takes the form

$$H_n(\mu) := \frac{\{\sqrt{n}(\bar{X}_n - \mu)\}^2}{n^{-2} \sum_{t=1}^n (S_{t,n} - t\bar{X}_n)^2}$$

and its limiting distribution is

$$H(\sigma(\cdot)) := \frac{B(1, \sigma(\cdot))^2}{\int_0^1 (B(r, \sigma(\cdot)) - rB(1, \sigma(\cdot)))^2 dr}, \quad (29)$$

which depends on the unknown infinitely dimensional nuisance parameter  $\{\sigma(u), u \in [0, 1]\}$ , and is thus not pivotal.

To approximate the distribution of  $H(\sigma(\cdot))$ , one can use the dependent wild bootstrap [Shao (2010b)]. In particular, let  $\hat{e}_{t,n} = X_{t,n} - \bar{X}_n$  be the residuals and the bootstrap sample  $X_{t,n}^* = \bar{X}_n + \hat{e}_{t,n}W_{t,n}$ , where  $\{W_{t,n}\}_{t=1}^n$  is a realization of mean zero  $l$ -dependent stationary time series. Here  $l = l(n)$  is a bandwidth parameter which plays a similar role as the block size in the block-based bootstrap. In practice, a convenient way to generate  $(W_{t,n})_{t=1}^n$  is to simulate a multivariate Gaussian vector with mean zero and  $\text{cov}(W_{t,n}, W_{t',n}) = a\{(t-t')/l\}$  for some kernel function  $a(\cdot)$  (say, the Bartlett kernel). Under some regularity conditions, it can be shown that

$$n^{-1/2} \sum_{t=1}^{\lfloor nr \rfloor} (X_{t,n}^* - \bar{X}_n) \Rightarrow B(r, \sigma(\cdot)) \text{ in probability.}$$

Let  $S_{t,n}^* = \sum_{j=1}^t X_{j,n}^*$  and  $\bar{X}_n^* = n^{-1} S_{n,n}^*$ . Using the continuous mapping theorem, we then have

$$H_n^* := \frac{\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)\}^2}{n^{-2} \sum_{t=1}^n (S_{t,n}^* - t\bar{X}_n^*)^2} \rightarrow_D H(\sigma(\cdot)) \text{ in probability,}$$

and the inference can be conducted based on the bootstrap approximation. In Rho and Shao (2013b), the consistency of the dependent wild bootstrap was rigorously justified under a more general regression setting.

A special case of the locally stationary process is

$$e_{t,n} = v(t/n)\mathcal{Z}_t, \quad (30)$$

where  $\mathcal{Z}_t = G(\mathcal{F}_t)$  is a stationary sequence with mean zero and  $v(\cdot)$  is a deterministic function on  $[0, 1]$  that captures the time varying unconditional heteroscedasticity. It was called modulated stationary process in Zhao (2011) and Zhao and Li (2013). In this case,  $\sigma^2(u) = v^2(u)V_{\mathcal{Z}}^2$ , where  $V_{\mathcal{Z}}^2 = \sum_{k=-\infty}^{\infty} \text{cov}(\mathcal{Z}_0, \mathcal{Z}_k)$  is the long run variance of the stationary process  $\mathcal{Z}_t$ . Consequently,  $B(r, \sigma(\cdot)) = V_{\mathcal{Z}} \int_0^r v(u)dB(u)$  and

$$H(\sigma(\cdot)) = \frac{\{\int_0^1 v(u)dB(u)\}^2}{\int_0^1 \{\int_0^r v(u)dB(u) - r \int_0^1 v(u)dB(u)\}^2 dr},$$

where  $V_Z$  is canceled out. Thus the temporal dependence in the error does not affect the limiting distribution of the SN quantity and the only source of non-pivotalness in the limiting distribution lies in the heteroscedasticity, which can be captured by using the traditional wild bootstrap [Wu (1986), Liu (1988)]. The consistency of wild bootstrap has been proved in Rho and Shao (2015) in a more general regression setting and for both modulated stationary process and heteroscedastic linear process [Cavaliere (2005), Cavaliere and Taylor (2007)].

**REMARK 3.4.** As we see from the discussions in Section 3.4 and Section 3.5, the long memory and time varying second order properties of a time series can make the limiting distribution of the self-normalized mean nonpivotal, depending on the magnitude of long memory or the whole long run variance function. The inference is thus made feasible by consistently approximating the nonpivotal limiting distributions using either subsampling or bandwidth-dependent bootstrap (e.g., the dependent wild bootstrap). When the locally stationary error process admits a modulated form (30), the limiting distribution of the SN quantity only depends on the heteroscedasticity but not on temporal dependence, so the bandwidth-free wild bootstrap suffices for a consistent approximation.

### 3.6 Near-integrated time series

In econometrics, time series of strong autocorrelation is often modeled as an autoregressive process with unit root or near unit root. Such nonstationary processes play an important role in the modeling and inference of econometric time series as most economic time series exhibit certain degree of unit root nonstationarity, as evident from the shape of their spectral densities [Granger (1966)]. As seen from the simulation results in Shao (2010a) and Kiefer and Vogelsang (2005), the coverage accuracy of the SN-based interval and the size accuracy of the fixed- $b$  based HAC test deteriorate as the positive autocorrelation gets stronger; see Zhang and Shao (2013) for some theoretical explanations based on edge-worth expansion of the distribution function of the studentized mean. Thus the fixed- $b$  approach, which was developed under weak dependence assumptions, may still lead to severely oversized tests when the autocorrelation is as strong as that of a near-integrated process.

For a location model with strongly dependent errors expressed as a Gaussian AR(1) model with near unit root, Müller (2014) proposed a class of tests for the mean that are

robust to strong autocorrelation and maximize a weighted average power criterion. Further extensions to regressions and models estimated by generalized method of moments are also mentioned in his paper. In Sun (2014, 2015), an alternative testing approach was developed to deal with strong autocorrelation by extending the fixed- $b$  asymptotic paradigm. The idea is to model the highly autocorrelated time series as a AR(1) process with near unit root, i.e., the AR(1) coefficient  $\rho_n = 1 - c_n/n$ . By fixing the local-to-unity parameter  $c_n = c$ , he derived the limit of the standardized partial sum process as a functional of the Ornstein-Uhlenbeck process. Under this so-called near-unity fixed-smoothing (i.e., fixed- $b$ , where  $b = l_n/n$  is treated as the smoothing parameter) asymptotics, he derived the limits of the  $F$ -test statistic and  $t$ -test statistic, which depend on both  $c$  and  $b$ . Since in practice  $b$  is known, and  $c$  can be estimated (see Sun (2015)), the critical values of the pivotal limiting distribution for a given pair of  $(b, c)$  can be approximated by simulation. It is worth mentioning that the near-unity fixed-smoothing asymptotics provides a smooth transition from the usual stationary fixed-smoothing asymptotics to the unit root fixed-smoothing asymptotics, and seems a natural extension of the fixed- $b$  asymptotics to the near unit root setting.

It is natural to ask whether the SN approach is still applicable to inference for near unit root process. The answer is unfortunately negative, at least in the setting of Müller (2014) and Sun (2015). However, as shown in Rho and Shao (2013a), the SN approach may still work after applying prewhitening to filter out strong autocorrelation. In practice, the choice of prewhitening filter is nontrivial and its effectiveness requires additional assumptions on the data generating process. To conclude, I would like to quote Müller (2014) and Vogelsang (2014) by saying that “Robust inference when autocorrelation is strong is not easy, and each method requires a set of assumptions in order for it to work well. In practice, the researchers in the practical field have to judge which set of regularity conditions makes the most sense for a specific problem”.

## 4 Functional time series

So far the discussion on the SN approach and its extensions has been limited to the univariate (or multivariate) time series setting. This section concerns inference for functional time series, or temporally dependent functional data. In Zhang, Shao, Hayhoe and Wuebbles (2011), the SN approach has been extended to test for a change point in the mean



function. The basic idea is to first apply functional principal component analysis (FPCA, hereafter) to project the data onto a space spanned by the first few principal components, and then apply the SN approach to the multivariate time series associated with principal component scores. In Zhang and Shao (2014a), the two sample inference problem was addressed using the SN approach, which is reviewed in Section 4.2.

#### 4.1 Testing for a change point in the mean function

Given a set of temporally dependent functional observations  $\{X_i(u)\}_{i=1}^N$ , we are interested in testing whether the mean function remains constant over time, i.e.,

$$H_{0,1} : \mathbb{E}[X_1(u)] = \mathbb{E}[X_2(u)] = \cdots = \mathbb{E}[X_N(u)], \quad u \in \mathcal{I}, \quad (31)$$

where  $\mathcal{I} = [0, 1]$  for simplicity. Under the null, we can write  $X_i(u) = \mu_1(u) + Y_i(u)$  with  $\mathbb{E}[Y_i(u)] = 0$ ,  $i = 1, 2, \dots, N$ . Under the alternative  $H_{a,1}$ , we assume there is a change point in the mean function, i.e.,

$$X_i(u) = \begin{cases} \mu_1(u) + Y_i(u) & 1 \leq i \leq k^*; \\ \mu_2(u) + Y_i(u) & k^* < i \leq N, \end{cases} \quad (32)$$

where  $k^* = \lfloor N\lambda \rfloor$  is an unknown change point for some  $\lambda \in (0, 1)$ ,  $\{Y_i(u)\}$  is a zero-mean functional sequence, and  $\mu_1(\cdot) \neq \mu_2(\cdot)$ .

To describe our methodology, we first introduce some useful notation commonly adopted in the literature of functional data. We consider the Hilbert space  $\mathbb{H}$  of square integrable functions defined on  $\mathcal{I} = [0, 1]$ . For any  $f, g \in \mathbb{H}$ , we define the inner product between  $f$  and  $g$  as  $\langle f, g \rangle = \int_{\mathcal{I}} f(u)g(u)du$  and denote  $\|\cdot\|$  as the corresponding norm, i.e.,  $\|f\| = \langle f, f \rangle^{1/2}$ . Let  $L^p$  be the space of real valued random variables with finite  $L^p$  norm, i.e.,  $(\mathbb{E}|X|^p)^{1/p} < \infty$ . We further define  $L_{\mathbb{H}}^p$  the space of  $\mathbb{H}$  valued random variables  $X$  such that  $v_p(X) := (\mathbb{E}\|X\|^p)^{1/p} < \infty$ . For  $X(\cdot) \in L_{\mathbb{H}}^2$ , let  $c(u, v) = \text{cov}\{X(u), X(v)\}$ ,  $u, v \in \mathcal{I}$  denote the covariance function. Then Mercer's Lemma implies that  $c(u, v)$  admits the spectral decomposition,

$$c(u, v) = \sum_{j=1}^{\infty} \lambda_j \phi_j(u) \phi_j(v), \quad (33)$$

where  $\lambda_j$  and  $\phi_j$  are the eigenvalue and eigenfunction respectively. The eigenvalues are ordered so that  $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ . By the Karhunen-Loève expansion [Bosq (2000, p. 26)],

we can write  $X_i(u) = \mathbb{E}[X_i(u)] + \sum_{j=1}^{\infty} \eta_{i,j} \phi_j(u)$ , where  $\{\eta_{i,j}\}$  are the principal components (scores) defined by  $\eta_{i,j} = \int_{\mathcal{I}} \{X_i(u) - \mathbb{E}[X_i(u)]\} \phi_j(u) du$ . A natural sample estimator of the covariance function  $c(u, v)$  is

$$\hat{c}(u, v) = \frac{1}{N} \sum_{i=1}^N \{X_i(u) - \bar{X}_N(u)\} \{X_i(v) - \bar{X}_N(v)\}, \quad (34)$$

where  $\bar{X}_N(u) = \frac{1}{N} \sum_{i=1}^N X_i(u)$  is the sample mean function. The eigenfunctions and corresponding eigenvalues of  $\hat{c}(u, v)$  are defined by

$$\int_{\mathcal{I}} \hat{c}(u, v) \hat{\phi}_j(v) dv = \hat{\lambda}_j \hat{\phi}_j(u), \quad j = 1, 2, \dots. \quad (35)$$

Then the empirical scores are given by

$$\hat{\eta}_{i,j} = \int_{\mathcal{I}} \{X_i(u) - \bar{X}_N(u)\} \hat{\phi}_j(u) du, \quad i = 1, 2, \dots, N; \quad j = 1, 2, \dots, K,$$

where  $K$  is the number of principal components we consider and is assumed to be fixed. In practice,  $K$  can be chosen by  $K = \inf\{J : \sum_{i=1}^J \hat{\lambda}_i / \sum_{i=1}^m \hat{\lambda}_i > \alpha\}$ , where  $m$  is the number of basis functions in smoothing and  $\alpha$  is a pre-specified number, say 85%.

Under the null, the score vector  $\eta_i = (\eta_{i,1}, \eta_{i,2}, \dots, \eta_{i,K})'$ ,  $i = 1, 2, \dots, N$  has a constant mean, whereas the mean changes under the alternative. Let  $\hat{\eta}_i = (\hat{\eta}_{i,1}, \dots, \hat{\eta}_{i,K})'$  and  $S_{N,\hat{\eta}}(t_1, t_2) = \sum_{i=t_1}^{t_2} \hat{\eta}_i$ , for  $1 \leq t_1 \leq t_2 \leq N$ . We can then define the CUSUM process as

$$T_{N,\hat{\eta}}(k, K) := \frac{1}{\sqrt{N}} \left\{ S_{N,\hat{\eta}}(1, k) - \frac{k}{N} S_{N,\hat{\eta}}(1, N) \right\}, \quad k = 1, 2, \dots, N. \quad (36)$$

The idea of testing for a change point in the mean function using the principal component scores was first proposed in Berkes, Gabrys, Horváth and Kokoszka (2009), who dealt with independent functional data. In Hörmann and Kokoszka (2010), a consistent long run variance estimator for the principal component scores was introduced to allow for weak dependence in functional time series. Zhang et al. (2011) extended the SN-based test developed in Shao and Zhang (2010) to functional setting by defining the self-normalized

process as

$$\begin{aligned}
V_{N,\hat{\eta}}(k, K) := & \frac{1}{N^2} \left[ \sum_{t=1}^k \left\{ S_{N,\hat{\eta}}(1, t) - \frac{t}{k} S_{N,\hat{\eta}}(1, k) \right\} \left\{ S_{N,\hat{\eta}}(1, t) - \frac{t}{k} S_{N,\hat{\eta}}(1, k) \right\}' \right. \\
& + \sum_{t=k+1}^N \left\{ S_{N,\hat{\eta}}(t, N) - \frac{N-t+1}{N-k} S_{N,\hat{\eta}}(k+1, N) \right\} \\
& \left. \left\{ S_{N,\hat{\eta}}(t, N) - \frac{N-t+1}{N-k} S_{N,\hat{\eta}}(k+1, N) \right\}' \right], \quad k = 1, \dots, N-1,
\end{aligned} \tag{37}$$

and the SN-based test statistic as

$$\begin{aligned}
G_{N,\hat{\eta}}(K) &= \sup_{k=1,2,\dots,N-1} \{ T_{N,\hat{\eta}}(k, K)' V_{N,\hat{\eta}}^{-1}(k, K) T_{N,\hat{\eta}}(k, K) \} \\
&= \mathcal{C}(N^{-1/2} S_{N,\hat{\eta}}(1, \lfloor Nr \rfloor), r \in [0, 1]),
\end{aligned} \tag{38}$$

where  $\mathcal{C}$  is the implicitly defined continuous mapping that corresponds to  $G_{N,\hat{\eta}}(K)$ . Under suitable assumptions, it can be shown that under the null hypothesis,

$$G_{N,\hat{\eta}}(K) \rightarrow_D \mathcal{L}(K) := \sup_{r \in [0,1]} \{ B_K(r) - r B_K(1) \}' V_K^{-1}(r) \{ B_K(r) - r B_K(1) \}, \tag{39}$$

where  $V_K(r)$  is the counterpart of  $V(r)$  (see (21)) with  $B(r)$  replaced by  $B_K(r)$ . The proof uses several important results presented in Hörmann and Kokoszka (2010) for a stationary weakly dependent functional sequence. In particular, let  $S_{N,\eta}(t_1, t_2) = \sum_{i=t_1}^{t_2} \eta_i$ . Under certain moment and weak dependence assumptions on  $(X_t(\cdot))_{t \in \mathbb{Z}}$ , Hörmann and Kokoszka (2010) showed that

$$\sup_{r \in [0,1]} \frac{1}{\sqrt{N}} |S_{N,\eta}(1, \lfloor Nr \rfloor) - S_{N,\hat{\eta}}(1, \lfloor Nr \rfloor)| = o_p(1),$$

which suggests that the estimation effect incurred by replacing  $\eta$  with  $\hat{\eta}$  is asymptotically negligible and it does not enter into the first order limiting distribution of the SN-based test statistic, as seen from (39).

Simulation results suggest that the SN-based test has less size distortion and stable size with respect to the choice of  $K$ , in comparison with the standard method developed in Hörmann and Kokoszka (2010). On the other hand, there is a power loss associated with the SN-based test, but the loss is fairly moderate. Similar to the univariate case, the nonmonotonic power problem was observed for Hörmann and Kokoszka's method when the bandwidth is chosen using Andrews' rule, but not for the SN-based test.

## 4.2 Two sample inference

Given two sequences of temporally dependent functional observations,  $\{X_i(u)\}_{i=1}^{N_1}$  and  $\{Y_i(u)\}_{i=1}^{N_2}$  defined on a common region  $\mathcal{I}$ , we are interested in comparing their second order properties. Suppose that the functional time series are second order stationary and that  $\mathbb{E}[X_i(u)] = \mathbb{E}[Y_i(u)] = 0$ . Define  $C_X = \mathbb{E}[\langle X_i, \cdot \rangle X_i]$  and  $C_Y = \mathbb{E}[\langle Y_i, \cdot \rangle Y_i]$  as the covariance operators of the two sequences respectively. For the convenience of presentation, we shall use the same notation for the covariance operator and the associated covariance function. Denote by  $\{\phi_X^j\}_{j=1}^\infty$  and  $\{\lambda_X^j\}_{j=1}^\infty$  the eigenfunctions and eigenvalues of  $C_X$ . Analogous quantities are  $\{\phi_Y^j\}_{j=1}^\infty$  and  $\{\lambda_Y^j\}_{j=1}^\infty$  for the second sample.

In Zhang and Shao (2014a), we considered testing the hypothesis that the two covariance operators or their associated eigenvalues and eigenfunctions are the same. The same testing problems have been considered in Benko, Härdle and Kneip (2009) but for two independent functional sequences with iid observations, where the authors used an iid bootstrap method which seems not applicable to the dependent case. To illustrate the idea, we focus on the test for the equality of the first few eigenvalues. By the Karhunen-Loève expansion (Bosq, 2000, p.26), we have

$$X_i(u) = \sum_{j=1}^{+\infty} \sqrt{\lambda_X^j} \beta_{X_i,j} \phi_X^j(u), \quad Y_i(u) = \sum_{j=1}^{+\infty} \sqrt{\lambda_Y^j} \beta_{Y_i,j} \phi_Y^j(u),$$

where  $\sqrt{\lambda_X^j} \beta_{X_i,j} = \int_{\mathcal{I}} X_i(u) \phi_X^j(u) du$  and  $\sqrt{\lambda_Y^j} \beta_{Y_i,j} = \int_{\mathcal{I}} Y_i(u) \phi_Y^j(u) du$  are the principal components (scores), which satisfy that  $\mathbb{E}[\beta_{X_i,j} \beta_{X_i,j'}] = \delta_{jj'}$  and  $\mathbb{E}[\beta_{Y_i,j} \beta_{Y_i,j'}] = \delta_{jj'}$ . For a prespecified positive integer  $K$ , we denote the vector of the first  $K$  eigenvalues by  $\lambda_X^{1:K} = (\lambda_X^1, \dots, \lambda_X^K)'$  and  $\lambda_Y^{1:K} = (\lambda_Y^1, \dots, \lambda_Y^K)'$ . We aim to test the null hypothesis  $H_{2,0} : \lambda_X^{1:K} = \lambda_Y^{1:K}$  versus the alternative  $H_{2,a} : \lambda_X^{1:K} \neq \lambda_Y^{1:K}$ .

Let  $N = N_1 + N_2$ . Recall that the recursive estimates of the eigenvalues  $\hat{\lambda}_{X,m}^j$  and  $\hat{\lambda}_{Y,m'}^j$  can be calculated based on the subsamples  $\{X_i(u)\}_{i=1}^m$  and  $\{Y_i(u)\}_{i=1}^{m'}$ . Let  $\hat{\theta}_k^j = \hat{\lambda}_{X, \lfloor kN_1/N \rfloor}^j - \hat{\lambda}_{Y, \lfloor kN_2/N \rfloor}^j$  and  $\hat{\theta}_k = (\hat{\theta}_k^1, \dots, \hat{\theta}_k^K)'$  with  $\lfloor N\epsilon \rfloor \leq k \leq N$  for some  $\epsilon \in (0, 1]$ , which is held fixed in the asymptotics. We consider the trimmed SN-based test statistic

$$G_{SN,N}(K) = N^3 \hat{\theta}_N' \left\{ \sum_{k=\lfloor N\epsilon \rfloor}^N k^2 (\hat{\theta}_k - \hat{\theta}_N)(\hat{\theta}_k - \hat{\theta}_N)' \right\}^{-1} \hat{\theta}_N. \quad (40)$$

For some technical reason, the trimming in the self-normalizer is required if the functional space in which the data lie is infinite dimensional. It can be seen from the proof in

Zhang and Shao (2014a) that the trimming is not necessary (i.e.,  $\epsilon$  can be set as 0) when functional data lie on a finite dimensional space. Under suitable technical conditions on the processes  $X_i(u)$  and  $Y_i(u)$ , the limiting null distribution of  $G_{SN,N}(K)$  can be shown to be  $B_K(1)'J_K(\epsilon)^{-1}B_K(1)$ , where  $J_K(\epsilon) = \int_{\epsilon}^1 (B_K(r) - rB_K(1))(B_K(r) - rB_K(1))'dr$ . Further the consistency of the test is also shown in Zhang and Shao (2014a).

A distinctive feature of this work is that it allows the dependence between two samples when the sample size  $N_1 = N_2$  (or approximately so). It is natural to ask if the SN method can be extended to the unbalanced two sample case. Unfortunately, as indicated by the following example for the two sample inference for the mean of a bivariate time series, the extension works only when the two samples are independent.

**EXAMPLE 4.1.** Consider two stationary time series  $\{X_i\}_{i=1}^{N_1}$  and  $\{Y_i\}_{i=1}^{N_2}$ . Let  $Z_k = \bar{X}_{\lfloor kN_1/N \rfloor} - \bar{Y}_{\lfloor kN_2/N \rfloor}$ ,  $\bar{X}_k = \sum_{i=1}^k X_i/k$ ,  $\bar{Y}_k = \sum_{i=1}^k Y_i/k$  and  $N = N_1 + N_2$ . The SN-based statistic for testing the equality of the means are given by

$$SN_{N_1, N_2} = \frac{NZ_N^2}{\frac{1}{N^2} \sum_{k=1}^N k^2 (Z_k - Z_N)^2}.$$

Assume the invariant principle holds for the bivariate process  $\{(X_i, Y_i)\}_{i \in \mathbb{Z}}$ , i.e.,

$$\frac{1}{\sqrt{N_0}} \sum_{i=1}^{\lfloor N_0 r \rfloor} \begin{pmatrix} X_i - \mu_X \\ Y_i - \mu_Y \end{pmatrix} \Rightarrow \begin{pmatrix} a & 0 \\ -c & -b \end{pmatrix} \begin{pmatrix} B^{(1)}(r) \\ B^{(2)}(r) \end{pmatrix}, \quad N_0 \rightarrow +\infty,$$

where  $B^{(1)}(r)$  and  $B^{(2)}(r)$  are two independent standard Brownian motions, and  $a, b \neq 0$  and  $c$  are unknown. Suppose  $N_1/N \rightarrow \gamma_1$  and  $N_2/N \rightarrow \gamma_2$ , where  $0 < \gamma_1, \gamma_2 < 1$ . By the continuous mapping theorem (from  $D^2[0, 1]$  to  $D^1[0, 1]$ ), we have that under the null

$$\begin{aligned} kZ_{\lfloor k \rfloor} / \sqrt{N} &= \frac{k}{\lfloor kN_1/N \rfloor} \left( \frac{1}{\sqrt{N}} \sum_{j=1}^{\lfloor kN_1/N \rfloor} X_j \right) - \frac{k}{\lfloor kN_2/N \rfloor} \left( \frac{1}{\sqrt{N}} \sum_{j=1}^{\lfloor kN_2/N \rfloor} Y_j \right) \\ &\Rightarrow V_{\gamma_1, \gamma_2}(r; a, b, c) := \frac{a}{\gamma_1} B^{(1)}(\gamma_1 r) + \frac{c}{\gamma_2} B^{(1)}(\gamma_2 r) + \frac{b}{\gamma_2} B^{(2)}(\gamma_2 r), \end{aligned}$$

with  $k = Nr$  for  $r \in (0, 1]$ . It therefore follows that

$$SN_{N_1, N_2} \rightarrow_D \frac{V_{\gamma_1, \gamma_2}^2(1; a, b, c)}{\int_0^1 (V_{\gamma_1, \gamma_2}(r; a, b, c) - rV_{\gamma_1, \gamma_2}(1; a, b, c))^2 dr},$$

where in general  $V_{\gamma_1, \gamma_2}(r; a, b, c)$  is not a Brownian motion and the limiting distribution depends on the unknown dependence structure. Notice that when  $\gamma_1 = \gamma_2$  (or  $c = 0$ ),  $V_{\gamma_1, \gamma_2}(r; a, b, c) = \frac{a+c}{\gamma_1}B^{(1)}(\gamma_1 r) + \frac{b}{\gamma_1}B^{(2)}(\gamma_1 r) =_D \sqrt{\frac{(a+c)^2 + b^2}{\gamma_1}}B^{(1)}(r)$  (or  $V_{\gamma_1, \gamma_2}(r; a, b, c) =_D \sqrt{\frac{a^2}{\gamma_1} + \frac{b^2}{\gamma_2}}B^{(1)}(r)$ ) and the limiting distribution is pivotal.

This example suggests that the limiting null distribution of the SN-based test statistic is not pivotal when sample sizes of the two sequences are not asymptotically equal unless the two sequences are independent. Thus the SN-based extension to the two sample case succeeds only when the two samples are independent or the sample sizes are asymptotically equal.

Since both covariance operator and its eigenfunctions are infinitely dimensional, to apply the SN method, we first project them onto the finite dimensional space spanned by certain basis functions. Typically, the basis functions are chosen to be the first few estimated eigenfunctions, but in the comparison of two eigenfunctions, the choice of basis functions should be made properly to ensure the asymptotic covariance matrix of the projection vector to be non-degenerate. See Remark 2.3 in Zhang and Shao (2014a).

**REMARK 4.1.** For both change point testing and two sample inference, the number of principal components (i.e.,  $K$ ) or the basis functions projected onto is assumed to be fixed as  $N \rightarrow \infty$  in our asymptotic framework. In practice, since  $K$  is usually chosen to make the first  $K$  principal components explain a certain percentage of variation, the magnitude of  $K$  critically depends on the prespecified threshold and the decay rate of the eigenvalues. It may be more meaningful to use the asymptotic results derived under the framework that  $K \rightarrow \infty$  but  $K/N \rightarrow 0$  as  $N \rightarrow \infty$ . This motivates us to ask whether the SN approach can be applied to the inference of a finite dimension parameter, where the dimension is not fixed but slowly growing with respect to sample size; see Fremdt, Horváth, Kokoszka and Steinebach (2014) for some related work. We shall leave this for future investigation.

## 5 Spatial data and spatio-temporal data

The extension of the SN approach to spatial setting was done in Zhang, Li and Shao (2014). Unlike a time series, spatial data does not have a natural one-directional ordering, which is essential in forming recursive estimates. For spatial lattice data with possibly irregular shaped sampling region, Zhang et al. (2014) defined a nested collection of subregions

which have the same shape as the whole region, applied recursive estimation and formed a self-normalizer. Under the increasing domain asymptotics, the limiting distribution of the self-normalized quantity was shown to depend on the irregular shape of the sampling region. A simulation based approach was adopted to obtain critical values. For nonlattice data, artificially ordering the spatial data into a time series in some way is shown to provide an effective means of generalizing the SN approach. However, the limiting distribution of the SN quantity is unknown, although a naive bootstrap method is seen to provide a decent approximation.

The SN approach has also been used to test for symmetry and separability of a space-time covariance function in Shao and Li (2009), as an alternative of the method developed in Li, Genton and Sherman (2007). In the latter paper, a test was developed based on the asymptotic joint normality of sample space-time covariances at a finite number of space-time lags. The asymptotic covariance matrix of sample space-time covariances was consistently estimated by using the subsampling method [Calstein (1986), Politis and Romano (1994)]. In Shao and Li (2009), recursive estimates of space-time covariances were used to form an inconsistent estimator of asymptotic covariance matrix and the resulting limiting null distribution of the self-normalized statistic is  $U_q$ , where  $q$  is the effective number of constraints on the space-time covariance function under the null hypothesis. Simulation results show that the SN-based test has superior size but loses some moderate power compared to the subsampling-based test in Li et al. (2007). It is worth noting that the setting in both Li et al. (2007) and Shao and Li (2009) assumes that the observations are taken from a fixed number of spatial locations at regularly spaced times  $\{1, \dots, n\}$ . Thus the space-time data can be viewed as a multivariate time series with the dimension of time series fixed and the length of time series going to infinity. This framework allows irregular nonlattice spatial configuration. For spatio-temporal data with regularly spaced lattice structure in space, the calculation of space-time covariances at certain space-time lags involves an average not only over time but also over space. For such data with a moderate amount of averaging in space, it may be beneficial to allow the number of points in both space and time going to infinity when deriving the asymptotic distribution of sample space-time covariances and self-normalized test statistic.

## 6 Related methods and extensions

As mentioned in Section 2, the existing inference approaches for time series data all involve a bandwidth parameter to accommodate the dependence when the unknown dependence is treated nonparametrically. The bandwidth is a smoothing parameter as it typically leads to a bias/variance tradeoff in variance estimation and a tradeoff between Type I error and Type II error in testing. In the early literature, the focus was on the consistency of these inference approaches (say, in terms of variance estimation) and a necessary condition to ensure the consistency is that the bandwidth goes to infinity but at a slower rate than sample size. In practice, different bandwidth parameters correspond to different finite sample distribution of test statistics (or studentized quantity), but the traditional first order asymptotic theory does not reflect different choices of bandwidth as the same reference distribution was used in the inference. To account for the influence of the bandwidth on the inference, Kiefer and Vogelsang (2005) advocated the fixed- $b$  asymptotics in the heteroscedasticity-autocorrelation robust testing context. The idea was to hold the ratio of bandwidth parameter  $l_n$  and sample size  $n$ , i.e.,  $b = l_n/n \in (0, 1]$ , as a fixed constant in the asymptotics. Under the fixed- $b$  asymptotics, the resulting limiting distribution of the studentized statistics (or quantity) is nonstandard, but is pivotal for a given  $b$ .

The fixed- $b$  idea has been generalized to subsampling and block bootstrap based inference by Shao and Politis (2013). The idea was to derive the limiting null distribution of the p-value under the fixed- $b$  asymptotics, which is no longer  $U(0, 1)$  (i.e., uniform distribution on  $[0, 1]$ ), as is the case under the small- $b$  asymptotics. It depends upon  $b$  and is pivotal in the scalar parameter case, but is non-pivotal in the case of a vector parameter or infinitely dimensional parameter. A calibration of the traditional subsampling or block-bootstrap based confidence set is then performed by either using simulated critical values corresponding to the  $b$ -dependent pivotal distribution when the parameter is a scalar or estimating the sampling distribution of the p-value using a second-level subsampling when the dimension of the parameter is more than 1. Simulation results indicate that the fixed- $b$  approach is more accurate than the traditional small- $b$  approach in terms of coverage accuracy, and the calibrated confidence sets tend to have smaller coverage errors than the uncalibrated counterparts.

In Zhang and Shao (2014b), the fixed- $b$  idea was extended to blockwise empirical likelihood (BEL, hereafter), which was first proposed by Kitamura (1997) as an extension



of Owen's (1988, 1990) empirical likelihood method to time series data. An important property of the EL is that minus twice of the log-EL ratio statistic evaluated at the true parameter converges to  $\chi^2$  limiting distribution (i.e., Wilks's theorem holds), which is the basis of EL ratio confidence region. Kitamura (1997) showed that Wilks's theorem still holds provided that the block size  $l_n$  goes to infinity but at a slower rate of sample size  $n$ . In the literature, there seems little guidance provided on the choice of the block size and the performance of BEL can be sensitive to the choice of block size. To capture the impact of block size on the finite sample distribution of log-EL ratio statistic, Zhang and Shao (2014b) adopted the moment condition model [Qin and Lawless (1994), Smith (2011)] and showed that the limiting distribution of BEL ratio evaluated at the true parameter is non-standard but pivotal under the fixed- $b$  asymptotics and the critical values can be obtained via simulations. In the simulation studies, it was shown that the confidence region based on the fixed- $b$  approach has more accurate coverage than the traditional  $\chi^2$ -based counterpart. The same phenomenon was observed when generalizing the fixed- $b$  approach to generalized blockwise empirical likelihood [Newey and Smith (2004)], which include exponential tilting and continuously updating generalized method of moments as special cases.

Nordman, Bunzel and Lahiri (2013) proposed a non-standard bandwidth-free empirical likelihood for the inference of time series, called expansive blockwise EL (EBEL). The EBEL uses recursive blocks, as the SN method does, so it may be viewed as an interesting marriage of self-normalization and empirical likelihood, but it was developed only for the smooth functional model, which is narrower than the framework adopted in Shao (2010a). The large sample distribution of log-EBEL ratio statistic is non-standard but pivotal. Through extensive simulations, Nordman et al. (2013) showed that EBEL generally exhibits comparable (or in some cases even better) coverage accuracy than the BEL with  $\chi^2$  approximation and suitable block size.

In addition to the extensions mentioned above, Kim and Zhao (2013) have developed SN-based unified inference for both sparse and dense longitudinal data. While their first method is based on a self-normalized central limit theorem which can adapt to both sparse and dense case, their second method uses recursive estimates, which can be regarded as an extension of Shao (2010a) to nonparametric context. Recently, the SN approach has been extended by Huang, Volgushev and Shao (2015) to inference for censored time series, where the limiting distribution of the SN quantity is obtained by using the new theory developed in Volgushev and Shao (2014), and a comparison with the blockwise empirical

likelihood approach of El Ghouch, Van Keilegom, and McKeague (2011) again confirms the coverage accuracy of the SN approach.

## 7 Conclusions

In this paper, we review some recent developments for the inference of time series data using the self-normalized approach. To summarize, we highlight a number of appealing features of self-normalization and mention several topics for future research. Among its desirable properties, we mention

(i) Implementational convenience. A key feature of SN in the time series context is to use recursive estimates, which can be readily computed without the need to design new algorithms. This is similar to blockwise jackknife or subsampling, where the first level estimate based on a subset of observations is computed, and no theoretical formula for asymptotic variance needs to be derived for practical implementation. If the limiting distribution of the SN quantity is pivotal, then we can either use the tabulated critical value or simulate the critical values. Otherwise, we may have to use resampling/subsampling methods or simulation-based approach to approximate the critical values.

(ii) Broad applicability. As demonstrated in this paper, SN is widely applicable to a number of important inference problems in time series analysis, ranging from univariate to functional time series, from stationary to locally stationary framework, and from short memory to long memory series. It can be viewed as a novel way of studentization or prepivoting [Beran (1987)], and it may be combined with other existing approaches (e.g., block bootstrap) to deliver more accurate inference [see Shao (2010a)].

Furthermore, the SN approach can facilitate the inference in quantile regression, in which the asymptotic variance for the quantile regression estimator often involves a nuisance parameter called the sparsity function [Koenker 2005, page 77], namely, the reciprocal of a (conditional) density evaluated at the true (conditional) quantile. Its consistent estimation typically involves smoothing and bandwidth selection. The SN approach offers a viable alternative [Zhou and Shao (2013)] and is worth further investigation in quantile-based inference. Consistent estimation of long run variance also comes up in other time series inference problems, such as unit root testing (e.g., Phillips (1987), Koenker and Xiao (2004)), and it would be interesting to explore the possibility for the SN-based alternatives.

(iii) Good theoretical properties. First order limiting distributions have been derived

for the SN quantities in most problems we mentioned except for the case of irregularly spaced spatial data [Zhang et al. (2014)]. In the derivation, the weak convergence of the sequential empirical process associated with recursive estimates play a crucial role; see Volgushev and Shao (2014). As we see from Section 3.2-Section 3.5, the limiting distribution of the SN quantity may not be pivotal, and thus require another level of approximation using resampling/subsampling or simulation-based approaches. The high order distribution theory has been derived in Jansson (2004), Sun et al. (2008) and Zhang and Shao (2013) for the self-normalized mean in the framework of Gaussian location model. The high order expansion of the finite sample distribution of studentized mean in Zhang and Shao (2013) provides theoretical insights on the magnitude and sign of size distortion, and the discussion and theory in Sun et al. (2008) greatly help to clarify the distinction between the fixed- $b$  and small- $b$  approximations. Of course, it would be interesting to develop higher order theory for the SN-based inference in non-Gaussian setting and for other problems. In the mean inference context, the testing-optimal bandwidth developed in Sun et al. (2008) tried to optimize a weighted average of type I and type II errors, and it would be interesting but may be very challenging to extend this idea and derive and implement the testing-optimal bandwidth for other problems.

(iv) Encouraging finite sample performance. In our finite sample simulations, we have demonstrated the size/coverage accuracy of the SN approach in comparison with the approaches that involve consistent variance estimation. The moderate power loss or wider interval associated with the SN approach has also been recognized in finite samples for most problems, which is consistent with the existing theory. As we commented in Remark 2.2, the SN approach may be viewed as a robust alternative to the existing ones with some mild sacrifice of efficiency.

That being said, we also want to point out some limitations of the SN approach and some areas for which SN has not been successfully extended to and more work is needed. In particular, we mention

(i) SN is not applicable to the inference of infinitely dimensional parameter, such as the marginal distribution function of a stationary time series  $X_t$ , but subsampling and block-based bootstrap do. Furthermore, when the dimension of a finite-dimensional parameter grows slowly with respect to sample size, there is no theory that supports or justifies the use of the SN approach. Therefore, it is better to view SN as a complement but not a replacement of the existing methods, since its applicability is different from the existing

ones.

(ii) The unidirectional ordering of a time series is implicitly used in forming recursive estimates, which is a key ingredient of the SN approach. It would be interesting to consider extensions to dependent data of other types, where no natural ordering exists; see Zhang et al. (2014) for an attempt to irregularly spaced spatial data. In Ibragimov and Müller (2010), a t-statistic based approach was proposed to deal with dependent data, which does not require a natural ordering. In addition, when the dependence in the time series is too strong (say, near-integrated time series), inference of certain parameter becomes difficult because the information aggregated over time does not accumulate quickly due to strong dependence. It would be desirable to extend the SN method to this setting in a principled way similar to Müller (2014).

(iii) When sample size  $n$  is large, the calculation of recursive estimates gets expensive. It would be interesting to develop a computationally scalable version of the SN method in the big data context. See Chen and Qu (2015) for a variant of SN that is less expensive to compute. Ideally, we hope this new version maintains the size/coverage accuracy of the SN approach while being computationally feasible for big data.

(iv) In some extensions, we have to introduce a trimming parameter  $\epsilon \in (0, 1)$  mostly for technical reasons. In other words, we only use recursive estimates starting from  $\hat{\theta}_{1, \lfloor \epsilon n \rfloor}$  in our self-normalizer. The choice of trimming parameter certainly affects the finite sample results, but its impact on the inference is captured by the first order limiting distribution and its approximation. This is in the same spirit of the fixed- $b$  approach, where the impact of  $b$  is reflected in the first order limiting distribution. Setting  $\epsilon = 0.1$  has been seen to perform quite well for various data generating processes and for the problems studied in Zhou and Shao (2013), Kim et al. (2015), and Huang et al. (2015). It would be interesting to (1) find a generally applicable rule to select  $\epsilon$  in the SN approach or  $b$  in the fixed- $b$  approach; (2) compare the trimmed SN approach with fixed- $b$  approach in the same general setting and find out the difference between the two and whether one dominates the other.

(v) Except for the extension to functional time series, the dimension of the time series is assumed to be finite and fixed in most applications of SN techniques. It is natural to ask whether SN can be extended to inference for high dimensional time series, which has attracted a lot of attention lately [see e.g., Fan, Lv and Qi (2011), Lam and Yao (2012)]. This would be an interesting avenue for future research.

## REFERENCES

- Abadir, K. M., Distaso, W. and Giraitis, L. (2009) Two estimators of the long-run variance: Beyond short memory. *Journal of Econometrics*, 120, 56-70.
- Abadir, K. M. and Paruolo, P. (1997) Two mixed normal densities from cointegration analysis. *Econometrica*, 65, 671-680.
- Andrews, D. W. K. (1991) Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59, 817-858.
- Aue, A. and Horváth, L. (2013) Structural breaks in time series. *Journal of Time Series Analysis*, 34, 1-16.
- Benko, M., Härdle, W. and Kneip, A. (2009) Common functional principal components. *Annals of Statistics*, 37, 1-34.
- Beran, R. (1987) Previoting to reduce level error of confidence sets. *Biometrika*, 74, 457-468.
- Berk, K. N. (1974) Consistent autoregressive spectral estimates. *Annals of Statistics*, 2, 489-502.
- Berkes, I., Gabrys, R., Horváth, L. and Kokoszka, P. (2009) Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society, Series B*, 71, 927-946.
- Billingsley, P. (1968) *Convergence of Probability Measures*, Wiley.
- Bosq, D. (2000). *Linear Process in Function Spaces: Theory and Applications*. New York: Springer.
- Bühlmann, P. (2002) Bootstraps for time series. *Statistical Science*, 17, 52-72.
- Calstein, E. (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, 14, 1171-1179.
- Cavaliere, G. (2005) Unit root tests under time-varying variances. *Econometric Reviews*, 23, 259-292.
- Cavaliere, G. and Taylor, A. M. R. (2007) Testing for unit roots in time series models with non-stationary volatility. *Journal of Econometrics*, 140, 919-947.
- Chan, N. H. and Wei, C. Z. (1987) Asymptotic inference for nearly nonstationary AR(1) processes. *The Annals of Statistics*, 15, 1050-1063.
- Chen, Y.-T. and Qu, Z. (2015) M tests with a new normalization matrix. *Econometric Review*, 34, 617-652.
- Crainiceanu, C. M. and Vogelsang, T. J. (2007) Spectral density bandwidth choice: source

- of nonmonotonic power for tests of a mean shift in a time series. *Journal of Statistical Computation and Simulation*, 77, 457-476.
- Davidson, J. and de Jong, R. M. (2000) The functional central limit theorem and weak convergence to stochastic Integrals II: fractionally integrated processes. *Econometric Theory*, 16, 643-666.
- de la Pena, V. H., Lai, T. L., and Shao. Q. M. (2009) *Self-Normalized Processes: Limit Theory and Statistical Applications*, Springer.
- Deng, A., and Perron, P. (2008) A non-local perspective on the power properties of the CUSUM and CUSUM of squares tests for structural change. *Journal of Econometrics*, 142, 212-240.
- Dobrushin, R.L., and Major, P. (1979) Non-central limit theorems for non-linear functionals of Gaussian fields, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 50, 827-842.
- El Ghouch, A., Van Keilegom, I., and McKeague, I. (2011) Empirical likelihood confidence intervals for dependent duration data. *Econometric Theory*, 27, 178-198.
- Fan, J., Lv, J. and Qi, L. (2011) Sparse high-dimensional models in economics. *Annual Review of Economics*, 3, 291-317.
- Fan, Z. (2010) *Statistical Issues and Developments in Time Series Analysis and Educational Measurement*. Ph.D. thesis, University of Illinois at Urbana-Champaign.
- Fremdt, S., Horváth, L., Kokoszka, P., and Steinebach, J. G. (2014) Functional data analysis with increasing number of projections. *Journal of Multivariate Analysis*, 124, 313-332.
- Geweke, J. and Porter-Hudak, S. (1983) The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4, 221-238.
- Granger, C. W. J. (1966) The typical spectral shape of an economic variable. *Econometrica*, 34, 150-161.
- Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995) On blocking rules for the bootstrap with dependent data. *Biometrika*, 82, 561-574.
- Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.
- Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38, 1845-1884.
- Huang, Y., Volgushev, S. and Shao, X. (2015) On self-normalization for censored dependent

- data. *Journal of Time Series Analysis*, 36, 109-124.
- Ibragimov, R. and Müller (2010)  $t$ -statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics*, 28, 453-468.
- Jach, A., McElroy, T. and Politis, D. N. (2012) Subsampling inference for the mean of heavy-tailed long memory time series. *Journal of Time Series Analysis*, 33, 96-111.
- Jansson, M. (2004) The error rejection probability of simple autocorrelation robust tests. *Econometrica*, 72, 937-946.
- Juhl, T. and Xiao, Z. (2009) Testing for changing mean monotonic power. *Journal of Econometrics*, 148, 14-24.
- Kiefer, N. M., and Vogelsang, T. J. (2002) Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica*, 70, 2093-2095.
- Kiefer, N. M., and Vogelsang, T. J. (2005) A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21, 1130-1164.
- Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000) Simple robust testing of regression hypotheses. *Econometrica*, 68, 695-714.
- Kim, S. and Zhao, Z. (2013) Unified inference for sparse and dense longitudinal models. *Biometrika*, 100, 203-212.
- Kim, S., Zhao, Z., and Shao, X. (2015) Nonparametric functional central limit theorem for time series regression with application to self-normalized confidence interval. *Journal of Multivariate Analysis*, 133, 277-290.
- Kitamura, Y. (1997) Empirical likelihood methods with weakly dependent processes. *Ann. Statist.*, 25, 2084-2102.
- Koenker, R. (2005) *Quantile Regression*. Cambridge: Cambridge University Press.
- Koenker, R. and Xiao, Z. (2004) Unit root quantile regression inference. *Journal of the American Statistical Association*, 99, 775-787.
- Künsch, H. (1987) Statistical aspects of self-similar processes. In Yu.A. Prohorov & V.V. Sazonov (eds.), *Proceedings of the 1st World Congress of the Bernoulli Society*, vol. 1, pp. 6774. Science Press.
- Künsch, H. (1989) The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17, 1217-1241.
- Lahiri, S. N. (2003) *Resampling Methods for Dependent Data*, Springer.
- Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: inference

- for the number of factors. *Annals of Statistics*, 40, 694-726.
- Lehmann, E. L. (1999) “Student” and small-sample theory. *Statistical Science*, 14, 418-426.
- Li, B., Genton, G. and Sherman, M. (2007) A nonparametric assessment of properties of space-time covariance functions. *J. Am. Statist. Assoc.*, 102, 736-744.
- Liu, R. Y. (1988) Bootstrap procedures under some non-i.i.d. models. *Annals of Statistics* 16, 1696-1708.
- Liu, R. Y., and Singh, K. (1992) Moving block jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, Ed. R. LePage and L. Billard, pp. 225-248. New York: John Wiley.
- Lobato, I. N. (2001) Testing that a dependent process is uncorrelated. *J. Am. Statist. Assoc.*, 96, 1066-1076.
- McElroy, T. and Politis, D. N. (2007) Computer-intensive rate estimation, diverging statistics, and scanning. *Ann. Statist.*, 35, 1827-1848.
- McElroy, T. and Politis, D. N. (2012) Fixed- $b$  asymptotics for the studentized mean from time series with short, long or negative memory. *Econometric Theory*, 28, 471-481.
- McElroy, T. and Politis, D. N. (2014) Spectral density and spectral distribution inference for long memory time series via fixed- $b$  asymptotics. *Journal of Econometrics*, 182, 211-225.
- Müller, U. K. (2007) A theory of robust long-run variance estimation. *Journal of Econometrics*, 141, 1331-1352.
- Müller, U. K. (2014) HAC corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics*, 32, 311-322. (with discussions and rejoinder)
- Newey, W. K. and Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219-255.
- Newey, W. K. and West, K. D. (1987) A simple, positive, semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 703-708.
- Newey, W. K., and West, K. D. (1994) Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61, 631-653.
- Nordman, D. J., Bunzel, H. and Lahiri, S. N. (2013) A non-standard empirical likelihood for time series. *Annals of Statistics*, 41, 3050-3073.
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.
- Owen, A. (1990). Empirical likelihood confidence regions. *Ann. Statist.*, 18, 90-120.



- Perron, P. (2006) Dealing with structural breaks. *Palgrave Handbook of Econometrics*, Vol. 1: Econometric Theory, K. Patterson and T.C. Mills (eds.), Palgrave Macmillan, 278-352.
- Phillips, P. C. B. (1987) Time series regression with a unit root. *Econometrica*, 55, 277-301.
- Politis, D. N. and Romano, J. P. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22, 2031-2050.
- Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer-Verlag, New York.
- Priestley, M. B. (1981) *Spectral Analysis and Time Series I, Univariate Series*. Academic Press, London.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, 22, 300-325.
- Quenouille (1949), Approximate tests of correlation in time series. *Journal of Royal Statistical Society, Series, B*, 11, 68-84.
- Rho, Y. and Shao, X. (2013a), Improving the bandwidth-free inference methods by prewhitening. *Journal of Statistical Planning and Inference*, 143, 1912-1922.
- Rho, Y. and Shao, X. (2013b) Inference for locally stationary time series regression models. Technical Report, University of Illinois at Urbana-Champaign.
- Rho, Y. and Shao, X. (2015) Inference for time series regression models with weakly dependent and heteroscedastic errors. *Journal of Business and Economic Statistics*, to appear.
- Robinson, P. M. (2005) Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction. *Econometric Theory*, 21, 171-180.
- Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, New York.
- Shao, X. (2010a) A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society, Series B* 72(3), 343-366. Corrigendum: 2010, 72(5), 695-696.
- Shao, X. (2010b) The dependent wild bootstrap. *Journal of the American Statistical Association*, 105, 218-235.
- Shao, X. (2012) Parametric inference for stationary time series models with dependent errors. *Scandinavian Journal of Statistics*, 39, 772-783.
- Shao, X. and Li, B. (2009) A tuning parameter free test for properties of space-time

- covariance functions. *Journal of Statistical Planning and Inference*, 139, 4031-4038.
- Shao, X. and Politis, D. N. (2013) Fixed- $b$  subsampling and block bootstrap: improved confidence sets based on p-value calibration. *Journal of the Royal Statistical Society, Series, B.*, 75, 161-184.
- Shao, X. and Zhang, X. (2010) Testing for change points in time series. *Journal of the American Statistical Association*, 105, 1228-1240.
- Simonoff, J. S. (1993) The relative importance of bias and variability in the estimation of the variance of a statistic. *The Statistician*, 42, 3-7.
- Smith, R. (2011) GEL criteria for moment condition models. *Econometric Theory*, 27, 1192-1235.
- Student (1908) The probable error of a mean. *Biometrika*, 6, 1-25.
- Sun, Y. (2014) Comment on “HAC corrections for strongly autocorrelated time series”. *Journal of Business and Economic Statistics*, 32, 330-334.
- Sun, Y. (2015) Fixed-smoothing asymptotics and asymptotic  $F$  and  $t$  tests in the presence of strong autocorrelation. To appear in *Advances in Econometrics: Essays in Honor of Peter C.B. Phillips*.
- Sun, Y., Phillips, P. C. B. and Jin, S. (2008) Optimal bandwidth selection in heteroscedasticity-autocorrelation robust testing. *Econometrica*, 76, 175-194.
- Taqqu, M.S. (1975) Weak convergence to fractional Brownian motion and to the Rosenblatt process, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31, 287-302.
- Taqqu, M.S. (1979) Convergence of integrated processes of arbitrary hermite rank, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 50, 53-83.
- Tukey (1958) Bias and confidence in not-quite large samples (Abstract). *Annals of Mathematical Statistics*, 29, 614.
- Velasco, C., and Robinson, P. M. (2001) Edgeworth expansions for spectral density estimates and studentized sample mean. *Econometric Theory*, 17, 497-539.
- Vogelsang, T. (1999) Sources of non-monotonic power when testing for a shift in mean of a dynamic time series. *Journal of Econometrics*, 88, 283-299.
- Vogelsang, T. (2014) Comment on “HAC corrections for strongly autocorrelated time series”. *Journal of Business and Economic Statistics*, 32, 334-338.
- Volgushev, S. and Shao, X. (2014) A general approach to the joint asymptotic analysis of statistics from sub-samples. *Electronic Journal of Statistics*, 8, 390-431.

- Wu, C. F. J. (1986) Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics* 14, 1261-1350.
- Wu, W. B. (2005) On the Bahadur representation of sample quantiles for dependent sequences. *Annals of Statistics*, 33, 1934-1963.
- Wu, W. B. and Shao, X. (2006) Invariance principles for fractionally integrated nonlinear processes, *IMS Lecture notes-Monographs series, Festschrift for Michael Woodroffe*, 50, 20-30.
- Zabell, S. L. (2008) On Student's 1908 article "The Probable Error of a Mean". *Journal of the American Statistical Association*, 103, 1-7.
- Zhang, X., Li, B. and Shao, X. (2014) Self-normalization for spatial data. *Scandinavian Journal of Statistics*, 41, 311-324.
- Zhang, X. and Shao, X. (2013) Fixed-smoothing asymptotics for time series. *Annals of Statistics*, 41, 1329-1349.
- Zhang, X. and Shao, X. (2014a) Two sample inference for temporally dependent functional data. *Bernoulli*, to appear.
- Zhang, X. and Shao, X. (2014b) Fixed- $b$  asymptotics for blockwise empirical likelihood. *Statistica Sinica*, 24, 1179-1194.
- Zhang, X., Shao, X., Hayhoe, K. and Wuebbles, D. (2011) Testing the structural stability of temporally dependent functional observations and application to climate projections. *Electronic Journal of Statistics*, 5, 1765-1796.
- Zhao, Z. (2011) A self-normalized confidence interval for the mean of a class of nonstationary processes. *Biometrika*, 98, 81-90.
- Zhao, Z. and Li, X. (2013) Inference for modulated stationary processes. *Bernoulli*, 19, 205-227.
- Zhou, Z. and Shao, X. (2013) Inference for linear models with dependent errors. *Journal of the Royal Statistical Society, Series B*, 75, 323-343.
- Zhou, Z. and Wu, W. B. (2009) Local linear quantile estimation for nonstationary time series. *Ann. Statist.*, 37, 2696-2729.

Table 1: The limits of the probability and ratio associated with interval lengths

$\alpha$	0.5	0.75	0.9	0.95	0.99
$P(\alpha)$	0.95	0.514	0.386	0.337	0.223
$R(\alpha)$	0.963	1.1	1.235	1.297	1.445

Table 2: Upper quantiles of the distribution of  $J_1$  simulated on the basis of 200000 monte carlo replications.

$\alpha$	0.8	0.9	0.95	0.975	0.99	0.995
$J_{1,\alpha}$	24.88	44.46	68.41	96.19	139.73	176.87

Figure 1: Approximate local powers for SN-based and traditional tests as a function of  $\delta$  under the local alternative  $\mu_n = \mu_0 + \delta s/\sqrt{n}$ . Equivalently,  $P_{SN}(\gamma, \delta)$  (solid line) and  $P_{TR}(\gamma, \delta)$  (dashed line) as a function of  $\delta$  at  $\gamma = 0.05$  and  $0.1$ .

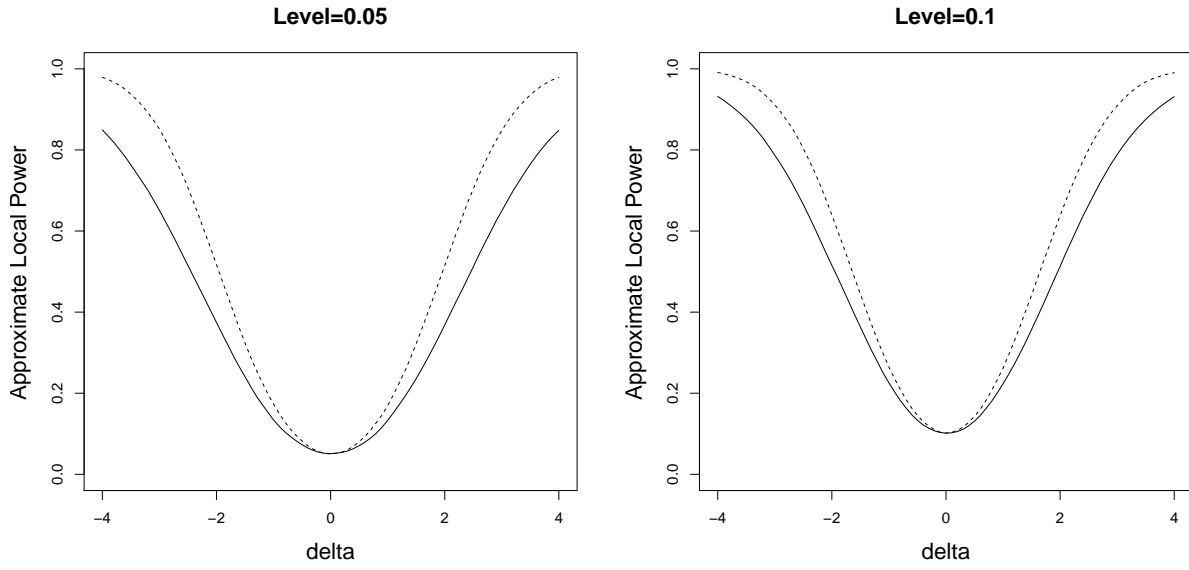


Table 3: Empirical coverage rates (in percentage) and ratios of average interval lengths for variants of SN method. The intervals are for marginal median of time series and the number of replication is 50000. Time series of length  $n$  is generated from Gaussian AR(1) model with AR(1) coefficient  $\rho$ . The variations  $SN_j$ ,  $j = 1, 2, 3$  of the self-normalized approach (denoted by SN) are introduced in Section 2.3.

$n$	$1-\alpha$	$\rho$	SN	$SN_1$	$SN_2$	$SN_3$	$\frac{len(SN)}{len(SN_3)}$	$\frac{len(SN_1)}{len(SN_3)}$	$\frac{len(SN_2)}{len(SN_3)}$
50	90%	0.2	86.7	86.9	89.9	90.4	.963	.967	.905
		0.5	85.7	85.5	87.6	88.8	.968	.971	.895
		0.8	80.1	80.1	79.4	83.1	.969	.971	.850
		-0.4	87.3	87.4	91.7	91.8	.955	.957	.914
	95%	0.2	91.9	92.0	94.4	94.9	.963	.967	.880
		0.5	91.2	91.0	92.9	93.8	.968	.971	.870
		0.8	86.8	86.7	86.2	89.4	.969	.971	.827
		-0.4	92.2	92.2	96.0	95.9	.955	.957	.890
200	90%	0.2	88.2	88.5	90.1	90.5	.978	.980	.909
		0.5	88.2	88.4	89.5	90.0	.984	.982	.908
		0.8	86.9	86.8	87.3	88.3	.986	.985	.897
		-0.4	88.4	88.6	90.9	91.5	.970	.972	.904
	95%	0.2	93.5	93.6	94.8	95.2	.978	.980	.884
		0.5	93.5	93.6	94.8	95.2	.978	.980	.884
		0.8	92.3	92.4	92.7	93.7	.986	.985	.873
		-0.4	93.2	93.5	95.3	95.6	.970	.972	.879

Table 4: Empirical coverage rates (in percentage) and ratio of average interval lengths for several inference methods. The intervals are for marginal mean of time series and the number of replications is 50000. Time series of length  $n$  is generated from Gaussian AR(1) model with AR(1) coefficient  $\rho$ . The symbols SN, TR, INF and SN<sub>2</sub> stand for the self-normalized approach, the traditional approach where long run variance is consistently estimated, the infeasible approach where true long run variance is used in normal approximation, and a variation of the self-normalized approach where all recursive subsample estimates are used in the self-normalizer, respectively.

$n$	$1 - \alpha$	$\rho$	SN	TR	INF	SN <sub>2</sub>	$\frac{\text{len}(\text{SN})}{\text{len}(\text{TR})}$	$\frac{\text{len}(\text{SN})}{\text{len}(\text{INF})}$	$\frac{\text{len}(\text{SN})}{\text{len}(\text{SN}_2)}$
50	90%	0.2	89.2	84.9	88.6	89.3	1.35	1.21	1.08
		0.5	87.1	79.0	86.7	86.8	1.44	1.18	1.10
		0.8	81.1	66.9	80.2	78.5	1.56	1.10	1.16
		-0.4	91.9	90.1	89.7	92.8	1.23	1.28	1.06
	95%	0.2	94.4	90.9	93.8	94.5	1.43	1.29	1.11
		0.5	92.9	85.9	92.1	92.6	1.53	1.26	1.12
		0.8	88.0	74.5	86.1	85.8	1.66	1.17	1.18
		-0.4	96.3	94.8	94.8	96.9	1.31	1.36	1.08
200	90%	0.2	89.8	87.2	89.7	89.7	1.30	1.22	1.09
		0.5	89.3	84.7	89.1	89.1	1.35	1.21	1.09
		0.8	87.5	79.6	87.0	86.6	1.43	1.19	1.11
		-0.4	90.5	91.0	90.1	90.7	1.18	1.24	1.08
	95%	0.2	94.8	92.9	94.6	94.8	1.38	1.30	1.11
		0.5	94.5	91.0	94.1	94.3	1.44	1.29	1.12
		0.8	93.1	86.5	92.4	92.5	1.52	1.26	1.13
		-0.4	95.3	95.4	94.9	95.5	1.26	1.31	1.10

Figure 2: Empirical rejection rates of the  $KS_n$  statistic as a function of the magnitude of change  $\eta$  under the model (19) when the bandwidth  $l_n$  is chosen using Andrews' plug in rule (2).

