Scientific Research Publishing

Applied Mathematics, 2021, 12, 1031-1037
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Scanning for Clusters of Large Values in Time Series: Application of the Stein-Chen Method

Tom Burr, Brad Henderson

Space Sciences, Los Alamos National Laboratory, New Mexico, USA
Email: tburr@lanl.gov

How to cite this paper: Burr, T. and Henderson, B. (2021) Scanning for Clusters of Large Values in Time Series: Application of the Stein-Chen Method. *Applied Mathematics*, **12**, 1031-1037.
https://doi.org/10.4236/am.2021.1211067

Received: October 24, 2021
Accepted: November 27, 2021
Published: November 30, 2021

Abstract

The purpose of this application paper is to apply the Stein-Chen (SC) method to provide a Poisson-based approximation and corresponding total variation distance bounds in a time series context. The SC method that is used approximates the probability density function (PDF) defined on how many times a pattern such as $I_t, I_{t+1}, I_{t+2} = \{1\,0\,1\}$ occurs starting at position t in a time series of length N that has been converted to binary values using a threshold. The original time series that is converted to binary is assumed to consist of a sequence of independent random variables, and could, for example, be a series of residuals that result from fitting any type of time series model. Note that if {1 0 1} is known to not occur, for example, starting at position $t = 1$, then this information impacts the probability that {1 0 1} occurs starting at position $t = 2$ or $t = 3$, because the trials to obtain {1 0 1} are overlapping and thus not independent, so the Poisson distribution assumptions are not met. Nevertheless, the results shown in four examples demonstrate that Poisson-based approximation (that is strictly correct only for independent trials) can be remarkably accurate, and the SC method provides a bound on the total variation distance between the true and approximate PDF.

Keywords

Clusters of Large Values, Stein-Chen Method

1. Introduction and Background

Suppose there is interest in the probability that a pattern such as {1 0 1} or {1 1 1} occurs in a sequence of $N = 10$ independent Bernoulli trails. The main interest in this paper is the case with a small Bernoulli success probability, $p = P(I_i = 1)$, consisting, for example, of whether a residual from a fitted time series model exceeds a threshold. A pattern such as {1 0 1} or {1 1 1} could indicate a depar-

ture from the fitted model, perhaps indicating that a signal of interest is present.

This paper considers scanning for {1 x 1} with $x = 0$ or 1, with $p = P(I_i = 1)$ being quite small, such as 0.10 or less. Then the probability of the pattern {1 x 1} is p^2, and there are $N - 2 = 8$ possible starting locations for the pattern in $N = 10$ trials. Because there are only $2^{10} = 1024$ possible patterns of 0's and 1's, all 1024 patterns could be listed, and the probabilities assigned to each set of 10 binary values that include {1 x 1} at least once could be summed to provide an exact calculation (Example 2 in Section 4). For larger values of N, this exact calculation is unwieldy, so an approximate method is desired, provided the approximation is highly accurate with provable error bounds.

This paper uses the SC approximation method (Section 3) to greatly simplify calculating the probability that a specified pattern occurs in a sequence of independent residuals in a time series context. The SC method approximates the probability density function (PDF) defined on how many times a pattern such as $I_t, I_{t+1}, I_{t+2} = \{1\,0\,1\}$ occurs starting at position t in a time series of length N that has been converted to binary values using a threshold. The original time series that is converted to binary is assumed to consist of a sequence of independent random variables, and could, for example, be a series of residuals that result from fitting any type of time series model. Note that if {1 0 1} is known to not occur, for example, starting at position $t = 1$, then this information impacts the probability that {1 0 1} occurs starting at position $t = 2$ or $t = 3$, because the trials to obtain {1 0 1} are overlapping and thus not independent, so the Poisson distribution assumptions are not met.

Figure 1 is an example of time series data, consisting of electric consumption recorded every hour for 14 days for a total of 336 measurements (the data named elec_load aggregated from 30 minute to 60 minute time steps from the TSrepr package in [1]). This electric consumption data is used here simply as an example of the type of time series that this paper considers. Figure 2 is a binary version of the series in Figure 1, with values 3.5 or larger set to 1 and values less than 3.5 set to 0. Approximately 2% (6 of 336) of the 336 values exceed 3.5.

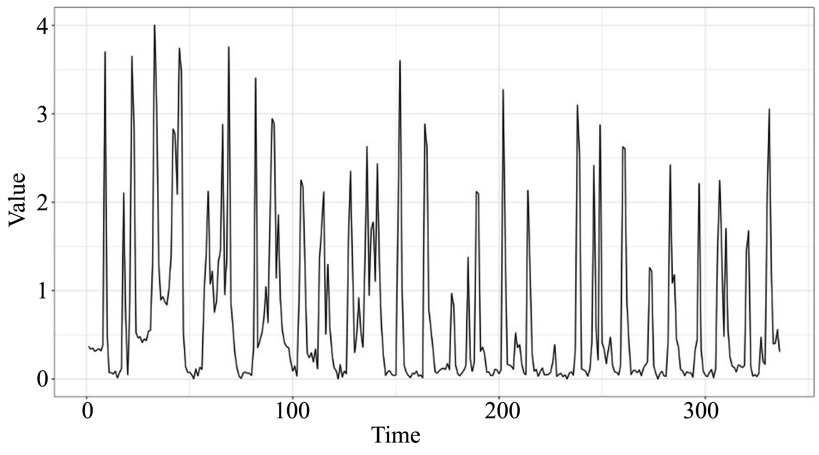

Figure 1. Time series data: 336 observations of electric consumption values in arbitrary units.

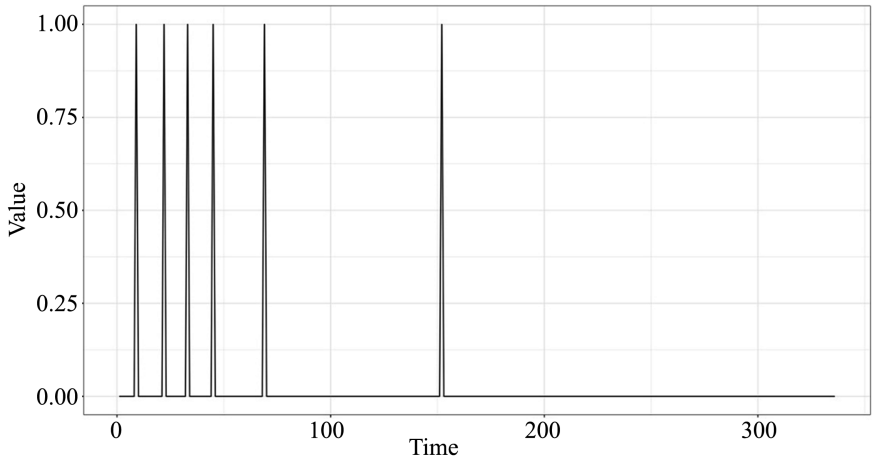

Figure 2. A binary version of the same data in **Figure 1**; 6 of 336 values exceed the 3.5 threshold.

Figure 2 is the type of data that motivated this case-study application of the SC method. In the applications of interest, the combinatorial counting can only be done for very short time series (and so it is done only in Example 2 below with a length 5 time series). Therefore, the application led the authors to apply and assess the SC bound for a corresponding simple Poisson approximation. The SC bound does not seem to be well known among practitioners; however, as this paper shows, the SC bound can defend the use of the simple Poisson approximation in some real applications, and can provide a very small bound on the approximation error.

The advantage of the SC method in this context is simplicity and tractability (as shown in Examples 1 to 4 below). The disadvantage is that the SC method is an approximate method for which the total variation distance bound must be calculated in order to assess the quality of the approximation under various conditions (as shown in Examples 1 to 4 below). Fortunately, the SC approximation quality is excellent in the applications of interest.

2. Methodology: Scanning for Specified Patterns

The $N = 336$ binary values in **Figure 2** are an example of the type of binary time series considered here. The binary values $I_1, I_2, I_3, \cdots, I_N$ are assumed to be independent and identically distributed with constant probability $p = P(I_i = 1)$. The probability p is the probability that the original time series X exceeds a threshold, and the I notation denotes an indicator or binary variable. As an aside, the *SC* method can also be applied if p is not constant over time, but the independence assumption is difficult to avoid [2]. Any type of time series model [3] can be fit to the series of interest, and then the resulting residuals become the original series that is thresholded to convert to binary; therefore, the application is quite general.

Suppose that large values of the original series are thought to rarely cluster, so, for example, a pattern such as {1 0 1} or {1 1 1} could indicate a departure from

the assumed time series model, perhaps indicating that a signal of interest is present. This paper will consider scanning for {1 x 1} with $x = 0$ or 1, with $p = P$ ($I_i = 1$) being quite small, such as 0.10 or less. Then the probability of the pattern {1 x 1} is $p_p = p^2$.

Start at index $i = 1$ and check whether {1 x 1} occurs in positions {1 2 3}, then start at index $i = 2$ and check whether {1 x 1} occurs starting at index 2 in positions {2 3 4}, then start at index 3, etc. Note, for example, that if {1 x 1} occurs starting at position $i = 1$, then the probability that {1 x 1} also occurs starting at index 3 is p. Clearly, there is a small neighborhood of dependence around each starting index, as just illustrated. This neighborhood of dependence violates the assumptions for a Poisson distribution (as a limit distribution for a sequence of N Bernoulli trials, each with small probability of success), but [2] shows that provided the dependence neighborhood is modest, the Poisson distribution can still provide an excellent approximation to the PDF defined on the number of times {1 x 1} occurs in a series of length N.

3. Stein-Chen Method

According to Theorem 2 in [2], the Poisson PDF with mean parameter $\lambda = (N-2)p_p$ provides an approximation Y to the true PDF W for the number of times {1 x 1} occurs in a series of length N. The value $(N-2)$ is used instead of N because the length-3 pattern could only be found starting at index $1, 2, \cdots, N-2$.

The quality of the Poisson (λ) approximation can be measured by computing the terms b_1 and b_2, where $b_1 = \sum_{i=1}^{N} \sum_{j \in N_i} p_i p_j$, with $N_i = \{i-2, i-1, i, i+1, i+2\}$ being the dependence neighborhood of index i, and $b_2 = \sum_{i=1}^{N} \sum_{j, j' \in N_i} E\{I_j I_{j'}\}$, where $j \neq j'$. The term b_3 is equal to 0 in Theorem 2 of [2] by construction of N_i in this example. Then, it is easily shown that the total variation distance (TVD) satisfies:

$$d_{TVD}(Y, W) \leq 4(b_1 + b_2) = 4(N-2)\left(9p_p^2 + 3p_p p\right) \quad (1)$$

The TVD is a quite general distance measure between two PDFs. The TVD is defined here as the maximum absolute difference between the probability assigned by Y and the probability assigned by W to any specified subset of possible integer values. In the current scanning context, the most important subset of possible values to consider is the single value {0}, which would imply that the pattern {1 x 1} never occurred (occurred 0 times) in the N - 2 overlapping trials. Then, the SC method in this context uses the Poisson approximation to assign a value to P {0} and the SC method ensures that the Poisson approximation to P {0} is quite accurate, as shown below.

According to the Poisson approximation, P ({1 x 1} never occurs) = $e^{-\lambda}$. For example, using $N = 1000$ and $p = 0.01$, $\lambda = 998p_p = 0.0998$, then $e^{-\lambda} = 0.905$ is the approximate probability that the pattern never occurs, with a SC-based bound of $4(N-2)\left(9p_p^2 + 3p_p p\right) = 0.0123$. Therefore, the maximum difference between the true probability defined by the Y random variable and the approx-

imate probability assigned to any subset of the possible number of occurrences of {1 x 1} defined by the approximating W (Poisson random variable) is 0.01236. So, for example, if the probability that ({1 x 1} never occurs) = $e^{-\lambda}$ = 0.905, then the true probability of 0 occurrences of the pattern is between 0.89 and 0.92. The next section uses simulation to confirm the quality of the SC approximation in Equation (1) in this context.

4. Simulation Results

This section provides simulation results for four examples. Example 1 was given in Section 3. Example 2 uses $N = 5$ and $p = 0.02$ where the exact PDF can be derived analytically by hand fairly easily. Example 3 uses $N = 336$ and $p = 0.02$, similar to that observed in the data in Figure 2. Example 4 is the same as Example 3, but increases N to $N = 10^4$ and is close to the actual application that motivated this investigation. Note that for small values of N such as $N = 5$, the 2^N possible patterns can be enumerated and the fraction of patterns for which the specified pattern of interest such as {1 1 1} starting at any position 1, 2, or 3 can be calculated exactly; therefore, simulation is not necessary (but also still can be done for comparison and completeness) in order to calculate the probability that {1 1 1} occurs at least once in a sequence of $N = 5$ binary values.

The simulation results have shown next each used 10^6 repeated sets of N Bernoulli trials. The **Appendix** provides example R code to do the simulations and calculations [1].

Example 1

Use $N = 1000$ and $p = 0.01$, then $\lambda = 998 p_p = 0.0998$, and $e^{-\lambda} = 0.905$.

The simulation-based P (0 occurrences of {1 x 1}) = 0.907, and the Poisson-based approximation gives 0.905 with a SC-based TVD bound from (1) of 0.0123.

Example 2

Use $N = 5$ and $p = 0.02$, then the exact PDF can be computed analytically fairly easily by finding the 17 of 32 possible patterns of 0 - 1 values in $N = 5$ positions. The analytically-derived exact (and the simulation-based) PDF both assign 0.999 to 0 occurrences of {1 x 1} and 0.001 to 1 occurrence. The SC-based Poisson approximation also assigns probability 0.999 to 0 occurrences and 0.001 to 1 occurrence. The SC-based TVD bound from (1) is 0.0003.

Example 3

As for the data in Figure 2, use $N = 336$ and $p = 0.02$. The simulation-based PDF assigns 0.88 to 0 occurrences of {1 x 1} and 0.12 to 1 occurrence. The SC-based Poisson approximation assigns probability 0.87 to 0 occurrences and 0.12 to 1 occurrence. The SC-based TVD bound from (1) is 0.034, in this case with $N = 336$ and p = 0.02. For the data in Figure 2, the pattern {1 x 1} actually occurred zero times in the 336 -2= 334 trials.

Example 4

Example 4 is the same as example 3, but $N = 10^4$ and $p = 0.0032$, so

$\lambda = 9998 p_p = 0.102$, which is nearly the same value of λ as in Example 1, but this example has quite large N and quite small p. The simulation-based PDF assigns 0.903 to 0 occurrences of {1 x 1} and 0.091.

Example 4 is close to the real application that motivated this paper, and for that length time series, the exact method's combinatorial counting (as was done in Example 2 where $N = 5$) is prohibitively unwieldy, so the SC bound becomes indispensable.

5. Conclusions and Summaries

This paper applied the SC method to approximate the PDF for the number of occurrences of an example pattern in an independent binary time series. In scanning for whether a pattern such as {1 x 1} occurs starting at index i, there are overlapping tries to achieve the pattern, resulting in many non-independent trials consisting of the values in three successive indices. As the time series length increases and the probability $p = P(I_i = 1)$ decreases, the SC method shows that the Poisson approximation is excellent, with a small total variation distance bound, just as in the case of many independent trials, each with small success probability.

The SC bound does not seem to be well known among practitioners; however, related references are available [4] [5] [6] [7]. Reference [4] applies the SC method to calculate coincidence probabilities. References [5] and [6] apply the SC method in different time series contexts than ours. Reference [7] applies the SC identity $Xf(X) = \mu E(f(X+1))$, where X is a Poisson(μ) random variable and E denotes expected value and $f()$ is any bounded function defined on the nonnegative integers, to simplify calculation of bivariate Poisson moments. The SC identify was used to develop the SC approximation method used in this paper.

The main contribution of this paper is to show that the SC bound can defend use of the simple Poisson approximation in real applications (as opposed to unwieldy combinatorial calculations as in Example 2 for larger time series lengths N), and provide a very small bound on the approximation error. Example 4 is close to the real application that motivated this paper, and for that length time series, the exact method's combinatorial counting (as was done in Example 2 where $N = 5$) is prohibitively unwieldy, so the SC bound becomes indispensable.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

[2] Arratia, R., Goldstein, L. and Gordon, L. (1990) Poisson Approximation and the Chen-Stein Methods. *Statistical Science*, **5**, 403-434. https://doi.org/10.1214/ss/1177012015

[3] Shumway, R. and Stoffer, D. (2016) Time Series Analysis and Its Applications with R Examples. 4th Edition, Springer, Pittsburgh. https://doi.org/10.1007/978-3-319-52452-8

[4] Sahatsathatsana, C. (2017) Applications of the Stein-Chen Method for the Problem of Coincidences. *International Journal of Pure and Applied Mathematics*, **116**, 49-59. https://doi.org/10.12732/ijpam.v116i1.5

[5] Kim, S. (2000) A Use of the Stein-Chen Method in Time Series Analysis. *Journal of Applied Probability*, **37**, 1129-1136. https://doi.org/10.1239/jap/1014843092

[6] Aleksandrov, B., Weis, C. and Jentsch, C. (2021) Goodness-of-Fit Tests for Poisson Count Time Series Based on the Stein-Chen Identity. *Statistica Neerlandica*, 1-30. https://doi.org/10.1111/stan.12252

[7] Weis, C. and Aleksandrov, B. (2020) Computing (Bivariate) Poisson Moments Using Stein-Chen Identities. *The American Statistician*, 1-6. https://doi.org/10.1080/00031305.2020.1763836

Appendix. Example R Code to Illustrate the Simplicity of the SC Method

```
p = 0.01; N = 10^3; nsim = 10^6; temp = numeric(nsim)
# record how many 1 0 1 occurrences in N − 2 overlapping trials.
for(isim in 1:nsim) {
  x = as.numeric(runif(N) < p) # simulated N independent Bernoulli trials
for(i in 1:(N − 2)) {
if(x[i]==1 && x[i + 2]==1) {temp[isim] = temp[isim] + 1}
  }
}
mean(temp==0)
lam = (N − 2) * p^2
exp(−lam)
# bound
p_p = p^2; 4*(N − 2) * (9 * p_p^2 + 3 * p_p * p)
```