Searching for Signal in Experimental Data

Ramana Dodla and Charles J. Wilson
Department of Biology, University of Texas at San Antonio, San Antonio, TX
[SciTopics. http://scitopics.com/Searching_for_signal_in_experimental_data.html]

Motivation and summary

One is often confronted with very brief data that is discrete and inhomogeneous in time to act upon. The number of data points could be as few as 10 or fewer. In neuroscience, determining the interdependence of timing of spike events is of central importance in temporal coding. The spike times which are recorded in many physiological experiments (in vivo/in vitro) are inhomogeneous in time, and in several occasions it is advantageous and economical to consider very few data samples and judge the data quality. Such discrete inhomogeneous data-sets may also occur in biophysics, geology, sociology, electrical circuit theory, nonlinear dynamical studies of chaos, signal processing, and other disciplines where collection of data is central. We describe here a phase function method that can detect periodicity or lack there of in data-sets that have very few data points which are not amenable for correlation methods. The method is not limited to treating only short data-sets, but can also be applied to large data sets, and when compared with the results obtained from auto and cross correlation function methods, the results using phase functions are significantly more rewarding in quality. The phase functions are also good candidates for instant computation of correlations during an experiment.

An example data set

An example data set from an experiment is shown in Table 1. These are the times (in seconds) at which a rat Globus Pallidus neuron showed voltage spiking in vitro. The data is time shifted such that the first spiking occurred at 0. (Such shift is however not necessary for our analysis.) Sequences of such spike times are usually collected from spiking brain neurons. Determining whether there is periodicity in the data or not is of central importance in understanding the functional circuit mechanisms between different brain areas. Often long sequences (from 100's to 1000's) of spike times are collected to analyze the nature of the data. For a neuron that spikes at 10 Hz, one collects data from 10's of seconds to minutes. In contrast, our data sample shown in Table 1 is collected for slightly more than half a second. The interest in the dynamics of such short samples is necessitated by the fact that several functions in the brain last only for a short duration.

Table 1: Spike events occurred at these times (in seconds). Are these events random or regular?
0
0.18595
0.33955
0.4997
0.61495

Mean measure and auto-correlation function are insufficient

The duration between successive data points, also called the interspike intervals, ISIs (0.18595, 0.1536, 0.16015, 0.11525) is not constant, and hence it is not obvious to judge whether this data represents a periodic signal or a random signal. The mean value of these intervals (0.1537375 s) does not indicate periodicity either because randomizing the order of these intervals still gives the same mean value.

A popular method to determine the underlying periodicity of a given data set is employing the autocorrelation function. This method can be extended to discrete data. But a reliable computation of auto- or cross correlations of data requires large data-sets. When applied to fewer data points, the results are unreliable because the computation requires a choice of a bin width that is used to count the number of data pairs that can be bracketed in a given bin, and the accuracy of the profile of the autocorrelation function is affected by the choice of the bin width. For a choice of very small bin width, the auto-correlation function computed for the data set of 5 points shown in Table 1 is shown in Fig. 1. The profile consists of delta-function looking impulses. Increasing the bin width does not provide a profile that is useable in predicting the periodicity or time constant of the correlations.

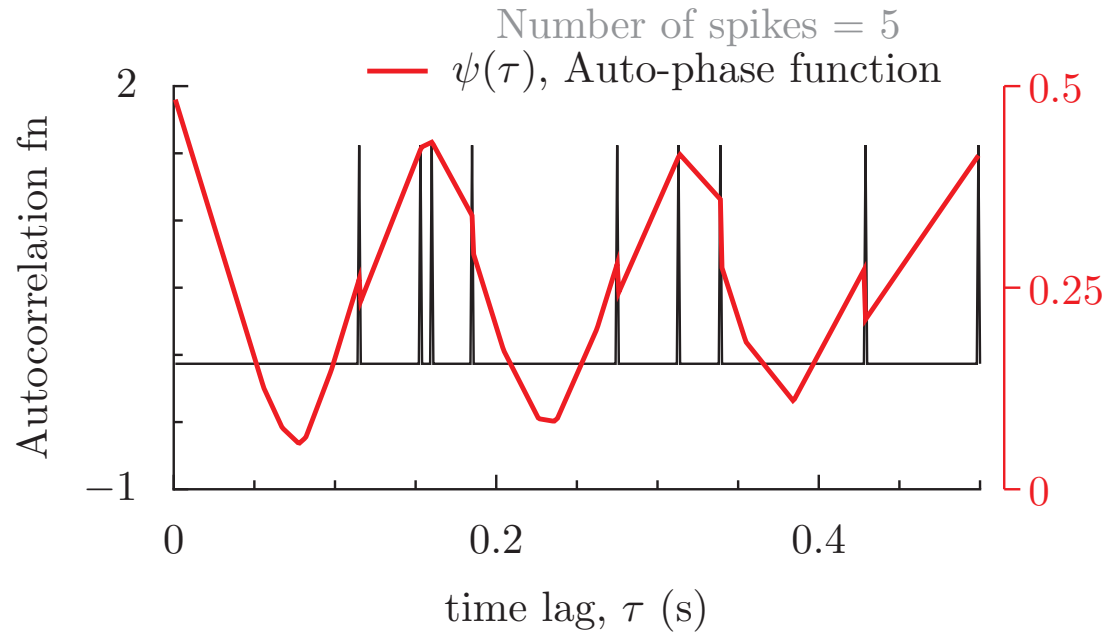

Figure 1: Comparing autocorrelation function (ACF) and autophase function for the data of Table 1 that has just 5 points. The functions are computed for $\tau > 0$. (At $\tau = 0$, the ACF acquires a value of 8.06, and the phase function acquires a value of 0.) The phase function is computed at τ increments of 0.001 s, and the autocorrelation function histogram uses a bin width of 0.001 s. Changing the histogram bin width changes the ACF profile but the phase function is unaltered by altering the resolution in τ. These parameters are the same for Figs. 2 and 3.

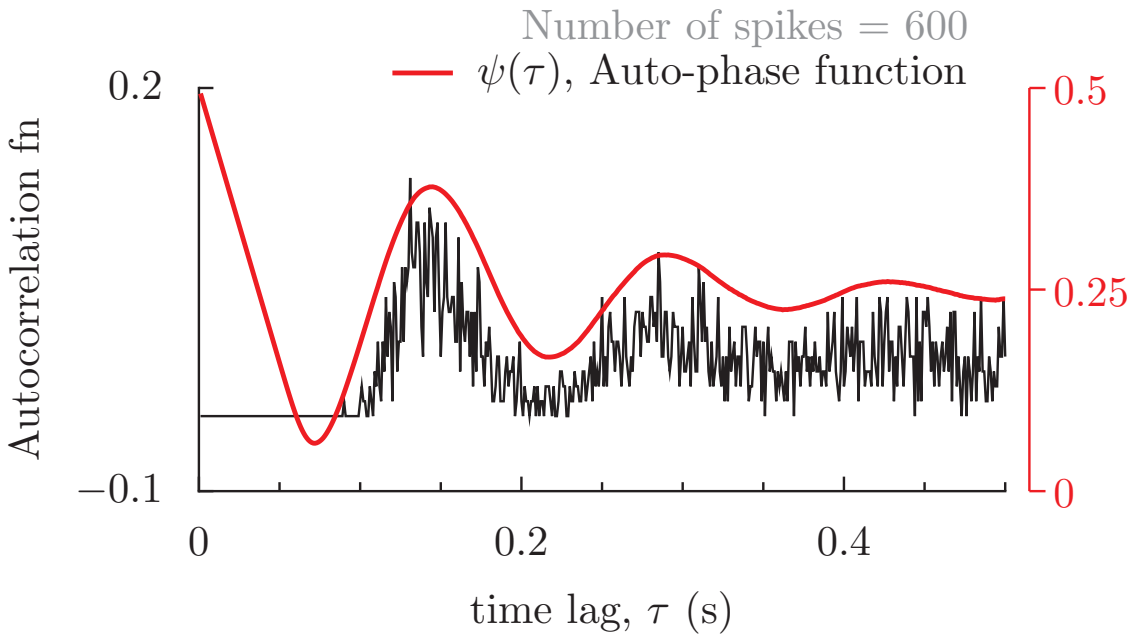

Figure 2: Comparing the autophase function and the autocorrelation function for a spike train sequence with 600 points. This is the same data set whose sample is shown and analyzed in Table 1 and Fig. 1. The functions are computed for $\tau > 0$. All the other parameters are as described in Fig. 1.

Introducing a "phase function"

We introduce the concept of a phase function [$\psi(\tau)$] to determine the correlations and the periodicity inherent in the underlying discrete data [1]. The phase function is termed "auto-phase function" when applied to single data set, and a "cross-phase function" when two different data-sets are used for computing the phase function between

them. The phase function does not use binning and hence difficulties introduced by bin width are absent. It does all what autocorrelation or cross-correlation functions do for discrete samples such as spike times, but does much better than them when the sample size is very small. The phase function is computed with τ as the lag parameter. When applied to the brief dataset of 5 points of Table 1, $\psi(\tau)$ displays a clear periodic profile (red line in Fig. 1) thus discovering a signal that could have been missed otherwise. When the method is applied to a much larger data set of 600 data points (Fig. 2), its profile completely recovers the autocorrelation function profile, but is now much more smoother than the corresponding auto-correlation function. There is no explicit smoothing carried out on the phase function. The decay time of the correlations can be estimated from the profile of the autophase function more easily than from the auto-correlation function.

A gallery of phase functions computed for different lengths of a single data set is shown in Fig. 3 and the corresponding auto-correlation functions are also shown. The phase function performed better, and is amenable for analysis that determines the frequency and timescale in the data.

Phase function method and properties

The phase function $[\psi(\tau)]$ is defined between two event trains. For an autophase function, the second train is derived from the first by time shifting (τ). The time shift/time lag is the same as the time lag in a correlation function computation. A phase pair is computed at event times and is represented as a point in a two dimensional phase plane. Then the phase function is defined as the average of the drift represented by each phase pair from the mean of the phase pairs. This is different than a "distance" between the spike trains. A "distance" usually involves taking the difference between the vectors represented by the event trains. Phase function can be computed for two spike trains that have different number of spike times.

The basic idea behind deriving phase pairs from the spike times can be described by assuming that the voltage trajectory of the neuron that is emitting the spike times is completing a full phase-space orbit between the successive spike times. So we ask what fraction of that orbit is completed at certain times. And those times are nothing but the spike times belonging to either spike train A or the spike train B. Such fractions are made into phase pairs. The computed phase pairs are represented in a two-dimensional phase plane. The absolute drift of the phase pairs with respect to the their mean is averaged over all the phase pairs, and the resultant quantity is normalized to represent the phase function value for the given configuration of the spike train pairs. The lag (τ) between the spike train pairs is a parameter and is increased in magnitude similar to that in computing a correlation function.

Lag $\tau = 0$ produces a singular behavior because each component of the phase pairs becomes identical to unity. Hence each phase pair point coincides with the mean itself. Thus at lag 0, the phase function becomes 0, i.e. $\psi(\tau = 0) = 0$. It can also be shown that $\psi(\tau \to 0^+) = \psi(\tau \to 0^-) = 0.5$. For an asynchronous state defined by a uniform distribution of the phases, we can show that $\psi(\tau \to \infty) = 0.25$. However not all asynchronous states can be described by this approximation. For a given pair of spike train sequences, the phase function reaches an appropriate state for $\tau \to \infty$. The oscillation of $\psi(\tau)$ for non-zero values indicates the inherent correlations. An absence of correlations is indicated by the absence of oscillations in the phase function. The initial decay seen near $\tau = 0^+$ does not indicate the correlations in the data, but the decay of the phase correlations from the asynchronous state represented by the phase configuration at $\tau = 0^+$.

Computing the phase pairs

We illustrate the computation of phase pairs with two example spike trains. Suppose we have two spike train sequences A and B which are time shifted by τ, and are represented by $A = \{A_1, A_2, A_3, A_4\}$, $B = \{B_1, B_2, B_3, B_4\}$. When time ordered, let the order be given by $A_1 \le B_1 \le B_2 \le B_3 \le A_2 \le A_3 \le A_4 \le B_4$. Then we compute phase pair points $Q(\gamma_i, \delta_i)$ at the spike times B_2, B_3, A_2, A_3, A_4 (i.e. leaving the first two and the last spike times) as follows. At B_2: the time elapsed for the $A-$train since a last spike time is: $B_2 - B_1$. But A's own spike period is: $A_2 - A_1$. We represent the resulting phase of the $A-$train at the spike time B_2 by $\gamma_1 = \frac{B_2 - B_1}{A_2 - A_1}$. Similarly the time elapsed for $B-$train since a last spike time is the same: $B_2 - B_1$. But B's own spike period is: $B_2 - B_1$. The resulting phase of the $B-$train at the spike time B_2 is given by $\delta_1 = \frac{B_2 - B_1}{B_2 - B_1}$. Hence the phase pair at B_2 is: $(\gamma_1, \delta_1) = (\frac{B_2 - B_1}{A_2 - A_1}, \frac{B_2 - B_1}{B_2 - B_1}) = (\frac{B_2 - B_1}{A_2 - A_1}, 1)$. Similarly, at B_3: $(\gamma_2, \delta_2) = (\frac{B_3 - B_2}{A_2 - A_1}, \frac{B_3 - B_2}{B_3 - B_2}) = (\frac{B_3 - B_2}{A_2 - A_1}, 1)$, at A_2: $(\gamma_3, \delta_3) = (\frac{A_2 - B_3}{A_2 - A_1}, \frac{A_2 - B_3}{B_4 - B_3}) = (\frac{A_2 - B_3}{A_2 - A_1}, \frac{A_2 - B_3}{B_4 - B_3})$, at A_3: $(\gamma_4, \delta_4) = (\frac{A_3 - A_2}{A_3 - A_2}, \frac{A_3 - A_2}{B_4 - B_3}) = (1, \frac{A_3 - A_2}{B_4 - B_3})$, and at A_4: $(\gamma_5, \delta_5) = (\frac{A_4 - A_3}{A_4 - A_3}, \frac{A_4 - A_3}{B_4 - B_3}) = (1, \frac{A_4 - A_3}{B_4 - B_3})$.

A phase cannot be computed for times before B_1 or after A_4; The end points cannot form phase pairs because at least one of the neurons has no evolution in that regime. However, the information in those points is not lost because for an appropriate τ value those points also contribute to the profile of $\psi(\tau)$.

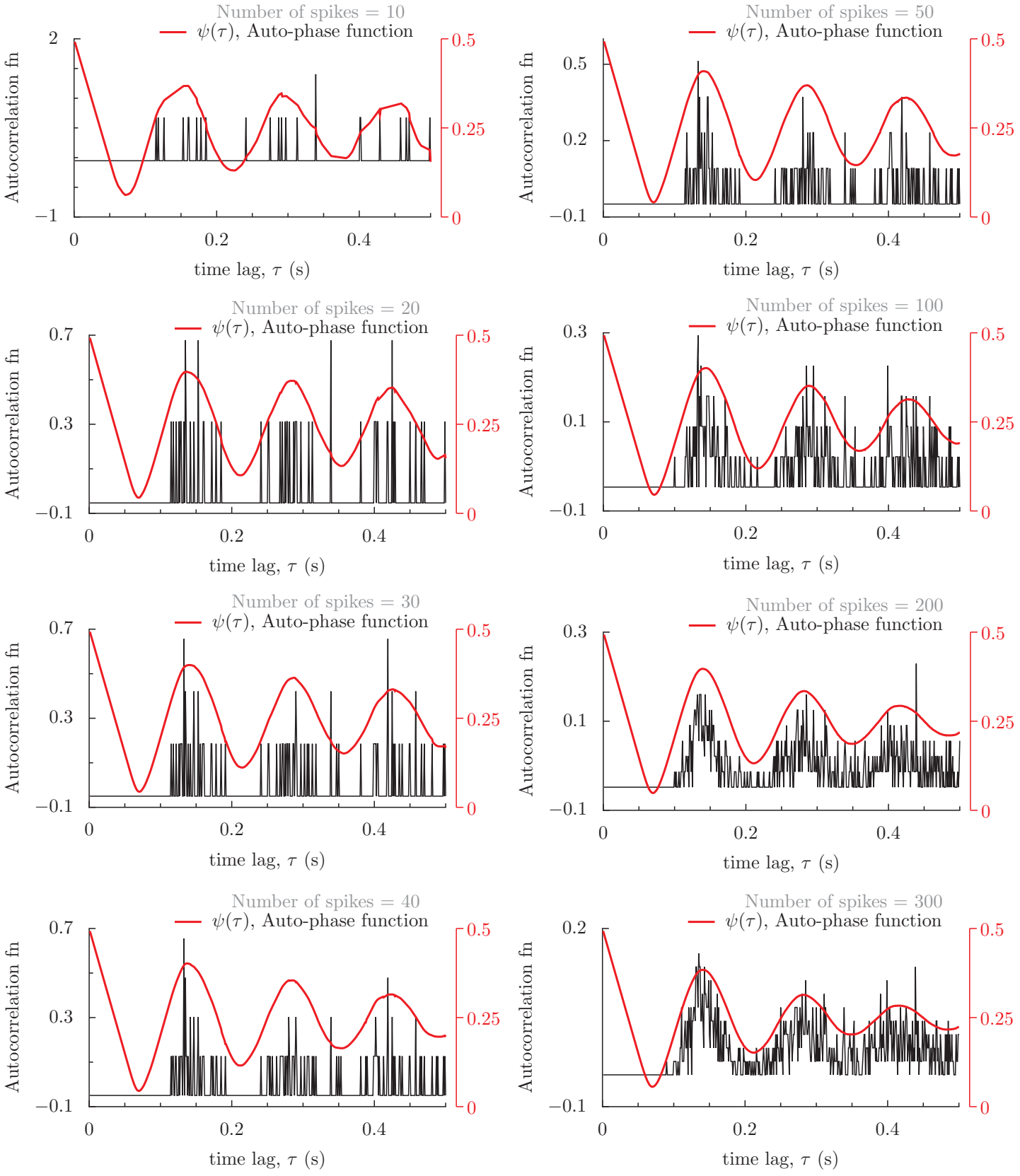

Figure 3: Comparing the autophase function and autocorrelation function computed for a single spike train by including in the computation increasing number of spike time events. The functions are computed for $\tau > 0$. The dataset used is the same as that whose sample is shown and analyzed in Table 1 and Fig. 1.

Computing the phase function

Once the phase pairs are computed, the phase function is evaluated by finding the average of "drift" experienced by each phase pair from their mean. This can be visualized geometrically by plotting the phase pairs in a two-dimensional phase plane. Imagine the phase pairs as a swarm of bees around a beehive located on the branch of a tree. The drift experienced by each bee is the absolute (outward or inward) distance along the branch they find themselves with respect to the center of the beehive. The phase function is now proportional to the mean of all those drifts normalized with the maximum such value a phase pair can ever affect ($\sqrt{2}$). The average phase point of the phase pairs is: $P(\langle\gamma\rangle, \langle\delta\rangle) = \left(\frac{1}{N}\sum_i^N \gamma_i, \frac{1}{N}\sum_i^N \delta_i\right)$, where $N = 4$ for the above example. Then the phase drift experienced by the phase pair (γ_i, δ_i) is $r_P - \frac{\gamma_i\langle\gamma\rangle + \delta_i\langle\delta\rangle}{r_P}$, where $r_P = \sqrt{\langle\gamma\rangle^2 + \langle\delta\rangle^2}$. We term this drift as $c_i(\tau)$ because the relative configuration of the spike trains is posited at a lag τ. The phase function is then defined as $\psi(\tau) = \frac{1}{\sqrt{2}N}\sum_{i=1}^N \mid c_i(\tau) \mid$.

Application to Poisson spike train sequence

Finally we present in Fig. 4 computation of the phase function for a sequence of events whose intervals are governed by a Poisson process of 1 Hz. For large number of samples the phase function shows a steady profile and has no oscillations indicating absence of detectable periodicity or correlations. For small number of samples, it is possible to discover local periodicity inherent in the spike times.

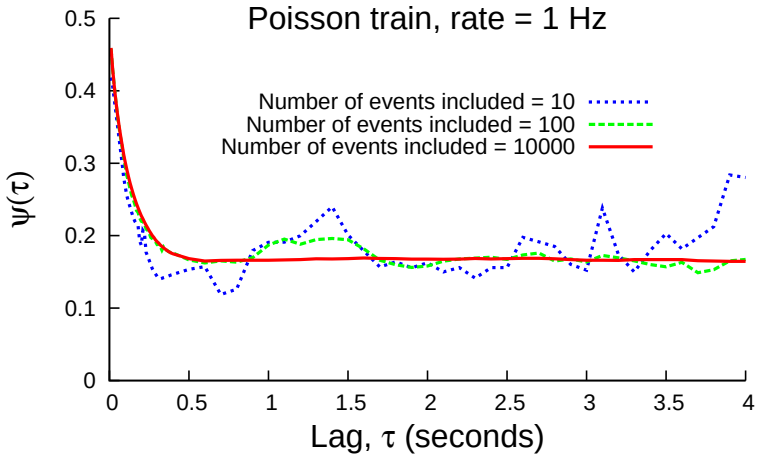

Figure 4: Autophase function for a 1 Hz Poisson spike train sequence when different lengths are included in the computation. No averaging over the results using multiple realizations is carried out. It is possible to find detectable correlations in the form of phase function modulations for brief data, but when large number of samples are included, the phase function shows no oscillations and reaches a steady profile indicating absence of any correlations. The initial decay seen near $\tau = 0$ does not indicate correlations in the data, but the decay from an asynchronous state near $\tau = 0$.

Reference

[1] Ramana Dodla, Charles J. Wilson. A phase function to quantify serial dependence between discrete samples. *Biophysical Journal* **98**:L05-L07, 2010.