Damaschke *et al.*
EURASIP Journal on Advances in Signal Processing (2024) 2024:17
https://doi.org/10.1186/s13634-024-01108-4

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Bias-free estimation of the covariance function and the power spectral density from data with missing samples including extended data gaps

Nils Damaschke[1], Volker Kühn[2] and Holger Nobach[3*]

*Correspondence:
holger.nobach@nambis.de

[1] Faculty of Computer Science and Electrical Engineering, Institute of General Electrical Engineering, University of Rostock, Albert-Einstein-Straße 2, 18059 Rostock, Germany
[2] Faculty of Computer Science and Electrical Engineering, Institute of Communications Engineering, University of Rostock, Richard-Wagner-Straße 31, 18119 Warnemünde, Rostock, Germany
[3] Max Planck Institute for Dynamics and Self-Organization, Am Faßberg 17, 37077 Göttingen, Germany

Abstract

Nonparametric estimation of the covariance function and the power spectral density of uniformly spaced data from stationary stochastic processes with missing samples is investigated. Several common methods are tested for their systematic and random errors under the condition of variations in the distribution of the missing samples. In addition to random and independent outliers, the influence of longer and hence correlated data gaps on the performance of the various estimators is also investigated. The aim is to construct a bias-free estimation routine for the covariance function and the power spectral density from stationary stochastic processes under the condition of missing samples with an optimum use of the available information in terms of low estimation variance and mean square error, and that independent of the spectral composition of the data gaps. The proposed procedure is a combination of three methods that allow bias-free estimation of the desired statistical functions with efficient use of the available information: weighted averaging over valid samples, derivation of the covariance estimate for the entire data set and restriction of the domain of the covariance function in a post-processing step, and appropriate correction of the covariance estimate after removal of the estimated mean value. The procedures abstain from interpolation of missing samples as well as block subdivision. Spectral estimates are obtained from covariance functions and vice versa using Wiener–Khinchin's theorem.

Keywords: Bias-free estimation, Covariance, Spectrum, Missing samples, Data gaps

1 Introduction

In normal operation, measurement instruments usually provide a continuous stream of equidistantly spaced samples of a physical quantity being observed. Common signal processing algorithms for statistical analysis usually rely on a continuous data stream, typically arranged in blocks of a defined duration. However, there are several reasons why normal operation may fail. There may be boundary conditions, under which the measurement system cannot operate. This includes cases, where the quantity under observation is temporarily inaccessible. For distributed measurement systems, the

communication channels may be temporarily disrupted. The instrument may also need to be reconfigured or shut down for maintenance. In other cases, the measurement principle involves a signal pre-processing that may fail under certain conditions, resulting in outliers or gaps in the data stream. A real case was a long-term environmental measurement with distributed sensors communicating over a wireless mesh. From time to time, packets of data got lost along the way, and occasionally, a microcontroller went suddenly through a reset cycle.

Data sets with missing samples or longer data gaps can be understood as the product of the uninterrupted process x_i with a specific sampling function w_i, which is the train of valid instances with unit amplitude. This way, the statistical properties of the data set with interruptions $w_i x_i$ deviate from those of the process under investigation. The discrete Fourier transform (DFT), e.g., obtained from such a signal is the convolution of the DFT of the process with the DFT of the sampling function. Therefore, the sampling function and its statistical properties have a direct influence on the estimated statistics of the measured signal as illustrated in Fig. 1. For missing samples occurring independently of each other, the covariance function and the power spectral density are different from those of longer data gaps. While the sampling function with independent outliers has no

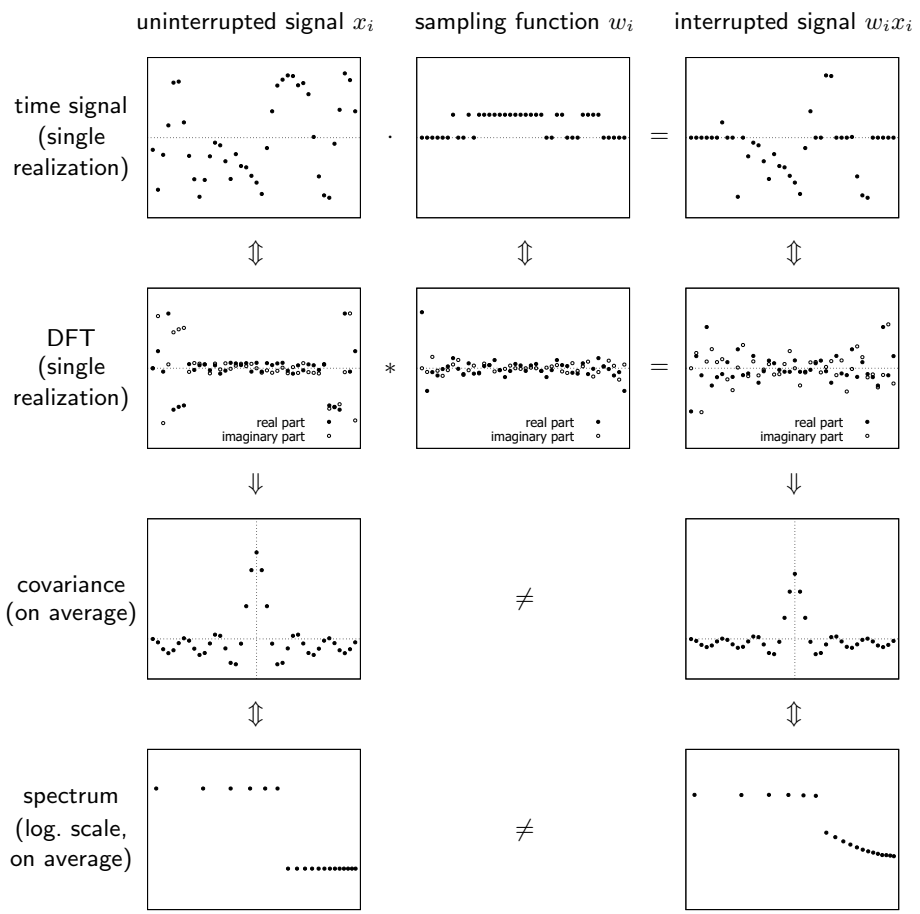

Fig. 1 Masking (multiplying) a signal with a sampling function corresponds to a convolution of the respective discrete Fourier transforms, finally leading to deviations in the covariance functions and power spectral densities between the uninterrupted and the interrupted signal

covariances between two different time instances and has a flat white spectrum, longer data gaps lead to temporal covariances and a colored spectrum as shown exemplarily in Fig. 2. The terms "independent outliers", "random outliers", "uncorrelated outliers" or "missing data points without correlation" are used as synonyms for this particular kind of a sampling function in the present paper. In contrast, "longer data gaps", "extended data gaps" or "correlated data gaps" are used for any kind of sampling function different than the first one.

An obvious way to circumvent the problem caused by missing data samples is to interpolate the signal and fill the gaps with predicted values, see [1]. The interpolation scheme can mime the statistical properties obtained from the valid parts of the signal and missing values can be predicted. Then the statistical quantities like the covariance function or the power spectral density are derived from the reconstructed signal consisting of a mixture of originally valid samples and the interpolated ones. Examples for this principle are Kalman interpolation in audio reconstruction [2, 3], the adaptive filter-bank approach [4] or the Karhunen–Loève procedure resp. proper orthogonal decomposition for gappy data [5] or in turbulence measurements [6]. However, even the best interpolation in terms of the minimum prediction error, understood as the minimum mean square error between the interpolated signal and the true signal, will lead to a significant

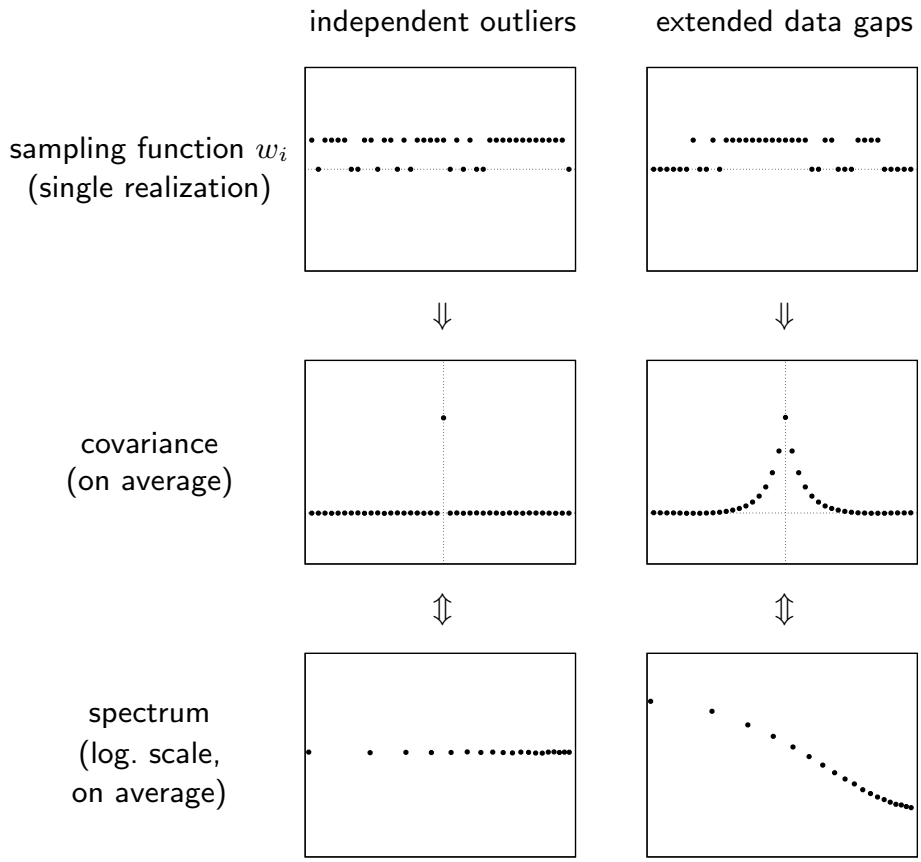

Fig. 2 Different statistical properties of the sampling function with information on missing data points: Independent outliers without correlations between each other yielding a white noise spectrum vs. extended data gaps with correlations yielding a colored spectrum

dynamic error depending on the probability of invalid data points. The prediction will ideally replace missing samples by something like the expectation of all possible continuations of the signal at the respective time instance. Then the interpolation itself would be bias-free; however, it will suppress parts of the fluctuations of the true signal. Finally, statistical properties of the partially interpolated signal deviate from those of the original signal. For rare and short gaps in the data sequence, this may work sufficiently; for more and longer gaps, the error may easily become unacceptable. Note that this holds for any interpolation scheme, even for those, which perfectly mimic the spectral composition of the signal under observation.

In [7] the expected dynamic error is derived particularly for sample-and-hold interpolation including empirically obtained parameters. In [8] the findings could be reproduced with the parameters derived algebraically. For an assumed covariance function of the uninterrupted signal, the covariance function after interpolation of missing data points is predicted based on the statistical properties of the occurrence of missing data points. The primary covariance function estimated from the interpolated signal then can be improved by the inverse of this correspondence. Under ideal conditions, this procedure entirely inverts the dynamic error caused by the interpolation. Therefore, the consideration and inversion of the influence of the interpolation step is a promising method to obtain bias-free estimates of the covariance function and the spectrum. However, the derivation of the particular correspondence depends on the specific statistical properties of the missing data points, namely a random occurrence of invalid samples which are independent of each other. If the statistics of the data gaps change, then the procedure needs substantial modification. Therefore, it is no universal solution.

In [9] no interpolation has been used. Instead the expected spectrum of the discrete Fourier transform has been derived for the signal with missing samples or data gaps of various statistical characteristics. From these expectations, a deconvolution could have been developed to improve the estimation of the spectra, similar to the deconvolution after interpolation in [7]. Instead, a procedure has been used avoiding the data gaps by rejecting all values past the first missing sample by means of zero padding. Depending on the probabilities of the occurrence of the first missing sample, the spectra get different resolution and the superposition of many spectra from individual data blocks becomes smeared. However, this method is very inefficient in using the information available, since significant amount of valid samples get rejected. It is limited to small amounts of missing samples anyway, since series of valid samples become too short for increasing amounts of missing samples.

There exists a wide variety of direct spectral estimators optimized for spectral estimation from a limited (typically small) number of unevenly sampled observations of signals [10–19]. They are widely counted as direct spectral estimators, since the amplitudes of the spectrum are obtained directly at any frequency from the sinusoidal fits. All these estimators potentially are able to process also signals with missing samples. However, the spectral composition of equidistantly sampled signals with independently missing samples deviates from random sampling in continuous time, not to mention correlated data gaps, which the above methods cannot handle accurately. Since Lomb–Scargle's method [10, 11] is widely used as a benchmark, it is included in the comparison below to prove it as biased.

Missing data samples in equidistant data streams have also been investigated broadly in [20–28], including also specific cases of correlated data gaps. These derivations strictly depend on the specific cases of missing data and are not robust against changes in spectral content of the data gaps. This also holds for [29–33], where parametric estimation has been performed. This way effective process identification is possible in a limited search space. However, bias-free estimation is not possible for unknown or changing spectral content of the data gaps.

All the methods mentioned above make bad compromises at the one or the other point. They are either biased, inefficient or they are limited to a specific sampling scheme. The present article introduces bias-free estimators for the covariance function and the power spectral density from equidistant data sets with interruptions of arbitrary spectral composition. It is a combination of three known but rarely used methods, namely a) weighted average taken from [25], except for any spectral or time windowing to circumvent any modulation of the spectrum by filtering resp. windowing, b) restriction of the domain of the covariance function, mentioned briefly in [34] as an appropriate means for reducing the estimation variance of the spectral estimates and c) correction of the covariance estimate after removal of the estimated mean value, adapted to the weighted average to work with gapped data [35–37]. This combination allows bias-free estimation of the covariance function and the power spectral density from signals with data gaps, independent of the spectral content of the data gaps and with an optimum use of the available information. The combination is especially advantageous, because the individual processing steps benefit from each other. Weighted averages discriminate missing data pairs, possibly leading to missing values in the covariance estimate, especially in the ranges of long time lags. Restricting the domain of the covariance function before its transformation into a spectrum first uses all valid pairs of data from the entire data set, resulting in better converged estimates of the individual values of the covariance estimate, and then only the most reliable are going into the transformation. Finally, nonzero mean values in combination with missing data lead to an increased estimation variance of the spectral estimate. Subtracting the mean value on the other side leads to biased covariance estimates. Here, the last part of the proposed method, the correction of the covariance estimate after removal of the estimated mean value, is the solution.

The invalid samples are assumed *a priori* known, given by an additional flag (weight) for each data sample indicating whether the sample is valid or invalid. No pre-knowledge about the characteristics of the process under investigation or the sampling scheme are needed. The procedures work independent of the statistical characteristics of the data gaps, except that pairs of data points exist at all required lag times for estimating the covariance function. It is assumed that the data gaps (randomly and independently occurring outliers as well as correlated data gaps) are not containing information from the data, namely that the validity of the samples is independent of the values of the observed process, so there is no preferential sampling.

Note that the introduced processing methods are suitable for signals from stationary stochastic processes only and for their statistical analysis. They are not suited for signals with time-dependent statistical properties, e.g., single pulses. For those, missing information cannot be restored by the methods presented here. Interpolation of

missing portions of the signal is not required and, neither interpolation of the data gaps nor any kind of reconstruction of the signal are intended.

Note further that the weighted averages make different discrimination of the data ensemble at different lag times. This and also bias correction for empirical mean removal result in correlation matrices, which potentially may violate the nonnegative definiteness. As a consequence, negative values occur in the corresponding power spectral density. Since the introduced procedures yield bias-free estimates for both, the covariance function as well as the spectrum, averages over multiple estimates of these functions will converge toward the true functions of the underlying process. However, since the procedure is consistent, the estimated functions will also converge toward the true functions if applied to single but longer data sets, without losing information between block boundaries. The ultimate solution of course is regularization. Since this inevitably introduces a bias in both the covariance function and the corresponding spectrum, regularization is not considered in the present paper, where bias-free estimation has priority.

The programs used here are available as supplementary material to the present paper at [38]. The implementation of the covariance and spectral estimations is done consistently based on FFT routines to achieve sufficient data throughput.

2 Methods

2.1 Weighted average

Assuming an equidistantly spaced signal x_i with N samples ($i = 1 \ldots N$), then the covariance function γ_k is

$$\gamma_k = \big\langle (x_i - \mu)(x_{i+k} - \mu) \big\rangle \quad k = -(N-1) \ldots N-1 \tag{1}$$

with the expectation $\langle \cdot \rangle$ of the product $\cdot$ and with the expectation of the signal $\mu = \langle x_i \rangle$. Note that x_i is the population of the generic signal instead of a particular realization. Note further that the autocovariance function is a function of the lag distance k between the samples x_i and x_{i+k}, and it is symmetrical about $k = 0$. That is why it is often given as the one-sided covariance function for $k = 0 \ldots N-1$. However, the estimation procedures can be adapted to cross-covariance functions between two signals also, where symmetry is not given.

If all samples x_i in a particular realization of the signal are given, then the covariance function at any lag distance k can be estimated from the average of all samples of the data set, which have the distance k. With missing samples, either reconstruction and filling the gaps will help for the price of a bias, as mentioned above, or the averaging process is restricted to those pairs of samples, where both x_i and x_{i+k} are available, as used, e.g., in [25]. This method has the potential to yield bias-free covariance estimates also from signals with gaps. However, the application of window functions in that publication introduces a new bias by modulating the values of the data sequence.

By leaving away the window function and the data modulation that goes with it, the covariance estimate can be formally written as the weighted average

$$C_k = \frac{\sum_{i=1}^{N-|k|} w_i w_{i+|k|}(x_i - \mu)(x_{i+|k|} - \mu)}{\sum_{i=1}^{N-|k|} w_i w_{i+|k|}} \tag{2}$$

with the weights w_i corresponding to the validity of the samples x_i ($w_i = 1$ for a valid sample and $w_i = 0$ for an invalid one). C_k denotes the covariance estimate and x_i and w_i in the sums are one particular realization of the signal and the corresponding weights.

This estimate of the covariance at lag distance k is the mean value of all products $(x_i - \mu)(x_{i+|k|} - \mu)$, which are available from the data set, where both values x_i and $x_{i+|k|}$ are valid, corresponding to $w_i w_{i+|k|}$ being one. In other words, this estimate calculates the sum of all such products of values with the respective distance available from the data set, and divides the sum through the number of the pairs counted in the sum in the numerator. As long as the sampling function w_i is independent of the signal x_i, the expectation of the estimate C_k then is

$$\langle C_k \rangle = \Big\langle (x_i - \mu)(x_{i+|k|} - \mu) \Big\rangle, \tag{3}$$

which is identical to the true covariance γ_k. Hence, C_k is a bias-free estimate of γ_k independent of the characteristics of the sampling function, as long the expectation μ of the signal is given and the sampling function is independent of the signal. This estimator is proposed and intensively investigated in terms of its characteristics in the present paper, especially under the condition of various spectral characteristics of the sampling function as formed by the series of weights. For the first tests for this estimator in Sect. 3, the expectation μ of the signal is assumed to be known a priori and correctly removed from the data before further processing. The specific role of an unknown expectation, the estimation and removal of an empirical mean value is discussed in Sect. 2.3. If the sampling function depends on the signal, appropriate weighting schemes are required additionally, as, e.g., in [19], which is not considered in the present paper.

2.2 Restricting the domain of the covariance function

The power spectral density obtained from a single data set has an unacceptable high estimation variance. A common means to reduce this variance is a subdivision of the entire data set into shorter blocks and average the respective power spectral densities obtained from the individual data blocks. This method is known as Bartlett's method, see [34, 39]. By using shorter data blocks, the spectral resolution is reduced accordingly, which leads to the desired effect of reduced estimation variance. A disadvantage of Bartlett's method is that correlations between samples at the end of one block and the beginning of the next are not counted. Furthermore, the wrap-around error may be increased if the assumption is made that the signal respectively the block is periodic. For too short blocks this may lead to significant deviations. In contrast, for longer blocks the reduction of the estimation variance becomes less effective. With Welch's method, see [40], where the statistical functions from overlapping blocks get averaged, correlations between block boundaries are counted. However, this also partially generates redundancy. This fact is taken into account by applying windowing functions to the data blocks prior to

their statistical analysis. Again, the additional data modulation by the window function leads to biased estimates of the statistical functions.

The reduction of the spectral resolution can also be obtained in a post-processing step from the entire data set without the necessity of block subdivision. Reducing the spectral resolution corresponds to a shorter domain of the covariance function. For a random process with a finite memory, the autocovariance function becomes zero at lag times beyond the memory length. In this case the autocovariance function can be shortened to the extent of the longest lasting covariance C_K with $K < N$ without losing information. Equation (4) then is determined for $k = -K \ldots K$ instead of $k = -(N-1) \ldots N-1$. For effectiveness, $K \ll N$ is proposed, where K must consider the correlation interval of the signal, and then in combination with a significantly longer data set. Due to the restricted domain of the autocovariance function, the spectral resolution is reduced accordingly, leading also to a significantly lower estimation variance of the spectrum without introducing new errors from too short a block subdivision or any amplitude modulation of the signal by window functions. The advantage of this method compared to usual block subdivision is that the correlations of all samples are considered from the entire data set without interruptions at block boundaries. Further suppression of the wrap-around error, e.g., via the application of a window function is not needed, avoiding additional modulation of the signal smearing the spectrum. This method is identical to [41, 42] used with a rectangular window applied to the covariance function estimated. Note that this method is known for a long time. It has been shortly mentioned in [34]. However, it has not gained acceptance, even if the results in reducing the estimation variance are comparable to block averaging without the risk of worsening by too short blocks and, it is superior in efficiently using the information available.

Principally, the power spectral density is obtained by means of the DFT from a shorter version of the covariance estimate than that one, which has been obtained by the covariance estimate before. Assuming the covariance estimate being bias-free and no further covariances exist outside the shorter interval from $-K$ through K, the power spectral density will also be bias-free, since the DFT is linear. If covariances exist beyond the interval from $-K$ through K, they will not be counted by the transform into the spectrum, which then gets biased.

2.3 Mean removal and Bessel's correction

Without mean removal, the covariance function would have an offset of μ^2 with μ being the expectation of the signal. Nevertheless, the spectral estimate is unbiased at all frequencies except for zero frequency even with $\mu \neq 0$ (Fig. 7a in advance). However, in contrast to uninterrupted equidistant sampling, in the case of missing samples, a nonzero expectation increases the estimation variance of the spectrum (Fig. 7b in advance). Therefore, it is recommendable to estimate and to remove the mean value from the data before further covariance or spectral analysis. If the empirical mean value $\bar{x}$ is used for that instead of the true expectation μ, a new bias occurs in the estimated covariance function as derived in [35–37], leading to a covariance estimate, which is asymptotically bias-free only. This new bias corresponds to the bias of the estimate of the data variance, if $\bar{x}$ is subtracted from the data instead of μ. For uncorrelated data, the appropriate correction is to divide the sum of

$(x_i - \bar{x})^2$ by $N - 1$ instead of N, which is widely known as Bessel's correction. However, the correction for correlated data is more complex. Fortunately, the mentioned papers also provide an appropriate correction. In reference to the variance estimate from uncorrelated data, the correction here is also denoted as Bessel's correction. Weighted averages are documented in the two last publications only. For the reader's convenience, a brief summary of the procedure applicable to the present case is given here.

Estimating the covariance function as

$$C_k = \frac{\sum_{i=1}^{N-|k|} w_i w_{i+|k|} (x_i - \bar{x})(x_{i+|k|} - \bar{x})}{\sum_{i=1}^{N-|k|} w_i w_{i+|k|}} \tag{4}$$

after estimation and removal of the empirical mean

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} \tag{5}$$

yields an expectation of the covariance function estimated, which is

$$\langle C \rangle = A\gamma \tag{6}$$

with the vector C, which contains all estimated covariance values C_k for $k = -K \ldots K$ and the vector γ, which is the counterpart with the true covariances γ_k for $k = -K \ldots K$. The square matrix A represents a linear relation between γ and $\langle C \rangle$ and it has the elements

$$a_{kj} = \delta_{k-j} + \frac{W_j}{D^2} - \frac{G_{kj} + H_{kj}}{D W_k} \tag{7}$$

with

$$\delta_{k-j} = \begin{cases} 1 & \text{for } k - j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

and

$$D = \sum_{i=1}^{N} w_i \tag{9}$$

$$W_j = \sum_{i=1}^{N-|j|} w_i w_{i+|j|} \tag{10}$$

$$G_{kj} = \sum_{i=\max(1,1-j,1-k)}^{\min(N,N-j,N-k)} w_i w_{i+j} w_{i+k} \tag{11}$$

$$H_{kj} = \sum_{i=\max(1,1-j,1+k-j)}^{\min(N,N-j,N+k-j)} w_i w_{i+j} w_{i+j-k} \tag{12}$$

and W_k similar to W_j with j replaced by k. Then the inverse A^{-1} of this matrix is applied to the primary, biased estimate of the covariance function C, finally yielding a refined estimate $\hat{C}$ of the covariance function

$$\hat{C} = A^{-1}C. \tag{13}$$

From Eq. 6, with A^{-1} multiplied on the left side, one obtains

$$A^{-1}\langle C\rangle = \gamma \tag{14}$$

and assuming A, which is defined by the sampling function w_i, being independent from C finally yields

$$\left\langle A^{-1}C\right\rangle = \langle \hat{C}\rangle = \gamma \tag{15}$$

meaning $\hat{C}$ is a bias-free estimate of γ. Note that $K < N - 1$ is required to ensure that A can be inverted, since at $K = N - 1$ the matrix is not full ranked. Furthermore, the derivation of the matrix A assumes that no covariances exist outside the investigated interval from $-K$ through K. If this is not given, the refined covariance estimate is not bias-free anymore; however, it will still be better than without this refinement. In [35] this is called a "nearly bias-free" covariance estimate. The requirement of having no further covariance outside the interval of lag times investigated coincides with the requirement for a bias-free estimate of the power spectral density derived hereof via the DFT.

3 Results and discussion

To test the performance of the proposed algorithm in comparison to established alternatives, a stochastic process is simulated generating data sequences with an artificial spectral composition from random white noise with an appropriate moving-average filter. The order of the filter can be chosen arbitrarily, in the present test an order of 25 has been used. The coefficients of the filter are chosen according to the desired spectrum, which also can be specified arbitrarily. To allow easy identification of various kinds of biases, which otherwise possibly would not appear obviously, in this test, the spectrum of the simulated process has an exponentially increasing slope, with a distinct dip in the observed frequency range. The parameters of the process and the values of the correlation function and the spectrum are provided as supplementary data with the present paper at [38]. The random generator provides independent signals of such spectral characteristics with a total length of 26tu (time units) to later obtain covariance estimates with lag times up to 25tu. This maximum lag time corresponds to the order of the moving-average filter, such that no correlations exist beyond the obtained covariance estimates and bias-free estimates of the covariance function with Bessel's correction and bias-free estimates of the power density spectra are possible. The signals have an expectation of $\mu = 4$au (amplitude units) and a standard deviation of 2au. These primary signals directly from the random generator have no interruptions.

To mimic the data gaps, for each sample of the primary signal appropriate weights are chosen by a second random process. To identify the performance of the covariance and spectral estimators under varying spectral characteristics of the data gaps, two different generators of the sampling functions have been realized. The first weighting routine generates individual samples marked as invalid, independently from each other, with a probability of 25%, resulting in a flat, white noise spectrum of the sampling function. The second weighting routine generates correlated series of invalid samples, where the length of invalid data sequences has an exponential distribution with a mean of 4tu. The probabilities of changes between valid and invalid samples are chosen such that this procedure also yields 25% invalid samples on average. In Fig. 3 individual realizations of the signals and the weights are shown. While in Fig. 3a only individual samples are marked as invalid independently from each other, in Fig. 3b longer sequences of invalid samples can be seen.

The generated signals then are analyzed by the proposed algorithm, and the results are compared to those of common alternatives in terms of the mean covariance and power spectral density and their respective estimation variances.

3.1 Weighted average

10,000 realizations of such signals have been simulated and analyzed for both weighting schemes. Wiener–Khinchin's theorem [43] is relating the covariance function C and the appropriate power spectral density S

$$S = \Delta t \cdot \mathrm{DFT}\{C\} \tag{16}$$

with the fundamental sampling step Δt of the signal and the discrete Fourier transform DFT. Both, C and S are two-sided functions, with their values arranged cyclic, modulo

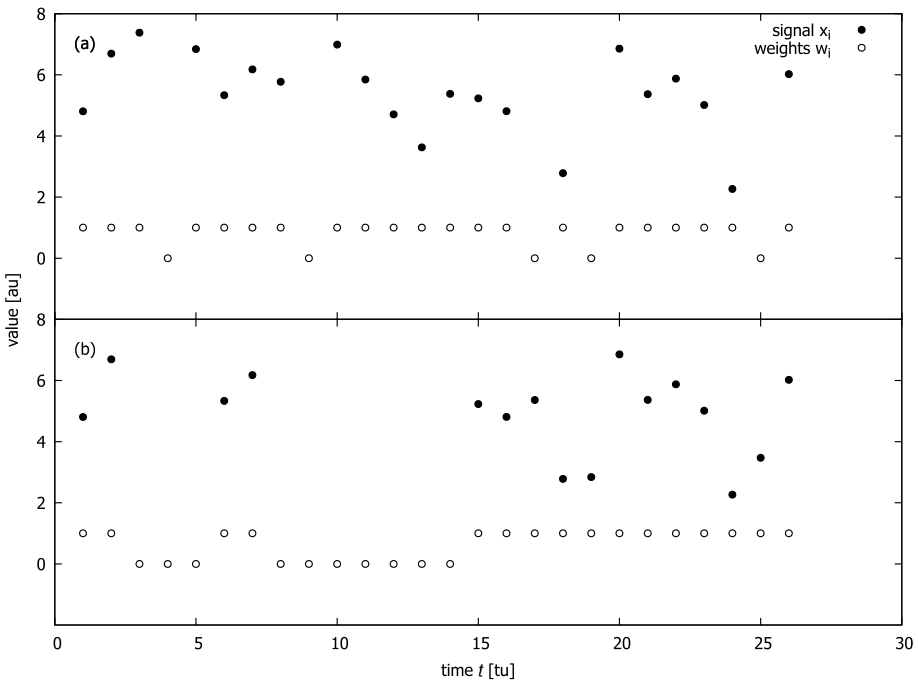

Fig. 3 Single realizations of the signal with **a** independent outliers and **b** with long gaps

their total length, which is in this first test $2N-1$, yielding a spectral increment of $\Delta f = \frac{1}{(2N-1)\Delta t}$. Provided that the covariance function is zero outside the determined rage of lag times, a bias-free covariance yields a bias-free power spectral density and vice versa, provided the spectrum is zero outside the range of frequencies determined. This way, both statistical functions can be used alternatively to represent possible deviations. Figure 4 shows the empirical mean of the power spectral density obtained from the 10,000 realizations of the covariance functions comparing the proposed weighted average with other widely used methods.

Sample-and-hold interpolation as a prominent example for interpolation attempts yields biased estimates for both cases, individual outliers as well as longer data gaps (Fig. 4a and b). Sample-and-hold interpolation with appropriate inversion of the dynamic error as in [7, 8] is constructed to correct the bias after interpolation for individual and independent outliers (Fig. 4a). Since the correction procedure is strictly bounded to the assumption of independent outliers, it fails in the case of correlated data gaps as in Fig. 4b. In contrast, the proposed weighted average yields bias-free estimates for both sampling schemes. Note that the known expectation μ has been subtracted from the data before estimating the covariance function, leading to bias-free estimates of the covariance function. Since no further correlations exist beyond the investigated range of lag times, the spectra obtained from the covariance estimates are also bias-free. This way, averages over multiple estimates of the covariance function or the power spectral density will converge toward the true functions of the underlying process. Additionally, results are shown for Lomb-Scargle's method [10, 11], which often is used as a reference algorithm for spectral estimation from irregularly sampled data. However, independent of the particular sampling scheme, a

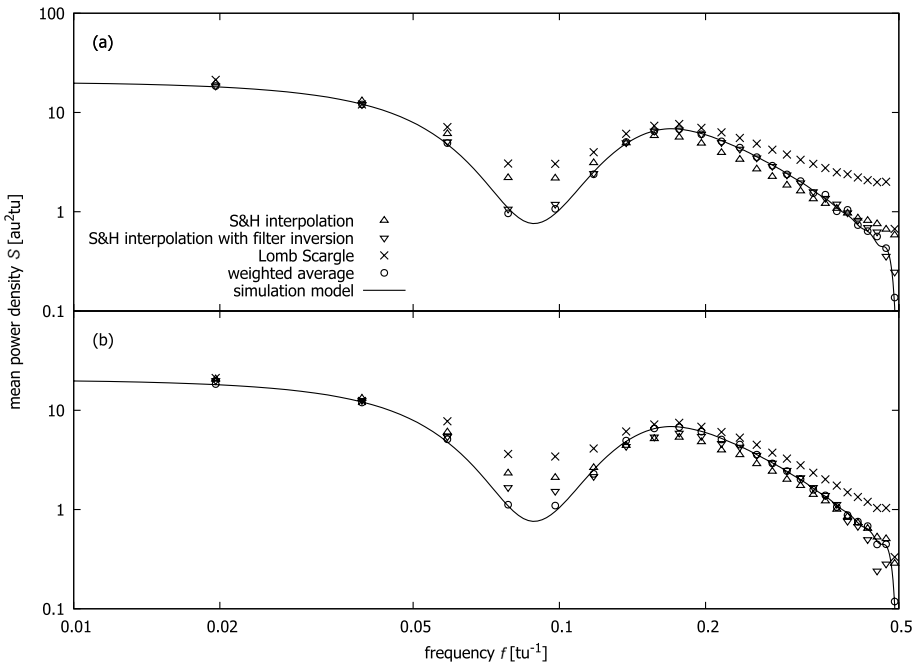

Fig. 4 Average of the power spectral density from 10,000 signals with **a** independent outliers and **b** with long gaps

distinct bias can be seen for that method, which is even larger than that for sample-and-hold interpolation.

The diagrams show the one-sided spectra in a logarithmic scale. This way neither the value at zero frequency can be seen nor negative values of the function. Even if negative spectral values physically make no sense, an unbiased estimate with some uncertainty may lead to values below zero. This corresponds to correlation matrices, which violate the nonnegative definiteness. Both, the discrimination of different data in weighted averaging as well as bias corrections (e.g., for sample-and-hold interpolation) potentially lead to such cases.

3.2 Restricting the domain of the covariance function

To demonstrate the performance of the restriction of the domain of the covariance function, the simulated signals have been extended in duration to 260tu each. From these signals, either 10 blocks of the initial signal duration of 26tu with no overlap or 19 blocks of the same duration with 50% overlap can be cut out. Then the covariance function is derived also from the entire signals and the domain is restricted to a maximum lag time of 25tu ($K = 25$), which is identical to the maximum lag time as achieved from the data blocks. The covariance function has been transformed into the power spectral density using again Wiener–Khinchin's theorem with the spectral increment of $\Delta f = \frac{1}{(2K-1)\Delta t}$ now. Again 10,000 realizations of the signal and the statistical analysis have been simulated. Figures 5 and 6 show the empirical mean and the empirical variance of the power spectral density from these 10,000 realizations for the two simulated weighting schemes. No systematic errors can be identified for the investigated estimators (Fig. 5), while the estimation variance for block averaging with overlap is lower than without overlap and

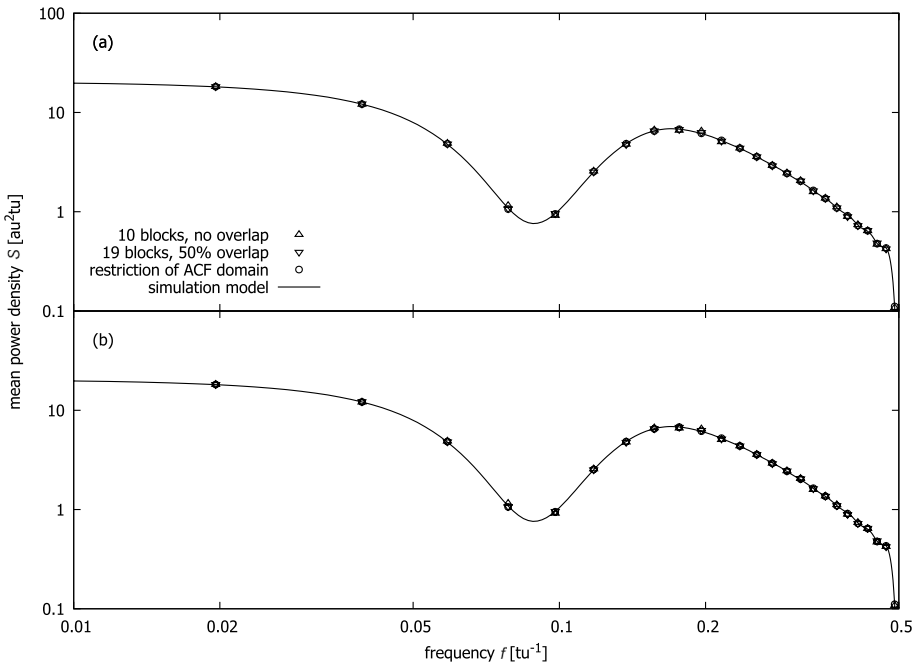

Fig. 5 Average of the power spectral density from 10,000 signals based on weighted average covariance estimation with **a** independent outliers and **b** with long gaps

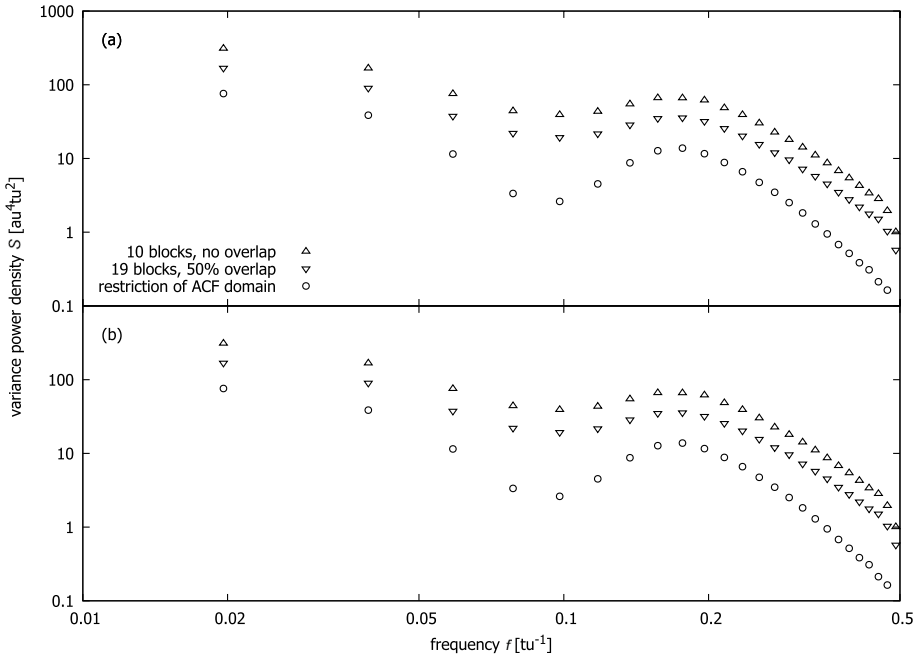

Fig. 6 Empirical estimation variance of the power spectral density from 10,000 signals based on weighted average covariance estimation with **a** independent outliers and **b** with long gaps

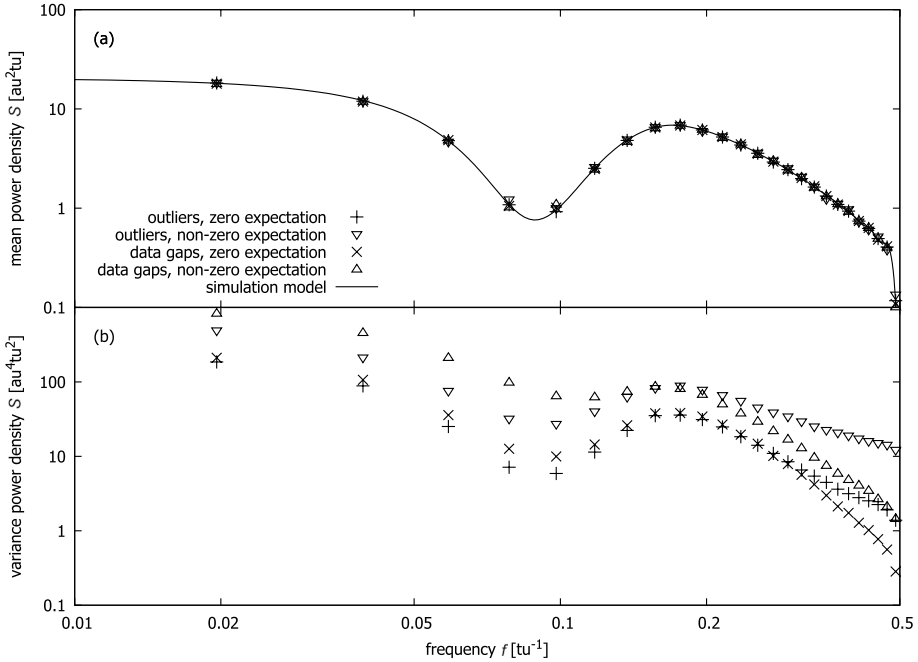

Fig. 7 **a** Empirical mean and **b** variance of the power spectral density from 10,000 signals based on weighted average covariance estimation

that of the method with domain restriction of the covariance function and without block subdivision is lowest (Fig. 6).

3.3 Mean removal and Bessel's correction

Figure 7 compares the results of the spectral estimation from data sets with zero and with nonzero expectation. The empirical mean of the spectra in Fig. 7a shows no difference between the various cases. A nonzero expectation μ of the data yields a covariance function with a constant offset of μ^2. It will affect the spectrum only at zero frequency, which is hidden in the logarithmic scaled one-sided spectrum shown in the graph. However, Fig. 7b shows that the estimation variance of the spectrum increases with a nonzero expectation. Therefore, a mean free signal would be preferred. Because normal applications lack the correct expectation μ, probably the empiric mean value $\bar{x}$ will be subtracted instead.

The expected bias due to the removal of the estimated mean value $\bar{x}$ instead of the correct but unknown expectation μ of the data mainly consists of an offset. Higher order terms exist, for symmetry reasons in autocovariance estimates, namely even orders. However, the constant offset is dominant. Therefore, the autocovariance estimates are better suited than the spectra to demonstrate this bias. Figure 8 shows the mean and the variance of the autocovariance estimates, obtained from 10,000 realizations of the signal for both weighting schemes. In this simulation, the signals had a duration of 100tu ($N = 100$) and the autocovariance estimates have been restricted to 25tu ($K = 25$), corresponding to the test cases above, to fulfill the requirement for the correction ($K < N - 1$) and to demonstrate the achievable effect.

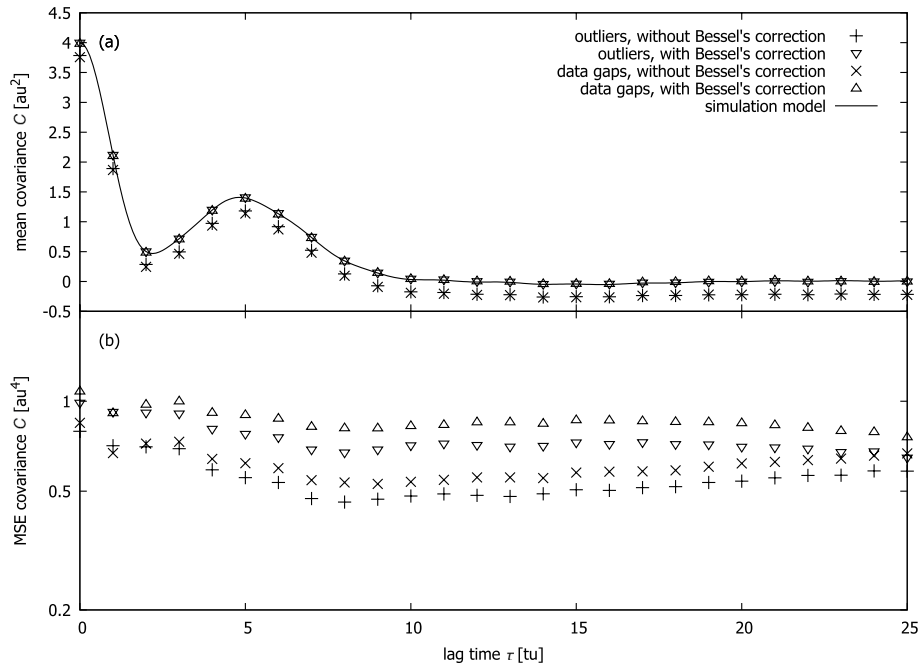

Fig. 8 **a** Empirical mean and **b** mean square error of the covariance estimates after removal of the empirical mean value from 10,000 signals using weighted averages

Figure 8a shows the empirical mean of the covariance estimates. The bias of the estimates without Bessel's correction is obvious. With the correction, the bias is effectively removed for both weighting schemes. However, the correction is achieved for the prize of an increased mean square error as shown in Fig. 8b, which depends on K chosen with respect to N and, it rapidly drops with smaller K. This perfectly coincides with the restriction of the domain of the covariance function above. Then, the mean square error still increases with Bessel's correction; however, it is acceptable if bias-free estimates are in focus.

4 Conclusion

Routines for nonparametric estimation of the covariance function and the power spectral densities from signals with invalid samples have been introduced for both cases, for independent individual outliers as well as for longer and hence correlated data gaps. The combination of ensemble averages over valid samples only, *a posteriori* restriction of the domain of the covariance function and Bessel's correction after estimation and removal of the estimated mean value, yields bias-free estimates of the statistical functions. This holds also for large amounts of data missing and it holds independent of the dynamic characteristics of the data gaps. While the first two parts of the whole procedure are clearly superior to alternative methods like interpolation or block subdivision, Bessel's correction yields bias-free estimates for the prize of an increased mean square error. Since the application of weighted average and restriction of the domain of the covariance function yields consistent estimates even with the removal of the estimated mean value, Bessel's correction as the final part of the introduced procedure is recommendable only, if strictly bias-free estimation is mandatory.

The implementation of the covariance and spectral estimations is done consistently based on FFT routines to achieve sufficient data throughput. All programs used here, the parameters of the simulated process and the values of the correlation function and the spectrum are available as supplementary material to the present paper at [38]. In this repository also programs for bias-free estimation of the cross-covariance and the cross-spectral density from two signals with missing data are available.

Abbreviations

au	Amplitude unit
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
tu	Time unit

Acknowledgements
Not applicable.

Author contributions
The research and the outcome of this specific publication are result of a long cooperation between the authors about fundamentals and applications of signal processing of unevenly sampled data. For the present manuscript, ND contributed to the definition of requirements and applications, VK contributed to the generalization of sampling cases, broadening of relevance due to new applications and writing, and HN contributed with methods, programming, simulations and writing.

Funding
Open Access funding enabled and organized by Projekt DEAL. This research had no specific funding. Open Access funding enabled and organized by Project DEAL.

Availability of data and materials
The procedures are available as Python source codes as supplementary material to this article together with all data sets under [38].

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 August 2023 Accepted: 8 January 2024
Published online: 25 January 2024

References

1. R. Vio, T. Strohmer, W. Wamsteker, On the reconstruction of irregularly sampled time series. Publ. Astron. Soc. Pac. **112**(767), 74–90 (2000). https://doi.org/10.1086/316495
2. M. Niedźwiecki, K. Cisowski, Adaptive scheme for elimination of background noise and impulsive disturbances from audio signals, in *Proceedings of the Quatrozieme Colloque GRETSI* (Juan-les-Pins, France), pp. 519–522 (1993)
3. S. Preihs, F.-R. Stöter, J. Ostermann, Low delay error concealment for audio signals, in *46th AES Conference on Audio Forensics* (2012)
4. P. Stoica, E.G. Larsson, J. Li, Adaptive filter-bank approach to restoration and spectral analysis of gapped data. Astron. J. **120**(4), 2163–2173 (2000). https://doi.org/10.1086/301572
5. R. Everson, L. Sirovich, Karhunen-Loève procedure for gappy data. J. Opt. Soc. Am. A **12**, 1657–1664 (1995). https://doi.org/10.1364/JOSAA.12.001657
6. D. Venturi, G.E. Karniadakis, Gappy data and reconstruction procedures for flow past a cylinder. J. Fluid Mech. **519**, 315–336 (2004). https://doi.org/10.1017/S0022112004001338
7. G. Plantier, S. Moreau, L. Simon, J.-C. Valière, A.L. Duff, H. Baillet, Nonparametric spectral analysis of wideband spectrum with missing data via sample-and-hold interpolation and deconvolution. Digit. Signal Process. **22**, 994–1004 (2012). https://doi.org/10.1016/j.dsp.2012.05.012
8. H. Nobach, Note on nonparametric spectral analysis of wideband spectrum with missing data via sample-and-hold interpolation and deconvolution. Digit. Signal Process. **87**, 19–20 (2019). https://doi.org/10.1016/j.dsp.2019.01.008
9. L. Sujbert, G. Orosz, FFT-based spectrum analysis in the case of data loss. IEEE Trans. Instrum. Meas. **65**, 968–976 (2016). https://doi.org/10.1109/TIM.2015.2508278
10. N.R. Lomb, Least-squares frequency analysis of unequally spaced data. Astrophys. Space Sci. **39**, 447–462 (1976). https://doi.org/10.1007/BF00648343
11. J.D. Scargle, Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. Astrophys. J. **263**, 835–853 (1982). https://doi.org/10.1086/160554
12. E. Masry, Spectral and probability density estimation from irregularly observed data, in *Statistical Analysis of Irregularly Observed Time Series in Statistics, Lecture Notes in Statistics*. ed. by E. Parzen (Springer, New York, 1983), pp.224–250. https://doi.org/10.1007/978-1-4684-9403-7_11
13. S. Ferraz-Mello, Estimation of periods from unequally spaced observations. Astron. J. **86**, 619–624 (1986). https://doi.org/10.1086/112924
14. G. Foster, Time series analysis by projection. I. Statistical properties of Fourier analysis. Astron. J. **111**, 541–554 (1996). https://doi.org/10.1086/117805
15. A. Mathias, F. Grond, R. Guardans, D. Seese, M. Canela, H.H. Diebner, Algorithms for spectral analysis of irregularly sampled time series. J. Stat. Softw. **11**, 1–27 (2004). https://doi.org/10.18637/jss.v011.i02
16. A. Rivoira, G.A. Fleury, A consistent nonparametric spectral estimator for randomly sampled signals. IEEE Trans. Signal Process. **52**, 2383–2395 (2004). https://doi.org/10.1109/TSP.2004.832002
17. P. Stoica, N. Sandgren, Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approach. Digit. Signal Process. **16**(6), 712–734 (2006). https://doi.org/10.1016/j.dsp.2006.08.012
18. P. Babu, P. Stoica, Spectral analysis of nonuniformly sampled data—a review. Digit. Signal Process. **20**, 359–378 (2010). https://doi.org/10.1016/j.dsp.2009.06.019
19. N. Damaschke, V. Kühn, H. Nobach, A fair review of non-parametric bias-free autocorrelation and spectral methods for randomly sampled data in laser Doppler velocimetry. Digit. Signal Process. **76**, 22–33 (2018). https://doi.org/10.1016/j.dsp.2018.01.018
20. R.H. Jones, Spectral analysis with regularly missed observations. Ann. Math. Stat. **33**(2), 455–461 (1962). https://doi.org/10.1214/aoms/1177704572
21. R.H. Jones, Spectral estimates and their distributions, part II. Scand. Actuar. J. **1962**(3–4), 135–153 (1962). https://doi.org/10.1080/03461238.1962.10405942
22. E. Parzen, On spectral analysis with missing observations and amplitude modulation. Sankhya Indian J. Stat. Ser. A **25**(4), 383–392 (1963)
23. P.A. Scheinok, Spectral analysis with randomly missed observations: The binomial case. Ann. Math. Stat. **36**(3), 972–977 (1965). https://doi.org/10.1214/aoms/1177700069
24. P. Bloomfield, Spectral analysis with randomly missing observations. J. R. Stat. Soc. Ser. B Stat. Methodol. **32**(3), 369–380 (1970)
25. R.H. Jones, Spectrum estimation with missing observations. Ann. Inst. Stat. Math. **23**(1), 387–398 (1971). https://doi.org/10.1007/BF02479238

26. R.H. Jones, Aliasing with unequally spaced observations. J. Appl. Meteorol. **11**(2), 245–254 (1972). https://doi.org/10.1175/1520-0450(1972)011<0245:AWUSO>2.0.CO;2
27. M.A. Ghazal, A. Elhassanein, Periodogram analysis with missing observations. J. Appl. Math. Comput. **22**(1–2), 209–222 (2006). https://doi.org/10.1007/BF02896472
28. C. Munteanu, C. Negrea, M. Echim, K. Mursula, Effect of data gaps: comparison of different spectral analysis methods. Ann. Geophys. **34**(4), 437–449 (2016). https://doi.org/10.5194/angeo-34-437-2016
29. P.M. Robinson, Estimation of a time series model from unequally spaced data. Stoch. Processes Their Appl. **6**(1), 9–24 (1977). https://doi.org/10.1016/0304-4149(77)90013-8
30. R.H. Jones, Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics **22**(3), 389–395 (1980). https://doi.org/10.2307/1268324
31. W. Dunsmuir, P.M. Robinson, Estimation of time series models in the presence of missing data. J. Am. Stat. Assoc. **76**(375), 560–568 (1981). https://doi.org/10.2307/2287513
32. W. Dunsmuir, P.M. Robinson, Parametric estimators for stationary time series with missing observations. Adv. Appl. Probab. **13**(1), 129–146 (1981). https://doi.org/10.2307/1426471
33. P.M. Robinson, Testing for serial correlation in regression with missing observations. J. R. Stat. Soc. Ser. B Stat. Methodol. **47**(3), 429–437 (1985)
34. M.S. Bartlett, Smoothing periodograms from time-series with continuous spectra. Nature **161**, 686–687 (1948). https://doi.org/10.1038/161686a0
35. T.J. Vogelsang, J. Yang, Exactly/nearly unbiased estimation of autocovariances of a univariate time series with unknown mean. J. Time Ser. Anal. **37**, 723–740 (2016). https://doi.org/10.1111/jtsa.12184
36. H. Nobach, Practical realization of Bessel's correction for a bias-free estimation of the auto-covariance and the cross-covariance functions (2023). https://doi.org/10.48550/arXiv.2303.11047. arXiv:2303.11047 [stat.ME]
37. N. Damaschke, V. Kühn, H. Nobach, Bias-free estimation of the auto- and cross-covariance and the corresponding power spectral densities from gappy data (2023). https://doi.org/10.48550/arXiv.2304.13997. arXiv:2304.13997 [eess.SP]
38. Supplementary Data Repository. Accessed 21 Aug 2023. http://www.nambis.de/publications/jaspgb.html
39. M.S. Bartlett, Periodogram analysis and continuous spectra. Biometrika **37**(1–2), 1–16 (1950). https://doi.org/10.1093/biomet/37.1-2.1
40. P.D. Welch, The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**(2), 70–73 (1967). https://doi.org/10.1109/TAU.1967.1161901
41. R.B. Blackman, J.W. Tukey, The measurement of power spectra from the point of view of communications engineering—part I. Bell Syst. Tech. J. **37**(1), 185–282 (1958). https://doi.org/10.1002/j.1538-7305.1958.tb03874.x
42. R.B. Blackman, J.W. Tukey, The measurement of power spectra from the point of view of communications engineering—part II. Bell Syst. Tech. J. **37**(2), 485–569 (1958). https://doi.org/10.1002/j.1538-7305.1958.tb01530.x
43. A. Khintchine, Korrelationstheorie der stationären stochastischen Prozesse. Math. Ann. **109**, 604–615 (1934). https://doi.org/10.1007/BF01449156

Publisher's Note