Estimating Time-Series Models From Irregularly Spaced Data

Piet M. T. Broersen and Robert Bos

***Abstract*—Maximum-likelihood estimation of the parameters of a continuous-time model for irregularly sampled data is very sensitive to initial conditions. Simulations may converge to a good solution if the true parameters are used as starting values for the nonlinear search of the minimum of the negative log likelihood. From realizable starting values, the convergence to a continuous-time model with an accurate spectrum is rare if more than three parameters have to be estimated. A discrete-time spectral estimator that applies a new algorithm for automatic equidistant missing-data analysis to irregularly spaced data is introduced. This requires equidistant resampling of the data. A slotted nearest neighbor (NN) resampling method replaces a true irregular observation time instant by the nearest equidistant resampling time point if and only if the distance to the true time is within half the slot width. It will be shown that this new resampling algorithm with the slotting principle has favorable properties over existing schemes such as NN resampling. A further improvement is obtained by using a slot width that is only a fraction of the resampling time.**

***Index Terms*—Continuous-time likelihood, nearest neighbor (NN) resampling, order selection, slotting, spectral estimation, unequally spaced, uneven sampling.**

I. Introduction

MANY estimation techniques for unevenly spaced data have been developed [1]. They can be divided in continuous-time and discrete-time spectral estimates. The maximum-likelihood (ML) estimator has been formulated for the estimation of continuous-time models from irregular data [2]. However, this estimator is known to be very sensitive to local minima and requires very good initial estimates [3].

Replacing the derivative operator in the continuous-time model by a discrete-time approximation is a method to identify continuous-time models from unevenly sampled data. Low-order autoregressive (AR) processes have been studied with this method [4]. All known variants inevitably suffer from bias.

For autoregressive moving average (ARMA) systems, the Cramér–Rao lower bound for the parameters has been derived [5], which depends on the actual irregular observations. This achievable accuracy computation uses a state space formulation of the ARMA model.

A new continuous ARMA method requires the explicit use of a sampling model for the irregular instants, for which the Poisson distribution is used [6]. The precise shape of that distribution is very important for the result, but it is almost impossible to establish it from practical data.

Several discrete-time methods have been described. The slotting technique estimates equidistant lags of the autocovariance function from irregularly sampled data. Many variants and improvements have been proposed [1]. No existing slotting method gives positive definite estimates for the autocorrelation function, which is necessary for a positive spectral density as Fourier transform. Sometimes reasonable results have been reported, but only if more than 100 000 observations are available.

Resampling techniques reconstruct a signal at equal time intervals. Those equidistant data can be analyzed with the usual signal processing methods, which can guarantee the positive definite property. However, spectral estimates at higher frequencies will be severely biased. Adrian and Yao [7] described the bias caused by sample-and-hold reconstruction as low-pass filtering of the signal, which is followed by adding noise. For Poisson-distributed measuring instants, these effects can in theory be eliminated using a refined sample-and-hold estimator [1]. Recent developments in refinement techniques are exclusively limited to Poisson distributions [8]. Undoing the bias is based on the asymptotic theory, neglecting the variance of the estimates. Therefore, very large data sets are required for this method. For practical data, it is not possible to reconstruct details with a magnitude below the theoretical bias level reliably. Nearest neighbor (NN) resampling and sample-and-hold have similar filtering and noise characteristics [9]. The resampled spectra are strongly biased for frequencies higher than about 15% or 20% of the mean data rate.

A new idea with time-series analysis can be perceived as searching for uninterrupted sequences of data that are almost equidistant [10] and using the Burg method for segments to estimate the spectrum with an AR model. In a similar approach, a slotted version of that Burg algorithm uses a modified NN resampling scheme to create an equidistantly resampled signal, with many empty places where no original observation fell inside the slot width. Slotting reduces the bias of NN resampling considerably. The reason is that slotting prevents an irregular observation to appear at multiple resampled time instants. A disadvantage of this method is that very large data sets are required to obtain some uninterrupted sequences of sufficient length for the Burg algorithm. It turned out that a nonlinear ML algorithm for missing-data problems, which have already been described by Jones [11], could sometimes give a better solution for slotted resampled data, also if much less data are available [12]. Whereas the slotted method required about 200 000 irregular observations, the ML method could

Manuscript received June 15, 2005; revised March 15, 2006.

P. M. T. Broersen is with the Department of Multi-Scale Physics, Delft University of Technology, Delft 2628 BW, The Netherlands (e-mail: broersen@tnw.tudelft.nl).

R. Bos is with Shell Netherlands, Den Haag 2501 CK, The Netherlands (e-mail: R.Bos@shell.com).

Digital Object Identifier 10.1109/TIM.2006.876389

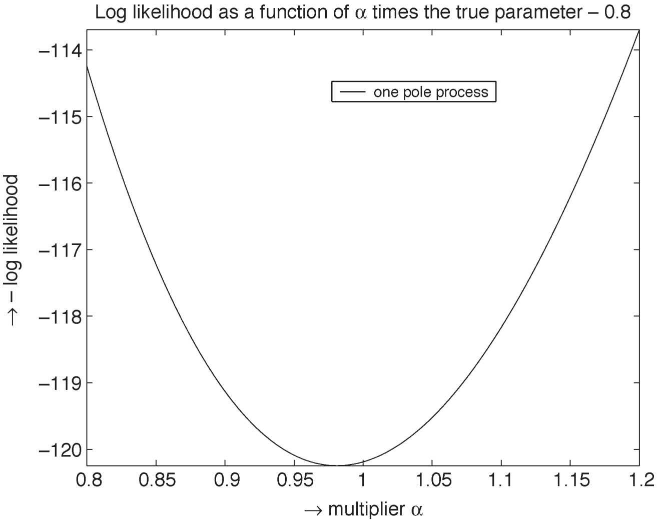


Fig. 1. Log likelihood of 1000 irregular observations of a continuous one-pole process as a function of α, where α is a multiplication factor for the true parameter.

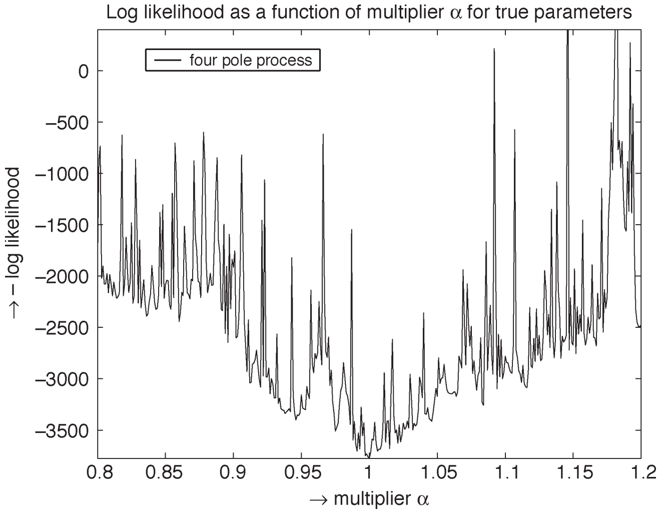


Fig. 2. Log likelihood of 1000 irregular observations of a continuous four-pole process as a function of α, where α is a simultaneous multiplication factor for the complete true parameter vector.

converge already to an accurate spectral estimate with 2000 or less observations.

Because of the many unsolved problems for continuous-time solutions and slotted autocovariance methods, this paper investigates the application of the best discrete-time method for equidistant missing-data problems [12] after that method is adapted to irregular data. A survey of existing missing-data methods and a robust version of the ML algorithm for autoregressive models of missing-data problems has been given [12], [13]. For missing-data problems, the performance of the robust and automatic ML algorithm outperforms all other methods.

The purpose of this paper is twofold. First, the ML principle for continuous-time models is investigated, but no robust algorithm has been found. Second, the continuous irregular problem is considered as an approximate equidistant missing-data problem. Modifications required to apply the existing automatic algorithm for equidistant missing data [12], [13] to irregularly sampled data are given. Irregular data are approximated by a number of shifted equidistant data sets. The choices of the grid time, the slot width, and the selection of the best discrete-time model for the irregular data are discussed.

II. Log Likelihood of AR Models

Experience shows that the estimation of first-order continuous AR models with a single negative real pole never creates a problem with continuous ML estimation. The surface of the likelihood as a function of the parameter is smooth, and any search method converges to the minimum whatever the start value might be. Fig. 1 gives an example for a real root at -0.8. It is not difficult to find the minimum from arbitrary initial values if the likelihood is completely regular.

Second-order simulations sometimes converged to the minimum of the likelihood. A simple sequential method starts with the AR(1) model. Then, the starting values for an AR(p) model are found by using the model estimated for order $p-1$ plus an additional negative initial value for the real pole of the new order p. In most two-pole simulations, the sequential starting values for orders 1 and 2 converged to the minimum of the likelihood. However, the sequential AR orders in a few runs failed to converge to a good spectral estimate with a peak. Using the true values of the parameters as starting values for the nonlinear search always found the model with a peak. This demonstrates that the ML estimate is very sensitive to the initial starting values for the nonlinear search [3]. A third-order example exists of a peak with a pair of complex damped conjugated poles and a negative pole, giving a constant slope. This three-pole example marks the transition from successful ML estimates and the impossibility to find initial values that are close enough to the global minimum to converge. The surface of the log likelihood was almost always rather rough with many local minima.

No runs of fourth-order processes with two complex conjugated pole pairs converged to the global minimum of the continuous log likelihood if sequential starting values were used. Only in exceptional cases, realizable starting values were close enough to the global minimum to obtain convergence to the minimum of the log likelihood with the nonlinear search. Sometimes, numerical problems even prevented convergence from the true process parameters as starting values. Matrices were close to singularity. Fig. 2 demonstrates a reason for the poor convergence. It is almost impossible to find the global minimum. Many local peaks and minima are seen. The number of peaks in Fig. 2 would grow with the number of likelihood evaluations, which is given by the density of the evaluation grid for α. Repeated simulation runs never delivered a four-pole realization, where the surface of the continuous log likelihood was smooth and without very large peaks.

Several ML scenarios have been attempted. Generally, the optimization will stop at the local minimum closest to the starting values. The shown appearance of the log-likelihood function is a combination of true likelihood properties, numerical singularities, and programming imperfections. No reports of successful ML algorithms for higher order models have been found in the literature. Only models with one or two poles can be determined reliably. It is clear that it will not be allowed to extrapolate a good ML behavior of an estimation method for one or two poles to arbitrary model orders.

III. Multishift Slotted NN Resampling

Discrete-time signal processing methods can be applied to equidistantly resampled irregular data with sample-and-hold [7] or NN resampling [9]. That gives a simple equidistant signal that has an unacceptable bias for frequencies above 0.15 or 0.2 times the mean data rate. The analysis of resampling methods shows that the bias is caused by the multiple use of the same observation and the shift of irregular observation times to a fixed grid. The multiple use of a single observation for more resampled data points is a serious problem. This creates a bias term in the estimated autocovariance function because the autocovariance $R(0)$ leaks to autocovariances estimated at nonzero lags. Multiple use of the same observation is eliminated in slotted NN resampling. That will produce a resampled signal with many empty places, which can be processed with a missing-data algorithm.

Assume that a signal $x(t)$ is measured at N irregular time instants $t_1, \ldots, t_N$. The average distance between samples T_0 is given by $T_0 = (t_N - t_1)/(N-1) = 1/f_0$, where f_0 denotes the mean data rate. The signal can be resampled on a grid at KN equidistant time instants at a grid distance of $T_r = T_0/K$. The resampled signal exists only for $t = nT_r$ with n as an integral number. The spectrum can be calculated up to the frequency $Kf_0/2$. The usual NN resampling substitutes at all grid points nT_r the closest irregular observation $x(t_i)$, with

$$|t_{i-1} - nT_r| > |t_i - nT_r| \quad |t_{i+1} - nT_r| > |t_i - nT_r|. \tag{1}$$

The uninterrupted resampled signal contains KN equidistant observations. For $K \gg 1$, that means that many of the original N irregular observations will be used for more resampled observations.

Slotted NN resampling only accepts a resampled observation at $t = nT_r$ if there is an irregular observation $x(t_i)$ with t_i within the time slot w, which can be expressed as

$$nT_r - 0.5w < t_i \leq nT_r + 0.5w. \tag{2}$$

If there is more than one irregular observation within a slot, the one closest to nT_r is selected for resampling; if there is no observation within the slot, the resampled signal at nT_r is left empty. For small T_r and for $w = T_r$, the number N_0 of nonempty resampled points nT_r becomes close to N because almost every irregular t_i falls into another time slot. For larger T_r with $K < 1$, more irregular observations may fall within one slot, and only the one closest to the grid point survives in the slotted NN resampled signal. The successive resampling times nT_r cover the whole continuous time axis for $w = T_r$.

Taking $w = T_r/M$ with M as an integer, gives disjunct intervals, where some irregular times t_i are not within any slot of (2). Therefore, multishift slotted NN resampling is introduced, where M different equidistant missing-data signals are extracted from one irregular data set. The sampling instants with nonempty places for the M signals are given as

$$nT_r + mw - 0.5w < t_i \leq nT_r + mw + 0.5w \qquad m = 0, 1, \ldots, M-1. \tag{3}$$

Now, all slots of width w are connected in time. The number of possible grid points is NMT_0/T_r. Hence, the fraction γ of points with an observation present is approximately given by $1/MK$. Experience with missing-data problems shows that time-series models can be easily estimated for $\gamma > 0.1$ [12], [13]. It may become difficult if γ is less than 0.01, unless the number of observations is very large. This limits the useful range of resampling time and slot width for a given number of observations.

The bias of multishift slotted NN resampling is strongly reduced in comparison with the usual NN. For Poisson sampling instants, its bias can be described with the probability density function $f(\tau)$ of the continuous-time lags τ that contribute to the resampled autocorrelation $R_{\text{res}}(nT_r)$. $f(\tau)$ is given as

$$\begin{aligned} f(\tau) =& 0.5f_0 \left\{ e^{-2f_0\tau} - e^{2f_0(\tau - w)} \right\} \\ &+ f_0^2 \tau e^{-2f_0\tau}, \qquad 0 < |\tau| < w/2 \\ =& (w-\tau) f_0^2 e^{-2f_0\tau}, \qquad w/2 \leq |\tau| \leq w \\ =& 0, \qquad |\tau| > w. \end{aligned} \tag{4}$$

With this result, the expectation of the resampled autocorrelation becomes

$$\begin{aligned} R_{\text{res}}(nT_r) &= \int_{-w}^{w} R(nT_r + \tau) f(\tau) d\tau, \qquad n \neq 0 \\ R_{\text{res}}(0) &= R(0). \end{aligned} \tag{5}$$

This type of bias due to the shift of irregular observation times to a resampled equidistant grid will be present in all equidistant evaluation methods for irregular data. This includes all slotted autocorrelation methods with fuzzy slotting or local normalization, as defined in [1]. The bias will be of the same order of magnitude as that obtained with (5) for $w = T_r$.

Fig. 3 shows the bias effects on the spectral density; the bias results from the variation of the autocorrelation function over the slot width. The example has a constant spectrum for frequencies f below $0.01f_0$, a constant slope in the double logarithmic presentation that descends at a rate of $\sim f^{-5/3}$ from $0.01f_0$, and an extra declining slope at a rate of $\sim f^{-7}$ for frequencies above $0.1f_0$. This type of spectra occurs in turbulent flow. The first figure shows the whole frequency range for a very low resampling rate. The resampling time $T_r = 2/f_0$ permits to compute spectra only up to $f_0/4$. The other figures are limited to the higher part of the frequency range to increase the visibility of the bias. The total frequency range increases inversely proportional to T_r. The bias becomes important in weak parts of the spectrum and becomes less if the slot width is reduced. For $w = 1$, the fraction of the total frequency range with an acceptable bias diminishes for smaller values of T_r. If

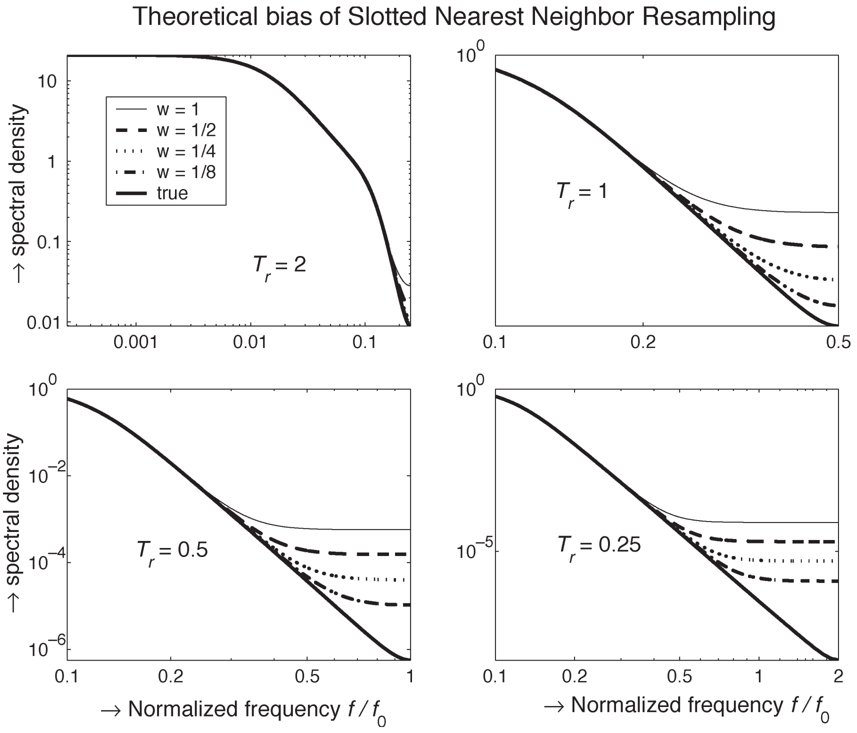


Fig. 3. Theoretical expectation of the spectral bias of slotted NN resampling for a turbulence spectrum for four resampling times T_r and four slot widths w, which is always given as a fraction of T_r for $f_0 = 1$.

the slot width is taken small enough, the bias will disappear eventually. That requires a very small slot if the dynamic range of the true spectrum is large, which is similar to that in Fig. 3 for $T_r = 0.25$. However, the small slot width also reduces the remaining data fraction γ because more resampling instants are used for the same number of irregular observations. A smaller slot reduces the bias of the spectral estimate, but it gives an increased variance because the remaining fraction γ becomes smaller and the estimation of parameters is more difficult.

As all resampling schemes give biased estimates, only the estimation of a continuous-time model can be unbiased for irregular data and can possibly approach the Cramér–Rao lower boundary for the achievable accuracy. Discrete-time spectra are intrinsically defined over a limited frequency range and can at best represent the data within that range.

IV. Equidistant Time-Series Models

Three different linear types of time-series models can be distinguished for equidistant observations x_n of a stationary stochastic process, namely 1) auto regressive or AR; 2) moving average or MA; and 3) combined ARMA models. An ARMA(p,q) model can be written as [14], [15]

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q} \tag{6}$$

where ε_n is a purely random process of independent identically distributed stochastic variables with zero mean and variance σ_ε^2. It is purely AR for $q = 0$ and MA for $p = 0$.

The estimated time-series model is a parametric estimator for the spectrum and the autocorrelation function. The power spectral density $h(\omega)$ of an ARMA(p,q) model is completely determined by the parameters in (6) together with the variance σ_ε^2 and is given as

$$h(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|B_q(e^{j\omega k})\right|^2}{\left|A_p(e^{j\omega k})\right|^2} = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|1 + \sum\limits_{k=1}^{q} b_k e^{-j\omega k}\right|^2}{\left|1 + \sum\limits_{k=1}^{p} a_k e^{-j\omega k}\right|^2}. \tag{7}$$

The autocovariance of x_n can be computed as the inverse Fourier integral transform of (7). It can easily be found directly with the standard theory [14], [15], showing that the parameters of a time-series model completely describe the power spectral density and the autocorrelation function of the data x_n.

Accuracy measures have been defined that can compare the quality of the discrete-time model with true or aliased continuous spectra [9]. With the spectral distortion SD, the integral of the squared difference between the logarithms of spectra can be computed for an arbitrary frequency range, which can be expressed as

$$SD = \frac{NT_r}{2\pi} \int\limits_{-\pi/T_r}^{\pi/T_r} \left[\ln\{h(\omega)\} - \ln\left\{\hat{h}(\omega)\right\}\right]^2 d\omega. \tag{8}$$

The hat indicates a spectral estimate. By limiting the integration to the definition area of the discrete spectrum, it is possible to attribute a single number to the accuracy of discrete-time approximations to continuous-time spectra [9].

An automatic ML program ARMA selection-missing (ARMAsel-mis) has been developed for the equidistant missing-data problem [12]. In simulations, the accuracy of the

ARMAsel-mis spectra in estimation when data are missing was better than the spectra obtained with many other methods from the literature. Examples have been given where the estimation of time-series models in missing-data problems was efficient, which means that the accuracy of the resulting model approached the limit of the achievable accuracy.

V. ARMA Selection-Irregular (ARMAsel-Irreg) Algorithm

Input for the algorithm for irregular data are the M equidistant missing-data sequences or signals obtained with the multishift slotted NN algorithm of (3). The signals are all derived from the original irregularly sampled observations in the same time interval. In principle, the data in the different signals are correlated and not independent. However, the most influential parts of each signal are found at places where only few data are missing. Generally, those places will be at different locations for the various signals, and the assumption that the signals are more or less independent is justified. An approximation for the likelihood is defined by computing it separately for each signal and adding them in the minimization procedure. Computing over more signals in the same time interval gives only a true likelihood if those signals are independent and uncorrelated. Otherwise, not all contributions to the true likelihood are taken into account because nearby observations may belong to different multishift resampled signals. The likelihood of the shifted signals is only equal to the sum of the likelihood of each of the individual signals if all shifted signals were independent. However, using the almost independent M signals together, each with about N/M observations, gives a much better accuracy than using only one of the resampled signals.

All elements for an automatic ARMAsel algorithm for irregular data can be copied from the algorithm that has been developed and described for missing data [12], [13], and are given in the following list.

- Apply multishift slotted NN resampling of (3) to replace the irregularly sampled signal into a number of equidistant missing-data signals that can be used in a dedicated missing-data algorithm that accepts the shifted signals with a slot width smaller than the resampling distance.
- Discrete-time "likelihood" for AR models is computed for every signal separately and added afterward. The exact method is used for $\gamma > 0.15$; else, an approximation is faster [12].
- Tangent of $\pi/2$ times the AR reflection coefficients is used in the minimization to guarantee estimated reflection coefficients with absolute values less than 1, which is a prerequisite for stationary AR models.
- Starting values for the AR$(p+1)$ model are the estimated reflection coefficients of the AR(p) model with an additional zero for the reflection coefficient of order $p+1$.
- AR(p) order selection uses the following as criterion:

$$GIC(p) = \text{the "log likelihood"} + \alpha p$$

  with $\alpha = 3$ for less than 25% missing, $\alpha = 5$ for less than 25% remaining, and $\alpha = 4$ otherwise.
- MA and ARMA models are estimated from the parameters of an intermediate AR model.
- Order of that intermediate AR model is chosen as the highest AR order with a spectrum close to the that of the selected AR model.
- Order selection for MA and ARMA models is based on the log likelihood plus three times the number of estimated parameters.
- Quantity γN can be considered as an effective number of observations. The fraction γ is determined by the choice of the resampling period and the slot width.

VI. Continuous ML and Discrete ARMAsel

The four-pole continuous AR process with two spectral peaks analyzed in Fig. 2 has been used as an example. The fitting of a continuous AR model with the ML method is very sensitive to the starting values. It has been demonstrated in Fig. 2 that the log-likelihood function is not well behaved and has many local minima [3]. The nonlinear ML optimization of the log likelihood uses the specific parameterization of Jones [2]. This is to make sure that the real parts of the roots of the estimated continuous model are negative and that the solution will always be stable. To verify that a good ML solution exists and that it has a low value of log likelihood, the true parameters have been used as starting values for the nonlinear minimization of the log likelihood. This optimization mostly converged to an ML spectral estimate that is very close to the true spectrum. However, those starting values are only possible in simulations where the truth is already known. It just shows that the true continuous ML solution with the global minimum of the log likelihood will be very attractive if good starting values can be found in practice.

In ML estimation in discrete-time missing-data problems, the model estimated for order $p-1$ with one additional zero has been used as the starting value for order p. Trying variants of this estimation with increasing orders for the continuous models were not successful if more than two parameters had to be estimated. They use the estimate of order $p-1$ as starting values for the solution at order p, with an additional pole at -0.2 or any other value. In addition, finite poles close to $-\infty$ have been tried for the additional pole. One of the problems is that the stability of the model prohibits the use of the value zero for the additional pole; some nonzero negative real pole must be used for the additional pole of order p. Generally, the likelihood of the starting model of order p becomes worse than the AR$(p-1)$ likelihood if that extra finite pole is added. That cannot happen in discrete-time modeling because the additional parameter with the value 0 for order p gives the same value of the likelihood. Moreover, the sequential optimization of the continuous log likelihood did not always converge to lower values for a model with more parameters. In no single simulation run with more than three parameters was a useful spectrum found with those sequential starting values for increasing model orders. Finding good starting values in practical situations is necessary. Fig. 2 shows that only a very small region around the true parameters will be good enough to obtain convergence to the global minimum.

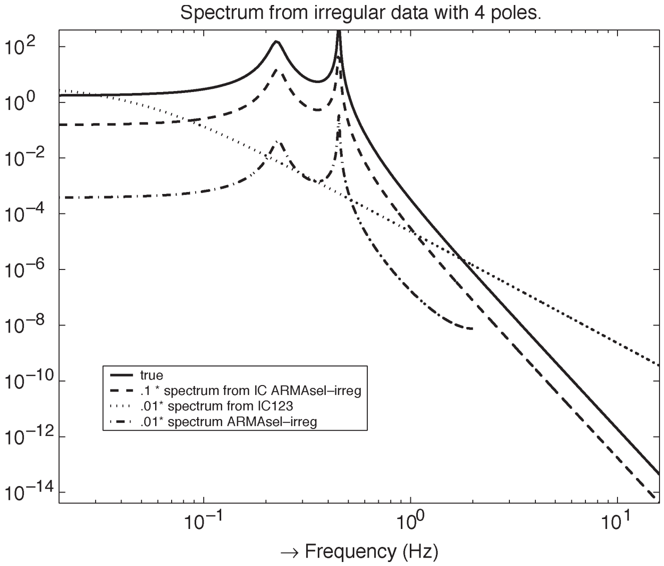


Fig. 4. True spectrum, ML spectrum found with realizable starting values derived from the ARMAsel-irreg solution and with sequential starting values denoted IC123, and the slotted NN estimate of ARMAsel-irreg. The number of irregular data was 1000, and the mean data rate was $f_0 = 1$. For ARMAsel-irreg, $T_r = 1/4$, and $w = 1/4$.

The ARMAsel-irreg algorithm gives a satisfactory estimate in the specific simulation run shown in Fig. 4, with two peaks at almost the true frequencies. That algorithm requires no user interaction, except for the choice of the resampling frequency T_r and the slot width w. The frequency range of the discrete-time solution is limited to below 2 Hz because the method uses an equidistant resampling scheme with $T_r = 1/4$. Therefore, the highest frequency range suffers from the bias that follows with (5). This is clearly seen in the spectrum of ARMAsel-irreg shown in Fig. 4, where the estimate becomes flat at the frequencies near 2.

The discrete-time model obtained with ARMAsel-irreg has been used to construct starting values for the continuous-time model. The zero-pole matching technique gives a transformation of a given discrete-time AR process to the poles of a continuous differential equation with only left-hand side terms [16]. The use of the bilinear transformation that is generally preferred would also introduce zeros in the continuous-time starting model. This would require a different likelihood algorithm. The continuous poles are then transformed to the starting values for the algorithm of Jones [2] for a nonlinear minimization of the log likelihood. Fig. 4 shows that the estimate of the continuous-time spectrum starting from ARMAsel-irreg was very good in this simulation run. The likelihood of the final model found with the optimization was about the same as the likelihood found from the true parameters as starting values. However, the improvement was not caused by the continuous ML optimization. Just transforming the estimated ARMAsel-irreg model to the continuous-time equivalent with the zero-pole matching did the job in this simulation run. The ML search program decided that it could not improve the likelihood further from those starting values. Only if the starting values had two peaks at approximately the correct frequencies would the continuous-time equivalent be as good as that in Fig. 4. It was remarkable that if an initial estimate has a peak at a wrong frequency, using that as the starting position resulted in a continuous ML solution with the peak at the same wrong frequency. The explanation is that the surface of the log likelihood is so rough that optimization will stop at a local minimum that is very nearby the starting point.

Looking at the surface of the discrete-time log likelihood in this example showed that this has a smooth surface. It has been verified by repeated simulations that the continuous log likelihood of the example with two peaks was always very rough, which is similar to that in Fig. 2, and the discrete-time log likelihood obtained with ARMAsel-irreg was smooth. If enough data are available, discrete-time models may have the bias of Fig. 3, but they do not have convergence problems because there are no peaks in the log likelihood.

The accuracy the ARMAsel-irreg spectrum is very good in comparison with many methods that use slotted estimates of autocorrelation functions to estimate spectra. Those methods mostly require 100 000 or more observations [1] and never deliver acceptable estimates with only 1000 observations.

VII. Simulations With Slotting

Whereas sample-and-hold or NN resampling without slotting always causes a filtering operation and additive noise in the frequency domain, this effect may disappear by using the slotting variant. As an example, the expectation of a white-noise spectrum remains white and unchanged after slotted NN resampling. This is clear with (5), where $R(0)$ is unchanged and all other contributions remain zero. Applying ordinary NN gives a colored-noise spectrum because the repeated resampled observations produce correlations at nonzero lags, which cause a colored spectrum.

The first simulation example has a constant slope in the double logarithmic presentation that descends at a rate of $\sim f^{-5/3}$ from $0.01 f_0$ and becomes flatter for very high frequencies above f_0 due to aliasing. This shape is inspired by turbulence data [17]. This true spectrum can be approximated very well with low-order AR models in the frequency range of Fig. 5. The theoretical bias is negligible, and the estimated and automatically selected AR(2) model of ARMAsel-irreg fits closely. The ARMAsel algorithm selected the AR(1) model for the equidistant data obtained with ordinary NN resampling. This estimated NN shows a bias that would also be present in the higher order AR models estimated from the NN signal. Like in white noise, this filtering effect of NN gives a downward bias. In this example, the performance of ARMAsel-irreg is very good.

The second example in Fig. 6 has an extra declining slope at a rate of $\sim f^{-7}$ for frequencies f above $0.1 f_0$. The bias of this example has been analyzed in Fig. 3. Fig. 6 gives the true continuous-time spectrum and two estimated discrete-time spectra. Here, the ARMAsel algorithm selected the AR(21) model for the equidistant data obtained with NN resampling. The ARMAsel-irreg algorithm selected the AR(5) model for the slotted NN data. Both estimated spectra are rather close to their respective biased expectations. Hence, slotting in combination with NN resampling gives a much lower bias, as might

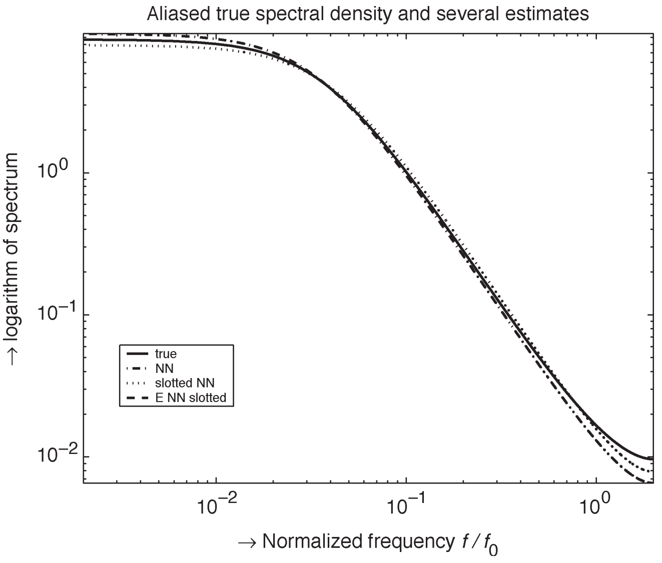


Fig. 5. True spectrum, the NN ARMAsel estimate, and the ARMAsel-irreg slotted NN estimate from 1000 irregular observations, with $T_r = T_0/4$ and $w = T_r/2$. The true spectrum and the slotted NN expectation coincide within the linewidth for this example.

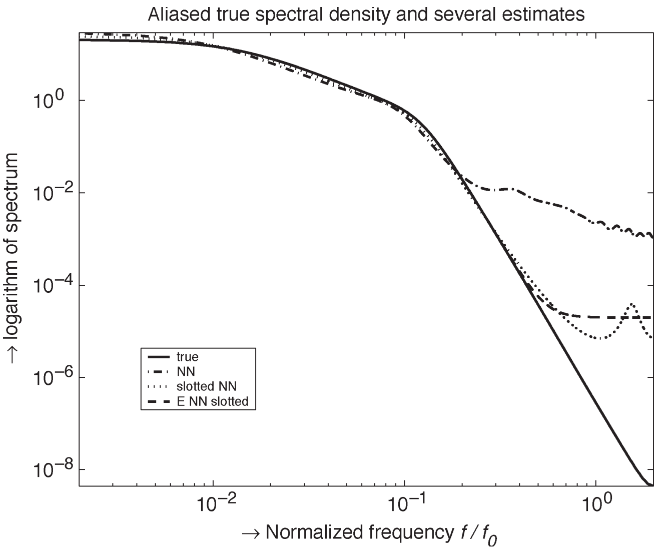


Fig. 6. True spectrum, the NN ARMAsel estimate, and the ARMAsel-irreg slotted NN estimate from 1000 irregular observations, with $T_r = T_0/4$ and $w = T_r/2$. The bias of the estimated spectra is close to the theoretical bias for NN with ARMAsel and to the slotted NN with ARMAsel-irreg.

be expected. Both NN and slotted NN are very accurate for $f < 0.2f_0$. In the range of $0.2f_0 < f < 0.5f_0$, the slotted estimate is still accurate, and the ordinary NN is biased. For still higher frequencies above $0.5f_0$, both estimates are no longer accurate. However, the bias reduction in the high-frequency range due to slotting is still more than a factor 100 better than that without slotting in this example with a steep spectral slope. Nevertheless, the slotted NN estimate also still has a considerable bias for high frequencies in Fig. 6. The improvement obtained with slotting is only negligible for frequencies below about $0.2f_0$, as has been predicted with theory [7].

VIII. Conclusion

A new estimator ARMAsel-irreg is introduced that fits a time-series model to multishift slotted NN resampled signals obtained from irregularly sampled data. The new ARMAsel-irreg algorithm combines a spectrum that is guaranteed to be positive with an improved accuracy at higher frequencies. The results in simulations with few data are much better than those that can be obtained from the same data but with other resampling techniques. The order and type of the best time-series model for the data can, in principle, be selected without user interaction. However, order selection may still fail in practice because of the likelihood properties.

Multishift slotted NN resampling with ARMAsel-irreg can give accurate discrete-time spectra if the dynamic spectral range of the signal is limited. For a large dynamic range, a small slot width can reduce the bias. That requires large data sets to obtain accurate estimates.

Continuous ML estimation may give good results for very low order AR models but only if the model order is known in advance and if the starting values of the nonlinear search are close enough to the true process parameters. Using the selected discrete-time model of ARMAsel-irreg as a starting value for the continuous likelihood minimization sometimes gives a good continuous-time model. However, ARMAsel-irreg is preferred for practical data.

References

[1] L. H. Benedict, H. Nobach, and C. Tropea, "Estimation of turbulent velocity spectra from laser Doppler data," *Meas. Sci. Technol.*, vol. 11, no. 8, pp. 1089–1104, Aug. 2000.

[2] R. H. Jones, "Fitting a continuous time autoregression to discrete data," in *Applied Time Series Analysis II*, D. F. Findley, Ed. New York: Academic, 1981, pp. 651–682.

[3] ——, "Fitting multivariate models to unequally spaced data," in *Time Series Analysis of Irregularly Spaced Data*, E. Parzen, Ed. New York: Springer-Verlag, 1983, pp. 158–188.

[4] E. K. Larsson and T. Söderström, "Identification of continuous-time AR processes from unevenly sampled data," *Automatica*, vol. 38, no. 4, pp. 709–718, Apr. 2002.

[5] E. K. Larsson and E. G. Larsson, "The CRB for parameter estimation in irregularly sampled continuous-time ARMA systems," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 197–200, Feb. 2002.

[6] E. Lahalle, G. Fleury, and A. Rivoira, "Continuous ARMA spectral estimation from irregularly sampled observations," in *Proc. IEEE/IMTC Conf.*, Como, Italy, 2004, pp. 923–927.

[7] R. J. Adrian and C. S. Yao, "Power spectra of fluid velocities measured by laser Doppler velocimetry," *Exp. Fluids*, vol. 5, no. 1, pp. 17–28, Jan. 1987.

[8] L. Simon and J. Fitzpatrick, "An improved sample-and-hold reconstruction procedure for estimation of power spectra from LDA data," *Exp. Fluids*, vol. 37, no. 2, pp. 272–280, 2004.

[9] S. de Waele and P. M. T. Broersen, "Error measures for resampled irregular data," *IEEE Trans. Instrum. Meas.*, vol. 49, no. 2, pp. 216–222, Apr. 2000.

[10] R. Bos, S. de Waele, and P. M. T. Broersen, "Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1289–1294, Dec. 2002.

[11] R. H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, vol. 22, no. 3, pp. 389–395, 1980.

[12] P. M. T. Broersen, S. de Waele, and R. Bos, "Autoregressive spectral analysis when observations are missing," *Automatica*, vol. 40, no. 9, pp. 1495–1504, 2004.

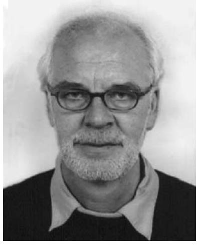
[13] ——, "Application of autoregressive spectral analysis to missing data problems," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 4, pp. 981–986, Aug. 2004.

[14] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1981.

[15] P. M. T. Broersen, *Automatic Autocorrelation and Spectral Analysis*. London, U.K.: Springer, 2006.

[16] G. F. Franklin, J. D. Powell, and M. Workman, *Digital Control of Dynamic Systems*. Menlo Park, CA: Addison-Wesley, 1998.

[17] W. K. Harteveld, R. F. Mudde, and H. E. A. van den Akker, "Estimation of turbulence power spectra for bubbly flows from laser Doppler anemometry signals," *Chem. Eng. Sci.*, vol. 60, pp. 6160–6168, 2005.



Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1968 and 1976, respectively,

He is currently with the Department of Multi-Scale Physics, Delft University of Technology. He developed a practical solution for the spectral and the autocorrelation analysis of stochastic data by the automatic selection of a suitable order and type for a time-series model of the data. His research interest is automatic identification on statistical grounds by letting measured data speak for themselves.

Robert Bos was born in Papendrecht, The Netherlands, in 1977. He received the M.Sc. degree in applied physics from Delft University of Technology, Delft, The Netherlands, in 2001. He is currently working toward the Ph.D. degree in the Delft Center for Systems and Control, Delft University of Technology.

He is currently a Reservoir Engineer with Shell Netherlands, Den Haag, The Netherlands. His research interests include monitoring using large-scale first-principles models.