Direct nonparametric estimation of the period and the shape of a periodic component in short duration signals

Ali Mohammad-Djafari
Laboratoire des signaux et systèmes,
CNRS-CentraleSupélec-Univ Paris Saclay, 91192 Gif-sur-Yvette, France
Email: djafari@lss.supelec.fr

***Abstract*—In this paper a new direct nonparametric estimation of the period and the shape of a periodic component in short duration signals is proposed and evaluated. Classical Fourier Transform (FT) methods lack precision and resolution when the duration of the signal is very short and the signal is noisy. The proposed method is based on the direct description of the problem as a linear inverse problem and a Bayesian inference approach with appropriate prior distributions. The expression of the joint posterior law of the period and the shape of the periodic component is obtained and used to determine both the period and the shape of the periodic component. Some results on synthetic data show the performance of the proposed method compared to the state of the art methods.**

keywords: Periodic signals; Period estimation; Short duration signals; Bayesian inference; Approximate Bayesian Computation (ABC)

I. Introduction

Detecting a periodic component in a short duration signal and estimating its period and shape is an important problem in many signal processing applications. As an example to mention is the biological applications, involving such time series as body temperature, activity level (circadian cycle) and many genes expression (cell cycle) [1]. Classically, Fourier Transform (FT) methods are used for this task, but when the duration of the signal is very short, for example only a few periods, these methods lack precision and resolution. Many parametric methods such as Fourier series decompositions can obtain better resolution [2], [3], [4], [5], but the precision of the period and the shape of the component is not often good enough when the shape of the periodic component is far from a sinusoidal form [6], [7], [8], [9]

In this work, a new method is presented to determine the period very precisely and estimate the nonparametric shape of that periodic component. The method is based on the description of the problem directly as a linear inverse problem and the Bayesian inference. A description of the forward problem formulation and a constrained Least Square method has been reported in [8], [10], a regularization based method in [11] and a *Maximum a posteriori* (MAP) estimation method in [12], [13], [14]. The main advantage of the Bayesian framework is the possibility of accounting explicitly for the errors and uncertainties such as the measurement noise and modelling uncertainties and can quantify the remaining uncertainties of the proposed estimates. However the computational costs of the exact computation of the solutions in these methods are very high.

II. Proposed method

A direct forward nonparametric model relating the observed data to the unknowns of the problem is illustrated in Figure 1 where a periodic component signal which is repeated a few times. Note that there is no need to have observed an exact multiple of the period. As we can see in this Figure, we have three complete periods and an extra fraction of it.

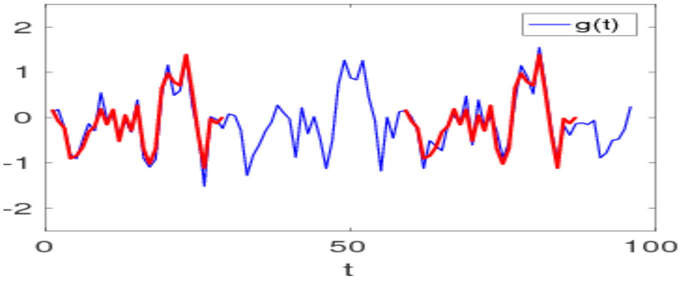


Fig. 1. A periodic signal and its periodic component. In this example 3 full periods and a fraction of it is observed.

The forward model is directly obtained by considering a length M time series $g(t), t = 0, \cdots, M-1$ in which we seek a length N periodic shape $f(t), t = 0, \cdots, N-1$. Noting by the vector $\boldsymbol{g}$ the M samples of $g(t)$ and by the vector $\boldsymbol{f}$ the N samples of the unknown periodic component, the relation between these two vectors becomes:

$$\begin{aligned}\boldsymbol{g} &= [g_1, \cdots, g_N, g_{N+1}, \cdots, g_{N+K}, \cdots, g_{KN+1}, \cdots, g_M]' \\ &= [f_1, \cdots, f_N, \; f_1 \quad, \cdots, \; f_N, \quad \cdots, \quad f_1, \qquad \cdots, \quad f_r]'\end{aligned} \tag{1}$$

where $M = KN+r$ with K the number of complete repetition of the periodic component and r is the length of the remainder.

For a given N, this relation can be written as a linear vector matrix relation $\boldsymbol{g} = \boldsymbol{H}_N \boldsymbol{f}$ with $\boldsymbol{H}_N$ a MxN matrix with the following structure

$$\boldsymbol{H}_N = [\boldsymbol{I}_N | \boldsymbol{I}_N | ... | \boldsymbol{I}_N | \boldsymbol{I}(:, 1:r)]' \tag{2}$$

where $\boldsymbol{I}_N$ is the identity matrix of size $(N \times N)$ and $\boldsymbol{I}_N(:,1:r)$ is its first r columns.

$\widetilde{\boldsymbol{f}} = \boldsymbol{H}'_N \boldsymbol{g}$ is a vector where

$$\widetilde{f}_j = \begin{cases} \sum_{k=1}^{K} g((k-1)N+j) & \text{for } j = 1, \cdots, r \\ \sum_{k=1}^{K-1} g((k-1)N+j) & \text{for } j = r+1, \cdots, N \end{cases} \tag{3}$$

The problem now is expressed in this way: from the time series observations $\boldsymbol{g}$, determine the period N and the shape $\boldsymbol{f}$ from the noisy data $\boldsymbol{g}$ related by:

$$\boldsymbol{g} = \boldsymbol{H}_N \boldsymbol{f} + \boldsymbol{\epsilon} \tag{4}$$

where the vector $\boldsymbol{\epsilon}$ represents the noise and all the other uncertainties.

From here, there are at least three approaches: Least Square (LS) methods, Regularization based methods and finally the Bayesian approach. In this paper, the third approach is followed.

A. Bayesian approach

A simple way to start is to consider the following two equations:

$$\begin{cases} \text{Forward model} \quad : & \boldsymbol{g} = \boldsymbol{H}_N \boldsymbol{f} + \boldsymbol{\epsilon} \\ \text{Regularity model} : & \boldsymbol{D}_N \boldsymbol{f} = \boldsymbol{\xi} \longrightarrow \boldsymbol{f} = \boldsymbol{D}_N^{-1} \boldsymbol{\xi} \end{cases} \tag{5}$$

where $\boldsymbol{\epsilon}$ contains all the measurement and modelling errors, $\boldsymbol{D}_N$ is a linear regularity operator and $\boldsymbol{\xi}$ is its error term. The prior distributions on $\boldsymbol{\epsilon}$, on $\boldsymbol{\xi}$, and on N are then assigned and the expressions of the likelihood $p(\boldsymbol{g}|\boldsymbol{f}, N)$ and the prior distributions $p(\boldsymbol{f}|N)$ and $p(N)$ are obtained. Then, they are used to obtain the posterior distribution $p(N, \boldsymbol{f}|\boldsymbol{g})$ from which the inference on N and $\boldsymbol{f}$ can be done. Let start by a very classical case by assigning Gaussian probability distributions on $\boldsymbol{\epsilon}$ and $\boldsymbol{\xi}$. Then, we have:

$$p(\boldsymbol{g}|\boldsymbol{f}, N, v_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}_N \boldsymbol{f}, v_\epsilon \boldsymbol{I}_M),$$

and

$$p(\boldsymbol{f}|N, v_f) = \mathcal{N}\left(\boldsymbol{f}|\boldsymbol{0}, v_f (\boldsymbol{D}_N \boldsymbol{D}'_N)^{-1}\right)$$

where, first we assume known v_ϵ and v_f. We will come back to this point later.

We also assign a uniform prior for N in an appropriate range $[N_{\text{min}}, N_{\text{max}}]$ for example $N_{\text{min}} = N/8$ and $N_{\text{max}} = N/2$. Then, we have:

$$\begin{array}{l} p(N, \boldsymbol{f}|\boldsymbol{g}, v_\epsilon, v_f) \propto p(\boldsymbol{g}|\boldsymbol{f}, N, v_\epsilon) p(\boldsymbol{f}|N, v_f) p(N) \\ \propto c(N) \exp\left[-\frac{1}{2v_\epsilon}\|\boldsymbol{g} - \boldsymbol{H}_N \boldsymbol{f}\|^2 - \frac{1}{2v_f}\|\boldsymbol{D}_N \boldsymbol{f}\|^2\right] \end{array} \tag{6}$$

with $c(N)$ a function which does not depend on $\boldsymbol{f}$. For a fixed value of N, we can recognize that:

$$p(\boldsymbol{f}|\boldsymbol{g}, N, v_\epsilon, v_f) \propto \exp\left[-\frac{1}{2v_\epsilon} J(\boldsymbol{f})\right] \tag{7}$$

with

$$J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}_N \boldsymbol{f}\|^2 + \lambda \|\boldsymbol{D}_N \boldsymbol{f}\|^2, \quad \text{with } \lambda = v_\epsilon / v_f \tag{8}$$

which makes the connection to the quadratic regularization methods. However, here, we can also recognize that

$$p(\boldsymbol{f}|\boldsymbol{g}, N, v_\epsilon, v_f) = \mathcal{N}(\boldsymbol{f}|\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\Sigma}}) \tag{9}$$

where

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\} = [\boldsymbol{H}'_N \boldsymbol{H}_N + \lambda \boldsymbol{D}'_N \boldsymbol{D}_N]^{-1} \boldsymbol{H}'_N \boldsymbol{g} \tag{10}$$

and

$$\widehat{\boldsymbol{\Sigma}} = v_\epsilon [\boldsymbol{H}'_N \boldsymbol{H}_N + \lambda \boldsymbol{D}'_N \boldsymbol{D}_N]^{-1} \tag{11}$$

which can be used to quantify the uncertainty of the solution.

For $\lambda = 0$ the obtained solution becomes the Least Squares (LS) solution: $\widehat{\boldsymbol{f}} = [\boldsymbol{H}'_N \boldsymbol{H}_N]^{-1} \boldsymbol{H}'_N \boldsymbol{g}$. Looking to the structure of $\boldsymbol{H}_N$ and $[\boldsymbol{H}'_N \boldsymbol{H}_N]$, it is easy to write it:

$$\widetilde{f}_j = \begin{cases} \frac{1}{K} \sum_{k=1}^{K} g((k-1)N+j), & \text{for } j = 1, \cdots, r \\ \frac{1}{K-1} \sum_{k=1}^{K-1} g((k-1)N+j), & j = r+1, \cdots, N. \end{cases} \tag{12}$$

When N is not known, the expression of the joint posterior is

$$\begin{array}{l} p(N, \boldsymbol{f}|\boldsymbol{g}, v_\epsilon, v_f) \propto c(N) p(\boldsymbol{f}|\boldsymbol{g}, N, v_\epsilon, v_f)\, p(N) \\ \propto (2\pi)^{-N/2} |\boldsymbol{\Sigma}_N|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{f} - \widehat{\boldsymbol{f}})' \widehat{\boldsymbol{\Sigma}}_N^{-1} (\boldsymbol{f} - \widehat{\boldsymbol{f}})\right] \\ \quad p(N) \end{array} \tag{13}$$

with

$$\begin{cases} \widehat{\boldsymbol{f}} = [\boldsymbol{H}'_N \boldsymbol{H}_N + \lambda \boldsymbol{D}'_N \boldsymbol{D}_N]^{-1} \boldsymbol{H}'_N \boldsymbol{g} \\ \widehat{\boldsymbol{\Sigma}}_N = v_\epsilon [\boldsymbol{H}'_N \boldsymbol{H}_N + \lambda \boldsymbol{D}'_N \boldsymbol{D}_N]^{-1}, \quad \lambda = v_\epsilon / v_f. \end{cases} \tag{14}$$

The Joint Maximum A posteriori (JMAP) solution can be defined as

$$(\widehat{N}, \widehat{\boldsymbol{f}}) = \arg\max_{(N, \boldsymbol{f})} \{p(N, \boldsymbol{f}|\boldsymbol{g}, v_\epsilon, v_f)\} \tag{15}$$

where we may note that, compared to the regularization approach, here a prior for N is included, and more importantly there are extra terms as functions of N which automatically apply the Ockham Razor principle. An alternate optimization of this criterion has been implementes in [14] which is summarized here:

JMAP Alternate Optimization Algorithm:

```
1) for any $N \in [N_{min}, N_{max}]$:
     – compute  $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_N$ given in (eq. 14).
     – compute  $J(N) = (\frac{N}{2}) \ln(2\pi) + \frac{1}{2} \ln |\widehat{\boldsymbol{\Sigma}}_N|$
   end
2) find $\widehat{N} = \arg\min_N \{J(N)\}$.
3) compute the final shape $\boldsymbol{f}$ and posterior covariance $\widehat{\boldsymbol{\Sigma}}_N$
   using (eq. 14) with $N = \widehat{N}$.
```

As we can see, the great computational cost of this algorithm is in the computation of $\widehat{\boldsymbol{\Sigma}}_N$ for different values of N. This is very costly. Here, we propose a faster version of this algorithm which takes advantage of FT spectrogram to propose an initial distribution to sample from and then obtain a good approximation of the posterior distribution $p(N|\boldsymbol{g})$ from which we can infer on N. This new algorithm is summarized as follows:

New Algorithm:

```
0) Initialization: compute $\boldsymbol{G} = |\text{FFT}(\boldsymbol{g})|^2$, normalize it and
   define the distribution $p(N) \propto \boldsymbol{G}$ in the range $N \in [N_{min}, N_{max}]$.
1) Generate samples from $p(N)$
   – compute $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_N$ given in (eq. 14).
   – compute $J(N) = (\frac{N}{2})\ln(2\pi) + \frac{1}{2}\ln|\widehat{\boldsymbol{\Sigma}}_N|$
   end
2) find $\widehat{N} = \arg\min_N \{J(N)\}$.
3) compute the final shape $\widehat{\boldsymbol{f}}$ and posterior covariance $\widehat{\boldsymbol{\Sigma}}_N$
   using (eq. 14) with $N = \widehat{N}$.
```

The steps 2 and 3 are the same as before. We can also use the criterion in step 2 to update the probability distribution $p(N)$ and iterate steps 1 and 2. This remark brings us to a Gibbs sampling scheme where becomes:

Gibbs sampling Algorithm:

```
0) as in previous algorithm and generate a sample from it
1) compute $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\Sigma}}_N$ given in (eq. 14) and generate a
   sample $\boldsymbol{f}$ from $p(\boldsymbol{f}|N, \boldsymbol{g}) = \mathcal{N}(\boldsymbol{f}|\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\Sigma}})$
2) compute $J(N) = (\frac{N}{2})\ln(2\pi) + \frac{1}{2}\ln|\widehat{\boldsymbol{\Sigma}}_N|$, normalize
   it and use it as $p(N|\boldsymbol{f}, \boldsymbol{g})$ and generate a sample $N$ from
   it.
```

After generating a great number of $(\boldsymbol{f}, N)$ from this Gibbs sampling scheme, we can compute $\widehat{\boldsymbol{f}}$ and $\widehat{N}$.

As an extra advantage of this approach is also the fact that it can be unsupervised by choosing appropriate prior distributions for v_ϵ and v_f (inverse gamma) and then obtaining the expression $p(\boldsymbol{f}, N, v_\epsilon, v_f|\boldsymbol{g})$. We can then use this to do inference on all the unknowns and then propose a semi-supervised method. One method is to use a Variational Bayesian Approximation (VBA). In this paper, we report only the results with fixed value of $v_\epsilon = .001$ and $v_f = 1$.

III. Results

To illustrate the performance of the proposed method, we simulated several signals containing a few periods of known shapes and obtained the period and the shape of these periodic components with different approaches. Figure 2 shows three examples in two contexts: High SNR=20dB and Low SNR=5dB. In these examples, $M = 96$, $N = 29$. In the first example, the shape is the sum of a sinusoid and its second harmonic. In the second example the shape is a sum of positive and negative Gaussian waveforms and in the third example, the shape is a random shape waveform.

To illustrate the difficulties of the period estimation of these signals, the magnitude of their DFT are shown in Figure 3. As we can see, when the periodic shape is very regular (first and second examples) and the SNR is good enough (left column), then the DFT can show a maximum pic which can be used to determine the period, but when the shape is not very regular (third exampe), then this method cannot give the right answer. We may also note that, due to the very limited observation length and the fact that this length is not a multiple of the period, the DFT has a bias (the right value is shown in color red).

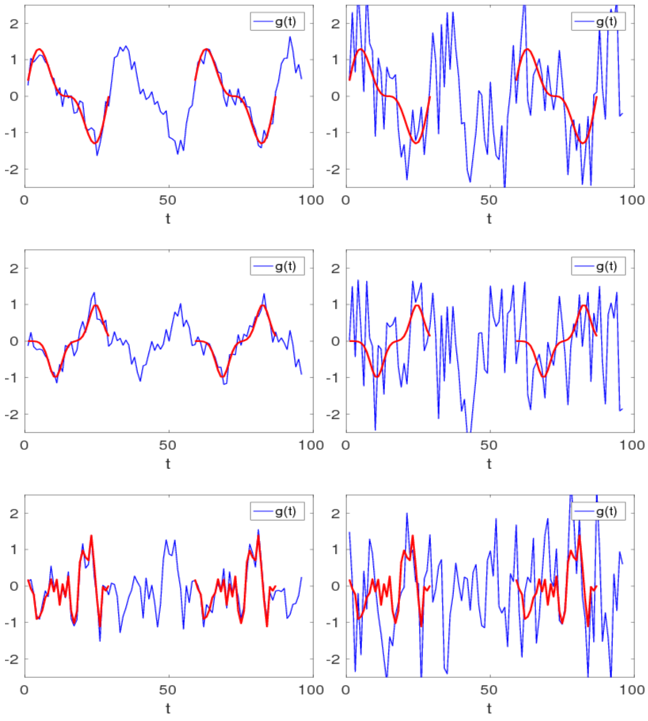


Fig. 2. Three examples of synthetically generated signals with two levels of noise: a) sum of a sinusoid and its second harmonic shape, b) sum of a positive and a negative Gaussian waveforms, c) a random shape waveform. Left: high SNR ratio (20 dB), Right: Low SNR (5dB).

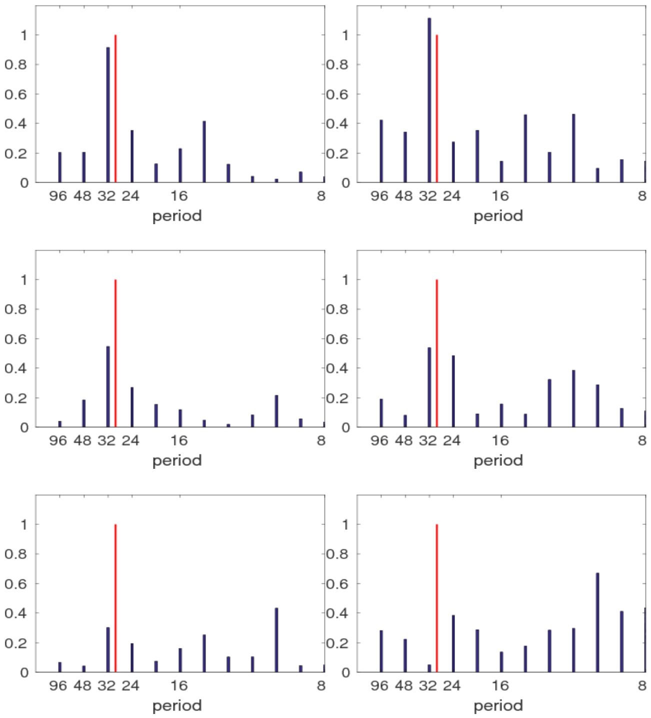


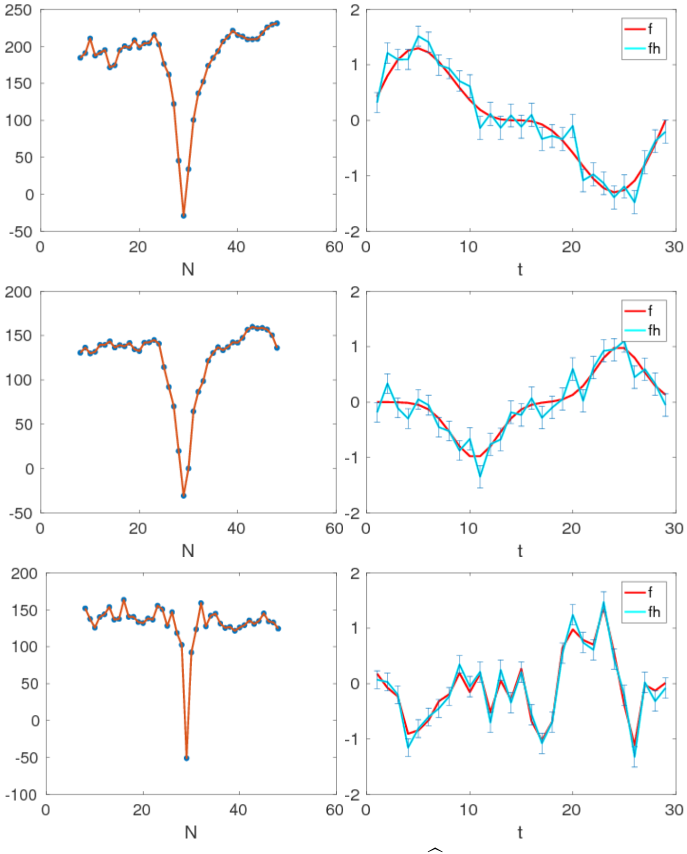
Fig. 3. The magnitude of the DFT of the three synthetically generated signals in Figure 2. When the periodic shape is very regular (first and second lines) and the SNR is good enough (left column), then the DFT can show a maximum pic which can be used to determine the period, but when the shape is not very regular (third line), then this method cannot give the right answer.

ISBN 978-0-9928626-7-1 © EURASIP 2017

In the following Figures, we illustrate the results obtained by the proposed algorithm, showing the observed data, the criterion for period selection and the estimated shapes compared to the truth with error bars computed using the diagonal elements of the posterior covariance.

A great number of other simulations are done to show:
– the performances of the algorithms as functions of SNR,
– the number $K = M/N$ of the periods in observed signal,
– sensitivity of the results with respect to the hyperparameters v_ϵ and v_f and their ratio $\lambda = \frac{v_\epsilon}{v_f}$, and
– sensitivity of the results with respect to different realization of the noise.

We also compared the convergency and the computational costs of the proposed algorithms. However, we cannot include them due to the lack of space, but they are available in a paper which will soon be submitted.

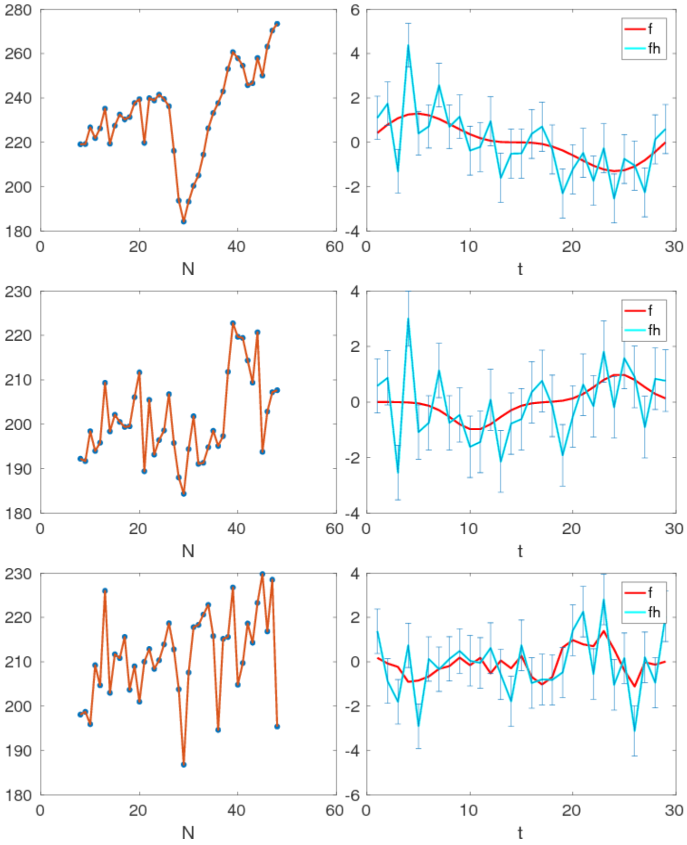


a) $J(N)$ b) $\widehat{f}(t)$ with error bars

Fig. 4. Three examples of period estimation results on simulated **high SNR data** shown on Fig 2: a) criterion for estimation of the period, b) Estimated shape

Figure 6 shows the performances of the proposed method as a function of SNR. As we can see, almost in all cases, the estimation of the period ($dn = \widehat{N} - N$) was exact, but the estimation error of the shape ($df = \frac{\|\widehat{\boldsymbol{f}} - \boldsymbol{f}\|}{\|\boldsymbol{f}\|}$) increases when SNR decreases.

Figure 7 shows the performances of the proposed method as a function of the relative length of the observed signal M over the period N. As we can see, here too, almost in all cases, except the case where M/N¡3, meaning that we have less than 3 periods in the the observed signal, the estimation of the period was exact. However, the estimation error of the shape does not change too much with M/N.



a) $J(N)$ $\widehat{f}(t)$ with error bars

Fig. 5. Three examples of period estimation results on simulated **low SNR data:** shown on Fig 2: a) criterion for estimation of the period, b) Estimated shape

Figure 8 shows the log-posterior $\ln p(N|\boldsymbol{g})$ for one of the examples

IV. Conclusion

We considered the problem of the estimation of an arbitrary shaped periodic component in a short duration noisy signal. As for such short signals, Discrete Fourier Transform (DFT) methods cannot give satisfactory results, we proposed a direct nonparametric forward model linking the unknown quantities (period and nonparametric shape of the periodic component) to the observed noisy signal. Based on this forward model, we proposed a Bayesian framework, where we obtained an expression for the joint posterior distribution of the period and the shape which is then used to estimate them. Using an approximate Bayesian computation approach, we proposed, in a first step, an algorithm which first estimates the period and then the shape of the periodic component. The cost of the computation of this algorithm is very high due to the exploration of all the possible values of the periods with given

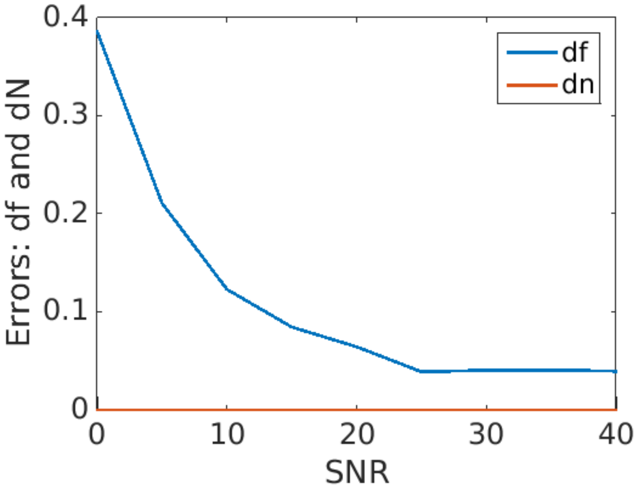


Fig. 6. Performances of the proposed method as a function of SNR. As we can see, almost in all cases, the estimation of the period was exact (dn=0), but the estimation error (df) of the shape decreases when SNR increases.

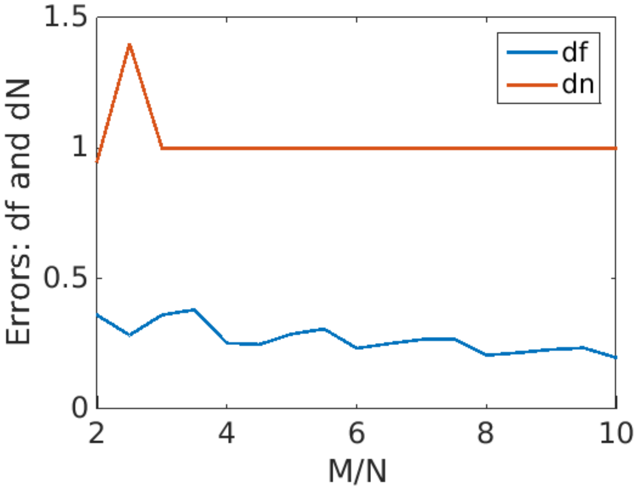


Fig. 7. Performances of the proposed method as a function of the ratio M/N. As we can see, almost in all cases, the estimation of the period was exact (dn=0), but the estimation error (df) of the shape decreases when SNR increases.

precision to compute its posterior probability, from which we look for the MAP estimate. When the period is estimated, then the estimation of the shape is easy. Trying to reduce this cost, a new faster algorithm is also proposed using the absolute periodogram of the signal as an initial probability distribution for the period from which samples are generated and used for a second stage where this distribution is updated. Finally, a Gibbs sampling scheme is proposed for exploring the the joint posterior space. Some simulation results with different simple and smooth or arbitrary random shapes showed the efficiency of the proposed method.

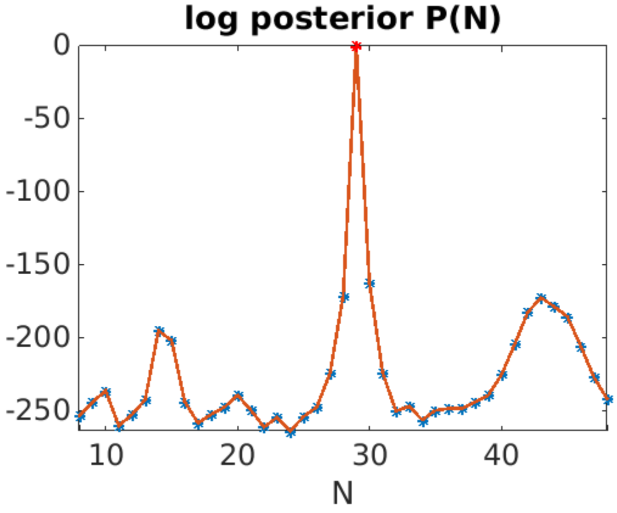


Fig. 8. $\ln p(N|\boldsymbol{g})$

References

[1] C. J. Sherr, "Cancer cell cycles.," *Science*, vol. 274, no. 5293, pp. 1672–1677, Dec 1996.

[2] Miika Ahdesmäki, Harri Lähdesmäki, Ron Pearson, Heikki Huttunen, and Olli Yli-Harja, "Robust detection of periodic time series measured from biological systems.," *BMC Bioinformatics*, vol. 6, pp. 117, 2005.

[3] Julien Epps, Hua Ying, and Gavin A. Huttley, "Statistical methods for detecting periodic fragments in DNA sequence data.," *Biol Direct*, vol. 6, pp. 21, 2011.

[4] Alan L. Hutchison, Mark Maienschein-Cline, Andrew H. Chiang, S M Ali Tabei, Herman Gudjonson, Neil Bahroos, Ravi Allada, and Aaron R. Dinner, "Improved statistical methods enable greater sensitivity in rhythm detection for genome-wide data.," *PLoS Comput Biol*, vol. 11, no. 3, pp. e1004094, Mar 2015.

[5] Mircea Dumitru, Ali Mohammad-Djafari, and Simona Baghai Sain, "Precise periodic components estimation for chronobiological signals through bayesian inference with sparsity enforcing prior.," *EURASIP J Bioinform Syst Biol*, vol. 2016, no. 1, pp. 3, Dec 2016.

[6] Paul F. Thaben and Pål O. Westermark, "Detecting rhythms in time series with rain.," *J Biol Rhythms*, vol. 29, no. 6, pp. 391–400, Dec 2014.

[7] Kuo-ching Liang, Xiaodong Wang, and Ta-Hsin Li, "Robust discovery of periodically expressed genes using the laplace periodogram.," *BMC Bioinformatics*, vol. 10, pp. 15, 2009.

[8] Daisuke Tominaga, "Periodicity detection method for small-sample time series datasets.," *Bioinform Biol Insights*, vol. 4, pp. 127–136, 2010.

[9] Michael L. Whitfield, Gavin Sherlock, Alok J. Saldanha, John I. Murray, Catherine A. Ball, Karen E. Alexander, John C. Matese, Charles M. Perou, Myra M. Hurt, Patrick O. Brown, and David Botstein, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors.," *Mol Biol Cell*, vol. 13, no. 6, pp. 1977–2000, Jun 2002.

[10] Sofia Wichert, Konstantinos Fokianos, and Korbinian Strimmer, "Identifying periodically expressed transcripts in microarray time series data.," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, Jan 2004.

[11] A. Mohammad Djafari, "Estimating the period and the shape of a periodic component in short duration signals," in *International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2016)*, 2016.

[12] Darya Chudova, Alexander Ihler, Kevin K. Lin, Bogi Andersen, and Padhraic Smyth, "Bayesian detection of non-sinusoidal periodic patterns in circadian expression data.," *Bioinformatics*, vol. 25, no. 23, pp. 3114–3120, Dec 2009.

[13] Emma Granqvist, Giles E D. Oldroyd, and Richard J. Morris, "Automated Bayesian model development for frequency detection in biological time series.," *BMC Syst Biol*, vol. 5, pp. 97, 2011.

[14] A. Mohammad-Djafari, "Efficient scalable variational bayesian approximation methods for inverse problems," in *SIAM Int. Conference on Uncertainty Quantification (UQ16), EPFL, April 4-8, 2016, Lausanne, Suisse*, 2016.