STOCHASTIC COMPLEXITY FOR THE DETECTION OF PERIODICALLY EXPRESSED GENES

Ciprian Doru Giurcăneanu

Tampere University of Technology, Institute of Signal Processing
P.O. Box 553, FIN-33101 Tampere, Finland
ciprian.giurcaneanu@tut.fi

ABSTRACT

The problem we address in this study is to decide, based on the available measurements, if a particular gene exhibits a periodic behavior. To this end we propose a principled method relying on the Stochastic Complexity (SC) whose computation is discussed for the generalized Gaussian distribution. We also investigate the relationship between SC, the well-known Minimum Description Length (MDL) formula, and the Bayesian Information Criterion (BIC). The performances of the SC-based approach are compared for simulated and real data with methods that are widely accepted in the bioinformatics community.

1. INTRODUCTION

The majority of the methods applied to investigate the periodicity of gene expression data either assume the frequency to be a priori known, or resort to finding the peaks of the periodogram. In [1], after identifying the peak of the periodogram computed from the time series available for a particular gene, its statistical significance is evaluated with the Fisher g-test. The final decision on the periodicity of the analyzed gene involves also the g-statistics of all other genes measured during the same experiment. It is important to mention that the expressions of thousands of genes are simultaneously measured, and only a very small subset of them are periodically expressed. The reference [2] proposes the use of the same test in conjunction with a robust spectral estimator that utilizes in the correlogram formula the correlations of the sample ranks instead of the usual estimates for the autocorellations. For simplicity, we will refer to the method from [1] as the Fisher Test (FT), and to the one from [2] as the Robust Method (RM). Remark that in TF and RM, drawing a conclusion on the cyclicity of a gene involves necessarily the measurements recorded for other genes during the same experiment.

In some of the previous studies, model-based methods have been applied for the detection of periodicity in gene expression data. For example, [3] relies on a linear combination of cubic B-spline basis to model the measurements. The shape of the function that describes the model is estimated from a training data set and further employed to analyze genes whose periodicity is not a priori known.

Since the training step prevents the use of such methods for poorly characterized organisms, a learning-free approach based on the Bayesian Information Criterion (BIC) was introduced in [4]. The derivation of the detection algorithm from [4] is done under the Gaussian hypothesis.

For gene expression data, it is of special interest to investigate the case when the distribution function of the noise has tails heavier than the Gaussian distribution with the same variance [2]. In this study, we propose a new method for the detection of the periodic signals in generalized Gaussian distribution (GGD). Note that GGD accommodates both the Gaussian and the heavy-tailed distributions [5]. The newly proposed method relies on the stochastic complexity (SC) [6] to decide if a data sequence is pure noise or periodic. For the particular Gaussian case, BIC is an asymptotic approximation of SC. As the analyzed data have small sample sizes, the non-asymptotic SC formula yields better results than BIC.

The rest of the paper is organized as follows. In Section 2 we briefly revisit the models for periodic gene expression data. Section 3 is focused on the computation of SC. The model selection performances of SC are evaluated in Section 4 with simulated and real data.

2. PROBLEM FORMULATION

The problem we address here is to decide if a gene is periodically expressed or not, based on the available measurements $\mathbf{y}^N = [y_{t_1} \dots y_{t_N}]^\top$. Similarly with the approaches from [1][2][4][7], the decision is recast as the problem of selection between the following two models:

$$\begin{aligned} &\mathcal{M}_0 : y_t = \mu + u_t, \\ &\mathcal{M}_1 : y_t = \mu + \alpha\cos(\omega t_n + \phi) + u_t, \end{aligned} \tag{1}$$

where $t \in \{t_1, \dots, t_N\}$, and $\mu, \alpha, \omega, \phi$ are non-random parameters that will be estimated from $\mathbf{y}^N$. The amplitude α is strict positive, the angular frequency ω belongs to the interval $(0, \pi)$, and the phase $\phi \in [-\pi, \pi)$.

In line with most of the studies on the periodicity of the gene expressions, we assume that the noise samples u_t are independent and identically distributed (i.i.d.). Hereafter the entries of the stochastic sequence $\mathbf{u}^N = [u_{t_1} \dots u_{t_N}]^\top$ are modeled with the GGD having zero mean, variance $\tau > 0$, and shape parameter $\nu > 1/2$:

$$p(u; \tau, \nu) = \frac{1}{2\Gamma(1+1/\nu)c_\nu} \exp\left(-\left|\frac{u}{c_\nu}\right|^\nu\right), u \in \mathbb{R}, \tag{2}$$

where $c_\nu = (\tau\Gamma(1/\nu)/\Gamma(3/\nu))^{1/2}$ and $\Gamma(\cdot)$ denotes the *Gamma* function [5].

GGD models a large family of symmetric distributions. For $\nu = 2$, GGD reduces to the Gaussian distribution, and it is easy to verify that GGD with $\nu = 1$ coincides with the Laplacian distribution, whereas GGD tends to the Uniform distribution for $\nu \to \infty$ [5]. Observe that the values of the shape parameter smaller than two correspond to peaked distributions that have tails heavier than the Gaussian distribution.

Assuming GGD for the additive noise u_t extends the Gaussian model from [4] and [7]. In [4], the proposed Bayesian detector was also tested for Laplacian and Uniform noise distributions, although it was designed under the Gaussian hypothesis.

This work was supported by the Academy of Finland, project No. 113572 and 213462.

Without loss of generality, we consider the re-parametrization $A = \alpha\cos\phi$, $B = -\alpha\sin\phi$, $C = \mu$, and (1) can be written equivalently as

$$\begin{aligned}\mathcal{M}_0 &: y_t = C + u_t,\\ \mathcal{M}_1 &: y_t = A\cos\omega t_n + B\sin\omega t_n + C + u_t,\end{aligned} \tag{3}$$

where $t \in \{t_1, \ldots, t_N\}$. Additionally we define $\beta = (2-\nu)/\nu$ [8], which allows to map the domain of the shape parameter from $(1/2, \infty)$ to $(-1, 3)$. The notations $\boldsymbol{\eta} = [C\ \tau\ \beta]$ and $\boldsymbol{\theta} = [A\ B\ C\ \omega\ \tau\ \beta]$ will be useful for the following discussion.

3. MODEL SELECTION WITH SC

Relying on the Minimum Description Length (MDL) principle, we select the model that minimizes the SC of the observations $\mathbf{y}^N$ [6]. SC is defined as the negative logarithm of the Normalized Maximum Likelihood, and it is recommended to be evaluated with the sharp formula from [9], especially when the sample size is small. The formula involves the integral of the squared root of the determinant of Fisher information matrix (FIM) over the parameter space. For the model selection problem (3), we do not have a priori bounds for all the parameters, and this posses troubles with the computation of the integral. Instead of using arbitrary bounds, we apply the results from [10], which leads to the following formulae

$$\begin{aligned}\mathrm{SC}(\mathbf{y}^N;\mathcal{M}_0) = &-\ln p(\mathbf{y}^N;\hat{\boldsymbol{\eta}}\,,\mathcal{M}_0) + \ln|\mathbf{J}_N(\hat{\boldsymbol{\eta}}\,,\mathcal{M}_0)|^{1/2}\\ &+\sum_{i=1}^{3}\ln(|\hat{\boldsymbol{\eta}}_{\,i}| + N^{-1/4}),\\ \mathrm{SC}(\mathbf{y}^N;\mathcal{M}_1) = &-\ln p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,,\mathcal{M}_1) + \ln|\mathbf{J}_N(\hat{\boldsymbol{\theta}}\,,\mathcal{M}_1)|^{1/2}\\ &+\sum_{i=1}^{6}\ln(|\hat{\boldsymbol{\theta}}_{\,i}| + N^{-1/4}),\end{aligned}$$

where $p(\mathbf{y}^N;\hat{\boldsymbol{\eta}}\,,\mathcal{M}_0)$ is the maximum likelihood (ML) under $\mathcal{M}_0$, and $p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,,\mathcal{M}_1)$ is the ML under $\mathcal{M}_1$. $\mathbf{J}_N(\boldsymbol{\eta}\,,\mathcal{M}_0)$ is the FIM whose expression is given by the celebrated formula $\mathbf{J}_N(\boldsymbol{\eta}\,,\mathcal{M}_0) = E\left[-\frac{\partial^2 \ln p(\mathbf{y}^N;\boldsymbol{\eta}\,,\mathcal{M}_0)}{\partial\boldsymbol{\eta}\,\partial\boldsymbol{\eta}^{\top}}\right]$. The definition of $\mathbf{J}_N(\boldsymbol{\theta}\,,\mathcal{M}_1)$ can be written similarly. We refer to [5] for results on the ML regularity conditions in various regions of the shape parameter space.

A discussion on the relationship between SC and other well-known selection rules is deferred to the Appendix.

Computation of the FIM We briefly investigate next how $|\mathbf{J}_N(\boldsymbol{\theta}\,,\mathcal{M}_1)|$ can be calculated. Since the signal parameters (ω, A, B, C) are independent of the noise parameters (τ, β), FIM is block diagonal and $|\mathbf{J}_N(\boldsymbol{\theta}\,,\mathcal{M}_1)| = |\mathbf{J}_N(\omega, A, B, C)|\times |\mathbf{J}_N(\tau,\beta)|$. Based on the results from [11] and [12], we write $\mathbf{J}_N(\omega, A, B, C) = \frac{\gamma_\beta}{\tau}\sum_{t=t_1}^{t_N}\mathbf{G}_t(\omega, A, B, C)$, where $\gamma_\beta = \frac{4}{(1+\beta)^2}\frac{\Gamma\left(\frac{3+3\beta}{2}\right)\Gamma\left(\frac{3-\beta}{2}\right)}{\Gamma^2\left(\frac{1+\beta}{2}\right)}$, and the entries of $\mathbf{G}_t(\omega, A, B, C)$ are given by the equations (24)-(33) from [13]. Remark that $\mathbf{J}_N(\omega, A, B, C)$ depends on the shape of the noise distribution only through the factor γ_β. It was proven in [11] that $\gamma_\beta \geq 1$ and attains its minimum for the Gaussian distribution ($\beta = 0$). The entries of $\mathbf{J}_N(\tau,\beta)$ can be easily calculated by resorting to the equations (19), (27) and (29) from [5]. Moreover, $\mathbf{J}_N(\boldsymbol{\eta}\,,\mathcal{M}_0)$ can be evaluated with the equations (15) and (28) from [5].

Estimation of the unknown parameters Obtaining the ML estimates for the parameters of a signal in GGD noise is difficult, especially when $\beta \in (1,3)$, or equivalently $\nu \in (1/2, 1)$ [5]. Since we do not known a priori the value of β, we have also to estimate the shape parameter from the available data, which makes the task even more difficult.

Input: The measurements $\mathbf{y}^N$, and $\beta_1, \ldots, \beta_M$, a set of non-zero values for the shape parameter.

1. Estimate $\hat{\omega}$ with a nonparametric method;

$$\mathbf{H} \leftarrow \begin{bmatrix}\cos\hat{\omega}t_1 & \sin\hat{\omega}t_1 & 1\\ \vdots & \ddots & \vdots\\ \cos\hat{\omega}t_N & \sin\hat{\omega}t_N & 1\end{bmatrix};$$

2. Least squares (LS) estimation

$\hat{\mathbf{x}} \leftarrow \left(\mathbf{H}^\top\mathbf{H}\right)^{-1}\mathbf{H}^\top\mathbf{y}^N$;
$\hat{\mathbf{u}} \leftarrow \mathbf{y}^N - \mathbf{H}\hat{\mathbf{x}}$;
$\mathbf{U} \leftarrow \mathrm{diag}(|\hat{\mathbf{u}}|)$;
$\hat{\tau} \leftarrow (\hat{\mathbf{u}}^\top\hat{\mathbf{u}})/N$;
$\hat{\boldsymbol{\theta}}\,[0] \leftarrow [\hat{\mathbf{x}}^\top\ \hat{\omega}\ \hat{\tau}\ 0]$;
$p[0] \leftarrow p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,[0],\mathcal{M}_1)$;

3. Weighted least squares (WLS) estimation

For $m = 1 : M$,

$\boldsymbol{\Upsilon} \leftarrow \mathbf{U}^{-2\beta_m/(1+\beta_m)}$;
$\hat{\hat{\mathbf{x}}} \leftarrow \left(\mathbf{H}^\top\boldsymbol{\Upsilon}\,\mathbf{H}\right)^{-1}\mathbf{H}^\top\boldsymbol{\Upsilon}\,\mathbf{y}^N$;
$\hat{\hat{\mathbf{u}}} \leftarrow \mathbf{y}^N - \mathbf{H}\hat{\hat{\mathbf{x}}}$;
$\hat{\hat{\tau}} \leftarrow \frac{\Gamma\left(\frac{3+3\beta_m}{2}\right)}{\Gamma\left(\frac{1+\beta_m}{2}\right)}\left(\frac{2\sum_{n=1}^{N}|\hat{\mathbf{u}}_n|^{2/(1+\beta_m)}}{N(1+\beta_m)}\right)^{1+\beta_m}$;
$\hat{\boldsymbol{\theta}}\,[m] \leftarrow [\hat{\hat{\mathbf{x}}}^\top\ \hat{\omega}\ \hat{\hat{\tau}}\ \beta_m]$;
$p[m] \leftarrow p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,[m],\mathcal{M}_1)$;

End

4. Estimation results

$m^* \leftarrow \arg\max_{0\leq m\leq M} p[m]$;
$\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}\,[m^*]$;
$\ln p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,,\mathcal{M}_1) \leftarrow \ln p(\mathbf{y}^N;\hat{\boldsymbol{\theta}}\,[m^*],\mathcal{M}_1)$;

Figure 1. One sine-wave in noise with incompletely characterized GGD: algorithm for the estimation of the signal and noise parameters.

We outline in Figure 1 the estimation algorithm proposed for the model $\mathcal{M}_1$. The pseudo-code aims to describe clearly the procedure, and it does not emphasize on the efficient numerical implementation. Some details are given below.

For estimating the frequency $\hat{\omega}$ at Step 1, one can pick the peak of the discrete Fourier transform of $\mathbf{y}^N$. Unfortunately, the gene expression data are unevenly sampled in most of the experiments, and when they are uniformly sampled, some data are occasionally missing. As the available data are not equidistantly spaced on the time-axis, it is recommended to use the CLEAN algorithm instead of the periodogram [14]. In our experiments we have applied both CLEAN and the frequency estimator utilized in the RM, and we noticed their similar performances. For sake of comparison, all the results reported in this study are obtained with the frequency estimator available with the implementation of the RM.

For the maximization of the likelihood function, we resort to a 1-D grid search in the space of the shape parameter. The number of grid points M is, as usual, a tradeoff between the estimation accuracy and the computational burden of the algorithm. In our experiments, we use in the space of the ν parameter a grid whose points are $0.6, 0.7, \ldots, 5.0$. The equivalent β values can be easily computed.

Once $\hat{\omega}$ is estimated and β is chosen to be a particular point of the grid, the estimation of $\mathbf{x} = [A\ B\ C]^\top$ reduces to an optimization problem that it is convex only for $\beta \leq 1$. We cannot restrict the search only to the values of the shape parameter that are smaller than one, because we are especially interested in heavy-tailed distributions that correspond to $\beta \in [1, 3)$. To circumvent the difficulties without resorting to computationally expensive methods, we apply a fast algorithm that was originally

introduced in [15]. Steps 2 and 3 implement the algorithm for our particular case: firstly the crude estimates $\hat{\mathbf{x}}$ are computed as the LS solution that maximizes the likelihood function for the Gaussian case ($\beta = 0$). The estimates are further refined for each point in the grid by solving a WLS problem whose weights depend on β.

The procedure described in Figure 1 returns both the estimated parameters and the maximum of the log-likelihood function. Remark that the algorithm can be easily modified to make use of prior knowledge. For example, in some studies on the periodicity of the gene expressions it is assumed that the frequency ω is known. To exploit this information, we can simply skip the first Step of the algorithm. Furthermore, when evaluating SC we remove the row and the column of FIM that correspond to the frequency parameter, and we also drop the term $\ln(|\hat{\omega}| + N^{-1/4})$.

As $\boldsymbol{\eta}$ parameters are a subset of $\boldsymbol{\theta}$, the estimation algorithm for model $\mathcal{M}_0$ can be immediately obtained from the pseudocode listed in Figure 1. There exist one important difference in our implementation: $\hat{C} = \text{median}(\mathbf{y}^N) \forall \beta \in [1, 3)$. More details on this estimator for C can be found in [5].

4. EXPERIMENTAL RESULTS

4.1. Simulated data

We illustrate the performances of the newly proposed method with an example inspired from [1] and [2]. We prefer to use in the description of the experiment the parametrization (1) because is more intuitive than (3). Due to the same reason, we employ for GGD the parameter ν and not β.

We investigate the model selection performances when the noise is GGD with unitary variance and $\nu \in \{0.6, 1.0, 2.0\}$. In our experimental settings, the parameter μ is zero. The synthetic data are assumed to be sampled at the time moments $1, 2, \ldots, N'$, where $N' \in \{30, 40, 50, 60, 70\}$.

For each value of N' and for each shape parameter ν, we generate 10^4 realizations. Half of them are non-periodic (model $\mathcal{M}_0$), and the rest are periodic with amplitude $\alpha = \sqrt{2}$ (model $\mathcal{M}_1$). To mimic the real case, for each periodic time series, the frequency $\omega/(2\pi)$ is chosen as an outcome of the Uniform distribution on $(0.08, 0.12)$, and the phase ϕ is an outcome of the Uniform distribution on $(-\pi, \pi)$. All the frequencies ω and the phases ϕ are statistically independent.

Since we are concerned with the influence of the missing points on the detection results, we remove $\lceil N'/4 \rceil$ measurements from each time series with length N'; the locations of the eliminated measurements are randomly chosen such that all the entries of $\{1, \ldots, N'\}$ have equal probability to be selected. The positions of the missing points are decided independently for each time series. Note that, for both the periodic and the non-periodic data sequences, the number of observations is decreased from N' to N. For completeness, we indicate between parentheses the value of N corresponding to each N': 30(22), 40(30), 50(37), 60(45), 70(52).

We use the synthetic data to test the model selection performances of the SC and of the crude MDL criterion given in the Appendix. When SC is applied in conjunction with the algorithm outlined in Figure 1, the resulting method is dubbed SCts because the core of the estimation algorithm is a two-step LS. If the model is selected with the MDL criterion, then the method is named MDLts.

Relying on the Gaussian hypothesis is equivalent with skipping the Step 3 of the algorithm in Figure 1. The estimates obtained under this hypothesis are further used in combination with either SC or MDL, and the resulting methods are named SCg and MDLg, respectively.

The model selection criteria are generally evaluated based on the probability of choosing the correct model when the ground truth is known. Therefore we are interested for all the methods on the empirical probability of choosing $\mathcal{M}_1$ when the simulated data are periodic. This probability is named P_D (detection probability), which is a term widely used in the engineering literature [11]. For the pure noise data we dot report the probability of selecting $\mathcal{M}_0$, and we prefer an equivalent measure, namely the probability of deciding $\mathcal{M}_1$ when the test data are non-periodic. This is named P_{FA} (probability of false alarm) with a term borrowed from the detection theory. The interested reader can find in [11] the equivalence between this nomenclature and the one used in the statistics literature.

We plot in Figure 2 the values of P_D and P_{FA} versus the number of available measurements N when the shape parameter of the noise takes three different values. We have also considered in our comparisons the RM and the FT. For clarity of the graphs, we do not plot the results yielded by the FT because they are significantly worse than those produced by the RM. For the RM, we pick the largest P_{FA} for each graph in the right-hand-side column to be the significance level α. Observe that a horizontal line is drawn for the value of α in all the graphs within the right column. For each ν and N, α is used together with the 5000 synthetic time series that are pure noise in order to "calibrate" the RM. After this step, RM is applied to the rest of 5000 periodic time series for estimating P_D plotted in the left-hand-side columns. Since we employ the same notation for both the amplitude and the significance level, the interpretation of α will be clear from the context.

As P_{FA} is small for all the tested methods, the difference in performances is given by P_D. Observe for $\nu = 0.6$ that P_D is larger for the MDLts than for the MDLg, but the difference decreases when $\nu = 1$, and the P_D becomes smaller for the MDLts than for the MDLg when $\nu = 2$. A similar trend can be observed when comparing the P_D of the SCts and the SCg for ν increasing from 0.6 to 2. Thus the use of the the two-step algorithm is recommended only when the distribution of the noise is heavy-tailed. SC compares favorably with the MDL in all the experimental situations, with the remarkable exception of the case when $\nu = 0.6$ and $N = 22$. We can conclude that applying SC instead of MDL for model selection has a positive impact on the results. It is interesting to note that RM is superior to both MDLg and MDLts for almost all sample sizes when the noise is Gaussian ($\nu = 2$).

4.2. Molecular data

In the experiment Thy-Thy 3 from [16] the epithelial cell line Hela S3 is measured during 46 hours with a uniform sampling period of 1 hour. After discarding the clones that have more than 30% missing values [2], we analyze the rest of the 41508 clones with the four methods that have been already tested with synthetic data. We mention that 1134 clones have been labeled as periodic in [16].

We report for each tested method the number of clones identified as periodic, and also how many of them can be found in the list with 1134 entries provided by the supplemental material at *http://genome-www.stanford.edu/Human-CellCycle/Hela/*. For example, the number of periodically expressed clones found by SCts was 2481. As 533 of them are also in the aforementioned list, we write for conciseness SCts (2481;533). With the same notation, we give the results for the other three methods: MDLts (2568;552), SCg (2535;551) and MDLg (2451;556).

The analysis of the entire data set was performed with our Matlab implementation on a Pentium IV at 3.2 GHz. The execution time was 17.5 min. when applying the two-step algorithm and only 3 min. when the parameter estimation was performed under the Gaussian hypothesis.

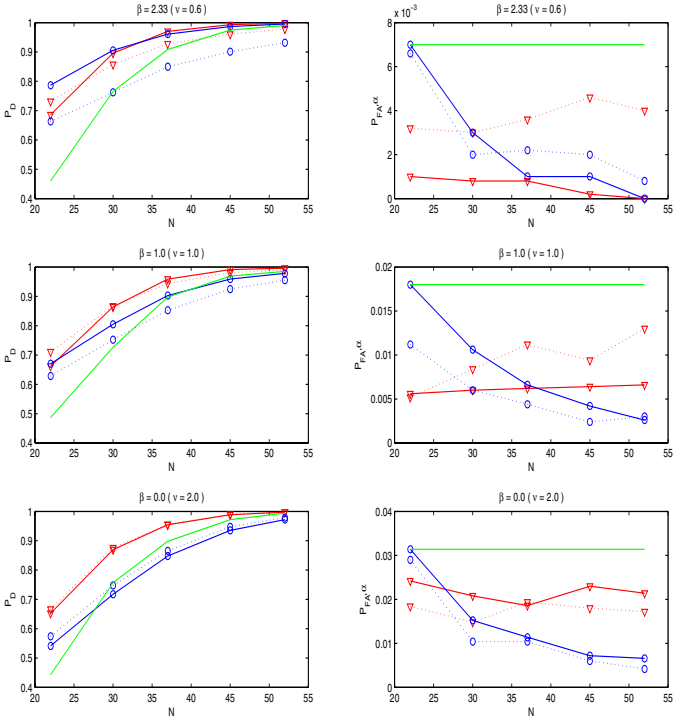

Figure 2. P_D and P_{FA} versus the sample size for five detection methods: SCts (continuous red line with a triangle), SCg (dottted red line with a triangle), MDLts (continuous blue line with a circle), MDLg (dotted blue line with a circle) and RM (continuous green line).

APPENDIX: MODEL SELECTION RULES

For simplicity, we consider only the case when the noise distribution is Gaussian, and consequently we do not account the contribution of β to the SC expression. Relying on results from [17], we can readily prove the asymptotic equivalence between SC and a crude MDL formula that amounts to the selection of the index $k \in \{0, 1\}$ to minimize

$$\text{MDL}(\mathbf{y}^N; \mathcal{M}_k) = -\ln p(\mathbf{y}^N; \hat{\boldsymbol{\zeta}}\,(k), \mathcal{M}_k) + \frac{5k+2}{2} \ln N,$$

where $\hat{\boldsymbol{\zeta}}\,(0) = [\hat{C}\ \hat{\tau}]^\top$ and $\hat{\boldsymbol{\zeta}}\,(1) = [\hat{\omega}\ \hat{A}\ \hat{B}\ \hat{C}\ \hat{\tau}]^\top$. The term $\frac{5k+2}{2} \ln N$ has the following significance: the penalty for the unknown frequency ω is $\frac{3}{2} \ln N$, and the penalty for each additional unknown parameter is $\frac{1}{2} \ln N$ [18].

It is well-known the equivalence between BIC and MDL, namely $\text{BIC}(\mathbf{y}^N; \mathcal{M}_k) = -\text{MDL}(\mathbf{y}^N; \mathcal{M}_k)$ [17]. Therefore, BIC decides that a gene is periodic whenever $\text{BIC}(\mathbf{y}^N; \mathcal{M}_1) - \text{BIC}(\mathbf{y}^N; \mathcal{M}_0) > 0$. A similar selection rule was applied in [4] to assign a gene as periodic: $\text{BIC}'(\mathbf{y}^N; \mathcal{M}_1) - \text{BIC}'(\mathbf{y}^N; \mathcal{M}_0) > \rho$, where the difference between the expressions of BIC′ and BIC is due to the penalty term. In BIC′, a penalty of $\frac{1}{2} \ln N$ is accounted for each parameter including the unknown frequency. We mention that previous studies have already shown the superiority of BIC in comparison with BIC′ [18]. To clarify the role of ρ, we quote from [4]: "Choice of detection threshold is arbitrary and must be decided by the investigator. Classical p-values for this test statistic may always be generated by means of resampling techniques but this is not a topic covered in this work."

MDL principle selects the model that leads to the shortest description length for the available data, thus a detection threshold is not needed, and we decide that a gene is periodic when $\text{SC}(\mathbf{y}^N; \mathcal{M}_1) < \text{SC}(\mathbf{y}^N; \mathcal{M}_0)$.

In the end of this Section, we mention that the similarities found in [19] between the Akaike's Information Criterion [17] and the Generalized Likelihood Ratio Test [11] can be extended to all the information theoretic criteria for which the penalty term depends only on the number of parameters and the number of the available measurements. A more elaborated discussion can be found in [6].

Acknowledgement The author is thankful to MSc M. Ahdesmäki for providing the Matlab implementations of the FT and the RM.

5. REFERENCES

[1] S. Wichert, K. Fokianos, and K. Strimmer, "Identifying periodically expressed transcripts in microarray time series," *Bioinformatics*, vol. 20, no. 1, pp. 5–20, 2004.

[2] M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, May 2005, http://www.biomedcentral.com/1471-2105/6/117.

[3] Y. Luan and H. Li, "Model-based methods for indentifying periodicaly expressed genes based on time course microarray gene expression data," *Bioinformatics*, vol. 20, no. 3, pp. 332–339, 2004.

[4] C. R. Anderssoon, A. Isaksson, and M. G. Gustafsson, "Bayesian detection of periodic mRNA time profiles without use of training examples," *BMC Bioinformatics*, Feb. 2006, http://www.biomedcentral.com/1471-2105/7/63.

[5] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *J. Acoust. Soc. Am.*, vol. 86, no. 4, pp. 1404–1415, 1989.

[6] J. Rissanen, *Information, complexity, and the MDL principle*, Springer Verlag, to appear.

[7] C. Zhou, J. Wakefield, and L. Breeden, "Bayesian analysis of cell-cycle gene expression data," UW Biostatistics Working Paper Series. Working Paper 276, Univ. of Washington, http://www.bepress.com/uwbiostat/paper276/, Dec. 2005.

[8] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*, Prentice Hall, 1993.

[9] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.

[10] G. Qian and H. R. Künsch, "Some notes on Rissanen's stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 782–786, Mar. 1998.

[11] S. M. Kay, *Fundamentals of statistical signal processing: detection theory*, Prentice Hall, 1998.

[12] A. Swami, "Cramer-Rao bounds for deterministic signals in additive and multiplicative noise," *Signal Processing*, vol. 53, pp. 231–244, 1996.

[13] P. Händel, "Properties of the IEEE-STD-1057 four-parameter sine wave fit algorithm," *IEEE Trans. Instr. Measur.*, vol. 49, no. 6, pp. 1189–1193, 2000.

[14] S. Baisch and G. H. R. Bokelmann, "Spectral analysis with incomplete time series: an example from seismology," *Computers & Geosciences*, vol. 25, pp. 739–750, 1999.

[15] D. Sengupta and S. M. Kay, "Parameter estimation and GLRT detection in colored non-Gaussian autoregressive processes," *IEEE Trans. Signal. Proces.*, vol. 38, no. 10, pp. 1661–1676, 1990.

[16] M. Whitfield, G. Sherlock, A. Saldanha, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown, and D. Botstein, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, pp. 1977–2000, 2002.

[17] P. Stoica and Y. Selen, "A review of information criterion rules," *IEEE Signal. Proces. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[18] P. M. Djuric, "A model selection rule for the sinusoids in white Gaussian noise," *IEEE Trans. Signal. Proces.*, vol. 44, no. 7, pp. 1744–1751, 1996.

[19] T. Söderström, "On model structure testing in system identification," *Int. J. Control*, vol. 26, no. 1, pp. 1–18, 1977.