Improved methods for the analysis of circadian rhythms in correlated gene expression data

Songklanakarin J. Sci. Technol.
40 (3), 692-700, May - Jun. 2018

http://www.sjst.psu.ac.th

Original Article

Improved methods for the analysis of circadian rhythms in correlated gene expression data

Wannapa Pukdee[1], Orathai Polsen[1], and Mohamed Fazil Baksh[2*]

[1] *Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bang Sue, Bangkok, 10800 Thailand*

[2] *Department of Mathematics and Statistics, School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, RG6 6FN United Kingdom*

Received: 7 October 2016; Revised: 21 February 2017; Accepted: 28 March 2017

Abstract

Circadian clocks regulate biological behaviours, such as sleeping and waking times, that recur naturally on an approximately 24-hour cycle. These clocks tend to be influenced by a variety of external factors, sometimes to the extent that it can have an impact on health. As an example in pharmacology, the effects of chemicals on the circadian rhythm in patients can be key to clarifying the relationship of drug efficacy and toxicity with dosing times. While pre-clinical experiments conducted to elucidate these effects may produce correlated data measured over time, such as gene expression profiles, existing methods for fitting parametric nonlinear regression models are, however, inadequate and can lead to unreliable, inconsistent parameter estimates and invalid inference. De-trending is widely used as a pre-processing step to address non-stationarity in the data, before fitting models based on the assumption of independence. However, as it is unclear that this approach properly accounts for the correlation structure, alternative methods that specifically model the correlation in the data based on conditional least squares and a two-stage estimation procedure are proposed and evaluated. A simulation study covering a wide range of scenarios and models shows that the proposed methods are more efficient and robust against model mis-specification than de-trending and, furthermore, they reduced estimation bias in the circadian period and provide more reliable confidence intervals.

Keywords: correlated gene expression data, de-trending method, nonlinear regression

1. Introduction

Most biological organisms, including humans, display an internal process (Erzberger *et al*., 2013) that regulates their behaviour according to the time of day. The internal clocks that determine the natural recurrence of biological processes, such as sleep and wake times, on a twenty-four-hour cycle are called circadian clocks (Cammack *et al*., 2006). In the study and development of drugs, circadian rhythms play a key role in understanding the relationship of efficacy and toxicity with dosing times (Paschos *et al*., 2010). Experimental adjustments of administration times of drugs can minimize the toxicity and maximize the efficacy of drugs. In addition, the cited reference provides examples of how circadian rhythms may affect the treatment of hypertension and cancer.

A gene is said to be expressed when it produces a functional product, such as protein molecules, used in an organism's cells. Bioluminescence is used in quantifying the gene expression of a cell. To generate bioluminescence an oxidative enzyme, in the case of circadian rhythms luciferase (Allard & Kopish, 2008), is implanted in the membrane of a living cell to produce light. The light emission is based on the conversion of chemical energy to radiation and is very efficient in terms of the released heat, i.e., most of the chemical energy is converted to radiation. Produced light intensity from the cells is then measured, and is used as a response variable in relevant experiments (Albert *et al*., 2008).

*Corresponding author
Email address: m.f.baksh@reading.ac.uk

These technologies allow scientists to detect changes in the expression of genes over time. Responses arising from the study of circadian gene expression are measurements of intensity, in relative units, over a course of time.

The data used in this paper were produced in the pre-clinical investigation phase of drug development by a pharmaceutical company. Human cells in a well-plate were treated with a chemical compound and the gene expression profiles were recorded. The experiment was replicated four times and the gene expression level in each well was measured every 1.5 hours for 78 hours. The cells within each well were synchronized because the measured expression is population average for a well, and our goal was to inspect circadian rhythms. This experimental design gives serially correlated observations for each well. Of interest in this paper is the development of models that efficiently capture the oscillatory time-pattern of gene expression while accounting for the correlation. Of particular interest is estimation of the period, as this provides information about the effects of the chemical compound on the circadian rhythm.

Usually the response level in a circadian gene expression experiment decreases with time. To adjust the observations for this trend, a pre-processing step is proposed by Yang and Su (2010) to remove the linear trend by using simple regression, and then the de-trended data are modelled by ordinary least squares (OLS). De-trending is widely used to fit models for correlated gene expression data, and it is assumed to produce independent errors based on stationarity assumptions. In order to address non-stationary correlated responses, the de-trended responses are fit with sinusoidal models (Izumo *et al*., 2003, 2006; Kyriacou & Hall, 1980; Maier *et al*., 2009) assuming independent errors. However, it is unclear that this de-trending (DET) method is adequate to account for the potential correlations in the responses, and further, de-trending produces correlated residuals. Properly accounting for correlated responses is important, as failure to do so can lead to biased parameter estimates and underestimation of their standard errors (Bender & Heinemann, 1995).

Conditional least squares (Bates & Watts, 1988) and two-stage estimation approach (Seber & Wild, 2003) are alternative strategies for fitting regression models to correlated data. These two methods are not based on time series assumptions, but rather they intend to address the correlation problem by explicitly modelling the correlation structure. Both conditional least squares and two-stage estimation methods utilize least squares procedures, and therefore benefit from the standard distributional properties of least squares estimators. They also tend to be computationally tractable. Neither method has previously been proposed in the literature for modelling circadian rhythms in correlated gene expression data.

This paper evaluates the conditional least squares and the two-stage estimation methods in nonlinear regression modelling of correlated gene expression data displaying an oscillatory pattern. The focus is on efficiency and reliability of these methods in estimating the oscillation period. The use of nonlinear models is novel to this application area. By directly modelling the trend and correlation pattern in the data, the limitations of the de-trending approach described above can be avoided. Comparisons of the proposed methods with the de-trending approach over a range of scenarios and models, including situations where the fitted model is incorrectly specified, are provided based on simulations.

2. Methods

Consider the nonlinear regression model of the relationship between an independent variable *t* and a dependent response variable *y* measured at *n* time points for each of *r* individuals,

$$\mathbf{y}_i = \mathbf{f}(\mathbf{t}_i;\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_i;\ i = 1,\ldots,r, \tag{1}$$

where $\mathbf{y}_i = (y_{i,1},\ldots,y_{i,n})'$ is the observed response vector for the *i*th individual, $\mathbf{t}_i = (t_{i,1},\ldots,t_{i,n})'$ is a time vector, $\mathbf{f}(\boldsymbol{\theta}) = (f(t_{i,1};\boldsymbol{\theta}),\ldots, f(t_{i,n};\boldsymbol{\theta}))'$ for some nonlinear function *f* of *t* with an unknown parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1},\ldots,\varepsilon_{i,n})'$ is an error vector. Assuming the repeated measures on each individual follow a stationary autoregressive process of order 1, AR(1), the error components then are linearly related between time points *j* and *j*-1

$$\varepsilon_{i,j} = \rho\varepsilon_{i,j-1} + \delta_{i,j};\ j = 1,\ldots,n, \tag{2}$$

where $|\rho| < 1$ is the correlation between $\varepsilon_{i,j-1}$ and $\varepsilon_{i,j}$, and $\delta_{i,j}$ are assumed to be normal, independent and identically distributed with zero mean and common variance σ^2.

Two possible ways to fit an AR(1) model when no assumptions are made on the joint distribution of the error terms are conditional least squares and the two-stage estimation method. Both methods, which are described below, fit the nonlinear regression model by least squares.

2.1 Conditional least squares estimation

The least squares estimation method is adapted to correlated responses by replacing the expected response from the model by a conditional expectation in the sum of squared deviations (Klimko & Nelson, 1978). In the case of correlated errors coming from a stationary AR(1) process in Equation (2) the conditional least squares (CLS) model can be shown to obey

$$\begin{aligned} y_{i,j} - f(t_{i,j};\boldsymbol{\theta}) &= \rho(y_{i,j-1} - f(t_{i,j-1};\boldsymbol{\theta})) + \delta_{i,j};\ j = 2,\ldots,n, \\ y_{i,j} &= \rho y_{i,j-1} + f(t_{i,j};\boldsymbol{\theta}) - \rho f(t_{i,j-1};\boldsymbol{\theta}). \end{aligned} \tag{3}$$

As normally distributed errors in an autoregressive model makes maximum likelihood equivalent to least squares estimation, the CLS method produces parameter estimates with similar properties as maximum likelihood estimators. In particular, the estimates obtained are consistent and asymptotically normal under mild regularity conditions (Klimko & Nelson, 1978). Note that the degrees of freedom for this model (3) are reduced by the first order autoregressive process, which impacts precision of the estimates. In addition, the increased number of model parameters increases the risk of convergence problems in iterative fitting.

2.2 Two-stage estimation

A two-stage (TS) approach that consists of two ordinary least squares (OLS) procedures, for estimating the parameters in nonlinear time series regression with autoregressive errors, has been proposed (Gallant & Goebel, 1976). Applied to the problem considered here, first the correlation structure is ignored and the model (1) is fitted by OLS to produce estimates $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ of $\boldsymbol{\theta}$ and fitted values $f(t_{i,j};\hat{\boldsymbol{\theta}}_{\text{OLS}})$. The residual vector for the *i*th individual,

$$\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathrm{f}(\mathbf{t}_i;\hat{\boldsymbol{\theta}}_{\text{OLS}}),$$

is used to produce an estimate of ρ (Park & Mitchell, 1980) given by

$$\hat{\rho}_i = \frac{\sum_{j=2}^{n} \hat{\varepsilon}_{i,j}\hat{\varepsilon}_{i,j-1}}{\sum_{j=2}^{n-1} \hat{\varepsilon}_{i,j}^2}.$$

In the second stage, by using the mean of $\hat{\rho}_1,\ldots,\hat{\rho}_r$, denoted $\hat{\rho}$, to estimate the common correlation ρ, a modified model (4)

$$z_{i,j} = g(t_{i,j};\boldsymbol{\theta}) + \delta_{i,j};\ i = 1,\ldots,r, \tag{4}$$

where

$$z_{i,j} = \begin{cases} (1-\hat{\rho}^2)^{\frac{1}{2}} y_{i,j} & ;\ j = 1 \\ y_{i,j} - \hat{\rho} y_{i,j-1} & ;\ j = 2,\ldots,n, \end{cases}$$

and

$$g(t_{i,j};\boldsymbol{\theta}) = \begin{cases} (1-\hat{\rho}^2)^{\frac{1}{2}} f(t_{i,j};\boldsymbol{\theta}) & ;\ j = 1 \\ f(t_{i,j};\boldsymbol{\theta}) - \hat{\rho} f(t_{i,j-1};\boldsymbol{\theta}) & ;\ j = 2,\ldots,n, \end{cases}$$

is constructed and fitted using OLS.

The TS procedure produces estimators with asymptotic properties similar to OLS estimators (Gallant & Goebel, 1976) and, unlike in CLS, no observations are excluded from the analysis.

2.3 Nonlinear functions

Although several functions can be found in the literature to model data displaying a sinusoidal pattern with a decreasing trend over time, in this paper the following three functions are considered as they display patterns consistent with real gene expression data. The one-sine function is a modified version of Izumo *et al*. (2003) with added decreasing trend

$$f(t_{i,j};\boldsymbol{\theta}) = \alpha + \beta t_{i,j} + a\exp(-dt_{i,j})\sin(\frac{2\pi t_{i,j}}{\tau} + \Phi),$$

where τ is the period, a is the amplitude, Φ represents the phase of the sine wave, d is a damping parameter, α is an intercept and β is a slope of the linear trend. The song-sine function modified from Kyriacou and Hall (1980) extends the one-sine function to allow a linear constant displacement a_s in the amplitude, and is given by

$$f(t_{i,j};\boldsymbol{\theta}) = \alpha + \beta t_{i,j} + (a_s + a\exp(-dt_{i,j}))\sin(\frac{2\pi t_{i,j}}{\tau} + \Phi).$$

Finally, in order to deal with the potential of more than one sinusoidal pattern, the two-sine with damping function

$$f(t_{i,j};\boldsymbol{\theta}) = \alpha + \beta t_{i,j} + a\exp(-dt_{i,j})\sin(\frac{2\pi t_{i,j}}{\tau} + \Phi) + b\sin(\frac{2\pi t_{i,j}}{\upsilon} + \Phi),$$

where b and υ are the amplitude and the period of the second sine term, respectively, is proposed as a novel function. Note that the possibility of more than one sine pattern has arisen in discussions with subject matter specialists.

3. Simulation Study

A simulation study was carried out to assess the methods in a variety of scenarios, including cases where the fitted model is incorrectly specified. In order to mimic the correlations in circadian gene expression over time, datasets were simulated with various levels of correlation ρ in the AR(1) process. In particular, the *i*th dataset ($i = 1,\ldots,r$) of size-*n* sample is generated from

$$y_{i,j} = \begin{cases} f(t_{i,j};\boldsymbol{\theta}) + \delta_{i,j} & ;\ j = 1 \\ \rho y_{i,j-1} + f(t_{i,j};\boldsymbol{\theta}) - \rho f(t_{i,j-1};\boldsymbol{\theta}) + \delta_{i,j} & ;\ j = 2,\ldots,n, \end{cases}$$

where $\delta_{i,j}$ are independent and identically distributed $N(0,\sigma^2)$.

Results presented in this paper are for simulated datasets generated under the parameter values $\boldsymbol{\theta}$ shown in Table 1. In addition, the AR (1) parameters are $\rho = (0, 0.25, 0.75)$ and $\sigma^2 = 25$. For each study, repeated measures are simulated for $r = 4$ independent individuals at times $t_{i,j} = 0, 1.5, \ldots, 78$, so that $n = 53$. The parameter values were selected so that the simulated datasets resemble observed circadian expression data.

For instance, the value $\tau = 24$ is in the range of circadian period length (20-28h) determined by Yang and Su (2010). Shown in Figure 1 are examples of synthetic datasets generated by the three models in the previous section.

For each simulation run, a total of 10,000 replicate studies are generated and analysed using R (R Core Team, 2013) with the nls function based on Gauss-Newton algorithm; see Ritz and Streibig (2008) and Crawley (2013) for

Table 1. The three sets of parameter values used in the simulations.

Model	$\boldsymbol{\theta}$								
	τ	υ	a_s	a	b	Φ	d	α	β
one-sine	24	-	-	180	-	0.31	0.07	330	-3
song-sine	24	-	0.5	180	-	0.31	0.07	330	-3
two-sine with damping	24	35	-	180	0.5	0.31	0.07	330	-3

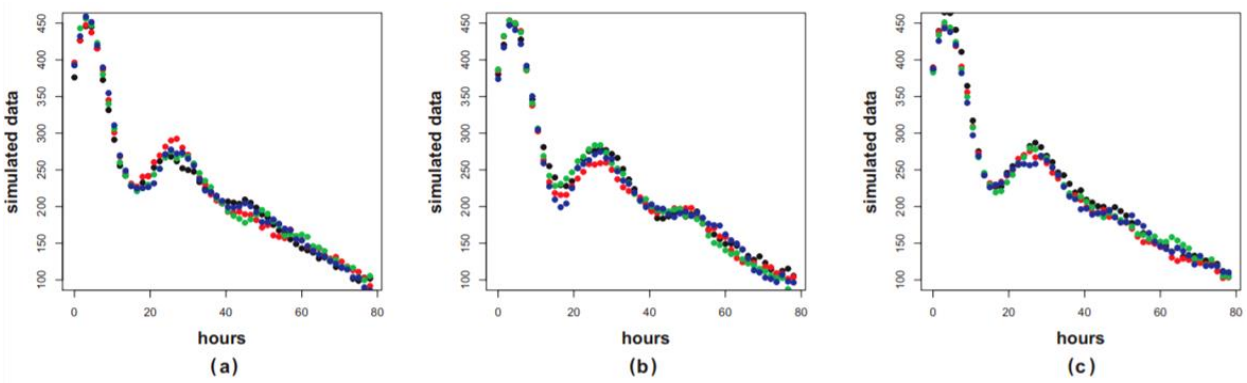

Figure 1. Example of the synthetic time-series datasets generated by the following functions: (a) one-sine, (b) song-sine and (c) two-sine with damping with AR(1) errors at $\rho = 0.75$ and $\sigma^2 = 25$.

details. In order to assess the efficacy of each method, parameter estimates were investigated and compared in terms of bias, relative difference between the standard deviation of estimates from replicate studies and the mean of standard errors produced by non-linear least squares fitting, root mean square errors, and coverage probability.

The main parameter of interest to identify from circadian rhythm data is the period τ, since it is used to predict the body's response to treatment and in the design of proper protocols for drug administration. Let $\hat{\tau}_m$ denote the period estimate from the *m*th simulation run, and let $\hat{\tau}$ be the average of $\hat{\tau}_m; m = 1, 2\ldots, M$. The bias of the estimator is defined as

$$\%\text{Bias} = 100\left(\frac{\text{mean}(\hat{\tau}) - \tau}{\tau}\right) = 100\left(\frac{\text{Bias}(\hat{\tau})}{\tau}\right).$$

Similarly, to assess the bias in variance estimates, the relative difference between the standard deviation and the standard error for the estimate is given by

$$\%\text{Diff} = 100\left(\frac{\text{SD}(\hat{\tau}) - \text{SE}(\hat{\tau})}{\text{SE}(\hat{\tau})}\right),$$

where $\text{SD}(\hat{\tau}) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\tau}_m - \text{mean}(\hat{\tau}))^2}$ and $\text{SE}(\hat{\tau}) = \frac{1}{M}\sum_{m=1}^{M}\text{SE}(\hat{\tau}_m)$, with $\text{SE}(\hat{\tau}_m)$ the standard error of the period estimate for the *m*th simulated dataset.

Efficiency of the method is measured by the root mean square error

$$\text{RMSE} = \sqrt{(\text{SD}(\hat{\tau}))^2 + (\text{Bias}(\hat{\tau}))^2}.$$

Finally, the estimate and the standard error are combined to construct the $100(1-\alpha)\%$ confidence interval (CI) for τ given by

$$\hat{\tau}_m - t_{\frac{\alpha}{2},\nu}\text{SE}(\hat{\tau}_m) \le \tau \le \hat{\tau}_m + t_{\frac{\alpha}{2},\nu}\text{SE}(\hat{\tau}_m),$$

where $t_{\frac{\alpha}{2},\nu}$ is the upper $\frac{\alpha}{2}$ quantile of student *t* distribution with ν degrees of freedom. How often the confidence interval covers the true value of τ provides an estimate of the coverage probability for τ and hence a measure of statistical inference validity.

Note that the Gauss-Newton algorithm does not necessarily convergence in all instances, so *M* is the total number of successful fits with converged parameters, and this differs between the different methods.

4. Results

This section presents simulation results from conditional least squares and two-stage methods in fitting the models described in Section 2.3. Also presented for comparison are the results from de-trending. Evaluations are presented both with the same type of model generating the data and fit to the data, as well as for cases with incorrectly specified fitted model. The latter cases reflect real-life conditions, where the data generating model is unknown, and help critically evaluate the robustness of the methods against model mis-specification.

Table 2 summarizes the performance in terms of bias (%Bias), relative difference (%Diff) and root mean square error (RMSE) for the methods, with the data generated by the one-sine model. The results show that for the correct model type at all ρ (0.00, 0.25 and 0.75), estimates from DET are negatively biased. Moreover, DET overestimates the variance of $\hat{\tau}$ and is consequently less efficient in terms of the RMSE. This leads to poor coverage probability, as shown in Figure 2 (a). On the other hand, CLS and TS produce unbiased estimates and good variance estimates. Consequently, their coverage probabilities are close to the expected value.

Table 2. Percentage bias, percentage relative difference and root mean square error of the period estimate $\hat{\tau}$ for DET, CLS and TS procedures when the true model is one-sine with $\tau = 24$.

Fitted model		ρ	DET			CLS			TS		
			%Bias	%Diff	RMSE	%Bias	%Diff	RMSE	%Bias	%Diff	RMSE
one-sine	τ	0.00	-1.3481	-57.8739	0.3460	0.0001	1.3136	0.1352	-0.0010	2.8511	0.1107
		0.25	-1.3402	-47.9025	0.3558	0.0057	2.1349	0.1857	0.0034	4.3817	0.1354
		0.75	-1.3114	-14.6151	0.4097	0.0239	4.9797	0.4021	0.0242	8.1855	0.2287
song-sine	τ	0.00	-0.7153	-43.1825	0.2461	0.0006	1.9455	0.1366	0.0001	3.5042	0.1116
		0.25	-0.6866	-24.5922	0.2866	0.0037	3.5934	0.1888	0.0027	5.5995	0.1367
		0.75	-0.2137	104.4706	0.6605	0.0341	10.9491	0.4228	0.0266	11.2017	0.2332
two-sine with damping	τ	0.00	-1.0407	39.7004	0.4935	-0.0189	4.4714	0.1463	-0.0398	6.8993	0.1229
		0.25	-0.9155	74.0320	0.5802	-0.0037	5.5841	0.2015	-0.0537	9.1860	0.1516
		0.75	-0.8384	96.0890	0.6624	0.1654	8.6487	0.4588	-0.0250	15.0083	0.2635
	υ	0.00	-163.4983	8013.8058	67.9124	34.7031	233.4452	24.0473	36.4217	279.8772	23.5832
		0.25	-159.3910	8328.1164	67.4969	33.7229	360.1641	31.1765	35.3224	290.9461	24.7837
		0.75	-142.6832	8043.6805	60.0745	34.2731	262.5797	30.4309	31.1163	383.1758	33.7434

Note: The period estimates for τ and υ when the fitted model is two-sine with damping.

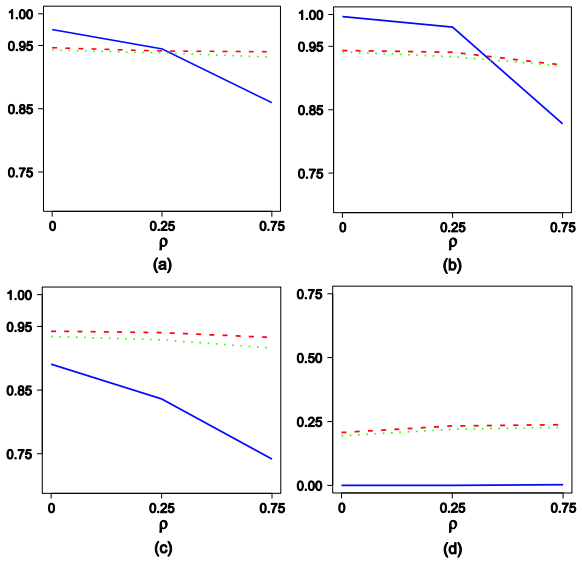

Figure 2. Plots of coverage probability of 95% confidence interval for the period τ using DET (solid line), CLS (dashed line) and TS (dotted line) when the true model is one-sine and the fitted model is (a) one-sine, (b) song-sine and (c) two-sine with damping. Coverage probability plots for υ in the two-sine with damping model are shown in (d).

For cases where the fitted model was mis-specified, when the simple song-sine function was used to fit the data, DET produced estimates that are less biased than when the correct model is fitted. However, the results also suggest that standard errors of parameter estimates are overestimated for low and moderate correlations and underestimated for strong correlations. This leads to overall poor coverage of the confidence intervals, as shown in Figure 2 (b). When the fitted model is two-sine with damping, both $\hat{\tau}$ and the standard error of $\hat{\tau}$ are underestimated by the DET method. On the other hand, both CLS and TS produced unbiased estimates in all cases but they tend to slightly underestimate the variances, especially when the data are strongly correlated, see Table 2. Also, as Figures 2 (a, b and c) show, CLS and TS procedures produce confidence intervals that are reasonably consistent with the theoretical expectations, albeit with slightly decreasing coverage as correlation increases.

All the methods perform poorly in estimation of the second period term υ when the fitted model is two-sine with damping. DET severely underestimates, whereas CLS and TS consistently overestimate υ, and all these methods underestimate the standard error. Not surprisingly, this estimation bias leads to the poor coverage probabilities shown in Figure 2 (d).

Following conclusions when the true models are song-sine and two-sine with damping can be drawn from Figures 3-4 and Tables 3-4. In the simulations with data generated under the song-sine model, results in Figure 3 and Table 3 show that DET again performs quite poorly, whether or not the fitted model is correctly specified. On the other hand, the findings for CLS and TS are consistent with the earlier results in Figure 2 and Table 2.

Table 4 shows simulation the results when the true model has an extra sine term with as second period ($\tau = 24$ and $\upsilon = 35$). The results again show that even though the fitted model was correctly specified, DET consistently underestimated both periods and produced variance estimates that are too small. In contrast, by explicitly modelling correlation in the data, the proposed CLS and TS methods perform far better in all cases. Moreover the coverage probabilities under CLS are close to 0.95 but, as TS produces slight underestimates of variances (as given by %Diff), its coverage probability shown in Figure 4 is slightly less than expected.

In summary, CLS and TS give more efficient estimates and are comparatively robust against model mis-specification. This reduces bias in estimates of the circadian period and gives better coverage probabilities. On the other hand, by not properly accounting for the correlation, DET has biases in estimates of period and standard error.

5. Example

To compare the DET method with the proposed CLS and TS methods in real-life situations, all three methods, were applied to data that, as explained in the introduction, comes from experiments run over 78 hours with a drug treat-

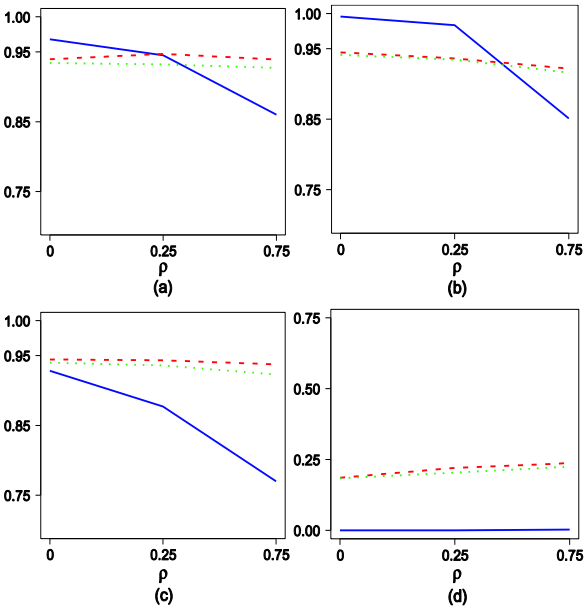

Figure 3. Plots of coverage probability of 95% confidence interval for the period τ using DET (solid line), CLS (dashed line) and TS (dotted line) when the true model is song-sine and the fitted model is (a) one-sine, (b) song-sine and (c) two-sine with damping. Coverage probability plots for υ in the two-sine with damping model are shown in (d).

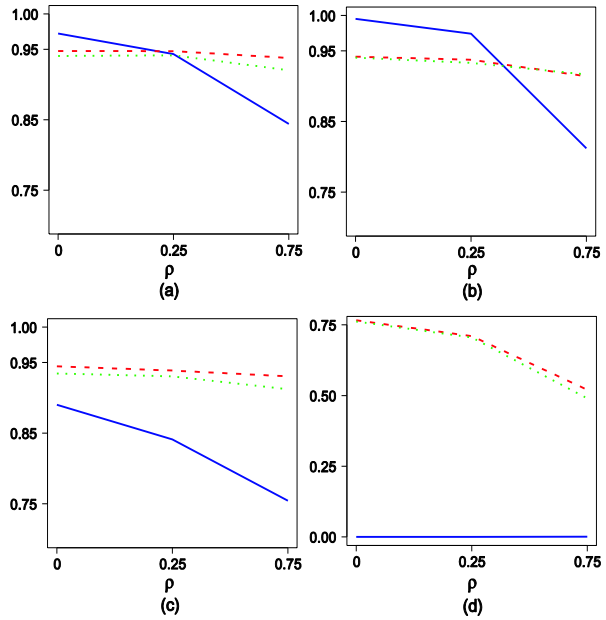

Figure 4. Plots of coverage probability of 95% confidence interval for the period τ using DET (solid line), CLS (dashed line) and TS (dotted line) when the true model is two-sine with damping and the fitted model is (a) one-sine, (b) song-sine and (c) two-sine with damping. Coverage probability plots for υ in the two-sine with damping model are shown in (d).

Table 3. Percentage bias, percentage relative difference and root mean square error of the period estimate $\hat{\tau}$ for DET, CLS and TS procedures when the true model is song-sine with $\tau = 24$.

Fitted model		ρ	DET			CLS			TS		
			%Bias	%Diff	RMSE	%Bias	%Diff	RMSE	%Bias	%Diff	RMSE
one-sine	τ	0.00	-1.2403	-58.5760	0.3205	0.0564	0.4097	0.1325	0.0364	2.4322	0.1089
		0.25	-1.2461	-48.5675	0.3336	0.0545	1.2011	0.1812	0.0299	4.5569	0.1337
		0.75	-1.2326	-14.8026	0.3923	0.0738	3.6574	0.3897	0.0291	8.7886	0.2266
song-sine	τ	0.00	-0.7335	-46.3534	0.2399	-0.0007	1.8875	0.1328	-0.0009	3.5155	0.1091
		0.25	-0.7045	-30.2979	0.2714	0.0034	3.6546	0.1833	0.0019	5.6232	0.1338
		0.75	-0.3208	87.5651	0.5989	0.0270	10.3293	0.4068	0.0259	11.1047	0.2282
two-sine with damping	τ	0.00	-1.1211	5.3314	0.4112	0.0470	2.4027	0.1397	-0.0049	4.3776	0.1163
		0.25	-0.9925	47.9338	0.5033	0.0857	3.6965	0.1940	-0.0143	7.0522	0.1449
		0.75	-0.8836	84.2489	0.6193	0.2289	7.4677	0.4459	0.0087	14.0669	0.2570
	υ	0.00	-153.3599	7803.6494	62.4126	39.6778	262.1625	26.2878	38.5423	239.0626	24.4712
		0.25	-146.9491	7327.3300	55.6965	33.7229	360.1641	31.1765	35.3224	269.1014	27.2147
		0.75	-133.8591	7826.7206	55.8487	38.7591	269.6656	31.7897	36.5362	337.7400	36.2693

Note: The period estimates for τ and υ when the fitted model is two-sine with damping.

Table 4. Percentage bias, percentage relative difference and root mean square error of the period estimate $\hat{\tau}$ for DET, CLS and TS procedures when the true model is two-sine with damping with $\tau = 24$ and $\upsilon = 35$.

Fitted model		ρ	DET			CLS			TS		
			%Bias	%Diff	RMSE	%Bias	%Diff	RMSE	%Bias	%Diff	RMSE
one-sine	τ	0.00	-1.2168	-57.5719	0.3177	0.2142	2.1370	0.1468	0.2055	3.3333	0.1224
		0.25	-1.2087	-47.5178	0.3290	0.2221	3.2454	0.1969	0.2073	4.8921	0.1457
		0.75	-1.1777	-13.8429	0.3895	0.2546	6.6532	0.4170	0.2190	8.7644	0.2371
song-sine	τ	0.00	-0.8361	-43.8445	0.2661	-0.0169	2.3665	0.1381	0.0535	3.7040	0.1133
		0.25	-0.7988	-21.9388	0.3098	-0.0404	4.4336	0.1914	0.0698	6.0290	0.1391
		0.75	-0.2610	116.3003	0.7000	-0.0677	12.716	0.4310	0.1225	12.5079	0.2389
two-sine with damping	τ	0.00	-1.2081	30.1877	0.4925	-0.0282	2.3439	0.1410	-0.0467	4.1905	0.1185
		0.25	-1.0585	64.8227	0.5701	0.0016	5.1759	0.1979	-0.0395	7.8307	0.1483
		0.75	-1.0024	83.7758	0.6399	0.1259	9.1935	0.4581	0.0296	15.9557	0.2655
	υ	0.00	-135.2289	8163.4433	74.4227	-3.1040	245.8983	17.2457	-2.5064	188.3904	15.7257
		0.25	-132.2588	8097.1445	70.3876	-5.3984	255.0431	23.2962	-4.0577	245.8808	20.4272
		0.75	-123.5109	8008.4038	65.1504	-6.7022	258.6963	29.4377	-10.5172	422.4302	36.1448

Note: The period estimates for τ and υ when the fitted model is two-sine with damping.

ment. As mentioned before, the same treatment was applied to four sets of cells, each measured every 1.5 hours. The intensity of bioluminescence was measured as indicator of a gene's expression level. In the data analysis, only those responses that showed an effect at 0h were included. A scatter plot of the data displays cyclic patterns with a linear decreasing trend over time, as seen in Figure 5. The sinusoidal functions described in Section 2 with autoregressive errors of order 1, AR (1), were tested for modelling these data.

In order to compare the performances of DET with the proposed methods, CLS and TS, Table 5 summarizes the analyses in terms of the 95% confidence interval (CI) for τ, and the residual standard errors $\hat{\sigma}$ from DET, CLS and TS approaches. Table 6 shows the lack of fit tests comparing residuals from the nonlinear models to residuals for one-way ANOVA models of the replicate observations at each time point that account for the correlation structure. Plots of the fitted models are given in Figure 6.

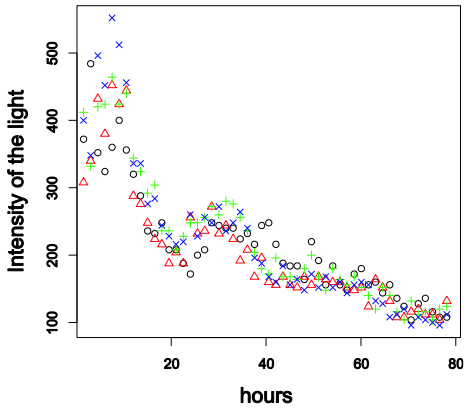

Figure 5. Circadian gene expression over time as measured by intensity of light, in relative units. The four replicates at each time point are shown with different symbols.

Table 5. Estimates and CI's of the circadian period in a real gene expression dataset obtained using three different models fitted by DET, CLS and TS procedures.

Fitted model		DET		CLS		TS	
		95% CI	$\hat{\sigma}$	95% CI	$\hat{\sigma}$	95% CI	$\hat{\sigma}$
one-sine	τ	24.74 ± 1.14	33.37	24.15 ± 1.89	26.39	26.50 ± 1.63	27.74
song-sine	τ	25.88 ± 1.10	32.99	23.97 ± 1.45	26.35	26.89 ± 1.73	27.73
two-sine with damping	τ	24.75 ± 1.15	33.43	24.89 ± 2.48	26.45	26.45 ± 1.46	27.44
	υ	-6.15 ± 0.23		29.38 ± 3.76		55.16 ± 8.54	

Table 6. Lack of fit test for one-sine, song-sine and two-sine with damping models fitted by DET, CLS and TS.

Fitted model	DET		CLS		TS	
	F	*p*-value	F	*p*-value	F	*p*-value
one-sine	2.561	8.703E-06	1.425	0.062	1.334	0.099
song-sine	2.451	2.467E-05	1.416	0.067	1.353	0.090
two-sine with damping	2.638	6.011E-06	1.464	0.052	1.141	0.274

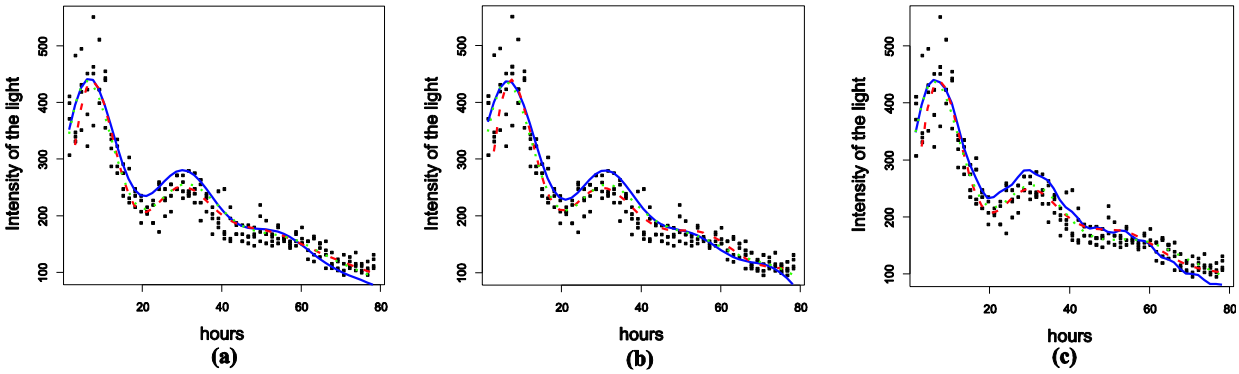

Figure 6. Fitted models (a) one-sine, (b) song-sine and (c) two-sine with damping to gene expression observations using DET (solid line), CLS (dashed line) and TS (dotted line) procedures.

The results show that for all the fitted models, the CLS estimates of the circadian periods are approximately 24 h with standard errors that are smaller than those obtained using DET and TS. The TS approach produces period estimates that are approximately 26 h with moderate residual errors. In contrast, the DET method produces period estimates around 25 h with the largest residual standard errors.

The lack of fit tests show that the CLS and TS methods provide good fits to the data, since there is no evidence of lack of fit. However DET fit the data poorly, as presented in Table 6. This is substantiated by plots of the fitted models, showing that the proposed methods produced the best fit to the observed cyclic pattern, as shown in Figure 6.

6. Conclusions

In this paper, we compared de-trending (DET) as the current baseline method for analyzing circadian rhythms in gene expression profiles to conditional least squares (CLS) and two-stage (TS) estimation as alternative methods. Simulation results clearly suggest that DET produced biased estimates of the circadian period and poor variance estimates, leading to invalid statistical inference. On the other hand, the proposed methods are not only much more efficient and robust against model mis-specification, but also had reduced bias in estimates of the circadian period and more reliable confidence intervals. The TS method produced slightly poorer confidence intervals than CLS in cases with high correlation, due to underestimated standard errors of parameter estimates. Although both proposed alternative methods provided good fits to real data, CLS produced more valid confidence intervals. In further work, we will propose methods for comparatively accurate variance estimation by maximum likelyhood, and will explore more sophisticated models capable of capturing complex data patterns.

The work here clearly illustrates de-trending to address non-stationarity of correlated data, although commonly used, should be undertaken with caution. In contrast, methods that explicitly account for the correlation, such as conditional least squares and two-stage estimation of non-

linear regression models, are viable and potentially more reliable and robust against model mis-specification. Finally, approaches such as CLS and TS are relatively straightforward to implement using standard statistical software packages, and their usage, for example in human drug development studies to understand circadian rhythms interfering with drug metabolism, should be encouraged.

Acknowledgements

We would like to thank the Ministry of Science and Technology (MOST) of Thailand for the financial support.

References

Albert, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2008). *Molecular biology of the cell*. New York, NY: Garland Science.

Allard, S. T. M., & Kopish, K. (2008). Luciferase reporter assays: powerful, adaptable tools for cell biology research. *Cell Notes Issue*, *21*(1), 1-4.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York, NY: John Wiley and Sons.

Bender, R., & Heinemann, L. (1995). Fitting nonlinear regression models with correlated errors to individual pharmacodynamics data using SAS software. *Journal of Pharmacokinetics and Biopharmaceutics, 23* (1), 87-100.

Cammack, R., Attwood, T. K., Campbell, P. N., Parish, J. H., Smith, A. D., Stirling, J. L., & Vella F. (2006). *Oxford dictionary of biochemistry and molecular biology*. New York, NY: Oxford University Press.

Crawley, M. J. (2013). *The R book.* West Sussex, England: John Wiley and Sons.

Erzberger, A., Hampp, G., Grannada, A. E., Albrecht, U., & Herzel, H. (2013). Genetic redundancy strengthens the circadian clock leading to a narrow entrainment range. *Journal of the Royal Society Interface, 10*(1), 1-11.

Gallant, A. R., & Goebel, J. J. (1976). Nonlinear regression with autoregressive errors. *Journal of the American Statistical Association, 71*(356), 961-967.

Izumo, M., Johnson, C. H., & Yamazaki, S. (2003). Circadian gene expression in mammalian fibroblasts in revealed by real-time luminescence reporting: Temperature compensation and damping. *The National Academy of Sciences of the USA, 100*(26), 16089-16094.

Izumo, M., Sato, T. R., Straume, M., & Johnson, C. H. (2006). Quantitative analyses of circadian gene expression in mammalian cell cultures. *PLoS Computational Biology, 2*(10), 1248-1261.

Klimko, L. A., & Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *The Annals of Statistics, 6*(3), 629-642.

Kyriacou, C. P., & Hall, J. C. (1980). Circadian rhythm mutations in Drosophila melanogaster affect short-term fluctuations in the male's courtship song. *The National Academy of Sciences of the USA, 77*(11), 6729-6733.

Maier, B., Wendt, S., Vanselow, J. T., Wallach, T., Reischl, S., Oehmke, S., . . . Kramer, A. (2009). A large-scale functional RNAi screen reveals a role for CK2 in the mammalian circadian clock. *Genes & Development, 23*(1), 708-718.

Park, R. E., & Mitchell, B. M. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics, 13*(2), 185-201.

Paschos, G. K., Baggs, J. E., Hogenesch, J. B., & FitzGerald, G. A. (2010). The role of clock genes in pharmacology. *The Annual Review of Pharmacology and Toxicology, 50*(1), 187-214.

R Core Team. (2014, October 31). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R, Use R!*. New York, NY: Springer.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear Regression*. New York, NY: Wiley Interscience.

Yang, R., & Su, Z. (2010). Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. *Bioresource Technology, 26*(1), 168-174.