Design of Experiments for Static Influences on Harmonic Processes

Winfried Theis,
Collaborative Research Center 475,
Universität Dortmund, Germany*

November 25, 2002

Abstract

In mechanics it happens that some fixed influencing factors determine the nature of a harmonic process. This can be modelled by regression of the influencing factors on periodogram ordinates of the relevant frequencies. Thereby the time-domain is bypassed, and static models can be applied. Since it is known that periodogram ordinates are (non-central) chi-squared distributed, when the noise process is gaussian, it seems to be natural to tackle the problem with generalised linear models. But in the case of harmonic processes the ordinates at the relevant frequencies typically show large non-centrality parameters and therefore a normal approximation may be an alternative.

Prior information about the error distribution, parameter estimates and the link function is needed to construct an optimal experimental design for a generalised linear model. Therefore it is of interest to assess the loss realised by using a normality assumption in the construction of the experimental design. This possible loss is investigated in a simulation study. The experimental design of the simulation study itself is chosen to span a wide range of possible situations.

1 Introduction

This study was motivated by a project aimed at the modeling of the BTA-deep-hole-drilling process. In the analysis of on-line measurements of the

*e-mail: theis@statistik.uni-dortmund.de

boring moment it turned out that the process is mainly dominated by several eigen-frequencies. So it is of interest in which way influencing factors effect these eigen-frequencies. The amplitudes of these eigen-frequencies change during the process, because the damping effects in the system change with the depth of the hole. To get a starting point for a further analysis of this process the simpler question of the modeling of static influences on amplitudes is considered. This is a necessary preliminary step to go on to the more complex question of time-varying influences or time-variation in the response.

Looking at the periodogram ordinates as estimators of the amplitude of frequencies in a harmonic process and then model the effect of influences on these periodogram ordinates seems to be a natural way of dimension reduction in this problem. When looking at the distribution of the periodogram ordinates of harmonic processes a normal approximation seemed to be quite promising. If this approximation holds, it makes the use of known optimal designs as a basis for designs for the models on periodogram ordinates possible.

The paper is organised as follows: First the model is considered in some detail, secondly the construction of the simulations to explore the goodness of the normal approximation in a variety of settings is presented. Finally results from this simulation are presented and conclusions for the next steps are drawn and those next steps are outlined.

2 Models on Periodogram ordinates

As pointed out in the introduction the main focus of this work are harmonic processes – or processes resulting in observed processes which resemble such harmonic processes. So the basic model for the time series in this context is the following:

$$G_t(\vec{x}) = \sum_{k=1}^{K} g_k(\vec{x}) \cos(2\pi f_k t + \varphi_k) + \varepsilon_t, \tag{i}$$

where $g_k : \mathbb{R}^d \longrightarrow \mathbb{R}$ are functions connecting the influencing parameters $\vec{x}$ to the amplitudes at the relevant frequencies $f_k \in (0, \pi/2)$, φ_k are the corresponding phases and finally $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)\, \forall t$. The aim is to estimate the functions $g_k : \mathbb{R}^d \longrightarrow \mathbb{R}; k = 1, \ldots, K$.

When applying the periodogram to this model the result is the following:

$$I_{G_t(\vec{x})}(f) = \begin{cases} n(|F_\varepsilon(f) + B|)^2 & \text{for } f \neq f_k\,,\ k = 1, \ldots, K \\ n(|g_k(\vec{x}) + F_\varepsilon(f)|)^2 & \text{for } f = f_k\,,\ k = 1, \ldots, K \end{cases}, \tag{ii}$$

where $B \neq 0$ is only true for frequencies near to one of the f_k, $k = 1, \ldots, K$, and $|\cdot|$ denotes the complex absolute value and F the finite Fourier transform.

Since F is a linear function F_ε is a normally distributed random variable and thereby it is easily seen that the periodogram ordinates are non-central χ^2-distributed. The amount of non-centrality is determined in this context by the number of observations n in the series, the closeness to the relevant frequencies f_k, and the functions g_k , $k = 1, \ldots, K$.

From this, one can derive that the expected value of the periodogram ordinates at the relevant frequencies is

$$E(I_{G_t(\vec{x})}(f)) = n(g_k(\vec{x})^2 + 2\sigma_\varepsilon^2) \text{ for } f = f_k \,,\ k = 1, \ldots, K. \qquad \text{(iii)}$$

This result uses mainly the fact that at the relevant frequencies the complex Fourier transform reduces to its real part for the harmonic process. Additionally the second additive part is χ_2^2-distributed, which is the same as an exponential distribution and so its expected value is $2\sigma_\varepsilon^2$.

Johnson et al. (1994) state several normal approximations of the non-central χ^2-distribution. All of these approximations are dependent on the value of the non-centrality parameter, which in this context depends on the value of the functions g_k , $k = 1, \ldots, K$ and the number of observations. So there is theoretical reason to use a normal approximation. The impact of this approximation is tested in the simulation study.

Furthermore, to get an estimate of the amplitude functions g_k it is necessary to transform the periodogram ordinates $I_{G_t(\vec{x})}(f)$ in the following way:

$$\sqrt{\frac{I_{G_t(\vec{x})}(f) - 2\sigma_\varepsilon^2}{n}},$$

which is a transformation leading to a normal approximation.

2.1 Estimating the variance of ε (σ_ε^2)

When trying to construct an estimator for the functions $g_k(\vec{x})$, $k = 1, \ldots, K$, it is necessary to find an estimate of σ_ε^2. If the frequencies are known, it is possible to fit a harmonic process in those frequencies at different values of the input variables $\vec{x}$ and use the least squares residuals to estimate σ_ε^2. There are two reasons against this procedure:

On the one hand one needs to know the relevant frequencies, which might not be the case, and on the other hand a very good estimate of the phase is needed to get an unbiased estimate.

So there is the need for another procedure, which does not rely on known relevant frequencies and is not as vulnerable to faulty presteps for its calculation.

Taking the special properties into account, which a harmonic process with a small number of relevant frequencies K compared to the number of observations n in each time series measured at some value of $\vec{x}$ does possess, the following procedure looks promising:

(1) Estimate the periodogram $I_{G_t(\vec{x_l})}$ for all input values x_l, $l \in 1, \dots, L$

(2) Merge all $I_{G_t(\vec{x_l})}(f)$ into one sample

(3) Calculate a robust estimator for the expected value of $I_{G_t(\vec{x_l})}(f)$, e.g. the standardized median ($med_{stand.}(X) = \frac{1}{\log(2)} med(X)$, Gather and Schultze (1999)) on the merged sample

Step 2 enlarges the data base for the robust estimate, because it is assumed that the observations at the different input values are independent and the realisations of $I_{G_t(\vec{x_l})}(f)$ for different Fourier-frequencies are independent due to the orthogonality relations of the Fourier transform (cp. e.g. Bloomfield (2000)). If $K \ll n$ and $\frac{K+K_l}{n}$, with K_l the number of frequencies influenced by leakage, is lower than the breakdown point of the robust estimator, which equals $\frac{1}{2}$ for the standardized median (Gather and Schultze, 1999), we get an estimator for $2\sigma_\varepsilon^2$.

Known frequencies

If the relevant frequencies are known, it suffices to calculate the regressions on periodogram ordinates of the corresponding Fourier-frequencies. If f_k are Fourier-frequencies themselves it suffices to do the calculations only at those frequencies. If they are not Fourier-frequencies the models should be estimated based on the amplitudes at the nearest Fourier-frequencies.

Unknown frequencies

In the case of unknown frequencies the relevant frequencies have to be estimated which can be done by using the estimate of $2\sigma_\varepsilon^2$ to test for significant frequencies in all time series from the experiments and then select those frequencies present in all of them. This expects the relevant frequencies to be present for all values of $\vec{x}$.

Estimating the models

For the evaluation of the regression models one proceedes as follows:

(1) Divide the periodogram ordinates by n

(2) substract $2\sigma_\varepsilon^2$ to eliminate the bias

(3) take the square root of the periodogram ordinate

(4) calculate the regression.

3 Simulation

In this section the simulation to evaluate the properties of the model of static influences on harmonic processes is described. The main goal was to assess these properties in a wide variety of settings. This goal was reached by constructing the simulation according to standard procedures in experimental design.

3.1 Design Considerations

When looking at the harmonic model with influenced amplitude, the following parameters determine this model:

- The functions g_k, $k = 1, \ldots, K$
- The relevant frequencies f_k, $k = 1, \ldots, K$
- The phases ϕ_k, $k = 1, \ldots, K$
- The error variance σ_ε^2

Of these parameters the phases are of no interest for the models on periodogram ordinates. For the models on periodogram ordinates the following properties are of interest:

- The length of the observed series n
- The number of relevant frequencies f_k, $k = 1, \ldots, K$
- The distance of the relevant frequencies to each other
- Whether the frequencies are Fourier frequencies or not
- The functions g_k, $k = 1, \ldots, K$
- The error variance σ_ε^2

The last two points are closely related, since they determine the signal-to-noise ratio in this model. To keep things simple for the simulation the chosen functions g_k, $k = 1, \ldots, K$, are all equal:

$$g_k(\vec{x}) := 2 + 3x_1 + 4x_2 + 0.001x_3$$

All influencing variables x_1, x_2, x_3 are set to levels 0 and 1 in a 2^3 full factorial design. Then the smallest signal is reached for $x_1 = x_2 = x_3 = 0$ and therefore the signal-to-noise ratio SNR may be calculated by

$$SNR = \frac{2}{\sigma_\varepsilon} \tag{iv}$$

For the lengths of the series powers of 2 are chosen to allow for real fast fourier transforms. The number of frequencies is varied between 1 and 5, the distance of the frequencies between 1 and 10 and for the signal-to-noise ratio values between 1.11 and 101.11 are chosen. The choice of 1.11 for the lower bound of the SNR is set to ensure that a signal is present although not prominent in the series, the chosen transformation leads then to the upper bound (see below). The property "fourier frequency" or "non-fourier frequency" is assured by selecting always as first frequency the fourier frequency $\frac{5}{n}$ and the next frequencies $\frac{5+\delta_f(k-1)}{n}$ and adding in the non-fourier case an $\frac{1}{\sqrt{2}}$. The underlying design is a sort of space-filling design, a Coffee-House-design (Müller, 2001), with 20 experiments constructed in a standard hypercube $[0,1]^5$ from which all variables are transformed to fit the requirements. The transformation functions are:

(1) For the number of frequencies: $[10x] \bmod 5 + 1$

(2) Fourier Frequencies yes/no ensured by additive factor: $\frac{[x]}{\sqrt{2}}$

(3) Distance of frequencies δ_f: $[10x] \bmod 10 + 1$

(4) Length of series n: $2^{[10x] \bmod 7 + 6}$

(5) Signal-to-Noise ratio: $x^3 * 100 + 1.11$

$[\cdot]$ denotes the step function produced by rounding to the next integer. The maximal length is dictated by the maximal available memory space of 512 MB on the used computer. All calculations were done on an Athlon 700Mhz PC running under Linux using the statistical package **R** (Ihaka and Gentleman, 1996).

3.2 Results

In the simulations it was found that in most cases the assumption of normal distribution of the regression parameters can not be rejected at a 5% level in a Shapiro-Wilk test (Shapiro et al., 1968), only 22 out of 368 estimated parameters showed a significant non-normal behaviour in 100 repetitions according to this test. So only in 6% of the considered samples some non-normality was found. At most two regression parameters were found to be distributed non-normal. By checking the situations in which such rejections of the normality assumption appeared, no recognizable pattern which parameter was affected could be found. Common to the experiments with rejections of normality in some parameters was that they had only a small number of observations, but several relevant frequencies and a low SNR.

The proposed estimator for the error variance turned out to be generally biased. The biasedness is dependent on the number of relevant frequencies and the number of observations, as expected. It seems that the amount of leakage is higher than presumed, even for only two relevant Fourier-frequencies σ_ε^2 is overestimated, as can be seen in Figure 1. Nevertheless the determination of relevant frequencies by using a significance test based on this estimate was found to be very efficient. In all situations at least all relevant frequencies were found and the number of additionally considered relevant frequencies was low. So the over-estimation allowed only very strong leakage frequencies to appear relevant.

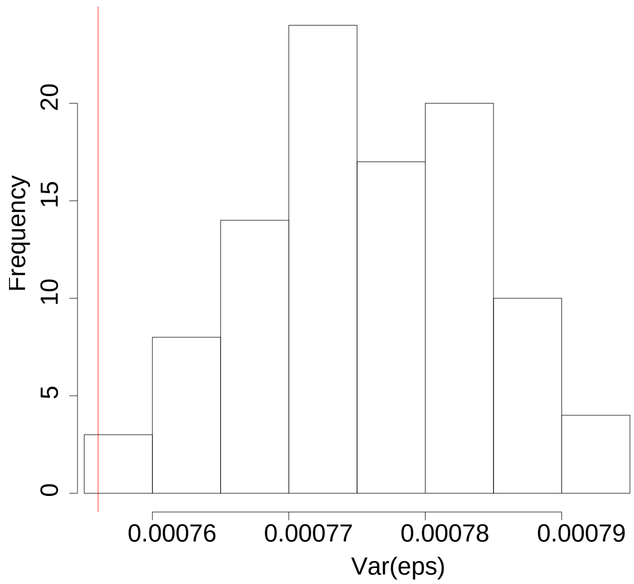

Figure 1: Estimated σ_ε^2 and true value, depicted by vertical line.

Looking at the estimated parameter values, it turned out that they are distributed around half the set value, which is correct, because the used design leads to the estimation of half-effects. Only a slight underestimation of the parameters is observable, which is due to the overestimation of σ_{ε}^2. Figure 2 shows as example the results from an experiment with only one relevant Fourier-frequence and the highest possible SNR. But at least the significance of the regression parameters is not heavily influenced by this finding, in all situations the parameter for variable x_3 was found insignificant.

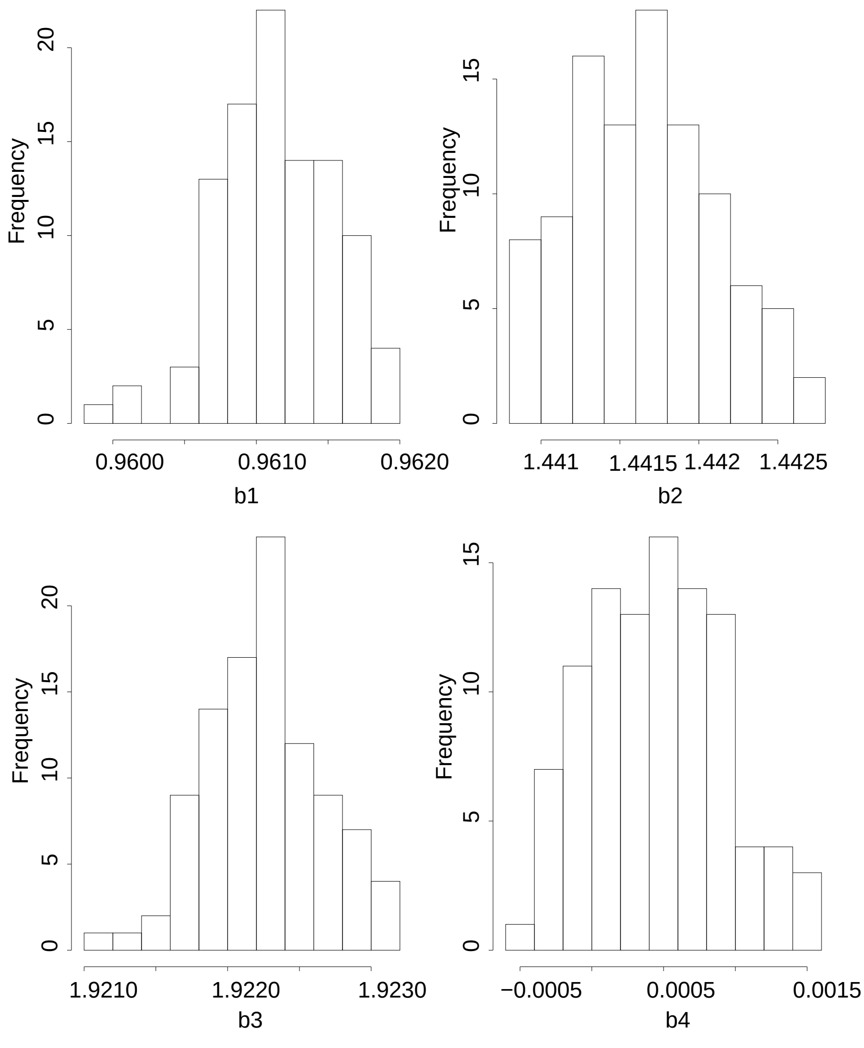

Figure 2: Histograms of parameter estimates for the linear model.

4 Conclusion

The simulation study showed that the usage of a normality approximation is justified in all situations, but dangerous when many relevant frequencies are present and the number of observtions is low. Furthermore, the determination of the relevant frequencies by using the adjusted median as an estimate for the error variance in the time series turned out to be very effective, although it led in all situations to overestimation.

As consequence for the design of experiments for the discussed model it can be stated, that the number of observations should be chosen as large as possible. On the one hand this ensures that the relevant frequencies are near to some of the fourier frequencies in the periodogram, on the other hand it enlarges the non-centrality parameter in the non-central χ^2-distribution and thereby improves the correctness of the normality approximation. Furthermore, a large number of observations work against the overestimation of σ_ε^2.

The next question to tackle will be the impact of the usage of filters in the frequency domain on the models. In this way the estimates could possibly be improved and also the approximations should become even more appropriate.

With the results of this simulation study it is now reasonable to explore the question of time-varying models on periodogram ordinates by using the normal approximation and known results from repeated measurements. Starting from these results the goal is to construct experimental designs for these models.

Acknowledgements

This work has been supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).

References

Peter Bloomfield. *Fourier Analysis of Time Series*. John Wiley & sons, New York, 2 edition, 2000.

U. Gather and V. Schultze. Robust estimation of scale of an exponential distribution. *Statistica Neerlandica*, 53(3):327–341, 1999.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*, volume 1,2. Wiley, 1994.

W. G. Müller. Coffee-house designs. In A. Atkinson et al., editor, *Optimum Design 2000*, pages 241–248, 2001.

S.S. Shapiro, M.B. Wilk, and M.J. Chen. A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63: 1343–1372, 1968.