RC24340 REVISED (W0802-035) February 8, 2008
Mathematics

IBM Research Report

On Asymptotic Normality of Nonlinear Least Squares for Sinusoidal Parameter Estimation

Ta-Hsin Li
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598
(thl@us.ibm.com)

Kai-Sheng Song
Department of Mathematics
University of North Texas
Denton, TX 76203-1430
(ksong@unt.edu)

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Abstract

This paper revisits the asymptotic normality of the nonlinear least-squares estimator for sinusoidal parameter estimation and fills two voids in the literature. First, it provides a complete proof of the asymptotic normality of the nonlinear least-squares estimator for sinusoidal signals in additive non-Gaussian white noise. Second, it uncovers the necessity of re-interpreting and re-defining the signal-to-noise ratio when applying the asymptotic theory to practical situations where the sample sizes are finite and the noise distribution has heavy tails. Simulation results are given to demonstrate the findings.

1 Introduction

Consider the problem of estimating the parameter $\boldsymbol{\theta} := [A_1, B_1, \omega_1, \ldots, A_p, B_p, \omega_p]^T \in \Theta_0 := (\mathbb{R} \times \mathbb{R} \times \Omega)^p$ from a data record $\{y_1, \ldots, y_n\}$ of length n that can be modeled as

$$y_t = \sum_{k=1}^{p} \{A_k \cos(\omega_k t) + B_k \sin(\omega_k t)\} + \varepsilon_t \quad (t = 1, \ldots, n) \tag{1}$$

where $p > 0$ is a known integer, $A_k \in \mathbb{R}$, $B_k \in \mathbb{R}$, and $\omega_k \in \Omega := (0, \pi)$, satisfying $\omega_k \neq \omega_{k'}$ for $k \neq k'$, are unknown constants, and $\{\varepsilon_t\}$ is a white noise process with mean zero and unknown variance $\sigma^2 > 0$. It is commonly believed that the nonlinear least-squares (NLS) estimator of $\boldsymbol{\theta}$, defined as a minimizer of

$$\ell_2(\boldsymbol{\vartheta}) := \sum_{t=1}^{n} \left| y_t - \sum_{k=1}^{p} \{\vartheta_{3k-2} \cos(\vartheta_{3k} t) + \vartheta_{3k-1} \sin(\vartheta_{3k} t)\} \right|^2 \tag{2}$$

with $\boldsymbol{\vartheta} := [\vartheta_1, \ldots, \vartheta_{3p}]^T$, has an asymptotic normal distribution with mean equal to $\boldsymbol{\theta}$ and covariance matrix equal to the Cramér Rao lower bound (CRLB) under the Gaussian white noise (GWN) assumption, which we denote by $\mathrm{CRLB}(\boldsymbol{\theta})$. In other words, $\hat{\boldsymbol{\theta}}_n := \arg\min \ell_2(\boldsymbol{\vartheta}) \overset{A}{\sim} N(\boldsymbol{\theta}, \mathrm{CRLB}(\boldsymbol{\theta}))$, where $\overset{A}{\sim}$ means "asymptotically distributed as."

This asymptotic normality (AN) assertion is believed to be true in the case of GWN because the NLS estimator then becomes the maximum likelihood (ML) estimator which typically has an asymptotic normal distribution with the CRLB as its asymptotic variance. For non-Gaussian noise, the AN assertion is largely based on an "approximation" argument originated from [1] and [2] because the well-known results for the general problem of nonlinear least-squares regression, such as [3] and [4], do not directly apply. In this argument, the NLS objective function (2) is first approximated by a new objective function of which the minimization leads to a periodogram maximizer for the frequency estimation and the discrete Fourier transform for the amplitude estimation (with the estimated frequencies in place of the true frequencies). From this approximation one tends to conclude that the NLS estimator should have the same asymptotic distribution as the minimizer of the new objective function which was proved to be asymptotically normal with the asserted mean and covariance matrix for any finite-variance white noise. However, this leap of conclusion is not automatically valid. In fact, the results in [1] were presented more as a justification of the periodogram maximizer than as an analysis of NLS itself.

The asymptotic variance of the NLS estimator was derived in [5] and shown to coincide with CRLB$(\boldsymbol{\theta})$ for any finite-variance white noise. It was done by using the standard technique of Taylor series expansion of the normal equations, but all higher-order terms were simply ignored in effect. Existence or nonexistence of a limiting distribution was not discussed. A proof of asymptotic normality was provided in [6] for complex sinusoids but under the GWN assumption rather than non-Gaussian conditions. An attempt was made more recently in [7] to give a direct proof of the AN assertion of the NLS estimator under (1) and non-Gaussian conditions. As in [3] and [4], the approach taken by [7] is the traditional two-step approach by which one first establishes consistency and then proves asymptotic normality using the Taylor expansion of the gradient function which equals zero at the minimizer. Unfortunately, the argument for both steps is flawed. See [8] for more detailed comments on [7] and related literature.

In this article, we provide a complete proof of the asymptotic normality of the NLS estimator by working directly with the NLS objective function and by taking a modern approach rooted in the analysis of local asymptotic normality (LAN) [9]. In addition, we point out a potential difficulty in applying the asymptotic theory to practical situations where the sample size is finite and the noise has heavy tails. This issue is largely absent in the literature because the noise used in most simulation studies that compare simulated results with the CRLB, including more recent ones [10]–[13], is invariably Gaussian. Our findings suggest that to ensure the validity of the theoretical results the notion of signal-to-noise ratio (SNR) in the CRLB should be more carefully interpreted for heavy-tailed but finite-variance noise, and that the sample variance instead of the theoretical variance should be used in these situations when evaluating the performance of the NLS estimator against the asymptotic theory.

2 Proof of Asymptotic Normality

Because of the presence of numerous local minima in the NLS objective function, few algorithms of practical value can produce the global minimizer of $\ell_2(\boldsymbol{\vartheta})$. Practical algorithms usually begin with an initial value, based on prior knowledge or an initialization procedure, and then find a local minimizer nearby as the final estimate. This method often lead to satisfactory results as long as the initial values are sufficiently close to the true parameter value. Therefore, it sufficies to consider the minimizer of $\ell_2(\boldsymbol{\vartheta})$ in a neighborhood of $\boldsymbol{\theta}$ and study the properties of this local minimizer. Critical to this line of inquiry is that the size of the neighborhood should be specified as precisely as possible in order to guide the practitioner in choosing the right initial values or initialization procedures. Our main theorem serves these purposes.

We define the NLS estimator as the minimizer of $\ell_2(\boldsymbol{\vartheta})$ in a closed neighborhood of $\boldsymbol{\theta}$ that shrinks at a certain rate as the sample size grows. The existence of the minimizer is guaranteed by the continuity of $\ell_2(\boldsymbol{\vartheta})$ as a function of $\boldsymbol{\vartheta}$. The shrinking neighborhood circumvents the consistency issue without requiring the minimizer to lie in the interior of the neighborhood or to have a zero gradient. It is not a limitation of our approach but a basic requirement for NLS to overcome the problem of spurious local extrema [5] [18]. Our theorem explicitly specifies the minimal rate of shrinkage and thus helps the practitioner to choose the right initialization procedures. For example, according to the theorem, a frequency estimator of accuracy $\mathcal{O}(n^{-1/2})$ is not good enough, but a frequency estimator of accuracy $\mathcal{o}(n^{-11/8})$ is. Note that the rate of shrinkage could be improved slightly (e.g., by assuming the existence of higher moments for the noise), but it must be faster than $\mathcal{O}(n^{-1})$ for the frequency parameter; otherwise, one would face the problem of spurious local extrema as demonstrated by many analytical and numerical studies [5] [18].

Theorem 1. *Let $\{y_t\}$ satisfy (1), where $\{\varepsilon_t\}$ is a white noise process with mean zero and finite variance $\sigma^2 > 0$. Let $\Theta_n := \{\boldsymbol{\vartheta} \in \Theta_0 : \|D_n^{-1}(\boldsymbol{\vartheta} - \boldsymbol{\theta})\| \le cn^{\alpha}\}$ for some constants $\alpha \in (0, \frac{1}{8})$ and $c > 0$, where D_n be a block-diagonal matrix of p blocks with each diagonal block equal to* $\mathrm{diag}\{n^{-1/2}, n^{-1/2}, n^{-3/2}\}$. *Then,*

$\hat{\boldsymbol{\theta}}_n := \arg\min\{\ell_2(\boldsymbol{\vartheta}) : \boldsymbol{\vartheta} \in \Theta_n\} \overset{A}{\sim} N(\boldsymbol{\theta}, CRLB(\boldsymbol{\theta}))$ *as* $n \to \infty$*, where* $CRLB(\boldsymbol{\theta})$ *is a block-diagonal matrix of* p *blocks with the kth diagonal block taking the form*

$$\Sigma_k := \frac{1}{\gamma_k} \begin{bmatrix} (A_k^2 + 4B_k^2)/n & -3A_kB_k/n & -6B_k/n^2 \\ & (4A_k^2 + B_k^2)/n & 6A_k/n^2 \\ \mathit{symmetry} & & 12/n^3 \end{bmatrix},$$

and with $\gamma_k := \frac{1}{2}(A_k^2 + B_k^2)/\sigma^2$ *being the SNR of the kth sinusoid.*

Proof. Let $s_t(\boldsymbol{\vartheta}) := \sum_{k=1}^{p}\{\vartheta_{3k-2}\cos(\vartheta_{3k}t) + \vartheta_{3k-1}\sin(\vartheta_{3k}t)\}$ and consider, according to the LAN approach, the random function $Z_n(\boldsymbol{\delta}) := (2\sigma^2)^{-1}\sum_{t=1}^{n}\{|y_t - s_t(\boldsymbol{\theta} + D_n\boldsymbol{\delta})|^2 - |\varepsilon_t|^2\}$. The objective is to prove that $Z_n(\boldsymbol{\delta})$ can be expressed as

$$Z_n(\boldsymbol{\delta}) = -\boldsymbol{\delta}^T\boldsymbol{\zeta}_n + \tfrac{1}{2}\boldsymbol{\delta}^T\tilde{I}_G(\boldsymbol{\theta})\boldsymbol{\delta} + R_n(\boldsymbol{\delta}), \tag{3}$$

where $R_n(\boldsymbol{\delta}) = \mathcal{O}_p(n^{-1/4+2\alpha})$ uniformly in $\boldsymbol{\delta} \in \Delta_n := \{\boldsymbol{\delta} : \boldsymbol{\theta} + D_n\boldsymbol{\delta} \in \Theta_0, \|\boldsymbol{\delta}\| \le cn^{\alpha}\} = \{\boldsymbol{\delta} : \boldsymbol{\theta} + D_n\boldsymbol{\delta} \in \Theta_n\}$ and $\boldsymbol{\zeta}_n \xrightarrow{D} N(\mathbf{0}, \tilde{I}_G(\boldsymbol{\theta}))$, with $\tilde{I}_G(\boldsymbol{\theta}) := \boldsymbol{D}_n\mathrm{CRLB}^{-1}(\boldsymbol{\theta})\boldsymbol{D}_n$, which does not depend on n, being the normalized asymptotic Fisher information matrix under the GWN assumption. Note that unlike some linear LAN problems for which it suffices to establish the uniformity of an expression analogous to (3) in a fixed compact set of $\boldsymbol{\delta}$, the nonlinear problem under consideration commands a greater demand that calls for the uniformity of (3) to hold in a compact set Δ_n which grows to the inifinite space $\mathbb{R}^{3p}$. The rate at which Δ_n is permitted to grow ultimately determines the required accuracy of initial values.

The quadratic function of $\boldsymbol{\delta}$ in (3) has a unique minimum $\tilde{\boldsymbol{\delta}}_n := \tilde{I}_G^{-1}(\boldsymbol{\theta})\boldsymbol{\zeta}_n$. The assertion of the theorem follows immediately if we can show that $\hat{\boldsymbol{\delta}}_n := \arg\min\{Z_n(\boldsymbol{\delta}) : \boldsymbol{\delta} \in \Delta_n\} = D_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ is $\mathcal{O}_p(1)$ away from $\tilde{\boldsymbol{\delta}}_n$, i.e., $\hat{\boldsymbol{\delta}}_n - \tilde{\boldsymbol{\delta}}_n \xrightarrow{P} 0$. Toward that end, we rewrite (3) as $Z_n(\boldsymbol{\delta}) = Z_n(\tilde{\boldsymbol{\delta}}_n) + \frac{1}{2}(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}_n)^T\tilde{I}_G(\boldsymbol{\theta})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}_n) + R_n(\boldsymbol{\delta}) - R_n(\tilde{\boldsymbol{\delta}}_n)$ and define $R_n := \max\{|R_n(\boldsymbol{\delta})| : \boldsymbol{\delta} \in \Delta_n\}$. For any constant $\mu > 0$, if $\tilde{\boldsymbol{\delta}}_n \in \Delta_n$, then $\inf\{Z_n(\boldsymbol{\delta}) : \boldsymbol{\delta} \in \Delta_n, \|\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}_n\| > \mu\} \ge Z_n(\tilde{\boldsymbol{\delta}}_n) + \frac{1}{2}\kappa\mu^2 - 2R_n$, where $\kappa > 0$ is the smallest eigenvalue of $\tilde{I}_G(\boldsymbol{\theta})$. Since $\tilde{\boldsymbol{\delta}}_n$ converges in distribution and $\Delta_n \to \mathbb{R}^{3p}$, we have $P(\tilde{\boldsymbol{\delta}}_n \notin \Delta_n) \to 0$. This, combined with $R_n \xrightarrow{P} 0$, implies $P(\|\hat{\boldsymbol{\delta}}_n - \tilde{\boldsymbol{\delta}}_n\| > \mu, \tilde{\boldsymbol{\delta}}_n \in \Delta_n) \to 0$, which, in turn, leads to $P(\|\hat{\boldsymbol{\delta}}_n - \tilde{\boldsymbol{\delta}}_n\| > \mu) \to 0$.

To establish (3), we use the Taylor expansion $s_t(\boldsymbol{\theta}+\boldsymbol{D}_n\boldsymbol{\delta}) = s_t(\boldsymbol{\theta}) + v_t + \tilde{r}_t$, where $v_t := (\boldsymbol{D}_n\boldsymbol{\delta})^T \boldsymbol{g}_t(\boldsymbol{\theta})$ and $\tilde{r}_t := \frac{1}{2}(\boldsymbol{D}_n\boldsymbol{\delta})^T \boldsymbol{H}_t(\boldsymbol{\theta}_{nt})\,(\boldsymbol{D}_n\boldsymbol{\delta})$, with $\boldsymbol{g}_t(\boldsymbol{\vartheta})$ being the gradient vector, $\boldsymbol{H}_t(\boldsymbol{\vartheta})$ the Hessian matrix, and $\boldsymbol{\theta}_{nt}$ an intermediate point between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}+\boldsymbol{D}_n\boldsymbol{\delta}$. Since $y_t = s_t(\boldsymbol{\theta}) + \varepsilon_t$, it follows that $Z_n(\boldsymbol{\delta}) = \sum_{i=1}^{5} Z_{ni}$, where $Z_{n1} := -\sum v_t \varepsilon_t/\sigma^2$, $Z_{n2} := \frac{1}{2}\sum v_t^2/\sigma^2$, $Z_{n3} := -\sum \tilde{r}_t \varepsilon_t/\sigma^2$, $Z_{n4} := \sum v_t \tilde{r}_t/\sigma^2$, and $Z_{n5} := \frac{1}{2}\sum \tilde{r}_t^2/\sigma^2$.

First, we show $Z_{n4} = \mathcal{O}(n^{-1/2+3\alpha})$ uniformly in $\boldsymbol{\delta} \in \Delta_n$ by using the Taylor expansion $\tilde{r}_t = r_t + u_t$, where r_t is defined in the same way as $\tilde{r}_t$ except that the true parameter value $\boldsymbol{\theta}$ is in place of the intermediate point $\boldsymbol{\theta}_{nt}$. In this expansion, u_t is a linear combination of the third partial derivatives of $s_t(\boldsymbol{\vartheta})$ evaluated at an intermediate point which may depend on t. These third partial derivatives are either zero (for those that involve differentiation with respect to the amplitudes more than once) or can be expressed as $\mathcal{O}(t^\beta)$ for some $\beta = 2,3$. Moreover, owing to the presence of $\boldsymbol{D}_n\boldsymbol{\delta}$ and to the fact that the components in $\boldsymbol{\theta}_{nt} - \boldsymbol{\theta}$ that correspond to the amplitudes can be expressed as $\mathcal{O}(n^{-1/2+\alpha})$ and those that correspond to the frequencies as $\mathcal{O}(n^{-3/2+\alpha})$, the coefficients of the third partial derivatives of the form $\mathcal{O}(t^\beta)$ take the form $\mathcal{O}(n^{-(\beta+1)-1/2+3\alpha})$. This, combined with $v_t = \mathcal{O}(n^{-1/2+\alpha})$, leads to $\sum v_t u_t = \mathcal{O}(n^{-1+4\alpha})$. Direct calculation also shows that $\sum v_t r_t = \mathcal{O}(n^{-1/2+3\alpha})$. This proves $Z_{n4} = \mathcal{O}(n^{-1/2+3\alpha})$. Similarly, we obtain $Z_{n5} = \mathcal{O}(n^{-1+4\alpha})$ because $\sum r_t^2 = \mathcal{O}(n^{-1+4\alpha})$, $\sum r_t u_t = \mathcal{O}(n^{-3/2+5\alpha})$, and $\sum u_t^2 = \mathcal{O}(n^{-2+6\alpha})$. Furthermore, straightforward calculation yields $Z_{n2} = \frac{1}{2}\boldsymbol{\delta}^T\{\tilde{I}_G(\boldsymbol{\theta}) + \mathcal{O}(n^{-1})\}\,\boldsymbol{\delta} = \frac{1}{2}\boldsymbol{\delta}^T \tilde{I}_G(\boldsymbol{\theta})\,\boldsymbol{\delta} + \mathcal{O}(n^{-1+2\alpha})$.

Next, we show $Z_{n3} = \mathcal{O}_p(n^{-1/4+2\alpha})$ uniformly in $\boldsymbol{\delta} \in \Delta_n$. Toward that end, we note that $Z_{n3} = -(\sum r_t \varepsilon_t + \sum u_t \varepsilon_t)/\sigma^2$. Because the expected value of $n^{-(\beta+1)-1/2+3\alpha}\sum t^\beta |\varepsilon_t|$ takes the form $\mathcal{O}(n^{-1/2+3\alpha})$, we obtain $\max|\sum u_t \varepsilon_t| = \mathcal{O}_p(n^{-1/2+3\alpha})$, where the max is over $\boldsymbol{\delta} \in \Delta_n$. Moreover, $|\sum r_t \varepsilon_t|$ can be upper bounded uniformly in $\boldsymbol{\delta} \in \Delta_n$ by a linear combination of $n^{-(\beta+1)+2\alpha}|\sum t^\beta \varepsilon_t \exp(i\omega_k t)|$ for $\beta = 1,2$, where $i := \sqrt{-1}$. It is easy to see that $|\sum t^\beta \varepsilon_t \exp(i\omega_k t)|^2$ does not exceed

$$\sum_{t=1}^{n} t^{2\beta}\varepsilon_t^2 + 2\sum_{s=1}^{n-1}\left|\sum_{t=1}^{n-s} t^\beta (t+s)^\beta \varepsilon_t \varepsilon_{t+s}\right|.$$

The expected value of the first term can be expressed as $\mathcal{O}(n^{2\beta+1})$; the expected value of the second term can be bounded, using the Cauchy-Schwarz inequality, by

$$2\sum_{s=1}^{n-1}\left\{E\left(\sum_{t=1}^{n-s}t^{\beta}(t+s)^{\beta}\varepsilon_t\varepsilon_{t+s}\right)^2\right\}^{1/2},$$

which, in turn, takes the form $\mathcal{O}(n^{2\beta+3/2})$. This proves that $n^{-(\beta+1)+2\alpha}|\sum t^{\beta}\varepsilon_t\exp(i\omega t)| = \mathcal{O}_p(n^{-1/4+2\alpha})$ uniformly in $\omega\in\mathbb{R}$. Therefore, $\max|\sum r_t\varepsilon_t| = \mathcal{O}_p(n^{-1/4+2\alpha})$, and the assertion follows.

Finally, let $\boldsymbol{\zeta}_n := D_n\sum\boldsymbol{g}_t(\boldsymbol{\theta})\varepsilon_t/\sigma^2$. It remains to show that $Z_{n1} = -\sum v_t\varepsilon_t/\sigma^2 = -\boldsymbol{\delta}^T\boldsymbol{\zeta}_n \xrightarrow{D} -\boldsymbol{\delta}^T\boldsymbol{\zeta}$ for any fixed $\boldsymbol{\delta}\neq\mathbf{0}$, where $\boldsymbol{\zeta}\sim N(\mathbf{0},\tilde{I}_G(\boldsymbol{\theta}))$. We follow the steps in [1]. First, we note that $E(\sum v_t\varepsilon_t) = 0$ and $\sigma_n^2 := Var(\sum v_t\varepsilon_t) = \sigma^2\sum v_t^2 \to \sigma^4\boldsymbol{\delta}^T\tilde{I}_G(\boldsymbol{\theta})\,\boldsymbol{\delta}$. Therefore, it suffices to verify the Lindeberg condition

$$\sum_{t=1}^{n}E\{(v_t\varepsilon_t/\sigma_n)^2I(|v_t\varepsilon_t|/\sigma_n > a)\}\to 0 \tag{4}$$

for any $a>0$. This can be done by showing $c_n^2 := \max(v_t^2)/\sum v_t^2 \to 0$. Indeed, because the left hand side of (4) can be written as $\sum(v_t/\sigma_n)^2E\{\varepsilon_t^2I(|\varepsilon_t| > a\sigma_n/|v_t|)\}$, and because $\sigma_n/|v_t|\geq\sigma/c_n$ so that $E\{\varepsilon_t^2I(|\varepsilon_t| > a\sigma_n/|v_t|)\}\leq E\{\varepsilon_t^2I(|\varepsilon_t| > a\sigma/c_n)\} = e_n := E\{\varepsilon_1^2I(|\varepsilon_1| > a\sigma/c_n)\}$, where the last expression is due to the identical distribution of ε_t, the quantity in the left hand side of (4) is upper bounded by $\sum(v_t/\sigma_n)^2e_n = e_n/\sigma^2$. If $c_n\to 0$, then $e_n\to 0$, hence the Lindeberg condition. That $c_n\to 0$ follows from the fact that $\max(|v_t|) = \mathcal{O}(n^{-1/2})$ and $\sum v_t^2\to\sigma^2\boldsymbol{\delta}^T\tilde{I}_G(\boldsymbol{\theta})\,\boldsymbol{\delta}>0$. □

3 Simulation

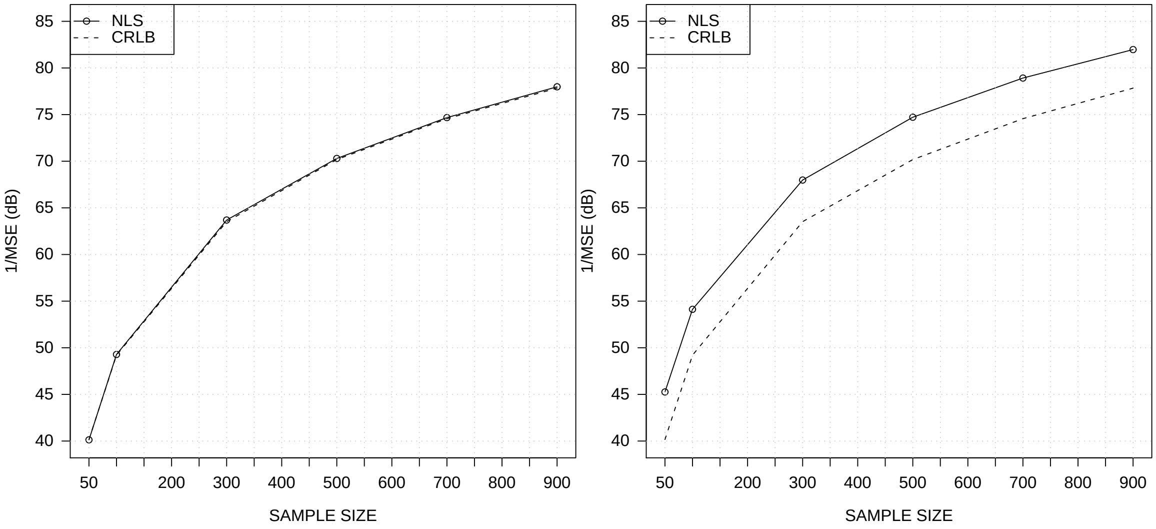
There are simulation studies in the literature that compare the NLS estimator against the CRLB under the GWN conditions. These studies show that the simulated mean-squared error (MSE) of the estimator closely follows the CRLB suggested by the asymptotic theory, even for relatively small sample sizes. In this section, we are interested in a similar comparison but under non-Gaussian conditions, especially under the condition of heavy-tailed noise. This is different from the estimation issues discussed in [14]–[16] under impulsive noise with infinite variance and the robustness issues discussed in [17].

For illustration purposes, consider the case of $p = 1$ where $y_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon_t$. According to Theorem 1, the NLS frequency estimator, $\hat{\omega}_n$, should be asymptotically distributed as $N(\omega, 12\gamma^{-1}n^{-3})$, where $\gamma := \frac{1}{2}(A^2+B^2)/\sigma^2$ is the SNR. More to the point, Theorem 1 promises that the asymptotic distribution is valid for any zero-mean white noise with heavy or light tails, as long as the noise has a finite variance. In particular, Theorem 1 promises that the MSE of $\hat{\omega}_n$ should be closely approximated by the CRLB, which equals $12\gamma^{-1}n^{-3}$, at least for large sample sizes. The purpose of our simulation study in this section is to find out how closely the MSE follows the CRLB under the condition of heavy-tailed noise.

To facilitate the study, we consider the family of Student's T distributions, denoted as T_ν, because the heaviness of its tails can be easily controlled by the degree-of-freedom parameter ν. Recall that T_ν has a finite fourth moment only if $\nu > 4$, a finite third moment only if $\nu > 3$, and a finite second moment, with variance $\nu/(\nu-2)$, only if $\nu > 2$. In general, the lower the order of the nonexistent moments the heavier the tails would become. Therefore, by varying ν, we can simulate noise data with a range of tail behavior.

The first simulation sets $\nu = 4.1$, so the noise does not have very heavy tails. For a given SNR γ, the noise data $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are generated by scaling a sample of i.i.d. T_ν-distributed random variables $\{x_1, \ldots, x_n\}$ such that $\varepsilon_t = c x_t/\sqrt{\nu/(\nu-2)}$, where $c := \sqrt{\frac{1}{2}(A^2+B^2)/\gamma}$. In theory, the variance of so-generated ε_t is $\sigma^2 = c^2$ and, according to Theorem 1, the MSE of $\hat{\omega}_n$ is expected to be close to $12\gamma^{-1}n^{-3} = 24n^{-3}\sigma^2/(A^2+B^2)$. In the simulation, we choose $A = 1$, $B = 0$, $\omega = 0.15 \times 2\pi$, and $\gamma = 1$ (SNR = 0 dB). The NLS estimates are calculated by the function `optim` of the software package R using a simplex algorithm. To avoid the problem of spurious local minima, we use the true parameter value as the initial guess. In practice, the initial value can be obtained from other estimators [1] [5] [11] [19], but this subject is not our focus here.

Fig. 1(a) shows the result of the first simulation where $1/\text{MSE}$ is plotted against the sample size, with MSE calculated from 10,000 independent Monte Carlo runs. The dotted line depicts $1/\text{CRLB}$. As we can see, in this case, the MSE follows the CRLB very closely, even for sample size 50. This result resembles the results in the literature under the GWN conditions.



(a)

(b)

Figure 1: Reciprocal MSE (in decibels) of the NLS estimates $\hat{\omega}_n/(2\pi)$ as a function of n. Dotted line represents the reciprocal CRLB. The noise is distributed as T_ν with (a) $\nu = 4.1$ and (b) $\nu = 2.1$.

In the next simulation, we reduce the degrees of freedom to $\nu = 2.1$, so the noise becomes very heavy tailed (third moment does not exist). Fig. 1(b) depicts the resulting MSE in the same way as Fig. 1(a). This experiment reveals that the MSE is no longer approximated well by the CRLB, even for sample sizes as large as 900. In fact, the simulated MSE is about 5 dB smaller than the CRLB predicted by Theorem 1. It is important to note that the noise distribution does not violate the assumptions in Theorem 1, because the noise remains to have a finite variance with $\nu = 2.1$. The only difference from the previous simulation is that the noise no longer possesses finite third and fourth moments. Clearly, in this case, the sample size must be extremely large in order for Theorem 1 to be relevant. (We tested it using sample sizes up to 20,000 and still observed a 2.8 dB discrepancy.)

To explain the huge discrepancy revealed in Fig. 1(b), one has to take a closer look at the proof of Theorem 1. As can be seen, the theoretical variance σ^2 plays a crucial role in calculating the asymptotic variance of the sum $\sum v_t \varepsilon_t$ that ultimately determines the asymptotic distribution of the NLS estimator. The

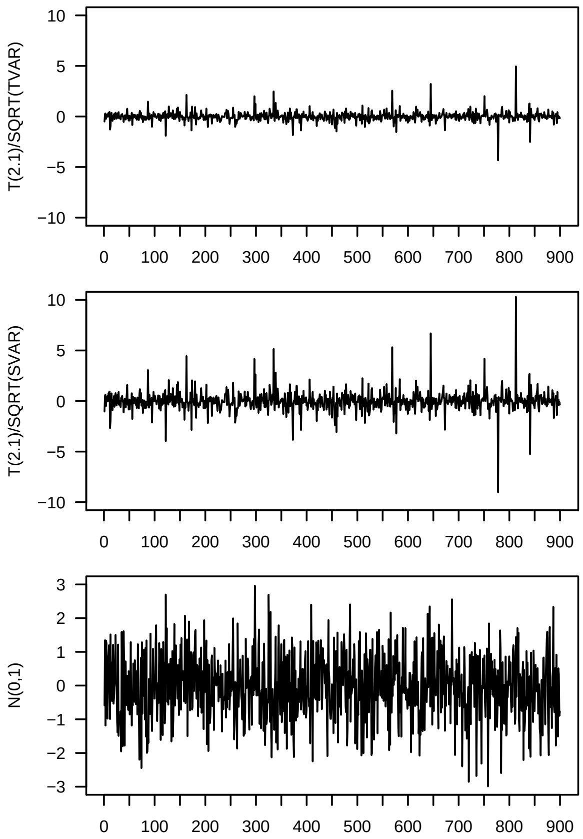


Figure 2: A sample of T_ν-distributed random noise with $\nu = 2.1$, scaled by theoretical variance (top) and by sample variance (middle). The bottom panel shows a random sample from $N(0,1)$ as reference.

variance σ^2 is interpreted as the power of the noise, but is calculated as an ensemble average of ε_t^2 over the imaginary statistical population or an infinite data record. For finite sample sizes, a more meaningful measure of the noise power is the sample average $s_n^2 := n^{-1}\sum \varepsilon_t^2$ of the noise data $\{\varepsilon_1, \ldots, \varepsilon_n\}$. Typically, σ^2 and s_n^2 do not differ very much (at least for reasonably large sample sizes) when the noise distribution behaves well. For example, if the noise has finite fourth moments, s_n^2 approaches σ^2 as $n \to \infty$. This is true in particular for the widely studied GWN cases. But, when the noise has heavy tails, σ^2 is no longer a good measure of the power of a specific realization of the noise, whereas s_n^2 remains so. In fact, σ^2 is usually much higher than s_n^2. Therefore, using σ^2 to define the SNR tends to considerably lower the true SNR of the data which is better defined as $\frac{1}{2}(A^2 + B^2)/s_n^2$. This explains why the CRLB, defined using σ^2, is much greater than the simulated MSE in Fig. 1(b).

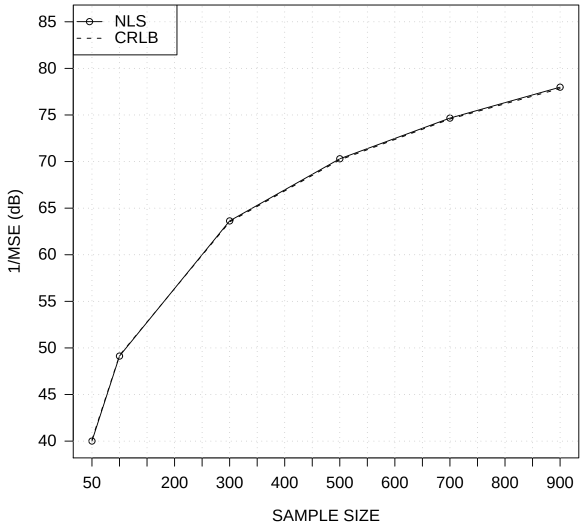


Figure 3: Same as Fig. 1(b) except the noise is scaled by sample variance instead of theoretical variance.

With this in mind, we replace the theoretical variance of the T_v variables with their sample variance in scaling them to obtain ε_t. This scaling technique ensures that the sample SNR is always equal to the design value γ for each realization of the noise. Its effect, as demonstrated in Fig. 2, is a properly boosted noise power. By repeating the experiment shown in Fig. 1(b) with the new scaling method, we find that the resulting MSE, depicted in Fig. 3, becomes well approximated by the CRLB, so Theorem 1 becomes valid again for these sample sizes. To close this section, we must point out that the intended use of the scaling technique is not for estimating the noise variance from the observed data $\{y_t\}$ but for simulation studies that compare simulated results with theoretical results.

References

[1] A. M. Walker, “On the estimation of a harmonic component in a time series with stationary independent residuals,” *Biometrika*, vol. 58, no. 1, pp. 21–36, 1971.

[2] D. C. Rife and R. R. Boorstyn, “Single-tone parameter estimation from discrete-time observations,” *IEEE Trans. Inform. Theory*, vol. 20, no. 5, pp. 591–598, 1974.

[3] R. Jennrich, “Asymptotic properties of nonlinear least squares estimation,” *Ann. Statist*, vol. 40, pp. 633–643, 1969.

[4] C.-F. Wu, “Asymptotic theory on nonlinear least squares estimation,” *Ann. Statist.*, vol. 9, no. 3, pp. 501–513, 1981.

[5] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, “Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 378–392, 1989.

[6] C. R. Rao and L. C. Zhao, “Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals,” *IEEE Trans. Signal Process.*, vol. 41, no. 3, pp. 1461–1464, 1993.

[7] D. Kundu, “Asymptotic theory of the least squares estimators of sinusoidal signal,” *Statistics*, vol. 30, pp. 221–238, 1997.

[8] K. S. Song and T. H. Li, “A note on asymptotics of least squares for nonlinear harmonic regression,” IBM Tech. Report, 2008.

[9] A. W. van der Vaart, *Asymptotic Statistics*, Chapt. 7, Cambridge University Press, Cambridge, UK, 1998.

[10] S. S. Abeysekera, “Efficient frequency estimation using the pulse-pair method at various lags,” *IEEE Trans. Commun.*, vol. 54, no. 9, pp. 1542–1546, 2006.

[11] E. Aboutanios and B. Mulgrew, “Iterative frequency estimation by interpolation on Fourier coefficients,” *IEEE Trans. Signal Process.*, vol. 53, no. 4, pp. 1237–1242, 2005.

[12] T. Brown and M. Wang, “An iterative algorithm for single-frequency estimation,” *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2671–2682, 2002.

[13] H. C. So and K. W. Chan, “Approximate maximum-likelihood algorithms for two-dimensional frequency estimation of a complex sinusoid,” *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 3231–3237, 2006.

[14] J. Friedman, H. Messer, and J.-F. Cardoso, “Robust parameter estimation of a deterministic signal in impulsive noise,” *IEEE Trans. Signal Processing*, vol. 48, no. 4, pp. 935–942, 2000.

[15] S. Nandi, S. Iyer, and D. Kundu, “Estimating the frequencies in presence of heavy tail errors,” *Statistics and Probability Letters*, vol. 58, pp. 265–282, 2002.

[16] A. Swami and B. M. Sadler, “On some detection and estimation problems in heavy-tailed noise,” *Signal Processing*, vol. 82, pp. 1829–1846, 2002.

[17] G. K. Smyth and D. M. Hawkins, “Robust frequency estimation using elemental sets,” *J. Computational and Graphical Statistics*, vol. 9, pp. 196–214, 2000.

[18] J. A. Rice and M. Rosenblatt, “On frequency estimation,” *Biometrika*, vol. 75, no. 3, pp. 477–484, 1988.

[19] T. H. Li and K. S. Song, “Asymptotic analysis of a fast algorithm for efficient multiple frequency estimation,” *IEEE Trans. Inform. Theory*, vol. 48, no. 10, pp. 2709–2720, 2002. Errata: vol. 49, no. 2, p. 529, 2003.