Bayesian Frequency Estimation

Douglas E. Johnston and Petar M. Djurić*

Technical Report #633
Department of Electrical Engineering
State University of New York
Stony Brook, NY 11794

14 August 1992

Abstract

This paper addresses the problem of frequency estimation for complex exponentials, or cisoids, embedded in general Gaussian noise using a Bayesian approach. The main contribution is that the a posteriori density of the frequencies of M complex exponentials in arbitrary Gaussian noise is derived. Contrary to previous research, the density is derived for colored noise, without ignoring terms, and a valid prior density is used to marginalize the phase and amplitudes of the signals. This results in the a posteriori density which, when used for point estimation, significantly outperforms the maximum likelihood estimator both for single cisoids in colored noise and closely seperated cisoids in arbitrary noise. Along the way, a number of both useful and interesting properties of the Bayesian technique are discovered.

*Supported in part by grant #MIP-9110628 from the NSF.

1 Introduction

The estimation of the frequencies of a number of sinusoids embedded in Gaussian noise has intrigued researchers for many years. From the direction of arrival problem in radar to the removal of cyclical components in time series, frequency estimation is entrenched as an integral part of disciplines as diverse as biomedicine, economics and radar. The essence of the problem is to determine, in some defined sense, the frequencies of multiple sinusoids given a finite number of observations contaminated with additive noise.

The problem was first formulated by Slepian [16] for the continuous time model and then Rife and Boorstyn [14] expanded on those results by deriving the maximum likelihood (ML) estimates, as well as the Cramer-Rao bound, for discrete sequences. Since then, a surge of research has occurred with the introduction of autoregressive modelling techniques [17], eigenvector methods [8] and computationally efficient means of computing the ML estimates [18].

More recently, there has been an interest in using a Bayesian approach [1] to solve the parameter estimation problem. Kramer and Sorenson [9] have reported on the importance of the Bayesian approach for dynamical state estimation and Bayesian techniques have been used in the direction of arrival problem [13] to construct direction estimates in arbitrary, unknown noise though the marginalized estimates were not derived. The use of the Bayesian approach for frequency estimation was first discussed by Jaynes [5] and then later by Bretthrost [2] and most recently by Quinn [12]. The benefit of the Bayesian approach is that the a posteriori density of the parameters given the observed data is derived and this is used as the inferential apparatus. In addition to point estimation, confidence intervals can be computed and new data can be easily incorporated in a fashion that is theoretically sound. In contrast, the likelihood function, which is used in classical estimation methods, does not equate to a probability density function and therefore has limited inferential use.

An important feature of the Bayesian approach, and one of some debate, is the choice of the prior density for the parameters. As pointed out by Jaynes [6], the prior density reflects one's knowledge, or belief, about the parameters prior to observing the data and is not to be construed as a relative frequency of occurrence. In fact, Cox [3] proved that degrees of belief obey the same rules as probabilities and therefore are the driving force in scientific inference. As such, the parameters may be determinisitic as long as we interpret probability in it's proper context which results in what is known as strong Bayesian inference. When nothing is known about the parameters in advance, a noninformative prior, constructed using Jeffrey's invariance rule [7], is used which corresponds to complete ignorence. In most circumstances, this corresponds to a uniform prior which is flat over the parameter space. In many cases, the prior is chosen based on intimate knowledge of the problem at hand. For example, the prior knowledge of the phase and amplitude of a sinusoid are usually modelled as a Uniform and Rayleigh density, respectively, for obvious reasons.

The paper proceeds as follows. In Section 2, the a posteriori density of the frequency of a single cisoid embedded in general Gaussian noise is derived. This not only acquaints the reader with the Bayesian methodology but also lays the groundwork for the multiple frequency case. Using the dervived density, the point estimation for the frequency is discussed and a number of interesting properties are discovered. In Section 3, the a posteriori density for multiple frequencies is derived. This was derived independently by Quinn [12] though our two results are in different forms. Our density was derived in a recursive form which is not only appealing from a computional standpoint but also is vital in our understanding of the point estimation of the frequencies. We discuss the point estimation of the frequencies and consider pathological condition when two frequencies become arbitrarily close. We consider both the MAP estimate, which is shown to theoretically not exist, as well the MMSE estimate. In Section 4, we report on the simulations that were conducted. Using a colored noise sequence, we show the improvement in estimation capabilities using the Bayesian approach as compared to the maximum likelihood technique both for the case of an isolated cisoid as well as for two closely seperated cisoids. Finally, we draw some conclusions in Section 5 and suggest some further research.

2 Bayesian Estimation of a Single Frequency

The problem that we are concerned with in this section can be formulated as follows. Let $d(t)$ denote a complex, stochastic process which can be decomposed as a sum of a harmonic term and noise. The continuous waveform is written as

$$d(t) = \rho e^{j(\omega t + \phi)} + n(t), \tag{1}$$

where the parameters of the model, ρ, ω and ϕ, are unknown quantities and $n(t)$ is a complex, zero mean, stationary Gaussian process whose real and imaginary parts are independently and identically distributed. The continuous waveform is sampled at N distinct points, $t_k, k = 0, \cdots, N-1$, which results in the discrete sequence

$$d(t_k) = \rho e^{j(\omega t_k + \phi)} + n(t_k), \qquad k = 0, \cdots, N-1. \tag{2}$$

Based on the discrete sequence, it is our goal to infer information concerning the true value of ω. In addition to the observed data, we have at our disposal prior information concerning the parameters which will be denoted as I. The inferences concerning ω may take the form of a point estimate, a confidence interval or the complete probabilistic description of ω given the observed data and our prior information. In any case, we must determine the a posteriori density of ω given the samples, $d(t_k), k = 0, \cdots, N-1$, and our prior information, I. We can write (2) in matrix form as

$$\mathbf{d} = \rho e^{j\phi}\mathbf{a}(\omega) + \mathbf{n} \tag{3}$$

with

$$\mathbf{d}^T = [d(t_0),\ d(t_1),\ \cdots,\ d(t_{N-1})] \tag{4}$$

and

$$\mathbf{a}(\omega)^T = \left[e^{j\omega t_0},\ e^{j\omega t_1},\ \cdots,\ e^{j\omega t_{N-1}}\right]. \tag{5}$$

When viewed as a function of ω, $\mathbf{a}(\omega)$ is commonly referred to as the signal manifold and is a one dimensional continuum within C^N, complex N-space. The noise, $\mathbf{n}$, is a complex, zero mean random vector with the real and imaginary parts being independently Gaussian distributed, both having identical covariance structure, $\mathbf{R}$, which is assumed known and included in our prior knowledge, I. In addition, the matrix $\mathbf{R}$ is assumed to be nonsingular.

Our interest is in determining $f(\omega|\mathbf{d}, I)$, the a posteriori density of ω given the data and our prior information, which can then be used for inference concerning ω. To determine $f(\omega|\mathbf{d}, I)$ we must begin with the joint conditional density of ρ, ω and ϕ, and then marginalize with respect to the unwanted, nuisance parameters. That is, we wish to first determine

$$f(\rho, \omega, \phi|\mathbf{d}, I) = \frac{f(\mathbf{d}|\rho, \omega, \phi, I) f(\rho, \omega, \phi|I)}{f(\mathbf{d}|I)} \tag{6}$$

and then integrate out the nuisance parameters, ρ and ϕ. Since the term $f(\mathbf{d}|I)$ in (6) is merely a normalization factor, we can rewrite (6) as

$$f(\rho, \omega, \phi|\mathbf{d}, I) \propto f(\mathbf{d}|\rho, \omega, \phi, I)\ f(\rho, \omega, \phi|I) \tag{7}$$

where $A \propto B$ translates to A is proportional to B. When the data is held constant and the model parameters are allowed to vary, $f(\mathbf{d}|\rho, \omega, \phi, I)$ is called the likelihood function. The second term, $f(\rho, \omega, \phi|I)$, is the a priori distribution of our model parameters and will be called the prior for simplicity. This represents our state of knowledge prior to observing the data. Given our assumed noise distribution, the likelihood function can be written as [4]

$$f(\mathbf{d}|\rho,\omega,\phi,I) \propto \exp[-\frac{1}{2}(\mathbf{d}-\rho e^{j\phi}\mathbf{a}(\omega))^H\mathbf{R}^{-1}(\mathbf{d}-\rho e^{j\phi}\mathbf{a}(\omega))]. \tag{8}$$

This can be written in a more convenient form (See Appendix A) as

$$f(\mathbf{d}|\rho,\omega,\phi,I) \propto \exp[-\frac{1}{2}\rho^2\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{a}(\omega)]\exp[\rho|\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|\cos(\phi+\alpha(\omega))] \tag{9}$$

where

$$\alpha(\omega) = \tan^{-1}[\frac{\Im(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega))}{\Re(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega))}] \tag{10}$$

and $\Re(x)$ and $\Im(x)$ correspond to the real and imaginary parts of the complex argument, x, respectively.

To marginalize the conditional density function, we must integrate out the nuisance parameters, ρ and ϕ. Namely,

$$f(\omega|\mathbf{d},I) \propto \int_\rho\int_\phi f(\mathbf{d}|\rho,\omega,\phi,I)\ f(\rho,\omega,\phi|I)d\phi d\rho. \tag{11}$$

If we assume that ρ, ϕ and ω are independent, then we can write

$$f(\omega|\mathbf{d},I) \propto f(\omega|I)\int_\rho f(\rho|I)\int_\phi f(\mathbf{d}|\rho,\omega,\phi,I)\ f(\phi|I)d\phi d\rho. \tag{12}$$

The inner integral will be referred to as the phaseless likelihood function,

$$L(\omega,\rho) = \int_\phi f(\mathbf{d}|\rho,\omega,\phi,I)\ f(\phi|I)d\phi. \tag{13}$$

In order to perform the integration in (13), the prior density of ϕ given our a priori information must be chosen. In most communication systems, a reference phase is used and a distribution is presumed centered at the reference phase. In a passive system, such as radar, the only realistic choice is a noninformative density. In the derivations that follow we will use a noninformative density which, in this case, is the uniform distribution defined over the interval $[0,2\pi]$. Using the uniform prior, the phaseless likelihood is

$$L(\omega,\rho) \propto G(\omega,\rho)\int_0^{2\pi}\frac{1}{2\pi}\exp[\rho|\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|\cos(\phi+\alpha(\omega))]d\phi, \tag{14}$$

where $G(\omega,\rho)$ represents all the terms not depending on ϕ. The integral in (14) is recognized as the modified Bessel function of the first kind, zero'th order,

$$I_0(x) \stackrel{\text{def}}{=} \int_0^{2\pi}\frac{1}{2\pi}\exp[x\cos\phi]d\phi. \tag{15}$$

Hence, we can write the phaseless likelihood more compactly as

$$L(\omega,\rho) \propto \exp[-\frac{1}{2}\rho^2 \mathbf{a}(\omega)^H \mathbf{R}^{-1}\mathbf{a}(\omega)] I_0(\rho|\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|). \tag{16}$$

At this point, it is worthwhile to note that if we were interested in infering information about both the amplitude and the frequency of the observed waveform, we would multiply the phaseless likelihood by our prior densities for ρ and ω and normalize to obtain the joint a posteriori density.

To perform the second integration in (12), we must decide on a prior to use for the amplitude of the signal. The Rayleigh density is a judicious choice since, combined with the uniform phase distribution, it results in a normally distributed complex amplitude for the harmonic. We will thus write

$$f(\rho|I) = \frac{\rho}{{\sigma_s}^2} e^{-\frac{\rho^2}{2{\sigma_s}^2}}. \tag{17}$$

Letting

$$S(\omega) = \mathbf{a}(\omega)^H \mathbf{R}^{-1}\mathbf{a}(\omega) \tag{18}$$

and

$$T(\omega) = |\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|, \tag{19}$$

we can write

$$f(\omega|\mathbf{d},I) \propto f(\omega|I)\int_0^\infty \rho e^{-\frac{\rho^2}{2}(S(\omega)+\frac{1}{{\sigma_s}^2})} I_0(\rho T(\omega))d\rho. \tag{20}$$

Using the Integral formula (See Appendix B)

$$\int_0^\infty x e^{-\frac{1}{2}ax^2} I_0(bx)dx = \frac{1}{a}e^{\frac{b^2}{2a}}, \tag{21}$$

we can write

$$f(\omega|\mathbf{d},I) \propto f(\omega|I)(S(\omega)+\frac{1}{{\sigma_s}^2})^{-1}\exp[\frac{1}{2}\frac{T(\omega)^2}{S(\omega)+\frac{1}{{\sigma_s}^2}}]. \tag{22}$$

If we assume that the probability of one frequency occurring does not outweight any other frequency, we may write the conditional density of ω given the data and our a priori information as

$$f(\omega|\mathbf{d},I) \propto (S(\omega)+\frac{1}{{\sigma_s}^2})^{-1}\exp[\frac{1}{2}\frac{T(\omega)^2}{S(\omega)+\frac{1}{{\sigma_s}^2}}]. \tag{23}$$

It is informative to investigate the two limiting forms of (23). The first case is when $\sigma_s \to 0$. Normalizing (23), we may write

$$f(\omega|\mathbf{d},I) = \frac{({\sigma_s}^2 S(\omega)+1)^{-1}\exp[\frac{1}{2}\frac{{\sigma_s}^2 T(\omega)^2}{{\sigma_s}^2 S(\omega)+1}]}{\int_0^{2\pi}({\sigma_s}^2 S(\omega)+1)^{-1}\exp[\frac{1}{2}\frac{{\sigma_s}^2 T(\omega)^2}{{\sigma_s}^2 S(\omega)+1}]d\omega}. \tag{24}$$

The function,

$$g(\omega, \sigma_s) = ({\sigma_s}^2 S(\omega) + 1)^{-1} \exp[\frac{1}{2} \frac{{\sigma_s}^2 T(\omega)^2}{{\sigma_s}^2 S(\omega) + 1}], \tag{25}$$

is a continuous function of ω on the closed interval, $[0, 2\pi]$, for all σ_s in $[0, \infty)$. In addition, $g(\omega, \sigma_s)$ is dominated by the function $\exp[\frac{1}{2}\frac{T(\omega)^2}{S(\omega)}]$ and converges to the constant function, equal to unity, on $[0, 2\pi]$ as $\sigma_s \to 0$. Combined, these facts imply that we may make use of Lebesgue's convergence theorem [15] to show

$$\lim_{\sigma_s \to 0} \frac{g(\omega)}{\int_0^{2\pi} g(\omega) d\omega} = \frac{\lim_{\sigma_s \to 0} g(\omega)}{\int_0^{2\pi} \lim_{\sigma_s \to 0} g(\omega) d\omega}. \tag{26}$$

This allows us to state that

$$\lim_{\sigma_s \to 0} f(\omega | \mathbf{d}, I) = \frac{1}{\int_0^{2\pi} d\omega} \tag{27}$$

or

$$\lim_{\sigma_s \to 0} f(\omega | \mathbf{d}, I) = \frac{1}{2\pi}. \tag{28}$$

Thus, the limiting form, as $\sigma_s \to 0$, of the a posteriori distribution is the uniform distribution over the interval $[0, 2\pi]$. This is intuitively pleasing since as $\sigma_s \to 0$, the observations do not contain any signal information and hence, our best bet is to rely on our prior information. The Bayesian approach is effectively letting us know that the observations are of no help in this case.

It is also interesting to investigate the form of (23) when $\sigma_s \to \infty$. This is the case when our prior information is quite vague and a noninformative prior is appropriate. In that case,

$$f(\omega | \mathbf{d}, I) \propto S(\omega)^{-1} \exp[\frac{1}{2} \frac{T(\omega)^2}{S(\omega)}]. \tag{29}$$

The term within the exponential,

$$\frac{T(\omega)^2}{S(\omega)} = \frac{|\mathbf{d}^H \mathbf{R}^{-1} \mathbf{a}(\omega)|^2}{\mathbf{a}(\omega)^H \mathbf{R}^{-1} \mathbf{a}(\omega)}, \tag{30}$$

is the squared length of the vector $\mathbf{R}^{-\frac{1}{2}}\mathbf{d}$ projected onto the vector $\tilde{\mathbf{a}}(\omega) = \mathbf{R}^{-\frac{1}{2}}\mathbf{a}(\omega)$. This is the output of a correlation receiver where the data has been pre-whitened to remove the noise correlations. This term will be maximal when the whitened data looks most like the matching signal, $\tilde{\mathbf{a}}(\omega)$. Thus, the exponential term contributes to the a posterior distribution the degree to which the data is fit by a particular harmonic. The additional term, $S(\omega)^{-1}$, is a biasing term which is largest when the length of the vector $\mathbf{R}^{-\frac{1}{2}}\mathbf{a}(\omega)$ is minimized.

To understand the effect of the term, $S(\omega)^{-1}$, consider evaluating the a posteriori probability at two different frequencies, ω_1 and ω_2. Suppose that $S(\omega_1) < S(\omega_2)$ so that the vector, $\tilde{\mathbf{a}}(\omega_1)$ is smaller, in length, than $\tilde{\mathbf{a}}(\omega_2)$. This implies that the energy of a complex sinusoid at ω_1 would be smaller than one at ω_2 after they have been pre-whitened. For the a posteriori probabilities to be equal, it is required that

$$\frac{T(\omega_2)^2}{S(\omega_2)} = \frac{T(\omega_1)^2}{S(\omega_1)} + 2\ln[\frac{S(\omega_2)}{S(\omega_1)}] \tag{31}$$

and thus, to achieve the same a posteriori probability, the data must have a higher degree of correlation with a harmonic at ω_2 than at ω_1. This is because the whitening filter will also distort the complex sinusoid and, hence, the observed signal to noise ratio will be lower if the actual harmonic is at ω_1 compared to if it is at ω_2. The Bayesian approach takes into account the fact that certain frequencies are favored by the pre-whitening process.

Equipped with the a posteriori density of ω, we are in a position to extract information concerning it's true value. By far, the most popular piece of information is the point estimate of ω. Based on an optimality criteria, an ω is chosen to minimize an associated cost, or risk, function. Of course, there are many alternatives to the point estimate, such as a confidence interval, and some information is lost when only the "best" estimate of ω is retained. The maximum a posteriori (MAP) estimate is a popular point estimator which is defined as

$$\hat{\omega}_{MAP} \stackrel{\text{def}}{=} \arg\left\{\max_{\omega} f(\omega|\mathbf{d}, I)\right\}. \tag{32}$$

As seen, from (23), the a posteriori density of ω is a complicated expression, particularly for a general noise correlation matrix, and a closed form solution for the MAP estimate will not exist. In addition, the MAP estimate requires prior knowledge in the form of σ_s. Of course, by letting $\sigma_s \rightarrow \infty$, we can plead complete ignorence about the parameters before observing the data. In the case of spherically white noise, $\mathbf{R} = \sigma_n^2\mathbf{I}$, the a posteriori density can be written as

$$f(\omega|\mathbf{d}, I) \propto \exp[\frac{1}{2}\frac{|\mathbf{d}^H\mathbf{a}(\omega)|^2}{N\sigma_n^{-2} + \sigma_s^{-2}}]. \tag{33}$$

and due to the monotonicity of the natural logarithm, the MAP estimate can be written as

$$\hat{\omega}_{MAP} = \arg\left\{\max_{\omega} |\mathbf{d}^H\mathbf{a}(\omega)|\right\}. \tag{34}$$

An interesting observation, based on (34), is that, for the spherical white noise case, our prior information concerning ρ has no bearing on our estimate for ω. In fact, a more profound statement is that no matter what prior distribution is assumed for the amplitude of the signal, we will obtain the same result. This can be proved as follows. For the case of spherically white noise, we may rewrite (20) as

$$f(\omega|\mathrm{d},I) \propto \int_\rho \exp[-\frac{\rho^2}{2}N\sigma_n^{-2}]I_0(\rho T(\omega))f(\rho|I)d\rho. \tag{35}$$

The modified Bessel function is a convex function, symmetric about the origin. This implies that

$$I_0(\rho T(\omega_1)) \geq I_0(\rho T(\omega_2)) \quad \forall\, T(\omega_1) \geq T(\omega_2) \text{ and } \forall\, \rho. \tag{36}$$

Multiplying both sides of the inequality in (36) by a non-negative function and integrating over ρ does not change the statement. That is, if $g(\rho) \geq 0 \,\forall\, \rho$ then

$$\int_\rho g(\rho)I_0(\rho T(\omega_1))d\rho \geq \int_\rho g(\rho)I_0(\rho T(\omega_2))d\rho \quad \forall\, T(\omega_1) \geq T(\omega_2). \tag{37}$$

Thus maximizing

$$H(\omega) = \int_\rho g(\rho)I_0(\rho T(\omega))d\rho \tag{38}$$

is equivalent to maximizing $T(\omega)$. In our case

$$g(\rho) = \exp[-\frac{\rho^2}{2}N\sigma_n^{-2}]f(\rho|I) \geq 0 \quad \forall\, \rho \tag{39}$$

and thus, regardless of our prior density of ρ, the MAP estimate for ω will always correspond to the value of ω that maximizes $|\mathbf{d}^H\mathbf{a}(\omega)|$. Intuitively, this is due to the fact that our observed amplitude distribution is independent and identical for each ω under consideration. Thus, once a sample ρ has been generated, the underlying distribution does not matter in terms of our decision. The only consideration is the the length of the projection of the data vector onto the signal manifold. It is also interesting to note that, since phase was uniformly integrated out, the distance metric between the observed vector and the signal manifold is no longer pertinent but rather the closeness between the observed vector space and the hypothesized subspace, which is spanned by $\mathbf{a}(\omega)$.

Another popular point estimate is the minimal mean square error estimate (MMSE) which is synonymous with the conditional expectation of ω. That is

$$\hat{\omega}_{MMSE} = \mathrm{E}(\omega|\mathrm{d},I). \tag{40}$$

Similar to the MAP estimator, a closed form solution (40) is not feasible though it can be easily computed numerically. The MMSE estimate will coincide with the MAP estimate when the a posteriori density of ω is symmetric about it's maximal value.

In addition to the two Bayesian point estimators, a very popular approach to estimating the frequency of the complex exponential is the maximum likelihood (ML) technique, which is based on the philosophy: "Let the data speak for themselves." Since there is no marginalization, one is forced to solve for all the model parameters at once though in some cases, such as ours, the parameters can be solved for in succession. The ML estimator for frequency can be written as

$$\hat{\omega}_{ML} = \arg\max_{\omega} \frac{|\mathbf{d}^H \mathbf{R}^{-1} \mathbf{a}(\omega)|}{\mathbf{a}(\omega)^H \mathbf{R}^{-1} \mathbf{a}(\omega)}. \tag{41}$$

Under the assumption that $\mathbf{R} = \sigma_n^2 \mathbf{I}$, the ML estimate becomes

$$\hat{\omega}_{ML} = \arg\max_{\omega} |\mathbf{d}^H \mathbf{a}(\omega)| \tag{42}$$

and it can be seen that the ML and the MAP estimates are equivalent and that they are both equivalent to a matched filtering operation on the data. It can be easily shown that, the sufficient statistic, $|\mathbf{d}^H \mathbf{a}(\omega)|$ is equivalent to the standard periodogram, properly zero padded, and thus, for a single complex exponential in white spherical noise, the peak of the periodogram is an optimal point estimate for ω.

3 Multiple Frequency Estimation

For the case of more than one complex exponential, the problem is inherently more difficult to solve. The data model used in (2) becomes

$$d(t_k) = \sum_{i=1}^{M} \rho_i e^{j(\omega_i t_k + \phi_i)} + n(t_k), \qquad k = 0, \cdots, N-1 \tag{43}$$

where M is the number of complex exponentials, which is assumed known and $\omega_i \neq \omega_j \; \forall \; i \neq j$. We can write (43) in matrix form by using the notation of the previous section along with the following. Let ρ_M, ω_M and ϕ_M denote the $M \times 1$ parameter vectors such that the i'th element is the amplitude, frequency and phase of the i'th complex exponential, respectively. Using this notation, ρ_{M-1} would be the $(M-1) \times 1$ vector which contains all of the amplitudes of the complex exponentials except the M'th. In addition, define the matrices

$$\mathbf{A}(\omega_M) = [\mathbf{a}(\omega_1) \vdots \cdots \vdots \mathbf{a}(\omega_M)] \tag{44}$$

and

$$\mathbf{s} = [\rho_1 e^{j\phi_1}, \cdots, \rho_M e^{j\phi_M}]^T, \tag{45}$$

so that we may write (43) as

$$\mathbf{d} = \mathbf{A}(\omega_M)\mathbf{s} + \mathbf{n}. \tag{46}$$

It is common to refer to $\mathbf{A}(\omega_M)$ as the steering matrix and, as before, $\mathbf{s}$ is modeled as having a density corresponding to a zero mean, complex Gaussian random vector which represents the phase and the amplitudes of the complex exponentials.

The likelihood function of the data given our parameters is written as

$$f(\mathbf{d}|\rho_M, \omega_M, \phi_M, I) \propto \exp[-\frac{1}{2}(\mathbf{d} - \mathbf{A}(\omega_M)\mathbf{s})^H \mathbf{R}^{-1}(\mathbf{d} - \mathbf{A}(\omega_M)\mathbf{s})] \tag{47}$$

and we must integrate out the nuisance parameters, ρ_M and ϕ_M. To avoid cumbersome notation, let us define

$$S_M(\omega_i, \omega_j) = \mathbf{a}(\omega_i)^H \mathbf{R}^{-1} \mathbf{a}(\omega_j) \tag{48}$$

and

$$T_M(\omega_i) = \mathbf{a}(\omega_i)^H \mathbf{R}^{-1} \mathbf{d}. \tag{49}$$

Ignoring terms not depending on the parameters, we may write

$$f(\mathbf{d}|\rho_M, \omega_M, \phi_M, I) \propto \exp[\sum_{k=1}^{M} \rho_k \Re(e^{-j\phi_k} T_M(\omega_k)) - \frac{1}{2} \sum_{k=1}^{M} \sum_{l=1}^{M} \rho_k \rho_l e^{-j\phi_k} e^{j\phi_l} S_M(\omega_k, \omega_l)]. \tag{50}$$

The second term within the exponential of (50) can be written as

$$\sum_{k=1}^{M} \sum_{l=1}^{M} \rho_k \rho_l e^{-j\phi_k} e^{j\phi_l} S_M(\omega_k, \omega_l) = \sum_{k=1}^{M} \rho_k^2 S_M(\omega_k, \omega_k) + \sum_{\substack{k=1 \\ k \neq l}}^{M} \sum_{l=1}^{M} \rho_k \rho_l e^{-j\phi_k} e^{j\phi_l} S_M(\omega_k, \omega_l). \tag{51}$$

We can use the fact (See Appendix C),

$$\sum_{\substack{k=1 \\ k \neq l}}^{M} \sum_{l=1}^{M} \beta(k,l) = \sum_{k=1}^{M} \sum_{l=1}^{k-1} [\beta(k,l) + \beta(l,k)], \tag{52}$$

by noting that in our case $\beta(k,l) = \beta^*(l,k)$ to write the second term in (51) as

$$\sum_{\substack{k=1 \\ k \neq l}}^{M} \sum_{l=1}^{M} \rho_k \rho_l e^{-j\phi_k} e^{j\phi_l} S_M(\omega_k, \omega_l) = 2 \sum_{k=1}^{M} \rho_k \Re(e^{-j\phi_k} \sum_{l=1}^{k-1} \rho_l e^{j\phi_l} S_M(\omega_k, \omega_l)). \tag{53}$$

Inserting (53) in (51) and the result into (50), we obtain

$$f(\mathrm{d}|\rho_M, \omega_M, \phi_M, I) \propto \exp[-\frac{1}{2}\sum_{k=1}^{M}\rho_k^2 S_M(\omega_k,\omega_k)]$$
$$\times \exp[\sum_{k=1}^{M}\rho_k \Re(e^{-j\phi_k}(T_M(\omega_k) - \sum_{l=1}^{k-1}\rho_l e^{j\phi_l} S_M(\omega_k,\omega_l)))]. \quad (54)$$

At this point, we must integrate out the nuisance parameters. We will make use of the same priors as in Section 2 and we will allow $\sigma_s \rightarrow \infty$ so that our prior information represents complete ignorance about both the phase and the amplitude of the complex exponentials. In addition, this noninformative prior naturally assumes the most conservative posture that the signals are incoherent. The fact that the signals might be coherent does not bother us since this prior merely represents what we know about the signals, prior to observing the data.

It should be noted that the route we proceed down to marginalize the parameters differs significantly from previous work [2,10]. In those researchers' work, the parameters were marginalized using a noninformative prior after the basis functions, which in our case are cisoids, were orthogonalized. This is not only invalid, since the prior information pertains to the original signals, not the orthogonalized ones, but also leads to point estimates that are equivalent to the maximum likelihood estimates. To properly marginalize, the Jacobian of the orthogonalizing transformation needs to be incorporated into the prior before integration. As we will see shortly, the choice of prior is significant in terms of the resulting marginalized density.

To begin, we integrate out the phase and amplitude for the M'th cisoid. After some tedious algebra (See Appendix D), we obtain the result

$$f(\mathrm{d}|\rho_{M-1}, \omega_M, \phi_{M-1}, I) \propto \frac{1}{S_M(\omega_M,\omega_M)}$$
$$\times \exp[\frac{1}{2}\frac{|T_M(\omega_M)|^2}{S_M(\omega_M,\omega_M)}] \exp[-\frac{1}{2}\sum_{k=1}^{M-1}\rho_k^2 S_{M-1}(\omega_k,\omega_k)]$$
$$\times \exp[\sum_{k=1}^{M-1}\rho_k \Re(e^{-j\phi_k}(T_{M-1}(\omega_k) - \sum_{l=1}^{k-1}\rho_l e^{j\phi_l} S_{M-1}(\omega_k,\omega_l)))] \quad (55)$$

where we have introduced the notation

$$S_{M-1}(\omega_k,\omega_l) = S_M(\omega_k,\omega_l) - \frac{S_M(\omega_k,\omega_M)S_M(\omega_M,\omega_l)}{S_M(\omega_M,\omega_M)} \quad (56)$$

and

$$T_{M-1}(\omega_k) = T_M(\omega_k) - T_M(\omega_M)\frac{S_M(\omega_k,\omega_M)}{S_M(\omega_M,\omega_M)}. \tag{57}$$

Comparing (55) to (54), it is apparent that a recursive relationship exists and that, assuming no prior knowledge concerning the frequencies, we may write the a posteriori density of ω as

$$f(\omega_M|\mathrm{d},I) \propto \prod_{i=1}^{M} S_i(\omega_i,\omega_i)^{-1}\exp[\frac{1}{2}\frac{|T_i(\omega_i)|^2}{S_i(\omega_i,\omega_i)}]. \tag{58}$$

The terms, $S_i(\omega_i,\omega_i)$ and $T_i(\omega_i)$ in (58) are defined recursively for $i = 1,\cdots,M-1$ as

$$S_i(\omega_i,\omega_j) = S_{i+1}(\omega_i,\omega_j) - \frac{S_{i+1}(\omega_i,\omega_{i+1})S_{i+1}(\omega_{i+1},\omega_j)}{S_{i+1}(\omega_{i+1},\omega_{i+1})} \tag{59}$$

and

$$T_i(\omega_i) = T_{i+1}(\omega_i) - T_{i+1}(\omega_{i+1})\frac{S_{i+1}(\omega_i,\omega_{i+1})}{S_{i+1}(\omega_{i+1},\omega_{i+1})} \tag{60}$$

with $S_M(\omega_i,\omega_j)$ and $T_M(\omega_i)$ defined in (48) and (49), respectively.

These expressions seem complicated and are quite difficult to interpret. To simplify matters, we must first introduce some notation. If $\mathbf{x}$ and $\mathbf{y}$ are $N\times 1$ complex vectors and $\mathbf{B}$ is an $N\times N$, square, complex matrix then let

$$<\mathbf{x},\mathbf{y}>_{\mathbf{B}} \stackrel{\text{def}}{=} \mathbf{x}^H\mathbf{B}\mathbf{y}. \tag{61}$$

If $\mathbf{B}$ is positive semi-definite, then (61) satisfies all the properties of an inner product [11] and we can introduce a norm, induced by the inner product which is denoted as

$$||\mathbf{x}||_{\mathbf{B}} = <\mathbf{x},\mathbf{x}>_{\mathbf{B}}, \tag{62}$$

as well as make use of concepts such as orthogonality, projection operators and so forth. Two vectors, $\mathbf{x}$ and $\mathbf{y}$, are said to be conjugate directions, with respect to $\mathbf{B}$, if $<\mathbf{x},\mathbf{y}>_{\mathbf{B}}=0$. Note that conjugacy is a generalization of orthogonality. Since $\mathbf{R}^{-1}$ is hermitian with non-negative eigenvalues, it is positive semi-definite and we may use the above notation. We now prove, by induction, the following:

$$S_i(\omega_i,\omega_j) = <\mathbf{P}_i^{\perp}\mathrm{a}(\omega_i),\mathbf{P}_i^{\perp}\mathrm{a}(\omega_j)>_{\mathbf{R}^{-1}} \tag{63}$$

and

$$T_i(\omega_i) = <\mathbf{P}_i^{\perp}\mathrm{a}(\omega_i),\mathbf{d}>_{\mathbf{R}^{-1}}. \tag{64}$$

where $\mathbf{P}_i^\perp$, $i = 1, \cdots, M$, is the projection operator, defined under $\mathbf{R}^{-1}$, onto the conjugate subspace spanned by $\{\mathbf{a}(\omega_j)\}$, $j = i+1, \cdots, M$, with $\mathbf{P}_M^\perp = \mathbf{I}$, the identity operator. That is

$$< \mathbf{a}(\omega_j), \mathbf{P}_i^\perp \mathbf{x} >_{\mathbf{R}^{-1}} = 0 \quad i = 1, \cdots, M-1 \quad j = i+1, \cdots, M. \tag{65}$$

For $i = M$, both (63) and (64) are true by definition. We assume that they are true for $i = k+1$ and try to prove them true for $i = k$ where $0 \leq k < M$. For (63), we can use (59) and the inductive hypothesis to write

$$\begin{aligned} S_k(\omega_k, \omega_j) &= < \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_k), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}} \\ &- \frac{< \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_k), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}) >_{\mathbf{R}^{-1}} < \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}}}{||\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1})||_{\mathbf{R}^{-1}}^2} \end{aligned} \tag{66}$$

which can be written as

$$S_k(\omega_k, \omega_j) = < \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_k), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) - \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}) \frac{< \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}}}{||\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1})||_{\mathbf{R}^{-1}}^2} >_{\mathbf{R}^{-1}} . \tag{67}$$

Noting that the term on the right side of the inner product in (67) is $\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j)$ orthogonalized with respect to $\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1})$, we can write

$$\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) - \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}) \frac{< \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1}), \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}}}{||\mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_{k+1})||_{\mathbf{R}^{-1}}^2} = \mathbf{P}_{\mathbf{a}(\omega_{k+1})}^\perp \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_j) \tag{68}$$

where $\mathbf{P}_{\mathbf{a}(\omega_{k+1})}^\perp$ is the projection operator onto the subspace conjugate to $\mathbf{a}(\omega_{k+1})$. It is obvious that

$$\mathbf{P}_{\mathbf{a}(\omega_{k+1})}^\perp \mathbf{P}_{k+1}^\perp = \mathbf{P}_k^\perp \tag{69}$$

and thus

$$S_k(\omega_k, \omega_j) = < \mathbf{P}_{k+1}^\perp \mathbf{a}(\omega_k), \mathbf{P}_k^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}} \tag{70}$$

which is equivalent to

$$S_k(\omega_k, \omega_j) = < \mathbf{P}_k^\perp \mathbf{a}(\omega_k), \mathbf{P}_k^\perp \mathbf{a}(\omega_j) >_{\mathbf{R}^{-1}} . \tag{71}$$

This result implies that (63) holds for all $k = 1, \cdots, M$. Of interest to us is the special case when $k = j$ and we have

$$S_k(\omega_k, \omega_k) = ||\mathbf{P}_k^\perp \mathbf{a}(\omega_k)||_{\mathbf{R}^{-1}}^2 \tag{72}$$

Likewise for (64), we can write

$$T_k(\omega_k) = <\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_k), \mathbf{d}>_{\mathbf{R}^{-1}} - \frac{<\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1}), \mathbf{d}>_{\mathbf{R}^{-1}} <\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_k), \mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1})>_{\mathbf{R}^{-1}}}{||\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1})||^2_{\mathbf{R}^{-1}}} \tag{73}$$

which can be written

$$T_k(\omega_k) = <\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_k) - \mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1})\frac{<\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1}), \mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_k)>_{\mathbf{R}^{-1}}}{||\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1})||^2_{\mathbf{R}^{-1}}}, \mathbf{d}>_{\mathbf{R}^{-1}} . \tag{74}$$

As in (68), the term on the right side of the inner product in (74) is $\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_k)$ orthogonalized with respect to $\mathbf{P}_{k+1}^{\perp}\mathbf{a}(\omega_{k+1})$ so that we may write

$$T_k(\omega_k) = <\mathbf{P}_{k}^{\perp}\mathbf{a}(\omega_k), \mathbf{d}>_{\mathbf{R}^{-1}} . \tag{75}$$

which implies that (64) holds for all $k = 1, \cdots, M$.

As a result of (63) and (64), we can write the a posteriori density of ω in a more convenient form as

$$f(\omega_M|\mathbf{d}, I) \propto \prod_{i=1}^{M} \frac{1}{||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}}} \exp[\frac{1}{2}\frac{|<\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i), \mathbf{d}>_{\mathbf{R}^{-1}}|^2}{||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}}}] \tag{76}$$

and for the special case when $M = 1$, it can be seen that (76) reduces to (29). A similar result was independently derived by Quinn [12] though in a form the does not illustrate the recursive nature of the density.

Given the a posteriori density, we can use it to extract the desired information concerning ω_M. Concentrating on the point estimation of ω_M, the MAP or the MMSE estimate may be used. In either case, an analytical solution is not possible so the estimates must be constructed numerically. The MAP estimate is the value of ω_M that maximizes (76). Theoretically, the MAP estimate does not exist because the density has an unbounded derivative as two frequencies become arbitrarily close. In practice, the density can be evaluated on a discrete grid in the parameter space and this problem can be somewhat alleviated. Taking the logarithm of (76), we have

$$\hat{\omega}_{M_{MAP}} \stackrel{\text{def}}{=} \arg\left\{\max_{\omega_M}\sum_{i=1}^{M}\frac{|<\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i), \mathbf{d}>_{\mathbf{R}^{-1}}|^2}{||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}}} - 2\ln(||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}})\right\} . \tag{77}$$

Recall that the maximum likelihood estimate is

$$\hat{\omega}_{M_{ML}} \stackrel{\text{def}}{=} \arg\left\{\max_{\omega_M}\sum_{i=1}^{M}\frac{|<\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i), \mathbf{d}>_{\mathbf{R}^{-1}}|^2}{||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}}}\right\} \tag{78}$$

which corresponds to finding the best fitting subspace to the data. The MAP and the ML estimates are intimitely related, the only difference being the penalty term, in (77),

$$\mathcal{P}(\omega_M) = 2\sum_{i=1}^{M} \ln(||\mathbf{P}_i^{\perp}\mathbf{a}(\omega_i)||^2_{\mathbf{R}^{-1}}) \tag{79}$$

which is a biasing term that takes into account the coloration of the noise as well as the effect of orthogonalization. As was the case with one complex exponential, frequencies that are close to the dominant noise directions are penalized less to account for their inherently poorer SNR after pre-whitening. If the noise is spherically white, then no frequency has an advantage over another due to the noise.

Contrary to the case of one cisoid, there is a second penalty effect which arises due to the natural orthogonalizing procedure that results from marginalizing the parameters. To understand this effect, consider the case of evaluating the MAP estimate for $M = 2$ in spherically white noise. The first term in (77), equivalent to the ML estimator, wants to choose the ω_1 and ω_2 that best fits the data. The penalty term depends only on ω_1 and is proportional to the length of $\mathbf{a}(\omega_1)$ after it has been projected onto the subspace which is orthogonal to $\mathbf{a}(\omega_2)$. The smaller the length of that vector, the less penalty that is applied to the overall cost function. The reason for this is that if there truly were two complex exponentials at ω_1 and ω_2 then, by virtue of orthogonalizing $\mathbf{a}(\omega_1)$ with respect to $\mathbf{a}(\omega_2)$, the only data that is available to fit $\mathbf{a}(\omega_1)$ is that portion of $\mathbf{d}$ which is orthogonal to $\mathbf{a}(\omega_2)$. Since, for closely spaced frequencies, a substantial amount of $\mathbf{d}$ which is due to $\mathbf{a}(\omega_1)$ is unavailable for it's use, the Bayesian approach accounts for this via the penalty term. The less data we have, the less we have to fit the data. The Bayesian approach accounts for the lower SNR, due to the orthogonalization process, and thus (77) can be viewed as a generalized subspace fitting method.

One drawback of the penalty factor is that it tends to dominate the density when two frequencies are hypothesized very close together. This affects the estimates greatly when only the peak of the density is used for point estimates. The MMSE estimate is somewhat immune to this effect since an integration over the parameter space is used to form the estimate. As mentioned previously, the MAP estimate does not truly exist due to the unbounded derivative as two frequencies get close.

This penalty effect was also discussed in [12] though it was interpreted as a parsimonious model term which under severe conditions (i.e., low SNR or closely seperated frequencies) would serve as an indication of over-parameterization. Of course, this is a reasonable explanation when the density is evaluated on a discrete grid though the conditions when the penalty term begins to dominate depends on the grid spacing, which is an undesirable feature. All in all, the MAP estimate has limited inferential capability using the marginalized density. Of course, this opens the door for future research to investigate this penalty term and to improve our model hypothesis to account for the case when the model degrades to one of lower dimension.

It should be noted that if the cisoids were orthogonal to begin with (i.e., widely spaced frequencies and/or large data records) then the MAP and the ML are equivalent and, in the case of spherically white noise, they both correspond to finding the M dominant peaks in the periodogram. This helps explain the widespread use of the periodogram for frequency estimation. We have neglected to discuss the MMSE estimate, not because of it's lack of importance, but rather because, in general terms, it is difficult to say much. In practice, it turns out that the MMSE estimator performs very well and deserves it's title as the minimum variance estimator.

4 Simulation Results

To illustrate the performance of the Bayesian approach, a number of experiments were performed. A complex, colored Gaussian noise sequence was created by driving a second order, all pole filter with zero mean, white Gaussian noise. The noise sequence is thus an autoregressive process and can be expressed by the recursive equation

$$n_k = a_1 n_{k-1} + a_2 n_{k-2} + w_k \tag{80}$$

where w_k is a zero mean, complex Gaussian noise process with $E(w_i w_j) = 2\sigma^2 \delta_{ij}$ with δ_{ij} being the Kronecker delta. The real and imaginary parts of the complex noise process are independent and the poles of the filter were chosen as a conjugate pair, located at $z = re^{\pm j\omega_0}$. This implies that

$$a_1 = 2r\cos(\omega_0) \tag{81}$$

and

$$a_2 = -r^2 \tag{82}$$

with $|r| < 1$ to insure stability.

For simplicity, and without loss of generality, ω_0 was chosen to be $\frac{\pi}{2}$. In this case, $a_1 = 0$ and it can be shown that the correlation sequence, $\mathcal{R}(i-j) = E(n_i n_j)$, is

$$\mathcal{R}(k) = 2\begin{cases} \frac{\sigma^2 r^{|k|}}{1-r^4} & k = 0,4,8,\ldots \\ -\frac{\sigma^2 r^{|k|}}{1-r^4} & k = 2,6,10,\ldots \\ 0 & \text{otherwise} \end{cases} \tag{83}$$

and the i,j'th element of the correlation matrix for the real and imaginary part of the noise is thus

$$[\mathbf{R}]_{i,j} = \frac{1}{2}\mathcal{R}(|i-j|). \tag{84}$$

The magnitude spectra for three different choices of r^2 are shown in Figure 1. The magnitude spectrum of the noise can be written as

$$\mathcal{S}(\omega) = \frac{\sigma^2}{1 - 2r^2\cos(2\omega) + r^4} \tag{85}$$

and the ordinate axis is in units of dB. As r approaches 1, the bandwidth of the noise decreases though it's maximum amplitude increases reflecting the poles' proximity to the unit circle.

Initially, a single complex sinusoid, or cisoid, was embedded within the complex Gaussian noise and the three estimators, ML, MAP and MMSE, were used to estimate the frequency of the cisoid. To justly compare performance, the estimators were compared as a function of both frequencey and signal to noise ratio (SNR), where the SNR, in dB, is defined as

$$\text{SNR} \stackrel{\text{def}}{=} 20\log_{10}\left(\frac{\rho}{\sqrt{\mathcal{R}(0)}}\right). \tag{86}$$

For the noise sequence that we have used, the SNR becomes

$$\text{SNR} = 20\log_{10}\left(\frac{\rho\sqrt{1-r^4)}}{\sqrt{2}\sigma}\right). \tag{87}$$

The phase of the cisoid was zero radians and the prior information, for the Bayesian estimators, was chosen to represent complete ignorance of the parameters which means (29) was used for the MAP and the MMSE estimates. Thus for our experiments, the parameters were deterministic and the prior distribution represents our a priori knowledge concerning the value of the parameters before the experiment was run. The estimators were run at 21 evenly spaced frequencies ranging from 0 to π. For each run, 250 trials were conducted to insure statistical significance and the RMS error, for each estimator, was computed.

In addition to comparing the performance of the estimators against each other, they are compared to the Cramer-Rao (CR) bound for an unbiased frequency estimator in colored Gaussian noise (See Appendix E.) In Figure 2 are the RMS curves, as a function of frequency, for a SNR of -6 dB and for $r^2 = .8$. The number of samples, N, used in the estimation procedure was 25 and they were equally spaced with a unit sampling rate.

As can be seen, for frequencies that are within the noise band, the MAP and the MMSE estimates both perform better than the ML estimates. In fact, at the point of closest appoach to the poles of the noise process, the MMSE estimator reduces the RMS error by a factor of 4 when compared to the ML estimator. As described in Section 2, this is because the Bayesian approach accounts for the fact that frequencies outside of the noise band are favored in the whitening process. For frequencies outside of the noise band, the estimators are essentially equivalent since the noise has little effect on the cisoids in this region. As expected, all of the estimators do not achieve the CR bound except in the frequency range were the noise is insignificant. It is interesting to note that the CR bound exhibits the same behavior as the estimators in the sense that the bound increase within the noise band.

In Figure 3 are the RMS curves for a SNR of 0 dB. The improvement in performance of the MAP and MMSE estimators is still evident though over a narrower frequency band. The increased SNR has decreased the range that the noise has an impact on the estimators. Within the noise band, the Bayesian estimators attain a significant improvement as compared to the ML estimator. In the band where the noise is essentially white, the estimators are equivalent which is a reflection of the analysis in Section 2. The performance of the estimators for a single cisoid at $\frac{\pi}{2}$ as a function of SNR is shown in Figure 5. As can be seen, the MAP and MMSE are superior to the ML estimator. To illustrate the overall performance as a function of SNR, the mean squared error over all frequencies was computed and plotted as a function of SNR in Figure 5. The MAP and MMSE estimators perform better than the ML estimator though all three achieve the CR bound at a high SNR. The MMSE performs better than the MAP estimator since the a posteriori density is generally not symmetric which induces a higher bias in the MAP estimates.

The next experiment that was conducted examined the performance of the estimators for multiple frequency estimation. Two complex exponentials were superimposed onto a colored noise sequence which was generated in the same manner as in the first experiment with $r^2 = .8$. The frequencies of the cisoids were, $\omega_1 = .5\pi$ and $\omega_2 = .52\pi$. The first cisoid had zero phase while the second one had a phase of $\pi/4$ and the amplitudes were chosen to create a stated signal to noise ratio. This is a standard signal used when examining frequency resolution and the continuous waveform is written as

$$d(t) = \rho e^{j.5\pi t} + \rho e^{j(.52\pi t + .25\pi)} + n(t). \tag{88}$$

The number of samples used was 25 and 100 trials were performed for each SNR. The mean squared error, MSE, was computed for each estimator and is plotted in Figure 6. As can be seen, at the higher SNR, all three estimators are essentially equivalent while, as the SNR decreases, the ML estimator degrades more rapidly than the MAP or the MMSE. At a SNR below 0 dB, the ML estimator exhibits poor performance, while the MAP and the MMSE still retain a reasonable degree of fidelity. The MAP and the MMSE estimates are higher than the Cramer-Rao bound implying that a bias exists for those estimators. These simulation results not only confirm our analysis of the previous sections but also indicate the significance of the Bayesian approach.

5 Conclusion

In this paper we have examined the Bayesian estimation of the frequencies of multiple cisoids embedded in general Gaussian noise. For both the single cisoid and the multiple frequency cases, the a posteriori density of the frequencies given the data and our prior information were derived. The unwanted parameters were properly marginalized and appropriate prior densities used in the integration.

For the single cisoid, the resulting density was shown to account for the coloration of the noise by taking into account that frequencies closer to the dominant noise directions were prejudiced by the matched filtering operation. Under spherical white noise assumptions, it was shown that the MAP and the ML estimates were equivalent and that our prior density did not matter in terms of the maximum of the density.

For the multiple frequency case, the derived density not only accounted for the coloration of the noise but also took into account the degree of orthogonality between the signal vectors. The closer two frequencies were hypothesized, the less penalty was applied. This resulted in an a posteriori density whose maximum does not exist within the parameter space and thus the MAP estimate is, theoretically, undefined. A number of simulations were performed and the Bayesian approach was shown to be significantly better than the ML for both a single cisoid in colored Gaussian noise as well as two closely seperated cisoids in arbitrary noise. This indicates that the Bayesian methodology offers significant advantages and is worthy of future investigations.

One idea for future research is to concentrate on the penalty term of the marginalized density to understand it's effect more thoroughly and to reevaluate our model hypothesis. One thought is that it may not be valid to imply that two frequencies can not be equal. It would be more logical to have the higher dimensional a posteriori density transition to the lower dimensional density along the space where two frequencies are equal. This will probably make us rethink our model as well as assumptions we have made.

A Appendix

Derivation of Equation (9)

Writing (8) again

$$f(\mathbf{d}|\rho,\omega,\phi,I) \propto \exp[-\frac{1}{2}(\mathbf{d}-\rho e^{j\phi}\mathbf{a}(\omega))^H\mathbf{R}^{-1}(\mathbf{d}-\rho e^{j\phi}\mathbf{a}(\omega))], \tag{89}$$

we can multiply through the quadratic form and discard terms that are independent of the parameters,

$$f(\mathbf{d}|\rho,\omega,\phi,I) \propto \exp[-\frac{1}{2}\rho^2\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{a}(\omega)]\exp[\frac{\rho}{2}(e^{j\phi}\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)+e^{-j\phi}\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{d})], \tag{90}$$

which can be written as

$$\begin{aligned} f(\mathbf{d}|\rho,\omega,\phi,I) \propto{}& \exp[-\frac{1}{2}\rho^2\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{a}(\omega)] \\ &\times\exp[\rho(\cos(\phi)\Re(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega))-\sin(\phi)\Im(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)))]. \end{aligned} \tag{91}$$

By letting

$$\alpha(\omega) = \tan^{-1}[\frac{\Im(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega))}{\Re(\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega))}], \tag{92}$$

we can write

$$\begin{aligned} f(\mathbf{d}|\rho,\omega,\phi,I) \propto{}& \exp[-\frac{1}{2}\rho^2\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{a}(\omega)] \\ &\times\exp[\rho|\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|(\cos(\phi)\cos(\alpha(\omega))-\sin(\phi)\sin(\alpha(\omega)))] \end{aligned} \tag{93}$$

and finally, using the cosine addition formula, we have

$$f(\mathbf{d}|\rho,\omega,\phi,I) \propto \exp[-\frac{1}{2}\rho^2\mathbf{a}(\omega)^H\mathbf{R}^{-1}\mathbf{a}(\omega)]\exp[\rho|\mathbf{d}^H\mathbf{R}^{-1}\mathbf{a}(\omega)|\cos(\phi+\alpha(\omega))] \tag{94}$$

B Appendix

Derivation of Integral Formula

Consider the integral

$$\int_0^\infty xe^{-\frac{1}{2}ax^2}I_0(bx)dx. \tag{95}$$

Using the change of variables, $u = \sqrt{a}x$, we can write

$$\int_0^\infty xe^{-\frac{1}{2}ax^2} I_0(bx)dx = \int_0^\infty \frac{u}{a} e^{-\frac{u^2}{2}} I_0(\frac{b}{\sqrt{a}}u)du \tag{96}$$

or going one step furthur,

$$\int_0^\infty xe^{-\frac{1}{2}ax^2} I_0(bx)dx = \frac{1}{a} e^{\frac{b^2}{2a}} \int_0^\infty ue^{-\frac{1}{2}(u^2+\frac{b^2}{a})} I_0(\frac{b}{\sqrt{a}}u)du. \tag{97}$$

The term within the integral on the right side of (97) is recognized as the Rician distribution and thus integrates to unity. This allows us to write the desired result,

$$\int_0^\infty xe^{-\frac{1}{2}ax^2} I_0(bx)dx = \frac{1}{a} e^{\frac{b^2}{2a}}. \tag{98}$$

C Appendix

Proof of Equation (52)

We shall prove

$$\sum_{k=1}^{M} \sum_{\substack{l=1 \\ k\neq l}}^{M} \beta(k,l) = \sum_{k=1}^{M} \sum_{l=1}^{k-1} [\beta(k,l) + \beta(l,k)], \tag{99}$$

by induction. For $M = 2$, we have

$$\sum_{k=1}^{2} \sum_{\substack{l=1 \\ k\neq l}}^{2} \beta(k,l) = \beta(1,2) + \beta(2,1) \tag{100}$$

and

$$\sum_{k=1}^{2} \sum_{l=1}^{k-1} [\beta(k,l) + \beta(l,k)] = \beta(2,1) + \beta(1,2) \tag{101}$$

so (99) holds for $M = 2$. We assume that (99) holds for $M - 1$ and prove it holds for M. We may write

$$\sum_{k=1}^{M} \sum_{\substack{l=1 \\ k\neq l}}^{M} \beta(k,l) = \sum_{k=1}^{M-1} \sum_{\substack{l=1 \\ k\neq l}}^{M} \beta(k,l) + \sum_{l=1}^{M-1} \beta(M,l) \tag{102}$$

or

$$\sum_{\substack{k=1 \\ k\neq l}}^{M}\sum_{l=1}^{M} \beta(k,l) = \sum_{\substack{k=1 \\ k\neq l}}^{M-1}\sum_{l=1}^{M-1} \beta(k,l) + \sum_{l=1}^{M-1}[\beta(M,l)+\beta(l,M)]. \tag{103}$$

Using the inductive hypothesis,

$$\sum_{\substack{k=1 \\ k\neq l}}^{M}\sum_{l=1}^{M} \beta(k,l) = \sum_{k=1}^{M-1}\sum_{l=1}^{k-1}[\beta(k,l)+\beta(l,k)] + \sum_{l=1}^{M-1}[\beta(M,l)+\beta(l,M)] \tag{104}$$

and the fact that

$$\sum_{l=1}^{M-1}[\beta(M,l)+\beta(l,M)] = \sum_{k=M}^{M}\sum_{l=1}^{k-1}[\beta(k,l)+\beta(l,k)], \tag{105}$$

we may write

$$\sum_{\substack{k=1 \\ k\neq l}}^{M}\sum_{l=1}^{M} \beta(k,l) = \sum_{k=1}^{M}\sum_{l=1}^{k-1}[\beta(k,l)+\beta(l,k)] \tag{106}$$

and thus (99) is true for all M.

D Appendix

Derivation of Equation (55)

Beginning with

$$f(\mathbf{d}|\rho_M,\omega_M,\phi_M,I) \propto \exp[-\frac{1}{2}\sum_{k=1}^{M}\rho_k^2 S_M(\omega_k,\omega_k)]$$
$$\times\exp[\sum_{k=1}^{M}\rho_k \Re(e^{-j\phi_k}(T_M(\omega_k)-\sum_{l=1}^{k-1}\rho_l e^{j\phi_l}S_M(\omega_k,\omega_l)))], \tag{107}$$

we can use the results from Section 2 to show that

$$f(\mathbf{d}|\rho_{M-1},\omega_M,\phi_{M-1},I) \propto \exp[-\frac{1}{2}\sum_{k=1}^{M-1}\rho_k^2 S_M(\omega_k,\omega_k)]$$
$$\times\exp[\sum_{k=1}^{M-1}\rho_k \Re(e^{-j\phi_k}(T_M(\omega_k)-\sum_{l=1}^{k-1}\rho_l e^{j\phi_l}S_M(\omega_k,\omega_l)))]$$
$$\times S_M(\omega_M,\omega_M)^{-1}\exp[\frac{1}{2}\frac{|T_M(\omega_M)-\sum_{l=1}^{M-1}\rho_l e^{j\phi_l}S_M(\omega_M,\omega_l)|^2}{S_M(\omega_M,\omega_M)}]. \tag{108}$$

The last term in (108) can be expanded as

$$\begin{aligned}
&\exp[\frac{1}{2}\frac{|T_M(\omega_M) - \sum_{l=1}^{M-1}\rho_l e^{j\phi_l} S_M(\omega_M,\omega_l)|^2}{S_M(\omega_M,\omega_M)}] \\
&= \exp[\frac{1}{2}\frac{|T_M(\omega_M)|^2}{S_M(\omega_M,\omega_M)}]\exp[-\sum_{k=1}^{M-1}\rho_k \Re(e^{-j\phi_k}\frac{T_M(\omega_M)S_M(\omega_k,\omega_M)}{S_M(\omega_M,\omega_M)})] \\
&\times \exp[\frac{1}{2}\sum_{k=1}^{M-1}\sum_{l=1}^{M-1}\rho_k\rho_l e^{-j\phi_k}e^{j\phi_l}\frac{S_M(\omega_k,\omega_M)S_M(\omega_M,\omega_l)}{S_M(\omega_M,\omega_M)}].
\end{aligned} \tag{109}$$

With the result from (52), we may write the last term of (109) as

$$\begin{aligned}
&\exp[\frac{1}{2}\sum_{k=1}^{M-1}\sum_{l=1}^{M-1}\rho_k\rho_l e^{-j\phi_k}e^{j\phi_l}\frac{S_M(\omega_k,\omega_M)S_M(\omega_M,\omega_l)}{S_M(\omega_M,\omega_M)}] \\
&= \exp[\frac{1}{2}\sum_{k=1}^{M-1}\rho_k^2\frac{|S_M(\omega_k,\omega_M)|^2}{S_M(\omega_M,\omega_M)}] \\
&\times \exp[\sum_{k=1}^{M-1}\rho_k\Re(e^{-j\phi_k}\sum_{l=1}^{k-1}\rho_l e^{j\phi_l}\frac{S_M(\omega_k,\omega_M)S_M(\omega_M,\omega_l)}{S_M(\omega_M,\omega_M)})].
\end{aligned} \tag{110}$$

Using the notation

$$S_{M-1}(\omega_k,\omega_l) = S_M(\omega_k,\omega_l) - \frac{S_M(\omega_k,\omega_M)S_M(\omega_M,\omega_l)}{S_M(\omega_M,\omega_M)} \tag{111}$$

and

$$T_{M-1}(\omega_k) = T_M(\omega_k) - T_M(\omega_M)\frac{S_M(\omega_k,\omega_M)}{S_M(\omega_M,\omega_M)}, \tag{112}$$

we may combine (109) - (112) and write (108) as

$$\begin{aligned}
f(\mathbf{d}|\rho_{M-1},\omega_M,\phi_{M-1},I) &\propto \frac{1}{S_M(\omega_M,\omega_M)} \\
&\times \exp[\frac{1}{2}\frac{|T_M(\omega_M)|^2}{S_M(\omega_M,\omega_M)}]\exp[-\frac{1}{2}\sum_{k=1}^{M-1}\rho_k^2 S_{M-1}(\omega_k,\omega_k)] \\
&\times \exp[\sum_{k=1}^{M-1}\rho_k\Re(e^{-j\phi_k}(T_{M-1}(\omega_k) - \sum_{l=1}^{k-1}\rho_l e^{j\phi_l}S_{M-1}(\omega_k,\omega_l)))].
\end{aligned} \tag{113}$$

E Appendix

Cramer-Rao Bound for Frequency Estimation

The Cramer-Rao bound for an unbiased estimate of a deterministic parameter vector, θ, can be obtained by inverting the Fisher information matrix which is defined as

$$[\mathbf{J}]_{i,j} \stackrel{\text{def}}{=} E\left\{\frac{\partial}{\partial\theta_i}\mathcal{L}(\theta)\frac{\partial}{\partial\theta_j}\mathcal{L}(\theta)\right\}, \tag{114}$$

where $\mathcal{L}(\theta)$ is the log-likelihood function of the observations given θ. This can be written in compact matrix form as

$$\mathbf{J} = E\left\{(\nabla_\theta\mathcal{L}(\theta))(\nabla_\theta\mathcal{L}(\theta))^H\right\}, \tag{115}$$

with ∇_θ being the gradient with respect to θ. For the Gaussian problem, where the covariance matrix $\mathbf{R}$ is known, the log-likelihood function can be written as

$$\mathcal{L}(\theta) = -\frac{1}{2}(\mathbf{d} - \mathbf{h}(\theta))^H\mathbf{R}^{-1}(\mathbf{d} - \mathbf{h}(\theta)). \tag{116}$$

The function, $\mathbf{h}(\theta)$, is the known mapping from our parameter space to the observation space and all terms not depending on the parameters have been dropped. We can write the gradient of the log-likelihood function as

$$\nabla_\theta\mathcal{L}(\theta) = -\Re(\mathcal{J}_{\mathbf{h}}^H\mathbf{R}^{-1}(\mathbf{d} - \mathbf{h}(\theta))) \tag{117}$$

where $\mathcal{J}_{\mathbf{h}}$ is the Jacobian matrix of the function $\mathbf{h}(\theta)$ and

$$[\mathcal{J}_{\mathbf{h}}]_{i,j} = \frac{\partial h_i(\theta)}{\partial\theta_j}. \tag{118}$$

It is a simple matter to compute (115) given (117) and we obtain

$$\mathbf{J} = \Re(\mathcal{J}_h^H\mathbf{R}^{-1}\mathcal{J}_h). \tag{119}$$

For the case of white, spherical noise, the Fisher information matrix becomes

$$\mathbf{J} = \frac{1}{\sigma^2}\Re(\mathcal{J}_h^H\mathcal{J}_h). \tag{120}$$

Once the Fisher information matrix is computed, it is inverted to obtain the Cramer-Rao bound,

$$Cov(\theta) \geq \mathbf{J}^{-1}, \tag{121}$$

with the diagonal elements representing the lower bound on any unbiased estimator for the parameter vector, θ.

References

[1] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley, 1973.

[2] L. Bretthrost, "Bayesian spectrum analysis and parameter estimation," Ph.D. Thesis, Washington University, St. Louis, 1987.

[3] R. T. Cox, "Probability, frequency and reasonable expectation," *Am. J. Phys.*, Vol. 14, pp. 1-13, 1946.

[4] N. R. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution," *Ann. Math. Stat.* , pp. 152-177, 1963.

[5] E. T. Jaynes, "Bayesian spectrum and chirp analysis," *Proc. of the Third workshop on maximum entropy and Bayesian methods*, Boston, MA: D. Reidel, 1987.

[6] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Sys. Sci. and Cybern.*, Vol. SSC-4, pp. 227-241, Sept. 1968.

[7] H. Jeffreys, *Theory of Probability*, New York, NY: Oxford University Press, 1967.

[8] D. H. Johnson, "The application of spectral estimation methods to bearing estimation problems," *Proc. of the IEEE*, Vol. 70, pp. 1018-1028, Sept. 1982.

[9] S. C. Kramer and H. W. Sorenson, "Bayesian parameter estimation," *IEEE Trans. Automat. Contr.*, Vol. 33, pp. 217-222, Feb. 1988.

[10] J. Lasenby and W. J. Fitzgerald, "A Bayesian approach to high-resolution beamforming," *IEE Proceedings*, Vol. 138, pp. 539-544, Dec. 1991.

[11] B. Noble and J. W. Daniel, *Applied Linear Algebra*, New Jersey: Prentice-Hall, 1988.

[12] A. Quinn,"The performance of Bayesian estimators in the superresolution of signal parameters," *Proc. 1992 Int. Conf. Acoust., Speech, Signal Process.*, Vol. 5, pp. 297-300, 1992.

[13] J. P. Reilly, K. M. Wong and P. M. Reilly, "Direction of arrival estimation in the presence of noise with unknown covariance matrices," *Proc. 1989 Int. Conf. Acoust., Speech, Signal Process.*, pp. 2609-2612, 1989.

[14] D. C. Rife and R. R. Boorstyn, "Single tone parameter estimation from discrete-time observations," *IEEE Trans. Infor. Theory*, Vol. IT-20, pp. 591-598, Sept. 1974.

[15] W. Rudin, *Principles of Mathematical Analysis*, New York, NY: McGraw-Hill, 1976.

[16] D. Slepian, "Estimation of signal parameters in the presence of noise," *IRE Trans. Infor. Theory*, PGIT-3, pp. 68-89, Mar. 1957.

[17] D.W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood," *Proc. of the IEEE*, Vol. 70, pp. 975-989, Sept. 1982.

[18] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust., Speech, Sig. Proc.*, Vol. 36, pp. 1553-1560, Oct. 1988.

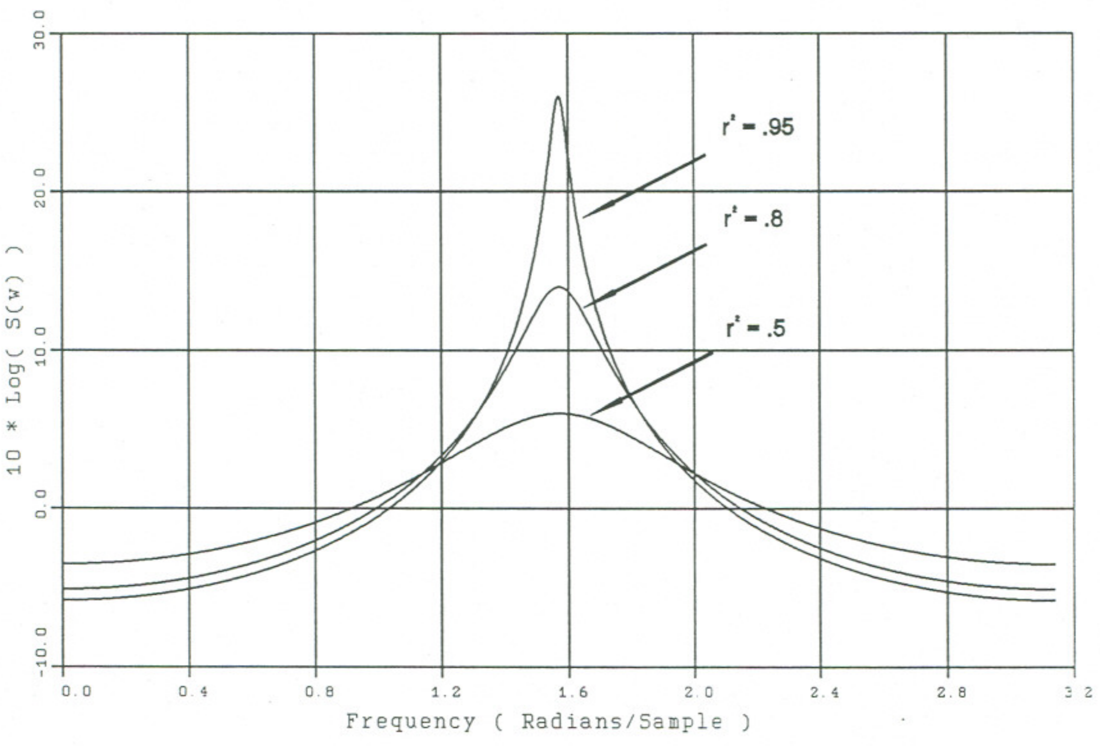

Figure 1: Spectra of real and imaginary noise process for r^2 = .5, .8 and .95.

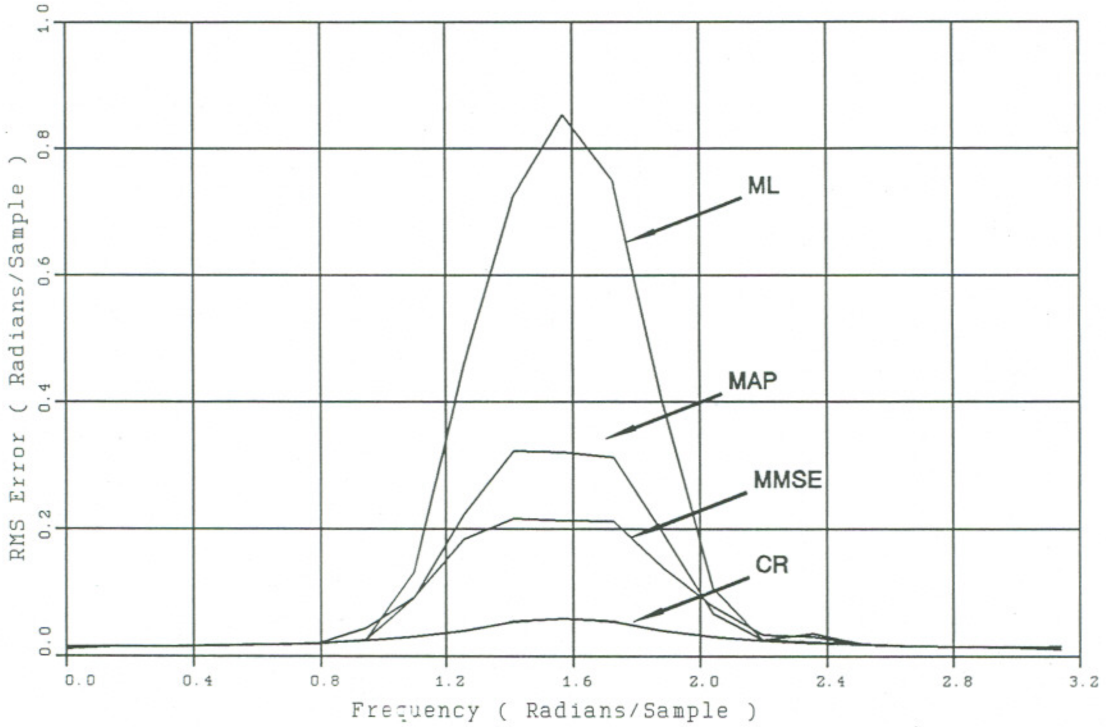

Figure 2: RMS error vs. Frequency for single cisoid. SNR = -6 dB with N = 25, $\bar{r}^2$ = .8 and 250 trials for each of 21 evenly spaced frequencies.

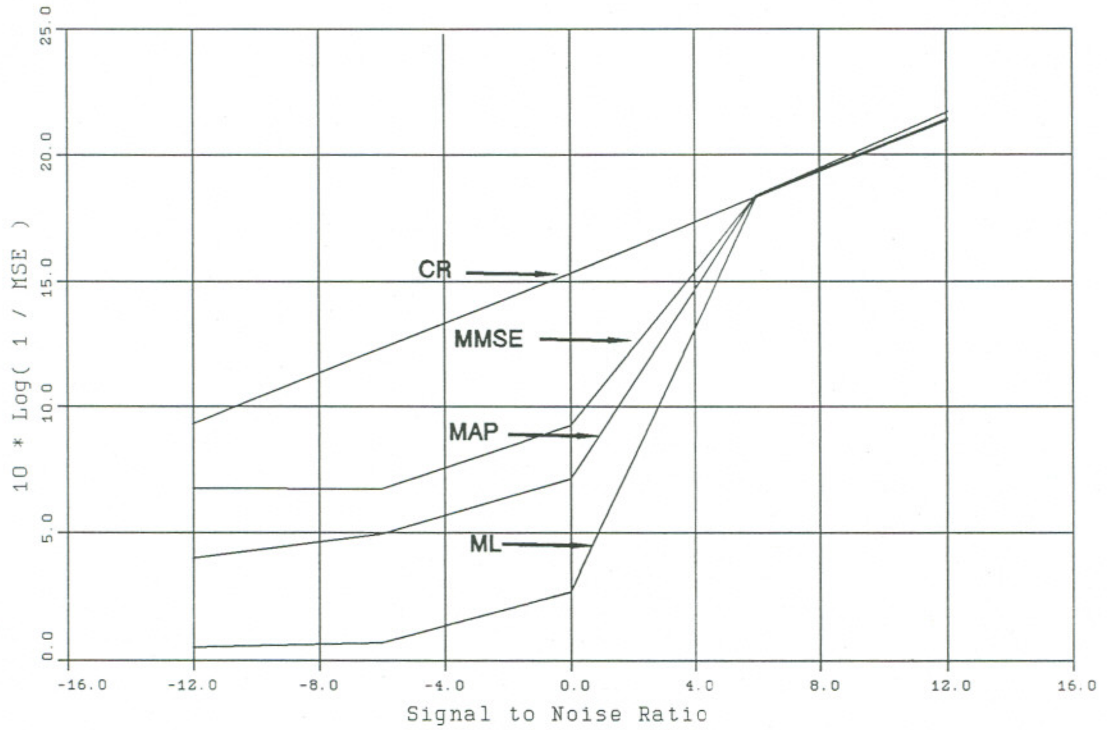

Figure 4: MSE vs. SNR for single cisoid at .5 PI. N = 25 and r^2 = .8

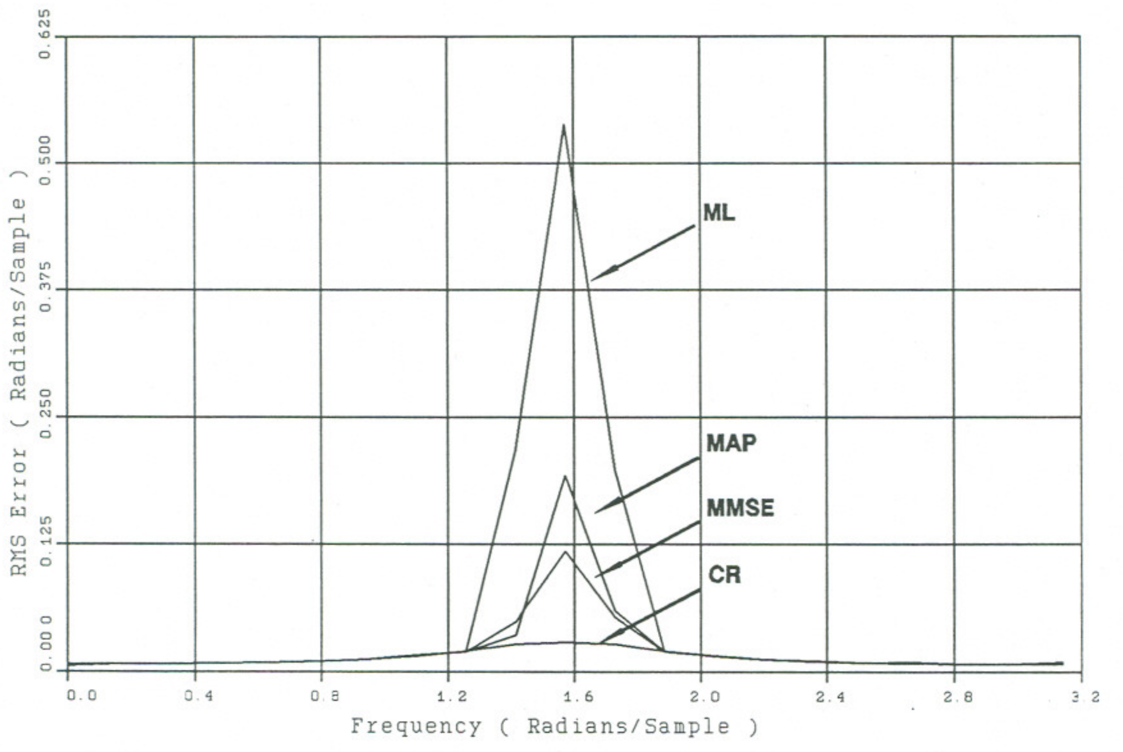

Figure 3: RMS error vs. Frequency for single cisoid. SNR = 0 dB with N = 25, r^2 = .8 and 250 trials for each of 21 evenly spaced frequencies.

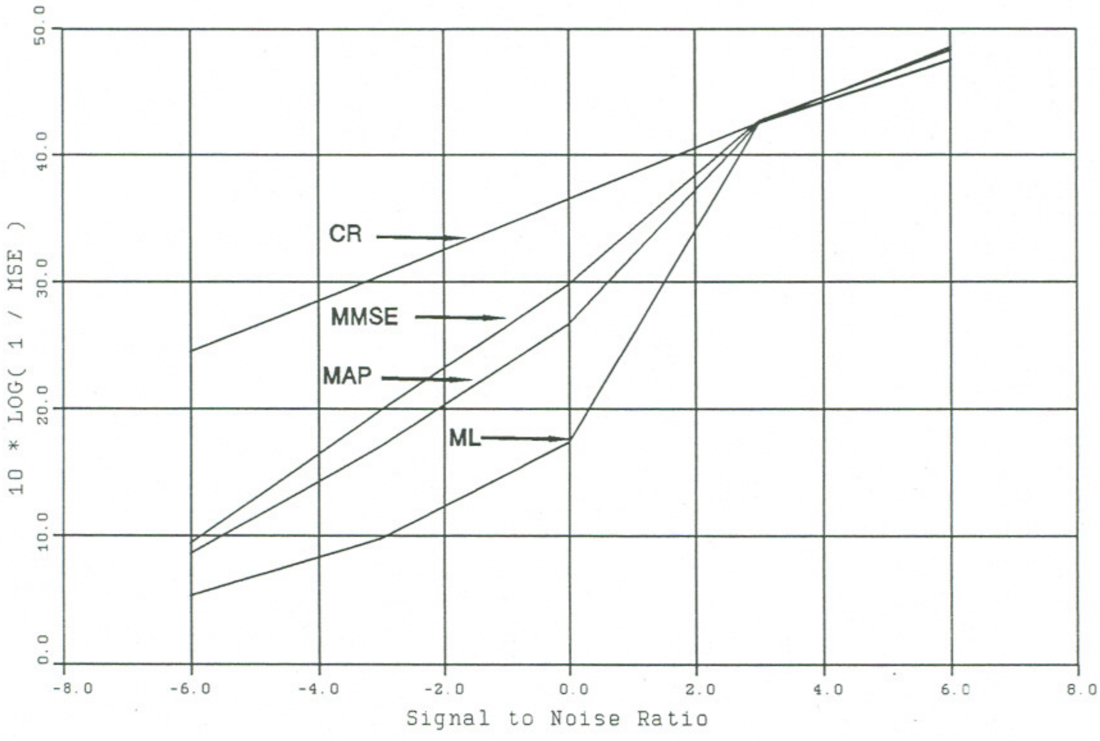

Figure 5: MSE vs. SNR for single cisoid, over all frequencies. N = 25 and r^2 = .8

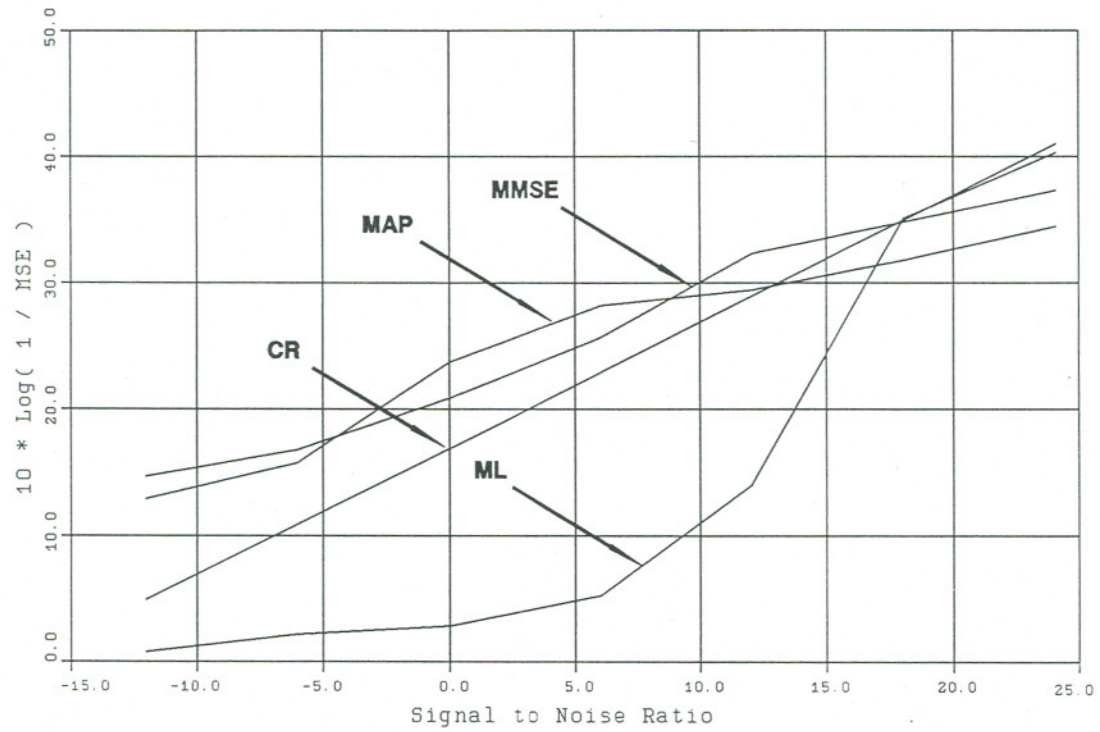

Figure 6: MSE vs. SNR for two closely spaced frequencies at .5 and .52 PI. N = 25, r^2 = .8 with 100 trials for each of 7 SNRs.