NOVEL PARAMETER PRIORS FOR BAYESIAN SIGNAL IDENTIFICATION

Anthony Quinn

Department of Electronic and Electrical Engineering, University of Dublin, Trinity College, Dublin 2, Ireland.

Phone: +353-1-6081863/6081580; Fax: +353-1-6772442; Email: aquinn@ee.tcd.ie

ABSTRACT

The problem of eliciting priors on the parameter space of a signal hypothesis is considered in this paper, and two lesser-known approaches are emphasized. Each yields conservative priors appropriate for data-dominated Bayesian parameter inference. They are based, respectively, on the principles of (i) a posteriori transformation invariance, and (ii) a priori maximum entropy. Novel priors on a wide class of signal models are deduced. Their ability to regularize inference of the difference frequency between closely spaced tones is considered, and they are compared with the Ockham Prior which was studied in previous work.

1. INTRODUCTION

In the Bayesian approach, inferences concerning unknown parameters are drawn from the a posteriori density, which is formed, from Bayes’ Theorem [1, 2], as the product of the Likelihood Function (LF) and the prior. It is the problem of assigning prior distributions in signal identification which is addressed in this paper. The prior not only quantifies any relevant information available before observations are made, but it can also be used to modify inferences so that they behave appropriately (Section 2.1).

The problem of prior elicitation remains the single most contentious area in all of mathematical inference. Many criteria have been proposed in the literature [1–7], some of which have been used in Bayesian signal identification:

(i) Bayes’ Postulate: this states that a uniform prior should be assigned when no prior information is available. Inferences are then drawn exclusively from the LF.

(ii) Jeffreys’ Priors: \( \Theta = \theta \) is a location parameter if the LF has the form

\[
p(x \mid \theta) = h(x - \theta).
\]

In this case, Jeffreys’ Prior for \( \Theta \) is uniform, as in (i). \( \Sigma = \sigma \) is a scale parameter if

\[
p(x \mid \sigma) = \frac{1}{\sigma} h\left(\frac{x}{\sigma}\right),
\]

for which Jeffreys’ Prior is \( p(\sigma) \propto 1/\sigma \). The former is employed widely as the prior for ‘signal’ parameters, the latter as the prior on the standard deviation of the additive ‘noise’ [1, 5, 8].

(iii) Conjugate Priors: the prior, \( q(\theta) \), is conjugate to the LF, \( p(x \mid \theta) \), if the resulting posterior, \( \pi(\theta \mid x) \), has the same form as \( q(\cdot) \). This property is useful in ensuring tractability in problems of sequential inference [2].

The sequel defines the problem of assigning priors consistently (i.e., “elicitation”), and presents two strategies which are not widely known in the signal identification literature.

2. PRIOR ELICITATION

Consider a parametric hypothesis, \( I \), which seeks to explain a set of observations, \( D = d \in \mathbb{C}^N \), in terms of a set of unknown parameters, \( \Theta = \theta \in \mathbb{C} \times \mathbb{C}^N \), where \( \mathbb{C} \) denotes the space of \( \theta \), and notation of the type \( X = x \) signifies a realization, \( x \), for the probabilistic parameter (p.p.) \( [\theta] \) vector, \( X \). \( D \) may be viewed as a vector of \( N \) discrete samples from a random process. The direct probability of \( d \) given candidate values of \( \theta \), i.e. \( p(d \mid \theta, I) \), may be interpreted as a function, \( l(\theta \mid d, I) \), of \( \theta \) itself, being the Likelihood Function (LF). In this paper, \( I \) additively decomposes \( d \) into a ‘signal’, \( s = g(\theta) \), and ‘noise’, \( e \):

\[
d = g(\theta) + e,
\]

where

\[
e \sim p_E(e \mid I).
\]

‘\( \sim \)’ denotes ‘is distributed as’, and \( g(\cdot) \) is the sampled signal parameter, parameterized by \( \theta \). It will be assumed that the analytical form of this function is known, so that the signal identification problem reduces to one of estimating \( \Theta \). The framework is readily extended to the case of model uncertainty [5, 9]. From (1, 2), the LF is given by

\[
p(d \mid \theta, I) \equiv l(\theta \mid d, I) = p_E(d - g(\theta) \mid I).
\]

Much of the orthodox signal identification literature is based on inferences from the LF. In contrast, the Bayesian approach recognizes that degrees of belief in a proposition are quantified by Probability. For example, the proposition that the (unknown) parameters, \( \Theta \), take on value \( \theta \)—i.e., \( \Theta = \theta \)—has a belief level given by \( p(\theta \mid d, I) \) once \( d \) is observed. This is the a posteriori (AP) pdf of \( \Theta \) [1–3, 5].
Bayes’ Rule inverts the direct probability (3) to yield the required AP pdf:

\[ p(\theta \mid d, I) \propto p(d \mid \theta, I)p(\theta) \mid I). \]  

(4)

\( p(\theta \mid I) \) is the prior on \( \theta \). Data-dominated (i.e., objective) inference may be achieved using uniform or diffuse priors, as discussed in Section 1. Then, \( p(\theta \mid d, I) \sim \ln([d \mid \theta, I]) \) from (4), but this fails to regularize the inference, as discussed next.

### 2.1. Prior Regularization

Normalizability of \( p_\theta(e \mid I) \) (2) implies that \( p_\theta(e \mid I) = f(\|e\|) \), where \( f(\cdot) \) is monotonically decreasing and \( \|e\| \) denotes some norm of \( e \). For example, the (complex) Gaussian noise pdf—to be explored in this paper—is \( p_\theta(e \mid I) \propto \exp(-e^H \Sigma^{-1} e) \), where \( e^H \Sigma^{-1} e \) is a Mahalanobis norm. The Maximum Likelihood (ML) estimate, from (3), is given by \( \hat{\theta}_{ML} \equiv \arg \min \|d - g(\theta)\| \). The criterion may overfit \( d \) with \( g(\cdot) \), particularly if \( k \ll N \). This manifests itself as thresholds in estimation and excessive order in model selection [5, 9]. Prior regularization can be accomplished via an appropriate non-uniform prior, to yield a Maximum a Posteriori (MAP) criterion of the type

\[ \hat{\theta}_{MAP} \equiv \arg \min_{\theta} \|d - g(\theta)\| + \beta(\theta), \]  

(5)

where \( \beta(\theta) \) is introduced via the prior (4). In tasks such as image segmentation [10], for example, Markovian priors may be adopted to encourage connectivity in the inferred label field. Two principles which yield novel regularizing priors for the signal identification task (1) are now presented.

### 3. Transformation Invariant Inference

Let \( h: \Omega \subseteq \mathbb{C} \rightarrow \Omega' \subseteq \mathbb{C}' \mid \eta = h(\theta) \) be a complex, bijective, holomorphic transformation with continuous first partial derivatives. The gradient matrix of the transformation is denoted by \( \nabla_{\theta} \eta^T \), whose determinant is assumed to be non-zero, \( \forall \theta \in \Omega \). Denoting the AP inference for \( \Theta \) by \( p(\theta \mid d, I) \) (assumed positive \( \forall \theta \)), then the inference for \( \eta \) is [1, 3, 4]

\[ q(\eta \mid d, I) \propto \frac{1}{|\nabla_{\theta} \eta^T|^2} p\bigl[h^{-1}(\eta) \mid d, I\bigr]. \]  

(6)

where \( h^{-1}(\cdot) \) denotes the inverse transformation. The determinant term, \( |\cdot| \), which is squared since the transformation is complex, is the hypervolume element expansion ratio for \( h(\cdot) \), and arises in (6) because of the normalization requirement of probabilities.

One consequence of (6) is that Bayesian inferences—as opposed to inferences based on non-measure functions such as the LF [5]—are not, in general, invariant with respect to a transformation. For example, the Maximum a Posteriori (MAP) estimate will not satisfy \( \eta_{MAP} = h(\theta_{MAP}) \), which is desirable. Thus, the MAP estimate of \( f \) in the signal model \( \sin(2\pi fn) \) will not be consistent with that of \( T \) in \( \sin\left(\frac{2\pi n}{T}\right) \), under the transformation \( T = 1/f \). However, the following theorem, based on work by Jeffreys [4], provides a sufficient condition for transformation invariance to be met.

**Theorem 1** Consider hypothesis \( I \), which explains observations \( d \in \Omega' \) in terms of parameters \( \Theta = \theta \in \Omega \subseteq \mathbb{C} \), via the LF, \( l(\theta \mid d, I) \). Let \( \eta = h(\theta) \) belong to the transformation class defined above. Then AP inferences are invariant with respect to \( h(\cdot) \) if the following prior is used:

\[ p(\theta \mid I) \propto |J(\theta \mid I)|. \]  

(7)

\( J(\theta \mid I) \) is the Fisher information matrix under \( I \), whose elements are

\[ J_{ij}(\theta \mid I) = -\frac{\partial^2 \ln l(\theta \mid I)}{\partial \theta_i \partial \theta_j} \]  

(8)

• A proof is given in [7]. The mild regularity conditions which must be met by \( l(\theta \mid d, I) \) are given in [5].

• The appropriate invariance prior for real parameter transformations is \( p(\theta \mid I) \sim |\nabla_{\theta} \eta^T|^T \).

#### 3.1. Data-Translated Likelihood

If the LF can be expressed in the form

\[ l(\theta \mid d, I) = q [\phi(\theta) - f(d)], \]  

(9)

then it is data-translated [1] in the transformed space, \( \Omega' \), of \( \phi \), since \( d \) influences only the location, and not the shape, of the LF. While this property often holds in the single parameter case for asymptotically large samples, it is necessary to weaken the definition in the multiparameter case to one of hypervolume invariance of the LF. It has been shown [1] that the prior (7) is consistent with the assignment of a uniform prior in the space, \( \Omega \), where this hypervolume invariance is achieved. This confers another justification for choosing (7), though it is important to recognize the weakness of this invariance constraint [4].

#### 3.2. Prior Stochastic Independence

If \( \Theta = (\psi^T, \Phi^T)^T \), and stochastic independence between \( \psi \) and \( \Phi \) is to be imposed a priori, then the prior factorization

\[ p(\theta \mid k, l, I) \propto J_1(\psi \mid k, I)J_2(\Phi \mid l, I) \]  

(10)

secures AP invariance under the decoupled transformations \( \eta_1 = h_1(\psi), \eta_2 = h_2(\Phi), \) where \( J_1(\psi \mid k, I) \) is calculated from \( l(\psi \mid k, I) \) via (8), and \( J_2(\Phi \mid l, I) \) is calculated from \( l(\Phi \mid l, I) \). \( k \) and \( l \) are the necessary hyperparameters of the prior [5].
4. INVARIANCE PRIOR FOR A WIDE SIGNAL CLASS

Consider the hypothesis, $I$, which analyzes observations, $d[n]$, in terms of $m$ basis functions, $G_i[n, \omega], \omega \in \mathbb{R}$:

$$d[n] = \sum_{i=1}^{m} b_i G_i[n, \omega] + \epsilon[n], \quad n = 0, \ldots, N - 1. \tag{11}$$

$b_i$ are the linear coefficients, and $\epsilon[n]$ are the residuals (‘noise’). This rich class [5,9,11] embraces a wide range of signal analysis techniques as special cases, including Fourier and wavelet analyses, etc. Gathering the $N$ observations into vector-matrix form, then

$$\mathbf{d} = \mathbf{s} + \mathbf{e} = \mathbf{G} (\omega) \mathbf{b} + \mathbf{e}, \tag{12}$$

where $\mathbf{G}(\omega) \in \mathbb{C}^{N \times m}$ with $(n,k)$th element $G_k[n,\omega]$, and $\mathbf{e} \sim \mathcal{N}(0, \Sigma)$, being complex, zero-mean, Gaussian noise with known covariance matrix $\Sigma$ (see Section 2.1). From (3,12):

$$l(\omega, \mathbf{b} | \mathbf{d}, \Sigma, I) \propto \exp \left[ -(\mathbf{d} - \mathbf{Gb})^H \Sigma^{-1} (\mathbf{d} - \mathbf{Gb}) \right], \tag{13}$$

where $\mathbf{G} \equiv \mathbf{G}(\omega)$ for convenience. From (7,8,13), matrix calculus yields the following prior for the signal parameter space [5]:

$$p(\omega, \mathbf{b} | \Sigma, I) \propto \left[ b^H \frac{\partial \mathbf{G}^H}{\partial \omega_i} \Sigma^{-1} \frac{\partial \mathbf{G}}{\partial \omega_j} \mathbf{b} - b^H \frac{\partial \mathbf{G}^H}{\partial \omega_i} \Sigma^{-1} \mathbf{G} \frac{\partial \mathbf{G}}{\partial \omega_j} \mathbf{b} - j b^H \frac{\partial \mathbf{G}^H}{\partial \omega_i} \Sigma^{-1} \mathbf{G} \frac{\partial \mathbf{G}}{\partial \omega_j} \mathbf{b} + j b^H \frac{\partial \mathbf{G}^H}{\partial \omega_i} \Sigma^{-1} \mathbf{G} \frac{\partial \mathbf{G}}{\partial \omega_j} \mathbf{b} \right]^\frac{1}{2}, \tag{14}$$

where $\alpha$ is a Lagrange multiplier. Standard manipulations yield $\alpha = m/N \gamma^2$ [5].

- It has been shown [5,11] that (19) is similar to the Ockham Prior for the signal structure (12); i.e., it favours objectively ‘simpler’ models, thereby regularizing the AP inference for $\omega$ and $\mathbf{b}$, in the sense explained by (5);

- (19) may be generalized to the case where there is testable information concerning signal autocorrelations to lag $N - 1$, i.e.,

$$\mathcal{E}[\text{ss}^H] = \mathbf{C}. \tag{20}$$

$s = \mathbf{Gb}$ (12) and $\mathbf{C} \in \mathbb{C}^{N \times N}$ is the known Hermitian correlation matrix for the signal. Maximization of (17), constrained by (20), yields the following entropic prior [5]:

$$p(\omega, \mathbf{b} | \mathbf{C}, I) \propto \exp \left[ -b^H \mathbf{G}^H \mathbf{C}^{-1} \mathbf{Gb} \right]. \tag{21}$$

- Prior independence is not engendered by the priors (19,21). If the second moments of $\mathbf{b}$ and $\omega$ are available independently a priori, then conventional independent Gaussian priors are obtained on the spaces, $\mathbb{R}$ and $\mathbb{W}$, of $\mathbf{b}$ and $\omega$ respectively, using the procedure above [5].
6. SUPERRESOLUTION OF CLOSELY SPACED FREQUENCIES

Consider the classic problem of inferring the difference frequency, $\omega$, in the following signal model:

$$d[n] = e^{j\omega n} \left[ b_1 + b_2 e^{j\omega n} \right] + e[n], \quad n = 0, \ldots, N - 1,$$

where $e[n] = e[n] \sim N(0, \sigma^2), \forall n$. $\omega_1$ is known a priori and $b_1, b_2$ and $\omega$ are to be inferred. In the sequel, marginal a priori densities for $\omega$,

$$p(\omega | x, I) = \int p(\omega, b | x, I) db,$$

are compared, where $x$ denotes any necessary hyperparameters, and the integrand on the right is the joint parameter prior which will be assigned using the strategies outlined in this paper.

6.1. Invariance Prior

Assuming $\omega$ and $b$ are independent a priori, then, from (15) or (16), $p(\omega | \sigma, k, I) \propto p(\omega, b | \sigma, k, I)$ (assuming a range constraint on $b$). Substituting (22) (via (11) and (12) into (15) or (16), noting that $\omega = \omega$ (a scalar) in this case, and letting $\Sigma = \sigma^2 I_N$ (i.e., diagonal), then $p(\omega | \sigma, k, I) \propto \text{const}.$ (see Fig. 1). In the Figure, one DFT bin is defined to be $2\pi/N$ rad/sample. If independence is relaxed a priori, then (14) must be employed, yielding a prior which is also illustrated in Fig. 1. It behaves similarly to the one proposed in [12], being approximately uniform for $\omega > 1$ DFT bin, and small for $\omega < 1$ bin, thereby eliminating the possibility of model rank deficiency a priori.

6.2. Entropic Prior

From (19), the marginal entropic prior for $\omega$ is [5]

$$p(\omega | \alpha, I) = \int p(\omega, b | \alpha, I) db \propto \left| G^H G \right|$$

$$= \left( N^2 - \sin^2 \left( \frac{Na}{2} \right) / \sin^2 \left( \frac{N\omega}{2} \right) \right)^{-1},$$

assuming $\bar{G} = \bar{G}^T$. This is also plotted in Fig. 1. As a regularizing prior, it outperforms the others; i.e., it is large for $\omega < 1$ bin, and has minima at integer DFT bins, $\omega = k\pi/N, k \in \mathbb{Z}$ (i.e., where the basis functions of (22) are orthogonal [5]). As such, this prior engenders Occam characteristics in the inference: if data support for the hypothesis (22) is insufficient, (23) dominates the AP inference, $p(\omega | d, \sigma, I)$, because of the singularity at $\omega = 0$, and this encourages the rejection of the hypothesis. If data support is sufficient, the LF will dominate the prior (23), and good estimates will result. As mentioned earlier, (19) and (23) have been deduced, from fundamentals, as the Occam Prior, whose excellent regularization properties, in both Estimation and Model Selection, have been explored in simulations [5, 9, 11]. Robust, threshold-free inferences are possible over a wide range of Signal-to-Noise Ratios (SNRs) and observation window lengths ($N$).

Figure 1: Invariance and Entropic Priors for $\omega$ in the 2-Cis Model (22).

References


