A Bifocal Measure of Expected Ambiguity in Bayesian Nonlinear Parameter Estimation

Emanuel WINTERFORS

Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 75252 Paris, Cedex 05 France
(*winterfors@ann.jussieu.fr*)
and
Grant Institute of Earth Science, School of Geoscience, Edinburgh University
Edinburgh EH9 3JW, United Kingdom

Andrew CURTIS

Grant Institute of Earth Science, School of Geoscience, Edinburgh University
Edinburgh EH9 3JW, United Kingdom
(*andrew.curtis@ed.ac.uk*)
and
ECOSSE (Edinburgh Collaborative of Subsurface Science and Engineering)
Edinburgh, United Kingdom

We present a novel approach to define and calculate the expected uncertainty of Bayesian parameter estimates, prior to collecting any observational data. This can be used to design investigation techniques or experiments that minimize expected uncertainty. Our approach accounts fully for nonlinearity in the parameter–observation relationship, which is neither the case for the Bayesian D- and A-optimality criteria most commonly used in experimental design, nor the case for most other derivative- or information matrix-based experimental design techniques. Our method is based on analyzing pairs of parameter estimates, thus forming a "bifocal" measure of ambiguity. Derivatives of observable data with respect to parameter values are neither required nor calculated. For linear models, our new measure is equivalent to expected posterior variance, and it is closely related to expected posterior variance in nonlinear models.

KEY WORDS: Bayesian methods; Decision theory; Frequency estimation; Microseismic location; Nonlinear models; Optimal design.

1. INTRODUCTION

Computationally efficient methods to estimate the expected uncertainty in Bayesian parameter estimates (prior to data collection) have been developed in the field of statistical experimental design. One can broadly divide approaches to optimal design of experiments into three categories:

(a) Linear methods
(b) Methods that invoke approximation by local linearization (e.g., Bayesian D- or A-optimality)
(c) Methods that do not use any linear approximations

Linear methods (a) are only applicable to linear models, but are by far the best studied. They are the subject of several books, for example Fedorov (1972), Silvey (1980), and Pukelsheim (1993). For nonlinear models, the vast majority of existing work is in category (b), which uses local derivatives (linearization) of the parameter–observation relationship to estimate uncertainty in parameter estimates, averaged with respect to a prior probability distribution over parameter space. A good review of these methods can be found in Chaloner and Verdinelli (1995). However, derivative-based design methods do not work for all nonlinear models, as demonstrated by the following example.

Consider a single-hand stopwatch where one wants to choose the angular velocity of the hand for optimal timing precision in a given time interval of $\Delta\theta = 60$ sec. The time θ is deduced from the angle y of the single stopwatch hand, which can be read with some uncertainty of $\pm\varepsilon_y$ degrees ($^\circ$). As long as the hand will not complete more than a full 360° turn in 60 sec, the uncertainty in time can be calculated as $\varepsilon_\theta = (dy/d\theta)^{-1}\varepsilon_y$, implying that time uncertainty will be lower, the higher the angular velocity $dy/d\theta$ used.

All derivative- or information matrix-based methods for estimating parameter uncertainty work by analyzing only the derivative $dy/d\theta$ at one or many points along the curve defined by $y(\theta)$ (see Figure 1). Such methods would here predict lowest uncertainty when using the highest angular velocity that the watch motor can produce. This clearly becomes problematic if the hand may have completed more than one 360° turn before being stopped. Then, for any angle measurement $y \pm \varepsilon_y$, several quite different time readings $\theta \pm \varepsilon_\theta$ are possible (Figure 1), leading to ambiguity in θ, with a total variance that *increases* monotonically with the magnitude of the derivative $dy/d\theta$. Any derivative- or information matrix-based method will select the design that maximizes the magnitude of $dy/d\theta$, and thus, will result in the highest possible variance and hence the worst possible experimental design.

The example above is chosen as the simplest possible case for which derivative-based nonlinear design methods fail to find the optimal design, but many others exist, such as Examples 4.1 and 4.2.

Methods in category (c) do not rely on local linearization and can correctly detect and account for parameter ambiguity, as in Figure 1. Such methods have been studied by DeGroot (1984), Müller and Parmigiani (1995), and van den Berg, Curtis, and

TECHNOMETRICS, MAY 2012, VOL. 54, NO. 2
DOI: 10.1080/00401706.2012.676953

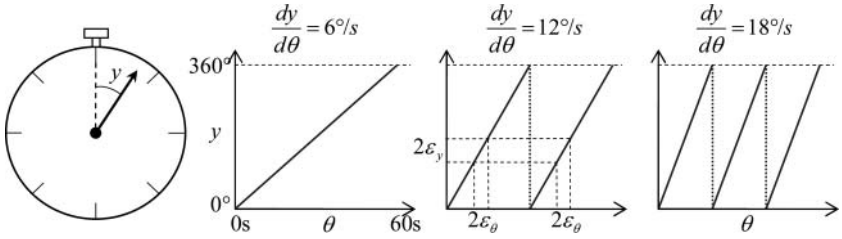

Figure 1. Schematic example of how high angular velocity of a single hand stopwatch can lead to ambiguous time readings. The central panel shows an interval $\pm\varepsilon_y$ around an observed angle y and two intervals $\pm\varepsilon_\theta$ around possible (but ambiguous) time readings.

Trampert (2003), among others. These methods are applicable to a wider class of problems than those in categories (a) and (b), but typically require substantially more computation than methods based on linearization. Therefore, design of problems involving more than two or three parameters and observable variables have in the past been numerically intractable.

Herein, we present a more efficient method to estimate expected parameter uncertainty without using linearization. The resulting ambiguity measure is based on a novel *bifocal* approach, replacing the advantages provided by local linearization with advantages afforded by analyzing pairs of parameter estimates simultaneously. We show that this results in a measure that is intuitive, efficiently calculable, and which accounts for full nonlinearity of any model, while being equivalent to the expected variance when applied to linear models with normally distributed uncertainties.

Section 2 introduces a formal definition of the observational ambiguity measure and relates it to other commonly used uncertainty measures. Section 3 deals with numerical aspects of evaluating the new measure using Monte Carlo (MC) methods and of optimizing the measure with respect to the design of an investigation technique. Two example applications can be found in Section 4: in one, other fully nonlinear uncertainty measures are evaluated for comparison; in the other, this is impossible—only our measure is computationally tractable to evaluate and optimize.

2. CONCEPTS AND FORMALISM

Consider a set of *parameters* θ one wishes to obtain knowledge about, and let Θ be the set of all values that θ can take. Through some *investigation technique*/experiment one can make *observations* y, related to θ through a known relationship predicting a probability measure over the set Ω of all possible observations, represented by a probability density $p(y|\theta,\xi)$, conditional on the parameters θ and the design ξ of the investigation technique.

Any preobservation knowledge of the actual values of the parameters θ is represented by a *prior* probability density $p(\theta)$ over the *parameter space* Θ, here assumed to be independent of the design ξ. After having made an observation y (i.e., acquired data), the state of knowledge about θ can be updated through Bayes' theorem, calculating a *posterior* probability density

$$p(\theta|y,\xi) = \frac{p(y|\theta,\xi)p(\theta)}{p(y|\xi)}, \tag{1}$$

where $p(y|\xi)$ is the marginal probability density over observation space

$$p(y|\xi) = \int_{\theta\in\Theta} p(y|\theta,\xi)p(\theta)d\theta. \tag{2}$$

2.1 Bifocal Analysis of Ambiguity

Consider two different points $\dot\theta \neq \ddot\theta$ in parameter space Θ, where the dots are used as identifiers only and do not denote time derivatives. It is then possible to define various measures of how likely these are to give rise to the same observation—and hence be indistinguishable given such an observation. This is determined by the extent to which their respective probability densities $p(y|\dot\theta,\xi)$ and $p(y|\ddot\theta,\xi)$ overlap (see Figure 2). The

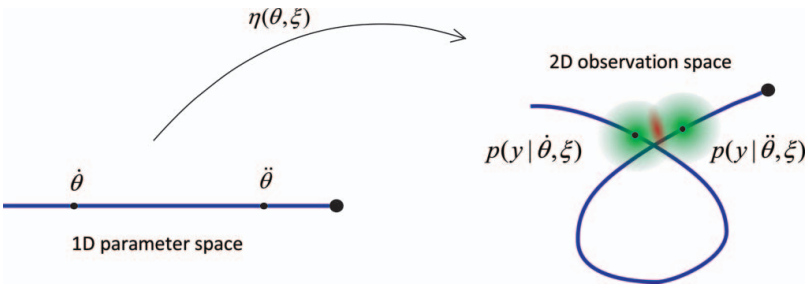

Figure 2. Schematic example problem of a single parameter θ to be estimated from two observed variables $[y^1, y^2]^T = y$ with normally distributed observational uncertainties. The bold line and curve show the finite parameter range (left) being mapped into the two-dimensional observation space (right) by $\eta(\theta,\xi) = \int yp(y|\theta,\xi)dy$. Two parameter values, $\dot\theta$ and $\ddot\theta$, are shown, for which $p(y|\dot\theta,\xi)$ and $p(y|\ddot\theta,\xi)$ (lightly shaded) partly overlap, so observations within the overlap area (darker shaded) are ambiguous and incapable of discriminating between parameter values $\dot\theta$ and $\ddot\theta$. The online version of this figure is in color.

most straightforward such measure is

$$Q(\dot{\theta}, \ddot{\theta}, \xi) = \int_{y \in \Omega} p(y|\dot{\theta}, \xi) p(y|\ddot{\theta}, \xi) dy. \quad (3)$$

When designing an investigation technique, $Q(\dot{\theta}, \ddot{\theta}, \xi)$ should ideally be small when $\dot{\theta}$ and $\ddot{\theta}$ are far apart. Equation (3) defines a *bifocal* measure, simultaneously focusing on two points $(\dot{\theta}, \ddot{\theta})$ in parameter space instead of only one, which is the most common approach. Bingham and Chipman (2007) studied a similar measure but instead integrated the square root of the product of the probability densities.

This measure does, however, have two disadvantages: first, $Q(\dot{\theta}, \ddot{\theta}, \xi)$ is always high for $\dot{\theta} = \ddot{\theta}$ even though this case does not contribute to uncertainty in estimates of the model parameters θ. Second, the units of $Q(\dot{\theta}, \ddot{\theta}, \xi)$ are the same as those of a probability density $p(y)$ over observation space, implying that $Q(\dot{\theta}, \ddot{\theta}, \xi)$ will, on average, increase with decreased observational uncertainty (the opposite should ideally be the case).

One way to overcome the first problem is to multiply $Q(\dot{\theta}, \ddot{\theta}, \xi)$ by the squared distance between $\dot{\theta}$ and $\ddot{\theta}$ (assuming that Θ is equipped with a metric d). The second problem can be addressed by dividing $Q(\dot{\theta}, \ddot{\theta}, \xi)$ by a measure D_Ω of average observational probability density

$$D_\Omega = \int_{y \in \Omega} p(y|\xi) p(y|\xi) dy. \quad (4)$$

This gives the ambiguity measure

$$R(\dot{\theta}, \ddot{\theta}, \xi) = \frac{d^2(\dot{\theta}, \ddot{\theta})}{D_\Omega} Q(\dot{\theta}, \ddot{\theta}, \xi). \quad (5)$$

The approach of analyzing point pairs $(\dot{\theta}, \ddot{\theta})$ with overlapping, or "intersecting," conditional observation densities, as well as weighting them by squared distance, is a generalization of the nonprobabilistic *intersection point-pair* concept developed by Winterfors and Curtis (2008).

2.2 Measure of Expected Observational Ambiguity

To create a global measure of the ambiguity of an investigation technique with a given design ξ, it is necessary to take the expectation of $R(\dot{\theta}, \ddot{\theta}, \xi)$ over all possible point pairs in parameter space, with respect to the prior density $p(\theta)$:

$$W(\xi) = \int_{\dot{\theta} \in \Theta} \int_{\ddot{\theta} \in \Theta} p(\dot{\theta}) p(\ddot{\theta}) \quad R(\dot{\theta}, \ddot{\theta}, \xi) \quad d\ddot{\theta} d\dot{\theta}. \quad (6)$$

The resulting expected observational ambiguity $W(\xi)$ is thus a measure of the average ambiguity of all possible observations, given an investigation technique with design ξ and a prior distribution $p(\theta)$ over parameter values. It is crucial to note that this is not the same as the ambiguity of a particular posterior distribution $p(\theta|y, \xi)$ that is typically calculated after observations y have been made. $W(\xi)$ can therefore be used in planning and designing the investigation technique described by ξ, prior to acquisition of any observations.

2.3 Defining an Appropriate Metric

The choice of metric $d(\dot{\theta}, \ddot{\theta})$ used in (5) plays an important role in evaluating the quality of any investigation technique. As a general guideline, $d(\dot{\theta}, \ddot{\theta})$ should be chosen to be proportional to the need to discriminate between two points $\dot{\theta}$ and $\ddot{\theta}$ in parameter space.

For example, in the linear scope of Euclidian metrics, one might define

$$d^2(\dot{\theta}, \ddot{\theta}) = \Delta\theta^T \Sigma^{-1} \Delta\theta, \quad (7)$$

where $\Delta\theta = \dot{\theta} - \ddot{\theta}$ and Σ^{-1} is a diagonal matrix of weights representing the relative importance of each of the component parameters of θ. If there are parameters affecting the observations but whose values are of no interest (often called nuisance parameters, see Silvey 1980), their relative weights can simply be set to zero. The matrix Σ^{-1} can also be nondiagonal, in which case it can be seen as the inverse of a reference covariance matrix Σ in parameter space, to which the covariance of the posterior distributions should ideally be proportional.

2.4 Relations to Variance

The expected observational ambiguity $W(\xi)$ relates to the expected posterior variance in a simple manner, demonstrated by the following theorems. All proofs can be found in the Appendix.

Lemma 1. Inserting (5) into (6), applying Bayes' rule, and changing the order of integration (assuming that both Θ and Ω are complete measure spaces, $W(\xi) < \infty$, and $d^2(\dot{\theta}, \ddot{\theta}) p(\dot{\theta}, y|\xi) p(\ddot{\theta}, y|\xi)$ is $\Theta \times \Theta \times \Omega$-measurable) gives

$$W(\xi) = \frac{2}{D_\Omega} \int_{y \in \Omega} p^2(y|\xi) \; V[\Theta|y, \xi] \; dy, \quad (8)$$

where $V[\Theta|y, \xi]$ is a functional of the posterior distribution $p(\theta|y, \xi)$:

$$V[\Theta|y, \xi] = \frac{1}{2} \int_{\dot{\theta} \in \Theta} \int_{\ddot{\theta} \in \Theta} d^2(\dot{\theta}, \ddot{\theta}) p(\dot{\theta}|y, \xi) p(\ddot{\theta}|y, \xi) d\ddot{\theta} d\dot{\theta}. \quad (9)$$

Theorem 1. If the distance function d is the standard Euclidian distance (i.e.,Θ is a subset of a vector space), the functional $V[\Theta|y, \xi]$ is equal to the variance of $p(\theta|y, \xi)$:

$$V[\Theta|y, \xi] = \int_{\theta \in \Theta} p(\theta|y, \xi) d^2(\theta, \mu) \, d\theta, \quad (10)$$

where μ is the expectation of $p(\theta|y, \xi)$.

Since $V[\cdot]$ is well defined for any space with a distance function d, it can be viewed as a generalization of the variance to all metric spaces. $V[\cdot]$ can be shown to be invariant under changes of coordinates (as long as the metric is so), but this article will only make use of the fact that $V[\cdot]$ can be calculated without the need for explicitly calculating the expectation of the distribution.

The expected observational ambiguity $W(\xi)$, as expressed in (8), is thus twice the expected variance of the posterior, weighted by $p(y|\xi)$ corresponding to the marginal probability density in observation space. Equivalently, $W(\xi)$ can be interpreted as

twice the expected posterior variance of a modified joint probability distribution $p\prime(\theta, y|\xi)$ with posteriors identical to those of $p(\theta, y|\xi)$, but with a modified marginal distribution over observation space:

$$W(\xi) = 2 \int_{y\in\Omega} p'(y|\xi)\ V[\Theta|y,\xi]\ dy, \tag{11}$$

$$p'(y|\xi) = \frac{p^2(y|\xi)}{\int_{y\in\Omega} p^2(y|\xi)dy}. \tag{12}$$

Using this modified marginal distribution can be thought of as emphasizing the observations with the highest probability at the expense of those with a low probability.

In linear models with normally distributed prior and observational uncertainties, the posterior variance $V[\Theta|y,\xi]$ will be the independent of the observation y. The inverse posterior covariance matrix $\Sigma^{-1}_{\Theta|y}$ can be calculated

$$\Sigma^{-1}_{\Theta|y} = \Sigma^{-1}_{\Theta} + \mathbf{J}^{\mathrm{T}}\Sigma^{-1}_{\Omega|\theta}\mathbf{J}, \tag{13}$$

where $\mathbf{J}$ is the Jacobian of the linear model, and Σ^{-1}_{Θ} and $\Sigma^{-1}_{\Omega|\theta}$ are inverse covariance matrices for the prior and observational uncertainty distributions, respectively. Setting $\Sigma^{-1}_{\Theta} = 0$ corresponds to the case where no prior information is present, in which case $\Sigma^{-1}_{\Theta|y}$ is equal to the Fisher information matrix. Taking the expectation with respect to any distribution $p(y)$, expected posterior variance will (trivially, since it is constant) also be equal to (13), implying that $W(\xi)/2$ equals the expected variance for models with normally distributed prior and observational uncertainties.

2.5 Relations to Shannon Entropy

Another commonly used measure of the uncertainty represented by any probability density $p(\theta)$ is the Shannon entropy (sometimes called differential Shannon entropy to distinguish from the case of a probability distribution over a discrete space)

$$H[\Theta] = -\int_{\theta\in\Theta} p(\theta)\ln(p(\theta))\,d\theta. \tag{14}$$

The expected Shannon entropy of posterior probability densities can then be defined as

$$H[\Theta|\Omega,\xi] = -\int_{y\in\Omega}\int_{\theta\in\Theta} p(\theta, y|\xi)\ln(p(\theta|y,\xi))\,d\theta dy. \tag{15}$$

Minimizing $H[\Theta|\Omega,\xi]$ with respect to design ξ will maximize the expected decrease in Shannon entropy of the investigation technique. Sebastiani and Wynn (2000) demonstrated that $H[\Theta|\Omega,\xi]$ can also be calculated using

$$H[\Theta|\Omega,\xi] = H[\Omega|\Theta,\xi] + H[\Theta] - H[\Omega|\xi], \tag{16}$$

where $H[\Omega|\Theta,\xi]$ is the expected entropy of the observational uncertainty, $H[\Theta]$ the entropy of the prior distribution, and $H[\Omega|\xi]$ the entropy of the marginal distribution in observation space. If the observational entropy is independent of ξ, so is its expectation $H[\Omega|\Theta,\xi]$. In such cases, the expected posterior entropy is only dependent on the design ξ through the marginal observational entropy $H[\Omega|\xi]$. Therefore, minimizing the expected posterior entropy requires maximizing $H[\Omega|\xi]$. Finding an optimal design ξ in this way is called maximum entropy sampling (Sebastiani and Wynn 2000).

The variance provides an upper bound on the Shannon entropy through the maximum entropy principle. When $\Theta = \mathbb{R}^n$, this inequality takes the form (Dembo, Cover, and Thomas 1991)

$$V[\Theta] \geq \frac{n}{2\pi e}\exp\left(\frac{2}{n}H[\Theta]\right), \tag{17}$$

where $V[\Theta]$ and $H[\Theta]$ denote the variance and entropy of any distribution over Θ, respectively. Left- and right-hand sides of (17) are equal only for normal distributions, with covariance matrix proportional to the identity matrix, in the following text referred to as *spherical* normal distributions. Writing the expected variance of the posterior as $V[\Theta|\Omega,\xi] = \int p(y|\xi)V[\Theta|y,\xi]dy$, Theorem 2 is easily derived:

Theorem 2. For the expected variance and expected Shannon entropy of the posterior probability distribution over $\Theta \subseteq \mathbb{R}^n$, the following inequality holds

$$V[\Theta|\Omega,\xi] \geq \frac{n}{2\pi e}\exp\left(\frac{2}{n}H[\Theta|\Omega,\xi]\right), \tag{18}$$

with equality if and only if all posterior densities $p(\theta|y,\xi)$ are spherical normal distributions with identical entropy (or equivalently, identical variance).

Combining (18) with (16) and assuming that observational uncertainty is independent of design, one obtains

$$V[\Theta|\Omega,\xi] \geq K\exp\left(-\frac{2}{n}H[\Omega|\xi]\right), \tag{19}$$

where K is a constant with respect to ξ:

$$K = \frac{n}{2\pi e}\exp\left(\frac{2}{n}H[\Omega|\Theta] + H[\Theta]\right). \tag{20}$$

Evaluating the expected posterior variance thus imposes an upper bound on the expected posterior entropy. Both these measures are computationally expensive, requiring numerical integration over both Θ and Ω. If one could instead provide an upper bound on expected posterior entropy using $W(\xi)$, this would require less computation. Fortunately, there exists such an upper bound, as demonstrated by the following theorem.

Theorem 3. The expected observational ambiguity $W(\xi)$ is related to the expected Shannon entropy of the posterior distribution over $\Theta \subseteq \mathbb{R}^n$ through the inequality

$$W(\xi) \geq \frac{n}{\pi e}\frac{\exp(-H[\Omega|\xi])}{D_\Omega}\exp\left(\frac{2}{n}H[\Theta|\Omega,\xi]\right), \tag{21}$$

with equality if and only if all possible posterior densities are spherical normal distributions with identical entropy, and $p(y)$ is uniform (i.e., constant for all y where it is nonzero).

Using (16) and assuming that observational uncertainty is independent of design, (21) can be written

$$W(\xi) \geq 2\frac{K}{D_\Omega}\exp\left(-\frac{2+n}{n}H[\Omega|\xi]\right), \tag{22}$$

where K is the constant defined in (20). Theorem 2 and Theorem 3 again demonstrate the similarity of $W(\xi)$ to the expected posterior variance, also with respect to how they impose an upper bound on expected posterior Shannon entropy.

2.6 Comparison With Bayesian and Classical A- and D-Optimality

Consider an observation–parameter relationship that can be written in the form

$$y = \eta(\theta, \xi) + \varepsilon, \tag{23}$$

where $\eta(\theta, \xi)$ gives the expected observation, and ε is a random, zero-mean observational error.

If $\eta(\theta, \xi)$ is nonlinear with respect to θ, the standard approach is to approximate posterior covariance matrices $\Sigma_{\Theta|y}$ using local linearization of (23) around a point θ in Θ: $\eta(\theta + \Delta\theta, \xi) \approx \mathbf{J}\Delta\theta + \eta(\theta, \xi)$, and then approximate the posterior covariance $\Sigma_{\Theta|y}$ using (13), for $y \approx \eta(\theta, \xi)$. Some scalar function Φ of $\Sigma_{\Theta|y}$, typically its variance ($\mathrm{tr}[\Sigma_{\Theta|y}]$, often referred to as the A-optimality criterion) or decrease in differential Shannon entropy $(\log|\Sigma_\Theta| - \log|\Sigma_{\Theta|y}|)/2$, often referred to as the D-optimality criterion, is then integrated over the prior distribution to obtain an approximate design criterion:

$$Q_\Phi(\xi) = \int_{\theta\in\Theta} p(\theta)\Phi(\Sigma_{\Theta|y})d\theta. \tag{24}$$

Such an approach works well as long as posterior distributions are approximately Gaussian. This is clearly not the case if $\eta(\theta, \xi)$ is not fairly linear with respect to θ within the width of the observational error ε. One can check if such strong nonlinearity is present by studying the curvature of $\eta(\theta, \xi)$ with respect to θ (Clyde 1993), but doing so does unfortunately still not guarantee approximate normality of all posterior distributions. If there are self-intersections present in the image of Θ mapped by $\eta(\theta, \xi)$ into Ω (as in Figure 2), observations near such intersections will correspond to posterior distributions containing multiple peaks [as illustrated later in Figure 4(c) and Figure 5(b)]. Even if the individual peaks may be approximately Gaussian, the full distribution will have very different properties that are ignored in designs using the above criterions (see Example 4.1).

3. NUMERICAL EVALUATION AND OPTIMIZATION

Section 2 shows that the expected ambiguity measure $W(\xi)$ is a viable measure of expected posterior uncertainty. This section presents methods for estimating and optimizing $W(\xi)$. The algorithms presented are most efficient for cases where the integral in $R(\dot\theta, \ddot\theta, \xi)$, as defined in (5), can be evaluated analytically, for example, when the shape of the observational uncertainty $p(y|\theta, \xi)$ is known in some closed form such as a Gaussian, Poisson, exponential, or Gamma distribution.

The possibility of evaluating the integral over observation space Ω analytically is what enables $W(\xi)$ to be evaluated with much higher numerical efficiency than other fully nonlinear methods for estimating expected uncertainty of parameter estimates, where the expectation (i.e., integral) over Ω has to be estimated using more costly numerical methods. The following sections will therefore focus on this special case when observational uncertainty is known in some closed form

$$p(y|\theta, \xi) = p\left(y|\eta(\theta, \xi)\right), \tag{25}$$

where $p(y|\eta)$ is a known closed-form probability density function with parameters η (e.g., a Gaussian distribution with mean η). Its parameters $\eta(\theta, \xi)$ depend in turn on θ and ξ, but this dependence does not have to be known in closed form. For example, $\eta(\theta, \xi)$ can be defined to be the solution of some set of differential equations that have to be approximated numerically.

3.1 MC Estimation Strategies

Inserting (25) into (3) one obtains

$$Q(\dot\eta, \ddot\eta) = \int_{y\in\Omega} p(y|\dot\eta)p(y|\ddot\eta)\, dy, \tag{26}$$

where $\dot\eta = \eta(\dot\theta, \xi)$ and $\ddot\eta = \eta(\ddot\theta, \xi)$. Again, the dots are used for discrimination only and do not denote derivatives. Expressions for $Q(\dot\eta, \ddot\eta)$ for some common observational uncertainty distributions are shown in Table 1 (for proofs and definitions, see the Appendix). The expected observational ambiguity $W(\xi)$ can then be estimated by MC integration, by approximating (6) as a

Table 1. Expressions for $Q(\dot\eta, \ddot\eta)$, given common observational uncertainty distributions

Distribution	$p(y\|\eta)$	$Q(\dot\eta, \ddot\eta)$
Multivariate normal	$\frac{1}{K}\exp(-\frac{1}{2}d^2(y,\eta))$, where $K = \sqrt{(2\pi)^n\|\Sigma\|}$ and $d^2(y,\eta) = (y-\eta)^T\Sigma^{-1}(y-\eta)$	$\frac{\sqrt{2^n}}{K}\exp(-\frac{1}{4}d^2(\dot\eta,\ddot\eta))$
Exponential	$\frac{1}{\eta}\exp(-\frac{y}{\eta})$ $y, \eta \geq 0$	$\frac{1}{\dot\eta + \ddot\eta}$
Gamma	$\frac{k^k y^{k-1}}{\eta^k\Gamma(k)}\exp(-\frac{ky}{\eta})$ $y, \eta, k \geq 0$	$\frac{k\Gamma(k-\frac{1}{2})}{\sqrt{\pi}\Gamma(k)}\frac{(4\dot\eta\ddot\eta)^{k-1}}{(\dot\eta+\ddot\eta)^{2k-1}}$
Poisson	$\frac{\eta^y}{y!}e^{-\eta}$ $y \in \mathbb{N}_0, \eta > 0$	$e^{-\dot\eta-\ddot\eta}I_0(2\sqrt{\dot\eta\ddot\eta})$

discrete sum

$$\hat{W}(\xi) = \frac{2}{N^2 \hat{D}_\Omega} \sum_{i=1}^{N} \sum_{j=1}^{i-1} d^2(\dot{\theta}_i, \ddot{\theta}_j) Q(\dot{\eta}_i, \ddot{\eta}_j) \frac{p(\dot{\theta}_i)}{s(\dot{\theta}_i)} \frac{p(\ddot{\theta}_j)}{s(\ddot{\theta}_j)}, \quad (27)$$

$$\hat{D}_\Omega = \frac{2}{N^2 + N} \sum_{i=1}^{N} \sum_{j=1}^{i} Q(\dot{\eta}_i, \ddot{\eta}_j) \frac{p(\dot{\theta}_i)}{s(\dot{\theta}_i)} \frac{p(\ddot{\theta}_j)}{s(\ddot{\theta}_j)}, \quad (28)$$

where $\{\theta_1, \theta_2, \ldots, \theta_N\} \sim s(\theta)$ is a set of random points in Θ sampled according to some sampling probability distribution $s(\theta)$. The symmetry properties of d and Q and the fact that $d(\theta, \theta) = 0$ have also been used to reduce the number of terms in the sums. Furthermore, the function η can be pre-evaluated at the points $\{\theta_1, \theta_2, \ldots, \theta_N\}$ creating the set $\{\eta_1, \eta_2, \ldots, \eta_N\}$, allowing for it to be evaluated $O(N)$ rather than $O(N^2)$ times in (27) and (28), reducing computation required for evaluating functions that may be numerically costly.

The most simple choice of sampling distribution is $s(\theta) = p(\theta)$, but many other MC methods for sampling of $\Theta \times \Theta$ can be constructed that may be more efficient, depending on the shape and distribution of $d^2(\dot{\theta}, \ddot{\theta}) Q(\dot{\eta}, \ddot{\eta}) p(\dot{\theta}) p(\ddot{\theta})$. The choice of algorithm for optimizing $W(\xi)$ with respect to the design ξ might also influence the most suitable sampling scheme in $\Theta \times \Theta$.

3.2 Sampling Based on Metric and Prior

Taking the shape of $Q(\dot{\theta}, \ddot{\theta}, \xi)$ into account when computing a sampling distribution $s(\theta)$ may be computationally expensive, but it is possible to construct one that is better than using $s(\theta) = p(\theta)$ by accounting only for the metric $d^2(\dot{\theta}, \ddot{\theta})$. Treating $Q(\dot{\theta}, \ddot{\theta}, \xi)$ similar to a nonnormalized probability density, having no information about $Q(\dot{\theta}, \ddot{\theta}, \xi)$, can be represented by setting it to a constant $Q(\dot{\theta}, \ddot{\theta}, \xi) = Q_0$, yielding

$$\tilde{s}(\dot{\theta}) = Q_0 p(\dot{\theta}) \int_{\ddot{\theta} \in \Theta} d^2(\dot{\theta}, \ddot{\theta}) p(\ddot{\theta}) d\ddot{\theta}, \quad (29)$$

which can be normalized as $s(\dot{\theta}) = \tilde{s}(\dot{\theta})/\mathrm{M}$, where $\mathrm{M} = Q_0 \int \tilde{s}(\dot{\theta}) d\dot{\theta}$.

If Θ is a subspace of a vector space with metric $d^2(\dot{\theta}, \ddot{\theta}) = [\dot{\theta} - \ddot{\theta}]^2$ where the inner product is defined as $\theta^2 = \theta^T \Xi \theta$, Ξ being some positive definite matrix, $s(\dot{\theta})$ takes on the simple form

$$s(\dot{\theta}) = \frac{1}{2} p(\dot{\theta}) + \frac{1}{2\mathrm{tr}[\Sigma \Xi]} \left[\dot{\theta} - E[\Theta]\right]^2 p(\dot{\theta}), \quad (30)$$

where Σ is the covariance matrix of $p(\theta)$ and $E[\Theta]$ its mean.

3.3 Design Optimization

For the optimization of the design ξ, many different approaches are possible. In general, the optimization of an MC estimate of an objective function—inherently possessing some uncertainty—presents a number of new difficulties in addition to those encountered in the optimization of deterministic objective functions that can be evaluated exactly. This problem has been treated by several authors, such as Robbins and Monro (1951), Kiefer and Wolfowitz (1952), Spall (2003), and Chen (2002) in a general optimization setting, as well as by Merlé and Mentré (1997), Müller and Parmigiani (1995), and Müller, Sanso, and De Iorio (2004) for applications in nonlinear experimental design optimization.

The main problem arises from the difficulty in obtaining gradients of the objective function. If discrete differentiation is used, the uncertainty in the gradient estimate will be inversely proportional to the length of the discrete differentiation interval. It will thus approach infinity as the differentiation interval length approaches zero.

One way of reducing uncertainty in the gradient estimates is to avoid resampling in $\Theta \times \Theta$ for the differentiation in each of the components of ξ (Gaivoronski 1998):

$$\frac{\partial}{\partial \xi} \hat{W}(\xi) \approx \frac{2}{N^2 \hat{D}_\Omega} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \frac{p(\dot{\theta}_i)}{s(\dot{\theta}_i)} \frac{p(\ddot{\theta}_j)}{s(\ddot{\theta}_j)} (d^2(\dot{\theta}_i, \ddot{\theta}_j) - \hat{W}(\xi)) \times \frac{\partial}{\partial \xi} Q(\dot{\eta}_i, \ddot{\eta}_j), \quad (31)$$

$\hat{W}(\xi)$ and $\hat{D}_\Omega$ defined in (27) and (28). A commonly used similar approach is to keep the same sample $\{\theta_1, \theta_2, \ldots, \theta_N\}$ of points in parameter space throughout the whole optimization with respect to ξ, not doing resampling $\{\theta_1, \theta_2, \ldots, \theta_N\}$ for each new evaluation of $\hat{W}(\xi)$. This allows for the use of standard deterministic optimization algorithms, but renders error due to MC sampling difficult to detect.

In the examples that follow, we adopt the latter approach and acknowledge the possible existence of MC-related errors. To mitigate against the effects of these errors, as well as for the existence of multiple local minima, we perform each optimization several times with different random sampling methods to check that similar optimal solutions are found.

4. EXAMPLES

4.1 Frequency and Amplitude Estimation

A common problem arising in numerous domains of experimental research is to estimate the frequency and amplitude of some oscillation present in a signal. One may parameterize the signal as

$$y = \theta_1 \sin(\theta_2 t + \theta_3) + \varepsilon, \quad (32)$$

where θ_1 and θ_2 are amplitude and frequency, respectively. The phase θ_3 will be considered a nuisance parameter whose value is of no interest. ε is a random, zero-mean normally distributed observational error.

The design problem in this case is to determine at what values of t to observe y to minimize uncertainty in estimates of θ_1 and θ_2. The most well-known result on the subject is the Nyquist–Shannon sampling theorem (Nyquist 1928), stating that for θ_2 to be uniquely determined, the sampling frequency ξ must be at least two times the highest possible frequency of the signal:

$$\xi \geq 2 \max(\theta_2), \quad (33)$$

where $2 \max(\theta_2)$ is usually referred to as the Nyquist frequency/limit. However, this result does not account for uncertainty in observations y. It is also in direct conflict with derivative-based design theory, which rather suggests that sampling frequency should be as high as possible since that will

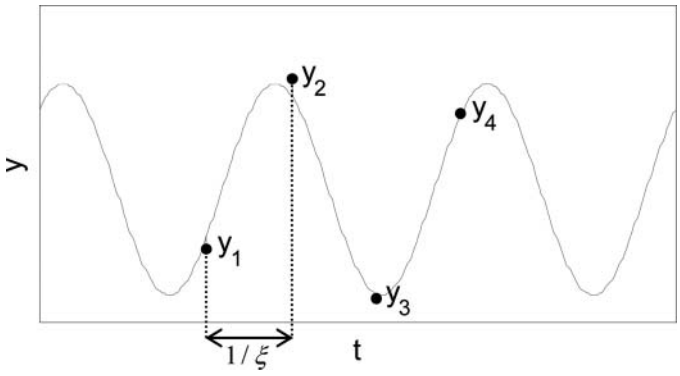

Figure 3. Example of the oscillating signal of (32) whose amplitude θ_1 and frequency θ_2 are to be determined by four regularly spaced noisy samples $y = [y_1, y_2, y_3, y_4]^T$ taken at a sampling frequency ξ. This is different from the example in the Introduction (Figure 1) in that the signal frequency θ_2 here is a sought parameter rather than a controllable design parameter.

maximize the sensitivity of the sampled values y to changes in the signal frequency θ_2.

Considering the case where the phase θ_3 of the signal is unknown and one wishes to determine the optimal sampling frequency ξ of only four samples $y = [y_1, y_2, y_3, y_4]^T$ regularly spaced by the time interval $1/\xi$, as shown in Figure 3. The expected observational ambiguity $W(\xi)$ was evaluated over a range of sampling frequencies $0\text{ Hz} < \xi < 6\text{ Hz}$, along with measures of expected posterior variance as well as expected gain in Shannon information evaluated using full MC sampling of posterior distributions over the two parameters of interest θ_1 and θ_2 (see the Appendix for numerical details). A uniform prior distribution $p(\theta)$ with bounds $2 \le \theta_1 \le 6$, $0 \le \theta_2 \le 1$, and $0 \le \theta_3 \le 2\pi$ was used, along with a normal observational uncertainty ε of variance $\sigma = 1/16$. The starting time of the sample sequence could also have been considered as a design parameter, but it is, in this case, of no significance due to the averaging over all phases $0 \le \theta_3 \le 2\pi$ of the signal (since the phase of the signal is unknown).

For comparison, Bayesian D_S and A_S criteria were also evaluated, approximating expected gain in Shannon information and expected posterior variance, respectively, using local linearization of the forward function. The "S" subscript indicates that only a part of the covariance matrix was used, corresponding to the subset of parameters of interest (i.e., the row and column corresponding to the nuisance parameter θ_3 were removed from the covariance matrix prior to evaluating trace or determinant).

Plots of the different design criteria as a function of sampling frequency are shown in Figure 4. Of the two criterion types, subplot (a) contains information measures that should be maximized, while subplot (b) contains uncertainty measures that should be minimized. When the sampling frequency is below the Nyquist criterion of 2 Hz, posterior distributions are multimodal. This shows up clearly in the posterior variance criterion, where the squared distances between the multiple peaks make the expected posterior variance very high for low sampling frequencies. The same effect can be seen in the expected observational ambiguity criterion $W(\xi)$.

The expected gain in Shannon information, however, remains nearly constant for all sampling frequencies, with a slight decreasing trend as sampling frequencies get high. This constancy is due to the fact that even though low sampling frequencies give multimodal posteriors, the area over which the probability is concentrated remains constant. Shannon information does not account for the shape or spread of a probability distribution, only the area over which the high probability is concentrated.

The Bayesian D_S and A_S approximations to expected Shannon information gain and expected posterior variance based on local linearization of the parameter-observation relationship deviate considerably from the true values, especially for low sampling frequencies. Both Bayesian D_S and A_S criteria indicate that the design is better, the lower the sampling frequency is, when in reality the design is getting worse. For low sampling frequencies, posterior distributions will be multimodal, as shown in Figure 4(c). Bayesian D- or A- criteria, however, measure only the properties of one local peak at a time, which may be orders of magnitude different from that of the full distribution (van den Berg, Curtis, and Trampert 2003). Any criterion based on information matrices (local linearization) should therefore not be used on nonlinear problems with multimodal posteriors.

4.2 Location of Wave Sources in a Medium of Inhomogeneous Velocity

The problem of locating sources of wave energy by measuring relative arrival times of the waves at a set of distant locations is common in earth science, as well as other disciplines. If the wave velocity is constant throughout the medium, the determination of the location of a source is fairly straightforward using simple geometrical considerations. When the wave velocity varies throughout the medium, the wave propagation must be simulated (using numerical methods) by solving the wave equation, which for an elastic isotropic medium, can be written

$$\rho \frac{\partial}{\partial t^2}\mathbf{u} = \mu\nabla^2\mathbf{u} + (\mu + \lambda)\nabla(\nabla \cdot \mathbf{u}) + \mathbf{F}, \qquad (34)$$

where $\mathbf{u}$ and $\mathbf{F}$ are displacement and body force vectors, respectively, ρ is the density, μ the shear modulus, and λ Lamé's first parameter of the medium. From simulated wave arrival times, one can then deduce potential source locations that give wave arrival times matching those observed.

In the case considered next, one seeks to locate microseismic events (small earthquakes or rock fractures) occurring in a subsurface medium with a velocity profile that has already been determined in a two-dimensional (2D) plane between two vertical boreholes. Any seismic event will generate both pressure and shear waves, the former traveling faster than the latter, depicted in Figure 5(a). The difference in arrival time between pressure wave and shear wave is then recorded at the seismic receivers, represented by squares in Figures 5 and 6, which can be located either on the earth surface (top) or in the two boreholes on the sides. A probability distribution for the location of the source can then be calculated using Bayes' rule, such as the one depicted in Figure 5(b). Due to the strong nonlinearity of the physics involved, the posterior distribution might have several local maxima, as in the case shown in Figure 5.

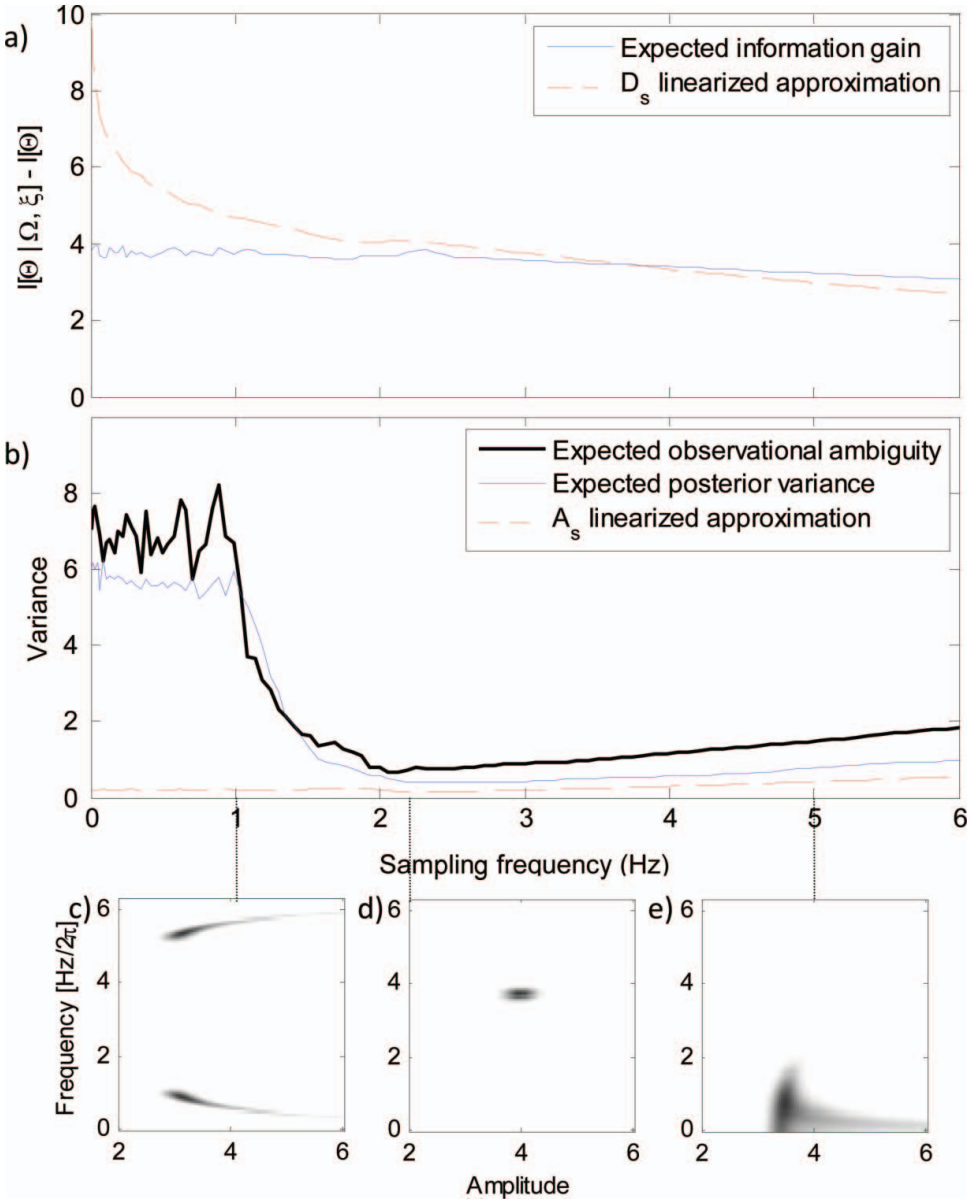

Figure 4. Subfigures (a) and (b) compare different design criteria as a function of sampling frequency: (a) represents two measures of expected gain in Shannon information, and (b) represents three measures related to posterior variance. Solid lines show criteria evaluated using full Monte Carlo sampling of all posterior distributions, dotted lines show Bayesian D_S and A_S criteria. The bold line in subfigure (b) shows the expected observational ambiguity $W(\xi)$, as defined in the text. Subfigures (c), (d), and (e) show example posterior distributions for random observations y at sampling frequencies 1 Hz, 2.2 Hz, and 5 Hz, respectively. The online version of this figure is in color.

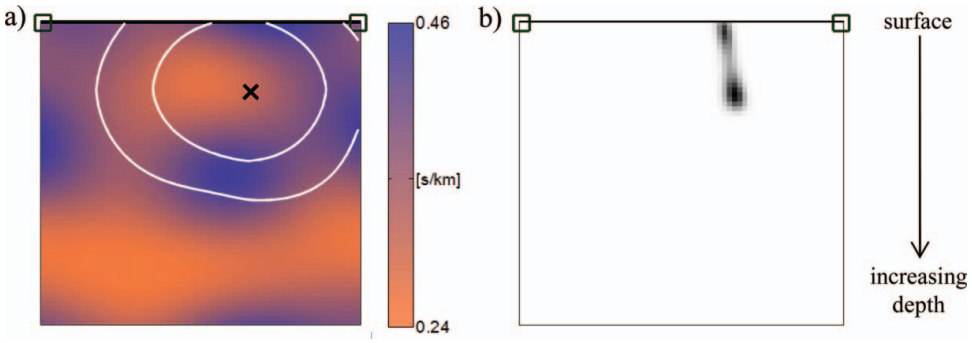

Figure 5. Subfigure (a) shows fast pressure (P) and slower shear (S) wavefronts emanating from a microseismic event (marked x), traveling toward two seismic receivers (squares). The shading of (a) represents the difference in inverse velocity between the two wavefronts throughout the subsurface medium. The shading of (b) represents the probability distribution for the source location given the (noisy) P- and S-wave arrival times recorded at the receivers. The online version of this figure is in color.

Here, the parameter space Θ will be spanned by both the source location θ_{loc} and velocity profile θ_{vel} (known approximately), that is, $\theta^{\text{T}} = [\theta_{\text{loc}}^{\text{T}}, \theta_{\text{vel}}^{\text{T}}]$. However, since we are only interested in determining the source location θ_{loc}, the natural metric $d(\dot{\theta}, \ddot{\theta})$ to use is the distance between two locations $d(\dot{\theta}, \ddot{\theta}) = d(\dot{\theta}_{\text{loc}}, \ddot{\theta}_{\text{loc}})$, ignoring any difference in velocity profiles (which thus constitute nuisance parameters).

The wave velocity profile θ_{vel} was parameterized by a 5×5 regular grid, with the difference in slowness (inverse velocity) between pressure wave and shear wave specified at each grid point. Slowness difference in between the grid points was calculated using interpolation, by taking the discrete 2D Fourier transform of the 5×5 velocities and extending the Fourier spectrum with zero-amplitude higher spatial frequencies before inverting the Fourier transform. The size of the modeled subsurface cross-section was 1 km $\times$ 1 km. The uncertainty in slowness $p(\theta_{\text{vel}})$ was represented by letting the slowness in each grid point vary independently according to a normal distribution with a standard deviation of 0.1 sec/km around an individual mean slowness, the latter represented by shading in Figures 5 and 6. For prior distribution of the wave field source locations $p(\theta_{\text{loc}})$, a uniform distribution over the 1 km $\times$ 1 km cross-section was used, making the total prior distribution over parameter space $p(\theta) = p(\theta_{\text{loc}})p(\theta_{\text{vel}})$.

The observational uncertainty in the detected arrival time difference $p(y|\theta, \xi)$ was assumed to be normally distributed, with a standard deviation of 6 msec for each receiver independently.

The expected observational ambiguity was approximated using (27) and (28) with an MC sample $\{\theta_i\}$ of 10,000 points in parameter space and a sampling distribution $s(\theta) = p(\theta)$. Expected arrival time differences $\eta(\theta, \xi)$, given source location θ_{loc}, velocity profile θ_{vel}, and receiver locations ξ, were approximated using a numerical solver, approximating (34) with an eikonal equation (Vidale 1988). Having normal observational uncertainties $p(y|\theta, \xi)$, $Q(\dot{\eta}_i, \ddot{\eta}_j)$ was calculated using Table 1.

Figure 6 shows three designs considered for the receiver locations. In the left column of the subfigures, lines have been drawn in between all location point pairs $(\dot{\theta}, \ddot{\theta})$ that can be formed out of the sample $\{\theta_i\}$ so that the amount of black ink of each line is proportional to $R(\dot{\theta}, \ddot{\theta}, \xi)$. This makes the total amount of black in each subfigure proportional to $W(\xi)$. When interpreting the plots, high density of lines in a particular area indicates that there will be high uncertainty in estimates of wave source locations in that area. Furthermore, the source location uncertainty will be oriented in the direction of the lines. This way of representing uncertainty in 2D location problems is a refinement of that the one used in Winterfors and Curtis (2008). The subfigures in the right column each contain six superimposed posterior distributions for typical source locations. To be able to separate the distributions, each one is encircled by a contour line.

Subfigures (a) and (b) represent a two-receiver design with both receivers on the surface, which is the minimal number of receivers required to position a wave source in the medium. Subfigures (c) and (d) represent a four-receivers design, with all receivers located in the deeper half of the two boreholes. This represents a commonly used design for surveying hydrocarbon reservoirs. Subfigures (e) and (f) show the design that minimizes the expected observational ambiguity $W(\xi)$ for the approximate velocity profile shown, with two receivers on the surface and one in each borehole.

Comparing the three designs, it is clear that (a) possesses considerable ambiguity ($W(\xi_a) = 14.5$ m^2) compared with (c) and (e), which is not surprising, considering it uses only half the number of receivers of the other two. Most ambiguity in locations is present near the surface, with near-vertical uncertainty directions in a fan-shape, but some uncertainty is also present deeper to the left in lateral directions. Subfigure (c), with an expected observational ambiguity of $W(\xi_c) = 2.45$ m^2, has some vertically oriented uncertainty in the high-velocity area in between the four receivers, as well as some more horizontally oriented uncertainty closer to the surface.

The optimal design in subfigure (e) has $W(\xi_e) = 1.98$ m^2, and uncertainty that is more evenly distributed throughout the cross-section, albeit slightly higher in high-velocity areas than in low-velocity ones. The optimal receiver locations were determined using a simplex optimization algorithm accounting for the constraints imposed by the finite length of the boreholes. Since the expected observational ambiguity may have many local minima, the optimization was repeated 500 times using different starting points (receiver configurations) so that a global minimum could be found with reasonable confidence. Each optimization procedure required about 30 sec of computation time on a 1.4-GHz Pentium IV CPU.

5. DISCUSSION AND CONCLUSIONS

Increasing computational efficiency of fully nonlinear Bayesian design methods has been the primary objective of this work. Expected posterior variance can be estimated by brute force numerical integration methods, but these are too inefficient to tackle even fairly low-dimensional problems. The main issue is computing the expectation over all possible observations, since observation space is usually of considerably higher dimension than the parameter space (lower dimension would make the parameter estimation problem underdetermined). By enabling all integration over observation space to be carried out analytically, $W(\xi)$ offers huge gains in computation efficiency compared with brute force methods.

The expected observational ambiguity $W(\xi)$ measures to what extent different sets of parameters can give rise to the same observations, that is, how much their respective conditional probability distributions in observation space overlap. This is the central property that makes it useful in design of investigation techniques, since it measures a fundamental undesirable property that should be minimized. The expected posterior variance measures a very similar property, but using a slightly different way of measuring the overlap of two probability distributions: $W(\xi)$ uses $R(\dot{\theta}, \ddot{\theta}, \xi) = d^2(\dot{\theta}, \ddot{\theta}) \int p(y|\dot{\theta}, \xi)p(y|\ddot{\theta}, \xi)dy / \int p^2(y|\xi)dy$, whereas the calculating expected posterior variance corresponds to using an overlap measure $R(\dot{\theta}, \ddot{\theta}, \xi) = d^2(\dot{\theta}, \ddot{\theta}) \int p(y|\dot{\theta}, \xi)p(y|\ddot{\theta}, \xi)/(2p(y|\xi))dy$. The former can often be evaluated analytically, whereas the latter cannot.

Theorem 1 shows another, equivalent way of interpreting $W(\xi)$, as a measure of expected posterior variance weighted by the marginal probability density $p(y|\xi)$ over observation space.

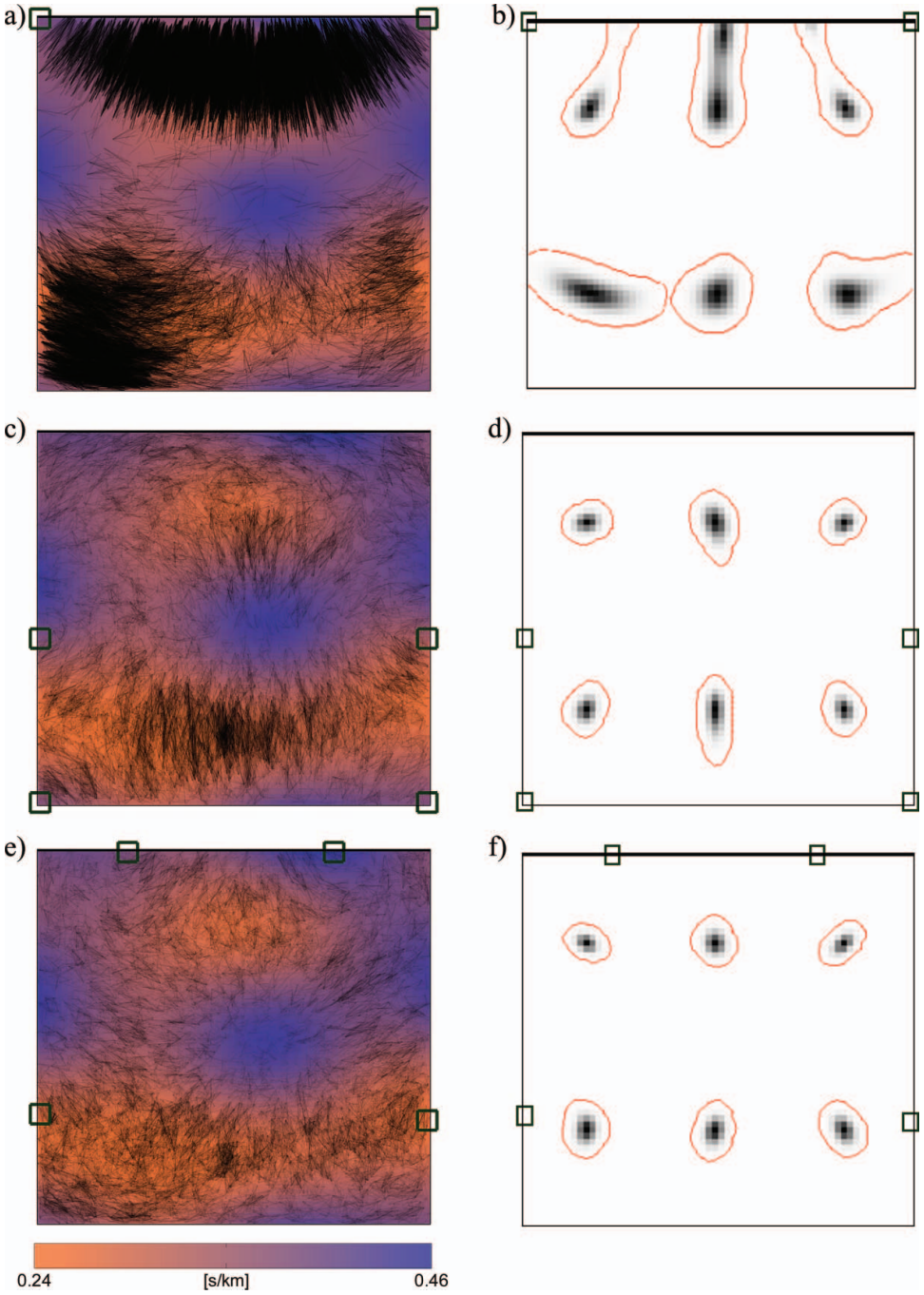

Figure 6. Subfigures (a), (c), and (e) represent areas of large uncertainty in source location, as well as the direction of uncertainty, using black lines connecting pairs of locations that cannot be discriminated by travel time data recorded by the receivers (squares). The upper edge is the earth's surface on each vertical subsurface intersection. Subfigures (b), (d), and (f) each show six typical posterior distributions for the source location, each of which is encircled by a contour line. The online version of this figure is in color.

It is also possible to use $W(\xi)$ as a design criterion when one is to use frequentist inference to interpret observations y. In a similar manner to when Bayesian D- and A-optimality criteria are used, one can integrate over some prior $p(\theta)$ over parameter space that will then not be used for inference. Just as in the fully Bayesian case, $W(\xi)$ will then measure prevalence of multiple, radically different parameter sets fitting the observations—and minimizing $W(\xi)$ will reduce or eliminate such problems.

The introductory example, as well as Example 4.1, demonstrates that Bayesian D- and A-optimality may fail when some of the posterior distributions are multimodal. It also shows that the expected observational ambiguity measure is a useful alternative to more computationally expensive MC measures of expected posterior variance.

Example 4.2 demonstrates the feasibility of using expected observational ambiguity as a design criterion for designing investigation techniques based on complex physical phenomena where evaluation of expected posterior variance is numerically intractable.

For future work on expected observational ambiguity, obvious areas of potential improvement are (a) refined strategies for sampling the parameter space where $R(\dot{\theta}, \ddot{\theta}, \xi)$ is relatively

high, and (b) constructing better algorithms for the optimization of uncertain estimates of $W(\xi)$ with respect to the design ξ.

APPENDIX

Proof of Theorem 1

Easily verified by inserting $d^2(\dot{\theta}, \ddot{\theta}) = [\dot{\theta} - \ddot{\theta}]^2$ (where $\theta^2 = \theta^T\theta$) into (9) and expanding the squared expressions, yielding

$$V[\Theta|y,\xi] = \int_{\theta\in\Theta} \theta^2 p(\theta|y,\xi)d\Theta - \left(\int_{\theta\in\Theta} \theta p(\theta|y,\xi)d\Theta\right)^2 \tag{A.1}$$

which is just an alternative expression for the variance.

Proof of Theorem 2

Taking the expectation of both sides of inequality (17) with respect to the marginal density $p(y)$, one obtains (A.2).

Jensen's inequality states that $f(E[X]) \le E[f(X)]$ for all strictly convex functions f, with equality if and only if $p(x)$ equals the Dirac delta function $\delta(x - \mu_X)$. Application to the right-hand side of (A.2) gives (A.3), since the exponential is strictly convex, proving inequality (18).

$$V[\Theta|\Omega,\xi] \ge \frac{n}{2\pi e}\int_{y\in\Omega} p(y)\exp\left(\frac{2}{n}H[\Theta|y,\xi]\right)dy \tag{A.2}$$

$$\ge \frac{n}{2\pi e}\exp\left(\frac{2}{n}H[\Theta|\Omega,\xi]\right). \tag{A.3}$$

Since inequality (17) has equality if and only if $p(\theta)$ is a spherical normal distribution, so will inequality (A.2). Applying Jensen's theorem guarantees equality if and only if $p(H[\Theta|y,\xi])$ is a Dirac function with respect to y; in other words, all possible posteriors must have identical entropy. These two conditions prove the equality condition of Theorem 2.

Proof of Theorem 3

The proof is along the same lines as that of Theorem 2. Multiplying both sides of inequality (17) with $p(y)$, one obtains

$$p(y)V[\Theta|y] \ge \frac{n}{2\pi e}\exp\left(\ln(p(y)) + \frac{2}{n}H[\Theta|y]\right). \tag{A.4}$$

Taking the expectation of both sides with respect to $p(y|\xi)$ and using Lemma 1 gives (A.5). Since the exponential is strictly convex, we can apply Jensen's inequality to the right-hand side, obtaining (A.6), which is equivalent to (A.7) :

$$\frac{W(\xi)}{L} \ge \int_{y\in\Omega} p(y|\xi)\exp\left(\frac{2}{n}H[\Theta|y,\xi] + \ln(p(y|\xi))\right)d\Omega \tag{A.5}$$

$$\ge \exp\left(\int_{y\in\Omega} p(y|\xi)\left(\frac{2}{n}H[\Theta|y,\xi] + \ln(p(y|\xi))\right)d\Omega\right) \tag{A.6}$$

$$= \exp\left(\frac{2}{n}H[\Theta|\Omega,\xi] - H[\Omega|\xi]\right) \tag{A.7}$$

where $L = n/(\pi e D_\Omega)$.

Inequality (17) has equality if and only if $p(\theta)$ is a spherical normal distribution, and so will inequality (A.4).

Applying Jensen's theorem guarantees equality if and only if $\ln(p(y|\xi)) + 2H[\Theta|y,\xi]/n$ is constant with respect to y when $p(y|\xi) > 0$. Combining these two conditions implies that (A.5)–(A.7) all have equality if and only if all possible posterior densities are spherical normal distributions with identical entropy, and $p(y|\xi)$ is constant for all $y : p(y|\xi) > 0$, which proves the equality condition of Theorem 3.

Proofs of Results in Table 1

Multivariate Normal. Given that $p(y|\theta,\xi)$ is Gaussian $p(y|\eta(\theta,\xi)) = \exp(-[y-\eta(\theta,\xi)]^2/2)/K$, where $K = \sqrt{(2\pi)^n|\Sigma|}$, the square operator $(^2)$ is defined as $y^2 = y^T\Sigma^{-1}y$ and Σ an $n\times n$ covariance matrix, the following simplifications of the expression for $Q(\dot{\eta},\ddot{\eta}) = \int_{y\in\Omega} p(y|\dot{\eta})\,p(y|\ddot{\eta})\,dy$, where $\dot{\eta} = \eta(\dot{\theta},\xi)$ and $\ddot{\eta} = \eta(\ddot{\theta},\xi)$, can be made:

$$Q(\dot{\eta},\ddot{\eta}) = \frac{1}{K^2}\int_{y\in\mathbb{R}^n}\exp\left(-\frac{1}{2}[y-\dot{\eta}]^2 - \frac{1}{2}[y-\ddot{\eta}]^2\right)dy \tag{A.8}$$

$$= \frac{1}{K^2}\int_{y'\in\mathbb{R}^n}\exp\left(-\frac{1}{2}\left(\left[y' + \frac{1}{2}\Delta\eta\right]^2 + \left[y' - \frac{1}{2}\Delta\eta\right]^2\right)\right)dy' \tag{A.9}$$

$$= \frac{1}{K^2}\int_{y'\in\mathbb{R}^n}\exp\left(-y'^2 - \frac{1}{4}\Delta\eta^2\right)dy' \tag{A.10}$$

$$= \frac{\sqrt{2^n}}{K}\exp\left(-\frac{1}{4}\Delta\eta^2\right) \tag{A.11}$$

Equation (A.9) is formed by a change of variable $y = y' + (\dot{\eta}+\ddot{\eta})/2$ in the integrand and by defining that $\Delta\eta = \ddot{\eta} - \dot{\eta}$. Expanding the squares gives (A.10), and using that $\int_y \exp(-y^\dagger\sigma^{-1}y)dM = \sqrt{(2\pi)^n|2\sigma|} = 2^n\sqrt{\pi^n|\sigma|}$ yields (A.11).

Gamma and Exponential. When $p(y|\theta,\xi)$ is a Gamma distribution $p(y|\eta) = k^k y^{k-1}\exp(-ky/\eta)/(\eta^k\Gamma(k))$, where $y,\eta,k \ge 0$, we obtain (A.12). The integrand has a well-known solution (A.13), whose insertion into (A.12) gives (A.14). The exponential distribution is the special case of a Gamma distribution with $k = 1$:

$$Q(\dot{\eta},\ddot{\eta}) = \frac{k^{2k}}{\dot{\eta}^k\ddot{\eta}^k\Gamma^2(k)}\int_{y=0}^{\infty} y^{2k-2}\exp\left(-\frac{\dot{\eta}+\ddot{\eta}}{\dot{\eta}\ddot{\eta}}ky\right)dy. \tag{A.12}$$

$$\int_{y=0}^{\infty} y^{2k-2}\exp\left(-\frac{\dot\eta+\ddot\eta}{\dot\eta\ddot\eta}ky\right)dy=\frac{\Gamma(k)}{\sqrt{\pi}4^k}\Gamma\left(k-\frac{1}{2}\right)\left(\frac{\dot\eta+\ddot\eta}{4\dot\eta\ddot\eta}k\right)^{1-2k} \tag{A.13}$$

$$Q(\dot\eta,\ddot\eta)=\frac{k\Gamma\,(k-1/2)}{\sqrt{\pi}\Gamma(k)}\frac{(4\dot\eta\ddot\eta)^{k-1}}{(\dot\eta+\ddot\eta)^{2k-1}}. \tag{A.14}$$

Poisson. When $p(y|\eta)=\eta^y e^{-\eta}/y!$ over the set of positive integers, $Q(\dot\eta,\ddot\eta)$ is defined through the discrete sum

$$Q(\dot\eta,\ddot\eta)=\sum_{k=0}^{\infty}\frac{\dot\eta^y}{y!}e^{-\dot\eta}\frac{\ddot\eta^y}{y!}e^{-\ddot\eta}=e^{-\dot\eta-\ddot\eta}$$
$$\sum_{k=0}^{\infty}\frac{\dot\eta^k\ddot\eta^k}{y!^2}=e^{-\dot\eta-\ddot\eta}I_0(2\sqrt{\dot\eta\ddot\eta}), \tag{A.15}$$

where the last equality is obtained by inserting the definition of the modified Bessel function of the first kind, I_0.

Numerical Evaluation of Design Criteria in Example 4.1

A random sample of N (in the example equal to 800, the smallest number that gave the same plotted curves over several runs) points $\{\theta_1,\theta_2,\ldots,\theta_N\}$ in parameter space was generated according to the prior distribution $p(\theta)$. From each point θ_i, one point y_i in observation space was generated according to $p(y|\theta_i,\xi)$, forming a set of N points $\{y_1,y_2,\ldots,y_N\}\sim p(y|\xi)$.

It is possible to estimate $p(y|\xi)$ at the points $\{y_1,y_2,\ldots,y_N\}$ by

$$\hat p(y_k|\xi)=\frac{1}{N-1}\left(\sum_{i\neq k}p(y_k|\theta_i,\xi)\right), \tag{A.16}$$

where $i\neq k$ implies summation over $i=1$ up to N, but excluding $i=k$ (it has to be excluded since y_k is not independent of θ_k).

The expected posterior variance was then estimated

$$\mathrm{Var}\,[\Theta|\Omega,\xi]\approx\frac{1}{N}\sum_k\left(\theta_k-\frac{\sum_{m\neq k}\theta_m p(y_k|\theta_m,\xi)}{(N-1)\hat p(y_k|\xi)}\right)^2 \tag{A.17}$$

The squaring of the bracketed terms on the right represents self-multiplication by an inner product $\theta^2=\theta^T\mathbf{M}\theta$, where $\mathbf{M}$ is a diagonal matrix with the first two diagonal entries equal to 1 and the third equal to zero (to remove influence of the nuisance parameter).

The expected gain in Shannon information over the space spanned by the first two components of the parameter vector θ was estimated as

$$I[\Theta|\Omega]-I[\Theta]\approx\frac{1}{N(N-1)}\sum_j\sum_{i\neq j}G_{ij}\ln(G_{ij}) \tag{A.18}$$

where $G_{ij}=\hat p(y_j|[\theta_i^1,\theta_i^2]^T,\xi)/\hat p(y_j|\xi)$ and $\hat p(y_j|[\theta_i^1,\theta_i^2]^T,\xi)=\sum_{k\neq j}p(y_j|[\theta_i^1,\theta_i^2,\theta_k^3]^T,\xi)/(N-1)$.

$$\hat p\left(y_j|\left[\theta_i^1,\theta_i^2\right]^T,\xi\right)=\frac{1}{N-1}\sum_{k\neq j}p\left(y_j|\left[\theta_i^1,\theta_i^2,\theta_k^3\right]^T,\xi\right).$$

The Bayesian D_S and A_S criteria were evaluated on the same random sample $\{\theta_1,\theta_2,\ldots,\theta_N\}$ using (13) to approximate posterior covariance matrices.

The expected observational ambiguity $W(\xi)$ was estimated using (27) on the sample $\{\theta_1,\theta_2,\ldots,\theta_N\}$. Computation time to generate Figure 4 was about 180 sec on a 1.4-GHz Pentium IV processor.

ACKNOWLEDGMENTS

We thank the referees and, in particular, the Associate Editor and the Editor, whose comments and suggestions have been very helpful in clarifying this presentation.

[Received March 2010. Revised February 2012.]

REFERENCES

Bingham, D. R., and Chipman, H. A. (2007), "Incorporating Prior Information in Optimal Design for Model Selection," *Technometrics*, 49(2), 155–163. [181]

Chaloner, K., and Verdinelli, I. (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10(3), 273–304. [179]

Chen, H. (2002), *Stochastic Approximation and Its Applications*, Dordrecht: Kluwer. [184]

Clyde, M. (1993), "Bayesian Optimal Design for Approximate Normality," Ph.D. dissertation, University of Minnesota. [183]

DeGroot, M. H. (1984), "Changes in Utility as Information," *Theory and Decision*, 17(3), 287–303. [179]

Dembo, A., Cover, T., and Thomas, J. (1991), "Information Theoretic Inequalities", *IEEE Transactions on Information Theory*, 17(6), 1501–1518. [182]

Fedorov, V. V. (1972), *Theory of Optimal Experiments*, London: Academic Press. [179]

Gaivoronski, A. (1988), "Stochastic Quasigradient Methods and Their Implementation," in *Numerical Techniques for Stochastic Optimization*, ed. Ermoliev and R. J.-B. Wets, New York: Springer, pp. 313–351. [184]

Kiefer, J., and Wolfowitz, J. (1952), "Stochastic Estimation of the Maximum of a Regression Function," *Annals of Mathematical Statistics*, 23, 462–466. [184]

Merlé, Y., and Mentré, F. (1997), "Stochastic Optimization Algorithms of a Bayesian Design Criterion for Bayesian Parameter Estimation of Nonlinear Regression Models: Application in Pharmacokinetics," *Mathematical Biosciences*, 144, 45–70. [184]

Müller, P., and Parmigiani, G. (1995), "Optimal Design via Curve Fitting of Monte Carlo Experiments," *Journal of the American Statistical Association*, 90(432), 1322–1330. [179,184]

Müller, P., Sanso, B., and De Iorio, M. (2004),"Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation," *Journal of the American Statistical Association*, 99(467), 788–798. [184]

Nyquist, H. (1928), "Certain Topics in Telegraph Transmission Theory," *Transactions of the American Institute of Electrical Engineers*, 47, 617–644. [184]

Pukelsheim, F. (1993), *Optimal Design of Experiments*, New York: Wiley. [179]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22(3), 400–407. [184]

Sebastiani, P., and Wynn, H. (2000), "Maximum Entropy Sampling and Optimal Bayesian Experimental Design," *Journal of the Royal Statistical Society,* Series B, 62(1), 145–157. [182]

Silvey, S. D. (1980), *Optimal Design*, London: Chapman & Hall. [179,181]

Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization*, New York: Wiley & Sons. [184]

van den Berg, J. A., Curtis, A., and Trampert, J. (2003), "Bayesian, Nonlinear Experimental Design Applied to Simple, Geophysical Examples," *Geophysical Journal International*, 55(2), 411–421. [179,185]

Vidale, J. (1988), "Finite-Difference Calculation of Travel Times," *Bulletin of the Seismological Society of America*, 78(6), 2062–2076. [187]

Winterfors, E., and Curtis, A. (2008), "Numerical Detection and Reduction of Non-Uniqueness in Nonlinear Inverse Problems," *Inverse Problems*, 24(2), 025016. [181,187]