

Efficient nonlinear Bayesian survey design using D_N optimization

Darrell Coles¹ and Andrew Curtis²

ABSTRACT

A new method for fully nonlinear, Bayesian survey design renders the optimization of industrial-scale geoscientific surveys as a practical possibility. The method, D_N optimization, designs surveys to maximally discriminate between different possible models. It is based on a generalization to nonlinear design problems of the D criterion (which is for linearized design problems). The main practical advantage of D_N optimization is that it uses efficient algorithms developed originally for linearized design theory, resulting in lower computing and storage costs than for other nonlinear Bayesian design techniques. In a real example in which we optimized a seafloor microseismic sensor network to monitor a fractured petroleum reservoir, we compared D_N optimization with two other networks: one proposed by an industrial contractor and one optimized using a linearized Bayesian design method. Our technique yielded a network with superior expected data quality in terms of reduced uncertainties on hypocenter locations.

INTRODUCTION

Statistical experimental design (SED) is the theory and practice of optimizing experiments to maximize the expected information in data observations. A good experiment is one in which recorded data are expected to discriminate maximally between different possible models. The virtue of SED is that experiments can be optimized before any data are collected, even where existing knowledge about the model parameters is limited. This surprising fact hinges on the use of prior information.

For our purposes, *prior information* describes any quantifiable knowledge relevant to the data-model relationship, the expected

uncertainty in data observations, and the range of probable model parameterizations. We restrict this article to model-oriented design problems in which a theoretical relationship between data and model is known and expressed as

$$\mathbf{d} = \mathbf{g}(\mathbf{m}, \boldsymbol{\xi}) + \boldsymbol{\varepsilon}(\boldsymbol{\xi}), \quad (1)$$

where \mathbf{g} is the theoretical function relating data and model, \mathbf{m} is the model parameter vector, $\boldsymbol{\xi}$ is a vector describing the design of the experiment used to observe data \mathbf{d} , and $\boldsymbol{\varepsilon}$ is the data noise, which may depend on the experimental design $\boldsymbol{\xi}$. The goal of SED is to optimize $\boldsymbol{\xi}$ before the experiment to obtain maximum information about \mathbf{m} after the experiment.

A Bayesian statement of the solution to the inverse problem of constraining model \mathbf{m} given any data set \mathbf{d} is

$$p(\mathbf{m}|\mathbf{d}, \boldsymbol{\xi}) = \frac{p(\mathbf{d}|\mathbf{m}, \boldsymbol{\xi})p(\mathbf{m})}{p(\mathbf{d}|\boldsymbol{\xi})}, \quad (2)$$

where $p(\mathbf{m}|\mathbf{d}, \boldsymbol{\xi})$ is the conditional posterior (postsurvey) model probability density function (PDF) given data \mathbf{d} and design $\boldsymbol{\xi}$, $p(\mathbf{d}|\mathbf{m}, \boldsymbol{\xi})$ is the conditional data PDF given model \mathbf{m} and $\boldsymbol{\xi}$, $p(\mathbf{m})$ is the prior (presurvey) model PDF (assumed to be independent of \mathbf{d} and $\boldsymbol{\xi}$), and $p(\mathbf{d}|\boldsymbol{\xi})$ is the marginal data PDF given $\boldsymbol{\xi}$ (which equals the integral of the numerator on the right-hand side over the entire model domain, i.e., this term normalizes the right side of equation 2 to make the left side a valid PDF). Equation 2 accommodates quantitative prior information on the model via the model prior PDF $p(\mathbf{m})$ and information on data uncertainties via the data-noise prior PDF $p(\mathbf{d}|\mathbf{m}, \boldsymbol{\xi})$. The model and data-noise priors usually can be estimated before final data collection. Because \mathbf{d} implicitly depends on the theoretical data-model function \mathbf{g} in equation 1, the conditional data PDF also incorporates this theoretical prior information.

In geophysics, \mathbf{g} normally expresses a nonlinear relation between \mathbf{d} and \mathbf{m} (e.g., the relationship between heterogeneity in subsurface electrical conductivity \mathbf{m} and electromagnetic measurements \mathbf{d} , where $\boldsymbol{\xi}$ could describe the locations where measurements are

Manuscript received by the Editor 20 August 2010; published online 24 March 2011.

¹University of Edinburgh, School of GeoSciences, Grant Institute, Edinburgh, U. K., and Geoscience Research Centre, Total E&P UK, Aberdeen, U. K. E-mail: darrell.coles@gmail.com.

²University of Edinburgh, School of GeoSciences, Grant Institute, Edinburgh, U. K. E-mail: andrew.curtis@ed.ac.uk.

© 2011 Society of Exploration Geophysicists. All rights reserved.

made on the surface). The nonlinearity of \mathbf{g} must be accounted for in the design of geoscientific experiments because it strongly affects post-inversion parameter uncertainties.

A few geoscientific design papers explicitly address the nonlinearity of \mathbf{g} (van den Berg et al., 2003; Winterfors and Curtis, 2008; Guest and Curtis, 2009, 2010). These papers avoid local linearization (approximation) of \mathbf{g} , which is prevalent in most other geoscientific design studies such as survey design for geoelectrical methods (e.g., Coles and Morgan, 2009), bathymetry inference (Narayanan et al., 2004), seismic borehole tomography (e.g., Curtis 1999a, 1999b; Haber et al., 2008), seismic network optimization (e.g., Steinberg et al., 1995), and oceanographic experimentation (Barth and Wunsch, 1990). Avoiding linearization is important because such a low-order approximation to \mathbf{g} can introduce systematic errors to the design optimization, compromising the quality of data from the resulting experiments.

The chief obstacle to nonlinear design is computing expense. This is because the optimum design ξ^* is that which maximizes the expected information, which is typically calculated by an integral operation on the posterior model PDF given in equation 2. We do not know which is the correct model \mathbf{m} a priori, so it is necessary to maximize the expectation or average a posteriori model information over all probable models according to the prior model PDF $p(\mathbf{m})$ and over all probable data sets $p(\mathbf{d}|\mathbf{m}, \xi)$. Estimating this expectation or average requires integration over all \mathbf{m} and \mathbf{d} , a very significant computation. Hence, any work that attempts to make nonlinear design practical must address this computing expense.

Winterfors and Curtis (2008) address computing expense by introducing an efficiently calculable measure of information in the model space. The efficiency derives from the fact that data uncertainty distributions describing variations of ε in equation 1 are usually assumed to follow well-understood analytic forms (e.g., Gaussian, Poisson). In such cases, part of the integration can be performed analytically rather than numerically. Guest and Curtis (2009, 2010) instead introduce a Monte Carlo integral method that performs more efficiently than previous methods. Its efficiency derives from successively removing from consideration regions from the space of possible designs that are statistically unlikely to contain ξ^* .

We introduce a method of nonlinear Bayesian design that is computationally efficient and, as with the methods above,

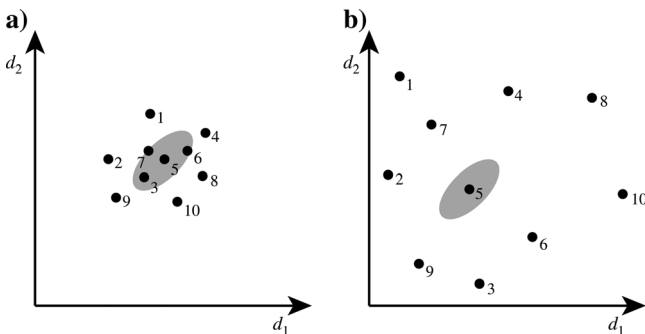


Figure 1. Ten data sets (points) corresponding to the same 10 distinct models as observed by two different experiments, (a) ξ_1 and (b) ξ_2 , of two observations in 2D data space (d_1 and d_2) corresponding to a single model. The ellipse represents a priori expected data uncertainty, in this case around data set 5.

requires no linearization of \mathbf{g} . We call the design objective function in this method the D_N criterion; the process of optimizing an experiment according to this objective is called D_N optimization. The D stands for the determinant criterion from linear design theory (cf. Pukelsheim, 2006), which we extend to nonlinear design problems, whence comes the subscript N . Shewry and Wynn (1987) show that a priori information in data space is functionally related to a posteriori information in model space. This makes it possible to optimize a design ξ^* in data space (through an appropriate choice of objective function) that is optimum over the expected posterior PDF of models in model space. The ability to optimize designs in data space is computationally expedient because it precludes the need to perform costly inversions as part of the optimization workflow (Shewry and Wynn, 1987; Guest and Curtis, 2009). The efficiency of our new method is derived from this fact.

We demonstrate the method on two design problems. The first examines a generic tomography problem in which the objective is to optimize a set of source-receiver pairs to query an unknown velocity model. The purpose of this example is to familiarize the reader with our theory. The second example designs a real, industrial-scale, microseismic monitoring survey; it shows that our method significantly outperforms designs constructed using current standard methods.

THE D_N CRITERION

To begin the development, we introduce a simple hypothetical scenario. Consider a set of 10 expected notional data sets, recorded by two distinct experiments with designs ξ_1 and ξ_2 of two observation points each. Each data set corresponds to a distinct model $\mathbf{m}_i \sim p(\mathbf{m})$ through a notional theoretical function \mathbf{g} . The notation $x \sim p(x)$ means that x has been sampled from, or is distributed according to, the PDF $p(x)$. The data sets can be plotted as points in a 2D data space as in Figure 1. Additionally, the data are expected to be noisy, so each data set has an uncertainty region, as depicted by the ellipse around data set 5, which represents the PDF of ε . In this example, experiment 1 causes the data sets to be close to one another in data space but experiment 2 does not. The problem with ξ_1 is that because the recorded data are expected to contain errors $\varepsilon(\xi_1)$, they may be consistent with several different models from the sample, (e.g., several data sets fall within the uncertainty region of data set 5). Therefore, data set 5 cannot be used to discriminate between the distinct models corresponding to each of these data sets.

By contrast, in experiment 2, the data sets are expected to be far enough from each other that the uncertainty region associated with ε around any particular data set contains the expected data corresponding to no other models from the sample. In this sense, experiment 2 offers better model discrimination than experiment 1 because it reduces model ambiguity; if any one of the 10 data sets were actually observed, the model corresponding to that data set would be readily identified (even accounting for data uncertainty from noise), which cannot be claimed of experiment 1. Thus we prefer design ξ_2 to design ξ_1 .

This heuristic example highlights that model ambiguity can be minimized by maximizing some measure of the expected distance between data sets, accounting for expected data uncertainties. In fact, this is the main thesis behind work by Shewry and Wynn (1987) on maximum entropy sampling: they prove that

when the data error is independent of the design, maximizing the entropy of expected data (effectively a measure of data scatter and hence, informally, a distance measure in data space) is equivalent to minimizing the entropy of a posteriori expected models (this is a well-known Bayesian SED objective that measures the a posteriori model uncertainty). Using that result, it is evident in Figure 1 that experiment 2 is superior to experiment 1, as the entropy (scatter) of the mediated data is plainly greater. Thus, design optimization can be carried out in data space rather than in model space, where it is traditional to address model uncertainty.

In reality, the set of permissible models is usually continuous and not discrete as in Figure 1, so some model ambiguity is unavoidable. Moreover, it is not just the distance between data sets (points in data space) that should be maximized but rather the degree to which the uncertainty distributions around potential recorded data sets overlap. By including these uncertainty distributions, the expected characteristics of the data noise are naturally integrated into the design problem.

Assuming $\mathbf{g}(\boldsymbol{\xi})$ is a multivariate Gaussian (hereafter *multinormal*), the conditional PDF of a data set corresponding to model \mathbf{m}_k is of the form $p_k(\mathbf{d}|\mathbf{m}_k, \boldsymbol{\xi}) \equiv N[\mathbf{g}(\mathbf{m}_k, \boldsymbol{\xi}), \boldsymbol{\Sigma}_d(\boldsymbol{\xi})]$, where $N(\boldsymbol{\mu}, \mathbf{C})$ denotes a multinormal distribution with mean $\boldsymbol{\mu}$ and covariance \mathbf{C} , and where $\boldsymbol{\Sigma}_d(\boldsymbol{\xi})$ is the expected data-noise covariance matrix of $\mathbf{g}(\boldsymbol{\xi})$, which in this work may be design dependent. The ellipse centered on data set 5 in Figure 1 represents an example p_k ; its size, orientation, and position symbolize the multinormal distribution $N[\mathbf{g}(\mathbf{m}_k, \boldsymbol{\xi}), \boldsymbol{\Sigma}_d(\boldsymbol{\xi})]$ from which the data are expected to derive.

Now consider a pair of data sets \mathbf{d}_i and \mathbf{d}_j with conditional PDFs p_i and p_j , respectively. Based on the foregoing development, the objective is to minimize the overlap between them. A natural way to do this is to maximize their relative entropy or Kullback-Leibler divergence D (cf. Cover and Thomas, 1991):

$$D(p_i(\mathbf{d})||p_j(\mathbf{d})) \equiv \int_{\Gamma} p_i(\mathbf{d}) \ln \frac{p_i(\mathbf{d})}{p_j(\mathbf{d})} d\mathbf{d}, \quad (3)$$

where Γ is the data domain over which p_i and p_j are defined (the notation in equation 3 is conventional and native to information theory). Informally, relative entropy is a nonnegative measure of the distance between two PDFs that only equals zero when the PDFs are identical. The relative entropy between multinormal distributions is analytic (Goldberger et al., 2003) and in the case of p_i and p_j is

$$D(p_i||p_j) = \frac{1}{2} [\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})]^T [\boldsymbol{\Sigma}_d(\boldsymbol{\xi})]^{-1} \times [\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})]. \quad (4)$$

Notably, equation 4 is an instance of the squared Mahalanobis distance (Mahalanobis, 1936), a general statistical distance measure equal to the squared Euclidean distance between probable data sets, normalized by the expected data uncertainties $\boldsymbol{\Sigma}_d(\boldsymbol{\xi})$.

Because equation 4 is valid for any pair of models sampled from $p(\mathbf{m})$, $\mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})$ and $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi})$ can be treated as random variables; so by virtue of the central limit theorem, $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})$ is more Gaussian than the distributions of $\mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})$ or $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi})$ for random variations in \mathbf{m}_i and \mathbf{m}_j (Hyvärinen et al., 2001). We henceforth assume that $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})$ is approximately multi-

normal over the domain of probable models and hence has (it is readily demonstrated) mean

$$E_{\mathbf{m}_i, \mathbf{m}_j} [\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})] = \mathbf{0} \quad (5)$$

and (it is also readily demonstrated) covariance

$$\begin{aligned} \boldsymbol{\Sigma}_g(\boldsymbol{\xi}) &= E_{\mathbf{m}_i, \mathbf{m}_j} \{ [\mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi})] [\mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi})]^T \} \\ &= 2E_{\mathbf{m}} \{ [\mathbf{g}(\mathbf{m}, \boldsymbol{\xi}) - \bar{\mathbf{g}}(\mathbf{m}, \boldsymbol{\xi})] [\mathbf{g}(\mathbf{m}, \boldsymbol{\xi}) - \bar{\mathbf{g}}(\mathbf{m}, \boldsymbol{\xi})]^T \}, \end{aligned} \quad (6)$$

where E_x is the expectation operator with respect to $p(x)$ and $\bar{\mathbf{g}}(\mathbf{m}, \boldsymbol{\xi}) \equiv E_{\mathbf{m}} \mathbf{g}(\mathbf{m}, \boldsymbol{\xi})$. Hence, $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi}) \sim N[\mathbf{0}, \boldsymbol{\Sigma}_g(\boldsymbol{\xi})]$. We call $\boldsymbol{\Sigma}_g(\boldsymbol{\xi})$ the *theoretical data covariance matrix* because it corresponds to the deterministic component of the data, governed by the theoretical function \mathbf{g} , and to distinguish it from the purely stochastic data component, which corresponds to the data-noise covariance $\boldsymbol{\Sigma}_d(\boldsymbol{\xi})$. The factor of two in equation 6 is henceforth suppressed without any loss of generality.

It is a property of multinormal variables (Wunsch, 1996) that for each $\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi}) \sim N[\mathbf{0}, \boldsymbol{\Sigma}_g(\boldsymbol{\xi})]$ there exists a $\boldsymbol{\delta}(\boldsymbol{\xi}) \sim N[\mathbf{0}, \mathbf{I}]$ such that

$$\boldsymbol{\delta}(\boldsymbol{\xi}) = [\boldsymbol{\Sigma}_g(\boldsymbol{\xi})]^{-1/2} [\mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi})], \quad (7)$$

where

$$[\boldsymbol{\Sigma}_g(\boldsymbol{\xi})]^{1/2} \boldsymbol{\delta}(\boldsymbol{\xi}) = \mathbf{g}(\mathbf{m}_j, \boldsymbol{\xi}) - \mathbf{g}(\mathbf{m}_i, \boldsymbol{\xi}), \quad (8)$$

Thus, expression 4 reduces to

$$D(p_i||p_j) \cong \frac{1}{2} \boldsymbol{\delta}(\boldsymbol{\xi})^T \boldsymbol{\Sigma}(\boldsymbol{\xi}) \boldsymbol{\delta}(\boldsymbol{\xi}), \quad (9)$$

where

$$\boldsymbol{\Sigma}(\boldsymbol{\xi}) \equiv [\boldsymbol{\Sigma}_g(\boldsymbol{\xi})]^{1/2} [\boldsymbol{\Sigma}_d(\boldsymbol{\xi})]^{-1} [\boldsymbol{\Sigma}_g(\boldsymbol{\xi})]^{1/2}. \quad (10)$$

We call $\boldsymbol{\Sigma}(\boldsymbol{\xi})$ the *nonlinear data covariance matrix*.

Equation 9 is (an approximation of) the relative entropy between any two conditional data PDFs, and it would seem natural to maximize its expectation over all $\boldsymbol{\delta}$ because relative entropy is a distance measure between data sets. This expectation is easily shown to be

$$E_{\boldsymbol{\delta}} [\boldsymbol{\delta}^T \boldsymbol{\Sigma} \boldsymbol{\delta}] = \int \boldsymbol{\delta}^T \boldsymbol{\Sigma} \boldsymbol{\delta} p(\boldsymbol{\delta}) d\boldsymbol{\delta} = \text{tr} \boldsymbol{\Sigma}. \quad (11)$$

However, expression 11 has a problem: the integral can be maximized even if $D(p_i||p_j) = 0$ (i.e., the integrand is zero) for some distinct pairs of models (recalling that $\boldsymbol{\delta}$ corresponds to model pairs through equation 7). This is undesirable; by maximizing $\text{tr} \boldsymbol{\Sigma}$, we would like all data sets corresponding to distinct models to be themselves as distinct as possible. Thus, an objective function based on the expected relative entropy between data sets, as in equation 11, is inadequate for SED because it does not prevent an optimum experiment from yielding a nonunique theoretical relationship between data and model (at least for some distinct model pairs).

The integrand in expression 11 can be zero if $\boldsymbol{\delta}(\boldsymbol{\xi}) = \mathbf{0}$ as a consequence of \mathbf{m}_i equaling \mathbf{m}_j (see equation 7) or if $\boldsymbol{\Sigma}(\boldsymbol{\xi})$ has a nontrivial null space. The former case is trivial because the relative entropy between the uncertainty PDFs of two data sets corresponding to the same model must be zero. However, the latter

case signifies undesirable nonuniqueness in the theoretical data-model function. It is possible to prevent the singularity of $\Sigma(\xi)$ by modifying equation 11 to penalize experiments ξ that allow it. One way to do this is to define the design objective function:

$$\begin{aligned}\Phi_{DN}(\xi) &\equiv E_{\delta}\{\delta(\xi)^T [\ln \Sigma(\xi)] \delta(\xi)\} \\ &= \text{tr} \ln \Sigma(\xi) = \ln \det \Sigma(\xi),\end{aligned}\quad (12)$$

where $\ln \Sigma$ is the matrix logarithm of Σ defined as $\mathbf{Q}(\ln \Lambda)\mathbf{Q}^T$. Here, $\mathbf{Q}\Lambda\mathbf{Q}^T$ is the spectral decomposition of Σ and $\ln \Lambda \equiv \text{diag}\{\ln \lambda_i\}$, with λ_i being the eigenvalues of Σ . Any experiment ξ for which $\Sigma(\xi)$ is singular will cause Φ_{DN} to be negative infinity because $\Sigma(\xi)$ then has one or more zero eigenvalues, causing $\ln \det \Sigma(\xi) = -\infty$ and irreversibly eliminating that experiment from candidacy as a maximizer.

Equation 12 is the D criterion (cf. Pukelsheim, 2006) common to linearized design theory, except that it has now been generalized for nonlinear Bayesian design. That is, Σ is defined with respect to the nonlinear function \mathbf{g} — no linearization is used to define the D_N criterion in equation 12 — and incorporates prior model information $p(\mathbf{m})$ and prior data uncertainty information $N[\mathbf{0}, \Sigma_d(\xi)]$ through its dependence on $\Sigma_{\mathbf{g}}$ and Σ_d , respectively. We call an experiment maximizing this criterion Bayesian D_N optimal or just D_N optimal for short. We could also define A_N , E_N , and T_N criteria, equivalent to linear A , E , and T criteria (cf. Atkinson et al., 2007), by using the appropriate operators from the design literature on Σ .

LINEARIZED SEQUENTIAL DESIGN

The utility of the D_N criterion lies in the fact that efficient sequential design algorithms from linearized SED (e.g., Curtis et al., 2004; Stummer et al., 2004; Coles and Morgan, 2009; Coles and Curtis, 2011; Khodja et al., 2010) exist for D optimization (the A_N , E_N , and T_N criteria remarked upon above could also be maximized/minimized using these algorithms). When combined with a linearized sequential design algorithm (LSDA), the D_N criterion renders nonlinear Bayesian design computationally feasible for large-scale industrial applications, a feature shared by no other geoscientific nonlinear design technique without recourse to cluster computing or reparameterization of the design space (e.g., Ajo-Franklin, 2009; Guest and Curtis, 2009, 2010). We expand upon this point in the Discussion. Additionally, Coles and Curtis (2011) show that LSDAs can be many orders of magnitude faster than global search techniques.

Details on implementing LSDAs can be found in Coles and Curtis (2011) or Coles and Morgan (2009). Briefly, LSDAs take three basic inputs: a matrix \mathbf{A} , a design criterion (here, the D_N criterion), and a scalar n that indicates the number of observation points desired. Technically, \mathbf{A} can contain any information relevant to experimental design, subject to the requirement that its rows correspond to the set of candidate observation points or types (henceforth just points) for the design problem. For example, if there are 100 total observation points to choose from in a design problem (of which some subset will be selected for the optimum experiment), then \mathbf{A} has 100 rows, each corresponding to one observation point in discrete experiment space. LSDAs operate by iteratively adding and/or deleting observations to/from an experiment, which amounts to rows of \mathbf{A} being switched on or off. The goal of each addition/deletion is to extremize the

quality of the experiment as measured by the specified design criterion (we refer to this as the *objective value*).

LSDAs are greedy algorithms that make local, rather than global, updates to an experiment undergoing optimization. Consequently, they trade global optimality for computational efficiency (Coles and Morgan, 2009; Coles and Curtis, 2011). Three LSDAs are commonly used: construction, decimation, and exchange. Construction adds observation points (one at a time or in groups) to the experiment, conditional on its current state, until it comprises n such points. Decimation deletes observation points from the experiment, again conditional on its current state, until n remain. The exchange algorithm cycles through the n observation points in the experiment, performing a test replacement with all candidate observation points; the test replacement that extremizes the objective value of the experiment is exchanged for the current observation point. As mentioned, sequential design algorithms do not guarantee global optimality, but exchange can approach this in practice (Coles and Curtis, 2011).

DESIGN WORKFLOW

Experiments optimized according to the D_N criterion can be found by linearized sequential design algorithms by executing the following workflow:

- 1) Generate an ensemble of probable models $\{\mathbf{m}_i \mid \mathbf{m}_i \sim p(\mathbf{m})\}$
- 2) Denote the set of all q candidate observation points Ξ ; project each model through the theoretical function and over Ξ to create an ensemble of probable theoretical data sets $\{\mathbf{d}_i \mid \mathbf{d}_i = \mathbf{g}(\mathbf{m}_i, \Xi)\} \sim p(\mathbf{d}|\Xi)$.
- 3) Numerically estimate the theoretical data covariance matrix $\Sigma_{\mathbf{g}}(\Xi) \in \mathbf{R}^{q \times q}$ of the data ensemble, as in equation 6.
- 4) Evaluate $\Sigma(\Xi)$ according to equation 10 (expedients to this and step 3 are discussed later).
- 5) Use an LSDA to find the experiment $\xi^* \subseteq \Xi$ using $s \leq q$ points that maximizes $\Phi_{DN}(\xi)$ in equation 12.

The covariance matrices are with respect to the set of all candidate observation points Ξ . This is by convention because very efficient LSDAs exist (Coles and Morgan, 2009; Khodja et al., 2010; Coles and Curtis, 2011) that require $\Sigma(\Xi)$ as an input (they actually require $[\Sigma(\Xi)]^{1/2}$).

EXAMPLE 1: GENERIC TOMOGRAPHY DESIGN

Our first example designs a D_N optimal experiment for a generic tomography problem. The purpose is to solidify understanding of the basic theoretical machinery before introducing a more sophisticated example. Figure 2 illustrates the problem, the objective of which is to design a tomography experiment consisting of four source-receiver (circles and squares, respectively) pairs along which traveltime measurements will be observed. The medium is divided into four square cells, each with unit-length edges and each assumed to span a region of constant velocity. The objective of the experiment is to estimate these velocities.

The traveltimes t_i are modeled as $t_1 = 1/V_1 + 1/V_2$; $t_2 = 1/V_3 + 1/V_4$; $t_3 = \sqrt{2}/V_1$; $t_4 = \sqrt{2}/V_2$; $t_5 = \sqrt{2}/V_3$; and $t_6 = \sqrt{2}/V_4$. These equations express the theoretical data-model relationship, given by \mathbf{g} in our notation. The set Ξ contains six candidate observation points (the source-receiver pairs that give

rise to the preceding six theoretical functions), and from this Ξ an experiment $\xi \subset \Xi$ will be designed. Four of the six will be chosen to comprise a D_N optimal experiment, so there are 15 distinct experiments to consider. Because the design problem is small, an exhaustive search can be conducted easily, so no LSDA is used in this case.

Individual elements of the data noise $\varepsilon(\Xi)$ are assumed to be independent and identically normally distributed, with zero mean and unit variance, so $\Sigma_d(\Xi) = \mathbf{I}$. The model parameters V_i form parameter vector $\mathbf{m} = [V_1, V_2, V_3, V_4]^T$ and are assumed to be independent and identically uniformly distributed over the interval from 2 to 5 km/s, which constitutes the model prior distribution $p(\mathbf{m})$.

In this example, the theoretical data covariance over all candidate source-receiver pairs (observation points) can be calculated analytically and is, rounded at the fourth decimal place,

$$\Sigma_g(\Xi) = \begin{bmatrix} 0.0134 & 0 & 0.0095 & 0.0095 & 0 & 0 \\ 0 & 0.0134 & 0 & 0 & 0.0095 & 0.0095 \\ 0.0095 & 0 & 0.0134 & 0 & 0 & 0 \\ 0.0095 & 0 & 0 & 0.0134 & 0 & 0 \\ 0 & 0.0095 & 0 & 0 & 0.0134 & 0 \\ 0 & 0.0095 & 0 & 0 & 0 & 0.0134 \end{bmatrix}, \quad (13)$$

where, to be clear, the ij th element of Σ_g is $\text{cov}(t_i, t_j)$ over the domain of $p(\mathbf{m})$. Because the data noise covariance in this case is the identity, the nonlinear covariance matrix $\Sigma(\Xi)$ is just $\Sigma_g(\Xi)$. Note that any time the data noise is assumed a priori to be independent and identically distributed, $\Sigma_d(\Xi)$ can be set to the identity because this only differs from the true covariance by a multiplicative factor, which leaves the critical points of Φ_{D_N} unchanged.

An evaluation of the D_N value of all 15 candidate experiments reveals that the D_N optimal experiment selects the four diagonal

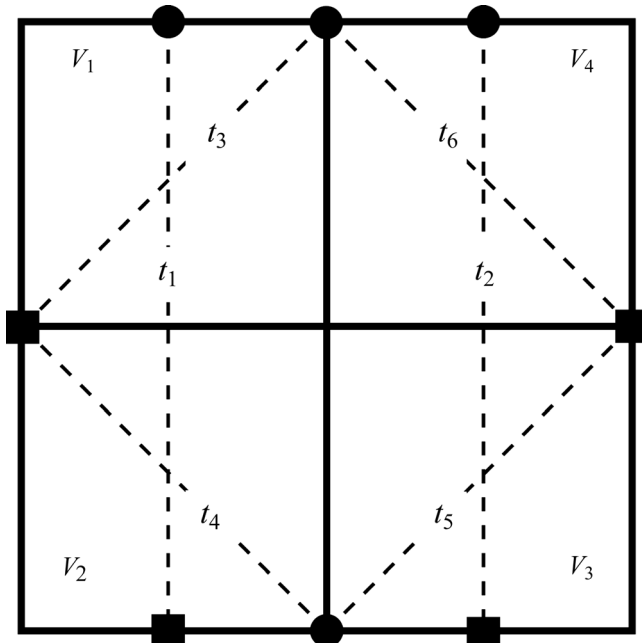


Figure 2. Generic tomography design problem in which four cells with unknown velocities v_1 – v_4 are to be optimally queried by four of the six source (circle)–receiver (square) pairs shown, which mediate traveltimes t_1 – t_6 along the raypaths shown (dashed).

source-receiver pairs corresponding to traveltimes $t_3, t_4, t_5,$ and t_6 (Table 1). This result is intuitive; each source-receiver pair in the D_N optimal survey constrains the velocity in exactly one cell, uncomplicated by sensitivity to adjacent cells. Six experiments have D_N values of negative infinity. It is evident upon inspection that each of these fails to constrain the velocity uniquely in at least one cell, demonstrating that the log-modified relative entropy measure (equation 12) indeed precludes experiments that yield nonunique data-model relationships.

In practice, $\Sigma_g(\Xi)$ usually will not be analytic and must be estimated by sampling the data marginal PDF, mediated by a sampling from $p(\mathbf{m})$, which raises the question, “How many samples are sufficient?” This is explored briefly in Table 2, which reports the rms error in the approximation of the elements of $\Sigma_g(\Xi)$ compared to the analytic matrix for different sample sizes. Also shown is the D_N optimum experiment, found by using each approximation of $\Sigma_g(\Xi)$. The approximation error reduces roughly as the inverse square root of the number of samples; importantly, the D_N optimum experiment in all cases is identical to the one arrived at by using the analytic theoretical data covariance matrix.

EXAMPLE 2: MICROSEISMIC MONITORING NETWORK DESIGN

A more realistic demonstration of the methodology is to optimize a seafloor microseismic receiver network for monitoring an offshore petroleum field. The model \mathbf{m} represents the hypocentral coordinates of a microseismic event, and data \mathbf{d} are the expected arrival times at a set of candidate receiver stations. Microseismic events are assumed to originate primarily around major faults, so $p(\mathbf{m})$ assigns uniform probability to events

Table 1. D_N value (after equation 12, using \log_{10} instead of \ln) of the 15 experiments of four source-receiver pairs possible in Figure 2. Left four columns indicate the subscripts of the traveltimes observed by each experiment. The D_N optimal experiment is highlighted.

Experiment				D_N value
1	2	3	4	$-\infty$
1	2	3	5	-8.1405
1	2	3	6	-8.1405
1	2	4	5	-8.1405
1	2	4	6	-8.1405
1	2	5	6	$-\infty$
1	3	4	5	$-\infty$
1	3	4	6	$-\infty$
1	3	5	6	-7.8395
1	4	5	6	-7.8395
2	3	4	5	-7.8395
2	3	4	6	-7.8395
2	3	5	6	$-\infty$
2	4	5	6	$-\infty$
3	4	5	6	-7.5384

along faults in the reservoir interval (4–6 km below the seafloor) and zero probability elsewhere.

Six hundred example microseismic sources were randomly sampled from $p(\mathbf{m})$ (Figure 3), forming the model ensemble in

Table 2. Misfit between analytic and estimated $\Sigma_g(\Xi)$ for various model space sample sizes, calculated as the rms error between the elements of the approximated and analytic matrices. Also shown is the D_N optimal experimental design in each case in brackets (bottom row), following the subscript convention described in Table 1.

# samples	16	64	256	1024	4096	16,384	65,536
Misfit	0.0285	0.0092	0.0066	0.0025	0.0014	0.0010	0.0005
Design	{3,4,5,6}	{3,4,5,6}	{3,4,5,6}	{3,4,5,6}	{3,4,5,6}	{3,4,5,6}	{3,4,5,6}

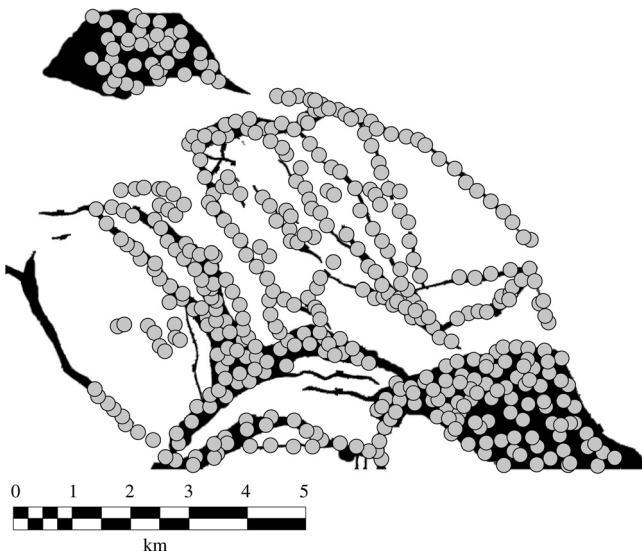


Figure 3. Map of a fractured and faulted (black) petroleum field with 600 possible microseismic sources (dots) randomly selected along faults in accordance with the model prior PDF.

step 1 of the design workflow, and their arrival times at each candidate receiver were calculated to form the data ensemble defined in step 2. The set of candidate receiver positions was a 41×41 grid with 1-km spacing on the seafloor (4 km above the top of the fracture network), centered on the reservoir. The overburden is treated as homogeneous. A modified version of Hutton and Boore's (1987) attenuation was used to model expected signal-to-noise ratios (S/N), which are accommodated in the data-noise covariance $\Sigma_d(\Xi)$. The modification causes S/N to drop off by three orders of magnitude at a distance of 20 km relative to a station positioned at the epicenter of a microseismic event. This builds in a trade-off between placing receivers far away to maximize the angular aperture of the array and nearby to maximize S/N and hence microseismic detectability.

We compare three receiver networks: a network of 13 receivers proposed by a well-known industrial contractor, designed using heuristics (rules of thumb); a comparably sized network found using the linear dependence reduction schema described by Curtis et al. (2004), called LDR optimal, which uses a linearized Bayesian sequential design method; and a comparably sized D_N optimal network optimized by the exchange algorithm. To compare the receiver networks, average post-inversion model variances were estimated for hypocenters at 5-km depths over a region (in map view) slightly larger than the footprint of the fracture network. Average uncertainty was calculated by taking the mean of the diagonal of the linearized model covariance matrix for each receiver network, $1/3\text{trace}(\mathbf{G}^T\mathbf{G})^{-1}$, at each point on a dense grid of potential hypocenters over the region described, where \mathbf{G} is the Jacobian of the traveltime function with respect to the hypocentral coordinates, evaluated for each receiver network and each potential hypocenter.

Results are shown in Figure 4; the fracture network is included for reference. Of the three networks, the D_N optimal network mediates the lowest overall expected hypocentral uncertainties over the fracture network (the region where

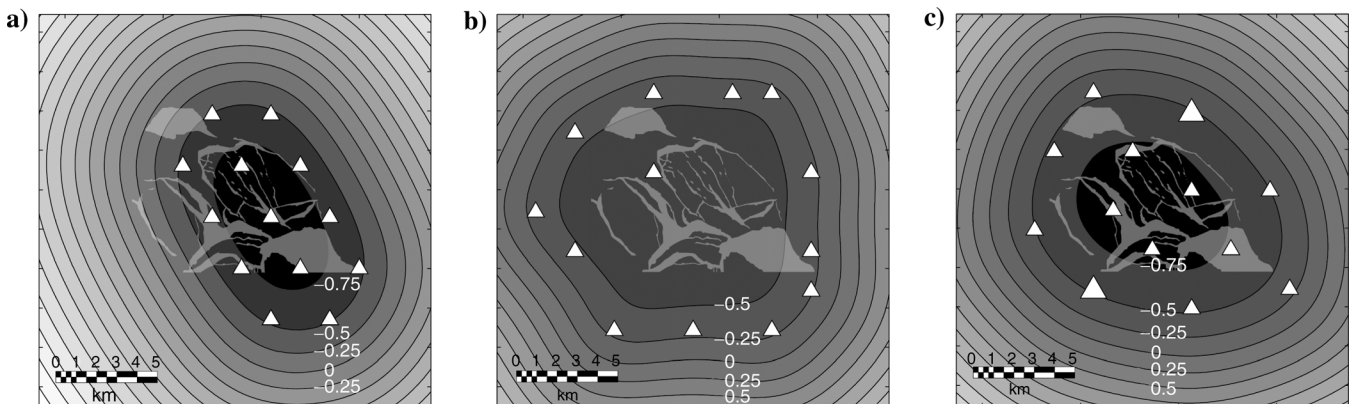


Figure 4. Contour maps of the expected post-inversion model variance (uncertainty) for three survey designs — (a) heuristic, (b) LDR optimal (Curtis et al., 2004), (c) D_N optimal — of 13 receivers (triangles). The maps are reported in \log_{10} variance ($\log_{10} \text{km}^2$), and isolines are labeled with white numbers. For example, the -0.5 isoline indicates the locations at which a hypocentral estimate would have expected model variance of $10^{-0.5} \text{km}^2$. The two larger triangles in (c) are referred to in Figure 5.

microseismic events are assumed to be most probable). The entire fracture network falls within the $10^{-0.5}\text{-km}^2$ isoline, meaning that the average expected uncertainty in the hypocentral estimate of any event originating on a fault is never greater than ± 562 m (i.e., the square root of $10^{-0.5}\text{ km}^2$). In contrast, the largest expected hypocentral estimation error for fault-originating events is approximately 750 m and >1 km for the LDR optimum and heuristic networks, respectively.

Table 3 reports the percent difference in the expected post-inversion standard deviation of each hypocentral coordinate estimate for the heuristic and LDR optimal networks relative to the D_N optimal network. These were estimated by calculating the percent difference in the mean root of the diagonals of the linearized inversion model covariance matrix $(\mathbf{G}^T\mathbf{G})^{-1}$ for each receiver network (over the set of 600 example microseismic events) relative to the mean root of that for the D_N optimum network. The D_N optimum design plainly produced the smallest expected uncertainty in each hypocentral coordinate, particularly in the depth coordinate, where the heuristic and LDR optimal networks yielded uncertainties at least 150% greater than the D_N optimum network on average.

In summary, the D_N optimal network clearly produced lower expected hypocentral uncertainties over the targeted fracture network as well as a more homogeneous distribution of uncertainty in the region of expected microseismicity.

DISCUSSION

Example 1 offers a heuristic case to familiarize the reader with our design theory, and it suggests an important possibility: that D_N optimization might be fairly robust to imperfect sampling of the marginal data PDF, as demonstrated by the D_N optimum designs in Table 2. It is possible that relatively small sample sizes are sufficient to optimize designs by the D_N criterion, which translates to added computational efficiency, especially for high-dimensional design problems. It would be useful to explore this possibility in future work.

In example 2, the D_N optimum network is plainly superior to the other two networks in terms of overall and coordinate hypocentral uncertainties. Example 2 demonstrates that the D_N criterion and attendant LSDA are suitable to optimize experiments efficiently when the data-model relationship is nonlinear and when the expected model parameterization and the expected data noise can be characterized probabilistically a priori.

The value of our methodology is that it offers a means to optimize large-scale nonlinear geoscientific designs in a fraction of the normal time. Consider that the largest (nonreparameterized) nonlinear geoscientific design published to date includes 10 observation points (Guest and Curtis, 2009). Optimizing a nonlinear design the size of example 2 by Guest and Curtis's method is at the computational limit of existing nonlinear design methods (Guest, personal communication, 2009). In contrast, the D_N optimal network in example 2 is optimized in a few seconds, including time needed to compute $\Sigma(\Xi)$. By borrowing from the efficiencies of linear methods, our method makes it computationally feasible to optimize experiments of many hundreds or possibly thousands of observation points in a matter of minutes, a major advance in nonlinear geoscientific SED. To demonstrate, we timed the D_N optimization workflow while designing a microseismic monitoring network (using the same priors as in

example 2) of 320 receiver stations, which took just over 4.5 minutes on a personal laptop. The linearized sequential design algorithms discussed previously are also theoretically parallelizable, so much larger D_N optimum experimental designs might be found using distributed computing.

It is notable that our method does not require evaluation of the Jacobian matrices (required for standard linear D optimal design, for example), which can be expensive to compute and formidable to store in memory, especially for Bayesian methods that require Monte Carlo integrations involving many Jacobian matrices (e.g., Chaloner and Verdinelli, 1995). Our technique requires only forward calculation and storage of the theoretical function \mathbf{g} and is therefore limited only by the expense of this calculation.

Continuing with the points of efficiency and storage, it is advisable to use a method to approximate $\Sigma(\Xi)$ (or $[\Sigma(\Xi)]^{1/2}$, which is used in practice with many LSDAs) directly and efficiently (avoiding step 3 of the design workflow) as successive random models are sampled from $p(\mathbf{m})$. We recommend recursive principal component analysis (PCA) (Peddani et al., 2004), which can update the estimate of $\Sigma(\Xi)$ iteratively and thereby avoid a bulk computation (after all samples have been collected) — a formidable task if many data samples are taken. PCA also facilitates data compression because it identifies the degrees of freedom of the nonlinear theoretical function over the domain of probable models. Properly applied, PCA can save on storage and boost the computational efficiency of LSDAs, especially if $[\Sigma(\Xi)]^{1/2}$ is used. This is because $[\Sigma(\Xi)]^{1/2}$ often can be expressed more compactly than $\Sigma(\Xi)$ because of a limited degree of freedom in \mathbf{g} , given the set of candidate observation points Ξ and the model prior $p(\mathbf{m})$. Recursive PCA was used in the microseismic example. It is also readily shown that the D_N criterion is identical to maximum entropy design criteria (Shewry and Wynn, 1987; Sebastiani and Wynn, 2000; van den Berg et al., 2003; Guest and Curtis, 2009, 2010) when data and model are related linearly and the data noise is Gaussian (c.f., Chaloner and Verdinelli, 1995).

The D_N criterion assumes that $\mathbf{g}(\mathbf{m}_i, \xi) - \mathbf{g}(\mathbf{m}_j, \xi)$ is approximately multinormal. We tested the validity of this assumption in the microseismic example by using a Shapiro-Wilk test for normality (Shapiro and Wilk, 1965). The correlation between the linear fit of the quantiles of $\mathbf{g}(\mathbf{m}_i, \xi) - \mathbf{g}(\mathbf{m}_j, \xi)$ (for each candidate receiver) and the quantiles of a normal distribution (of the same mean and variance) was always greater than 0.99. Thus, the multinormal assumption was valid in this case, as shown in Figure 5. Even if the multinormal assumption were invalid, it would only mean that Φ_{D_N} is a poor approximation of the relative entropy. The D_N criterion is nonetheless a measure of the expected Mahalanobis distance between data sets; so the D_N criterion is potentially a viable design objective regardless of multinormality.

Table 3. Percentage difference in the mean post-inversion uncertainty of each hypocentral coordinate using Heuristic and LDF optimal designs relative to those mediated by the D_N optimal network. Positive numbers represent increased certainty from the D_N -optimal designs.

	x	y	z
Heuristic	+44.7%	+16.7%	+174.0%
LDR optimal	+88.3%	+63.5%	+164.2%

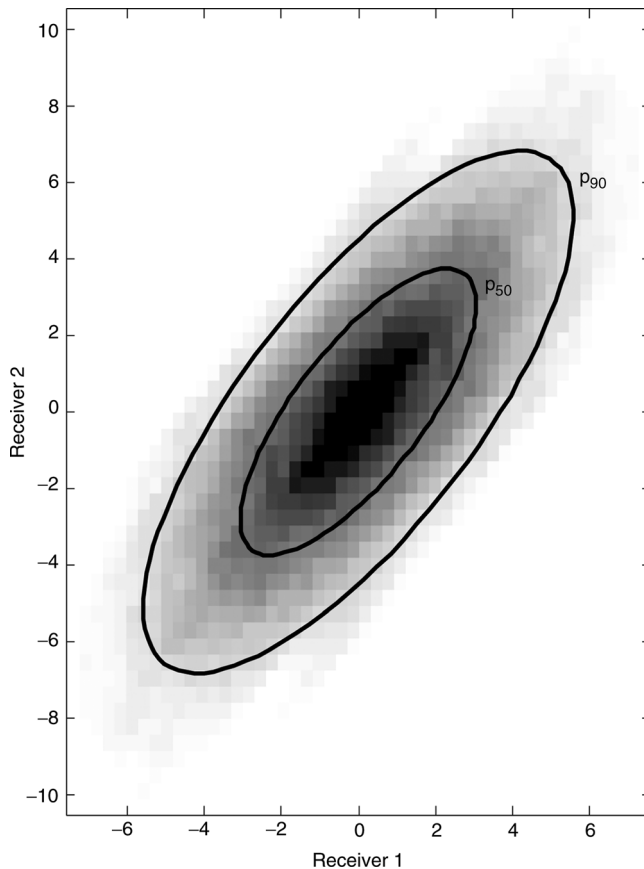


Figure 5. The expected joint distribution (gray region; dark is high density) of $\mathbf{g}(\mathbf{m}_i, \xi) - \mathbf{g}(\mathbf{m}_j, \xi)$ for two receivers ξ (enlarged triangles, Figure 4c), estimated over the set of example microseismic events in Figure 3. Overlain are 50th- and 90th-percentile regions of the best-fitting multinormal distribution.

CONCLUSION

We have presented a new method for nonlinear Bayesian statistical experimental design based on a generalization of the D criterion to nonlinear Bayesian design. The method takes advantage of efficient linear methods and has lower data-storage requirements than other nonlinear algorithms. It also appears to be robust to sampling errors, although more research is needed to confirm this. The method makes robust, industrial-scale, geoscientific survey optimization computationally feasible for nonlinear problems. When optimizing a seafloor microseismic monitoring network, our technique demonstrably reduces spatial bias and hypocentral uncertainty more than networks optimized using a linearized Bayesian design method and those designed heuristically by an industrial contractor.

REFERENCES

- Ajo-Franklin, J., 2009, Optimal experiment design for time-lapse travel-time tomography: *Geophysics*, **74**, no.4, Q27–Q40, doi:10.1190/1.3141738.
- Atkinson, A. C., A. N. Donev, and R. D. Tobias, 2007, Optimum experimental designs, with SAS: Oxford University Press.
- Barth, N., and C. Wunsch, 1990, Oceanographic experiment design by simulated annealing: *Journal of Physical Oceanography*, **20**, no.9, 1249–1263, doi:10.1175/1520-0485(1990)020<1249:OEDBSA>2.0.CO;2.
- Chaloner, K., and I. Verdinelli, 1995, Bayesian experimental design: A review: *Statistical Science*, **10**, no. 3, 273–304, doi:10.1214/ss/1177009939.
- Coles, D., and A. Curtis, 2011, A free lunch in linearized experimental design?: *Computers and Geosciences*, doi:10.1016/j.cageo.2010.09.012.
- Coles, D., and F. D. Morgan, 2009, A method of fast, sequential experimental design for linearized geophysical inverse problems: *Geophysical Journal International*, **178**, no. 1, 145–158, doi:10.1111/j.1365-246X.2009.04156.x.
- Cover, T., and J. Thomas, 1991, *Elements of information theory*: Wiley Interscience.
- Curtis, A., 1999a, Optimal experiment design: Cross-borehole tomographic examples: *Geophysical Journal International*, **136**, no. 3, 637–650, doi:10.1046/j.1365-246x.1999.00749.x.
- , 1999b, Optimal design of focused experiments and surveys: *Geophysical Journal International*, **139**, no.1, 205–215, doi:10.1046/j.1365-246X.1999.00947.x.
- Curtis, A., A. Michelini, D. Leslie, and A. Lomax, 2004, A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys: *Geophysical Journal International*, **157**, no. 2, 595–606, doi:10.1111/j.1365-246X.2004.02114.x.
- Goldberger, J., S. Gordon, and H. Greenspan, 2003, An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures: *Proceedings of the 9th International Conference on Computer Vision, IEEE*, 487–493.
- Guest, T., and A. Curtis, 2009, Iteratively constructive sequential design of experiments and surveys with nonlinear parameter-data relationships: *Journal of Geophysical Research — Solid Earth and Planets*, **114**, no. B4, B04307, doi:10.1029/2008JB005948.
- , 2010, Optimal trace selection for AVA processing of shale-sand reservoirs: *Geophysics*, **75**, no. 4, C37–C47.
- Haber, E., L. Horesh, and L. Tenorio, 2008, Numerical methods for experimental design of large-scale linear ill-posed inverse problems: *Inverse Problems*, **24**, no. 5, 055012, doi:10.1088/0266-5611/24/5/055012.
- Hutton, L. K., and D. M. Boore, 1987, The ML scale in southern California: *Bulletin of the Seismological Society of America*, **77**, 2074–2094.
- Hyvärinen, A., J. Karhunen, and E. Oja, 2001, *Independent component analysis*: John Wiley & Sons, Inc.
- Khodja, M. R., M. Prange, and H. Djikpesse, 2010, Guided Bayesian optimal experimental design: *Inverse Problems*, **26**, no. 5, 055008, doi:10.1088/0266-5611/26/5/055008.
- Mahalanobis, P., 1936, On the generalised distance in statistics: *Proceedings of the National Institute of Science, India* **2**, 49–55.
- Narayanan, C., V. N. R. Rao, and J. M. Kaihatu, 2004, Model parameterization and experimental design issues in nearshore bathymetry inversion: *Journal of Geophysical Research — Oceans*, **109**, no. C8, C08006, doi:10.1029/2002JC001756.
- Peddani, H., D. Erdogmus, Y. N. Rao, A. Hegde, and J. C. Principe, 2004, Recursive principal components analysis using eigenvector matrix perturbation: *Proceedings of the IEEE Signal Processing Society Workshop*, 83–92.
- Pukelsheim, F., 2006, *Optimal design of experiments*: Society for Industrial and Applied Mathematics.
- Sebastiani, P., and H. P. Wynn, 2000, Maximum entropy sampling and optimal Bayesian experimental design: *Journal of the Royal Statistical Society Series B — Statistical Methodology*, **62**, no. 1, 145–157, doi:10.1111/1467-9868.00225.
- Shapiro, S. S., and M. B. Wilk, 1965, An analysis of variance test for normality: *Biometrika*, **52**, 591–611.
- Shewry, M. C., and H. P. Wynn, 1987, Maximum entropy sampling: *Journal of Applied Statistics*, **14**, no. 2, 165–170, doi:10.1080/02664768700000020.
- Steinberg, D. M., N. Rabinowitz, Y. Shimshoni, and D. Mizrahi, 1995, Configuring a seismographic network for optimal monitoring of fault lines and multiple sources: *Bulletin of the Seismological Society of America*, **85**, 1847–1857.
- Stummer, P., H. Maurer, and A. G. Green, 2004, Experimental design: Electrical resistivity data sets that provide optimum subsurface information: *Geophysics*, **69**, 120–139, doi:10.1190/1.1649381.
- van den Berg, J., A. Curtis, and J. Trampert, 2003, Optimal nonlinear Bayesian experimental design: An application to amplitude versus offset experiments: *Geophysical Journal International*, **155**, 411–421, doi:10.1046/j.1365-246X.2003.02048.x.
- Winterfors, E., and A. Curtis, 2008, Numerical detection and reduction of non-uniqueness in nonlinear inverse problems: *Inverse Problems*, **24**, no. 2, 025016, doi:10.1088/0266-5611/24/2/025016.
- Wunsch, C., 1996, *The ocean circulation inverse problem*: Cambridge University Press.