Learning functions and approximate Bayesian computation design: ABCD

Entropy **2014**, *16*, 4353-4374; doi:10.3390/e16084353

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Learning Functions and Approximate Bayesian Computation Design: ABCD

Markus Hainy [1], Werner G. Müller [1] and Henry P. Wynn [2,*]

[1] Department of Applied Statistics, Johannes Kepler University, 4040 Linz, Austria;
E-Mails: Markus.Hainy@jku.at (M.H.); Werner.Mueller@jku.at (W.G.M.)
[2] Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK

* Author to whom correspondence should be addressed; E-Mail: h.wynn@lse.ac.uk;
Tel.: +44-(0)20-7955-6116.

Received: 25 April 2014; in revised form: 18 July 2014 / Accepted: 28 July 2014 /
Published: 4 August 2014

Abstract: A general approach to Bayesian learning revisits some classical results, which study which functionals on a prior distribution are expected to increase, in a preposterior sense. The results are applied to information functionals of the Shannon type and to a class of functionals based on expected distance. A close connection is made between the latter and a metric embedding theory due to Schoenberg and others. For the Shannon type, there is a connection to majorization theory for distributions. A computational method is described to solve generalized optimal experimental design problems arising from the learning framework based on a version of the well-known approximate Bayesian computation (ABC) method for carrying out the Bayesian analysis based on Monte Carlo simulation. Some simple examples are given.

Keywords: learning; Shannon information; majorization; optimum experimental design; approximate Bayesian computation

1. Introduction

A Bayesian approach to the optimal design of experiments uses some measure of preposterior utility, or information, to assess the efficacy of an experimental design or, more generally, the choice of sampling distribution. Various versions of this approach have been developed by Blackwell [1], and Torgerson [2]

gives a clear account. Renyi [3], Lindley [4] and Goel and DeGroot [5] use information-theoretic approaches to measure the value of an experiment; see also the review paper by Ginebra [6]. Chaloner and Verdinelli [7] give a broad discussion of the Bayesian design of experiments, and Wynn and Sebastiani [8] also discuss the Bayes information-theoretic approach. There is wider interest in these issues in cognitive science and epistemology; see Chater and Oaksford [9].

When new data arrives, one can expect to improve the information about an unknown parameter θ. The key theorem, which is Theorem 2 here, gives conditions on informational functionals for this to be the case, and then, they will be called learning functionals. This class includes many special types of information, such as Shannon information, as special cases.

Section 2 gives the main theorems on learning functionals. We give our own simple proofs for completion, and the material can be considered as a compressed summary of what can be found in quite a scattered literature. We study two types of learning function, those of which we shall call the Shannon type and, in Section 3, those based on distances. For the latter, we shall make a new connection to the metric embedding theory contained in the work of Schoenberg with a link to Bernstein functions [10,11]. This yields a wide class of new learning functions. Following two, somewhat provocative, counter-examples and a short discussion of surprise in Section 4, we relate learning functions of the Shannon type to the theory of majorization in Section 5. Section 6 specializes learning functions on covariance matrices.

We shall use the classical Bayes formulation with θ as an unknown parameter with a prior density $\pi(\theta)$ on a parameter space Θ and a sampling density $f(x|\theta)$ on an appropriate sample space. We denote by $f_{X,\theta}(x,\theta) = f(x|\theta)\pi(\theta)$ the joint density of X and θ and use $f_X(x)$ for the marginal density of X. The nature of expectations will be clear from the notation. To make the development straightforward, we shall look at the case of distributions with densities (with respect to Lebesgue measure) or, occasionally, discrete distributions with finite support. All necessary conditions for conditional densities, integration and differentiation will be implicitly assumed.

In Section 7, approximate Bayesian computation (ABC) is applied to problems in optimal experimental design (hence, ABCD). We believe that an understanding of modern optimal experimental design and its computational aspects needs to be grounded in some understanding of learning. At the same time, there is added value in taking a wide interpretation of optimal design as a choice, with constraints, of the sampling distribution $f(x|\theta)$. Thus, one may index $f(x|\theta)$ by a control variable z and write $f(x|\theta, z)$ or $f(x(z)|\theta)$. Certain aspects of the distribution may depend on z, others not. An experimental design can be taken as the choice of a set of z, at each of which we take one or more observations, giving a multivariate distribution. In areas, such as search theory and optimization, z may be a site at which one measures or observes with error. In spatial sampling, one may also use the term "site" for z. However, z could be a simple flag, which indicates one or another of somewhat unrelated experiments to estimate a common θ. In medicine, for example, one discusses different types of "intervention" for the same patient.

2. Information-Based Learning

The classical formulation proceeds as follows. Let U be a random variable with density $f_U(u)$. Let $g(\cdot)$ be a function on $R+ \rightarrow R$ and define a measure of information of the Shannon type for U with respect to g as

$$I_g(U) = \mathrm{E}_U(g(f_U(U))).$$

When $g(u) = \log(u)$, we have Shannon information. When $g(u) = \frac{u^\gamma - 1}{\gamma}$, $(\gamma > -1)$, we have a version similar to Renyi information, which is sometimes called Tsallis information [12].

If X represents the future observation, we can measure the preposterior information of the experiment (query, *etc.*), which generates a realization of X, by the prior expectation of the posterior information, which we define as:

$$I_g(\theta; X) = \mathrm{E}_X \mathrm{E}_{\theta|X}(g(\pi(\theta|X))) = \mathrm{E}_{X,\theta}(g(\pi(\theta|X))).$$

In the second term, the inner expectation is with respect to the posterior (conditional) distribution of θ given X, namely $\pi(\theta|X)$, and the outside expectation is with respect to the marginal distribution of X. In the last term, the expectation is with respect to the full joint distribution of X and θ. We wish to compare $I_g(\theta; X)$ with the prior information:

$$I_g(\theta) = \mathrm{E}_\theta(g(\pi(\theta))).$$

Theorem 1. *For fixed $g(u)$ and the standard Bayesian set-up, the pre-posterior quantity $I_g(\theta, X)$ and prior value, $I_g(\theta)$, satisfy:*

$$I_g(\theta; X) \geq I_g(\theta) = E_\theta(g(\pi(\theta))),$$

for all joint distributions $f_{X,\theta}(x, \theta)$ if and only if $h(u) = ug(u)$ is convex on R^+.

We shall postpone the proof of Theorem 1 until after a more general result for functionals on densities:

$$\phi : \pi(\theta) \mapsto R.$$

Theorem 2. *For the standard Bayesian set-up and a functional $\phi(\cdot)$,*

$$\phi(\pi(\theta)) \leq E_X \phi(\pi(\theta|X))$$

for all joint distributions $f_{X,\theta}(x, \theta)$ if and only if ϕ is convex as a functional:

$$\phi((1-\alpha)\pi_1 + \alpha\pi_2) \leq (1-\alpha)\phi(\pi_1) + \alpha\phi(\pi_2),$$

for $0 \leq \alpha \leq 1$ and all π_1, π_2.

Proof. Note that taking expectations with respect to the marginal distribution of X amounts to a convex mixing, not dependent on θ. Thus, using Jensen's inequality:

$$\begin{aligned} \mathrm{E}_X(\phi(\pi(\theta|X))) &\geq \phi(\mathrm{E}_X(\pi(\theta|X))) \\ &= \phi(\pi(\theta)). \end{aligned}$$

The necessity comes from a special construction. We show that given a functional $\phi(\cdot)$ and a triple $\{\pi_1, \pi_2, \alpha\}$, such that:

$$\phi((1-\alpha)\pi_1 + \alpha\pi_2) > (1-\alpha)\phi(\pi_1) + \alpha\phi(\pi_2),$$

we can find a pair $\{f(x,\theta),\pi(\theta)\}$, such that

$$\phi(\pi(\theta)) > \mathrm{E}_X \phi(\pi(\theta|x)). \tag{1}$$

Thus, let X be a Bernoulli random variable with marginal distribution $(\text{prob}\{X=0\},\text{prob}\{X=1\}) = (1-\alpha,\alpha)$. Then, it is straightforward to choose a joint distribution of θ and X, such that:

$$\pi(\theta|X=0) = \pi_1(\theta),\ \pi(\theta|X=1) = \pi_2(\theta),$$

from which we obtain (1). □

Proof. (of Theorem 1). We now show that Theorem 1 is a special case of Theorem 2.

Write $\pi_\alpha(\theta) = (1-\alpha)\pi_1(\theta) + \alpha\pi_2(\theta)$. If $h(u) = ug(u)$ is convex as a function of its argument u:

$$\int h(\pi_\alpha(\theta))d\theta \leq \int \left((1-\alpha)h(\pi_1(\theta)) + \alpha h(\pi_2(\theta))\right) d\theta \tag{2}$$

$$= (1-\alpha)\int h(\pi_1(\theta))d\theta + \alpha\int h(\pi_2(\theta))d\theta, \tag{3}$$

proving one direction.

The reverse is to show that if I_g is convex for all π, then h is convex. For this, again, we need a special construction. We carry this out on one dimension, the extension to more than one dimension being straightforward. For ease of exposition, we also make the necessary differentiability conditions. The second directional derivative of $I_g(\theta)$ in the space of distributions (which is convex) at π_1 towards π_2 is:

$$\frac{\partial^2}{\partial\alpha^2}\int g(\pi_\alpha(\theta))\pi_\alpha(\theta)d\theta\Big|_{\alpha=0} = \int (\pi_1-\pi_2)^2(g''(\pi_1)\pi_1 + 2g'(\pi_1))d\theta.$$

Let π_1 represent a uniform distribution on $[0,\frac{1}{z}]$, for some $z \geq 0$, and let π_2 be a distribution with support contained in $[0,\frac{1}{z}]$. Then, the above becomes:

$$\int_0^{\frac{1}{z}} (z-\pi_2(\theta))^2(g''(z)z + 2g'(z))d\theta = (g''(z)z + 2g'(z))\int_0^{\frac{1}{z}} (z-\pi_2(\theta))^2 d\theta.$$

Now, assume that $h(z) = zg(z)$ is not convex at z; then $h''(z) = g''(z)z + 2g'(z) < 0$ and any choice of π_2, which makes the integral on the right-hand side positive, shows that $I_g(\theta)$ is not convex at z. This completes the proof. □

Theorem 2 has a considerable history of discovery and rediscovery and, in its full version, should probably be attributed to DeGroot [13]; see Ginebra [6]. The early results concentrated on functionals of the Shannon type, basically yielding Theorem 1. Note that the condition $h(u) = ug(u)$ being convex on R^+ is equivalent to $g(\frac{1}{u})$ being convex, which is referred to as $g(u)$ being "reciprocally convex" by Goldman and Shaked [14]; see also Fallis and Lyddell [15].

3. Distance-Based Information Functions

Shannon type information functionals take no account of metrics. Intuitively, if mass is moved around, the information stays the same. Let Z_1, Z_2 be independent copies from $\pi(z)$, and let $d(z_1, z_2)$ be a distance or metric. Define d-information as:

$$\phi(\pi) = -\mathrm{E}_{Z_1,Z_2}(d(Z_1, Z_2)^2).$$

Now, with $\pi_\alpha(z) = (1-\alpha)\pi_1(z) + \alpha\pi_2(z)$,

$$\phi(\pi_\alpha) = -\int\int d(z_1, z_2)^2((1-\alpha)\pi_1(z_1) + \alpha\pi_2(z_1))((1-\alpha)\pi_1(z_2) + \alpha\pi_2(z_2))dz_1dz_2. \tag{4}$$

The condition for convexity, again using the second directional derivative with respect to α, is

$$-\int\int d(z_1, z_2)^2(\pi_1(z_1) - \pi_2(z_1))(\pi_1(z_2) - \pi_2(z_2))dz_1dz_2 \geq 0. \tag{5}$$

Noting that $\int(\pi_1(z_1) - \pi_2(z_1)) = 0$, (5) is a generalized version of the following condition:

$$-\sum_{ij} d(z_i, z_j)z_iz_j \geq 0, \text{ for all } z, \sum z_i = 0. \tag{6}$$

Condition (6), considered as a condition on a distance matrix $d_{ij} = d(z_i, z_j)$, is called almost positive and is the necessary and sufficient condition for an abstract set of points $P_1, \ldots, P_k$, with interpoint distances $\{d_{ij}\}$, to be embedded in Euclidean space.

Theorem 3. *If $d_{ij} = d_{ji}$, $1 \leq i < j \leq n$, are $\frac{1}{2}n(n-1)$ positive quantities, then a necessary and sufficient condition that the d_{ij} are the interpoint distances between points P_i, $i = 1, \ldots, n$, in R^n is that the distance matrix $D = -\{d_{ij}\}$ is an almost positive matrix.*

This is a special case of metric embedding, sometimes called metric multi-dimensional scaling, in statistics; see, for example, Torgeson [16], Gower [17,18]. A more general result is:

Theorem 4. *Let S be a separable metric with metric space with metric $d(x, y)$, then S can be isometrically embedded in l_2 if and only if $A(x, y) = -d(x, y)$ is an almost positive matrix.*

It is a task to identify the functions $B(d(x, y)^2)$, such that, when $d(x, y)$ is a Euclidean or Hilbert space metric, the space with the new metric can still be embedded into the Hilbert space. Schoenberg [10] gives the following major result that such $B(\cdot)$ comprise the Bernstein function defined as follows (see Theorem 12.14 in [11]):

Definition 1. *A function $B : (0, \infty) \mapsto R$ is a Bernstein function if it is C^∞, $f(\lambda) \geq 0$ for all $\lambda > 0$ and the derivatives $f^{(n)}$ satisfy $(-1)^{n-1}f^{(n)} \geq 0$ for all positive integers n and all $\lambda > 0$.*

Note that this says that f' is a completely monotone function.

Theorem 5. *(Schoenberg) The following are equivalent:*

(1) $B(||x-y||^2)$ $(x, y \in H)$ *is the square of a distance function, which isometrically embeds into Hilbert space* H*, i.e., there exists a* $\phi : H \mapsto H$*, such that:*

$$B(||x-y||^2) = ||\phi(x) - \phi(y)||^2 . \tag{7}$$

(2) B *is a Bernstein function.*

(3) $e^{-B(t)}$ *is the Laplace transform of an infinitely divisible distribution, i.e.,*

$$B(t) = -\log \int_0^\infty \frac{e^{-tu}}{u} d\gamma(u),$$

where γ *is an infinitely divisible distribution.*

(4) B *has the Lévy-Khintchine representation:*

$$B(t) = B_{\mu,b}(t) = bt + \int_0^\infty (1 - e^{-tu}) d\mu(u) \tag{8}$$

for some $b \geq 0$ *and a measure* μ*, such that* $\int_0^\infty (1 \wedge t) d\mu(t) < \infty$*, with the condition that* $B_{\mu,b}(t) > 0$ *for* $t > 0$*.*

We now combine the above discussion with Schoenberg's theorem.

Theorem 6. *If* $B(\cdot)$ *is a Bernstein function with* $B(0) = 0$ *and* $d(z_1, z_2)$ *is a Euclidean distance, then* $\phi(\pi) = -E_{Z_1,Z_2}(B(d(Z_1, Z_2)^2))$ *is a learning function.*

In the univariate case the negative of the variance of the distribution is a learning function since:

$$\text{var}(Z) = \frac{1}{2}\text{E}_{Z_1,Z_2}(Z_1 - Z_2)^2.$$

When Z is multivariate, we again take independent copies Z_1, Z_2 of Z and use Euclidean distance, and we have that minus the trace of the covariance matrix of Z, Γ, is a learning function:

$$\frac{1}{2}\text{E}_{Z_1,Z_2}(||Z_1 - Z_2||^2) = \text{trace}(\Gamma).$$

Schilling *et al.* [11] (Chapter 15) list 138 Bernstein functions, each of which will lead to a learning functional of the distance type. We give a small selection of Bernstein functions $B(\lambda)$, which then, applied with $\lambda = d(z_1, z_2)^2$, give a learning function:

$$\begin{aligned} \lambda^\alpha, &\quad 0 < \alpha < 1, \\ (1+\lambda)^\alpha - 1, &\quad 0 < \alpha < 1, \\ 1 - (1+\lambda)^{\alpha-1}, &\quad 0 < \alpha < 1, \\ \frac{\lambda}{\lambda + \alpha}, &\quad \alpha > 0. \end{aligned}$$

4. Counterexamples

We show first that it is not true that information always increases. That is, it is not true that the posterior information is always more than the prior information:

$$I_g(\theta) \leq \mathrm{E}_{\theta|X}(g(\pi(\theta|X))).$$

A simple discrete example runs as follows. I have lost my keys. With high prior probability, p, I think they are on my desk. Suppose I have a uniform prior over all k likely other locations. However, suppose when I look on the desk that my keys are not there. My posterior distribution is now uniform on the other locations. Under certain conditions on p and k, Shannon information has gone down. For fixed p, the condition is $k > k^*$ where:

$$\begin{aligned} k^* &= \frac{(1-p)^{1-\frac{1}{p}}}{p} \\ &= e \cdot \left(\frac{1}{p} - \frac{1}{2} + O(p) \right), \end{aligned}$$

by expanding pk^* in a Taylor expansion. When $p = \frac{1}{2}$, $k^* = 4$ and $pk^* \to e, 1$ when $p \to 0, 1$. This example is captured by the somewhat self-doubting phrase "if my keys are not on my desk, I don't know where they are". Note, however, that something has improved: the support size is reduced from $k+1$ to k.

There is a simple way of obtaining a large class of examples, namely to arrange that there are x-values for which the posterior distribution is approximately uniform. Then, because the uniform distribution typically has low information, for such x, we can have a decrease in information. Thus, we construct examples in which $f(x|\theta)\pi(\theta)$ happens to be approximately constant for some x. This motivates the following example.

Let $\Theta \times \mathcal{X} = [0,1]^2$ with joint distribution having support on $[0,1]^2$. Let $\pi(\theta)$ be the prior distribution and define a sampling distribution:

$$f(x|\theta) = a(\theta)(1-x) + \frac{x}{\pi(\theta)}.$$

Note that we include the prior distribution into the sampling distribution as a constructive device, not as some strange new general principle. We have in mind, in giving this construction, that when $x \to 1$, the first term should approach zero and the second term, after multiplying by $\pi(\theta)$, should approach unity. Solving for $a(\theta)$ by setting $\int_0^1 f(x|\theta)dx = 1$, we have $a(\theta) = \frac{2\pi(\theta)-1}{\pi(\theta)}$ so that:

$$f(x|\theta) = \frac{(2\pi(\theta)-1)(1-x)+x}{\pi(\theta)}.$$

The joint distribution is then:

$$f(x|\theta)\pi(\theta) = (2\pi(\theta)-1)(1-x)+x. \tag{9}$$

The marginal distribution of X is $f_X(x) = 1$ on $[0,1]$, since the integral of (9) is unity, so that (9) is also the posterior distribution $\pi(\theta|x)$. Note that, in order for (9) to be a proper density, we require that $\pi(\theta) \geq \frac{1}{2}$ for $0 \leq \theta \leq 1$.

The Shannon information of the prior is:

$$I_0 = \int_0^1 \pi(\theta) \log \pi(\theta) d\theta,$$

and of the posterior is

$$I_1 = \int_0^1 ((2\pi(\theta) - 1)(1 - x) + x) \log((2\pi(\theta) - 1)(1 - x) + x) d\theta.$$

When $x = \frac{1}{2}$, the integrands of I_1 and I_0 are equal and $I_0 = I_1$. When $x = 1$, the integrand of I_1 is zero, as expected. Thus, for a non-uniform prior, we have less posterior information in a neighborhood of $x = 1$, as we aimed to achieve.

Specializing $\pi(\theta) = \frac{1}{2} + \theta$ on $[0, 1]$ gives:

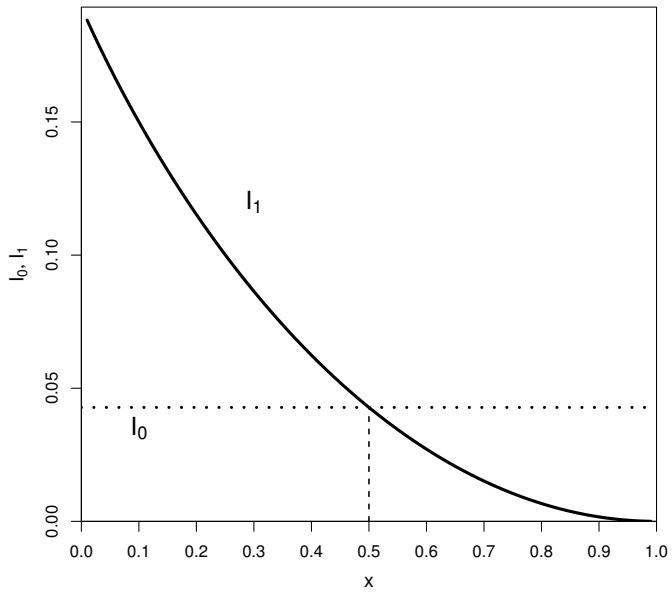
$$\begin{aligned} I_0 &= \tfrac{9}{8} \log 3 - \log 2 - 1/2 \\ I_1 &= \tfrac{1}{4(1-x)}((2 - x)^2 \log(2 - x) - x^2 \log(x) + 2x - 2) \end{aligned}$$

Information I_1 decreases from a maximum of $\log(2) - \frac{1}{2}$ at $x = 0$, through the value I_0 at $x = \frac{1}{2}$, to the value zero at $x = 1$; see also Figure 1. Thus, $I_0 > I_1$ for $\frac{1}{2} < x \leq 1$. Since the marginal distribution of X is uniform on $[0, 1]$, we have the challenging fact that:

$$\text{prob}_X\{I_1 < I_0\} = \frac{1}{2}.$$

Namely, with prior probability equal to one half, there is less Shannon information in the posterior than the prior. The Renyi entropy exhibits the same phenomenon, but we omit the calculations. We might say that $f(x|\theta)$ is not a good choice sampling distribution to learn about θ.

Figure 1. Shannon information of the prior, I_0, and of the posterior, I_1, depending on x.



4.1. Surprise and Ignorance

The conflict between prior beliefs and empirical data, demonstrated by these examples, lies at the heart of debates about inference and learning, that is to say epistemology. This has given rise to formal theories of surprise, which seek to take account of the conflict. Some Bayesian theories are closely related to the learning theory discussed here and measure surprise quantities, such as the difference:

$$S(\pi, f) = I_g(\theta) - \mathrm{E}_{\theta|X} g(\pi(\theta|X)).$$

Since, under the conditions of Theorem 1, S is expected to be negative, a positive value is taken to measure surprise; see Itti and Baldi [19].

Taking a subjective view of these issues, we may stray into cognitive science, where there is evidence that the human brain may react in a more focused way than normal when there is surprise. This is related to wider computational models of learning: given the finite computational capacity of the brain, we need to use our sensing resources carefully in situations of risk or utility. One such body of work emanates from the so-called "cocktail party effect": if the subject matter is of sufficient interest, such as the mention of one's own name across a crowded room, then one's attention is directed towards the conversation. Discussions about how the attention is first captured are closely related to surprise; see Haykin and Chen [20].

4.2. Minimal Information Prior Distributions

It is clear that if the prior distribution has minimal information (maximum entropy), then there is no surprise, because S, as defined above, is never positive. The use of such prior distributions has been advocated for many years and is incorporated into objective Bayesian analysis by some researchers. One key idea is to use Jeffrey's prior distributions, that is those which are invariant under a suitable group (Haar measure); for a discussion, see Berger [21].

An unresolved issue is that the minimal information distribution depends on the learning function. A simple example is that for Shannon information, the minimal information distribution with support on $[0, 1]$ is the uniform distribution, whereas the maximum variance distribution has mass $\frac{1}{2}$ at each of $\{0, 1\}$ and variance $\frac{1}{4}$, which is achievable for the Beta(α, β) distribution as $\alpha, \beta \to 1$. The variance of the uniform distribution, on the other hand, is $\frac{1}{12} < \frac{1}{4}$.

Consider the standard beta-binomial Bayesian set-up, where the sampling distribution is Bin(n, θ) and the (conjugate) prior is Beta(α, β). If x is the data, the posterior distribution is Beta$(\alpha + x, \beta + n - x)$, and the posterior mean, which is the Bayes estimator with respect to quadratic loss, is $\hat{\theta} = \frac{\alpha+x}{\alpha+\beta+n}$. The minimal Shannon information is achieved for the uniform distribution when $\alpha, \beta \to 1$, in which case we have $\hat{\theta} = \frac{1+x}{2+n}$. However, if we take $\alpha, \beta \to 0$, giving, as mentioned, the minimal information with respect to the variance, we obtain in the limit the maximum likelihood estimator $\frac{x}{n}$. The same feature arises with the Dirichlet-multinomial case, with the Dirichlet prior distribution: $\pi(\theta_1, \ldots, \theta_k) = \frac{\prod \theta_i^{\alpha_i - 1}}{\mathrm{Beta}(\alpha_1, \ldots, \alpha_k)}$. The minimal Shannon information is uniform when all $\alpha_i = 1$, but the minimal trace of the covariance matrix is for mass $\frac{1}{k}$ at each corner of the simplex $\sum \theta_i = 1$.

5. The Role of Majorization

We concentrate here on Shannon-type learning functions. The analysis of the last section leads to the notion that for two distributions $\pi_1(\theta)$ and $\pi_2(\theta)$, the second is more peaked than the first if and only if:

$$\int_\Theta h(\pi_1(\theta))d\theta \leq \int_\Theta h(\pi_2(\theta))d\theta \text{ for all convex } h(u) = ug(u) \text{ on } R^+. \tag{10}$$

The statement (10) defines a partial ordering between π_1 and π_2.

For Bayesian learning, we may hope that the ordering holds when π_1 is the prior distribution and π_2 is the posterior distribution. We have seen from the counterexamples that it does not hold in general, but, loosely speaking, always holds in expectation, by Theorem 1. However, it is natural to try to understand the partial ordering, and we shall now indicate that the ordering is equivalent to a well-known majorization ordering for distributions.

Consider two discrete distributions with n-vectors of probabilities $\pi_1 = (\pi_1^{(1)}, \ldots, \pi_n^{(1)})$ and $\pi_2 = (\pi_1^{(2)}, \ldots, \pi_n^{(2)})$, where $\sum_i \pi_i^{(1)} = \sum_i \pi_i^{(2)} = 1$. First, order the probabilities:

$$\tilde{\pi}_1^{(1)} \geq \ldots \geq \tilde{\pi}_n^{(1)}, \quad \tilde{\pi}_1^{(2)} \geq \ldots \geq \tilde{\pi}_n^{(2)}.$$

Then, π_2 is said to majorize π_1, written $\pi_1 \preceq \pi_2$, when:

$$\sum_{i=1}^{j} \tilde{\pi}_i^{(1)} \leq \sum_{i=1}^{j} \tilde{\pi}_i^{(2)}$$

for $j = 1, \ldots, n$ (equality for $j = n$). The standard reference is Marshall and Olkin [22], where one can find several equivalent conditions. Two of the best known are:

A1. there is a doubly stochastic matrix $P_{n\times n}$, such that $\pi_1 = P\pi_2$;
A2. $\sum_i^n h(\pi_i^{(1)}) \leq \sum_i^n h(\pi_i^{(2)})$ for all continuous convex functions $h(x)$.

Condition A2 shows that, in the discrete case, the partial ordering (10) is equivalent to the majorization of the raw probabilities.

We now extend this to the continuous case. This generalization, which we shall also call $\preceq$, to save notation, has a long history, and the area is historically referred to as the theory of the "rearrangements of functions" to respect the terminology of Hardy *et al.* [23]. It is particularly well-suited to probability density functions, because $\int \pi_1(\theta)d\theta = \int \pi_2(\theta)d\theta = 1$. The natural analogue of the ordered values in the discrete case is that every density π has a unique density $\tilde{\pi}$, called a "decreasing rearrangement", obtained by a reordering of the probability mass to be non-increasing, by direct analogy with the discrete case above. In the theory, π and $\tilde{\pi}$ are then referred to as being equimeasurable, in the sense that the supports are transformed in a measure-preserving way.

There are short sections on the topic in Marshall and Olkin [22] and in Müller and Stoyan [24]. A key paper in the development is Ryff [25]. The next paragraph is a brief summary.

Definition 2. *Let $\pi(z)$ be a probability density and define $m(y) = \mu\{z : \pi(z) \geq y\}$. Then:*

$$\tilde{\pi}(t) = \textit{sup}\{y : \ m(y) > t\}, \ t > 0$$

is called the decreasing rearrangement of $\pi(z)$.

The picture is that the probability mass (in infinitely small intervals) is moved, so that a given mass is to the left of any smaller mass. For example, for the triangular distribution:

$$\pi(\theta) = \begin{cases} 4\theta, & 0 \leq \theta < \frac{1}{2} \\ 4(1-\theta), & \frac{1}{2} \leq \theta \leq 1 \end{cases}$$

we have:

$$\tilde{\pi}(\theta) = 2(1-\theta),\ 0 \leq \theta \leq 1.$$

Definition 3. *We say that π_2 majorizes π_1, written $\pi_1 \preceq \pi_2$, if and only if, for the decreasing rearrangements,*

$$\int_0^c \tilde{\pi}_1(z)dz \leq \int_0^c \tilde{\pi}_2(z)dz$$

for all $c > 0$.

Define a doubly stochastic kernel $P(x, y) \geq 0$ on $(0, \infty)$, that is:

$$\int_x P(x, y) = \int_y P(x, y) = 1.$$

There is a list of key equivalent conditions to $\preceq$, which are the continuous counterparts of the discrete majorization conditions. The first two generalize A1 and A2 above.

B1. $\pi_1(\theta) = \int_\Theta P(\theta, z)\pi_2(z)dz$ for some non-negative doubly stochastic kernel $P(x, y)$.
B2. $\int_\Theta h(\pi_1(z))dz \leq \int_\Theta h(\pi_2(z))dz$ for all continuous convex functions h.
B3. $\int_\Theta (\pi_1(z) - c)_+ dz \leq \int_\Theta (\pi_2(z) - c)_+ dz$ for all $c > 0$.

Condition B2 is the key, for it shows that in the univariate case, if we assume that $h(u) = ug(u)$ is continuous and convex, (10) is equivalent to $\pi_1(\theta) \preceq \pi_2(\theta)$. We also see that $\preceq$ is equivalent to standard first order stochastic dominance of the decreasing rearrangements, since $\tilde{F}(\theta) = \int_0^\theta \tilde{\pi}(z)dz$ is the cdf corresponding to $\tilde{\pi}(\theta)$. Condition B3 says that the probability mass under the density above a "slice" at height c is more for π_2 than for π_1.

We can summarize this discussion by the following.

Proposition 1. *A functional is a learning functional of the Shannon type (under mild conditions) if and only if it is an order-preserving functional with respect to the majorization ordering on distributions.*

The role of majorization has been noticed by DeGroot and Fienberg [26] in the related area of proper scoring rules.

The classic theory of rearrangements is for univariate distributions, whereas, as stated, we are interested in θ of arbitrary dimension. In the present paper, we will simply make the claim that the interpretation of our partial ordering in terms of decreasing rearrangements can indeed be extended to the multivariate case. Heuristically, this is done as follows. For a multivariate distribution, we may create a univariate rearrangement by considering a decreasing threshold and "squashing" all of the multivariate mass for which the density is above the threshold to a univariate mass adjacent to the origin. Since we are transforming multivariate volume to area, care is needed with Jacobians. We can then use the univariate

development above. It is an instructive exercise to consider the univariate decreasing rearrangement of the multivariate normal distribution, but we omit the computations here.

6. Learning Based on Covariance Functions

If we restrict our functionals to those which are only functionals of covariance matrices, then we can prove wider results than just for the trace. Dawid and Sebastiani [27] (Section 4) refer to dispersion-coherent uncertainty functions and, where their results are close to ours, we differ only by assumptions.

We use the notation $A \geq 0$ to mean that a symmetric matrix is non-negative definite.

Definition 4. *For two $n \times n$ symmetric non-negative definite matrices A and B, the Loewner ordering $A \geq B$ holds when $A - B \geq 0$.*

Definition 5. *A function $\phi : A \mapsto R$ on the class of non-negative definite matrices A is said to be Loewner increasing (also called matrix monotone) if $A \geq B \Rightarrow \phi(A) \geq \phi(B)$.*

Theorem 7. *A function ϕ is Loewner increasing and concave on the class of covariance matrices $\Gamma(\pi)$ if and only if $-\phi$ is a learning function on the corresponding distributions.*

Proof. Assume ϕ is Loewner increasing. To simplifying the notation, we call $\mu(\pi)$ and $\Gamma(\pi)$ the mean vector and covariance matrix, respectively, of the random variable Z with distribution π. Now, consider a mixed density $\pi_\alpha = (1-\alpha)\pi_1 + \alpha\pi_2$. Then, with obvious notation,

$$\begin{aligned}
\Gamma(\pi_\alpha) &= \mathrm{E}_\alpha(ZZ^T) - \mu_\alpha\mu_\alpha^T \\
&= (1-\alpha)\Gamma_1 + \alpha\Gamma_2 + (1-\alpha)\mu_1\mu_1^T + \alpha\mu_2\mu_2^T - ((1-\alpha)\mu_1 + \alpha\mu_2)((1-\alpha)\mu_1 + \alpha\mu_2)^T \\
&= (1-\alpha)\Gamma_1 + \alpha\Gamma_2 + \alpha(1-\alpha)(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \\
&\geq (1-\alpha)\Gamma_1 + \alpha\Gamma_2,
\end{aligned}$$

for $0 \leq \alpha \leq 1$, since $(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ is non-negative definite. Then, since ϕ is Loewner increasing and concave, $\phi(\Gamma(\pi_\alpha)) \geq \phi\left((1-\alpha)\Gamma(\pi_1) + \alpha\Gamma(\pi_2)\right) \geq (1-\alpha)\phi(\Gamma(\pi_1)) + \alpha\phi(\Gamma(\pi_2))$, and by Theorem 2, $-\phi$ is a learning function.

We first prove the converse for matrices Γ and $\tilde{\Gamma} = \Gamma + zz^T$, for some vector z. Take two distributions with equal covariance functions, but with means satisfying $\mu_1 - \mu_2 = 2z$. Then,

$$\begin{aligned}
\Gamma(\pi_{\frac{1}{2}}) &= \Gamma + \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \\
&= \Gamma + zz^T \\
&= \tilde{\Gamma}.
\end{aligned}$$

Now assume $-\phi$ is a learning function. Then, by concavity,

$$\begin{aligned}
\phi(\tilde{\Gamma}) &= \phi(\pi_{\frac{1}{2}}) \\
&\geq \frac{1}{2}\phi(\Gamma) + \frac{1}{2}\phi(\Gamma) \\
&= \phi(\Gamma).
\end{aligned}$$

In general, we can write any $\tilde{\Gamma} \geq \Gamma$ as $\tilde{\Gamma} = \Gamma + \sum_{i=1}^{m} z^{(i)} z^{(i)T}$, for a sequence of vectors $\{z^{(i)}\}$, $i = 1, \ldots, m$, and the result follows by induction from the last result. □

Most criteria used in classical optimum design theory (in the linear regression setting) when applied to covariance matrices are Loewner increasing. If, in addition, we can claim concavity, then by Theorem 7, the negative of any such function is a learning function. We have seen in Section 3 that $-\text{trace}(\Gamma)$ is a learning function, while $-\log\det(\Gamma)$ corresponding to D-optimality is another example.

For the normal distribution, we can show that for two normal density functions, π_1 and π_2, with covariance Γ_1 and Γ_2, respectively, we have that for any Shannon-type learning function $I_g(\theta_1) \leq I_g(\theta_2)$ if and only if $\det(\Gamma_1) \geq \det(\Gamma_2)$. We should note that in many Bayesian set-ups, such as regression and Gaussian process prediction, we have a joint multivariate distribution between x and θ. Suppose that, with obvious notation, the joint covariance matrix is:

$$\Gamma_{\theta,X} = \begin{pmatrix} \Gamma_\theta & \gamma_{\theta,X} \\ \gamma_{\theta,X}^T & \Gamma_X \end{pmatrix}.$$

Then, the posterior distribution for θ has covariance $\Gamma_\theta - \gamma_{\theta,X}\Gamma_X^{-1}\gamma_{\theta,X}^T \leq \Gamma_\theta$. Thus, for any Loewner increasing ϕ, it holds that $-\phi(\pi(\theta)) \leq -\mathrm{E}_X(\phi(\pi(\theta|X)))$, by Theorem 7. However, as the conditional covariance matrix does not depend on X, we have learning in the strong sense; $-\phi(\pi(\theta)) \leq -\phi(\pi(\theta|X))$. Classifying learning functions for θ and $\Gamma_{\theta,X}$ in the case where they are both unknown is not yet fully developed.

7. Approximate Bayesian Computation Designs

We now present a general method for performing optimum experimental design calculations, which, combined with the theory of learning outlined above, may provide a comprehensive approach. Recall that in our general setting, a decision about experimentation or observation is essentially a choice of the sampling distribution. In the statistical theory of the design of experiments, this choice typically means a choice of observation sites indexed by a control or independent variable z.

Indeed, we will have examples below in this category. However, the general formulation is that we want to maximize ψ over some restricted set of sampling distributions $f(x|\theta) \in \mathcal{F}$. A choice of f we call generalized design. Below, we will have one non-standard example based on selective sampling. Note that we shall always assume that the prior distribution $\pi(\theta)$ is fixed, which is independent of the choice of f. Then, recalling our general information functional as $\phi(\pi)$, the design optimization problem is (for fixed π):

$$\max_{f \in \mathcal{F}} \psi(f) = \mathrm{E}_{X_D}\phi(\pi(\theta|X_D)), \tag{11}$$

where we stress the dependence of the random variable X on the design and, thereby, on the sampling distribution f, by adding the subscript D.

If the set of sampling distributions f is specified by the control variable z, that is the choice of the sampling distribution $f(x|\theta, z)$ amounts to selecting $z \in \mathcal{Z}$, then the maximization problem is:

$$\max_{z \in \mathcal{Z}} \psi(f) = \mathrm{E}_{X_D}\phi(\pi(\theta|X_D, z)).$$

In the examples that we consider below, the sampling distribution will be indexed by a control variable z.

An important distinction should be made between what we shall here call linear and non-linear criteria. By a more general utility problem being linear, we mean that there is a utility function $U(\theta, x)$, such that, when we seek to minimize, again over choice f,

$$\mathrm{E}_{X_D}\mathrm{E}_{\theta|X_D}U(X_D,\theta) = \mathrm{E}_{X_D,\theta}U(X_D,\theta),$$

where the last expectation is with respect to the joint distribution of X_D and θ. In terms of integration, this only requires a single double integral. The non-linear case requires the evaluation of an "internal" integral for $\mathrm{E}_{\theta|X_D}U(X_D,\theta)$ and an external integral for E_{X_D}. It is important to note that Shannon-type functionals are special types of linear functionals where $U(\theta, X_D) = g(\pi(\theta|X_D))$. The distance-based functionals are non-linear in that they require a repeated single integral.

This distinction is important when other costs or utilities are included in addition to those coming from learning. Most obvious is a cost for the experiment. This could be fixed, so that no preposterior analysis is required, or it might be random in that it depends on the actual observation. For example one might add an additional utility $U(X_D)$ solely dependent of the outcome of the experiment: if it really does snow, then snow plows may need to be deployed. The overall (preposterior) expected value of the experiment might be:

$$\mathrm{E}_{X_D}\mathrm{E}_{\theta|X_D}U(X_D,\theta) + \mathrm{E}_{X_D}U(X_D).$$

In this way, one can study the exploration-exploitation problem, often referred to in search and optimization.

We now give a procedure to compute ψ for a particular choice of sampling distribution $f \in \mathcal{F}$. We assume that $f(x|\theta)$ and $\pi(\theta)$ are known. If the functional ϕ is non-linear, we have to obtain the posterior distribution $\pi(\theta|X_D)$ before evaluating ϕ. For simplicity, we use ABC rejection sampling (see Marjoram *et al.* [28]) to obtain an approximate sample from $\pi(\theta|X_D)$ that allows us to estimate the functional $\phi(\pi(\theta|X_D))$. In many cases, it is hard to find an analytical solution for $\pi(\theta|X_D)$, especially if $f(x|\theta)$ is intractable. These are the cases where ABC methods are most useful. Furthermore, ABC rejection sampling has the advantage that it is easily possible to re-compute $\hat{\phi}(\pi(\theta|X_D))$ for different values of X_D, which is an important feature, because we have to integrate over the marginal distribution of X_D in order to obtain $\psi(f) = \mathrm{E}_{X_D}\phi(\pi(\theta|X_D))$.

For a given $f \in \mathcal{F}$, we find the estimate $\hat{\psi}$ by integrating over $\hat{\phi}(\pi(\theta|X_D))$ with respect to the marginal distribution f_X. We can achieve this using Monte Carlo integration:

$$\psi(f) \approx \hat{\psi} = \frac{1}{G}\sum_{i=1}^{G}\hat{\phi}(\pi(\theta|x_D^{(i)}))$$

for $x_D^{(i)} \sim f_X$. The ABC procedure to obtain the estimate $\hat{\phi}(\pi(\theta|x_D))$ given x_D is as follows.

(1) Sample from $\pi(\theta)$: $\{\theta_1, \ldots, \theta_H\}$.
(2) For each θ_i, sample from $f(x|\theta_i)$ to obtain a sample: $x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$. This gives a sample from the joint distribution: $f_{X,\theta}$.
(3) For each θ_i, compute a vector of summary statistics: $T(x^{(i)}) = (T_1(x^{(i)}), \ldots, T_m(x^{(i)}))$.
(4) Split T-space into disjoint neighborhoods $\mathcal{N}_k$.

(5) Find the neighborhood $\mathcal{N}_k$ for which $T(x_D) \in \mathcal{N}_k$ and collect the θ_i for which $T(x^{(i)}) \in \mathcal{N}_k$, forming an approximate posterior distribution $\tilde{\pi}(\theta|T)$, which if T is approximately sufficient, should be close to $\pi(\theta|x_D)$. If T is sufficient, we have that $\tilde{\pi}(\theta|T) \to \pi(\theta|x_D)$ as $|\mathcal{N}_k| \to 0$.
(6) Approximate $\pi(\theta|x_D)$ by $\tilde{\pi}(\theta|T)$.
(7) Evaluate $\phi(\pi(\theta|x_D))$ by integration (internal integration).

Steps 1–4 need to be conducted only once at the outset for each $f \in \mathcal{F}$; only Steps 5–7 have to be repeated for each $x_D \sim f_X$.

For the linear functional, explained above, we do not even need to compute the posterior distribution, $\pi(\theta|x_D)$, if we are happy to use the naive approximation to the double integral:

$$\psi(f) \approx \frac{1}{G}\sum_{i=1}^{G} U(x_i, \theta_i),$$

where $\{x_i, \theta_i\}_{i=1}^{N}$ are independent draws from the joint distribution $f(x, \theta) = f(x|\theta)\pi(\theta)$.

The optimum $\psi(f)$ for $f \in \mathcal{F}$ may be found by employing any suitable optimization method. In this paper, we intend to focus on the computation of $\hat{\psi}(f)$. Therefore, in the illustrative examples below, we take a "crude" optimization approach, that is we estimate $\psi(f)$ for a fixed set of possible choices for f and compare the estimates.

The basic technique of ABCD was introduced in Hainy *et al.* [29], but here, we present it fully embedded into statistical learning theory. Note that related different procedures utilizing MCMC chains were independently developed in Drovandi and Pettitt [30] and Hainy *et al.* [31].

We now present two examples that are meant to illustrate the applicability of ABCD to very general design problems using non-linear design criteria. Although these examples are rather simple and may also be solved by analytical or numerical methods, their generalizations become intractable using traditional methods.

7.1. Selective Sampling

When the background sampling distribution is $f(x|\theta)$, we may impose prior constraints of which data we accept to use. Such models in greater generality may occur when observation is cheap, but the use of observation is expensive, for example computationally. We can call this "selective sampling", and we present a simple example.

Suppose in a one-dimensional problem that we are only allowed to accept observations from two slits of equal width at z_1 and z_2. Here, the model is equivalent (in the limit as the slit widths become small) to replacing $f(x|\theta)$ by the discrete distribution:

$$f(x = i|\theta, z_1, z_2) = \frac{f(z_i|\theta)}{f(z_1|\theta) + f(z_2|\theta)}, \; i = 1, 2.$$

If we have a prior distribution $\pi(\theta)$ and $f(x|z_1, z_2) = \int f(x|\theta, z_1, z_2)\pi(\theta)d\theta$ denotes the marginal distribution of x, the posterior distribution is given by:

$$\pi(\theta|x = i, z_1, z_2) = \frac{f(x = i|\theta, z_1, z_2)\pi(\theta)}{f(x = i|z_1, z_2)}, \; i = 1, 2.$$

To simplify even further, we take as a criterion:

$$\phi(\pi(\theta|x, z_1, z_2)) = \max_{\theta} \pi(\theta|x, z_1, z_2).$$

The maximum is a limiting version of Tsallis entropy and is a learning functional.

Now consider a special case:

$$\begin{aligned} z|\theta &\sim \mathcal{N}(\theta, 1), \\ \theta &\sim U[-1, 1]. \end{aligned}$$

The preposterior:

$$\psi(z_1, z_2) = \sum_{i=1}^{2} \phi(\pi(\theta|x = i, z_1, z_2)) f(x = i|z_1, z_2)$$

can be calculated explicitly. If $z_2 \geq z_1$ and $z_i \in [-a, a]$, then:

$$\begin{aligned} \max_{z_1, z_2} \psi(z_1, z_2) &= \psi(-a, a) \\ &= \frac{1}{1 + \exp(-2a)} \\ &= \begin{cases} \frac{1}{2} & a \to 0 \\ 1 & a \to \infty \end{cases}. \end{aligned}$$

Next, we show how this example can be solved using ABCD. Due to the special structure of the sampling distribution in this example, we modified our ABC sampling strategy slightly.

(1) For fixed z_1 and z_2, sample H numbers $\{\theta^{(j)}, j = 1, \ldots, H\}$ from the prior.

(2) For each $\theta^{(j)}$, repeat:

   (a) sample $z^{(k)} \sim \pi(z|\theta^{(j)})$ until $\#\left\{z^{(k)} \in \{N_\epsilon(z_1), N_\epsilon(z_2)\}\right\} = K_z$, where $N_\epsilon(z) = [z - \epsilon/2, z + \epsilon/2]$;

   (b) drop all $z^{(k)} \notin \{N_\epsilon(z_1), N_\epsilon(z_2)\}$;

   (c) sample $x^{(j)}$ from discrete distribution with probabilities $\Pr(x^{(j)} = i) = \dfrac{\#\{z^{(k)} \in N_\epsilon(z_i)\}}{K_z}$, $i = 1, 2$.

(3) For $i = 1, 2$, select all $\theta^{(j)}$ for which $x^{(j)} = i$, compute kernel density estimate for these $\theta^{(j)}$ and obtain maximum $\to \hat{\phi}(\hat{\pi}(\theta|x = i, z_1, z_2))$.

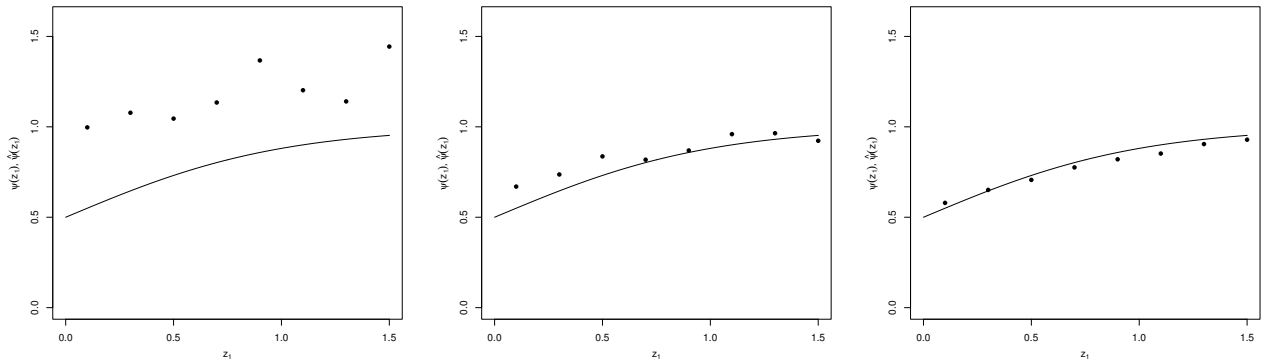
(4) $\hat{\psi}(z_1, z_2) = \sum_{i=1}^{2} \hat{\phi}(\hat{\pi}(\theta|x = i, z_1, z_2)) \dfrac{\#\{x^{(j)} = i\}}{H}$.

We performed our ABC sampling strategy for this example for a range of parameters for the slit neighborhood length ϵ ($\epsilon = 0.005, 0.01, 0.05$), H (H = 100, 1,000, 10,000) and K_z (K_z = 50, 100, 200) in order to assess the effect of these parameters on the accuracy of the ABC estimates of the criterion ψ. The most notable effect was found for the ABC sample size H.

Figure 2 shows the estimated values of the criterion, $\hat{\psi}$, for the special case where $z_2 = -z_1$ when $a = 1.5$. We set $\epsilon = 0.01$, $K_z = 100$. The ABC sample size H is set to $H = 100$

(left), $H = 1,000$ (center), and $H = 10,000$ (right). The criterion was evaluated at the eight points ($z_1 = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$). The theoretical criterion function $\psi(z_1)$ is plotted as a solid line.

Figure 2. Estimated values of the criterion $\hat{\psi}(z_1)$ (points) and theoretical criterion function $\psi(z_1)$ (solid line) for $\epsilon = 0.01$, $K_z = 100$, and $H = 100$ (**a**), $H = 1,000$ (**b**), $H = 10,000$ (**c**).



(**a**) (**b**) (**c**)

7.2. Spatial Sampling for Prediction

This example is also a simple version of an important paradigm, namely optimal sampling of a spatial stochastic process for good prediction. Here, the stochastic process labeled X is indexed by a space variable z, and we write $X_i = X_i(z_i),\ i = 1, \ldots, n$ to indicate sampling at sites (the design) $D_n = \{z_1, \ldots, z_n\}$. We would typically take the design space, $\mathcal{Z}$, to be a compact region.

We wish to compute the predictive distribution at a new site z_{n+1}, namely $x_{n+1}(z_{n+1})$, given $x_D = x(D_n) = (x_1(z_1), \ldots, x_n(z_n))$. In the Gaussian case, the background parameter θ could be related to a fixed effect (drift) or the covariance function of the process, or both. In the analysis, x_{n+1} is regarded as an additional parameter, and we need its (marginal) conditional distribution.

The criterion of interest is the maximum variance of the (posterior) predictive distribution over the design space:

$$\begin{aligned} -\phi(x(D_n)) &= \max_{z_{n+1} \in \mathcal{Z}} \text{var}(X_{n+1}(z_{n+1}) | x(D_n)) \\ &= \max_{z_{n+1} \in \mathcal{Z}} \int (x_{n+1} - \mu_{x_{n+1}})^2 \pi(x_{n+1} | x(D_n), z_{n+1}) dx_{n+1}. \end{aligned}$$

This functional is learnable, since it is is a maximum of a set of variances, each one of which is learnable.

Referring back to how the general design optimization problem that was stated in (11), the posterior predictive distribution of x_{n+1} may be interpreted as the posterior distribution in (11). The optimality criterion ψ is found by integrating ϕ with respect to $X_1, \ldots, X_n$.

The strategy is to select a design D_n and then perform ABC at each test point z_{n+1}. The learning functional $\phi(x_D)$ is estimated by generating the sample $I = \{x_D^{(j)}, x_{n+1}^{(j)}\}_{j=1}^H = \{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}, x_{n+1}^{(j)}\}_{j=1}^H$ at the sites $\{z_1, z_2, \ldots, z_n, z_{n+1}\}$ and calculating:

$$-\hat{\phi}(x_D) = \max_{z_{n+1} \in \mathcal{Z}} \frac{1}{|J_\epsilon(x_D)|} \sum_{j \in J_\epsilon(x_D)} (x_{n+1}^{(j)} - \bar{x}_{n+1})^2,$$

where $J_\epsilon(x_D) = \{j \in \{1,\ldots,H\} : x_D^{(j)} \in N_\epsilon(x_D)\}$, we have $x_D^{(j)} \in N_\epsilon(x_D)$ if $|x_i^{(j)} - x_i| \leq \epsilon \; \forall \; i = 1,\ldots,n$, and $\bar{x}_{n+1} = (1/|J_\epsilon(x_D)|) \sum_{j \in J_\epsilon(x_D)} x_{n+1}^{(j)}$.

In order to estimate $\psi(D_n) = \mathrm{E}_{X_D}(\phi(X_D))$, we obtain a sample $O = \{x_D^{(i)}\}_{i=1}^G$ from the marginal distribution of the random field at the design D_n and perform Monte Carlo integration:

$$\hat{\psi}(D_n) = \frac{1}{G} \sum_{i=1}^{G} \hat{\phi}(x_D^{(i)}). \tag{12}$$

For each $x_D^{(i)} \in O$ from the marginal sample, we use the sample I to compute $\hat{\phi}(x_D^{(i)})$ in order to save computing time. We then vary the design using some optimization algorithm.

A simple example is adopted from Müller *et al.* [32]. The observations $(x_1(z_1), x_2(z_2), x_3(z_3), x_4(z_4))$ are assumed to be distributed according to a one-dimensional Gaussian random field with mean zero, a marginal variance of one and $z_i \in [0,1]$. We want to select an optimal design $D_3 = (z_1, z_2, z_3)$, such that:

$$-\psi(D_3) = \mathrm{E}_{X_{1:3}(D_3)} \left[\max_{z_4 \in [0,1]} \mathrm{var}\left(X_4(z_4) | X_{1:3}(D_3)\right) \right]$$

is minimal.

We assume the Ornstein–Uhlenbeck process with correlation function $\rho(|s-t|;\, \theta) = e^{-\theta|s-t|}$. Two prior distributions for the parameter θ are considered. The first one is a point prior at $\theta = \log(100)$, so that $\rho(h) = \rho(h;\, \log(100)) = 0.01^h$. This is the correlation function used by Müller *et al.* [32] in their study of empirical kriging optimal designs. The second prior distribution is an exponential prior for θ with scale parameter $\lambda = 10$ (*i.e.*, $\theta \sim \mathrm{Exp}(10)$). The scale parameter λ was chosen, such that the average correlation functions of the point and exponential priors are similar. By that, we mean that the average of the mean correlation function for the exponential prior over all pairs of sites s and t, $\mathrm{E}_{s,t}[\mathrm{E}_\theta\{\rho(|s-t|;\, \theta)|\theta \sim \mathrm{Exp}(\lambda)\}] = \mathrm{E}_{s,t}[1/(1+\lambda|s-t|)]$, matches the average of the fixed correlation function $\rho(|s-t|;\, \log(100)) = 0.01^{|s-t|}$ over all pairs of sites s and t, $\mathrm{E}_{s,t}[0.01^{|s-t|}]$. The sites are assumed to be uniformly distributed over the coordinate space.

To be more specific, first, for each site $s \in \mathcal{X}$, the average correlation to all other sites $t \in \mathcal{X}$ is computed. Then, these average correlations are averaged over all sites $s \in \mathcal{X}$. For the point prior, the average correlation is $\mathrm{E}_{s,t}[\rho(|s-t|;\, \log(100))] = \frac{2}{\log(100)^2}(\log(100) - (1 - \frac{1}{100})) = 0.3409$, and for the exponential prior, the value is $\mathrm{E}_{s,t}[\mathrm{E}_\theta\{\rho(|s-t|;\, \theta)|\theta \sim \mathrm{Exp}(\lambda)\}] = \frac{2}{\lambda^2}[(1+\lambda)\log(1+\lambda) - \lambda]$. If $\lambda = 10$, we have $\mathrm{E}_{s,t}[\mathrm{E}_\theta\{\rho(|s-t|;\, \theta)|\theta \sim \mathrm{Exp}(10)\}] = 0.3275$.

Figure 3 depicts the distributions of the correlation function $\rho(h;\, \theta) = \exp(-\theta h)$ under the two prior distributions. The solid line corresponds to the fixed correlation function $\rho(h;\, \theta = \log(100)) = 0.01^h$. The dotted line and the two dashed lines represent the mean correlation function and the 0.025- and 0.975-quantile functions for $\rho(h;\, \theta)$ under the prior $\theta \sim \mathrm{Exp}(10)$.

We estimated the criterion on a grid with spacing 0.05 for the design points z_1 and z_3 (z_2 is fixed at $z_2 = 0.5$). We set $G = 1,000$, $H = 5 \cdot 10^6$ and $\epsilon = 0.01$ for each design point. The sample $\{x^{(j)}(z) :\; z \in \mathcal{Z}\}_{j=1}^H$ is simulated at all points z of the grid prior to the actual ABC algorithm. In order to accelerate the computations, it is then reused for all possible designs D_3 to estimate each $\hat{\phi}(x_D^{(i)})$, $i = 1,\ldots,G$, in (12). The sample size $H = 5 \cdot 10^6$ was deemed to provide a sufficiently exhaustive sample from the four-dimensional normal vector $(x_1(z_1), x_2(z_2), x_3(z_3), x_4(z_4))$ for any $z_i \in \mathcal{Z}$, so that

the distortive effect of using the same sample for the computations of all $\hat{\phi}(x_D^{(i)})$ is only of negligible concern for our purposes of ranking the designs.

Figure 3. Prior distributions of correlation function $\rho(h;\ \theta)$: correlation function $\rho(h) = 0.01^h$ under point prior $\theta = \log(100)$ (solid line); mean correlation function (dotted line) and 0.025- and 0.975-quantile functions (dashed lines) for $\rho(h;\ \theta)$ under the prior $\theta \sim \text{Exp}(10)$.

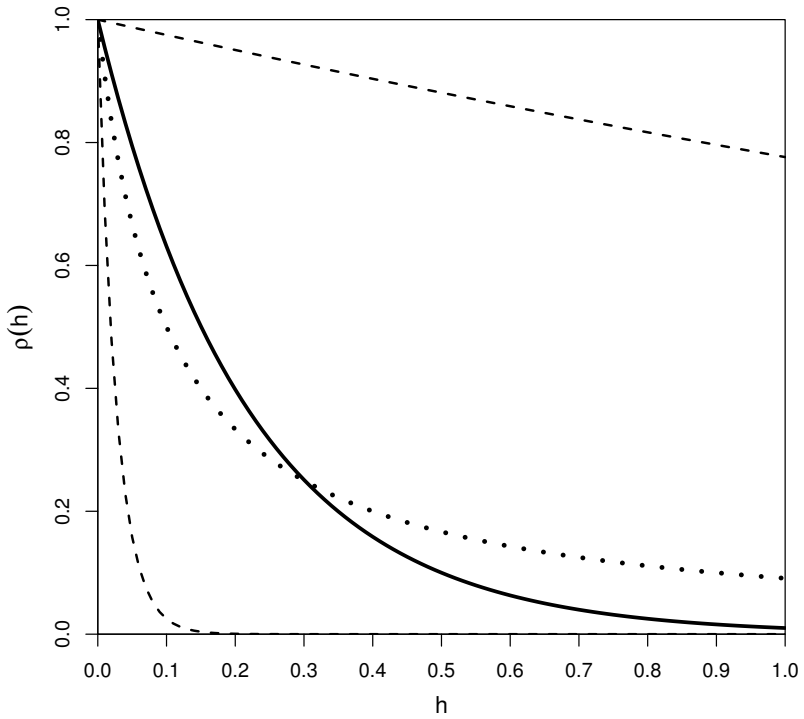
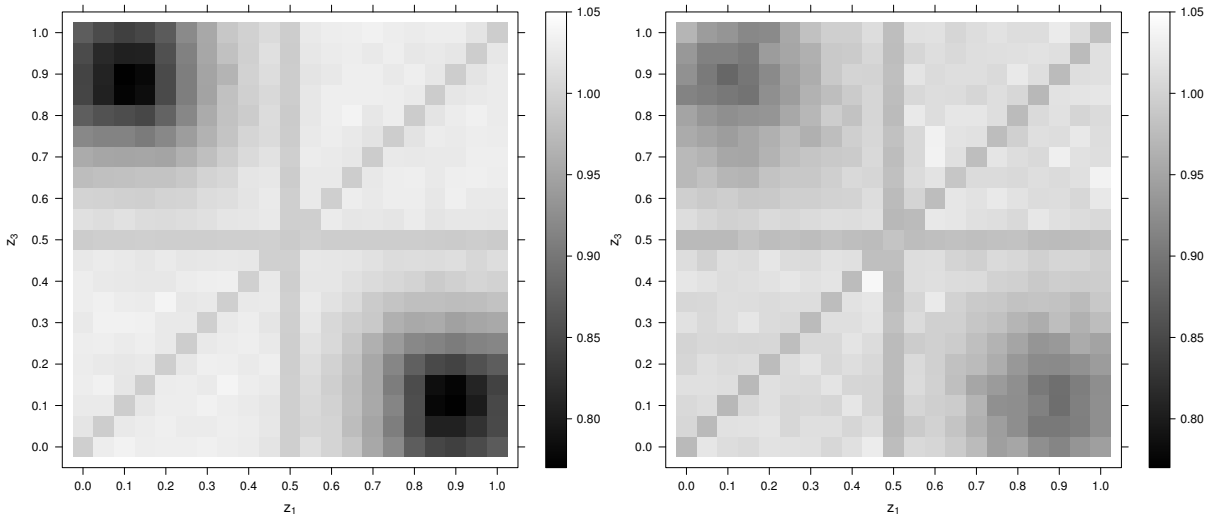


Figure 4 (left) shows the map of estimated criterion values, $-\hat{\psi}(D_3)$, when the prior distribution of θ is the point prior at $\theta = \log(100)$. It can be seen that the minimum of the criterion is attained at about $(z_1, z_3) = (0.9, 0.1)$ or $(z_1, z_3) = (0.1, 0.9)$, which is comparable to the the results obtained in Müller *et al.* [32] for empirical kriging optimal designs. Note that the diverging criterion values at the diagonal and at $z_1 = 0.5$ and $z_3 = 0.5$ are attributable to a specific feature of the ABC method used. At these designs, the actual dimension of the design is lower than three, so for a given ϵ, there are more elements in the neighborhood than for the other designs with three distinctive design points. Hence, a much larger fraction of the total sample, $\{x_{n+1}^{(j)}\}_{j=1}^{H}$, is included in the ABC sample, $\{x_{n+1}^{(j)} : j \in J_\epsilon(y_D)\}$. Therefore, the values of the criterion get closer to the marginal variance of one. In order to avoid this effect, the parameter ϵ would have to be adapted in these cases. Alternatively, one could use other variants of ABC rejection, where the fixed number of N elements of $I = \{x_D^{(j)}, x_{n+1}^{(j)}\}_{j=1}^{H}$ with the smallest distance to the draw $x_D^{(i)} \in O$ are constituting the ABC posterior sample, making it necessary to compute and sort out the distances for each $x_D^{(i)} \in O$.

Figure 4 (right) gives the estimated criterion values, $-\hat{\psi}(D_3)$, when the prior of θ is $\theta \sim \text{Exp}(10)$. Due to the uncertainty of the prior parameter θ, the optimal design points for z_1 and z_3 slightly move to the edges, which is also in accordance with the findings of Müller *et al.* [32].

Figure 4. Spatial prediction criterion map for the point prior at $\theta = \log(100)$ (**left**) and for the exponential prior $\theta \sim \text{Exp}(10)$ (**right**).



8. Conclusions

There are some fundamental results in Bayesian learning which provide important background to fields like the optimal design of experiments. Functionals of prior distributions which are learnable, via observation, in a wide sense, are convex. Shannon information is an example but there are many others and the paper points to some wide classes with connections to other fields. It combines the theory of learning with an effective method for the optimal design of experiments based on simulation: ABCD. It is suggested that the method should prove useful in non-standard situations, such as non-linear, non-Gaussian models and for complex problems where the sampling distribution is intractable but one can still draw samples from it, for given parameter values. A simple message is that the learning theory and simulation method applies to a generalized notion of an experiment as a choice of sampling distribution, under restrictions.

Acknowledgments

The research of the first author has been partially supported by the French Science Fund (ANR) and Austrian Science Fund (FWF) bilateral grant I-833-N18. The last author is grateful for the award of Exzellenzstipendium des Landes Oberösterreich by the governor of Upper-Austria, in 2012.

Author Contributions

The background sections were mainly authored by Henry P. Wynn; ABCD was jointly conceived by Werner G. Müller and Henry P. Wynn; All computations for the examples were performed by Markus Hainy. All authors have read and approved the final published manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Blackwell, D. Comparison of Experiments. In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 31 July–12 August 1950; University of California Press: Berkeley, CA, USA, 1951; pp. 93–102.
2. Torgersen, E. *Comparison of Statistical Experiments*; Encyclopedia of Mathematics and its Applications 36; Cambridge University Press: Cambridge, UK, 1991.
3. Rényi, A. On Measures of Entropy and Information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
4. Lindley, D.V. On a Measure of the Information Provided by an Experiment. *Ann. Math. Stat.* **1956**, *27*, 986–1005.
5. Goel, P.K.; DeGroot, M.H. Comparison of Experiments and Information Measures. *Ann. Math. Stat.* **1979**, *7*, 1066–1077.
6. Ginebra, J. On the measure of the information in a statistical experiment. *Bayesian Anal.* **2007**, *2*, 167–211.
7. Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Stat. Sci.* **1995**, *10*, 273–304.
8. Sebastiani, P.; Wynn, H.P. Maximum entropy sampling and optimal Bayesian experimental design. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **2000**, *62*, 145–157.
9. Chater, N. The Probability Heuristics Model of Syllogistic Reasoning. *Cogn. Psychol.* **1999**, *38*, 191–258.
10. Schoenberg, I.J. Metric Spaces and Positive Definite Functions. *Trans. Am. Math. Soc.* **1938**, *44*, 522–536.
11. Schilling, R.L.; Song, R.; Vondracek, Z. *Bernstein Functions: Theory and Applications*; De Gruyter Studies in Mathematics 37; De Gruyter: Berlin, Germany, 2012.
12. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
13. DeGroot, M.H. *Optimal Statistical Decisions*, WCL edition; Wiley-Interscience: Hoboken, NJ, USA, 2004.
14. Goldman, A.I.; Shaked, M. Results on inquiry and truth possession. *Stat. Probab. Lett.* **1991**, *12*, 415–420.
15. Fallis, D.; Liddell, G. Further results on inquiry and truth possession. *Stat. Probab. Lett.* **2002**, *60*, 169–182.
16. Torgerson, W.S. *Theory and Methods of Scaling*; John Wiley and Sons, Inc.: New York, NY, USA, 1958.
17. Gower, J.C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **1966**, *53*, 325–338.
18. Gower, J.C. Euclidean distance geometry. *Math. Sci.* **1982**, *7*, 1–14.
19. Itti, L.; Baldi, P. Bayesian surprise attracts human attention. *Vis. Res.* **2009**, *49*, 1295–1306.

20. Haykin, S.; Chen, Z. The Cocktail Party Problem. *Neural Comput.* **2005**, *17*, 1875–1902.
21. Berger, J. The case for objective Bayesian analysis. *Bayesian Anal.* **2006**, *1*, 385–402.
22. Marshall, A.W.; Olkin, I.; Arnold, B.C. *Inequalities: Theory of Majorization and Its Applications*, 2nd ed; Springer Series in Statistics; Springer: Berlin, Germany, 2009.
23. Hardy, G.H.; Littlewood, J.E.; Pólya, G. *Inequalities*, 2nd ed.; Cambridge Mathematical Library; Cambridge University Press: Cambridge, UK, 1988.
24. Müller, A.; Stoyan, D. *Comparison Methods for Stochastic Models and Risks*, 1st ed. Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2002.
25. Ryff, J.V. Orbits of l^1-functions under doubly stochastic transformations. *Trans. Am. Math. Soc.* **1965**, *117*, 92–100.
26. DeGroot, M.H.; Fienberg, S. Comparing probability forecasters: Basic binary concepts and multivariate extensions. In *Bayesian Inference and Decision Techniques*; Goel, P., Zellner, A., Eds.; North-Holland: Amsterdam, The Netherlands, 1986; pp. 247–264.
27. Dawid, A.P.; Sebastiani, P. Coherent dispersion criteria for optimal experimental design. *Ann. Stat.* **1999**, *27*, 65–81.
28. Marjoram, P.; Molitor, J.; Plagnol, V.; Tavaré, S. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 15324–15328.
29. Hainy, M.; Müller, W.; Wynn, H. Approximate Bayesian Computation Design (ABCD), an Introduction. In *mODa 10—Advances in Model-Oriented Design and Analysis*; Ucinski, D., Atkinson, A.C., Patan, M., Eds.; Contributions to Statistics, Springer International Publishing: Heidelberg/Berlin, Germany, 2013; pp. 135–143.
30. Drovandi, C.C.; Pettitt, A.N. Bayesian Experimental Design for Models with Intractable Likelihoods. *Biom* **2013**, *69*, 937–948.
31. Hainy, M.; Müller, W.G.; Wagner, H. *Likelihood-free Simulation-based Optimal Design*; Technical Report; Johannes Kepler University: Linz, Austria, 2013.
32. Müller, W.G.; Pronzato, L.; Waldl, H. Beyond space-filling: An illustrative case. *Procedia Environ. Sci.* **2011**, *7*, 14–19.