entropy

MDPI

Article

Bayesian Input Design for Linear Dynamical Model Discrimination

Piotr Bania

Department of Automatic Control and Robotics, AGH University of Science and Technology, Al. A. Mickiewicza 30, 30-059 Krakow, Poland; pba@agh.edu.pl; Tel.: +48-12-617-28-34

Received: 16 January 2019; Accepted: 27 March 2019; Published: 30 March 2019

Abstract: A Bayesian design of the input signal for linear dynamical model discrimination has been proposed. The discrimination task is formulated as an estimation problem, where the estimated parameter indexes particular models. As the mutual information between the parameter and model output is difficult to calculate, its lower bound has been used as a utility function. The lower bound is then maximized under the signal energy constraint. Selection between two models and the small energy limit are analyzed first. The solution of these tasks is given by the eigenvector of a certain Hermitian matrix. Next, the large energy limit is discussed. It is proved that almost all (in the sense of the Lebesgue measure) high energy signals generate the maximum available information, provided that the impulse responses of the models are different. The first illustrative example shows that the optimal signal can significantly reduce error probability, compared to the commonly-used step or square signals. In the second example, Bayesian design is compared with classical average D-optimal design. It is shown that the Bayesian design is superior to D-optimal design, at least in this example. Some extensions of the method beyond linear and Gaussian models are briefly discussed.

Keywords: bayesian experimental design; model discrimination; information; entropy

1. Introduction

Discrimination of various dynamical models of the same process has a wide area of applications, especially in multiple-model fault detection and isolation [1–4] and, in many other estimation and control problems [5–7], it is necessary to choose the most likely dynamical model from a finite set. The discrimination task can be formulated as an estimation problem, where the estimated parameter θ indexes particular models. The problem can also be considered as a finite-dimensional approximation of more general identification tasks [8]. As the error probability or variance of the estimator of θ usually depends on the input signal, it is important to select a signal that minimizes error probability or maximizes a utility function that encodes the purpose of the experiment. Selection of an input signal that maximizes a utility function is strongly related to optimal experimental design [9].

Experimental design methods can be divided into classical and Bayesian. The classical methods, also called optimal experimental design, typically use various functionals of the Fisher information matrix as a utility function. These methods are widely described in the literature and work well if the model is linear in its parameters (see [8,10–12] and the review article [9]). Unfortunately, in typical identification tasks, the solution of model equation and the covariance depends non-linearly on θ, even if the model equation is linear. This implies that the information matrix and the utility function depend on the parameter θ to be estimated. Therefore, only locally-optimal design can be obtained [13]. To obtain more robust methods, an averaging over the prior parameter distribution or minimax design [14], [8] (Section 6.1) are commonly used, but these methods are not fully Bayesian.

Bayesian optimal design uses the utility function, a functional of the posterior distribution (see [15,16] and the review articles [13,17,18]). The most commonly used utility functions are mutual

information between parameters and model output, Kullback-Leibler divergence between the prior and posterior distributions, and the determinant of the posterior covariance matrix [13,16]. In contrast to classical methods, in Bayesian design the utility function does not depend on the parameters to be estimated. Hence, the method can cope with non-linear problems. The utility function, which is suitable for model discrimination, is the error probability of the MAP estimator of θ [19]. Such a utility function is generally difficult to calculate (see [19]), but the result of Feder & Merhav [20] implies that the error probability of the MAP estimator is upper-bounded by some decreasing function of mutual information between θ and the output of the system. Hence, the maximization of mutual information creates the possibility of reducing the error probability, provided that appropriate estimator is used. However, the most serious problem that inhibits the development of this idea is great computational complexity in calculating the mutual information.

The main contribution of this article is a fully-Bayesian (in the terminology of [13]) method for finding an input signal that maximizes the mutual information between θ and the system output. Maximization of information or, equivalently, maximization of the output entropy has been proposed by many authors (see, e.g., [13,15,17,18,21–23]), but the mutual information is very hard to compute and the problem is often intractable. To overcome this serious difficulty, instead of mutual information the lower bound, given by Kolchinsky & Tracey [24], has been used. This is a pairwise-distance based entropy estimator and it it useful here, since it is differentiable, tight, and asymptotically reaches the maximum possible information (see [24] (Sections 3.2, 4, 6)). Maximization of such a lower bound, under the signal energy (i.e., the square of the signal norm) constraints, is much simpler, gives satisfactory solutions, and allows for practical implementation of the idea of maximizing information. This is illustrated with examples. Moreover, it is shown that, for certain cases, this problem reduces to a solution of a certain eigenproblem.

The article is organized as follows. In Section 2, the estimation task is formulated and the upper bound of the error probability and the lower bound of the mutual information are given. In Section 2.1, a selection between two models is discussed and an exact solution is given. Design of input signals with small energy, which is required in some applications, is described in Section 2.2. In Section 2.3, the large energy limit is discussed. An application to linear dynamical systems with unknown parameters is given in Section 3. An example of finding the most likely model among three stochastic models with different structures is given in Section 4. Comparison with classical D-optimal design is performed in Section 5. The article ends with conclusions and references.

2. Maximization of Mutual Information between the System Output and Parameter

Let us consider a family of linear models

$$Y = F_\theta U + Z, \tag{1}$$

where $\theta \in 1, 2, ..., r$, $Y, Z \in R^{n_Y}$, and $U \in R^{n_U}$. The matrices F_θ are bounded. The parameter θ is unknown. The prior distribution of θ is given by

$$P(\theta = i) = p_{0,i}, i = 1, ..., r. \tag{2}$$

The random variable Z is conditionally normal (i.e., $p(Z|\theta) = N(Z, 0, S_\theta)$), where the covariance matrices S_θ are given a priori and $S_\theta > 0$, for all θ. The variable U is called the input signal. In all formulas below, the input signal U is a deterministic variable. The set of admissible signals is given by

$$\mathbf{S}_\varrho = \{U \in R^{n_U}; U^T U \leqslant \varrho\}. \tag{3}$$

Under these assumptions, and after applying Bayes rule:

$$p(Y|U) = \sum_{\theta=1}^{r} p_{0,\theta} N(Y, F_\theta U, S_\theta), \tag{4}$$

$$p(Y|\theta, U) = N(Y, F_\theta U, S_\theta), \tag{5}$$

$$p(\theta|Y, U) = \frac{p_{0,\theta} N(Y, F_\theta U, S_\theta)}{\sum_{j=1}^{r} p_{0,j} N(Y, F_j U, S_j)}. \tag{6}$$

The entropies of Y and θ and the conditional entropies are defined as

$$H(\theta) = -\sum_{\theta=1}^{r} p_{0,\theta} \ln p_{0,\theta}, \tag{7}$$

$$H(\theta|Y, U) = -\int p(Y|U) \left(\sum_{\theta=1}^{r} p(\theta|Y, U) \ln p(\theta|Y, U) \right) dY, \tag{8}$$

$$H(Y|U) = -\int p(Y|U) \ln p(Y|U) dY, \tag{9}$$

$$H(Y|\theta) = \tfrac{1}{2} \sum_{\theta=1}^{r} p_{0,\theta} \ln \left((2\pi e)^{n_y} |S_\theta| \right). \tag{10}$$

The mutual information between θ and Y is defined as (see [25] (pp. 19, 250))

$$I(Y; \theta|U) = H(\theta|U) - H(\theta|Y, U) = H(Y|U) - H(Y|\theta, U).$$

As $H(\theta|U) = H(\theta)$ and $H(Y|\theta, U) = H(Y|\theta)$, then $I(Y; \theta|U)$ is given by

$$I(Y; \theta|U) = H(\theta) - H(\theta|Y, U) = H(Y|U) - H(Y|\theta). \tag{11}$$

The MAP estimator of θ is defined as

$$\hat{\theta}(Y, U) = \arg \max_{\theta \in \{1,\ldots,r\}} p(\theta|Y, U).$$

The error probability of $\hat{\theta}$ is given by (see [20])

$$P_e(U) = 1 - \int \left(\max_{\theta \in \{1,\ldots,r\}} p(\theta|Y, U) \right) p(Y|U) dY.$$

It follows from Fano's inequality ([25] (p. 38)), that P_e is lower bounded by an increasing function in $H(\theta|Y, U)$. Feder & Merhav [20] proved that $2P_e(U) \leqslant H(\theta|Y, U) \log_2 e$. As $H(\theta|Y, U) = H(\theta) - I(Y; \theta|U)$ and $H(\theta)$ does not depend on U, then the maximization of $I(Y; \theta|U)$ creates the possibility of reducing P_e, and the optimal signal is given by

$$U^*(\varrho) = \arg \max_{U \in \mathbf{S}_\varrho} I(Y; \theta|U). \tag{12}$$

To overcome the problems associated with the calculation of $I(Y; \theta|U)$, we will use its lower bound.

Lemma 1. *(Information bounds). For all* $U \in R^{n_U}$,

$$I_l(U) \leqslant I(Y; \theta|U) \leqslant H(\theta), \tag{13}$$

where

$$I_l(U) = -\sum_{i=1}^{r} p_{0,i} \ln \left(\sum_{j=1}^{r} p_{0,j} e^{-D_{i,j}(U)} \right), \tag{14}$$

$$D_{i,j}(U) = \tfrac{1}{4} U^T Q_{i,j} U + \tfrac{1}{2} \ln |\tfrac{1}{2}(S_i + S_j)| - \tfrac{1}{4} \ln |S_i||S_j|, \text{ and} \tag{15}$$

$$Q_{i,j} = (F_i - F_j)^T (S_i + S_j)^{-1} (F_i - F_j). \tag{16}$$

Proof. According to (4), $p(Y|U)$ is finite Gaussian mixture. For such mixtures, the information bounds are known. A detailed proof, based on Chernoff α-divergence, is given in [24] (Section 4). □

Lemma 2. *Let $\hat{\theta}(Y, U) = \arg\max_{\theta \in \{1,\dots,r\}} p(\theta|Y, U)$ be the MAP estimator of θ, and let $P_e(U)$ denote its error probability. There exists a continuous, increasing, and concave function $f : [0, H(\theta)] \to [0, 1 - r^{-1}]$, such that*

$$P_e(U) \leqslant f(H(\theta) - I_l(U)) \leqslant \tfrac{1}{2}(H(\theta) - I_l(U)) \log_2 e. \tag{17}$$

Proof. Feder & Merhav [20] (see Theorem 1 and Equation (14)) proved that there exists an increasing, continuous, and convex function $\phi : [0, 1 - r^{-1}] \to [0, H(\theta) \log_2 e]$, such that

$$2P_e(U) \leqslant \phi(P_e(U)) \leqslant H(\theta|Y, U) \log_2 e. \tag{18}$$

As $H(\theta|Y, U) = H(\theta) - I(Y; \theta|U)$ and $I_l(U) \leqslant I(Y; \theta|U)$, then $\phi(P_e(U)) \leqslant (H(\theta) - I_l(U)) \log_2 e$. The function $g = \phi^{-1}$ is increasing, continuous, concave, and it follows from (18) that $2g(\eta) \leqslant \eta$. Hence, $P_e(U) \leqslant \phi^{-1}((H(\theta) - I_l(U)) \log_2 e) = g((H(\theta) - I_l(U)) \log_2 e) \leqslant \frac{1}{2}(H(\theta) - I_l(U)) \log_2 e$. Taking $f(\eta) = g(\eta \log_2 e)$ we obtain the result. □

Now, the approximate solution of (12) is given by

$$U^*(\varrho) = \arg \max_{U \in \mathbf{S}_\varrho} I_l(U). \tag{19}$$

As I_l is smooth and $\mathbf{S}_\varrho$ is compact, (19) is well-defined.

2.1. Selection between Two Models

Suppose that θ takes only two values, 1 and 2, with prior probabilities $p_{0,1}$ and $p_{0,2} = 1 - p_{0,1}$, respectively. It's easy to check, by direct calculation, that

$$e^{-I_l(U)} = (p_{0,1} + p_{0,2} e^{-D_{1,2}(U)})^{p_{0,1}} (p_{0,1} e^{-D_{1,2}(U)} + p_{0,2})^{p_{0,2}}. \tag{20}$$

Equation (20) implies that the maximization of I_l is equivalent to the maximization of $D_{1,2}$. On the basis of (15), we have the following optimization: task

$$\max_{U^T U \leqslant \varrho} U^T Q_{1,2} U. \tag{21}$$

The solution of (21) is the eigenvector of $Q_{1,2}$ corresponding to its largest eigenvalue; that is,

$$Q_{1,2} U^* = \lambda_{max}(Q_{1,2}) U^*, ||U^*||^2 = \varrho. \tag{22}$$

2.2. Small Energy Limit

In many practical applications, the energy of an excitation signal must be small. The second order Taylor expansion of (14) gives

$$I_l(U) = \tfrac{1}{4}U^T Q U - \sum_{i=1}^{r} p_{0,i} \ln \sum_{j=1}^{r} \alpha_{i,j} + o(||U||^2), \tag{23}$$

where

$$Q = \sum_{i=1}^{r} p_{0,i} \left(\frac{\sum\limits_{j=1}^{r} \alpha_{i,j} Q_{i,j}}{\sum\limits_{j=1}^{r} \alpha_{i,j}} \right), \text{ and} \tag{24}$$

$$\alpha_{i,j} = p_{0,j} e^{-D_{i,j}(0)}. \tag{25}$$

If the value of ϱ (see (3)) is small, then the last term in (23) can be omitted. As the second term does not depend on U, we have the following optimization task:

$$\max_{U^T U \leqslant \varrho} U^T Q U. \tag{26}$$

The solution of (26) is the eigenvector of Q corresponding to its largest eigenvalue.

2.3. Large Energy Limit

We will investigate asymptotic behaviour of $I(Y;\theta|U)$ when $||U|| \to \infty$. On the basis of Lemma 1, the condition

$$\min_{i \neq j} U^T Q_{i,j} U > 0 \tag{27}$$

guarantees that $\lim_{\varrho\to\infty} I(Y;\theta|\varrho U) = H(\theta)$. It is also possible that $\lim_{\varrho\to\infty} I(Y;\theta|\varrho U) < H(\theta)$, for some U. Such signals are weakly informative and they cannot generate the maximum information, even if their amplitude tends to infinity. Let $\mathbf{S}_1$ denote the unit ball in R^{n_U} and let μ be the Lebesgue measure on $\mathbf{S}_1$. The set of weakly informative signals is defined as

$$\Omega = \{U \in \mathbf{S}_1 : \lim_{\varrho\to\infty} I(Y;\theta|\varrho U) < H(\theta)\}. \tag{28}$$

Theorem 1. *$\mu(\Omega) = 0$ if and only if $F_i \neq F_j$, for all $i \neq j$.*

Proof. $\Leftarrow$: On the basis of Lemma 1 and (28), $\Omega = \cup_{i\neq j}\Omega_{i,j}$, where $\Omega_{i,j} = \{\xi \in \mathbf{S}_1 : \xi^T Q_{i,j} \xi = 0\}$. Since $S_i + S_j$ is positive-definite and $F_i \neq F_j$ then, on the basis of (16), the matrix $Q_{i,j}$ has at least one positive eigenvalue. Hence, $\mu(\Omega_{i,j}) = 0$ and $\mu(\Omega) = \sum\limits_{i\neq j} \mu(\Omega_{i,j}) = 0$. $\Rightarrow$: The condition $\mu(\Omega) = 0$ implies that $\mu(\Omega_{i,j}) = 0$. Hence, $Q_{i,j}$ has at least one positive eigenvalue, which is possible only if $F_i \neq F_j$. □

As a conclusion, we have the following result.

Theorem 2. *Let $\hat{\theta}(Y,U) = \arg\max_{\theta\in\{1,\ldots,r\}} p(\theta|Y,U)$, be the MAP estimator of θ and let $P_e(U)$ denote its error probability. If $F_i \neq F_j$ for all $i \neq j$, then, for any $\epsilon > 0$, there exists a number $\varrho > 0$ and a signal $U \in \mathbf{S}_\varrho$, such that $P_e(U) < \epsilon$.*

Proof. By the assumption, and from Theorem 1, the set $\mathbf{S}_1 \setminus \Omega$ is non-empty. If $U \in \mathbf{S}_1 \setminus \Omega$, then $\min_{i\neq j} U^T Q_{i,j} U > 0$ and, from Lemma 1, we get $\lim_{\varrho\to\infty} I_l(\varrho U) = \lim_{\varrho\to\infty} I(Y;\theta|\varrho U) = H(\theta)$. Now, Lemma 2 implies that $2P_e(U) \leqslant (H(\theta) - I_l(\varrho U)) \log_2 e < \epsilon$, for sufficiently large ϱ. □

3. Application to Linear Dynamical Systems

Consider, now, the family of linear systems

$$x_{k+1} = A_\theta x_k + B_\theta u_k + G_\theta w_k, k = 0, 1, 2, ..., N-1, \tag{29}$$

$$y_k = C_\theta x_k + D_\theta v_k, k = 1, 2, ..., N, \tag{30}$$

where the prior distribution of θ is given by (2) and $x_k \in R^n, y_k \in R^m, w_k \in R^{n_w}, v_k \in R^m$, $w_k \sim N(0, I_{n_w})$, and $v_k \sim N(0, I_m)$. The variables $w_0, ..., w_{N-1}, v_1, ..., v_N$ are mutually independent. The initial condition is zero. The solution of (29) with initial condition $x_0 = 0$ has the form

$$x_k = \sum_{i=0}^{k-1} A_\theta^{k-i-1} B_\theta u_i + \sum_{i=0}^{k-1} A_\theta^{k-i-1} G_\theta w_i. \tag{31}$$

If we denote $X = \text{col}(x_1, ..., x_N)$, $Y = \text{col}(y_1, ..., y_N)$, $U = \text{col}(u_0, ..., u_{N-1})$, $W = \text{col}(w_0, ..., w_{N-1})$, and $V = \text{col}(v_1, ..., v_N)$, then (31) and (30) can be rewritten as

$$X = \mathcal{B}_\theta U + \mathcal{G}_\theta W, \text{ and} \tag{32}$$

$$Y = \mathcal{C}_\theta X + \mathcal{D}_\theta V, \tag{33}$$

where the matrices $\mathcal{B}_\theta$, $\mathcal{G}_\theta$, $\mathcal{C}_\theta$, and $\mathcal{D}_\theta$ follow forms (30) and (31). The variables W and V are independent, where $W \sim N(0, I_{Nn_w})$ and $V \sim N(0, I_{Nm})$. Substituting (32) into (33) we get Equation (1), where $F_\theta = \mathcal{C}_\theta \mathcal{B}_\theta$, $Z = \mathcal{C}_\theta \mathcal{G}_\theta W + \mathcal{D}_\theta V$. The conditional density of Z has the form $p(Z|\theta) = N(Z, 0, S_\theta)$, where the covariance matrix is given by

$$S_\theta = \mathcal{D}_\theta \mathcal{D}_\theta^T + \mathcal{C}_\theta \mathcal{G}_\theta \mathcal{G}_\theta^T \mathcal{C}_\theta^T. \tag{34}$$

Hence, the results of Section 2 can be applied to the dynamical system (29) and (30).

4. Example

In some fault detection and isolation problems [3,4], there is a need to determine which of the known models of the process is the most adequate. It is, therefore, important to find a signal that emphasizes the differences between these various models. As an example of this type of problem, let us consider three stochastic continuous-time models:

$$dx = (A_\theta x + B_\theta u)dt + G_\theta dw, \tag{35}$$

where $\theta \in \{1, 2, 3\}$, $x(t) \in R^\theta$, $x(0) = 0$, $u(t), w(t) \in R$, w is a standard Wiener process, and

$$A_1 = -1, B_1 = 1, G_1 = 0.05, \tag{36}$$

$$A_2 = \begin{bmatrix} 0 & 1 \\ -3 & -2.5 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, G_2 = \begin{bmatrix} 0 \\ 0.05 \end{bmatrix}, \tag{37}$$

$$A_3 = \begin{bmatrix} 0 & 1 & 0 \\ -3 & -3.5 & 1 \\ 0 & 0 & -10 \end{bmatrix}, B_3 = \begin{bmatrix} 0 \\ 0 \\ 30 \end{bmatrix}, G_3 = \begin{bmatrix} 0 \\ 0 \\ 0.05 \end{bmatrix}. \tag{38}$$

The step responses of these models are similar and they are difficult to experimentally distinguish from each other if the noise level is significant. The observation equation has the form

$$y_k = x_1(t_k) + 0.05 v_k, k = 1, 2, ..., N, \tag{39}$$

where $v_k \sim N(0,1)$, $t_k = kT_0$, $T_0 = 0.1$, and x_1 is the first component of $x(t)$. If $x_k = x(t_k)$ and $u(t) = u_k, t \in [t_{k-1}, t_k)$, then, after discretization, the state x_k and the output y_k are described by (29) and (30), with appropriate matrices A_θ, B_θ, C_θ, G_θ, and D_θ. The matrices F_θ and S_θ are calculated by using (31)–(34). Let us observe that, although the orders of the systems are different, the size of both F_θ and S_θ is always $N \times N$. We are interested in the maximization of $I_l(U)$. The solutions of (19) and (26) with a uniform prior, $\varrho = N$, and $N = 200$ steps, are shown in the upper part of Figure 1. The step responses and the optimal responses are shown in the bottom part of Figure 1.

Let us observe that, in contrast to the step signal, the optimal signal clearly distinguishes the systems—although the energy of all input signals was the same and equal to N.

Let $U_{st}, U_{sq} \in \partial \mathbf{S}_1$ denote the normalized step and square signal with period of three, respectively, and let $U^*(\varrho)$ denote the optimal signal. To check the validity of the results, the error probabilities $P_e(\varrho U_{st})$, $P_e(\varrho U_{sq})$, and $P_e(U^*(\varrho))$ were estimated by Monte Carlo simulation with 10^6 trials and $N = 50$ steps. The results are shown in Figure 2. It was observed that the optimal signal gives an error probability several thousand times smaller than the step or square signal with the same energy. The second observation is that $P_e(\varrho U_{sq})$ initially increased with ϱ. To explain this, let us note that Inequality (17) does not guarantee that P_e is decreasing function of ϱ. Hence, it is possible that P_e increases in certain directions, although Theorem 2 guarantees that P_e tends to zero, provided that signal norm tends to infinity.

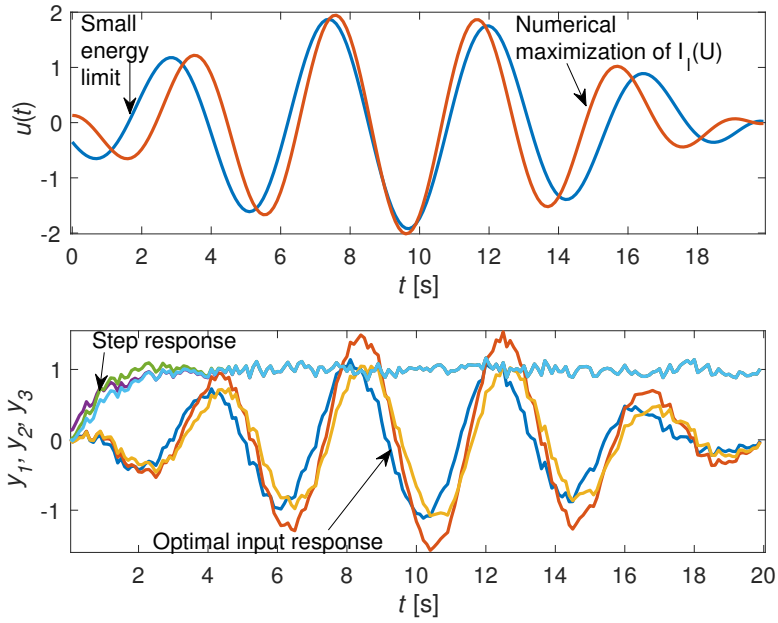

Figure 1. (**Top**) Numerical solution of (19) and the small energy approximation (26), for $\varrho = N = 200$. (**Bottom**) Step responses and optimal responses of all systems.

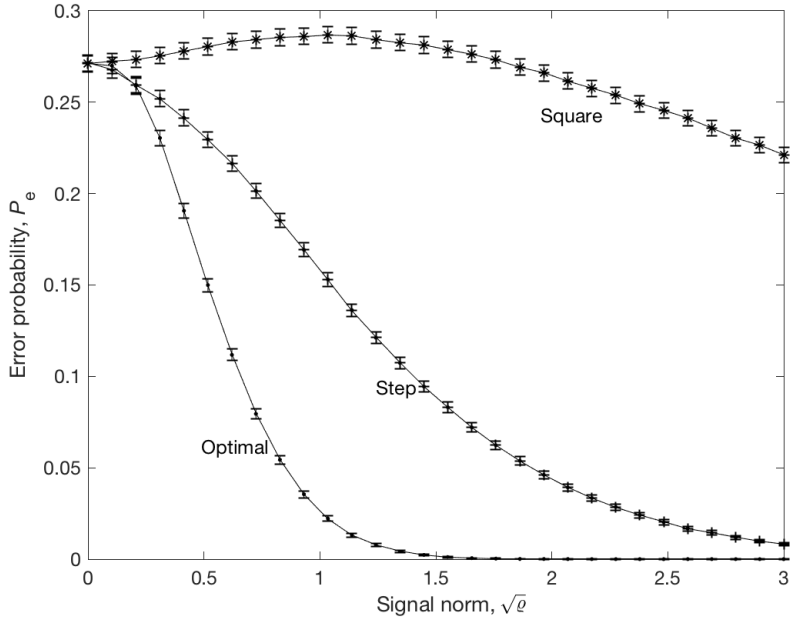

Figure 2. Error probability of the MAP estimator for the optimal signal (.), step signal (+), and square (*) signal with period of three. The number of steps is $N = 50$. The error probability has been estimated by a Monte Carlo method with 10^6 trials. Standard error bars were multiplied by factor of 10 for better visibility.

5. Comparison with the Average D-Optimal Design

Classical methods of signal design for parameter identification use various functionals of the Fisher information matrix as a utility function. One of the most popular is D-optimal design, which consists of finding a signal that maximizes the determinant of the information matrix (see [10–12] and the review article [9]). These methods are well-suited to models that are linear in their parameters. Unfortunately, in typical identification and discrimination tasks, the output is a non-linear function of the parameters and the information matrix depends on unknown parameters to be identified. One of the possibilities for avoiding this problem is the averaging of the utility function over the prior parameter distribution. This method is called average D-optimal design (see [14,26] and [9] (Sections 5.3.5 and 6), for details). The Bayesian design, described in the previous sections, will be compared with the average D-optimal design. To that end, let us consider a finite family of linear models (see also [12] (pp. 91–93))

$$y_k = \frac{b_\theta z^{-1}}{1 - a_\theta z^{-1}} u_k + \sigma_v v_k, \tag{40}$$

where $\theta \in \{1, 2, 3, 4\}$, $a_\theta = 0.6 + 0.1(\theta - 1)$, $b_\theta = 1 - a_\theta$, $\sigma_v = 0.1$, and $v_k \sim N(0, 1)$. The prior distribution of θ is uniform (i.e., $p_{0,\theta} = 0.25$). The state space representation of (40) has the form

$$x_{k+1} = a_\theta x_k + b_\theta u_k, \tag{41}$$

$$y_k = x_k + \sigma_v v_k, \tag{42}$$

which is consistent with (29) and (30). The Fisher information matrix is given by

$$M_F(\theta, U) = \frac{1}{N\sigma_v^2} \sum_{k=1}^{N} d_k d_k^T, \tag{43}$$

where $d_k = (\xi_k, \eta_k)^T$ and $\xi_k = \frac{\partial y_k}{\partial a_\theta}$, $\eta_k = \frac{\partial y_k}{\partial b_\theta}$, denote the sensitivity of the output y_k to changes in parameters a and b, respectively. The derivatives ξ_k and η_k fulfil the sensitivity equations

$$\xi_k = 2a_\theta \xi_{k-1} - a_\theta^2 \xi_{k-2} + b_\theta u_{k-2}, \tag{44}$$

$$\eta_k = a_\theta \eta_{k-1} + u_{k-1}, k = 1, 2, ..., N, \tag{45}$$

with zero initial conditions. The average D-optimal design consists in finding a signal U that maximizes the expectation of the determinant of the information matrix (see [9] (Sections 5.3.5 and 6), [14], [11] (Chapter 6), and [12] for details). Hence, the utility function to be maximized has the form

$$J(U) = \sum_{\theta=1}^{4} p_{0,\theta} |M_F(\theta, U)|, \tag{46}$$

with the energy constraints given by (3). Maximization of the utility function (46) has been performed for various signal energies and the error probability of the MAP estimator was estimated by Monte Carlo with 10^5 trials. The same procedure was repeated using Bayesian design for (41) and (42). The results are shown in Figure 3. The error rate of Bayesian method is significantly smaller when compared to D-optimal design, at least in this example. In particular, the signal shown in the upper-right part of Figure 3 gives an error probability approximately three times smaller than that of D-optimal signal, although the energy of both signals was the same.

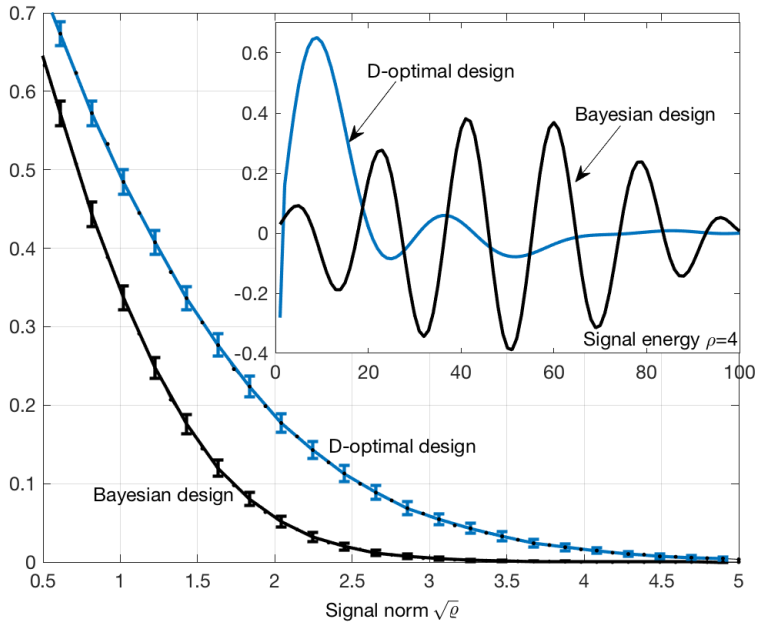

Figure 3. Error probability of theMAP estimator (see Section 2), as a function of signal norm and the exemplary input signals (top right) generated by D-optimal and Bayesian methods. The error probability was calculated by a Monte Carlo method with 10^5 trials. Standard error bars were multiplied by a factor of 10 for better visibility.

6. Possible Extensions of the Results

In this section, we will briefly discuss some possible extensions of the results to an infinite set of parameters and beyond linear and Gaussian models.

6.1. Non-Linear Models

Although the article refers to linear models, it is possible to extend the results to non-linear models of the form

$$Y = F_\theta(U) + Z, \tag{47}$$

where the conditional density of variable Z is given by

$$p(Z|\theta, U) = N(Z, 0, S_\theta(U)) \tag{48}$$

and $S_\theta(U) > 0$, for all $U \in U_{ad}$, $\theta \in \{1, ..., r\}$. Under these assumptions, the density of Y still remains a Gaussian mixture and the information lower bound takes the form

$$I_l(U) = -\sum_{i=1}^{r} p_{0,i} \ln \left(\sum_{j=1}^{r} p_{0,j} e^{-D_{i,j}(U)} \right), \tag{49}$$

where

$$\begin{aligned} D_{i,j}(U) &= \tfrac{1}{4}(F_i(U) - F_j(U))^T (S_i(U) + S_j(U))^{-1} (F_i(U) - F_j(U)) + \\ &+ \tfrac{1}{2} \ln |\tfrac{1}{2}(S_i(U) + S_j(U))| - \tfrac{1}{4} \ln |S_i(U)||S_j(U)|. \end{aligned} \tag{50}$$

6.2. Non-Gaussian Models

If $p(Z|\theta, U)$ is non-Gaussian distribution, then it is possible, on the basis of Equation (10) in [24], to construct an information lower bound of the form

$$I_l(U) = -\sum_{i=1}^{r} p_{0,i} \ln \left(\sum_{j=1}^{r} p_{0,j} e^{-C_\alpha(p_i||p_j)} \right), \tag{51}$$

where

$$C_\alpha(p_i||p_j) = -\ln \int p(Z|i, U)^\alpha p(Z|j, U)^{1-\alpha} dZ \tag{52}$$

is the Chernoff α-divergence and $\alpha \in [0, 1]$. Unfortunately, calculation of (52) is difficult if n_Y is large.

6.3. Infinite Set of Parameters

Let us consider following model:

$$Y = F(\theta)U + Z, \tag{53}$$

where $p(Z|\theta) = N(Z, 0, S(\theta))$ and $\theta \in R^p$. If we assume that prior density of θ is Gaussian; that is,

$$p_0(\theta) = N(\theta, m_\theta, S_\theta), S_\theta > 0, \tag{54}$$

then

$$p(Y) = \int p_0(\theta) N(Y, F(\theta)U, S(\theta)), \tag{55}$$

and $p(Y)$ can be approximated by a finite Gaussian mixture

$$p(Y|U) = \int p_0(\theta) N(Y, F(\theta)U, S(\theta)) d\theta \approx \sum_{j=1}^{N_a} p_{0,j} N(Y, F(\theta_j)U, S(\theta_j)), \tag{56}$$

where $p_{0,j} \geqslant 0$ and $\sum_{j=1}^{N_a} p_{0,j} = 1$. It's possible to calculate weights and nodes in (56) by using multidimensional quadrature. If n_θ is large, then an appropriate sparse grid should be used. To illustrate the method, we will show only a simple, second-order quadrature with $2n_\theta$ points.

Lemma 3. *The approximate value of the integral* $J(f) = \int N(\theta, m_\theta, S_\theta) f(\theta) d\theta$ *is given by*

$$J(f) \approx \frac{1}{2n_\theta} \sum_{j=1}^{2n_\theta} f(\theta_j), \tag{57}$$

where

$$\theta_{2i-1} = m_\theta - S_\theta^{0.5} e_i, \theta_{2i} = m_\theta + S_\theta^{0.5} e_i, i = 1, ..., n_\theta \tag{58}$$

and e_i *is* i^{th} *basis vector. If* $f(\theta) = \frac{1}{2}\theta^T A\theta + b^T\theta + c$, *then equality holds in (57).*

Proof. Direct calculation. □

Application of Lemma 3 to (56) gives $p_{0,j} = p_0 = (2n_\theta)^{-1}$, $N_a = 2n_\theta$. Now, since (56) is a Gaussian mixture, the results of Section 2 can be utilized and the information lower bound takes the form

$$I_l(U) = -p_0 \sum_{i=1}^{r} \ln \left(\sum_{j=1}^{r} e^{-D_{i,j}(U)} \right) - \ln p_0, \tag{59}$$

where $D_{i,j}$ and θ_j are given by (15), (16), and (58), respectively, and $F_j = F(\theta_j)$, $S_j = S(\theta_j)$. The approximate solution of (12) can be found by maximization of (59) with the constraints (3).

7. Discussion and Conclusions

An effective Bayesian design method for linear dynamical model discrimination has been developed. The discrimination task is formulated as an estimation problem, where the estimated parameter θ indexes particular models. To overcome the computational complexity, instead of $I(Y; \theta|U)$, its lower bound ($I_l(U)$), proposed by Kolchinsky & Tracey [24], was used as a utility function. This bound is especially useful, as it is differentiable, tight, and reaches the maximum available information $H(\theta)$. It has been proved, on the basis of the results of Feder & Merhav [20], that the error probability of the MAP estimator (see Lemma 2) is upper bounded by $\frac{1}{2}(H(\theta) - I_l(U)) \log_2 e$. The maximization of $I_l(U)$ has been considered under the signal energy constraint, but other kinds of constraints can also be easily implemented. It was shown that the maximization of $I_l(U)$, in the case of two parameters, is equivalent to maximization of a quadratic form on the sphere (see also [3]). Next, the small energy limit was analyzed. It was proved that the solution is given by an eigenvector corresponding to maximal eigenvalue of some Hermitian matrix. This result can serve as starting point for numerical maximization of $I_l(U)$. If the energy of the signal tends to infinity, then almost all (in the sense of the Lebesgue measure) signals generate maximum information, provided that the impulse responses of the models are pairwise different. Under these conditions, it was proved that P_e of the MAP estimator tends to zero.

An example of discrimination of three stochastic models with different structures was given. It is easy to observe, from Figure 1, that, in contrast to the step signal, the optimal signal clearly distinguished the systems, although the energy of both signals was the same. The P_e of the MAP estimator was calculated by Monte Carlo simulation. It was observed that the square signal gave an error probability several thousand times greater than the optimal signal with the same energy. Hence, we conclude that the error probability and the accuracy of MAP estimator depends very strongly on the excitation signal. Although Theorem 2 implies that $\lim_{\varrho \to \infty} P_e(\varrho U) = 0$ for almost all U, there exist signals such that $P_e(\varrho U)$ is locally increasing. This is the case of a high-frequency square signal, as illustrated in Figure 2.

It was shown, in Section 5 (see Figure 3), that P_e of the MAP estimator corresponding to Bayesian design was a few times smaller than P_e generated by D-optimal design, at least in the analyzed example. This result suggests that Bayesian design can be applied to non-linear problems and that it is superior to classical D-optimal design.

Some extensions of the results to the infinite set of parameters and beyond linear and Gaussian assumptions were briefly discussed in Section 6. Extension to non-linear models seems to be easy, but the non-Gaussian case is difficult and a deeper analysis is required. The case of an infinite set of parameters was discussed in Section 6.3. It was shown that the measurement density can be approximated by a finite Gaussian mixture, after which the results of Section 2 could be directly applied. The general conclusion of this analysis is that the information bounds can be easily constructed, as long as the measurement density is a mixture of Gaussian distributions.

An analytical gradient formula must be provided for the effective numerical maximization of $I_l(U)$. The matrix inversions and the determinants in (15) and (16) should be calculated by SVD. To reduce the computational complexity and required memory resources, the symmetries appearing in (15) and (16), and the fact that $D_{i,i} = 0$, should be utilized. The determinants in (15) and the matrices F_i, S_i can be calculated off-line, but the matrices $Q_{i,j}$ may require too much memory if N and r are large. Therefore, $Q_{i,j}$ was calculated on-line.

Applications of the presented methods in dual control problems [5] are expected as part of our future work. An additional area in applications is the issue of automated testing of dynamical systems.

Funding: This research was financed from the statutory subsidy of the AGH University of Science and Technology, No. 16.16.120.773.

Acknowledgments: I would like to thank Jerzy Baranowski for discussions and comments.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

$\xi \sim N(m, S)$ means that ξ has normal distribution with mean m and covariance S. The density of a normally-distributed variable is denoted by $N(x, m, S) = 2\pi^{-\frac{n}{2}}|S|^{-\frac{1}{2}} \exp(-0.5(x-m)^T S^{-1}(x-m))$. The symbol $\mathrm{col}(a_1, a_2, ..., a_n)$ denotes a column vector. The set of symmetric, positive-definite matrices of dimension n is denoted by $\mathbf{S}^+(n)$.

References

1. Bania, P.; Baranowski, J. Bayesian estimator of a faulty state: Logarithmic odds approach. In Proceedings of the 22nd International Conference on Methods and Models in Automation and Robotics (MMAR), Miedzyzdroje, Poland, 28–31 August 2017; pp. 253–257.
2. Baranowski, J.; Bania, P.; Prasad, I.; Cong, T. Bayesian fault detection and isolation using Field Kalman Filter. *EURASIP J. Adv. Signal Process.* **2017**, *79*. [CrossRef]
3. Blackmore, L.; Williams, B. Finite Horizon Control Design for Optimal Model Discrimination. In Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 12–15 December 2005; doi:10.1109/CDC.2005.1582753. [CrossRef]
4. Pouliezos, A.; Stavrakakis, G. *Real Time Fault Monitoring of Industrial Processes*; Kluwer Academic: Boston, MA, USA, 1994.
5. Bania, P. Example for equivalence of dual and information based optimal control. *Int. J. Control* **2018**. [CrossRef]
6. Lorenz, S.; Diederichs, E.; Telgmann, R.; Schütte, C. Discrimination of dynamical system models for biological and chemical processes. *J. Comput. Chem.* **2007**, *28*, 1384–1399. [CrossRef] [PubMed]
7. Ucinski, D.; Bogacka, B. T-optimum designs for discrimination between two multiresponse dynamic models. *J. R. Stat. Soc. B* **2005**, *67*, 3–18. [CrossRef]
8. Walter, E.; Pronzato, L. Identification of Parametric Models from Experimental Data. In *Series: Communications and Control Engineering*; Springer: Berlin/Heidelberg, Germany, 1997.

9. Pronzato, L. Optimal experimental design and some related control problems. *Automatica* **2008**, *44*, 303–325. [CrossRef]
10. Atkinson, A.C.; Donev, A.N. *Optimum Experimental Design*; Oxford University Press: Oxford, UK, 1992.
11. Goodwin, G.C.; Payne, R.L. *Dynamic System Identification: Experiment Design and Data Analysis*; Academic Press: New York, NY, USA, 1977.
12. Payne, R.L. Optimal Experiment Design for Dynamic System Identification. Ph.D. Thesis. Department of Computing and Control, Imperial College of Science and Technology, University of London, London, UK, February 1974.
13. Ryan, E.G.; Drovandi, C.C.; McGree, J.M.; Pettitt, A.N. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *Int. Stat. Rev.* **2016**, *84*, 128–154. [CrossRef]
14. Fedorov, V.V. Convex design theory. *Math. Operationsforsch. Stat. Ser. Stat.* **1980**, *1*, 403–413.
15. Lindley, D.V. *Bayesian Statistics—A Review*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 1972.
16. Ryan, E.G.; Drovandi, C.C.; Pettitt, A.N. Fully Bayesian Experimental Design for Pharmacokinetic Studies. *Entropy* **2015**, *17*, 1063–1089. [CrossRef]
17. Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Stat. Sci.* **1995**, *10*, 273–304. [CrossRef]
18. DasGupta, A. *Review of Optimal Bayes Designs*; Technical Report; Purdue University: West Lafayette, IN, USA, 1995.
19. Routtenberg, T.; Tabrikian, J. A general class of lower bounds on the probability of error in multiple hypothesis testing. In Proceedings of the 25th IEEE Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel, 3–5 December 2008; pp. 750–754.
20. Feder, M.; Merhav, N. Relations between entropy and error probability. *IEEE Trans. Inf. Theory* **1994**, *40*, 259–266. [CrossRef]
21. Arimoto, S.; Kimura, H. Optimum input test signals for system identification—An information-theoretical approach. *Int. J. Syst. Sci.* **1971**, *1*, 279–290. [CrossRef]
22. Fujimoto, Y.; Sugie, T. Informative input design for Kernel-Based system identification. In Procedings of the 2016 IEEE 55th Conference on Decision and Control (CDC), Las Vegas, NV, USA, 12–14 December 2016; pp. 4636–4639.
23. Hatanaka, T.; Uosaki, K. Optimal Input Design for Discrimination of Linear Stochastic Models Based on Kullback-Leibler Discrimination Information Measure. In Proceedings of the 8th IFAC/IFOORS Symposium on Identification and System Parameter Estimation 1988, Beijing, China, 27–31 August 1988; pp. 571–575.
24. Kolchinsky, A.; Tracey, B.D. Estimating Mixture Entropy with Pairwise Distances. *Entropy* **2017**, *19*, 361. [CrossRef]
25. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006.
26. Atkinson, A.C.; Fedorov, V.V. Optimal design: Experiments for discriminating between several models. *Biometrika* **1975**, *62*, 289–303.