The University of Georgia

DEPARTMENT OF STATISTICS
Technical Report

TECHNICAL REPORT NUMBER 2007-6

MULTI-OBJECTIVE OPTIMAL EXPERIMENTAL DESIGNS FOR EVENT-RELATED FMRI STUDIES

By

Ming-Hung Kao
Department of Statistics
University of Georgia
Athens, Georgia 30602

Abhyuday Mandal
Department of Statistics
University of Georgia
Athens, Georgia 30602

Nicole Lazar
Department of Statistics
University of Georgia
Athens, Georgia 30602

John Stufken
Department of Statistics
University of Georgia
Athens, Georgia 30602

October, 2007

Multi-objective Optimal Experimental Designs for Event-Related fMRI Studies

By

Ming-Hung Kao, Abhyuday Mandal, Nicole Lazar and John Stufken

Department of Statistics

University of Georgia

Abstract

The nature of ER-fMRI studies impels the need for multi-objective designs that simultaneously accomplish statistical goals, circumvent psychological confounds, and fulfill customized requirements. Incorporating knowledge about ER-fMRI designs, we propose an efficient algorithm to search for optimal multi-objective designs. This algorithm significantly outperforms previous search algorithms in terms of achieved efficiency, computation time and convergence rate. In addition, our design criterion, which serves as the objective function of the search algorithm, allows consistent design comparisons. This consistency is crucial to the success of the search algorithms. Furthermore, the parametrization utilized in our underlying model allows parameters that are interpretable and faithfully reflects the fluctuation of the hemodynamic response function.

KEY WORDS: compound design criterion; design efficiency; normalization; counterbalancing; discretization interval; genetic algorithms; weighted mutation; double-loop algorithms.

1 Introduction

Event-related functional magnetic resonance imaging (ER-fMRI) is one of the leading technologies for studying human brain activity in response to mental stimuli (Josephs et al., 1997; Rosen et al., 1998; Dale, 1999; Bandettini and Cox, 2000). Before conducting an ER-fMRI experiment, a design sequence consisting of stimuli of one or more types interlaced

with rests is prepared. This sequence is presented to an experimental subject while the MR scanner scans his/her brain every few seconds. The blood oxygenation level dependent (BOLD) time series at each brain voxel is collected for the end purpose of statistical inference. The design issue here is to best allocate the stimuli so that statistical inference is precise and valid.

Two common statistical goals in ER-fMRI are to estimate the hemodynamic response function (HRF; the noise-free BOLD time series triggered by a single, brief stimulus), and to detect brain activations; see also Buxton et al. (2000) and Birn et al. (2002). Considering both goals in one experiment is not uncommon, but it requires a good multi-objective design that simultaneously achieves high efficiencies on both dimensions. In addition, statistical efficiency is not the only concern while planning fMRI design sequences. Psychology plays an important, even crucial, role. When a design sequence is patterned or easy to predict, psychological effects such as habituation or anticipation might arise to confound stimulus effects (Rosen et al., 1998; Dale, 1999). Good designs should avoid these psychological confounds while retaining high efficiency in making statistical inference. Moreover, customized requirements such as a required frequency for each stimulus type might also arise to further complicate the design problem. With all these inherent requirements and limitations, finding a good, multi-objective design is hard, but inevitable for ER-fMRI; this necessitates a well-defined multi-objective design criterion (or MO-criterion for short) for evaluating competing designs.

An additional difficulty for this design problem comes from the size and shape of the design space. Consisting of all possible ER-fMRI designs, the design space is enormous and irregular (Buračas and Boynton, 2002; Liu, 2004). Searching over this huge space for an optimal design is an arduous task. Therefore, an efficient search algorithm is as well crucial.

Wager and Nichols (2003), referred to as WN henceforward, put forward a framework for finding optimal designs for ER-fMRI. By formulating a weighted-sum MO-criterion, they transfer the multi-objective design problem into a single-objective one. This largely simplifies the problem; see also Miettinen (1999), Wong (1999), and Deb (2001). They also propose

a modified genetic algorithm (or WN's GA) to search for (near-)optimal multi-objective designs. This path-breaking framework is influential and WN's GA has been applied in many studies over the last few years (e.g., Callan et al., 2006; Ramautar et al., 2006; Summerfield et al., 2006; Wang et al., 2007).

Working in WN's pioneering framework, we develop an efficient algorithm which has several advantages over WN's GA. First, knowledge for the performance of the well-known ER-fMRI designs is incorporated to expedite the search. Second, the design criteria are defined to allow fair, consistent design comparisons. While crucial to the success of the search algorithms, WN's approach does not always achieve this. Third, our underlying model faithfully reflects the fluctuation in the HRF, and our HRF parameters are interpretable. By contrast, WN's model is less rigorous. Each parameter in their models may simultaneously represent more than one height of the HRF when the ISI (inter-stimulus interval, time interval between consecutive event onsets) is not a multiple of the TR (time to repetition or sampling rate); i.e., $ISI \neq mTR$ for any positive integer m. With these advantages, our GA outperforms WN's GA. In addition, the deigns found by our GA are superior to the designs that are currently in use by researchers.

Though taking less computation time than WN's approach, our algorithm achieves designs with significantly higher efficiencies in terms of both our and WN's criteria. As shown in the simulation, the convergence rate of our algorithm is also much faster than that of WN's. A dramatic improvement is evident. Furthermore, our designs yield an advantageous trade-off between estimation efficiency and detection power, compared to WN's designs and other well-known ER-fMRI designs including m-sequence-based designs, block designs, mixed designs, clustered m-sequences and permuted block designs. Moreover, our algorithm consistently yields designs that are in good agreement with the (approximated) optimal stimulus proportion, the proportion of each stimulus type, of Liu and Frank (2004). As mentioned in Liu and Frank (2004), optimizing the design efficiency hinges on this proportion.

We also propose tuning methods to the algorithmic parameter α introduced by WN. This parameter links to selective pressure and its value impacts the population diversity of

3

WN's GA. Based on their observation of the GA's performance, WN propose an effective value for α. While this value might work well for their approach, it is desirable to adapt the value of α to different GA environments. We thus borrow ideas from double-loop algorithms to allow self-adapting α; that is to adjust the value of α according to the performance of the past generations of our GA. However, the self-adaptive mechanisms complicate the search algorithm. Since the effectiveness of α in our algorithm is limited (as shown in our simulations), we suggest not to include this parameter for simplicity.

The rest of the article is organized as follows. We start with the important features of our approach in Section 2. Details of our proposed algorithm is then presented in Section 3. Simulations are provided in Section 4. Conclusions and a discussion are in Section 5.

2 Features of Our Algorithm

We aim at finding optimal multi-objective ER-fMRI designs. Here, an ER-fMRI design is an alignment of events, including stimuli and rests. For convenience, the symbols $0, 1, ..., Q$ are used to represent the events, where Q is the total number of stimulus types. Rests are indicated by 0s and a type-i stimulus is denoted by i, $i = 1, ..., Q$. A design, denoted by ξ, looks like $\xi = \{101201210...\}$.

While being presented to the subject, each event in the design sequence lasts for a short period of time relative to the ISI, the fixed time interval between event onsets. The "off period" fills up the remaining time before the onset of the next event. Rests (or 0s) are indistinguishable from the "off period". They are "pseudo-events" to facilitate our approach.

Our proposed approach mainly focus on searching for designs that efficiently achieve the four important objectives of ER-fMRI: 1) estimating the HRF, 2) detecting brain activation, 3) avoiding psychological confounds, and 4) maintaining the desired frequency for each stimulus type. Although these four dimensions are the focus of the current work, our approach is general enough to accommodate other objectives as well.

The important features of our approach are detailed in the rest of this section. A complete

accounting of this approach is deferred to Section 3.

2.1   Diverse Design Patterns

The first feature of our approach lies in the diverse design patterns utilized in our algorithm. The motivation comes from observing Figure 1, which shows a scatter plot of estimation efficiency versus detection power for 10,000 randomly generated designs. As defined in Section 3, these two "larger-the-better" design criteria evaluate the ability of a design in achieving the two statistical goals — estimation and detection. In Figure 1, the two outstanding designs are the block design, where events of the same type are clustered together, and the m-sequence (a pseudo-random sequence; e.g., Buračas and Boynton, 2002; Liu, 2004; Barker, 2004); they are generated systematically. In that Figure, none of the 10,000 random designs reaches as a high estimation efficiency as the m-sequence, nor do they attain high detection power. Furthermore, the desirable region in the upper right corner is empty altogether. Thus, the performance of an algorithm involving only random designs (such as WN's GA) may be hindered. Taking advantage of the systematic designs in broadening and expediting the search seems natural and promising (see also Liu, 2004, p. 412). We note that Figure 1 also implicitly indicate a well-known trade-off between estimation efficiency and detection power (Liu et al., 2001; Liu and Frank, 2004; Liu, 2004). While the m-sequence or random designs have high estimation efficiencies, their abilities in detecting activation are scarcely comparable to the block design. On the other hand, the block design has the highest detection power, yet this superiority comes form a sacrifice on its capability for estimating the HRF. This trade-off can be readily seen in Figure 6.

Figure 2 presents the design patterns included in our algorithm. In that figure, different shades in each design represent events of different types; white means rest. The area of each shade is proportional to the number of events involved. The designs considered are m-sequences or m-sequence-based designs, random designs, block designs, and their concatenations, where junctions could occur anywhere in the design. Good properties are observed
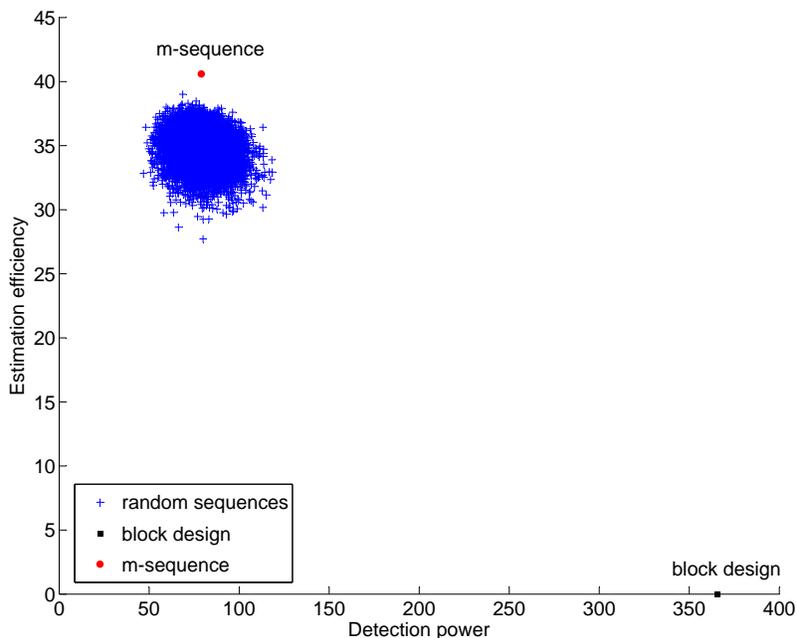
5

Figure 1: Estimation efficiency vs. detection power for 10,000 randomly generated designs, an m-sequence and a block design. Estimating the HRF as well as detecting activation for each stimulus type are of concern. Other settings are the same as the simulations in Sections 4.1 and 4.2; details are presented there.

in these designs. Block designs have high detection powers and m-sequences possess high estimation efficiencies (Figure 1; see also Liu, 2004). Random designs have the potential to reach any corner in the design space, though they might take a long time to achieve this. The concatenations of block designs and m-sequences (or random designs), also known as mixed designs, can provide good designs achieving good trade-offs for the two competing statistical goals (Liu, 2004).

2.2   MO-criterion

The second feature of our approach lies in our multi-objective design criterion, which allows fair, consistent design comparisons. The motivation for our carefully defining the criterion comes from a pitfall we observed in WN's MO-criterion. Their MO-criterion is a weighted sum of the normalized individual criteria, each individual criterion evaluates the
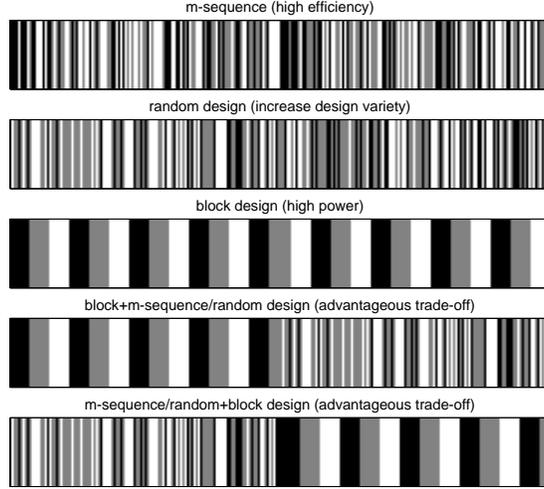
Figure 2: Different design patterns; different shades indicate different event types; white means rest. The area of each shade is proportional to the number of events involved.

achievement of a design with respect to a study objective. The normalization used by WN is

$$\tilde{F}_i^{(g)}(\xi) = \frac{F_i(\xi) - mean_g(F_i)}{SD_g(F_i)}, \tag{1}$$

where $F_i(\xi)$ represents the value of the ith individual criterion for a design ξ in the gth generation of WN's GA, and $mean_g(F_i)$ and $SD_g(F_i)$ are the mean and the standard deviation of F_i-values over the same generation. Designs vary over generations, and the same is true for $mean_g(F_i)$ and $SD_g(F_i)$. As a consequence, the resulting MO-criterion is a moving target during the evolution of WN's GA and fair, consistent design comparisons cannot be achieved. To demonstrate this drawback, an illustrative example is provided below. We use the superscript g to indicate the dependence of WN's design criterion on the GA generation.

Example 2.1. *In Table 1, we compare the two designs ξ_1 and ξ_2 with respect to an MO-criterion $F^{(g)} = 0.5\tilde{F}_1^{(g)} + 0.5\tilde{F}_2^{(g)}$, where $\tilde{F}_i^{(g)}$ is defined in (1). With the hypothetical values of the individual criteria, F_1 and F_2, the mean and the standard deviation of the F_i within each generation can be calculated. For example, $mean_1(F_2) = 6$, $SD_1(F_2) = 1$; $mean_2(F_2) = 6.5$, $SD_2(F_2) = 0.5$. The $F^{(g)}$-value for each design can thus be obtained; i.e., the last column of*

7

the table.

With respect to $F^{(g)}$, ξ_1 is judged to be the best design in the first generation. However, ξ_2 becomes the best in the second generation. The comparison heavily depends on the third designs, ξ_3 for Generation 1 and ξ_3^* for Generation 2. The comparing results are therefore inconsistent.

Table 1: Inconsistent Design Comparisons

Generation 1	(F_1, F_2)	$(\tilde{F}_1^{(g)}, \tilde{F}_2^{(g)})$	$F^{(g)}$	
ξ_1	(20.0, 6.0)	(1.15, 0.00)	0.58	✓
ξ_2	(18.0, 7.0)	(-0.58, 1.00)	0.21	
ξ_3	(18.0, 5.0)	(-0.58,-1.00)	-0.79	
Generation 2				
ξ_1	(20.0, 6.0)	(1.15,-1.00)	0.08	
ξ_2	(18.0, 7.0)	(-0.58, 1.00)	0.21	✓
ξ_3^*	(18.0, 6.5)	(-0.58, 0.00)	-0.29	

F_i: individual design criterion
$\tilde{F}_i^{(g)}$: normalized design criterion using (1)
$F^{(g)} = 0.5\tilde{F}_1^{(g)} + 0.5\tilde{F}_2^{(g)}$
✓: the best design of a generation

Instead of (1), we use the following standardization for larger-the-better design criteria:

$$\tilde{F}_i^*(\xi) = \frac{F_i(\xi) - \min(F_i)}{\max(F_i) - \min(F_i)},$$

where $\min(F_i)$ and $\max(F_i)$ are the global minimum and maximum for F_i. Applying this to the previous example and assuming $0 \leq F_1 \leq 20$ and $0 \leq F_2 \leq 7$, the best design is ξ_2 in both generations; i.e., using $F^* = 0.5\tilde{F}_1^* + 0.5\tilde{F}_2^*$ to replace $F^{(g)}$, we have $F^*(\xi_1) = 0.93$, $F^*(\xi_2) = 0.95$, $F^*(\xi_3) = 0.81$, and $F^*(\xi_3^*) = 0.92$.

This standardization only depends on the extreme values of the individual criteria, which are fixed. Thus, an MO-criterion allowing consistent design comparisons can be defined and the pitfall is remedied. However, obtaining the extreme values might be difficult for some design criteria. One possible way is to use our algorithm to find approximations as detailed in

Section 4. While these additional searches increase the computational burden, our algorithm remains much faster than available competitors (see Section 4.3).

The standardization method also guarantees range comparability among individual criteria. The standardized value of each individual criterion is between 0 and 1 with one corresponding to the optimal design(s) on that dimension and zero the worst. This is a preferred property while combining different criteria (Imhof and Wong, 2000). Clearly, the combination between a criterion ranging from 0.1 to 0.01 with another criterion ranging from 10 to 100 is not desirable.

2.3 Model Formulation

The third feature of our approach lies in the formulation of the statistical model. When estimating the HRF, we follow WN to consider the linear model with a parameter vector describing the HRF. The major advantage of our model lies in the use of the discretization interval (Dale, 1999) when parametrizing the HRF. This interval helps to discretize the continuous HRF so that a finite set of parameters can be used to capture the fluctuation of the HRF over time. Denoting the length of the discretization interval by ΔT, we set ΔT to the greatest value dividing both the ISI and the TR. Our HRF parameters is then associated with the heights of the HRF evaluated every ΔT following the stimulus onset; these heights form a minimal set of heights that could possibly contribute to the observed responses. As a result, the fluctuation of the HRF is faithfully accounted for by our model.

By contrast, the ith HRF parameter in WN's model corresponds to the height of the HRF at the ith scan following the stimulus onset. Each parameter in their models may simultaneously represent more than one height of the HRF when ISI $\neq m$TR. The need for ΔT as well as its use in discretizing the HRF is illustrated in the following example.

Example 2.2. *In Figure 3, we consider one stimulus type ($Q = 1$). The time interval between two consecutive events is 1.5s (ISI = 1.5s), and that between two successive scans is 2s (TR = 2s). An illustrative design is $\xi = \{110100...\}$ with three stimuli taking place at*

9

1.5s, 3s and 6s, respectively. The three stimuli are the blue, green and red vertical bars in the figure. The curve following each of these bars is the HRF evoked. The three HRFs are assumed to be the same because they are triggered by the same stimulus type.

The five black, vertical lines corresponds to the first five MR scans, at which the BOLD signals are collected. The heights of the HRFs, or equivalently the effects of the stimuli, that contribute to the signals collected are indicated by the dots on the lines. These heights are different. Therefore, we need different parameters to represent them as well as any other heights that could possibly contribute to the responses.

Under the current combination of ISI and TR, the heights of our interest are those occur every 0.5s on the HRF curve. They are the red dots and squares on the third curve in Figure 3. We note that 0.5 is the greatest value diving both the ISI and the TR. Setting ΔT to this value, our HRF parameters then describe the discritized HRF, $h(j\Delta T)$, $j = 0, 1, \dots$. Here, $h(t)$ is the HRF at (peri-stimulus) time t with $t = 0$ corresponds to the stimulus onset. All the heights that could possibly make contribution are taken into account. In addition, irrelevant heights that will never contribute to the responses (e.g., heights at 0.1s and 0.2s) are left out.

Another parametrization method is used in WN's program (http://www.columbia.edu/cu/psychology/tor/index.html). They link each of their parameters to the HRF evaluated at each scan following the stimulus onset, regardless of the time interval between the onset and its next scan. This time interval may vary when ISI $\neq m$TR. Thus, each parameter can simultaneously represent more than one height of the HRF. For the previous example, this method uses one parameter to simultaneously represent the heights $h(0)$, $h(0.5)$ and $h(1)$. These heights correspond to the first dots (scans) following stimulus onsets in Figure 3; they are different. Similarly, another parameter is used to describe the second scans following the three onsets. This can reduce the number of parameters. However, the shape of the HRF is not fully described. In addition, each parameter has more than one meaning. We note that the two parametrization methods are equivalent if ISI $= m$TR for a positive integer m.
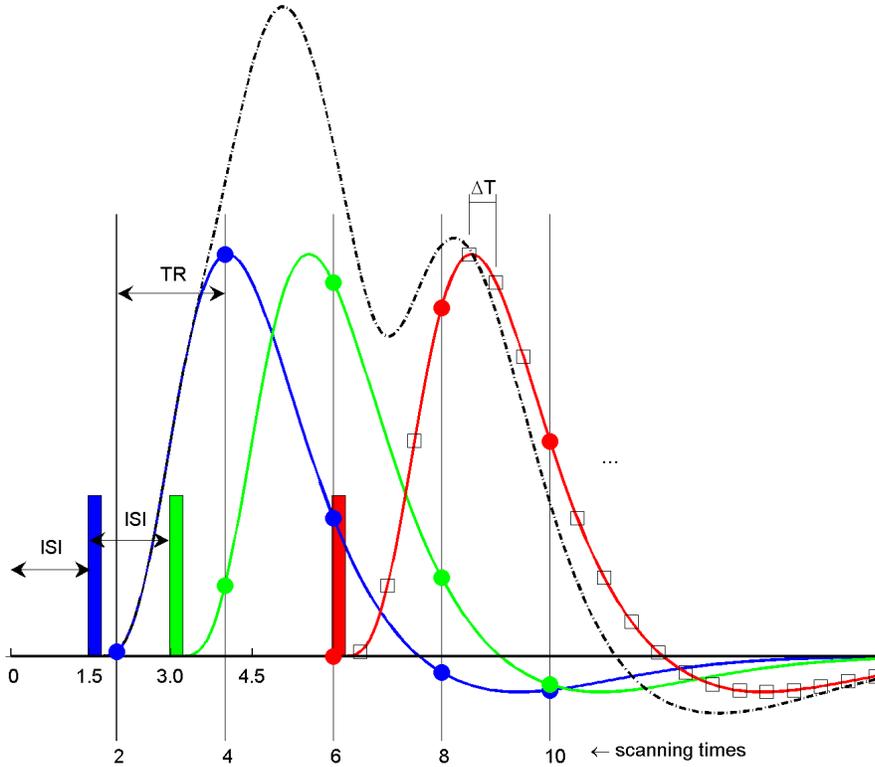
Figure 3: HRF parametrization (see Example 2.2 for details). Here, $Q = 1$, ISI $= 1.5$s, TR $= 2$s, and $\Delta T = 0.5$. The three vertical bars are the stimulus onsets. Solid curves represent the HRF of this stimulus type. Dots are the heights of the HRF that contribute to the responses at the first five scans under $\xi = \{110100...\}$. Squares and dots on the third HRF compose the parameter vector, including all the heights that could possibly contribute to the observed response under this combination of the ISI and TR. The dash-dot curve represents the noise-free, trend-free BOLD responses induced by the three stimuli.

3   Methodology

Our proposed method is detailed in this section. The underlying models for the two statistical goals are introduced. Base on these models, we define the individual design criterion for each study objective, and formulate our MO-criterion. The proposed algorithm is also described in this section.

3.1 Model

As in WN, the underlying models considered for analyzing the BOLD time series are general linear models (see also, Friston et al., 1995; Worsley and Friston, 1995; Dale, 1999; Liu and Frank, 2004). Although the assumptions of linearity and additivity might not hold for dense events, or for a very short ISI, linear models are proven to be powerful and are popular for fMRI studies (e.g., Boynton et al., 1996; Dale and Buckner, 1997; Fuhrmann Alpert et al., 2007; Lindquist et al., 2007).

To estimate the HRF, the following model is used.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{h} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{Y}$ is the voxel-wise BOLD time series, $\boldsymbol{h} = (\boldsymbol{h}_1', ..., \boldsymbol{h}_Q')'$ consists of Q parameter vectors, each representing the HRF of a stimulus type, $\boldsymbol{X} = [\boldsymbol{X}_1 \cdots \boldsymbol{X}_Q]$ is the design matrix, $\boldsymbol{S}\boldsymbol{\gamma}$ is the nuisance term describing the overall mean, trend or drift of $\boldsymbol{Y}$ with $\boldsymbol{\gamma}$ being the nuisance parameter, and $\boldsymbol{e}$ is noise.

The HRF parametrization introduced in Section 2.3 is applied. The parameter vector $\boldsymbol{h}_i = (h_{1i}, ..., h_{ki})'$ represents the HRF, $h_i(t)$, for the type-i stimulus. With the ΔT defined above, we use h_{ji} to describe the height $h_i((j-1)\Delta T)$; $j = 1, ..., k$. Here, the length of $\boldsymbol{h}_i$ is $k = 1 + \lfloor \frac{K}{\Delta T} \rfloor$, where $\lfloor a \rfloor$ is the greatest integer less than or equal to a and K is the duration of the HRF, counting from the stimulus onset to the complete return of the HRF to baseline. Assuming the same HRF duration for all the Q stimulus types is not uncommon (e.g., Liu and Frank, 2004), but our approach can easily be generalized to accommodate HRFs of different durations.

The matrix $\boldsymbol{X}$ is determined by both the design sequence and the HRF parametrization. The matrix corresponding to Example 2.2 is provided below as an illustration. Each column is linked to an h_{j1} ($Q = 1$) and is labeled by $t_j = (j-1)\Delta T$ ($\Delta T = 0.5$). Rows are labeled by scanning times, which are multiples of TR.

12

$$\boldsymbol{X} = \begin{array}{r} \\ \\ 2s \rightarrow \\ 4s \rightarrow \\ 6s \rightarrow \\ 8s \rightarrow \\ 10s \rightarrow \\ 12s \rightarrow \\ \end{array} \begin{array}{cccccccccccccc} 0s & 0.5s & 1s & 1.5s & 2s & 2.5s & 3s & 3.5s & 4s & 4.5s & 5s & 5.5s & \dots \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}.$$

For activation detection, the following linear model is assumed.

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{\eta}. \tag{3}$$

A basis function, $\boldsymbol{h}_0$, for the HRFs is always assumed for detection. Throughout this article, we consider $\boldsymbol{h}_0$ to be the canonical HRF of the software SPM2 (http://www.fil.ion.ucl.ac.uk/spm/software/spm2/), which is a combination of two gamma distributions and is scaled to have a maximal value of one. In model (3), the matrix $\boldsymbol{Z}$, representing the convolution of the stimuli with $\boldsymbol{h}_0$, is thus $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{h}_0$, assuming impulse-like neuronal activity; for details see, e.g., Josephs et al. (1997). The parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_Q)'$ is the response amplitudes, or the maximal heights for the Q HRFs. The term $\boldsymbol{S}\boldsymbol{\gamma}$ is again a nuisance term, and $\boldsymbol{\eta}$ is noise. Detecting activations is equivalent to testing the significance of $\boldsymbol{\theta}$ or its linear combinations based on the researcher's interest.

In this study, the same basis function $\boldsymbol{h}_0$ is assumed for all the Q stimulus types. Only the amplitudes are allow to vary. However, incorporating different basis functions for different stimuli is also possible in our approach. One can take $\boldsymbol{Z} = [\boldsymbol{X}_1\boldsymbol{h}_0^{(1)} \, ... \, \boldsymbol{X}_Q\boldsymbol{h}_0^{(Q)}]$ to proceed, yet this setting is beyond the scope of the current work.

The two linear models are also considered in WN, but with a different HRF parametrization as previously mentioned. Their use can also be found in Liu and Frank (2004). Although the assumptions of linearity might be violated for certain situations, this linear model framework is popular and is powerful for many cases. Therefore, we restrict ourselves

13

to this framework in the current work.

3.2 Design Criteria

With the general linear models, we define our design criteria including four individual design criteria and the MO-criterion. We denote by $\boldsymbol{I}_a$ and $\boldsymbol{0}_{a \times b}$ the $a \times a$ identity matrix and the $a \times b$ matrix of zeros, respectively. T is the total number of scans.

3.2.1 Estimation Efficiency

The first design criterion pertains to estimating estimable linear combinations of the HRF parameters, say $\boldsymbol{C}_x \boldsymbol{h}$. We assume a known $T \times T$ matrix $\boldsymbol{V}$ such that $\boldsymbol{V} \boldsymbol{e}$ is white noise with variance σ^2. The variance-covariance matrix of the generalized least squares (GLS) estimator of $\boldsymbol{C}_x \boldsymbol{h}$ is

$$Cov(\boldsymbol{C}_x \hat{\boldsymbol{h}}_{GLS}) = \sigma^2 \boldsymbol{M}_x = \sigma^2 \boldsymbol{C}_x [\boldsymbol{X}' \boldsymbol{V}' (\boldsymbol{I}_T - P_{\boldsymbol{V}\boldsymbol{s}}) \boldsymbol{V} \boldsymbol{X}]^- \boldsymbol{C}'_x, \tag{4}$$

where $P_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^- \boldsymbol{A}'$ is the orthogonal projection on the vector space spanned by column vectors of $\boldsymbol{A}$, and $\boldsymbol{A}^-$ is any generalized inverse matrix of $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$. When $\boldsymbol{C}_x = \boldsymbol{I}_{Qk}$, the estimation of each HRF is considered. When interest is in the difference between the ith HRF and the jth HRF $(i < j)$, we have

$$\begin{aligned}
\boldsymbol{C}_x &= (\delta'_i - \delta'_j) \otimes \boldsymbol{I}_k \\
&= \begin{cases}
\left[\boldsymbol{0}_{k \times k(i-1)}, \boldsymbol{I}_k, \boldsymbol{0}_{k \times k(j-i-1)}, -\boldsymbol{I}_k, \boldsymbol{0}_{k \times k(Q-j)} \right], & j - i > 1; \\
\left[\boldsymbol{0}_{k \times k(i-1)}, \boldsymbol{I}_k, -\boldsymbol{I}_k, \boldsymbol{0}_{k \times k(Q-j)} \right], & j - i = 1,
\end{cases}
\end{aligned}$$

where δ_i is a $Q \times 1$ vector with one at the ith position and zeros elsewhere, and $\otimes$ is the Kronecker product. Other $\boldsymbol{C}_x$ can also be considered.

In the theory of optimal designs, the A- and D-optimality criteria are popular in evaluating estimation efficiency (e.g., Atkinson and Donev, 1992). An A-optimal design minimizes

the sum or, equivalently, the average of the variances of the parameter estimators. D-optimality endeavors to minimize the generalized variance of the parameter estimators and thus the volume of the confidence ellipsoid of the parameters. Formulated as a larger-the-better criterion, the design criterion for the efficiency of estimating $\boldsymbol{C}_x \boldsymbol{h}$ can be defined as one of these two optimality criteria:

$$\Phi_A(\boldsymbol{M}_x) = \frac{p}{trace(\boldsymbol{M}_x)}, \text{ for } A\text{-optimality;}$$

$$\Phi_D(\boldsymbol{M}_x) = det(\boldsymbol{M}_x)^{-1/p}, \text{ for } D\text{-optimality,}$$

where $det(\cdot)$ stands for the determinant, and p is the dimension of $\boldsymbol{M}_x$. Directly applying these two criteria, our first design criterion is defined below. We use the term "estimation efficiency" to indicate its value.

Definition 3.1. *The design criterion for estimating $\boldsymbol{C}_x \boldsymbol{h}$ using model (2) is*

$$F_e = \Phi(\boldsymbol{M}_x),$$

where $\Phi(\cdot)$ is $\Phi_A(\cdot)$ for A-optimality or $\Phi_D(\cdot)$ for D-optimality, and $\boldsymbol{M}_x$ is defined in (4).

3.2.2 Detection Power

The second design criterion is based on model (3) for detecting activation. Similar to the estimation problem, we work on $Cov(\boldsymbol{C}_z \hat{\boldsymbol{\theta}}_{GLS})$, the variance-covariance matrix of the GLS estimator of $\boldsymbol{C}_z \boldsymbol{\theta}$; see also, Pukelsheim (1993, p. 71), Birn et al. (2002), and Liu and Frank (2004). A- or D-optimality can again be utilized. Similar to (4), we have ($\boldsymbol{V}\eta$ is white noise),

$$Cov(\boldsymbol{C}_z \hat{\boldsymbol{\theta}}_{GLS}) = \sigma^2 \boldsymbol{M}_z = \sigma^2 \boldsymbol{C}_z [\boldsymbol{Z}' \boldsymbol{V}' (\boldsymbol{I}_T - P_{\boldsymbol{V}s}) \boldsymbol{V} \boldsymbol{Z}]^- \boldsymbol{C}_z'. \tag{5}$$

When $\boldsymbol{\theta}$ itself is of concern, we take $\boldsymbol{C}_z = \boldsymbol{I}_Q$. When the difference between the ith and the jth amplitudes is considered, we have $\boldsymbol{C}_z = \delta_i' - \delta_j'$. The second design criterion is then

15

defined as:

Definition 3.2. *The design criterion for detecting activation using model (3) is*

$$F_d = \Phi(\boldsymbol{M}_z),$$

where $\Phi(\cdot)$ is $\Phi_A(\cdot)$ for A-optimality or $\Phi_D(\cdot)$ for D-optimality, and $\boldsymbol{M}_z$ is defined in (5).

The term "detection power" is used to indicate the F_d-value.

3.2.3 Unpredictability

The third design criterion evaluates unpredictability of a design sequence. We would like a sequence that makes it difficult for a subject to anticipate future stimuli based on past stimuli. To achieve this, the Rth order counterbalancing (or R-CB for short) property of WN is considered, where R is a given integer. This property is defined on a sub-design of the original design obtained by keeping only the stimuli but deleting all rests. For any $r \in \{1, ..., R\}$, we count the pairs of stimuli that appear in positions $(t, t + r)$ in the sub-design, $t = 1, .., (n - r)$; n is the length of the sub-design. The Rth order counterbalancing aims at having each pair appear a number of times that is proportional to the product of the specified proportions for the stimuli in the sub-design.

Let $(i, j)_r$ be a pair in the sub-design with a type-j stimulus following a type-i stimulus in r lags ($r - 1$ stimuli in between). The original design criterion defined by WN is

$$1 - \frac{1}{RQ^2} \sum_{i,j}^{Q} \sum_{r}^{R} \left[n_{ij}^{(r)} - E(N_{ij}^{(r)}) \right]^2,$$

where $n_{ij}^{(r)}$ is the observed number of $(i, j)_r$, with $E(N_{ij}^{(r)}) = nP_iP_j$ being its expected value, n is the length of the sub-design, and P_i is the pre-specified proportion of type-i stimuli in the sub-design, for $i = 1, ..., Q$.

The finite length of the design is not accounted for in their definition. Thus, $E(N_{ij}^{(r)})$ is inflated — it should be $(n - r)P_iP_j$. We, however, include this information to define our

16

third design criterion.

Definition 3.3. *The design criterion for the Rth order counterbalancing is*

$$F_c = \sum_{r=1}^{R} \sum_{i,j \in \{1,\ldots,Q\}} \lfloor |n_{ij}^{(r)} - (n-r)P_iP_j| \rfloor,$$

where $n_{ij}^{(r)}$ is the observed number of $(i,j)_r$ in the sub-design, n is the length of the sub-design, and $\lfloor |a| \rfloor$ is the greatest integer less than or equal to the absolute value of a.

Note that, when researchers have no preference for the P_i, $P_i = 1/Q$ can be used for $i = 1, \ldots, Q$. This criterion then measures the departure of the sub-design from the situation where all the Q^2 lag-r pairs $(i,j)_r$ appear as equally often as possible for $r = 1, \ldots, R$.

3.2.4   Desired Stimulus Type Frequency

The fourth design criterion is to evaluate the fulfilment of the desired frequency for each stimulus type. By considering again the sub-design, we define the following design criterion.

Definition 3.4. *The design criterion for the desired stimulus type frequency is*

$$F_f = \sum_{i \in \{1,\ldots,Q\}} \lfloor |n_i - nP_i| \rfloor$$

where nP_i is the desired frequency of the ith stimulus type, n_i is the observed frequency, and n is the length of the sub-design.

This criterion measures the departure of the observed stimulus type frequencies of all stimulus types from the desired ones.

3.2.5 MO-criterion

Since F_e and F_d are larger-the-better and F_c and F_f are smaller-the-better, we define our standardized individual criteria as:

$$\tilde{F}_i^* = \begin{cases} \frac{F_i - min(F_i)}{max(F_i) - min(F_i)}, & i = e, d; \\[2ex] 1 - \frac{F_i - min(F_i)}{max(F_i) - min(F_i)}, & i = f, c. \end{cases}$$

Our MO-criterion is then defined as

$$F^* = w_c \tilde{F}_c^* + w_d \tilde{F}_d^* + w_e \tilde{F}_e^* + w_f \tilde{F}_f^*,$$

where w_i is the weight for each individual design criterion, $i = c, d, e, f$. These weights are determined based on the researcher's interest. Without loss of generality, we let $w_c + w_d + w_e + w_f = 1$.

3.3 Search Algorithm

We propose an algorithm to search for the best multi-objective design. Our algorithm is based on the genetic algorithm technique. We briefly introduce this technique followed by our proposed algorithm.

3.3.1 Genetic Algorithms

GAs (Holland, 1975; 1992) are popular for solving optimization problems. The basic idea is to solve the problems via an evolutionary process which results in better solutions (offsprings) based on good solutions (parents).

By mimicking Darwin's theory of evolution, GAs incorporate natural selection and gene variation in the search for optimal solutions. Natural selection helps progress toward the best solution, and gene variation not only broadens the exploration over the solution space but also facilitates moves out of local optima.

The simplest GA starts with a group of initial chromosomes, called parents. In our context, these chromosomes are design sequences (e.g., $\xi = \{101201210...\}$), and each event in a design is called a gene. Parents are paired to give birth to offsprings using crossover — exchanging the corresponding fractions of paired chromosomes. Mutations and immigrants are often introduced to increase gene variety. The mutation mechanism perturbs randomly chosen genes in offsprings (e.g., replaces these randomly chosen genes by randomly generated ones) and immigrants are newly generated chromosomes. The population is then enlarged to include parents, offsprings and immigrants. To evaluate the "quality" or fitness of each chromosome, a fitness function is specified as per the objective(s) of the study and each chromosome is assigned a fitness score. Based on fitness scores, natural selection prunes the enlarged population to maintain a constant population size. Only the fittest survive to the next generation. The process, when repeated, ensures the preservation of good traits and the optimal chromosome or solution can be expected. This optimization technique is widely applied in different disciplines and many advanced variants are still being proposed (e.g., Mandal et al., 2006; Basokur et al., 2007).

3.3.2 Proposed Algorithm

Our proposed algorithm is detailed below.

Step 1. (Initial designs) Generate G initial designs consisting of random designs, an m-sequence or m-sequence-based design, a block design and their combinations. Use the objective function to evaluate the fitness of each initial design.

Step 2. (Crossover) With probability proportional to fitness, draw with replacement $G/2$ pairs of designs to crossover — select a random cut-point and exchange the corresponding fractions of "genetic material" in paired designs. Here, the "genetic material" is the design sequence.

Step 3. (Mutation) Randomly select $q\%$ of the events from the G offspring designs. Replace these events by randomly generated ones. Here, an event is a stimulus or a rest.

Step 4. (Immigration) Add to the population another I designs drawn from random designs, block designs and their combinations.

Step 5. (Fitness) Obtain the fitness scores of the offsprings and immigrants.

Step 6. (Natural selection) Keep the best G designs according to their fitness scores to form the parents of the next generation. Discard the others.

Step 7. (Stop) Repeat steps 2 through 6 until a stopping rule is met (e.g., after M generations). Keep track of the best design over generations.

Initial Designs and Immigrants

In Step 1, m-sequences or m-sequence-based designs are generated following Liu (2004); see also Buračas and Boynton (2002). These designs are well-known for their high estimation efficiencies. Since they are not always available, concatenations or truncations of the existing ones are also considered. We include the one yielding the highest estimation efficiency as one of the initial designs.

The initial block design has the highest detection power among designs of differing numbers of blocks and of two different patterns. In this pool of candidate block designs, the number of blocks for each stimulus type ranges among one to five, 10, 15, 20, 25, 30, and 40. The two patterns include repetitions of NABC and NANBNC, where N is a block of rests and A, B and C represent blocks of stimuli of different types. In addition to the initial block design, immigrants in Step 4 ensure a steady supply of blocks of different sizes.

The combination of a block design with an m-sequence-based design or a random design is obtained through crossover. These mixed designs constitute a portion, e.g., one-third, of the initial designs. The remaining initial designs are formed by random designs.

Objective Function

The objective function used in Step 1 and Step 5 of our GA evaluates the fitness or "goodness" of the designs. Based on the goal of the search, the objective function can be taken as a single F_i or as the MO-criterion with weights selected by the researcher's interest.

Note that the extreme values of the F_is are required to use the MO-criterion.

Theoretical values of $\max(F_e)$ and $\max(F_d)$ are generally not available. They can be approximated by performing a "pre-run" of our GA using the non-standardized function F_e (or F_d) as the objective function. The values of $\min(F_e)$ and $\min(F_d)$ are set to zero, corresponding to designs for which the parameters of interest are non-estimable. Both $\min(F_c)$ and $\min(F_f)$ are zero. Their maximal values are attained by the design containing only the stimulus type with the smallest specified proportion P_i. With these extreme values, the MO-criterion F^* is well-defined and serves as the objective function for finding optimal multi-objective designs.

In our simulations, $G = 20$, $q = 1$, $I = 4$ and $M = 10,000$ (chosen because a larger value does not seem to lead to significantly better designs).

3.3.3 Tuning Parameter

WN propose a fitness function consisting of random and deterministic components:

$$F_t(\alpha) = U(0, 1) + \frac{1}{1 + e^{-\alpha F}}, \tag{6}$$

where $U(0, 1)$ is a uniformly distributed random number, F represents WN's MO-criterion, and α is an artificial parameter to control the degree of "randomness" of natural selection. When α is close to zero, $F_t(\alpha) \approx U(0, 1) + (1/2)$. Natural selection using this $F_t(\alpha)$ is nearly completely random. On the other hand, increasing α decreases the survival rate of inferior designs (in terms of F); F starts to dominate when α becomes large. An effective value of α suggested by WN is 2.1.

The function $F_t^*(\alpha)$ (using F^* to replace F in (6)) is also considered in our algorithm to calculate the fitness in Step 5. We adapt the concept of double-loop algorithms (e.g., Jin et al., 2005) to provide tuning methods for this artificial parameter α. Our algorithm then consists of an outer loop and an inner loop. The inner loop is still our algorithm proposed in Subsection 3.3.2, but use $F_t^*(\alpha)$ in Step 5. The outer loop adjusts the value of α. After

21

$M_I(= 100)$ generations in the inner loop, the outer loop changes the value of α and sends it back to the inner loop for another run. We repeat this process for $M_O(= 100)$ times. A total of $M_I M_O$ generations has been searched.

The following three tuning methods can be utilized in the outer loop.

- **Cooling:** always decrease α;

- **Warming:** always increase α;

- **Adaptive:** change α according to the search performance.

The first two methods are straightforward. To implement the adaptive tuning method, we collect two quantities from each generation of our GA (the inner loop): 1) the proportion of inferior designs survived, and 2) the improvement status of each generation ($= 1$, when a better design is found in this generation; $= 0$, otherwise). With the fact that the first quantity is negatively correlated to α, we regress the second quantity on the first one using logistic regression. When the regression coefficient is significantly positive (type I error rate $= 0.1$), we increase the proportion by decreasing α since high survival rate seems helpful. When the coefficient is significantly negative, α is increased. The value of α is unchanged for insignificant results.

When no improvement is observed in an inner loop, i.e., the second quantities are all zeros, the adjustment to α depends on the average of the first quantity over this inner loop. When the averaged proportion is greater than a pre-specified upper bound (say 0.5), increase α to force the proportion to go down. A reverse alternation to α is made when the averaged proportion is smaller than a pre-specified lower bound (say 0.1). No adjustment is performed when the averaged proportion falls between the two specified bounds.

The effectiveness of α is studied by comparing our algorithm with its own variants, defined by different tuning methods. When ISI $= m$TR (m is an integer), WN's GA is also compared to our algorithm. When ISI $\neq m$TR, a random search is used instead since WN applies a different HRF parametrization method (Section 2.3). In lieu of reproducing offsprings based on good parents, the random search generates new designs in each generation. The best

22

Table 2: Algorithms compared in the simulations

Name	Algorithm description
Our GA	Our proposed genetic algorithm using the deterministic objective function
Fixed	Our GA using $F_t^*(\alpha)$ for natural selection; $\alpha = 2.1$
Adaptive	Our GA using $F_t^*(\alpha)$ for natural selection; adjust α adaptively
Cooling	Our GA using $F_t^*(\alpha)$ for natural selection; always decrease α
Warming	Our GA using $F_t^*(\alpha)$ for natural selection; always increase α
WN's GA	the GA utilized by WN
Random	random search

design is still tracked over generations. Table 2 lists the algorithms to be compared. In addition to Table 2, we also apply the weighted mutation technique of Mandal et al. (2006) to our algorithm in one simulation. This technique provides an alternative to perform mutation, yet it increases computational burden.

4 Simulations

4.1 Detection Power

The first simulation is to find a design with high detection power. The error $\boldsymbol{e}$ is assumed to be white noise and $\boldsymbol{S}$ is a vector of ones. Time duration of the experiment is set to 486 seconds and ISI = TR = ΔT = 2s with $Q = 2$ (i.e., two stimulus types). There are $L = 243$ events (rests plus stimuli) in the design sequence. Model (3) is considered and $\boldsymbol{C}_z = [1, -1]$; this is to investigate the difference between the two stimulus types. Only detection power is considered (i.e., $w_d = 1$), and F_d is reported without standardization. The A-optimality criterion is utilized, which is equivalent to using D-optimality since $\boldsymbol{M}_z$ is a scalar. The setting is similar to that in WN, and a block design is expected to be the best design.

A total of 10,000 generations with 20 designs plus four immigrants each is searched in our GA. To match this setting, the variants of our GA have 100 iterations in the outer loop and 100 generations in the inner loop. For comparison, WN's GA is also implemented

and each generation in their algorithm contains 24 designs since WN's GA does not include immigrants.

Figure 4 (A1) presents the final designs found by our algorithms (the original one plus its variants) and WN's GA. The F_d-values achieved are also provided. Although the search starts (not shown) with designs containing both rests and stimuli, almost all the rests (white) are excluded in the final designs. This fits the main aim of this simulation, which is to detect the difference between the two stimulus types without taking into account the psychological effects. Figure 4 (A2) presents the achieved F_d-value over generations. Our GA and its variants outperform WN's GA. The algorithmic parameter α has little effect on search performance — our GA and its variants are comparable.

From Figure 4, our GA and its variants converge much faster than WN's GA. The final design found by our GA ($F_d = 365.8295$) is better than that of WN's GA ($F_d = 319.6362$). A systematic search over designs of two to 40 blocks found the best design of 24 blocks yielding an F_d-value of 365.4099. With the fixed block size of ten, this systematic design has three rests at the end. Our final design is superior since our algorithm is more flexible than systematic search.

The computation time of our GA is much shorter than that of WN's GA. On a Pentium Dual 3.20/3.19 GHz computer with 3.5 Gb of RAM, our GA spends six minutes and WN's GA uses 99.73 minutes for this simulation. Among the variants of our GA, the most time-consuming is the adaptive method, which takes 6.52 minutes.

4.2   Estimation Efficiency

The second search is to find the best design for HRF estimation, i.e., the design that yields the maximum of F_e ($w_e = 1$) under $\boldsymbol{C}_x = \boldsymbol{I}_{34}$ ($Q = 2$; $k = 1 + 32/2 = 17$). Model (2) is utilized and all the other settings are the same as in the previous simulation. The final designs are given in Figure 4 (B1) with their estimation efficiencies. Figure 4 (B2) provides the achieved F_e-values over generations. Our GA and its variants are comparable and they

24

(A1)

Fixed   365.9117

our GA   365.8295

Cooling   365.7792

Warming   365.6898

Adaptive   365.6774

WN's GA   319.6362

(B1)

Warming   41.2811

our GA   41.2739

Fixed   41.2495

Adaptive   41.2313

Cooling   41.2108
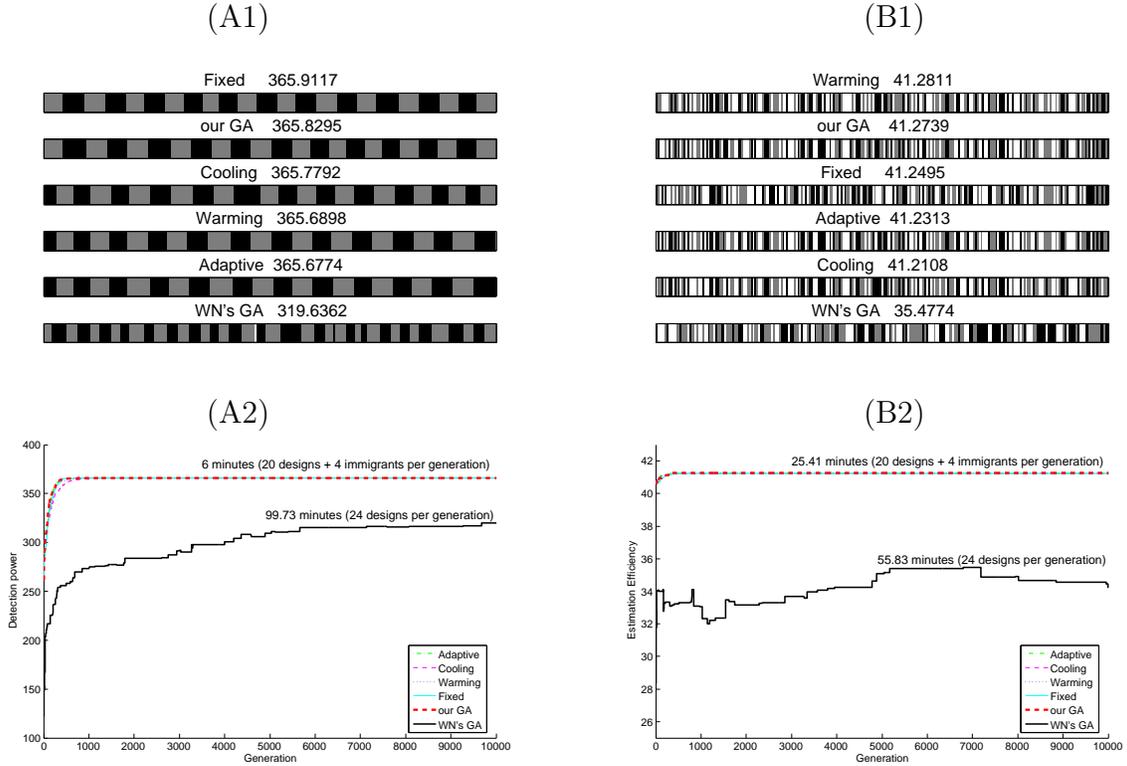
WN's GA   35.4774

(A2)

(B2)

Figure 4: Results of efficient designs for (A) activation detection and (B) HRF estimation. The design for each algorithm is presented with achieved detection power in (A1) and estimation efficiency in (B1). Different shades indicate different events and white represents rests. (A2) and (B2) are evolutions of F_d and F_e over generations, respectively. To finish these simulations, WN's GA takes 99.73 minutes for (A) and 55.83 minutes for (B). Our GA spends 6.00 minutes for (A) and 25.41 minutes for (B).

outperform WN's GA. Our GA uses 25.41 minutes for this simulation whereas WN's GA spends 55.83 minutes. The adaptive method spends 26.68 minutes to finish this simulation.

The estimation efficiency of an m-sequence under this setting is 40.5921, which is slightly lower than that achieved by our GA (F_e= 41.2739). Our GA thus provides a way to obtain near-optimal designs for parameter estimation, without high computational cost. Note that an m-sequence only exists when $Q + 1$ is a prime or a prime power. Therefore, our approach is especially valuable when m-sequences do not exist.

25

4.3   Multi-objective Design

This simulation searches for efficient multi-objective designs by assigning equal weights to the four criteria. The design consists of 255 events. Both model (2) and (3) are considered with a stationary AR(1) error (ρ, the correlation coefficient, is 0.3). The square of the whitening matrix is

$$
V^2 = \begin{pmatrix}
1 & -\rho & 0 & \dots & 0 & 0 \\
-\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\
0 & 0 & 0 & \dots & -\rho & 1
\end{pmatrix}.
$$

In this simulation, $Q = 3$, $C_x = I_{51}$ (again $k = 1 + 32/2 = 17$), $C_z = I_3$, a third-order ($R = 3$) counterbalancing property is considered, and equal frequencies for the three stimulus types are required; i.e., $P_i = 1/3$, $i = 1, 2, 3$.
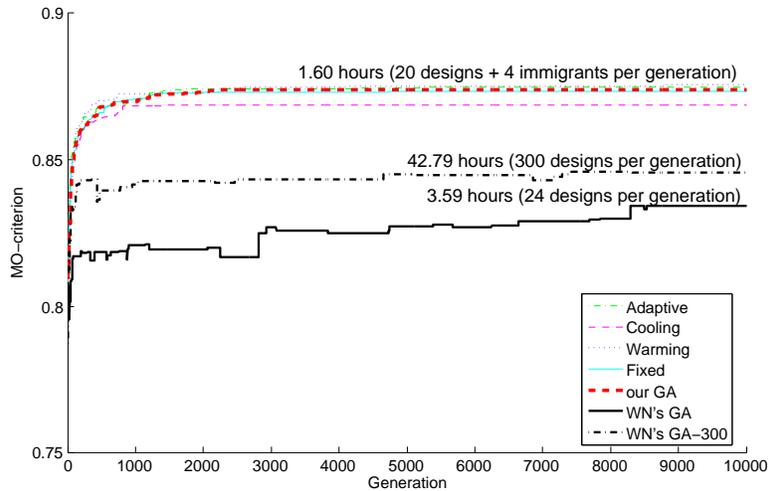

Figure 5: Achieved value of MO-criterion vs. generation for each algorithm in finding efficient multi-objective designs with equal weights assigned to F_c, F_d, F_e, and F_f. Our GA takes 1.60 hours, which includes finding near-maximum value for F_e (0.56 hours), and for F_d (0.14 hours) and performing the main search (0.90 hours). Our design achieves an F^* of 0.875, while WN's GA achieves 0.823 and WN's GA-300 attains 0.834.

Figure 5 compares the estimation efficiencies of the various algorithms. WN's GA does not perform well when each generation has only 24 designs. Using 300 designs per generation (as in WN) improves the result, however, it consumes much more CPU time (42.79 hours for 10,000 generations). A total of 1.60 hours is spent by our GA, including finding $\max(F_e)$ (0.56 hours) and $\max(F_d)$ (0.14 hours). Note that the curve of achieved F_d-value of WN's GA is not monotone, which points at the drawback of their normalization method. This phenomenon is also observed in the previous simulations.

In Figure 6, designs ($\bullet$) found by our GA under different weighting schemes — w_d increased from 0 to 1 by 0.025 and $w_c = w_f = 0$ — are compared with WN's designs (w_d increased from 0 to 1 by 0.05), mixed designs, clustered m-sequences and permuted block designs Liu and Frank (2004). A better trade-off between these two criteria is found by our GA. This frontier is very similar to that (not shown) obtained by increasing w_d from 0 to 0.5 by 0.025 and fixing $w_c = w_f = 0.25$. As expected, smaller efficiencies are found in the latter setting. However, the difference is negligible.
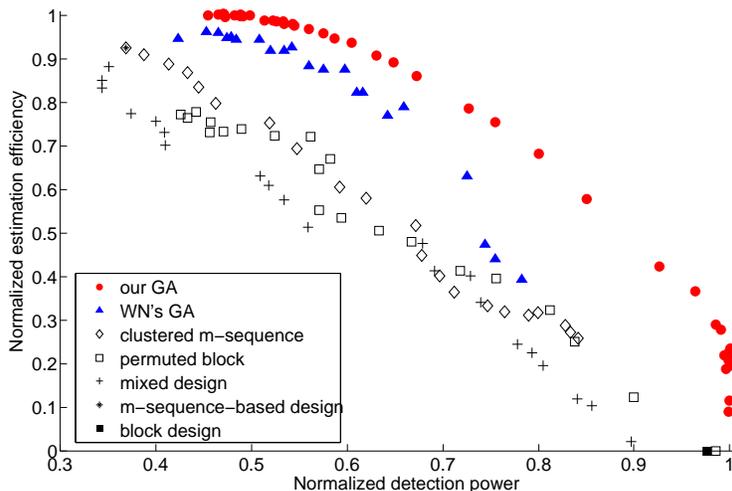

Figure 6: Estimation efficiency vs. detection power for different designs.

Table 3: Comparisons of final designs using WN's criteria

	WN's GA (24)	WN's GA (300)	our GA
Detection power	48.200	47.326	57.207
Estimation efficiency	27.920	29.317	29.870
Counterbalancing	98.293	98.127	95.884
Desired frequency	0.915	0.917	0.913

For comparison, the final design with equal weights obtained from our GA is also inserted to WN's program to evaluate the efficiencies using their design criteria. The results are listed in Table 3. With respect to the two statistical goals, efficiencies of the design found by our GA are higher than those reached by WN's design. The desired frequency is comparable, and WN's design is slightly more counterbalanced. Simulation results for other weighting schemes are presented in Table 4. As presented there, when more weight is assigned to counterbalancing, the F_c values of the designs found by our GA increase while values for the other criteria remain superior to those obtained for WN's GA.

Table 4: Comparisons of algorithms under unequal weights

	Our GA	Adaptive	Cooling	Warming	Fixed	WN's GA
$F^* = (2/6)\tilde{F}_c^* + (2/6)\tilde{F}_d^* + (1/6)\tilde{F}_e^* + (1/6)\tilde{F}_f^*$						
F^*	0.8531	0.8497	0.8505	0.8490	0.8508	0.7836
$\tilde{F}_c^*$	0.9668	0.9652	0.9613	0.9668	0.9652	0.9700
$\tilde{F}_d^*$	0.6859	0.6576	0.6931	0.6648	0.6705	0.4874
$\tilde{F}_e^*$	0.8130	0.8524	0.7942	0.8309	0.8334	0.7865
$\tilde{F}_f^*$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$F^* = (3/6)\tilde{F}_c^* + (2/6)\tilde{F}_d^* + (1/6)\tilde{F}_e^*$						
F^*	0.8407	0.8449	0.8411	0.8461	0.8464	0.7786
$\tilde{F}_c^*$	0.9842	0.9826	0.9842	0.9803	0.9858	0.9668
$\tilde{F}_d^*$	0.6175	0.6196	0.6229	0.6469	0.6305	0.4623
$\tilde{F}_e^*$	0.8565	0.8826	0.8484	0.8418	0.8600	0.8468

4.4   Optimality of the m-sequence

Buračas and Boynton (2002) demonstrate the (near-)optimality of the m-sequence when

the main interest is to estimate the HRF and the model is with white noise but with neither drift nor trend. In this simulation, we follow Buračas and Boynton (2002) to work on white noise and set $\boldsymbol{S}$ in model (2) to a vector of ones, accounting for the overall mean of the fMRI time series. The main focus is to estimate the HRF, $\boldsymbol{h}$, so that $\boldsymbol{C}_x$ is the identity matrix. Different combinations of Q and L used by Liu (2004) are considered. Our GA then finds designs optimizing the estimation efficiency; i.e., $w_e = 1$. For this comparison, we include only random designs as initial designs in our GA. Due to the computation time, here we let the algorithm run for only 2,000 generations at each combination.

We compare our designs to m-sequences, which are demonstrated by Buračas and Boynton (2002) to have high estimation efficiencies. The values of F_e achieved by our designs and by m-sequences are presented in Table 5. The CPU time spent by our GA is also provided. Even without the help of the m-sequence, our GA consistently finds better designs. As shown in Table 5, the proportions of the stimuli in our designs are in good agreement with those optimal proportions approximated by Liu and Frank (2004).

In addition to that introduced in Step 3 of our GA, the weighted mutation proposed by Mandal et al. (2006) provides an alternative method for mutation; see Appendix. This technique can be easily applied in our GA and is applied to find designs under the setting of this simulation. As shown in Table 5, the weighted mutation slight improve the searching result. However, the CPU time spent increases drastically with the length of the design.

Although m-sequences are well-known for their high estimation efficiencies, they are lacking in flexibility. First, these designs are known to exist only when $Q + 1$ is a prime or a prime power. However, our GA can accommodate any number of stimulus types. Second, their high estimation efficiencies hinges on model assumptions. While m-sequences are demonstrated to be (near-)optimal under white noise, they can be outperformed by the designs with the "decorrelation" property described in Buračas and Boynton (2002) when a more realistic assumption of correlated noise is considered. Our GA can take into account any assumptions to the noise and finds better designs. In addition, our approach accommodates any assumption to the trend or drift of the fMRI time series. Furthermore, the proportion

29

Table 5: The F_e^*-values and the proportions of the stimuli

number of types (Q)	3	4	6	7	8	10	12
length of design (L)	255	624	342	511	728	1330	2196
F_e^*-value:							
our GA	33.34	68.39	26.76	36.20	46.68	72.73	105.18
weighted mutation	33.39	68.88	27.06	36.94	48.10	75.07	108.31
m-sequence	31.80	63.08	24.33	31.94	40.72	61.38	85.39
Stimulus proportion:							
our GA	0.21-0.23	0.17	0.12-0.13	0.10-0.11	0.09-0.10	0.08	0.06-0.07
weighted mutation	0.22-0.23	0.17	0.12	0.10-0.11	0.09-0.10	0.08	0.06-0.07
optimum	0.21	0.17	0.12	0.10	0.09	0.08	0.06
CPU time (hours):							
our GA	0.11	0.46	0.37	0.68	1.35	4.51	13.09
weighted mutation	0.57	6.26	1.67	5.55	11.25	49.63	145.75

of the stimuli in an m-sequence is always $1/(Q+1)$. When this proportion is not optimal, e.g., in the current simulation, m-sequences can be sub-optimal. In contrast, our GA finds designs concurring with the approximated optimal stimulus proportion of Liu and Frank (2004).

For this simulation and quite a few other situations, we can find designs yielding higher estimation efficiencies than m-sequences without the benefit of an m-sequence among the initial designs. However, this can be hard when both $\boldsymbol{h}$ and pairwise differences between $\boldsymbol{h}_i$s are of interest, and the model is with white noise but with neither drift nor trend. For that particular situation, the optimal stimulus proportion is $1/(Q+1)$ and finding a design to outperform the m-sequence is hard (Liu and Frank, 2004; Liu, 2004).

4.5 ISI $\neq$ TR

This simulation considers the case when ISI and TR are not equal. The settings are the same as for the previous simulation but with ISI $= 1.5$s, TR $= 2$s and $\Delta T = 0.5s$. In Figure 7, the efficiencies of our GA and its variants are compared to a random search. Random search does not improve much along the way, in terms of the MO-criterion (equal weights are

applied). Each of these algorithms takes about five hours of CPU time (the longest is 5.18 hours). The lengthened computation time is due to the increased length of the parameter vector $\boldsymbol{h}$, and hence the expanded dimension of $\boldsymbol{X}$. Calculating the inverse of the higher dimensional matrix (in obtaining $\boldsymbol{M}_x$) consumes more computational resources. WN's GA is not considered for this simulation since they employ a different HRF parametrization.
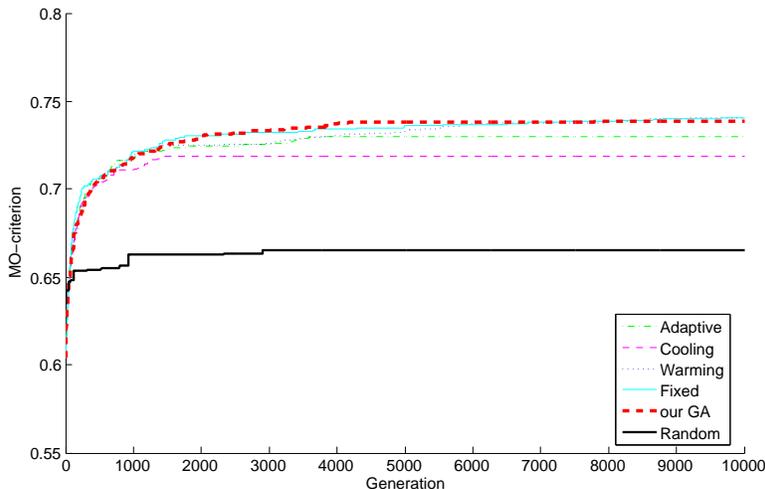

Figure 7: Achieved MO-criterion vs. generation for each algorithm in finding efficient multi-objective designs when ISI = 1.5s and TR = 2s.

5   Conclusions and Discussion

We propose an efficient algorithm to find optimal or near-optimal fMRI designs, which can be used for single or multiple objectives with single or multiple stimulus types. This algorithm outperforms others currently in use by researchers. We consider four particular objectives in this study, yet our algorithm is flexible enough to accommodate other objectives as well.

The Four objectives are considered: HRF estimation, activation detection, unpredictabil-

ity of the design and fulfillment of desired stimulus type frequency. For the first two objectives, A- and D-optimality are introduced. Although the choice between these two criteria should be based on scientific goals, some researchers recommend the D-optimality criterion (e.g. Goos, 2002, p. 37), since D-optimal designs usually perform well with respect to the A-optimality criterion but the reverse relationship is not always true. For comparison, we use A-optimality, as do Wager and Nichols (2003), in our simulations. However, both optimality criteria can easily be implemented in our program.

The Rth order counterbalancing is used to evaluate the achievement of the third objective — unpredictability of the design. This criterion focuses on a sub-design excluding rests. When only one stimulus type is considered, this criterion is of no use and we no longer ask for the third objective. Since this property is defined on the sub-design, the allocation of rests is decided in favor of other study objectives. A stochastic variation or "jitter" in the number of rests between stimuli (and hence the time interval between stimuli) is thus allowed. This stochastic variation is favorable for statistical objectives (e.g. Serences, 2004) and might also provide protection against psychological effects since it is hard to predict the onset of the next stimulus (Amaro and Barker, 2006). Note that a large order of R is believed to be unnecessary due to limited human memory capacity (Liu and Frank, 2004).

We include the desired frequency of each stimulus type as a customized requirement. This property is also defined on the sub-design, and the frequency of rests is adjusted by our GA to yield high efficiencies with respect to other criteria. Consistently, designs found by our GA are in good agreement with the approximated optimal stimulus proportion of Liu and Frank (2004). This optimal proportion is a cornerstone for optimizing the statistical efficiencies.

Other customized requirements, such as a restriction on the number of stimuli in a row before a rest, can also be considered. As long as a design criterion (e.g., observed number minus expected number) can be defined, the criterion can easily be included in the algorithm. In addition, the "forbidden array" introduced in Mandal et al. (2006) might also be useful. This methodology expels all the "forbidden" designs that do not satisfy the requirements.

Our well-defined design criterion ensures that our GA, when it evolves, finds a better design. As pointed out previously, WN's design criterion is a moving target during the search. Achieving a better design is thus not guaranteed. By contrast, our MO-criterion provides a stable, clear target for the search algorithm.

However, using our MO-criterion requires values for $\max(F_e)$ and $\max(F_d)$. These are numerically approximated via our GA. A possible alternative is to follow Liu and Frank (2004) to find analytical approximations. This approach is not applied here because their approximations can be either too wide or too tight for certain cases and it is unknown whether their approximated bounds can actually be achieved by any design; see also Liu (2004). Moreover, the requisite bounds should adapt to a wide range of conditions, such as different correlation structures and nuisance terms. This makes it even harder to derive bounds best suited to each situation. Instead, our approach provides an easy way to numerically approximate the extreme values.

Building upon our well-defined design criterion, we take advantage of good ER-fMRI designs to develop an efficient search algorithm. Conceptually, our algorithm follows Holland's (1975) notion of building blocks; see also Goldberg (1989); Michalewicz (1996). Rooted in the fundamental theorem of GAs, also known as the schema theorem, the building block hypothesis views these constructs as the driving engine for GAs (Goldberg, 1989). Ensuring a good supply of these building blocks is thus one of the key steps for developing good GAs (Goldberg, 2002; Ahn, 2006). The inclusion of good ER-fMRI designs as both initial designs and immigrants follows this concept. Furthermore, using good ER-fMRI designs as initial designs also means that our algorithm starts from a good position.

We also consider the artificial tuning parameter α of WN and propose tuning methods to adaptively adjust its value. However, it does not yield much in terms of efficiency to our algorithm. Further, no dominating tuning method is revealed. Therefore, we recommend the simplest version of the algorithm (i.e., without α). In addition, the weighted mutation technique tends to yield designs with slightly higher efficiencies than the random mutation. However, the CPU time spent increases drastically with the length of the design sequence.

Again, we suggest the use of the simplest version of our GA. An efficiently way to incorporate the weighted mutation technique to our algorithm is an interesting future research.

References

Ahn, C. W. (2006), *Advances in Evolutionary Algorithms: Theory, Design and Practice*, Studies in computational intelligence,, Berlin; New York: Springer.

Amaro, E. J. and Barker, G. J. (2006), "Study Design in MRI: Basic Principles," *Brain and Cognition*, 60, 220–232.

Atkinson, A. C. and Donev, A. N. (1992), *Optimum Experimental Designs*, Oxford statistical science series, Oxford [England] New York: Clarendon Press; Oxford University Press.

Bandettini, P. A. and Cox, R. W. (2000), "Event-Related fMRI Contrast When Using Constant Interstimulus Interval: Theory and Experiment," *Magnetic Resonance in Medicine*, 43, 540–548.

Barker, H. A. (2004), "Primitive Maximum-Length Sequences and Pseudo-Random Signals," *Transactions of the Institute of Measurement & Control*, 26, 339–348.

Basokur, A. T., Akca, I., and Siyam, N. W. A. (2007), "Hybrid Genetic Algorithms in View of the Evolution Theories with Application for the Electrical Sounding Method," *Geophysical Prospecting*, 55, 393–406.

Birn, R. M., Cox, R. W., and Bandettini, P. A. (2002), "Detection versus Estimation in Event-Related fMRI: Choosing the Optimal Stimulus Timing," *NeuroImage*, 15, 252–264.

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996), "Linear Systems Analysis of Functional Magnetic Resonance Imaging in Human V1," *J. Neurosci.*, 16, 4207–4221.

Buračas, G. T. and Boynton, G. M. (2002), "Efficient Design of Event-Related fMRI Experiments Using M-Sequences," *NeuroImage*, 16, 801–813.

Buxton, R. B., Liu, T. T., Martinez, A., Frank, L. R., Luh, W. M., and Wong, E. C. (2000), "Sorting Out Event-Related Paradigms in fMRI: The Distinction Between Detecting an Activation and Estimating the Hemodynamic Response," *NeuroImage*, 11, S457.

Callan, A. M., Callan, D. E., Tajima, K., and Akahane-Yamada, R. (2006), "Neural Processes Involved with Perception of Non-Native Durational Contrasts," *Neuroreport*, 17, 1353–1357.

Dale, A. M. (1999), "Optimal Experimental Design for Event-Related fMRI," *Human Brain Mapping*, 8, 109–114.

Dale, A. M. and Buckner, R. L. (1997), "Selective Averaging of Rapidly Presented Individual Trials Using fMRI," *Human Brain Mapping*, 5, 329–340.

Deb, K. (2001), *Multi-Objective Optimization Using Evolutionary Algorithms*, Wiley-Interscience Series in Systems and Optimization, Chichester; New York: John Wiley & Sons, 1st ed.

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., and Turner, R. (1995), "Analysis of fMRI Time-Series Revisited," *NeuroImage*, 2, 45–53.

Fuhrmann Alpert, G., Sun, F. T., Handwerker, D., D'Esposito, M., and Knight, R. T. (2007), "Spatio-Temporal Information Analysis of Event-Related Bold Responses," *NeuroImage*, 34, 1545–1561.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Reading, Massachusetts: Addison-Wesley.

— (2002), *The Design of Innovation : Lessons from and for Competent Genetic Algorithms*, Boston: Kluwer Academic Publishers.

Goos, P. (2002), *The Optimal Design of Blocked and Split-Plot Experiments*, New York: Springer.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications To Biology, Control, and Artificial Intelligence*, Ann Arbor: University of Michigan Press.

— (1992), *Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Complex adaptive systems, Cambridge, Mass.: MIT Press, 1st ed.

Imhof, L. and Wong, W. K. (2000), "A Graphical Method for Finding Maximin Efficiency Designs," *Biometrics*, 56, 113–117.

Jin, R., Chen, W., and Sudjianto, A. (2005), "An Efficient Algorithm for Constructing Optimal Design of Computer Experiments," *Journal of Statistical Planning and Inference*, 134, 268–287.

Josephs, O., Turner, R., and Friston, K. (1997), "Event-Related fMRI," *Human Brain Mapping*, 5, 243–248.

Lindquist, M. A., Waugh, C., and Wager, T. D. (2007), "Modeling State-Related fMRI Activity Using Change-Point Theory," *NeuroImage*, 35, 1125–1141.

Liu, T. T. (2004), "Efficiency, Power, and Entropy in Event-Related fMRI with Multiple Trial Types: Part II: Design of Experiments," *NeuroImage*, 21, 401–413.

Liu, T. T. and Frank, L. R. (2004), "Efficiency, Power, and Entropy in Event-Related fMRI with Multiple Trial Types: Part I: Theory," *NeuroImage*, 21, 387–400.

Liu, T. T., Frank, L. R., Wong, E. C., and Buxton, R. B. (2001), "Detection Power, Estimation Efficiency, and Predictability in Event-Related fMRI," *NeuroImage*, 13, 759–773.

Mandal, A., Jeff Wu, C. F., and Johnson, K. (2006), "SELC: Sequential Elimination of Level Combinations by Means of Modified Genetic Algorithms," *Technometrics*, 48, 273–283.

Michalewicz, Z. (1996), *Genetic Algorithms + Data Structures = Evolution Programs*, Berlin; New York: Springer-Verlag, 3rd ed.

Miettinen, K. (1999), *Nonlinear Multiobjective Optimization*, International Series in Operations Research & Management Science, Boston: Kluwer Academic Publishers.

Pukelsheim, F. (1993), *Optimal Design of Experiments*, Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics, New York: Wiley.

Ramautar, J. R., Slagter, H. A., Kok, A., and Ridderinkhof, K. R. (2006), "Probability Effects in the Stop-Signal Paradigm: The Insula and the Significance of Failed Inhibition," *Brain Research*, 1105, 143–154.

Rosen, B. R., Buckner, R. L., and Dale, A. M. (1998), "Event-Related functional MRI: Past, Present, and Future," *PNAS*, 95, 773–780.

Serences, J. T. (2004), "A Comparison of Methods for Characterizing the Event-Related BOLD Timeseries in Rapid fMRI," *NeuroImage*, 21, 1690–1700.

Summerfield, C., Egner, T., Mangels, J., and Hirsch, J. (2006), "Mistaking a House for a Face: Neural Correlates of Misperception in Healthy Humans," *Cerebral Cortex*, 16, 500–508.

Wager, T. D. and Nichols, T. E. (2003), "Optimization of Experimental Design in fMRI: A General Framework Using A Genetic Algorithm," *NeuroImage*, 18, 293–309.

Wang, Y. P., Xue, G., Chen, C. S., Xue, F., and Dong, Q. (2007), "Neural Bases of Asymmetric Language Switching in Second-Language Learners: An ER-fMRI Study," *Neuroimage*, 35, 862–870.

Wong, W. K. (1999), "Recent Advances in Multiple-Objective Design Strategies," *Statistica Neerlandica*, 53, 257–276.

Worsley, K. J. and Friston, K. J. (1995), "Analysis of fMRI Time-Series Revisited–Again," *NeuroImage*, 2, 173–181.

Appendix

Weighted Mutation. While the mutation (in Step 3) of our GA is guided by a random mechanism, weighted mutation makes use of the information obtained from the previous generations of our GA. To apply this technique, we define a factor as a location in the design sequence. The weighted mutation scheme then identifies factors (locations) that have significant impacts on fitness. If a factor L_j is significant, the mutation probability, denoted by $p_{j\ell}$, for the event at L_j to "mutate" to the ℓth event type is

$$p_{j\ell} \propto \bar{F}^*(L_j = l), \quad \text{for } j = 1, 2, \ldots, L, \text{ and } \ell = 0, 1, \ldots, Q;$$

i.e., $p_{j\ell}$ is proportional to the averaged fitness over the design sequences with a type-ℓ event located at L_j. Note that the design sequences that guide this mutation are collected from the GA generations performed so far. If L_j does not have a significant main effect, the mutation is again driven by the random mechanism introduced in Step 3; i.e., $p_{j1} = \ldots = p_{jQ}$. Details about weighted mutation can be found in Mandal et al. (2006).