ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 21 (2004) 387–400

Efficiency, power, and entropy in event-related fMRI with multiple trial types
Part I: theory

Thomas T. Liu[a,*] and Lawrence R. Frank[a,b]

[a] *Center for Functional Magnetic Resonance Imaging, University of California San Diego, La Jolla, CA, USA*
[b] *Veterans Administration San Diego Healthcare System, La Jolla, CA, USA*

Received 27 June 2003; revised 29 August 2003; accepted 10 September 2003

Experimental designs for functional magnetic resonance imaging (fMRI) experiments can be characterized by their estimation efficiency, which is a measure of the variance in the estimate of the hemodynamic response function (HRF), and their detection power, which is a measure of the variance in the estimate of the amplitude of functional activity. Previous studies have shown that there exists a fundamental trade-off between efficiency and power for experiments with a single trial type of interest. This paper extends the prior work by presenting a theoretical model for the relation between detection power and estimation efficiency in experiments with multiple trial types. It is shown that the trade-off between efficiency and power present in multiple-trial-type experiments is identical in form to that observed for single-trial-type experiments. Departures from the predicted trade-off due to the inclusion of basis function expansions and the assumption of correlated noise are examined. Finally, conditional entropy is introduced as measure for the randomness of a design, and an empirical relation between entropy and estimation efficiency is presented.

Keywords: Estimation efficiency; Conditional entropy; fMRI

Introduction

Event-related experimental designs for functional magnetic resonance imaging (fMRI) are useful for many cognitive experiments, in part because of their ability to avoid the confounds, such as habituation and anticipation, of more traditional block designs (Rosen et al., 1998). In addition, event-related designs offer the opportunity to rapidly estimate the hemodynamic response function (HRF) to a short stimulus. It has been previously shown that for experimental designs with one trial type, there is a fundamental trade-off between estimation efficiency, the ability to estimate the HRF, and detection power, the ability to detect functional activation (Birn et al., 2002; Liu et al., 2001). A third factor, the perceived randomness of a design, is also important in selecting a design. Randomness can be critical for reducing confounds that arise when a subject can too easily predict the stimulus pattern, and was found to increase with efficiency for experiments with one trial type (Liu et al., 2001).

Block designs in which individual events are clustered into "on" periods of activation alternated with "off" control periods offer nearly optimal detection power at the expense of poor estimation efficiency and randomness. By contrast, designs in which the interstimulus intervals between events are properly randomized offer nearly optimal estimation efficiency (Dale, 1999) but relatively low detection power. To achieve intermediate trade-offs between randomness, estimation efficiency, and detection power, other types of designs, such as stochastic, mixed, and semi-random designs, have been proposed (Friston et al., 1999; Liu et al., 2001). For example, a semi-random design can double the detection power of a randomized design with only a 20% decrease in estimation efficiency and a 10% decrease in randomness (Liu et al., 2001). Such a design would be useful in experiments where a degree of randomness is required for the cognitive paradigm but where the emphasis is on detection of functional activity as opposed to estimation of the HRF.

Although experimental designs with one trial type have found widespread use, especially for the study of sensory areas, many cognitive experiments require the use of designs with multiple trial types. For example, in an fMRI study of face perception and memory, Clark et al. (1998) used an experimental design with three trial types (images of a target face, novel faces, and nonsense scrambled faces) plus a blank screen image as a control state. Experiments with multiple trial types allow for the statistical assessment of each trial type versus the control state and also for contrasts between trial types. As in experiments with a single trial type, the goals of a multiple-trial-type experiment may require either maximal detection power, maximal estimation efficiency, maximal randomness, or some intermediate trade-off between power, efficiency, and randomness.

In this paper, we consider the problems of estimation and detection using the framework of the general linear model (GLM) for experiments with multiple trial types. We present metrics for

* Corresponding author. Center for Functional Magnetic Resonance Imaging, University of California San Diego, 9500 Gilman Drive, MC 0677 La Jolla, CA 92093-0677. Fax: +1-858-822-0605.
E-mail address: ttliu@ucsd.edu (T.T. Liu).
Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.09.030

estimation efficiency and detection power and derive a model describing the fundamental trade-off between the two metrics. We then show that the form of the trade-off is identical to that previously derived for experiments with a single trial type. We also introduce conditional entropy as a measure for the randomness of a design, and present an empirical relation between entropy and estimation efficiency.

The relation between estimation efficiency and detection power is further explored by considering a generalized definition of estimation efficiency in which the HRF is assumed to be a linear combination of basis functions. Basis function expansions represent one means of taking advantage of a priori knowledge about the shape of the HRF (Friston et al., 1998). It is shown that estimation efficiency and detection power represent two limiting cases of the generalized definition, with estimation efficiency corresponding to the use of as many basis functions as there are unknown HRF parameters, and detection power corresponding to the use of only one basis function. The efficiency obtained with intermediate choices of basis function expansions is examined both theoretically and with numerical simulations. Finally, the presence of correlated noise in fMRI experiments is now well documented (Burock and Dale, 2000), and we examine its effect on estimation efficiency and detection power.

The implications of our findings for the optimal design of fMRI experiments are explored further in the companion paper (Liu, 2004).

Theory

General linear model

The GLM provides a flexible framework for analyzing fMRI signals (Dale, 1999; Friston et al., 1995). We assume that there are Q trial types and HRFs of interest, with each HRF of length k. In matrix notation, the signal model is

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{S}\mathbf{b} + \mathbf{n} \tag{1}$$

where $\mathbf{y}$ is a $N \times 1$ vector that represents the observed fMRI time series, $\mathbf{X}$ is a $N \times (kQ)$ design matrix, $\mathbf{h}$ is a $(kQ) \times 1$ vector with the Q HRFs, $\mathbf{S}$ is a $N \times l$ matrix consisting of nuisance model functions, $\mathbf{b}$ is a $l \times 1$ vector of nuisance parameters, and $\mathbf{n}$ is a $N \times 1$ vector that represents additive Gaussian noise with covariance matrix $\mathbf{\Sigma}$. The matrix square root of $\mathbf{\Sigma}$ is denoted by $\mathbf{\Sigma}^{1/2}$, where $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{1/2})^T$. The design matrix may be written as $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_Q]$ and the vector of HRFs as $\mathbf{h} = [\mathbf{h}_1^T \cdots \mathbf{h}_2^T \cdots \mathbf{h}_Q^T]^T$, where each $N \times k$ matrix $\mathbf{X}_q$ consists of shifted binary stimulus patterns for the qth trial type and $\mathbf{h}_q$ is the $k \times 1$ vector for the corresponding HRF.

Estimation efficiency

The minimum variance unbiased estimator (MVUB) of $\mathbf{h}$ can be obtained by multiplying Eq. (1) by $\mathbf{\Sigma}^{-1/2}$ and then applying the theory of oblique projections (Liu et al., 2001; Scharf and Friedlander, 1994; Seber, 1977) to yield:

$$\hat{\mathbf{h}} = (\mathbf{X}_\perp^T\mathbf{X}_\perp)^{-1}\mathbf{X}_\perp^T\mathbf{\Sigma}^{-1/2}\mathbf{y} \tag{2}$$

where $\mathbf{X}_\perp = \mathbf{P}_{\tilde{\mathbf{S}}}^{\perp}\mathbf{\Sigma}^{-1/2}\mathbf{X}$ is the whitened design matrix with whitened nuisance effects removed from each column by the projection matrix $\mathbf{P}_{\tilde{\mathbf{S}}}^{\perp} = \mathbf{I} - \tilde{\mathbf{S}}(\tilde{\mathbf{S}}^T\tilde{\mathbf{S}})^{-1}\tilde{\mathbf{S}}^T$ with whitened nuisance matrix matrix $\tilde{\mathbf{S}} = \mathbf{\Sigma}^{-1/2}\mathbf{S}$. As a practical consideration, Eq. (2) may be rewritten without matrix square roots as: $\hat{\mathbf{h}} = (\mathbf{X}^T\mathbf{K}\mathbf{X})^{-1}\mathbf{X}\mathbf{K}\mathbf{y}$, where $\mathbf{K} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{S}(\mathbf{S}^T\mathbf{\Sigma}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{\Sigma}^{-1}$. The covariance of the estimate is $\mathbf{C}_{\hat{\mathbf{h}}} = (\mathbf{X}_\perp^T\mathbf{X}_\perp)^{-1} = \mathbf{J}^{-1}$, where $\mathbf{J} = \mathbf{X}_\perp^T\mathbf{X}_\perp$ is the Fisher information matrix for $\hat{\mathbf{h}}$ (Scharf, 1991). The Fisher information matrix is a measure of the amount of information about $\mathbf{h}$ that is present in the data (Cover and Thomas, 1991). In a later section, we present an approximation for the information matrix that is critical for elucidating the relation between estimation efficiency and detection power.

In this paper, we consider the problem of estimating the HRF for each trial type and all pairwise differences between HRFs. To do so, we define the contrast matrix $\mathbf{L}_{ij} = \mathbf{D}_{ij} \otimes \mathbf{I}_k$, where

$$\mathbf{D}_{ij} = \begin{cases} \delta_i & \text{for } i = j \\ (\delta_i - \delta_j) & \text{for } i \neq j \end{cases} \tag{3}$$

is a $1 \times Q$ row vector, δ_i is a $1 \times Q$ Kronecker delta row vector with a 1 in the ith column and zeros everywhere else, $\mathbf{I}_k$ is the $k \times k$ identity matrix, and the symbol $\otimes$ denotes a Kronecker product as defined in Appendix A1. With this notation, the estimate of the HRF for the ith trial type is given by $\hat{\mathbf{h}}_i = \mathbf{L}_{ii}\hat{\mathbf{h}}$ and the estimate of the contrast between the HRFs for the ith and jth trial types is $\hat{\mathbf{h}}_{ij} = \mathbf{L}_{ij}\hat{\mathbf{h}}$. The covariance for each contrast is

$$\mathbf{C}_{ij} = \mathbf{L}_{ij}\mathbf{C}_{\hat{\mathbf{h}}}\mathbf{L}_{ij}^T = \mathbf{L}_{ij}(\mathbf{X}_\perp^T\mathbf{X}_\perp)^{-1}\mathbf{L}_{ij}^T \tag{4}$$

with total variance given by the trace $Tr[\mathbf{C}_{ij}]$ of the covariance. We define the overall estimation efficiency as the inverse of the variances averaged across trial types and contrasts:

$$\xi_{\text{tot}} = \frac{1}{\frac{1}{N_Q}\sum_{i \leq j} Tr[\mathbf{C}_{ij}]} \tag{5}$$

where $N_Q = Q + Q(Q-1)/2$ is the number of trial types plus the number of unique pairwise contrasts. Other metrics, such as the minimum efficiency across contrasts $\xi_{\min} = \min_{i,j}(1/Tr[\mathbf{C}_{ij}])$, may also be useful in quantifying the overall efficiency of a design.

Detection power

For each contrast of interest, the process of detection can be formally stated as the test of the null hypothesis $H_0 : \mathbf{L}_{ij}\mathbf{h} = 0$. The test can be accomplished with a F-statistic, which follows a non-central F distribution with non-centrality parameter η_{ij} when the null hypothesis is not true. In Liu et al. (2001), it was shown that the non-centrality parameter serves as a metric for detection power. The non-centrality parameter depends on the shape of the HRF (Liu et al., 2001), and thus to assess the detection power of a design, it is useful to make a priori assumptions about the shape of the HRF for each trial type. This is equivalent to replacing each HRF $\mathbf{h}_i$ in the GLM with $\mu_i\bar{\mathbf{h}}_i$, where $\bar{\mathbf{h}}_i$ is the assumed shape of the HRF for the ith trial type and μ_i is the response amplitude. With this assumption, Eq. (1) may be rewritten as

$$\mathbf{y} = \mathbf{Z}\mu + \mathbf{S}\mathbf{b} + \mathbf{n} \tag{6}$$

where $\mathbf{Z} = [\mathbf{z}_1^T\ \mathbf{z}_2^T \cdots \mathbf{z}_Q^T]^T$, $\mathbf{z}_i = \mathbf{X}_i\bar{\mathbf{h}}_i$ is the regressor for the ith trial type obtained by convolving the stimulus pattern with the assumed HRF shape, and $\mu = [\mu_1\ \mu_2 \cdots \mu_Q]^T$ is the vector of unknown

response amplitudes. The MVUB for the amplitude vector is $\hat{\mu} = (\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{Z}_\perp^T\Sigma^{-1/2}\mathbf{y}$, where $\mathbf{Z}_\perp = \mathbf{P}_{\tilde{\mathbf{S}}}^\perp\Sigma^{-1/2}\mathbf{Z}$, and the variance for the ijth contrast estimate $\hat{\mu}_{ij} = \mathbf{D}_{ij}\hat{\mu}$ is $C_{ij} = \mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T$, where C_{ij} is a scalar. The efficiency for each contrast is defined as

$$\xi_{ij} = [\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T]^{-1} \tag{7}$$

The F statistic for the ijth contrast is

$$F_{ij} = \frac{N-Q-l}{1}\frac{\hat{\mu}_{ij}^2[\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T]^{-1}}{\mathrm{RSS}} \tag{8}$$

where $\mathrm{RSS} = (\mathbf{y} - \mathbf{Z}\hat{\mu} - \mathbf{S}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{Z}\hat{\mu} - \mathbf{S}\hat{\mathbf{b}})$ is the residual sum of squares for the full model (Seber, 1977) with non-centrality parameter (Liu et al., 2001; Scharf and Friedlander, 1994)

$$\eta_{ij} = \mu_{ij}^2[\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T]^{-1} \tag{9}$$

where $\mu_{ij} = \mathbf{D}_{ij}\mu$. Eqs. (7) and (9) show that the non-centrality parameter is proportional to the efficiency of estimating response amplitudes with a priori assumptions about the shape of the HRF for each trial type (Birn et al., 2002; Liu et al., 2001).

In comparing designs, a reasonable a priori assumption is that the shape of the HRFs is constant across trial types so that $\bar{\mathbf{h}}_i = \bar{\mathbf{h}}_0$, where $\bar{\mathbf{h}}_0$ is the assumed base HRF shape. In addition, it is useful to normalize η_{ij} by $\mu_{ij}^2\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0$ to obtain the Rayleigh quotient (Strang, 1980):

$$R_{ij} = \frac{[\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T]^{-1}}{\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0} \tag{10}$$

We define the overall detection power as the harmonic mean of the Rayleigh quotients

$$R_{\mathrm{tot}} = \frac{1}{\frac{1}{N_Q}\sum_{i\le j}\frac{1}{R_{ij}}} = \frac{1}{\frac{\mathbf{h}_0^T\mathbf{h}_0}{N_Q}\sum_{i\le j}C_{ij}} \tag{11}$$

which is equivalent to the overall efficiency, normalized by $\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0$, for the amplitude estimates μ_{ij}. As with estimation efficiency, other metrics, such as the minimum detection power across contrasts $R_{\mathrm{min}} = \min_{i,j} R_{ij}$ may also be useful.

The structure of the Fisher information matrix $\mathbf{X}_\perp^T\mathbf{X}_\perp$

The key term in the expressions for estimation efficiency and detection power is the Fisher information matrix $\mathbf{X}_\perp^T\mathbf{X}_\perp$. In this section, we present an approximate model for the Fisher information matrix. Recall that the Fisher information matrix is the inverse of the covariance of the estimate $\hat{\mathbf{h}}$. We focus our analysis on experimental designs in which the stimuli from different trial types do not overlap. That is, at each time point in the design, only one trial type may have a stimulus present. Note that even with a non-overlapping design, the HRFs from different trial types can overlap because the spacing between stimuli can be smaller that the temporal widths of the HRFs. We also assume that the number of events per trial type is the same across trial types. Thus, for a design with N points, Q trial types, and m events per trial type, there are N - Qm null events represented by zeros in the design matrix.

The Fisher information matrix consists of block diagonal terms of the form $\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,q}$ and block off-diagonal terms of the form $\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,r}$, where $\mathbf{X}_{\perp,q} = \mathbf{P}_{\tilde{\mathbf{S}}}^\perp\Sigma^{-1/2}\mathbf{X}_q$ is the qth submatrix of $\mathbf{X}_\perp$ and $\mathbf{X}_q$ was defined previously as the design matrix corresponding to the qth trial type. For example, with two trial types, the Fisher information matrix is

$$\mathbf{X}_\perp^T\mathbf{X}_\perp = \begin{bmatrix}\mathbf{X}_{\perp,1}^T\\ \mathbf{X}_{\perp,2}^T\end{bmatrix}[\mathbf{X}_{\perp,1} \quad \mathbf{X}_{\perp,2}] = \begin{bmatrix}\mathbf{X}_{\perp,1}^T\mathbf{X}_{\perp,1} & \mathbf{X}_{\perp,1}^T\mathbf{X}_{\perp,2}\\ \mathbf{X}_{\perp,2}^T\mathbf{X}_{\perp,1} & \mathbf{X}_{\perp,2}^T\mathbf{X}_{\perp,2}\end{bmatrix}$$

The block diagonal terms are the autocorrelation matrices for the whitened and detrended stimulus patterns for each trial type, while the block off-diagonal terms are the cross-correlation matrices for stimulus patterns of different trial types. If the stimulus patterns for the different trial types are similar, that is, they are all block designs or all random designs, then a reasonable first order approximation is that the block diagonal terms $\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,q}$ have similar forms and can be represented by an average autocorrelation matrix $\boldsymbol{A}_k \approx \frac{C}{Q}\Sigma_{q=1}^{Q}\boldsymbol{X}_{\perp,q}^T\boldsymbol{X}_{\perp,q}$, where C is a scaling factor. In addition, the block off-diagonal terms can be approximated by $\rho(q,r)\mathbf{A}_k$, where $\rho(q,r)$ represents the cross-correlation between the whitened and detrended stimulus patterns of different trial types, that is, $\rho(q,r) \approx (\mathbf{X}_{\perp,q}(:,1))^T\mathbf{X}_{\perp,r}(:,1)$, where the notation $\mathbf{X}_{\perp,r}(:,1)$ denotes the first column of $\mathbf{X}_{\perp,r}$. With these approximations, the Fisher information matrix may be written as the Kronecker product

$$\mathbf{X}_\perp^T\mathbf{X}_\perp \approx \mathbf{E}_Q \otimes \mathbf{A}_k \tag{12}$$

where $\mathbf{E}_Q$ is a $Q \times Q$ matrix composed of the $\rho(q,r)$ cross-correlation terms.

To further motivate the form of the approximate model stated in Eq. (12) and develop its properties, we consider two limiting cases: (i) one unknown parameter per trial type but multiple trial types, and (ii) multiple parameters per trial type but only one trial type. In doing so, we assume that the additive noise is uncorrelated with covariance matrix $\Sigma = \sigma^2\mathbf{I}_N$ and that the nuisance model matrix $\mathbf{S}$ consists of a single constant vector.

In the first limiting case, we examine the structure of the Fisher information matrix when there is only one unknown parameter per trial type, that is, $k = 1$ and $Q > 1$. For this case, $\mathbf{A}_k$ is a scalar so that the Fisher information matrix reflects the form of $\mathbf{E}_Q$. The elements of the Fisher information matrix are the inner products between submatrices $\mathbf{X}_{\perp,q}$, which are $N \times 1$ column vectors representing the stimulus patterns with the means removed. It is shown in Appendix A2 that the diagonal terms are $\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,q} = Np(1-p)$ and the off-diagonal terms are $\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,r} = -Np^2$, where $p = m/N$ is the frequency of occurrence of an event, and it is assumed that p is the same for all trial types. Without loss of generality, we may set the scalar term $\mathbf{A}_k$ equal to unity so that $\mathbf{X}_\perp^T\mathbf{X}_\perp = \mathbf{E}_Q \otimes \mathbf{A}_k = \mathbf{E}_Q$, where

$$\mathbf{E}_Q = Np\cdot\begin{bmatrix}1-p & -p & \cdots & -p\\ -p & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & -p\\ -p & \cdots & -p & 1-p\end{bmatrix}$$
$$= Np(\mathbf{I}_Q - p\mathbf{1}_Q\mathbf{1}_Q^T) \tag{13}$$

with $\mathbf{1}_Q$ defined as the $Q \times 1$ vector with all elements equal to 1.

For the second limiting case, we assume that there is only one trial type so that $\mathbf{E}_Q$ is now a scalar and the Fisher information

matrix reflects the form of $\mathbf{A}_k$. Eq. (13) then simplifies to yield $\mathbf{E}_Q = Np(1-p)$ and therefore $\mathbf{X}_\perp^T\mathbf{X}_\perp \approx \mathbf{E}_Q \otimes \mathbf{A}_k = Np(1-p)\mathbf{A}_k$. We can infer several properties of $\mathbf{A}_k$ by rewriting the relation as $\boldsymbol{A}_k = 1/(Np(1-p))\boldsymbol{X}_\perp^T\boldsymbol{X}_\perp$. The first diagonal element of $\mathbf{X}_\perp^T\mathbf{X}_\perp$ is the magnitude squared of the detrended stimulus pattern which, as stated above, is equal to $Np(1-p)$. The remaining diagonal elements $\mathbf{X}_\perp^T\mathbf{X}_\perp$ will be less than or equal to $Np(1-p)$ because they are formed from shifted and detrended versions of the stimulus pattern. The properties of the diagonal elements leads us to conclude that $\mathbf{A}_k(1,1) = 1$ and $\mathbf{A}_k(i,i) \le 1$. In addition, $\mathbf{A}_k$ is symmetric and positive definite because of the quadratic form of $\mathbf{X}_\perp^T\mathbf{X}_\perp$.

Having derived the properties of $\mathbf{E}_Q$ and $\mathbf{A}_k$ from the limiting cases, we may write the block diagonal and off-diagonal terms of the Kronecker product as:

$$\mathbf{X}_{\perp,q}^T\mathbf{X}_{\perp,r} \approx \begin{cases} Np(1-p)\mathbf{A}_k & \text{for } q = r \\ -Np^2\mathbf{A}_k & \text{for } q \neq r \end{cases} \tag{14}$$

As an example of the approximation, let us assume that the length of the HRF is $k = 3$, the number of trial types is $Q = 2$, and the average autocorrelation matrix for the two trial types is

$$\mathbf{A}_k = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

where $|\rho| < 1$. The Fisher information matrix is given by:

$$\mathbf{X}_\perp^T\mathbf{X}_\perp \approx Np \times \begin{bmatrix} 1-p & -p \\ -p & 1-p \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

$$= Np \times \begin{bmatrix} 1-p & \rho(1-p) & \rho^2(1-p) & -p & -\rho p & -\rho^2 p \\ \rho(1-p) & 1-p & \rho(1-p) & -\rho p & -p & -\rho p \\ \rho^2(1-p) & \rho(1-p) & 1-p & \rho^2 p & -\rho p & -p \\ -p & -\rho p & -\rho^2 p & 1-p & \rho(1-p) & \rho^2(1-p) \\ -\rho p & -p & -\rho p & \rho(1-p) & 1-p & \rho(1-p) \\ -\rho^2 p & -\rho p & -p & \rho^2(1-p) & \rho(1-p) & 1-p \end{bmatrix}$$

For $\rho = 0$, this reduces to

$$\mathbf{X}_\perp^T\mathbf{X}_\perp = Np \times \begin{bmatrix} 1-p & 0 & 0 & -p & 0 & 0 \\ 0 & 1-p & 0 & 0 & -p & 0 \\ 0 & 0 & 1-p & 0 & 0 & -p \\ -p & 0 & 0 & 1-p & 0 & 0 \\ 0 & -p & 0 & 0 & 1-p & 0 \\ 0 & 0 & -p & 0 & 0 & 1-p \end{bmatrix},$$

which corresponds to the ideal case in which the detrended stimulus pattern for each trial type is uncorrelated with all shifts of itself and is uncorrelated with all non-zero shifts of the stimulus patterns for other trial types.

Relation between efficiency and power

The approximations for the Fisher information matrix stated in Eq. (14) may be used to derive the following expressions for estimation efficiency and detection power from Eqs. (5) and (11):

$$\xi_{\text{tot}} \approx \frac{N \cdot f(p,Q)}{Tr[\mathbf{A}_k^{-1}]} \tag{15}$$

$$R_{\text{tot}} \approx N \cdot f(p,Q)\frac{\mathbf{h}_0^T\mathbf{A}_k\mathbf{h}_0}{\mathbf{h}_0^T\mathbf{h}_0} \tag{16}$$

where

$$f(p,Q) = \frac{pN_Q}{\left(\frac{Q(1-(Q-1)p)}{1-Qp} + Q(Q-1)\right)} \tag{17}$$

Details of the derivation are provided in Appendix A4.

To understand the relation between efficiency and power, it is useful to examine the distribution of the eigenvalues of $\mathbf{A}_k$. As previously discussed in Liu et al. (2001), $Tr[\mathbf{A}_k^{-1}]$ in Eq. (15) is minimized when the eigenvalues of $\mathbf{A}_k$ are equal, whereas $\bar{\mathbf{h}}_0^T\mathbf{A}_k\bar{\mathbf{h}}_0/\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0$ in Eq. (16) is maximized when there is only one non-zero eigenvalue of $\mathbf{A}_k$. To model the distribution of eigenvalues between these two extremes, we adopt the model presented in Liu et al. (2001) in which the maximum eigenvalue is $\lambda_1 = \alpha M$ and the remaining eigenvalues are $\lambda_1 = (1-\alpha)M/(k-1)$, where α ranges from $1/k$ to 1 and $M = Tr[\mathbf{A}_k]$. When $\alpha = 1$, there is only one non-zero eigenvalue, and when $\alpha = 1/k$, the eigenvalues are equally spread. This model results in the approximations

$$\frac{\mathbf{h}_0^T\mathbf{A}_k\mathbf{h}_0}{\mathbf{h}_0^T\mathbf{h}_0} \approx R(\alpha,\theta) = \left(\alpha\cos^2\theta + \frac{1-\alpha}{k-1}\sin^2\theta\right)M \tag{18}$$

$$Tr[\mathbf{A}_k^{-1}] \approx \xi(\alpha) = \frac{\alpha(1-\alpha)M}{1+\alpha(k^2-k)} \tag{19}$$

where θ is the angle between $\mathbf{h}_0$ and the maximum eigenvalue when $\alpha = 1$ (Liu et al., 2001).

The angle θ is used to empirically model the relative detection power of block designs, which have very low estimation efficiency (Liu et al., 2001). Eq. (18) states that when $\alpha = 1$, the detection power is equal to $\cos^2\theta$ times the maximum attainable power, and the estimation efficiency is identically zero. As an example, for a design with one trial type, Liu et al. (2001) found that a block design with 32 blocks (each block consisted of 2 s of stimulus followed by 2 s of rest) had almost zero detection power corresponding to an angle of 90°, while a design with 1 block (64 s of stimulus followed by 64 s of rest) corresponded to an angle of 45° and therefore achieved only half of the maximum attainable power. To model the fact that the maximum detection power may not be achievable with a binary stimulus pattern, it is useful to define $\theta_{\min}$ as the minimum angle achievable between $\bar{\mathbf{h}}_0$ and the dominant eigenvector of $\mathbf{A}_k$, such that the maximum achievable detection power is $\cos^2\theta_{\min}$ times the maximum theoretical value

(Liu et al., 2001). In practice, θ_{min} is estimated by looking at the maximum detection powers of low frequency block designs. When the nuisance functions are limited to a constant term, the maximum is typically achieved by a design with 1 block (see, e.g., Fig. 1 in the Simulations section). For the range of parameters used in this paper, θ_{min} is about 45°.

Substitution of the approximations stated in Eqs. (18) and (19) into Eqs. (15) and (16) yields the following expressions for efficiency and power

$$R_{tot} \approx N \cdot f(p,Q) R(\alpha,\theta) \tag{20}$$

$$\xi_{tot} \approx N \cdot f(p,Q) \xi(\alpha) \tag{21}$$

Because the leading term $N{\cdot}f(p,Q)$ is common to both expressions, the relation between efficiency and power is determined solely by $\xi(\alpha)$ and $R(\alpha,\theta)$. These expressions are identical to those previously presented for experiments with one trial type. Parametric curves demonstrating the fundamental trade-off between $\xi(\alpha)$ and $R(\alpha,\theta)$ for various values of k are shown in Fig. 3 of Liu et al. (2001).

Optimal frequency of occurrence

To maximize either efficiency or power, a necessary condition is to maximize $f(p,Q)$ with respect to p. The maximum occurs at a frequency of occurrence $p = 1/(Q+1)$ and is given by $f(p = 1/(Q+1), Q) = 1/(2(Q+1))$. This is also the frequency of occurrence for which the variance and detection powers for individual events are equal to those for pairwise contrasts, that is, $Tr[\mathbf{C}_{ij}]\,|_{i\neq j} = Tr[\mathbf{C}_{ii}]$ and $R_{ij}\,|_{i\neq j} = R_{ii}$ for $p = 1/(Q+1)$. The fact that $p = 1/(Q+1)$ equalizes efficiencies for both individual event types and contrasts has been previously stated using the concept of null events (Burock et al., 1998; Friston et al., 1999).

In some experiments, it may be of interest to provide different weights for the events and pairwise contrasts. As an example of how this might be done, we add a weight term to $f(p,Q)$ to obtain

$$f(p,Q,k_1) = \frac{pN_Q}{\left(\dfrac{k_1 Q(1-(Q-1)p)}{1-Qp} + (1-k_1)Q(Q-1)\right)} \tag{22}$$

where $0 \le k_1 \le 1$ and $1 - k_1$ are the relative weights for the individual events and pairwise contrasts, respectively. For example, if only individual events are of interest ($k_1 = 1$), the optimal frequency of occurrence is $p = (Q - \sqrt{Q})/(Q^2 - Q)$ for $Q > 1$, whereas if only contrasts are of interest ($k_1 = 0$), the optimal frequency of occurrence is $p = 1/Q$. Note that $(Q - \sqrt{Q})/(Q^2 - Q) < (1/(Q+1) < (1/Q)$ for $Q > 1$. For intermediate values of k_1, the optimal frequency of occurrence will lie in the range $(Q - \sqrt{Q})/(Q^2 - Q) < p < (1/Q)$ and is given by

$$p = \frac{Q(2k_1-1) + Q^2(1-k_1) + k_1^{1/2}(Q(2k_1-1) + Q^2(1-k_1))^{1/2}}{Q(Q-1)(k_1 Q - Q - k_1)} \tag{23}$$

Bounds on efficiency and power

At the overall optimal frequency of occurrence $p = 1/(Q+1)$, the efficiency and power are

$$\xi_{tot} \approx \frac{N}{2(Q+1)} \xi(\alpha) \tag{24}$$

$$R_{tot} \approx \frac{N}{2(Q+1)} R(\alpha,\theta). \tag{25}$$

The bound on estimation efficiency may be derived by setting $\alpha = 1/k$, which is the value for which the eigenvalues are equally spread, and noting that $M \le k$ because $\mathbf{A}_k(i,i) \le 1$. The resultant bound is

$$\xi_{tot} \le \frac{N}{2(Q+1)} \frac{M}{k^2} \le \frac{N}{2(Q+1)k} \tag{26}$$

The bound on detection power may be derived by setting $\alpha = 1$, using the bound $M \le k$, and also using the minimum angle θ_{min} defined below Eq. (19). The resultant bound is

$$R_{tot} \le \frac{NM}{2(Q+1)} \cos^2\theta_{min} \le \frac{Nk}{2(Q+1)} \cos^2\theta_{min} \tag{27}$$

where the $\cos^2\theta_{min}$ term reflects the fact that the maximum detection power achievable in practice may be less than the maximum bound corresponding to $\theta_{min} = 0$. As in discussed in Liu et al. (2001), the bound $M \le k$ is typically not a tight bound because events in the delayed versions of each stimulus pattern are shifted out in the design matrix. An approximation for M is given in the appendix of Liu et al. (2001).

Departures from the predicted trade-off

Basis function expansions

The HRF observed in fMRI experiments exhibits a characteristic shape that reflects the dynamics of the neural and vascular systems. Although the exact shape can show great variability across subjects (Aguirre et al., 1998), the HRF can generally be characterized as a smooth function with a full width half-maximum of about 5–6 s. One means of taking advantage of this a priori knowledge is to confine the HRF to a subspace spanned by a set of smooth basis functions (Dale, 1999). That is, we assume that the HRF for each trial type can be expressed as $\mathbf{h}_q = \mathbf{B}\mathbf{c}_q$, where $\mathbf{B}$ is a $k \times s$ matrix with s basis functions as columns with $s < k$ and $\mathbf{c}_q$ is a $s \times 1$ vector of expansion coefficients. Without loss of generality, the basis functions can be chosen to be orthonormal, that is, $\mathbf{B}^T\mathbf{B} = \mathbf{I}$. The estimate of the HRF vector is given by $\hat{\mathbf{h}} = \tilde{\mathbf{B}}\hat{\mathbf{c}}$, where $\tilde{\mathbf{B}} = \mathbf{I}_Q \otimes \mathbf{B}$ and $\hat{\mathbf{c}} = (\tilde{\mathbf{B}}^T\mathbf{X}_\perp^T\mathbf{X}_\perp\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}^T\mathbf{X}_\perp^T\boldsymbol{\Sigma}^{-1/2}\mathbf{y}$. The covariance of the estimate is $\mathbf{C}_{\hat{\mathbf{h}}} = \tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\mathbf{X}_\perp^T\mathbf{X}_\perp\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}^T$, and estimation efficiency is obtained by using this expression in Eqs. (4) and (5).

The expression for detection power is unaffected by the use of a basis function expansion, since we assume that the base HRF shape $\bar{\mathbf{h}}_0$ lies within the range of $\mathbf{B}$. In other words, each HRF $\mathbf{h}_i$ in the GLM is replaced by $\mu_i\bar{\mathbf{h}}_0 = \mu_i\mathbf{B}\bar{\mathbf{c}}_0$, where $\bar{\mathbf{c}}_0$ is the vector of expansion coefficients for $\bar{\mathbf{h}}_0$, and the detection power is inversely proportional to the estimates of the unknown amplitudes μ_i.

To understand the dependence of estimation efficiency on the choice of basis functions, it is instructive to first examine two limiting cases. In the first limiting case where there is only one basis function equal to the assumed HRF, that is, $\mathbf{B} = \bar{\mathbf{h}}_0$, it is shown in Appendix A5 that estimation efficiency is equal to detection power. This result emphasizes the fact that estimation efficiency and detection power are both measures of statistical efficiency and the distinction between the two measures depends on the a priori assumptions about the shapes of the HRFs. As defined in this paper, detection power always assumes complete knowledge of the shape of the HRF, while estimation efficiency can reflect a range of assumptions as expressed by the choice of basis functions. It is also important to note that detection power is inversely proportional to the average variance of the estimates of the unknown amplitudes of the HRFs while estimation efficiency is inversely proportional to the average variance of the unknown coefficients of the HRFs.

In the second limiting case, when no a priori knowledge about the shape of the HRF is assumed, the basis function matrix is $\mathbf{B} = \mathbf{I}_k$. This is equivalent to assuming that no basis functions are used and therefore estimation efficiency is given by Eq. (5). Thus the fundamental trade-off between estimation efficiency and detection power described in previous sections holds. In addition, from the bounds stated in Eqs. (26) and (27), the maximum estimation efficiency is equal to the maximum detection power scaled by the factor $1/k^2$, that is, $\max(\xi_{\text{tot}}) = (1/k^2)\max(R_{\text{tot}})$.

For intermediate cases, where the dimension s of the subspace spanned by the basis functions lies between 1 and k, it is shown in Appendix A6 that the maximum achievable efficiency is k^2/s^2 greater than that obtained when no a priori information is assumed, or equivalently, $\max(\xi_{\text{tot}}) = (1/s)^2\max(R_{\text{tot}})$. The gain in efficiency is due to two factors. A factor of k/s results from the fact that fewer parameters are estimated when using basis functions. A second factor of k/s reflects the gain that is possible with a design that amplifies only those signals that lie within the basis function subspace. Whether the second gain factor is attainable in practice is an open question, but in general, we expect the gain to be less than the maximum theoretical gain. Examples of the gains in efficiency when using basis functions are provided in the Simulations section.

Correlated noise

In developing the theoretical approximations for efficiency and power, we have assumed that the noise is uncorrelated. Although a detailed treatment of the effects of correlated noise is beyond the scope of this paper, it is possible to make some preliminary observations. First, the Kronecker product approximation for the Fisher information matrix implies that the effect of correlated noise on multiple trial type experiments is fairly well reflected by its effect on a single-trial-type experiment. For a single-trial-type experiment, the estimation efficiency is $\xi_{\text{tot}} = 1/Tr[\mathbf{J}^{-1}]$, where the Fisher information matrix is $\mathbf{J} = \mathbf{X}_{\perp}^T\mathbf{X}_{\perp} = \mathbf{X}^T(\boldsymbol{\Sigma}^{-1/2})^T\mathbf{P}_{\mathbf{S}}^{\perp}\boldsymbol{\Sigma}^{-1/2}\mathbf{X}$. We have found that that the approximation $\mathbf{J} \approx \mathbf{X}^T\mathbf{P}_{\mathbf{S}}^{\perp}\boldsymbol{\Sigma}^{-1}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{X}$ works fairly well in practice. The approximation reflects the fact that for parameters typically found in fMRI experiments, the noise whitening step and the removal of low frequency nuisance terms are, to first order, interchangeable filtering operations. Efficiency is maximized when the eigenvalues of $\mathbf{J}$ are equally distributed. For a given noise covariance matrix $\boldsymbol{\Sigma}$, this can achieved by starting with a design matrix $\mathbf{X}$ that achieves maximal efficiency when the noise is uncorrelated, so that $\mathbf{X}^T\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{X}$ has equal eigenvalues. This corresponds to a stimulus pattern with a flat power spectrum (Haykin, 1996). The next step is to multiply the design by a coloring matrix $\boldsymbol{\Sigma}^{1/2}$ so that the information matrix is $\mathbf{X}^T(\boldsymbol{\Sigma}^{1/2})^T\mathbf{P}_{\mathbf{S}}^{\perp}\boldsymbol{\Sigma}^{-1}\mathbf{P}_{\mathbf{S}}^{\perp}\boldsymbol{\Sigma}^{1/2}\mathbf{X} \approx \mathbf{X}^T\mathbf{P}_{\mathbf{S}}^{\perp}(\boldsymbol{\Sigma}^{1/2})^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{1/2}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{X} = \mathbf{X}^T\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{X}$, where we have used the approximate interchangeability of the coloring and nuisance term removal steps. This corresponds to a stimulus pattern with a colored power spectrum with the same spectral shape as the correlated noise. Thus, efficiency is maximized by increasing the spectral power of the stimulus pattern at frequencies where the noise power is high and decreasing the spectral power at frequencies where the noise power is low. In the temporal domain, this implies that the efficiency is maximized by a stimulus pattern with a correlation structure that is similar to that of the noise.

The detection power for an experiment with one trial type is given by the Rayleigh quotient $(\bar{\mathbf{h}}_0^T\mathbf{J}\bar{\mathbf{h}}_0)/(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0) \approx (\bar{\mathbf{h}}_0^T\mathbf{X}^T\mathbf{P}_{\mathbf{S}}^{\perp}\boldsymbol{\Sigma}^{-1}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{X}\bar{\mathbf{h}}_0)/(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)$. The numerator has the form of a weighted inner product (Kailath et al., 2000), and so to preserve the properties of the Rayleigh quotient, it is reasonable to replace the denominator with a weighted inner product $\bar{\mathbf{h}}_0^T(\boldsymbol{\Sigma}^{-1})_k\bar{\mathbf{h}}_0$, where $(\boldsymbol{\Sigma}^{-1})_k$ is the central $k \times k$ submatrix of $\boldsymbol{\Sigma}^{-1}$. In other words, multiplication of the detection power by a factor of $(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)/(\bar{\mathbf{h}}_0^T(\boldsymbol{\Sigma}^{-1})_k\bar{\mathbf{h}}_0)$ will make the metric approximately invariant with respect to $\boldsymbol{\Sigma}$. This in turn implies that the detection power will be reduced by a factor of $\bar{\mathbf{h}}_0^T(\boldsymbol{\Sigma}^{-1})_k\bar{\mathbf{h}}_0/\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0$. The physical interpretation is that the detection power is to first order reduced by the amount to which the whitening process removes energy from the HRF. As shown in the Simulations section, this approximation works fairly well in practice.

Conditional entropy

In addition to estimation efficiency and detection power, the perceived randomness of an experimental design is an important factor to consider when designing an experiment. Having a sufficient degree of randomness can be critical for circumventing experimental confounds such as a habituation and anticipation (Rosen et al., 1998). Controlling the randomness of a design can also be useful for determining which brain regions are most sensitive to changes in randomness (Bischoff-Grethe et al., 2001). Various metrics for randomness or its converse, predictability, have been presented in the fMRI literature. For experiments with one trial type, Liu et al. (2001) defined predictability as the average probability of correctly guessing the next event in a design using a binary string prediction program based on conditional probabilities. Wager and Nichols (2003) introduced a counterbalancing score which measured the deviation of the conditional probabilities of a design with an arbitrary number of trial types from the conditional probabilities of a perfectly random design. Bischoff-Grethe et al. (2001) used mutual information as a measure of predictability for an experimental design with four trial types. These metrics are all related since they are based upon the conditional probabilities associated with a design. In this paper, we use conditional entropy as a measure of the perceived randomness of a design. As with the previous metrics, it is based upon conditional probabilities, and has the additional advantage of being a fundamental and commonly used metric for randomness (Cover and Thomas, 1991). The rth order conditional entropy H_r is a measure of the average uncertainty in the trial type of the next event given knowledge of the trial types of the r previous events. For a stimulus pattern of the form $a_1,a_2,\cdots,a_N$, where the trial type a_1 for the ith event is

drawn from an alphabet A, the expression for conditional entropy is

$$H_r = -\sum_{a_i \in A} \sum_{a_{i-1} \in A} \cdots \sum_{a_{i-r} \in A} p(a_{i-r}, \ldots, a_{i-1}, a_i) \times \log_2 p(a_i \mid a_{i-r}, \ldots, a_{i-1}) \quad (28)$$

where $p(a_{i-r},\ldots,a_{i-1},a_i)$ is the probability of the sequence $a_{i-r},\ldots,a_{i-1},a_i$ of $(r+1)$ trial types occurring and $p(a_i|a_{i-r},\cdots,a_{i-1})$ is the conditional probability of trial type a_i occurring after the sequence of r trial types $a_{i-r},\ldots,a_{i-1}$. For example, in an experiment with $Q=4$ trial types, the alphabet can be defined as $A=\{0,1,2,3,4\}$, where 0 denotes the null trial type, $p(0,3,4)$ is the probability of observing a sequence consisting of a null trial type followed by trial types 3 and 4, and $p(0|3,4)$ is the conditional probability of observing a null trial type after observing the sequence consisting of trial type 3 followed by trial type 4. Note that the order of the trial types matters, so that in general, $p(0|3,4)$ does not equal $p(0|4,3)$. Conditional entropy is measured in *bits* and corresponds to the average number of binary (e.g., yes/no) questions required to determine the next trial type given knowledge of the r previous trial types (Cover and Thomas, 1991).

Conditional entropy is related to the mutual information metric used in Bischoff-Grethe et al. (2001) by the expression $I_r = \log_2(Q+1) - H_r$, where we have assumed that the probability for each trial type is $p = 1/(Q+1)$ and the mutual information I_r has been generalized to include r previous events. The predictability metric used in Liu et al. (2001) is related through the expression $P_r = 1/2^{H_r}$. For example, a completely random sequence used in a 1 trial type experiment has a predictability of 0.5 and a conditional entropy of 1 bit. The counterbalancing metric of Wager and Nichols (2003) uses a sum of squared differences of conditional probabilities. It increases with entropy, but determining an exact relation between the two metrics is beyond the scope of this paper. Finally, it is convenient to use the quantity 2^{H_r} when comparing different designs. This is the inverse of predictability and is in essence a linear measure of randomness. Thus, a design with a conditional entropy of 2 bits is twice as random as a design with an entropy of 1 bit, since there are on average twice as many random outcomes.

It was shown in a previous section that the relation between estimation efficiency and detection power could be modeled by examining the spread of the eigenvalues of the matrix $\mathbf{A}_k$. In the presence of uncorrelated additive noise and a constant nuisance term, $\mathbf{A}_k$ is the average of the autocovariance matrices of the stimulus patterns. For completely random stimulus patterns, $\mathbf{A}_k$ is proportional to the identity matrix and therefore has an equal spread of eigenvalues. As the patterns become less random, the spread of the eigenvalues increases (Haykin, 1996). On the basis of this qualitative description of the eigenvalue spread, we expect that estimation efficiency should increase with entropy. For experiments with one trial type, simulations have been used to show that efficiency does increase with entropy (Liu et al., 2001). The relation between efficiency and entropy can be further motivated by noting that, for Gaussian sequences, the conditional entropy is related to the determinant of the autocovariance matrix, which is equal to the product of the eigenvalues (Cover and Thomas, 1991). As pointed out in Liu et al. (2001), the trace of the autocovariance matrix is approximately constant across experimental designs. With the trace of the autocovariance matrix held constant, the product of the eigenvalues is maximized when the eigenvalues are equally spread. As a simple example, consider a matrix with two equal eigenvalues, both with value λ_0. An increase in one eigenvalue by some amount Δ is balanced by a decrease in the other eigenvalue by Δ. The product of the two unequal eigenvalues is less than the product of the equal eigenvalues, since $(\lambda_0 - \Delta)(\lambda_0 + \Delta) \leq \lambda_0^2$. Thus, for Gaussian sequences, an equal spread of eigenvalues yields maximum entropy. Although the stimulus patterns in this paper consist of a discrete number of trial types and are not Gaussian sequences, the theoretical insight gained from the Gaussian case provides useful support for the qualitative description.

Simulations

To evaluate the performance of the proposed model and gain further insight into the relation between estimation efficiency, detection power, and conditional entropy, a series of numerical simulations was performed. Estimation efficiencies and detection powers were calculated using Eqs. (5) and (11) for a GLM with $k=15$, $N=240$, Q ranging from 2 to 5, and $p = 1/(Q+1)$. The conditional entropy was computed using Eq. (28). Semi-random stimulus patterns were obtained by permuting block designs, which were of the form ABCNABCN, where A, B, and C represent the blocks for each trial type and N represents the blocks for the control state. We used a range of block designs with 1 to $n(Q)$ equally spaced blocks per trial type, where $n(Q)$ is the maximum number of blocks and ranged from $n(Q)=20$ for $Q=5$ to $n(Q)=40$ for $Q=2$. For example, for $Q=3$, block designs with 1, 2, 4, 10, 15, and 30 equally spaced blocks were used. The 1-block design was of the form ABCN, where each block contained 60 events, while the 2-block design was of the form ABCNABCN, where each block contained 30 events. There were a total of 200 permutation steps, where at each step the positions of two randomly chosen events were exchanged. A figure illustrating the permutation process is provided in Liu (2004).

For calculation of detection power, $\bar{\mathbf{h}}_0$ was a gamma density function of the form $h[j] = (\tau n!)^{-1}((j-1)\Delta t/\tau)^n e^{-j\Delta t/\tau}$ for $j \geq 1$ and 0 otherwise (Boynton et al., 1996) with $\Delta t = 1$, $\tau = 1.2$, and $n = 3$. In addition, for each value of Q, 1000 random patterns with a uniform distribution of trials types were generated and the pattern with the greatest estimation efficiency was selected for display.

Designs based upon prime m-sequences have been previously shown to achieve much higher estimation efficiency than designs obtained via a random search, especially as the number of trial types increases (Buracas and Boynton, 2002). The gain in efficiency reflects the fact that the m-sequences exhibit almost ideal autocorrelation properties, that is, they are nearly orthogonal to time-shifted versions of themselves. The designs presented in Buracas and Boynton (2002) were based on prime m-sequences for 2, 3, and 5 levels, and the addition of a 2-level m-sequence to a shifted version of itself to obtain a sequence with 4 levels. An alternative for a 4-level sequence is to use a power of primes m-sequence (e.g., $2^2 = 4$) (Godfrey, 1993). We used the power of prime sequence because its conditional entropy was slightly higher than that obtained by summing shifted m-sequences. Designs for $Q = 2$, 3, and 4 were generated using 3, 4, and 5 level m-sequences, respectively, by using the m-sequence to assign the trial type for each event in the design. For example, for a 3-level m-sequence of the form $\cdots$,0,1,2,1,$\cdots$, a 0 corresponds to a null or control event,

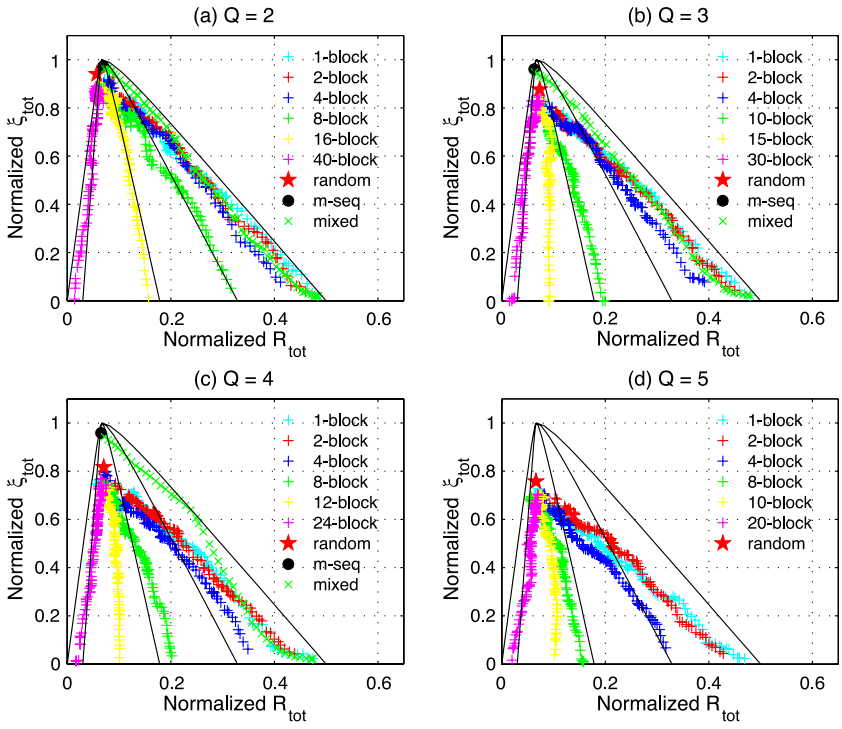

Fig. 1. Simulation results for $Q = 2$ to 5 showing estimation efficiency versus detection power. The nuisance function matrix consists of a constant term and the noise is assumed to be uncorrelated. Theoretical curves (solid lines) for θ equal to 45°, 55°, 65°, 80°, and 90° (progressing from right to left) are also shown. For the permuted block designs, each permutation path corresponds to one realization of the random permutation process, and the '+' symbols represent designs along that path. Each permutation path begins with a block design that exhibits nearly zero estimation efficiency.

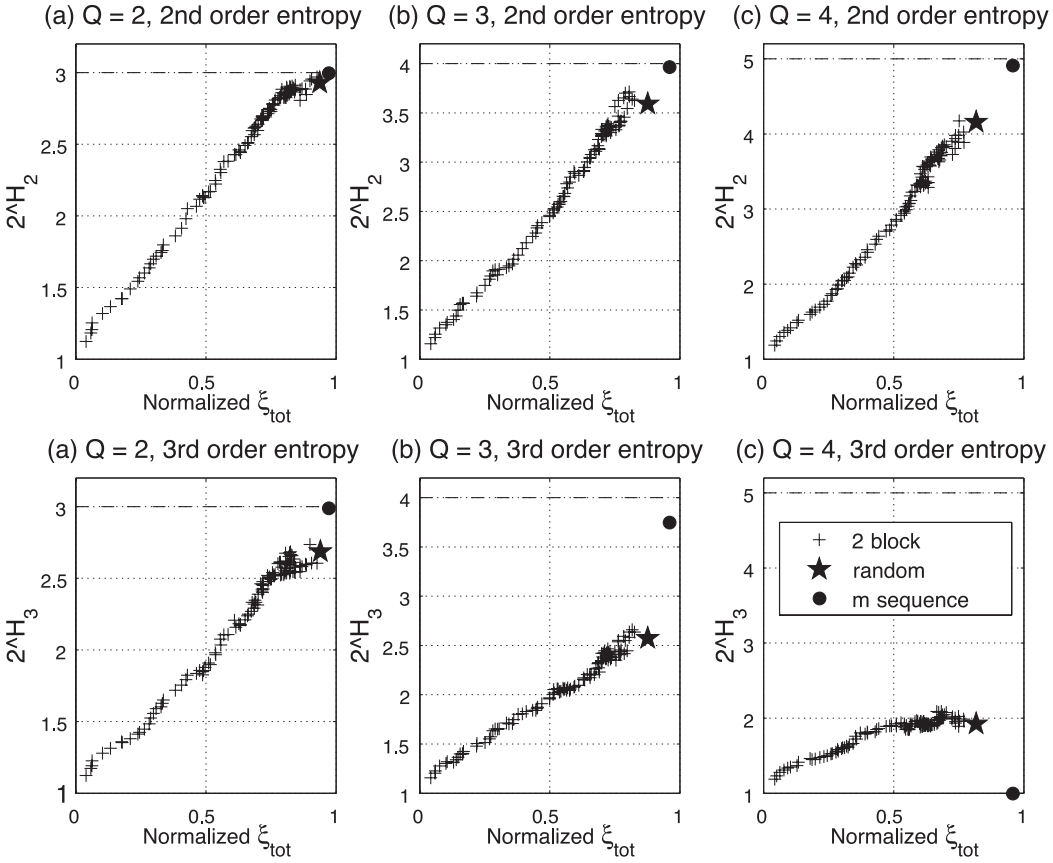

Fig. 2. Relation between second and third order conditional entropy and estimation efficiency for $Q = 2$, 3, and 4. The vertical axis is in units of $2\hat{}H_r$, where H_r is the rth order conditional entropy. The upper bound on $2\hat{}H_r$, corresponding to the maximum conditional entropy for a sequence with $(Q + 1)$ discrete values, is shown by the dashed line.

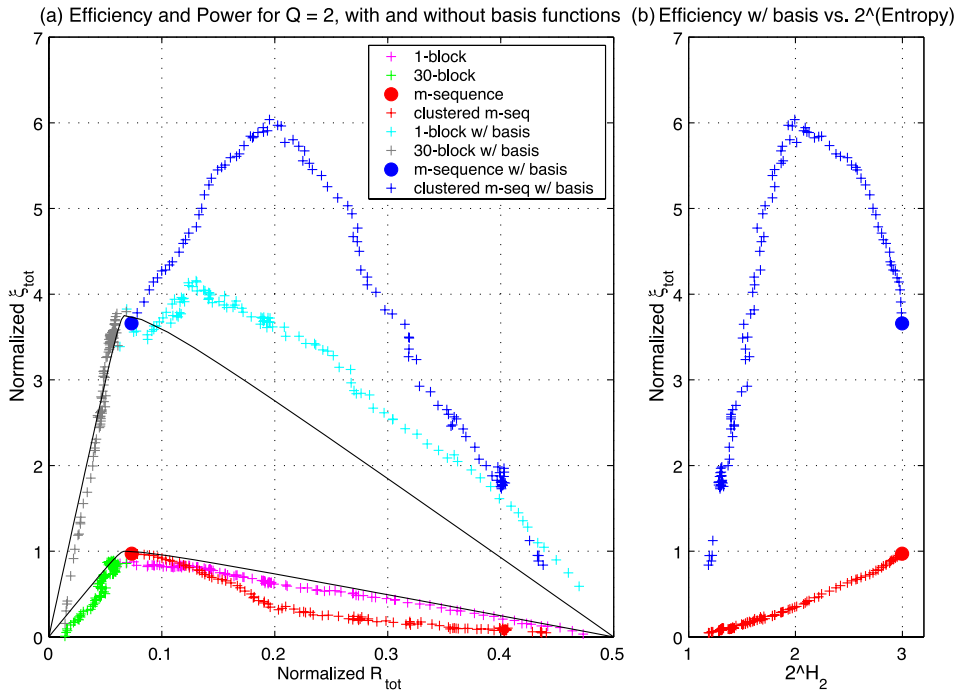

Fig. 3. (a) Estimation efficiency versus detection power for $Q = 2$ trial types showing the effect of using basis functions on efficiency. Theoretical curves (solid lines) for θ equal to 45° and 90° are also shown. (b) Estimation efficiency versus $2\hat{}H_2$ for clustered m-sequences using basis functions, where H_2 is the second order conditional entropy. For the permuted block and clustered m-sequence designs, each path corresponds to one realization of the random permutation or clustering process, and the '+' symbols represent designs along that path. Each block permutation path begins with a block design that exhibits nearly zero estimation efficiency, while the clustering m-sequence path begins with the m-sequence design.

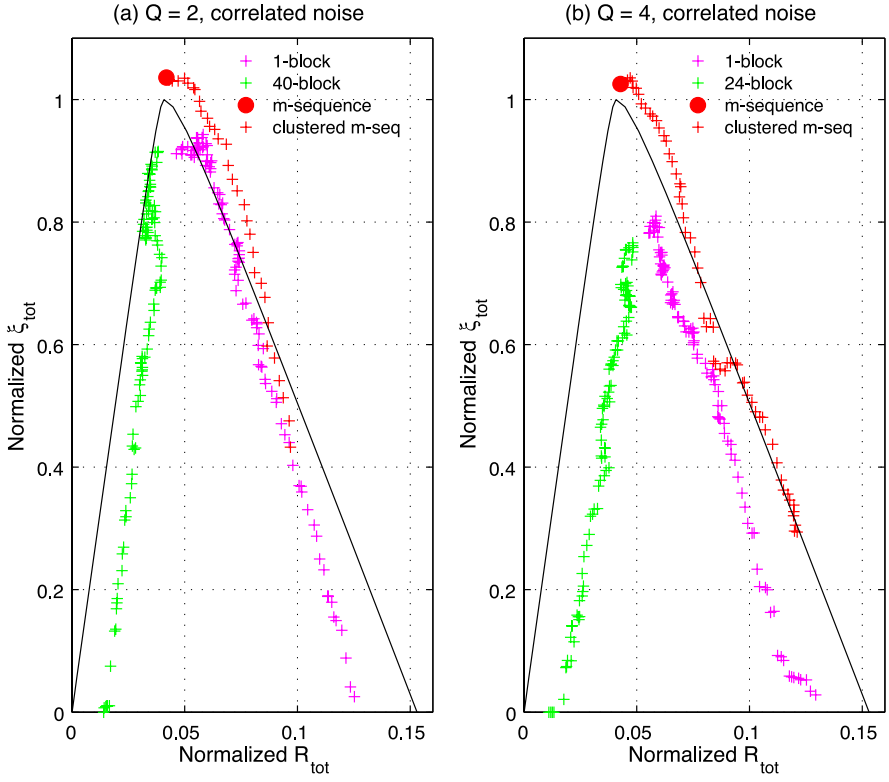

Fig. 4. Effect of correlated noise for $Q = 2$ and 4 trial types. Theoretical curves (solid lines) for θ equal to 60° and 90° are also shown.

a 1 corresponds to trial type 1, and a 2 corresponds to trial type 2. For $Q = 2$ and 3, the lengths of the 3- and 4-level m-sequences were 242 and 255, respectively, and these sequences were truncated to $N = 240$ points. For $Q = 4$, the 5-level 124 point m-sequence was repeated and then truncated to 240 points. We are not aware of a m-sequence-based design for $Q = 5$. Additional examples of m-sequence-based designs are provided in Liu (2004).

To take advantage of the high detection power of a block design and the high estimation efficiency of a m-sequence design, a mixed design can be formed from the concatenation of a block design and a m-sequence design (Liu et al., 2001). For $Q = 2$, 3, and 4 trial types, mixed designs were formed by concatenating a 1-block design of length $N - L$ with a m-sequence design truncated to length L, where L was varied from $Q + 1$ to N in increments of $Q + 1$ with $N = 240$. A figure showing mixed design examples is presented in Liu (2004).

Fig. 1 shows the paths of estimation efficiency versus detection power for m-sequence designs, randomly generated designs, mixed designs, and various permuted block designs for $Q = 2$ to 5. For the permuted block designs, the designs shown correspond to one realization of the random permutation process, and the starting points of the permutation paths are block designs, which exhibit nearly zero estimation efficiency. Both estimation efficiency and detection power are normalized by the maximum bounds presented in Eqs. (26) and (27) with $\theta_{\min} = 0°$. The dimension of the nuisance function matrix was $l = 1$, corresponding to a constant term for the nuisance function, and the noise was assumed to be uncorrelated. Parametric curves showing the relation between $\xi(\alpha)$ and $R(\alpha,\theta)$ as function of α with θ equal to 45°, 55°, 65°, 80°, and 90°are also plotted. For $Q = 2$, the paths taken by the permuted block designs are well modeled by the theoretical curves. The agreement between the theoretical curves and the permuted block designs decreases as the number of trial types Q increases. In addition, the estimation efficiency of the randomly generated designs decreases with respect to the theoretical bound as Q increases. By contrast, the performance of the m-sequence-based designs is very close to the theoretical bound. For $Q = 3$ and 4 , the mixed designs come significantly closer to the theoretical bound than the permuted block designs and thus lend support to the validity of the theoretical model.

Entropy

The relation between conditional entropy and estimation efficiency for $Q = 2$, 3, and 4 is shown in Fig. 2 for m-sequence-based designs, randomly generated designs, and permuted 2-block designs. For the permuted 2-block designs, the designs shown correspond to one realization of the random permutation process. The second order conditional entropies increase with estimation efficiency and exhibit the approximate relation

$$H_r \approx \log_2(Q\xi_{\text{tot,norm}} + 1) \tag{29}$$

where $\xi_{\text{tot,norm}}$ is the estimation efficiency normalized by its upper bound. The entropies for the m-sequence-based designs approach the upper bound of $\log_2(Q + 1)$ bits, which corresponds to a completely random sequence. Although not shown, the first order conditional entropies show similar behavior. The third order conditional entropies for $Q = 2$ also follow the approximate relation stated in Eq. (29), but the behavior for $Q = 3$ and 4 is more complicated. The permuted block designs increase with efficiency but the slope of the relation is less than that predicted by Eq. (29), and the maximum entropy achieved is comparable to that obtained by the randomly generated sequence with the highest efficiency. This behavior reflects the difficulty in obtaining sequences that achieve the upper bound on conditional entropy using random search techniques. For $Q = 3$, the m-sequence design is close to the upper bound, but for $Q = 4$, the third order conditional entropy for the m-sequence design is identically zero, indicating that this sequence is completely predictable. The m-sequence for $Q = 4$ is based on a m-sequence corresponding to a three-stage shift register with pentary logic (Buracas and Boynton, 2002). Thus, knowledge of the three previous trial types completely determines the subsequent trial type, and the third order conditional entropy is zero. In practice, however, the perceived entropy of the sequence is likely to be higher than the theoretical entropy. The theoretical entropy assumes perfect memory of the frequency with which each trial type occurred after any one of $5^3 = 125$ combinations of three previous trial types. A subject in an fMRI experiment is unlikely to remember this many combinations. In general, we expect that the appropriate order to use for the conditional entropy metric probably decreases with the number of trial types, and that the first or second order conditional entropy is probably a sufficient metric for most studies with more than two trial types.

Basis functions

The effect of basis function expansions was assessed using the basis set described in Friston et al. (1998). This basis set consists of two gamma density functions and their temporal derivatives, where the parameters of the gamma density functions are $\Delta t = 1$, $\tau = 1$, and $n = 3$ and 7. The basis set was orthogonalized using a singular value decomposition so that $\mathbf{B}^T\mathbf{B} = \mathbf{I}$. Fig. 3a shows the estimation efficiency versus detection power for $Q = 2$ trial types using permuted block designs (1 and 30 block) and m-sequence-based designs. In addition, a new class of designs called *clustered m-sequences* was also used. These designs start with the nearly optimal estimation efficiency achieved by m-sequences and attempt to gradually reduce estimation efficiency while increasing detection power. They are generated by randomly permuting a m-sequence such that at each permutation step, the sequence becomes more "blocky". Details of the algorithm are provided in Liu (2004). For both the permuted block and clustered m-sequence designs, the designs shown correspond to one realization of the permutation or clustering process, respectively. For permuted block sequences, the permutation process begins with block designs that exhibit minimum estimation efficiency, while for the clustered m-sequences, the clustering process begins with the m-sequence design. The estimation efficiency for the m-sequence-based design with basis functions is increased by a factor of approximately $k/s = 15/4 = 3.75$ with respect to the efficiency without basis functions. This is in agreement with the expected gain due to the decrease in the number of parameters. Clustered m-sequences offer significant increases in estimation efficiency, achieving up to 1.9 times greater efficiency than the m-sequence design, as compared to the maximal theoretical gain of 3.75. At the point of maximal efficiency, the clustered m-sequence design also achieves more than double the detection power of the m-sequence. The entropy is significantly decreased, however, with the maximal efficiency design resulting in a 33% decrease in randomness with respect to the m-sequence.

Correlated Noise

For simulations with correlated noise, we adopted the noise covariance model presented in Burock and Dale (2000) which has the form $R_N[n] = \sigma^2(\lambda\delta[n] + (1 - \lambda)\rho^{|n|})$ and parameters $\rho = 0.88$ and $\lambda = 0.75$. The effect of correlated noise is shown in Fig. 4 for experiments with two and four trial types. The maximum estimation efficiency is higher in the presence of correlated noise as compared to the uncorrelated noise case and in fact exceeds the theoretical bound. This effect has been previously reported in Birn et al. (2002) and Buracas and Boynton (2002). Note, however, that this effect assumes that the total noise variance is constant, such that the correlated noise has more energy at low frequencies and less at higher frequencies as compared to the uncorrelated noise case. Another noise normalization approach, such as matching the variance of the high frequency components, would result in higher total noise variance and hence a reduction in the maximum estimation efficiency in the presence of correlated noise. In the presence of correlated noise, clustered m-sequence designs provide higher detection power and equal or greater estimation efficiency than the m-sequence-based designs. The detection power of all designs is greatly reduced in the presence of correlated noise, as can be seen by comparing the horizontal axes in Fig. 4 versus those in Fig. 1. The reduction factor is approximately equal to the term $\tilde{\mathbf{h}}_0^T(\mathbf{\Sigma}^{-1})_k\tilde{\mathbf{h}}_0/(\tilde{\mathbf{h}}_0^T\tilde{\mathbf{h}}_0)$ presented in the Theory section. For comparison, the theoretical curves have also been scaled by this factor.

Discussion

Statistical efficiency is a critical factor to consider in the design of fMRI experiments because it has a direct impact on the quality of the acquired data. Two important metrics of statistical efficiency are detection power and estimation efficiency. Detection power is important in experiments that focus on mapping the loci of activation or that attempt to compare the magnitudes of activation between different brain regions, tasks, or populations. As stated in the Theory section, detection power is equivalent to the estimation efficiency for response amplitudes. In the context of this paper, estimation efficiency refers to the ability to estimate the shapes of the HRFs and is important for experiments that try to assess the variation of the shape across brain regions, tasks, or patient populations. A third metric is the conditional entropy of a design, which is a measure of perceived randomness. Maximizing the entropy can be critical for minimizing confounds such as habituation and anticipation. An empirical relation between conditional entropy and estimation efficiency was presented. The combination of this empirical relation with the theoretical model for the relation between estimation efficiency and detection power offers a comprehensive framework for understanding the trade-off between the three metrics. The optimum balance depends on the goals and constraints of the experiment. For example, an experiment may simultaneously require a high degree of entropy to minimize confounds and sufficient detection power to compare response amplitudes, while having no requirement on desired estimation efficiency.

The theoretical model presented in this paper provides insight into the relation between detection power and estimation efficiency for experiments with multiple trial types. In particular, it was shown that the fundamental trade-off between efficiency and power has the same form as that previously demonstrated for experiments with a single trial type (Liu et al., 2001). The model shows good agreement with numerical simulations for experiments with two trial types. For experiments with three and four trial types, there is greater deviation between the theoretically predicted performance and that obtained by permuted block designs, but mixed designs come reasonably close to the predicted trade-off in regions where the permuted block designs do not perform well. One possible explanation for the observed discrepancy is that the model makes a number of approximations that become less valid as the number of trial types increases. A second explanation is that the predicted theoretical performance becomes harder to achieve with an increasing number of trial types. The second explanation is supported by the observation that m-sequence designs come very close to achieving the bound on estimation efficiency. This contrasts sharply with the decrease in the efficiency of randomly generated designs as the number of trial types increases. As pointed out by Buracas and Boynton (2002), the probability of attaining the estimation efficiency of a m-sequence design via a random search of designs is very low. This probability diminishes rapidly with the number of trial types since the number of possible designs is $(Q + 1)^N$. The second explanation is also supported by the observation that increasing the search space for both permuted block and clustered m-sequence designs yields designs that come closer to the theoretical bounds (Liu, 2004). Other approaches, such as genetic algorithms (Wager and Nichols, 2003), may also be useful in coming closer to the bounds.

The incorporation of basis function expansions was shown to lead to a generalized definition of estimation efficiency that includes the definitions of estimation efficiency and detection power presented in Eqs. (5) and (11) as specific cases. Thus, detection power is equal to the generalized estimation efficiency when there is only one basis function, and estimation efficiency is equal to the generalized estimation efficiency when the set of basis functions consists of k Kronecker delta functions. The upper bound on estimation efficiency was shown to be proportional to the ratio $1/s^2$, where s is the number of basis functions used. In practice, the gain in estimation efficiency obtained by going to a smaller set of basis functions is less than the expected theoretical gain. A theoretical model that better explains the observed gain in efficiency is a topic for future work. In addition, the selection of an optimal set of basis functions that spans the space of all possible hemodynamic responses with the fewest number of dimensions is an area for further investigation.

With either basis function expansions or the assumption of correlated noise, the maximum estimation efficiency is not achieved by m-sequence-based designs, but by designs that take advantage of either the a priori assumptions about the HRF or the structure of the noise. In addition, these designs also have higher detection power since they are less random. As an example, a new class of designs based on clustered m-sequences was found to offer both higher estimation efficiency and detection power than m-sequence designs in the presence of basis function expansions or correlated noise. The construction of these designs is described in greater detail in Liu (2004).

The analysis in this paper has focused on non-overlapping designs in which the stimuli from different trial types do not overlap, although the responses from the stimuli may overlap due to the long temporal duration of the HRFs. Such designs are widely used in cognitive experiments, for example, Clark et al. (1998). Overlapping designs similar to those used for receptive field mapping in vision science have been also been proposed (Buracas and Boynton, 2002). In these designs, there is no constraint on the

overlap of stimuli from different trial types. Semi-overlapping designs in which there can be some overlap between stimuli from different trial types may also be useful, especially for modeling experiments in which cognitive trial types overlap in some restricted fashion. For example, in a decision-making experiment, it may be useful to explicitly model an overlap occurring between the end of the information assessment stage and the beginning of the decision-making stage. The performance of overlapping and semi-overlapping designs is an area for further investigation.

In addition, the analysis presented here has focused on binary designs, where each trial type may be either present or absent. In certain experiments, such as those focused on the visual system, the stimulus for each trial type may take on multiple discrete levels or a range of continuous levels. An extension of the current framework to handle designs with both multiple trial types and levels would be useful for these experiments.

Finally, our analysis has assumed that the neuronal and hemodynamic pathway from the stimulus to the measured response is well modeled by a linear time-invariant system, so that the measured response is the convolution of the stimulus with the HRF. There is growing evidence that nonlinear models, such as Volterra kernel expansions (Friston et al., 1998; Liu et al., 2002), provide a better representation and an extension of the current framework to incorporate nonlinear effects would be useful.

Acknowledgments

This work was supported in part by a Biomedical Engineering Research Grant from the Whitaker Foundation and by Merit Review SA321 from the Veterans Administration. We thank Giedrius Buracas for helpful discussions regarding m-sequences.

Appendix A

A.1. Kronecker product definitions and identities

The Kronecker product is defined as (Brewer, 1978; Moon and Stirling, 2000):

$$\mathbf{A}\otimes\mathbf{B}\equiv\begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1q}\mathbf{B} \\ a_{21}\mathbf{B} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{p1}\mathbf{B} & \cdots & \cdots & a_{pq}\mathbf{B} \end{bmatrix} \tag{A1}$$

Useful identities are:

$$(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1} \tag{A2}$$

$$\text{trace}[\mathbf{A}\otimes\mathbf{B}]=\text{trace}[\mathbf{A}]\text{trace}[\mathbf{B}] \tag{A3}$$

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=\mathbf{AC}\otimes\mathbf{BD} \tag{A4}$$

$$(\mathbf{A}\otimes\mathbf{B})^{T}=\mathbf{A}^{T}\otimes\mathbf{B}^{T} \tag{A5}$$

A.2. Derivation of $\boldsymbol{X}_{\perp,q}^{T}\boldsymbol{X}_{\perp,r}$

When $q = r$, the product $\mathbf{X}_{\perp,q}^{T}\mathbf{X}_{\perp,r}$ is the variance of a binary stimulus pattern with m events out of N total time points. This variance is $(1 - m/N)m$. When $q \neq r$, $\mathbf{X}_{\perp,q}^{T}\mathbf{X}_{\perp,r}$ is the dot product, after removal of constant terms, between two stimulus patterns, each with m events, where the events of the two patterns do not overlap. The dot product consists of $2m$ terms of the form $(-m/N)(1 - m/N)$ and $N - 2m$ terms of the form m^2/N^2 yielding the result $-m^2/N$.

A.3. Inverse of $\boldsymbol{E}_{Q}$

The inverse of $\mathbf{E}_{Q}$ can be derived with the matrix inversion lemma (Moon and Stirling, 2000)

$$(\mathbf{A}+\mathbf{BCD})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1}+\mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

to obtain

$$\begin{aligned}\boldsymbol{E}_{Q}^{-1}&=\frac{1}{pN(1-Qp)}\begin{bmatrix} 1-(Q-1)p & p & \cdots & p \\ p & & \ddots & \vdots \\ \vdots & & \ddots & p \\ p & \cdots & p & 1-(Q-1)p \end{bmatrix} \\ &=\frac{1}{pN(1-Qp)}\left((1-Q)\mathbf{I}_{Q}+p\mathbf{1}_{Q}\mathbf{1}_{Q}^{T}\right)\end{aligned}$$

A.4. Derivation of expressions for efficiency and power

The expression for estimation efficiency contains $Tr[\mathbf{C}_{ij}]$ terms that, with the use of Kronecker product identities, may be expanded as

$$\begin{aligned}Tr[\mathbf{C}_{ij}]&=Tr[\mathbf{L}_{ij}(\mathbf{X}_{\perp}^{T}\mathbf{X}_{\perp})^{-1}\mathbf{L}_{ij}^{T}] \\ &\approx Tr[(\mathbf{D}_{ij}\otimes\mathbf{I}_{k})(\mathbf{E}_{Q}^{-1}\otimes\mathbf{A}_{k}^{-1})(\mathbf{D}_{ij}^{T}\otimes\mathbf{I}_{k})] \\ &=Tr[(\mathbf{D}_{ij}\mathbf{E}_{Q}^{-1}\mathbf{D}_{ij}^{T})\otimes\mathbf{A}_{k}^{-1}] \\ &=Tr[\mathbf{D}_{ij}\mathbf{E}_{Q}^{-1}\mathbf{D}_{ij}^{T}]Tr[\mathbf{A}_{k}^{-1}]\end{aligned}$$

where the Kronecker product approximation for the Fisher information matrix is used in the second line. Substitution of the expression for $\mathbf{E}_{Q}^{-1}$ from Appendix A3 yields

$$Tr[\mathbf{C}_{ij}]=\begin{cases}\dfrac{1-(Q-1)p}{Np(1-Qp)}Tr[\mathbf{A}_{k}^{-1}] & \text{for } i=j \\ \dfrac{2}{Np}Tr[\mathbf{A}_{k}^{-1}] & \text{for } i\neq j\end{cases}$$

Substitution of the $Tr[\mathbf{C}_{ij}]$ terms into Eq. (5) leads to Eq. (15).

The expressions for detection power (Eqs. (10) and (11)) contain the term $(\mathbf{Z}_{\perp}^{T}\mathbf{Z}_{\perp})^{-1}$.

With the assumption that $\bar{\mathbf{h}}_{i}=\bar{\mathbf{h}}_{0}$, we may write $\mathbf{Z}_{\perp}=[\mathbf{X}_{\perp,1}\bar{\mathbf{h}}_{1}\ \mathbf{X}_{\perp,2}\bar{\mathbf{h}}_{2}\cdots\mathbf{X}_{\perp,Q}\bar{\mathbf{h}}_{Q}]=\mathbf{X}_{\perp}(\mathbf{I}_{Q}\otimes\bar{\mathbf{h}}_{0})$.

Applying the Kronecker product approximation for the Fisher information matrix, and identities (A4) and (A5), we find that

$$
\begin{aligned}
R_{ij} &= (\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T)^{-1}(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)^{-1} \\
&= (\mathbf{D}_{ij}((\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0)^T\mathbf{X}_\perp^T\mathbf{X}_\perp(\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0))^{-1}\mathbf{D}_{ij}^T)^{-1}(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)^{-1} \\
&\approx (\mathbf{D}_{ij}((\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0)^T(\mathbf{E}_Q\otimes\mathbf{A}_k)(\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0))^{-1}\mathbf{D}_{ij}^T)^{-1}(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)^{-1} \\
&= (\mathbf{D}_{ij}(\mathbf{E}_Q\bar{\mathbf{h}}_0^T\mathbf{A}_k\bar{\mathbf{h}}_0)^{-1}\mathbf{D}_{ij}^T)^{-1}(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)^{-1} \\
&= (\mathbf{D}_{ij}\mathbf{E}_Q^{-1}\mathbf{D}_{ij}^T)^{-1}\bar{\mathbf{h}}_0^T\mathbf{A}_k\bar{\mathbf{h}}_0(\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0)^{-1} \\
&= \begin{cases} \dfrac{Np(1-Qp)}{1-(Q-1)p}\dfrac{\bar{\mathbf{h}}_0^{\mathrm{T}}\mathbf{A}_{\mathrm{k}}\bar{\mathbf{h}}_0}{\bar{\mathbf{h}}_0^{\mathrm{T}}\bar{\mathbf{h}}_0} & \text{for } i=j \\ \dfrac{Np}{2}\dfrac{\bar{\mathbf{h}}_0^T\mathbf{A}_k\bar{\mathbf{h}}_0}{\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0} & \text{for } i\neq j \end{cases}
\end{aligned}
$$

Substitution of the R_{ij} terms into Eq. (11) yields Eq. (16).

A.5. Estimation efficiency equals detection power when $\boldsymbol{B} = \bar{\boldsymbol{h}}_0$

With the assumption that there is only one basis function such that $\mathbf{B} = \tilde{\mathbf{h}}_0$, the trace of each covariance term may be written as:

$$
\begin{aligned}
Tr[\mathbf{C}_{ij}] &= Tr[\mathbf{L}_{ij}\tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\mathbf{X}_\perp^T\mathbf{X}_\perp\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}^T\mathbf{L}_{ij}^T] \\
&= Tr[\mathbf{L}_{ij}(\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0)(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}(\mathbf{I}_Q\otimes\bar{\mathbf{h}}_0^T)\mathbf{L}_{ij}^T] \\
&= Tr[(\mathbf{D}_{ij}\otimes\mathbf{I}_k)((\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\otimes\bar{\mathbf{h}}_0\bar{\mathbf{h}}_0^T)(\mathbf{D}_{ij}^T\otimes\mathbf{I}_k)] \\
&= Tr[(\mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T)\otimes(\bar{\mathbf{h}}_0\bar{\mathbf{h}}_0^T)] \\
&= \mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T Tr[\bar{\mathbf{h}}_0\bar{\mathbf{h}}_0^T] \\
&= \mathbf{D}_{ij}(\mathbf{Z}_\perp^T\mathbf{Z}_\perp)^{-1}\mathbf{D}_{ij}^T\bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0 = \bar{\mathbf{h}}_0^T\bar{\mathbf{h}}_0 C_{ij}
\end{aligned}
$$

Substitution of the $Tr[\mathbf{C}_{ij}]$ terms into Eq. (5) yields Eq. (11), so that $\xi_{\mathrm{tot}} = R_{\mathrm{tot}}$. Note that this proof does not rely on the Kronecker product approximation for the Fisher information matrix.

A.6. Gain in estimation efficiency obtained with basis function expansions

When basis function expansions are used, the $Tr[\mathbf{C}_{ij}]$ terms can be written as

$$
\begin{aligned}
Tr[\mathbf{C}_{ij}] &= Tr[\mathbf{L}_{ij}\tilde{\mathbf{B}}(\tilde{\mathbf{B}}^T\mathbf{X}_\perp^T\mathbf{X}_\perp\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}^T\mathbf{L}_{ij}^T] \\
&\approx Tr[\mathbf{L}_{ij}\tilde{\mathbf{B}}((\mathbf{I}_Q\otimes\mathbf{B}^T)(\mathbf{E}_Q\otimes\mathbf{A}_k)(\mathbf{I}_Q\otimes\mathbf{B}))^{-1}\tilde{\mathbf{B}}^T\mathbf{L}_{ij}^T] \\
&= Tr[\mathbf{L}_{ij}(\mathbf{I}_Q\otimes\mathbf{B})(\mathbf{E}_Q\otimes(\mathbf{B}^T\mathbf{A}_k\mathbf{B}))^{-1}(\mathbf{I}_Q\otimes\mathbf{B}^T)\mathbf{L}_{ij}^T] \\
&= Tr[(\mathbf{D}_{ij}\otimes\mathbf{I}_k)(\mathbf{E}_Q^{-1}\otimes(\mathbf{B}(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}\mathbf{B}^T))(\mathbf{D}_{ij}^T\otimes\mathbf{I}_k)] \\
&= Tr[(\mathbf{D}_{ij}\mathbf{E}_Q^{-1}\mathbf{D}_{ij}^T)\otimes(\mathbf{B}(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}\mathbf{B}^T)] \\
&= Tr[\mathbf{D}_{ij}\mathbf{E}_Q^{-1}\mathbf{D}_{ij}^T]Tr[\mathbf{B}(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}\mathbf{B}^T] \\
&= Tr[\mathbf{D}_{ij}\mathbf{E}_Q^{-1}\mathbf{D}_{ij}^T]Tr[(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}]
\end{aligned}
$$

where we have made use of Kronecker product approximation for the Fisher information matrix, the trace identity $Tr[\mathbf{AB}] = Tr[\mathbf{BA}]$ and our assumption that $\mathbf{B}^T\mathbf{B} = \mathbf{I}$. The final expression is identical to the expression for $Tr[\mathbf{C}_{ij}]$ without the use of basis functions with the term $Tr[\mathbf{A}_k^{-1}]$ replaced by the term $Tr[(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}]$. The gain in efficiency can be assessed by comparing these terms. In the absence of basis functions, the optimal design has $\mathbf{A}_k = \mathbf{I}_k$. The gain in efficiency observed by using this design with basis functions is then $Tr[\mathbf{I}_k^{-1}]/Tr[(\mathbf{B}^T\mathbf{I}_k\mathbf{B})^{-1}] = k/s$, which reflects the fact that s/k fewer parameters are estimated. Estimation efficiency can be further increased by choosing a design that minimizes $Tr[(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}]$. For the optimal design, the eigenvalues of $\mathbf{B}^T\mathbf{A}_k\mathbf{B}$ are equally distributed and have maximum amplitude (Liu et al., 2001). Consider the eigen decomposition $\mathbf{A}_k = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{V}$ is the unitary matrix composed of the eigenvectors of $\mathbf{A}_k$ and $\mathbf{\Lambda}$ is the diagonal matrix with eigenvalues of $\mathbf{A}_k$. The eigenvalues of $\mathbf{B}^T\mathbf{A}_k\mathbf{B} = \mathbf{B}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{B}$ can be maximized if the first s eigenvectors of $\mathbf{A}_k$ are equal to the columns of $\mathbf{B}$. Then $\mathbf{V}^T\mathbf{B} = \begin{bmatrix}\mathbf{I}_s \\ 0_{(k-s)\times s}\end{bmatrix}$ and $\mathbf{B}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{B} = \mathbf{\Lambda}_s$, where $\mathbf{\Lambda}_s$ is the $s \times s$ diagonal matrix composed of the first s eigenvalues of $\mathbf{A}_k$. Thus, the eigenvalues of $\mathbf{B}^T\mathbf{A}_k\mathbf{B}$ are given by the first s eigenvalues of $\mathbf{A}_k$. From the definition of $\mathbf{A}_k$ in the section on the Fisher information matrix, the upper bound on the sum of eigenvalues of $\mathbf{A}_k$ is $\Sigma_{i=1}^k\lambda_i = Tr[\mathbf{A}_k] \leq k$. Equal distribution of this sum over the first s eigenvalues yields $Tr[(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}] = \Sigma_{i=1}^s\frac{1}{\lambda_i} = \Sigma_{i=1}^s\frac{s}{k} = \frac{s^2}{k}$.

For comparison, when no basis functions are used, efficiency is maximized when the eigenvalues of $\mathbf{A}_k$ are equally distributed, or equivalently $\mathbf{A}_k = \mathbf{I}_k$, so that $Tr[\mathbf{A}_k^{-1}] = k$. Thus, the relative gain in maximum achievable efficiency is given by $\min(Tr[\mathbf{A}_k^{-1}])/\min(Tr[(\mathbf{B}^T\mathbf{A}_k\mathbf{B})^{-1}]) = k^2/s^2$.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8, 360–369.

Birn, R.M., Cox, R.W., Bandettini, P.A., 2002. Detection versus estimation in event-related fMRI: choosing the optimal stimulus timing. NeuroImage 15, 252–264.

Bischoff-Grethe, A., Martin, M., Mao, H., Berns, G.S., 2001. The context of uncertainty modulates the subcortical response to predictability. J. Cogn. Sci. 13, 986–993.

Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. 16, 4207–4221.

Brewer, J.W., 1978. Kronecker products and matrix calculus in system theory. IEEE Trans. Circuits Syst. 25 (9), 772–781.

Buracas, G.T., Boynton, G.M., 2002. Efficient design of event-related fMRI experiments using M-sequences. NeuroImage 16, 801–813.

Burock, M.A., Dale, A.M., 2000. Estimation and detection of event-related fMRI signals with temporally correlated noise: a statistically efficient and unbiased approach. Hum. Brain Mapp. 11, 249–260.

Burock, M.A., Buckner, R.L., Woldorff, M.G., Rosen, B.R., Dale, A.M., 1998. Randomized event-related experimental designs allow for extremely rapid presentation rates using functional MRI. NeuroReport 9, 3735–3739.

Clark, V.P., Maisog, J.M., Haxby, J.V., 1998. fMRI study of face perception and memory using random stimulus sequences. J. Neurophysiol. 79, 3257–3265.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley, New York.

Dale, A.M., 1999. Optimal experimental design for event-related fMRI. Hum. Brain Mapp. 8, 109–114.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.

Friston, K.J., Josephs, O., Rees, G., Turner, R., 1998. Nonlinear event-related responses in fMRI. Magn. Reson. Med. 39, 41–52.

Friston, K.J., Zarahn, E., Josephs, O., Henson, R.N.A., Dale, A.M., 1999. Stochastic designs in event-related fMRI. NeuroImage 10, 607–619.

Godfrey, K., 1993. Perturbation Signals for System Identification. Prentice Hall, Englewood Cliffs, NJ.

Haykin, S., 1996. Adaptive Filter Theory. Prentice Hall, Upper Saddle River, NJ.

Kailath, T., Sayed, A.H., Hassibi, B., 2000. Linear Estimation. Prentice Hall, Upper Saddle River, NJ.

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., Frank, L.R., Wong, E.C., Buxton, R.B., 2001. Detection power, estimation efficiency, and predictability in event-related fMRI. NeuroImage 13, 759–773.

Liu, T.T., Buxton, R.B., Ghobrial, E., 2002. Unbiased Volterra kernel analysis of event-related fMRI data. Proceedings of the 10th Meeting Int. Soc. Magn. Reson. Med., Honolulu, Hawaii, 752.

Moon, T.K., Stirling, W.C., 2000. Mathematical Methods and Algorithms for Signal Processing. Prentice Hall, Upper Saddle River, NJ.

Rosen, B.R., Buckner, R.L., Dale, A.M., 1998. Event-related functional MRI: past, present, and future. Proc. Natl. Acad. Sci. U. S. A. 95, 773–780.

Scharf, L.L., 1991. Statistical Signal Processing: Detection, Estimation, and Time Series Analysis. Addison-Wesley, Reading, MA.

Scharf, L.L., Friedlander, B., 1994. Matched subspace detectors. IEEE Trans. Signal Process. 42 (8), 2146–2157.

Seber, G.A.F., 1977. Linear Regression Analysis. Wiley, New York.

Strang, G., 1980. Linear Algebra and Its Applications. Harcourt Brace Jovanovich, San Diego.

Wager, T.D., Nichols, T.E., 2003. Optimization of experimental design in fMRI: a general framework using a genetic algorithm. NeuroImage 18, 293–309.