Statistical modeling: Single subject analysis

The goal of an fMRI data analysis is to analyze each voxel's time series to see whether the BOLD signal changes in response to some manipulation. For example, if a stimulus was repeatedly presented to a subject in a blocked fashion, following the trend shown in the red line in the top panel of Figure 5.1, we would search for voxel time series that match this pattern, such as the BOLD signal shown in blue. The tool used to fit and detect this variation is the general linear model (GLM), where the BOLD time series plays the role of dependent variable, and the independent variables in the model reflect the expected BOLD stimulus timecourses. Observe, though, that square wave predictor in red doesn't follow the BOLD data very well, due to sluggish response of the physiology. This leads to one major focus of this chapter: Using our understanding of the BOLD response to create GLM predictors that will model the BOLD signal as accurately as possible. The other focus is modeling and accounting for BOLD noise and other sources of variation in fMRI time series.

Throughout this chapter the models being discussed will refer to modeling the BOLD signal in a single voxel in the brain. Such a voxel-by-voxel approach is known as a mass univariate data analysis, in contrast to a multivariate approach (see Chapters 8 and 9 for uses of multivariate models). We assume that the reader has a basic understanding of the general linear model; for a review see Appendix A. Once the data in all of the voxels are analyzed separately, they are combined across subjects for a group analysis (as described in Chapter 6) and then the statistics are assessed as an image as part of inference (as described in Chapter 7).

5.1 The BOLD signal

As described in Chapter 1, the BOLD signal arises from the interplay of blood flow, blood volume, and blood oxygenation in response to changes in neuronal activity. In short, under an active state, the local concentration of oxygenated hemoglobin increases, which increases homogeneity of magnetic susceptibility, resulting in an
5.1 The BOLD signal

Figure 5.1. Illustration of BOLD fMRI time series in active voxel and illustration of an unconvolved signal used to model the signal. The BOLD signal for an active voxel (blue) and the stimulus time series (red) is shown.

increase in T2*-weighted MRI signal. As shown in Figure 5.1 the BOLD signal (blue) does not increase instantaneously and does not return to baseline immediately after the stimulus ends (red). Because these changes in blood flow are relatively slow (evolving over several seconds), the BOLD signal is a blurred and delayed representation of the original neural signal.

The hemodynamic response can be described as the ideal, noiseless response to an infinitesimally brief stimulus. It has a number of important characteristics, shown in Figure 5.2:

- **Peak height:** This is the most common feature of interest, since it is most directly related to the amount of neuronal activity in the tissue (Logothetis et al., 2001). For BOLD fMRI, the maximum observed amplitude is about 5% for primary sensory stimulation, whereas signals of interest in cognitive studies are often in the 0.1–0.5% range.
- **Time to peak:** The peak of the HRF generally falls within 4–6 seconds of the stimulus onset.
- **Width:** The HRF rises within 1–2 seconds and returns to baseline by 12–20 seconds after the stimulus onset.
- **Initial dip:** Some studies have identified an initial dip in the BOLD signal that occurs within the first 1–2 seconds and is thought to reflect early oxygen consumption before changes in blood flow and volume occur (Buxton, 2001). Many studies have not found the initial dip, and when it is observed, it is generally a very small signal in comparison to the peak positive BOLD response. It is generally ignored in most models of fMRI data.
- **Poststimulus undershoot:** The HRF generally shows a late undershoot, which is relatively small in amplitude compared to the positive response and persists up to 20 seconds or more after the stimulus.

Importantly, there is substantial variability in each of these features of the HRF across brain areas and across individuals. For example, in the work of Kruggel & von Cramon (1999), the time until peak varied between 6 and 11 seconds.
Figure 5.2. Characteristics of the hemodynamic response. The shape of the HRF function can be described by a variety of characteristics including the time from the stimulus until peak (TP), height of response (H), the width of the HRF at half the height (W), poststimulus undershoot (PSU) and in some cases an initial dip (ID).

across voxels in a single subject. In Handwerker et al. (2004), a study of the HRF shape revealed that both the time until peak and width of the HRF varied within subjects across different regions of the brain and across subjects, with intersubject variability higher than intrasubject variability. And D’Esposito et al. (2003) reviewed a number of studies that compared the BOLD response in healthy young and elderly subjects and found, while the shape of the HRF was similar between the groups, elderly had reduced signal-to-noise ratios in the response magnitudes.

5.1.1 Convolution

An important characteristic of the BOLD signal is that the relationship between the neural response and the BOLD signal exhibits linear time invariant (LTI) properties. The meaning of linearity is that if a neural response is scaled by a factor of \( a \), then the BOLD response is also scaled by this same factor of \( a \). As shown in panel A of Figure 5.3, when the neural response (red) is doubled in magnitude, the expected BOLD response (blue) would then be doubled in magnitude. Linearity also implies additivity, in that if you know what the response is for two separate events, if the events were to both occur close together in time, the resulting signal would be the sum of the independent signals. This is illustrated in panel B of Figure 5.3, where neural responses for separate events (green) add linearly to create the expected BOLD response. Time invariant means that if a stimulus is shifted by \( t \) seconds, the BOLD response will also be shifted by this same amount.
5.1 The BOLD signal

Figure 5.3. Examples of linear time invariance. Panel A illustrates that when a neural signal is twice another, the resulting BOLD activation is also twice as large. Panel B shows how the signals for separate trials, shown in green, add linearly to get the BOLD activation.

Box 5.1.1 What is linear about the hemodynamic response?

The GLM approach to fMRI analysis relies crucially on the assumption that the hemodynamic response is a linear transformation of the underlying neuronal signal, and the question of linearity has been examined in detail in the fMRI literature.

It is first important to point out that there are two possible sources of nonlinearity between stimulation and BOLD responses. There could be nonlinearity between the BOLD response and the neuronal signal, which is the main focus of our review here. However, it is well known that there can also be nonlinearities in the relationship between stimuli and neuronal responses. A striking example occurs in the auditory system, where sustained trains of sounds exceeding a particular frequency are associated not with sustained neuronal activity but rather with phasic bursts of activity at the onset and offset of the train (Harms & Melcher, 2002). In other cases, such nonlinearities can occur due to
neuronal adaptation, whereby the neuronal response to the same stimulus becomes decreased upon repetition of that stimulus. Thus, it is always important to take into account known neurophysiological data when devising models for fMRI studies.

In early work by Dale & Buckner (1997), stimuli were presented in rapid succession and the assumption of linearity was tested by examining whether the estimated response to multiple stimuli matched the response to a single stimulus. This work showed that the response was indeed largely linear, and that the estimated hemodynamic responses to subsequent trials were very similar to the responses to a single trial. However, this match was not exact; in particular, the estimates on subsequent trials were somewhat compressed relative to the first trial. Further work has confirmed this nonlinearity, particularly for stimuli that occur less than 2 seconds apart (e.g., Wager et al., 2005). Another nonlinearity that has been noted relates to stimulus duration, whereby very brief stimuli exhibit much larger BOLD responses than would be expected based on longer stimuli. For example, Yeşilyurt et al. (2008) found that the BOLD response to a 5-millisecond visual stimulus was only half as large as the response to a 1,000-millisecond stimulus. Fortunately, while these nonlinearities are clearly important, for the range in which most cognitive fMRI studies occur, they will have relatively small impact.

The LTI properties of the BOLD signal have been studied in detail (see Box 5.1.1.), and there is a general consensus that the transform from neuronal activity to BOLD signal is largely LTI. Because of this, a natural approach to creating an expected BOLD signal from a given neural input is to use the convolution operation. Convolution is a way of blending two functions together in an LTI fashion. Specifically, the stimulus onset time series, \( f \) (such as the red trend in Figure 5.1) is blended with an HRF, \( h \), creating a shape that more closely represents the shape of the BOLD response. The operation is given by

\[
(h * f)(t) = \int h(\tau)f(t - \tau)d\tau
\]

(5.1)

Recall that in Section 3.7 convolution was used in a slightly different context when data were \textit{spatially} smoothed by convolving the data with a Gaussian kernel.

Choosing an appropriate HRF function is key in capturing the shape as best as possible and will ensure a good fit of the GLM regressors to the BOLD time series when signal is present.

5.1.1.1 Characterizing the hemodynamic response function

To obtain the predicted BOLD response using convolution, we need an estimate of the HRF. One way to estimate the shape of the response is to present stimuli that are...
5.1 The BOLD signal

Figure 5.4. Example of selective averaging. Panel A shows the original time series, highlighting the 22-second window around each stimulus presentation. Panel B overlays the windowed BOLD time series, and panel C shows the average of the windowed time series.

widely spaced in time (e.g., every 30 seconds) and then simply to average the evoked responses at each point in time with respect to the stimulus (referred to variously as peristimulus time averaging, selective averaging, or deconvolution). Figure 5.4 shows an example of such a process. Panel A shows the original time series, where 22-second windows around each stimulus (starting 2 seconds before the stimulus and lasting for 20 seconds after) are highlighted in different colors. The 11 windowed segments are overlaid in panel B and after averaging the timecourses (panel C) a less noisy image of the response function is obtained.

The work of Friston et al. (1994a) and Lange & Zeger (1997) applied deconvolution models to BOLD data to characterize the HRF and found that, in general, it was approximately described by a gamma function. The choice of the “canonical” HRF using a single gamma function was common until it was realized that the model fit could be further improved by accounting for the poststimulus undershoot, which is not modeled with a single gamma HRF. For this reason, a canonical HRF based on the combination of two gamma functions, known as a double-gamma HRF, was adopted (Friston et al., 1998; Glover, 1999). The first gamma function models the shape of the initial stimulus response and the second gamma function models the undershoot. Each analysis software package has default parameter settings that generate a
Statistical modeling: Single subject analysis

Figure 5.5. Illustration of different canonical hemodynamic response functions. The black line is the average BOLD response over multiple presentations of a block of stimuli. The red line is the unconvolved time course, blue uses the gamma HRF, whereas the green line is the double gamma, which fits the data best as it accounts for the poststimulus undershoot.

canonical HRF. For example, the default double gamma HRF in SPM has a delay of response set to 6 seconds and a delay of undershoot (relative to onset) of 16 seconds, among five other parameters. These defaults are used in all analyses unless specified otherwise by the user. The only parameter that is free to vary, which is actually estimated in the linear model, is the height of the response. Figure 5.5 illustrates the model fit to data (black) for the boxcar regressor (red), boxcar convolved with a gamma HRF (blue), and boxcar convolved with a double gamma HRF (green). The double gamma has a better fit, since it models the poststimulus undershoot. If a gamma is used instead of a double gamma when there is a strong poststimulus undershoot, the undershoot may pull down the baseline of the model, and as a result the height of the response may be underestimated.

5.1.2 Beyond the canonical HRF

The approach described previously assumed that the HRF could be accurately described using a single canonical response function. However, as was seen in Figure 1.2, hemodynamic responses may differ substantially between individuals in their shape, time to peak, and other aspects. In addition, a number of studies have shown that the shape of the hemodynamic response also differs across different brain regions for the same individual. If we use the canonical HRF, then we are biased to only find responses that are similar to that function. On the other hand, if we use a more complicated model allowing for more flexibility in the shape of the HRF by incorporating more parameters, we will have more variability in the estimates. This is what is often referred to as the bias-variance tradeoff. Because of the known variability in the HRF, it is common to use more complicated and flexible models in fMRI analysis.

If we do need a more flexible HRF model, there are a number of approaches that can be used to capture a broader range of response profiles. A popular approach is to use a set of HRF basis functions, functions that when combined linearly give a range
5.1 The BOLD signal

of expected shapes for the hemodynamic response. Using just the single canonical response function is the special case of using just one basis function. A two basis function example is the use of a single canonical response function as well as its derivative, allowing for a slight temporal shift. Some early work modeling fMRI data used Fourier sets (or sets of sine and cosine functions) to model the hemodynamic response (Josephs et al., 1997). Other basis sets include the finite impulse response model (FIR) set and constrained basis sets. We now review each of these approaches in turn.

5.1.2.1 Modeling the derivative

Probably the most commonly used basis set for fMRI analysis is the “canonical HRF plus derivatives” approach developed by Friston et al. (1998). The rationale for including the temporal derivative is that this basis can capture small offsets in the time to peak of the response. Given the expected signal, \( Y(t) = \beta X(t) \), a time-shifted version of the hemodynamic response function can be described as \( Y(t) = \beta X(t + \delta) \). However, the shift \( \delta \) is not linear and hence cannot be estimated with the GLM. However, a Taylor series expansion of \( X(t + \delta) \) with respect to \( \delta \) gives a linear approximation, \( Y(t) = \beta X(t) + \delta X'(t) + \cdots \), implying a GLM of \( Y(t) \approx \beta_1 X(t) + \beta_2 X'(t) \). The \( \beta_2 \) term is not directly interpretable as a delay parameter, but linear combinations of \( X(t) \) and \( X'(t) \) will model small time shifts of the HRF. Figure 5.6 shows the standard regressor, consisting of the convolution of the stimulus onset and the canonical HRF, its first derivative and the sum of the two, illustrating how adding the two results in a slight shift to the left.

The same Taylor series approach can be taken with respect to other parameters that describe the canonical HRF. For example, the SPM software offers to include “time & dispersion” derivatives. In this case, a third HRF basis is used that is the derivative of \( X(t) \) with respect to width of the response (parameter \( W \) shown in

![Figure 5.6](image-url)

Figure 5.6. The stimulus convolved with the HRF (blue), its derivative (red), and the sum of the two (green), illustrating that including a derivative term in your linear model can adjust for small shifts in the timing of the stimulus.
Statistical modeling: Single subject analysis

Figure 5.2). Similar to a temporal derivative that models temporal shifts, this allows the width of the fitted HRF to vary slightly from the canonical HRF.

5.1.2.2 Finite impulse response models

The most flexible model for capturing the HRF shape for a response is the FIR basis set. In this approach, a window around the stimulus is chosen and the GLM is used to model the response at each time point within this window. Figure 5.7 shows an example for a study where one trial type was present, where the left side of the figure illustrates the model and the right panel shows how the parameter estimates from the model reveal the HRF shape. Each regressor models a specific time point in the window surrounding the trial presentation (often called peristimulus time). In this case the first regressor corresponds to 2 seconds prior to stimulus presentation, the second is during the stimulus presentation and so on, continuing until 20 seconds after the trial onset. Note that in some cases a single time point (row of the design matrix) may be modeled by two different parameters when trials are sufficiently close in time. For example, in this design the first two time points of a trial window overlap with the last two time points of the previous trial window. The GLM will appropriately account for this overlap by virtue of the additivity of the BOLD response.

The flexibility of the FIR model to capture the shape of the HRF comes at the cost of an increase in the variability of the estimates (i.e., a bias–variance tradeoff). While we are less biased about the shape of the HRF, the variability of our estimates

![Model: $Y = X\theta$](image)

![Parameter Estimates](image)

Figure 5.7. Model and estimates corresponding to an FIR model. On the left is the model used to estimate the FIR parameters, where the BOLD time series is the outcome, the design matrix consists of 0s (black) and 1s (white), and there is a separate parameter estimated for each point in the window around the stimulus. For example, the first regressor and parameter ($\theta_{\text{ons-2}}$) correspond to the time point 2 seconds prior to onset time, the second regressor is during onset and so on. The TR in this case is 2 seconds, so the regressors differ by increments of 2 seconds. The right panel illustrates how the parameter estimates from the FIR model capture the shape of the HRF.
increases since fewer data points are contributing to each parameter’s estimate. Additionally, collinearity between the regressors, due to overlapping stimuli, can increase the variability of the estimates further.

One consideration when using the FIR model is how the results of the model fit are used at the next level of analysis. It is common for researchers to fit an FIR model to their data, with reasoning like “I don’t want to assume a shape of the hemodynamic response.” This is true, but it only holds if one actually uses the entirety of the FIR fit in the higher-level analysis. As discussed in Chapter 6, the standard group modeling approach assumes the data have been distilled down to one BOLD measure per subject (e.g., the parameter for an event modeled with a canonical HRF). However, a FIR fit consists of many measures (e.g., 12 values in the example in Figure 5.7), and this requires a type of multivariate model to fit all of these responses at the group level. Some authors simply select the parameter from a single FIR bin to take to a group analysis, however this itself implies an assumption that the peak BOLD response falls at that time point. In general, FIR models are most appropriate for studies focused on the characterization of the shape of the hemodynamic response, and not for studies that are primarily focused on detecting activation.

5.1.2.3 Constrained basis sets

At one end of the spectrum there is the lower-bias, higher-variance FIR basis set and at the other end is the higher-bias, lower-variance canonical HRF basis function. An approach that falls somewhere in the middle of the bias–variance spectrum is the use of constrained basis sets, which takes into account known features of the hemodynamic response (such as the fact that it is smooth, starts at zero, ramps slowly, and returns to zero) but still allows flexibility to fit a range of possible responses. Instead of convolving the stimulus onset time series with a single canonical HRF, a set of functions that capture different aspects of the shape of the HRF are used. One approach to constructing a set of basis functions is to first generate a set of HRF shapes that are reasonable, say by varying some of the parameters outlined in Figure 5.2, and then using principal components analysis to extract a set of basis functions that describe this set well. This is the approach taken by Woolrich et al. (2004a) in the FMRIB Linear Optimal Basis Set (FLOBS) algorithm. Figure 5.8 shows an example of four-function basis set developed using this approach (panel A), and some examples of HRF shapes that can be created using linear combinations of these functions (panel B). Although some shape is imposed, different linear combinations of these shapes allow for a larger variety of HRF shapes than using a single canonical basis function.

The benefit of this basis set is that instead of the, say, 12 parameters that must be estimated for the FIR model, only 4 parameters must be estimated for this basis set, which helps reduce the variability of the estimates. Figure 5.9 illustrates the fit of three approaches: double gamma canonical HRF (blue), a basis set with four basis functions (red), and the FIR model (green). The canonical HRF model is fitting a
Statistical modeling: Single subject analysis

Figure 5.8. Examples of constrained basis sets (panel A) and linear combinations of the basis (panel B). The four basis functions in panel A were produced using the FMRIB linear optimal basis set algorithm. These functions would be convolved with the stimulus onset to create four regressors for modeling the overall HRF shape, compared with the ten regressors shown in the FIR model in Figure 5.7. The right panel illustrates 4 different linear combinations of the basis functions, illustrating the variations in HRF shape that the basis functions are able to model.

Figure 5.9. Examples of fitting data with a canonical HRF (blue), set of three basis functions (red) and FIR model with ten time points (green). In this case the peak of the double gamma seems to be too early compared to the fits of the more flexible basis function and FIR models. The double gamma fit is smoothest while the FIR model fit is the noisiest.

single parameter and has the smoothest fit, whereas the ten-parameter FIR fit is the noisiest. Both the FIR and four-basis function models illustrate that the imposed time to peak of the double gamma function may be a bit too early and therefore the peak amplitude is underestimated.
5.1 The BOLD signal

In summary, although it would seem that we always want unbiased estimates, often accepting a small amount of bias will result in a greater reduction in variance, especially when that bias fits with our preexisting knowledge about the underlying data. When using models with a large set of basis functions, we must be aware that the flexibility of the model may fit unrealistic HRFs, for example, a shape with two separate positive bumps. In fMRI this means that we usually accept imperfect fit of the canonical HRF in exchange for greater precision and interpretability in our parameter estimates.

5.1.3 Other modeling considerations

Time resolution of the model. The canonical HRF is usually plotted as a smooth curve, as in Figure 5.2, with very fine time resolution. In practice, BOLD time series have time resolution (i.e., repetition time or TR) of 2–3 seconds, yet the events do not usually fall exactly at the beginning of an image acquisition nor last the entire TR. At coarse time resolution, the stimulus time course for any event or combination of events occurring within the time frame of a single TR would look the same at a time resolution of TR seconds, the convolved signal would look the same. For example, if the TR = 3 seconds, the following four types of trial presentations would have the same representation at the time resolution of 3 seconds: a stimulus starting at 0 second and lasting 1 second, a stimulus starting at 0 second lasting 2 seconds, a stimulus starting at 0 second and lasting 3 seconds, and a stimulus starting at 2 seconds lasting 1 second. Specifically, the convolved signal for all four of these tasks in a space with a single unit time resolution of 3 seconds would be the dashed black line shown in the right panel of Figure 5.10.

Figure 5.10. Illustrating why convolution is performed in a higher-resolution time domain than the TR. The left panel shows different stimuli, with timings indicated in the legend, convolved with the double gamma HRF in a high-resolution time domain. The right panel illustrates the result that would be acquired for all four stimuli if convolution was done in the lower time resolution of 3 seconds (black dashed line) compared to the signals that are acquired by down-sampling the higher-resolution signals from the left-hand panel (solid colored lines).
Statistical modeling: Single subject analysis

To more accurately model the predicted HRF, most software packages first up-sample the stimulus presentation times into a higher time resolution, convolve them with the HRF, and then down-sample the convolved functions to the original time resolution. By doing so, the timing of the stimuli is more accurately represented and since a higher resolution HRF is used, the shape of the response is more accurate. The left-hand side of Figure 5.10 shows the four time courses described previously, convolved in a high time resolution space of 0.1 second. The right panel of the figure shows the down-sampled versions of the time series (solid lines) as well as the signal that would result if the convolution was done in the 3-seconds time resolution space (dotted line). The reason the dotted black line does not match the 3-second event (yellow line) is a result of differences in accuracies of HRF shapes in the two time domains.

Modeling parametric modulation. In many cases, some feature of a stimulus or task can be parametrically varied, with the expectation that this will be reflected in the strength of the neural response. For example, Boynton et al. (1996) parametrically modulated the contrast of a visual stimulus, and showed that neural responses in area V1 responded linearly to this modulation. In designs like this, the parametric modulation can be modeled using an additional regressor in the design matrix, known as a parametric regressor. A stimulus onset time series consists of stick functions, usually of equal height. To create a parametric regressor, each onset’s stick function has a height reflecting the strength of the stimulus for that trial. Figure 5.11 shows an example of how a parametrically modulated regressor is created. The top panel shows the timing of the stimuli, where the numbers above the stimuli refer to the modulation value. The unmodulated regressor is shown in panel B and panel C shows the modulated version. It is important that the height values are demeaned prior to creating the regressor, in order to ensure that the parametric regressor is not correlated with the unmodulated regressor. It is also important that the GLM always include an unmodulated regressor in addition to the parametrically modulated regressor. This is analogous to the need to include an intercept term in a linear regression model along with the slope term, since without an intercept it is assumed that the fitted line will go through the origin.

Modeling behavioral response times. Historically, most fMRI researchers have not taken response times of the subject’s behavioral responses into account in modeling fMRI data; instead, they have used a constant-duration impulse for all trials. However, one of the general maxims of good statistical modeling practice is that if we know of a factor that could affect the data, we should include it in the model. Due to linearity of the BOLD response, we know that twice the neural response results in a BOLD signal that has twice the magnitude. However, as shown in Figure 5.10 a stimulus that is twice as long will also have a BOLD response that is about twice as high (red versus blue line). Thus, trials with longer processing times could have
5.1 The BOLD signal

Figure 5.11. Construction of a parametrically modulated regressor. Panel A shows the timing of the stimuli, where the numbers above the stimuli correspond to the modulation value for that stimulus. Panel B shows the unmodulated regressor, and Panel C shows the modulated regressor. Note that both the modulated and unmodulated regressors would be included in the linear regression model.

much greater activation simply due to the amount of time on the task, rather than reflecting any qualitative difference in the nature of neural processing. In fact, it is likely that many differences in activation between conditions observed in functional neuroimaging studies are due simply to the fact that one condition takes longer than the other.

One alternative (recommended by Grinband et al., 2008) is to create the primary regressors in the model using the actual duration of each trial, rather than a fixed duration across trials. This will increase sensitivity for effects that vary with response
Statistical modeling: Single subject analysis

time but will decrease sensitivity for effects that are constant across trials. A second alternative, which we prefer, is to create the primary regressor using a constant duration, but then include an additional parametric regressor that varies according to response time. This will ensure that the effects of response time are removed from the model and also allows the separate interrogation of constant effects and effects that vary with response time.

Modeling motion parameters. As described in Section 3.6, head motion during the scan can cause artifacts in the data even after applying image registration methods. As a result, it is a good idea to include motion regressors in the model to account for artifacts and motion-related variance. This is done by including the six time courses of the translation and rotation parameters as nuisance regressors in the model. The term nuisance is used to describe regressors that are included in the model to pick up extra variability in the data when there is no interest in carrying out inference on the corresponding parameters. Additionally, it is often beneficial to include the derivatives of the motion parameters, as they can help model motion-related noise and spikes in the data.

Generally we expect that the inclusion of motion regressors will reduce error variance and improve detection power. However, if the motion is correlated with the task, inclusion of the motion regressors may eliminate significant regions seen otherwise. This is because the GLM bases the significance of experimental effects only on the variability uniquely attributable to experimental sources. When such variability is indistinguishable between motion and experimental sources, the significance of the results is reduced, which prevents movement-induced false positives.

Orthogonalization. A common occurrence in fMRI studies is to have a design that includes regressors that are correlated with each other to some extent. For example, if you were modeling the presentation of a stimulus as well as the subject’s response, these two events occur in close proximity, and so the corresponding regressors in the GLM will be highly correlated. The second panel of Figure 5.12 shows an example of correlated regressors ($r = 0.59$), where the green time series represents the stimulus presentation and the red time series models the subject’s response 2 seconds later. As just mentioned, the GLM has an essential property in that only the variability unique to a particular regressor drives the parameter estimate for that regressor. The variability described by two regressors $X_1$ and $X_2$ can be thought of as having three portions, that unique to $X_1$, that unique to $X_2$, and the shared variability. In cases where two regressors are orthogonal (uncorrelated), there is no shared variability component, whereas when they are highly correlated, the unique portion for each regressor is small. This results in unstable and highly variable parameter estimates, which leads to a loss of statistical power.

The data in the top panel of Figure 5.12 contains stimulus- and response-related effects; the second panel of Figure 5.12 shows the model and t-statistics for the
5.1 The BOLD signal

Figure 5.12. Illustration of regressor orthogonalization. The top panel shows simulated data corresponding to a positive response to both the stimulus and response. The second panel shows the highly correlated regressors that would be used to model the stimulus (green) and response (red). The following two panels illustrate what each of the regressors looks like when orthogonalized with respect to the other; the right column shows the t-statistics in each case. Note that the t-statistic for the non-orthogonalized regressor is what changes in each case.

model with unaltered stimulius and response variables. Even though there is signal, due to correlation between regressors, only one t-statistic is significant. The third panel of Figure 5.12 shows the model after orthogonalizing the stimulus regressor with respect to the response regressor (essentially just the residuals from regressing the red regressor on the green). Now the response t-statistic is very large as we’ve forced it to take all variation it can; note that the stimulus t-statistic is the same, as its interpretation hasn’t changed; it still measures the unique variation attributable to response. The bottom panel of Figure 5.12 shows the reverse situation, after
Statistical modeling: Single subject analysis

orthogonalizing the response regressor with respect to the stimulus regressor. Now the response t-statistic is very large, and the stimulus regressor’s t-statistic matches its original value.

It is because of the arbitrary apportionment of variability and significance just demonstrated that we normally recommend against orthogonalization. Only in cases where variables are clearly serving a supplementary role should orthogonalization be used. For example, when a canonical pulse temporal derivative HRF basis is used, the temporal derivative regressor is orthogonalized with respect to the canonical regressor. This is appropriate because the temporal derivative is only present to reduce error variance, and any shared experimental variance between the two regressors can safely be attributed to the first regressor.

5.2 The BOLD noise

The previous section described how the BOLD signal is modeled, and this section focuses on the other type of variability in the data, the noise. In general the term noise is used to describe any variability in the data that is not related to the experimental design. There are two categories of noise. One type is white noise, which is broadband and not focused at any particular frequencies. The other, referred to as structured noise, reflects coherent sources of variability such as physiological fluctuations that occur at particular frequencies, and thus is colored noise. By characterizing the structure of the noise, it can be incorporated into the GLM, improving the fit of the model. Structured noise can result in violations of the assumption of the GLM that observations are not correlated, and false positive rates may increase if it is ignored.

5.2.1 Characterizing the noise

The most obvious characteristic of noise in BOLD fMRI data is the presence of low-frequency drift. Figure 5.13 shows an example of drift in an fMRI time series in both the time domain (left) and the power spectrum in the Fourier domain (right). The power spectrum is acquired by taking the Fourier transform of the time series; the $X$ axis of the plot refers to different frequencies, whereas the $Y$ axis refers to the power, or strength, of this frequency in the data. For example, this voxel is active, and since the frequency of the stimulus is one cycle every 40 seconds, there is a spike in the power spectrum at $1/40$ seconds $= 0.025$ Hz. This time series also exhibits a slowly increasing trend in the time domain, which, since it is low frequency, contributes to the power at lower frequencies of the power spectrum. The shape of the power spectrum is often referred to as the 1/f, or inverse frequency, function (Zarahn et al., 1997).

Initially the source of the 1/f noise structure in fMRI data was not clear. The noise structure was studied in great detail by Aguirre et al. (1997) and Zarahn et al. (1997) in both humans and water phantoms to determine whether the 1/f noise was physiologic or due to the scanner. Additionally, they examined a variety of computers and...
5.2 The BOLD noise

Figure 5.13. Noise structure of fMRI data. The left panel shows the BOLD time series for a voxel that exhibits both signal (blocked pattern is visible) as well as low-frequency drift in the form of a slow uphill trend. The right panel shows the same data in the frequency domain. The red line corresponds to the shape of a fitted 1/f function, which matches the shape of the power spectrum well, and the spike in the power spectrum at 0.025 Hz corresponds to the frequency of the task (once every 40 seconds) in this experiment.

Equipment to determine whether the 1/f noise was due to radiofrequency contamination, but in all cases the 1/f structure prevailed. Some of the noise appears to be due to effects of subject movement that remain after motion correction, or cardiac and respiratory effects. However, even when phantoms and cadavers (Smith et al., 1999) are scanned, low-frequency noise persists, indicating that the scanner itself is an additional source of structured noise.

Since low-frequency noise is always present in fMRI data, it is important that when planning a study the frequency of the task does not fall into the range between 0 and 0.015 Hz where the low-frequency noise has typically been found, meaning that the frequency of trial presentation should be faster than one cycle every 65–70 seconds (i.e., a block length of no more than about 35 seconds for an on–off blocked design). If stimuli are grouped together in long blocks, the signal will be lost in the lower frequencies, where the task signal cannot be separated from the noise. If it is necessary to examine effects that change more slowly, one alternative is to use fMRI techniques such as arterial spin labeling (ASL), which do not exhibit such low-frequency fluctuations (e.g., Detre & Wang, 2002).

Removing the low-frequency trends is handled using a combination of two approaches. First, a high-pass filter is used to remove low-frequency trends from the data. However, after high-pass filtering, fMRI time series are still correlated over time. As discussed in Appendix A, one assumption of the GLM is that the data are not temporally autocorrelated and that the variance of the data is constant over observations. When these assumptions are violated, the inferences based on the GLM are biased and can result in an elevated false positive rate. Thus, a second step attempts to estimate and undo the correlation structure of the data. The current standard approach is to prewhiten the data to remove the temporal autocorrelation, but another approach that was initially considered, precoloring, will also be discussed.
5.2.2 High-pass filtering

The most common approach for removing low-frequency trends is to apply what is known as a high-pass filter. One approach of high-pass filtering is to add a discrete cosine transform (DCT) basis set to the design matrix, such as the example shown in Figure 5.14. Figure 5.15 shows the original time series (top) and fit of the DCT basis functions to the data (green). The middle panel shows the original time series (blue) and the high-pass filtered time series using the DCT basis set (green), and the bottom panel shows the same, but in the Fourier domain. After high-pass filtering, the drifting at the beginning and end of the time series is removed. When using a DCT basis set, the highest frequency cosine function that should be included would correspond to the highest frequency that is desired to be removed from the data, which is chosen to avoid the frequency of the experimental task that is also being modeled. As a rough rule of thumb, the longest period of the drift DCT basis should be at least twice the period of an on–off block design.

Another approach to removing the low-frequency drift, which is used by the FSL software package, is to fit a locally weighted scatterplot smoothing, or LOWESS, model to the time series and remove the estimated trend from the data. A LOWESS model fits a local linear regression over a window of the data weighting points in the middle of the window rather than around the edges, for example using a Gaussian function as a weight. For details on the LOWESS model, see Cleveland (1979). The result is that the low-frequency trends in the data are picked up by the LOWESS fit, as shown in the top panel of Figure 5.15 (red). After this trend is fit to the data, it is subtracted, resulting in a high-pass filtered data set shown in the middle panel of Figure 5.15. The larger the window, the higher the frequencies that are filtered out of the data, so the window should be chosen to only remove frequencies in the low end of the spectrum far from the frequency of the task. Figure 5.15 shows that in the time domain, the DCT basis function and LOWESS approaches yield very similar results. In the spectral domain, however, the power of the data filtered using the DCT basis set drops to 0 as the frequency decreases, whereas the LOWESS fit has a more gentle roll off as frequency decreases.

![Figure 5.14. Discrete cosine transform basis set. The five lines correspond to the first five discrete cosine transform basis functions, starting with a constant term and then increasing the frequency of the cosine by a half period with each additional curve.](image-url)
5.2 The BOLD noise

Figure 5.15. Illustration of high-pass filtering. The top panel shows the original time series (blue) as well as the HP filter fit using a LOWESS curve (red) and a DCT basis set (green). In both cases the fit is very similar. The middle panel shows the original data (blue) as well as the filtered data using the two methods. The bottom panel shows the power spectrum of the original data (blue) as well as the filtered data, illustrating that the original time series exhibited low-frequency noise, which is removed by the high-pass filters. Note that the DCT basis set tends to have a sharper decrease at the lowest frequencies, whereas the LOWESS filter has a more gentle roll-off at the edges.
5.2.3 Prewhitening

After high-pass filtering, the fMRI time series are still temporally autocorrelated, where the correlation increases as the temporal proximity of two data points increases. As discussed in Appendix A, in order for the general linear model estimates to be unbiased and have minimum variance among all unbiased estimators, a very appealing quality in an estimate, the data going into the model cannot be correlated, and the variance at each time point must be identical. When data are temporally autocorrelated, prewhitening removes this correlation from the GLM prior to estimation. The prewhitening process is generally carried out in two steps. In the first step, the GLM is fit ignoring temporal autocorrelation to obtain the model residuals, which is the original data with all modeled variability removed. The residuals are then used to estimate the autocorrelation structure, and then model estimation is carried out after prewhitening both the data and the design matrix. In general, the GLM with the original data has the following form:

\[ Y = X\beta + \epsilon, \quad (5.2) \]

where \( Y \) is the BOLD time series, \( X \) is the design matrix, \( \beta \) is the vector of parameter estimates, and \( \epsilon \) is the error variance, which is assumed to be normally distributed with a mean of 0 and a covariance of \( \sigma^2 V \). The GLM estimates for \( \beta \) are only optimal when \( V \) is the identity matrix, but since the BOLD data are autocorrelated, \( V \) is not the identity; it may have nonzero off-diagonal elements representing cross-correlation and may have varying values along the diagonal, representing differences in variance across time. Prewhitening involves finding a matrix, \( W \), such that \( WW' = I_T \), where \( I_T \) is an identity matrix with the same number of rows and columns as the time series. Prewhitening involves premultiplying both sides of the GLM by \( W \) to give

\[ WY = WX\beta + W\epsilon, \quad (5.3) \]

where now the covariance of \( W\epsilon \) is \( \text{Cov}(W\epsilon) = \sigma^2 WW' = \sigma^2 I_T \). Therefore, the whitened model has independent error terms, and the GLM estimates from this model are optimal. More details can be found in Appendix A.

The reason for using the residuals is so that the prewhitening process only removes temporal autocorrelation present in the data that is not related to the task of interest; that is, it removes autocorrelation that is part of the noise rather than part of the signal. Although this process is straightforward, it depends on the model of the temporal autocorrelation being exactly correct. If the correlation is not removed perfectly, the assumptions of the GLM that the data are uncorrelated with equal variance for each time point will be violated, and hence inferences may be biased.
5.2 The BOLD noise

There are different models for BOLD noise that have been found to describe the correlation well. The simplest is known as the AR(1) model and simply assumes the variance of each time point is 1, and the correlation between pairs of data points decreases geometrically as the data are further apart (\( \text{cor}(y_i, y_{i+a}) = \rho^a \)). A slightly more complicated correlation model adds an additional variance parameter and is known as the AR(1) + white noise (AR(1) + WN). The added variance is the white noise. This is a special case of a more general correlation model known as the autoregressive, moving average or ARMA model (Box et al., 2008). Specifically the ARMA(1,1), which has one autoregressive parameter like the AR(1) with an additional moving average parameter. All three of these correlation models fit well to the 1/f trend found in the power spectra of fMRI time series. Another option is to use an unstructured covariance estimate, where the correlation for each lag in the time series is estimated separately. This tends to have more parameters than the three parameter AR(1) + WN, making it less biased, but the correlation estimates are less stable (more variable).

It is important that the correlation model used in fMRI strikes a balance between having enough parameters to describe the correlation structure accurately, but not having too many parameters, since this will cause the estimates to be more variable due to a lack of degrees of freedom. If a modeling approach produces estimates that are too variable, the typical solution to increase the degrees of freedom. Since the degrees of freedom are roughly equal to the number of data points minus the number of parameters being estimated, they can be increased by either increasing the amount of data (e.g., pooling data across multiple voxels) or decreasing the number of parameters being estimated by using a simpler correlation model. One drawback of estimating the temporal autocorrelation model by pooling across multiple voxels is that the temporal autocorrelation structure is not exactly the same for all regions of the brain (Worsley et al., 2002; Zarań et al., 1997). Likewise, using only a few parameters in the autocorrelation model may not capture the trend thoroughly. All software packages use a combination of both of these tactics. In SPM, for example, a simple AR(1) + WN model is fit globally, pooling information from all voxels found to be significant in a preliminary fit. The model captures the correlation structure well, and by using a global approach, the estimate is not highly variable. FSL, on the other hand, uses a slightly different approach with an unstructured autocorrelation estimate, meaning a nonparametric approach is used to estimate the correlation for each lag of the time series. Instead of having a global estimate of the correlation, FSL instead smooths the correlation estimate spatially. Additionally, since the correlations for high lags have very few time points, these estimates are highly variable. To solve this issue, FSL uses a Tukey Taper, which smooths the correlation estimate in the spectral domain to the point where correlations at high lags are set to 0. Typically around 6–12 parameters are used for this voxelwise correlation estimation.
5.2.4 Precoloring

Whether fMRI time series should be whitened or, the opposite, temporally smoothed, was at one point a contentious topic in the fMRI analysis literature. Even though it is theoretically optimal, prewhitening was not commonly used initially. This was in part due to inaccurate correlation models that gave biased variance estimates and could inflate false positive risk (Friston et al., 2000). Temporal smoothing or low-pass filtering, dubbed precoloring was initially considered as a better approach to modeling temporal autocorrelation as it did not have the same problems with bias that prewhitening had (prior to the employment of the regularization strategies discussed above).

The general idea behind precoloring is that since we do not know the true structure of the temporal autocorrelation, we impose more autocorrelation with a low-pass filter. While strictly the autocorrelation in the data is now a combination of an unknown intrinsic autocorrelation and the smoothing-induced autocorrelation, the latter swamps the former, and we assume the autocorrelation is known and equal to that from the low-pass filtering. With a known autocorrelation, the standard errors and the degrees of freedom can then be adjusted to eliminate bias and produce valid inferences. Figures 5.16 and 5.17 illustrate all types of filtering for blocked stimuli and random stimuli, respectively. The top panels show the original data in the time (left) and spectral (right) domains, the second and third panels show high-pass and low-pass filtering, and the bottom panel show bandpass filtering, which is both high and low-pass filtering. Bandpass filtering was introduced earlier (Section 2.4.2) in terms of spatial filtering of fMRI data. Just as high-pass filtering the data removes low-frequency trends in the data, low-pass filtering removes high-frequency trends in the data.

The major problem with low-pass filtering the data is that in many fMRI experiments the stimuli are presented in a random fashion, meaning that the signal covers a wide range of frequencies, including some high-frequencies. When data are low-pass filtered, the high-frequency components are removed from the data, which means signal is being removed from the data. As Figure 5.16 shows with a block design, since most of the power for the task is focused at lower frequencies, the high-pass filter does not remove the task-based trends from the data. On the other hand, Figure 5.17 shows that for an event-related study, the power for the task is spread across a wide range of frequencies and so the low-pass filter ends up removing some of the task signal. Because of this and also due to the development of better regularization approaches for prewhitening, precoloring is not a recommended solution for dealing with temporal autocorrelation.

5.3 Study design and modeling strategies

Depending on what the investigator is interested in measuring, different study designs and modeling strategies are required. On one end of the spectrum, the
5.3 Study design and modeling strategies

Figure 5.16. Illustration of different filtering approaches on data with blocked trials. The left column shows the result in the time domain, and the right column, in the Fourier domain. The first row is the original data followed by high-pass (HP), low-pass (LP), and bandpass (BP) filtered data, respectively. When trials are blocked, most of the task-based frequency is focused at one point (0.25 Hz in this case), while the rest is aliased to other frequencies. The HP filter removes low-frequency drift, as shown before, whereas the LP filter removes high-frequency trends. Applying both in the BP filtered case removes variability at both low and high frequencies but preserves the bulk of the signal at 0.25 Hz.
Figure 5.17. Illustration of different filtering approaches on data with randomly spaced trials. The left column shows the result in the time domain, and the right column, in the Fourier domain. The first row is the original data followed by high-pass (HP), low-pass (LP), and bandpass (BP) filtered data, respectively. Since the trials are spaced randomly in time, the task energy is spread over a wide range of higher frequencies. In the LP-filtered case, only low-frequency noise is removed, but in the LP and BP filtered cases, much of the higher frequency signal is removed by the filter, which will have a negative impact on our GLM inferences since there is little signal left to detect.
investigator may only be interested in detecting and estimating the magnitude of activation (referred to as detection). On the other hand, the investigator may be interested in estimating the precise shape of the hemodynamic response associated with the task (referred to as estimation). Not only are different models required to test each of these hypotheses, but different study designs are more appropriate for testing each; in fact, as we will see later, designs that are optimized to test for one of them will in general be relatively poor at testing the other. Additionally, when multiple task types are present in a study, some study designs are more optimal than others. These issues will be discussed in the following sections.

5.3.1 Study design: Estimation and detection

Imagine you have a small pile of sand on your kitchen floor (the size of a handful). From a distance it is easy to deduce that there is indeed a pile of sand on your floor. If you kneel down and look very closely at the sand without touching it you would be able to see the shapes of some of the granules that were on the outside of the pile, but you cannot study the shape of the granules in the middle. In this case we would say our ability to detect the presence of sand was quite high, but our ability to estimate the shape of the granules of the sand was not as good. Now imagine that you spread the sand around on the floor with your hand. Once it is spread out, you can look closely and see almost every side of each granule of sand and you can gather more information on the shape of the individual granule. On the other hand, if you stood up and backed away, it would not be as easy to detect that there was sand on the floor and where, specifically, the sand was. So, in this case our ability to estimate the shape is good, but our detection ability is poor.

This analogy can also be applied to trials in an fMRI experiment. If the investigator is mostly concerned with detecting the presence of a response, the best way to ensure this is to block the trials together. Note, when this is done, we only get a little bit of information about the shape of the HRF from the few trials at the beginning and the end of the block, whereas the ones in the middle do not tell us anything about the shape but do contribute to the ability to detect the response. On the other hand, if we separate the trials with a little bit of time, it is much easier to capture the specific shape of the HRF from our data. Note that if the trials have too much time between them and occur at regular intervals, there won’t be enough trials to estimate the HRF shape very well, and the study will be very boring for the test subject, which will decrease the quality of the data. The solution is to use randomly spaced trials. The GLM model is able to pick out the shape of the HRF through the FIR model, or other models described in Section 5.1.2, and since more trials are included the estimates will be less variable.

To strike a balance between estimation and detection, one suggestion is to use a semi-random design, which can be constructed by starting with the trials blocked and then randomly moving trials to break up the blocked shape (Liu et al., 2001). It should be noted that if a design has trials that are blocked as well as randomly
spaced, the blocked trials should be modeled separate from the spaced trials. This is because the mental processes for separated stimuli may be different than when the stimuli are rapidly presented and so the interpretation of the two types of stimulus presentation differ.

### 5.3.2 Study design: Multiple stimulus types

The previous section primarily concerns a study where there is only a single type of stimulus presented, but in most cases two or more stimulus or trial types are involved, and then it must be determined what order these stimuli will be presented and how much time will be between each stimulus. In some cases, the order of the stimuli is fixed, for example when a cue is given followed by a stimulus that the subject then responds to, the events always occur in the order of cue, stimulus, response. When regressors in a GLM are highly correlated, the estimates of the parameters are quite unstable and hence highly variable. Another way of stating that an estimate has high variability is to say it has low efficiency. Efficiency is calculated as the inverse of the parameter estimate's variance. In general, if the design matrix is given by $X$ and if $\beta$ is the vector of parameters to be estimated in the GLM, the covariance of the parameter estimates is given by

$$
\text{Cov}(\hat{\beta}) = (X'X)^{-1} \sigma^2
$$

(5.4)

where $\sigma^2$ is the error variance. (Here we neglect the whitening, but the equation still works if we replace $X$ with $WX$.) Design efficiency typically only refers to the variance due to the design, or $(X'X)^{-1}$, so for a given contrast of parameters, $c$, the efficiency for that contrast is defined as

$$
\text{eff}(c\hat{\beta}) = \frac{1}{c(X'X)^{-1}c'}
$$

(5.5)

In the simplest case, if there were two parameters in the model and the interest was in the efficiency of the first parameter, $\beta_1$ (corresponding to $H_0: \beta_1 = 0$) and so the contrast would be $c = [1 \ 0]$. There is not a meaningful cutoff for an acceptable level of efficiency, but instead the efficiencies for a variety of models can be estimated and then the most efficient model would be chosen. A measure related to efficiency with a more meaningful cutoff is the power of a statistical test. Since this calculation requires knowledge of $\sigma^2$, it is more complicated to estimate and isn’t necessary if one simply wants to rank designs to choose the most efficient one, since the most efficient design will by default also have the highest power. A more detailed description of statistical power for fMRI studies is given in Chapter 6.

The top panel of Figure 5.18 shows an example of two correlated regressors, where the timing for one stimulus (blue) always occurs 6 seconds prior to the presentation of a second stimulus (green). In this case, the correlation between the two regressors is very high ($\text{corr} = -0.61$), meaning the estimates of the parameters corresponding
5.3 Study design and modeling strategies

Figure 5.18. How jittering can improve model fit. In the top panel, the time between the first stimulus (blue) and the second (green) is always fixed at 6 seconds, whereas in the bottom panel the time between stimuli is randomly chosen between 1 and 10 seconds (so on average it is 6 seconds). The regressors in the top panel are more highly correlated ($\text{corr} = -0.61$); the correlation between regressors in the bottom panel is much smaller in magnitude ($\text{corr} = -0.15$). As a result, the design in the bottom panel is more efficient at estimating the magnitudes of each stimulus.

to each trial type are highly variable. The value of $(X'X)^{-1}$ for this fixed ISI design $(X_F)$ is given by

$$
(X_F'X_F)^{-1} = \begin{pmatrix}
0.5632 & 0.3465 \\
0.3465 & 0.5703
\end{pmatrix}
$$

so the efficiency for estimating $\beta_1$ (corresponding to the blue stimulus) would be $\frac{1}{0.5632} = 1.76$ and the efficiency for estimating $\beta_2$ (corresponding to the green stimulus) is $\frac{1}{0.5703} = 1.75$. In the bottom panel of Figure 5.18, the time between the first stimulus, and the second stimulus is jittered randomly, where the timing is sampled from a uniform distribution ranging between 1 and 11 seconds. Note, on average, the timing between the first and second trial types is equivalent to what is shown in the top panel (6 seconds). In this case, the correlation between regressors has a smaller magnitude of $-0.15$, and therefore the estimates for the parameters for each trial type are much less variable. In this case the model variance corresponding to the jittered design, $X_j$, is given by

$$
(X_j'X_j)^{-1} = \begin{pmatrix}
0.3606 & 0.0602 \\
0.0602 & 0.4640
\end{pmatrix}
$$

Note the off-diagonal entries, corresponding to the covariance, are much smaller in this case, and the diagonal elements, corresponding to the variances, are much smaller since there is less collinearity between the regressors. The efficiency for
estimating $\beta_1$ would be $\frac{1}{0.3606} = 2.77$ and the efficiency for estimating $\beta_2$ is $\frac{1}{0.4640} = 2.16$. Compared to the efficiencies of the fixed ISI designs, this random ISI design is 57% more efficient in estimating $\beta_1$ and 23% more efficient in estimating $\beta_2$. This relates back to the earlier discussion about orthogonality, since in the fixed ISI design regressors are correlated, there is little unique variance for each regressor to be used to estimate the corresponding parameter and hence the efficiency is lower.

### 5.3.3 Optimizing fMRI designs

When developing a new study design, one must consider a variety of designs and look at both the estimation and detection abilities of those designs and choose the design that is most suitable for the purposes of the study. Additionally, it is important to consider the psychological factors of the experiment. If the ordering of the stimuli is easy to predict there will be problems with habituation of the subject to the task. As more trial types are added to the study the number of possible designs, including different orderings of stimuli and timings between stimuli, is quite large and searching over all designs and checking the estimation and detection abilities of each is not a feasible task. Instead, it is helpful to develop a search algorithm that methodically chooses designs. The simplest example of a search algorithm is that of the permuted block design (Liu, 2004). As described earlier, when trials are blocked, the ability to detect the activation corresponding to the blocked stimuli is quite great, but the ability to estimate the shape of the HRF is not great. Additionally, blocked stimuli may not be psychologically interesting for the investigator to study. The permuted block design starts with the stimuli blocked by task type and then TRs are randomly chosen and the stimuli in the TRs are swapped, hence breaking up the blockiness of the design. This is then repeated, and after many iterations the stimuli will be randomly ordered, corresponding to a design where detection has decreased (from the original block design) and estimation of the HRF has increased. Designs along the continuum can be selected for the study, depending on the desired amount of detection and estimation, while keeping the study interesting for the subject.

Another approach to selecting the ordering of stimuli in an experiment is to use what are called maximal length sequences, or M-sequences (Liu & Frank, 2004; Liu, 2004). M-sequences are series of 0s and 1s that are maximally uncorrelated with shifted versions of themselves. Since an FIR model is nothing more than multiple copies of the stimulus stick functions shifted in time, this means that M-sequences will produces FIR models with covariates that have the least amount of correlation between them as possible, which will lead to highly efficient estimates of each lag of the FIR model. This also means that M-sequences are ideal for estimating the shape of the HRF, but not necessarily very good at building designs with high detection power for an assumed canonical HRF.

So far the discussion of optimal designs has not mentioned psychological factors of the design, like predictability and counterbalancing, and other arbitrary constraints
5.3 Study design and modeling strategies

on the design, like ensuring that equal number of trials are presented for different trial
types. Instead of just optimizing statistical properties, by maximizing a cost function
that is a combination of different design "fitness" properties, designs that have good
efficiency and psychological properties can be found. The cost function will be quite
complicated, however, making traditional optimization methods challenging. This
has motivated the use of evolutionary or "genetic algorithms" for finding event-
related fMRI designs that maximize arbitrary fitness criterion (Wager & Nichols,
2003). Recent work has improved upon this approach by using M-sequences to find
good initial guesses of designs (Kao et al., 2009).