The General Linear Model and fMRI: does love last forever?

Jean-Baptiste Poline[a,b], Matthew Brett[b]

[a]*Neurospin, Bat. 145, CEA, Gif-sur-Yvette, 91191, France*
[b]*Henry Wheeler Brain Imaging Center, 10 Giannini Hall, UC Berkeley Berkeley, CA 94720, USA*

Keywords: General Linear Model, functional MRI

Abstract

In this review, we first set out the General Linear Model (GLM) for the non technical reader, as a tool able to do both linear regression and ANOVA within the same flexible framework. We present a short history of its development in the fMRI community, and describe some interesting examples of its early use. We offer a few warnings, as the GLM relies on assumptions that may not hold in all situations. We conclude with a few wishes for the future of fMRI analyses, with -or without- the GLM. The appendix develops some aspects of use of contrasts for testing for the more technical reader.

Introduction

The General Linear Model (GLM) has been at the heart of functional Magnetic Resonance Imaging analyses for the past last 20 years. While methods have evolved rapidly, most of the papers published in the field use this technique, and it is likely that this will continue, for better or worse. The main reason for this is the conceptual simplicity of the GLM, the fact that it implements standard statistics used in biomedical research, and that it can provide some answers to most of the standard questions put to the data. While one may wish that the research community would be prompt to adopt new statistical frameworks – for instance the Bayesian framework – like many communities, it has stayed close to the familiar analyses that the GLM provides.

There seems to be something special about the GLM in fMRI. A PubMed search for “general linear model” in article titles finds neuroimaging papers in the large majority. Here we will argue that the GLM has a special flavor in functional imaging because the typical interface to the GLM in neuroimaging is close to the details of its implementation. This lends a “low level” flavor to the use of the GLM in fMRI.

In this review, we will first explain what the GLM is, in the form of a introductory tutorial. We hope our readers will not be offended by the basic level of our tutorial. Of course the GLM is very widely used; fMRI analyses are rarely done with any other technique except for resting state fMRI (although, resting state fMRI is most often analyzed with correlation or partial correlation, which can be seen as a special case of the GLM). The reason we have written our article in this way is that, despite its wide use, it is not always completely understood. While starting with the very basic, we hope that even those with some background in statistics will learn some useful facts (perhaps in the appendices). We will include some examples of cases that have been confusing, sometimes shamelessly stolen from other references (and cite our sources[1].) Most of these examples are not specific to fMRI, but the fMRI community or at least a good part of it has adopted a specific way of presenting the GLM in matrix form. This allows great flexibility and better understanding of the inner mechanism of the GLM, but it can be hard to see the relationship of the neuroimaging GLM with the same underlying machinery in other statistical tools. We will try to explain the different presentations and how they correspond.

[1] but only when those come from friends or important members of our community.

In the second section, we will briefly review the history of the introduction of the GLM into the fMRI community. In the third section, we offer a few examples to show how the GLM was used to answer some new questions arising in fMRI. Finally, we conclude with a few wishes for the future.

1. What is the general linear model?

To explain the general linear model, we start with an extremely basic review of notation for linear regression.

Linear regression for those dumber than dummies

Let us say that 9 authors are submitting papers for a special edition of NeuroImage. The edition is disastrously late and the editors of NeuroImage are interested to know why. The editors wonder whether older authors are submitting their articles later than younger authors.

Let's give the nine authors numbers $i = 1 \ldots 9$.

For each author we have:

y_i = Days after the deadline that author i submitted their article

x_i = age of author i

At a first pass the editors predict that there is a straight line relationship between number of days late y_i and age x_i.

Because the editors are rigorous they specify this as a *model*, thus:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

β_0 is the intercept (the number of days late for an author of age 0) and β_1 is the slope (the number of days late attributable to each year of someone's age). ϵ_i is the remaining unexplained data in y_i after subtracting $\beta_0 + \beta_1 x_i$.

At the moment we do not have a fully specified model, we have only rephrased our data to state that they are to be a sum of terms. For the model to be fully specified we need some assumption on ϵ_i. Let us say that we believe the values in ϵ arise as independent samples from a Gaussian distribution with zero mean and variance of σ^2. *Independent* means that knowing the value for ϵ_i gives you no further information about ϵ_j for any i, j. The values are *identically* distributed when each error ϵ_i arises from a Gaussian distribution with the same variance σ^2, regardless of i. Here our model of the errors is that they are zero mean, independent and identically distributed, written as $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The model is characterized *both* by the formula explaining the data ($y_i = \beta_0 + \beta_1 x_i + \epsilon_i$) and by the assumptions on ϵ.

We now have a statistical model expressed as a formula in terms of x_i. To get to the classic matrix formulation we apply a trick, which is to make a new vector of ones, called x_0, where $x_{0i} = 1$ for all i. We rename our previous x_i as x_{1i}:

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \epsilon_i \tag{2}$$

This is the same as model 1 above because $\beta_0 x_{0i} = \beta_0 \times 1 = \beta_0$. We can think of y, x_0, x_1, ϵ as four column vectors of length 9, and reformulate the model as matrix addition:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix}$$

Finally we reach the matrix formulation of the general linear model by rearranging the matrix addition as matrix multiplication:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \\ 1 & x_8 \\ 1 & x_9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix} \tag{3}$$

Call the y_i vector Y, call the stack of vectors x_0, x_1 the *design matrix* X, call the parameters β_0, β_1 the *parameter vector* β and call ϵ the *error vector*:

$$Y = X\beta + \epsilon \tag{4}$$

with ϵ a zero mean Gaussian noise. Returning to our late articles, we may also have the hypothesis that authors who are more widely cited are more relaxed about sending in their articles late. We get a measure of citation for each author and call this x_{2i} for author i. The statistical formula is:

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \tag{5}$$

In the matrix formulation of equation 4, the new term x_{2i} can be expressed as an extra column of x_{2i} values in X, and an extra parameter β_2 in the vector β. By extension the matrix formulation can handle any multiple regression model with p parameters (Scheffé, 1959, p. 4):

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \tag{6}$$

Equation 4 is the matrix formulation of equation 6.

In section 7.1 of the Appendix we describe how this can be used in the analysis of intra subject data. Next we show that we can include analysis of variance within the same framework.

The general linear model and the analysis of variance

Let us imagine that we do not believe that age or citation rate are important, but we do believe that authors from the USA, UK and France have different tendencies to submit their articles late. The first 3 authors (1..3) are from the USA, authors 3..6 are from the UK and authors 7..9 are from France. Maybe it is more acceptable to be late if you are from the UK (not France). Now we have a new model which looks like this:

$$y_i = \beta_1 + \epsilon_i, \text{ for authors i from the USA}$$
$$y_i = \beta_2 + \epsilon_i, \text{ for authors i from the UK}$$
$$y_i = \beta_3 + \epsilon_i, \text{ for authors i from France}$$

Let us assume again that the errors ϵ_i are independent. If we want to find $\beta_1, \beta_2, \beta_3$ that will result in the smallest sum of ϵ_i^2 values ($\sum_i \epsilon_i^2$), then it may be obvious that $\beta_1, \beta_2, \beta_3$ must be the means of the y_i lateness scores for US authors, the UK authors and the French authors, respectively.

The statistician writes the new model like this:

$$y_{ij} = \beta_j + \epsilon_{ij} \tag{7}$$

where the $j = 1..3$ index the group. In matrix form, this will look like:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix} \tag{8}$$

Let us call each column of X a *regressor*. We have three regressors; let us call these r_1, r_2, r_3. $r_{1i} = 1$ when the author number i is from the USA and zero otherwise. The value of r_{1i} therefore indicates membership of author i in the category "USA". r_1, r_2, r_3 are often called *indicator* variables (or regressors). They may also be called *dummy* variables (or regressors). Encoding group membership with dummy variables allows us to express analysis of variance and covariance in the framework of the linear model (equations 6 and 4).

As before, if our errors ϵ_i are independent, and we solve our equation to give us the β vector that minimizes $\sum_i \epsilon_i^2$, then the three entries of the β vector will be the means of the US, UK and French author late scores, as for equation 7.

If we decide we want to add back the effects of age to our model, we have analysis of covariance (ANCOVA). This is as simple as combining the columns of the design matrix of equation 3 and equation 8:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & x_1 \\ 1 & 1 & 0 & 0 & x_2 \\ 1 & 1 & 0 & 0 & x_3 \\ 1 & 0 & 1 & 0 & x_4 \\ 1 & 0 & 1 & 0 & x_5 \\ 1 & 0 & 1 & 0 & x_6 \\ 1 & 0 & 0 & 1 & x_7 \\ 1 & 0 & 0 & 1 & x_8 \\ 1 & 0 & 0 & 1 & x_9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix} \tag{9}$$

Note that we have added the constant column from the slope and intercept model (the first column of all 1). In fact this makes the design redundant, a point to which we return in the appendix 7.2.

What is the general linear model, really?

The term *general linear model* refers to a linear model of form $Y = X\beta + \epsilon$ (equations 6 and 4) in which:

- There may be analysis of variance coded with dummy variables
- The errors ϵ are drawn from a zero mean Gaussian distribution

In retrospect, it is surprising that the GLM was not introduced earlier, as it was already presented as such in several statistical textbooks. However, there seems to be some inertia in the teaching of statistics, and statistical techniques such as t-test, ANOVA or ANCOVA are still today sometimes taught separately.

The matrix formulation of multiple regression appears to date from 1935 with the work of Alexander Aitken (Aitken, 1935; Seal, 1967). Scheffé's classic monograph "The Analysis of Variance" has a section entitled "Deriving the formulas for an analysis of covariance from those from a corresponding analysis of variance." (Scheffé, 1959). In effect, this is the general linear model, although he does not name it as such. The least square technique itself was first published by Legendre in 1805 (Legendre, 1805), and further developed by Gauss, Laplace, and others (Lehmann, 2008). The importance of the normality assumption was emphasized by the works of Pearson and Fisher during the first half of the twentieth century in the development of inference procedures[2].

The general linear model is not always covered explicitly in introductory statistical texts. This may be because analysis of variance is easier to explain by showing direct subtraction of group means. However, the use of dummy variables in multiple regression gets its own section in introductory statistical texts at least as early as 1972 (Wonnacott and Wonnacott, 1972). In the preface to the third edition of "Statistics" by William L. Hays (Hays, 1981), Hays comments on the addition of a new chapter called "The general linear model and the analysis of variance":

> The ready availability of computer programs for multiple regression and for multivariate analysis generally is giving such methods a far more ubiquitous role in research then they formerly enjoyed. In order to understand and to take advantage of the many options these methods represent, the student needs some early groundwork in the general linear model, and especially in the essential connections between multiple regression and the analysis of variance.

This quote suggests another explanation for the visibility (or otherwise) of the general linear model. The general linear model in its matrix form is also the form in which a computer program is likely to solve the estimation problem. It seems likely to us that the general linear model in its matrix form will have a more direct explanatory appeal to those writing code to estimate the model parameters, because the solution can be very concisely expressed with matrices.

The `lm` linear model in the S / R programming languages implements the general linear model with a formula interface (Chambers et al., 1992). Other packages with explicit interfaces to the general linear model include SPSS, SAS, and minitab.

Estimating and interrogating the model

A quick summary on how to estimate the parameters β is presented in the appendix section (7.7). Once estimated on some data, the β are noted $\hat{\beta}$.

We can phrase questions of interest in terms of the model β values. Returning to model 9, one question might be if our data (here the number of days after the deadline) changes with age. In other words, from model 9: is $\beta_4 = 0$? (is the effect of age equals to zero? Note that the fifth column corresponds to β_4). From the same model we may want to know if the average number of days late differs between groups: is β_1 equal to β_2, and β_2 equal to β_3, or equivalently is $\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$?

When we ask questions such as $\beta_4 = 0$?, we know that we do not in fact have β_4 — the ideal parameter — but $\hat{\beta}_4$ — an estimate of the parameter. So, our tests take into account that $\hat{\beta}_4$ will have some error in its estimate, related to the size of the residuals $\hat{\epsilon}$ and the correlations between the columns of the design matrix[3] X.

More on the interpretation of the $\hat{\beta}$ can be found in the appendix section (7.2).

To summarize

- The GLM is . . . a model. The word model can have several definitions in different contexts but here this says that we know something about how our data (days late of article submission) are related to other data or knowledge (such as the age of a subject, or which country they come from). The model is the expression of the form of this belief with mathematical symbols.

- The GLM is "Linear". This simply refers to the belief (expressed in the model) that our data formed with addition or subtractions of weighted known data. In other words, this is a simple multiple regression, in which a series of numbers is approximated by the weighted sum of other numbers. The

[2] The term multiple regression is attributed to Pearson by the online `www.statsoft.com` which states that "The general purpose of multiple regression (the term was first used by Pearson, 1903) is to analyze the relationship between several independent or predictor variables and a dependent or criterion variable."

[3] The higher the correlations between the columns of X, the higher is the variance of the $\hat{\beta}$

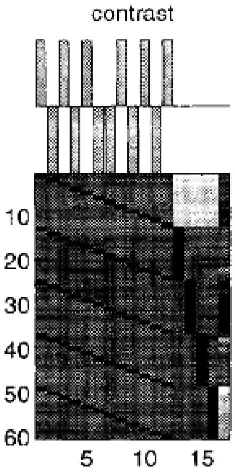

Figure 1: **The design matrix of Friston et al, 1995**: This is (probably) the first time the design was displayed as a design matrix in the neuroimaging literature (in this instance for PET data). It shows that the first implementation made the design matrix non-redundant by removing one column of the subject effects and centering these. Testing for the subject effect could simply be accomplished by a contrast built as the identity matrix aligned above the columns

data approximated are often called the dependent variable - and in our case will be article submission times - and the data from which it is approximated are called the regressors or concomitant variables (or sometimes, confusingly, the "independent" variables).

- "General". It is called general because the regressors are not necessarily measured data, but can code for the belonging to a group or to a condition. The GLM encompasses linear regression *and* ANOVA.

In the literature, the GLM also generally encompasses multivariate methods (severals measurements y per individual or line of X), therefore more than one column in the Y matrix. We chose not to discuss this aspect here to keep this review simple, but see for instance Worsley et al. (1997).

2. A brief and biased history of the development of the GLM in neuroimaging, and some lessons

The first fMRI papers comparing two experimental conditions could have phrased their methods in terms of the GLM, but it was not until 1995 that Friston et al (Friston et al., 1995) proposed the GLM as such to the functional imaging community.

In the implementations of SPM prior to SPM94, analyses such as analysis of variance, t-tests and regressions required separate routines. Once introduced, the GLM allowed SPM to simplify the code considerably. SPM is written in MATLAB that is a particularly good match for the GLM as the matrix expressions found in textbooks could be implemented directly, so that the code almost reproduces the formulas. The good match between the code and the formulas made the code easier to understand. This was an important factor for the rapid dissemination of the ideas.

The design matrix display was introduced in SPM94. At the time, Keith Worsley reported that he did not see the use of the GLM formulation as a great addition to the field of statistics, but he was enthusiastic about the idea of *displaying* the design matrix. Displaying the design as a matrix was not and is still not common in statistical packages, but the information contained in the visual representation makes it easier to think about possible comparisons that could be performed. Keith thought this was a significant addition to the field of applied statistics. Figure 1 shows the historic design matrix of (Friston et al., 1995).

As fMRI methods developed, it became clear that the temporal correlation of the signal was an issue: the noise could not be considered independent between scans, so models assuming independent errors did not apply. One idea at the time was to apply some temporal smoothing both to the data Y and to the design matrix X, so that the noise variance-covariance could be approximated by the effect of the applied smoothing filter rather than from the data, and the degrees of freedom corrected accordingly (or, so it was believed ...). This was not an efficient approach, and clearly the theory of the GLM was not completely mastered (including by the first author of this review). This was corrected quickly by Keith Worsley (Worsley and Friston, 1995), and soon the better solution of whitening the signal using an auto-regressive model was implemented by several in the field (Bullmore et al., 1996; Woolrich et al., 2001).

The flexibility of the GLM framework quickly resulted in ingenious ways to solve a variety of interesting problems. For instance, with the advent of event related fMRI, we needed flexible models for the hemodynamic response function (HRF). Up until this point the most common method of constructing an estimated fMRI time course was to convolve a time course of delta functions representing the event onsets with a canonical[4] HRF derived from the auditory or visual cortex BOLD observations. SPM96 offered the possibility of adding other regressors to the event model that could capture variation in the shape of the response. One popular set of functions was the Taylor-series expansion of the HRF, adding to the HRF its derivative in time and the derivative with respect to the parameter controlling for the width of the HRF (see the next section for some additional examples of the use of the GLM in fMRI). The unknown lag of the response and the variation of its shape could therefore be accommodated in a very simple manner[5]. The estimated parameters of basis functions could be used to estimate the delay of the HRF relative to the canonical response (Liao et al., 2002; Henson et al., 2002).

The GLM used to take a significant time to compute on long time series. In order to do an F-test we need to compare residual variance between a full and a reduced model. The early versions of SPM implemented the F-test by running the full and reduced model separately. We soon realized that, in fact, any desired test within the specified model could be implemented by computation involving only the parameters estimated from the full model. A decrease in computation time was one outcome; it was also a more principled way to design the F-test via contrasts in the space of the model parameters (Poline et al., 2007).

Roger Woods appears to have been the first in our field to realize that the variance estimates from models including multiple subjects that included all the scans from all subjects should not be used to test for group effects (Woods, 1996; Holmes and Friston, 1998). This was an issue in PET analyses from the early 1990's, but it became obvious to the neuroimaging community only when the number of scans per subject started to get close to 1000. In those cases the 32-bit address space of most CPUs became the limiting factor for analyses, motivating us to think harder about the right way to do group analyses. Woods (Woods, 1996) pointed out the distinction between across scan and across subject variance to attribute the correct degrees of freedom for population inference. Again, in retrospect, it is rather puzzling that we had not interacted with the statistical community enough to be aware of the issue sooner.

Identification of the several levels of variance led to the development of hierarchical models, variance component estimation, and Bayesian techniques. Here the fMRI community started to catch up with the standards in biostatistics (Friston et al., 2002; Beckmann et al., 2003; Woolrich et al., 2004). The implementation framework proposed for instance in SPM makes the GLM even more general, by parameterizing the variance-covariance as a linear combination of weighted symmetric matrices. These weights are then found using Restricted Maximum Likelihood in an Expectation Maximization scheme allowing correct inference on repeated measures designs. Some packages, such as FSL, implement a mixed effect model that takes the first level variance into account in the estimation of the group results. For instance, a subject with a high level of noise would have less influence on the group result than a subject with less noise[6]. AFNI also implements mixed effect analysis in specific command line programs such as `3dMEMA`.

[4] The word canonical here means the standard (most commonly used) model of the hemodynamic response, based on a combination of gamma functions

[5] In general this means comparing all the functions with an F-test when comparing two conditions. Since this leads to an unsigned test and is less easy to manipulate at the group level, researchers often only consider the parameter weighting the HRF function. This is suitable only if the other functions have similar weights between the conditions. If not, a more delayed response may appear weaker, even if its amplitude is actually higher

[6] Often the number of scans per subject is high and this makes the intra-subject variance negligible

To conclude this section, we believe we have learnt two main lessons from this short and biased review of recent history. First, a few missteps made while developing fMRI analysis techniques could have been avoided by closer interaction with the statistical community, and the field could have moved even more quickly. While the statistics community is unlikely to be impressed by techniques such as the GLM that have been considered standard for some time, many statisticians are interested in fMRI data for the challenge they offer and because they create pretty images of the brain at work. Through this interaction, we may be able to gather better techniques. The rider is always that, in order to be useful, the techniques have to be implemented in standard packages that can work with neuroimaging data, which is not always the case for the latest methods. Many of the most advanced techniques are still difficult to use, as the work to make these available to the community of researchers in cognitive neuroscience is not well recognized or evaluated.

Second, it is remarkable to see how the term "GLM" and the overall framework has impacted neuroimaging. The vast majority of articles found with a PubMed search for "General Linear Model" or "GLM" in the title are either from neuroimaging or are studies of a gene called GLM. Clearly, neuroimaging has embraced this framework following the Friston et al. (1995) paper, and there is little chance that the use of the GLM will disappear in the near future. Other techniques such as classification or machine learning are likely to get more and more used, but often these will start with some processing involving the linear model, or relate the results to those of the GLM. Inference will probably move to non-parametric statistics because of the richness of the tests that can be performed and the more exact estimation of the risk of errors, but the statistics computed are likely to be based on the GLM, because of its (in general) easy interpretation and fast computation.

3. A few examples on the early use of the GLM in fMRI - and some warnings

We have chosen a few examples on the way that the GLM was used in the early development of fMRI methodology to illustrate its versatility. There are many possible illustrations of this, and the following are the ones that first came to mind because of their simplicity and usefulness.

Modeling low frequency drifts

The first attractive and very practical use of the GLM was modeling of the low frequency noise found in the fMRI signal by the use of cosine functions. Engineering schools would traditionally teach their students how to implement a digital filter for electrical signals that are not short time series, but continuous signals sampled by analog-to-digital chips. These filters were not adapted to the shorter fMRI time series. In the GLM, low pass filtering simply amounts to including the low frequency to be removed in the model. For example, we can assemble a matrix L composed of column vectors containing cosine functions up to a certain frequency. By appending the L matrix to the design matrix X, we entirely decorrelate the data from these low frequencies in the fit. The interpretation then becomes very easy in terms of which frequencies or period of time has been removed. As the fMRI run can be quite long the L matrix could include many columns, and this produced design matrices with many parameters, also taking a lot of the space on the display of the design. An alternative but equivalent method is to pre-multiply both the design X and the data Y with a matrix $K = I - LL^{-}$ that projects onto the space orthogonal to L. Then, the structure induced in the noise ϵ has to be taken into account, as ϵ would then have KK^T (which is equal to K in this instance) as variance-covariance structure.

Including the estimated movement in the model

It has become common practice to include the estimated movement parameters in the design matrix as nuisance regressors. Clearly, a subject's movements induce considerable changes in signal at the voxel level, but the idea to include the movement parameter estimates in the design matrix was not necessarily straightforward as it is not clear that these changes are linear with the amount of translation in a particular direction or rotation about a particular axis. In fact, the signal induced by movement is not linear with the movement parameters. To approximate non-linear behavior, some models also include the square and sometimes the cube of movement parameters. The parameters can also shifted by one scan up and one scan down to model the effect of lag between the movement and signal. Which set of regressors is sufficient to remove most of the noise induced by actual movement is a difficult question that has no universal answer. In part, this is an empirical question that can be addressed with an F-test, but in general choosing the best set of regressors in a GLM is not yet common practice. Another solution often seen in the literature is to include time courses of regions that are particularly sensitive to movement as regressors in the model. For example it is common to include a regressor derived from the time course of signal in the ventricles.

Finite Impulse Response (FIR) and other basis functions

As we mentioned above, a set of basis functions gives some flexibility to better fit the effect of the Hemodynamic Response Function. One particularly interesting set of functions is the Finite Impulse Response (FIR) basis functions (Dale, 1999; Glover, 1999). This basis set is used in the GLM to obtain a noisy but simple estimate of the shape of the hemodynamic response after a given experimental condition. For instance if we want to estimate the HRF over 20 seconds and we have a TR of 2 seconds, the FIR basis set would consist of 10 regressors (per condition), each regressor consisting of a "1" for the lag at which one would like to estimate the response and "0" otherwise. Before this solution was devised, selective averaging was used and this led to complications when the conditions overlapped and were not balanced or entirely randomized across the run. The FIR model gives an elegant solution to a common problem. However, when the time series are not very long and the number of conditions is large, this leads to a large number of parameters to estimate. Also, the experimental conditions can be partly correlated, and in these two situations the estimation can be very noisy. As usual, we do not usually know *a priori* the correct period over which we should estimate the HRF. More recent work (also based on linear models) with priors or regularization schemes provide reasonable solutions to this estimation problem (see for instance (Ciuciu et al., 2003), amongst several). Correct estimation of the HRF and ways of incorporating this estimation into the detection step without increasing the risk of false positives has become an important sub-field of fMRI. Models with FIR basis functions are not usually adequate for detection purposes - meaning that the large number of parameters makes the estimation noisy enough that it can be difficult to detect significant signal related to particular event types (conditions).

Psychophysiological Interactions: "PPI"

This technique is yet another clever use of the GLM (Friston et al., 1997). Imagine that your experimental paradigm[7] is made up of blocks of 30s of condition A: "think", and 30s of condition B: "do not think"[8]. Now extract the time series $R(t)$ from your favorite brain region R, multiply this time series by -1 only during condition (B), call this $PPi(t)$. What are the regions showing correlation during A and anti-correlation during B with your favorite region? To get the answer, include the constructed regressor in X, also include $R(t)$, and test within each voxel for the β corresponding to $PPi(t)$. This map should show regions that correlate differently with R during A compared to B.

GLM for simultaneous EEG and fMRI

As a last quick example, the first analysis of simultaneous EEG and fMRI used the following strategy. First, the EEG signal was preprocessed to remove the huge artefactual signal induced by the gradients during Echo Planar Imaging, a step that can use the GLM (Vincent et al., 2007). Second, the power of the alpha band (8-12 Herz) was calculated within a short and sliding time window, and this measure was convolved with the HRF (or with a set of basis functions to model variation of the HRF). This convolved signal was then sampled at each TR to form a regressor, which was finally included in the model X (see for instance (Goldman et al., 2000; Laufs et al., 2003), amongst others). Many epilepsy researchers also investigated how to best estimate the HRF after inter-ictal spikes, evoked response potentials (ERPs), or spectral EEG variations, using the GLM or techniques derived from it.

[7] The experimental paradigm is the set and timing of experimental conditions in this context.

[8] Some cognitive psychologists would object to these experimental conditions, but the authors believe that with sophisticated statistical techniques and enough data they should be able to find some differences between the two conditions

Take-home message

There are an infinite number of clever ways to use the GLM to answer questions with fMRI data, it would be hard to list them all.

Reflecting on our previous examples, it seems that the GLM has given us ways to find quick solutions for many common problems in neuroimaging. To summarize, the GLM is 1) conceptually simple 2) readily available in standard packages 3) an incredibly flexible tool 4) implements the standard statistical testing framework 5) does not require heavy computation and can be implemented in only a few lines of MATLAB or Python code. This combination is so attractive to methodologists and neuroimagers that is almost impossible to resist.

However, the solution provided can be "too quick". Generally, a better method can be designed that may include a more complex model with regularization terms, and variable or dimension selection with cross validation. Often the computation of these more complex solutions is iterative and may involve statistical sampling. These better-adapted solutions may already exist, but the software packages to use them are not very easy to install or use. The GLM in its simple form has – for better or worse – bright days ahead and for the foreseeable future.

The GLM relies on assumptions. First the matrix X should contain the appropriate regressors, whether at the first (across scan) or at the second (typically, across subject) level. Too few, or too many regressors (or put differently, a design matrix space that is too large or too narrow) will lead to either loss in sensitivity or in specificity. Second, the normality assumption should hold. This is difficult to check at each and every voxel, but see the work of many that have implemented non parametric tests, robust techniques or outlier detection in neuroimaging ((Holmes et al., 1996), followed by many). Third, the assumptions on the variance covariance structure of the noise should hold as well. These assumptions are difficult to check. Lehmann (2008) summarizes the history of the linear model:

> In the 1920's it was developed in a new direction by R.A. Fisher whose principal applications were in agriculture and biology. Finally, beginning in the 1930's and 40's it became an important tool for the social sciences. As new areas of applications were added, the assumptions underlying the model tended to become more questionable, and the resulting statistical techniques more prone to misuse.

Summary measures and visualization tools for the residuals of the model are therefore crucial, and again, too rarely seen in current mainstream packages. A good review of the assumptions and pitfalls of the GLM can be found in (Monti, 2011).

In any case, it seems clear that it will be valuable for researchers using fMRI to master the swiss army knife tool of the GLM, and understand its assumptions and limitations.

4. Conclusions: three wishes and one prediction

We propose three wishes and one prediction about the GLM and the future of neuroimaging.

- Wish N^o1. We wish that assumptions made by the GLM would be more often checked. This implies that more people would use the diagnostic tools such as (Zhang et al., 2006), and that there should be more development of these tools. To be used more often, the diagnostic tools should be part of the main analysis packages. We wish there was more time spent on checking the results and less on rushing to publish, but we do not expect this to happen any time soon.

- Wish N^o2. There are still improvements that can be made in the current implementation of the GLM, at least in the packages we know well. For instance, in SPM, while the term σ^2 is specific to each voxel to allow for the magnitude of the noise to change, the AR coefficient and the structure of the covariance are assumed to be identical across "activated"[9] voxels, to speed up computations and allow

[9] The variance structure is estimated across the voxels for which the effects of interest in the model are "significant" at the α level 0.001

the use of random field theory. There is still room for improvement in the current packages, but as this work is not necessarily fancy or exciting it is rarely undertaken. Another direction of improvement is to develop tools to help in choosing the appropriate model (see (Kherif et al., 2002) for an attempt). The standard approach is to assume the model X itself is constant across regions of the brain despite apriori knowledge that this should not in general be the case.

- Wish N^o3. We wish that models that match the data better than the GLM would be used more often, and therefore that software packages proposing those would be more widely used. In particular, non parametric, robust and Bayesian procedures could be employed more often. This often involves longer and more complicated calculations, and the methods are harder to explain to reluctant reviewers, who prefer to see methods they understand even when they may not be right for the data. Better teaching and tighter interactions between research communities are needed.

- Prediction. We predict the GLM will still be around after the middle-aged authors of this paper are long gone.

To conclude, the GLM has provided us with an extraordinary range of solutions to problems in the field and yielded a wealth of results, and will continue to do so in the future. Let's hope that this venerable giant will not prevent the development of more complex but better techniques in fMRI.

5. Acknowledgements

We are deeply grateful for the careful reading and excellent suggestions from P.J. Toussaint.

6. References

Aitken, A., 1935. On least squares and linear combinations of observations. Proceedings of the Royal Society of Edinburgh 55, 42–48.

Beckmann, C. F., Jenkinson, M., Smith, S. M., Oct. 2003. General multilevel linear modeling for group analysis in FMRI. NeuroImage 20 (2), 1052–63.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., Feb. 1996. Statistical methods of estimation and inference for functional MR image analysis. Magnetic Resonance in Medicine 35 (2), 261–77.

Chambers, J., Hastie, T., et al., 1992. Statistical models in S. Chapman & Hall London.

Christensen, R., 2002. Plane answers to complex questions. the theory of linear models.

Ciuciu, P., Poline, J.-B., Marrelec, G., Idier, J., Pallier, C., Benali, H., Oct. 2003. Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. IEEE transactions on medical imaging 22 (10), 1235–51.

Dale, A. M., Jan. 1999. Optimal experimental design for event-related fMRI. Human brain mapping 8 (2-3), 109–14.

Friston, K., Holmes, A., Worsley, K., Poline, J.-B., Frith, C., Frackowiak, R., 1995. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapping 2, 189–210.

Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., Dolan, R. J., Oct. 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6 (3), 218–29.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., Jun. 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16 (2), 465–83.

Glover, G. H., Apr. 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9 (4), 416–29.

Goldman, R. I., Stern, J. M., Engel, J., Cohen, M. S., Nov. 2000. Acquiring simultaneous EEG and functional MRI. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 111 (11), 1974–80.

Hays, W., 1981. Statistics, 3rd Edition. CBS college publishing.

Henson, R. N. A., Price, C. J., Rugg, M. D., Turner, R., Friston, K. J., Jan. 2002. Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations. NeuroImage 15 (1), 83–97.

Holmes, A. P., Blair, R. C., Watson, J. D., Ford, I., Jan. 1996. Nonparametric analysis of statistic images from functional mapping experiments. Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism 16 (1), 7–22.

Holmes, A. P., Friston, K. J., 1998. Generalisability, Random Effects and Population Inference. In: NeuroImage. Vol. 7. p. S754.

Kherif, F., Poline, J.-B., Flandin, G., Benali, H., Simon, O., Dehaene, S., Worsley, K. J., Aug. 2002. Multivariate model specification for fMRI data. NeuroImage 16 (4), 1068–83.

Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., Krakow, K., Aug. 2003. EEG-correlated fMRI of human alpha activity. NeuroImage 19 (4), 1463–76.

Legendre, A. M., 1805. Nouvelles Méthodes pour la Détermination des Orbites des Comètes. International review of neurobiology, 72–75.

Lehmann, E. L., 2008. On the history and use of some standard. Institute of Mathematical Statistics Collections 2, 114–126.

Liao, C. H., Worsley, K. J., Poline, J.-B., Aston, J. A. D., Duncan, G. H., Evans, A. C., Jul. 2002. Estimating the delay of the fMRI response. NeuroImage 16 (3 Pt 1), 593–606.

Monti, M. M., Jan. 2011. Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. Frontiers in human neuroscience 5, 28.

Poldrack, R. A., Mumford, J. A., Nichols, T. E., 2011. Handbook of Functional MRI Data Analysis. Cambridge University Press.

Poline, J.-B., Kherif, F., Pallier, C., Penny, W., 2007. Contrasts and classical inference. . Elsevier: Amsterdam. In: Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.), Statistical parametric mapping: the analysis of functional brain images. Elsevier, Amsterdam, pp. 126 – 139.

Scheffé, H., 1959. The Analysis of Variance, first edit Edition. John Wiley \& Sons, New York.

Seal, H., 1967. Studies in the history of probability and statistics. xv the historical development of the gauss linear model. Biometrika 54 (1-2), 1–24.

Vincent, J. L., Larson-Prior, L. J., Zempel, J. M., Snyder, A. Z., May 2007. Moving GLM ballistocardiogram artifact reduction for EEG acquired simultaneously with fMRI. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 118 (5), 981–98.

Wonnacott, T., Wonnacott, R., 1972. Introductory statistics, 2nd Edition. "John Wiley & Sons".

Woods, R. P., Dec. 1996. Modeling for intergroup comparisons of imaging data. NeuroImage 4 (3 Pt 3), S84–94.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., Smith, S. M., Apr. 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21 (4), 1732–47.

Woolrich, M. W., Ripley, B. D., Brady, M., Smith, S. M., Dec. 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–86.

Worsley, K. J., Friston, K. J., Sep. 1995. Analysis of fMRI time-series revisited–again. NeuroImage 2 (3), 173–81.

Worsley, K. J., Poline, J. B., Friston, K. J., Evans, A. C., Nov. 1997. Characterizing the response of PET and fMRI data using multivariate linear models. NeuroImage 6 (4), 305–19.

Zhang, H., Luo, W.-L., Nichols, T. E., May 2006. Diagnosis of single-subject and group fMRI data with SPMd. Human brain mapping 27 (5), 442–51.

7. Appendix

7.1. The intra subject example.

This a “regression” example. If you measure the signal during an Echo Planar Imaging sequence, you may want to see how it correlates with the experimental paradigm presented to the subject during scanning. Your belief is that in some regions of the brain, the stimulation will induce some neural activity, the neural activity will induce some hemodynamic response, therefore the data (time series) acquired $y(t)$ is correlated with an ideal — noise free — response $x(t)$. In other words, $y(t)$ is equal to this perfect response $x(t)$ times a weighting coefficient β, plus some random noise $\epsilon(t)$.

The model is therefore $y(t) = \beta x(t) + \epsilon(t)$, with $\epsilon(t)$ following a normal distribution with zero mean, (if the distribution is not Gaussian we are in the case of a *generalized* linear model), and β is the weight on the ideal response[10]. If more than one response $x(t)$ is expected (there are several conditions in the experimental paradigm), other terms are added and the model is a multiple linear regression model.

7.2. Interpretation of the β and reparametrization of the model

Working directly with the parameter estimates is a powerful framework that allows us direct access to any questions that can be put to the model. However, interpretation of the $\hat{\beta}$ is not always straightforward. We take the simple example of model (1) to show this.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10}$$

[10] It is important to see that here β is an “ideal” parameter. With some specific data, the *estimation* of β, denoted $\hat{\beta}$ will take some specific value. But if the experiment is repeated, the data y will be different, and the estimation of $\hat{\beta}$ will differ. On average, we should have $y(t) = x(t)\beta$.

The parameter β_0 can be interpreted as the number of late days for an author of age 0, while the β_1 is the number of late days attributable to each year of someone's age (the slope of the regression). Let us center the age regressor around zero, i.e. remove the mean age from each value, with: $\tilde{x_i} = x_i - \sum_i x_i/n$. Now replace x by $\tilde{x}$ in X, then the parameter β_0 will be interpreted as the number of days late at the average author's age, which may be a more useful value. The parameter β_1 in this reparametrized model does not change, and is still the number of days late attributable to each year of someone's age. Removing the mean of x is "decorrelating", in other words "orthogonalizing" x with respect to the first column of ones. Note that with this model parametrization $\tilde{X} = [\mathbf{1}\ \tilde{x}]$, the estimation of the parameter β_1 does not change, but the estimation of β_0 does. The test on $\hat{\beta}_1$ does not change, but the test on $\hat{\beta}_0$ does.

Let us take again the example of model (9) with a constant term, indicator variables coding for the three groups (US,UK,FR), and the covariate age x_i. Our data are still the time of submission after deadline. We have the following design matrix:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & x_1 \\ 1 & 1 & 0 & 0 & x_2 \\ 1 & 1 & 0 & 0 & x_3 \\ 1 & 0 & 1 & 0 & x_4 \\ 1 & 0 & 1 & 0 & x_5 \\ 1 & 0 & 1 & 0 & x_6 \\ 1 & 0 & 0 & 1 & x_7 \\ 1 & 0 & 0 & 1 & x_8 \\ 1 & 0 & 0 & 1 & x_9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix} \tag{11}$$

We first note that in this model the parameters are not uniquely defined, as the first column is the sum of the second, third and fourth columns. In that case, the parameters $\hat{\beta}_{1,2,\mathrm{or}\,3}$ are only defined up to an arbitrary additive constant term. The only interesting questions that can be put to the model involving the second through the fourth columns will take the form of the difference between parameters $\hat{\beta}_{1,2,\mathrm{or}3}$, eliminating the constant term, for instance: $\beta_1 - \beta_2 = 0$? The difference between "US" versus "mean of UK and FR" would be written as: $\beta_1 - (\beta_2 + \beta_3)/2 = 0$.

If on average the "UK" and "FR" authors are older than the US authors — their countries are — then we may find our test is not significant because the difference in the Y between the two groups may largely be explained by the difference in age of authors from the old continent compared to the new continent. This is correct, as we don't want to attribute to the factor "country" those differences that could in fact be due to "age".

One could have written model (11) differently to eliminate the redundancy, for instance removing the first column. In that case, we could test for the first parameter, which would be interpreted as the days of submission after deadline for an hypothetical "US" author of age 0, but this is hard to interpret and in general you may be more interested in the difference between the group means rather than whether the absolute value of the mean of a particular group is different from zero.

Another solution would be to code for the *difference* between "US" versus "UK", and the difference "UK" versus "FR"[11]. Each parameter of this new version is estimable, and can be interpreted as the difference of the means of the groups, *accounting for* age, but the (interesting) comparison "US" versus ("UK" and "FR") would be more difficult to formulate. It is often easier to use a simple parametrization and set up a valid contrast.

Note that these are only three versions of the **same** model: a re-parametrization that makes it more or less easy to interpret the parameters, and test for some effect. However, the three models are the same: the fitted data, the estimated error and the information that can be extracted from the models are identical in the three cases.

See Poldrack et al. (2011) for a similar example with more detailed[12] explanations and using a model that includes interaction terms (we highly recommend this reference).

[11] This is the parametrization that R would use.

[12] and possibly clearer

To conclude this section, while simple, the different possible parametrization of the model and the corresponding contrasts and tests that can be put to the model can be a little tricky to interpret. The most crucial aspect to remember is that whenever an effect is tested, all other effects present in the model are are removed from the data, whether these are confound variables or other effects that may be of interest, such that the tested effects are "decorrelated" from any other modeled effect. Again, this can be thought as testing the full model against a reduced model that contains all non-tested effects. For a deeper understanding of this, appendix sections (7.5 to 7.9) should help.

7.3. Notation

Y	:	Data	The $(n, 1)$ time series, where n is the number of time points or scans. y_i : one of those measures.
λ, Λ	:	Contrast	vector or matrix of weights (λ or Λ) or of the parameter estimates used to form the (numerator) of the statistics (denoted c in some other references)
X	:	Design matrix or design model	the (n, p) matrix of regressors .
β	:	Model parameters	The true (unobservable) coefficients such that the weighted sum of the regressors is the expectation of our data (if X is correct)
$\hat{\beta}$	:	Parameter estimates	The computed estimation of the β using the data Y : $\hat{\beta} = (X^TX)^-X^TY$
$C(X)$	:	Vector space spanned by X	Given a model X, the vector space spanned by X are all vectors v that can be written as $v = X\lambda$
P_X	:	The orthogonal projector onto X	$P_X = X(X^TX)^-X^T$
R_X	:	Residual forming matrix	Given a model X, the residual forming matrix $R_X = I_n - XX^-$ transforms the data Y into the residuals $r = R_XY$, in the space orthogonal to X
σ^2V	:	scan × scan covariance	This (n, n) matrix describes the (noise) covariance between scans

7.4. GLM, a quick summary

The GLM can be written in matrix form with:

$$Y = X\beta + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The best linear unbiased estimate of $E(Y)$ is the ordinary least square solution:

$$\begin{aligned} \hat{Y} = X\hat{\beta} &= X(X^TX)^-X^TY = P_XY, \text{ with } P_X = X(X^TX)^-X^T \\ \text{we have: } \hat{\beta} &= (X^TX)^-X^TY \end{aligned} \tag{12}$$

The residual noise is estimated with:

$$e = Y - \hat{Y} = Y - P_XY = R_XY, \text{ with } R_X = I - P_X$$

If the error is not independent or identically distributed, but follows $\epsilon \sim \mathcal{N}(0, \sigma^2 V)$, the best solution is to "whiten" the data by pre-multiplying Y and X by K chosen such that $KVK^T = I$, and the model is:

$$KY = KX\beta + K\epsilon \text{ , with } K\epsilon \sim \mathcal{N}(0, \sigma^2 I) \text{ , since } \mathrm{Var}(K\epsilon) = \sigma^2 KVK^T = \sigma^2 I$$

This model is then solved as above, leading to the estimate:

$$\hat{Y} = X\hat{\beta} = X(X^TV^{-1}X)^-X^TV^{-1}Y \tag{13}$$

7.5. Subspaces

Let us consider a set of p vectors x_i of dimension $(n, 1)$ (with $p < n$), such as regressors in an fMRI analysis. The space spanned by this set of vectors is formed of all possible vectors (say u) that can be expressed as a linear combination of the x_i : $u = \alpha_1 x_1 + \alpha_2 x_2 + ...\alpha_p x_p$, or equivalently, there is a $(p, 1)$ vector α such that $u = X\alpha$. If the matrix X is formed with the x_i : $X = [x_1 \; x_2 ... \; x_p]$, we note this space as $\mathcal{C}(X)$.

All vectors v such that the correlation of v with any of the x_i is zero, are said to belong to the space orthogonal of X, denoted by: $v \in \mathcal{C}(X)^\perp$

Not all the x_i may be necessary to form $\mathcal{C}(X)$. The minimal number needed is called the rank of the matrix X. If only a subset of the x_i are selected, say that they form the smaller matrix X_0, the space spanned by X_0, $\mathcal{C}(X_0)$ is called a subspace of X. A contrast defines two complementary subspaces of the design matrix X : one that is tested and one corresponding to the reduced model.

7.6. Orthogonal projection

The orthogonal projection of a vector x onto the space of a matrix A, is a vector x_p, such that 1) x_p is formed with a linear combination of the columns of A, and 2) x_p is as close as possible to x, in the sense that the sum of squares of $x - x_p$ is minimal. The projector onto A, denoted P_A, is the matrix such that $x_p = P_A x$. P_A is unique and can be computed with $P_A = AA^-$, with A^- denoting the Moore-Penrose pseudo inverse[13] of A. For instance, in appendix section 7.7, the fitted data $\hat{Y}$ can be computed with:

$$\hat{Y} = P_X Y = XX^- Y = X(X^T X)^- X^T Y = X\hat{\beta} \tag{14}$$

Most of the operations needed when working with linear models only involve computations in the parameter space, as shown in equation 17. For a further gain, if the design is degenerate, one can work in an orthonormal basis of the space of X. This is how the SPM code is implemented.

7.7. Parameters of the model are not necessarily unique: estimability issues.

If the design matrix X has some redundancy in one or several of its columns can be formed from a linear combination of the other columns. Two models X_1 and X_2 are equivalent (and in some sense, are equal) if any column of X_1 can be formed with a combination of columns of X_2. In other words, their vector space — the set of linear combinations of their columns - are equal. We denote the space of X as $\mathcal{C}(X)$.

If X has some redundancy, the minimal set with which the columns of X can be formed has fewer vectors (columns) than the number of columns in X, and the parameters β that can be constructed by combinations of others are not uniquely defined. Why is this important? This points to the fundamental aspect of the GLM: the parameters are not necessarily the most important aspect of the model, it is the set of vectors that can be formed with the columns of X. What model (4) in fact means is that the expectation of Y is $X\beta$, therefore a linear combination of the columns of X: $E(Y) = X\beta$. So if we want to *estimate* $E(Y)$ from the Y in the least squares sense, we will need to look for the orthogonal projection of Y onto X (see appendix 7.6 for the definition of projectors). If X has some redundancy, the betas are computed (for instance) using the Moore-Penrose pseudo inverse ($\hat{\beta} = (X^T X)^- X^T Y$).

Estimability. The only interesting information we can get from the model is from the rows of $E(Y)$, or from a linear combination of those. The form of an estimable function is therefore $h^T E(Y)$, or $h^T X\beta$. In practice, the vector h is rarely seen, but it can be seen from this that a contrast of the parameters of the model will have the form $\lambda = h^T X$, and therefore is a linear combination of the rows of X. For instance, the SPM package allows you to test linear combinations of the β, but would only accept those that are estimable (linear combinations of the rows of X).

[13] Any generalized inverse could be used.

7.8. Testing effects with the model parameters is equivalent to testing a reduced model

Testing whether there is an effect of some factor or of some covariate in the model is specified by setting a linear combination of the β to zero, or testing whether it is positive/negative. The null hypothesis is therefore that, for instance, $\lambda^T\beta = 0$, or $\lambda^T\beta > 0$. In the "formula" framework, it may be difficult to ask questions about the difference between the individual levels (for instance if a factor has three levels, what is the difference between level 1 and level 3?)[14].

To reject (or accept) the null hypotheses stated below in the case where λ involves only one linear combination, the standard t or F statistic can be formed:

$$t_{df} = \lambda^T\hat{\beta}/\mathrm{SD}(\lambda^T\hat{\beta}) \tag{15}$$

where $\mathrm{SD}(\lambda^T\hat{\beta})$ denotes the standard deviation of $\lambda^T\hat{\beta}$ and is computed from the variance

$$\mathrm{var}\big[\lambda^T\hat{\beta}\big] = \hat{\sigma^2}\lambda^T(X^TX)^-\lambda \tag{16}$$

For independent Gaussian errors t_{df} follows a Student distribution with degrees of freedom given by $df = \mathrm{tr}\big[R_X\big] = n - p$ with n the number of rows in X and p the number of independent regressors in X[15]. At the voxel level, the value t_{df} is tested against the likelihood of this value under the null hypothesis.

If a number of columns in X have to be tested conjointly, as in the ANOVA example, the F statistics can be formed by testing the sum of squares of the residuals under the reduced model (the model without the regressors corresponding to the effects being tested) compared to the sum of squares of the residuals under the full model.

$$\begin{aligned} F_{\nu_1,\nu_2} &= \frac{(Y^T(I-P_{X_0})Y - Y^T(I-P_X)Y)/\nu_1}{Y^T(I-P_X)Y/\nu_2} && (17)\\ &= \frac{(SSR(X_0) - SSR(X))/\nu_1}{SSR(X)/\nu_2} && (18)\end{aligned}$$

with $SSR(X)$, $SSR(X_0)$, the sum of squares for error of models X, X_0 respectively, and:

$$\begin{aligned} \nu_1 &= \mathrm{tr}\,(P_X - P_{X_0}) = \mathrm{tr}\,(R_0 - R_X) && (19)\\ \nu_2 &= \mathrm{tr}\,(I - P_X) = \mathrm{tr}\,(R_X) \end{aligned}$$

While very useful for the understanding of an F test as the difference of sum of squares (this is probably how most fMRI packages implemented the F test), there are two serious limitations with this formulation. First, it requires that we compute and fit two linear models. When dealing with thousands of voxels, and sometimes long time series (and multiple runs), the computational cost was very high, and clearly limited the number of interesting comparisons that cognitive neuroscientists would happily do. The other limiting aspect is that if we want to test several linear combinations of the β, such as for instance if $\beta_1 = \beta_2$, and $\beta_2 = \beta_3$, these may not necessarily be columns of X directly. The next section explains how to go from a contrast of interest to testing the associated reduced model.

7.9. What is the reduced model associated with a specific contrast?

We assume from now on that the contrast of parameters we would like to test may concern multiple constraints on the β, and will denote those constraints by $\Lambda^T\beta = 0$.

[14] Those questions should usually come after it has been seen that the factor has an effect.

[15] When the variance covariance V cannot be inverted easily but an estimation is known, the degrees of freedom are approximated (Worsley and Friston, 1995)

As seen above, setting $\Lambda^T\beta = 0$ is also setting $H^T X\beta = 0$ for some matrix H (the number of columns in H is the number of columns in Λ). If we have a (legitimate) contrast of the parameters, what does H look like? How does this relate to a reduced model? Why is all this relevant for fMRI? It is interesting to understand what H is. While Λ puts some constraints on the parameters of the model, H is the equivalent constraint on the space of X Christensen (2002). The reduced model that is to be tested is $Y = X\beta + \epsilon$ **and** $H^T X\beta = 0$, or, equivalently, that $E(Y) \in C(X)$ **and** $E(Y) \in C(H)^\perp$ where $C(H)^\perp$ is the space orthogonal to H ([16]). The reduced model should therefore be a matrix X_0 such that $X_0 \in C(X) and X_0 \in C(H)^\perp$.

By choosing $H = X^{T-}\Lambda$, we impose the constraint on X that corresponds to $\Lambda^T\beta = 0$. Since $H^T E(Y) = H^T X\beta = \Lambda^T X^- X\beta = \Lambda^T\beta = 0$, we have $E(Y) \in C(H)^\perp$ with this particular choice of H. Also, $H \in C(X)$, so H is a matrix that imposes a constraint on X within the space of X, and is the part of X that is being tested. The part of X that is not being tested, the reduced model, can be found easily by projecting the orthogonal space of H onto X , so we have $X_0 = P_X R_H = P_X(I - P_H) = P_X - P_H$. Testing for $\Lambda^T\beta = 0$ is equivalent to testing for the reduced model $Y = X_0\gamma + \epsilon$ with X_0 as defined above. Having defined X_0 as above, we have $P_X = P_{X_0} + P_H$.

Now, the numerator of the F test 17 can be rewritten in a much simpler way, as a function of Λ only:

$$\begin{aligned} Y^T(P_H)Y &= Y^T X(X^T X)^- X^T H(H^T H)^- H^T X(X^T X)^- X^T Y/r(\Lambda) && (20)\\ &= \hat{\beta}^T\Lambda(\Lambda^T(X^T X)^-\Lambda)^-\Lambda^T\hat{\beta}/r(\Lambda) && (21)\end{aligned}$$

Where $r(\Lambda)$ is the rank of Λ, and the F test can be rewritten as:

$$F_{\nu_1,\nu_2} = \frac{\hat{\beta}^T\Lambda(\Lambda^T(X^T X)^-\Lambda)^-\Lambda^T\hat{\beta}/\nu_1}{MSE} \qquad (22)$$

where the MSE is the mean square error $SSR(X)/\nu_2$.

[16] This is the set of vectors with zero correlation with any vector of H