Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 20 (2003) 1346–1355

www.elsevier.com/locate/ynimg

Bayesian second-level analysis of functional magnetic resonance images

Jane Neumann* and Gabriele Lohmann

Max–Planck–Institute of Cognitive Neuroscience, Stephanstrasse 1a, D-04103 Leipzig, Germany

Received 17 April 2003; revised 1 July 2003; accepted 14 July 2003

Abstract

We propose a new method for the second-level analysis of functional MRI data based on Bayesian statistics. Our method does not require a computationally costly Bayesian model on the first level of analysis. Rather, modeling for single subjects is realized by means of the commonly applied General Linear Model. On the basis of the resulting parameter estimates for single subjects we calculate posterior probability maps and maps of the effect size for effects of interest in groups of subjects. A comparison of this method with the conventional analysis based on *t* statistics shows that the new approach is more robust against outliers. Moreover, our method overcomes some of the severe problems of null hypothesis significance tests such as the need to correct for multiple comparisons and facilitates inferences which are hard to formulate in terms of classical inferences.

1. Introduction

The most widely used methods for the statistical analysis of fMRI data are based on a linear model of the hemodynamic response function (Friston, 1994). The detection of activated voxels is realized most commonly by means of statistical null hypothesis significance tests (NHST) based on frequentist or classical[1] *t* and *F* statistics (Worsley and Friston, 1995; Ardekani and Kanno, 1998). These tests are typically performed on different levels of analysis. On the first level, values indicating the significance of an effect are obtained for individual subjects. On the second level of analysis, statistical tests facilitate multisubject or multisession comparisons.

In recent years Bayesian techniques have been introduced to the field of functional MRI, providing a powerful alternative to linear modeling and NHST (Frank et al., 1998; Kershaw et al., 1999; Genovese, 2000; Høojen-Søorensen et al., 2000; Gössl et al., 2001a, 2001b; Friston et al., 2002a, 2002b; Marrelec et al., 2003). These methods are aimed at a complete Bayesian analysis of the functional data on all levels of analysis. They facilitate Bayesian models for the estimation of the hemodynamic response and apply Bayesian inference for the detection of functional activation in both single subjects and groups of subjects.

In this article, we propose an efficient new method that applies Bayesian techniques to the second level of analysis only. Our approach does not require a computationally expensive fully Bayesian analysis and modeling of the hemodynamic response on the first level. Rather, modeling on the level of single subjects is based on classical least-squares estimates of parameters for the General Linear Model (GLM). These parameter estimates are then viewed within a Bayesian framework as evidence for the presence or absence of some effect of interest in a group of subjects on the second level of analysis.

Traditionally, the analysis of functional MRI data has been viewed *either* within the frequentist *or* within the Bayesian framework. Our new approach is to draw on methodologies from each framework in different parts of the analysis. It is important to be clear that using Bayesian techniques on the second level of analysis does not presup-

* Corresponding author. Fax: +49 (0)341 9940 221.
E-mail address: neumann@cns.mpg.de (J. Neumann).

[1] Following other authors such as Friston et al. (2002b) we will use the terms “classical” or “conventional” to refer to analysis methods based on a Fisher statistic and a frequentist interpretation of probability, although Bayesian methods were developed much earlier in the history of science.

1053-8119/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S1053-8119(03)00443-9

pose Bayesian modeling and parameter estimation on the level of single subjects. Neither entails the classical treatment of the GLM frequentist methodologies on the level of inference. In our proposed method we view modeling and parameter estimation for single subjects within the classical, and statistical inference for groups of subjects within the Bayesian framework. This way we combine the relative simplicity of the GLM on the first level with the power and flexibility of Bayesian inference on the second level of analysis. Most notably, after parameter estimation of the GLM for single subjects, computation times for the Bayesian second-level analysis are under 10 s for a typical-sized group of subjects on a standard UNIX workstation.

The analysis of groups of subjects is of particular importance to the statistical evaluation of fMRI data. It has been shown by a large number of studies that results for individual sessions can vary considerably from session to session and from subject to subject (Aguirre et al., 1998; Miezin et al., 2000; McGonigle et al., 2000; Neumann et al., 2003). Single session results have thus to be treated with caution, as they represent only a sample of a subject's response. Consequently, only the analysis of groups of subjects allows meaningful generalizations to expected activations in the population. This requires statistical inference methods for groups of subjects that reflect the commonalities in the responses of different subjects while, at the same time, being robust against differences between subjects caused by their neuroanatomical or physiological variability.

Bayesian approaches are attractive alternatives to classical analysis methods, because they overcome a number of severe shortcomings of NHST. *P* values resulting from NHST describe the estimated probability of obtaining the observed data provided the null hypothesis of zero activation for an effect of interest is true. Consequently, sufficiently small *P* values are used to reject the null hypothesis of zero activation. It is in the nature of the test that given a large enough sample size, *P* will always be small enough to reject the null hypothesis. On the other hand, the alternative to the null hypothesis can never be rejected. In other words, although we are testing against the hypothesis of zero activation, the test does not allow us to infer that no activation has occurred. The somewhat counterintuitive way of reading the results of a NHST has more than once led to false interpretations of the observed data (Krueger, 2001; Gigerenzer, 1993; Oakes, 1986). In contrast, Bayesian inference provides a means of directly assessing the probability for an effect of interest to take on a certain range of values. For example, it allows us to infer the probability that a contrast between two experimental conditions is larger than zero.

Equally problematic for the application of NHST is the need for adjusting *P* values according to the search volume of a test statistic in order to account for the multiple comparison problem. This problem arises when repeatedly applying a *t* test to assess the significance of activation in different voxels. The threshold indicating significance increases with the number of examined voxels (Friston et al., 2002b). The required adjustment implicates that inferences about some part of the brain depend on whether other parts have been inspected, which is not very plausible. The need to address this multiple comparison problem does not arise from the application of Bayesian inferences to individual voxels. The probability of activation in one cortical area is independent of the inspection of other cortical regions.

Finally, in addition to the localization of activated cortical areas, more complex research questions such as the detection of functional dependencies or the analysis of the temporal behavior of the BOLD response become more and more the focus of ongoing research in fMRI. Such questions which are often hard or impossible to formulate in terms of a traditional NHST can be directly addressed using Bayesian inference.

The main principles of Bayesian inference and the GLM which are essential for the further understanding of the article are summarized in the following section. A comprehensive introduction to the GLM in the context of fMRI data analysis can be found in Friston et al. (1994); Friston (1994); Worsley and Friston (1995); Zarahn et al. (1997) and Lohmann et al. (2001). Excellent introductions to Bayesian data analysis and Bayesian inference in statistical analysis are provided by Gelman et al. (2000) and Box and Tiao (1992).

2. Methods

General Linear Model

For the General Linear Model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{1}$$

with data vector **Y,** design matrix **X,** and independently and identically normally distributed errors ε, it can be shown that the sampling distribution of the least-squares estimates for the parameters $\hat{\beta}_i$, $i = 0,1, \ldots, k$, is normal with $E(\hat{\beta}_i) = \beta_i$ and $\mathrm{Var}(\hat{\beta}_i) = \sigma^2 c_{ii}$, where c_{ii} are the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$ (Seber, 1977). An extended GLM accounting for serial autocorrelation in the observation data is

$$\mathbf{KY} = \mathbf{G}\beta + \mathbf{K}\varepsilon, \tag{2}$$

where **K** is a convolution matrix using a Gaussian kernel and $\mathbf{G} = \mathbf{KX}$ is the convolved design matrix. Here the variance of $\hat{\beta}$ extends to

$$\mathrm{Var}(\hat{\beta}) = \hat{\sigma}^2\mathbf{G}^+\mathbf{V}(\mathbf{G}^+)^\mathbf{T}, \tag{3}$$

where $\mathbf{V} = \mathbf{KK}^+$, $\mathbf{G}^+ = (\mathbf{G}^\mathbf{T}\mathbf{G})^{-1}\mathbf{G}^\mathbf{T}$ is the so-called *Moore-Penrose inverse* of **G,** and $\hat{\sigma}^2$ is an unbiased estimator for the variance σ^2. After fitting the linear model to the observed data, effects of interest can be expressed by means of *contrasts* $\mathbf{c}\hat{\beta}$ which are linear combinations of the parameter estimates. The vector **c** is a set of weights that usually sum to zero. In case of a single effect $\mathbf{c} = 1$. The estimated variance of a contrast is

$$\mathrm{Var}(\mathbf{c}\hat{\beta}) = \mathbf{c}\mathrm{Var}(\hat{\beta})\mathbf{c}^{\mathbf{T}} = \hat{\sigma}^2\mathbf{c}\mathbf{G}^{+}\mathbf{V}(\mathbf{c}\mathbf{G}^{+})^{\mathbf{T}}. \tag{4}$$

It is important to note that the calculation of contrasts up to this point does not involve any statistical evaluation of the data. A contrast merely contains information about how one or more covariates correspond to the experimental design. However, contrast images can serve as input for further statistical analysis. Within a classical framework, for example, statistical parametric maps SPM{t} can be constructed from contrast images by conducting a one sample *t* test that assesses the null hypothesis of zero response. Obtained *t* values are typically transformed into *z* values, giving a SPM{z} which, in order to detect and visualize activated voxels, is often thresholded at a commonly accepted but arbitrary level, e.g., $z = 3.09$, corresponding to $P = 0.001$. In our approach, we use contrasts obtained for single subjects and their respective estimated variances as input to Bayesian inferences over groups of subjects on the second level of statistical analysis.

Bayes' theorem

Bayesian inference rests upon the posterior probability distribution of model parameters given some observed data. For a model parameter θ with the probability distribution $p(\theta)$ and the observed data *y*, the posterior probability distribution of the parameter given the data $p(\theta|y)$ can be calculated according to Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{5}$$

$p(\theta)$ is called the prior probability distribution or simply *prior* of the parameter θ, expressing our initial knowledge or belief about the value of the parameter. $p(\theta|y)$ is called the posterior probability distribution or *posterior* of the parameter, expressing our belief about the parameter in the light of evidence from the data *y*. Since the data *y* are known and the true parameter θ is unknown, it is more convenient to express the conditional probability distribution $p(y|\theta)$ as the *likelihood function* of θ for given data *y*, which is written as $l(\theta|y)$ (Box and Tiao, 1992). Moreover, $p(y)$ is a function of the known data only and is constant with respect to the parameter θ. Equation (5) can thus be reformulated as

$$p(\theta|y) \propto l(\theta|y)p(\theta). \tag{6}$$

In words, the posterior probability distribution for a parameter θ given the data *y* is proportional to the product of the distribution of θ prior to the data and the likelihood of the parameter given the data. Bayes' theorem thus provides a mathematical means of combining previous knowledge with new evidence. This becomes particularly apparent when it is applied in an iterative way. Assume some parameter θ and some initial data y_1. According to Bayes' theorem, the posterior of θ can be expressed as

$$p(\theta|y_1) \propto l(\theta|y_1)p(\theta). \tag{7}$$

For a second observation y_2 with a distribution independent of y_1 we can state

$$\begin{aligned} p(\theta|y_2,y_1) &\propto l(\theta|y_2)l(\theta|y_1)p(\theta) \\ &\propto l(\theta|y_2)p(\theta|y_1) \end{aligned}. \tag{8}$$

The posterior calculated in Eq. (7) plays the role of the prior in Eq. (8). The new posterior can in turn serve as prior in a subsequent step, as new data come in. For this iterative process the posterior probability distribution can be easily calculated, if both prior and likelihood are normally distributed. If we suppose *a priori* that the parameter θ is distributed as

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\theta-\theta_0}{\sigma_0}\right)^2\right], \quad -\infty < \theta < \infty, \tag{9}$$

and the likelihood function of the parameter is

$$l(\theta|y) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta-y}{\sigma_1}\right)^2\right] \tag{10}$$

for an observation *y*, then it can be shown that the posterior distribution of θ is

$$p(\theta|y) = \frac{(\sigma_0^{-2}+\sigma_1^{-2})^{1/2}}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\sigma_0^{-2}+\sigma_1^{-2}) \times (\theta-\bar{\theta})^2\right], \quad -\infty < \theta < \infty, \tag{11}$$

which is the Normal distribution $N[\bar{\theta}, \bar{\sigma}^2]$ where

$$\bar{\theta} = \frac{1}{\sigma_0^{-2}+\sigma_1^{-2}}(\sigma_0^{-2}\theta_0 + \sigma_1^{-2}y) \tag{12}$$

$$\bar{\sigma}^2 = (\sigma_0^{-2}+\sigma_1^{-2})^{-1}. \tag{13}$$

The full proof is given in Box and Tiao (1992). See also Lee (1997) and Gelman et al. (2000) for discussion. Note that the resulting posterior mean can be interpreted as weighted average of the prior mean and the observed data, with weights proportional to the inverse variance.

Putting things together

Given the normal sampling distribution of the parameter estimates $\hat{\beta}_i$, i = 0,1, . . . , k, in the GLM, we can use the formalism described above to infer about the mean of a contrast in a group of subjects. Parameter estimates obtained for single subjects can be combined in an iterative process outlined by Eq. (7) to (13), given the same underlying model specification for all subjects. After establishing some prior which represents our initial belief about the mean of the contrast in the group of subjects, we can view the parameter estimates obtained from individual subjects as "data" or evidence modifying this belief. As prior we use the probability distribution of the contrast estimated for a

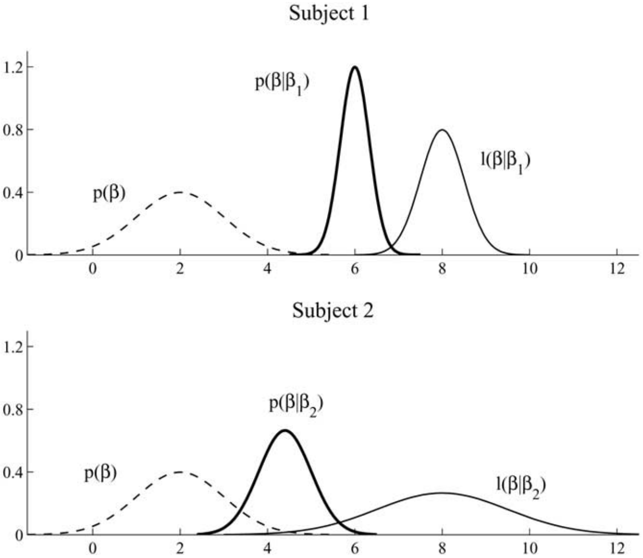


Fig. 1. The likelihood functions $l(\beta|\beta_1)$ and $l(\beta|\beta_2)$ representing data from two subjects are combined with the same prior $p(\beta)$. The observed data have the same mean but different variances. Data with small variance (Subject 1) have a larger influence on the prior than data with larger variance (Subject 2). Consequently, the posterior $p(\beta|\beta_1)$ is shifted more toward the mean of the observed data than $p(\beta|\beta_2)$, and the variance of $p(\beta|\beta_1)$ is smaller than that of $p(\beta|\beta_2)$.

randomly chosen subject. In other words, we assume that the contrast of a randomly chosen subject is a good representative of the mean contrast in the entire group. This initial prior is then combined with the estimates for the contrast of a second subject, according to Bayes' theorem. Thus, we are updating our belief which is based on the observation of the first subject by evidence from the second subject.

More specifically, given the contrast of interest estimated for two subjects, $\mathbf{c}\hat{\beta}_1$ and $\mathbf{c}\hat{\beta}_2$, with respective variances $\sigma_1 = \mathrm{Var}(\mathbf{c}\hat{\beta}_1)$ and $\sigma_2 = \mathrm{Var}(\mathbf{c}\hat{\beta}_2)$, the posterior probability distribution of the combined contrast is the Normal distribution $N[\overline{\mathbf{c}\hat{\beta}}, \bar{\sigma}^2]$, where

$$\overline{\mathbf{c}\hat{\beta}} = \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}}(\sigma_1^{-2}\mathbf{c}\hat{\beta}_1 + \sigma_2^{-2}\mathbf{c}\hat{\beta}_2) \tag{14}$$

$$\bar{\sigma}^2 = (\sigma_1^{-2} + \sigma_2^{-2})^{-1} \tag{15}$$

This probability distribution reflecting evidence from two subjects then serves as prior for the subsequent step of the iteration, where it is combined with the contrast estimated for the next subject. The result of the iterative process is the posterior probability distribution of the weighted mean effect for the whole group of subjects.

If the number of subjects is known in advance, the iterative process can be replaced by a single step. For normally distributed estimated contrasts of interest $\mathbf{c}\hat{\beta}_i$ with respective variances $\sigma_i = \mathrm{Var}(\mathbf{c}\hat{\beta}_i)$, $i = 0,1, \ldots, k$, obtained from k single subjects, the posterior of the combined contrast is the Normal distribution $N[\overline{\mathbf{c}\hat{\beta}}, \bar{\sigma}^2]$, where

$$\overline{\mathbf{c}\hat{\beta}} = \frac{\sum_k \sigma_k^{-2}\mathbf{c}\hat{\beta}_k}{\sum_k \sigma_k^{-2}} \tag{16}$$

$$\bar{\sigma}^2 = \frac{1}{\sum_k \sigma_k^{-2}}. \tag{17}$$

These equations allow a very clear interpretation of the result. The mean of the posterior is the sum of the means of the individual parameter estimates weighted by their respective inverse variance. The resulting variance represents the pooled within-subject variance and is a measure of our certainty about the population mean. It further becomes obvious that the choice of the subject for the initial prior and the order of the remaining subjects do not influence the result of the iterative calculation of the posterior.

It is an important point to note from Eq. (14) and (15) that the influence of each individual subject on the posterior for the entire group is determined by the estimated variance of the contrast specific to this subject. More precisely, the smaller the within-subject variance, the larger the influence of this subject on the posterior. This is illustrated in Fig. 1, where the same prior $p(\beta) \sim N(2,1)$ is combined with parameter estimates β_1 and β_2 from two subjects, whereby $l(\beta|\beta_1) \sim N(8,0.5)$ and $l(\beta|\beta_2) \sim N(8,1.5)$ for Subject 1 and Subject 2, respectively. The relatively small variance of β_1 causes the mean of the posterior to move toward the mean of β_1 and also results in a considerable decrease of variance from the prior to the posterior of β. The resulting posterior is $p(\beta|\beta_1) \sim N(6, 0.33)$, calculated using Eq. (12) and (13). In comparison, the posterior for the second subject is $p(\beta|\beta_2) \sim N(4.4, 0.6)$. The larger variance of β_2 causes the mean of the posterior to move less toward the mean of β_2 and the posterior still shows a relatively high variance. This result is intuitively plausible, since the estimated variance of the parameters represents the stability of the corresponding contrast in the obtained measurement. The smaller this variance, i.e., the higher the stability of the measurements, the higher should our certainty be about the observed effect and, consequently, the more should the observed data influence or correct our belief about the true value of the effect of interest.

Bayesian inference

In the most simple case of a second-level analysis of fMRI data we wish to make inferences about the presence or absence of an effect of interest in a group of subjects. Such effect is usually the activation corresponding to an experimental condition or a contrast, i.e. the difference in activation between two conditions. With a NHST the null hypothesis of zero activation or zero contrast is assessed based on estimated contrasts for individual subjects. The P values resulting from the test describe the estimated probability of obtaining these individual contrasts provided the null hypothesis of zero activation is true. If this probability is small enough, we reject the null hypothesis. The test does not tell us, however, how certain we can be that the effect is present in the group, if the null hypothesis was rejected. This question can be directly addressed using Bayesian inference.

Bayesian inference is based on posterior probability distributions. From the posterior mean and variance of an effect of interest we can calculate the probability for the effect to take on a range of values. This provides us with a measure of certainty about the effect in the group. For any probability distribution function *f*(θ) the probability that the random variate θ takes on a value in the interval [a,b] is

$$P(a \leq \theta \leq b) = \int_a^b f(\theta)d\theta. \tag{18}$$

For example, given estimated contrasts[2] $\hat{\beta}^1,\hat{\beta}^2, \ldots, \hat{\beta}^n$ of n individual subjects, the probability of a positive contrast β for the entire group is

$$P(\beta > 0) = \int_0^\infty p(\beta|\hat{\beta}^1\hat{\beta}^2, \ldots, \hat{\beta}^n)d\beta. \tag{19}$$

Note that while the mean β of the posterior tells us something about the size of the effect of interest, the posterior probability enables us to mathematically express the strength of evidence for the effect (Genovese, 2000). The ability to calculate this probability thus facilitates statements such as "Given the evidence of the observed data, we believe that in the investigated region the contrast is positive with 95% probability." Such statements very directly address the question of localizing regions with stimulus-related activation in the brain, and they are a much more intuitively plausible interpretation of the observed data than the rejection of a null hypothesis.

By means of Bayesian inference we can also tackle more complex research questions that are hard or impossible to formulate in terms of traditional hypothesis testing. This has been demonstrated, for example, for the comparison of activation amplitudes in different voxels (Frank et al., 1998) or the assessment of monotonicity of experimental conditions (Genovese, 2000). In the latter case, using NHST in order to assess the monotonicity of, say, four parameter estimates $\beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4$ requires the repeated application of single tests for the hypotheses $\beta_2 - \beta_1 \geq 0$ and $\beta_3 - \beta_2 \geq 0$ and $\beta_4 - \beta_3 \geq 0$. The resulting *P* values have in turn to be corrected for multiple comparisons. In contrast, within a Bayesian framework the posterior probability distributions for the parameters β_1, β_2, β_3, and β_4 allow the direct computation of P ($\beta_1 \leq \beta_2 \leq \beta_3 \leq \beta_4 \mid y$).

When examining the results of fMRI experiments, we are often interested in comparisons of different groups of subjects, for example, left-handed and right-handed participants of our study or subjects of different age or sex. The posterior distributions resulting from our analysis can directly be used to infer about differences between the means of two groups of subjects. Given the two probability distributions $p(\beta_1|\hat{\beta}^1 \ldots \hat{\beta}^k)$ and $p(\beta_2|\hat{\beta}^{k+1} \ldots \hat{\beta}^n)$ for a contrast of interest in two groups of k and n − k subjects, respectively, and assuming independence of the two groups, the joint distribution of the contrast is

$$p(\beta_1,\beta_2|\hat{\beta}) = p(\beta_1|\hat{\beta}^1, \ldots, \hat{\beta}^k)p(\beta_2|\hat{\beta}^{k+1}, \ldots, \hat{\beta}^n), \tag{20}$$

with $\hat{\beta} = \{\hat{\beta}^1, \ldots, \hat{\beta}^k,\hat{\beta}^{k+1}, \ldots, \hat{\beta}^n\}$. The posterior of the difference in means $d = \beta_2 - \beta_1$ is then the correlation of the two independent distributions (Frank et al., 1998). For normally distributed $p(\beta_1) \sim N(\beta_1,\sigma_1^2)$ and $p(\beta_2) \sim N(\beta_2,\sigma_2^2)$ this correlation has the analytical form

$$p(d|\hat{\beta}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{d-\mu}{\sigma}\right)^2\right], \tag{21}$$

with $\mu = \beta_2 - \beta_1$ and $\sigma = (\sigma_1^2 + \sigma_2^2)^{1/2}$ (Box and Tiao, 1992; Frank et al., 1998). The probability for a difference in means between groups can then be calculated by integrating the posterior distribution in the respective interval. Other tests for differences in means such as comparisons of different contrasts in a single subject or within one group of subjects can be calculated in the same manner.

3. Experimental results

Posterior probability distributions can be summarized and visualized in various ways. Maps of the posterior mean of an effect provide estimates for the effect size in every voxel. Posterior probability maps reflect the probability with which we can expect the effect to be found in a population. As one is usually interested in both, how large is the effect of interest and how likely is it to occur, we will in the following present maps of the posterior means together with the corresponding posterior probability maps.

Our method was tested on data obtained for a previous experiment on differences between the left and the right occipital cortex in response to spatial cueing (Pollmann and Morillo, 2003). Twelve subjects were first presented with a small or big visual cue on the left or the right side of a screen. Cue presentation was followed by the presentation of a target either in the cue location, i.e., in the same visual hemifield as the cue (valid trials), or in the contralateral visual hemifield (invalid trials). Subjects were instructed to focus their attention to the cued area while fixating a marker in the center of the screen. After the presentation of the target, subjects had to perform a simple target discrimination task.

The obtained data were processed with the software LIPSIA (Lohmann et al., 2001). This software package contains tools for preprocessing, registration, statistical evaluation, and presentation of fMRI data. In LIPSIA sta-

[2] For the sake of simplicity we will in the following refer to both a single parameter and a contrast of parameters as β.

Table 1
Talairach coordinates, z values, posterior means, and standard deviations of the posteriors for five cortical regions

ROI	Location	z value (NHST)	Mean	Standard deviation
1 Lingual gyrus	R (13 −83 −4)	4.36	0.23	0.02
2 Lingual gyrus	L (−14 80 −1)	−4.79	−0.25	0.02
3 Lateral occipital gyrus	R (31 −70 2)	3.83	0.10	0.01
4 Lateral occipital gyrus	L (−38 −70 2)	−4.51	−0.09	0.01
5 IPS/TOS	R (28 −71 23)	4.13	0.12	0.01

Note. Centers of the most significant activations for the contrast between valid-left and valid-right trials. z values resulting from a classical analysis and means and standard deviations from the second-level Bayesian analysis are shown. While z values are comparable for all regions, differences between the lingual gyri and the remaining areas can be found for the posterior means and standard deviations.

tistical evaluation is implemented as a two-stage random-effect analysis (Holmes and Friston, 1998) based on the least-squares parameter estimation of a GLM for serially autocorrelated observations (Friston, 1994; Worsley and Friston, 1995; Zarahn et al., 1997). More specific information on the model as well as the experimental design, hypotheses, and detailed results of the classical analysis can be found in Pollmann and Morillo (2003). The Bayesian second-level analysis described above was performed on the parameter estimates from the first stage of the random-effect model. The method was implemented in C and computations were performed with an AMD Athlon(TM) XP1800+ processor and 768MB working memory. Computation time for the complete second-level analysis was under 10 s for each contrast including the extraction of individual contrast means and variances from the results of the first-level analysis.

For the most prominent contrast, valid-left against valid-right trials, the pattern of activation obtained from the classical analysis was replicated with our Bayesian approach. The most significant activations were found in the left and right lingual gyrus, in the lateral occipital gyri of both hemispheres, and in the junction of the right intraparietal sulcus and the transverse occipital sulcus (IPS/TOS). For the centers of activation in these regions, coordinates in the Talairach stereotactic space (Talairach and Tournoux, 1988), z values obtained from the classical analysis,[3] and means and standard deviations of the posteriors from the Bayesian analysis are listed in Table 1. The regions are visualized in Fig. 2. The top row of the figure shows the posterior mean values of the contrast obtained from the Bayesian second-level analysis (left) and the posterior probability maps for $P\,(\mathbf{c}\beta > 0)$ (middle) and $P\,(\mathbf{c}\beta < 0)$ (right). A threshold of 99.9% was applied to the probability maps for better visualization. The bottom row shows the corresponding SPM{z} from the classical analysis threshold at $z = 3.09$.

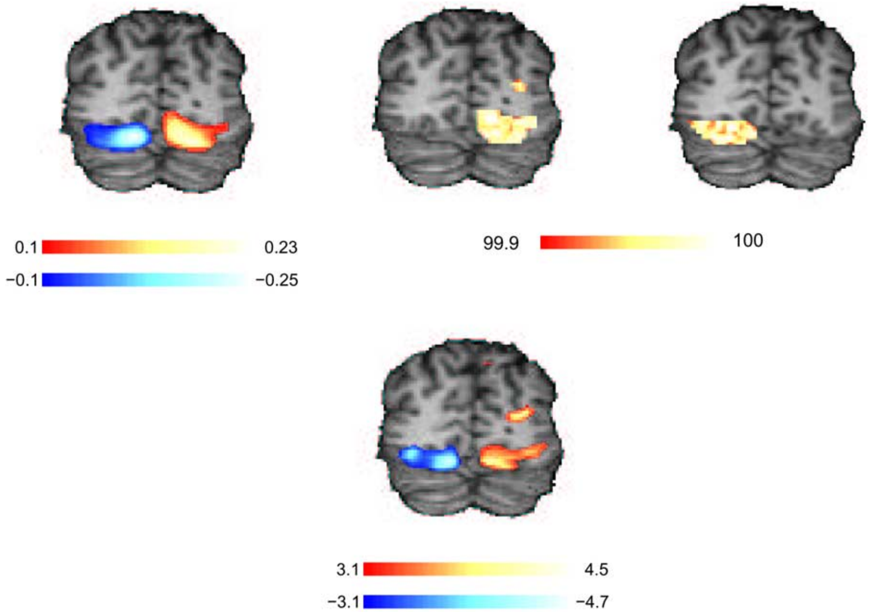
As can be seen, the posterior probability maps correspond well with the SPM{z}. $P(\mathbf{c} > 0)$ and $P(\mathbf{c} < 0)$ exceed 99.9% in regions with significant positive and negative z values, respectively. However, whereas the SPM{z} suggests five centers of activation within the significantly activated areas, the posterior probabilities are more homogeneously distributed over the significantly activated areas. In other words, the probability for activation is very high both in the centers of activation detected with the classical method and in their surrounding voxels. This seems intuitively plausible, since we would expect that the probability of activation in voxels close to an activation focus is still very high, even if the strength of the activation is smaller than in the center.

Differences between the five centers of activation which are not obvious from the SPM{z} can be detected from the posterior means. The posterior means in the lingual gyri are considerably higher than in the other regions (see also Table 1) suggesting that the contrast between valid-left and the valid-right trials is much stronger there than in the remaining activated areas. Note that the value of the posterior mean in our model is proportional to the difference in the maximum signal amplitude for valid-left and valid-right trials. The differences in the posterior means found here support the results from time course analyses of these regions presented by Pollmann and Morillo (2003) (see in particular their Fig. 2). Maximum amplitude differences in the time courses between valid-left and valid-right trials were about 0.2% signal change for both lingual gyri. In the lateral occipital gyri and the IPS/TOS this difference was considerably lower.

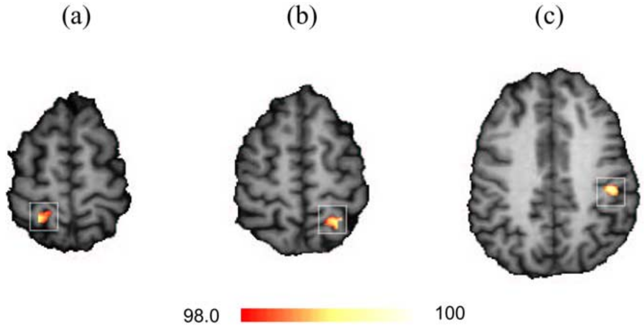
A second possible contrast in the experimental design arises from the presentation of small or big cues followed by the target in the same visual hemifield (valid-small against valid-big trials). The behavioral data suggested, if at all, only a very small effect. However, the Bayesian analysis revealed a number of cortical regions with posterior probabilities similar to those estimated for the previous contrast. Three of these regions are visualized in Fig. 3. Here, the Bayesian analysis yielded posterior probabilities of 98% or higher, whereas the values in the classical SPM{z} did not exceed the threshold of $z = 3.09$.

One reason for the different results obtained with both methods becomes visible when analyzing the contrast found

[3] Note that the values reported here differ slightly from the results in Pollmann and Morillo (2003). These differences were caused by the use of an updated version of the software package LIPSIA with slightly modified parameter settings.



2



3

Fig. 2. Posterior means (top left) and posterior probability maps for P ($\mathbf{c}\beta > 0$) (top middle) and P ($\mathbf{c}\beta < 0$) (top right) for the contrast between valid-left and valid-right trials. The corresponding SPM{z} is shown in the bottom row. Note that negative z values and posterior means (blue) indicate activations for the inverse contrast, i.e., valid-right against valid-left trials.

Fig. 3. Posterior probability maps for the contrast between valid-small and valid-big trials. The contrasts for all individual subjects in the three marked regions are further analyzed in Figs. 4a–c.

for the individual subjects. Figs. 4a–c show the contrast for all individual subjects for the centers of activation in the three cortical areas marked in Figs. 3a–c, respectively. For comparison, Fig. 4d shows data for the center of activation in the inferior temporal sulcus obtained for the previous contrast. Here, the classical method yielded a z value of 4.49 which is usually regarded as significant activation.

For all four voxels in Fig. 4 the estimated contrast is larger than zero for the majority of subjects. However, two subjects in voxel (a) and one subject in voxels (b) and (c) differ from the general pattern with large negative contrasts. Such differences in the estimated contrasts can be caused, for example, by large anatomical variations between subjects which cannot fully be accounted for by the preprocessing procedures. Outliers can also result from differences in the temporal behavior of the BOLD response of individual subjects. Relatively large temporal offsets between the observed data and the model function for the hemodynamic response can lead to poor fitting and estimation of the model parameters, in particular if the same model function is applied to all subjects, which is a prerequisite for both the two-stage random-effect analysis and our Bayesian method.

In our data, the estimated variances of the outliers are larger than the variances of the estimates for most other subjects. As we have already seen from Eq. (14) and (15) and from Fig. 1, with our approach the influence of each individual subject on the posterior for the entire group is determined by the estimated variance of the contrast specific to this subject. Given the relatively large variances of their estimates, the influence of the outliers' contrasts is not large enough to move the posterior mean close to zero. Consequently, the posterior probability $P(\mathbf{c}\beta > 0)$ is still very high, whereas the significance values from the classical method do not exceed the required threshold due to the high between-subject variance caused by the outliers. For comparison, the classical second-level analysis was repeated for

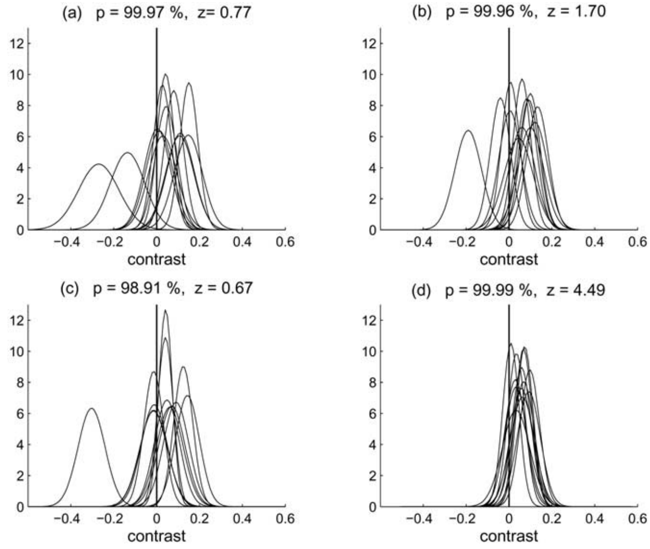


Fig. 4. The sampling distribution of the contrast between valid-small and valid-big trials estimated for all individual subjects in three voxels (a)–(c). These three voxels have the highest posterior probability in the cortical areas marked in Figs. 3a–c, respectively. These areas show no significant activation in the classical method. For comparison, (d) shows the contrast between valid-left and valid-right trials for all individual subjects for the center of activation in the inferior temporal sulcus. This region was also found significantly activated using the classical method. Posterior means and standard deviations were similar for all four voxels: (a) 0.052, 0.015, (b) 0.048, 0.014, (c) 0.031, 0.014, and (d) 0.054, 0.013, respectively.

the three regions (a), (b), and (c), now omitting the two outliers for (a) and the single one for (b) and (c). The *z* values for the three voxels increased to 3.14, 3.19, and 2.61, respectively. Although these values are still comparatively small, the large increase relative to the initial analysis shows clearly that the outliers had a huge impact on the results of the classical analysis.

The complete Bayesian second-level analysis of all experimental conditions and the exact neuropsychological interpretation of the obtained results are beyond the scope of this article and have to be left to the original experimenters. However, we belief that the data presented here demonstrate that Bayesian analysis is a promising alternative to conventional methods, in particular for experimental paradigms which address very subtle differences between conditions, where classical methods are likely to fail due to too conservative thresholding.

4. Discussion

We have introduced a Bayesian method for second- and higher-level analyses of fMRI data which is based on modeling the obtained measurements for single subjects by means of the GLM. The method is easy to implement and computationally inexpensive. The required computation time is on the order of seconds for a complete second-level analysis following a relatively simple classical modeling on the first level. This is in stark contrast to alternative approaches using nonlinear or hierarchical Bayesian modeling. Complex Bayesian models can be computationally expensive, and computation times on the order of hours or even days for single subjects have been reported (Genovese, 2000). Despite its simplicity, our method overcomes some of the drawbacks of NHST such as the need to address the problem of multiple comparisons. It provides estimates for both the size of an effect of interest and the probability of the effect to occur in the population. The results are easy to interpret and intuitively more plausible than results of classical NHST. Like other Bayesian approaches, our method permits complex inferences which are hard to derive from NHST.

Note that our Bayesian second-level analysis could also be combined with a different first-level analysis and is not restricted to the use of the GLM on the first level. One could, for example, conceive of a nonlinear model on the first level resulting in parameters with a clear physical interpretation such as a direct estimate of the amplitude or time delay of the observed signals. The only prerequisite for the straightforward application of Bayes' theorem presented here is that the effects of interest are described for single subjects as normally distributed variables.

NHST is based on a frequentist interpretation of probability. The probability of an event is defined as its relative frequency and is therefore, when viewed over a large number of trials, a constant. Consequently, a hypothesis about the event can be only true or false, and the observed data help us to decide between these two possibilities. They do not allow us, however, to assess the probability of our hypothesis to be correct, although *P* values resulting from NHST are often wrongly interpreted this way (Krueger, 2001; Gigerenzer, 1993; Oakes, 1986). Bayesian inference on the other hand provides us with exactly this information. Here probability is viewed as an individual's belief about an event which is modified by the observed data. Our initial belief (or hypothesis) about the event, represented as prior probability, is modified by the observed data whereby we become more certain about the true nature of the event the more data we encounter.

Critics of Bayesian techniques often stress the subjectivity inherent in the methods. Clearly, the posterior probability of an event crucially depends on the chosen prior, i.e., on the experimenter's belief about the event in question. However, we would argue in line with Lange (1997) and others that this subjective element should be regarded a virtue rather than a disadvantage. It provides us with a method to incorporate knowledge and experiences from previous studies or subjects into our model and combine them with newly acquired data. Making use of this virtue, we take the probability distribution estimated for one subject as our initial belief about the true distribution of the parameter in the entire group, i.e., as the first prior in the iterative application of Bayes' theorem. This way only actually observed data enter into the calculation of the posterior for the group of subjects.

It is also important to note that some degree of subjectivity enters into non-Bayesian models, too, when choosing the experimental design, formulating hypotheses, or selecting model parameters (Petersson et al., 1999; Krueger, 2001). As Gössl et al. (2001a) point out, model specifications such as the choice of basis functions in the GLM provide even harder constraints on the solutions than Bayesian priors. While the former specify a subspace in which the solution must lie, the latter impose only soft constraints on the solutions which can be violated, if a sufficient amount of acquired data provide appropriate evidence.

When assessing functional neuroimaging data, a considerable degree of variability in the measured signal can be observed both in individual scans of the same subject and in different subjects of a group (Aguirre et al., 1998; Miezin et al., 2000; McGonigle et al., 2000; Duann et al., 2002; Neumann et al., 2003). Different classical models allow the within-subject and the between-subject variance of the observed responses to enter the second-level analysis to varying degrees. Classical fixed-effect models do not take into account the between-subject variability of the responses. Rather, they are based on the assumption that all subjects respond with the same variance and thus utilize the within-subject variance as the only variance component. In other words, the acquired data are treated as if coming from a single subject. While the resulting large number of degrees of freedom facilitate a high sensitivity of the method, a classical fixed-effect model can produce significant effects by virtue of a single subject.

Mixed- or random-effect models are designed to take into account both the within-subject and between-subject variability of the responses. Subjects are viewed as randomly sampled from a population and the effects estimated for each subject are treated as random variables. The resulting variance of the estimated response across subjects contains both within- and between-subject variance components in a proportion determined by the ratio of scans per subject to the number of subjects (Friston et al., 1999). However, the analysis of random-effect models is often difficult (Searle et al., 1992) and the usually small number of subjects in fMRI experiments results in a low power of such analyses. A relatively simple two-stage model implementing a random-effect analysis can be applied under the conditions that the model is balanced, i.e., the same experimental design is used for all subjects, and that the model is separable by subject, i.e., the parameter estimates for each subject are independent (Holmes and Friston, 1998). This implementation, as realized for example in SPM and LIPSIA, builds upon the idea of simple summary statistics (Frison and Pocock, 1992). It should be noted, however, that the model rests on the simplifying assumption that the within-subject residual variance is constant for all subjects (McGonigle et al., 2000). This means that while the within-subject variance is one component of the overall variance, differences in the within-subject variance between subjects are neglected in the analysis. As our data have shown, these models are still relatively sensitive to outliers which cause a high between-subject variance.

In our Bayesian approach the variance of the resulting posterior is the pooled within-subject variance of all subjects (Worsley et al., 2002). The between-subject variance is expressed by the spread of the means of the estimated parameters for individual subjects. Note that while the actual location of the means clearly influences the posterior mean for the group, their variance does not necessarily do so. In this respect our method implements a fixed-effect analysis. However, unlike in classical fixed-effect analyses, measurements are not viewed as coming from a single subject. Modeling on the first level is performed independently for each individual subject and the Bayesian inference allows for different within-subject variances. Most importantly, the within-subject variance, i.e., the stability of the measurements obtained for the individual trials and the goodness of the model fit, determines the influence of a subject on the posterior for the group. The influence of a few outliers on the group result is small as long as their within-subject variance is not considerably smaller than those of the remaining subjects. This is a large advantage over conventional methods where the influence of each individual subject is not weighted by its within-subject variance. Consequently, outliers can cause the between-subject variance to increase considerably independent of their within-subject variance, which in turn results in small *t* and subsequent *z* values. Such outliers cannot always and completely be avoided given the large anatomical and physiological variability in the population. Therefore, robustness against such outliers is a prerequisite of powerful analytical tools for the evaluation of fMRI data. Our Bayesian second-level analysis meets this prerequisite.

Our experimental results highlight the fact that it is important to consider both the effect size and the significance or strength of evidence for the effect. This is of course equally true for both Bayesian techniques and classical analyses. For the latter the presentation of effect size together with the significance of the activation is too often neglected. Within the Bayesian framework, posterior probability distributions provide sufficient means to report both aspects of the obtained results. Moreover, Bayesian inference based on posterior probability distributions does not depend on any form of thresholding (Friston et al., 2002a) which is typically required in classical approaches. In fact, using posterior mean and variance of an effect, posterior probability maps for any threshold of interest can be derived. This becomes particularly important for complex scientific hypotheses and, most notably, for analyses based on model functions whose estimated parameters have a clear physical interpretation.

Finally, a comparison between our Bayesian approach and a classical analysis on the same data revealed that the latter might disregard a number of activations on the basis of their relatively low significance. We agree with Friston et al. (2002b), who observed that "there is no magical increase in power afforded by a Bayesian approach." However, we would also argue that we must not ignore cortical areas for the neuropsychological interpretation of experimental results, where the posterior probability of an effect is as high as 99% or above, just because they have missed a more or less arbitrary threshold in the classical analysis. We believe that they at least demand a closer look.

Acknowledgments

We thank Prof. D. Yves von Cramon, Prof. Stefan Pollmann, and two anonymous reviewers for invaluable help and many insightful comments on earlier drafts of the manuscript. Prof. Pollmann also kindly provided the data for our analysis. Further thanks go to Dr. Jöran Lepsien for help with the classical statistical analysis of the data.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human BOLD hemodynamic responses. NeuroImage 8 (4), 360–369.

Ardekani, B.A., Kanno, I., 1998. Statistical methods for detecting activated regions in functional MRI of the brain. Magn. Reson. Imaging 16 (10), 1217–1225.

Box, G.E.P., Tiao, G.C., 1992. Bayesian Inference in Statistical Analysis. Wiley, New York.

Duann, J.-R., Jung, T.-P., Kuo, W.-J., Yeh, T.-C., Makeig, S., Hsieh, J.-C., Sejnowski, T.J., 2002. Single-trial variability in event-related BOLD signals. NeuroImage 15, 823–835 doi:10.1006/nimg.2001.1049.

Frank, L.R., Buxton, R.B., Wong, E.C., 1998. Probabilistic analysis of functional magnetic resonance imaging data. Magn. Reson. Med. 39, 132–148.

Frison, L., Pocock, S.J., 1992. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Statist. Med. 11, 1685–1704.

Friston, K.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Human Brain Mapp. 2, 189–210.

Friston, K.J., Glaser, D.E., Henson, R.N.A., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: applications. NeuroImage 16, 483–512 doi:10.1006/nimg.2002.1091.

Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. Comments and controversies: how many subjects constitute a study? NeuroImage 10, 1–5 doi:10.1006/nimg.1999.0439.

Friston, K.J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time series. Human Brain Mapp. 1, 153–171.

Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483 doi:10.1006/nimg.2002.1090.

Gelman, A., Carlin, J., Stern, H.S., Rubin, D.B., 2000. *Bayesian Data Analysis,* Chapman & Hall–CRC.

Genovese, C.R., 2000. A Bayesian time-course model for functional magnetic resonance imaging data. J. Am. Stat. Assoc. 95 (451), 691–703.

Gigerenzer, G., 1993. The superego, the ego and the id in statistical reasoning. in: Keren, G., Lewis, C. (Eds.), A Handbook for Data Analysis in the Behavioural Sciences: Methodological Issues. Erlbaum, Hillsdale, NJ, pp. 311–339.

Gössl, C., Auer, D.P., Fahrmeir, L., 2001a. Bayesian spatiotemporal inference in functional magnetic resonance imaging. Biometrics 57, 554–562.

Gössl, C., Fahrmeir, L., Auer, D.P., 2001b. Bayesian modeling of the hemodynamic response function in BOLD fMRI. NeuroImage 14, 140–148 doi:10.1006/nimg.2001.0795.

Højen-Sørensen, P., Hansen, L., Rasmussen, C., 2000. Bayesian modelling of fMRI time series, in: Solla, S., Leen, T., Müller, K.-R. (Eds.). *Advances in Neural Information Processing Systems,* Vol. 12, MIT Press, pp. 754–760.

Holmes, A.P., Friston, K.J., 1998. Generalizability, random effects, and population inference. NeuroImage 7, S754.

Kershaw, J., Ardekani, B.A., Kanno, I., 1999. Application of Bayesian inference to fMRI data analysis. IEEE Trans. Med. Imaging. 18 (12), 1138–1153.

Krueger, J., 2001. Null hypothesis significance testing: on the survival of a flawed method. Am. Psychologist 56 (1), 16–26 doi:10.1037//0003-066X.56.1.16.

Lange, N., 1997. Empirical and substantive models, the Bayesian paradigm, and meta-analysis in functional brain imaging. Human Brain Mapp. 5 (4), 259–263.

Lee, P.M., 1997. *Bayesian Statistics: An Introduction,* Oxford University Press.

Lohmann, G., Müller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., Zysset, S., von Cramon, D.Y., 2001. LIPSIA—a new software system for the evaluation of functional magnetic resonance images of the human brain. Comput. Med. Imaging Graphics 25 (6), 4498.

Marrelec, G., Benali, H., Ciuciu, P., Plgrini-Issac, M., Poline, J.-B., 2003. Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. Human Brain Mapp. 19 (1), 1–17 doi:10.1002/hbm.10100.

McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S.J., Holmes, A.P., 2000. Variability in fMRI: an examination of intersession differences. NeuroImage 11, 708–734 doi:10.1006/nimg.2000.0562.

Miezin, F.M., Maccotta, L., Ollinger, J.M., Petersen, S.E., Buckner, R.L., 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. NeuroImage 11, 735–759 doi:10.1006/nimg.2000.0568.

Neumann, J., Lohmann, G., Zysset, S., von Cramon, D.Y., 2003. Within-subject variability of BOLD response dynamics. NeuroImage, 19, 784–796.

Oakes, M., 1986. Statistical Inference: A Commentary for the Social and Behavioral Sciences. Wiley, New York.

Petersson, K.M., Nichols, T.E., Poline, J.-B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging. I. Non-inferential methods and statistical models. Phil. Trans. R. Soc. Lond. B 354, 1239–1260.

Pollmann, S., Morillo, M., 2003. Left and right occipital cortices differ in their response to spatial cueing. NeuroImage 18, 273–283 doi:10.1016/S1053-8119(02)00039-3.

Searle, S.R., Casalla, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

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

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain. Thieme, Stuttgart.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—again. NeuroImage 2, 173–181.

Worsley, K.J., Liao, C., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15, 1–15.

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics. NeuroImage 5, 179–197.