Free estimation of the Hemodynamic Response Function by cross validation in functional Magnetic Resonance Imaging data

A solution for the wrong assumption of an equal Hemodynamic Response Function between and within subjects

---

Bram Rohlfs

Master's Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: Juli 12 2019

Student number: s1236466

Supervisor: Dr. W.D. Weeda

Abstract

Neural activity is estimated by measuring the Blood Oxygen Level Dependent (BOLD) in functional Magnetic Resonance Imaging (fMRI) research. The time course of the BOLD response is often called the Hemodynamic Response described by the Hemodynamic Response Function (HRF). Several studies have shown that the shape of the HRF can vary in shape between persons and within persons between areas of the brain. Therefore, when analysing fMRI data, the method used should account for this variability. However, commonly used methods assume no or only little variability in the HRF.

This study proposes a method that treats the BOLD response on single trials within one time series as separate time series. Doing so allows for the use of cross validation to freely estimate the HRF and thus account for the variability in the data. The free HRF estimation is done by averaging the BOLD response over randomly selected trials multiple times. In each time, trials are either assigned to HRF estimation (in training data) or to evaluate the estimated HRF (in test data). The testing for activation is done by comparing the freely estimated HRF with an intercept only model on the trials assigned to evaluate the estimated HRF. A simulation study is conducted to test this method and compare it to a basic approach using a pre-specified HRF in a general linear model (GLM). Sensitivity, specificity and amplitude estimation are measured, discussed and compared between methods.

The simulations showed that the free HRF estimation indeed accounts for variability of the HRF in the data while the basic BGLM approach suffers increasingly on sensitivity and amplitude estimation as the difference between the pre-specified HRF and the HRF in the data increases. On the account of specificity, the free HRF estimation performs slightly worse compared to the basic BGLM approach. It is concluded that the proposed method accounts for the variability of the HRF well but that improvements have to be made to make the method applicable in practice. Lastly it is concluded that treating the trials within one time series as separate time series opens up to different methods of evaluating the HRF in the future.

Table of Content

Introduction

With the rise of functional magnetic resonance imaging (fMRI) as a method to measure brain activity (Kwong et al., 1992; Ogawa et al. 1992), came the need for methods to analyse fMRI data. A common goal of these methods is to detect brain activity by estimating the "Blood Oxygen Level Dependent" (BOLD) over time. When areas in the brain become active, an increase of blood flow occurs. Since this increase in blood flow occurs together with a change in blood oxygen level, measuring the oxygen level in the bloodstream (the BOLD) can be seen as an indirect way to measure brain activity (Logothetis, 2008). The time course of the BOLD is often called the Hemodynamic Response (HR) described by the Hemodynamic Response Function (HRF). Hence, the HRF can be seen as the association between the BOLD response and the underlying neural activity, which is the construct of interest in fMRI analyses. Together with the interest in the HR, comes the interest in distinguishing the HR between different parts of the brain. The ability to measure neural activity for different areas in the brain is important when researchers want to know which areas of the brain are used, when performing certain tasks or being exposed to certain stimuli. To be able to do so, the brain is divided into cubes called voxels. The volume of the voxels depends on the spatial resolution of the fMRI scanner. A common resolution is 3 x 3 x 3 mm^2 but higher resolutions can go from 100 to 150 μm (Goense, Bohraus & Logothetis, 2016). With an average brain volume of around 1200 cm^2 (e.g. Allen, Damasio & Grabowski, 2002), depended on resolution, this can lead to many voxels per brain that need to be analyzed for neural activity (for example 400.000 voxels for a resolution of 3 x 3 x 3 mm^2 and 1.200.000 voxels for a resolution of 1 x 1 x 1 mm^2).

In general, when a subject receives a stimulus or performs a task in the fMRI scanner, a BOLD signal is estimated for each voxel in the brain. When the BOLD time course (or HR) of a voxel changes with the introduction of the stimulus or task, those voxels will then be seen as related to the stimulus or task (Logothetis & Wandell, 2004).

The General Linear Model

The General Linear Model is a commonly used method to analyse the fMRI data (Pernet, 2014). When using a general linear model (GLM) to analyse fMRI data, an HRF is modelled to resemble the expected HRF in activated voxels. This HRF is modelled by convolving a stick-function, which describes when the neural activation is expected, with an HRF function, which describes the shape of the HRF that is expected. This expected HRF is then fitted

through a GLM on each voxel to obtain a regression coefficient and matching p-value per voxel. The p-value indicates the chance of no match between the expected HRF and the voxel's time course and the regression coefficient provides an estimation of the HRF amplitude in the data. When modelling the expected HRF to match the HRF in the data, strong a priori assumptions are made about the shape of the HRF. First, it is assumed that the shape of the modelled HRF matches the shape of the HRF in the data. Second, it is assumed that the shape of the HRF in the data is the same between persons and within persons (between voxels) (Friston, Holmes, Worsley, Poline, Frith & Frackowiak, 1994; Friston et al, 1995). Since the modelled HRF is fitted on the data, the power to detect and estimate activation is hugely dependent on the match between the modelled HRF and the true HRF in the data (Lindquist, Meng Loh, Atlas & Wager, 2009). Therefore, the validity of the analyses is dependent on these assumptions being tenable when fitting the modelled HRF to the data.

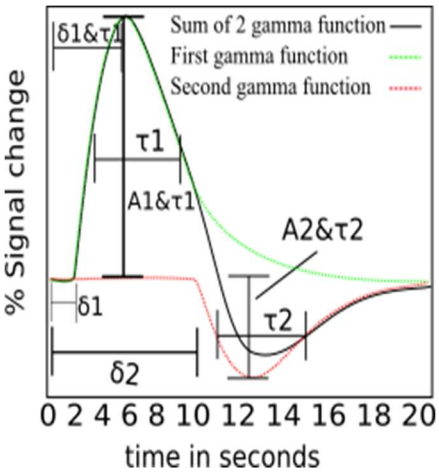


Figure 1. "Parameters for the sum of two gamma function model of the hemodynamic response. A_1 and A_2 model the magnitude of the peak and undershoot terms respectively; τ_1 and τ_2 model the width, peak height, and time-to-peak; δ_1 and δ_2 model the time-to-onset (or latency)". (Adapted from Handwerker, Ollinger, & D'esposito, 2003, figure 1)

Figure 1 shows the shape and parameters of a double gamma HRF often used to model the expected HRF. Regarding the assumption of an equal HRF between persons and within persons (between voxels), when using a predefined shape and parameters for the HRF, possibilities of all other shapes, parameters or any variations between persons or within persons (between voxels) are disregarded. This can become a problem when the modelled HRF does not match the HRF in the data or if the HRF in the data varies in shape between or within persons (between voxels). If this happens, the modelled HRF can be seen as mis specified since it no longer matches the shape of the HRF in the data. For example, when the

latency (or time to onset) of the expected HRF is predefined on one second but in the data, there is actually a three or maybe four seconds latency, the model does not resemble the data anymore and might fail to capture the magnitude of the actual neuronal activity. Furthermore, when in one area of the brain the shape of the HRF differs from another part of the brain, a single model for the expected HRF might not be sufficient to capture all neuronal activity in the brain. Besides these examples, there are many other parameters that can be misspecified when modelling the expected HRF depending on the function chosen to shape the HRF.

According to several studies, (Lee, Meyer & Glover, 1995; Aguirre, Zarahn, & D'esposito,1997; Handwerker et al. , 2003) variations in HRF between persons and within persons (between voxels) are indeed present and can have a significant impact on the analysis. Therefore, the assumption of an equal HRF between persons and within persons (between voxels) is false. Hence, as described above, ignoring the variability of the HRF between persons and within persons (between voxels) might lead to a misspecified model. In turn, using a misspecified model for analyses can lead to biased results and thus lower the validity of the research. Therefore, to increase the validity of the research, the method used must have some way to account for this variability.

To do so, several adaptations to the above-described use of the GLM have been proposed. An often-used method is adding derivatives of the modelled HRF in the GLM to increase the variability of the model. For example, by adding the first temporal derivative of the HRF (orthogonalized or not) the GLM becomes especially more sensitive to differences in latency and time to peak between the modelled HRF and the HRF in the data. Although adding derivatives to the model can increase the variability of the modelled HRF, the exact specification of the model and the study-design itself can influence results between different models (Pernet, 2014). Another adaptation is adding mixed (or random) effects to the GLM. Depending on the specification, adding mixed effects to the model can only partly account for the within and between subject variability (Friston, Stephan, Lund, Morcom, & Kiebel, 2005).

Alternative Methods for HRF Estimation

Solutions to the variability of the HRF in the data can also be found outside the framework of a pre-defined HRF by releasing the assumptions of and restrictions on the shape of the HRF as done in several existing methods. Lazar (2008) mentions several approaches. Burock and Dale (2000) propose to add the HRF as a parameter in a GLM, Marrelec, Benali, Ciuciu, P´el´egrini-Issac and Poline (2003) extend this idea using Bayesian statistics to estimate the HRF, Gibbons et al. (2004) fit a cubic polynomial over averaged trials to estimate the HRF

and Kapur et al. (2009) extend this approach by instead of averaging over trials, fitting a fourth-degree spline on the voxel time series.

Burock and Dale (2000) propose to add the HRF as a parameter in a GLM. By adding the HRF as unknown parameter in a model where the BOLD signal and time of stimuli are added as known parameters, the HRF can be estimated by maximum likelihood methods. This way they avoid having to specify any form for the HRF. However, when the HRF is estimated, hypotheses about the shape of the HRF must still be specified to be able to allow for statistical testing. Marrelec et al. (2003) extend this idea using Bayesian statistics to estimate the HRF but the HRF still needs to be evaluated after estimating and not all researchers might be familiar with Bayesian statistics. Gibbons et al. (2004) fit a cubic polynomial over averaged trials to estimate an average HRF. They then add random effects for each voxel to allow for deviations from the averaged HRF. Kapur et al. (2009) extend this approach by instead of averaging over trials, fitting a fourth-degree spline on the voxel time series. However, by using this approach not all variability of the HRF between voxels might be accounted for.

Furthermore, the methods mentioned above can be complex and there might be a need for more simpler intuitive approaches within this area of science. Also, more complex models can be more sensitive to noise and with more complexity in the model comes the chance of overfitting the models. Therefore, methods designed to account for the variability in the HRF should aim to avoid prespecifying the HRF or having to specify hypotheses about the HRF after estimation while still remaining simple.

Proposed Method

This study proposes a new method that accounts for the variability in the HRF using a more intuitive approach. By treating the trials as separate time series, multiple models can be fitted and evaluated through cross validation. This allows for the use of basic statistics to test for activation while using a freely estimated HRF. The HRF in this method is freely estimated by averaging the signal over randomly selected trails. This way avoiding to specify hypotheses about the estimated HRF while still using the GLM to estimate the HRF.

Basically, the HRF is estimated from the data itself per voxel time series by averaging the signal over the trials. Furthermore, by assigning trials to either the independent variable or the dependent variable, the averaged trials can be fit using a GLM. By doing this multiple times, different models can be fitted and evaluated. The multiple model fitting is done based on the general concept of cross validation. The cross validation method used in this study

assigns the trials to a pre-defined number of "folds". By leaving out or including a fold different models are obtained when fitting the GLM. This is done to get a more realistic representation of the (de)regularities in the data compared to only fitting one model. Furthermore, by using cross validation, the models can be evaluated on trials excluded from the model. This might decrease the chance of false positives or overfitting since the final data used to evaluate the model was not used to fit the model.

This method omits to have to specify hypotheses about the HRF as done in Burock and Dale (2000). However, by averaging over trials, the assumption is made that the shape of the HRF is the same between trials within one voxel time series. Averaging over trials also makes it inevitable that some noise will end up in the model. To deal with this problem, the proposed method makes the assumption that the shape of the HRF is smooth. By smoothening the averaged signal obtained from the trials the proposed method might be more robust to noise. The procedure of the proposed method is further explained and illustrated in the method section and is further referred to as Free HRF Estimation (FHRFE).

Research Question and Hypotheses

The aim of this study is to explore the effectiveness of the FHRFE and compare the sensitivity, specificity and amplitude estimation to a basic GLM in which the HRF is pre specified. It is hypothesized that the FHRFE outperforms the basic GLM method on sensitivity and amplitude estimation when the HRF is wrongly specified in the basic GLM model.

To test and compare methods, a simulation study is done where the time series are simulated under different conditions on a single trial level. By this simulation, the existing basic GLM can be compared to the Free HRF Estimation in sensitivity, specificity and amplitude estimation. Further expectations are described in the method section.

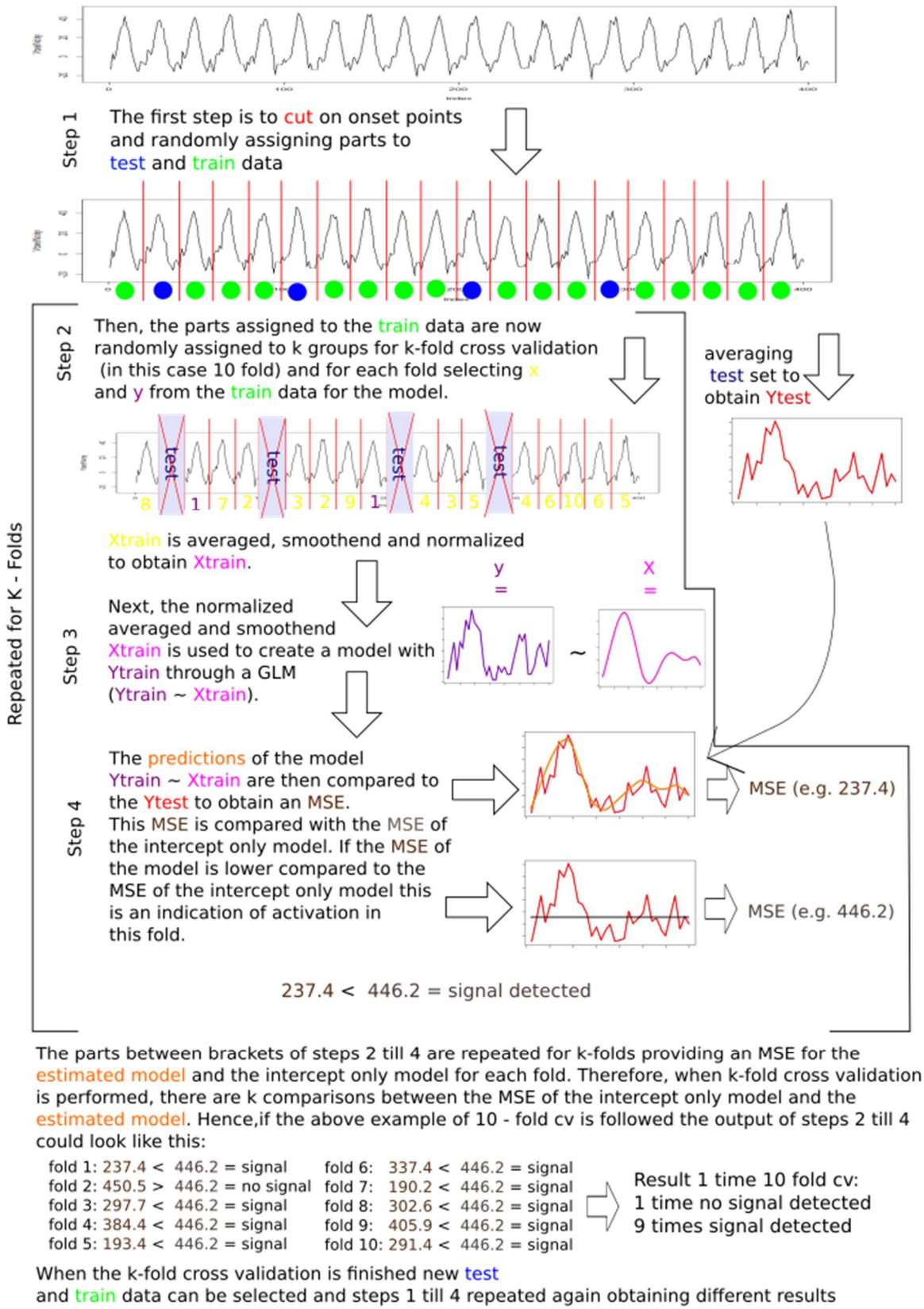
Methods

In this section, the method used to answer the research questions is described. First, the Free HRF Estimation is explained in more detail. Second, the materials and the general procedure is described. Third, the simulation of the data is described and fourth, the method of evaluation is described.

Free HRF Estimation Explained

Figure 2 and 3 show a visual representation of the steps taken by the Free HRF Estimation to detect the signal and estimate the amplitude in the voxel time series. The Free HRF Estimation uses multiple cross validations per time series to obtain the shape of the HRF and compares the obtained model to an intercept-only model on test data. As briefly explained in the introduction, firstly, by using cross validation, the HRF is estimated by multiple models which provide a distribution of the model results. This distribution can be used to provide a measure of certainty of which the null hypotheses of no activation can be rejected. Secondly, by using cross validation, the models can be evaluated on data not used to create the models with (the test data). This allows for a different way of evaluating the models compared to when only fitting a single model and might counter overfitting or false positives. The Free HRF Estimation has two variations which are tested. The first is signal detection based on MSE value and the second is signal detection based on chi-square value (X^2). In both variations, the general idea is the same but the way the models are fitted and evaluated differs. This is further explained in this section later.

In general, to estimate the HRF the Free HRF Estimation goes through several steps. First, the time series of the voxel is split per trial into parts of 20 seconds long where a stimulus is presented. Second, these pieces are assigned to the training (train) or test data (test) randomly. The data in the pieces assigned to the test data is averaged to obtain the averaged HRF of the test data (Y_{test}). The data assigned to the train data is used to create the models of the HRF. For each fold, in the cross validation, a different model of the HRF is obtained and evaluated on the test data (Y_{test}). This can be done multiple times by using different train and test data for each cross validation. If the repeated cross validation is finished the evaluation of the models can be averaged to obtain an indication for the activation in the voxel time series. For the two different approaches, this is done in slightly different ways as explained after figure 2 and 3.



The parts between brackets of steps 2 till 4 are repeated for k-folds providing an MSE for the estimated model and the intercept only model for each fold. Therefore, when k-fold cross validation is performed, there are k comparisons between the MSE of the intercept only model and the estimated model. Hence,if the above example of 10 - fold cv is followed the output of steps 2 till 4 could look like this:

fold 1: 237.4 < 446.2 = signal	fold 6: 337.4 < 446.2 = signal
fold 2: 450.5 > 446.2 = no signal	fold 7: 190.2 < 446.2 = signal
fold 3: 297.7 < 446.2 = signal	fold 8: 302.6 < 446.2 = signal
fold 4: 384.4 < 446.2 = signal	fold 9: 405.9 < 446.2 = signal
fold 5: 193.4 < 446.2 = signal	fold 10: 291.4 < 446.2 = signal

⇨ Result 1 time 10 fold cv:
1 time no signal detected
9 times signal detected

When the k-fold cross validation is finished new test and train data can be selected and steps 1 till 4 repeated again obtaining different results

Figure 2. Illustration of the Free HRF Estimation using MSE value to evaluate the obtained models.

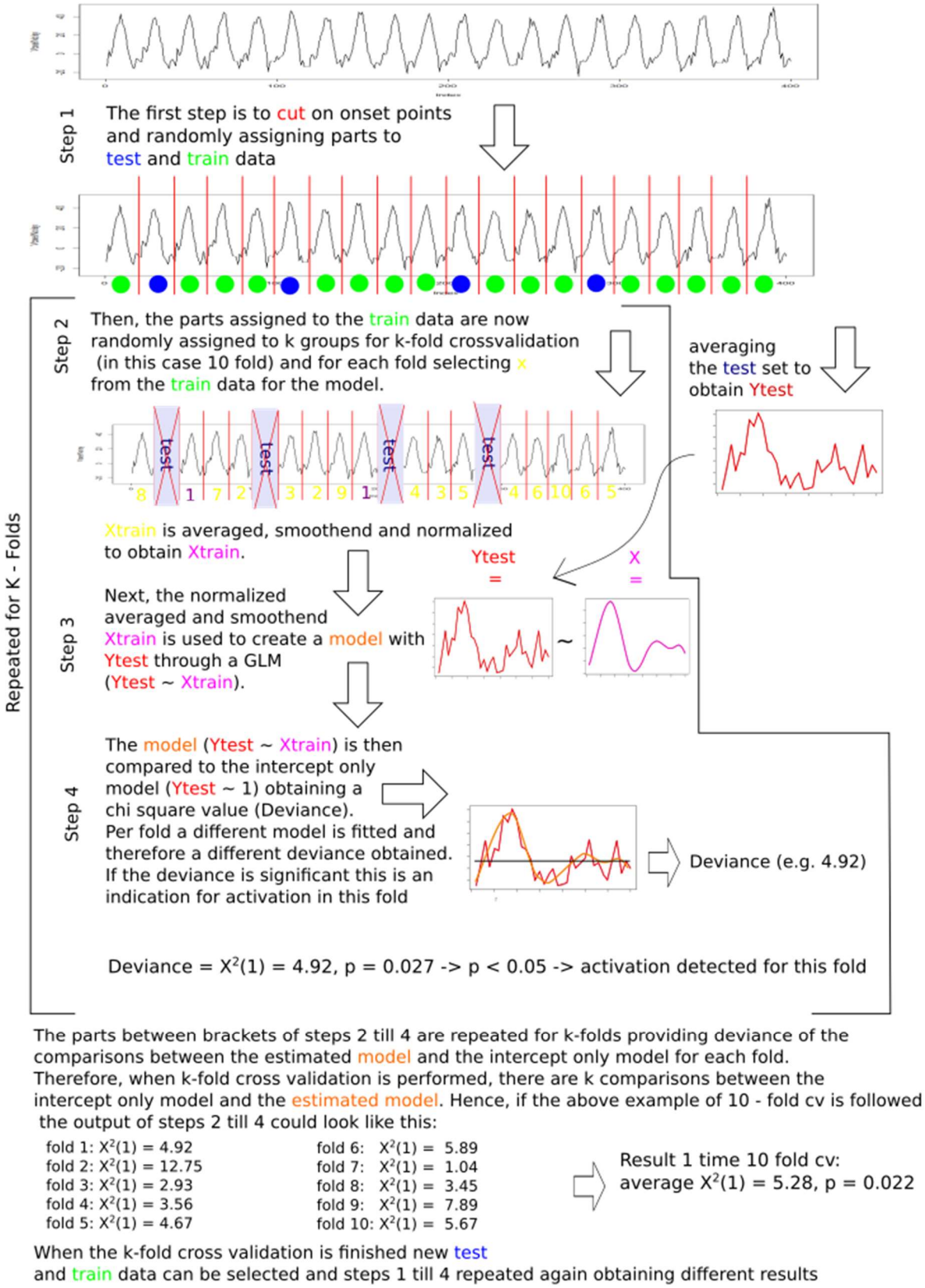


Figure 3. Illustration of the Free HRF Estimation using chi square value to evaluate the obtained models

Signal detection based on MSE value. For signal detection based on MSE value, as shown in step 2 and 3 in figure 2, in each fold of the cross-validation an HRF is estimated by using some pieces of training data as independent variable X_{train}, and other pieces as dependent variable Y_{train} in a GLM. X_{train} is created by averaging the data in the pieces assigned to X_{train}. After averaging, X_{train} is normalized and smoothed by using polynomial basis functions. Y_{train} is created by averaging the data in the pieces assigned to Y_{train}. Per fold, X_{train} and Y_{train} consist of different data and thus a different model is provided per fold. The models obtained from each fold provide an estimate of HRF amplitude. The predictions of the models obtained from this process are then compared to the test data (Y_{test}) to obtain an MSE per model. This MSE is then later compared to the MSE of the intercept-only model as shown in step 4 of figure 1. To obtain more stable results the cross-validation is repeated with different train and test data.

The repeated cross validation produces a different model and a matching MSE per fold. Then, a count is done for how many times the intercept-only model has a higher MSE compared to the obtained model. When this is the case for 95% of the models created in the folds the null hypothesis of no signal is rejected. To do this properly, there should be a considerable amount of models (and thus folds) to obtain stable results. When the Free HRF Estimation, for example, is done with 10 fold cross-validation, each fold will provide 1 model. Therefore, to obtain 100 models and their matching MSE, the 10 fold cross-validation should be performed 10 times.

Signal detection based on X^2 value. Figure 3 shows signal detection based on X^2 value. In each fold of the cross-validation, an HRF is estimated by using pieces of training data as independent variable X_{train}. X_{train} is created by averaging the data in the pieces assigned to the train data. After averaging, X_{train} is normalized and smoothed by using polynomial basis functions. However, Y now consists out of the test data (Y_{test}) and therefore, only X_{train} consists of different pieces per fold and Y stays the same. Per fold, the obtained model is compared to the intercept-only model to obtain a deviance.

By repeating cross validation, multiple models and deviances are obtained. The average of these deviances is then compared to the chi-square distribution to obtain a p-value. When smaller than 0.05 the null hypothesis of no signal is rejected.

Estimation of the amplitude. The estimated amplitude is provided by taking the average maximum predicted signal of all generated models in the folds.

Basic GLM explained

The method that the FHRFE is compared to, fits one model through a GLM. Within this model, X is defined with a double – gamma function (also called canonical HRF) and Y is defined by the data. When fitting X on Y yields an F value with matching p-value lower than 0.05, the null hypotheses of no signal is rejected. Amplitude estimation can be done with this method by interpreting the beta obtained from the model. In the next sections, this method is referred to as basic GLM (BGLM)

Materials and General Procedure

All simulations and analysis are done within the R environment version 3.4.1 with Rstudio (R core Team, 2017). The "neuRosim" package (Welvaert, Durnez, Moerkerke, Verdoolaege & Roseel, 2011) is used for all fMRI related functions.

To compare and validate the FHRFE, this study used simulated fMRI data opposed to real fMRI data. Since the properties of the FHRFE are not yet known this study can be seen as a first step to explore this approach. In real fMRI data, the true signal is not known and therefore using this data cannot indicate the false or true positives and negatives.

To save time during the simulation, the simulations are split up into four parts and run on a cluster. The simulations where no signal is present in the data is the first part and the simulations where a signal was present in the data is split into the other three parts based on the type of noise. The seeds were randomly picked between 0 and 1000. For the no-signal simulations, the seed is set on 221, the white noise signal present seed is set on 402, the autocorrelated noise of second order signal present seed is set on 871 and the autocorrelated noise of sixth order signal present seed is set on 664. On each cluster, a wrapper is used to first simulate the time-series and then analyse each time series with all methods.

Simulations

The methods are tested and compared on multiple differently simulated time-series. These simulations indicate when the methods detect activation and how well they estimate the amplitude of the HRF. The factors manipulated in the simulation of the time series are shown in table 1. Noise was selected as a factor because it is known that even after cleaning the fMRI data several types of noise still exist. Therefore it is important to know how the FHRFE performs with different types of noise. Stimulus interval is selected as a factor since the FHRFE might perform differently on different study designs. In this study the stimulus

interval is the time in seconds between the stimuli (or trials) in the study design. More about the stimulus interval is explained later. Especially when the interval is short and noise is autocorrelated the FHRFE might need more time points to detect the HRF in the data. SNR is selected as a factor because testing the FHRFE on different amounts of noise can indicate the properties of the FHRFE under different amounts of noise.

The factors manipulated for the methods are also shown in table 1. To simulate the variability of the HRF this study used a variable latency and function to shape the HRF. The variations in HRF should impact the BGLM method but not the FHRFE since the first assumes the wrong shape of the HRF and the last doesn't assume a pre-defined shape. This should show the benefits of using the FHRFE over using the BGLM method. Furthermore, the number of folds, times of cross validation and size of test set are explored for the FHRFE to see how these settings influence the method.

Data. This study aims to simulate already cleaned fMRI data and uses a temporal resolution of one second. The time-series are simulated under different conditions as shown in table 1. In total, there are 108 conditions where a signal is present. The conditions where the stimulus interval is 40 seconds or where SNR is five are considered as ideal but not realistic conditions. These are mainly used to see if the methods actually work but not used to indicate and compare sensitivity or amplitude estimation. In total there are 18 conditions where no signal is present. Since in these time-series no signal is present here, the maximum interval and number of trials only impact time-series length. All conditions are simulated 100 times.

Noise. The time-series are simulated within three conditions of noise. Although it is known that multiple types of noise can exist in fMRI data (Lund, Madsen, Sidaros, Luo & Nichols, 2005), to simplify interpretations of the results, the types of noise are not mixed and limited to autocorrelated noise and white noise. It is expected that the autocorrelated noise decreases the sensitivity, specificity and amplitude estimation of the FHRFE.

White noise. In this study, the white noise added to the time series is randomly distributed with a mean of zero and a standard deviation dependent on SNR. This leads to independent errors.

Autocorrelated noise. In this study, the autocorrelated noise added to the time-series generates dependent errors which are not beneficial for the methods, since all methods are based on the assumption of independent errors. The coefficients for the autocorrelated noise are adapted from Sahib et all. (2016). For the autocorrelated noise of the second order, the coefficients 0.624 and 0.003 are used. For the autocorrelated noise of the sixth order the coefficients 0.641, 0.068, -0.088, -0.035, 0.074, -0.010 are used.

Stimulus intervals. In this study, the stimulus interval is the time in seconds between the stimuli (or trials) in the study design. The stimulus intervals chosen are either between 6 and 8 seconds, 10 and 12 seconds or 40 seconds (referred to respectively as "maximum interval" 8, 12 and 40). For the 6-8 and 10-12 second intervals, the intervals are randomly jittered with uniform distribution. This is normally done to have more chance of capturing the peak of the HRF. When not jittered it might occur that the peak of the HRF falls in between scans and is therefore missed. It is expected that the FHRFE performs better when maximum intervals are bigger. This because the FHRFE then has a longer time course to average out the noise.

SNR. In this study, the SNR is defined so that the standard deviation of the noise is equal to the amplitude of the HRF divided by the SNR. Therefore, when SNR increases the amount of noise decreases.

Analysis. Each simulated time series is analysed with all methods. Within the FHRFE there are 16 different settings. Within the BGLM method, there are five different settings. In this study, the latency is simulated by shifting the onsets passed on to the methods while the data stays the same. For example: when one second difference in latency is passed to the method, the onsets in the method are equal to the onsets in the data minus one second.

Latency. As earlier described, several studies showed that there is variability in the HRF between and within subjects. The variable latency of the HRF is approached in this study by altering the onsets passed to the methods. For example, when introducing a difference in latency of one second between the data and the method, the onsets passed to the method are the onsets in the data minus one second. Although it is shown that the difference in latency of the HRF can be more than 4 seconds, especially when only looking at the time of the peak (Lee et. al., 1995), this study used a difference in latency of zero to four seconds to keep the results comprehensible.

Size of the train set. The size of the train set is tested with 90% and 70%. It is expected that a train set of 70% is more sensitive but less specific compared to a test set of 90%. This is expected because the FHRFE averages trials and when more trials are averaged less noise is left. Therefore the chance of a good fit of the estimated model is higher.

K-folds. The number of folds are tested with 5 and 10 folds. It is expected that 5 fold is less sensitive but more specific compared to 10 folds as the same principle occurs here as described for the size of the test set above.

Multiple cross validations. The number of cross validations is tested with 10 and 50 times. 10 times cross validation is expected to be more variable in results compared to 50 times since small deviations are then not as influential to the final result.

Table 1.

Manipulated factors for simulation and analysis

Factors	**Factor values**				
Data					
Noise	White	Ar2	AR6		
Trials	20	50			
Stimulus intervals	40 sec	6 – 8 sec	10-12 sec		
SNR*	1	2	5		
Shape HRF in data*	Double-gamma	Gamma			
Signal	Yes	No			
Analyses					
Size test set	10%	30%			
K - folds	10	5			
Multiple cross validations	10	50			
Difference in latency*	0	1	2	3	4

* Not used when the signal is “No”

Evaluation

Statistical tests. To compare methods, sensitivity, specificity and amplitude estimation are measured of all methods with and without mis specifications. To measure sensitivity, for every condition of the simulation, a count is done how many times the methods indicate a signal when there is a signal in the data. To measure specificity, for every condition of the simulation, a count is done how many times the methods indicate no signal when there is no signal in the data. Since the methods are using the same data, this study can be described as a paired sample design (using the different methods to analyse the same time series).

Therefore, when comparing sensitivity or specificity between methods, a contingency table can show differences between methods. To show actual differences between methods, the marginal proportions of discordant "pairs" are more informative rather than looking at total sensitivity or specificity. Discordant "pairs" in this study are time-series where one method indicated a signal present when the other method did not and vice versa. The matching statistical test to indicate a difference in sensitivity or specificity here is McNemar's test for paired samples. Since McNemar's test can only describe differences between two methods while more than 2 methods are compared in this study, Cochran's Q test is used to test for significant differences between methods in one analysis (McCrum-Gardner, 2008; Dietterich, 1998). Furthermore, post hoc McNemar tests are used to indicate the specific differences between methods.

To compare amplitude estimation, this study uses repeated measures one-way ANOVA to indicate significant differences in mean amplitude estimation between the different latency conditions and different methods. Furthermore, post hoc t-tests are used to indicate the specifiv differences between methods.

To account for multiple testing, this study uses Bonferonni-Holm adjustments to the p-values of the tests. This correction is adapted from Holm (1979) and intends to control for the family-wise error rate and therefore protects against type one errors (rejecting the null hypothesis falsely).

Lastly, it should be noted that all statistical tests used here are dependent on sample size and as described before, when averaging over conditions the sample size in this study can be large. Therefore, extra emphasis should be on the impact of the differences rather than the significance of the differences which will be discussed in the discussion section.

Settings for the Free HRF Estimation. The 16 different settings within the FHRFE are explored and the most beneficial setting is chosen to compare the BGLM method with. To decide which settings are most beneficial for the Free HRF Estimation the different settings are compared on sensitivity and specificity. First, averaging over all settings and conditions, sensitivity and specificity based on MSE and X^2 values are explored. Based on the results a decision is made on which method would be best to compare to the BGLM method.

Second, number of folds, times of cross-validation and size of the training set are visually explored. From this exploration, a decision is made for which settings to choose for the FHRFE.

Results

In this section, the results of the sensitivity and specificity analyses and the analyses of the amplitude estimations for the different methods are presented. Sensitivity refers to the proportion of true positives when the method is applied to data in which activation is simulated. Specificity refers to the proportion of true negatives when there is no activation simulated in the data. First, the performance of the FHRFE and BGLM is explored on ideal conditions. Second, the settings for the Free HRF Estimation are explored and third, the FHRFE is compared to the BGLM method.

Exploration of performance FHRFE and BGLM

Figure 4 shows an example of how the FHRFE estimates the HRF compared to the BGLM when the difference in latency is 3 seconds. The example uses the ideal conditions of an SNR of 5, a maximum interval of 40 and white noise. The figure shows that under these conditions, the estimated HRF by the BGLM differs from the actual HRF in the data. The time of the peak is earlier and the magnitude of the amplitude smaller. The HRF estimated by FHRFE however, closely resembles the HRF in the data.

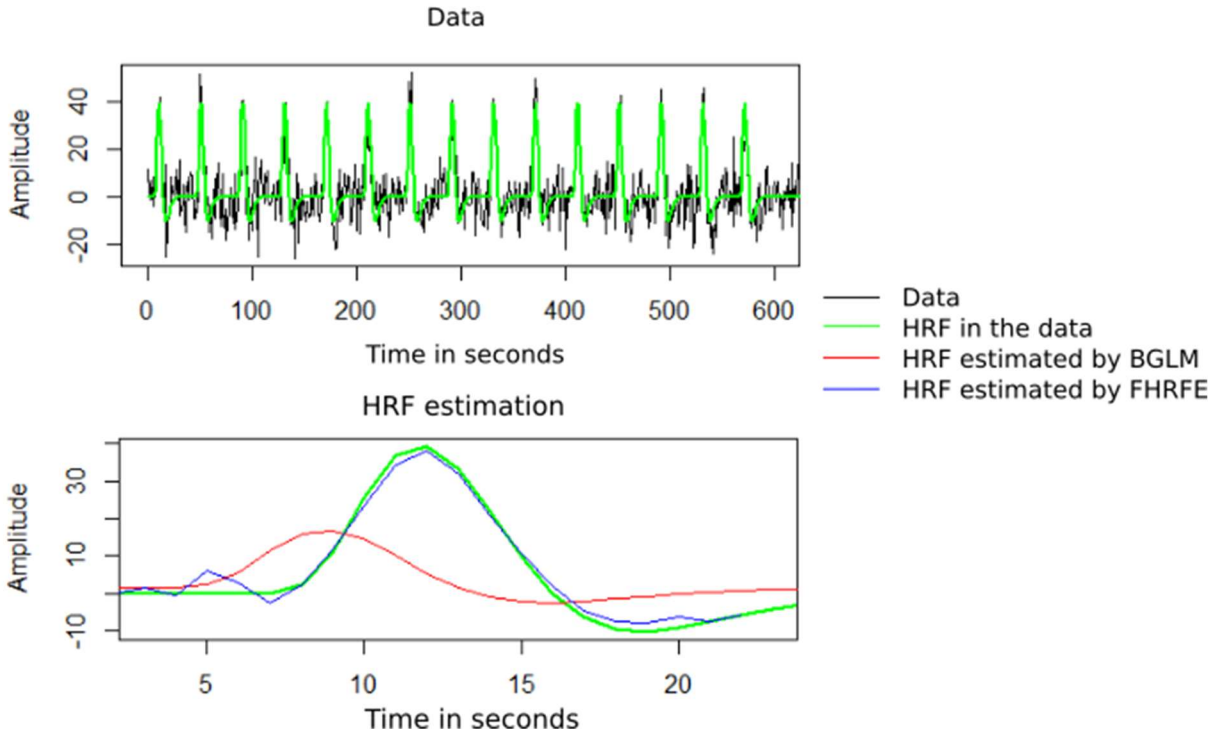


Figure 4. Example of HRF estimation by FHRFE using the MSE method and BGLM. Difference in latency is 3 seconds, SNR is 5, max interval is 40 and noise is white.

Figure 5 shows that all methods perform very well on sensitivity when SNR is 5 and the maximum interval is 40. However, the BGLM shows a lower sensitivity when the difference in latency is 4 seconds while the FHRFE shows almost no change in sensitivity between when the difference in latency is 4 seconds. When looking at amplitude estimation, the FHRFE does not seem to suffer from the difference in latency while the BGLM method shows to worsen in amplitude estimation as the difference in latency increases.

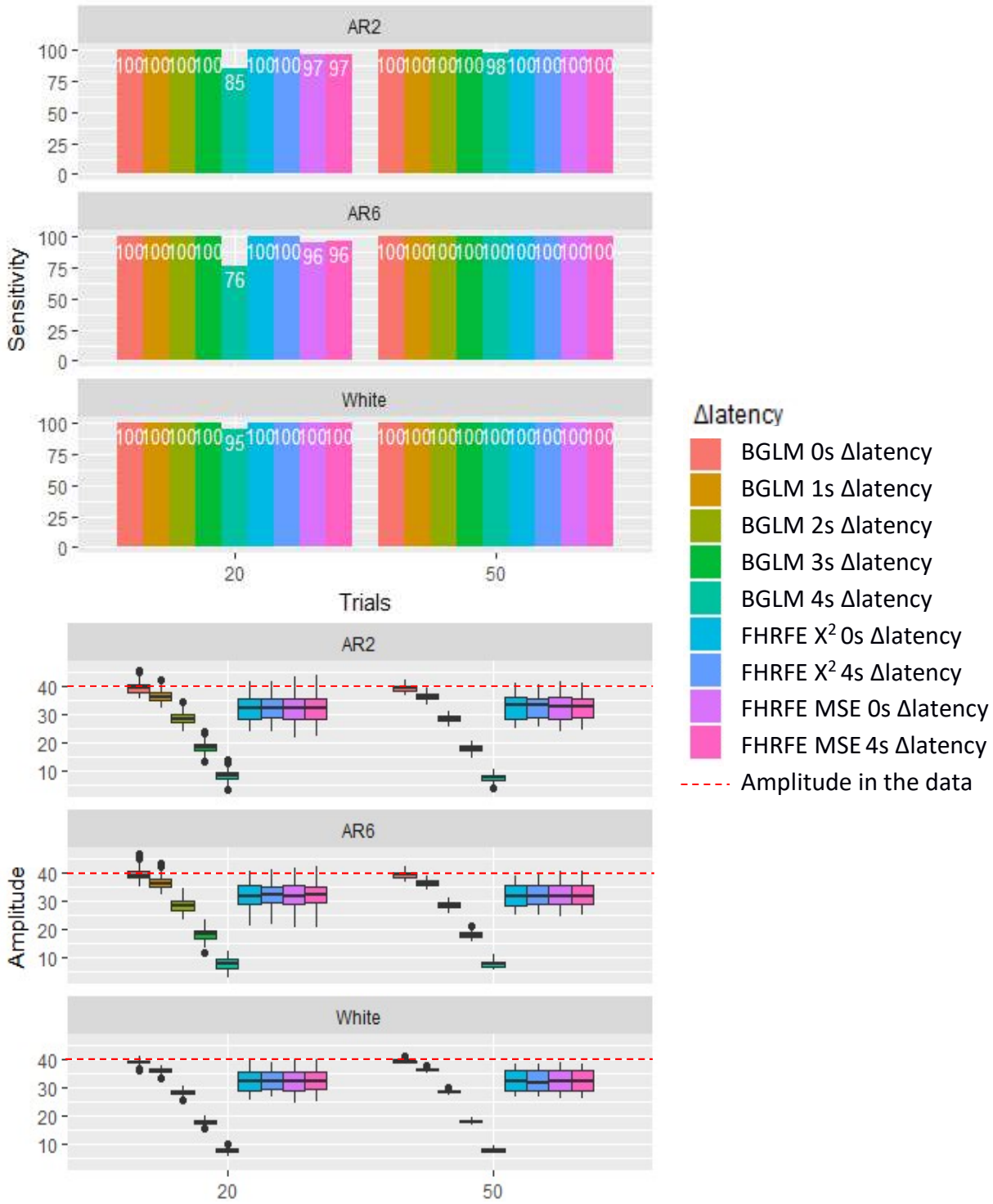


Figure 5. Sensitivity and amplitude estimation of the methods with and without difference in latency for the condition with an SNR of 5 and an max interval of 40 and a double gamma HRF . The amplitude in the data is 40. (n = 600 time series)

Although an SNR of five and an interval of 40 seconds are ideal conditions for all methods, these are not realistic settings for a real study and are only used to give an initial indication of the performance of the methods. Therefore the conditions where SNR is 5 or maximum interval is 40 are excluded from the analyses in further sections.

Best settings for the Free HRF Estimation

In this section of the results, the settings for the Free HRF Estimation are explored and optimal settings are chosen with respect to sensitivity and specificity.

MSE vs X^2.

Sensitivity. Earlier, it was stated that signal detection based on MSE was expected to be less sensitive compared to signal detection based on X^2. Figure 6 and table 1 show that averaged over all settings and conditions, detection of signal based on the X^2 value is more sensitive compared to detection based on MSE ($\Delta P = -62.86$). Detection based on MSE detects a signal when detection based on X^2 doesn't for 0.15% of the time while it's 62.91% otherwise. Figure 3 also shows that when the maximum interval increases from eight to twelve, number of trials increases from 20 to 50 or SNR increases from one to two, the sensitivity also increases for both methods. When SNR is two and when maximum interval is 12 seconds, signal detection based on X^2 value reaches a sensitivity of 100% at 50 trials.

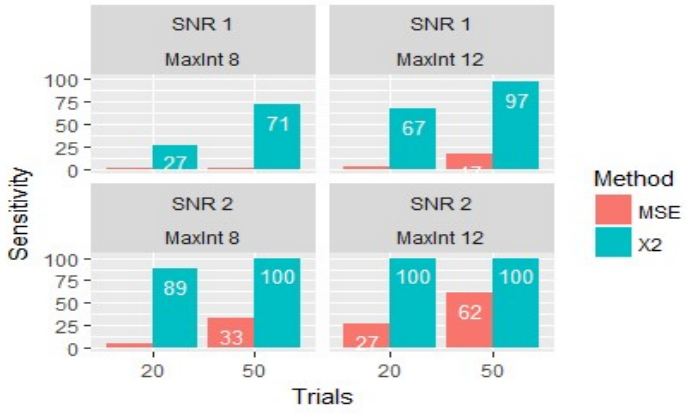


Figure 6. Sensitivity of signal detection based on MSE and X^2 values. Averaged over noise (White, AR2, AR6) conditions and over folds (5,10), times (10, 50) and size of training set (70%, 90%) settings. *(n = 8 settings x 1200 time series = 9600 analyses per method).* (n = 8 settings x 4800 time series = 38400 analysis for each method).

Table 1.

X2 and MSE Sensitivity

		X^2		
		Sign	***Non - Sign***	
	Sign	18.41	0.15	18.56
MSE	***Non - Sign***	62.91	18.53	81.56
		81.32	18.68	

Notes. All values are in proportions of total and averaged over all conditions. Difference in marginal proportions: ΔP = - 62.86. Averaged over number of trials (20,50), noise (White, AR2, AR6), SNR (1, 2) and max interval (8, 12) conditions and over folds (5,10), times (10, 50) and size of training set (70%, 90%) settings. (n = 8 settings x 1200 time series = 9600 analyses per method). (n = 8 settings x 4800 time series = 38400 analysis for each method).

Specificity. Earlier it was stated that that signal detection based on MSE was expected to be more specific compared to signal detection based on X^2. Figure 7 and table 2 show that averaged over all settings and conditions, signal detection based on MSE is more specific than detection of signal based on X^2 when looking at figure 4. Detection based on MSE doesn't detect a signal when detection based on X2 does for 16.50% of the time while it's 0.83% otherwise. Figure 4 also shows that in general, signal detection based on X^2 is less specific when the noise is autocorrelated compared to when the noise is white. Also, when space between intervals increases from a maximum interval of eight to twelve, the specificity of signal detection based on X2 value also increases in general. The number of trials doesn't seem to have a big impact on specificity.

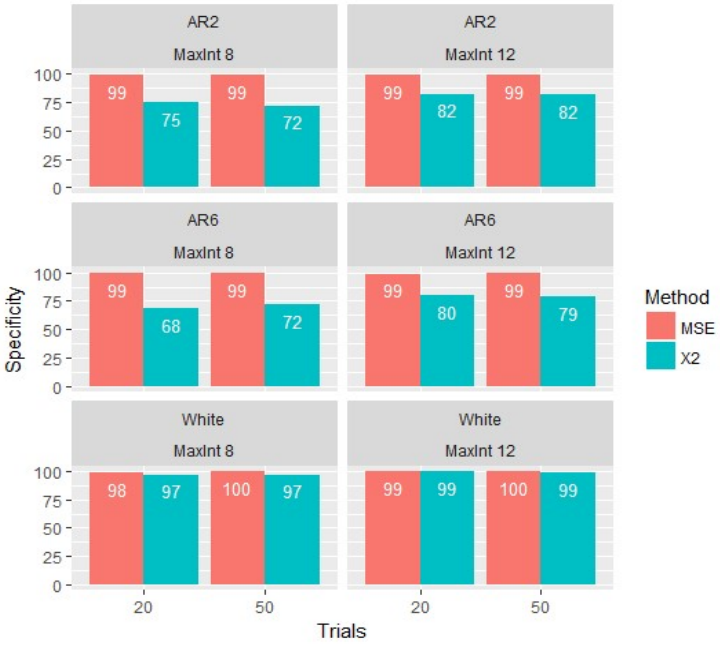


Figure 7. Specificity of signal detection based on MSE and X^2 values. The trials and maximum interval here are only referring to the length of the simulated time series since no signal was present in the data. Averaged over number of trials (20,50), noise (White, AR2, AR6) and max interval (8, 12) conditions and over folds (5,10), times (10, 50) and size of training set (70%, 90%) settings. (n = 8 settings x 1200 time series = 9600 analyses per method).

Table 2.

Specificity of signal detection based on X^2 and MSE

		X2		
		Sign	Non - Sign	
MSE	Sign	0.01	0.83	0.84
	Non - Sign	16.50	82.66	99.16
		16.51	83.49	

Notes. All values are in proportions of total. Difference in marginal proportions: ΔP = -15.67 Averaged over number of trials (20,50), noise (White, AR2, AR6) and max interval (8, 12) conditions and over folds (5,10), times (10, 50) and size of training set (70%, 90%) settings. (n = 8 settings x 1200 time series = 9600 analyses per method).

X2 over MSE. Bared on sensitivity (true positive rate), this study prefers to base signal detection on the X^2 value (MSE: 18.56% vs X^2 : 81.36%).. In contrary, based on specificity (true negative rate) this study prefers to base signal detection on MSE-value (MSE: 99.16% vs X^2 : 83.49%). When looking at both sensitivity and specificity, this study preferred to base signal detection on the X^2 value over the MSE value. This choice was made since, although very specific, basing the signal detection on MSE value has shown to lack a lot in sensitivity, especially when SNR was one. Therefore, to keep further sections clear, the MSE method is excluded in the following results. The results of signal detection based on MSE-value are shown in appendix A and B.

Folds, times and size of the training set.

Sensitivity. Earlier it was stated that number of folds, times of cross-validating, and size of training data could have an effect on sensitivity. It was expected that five folds would be less sensitive than 10 folds, 10 times less sensitive than 50 times and a training set of 70% was expected to be more sensitive than 90%. Figure 8 shows that, when SNR is two, the settings of the Free HRF Estimation don't seem to make a big difference when looking at sensitivity. However, when SNR is one, figure 8 shows that sensitivity is highest when using 70% of the data for training. Number of folds and times of cross validation don't seem to have a big impact on sensitivity in general. When SNR increases from one to two, figure 8 also shows that sensitivity for all settings increases from a minimum of 56% to a minimum of 95%.

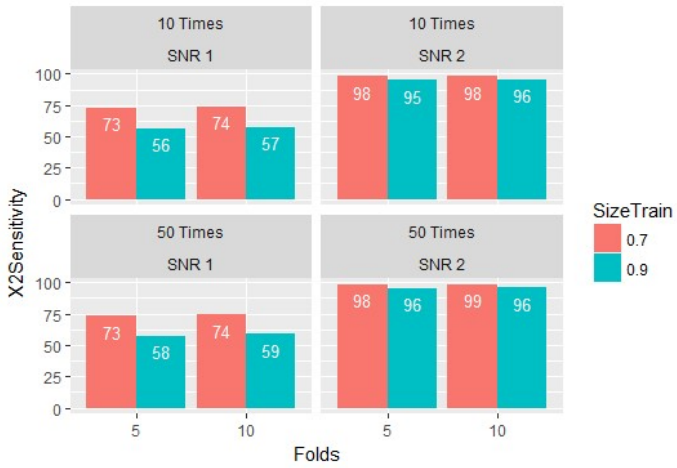


Figure 8. Sensitivity of signal detection based on X^2, the different settings compared. Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 8 settings x 4800 time series = 38400 analyses).

Specificity. Earlier it was stated that number of folds, times of cross-validating, and size of training data could have an effect on specificity. It was expected that five folds would be more specific than 10 folds, 10 times more specific than 50 times and a training set of 70% was expected to be less specific than 90%. Figure 9 shows that folds and times don't seem to have a big impact on specificity. However, figure 9 does show that when 90% of the data is used for training the method is most specific. Also, figure 9 shows that specificity decreases when the noise is autocorrelated compared to white noise for all settings.

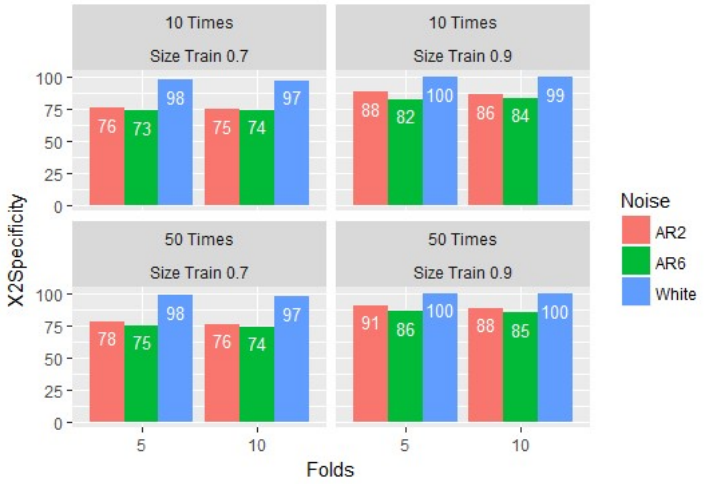


*Figure 9. S*pecificity of signal detection based on X^2 with the different settings compared. Averaged over number of trials (20,50) and max interval (8, 12) conditions. (n = 8 settings x 1200 time series = 9600 analyses).

5 folds 10 times with training on 70% of data preferred. Based on the results above, the 5 folds, 10 times with training on 70% of the data settings are preferred by this study. This choice is made since the 5 fold 10 times is less computationally intensive and the 70% training data is preferred since this setting was more sensitive. The preferred method has a general sensitivity of 85.5% and a specificity of 82.3% when averaged over all conditions. Figure 10 shows that the sensitivity of the FHRFE seems to suffer from auto-correlated noise when SNR is low or the maximum interval is 8. To keep further sections clear, only this method is used in the following results.

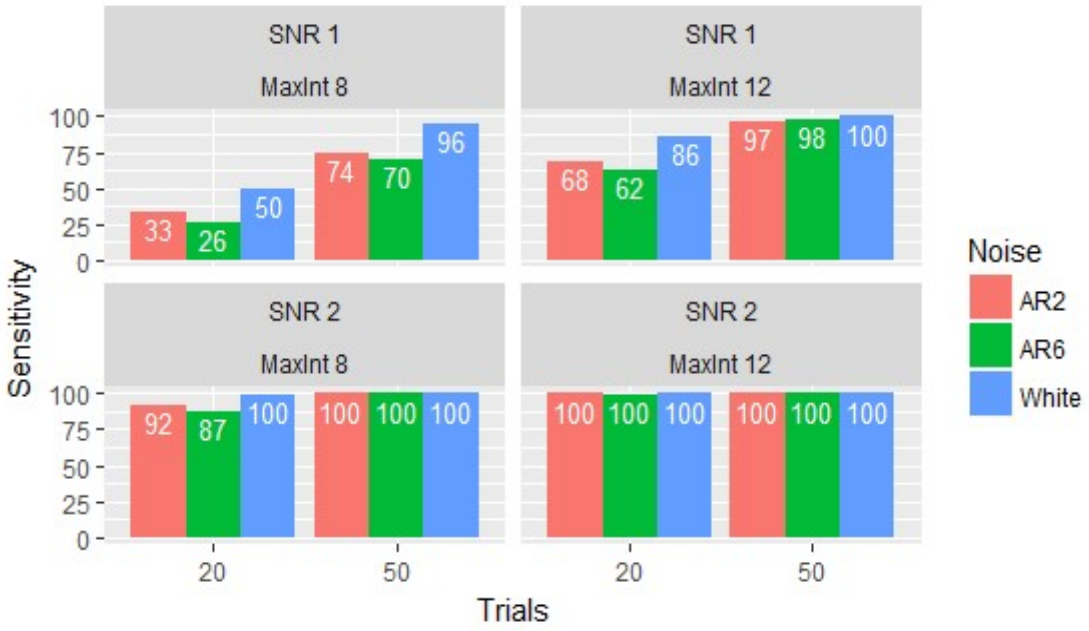


Figure 10. Sensitivity of the preferred settings for the FHRFE method (5 fold, 10 times and 70% training data). Averaged over HRF shape (double-gamma, single-gamma) condition. (n = 4800 time series).

GLM general evaluation

In this section of the results, the general sensitivity and specificity of the BGLM are evaluated.

Sensitivity. Figure 11 shows the general sensitivity of the BGLM method. Earlier it was stated that the BGLM is expected to be very sensitive overall conditions averaged when the shape of the HRF in the data is a double-gamma. When the shape of the HRF in the data is a single-gamma, it is expected that the BGLM would be less sensitive since the shape fitted doesn't match the shape in the data.

Figure 11 shows that sensitivity is very high in all conditions when she shape of the HRF in the data is a double-gamma. When the shape of the HRF in the data is a single-gamma, figure 11 shows that the sensitivity can vary between conditions with a sensitivity between 29% and 100%. Sensitivity is especially lowest when the maximum interval is eight seconds, with an SNR of one, 20 trials and white noise. Sensitivity increases in general when SNR increases from one to two, maximum interval increases from eight to 12 and trials increase from 20 to 50.

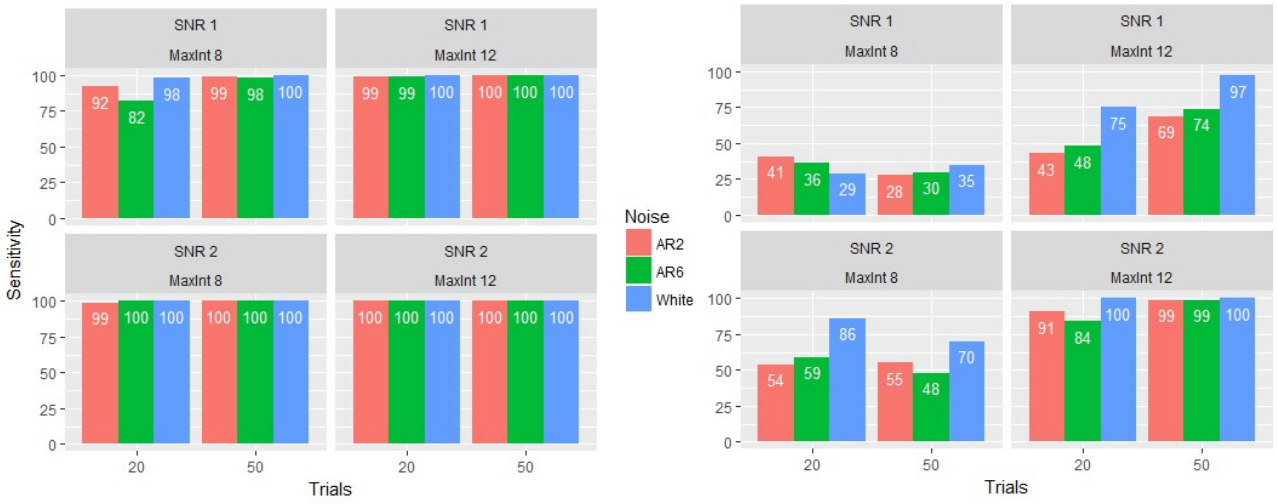


Figure 11. Sensitivity BGLM. The bar plots on the left are for the double-gamma and the bar plots on the right are for the single-gamma HRF in the data. (n = 2400 time series)

Specificity. Figure 12 shows that specificity is high for the white noise condition and that the BGLM method suffers from autocorrelated noise. The length of the time-series, represented by the "maximum interval" and "number of trials" doesn't seem to have a big impact on specificity.

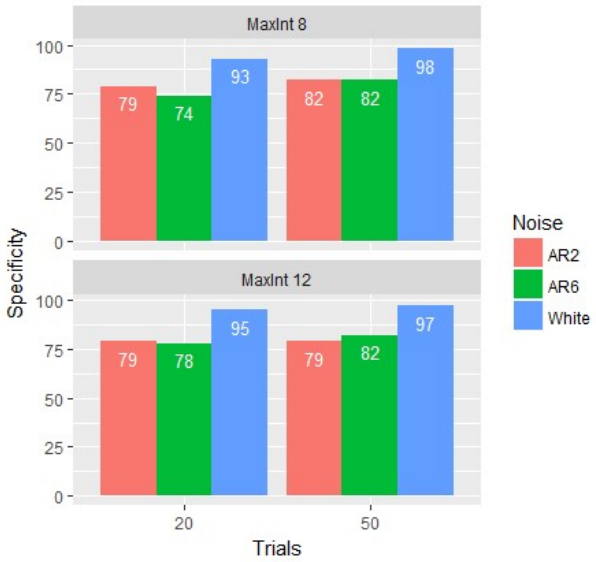


Figure 12. Specificity of the BGLM. (n = 1200 time series)

Effect of difference in latency and shape of the HRF on the methods

In this section of the results, the effect of the difference in latency on the sensitivity of the Free HRF Estimation and on the BGLM are shown.

Free HRF Estimation. To test whether the difference in latency and shape of the HRF in the data have an impact on the sensitivity of the Free HRF Estimation, sensitivity was compared between the Free HRF Estimation with zero and four seconds difference in latency. On this regard, the Free HRF Estimation was expected to show its greatest benefit. Namely, the difference in latency and shape should not have any influence on the sensitivity of the Free HRF Estimation. Table 3 and table 4 show a small difference in marginal proportions when averaging over all conditions for the Free HRF Estimation when comparing no difference in latency and four seconds difference in latency. When the shape of the HRF in the data is a double-gamma, table 3 shows a general sensitivity of 86.29% for zero and 85.13% for four seconds difference in latency with a difference in proportions of 1.17%. When the shape of the HRF is a single-gamma, table 4 shows a general sensitivity of 85.29% for zero and 83.33% for four seconds difference in latency with a difference in proportions of 1.96%.

Table 3.

Effect of difference in latency on signal detection for the Free HRF Estimation. HRF shape: double-gamma

		Four seconds difference in latency		
		Sign	***Non - Sign***	
No difference in latency	***Sign***	82.75	3.54	86.29
	Non - Sign	2.38	11.33	13.71
		85.13	14.88	

Notes. All values are in proportions of total.

Difference in marginal proportions: $\Delta P = 1.17$.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

Table 4.

Effect of difference in latency on signal detection for the Free HRF Estimation. HRF shape: single-gamma

		Four seconds difference in latency		
		Sign	***Non - Sign***	
No difference in latency	***Sign***	80.88	4.42	85.29
	Non - Sign	2.46	12.25	14.71
		83.33	16.67	

Notes. All values are in proportions of total.

Difference in marginal proportions: ΔP = 1.96.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

GLM. Table 5 and Table 6 show the comparison between the BGLM with no difference in latency and with one, two, three and four seconds difference in latency. Earlier it was stated that difference in latency is expected to decrease the sensitivity of the BGLM. When the shape of the HRF in the data is a double-gamma, table 5 shows that three seconds difference in latency has the biggest impact on sensitivity with 98.58% sensitivity for no difference in latency and 18.04% sensitivity for three seconds difference in latency (ΔP = 80.54). Here, the BGLM with no difference in latency detects a signal when the BGLM with three seconds difference in latency doesn't for 80.96% of the time while it's 0.42% otherwise. One second difference in latency doesn't seem to have a big influence on sensitivity with 96.33% sensitivity for the BGLM with one second difference in latency (ΔP = 2.25). Here, the BGLM with no difference in latency detects a signal when the BGLM with one second difference in latency doesn't for 2.33% of the time while it's 0.08% otherwise.

When the shape of the HRF in the data is a single-gamma, table 6 shows that the difference in latency actually benefits the model when looking at sensitivity. Two seconds difference in latency seems to be the most beneficial with 64.58% sensitivity for the BGLM with no difference in latency and 98.58% sensitivity for the BGLM with two seconds difference in latency ($\Delta P = -34.00$). Here, the BGLM with no difference in latency detects a signal when the BGLM with two seconds difference in latency doesn't for 0.08% of the time while it's 34.08% otherwise.

Table 5.

Effect of difference in latency on signal detection for the BGLM. HRF shape: double-gamma.

		Seconds difference in latency								**Total**
		1		**2**		**3**		**4**		
		S	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	
No diff	***S***	96.33	2.33	75.08	23.50	17.62	80.96	67.17	31.42	98.58
	NS	0.08	1.33	0.21	1.21	0.42	1.00	0.63	0.79	1.42
Total		96.33	3.67	75.29	24.71	18.04	81.96	67.79	32.21	

Notes. All values are in proportions of total.

S = significant detection of signal, NS = non-significant detection of signal

Differences between no difference in latency and difference in latency in marginal proportions: one second difference in latency: $\Delta P = 2.25$; two seconds difference in latency: $\Delta P = 23.29$; three seconds difference in latency: $\Delta P = 80.54$; four seconds difference in latency: $\Delta P = 30.79$.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

Table 6.

Effect of difference in latency on signal detection for the BGLM. HRF shape: single-gamma.

		Seconds difference in latency								Total
		1s		2s		3s		4s		
		S	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	
No diff	***S***	64.29	0.29	64.50	0.08	63.75	0.83	56.38	8.21	64.58
	NS	31.13	4.29	34.08	1.33	33.12	2.29	26.38	9.04	35.42
Total		95.42	4.58	98.58	1.42	96.88	3.12	82.75	17.25	

Notes. All values are in proportions of total.

S = significant detection of signal, NS = non-significant detection of signal

Differences between no difference in latency and difference in latency in marginal proportions: one second difference in latency: $\Delta P = -30.83$; two seconds difference in latency: $\Delta P = -34.00$; three seconds difference in latency: $\Delta P = -32.29$; four seconds difference in latency: $\Delta P = -18.17$.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

BGLM and Free HRF Estimation Compared.

In this section of the results, the Free HRF Estimation is compared to the BGLM. It was decided to use the Free HRF Estimation with four seconds difference in latency for comparison with the BGLM. This to be sure that even with the difference in latency the Free HRF Estimation offers the solution it was designed for.

Sensitivity. Earlier it was stated that the Free HRF Estimation would be more sensitive compared to the BGLM when the shape of the HRF and time of onset in the data don't match the shape and time of onset in the BGLM. Table 7 and figure 13 show that when the shape of the HRF in the data is a double-gamma the Free HRF Estimation has a higher sensitivity than the BGLM when the difference in latency in the BGLM is two, three or four seconds. When the shape of the HRF in the data is a single-gamma, table 8 and figure 13 show that the Free HRF Estimation only has a notable higher sensitivity compared to the BGLM when there is actually no difference in latency in the BGLM. When there's one, two

or three seconds difference in latency in the BGLM the BGLM has notable higher sensitivity compared to the Free HRF Estimation.

Table 7.

Sensitivity of signal detection with BGLM and Free HRF Estimation for the difference in latency. Shape HRF: double-gamma

		BGLM										**FHRFE**		
		0s diff		**1s diff**		**2s diff**		**3s diff**		**4s diff**		**0s diff**		
		S	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	**Total**
FHRFE	***S***	85.00	0.12	84.08	1.04	69.04	16.08	16.25	68.88	65.21	19.92	82.75	2.38	85.12
	NS	13.58	1.29	12.25	2.62	6.25	8.62	1.79	13.08	2.58	12.29	3.54	11.33	14.87
Total		98.58	1.42	96.33	3.67	75.29	24.71	18.04	81.96	67.79	32.21	86.29	13.71	

Notes. All values are in proportions of total and the Free HRF Estimation is with four seconds difference in latency.

Cochran's Q $X^2(6) = 6228.871$, $p < 0.001$.

Differences between FHRFE with four seconds difference in latency and BGLM method:

no difference in latency: $\Delta P = -13.46$, OR = 0.009, $p < 0.001$,

one second difference in latency: $\Delta P = -11.21$, OR = 0.085, $p < 0.001$

two seconds difference in latency: $\Delta P = 9.83$, OR = 2.573, $p < 0.001$

three seconds difference in latency: $\Delta P = 67.08$, OR = 38.442, $p < 0.001$

four seconds difference in latency: $\Delta P = 17.33$, OR = 7.710, $p < 0.001$

FHRFE no difference in latency: $\Delta P = -13.46$, OR = 0.671, $p = 0.023$, $p_{adjusted} = 0.023$

All p-values < 0.001 were also < 0.001 after adjustment.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

Table 8.

Sensitivity of signal detection with BGLM and Free HRF Estimation for the difference in latency. Shape HRF: single-gamma

		BGLM										FHRFE		
		0s diff		1s diff		2s diff		3s diff		4s diff		0s diff		
		S	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	***S***	***NS***	**Total**
FHRFE	***S***	57.67	25.67	81.54	1.79	83.17	0.17	82.71	0.62	73.33	10.00	80.88	2.46	83.33
	NS	6.92	9.75	13.88	2.79	15.42	1.25	14.17	2.50	9.42	7.25	4.42	12.25	16.77
Total		64.58	35.42	95.42	4.58	98.58	1.42	96.88	3.12	82.75	17.25	85.29	14.71	

Notes. All values are in proportions of total and the Free HRF Estimation is with four seconds difference in latency.

Cochran's Q: $X^2(6) = 2190.418$, $p < 0.001$.

Differences between FHRFE with four seconds difference in latency and BGLM method:

no difference in latency: P = 18.75, OR = 3.711, $p < 0.001$,

one second difference in latency: $\Delta P = -12.09$, OR = 0.129, $p < 0.001$

two seconds difference in latency: $\Delta P = -15.25$, OR = 0.011, $p < 0.001$

three seconds difference in latency: $\Delta P = -13.55$, OR = 0.044 , $p < 0.001$

four seconds difference in latency: $\Delta P = 0.58$, OR = 1.062 , $p = 0.547$

FHRFE no difference in latency: $\Delta P = -1.96$, OR = 0.557, $p < 0.001$

All significant p values were also < 0.001 after adjustment

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions.

(n = 2400 time series)

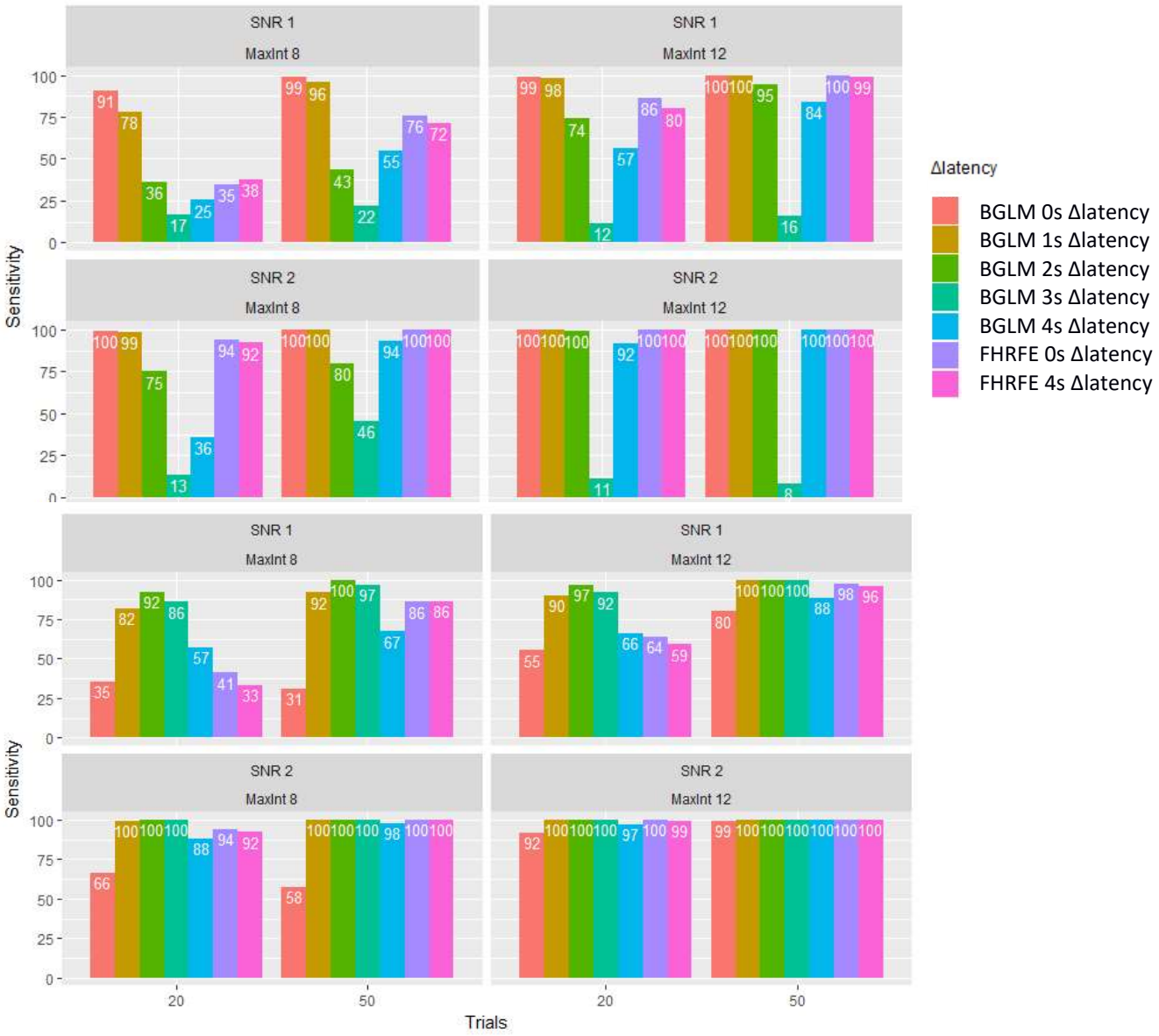


Figure 13. Sensitivity Free HRF Estimation and BGLM with zero to four seconds lag. The top four plots are with a double-gamma shape in the data and the bottom four with a single-gamma shape in the data. Averaged over noise (White, AR2, AR6) conditions. (n = 4800)

Specificity. Earlier no specific statement was made about the expected specificity. Table 9 shows that overall conditions averaged, the BGLM method is more specific with 84.83% versus 77.83% specificity for the Free HRF Estimation (ΔP = -7.00). Here the Free HRF Estimation detects a signal when the BGLM method doesn't for 19.25% of the time while it's 12.25% otherwise. Figure 14 also shows that autocorrelated noise reduces specificity for both methods compared to white noise. Also, the Free HRF Estimation seems to benefit from a longer interval between supposed onsets which basically means a longer time series.

Table 9.

Specificity Free HRF Estimation and BGLM.

		GLM		
		Sign	***Non - Sign***	
FHRFE	***Sign***	2.92	19.25	22.17
	Non - Sign	12.25	65.58	77.83
		15.17	84.83	n = 1200

Notes. All values are in proportions of total.

McNemar test: $X^2(1) = 18.225$, $p < 0.001$, p adjusted < 0.001

Difference in marginal proportions: ΔP = 7.00, OR = 0.63.

Averaged over number of trials (20,50), noise (White, AR2, AR6) and max interval (8, 12) conditions. (n = 1200 time series)

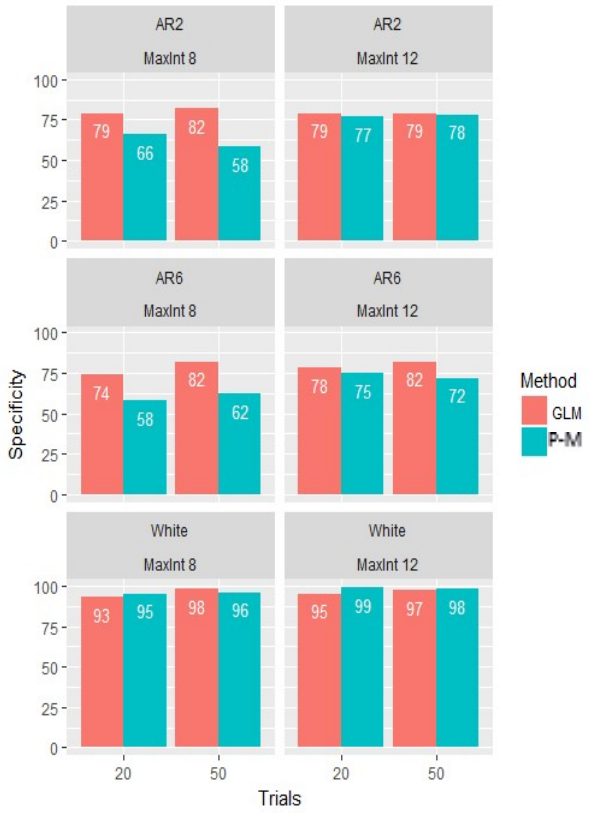


Figure 14. Specificity Free HRF Estimation and BGLM.

Amplitude estimation. Earlier it was stated that the Free HRF Estimation was expected to be more precise in estimating the amplitude compared to the BGLM when the shape of the HRF and time of onset in the data didn't match the shape and time of onset in the BGLM. Since the amplitude in the data was 40, the closer to 40 the more precise.

Table 10 and figure 15 show that in general, when the shape of the HRF in the data is a double-gamma, the Free HRF Estimation is more precise in amplitude estimation compared to the BGLM when the difference in latency in the BGLM is two seconds or higher (2s, $\Delta M = 6.78$; 3s, $\Delta M = 15.57$; 4s, $\Delta M = 12.13$). Also, figure 15 shows that when the number of trials increases from 20 to 50 or the SNR increases from one to two, the spread of the estimated amplitudes seems to decrease.

When the shape of the HRF in the data is a single-gamma, the Free HRF Estimation is only more precise compared to the BGLM with zero and four seconds Difference in latency (0s, $\Delta M = 24.89$;4s, $\Delta M = 28.83$). Also, when the number of trials increases from 20 to 50 or the SNR increases from one to two, the spread of the estimated amplitudes seems to decrease.

Table 10.

Amplitude estimation BGLM in comparison to Free HRF Estimation Shape HRF: double-gamma.

	GLM					FHRFE
	0s difference	***1s difference***	***2s difference***	***3s difference***	***4s difference***	***0s difference***
Mean (*SD*)	39.17 (7.03)	34.57 (6.99)	23.12 (6.72)	14.33 (6.15)	17.77 (6.15)	30.68 (8.09)
Mean$_{dif}$ (*SD*)	-9.26 (4.91)	-4.66 (5.46)	6.78 (6.79)	15.57 (8.42)	12.13 (7.40)	-0.78 (4.34)
CohensD	-1.89	-0.85	0.99	1.85	1.64	-0.17

Note. The mean difference is compared to the mean amplitude estimation of the Free HRF Estimation (*M* = 29.90, *SD* = 6.96).

Differences between amplitudes significant with F(6) = 1107, p < 0.001, p adjusted < 0.001. All paired t tests were significant with normal and adjusted p < 0.001.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

Table 11.

Amplitude estimation BGLM in comparison to Free HRF Estimation; shape HRF: single-gamma

	GLM					FHRFE
	0s difference	***1s difference***	***2s difference***	***3s difference***	***4s difference***	***0s difference***
Mean (*SD*)	24.89 (6.77)	36.22 (7.72)	42.10 (8.59)	38.99 (8.65)	28.83 (7.89)	35.32 (7.22)
Mean$_{dif}$ (*SD*)	9.28(7.48)	-2.04 (6.38)	-7.93 (6.03)	-4.82 (6.20)	5.35 (6.77)	-1.15 (4.02)
CohensD	1.24	-0.32	-1.32	-0.77	0.79	0.26

Note. Mean difference is compared to the mean amplitude estimation of the Free HRF Estimation with four seconds difference in latency. (*M* = 34.18, *SD* = 7.24).

Differences between amplitudes significant with F(6) = 379.5, p < 0.001. All paired t-tests were significant with normal and adjusted p < 0.001.

Averaged over number of trials (20,50), noise (White, AR2, AR6), max interval (8, 12) and SNR (1,2) conditions. (n = 2400 time series)

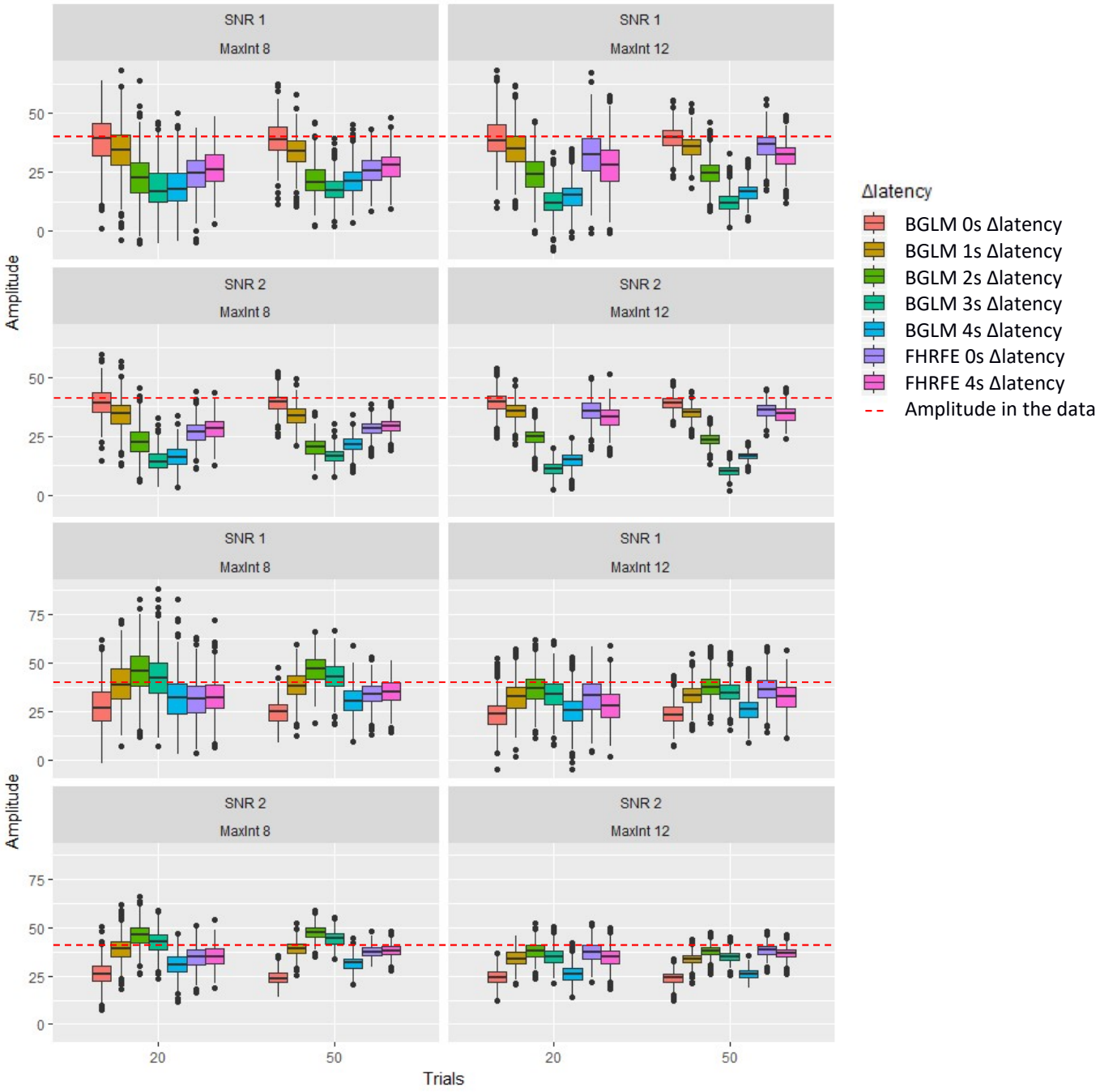


Figure 15. Amplitude estimation for the Free HRF Estimation and the BGLM with zero to four seconds difference in latency. The top four plots are with a double-gamma shape in the data and the bottom four with a single-gamma shape in the data. Averaged over all noise (White, AR2, AR6) conditions. (n = 4800 time series).

Discussion

In this section, the results and the limitations of this study are summarized and discussed and recommendations are done to improve the FHRFE and for future research.

General results

When testing the FHRFE on "ideal" data with an SNR of five and an interval of 40 both signal detection based on MSE as X^2 value showed promising results. However, differences showed when inspecting the results of the methods applied to less ideal data. The results showed that signal detection based on MSE value was a lot less sensitive but more specific compared to signal detection based on X^2 value. The lack in sensitivity of signal detection based on MSE value outweighs the high specificity and therefore, this study chose to continue with signal detection based on X^2 value.

The low sensitivity and the high specificity of signal detection based on MSE value indicate that this method might be too strict when evaluating the time-series. As discussed in more detail in the method section, signal detection based on MSE value compares the obtained model with the intercept-only model for each fold. When the obtained model has a lower MSE compared to the intercept-only model, one vote is added in favour of activation. In the final evaluation of the models, 95% of the votes have to be in favour of activation to reject the null hypotheses of no activation

Regarding the times, folds and size of training set, this study prefers to use the FHRFE with 5 folds, 10 times and 70% training set. Since only the size of the training set seemed to impact sensitivity and specificity, 5 folds and 10 times is least computationally intensive. The bar plots indicated that a training set of 70% was more sensitive but less specific compared to a training set of 90%. This difference in sensitivity seemed bigger when the SNR was one compared to two. This study chose to prefer a training set of 70% since this method seemed to have a slightly better sensitivity versus specificity ratio. However, since no specific test was used, the true difference remains arbitrary. The difference in sensitivity and specificity can best be explained by the size of the test set. When using 90% to train, 10% is left for the test set. Since the parts in the test set are averaged, the fewer parts assigned to the test set, the more variable the test set becomes. This makes a match between the test set and the training set less likely and thus, there is less chance of rejecting the null hypotheses of no activation.

The general evaluation of the BGLM method shows, as expected, a very high sensitivity when the shape in the data is a double-gamma. Only when the maximum interval was eight with an SNR of 1 the sensitivity seemed slightly lower when noise was

autocorrelated. This was expected since the data is analysed with exactly the same model as the data was simulated with. When the shape of the HRF in the data was a single-gamma the sensitivity of the BGLM method dropped. Especially when the maximum interval was short and the noise was high. The specificity of the BGLM method was high for white noise but seemed, although still acceptable, to suffer from autocorrelated noise.

When evaluating the effect of the difference in latency on the FHRFE the FHRFE with four seconds difference in latency performs slightly worse on sensitivity compared to zero seconds difference in latency for both HRF shapes in the data. The effect of the difference in latency on the BGLM method, however, seems bigger. When the shape of the HRF in the data is a double-gamma, 1 second difference in latency doesn't seem to impact sensitivity much, 2 seconds and 4 seconds difference in latency decrease sensitivity a lot and three seconds difference in latency decreases the sensitivity so much it drops to 18%. The higher sensitivity when there is 4 seconds difference in latency in the model compared to 3 seconds can best be explained in figure 16. Figure 16 shows that when the model is specified with 4 seconds difference in latency, the peak of the HRF actually captures a part of the undershoot of the data. The result is a bigger chance of significance but a negative Beta. Thus, although four seconds difference in latency seems to decrease sensitivity less compared to 3 seconds difference in latency, it does so with a wrong model.

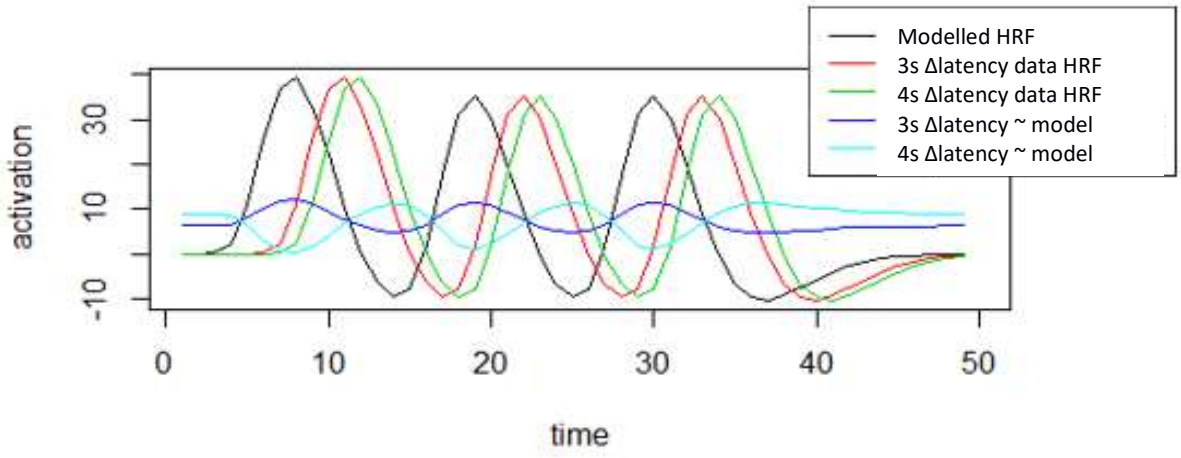


Figure 16. Illustration of the effect of difference in latency for the BGLM

When the shape of the HRF in the data is a single gamma, the difference in latency is actually beneficial to the model. No difference in latency has lower sensitivity compared to 1 to 4 seconds difference in latency. This might be due to the fact that the model analyses the single-gamma data with a double-gamma. Since, the peak of the double-gamma occurs later compared to the peak of the single-gamma, introducing differences in latency to the model

improves sensitivity in this case. These results show how sensitive the BGLM method can be to misspecification of the model.

When comparing the FHRFE with zero and four seconds difference in latency and the BGLM method with zero to four seconds difference in latency, Cochran's Q test indicated a significant difference between the methods when the shape of the HRF in the data is double-gamma. Also, post hoc McNemar tests showed that all methods differed significantly. Meaning, there is a difference in sensitivity for all difference in latency conditions within and between methods. Although not expected, the sensitivity of the FHRFE with no difference in latency is significantly higher compared to the FHRFE with 4 seconds difference in latency. However, the actual difference ($\Delta P = 1.17$) can be considered as small compared to the impact that difference in latency has on the BGLM method (one second difference in latency: $\Delta P = 2.25$; two seconds difference in latency: $\Delta P = 23.29$; three seconds difference in latency: $\Delta P = 80.54$; four seconds difference in latency: $\Delta P = 30.79$). Furthermore, the differences in sensitivity showed that when the shape of the HRF in the data is a double-gamma the FHRFE only outperformed the BGLM method when the difference in latency in the BGLM method was 2, 3 or 4 seconds.

When the shape of the HRF in the data is a single-gamma. Cochran's Q test indicated a significant difference between the methods in sensitivity. The post hoc McNemar tests showed a significant difference in sensitivity between all methods except between the FHRFE with no difference in latency and the FHRFE with 4 seconds difference in latency. Here, the FHRFE with four seconds difference in latency only outperformed the BGLM method when there was no difference in latency in the BGLM method. This can be explained by the fact that the difference in latency actually seems to benefit the BGLM method when the shape of the HRF in the data is single-gamma because the peak of the single-gamma HRF is later.

When comparing specificity between the FHRFE and the BGLM method, the BGLM method is more specific. The FHRFE especially seems to suffer on this regard, when noise was autocorrelated and the time series was short. An explanation could be that when using the X^2 evaluation the models are fitted and tested on the same data. This can lead to overfitting and, as shown, a lower specificity compared to the BGLM method (77.83% vs 84.43%, $\Delta P = 7.00$).

A repeated measures ANOVA showed significant differences in amplitude estimation between methods for both HRF shapes in the data. Also, for both shapes post hoc t-tests show significant differences between all methods. As also when comparing sensitivity, the estimation of the amplitude significantly differed between the FHRFE with no difference in

latency and the FHRFE with four seconds latency (MΔamplitude = 0.78). However, the difference itself can be considered as small compared to the impact of difference in latency on the BGLM method (one second difference in latency: MΔamplitude = 4.6; two seconds difference in latency: MΔamplitude = 16.05; three seconds difference in latency: MΔamplitude = 24.84; four seconds difference in latency: MΔamplitude = 21.4). Furthermore, when the shape of the HRF in the data is a double-gamma, the FHRFE is more precise in amplitude estimation compared to the BGLM method when the difference in latency in the BGLM method is two seconds or higher.

When the shape of the HRF in the data is a single-gamma, the FHRFE is more precise in amplitude estimation compared to the BGLM method when there is no or four second difference in latency in the BGLM method. The boxplots indicate that an increase in SNR, number of trials or maximum interval mainly seemed to decrease the standard deviation of the amplitude estimation but doesn't seem to change the mean estimate.

The plots indicate that study design and noise can influence the sensitivity-specificity and amplitude estimation of the FHRFE. Figure 10 showed that the Free HRF Estimation becomes more sensitive and specific when the number of trials and maximum interval increases. Also, sensitivity and specificity seem to suffer more from autocorrelated noise compared to white noise.

Limitations and recommendations

The first obvious limitation of this study is the use of simulated data. Although the use of simulated data has the benefit that the true signal is known, there can still be unknown differences between the simulated data and real data. When testing on simulated data, characteristics of the tested method might be missed that could be a problem when used in real data. Therefore, a comparison of the FHRFE with another established method on real data would complement the current study. Second, choices were made about the conditions the methods were tested. This had to be done to keep computation time feasible and the results clear for this study. One factor that could be further investigated is noise conditions. This study used conditions where autocorrelated noise and white noise were separated. Furthermore, no other types of noise were included. However, Lund et al. (2005) describe that noise in fMRI data is often a combined form. Future studies could add different noise conditions to make the simulations more realistic and therefore make the results more generalizable to real fMRI data.

Another limitation is that all simulations were done with a TR of 1. Not all studies use a TR of 1 and therefore results of this study might be less generalizable to studies using a different TR.

Regarding the shapes of the HRF used in this study more variability could be introduced besides latency and a double-gamma or single-gamma HRF. As shown by Handwerker et al. (2004) the HRF can vary in many parameters and therefore more options remain to explore the impact on existing methods and may show more benefits of using the FHRFE.

Also, when comparisons were made, this study only compared the FHRFE with a basic BGLM in which the HRF was prespecified. As mentioned in the introduction, several other methods exist to improve the analyses either using the GLM or other methods to account for the variability of the HRF. These methods were not used in this study and therefore future research can compare the FHRFE to other approaches.

Regarding the FHRFE, improvements could be made. As shown in the results, signal detection based on X^2 value was preferred over MSE value because signal detection based on MSE value lacked in sensitivity. However, the specificity of signal detection based on X^2 value indicates that this method overfits the models and therefore had more false positives compared to the BGLM method with a difference in specificity of 7%. Therefore, if the sensitivity of signal detection based on MSE value can be improved, this method might have a better sensitivity-specificity ratio compared to the X^2 method. A possibility for the low sensitivity of the MSE method is that it uses less data to train the model compared to the X^2 method. An indication for this was that with more trials and thus more data, the MSE method performed better on sensitivity. Another reason that could explain the low sensitivity can be found in the evaluation of the models. When using MSE value for signal detection the FHRFE uses an absolute evaluation of the models. Meaning, when through 5 fold 10 times cross validation, 47 of 50 models obtained models have a higher MSE compared to the intercept-only model, 94% of the models indicate activation in the voxel time series. In this study, 94% was seen as too little proof of activation since the threshold was set on 95%. One way to improve sensitivity can be lowering this threshold. By lowering the threshold the method will reject the null hypotheses of no signal faster and therefore become more sensitive. However, the higher sensitivity could come at the cost of specificity. By lowering the threshold the method would probably give more false positives. Another way to improve sensitivity can be changing the final evaluation of the models from absolute (higher or lower MSE to continuous (MSE difference). Meaning that when using this method the mean difference

between the estimated models MSE and the intercept only MSE can be used for indication of activation. This way, bad models can be compensated by better models which might improve sensitivity.

Furthermore, the FHRFE still uses the GLM as a step to obtain p-values and models when using X^2 value to evaluate the models. When using MSE value to evaluate activation this step can be skipped. This can be done by directly fitting the averaged trials in the training data to the test data. However, to do this successfully, the sensitivity of the evaluation with MSE value does need to improve.

Also, since folds and times didn't seem to make a very big difference for the FHRFE this might be an indication that the computation time can be decreased by using even fewer folds and times. Future research could aim to study what the optimal number of models is for this method. However, the more folds and times of cross validations are used, the more estimations are made. This has the benefits that a distribution of the amplitude estimation by the models can be examined.

Another recommendation for future research is the comparison of the FHRFE with other methods that account for the variability of the HRF. Since other approaches exist to deal with the problem of between and within persons variability of the HRF it would be informative to know how the FHRFE compares to these methods on the same data.

Lastly, based on the current results, this study recommends using the FHRFE over using the BGLM. Since the HRF has shown to be variable on even more parameters than tested in this study, the validity of the analyses can be greatly improved by freely estimating the HRF.

Conclusion

This study aimed to explore a new method to account for the within and between subjects variability of the Hemodynamic Response Function in fMRI data. The HRF was freely estimated by averaging over trials. Because the trials were treated as separate time series, cross validation could be applied and multiple models could be fitted.

The Free HRF Estimation was compared to a basic approach where the HRF was prespecified in a GLM. In this study, the variability of the HRF was introduced by mis specifying the latency and HRF shape in the analyses. A simulation study showed that when the HRF in the model was mis specified, the FHRFE only suffered minimally in sensitivity and amplitude estimation and although the basic BGLM method could handle some misspecification, it increasingly worsened on sensitivity and amplitude estimation as the difference in latency increased. Regarding specificity, the basic BGLM slightly outperformed the FHRFE. However, the fact that the FHRFE is not affected by a variable HRF could outweigh the lower specificity.

Furthermore, several improvements to the FHRFE can still be made. The method of evaluating the models on MSE value can be altered in such way that it becomes more sensitive. Also, evaluating the models of the HRF can be done outside the framework of the BGLM. This can be done by avoiding fitting the averaged trials with a GLM and directly comparing these to the intercept only model on the test data. These adaptations can be tested in future research.

Although, this study showed promising results of the FHRFE, it would be too early to say if this method should be preferred over others. The FHRFE has not yet been tested on real data and has not yet been compared to other methods that freely estimate the HRF.

Lastly, it is indicated that treating the trials as a separate time series can work when estimating and evaluating the HRF in the data. This possibility can open up to many different approaches in the future. Also, Since the HRF has shown to be variable on even more parameters than tested in this study, the validity of the analyses can be greatly improved by freely estimating the HRF.

References

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

Allen, J. S., Damasio, H., & Grabowski, T. J. (2002). Normal neuroanatomical variation in the human brain: An MRI-volumetric study. *American Journal of Physical Anthropology*, *118*(4), 341–358.

Burock, M., & Dale, A. (2000). Estimation and detection of event-related fMRI signals with temporally Correlated Noise: A Statistically Efficient and Unbiased Approach. *Human Brain Mapping*, *260*, 249–260.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1924.

Friston, K. J., Holmes, A., Worsley, K., Poline, J.-P., Frith, C., & Frackowiak, R. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(081), 189–210.

Friston, K. J., Holmes, A., Poline, J.-B., Grasby, P., Williams, S., Frackowiak, R. S. J., Turner, R. (1995). Analysis of time-series revised. *NeuroImage 2,* 45–53.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, *24*(1), 244–252.

Genovese, C. R., & Genovese, C. R. (2000). A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data. *Journal of the American Statistical Association*, *95*(451), 691–703.

Gibbons, R. D., Lazar, N. A., Bhaumik, D. K., Sclove, S. L., Chen, H. Y., Thulborn, K. R., … Patterson, D. (2004). Estimation and classification of fMRI hemodynamic response patterns. *NeuroImage*, *22*(2), 804–814.

Goense, J., Bohraus, Y., & Logothetis, N. K. (2016). fMRI at High Spatial Resolution: Implications for BOLD-Models. *Frontiers in Computational Neuroscience*, *10*(June), 1–13.

Handwerker, D. A., Ollinger, J. M., & D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, *21*(4), 1639–1651.

Kapur, K., Roy, A., Bhaumik, D. K., Gibbons, R. D., Lazar, N. A., Sweeney, J. A., Aryal, S. & Patterson, D. (2009). Estimation and classification of BOLD responses over multiple trials. *Communications in Statistics - Theory and Methods*, *38*(16–17), 3099–3113.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., Kennedy, D. N., Hoppel, B. E., Cohen, M. S., Turner, R., Cheng, H., Brady, T. J., & Rosen, B. R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, *89*(12), 5675–5679.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W. L., & Nichols, T. E. (2005). Non-white noise in fMRI: Does modelling have an impact. *Neuroimage*, *29*(July 2005), 54–66.

Lazar, N. A. (2008). *The Statistical Analysis of Functional MRI Data*. (M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, & W. Wong, Eds.) (1st ed.). Springer.

Lee, A., Meyer, C., & Glover, G. (1995). Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neu-oimaging. *Magn Reson Med*, *33*, 745 – 754.

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD Signal. *Annual Review of Physiology*, *66*(1), 735–769.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–78.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W. L., & Nichols, T. E. (2005). Non-white noise in fMRI: Does modelling have an impact. *Neuroimage*, *29*(July 2005), 54–66.

Marrelec, G., Benali, H., Ciuciu, P., P´el´egrini-Issac, M., and Poline, J.-B. (2003). Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Human Brain Mapping, 19*, 1–17.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, *46*(1), 38–41.

Marijke Welvaert, Joke Durnez, Beatrijs Moerkerke, Geert Verdoolaege, Yves Rosseel (2011). *neuRosim: An R Package for Generating fMRI Data. Journal of Statistical Software*, 44(10), 1-18. URL http://www.jstatsoft.org/v44/i10/.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, *89*(13), 5951–5955.

Pernet, C. R. (2014). Misconceptions in the use of the General Linear Model applied to functional MRI: A tutorial for junior neuro-imagers. *Frontiers in Neuroscience*, *8*(8 JAN), 1–12.

R Core Team (2015). *R: A language and environment for statistical computing. R Foundation for Statistical* Computing, Vienna, Austria.URL http://www.R-project.org/.

Sahib, A. K., Mathiak, K., Erb, M., Elshahabi, A., Klamer, S., Scheffler, K., Ethofer, T. (2016). Effect of temporal resolution and serial autocorrelations in event-related functional MRI. *Magnetic Resonance in Medicine*, *76*(6), 1805–1813.

Appendices

Appendix A. Performance of FHRFE using MSE - value for signal detection

Figure 1. Sensitivity for different settings within the FHRFE using MSE-value for signal detection. Averaged over number of trials (20, 50), max intervals (8, 12, 40), noise conditions (White, AR2, AR6) and HRF shape (double-gamma, single-gamma). (n = 4800 time series)

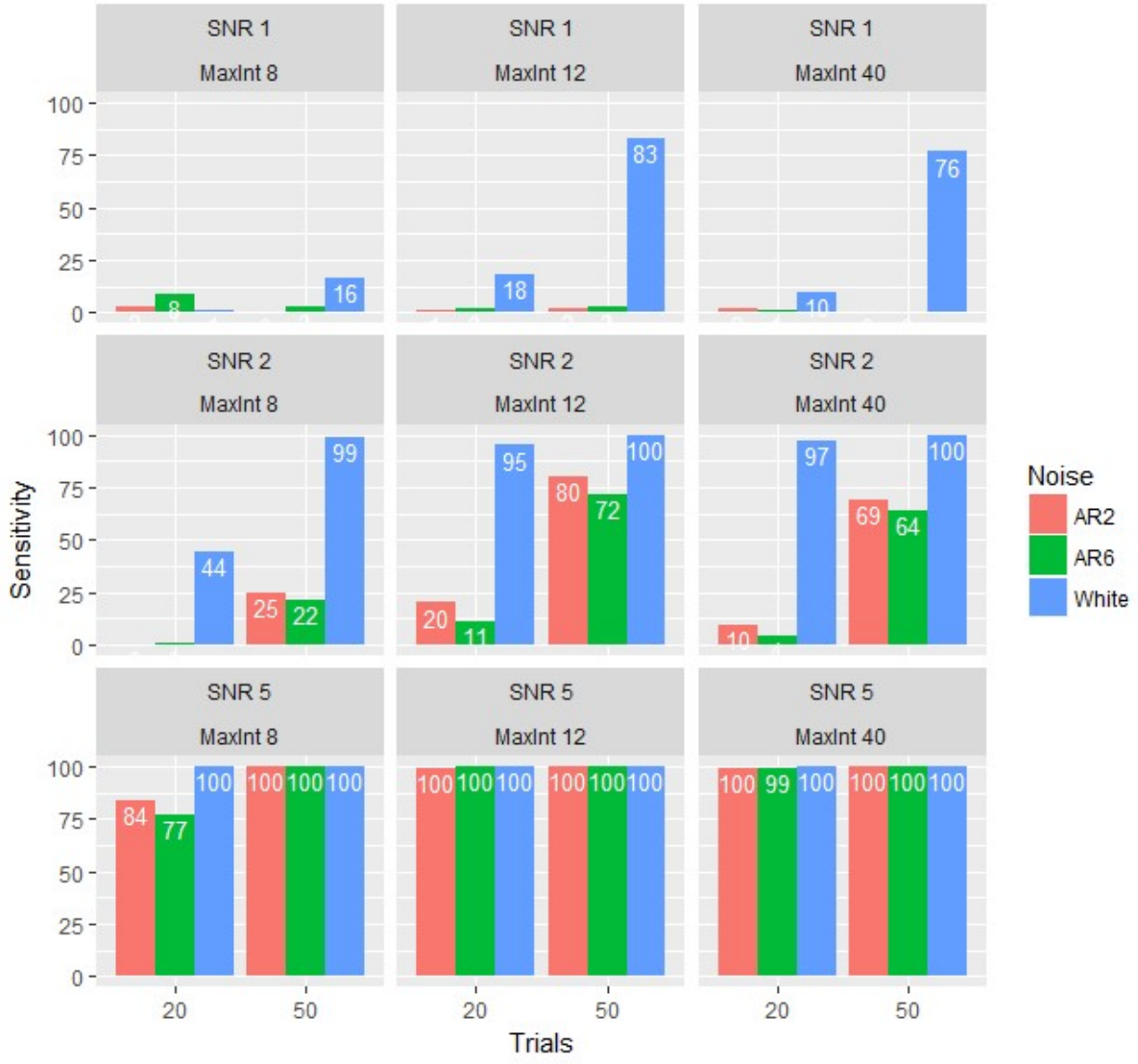


Figure 2. Sensitivity in different conditions for the FHRFE using MSE-value for signal detection. 5 fold 10 times with a training set of 70% used as settings for the FHRFE. Averaged over and HRF shape (double-gamma, single-gamma) condition. (n = 4800 time series)

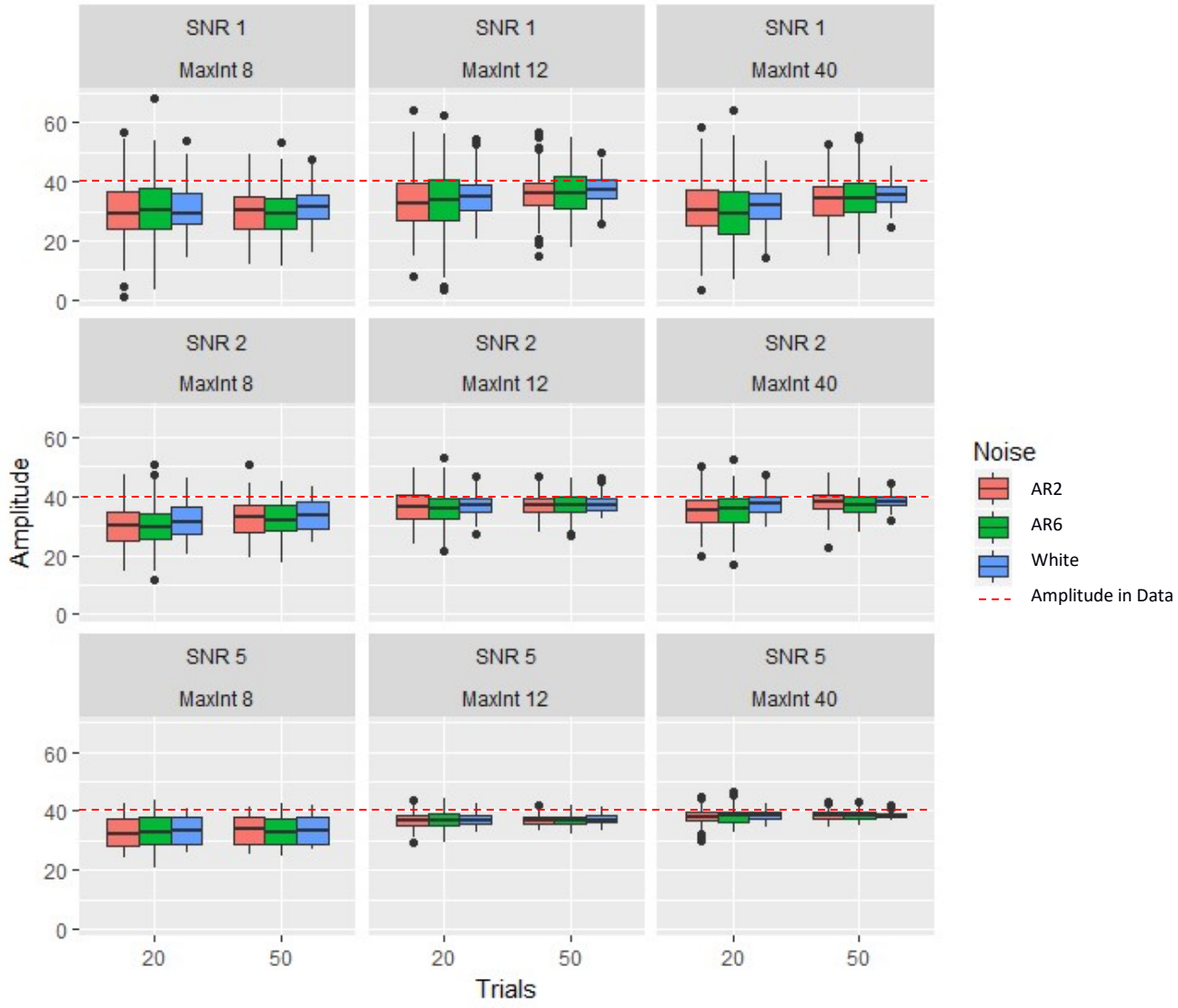


Figure 3. Amplitude estimation in different conditions for the FHRFE using MSE-value for signal detection. 5 fold 10 times with a training set of 70% used as settings for the FHRFE. Averaged over and HRF shape (double-gamma, single-gamma) condition. (n = 4800 time series)