Journal of Integrative Neuroscience, Vol. 9, No. 4 (2010) 429–452

DOI: 10.1142/S0219635210002494

Imperial College Press
www.icpress.co.uk

OPTIMAL HRF AND SMOOTHING PARAMETERS FOR FMRI TIME SERIES WITHIN AN AUTOREGRESSIVE MODELING FRAMEWORK

ANDREAS GALKA[*,†,‡], MICHAEL SINIATCHKIN[‡], ULRICH STEPHANI[‡], KRISTINA GROENING[‡], STEPHAN WOLFF[§], JORGE BOSCH-BAYARD[¶] and TOHRU OZAKI[‖]

[†]*Department of Neuropediatrics, University of Kiel*
24098 Kiel, Germany

[‡]*Institute of Experimental and Applied Physics*
University of Kiel, 24098 Kiel, Germany

[§]*Department of Neuroradiology, University of Kiel*
24098 Kiel, Germany

[¶]*Cuban Neuroscience Center*
Ave 25 No. 5202 esquina 158 Cubanacán
POB 6880, 6990, Ciudad Habana, Cuba

[‖]*Tohoku University, 28 Kawauchi*
Aoba-ku, Tokyo 980-8576, Japan
[*]*a.galka@neurologie.uni-kiel.de*

Received 1 August 2010
Accepted 17 September 2010

The analysis of time series obtained by functional magnetic resonance imaging (FMRI) may be approached by fitting predictive parametric models, such as nearest-neighbor autoregressive models with exogeneous input (NNARX). As a part of the modeling procedure, it is possible to apply instantaneous linear transformations to the data. Spatial smoothing, a common preprocessing step, may be interpreted as such a transformation. The autoregressive parameters may be constrained, such that they provide a response behavior that corresponds to the canonical haemodynamic response function (HRF). We present an algorithm for estimating the parameters of the linear transformations and of the HRF within a rigorous maximum-likelihood framework. Using this approach, both the optimal amount of spatial smoothing and the optimal HRF can be estimated simultaneously for a given FMRI data set. An example from a motor-task experiment is discussed. It is found that, for this data set, weak, but non-zero, spatial smoothing is optimal. Furthermore, it is demonstrated that activated regions can be estimated within the maximum-likelihood framework.

Keywords: FMRI; haemodynamic response function; autoregressive modeling; NNARX; spatial whitening; spatial smoothing; maximum-likelihood; model comparison; circular reasoning.

[*]Corresponding author.

1. Introduction

Functional magnetic resonance imaging (FMRI) provides temporally and spatially resolved information regarding the blood oxygen level dependent (BOLD) signals of neural tissue, thereby offering a valuable approach to the study of information processing in the human brain [1]. FMRI data consists of multivariate time series of very high dimension, as given by the number of voxels (typically $10^4 - 10^5$), and limited length, as given by the number of brain scans (typically a few hundred).

Analysis of FMRI time series by the rigorous methodology that has been developed in statistical time series analysis is challenging due to high data dimensions. The methods for FMRI analysis that have become a quasi-standard so far, such as voxel-wise regression models [2–4], have been designed mostly by heuristic procedures, in an attempt to reach a compromise between prior physiological knowledge, properties of the available data and computational time demands. With gradual progress in the development of both methodology and computational power, it has become possible to design refined algorithms that aim at extracting more dynamic information from the data, while imposing less heuristic assumptions and constraints.

Statistical analysis of data always has to address the issue of optimal merging of the information contained in the data with well-established information. In a perfect world, no prior information would be required, and all relevant information could be extracted, or reproduced, from the data. In practice, this is rarely practicable, and constraints have to be imposed, based on experience. The danger of these constraints lies with the risk of circular reasoning. The constraints and other details of the design of the analysis procedures design are modified until the previous expectations are "confirmed" by the results of the analysis. Recently, the danger of "circularity bias" in neuroscience has begun to attract increased attention [5].

In this paper, we extend earlier work on modeling FMRI time series by autoregressive models [6], thereby employing a class of predictive models which are well-established in statistical time series analysis. Particular emphasis is put on the estimation of model parameters by the maximum-likelihood method [7].

In most cases, a time-dependent stimulus is present during acquisition of the FMRI data. This is ensured either by asking the subject to perform a specific task, or by applying sensory input. This time-dependent stimulation is represented by a stimulus function, that can be used as a known external input of the predictive model fitted to the data. The stimulus function contains the timing, and possibly, the intensity of the applied stimulus. The analysis then aims at estimating brain areas that show the strongest activation to the stimulus. For this purpose, it is necessary to employ a model of the response function of the BOLD signal with respect to stimulation; again, this function can be imposed as prior knowledge, or estimated from the data. In this paper, we will demonstrate how such an estimation can be performed within a maximum-likelihood framework.

When estimating activated brain areas by voxel-wise regression models, spatial smoothing of the data has become a commonly applied preprocessing step. The

amount of spatial smoothing is usually selected according to established tradition. Spatial smoothing represents an instantaneous linear transformation that may be interpreted as playing a specific role within the autoregressive modeling framework. The same is also true for the Laplacian transformation, an instantaneous linear transformation that forms an important element within autoregressive modeling of FMRI data. We will demonstrate how both transformations can be simultaneously applied and optimized within the maximum-likelihood framework.

2. FMRI Modeling

2.1. *FMRI modeling by the general linear model*

Let $y(t, v)$ denote the FMRI data at voxel v, where $v = 1, \ldots, N_v$, and time point $t\,\Delta t$, where $t = 1, \ldots, N_t$; Δt denotes the sampling time of the data (usually denoted as "TR"). Let $s(t)$ denote the stimulus function at time point $t\,\Delta t$.

We assume that certain basic preprocessing of the raw data has taken place, such as motion correction. Time slice correction may be performed, although it is not essential. For standard analysis, spatial smoothing is commonly applied; we will discuss this step later in greater detail.

A well-established standard procedure consists of analyzing FMRI time series by fitting a *general linear model* (GLM) given by

$$y(t, v) = \theta(v) \sum_{\tau=0}^{\tau_{\max}} h(\tau\,\Delta t) s(t - \tau\,\Delta t) + \epsilon(t, v), \tag{1}$$

where $h(t\,\Delta t)$ denotes the stimulus response function, i.e., the haemodynamic response function (HRF), $\epsilon(t, v)$ denotes a voxel-dependent residual noise component, and $\theta(v)$ denotes a voxel-dependent least-squares-fit regression coefficient. Note that the regression model given by Eq. (1) may contain further terms, such as a constant regressor or motion-related regressors, which have been omitted for simplicity. The summation index τ denotes the time points of previous measurements, relative to the current time t and is expressed as multiples of the sampling time Δt. The least-squares-fit coefficients $\theta(v)$ can be estimated by standard least-squares regression techniques.

A typical choice for the HRF is given by a "gamma"-type function [8]

$$h(t\,\Delta t) = \left(\frac{t\,\Delta t}{\delta}\right)^{\gamma} \exp(-\lambda(t\,\Delta t - \delta)), \tag{2}$$

where $\delta = \gamma/\lambda$ represents the time delay of the peak of the stimulus response, or by a difference of two such functions [9],

$$h(t\,\Delta t) = \left(\frac{t\,\Delta t}{\delta_1}\right)^{\gamma_1} \exp(-\lambda_1(t\,\Delta t - \delta_1)) - k\,\left(\frac{t\,\Delta t}{\delta_2}\right)^{\gamma_2} \exp(-\lambda_2(t\,\Delta t - \delta_2)). \tag{3}$$

Typical values for the parameters are $\gamma_1 = 6$, $\gamma_2 = 16$, $\lambda_1 = \lambda_2 = 1$, $k = 1/6$; these values are used in the SPM8 software [2]. Note that for a difference of two

"gamma"-functions, the parameters δ_1 and δ_2 will not precisely describe the time delays of the peaks, as the superposition of these two functions will shift the peaks; in extreme cases the negative peak may be completely suppressed.

In many cases, the model given by Eq. (1) may be unable to account for all temporal correlations present in the data. Consequently the residual noise term $\epsilon(t, v)$ will have a non-white power spectrum. In order to describe and remove this residual correlation, further steps may be added to the modeling approach, such as whitening of $\epsilon(t, v)$ by an additional autoregressive model [10, 11].

2.2. *Multivariate AR/ARX models*

Linear autoregressive (AR) models represent a class of predictive parametric models which are widely applied to time series modeling [12]. In the previous section, we have briefly mentioned the possibility of using an AR model for whitening the residual noise term of voxel-wise fitting a GLM. As an alternative, we may, for a moment, ignore the standard HRF and try to directly describe all temporal correlations contained in a given FMRI time series by a suitable AR model.

If an AR model is formulated for the complete data vector of all voxels at a given point of time, $\mathbf{y}(t) = (y(t, 1), \ldots, y(t, N_v))^\dagger$, we would obtain a multivariate ARX model (where the "X" represents the exogeneous input term):

$$\mathbf{y}(t) = \sum_{\tau=1}^{p} \mathsf{A}(\tau)\mathbf{y}(t-\tau) + \sum_{\tau=1}^{q} \mathsf{B}_s(\tau)s(t-\tau) + \eta(t), \tag{4}$$

where p and q denote positive integer model orders, $\mathsf{A}(\tau)$ denotes a set $(N_v \times N_v)$-dimensional AR parameter matrices, $\mathsf{B}_s(\tau)$ denotes a set N_v-dimensional input gain parameter vectors, and $\eta(t)$ denotes a N_v-dimensional vector of driving noise, assumed to be white and distributed according to a multivariate Gaussian distribution with zero mean and $(N_v \times N_v)$-dimensional covariance matrix S_η.

Equation (4) describes a dynamical model which is driven by two input signals, $s(t)$ and $\eta(t)$. The model is an autoregressive model of order p with respect to the driving noise input $\eta(t)$, to be denoted as AR(p), and an autoregressive moving-average model of orders p, q with respect to the stimulus input $s(t)$, to be denoted as ARMA(p, q); for the latter model, a constraint $\mathsf{B}_s(0) = 0$ is required.

We remark that it would also be possible to generalize the model such that it would become an ARMA model with respect to the driving noise input, such that Eq. (4) would become

$$\mathbf{y}(t) = \sum_{\tau=1}^{p} \mathsf{A}(\tau)\mathbf{y}(t-\tau) + \sum_{\tau=1}^{q_s} \mathsf{B}_s(\tau)s(t-\tau) + \sum_{\tau=0}^{q_\eta} \mathsf{B}_\eta(\tau)\eta(t-\tau). \tag{5}$$

Here, a constraint $\mathsf{B}_\eta(0) = 1$ is required. While the model of Eq. (4) can be conveniently estimated by least-squares regression, this is impossible for Eq. (5), since the regressors $\eta(t-\tau)$ are unknown; the standard way to estimate moving-average models for unknown regressors employs state space modeling and Kalman filtering,

within some numerical parameter estimation procedure. We will discuss this point further in Sec. 5.3.

In the model of Eq. (4) the off-diagonal elements of the covariance matrix of the driving noise, S_η, describe instantaneous correlations between pairs of voxels, i.e., spatial correlations, while the off-diagonal elements of the AR parameter matrices $\mathsf{A}(\tau)$, $\tau = 1, \ldots, p$, describe delayed correlations between pairs of voxels, i.e., spatiotemporal correlations. The diagonal elements of S_η and $\mathsf{A}(\tau)$ describe instantaneous and delayed correlations, respectively, locally at each voxel. Note that, this model formally aims at explicitly modeling the interaction between every pair of voxels, while in the GLM there is no such possibility, rather the data at each voxel is modeled independently of the remaining voxels.

In reality due to the large number of voxels N_v, fitting the model of Eq. (4) to real FMRI data is impracticable, as the number of model parameters would exceed the number of available data values. It is for this reason that the NNARX model has been introduced. Further details of the NNARX model will be reviewed in the next section.

2.3. *Nearest-neighbor autoregressive modeling*

If the number of model parameters of the model of Eq. (4) is to be reduced, the huge parameter matrices $\mathsf{A}(\tau)$ and S_η should be made *sparse*. Sparseness may be enforced by an explicit model constraint; as an example, only pairs of voxels which are direct spatial neighbors within the grid of voxels, may interact via the matrices $\mathsf{A}(\tau)$, while all other elements of the $\mathsf{A}(\tau)$ are set to zero. It is reasonable to assume that the neighboring voxels will contain useful information for predicting the future activity of a given voxel. This model is known as the *Nearest-Neighbor Autoregressive model with exogeneous input* (NNARX) [6].

We remark that the sparseness of $\mathsf{A}(\tau)$ and S_η can also be enforced without imposing an explicit nearest-neighbor assumption, such that a subset of non-neighbor voxel-voxel interactions is retained instead. Suitable algorithms for this purpose have recently been developed by Valdés-Sosa and coworkers [13, 14]; these algorithms are based on incorporating sparseness constraints into the least-squares regression step, within a Bayesian framework. They offer the benefit of identifying important non-neighbor voxel-voxel interactions directly from the data. We will briefly return to this generalization in Sec. 8.

At the level of an individual voxel, the local model corresponding to the global NNARX model is given by

$$y(t,v) = \sum_{\tau=1}^{p_d} a(\tau,v,v)\, y(t-\tau,v) + \sum_{w\in\mathcal{N}(v)} \sum_{\tau=1}^{p_n} a(\tau,v,w)\, y(t-\tau,w) + \sum_{\tau=1}^{q} b_s(\tau,v)\, s(t-\tau) + \eta(t,v), \tag{6}$$

where $\mathscr{N}(v)$ denotes the set of labels of those voxels which are neighbors of voxel v. In Eq. (6) we allow for different AR model orders of the "diagonal" term (first term on the right-hand side) and of the neighborhood term (second term), denoted by p_d and p_n; for convenience, we shall stipulate that $p_d \geq p_n$. The set of AR parameters $a(\tau, v, w)$ represents the non-zero elements of the *sparse* global AR parameter matrices $\mathsf{A}(\tau)$.

With respect to the parameters $a(\tau, v, v)$, $a(\tau, v, w)$ and $b_s(\tau, v)$, the set of voxel-wise models of Eq. (6) can be fitted to a given FMRI time series very efficiently by the standard least-squares regression approach. However, the elements of the covariance matrix S_η pose a problem, which will be addressed in the next section.

3. Instantaneous Transformations

3.1. *Spatial whitening and Laplacian transformation*

By decomposing the high-dimensional multivariate AR model of Eq. (4) into a set of univariate models, as given by Eq. (6), it is implicitly assumed that S_η was diagonal. This may be an unrealistic assumption for FMRI time series data, given its low sampling rate; however, estimating all off-diagonal elements of S_η would massively increase the number of model parameters again and, in addition, render the decomposition into univariate local models and thereby the use of fast regression methods impossible.

As a suggested solution, an instantaneous linear transformation can be applied to the data prior to modeling [6, 15, 16]:

$$\tilde{\mathbf{y}}(t) = \mathsf{L}\mathbf{y}(t), \tag{7}$$

where L denotes a Laplacian matrix, i.e., a discretization of a second-order spatial derivative operator. For a given voxel set with neighborhood structure given by a set of neighbor label sets $\mathscr{N}(v)$, $v = 1, \ldots, N_v$, the Laplacian matrix is defined as [17]

$$\mathsf{L} = \mathsf{I}_{N_v} + c\mathsf{N}. \tag{8}$$

Here I_{N_v} denotes the N_v-dimensional identity matrix, and N denotes a $(N_v \times N_v)$-dimensional matrix having $\mathsf{N}_{vw} = 1$ if $w \in \mathscr{N}(v)$, $v \neq w$, and 0 otherwise. For the parameter c a value of $-1/6$ would be expected; alternatively c may be treated as a model parameter to be estimated by maximum-likelihood estimation. For $c = 0$ the case without the Laplacian transformation is retrieved.

Note that L is a sparse matrix; in fact, the non-zero elements of L occupy the same positions as the non-zero elements of the AR parameter matrices $\mathsf{A}(\tau)$.

The Laplacian transformation serves the purpose of removing instantaneous correlations between neighboring voxels, therefore it is known as *spatial whitening*. After the Laplacian matrix has been created, it is multiplied once with each data vector $\mathbf{y}(t)$ and the resulting transformed vectors $\tilde{\mathbf{y}}(t)$ are modeled by the NNARX model, as described above.

Spatial whitening approximately corresponds to assuming a particular non-diagonal shape for the driving noise covariance matrix [15, 16]

$$\mathsf{S}_\eta(\mathbf{y}) = \mathsf{L}^{-1}\mathsf{S}_\eta(\tilde{\mathbf{y}})(\mathsf{L}^{-1})^\dagger, \tag{9}$$

where $\mathsf{S}_\eta(\tilde{\mathbf{y}})$ denotes a diagonal covariance matrix which is introduced for the purpose of describing the covariance structure of the driving noise corresponding to the transformed data. We emphasize that diagonality of this covariance matrix is primarily not an assumption or approximation, but a model design decision. The diagonal elements of $\mathsf{S}_\eta(\tilde{\mathbf{y}})$ shall be denoted by $\sigma_\eta^2(\tilde{y}(v))$, $v = 1, \ldots, N_v$. A detailed discussion of the *spatial whitening* approach to modeling a non-diagonal covariance matrix can be found in [16].

A graphical visualization of the Laplacian transformation is shown in the left panel of Fig. 1.

3.2. *Spatial smoothing transformation*

As previously mentioned in Sec. 2.1, a spatial smoothing transformation is commonly applied before fitting the GLM. This transformation serves the purpose of improving the signal-to-noise ratio. Furthermore, it plays a role within the theory of Gaussian random fields, which is applied to the GLM. Spatial smoothing corresponds again to an instantaneous linear transformation, as given by Eq. (7), but with a smoothing matrix M instead of the Laplacian matrix L. If smoothing is performed by a Gaussian kernel, the elements of M are given by

$$\mathsf{M}_{vw} = \exp\left(-\frac{d(v,w)^2}{2\sigma_m^2}\right), \tag{10}$$

where $d(v, w)$ denotes the Euclidean distance between voxels v and w. This smoothing matrix depends on the variance parameter σ_m^2; for larger values we have strong smoothing, while for $\sigma_m^2 \to 0$ the case without the smoothing transformation is retrieved. As with L, all elements on the diagonal of M are ones.

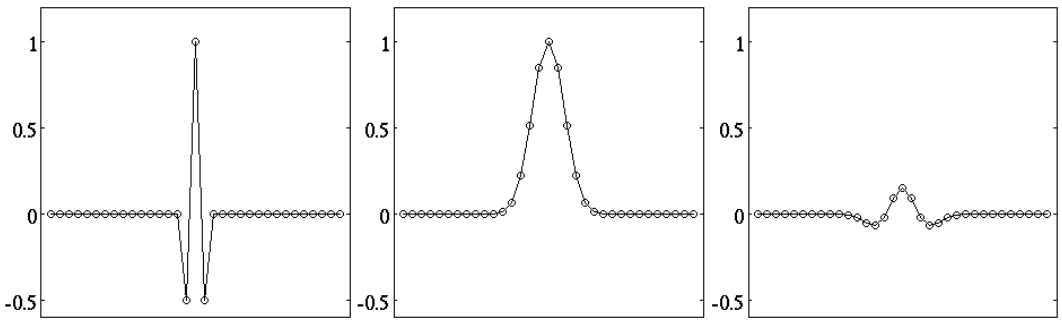

Fig. 1. Graphical visualization of the Laplacian transformation (left panel), the spatial smoothing transformation (center panel) and the joint spatial smoothing and Laplacian transformation (right panel) for the case of a 1-dimensional chain of voxels; parameters are $c = -0.5$ and $\sigma_m^2 = 3.0$.

According to Eq. (10), the smoothing matrix M is not a sparse matrix as all elements are positive. In practical work, it is convenient to define a threshold, such that elements smaller than this threshold are replaced by zero. In this paper we use a threshold of 10^{-4}; for suitably small values of σ_m^2 the smoothing matrix will then be sparse, although the number of its non-zero elements will typically be larger than that of the Laplacian matrix.

A graphical visualization of the spatial smoothing transformation is shown in the center panel of Fig. 1.

3.3. *Joint smoothing and Laplacian transformation*

Smoothing is commonly regarded as a preprocessing step, while the Laplacian transformation forms part of the data modeling procedure, as previously discussed; therefore, these two transformations may be regarded as mutually non-interfering steps that can peacefully coexist. This was the position of the earlier work on NNARX modeling [6], where the smoothing transformation was hardly mentioned.

On the other hand, it is obvious that these two transformations pursue contrarious aims: The Laplacian sharpens differences between neighboring voxels by performing a differentiation step, while the smoothing matrix represents an integration step by averaging over local neighborhoods within the data. Therefore it is possible to choose parameters c and σ_m^2 such that these two transformations almost cancel out. However, for somewhat larger variance, the Laplacian cannot cancel out the action of the smoothing matrix, since it is confined to the nearest neighbors, while the smoothing matrix reaches further out into space. In this case, the two transformations do not cancel one another, but rather superimpose and form a new transformation, which in the field of image processing is known as the "Laplacian-of-Gaussian" operator, or as the Hildreth-Marr operator.

Within the framework of NNARX modeling, we may interpret this combined transformation as a modified model for the non-diagonal driving noise covariance matrix S_η, which now contains two parameters, c and σ_m^2, instead of just one, c.

A graphical visualization of the joint smoothing and Laplacian transformation is shown in the right panel of Fig. 1. The figure demonstrates the superimposition of the properties of the Laplacian and smoothing transformations to form a joint transformation.

4. Describing the HRF within an Autoregressive Model

The standard HRF of Eq. (3) can be reproduced with an appropriate ARMA model. A specific advantage of ARMA models, as compared to AR models, is their ability to describe almost any impulse response behavior using a model with finite model orders p and q. The HRF represents the impulse response behavior of the BOLD signal with respect to the stimulus input $s(t)$. By applying a suitable estimation procedure to the HRF, such as the iterative method proposed by Steiglitz

and McBride [18], the corresponding sets of AR parameters $a(\tau)$, $\tau = 1, \ldots, p$, and of MA parameters $b_s(\tau)$, $\tau = 1, \ldots, p-1$, can be estimated.

The resulting model can then be employed at each brain voxel of a given voxel set; consequently, the local voxel-wise NNARX models will not differ as they all refer to the same global HRF. For this reason we denote the AR/MA parameters by $a(\tau)$ and $b_s(\tau)$ instead of $a(\tau, v, w)$ and $b_s(\tau, v)$. Nearest-neighbor AR parameters $a(\tau, v, w), v \neq w$, are set to zero for this model. The corresponding ARMA model provides a prediction for each voxel and time point, given the stimulus function and the previous data at this voxel. The time series of these predictions can be used as a regressor within a linear regression step; the result is a regression parameter at each voxel, which directly corresponds to the parameter $\theta(v)$ in the GLM, see Eq. (1). This particular voxel-wise AR model shall be denoted as HRF-ARX.

We will now demonstrate the practical modeling of the HRF using an ARMA model. We select the standard parameter values for Eq. (3); the resulting continuous function is shown in Fig. 2 (solid line). We assume a sampling time of 2.5 seconds and choose an AR model order of $p = 10$ and an MA model order of $q = 9$. Choosing model orders that fulfill $q = p-1$ offers advantages for the practical implementation. Application of the Steiglitz and McBride method yields the AR/MA parameters shown in the smaller inset of Fig. 2. Note the semilogarithmic plot and the different

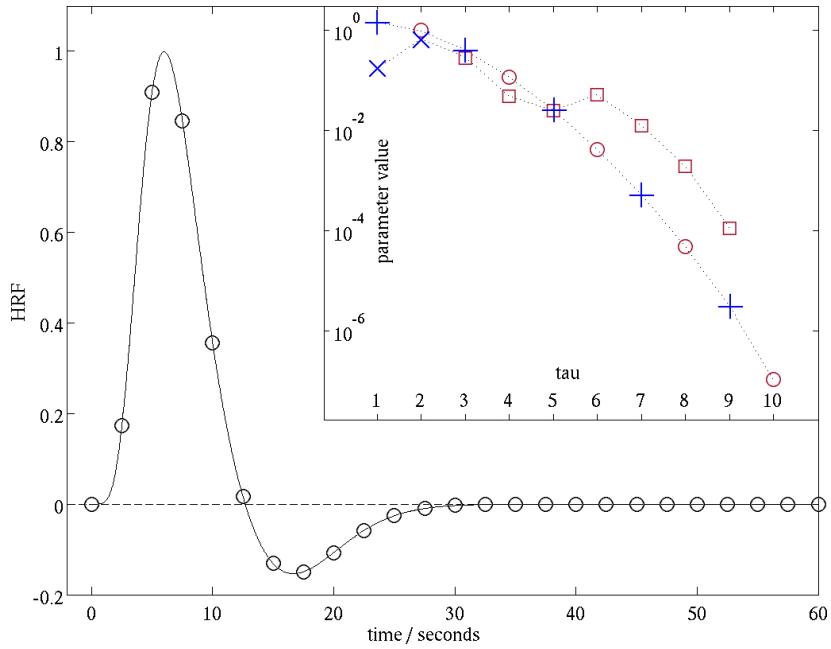

Fig. 2. Haemodynamic response function (HRF) for standard SPM8 parameters (solid line) and corresponding ARMA(10,9) impulse response function (circles); smaller picture: semilogarithmic plot of AR parameters ("+" and "○" symbols) and MA parameters ("×" and "□" symbols) of the ARMA(10,9) model; "+" and "×" symbols represent positive parameters, while "○" and "□" symbols represent negative parameters.

symbols used for positive and negative parameters. The ARMA model defined by these parameters corresponds to the discrete impulse response function shown in the main panel of the figure by circles. Note that the response function is only defined at these circles. From the figure, it can be observed that the discrete impulse response function of the ARMA(10,9) model fits the continuous HRF very well.

5. The Framework for Model Fitting and Comparison

5.1. *The log-likelihood of NNARX models with instantaneous linear transformations*

The models discussed thus far are characterized by several sets of parameters; values for these parameters may be chosen from prior knowledge, or preferably from a statistical criterion applied to the data. Here we choose the well-known criterion of maximizing the logarithmic likelihood, abbreviated as log-likelihood and denoted by $\log L$. Let $\nu_y(t, v)$ denote the data prediction errors of the model at voxel v and time t; these prediction errors are also known as *innovations* [19]. If a Gaussian distribution is assumed for the distribution of the innovations, then for NNARX (and also other models fitted by least-squares regression), the log-likelihood is given by

$$\log L(\mathbf{y}(1), \ldots, \mathbf{y}(N_t)) = -\frac{N_t}{2} \sum_{v=1}^{N_v} \left(\log \sigma_y^2(v) + 1 + \log 2\pi \right) + N_t \log |\mathsf{L}(c)| + N_t \log \left| \mathsf{M}(\sigma_m^2) \right|, \tag{11}$$

where $\sigma_y^2(v)$ denotes an estimate of the voxel-wise innovation variance, given by

$$\sigma_y^2(v) = \frac{1}{N_t - 1} \sum_{t=1}^{N_t} \nu_y^2(t, v). \tag{12}$$

Equation (11) also contains a correction with respect to the instantaneous linear transformations introduced above, represented by the Laplacian matrix $\mathsf{L}(c)$ and the smoothing matrix $\mathsf{M}(\sigma_m^2)$. It follows from the theory of probabilities that this correction is given by adding, at each time point, the logarithm of the determinant of each of these matrices. With this correction term, the log-likelihood explicitly depends on the parameters c and σ_m^2 (note that also the innovations $\nu_y(t, v)$ implicitly depend on these parameters).

Since both $\mathsf{L}(c)$ and $\mathsf{M}(\sigma_m^2)$ are large matrices — a typical dimension would be $35{,}000 \times 35{,}000$, corresponding to a 3 mm-grid of voxels — computing these log-determinants poses numerical challenges. If the Cholesky decomposition or at least the LU decomposition can be computed, they provide convenient solutions; if not, subpartitioning of the matrices should be applied, such that the sparse structure can be exploited. Furthermore, we have found that applying *approximate minimum degree ordering* [20] to these matrices considerably reduces computation time.

For practical optimization of the log-likelihood, we recommend replacing the computation of the full log-determinant by a parametrization of the functions which map the model parameters, i.e., c or σ_m^2, to the log-determinants. In the case of the Laplacian, this function turns out to be given by a power law:

$$\log |\mathsf{L}(c)| = \alpha\, c^{\beta}, \tag{13}$$

where α and β are parameters that depend on the given FMRI data set and the chosen set of active voxels. By computing $\log |\mathsf{L}(c)|$ for several values of c by the above-mentioned techniques, α and β can be easily estimated.

In the case of the smoothing matrix, we have not found an analytic form for the dependence of $\log |\mathsf{M}|$ on σ_m^2. Thus, we have resorted to employing interpolation and extrapolation by cubic splines to a set of explicitly computed values. Note that the thresholded smoothing matrix M could formally be described as the weighted sum of several "higher-order" Laplacian matrices, such that the first Laplacian matrix refers to the set of nearest neighbors, the second to the set of second-nearest neighbors, and so on. For each of these matrices, there would be a power law for the dependence of its log-determinant on the smoothing parameter. Unfortunately, there exists no simple relationship between the determinants of a set of matrices and the determinant of the sum of these matrices. Therefore we see little hope for finding an analytical expression for the case of the smoothing matrix.

5.2. *Overfitting control*

It is well known that the maximum-likelihood method itself will always favor larger models over smaller as it only evaluates the variance of the innovations. As a compromise between the accuracy of the predictions and the size of the model, information criteria can be introduced, such as the Akaike Information Criterion (AIC) [21, 22] which is defined by

$$\mathrm{AIC} = -2 \log L(\mathbf{y}(1), \ldots, \mathbf{y}(N_t)) + 2N_{\mathrm{par}} \frac{N_t}{N_t - N_{\mathrm{par}} - 1}, \tag{14}$$

where N_{par} denotes the total number of model parameters. The maximum-likelihood method is then replaced by a corresponding minimum-AIC method. AIC was derived as an unbiased variant of the log-likelihood (multiplied by -2) [21].

We remark that the definition given here for AIC differs from the commonly used definition by the fraction $N_t/(N_t - N_{\mathrm{par}} - 1)$ which was derived as a correction for the case of short time series [23]; it is omitted in most applications, but we have found that for the case of FMRI time series this correction is necessary, as will be demonstrated below.

For a complete FMRI time series, the value of AIC has a tendency to be quite large, therefore, in this paper we will always report the average voxel-wise AIC, i.e., AIC/N_v.

Note that the AIC, either in its corrected or in its standard form, represents only one particular example selected from numerous information criteria that have

been proposed for the purpose of overfitting control. A well-known alternative to AIC is the Bayesian Information Criterion (BIC), also known as Schwarz Information Criterion, which imposes a stronger penalty for model complexity than AIC. Another alternative is given by "free energy" [24] which has recently attracted considerable attention. In this paper we refrain from a detailed comparison of these criteria, but we should mention that certain conclusions with respect to the relative performance of given models may depend on the choice of the information criterion.

For the NNARX model discussed in Sec. 2.3, the number of model parameters is given by

$$N_{\mathrm{par}} = p_d N_v + \langle k \rangle p_n N_v + q N_v + N_v + 2, \tag{15}$$

where $\langle k \rangle$ denotes the average number of neighbors of a voxel; $\langle k \rangle$ will be smaller than 6. The five terms of this expression correspond to the following five groups of parameters:

$$a(\tau, v, v)$$
$$a(\tau, v, w), \quad v \neq w$$
$$b_s(\tau, v)$$
$$\sigma_\eta^2(\tilde{y}(v))$$
$$\{c, \sigma_m^2\}.$$

For the HRF-ARX model discussed in Sec. 4, the number of model parameters differs, even if the model orders p and q are the same as that in the NNARX model; this results from the fact that all AR/MA parameters are determined by the HRF. For this case, N_{par} is given by

$$N_{\mathrm{par}} = N_v + N_v + 2 + 5. \tag{16}$$

The four terms of this expression correspond to the following four groups of parameters:

$$\theta(v)$$
$$\sigma_\eta^2(\tilde{y}(v))$$
$$\{c, \sigma_m^2\}$$
$$\{\gamma_1, \gamma_2, \lambda_1, \lambda_2, k\}.$$

Note that the global parameters of the HRF have replaced the sets of local AR/MA parameters.

5.3. *Practical model fitting*

Most of the parameters of NNARX and HRF-ARX models, as listed in the previous subsection, can be fitted by voxel-wise standard least-squares regression. However, this is not possible for the *global* parameters, i.e., the Laplacian parameter c, the

smoothing parameter σ_m^2 and the HRF parameters $\gamma_1, \gamma_2, \lambda_1, \lambda_2, k$. On the contrary, estimates for the global parameters need to be given, before least-squares regression can be performed. For this reason, algorithms for numerical optimization need to be applied.

We have chosen to employ the BFGS quasi-Newton algorithm [25] and the Nelder-Mead simplex algorithm [26], both of which are very well established in the field of numerical optimization. For any given set of values for the global parameters, the corresponding instantaneous transformations are applied to the data; for the HRF-ARX model, the HRF also is computed and the corresponding ARMA parameters are estimated by the Steiglitz-McBride method. The voxel-wise standard least-squares regression step is then performed for the appropriately transformed data and recomputed regressors; from the resulting innovations, a log-likelihood and an AIC value can be computed.

The numerical optimization algorithms receive this AIC value and try to find estimates for the global parameters that minimize the AIC; they are unaware of the voxel-wise least-squares regression step which continues to recompute all local model parameters for each new set of the global parameters. For the HRF-ARX model, the Steiglitz-McBride iteration is repeated for each new set of the global HRF parameters, as part of the computation of a new AIC value. In this way, the number of parameters to be estimated by numerical optimization is kept to a minimum.

The numerical optimization consists of several steps, some of which aim at optimizing only a subset of the global parameters, e.g., only c, σ_m^2 or only $\gamma_1, \gamma_2, \lambda_1, \lambda_2, k$, while others aim at jointly optimizing the complete set. The BFGS quasi-Newton and Nelder-Mead simplex algorithms are used alternatingly; sometimes, the simplex algorithm succeeds in cases where the quasi-Newton algorithm fails, due to numerical problems.

In Sec. 2.2 we had briefly mentioned the option of introducing moving-average (MA) terms with respect to the driving noise term $\eta(t)$, compare Eq. (5). In contrast to AR and stimulus gain parameters, it is not possible to estimate MA parameters by least-squares regression; instead numerical optimization would be required, such as the method proposed by Melard [27]. Since these parameters would remain local parameters, their estimation cannot be merged with the estimation of the global parameters by numerical optimization; on the contrary, the estimation of the MA parameters would have to be repeated for all voxels, each time a new set of the global parameters is evaluated. For this reason, we expect that such a procedure would be very time-consuming.

In addition to sparseness constraints, Valdés-Sosa and coworkers [13, 14] have also applied smoothness constraints to the spatial maps of estimated AR parameters, again by incorporating them into the least-squares regression step, within a Bayesian framework. In this paper we have refrained from imposing such constraints, but we would like to mention its possibility.

6. Estimation of Activated Voxels

Within the framework of the GLM, t-test maps are usually employed for the estimation of the brain areas that show the strongest activation with respect to the stimulus. A similar approach can be designed, in the spirit of likelihood ratio testing (LRT) [28], within the maximum-likelihood, or minimum-AIC, approach. For this purpose, the voxel-wise contributions to the log-likelihood, or the AIC, are compared for two models, first the full model, including the stimulus as an external input term, second the same model, but without the stimulus input. In practical work, the second model can be fitted using the same routine as used for the first, with the exception of replacing the stimulus function with zeros, to ensure that it does not contribute to the predictions. NNARX and HRF-ARX model parameters are refitted by voxel-wise standard least-squares regression, while parameters of the instantaneous transformations and the HRF are kept at their optimal values within the full model.

The voxel-wise log-likelihood is essentially given by the expression $\log \sigma_y^2(v)$ in Eq. (11). The innovation variance $\sigma_y^2(v)$ will be smallest for the full model. If the stimulus input contributes nothing to improving the predictions, $\sigma_y^2(v)$ will have near similar values for both models. A useful LRT-type statistic is then given by

$$D(v) = N_t(\log \sigma_y^2(v; (\text{model without stimulus})) - \log \sigma_y^2(v; (\text{full model}))). \quad (17)$$

This statistic corresponds to a difference of AIC values, but without corrections with respect to the number of model parameters. In order to estimate the set of voxels with the strongest activation, a threshold may be directly defined, or indirectly by choosing a fraction of voxels to be selected; by the latter approach constant offsets of $D(v)$, due to the number of model parameters, have no effect.

Alternatively, it is also possible to directly compute t-test maps for the stimulus gain parameters $b(\tau, v)$, since these parameters are obtained by the same least-squares regression step as the GLM regression coefficients. It turns out that maps of activated voxels resulting from this approach are very similar to those resulting from the difference of AIC values.

7. A Practical Example: Motor-Task Data

In order to demonstrate the practical application of the models discussed above, we will now fit these models to a single FMRI time series.

7.1. *Experimental setup and preprocessing*

The chosen data set was recorded from a 38-year-old healthy male in the awake state, performing a block-design finger-tapping task; block length was 60 seconds task and 60 seconds rest.

FMRI was performed with a 3-Tesla MR scanner (Philips Achieva, Philips, Best, The Netherlands) and a standard 8-channel SENSE head coil. A single-shot T2*-weighted gradient-echo planar imaging sequence was used (sampling time $\Delta t = 2500\,\text{ms}$, TE $= 45\,\text{ms}$, 30 slices, 64×64 matrix, slice thickness $= 3.5\,\text{mm}$, FOV $= 200\,\text{mm}$, flip angle $= 90^{\text{o}}$). $N_t = 500$ brain volumes were acquired during the experiment.

Preprocessing of the FMRI time series was done by motion correction and by removing slow activity at each voxel by fitting a cosine basis set, using *SPM8* for both tasks; no time slice alignment was performed. In order to remove weak-signal voxels (assumed to lie outside of gray matter) from the analysis, a suitable threshold for the local variance was defined by inspection of the distribution of variances of the local time series of the original $64 \times 64 \times 30$ voxels. The number of voxels with variance above the chosen threshold was $N_v = 36{,}468$.

7.2. *Modeling of data*

We compare the following models and sets of models:

(a) NNARX$(p, 1)$ with $p = 1, \ldots, 20$;
non-optimized values for the parameters of the instantaneous transformations: $c = -1/6$, $\sigma_m^2 = 2.0$; smoothing may be omitted by choosing $\sigma_m^2 = 0.0$
(b) NNARX$(p, 1)$ with $p = 1, \ldots, 20$;
optimized values for the parameters c and σ_m^2 (see below)
(c) NNARX$(p, p-1)$ with $p = 1, \ldots, 20$;
non-optimized values $c = -1/6$, $\sigma_m^2 = 2.0$
(d) NNARX$(p, p-1)$ with $p = 1, \ldots, 20$;
optimized values for c and σ_m^2 (see below)
(e) HRF-ARX(10, 9);
non-optimized values $c = -1/6$, $\sigma_m^2 = 2.0$; smoothing may be omitted by choosing $\sigma_m^2 = 0.0$; HRF parameters are chosen at their SPM8 values, i.e., at non-optimized values
(f) HRF-ARX(10, 9);
optimized values for c and the HRF parameters (see below), but $\sigma_m^2 = 2.0$ is fixed
(g) HRF-ARX(10, 9);
optimized values for c, σ_m^2 and the HRF parameters (see below)

7.3. *Results of modeling: NNARX*

Results for the sets of models labeled (a), (b), (c) and (d) are displayed in Fig. 3: Values of AIC are shown versus model order p. We should mention that, due to the long sampling time of FMRI, we would expect the minimum of AIC to occur at rather small model orders. The figure shows both AIC values including the small-sample correction, as given by Eq. (14), and AIC values without this correction,

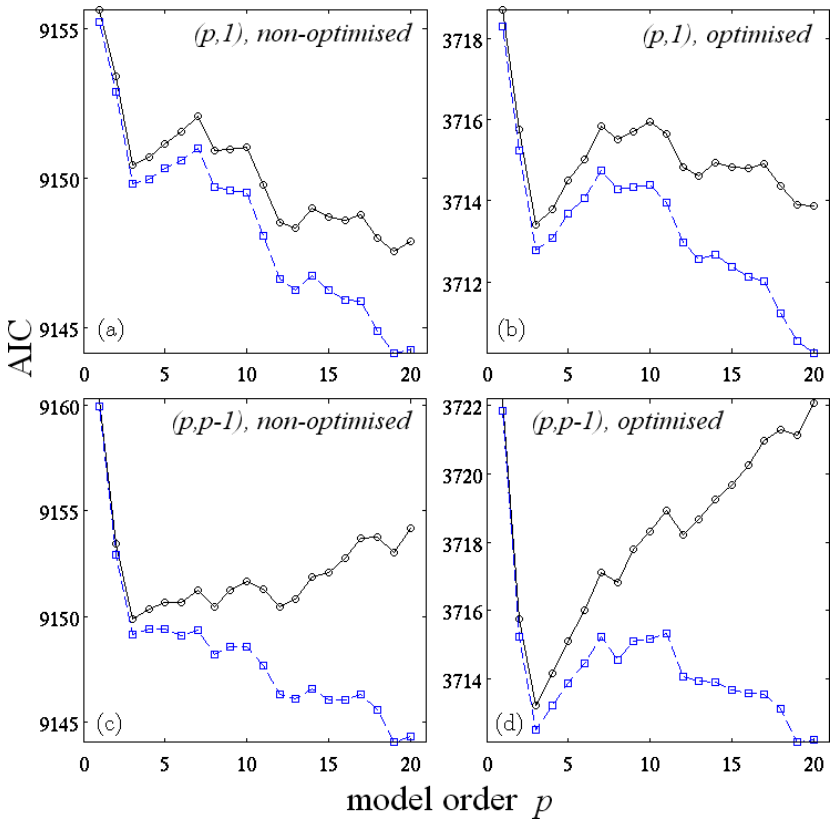

Fig. 3. Corrected AIC (solid lines and circles) and uncorrected AIC (dashed lines and squares) versus AR model order p for NNARX$(p,1)$ models (upper panels) and NNARX$(p,p-1)$ models (lower panels), using optimized (left panels) and non-optimized (right panels) values for the parameters of the instantaneous transformations c and σ_m^2. Labels (a)–(d) refer to the list of models in the text.

i.e., without the factor $N_t/(N_t - N_{\mathrm{par}} - 1)$. It can be seen that, for all four sets of models, the uncorrected AIC reaches the smallest values for the highest model orders, in contradiction to the expected behavior, while for sets (b), (c) and (d) the corrected AIC reaches its minimum at a model order of $p = 3$ and rises again for larger model orders. From this result, we conclude that the corrected AIC should be employed for this application, as it was also done in [6].

However, for set (a), i.e., the case of NNARX$(p,1)$ models with non-optimized values for the parameters of the instantaneous transformations, even the corrected AIC drops to smaller values for $p = 20$ than for $p = 3$. Also for the optimized case (b), the corrected AIC rises only weakly at $p > 3$, probably indicating the presence of considerable long-delay correlations within the data. Using a criterion with a stronger penalty for model complexity, such as BIC, would yield a minimum at $p = 3$ for all cases.

Optimized estimates for c and σ_m^2 show very weak dependence on model order p. Thus, results shown in Fig. 3 were obtained using values of c and σ_m^2 which were optimal for $p = 3$. The corresponding estimates for $p = 3$ and $p = 10$ are given

in Table 1, which provides a comparison of voxel-wise AIC results and optimized parameter estimates for NNARX and HRF-ARX models.

7.4. *Results of modeling: Instantaneous transformations*

In Table 1, results for the voxel-wise AIC are compared for the case of "strong" smoothing, with a variance parameter of $\sigma_m^2 = 2.0$, the case of optimal smoothing, according to the minimum-AIC criterion, and the case of no smoothing, $\sigma_m^2 = 0.0$. "Strong" smoothing corresponds to the amount of spatial smoothing that is commonly applied as part of preprocessing.

When comparing the AIC values, it is immediately evident that strong smoothing yields a much higher (and therefore worse) AIC than no smoothing or optimal smoothing. On the other hand, the improvement obtained by optimizing σ_m^2, and also c, is small, compared with the choice $c = -1/6, \sigma_m^2 = 0.0$. While the optimal value of the Laplacian parameter c is close to its initial value of $-1/6$, the optimal value of σ_m^2 is much smaller than 2.0, corresponding to only a weak amount of spatial smoothing. When visualizing the joint Laplacian and smoothing transformation in the same style as in Fig. 1, there is hardly any visible difference between the pure Laplacian transformation and the joint transformation.

This result can also be illustrated by substituting $\sigma_m^2 = 0.14$ and $d(v, w) = 1$ in Eq. (10). A value of $m(v, w) = 0.0281$ is obtained for the largest non-diagonal smoothing coefficient; obviously this coefficient will provide only very weak smoothing.

7.5. *Results of modeling: HRF-ARX*

Table 1 also shows results for four HRF-ARX models. In the case of these models, model orders p and q are determined by the ARMA representation that was

Table 1. Comparison of voxel-wise AIC and parameter estimates for several models; an asterisk denotes an optimized parameter estimate. Labels (a)–(g) in the first column refer to the list of models in the text.

Model	p	q	AIC/N_v	c	σ_m^2	γ_1	γ_2	λ_1	λ_2	k
(a) NNARX	3	1	3732.8	$-1/6$	0.0	—	—	—	—	—
(a) NNARX	10	1	3735.4	$-1/6$	0.0	—	—	—	—	—
(a) NNARX	3	1	9150.4	$-1/6$	2.0	—	—	—	—	—
(a) NNARX	10	1	9151.1	$-1/6$	2.0	—	—	—	—	—
(b) NNARX	3	1	3713.4	$-0.14061*$	$0.1425*$	—	—	—	—	—
(b) NNARX	10	1	3715.9	$-0.14025*$	$0.1421*$	—	—	—	—	—
(c) NNARX	10	9	3737.8	$-1/6$	0.0	—	—	—	—	—
(c) NNARX	10	9	9151.7	$-1/6$	2.0	—	—	—	—	—
(d) NNARX	10	9	3718.3	$-0.14007*$	$0.14184*$	—	—	—	—	—
(e) HRF-ARX	10	9	3777.6	$-1/6$	0.0	6.0	16.0	1.0	1.0	1/6
(e) HRF-ARX	10	9	9216.8	$-1/6$	2.0	6.0	16.0	1.0	1.0	1/6
(f) HRF-ARX	10	9	9047.7	$-0.17546*$	2.0	$10.337*$	$10.353*$	$1.626*$	$1.530*$	$0.354*$
(g) HRF-ARX	10	9	3736.3	$-0.13364*$	$0.12814*$	$9.019*$	$11.419*$	$1.332*$	$1.138*$	$0.318*$

chosen for the HRF, while the HRF parameters, as well as the parameters of the instantaneous transformations, may be optimized by minimum-AIC.

Table 1 demonstrates that also in HRF-ARX models strong smoothing yields higher AIC. The improvement of AIC by optimizing all parameters is small, compared to the effect of removing strong smoothing, but still clearly significant. A natural significant threshold of AIC is given by a value of two; here we find an average improvement of about 41 at each voxel.

The optimal values for c and σ_m^2 do not differ much from the corresponding NNARX values. The values of the HRF parameters may be compared with the initial/standard SPM8 values by plotting the corresponding HRFs themselves; this is done in Fig. 4.

From Fig. 4 it can be seen that the HRF corresponding to the optimized parameters differs from the standard SPM8 HRF by a somewhat weaker positive activation peak with a slightly larger time delay of 6.4 seconds, and by a negative activation peak with a considerably smaller time delay of 13.1 seconds. The corresponding parameters are $\delta_1 = 6.77$ and $\delta_2 = 10.03$; as mentioned in Sec. 2.1, the superimposition of two "gamma" functions causes shifts of the peaks.

The figure also shows the HRF corresponding to the constrained model with $\sigma_m^2 = 2.0$; for this HRF, the positive activation peak is still weaker than for the standard SPM8 HRF, and the negative activation peak is completely missing. Parameters are $\delta_1 = 6.36$ and $\delta_2 = 6.76$, but the negative activation peak is suppressed by the positive peak.

We remark that the chosen FMRI data set was recorded from a block-design experiment. It can be expected that the estimation of the optimal HRF shape would benefit from choosing an event-related design; this will be done in future work.

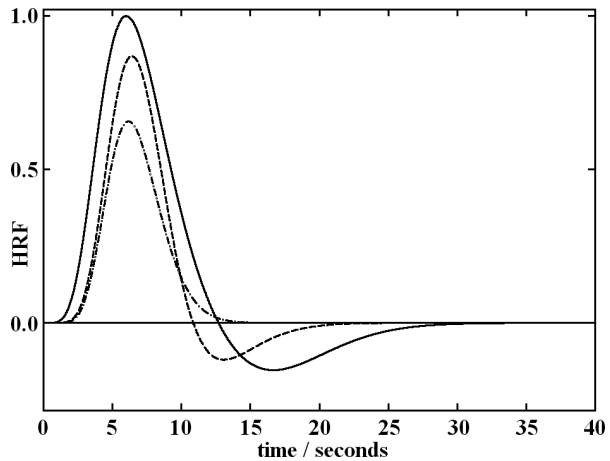

Fig. 4. Haemodynamic response function (HRF) for standard SPM8 parameters (solid line), for optimized parameters, if also the Laplacian parameter c and the smoothing parameter σ_m^2 were optimized (dashed line), and for optimized parameters, if c was also optimized, but σ_m^2 was fixed at a value of 2.0 (dash-dotted line).

7.6. *Results of modeling: Activated voxels*

In Fig. 5 we compare stimulus-related activations corresponding to the models discussed in this paper. We begin with the t-maps of a standard GLM analysis. The threshold is chosen such that the 50 voxels with the strongest activation are selected. Choosing a fixed number of activated voxels allows us to compare with the statistic based on AIC differences, as defined in Eq. (17).

The first row of panels of Fig. 5 shows the GLM results. The activated voxels form a clear cluster in the vicinity of the motor cortex of the left hemisphere; this is

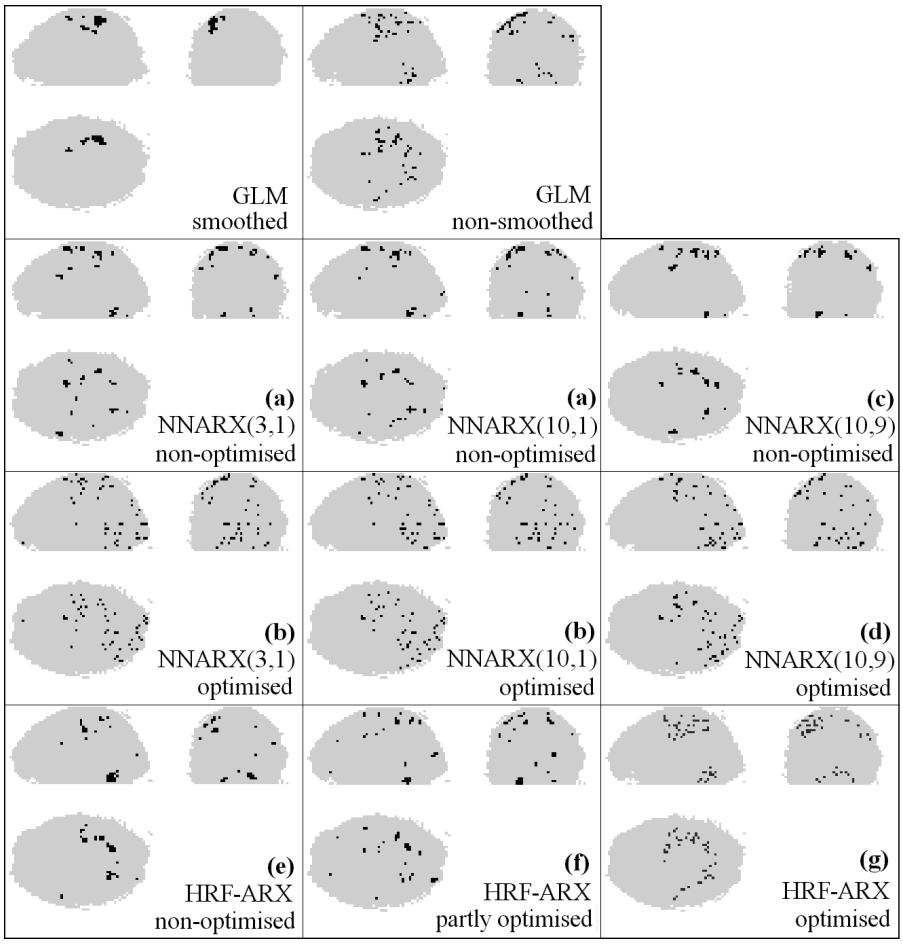

Fig. 5. Glass-brain visualization of the set of 50 voxels with strongest stimulus-related activation for motor-task FMRI data using the following models: GLM with smoothing ($\sigma_m^2 = 2.0$) and without smoothing ($\sigma_m^2 = 0.0$); NNARX with non-optimized standard values for c and σ_m^2, and model orders $p = 3, q = 1$, $p = 10, q = 1$ and $p = 10, q = 9$; NNARX with optimized values for c and σ_m^2, and the same model orders; and HRF-ARX with non-optimized standard values for c, σ_m^2 and the HRF parameters, with optimized values for c and the HRF parameters, but $\sigma_m^2 = 2.0$ (denoted by "partly optimized"), and with optimized values for all parameters. Model orders for HRF-ARX are always $p = 10, q = 9$. Labels (a)–(g) refer to the list of models in the text.

the expected result for a right-hand motor task. If spatial smoothing is omitted, the same analysis yields a cloud of mostly non-connected voxels, the majority of which is again located in the left motor cortex, while some activated voxels are visible in the vicinity of the right motor cortex and some in the cerebellum.

The second row of panels shows results for NNARX(3,1), NNARX(10,1) and NNARX(10,9) models with non-optimized standard values for c and σ_m^2, i.e., with strong smoothing. All models yield several activations in the cortex and the cerebellum. For the NNARX(10,1) and NNARX(10,9) models, the activations are focusing on the left motor cortex, while still other activations persist in the right hemisphere and the cerebellum.

The third row of panels shows results for the same NNARX models, but with optimized values for c and σ_m^2, i.e., with very weak smoothing (compare Table 1). In all three cases we see somewhat diffuse clouds of voxels. As can be seen from the sagittal projections, a subset of these voxels is located in the vicinity of the left motor cortex, especially for the NNARX(10,9) model. There is no corresponding activation in the right motor cortex.

The fourth row of panels shows results for HRF-ARX models, with non-optimized standard values for c, σ_m^2 and the HRF parameters (left), with optimized values for c and the HRF parameters, but $\sigma_m^2 = 2.0$ (center), and with all parameters optimized (right). For the non-optimized model, most activated voxels are situated in the vicinity of the left motor cortex and in the cerebellum. The results for the partly optimized model appear somewhat blurred, despite the strong smoothing. Finally, the fully optimized model corresponds to very weak smoothing, whence the activated voxels form a cloud of mostly non-connected voxels. Nevertheless, it is interesting to note, that this cloud seems to localize rather well around the left motor cortex and the cerebellum. Note the similarity with the result of the non-smoothed GLM.

8. Discussion and Conclusion

In this paper we have compared two classes of autoregressive models for modeling FMRI time series, and we have demonstrated how standard concepts from the field of FMRI modeling, namely a prespecified shape for the HRF and preprocessing by spatial smoothing, may be incorporated into this modeling approach, within a rigorous maximum-likelihood framework.

There are a number of similarities between standard GLM analysis and analysis by NNARX and HRF-ARX model fitting. Both approaches are based on voxel-wise least-squares regression, employ autoregressive modeling for describing temporal autocorrelations, and can be applied to spatially smoothed data. Furthermore both approaches provide local parameters for quantifying the effect of the stimulus on a given voxel, i.e., in principle any voxel is given a chance to respond strongly to the stimulus.

On the other hand, the Laplacian transformation does not have a corresponding element in the GLM, and both the amount of spatial smoothing and the shape of the

HRF need to be predefined for GLM analysis; for HRF-ARX and NNARX model fittings the parameters of the two instantaneous transformations can be optimized for the given data, and for HRF-ARX models the HRF can also be optimized. This optimization is based on the prediction of all available data, regardless of the fact that only specific areas will actually be activated by the stimulus; therefore it can be assumed that the parameters resulting from this optimization reflect fundamental behavior of "background" BOLD dynamics and much less specific aspects of stimulus response behavior. This aspect is also illustrated by our reinterpretation of the spatial smoothing parameter σ_m^2 as an additional parameter of the non-diagonal driving noise covariance matrix S_η. The LRT-type statistic given by Eq. (17) is designed in a way such that effects of the "background" BOLD dynamics approximately cancel out.

GLM analysis, however, has only very limited ability to characterize the "background" BOLD dynamics, namely through autoregressive modeling of the regression residuals. Nevertheless, it may benefit from optimized HRF-ARX models, e.g., by employing the optimal HRF parameters.

The concept of investigating "background" BOLD dynamics through predictive modeling, regardless of the presence or absence of stimuli, presents an interesting direction for further work; as examples we mention voxel-voxel connectivity and "resting state networks". It is sometimes conjectured that a large fraction of the potentially useful information contained in FMRI time series has not yet been accessed [29]; predictive modeling within the innovation approach [19] can be expected to play an important role in this respect, since it allows the decomposition of the spatiotemporal correlation structure of the data into different layers [16]. As an example, we expect that important non-neighbor voxel-voxel interactions can be extracted from the innovations more easily than from the original data.

The quality of NNARX and HRF-ARX models may be compared by, at least, two different criteria: by the minimum-AIC criterion (where the AIC could be replaced by some alternative information criterion) and by the maps of estimated activated voxels. We see from Table 1 that, according to the minimum-AIC criterion, the NNARX(3,1) model with optimized values for the parameters of the instantaneous transformations performs best among the set of all models which were tested in this paper. The optimized values correspond to the case of very weak smoothing; the value of the Laplacian transformation parameter is close to the standard value, $c = -1/6$. For this optimal model, the map of activated voxels consists of a cloud of non-connected voxels part of which fall into the left motor cortex, which is the expected region.

As has been demonstrated, the set of activated voxels may be better focused onto the expected region in at least three ways: by increasing the model order, by reverting to stronger spatial smoothing, or by imposing a particular standard shape of the HRF, taken from prior knowledge, onto the model (such that it becomes closer to the standard GLM analysis). However, all of these changes yield worse values of

AIC; this applies particularly to stronger spatial smoothing. This result illustrates the inherent problems with steps for data processing that are justified mainly by "tradition", not by a quantitative statistical criterion.

We would like to argue that spatially smoothing the data prior to any further analysis represents a dubious way to treat the data. By taking such a step, data that is obtained with very good spatial resolution, is artificially blurred, neglecting anatomical boundaries within the body and the high spatial resolution is sacrificed to noise reduction. It would be difficult to believe that such blurring would facilitate subsequent modeling, therefore we regard our result that strong smoothing is discouraged by the minimum-AIC criterion as natural.

The question then arises, how the clouds of activated voxels which result from very weak or absent smoothing, are to be interpreted? A few well-connected clusters of activated voxels, preferably close to the expected regions, seem more desirable than diffuse clouds. We are currently unable to decide whether these clouds indicate an improved spatial resolution, thereby possibly hinting at meaningful information, or result from nuisance effects, such as noise components of physiological or technical origin. It is likely that both possibilities play a role.

But we need to reiterate the problem of circular reasoning. The standard GLM analysis, including spatial smoothing, has gained credibility because it reproduces the expected activations in many data sets. While such expectations are generally justified, they cannot be easily extended to spatial details at high spatial resolution, except in favorable cases. Furthermore, results from GLM analysis have almost become "benchmark" results, even in cases where no certain independent information on the true locations of the activated areas is available.

Unfortunately, it is very difficult to generate reliable benchmark data sets for this situation, since it would require the existence of a method to gain certain knowledge of the true activations within that benchmark data. Simulations may offer a solution here, provided they succeed in imitating the physics of the data acquisition process very closely. It is important to understand that simplistic autoregressive simulations that could be implemented with limited effort, would not suffice for this purpose.

Expressing the same point in different words: How do we know that for a particular data set, the map with few well-connected clusters is closer to the unknown truth than the map with the widely dispersed cloud? It may be easier to interpret, but that alone does not justify acting against the minimum-AIC criterion.

It is obvious that well-established prior knowledge needs to be accessed for modeling, but based on the results presented in this paper, we doubt that strong spatial smoothing can be justified along these lines. We are more optimistic in the case of the HRF. For the HRF-ARX models, the same global HRF is imposed on all voxels, while for the NNARX models, each voxel chooses, purely data-driven, its own local HRF, independent of its neighbor voxels. Both of these are extreme cases, and future work should be directed towards developing compromises between the extremes.

Acknowledgments

The work of A. Galka was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through Sonderforschungsbereich SFB 855 and by the Japanese Society for the Promotion of Science (JSPS) through fellowship ID No. P 03059. The work of T. Ozaki was supported by the Japanese Society for the Promotion of Science (JSPS) through grants KIBAN B No. 173000922301 and WAKATE B No. 197002710002. The authors are grateful to Oliver Granert for drawing their attention to the Hildreth-Marr operator.

References

[1] Huettel SA, Song AW, McCarthy G, *Functional Magnetic Resonance Imaging*, Sinauer Assoc., Sunderland, 2004.
[2] Friston KJ, Jezzard P, Turner R, The analysis of functional MRI time series, *Hum Brain Mapp* **1**:153–171, 1994.
[3] Cox RW, AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages, *Comput Biomed Res* **29**:162–173, 1996.
[4] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM, Advances in functional and structural MR image analysis and implementation as FSL, *Neuroimage* **23**:208–219, 2004.
[5] Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI, Circular analysis in systems neuroscience: The dangers of double dipping, *Nat Neurosci* **12**:535–540, 2009.
[6] Riera JJ, Bosch-Bayard J, Yamashita O, Kawashima R, Sadato N, Okada T, Ozaki T, fMRI activation maps based on the NN-ARX model, *Neuroimage* **23**:680–697, 2004.
[7] Schweppe F, Evaluation of likelihood functions for gaussian signals, *IEEE Trans Inf Theory* **11**:61–70, 1965.
[8] Lange N, Zeger SL, Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging, *Appl Stat* **46**:1–29, 1997.
[9] Friston KJ, Fletcher P, Josephs O, Holmes AP, Rugg MD, Turner R, Event-related fMRI: Characterising differential responses, *Neuroimage* **7**:30–40, 1998.
[10] Bullmore E, Brammer M, Williams SCR, Rabe-Hesketh S, Janot N, David A, Mellers J, Howard R, Sham P, Statistical methods of estimation and inference for functional MR image analysis, *Magn Reson Med* **35**:261–277, 1996.
[11] Worsley KJ, Liao CH, Aston J, Petre V, Duncan GH, Morales F, Evans AC, A general statistical analysis for fMRI data, *Neuroimage* **15**:1–15, 2002.
[12] Box GEP, Jenkins GM, *Time Series Analysis, Forecasting and Control*, 2nd ed., Holden-Day, San Francisco, 1976.
[13] Valdés-Sosa PA, Spatio-temporal autoregressive models defined over brain manifolds, *Neuroinformatics* **2**:239–250, 2004.
[14] Valdés-Sosa PA, Bornot-Sánchez JM, Vega-Hernández M, Melie-García L, Lage-Castellanos A, Canales-Rodríguez E, Granger causality on spatial manifolds: Applications to neuroimaging, in Schelter B, Winterhalder M, Timmer J (eds.), *Handbook of*

Time Series Analysis: Recent Theoretical Developments and Applications, Wiley-VCH, Weinheim, pp. 461–491, 2006.

[15] Galka A, Yamashita O, Ozaki T, Biscay R, Valdés-Sosa PA, A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering, *Neuroimage* **23**:435–453, 2004.

[16] Galka A, Ozaki T, Bosch-Bayard J, Yamashita O, Whitening as a tool for estimating mutual information in spatiotemporal data sets, *J Stat Phys* **124**:1275–1315, 2006.

[17] Penny WD, Trujillo-Barreto NJ, Friston KJ, Bayesian fMRI time series analysis with spatial priors, *Neuroimage* **24**:350–362, 2005.

[18] Steiglitz K, McBride L, A technique for the identification of linear systems, *IEEE Trans Automat Contr* **10**:461–464, 1965.

[19] Kailath T, An innovations approach to least-squares estimation — Part I: Linear filtering in additive white noise, *IEEE Trans Automat Contr* **13**:646–655, 1968.

[20] Amestoy P, Davis TA, Duff IS, Algorithm 837: Amd, an approximate minimum degree ordering algorithm, *ACM Trans Math Softw* **30**:381–388, 2004.

[21] Sakamoto Y, Ishiguro M, Kitagawa G, *Akaike Information Criterion Statistics*, D. Reidel Publishing Comp, Dordrecht, 1986.

[22] Bernasconi C, König P, On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings, *Biol Cybern* **81**:199–210, 1999.

[23] Hurvich CM, Tsay C-L, Regression and time series model selection in small samples, *Biometrika* **76**:297–307, 1989.

[24] Friston KJ, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W, Variational free energy and the laplace approximation, *Neuroimage* **34**:220–234, 2007.

[25] Dyrholm M, Makeig S, Hansen LK, Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG, *Neural Computation* **19**:934–955, 2007.

[26] Baldick R, *Applied Optimization: Formulation and Algorithms for Engineering Systems*, Cambridge University Press, Cambridge, 2006.

[27] Melard G, A fast algorithm for the exact likelihood of autoregressive-moving average models, *J R Stat Soc Ser C Appl Stat* **33**:104–114, 1984.

[28] Buse A, The likelihood ratio, Wald, and Lagrange multiplier test: An expository note, *Am Stat* **36**:153–157, 1982.

[29] Haynes J-D, Rees G, Decoding mental states from brain activity in humans, *Nat Rev Neurosci* **7**:523–534, 2006.