NeuroImage xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

A semi-parametric nonlinear model for event-related fMRI

Tingting Zhang [a,*], Fan Li [b], Marlen Z. Gonzalez [c], Erin L. Maresh [c], James A. Coan [c]

[a] Department of Statistics, University of Virginia, Charlottesville, VA 22904, USA
[b] Department of Statistical Science, Duke University, Durham, NC 27708, USA
[c] Department of Psychology, University of Virginia, Charlottesville, VA 22904, USA

A R T I C L E I N F O

Article history:
Accepted 4 April 2014
Available online xxxx

Keywords:
fMRI
Hemodynamic response function
GLM
Multi-subject
Nonlinearity
Spline
Volterra series

A B S T R A C T

Nonlinearity in evoked hemodynamic responses often presents in event-related fMRI studies. Volterra series, a higher-order extension of linear convolution, has been used in the literature to construct a nonlinear characterization of hemodynamic responses. Estimation of the Volterra kernel coefficients in these models is usually challenging due to the large number of parameters. We propose a new semi-parametric model based on Volterra series for the hemodynamic responses that greatly reduces the number of parameters and enables "information borrowing" among subjects. This model assumes that in the same brain region and under the same stimulus, the hemodynamic responses across subjects share a common but unknown functional shape that can differ in magnitude, latency and degree of interaction. We develop a computationally-efficient strategy based on splines to estimate the model parameters, and a hypothesis test on nonlinearity. The proposed method is compared with several existing methods via extensive simulations, and is applied to a real event-related fMRI study.

Introduction

The existence of nonlinearities in evoked responses in blood oxygen level-dependent (BOLD) fMRI, particularly in event-related designs, has been widely recognized in the literature (e.g., Buxton et al., 1998; Friston et al., 1998b, 2000; Miller et al., 2001; Soltysik et al., 2004; Vazquez and Noll, 1998; Wager et al., 2005). The extent of nonlinearity usually varies across brain regions and stimuli, and shorter intervals between stimuli lead to stronger nonlinearity than longer ones (Buckner, 1998; Dale and Buckner, 1997; Liu and Gao, 2000; Vazquez and Noll, 1998). These nonlinearities are believed to arise from nonlinearities both in the vascular response and at the neuronal level, and are commonly expressed as interactions among stimuli. Though the importance of adjusting for nonlinear interactions in estimating hemodynamic responses has been demonstrated (a compelling example is given in Wager et al. (2005)), reliable quantification of nonlinearity is challenging in practice. Two main types of nonlinear models for fMRI have been developed: the dynamical Ballon model (Buxton and Frank, 1997; Buxton et al., 1998; Mandeville et al., 1999) and the Volterra series based models (Friston et al., 1998b, 2000), the connection between which is established in Friston et al. (2000). These models are flexible in accommodating various interaction effects, but their implementation is often hampered by model complexity. For instance, the Volterra series models generally involve a large number of free parameters, which pose difficulty in obtaining stable estimates due to over-fitting and loss of power given limited available data. This motivates us to propose a parsimonious semi-parametric Volterra series model that enables efficient presentation and estimation of nonlinearities in this article.

The Volterra series model is an extension from the general linear model (GLM; Friston et al., 1995; Worsley and Friston, 1995), where the observed BOLD time series for each voxel is modeled as the linear convolution between the stimulus function and the unknown hemodynamic response function (HRF). The GLM assumes linear time invariant system, and thus is not applicable in the presence of significant deviation from expected linear system behavior. The Volterra series, a series of infinite sum of multidimensional convolutional integrals, is essentially a higher-order extension of linear convolutions. For simplicity, second-order Volterra series are most commonly used for characterizing pairwise interactions between stimuli. Represented by two-dimensional spline bases in a fully nonparametric manner (Friston et al., 1998b), the second-order Volterra series is very flexible to accommodate a variety of nonlinear hemodynamic behaviors across different regions, stimuli and subjects. Moreover, under the spline representation, the extended GLM based on Volterra series is converted to a linear regression, the computation of which is straightforward. The ensuing parameter estimates, however, have large variances, especially when obtained from a single individual's data.

In Zhang et al. (2013), we proposed a semi-parametric HRF model within the GLM framework for multi-subject fMRI data. By assuming that for a fixed voxel and stimulus the HRFs share a common but unknown functional shape, and differ in magnitude and latency across subjects, this model allows for combining multi-subject data information for HRF estimation. Thus, the estimation efficiency can be

* Corresponding author at: Halsey Hall 111, University of Virginia, Charlottesville, VA 22904, USA. Fax: +1 434 924 3076.
E-mail address: tz3b@virginia.edu (T. Zhang).

http://dx.doi.org/10.1016/j.neuroimage.2014.04.017
1053-8119/© 2014 Published by Elsevier Inc.

significantly increased in contrast to analyzing each individual subject's data independently. We extend such "information borrowing" idea to the second-order Volterra series model. Specifically, in addition to using the semi-parametric HRF model, here we also assume that for a fixed voxel and a pair of stimuli, their associated second-order Volterra kernel has a common and unknown functional sphere, and differs in the extent of interaction across subjects. We develop a computationally-efficient strategy based on nonparametric spline expansions (De Boor, 2001; Eubank, 1988; Parker and Rice, 1985; Ruppert et al., 2003; Wahba, 1990) to estimate subject-specific and population-common characteristics. We also propose a hypothesis test on the sample average of second-order Volterra kernel estimates for assessing population interaction effect. Performance of the method is examined by both simulations and a real fMRI study.

Section Materials and methods presents the new method: Section Model introduces the semi-parametric model based on Volterra series; Section Spline-based estimation describes a new spline-basis-based regularized estimation strategy for estimating the model parameters and discusses the selection of functional basis and penalty parameter; and Section Hypothesis testing on nonlinearity develops a hypothesis test on nonlinearity. We then apply the proposed method to a real event-related fMRI study in Section Real data example and compare the method with several existing methods via simulations in Section Simulations. Section Discussion concludes.

Materials and methods

Model

We adopt the standard massive univariate approach; since the same approach applies to each voxel, the subscript for voxel is omitted here. For subject i ($i = 1,\cdots, n$), let $y_i(t)$ for $t = \delta,\cdots, T \cdot \delta$ be the observed fMRI time series of a given brain voxel, where δ is the experiment time unit when each fMRI scan is captured, usually ranging from 0.5 to 2 s. Also for subject i and stimulus k ($k = 1,\cdots, K$), let $v_{i,k}(t)$ be the known stimulus function which equals 1 if the kth stimulus evoked at $t(>0)$ in the experimental design for subject i, and 0 otherwise. The Volterra series is an extension of the Taylor series representation of the nonlinear system where the output of the nonlinear system depends on the past history of the input to the system. Friston et al. (1998b) proposed to use the second-order Volterra series to characterize nonlinearity in evoked hemodynamic responses as follows:

$$
\begin{aligned}
y_i(t) = \mathbf{d}_i(t)\cdot\beta_i + \sum_{k=1}^{K}\int_0^m h_{i,k}(u)\cdot v_{i,k}(t-u)du \\
+ \sum_{k_1,k_2=1}^{K}\int_0^m V_{i,k_1k_2}(u_1,u_2)\cdot v_{i,k_1}(t-u_1)\cdot v_{i,k_2}(t-u_2)du_1du_2 + \varepsilon_i(t),
\end{aligned}
\tag{1}
$$

where $\mathbf{d}_i(t)$ is a lower-order polynomial accounting for the low-frequency drift due to physiological noise or subject motion in the fMRI (Brosch et al., 2002; Luo and Puthusserypady, 2008; Smith et al., 1999); $h_{i,k}(t)$ is the hemodynamic response function (HRF) corresponding to the kth stimulus for subject i; $V_{i,k1k2}(t_1,t_2)$ is the 2nd-order Volterra kernel function that models the interaction between the hemodynamic responses under stimuli k_1 and k_2 for subject i; m is a fixed constant defining the domain of the HRF; and $\varepsilon_i(t)$ is the error term. Following a common practice in the literature, we adopt a 2nd-order polynomial for the drifting term $\mathrm{d}_i(t) = (1, t, t^2)$ with parameters $\beta_i = (\beta_{i,0}, \beta_{i,1}, \beta_{i,2})'$. Though it is possible to use higher order Volterra kernels, we focus on the second order for simplicity. The height, time to peak, and width of a HRF is commonly interpreted as magnitude, reaction time, and duration, respectively, of subjects' neuronal activity in response to stimuli. A typical HRF shape is shown in Fig. 4(a), having onset at the stimulus-evoked time, reaching peak between 5 and 8 s, and declining afterward to the baseline (zero). Model (1) without the term of the 2nd-order Volterra kernel is the GLM (Friston et al., 1995). There is a vast literature on the estimation of the HRF $h_{i,k}(t)$, including parametric methods (e.g., Friston et al., 1998a; Glover, 1999; Henson et al., 2002; Lindquist and Wager, 2007; Lindquist et al., 2009; Riera et al., 2004; Worsley and Friston, 1995) and nonparametric methods (e.g., Aguirre et al., 1998; Bai et al., 2009; Dale, 1999; Lange et al., 1999; Vakorin et al., 2007; Wang et al., 2011; Woolrich et al., 2004; Zarahn, 2002). Estimation of $V_{i,k_1k_2}(t_1,t_2)$ is more challenging than that of the HRF, because the Volterra kernel function, defined on the two-dimensional space, involves many more parameters, while the number of observations, T, for each subject is usually limited.

Model (1) can be viewed as a special case of linear functional models, with slope functions $h_{i,k}$ and interaction functions $V_{i,k1,k2}$. In the neuropsychological studies we consider, the underlying slope functions, the HRFs, vary across subjects in height, time to peak, and width. Therefore, the common practice of assuming identical parameter functions does not apply here. In fact, extracting subject-specific characteristics is often one of the main goals in multi-subject fMRI studies. To simultaneously model population-wide and subject-specific characteristics of brain activity, and to "borrow information" across subjects, we assume a semi-parametric form for both h and V:

$$
h_{i,k}(t) = A_{i,k}\cdot f_k\left(t + D_{i,k}\right), \tag{2}
$$

$$
V_{i,k_1k_2}(t_1,t_2) = M_{i,k_1k_2}\cdot V_{k_1k_2}(t_1,t_2), \tag{3}
$$

where $A_{i,k}$, $D_{i,k}$ and $M_{i,k1k2}$ are unknown fixed parameters, representing magnitude and latency of brain's reaction to the kth stimulus, and intensity of the interaction between the k_1th and k_2th stimuli, respectively, for subject i; $f_k(t)$ is the population average HRF corresponding to the kth stimulus, and V_{k1k2} is the population average interaction function between the k_1th and k_2th stimuli. Model (3) assumes that the interaction pattern between hemodynamic responses of a given pair of stimuli is identical, but differs in intensity across subjects. No parametric assumption except for differentiability is imposed on f_k and V_{k1k2}. By assuming that all the subjects have a common functional form of the HRFs and their interactions, Models (2) and (3) greatly reduce the number of parameters and also enable efficient information sharing across subjects. Note that Model (3) does not account for interaction effects on the onset and time to peak of hemodynamic responses, which are generally too complicated to be quantified for a two-dimensional function, whereas subject-specific interaction intensity is much easier to interpret. Model (2) was previously proposed in Zhang et al. (2013) in the context of GLM. When direct observations of $h_{i,k}(t)$ are available, Model (2) is referred to as "shift and magnitude registration" by Ramsay and Silverman (2005). A similar shape-invariant model for longitudinal data analysis has been also discussed in Lindstrom (1995). In GLM, however, one needs to address the additional challenge of deconvoluting $h_{i,k}(t)$ from the observed time series.

Spline-based estimation

We now develop a spline-basis-based regularized strategy to estimate the parameters in the proposed model. Assuming that the latency $D_{i,k}$ is smaller than the experimental time unit, we use a first-order Taylor expansion to approximate Model (2), converting $h_{i,k}(t)$ to a linear presentation in terms of subject-specific parameters $A_{i,k}$ and $D_{i,k}$:

$$
h_{i,k}(t) \approx A_{i,k}\cdot f_k(t) + C_{i,k}\cdot f_k^{(1)}(t), \tag{4}
$$

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017

where $C_{i,k} = A_{i,k} \cdot D_{i,k}$. Then we represent $f_k(t)$ by cubic B-spline bases: $f_k(t) = \sum_{l=1}^{L} a_{kl} \cdot b_l(t)$, where the basis functions $b_l(t)$ are chosen based on a partition $\Lambda_q = (t_0 = 0, t_1, \cdots, t_q = m)$ of the interval [0, m]. Selection of the knots Λ_q is discussed later. Given the boundary condition that $h_{i,k}(0) = h_{i,k}(m) = 0$, we let $a_{1k} = a_{Lk} = 0$.

Similarly, we represent the bivariate function $V_{k_1k_2}(t_1, t_2)$ by cubic spline bases:

$$V_{k_1k_2}(t_1, t_2) = \sum_{l_1l_2=1}^{L} Z_{k_1k_2l_1l_2} \cdot b_{l_1}(t_1) \cdot b_{l_2}(t_2).$$

It is known that nonlinearity disappears if events are spaced at least 5 s apart (Miezin et al., 2000), implying that $V_{k_1k_2}(t_1, t_2) = 0$ for $|t_1 - t_2| \geq 5$. Using this fact and cubic spline bases with equally-spaced knots, the number of free parameters can be reduced by letting $Z_{k_1k_2l_1l_2} = 0$ for $|l_1 - l_2| \geq 4 + 5/m \cdot (L-2)$. This fact also indicates that some $V_{k_1k_2}$'s, whose associated pairs of stimuli are always more than 5 s apart in the experiment, equal zeros in the model. Moreover, in many event-related experiments, pairs of stimuli are separated at certain values, implying that some values of $V_{k_1k_2}(t_1, t_2)$ are not observable. In this case, because the spline bases $b_l(t)$'s only cover a short period of the domain [0, m], some coefficients $Z_{k_1k_2l_1l_2}$ are not observable and should not be included in the model, which can further reduce the number of free parameters.

Letting $\mathcal{L}^2 = \{(l_1, l_2) : 1 \leq l_1, l_2 \leq L; |l_1 - l_2| \geq 4 + 5/m \cdot (L-2)\}$ and $\mathcal{K}^2 = \{(k_1, k_2) :$ there exists at least one $(u_1, u_2) \in (0, m)^2$ such that $v_{i,k_1}(t-u_1) = v_{i,k_2}(t-u_2) = 1$ for at least one subject i and $|u_1 - u_2| < 5\}$. The nonlinear functional Model (1) is transformed to the following bilinear model:

$$y_i(t) = \mathrm{d}_i(t) \cdot \beta_i + \sum_{k=1}^{K} \sum_{l=2}^{L-1} \omega_{i,kl} \cdot \rho_{i,kl}(t) + \sum_{k=1}^{K} \sum_{l=2}^{L-1} \phi_{i,kl} \cdot \varrho_{i,kl}(t) + \sum_{(k_1,k_2) \in \mathcal{K}^2} \sum_{(l_1,l_2) \in \mathcal{L}^2} \nu_{i,k_1k_2l_1l_2} \cdot \psi_{k_1k_2l_1l_2}(t) + \varepsilon_i(t), \qquad (5)$$

where $\omega_{i,kl} = A_{i,k} \cdot a_{kl}$, $\phi_{i,kl} = C_{i,k} \cdot a_{kl}$, $\nu_{i,k_1k_2l_1l_2} = M_{i,k_1k_2} \cdot Z_{k_1k_2l_1l_2}$, $\rho_{i,kl}(t) = \int_0^m b_l(u) \cdot v_{i,k}(t-u)du$, $\varrho_{i,kl}(t) = \int_0^m b_l(u) \cdot u_{i,k}(t-u)du$, and $\psi_{k_1k_2l_1l_2}(t) = \int_0^m \int_0^m b_{l_1}(u_1) \cdot b_{l_2}(u_2) \cdot v_{i,k_1}(t-u_1) \cdot v_{i,k_2}(t-u_2)du_1du_2$ are known functions. Here subject-specific parameters $A_{i,k}$, $C_{i,k}$, a_{kl}, M_{i,k_1k_2}, $Z_{k_1k_2l_1l_2}$ are not directly identifiable, but their products $\omega_{i,kl}$, $\phi_{i,kl}$ and $\nu_{i,k_1k_2l_1l_2}$ are unique. Therefore, the estimates of subject-specific HRFs and second-order Volterra kernels are still unique. Notations of the key parameters are listed in Table 1.

Table 1
Notations of key parameters.

Parameter	Description
$\mathbf{d}_i(t)$	A vector of known time-varying covariates
β_i	Coefficients of $\mathbf{d}_i(t)$
$A_{i,k}$	Subject-specific magnitude of the kth HRF
$D_{i,k}$	Subject-specific latency of the kth HRF
M_{i,k_1k_2}	Subject-specific degree of interaction between stimuli k_1 and k_2
a_{kl}	Coefficients of the spline bases representing the kth common function $f_k(t)$
$Z_{k_1k_2l_1l_2}$	Coefficients of the spline bases representing the 2nd-order Volterra kernel $V_{k_1k_2}(t_1, t_2)$
$C_{i,k}$	Product $A_{i,k} \cdot D_{i,k}$
$\omega_{i,kl}$	Product $A_{i,k} \cdot a_{kl}$
$\phi_{i,kl}$	Product $C_{i,k} \cdot a_{kl}$
$\nu_{i,k_1k_2l_1l_2}$	Product $M_{i,k_1k_2} \cdot Z_{k_1k_2l_1l_2}$
$\rho_{i,kl}(t)$	Known functions $\int_0^m b_l(u) \cdot v_{i,k}(t-u)du$
$\varrho_{i,kl}(t)$	Known functions $\int_0^m b_l^{(1)}(u) \cdot v_{i,k}(t-u)du$
$\psi_{i \cdot k_1k_2l_1l_2}(t)$	Known functions $\int_0^m \int_0^m b_{l_1}(u_1) \cdot b_{l_2}(u_2) \cdot v_{i,k_1}(t-u_1) \cdot v_{i,k_2}(t-u_2)du_1du_2$

A standard approach to estimating parameters in a bilinear model is through minimizing the mean squared error (MSE) of fMRI time series via the alternating least squares (ALS) algorithm, an iterative optimizing procedure. Iterative procedures often lead to slow convergence and volatile estimates, particularly in the cases with a large number of parameters and low signal-to-noise ratio. Therefore, below we propose a new noniterative estimation strategy based on regularization:

Step 1. If the latency $D_{i,k}$ is close to zero, parameters $\phi_{i,kl}$'s should be much smaller than $\omega_{i,kl}$'s and have little effect on estimating $h_{i,k}$. Given this, we first omit the term $\phi_{i,kl} \cdot \varrho_{i,kl}(t)$ involving the first-order derivative of f_k in Model (5) and obtain parameter estimates $\hat{\beta}_i$, $\hat{\omega}_{i,kl}$ and $\hat{\nu}_{i,k_1k_2l_1l_2}$ for each subject i, by minimizing the penalized MSE (PMSE) of $y_i(t)$,

$$\mathrm{PMSE}_i = \sum_{t=\delta}^{T \cdot \delta} \left[y_i(t) - \mathrm{d}_i(t) \cdot \beta_i - \sum_{k=1}^{K} \sum_{l=2}^{L-1} \omega_{i,kl} \cdot \rho_{i,kl}(t) - \sum_{k_1,k_2} \sum_{l_1,l_2} \nu_{i,k_1k_2l_1l_2} \cdot \psi_{i,k_1k_2l_1l_2}(t) \right]^2 + \lambda \left[\sum_k \int \left(\sum_l \omega_{i,kl} \cdot b_l^{(2)}(u) \right)^2 du + \sum_{k_1,k_2} \iint \left(\sum_{l_1,l_2} \nu_{i,k_1k_2l_1l_2} \cdot b_{l_1}^{(1)}(u_1) \cdot b_{l_2}^{(1)}(u_2) \right)^2 du_1 du_2 \right]. \qquad (6)$$

Step 2. Estimate $f_k(t)$ and $V_{k_1k_2}(t_1, t_2)$ respectively by $\hat{f}_k(t) = \sum_{i=1}^{n} \hat{h}_{i,k}(t)/n$ and $\hat{V}_{k_1k_2}(t_1, t_2) = \sum_{i=1}^{n} \hat{V}_{i,k_1k_2}(t_1, t_2)/n$, where $\hat{h}_{i,k}(t) = \sum_{l=2}^{L-1} \hat{\omega}_{i,kl} \cdot b_l(t)$ and $\hat{V}_{i,k_1k_2}(t_1, t_2) = \sum_{l_1l_2} \hat{\nu}_{i,k_1k_2l_1l_2} \cdot b_{l_1}(t_1) \cdot b_{l_2}(t_2)$.

Step 3. Given $\hat{a}_{kl} = \sum_{i=1}^{n} \hat{\omega}_{i,kl}/n$ and $\hat{Z}_{k_1k_2l_1l_2} = \sum_i \hat{\nu}_{i,k_1k_2l_1l_2}/n$ from Step 2, re-evaluate $A_{i,k}$, $C_{i,k}$ and M_{i,k_1k_2} through ordinary least square regression (OLS) of Model (5).

Step 1 is equivalent to estimating each subject's HRFs and the 2nd-order Volterra kernel in a fully nonparametric manner under spline-basis representations: $h_{i,k}(t) = \sum_{l=2}^{L-1} \omega_{i,kl} \cdot b_l(t)$, and $V_{i,k_1k_2}(t_1, t_2) = \sum_{l_1,l_2} \nu_{i,k_1k_2l_1l_2} \cdot b_{l_1}(t_1) \cdot b_{l_2}(t_2)$. The penalty in PMSE_i is used to regularize the roughness of the nonparametric estimates. The analytic minimizer of PMSE_i is essentially a Tikhonov-regularized regression estimator, because the MSE, the first term in Eq. (6), is quadratic of the parameters $(\beta_i, \omega_{i,kl}, \nu_{i,k_1k_2l_1l_2})$ and the penalty is quadratic of the parameters $\omega_{i,kl}$ and $\nu_{i,k_1k_2l_1l_2}$. We believe that the average of subjects' nonparametric HRF estimates can approximate the population mean HRF shape well in Step 2 for two reasons. First, the point-wise average of subjects' HRFs is close to the population mean HRF shape, if the underlying HRFs indeed follow the proposed semi-parametric model; second, empirically we found that though individual subject's nonparametric estimates may vary significantly in shape due to large data noise, the shape of their average is generally stable.

In the literature knot or basis selection it is typically performed with direct observations of a single target function (Zhou and Shen, 2001), whereas in our study we need to estimate multiple $h_{i,k}$'s and V_{i,k_1k_2}'s simultaneously without any direct observations. For simplicity, we use equally-spaced knots for both $h_{i,k}$ and V_{i,k_1k_2}, and select a set of bases from two choices—with knots separated by 1 and 1/2, respectively—by a ten-fold cross-validation (TFCV) procedure. Distinct from the standard approach, the TFCV here is carried out by dividing all subjects' fMRI data into ten time periods of equal length instead of ten sub-samples. Specifically, each time data in one period is removed, the model constructed based on the of rest of the data is used to predict the left-out data, and the overall prediction error summed up over ten periods is used as the criterion for knot selection.

As for penalty parameter selection, available methods include ordinary cross-validation (OCV), generalized cross-validation (GCV; Wahba, 1990), GCV for functional data analysis by Reiss and Ogden (2007, 2009), and restricted maximum likelihood (Wood, 2011), among many others. In our case, since penalty parameter selection confounds knot selection, the two are performed together by the modified TFCV above.

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017

Hypothesis testing on nonlinearity

To detect deviation from the linear time-invariant system, we propose an easy-to-implement test on estimated $\hat{V}_{k_1k_2}$ based on Hotelling's T-squared distribution. Under normality assumption of the error term or with long enough observation time T in Model (1), the estimates $\hat{\nu}_{i,k_1k_2} = \left(\hat{\nu}_{i,k_1k_2l_1l_2}, (l_1, l_2) \in \mathcal{L}^2\right)'$ from Step 1 for each subject i approximately follows a normal distribution N $(\nu_{i,k_1k_2}, \Delta_i)$, where the variance–covariance matrix Δ_i depends on convolutions $\rho_{i,kl}(t)$, $\varrho_{i,kl}(t)$ and $\psi_{i,k_1k_2l_1l_2}(t)$, and $\sigma_i^2 = \mathrm{var}(\varepsilon_i(t))$. Assuming that across population $\nu_{i,k_1k_2} \sim \mathrm{N}\left(\mu_{k_1k_2}, \Lambda\right)$, where $\mu_{k_1k_2}$ denotes the parameters for the population mean interaction function $V_{k_1k_2}$, then the population-wise $\hat{\nu}_{i,k_1k_2} \sim \mathrm{N}\left(\mu_{k_1k_2}, \Upsilon\right)$, where Υ is the variance and covariance matrix of $\hat{\nu}_{i,k_1k_2}$ across population. Then the test of nonlinearity is reduced to test whether $\mu_{k1k2} = 0$.

To test the mean of independent and identically distributed multivariate (p-dimensional) Gaussian random variables, $\mathrm{x}_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \Sigma)$, it is standard to use the Hotelling's T-squared statistic, defined by

$$(\overline{\mathrm{x}}-\mu)'\mathbf{W}^{-1}(\overline{\mathrm{x}}-\mu)\frac{n(n-p)}{(n-1)p}, \quad \text{with} \quad \mathbf{W} = \sum_{i=1}^{n}(\mathrm{x}_i-\overline{\mathrm{x}})'(\mathrm{x}_i-\overline{\mathrm{x}})/(n-1),$$

which follows an F distribution with degrees of freedom p and $n - p$. Based on this, we propose to test $H_0 : \mu_{k_1k_2} = 0$ vs. $H_A : \mu_{k_1k_2} \neq 0$ by using the statistic

$$\mathcal{T}^2 = \left(\sum_{i=1}^{n}\hat{\nu}_{i,k_1k_2}/n\right)'\hat{\Upsilon}^{-1}\left(\sum_{i=1}^{n}\hat{\nu}_{i,k_1k_2}/n\right),$$

where $\hat{\Upsilon}$ is the sample variance–covariance matrix of $\hat{\nu}_{i,k_1k_2}$. We reject the null hypothesis if $\mathcal{T}^2 > F_{p,n-p}^{1-\alpha}$, where $F_{p,n-p}^{1-\alpha}$ is the $1 - \alpha$ percentile of an F distribution with degrees of freedom p and $n - p$. In practice, with many functional bases used to represent the kernel function V_{k1k2}, however, p can be even larger than n, or comparable to n, leading close to singular $\hat{\Upsilon}$ and thus low detection power. To address this issue, we use only a subset of (l_1, l_2) in $\hat{\nu}_{i,k_1k_2l_1l_2}$ to significantly reduce p. Specifically, we perform a test on equally spaced elements of $\hat{\nu}_{i,k_1k_2l_1l_2}$, given that V_{k1k2} is smooth and $\nu_{i,k_1k_2l_1l_2}$'s corresponding to spatially-close regions usually have similar values. Simulations in Section Simulations shows that such a test has a high power with type I error preserved at the specified significance level.

Results

Real data example

Data

We analyze the fMRI data collected from the Monetary Incentive Delay (MID) Experiment, which measures subjects' brain activity related to reward and penalty processing (Knutson et al., 2000). In this experiment, 19 subjects (10 male, 9 female) between 22 and 25 years of age were recruited from a larger representative longitudinal community sample (Allen et al., 2007).

In the MID task, each participant completed a protocol comprised of 72 6-second trials involving either no monetary outcome (control/neutral task), a potential reward (reward task), or a potential penalty (penalty task). The fMRI scans were acquired at every 2 s (TR), leading to $T = 219$ frames of data for each subject. In each trial, participants were first shown a cue shape for 500 ms (anticipation condition), then waited a variable interval of between 2500 and 3500 ms, and were shown a white target square lasting between 160 and 260 ms (response condition). The cue shape (circle, square or triangle) shown at the start of each trial signals the type of the trial (reward, penalty or no incentive) to be implemented, and the white target shown at the end of each trial indicates button press from the participants, who were also told that their reaction times would affect the amount of money they receive in the monetary reward trial or lose in the penalty trial. In total, there were six stimuli involved in the experiment: three signal stimuli for the three types of monetary outcomes and three response stimuli to which the participants were required to respond. The six stimuli are henceforth referred to as neutral signal, reward signal, penalty signal, neutral response, reward response, and penalty response. The order of trials in the protocol for each participant was randomized, with 25% of them control trials, 37.5% reward trials, and 37.5% punishment trials. During the experiment, we used a Siemens 3.0 T MAGNETOM Trio high-speed magnetic imaging device at UVA's Fontaine Research Park to acquire fMRI data, with a CP transmit/receive head coil with integrated mirror. Two hundred twenty-four functional T2*-weighted Echo Planar images (EPIs) sensitive to BOLD contrast were collected per block, in volumes of 28 3.5-mm transversal echo-planar slices (1-mm slice gap) covering the whole brain (1-mm slice gap, TR = 2000 ms, TE = 40 ms, flip angle = 90°, FOV = 192 mm, matrix = 64 × 64, voxel size = 3 × 3 × 3.5 mm). More details of the experimental design, fMRI data acquisition and preprocessing can be found in Zhang et al. (2012).

Statistical analysis and discussion

We apply the proposed methods to four regions of interest (ROI): right putamen (2144 voxels), right amygdala (1587 voxels), right pallidum (1246 voxels), and right caudate (2504 voxels). These were determined structurally using the Harvard subcortical brain atlas, and were chosen for their likely involvement in affective neural processing based on previous studies (e.g., Knutson et al., 2000). For each voxel, we include in Model (1) six HRFs corresponding to the six stimuli. For each of the three tasks (neutral, reward and penalty), we use a 2nd-order Volterra kernel to characterize the interaction between the corresponding signal and response stimuli. Using the proposed noniterative estimation strategy, we evaluate the HRFs and their interactions. Statistical significance of the nonlinear term is tested using the Hotelling's T-squared test in Section Hypothesis testing on nonlinearity.

Fig. 1 displays the heat maps of P-values (P-values above 0.2 are not shown) of ROI voxels in testing interactions between signal and response stimuli. No significant interaction pattern is identified in right caudate and right amygdala, and thus the related results are omitted. There is almost no interaction between neutral signal and response stimuli across all the ROIs, which is intuitive, because neutral signal stimulus indicates that the following response is not required and does not affect any final gain. The most significant interaction is between monetary penalty signal and response stimuli, especially in the right putamen and right pallidum. Table 2 summarizes the percentages of voxels identified to be significant in the test of interaction between reward and penalty stimuli in these two regions at different significance thresholds. We used the empirical Bayes approach by Efron (2008) to evaluate the false discovery rates of the multiple hypothesis testing. An alternative approach is to use Benjamini–Hochberg (BH) threshold (Benjamini and Hochberg, 1995) to control for the false discovery rate (FDR) at different rates. Since the signal and response stimuli are not closely presented with inter-stimulus-interval (ISI) ranging from 2.5 s to 3.5 s, the interaction effect is not as intense as those with ISIs for no more than 1 s. In addition, the power Q8 of detecting nonlinearity is further diminished by the small sample size and large noise of fMRI data, and thus there are moderate FDRs in the multiple hypothesis tests of voxels. Nevertheless, a large proportion of voxels were still detected with significant interactions in the penalty task. In contrast, there is little interaction detected under the reward task. The reasons that interactions between negative signal and response stimuli are most prominent, and they are mainly in the right putamen and right pallidum are two-fold. First, the putamen and

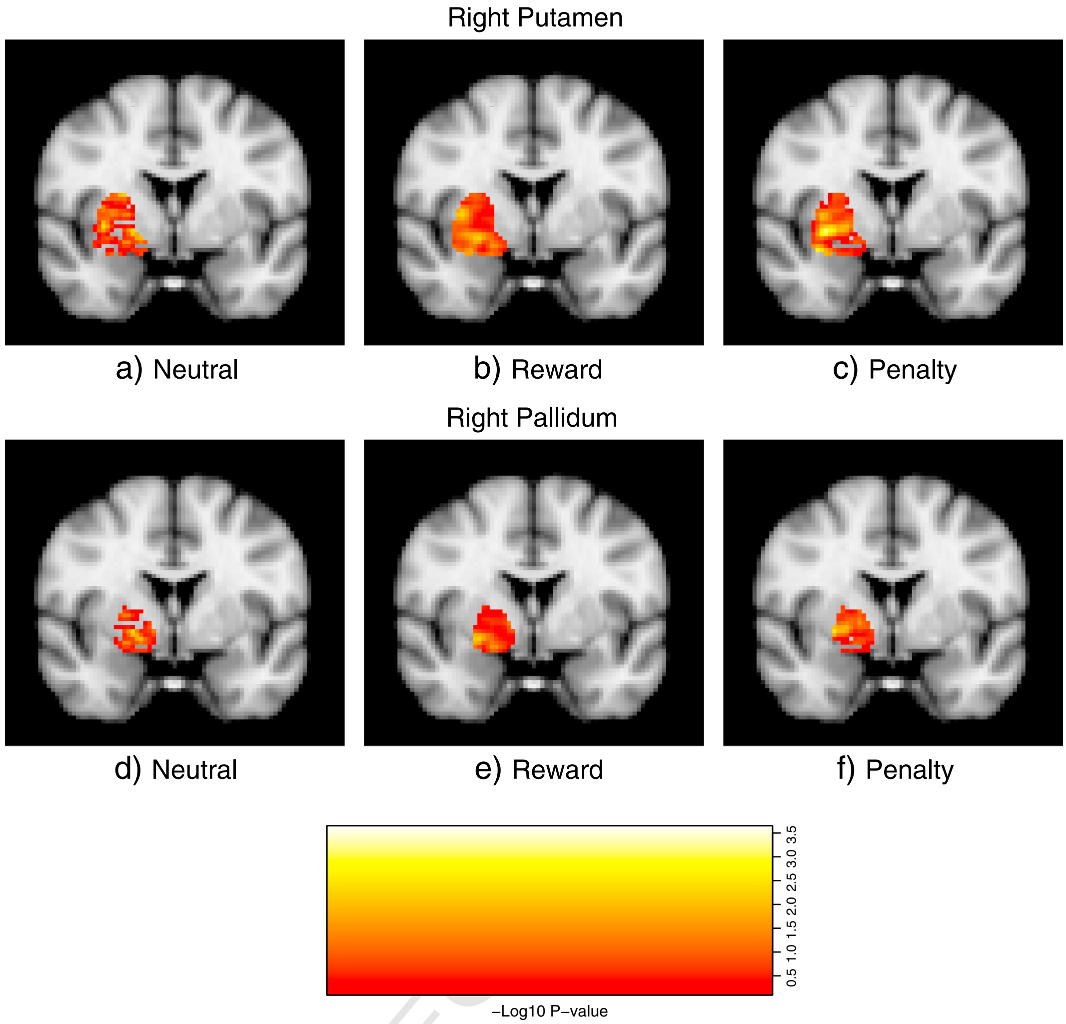

Fig. 1. Heat maps of P-values of voxels in ROIs. P-values of nonlinear tests of interactions respectively between neutral, monetary reward, and monetary penalty signal and response stimuli in right putamen and pallidum. The P-values are presented in −log10 scale.

pallidum are both regions of the basal ganglia, a subcortical network that is involved in, among other things, voluntary control of motor movements. Activation of these areas during signal presentation suggests preparatory motor activity in anticipation of the response cue. Second, such activation is more prominent in the penalty task which is not surprising, given the large body of work in psychology indicating that individuals react more strongly to negative stimuli than to positive stimuli (e.g., Baumeister et al., 2001). For example, brains are generally more active under negative stimuli (Cacioppo et al., 1997) and negative interactions more strongly define our attitudes about relationships (e.g., Gottman, 1994; Huston and Vangelisti, 1991).

To inspect the interaction effects, for each voxel with a P-value smaller than 5% in the right putamen and pallidum, we calculate the averaged 2nd-order Volterra kernel estimates across time and subjects, $\sum_i \sum_{t_1} \sum_{t_2} \hat{V}_{i,k_1k_2}(t_1,t_2)/(n \cdot m^2)$, histograms of which are presented in Figs. 2(a) and (c). To give a more explicit view of the detected nonlinearity, Figs. 2(b) and (d) respectively shows the estimated population mean $\hat{V}_{k_1k_2}(t_1,t_2)$ of the voxel with the most significant nonlinear behavior in the two regions. The color scale is arbitrary; light yellow is positive, and dark red is negative. Since intervals between consecutive stimuli in this experimental design are between 2 and 4 s, nonzero values of $V_{k_1k_2}(t_1,t_2)$ only appear in the off-diagonal band where $|t_1 - t_2|$ is between 2 and 4 s, and the values at other points are not observable. The interactive effect of penalty tasks, especially in the right putamen, tends to be negative. One possible explanation is that the signal stimulus prepares the subjects for the response, leading to less intensive reactivity when response stimulus is presented. Such a negative interaction effect was also reported in Friston et al. (1998b). In terms of data analysis, the magnitude of the HRF would be underestimated if significant nonlinearity in the underlying hemodynamic responses exists but is not taken into account in the estimation.

Fig. 3 displays the estimated population mean HRF f_k (dark line) and individual HRF $h_{i,k}$ (broken lines) of several randomly selected subjects for the voxel in the right putamen that has the most significant interaction of the penalty task. The effect of "borrowing" information across subjects can be clearly seen here as $\hat{f}_k$ is much less variant than the $\hat{h}_{i,k}$'s, though they share a similar shape, for each of the six stimuli. In general, the response stimuli evoked stronger and stabler activity across subjects than the signal stimuli, since subjects' response affected the

Table 2
The percentages and associated false discovery rates (FDR, in parentheses) of voxels identified in the ROIs by the test on nonlinearity at different significance levels.

Significance level	Right putamen		Right pallidum	
(FDR) (%)	Reward	Penalty	Reward	Penalty
5%	7.4	20.7	5.6	13.8
FDR	(67.7)	(24.1)	(89.0)	(36.2)
10%	18.1	33.4	12.4	23.5
FDR	(55.3)	(30.0)	(80.4)	(42.5)

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017

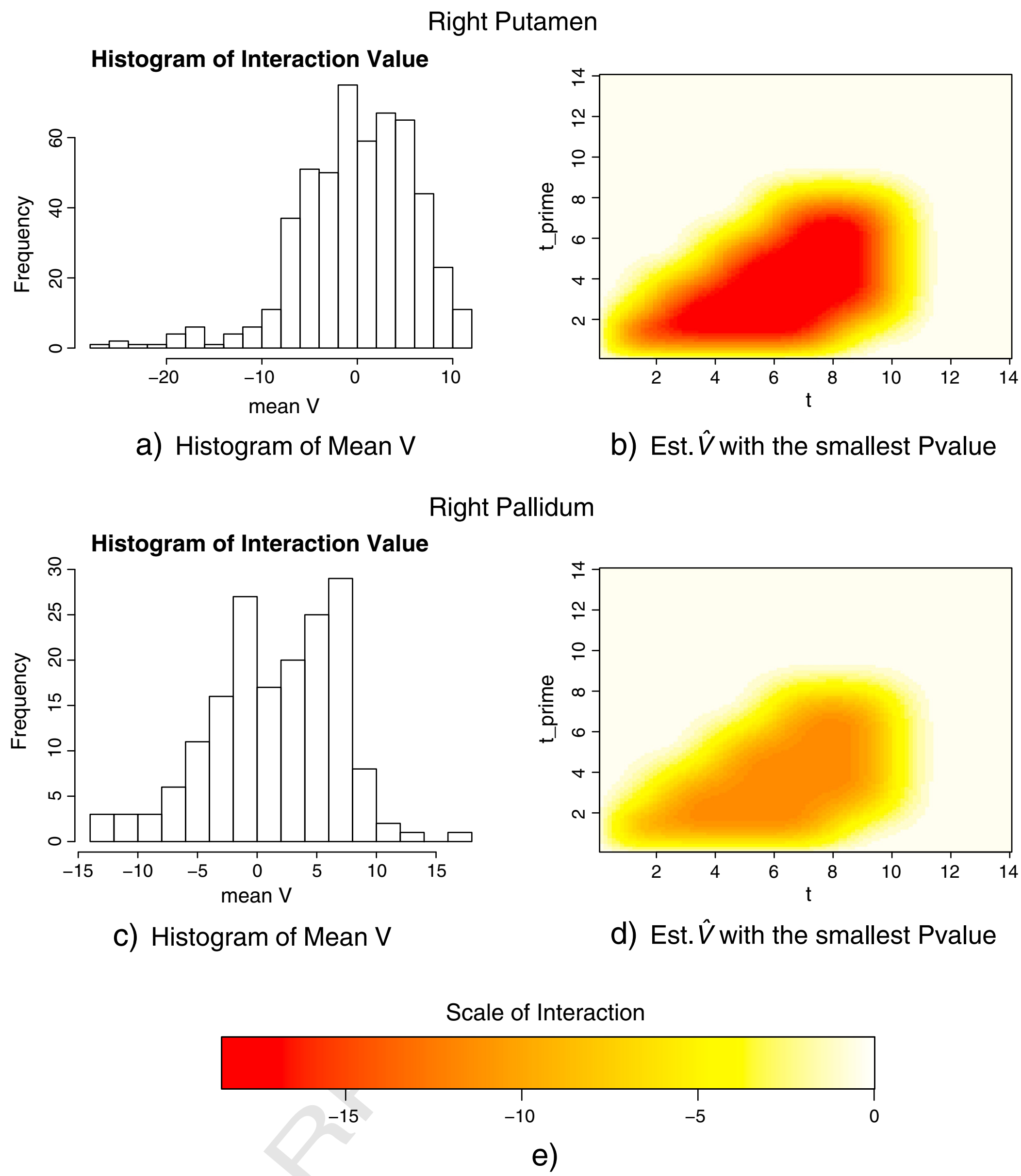

Fig. 2. Histograms of $\sum_i\sum_{t_1}\sum_{t_2}\hat{V}_{i,k_1k_2}(t_1,t_2)/(n\cdot m^2)$ for modeling interactions between penalty signal and response stimuli of all voxels with a P-value smaller than 5% in right putamen (a) and right pallidum (c). Estimated population average interaction function $\sum_i\hat{V}_{i,k_1k_2}(t_1,t_2)/n$ between penalty signal and response stimuli of the voxel in right putamen (b) and right pallidum (d) with the smallest P-value.

ensuing monetary gain or losses. The mental activity caused by the signal stimulus has a large variation across subjects. Such findings are in keeping with previous work indicating that passive viewing or "resting" generally produces noisier data than those that require a response from subjects. One model suggests that this "noise" may be a product of interactions between individual differences in cognitive and affective styles with uncontrolled portions of the experiment (Coan et al., 2006). So while the response cue elicits the same motor response from everyone (and thus a more coherent neural response), passive cue viewing may elicit similar, but relatively less coherent mental actions.

Simulations

Simulation design

We conduct simulations to further examine the properties of the proposed semi-parametric model in HRF estimation and also to compare with four existing methods: the linear semi-parametric model for HRF without the 2nd-order Volterra kernels proposed by Zhang et al. (2013), referred to as the linear spline-based method; a parametric approach representing HRF by a linear combination of canonical HRF and its first derivative, called canonical method hereafter; nonparametric Tikhonov-regularized estimate with penalty parameter selected by generalized cross validation (Tik-GCV, Casanova et al., 2008); and nonparametric smooth finite impulse response (SFIR) method (Goutte et al., 2000).

We generate time series data using the experimental design identical to that in the MID experiment with six stimuli for $n = 19$ subjects and three interaction effects. The HRFs $h_{i,k}(t)$ follow Model (2) with the population mean HRF f_k being a mixture of two gamma functions that have the same mathematical expression as the canonical HRF (Worsley and Friston, 1995):

$$f_k(t) = b_{1,k}^{a_{1,k}}\frac{t^{a_{1,k}-1}\cdot e^{-b_{1,k}t}}{\Gamma\left(a_{1,k}\right)} - c_k\cdot b_{2,k}^{a_{2,k}}\frac{t^{a_{2,k}-1}\cdot e^{-b_{2,k}t}}{\Gamma\left(a_{2,k}\right)},\quad k = 1,\ldots,6. \tag{7}$$

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017

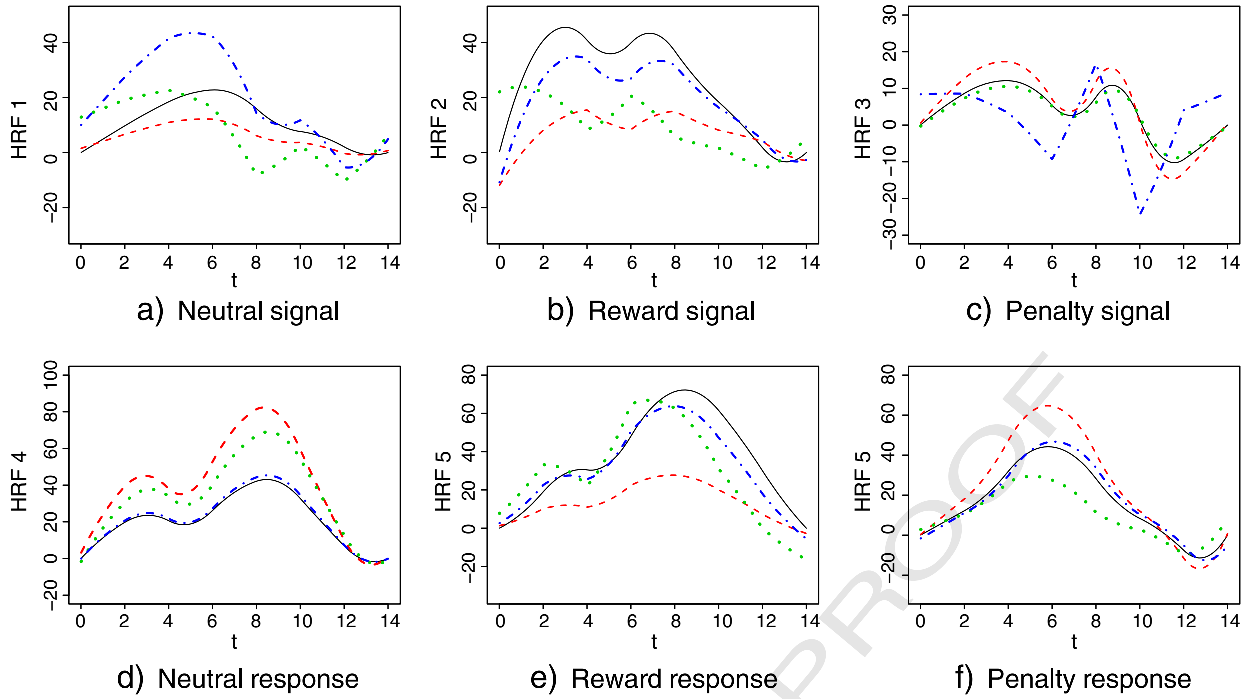

Fig. 3. Estimated HRFs of a voxel in right putamen with significant interactions in monetary penalty task. The black lines are the estimated f_k while the three broken lines are the estimated $h_{i,k}$ for three randomly selected subjects.

By assigning different values to the parameters, the six f_k's have distinct shapes. The parameters for the six HRFs are given in Table 3, and several simulated HRFs for each stimulus are displayed in Fig. 4. The first two HRFs follow a canonical shape, but differ in the range of variation in latency. The third and fourth HRFs have distinct shapes from the canonical one, but still follow the proposed semi-parametric model. The last two HRFs violate the model assumption, having a large variation both in latency and magnitude. To mimic the MID experiment, we consider three types of nonlinearity, respectively characterized by three second-order Volterra kernels:

$$V_{1,4}(t_1,t_2) = 8\exp\{|t_1/1500 + t_2/1000|\},$$
$$V_{2,5} = 2\exp\{-|t_1/1500 - t_2/1000|\}, V_{3,6} = 0,$$

for $|t_1 - t_2| \le 3.5$ and $t_1 \le 8$, and the kernels equal zero at the rest of (t_1, t_2). These kernels are chosen such that their values are close to zero at the boundary of domain $|t_1 - t_2| \le 3.5$, beyond which very few observations are available. The associated subjects' intensities of interaction, $M_{i,14}$ and $M_{i,25}$ are generated from uniform distributions with ranges $(-200, -100)$ and $(-150, -100)$, respectively, to represent negative interactions observed in many practical cases.

Table 3
Parameters of the simulated HRFs $h_{i,k}$, where $U(a, b)$ denotes uniform distribution defined on interval (a, b), and $N(\mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2.

HRF k	A_{ik}	D_{ik}	$a_{1,i}$	$a_{2,i}$	$b_{1,i}$	$b_{2,i}$	c
1	$N(700, 300^2)$	$U(-1.5,1.0)$	6	16	1	1	1/6
2	$N(500, 200^2)$	$U(-1.0,1.0)$	6	16	1	1	1/6
3	$N(400, 150^2)$	$U(0.0,4.0)$	19	20	2	2	2/3
4	$U(500, 1500)$	$U(1.0,4.0)$	20	22	2	2	9/10
5	$U(100, 500)$	$U(-3.0,0)$	$U(6,8)$	$U(15,18)$	$U(1,3)$	$U(1,3)$	1/6
6	$U(100, 500)$	$U(-2.0,1.0)$	$U(18,22)$	$U(9,25)$	$U(3,4)$	$U(3,4)$	1/4

The error terms $\varepsilon_i = (\varepsilon_i(1), \ldots, \varepsilon_i(T))'$ are simulated from an auto-regressive model of order 4 (AR(4)) with lag − 1 correlation of 0.45 and lag − 2 correlation of 0.35:

$$\varepsilon_i(t) = 0.37\varepsilon_i(t-1) + 0.14\varepsilon_i(t-2) + 0.05\varepsilon_i(t-3) + 0.02\varepsilon_i(t-4) + e_i(t),$$

where $e_i(t) \overset{i.i.d}{\sim} N\left(0, \sigma_i^2\right)$. To reflect the heteroscedastic variances across subjects, we let σ_i^2 vary across subjects, following $Ga(2,1/25) + 50$ so that generated data have a weak signal-to-noise ratio. For each simulated example below, we first generate $h_{i,k}$, $V_{i,k1k2}$ for $i = 1,\cdots, n$, $k = 1,2,\cdots, 6$ and $(k_1, k_2) \in \{(1, 3), (2, 4), (3, 6)\}$, and random second order polynomials $\mathbf{d}_i(t)\beta_i$ with, $\beta_{i,1} \sim U(-0.1, 0.1)$, $\beta_{i,2} \sim U(-0.05, 0.05)$ for each i. Then based on these, $y_i(t)$ is simulated given the design and the stimulus functions.

We use the root mean square error (RMSE) of subjects' HRF estimates and average relative errors (ARE) of the height (HR) of the estimated HRFs as the criterion for comparison:

$$e(\mathrm{HR}_k) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\mathrm{HR}_{i,k} - \widehat{\mathrm{HR}}_{i,k}\right|}{\mathrm{HR}_{i,k}}, \quad \mathrm{RMSE}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\left\|h_{i,k} - \hat{h}_{i,k}\right\|}{\left\|h_{i,k}\right\|},$$

where $\|\cdot\|$ is the L^2 norm.

Analysis and results

We evaluated the type I and type II errors of the proposed hypothesis tests on nonlinearity, and showed the histograms of P-values in testing values of $V_{1,4}$, $V_{2,5}$, and $V_{3,6}$ in Fig. 1. For zero interaction in the case of $V_{3,6}$, the histogram of P-values is close to be flat, indicating that the type I error of the test is preserved at the specified level. As shown in Figs. 5(a) and (b), the test on $V_{1,4}$ has a power close to one with all the P-values strictly less than 1%. The test on $V_{2,5}$ though has a smaller power due to its smaller value and still detects nonlinearity 36 times out of 100 simulations with threshold at 5%.

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017

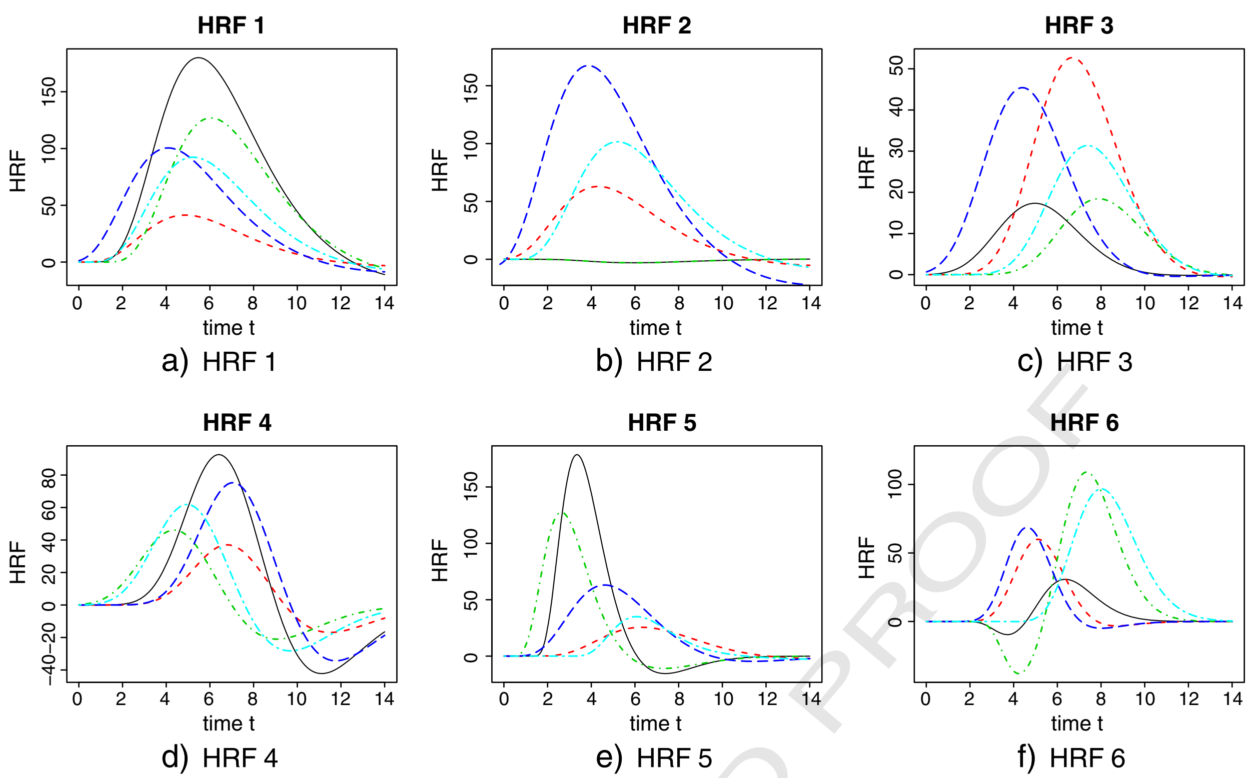

Fig. 4. Simulated HRFs for six stimuli.

Table 4 summarizes the ARE of HR and RMSE of the six HRFs obtained from the five methods, where the cubic-spline-based methods use knots equally separated by 2 unit time. Among these methods, the proposed nonlinear model generally performs the best with reasonably small errors both in estimating functional shape and HR, the linear spline model is the second best, followed by the SFIR and Tik-GCV, while the canonical method performs the worst, even when the underlying HRFs follow the canonical form (HRFs 1–2). This is not surprising given that the proposed nonlinear model is the only method that accounts for the interactions between stimuli. However, in terms of estimating a single value HR, the nonlinear and linear models have comparable performance, though the former recovers the entire curve with a much smaller error. This is probably because with the large variation of the fMRI data, the variation of the maximum value of the HRF estimates under the linear and nonlinear models is comparable, though the locations of maximum can vary significantly. The underperformance of the canonical method, especially for HRFs 3–6, is likely due to the huge overall model fitting error coming from the misspecified functional shapes of the HRFs.

Discussion

We proposed a semi-parametric nonlinear characterization of hemodynamic responses for multi-subject fMRI data based on the Volterra series. The new model is flexible to accommodate variation of brain activity across different stimuli and voxels, and allows "borrowing" information across subjects to increase estimation efficiency. Using first-order Taylor expansion and spline basis representation, the nonlinear model is converted to a bilinear one, for which we developed a fast noniterative estimation strategy. Applying the proposed method to the event-related MID study, we identified a deviation from the commonly assumed linear time-invariant system in various brain regions due to interactions between stimuli. Through Monte Carlo simulation, we also showed that the proposed method outperforms several existing methods for HRF estimation when the nonlinear effect is significant.

It is natural to extend the information-borrowing idea to spatial context, that is, information can be borrowed from neighboring voxels. In fact, spatial information has been taken into account in the pre-processing stage of fMRI data analysis, which usually involves spatial

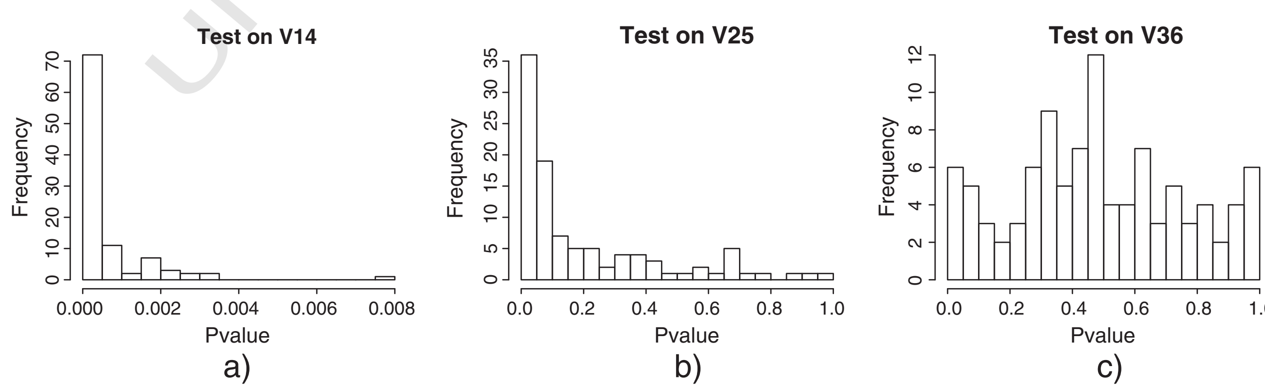

Fig. 5. Histograms of P-values for testing nonzero $V_{1,4}$, $V_{2,5}$, and zero $V_{3,6}$ respectively.

Table 4
Mean AREs for estimating HR and RMSE of the simulated HRFs from the simulated example by different methods, where the spline-based methods use knots equally spaced by 1.

	HRF	Spline-based strategies		Can.	Tik-GCV	SFIR
	k	Nonlinear	Linear			
RMSE	1	2.16	10.17	8.34	7.34	3.56
	2	1.87	4.18	9.38	4.50	2.93
	3	4.50	4.76	13.48	11.76	8.12
	4	2.16	3.67	12.16	7.08	4.18
	5	1.69	2.89	8.23	4.08	2.18
	6	1.84	2.64	10.56	5.17	2.37
e(HR)	1	3.41	3.62	29.87	5.09	29.87
	2	2.84	1.98	10.58	3.30	10.58
	3	6.29	10.02	6.33	32.45	6.33
	4	0.92	0.80	10.72	1.77	10.72
	5	0.63	0.80	7.07	1.09	7.07
	6	0.68	0.78	7.37	1.26	7.37

smoothing. Consequently, the fMRI time series at spatially-close voxels usually have similar values and the resulting parameter estimates for spatially-close voxels are very similar. In the analysis stage, it is possible to conduct another step of spatial smoothing over the estimates from the proposed model using existing methods in the literature. For example, Polzehl and Spokoiny (2000) developed a locally adaptive weight smoothing method for imaging denoising and enhancement in univariate situations where each data point associated with each image pixel/voxel can be well approximated by a local constant function depending only on the spatial location of the pixel/voxel. Li et al. (2011) extended this approach further and developed multiscale adaptive regression models for multi-subjects' vectors of image measurements. This method integrates imaging smoothing with spatial data analysis of the smoothed data. Arias-Castro et al. (2012) characterized the performance of nonlocal means and related adaptive kernel-based image denoising methods by providing theoretical bounds on the estimation errors of these methods, which depend on the number of observed pixels and the underlying imaging features. Readers are referred to Yue et al. (2010) for a more detailed overview of smoothing methods used in the neuroimaging literature.

A nontrivial number of parameters are usually required to characterize nonlinearity, which may substantially increase the variance of the estimates and thus reduce power of detecting activation when the sample size is small. On the other hand, when strong nonlinear effects present, our simulations show that estimation of the additional nonlinearity parameters does not undermine estimation of the HRFs, and in fact, ignoring them introduces large bias in the HRF estimates. Our approach to this bias-variance tradeoff is to limit the number of functional bases (and thus the number of free parameters) representing subject-specific HRFs and the 2nd-order Volterra kernel. Through simulations, we found that our approach is the most efficient when (1) the nonlinear effect is strong, and/or (2) the sample size is large, and/or (3) the number of parameters characterizing interactive effects is small. For example, in the MID application, only a small area of $V_{k_1k_2}$ was observed, which significantly reduced the number of free parameters. Consequently, the proposed nonlinear model performed well though three different types of interactions were modeled. More generally, in studies where a considerable number of pairs of interactions are modeled, estimation errors can still be reduced by utilizing the prior knowledge of the small domain of $V_{k_1k_2}$. As a practical guideline, we recommend to model nonlinearity only when the interaction effect is of interest, or is expected to be strong (e.g., in event-related designs with short ISIs).

In our estimation strategy, we only impose regularity on the 1st-order derivatives of the two arguments of $V_{i,k_1k_2}(t_1,t_2)$, without assuming high-order differentiability; estimation errors of the model may be further reduced by imposing a different roughness constraint. Moreover, different penalty parameters can be considered for roughness constraints on HRF and Volterra kernels.

We neglect the variation of interaction effect on response latency across subjects in our model for V_{i,k_1k_2} for simplicity and easy interpretation. With sufficient data, it is possible to evaluate such subject-specific interaction effect on latency by, for instance, the following semi-parametric Volterra series model, $\widetilde{V}_{i,k_1k_2}(t_1,t_2) = M_{i,k_1k_2} \cdot V_{k_1k_2}(t_1, t_2 + L_{i,k_1k_2})$ for $t_2 > t_1$, where L_{i,k_1k_2} characterizes the subject-specific latency change. Similar to the estimation of the latency coefficient $D_{i,k}$ in the HRF $h_{i,k}(t)$, we can use a first-order Taylor expansion to approximate and simplify the estimation:

$$\widetilde{V}_{i,k_1k_2}(t_1,t_2) \approx M_{i,k_1k_2} \cdot V_{k_1k_2}(t_1,t_2) + M_{i,k_1k_2} \cdot L_{i,k_1k_2} \cdot V^{(0,1)}_{k_1k_2}(t_1,t_2),$$

where the superscripts (0, 1) stand for the first order partial derivative on t_2. Based on spline representations of V_{k1k2} and f_k, we can also use a noniterative procedure to estimate $\widetilde{V}_{i,k_1k_2}(t_1,t_2)$: first estimate f_k and $V_{k_1k_2}$ through the same Steps 1–2; then evaluate subject-specific parameters $A_{i,k}$, $D_{i,k}$, M_{i,k_1k_2} and L_{i,k_1k_2} by the OLS estimates given the estimated f_k and $V_{k_1k_2}$. We can impose $\sum_i L_{i,k_1k_2} = 0$ to avoid identifiability issue. Under this restriction, if the interest is mainly on the extent of interaction, it is reasonable to use the model for $V_{i,k_1k_2}(t_1,t_2)$ proposed in the article, where subject-specific interaction effects on latency, with zero means, are incorporated into the error terms.

Higher-order, say 3rd-order, Volterra kernels can in principle be used for evaluating interactions between more than two stimuli. For the experiment with inter-stimulus interval larger than 2 s, however, this may not be beneficial because: first, the ensuing model will be overly complicated; second, biologically high-order interactions most likely will be negligible in comparison to lower-order ones if the interval between nonconsecutive stimuli is larger than 4 s.

Acknowledgments

We thank two reviewers for insightful comments. Zhang's research is partially funded by the U.S. NSF DMS grants 1209118 and 1120756. Li's research is partially funded by the U.S. NSF DMS grant 1208983. Coan's research was partially funded by grant R01MH080725 from the U.S. National Institute of Mental Health (NIMH). The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF or NIMH. Q12

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8, 360–369.
Allen, J.P., Porter, M., McFarland, F.C., McElhaney, K.B., Marsh, P., 2007. The relation of attachment security to adolescents' paternal and peer relationships, depression, and externalizing behavior. Child Dev. 78, 1222–1239.
Arias-Castro, E., Salmon, J., Willett, R., 2012. Oracle inequalities and minimax rates for non-local means and related adaptive kernel-based methods. SIAM J. Imaging Sci. 5, 944–992.
Bai, P., Truong, Y., Huang, X., 2009. Nonparametric estimation of hemodynamic response function: a frequency domain approach. IMS Lecture Notes—Monograph Series. Optimality: The Third Erich L. Lehmann Symposium, 57, pp. 190–215.
Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D., 2001. Bad is strong than good. Rev. Gen. Psychol. 5 (4), 323–370.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57 (1), 289–300.
Brosch, J., Talavage, T., Ulmer, J., Nyenhuis, J., 2002. Simulation of human respiration in fMRI with a mechanical model. IEEE Trans. Biomed. Eng. 49, 700–707.
Buckner, R.L., 1998. Event-related fMRI and the hemodynamic response. Hum. Brain Mapp. 6, 373–377.
Buxton, R.B., Frank, L.R., 1997. A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. J. Cereb. Blood Flow Metab. 17, 64–72.
Buxton, R.B., Wong, E.C., Frank, L.R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the Balloon model. Magn. Reson. Med. 39, 855–864.
Cacioppo, J.T., Gardner, W.L., Berntson, G.G., 1997. Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. Personal. Soc. Psychol. Rev. 1, 3–25.
Casanova, R., Ryali, S., Serences, J., Yang, L., Kraft, R., Laurienti, P.J., Maldjian, J.A., 2008. The impact of temporal regularization on estimates of the BOLD hemodynamic response function: a comparative analysis. NeuroImage 40 (4), 1606–1618.

Coan, J.A., Allen, J.J., McKnight, P.E., 2006. A capability model of individual differences in frontal EEG asymmetry. Biol. Psychol. 72, 198–207.

Dale, A., 1999. Optimal experimental design for event-related fMRI. Hum. Brain Mapp. 8, 109–114.

Dale, A.M., Buckner, R.L., 1997. Selective averaging of rapidly presented individual trials using fMRI. Hum. Brain Mapp. 5, 329–340.

De Boor, C., 2001. A Practical Guide to Splines (Revised Edition). Springer.

Efron, B., 2008. False discovery rates and the James–Stein estimator. Stat. Sin. 18, 805–816.

Eubank, R.L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, Inc., New York.

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

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R., 1998a. Event-related fMRI: characterizing differential responses. NeuroImage 7, 30–40.

Friston, K.J., Josephs, O., Rees, G., Turner, R., 1998b. Nonlinear event-related responses in fMRI. Magn. Reson. Med. 39, 41–52.

Friston, K.J., A Mechelli, A., Turner, E., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. NeuroImage 12, 466–477.

Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. NeuroImage 9, 416–429.

Gottman, J.M., 1994. What Predicts Divorce? Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Goutte, C., Nielsen, F.A., Hansen, L.K., 2000. Modeling the haemodynamic response in fMRI using smooth FIR filters. IEEE Trans. Med. Imaging 19, 1188–1201.

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

Huston, T.L., Vangelisti, A.L., 1991. Socioemotional behavior and satisfaction in marital relationships. J. Pers. Soc. Psychol. 61, 721–733.

Knutson, B., Westdorp, A., Kaiser, E., Hommer, D., 2000. FMRI visualization of brain activity during a monetary incentive delay task. NeuroImage 12, 20–27.

Lange, N., Strother, S.C., Anderson, J.R., Nielsen, F.A., Holmes, A.P., Kolenda, T., Savoy, R., Hansen, L.K., 1999. Plurality and resemblance in fMRI data analysis. NeuroImage 10, 282–303.

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J.H., Ibrahim, J.G., 2011. Multiscale adaptive regression models for neuroimaging data. J. R. Stat. Soc. Ser. B 73, 559–578.

Lindquist, M.A., Wager, T.D., 2007. Validity and power in hemodynamic response modelling: a comparison study and a new approach. Hum. Brain Mapp. 28, 764–784.

Lindquist, M.A., Loh, J.M., Atlas, L., Wager, T.D., 2009. Modeling the hemodynamic response function in fMRI: efficiency, bias and Mis-modeling. NeuroImage 45, S187–S198.

Lindstrom, M., 1995. Self modeling with random scale and shift parameters and a free-knot spline shape function. Stat. Med. 14, 2009–2021.

Liu, H., Gao, J., 2000. An investigation of the impulse functions for the nonlinear BOLD response in functional MRI. Magn. Reson. Imaging 18, 931–938.

Luo, H., Puthusserypady, S., 2008. Analysis of fMRI data with drift: modified general linear model and Bayesian estimator. IEEE Trans. Biomed. Eng. 55, 1504–1511.

Mandeville, J.B., Marota, J.J., Ayata, C., Zararchuk, G., Moskowitz, M.A., Rosen, B., Weisskoff, R.M., 1999. Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. J. Cereb. Blood Flow Metab. 19, 679–689.

Miezin, F.M., Maccotta, L., Ollinger, J.M., Petersen, S.E., Buckner, R.L., 2000. Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. NeuroImage 11 (6), 735–759.

Miller, K.L., Luh, W.M., Liu, T.T., Martinez, A., Obata, T., Wong, E.C., Frank, L.R., Buxton, R.B., 2001. Nonlinear temporal dynamics of the cerebral blood flow response. Hum. Brain Mapp. 13, 1–12.

Parker, R.L., Rice, J.A., 1985. Discussion of “Some aspects of the spline smoothing approach to nonparametric regression curve fitting” by B. W. Silverman. J. R. Stat. Soc. Ser. B 47, 40–42.

Polzehl, J., Spokoiny, V.G., 2000. Adaptive weights smoothing with applications to image restoration. J. R. Stat. Soc. Ser. B 62, 335–354.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second edition. Springer.

Reiss, P.T., Ogden, R.T., 2007. Functional principal component regression and functional partial least squares. J. Am. Stat. Assoc. 102, 984–996.

Reiss, P.T., Ogden, R.T., 2009. Smoothing parameter selection for a class of semiparametric linear models. J. R. Stat. Soc. Ser. B 71, 505–523.

Riera, J.J., Watanabe, J., Kazuki, I., Naoki, M., Aubert, E., Ozaki, T., Kawashima, R., 2004. A state-space model of the hemodynamic approach: nonlinear filtering of bold signals. NeuroImage 21, 547–567.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press.

Smith, A., Lewis, B., Ruttinmann, U., et al., 1999. Investigation of low frequency drift in fMRI signal. NeuroImage 9, 526–533.

Soltysik, D.A., Peck, K.K., White, K.D., Crosson, B., Briggs, R.W., 2004. Comparison of hemodynamic response non-linearity across primary cortical areas. NeuroImage 22, 1117–1127.

Vakorin, V.A., Borowsky, R., Sarty, G.E., 2007. Characterizing the functional MRI response using Tikhonov regularization. Stat. Med. 26 (21), 3830–3844.

Vazquez, A.L., Noll, D.C., 1998. Nonlinear aspects of the BOLD response in functional MRI. NeuroImage 7, 108–118.

Wager, T.D., Vazquez, A., Hernandez, L., Nollb, D.C., 2005. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. NeuroImage 25, 206–218.

Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.

Wang, J., Zhu, H., Fan, J.Q., Giovanello, K., Lin, W.L., 2011. Multiscale adaptive smoothing model for the hemodynamic response function in fMRI. MICCAI, LNCS 6892 pp. 269–276.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B 73, 3–36.

Woolrich, M.W., Behrens, T.E., Smith, S.M., 2004. Constrained linear basis sets for HRF modelling using variational Bayes. NeuroImage 21, 1748–1761.

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

Yue, Y., Loh, J.M., Lindquist, M.A., 2010. Adaptive spatial smoothing of fMRI images. Stat. Interface 3, 3–13.

Zarahn, E., 2002. Using larger dimensional signal subspaces to increase sensitivity in fMRI time series analyses. Hum. Brain Mapp. 17, 13–16.

Zhang, T., Li, F., Beckes, L., Brown, C., Coan, J.A., 2012. Nonparametric inference of hemodynamic response for multi-subject fMRI data. NeuroImage 63, 1754–1765.

Zhang, T., Li, F., Beckes, L., Coan, J.A., 2013. A semi-parametric model of the hemodynamic response for multi-subject fMRI data. NeuroImage 75, 136–145.

Zhou, S., Shen, X., 2001. Spatially adaptive regression splines and accurate knot selection schemes. J. Am. Stat. Assoc. 96, 247–259.

Please cite this article as: Zhang, T., et al., A semi-parametric nonlinear model for event-related fMRI, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.017