Wavelet minimum description length detrending for near-infrared spectroscopy

Kwang Eun Jang
Sungho Tak
Jinwook Jung
Jaeduck Jang
Korea Advanced Institute of Science and Technology
  (KAIST)
Department of Bio and Brain Engineering
Bio Imaging Signal Processing Laboratory
373-1 Guseong-dong
Yuseong-gu, Daejeon 305-701
Korea

Yong Jeong
Korea Advanced Institute of Science and Technology
  (KAIST)
Department of Bio and Brain Engineering
Cognitive and Behavioral Neuroscience Laboratory
373-1 Guseong-dong
Yuseong-gu, Daejeon 305-701
Korea

Jong Chul Ye
Korea Advanced Institute of Science and Technology
  (KAIST)
Department of Bio and Brain Engineering
Bio Imaging Signal Processing Laboratory
373-1 Guseong-dong
Yuseong-gu, Daejeon 305-701
Korea

Abstract. Near-infrared spectroscopy (NIRS) can be employed to investigate brain activities associated with regional changes of the oxy- and deoxyhemoglobin concentration by measuring the absorption of near-infrared light through the intact skull. NIRS is regarded as a promising neuroimaging modality thanks to its excellent temporal resolution and flexibility for routine monitoring. Recently, the general linear model (GLM), which is a standard method for functional MRI (fMRI) analysis, has been employed for quantitative analysis of NIRS data. However, the GLM often fails in NIRS when there exists an unknown global trend due to breathing, cardiac, vasomotion, or other experimental errors. We propose a wavelet minimum description length (Wavelet-MDL) detrending algorithm to overcome this problem. Specifically, the wavelet transform is applied to decompose NIRS measurements into global trends, hemodynamic signals, and uncorrelated noise components at distinct scales. The minimum description length (MDL) principle plays an important role in preventing over- or underfitting and facilitates optimal model order selection for the global trend estimate. Experimental results demonstrate that the new detrending algorithm outperforms the conventional approaches. © *2009 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.3127204]

Keywords: Near infrared spectroscopy; Wavelet transforms; Signal processing; detrending; minimum description length principle.

Paper 08052RR received Feb. 12, 2008; revised manuscript received Mar. 12, 2009; accepted for publication Mar. 12, 2009; published online May 11, 2009. This paper is a revision of a paper presented at the SPIE conference on Multimodal Biomedical Imaging III, January 2008, San Jose, California. The paper presented there appears (unrefereed) in SPIE Proceedings Vol. 6850.

1 Introduction

Near-infrared (NIR) light, with a wavelength between 650 nm and 950 nm, is capable of penetrating deeply through biological tissues. This is because NIR light is weakly absorbed by biological chromophores such as hemoglobin, myoglobin, and cytochrome c oxidase.[1] The relatively deep penetration depth of NIR light in the human brain makes it possible to measure brain activities associated with regional changes of oxy- and deoxy hemoglobin concentrations.[2] This spectroscopic technique using NIR light for monitoring brain activities is known as functional near-infrared spectroscopy (NIRS).[3]

NIRS is regarded as a promising neuroimaging modality owing to a number of advantages over other neuroimaging modalities such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI).[3] For example, there is no theoretical limitation on temporal resolution (while of course there are practical limits arising from, for example, the speed of the analog-to-digital converter). The temporal resolution of NIRS, therefore, is sufficient to investigate hemodynamic responses due to brain activations as well as other fast varying physiological conditions. Furthermore, NIRS does not require that the subject lie on his/her back in a confined environment during experiments, thereby making it possible to investigate subjects that would normally be difficult to examine using fMRI or PET, including infants, children, and patients with psychological issues. Moreover, NIRS is highly flexible, portable, and relatively low cost.

However, there remain several theoretical and practical difficulties for quantitative analysis of NIRS data. For example, the differential path length factor (DPF)[4] depends on various parameters such as the age of the subject,[5] wavelength of the imaging system,[6] and position within a brain.[7] Time-resolved or frequency domain NIRS equipment may be used to estimate the mean optical path length; however, continuous wave (CW) systems are more commonly used due to several practical concerns such as cost of implementation.[7] Other subject-dependent parameters such as depth of the skull and optical properties of hair may influence measurement of optical parameters.

Recently, many researchers have been developing statistical analysis toolboxes for NIRS based on the generalized linear model (GLM).[8–11] The GLM is a statistical linear model that explains data as a linear combination of explanatory variables plus an error term. Since the GLM analysis relies on the temporal variational pattern of signals, it is more robust to differential path length factor (DPF) variation, optical scattering, or poor contact. Furthermore, statistical parameter mapping (SPM) using the GLM is a standard method for analyzing fMRI data,[12] and thus integration of NIRS and fMRI

within the same SPM framework may offer the advantage of modeling both types of data in the same mathematical framework to make inferences. Based on these observations, we have developed a new public domain statistical toolbox called NIRS-SPM.[11] By incorporating the GLM with the *p*-value calculation using Sun's tube formula,[13,14] NIRS-SPM not only enables calculation of activation maps of oxy-, deoxy, and total hemoglobin but also allows for super resolution localization, which is not possible using conventional analysis tools.[11]

Note that the GLM often fails when there exist global drifts in the NIRS measurements due to various reasons, such as subject movement, blood pressure variation, and instrumental instability. This global trend causes a low-frequency bias in NIRS measurements. Moreover, the amplitude of the global drift is often comparable to that of the signal from brain activation, which degrades the signal-to-noise ratio. In order to eliminate the global trend and thus improve the signal-to-noise ratio, high-pass filtering is often used in practice.[15] However, the signals from brain activations are often degraded by a simple filtering, since the frequency response of the hemodynamic response can also be affected during high-pass filtering.

In order to overcome this problem, this paper investigates a wavelet-based detrending algorithm for fMRI[16] and adapts it for NIRS applications. Specifically, the wavelet transform was applied to decompose NIRS measurements into global trends, hemodynamic signals, and noise components at distinct scales. However, unlike fMRI data, the NIRS time series is considerably long due to the fast sampling frequency. Hence, we observed that direct application of the model order selection rule in Ref. [16] leads to erroneous results in NIRS. To remedy this problem, the minimum description length (MDL) principle with universal prior of integers[17] is found to be suitable for the NIRS time series, as it avoids over- and underfitting of the global trend estimate, thanks to the asymptotic optimality of MDL. Experimental results confirm that the new detrending algorithm outperforms the conventional approaches. We have therefore incorporated the proposed Wavelet-MDL detrending algorithms within our NIRS-SPM framework, which will soon be publicly available at the website of the authors (http://bisp.kaist.ac.kr/NIRS-SPM). We observed that the Wavelet-MDL detrending method within NIRS-SPM provides more specific localizations of the neuronal activation than the standard high-pass filtering approach based on the discrete cosine transform (DCT).

2 NIRS Measurement Model

The modified Beer-Lambert law (MBLL), which describes optical attenuation in a highly scattering medium such as biological tissue,[4] provides a relation between raw optical density (OD) data and changes of chromophore concentrations. According to the MBLL, the change in $OD(\lambda, r, t)$ for the wavelength λ at the cerebral cortex position $r \in \mathfrak{R}^3$ at time t due to the N_c number of chromophore concentration changes $[\Delta c^{(i)}(r,t)]_{i=1}^{N_c}$ is described as

$$\Delta OD(\lambda, r, t) = -\ln\left(\frac{I_F}{I_o}\right) = \sum_{i=1}^{N_c} a_i(\lambda)\Delta c^{(i)}(r,t)d(r)l(r), \quad (1)$$

where I_F denotes the final measured optical intensity, I_o denotes the initial measured optical intensity, $a_i(\lambda)$ is the extinction coefficient of the *i*'th chromophore at wavelength λ, $d(r)$ is the DPF, and $l(r)$ is the distance between the source and detector at position r, respectively. Assuming that oxy- and deoxyhemoglobin are the major two choromophores, the noisy measured optical density is then described by the following matrix formulation:

$$\begin{bmatrix} \Delta OD(r,t;\lambda_1) \\ \Delta OD(r,t;\lambda_2) \end{bmatrix} = d(r)l(r) \begin{bmatrix} a_1(\lambda_1) & a_2(\lambda_1) \\ a_1(\lambda_2) & a_2(\lambda_2) \end{bmatrix} \begin{bmatrix} \Delta c_{HbO}(r,t) \\ \Delta c_{HbR}(r,t) \end{bmatrix}$$
$$+ \begin{bmatrix} w(r,t;\lambda_1) \\ w(r,t;\lambda_2) \end{bmatrix}, \quad (2)$$

where $\Delta c_{HbO}(r,t)$ and $\Delta c_{HbR}(r,t)$ denote the time series of the chromophore changes for the oxy- and deoxyhemoglobin, and $w(r,t;\lambda_i)$ is the additive noise for the wavelength λ_i, respectively. Here, we assume that $d(r)$ is the same at both wavelengths and the equality assumption does not affect the validity or applicability of what follows. Then, by multiplying the inverse matrix of the extinction coefficients with Eq. (2), we can derive the expression of the noisy oxy- and deoxyhemoglobin signals:

$$\begin{bmatrix} y_{HbO}(r,t) \\ y_{HbR}(r,t) \end{bmatrix} = d(r)l(r) \begin{bmatrix} \Delta c_{HbO}(r,t) \\ \Delta c_{HbR}(r,t) \end{bmatrix} + \begin{bmatrix} \epsilon_{HbO}(r,t) \\ \epsilon_{HbR}(r,t) \end{bmatrix}, \quad (3)$$

where $\epsilon_{HbO}(r,t)$ and $\epsilon_{HbR}(r,t)$ is the additive zero mean Gaussian noise for the oxy- and deoxy- channels, respectively. Although the DPF parameter $d(r)$ can be measured using time-domain or frequency-domain systems by calculating the temporal point spread function,[7] this information is not obtainable in commonly available CW systems. Furthermore, NIRS data acquisition is considerably affected by a variety of measurement conditions, such as the color of hair and the scalp depth, which introduces position- and subject-dependent scattering effects. For these reasons, analyzing NIRS data using the magnitude of chromophore concentration changes is often problematic.

3 General Linear Model for NIRS

In the fMRI domain, the validity of the GLM has been extensively tested, and the GLM has been established as a standard analyzing method. Statistical parametric mapping (SPM)[18] and analysis of functional neuro images (AFNI)[19] are widely used programs based on the GLM. Analysis based on the GLM consists of three steps: model specification, parameter estimation, and statistical inference.[15] In this section, we review the GLM approach for NIRS applications.[11]

The GLM describes a measurement $y_{HbX}(r,t)$ (i.e., y_{HbO} or y_{HbR}) in terms of a linear combination of L explanatory variables plus an error term:

$$y_{HbX}(r,t) = x_1(t)\beta_1 + \cdots + x_L(t)\beta_L + \epsilon_{HbX}(r,t). \quad (4)$$

Here, β_i denotes an unknown strength of response, and $x_i(t)$ is an explanatory variable originating from a model of hemody-

namic responses. Now, let y and ϵ denote the vector of the time series of the hemodynamic signal and noise at the location r, respectively:

$$y = [y_{HbX}(r,t_1)\ y_{HbX}(r,t_2)\ \cdots\ y_{HbX}(r,t_N)]^T, \qquad (5)$$

$$\epsilon = [\epsilon_{HbX}(r,t_1)\ \epsilon_{HbX}(r,t_2)\ \cdots\ \epsilon_{HbX}(r,t_N)]^T. \qquad (6)$$

The corresponding GLM model in a matrix form is then given by:

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad (7)$$

where y is an N-dimensional column vector whose elements are the sampled NIRS data at N time points, ϵ denotes an error vector, and β is an L-dimensional column vector that represents unknown strengths of the response. Usually, the $N \times L$ matrix X is called a design matrix and serves as a predictor for the measured signal.[15]

For fMRI signals, Boynton et al. showed that the BOLD signal can be approximated as a convolution model between a stimulus function and a hemodynamic response function (HRF).[20] Based on a similar argument, several statistical analysis toolboxes using the GLM are currently available for NIRS.[8–11] The stick function or the boxcar function is typically used for the stimulus function. For the HRF, there are a number of possible models. In this paper, we follow fMRI approaches and employ the so-called canonical HRF, which is composed of two gamma functions.[15] Additionally, the derivatives of the HRF with respect to delay and dispersion can be used to mitigate the problem that the precise shape of the HRF varies across the brain.[21] An adaptive estimation of HRF using multiple gamma functions can also be used in NIRS to account for oxygen species–dependent hemodynamics variation.[11]

After the model specification, the least-squares parameter estimator is derived using the ordinary least squares. If the design matrix X is of full rank, the least-squares estimate is:

$$\hat{\boldsymbol{\beta}} = (X^T X)^- X^T y, \qquad (8)$$

where X^- denotes the pseudo-inverse of X. With the obtained least-squares estimates, one can construct statistics for the statistical inference. In most cases, we consider a linear combination of the parameter estimates:

$$c_1\hat{\beta}_1 + \cdots + c_L\hat{\beta}_L = c^T\hat{\boldsymbol{\beta}}, \qquad (9)$$

where the vector c is called a contrast vector.[15] Similar to the fMRI-SPM analysis,[15] the error ϵ in Eq. (7) is assumed to be normally distributed with a temporal covariance matrix Σ; hence, we have

$$c^T\hat{\boldsymbol{\beta}} \sim N[c^T\boldsymbol{\beta}, c^T(X^TX)^-X^T\Sigma X(X^TX)^-c]. \qquad (10)$$

Thus, t-statistics for the null hypothesis that asserts no activation is given by

$$t = \frac{c^T\hat{\boldsymbol{\beta}}}{[c^T(X^TX)^-X^T\Sigma X(X^TX)^-c]^{1/2}}, \qquad (11)$$

where t denotes a random variable with a Student's t-distribution with degree of freedom df given as follows:[15]

$$df = \frac{tr[R\Sigma]^2}{tr[R\Sigma R\Sigma]}, \quad R = I_{N \times N} - X(X^TX)^-X^T, \qquad (12)$$

where $I_{N \times N}$ denotes the $N \times N$ identity matrix.

The estimate of the temporal correlation matrix Σ hence affects the overall t-value and corresponding inferences. For a detailed discussion on the estimation of Σ, readers can refer to our previous work on this issue.[11] If the calculated t-value is larger than a certain threshold value, then the inference steps abandon the null hypothesis and we declare the area to be activated. The threshold value is calculated by fixing a p-value in the range of 0.001 to 0.05. A smaller p-value provides a higher threshold value. The nonlinear relationship between the p-value and threshold can be explicitly represented using the tube formula, as described in our companion paper.[11]

4 Wavelet-MDL Detrending

In this section, we develop a novel detrending algorithm that is designed to address the global-drift issues described in the Introduction. Our point of departure is an algorithm developed for detrending the fMRI time series.[16]

4.1 Notation

We first introduce the notation associated with a discrete wavelet transform by following the standard conventions.[22] Let $\psi(t)$ denote a wavelet associated with the multiresolution analysis.[22] Let $\Phi(t)$, h, and g be the scaling function, the low-pass filter, and the high-pass filter associated with this wavelet transform, respectively. With a slight abuse of notation, let $\boldsymbol{\theta} = \{\theta[n]\}, n = 0, \ldots, N-1$ be a discrete version of a continuous signal $\boldsymbol{\theta}(t)$. For simplicity, we assume $N = 2^J$, where J is the maximum level of wavelet decomposition. The wavelet coefficients composed of approximation coefficients $\{a\theta_j[k]\}_{j,k}$ and detail coefficients $\{d\theta_j[k]\}_{j,k}$ are defined by the following recursions:[22]

$$a\theta_0[k] = \theta[k], \quad k = 0, \ldots, N-1,$$

$$a\theta_{j+1}[k] = \sum_n h[n-2k]a\theta_j[n], \quad k = 0, \ldots, 2^{-j-1}N-1,$$

$$d\theta_{j+1}[k] = \sum_n g[n-2k]a\theta_j[n], \quad k = 0, \ldots, 2^{-j-1}N-1,$$

where $j = 0, \ldots, J-1$. We introduce a matrix W to represent the discrete wavelet transform:

$$W\boldsymbol{\theta} = \{a\theta_J[0], d_J, d_{J-1}, \ldots, d_1\}^T, \qquad (13)$$

where each submatrix is given by

$$d_J = \{d\theta_J[0]\} \in \mathfrak{R}^{1 \times 1},$$

$$d_{J-1} = \{d\theta_{J-1}[0], d\theta_{J-1}[1]\}^T \in \Re^{2 \times 1},$$

$$d_j = \{d\theta_j[0], \ldots, d\theta_j[2^{-j}N - 1]\}^T \in \Re^{2^{-j}N \times 1},$$

$$d_1 = \{d\theta_1[0], \ldots, d\theta_1[2^{-1}N - 1]\}^T \in \Re^{2^{-1}N \times 1}.$$

4.2 Modified GLM with Baseline Drift

In the wavelet detrending algorithm for fMRI,[16] the baseline drift is included as part of the GLM:

$$y = X\beta + \epsilon + \theta, \tag{14}$$

where y indicates the measured BOLD signal, X and ϵ denote the predictor and the additive noise with the temporal covariance matrix Σ, and θ is the additional global drift, respectively. A similar argument may be applied to the NIRS case. We introduce trend terms in Eq. (2) as follows:

$$\begin{bmatrix} \Delta OD(r,t;\lambda_1) \\ \Delta OD(r,t;\lambda_2) \end{bmatrix} = d(r)l(r) \begin{bmatrix} a_1(\lambda_1) & a_2(\lambda_1) \\ a_1(\lambda_2) & a_2(\lambda_2) \end{bmatrix} \begin{bmatrix} \Delta c_{HbO}(r,t) \\ \Delta c_{HbR}(r,t) \end{bmatrix}$$
$$+ \begin{bmatrix} w(r,t;\lambda_1) \\ w(r,t;\lambda_2) \end{bmatrix} + \begin{bmatrix} \widetilde{\theta}(r,t;\lambda_1) \\ \widetilde{\theta}(r,t;\lambda_2) \end{bmatrix}, \tag{15}$$

where $\widetilde{\theta}(r,t;\lambda_i)$ denotes the global trend for wavelength λ_i at the location r. If we multiply the inverse matrix of the extinction coefficients, the equation of the noisy oxy- and deoxyhemoglobin is given as

$$\begin{bmatrix} y_{HbO}(r,t) \\ y_{HbR}(r,t) \end{bmatrix} = d(r)l(r) \begin{bmatrix} \Delta c_{HbO}(r,t) \\ \Delta c_{HbR}(r,t) \end{bmatrix} + \begin{bmatrix} \epsilon_{HbO}(r,t) \\ \epsilon_{HbR}(r,t) \end{bmatrix}$$
$$+ \begin{bmatrix} \theta_{HbO}(r,t) \\ \theta_{HbR}(r,t) \end{bmatrix}, \tag{16}$$

where $\theta_{HbO}(r,t) = 1/C[a_2(\lambda_2)\widetilde{\theta}(r,t;\lambda_1) - a_2(\lambda_1)\widetilde{\theta}(r,t;\lambda_2)]$, $\theta_{HbR}(r,t) = 1/C[-a_1(\lambda_2)\widetilde{\theta}(r,t;\lambda_1) + a_1(\lambda_1)\widetilde{\theta}(r,t;\lambda_2)]$, and $C = [a_1(\lambda_1)a_2(\lambda_2) - a_2(\lambda_1)a_1(\lambda_2)]$. Let y_{HbX}, ϵ_{HbX}, and θ_{HbX} denote vectors of the hemodynamic signal, additive noise, and global trend signal, respectively. If we introduce the GLM for each chromophore, we can derive the modified GLM for NIRS as follows:

$$y_{HbX} = X_{HbX}\beta_{HbX} + \epsilon_{HbX} + \theta_{HbX}. \tag{17}$$

For simplicity, we remove the subscript HbX from Eq. (17) and use the general form given by Eq. (14).

The global trend signal varies smoothly in most cases. Based on this observation, conventional detrending algorithms use filtering to remove the low-frequency trend signal. The problem of this approach, however, is that the hemodynamic signal often has a low-frequency varying component that can be erroneously removed during the filtering. In order to deal with this artifact, in our wavelet detrending, the unknown trend is modeled as a signal restricted in a subspace spanned by coarse scale wavelets.[16] More specifically, the trend is modeled as:[16]

$$\theta(t) = a\theta_J[0]\Phi(2^{-J}t) + \sum_{j=J_0}^{J} \sum_{k=0}^{2^{-j}N-1} d\theta_j[k]\psi(2^{-j}t - k), \tag{18}$$

where Φ, ψ, $a\theta_J$, $d\theta_j$, J, and N are defined as earlier, and J_0 denotes the finest scale that determines the smoothness of the trend. Note that the detail coefficients $d\theta_j[k]$ are all zero for fine scales, i.e., $1 \leq j \leq J_0 - 1$.

Using the discrete wavelet transform (DWT) matrix W defined in Eq. (13), we can represent the wavelet transform of a global trend signal as follows:

$$W\theta = [a\theta_J[0], d_J, \ldots, d_{J_0}, 0, \ldots, 0]^T. \tag{19}$$

The maximum likelihood estimates for the trend θ and the unknown signal strength β in Eq. (14) are then given by:[16]

$$\hat{\xi} = [A^T\Sigma^{-1}A]^{-1}A^T\Sigma^{-1}Wy, \tag{20}$$

where $\xi = \{a\theta_J[0], d\theta_J[0], \ldots, d\theta_{J_0}[2^{-J_0}N - 1], \beta\}^T$, and

$$A = \begin{bmatrix} 1 & & & 0 & ax_J^{(1)}[0] & \cdots & ax_J^{(L)}[0] \\ & 1 & & 0 & x_J^{(1)} & & x_J^{(L)} \\ & \cdots & & \vdots & \vdots & \cdots & \vdots \\ & & 1 & 0 & \vdots & & \vdots \\ & & & 1 & x_{J_0}^{(1)} & \cdots & x_{J_0}^{(L)} \\ & & & 0 & x_{J_0-1}^{(1)} & & x_{J_0-1}^{(L)} \\ & & & \vdots & \vdots & \cdots & \vdots \\ & & & 0 & x_1^{(1)} & \cdots & x_1^{(L)} \end{bmatrix}$$
$$= \begin{bmatrix} I_{n_0 \times n_0} & & \\ & WX & \\ 0_{(N-n_0) \times n_0} & & \end{bmatrix},$$

where $n_0 = 2^{-J_0+1}N$ denotes the number of nonzero coefficients that describe the trend, $I_{n_0 \times n_0}$ denotes a $n_0 \times n_0$ identity matrix, $0_{(N-n_0) \times n_0}$ denotes a $(N-n_0) \times n_0$ matrix whose elements are zero, A is a $N \times (n_0+L)$ matrix, Σ denotes the $N \times N$ noise covariance matrix, W is the DWT matrix, and $x_i^{(k)}$ is the i'th wavelet coefficient of the k'th column of design matrix, respectively.

Note that Σ in Eq. (20) is the covariance matrix of noise in wavelet domain. Even if the original time series is highly correlated, the array of wavelet coefficients exhibits much less correlation.[23] This decorrelating (or whitening) feature of the wavelet transform has been studied in Refs. 23–25. Especially for the fractional Brownian motion (fBM) process,[26] which is a popular model for $1/f$-type noise, the decay rate of the correlation of wavelet coefficients in j'th scale is derived as:[24]

$$E\{d\theta_j[m], d\theta_j[n]\} = O(|2^{-j}(m-n)|^{2(H-p)}), \tag{21}$$

where $E\{\cdot\}$ is an expectation operator, $H \in (0.5, 1)$ is a constant that defines the degree of correlation, and p denotes the number of vanishing moments of a wavelet transform. Therefore, for a wavelet transform that has a large enough p, we may assume that the noise in wavelet coefficients has the

uncorrelated Gaussian distribution whose covariance matrix is given as:

$$\mathbf{\Sigma} = \text{diag}\{\sigma_J^2, \sigma_J^2, \ldots, \sigma_1^2\}, \tag{22}$$

where σ_j^2 is the variance in the j'th decomposition level. To estimate σ_j^2, we use the median absolute deviation of wavelet coefficients:[23,27]

$$\hat{\sigma}_j^2 = \text{Median}\{d\theta_j[0], \ldots, d\theta_j[2^{-j}N - 1]\}/0.6745, \tag{23}$$

where the number 0.6745 is the calibration factor for Gaussian distribution.[23,27]

In the wavelet detrending method for a fMRI time series,[16] the complexity of the unknown global bias is solely determined by J_0. In other words, there is an implicit assumption that all coefficients at the same scale are all zero or all nonzero simultaneously. However, in the case of NIRS, the number of data points N is much larger than that of the fMRI time series, and thus this simple scheme may cause over- or underfitting. Therefore, for a fine-tuned estimate of model order, the wavelet coefficients at the same scale are sorted in order of descending magnitude and are included one by one until a suitable complexity is found by the model order selection criterion. Note that this procedure is similar to MDL thresholding in signal restoration.[28,29]

The wavelet detrending method has several advantages over the conventional filtering approach. Note that the conventional filtering cannot prevent the removal of the hemodynamic signal, since it decides the cutoff frequency using only the paradigm repetition frequency, and the nonnegligible amount of the hemodynamic signals can often be filtered out. This is especially true for the event-related paradigm[30] since it does not have a well-defined cutoff frequency. Compared with the conventional method, the wavelet-based detrending algorithm includes the design matrix X in the least-squares estimation process described in Eq. (20). Hence, the canonical hemodynamics time series is fully utilized in the estimation process, and the algorithm is more robust. Furthermore, the wavelet-based approach is more effective in removing relatively fast varying trends thanks to the optimality of wavelet transform in describing the transient changes of fast varying signals.[22]

4.3 Minimum Description Length Principle for Model Order Selection

In Eq. (20), we have shown a simple one-step method for simultaneous estimation of the global trend $\boldsymbol{\theta}$ and the signal strength $\boldsymbol{\beta}$ for a *given* hyperparameter J_0 in terms of maximum likelihood estimation. The remaining issue is the determination of the hyperparameter J_0. Even though J_0 is a single integer valued parameter, it affects the whole behavior of the estimated trend signal, since J_0 determines the number of wavelet coefficients n_0 that describe the trend. More specifically, if n_0 is an inappropriately large number, the hemodynamic response is distorted due to the overfitted trend estimate, whereas a smaller n_0 value may not capture the unknown trend properly. Therefore, the accuracy of the order estimate J_0 (or n_0) ultimately determines the quality of the trend estimation.

The problem of estimating the order J_0 (or n_0) is formally called the model order selection problem.[17,31,32] A good model order selection criterion should satisfy two conflicting goals simultaneously: goodness of fit and concision of the model. Balancing these two goals is crucial in preventing the over- or underfitting. Akaike information criterion (AIC),[33] Schwartz information criterion (SIC),[32] and minimum description length (MDL) suggested by Rissanen[17] are the most popular criteria currently available. The best model is then selected as the one that gives the smallest cost among several plausible models.

These three methods assume that the probability density function (pdf) is a function of unknown model order n_0. Thus, the distance between the true pdf and the estimated pdf induced from the parameter estimate $\hat{n}_0$ plays an important role. More specifically, AIC and SIC are closely related to the Kullback-Leibler (K-L) distance,[34] defined as follows:

$$I(f,g) = \int f(\mathbf{y})\log\left[\frac{f(\mathbf{y})}{g(\mathbf{y}|n_0)}\right]d\mathbf{y} = E_f\left\{\log\left[\frac{f(\mathbf{y})}{g(\mathbf{y}|n_0)}\right]\right\}, \tag{24}$$

where $E_f[\cdot]$ stands for the expectation with respect to f, f denotes a true pdf for data vector $\mathbf{y}$, and g denotes an approximated pdf governed by the model order n_0, respectively. The K-L distance can serve as a measure of the difference between the true pdf and an estimated pdf. The K-L distance gives a nonnegative value in general, but zero if and only if $f = g(\mathbf{y}|n_0)$. The basic concept of AIC and SIC is to minimize the K-L distance. However, it is not trivial to evaluate the K-L distance directly, since it is hard to define the true pdf f and $g(\mathbf{y}|n_0)$. Therefore, asymptotic approximations for the K-L distance have been derived. For example, AIC uses a cross-validation perspective, and SIC employs a prior pdf for the model order n_0 (Refs. 35). An improved version of AIC that reduces the probability of overfitting is also available,[31] and denoted by AIC_C. In many cases, AIC_C outperforms the conventional AIC.[35] The AIC_C and SIC cost functions for detrending are then summarized as:

$$\text{AIC}_C = -\log P(\mathbf{y}|n_0) + \frac{1}{2}\frac{N+n_0}{N-n_0-2},$$

$$\text{SIC} = -\log P(\mathbf{y}|n_0) + \frac{1}{2}n_0\log N, \tag{25}$$

where $\log P(\mathbf{y}|n_0)$ denotes the log-likelihood of $\mathbf{y}$ under the model order n_0. The first term $-\log P(\mathbf{y}|n_0)$ can be easily computed under an assumption that the noise is uncorrelated, independent, and normally distributed:

$$-\log P(\mathbf{y}|n_0) = -\log\left[(2\pi\sigma^2)^{-N/2}\exp\left(-\frac{\|\mathbf{y}-X\boldsymbol{\beta}-\boldsymbol{\theta}\|^2}{2\sigma^2}\right)\right], \tag{26}$$

where σ^2 denotes the unknown noise variance. Using the partial derivative of Eq. (26) with respect to σ^2, the maximum likelihood estimate $\hat{\sigma}^2$ of the unknown noise variance is given as
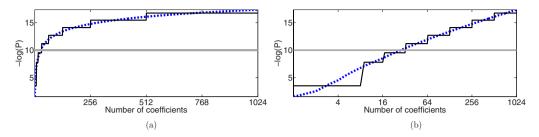
Fig. 1 (a) A prior code length: $-\log_2(P)$. (b) The log-scale view of (a).

$$\hat{\sigma}^2(n_0) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\theta}\|^2, \qquad (27)$$

where we explicitly represent the dependency of $\hat{\sigma}^2$ on the parameter n_0. Substituting Eq. (27) into Eq. (26), we get

$$-\log P(\mathbf{y}|n_0) = \frac{N}{2}\log \hat{\sigma}^2(n_0) + const, \qquad (28)$$

where *const* stands for a term that is not related to the model selection problem. Then, by substituting Eq. (28) into Eq. (25), we can find the optimal model order n_0 that minimizes the cost functions AIC_C and SIC, respectively.

The MDL principle follows a different philosophy compared with AIC and SIC. Formally, it is based on the Kolmogorov's descriptive complexity,[36,37] which defines the amount of complexity as the length of the shortest binary computer program that is able to describe an object. In more intuitive terms, the MDL principle is based on the so-called Occam's razor, interpreted as "If there are many explanations consistent with the observed data, choose the simplest."[37] However, it is hard to compute to Kolmogorov's descriptive complexity.[37] Rissanen suggested that the description length can be regarded as the number of binary bits used for data transmission between a hypothetical pair of an encoder and a decoder.[17] This idea is heavily supported by Shannon's source coding theorem, since the expected description length of data is minimum on average when the true model parameter of the pdf is used.[37,38] Note that this interpretation is parallel to K-L distance, which is zero if and only if the model parameter is true.

According to MDL, our task is to measure the total expected codelength to encode the measurement and the model. More specifically, the total code is composed of two parts: one for encoding measurements based on a model and the other for encoding the model itself. Hence, the total code length is described as

$$MDL(n_0) = -\log_2 P(\mathbf{y}|n_0) + L(n_0), \qquad (29)$$

where $-\log_2 P(\mathbf{y}|n_0)$ is related to the goodness-of-fit with respect to $\log_2$ basis to translate it into the code length in bits. The code length $L(n_0)$, however, is distinct from AIC and SIC. In MDL detrending, the hypothetical encoder should encode the position of each wavelet coefficient as well as its magnitude. If we assume the uniform distribution of the position of coefficients and the optimized truncation for real-valued magnitude,[39] the code length $L(n_0)$ for MDL criterion can be written as:[28]

$$L(n_0) = \frac{1}{2}n_0 \log_2 N + n_0 \log_2 N = \frac{3}{2}n_0 \log_2 N, \qquad (30)$$

where $n_0 \log_2 N$ and $(1/2)n_0 \log_2 N$ denote the code length for the location and magnitude of the wavelet coefficients, respectively. Since Eq. (30) was originally derived by Saito for a wavelet image compression problem,[28] we call this Saito's MDL.

Note that the code length for encoding positions of coefficients in Saito's MDL can be written as

$$(1/2)n_0 \log_2 N = n_0 \left[-\log_2 \left(\frac{1}{N} \right) \right], \qquad (31)$$

which implies that locations are encoded using Shannon's coding scheme under the assumption of uniformly distributed coefficients along all decomposition levels. However, considering the smoothness of a global trend, it is not appropriate to assume that all wavelet coefficients have identical probability of being components of the trend estimate regardless of their scale. In practice, we can easily conjecture that coarser scale wavelet coefficients are more likely to be included in the trends.
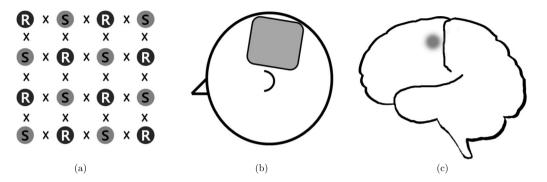
Therefore, we consider an alternative *a priori* distribution for the wavelet coefficients, whose probability varies across scale. Note that the new code should satisfy Kraft's inequality, defined as:

$$\sum_{x \in X} 2^{-L(x)} \le 1, \qquad (32)$$

where X denotes a set of binary codes, and L denotes the length of the binary codes. Kraft's inequality is the sufficient and necessary condition for the prefix code that guarantees the unique translation of received code words.[37,38] This is a basic constraint for pdf to be used in the MDL framework. Among a number of pdfs that have scale dependent probabilities, we select the universal prior for integers proposed by Rissanen:[39]

$$P_u(n) = 2^{-L_u(n)}, \quad n > 0, \quad L_u(n) = \log_2^* n + \log_2 c,$$
$$c \approx 2.865064, \qquad (33)$$

where $\log_2^* n = \log_2 n + \log_2(\log_2 n) + \log_2[\log_2(\log_2 n)] + \dots$, the sum involves only the nonnegative terms, and c is defined to bring the left side of Kraft's inequality to unity. Interestingly, the corresponding code length for this prior assigns a shorter code length for coarser scale wavelet coefficients, as shown in Figs. 1(a) and 1(b). On the contrary, the uniform

Fig. 2 RFT experiment setup. (a) Arrangement of optodes. Eight black circles denote receivers, and eight gray circles denote sources. *X* signs correspond to 24 source and receiver pairs. (b) Overall position of optodes. (c) Target area for right-finger tapping experiments, which is located at the motor cortex of the left hemisphere.

prior in Saito's MDL assigns the same code length for all scales. Therefore, the universal prior for integers provides a larger penalty for choosing a finer scale; hence, it prefers a coarser scale trend estimate.

By extending this concept, it is reasonable to assign the same code length for wavelet coefficients within the same scale. A slight modification is required in order to give an identical probability to coefficients within the scale. Specifically, suppose m_j denotes the number of wavelet coefficients in the j'th scale. The code length $\hat{L}(j)$ from the modified universal prior corresponding to the j'th scale is then given by

$$\hat{L}(j) = \frac{1}{m_j} \sum_{n=0}^{m_j-1} L_u(n + s_j), \qquad (34)$$

where $s_j = 1 + \sum_{k=j}^{j+1} m_k$ denotes the starting index of the j'th scale wavelet coefficients, J denotes the coarsest scale, and $L_u(\cdot)$ is defined in Eq. (33), respectively. This modified universal prior is illustrated in Fig. 1, which exhibits a staircase-like increasing behavior in the code length. The final MDL criterion can therefore be summarized as follows:

$$MDL(n_0) = \frac{N}{2} \log_2 \hat{\sigma}^2(n_0) + \frac{1}{2} n_0 \log_2 N + \sum_{j=J_0}^{J} m_j \hat{L}(j), \qquad (35)$$

where $n_0 = \sum_{j=J_0}^{J} m_j$. Here, $(1/2)n_0 \log_2 N$ encodes the magnitude, whereas $\sum_{j=J_0}^{J} m_j \hat{L}(j)$ encodes the location.

We have observed that both AIC_C and SIC tend to give an overfitted model for the NIRS signal, as described in the experimental results. This is because the number of temporal NIRS sequences is much larger than that of fMRI data, and it is known that the probability of overfitting is more than zero under quite general conditions, as $N \to \infty$ in these priors.[35] However, the MDL criterion does not exhibit such overfitting thanks to its asymptotic optimality.

4.4 Implementation Issues

In order to make at least one wavelet coefficient free of boundary artifacts, the maximum level of decomposition J should allow at least M coefficients at the last decomposition level, where M denotes the support length of the wavelet. Under this constraint, the maximum level of wavelet decomposition is given by:
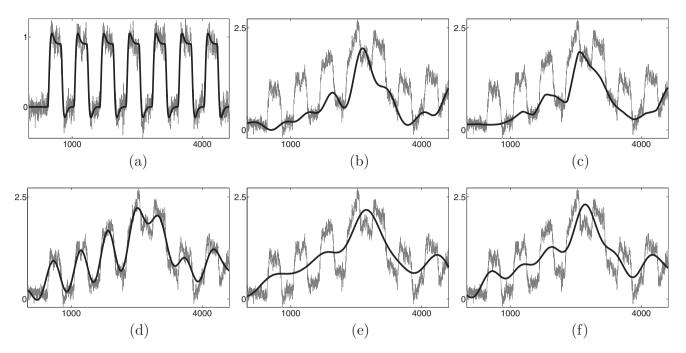
$$J_p = \left\lfloor \log_2 \frac{N}{M-1} \right\rfloor, \qquad (36)$$

where $\lfloor X \rfloor$ denotes an operator that truncates X to the nearest integers toward zero. In case the number of wavelet coefficients at the J_p scale is larger than M, we allow an additional level of wavelet decomposition by boundary extension of the data so that the number of wavelet coefficients at the coarsest level always becomes M. This constraint allows us to use the same form of the modified universal prior of integers regardless of the length of the NIRS time series. More specifically, if the total number of data elongations is $N_{ext} = (2M - m_{J_p}) \times 2^{J_p}$, where $m_{J_p} = \lfloor N \times 2^{-J_p} \rfloor$ denotes the number of coefficients in the J_p'th scale, then we have $m_{(J_p+1)} = \lfloor N \times 2^{-(J_p+1)} \rfloor = M$. Therefore, the maximum decomposition level of the NIRS signal becomes (J_p+1), and the number of wavelet coefficients at the finest scale is M. For the specific boundary extension, we employ a symmetric extension.[40]

In order to implement wavelet detrending, the wavelet should be compact. Daubechies wavelets and Symlets[22] with small orders are therefore reasonable candidates. However, due to the small vanishing moment of these wavelets, the trend estimate often violates the assumption of smoothness. Hence, we selected a CDF 9/7 biorthogonal filter[41] with a vanishing moment of 9, which gives a reasonable trade-off between the vanishing moment and the wavelet support. Note that CDF 9/7 is a standard wavelet filter in lossy compression of JPEG 2000 (Ref. 42), due to the compactness and sufficient vanishing moments.

5 Experimental Results

We performed NIRS experiments using Oxymon Mk III (Artinis, Netherlands), which has eight laser diodes and four detectors. In this system, two continuous wave lights (856 nm and 781 nm) are emitted at each source fiber. A suitable arrangement using 24 pairs of a source and a detector are illustrated in Fig. 2(a). Sources and detectors were attached by optical fibers on the scalp covering the motor cortex, as illustrated in Figs. 2(b) and 2(c). The distance between the source

Fig. 3 (a) A synthetic hemodynamic response (black line) and a noise added signal (gray line). (b) The overall simulated signal (gray line) and a ground-truth trend. Trend estimates using (c) Wavelet-MDL, (d) FIR with cutoff frequency 0.02 Hz, (e) FIR with cutoff frequency 0.015 Hz, and (f) DCT with cutoff frequency 0.015 Hz. Wavelet-MDL gives a closer estimate for the unknown trend.

and detector was 3.5 cm. A 3.0T MRI scanner (ISOL, Korea) was also used to simultaneously measure the BOLD signal. The echo planar imaging (EPI) sequence was used with TR/TE=3000/35 ms, flip angle=80 deg, 35 slices, and 4-mm slice thickness. In the subsequent anatomical scanning session, T1-weighted structural images were acquired using the same system.

To evaluate the proposed detrending algorithm, nine healthy, right-handed male adults were examined using a right-finger tapping (RFT-1) task. The 21-s periods of activation were alternated with 30-s periods of rest. During the activation periods, subjects were instructed to tap right-hand fingers. Total recoding time was 552 s. In addition, one subject was examined under the same conditions; however, in this time, the subject did not receive any stimulus during the recording period, which we call the baseline. This baseline data is used for a subsequent simulation study. No subject had any history of neurological disorders. All subjects were informed about the whole experiment process. The investigation was approved by the Institutional Review Board of Korea Advanced Institute of Science and Technology (KAIST).

5.1 Simulation Study

For comparison, two detrending methods based on the conventional filtering[8] were implemented. First, finite impulse response (FIR) filters were designed using the Kaiser window method, whose cutoff frequencies were appropriately adjusted for each case.[43] In addition, we also compared our method with DCT-based filtering, a standard detrending technique in SPM.[15] DCT coefficients that are below the cutoff frequency were declared as trend estimates. Note that we utilized the symmetric extension at both ends of the NIRS data to reduce boundary distortion during filtering.

A simulated NIRS time series was constructed to have a global bias, a simulated task-related time series, and AR(1) noise (auto-regressive noise of order 1). Note that the autoregressive (order 1) plus white noise model is a standard model for serial correlations in fMRI-SPM (Refs. 15 and 44) from the empirical perspectives, and the AR(1) model has successfully explained the short-range correlations.[44] Hence, this simulation also employs the AR(1) model to account for the short-range correlations in NIRS time series. The global bias was extracted from the real NIRS experiment under the baseline condition using a low-pass filter whose stopband edge was 0.02 Hz. The hemodynamic response was modeled by the canonical HRF with its derivatives. The extracted trend was normalized to [0,2], and the magnitude of the hemodynamic response was set to 1.05. The AR(1) noise ϵ was generated by $\epsilon[m]=\alpha \cdot \epsilon[m-1]+e[m]$, where α was set to 0.8, and e is the vector whose elements are zero mean white Gaussian variables with variance $\sigma^2=3.6\times10^{-3}$. The noise covariance matrix is then given by:

$$[\boldsymbol{\Sigma}]_{m,m+k}=E\{\boldsymbol{\epsilon}[m]\boldsymbol{\epsilon}[m+k]\}=\frac{\sigma^2}{1-\alpha^2}\alpha^{|k|}.$$

This covariance matrix was used in calculating t-scores in Eq. (11).

Table 1 t-scores with distinct detrending methods.

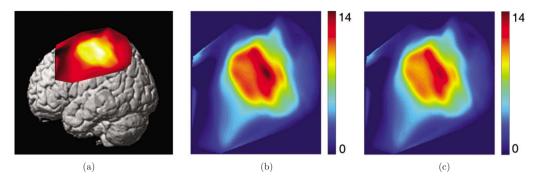|	Wavelet-MDL	Kaiser-A	Kaiser-B	DCT
t-score	73.22	16.89	47.57	45.73

Fig. 4 Deoxyhemoglobin *t*-maps for the RFT experiment. (a) Result from NIRS-SPM. Magnified *t*-maps using (b) Wavelet-MDL and (c) the conventional method. The wavelet-MDL detrending method provides a higher *t*-value centered at the target area.

In Figs. 3(a) and 3(b), the simulated NIRS time series and its components are illustrated. Black lines in Figs. 3(c), 3(d), and 3(f) show global trend estimates using various detrending methods. The same cutoff frequency for simulating the ground-truth trend was used in Fig. 3(d), which represents the case where the frequency content of a global trend is exactly known. In Fig. 3(e), a FIR filter with a stopband edge of 0.015 Hz is used. Figure 3(f) is the result of DCT-based filtering whose cutoff frequency is 0.015 Hz. Even in the case where the exact cutoff frequency is known, as illustrated in Fig. 3(d), it was hard to distinguish between the hemodynamic signal and the global trend when both signals have the same frequency contents. Indeed, the hemodynamic signal is an extremely low frequency signal considering the sampling frequency of NIRS, and hence it is sometimes similar to the global bias in its frequency content. We can easily see that the Wavelet-MDL-based detrending method (Wavelet-MDL) gives a much closer estimate for the unknown bias, as illustrated in Fig. 3(c). For a quantitative analysis, Table 1 shows the *t*-score using Eq. (11), which confirms that the proposed detrending method shows the best performance in estimating the hemodynamics signals.

Recall that the α value in the AR(1) model denotes the degree of correlation between the neighborhood samples. Even though this paper chose $\alpha=0.8$ to represent relatively strong correlation between the neighboring samples, similar behaviors have been obtained for a wide range of α values, and our Wavelet-MDL framework has been observed to outperform the other methods for all cases.

5.2 Experimental NIRS Data

We now apply our algorithm to real NIRS measurements. For calculating *t*-values, we used NIRS-SPM software, which has been developed by our group.[11] NIRS-SPM allows the estimation of the temporal correlation and determines a *p*-value using the tube formula. For a detailed discussion on the estimation of the temporal correlation, *p*-value, and related threshold value, readers can refer to our previous work on this issue.[11] In NIRS-SPM, the conventional filtering based on DCT is implemented. We set the cutoff frequency as 1/60 Hz, which is below the frequency of RFT paradigm repetition frequency of 0.018 Hz.

In Fig. 4(a), the deoxy-hemoglobin (HbR) *t*-map from NIRS-SPM using Wavelet-MDL is illustrated. For clear comparison, the magnified figure is illustrated in Fig. 4(b), whereas Fig. 4(c) shows the result of the conventional DCT filtering approach. The temporal covariance was estimated using a prewhitening method.[11] While the task-related activation is observable in both cases, Wavelet-MDL provides a higher *t*-value centered at the target area. The maximum *t*-values were 13.95 and 13.16 for Wavelet-MDL and the DCT-based filtering, respectively. The higher *t*-values indicate that the proposed method provides a statistically more significant estimate of the activation map.

An oxyhemoglobin time series and trend estimates are illustrated in Fig. 5. Since the design matrix is explicitly incorporated during the detrending process of Wavelet-MDL, as described in Eq. (20), the oxyhemoglobin signals were almost
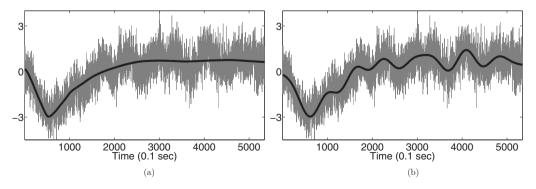

Fig. 5 Oxyhemoglobin measurement during the RFT experiment and estimated trend with (a) Wavelet-MDL and (b) the conventional method. The task-related signals are not removed for the case of Wavelet-MDL.
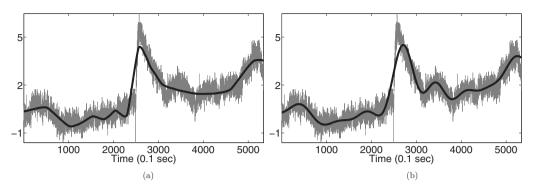
Fig. 6 Oxyhemoglobin measurement during the RFT experiment and estimated trend with (a) Wavelet-MDL and (b) the conventional method. Wavelet-MDL is capable of removing fast-varying trends without damaging task-related signals.

conserved, as shown in Fig. 5(a). However, in the case of the conventional approach, some components of oxyhemoglobin signals were classified as global trends, as illustrated in Fig. 5(b). Figure 6 shows another oxyhemoglobin signal that contains a rapidly varying bias. If we use the conventional detrending method, the residual of the fast-varying transient trend can be classified as an oxyhemoglobin signal. However, Wavelet-MDL can estimate the unknown trend even if it has high-frequency transient signal, since it is capable of capturing the transient signal and adjusting the complexity of a global trend automatically, as illustrated in Fig 6(a). However, in Fig. 6(b), the fast varying trend was not fully captured by the conventional approach, and the hemodynamic signal was damaged.

Last, we compared MDL with AIC_C, SIC, and Saito's MDL. Figures 7(b)–7(d) show that the estimated trends based on AIC_C, SIC, and Saito's MDL result in overfitting, whereas our algorithm correctly captures the trends. This implies that the modified universal prior of the integer is effective in model order selection.

5.3 Group Analysis

Group analyses were performed using NIRS-SPM software. NIRS data as well as simultaneously recorded fMRI data from nine subjects were used. Figure 8 shows group activation maps for the HbR signal at $p = 0.05$, which is overlayed on the fMRI activation map at $p = 0.005$. The Wavelet-MDL detrending method gives more specific localization of the target motor cortex compared to the conventional DCT-based high-pass filtering. To quantify the accuracy of localization, we calculate the receiver operating characteristic (ROC) for the spatial correlation of the NIRS activation map with that of fMRI BOLD. An ROC curve is a graphical plot of "sensitivity" versus "1-specificity" for a binary classifier. In ROC, (0,1) is the point of an ideal classifier, and the diagonal line represents the performance of random guessing. Therefore, the area under the ROC curve is a good indicator of the performance of a classifier. We evaluated our detrending method by measuring the spatial correlation between NIRS activation maps and the fMRI BOLD map. The BOLD map with $p = 0.005$ was
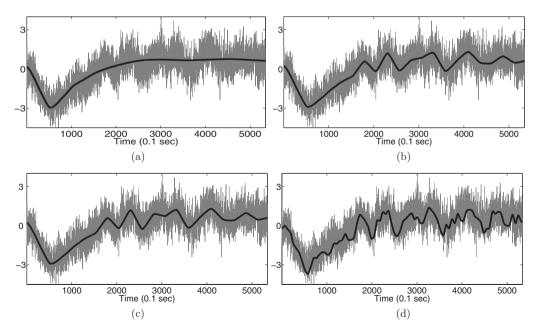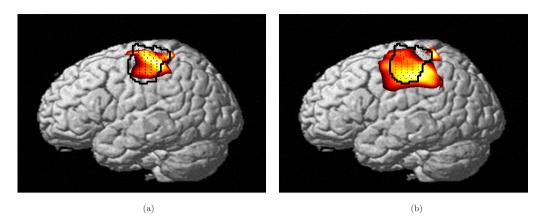

Fig. 7 Trend estimate using (a) the proposed MDL, (b) Saito's MDL, (c) SIC, and (d) AIC_C. Compared with other model order criteria, the proposed method gives the most accurate estimate.

(a)   (b)

Fig. 8 NIRS-SPM group analysis result at $p=0.05$. (a) Wavelet-MDL detrending and (b) DCT based detrending. Wavelet-MDL provides a more specific activation map. The mesh image corresponds to the fMRI activation map at $p=0.005$.

assumed as the ground truth in Fig. 9. The sensitivity and specificity were calculated by changing the threshold values. The ROC analysis in Fig. 9 shows that the proposed detrending algorithm outperforms the conventional high-pass filtering since the area under the ROC curve is largest for Wavelet-MDL.

6 Discussion

6.1 Temporal Covariance in MDL

In Eq. (26), we assume that the noise is uncorrelated, independent, and normally distributed. For the correlated noise, the covariance matrix Σ should be considered in Eqs. (26)–(28). Specifically, consider the negative log-likelihood for measurement in the wavelet domain. Let $\mathbf{r}(n_0)$ denote the residual vector for model order n_0 in the wavelet domain, i.e., $\mathbf{r}(n_0)=W(\mathbf{y}-X\boldsymbol{\beta}-\boldsymbol{\theta})$. Since the dependency for $\mathbf{r}$ on the model order n_0 is clear in the context, we omit the letter n_0 in the sequel. If we model the correlated noise using the fractional Brownian motion process[26] as in the fMRI case,[16] the pdf of measurement $\mathbf{y}$ in the wavelet domain is given by

$$p(W\mathbf{y}|n_0) = \frac{1}{(2\pi)^{2N}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}r^T\Sigma^{-1}r\right), \quad (37)$$

where Σ denotes a diagonal covariance matrix described in Eq. (22). If we substitute Eq. (22) for Σ in Eq. (37), the negative log-likelihood for measurement is then given by

$$-\log p(W\mathbf{y}|n_0) = \frac{1}{2}\sum_{j=1}^{J} [N_j \log \sigma_j^2(n_0) + \sigma_j^{-2}(n_0)r_j^T r_j] + c, \quad (38)$$

where $N_j=2^{-j}N$ is the number of wavelet coefficients in the j'th scale, $\mathbf{r}_j$ is the subvector of $\mathbf{r}$ corresponding to the j'th scale, and c is a constant not related to model order selection. The maximum likelihood (ML) estimator for level-dependent variance $\sigma_j^2(n_0)$ is easily obtained by a straightforward calculation as follows:

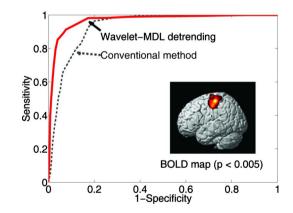$$\hat{\sigma}_j^2(n_0) = \frac{1}{N_j}r_j^T r_j, \quad (39)$$

which is identical with the ML estimates of the residual variance in the j'th scale. If we substitute Eq. (39) for $\sigma_j^2(n_0)$ in Eq. (38):

$$-\log p(W\mathbf{y}|n_0) = \sum_{j=1}^{J} \frac{N_j}{2} \log \hat{\sigma}_j^2(n_0) + c'', \quad (40)$$

where c'' is a constant not related to model order selection. Now, we employ an exponentially decaying variance model along decomposition levels that is popularly used for modeling the correlated noise in the wavelet domain:[24,45,46]

$$\sigma_j^2(n_0) = \tau^2(n_0)2^{\alpha j}, \quad (41)$$

where α is the spectral decay rate of the correlated noise model and τ denotes a constant that reflects the overall noise

Fig. 9 ROC curves for activation maps using Wavelet-MDL and DCT-based detrending. The fMRI activation map at $p=0.005$ shown as an inset is used as a ground truth, and the sensitivity and the specificity were calculated by changing the threshold value. The ROC analysis shows that the area under the ROC curve for Wavelet-MDL was largest, indicating that the proposed algorithm outperforms the conventional method.

variance. We can estimated τ in the zeroth, decomposition level, i.e., in the time domain, as follows:

$$\hat{\tau}^2(n_0) = \hat{\sigma}_0^2(n_0) = \frac{1}{N}\|\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{\theta}(n_0)\|^2. \qquad (42)$$

By substituting Eq. (41) and Eq. (42) for $\hat{\sigma}_j^2(n_0)$ in Eq. (40), we obtain

$$-\log p(\boldsymbol{Wy}|n_0) = \frac{N}{2}\log \hat{\sigma}_0^2(n_0) + c''', \qquad (43)$$

where c''' is a constant not related to model order selection. Hence, Eq. (43) is identical to Eq. (28).

6.2 *t-Value versus Activation Map*

Recall that *t*-value is often used to quantify the performance of the detrending algorithms.[16] However, more correct quantification should be based on the accuracy of the activation map at the same *p*-value. For example, consider a case where a fixed *p*-value provides an identical threshold value. Then, in obtaining the activation maps, the overall increase in *t*-values may imply that the resulting activation map becomes significantly larger. Hence, to guarantee the accurate localization, the higher contrast of *t*-value between activated and background region is more important. In this perspective, our group analysis result confirms that wavelet-MDL detrending is more accurate in localizing the activation maps.

6.3 *Precoloring versus Prewhitening*

In Ref. 11, we have proposed two different methods to remove the temporal covariance: precoloring and prewhitening. Precoloring tries to remove the temporal correlation by filtering with HRF filter, whereas prewhitening attempts to estimate $\boldsymbol{\Sigma}$ based on the assumption that the noise is the AR(1) process. In this perspective, the prewhitening provides more accurate estimation of $\boldsymbol{\Sigma}$ as long as the AR(1) assumption is correct. The *t*-value in Fig. 4 was hence calculated using prewhitening. However, as demonstrated in Ref. 11, precoloring was more robust in removing the temporal correlation, and the final activation maps from precoloring were observed to be better due to the limited number of measurements compared to its fMRI counterpart.[11] Hence, this paper calculates the group activation map using the precoloring method, as in. Ref. 11.

7 Conclusion

We developed a Wavelet-MDL-based detrending method that is robust under motion and physiological variation. In Wavelet-MDL detrending, a wavelet transform is applied to the NIRS time series to decompose it into bias, hemodynamic signal, and noise components in distinct scales. With the MDL criterion using a modified universal prior for the integer, the optimal model order could be easily estimated, and the unknown drift signal in NIRS data was successfully removed. Experimental results demonstrated that the new detrending algorithm outperforms the conventional approaches.

References

1. F. F. Jobsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science* **198**, 1264–1267 (1977).
2. S. G. Diamond, T. J. Huppert, V. Kolehmainen, M. A. Franceschini, J. P. Kaipio, S. R. Arridge, and D. A. Boas, "Dynamic physiological modeling for functional diffuse optical tomography," *Neuroimage* **30**(1), 88–101 (2006).
3. Y. Hoshi, "Functional near-infrared optical imaging: utility and limitations in human brain mapping," *Psychophysiology* **40**, 511–520 (2003).
4. M. Cope and D. T. Delpy, "System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination," *Med. Biol. Eng. Comput.* **26**, 289–294 (1988).
5. M. Essenpreis, M. Cope, C. E. Elwell, S. R. Arridge, P. van der Zee, and D. T. Delpy, "Wavelength dependence of the differential pathlength factor and the log slope in time-resolved tissue spectroscopy," *Adv. Exp. Med. Biol.* **333**, 9–20 (1993).
6. A. Duncan, J. H. Meek, M. Clemence, C. E. Elwell, L. Tyszczuk, M. Cope, and D. T. Delpy, "Optical pathlength measurements on adult head, calf, and forearm and the head of the newborn infant using phase resolved optical spectroscopy," *Phys. Med. Biol.* **40**, 295–304 (1995).
7. H. Zhao, Y. Tanikawa, F. Gao, Y. Onodera, A. Sassaroli, K. Tanaka, and Y. Yamada, "Maps of optical differential pathlength factor of human adult forehead, somatosensory motor, and occipital regions at multi-wavelengths in NIR," *Phys. Med. Biol.* **47**(12), 2075–2093 (2002).
8. M. L. Schroeter, M. M. Bucheler, K. Muller, K. Uludag, H. Obrig, G. Lohmann, M. Tittgemeyer, A. Villringer, and D. Yves von Cramon, "Towards a standard analysis for functional near-infrared imaging," *Neuroimage* **21**, 283–290 (2004).
9. M. M. Plichta, S. Heinzel, A. C. Ehlis, P. Pauli, and A. J. Fallgatter, "Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data: a parametric validation study," *Neuroimage* **35**, 625–634 (2007).
10. P. H. Koh, D. E. Glaser, G. Flandin, S. Kiebel, B. Butterworth, A. Maki, D. T. Delpy, and C. E. Elwell, "Functional optical signal analysis: a software tool for near-infrared spectroscopy data processing incorporating statistical parametric mapping," *J. Biomed. Opt.* **12**, 1–13 (2007).
11. J. C. Ye, S. Tak, K. E. Jang, J. Jung, and J. Jang, "NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy," *Neuroimage* **44**(2), 428–447, (2009).
12. K. J. Worsley and K. J. Friston, "Analysis of fMRI time-series: revisited again," *Neuroimage* **2**, 173–181 (1995).
13. J. Sun, "Tail probabilities of the maxima of Gaussian random fields," *Ann. Probab.* **21**(1), 34–71 (1993).
14. J. Cao and K. J. Worsley, "The geometry of the Hotellings random field with applications to the detection of shape changes," *Ann. Stat.* **27**(3), 925–942 (1999).
15. K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, Academic Press, San Diego, CA (2006).
16. F. G. Meyer, "Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series," *IEEE Trans. Med. Imaging* **22**(3), 315–322 (2003).
17. J. Rissanen, "Modeling by shortest data description," *Automatica* **14**(5), 465–471 (1978).
18. K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Hum. Brain Mapp* **2**(4), 189–210 (1995).
19. R. W. Cox, "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages," *Comput. Biomed. Res.* **29**, 162–173 (1996).

20. G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger, "Linear systems analysis of functional magnetic resonance imaging in human V1," *J. Neurosci.* **16**, 4207–4221 (1996).
21. R. N. A. Henson, C. J. Price, M. D. Rugg, R. Turner, and K. J. Friston, "Detecting latency differences in event-related BOLD responses: application to words versus nonwords and initial versus repeated face presentations," *Neuroimage* **15**, 83–97 (2002).
22. S. G. Mallat, *A Wavelet Tour of Signal Processing*, Academic, New York (1999).
23. I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. R. Stat. Soc. Ser. B (Methodol.)* **59**(2), 319–351 (1997).
24. P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Inf. Theory* **38**(2), 910–917 (1992).
25. D. Ville, M. L. Seghier, F. Lazeyras, T. Blu, and M. Unser, "WSPM: wavelet-based statistical parametric mapping," *Neuroimage* **37**(4), 1205–1217 (2007).
26. B. B. Mandelbrot and J. W. Ness, "Fractional Brownian motions, fractional noises, and applications," *SIAM Rev.* **10**(4), 422–437 (1968).
27. D. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika* **81**(3), 425–455 (1994).
28. N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion," *Wavelets Geophys.* 299–324 (1994).
29. J. Rissanen, "MDL denoising," *IEEE Trans. Inf. Theory* **46**(7), 2537–2543 (2000).
30. K. J. Friston, P. Fletcher, O. Josephs, A. Holmes, M. D. Rugg, and R. Turner, "Event-related fMRI: characterizing differential responses," *Neuroimage* **7**, 30–40 (1998).
31. K. L. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika* **76**(2), 297–307 (1989).
32. G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.* **6**, 461–464 (1978).
33. H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control* **AC-19**(6), 716–723 (1974).
34. S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.* **22**(1), 79–86 (1951).
35. P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.* **21**, 36–47 (2004).
36. A. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inf. Transm.* **1**, 4–7 (1965).
37. T. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Inc., New York (1991).
38. M. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Am. Stat. Assoc.* **96**(454), 746–774 (2001).
39. J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Stat.* **11**(2), 417–431 (1983).
40. M. Unser, "A practical guide to the implementation of the wavelet transform," chapter 2 in A. Aldroubi and M. Unser, Eds., *Wavelets in Medicine and Biology*, pp. 37–73, CRC, Boca Raton, FL (1996).
41. A. Cohen, I. Daubeches, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.* **45**(5), 485–560 (1992).
42. B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: foundations of JPEG 2000," *IEEE Signal Process. Mag.* **18**(5), 22–35 (2001).
43. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Upper Saddle River, NJ (1989).
44. P. L. Purdon and R. M. Weisskoff, "Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI," *Hum. Brain Mapp* **6**(4), 239—249 (1998).
45. E. Bullmore, C. Long, J. Suckling, J. Fadili, G. Calvert, F. Zelaya, T. A. Carpenter, and M. Brammer, "Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains," *Hum. Brain Mapp* **12**(2), 61–78 (2001).
46. J. M. Fadili and E. Bullmore, "Penalized partially linear models using sparse representations with an application to fMRI time series," *IEEE Trans. Signal Process.* **53**(9), 3436–3448 (2005).