frontiers in Neuroscience

METHODS
published: xx October 2015
doi: 10.3389/fnins.2015.00375

Detecting the subtle shape differences in hemodynamic responses at the group level

Gang Chen[1]*, *Ziad S. Saad*[1], *Nancy E. Adleman*[2], *Ellen Leibenluft*[3] and *Robert W. Cox*[1]

[1] Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA, [2] Department of Psychology, The Catholic University of America, Washington, DC, USA, [3] Section on Bipolar Spectrum Disorders, Emotion and Development Branch, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA

The nature of the hemodynamic response (HDR) is still not fully understood due to the multifaceted processes involved. Aside from the overall amplitude, the response may vary across cognitive states, tasks, brain regions, and subjects with respect to characteristics such as rise and fall speed, peak duration, undershoot shape, and overall duration. Here we demonstrate that the fixed-shape (FSM) or adjusted-shape (ASM) methods may fail to detect some shape subtleties (e.g., speed of rise or recovery, or undershoot). In contrast, the estimated-shape method (ESM) through multiple basis functions can provide the opportunity to identify some subtle shape differences and achieve higher statistical power at both individual and group levels. Previously, some dimension reduction approaches focused on the peak magnitude, or made inferences based on the area under the curve (AUC) or interaction, which can lead to potential misidentifications. By adopting a generic framework of multivariate modeling (MVM), we showcase a hybrid approach that is validated by simulations and real data. With the whole HDR shape integrity maintained as input at the group level, the approach allows the investigator to substantiate these more nuanced effects through the unique HDR shape features. Unlike the few analyses that were limited to main effect, two- or three-way interactions, we extend the modeling approach to an inclusive platform that is more adaptable than the conventional GLM. With multiple effect estimates from ESM for each condition, linear mixed-effects (LME) modeling should be used at the group level when there is only one group of subjects without any other explanatory variables. Under other situations, an approximate approach through dimension reduction within the MVM framework can be adopted to achieve a practical equipoise among representation, false positive control, statistical power, and modeling flexibility. The associated program `3dMVM` is publicly available as part of the AFNI suite.

Keywords: hemodynamic response, basis function, multivariate general linear model, linear mixed-effects model, FMRI group analysis, AFNI

OPEN ACCESS

Edited by:
Brian Caffo,
Johns Hopkins University, USA

Reviewed by:
Xi-Nian Zuo,
Chinese Academy of Sciences, China
Théodore Papadopoulo,
French Institute for Research in Computer Science and Automation, France

***Correspondence:**
Gang Chen
gangchen@mail.nih.gov

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *22 May 2015*
Accepted: *28 September 2015*
Published: *xx October 2015*

Citation:
Chen G, Saad ZS, Adleman NE, Leibenluft E and Cox RW (2015) Detecting the subtle shape differences in hemodynamic responses at the group level. Front. Neurosci. 9:375. doi: 10.3389/fnins.2015.00375

INTRODUCTION

When a region in the brain is activated, oxygen and glucose demands lead to blood vessel dilation, followed by increased blood to the tissue (neurons and astrocytes) under stress. The onset of a neuronal activity triggers a sequence of physiological events in the blood vessels of the surrounding area, typically characterized by the changes in cerebral blood flow as well as concentration fluctuations of deoxyhemoglobin and oxyhemoglobin. The blood oxygenation level dependent (BOLD) signal from the FMRI scanning mainly captures the concentration changes of deoxyhemoglobin; that is, the BOLD signal is a surrogate and signature of neuronal activations plus various sources of noise (e.g., physiological and random fluctuations). As an indirect measure of neuronal activity, the shape of the BOLD response may hold some crucial features about brain function. However, the cascade of events from neural activation to measurable MRI signal is complex and nonlinear under certain regimes (Friston et al., 1998b; Birn et al., 2001; Logothetis and Wandell, 2004; Logothetis, 2008; Magri et al., 2012): Even though the BOLD response is simply interpreted as changes in neuronal processing, the same neuronal activity may evoke different hemodynamic response (HDR) shape across trials, regions, conditions/tasks, subjects, and groups. For example, neurophysiological confounds such as neurovascular coupling or energy consumption changes could lead to different BOLD response features, potentially explaining the HDR variability in magnitude and shape across brain regions, cognitive conditions and populations (e.g., children with autism vs. controls, Reynell and Harris, 2013). Nevertheless, meaningful interpretation as well as detection power in FMRI data analysis may depend on the accurate modeling of the BOLD response both at the individual subject and group levels (e.g., Buxton et al., 2004; Handwerker et al., 2004; Stephen et al., 2007; Barbé et al., 2012; Badillo et al., 2013).

Under an experimentally-manipulated situation, the subject typically performs some tasks or is put under certain conditions in an event-related design, with each trial lasting for 2 s or less, and the HDR to each trial can be mathematically characterized by an impulse response function (IRF) that corresponds to a stimulus with a theoretically instantaneous duration and unit intensity. The voxel-wise EPI signal is then modeled through time series regression with explanatory variables (or regressors) of interest, each of which is constructed through the convolution between the stimulus timing and the IRF. In a block design, each task or condition has a duration of more than two seconds. As each block can be approximately considered as a sequence of events with an interval of scanning repetition time (TR), the theoretical HDR is usually hypothesized as the integral or linear summation of the consecutive IRFs, or the convolution of IRF over the stimulus duration.

We typically adopt some formative mathematical functions (usually called HDR functions or HRFs) to approximate the HDR based on the experimental data with the assumption of linearity and time-invariance (or stationarity) (Marrelec et al., 2003), and consider three common approaches to modeling the average HDR across trials. The first one presumes a fixed shape IRF (e.g., gamma variate or wave form in AFNI, Cohen, 1997; canonical IRF in SPM, FSL, and NIPY, Friston et al., 1998a). With this model-based or fixed-shape method (FSM), the regression coefficient or β associated with each condition in the individual subject analysis reflects the major HDR magnitude (e.g., percent signal change). The second approach makes no assumption about the IRF's shape and estimates it with a set of basis functions. The number of basis functions varies depending on the kernel set and the duration over which the response is being modeled. A common approach to this estimated-shape method (ESM) consists of using a set of equally-spaced TENT (piecewise linear) functions or linear splines, and each of the resulting regression coefficient represents an estimate of the response amplitude at some time after stimulus onset. Regardless of the kernel set, however, ESM generates the same number of regressors as the number of basis functions (e.g., m) per condition or task, resulting in m regression coefficients which need to be considered simultaneously at the group level. In addition to the aforementioned TENT basis set, options for ESM at the voxel level include cubic splines, Legendre polynomials, sines, or user-defined functions in AFNI, and finite impulse function (FIR) in SPM, FSL, and NIPY, inverse logit (Lindquist et al., 2009), and high-order B-splines (Degras and Lindquist, 2014). In addition, the python package PyHRF offers an ESM at the parcel level through the joint detection-estimation framework (Vincent et al., 2014). It is of note that one significant advantage of adopting basis functions such as TENT or cubic splines is the flexibility of creating regressors through piecewise interpolation when the stimulus onset times are not aligned with the TR grids (e.g., the acquisition time is shorter than TR if one wants to present "silent trials" as a control condition to speech or other auditory stimulus). The third approach lies between the two extremes of FSM and ESM, and uses a set of two or three basis functions (Friston et al., 1998b). In this adjusted-shape method (ASM), the first basis (canonical IRF) captures the major HDR shape, and the second basis, the time derivative of the canonical IRF, provides some flexibility in modeling the delay or time-to-peak, while the third basis, dispersion curve (derivative relative to the dispersion parameter in the canonical IRF), allows the peak duration to vary.

With one parameter per condition, FSM is the most efficient[1] and statistically powerful among the three, if the presumed shape is reasonably close to the ground truth, and the group analysis strategies have been developed to reasonable maturity: The β values at the individual level are typically brought to the group level using the Student's t-test, permutation tests (Nichols and Holmes, 2002; Dehaene-Lambertz et al., 2006; Mériaux et al., 2006; Winkler et al., 2014), AN(C)OVA, general linear model (GLM) (Poline and Brett, 2012), multivariate modeling (MVM) (Chen et al., 2014), linear mixed-effects (LME) method (Bernal-Rusiel et al., 2013; Chen et al., 2013), or mixed-effect multilevel analysis (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012), with the assumption that each effect estimate is equally reliable across all subjects. However, deviations of the HDR from the presumed shape would result in biased estimates of the

[1] The efficiency in the statistics context measures the optimality of a testing method. A more efficient test requires a smaller sample size to attain a fixed power level.

amplitude, in addition to failing to capture differences in shape such as during the undershoot or recovery phase. ESM is the most flexible among the three methods in terms of providing a more accurate characterization of the BOLD response and can achieve higher activation detection power in individuals. In addition, the estimated HDR curve with a unique signature shape offers much stronger support for the existence of activation than a single scaling factor or β value with FSM or ASM. Compared with FSM, ASM also results in a less biased response amplitude for the principal kernel, and can account for more variance compared to FSM; however, the common practice of using only the principal kernel's coefficient at the group level will not allow the detection of shape changes between conditions and or groups when those exist.

Difficulties with using ESM (and to a lesser degree ASM) include the need for a larger number of kernel coefficients that need to be estimated. They requires m times more regressors than FSM in the individual subject analysis, which translates to more data points and scanning time to reach similar statistical power in individuals. Secondly, the risk of over-fitting exists when some confounding effects such as head motion and physiological noise are stimulus-locked and not fully accounted for. Lastly, the most challenging step lies at the group level: How to simultaneously handle those m effect estimates? And how to summarize and interpret the results? To avoid the complexity involved in the multiple effect estimates from ESM or ASM, the popular approach at the group level is dimensional reduction, condensing the shape information over the multiple values into one number. For ESM, one method is to sum over all or a subset of effect estimates (e.g., ignoring a few time points at the beginning and the end) to obtain the area under the curve (AUC) (e.g., Beauchamp et al., 2003; Greene et al., 2007; McGregor et al., 2013). As the BOLD response curve can be characterized by parameters such as amplitude (or height), delay (or time-to-peak), duration (or HWFM), another dimensional reduction proposal is to perform the group analysis on such a derived parameter from the estimated HDR (Lindquist et al., 2009; Degras and Lindquist, 2014). With two or three effect estimates per condition from ASM at the group level, the popular approach focuses on the β value of the canonical HDR while ignoring the parameters for the shape adjustments (i.e., the function of these other parameters is to absorb minor shape fluctuations that would otherwise be modeled as "noise"). One alternative is to estimate the HDR height using the Euclidean or L^2-norm distance (L2D) of the two or three effect estimates (Calhoun et al., 2004; Lindquist et al., 2009; Steffener et al., 2010). Essentially, these dimensional reduction methods transform the effect estimates in an k-dimensional space $\mathbb{R}^k$ to one-dimensional $\mathbb{R}^1$. As information loss is unavoidable in the process, statistical power in activation identification would suffer. This raises the question of whether a more preferable approach to significance testing might better exploit the information in the HDR shape at the group level.

A Motivational Example

To demonstrate and compare various modeling approaches at the group level, we adopt the same experimental data used in our previous paper (Chen et al., 2014), with a typical group design that accounts for a confounding effect: varying age across subjects. Briefly, the experiment involved one between-subjects factor, group (two levels: 21 children and 29 adults) and one within-subject factor (two levels: congruent and incongruent conditions). Stimuli were large letters (either "H" or "S") composed of smaller letters ("H" or "S"). For half of the stimuli, the large letter and the component letters were congruent (e.g., "H" composed of "H"s) and for half they were incongruent (e.g., "H" composed of "S"s). Parameters for the whole brain BOLD data on a 3.0 T scanner were: voxel size of 3.75 × 3.75 × 5.0 mm^3, 24 contiguously interleaved axial slices, and TR of 1250 ms (TE = 25 ms, FOV = 240 mm, flip angle = 35°). Six runs of EPI data were acquired from each subject, and each run lasted for 380 s with 304 data points. The task followed an event-related design with 96 trials in each run, with three runs of congruent stimuli interleaved with three runs of incongruent stimuli (order counterbalanced across subjects). Subjects used a two button box to identify the large letter during global runs and the component letter during local runs. Each trial lasted 2500 ms: the stimulus was presented for 200 ms, followed by a fixation point for 2300 ms. Inter-trial intervals were jittered with a varying number of TRs, allowing for a trial-by-trial analysis of how the subject's BOLD response varied with changes in reaction time (RT). The experiment protocol was approved by the Combined Neuroscience Institutional Review Board at the NIMH, and the National Clinical Trials Identifier is NCT00006177.

The EPI time series went through the following preprocessing steps: slice timing and head motion corrections, spatial alignment to a Talairach template (TT_N27) at a voxel size of 3.5 × 3.5 × 3.5 mm^3, smoothing with an isotropic FWHM of 6 mm, and scaling each voxel time series by its mean value. The scaling step during preprocessing enables one to interpret each regression coefficient of interest as an approximate estimate of percent signal change relative to the temporal mean. The six runs of data were concatenated for the individual regression analysis with the discontinuities across runs properly handled (Chen et al., 2012). To capture the subtle HDR shape under a condition, two modeling approaches were adopted, ESM and ASM, for model comparison. With ESM, each trial was modeled with 10 tent basis functions, each of which spanned one TR (or 1.25 s). The subject's RT at each trial was incorporated as a per-trial modulation variable. In other words, two effects per condition were estimated in the time series regression at the individual level: one revealed the response curve associated with the average RT while the other showed the marginal effect of RT (response amplitude change when RT increases by 1 s) at each time point subsequent to the stimulus. In addition, the following confounding effects were included in the model for each subject, for each run: third-order Legendre polynomials accounting for slow drifts, incorrect trials (misses), censored time points with extreme head motion, and the six head motion parameters. The modeling strategy remained the same with ASM except that the three SPM basis functions (canonical IRF plus time and dispersion derivatives) were employed to model the BOLD responses instead of the 10 tents.

At the group level, it is the BOLD effects associated with the average RT that are of interest here. In addition

to the estimated HDR profiles, three other explanatory variables considered are: a) between-subjects factor, Group (two levels: children and adults), b) within-subject factors, Condition (two levels: congruent and incongruent), and c) quantitative covariate, age. The focus is on the interaction of HDR between Group and Condition: Do the two groups differ in the HDR profile contrast between the two conditions?

Preview

This paper is a sequel to our previous exploration (Chen et al., 2014) of the multivariate modeling (MVM) approach for FMRI group analysis. The layout is as follows. First, we explore and review various hypothesis testing strategies at the group level when the HDR is estimated through multiple basis functions. Second, simulation data were generated to reveal how each methodology performs in terms of controllability for false positives and false negatives, and the performance of these methods was assessed when they were applied to the experimental dataset at both individual and group levels. Finally, we compare all the modeling methodologies for ASM and ESM as well as with and without dimension reduction. The modeling strategies and testing methods discussed here are all performed at the voxel level. Multiple testing correction can be applied in the conventional fashion by controlling the false positive rate (Benjamini and Hochberg, 1995) or the family-wise error through Monte Carlo simulations (`3dClustSim` in AFNI, Forman et al., 1995) or random field theory (Worsley et al., 1992).

Our major contribution here is to demonstrate the importance of accounting for shape differences and to offer testing approaches at the group level within an MVM platform with the modeling flexibility that would not be available under the conventional GLM. Through our demonstration we propose that ESM should be adopted whenever appropriate or possible to identify the nuanced differences in HDR shape that would be difficult or unlikely to be revealed through FSM or ASM. Furthermore, we recommend that the investigator report the effect estimates such as the HDR curves to substantiate the results in addition to the statistical significance. The modeling framework and functionality are available in the program `3dMVM` for public use in the AFNI suite (Cox, 1996).

Throughout this article, regular italic letters (e.g., α) stand for scalars, boldfaced italic letters in lower ($\boldsymbol{a}$) and upper ($\boldsymbol{X}$) cases for column vectors and matrices respectively. The word *multivariate* is used here in the sense of treating the effect estimates from the same subject or from the levels of a within-subject factor as the instantiations of simultaneous response (or outcome) variables (e.g., the effect estimates for the HDR). This usage differs from the popular connotation in the FMRI field when the spatial structure (multiple voxels) is modeled as the simultaneous response variables, including such methods as multivariate pattern analysis (Haxby, 2012), independent component analysis, and machine learning methods such as support vector machines. Major acronyms used in the paper are listed in Appendix A.

METHODS

As shown in Chen et al. (2014), we formulate the group analysis under a multivariate GLM or MVM platform that is expressed from a subject-wise perspective, $\boldsymbol{\beta}_i^T = \boldsymbol{x}_i^T \boldsymbol{A} + \boldsymbol{\delta}_i^T$, or through the variable-wise pivot, $\boldsymbol{b}_j = \boldsymbol{X}\boldsymbol{a}_j + \boldsymbol{d}_j$, or in the following concise form,

$$\boldsymbol{B}_{n\times m} = \boldsymbol{X}_{n\times q}\, \boldsymbol{A}_{q\times m} + \boldsymbol{D}_{n\times m}. \quad (1)$$

The n rows of the response matrix $\boldsymbol{B} = (\beta_{ij})_{n\times m} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_n^T)^T = (\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_m)$ represent the data from the n subjects while the m columns correspond to the levels of within-subject factor(s). For example, the effect estimates from the multiple basis functions under ESM or ASM can be considered the response values associated with the levels of a within-subject or repeated-measures factor (termed Component hereafter). When multiple within-subject factors occur, all their level combinations for each subject are *flattened* from a multi-dimensional space onto a one-dimensional row of $\boldsymbol{B}$. It is noteworthy that the within-subject factors are expressed as columns in $\boldsymbol{B}$ on the left-hand side of the model (1), and only between-subjects variables such as subjects-grouping factors (e.g., sex, genotypes), subject-specific measures (e.g., age, IQ) and their interactions are treated as q explanatory variables on the right-hand side. The same linear system is assumed for all the m response variables, which share the same design matrix $\boldsymbol{X} = (x_{ih}) = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)^T$. Without loss of generality, $\boldsymbol{X}$ is assumed to have full column-rank q. Each column of the regression coefficient matrix $\boldsymbol{A} = (\alpha_{hj})$ corresponds to a response variable, and each row is associated with an explanatory variable. Lastly, the error matrix $\boldsymbol{D} = (\delta_{ij})_{n\times m} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, ..., \boldsymbol{\delta}_n)^T = (\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_m)$ is assumed nm-dimensional Gaussian: $vec(\boldsymbol{D}) \sim N(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$, where vec and $\otimes$ are column stacking and direct (or Kronecker) product operators respectively. As in univariate modeling (UVM), the assumptions for model (1) are linearity, Gaussianity and homogeneity of variance-covariance structure (same $\boldsymbol{\Sigma}$ across all the between-subjects effects). When only one group of subjects is involved ($q = 1$), the parameter matrix $\boldsymbol{A}$ becomes a row vector $(\alpha_1, \alpha_2, ..., \alpha_m)$ that is associated with the m levels of a within-subject factor.

As demonstrated in Chen et al. (2014), MVM has a few advantages over its univariate counterpart. When the data are essentially multidimensional like the multiple effect estimates from ESM or ASM, MVM has a crucial role in formulating hypothesis testing. In addition, it characterizes and quantifies the intercorrelations among the variables based on the data rather than a presumed variance-covariance structure as in UVM. Furthermore, MVM in general provides a better control for false positives than UVM. Lastly, the conventional univariate testing (UVT) under GLM can be easily performed under the MVM framework with a few extra advantages. Here we discuss one aspect by which the group analysis of neuroimaging data will benefit from the MVM facility when the HDR profile is estimated from multiple basis functions instead of being presumed to have a fixed shape. Then in the section Simulations and Real Experiment Results, we elaborate and compare a few testing alternatives in

terms of power and false positives, using simulations and in terms of performance with real data.

Different Testing Strategies

Here we exemplify two simple and prototypical cases with the HDR profile modeled by m basis functions at the individual subject level: a) one group of subjects with the associated effects at the group level expressed as $\alpha_1, \alpha_2, ..., \alpha_m$ under (1), and b) either two groups or two conditions and the two sets of effect estimates for HDR are α_{1j} and α_{2j} respectively, $j = 1, 2, ..., m$. To simplify geometric representations, we assume equal number of subjects across groups in the case of group comparison, but the assumption is not required from the modeling perspective. The various modeling strategies discussed below for these two cases can be easily extended to situations with more explanatory variables, including factors and quantitative covariates.

Multivariate Testing (MVT)

As the analogs of one- and two-sample or paired t-tests under UVT, the two prototypes can be expressed with the following null hypotheses,

$$H_{01}^{MVT} : \alpha_1 = 0, \alpha_2 = 0, ..., \alpha_m = 0, \quad (2a)$$

$$H_{02}^{MVT} : \alpha_{11} = \alpha_{21}, \alpha_{12} = \alpha_{22}, ..., \alpha_{1m} = \alpha_{2m}. \quad (2b)$$

In other words, the m regression coefficients associated with the m basis functions from each subject are brought to the group level and treated as the instantiated values of m simultaneous variables. When the effect estimates associated with the basis functions of ESM or ASM are treated as the values of m simultaneous response variables, the hypothesis (2a) or (2b) can be analyzed through MVT under the model (1). Geometrically, the data for H_{01}^{MVT} represent the group centroid $(\alpha_1, \alpha_2, ..., \alpha_m)$ in the m-dimensional real coordinate space $\mathbb{R}^m$ (**Table 1**), and the associated one-sample Hotelling T^2-test is performed to reveal whether the group centroid lies in the rejection region (outside of an m-dimensional ellipse centering around the origin in the case of H_{01}^{MVT}). Similarly, the data for H_{02}^{MVT} are expressed as two group centroids, $(\alpha_{11}, \alpha_{12}, ..., \alpha_{1m})$ and $(\alpha_{21}, \alpha_{22}, ..., \alpha_{2m})$, and the corresponding two-sample Hotelling T^2-test is conducted to see if the hypothesis (2b) about the two centroids can be rejected. The hypothesis (2b) can be easily generalized to the situation with more than two groups of subjects (e.g., three genotypes) as well as more than one subject-grouping variable (e.g., sex, genotypes, and handedness) through the formulation of general linear testing (Chen et al., 2014). One noteworthy feature of MVT is that it allows those simultaneous effects to have different scales or units, unlike the traditional AN(C)OVA or univariate GLM in which all the levels of a factor are usually of the same dimension.

Linear Mixed-effects Modeling (LME)

As demonstrated in Chen et al. (2013), linear mixed-effects modeling (LME) can be adopted for group analysis when the HDR is estimated through multiple basis functions. Specifically, the m regression coefficients from each subject associated with the m basis functions are modeled as values corresponding to m levels of a within-subject factor under the LME framework. When no other explanatory variables are present in the model, the LME methodology can be formulated by (2a) with an intercept of 0. That is, the m effects are coded by m indicator variables instead of any conventional contrast coding. Suppose that the m effect estimates associated with the m basis functions from the ith subject are $\beta_{i1}, \beta_{i2}, ..., \beta_{im}$, the LME model can be specified as,

$$\beta_{ij} = \alpha_j x_{ij} + \delta_i + \epsilon_{ij}, i = 1, 2, ..., n, j = 1, 2, ..., m.$$

where the random effect δ_i characterizes the deviation or shift of the ith subject's HDR from the overall group HDR, the residual term ϵ_{ij} indicates the deviation of each effect estimate β_{ij} from the ith subject's HDR, and the indicator variables x_{ij} take the cell mean coding,

$$x_{ij} = \begin{cases} 1, & \text{if } i\text{th subject is at } j\text{th level,} \\ 0, & \text{otherwise.} \end{cases}$$

so that the parameters $\alpha_j, j = 1, 2, ..., m$ capture the overall group HDR. The significance of the overall HDR at the group level can be tested through LME on the same hypothesis as (2a),

$$H_0^{LME} : \alpha_1 = 0, \alpha_2 = 0, ..., \alpha_m = 0. \quad (3)$$

It is of note that the LME approach does not work when other explanatory variables (multiple groups, conditions, or quantitative covariates) are involved because (2a) or (2b) cannot be formulated due to the parameterization constraint through dummy coding. For instance, when there are two groups involved, the typical contrast coding for the two groups renders one dummy variable (e.g., the contrast of one group vs. the other when effect coding is adopted); however, such a coding strategy relies on the existence of an intercept in the model. If the two groups are coded by two indicator variables, the model matrix would become overparameterized.

Area-under-the-Curve (AUC)

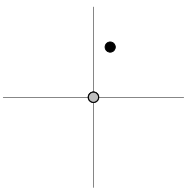
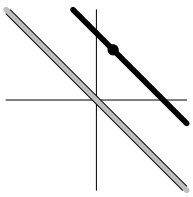
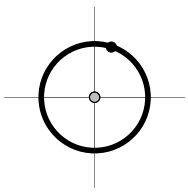
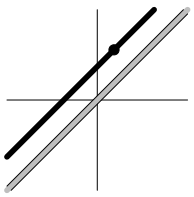
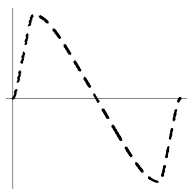
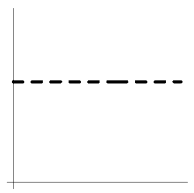
The multiple estimates associated with the multiple basis functions can be reduced to a single value, which is the area under the curve of the estimated response function. The AUC hypotheses for the two prototypes (2a) and (2b) become

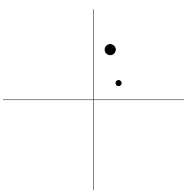
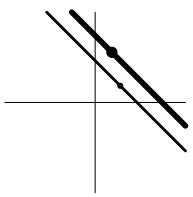
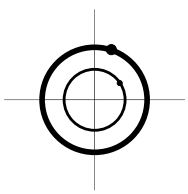
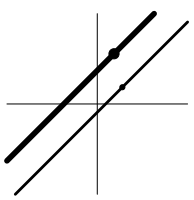
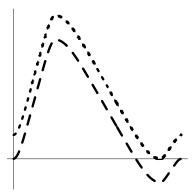
$$H_{01}^{AUC} : \sum_{j=1}^{m} \alpha_j = 0, \quad (4a)$$

$$H_{02}^{AUC} : \sum_{j=1}^{m} \alpha_{1j} = \sum_{j=1}^{m} \alpha_{2j}. \quad (4b)$$

That is, the sum of the m coefficients (or area under the HDR curve) is used to summarize the overall response amplitude per subject in one- or two-sample t-test at the group level. The AUC hypotheses (4a) and (4b) are essentially a zero-way interaction (or intercept) and a one-way interaction (or the main effect of Group or Condition) respectively and can be performed under the AN(C)OVA, GLM, or MVM framework. Their geometrical interpretations are as follows (cf. **Table 1**). The data for H_{01}^{AUC} lie on an $\mathbb{R}^{m-1}$ isosurface (or hyperplane) $\alpha_1 + ... + \alpha_m = c$, and the

TABLE 1 | Schematic comparisons among various testing methods.

One-sample				
Method[a]	**MVT/LME**	**AUC**	**L2D**	**EXC (XUV and XMV)**
H_0	$\alpha_1 = ... = \alpha_m = 0$	$\alpha_1 + ... + \alpha_m = 0$	$(\alpha_1^2 + ... + \alpha_m^2)^{1/2} = 0$	$\alpha_1 = ... = \alpha_m$
Dimensions in $\mathbb{R}^m$	0	$m-1$	$m-1$	1
DFs for F-statistic[b]	$m, n-m-q+1$	$1, n-q$	$1, n-q$	$m-1, (m-1)(n-q)$
Geometric representation[c] of H_0 and H_1 ($m=2$)				
Geometric representation[d] of HDR when detection failure occurs due to improper H_0 formulation	no		no	

Two-sample or paired				
Method	**MVT**	**AUC**	**L2D**	**EXC (XUV and XMV)**
H_0	$\alpha_{11} = \alpha_{21}, ..., \alpha_{1m} = \alpha_{2m}$	$\sum_{j=1}^m \alpha_{1j} = \sum_{j=1}^m \alpha_{2j}$	$(\sum_{j=1}^m \alpha_{1j}^2)^{1/2} = (\sum_{j=1}^m \alpha_{2j}^2)^{1/2}$	$\alpha_{11} - \alpha_{21} = ... = \alpha_{1m} - \alpha_{2m}$
Dimensions in $\mathbb{R}^m$	0	$m-1$	$m-1$	1
DFs for F-statistic	$m, n-m-q+1$	$1, n-q$	$1, n-q$	$m-1, (m-1)(n-q)$
Geometric representation[e] of H_0 and H_1				
Geometric representation[f] of HDR when detection failure occurs due to improper H_0 formulation	no			

[a] *The table is meant to show the dimensions of each null hypothesis and an instantiation in the rejection domain while the whole rejection domain is not represented here. For example, the reject region of one-sample Hotelling T^2-test for MVT (2a) is outside of an m-dimensional ellipse.*

[b] *An interesting fact is that the numerator degrees of freedom for the F-statistic under MVT and UVT are the dimensions of the complementary space to the associated null hypothesis H_0, or the dimensions of the alternative hypothesis H_1.*

[c] *The two axes represent the two weights associated with the two basis functions. The whole rejection regions are not shown here, and the shaded (gray) and solid (black) areas correspond respectively to the null hypothesis H_0 space and an instantiation (and its dimension) in the alternative hypothesis H_1 space. Detection failure occurs when the group centroid falls on the diagonal line other than the origin under AUC and EXC.*

[d] *The horizontal and vertical axes represent time and the amplitude of HDR curve (dashed line).*

[e] *The two axes represent the two weights associated with the two basis functions. The whole rejection regions are not shown here, and the shaded and sold areas correspond respectively to the null hypothesis H_0 space and an instantiation (and its dimension) in the alternative hypothesis H_1 space. The two types of line thickness (or dot size) differentiate the two groups (or conditions).*

[f] *The horizontal and vertical axes represent time and the amplitude of HDR curves. The two line types, dashed and dotted, differentiate the two groups or conditions.*

associated test for AUC (4a) is executed on the distance between the data isosurface and the null isosurface $\alpha_1+...+\alpha_m = 0$. As the correct null hypothesis for MVT (2a) is only a subset of AUC (4a), the rejection domain of AUC (4a) is only a subset of the rejection domain for MVT (2a), leading to a misrepresentation in (4a) and a detection failure when a data point lies on $\alpha_1 + ... + \alpha_m = 0$ but not at the origin (i.e., the HDR curve has roughly equal area below and above the x-axis, e.g., a large undershoot). Similarly for H_{02}^{AUC}.

Euclidean Distance (L2D)

As an alternate dimension reduction approach, the null hypotheses associated with the Euclidean or L^2 distance (L2D) for ESM can be formulated respectively as

$$H_{01}^{L2D} : (\sum_{j=1}^{m} \alpha_j^2)^{1/2} = 0, \tag{5a}$$

$$H_{02}^{L2D} : (\sum_{j=1}^{m} \alpha_{1j}^2)^{1/2} = (\sum_{j=1}^{m} \alpha_{2j}^2)^{1/2}. \tag{5b}$$

In other words, one captures the overall magnitude for each subject using the L^2-distance of the m regression coefficients from no response, and then performs one- or two-sample t-test on the distances.

For ASM, the null hypotheses with the focus on the canonical basis are

$$H_0^{CAN} : \alpha_1 = 0, \tag{6a}$$

$$H_0^{CAN} : \alpha_{11} = \alpha_{21}. \tag{6b}$$

And the null hypotheses for L2D (Calhoun et al., 2004; Steffener et al., 2010) are tested with the first two bases,

$$H_0^{L2D} : sgn(\alpha_1)(\alpha_1^2 + \alpha_2^2)^{1/2} = 0, \tag{7a}$$

$$H_0^{L2D} : sgn(\alpha_{11})(\alpha_{11}^2 + \alpha_{12}^2)^{1/2} = sgn(\alpha_{21})(\alpha_{21}^2 + \alpha_{22}^2)^{1/2} \tag{7b}$$

or with all the three bases,

$$H_0^{L2D} : sgn(\alpha_1)(\alpha_1^2 + \alpha_2^2 + \alpha_3^2)^{1/2} = 0, \tag{8a}$$

$$H_0^{L2D} : sgn(\alpha_{11})(\alpha_{11}^2 + \alpha_{12}^2 + \alpha_{13}^2)^{1/2} = sgn(\alpha_{21}) (\alpha_{21}^2 + \alpha_{22}^2 + \alpha_{23}^2)^{1/2}, \tag{8b}$$

where *sgn* is the sign function. That is, the L2D for ASM is similar to the L2D for ESM, but using the two or three weights associated with the two or three basis functions in ASM and assigning the sign of the canonical response to the resultant L^2-distance.

Their geometrical interpretations are as follows (**Table 1**). The data for H_{01}^{L2D} lie on an $\mathbb{R}^{m-1}$ iso-sphere, and the associated test for (5a) is executed on the radius of the $\mathbb{R}^{m-1}$ iso-sphere, leading to no geometrical distortion (but not necessarily true statistically). On the other hand, the data for H_{02}^{L2D} are on two $\mathbb{R}^{m-1}$ iso-sphere surfaces, and the associated test for (5b) acts on the radius difference between the two $\mathbb{R}^{m-1}$ iso-spheres, resulting a detection failure when the two HDR curves have roughly the same radius.

Effect-by-Component Interaction (EXC: XUV and XMV)

By treating the m effect estimates from ESM as m levels of a within-subject factor Component, one can test the hypothesis for the effect-by-component interaction (EXC); that is, the m regression coefficients associated the m basis functions are taken to the group level without any condensation:

$$H_{01}^{EXC} : \alpha_1 = \alpha_2 = ... = \alpha_m, \tag{9a}$$

$$H_{02}^{EXC} : \alpha_{11} - \alpha_{21} = \alpha_{12} - \alpha_{22} = ... = \alpha_{1m} - \alpha_{2m}. \tag{9b}$$

As discussed in Chen et al. (2014), EXC (9) can be tested through two methods, one univariate testing for the interaction (XUV), and one multivariate testing for the interaction (XMV). More specifically, with XUV one tests the equality among the m components in (9) by treating them as the m levels of a within-subject factor in an AN(C)OVA or univariate GLM platform. In contrast, the equality among the m components in (9) is tested in XMV as m simultaneous variables in an MAN(C)OVA or multivariate GLM (Appendix B).

The geometrical interpretations of the hypotheses are the following (**Table 1**). EXC (9a) tests the main effect (or first-way interaction) of Component, representing a straight line in $\mathbb{R}^m$. The associated test for (9a) is executed on the distance between the data line and the null line (a diagonal line through the origin). As the correct null hypothesis (2a) is only a subset of H_{01}^{EXC}, its rejection domain is only a subset of the rejection domain for MVT (2a), leading to a misrepresentation in (9a) and a detection failure when the group centroid lies on the null line but not at the origin (i.e., the HDR curve is roughly a flat line). Similarly, EXC (9b) as a two-way interaction between Group/Condition and Component is represented by two lines, and the corresponding test acts on the distance between the two lines: are the HDR profiles parallel with each other between the two groups or conditions? As the correct null hypothesis (2b) is only a subset of EXC (9b), the rejection domain of EXC (9b) is only a subset of MVT (2b), resulting in a misrepresentation in (9b) and a detection failure when the two HDR curves are roughly parallel with each other (**Table 1**).

SIMULATIONS AND REAL EXPERIMENT RESULTS

Among all the testing strategies, LME and MVT are the most precise (points in **Table 1**). Among all the dimensional reduction methods, the two EXC methods, XUV and XMV, are of the closest approximation to the null hypothesis (lines), while AUC and L2D are the least accurate ($\mathbb{R}^{m-1}$ planes and sphere surfaces respectively). We need to address the question of whether the geometric accuracy order translates to statistical power through simulations and to performance when the methods are applied to real data.

Simulations of Group Analysis with Different Testing Methods

As the spatial extent of FMRI data analysis is independently controlled through false positive rate or family-wise error, the

simulations here were performed at a voxel to examine and compare the false positives and power performance among the testing methods. Simulated data were generated with the following parameters, imitating a typical FMRI group analysis with six scenarios (top row in **Figure 1**): a) one group of subjects with a small undershoot at the end of HDR curve; b) one group of subjects with a moderate undershoot at the end; c) two homoscedastic groups (same variance between groups) with equal number of subjects in each with a similar HDR profile but a factor of 2 difference in amplitude; d) two homoscedastic groups with equal number of subjects in each with HDR having the same amplitude but with a 2 s difference in peak location; e) two heteroscedastic groups (different variance between groups) with equal number of subjects in each with a similar HDR profile but a factor of 2 difference in amplitude; and f) two heteroscedastic groups with equal number of subjects in each with HDR having the same amplitude but with a 2 s difference in peak location. The HDRs are presumably estimated through 7 basis functions (e.g., TENT in AFNI) at the individual level, and the associated 7 effect components $\{\beta_i, i = 1, 2, ..., 7\}$ at the TR grids are assumed to follow a multivariate Gaussian distribution with a first order autoregressive AR(1) structure for their variance-covariance matrix

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^6 \\ \rho & 1 & \rho & ... & \rho^5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^6 & \rho^5 & \rho^4 & ... & 1 \end{bmatrix}.$$

The choice of a simple Σ structure here is to allow manageable number of simulations while in the same time providing a reasonable structure similar to the one adopted for the Gaussian prior in Marrelec et al. (2003) that guarantees the HDR smoothness. To explore the impact of sample size, the number of subjects in each group was simulated at $n = 9$, 12, 15, 18, 21, 24, 27, 30 with $\rho = 0.3$ for each of the six scenarios. The standard error σ varied (shown in **Figure 1**) across the scenarios to obtain comparable power for each n. 5000 datasets were simulated, each of which was analyzed through `3dMVM` with two explanatory variables, Group (between-subjects factor with 2 levels) and Component (within-subject factor with 7 levels that are associated with the 7 basis functions). False positive rate (FPR) and power were assessed by counting the datasets with their respective F- or t-statistic surpassing the threshold corresponding to the nominal significance level of 0.05. Similarly, one- or two-sample t-test was performed on the AUC and L2D values respectively.

Among the six scenarios, all the testing methods showed proper control of FPR except for L2D with one group of subjects. L2D exhibits high power but at the cost of poor FPR control. This is in part due to the reduction of effect estimates to a positive value regardless the signs of the individual components in ESM. It is possible to reduce this problem in ASM when the sign of the principal kernel is assigned to the resulting L2D measure as shown in (7) and (8). Also, L2D achieved the lowest power with two groups of subjects. AUC simply sums over all the components, significantly misrepresenting the effects when the undershoot becomes moderate. This is reflected in the results where reasonable power is achieved when the undershoot is small and lower power is obtained when the undershoot is moderate. With two groups, AUC performed well in power when the two groups had the same HDR shape, but behaved as poorly as L2D when the two groups had different HDR shapes. As expected, AUC is only sensitive to peak amplitude differences, but is insensitive to shape subtleties. Except for L2D and AUC, the other methods tend to converge in power when the sample size is large enough (e.g., 30 or more). With one group, LME outperformed all other candidates. XUV had a balanced performance on power among all the scenarios, constantly surpassing XMV. Lastly, MVT was slightly more powerful than XUV with two groups when their HDRs were of the same shape with a large number of subjects (e.g., 20 or more per group).

In summary, our simulations show that LME is preferred when there is only one group of subjects with no other explanatory variables present. Under other circumstances, XUV is the preferred choice, especially with the typical sample size of most studies, while MVT, AUC, and XMV may provide some auxiliary detection power.

Results with Experimental Data

How do the testing approaches perform when applied to real data? Would their performances be consistent with the simulations? To address these questions, we ran `3dMVM` on the ESM data presented in the Introduction section with $n = 50$ (2 groups: 21 children and 29 adults), $m = 20$ (2 conditions with each having 10 component estimates at 10 TR grids) and design matrix $\boldsymbol{X}$ of $q = 4$ columns in the MVM (1): all ones (intercept associated with the average effect across groups), effect coding for the two groups, the average age effect between the two groups, and the interaction group:age (or group difference in age effect). The age values were centered within each group so that the group effect can be interpreted as the difference between the two groups at their respective average age. The effect of interest was on the interaction of group and condition: Did the two groups have the same HDR profile difference between the two conditions? Five F-statistics from MVT, XUV (with sphericity correction), AUC, L2D, and XMV, were obtained and then, due to different degrees of freedom, converted to Z-values for direct comparisons (**Figure 2A**). To take advantage of the geometrical representation in **Table 1** when interpreting the effect of interest, we reduce the within-subject factor Condition to the contrast between the two conditions, so that the interaction effect essentially becomes the group contrast in terms of the HDR profile difference between the two conditions (**Figure 2C**).

Consistent with the simulation results, XUV achieved the highest detection power in most regions (**Figure 2A** top) while L2D showed low power (and likely high FPR) due to no differentiation between the positive and negative effect estimates for ESM. All the other three methods, MVT, AUC, and XMV, were generally less powerful than XUV. The strong performance of XUV can be seen in the estimated HDR curves at Voxel 1 (**Figures 2B** left,**C**) extracted from a cluster (left postcentral gyrus). More specifically, the adults had roughly the same HDR profile between the two conditions except for a faster recovery

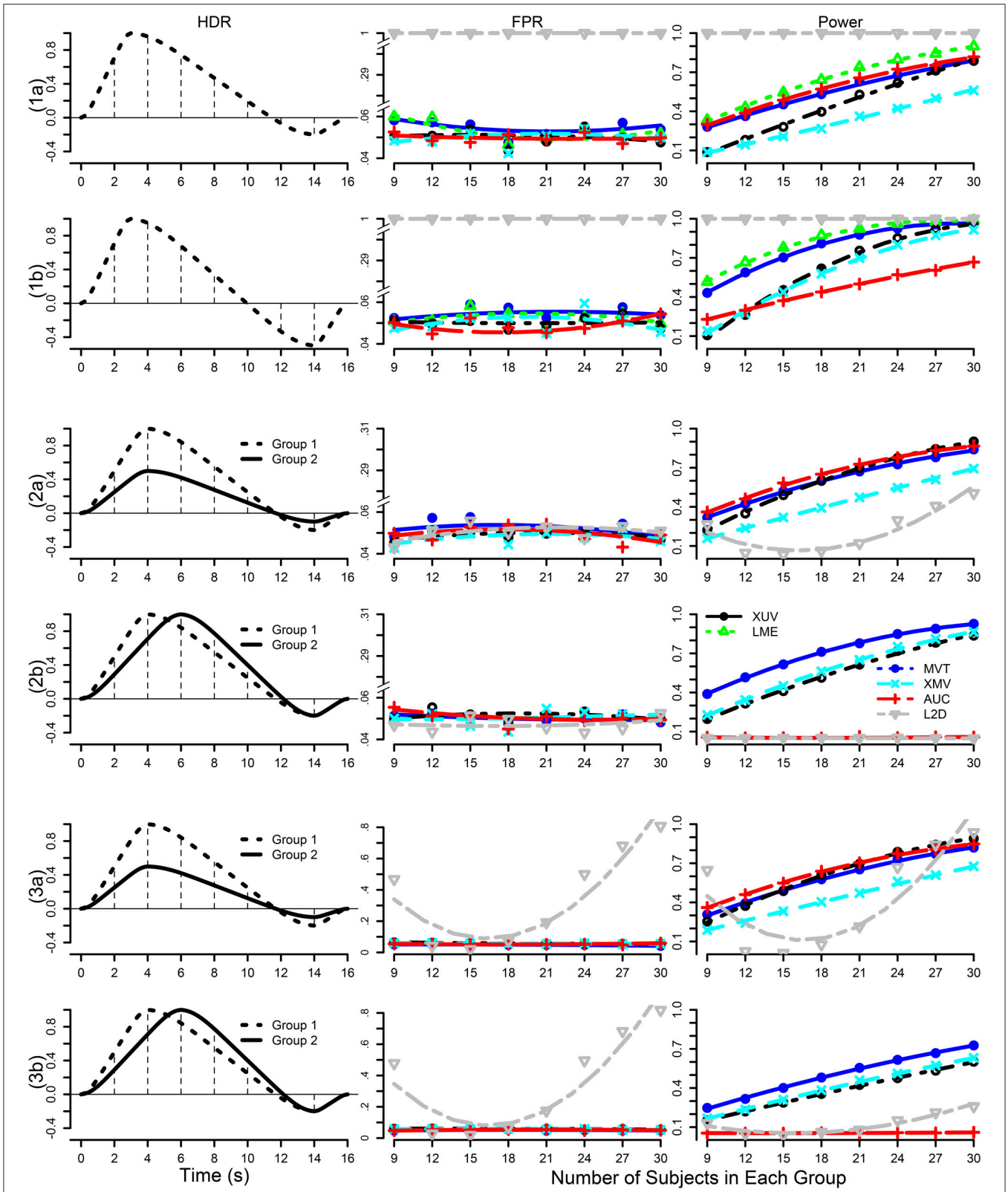


FIGURE 1 | Simulation parameters and results. The six rows correspond to the scenarios in which the presumed HDRs (first column) with a poststimulus undershoot were generated by the convolution program `waver` in AFNI, and sampled at TR = 2 s (shown with vertical dotted lines): (1) one group with a small

(Continued)

FIGURE 1 | Continued
(1a, $\sigma = 1.8$) and a moderate (1b, $\sigma = 1.8$) undershoot, (2) two homoscedastic groups with the same HDR shape but different amplitudes (2a, $\sigma = 0.5$) and with same peak amplitude but a difference of two seconds in peak location (2b, $\sigma = 0.3$), (3) two heteroscedastic groups with the same HDR shape but different amplitudes (3a, $\sigma = 0.3$) and with same peak amplitude but a difference of two seconds in peak location (3b, $\sigma = 0.3$). FPR and power are shown in the second and third columns with a varying number of subjects in each group at a temporal correlation coefficient ρ of 0.3 under six testing approaches: XUV, LME, MVT, XMV, AUC, and L2D. The curves for FPR and power were fitted to the simulation results (plotting symbols) through LOESS smoothing with second order local polynomials.

phase under the Congruent condition than the Incongruent condition; in contrast, the upstroke and peak were more elevated under the Congruent condition in the children than the Incongruent condition except for the recovery phase during the last 3 TRs. Geometrically, the interaction effect between Group and Condition at Voxel 1 is represented by the fact that the HDR profiles of condition difference were intersecting between the two groups (**Figure 2C**). MVT and XMV achieved a moderate power while AUC and L2D failed to reach the significance level of 0.05 at Voxel 1 (**Figure 2B** left). On the other hand, the detection failure of XUV at Voxel 2 (left precuneus) was caused by the fact that the condition contrast was roughly parallel between the two groups (**Figure 2C**), as geometrically demonstrated in **Table 1**. MVT, AUC, and XMV showed their auxiliary role when XUV failed (**Figure 2B** left).

With the ASM analysis results, five tests were performed using `3dMVM`. First, the popular approach of focusing on the effect estimate β_0 associated with the first basis (canonical) function through the hypothesis (6b) was adopted (**Figure 2A** bottom). Secondly, the L2D approach (7) was used on the first two basis functions (not shown here) as well as all three. Thirdly, MVT was performed using (2b) with the three coefficients. Lastly, the HDR curve at each condition was reassembled for each subject using the three coefficients, and the reconstructed effect estimates only at the first 10 TRs were analyzed with `3dMVM` for two reasons: a) with the three SPM curves covering 32 s or 25 TRs, the model would contain too many parameters relative to the data size; b) the effect estimates after the first 10 TRs were mostly negligible. Two tests, XUV and AUC, were performed while MVT and XMV were impossible because the rank was 3 among the 10 effect estimates from the linearly reconstructed HDR per condition.

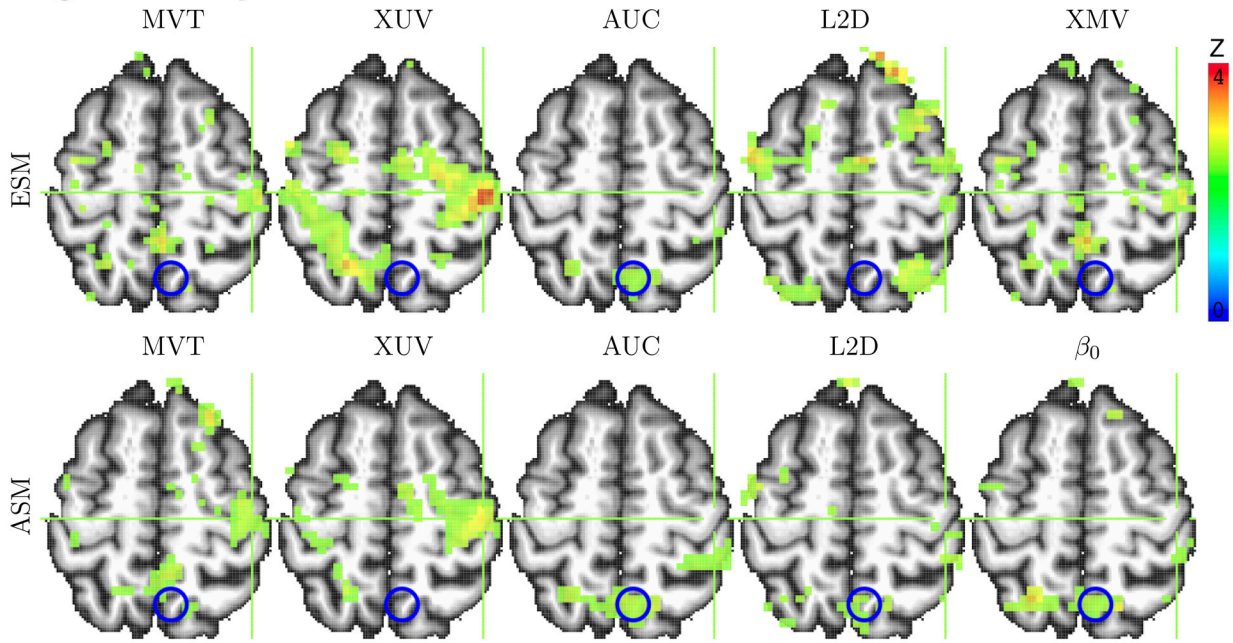
The detection power for both β_0 and L2D with ASM was very low (**Figure 2A** bottom), illustrating the fact that focusing on the peak or the combined effects associated the two or three basis functions would largely fail to detect subtle differences during the BOLD uprising and recovery phases. In contrast, MVT (with the coefficients from three basis functions of ASM), XUV and AUC (with the reconstructed HDRs from ASM) outperformed the conventional approaches of β_0 and L2D in SPM. Such failure of ASM is specifically demonstrated at Voxel 1 where the peak alone or the summarized values from the three coefficients were not as powerful as the reassembled HDR profiles (**Figure 2B** right). It is noteworthy that XUV with ASM was less powerful than its ESM counterpart, showcasing the coarser characterization with three parameters in ASM than the estimation at every time point in ESM. Furthermore, for both ESM and ASM, even though XUV was mostly more powerful than the alternatives, MVT and AUC (as well as XMV for ESM and β_0 for ASM) played a supplementary role when XUV failed (Voxel 2 in **Figure 2B** right).

To recapitulate the performance of the five testing methods in situations when LME cannot be applied, ESM provided a more accurate estimation for the HDR curves than ASM, leading to a higher success in detection power. In addition, with the typical sample size in most studies, XUV as an approximate approach had the lowest power loss at the group level compared to other dimensional alternatives as well as the test with the most accurate hypothesis formulation, MVT. However, MVT plus the lesser accurate approximations such as AUC and XMV may play an auxiliary or even irreplaceable role in situations when XUV suffers from power loss (e.g., **Table 1** or Voxel 2 in **Figure 2**).

DISCUSSION

There are many characteristics that could describe the HDR shape: onset latency, onset-to-peak, peak location, peak duration, magnitude or shape of the undershoot after the onset or during the recovery phase, and habituation or saturation effect. Because of the multiple facets of HDR shape, a lot of effects may well have gone undetected at both individual and group levels in most neuroimaging data analyses, and the failures to capture the shape nuances might have partially contributed to the poor reliability and reproducibility in the field. With a few exceptions, most analyses adopt FSM or ASM mainly for the simplicity of group analysis, as each condition or task is associated with one effect estimate, while other coefficients (e.g., time and dispersion derivatives in ASM) are a priori ignored. That is, activation detection intuitively focuses on the estimated magnitude around the activation peak while statistical inference on the whole HDR shape is generally considered a daunting hurdle. FSM may work well for situations such as a contrast between a condition and fixation. However, it would fail to detect shape subtleties such as prolonged plateau at the peak, slower or faster rise or fall, bigger or longer undershoot, or overall duration. Therefore, FSM through a presumed HDR (gamma variate in AFNI, canonical function in FSL and SPM) is very crude even in an experiment with a block design (Saad et al., 2006; Shan et al., 2013). ASM is an improvement over FSM; however, its flexibility is still limited. For instance, when one is interested in contrasting two conditions (or groups) or in investigating higher-order interactions, the three ASM basis functions may still not be enough in capturing the undershoot subtleties. In addition, characterizing the whole HDR curve with its peak value from ASM for group analysis may suffer from significant power loss, as demonstrated in our real experimental data. Response shapes can vary considerably over space (e.g., Handwerker et al., 2004; Gonzalez-Castillo et al., 2012; Badillo et al., 2013), and we believe it is important to model

A **Significance maps**



B **Significance level at two voxels**

Voxel		ESM: p-value					ASM: p-value				
No.	coordinates	MVT	XUV	AUC	L2D	XMV	MVT	XUV	AUC	L2D	β_0
1	41 -22 54	0.020	0.0002	0.67	0.18	0.013	0.015	0.0025	0.34	0.14	0.42
2	6 -64 46	0.0017	0.119	0.009	0.482	0.0012	0.037	0.22	0.0040	0.092	0.010

C **Estimated HDR curves and the group-by-condition interaction effect at two voxels**

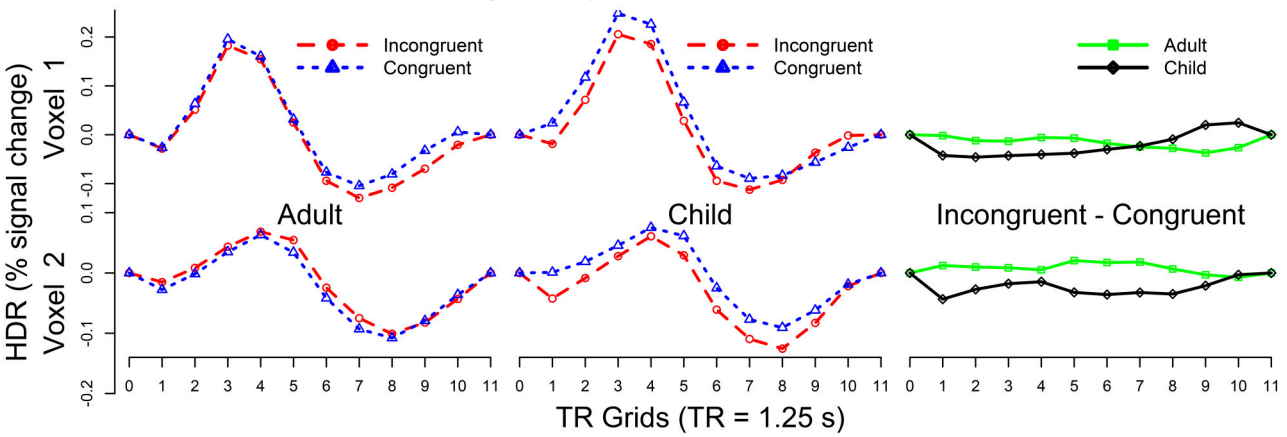


FIGURE 2 | Analysis results of experimental data. (A) Five tests for ESM and ASM are illustrated at an axial slice ($Z = 54$ mm) at $p = 0.05$ level with the radiological convention (left is right). To demonstrate the subtle differences among the methods, the raw results are shown here without multiple testing correction applied. When family-wise error correction through Monte Carlo simulations was adopted, a minimum cluster of 140 voxels for a voxel-level significance of 0.05 led to a surviving cluster at the crosshair (Voxel 1) for XUV for ESM and XUV for ASM. For the cluster labeled with blue circles (Voxel 2), the surviving tests were AUC for ESM, AUC and β_0 for ASM. **(B)** The power differences (p-values in blue when below 0.05) among the five tests are demonstrated at Voxels 1 and 2, whose approximate locations (left postcentral gyrus and left precuneus) are marked with the green crosshair and blue circle respectively in the axial views in **(A)**. **(C)** The estimated HDRs through ESM are shown for the two conditions (first two columns) and their differences (third column) at Voxels 1 and 2. Each HDR profile spans over 11 TRs or 13.75 s. The profile patterns at Voxels 1 and 2 are shared by their neighboring voxels in their respective clusters. In addition to the statistical significance in **(A)** and **(B)**, the HDR signature profiles provide an extra evidence for the associated effects at these voxels.

more accurately the HDRs at the individual level and test for shape rather just amplitude at the group level, particularly when detecting subtle differences between conditions or groups. The dominant adoption of FSM or ASM with a relatively rigid HDR shape reflects the daunting challenge in adopting ESM at the group level, and it is this challenge that motivated our exploration of various group analysis strategies with ESM.

Overview of the Testing Methodologies

Among all the testing strategies for ESM (**Table 1**), MVT and LME maintain an accurate characterization for the hypothesis. In contrast, the dimensional reduction methods AUC, L2D, and EXC (XUV and XMV) project the original space of the alternative hypothesis from $\mathbb{R}^m$ to $\mathbb{R}^1$, $\mathbb{R}^1$, and $\mathbb{R}^{m-1}$, respectively. Any dimensional reduction usually translates to information loss or geometrical distortion. Based on the results from our simulations and real data applications, we believe that the major testing methods for ESM are LME, XUV, MVT, XMV, and AUC, which all have the proper controllability for FPR. If sample size is not an issue in FMRI studies, MVT (e.g., hypothesis 2a or 2b) would be the most accurate approach in terms of hypothesis characterization. However, in practice the number of subjects is usually not large enough for MVT due to resource limitations (e.g., financial cost, time, and manpower), leading to an underpowered performance of MVT as shown in our simulations and real data. Among all the workaround methods through dimensional reduction, XUV has the least hypothesis distortion and the lowest power loss. With one group of subjects and no other explanatory variables present, XUV surpasses MVT, XMV, and AUC in power. However, with an accurate representation of the hypothesis, LME is slightly more efficient than XUV, and should be considered as the first choice (e.g., Alvarez et al., 2008). For all other situations, LME modeling is not feasible due to the constraint of variable parameterization, and we opt for the workaround methods through dimensional reduction, among which AUC is insensitive to subtle shape differences while XMV mostly underperforms unless when the temporal correlation is relatively high (e.g., 0.65 or higher; Chen et al., 2014). XUV achieves the best balance between dimensional reduction and statistical power. However, as XUV tests for parallelism, not exactly the same as the accurate representation characterized in MVT, it may fail in detecting the situation where the HDR profiles are roughly parallel. To compensate for the occasions when XUV fails, other dimensional reduction methods (MVT, AUC, XMV) may offer some complementary detection power.

In light of the discussion here, we strongly encourage the adoption of the ESM approach to achieving two goals: detecting activations and estimating the hemodynamics by characterizing the HDR shape. In addition to the large power gain at both individual and group levels, ESM provides the estimated HDR shape information at the group level, providing an extra layer of validation about the effect veracity through the graphical display of the familiar HDR shape, and alleviating the misconceptions and malpractices prevalent in statistical analysis (e.g., P-hacking, graphical presentation of statistic values instead of effect estimates, overuse of statistical significance; Motulsky, 2014). The HDR profile information from ESM offers a precious boost especially when a cluster fails to survive the typical stringent thresholding for multiple testing correction but still reaches the significance level of 0.05 at the voxel level. Such a reassuring support of ESM is unavailable from the alternatives of FSM and ASM, with which typically the investigator would be only able to report the peak HDR magnitude or statistic values at a region.

Our recommendation of adopting ESM not only applies to event-related experiments, but also are adaptable to modeling the attenuation or habituation effect in block designs (Saad et al., 2006). In addition, this approximation modeling methodology of XUV assisted with MVT, AUC, and XMV has been applied to DTI data in which the simultaneous variables (white matter network groups such as corpus callosum, corona radiata, left and right hemispheric projection fibers, left and right hemispheric association fibers) were modeled by multiple explanatory variables (e.g., sex, age, behavioral measures) for each response variable such as fractional anisotropy, axial diffusivity, mean diffusivity, radial diffusivity, T1 relaxation time, proton density, and volume (Taylor et al., 2015).

The proposed modeling strategies have been implemented into the open-source program `3dMVM` in AFNI, which offers the investigator all the testing results in the output including XUV and the auxiliary approaches (MVT, XMV, and AUC). MVT for the components from ESM presents a unique challenge when one or more within-subject factors are included in the model, and we offer a testing strategy that still fits in the MVM framework (Appendix B). As an alternative, these tests could be conducted in the traditional univariate GLM except for the two multivariate methods, MVT and XMV. In other words, some of the testing methods (MVT and XMV) are truly multivariate, while others (XUV, AUV, and L2D) are essentially univariate. However, as we demonstrated in Chen et al. (2014), these univariate tests are sometimes difficult to perform under the univariate framework, as shown by the implementation challenges faced by some of the neuroimaging packages. Instead, these univariate tests can be more conveniently formulated under the MVM platform by treating the levels of each within subject factor as simultaneous variables in (1) and then constructing the univariate testing statistics through a conversion process. For example, those univariate tests presented in **Figure 2** cannot be performed under the univariate GLM framework due to the incorporation of a covariate (age) in the presence of two within subject factors (Condition and HDR effects). It is in this sense that we frame our discussion here under the MVM perspective.

Limitations of the ESM Approach

It is noteworthy that the reliability information from the individual subject analysis is not considered at the group level with the modeling methods discussed here, unlike the mixed-effect multilevel analysis (Worsley et al., 2002; Woolrich et al., 2004; Chen et al., 2012). In addition, the number of basis functions monotonically increases among FSM, ASM, and ESM, therefore it is expected that the goodness of fit at the individual subject analysis level improves across the three methods. On the other hand, as each condition is characterized through multiple

(e.g., $\geq$7) basis functions in ESM, a reliable estimation of the HDR curve at the individual level pays a price through the lower degrees of freedom and requires enough (e.g., 20 or more) trials per condition, and may encounter the risk of numerical instability due to high correlations or even multicollinearity among the regressors. These latter issues can be exacerbated by poor stimulus timing designs. In addition, the typical regression analysis at the individual level assumes the linearity of HDR across trials. Although available (e.g., `3dNLfim` in AFNI), a non-linear approach is usually difficult to handle and still requires some extent of prior information about the HDR shape. Furthermore, the ESM approach is generally considered to be susceptible to noise or effects unrelated to the effects of interest (e.g., head motion, physiological confounds). In other words, the confounding effects may leak into the HDR estimation through over-fitting. However, the false positives from the potential over-fitting at the individual level is less a concern at the group level for the following reasons: a) the likelihood is reduced unless most subjects systematically have similar or same confounding effects; b) cluster-based inferences may reduce the risk of false positives; and most importantly c) examination of the estimated HDR profiles offer an extra safeguard to filter out the potential false positives.

Comparisons with Other Modeling Approaches

Some (not all) of the dimensional reduction methods for ESM discussed here have been sporadically and individually applied to real data in the literature. For example, a popular practice with ASM is to solely focus on the coefficient of the principal basis function (e.g., canonical curve in SPM) with other coefficients (e.g., time and dispersion derivatives) being a priori abandoned. As our results with real data showed, the investigator may fail to detect most activations when the effect lies in the HDR shape nuances but not the peak. One suggestion for ASM was to extend the definition of amplitude in (6) to the L^2-distance by including either the effect for the time derivative (7) or the effects for both time and dispersion derivatives (8) (Calhoun et al., 2004; Worsley and Taylor, 2006; Steffener et al., 2010). A similar approach was to express the effect estimates from the first two basis functions of ASM as a complex number (Wang et al., 2012). However, the potential issues with L2D or its analogs (e.g., Worsley and Taylor, 2006) are the following. a) The definition of amplitude extension in (7) and (8) is under the premise that all the three basis functions are orthogonal with each other (Calhoun et al., 2004). However, only the first two basis functions are orthogonal with each other, but not the third one. b) The second and third basis functions are not normalized; that is, they are not scaled to have a maximum value of 1, unlike the first basis function. In addition, the three effect estimates have different dimensions: the first is of percent signal change while the other two of percent signal change by the unit of time. Therefore, it is difficult to render a physically meaning interpretation with the L2D measures. c) All the effect estimates including negative values are folded into a positive L2D measure, which cannot be differentiated among those effect estimates on the same circle or sphere (see **Table 1**). In addition, it may lead to the violation of the Gaussian distribution assumption, as illustrated in the poor controllability of FPR (**Figure 1**). d) Their power performance is not satisfactory (**Figures 1**, **2**). As an alternative, MVT or LME through the hypothesis (2a) or (2b) on the two or three effect estimates from ASM, as shown in **Figure 2A**, provides a more accurate characterization because it allows for different units or dimensions across the effects.

Similarly for ESM, two dimensional reduction methods have separately been adopted in data analyses. For example, AUC was employed in Beauchamp et al. (2003), Greene et al. (2007), and McGregor et al. (2013). Although not explicitly stated, XUV was used in several real applications to identify the HDR effect under a condition through the main effect (or one-way interaction) of the ESM components in a one-way within-subject ANOVA (Weissman et al., 2006; Geier et al., 2007; Church et al., 2008), to detect the group or condition differences in the overall HDR shape through the group-by-component or condition-by-component interaction in a two-way ANOVA (e.g., Schlaggar et al., 2002; Church et al., 2008; Shuster et al., 2014), and to explore the three-way group-by-task-by-component interaction (Church et al., 2008). However, two limitations were not addressed in those analyses: the potential identification failure of XUV (**Table 1** and Voxel 2 in **Figure 2**), and the limited applicability of univariate GLM.

Some comparisons were performed in terms of amplitude, peak latency, and duration in the estimated HDR among various modeling methods (e.g., FSM, L2D, ESM, a nonlinear model, and inverse logit model; Lindquist et al., 2009). The inverse logit model was deemed the best among the candidates in both simulations and real data, and slightly more powerful than ESM. However, the comparisons were not optimal. First, the dimensional reduction from the HDR shape in $\mathbb{R}^m$ to the three quantities (amplitude, delay, and duration) in $\mathbb{R}^3$ might be compromised in power when detecting the shape subtleties—this point can be highly dependent on the experiment. Secondly, the reliability for the estimation of the three characteristics was suboptimal. For example, the lackluster performance of ESM in Lindquist et al. (2009) might be caused by the inaccurate amplitude based on the first local peak because such an approach could be misleading especially when more than one local peak occurs. Lastly, the final group analyses were still focused on the amplitude with the Student's t-test, an effective dimensional reduction from $\mathbb{R}^m$ to $\mathbb{R}^1$.

A multivariate approach (Zhang et al., 2012) was previously proposed, analogous to our method except for the following differences. It was demonstrated among the voxels within only five structurally pre-defined regions; smoothing the estimated HDR from each subject by a Gaussian kernel and imposing regularization on the smoothed HDR were performed to improve the temporal continuities of the HDR; and group analysis was run through multivariate testing of one-sample or pair-wise comparisons among conditions, equivalent to MVT (2a or 2b) discussed here. Another approach (Zhang et al., 2013) assumed that the HDR under each condition would only vary in amplitude and latency across subjects; that is, the HDR shape was presumed same across all subjects. Specifically, the HDR curve for each condition was characterized at the group level by two parameters:

one was of interest (amplitude) and the other of no interest (delay). In addition, the HDR shape (fixed across subjects) was modeled by cubic splines plus their time derivatives. Once the amplitude was estimated for each subject in a one-tier model that incorporated both within- and across-subject variances, a second round of group analysis was performed only on the amplitudes (ignoring the delay) through typical one-sample or paired *t*-test to make inference about a condition or contrast. The approach was demonstrated among the voxels within only three structurally predefined regions.

Recently, a hierarchical approach was proposed for ESM through integrating both individual and group levels into one model (Degras and Lindquist, 2014) in which the HDR curves were captured through multiple higher-order B-spline functions. Even though only demonstrated on one slice of data, the approach is appealing because the variability at both levels is accounted for. However, the current implementation in Matlab is hindered by the following constraints or limitations. a) Spatial parcellation based on anatomical structure was required to determine the temporal correlation structure in the noise component. More applicable approaches would be based on a priori regions that are functionally parcellated through, for example, hierarchical clustering (Thirion et al., 2006; Ji, 2010), joint parcellation detection-estimation (Badillo et al., 2014), consensus clustering (Badillo et al., 2013), k-means clustering (Ji, 2010), etc. (b) The HDR shape may vary across different stimulus conditions under some scenarios (e.g., Ciuciu et al., 2003), and a presumption of the same shape HDR as in Degras and Lindquist (2014) may decrease the detection power when the shape subtleties are of interest. The same HDR assumption is reasonable under other circumstances and has proven sufficient for encoding or decoding the brain activity (Pedregosa et al., 2015). c) Final statistical inference in Degras and Lindquist (2014) through an asymptotic *t*-test was still based on the scaling factors of the same HDR curve shared by all conditions, a dimensional reduction approach from $\mathbb{R}^m$ to $\mathbb{R}^1$. An alternative approach is the incorporation of both individual and group levels in a mixed-effects model under the Bayesian framework (Chaari et al., 2013; Badillo et al., 2014). Applied at a priori regions that are functionally parcellated, this jointed detection and estimation method may render a robust procedure less sensitive to outliers than the conventional two-tier methods under the assumption that all the voxels share the same HDR within a region or parcel.

CONCLUSION

Here we demonstrate with simulations and experimental data that the fixed-shape (FSM) or adjusted-shape (ASM) method may fail to detect most of the shape subtleties (e.g., the speed of rise or recovery, undershoot) in hemodynamic response (HDR). In contrast, the estimated-shape method (ESM) through multiple basis functions would more accurately characterize the cerebral blood flow regulation, and significantly improve the detection power at both individual and group levels. In addition, we propose an analysis scheme for ESM that still fits within the conventional two-tier analysis pipeline and achieves higher statistical power than the alternatives: one performs regression time series analysis separately for each individual subject, and then conducts group analysis with the individual effect estimates. For one group of subjects, a linear mixed-effects (LME) model is preferred if no other explanatory variables are present. In all other scenarios, statistical inferences on the HDR shape can be achieved through a hybrid combination of multivariate testing (MVT) and dimensional reduction approaches with a multivariate model (MVM). Simulations are shown in terms of controllability for false positive rate (FPR) and power achievement among various testing methods. The strategy was applied to a dataset from a real experiment to compare among different testing strategies in terms of power assessment. In addition, we showcase that the MVM flexibility allows any number of explanatory variables including between- and within-subject factors as well as between-subjects covariates.

ACKNOWLEDGMENTS

Our work benefited significantly from the statistical computational language and environment *R*, its many packages, and the great support of the *R* community. All the plots were created in *R* with the base graphics library. Special thanks are due to Helios de Rosario for his help in technical details of using the *R* package *phia*. The research and writing of the paper were supported by the NIMH and NINDS Intramural Research Programs of the NIH/HHS, USA.

REFERENCES

Alvarez, R. P., Biggs, A., Chen, G., Pine, D. S., and Grillon, C. (2008). Contextual fear conditioning in humans: cortical-hippocampal and amygdala contributions. *J. Neurosci.* 28, 6211–6219. doi: 10.1523/JNEUROSCI.1246-08.2008

Badillo, S., Desmidt, S., Ginisty, C., and Ciuciu, P. (2014). "Multi-subject Bayesian joint detection and estimation in fMRI," in *PRNI* (Tübingen), 1–4.

Badillo, S., Varoquaux, G., and Ciuciu, P. (2013). "Hemodynamic estimation based on consensus clustering," in *IEEE Pattern Recognition in Neuroimaging (PRNI)*, (Philadelphia, PA), 211–215.

Barbé, K., Van Moer, W., and Nagels, G. (2012). Fractional-order time series models for extracting the haemodynamic response from functional magnetic resonance imaging data. *IEEE Trans. Biomed. Eng.* 59, 2264–2272. doi: 10.1109/TBME.2012.2202117

Beauchamp, M. S., Lee, K. E., Haxby, J. V., and Martin, A. (2003). FMRI responses to video and point-light displays of moving humans and manipulable objects. *J. Cogn. Neurosci.* 15, 991–1001. doi: 10.1162/089892903770007380

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Bernal-Rusiel, J. L., Greve, D. N., Reuter, M., Fischl, B., Sabuncu, M. R., and Alzheimer's Disease Neuroimaging Initiative, (2013). Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. *Neuroimage* 66, 249–260. doi: 10.1016/j.neuroimage.2012.10.065

Birn, R. M., Saad, Z. S., and Bandettini, P. A. (2001). Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. *Neuroimage* 14, 817–826. doi: 10.1006/nimg.2001.0873

Buxton, R. B., Uludağ, K., Dubowitz, D. J., and Liu, T. T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage* 23(Suppl. 1), S220–S233. doi: 10.1016/j.neuroimage.2004.07.013

Calhoun, V. D., Stevens, M. C., Pearlson, G. D., and Kiehl, K. A. (2004). fMRI analysis with the general linear model: removal of latency-bias by incorporation of hemodynamic derivative terms. *Neuroimage* 22, 252–257. doi: 10.1016/j.neuroimage.2003.12.029

Chaari, L., Vincent, T., Forbes, F., Dojat, M., and Ciuciu, P. (2013). Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE Tran. Med. Imaging* 32, 821–837. doi: 10.1109/TMI.2012.2225636

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., and Cox, R. W. (2014). Applications of multivariate modeling to neuroimaging group analysis: a comprehensive alternative to univariate general linear model. *Neuroimage* 99, 571–588. doi: 10.1016/j.neuroimage.2014.06.027

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., and Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *Neuroimage* 73, 176–190. doi: 10.1016/j.neuroimage.2013.01.047

Chen, G., Saad, Z. S., Nath, A. R., Beauchamp, M. S., and Cox, R. W. (2012). FMRI group analysis combining effect estimates and their variances. *Neuroimage* 60, 747–765. doi: 10.1016/j.neuroimage.2011.12.060

Church, J. A., Coalson, R. S., Lugar, H. M., Petersen, S. E., and Schlaggar, B. L. (2008). A developmental fMRI study of reading and repetition reveals changes in phonological and visual mechanisms over age. *Cereb. Cortex* 18, 2054–2065. doi: 10.1093/cercor/bhm228

Ciuciu, P., Poline, J. B., Marrelec, G., Idier, J., Pallier, C., and Benali, H. (2003). Unsupervised robust nonparametric estimation of the hemodynamic response function for any fMRI experiment. *IEEE Trans. Med. Imaging* 22, 1235–1251. doi: 10.1109/TMI.2003.817759

Cohen, M. S. (1997). Parametric analysis of fMRI data using linear systems methods. *Neuroimage* 6, 93–103. doi: 10.1006/nimg.1997.0278

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi: 10.1006/cbmr.1996.0014

Degras, D., and Lindquist M. A. (2014). A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. *Neuroimage* 98, 61–72. doi: 10.1016/j.neuroimage.2014.04.052

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., et al. (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14240–14245. doi: 10.1073/pnas.0606302103

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., and Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* 33, 636–647. doi: 10.1002/mrm.1910330508

Fox, J., Friendly, M., and Weisberg S. (2013). Hypothesis tests for multivariate linear models using the *car* package. *The R. Journal* 5, 2013, 39–52.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998a). Event-related fMRI: characterizing differential responses. *Neuroimage* 7, 30–40. doi: 10.1006/nimg.1997.0306

Friston, K. J., Josephs, O., Rees, G., and Turner, R. (1998b). Nonlinear event-related responses in fMRI. *Magn. Reson. Med.* 39, 41–52. doi: 10.1002/mrm.1910390109

Geier, C. F., Garver, K. E., and Luna, B. (2007). Circuitry underlying temporally extended spatial working memory. *Neuroimage* 35, 904–915. doi: 10.1016/j.neuroimage.2006.12.022

Gonzalez-Castillo, J., Saad, Z. S., Handwerker, D. A., Inati, S. J., Brenowitz, N., and Bandettini, P. A. (2012). Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5487–5492. doi: 10.1073/pnas.1121049109

Greene, A. J., Gross , W. L., Elsinger , C. L., and Rao , S. M. (2007). Hippocampal differentiation without recognition: an fMRI analysis of the contextual cueing task. *Learn. Mem.* 14, 548–553. doi: 10.1101/lm.609807

Greenhouse, S. W., and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika* 24, 95–112. doi: 10.1007/BF02289823

Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage* 21, 1639–1651. doi: 10.1016/j.neuroimage.2003.11.029

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage* 62, 852–855. doi: 10.1016/j.neuroimage.2012.03.016

Huynh, H., and Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomised block and split-plot designs. *J. Educ. Stat.* 1, 69–82. doi: 10.2307/1164736

Ji, Y. (2010). *Data-driven fMRI data analysis based on parcellation*. Ph.D. thesis, The University of Nottingham, England.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage* 45, S187–S198. doi: 10.1016/j.neuroimage.2008.10.065

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature* 453, 869–878. doi: 10.1038/nature06976

Logothetis, N. K., and Wandell, B. A. (2004). Interpreting the BOLD signal. *Annu. Rev. Physiol.* 66, 735–769. doi: 10.1146/annurev.physiol.66.082602.092845

Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., and Poline, J. B. (2006). Combined permutation test and mixed-effect model for group average analysis in fMRI. *Hum. Brain Mapp.* 27, 402–410. doi: 10.1002/hbm.20251

Magri, C., Schridde, U., Murayama, Y., Panzeri, S., and Logothetis, N. K. (2012). The amplitude and timing of the BOLD signal reflects the relationship between local field potential power at different frequencies. *J. Neurosci.* 32, 1395–1407. doi: 10.1523/JNEUROSCI.3985-11.2012

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

McGregor, K. M., Nocera, J. R., Sudhyadhom, A., Patten, C., Manini, T. M., Kleim, J. A., et al. (2013). Effects of aerobic fitness on aging-related changes of interhemispheric inhibition and motor performance. *Front. Aging Neurosci.* 5:66. doi: 10.3389/fnagi.2013.00066

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *J. Pharmacol. Exp. Ther.* 351, 200–205. doi: 10.1124/jpet.114.219170

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., and Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *Neuroimage* 104, 209–220. doi: 10.1016/j.neuroimage.2014.09.060

Poline, J. B., and Brett, M. (2012). The general linear model and fMRI: does love last forever? *Neuroimage* 62, 871–880. doi: 10.1016/j.neuroimage.2012.01.133

Rencher, A. C., and Christensen, W. F. (2012). *Methods of Multivariate Analysis, 3rd Edn.* Hoboken, NJ: John Wiley & Sons, Inc.

Reynell, C., and Harris, J. J. (2013). The BOLD signal and neurovascular coupling in autism. *Dev. Cogn. Neurosci.* 6, 72–79. doi: 10.1002/9781118391686

Saad, Z. S., Chen, G., Reynolds, R. C., Christidis, P. P., Hammett, K. R., Bellgowan, P. S. F., et al. (2006). Functional imaging analysis contest (FIAC) analysis according to AFNI and SUMA. *Hum. Brian Mapp.* 27, 417–424. doi: 10.1002/hbm.20247

Schlaggar, B. L., Brown, T. T., Lugar, H. M., Visscher, K. M., Miezin, F. M., and Petersen, S. E. (2002). Functional neuroanatomical differences between adults and school-age children in the processing of single words. *Science* 296, 1476–1479. doi: 10.1126/science.1069464

Shan, Z. Y., Wright, M. J., Thompson, P. M., McMahon, K. L., Blokland, G. G. A. M., de Zubicaray, G. I., et al. (2013). Modeling of the hemodynamic responses in block design fMRI studies. *J. Cereb. Blood Flow Metab.* 34, 316–324. doi: 10.1038/jcbfm.2013.200

Shuster, L. I., Moore, D. R., Chen, G., Ruschello, D. M., and Wonderlin, W. F. (2014). Does experience in talking facilitate speech repetition? *Neuroimage* 87, 80–88. doi: 10.1016/j.neuroimage.2013.10.064

Steffener, J., Tabert, M., Reuben, A., and Stern, Y. (2010). Investigating hemodynamic response variability at the group level using basis functions. *Neuroimage* 49, 2113–2122. doi: 10.1016/j.neuroimage.2009.11.014

Stephen, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. (2007). Comparing hemodynamic models with DCM. *Neuroimage* 38, 387–401. doi: 10.1016/j.neuroimage.2007.07.040

Taylor, P. A., Jacobson, J. L., van der Kouwe, A., Molteno, C. D., Chen, G., Wintermark, P., et al. (2015). A DTI-based tractography study of structural brain changes associated with prenatal alcohol exposure in newborn. *Hum. Brain Mapp.* 36, 170–186. doi: 10.1002/hbm.22620

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., and Poline, J.-B. (2006). Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* 27, 678–693. doi: 10.1002/hbm.20210

Vincent, T., Badillo, S., Risser, L., Chaari, L., Bakhous, C., Forbes, F., et al. (2014). Flexible multivariate hemodynamics fMRI data analyses and simulations with PyHRF. *Front. Neurosci.* 8:67. doi: 10.3389/fnins.2014.00067

Wang, P., Wang, Z., and He, L. (2012). Complex-number representation of informed basis functions in general linear modeling of functional magnetic resonance imaging. *J. Neurosci. Methods* 205, 28–35. doi: 10.1016/j.jneumeth.2011.12.016

Weissman, D. H., Roberts, K. C., Visscher, K. M., and Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nat. Neurosci.* 9, 971–978. doi: 10.1038/nn1727

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage* 92, 381–397. doi: 10.1016/j.neuroimage.2014.01.060

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage* 21, 1732–1747. doi: 10.1016/j.neuroimage.2003.12.023

Worsley, K. J., and Taylor, J. E. (2006). Detecting fMRI activation allowing for unknown latency of the hemodynamic response. *Neuroimage* 29, 649–654. doi: 10.1016/j.neuroimage.2005.07.032

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., et al. (2002). A general statistical analysis for fMRI data. *Neuroimage* 15, 1–15. doi: 10.1006/nimg.2001.0933

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* 12, 900–918. doi: 10.1038/jcbfm.1992.127

Zhang, T., Li, F., Beckes, L., and Coan, J. A. (2013). A semi-parametric model of the hemodynamic response for multi-subject fMRI data. *Neuroimage* 75, 136–145. doi: 10.1016/j.neuroimage.2013.02.048

Zhang, T., Li, F., Beckes, L., Brown, C., and Coan, J. A. (2012). Nonparametric inference of the hemodynamic response using multi-subject fMRI data. *Neuroimage* 63, 1754–1765. doi: 10.1016/j.neuroimage.2012.08.014

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

List of Acronyms used in the Paper

AN(C)OVA	Analysis of (co)variance
ASM	Adjusted-shape method
AUC	Are under the curve
ESM	Estimated-shape method
EXC	Effect-by-component interaction
FPR	False positive rate
FSM	Fixed-shape method
GLM	General linear model
HDR	Hemodynamic response
IRF	Impulse response function
L2D	Euclidian (L^2) distance
LME	Linear mixed-effects
MAN(C)OVA	Multivariate analysis of (co)variance
MVM	Multivariate modeling
MVT	Multivariate testing
UVM	Univariate modeling
UVT	Univariate testing
XMV	Multivariate testing for interaction
XUV	Univariate testing for interaction.

APPENDIX B

FORMULATION OF MULTIVARIATE TESTING IN THE PRESENCE OF ONE OR MORE WITHIN-SUBJECT FACTORS

As discussed in Chen et al. (2014), all the within-subject factors are flattened into $\mathbb{R}^1$ under the multivariate model (MVM) formulation (1). Once the regression coefficient matrix $\boldsymbol{A}$ is estimated through solving the MVM system (1) with the least squares principle, each general linear test (GLT) can be expressed as a function of $\boldsymbol{A}$,

$$H_0 : \boldsymbol{L}_{u\times q}\,\boldsymbol{A}_{q\times m}\,\boldsymbol{R}_{m\times v} = \boldsymbol{0}_{u\times v}, \tag{A1}$$

where the *hypothesis matrix* $\boldsymbol{L}$, through premultiplying, specifies the weights among the rows of $\boldsymbol{A}$ that are associated with the between-subjects variables (groups or subject-specific quantitative covariates), and the *response transformation matrix* $\boldsymbol{R}$, through postmultiplying, formulates the weighting among the columns of $\boldsymbol{A}$ that correspond to the m response variables. It is assumed that $\boldsymbol{L}$ and $\boldsymbol{R}$ are full of row- and column-rank respectively, and $u \le q$, $v \le m$. The matrix $\boldsymbol{L}$ (or $\boldsymbol{R}$) plays a role of contrasting or weighted averaging among the groups of a between-subjects factor (or the levels of a within-subject factor).

The conventional multivariate test (MVT) can be performed through any of the four multivariate statistics (Wilks' λ, Pillai-Bartlett trace, Lawley-Hotelling trace, and Roy's largest root) with $\boldsymbol{R} = \boldsymbol{I}_m$ once the hypothesis matrix $\boldsymbol{L}$ in (A1) is constructed (Appendix B in Chen et al., 2014). For instance, suppose that we consider an m-variate model with the following explanatory variables: three genotypes of subjects, age and their interactions. Via effect coding with the first genotype as reference, the model matrix $\boldsymbol{X}$ in (1) is of $q = 6$ columns: one for the intercept, two for the three genotypes, one for age, and two for their interactions. Accordingly, the $q = 6$ rows in $\boldsymbol{A}$ represent the overall mean, the respective effects for the second and third genotypes relative to the overall mean, the age effect associated with the overall mean, and the respective age effects for the second and third genotypes relative the average age effect. MVT for the main effect of genotypes, the genotype-by-age interaction, and the age effect for the first genotype can be obtained under (A1) respectively with

$$\boldsymbol{L}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{L}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$
$$\boldsymbol{L}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & -1 \end{bmatrix}, \boldsymbol{R}_1 = \boldsymbol{R}_2 = \boldsymbol{R}_3 = \boldsymbol{I}_m.$$

Similarly, both univariate and within-subject multivariate tests can be formulated by obtaining both the hypothesis matrix $\boldsymbol{L}$ and the response transformation matrix $\boldsymbol{R}$ in (A1) (Appendix C in Chen et al., 2014). In addition, all the *post-hoc* t- and F-tests (options -gltCode and -glfCode respectively in `3dMVM`) are also constructed as MVT under the platform (A1). For instance, the effect under a specific level and the contrast between two levels of a within-subject factor through -gltCode are evaluated essentially by a one-sample and a paired t-test respectively, while the main effect of a within-subject factor through -glfCode is assessed by a within-subject multivariate test.

When $\boldsymbol{R} = \boldsymbol{1}_{m\times 1}$, the hypothesis (A1) solely focuses on the between-subjects explanatory variables (columns in the model matrix $\boldsymbol{X}$ of MVM; 1) while the effects among the levels of the within-subject factors are averaged (or collapsed). Therefore, the AUC approach (4) can be conceptually tested under the multivariate framework (A1), respectively for one group,

$$\boldsymbol{L}_4 = 1, \boldsymbol{R}_4 = \boldsymbol{1}_{m\times 1},$$

and two groups,

$$\boldsymbol{L}_5 = (0, 1), \boldsymbol{R}_5 = \boldsymbol{1}_{m\times 1},$$

even though they would be readily performed through the conventional one- and two-sample t-tests.

When applied to the effect-by-component interaction (9a or 9b) with ESM (EXC in **Table 1**), the MVM framework offers both univariate (XUV) and multivariate (XMV) approaches, which are tested under the same formulation, respectively for one group (A1),

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m,$$
$$\boldsymbol{L}_6 = 1, \boldsymbol{R}_6 = \begin{bmatrix} \boldsymbol{I}_{m-1} \\ -\boldsymbol{1}_{1\times(m-1)} \end{bmatrix},$$

and two groups,

$$H_0 : \alpha_{11} - \alpha_{21} = \alpha_{12} - \alpha_{22} = ... = \alpha_{1m} - \alpha_{2m},$$
$$\boldsymbol{L}_7 = (0, 1), \boldsymbol{R}_7 = \boldsymbol{R}_6.$$

For XMV, standard multivariate testing statistics (Wilks' λ, Pillai-Bartlett trace, Lawley-Hotelling trace, Roy's largest root) are constructed through the eigenvalues of the "ratio" $\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{E})^{-1}$ between the SSPH matrix $\boldsymbol{H}$ for the hypothesis (A1) against the SSPE matrix $\boldsymbol{E}$ for the errors in the full model (Rencher and Christensen, 2012). In contrast, the univariate approach XUV is tested through the formulation of an F-statistic with the numerator and denominator sums of squares being as $tr(\boldsymbol{H}(\boldsymbol{R}^T\boldsymbol{R})^{-1})$ and $tr(\boldsymbol{E}(\boldsymbol{R}^T\boldsymbol{R})^{-1})$ under the sphericity assumption (Fox et al., 2013), and the F-value can be adjusted through the Greenhouse and Geisser (1959) or Huynh and Feldt (1976) correction if the sphericity assumption is violated.

All the applications so far in the literature have been focused on either MVT or UVT. In other words, a strict MVT applies to the situations of truly multivariate nature while a purely UVT is adopted to the conventional AN(C)OVA or GLM. However, if we treat the components from ESM as simultaneous response variables, the presence of one or more within-subject factors (e.g., two task conditions in the experimental data of this paper) necessitates a partial MVT. Here we demonstrate a strategy to formulate partial MVT with the construction of $\boldsymbol{L}$ and $\boldsymbol{R}$ using a template of two-way within-subject ANOVA with factors A and B of a and b levels respectively. Suppose that we want to model the levels of factor A as a simultaneous response variables (e.g., components or effect estimates from ESM) while factor B is considered as an explanatory variable (e.g., conditions). MVT for the effect of B can be achieved through the following specifications in (A1),

$$\boldsymbol{L} = \boldsymbol{I}_q, \boldsymbol{R} = \boldsymbol{I}_a \otimes \boldsymbol{R}^{(B)}.$$

Similarly, if the levels of factor B are modeled as b simultaneous response variables while factor A is considered as an explanatory variable, we have the following MVT specifications for the effect of A,

$$\boldsymbol{L} = \boldsymbol{I}_q, \boldsymbol{R} = \boldsymbol{R}^{(A)} \otimes \boldsymbol{I}_b.$$

The notations $\boldsymbol{R}^{(A)} = \begin{bmatrix} \boldsymbol{I}_{a-1} \\ -1_{1\times(a-1)} \end{bmatrix}$ and $\boldsymbol{R}^{(B)} = \begin{bmatrix} \boldsymbol{I}_{b-1} \\ -1_{1\times(b-1)} \end{bmatrix}$ above are conveniently the effect coding matrices for factors A and B respectively.