Received: 15 April 2021 | Accepted: 10 January 2022

DOI: 10.1111/rssc.12545

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series C Applied Statistics

Regularized regression on compositional trees with application to MRI analysis

Bingkai Wang[1] | Brian S. Caffo[1] | Xi Luo[2] | Chin-Fu Liu[3] | Andreia V. Faria[4] | Michael I. Miller[3] | Yi Zhao[5] | for the Alzheimer's Disease Neuroimaging Initiative*

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[2]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Houston, Texas, USA

[3]Center for Imaging Science, Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA

[4]Department of Radiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[5]Department of Biostatistics, Indiana University School of Medicine and for the Alzheimer's Disease Neuroimaging Initiative, Indianapolis, Indiana, USA

Correspondence
Bingkai Wang, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA.
Email: bingkai.w@gmail.com

Funding information
NIH, Grant/Award Numbers: P30AG072976, U54AG065181, R01EB029977, P41EB031771, U54DA049110

Abstract

A compositional tree refers to a tree structure on a set of random variables where each random variable is a node and composition occurs at each non-leaf node of the tree. As a generalization of compositional data, compositional trees handle more complex relationships among random variables and appear in many disciplines, such as brain imaging, genomics and finance. We consider the problem of sparse regression on data that are associated with a compositional tree and propose a transformation-free tree-based regularized regression method for component selection. The regularization penalty is designed based on the tree structure and encourages a sparse tree representation. We prove that our proposed estimator for regression coefficients is both consistent and model selection consistent. In the simulation study, our method shows higher accuracy than competing methods under different scenarios. By analysing a brain imaging data set from studies of Alzheimer's disease, our method identifies meaningful associations between memory decline and volume of brain regions that are consistent with current understanding.

KEYWORDS

composition, hierarchical tree, regularized regression

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1 | INTRODUCTION

Compositional data refer to a type of data where data points are non-negative and the data vector of each subject or observational unit sums up to one. Compositional data appear in a wide range of disciplines, such as econometrics (Mullahy, 2015), geology (Pawlowsky-Glahn & Egozcue, 2006) and epidemiology (Leite, 2016).

In the area of brain imaging, structural magnetic resonance imaging (MRI) and anatomical brain segmentation produce compositional data. For example, using three-dimensional images acquired via structural MRI, a five-level brain segmentation introduced by Mori et al. (2016) can partition the whole brain into regions at five granularity levels. At the most coarse level, the whole brain is segmented into telencephalon (left and right), diencephalon (left and right), metencephalon, mesencephalon and cerebrospinal fluid (CSF). At the finest level of the brain segmentation, the whole brain is segmented into 236 brain regions. The compositional data are then the fractional volumes of the 236 brain regions relative to the intracranial volume (ICV).

In addition to composition, the volumetric data have a tree structure. In the first step of segmentation, the whole brain is partitioned into seven brain regions. In the second step, each of the seven brain regions created by the first step is further partitioned into smaller regions, which can be thought of as tree branching. Applied to all brain segmentation steps, this analogy makes a tree structure that is rooted at the whole brain and has 236 leaves, which are the brain regions at the finest segmentation. The tree structure has 321 nodes in total, and is shown in Figure 1. A key feature of this tree structure is that the volume of a brain region is equal to the combined volume of its subregions (after one segmentation), which introduces extra composition among variables. We refer to this data structure as 'compositional tree'. We note that the structure of compositional data is a special case of compositional trees, which only have leaves and a root.

Compositional trees appear in many applications. For example, Wang and Zhao (2017b) presented a compositional tree of microbiome data, where the compositional tree is formed by bacterial taxa at multiple taxonomic levels. Another example is the fractional market capitalization of stocks in the S&P 500 index (relative to the total market capitalization of S&P 500), where all 500 stocks are partitioned into 11 sectors and each sector is further broken down into industries according to the Global Industry Classification Standard (MSCI, 2020). The fractional market capitalization of a sector (or industry) is the summation of fractional market capitalization of stocks that are categorized into this sector (or industry). Compared with compositional data, compositional trees provide more information about the relationships among variables and suggest grouping effects at different levels.

Although methods for analysing compositional data or tree-structured data have been developed, little is known about how to deal with compositional trees. Our goal is to study the association between an outcome of interest and covariates, where the covariates have a compositional tree structure and the dimension of covariates is large. In the MRI application, the outcome is the memory score of subjects from studies of Alzheimer's disease (AD), and the covariates are the volumetric MRI data of these subjects. Understanding the association between memory and volumes of brain regions can help with diagnosis and treatment for the AD. We provide a detailed description of the data in Section 2.

Lin et al. (2014), Fiksel et al. (2020) and Ma and Zhang (2020) studied regression methods for compositional data with or without regularization, but their results cannot be directly generalized to handle compositional trees. Kim and Xing (2012) proposed a tree lasso for estimating a sparse multiresponse regression function, which did not consider compositional data.

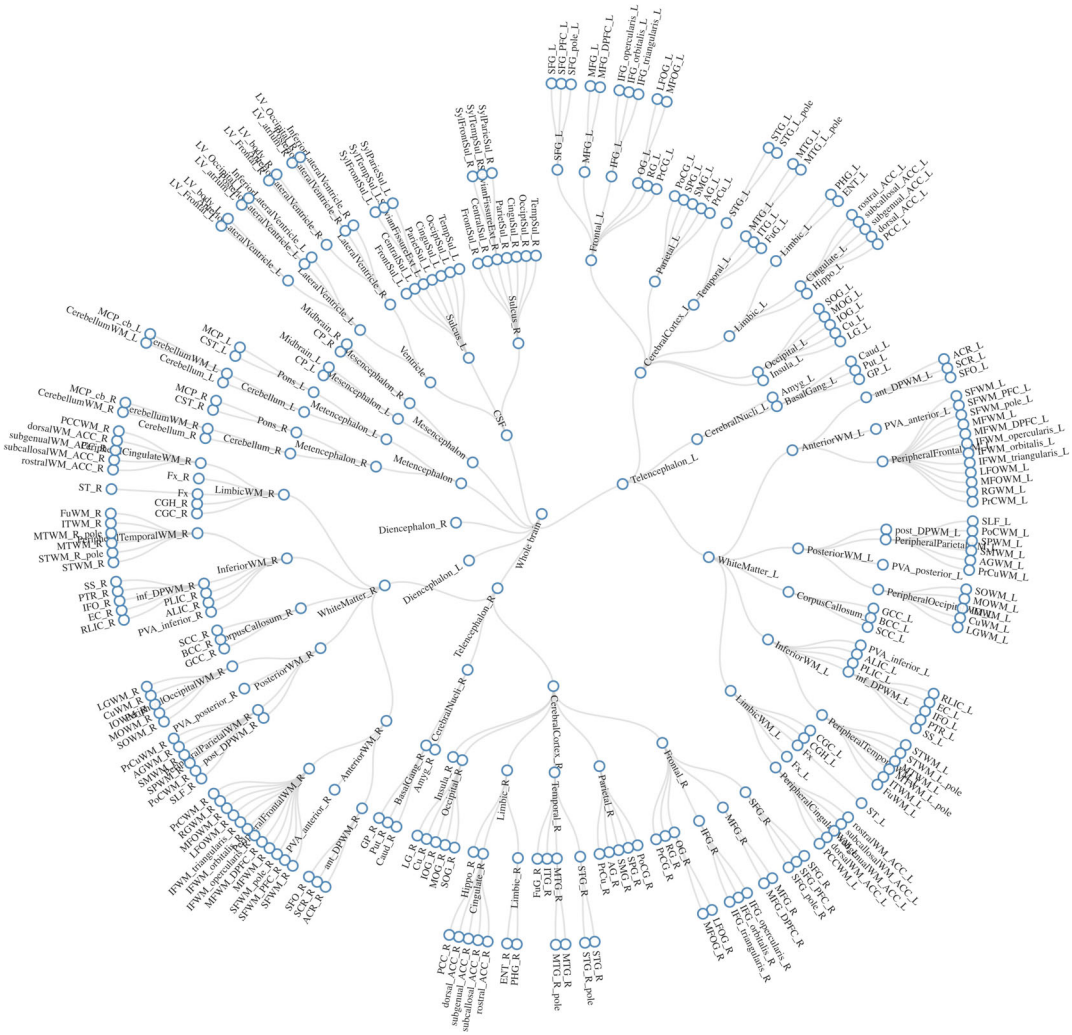

FIGURE 1 The compositional tree structure of the magnetic resonance imaging data example in a radial shape. Each blue circle represents a brain region (a node of the tree). The tree is rooted at the whole brain (center of the figure). Each grey curved segment connects two nodes, where the node closer to root represents the parent node, and the other is the child node. Each leaf node is connected by only one curved segment. A brain region with suffix '_L' or '_R' indicates that the region is in the left or right hemisphere of the brain [Colour figure can be viewed at wileyonlinelibrary.com]

More recently, Yan and Bien (2021) developed a tree-guided regularized regression method that aggregates rare features to improve the accuracy of prediction; this method, however, also focused on non-compositional data and cannot directly apply to compositional trees. To the best of our knowledge, two primary competitive works are Wang and Zhao (2017a, b), which developed tree-guided regularization methods for predictive feature construction and structured subcomposition selection, respectively. Both works focused on a tree structure with composition on leaves, which is different from the compositional tree, where composition exists at each node of the tree. Furthermore, neither work covered asymptotic properties, and Wang and Zhao (2017b) cannot handle the issue raised by boundary points (zeros or ones) in the data.

In this paper, we propose a regularized regression method to estimate the association between an outcome and covariates that have a compositional tree structure. The regularization term is constructed from the tree structure, which is assumed to be known, and designed to achieve sparsity in both marginal and conditional effects from covariates. Our model is transformation-free and able to handle boundary points (zeros or ones) in the data. We also establish consistency and model selection consistency of our estimators building on results from Lee et al. (2015).

In the next section, we introduce an MRI data example. In Section 3, we define the compositional tree and regression model. In Section 4, we present our proposed method to estimate the regression coefficients. We evaluate the performance of our proposed method through simulations in Section 5. The MRI data application is provided in Section 6. Section 7 discusses future directions.

2 | DATA EXAMPLE

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. We focus on the data set acquired by Liu et al. (2019) from the ADNI database.

The ADNI data set contains 819 subjects, who were diagnosed at the baseline as cognitive normal (229 subjects), MCI (402 subjects) or AD (188 subjects). For each subject, the composite memory score, MRI data, and a wide variety of demographic, behavioural and other non-imaging covariates were collected at several time points. We focus on the composite memory score and MRI data. The composite memory score was measured using data from the ADNI neuropsychological battery and validated by Crane et al. (2012), with higher scores indicating better memory. The MRI data consist of high-resolution T1-weighted images, which are pre-processed and segmented through MRICloud (https://www.MRICloud.org, Mori et al., 2016), a public platform for multi-contrast imaging segmentation and quantification. The pre-processing steps included orientation, inhomogeneity correction and histogram matching following with large deformation diffeomorphic metric mapping (LDDMM), and were described in detail in Glasser et al. (2013). After pre-processing, segmentations were obtained by fusing the multiatlas labelling method (Tang et al., 2013).

For the MRI data, the five-level brain segmentation defines 321 brain regions (aggregating brain regions from all five levels), which form a tree structure. At the first level of brain segmentation, the whole brain is partitioned into seven brain regions. At the second level of brain segmentation, each of the seven brain regions is further segmented into smaller regions. At the finest level, there are 236 brain regions. Figure 1 displays the tree structure of the 321 brain regions. For each brain region, we extracted its volume. Based on the five-level brain segmentation procedure, the volume of a brain region is equal to the combined volume of its subregions (after one segmentation). Furthermore, the combined volume of brain regions at the finest level is equal to the ICV.

Structural MRI data have been commonly used to identify biomarkers of AD (Vemuri & Jack, 2010). For example, the density of neurofibrillary tangles is an established pathological hallmark of AD, which can be reflected by MRI. We hence focus on the association between memory

decline, a common symptom of AD, and brain volumes. For each subject, we use the MRI data acquired at the initial screening, i.e. first time point, and the composite memory score acquired on the same day as the MRI scan or the first post-imaging measurement. We note that the response is the memory score measured at the first time point, instead of the change of memory between two time points.

3 | MODEL AND ASSUMPTIONS

3.1 | Compositional tree

We first define tree structure using notation from graph theory. Let $V = \{X_1, \ldots, X_p\}$ be a set of random variables with $0 \le X_j \le 1$ for $j = 1, \ldots, p$. Let E be a set of directed edges among $X_1, \ldots, X_p$ with $E \subset \{(X_j \to X_k) : X_j, X_k \in V\}$. For each edge $(X_j \to X_k) \in E$, we call X_j the parent of X_k, and X_k the child of X_j. X_j is a leaf node if it has no child and a root node if it has no parent. X_j is an ancestor of X_k if the directed edges in E can form a directed path from X_j to X_k, for example, $(X_j \to X_s), (X_s \to X_k) \in E$.

Definition 1 (V, E) forms a tree if (1) no X_j is an ancestor of itself (i.e., E not containing any directed cycle), (2) V contains only one root node and (3) each X_j has at most one parent.

In Definition 1, condition (1) defines a directed acyclic graph, and conditions (2) and (3) are often made in defining a rooted tree in graph theory. Figure 2 gives an example of tree structure with $V = \{X_1, \ldots, X_{10}\}$ and $E = \{(X_{10} \to X_9), (X_{10} \to X_8), (X_9 \to X_1), (X_9 \to X_7), (X_7 \to X_2), (X_7 \to X_3), (X_8 \to X_4), (X_8 \to X_5), (X_8 \to X_6)\}$.

In our data example, we can define a tree given the hierarchical brain segmentation. Let each X_j, $j = 1, \ldots, 321$, represent the volume of a brain region j and let V be the set of all X_j, where 321 is the total number of nodes in the tree structure shown in Figure 1. We regard X_j as the parent of X_k if brain region k is a subregion of j defined by one-step segmentation (i.e. there is no other subregion of j that contains k). If X_j is the parent of X_k, we also call brain region j the parent of brain region k. Then the edge set E is defined as the collection of all parent-child relationships among brain regions and the (only) root node is the ICV. For the (V, E) defined above, condition (1) of Definition 1 holds by construction, condition (2) follows because the root node is the ICV and condition (3) results from the fact that a region cannot be part of two disjoint bigger regions.

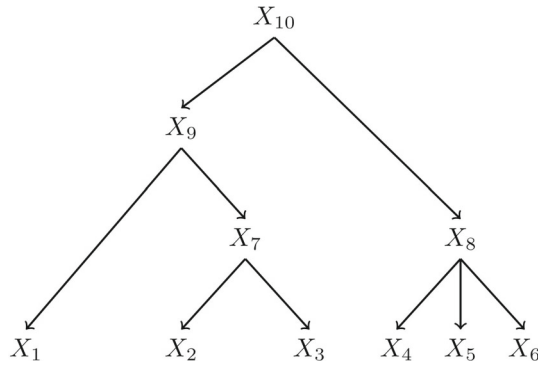

FIGURE 2 An example of a tree with $p = 10$

Although we define the tree structure using notations of graph theory, we emphasize that we do not associate the tree structure with conditional independence or causal diagrams, such as in graphical probabilistic models (Pearl, 2009). Our tree solely represents the hierarchical structure among $X_1, \ldots, X_p$ and is used to add compositional constraints, as described below. Our goal is to study the association between an outcome of interest and covariates X_j, $j = 1, \ldots, p$, instead of the relationships among covariates.

Consider compositional constraints on $(X_1, \ldots, X_p)$ complying with the tree structure. Denoting q as the number of leaf nodes, we can arrange the indices of $X_1, \ldots, X_p$ such that the first q variables $(X_1, \ldots, X_q)$ are the leaf nodes. For each $j = 1, \ldots, p$, let $c(j) = \{k : (X_j \to X_k) \in E\}$ denote the index set of children of X_j and let $|c(j)|$ denote the cardinality of $c(j)$ (i.e. the number of children of X_j). We then have the following definition of a compositional tree.

Definition 2 Assume (V, E) forms a tree and $X_1, \ldots, X_q$ are the leaf nodes. Then (V, E) forms a compositional tree if (1) $\sum_{j=1}^{q} X_j = 1$ and (2) $X_j = \sum_{k \in c(j)} X_k$ for each $j > q$.

In Definition 2, condition (1) imposes a compositional constraint on the leaf nodes. Condition (2) requires that each parent node is equal to the summation of its children. Conditions (1) and (2) together imply that the root node is a constant 1. In the example shown in Figure 2, the constraints for a compositional tree are $X_7 = X_2 + X_3$, $X_8 = X_4 + X_5 + X_6$, $X_9 = X_1 + X_7$ and $X_{10} = X_8 + X_9 = 1$. For the case that X_j has only one child X_k, Definition 2 implies that $X_j = X_k$ and we hence drop X_k to avoid any replicate. In this paper, we assume the compositional tree (V,E) for a column vector of random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ is known.

Compositional trees generalize compositional data by allowing more constraints on $\boldsymbol{X}$. Although $X_{q+1}, \ldots, X_p$ are linear combinations of leaf nodes, they still provide information on the structure of $\boldsymbol{X}$ and can help interpret conditional effects (defined in Section 3.2 below). To simplify notation, we say that $\boldsymbol{X}$ has a compositional tree structure if the associated (V, E) forms a compositional tree.

In our data example, brain regions $\boldsymbol{X}$ defined above have a compositional tree structure. Since the summation of all leaf node volumes is the ICV, then condition (1) of Definition 2 requires that the volumetric data are normalized by the ICV such that each person has a total brain volume 1. This is common practice in MRI analysis, since the ICV is typically only meaningfully related to physical size. Alternative strategies remove ventricular volumes and then study regional volumes relative to total brain volume (i.e. studying tissue composition). We include the ventricular volumes and normalize by ICV, since they are an important aspect of understanding progressive tissue loss in a disorder like AD. Condition (2) of Definition 2 states that the volume of each brain region is equal to the combined volume of all its children, since, by definition, each brain region is partitioned into subregions with no volume left undefined.

When dealing with compositional data, most current models work on a transformed space, for example isometric log ratio transformations (Egozcue et al., 2003) or log ratio transformations (Papke & Wooldridge, 1996). Although such transformations provide convenience in estimation, they cannot handle boundary values in $\boldsymbol{X}$ and add difficulty to interpretation (Fiksel et al., 2020). We hence work on the original space $\{\boldsymbol{X} : X_j \geq 0, j = 1, \ldots, p; \sum_{j=1}^{q} X_j = 1; X_j = \sum_{k \in c(j)} X_k, j = q + 1, \ldots, p\}$.

For a compositional tree, the vector space spanned by $\boldsymbol{X}$ has dimension (at most) $q < p$, which causes rank deficiency in many regression models. An alternative way is to model $X_j = \sum_{k \in c(j)} X_k + \varepsilon_j$, where ε_j is an independent Gaussian noise, following the method of Shojaie and Michailidis (2010). Although this method does not have the issue of rank deficiency, as long as the

covariance matrix of $(\varepsilon_1, \ldots, \varepsilon_p)$ is positive definite, we have $\varepsilon_j = 0$ almost always, which violates Gaussian modelling assumptions.

3.2 | Linear model, parameter identifiability and interpretation

Let Y be the outcome of interest. We assume the following linear model

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon = \boldsymbol{\beta}^\top \boldsymbol{X} + \varepsilon, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a column vector of unknown parameters, $\boldsymbol{X}$ has a compositional tree structure with q leaf nodes, and $\varepsilon \sim N(0, \sigma^2)$ is independent of $\boldsymbol{X}$. Since the root node of a compositional tree is a constant 1 and included in $\boldsymbol{X}$, the intercept term is omitted from model (1). For $i = 1, \ldots, n$, let $(\boldsymbol{X}_i, \varepsilon_i)$ be independent, identically distributed samples from the joint distribution of $(\boldsymbol{X}, \varepsilon)$ and let $Y_i = \boldsymbol{\beta}^\top \boldsymbol{X}_i + \varepsilon_i$.

Since $\boldsymbol{X}$ is rank deficient (with rank at most q), $\boldsymbol{\beta}$ is not unique. Due to this fact, each $\beta_j, j = 1, \ldots, p$ is not interpretable without further assumptions. To overcome this difficulty, we impose the following $p - q$ linear constraints on $\boldsymbol{\beta}$:

$$\sum_{k \in c(j)} \beta_k = 0 \text{ for all } j > q, \tag{2}$$

which uniquely define a $\boldsymbol{\beta}$ (as shown in the Supplementary Material). Linear constraints (2) require that, for each X_j that is not a leaf node, the average effect of its children on Y is 0. Then, each β_k can be interpreted as the deviation effect of X_k from the effect of its parent, X_j, on Y. To show this, consider the following derivation using Definition 2:

$$\beta_j X_j + \sum_{k \in c(j)} \beta_k X_k = \left(\beta_j + \frac{1}{|c(j)|} \sum_{l \in c(j)} \beta_l \right) X_j + \sum_{k \in c(j)} \left(\beta_k - \frac{1}{|c(j)|} \sum_{l \in c(j)} \beta_l \right) X_k,$$

which implies that the average coefficient of children of X_j can be absorbed into the coefficient of X_j and hence the remaining coefficients of $X_k, k \in c(j)$ are the deviations from X_j. By repeating this procedure recursively from leaf nodes to the root node, we get all coefficients satisfying linear constraints (2) with the desired interpretation. For conciseness, β_k is referred to as the 'conditional deviation effect' throughout, since its interpretation is conditioning on the parent of X_k, that is the parent of X_k held constant.

Let X_p denote the root node and $a(j)$ be the index set of ancestors of X_j. Then the linear model (1) with constraints (2) can be formulated as:

$$Y = \beta_p + \sum_{j=1}^{q} \alpha_j X_j + \varepsilon = \beta_p + \boldsymbol{\alpha}^\top \boldsymbol{X}_{leaf} + \varepsilon, \tag{3}$$

$$\text{subject to } \sum_{j=1}^{q} \alpha_j = 0, \tag{4}$$

where β_p is the regression coefficient of the root node X_p and serves as the intercept, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ with $\alpha_j = \beta_j + \sum_{k \in a(j) \setminus \{p\}} \beta_k$, and $\boldsymbol{X}_{leaf} = (X_1, \ldots, X_q)^\top$ is the vector of leaf nodes. We

assume that the only linear constraint on $\boldsymbol{X}_{leaf}$ is $\sum_{j=1}^{q} X_j = 1$, that is no component of $\boldsymbol{X}_{leaf}$ being a linear combination of the others. Model (3) not only provides direct interpretation of marginal associations between Y and $\boldsymbol{X}_{leaf}$ (which we introduce below), but is also useful for estimating $\boldsymbol{\beta}$ in Section 4.

Compared to model (1), model (3) only uses the leaf nodes. Each α_j is the aggregation of the conditional deviation effects of X_j and ancestors of X_j excluding the root node. If X_j is increased by δ at the expense of another leaf node, X_k, that is X_k decreased by δ, then Y is changed by $(\alpha_j - \alpha_k)\delta$. Hence, $\alpha_j - \alpha_k$ is interpreted as the relative effect between X_j and X_k. If X_j is relatively increased by δ at the expense of all other leaf nodes evenly, that is X_j increased by $(1 - 1/q)\delta$ and X_k decreased by δ/q for all $k \leq q$, $k \neq q$ such that $X_j - X_k = \delta$, then Y is changed by, using the constraint (4),

$$\alpha_j(X_j + \frac{q-1}{q}\delta) + \sum_{1\leq k\leq q, k\neq j} \alpha_k(X_k - \frac{1}{q}\delta) - \sum_{k=1}^{q} \alpha_k X_k = \alpha_j \delta. \tag{5}$$

This implies that α_j can be interpreted as the effect of X_j relative to all leaf nodes, and we refer to α_j as the 'marginal deviation effect' throughout for conciseness. Without the constraint (4), α_j is not identifiable, since $\sum_{j=1}^{q} X_j = 1$. However, $\beta_p + \alpha_j$ would still be identifiable, a fact that we use for estimation in Section 4. This statement is formally described in Proposition 1, which is proven in the Supplementary Material.

Proposition 1 *Assume that* $\sum_{j=1}^{q} X_j = 1$ *and no component of* $\boldsymbol{X}_{leaf}$ *is a linear combination of the others. If two sets of parameters* $(\beta_p, \alpha_1, \ldots, \alpha_q)$ *and* $(\beta_p, \alpha_1, \ldots, \alpha_q)$ *both satisfy model (3), then* $\beta_p + \alpha_j = \beta_p + \alpha_j$ *for each* $j = 1, \ldots, q$.

Compared with a classical linear regression model where the design matrix has full rank, linear models on compositional trees require more careful and subtle interpretation of the coefficients. The marginal deviation effect, α_j, is not the association between X_j and Y, but the relative effect of X_j on Y compared with the rest leaf nodes as demonstrated in Equation (5). Such an interpretation of α_j is common for all linear models on compositional data. The conditional deviation effect β_j is not the association between X_j and Y either, but the relative effect of X_j on Y compared with the 'siblings' of X_j, conditioning on the parent of X_j. In our data example, both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are scientifically meaningful. The marginal deviation effect represents the effect of the fractional volume of a leaf region on memory, while the conditional deviation effect is the residual effect of the fractional volume of a brain region on memory after removing the effect of its ancestors.

4 | ESTIMATION

Let $\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*$ denote the true parameters that satisfy model (1) with constraints (2) and model (3) with constraint (4), respectively. Our goal is to estimate $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$. Since p and q are potentially large ($p = 321$ and $q = 236$ for our data example), we propose a new regularization term based on lasso for component selection (Lin et al., 2014) and fused lasso (Tibshirani et al., 2005) and perform regularized regression to achieve sparsity in both $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. In the method described below, we first estimate $\boldsymbol{\alpha}^*$ using the generalized lasso (Tibshirani & Taylor, 2011) and then calculate $\hat{\boldsymbol{\beta}}$ based on $\hat{\boldsymbol{\alpha}}$ by solving linear systems. We call our method for estimating $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ the compositional tree-guided LASSO (CT-LASSO).

4.1 | Regularization

For any $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\alpha} \in \mathbb{R}^q$, consider the regularization term

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta) = \eta P_1(\boldsymbol{\alpha}) + (1-\eta) P_2(\boldsymbol{\beta}),$$

where $\eta \in [0, 1]$ is a tuning parameter adjusting the weight between $P_1(\boldsymbol{\alpha})$ and $P_2(\boldsymbol{\beta})$,

$$P_1(\boldsymbol{\alpha}) = \sum_{j=1}^{q} \left| \alpha_j - \frac{1}{q} \sum_{k=1}^{q} \alpha_k \right|,$$
$$P_2(\boldsymbol{\beta}) = \sum_{j=q+1}^{p} \sum_{s=1}^{|c(j)|-1} |\beta_{j_s} - \beta_{j_{s+1}}|,$$

where $c(j)$ is the index set of children of X_j with the elements in $c(j)$ encoded as $j_1, \ldots, j_{|c(j)|}$. $P_1(\boldsymbol{\alpha})$ selects leaf nodes with non-zero marginal deviation effects. If $\alpha_j = \frac{1}{q}\sum_{k=1}^{q} \alpha_k$, then changing X_j at the expense of all other leaf nodes evenly will not result in changes of Y. This penalty is known as the lasso for component selection, and is also seen in Wang and Zhao (2017b) for dealing with compositional data. In $P_2(\boldsymbol{\beta})$, for each X_j with $j > q$, we penalize the difference among coefficients of its children using the fused lasso penalty. If $|\beta_{j_s} - \beta_{j_{s+1}}| = 0$ for all $s = 1, \ldots, |c(j)| - 1$, which means all children of X_j have no conditional deviation effect, then the component $\sum_{k \in c(j)} \beta_k X_k = \beta_{j_1} X_j$, resulting in a sparse representation of linear model (1). Combined with the linear constraints (2), the above case is also equivalent to $\beta_k = 0$ for all $k \in c(j)$. $P_2(\boldsymbol{\beta})$ also shares the expression with the regularization terms of 'tree-guided fused lasso 2' (TFL-2) by Wang and Zhao (2017a), which was designed to construct predictive features for compositional microbiome data. The following proposition gives some properties of $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta)$.

Proposition 2 *Given linear constraints (2), there exists a matrix* $\boldsymbol{D}(\eta) \in \mathbb{R}^{(2q-1)\times q}$ *such that* $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta) = \|\boldsymbol{D}(\eta)\boldsymbol{\alpha}\|_1$ *and* $\boldsymbol{D}(\eta)\mathbf{1}_q = \mathbf{0}_q$, *where* $\|\cdot\|_1$ *is the* L_1-norm, $\mathbf{1}_q, \mathbf{0}_q \in \mathbb{R}^q$ *are column vectors with all entries* 1, 0 *respectively*.

Proposition 2 implies that the penalty $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta)$ can be formulated as a function of $\boldsymbol{\alpha}$ and η, making it possible to perform regularized regression based on model (3), which does not involve $\boldsymbol{\beta}$. Furthermore, this penalty is invariant with respect to constant change of $\boldsymbol{\alpha}$ (i.e. $\|\boldsymbol{D}(\eta)\boldsymbol{\alpha}\|_1 = \|\boldsymbol{D}(\eta)(\boldsymbol{\alpha} + C\mathbf{1}_q)\|_1$ for any $C \in \mathbb{R}$), which makes it equivalent to penalize $\boldsymbol{\alpha} + \beta_q \mathbf{1}_q$ as we do in Section 4.2 below. We prove Proposition 2 and show how $\boldsymbol{D}(\eta)$ is constructed in the Supplementary Material.

For the regularization terms on $\boldsymbol{\beta}$, we also consider the pairwise fused lasso (She, 2008), which is defined as

$$\tilde{P}_2(\boldsymbol{\beta}) = \sum_{j=q+1}^{p} \sum_{1 \le s_1 < s_2 \le c(j)} \left| \beta_{j_{s_1}} - \beta_{j_{s_2}} \right|.$$

We refer to our method with $P_2(\boldsymbol{\beta})$ substituted by $\tilde{P}_2(\boldsymbol{\beta})$ as 'CT-LASSO-p'. Compared with $P_2(\boldsymbol{\beta})$, $\tilde{P}_2(\boldsymbol{\beta})$ adds more terms to penalize the difference of the conditional deviation effects and further encourages the conditional deviation effects to be identical. Wang and Zhao (2017a), however, pointed out that the pairwise fused lasso suffers from instability and long computation time due to the large number of rows in the regularization matrix when using the 'genlasso' R package. For

compositional trees, this could happen when a node has many children. Our simulation study in Section 5 shows that CT-LASSO and CT-LASSO-p yield similar performance, while CT-LASSO is twice faster than CT-LASSO-p.

In the regularized regression method by Yan and Bien (2021), the penalty is $\eta\sum_{j=1}^{q} w_j|\alpha_j| + (1-\eta)\sum_{j=q+1}^{p-1} w_j|\beta_j|$, where w_j are pre-specified weights. This penalty differs from $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \eta)$ on two aspects. First, our $P_1(\boldsymbol{\alpha})$ penalizes the deviation of α_j from their average, which is designed for compositional data and different from $\sum_{j=1}^{q} w_j|\alpha_j|$. Second, $\sum_{j=q+1}^{p-1} w_j|\beta_j|$ penalizes the magnitude of the conditional deviation effects, while our $P_2(\boldsymbol{\beta})$ encourages the conditional deviation effects to be equal even when they are non-zero; hence, the latter regularization term has the advantage to group non-zero conditional deviation effects, which facilitates the interpretation of compositional tree models.

4.2 | Estimating $\boldsymbol{\alpha}^*$

We estimate $\boldsymbol{\alpha}^*$ by $\hat{\boldsymbol{\alpha}} = \hat{\tilde{\boldsymbol{\alpha}}} - \mathbf{1}_q\mathbf{1}_q^\top\hat{\tilde{\boldsymbol{\alpha}}}$, where

$$\hat{\tilde{\boldsymbol{\alpha}}} = \arg\ \min_{\tilde{\boldsymbol{\alpha}}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \tilde{\boldsymbol{\alpha}}^\top \boldsymbol{X}_{leaf,i}\right)^2 + \lambda\|\boldsymbol{D}(\eta)\tilde{\boldsymbol{\alpha}}\|_1 \tag{6}$$

with $\boldsymbol{X}_{leaf,i} = (X_{i1}, \ldots, X_{iq})^\top$, $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + \beta_p\mathbf{1}_q$ and $\lambda > 0$ being the tuning parameter. In Equation (6), $\hat{\tilde{\boldsymbol{\alpha}}}$ is an estimate of $\tilde{\boldsymbol{\alpha}}$, which is identifiable as discussed in Section 3.2 and does not involve any linear constraints. Then, $\hat{\boldsymbol{\alpha}}$ is constructed by imposing the constraint (4), that is centring $\hat{\tilde{\boldsymbol{\alpha}}}$. We note that the regularization term $\lambda\|\boldsymbol{D}(\eta)\tilde{\boldsymbol{\alpha}}\|_1$ imposes the desired sparsity on $\boldsymbol{\alpha}$, since $\lambda\|\boldsymbol{D}(\eta)\tilde{\boldsymbol{\alpha}}\|_1 = \lambda\|\boldsymbol{D}(\eta)\boldsymbol{\alpha}\|_1$ given Proposition 2.

For any $\boldsymbol{\alpha} \in \mathbb{R}^q$ and given $\eta \in [0, 1]$, let $\mathcal{S}(\boldsymbol{\alpha})$ be the support of $\boldsymbol{D}(\eta)\boldsymbol{\alpha}$, i.e., $\mathcal{S}(\boldsymbol{\alpha}) = \{j \in \{1, \ldots, 2q-1\} : \boldsymbol{e}_j^\top\boldsymbol{D}(\eta)\boldsymbol{\alpha} \neq 0\}$ with $\boldsymbol{e}_j \in \mathbb{R}^{2q-1}$ being a column vector with the j-th entry 1 and the rest 0. Let $\mathcal{M} = \{\boldsymbol{\alpha} : \mathcal{S}(\boldsymbol{\alpha}) \subset \mathcal{S}(\boldsymbol{\alpha}^*)\}$ denote the model subspace of interest. That is, for $\boldsymbol{\alpha} \in \mathcal{M}$, an entry of $\boldsymbol{D}(\eta)\boldsymbol{\alpha}$ is non-zero only if the corresponding entry of $\boldsymbol{D}(\eta)\boldsymbol{\alpha}^*$ is non-zero. The following theorem, adapted from Corollary 4.2 of Lee et al. (2015), gives consistency and model selection consistency of $\hat{\boldsymbol{\alpha}}$.

Theorem 1 *Given $\eta \in [0,1]$, we assume $\{\boldsymbol{X}_{leaf,i}\}_{i=1}^n$ satisfies restricted strong convexity (RSC) on $\mathcal{M}$ and irrepresentability, which we define in the Supplementary Material. For $\lambda = C_1\sigma\sqrt{\frac{\log q}{n}}$, $\hat{\boldsymbol{\alpha}}$ is unique and, with probability at least $1 - 2/q$,*

(a) *(consistency)* $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|_2 \leq C_2\sigma\sqrt{\frac{\log q}{n}}$,
(b) *(model selection consistency)* $\hat{\boldsymbol{\alpha}} \in \mathcal{M}$,

where $\|\cdot\|_2$ is the L_2-norm and C_1, C_2 are known constants given in the Supplementary Material.

Theorem 1 implies that when q and $n/\log(q)$ are large, then, with high probability, our estimate $\hat{\boldsymbol{\alpha}}$ is close to the truth and does not contain false positives (non-zero effect of inactive predictors with respect to $\boldsymbol{D}(\eta)$). The RSC assumption is typically satisfied when $X_{leaf,i}$ follows a multivariate normal distribution (Raskutti et al., 2010). The irrepresentability assumption requires that the active predictors (with respect to $\boldsymbol{D}(\eta)$) are orthogonal or nearly-orthogonal to the inactive

predictors (Lee et al., 2015). We provide a detailed description and discussion of these assumptions in the Supplementary Material.

Given η, the optimization problem (6) can be solved by the *genlasso* package (Tibshirani & Taylor, 2011) in R software. To select the tuning parameter λ, we propose to use the Akaike information criterion (AIC, Akaike et al., 1998) or Bayesian information criterion (BIC, Schwarz, 1978). Let

$$IC_\gamma(\eta, \lambda) = n \log \left\{ \sum_{i=1}^{n} \left(Y_i - \hat{\tilde{\boldsymbol{\alpha}}}^\top \boldsymbol{X}_{leaf,i} \right)^2 \right\} + \gamma \, \mathrm{df}(\eta, \lambda),$$

where γ is a complexity factor, $\mathrm{df}(\eta, \lambda)$ is the effective number of parameters in $\hat{\tilde{\boldsymbol{\alpha}}}$. $IC_\gamma(\eta, \lambda)$ refers to AIC if $\gamma = 2$ and BIC if $\gamma = \log(n)$. For any $\eta \in [0, 1]$, define $\hat{\lambda}(\eta) = \arg\min_{\lambda \geq 0} IC_\gamma(\eta, \lambda)$. We select the tuning parameters $\hat{\eta} = \arg\min_{\eta \in [0,1]} IC_\gamma(\eta, \hat{\lambda}(\eta))$ and $\hat{\lambda} = \hat{\lambda}(\hat{\eta})$. An alternative method to tune parameters is cross-validation, but we do not consider it here, since it would dramatically increase the computation complexity and performs similarly to AIC.

4.3 | Estimating $\boldsymbol{\beta}^*$

Given $\hat{\boldsymbol{\alpha}}$, we calculate $\hat{\boldsymbol{\beta}}$ as follows. Since $\alpha_j = \beta_j + \sum_{k \in a(j)} \beta_k - \beta_p$ for $j = 1, \ldots, q$, we can construct a matrix $\boldsymbol{Q}_1 \in \mathbb{R}^{q \times p}$ such that $\boldsymbol{Q}_1 \boldsymbol{\beta} = \boldsymbol{\alpha}$. Since $\boldsymbol{\beta}$ also satisfies linear constraints (2), we can construct another matrix $\boldsymbol{Q}_2 \in \mathbb{R}^{(q-p) \times p}$ such that $\boldsymbol{Q}_2 \boldsymbol{\beta} = \boldsymbol{0}_{p-q}$. Denoting $\boldsymbol{Q} = (\boldsymbol{Q}_1^\top, \boldsymbol{Q}_2^\top)^\top$, then $\hat{\boldsymbol{\beta}}$ is calculated by solving the linear system

$$\boldsymbol{Q}\boldsymbol{\beta} = \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \boldsymbol{0}_{p-q} \end{pmatrix}. \tag{7}$$

The following theorem implies that $\hat{\boldsymbol{\beta}}$ is uniquely determined by $\hat{\boldsymbol{\alpha}}$ (i.e. $\boldsymbol{Q}$ is invertible) and is consistent and model selection consistent under the same conditions as $\hat{\boldsymbol{\alpha}}$.

Theorem 2 Let C_1, C_2, λ and $\mathcal{M}$ be the quantities defined in Theorem 1. Given the same assumptions made in Theorem 1, $\hat{\boldsymbol{\beta}}$ is uniquely determined by $\hat{\boldsymbol{\alpha}}$ and, with probability at least $1 - 2/q$,

(a) (consistency) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{Q}^{-1}\|_2 C_2 \sigma \sqrt{\frac{\log q}{n}}$,
(b) (model selection consistency) $\boldsymbol{Q}_1 \hat{\boldsymbol{\beta}} \in \mathcal{M}$.

An alternative method to estimate $\boldsymbol{\beta}$ is solving a constrained optimization problem following Lin et al. (2014):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}^\top \boldsymbol{X}_i \right)^2 + \lambda P(\alpha, \boldsymbol{\beta}, \eta),$$
$$\text{subject to} \quad \text{linear constraints (2)}.$$

However, this method has to handle the rank deficiency of $\boldsymbol{X}$ and $p - q$ linear constraints. If q is much smaller than p, then the number of linear constraints can be large, which may cause bias and increased computational complexity. In our proposed method, these two issues are avoided by using two steps to estimate $\boldsymbol{\beta}^*$ (first estimating $\boldsymbol{\alpha}^*$ and then $\boldsymbol{\beta}^*$).

To the best of our knowledge, we are the first to study regularized regression on a compositional tree and provide consistency and model selection consistency. Kim and Xing (2012) developed a tree-guided group lasso method, but their goal was to analyse multiresponse data and they did not consider composition. Lin et al. (2014) used the lasso for component selection in compositional data and their optimization problem is a special case of ours, setting $\eta = 1$. Wang and Zhao (2017a, b) performed penalized regression on compositional data with a hierarchical tree structure, but they did not consider compositional trees or provide asymptotic results.

5 | SIMULATION STUDY

5.1 | Simulation settings

In this simulation study, we consider four data generating distributions, which cover combinations of the following settings: a binary compositional tree or the MRI-motivated compositional tree, and leaf or stem effects. A binary compositional tree is a compositional tree where each parent has two children, while the MRI-motivated compositional tree represents the same tree structure as our data example (where a parent node may have more than two children). Leaf effects stand for linear models where the true effects (non-zero β_j) are only from nodes near the leaves, while stem effects mean that true effects are only from nodes near the root. Different from the leaf effects where both $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ are sparse, stem effects will lead to non-sparse $\boldsymbol{\alpha}^*$.

The first scenario (Scenario 1) has a binary compositional tree and leaf effects. The tree structure is shown in Cf Figure 3, where $p = 255$, $q = 128$ and $\boldsymbol{X}_{leaf} = (X_1, X_2, \ldots, X_q)$. Letting $n = 120$ (i.e. $n < q$), we independently generate $\boldsymbol{X}_{leaf,i}, i = 1, \ldots, n$ by first independently sampling $\widetilde{\boldsymbol{X}}_{leaf,i}$ from a multivariate log-normal distribution that is $\widetilde{\boldsymbol{X}}_{leaf,i} = \exp(\boldsymbol{Z}_i)$ with $\boldsymbol{Z}_i$ following a multivariate normal distribution. We assume $\boldsymbol{Z}_i$ has mean $\mathbf{0}_q$ and variance $\boldsymbol{\Sigma} = (\sigma_{ij})_{q\times q}$, where $\sigma_{ij} = 0.2^{|i-j|}$ is the (i,j)-th column entry of $\boldsymbol{\Sigma}$. We then define $\boldsymbol{X}_{leaf,i} = \widetilde{\boldsymbol{X}}_{leaf,i}/\mathbf{1}_q^\top \widetilde{\boldsymbol{X}}_{leaf,i}$ to satisfy the composition condition. For $j > q$, we generate X_{ij} following the definition of compositional trees using $\boldsymbol{X}_{leaf,i}$. We define, for $i = 1, \ldots, n$,

$$Y_i = 3 + X_{i,1} - X_{i,2} + X_{i,129} - X_{i,130} + \varepsilon_i = 3 + 2X_{i,1} - X_{i,3} - X_{i,4} + \varepsilon_i,$$

where ε_i is an independent sample from $N(0, \sigma^2)$ and σ^2 is chosen such that $Var(\boldsymbol{X}^\top\boldsymbol{\beta}) = Var(\varepsilon)$. This model only involves the left bottom corner in the tree shown in Figure 3. The non-zero conditional deviation effects are $\beta_1^* = \beta_{129}^* = 1, \beta_2^* = \beta_{130}^* = -1$ and the non-zero marginal deviation effects are $\alpha_1^* = 2, \alpha_3^* = \alpha_4^* = -1$.

The second scenario (Scenario 2) has a binary compositional tree and stem effects, where the binary compositional tree and $\boldsymbol{X}_i, i = 1, \ldots, n$ is the same as in Scenario 1. For the stem effect, we define

$$Y_i = 3 + X_{i,249} - X_{i,250} + X_{i,253} - X_{i,254} + \varepsilon_i = 3 + 2\sum_{j=1}^{32} X_{ij} - \sum_{j=65}^{128} X_{ij} + \varepsilon_i,$$

where ε_i is defined in the same way as in Scenario 1, and the second equality results from the compositional tree structure (Figure 3) which implies $X_{i,249} = \sum_{j=1}^{32} X_{ij}$, $X_{i,250} = \sum_{j=33}^{64} X_{ij}$, $X_{i,253} = \sum_{j=1}^{64} X_{ij}$, and $X_{i,254} = \sum_{j=65}^{128} X_{ij}$. Unlike Scenario 1, this model only involves the conditional deviation effects from the top part in the tree (nodes near the root), which are $\beta_{249}^* = \beta_{253}^* = 1$,

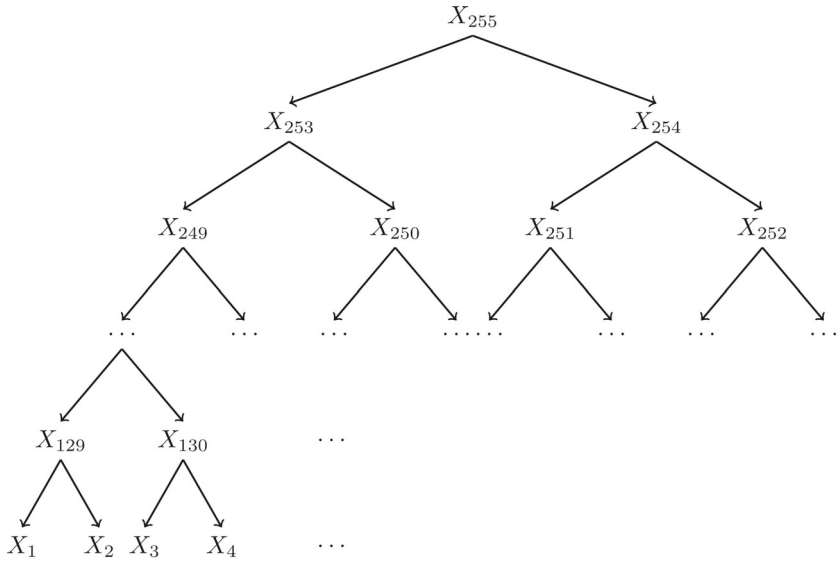

FIGURE 3 The binary compositional tree considered in Scenarios 1 and 2 of the simulation study with $p = 255$ and $q = 128$

$\beta^*_{250} = \beta^*_{254} = -1$. Furthermore, the marginal deviation effect $\boldsymbol{\alpha}$ is no longer sparse because $\alpha^*_j = 2$ for $j = 1, \ldots, 32$ and $\alpha^*_j = -1$ for $j = 65, \ldots, 128$.

In the third scenario (Scenario 3), we consider leaf effects, an MRI-motivated compositional tree, and covariate data that are re-sampled from the real MRI data. The MRI-motivated compositional tree is shown in Figure 1, where $p = 321$ and $q = 236$. For our MRI data example, $n = 819$ and we denote the empirical distribution of $(\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_n)$ as F_n, where $\widetilde{\boldsymbol{X}}_i$ contains the fractional brain volumetric data of participant i. Let $\boldsymbol{X}_i, i = 1, \ldots, n$ be independent samples from F_n. We model, for $i = 1, \ldots, n$,

$$Y_i = 3 + X_{i,\ \text{SFG-L}} - X_{i,\ \text{SFG-PFC-L}} + \varepsilon_i,$$

where SFG-L and SFG-PFC-L are leaf nodes and subregions of the superior frontal gyrus left hemisphere and ε is as defined in Scenario 1. In this model, we have $\beta^*_{\text{SFG-L}} = \alpha^*_{\text{SFG-L}} = 1$ and $\beta^*_{\text{SFG-PFC-L}} = \alpha^*_{\text{SFG-PFC-L}} = -1$.

In the last scenario (Scenario 4), we consider the MRI-motivated compositional tree again but with stem effects. We use the same compositional tree and $\boldsymbol{X}_i$, $i = 1, \ldots, n$ as in Scenario 3. Let

$$Y_i = 3 + X_{i,\ \text{Telencephalon-L}} - X_{i,\ \text{Telencephalon-R}} + \varepsilon_i,$$

where Telencephalon-L and Telencephalon-R represent the telencephalon located in the left and right hemisphere, respectively, and both are children of the ICV. Different from Scenario 3, this model has $\beta^*_{\text{Telencephalon-L}} = -1$, $\beta^*_{\text{Telencephalon-R}} = 1$ and 200 non-zero entries in $\boldsymbol{\alpha}^*$.

For each of the four scenarios, we simulate m=1000 data sets. For each data set, we implement our proposed methods (CT-LASSO and CT-LASSO-p), TFL-2 (Wang & Zhao, 2017a), TASSO (Wang & Zhao, 2017b), and lasso for component selection (abbreviated as CLASSO throughout, Lin et al., 2014). For TASSO, we use their default settings to estimate $\boldsymbol{\alpha}^*$ and calculate $\hat{\boldsymbol{\beta}}_{TASSO}$ by

solving Equation (7). The only difference is that the natural log-transformation is not performed, as described in Section 3. Since CLASSO (without log-transformation as we consider here) and TFL-2 are special cases of our CT-LASSO, we calculate $\hat{\boldsymbol{\beta}}_{CLASSO}$ and $\hat{\boldsymbol{\beta}}_{TFL-2}$ following the same procedure as CT-LASSO setting $\eta = 1$ and 0, respectively.

In our simulation, the covariates $\boldsymbol{X}_{leaf}$ are not log-transformed as in the simulations of Lin et al. (2014) and Wang and Zhao (2017b) for CLASSO and TASSO. This is because the MRI data, for example Scenarios 3 and 4, contain zeros due to measurement error. Hence log-transformation is not implemented when using CLASSO and TASSO. Such modification on the distribution of $\boldsymbol{X}_{leaf}$ does not harm the performance of CLASSO or TASSO, since they are implemented to capture the true relationship between Y and $\boldsymbol{X}_{leaf}$. In the simulation results of Scenarios 1 and 3 below, CLASSO and TASSO have high accuracy when using BIC tuning.

For all five methods, we use AIC or BIC to select the tuning parameters. The following metrics are used to compare their performances: (1) sensitivity, defined as $|\{j : \hat{\beta}_j \neq 0, \beta_j^* \neq 0\}|/|\{j : \beta_j^* \neq 0\}|$, (2) specificity, defined as $|\{j : \hat{\beta}_j = \beta_j^* = 0\}|/|\{j : \beta_j^* = 0\}|$ and (3) L_2-loss, defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. For each of the above metrics, we report its average and standard error over the m data sets. Since Scenarios 1 and 2 have $n < q$, a small L_2-penalty (0.0001) is added when solving the optimization problem (6).

5.2 | Simulation results

Table 1 gives the simulation results for Scenarios 1–4. We first compare CT-LASSO and CT-LASSO-p, both of which are our proposed methods but with different versions of the fused lasso. Across all scenarios with BIC tuning, CT-LASSO and CT-LASSO-p have high accuracy and similar performance. In Scenarios 1 and 2, CT-LASSO and CT-LASSO-p are identical since each non-leaf node has two children; in Scenarios 3 and 4, the regularization matrix $\boldsymbol{D}(\eta)$ has 471 rows for CT-LASSO and 847 rows for CT-LASSO-p, which leads to different selection of the tuning parameter η. In addition, for a single model fit on simulated MRI data using a 2.6 GHz CPU, CT-LASSO uses 7.3 s, while CT-LASSO-p uses 21.9 s. Given that CT-LASSO has comparable performance but shorter computation time than CT-LASSO-p in the MRI-based simulation, we use CT-LASSO for the MRI data application.

We next compare CT-LASSO with TFL-2, which is a special case of CT-LASSO with $\eta = 0$. In Scenarios 2 and 4 where the main effects come from nodes near the root, CT-LASSO and TFL-2 both perform well with BIC tuning, and CT-LASSO almost always selects $\hat{\eta} = 0$, leading to the similar performance of CT-LASSO and TFL-2. However, when the main effects come from nodes near the leaves as in Scenarios 1 and 3, TFL-2 fails to achieve sparsity in $\boldsymbol{\alpha}^*$ and is hence less accurate than CT-LASSO. In Scenario 1, TFL-2 has low sensitivity due to the failure of identifying all the true effects, and, in Scenario 3, TFL-2 has the largest L_2-loss because it falsely identifies noises in order to minimize the difference among β's.

Next, we compare the performance of CT-LASSO, TASSO and CLASSO. In Scenarios 1 and 3, the true parameter $\boldsymbol{\alpha}^*$ is sparse and all three methods perform well, as expected. CT-LASSO has slightly better performance than TASSO and CLASSO by penalizing both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, compared with penalization of only $\boldsymbol{\alpha}$ in TASSO or CLASSO. In Scenarios 2 and 4, since the true parameter, $\boldsymbol{\alpha}^*$, is not sparse, CT-LASSO outperforms the other two methods on all performance metrics. In such cases, the L_1 penalty on $\boldsymbol{\alpha}$ does not help. Hence, TASSO and CLASSO tend to over-penalize (low sensitivity, high specificity) or under-penalize (high sensitivity, low specificity) on $\boldsymbol{\alpha}$, either

TABLE 1 Simulation results for Scenarios 1–4 comparing CT-LASSO (our proposed method), CT-LASSO-p (our proposed method with pairwise fused lasso), TFL (Wang & Zhao, 2017a), TASSO (Wang & Zhao, 2017b) and CLASSO (LASSO for component selection). The numbers for sensitivity, specificity and L_2-loss are the averages and standard errors (in the parenthesis) over 1000 simulated data sets

	Method	Tuning	Sensitivity	Specificity	L_2-loss	η
Scenario 1	CT-LASSO	AIC	1.00 (0.02)	0.04 (0.02)	4.47 (0.51)	0.33 (0.38)
		BIC	0.98 (0.11)	0.97 (0.06)	0.76 (0.38)	0.50 (0.17)
	CT-LASSO-p	AIC	1.00 (0.02)	0.04 (0.02)	4.47 (0.51)	0.33 (0.38)
		BIC	0.98 (0.11)	0.97 (0.06)	0.76 (0.38)	0.50 (0.17)
	TFL-2	AIC	1.00 (0.02)	0.05 (0.02)	4.44 (0.51)	–
		BIC	0.77 (0.30)	0.95 (0.10)	1.22 (0.51)	–
	TASSO	AIC	1.00 (0.00)	0.04 (0.02)	4.27 (0.49)	–
		BIC	0.97 (0.17)	0.96 (0.07)	0.98 (0.39)	–
	CLASSO	AIC	1.00 (0.00)	0.04 (0.02)	4.35 (0.51)	–
		BIC	0.98 (0.12)	0.92 (0.10)	0.96 (0.41)	–
Scenario 2	CT-LASSO	AIC	1.00 (0.02)	0.01 (0.01)	29.29 (3.57)	0.36 (0.37)
		BIC	1.00 (0.00)	0.99 (0.06)	0.54 (1.51)	0.01 (0.03)
	CT-LASSO-p	AIC	1.00 (0.02)	0.01 (0.01)	29.29 (3.57)	0.36 (0.37)
		BIC	1.00 (0.00)	0.99 (0.06)	0.54 (1.51)	0.01 (0.03)
	TFL-2	AIC	1.00 (0.02)	0.02 (0.01)	29.1 (3.55)	–
		BIC	1.00 (0.00)	0.99 (0.06)	0.53 (1.51)	–
	TASSO	AIC	1.00 (0.02)	0.01 (0.01)	28.72 (3.43)	–
		BIC	0.05 (0.21)	0.95 (0.21)	3.31 (4.99)	–
	CLASSO	AIC	1.00 (0.02)	0.01 (0.01)	29.11 (3.56)	–
		BIC	0.40 (0.48)	0.88 (0.27)	4.33 (6.45)	–
Scenario 3	CT-LASSO	AIC	1.00 (0.00)	0.93 (0.07)	0.23 (0.15)	0.86 (0.18)
		BIC	1.00 (0.00)	0.98 (0.02)	0.27 (0.14)	0.89 (0.14)
	CT-LASSO-p	AIC	1.00 (0.00)	0.94 (0.05)	0.22 (0.14)	0.00 (0.00)
		BIC	1.00 (0.00)	0.99 (0.02)	0.36 (2.99)	0.00 (0.00)
	TFL-2	AIC	0.98 (0.12)	0.78 (0.11)	0.53 (0.17)	–
		BIC	0.98 (0.12)	0.98 (0.02)	0.73 (0.09)	–
	TASSO	AIC	1.00 (0.00)	0.95 (0.03)	0.54 (0.21)	–
		BIC	1.00 (0.00)	0.98 (0.01)	0.51 (0.21)	–
	CLASSO	AIC	1.00 (0.00)	0.92 (0.03)	0.57 (0.05)	–
		BIC	1.00 (0.00)	0.96(0.02)	0.57 (0.05)	–

(Continues)

TABLE 1 (Continued)

	Method	Tuning	Sensitivity	Specificity	L_2-loss	η
Scenario 4	CT-LASSO	AIC	1.00 (0.00)	0.92 (0.08)	0.65 (0.59)	0.00 (0.00)
		BIC	1.00 (0.00)	0.98 (0.01)	0.41 (0.34)	0.00 (0.00)
	CT-LASSO-p	AIC	1.00 (0.00)	0.93 (0.07)	2.95 (32.78)	0.00 (0.00)
		BIC	1.00 (0.00)	0.97 (0.01)	0.48 (0.38)	0.00 (0.00)
	TFL-2	AIC	1.00 (0.00)	0.92(0.08)	0.64 (0.59)	–
		BIC	1.00 (0.00)	0.98 (0.01)	0.41(0.34)	–
	TASSO	AIC	0.91 (0.21)	0.35 (0.12)	6.04 (3.27)	–
		BIC	0.51 (0.46)	0.84 (0.12)	3.36 (1.43)	–
	CLASSO	AIC	1.00 (0.02)	0.25 (0.05)	4.71 (0.31)	–
		BIC	0.99 (0.10)	0.59 (0.11)	3.73 (0.65)	–

leading to high SSEs. In contrast, CT-LASSO always selects $\eta = 0$ under BIC tuning, implying it only penalizes differences of conditional deviation effects.

BIC tends to perform as well as AIC does or better. Especially, in Scenarios 1 and 2, BIC is substantially better than AIC mirroring the simulation results of Wang and Zhao (2017b). This is due to the well-known fact that AIC tends to identify more variables than the true model leading to better prediction performance, while BIC is consistent in selecting the true model (Yang, 2005; Zou et al., 2007). Since our goal is to identify non-zero associations between $\boldsymbol{X}$ and Y instead of prediction, we only use the metrics of evaluating model selection consistency for comparison of methods. In Scenario 3, although the L_2-loss of BIC is slightly larger than AIC, BIC has higher specificity, suggesting that AIC will lead to more false positives.

In the Supplementary Material, we provide an additional simulation study, where noisier data is used to stress-test the method. In particular, σ^2 is set such that $\text{Var}(\varepsilon) = 10\text{Var}(\boldsymbol{X}^\top\boldsymbol{\beta})$ while all of the other settings of Scenarios 1-4 are kept constant. In this simulation, all methods perform worse, because of the weaker signal relative to the noise. CT-LASSO, however, still outperforms TASSO and CLASSO and the findings described in other simulations still hold. In addition, simulation results for Scenarios 1 and 2 setting $n = 1000$ are provided, which show similarly good relative performance.

6 | MRI DATA APPLICATION

We applied our proposed method to the data example introduced in Section 2. The outcome Y is the composite memory score while $\boldsymbol{X}$ is the brain volumes resulting from the five-level brain segmentation. Since the simulation study shows that BIC outperforms AIC on the compositional tree of the data example, we used BIC to tune the hyperparameters and obtained $\hat{\eta} = 0.405$.

CT-LASSO identified 77 non-zero marginal deviation effects ($\hat{\boldsymbol{\alpha}}$) from the 236 leaf brain regions. Table 2 displays the 10 largest effects, which account for 48% of $\|\hat{\boldsymbol{\alpha}}\|_1$. Among the 10 largest effects, Hippo-L represents the hippocampus in the left hemisphere, which is a limbic subregion and whose atrophy is well established and studied in the progression of AD (Pini et al., 2016). InferiorLV-R is the inferior pars of the right lateral ventricle (LV). Evidence has shown that

TABLE 2 Top 10 regression coefficients (α_j) of magnetic resonance imaging (MRI) application

ROI	α_j	ROI	α_j
Hippo-L	518.64	IOG-L	90.32
InferiorLV-R	−261.88	Cu-L	−70.34
Amyg-R	172.73	LG-L	−70.34
SOG-L	90.32	SylParieSul-L	69.06
MOG-L	90.32	MTG-L	62.14

its enlargement is related to MCI and AD (Nestor et al., 2008). Amyg-R stands for the amygdala in the right hemisphere. A recent study (Poulin et al., 2011) on this region suggested that 'the magnitude of amygdala atrophy is comparable to that of the hippocampus in the earliest clinical stages of AD, and is related to global illness severity'. SOG-L, MOG-L and IOG-L represent the left superior, middle and inferior occipital gyri, respectively, and are identified as a group (same marginal deviation effect) by CT-LASSO. Cu-L and LG-L are the cuneus and lingual gyrus in the left occipital region, respectively, and also have the same marginal deviation effect. Although the occipital subregions have opposite signs of marginal deviation effects, their conditional deviation effects (i.e. β_j) cancel off when combined, resulting in a positive marginal deviation effect (25.06) of the occipital region on memory. Holroyd et al. (2000) showed that occipital atrophy is associated with visual hallucinations (the most common type of hallucination) in AD. However, less is known about the different roles of occipital subregions in AD. SylParieSul-L represents the sylvian parietal sulcus in the left hemisphere. To the best of our knowledge, its enlargement is associated with progression of AD (Liu et al., 2012), which is contrary to our finding. However, we note that this region is also identified as a positive marginal deviation effect by TASSO (43.87), which may suggest a false-positive result of the variable selection methods or a special structure of the data set. MTG-L stands for the left middle temporal gyrus. Its atrophy has been associated with AD (Pini et al., 2016). However, it is important to emphasize that these results are exploratory in nature, since the method investigates a large possible collection of potential relationships and we did not pre-register any specific hypotheses.

The 77 marginal deviation effects are aggregations of 109 conditional deviation effects (β). For the 10 largest marginal deviation effects, we decomposed them into conditional deviation effects using the definition of $\boldsymbol{\alpha}$ (Section 3) and displayed the results in Figure 4. All 10 effects are from CSF and telencephalon. The effects from the ventricle are negative and the effects from the limbic region are positive, both of which are consistent with existing scientific findings (Nestor et al., 2008; Pini et al., 2016). Complete results for marginal and conditional deviation effects are given in the Supplementary Material.

In addition to CT-LASSO, we also run TFL-2, TASSO and CLASSO with BIC tuning. TFL-2 identifies 236 marginal deviation effects and 37 conditional deviation effects, suggesting an under-penalization of the marginal deviation effects. In addition, TFL-2 fails to identify the effect from the hippocampus, which is well known to be associated with AD. Twenty non-zero marginal deviation effects and 73 non-zero conditional deviation effects were identified by CT-LASSO, TASSO and CLASSO, including brain regions in the ventricles and the temporal lobe, although the magnitude of these effects differs substantially among methods. Especially, the Amyg-R (right hemisphere amygdala) region is identified only by CT-LASSO. Compared with CT-LASSO, TASSO and CLASSO identify fewer marginal effects (40 non-zero entries in $\hat{\boldsymbol{\alpha}}_{TASSO}$ and 27 non-zero entries in $\hat{\boldsymbol{\alpha}}_{CLASSO}$), but they have larger BIC (5114 for CT-LASSO, 5134 for TASSO and

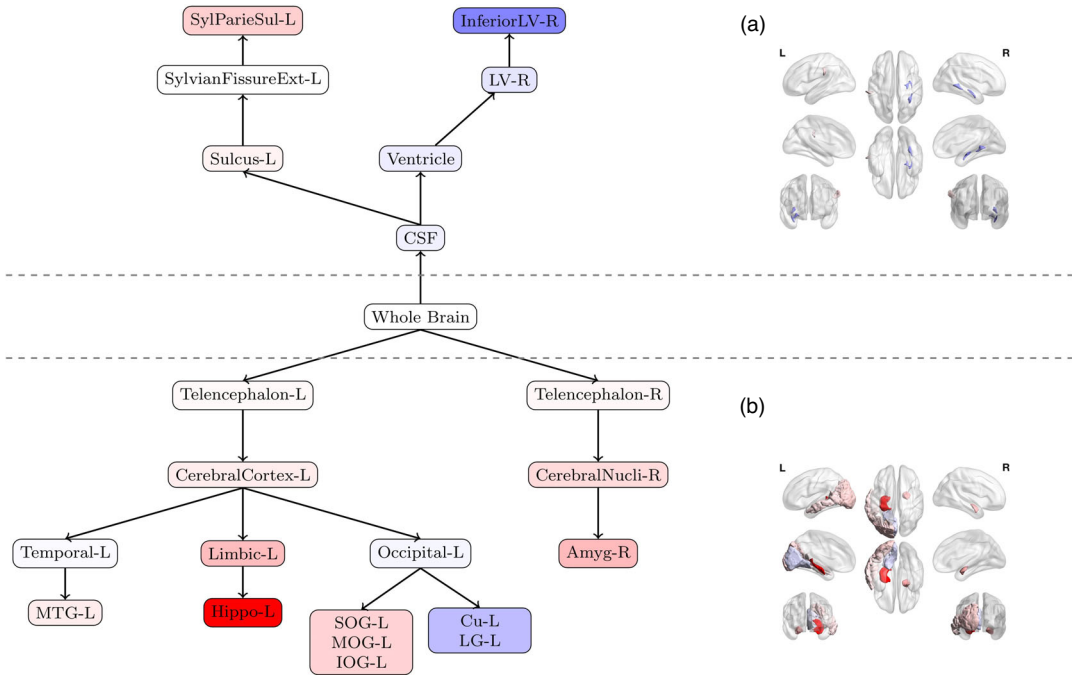

FIGURE 4 Conditional deviation effects $\boldsymbol{\beta}$ related to the 10 largest marginal deviation effects $\boldsymbol{\alpha}$. Suffix '-L' or '-R' refers to the left or right hemisphere of the brain, respectively. The red (blue) colour represents the positive (negative) sign of β with darker colour indicating a larger value of $|\beta|$. Panels (a) and (b) show the aggregated conditional effects in cerebrospinal fluid and telencephalon in three-dimensional template brain space [Colour figure can be viewed at wileyonlinelibrary.com]

5120 for CLASSO), indicating a larger residual error. In addition, CT-LASSO tends to group effects together, for example the left hemisphere occipital subregions, which can facilitate the interpretation of marginal and conditional deviation effects. For all three methods, the effects from left and right hemispheres are generally not equal, potentially suggesting a laterally asymmetric correlation between volume and memory.

7 | DISCUSSION

The linear model in Section 3 also allows for including additional covariates, in addition to covariates associated with the compositional tree. However, when interaction terms are added, the linear constraints (2) can only handle interactions between additional covariates and the whole compositional tree.

In our method, for estimating $\boldsymbol{\alpha}^*$, we assume that no components of the leaf nodes are linear combinations of the others such that $\tilde{\boldsymbol{\alpha}}^*$ is identifiable. This assumption generally holds if no further linear constraints are made on the leaf nodes. When this assumption is not true, one can add a small L_2-penalty to the right side of Equation (6) and run the model, otherwise unmodified. In this case, point estimates of $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ may be biased because of the L_2-penalty.

Our proposed method also assumes that the outcome is continuous. If the outcome is binary or a count, then relatively minor modifications could use generalized linear models. However,

since the loss function is no longer linear, how to consistently estimate $\boldsymbol{\alpha}^*$ with generalized lasso penalty remains future research.

We provide the R code for reproducing the simulations and data analyses on Github at https://github.com/BingkaiWang/compositional-hierarchical-tree-regression.

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This work was supported by NIH grants (P30AG072976, U54AG065181, R01EB029977, P41EB031771, and U54DA049110).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the LONI Image and Data Archive at https://urldefense.com/v3/__ https://ida.loni.usc.edu/__;!!N11eV2iwtfs!s7vhACgBsnIamFuRrItB5EicSzWAyp9h86KEo7gs7tiIK2yVDpumqSvzQYIkanc6xQKVJYRDUafADIZNOw$.

ORCID

Bingkai Wang https://orcid.org/0000-0002-9349-2336
Xi Luo https://orcid.org/0000-0002-0909-9372
Yi Zhao https://orcid.org/0000-0003-4766-5934

REFERENCES

Akaike, H., Parzen, E., Tanabe, K. & Kitagawa, G. (1998) *Selected papers of Hirotugu Akaike*. Berlin: Springer Science & Business Media.

Crane, P.K., Carle, A., Gibbons, L.E., Insel, P., Mackin, R.S., Gross, A. et al. (2012) Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Brain Imaging and Behavior*, 6, 502–516.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. (2003) Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Fiksel, J., Zeger, S. & Datta, A. (2020) A transformation-free linear regression for compositional outcomes and predictors. *arXiv preprint arXiv:2004.07881*.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L. et al. (2013) The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80, 105–124.

Holroyd, S., Shepherd, M.L. & Downs III, J.H. (2000) Occipital atrophy is associated with visual hallucinations in Alzheimer's disease. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 12, 25–28.

Kim, S. & Xing, E.P. (2012) Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics*, 6, 1095–1117.

Lee, J.D., Sun, Y. & Taylor, J.E. (2015) On model selection consistency of regularized m-estimators. *Electronic Journal of Statistics*, 9, 608–642.

Leite, M.L.C. (2016) Applying compositional data methodology to nutritional epidemiology. *Statistical Methods in Medical Research*, 25, 3057–3065.

Lin, W., Shi, P., Feng, R. & Li, H. (2014) Variable selection in regression with compositional covariates. *Biometrika*, 101, 785–797.

Liu, T., Lipnicki, D.M., Zhu, W., Tao, D., Zhang, C., Cui, Y. et al. (2012) Cortical gyrification and sulcal spans in early stage Alzheimer's disease. *PloS One*, 7, e31083.

Liu, C.-F., Padhy, S., Ramachandran, S., Wang, V.X., Efimov, A., Bernal, A. et al. (2019) Using deep Siamese neural networks for detection of brain asymmetries associated with Alzheimer's disease and mild cognitive impairment. *Magnetic Resonance Imaging*, 64, 190–199.

Ma, X. & Zhang, P. (2020) Quantile regression for compositional covariates. *arXiv preprint arXiv:2006.00789*.

Mori, S., Wu, D., Ceritoglu, C., Li, Y., Kolasny, A., Vaillant, M.A. et al. (2016) Mricloud: delivering high-throughput MRI neuroinformatics as cloud-based software as a service. *Computing in Science & Engineering*, 18, 21–35.

MSCI. (2020) Global industry classification standard (gics) methodology. Available from: https://www.msci.com/documents/1296102/11185224/GICS+Methodology+2020.pdf/

Mullahy, J. (2015) Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods*, 4, 71–100.

Nestor, S.M., Rupsingh, R., Borrie, M., Smith, M., Accomazzi, V., Wells, J.L. et al. (2008) Ventricular enlargement as a possible measure of Alzheimer's disease progression validated using the Alzheimer's Disease Neuroimaging Initiative database. *Brain*, 131, 2443–2454.

Papke, L.E. & Wooldridge, J.M. (1996) Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11, 619–632.

Pawlowsky-Glahn, V. & Egozcue, J. J. (2006) Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, 264, 1–10.

Pearl, J. (2009) *Causality*. Cambridge: Cambridge University Press.

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E. et al. (2016) Brain atrophy in Alzheimer's disease and aging. *Ageing Research Reviews*, 30, 25–48.

Poulin, S.P., Dautoff, R., Morris, J.C., Barrett, L.F., Dickerson, B.C. & Alzheimer's Disease Neuroimaging Initiative (2011) Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research: Neuroimaging*, 194, 7–13.

Raskutti, G., Wainwright, M.J. & Yu, B. (2010) Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11, 2241–2259.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

She, Y. (2008) *Sparse regression with exact clustering*. Stanford, CA: Stanford University.

Shojaie, A. & Michailidis, G. (2010) Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97, 519–538.

Tang, X., Oishi, K., Faria, A.V., Hillis, A.E., Albert, M.S., Mori, S. et al. (2013) Bayesian parameter estimation and segmentation in the multi-atlas random orbit model. *PloS One*, 8, e65591.

Tibshirani, R.J. & Taylor, J. (2011) The solution path of the generalized lasso. *The Annals of Statistics*, 39, 1335–1371.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Vemuri, P. & Jack, C.R. (2010) Role of structural MRI in Alzheimer's disease. *Alzheimer's Research & Therapy*, 2, 23.

Wang, T. & Zhao, H. (2017a) Constructing predictive microbial signatures at multiple taxonomic levels. *Journal of the American Statistical Association*, 112, 1022–1031.

Wang, T. & Zhao, H. (2017b) Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11, 771–791.
Yan, X. & Bien, J. (2021) Rare feature selection in high dimensions. *Journal of the American Statistical Association*, 116, 887–900.
Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92, 937–950.
Zou, H., Hastie, T. & Tibshirani, R. (2007) On the 'degrees of freedom' of the lasso. *The Annals of Statistics*, 35, 2173–2192.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wang, B., Caffo, B.S., Luo, X., Liu, C.-F., Faria, A.V., Miller, M.I. et al. (2022) Regularized regression on compositional trees with application to MRI analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(3), 541–561. Available from: https://doi.org/10.1111/rssc.12545