Statistica Sinica Preprint No: SS-2025-0314	
Title	Transfer Learning for High-dimensional Regression with Compositional Covariates: Application to Microbiome Studies
Manuscript ID	SS-2025-0314
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0314
Complete List of Authors	Qinqin Hu, Xiaojing Luo, Chencheng Ma and Wang Zhou
Corresponding Authors	Qinqin Hu
E-mails	qqhu@sdu.edu.cn
Notice: Accepted author version.	

Statistica Sinica

Transfer Learning for High-Dimensional Regression with Compositional Covariates: Application to Microbiome Studies

Qinqin Hu [1*], Xiaojing Luo[1], Chencheng Ma[2] and Wang Zhou[3]

[1]*Shandong University at Weihai* , [2]*Shandong University and* [3]*National University of Singapore.*

Abstract:

High-dimensional compositional regression is common in microbiome studies, where covariates are relative abundances and the number of taxa often exceeds the sample size. Standard regression methods may be invalid for such data. While the centered log-ratio transformed linear model offers a principled framework, it yields unstable estimates with limited target samples. We develop transfer learning procedures that borrow information from auxiliary source studies for high-dimensional compositional regression with subcomposition structures and additional non-compositional covariates. The proposed methods incorporate the compositional linear constraint through constrained ℓ_1 -regularized estimation and allow both model and covariate shifts across studies. We propose Oracle-Trans-sub-Coda-Lasso, for known informative sources, and Trans-sub-Coda-Lasso, which detects informative sources using marginal screening statistics. Under suitable regularity and similarity conditions, we establish the ℓ_2-norm error convergence rate of the oracle estimator and the consistency of the source-detection procedure. Simulations and an application to ulcerative colitis gut microbiome data for body mass index prediction demonstrate the improved performance of the proposed methods.

Key words and phrases: Transfer learning, Multisource data, Compositional data, Centered log-ratio model, High-dimensional regression.

*Corresponding author. E-mail:qqhu@sdu.edu.cn

1. Introduction

Regression with compositional covariates arises in many scientific fields, including microbiome studies, geochemistry, environmental science, and nutritional epidemiology. Compositional covariates are proportions or relative abundances whose components are nonnegative and sum to a constant, and hence lie in a simplex rather than in the usual Euclidean space. Directly applying standard regression methods to raw compositions may therefore lead to invalid inference and misleading interpretation. This issue is particularly important in microbiome studies, where microbial abundances are observed as relative abundances and the number of taxa often exceeds the sample size.

A principled framework for compositional regression is the log-contrast model of Aitchison and Bacon-Shone (1984), which respects the scale-invariance property of compositional data by imposing a zero-sum constraint on the regression coefficients. For high-dimensional compositional covariates, Lin et al. (2014) introduced benchmark variables into their model, shifting constraints from covariates to coefficients. They then proposed estimating parameters via a constrained ℓ_1-regularized estimator. Subsequent extensions incorporated tree-structured hierarchies (Shi et al., 2016; Wang and Zhao, 2017), analyzed optimization geometry (Combettes and Müller, 2021), addressed robustness (Mishra and Müller, 2022), and handled measurement error (Shi et al., 2022; Tan et al., 2024). Longitudinal extensions have also been developed (Ma et al., 2023).

Despite these advances, limited sample size remains a major challenge in high-dimensional compositional regression, especially for target populations such as specific disease subgroups or clinical cohorts. Penalization exploits sparsity but does not increase the amount of information available for estimating the target model. In contrast, related datasets from external

cohorts or studies are increasingly available in biomedical applications. This motivates transfer learning, which borrows information from auxiliary source datasets to improve estimation and prediction for a target task.

Transfer learning has recently been studied extensively in high-dimensional statistics, including linear regression with single or multiple sources (Bastani, 2021; Li et al., 2022), adaptations for concurrent model and covariate shifts (He et al., 2024b), simultaneous estimation and source selection (Li et al., 2024a), and extensions to generalized linear models (Tian and Feng, 2023; Li et al., 2024c), quantile regression (Huang et al., 2023; Jin et al., 2024; Qiao et al., 2024), graphical models (He et al., 2022), and dimensionality reduction (Li et al., 2024b; He et al., 2025, 2024a). Reviews can be found in Pan and Yang (2010), Weiss et al. (2016), and Cheng et al. (2020). However, existing methods are primarily designed for ordinary Euclidean covariates and do not directly address the constraints and dependence structures induced by compositional data.

In this paper, we develop transfer-learning procedures for high-dimensional regression with subcomposition-structured compositional covariates and additional non-compositional covariates. The proposed model allows compositional variables to be divided into multiple subcompositions, each satisfying its own simplex constraint, and imposes a general linear constraint $\mathbf{C}_s^\top \boldsymbol{\theta}^* = \mathbf{0}$ on the compositional regression coefficients, where $\boldsymbol{\theta}^*$ denotes the compositional component of the regression coefficient vector. The standard compositional regression model without subcomposition structure is included as a special case. The framework also allows ordinary covariates, such as demographic, clinical, or environmental variables, to be modeled jointly with transformed compositional covariates, while the constraint is imposed only on the compositional component.

We propose two estimators. The Oracle-Trans-sub-Coda-Lasso is designed for the ideal case where informative source studies are known in advance. The Trans-sub-Coda-Lasso handles the more realistic setting where informative sources must be identified from the data. Our source-detection procedure uses marginal statistics to screen candidate source studies, avoiding repeated high-dimensional constrained estimation over source subsets. It is therefore simpler to implement than the source-selection strategy of Li et al. (2022) and is closer in spirit to Tian and Feng (2023) and Jin et al. (2024), although those works do not consider constrained regression with compositional covariates.

The proposed framework is not a direct extension of existing transfer-learning methods with an added linear constraint. After the centered log-ratio (clr) or subcomposition-wise centered log-ratio transformation, the transformed covariates are linearly dependent within each composition or subcomposition, making the covariance matrix singular by construction. This singularity creates substantial technical difficulties, since inverse-based arguments commonly used for ordinary covariates cannot be directly applied. Our analysis therefore requires arguments that accommodate both the compositional constraint and the singular design structure.

The main contributions of this paper are threefold. First, we introduce a penalized transfer-learning framework that unifies subcomposition-structured compositional covariates and ordinary covariates under a general linear constraint. Second, we develop both an oracle transfer-learning procedure and a data-driven procedure with informative-source detection. Third, under suitable regularity and similarity conditions, we establish the convergence rate of the Oracle-Trans-sub-Coda-Lasso and show that it can achieve a faster ℓ_2-norm error rate than the target-only benchmark estimator. We also prove the consistency of the pro-

posed source-detection procedure. Extensive simulations and an application to body mass index prediction using gut microbiome data from ulcerative colitis cohorts demonstrate the practical utility of the proposed methods.

The rest of the paper is organized as follows. Section 2 presents the proposed methodology. Section 3 establishes theoretical properties. Section 4 and Section 5 report simulation studies and the ulcerative colitis microbiome application, respectively. In Section 6, we review our contributions and discuss some future research directions. Additional numerical results and proofs are provided in the supplementary material.

Before we start the formal introduction, let us first summarize some notations that will be used frequently in this article. For two constants c_1 and c_2, denote $c_1 \vee c_2 = \max(c_1, c_2)$ and $c_1 \wedge c_2 = \min(c_1, c_2)$. For a generic set $\mathcal{S}$, the cardinality and complement set are defined by $|\mathcal{S}|$ and $\mathcal{S}^c$, respectively. For a vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$, denote the ℓ_1 norm by $\|\boldsymbol{x}\|_1 = \sum_{j=1}^p |x_j|$, the ℓ_2 norm by $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^p x_j^2}$ and the infinity norm by $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. Let $\boldsymbol{x}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$ be the sub-vector formed by the index in the set $\mathcal{S}$, and $\mathbf{0}_p$ and $\mathbf{1}_p$ be the p-dimensional vector whose elements are all zero and one, respectively. $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix. For a matrix $\mathbf{A} = (a_{i,j})_{p\times q} \in \mathbb{R}^{p\times q}$, define the minimum and maximum eigenvalues as $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively. And its 2-norm, infinity norm and max-norm are defined as $\|\mathbf{A}\|_2 = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \|\mathbf{A}\boldsymbol{x}\|_2$,$\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^p |a_{ij}|$ and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, respectively.

2. Methodology

In this section, we first introduce the centered log-ratio data (clr) formation model. We then develop two procedures for estimating the above model on the target data using transfer

learning.

2.1 Model

Consider a composition $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$ that takes values in the $(q-1)-$dimensional unit simplex $S^{q-1} = \{(x_1, \ldots, x_q) : x_j > 0, j = 1, \ldots, q, \sum_{j=1}^{q} x_j = 1\}$. The simplex structure imposes dependence between the components of the compositional data. Thus, the traditional methodology defined for spaces of real numbers cannot be applied. To overcome this problem, Aitchison (1982) proposed the additive log-ratio (alr) transformation which requires the choice of a reference. The log-contrast model is written as,

$$\boldsymbol{y} = \tilde{\mathbf{Z}}^q \boldsymbol{\theta}^*_{\setminus q} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\tilde{\mathbf{Z}}^q = \{\log(x_{ij}/x_{iq})\}$ is the $n \times (q-1)$ matrix whose $q-$th component is the reference component , $\boldsymbol{\theta}^*_{\setminus q} = (\theta^*_1, \ldots, \theta^*_{q-1})^\top \in \mathbb{R}^{q-1}$ is the corresponding regression coefficient vector, and $\boldsymbol{\varepsilon}$ is an $n-$vector of independent noise. Since the log-contrast model employed an arbitrary reference variable for all other variables, the solution changes depending on the selection of the reference. By expressing $\theta^*_q = -\sum_{i=1}^{q-1} \theta^*_j$, we can rewrite model (2.1) into a symmetric form as

$$\boldsymbol{y} = \tilde{\mathbf{Z}}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \sum_{j=1}^{q} \theta^*_j = 0, \tag{2.2}$$

where $\tilde{\mathbf{Z}} = \{\log(x_{ij})\}$ is the $n \times q$ log-transformed matrix (Aitchison and Bacon-Shone, 1984) and $\boldsymbol{\theta}^* = (\theta^*_1, \ldots, \theta^*_q)^\top \in \mathbb{R}^q$. The linear constraint in (2.2) ensures that, after model fitting, the response can be equivalently expressed as linear combinations of log-ratios of the original compositions (Aitchison, 2003; Lin et al., 2014; Combettes and Müller, 2021).

When grouping information for the predictors is available, the zero-sum constraint in (2.2) allows for a natural generalization. In the context of microbiome research, each predictor corresponds to a taxonomic or phylogenetic unit, and the hierarchical relationships among these units are typically encoded in a tree structure with q leaves and G hierarchical levels. Following Shi et al. (2016) , this information can be incorporated into constraint (2.2) via a linear equality constraint. To illustrate, suppose that microbiome data are analyzed at a fixed taxonomic level. The q taxa can be naturally partitioned into G disjoint groups according to a higher-level taxonomic or phylogenetic classification. Specifically, the g-th group contains p_g taxa belonging to the same higher-level taxon. Each set contains taxa belonging to the same phylum.

Let X_{gs} represent the relative abundance of the s-th taxon belonging to the g-th taxonomic group for $g = 1, \ldots, G$ and $s = 1, \ldots, p_g$, such that

$$\sum_{s=1}^{p_g} X_{gs} = 1, \quad \text{for } g = 1, \ldots, G.$$

Let $\mathbf{X}_g$ (an $n \times p_g$ matrix) denote n samples of the p_g taxa subcomposition. The model linking the subcomposition to the response $\boldsymbol{y}$ is

$$\boldsymbol{y} = \sum_{g=1}^{G} \tilde{\mathbf{Z}}_g \boldsymbol{\theta}_g + \boldsymbol{\varepsilon}, \quad \mathbf{1}_{p_g}^\top \boldsymbol{\theta}_g = \sum_{s=1}^{p_g} \theta_{gs} = 0 \quad \text{for } g = 1, \ldots, G, \tag{2.3}$$

with $\tilde{\mathbf{Z}}_g = \log(\mathbf{X_g}) = (\tilde{Z}_{g1}, \ldots, \tilde{Z}_{gp_g}) = (\log(X_{g1}), \ldots, \log(X_{gp_g})) \in \mathbb{R}^{n \times p_g}$ and $\boldsymbol{\theta}_g = (\theta_{g1}, \ldots, \theta_{gp_g})^\top$.

Without loss of generality, we assume that the taxa have been sorted according to their

groups. We define the subcomposition matrix $\mathbf{C}_s$ as follows

$$\mathbf{C}_s^\top = \begin{bmatrix} \mathbf{1}_{p_1}^\top & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{p_2}^\top & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{p_G}^\top \end{bmatrix}_{G\times p}.$$

Then model (2.3) can be rewritten by including the subcomposition matrix $\mathbf{C}_s$,

$$\boldsymbol{y} = \tilde{\mathbf{Z}}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \mathbf{C}_s^\top \boldsymbol{\theta}^* = \mathbf{0}, \tag{2.4}$$

where $\tilde{\mathbf{Z}} = (\tilde{\mathbf{Z}}_1, \dots, \tilde{\mathbf{Z}}_G)$ and $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_G^\top)^\top$.

We next adopt the grouped centered log-ratio(clr) transformation of the original taxonomic composition (Aitchison, 1982) for the subsequent analysis. For the g-th group, define

$$\tilde{\tilde{\mathbf{Z}}}_g = \mathrm{clr}(\mathbf{X}_g) = \log\{\mathbf{X}_g/(\Pi_{s=1}^{p_g} X_{gs})^{1/p_g}\},$$

which naturally satisfies $\mathbf{1}_{p_g}^\top \tilde{\tilde{\mathbf{Z}}}_g = \mathbf{0}$. Under the within-group zero-sum constraint $\mathbf{1}_{p_g}^\top {\boldsymbol{\theta}^*}_g = \mathbf{0}$, we have $\tilde{\tilde{\mathbf{Z}}}_g \boldsymbol{\theta}_g^* = \tilde{\mathbf{Z}}_g \boldsymbol{\theta}_g^*$. Therefore, under $\mathbf{C}_s^\top \boldsymbol{\theta}^* = 0$, there exists $\tilde{\tilde{\mathbf{Z}}}\boldsymbol{\theta}^* = \tilde{\mathbf{Z}}\boldsymbol{\theta}^*$, where $\tilde{\tilde{\mathbf{Z}}} = (\tilde{\tilde{\mathbf{Z}}}_1, \dots, \tilde{\tilde{\mathbf{Z}}}_G)$ is the $n \times q$ matrix, which confirms that the grouped centered log-ratio data formation (clr) model with G linear constraints on the coeffiecients,

$$\boldsymbol{y} = \tilde{\tilde{\mathbf{Z}}}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \mathbf{C}_s^\top \boldsymbol{\theta}^* = \mathbf{0}, \tag{2.5}$$

is algebraically equivalent to the grouped log-contrast model(2.4).

This reparameterization is not only algebraically convenient, but also distributionally motivated. In particular, for each group g, suppose that there exists a latent positive basis vector $\mathbf{V}_g$ such that $X_{gs} = V_{gs}/\sum_{t=1}^{p_g} V_{gt}$, and let $\mathbf{U}_g = \log(\mathbf{V}_g)$ be one admissible log-basis representation. It is worth noting that this representation is not unique: $\mathbf{V}_g$ is identifiable only up to a common positive scaling factor within group g, or equivalently, $\mathbf{U}_g$ is identifiable only up to an additive constant on the log scale. If $\mathbf{U}_g$ is multivariate normal, then $\mathbf{X}_g$ follows a logistic-normal model. Although $\mathbf{V}_g$ is not identifiable up to a common scaling factor, this ambiguity disappears after clr transformation, since $\mathrm{clr}(\mathbf{X}_g) = \mathbf{U}_g - \bar{U}_g \mathbf{1}_{p_g}$, where $\bar{U}_g = p_g^{-1}\sum_{s=1}^{p_g} U_{gs}$ denotes the average of the components of $\mathbf{U}_g$ within group g. Hence the grouped clr-transformed covariates inherit the Gaussian property of the latent log-basis vector, whereas the raw log-components $\log(X_{gs})$ do not admit such a direct characterization. This provides an additional justification for working with the grouped clr formulation.

When q' additional non-compositional covariates $\mathbf{N} \in \mathbb{R}^{n\times q'}$, such as habitat and host-associated factors or other control variables, are available, we can extend model (2.5) to

$$\boldsymbol{y} = \tilde{\tilde{\mathbf{Z}}}\boldsymbol{\theta}^* + \mathbf{N}\boldsymbol{a}^* + \boldsymbol{\varepsilon}, \quad \mathbf{C}_s^\top \boldsymbol{\theta}^* = \mathbf{0}, \tag{2.6}$$

where $\boldsymbol{a}^* \in \mathbb{R}^{q'}$ is the coefficient vector for all non-compositional variables. We refer to the proposed model (2.6) as the adjusted group-constrained centered log-ratio regression model, where "adjusted" indicates the inclusion of additional non-compositional covariates. By fusing the compositional and non-compositional covariates into the general design matrix $\mathbf{Z} = [\tilde{\tilde{\mathbf{Z}}}\,\mathbf{N}]_{n\times p}$, we can denote the corresponding model coefficients by $\boldsymbol{w}^* = (\boldsymbol{\theta}^*\,\boldsymbol{a}^*) \in \mathbb{R}^p$, where $p = q + q'$. Adding the linear constraint matrix $\mathbf{C}_s$ by a zero matrix $G \times q'$, denoted

by $\mathbf{C} = [\mathbf{C}_s^\top \mathbf{0}_{G\times q'}]^\top$, the model (2.6) can be simplified to

$$\boldsymbol{y} = \mathbf{Z}\boldsymbol{w}^* + \boldsymbol{\varepsilon}, \quad \mathbf{C}^\top \boldsymbol{w}^* = \mathbf{0}, \tag{2.7}$$

Our main interest is to estimate the $\boldsymbol{w}^* = (w_1^*, \ldots, w_p^*)^\top$, which is the true regression coefficient for the target dataset. In addition to the target dataset, suppose we also have access to K source datasets. For distinction, we denote the target dataset as $(\mathbf{X}^{(0)}, \mathbf{N}^{(0)}, \boldsymbol{y}^{(0)})$ and K source datasets with the k th source denoted as $(\mathbf{X}^{(k)}, \mathbf{N}^{(k)}, \boldsymbol{y}^{(k)})$, where $\mathbf{X}^{(k)} \in \mathbb{R}^{n_k\times q}, \mathbf{N}^{(k)} \in \mathbb{R}^{n_k\times q'}, \boldsymbol{y}^{(k)} \in \mathbb{R}^{n_k}$ for $k = 1, \ldots, K$. We assume that the adjusted group-constrained centered log-ratio regression model holds for target and source datasets,

$$\boldsymbol{y}^{(k)} = \mathbf{Z}^{(k)}\boldsymbol{w}^{*(k)} + \boldsymbol{\varepsilon}^{(k)}, \quad \mathbf{C}^\top \boldsymbol{w}^{*(k)} = \mathbf{0}, \tag{2.8}$$

for $k = 0, \ldots, K$, where $\boldsymbol{y}^{(k)} \in \mathbb{R}^{n_k}$ is the response vector and $\mathbf{Z}^{(k)} = [\tilde{\tilde{\mathbf{Z}}}^{(k)}\mathbf{N}^{(k)}]_{n_k\times p}$ is the $n_k \times p$ matrix, $\tilde{\tilde{\mathbf{Z}}}^{(k)} = (\mathrm{clr}(\mathbf{X}_1^{(k)}), \ldots, \mathrm{clr}(\mathbf{X}_G^{(k)}))$, $\mathbf{X}_g^{(k)}$ denotes the compositional data of the g-th group in the k-th source. The ith row of $\mathbf{Z}^{(k)}$ and the i-th element of $\boldsymbol{y}^{(k)}$ are denoted as $\boldsymbol{z}_i^{(k)\top}$ and $y_i^{(k)}$, respectively. For convenience, we assume that the covariates and responses are mean-centered. $\boldsymbol{w}^{*(k)} = (w_1^{*(k)}, \ldots, w_p^{*(k)})^\top \in \mathbb{R}^p$ is the regression coefficient vector for target model $(k = 0)$ and k-th source model $(k = 1, \ldots, K)$. Similarly, we assume $\boldsymbol{w}^{*(k)}$ has s_k nonzero elements and that the error term $\boldsymbol{\varepsilon}^{(k)}$ is independent noise $(k = 0, 1, \ldots, K)$.

The source regression coefficient vector $\boldsymbol{w}^{*(k)}(k \neq 0)$ can be different from that of the target $\boldsymbol{w}^{*(0)}$. Intuitively, the k-th source could be useful for transfer learning when $\boldsymbol{w}^{*(k)}$ is close to $\boldsymbol{w}^{*(0)}$. Define the k-th contrast $\boldsymbol{\delta}^{(k)} = \boldsymbol{w}^{*(0)} - \boldsymbol{w}^{*(k)}$. The set of informative source

datasets is those whose contrasts satisfy that

$$\mathcal{A} = \{1 \leq k \leq K : \|\boldsymbol{\delta}^{(k)}\|_1 \leq h\},$$

where $\|\cdot\|_1$ denotes the l_1 norm. The certain threshold $h > 0$ determines the transferring level of the informative set $\mathcal{A}$. In practice, the informative set $\mathcal{A}$ is unknown most of the time. So in the following transfer learning algorithm, we will divide it into two parts. The first part is the oracle algorithm with the known informative set $\mathcal{A}$. In the second part, we will first introduce the selection of available transfer learning sources and then the transfer learning algorithms with the estimated informative set $\mathcal{A}$.

Before deriving the transfer learning procedure, let us first introduce an estimator of the target regression coefficient vector $\boldsymbol{w}^{*(0)}$, using only the target data. Specifically, we consider the following constrained Lasso estimator,

$$\hat{\boldsymbol{w}}^{(0)} = \arg\min_{\boldsymbol{w}} \left(\frac{1}{2n_0} \|\boldsymbol{y}^{(0)} - \mathbf{Z}^{(0)}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 \right), \quad \text{subject to } \mathbf{C}^\top \boldsymbol{w} = \mathbf{0}, \tag{2.9}$$

where $\lambda > 0$ is a regularization parameter, and $\|\cdot\|_2$ denotes the l_2 norm. This estimator enjoys model selection consistency and uniform estimation consistency under some regular conditions on the target data.

2.2 Oracle Trans-sub-Coda-Lasso algorithm

Given an informative auxiliary set $\mathcal{A}$, we propose Oracle Trans-sub-Coda-Lasso—a transfer learning algorithm for high-dimensional linear regression with subcomposition-structured compositional covariates and additional non-compositional covariates. Building on method-

ologies from Bastani (2021), Li et al. (2022), Tian and Feng (2023) and Jin et al. (2024), our approach implements transfer learning through a targeted two-stage procedure: a transferring step followed by a debiasing step.

More specifically, given the informative set $\mathcal{A}$, which identifies source coefficient vectors $\boldsymbol{w}^{*(k)}$ $(k \neq 0)$ that are close to the target coefficient vector $\boldsymbol{w}^{*(0)}$ under a predetermined threshold, we leverage this information through data integration to achieve the goal of transferring. In particular, when $\boldsymbol{w}^{*(k)} = \boldsymbol{w}^{*(0)}$ for $k \in \mathcal{A}$, simultaneous estimation of shared parameters using combined target and source data improves estimation accuracy from $\mathcal{O}(s\log(p)/n_0)$ to $\mathcal{O}(s\log(p)/(n_0 + n_{\mathcal{A}}))$, where $s = \|\boldsymbol{w}^{*(0)}\|_0$ and n_0 and $n_{\mathcal{A}}$ denote target and aggregated source sample size, respectively. Therefore, we first fit the adjusted group-constrained centered log-ratio regression model with constrained Lasso by pooling target and all source samples in a given informative set $\mathcal{A}$. On the other hand, $\boldsymbol{w}^{*(k)}(k \in \mathcal{A})$ is merely very close to $\boldsymbol{w}^{*(0)}$, not exactly equal to $\boldsymbol{w}^{*(0)}$. We then fit the contrast in the second step using only the target data.

The detailed algorithm (Oracle Trans-sub-Coda-Lasso) is presented in Algorithm 1. In step 1, $\hat{\boldsymbol{w}}^{\mathcal{A}}$ converges to $\boldsymbol{w}^{\mathcal{A}}$ which can be defined via the following constrained optimization problem

$$\min_{w} \sum_{k\in\{0\}\bigcup\mathcal{A}} \alpha_k \mathbb{E}\left[(\boldsymbol{y}^{(k)} - \mathbf{Z}^{(k)}\boldsymbol{w})^\top(\boldsymbol{y}^{(k)} - \mathbf{Z}^{(k)}\boldsymbol{w})\right], \quad \text{subject to } \mathbf{C}^\top\boldsymbol{w} = \mathbf{0}, \tag{2.10}$$

where $\alpha_k = \frac{n_k}{n_{\mathcal{A}}+n_0}$ and $\boldsymbol{w}^{\mathcal{A}}$ can be expressed as a linear transformation of the true parameter $\boldsymbol{w}^{*(k)}$, that is, $\boldsymbol{w}^{\mathcal{A}} = \boldsymbol{\Omega}_c \sum_{k\in\{0\}\bigcup\mathcal{A}} \alpha_k \boldsymbol{\Sigma}^{(k)} w^{*(k)}$, where $\boldsymbol{\Sigma}^{(k)} = \mathbb{E}[(\mathbf{Z}^{(k)})^\top\mathbf{Z}^{(k)}]$, $\boldsymbol{\Omega}_{\mathbf{c}}$ is the Moore-Penrose inverse of $\boldsymbol{\Sigma}$ associated with the constraint matrix $\mathbf{C}$ and $\boldsymbol{\Sigma} = \sum_{k\in\{0\}\bigcup\mathcal{A}} \alpha_k \boldsymbol{\Sigma}^{(k)}$. To highlight our target parameter, we use $\boldsymbol{\beta}^*$ to denote the target parameter in the following.

Algorithm 1 : Oracle Trans-sub-Coda-Lasso

Input: target data $(\mathbf{Z}^{(0)}, \boldsymbol{y}^{(0)})$, source data $\{(\mathbf{Z}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{K}$, penalty parameters λ_w and λ_δ, transferring set $\mathcal{A}$

Output: the estimated coefficient vector $\hat{\boldsymbol{\beta}}$

step 1. Compute

$$\hat{\boldsymbol{w}}^{\mathcal{A}} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \left\{ \frac{1}{2(n_{\mathcal{A}} + n_0)} \sum_{k\in\{0\}\bigcup\mathcal{A}} \|\boldsymbol{y}^{(k)} - \mathbf{Z}^{(k)}\boldsymbol{w}\|_2^2 + \lambda_w \|\boldsymbol{w}\|_1 \right\}, \text{s.t. } \mathbf{C}^\top \boldsymbol{w} = \mathbf{0}.$$

step 2. Compute

$$\hat{\boldsymbol{\delta}}^{\mathcal{A}} = \arg\min_{\boldsymbol{\delta}\in\mathbb{R}^p} \left\{ \frac{1}{2n_0} \|\boldsymbol{y}^{(0)} - \mathbf{Z}^{(0)}(\hat{\boldsymbol{w}}^{\mathcal{A}} + \boldsymbol{\delta})\|_2^2 + \lambda_\delta \|\boldsymbol{\delta}\|_1 \right\}, \text{s.t. } \mathbf{C}^\top \boldsymbol{\delta} = \mathbf{0}.$$

step 3. Let $\hat{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{w}}^{\mathcal{A}} + \hat{\boldsymbol{\delta}}^{\mathcal{A}}$

Then the pooling contrast $\boldsymbol{\delta}^{\mathcal{A}} = \boldsymbol{w}^{A} - \boldsymbol{\beta}^* = \boldsymbol{\Omega}_c \sum_{k\in\{0\}\cup\mathcal{A}} \alpha_k \boldsymbol{\Sigma}^{(k)} \boldsymbol{\delta}^{(k)}$, where $\boldsymbol{\delta}^{(k)} = \boldsymbol{\beta}^* - \boldsymbol{w}^{*(k)}$ is the kth contrast. The derivation of the expression and the details of Algorithm 1 are provided in Section S1 of the Supplementary Materials.

2.3 Informative set selection

As we described, the Oracle Trans-sub-Coda-Lasso is based on knowledge of the informative set $\mathcal{A}$, which in practice may not be given. Motivated by Tian and Feng (2023) and Jin et al. (2024), we develop the following effective procedure to select the informative set $\mathcal{A}$, which is similar to the idea of cross-validation.

The proposed selection method has three steps. First, we partition the full target dataset into two disjoint subsets of equal size, one for training ($\mathcal{I}$) and another for testing ($\mathcal{I}^c$). Second, we run the pooling-training procedure with each source data and the target training data, to produce a set of single-source pooling estimators. These pooling estimators are evaluated on the target testing data. Finally, we select informative sources by comparing

Algorithm 2 : Trans-sub-Coda-Lasso

Input: target data $(\mathbf{Z}^{(0)}, \boldsymbol{y}^{(0)})$, source data $\{(\mathbf{Z}^{(k)}, \boldsymbol{y}^{(k)})\}_{k=1}^{K}$, penalty parameters $\{\lambda^{(k)}\}_{k=0}^{K}$ and a constant $\epsilon_0 \in (0, c_\epsilon]$ controls the strictness of selection, where c_ϵ is a constant that satisfies Assumption 5 below.
Output: the estimated coefficient vector $\widetilde{\boldsymbol{\beta}}$ and the selected transferring set $\hat{\mathcal{A}}$.
step 1. Randomly partition target data $(\mathbf{Z}^{(0)}, \boldsymbol{y}^{(0)})$ into two disjoint subsets $\mathcal{D}_1 = (\mathbf{Z}^{(0)}_{\mathcal{I}}, \boldsymbol{y}^{(0)}_{\mathcal{I}})$ and $\mathcal{D}_2 = (\mathbf{Z}^{(0)}_{\mathcal{I}^c}, \boldsymbol{y}^{(0)}_{\mathcal{I}^c})$, where $\mathcal{I}$ is a subset of $\{1, \ldots, n_0\}$ with cardinality $|\mathcal{I}| \approx n_0/2$.
step 2. Compute $\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{I}}$ by fitting the constrained Lasso on $\mathcal{D}_1$ with penalty parameter $\lambda^{(0)}$.
step 3. Obtain $\hat{\boldsymbol{w}}^{(0,k)}$ by running Step 1 in Algorithm 1 with $(\mathbf{Z}^{(0)}_{\mathcal{I}}, \boldsymbol{y}^{(0)}_{\mathcal{I}}) \bigcup (\mathbf{Z}^{(k)}, \boldsymbol{y}^{(k)})$ and penalty parameter $\lambda^{(k)}$ for all $k \neq 0$.
step 4. Calculate the error $\hat{Q}(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{I}})$ and $\hat{Q}(\hat{\boldsymbol{w}}^{(0,k)})$ for $k = 1, \ldots, K$.
step 5. Obtain the estimated informative set as follows

$$\hat{\mathcal{A}} = \{1 \leq k \leq K : \hat{Q}(\hat{\boldsymbol{w}}^{(0,k)}) \leq (1 + \epsilon_0)\hat{Q}(\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{I}})\}.$$

step 6. Obtain $\widetilde{\boldsymbol{\beta}}$ by Algorithm 1 using $\{(\mathbf{Z}^{(k)}, \boldsymbol{y}^{(k)})\}_{k \in \hat{\mathcal{A}}}$
step 7. Output $\widetilde{\boldsymbol{\beta}}$ and $\hat{\mathcal{A}}$.

the error incurred by transfer learning estimators with that of the target-only estimator. We consider the squared prediction error function. For any coefficient estimate $\boldsymbol{\beta}$, the average squared prediction error on the target testing data $\mathcal{I}^c$ is

$$\hat{Q}(\boldsymbol{\beta}) = \frac{2}{n_0} \sum_{i \in I^c} (y_i^{(0)} - \boldsymbol{z}_i^{(0)\top} \boldsymbol{\beta})^2. \tag{2.11}$$

The detailed algorithm is presented as Algorithm 2.

It is worth noting that we can estimate c_ϵ in Algorithm 2 in practice as

$$c_\epsilon = \inf_{k \in \{1, \ldots, K\}} \frac{\left(\hat{\boldsymbol{w}}^{(0,k)} - \hat{\boldsymbol{\beta}}^{(0)}\right)^\top \hat{\boldsymbol{\Sigma}}^{(0)} \left(\hat{\boldsymbol{w}}^{(0,k)} - \hat{\boldsymbol{\beta}}^{(0)}\right)}{\hat{Q}(\hat{\boldsymbol{\beta}}^{(0)})}$$

where $\hat{\boldsymbol{\beta}}^{(0)}$ is the initial single-target estimator in (2.9)

3. Theoretical results

In this section, we display the theoretical guarantees for the two proposed algorithms. We assume the following conditions hold in our theoretical analysis.

Assumption 1 .(Sub-Gaussian random errors). For each $k \in \mathcal{A} \cup \{0\}$, $\mathbb{E}[(y_i^{(k)})^2]$ is finite and the random noise $\varepsilon_i^{(k)}$ are i.i.d. sub-Gaussian with mean zero and variance σ_k^2.

Assumption 2 (Sub-Gaussian log-basis and non-compositional covariates). For each $k \in \mathcal{A} \cup \{0\}$, let $\mathbf{U}^{(k)} = \left(U_1^{(k)}, \ldots, U_G^{(k)}\right) \in \mathbb{R}^{n_k \times q}$ denote one admissible latent log-basis matrix associated with the grouped compositional observations. We assume that the rows of $\mathbf{U}^{(k)}$ and non-compositional $\mathbf{N}^{(k)}$ are i.i.d. sub-Gaussian with mean zero and covariance matrix $\mathbf{\Sigma}_U^{(k)}$ and $\mathbf{\Sigma}_N$, respectively. Moreover, the eigenvalues of $\mathbf{\Sigma}_U^{(k)}$ and $\mathbf{\Sigma}_N$ are bounded away from 0 and ∞. Since the latent log-basis is identifiable only up to groupwise additive constants, this assumption is imposed on one admissible representative of the latent log-basis class.

With the potential distributional shift, that is, the augmented design matrices are moderately heterogeneous, The following measurements are defined to characterize the maximum difference between $\mathbf{\Sigma}^{(0)}$ and $\mathbf{\Sigma}^{(k)}$, $k \in \mathcal{A}$:

$$C_\Sigma^{\mathcal{A}} = 1 + \max_{1 \le j \le p} \max_{k \in \mathcal{A}} \|e_j^\top (\mathbf{\Sigma}^{(k)} - \mathbf{\Sigma}^{(0)}) \mathbf{\Omega}_c\|_1 \tag{3.12}$$

where e_j is the unit vector with 1 on the j th position, $\mathbf{\Sigma}^{(k)} = \mathbb{E}[\mathbf{Z}^{(k)\top} \mathbf{Z}^{(k)}]$. Notice that $C_\Sigma^{\mathcal{A}}$ can be further bounded, as shown in Li et al. (2022) and Tian and Feng (2023). The

parameter space we consider is

$$\Theta(s,h)=\left\{\boldsymbol{\beta}^*,\{\boldsymbol{w}^{*(k)}\}_{k\in\mathcal{A}}:\|\boldsymbol{\beta}^*\|_0\le s,\sup_{k\in\mathcal{A}}\|\boldsymbol{w}^{*(k)}-\boldsymbol{\beta}^*\|_1\le h\right\}$$

Assumption 3 .(Sample size and finite source). We assume $s\log p/(n_0+n_{\mathcal{A}})+C_{\Sigma}^{\mathcal{A}}h\sqrt{\log p/n_0}=o(1)$ and $K=O(n_0)$, $|\mathcal{A}|$ is bounded by a constant.

Assumption 1 assumes sub-Gaussian random noises for target and source samples, and the second moment of the response vector is finite.

Assumption 2 is motivated by the logistic-normal model for compositional data introduced by Aitchison (1982, 2003), under which a composition arises from a latent positive basis through normalization. Moreover, Assumption 2 implies that the grouped clr-transformed covariates are sub-Gaussian. From another perspective, one may also directly impose sub-Gaussian assumptions on the grouped clr-transformed covariates, as in Shi et al. (2016) and Yuan et al. (2024). The sub-Gaussian tail condition and the bounded-spectrum covariance assumption are standard regularity conditions in high-dimensional theory and are imposed here to facilitate concentration arguments and theoretical error control. Thus, Assumption 2 may be viewed as a natural grouped extension of basis-based logistic-normal modeling for compositional data.

Assumption 3 amounts to requiring the true model is sufficiently sparse and the discrepancy in regression coefficients between the target and the sources is not too large.

The following theorem gives the convergence rate for the Oracle Trans-sub-Coda-Lasso estimator under the aforementioned conditions.

Theorem 1. ***(Convergence rate of Oracle Trans-sub-Coda-Lasso).*** *Assume that*

$\boldsymbol{\beta}^*, \{\boldsymbol{w}^{*(k)}\}_{k\in\mathcal{A}} \in \Theta(s,h)$ and Assumptions 1-3 hold. Suppose $\mathcal{A}$ is known with $C_\Sigma^{\mathcal{A}} h \lesssim s\sqrt{\log p/n_0}$, and $n_0 \lesssim n_{\mathcal{A}}$. Let the Oracle Trans-sub-Coda-Lasso estimator $\hat{\boldsymbol{\beta}}$ be computed with $\lambda_w = C_w\sqrt{\log p/(n_0+n_{\mathcal{A}})}$ and $\lambda_\delta = C_\delta\sqrt{\log p/n_0}$, where C_w and C_δ are sufficiently large positive constants. Then, with the probability at least $1-\exp\{-c\log p\}$ for some constant c, we have

$$\frac{1}{n_0}\|\mathbf{Z}^{(0)}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2^2 \vee \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2^2 = O\left(\frac{s\log p}{n_0+n_{\mathcal{A}}} + \frac{s\log p}{n_0} \wedge C_\Sigma^{\mathcal{A}} h\sqrt{\log p/n_0} \wedge (C_\Sigma^{\mathcal{A}} h)^2\right).$$

Theorem 1 demonstrates that the convergence rate of Oracle Trans-sub-Coda-Lasso estimator relies on $n_{\mathcal{A}}$ and $C_\Sigma^{\mathcal{A}} h$, which represent the number of informative source samples and the similarity between informative source studies and the target study, respectively. Ignoring the non-compositional part and grouping structure, the target-only estimator whose rate of ℓ_2 error bounded on $\boldsymbol{\beta}$ is $s\log p/n_0$ (Lin et al. (2014)). Under the conditions $n_{\mathcal{A}} \gg n_0$ and $C_\Sigma^{\mathcal{A}} h \ll s\sqrt{\log p/n_0}$, the oracle transfer learning estimator enjoys a faster convergence rate. Even in the worst-case scenario that there is zero informative source, that is, $\mathcal{A}=\emptyset$, the convergence rate of the oracle transfer learning estimator is no slower than that of the target-only estimator. Theorem 1 further indicates that model performance improves as the heterogeneity between the informative sources and the target decreases, as quantified by $C_\Sigma^{\mathcal{A}}$.

Next we will consider informative set $\mathcal{A}$ selection problem. The corresponding population version of $\hat{Q}(\boldsymbol{\beta})$ defined in (2.11) is

$$Q(\boldsymbol{\beta}) = \mathbb{E}\{(\boldsymbol{y}^{(0)}-\mathbf{Z}^{(0)}\boldsymbol{\beta})^\top(\boldsymbol{y}^{(0)}-\mathbf{Z}^{(0)}\boldsymbol{\beta})\},$$

where the expectation is taken with respect to the target distribution. We define $\boldsymbol{w}^{(0,k)}$ as the underlying fusion coefficient vector between the target and the source k arising from Step 3 of the Algorithm 2 ,

$$\boldsymbol{w}^{(0,k)} = \boldsymbol{\Omega}_c^{(0,k)} \sum_{j=0,k} \alpha_j \boldsymbol{\Sigma}^{(j)} \boldsymbol{w}^{*(j)},$$

where $\boldsymbol{\Omega}_c^{(0,k)}$ is the Moore-Penrose inverse of $\boldsymbol{\Sigma}^{(0,k)} = \sum_{j=0,k} \alpha_j \boldsymbol{\Sigma}^{(j)}$.

In order to ensure consistent informative set selection, we impose a general assumption.

Assumption 4 .(Weak sparsity condition) With s' such that $\|\boldsymbol{w}^{*(k)}\|_0 \leq s'$ for all $k \in \mathcal{A}^c$, there exists a set $S'_k \subset \{1, \ldots, p\}$ such that $|S'_k| \leq s'$ and $\|\boldsymbol{w}^{(0,k)}_{S'^c_k}\|_1 \leq h'$ for any $k \in \mathcal{A}^c$ with $h' = o(1)$.

Assumption 4 assumes that the sparsity pattern of $\boldsymbol{w}^{*(k)}$, $k \in \mathcal{A}^c$ is similar to $\boldsymbol{\beta}^*$, which implies that the corresponding fusion coefficient $\boldsymbol{w}^{(0,k)}$ remain to be "weakly" sparse (Tian and Feng, 2023; Jin et al., 2024). To identify informative auxiliary studies, our selection rule compares the target testing error of the pooled estimator $\hat{\boldsymbol{w}}^{(0,k)}$ with that of the target-only estimator $\hat{\boldsymbol{\beta}}^{(0)}_{\mathcal{I}}$. Therefore, at the population level, the key quantity is the discrepancy between the pooled population minimizer $\boldsymbol{w}^{*(0,k)}$ and $\boldsymbol{\beta}^*$. Intuitively, for $k \in \mathcal{A}$, pooling study k with the target study should only introduce a small perturbation, so that $Q(\boldsymbol{w}^{*(0,k)}) - Q(\boldsymbol{\beta}^*)$ remains small; in contrast, for $k \in \mathcal{A}^c$, the corresponding discrepancy should be sufficiently large so that the resulting increase in target prediction risk can still be detected after accounting for stochastic errors.

To formulate this idea in an interpretable way, we express the separation condition through the discrepancy between the auxiliary coefficient $\boldsymbol{w}^{*(k)}$ and the target coefficient $\boldsymbol{\beta}^*$, and then connect it to the pooled-target discrepancy through a transfer factor. This leads

to the following identifiability assumption.

Assumption 5 .(Identifiability of $\mathcal{A}$). Denote $s^* = s \vee s'$, $h^* = C_\Sigma^{\mathcal{A}} h \vee h'$ and $\underline{n} = \min_{k=1}^K n_k$, $\eta^2 = o(n_0)$, λ_{min} as the minimum eigenvalue of the matrix $\mathbf{\Sigma}^{(0)} + \rho\tilde{\mathbf{C}}\tilde{\mathbf{C}}^\top$ for any constants $\rho > 0$, and

$$\kappa_1 = \sqrt{\frac{\log p}{n_0}}\left(s^*\sqrt{\frac{\log p}{n_0+\underline{n}}} + h^*\right) + \sqrt{s^*}\left(1+\sqrt{\frac{\eta^2}{n_0}}\right)\sqrt{s^*\frac{\log p}{n_0+\underline{n}} + h*\sqrt{\frac{\log p}{n_0+\underline{n}}}}$$

and

$$\kappa_2 = s^*\sqrt{\frac{\log p}{n_0}}.$$

For any $k \in \mathcal{A}^c$, there exists a positive constant c_ϵ such that

$$\|\boldsymbol{w}^{*(k)} - \boldsymbol{\beta}^*\|_2^2 \geq \lambda_{min}^{-1}\{c_\epsilon Q(\boldsymbol{\beta}^*) + 4(\kappa_1 \vee \kappa_2))\}\kappa_3,$$

where $\kappa_3 = \max_{k\in\mathcal{A}^c}\{\|\alpha_k^{-1}\mathbf{\Omega}_c^{(k)}\mathbf{\Sigma}^{(0,k)}\|_2^2\}$, $\alpha_k = n_k/(n_0+n_k)$, $\mathbf{\Sigma}^{(0,k)} = \sum_{j=0,k}\alpha_j\mathbf{\Sigma}^{(j)}$ and $\mathbf{\Omega}_c^{(k)}$ is the Moore-Penrose inverse of $\mathbf{\Sigma}^{(k)}$. Meanwhile, we require $h^2 = o(Q(\boldsymbol{\beta}^*))$.

Assumption 5 is an identifiability condition ensuring that non-informative auxiliary studies can be distinguished from informative ones. It is mainly imposed as a sufficient condition for the consistency of the proposed source detection procedure. We emphasize that exact recovery of the informative set $\mathcal{A}$ is not necessarily required for improving estimation or prediction accuracy; in practice, some auxiliary studies outside $\mathcal{A}$ may still contain partially transferable information.

The quantities κ_1 and κ_2 represent stochastic error levels: κ_1 controls the estimation error incurred when replacing the pooled population minimizer by its sample counterpart,

while κ_2 controls the discrepancy between the empirical testing error $\hat{Q}(\cdot)$ and the population risk $Q(\cdot)$. The quantity κ_3 is a transfer constant that links the auxiliary-target discrepancy to the pooled-target discrepancy. In particular, κ_3 is motivated by the inequality $\|\boldsymbol{w}^{*(k)} - \boldsymbol{\beta}^*\|_2^2 \leq \left\|\alpha_k^{-1}\boldsymbol{\Omega}_c^{(k)}\boldsymbol{\Sigma}^{(0,k)}\right\|_2^2 \|\boldsymbol{w}^{*(0,k)} - \boldsymbol{\beta}^*\|_2^2$. Therefore, a sufficiently large auxiliary-target separation for $k \in \mathcal{A}^c$, after accounting for the transfer effect κ_3, implies a detectable pooled-target separation at the population level. Combined with the fact that the pooled-target discrepancy is of order $O(h^2)$ for $k \in \mathcal{A}$, Assumption 5 guarantees that the empirical testing errors of informative and non-informative auxiliary studies remain separable with high probability.

In particular, for any $k \in \mathcal{A}^c$, Assumption 5 implies that $c_\epsilon \leq \lambda_{\min}\|\boldsymbol{w}^{(0,k)} - \boldsymbol{\beta}^*\|_2^2/Q(\boldsymbol{\beta}^*)$. This implication motivates the practical estimation of c_ϵ in Algorithm 4.

The following theorem shows selection consistency property.

Theorem 2. ***(Selection consistency of $\hat{\mathcal{A}}$)*** *For Algorithm 2 (Trans-sub-Coda-Lasso), suppose Assumptions 1-5 hold. When $\kappa_1 \vee \kappa_2 = o(1)$, $\hat{\mathcal{A}}$ obtained in Algorithm 2 is consistent in identifying $\mathcal{A}$, such that*

$$Pr(\hat{\mathcal{A}} = \mathcal{A}) \geq 1 - \tilde{c}_1 \exp\{-c_1\eta^2\} - \tilde{c}_2 \exp\{-c_2 \log p\}$$

where c_1 ,$\tilde{c}_1$, c_2 and $\tilde{c}_2$ are some positive constants.

Theorem 2 guarantees that Algorithm 2 (Trans-sub-Coda-Lasso) has the same high-probability upper bounds of ℓ_1/ℓ_2-estimation error as those in Theorem 1 under the same conditions. All proofs are provided in Supplementary Material.

4. Numerical simulations

In this section, we empirically evaluate the performance of the proposed methods: Oracle-Trans-subCodalasso with known informative set (proposed in Algorithm 1) and Trans-subCodalasso with unknown informative set (proposed in Algorithm 2). We also compare them with several existing approaches. The subCodalasso estimator defined in (2.9) uses only target samples.The Pooled-sub-Codalasso estimator and Naive-Trans-subCodalasso estimator use all sources for subCodalasso and Oracle-Trans-subCodalasso, respectively. Trans-Lasso/GLM-alr means the estimators from Li et al. (2022) and Tian and Feng (2023) using additive log-ratio (alr) transformation of compositional data, respectively. The R code for this study can be found at https://github.com/luoxiaojing2578/TLsubCodalasso. More implementation details can be found in Section S.5.1 in the Supplementary Materials.

4.1 Numerical setup

In our simulation, we generate data from the true model:

$$\boldsymbol{y}^{(k)} = \mathbf{N}^{(k)}\boldsymbol{a}^{*(k)} + \tilde{\tilde{\mathbf{Z}}}^{(k)}\boldsymbol{\theta}^{*(k)} + \boldsymbol{\varepsilon}^{(k)}, \qquad \mathbf{C}_s^{\top}\boldsymbol{\theta}^{*(k)} = \mathbf{0},$$

for $k = 0, 1, \ldots, K$, where $k = 0$ for our target model and $k = 1, \ldots, K$ for source models. $\mathbf{N}^{(k)}$ contains an intercept and $q' - 1$ binary covariates generated from a Bernoulli distribution with probability 0.5. We follow a similar approach to generate compositional data as in Lin et al. (2014) and Shi et al. (2016). We generate an $n_k \times q$ data matrix $\mathbf{U}^{(k)} = (u_{ij}^{(k)})$ from a multivariate normal distribution $N_q(\boldsymbol{\nu}, \Sigma_U)$, and obtain compositional data $\mathbf{X}^{(k)} = (x_{i,gs}^{(k)})$ by transformation $x_{i,gs}^{(k)} = \frac{\exp(u_{i,gs}^{(k)})}{\sum_{s=1}^{p_g}\exp(u_{i,gs}^{(k)})}$, for $g = 1, \ldots, G$ and the covari-

ates matrix $\tilde{\tilde{\mathbf{Z}}}^{(k)} = (\tilde{\tilde{\mathbf{z}}}_{i,gs}^{(k)})$ by $\tilde{\tilde{\mathbf{z}}}_{i,gs} = \log\{x_{i,gs}^{(k)}/(\Pi_{s=1}^{p_g} x_{i,gs}^{(k)})^{1/p_g}\}$. To reflect that the components of a composition differ by order of magnitude, we also take $\boldsymbol{\nu} = (\nu_j)$ with $\nu_j = \log(0.5q)$ for $j = 1, \ldots, 5$ and $\nu_j = 0$ otherwise. The target model is set as $\boldsymbol{a}^{*(0)} = (0.5, -0.1, 0, \ldots, 0)^\top$ and $\boldsymbol{\theta}^{*(0)} = (1, -0.8, 0.4, 0, 0, -0.6, 0, 0, 0, 0, -1.5, 0, 1.2, 0, 0, 0.3, 0, \ldots, 0)^\top$. The regression coefficient $\boldsymbol{\theta}^{*(0)}$ used in the simulation satisfies the following 8 linear constraints:

$$\sum_{j=1}^{10} \theta_j^{*(0)} = 0, \sum_{j=11}^{16} \theta_j^{*(0)} = 0, \sum_{j=17}^{20} \theta_j^{*(0)} = 0, \sum_{j=21}^{23} \theta_j^{*(0)} = 0,$$

$$\sum_{j=24}^{30} \theta_j^{*(0)} = 0, \sum_{j=31}^{32} \theta_j^{*(0)} = 0, \sum_{j=33}^{40} \theta_j^{*(0)} = 0, \sum_{j=41}^{q} \theta_j^{*(0)} = 0.$$

$\boldsymbol{\varepsilon}^{(k)}$ is an n_k-dimensional vector of independent noise term, each component following a normal distribution with mean zero and standard deviation $\sigma^{(k)} = \left\|\mathbf{N}^{(k)}\boldsymbol{a}^{*(k)} + \tilde{\tilde{\mathbf{Z}}}^{(k)}\boldsymbol{\theta}^{*(k)}\right\| / \left(3\sqrt{n_k}\right)$. In the simulation studies, both the covariates and responses are mean-centered. Consider different numbers of informative sources $|\mathcal{A}| \in \{0, 5, 10, 15, 20\}$. To simulate model and covariate shift, we consider the following different configurations for the source model.

Model Shift. To examine the impact of source task characteristics, the coefficient vector $(\boldsymbol{a}^{*(k)}, \boldsymbol{\theta}^{*(k)})$ for each source task is constructed as follows. We randomly select d_a non-compositional variables, excluding the intercept term, and add perturbations to their coefficients. The compositional coefficients for the source is set as follows.

(1) Fixed setting. For given $\mathcal{A}$, if $k \in \mathcal{A}$ and $g \in H$, let

$$\theta_{gs}^{*(k)} = \theta_{gs}^{*(0)} + \xi_s I(s = H_{g1}^{(k)}) - \xi_s I(s = H_{g2}^{(k)}),$$

and if $k \notin \mathcal{A}$ and $g \in \{1, \ldots, G\}$,

$$\theta_{gs}^{*(k)} = \theta_{gs}^{*(0)} + \zeta_s I(s = H_{g3}^{(k)}) - \zeta_s I(s = H_{g4}^{(k)}),$$

where H is random subsets of $\{1, \ldots, G\}$ with $|H| = d/2 = \{2, 6\}$, and $(H_{g1}^{(k)}, H_{g2}^{(k)})$ is sampled uniformly without replacement from the positions in the g-th group, and $(H_{g3}^{(k)}, H_{g4}^{(k)})$ is sampled in the same way, overlap between the two pairs is allowed. ξ_s and ζ_s denote the quality of effective and ineffective sources, respectively. They are set according to one of the following two ways: **Scenario I.** $\xi_s \sim \mathcal{U}(0,\, 0.05)$ and $\zeta_s \sim \mathcal{U}(0.5,\, 1)$ which is good effective sources and poor ineffective sources; **Scenario II.** $(\xi_s, \zeta_s)^\top = (0.1, 0.5)^\top$, which is not too good effective sources and not too bad ineffective sources. The perturbation magnitudes of the non-compositional coefficients are consistent with those of the compositional coefficients.

(2) Random setting. For a given $\mathcal{A}$, and $g \in \{1, \ldots, G-1\}$, let

$$\theta_{gs}^{*(k)} = \theta_{gs}^{*(0)} + \xi_s I(s = H_{g1}^{(k)}) - \xi_s I(s = H_{g2}^{(k)}),$$

and

$$\theta_{Gs}^{*(k)} = \theta_{Gs}^{*(0)} + \xi_s I(s \in H_{G5}^{(k)}) - \xi_s I(s \in H_{G6}^{(k)}),$$

where $|H_{G5}^{(k)}| = |H_{G6}^{(k)}| = q/4 - 2(G-1)$ and $\xi_s \sim \mathcal{U}(0,\, d/q)$, $d = 4, 12$.

if $k \notin \mathcal{A}$ and $g \in \{1, \ldots, G-1\}$, let

$$\theta_{gs}^{*(k)} = \theta_{gs}^{*(0)} + \xi_s I(s = H_{g3}^{(k)}) - \xi_s I(s = H_{g4}^{(k)}),$$

and

$$\theta_{Gs}^{*(k)} = \theta_{Gs}^{*(0)} + \xi_s I(s \in H_{G7}^{(k)}) - \xi_s I(s \in H_{G8}^{(k)}),$$

where $|H_{G7}^{(k)}| = |H_{G8}^{(k)}| = q/2 - 2(G-1)$ and $\xi_s \sim \mathcal{U}(0.5,\, 0.5+5d/q)$, $d = 4, 12$. The perturbation magnitudes of the non-compositional coefficients follow the Scenario I setting described above.

Covariate Shift. To demonstrate the robustness of the proposed methods to covariate shifts, we consider two settings with different covariate distributions.

(a) Homogeneous design. Each $\boldsymbol{u}_i^{(k)} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma}_U)$, for $k = 0, \ldots, K$, where $\boldsymbol{\Sigma}_U = (\rho^{|i-j|})$ with $\rho = 0.2$ or 0.5.

(b) Heterogeneous design. Each $\boldsymbol{u}_i^{(k)} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma}_U^{(k)})$, where $\boldsymbol{\Sigma}_U^{(k)} = (\mathbf{A}^{(k)})^\top \mathbf{A}^{(k)} + \mathbf{I}$. Here $\mathbf{A}^{(k)}$ is a random matrix with each entry equal to 0.3 with probability 0.3 and equal to 0 with probability 0.7 and $\boldsymbol{\Sigma}_U^{(0)} = (\rho^{|i-j|})$ with $\rho = 0.5$.

To evaluate estimation accuracy, we calculate two metrics based on $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ for each method: the ℓ_2 estimation error (L_2-error) and the prediction error (PE), defined as

$$L_2\text{-error} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2, \quad \text{PE}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{y}^{(0)} - \mathbf{Z}^{(0)}\hat{\boldsymbol{\beta}}\|_2^2/n_0.$$

To assess variable selection performance, we report the true positive rate (TPR) and false discovery rate (FDR). Let $S = \{j : \beta_j^* \neq 0\}$ denote the true active set and $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ denote the estimated active set. The TPR and FDR are defined as

$$\text{TPR} = \frac{|S \cap \hat{S}|}{|S|}, \quad \text{FDR} = \frac{|S^c \cap \hat{S}|}{|\hat{S}| \vee 1}.$$

A larger TPR indicates better recovery of true signals, whereas a smaller FDR indicates better control of false discoveries.

4.2 Simulation results

In this simulation, different dimension combinations $(q', q) = (10, 100), (0, 100), (10, 50)$ and $(10, 200)$ are considered. We set $n_0 = n_1 = \cdots = n_K = 100$ with $K = 20$. To evaluate the performance of Trans-subCodalasso under the varying number of informative source domains, we consider $|\mathcal{A}| \in \{0, 5, 10, 15, 20\}$. We add perturbations to the one coefficient of non-compositional variables, i.e $d_a = 1$, when $q' \neq 0$. In the main text, we use the setting $(q', q) = (10, 100)$ to illustrate the results. Additional results under other dimension combinations and simulation settings are provided in Section S5.2 of the Supplementary Material.

We first evaluate the ℓ_2 estimation errors of these methods. In Figure 1-3, we report the ℓ_2 estimation errors under all considered source-domain settings for both homogeneous and heterogeneous designs. Each point is average from 100 independent simulations.

As expected, the estimation performance of subCodalasso using only the target samples remains almost unchanged as the number of informative source studies, $|\mathcal{A}|$, increases. In contrast, the pooled method and the other five transfer learning methods show decreasing estimation errors as $|\mathcal{A}|$ increases, indicating that they can effectively borrow information from transferable source studies. As d increases, the difference between the target and the sources increases and the ℓ_2 estimation error of the six methods increases. When informative sources are available ($i.e. |\mathcal{A}| > 0$), all transfer learning methods substantially outperform subCodalasso using only the target samples.

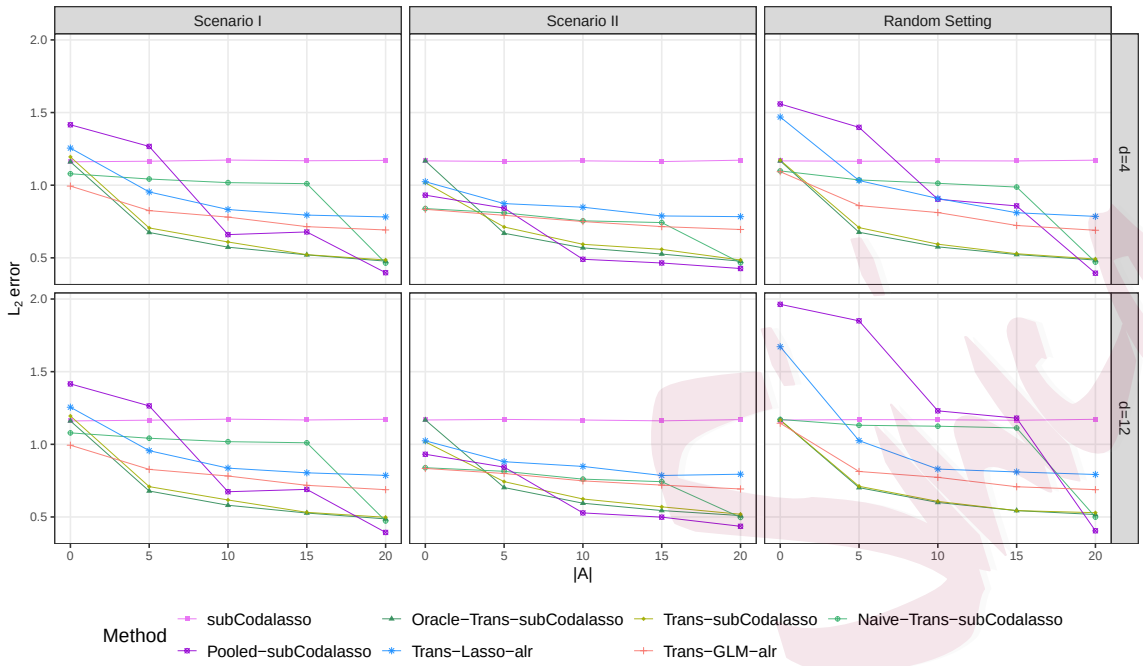

Figure 1: The average ℓ_2 estimation error of the seven methods under different settings for coefficient vector with homogeneous covariance matrices ($\rho = 0.2$, $(q', q) = (10, 100)$).

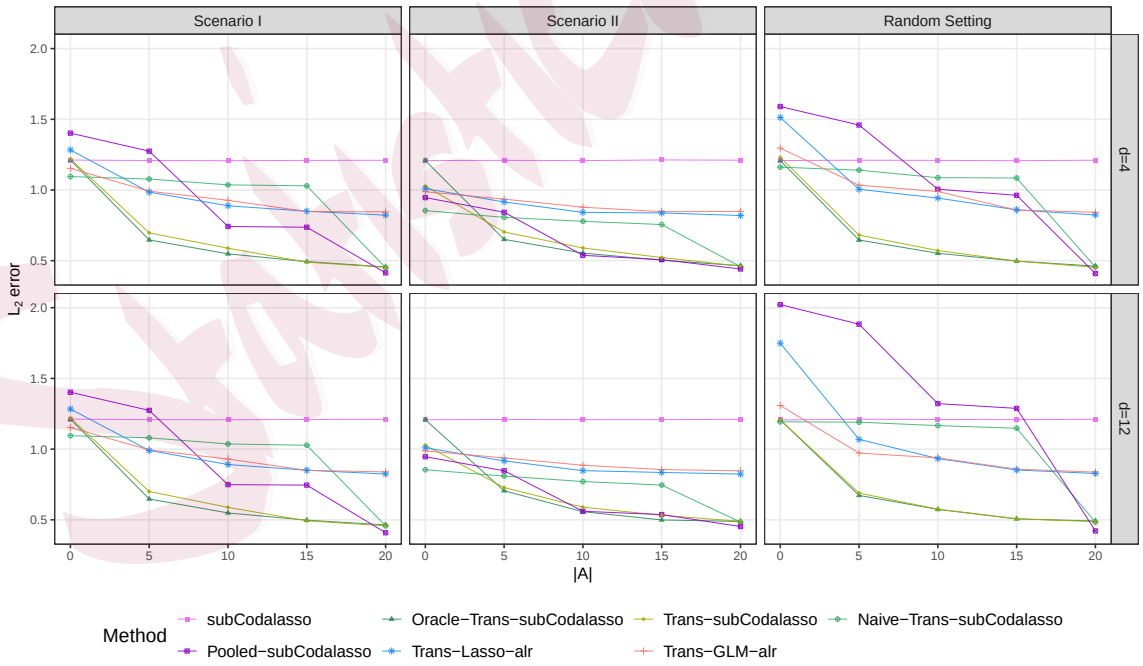

Figure 2: The average ℓ_2 estimation error of the seven methods under different settings for coefficient vector with homogeneous covariance matrices ($\rho = 0.5$, $(q', q) = (10, 100)$).

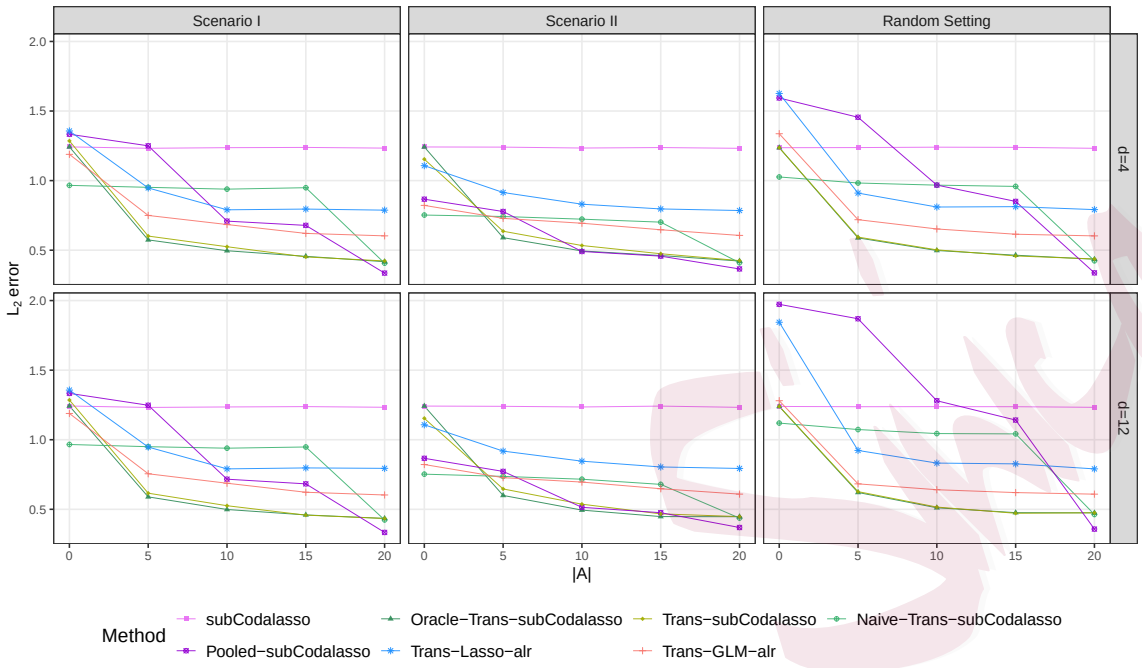

Figure 3: The average ℓ_2 estimation error of the seven methods under different settings for coefficient vector with heterogenous covariance matrices and $(q', q) = (10, 100)$.

Moreover, Oracle-Trans-subCodalasso achieves the best or nearly the best performance, which is expected because it assumes prior knowledge of the informative source set. Trans-subCodalasso closely approaches the oracle performance, particularly for moderate or large $|\mathcal{A}|$, and substantially improves over subCodalasso, demonstrating the advantage of transfer learning. When the difference between informative and non-informative sources is relatively large, as in Scenario I, or when the perturbations are randomly dense, as in the random setting, the proposed Oracle-Trans-subCodalasso and Trans-subCodalasso show more pronounced advantages over the competing transfer learning methods. In these cases, the performance gap between the naive method and Trans-subCodalasso further demonstrates the importance and effectiveness of source selection.

We also note that the auxiliary sources labeled as "non-informative" in the simulations are not necessarily entirely non-informative for transfer. For instance, in Scenario II where

the coefficient shift is 0.5 (the second column of Figure 2), Trans-subCodalasso performs better than the oracle method when $|\mathcal{A}| = 0$. This is mainly because the discrepancy between these auxiliary sources and the target study is relatively small. Although such sources are excluded from the oracle informative set $\mathcal{A}$, they may still carry useful transferable information. Therefore, by adaptively incorporating some mildly heterogeneous sources, Trans-subCodalasso may achieve performance comparable to, or even slightly better than, Oracle-Trans-subCodalasso. The behavior of Naive-Trans-subCodalasso under the same scenarios provides further support for this interpretation. On the other hand, ℓ_2 estimation error of Trans-subCodalasso estimator is slightly larger than that of subCodalasso estimator when all sources are poor ineffective quality (Scenario I under homogeneous with $\rho = 0.2$ and heterogeneous setting). A similar phenomenon can also be observed for Trans-Lasso-alr.

When all auxiliary sources are informative ($|\mathcal{A}| = K = 20$), the pooled method surpasses all transfer learning methods, including our proposed method. For two-step transfer learning methods, the second-step correction using target data is susceptible to auxiliary source bias. Specifically, the correction brings limited gains under small bias (e.g., $|\mathcal{A}| = K = 20$ with bias of $0 \sim 0.05$ and 0.1, comparing the pooled method with Oracle-Trans-subCodalasso), whereas its advantages become prominent under large bias (e.g., $|\mathcal{A}| = 0$ with bias ranging from 0.5 to 1.0,comparing the pooled method with Naive-Trans-subCodalasso). Furthermore, the performance of this correction procedure is based on the quality of the target data, as elaborated in Section S5.2.3 in Supplementary Materials.

It is worth noting that, under the homogeneous setting with $\rho = 0.2$ and Scenario II, the pooled method also performs well once the number of informative source domains $|\mathcal{A}|$ exceeds 10. This is mainly because, in this case, a large proportion of the source domains are

Table 1: Means and standard errors (in parentheses) of true positive rate (TPR) and false discovery rate (FDR) for seven methods under random setting for coefficient vector with heterogeneous covaraiance matrices and $q = 100$ based on 100 replicates.

Type	Method	TPR						FDR					
		$d = 4$			$d = 12$			$d = 4$			$d = 12$		
		$\lvert\mathcal{A}\rvert = 5$	$\lvert\mathcal{A}\rvert = 10$	$\lvert\mathcal{A}\rvert = 15$	$\lvert\mathcal{A}\rvert = 5$	$\lvert\mathcal{A}\rvert = 10$	$\lvert\mathcal{A}\rvert = 15$	$\lvert\mathcal{A}\rvert = 5$	$\lvert\mathcal{A}\rvert = 10$	$\lvert\mathcal{A}\rvert = 15$	$\lvert\mathcal{A}\rvert = 5$	$\lvert\mathcal{A}\rvert = 10$	$\lvert\mathcal{A}\rvert = 15$
$q' = 0$	sC	0.98(0.06)	0.98(0.06)	0.98(0.06)	0.98(0.06)	0.98(0.06)	0.98(0.06)	0.83(0.05)	0.83(0.05)	0.83(0.05)	0.83(0.05)	0.82(0.06)	0.83(0.05)
	OTC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.59(0.21)	0.51(0.24)	0.41(0.25)	0.62(0.22)	0.55(0.23)	0.46(0.25)
	TC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	**0.60(0.22)**	**0.50(0.24)**	**0.39(0.25)**	**0.62(0.22)**	**0.56(0.22)**	**0.47(0.25)**
	NTC	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.01)	1.00(0.01)	1.00(0.00)	0.81(0.07)	0.79(0.09)	0.76(0.13)	0.81(0.07)	0.81(0.07)	0.78(0.09)
	PC	0.99(0.03)	1.00(0.03)	1.00(0.00)	0.93(0.10)	0.96(0.08)	1.00(0.03)	0.57(0.18)	0.46(0.23)	0.34(0.22)	0.54(0.21)	0.45(0.24)	0.32(0.22)
	TL-a	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.86(0.15)	0.90(0.07)	0.88(0.10)	0.78(0.25)	0.87(0.12)	0.86(0.16)
	TG-a	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.39(0.20)	0.31(0.22)	0.26(0.24)	0.40(0.20)	0.34(0.21)	0.31(0.24)
$q' = 10$ compositional covariates	sC	0.98(0.06)	0.97(0.07)	0.97(0.06)	0.97(0.06)	0.97(0.07)	0.97(0.06)	0.85(0.06)	0.85(0.06)	0.85(0.06)	0.85(0.06)	0.85(0.06)	0.85(0.06)
	OTC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.52(0.24)	0.45(0.27)	0.36(0.27)	0.55(0.22)	0.49(0.25)	0.40(0.27)
	TC	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	**0.52(0.24)**	**0.45(0.27)**	**0.36(0.27)**	**0.55(0.22)**	**0.49(0.25)**	**0.40(0.26)**
	NTC	1.00(0.01)	1.00(0.00)	1.00(0.00)	0.99(0.04)	1.00(0.02)	1.00(0.01)	0.65(0.19)	0.62(0.20)	0.50(0.26)	0.70(0.17)	0.68(0.17)	0.60(0.23)
	PC	0.97(0.06)	0.99(0.03)	1.00(0.01)	0.86(0.13)	0.92(0.10)	0.99(0.05)	0.40(0.23)	0.35(0.24)	0.21(0.20)	0.35(0.26)	0.32(0.24)	0.16(0.20)
	TL-a	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.82(0.22)	0.82(0.21)	0.75(0.28)	0.75(0.25)	0.83(0.18)	0.78(0.23)
	TG-a	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.43(0.21)	0.38(0.21)	0.35(0.21)	0.45(0.19)	0.42(0.18)	0.38(0.20)
$q' = 10$ all covariates	sC	0.93(0.08)	0.92(0.08)	0.92(0.08)	0.92(0.08)	0.92(0.08)	0.92(0.08)	0.84(0.06)	0.84(0.06)	0.84(0.06)	0.84(0.06)	0.84(0.06)	0.84(0.06)
	OTC	0.90(0.03)	0.90(0.04)	0.90(0.03)	0.90(0.03)	0.90(0.03)	0.90(0.03)	0.51(0.22)	0.43(0.26)	0.35(0.27)	0.54(0.22)	0.47(0.25)	0.39(0.26)
	TC	0.90(0.03)	0.90(0.04)	0.90(0.03)	0.90(0.03)	0.90(0.03)	0.90(0.03)	**0.51(0.22)**	**0.43(0.26)**	**0.35(0.26)**	**0.54(0.22)**	**0.47(0.25)**	**0.39(0.26)**
	NTC	0.91(0.04)	0.91(0.04)	0.90(0.04)	0.91(0.06)	0.91(0.05)	0.91(0.04)	0.63(0.19)	0.60(0.20)	0.49(0.26)	0.68(0.17)	0.66(0.17)	0.58(0.23)
	PC	0.86(0.05)	0.88(0.03)	0.89(0.01)	0.78(0.10)	0.83(0.08)	0.88(0.04)	0.37(0.22)	0.33(0.23)	0.19(0.19)	0.33(0.24)	0.30(0.22)	0.15(0.19)
	TL-a	0.82(0.08)	0.86(0.05)	0.86(0.05)	0.82(0.07)	0.85(0.05)	0.85(0.05)	0.68(0.31)	0.80(0.20)	0.78(0.18)	0.62(0.31)	0.73(0.24)	0.77(0.18)
	TG-a	0.77(0.05)	0.78(0.03)	0.78(0.02)	0.79(0.03)	0.78(0.03)	0.78(0.02)	0.41(0.22)	0.37(0.23)	0.31(0.20)	0.41(0.21)	0.37(0.24)	0.30(0.20)

Note: sC: subCodalasso; OTC: Oracle-Trans-subCodalasso; TC: Trans-subCodalasso; NTC: Naive-Trans-subCodalasso; PC: Pooled-subCodalasso; TL-a: Trans-Lasso-alr; TG-a: Trans-GLM-alr.

informative, which favors the pooled estimator. Moreover, under Scenario II, the coefficient discrepancy between the non-informative source domains and the target domain is set to 0.5, indicating that the non-informative sources are not severely biased from the target domain. Consequently, pooling all source domains can still yield competitive performance.

Next, we turn our attention to the performance of the variable selection. We consider three different numbers of informative sources, $|\mathcal{A}| = 5, 10, 15$. Table 1 presents the TPR and FDR under model shift in the random setting and covariate shift in the heterogeneous design for $(q', q) = (0, 100)$ and $(q', q) = (10, 100)$. For $(q', q) = (10, 100)$, we report the results based on compositional covariates only as well as on all covariates.

According to the results in Table 1, the Trans-subCodalasso method performs remarkably similarly to the oracle benchmark method in all simulation settings. This indicates that the proposed method can effectively identify and utilize informative source domains without relying on prior knowledge. From the perspective of TPR, all methods perform remarkably

well for the compositional covariates. After incorporating non-compositional covariates, the TPR for total variables of all methods declines to some extent; nevertheless, the proposed method exhibits a much smaller drop compared with its competitors. In terms of FDR, the Pooled method and Trans-GLM-alr alternately yield the best performance, while our method closely follows with robust and stable FDR control. Notably, when focusing only on the compositional covariates, the FDR of the proposed method decreases after adding non-compositional variables. By contrast, the FDR of Trans-GLM-alr increases rather than decreases. This finding indicates that the introduction of non-compositional covariates does not interfere with the variable selection of compositional predictors in our method. Instead, by providing additional explanatory information — especially a small number of truly relevant non-compositional variables — it reduces the probability of falsely selecting irrelevant compositional variables.

We further evaluate Trans-subCodalasso under several simulation settings in Section S5.2 of the Supplementary Materials, focusing on prediction performance, compositional covariate constraints, exclusion of non-compositional covariates, and target-domain sample size.

5. Application to gut microbiome data

5.1 Data description and preprocessing

We apply the proposed method to gut microbiome data reported by Pasolli et al. (2016), publicly available from the http://segatalab.cibio.unitn.it/tools/metaml. The data consists of two studies involving subjects with ulcerative colitis and healthy controls, from whom fecal samples were collected for shotgun metagenomic profiles and demographic variables, including BMI, age, and sex, were recorded. We focus on genus-level microbial compositions

and use phylum-level information to define subcompositional structures.

Among the original 292 genera, those with zero counts in more than 90% of samples are removed, and phyla containing only one genus after filtering are excluded, leaving 81 genera from four phyla. After removing subjects with missing values, the ulcerative colitis and healthy datasets contain 111 and 628 subjects, respectively. Since sex-dependent differences in the human microbiome have been reported, we further stratify the data by disease status and sex into four subgroup datasets: UC male, UC female, Normal male and Normal female, with sample sizes 39, 72, 310 and 318, respectively. The zero counts in the data are replaced with 0.5, which corresponds to the maximum rounding error (Aitchison (2003), §11.5), and the resulting counts are transformed into compositional data. In the transfer-learning analysis, each subgroup dataset is used in turn as the target dataset, with the remaining three treated as source datasets.

We also incorporate age as an additional covariate, since gut microbiome composition may vary with age and affect host metabolic status (Bradley and Haran, 2024). For the subcomposition-based analysis, the 81 genera are grouped by phylum into Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria, containing 9, 12, 44 and 16 genera, respectively. The genus-level count data are normalized within each phylum to form subcompositional data, and subcompositional constraints are imposed separately within each phylum. Accordingly, we consider four model settings: subcomposition with age, subcomposition without age, composition with age, and composition without age.

To motivate our transfer learning framework, we perform preliminary analyses of similarity and heterogeneity across the four subgroup datasets. (see Section 5.3.1 in Supplementary Material for details). Our results indicate that while the datasets share common features

such as comparable BMI distributions, they exhibit substantial differences in age characteristics and BMI-microbiome association patterns. These findings confirm that the datasets are related yet heterogeneous, justifying the need for a transfer learning framework that can adaptively borrow information from relevant sources instead of naively pooling all data together. This provides clear motivation for our proposed Trans-subCodalasso method with transferable source detection.

5.2 Prediction performance

We next evaluate the performance of Trans-subCodalasso and other competing methods in predicting BMI using gut microbial compositional data. In each analysis, one subgroup dataset is treated as the target dataset and the remaining three subgroup datasets are used as source datasets. The 81 genus-level compositional variables are used as predictors ($p = 81$), and BMI is taken as the response. We compare the proposed Trans-subCodalasso with subCodalasso, Pooled-subCodalasso, Naive Trans-subCodalasso, Trans-Lasso-alr, and Trans-GLM-alr. For each target subgroup, we repeat the BMI prediction experiment over 100 random splits of the target dataset, using 70% of the observations for training and the remaining 30% for testing. Prediction accuracy is summarized by the average test error across 100 splits.

We evaluate prediction performance under four modeling settings and summarize each method by its prediction error relative to the corresponding target-only model within the same setting, as shown in Figure 4.

Trans-subCodalasso achieves lower prediction error than the corresponding target-only model in nearly all cases, with the only exception being the UC male target under the

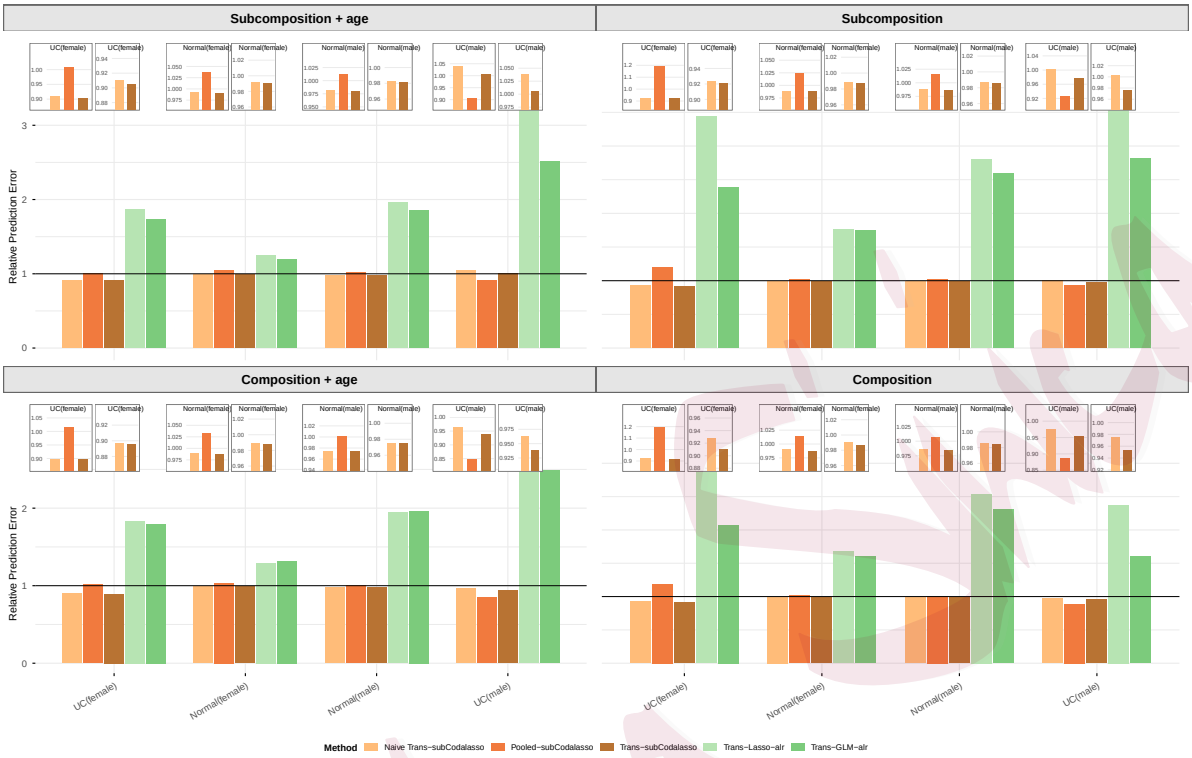

Figure 4: Relative prediction errors of different transfer learning methods relative to subCodalasso under four model settings

sub-compositional model with age adjustment. For UC male, the pooled method performs best across all four settings, possibly because this subgroup has the smallest target sample size ($n = 39$), compared with UC female ($n = 72$), Normal female ($n = 318$), and Normal male ($n = 310$). In such small-target settings, direct pooling may provide stronger variance reduction. We also observe that Trans-subCodalasso and Naive-Trans-subCodalasso yield similar performance for some targets, particularly the two Normal subgroups, suggesting that the additional benefit of adaptive source selection may be less pronounced when the target sample size is relatively large or when most candidate sources are transferable. In contrast, Trans-Lasso-alr and Trans-GLM-alr consistently perform worse than target-only across all settings, suggesting potential negative transfer under the alr-based representations. These results indicate that Trans-subCodalasso is generally effective and more robust than the alr-

based transfer approaches, although its empirical advantage may depend on target sample size, source-target similarity, and model complexity.

Original prediction errors under the four modeling settings are reported in the Supplementary Materials (Table 19). Adding age generally reduces prediction error for most target subgroups, except for UC male, with the largest improvements observed for the Normal targets. In contrast, the compositional and sub-compositional specifications show no consistent difference in prediction error, suggesting that the benefit of sub-composition modeling may depend on the target subgroup and transfer setting.

o further interpret the results, we examine the source-selection behavior of Trans-subCodalasso and the microbial features selected by the proposed method, with details provided in Sections S5.3.2 and S5.3.3 of the Supplementary Materials, respectively.

6. Discussions

This paper develops a transfer learning framework for high-dimensional regression with both subcompositions and non-compositional covariates. To accommodate the structural constraints induced by subcompositional predictors, we propose two two-step procedures, Oracle-Trans-sub-Coda-Lasso and Trans-sub-Coda-Lasso, which combine information borrowing across domains with constrained sparse estimation. The oracle version assumes that the transferable source domains are known, while the data-driven version further incorporates a source-detection step for the practically relevant case where such information is unavailable.

Our theoretical analysis establishes statistical guarantees for the proposed estimators under high-dimensional scaling and shows that transfer learning can lead to substantial

gains when the informative source domains are sufficiently aligned with the target domain. The simulation studies support these findings and demonstrate the empirical advantages of the proposed methods over target-only procedures and competing alternatives across a range of settings involving both subcompositions and non-compositional covariates.

There are several promising directions for future research. First, the current theory still relies on the condition $C_{\Sigma}^{A} < \infty$, which mainly arises from Step 1 of our procedure, where data from multiple sources are pooled for estimation. This requirement may become restrictive when covariate distributions differ substantially across sources. Simulation results in the supplementary material demonstrate that simple attempts to relax this condition, such as the constrained Trans-Fusion method, perform poorly under compositional data settings. This highlights the necessity of developing more robust cross-source covariance theories tailored specifically to the inherent characteristics of compositional data.

Second, ℓ_0-based methods such as the SDAR algorithm and best subset selection approaches are important alternatives for high-dimensional sparse regression. To further assess the relevance of ℓ_0-type alternatives, we additionally examined two linearly constrained ℓ_0-type competitors in the Supplementary Material. Additional simulations suggest that constrained ℓ_0-type methods can be advantageous in strong-signal and sparse settings, especially for coefficient estimation. However, they may also suffer from high computational cost, tuning sensitivity, or instability under weak-signal and less sparse regimes. By comparison, our method exhibits more stable performance across different regimes and yields fewer false negatives in variable selection, which is important for compositional data analysis. Developing efficient and tuning-stable linearly constrained ℓ_0-type methods for compositional transfer learning is an interesting direction for future work.

Another promising direction is to develop more general approaches for characterizing target–source relationships beyond pairwise comparisons. Such extensions may be useful in settings where transfer learning depends on the joint contribution of multiple source domains, for example when individual sources are all far from the target but their collective "center" is nevertheless close to it.

Supplementary Materials

The Supplementary Material includes the following sections: S1, the details of Algorithm 1S2, proof for the theorem 1; S3, proof for the theorem 2; S4, additional comparison with other constrained methods; S5, additional results of our numerical studies and real studies.

Acknowledgements

The authors wish to thank the Co-Editor, the Associate Editor and three reviewers for their many helpful and insightful comments and suggestions that greatly improved the paper. This work was supported in part by Shandong Provincial Natural Science Foundation (ZR2022MA065).

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44: 139-160.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell: The Blackburn Press.

Aitchison, J. and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika*, pages 323–330.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984.

Bertsekas, D. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific.

Bolnick, D. I., Snowberg, L. K., Hirsch, P. E., Lauber, C. L., Org, E., Parks, B., Lusis, A. J., Knight, R., Caporaso, J. G., and Svanbäck, R. (2014). Individual diet has sex-dependent effects on vertebrate gut microbiota. *Nature communications*, 5(1):4500.

Bradley, E. and Haran, J. (2024). The human gut microbiome and aging. *Gut Microbes*, 16(1):2359677. PMID: 38831607.

Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: Methods, theory and applications. *Springer Science & Business Media*.

Cheng, L., Wang, K., and Tsung, F. (2020). A hybrid transfer learning framework for in-plane freeform shape accuracy control in additive manufacturing. *IISE Transactions*, 53(3):298–312.

Combettes, P. L. and Müller, C. L. (2021). Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, 13(2):217–242.

He, Y., Li, Q., Hu, Q., and Liu, L. (2022). Transfer learning in high-dimensional semiparametric graphical models with application to brain connectivity analysis. *Statistics in Medicine*, 41(21):4112–4129.

He, Y., Li, Z., Liu, D., Qin, K., and Xie, J. (2024a). Representational transfer learning for matrix completion.

He, Y., Liu, D., Sun, Y., and Wang, Y. (2025). Transpca for large-dimensional factor analysis with weak factors: Power enhancement via knowledge transfer.

He, Z., Sun, Y., Liu, J., and Li, R. (2024b). Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In Dasgupta, S., Mandt, S., and Li, Y., editors, *International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*. Valencia, Spain, May 02–04, 2024.

Huang, J., Wang, M., and Wu, Y. (2023). Estimation and inference for transfer learning with high-dimensional

quantile regression.

Jin, J., Yan, J., Aseltine, R. H., and Chen, K. (2024). Transfer learning with large-scale quantile regression. *Technometrics*, 66(3):381–393.

Li, S., Cai, T. T., and Li, H. (2022). Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173.

Li, Z., Liu, D., He, Y., and Zhang, X. (2024a). Simultaneous estimation and dataset selection for transfer learning in high dimensions by a non-convex penalty.

Li, Z., Qin, K., He, Y., Zhou, W., and Zhang, X. (2024b). Knowledge transfer across multiple principal component analysis studies.

Li, S., Zhang, L., Cai, T. T., and Li, H. (2024). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119(546):1274–1285.

Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.

Ma, H., Zheng, Q., Zhang, Z., and Lai, H.and Peng, L. (2023). Globally adaptive longitudinal quantile regression with high dimensional compositional covariates. *Statistica Sinica*, 33(Spec Issue):1295–1318.

Mishra, A. and Müller, C. L. (2022). Robust regression with compositional covariates. *Computational Statistics & Data Analysis*, 165:107315.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. *PLOS Computational Biology*, 12(7):1–26.

Qiao, S., He, Y., and Zhou, W. (2024). Transfer learning for high-dimensional quantile regression with statistical

guarantee. *Transactions on Machine Learning Research.*

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10:1019–1040.

Shi, P., Zhou, Y., and Zhang, A. R. (2022). High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. *Biometrika*, 109(2):405–420.

Sun, Z., Xu, W. L., Cong, X. M., Li, G., and Chen, K. (2020). Log-contrast regression with functional compositional predictors: Linking preterm infant's gut microbiome trajectories to neurobehavioral outcome. *Annals of Applied Statistics*, 14(3):1535–1556.

Tan, W., Xue, L., Yang, S., and Zhan, X. (2024). High-dimensional log contrast models with measurement errors. *arXiv preprint arXiv:2407.15084*.

Tian, Y. and Feng, Y. (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697.

Torrey, L. and Shavlik, J. (2010). Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264, Hershey, PA: IGI Global.

Wang, T. and Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *The Annals of Applied Statistics*, 11(2):771 – 791.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3:1–40.

Yuan, P., Jin, C., and Li, G. (2024). Fdr control for linear log-contrast models with high-dimensional compositional covariates. *Computational Statistics & Data Analysis*, 197:107973.

Zhang, S., Wang, H., and Lin, W. (2025). Care: Large precision matrix estimation for compositional data. *Journal of the American Statistical Association*, 120(549): 305–317.

Qinqin Hu

School of Mathematics and Statistics, Shandong University at Weihai, China.

E-mail: qqhu@sdu.edu.cn

Xiaojing Luo

School of Mathematics and Statistics, Shandong University at Weihai, China.

E-mail: lxiaojing2578@163.com

Chencheng Ma

School of Mathematics, Shandong University, China.

E-mail: 202311964@mail.sdu.edu.cn

Wang Zhou

Department of Statistics and Data Science, National University of Singapore, Singapore.

E-mail: wangzhou@nus.edu.sg