



Article

Multi-Task Learning for Compositional Data via Sparse Network Lasso

Akira Okazaki ^{1,*}  and Shuichi Kawano ² 

¹ Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu 182-8585, Tokyo, Japan

² Faculty of Mathematics, Kyushu University, 744 Motoooka, Nishi-ku 819-0395, Fukuoka, Japan

* Correspondence: okazaki.akira@ai.lab.uec.ac.jp

Abstract: Multi-task learning is a statistical methodology that aims to improve the generalization performances of estimation and prediction tasks by sharing common information among multiple tasks. On the other hand, compositional data consist of proportions as components summing to one. Because components of compositional data depend on each other, existing methods for multi-task learning cannot be directly applied to them. In the framework of multi-task learning, a network lasso regularization enables us to consider each sample as a single task and construct different models for each one. In this paper, we propose a multi-task learning method for compositional data using a sparse network lasso. We focus on a symmetric form of the log-contrast model, which is a regression model with compositional covariates. Our proposed method enables us to extract latent clusters and relevant variables for compositional data by considering relationships among samples. The effectiveness of the proposed method is evaluated through simulation studies and application to gut microbiome data. Both results show that the prediction accuracy of our proposed method is better than existing methods when information about relationships among samples is appropriately obtained.

Keywords: clustering; log-contrast model; multi-task learning; symmetric form; variable selection



Citation: Okazaki, A.; Kawano, S. Multi-Task Learning for Compositional Data via Sparse Network Lasso. *Entropy* **2022**, *24*, 1839. <https://doi.org/10.3390/e24121839>

Academic Editor: Joaquín Abellán

Received: 31 October 2022

Accepted: 15 December 2022

Published: 17 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Multi-task learning is a statistical methodology that assumes a different model for each task and jointly estimates these models. By sharing the common information between them, the generalization performance of estimation and prediction tasks is improved [1]. Multi-task learning has been used in various fields of research, such as computer vision [2], natural language processing [3], and life sciences [4]. In life sciences, the risk factors may vary from patient to patient [5], and a model that is common to all patients cannot sufficiently extract general risk factors. In multi-task learning, each patient can be considered as a single task, and different models are built for each patient to extract both patient-specific and common factors for the disease [6]. Localized lasso [7] is a method that performs multi-task learning using network lasso regularization [8]. By treating each sample as a single task, localized lasso simultaneously performs multi-task learning and clustering in the framework of a regression model.

On the other hand, compositional data, which consist of the individual proportions of a composition, are used in the fields of geology and life sciences for microbiome analysis. Compositional data are constrained to always take positive values summing to one. Due to these constraints, it is difficult to apply existing multi-task learning methods to compositional data. In the field of microbiome analysis, studies on gut microbiomes [9,10] have suggested that there are multiple types of gut microbiome clusters that vary from individual to individual [11]. In the case of such data where multiple clusters may exist, it is difficult to extract sufficient information using existing regression models for compositional data.

In this paper, we propose a multi-task learning method for compositional data, focusing on the network lasso regularization and the symmetric form of the log-contrast model [12], which is a linear regression model with compositional covariates. The symmetric form is extended to the locally symmetric form in which each sample has a different regression coefficient vector. These regression coefficient vectors are clustered by the network lasso regularization. Furthermore, because the dimensionality of features in compositional data has been increasing, in particular in microbiome analysis [13], we use an ℓ_1 -regularization [14] to perform variable selection. The advantage of using ℓ_1 -regularization is being able to perform variable selection even if the number of parameters exceeds the sample size. In addition, ℓ_1 -regularization is formulated by convex optimization, which leads to feasible computation, while classical subset selection is not. The estimation of the parameters included in the model is performed using an estimation algorithm based on the alternating direction method of multipliers [15], because the model includes non-differentiable points in the ℓ_1 -regularization term and zero-sum constraints on the parameters. The constructed model includes regularization parameters, which are determined by cross-validation (CV).

The remainder of this paper is organized as follows. Section 2 introduces multi-task learning based on a network lasso. In Section 3, we describe the regression models for compositional data. We propose a multi-task learning method for compositional data and its estimation algorithm in Section 4. In Section 5, we discuss the effectiveness of the proposed method through Monte Carlo simulations. An application to gut microbiome data is presented in Section 6. Finally, Section 7 summarizes this paper and discusses future work.

2. Multi-Task Learning Based on a Network Lasso

Suppose that we have n observed p -dimensional data $\{x_i; i = 1, \dots, n\}$ and n observed data for the response variable $\{y_i; i = 1, \dots, n\}$ and that these pairs $\{(y_i, x_i), i = 1, \dots, n\}$ are given independently. The graph $R = R^T \in \mathbb{R}^{n \times n}$ is also given, where $(R)_{ij} = r_{ij} \geq 0$ represents the relationship between the sample pair (y_i, x_i) and (y_j, x_j) , and thus the diagonal components are zero.

We consider the following linear regression model:

$$y_i = \mathbf{x}_i^T \mathbf{w}_i + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\mathbf{w}_i = (w_{i1}, \dots, w_{ip})^T \in \mathbb{R}^p$ is the p -dimensional regression coefficient vector for sample x_i , and ϵ_i is an error term distributed as $N(0, \sigma^2)$ independently. Note that we exclude the intercept w_0 from the model, because we assume the centered response and the standardized explanatory variables. Model (1) comprises a different model for each sample. In classical regression models, the regression coefficient vectors are assumed to be identical (i.e., $\mathbf{w}_1 = \mathbf{w}_2 = \dots = \mathbf{w}_n$).

For Model (1), we consider the following minimization problem:

$$\min_{\mathbf{w}_i \in \mathbb{R}^p, i=1, \dots, n} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{w}_i)^2 + \lambda \sum_{m>l}^n r_{m,l} \|\mathbf{w}_m - \mathbf{w}_l\|_2 \right\}, \quad (2)$$

where $\lambda (> 0)$ is a regularization parameter. The second term in (2) is the network lasso regularization term [8]. For coefficient vectors \mathbf{w}_m and \mathbf{w}_l , the network lasso regularization term induces $\mathbf{w}_m = \mathbf{w}_l$. If these vectors are estimated to be the same, then the m -th and l -th samples are interpreted as belonging to the same cluster. In the framework of multi-task learning, the minimization problem (2) considers one sample as one task by setting a coefficient vector for each sample. This allows us to extract the information of the regression coefficient vectors separately for each task. In addition, by clustering the regression coefficient vectors using the network lasso regularization term, we can extract the common information among tasks.

Yamada et al. [7] proposed the localized lasso for minimization problem (2) by adding an $\ell_{1,2}$ -norm regularization term [16] as follows:

$$\min_{\mathbf{w}_i \in \mathbb{R}^p, i=1, \dots, n} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^T \mathbf{w}_i)^2 + \lambda_1 \sum_{m>l} r_{m,l} \|\mathbf{w}_m - \mathbf{w}_l\|_2 + \lambda_2 \sum_{i=1}^n \|\mathbf{w}_i\|_1^2 \right\}. \quad (3)$$

The $\ell_{1,2}$ -norm regularization term induces group structure and intra-group level sparsity: several regression coefficients in a group are estimated to be zero, but at least one is estimated to be non-zero by squaring over the ℓ_1 -norm. In the localized lasso, each regression coefficient vector \mathbf{w}_i is treated as a group in order to remain $\mathbf{w}_i \neq \mathbf{0}$. The localized lasso is used for multi-task learning and variable selection.

3. Regression Modeling for Compositional Data

The p -dimensional compositional data $\mathbf{x} = (x_1, \dots, x_p)^T$ are defined as proportional data in the simplex space:

$$\mathbb{S}^{p-1} = \left\{ (x_1, \dots, x_p) : x_j > 0 \quad (j = 1, \dots, p), \sum_{j=1}^p x_j = 1 \right\}. \quad (4)$$

This structure imposes dependence between the features of the compositional data. Thus, statistical methods defined for spaces of real numbers cannot be applied [17]. To overcome this problem, Aitchison and Bacon-Shone [12] proposed the log-contrast model, which is a linear regression model with compositional covariates.

Suppose that we have n observed p -dimensional compositional data $\{\mathbf{x}_i; i = 1, \dots, n\}$ and n objective variable data $\{y_i; i = 1, \dots, n\}$ and these pairs $\{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ are given independently. The log-contrast model is represented as follows:

$$y_i = \sum_{j=1}^{p-1} \log \frac{x_{ij}}{x_{ip}} \beta_j + \epsilon_i, \quad i = 1, \dots, n, \quad (5)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p-1})^T \in \mathbb{R}^{p-1}$ is a regression coefficient vector. Because the model uses an arbitrary variable as a reference for all other variables, the solution changes depending on the selection of the reference. By introducing $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, the log-contrast model is equivalently expressed in symmetric form as:

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \text{s.t.} \quad \sum_{j=1}^p \beta_j = 0, \quad i = 1, \dots, n, \quad (6)$$

where $\mathbf{z}_i = (\log x_{i1}, \dots, \log x_{ip})^T$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ is a regression coefficient vector. Lin et al. [13] proposed the minimization problem to select relevant variables in symmetric form by adding an ℓ_1 -regularization term [14]:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \quad \text{s.t.} \quad \sum_{j=1}^p \beta_j = 0. \quad (7)$$

Other models that extend this symmetric form of the problem have also been proposed [18–21].

4. Proposed Method

In this section, we propose a multi-task learning method for compositional data based on the network lasso and the symmetric form of the log-contrast model.

4.1. Model

We consider the locally symmetric form of the log-contrast model:

$$y_i = z_i^T w_i + \epsilon_i, \quad \text{s.t.} \quad \sum_{j=1}^p w_{ij} = 0, \quad i = 1, \dots, n, \tag{8}$$

where $z_i = (\log x_{i1}, \dots, \log x_{ip})^T$, and $w_i = (w_{i1}, \dots, w_{ip})^T$ is the regression coefficient vector for i -th sample of compositional data x_i . For Model (8), we consider the following minimization problem:

$$\begin{aligned} \min_{w_i \in \mathbb{R}^p, i=1, \dots, n} & \left\{ \sum_{i=1}^n (y_i - z_i^T w_i)^2 + \lambda_1 \sum_{m>l}^n r_{m,l} \|w_m - w_l\|_2 + \lambda_2 \sum_{i=1}^n \|w_i\|_1 \right\}, \\ \text{s.t.} & \quad \sum_{j=1}^p w_{ij} = 0, \quad i = 1, \dots, n, \end{aligned} \tag{9}$$

where $\lambda_1, \lambda_2 (> 0)$ are regularization parameters. The second term is the network lasso regularization term, which performs the clustering of the regression coefficient vectors. The third term is the ℓ_1 -regularization term [14]. This term is interpreted as a special case of the $\ell_{1,2}$ -regularization term used in Model (3). Unlike the ℓ_1 -regularization term, it is difficult to optimize the $\ell_{1,2}$ -regularization directly, because it does not have a closed form of the updates. To construct the estimation algorithm that performs variable selection and preserves the constraints for regression coefficient vectors simultaneously, we employ the ℓ_1 -regularization term. Since variable selection is performed by the ℓ_1 -regularization term, we refer to the combination of the second term and the third term as sparse network lasso after sparse group lasso [22].

For Model (8), when a new data point z_i^* is obtained after the estimation, there is no corresponding regression coefficient vector w_{i^*} for z_i^* . Thus, it is necessary to estimate the coefficient vector for predicting the response. Hallac et al. [8] proposed solving the following minimization problem:

$$\min_{w_{i^*} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n r_{i^*,i} \|w_{i^*} - \hat{w}_i\|_2 \right\}, \tag{10}$$

where \hat{w}_i is the estimated regression coefficient vector for the i -th sample. This problem is also known as the Weber problem. The solution of this problem is interpreted as the weighted geometric median of \hat{w}_i . For our proposed method, we consider solving the constrained Weber problem with the zero-sum constraint in the form:

$$\min_{w_{i^*} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n r_{i^*,i} \|w_{i^*} - \hat{w}_i\|_2 \right\}, \quad \text{s.t.} \quad \sum_{j=1}^p w_{i^*j} = 0. \tag{11}$$

4.2. Estimation Algorithm

To construct the estimation algorithm for the proposed method, we rewrite minimization problem (9) as follows:

$$\begin{aligned} \min_{w_i, a_m, b_i \in \mathbb{R}^p, i=1, \dots, n} & \left\{ \sum_{i=1}^n (y_i - z_i^T w_i)^2 + \lambda_1 \sum_{m>l}^n r_{m,l} \|a_{m,l} - a_{l,m}\|_2 + \lambda_2 \sum_{i=1}^n \|b_i\|_1 \right\}, \\ \text{s.t.} & \quad w_m = a_{m,l}, \quad w_i = b_i, \quad \mathbf{1}_p^T w_i = 0, \quad i, m, l = 1, \dots, n, \end{aligned} \tag{12}$$

where $\mathbf{1}_p$ is the p -dimensional vector of ones. The augmented Lagrangian for (12) is formulated as:

$$\begin{aligned}
 L(\Theta, Q)_\Omega = & \sum_{i=1}^n (y_i - \mathbf{z}_i^T \mathbf{w}_i)^2 \\
 & + \sum_{m>l}^n \{ \lambda_1 r_{m,l} \| \mathbf{a}_{m,l} - \mathbf{a}_{l,m} \|_2 \\
 & + \frac{\rho}{2} (\| \mathbf{w}_m - \mathbf{a}_{m,l} + \mathbf{s}_{m,l} \|_2^2 + \| \mathbf{w}_l - \mathbf{a}_{l,m} + \mathbf{s}_{l,m} \|_2^2) - \frac{\rho}{2} (\| \mathbf{s}_{m,l} \|_2^2 + \| \mathbf{s}_{l,m} \|_2^2) \} \\
 & + \sum_{i=1}^n \left\{ \lambda_2 \| \mathbf{b}_i \|_1 + \mathbf{t}_i^T (\mathbf{w}_i - \mathbf{b}_i) + \frac{\phi}{2} \| \mathbf{w}_i - \mathbf{b}_i \|_2^2 \right\} \\
 & + \sum_{i=1}^n \left\{ u_i \mathbf{1}_p^T \mathbf{w}_i + \frac{\psi}{2} \| \mathbf{1}_p^T \mathbf{w}_i \|_2^2 \right\},
 \end{aligned} \tag{13}$$

where $\mathbf{s}_{m,l}, \mathbf{t}_i, u_i$ are the Lagrange multipliers and $\rho, \phi, \psi (> 0)$ are the tuning parameters. For simplicity of notation, the parameters in the model $\mathbf{w}_i, \mathbf{a}_{i,j}, \mathbf{b}_i$ are collectively denoted as Θ , the Lagrange multipliers are collectively denoted as Q , and the tuning parameters are collectively denoted as Ω .

The update algorithm for Θ with the alternating direction method of multipliers (ADMM) is obtained from the following minimization problem:

$$\begin{aligned}
 \mathbf{w}^{(k+1)} &= \arg \min_{\mathbf{w}} L(\mathbf{w}, \mathbf{a}^{(k)}, \mathbf{b}^{(k)}, Q^{(k)})_\Omega, \\
 \mathbf{a}^{(k+1)} &= \arg \min_{\mathbf{a}} L(\mathbf{w}^{(k+1)}, \mathbf{a}, \mathbf{b}^{(k)}, Q^{(k)})_\Omega, \\
 \mathbf{b}^{(k+1)} &= \arg \min_{\mathbf{b}} L(\mathbf{w}^{(k+1)}, \mathbf{a}^{(k+1)}, \mathbf{b}, Q^{(k)})_\Omega,
 \end{aligned} \tag{14}$$

where k denotes the repetition number. By repeating the updates (14) and the update for Q , the estimation algorithm for (12) is represented by Algorithm 1. The estimation algorithm for (11) is represented by Algorithm 2 with the update from ADMM. The details of the derivations of the estimation algorithms are presented in Appendices A and B.

Algorithm 1 Estimation algorithm for (12) via ADMM

Require: Initialize $\mathbf{w}^{(0)}, \mathbf{a}^{(0)}, \mathbf{b}^{(0)}, \mathbf{s}^{(0)}, \mathbf{t}^{(0)}, u^{(0)}$.

while convergence **do**
for $i = 1 \dots, n$ **do**

$$\begin{aligned}
 \mathbf{w}_i^{(k+1)} = & \left\{ 2\mathbf{z}_i \mathbf{z}_i^T + (\rho(n-1) + \phi) I_p + \psi \mathbf{1}_p \mathbf{1}_p^T \right\}^{-1} \\
 & \left\{ 2y_i \mathbf{z}_i + \rho \sum_{m>l}^n (\mathbf{a}_{m,l}^{(k)} - \mathbf{s}_{l,m}^{(k)}) - \mathbf{t}_i^{(k)} + \phi \mathbf{b}_i^{(k)} - u_i \mathbf{1}_p \right\}
 \end{aligned}$$

end for

for $m, l = 1, \dots, n, (m > l)$ **do**

$$\begin{aligned}
 \theta &= \max \left(1 - \frac{\lambda_1 r_{m,l}}{\rho \| (\mathbf{w}_m^{(k+1)} + \mathbf{s}_{m,l}^{(k)}) - (\mathbf{w}_l^{(k+1)} + \mathbf{s}_{l,m}^{(k)}) \|_2}, 0.5 \right) \\
 \mathbf{a}_{m,l}^{(k+1)} &= \theta (\mathbf{w}_m^{(k+1)} + \mathbf{s}_{m,l}^{(k)}) + (1 - \theta) (\mathbf{w}_l^{(k+1)} + \mathbf{s}_{l,m}^{(k)}) \\
 \mathbf{a}_{l,m}^{(k+1)} &= (1 - \theta) (\mathbf{w}_m^{(k+1)} + \mathbf{s}_{m,l}^{(k)}) + \theta (\mathbf{w}_l^{(k+1)} + \mathbf{s}_{l,m}^{(k)})
 \end{aligned}$$

Algorithm 1 Cont.

```

end for
for  $i = 1, \dots, n, j = 1, \dots, p$  do
     $b_{ij}^{(k+1)} = S(w_{ij}^{(k+1)} + \frac{1}{\phi} t_{ij}^{(k)}, \frac{\lambda_2}{\phi})$ 
end for
for  $m, l = 1, \dots, n, (m \neq l)$  do
     $s_{m,l}^{(k+1)} = s_{m,l}^{(k)} + \rho(w_m^{(k+1)} - a_{m,l}^{(k+1)})$ 
end for
for  $i = 1 \dots, n$  do
     $t_i^{(k+1)} = t_i^{(k)} + \phi(w_i^{(k+1)} - b_i^{(k+1)})$ 
     $u_i^{(k+1)} = u_i^{(k)} + \psi \mathbf{1}_p^T w_i^{(k+1)}$ 
end for
end while
Ensure:  $b_i, i = 1, \dots, n.$ 

```

Algorithm 2 Estimation algorithm for constrained Weber problem (11) via ADMM

```

Require: Initialize  $w_{i^*}^{(0)}, e^{(0)}, u^{(0)}, v^{(0)}$ .
while convergence do
    for  $i = 1 \dots, n$  do
         $e_i^{(k+1)} = \min\left(\frac{r_{i^*i}}{\mu}, \|w_{i^*}^{(k)} - \frac{1}{\mu} u_i^{(k)} - \widehat{w}_i\|_2\right) \frac{w_{i^*}^{(k)} - \frac{1}{\mu} u_i^{(k)} - \widehat{w}_i}{\|w_{i^*}^{(k)} - \frac{1}{\mu} u_i^{(k)} - \widehat{w}_i\|_2}$ 
    end for
     $w_{i^*}^{(k+1)} = (\mu n I_p + \eta \mathbf{1}_p \mathbf{1}_p^T)^{-1} \left\{ \mu \sum_{i=1}^n (e_i^{(k+1)} + \frac{1}{\mu} u_i^{(k)}) - v^{(k)} \mathbf{1}_p \right\}$ 
    for  $i = 1 \dots, n$  do
         $u_i^{(k+1)} = u_i^{(k)} + \mu (e_i^{(k+1)} - w_{i^*}^{(k+1)})$ 
    end for
     $v^{(k+1)} = v^{(k)} + \eta \mathbf{1}_p^T w_{i^*}^{(k+1)}$ 
end while
Ensure:  $w_{i^*}$ 

```

5. Simulation Studies

In this section, we report simulation studies conducted with our proposed method using artificial data.

In our simulations, we generated artificial data from the true model:

$$y_i = \begin{cases} z_i^T w_{(1)}^* + \epsilon_i, & (i = 1, \dots, 40), \\ z_i^T w_{(2)}^* + \epsilon_i, & (i = 41, \dots, 80), \\ z_i^T w_{(3)}^* + \epsilon_i, & (i = 81, \dots, 120), \end{cases} \tag{15}$$

where $z_i = (\log x_{i1}, \dots, \log x_{ip})^T$, $x_i = (x_{i1}, \dots, x_{ip})^T$ is p -dimensional compositional data, $w_{(1)}^*, w_{(2)}^*, w_{(3)}^* \in \mathbb{R}^p$ are the true regression coefficient vectors, and ϵ_i is an error term distributed as $N(0, \sigma^2)$ independently. We generated compositional data $\{x_i; i = 1, \dots, 120\}$ as follows. First, we generated the data $\{c_i, i = 1, \dots, 120\}$ from the p -dimensional multivariate normal distribution $N_p(\omega, \Sigma)$ independently, where $(\omega)_j = \omega_j$, $(\Sigma)_{ij} = 0.2^{|i-j|}$, and

$$\omega_j = \begin{cases} \log(0.5p), & (j = 1, \dots, 5), \\ 0, & (j = 6, \dots, p). \end{cases} \tag{16}$$

Then, the data $\{c_i, i = 1, \dots, 120\}$ were converted to the compositional data $\{x_i; i = 1, \dots, 120\}$ by the following transformation:

$$x_{ij} = \frac{\exp(c_{ij})}{\sum_{k=1}^p \exp(c_{ik})}, \quad i = 1, \dots, 120, j = 1, \dots, p. \tag{17}$$

The true regression coefficient vectors were set as:

$$\begin{aligned} \mathbf{w}_{(1)}^* &= (1, -0.8, 0.6, 0, 0, -1.5, -0.5, 1.2, \mathbf{0}_{p-8}^T)^T, \\ \mathbf{w}_{(2)}^* &= (0, -0.5, 1, 1.2, 0.1, -1, 0, -0.8, \mathbf{0}_{p-8}^T)^T, \\ \mathbf{w}_{(3)}^* &= (0, 0, 0, 0.8, 1, 0, -0.8, -1, \mathbf{0}_{p-8}^T)^T. \end{aligned} \tag{18}$$

We also assumed that the graph $R \in \{0, 1\}^{120 \times 120}$ was observed. In the graph, the true value of each element was obtained with probability P_R . We calculated MSE as $\sum_{i^*=1}^{n^*} (y_{i^*} - \mathbf{z}_{i^*}^T \hat{\mathbf{w}}_{i^*})^2 / n^*$, dividing the 120 samples into 100 training data and 20 validation data. Here, n^* indicates the number of samples for validation data (i.e., 20). The regression coefficient vectors for the validation data were estimated based on the constrained Weber problem (11). To evaluate the effectiveness of our proposed method, it is compared with both Model (7) and the model obtained by removing the zero-sum constraint from minimization problem (9). We refer to the latter two comparison methods as compositional lasso (CL) and sparse network lasso (SNL), respectively. To the best of our knowledge, there are no studies that simultaneously perform regression and clustering on compositional data. Therefore, we compared with the CL and SNL models; CL assumes the situation where the existence of the multiple clusters is not considered, while SNL considers their existence while the nature of the compositional data is ignored.

The regularization parameters were determined by five-fold CV for both the proposed method and the comparison methods. The values of tuning parameters $\rho, \phi, \psi, \mu, \eta$ for ADMM were all set to one. We considered several settings: $p = \{30, 100, 200\}$, $\sigma = \{0.1, 0.5, 1\}$, $P_R = \{0.99, 0.95, 0.90, 0.80, 0.70\}$. We generated 100 datasets and computed the mean and standard deviation of MSE from the 100 repetitions.

Tables 1–3 show the results for the mean and standard deviation of MSE for each σ . The proposed method and SNL show better prediction accuracy than CL in almost settings. The reason for this may be that CL assumes only a single regression coefficient vector and thus fails to capture the true structure, which consists of three clusters. A comparison between the proposed method and SNL shows that the proposed method has higher prediction accuracy than SNL when $P_R = 0.99, 0.95$, and 0.90 , whereas SNL shows better results in most cases when $P_R = 0.80, 0.70$. This means that the proposed method is superior to SNL when the structure of the graph R is similar to the true structure. On the whole, the prediction accuracy deteriorates as P_R decreases for both the proposed method and SNL, but this deterioration is more extreme for the proposed method. For both the proposed method and SNL, which assume multiple regression coefficient vectors, the standard deviation is somewhat large.

Table 1. Mean and deviation of MSE in $\sigma = 0.1$ for simulations.

Method	$p = 30$	$p = 100$	$p = 200$
CL	5.20(1.44)	6.99(1.81)	8.75(2.69)
	$P_R = 0.99$		
Proposed	0.51 (0.58)	2.13 (2.35)	2.70 (1.92)
SNL	2.54(0.97)	3.73(1.31)	6.55(3.94)
	$P_R = 0.95$		
Proposed	2.74 (1.19)	2.96 (1.33)	3.68 (2.71)
SNL	3.19(0.98)	3.87(1.35)	5.26(4.15)

Table 1. Cont.

Method	$p = 30$	$p = 100$	$p = 200$
		$P_R = 0.90$	
Proposed	3.29 (1.38)	3.80 (1.33)	4.49 (1.81)
SNL	3.40(1.23)	4.25(1.49)	4.75(1.49)
		$P_R = 0.80$	
Proposed	4.20(1.58)	5.53(2.30)	7.49(3.98)
SNL	3.87 (1.41)	4.86 (1.52)	5.70 (2.00)
		$P_R = 0.70$	
Proposed	4.13 (1.56)	6.57(2.55)	7.66(2.28)
SNL	4.56(1.55)	5.63 (1.72)	6.69 (2.25)

Table 2. Mean and deviation of MSE in $\sigma = 0.5$ for simulations.

Method	$p = 30$	$p = 100$	$p = 200$
CL	5.64(1.42)	7.94(2.31)	10.02(2.96)
		$P_R = 0.99$	
Proposed	1.02 (0.75)	2.13 (1.50)	3.07 (1.61)
SNL	2.97(1.23)	3.73(1.32)	5.98(3.95)
		$P_R = 0.95$	
Proposed	3.05 (1.28)	3.36 (1.05)	4.37 (3.47)
SNL	3.50(1.20)	4.19(1.35)	5.18(2.54)
		$P_R = 0.90$	
Proposed	3.55 (1.40)	4.83(1.61)	5.13 (2.99)
SNL	3.83(1.30)	4.43 (1.26)	5.21(3.42)
		$P_R = 0.80$	
Proposed	4.10(1.47)	5.21 (1.89)	6.70(2.51)
SNL	4.06 (1.35)	5.32(1.88)	6.06 (1.94)
		$P_R = 0.70$	
Proposed	4.33 (1.39)	6.59(2.62)	8.58(3.16)
SNL	4.53(1.54)	5.71 (1.78)	7.37 (2.24)

Table 3. Mean and deviation of MSE in $\sigma = 1$ for simulations.

Method	$p = 30$	$p = 100$	$p = 200$
CL	6.50(1.78)	8.05(2.57)	10.41(3.10)
		$P_R = 0.99$	
Proposed	2.34 (1.29)	3.25 (1.87)	3.87 (1.97)
SNL	3.94(1.47)	4.98(1.67)	5.90(3.12)
		$P_R = 0.95$	
Proposed	3.43 (1.24)	3.96 (1.61)	4.73 (2.35)
SNL	4.36(1.31)	4.73(1.29)	5.39(1.64)
		$P_R = 0.90$	
Proposed	4.28 (1.55)	4.83 (1.61)	5.09 (1.88)
SNL	4.71(1.49)	5.25(2.08)	5.75(2.29)
		$P_R = 0.80$	
Proposed	5.67(1.80)	6.61(2.05)	7.77(3.68)
SNL	5.20 (1.79)	6.06 (2.01)	6.77 (2.07)
		$P_R = 0.70$	
Proposed	5.73(1.87)	8.21(2.71)	8.86(3.57)
SNL	5.31 (1.84)	7.02 (2.29)	7.97 (2.56)

6. Application to Gut Microbiome Data

The gut microbiome affects human health/disease, for example, in terms of obesity. Gut microbiomes may be exposed to inter-individual heterogeneity from environmental factors such as diet as well as from hormonal factors and age [23,24]. In this section, we applied our proposed method to the real dataset reported by [9]. This dataset was obtained from a cross-sectional study of 98 healthy volunteers conducted at the University of Pennsylvania to investigate the connections between long-term dietary patterns and gut microbiome composition. Stool samples were collected from the subjects, and DNA samples were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments of the V1–V2 region. In the results, the counts for more than 17,000 species-level OTUs were obtained. Demographic data, such as body mass index (BMI), sex, and age, were also obtained.

We used centered BMI as the response and the compositional data of the gut microbiome as the explanatory variable. In order to reduce their number, we used single-linkage clustering based on an available phylogenetic tree to aggregate the OTUs, which is provided as the function `tip_glom` in the R package “phyloseq” see [25]. In this process, some closely related OTUs defined on the leaf nodes of the phylogenetic tree are aggregated into one OTU when the cophenetic distances between the OTUs are smaller than a certain threshold. We set the threshold at 0.5. As a result, 166 OTUs were obtained after the aggregation. Since some OTUs had zero counts, making it impossible to take the logarithm, these counts were replaced by the value one before converting them to compositional data.

We computed the graph $R \in \mathbb{R}^{98 \times 98}$ as follows:

$$R = \frac{S^T + S}{2}, \quad (S)_{ij} = \begin{cases} 1 & j\text{-th sample is a 5-NN of } i\text{-th sample with } D_{ij}, \\ 0 & \text{Otherwise,} \end{cases} \quad (19)$$

where D_{ij} is the distance between the i -th and j -th samples. Distance D_{ij} was calculated in the following two ways: (i) Gower distance [26] was calculated using the sex and age data of the subjects. (ii) Log-ratio distance (e.g., see [27]) was calculated using the explanatory variable, as follows:

$$D_{ij} = \sqrt{\sum_{l=1}^p (z_{il}^c - z_{jl}^c)^2}, \quad (20)$$

where $z_{ij}^c = \log x_{ij} - \frac{1}{p} \sum_{j=1}^p \log x_{ij}$. Using these two ways of calculating distance, we obtained two different R and estimation results. We refer to these two methods as Proposed (i) and Proposed (ii), respectively. Equation (19) is the same as the one used in Yamada et al. [7] in the application to real datasets.

To evaluate the prediction accuracy of our proposed method, we calculated MSE by randomly splitting the dataset into 90 samples as the training data and eight samples as the validation data. We again used the method of Lin et al. (2014), which is referred to as compositional lasso (CL), as a comparison method. The regularization parameters were determined by five-fold CV for both our proposed method and CL. The mean and standard deviation of MSE were calculated from 100 repetitions.

Table 4 shows the mean and standard deviation of MSE in the real data analysis. We observe that Proposed (i) has the smallest mean and standard deviation of MSE. This indicates that our proposed method captures the inherent structure of the compositional data by providing an appropriate graph R . However, the standard deviation is large even for Proposed (i), which indicates that the prediction accuracy may strongly depend on the assignments of samples to the training data and the validation data.

Table 4. Mean and standard deviation of MSE for real data analysis (100 repetitions).

Value	Proposed (i)	Proposed (ii)	CL
MSE (SD)	23.01(16.62)	31.59(22.44)	30.96(23.36)

Table 5 shows the coefficient of determination R^2 using leave-one-out cross-validation (LOOCV) for Proposed (i) and CL. The fittings of the observed and predicted BMI are plotted in Figure 1a,b for Proposed (i) and CL, respectively. The horizontal axis represents the centered observed BMI values, and the vertical axis represents the corresponding predicted BMI. As shown, CL does not predict data with centered observed values between -10 and -5 as being in that interval, whereas Proposed (i) predicts these data correctly to some extent.

Table 5. Coefficients of determination using LOOCV.

Value	Proposed (i)	CL
LOOCV R^2	0.245	0.083

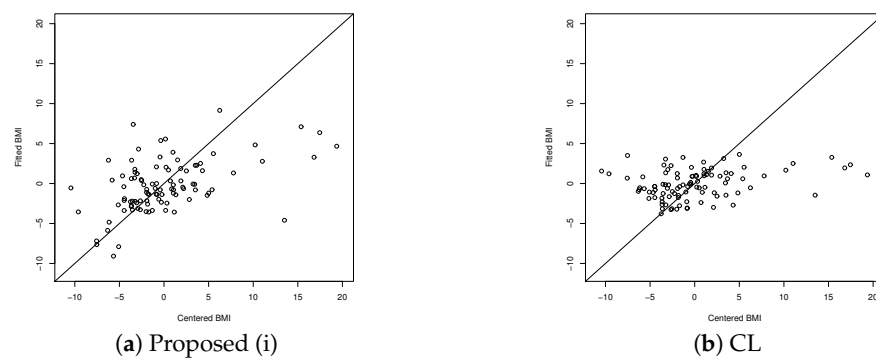


Figure 1. Observed and predicted BMI using LOOCV.

Figure 2 shows the regression coefficients estimated by Proposed (i) for all samples, where the regularization parameters are determined by LOOCV. To obtain the results in Figure 2, we used hierarchical clustering to group together similar regression coefficient vectors, setting the number of clusters as five.

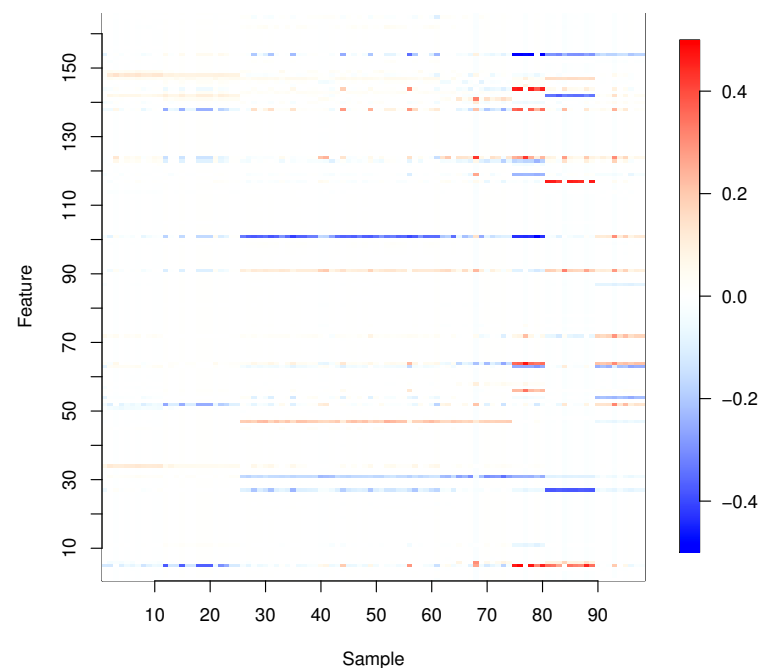


Figure 2. Estimated regression coefficients for all samples.

With our proposed method, many of the estimated regression coefficients were not exactly zero but close to zero. Thus, we will treat estimated regression coefficients $|\hat{w}_{ij}| < 0.05$

as being exactly zero to simplify the interpretation. Figure 3 shows only those coefficients that satisfy $|\hat{w}_{ij}| \geq 0.05$ in at least one sample, where the corresponding variables are listed in Table 6.

It is reported that the human gut microbiome can be classified into three clusters, called enterotypes, which are characterized by three predominant genera: *Bacteroides*, *Prevotella*, and *Ruminococcus* [11]. In the dataset, OTUs of genus levels *Prevotella* and *Ruminococcus* were aggregated into the OTUs of family levels *Prevotellaceae* and *Ruminococcaceae* by the single-linkage clustering. In Figure 3, *Bacteroides* correspond to OTU5, 6, 7, 8, 9, and 10; *Prevotellaceae* corresponds to OTU12; and *Ruminococcaceae* corresponds to OTU30 and 31. For these OTUs, the differences are clear between OTU6, 9, 10 and OTU30, 31 among samples 81–90, in which only *Bacteroides* are correlated to the response. On the other hand, the differences among samples 65–74 are also indicated, in which only *Bacteroides* do not affect BMI. These results suggest that *Bacteroides*, *Prevotellaceae*, and *Ruminococcaceae* may have different effects on BMI that are associated with enterotypes. In addition, it is reported that women with a higher abundance of *Prevotellaceae* are more obese [28]. The regression coefficients of non-zero *Prevotellaceae* are all positive, and the eight corresponding samples are all females. On the other hand, in OTU29 indicating *Roseburia*, 9 samples out of 10 are negatively associated with BMI. *Roseburia* is also reported to be negatively correlated with indicators of body weight [29].

Table 6. Variables with estimated regression coefficients $|\hat{w}_{ij}| \geq 0.05$ for at least one sample.

Variable	Kingdom	Phylum	Class	Order	Family	Genus	Species
OTU1	Bacteria						
OTU2	Bacteria						
OTU3	Bacteria	Bacteroidetes					
OTU4	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales			
OTU5	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU6	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU7	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU8	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU9	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU10	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU11	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Porphyromonadaceae	Parabacteroides	
OTU12	Bacteria	Bacteroidetes	Bacteroidetes	Bacteroidales	Prevotellaceae		
OTU13	Bacteria	Firmicutes	Clostridia				
OTU14	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU15	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU16	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU17	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU18	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU19	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU20	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU21	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU22	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU23	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU24	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU25	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU26	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU27	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU28	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae		
OTU29	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia	
OTU30	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae		
OTU31	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae		
OTU32	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Catenibacterium	
OTU33	Bacteria	Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	Erysipelotrichaceae. Incertae.Sedis	
OTU34	Bacteria	Proteobacteria					
OTU35	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae		

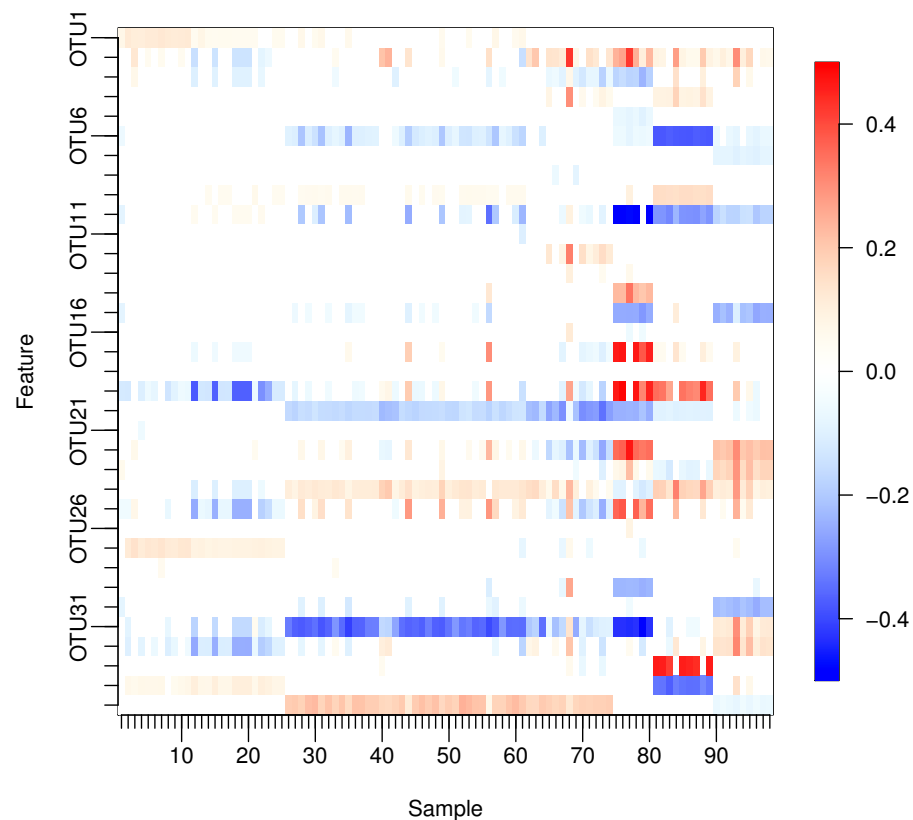


Figure 3. Only the estimated regression coefficients with $|\hat{w}_{ij}| \geq 0.05$ for at least one sample.

7. Conclusions

We proposed a multi-task learning method for compositional data based on a sparse network lasso and log-contrast model. By imposing a zero-sum constraint on the model corresponding to each sample, we could extract the information of latent clusters in the regression coefficient vectors for compositional data. In the results of simulations, the proposed method worked well when clusters existed for the compositional data and an appropriate graph R was obtained. In a human gut microbiome analysis, our proposed method is better than the existing method in prediction accuracy by considering the heterogeneity from age and sex. In addition, cluster-specific OTUs such as ones related to enterotypes were detected in terms of effects on BMI.

Although our proposed method shrinks some regression coefficients that do not affect response to zero, many coefficients close to zero remain. Furthermore, in both the simulations and human gut microbiome analysis, the prediction accuracy of the proposed method deteriorated significantly when the obtained R did not capture the true structure. Moreover, the standard deviations of MSE were high in almost all cases. We leave these as topics of future research.

Author Contributions: Conceptualization, A.O.; methodology, A.O.; formal analysis, A.O.; data curation, A.O.; writing—original draft preparation, A.O. and S.K.; supervision, S.K.; funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Numbers JP19K11854 and JP20H02227.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the gut microbiome data not being personally identifiable.

Informed Consent Statement: Patient consent was waived due to the gut microbiome data not being personally identifiable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, A.O., upon reasonable request. The source codes of the proposed method are available at <https://github.com/aokazaki255/CSNL>.

Acknowledgments: The authors would like to thank the Associate Editor and the reviewers for their helpful comments and constructive suggestions. The authors would like to also thank Tao Wang of Shanghai Jiao Tong University for providing the gut microbiome data used in Section 6. Supercomputing resources were provided by the Human Genome Center (the Univ. of Tokyo).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations is used in this manuscript:

ADMM alternating direction method of multipliers

Appendix A. Derivations of Update Formulas in ADMM

Appendix A.1. Update of w

In the update of w , we minimize the terms of the augmented Lagrangian (13) depending on w_i as follows:

$$w_i^{(k+1)} = \arg \min_{w_i \in \mathbb{R}^p} \left\{ (y_i - z_i^T w_i)^2 + \frac{\rho}{2} \sum_{m \neq i}^n \|w_i - a_{i,m}^{(k)} + s_{i,m}^{(k)}\|_2^2 + t_i^{(k)T} (w_i - b_i^{(k)}) + \frac{\phi}{2} \|w_i - b_i^{(k)}\|_2^2 + u_i^{(k)} \mathbf{1}_p^T w_i + \frac{\psi}{2} \|\mathbf{1}_p^T w_i\|_2^2 \right\}. \tag{A1}$$

From $\frac{\partial L}{\partial w_i} = 0$, we obtain the update:

$$w_i^{(k+1)} = \left\{ 2z_i z_i^T + (\rho(n-1) + \phi)I_p + \psi \mathbf{1}_p \mathbf{1}_p^T \right\}^{-1} \left\{ 2y_i z_i + \rho \sum_{m \neq i}^n (a_{i,m}^{(k)} - s_{i,m}^{(k)}) - t_i^{(k)} + \phi b_i^{(k)} - u_i^{(k)} \mathbf{1}_p \right\}. \tag{A2}$$

Appendix A.2. Update of a

The update of a is obtained by the joint minimization of $a_{m,l}$ and $a_{l,m}$ as follows:

$$a_{m,l}^{(k+1)}, a_{l,m}^{(k+1)} = \arg \min_{a_{m,l}, a_{l,m} \in \mathbb{R}^p} \left\{ \lambda_1 r_{m,l} \|a_{m,l} - a_{l,m}\|_2 + \frac{\rho}{2} (\|w_m^{(k+1)} - a_{m,l} + s_{m,l}^{(k)}\|_2^2 + \|w_l^{(k+1)} - a_{l,m} + s_{l,m}^{(k)}\|_2^2) \right\}. \tag{A3}$$

In [8], the analytical solution was given as:

$$a_{m,l}^{(k+1)} = \theta (w_m^{(k+1)} + s_{m,l}^{(k)}) + (1 - \theta) (w_l^{(k+1)} + s_{l,m}^{(k)}), \tag{A4}$$

$$a_{l,m}^{(k+1)} = (1 - \theta) (w_m^{(k+1)} + s_{m,l}^{(k)}) + \theta (w_l^{(k+1)} + s_{l,m}^{(k)}),$$

where

$$\theta = \max \left(1 - \frac{\lambda_1 r_{m,l}}{\rho \| (w_m^{(k+1)} + s_{m,l}^{(k)}) - (w_l^{(k+1)} + s_{l,m}^{(k)}) \|_2}, 0.5 \right). \tag{A5}$$

Appendix A.3. Update of \mathbf{b}

The update of \mathbf{b} is obtained by the following minimization problem:

$$\mathbf{b}_i^{(k+1)} = \arg \min_{\mathbf{b}_i \in \mathbb{R}^p} \sum_{i=1}^n \left\{ \lambda_2 \|\mathbf{b}_i\|_1 + \mathbf{t}_i^T (\mathbf{w}_i - \mathbf{b}_i) + \frac{\phi}{2} \|\mathbf{w}_i - \mathbf{b}_i\|_2^2 \right\}. \tag{A6}$$

Because the minimization problem with respect to \mathbf{b}_i contains a non-differentiable point in the ℓ_1 norm of \mathbf{b}_i , we consider the subderivative of $|b_{ij}|$. Then, we obtain the update:

$$b_{ij}^{(k+1)} = S \left(w_{ij} + \frac{t_{ij}}{\phi}, \frac{\lambda_2}{\phi} \right), \quad j = 1, \dots, p, \tag{A7}$$

where $S(\cdot, \cdot)$ is the soft-thresholding operator given by $S(x, \lambda) := \text{sign}(x)(|x| - \lambda)_+$.

Appendix A.4. Update of Q

The updates for the Lagrange multipliers denoted as Q are obtained by gradient descent as follows:

$$\begin{aligned} \mathbf{s}_{m,l}^{(k+1)} &= \mathbf{s}_{m,l}^{(k)} + \rho (\mathbf{w}_m^{(k+1)} - \mathbf{a}_{m,l}^{(k+1)}), \quad m, l = 1, \dots, n \ (i \neq j), \\ \mathbf{t}_i^{(k+1)} &= \mathbf{t}_i^{(k)} + \phi (\mathbf{w}_i^{(k+1)} - \mathbf{b}_i^{(k+1)}), \quad i = 1, \dots, n, \\ \mathbf{u}_i^{(k+1)} &= \mathbf{u}_i^{(k)} + \psi \mathbf{1}_p^T \mathbf{w}_i^{(k+1)}, \quad i = 1, \dots, n. \end{aligned} \tag{A8}$$

Appendix B. Update Algorithm for Constrained Weber Problem via ADMM

We consider the updates for the following constrained Weber problem via ADMM based on: [30].

$$\min_{\mathbf{w}_i^* \in \mathbb{R}^p} \left\{ \sum_{i=1}^n r_{i^*,i} \|\mathbf{w}_i^* - \widehat{\mathbf{w}}_i\|_2 \right\}, \quad \text{s.t.} \quad \sum_{j=1}^p w_{i^*j} = 0. \tag{A9}$$

The minimization problem (A9) is equivalently represented as:

$$\begin{aligned} \min_{\mathbf{w}_i^*, \mathbf{e}_1, \dots, \mathbf{e}_n \in \mathbb{R}^p} & \left\{ \sum_{i=1}^n r_{i^*,i} \|\mathbf{e}_i - \widehat{\mathbf{w}}_i\|_2 \right\}, \\ \text{s.t.} & \quad \mathbf{1}_p^T \mathbf{w}_i^* = 0, \mathbf{e}_i = \mathbf{w}_i^*, \quad i = 1, \dots, n. \end{aligned} \tag{A10}$$

The augmented Lagrangian for (A10) is formulated as:

$$\begin{aligned} L_{\rho, \phi}(\mathbf{w}_i^*, \mathbf{e}_1, \dots, \mathbf{e}_n) &= \sum_{i=1}^n \left\{ r_{i^*,i} \|\mathbf{e}_i - \widehat{\mathbf{w}}_i\|_2 + \mathbf{u}_i^T (\mathbf{e}_i - \mathbf{w}_i^*) + \frac{\mu}{2} \|\mathbf{e}_i - \mathbf{w}_i^*\|_2^2 \right\} \\ &+ v \mathbf{1}_p^T \mathbf{w}_i^* + \frac{\eta}{2} \|\mathbf{1}_p^T \mathbf{w}_i^*\|_2^2, \end{aligned} \tag{A11}$$

where \mathbf{u}_i, v are Lagrange multipliers and $\mu, \eta (> 0)$ are tuning parameters.

Appendix B.1. Update of \mathbf{w}_i^*

In the update of \mathbf{w}_i^* , we minimize the terms of the augmented Lagrangian (A11) depending on \mathbf{w}_i^* as follows:

$$\mathbf{w}_{i^*}^{(k+1)} = \arg \min_{\mathbf{w}_i^* \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \left\{ \frac{\mu}{2} \left\| \mathbf{w}_i^* - \mathbf{e}_i^{(k)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} \right\|_2^2 \right\} + v^{(k)} \mathbf{1}_p^T \mathbf{w}_i^* + \frac{\eta}{2} \|\mathbf{1}_p^T \mathbf{w}_i^*\|_2^2 \right\}. \tag{A12}$$

From $\frac{\partial L}{\partial \mathbf{w}_i^*} = \mathbf{0}$, we obtain the update:

$$\mathbf{w}_{i^*}^{(k+1)} = (\mu n \mathbf{1}_p + \eta \mathbf{1}_p \mathbf{1}_p^T)^{-1} \left\{ \mu \sum_{i=1}^n \left(\mathbf{e}_i^{(k)} + \frac{1}{\mu} \mathbf{u}_i^{(k)} \right) - v^{(k)} \mathbf{1}_p \right\}. \quad (\text{A13})$$

Appendix B.2. Update of \mathbf{e}

In the update of \mathbf{e}_i , we minimize the terms of augmented Lagrangian (A11) depending on \mathbf{e}_i as follows:

$$\mathbf{e}_i^{(k+1)} = \arg \min_{\mathbf{e}_i \in \mathbb{R}^p} \left\{ r_{i^*,i} \|\mathbf{e}_i - \mathbf{w}_{i^*}^{(k+1)}\|_2 + \frac{\mu}{2} \left\| \mathbf{e}_i - \left(\mathbf{w}_{i^*}^{(k+1)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} \right) \right\|_2^2 \right\}, \quad (\text{A14})$$

$$i = 1, \dots, n,$$

The minimization problem (A14) is equivalently expressed as:

$$\mathbf{e}_i^{(k+1)} = \arg \min_{\mathbf{e}_i \in \mathbb{R}^p} \left\{ \frac{r_{i^*,i}}{\mu} \|\mathbf{e}_i - \mathbf{w}_{i^*}^{(k)}\|_2 + \frac{1}{2} \left\| \mathbf{e}_i - \left(\mathbf{w}_{i^*}^{(k)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} \right) \right\|_2^2 \right\}, \quad (\text{A15})$$

$$i = 1, \dots, n.$$

In [30], because the right-hand side of (A15) is the proximal map of the function $f(\mathbf{e}_i) = \frac{r_{i^*,i}}{\rho} \|\mathbf{e}_i - \mathbf{w}_{i^*}\|_2$, by using Moreau's decomposition (e.g., [31]), the updates of \mathbf{e} are obtained by:

$$\mathbf{e}_i^{(k+1)} = \min \left(\frac{r_{i^*,i}}{\mu}, \left\| \mathbf{w}_{i^*}^{(k)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} - \hat{\mathbf{w}}_i \right\|_2 \right) \frac{\mathbf{w}_{i^*}^{(k)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} - \hat{\mathbf{w}}_i}{\left\| \mathbf{w}_{i^*}^{(k)} - \frac{1}{\mu} \mathbf{u}_i^{(k)} - \hat{\mathbf{w}}_i \right\|_2}. \quad (\text{A16})$$

Appendix B.3. Update of \mathbf{u} and v

The updates for Lagrange multipliers \mathbf{u}_i and v are obtained by gradient descent as follows:

$$\mathbf{u}_i^{(k+1)} = \mathbf{u}_i^{(k)} + \mu (\mathbf{e}_i^{(k+1)} - \mathbf{w}_{i^*}^{(k+1)}), \quad i = 1, \dots, n, \quad (\text{A17})$$

$$v^{(k+1)} = v^{(k)} + \eta \mathbf{1}_p^T \mathbf{w}_{i^*}^{(k+1)}.$$

References

- Argyriou, A.; Evgeniou, T.; Pontil, M. Convex multi-task feature learning. *Mach. Learn.* **2008**, *73*, 243–272. [\[CrossRef\]](#)
- Abdulnabi, A.H.; Wang, G.; Lu, J.; Jia, K. Multi-Task CNN Model for Attribute Prediction. *IEEE Trans. Multimed.* **2015**, *17*, 1949–1959. [\[CrossRef\]](#)
- Luong, M.T.; Le, Q.V.; Sutskever, I.; Vinyals, O.; Kaiser, L. Multi-task Sequence to Sequence Learning. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
- Lengerich, B.J.; Aragam, B.; Xing, E.P. Personalized regression enables sample-specific pan-cancer analysis. *Bioinformatics* **2018**, *34*, i178–i186. [\[CrossRef\]](#)
- Cowie, M.R.; Mosterd, A.; Wood, D.A.; Deckers, J.W.; Poole-Wilson, P.A.; Sutton, G.C.; Grobbee, D.E. The epidemiology of heart failure. *Eur. Heart J.* **1997**, *18*, 208–225. [\[CrossRef\]](#) [\[PubMed\]](#)
- Xu, J.; Zhou, J.; Tan, P.N. FORMULA: FactORized MULti-task LeArning for task discovery in personalized medical models. In Proceedings of the 2015 SIAM International Conference on Data Mining (SDM), Vancouver, BC, Canada, 30 April–2 May 2015; pp. 496–504.
- Yamada, M.; Koh, T.; Iwata, T.; Shawe-Taylor, J.; Kaski, S. Localized Lasso for High-Dimensional Regression. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 325–333.
- Hallac, D.; Leskovec, J.; Boyd, S. Network Lasso: Clustering and Optimization in Large Graphs. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 387–396.
- Wu, G.D.; Chen, J.; Hoffmann, C.; Bittinger, K.; Chen, Y.Y.; Keilbaugh, S.A.; Bewtra, M.; Knights, D.; Walters, W.A.; Knight, R.; et al. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science* **2011**, *334*, 105–108. [\[CrossRef\]](#) [\[PubMed\]](#)

10. Dillon, S.M.; Frank, D.N.; Wilson, C.C. The gut microbiome and HIV-1 pathogenesis: A two-way street. *AIDS* **2016**, *30*, 2737–2751. [[CrossRef](#)]
11. Arumugam, M.; Raes, J.; Pelletier, E.; Le Paslier, D.; Yamada, T.; Mende, D.R.; Fernandes, G.R.; Tap, J.; Bruls, T.; Batto, J.M.; et al. Enterotypes of the human gut microbiome. *Nature* **2011**, *473*, 174–180. [[CrossRef](#)]
12. Aitchison, J.; Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika* **1984**, *71*, 323–330. [[CrossRef](#)]
13. Lin, W.; Shi, P.; Feng, R.; Li, H. Variable selection in regression with compositional covariates. *Biometrika* **2014**, *101*, 785–797. [[CrossRef](#)]
14. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **1996**, *58*, 267–288. [[CrossRef](#)]
15. Boyd, S.; Parikh, N.; Chu, E. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*; Now Publishers Inc: Delft, The Netherlands, 2011.
16. Kong, D.; Fujimaki, R.; Liu, J.; Nie, F.; Ding, C. Exclusive Feature Learning on Arbitrary Structures via $\ell_{1,2}$ -norm. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1655–1663.
17. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc.* **1982**, *44*, 139–160. [[CrossRef](#)]
18. Shi, P.; Zhang, A.; Li, H. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* **2016**, *10*, 1019–1040. [[CrossRef](#)]
19. Wang, T.; Zhao, H. Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.* **2017**, *11*, 771–791. [[CrossRef](#)]
20. Bien, J.; Yan, X.; Simpson, L.; Müller, C.L. Tree-aggregated predictive modeling of microbiome data. *Sci. Rep.* **2021**, *11*, 14505. [[CrossRef](#)]
21. Combettes, P.L.; Müller, C.L. Regression Models for Compositional Data: General Log-Contrast Formulations, Proximal Optimization, and Microbiome Data Applications. *Stat. Biosci.* **2021**, *13*, 217–242. [[CrossRef](#)]
22. Friedman, J.; Hastie, T.; Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv* **2010**, arXiv:1001.0736.
23. Haro, C.; Rangel-Zúñiga, O.A.; Alcalá-Díaz, J.F.; Gómez-Delgado, F.; Pérez-Martínez, P.; Delgado-Lista, J.; Quintana-Navarro, G.M.; Landa, B.B.; Navas-Cortés, J.A.; Tena-Sempere, M.; et al. Intestinal microbiota is influenced by gender and body mass index. *PLoS ONE* **2016**, *11*, e0154090. [[CrossRef](#)]
24. Saraswati, S.; Sitaraman, R. Aging and the human gut microbiota—from correlation to causality. *Front. Microbiol.* **2015**, *5*, 764. [[CrossRef](#)]
25. McMurdie, P.J.; Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. [[CrossRef](#)]
26. Gower, J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857–871. [[CrossRef](#)]
27. Greenacre, M. *Compositional Data Analysis in Practice*; CRC Press: Boca Raton, FL, USA, 2018.
28. Cuevas-Sierra, A.; Riezu-Boj, J.I.; Gुरुceaga, E.; Milagro, F.I.; Martínez, J.A. Sex-Specific Associations between Gut Prevotellaceae and Host Genetics on Adiposity. *Microorganisms* **2020**, *8*, 938. [[CrossRef](#)] [[PubMed](#)]
29. Zeng, Q.; Li, D.; He, Y.; Li, Y.; Yang, Z.; Zhao, X.; Liu, Y.; Wang, Y.; Sun, J.; Feng, X.; et al. Discrepant gut microbiota markers for the classification of obesity-related metabolic abnormalities. *Sci. Rep.* **2019**, *9*, 13424. [[CrossRef](#)] [[PubMed](#)]
30. Chaudhury, K.N.; Ramakrishnan, K.R. A new ADMM algorithm for the Euclidean Median and its application to robust patch regression. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015.
31. Parikh, N.; Boyd, S. Proximal Algorithms. *Found. Trends Optim.* **2014**, *1*, 127–239. [[CrossRef](#)]