A compositional-group LASSO for compositional covariates

Xuejun Ma * Yao Dong †

Abstract

In this paper, we propose a compositional-group LASSO to deal with the selection of group variables for high-dimensional compositional data. We introduce an algorithm to solve this convex optimization problem via generalized gradient descent. Simulation studies show the effectiveness and the flexibility of the proposed method. Finally, we study the gut microbiome data using the proposed method.

Key words: compositional data, compositional-group LASSO, log-contrast model.

MSC2010 subject classifications: Primary 62J05; secondary 62J07.

1 Introduction

With the advent of modern technology for data collection, researchers are able to collect compositional data, which strictly positive and multivariate that are constrained to have a unit sum, in diverse fields of scientific research, such as geochemical compositions of rocks in geology and species compositions of biological communities in ecology. The data also referred to as mixture data [Aitchison and Bacon-Shone (1984), Cornell (2002), Snee (1973) and Lin et al. (2014)]. Linear regression is a fundamental and commonly used technique for characterizing the relationship between a response variable, said Y, and a group of predictors variables, said X. Compositional predictors variables need to account of the intrinsic multivariate nature. Cornell (2002) applied the log-ratio transformation[Aitchison and Bacon-Shone (1984)], and developed the linear log-contrast model.

Variable selection plays a central role in high-dimensional data analysis. Many methods have been proposed, and become increasingly frequent and important in various research fields. These methods include, but are not limited to, LASSO [Tibshirani (1996)], adaptive LASSO[Zhou et al. (2006)], SCAD [Fan et al. (2001)] and elastic net [Zou et al. (2005)] When predictor variables are divided into different groups, Yuan et al. (2006) proposed the group LASSO to select group variables. However, the method can not identify the sparse solution within a group. Simon et al. (2013) proposed sparse-group LASSO to overcome this problem .

For compositional data, the conventional regularization methods may not perform well since the linear constraints on regression coefficients. The above variable selection method can not be simply extended to compositional data. Lin et al. (2014) proposed a variable selection method with ℓ_1 regularization based on the log-contrast transformation. Shi et al. (2016) introduced more general high-dimensional linear model with many linear constraints on coefficients proposed subcompositional regression model. However, above methods do not work for group variables.

*School of Mathematical Sciences, Soochow University, 215006, Suzhou, China, xuejunma@suda.edu.cn

†The corresponding author, School of Mathematical Sciences, Soochow University, 215006, Suzhou, China, 20184207042@stu.suda.edu.cn

Motivated by Simon et al. (2013), we consider compositional-group LASSO which can describe effect of group-wise and within group sparsity in high dimensional linear log-contrast model. Furthermore, we show an iterative algorithm based on the subgradient method.

The paper is organized as follows. In Section 2, we introduce the proposed variable selection method, and present the computational algorithm and a method for selecting the tuning parameter. Section 3 shows some results from simulation studies for assessing the finite sample properties of the proposed approach, which suggests that it works well for practical situations. In Section 4, the proposed method is applied to the gut microbiome data set to identify bacterial genes are associated with BMI.

2 Regression model for compositional data

2.1 Some Preliminaries

Let $y = (y_1, ..., y_n)^\top$ be the response vector. $X = (x_{ij})$ is an $n \times p$ design matrix (covariate matrix). Each row of the matrix is in the $(p-1)$-dimension positive simplex $S^{p-1} = \{(x_{i1}, \cdots, x_{ip}) : x_{ij} > 0,\ \sum_{j=1}^{p} x_{ij} = 1,\ i = 1, \ldots, n;\ j = 1, \ldots, p\}$. Based on the log-ratio transformation of Aitchison (1982), Aitchison and Bacon-Shone (1984), we develop the linear log-contrast model as follows:

$$y = Z^p \beta_{\backslash p} + \varepsilon \tag{1}$$

where $Z^p = \{\log(x_{ij}/x_{ip})\}$ is an $n \times (p-1)$ log-ratio matrix. Here the pth component is the reference component. $\beta_{\backslash p} = (\beta_1, \ldots, \beta_{p-1})^\top$ is the parameter. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the error term. The selection of the reference component is crucial in model (1), especially in high dimensional data. As Lin et al. (2014), we let $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, and model (1) can be expressed as

$$y = Z\beta + \varepsilon, \qquad \text{subject to } \mathbf{1}_p^\top \beta = 0, \tag{2}$$

where $\mathbf{1}_p = (1, \ldots, 1)^\top \in \mathbb{R}^p$, $Z = \log(x_{ij}) \in \mathbb{R}^{n \times p}$, and $\beta = (\beta_1, \ldots, \beta_p)^\top$. Furthermore, they proposed an ℓ_1 regularization method for the linear log-contrast model

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1\right), \qquad \text{subject to} \sum_{j=1}^{p} \beta_j = 0$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. λ is the tuning parameter.

2.2 Compositional-group LASSO

The matrix design matrix Z is divided into m different groups $Z^{(1)}, \ldots, Z^{(m)}$. Here $Z^{(l)}$ is an $n \times p_l$ matrix where p_l is the number of covariates in group l. Now, we consider the compositional-group LASSO which is to minimize the following objective function

$$\begin{aligned} &\min_{\beta} \frac{1}{2n}\|y - \sum_{l=1}^{m} Z^{(l)}\beta^{(l)}\|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l}\|\beta^{(l)}\|_2 + \alpha\lambda\|\beta\|_1 \\ &s.t. \sum_{j=1}^{p} \beta_j = 0 \end{aligned} \tag{3}$$

where $\alpha \in [0, 1]$ is a trade-off between LASSO and group LASSO penalties. The method can get both sparsity of groups and within each group. When $\alpha = 1$, it is the LASSO of Lin et al. (2014). When $\alpha = 0$, it is the group LASSO of Yuan et al. (2006).

2.3 Algorithm

Using lagrange multiplication, (3) can be expressed as

$$\frac{1}{2n}\|y-\sum_{l=1}^{m}Z^{(l)}\beta^{(l)}\|_2^2+(1-\alpha)\lambda\sum_{l=1}^{m}\sqrt{p_l}\|\beta^{(l)}\|_2+\alpha\lambda\|\beta\|_1+\omega\sum_{j=1}^{p}\beta_j \tag{4}$$

For the convenience of calculation, we use $(1-\alpha)\lambda$ instead of $(1-\alpha)\lambda\sqrt{p_l}$. (4) is a convex function. Since $\|\beta\|_1$ is not differentiable at 0, the subgradient method of Boyd et al. (2003) is used to find the optimal solution β. If $\hat{\beta}$ is the optimal solution, for group k, $\hat{\beta}^{(k)}$ must satisfy:

$$\frac{1}{n}Z^{(k)\top}\Big(y-\sum_{l=1}^{m}Z^{(l)}\hat{\beta}^{(l)}\Big)=(1-\alpha)\lambda u+\alpha\lambda v+\omega\mathbf{1}_{p_k}$$

Here u and v represent the subgradients of $\|\beta^{(k)}\|_2$ and $\|\beta^{(k)}\|_1$ respectively. $\omega\mathbf{1}_{p_k}$ denotes the gradients of $\sum_{j=1}^{p}\beta_j$. $\mathbf{1}_{p_k}$ is the unit column vector of length p_k.

$$u=\begin{cases}\frac{\hat{\beta}^{(k)}}{\|\hat{\beta}^{(k)}\|_2}, & \text{if}\quad \hat{\beta}^{(k)}\neq\mathbf{0}\\ \in\{u:\|u\|_2\leq 1\}, & \text{if}\quad \hat{\beta}^{(k)}=\mathbf{0}\end{cases}$$

$$v=\begin{cases}\text{sign}(\hat{\beta}_j^{(k)}), & \text{if}\quad \hat{\beta}_j^{(k)}\neq 0\\ \in\{v:\|v\|_1\leq 1\}, & \text{if}\quad \hat{\beta}_j^{(k)}=0\end{cases}$$

If $\hat{\beta}^{(k)}=\mathbf{0}$, elementary calculations show that

$$\Big\|S(Z^{(k)\top}r_{(-k)}/n,\alpha\lambda)\Big\|_2\leq(1-\alpha)\lambda+\omega \tag{5}$$

where

$$r_{(-k)}=y-\sum_{l\neq k}Z^{(l)}\hat{\beta}^{(l)}$$

and S$(\cdot)$ is the coordinate-wise soft thresholding operator

$$S(\alpha,\beta)=\text{sign}(\alpha)(|\alpha|-\beta)_+$$

We extend the method of Simon et al. (2013), using the blockwise descent method to solve the compositional-group LASSO problem, which is simple since our punishment is separable between groups. Now, we consider the effect of the kth group, and suppose the coefficients of the other groups as known constants. Minimizing equation (4) is equivalent to finding $\beta^{(k)}$, minimizing

$$\frac{1}{2n}\|r_{(-k)}-Z^{(k)}\beta^{(k)}\|_2^2+(1-\alpha)\lambda\|\beta^{(k)}\|_2+\alpha\lambda\|\beta^{(k)}\|_1+\omega\sum_{s=1}^{p_k}\beta_{ks} \tag{6}$$

Let $\ell(r_{(-k)},\beta^{(k)})=\frac{1}{2n}\|r_{(-k)}-Z^{(k)}\beta^{(k)}\|_2^2$, by Taylor expansion, we have

$$\ell(r_{(-k)},\beta^{(k)})\leq\ell(r_{(-k)},\beta_0)+(\beta^{(k)}-\beta_0)^\top\nabla\ell(r_{(-k)},\beta_0)+\frac{1}{2t}\|\beta^{(k)}-\beta_0\|_2^2 \tag{7}$$

where t is sufficiently small that the quadratic term dominates the Hessian of our loss. $\nabla\ell(r_{(-k)},\beta_0)$ refers only to the subgradient of the group k. Minimizing this function

would give us our usual gradient step(with stepsize t) in the unpenalized case. Add (7) to objective function (6),

$$M(\beta^{(k)}) = \ell(r_{(-l)}, \beta_0^{(k)}) + (\beta^{(k)} - \beta_0^{(k)})^T \nabla \ell(r_{(-l)}, \beta_0^{(k)}) + \frac{1}{2t}\|\beta^{(k)} - \beta_0^{(k)}\|_2^2 \\ +(1-\alpha)\lambda\|\beta^{(k)}\|_2 + \alpha\lambda\|\beta^{(k)}\|_1 + \omega\sum_{s=1}^{p_k}\beta_{ks} \tag{8}$$

where $\beta_0^{(k)}$ is the initial solution of equation object function (6). Our goal is to find $\hat{\beta}^{(k)}$ to minimize $M(\beta^{(k)})$, which is equivalent to minimizing

$$\frac{1}{2n}\|(\beta^{(k)} - \beta_0^{(k)}) - t\nabla\ell(r_{(-l)}, \beta_0^{(k)})\|_2^2 + (1-\alpha)\lambda\|\beta^{(k)}\|_2 \\ +\alpha\lambda\|\beta^{(k)}\|_1 + \omega\sum_{s=1}^{p_k}\beta_{ks} \tag{9}$$

If $\beta^{(k)} = \mathbf{0}$,

$$\left\|S\Big(\beta_0^{(k)} - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}), t\alpha\lambda\Big)\right\|_2 \leq t(1-\alpha)\lambda + \omega$$

If $\beta^{(k)} \neq \mathbf{0}$,

$$\Big(1 + \frac{t(1-\alpha)\lambda}{\|\beta^{(k)}\|_2}\Big)\beta^{(k)} = S(\beta_0 - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda) \tag{10}$$

Taking the norm of both sides we can obtain

$$\|\beta^{(k)}\|_2 = \Big(\|S(\beta_0 - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda)\|_2 - t(1-\alpha)\lambda\Big)_+$$

We plug this into (10), and obtain

$$\beta^{(k)} = \Big(1 - \frac{t(1-\alpha)\lambda}{\|S(\beta_0^{(k)} - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda)\|_2}\Big)_+ S(\beta_0^{(k)} - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda) \tag{11}$$

If we update equation (11), recenter each pass at $(\beta^{(k)})_{new} = (\beta_0^{(k)})_{old}$, then we can get the optimal $\beta_{(k)}$ with fixed other coefficients. We use the above formula to calculate β of each group, and converge to the global optimal solution. $U(\beta_0^{(k)}, t)$ denotes our update formula

$$U(\beta_0^{(k)}, t) = \Big(1 - \frac{t(1-\alpha)\lambda}{\|S(\beta_0^{(k)} - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda)\|_2}\Big)_+ S(\beta_0 - t\nabla\ell(r_{(-k)}, \beta_0^{(k)}) - \omega, t\alpha\lambda)$$

Algorithm Using coordinate descent method to solve the problem (3)
Step 1. Initialize β_0 with 0 or a warm start, $k = 0$.
Step 2. For $j = 1, 2, \ldots, m$, if $\beta^{(j)}$ satisfies

$$\|S(Z^{(j)\top} r_{(-j)}/n, \alpha\lambda)\|_2 \leq (1-\alpha)\lambda + \omega$$

then let $\beta_{k+1}^{(j)} = 0$. If not, within the group apply *Step 3.*
Step 3. Iterate θ until convergence
(a) Update $\theta \leftarrow \beta_k^{(j)}$
(b) Update $\beta_k^{(j)}$ by

$$\beta_k^{(j)} \leftarrow U(\theta, t)$$

Let $\beta_{k+1}^{(j)} = \theta$, and return *Step 2*.

Step 4. Update $k+1 \leftarrow k$ and repeat Steps 2 and 3 until convergence, output $\hat{\beta} = \beta_{k+1}$.

Step 2 is mainly to judge the importance of the group. *Step 3* is to judge the importance of the variables within the group .

Remark: Suppose β_{sgl} is the solution of sparse-group LASSO, and the solution of compositional-group LASSO is β_{cgl}. By calculating the simple scaling, we can know that $\beta_{cgl} = S(\beta_{sgl}, \tilde{\omega})$.

2.4 Selection of λ, α and ω

Now, we consider the choices of LASSO and group LASSO trade-off parameter α, penalty rate parameter λ, and compositional data coefficient linear constraints parameter ω. If we use grid search to select parameters, there are many parameters choices. Hence the efficiency of this method is not good. Here we introduce a simple way. Firstly, we study ω. ω is determined by data. We set $n\omega \in [1, 2]$ for low signal-to-noise ratio, and $n\omega \in [2, 3]$ for the other cases. From the simulations, the method performs well. For the choice of λ and α, we apply the method of Simon et al. (2013).

3 Simulations

As Lin et al. (2014), we generate the covariate data in the following way. We first generate an $n \times p$ data matrix $X = (x_{ij})$ from a multivariate normal distribution $N_p(\theta, \Sigma)$, and then obtain the covariate matrix $Z = (z_{ij})$ by the transformation $z_{ij} = \exp(x_{ij}) / \sum_{k=1}^{p} \exp(x_{ik})$. To reflect the differences of component data, we let $\theta = (\theta_j)$ with $\theta_j = \log(0.5p)$ for $j = 1, \ldots, 5$ and $\theta_j = 0$ otherwise. To describe different levels of correlations among the components, we let $\Sigma = (\rho^{|i-j|})$ with $\rho = 0.2$ or 0.5. We consider the following linear model.

$$y = X\beta + \varepsilon$$

The error term is generated from five distributions: (1) standard normal distribution: $N(0, 1)$; (2) t-distribution: $t(3)$; (3) standard laplace distribution: $Laplace(0, 1)$; (4) gamma distribution: $Ga(2, 2)$; (5) standard exponential distribution: $Exp(1)$. β is generated from three cases:

Case 1 Three variables is one group. Sparsity rate within group is 1/3.

$$\beta = (1, -0.8, 0, 0.6, 0, 0, -1.5, -0.5, 1.2, 0, \ldots, 0)^\top.$$

Case 2 Eight variables is one group. Sparsity rate within group is 45%.

$$\beta = (1, -0.8, 0.4, 0, 0, -0.6, 0, 0, 0, 0, -1.5, 0, 1.2, 0, 0, 0.3, 0, \cdots, 0)^\top$$

Case 3 Eight variables is one group. Sparsity rate within group is 0%.

$$\beta = (1, -0.8, 0.4, 0.3, -0.2, -0.6, 0.5, -0.3, 1.2, -0.7, -1.5, 0.45, 1.2, 0.7, -1.5, -0.15, 0, \ldots, 0)^\top.$$

We mainly consider variable selection in Case 1 where n is bigger than p, and set $(n, p) = (50, 10), (100, 10), (100, 20)$ and $(200, 20)$. In Cases 2 and 3, we consider $n < p$, and set $(n, p) = (50, 100), (100, 500)$ and $(100, 1000)$. Group Sparsity rate within group is different between Cases 2 and 3. The latter is an extreme case which has no sparsity within group. The procedure is repeated 500 times. We evaluate the performance through the following four criteria:

(1)

$$GRCI = \frac{1}{500}\sum_{i=1}^{500}(GCI_i),$$

where GCI_i is based on $i-th$ sample. If all groups are completely correctly identified, $GCI_i = 1$. Otherwise, $GCI_i = 0$. Therefore, the closer GRCI is to 1, the better model performance.

(2)

$$GII = GTN + GFP$$

where GTN is the average number of groups true negatives. GFP is the average number of groups false positives. So, the closer GII is to 0, the better model performance.

(3)

$$RCI = \frac{1}{500}\sum_{i=1}^{500}(CI_i)$$

where CI_i is based on $i-th$ sample. If all variables completely correctly identified, $CI_i = 1$; $CI_i = 0$ otherwise. Hence, the closer GCI is to 1, the better model performance.

(4)

$$II = TN + FP$$

where TN is the average number of true negatives. FP is the average number of false positives. Therefore, the closer II is to 0, the better model performance.

Tables 1-3 report the results. CGL outperforms than SGL regardless of p and the error distribution since SGL do not consider the constraint on data, which is more inclined to select the important variables or groups. CRCI is superior to RCI, which implies that the identification of group is better than within the group since signal-to-noise ratio is very high. CGL can work in low and high dimensional data. Furthermore, the performances of two methods increase gradually with n.

Table 1: *Simulation results for Case 1*

(n,p)	Distribution	Method	$\rho=0.2$				$\rho=0.5$			
			GRCI	GII	RCI	II	GRCI	GII	RCI	II
(50, 10)	$N(0,1)$	CGL	0.968	0.016	0.776	0.246	0.956	0.026	0.614	0.452
		SGL	0.627	0.376	0.000	3.180	0.598	0.402	0.000	3.208
(100, 10)		CGL	0.922	0.008	0.816	0.204	0.976	0.012	0.724	0.300
		SGL	0.700	0.300	0.000	3.112	0.756	0.244	0.000	2.968
(100, 20)		CGL	0.996	0.004	0.760	0.256	0.092	0.004	0.808	0.220
		SGL	0.840	0.500	0.000	3.450	0.436	0.820	0.000	4.904
(200, 20)		CGL	0.922	0.008	0.816	0.234	1.000	0.000	0.752	0.276
		SGL	0.700	0.300	0.000	3.024	0.778	0.237	0.000	3.348
(50, 10)	$t(3)$	CGL	0.832	0.124	0.448	0.792	0.696	0.124	0.320	1.064
		SGL	0.472	0.528	0.000	3.400	0.472	0.516	0.000	3.324
(100, 10)		CGL	0.924	0.072	0.484	0.660	0.864	0.100	0.500	0.680
		SGL	0.552	0.448	0.000	3.280	0.612	0.388	0.000	3.228
(100, 20)		CGL	0.772	0.292	0.456	0.940	0.764	0.220	0.448	0.844
		SGL	0.112	1.776	0.000	7.440	0.064	2.064	0.000	8.320
(200, 20)		CGL	0.916	0.088	0.460	0.744	0.916	0.088	0.492	0.716
		SGL	0.248	1.300	0.000	6.196	0.122	1.230	0.000	6.480
(50, 10)	$Lap(0,1)$	CGL	0.916	0.048	0.552	0.556	0.856	0.080	0.392	0.840
		SGL	0.528	0.472	0.000	3.340	0.584	0.416	0.000	3.286
(100, 10)		CGL	0.956	0.044	0.524	0.568	0.924	0.360	0.584	0.516
		SGL	0.620	0.380	0.000	3.184	0.740	0.260	0.000	3.104
(100, 20)		CGL	0.884	0.120	0.536	0.592	0.896	0.080	0.580	0.528
		SGL	0.276	1.256	0.000	6.048	0.144	1.769	0.000	7.616
(200, 20)		CGL	0.984	0.016	0.596	0.480	0.984	0.016	0.584	0.488
		SGL	0.568	0.532	0.000	4.036	0.568	0.532	0.000	4.036
(50, 10)	$Ga(2,2)$	CGL	0.976	0.004	0.892	0.128	0.832	0.012	0.756	0.272
		SGL	0.712	0.288	0.000	3.052	0.760	0.232	0.000	2.980
(100, 10)		CGL	1.000	0.000	0.880	0.132	0.996	0.006	0.860	0.144
		SGL	0.844	0.156	0.000	2.872	0.812	0.188	0.000	2.888
(100, 20)		CGL	1.000	0.000	0.892	0.136	0.992	0.004	0.912	0.092
		SGL	0.822	0.330	0.000	2.508	0.752	0.280	0.000	4.408
(200, 20)		CGL	1.000	0.000	0.960	0.040	1.000	0.000	0.864	0.098
		SGL	0.972	0.026	0.000	2.812	0.972	0.032	0.000	2.736
(50, 10)	$Exp(1)$	CGL	0.964	0.02	0.776	0.264	0.812	0.040	0.776	0.264
		SGL	0.644	0.352	0.000	3.152	0.622	0.348	0.000	3.167
(100, 10)		CGL	1.000	0.000	0.788	0.224	0.968	0.040	0.768	0.240
		SGL	0.748	0.252	0.004	2.956	0.704	0.276	0.004	3.214
(100, 20)		CGL	0.984	0.016	0.828	0.160	0.972	0.020	0.798	0.172
		SGL	0.432	0.908	0.000	3.092	0.380	0.648	0.000	4.332
(200, 20)		CGL	1.000	0.000	0.776	0.260	1.000	0.000	0.768	0.267
		SGL	0.844	0.176	0.000	3.112	0.820	0.204	0.000	3.188

Table 2: *Simulation results for Case 2*

(n,p)	Distribution	Method	$\rho=0.2$				$\rho=0.5$			
			GRCI	GII	RCI	II	GRCI	GII	RCI	II
(50, 100)	$N(0,1)$	CGL	0.844	0.156	0.200	1.756	0.822	0.212	0.070	2.684
		SGL	0.700	0.560	0.000	13.00	0.200	2.060	0.000	23.04
(100, 500)		CGL	0.990	0.010	0.048	0.750	0.970	0.030	0.060	1.940
		SGL	0.140	5.700	0.000	51.32	0.000	3.780	0.000	36.40
(100, 1000)		CGL	0.980	0.020	0.386	0.950	0.932	0.018	0.086	2.030
		SGL	0.200	3.780	0.000	36.40	0.000	8.820	0.000	76.54
(50,100)	$t(3)$	CGL	0.756	0.314	0.126	3.370	0.750	0.334	0.032	3.366
		SGL	0.200	2.100	0.000	22.96	0.020	3.460	0.000	32.90
(100, 500)		CGL	0.805	0.189	0.316	1.589	0.708	0.424	0.036	3.458
		SGL	0.620	1.320	0.000	17.58	0.100	4.880	0.000	44.52
(100, 1000)		CGL	0.857	0.123	0.341	1.751	0.817	0.520	0.341	3.665
		SGL	0.520	1.640	0.000	20.38	0.080	5.400	0.000	49.04
(50,100)	$Lap(0,1)$	CGL	0.850	0.174	0.146	1.852	0.750	0.334	0.032	3.340
		SGL	0.140	1.880	0.000	21.64	0.080	2.780	0.000	28.24
(100, 500)		CGL	0.983	0.019	0.438	0.816	0.910	0.122	0.106	2.292
		SGL	0.700	0.380	0.000	10.70	0.180	23.20	0.000	25.66
(100, 1000)		CGL	0.982	0.018	0.488	0.754	0.894	0.142	0.086	2.502
		SGL	0.660	0.940	0.000	15.06	0.040	4.560	0.000	42.60
(50, 100)	$Ga(2,2)$	CGL	1.000	0.000	0.542	0.547	1.000	0.000	0.254	0.327
		SGL	0.600	0.040	0.000	7.960	0.060	2.120	0.000	23.48
(100, 500)		CGL	1.000	0.000	0.764	0.284	1.000	0.000	0.382	1.090
		SGL	0.760	0.040	0.000	7.960	0.200	0.720	0.000	13.32
(100, 1000)		CGL	0.998	0.002	0.742	0.300	0.957	0.063	0.296	1.130
		SGL	0.700	0.120	0.000	8.580	0.170	1.360	0.000	18.70
(50, 100)	$Exp(1)$	CGL	0.980	0.020	0.340	0.950	0.950	0.060	0.190	1.590
		SGL	0.320	1.300	0.000	16.76	0.040	2.460	0.000	25.80
(100, 500)		CGL	0.990	0.010	0.720	0.290	0.930	0.070	0.410	0.870
		SGL	0.880	0.140	0.000	8.860	0.340	1.240	0.000	17.46
(100, 1000)		CGL	0.978	0.024	0.684	0.368	0.959	0.044	0.418	0.852
		SGL	0.800	0.200	0.000	9.020	0.200	2.360	0.000	25.86

Table 3: *Simulation results for Case 3*

(n,p)	Distribution	Method	$\rho = 0.2$				$\rho = 0.5$			
			GRCI	GII	RCI	II	GRCI	GII	RCI	II
(50, 100)	$N(0,1)$	CGL	0.890	0.122	0.120	1.766	0.740	0.334	0.520	2.644
		SGL	0.560	0.760	0.560	5.560	0.520	0.800	0.050	5.900
(100, 500)		CGL	0.974	0.026	0.710	0.318	0.948	0.056	0.504	0.658
		SGL	0.400	1.700	0.400	13.10	0.200	1.680	0.200	14.22
(100, 1000)		CGL	0.988	0.012	0.692	0.354	0.970	0.032	0.404	0.862
		SGL	0.120	0.920	0.120	17.96	0.110	3.340	0.110	26.06
(50, 100)	$t(3)$	CGL	0.424	1.070	0.040	3.393	0.325	1.250	0.000	5.238
		SGL	0.020	2.620	0.030	20.14	0.060	3.700	0.060	27.32
(100, 500)		CGL	0.614	0.726	0.272	1.844	0.480	1.03	0.110	3.000
		SGL	0.260	2.200	0.100	15.66	0.100	3.960	0.080	30.40
(100, 1000)		CGL	0.604	0.746	0.238	1.938	0.510	1.140	0.100	3.580
		SGL	0.120	3.500	0.060	30.21	0.040	5.860	0.040	45.66
(50,100)	$Lap(0,1)$	CGL	0.476	0.792	0.032	3.280	0.396	1.058	0.026	4.456
		SGL	0.140	1.040	0.440	17.30	0.060	3.620	0.060	26.76
(100, 500)		CGL	0.750	0.328	0.368	1.052	0.630	0.218	0.238	1.392
		SGL	0.220	2.300	0.220	17.30	0.080	3.000	0.080	23.16
(100, 1000)		CGL	0.766	0.308	0.370	1.076	0.682	0.410	0.150	1.948
		SGL	0.160	3.540	0.160	27.56	0.040	4.600	0.040	36.02
(50,100)	$Ga(2,2)$	CGL	0.476	0.792	0.032	3.280	0.396	1.058	0.026	4.456
		SGL	0.140	1.040	0.440	17.30	0.060	3.620	0.060	26.76
(100, 500)		CGL	0.750	0.328	0.368	1.052	0.630	0.218	0.238	1.392
		SGL	0.220	2.300	0.220	19.50	0.080	3.000	0.080	23.16
(100, 1000)		CGL	0.766	0.308	0.370	1.076	0.682	0.410	0.150	1.948
		SGL	0.160	3.540	0.160	27.56	0.040	4.600	0.040	36.06
(50, 100)	$Exp(1)$	CGL	0.830	0.200	0.148	1.838	0.760	0.303	0.062	2.686
		SGL	0.340	1.300	0.340	9.000	0.060	2.820	0.060	10.82
(100, 500)		CGL	0.972	0.030	0.702	0.338	0.944	0.068	0.068	0.454
		SGL	0.540	1.040	0.540	7.720	0.400	1.700	0.400	13.20
(100, 1000)		CGL	0.950	0.050	0.668	0.382	0.948	0.058	0.400	0.940
		SGL	0.440	1.120	0.540	8.540	0.070	3.120	0.020	24.42

4 Gut microbiome data

Gut microbiome composition is an important role in food digestion and nutrition. Wu et al. (2011) reported a cross-sectional study of 98 healthy volunteer carriers to investigate the connections between micronutrients and gut micirbiome composition. The DNAs from fecal samples were analysed by 454/Roche pyrosequencing of 16S rRNA gene segment from the V1-V2 region. After the pyrosequences were denoised, we obtained with an average of 9168 reads per sample with a standard deviation of 3864, and 3068 operaional taxonomic units(OUTs) were obtained. The OTUs were combined into 89 genera that appeared in at least one sample. Out of these 87 genera, 42 genera had zero counts in more than 90% of the samples and were removed from our analysis. The remaining 160 common genera come from eight phylum, *Actinobacteria*, *Bacteroidetes*, *Cyanobacteria*, *Firmicutes*, *Fusobacteria*, *Proteobacteria*, *Tenericutes* and *Verrucomicrobia*. Since dysbiosis of gut microbiome has associated with BMI [Ley et al. (2005), Ley et al. (2006) and Turnbaugh et al. (2006)]

and we are interested in identify the bacterial genera that are associated with *BMI* after adjustinng for total fat and caloric intacks. We transform them into compositional data after replacing zero count by the maximum rounding error 0.5 [Aitchison (1982)]. So we set the following model as

$$E(BMI) = \sum_{g=1}^{9} \sum_{s=1}^{m_g} \beta_{gs} \log(X_{gs}).$$

We set $\alpha = 0.9$, and divide covariates into eight groups based on Phylum where the variable is Genus. $\log(X_{gs})$ is the logarithm of the relative abundance of sth genus of gth phylum. We apply the proposed model to the dataset with BMI as the response.

We use the bootstrap method, and get the selection probability to assess the importance of group and variable within the group. The replication times is 100. Selection probability is the proportion of selecting Phylum and Genus.

Table 4: *Selection probilities of five genera in the gut microbiome data*

Phylum	Class	Genus	Selection probability
Actinobacteria	Coriobacteriia	Enterorhabdus	73%
Bacteroidetes	Bacteroidia	Alistipes	83%
Firmicutes	Clostridia	Oscillibacter	95%
Firmicutes	Bacilli	Streptococcus	86%
Firmicutes	Clostridia	Faecalibacterium	78%

From Table 4, the proposed method selects *Actinobacteria*, *Bacteroidetes* and *Firmicutes* as being associated with BMI at the phylum level, which is consistent with Koliada et al. (2017). They studied Ukrainian adult data and indicates that obese persons have a significantly higher level of *Firmicutes* and lower level of *Bacteroidetes* compared to normal-weight and lean adults.

Moreover, we get that bacterial genus *Oscillibacter* is more likely to be associated with BMI at the genus level. A recent study also found that when *Bacteroides* and *Faecalibacterium* were equally abundant, the abundance of *Oscillibacter* was the major determinant of obese or normal status [Hu et al. (2015)]. In other words, these clearly demonstrate the effectiveness and the flexibility of the proposed method.

5 Discussion

This paper discusses variable selection of high dimensional compositional data, especially for group variable selection . Since we add ℓ_1 and ℓ_2 regular terms , the estimation of the coefficients is biased. Breheny et al. (2009, 2015) proposed an unbiased estimation of group variables. We are interested in extending our method to obtain unbiased estimation in our future research..

References

Aitchison, J., and Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. Biometrika, 71(2), 323-330.

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B, 44(2), 139-160.

Breheny, P., and Huang, J. (2009). Penalized methods for bi-level variable selection. Statistics and Its Interface, 2(3), 369-380.

Breheny, P., and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing, 25(2), 173-187.

Boyd, S., Xiao, L., and Mutapcic, A. (2003). Subgradient methods. Lecture Notes of EE392o, Stanford University, Autumn Quarter, 2004, 2004-2005.

Cornell, J. (2002). Experiments with mixtures: designs, models, and the analysis of mixture data, 3rd ed. Wiley, New York. MR1882356.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360.

Hu, H. , Park, S. , Jang, H. , Choi, M. , Park, K. , Kang, J. , Park, S. , Lee, H. , and Cho, S. (2015). Obesity alters the microbial community profile in Korean adolescents. PloS one, 10(7), e0134333.

Koliada, A., Syzenko, G., Moseiko, V., Budovska, L., Puchkov, K., Perederiy, V., Gavalko, Y., Dorofeyev, A., Romanenko, M., Tkach, S., et al. (2017). Association between body mass index and firmicutes/bacteroidetes ratio in an adult ukrainian population. BMC Microbiology, 17(1), 120.

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

Ley, R., Backhed, F., Turnbaugh, P., Lozupone, C., Knight, R., and Gordon, J. (2005). Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America, 102(31), 11070-11075.

Ley, R., Turnbaugh, P., Klein, S., and Gordon, J. (2006). Human gut microbes associated with obesity. Nature, 444(7122), 1022-1023.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. Journal of Computational and Graphical Statistics, 22(2), 231-245.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. The Annals of Applied Statistics, 10(2), 1019-1040.

Snee, R. (1973). Techniques for the analysis of mixture data. Technometrics, 15(3), 517-528.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

Turnbaugh, P., Ley, R., Mahowald, M., Magrini, V., Mardis, E., and Gordon, J. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444(7122), 1027-1031.

Wu, G., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y., Keilbaugh, S., Bewtra, M., Knights, D., Walters, W., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F., and Lewis, J. (2011). Linking long-term dietary patterns with gut microbial enterotypes. Science, 334(6052), 105-108.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B, 68(1), 49-67.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418-1429.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301-320.