Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis

Jun Chen, Hongzhe Li

(Submitted on 23 May 2013)

With the development of next generation sequencing technology, researchers have now been able to study the microbiome composition using direct sequencing, whose output are bacterial taxa counts for each microbiome sample. One goal of microbiome study is to associate the microbiome composition with environmental covariates. We propose to model the taxa counts using a Dirichlet-multinomial (DM) regression model in order to account for overdispersion of observed counts. The DM regression model can be used for testing the association between taxa composition and covariates using the likelihood ratio test. However, when the number of covariates is large, multiple testing can lead to loss of power. To address the high dimensionality of the problem, we develop a penalized likelihood approach to estimate the regression parameters and to select the variables by imposing a sparse group ℓ_1 penalty to encourage both group-level and within-group sparsity. Such a variable selection procedure can lead to selection of the relevant covariates and their associated bacterial taxa. An efficient block-coordinate descent algorithm is developed to solve the optimization problem. We present extensive simulations to demonstrate that the sparse DM regression can result in better identification of the microbiome-associated covariates than models that ignore overdispersion or only consider the proportions. We demonstrate the power of our method in an analysis of a data set evaluating the effects of nutrient intake on human gut microbiome composition. Our results have clearly shown that the nutrient intake is strongly associated with the human gut microbiome.

Comments:	Published in at this http URL the Annals of Applied Statistics (this http URL) by the Institute of Mathematical Statistics (this http URL)
Subjects:	**Applications (stat.AP)**
Journal reference:	Annals of Applied Statistics 2013, Vol. 7, No. 1, 418-442
DOI:	10.1214/12-AOAS592
Report number:	IMS-AOAS-AOAS592

Submission history

From: Jun Chen [view email]
[v1] Thu, 23 May 2013 09:20:53 GMT (2062kb)