Regularized Generalized Linear Models to Disclose Host-Microbiome Associations in Colorectal Cancer

Eliana Ibrahimi*
Department of Biology, University of Tirana
eliana.ibrahimi@fshn.edu.al

Mina Norouzirad
Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA SST)
mina.norouzirad@gmail.com

Melisa Meto
Department of Biology, University of Tirana, Albania
melisa.meto@fshnstudent.info

Marta B. Lopes
Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology (NOVA SST); Research and Development Unit for Mechanical and Industrial Engineering (UNIDEMI), NOVA School of Science and Technology (NOVA SST)
marta.lopes@fct.unl.pt

ABSTRACT

Recent studies have shown that gut microbiome is associated with colorectal cancer (CRC) progression and anti-cancer therapy efficacy. This study aims to optimize the ridge, elastic net, and lasso regularized generalized linear models (GLM), widely used for supervised machine learning, for multiclass classification tasks (healthy/adenoma/carcinoma). The models are applied to a benchmark gut microbiome dataset using raw and transformed data. A cross-validation procedure is used to select an optimal value for the shrinkage parameter, λ. The results show a higher accuracy of the ridge and elastic net models compared to the lasso model. We confirm known associations of several microbiome genera with CRC and adenoma. These findings are expected to contribute to the definition of CRC-microbiome signatures to be further validated in microbiome-related therapy studies.

CCS CONCEPTS

• **Mathematics of computing** → Probability and statistics; Statistical paradigms; Regression analysis.

KEYWORDS

Generalized Linear Models, Lasso, Ridge, Elastic net, Gut Microbiome, Colorectal Cancer

ACM Reference Format:
Eliana Ibrahimi, Mina Norouzirad, Melisa Meto, and Marta B. Lopes. 2023. Regularized Generalized Linear Models to Disclose Host-Microbiome Associations in Colorectal Cancer. In *2023 6th International Conference on Mathematics and Statistics (ICoMS 2023), July 14–16, 2023, Leipzig, Germany.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3613347.3613362

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICoMS 2023, July 14–16, 2023, Leipzig, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0018-7/23/07…$15.00
https://doi.org/10.1145/3613347.3613362

1 INTRODUCTION

Colorectal cancer (CRC) is considered one of the most highly spread malignant tumors, and the third in the global mortality rate ranking [1]. The development of colorectal cancer occurs gradually, starting with the appearance of hyperproliferation that leads to the formation of adenomas and, in the most severe cases, leads to degradation to the stage of carcinomas [2]. Recently it has been concluded that one of the main factors that affect the development of colorectal cancer is the gut microbiome, so the interest in its study has increased among researchers. The human intestine is populated by more than 1000 different species of these symbiotic microorganisms, which play a crucial role in maintaining the optimal conditions of the environment they inhabit, as well as serving as neutralizers of pathogens that can enter the body from the outside environment, helping in defense against invaders [3]. The intestinal microbiome contributes to the appearance and development of CRC when it loses the ability to control the homeostasis of the environment, promoting the production of cancer-associated metabolites and the immune response and increasing the synthesis of genotoxic virulence factors [4]. In studies that have been carried out by observing different groups of individuals, it has been noticed that CRC patients showed an abnormal structure of the gut microbiome, compared to healthy individuals who presented a normal structure [5].

Microbiome data analysis is challenging due to some data characteristics such as high dimensionality, sparsity, zero inflation, and compositionality. Appropriate statistical and machine learning methods are required to properly analyze these data and to select the most relevant features that play a role in separating patient groups.

Recent research has demonstrated the effectiveness of regularized regression models to select microbial features that are associated with the development of colorectal cancer and other diseases [6–8]. However, it is important to note that each study has its own limitations, and that further research is needed to validate these

findings and identify potential biomarkers for the early detection and prevention of colorectal cancer.

Our study aims to explore regularized generalized linear models (GLM), widely used for supervised machine learning, for multiclass (healthy/adenoma/carcinoma) classification, and feature selection tasks. The results obtained are expected to provide insight into the role of the gut microbiome in CRC progression and treatment.

2 METHODS

2.1 The data

In this study, we used microbiome 16S data from Zeller et al. [5], which focuses on the association of gut microbiome with CRC. The data considered for analysis consisted of Operational Taxonomic Unit (OTU) counts from 129 samples, from which 38 were adenoma samples, 41 were carcinoma, and 50 were healthy subjects. OTUs, with a relative frequency of counts smaller than 0.0001 among all samples, were not included in the analysis. The relative abundance for each OTU was calculated by dividing the sum of the counts from all samples by the overall sum of counts. This resulted in the final 153 genera, which could then be used for statistical analysis.

2.2 Modeling

In this study, we applied regularized GLMs [9], including the ridge, lasso, and elastic net estimators, for multiclass (healthy/adenoma/carcinoma) classification tasks.

Let us consider a model with a univariate response $Y = (Y_1, \ldots, Y_n)^\top$ and p-dimensional covariates $X = (x_1, \ldots, x_p)$ where for $i \in \{1, \ldots, p\}$, $x_i = (x_{i1}, \ldots, x_{in})^\top$ as $g(E[Y_i|X_i = x]) = \beta_0 + \sum_{i=1}^{p} \beta_j x_j$ where $g(\cdot)$ is a real-valued and known link function, β_0 is an intercept and the covariates x_i are either fixed or random. An implicit assumption of GLM is that $E[Y_i|X_i = x]$ depends on $X_i$only through the function g. That is, the (conditional) probability or density of $Y|X = x$ is of the form $p(y|x) = p_{\beta_0,\beta}(y|x)$.

Since our categorical response variable Y has more than two levels ($K > 2$), we used a multinomial logistic regression, where,

$$P(Y = j|x) = \frac{exp\left\{\beta_{0y} + x^\top \beta_y\right\}}{\sum_{j=1}^{K} exp\left\{\beta_{0j} + x^\top \beta_j\right\}}, \quad j = 1, \ldots, K-1 \quad (1)$$

The regularized estimator for GLMs is constructed by adding a penalty to the (scaled) negative log-likelihood, which is

$$\rho(x, y) = \rho_{\beta_0,\beta_1}(x, y) = -n^{-1} \sum_{i=1}^{n} log(p_{\beta_0,\beta}(Y_i|X_i) \quad (2)$$

The ridge regression [10] then defined as

$$\hat{\beta}_0(\lambda), \hat{\beta}(\lambda) = \underset{\beta,\beta_0}{arg\,min} \left\{ \rho(x, y) + \lambda \sum_{i=1}^{p} \beta_j^2 \right\} \quad (3)$$

where λ is a tuning parameter.

The ridge estimator shrinks the coefficients of correlated predictors equally toward zero but not reaching zero. To identify essential predictors and reduce the number of predictors in a GLM, the lasso estimator [11] can be used. It is defined as

$$\hat{\beta}_0(\lambda), \hat{\beta}(\lambda) = \underset{\beta,\beta_0}{arg\,min} \left\{ \rho(x, y) + \lambda \sum_{i=1}^{p} |\beta_j| \right\} \quad (4)$$

where λ is a tuning parameter. Unlike ridge regression, as the penalty term increases, the lasso estimator is an alternative to the stepwise regression method and can create a smaller model with fewer predictors.

When there are several highly correlated variables, it is recommended to use the elastic net estimator [12], a hybrid of ridge and lasso regularizations. For α strictly between 0 and 1, and nonnegative λ, the elastic net estimator is defined as

$$\hat{\beta}_0(\alpha, \lambda), \hat{\beta}(\alpha, \lambda) = \underset{\beta,\beta_0}{arg\,min} \left\{ \rho(x, y) + \lambda \left(\alpha \sum_{i=1}^{p} |\beta_j| + (1-\alpha) \sum_{i=1}^{p} \beta_j^2 \right) \right\} \quad (5)$$

where λ and α are tuning parameters. This estimator is like lasso when α=1, and when α shrinks towards zero, it approaches ridge regression.

Regularized GLMs were applied to the microbiome 16S dataset [5] The analysis was performed in the R software through the glmnet package [13]. The glmnet function standardizes the predictors and response by normalizing and centering them before the analysis, but the result is back transformed in the original scale. To select an optimal value for the tunning parameter, λ, a 10-fold cross-validation procedure was performed. For each λ, a predictive model is fitted in the training set (60% of the data) and then used to predict the outcome value of each sample in the test set (40%). The model with optimal λ that gives the best accuracy is used for future prediction.

When regularization methods are applied to microbiome data, the random partitioning of the data is of concern. Thus, the impact of data partitioning in cross-validation on the λ selection was further investigated by repeating the cross-validation procedure ten times. In each cross-validation, training and validation sets were built randomly and differently. The analysis is performed on standardized (mean 0 and standard deviation 1) and centered log ratio (CLR) transformed data.

3 RESULTS AND DISCUSSION

The first aim of this work was to estimate the optimum λ value which gives the highest model accuracy or lowest mean squared error (MSE). It is known that different values of λ give different coefficient estimates. The 10-fold cross-validation procedure performed to select an optimal value for the shrinkage parameter λ on standardized and CLR transformed data showed that λ was quite similar for all models (Table 1; Figure 1). Using the cross-validation results, we get the optimum values for λ given in table 1.

Using glmnet, we easily visualized how the estimates for the coefficients vary depending on λ in a sample (Figure 2). As it can be seen, for ridge the number of non-zero coefficients do not change for all the λ values and represents the number of predictors in the model. The lasso approach tends to "shrink" the coefficients to zero as λ increases, clearly shown by the numbers in the upper part of the plot. Going from ridge to lasso, as α increases, the number of non-zero coefficients significantly reduce from 153 to 18 for all three classes.

The coefficients from the ridge and elastic net regression models showed that in both healthy and adenoma groups, the genera which

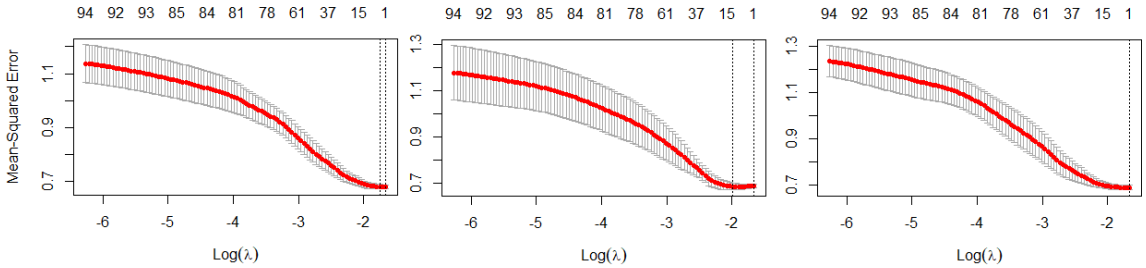

Figure 1: Plots of the cross-validated estimate of MSE as a function of log(λ) value for ridge (left), elastic net (α =0.6) (center), and lasso (right), results from CLR transformed data. The upper part of the graphs gives the number of non-zero coefficients for a given value of log(λ). The gray bars at each point give the MSE ± the standard error for a specific λ. The dashed lines give the position of the smallest MSE.

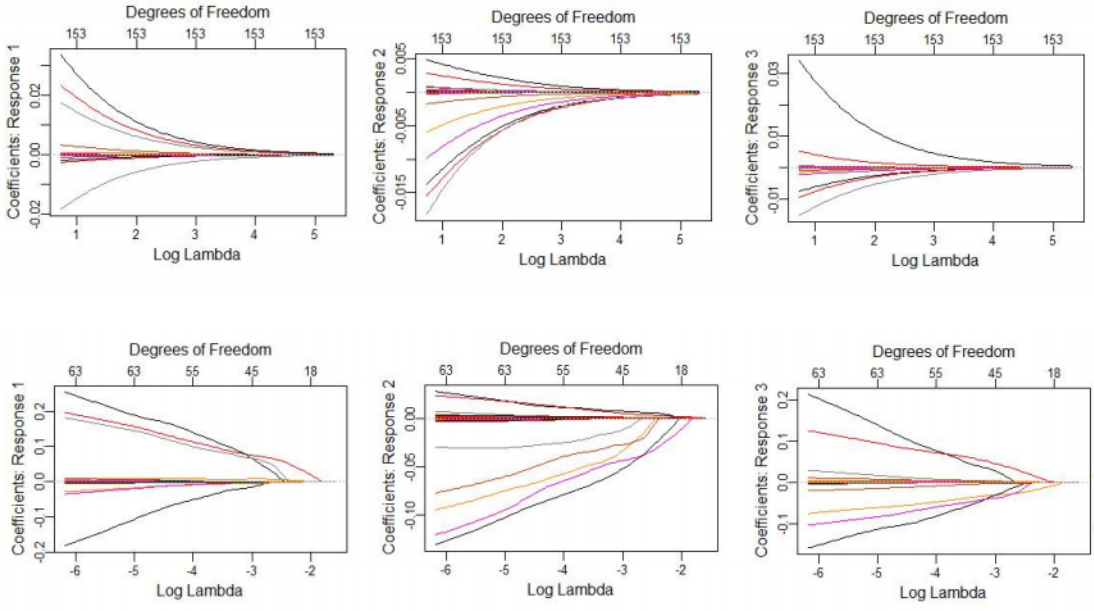

Figure 2: The coefficients for the three categories of the response (response 1=adenoma; response 2=cancer; response 3=healthy) obtained by ridge (above) and LASSO (below) for the Zeller et al., 2014 CLR transformed data, plotted versus log λ. The upper part of the plot gives the number of non-zero coefficients in the model for a given log λ.

most contributed to the model (retaining non-zero coefficients) belong to Actinobacteria and Proteobacteria phyla. In contrast, for the carcinoma group, the genera from Bacteroidetes and Fusobacteria were the most informative to the model. Similar results are reported from other studies [5, 14] where Fusobacteria were associated with the CRC phenotype.

The MSE estimates indicate that ridge outperforms the elastic net and lasso GLMs showing a lower MSE (Table 1). The CLR transformation yielded better results, lower MSE, compared to the normalized transformation of the data. As α increases the MSE tends to increase, showing that when more coefficients estimates are put to zero the accuracy of the model decreases for the dataset in use.

To visualize the performance of the models, we plotted the ROC curves for all classes (adenoma, carcinoma, healthy) using the ROCR R package (Figure 3). From these plots we can see that the ridge and the elastic net models have a better performance compared to the lasso model, especially for the adenoma class.

Our results are in line with other studies which have reported that ridge regression may perform slightly better to lasso regularization in microbiome data modelling [15, 16] in some scenarios.

However, many studies that applied regularized regression models in predicting microbial taxa associated with colorectal cancer, reported a great performance with very high accuracy for lasso [17–19]. These findings show that the choice between lasso and ridge regression in modeling microbiome data may depend on the specific research question and dataset characteristics. In our case

Table 1: Summary results for the ridge, elastic net, and lasso GLMs performed on normalized and CLR transformed data (α and λ are the parameters that control the mixing between the ridge and the lasso, and the severity of the penalty, respectively).

		Normalized data		CLR transformed data	
Regularized	α	λ optimum	MSE	λ optimum	MSE
Ridge	0	0.059	0.519	0.173	0.461
Elastic net	0.2	0.094	0.584	0.190	0.538
	0.4	0.036	0.635	0.151	0.537
	0.6	0.092	0.615	0.165	0.531
	0.8	0.062	0.653	0.191	0.564
Lasso	1	0.049	0.654	0.190	0.563

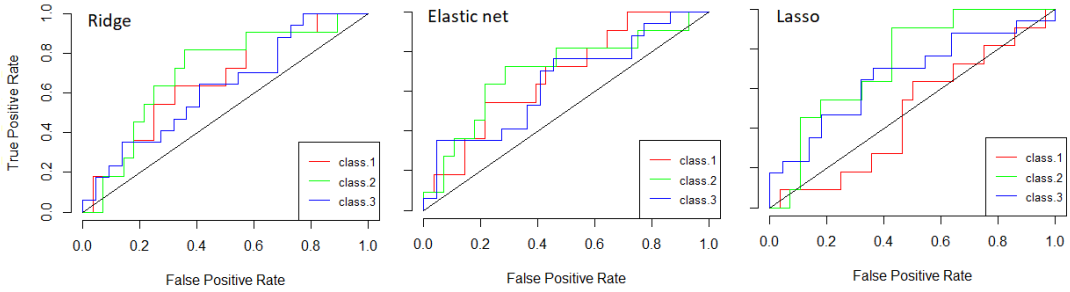

Figure 3: Plot of the ROC curves for all classes (class 1=adenoma; class 2=carcinoma; class3=healthy) for ridge, elastic net (α =0.6), and Lasso.

the lasso penalty might be too stringent in this dataset leading to a very sparse solution and decreased accuracy. Ridge regression can outperform lasso in classification accuracy but maybe not in identifying features, because all are used in the end for the classification, and comparing features' importance between models is not straightforward.

Our future work will focus on validating the models in a larger dataset where data from other studies and populations are included. We also aim to add the coda-lasso approach [20] in our analysis to count for the compositional nature of microbiome data correctly.

4 CONCLUSIONS

Our findings suggest that ridge and elastic net models can outperform lasso when applied to 16S microbiome data. Despite that, we recommend that in the search for an accurate sparse and interpretable solution researchers consider factors such as the number of predictors, the sample size, and the desired level of sparsity in the final model when selecting a regularization method.

ACKNOWLEDGMENTS

We are grateful to ML4Microbiome CA18131 COST Action 'Statistical and machine learning methods for microbiome data' for supporting this study with short scientific mission grant. We also thank Fundação para a Ciência e a Tecnologia (FCT) with references CEECINST/00102/2018, UIDB/00297/2020 and UIDB/00297/2020 (NOVA MATH) and UIDP/04516/2020 (NOVA LINCS).

REFERENCES

[1] W. Chen, F. Liu, Z. Ling, X. Tong, C. Xiang, Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer, PLoS One. 7 (2012) e39743. https://doi.org/10.1371/journal.pone.0039743.

[2] J. Kim, H.K. Lee, Potential Role of the Gut Microbiome In Colorectal Cancer Progression, Frontiers in Immunology. 12 (2022). https://www.frontiersin.org/articles/10.3389/fimmu.2021.807648 (accessed August 5, 2023).

[3] S. Temraz, F. Nassar, R. Nasr, M. Charafeddine, D. Mukherji, A. Shamseddine, Gut Microbiome: A Promising Biomarker for Immunotherapy in Colorectal Cancer, Int J Mol Sci. 20 (2019) 4155. https://doi.org/10.3390/ijms20174155.

[4] K. Wieczorska, M. Stolarek, R. Stec, The Role of the Gut Microbiome in Colorectal Cancer: Where Are We? Where Are We Going?, Clin Colorectal Cancer. 19 (2020) 5–12. https://doi.org/10.1016/j.clcc.2019.07.006.

[5] G. Zeller, J. Tap, A.Y. Voigt, S. Sunagawa, J.R. Kultima, P.I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D.R. Mende, M.A. Schneider, P. Schrotz-King, C. Tournigand, J. Tran Van Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor, C.M. Ulrich, M. von Knebel Doeberitz, I. Sobhani, P. Bork, Potential of fecal microbiota for early-stage detection of colorectal cancer, Molecular Systems Biology. 10 (2014) 766. https://doi.org/10.15252/msb.20145645.

[6] M. Dong, L. Li, M. Chen, A. Kusalik, W. Xu, Predictive analysis methods for human microbiome data with application to Parkinson's disease, PLOS ONE. 15 (2020) e0237779. https://doi.org/10.1371/journal.pone.0237779.

[7] D. Knights, E.K. Costello, R. Knight, Supervised classification of human microbiota, FEMS Microbiol Rev. 35 (2011) 343–359. https://doi.org/10.1111/j.1574-6976.2010.00251.x.

[8] L. Berbert, A. Santos, D.O. Magro, D. Guadagnini, H.B. Assalin, L.H. Lourenço, C. a. R. Martinez, M.J.A. Saad, C.S.R. Coy, Metagenomics analysis reveals universal signatures of the intestinal microbiota in colorectal cancer, regardless of regional differences, Braz J Med Biol Res. 55 (2022) e11832. https://doi.org/10.1590/1414-431X2022e11832.

[9] McCullagh, P, Nelder, J. A, Generalized Linear Models | P. McCullagh | Taylor & Francis eBooks, Re, (n.d.). https://www.taylorfrancis.com/books/mono/10.1201/9780203753736/generalized-linear-models-mccullagh (accessed August 5, 2023).

[10] A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics. 12 (1970) 55–67. https://doi.org/10.2307/1267351.
[11] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological). 58 (1996) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[12] H. Zou, T. Hastie, Regularization and Variable Selection Via the Elastic Net, Journal of the Royal Statistical Society Series B: Statistical Methodology. 67 (2005) 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
[13] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, J Stat Softw. 33 (2010) 1–22.
[14] T. Tahara, E. Yamamoto, H. Suzuki, R. Maruyama, W. Chung, J. Garriga, J. Jelinek, H. Yamano, T. Sugai, B. An, I. Shureiqi, M. Toyota, Y. Kondo, M.R.H. Estécio, J.-P.J. Issa, Fusobacterium in colonic flora and molecular features of colorectal carcinoma, Cancer Res. 74 (2014) 1311–1318. https://doi.org/10.1158/0008-5472.CAN-13-1865.
[15] D. Sharma, W. Xu, phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data, Bioinformatics. 37 (2021) 3707–3714. https://doi.org/10.1093/bioinformatics/btab482.
[16] K. Shestopaloff, M. Dong, F. Gao, W. Xu, DCMD: Distance-based classification using mixture distributions on microbiome data, PLOS Computational Biology. 17 (2021) e1008799. https://doi.org/10.1371/journal.pcbi.1008799.
[17] O. Queen, S.J. Emrich, LASSO-based feature selection for improved microbial and microbiome classification, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021: pp. 2301–2308. https://doi.org/10.1109/BIBM52615.2021.9669485.
[18] Y.-H. Zhou, G. Sun, Improve the Colorectal Cancer Diagnosis Using Gut Microbiome Data, Frontiers in Molecular Biosciences. 9 (2022). https://www.frontiersin.org/articles/10.3389/fmolb.2022.921945 (accessed August 5, 2023).
[19] S. Bosch, A. Acharjee, M.N. Quraishi, I.V. Bijnsdorp, P. Rojas, A. Bakkali, E.E. Jansen, P. Stokkers, J. Kuijvenhoven, T.V. Pham, A.D. Beggs, C.R. Jimenez, E.A. Struys, G.V. Gkoutos, T.G. de Meij, N.K. de Boer, Integration of stool microbiota, proteome and amino acid profiles to discriminate patients with adenomas and colorectal cancer, Gut Microbes. 14 (n.d.) 2139979. https://doi.org/10.1080/19490976.2022.2139979.
[20] A. Susin, Y. Wang, K.-A. Lê Cao, M.L. Calle, Variable selection in microbiome compositional data analysis, NAR Genom Bioinform. 2 (2020) lqaa029. https://doi.org/10.1093/nargab/lqaa029.