Understanding the Interplay Between Gut Microbiome and Immune System

Using Elastic Net to Predict Immune Phenotypes In the Gut Microbiome

AUTHOR
M. Knigge

INSTITUE
Hanze University
Life Science and Technology

ORGANIZATION
Eriba
Department of Genetics

DATE
10, Juli 2018

Understanding the Interplay Between Gut Microbiome and Immune System

Using Elastic Net to Predict Immune Phenotypes In the Gut Microbiome

Matthijs Knigge[1], **Serena Sanna**[2], **Yang Li**[2]

1. Hanze University of Groningen, Institute for Life Science & Technology, Bioinformatics, University of Medical Center Groningen, Department of Genetics, Groningen, The Netherlands.

2. University of Groningen, University of Medical Center Groningen, Department of Genetics, Groningen, The Netherlands.

Hanze University of Applied Sciences Groningen

umcg

AUTHOR
M. Knigge

SUPERVISORS
Asst. Prof. Dr. S. Sanna
Asst. Prof. Dr. Y. Li

LECTURER
Dr. M. A. Noback

INSTITUE
Hanze University
Life Science and Technology

ORGANIZATION
Eriba
Department of Genetics

DATE
10, Juli 2018

ABSTRACT

ABBREVIATIONS

GI	Gastrointestinal tract
MWAS	Microbiome wide-association study
IBD	Inflammatory Bowel Disease
BMI	Body Mass Index
500FG	500 Functional Genomes
LL-Deep	LifeLines-Deep
MR	Mendelian Randomization
Ag	Antigen
OLS	Ordinary Least Squares
TSS	Total Sum of Squares

TABLE OF CONTENTS

INTRODUCTION

Evidence suggests that the human gut microbiome plays an important role in metabolic and immune function.[1] Empirical evidence of Microbiome-Wide Association Studies (MWAS) in population cohorts show significant associations identified systematically between the gut microbiome and multiple complex traits and intrinsic factors, like Inflammatory Bowel syndrome (IBD), body mass index (BMI), food allergies and blood cells composition.[2] Discovery of these links is key to understanding disease, and potentially disease treatment.

The body of a human is a complex interconnected ecosystem, and the gut is where the body acts as a first line of defense. Where it interacts with the "outside world", functioning as a frontline of the immune system, which is constantly exposed to new microbes and molecules.[3] The whole collection of microbes and molecules that are present in and on the human body is known as the microbiota.[4] The microbiome refers to the whole set of genes within these microbes. The role of the microbiome composition/function is considered as an acting organ in the body's operation. It is presumed that it has an effect on aging, digestion, mood, cognitive function, and immune system.[5] The immune is a defensive system from the host entailing many biological structures and processes within an organism to protect against diseases, and infection. The function of the immune system relies on the ability to detect and distinguish a wide arrange of agents known as pathogens, viruses, and parasites from self and non-self.[6]

The aim of this research is the identification of causality links between the microbiome composition/function and immune system: does the microbiome influence the immune system (cell counts, cytokines, globulin levels), or/and does the immune system influence the microbiome. We have access to the largest population cohort with gut metagenomic sequencing (LL-DEEP), but this cohort has not been characterized in depth for immune traits.And we have access to another cohort, the human function genome project 500FG[7] which was specifically designed to assess the immune and metabolic system (422 phenotypes available), and also microbiome, but it is very small. This cohort has also genetic and gene expression available. By using gene expression data, genetic data, and transcriptomic data from the 500 Functional Genomic (500FG) cohort a linear model is constructed that explains immune traits/functions between gene expression data, genetic data, genetic data combined gene expression data and the 500FG cohort with Elastic Net Regularization.[8] This constructed model, that is based on genetic data and gene expression data is used to predict immune traits/functions in LL-Deep data which contains genetic, and gene expression data from a large number of individuals which lacks immunogenic information. These predicted immune phenotypes will be used to associate a link with microbial composition/function, see figure 1.

We expect to find links between microbiome composition/function and the immune system that can help us understand the interplay between the gut microbiome and the immune system, and ultimately help understand disease, and develop disease treatment; restoring microbiome composition and/or function through personalized nutrition or treatments. .

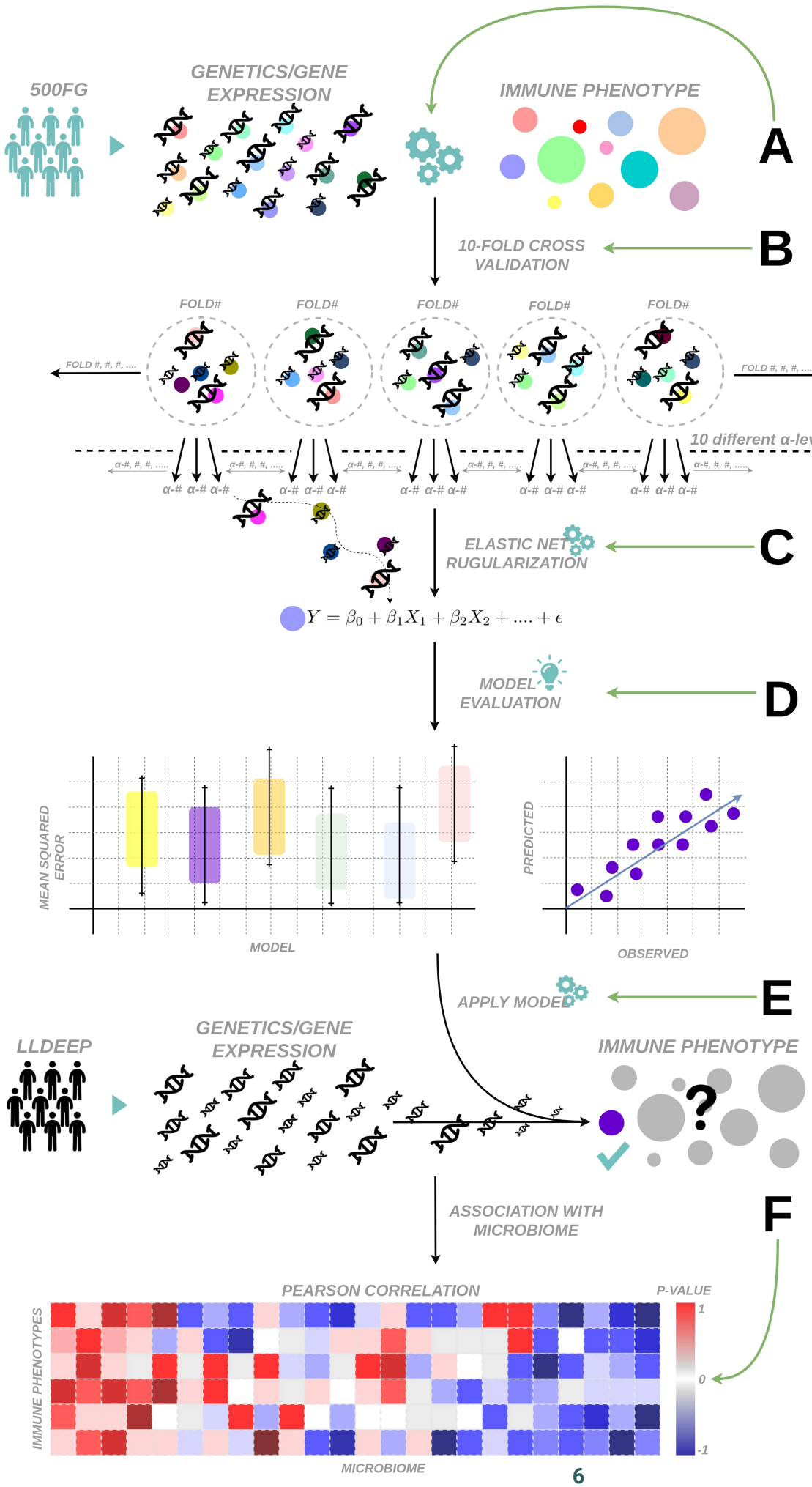

Figure 1: Schematic overview of the research project. As described, the goal is to build a model that can predict immune phenotypes. The data used (**A**) for the this model is gene expression and genetics, and measured immune phenotypes in the 500FG cohort. This data will be processed and splitted randomly in 10 folds (**B**), which is shown by the different colors. For every fold, this data will enter the elastic net function ten times with different alpha levels (**C**), so in total this method will be applied 100 times in order to construct a model that can explain a certain immune phenotype. After the model building process, it has to be evaluated which combination of data has the best prediction power (**D**): genetics, gene expression, or genetics combined with gene expression. When the an appropriate model is chosen, it will be used to predict immune phenotypes in LLDeep (**E**) with gene expression, genetics, or genetics combined with gene expression data. This will produce immune phenotypes that where not available in this cohort before, and can be used to forecast links between the immune system and microbiome (**F**). From that point onward it is possible to investigate possible links that can help understand the interplay between the immune system and the microbiome which hopefully can help understand diseases, disease manifestation, or help develop disease treatment which can restore microbiome composition or function.

1. THEORY

1.1 GUT MICROBIOME

The gut microbiome is an acting barrier against harmful microbes in terms of competition for nutrients, and production of antimicrobial substances. Multiple antimicrobial compounds, like defensins[9], cathelicidins[10], and C-type lectins[11], which are produced in the Gastrointestinal tract (GI tract)[12]. The presence of bacteria or their structural components, and the presence of the products of the metabolism of bacteria has the potential to induce the expression and activation of antimicrobial substances which in turn contribute to the host protection against invading pathogens.[13]

The gut microbiome is consider a key player in human health, disease and is involved in host metabolism and immune function. It supplies an important role in genetic and metabolic role to the human body as a "super organism" in the form of gene functions encoded by the metagenome which out-sums human genes by a multiplication of 100.[14] Empirical evidence from a broad range of human studies and model systems suggests that the gut microbiome has a critical functional within this human "super-organism," it regulates both metabolic and inflammatory processes which do not only mediate chronic disease (metabolic syndrome, diabetes, cardiovascular disease) but also autoimmune diseases, dementia, and it is speculated that it influences the aging processes. Diet plays a key role in shaping the gut microbiome and also in shaping the ability to regulate host metabolism and immune function through production of metabolites.[15]

1.2 IMMUNE SYSTEM

The immune system function is to distinguish non-self from self to eliminate harmful non-self cells and molecules from the system. Beside self, non-self recognition, the immune system has the ability to recognize destroy abnormal cells that are derived from host tissue. Any molecule that can be recognized by the immune system is considered to be an antigen (Ag).[16]

This immune system has evolved to defend the host against an universe of pathogenic microbes that are constantly evolving. The immune system helps the host with removal of toxic or allergenic compounds that enter the host through mucosal surfaces. Primarily, the immune system's ability is to trigger a response to an invading pathogen, toxic, or allergen. The immune system uses two mechanisms to detect self from non-self: innate and adaptive mechanisms. Both these system rely on discrimination between self and non-self to eliminate pathogenic microbes.[17, 18]

The mechanisms contributing to recognition of microbial, toxic, or allergenic molecules can be divided into two categories: hard-wired responses that are encoded in the host's germline that recognize molecular patterns shared by microbes and toxins that are not in the body, and responses that are encoded by elements that somatically rearrange to assemble into antigen-binding molecules with specificity for individual unique foreign structures.[19]

The first response belongs to the innate immune response. Because the recognition molecules from the innate system are expressed on a large number of cells, this system acts rapidly after an invading pathogen or toxin is encountered. The second response belongs to the adaptive immune response. The adaptive system consists of small numbers of

cells with specificity for any individual pathogen, toxin, or allergen. These responding cells must proliferate after recognition of an antigen in order to acquire sufficient numbers to trigger an effective response against the microbe or the toxin. The adaptive response generally expresses itself temporally after the innate response. A key feature of the adaptive system is that it produces long-lived cells that persist in an dormant state, that can be re-express their effector functions rapidly after another encounter with a specific antigen. This provides the adaptive response with the ability to manifest immune memory.[20, 21, 22]

1.3 LINEAR REGRESSION

Regression analysis is the science of fitting straight lines to patterns of data, it is used for prediction and forcasting. In a linear regression model, the variable of interest is the dependent variable, which is predicted from a single or more variables, the independent variables, using a linear formula. Regression analysis is also used to found out which among the independent variables are related to the dependent variables, and also to explore the forms of these relations. The simple regression model can be represented as follows (equation 1):

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad (1)$$

Where β_0 is the Y intercept value, the coefficient β_1 is the slope of the line, the X_1 is an independent variable, and ϵ is the error term. The error term is used for correcting a prediction error between observed and the predicted value.[23, 24]

1.4 ORDINARY LEAST SQUARES

Ordinary Least Squares (OLS) is a method that is used for estimating/predicting parameters that are unknown in a linear regression model. OLS chooses the parameters of a linear function in a set of independent variables by minimizing the total sum of squares (TSS) of the differences between the observed independent variable in the given data set and those predicted by the linear function. This can be seen as a square drawn between the predicted and observed, see figure 2.

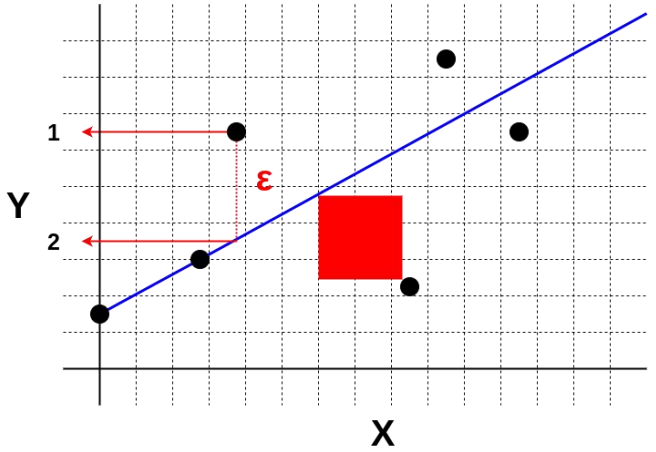

Figure 2: Visually overview of the OLS method. There is a square drawn between the observed value of Y for X (1), and the predicted value of Y for X (2). The total sum of squares represents the error in an OLS regression model. The smaller the differences (square size), the better the model can be fitted to the data.

The TSS is a representation of the error in the OLS regression model. In linear regression models, the error can be composed into two components: (1) error due to bias, and (2) error due to variance. The goal of OLS is to minimize the difference between the observed and the predicted with a linear approximation.[25, 26]

1.5 BIAS & VARIANCE

Understanding the error due to bias and the error due to variance is key to understanding the behavior of prediction models. It is important to understand how different sources of error can lead to error in bias and variance, this can help improve the data fitting process which results in more accurate models. The error due to bias is the difference between the expected prediction of the model and the correct (true) value which is the model tries to predict. The error due to variance is the variability of a model prediction for a given data

point. The model building process can be repeated multiple times, and this can be visualized as the variation for a prediction of a data point between different realizations of the model (figure 3).

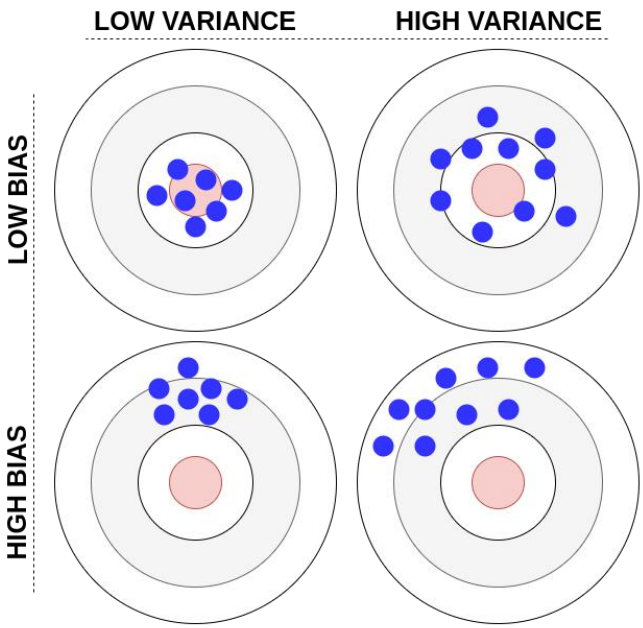

Figure 3: Error due to bias is the difference between the predicted value and the observed value. Error due to variance is the difference between points given for different realizations of the model. When there is a low bias and variance, the is little different between predicted and the correct value, and little variability between data points between different realizations of the model

In figure 3, these targets show the relation between variance and bias. If there is a model that is very accurate which means that the errors will be very low, it is expected that the model will have low bias and low variance. When there is low bias, it means that there is little difference between the predicted value and the observed value, and when there is low variance, there is little variability between the points in different realizations of the model. If there is low variance, but high bias there will be little variation between data points for different realizations of the model, but difference between the observed and predicted. Another example for a model with low bias and high variance, this can visually been seen as a line that can very accurate been drawn through a data pattern, but will vary a lot between different realizations of this model.[27, 28, 29]

1.6 GUASS MARKOV THEOREM

There is trade-off between the ability of a model to minimize bias and variance. For any model, the best spot is at a point where increase in bias is the same as the reduction in variance. This means that where bias is reduced, the variance increases, see figure 4. As shown, when the complexity of a model increases, by the addition of terms to the linear model, the bias will reduce and the variance will increase.[30, 31]

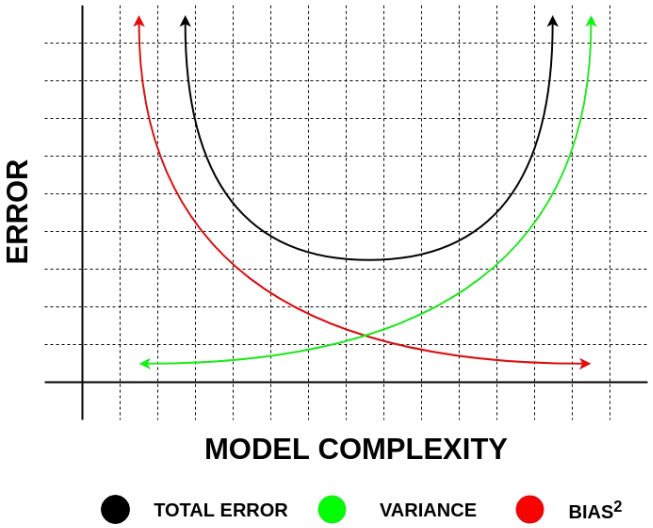

Figure 4: The trade-off between variance and bias. The best situation for any model is level of complexity where the increase of bias is equal to the reduction of variance. This figure shows that when the complexity of model rises by the addition of terms to the model that the bias will decrease and that the variance will increase.

The Gauss Markov theorem states that from all linear regression model estimates, that the OLS regression has the smallest variance. This means that there can be a model that is biases but has a smaller Mean Squared Error (MSE), in simple terms there can be added bias to the model which will lower the variance which in turn will give a better MSE. The adding of bias is done with shrinkage estimators.[32, 33, 34]

1.7 SHRINKAGE ESTIMATORS

A shrinkage estimator is an estimator that incorporates the effects of shrinkage. This means that a raw estimate is improved by the incorporation of other information. This incorporation of information creates a new estimate that is made closer to a supplied value.[35]

As stated by the Gauss Markov Theorem, from all unbiased estimators OLS has regression has the smallest variance. Which implies that there can be a better model where bias is increased and variance is lowered, thus improving the MSE. This can be done by adding information with shrinkage estimators to shrink the estimators towards zero (or some other fixed constant). For example, in a model where there is a estimate that is nonzero, an other estimate can be obtained by multiplying the raw estimate by a predefined parameter in order to lower the MSE. The effect here is that an unbiased raw estimate is converted into a biased estimate with a lower MSE.[36]

The shrinkage estimators replace the OLS estimates with (equation 2):

$$\beta'_k = \frac{1}{1+\lambda}\beta_k \quad (2)$$

Where β'_k is the new estimate, β_k is the raw estimate, and λ is the lambda parameter. λ is denoted as the shrinkage estimator (ridge constant). What this shows, is that if the lambda parameter is zero, the raw estimate will return, and if lambda gets large the raw estimate will approach a minimal value of zero, see figure 5. This figure explains that when the lambda parameter increases information is added to the estimate and increases bias and decreases variance, which produces a better net MSE.[37]

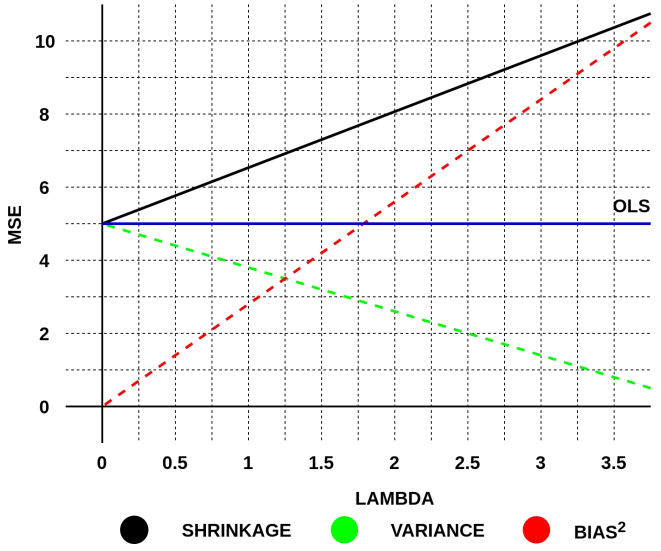

Figure 5: The lambda parameter shows that when it is zero that it will return the raw estimate, but when lambda increases a new estimate will be produced. If λ increases the bias increases and the variance decreases resulting in a lower MSE.

1.8 RIDGE REGRESSION

Ridge regression is modeling technique that tries to undermine the multicollinearity problem in OLS models by the incorporation of the λ parameter.[38] Multicollinearity is the phenomenon where one predictor variable in multiple regression analysis can be predicted nearly perfect accurately from other variables. This collinearity is a linear relationship between two variables. Two variables are perfectly collinear if there is an exact linear association. Multicollinearity is a being of having very high intercorrelations among the independent variables, this is a type of disturbance in the data that make the statistical inferences not reliable. And this can create the following problems[39, 40]:

- the regression coefficient can not be computed exactly
- the sign and magnitude of the coefficient can change from sample to sample

Thus, Ridge regression is a method which can create a model when there are more predictors than

observations, or when the data suffers from multicollinearity. In comparison to OLS, this method can not differentiate between predictors that are "important" or "less-important", so it includes all predictors. This can lead to over fitting, and OLS can not deal with multicollinearity. It is true that OLS produces unbiased estimates, but it can create variance so large that estimates can be inaccurate. Ridge regression adds bias to the estimates to lower the variance which makes the estimates more reliable. The Ridge regression uses shrinkage estimators to produce new estimators that are shrunken towards the "truth". The Ridge regression is especially good in improving OLS estimates when multicollinearity is present in the data.[41, 42]

Ridge regressions is a method that uses a tool called ℓ_2 regularization. An other form of regularization is the ℓ_1 type. This method adjust the size of the coefficients by a penalty equal to the absolute value of the magnitude of the coefficients. This ℓ_1 penalty can result in the elimination of coefficients. Whereas the ℓ_2 regularization adds a penalty to the coefficients that is equal to the square of the magnitude of the coefficients, which thus shrinks all coefficients by the same, therefor not eliminating any coefficient. In this regularization the λ parameter is used to control the strength of the penalty term, if the λ is set equal to zero, the Ridge regression is equal to OLS, and if this term is set to ∞ all coefficients will be set to zero.[43]

1.9 REGULZARIZATION

Regularization is method for avoiding overfitting of models by penalizing regression coefficients, this method reduces parameters and shrinks models. These simplified models will perform better at predictions. This regularization is necessary because the residual sum of squares (RSS) in OLS can be unstable, for sure when multicollinearity is present. The RSS a way of assessing the quality of a fit. It measurement of the overall difference between the data and the predicted data (residual is the distance from a data point to the regression line).[44] The science of fitting a model through a data pattern comes with pitfalls: any model can be fitted to a data set, see figure 6. In this example are seven data points, and there are five models fitted through this data pattern. The simplest model is a linear model through all data points, and multiple n-degree polynomial models, which shows that any model could fit this.

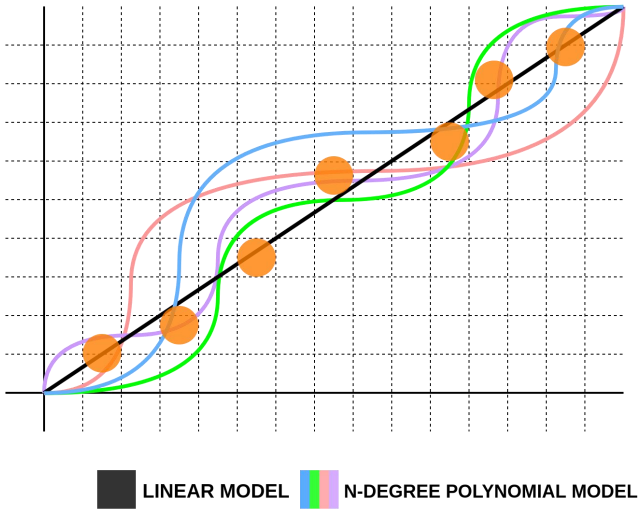

Figure 6: An example of multiple models fitted over seven data points. The black line shows a linear model, and the other ones are n-degree polynomial models.

With a small set of data is easy to create a model that is too complex and is overfitted, or a model that is too simple and may underfit. Regularization penalizes models that are to complex in favor of simpler models. The penalty terms used are ℓ_1, and ℓ_2.[45]

1.10 LEAST ABSOLUTE SELECTION AND SHRINKAGE OPERATOR

Least Absolute Selection and Shrinkage Operator (LASSO) is a regression method that can perform

both variable selection (Ridge does not) and regularization in order to yield better prediction. This type of regression analysis can perform very well on data data that is multicollinear. The LASSO methods the ℓ_1 regularization, this adds a penalty that is equal to the absolute value of the magnitude of the coefficients. This type of regularization can create models with few coefficients because coefficients can be set to zero, which is variable selection. The LASSO method only uses information that is useful for the prediction. Like Ridge regression, LASSO also uses the λ parameter to control the strength of the penalty (ℓ_1). The λ parameter is basically the amount of shrinkage. For LASSO, when λ is zero no parameters will be removed from the model. When the λ parameter increases more and more coefficients will be set to zero and removed from the model. If λ is set to ∞ all coefficients are removed. If λ increases, bias will become larger, while λ decreases the variance will increase.[46]

1.11 ELASTIC NET REGULARIZATION

Elastic net is a regularization method is a hybrid form based on a combined penalty of the ℓ_1 and ℓ_2 penalties from LASSO and Ridge regression. The penalty parameter for Elastic net (α) determines how much weight should be given the LASSO or Ridge regression. When the α parameter is set to 0, it performs like Ridge regression, and when it is set to 1 it perform in a same matter like the LASSO method.[47] In summary. The objective function of the Elastic net penalty looks like (equation 3)[48]:

$$RSS + \alpha\,\ell_2 + (1-\alpha)\,\ell_1 \tag{3}$$

Here, RSS is the residual sum of squares from the OLS, α is Elastic net penalty parameter, ℓ_2 is the Ridge penalty parameter, and ℓ_1 is the LASSO penalty parameter.

1.12 K-FOLD CROSS VALIDATION

K-fold cross validation is method for validating models which can access how the results will generalize to an independent data set. It is mostly used to estimate how accurately a prediction model is. In a prediction setting a model is given a dataset of known data that is used for "training" on which the model is trained, and a "known" dataset against which the model is tested. The goal of cross validation is to have a test dataset to test the model in order to overcome overfitting, and to give information about how well the model can be generalized to an independent dataset.[49]

A iteration of cross-validation entails the partitioning of a sample data set into subsets of data, then the analysis will be trained on one subset, and validated on the other subsets, see figure 7. To reduce the variability in this process, multiple iterations of cross validation are needed to be performed using different partitions, and then the resulting validations will be combined from all iterations to estimate the final predictive model.

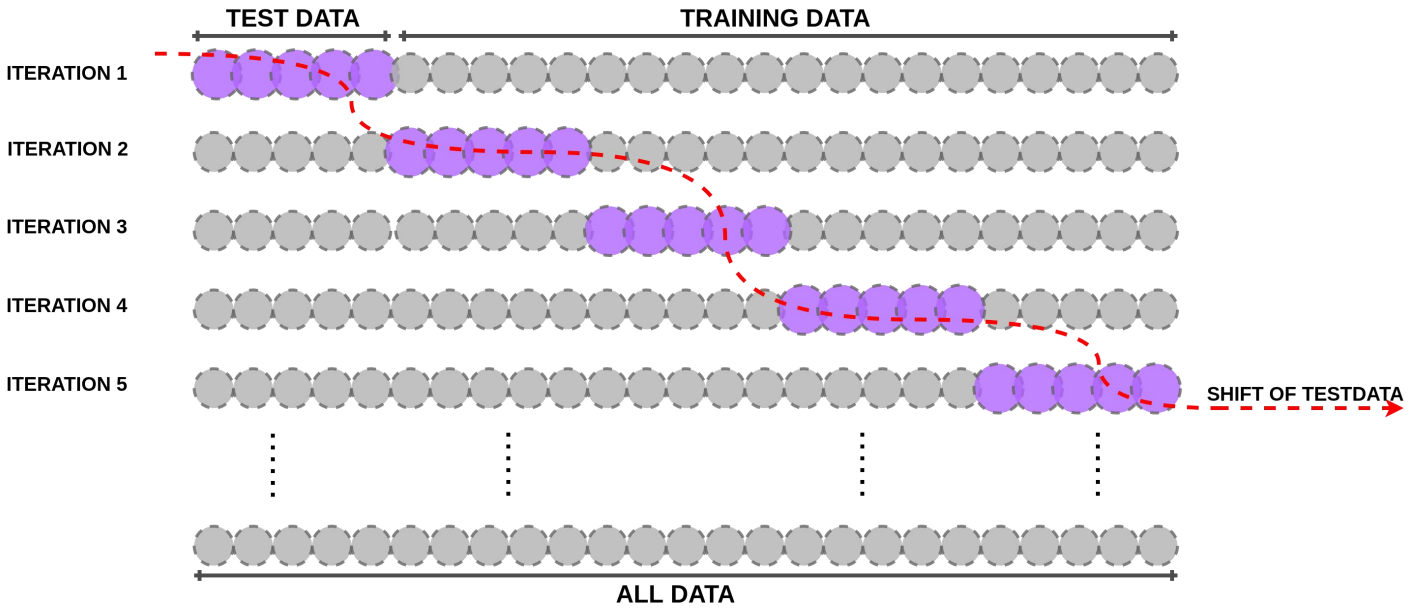

Figure 7: A visual overview of k-fold cross validation. This example has five iterations where in every iteration the data is subsetted in a test set where the model is validated and a subset where the model is trained.

2. MATERIAL & METHODS

2.1 DATA

One of the important goals of this research is to impute immunogenic phenotypes in the in order to infer associations between the immune system and gut microbiome. To impute immune phenotypes several components are needed in this study, a cohort with genetic, and or gene expression data, and measured immunogenic information, and in the cohort that lacks immunogenic information genetic and or gene expression data is needed. The data cohort used for this are 500FG and LLDeep.

2.1.1 500 FUNCTIONAL GENOMICS

The 500-Human Functional Genomics (500FG) cohort is part of the Human Function Genomics Project (HFGP) which has the aim to understand the factors that determine the variability in immune response. 500FG has the focus on gaining a better understanding of variability in the human cytokine responses. This project collected from 500 induviduals blood samples and isolated DNA which was hybridized on the HumanCoreExome SNP chip to gain genotype information on eight million SNPs. Besides the genotyping, blood also has been used to perform a stimulation experiment on with major human pathogens to profile cytokine response. Besides cytokines, in 500FG in peripheral blood cell counts have been measured, immunoglobulin levels, and platelet activation profiles where measured using whole blood flow cytometery. Also an integrative genomics analysis has been performed by using RNA-sequencing from blood samples to collect gene expression.

Thus, the 500FG cohort offers genotype information, gene expression, and measured cytokines stimulated on different time-points by major pathogens, and baseline molecular phenotypes like cell counts, platelets, and immunoglobulin levels.[50, 51]

2.1.2 LIFELINESDEEP

LifeLines s a population cohort of more than 165.000 participants that covers multiple generations of participating families and its main focus is on the determination of multifactorial diseases or traits. The LifeLines cohort has detailed information about phenotypic and environmental factors, and health status. For about ten percent of this cohort genetic information is available.[52, 53]

A subset from this data cohort is also part LifeLines-Deep (LLDeep) which consist of 1500 participants. The participants in this cohort are examined more thoroughly, especially with respect for molecular data, which constitutes to a more in-depth research perspective for associating genetic and phenotypic variants. In this cohort, additional biological information is collected. Like genome-wide transcriptomics, methylation patterns, metabolites, biomarkers and the gut microbiome is assessed.[54]

2.2 COMPUTATIONAL INFRASTRUCTURE

To ensure reproducibility and efficiency of the analyses, all steps are implemented in a software package for the R open-source environment.[55] The package, named

"*Interplay.Between.Gut.Microbiome.and.Immune.System*", performs elastic net regularization using genetic data, gene expression, and measured immunogenic information. Including data pre-processing and post-results sensitivity analyses as described in the following paragraphs. The "*Interplay.Between.Gut.Microbiome.and.Immune.System*"

package can be downloaded from the web-based hosting service Bitbucket:

https://bitbucket.org/MatthijsKnigge/Interplay.Between.Gut.Microbiome.and.Immune.System/

Or can be directly installed into R by using libraries that are described in the tutorial on the source page. This package provides the user with several options to perform and replicate the same or a similar study that has been conducted in this research. The core functionality that is implemented is the elastic net regularization method. However, there is also the functionality to pre-process the data, visualization, and other functionalities which are discussed in detail on the source page of the package

RESULTS

CONCLUSION

DISCUSSION

REFERENCES

[1] Zhernakova, A., Kurilshikov, A., Bonder, M., Tigchelaar, E., Schirmer, M., & Vatanen, T. et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science, 352(6285), 565-569. doi:10.1126/science.aad3369

[2] Schirmer, M., Smeekens, S., Vlamakis, H., Jaeger, M., Oosting, M., & Franzosa, E. et al. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell, 167(4), 1125-1136.e8. doi:10.1016/j.cell.2016.10.020

[3] Andrew L. Kau, Jeffrey I. Gordon. 2018. "Human Nutrition, The Gut Microbiome, And Immune System: Envisioning The Future". Nature 474 (7351): 327. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298082/.

[4] Ursell, L., Metcalf, J., Parfrey, L., & Knight, R. (2012). Defining the human microbiome. Nutrition Reviews, 70(suppl_1), S38-S44. Retrieved from https://academic.oup.com/nutritionreviews/article-abstract/70/suppl_1/S38/1921538?redirectedFrom=fulltext

[5] Josef Neu, J. (2011). Cesarean versus Vaginal Delivery: Long term infant outcomes and the Hygiene Hypothesis. Clinics In Perinatology, 38(2), 321. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3110651/

[6] Massimo Mangino, T. (2017). Innate and adaptive immune traits are differentially affected by genetic and environmental factors. Nature Communications, 8. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5227062/

[7] 500 Functional Genomics Project. (2018). Human Functional Genomics Project. Retrieved 19 April 2018, from http://www.humanfunctionalgenomics.org/site/?page_id=82

[8] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal Of The Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x

[9] Definition of Defensin. (2018). MedicineNet. Retrieved 20 April 2018, from https://www.medicinenet.com/script/main/art.asp?articlekey=26300

[10] cathelicidin. (2018). TheFreeDictionary.com. Retrieved 20 April 2018, from https://medical-dictionary.thefreedictionary.com/cathelicidin

[11] Cummings, R., & McEver, R. (2009). C-type Lectins. Cold Spring Harbor Laboratory Press. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK1943/

[12] Gastrointestinal Tract - National Library of Medicine - PubMed Health. (2018). PubMed Health. Retrieved 27 April 2018, from https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0022855/

[13] Kelly, D., Yang, L., & Pei, Z. (2017). A Review of the Oesophageal Microbiome in Health and Disease. Methods In Microbiology, 19-35. doi:10.1016/bs.mim.2017.08.001

[14] Gut Microbiota–Immune System Crosstalk: Implications for Metabolic Disease - Diet-Microbe Interactions in the Gut - Chapter 9 . (2018). Sciencedirect.com. Retrieved 20 April 2018, from https://www.sciencedirect.com/science/article/pii/B9780124078253000095

[15] The Microbiota in Gastrointestinal Pathophysiology - ScienceDirect . (2018). Sciencedirect.com. Retrieved 20 April 2018, from https://www.sciencedirect.com/science/book/9780128040249

[16] Overview of the Immune System - Immunology; Allergic Disorders - MSD Manual Professional Edition. (2018). MSD Manual Professional Edition. Retrieved 27 April 2018, from https://www.msdmanuals.com/professional/immunology-allergic-disorders/biology-of-the-immune-system/overview-of-the-immune-system

[17] Chaplin, D. (2003). 1. Overview of the immune response. Journal Of Allergy And Clinical Immunology, 111(2), S442-S459. doi:10.1067/mai.2003.125

[18] BONILLA, F., & GEHA, R. (2006). 2. Update on primary immunodeficiency diseases. Journal Of Allergy And Clinical Immunology, 117(2), S435-S441. doi:10.1016/j.jaci.2005.09.051

[19] Thornton, C., & Morgan, G. (2009). Innate and Adaptive Immune Pathways to Tolerance. Microbial Host-Interaction: Tolerance Versus Allergy, 45-61. doi:10.1159/000235782

[20] Wang, R., Miyahara, Y., & Wang, H. (2008). Toll-like receptors and immune regulation: implications for cancer therapy. Oncogene, 27(2), 181-189. doi:10.1038/sj.onc.1210906

[21] WANG, R. (2006). Immune suppression by tumor-specific CD4+ regulatory T-cells in cancer. Seminars In Cancer Biology, 16(1), 73-79. doi:10.1016/j.semcancer.2005.07.009

[22] Dabbagh, K., & Lewis, D. (2003). Toll-like receptors and T-helper-1/t-helper-2 responses. Current Opinion In Infectious Diseases, 16(3), 199-204. Retrieved from https://insights.ovid.com/pubmed?pmid=12821808

[23] Comput., S. (2018). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses | SIAM Journal on Scientific and Statistical Computing | Vol. 5, No. 3 | Society for Industrial and Applied Mathematics . SIAM Journal On Scientific And Statistical Computing. Retrieved from https://epubs.siam.org/doi/10.1137/0905052

[24] Data Science - Part XV - MARS, Logistic Regression, & Survival Analy…. (2018). Slideshare.net. Retrieved 7 May 2018, from https://www.slideshare.net/DerekKane/data-science-part-xv-mars-logistic-regression-survival-analysis

[25] de Souza, S., & Junqueira, R. (2005). A procedure to assess linearity by ordinary least squares method. Analytica Chimica Acta, 552(1-2), 25-35. doi:10.1016/j.aca.2005.07.043

[26] Rzhetsky, A., & Nei, M. (1992). Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. Journal Of Molecular Evolution, 35(4), 367-375. doi:10.1007/bf00161174

[27]

[28]

[29]

[30]

[31]

[32]

[33]

[34]

[35]

[36]

[37]

[38] A comprehensive beginners guide for Linear, R., A comprehensive beginners guide for Linear, R., & Jain, S. (2017). A comprehensive beginners guide for Linear, Ridge and Lasso Regression. Analytics Vidhya. Retrieved 9 May 2018, from https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

[39] Multicollinearity - Statistics Solutions. (2018). Statistics Solutions. Retrieved 9 May 2018, from https://www.statisticssolutions.com/multicollinearity/

[40] Multicollinearity: Definition, Causes, Examples. (2016). Statistics How To. Retrieved 9 May 2018, from http://www.statisticshowto.com/multicollinearity/

[41] de Vlaming, R., & Groenen, P. (2015). The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. Biomed Research International, 2015, 1-18. doi:10.1155/2015/143712

[42] Cule, E., & De Iorio, M. (2013). Ridge Regression in Prediction Problems: Automatic Choice of the Ridge Parameter. Genetic Epidemiology, 37(7), 704-714. doi:10.1002/gepi.21750

[43] Liang, Y., Liu, C., Luan, X., Leung, K., Chan, T., Xu, Z., & Zhang, H. (2013). Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC Bioinformatics, 14(1), 198. doi:10.1186/1471-2105-14-198

[44] Sum of Squares: Residual Sum, Total Sum, Explained Sum, Within. (2014). Statistics How To. Retrieved 10 May 2018, from http://www.statisticshowto.com/residual-sum-squares/

[45] Regularization: Simple Definition, L1 & L2 Penalties. (2016). Statistics How To. Retrieved 10 May 2018, from http://www.statisticshowto.com/regularization/

[46] Lasso Regression: Simple Definition. (2015). Statistics How To. Retrieved 10 May 2018, from http://www.statisticshowto.com/lasso-regression/

[47] Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers In Genetics, 4. doi:10.3389/fgene.2013.00270

[48] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net." J. R. Statistic. Soc., vol 67 (2005), Part 2., pp. 301-320.

[49] Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Journal Of Machine Learning Research, 5(Sep), 1089-1105. Retrieved from http://www.jmlr.org/papers/v5/grandvalet04a.html?92f58540

[50] Li, Y., Oosting, M., Smeekens, S., Jaeger, M., Aguirre-Gamboa, R., & Le, K. et al. (2016). A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. Cell, 167(4), 1099-1110.e14. doi:10.1016/j.cell.2016.10.017

[51] Aguirre-Gamboa, R., Joosten, I., Urbano, P., van der Molen, R., van Rijssen, E., & van Cranenbroek, B. et al. (2016). Differential Effects of Environmental and Genetic Factors on T and B Cell Immune Traits. Cell Reports, 17(9), 2474-2487. doi:10.1016/j.celrep.2016.10.053

[52] Stolk, R., Rosmalen, J., Postma, D., de Boer, R., Navis, G., & Slaets, J. et al. (2007). Universal risk factors for multifactorial diseases. European Journal Of Epidemiology, 23(1), 67-74. doi:10.1007/s10654-007-9204-4

[53] Scholtens, S., Smidt, N., Swertz, M., Bakker, S., Dotinga, A., & Vonk, J. et al. (2014). Cohort Profile: LifeLines, a three-generation cohort study and biobank. International Journal Of Epidemiology, 44(4), 1172-1180. doi:10.1093/ije/dyu229

[54] Tigchelaar, E., Zhernakova, A., Dekens, J., Hermes, G., Baranska, A., & Mujagic, Z. et al. (2015). Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open, 5(8), e006772. doi:10.1136/bmjopen-2014-006772

[55] R: What is R?. (2018). R-project.org. Retrieved 15 May 2018, from https://www.r-project.org/about.html