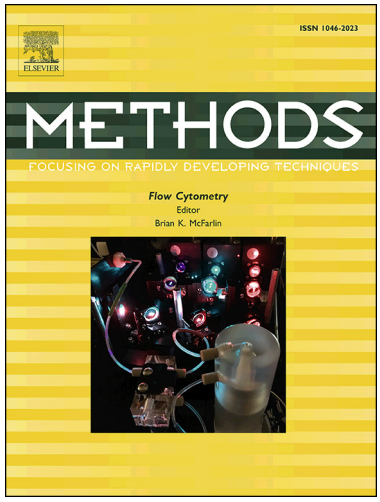
Domain intelligible models



Domain Intelligible Models

Sultan Imangaliyev, Andrei Prodan, Max Nieuwdorp, Albert K. Groen, Natal A.W. van Riel, Evgeni Levin

PII: S1046-2023(17)30423-1
DOI: https://doi.org/10.1016/j.ymeth.2018.06.011
Reference: YMETH 4510

To appear in: *Methods*

Please cite this article as: S. Imangaliyev, A. Prodan, M. Nieuwdorp, A.K. Groen, N.A.W. van Riel, E. Levin, Domain Intelligible Models, *Methods* (2018), doi: https://doi.org/10.1016/j.ymeth.2018.06.011

Domain Intelligible Models

Sultan Imangaliyev[1,2], Andrei Prodan[1,2], Max Nieuwdorp[1], Albert K. Groen[1], Natal A. W. van Riel[1], Evgeni Levin[1,2]

1 Department of Vascular Medicine, Academic Medical Center, University of Amsterdam, 1105 AZ Amsterdam, The Netherlands

2 Horaizon BV, 2625 GZ Delft, The Netherlands

Abstract

Mining biological information from rich "-omics" datasets is facilitated by organizing features into groups that are related to a biological phenomenon or clinical outcome. For example, microorganisms can be grouped based on a phylogenetic tree that depicts their similarities regarding genetic or physical characteristics. Here, we describe algorithms that incorporate auxiliary information in terms of groups of predictors and the relationships between them into the metagenome learning task to build intelligible models. In particular, our cost function guides the feature selection process using auxiliary information by requiring related groups of predictors to provide similar contributions to the final response. We apply the developed algorithms to a recently published dataset analyzing the effects of fecal microbiota transplantation (FMT) in order to identify factors that are associated with improved peripheral insulin sensitivity, leading to accurate predictions of the response to the FMT.

1. Introduction

With ever increasing amount of biomedical data generated via high-throughput technologies, statistical machine learning techniques play a key role in aiding the analysis of these complex, heterogeneous, high-dimensional, and usually noisy datasets. Novel algorithms for large-scale "-omics" analysis have been widely used to generate models that are able to predict a certain health outcome or to identify the features that contribute the most to such prediction. Thus, the

obtained models may not only provide predictions, but also help researchers deepen their understanding of the biological phenomenon.

The majority of popular machine learning methods [1] are not naturally suitable for the analysis of high-dimensional clinical metagenome data. This is particularly true for feature selection algorithms; due to intrinsic nature of relationships present among the predictors, it is challenging to ensure that all clinically important features are selected.

Sparse linear models such as lasso, group-lasso, elastic net, and their extensions [2] are usually considered to be the golden standard of feature selection and interpretability in statistical machine learning. These methods are particularly suitable for building high-dimensional regression models of the form $Y_i = X_i^T \beta + \epsilon_i$ for $i = 1, \ldots, n$, where Y_i is a real-valued response and X_i is a $p-$dimensional feature vector. Sparsity enforcing methods usually minimize the l_1 penalized sum of squares, therefore leading to sparse solutions where many components of β are zeros. Such regularization strategy has shown significant empirical success leading to automatic feature selection, as well as addressing the overfitting problem that is a common issue in high-dimensional regression tasks. On the other hand, sparse models such as lasso are based on the strong assumptions imposed by a linear model and they are not designed to learn non-linear interactions. The non-parametric additive models [3], having the form of $Y_i = \sum_{j=1}^{p} f_j(X_{i,j}^T) + \epsilon_i$, where f_j is a general smooth function, relax these assumptions and allow for much more flexibility in fitting the response variables. In this work, we study extensions of these sparse and additive models for biomarker selection in high-dimensional microbial data.

Making use of auxiliary information usually improves performance and leads to reliable solution for the feature selection model. For example, this can be done by introducing information about the relationship among sets of predictors into the model, such as arranging them in groups [4, 5], leading to the implementation of the group-lasso and the sparse-group lasso, respectively.

However, there is often more information available that could be used to guide the learning process. In the data-rich "-omics" fields, for instance, predictors can

be organized in groups that are expected to be related to a given phenomenon in the same way [6]. Another limitation of the aforementioned techniques is the impossibility of incorporating unlabeled data into the learning process. Our algorithm [7] addresses these issues by imposing semi-supervised co-regularization [9] of related groups of predictors. We assume that there is a known relationship among the predictors, and use it to (a) divide them into groups and (b) determine the relative "distance" or similarity between each pair of groups. We then impose that each group contributes individually to the prediction of the response variable, and that groups that are closely related should provide more similar contributions. This constraint is introduced as a bias in the cost function by enforcing the predictions of pairs of groups to be similar under the L_2 norm.

We use an extension of the sparse-group lasso algorithm, as well as the additive intelligible model to analyze clinical metagenome samples obtained during the FATLOSE project [10] and to identify biomarkers that are associated with improved peripheral insulin sensitivity. In patients with metabolic syndrome, a fecal microbiota transplantation (FMT) from a healthy lean donor can result in a transient increase of insulin sensitivity. However, it is still unknown why some patients show improvement in insulin sensitivity after intervention (responder patients), while others reveal no effect (non-responder patients). To address this question, we apply domain intelligible models to the collected dataset and achieve accurate predictions of the FMT treatment, as well as identification of the most relevant microbial biomarker species associated with the response to the intervention.

This paper is organized as follows. The algorithmic methods are described in sections 2, 3, 4 in which we provide derivations of the co-regularized sparse-group lasso and additive models with auxiliary information. The dataset structure, experimental settings, results of experiments, and concluding remarks are described in section 5.2.

2. Co-regularized sparse-group lasso

Let $S = (\mathbf{X}, \mathbf{y})$ denote a dataset where $\mathbf{y}$ is the $n \times 1$ response vector and $\mathbf{X}$ is the $n \times p$ predictor matrix with n observations and p predictors. After a location and scale transformation, we can assume that the response vector is centered and the predictors are standardized. We will seek to generate a model fitting procedure that produces the vector of coefficients $\hat{\boldsymbol{\beta}}$ $(p \times 1)$ that minimizes a given cost function $L(\boldsymbol{\beta})$:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} (L). \tag{1}$$

For all discussed methods, α and λ are non-negative regularization parameters that control the amount of induced sparsity in the vector of coefficients and the complete regularization term, respectively. The l_a norm of a vector $\boldsymbol{\beta}$ is considered to be $\|\boldsymbol{\beta}\|_a = \left(\sum_i |\beta_i|^a\right)^{1/a}$. In the case in which the predictors are divided into g groups, we re-write:

- The predictor matrix $\mathbf{X} = (\mathbf{X}^1|...|\mathbf{X}^g)$, where each $\mathbf{X}^{(v)}$ is a $n \times p^{(v)}$ sub-matrix, where $p^{(v)}$ is the number of predictors in group v.
- The vector of coefficients: $\boldsymbol{\beta} = (\beta_1, ..., \beta_p) = (\boldsymbol{\beta}^{(1)}|...|\boldsymbol{\beta}^{(g)})$, where $\boldsymbol{\beta}^{(v)}$ $(p^{(v)} \times 1)$ is the vector of coefficients of the predictors of group v.

We consider the following co-regularized sparse-group lasso cost function:

$$\begin{aligned} L_{CSGL}(\boldsymbol{\beta}, \alpha, \lambda, \boldsymbol{\Gamma}) = & \frac{1}{2n} \left\| \mathbf{y} - \sum_{l=1}^{g} \mathbf{X}^{(l)} \boldsymbol{\beta}^{(l)} \right\|_2^2 \\ & + \frac{1}{2n} \sum_{v,l=1}^{g} \gamma_{lv} \left\| \mathbf{X}^{(l)} \boldsymbol{\beta}^{(l)} - \mathbf{X}^{(v)} \boldsymbol{\beta}^{(v)}) \right\|_2^2 \\ & + \alpha\lambda \|\boldsymbol{\beta}\|_1 + (1-\alpha)\lambda \sum_{l=1}^{g} \sqrt{p^{(l)}} \left\| \boldsymbol{\beta}^{(l)} \right\|_2, \end{aligned} \tag{2}$$

where γ_{lv} is the co-regularization coefficient between groups l and v. Given g groups of predictors, the co-regularization coefficients can be arranged in the matrix $\boldsymbol{\Gamma}$ $(g \times g)$ with $\gamma_{1,1}, \gamma_{2,2}, ..., \gamma_{g,g}$ having no effect, since the co-regularization term will always go to zero when $l = v$. The proposed cost function is an

extension of the sparse-group lasso by including the co-regularization term: $\frac{1}{2n}\sum_{l,v=1}^{g}\gamma_{lv}\left\|\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)}-\mathbf{X}^{(v)}\boldsymbol{\beta}^{(v)})\right\|_2^2$. The prediction given by different views representing the same phenomenon should be similar to each other. In the case of groups defined by domain knowledge, this term works by encouraging the intrinsic relationship among predictors to be transferred to their contribution in the final prediction. Suppose that in addition to the training set $S=(\mathbf{X},\mathbf{y})$ with n labeled examples, we have a training set $\widetilde{S}=(\widetilde{\mathbf{X}})$ with m unlabeled samples. It is clear from the equation (2) that the proposed co-regularization term is independent of labels, and can therefore be used to incorporate available unlabeled data. In this case, the cost function can be re-written by substituting $\mathbf{X}$ with $\overline{\mathbf{X}}$, the $(n+m)\times p$ matrix, that results from the concatenation of matrices $\mathbf{X}$ and $\widetilde{\mathbf{X}}$.

3. Algorithm

The aim of the co-regularized sparse-group lasso algorithm is to find $\hat{\boldsymbol{\beta}}$ that minimizes the cost function. Because the function is not differentiable for every $\boldsymbol{\beta}$, we resort to proximal gradient methods [11] to solve the minimization problem. We briefly describe the minimization procedure when using proximal methods, show that it can be applied to the proposed cost function and use it to derive the algorithm to find the solution for the co-regularized sparse-group lasso.

3.1. The block-wise proximal gradient method

The proximal gradient method consists of splitting the cost function into two functions F and R and minimizing them sequentially. When minimizing R, a constraint is imposed in order to enforce that its minimum is kept close to the minimum of F under a squared loss. By construction, F is convex and differentiable and its minimum can therefore be found using standard gradient descent. R is convex and possibly non-differentiable. In our case,

$$F(\boldsymbol{\beta})=\frac{1}{2n}\left\|\mathbf{y}-\sum_{l=1}^{g}\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)}\right\|_2^2+\frac{1}{2n}\sum_{l,v=1}^{g}\gamma_{lv}\left\|\overline{\mathbf{X}}^{(l)}\boldsymbol{\beta}^{(l)}-\overline{\mathbf{X}}^{(v)}\boldsymbol{\beta}^{(v)})\right\|_2^2. \quad (3)$$

$$R(\boldsymbol{\beta}) = \alpha\lambda \left\|\boldsymbol{\beta}\right\|_1 + (1-\alpha)\lambda \sum_{l=1}^{g} \sqrt{p^{(l)}} \left\|\boldsymbol{\beta}^{(l)}\right\|_2 . \tag{4}$$

In the context of grouped predictors, it has been shown that a global minimum can be found by block-wise gradient descent [5, 14, 15], if the cost function is convex and separable between groups. Therefore, the final algorithm consists of updating the predictors of a given group (or block) while keeping all other predictors constant and cycling through all groups until convergence is reached.

3.2. Brief derivation of update rules

We describe the proximity operator for the cost function which was introduced in subsection 3.1. Consider:

$$\text{prox}_R(\mathbf{b}_t^{(l)}) := \min_{\boldsymbol{\beta}^{(l)}} \left(R(\boldsymbol{\beta}_t^{(l)}) + \frac{1}{2t} \left\|\boldsymbol{\beta}_t^{(l)} - \mathbf{b}_t^{(l)}\right\|_2^2 \right), \tag{5}$$

where R is given by equation 4, t is a step size of the proximal gradient descent, and $\mathbf{b}_t^{(l)} = \boldsymbol{\beta}_{t-1}^{(l)} - t\nabla_{(l)} F(\boldsymbol{\beta}_{t-1}^{(l)})$. For the particular case of the co-regularized sparse-group lasso, F is given by equation 3. The subgradient of the minimization problem defined by equation 5 is given by:

$$0 \in \frac{1}{t}(\boldsymbol{\beta}^{(l)} - \mathbf{b}_t^{(l)}) + (1-\alpha)\lambda\partial\left\|\boldsymbol{\beta}^{(l)}\right\|_2 + \alpha\lambda\partial\left\|\boldsymbol{\beta}^{(l)}\right\|_1, \tag{6}$$

which leads to:

$$J = \left\|S(\mathbf{b}^{(l)}, t\alpha\lambda)\right\|_1 \leqslant t(1-\alpha)\lambda \Leftrightarrow \boldsymbol{\beta}^{(l)} = \mathbf{0}, \tag{7}$$

where $S(\mathbf{b}^{(l)}, t\alpha\lambda)$ is the soft-thresholding operator applied to each component $b_j^{(l)}$ of $\mathbf{b}^{(l)}$:

$$(S(\mathbf{b}^{(l)}, t\alpha\lambda))_j = \begin{cases} \text{sgn}(b_j^{(l)})(|b_j^{(l)}| - t\alpha\lambda) & \text{if } |b_j^{(l)}| > t\alpha\lambda \\ 0 & \text{if } |b_j^{(l)}| \leqslant t\alpha\lambda. \end{cases} \tag{8}$$

In the case where condition 7 is not satisfied, a general update rule for $\boldsymbol{\beta}^{(l)}$ needs to be derived. Imposing that $\boldsymbol{\beta}^{(l)} \neq \mathbf{0}$ in equation 6 and re-arranging the terms, we get:

$$\left(1 + \frac{t(1-\alpha)\lambda}{\left\|\boldsymbol{\beta}^{(l)}\right\|_2}\right)\beta_j^{(l)} = b_j^{(l)} - t\alpha\lambda \ \mathrm{sgn}(\beta_j^{(l)}). \tag{9}$$

We can see that equation 9 just holds for $\beta_j^{(l)} \neq 0$ if $\mathrm{sgn}(\beta_j^{(l)}) = \mathrm{sgn}(b_j^{(l)})$ and $|b_j^{(l)}| > t\alpha\lambda$. This condition is equivalent to writing the right side of equation 9 as:

$$\left(1 + \frac{t(1-\alpha)\lambda}{\left\|\boldsymbol{\beta}^{(l)}\right\|_2}\right)\beta_j^{(l)} = (S(\mathbf{b}^{(l)}, t\alpha\lambda))_{j,+}. \tag{10}$$

where $[\cdot]_+ = max\{\cdot, 0\}$. By squaring both sides of equation 10 and solving it for $\left\|\boldsymbol{\beta}^{(l)}\right\|_2$, we obtain the final update rule:

$$\boldsymbol{\beta^{(l)}} = \left(1 - \frac{t(1-\alpha)\lambda}{\left\|S(\mathbf{b}^{(l)}, t\alpha\lambda)\right\|_2}\right)_+ (S(\mathbf{b}^{(l)}, t\alpha\lambda))_+. \tag{11}$$

This leads us to the description of the algorithm:

Algorithm 1 Co-regularized sparse-group lasso algorithm

Require: Cost function split into $F(\boldsymbol{\beta})$ (convex and differentiable) and $R(\boldsymbol{\beta})$ (convex and possibly non-differentiable), $\mathbf{y}$, $\mathbf{X} = (\mathbf{X}^{(1)}, ..., \mathbf{X}^{(g)})$,$\overline{\mathbf{X}} = (\overline{\mathbf{X}}^{(1)}, ..., \overline{\mathbf{X}}^{(g)})$, regularization parameters λ, α, co-regularization matrix $\boldsymbol{\Gamma}$ and maximum number of iterations k_{max}.

Ensure: $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{(1)}|...|\boldsymbol{\beta}_0^{(g)}) \leftarrow \mathbf{0}$.

while Number of iterations $k < k_{max}$ **do**
  for Each group $l = 1, ..., g$ **do**
    Check if group l can be eliminated by checking condition 7
    if yes **then**
      $\boldsymbol{\beta}^{(l)} \leftarrow \mathbf{0}$ and proceed to the next group
    else
      Until convergence, update $\boldsymbol{\beta}^{(l)}$ using the update rule 11.

3.3. Determination of regularization parameters

For the co-regularized method, the matrix $\mathbf{\Gamma}$ also needs to be specified. The proposed cost function is an extension of the sparse-group lasso by including the co-regularization term: $\frac{1}{2n}\sum_{l,v=1}^{M}\gamma_{lv}(\mathbf{X}^{(l)}\boldsymbol{\beta}^{(l)} - \mathbf{X}^{(v)}\boldsymbol{\beta}^{(v)})^2$. This term is analogous to a multi-view problem [8, 9, 23], in which the prediction given by different views representing the same phenomenon should be similar to each other. In the case of groups defined by domain knowledge, this term works by encouraging the intrinsic relationship among predictors to be transferred to their contribution in the final prediction. The value chosen for γ_{lv} will therefore reflect the extent to which two groups l and v are expected to give similar contributions to $\mathbf{y}$, where $\gamma_{lv} = 0$ represents the case where no similarity at all is expected. Ideally, domain knowledge should be the main driver to determine the value of each element of the matrix $\mathbf{\Gamma}$. The optimal values of the regularization hyper-parameters α and λ for co-regularized sparse-group lasso algorithm can be determined via k-fold cross-validation. Fine-tuning of values can be done by methods such as grid search, but this procedure can become computationally extensive depending of the number of predictor groups. Therefore, the use of domain knowledge can help decrease the number of degrees of freedom by setting the co-regularization term between groups that are not expected to be related to zero. Such an approach was adapted in a biological problem [22], where groups and group relationships were determined based on domain knowledge from molecular biology. In experiments performed on synthetic datasets [7], the value of γ was determined as follows: a relatively small grid of values was chosen and a model was fitted for each of them. Next, the quality of all models was analyzed regarding their ability to retrieve all relevant predictors in order to evaluate the effect of the co-regularization term.

4. Additive models with auxiliary information

Introduced by Hastie *et al.* [3], Generalized Additive Models (GAM), also referred to as Intelligible Models, lie in between fully parametric and nonpara-

metric models. GAM are less general in comparison to "fully" nonparametric models, but have a notable advantage of being readily interpretable and easier to estimate using a simple backfitting algorithm. Recently, additive models have been successfully applied in the biomedical domain [18], and they can be naturally extended to include pairwise interactions among predictors [17].

To characterize and define our additive model with auxiliary information, we first introduce some basic notation and adaptation in order to apply additive models to a high-dimensional learning setting. In case of fully nonparametric setting we consider estimation of the regression functions $f(\mathbf{X}) = f(X_1, \ldots, Xp) = \mathbb{E}[Y|X_1, \ldots, Xp]$ learned from a set of n data samples $\{\mathbf{x}^{(i)}, y^{(i)} : x^{(i)} \in \mathbb{R}^p, y^{(i)} \in \mathbb{R}\}$. Without assuming any parametric form of $f(X)$ we can write $Y = f(X) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$. In case of single covariate $f(X) = \mathbb{E}[Y|X]$, the function is known as the orthogonal projection of Y onto the linear space of all measurable functions of X and can be written as $f(X) = PY$, where P is the conditional expectation operator $\mathbb{E}[\cdot|X]$. By making the assumption that $f(x) = \mathbb{E}[Y|X = x]$ is a smooth function of x we can estimate $f(x)$ using a class of smoothing estimators, *e.g.* $f(x) = \sum_{i=1}^{n} l(x_i)y^i = l(x)^T\mathbf{y}$, where $l(x)$ is any smoothing function such as kernel smoother, spline, *etc.* Now let $\mathbf{y}^* \in \mathbb{R}$ be a vector of fitted values $y^{(i)}$ at $x^{(i)}$. Then the linear smoother has the form of $\mathbf{y}^* = \mathbf{S}\mathbf{y}$, where $\mathbf{S}_{i,j} = l_j(x^i)$ with $i, j = 1, \ldots n$.

Although it is straightforward to generalize one-dimensional smoothers to p-dimensional case, it is well known that smoothers break down in high dimensions due to the curse of dimensionality. This shortcoming motivates the study of additive models [3]:

$$g(X_1, \ldots, X_p) = \sum_{j}^{p} f_j(X_j) + \alpha, \tag{12}$$

where $f_1, ..., f_p$ are one-dimensional smooth component functions, one for each covariate. For simplicity and identification purposes, we assume $\alpha = 0$ and $f_j \in \mathcal{H}_j$, where $\mathcal{H}_j = \{f_j|\mathbb{E}[f_j(X_j)] = 0\}$. The optimization problem of additive

models in the population setting is to minimize:

$$L(\mathbf{f}) = \mathbb{E}\left[\left(Y - \sum_{j}^{p} f_j(X_j)\right)^2\right] \tag{13}$$

over $\{\mathbf{f} : f_j \in \mathcal{H}_j\}$ and the minimizer can be shown to satisfy:

$$f_j = \mathbb{E}\left[Y - \sum_{k \neq j} f_k | X_j\right] := P_j\left(Y - \sum_{k \neq j} f_k\right), \tag{14}$$

where $P_j = \mathbb{E}[\cdot|X_j]$ is the projection operator onto $\mathcal{H}_j$. Replacing P_j by a linear smoother with smoother matrix Sj immediately leads to a sample version of the above iterative procedure for fitting additive models:

$$\mathbf{f}_j^* \leftarrow \mathbf{S}_j\left(\mathbf{y} - \sum_{k \neq j} \mathbf{f}_k^*\right), j = 1, \ldots, p. \tag{15}$$

This simple iterative algorithm is known as backfitting. There have been multiple extensions proposed for additive models in order to be applicable to high-dimensional learning tasks. In particular, ideas to enforce sparsity [19] allow application of additive models to complex bioinformatics tasks. In this case, we consider a sparsity constraint on functions via regularization $\min_{\mathbf{f}}(L(\mathbf{f}) + \lambda\Omega(\mathbf{f}))$ where λ is the regularization parameter and $\Omega(\mathbf{f}) = \sum_{j=1}^{p} |f_j|$ encourages functional sparsity of the components. Then the stationary conditions can be given by:

$$f_j = \left[1 - \frac{\lambda}{|P_j R_j|}\right]_+ P_j R_j, \tag{16}$$

where $R_j = Y - \sum_{k \neq j} f_k$ is the partial residual and $[\cdot]_+ = max\{\cdot, 0\}$.

5. Results and Conclusion

5.1. Comparison with GAM

To evaluate predictive performance of the intelligible sparse additive model we first conducted an experiment on a synthetic dataset. The synthetic dataset was generated using scikit-learn package [24]. The dataset contains 250 examples and 500 features, among which only 10 are useful to predict target regression variable

(the rest are noisy irrelevant features). Part of the dataset, namely 80% is used for training, while the rest 20% was used for testing. We compared the model with a non-sparse Generalized Additive Model (GAM) [3]. GAM was implemented in pygam package[1]. Hyperparameters of all models were selected via 10-fold cross-validation when applicable. As a performance measure mean squared error (MSE) was computed on the test dataset. Results of this experiment demonstrate that in terms of intelligibility and predictive performance sparse additive model notably outperforms standard GAMs: 0.1 MSE vs. 0.6 MSE. Also, sparse additive model correctly picks up 9 out of 10 important features, while standard GAM fails to identify these most predictive variables.

5.2. Experiments on biological dataset

We used the two techniqes: a) co-regularized algorithm in regression and classification setting with smoothed hinge loss optimized via stochastic gradient and b) the intelligible sparse additive model to analyze clinical metagenome samples in order to identify biomarkers that are associated with improved peripheral insulin sensitivity after FMT. Our dataset and experimental setup are described in the previous study [10]. Specifically, the dataset contains labeled microbial data from thirty-eight subjects measured at the baseline time point, as well as the corresponding rate of glucose disappearance (RD) measurements used as a clinical measure for peripheral insulin sensitivity. In total, the microbial dataset consists of one thousand twenty-five features. To compensate for variance in feature amplitudes, all input features were zero-mean unit-variance scaled, while RD measurements were only centered.

The features were selected based on their stability after thirty runs of the feature selection model on randomly selected 80% data shuffles using the stability selection procedure [20]. The performance measure used for the regression task was the mean squared error (for RD at the baseline). The regularized feature selection model was parameterized by the hyper-parameter $\alpha \in [0, 1]$, which

[1] https://github.com/dswah/pyGAM

modulates the emphasis of sparsity induced on features. The shape functions are parameterized by the hyper-parameter $f \in [0, 1]$, which modulates the smoothness of locally weighted regression by varying the fraction of the data used when estimating y. Both of the hyper-parameters were specified in grid of values between 0.1 and 0.9 with step size of 0.2. The backfitting procedure was performed with a fixed number of fifteen iterations. Features which were selected in more than 70% of the stability runs were considered as important, and their average values were visualized using smoothed conditional means implemented in the R *ggplot2* package [21]. The trained smooth component functions for four selected features are presented in Figure 1 and Figure 2. These figures depict zero-mean unit-variance scaled microbial abundance values on horizontal axes and corresponding contributions to predicted rate of glucose disappearance measurements on vertical axes. For comparison purposes, only those four features were selected, which were reported in the previous study [10]. The gray areas represent standard deviations of smooth curves calculated using stability selection shuffles.

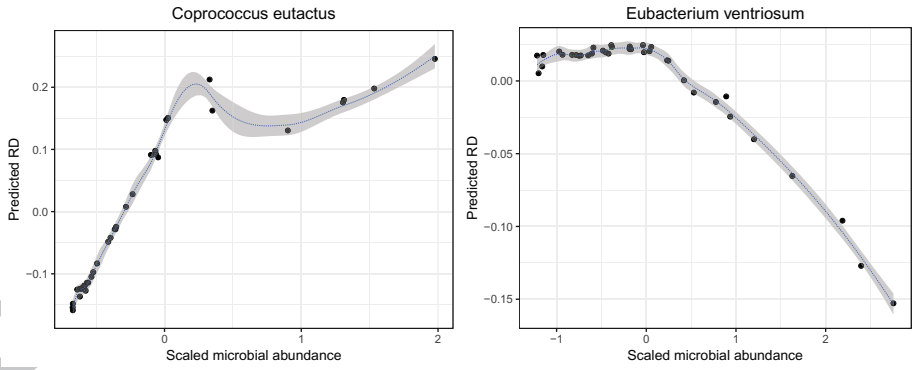


Figure 1: Predicted rate of glucose disappearance (RD) profile of the *Coprococcus eutactus* and *Eubacterium ventriosum* biomarker species. We have observed that abundance of *Eubacterium ventriosum* was lower in baseline fecal samples of responders.

Using the model, we found that metabolic responders, *i.e.* patients who show improvement after intervention, were characterized by lower initial fecal micro-

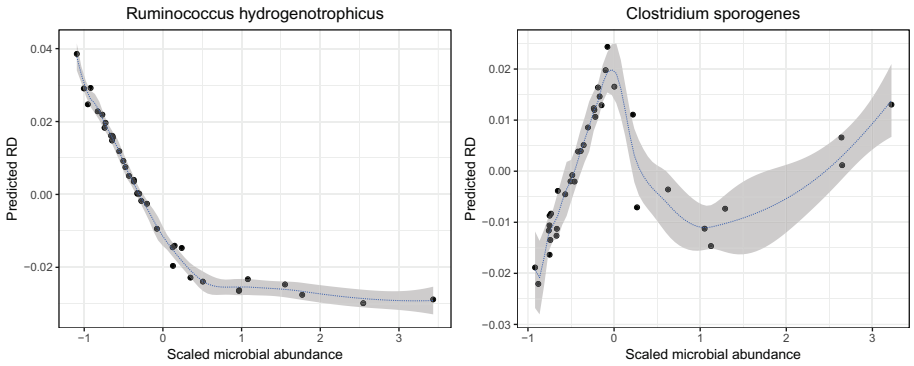


Figure 2: Predicted rate of glucose disappearance (RD) profile of the *Ruminococcus hydrogenotrophicus* and *Clostridium sporogenes* biomarker species. The increased abundance of the species *Ruminococcus* in the baseline fecal samples of non-responders has been previously linked to adverse intestinal health.

biota diversity. This was combined with higher abundance of *Subdoligranulum variabile* and *Dorea longicatena* in comparison to non-responders (*i.e.* patients on whom intervention had no effect), whereas abundance of *Eubacterium ventriosum* and *Ruminococcus torques* was lower in baseline fecal samples of responders. We also identified that the majority of the predictive fecal microbiota comprised abundance of four different species. In line with previous reports on metabolic response upon dietary intervention, our responders were characterized by increased pre-treatment abundance of *Subdoligranulum variabile*. In contrast, the increased abundance of the species *Ruminococcus* in the baseline fecal samples of non-responders has been previously linked to adverse intestinal health and aberrant production of fatty acid chain-containing metabolites. Based on these findings we conclude that for future interventions, determining baseline fecal microbiota composition might aid in predicting efficacy of treatment.

In conclusion, our analysis and the obtained results underscore suggestions that pre-treatment fecal microbiota biomarkers might regulate engraftment of lean-donor-derived bacterial species and thus predict treatment success. Disentangling such a specific signature of intestinal microbiota involved in ben-

eficial functional metabolic shifts might help to apply approaches aiming for improved prediction of insulin resistance development, as well as to design targeted microbiota-based interventions in obese individuals. The discovered species together with the developed model may pave the way for more personalized and targeted therapies for patients with metabolic syndrome. A main advantage of the domain intelligible models is increased interpretability, which is critical in biomedical tasks. We emphasise that in clinical practice, model interpretation is more important than model accuracy, thus less accurate models might be chosen instead of possibly more accurate "black-box" algorithms [18].

References

[1] Efron, B., Hastie, T: Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press (2016).

[2] Hastie, T., Tibshirani, R., and Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC (2015).

[3] Hastie, T. J. and Tibshirani, R. J.: Generalized additive models. Statistical Science (1986).

[4] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Statist. Soc. B 68, Part 1, 49–67 (2006).

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

[6] Rossello-Mora, R.: Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Taxonomy and Biodiversity 14 (2), 318–334 (2012).

[7] Santos, P.L.A., Imangaliyev, S., Schutte, K., Levin, E.: Feature Selection via Co-regularized Sparse-Group Lasso. International Workshop on Machine Learning, Optimization and Big Data, 118–131. Springer (2016)

[8] Ruijter, T., Tsivtsivadze, E., Heskes, T.: Online Co-Regularized Algorithms. Algorithmic Learning Theory/Discovery Science (2012).

[9] Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semisupervised learning with multiple views. Proceedings of ICML Workshop on Learning with Multiple Views (2005).

[10] Kootte, R., Levin, E., Salojarvi J., Smits, L. , Hartstra, A., Udayappan S., Hermes G., Bouter, K., Koopen, A., Holst, J., Knop, F., Blaak, E., Zhao, J., Smidt, H., Harms, A., Hankemeijer, T., Bergman, J., Romijn, H., Schaap F., Olde Damink, S., Ackermans, M., Dallinga-Thie, G., Zoetendal, E., de Vos, W., Serlie, M., Stroes, E., Groen, A., Nieuwdorp, M.: Improvement of Insulin Sensitivity after Lean Donor Feces in Metabolic Syndrome Is Driven by Baseline Intestinal Microbiota Composition. Cell Metabolism , Volume 26, Issue 4, 611–619 (2017)

[11] Parikh, N., Boyd, S.: Proximal Algorithms. Now Publishers Inc (2013).

[12] Hea, Z., Weichuan Yub, W.:Stable feature selection for biomarker discovery. Computational Biology and Chemistry 34, 215–225 (2010)

[13] Guyon, I. Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157–1182 (2003).

[14] Simon, N., Friedman, J., Hastie, T.: A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression (2013).

[15] Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso (2010).

[16] Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 150–158 (2012).

[17] Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 623–631 (2013).

[18] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD international conference on Knowledge discovery and data mining, 1721–1730 (2015).

[19] Ravikumar, P., Liu, H., Lafferty, J., Wasserman, L.: Spam: Sparse additive models. Proceedings of the 20th International Conference on Neural Information Processing Systems, 1201–1208 (2007).

[20] Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)

[21] Wickham, H.: ggplot2: elegant graphics for data analysis. Springer (2016)

[22] Imangaliyev, S., Matse, J.H., Bolscher, J.G.M., Brakenhoff, R.H., Wong, D.T.W., Bloemena, E., Veerman, E.C.I., Levin, E.: Discovery of salivary gland tumors' biomarkers via co-regularized sparse-group lasso. International Conference on Discovery Science, 298–305. Springer (2017)

[23] Imangaliyev, S., Levin, E.: Unsupervised Multi-View Feature Selection for Tumor Subtype Identification. ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 491-499. ACM (2017)

[24] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.: Scikit-learn Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830 (2011)

In the data-rich ``-omics'' fields features can be organised in groups that are related to a biological phenomenon or clinical outcome in the same way. For example, microorganisms can be grouped based on a phylogenetic tree that depicts their similarities regarding genetic or physical characteristics. We describe the algorithms that allows building intelligible models as well as incorporation of auxiliary information into the metagenome learning task in terms of groups of predictors and the relationships between those groups. In particular, our cost function guides the feature selection process using phylogenetic information by requiring related groups of predictors to provide similar contributions to the final response. We apply the algorithms to recently collected and published data on microbial effects of fecal microbial transplantation leading to accurate predictions of the response to the FMT treatment.