Cornelia Fütterer, Malte Nalenz and Thomas Augustin

Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection

Technical Report Number 239, 2021
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Discriminative Power Lasso – Incorporating Discriminative Power of Genes into Regularization-Based Variable Selection

Cornelia Fuetterer[+*], Malte Nalenz[+*], and Thomas Augustin[+]

Abstract—In precision medicine, it is known that specific genes are decisive for the development of different cell types. In drug development it is therefore of high relevance to identify biomarkers that allow to distinguish cell-subtypes that are connected to a disease. The main goal is to find a sparse set of genes that can be used for prediction. For standard classification methods the high dimensionality of gene expression data poses a severe challenge. Common approaches address this problem by excluding genes during preprocessing. As an alternative, L1-regularized regression (Lasso) can be used in order to identify the most impactful genes.

We argue to use an adaptive penalization scheme, based on the biological insight that decisive genes are expressed differently among the cell types. The differences in gene expression are measured as their *discriminitive power* (DP), which is based on the univariate compactness within classes and separation between classes. ANOVA based measures, as well as measures coming from clustering theory, are applied to construct the covariate specific DP.

The resulting model, that we call *Discriminative Power Lasso* (DP-Lasso), incorporates the DP as covariate specific penalization into the Lasso. Genes with a higher DP are penalized less heavily and have a higher chance for being part of the final model. With that the model can be guided towards more promising and trustworthy genes, while the coefficients of uninformative genes can be shrunken to zero more reliably.

We test our method on single-cell RNA-sequencing data as well as on simulated data. DP-Lasso leads on average to significantly sparser solutions compared to competing Lasso-based regularization approaches, while being competitive in terms of accuracy.

Keywords—Penalized Regression, Variable Selection, Clustering validation metrics, scRNA-sequencing data.

I. Introduction

In personalized medicine, it is important to identify genes, which can be used to accurately predict the individual outcomes. For the development of biomarkers, a lower number of covariates means less effort in its subsequent clinical testing. As in high-dimensional settings many genes are often noise, the challenge is to select only the covariates that are relevant in terms of prognostic, predictive or biological impact to the drug or the disease [19]. In case of non-small cell lung cancer (NSCLC), the detection of the biomarker EML4-ALK fusion gene [27] led to the development of the drug crizotinib, which is used for patients carrying an ALK-fusion. In contrast to the earlier low response, crizotinib dramatically raised the response rate in NSCLC [19].

In general, the transition of healthy cells into cancerous cells affects changes in gene expression that can be measured. It is therefore common practice to investigate single-cell RNA sequencing data, introduced by [30], which allows insights into the different cell types of single cells. In the case of a cell cycle, the cell passes from the DNA synthesis (S-phase) to the mitosis (M-phase), including the gap phases (G1 and G2) in between. These different phases can be distinguished by its measured gene expression of a synchronized cell population. For example, a high score at the G2M checkpoint can be an indicator of metastasis tumor [21]. Testing whether genes are differentially expressed among different cell types might therefore lead to valuable insights.

From a biological point of view, it is therefore of relevance to extract a sparse set of genes that can be used to classify and characterize the subpopulations [11]. One common approach is to use penalized regression models, such as the Lasso [31] that find a trade-off between model fit and model complexity. The advantage of the Lasso is that it provides variable selection, by setting coefficients to exactly zero. An extension is the adaptive Lasso [36] which uses covariate specific penalization terms. The penalization terms are inversely proportional to the ordinary least square (OLS) estimates from a multivariate regression model.

In this article, we combine the concepts of regularized regression with the biological background of differentially expressed genes. Genes that differ univariately with respect to the target, should be penalized less heavily.

We therefore introduce the term discriminative power (DP), which allows a covariate specific evaluation of compactness and separation with regard to the outcome. Discriminative power is measured by means of clustering indices [3], as well as by the classic concept of analysis of variance (ANOVA) [12].

The discriminative power is directly incorporated as covariate specific penalization into the adaptive Lasso, resulting in our approach Discriminative Power Lasso (DP-Lasso).

Using the DP as penalization weights in a L1-regularized model can be seen as a soft filtering as we do not exclude any covariates before performing regression, but favour genes with good univariate properties. The idea is to give covariates with low univariate DP a higher penalty, while reducing the penalty on the more promising covariates.

[+]Ludwig-Maximilians-University, Munich. Department of Statistics.

[*]These authors contributed equally to this work.

This paper is structured as follows. In Section II we introduce notations give an overview over commonly used regularization based methods. Section III introduces the DP-Lasso model. In Section IV and Section V we test the performance of DP-Lasso on scRNA-sequencing datasets as benchmark datasets, and on simulated data. Section VI concludes and provides an outlook.

II. Methods

In supervised learning, the goal is to estimate the underlying function that maps the p-dimensional covariate space to the outcome. As training data, we are given a matrix X, composed of p covariate vectors each containing the values of the N observations. This leads to the covariate matrix $X = (x_1, \cdots, x_p), j = 1, \cdots, p$, and the vector Y containing the N outcomes. x_{ij} denotes the value of observation i for covariate j, x_j the N values of covariate j, and $x_{i\cdot}$ the p dimensional observation vector for observation i. Given that the outcome is continuous, a common approach is to estimate the linear model

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad , \tag{1}$$

where β is the p-dimensional vector of regression coefficients. In the following categorical outcomes $Y \in \{1, \cdots, K\}$ are considered. In this case a generalized linear model (glm) is appropriate, which uses a linear structure as in Equation 1 and connects it to the target through a link function [10]. Thus, for binary outcomes $Y \in \{0, 1\}$ logistic regression is used and for $K > 2$ classes the multinomial-logit model. However, for ease of notation in the following the linear model is used in the description of the methods.

In high dimensional data and especially $p >> N$ glms cannot reliably be estimated, due to the problems of multicollinearity and perfect separation [1, 14]. Also glms can not deal efficiently with irrelevant predictors, as no variable selection is performed. It is therefore common practice to reduce the number of genes before analysis.

For this purpose, the univariate filtering approach selects covariates based on (adjusted) p-values of univariate tests or biological reasoning. The final result highly depends on the researcher's choice, because a threshold or number of genes kept for the analysis has to be specified.

Alternatively, one can use regularized regression models for parameter estimation, that find a trade-off between model fit and model complexity. Regularized regression models also lead to more stable solutions for β coefficients in $p >> N$, as extreme behaviour is penalized [15]. This allows to find a unique solution in situations where glms might fail, such as perfect seperability and multicolinearity.

In regularized regression models, the overall loss function is decomposed in the discrepancy of the observed target and the model prediction and a penalty term that controls the complexity of the model. In case of the classical Lasso, the penalty is equal to the L1-norm of the coefficients β, leading to the overall loss function [31]:

$$L(y, X, \beta, \lambda, w) = \underbrace{\sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2}_{\text{SSE}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{Penalty Term}}, \tag{2}$$

for linear regression. The degree of shrinkage and sparsity is controlled by a global shrinkage parameter λ, which is usually chosen via cross-validation.

Lasso regression allows to shrink coefficients to exactly zero, which leads to a covariate selection. Lasso has efficient solvers available, making it a good choice for high dimensional datasets. However, the Lasso has the known deficiency of overshrinkage: To remove a large number of uninformative covariates, a high penalty parameter needs to be chosen. This in return will also shrink the coefficients of informative predictors to some extent. To counteract, the Lasso will take in correlated predictors, to substitute for the overshrinkage [35]. This makes the interpretation of covariates left in the final model somewhat dubious, as it is unclear if the covariate itself is important or just as a substitute for the overshrinkage of another covariate.

If predictive performance is the primary objective, Ridge regression (L2-penalty) is a popular alternative. L2-penalty limits the influence of individual covariates, by penalizing high β's strongly, but shrinks no coefficient to exactly zero [15].

The Elastic Net (Zou and Hastie 2005) uses a mixture of the L1-norm (Lasso) and the L2-norm (Ridge). The loss function of the Elastic Net can be written as

$$L(y, X, \beta, \lambda, w) = \sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2 + \alpha \sum_{j=1}^{p} \lambda_j |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \lambda_j \beta_j^2, \tag{3}$$

where α is a mixing parameter that controls the proportion of L1 and L2-penalty that is put on the coefficients.
Elastic Net often shows better predictive performance than Lasso, while also being able to set coefficients to exactly zero.

To reduce the amount of over-shrinkage and improve variable selection consistency, the adaptive Lasso [36] was proposed. Instead of using the same global shrinkage λ on every coefficient, the adaptive Lasso uses a covariate specific shrinkage parameter λ_j, which allows a separate penalty for each covariate. This leads to the loss function of the adaptive Lasso [36]

$$L(y, X, \beta, \lambda, w) = \sum_{i=1}^{N} (y_i - x_{i\cdot}\beta)^2 + \sum_{j}^{p} \lambda_j |\beta_j|, \tag{4}$$

where $\lambda_j = \lambda w_j$ is the covariate specific penalty and w_j discount factors that increase or decrease the amount of penalization for covariate j. In the original adaptive Lasso, w_j is calculated as the inverse of the parameter estimates of the ordinary least squares (OLS) regression, hence $w_j = 1/\hat{\beta}_j^{(OLS)}$. This approach can be shown to improve the model selection consistency under certain assumptions [36]. More concretely this results in less penalization of important covariates with high $\hat{\beta}_j^{(OLS)}$, which allows the final coefficients to become large, mitigating the over-shrinkage effect. In case of $p >> N$, the covariate specific weighting can be obtained by a ridge regression instead of the OLS estimates.

Several other extensions of the Lasso have been proposed, such as the fused Lasso [32], group Lasso [20], Bayesian Lasso [22] and Bayesian shrinkage priors [2].

Another commonly used approach for gene selection is the usage of tree ensembles, such as random forests [8]. Random forests [4], that combine several decision trees, are a popular choice for genetic classification data, as they posses strong predictive performance and do not require further assumptions. Measures, such as (unbiased) variable importance [29] and SHAP values [17] can be used to assess the importance of individual covariates, to rank covariates and to identify the most impactful genes.

III. Discriminative Power Lasso

In $p >> N$ situations, where the number of covariates exceeds the number of observations, there always exists an infinite amount of solutions for the regression hyperplane defined by the regression coefficients. While regularization helps to promote sparsity and limit extreme behaviour, we argue that additional information can guide the model towards more robust and reliable solutions. In contrast to the original adaptive Lasso, we want to limit the impact of covariates that only work well in a multivariate model, but are not discriminative univariatly. If enough data is available, such interplay between different covariates can be reliably estimated. However, with limited training data, the chance of over-fitting on spurious relationships is high, when learning multivariate models. Therefore, we suggest to instead promote genes that decompose the data in ‘natural’ groups, measured by the univariate discriminative power based on the conditional distribution $f(X_j|Y), j = 1, ..., p$.

The construction of the DP can be motivated by the concept of analysis of variance that measures the impact of a grouping variable on a numeric outcome by their differences in means. Therefore, for the construction of the DP we use the dependent variable Y as independent variable that we condition on to explain the differences in X. This change in perspective adds new information that is unavailable in a purely supervised regression approach. Secondly, cluster validation measures that have been developed in unsupervised clustering theory can be applied. Instead of using the outputted cluster labels as groups, as it is usually done in unsupervised learning, we directly use the target labels Y as grouping. The discriminative power therefore measures how well a covariate decomposes the underlying groups in terms of compactness and separation.

A. Target Adaptive Regularization

We implement the preference towards covariates with high discriminative power by discounting their penalty, similarly to the adaptive Lasso. The overall loss function of DP-Lasso can be written as

$$L(y, X, \beta, \lambda, w) = \mathcal{E}(\hat{y}, y, \beta) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{5}$$

where $\mathcal{E}$ is an appropriate loss function measuring the deviation from the fitted response vector $\hat{y}$ and the true values y, using a suitable link function. For logistic regression deviance or log-loss are common choices for $\mathcal{E}$. In case of a linear model the model takes the form of Equation 4. We propose to chose the covariate specific penalty as $\lambda_j^{(DP)} := \lambda w_j^{(DP)}$ and $w_j^{(DP)} = 1/DP_j$, where DP_j is the discriminative power of gene j. This gives the model a gentle push towards covariates that appear more natural and reliable, based on their DP. Note that both the calculation of DP and the following regularized regression model are based on all N observations of the training data.

Combining the DP with the supervised approach enriches the regression model with new information. Covariates with high DP are more likely to be selected in the final model, whereas covariates, that only work well in a multivariate model, but have a low individual DP are more likely to be removed. The adaptive shrinkage parameter also counteracts the over-shrinkage. Coefficients of covariates that work well in the multivariate model and also appear as good candidates, based on their DP, will be penalized less heavily and will be allowed to become large. On the other hand, clearly uninformative covariates with a low DP will receive an even higher penalty and can be removed more easily in the regularization step. Lastly, if several solutions to Equation 5 are similarly good, our approach gives a gentle push towards covariates that appear more trustworthy.

B. Characterization of natural groupings

This section motivates the construction of our DP measures. In general, we assume covariates X_j in which the underlying groups Y are homogeneous and well separated from the other groups as more promising . This reflects the idea that relevant genes should express differently among the K classes. Figure 1 shows the distribution of two example genes from the later used single-cell RNA-sequencing dataset EMTAB2805 of [5]. For the gene on the left side, we can see that the two underlying classes show clear differences in their distribution. Also the two groups are relatively compact and their group-means well separated. For the gene on the right side, the two groups show a stronger overlap, and they are less separated. Therefore, the gene on the left side appears to be a more natural candidate for a decisive gene and should have a higher chance of being selected.

The same rationale can be used for $K > 2$. Figure 2 shows

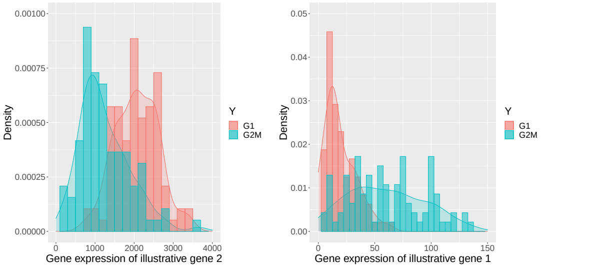

Fig. 1. Univariate distributions of two genes. The colors indicate the two groups. Left side: the two classes show clear differences in their distribution. Right side: the distributions are strongly overlapping with no clear difference.

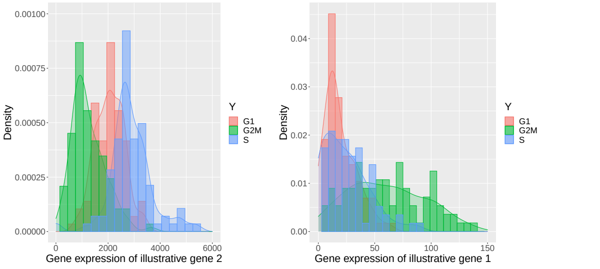

Fig. 2. Univariate distributions of two genes. The colors indicate the three groups. Left side: the three classes show clear differences in their distribution. Right side: the distributions are strongly overlapping with no clear difference.

the univariate distributions for three classes on the same genes, which can be used to assess the compactness and separation. The idea of DP-Lasso is therefore to prefer genes that decompose nicely into the underlying classes with regard to compactness and separation. We call this concept of 'natural grouping' the discriminative power DP. Genes with a high discriminative power will be favored in the regularization step (see Section III-A).

When using for example a logistic regression model, compactness of the groups (as an indication of naturality of the group) is not directly evaluated. The same goes for the distance between groups (or their means): As long as the groups are perfectly separable by a hyperplane, as is the case in $p >> N$, the margin to the discrimination plane is typically not considered in the loss function. Figure 3 shows two simulated covariates with a similar slope from a logistic regression model. While the two classes can be separated similarly good in both covariates, intuitively we would prefer the covariate shown on in right side, due to its distribution. Here the two classes express differently and the two groups are both compact and well separated, whereas the distribution on the left side appears more likely to be random. These descriptive illustrations aim to motivate the inclusion of additional information into the penalization by the discriminative power, which is described in the following.

The natural decomposition can be formalized by the concepts of compactness and separation with respect to the response.

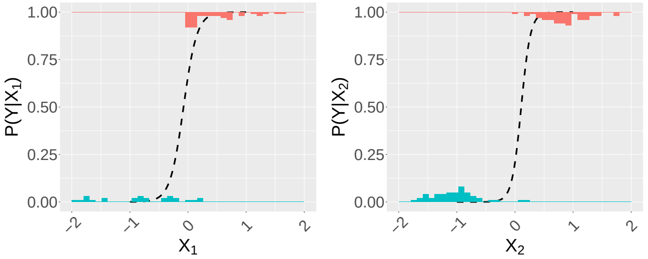

Fig. 3. The graph shows simulated genes, that can similarly well discriminated by a logistic regression. Left side: the clusters appear unnatural. Right side: compact groups with well separated group means.

C. Measures of discriminative power (DP)

In the following we describe three interesting options to measure the discriminative power. The goal is to capture information about the compactness and seperation between classes in each gene. The discriminative power is therefore calculated univariatly over each covariate j using the target variable y as grouping. In the following

$$x_j^{(k)} = \{x_{ij} : y_i = k\}_{i=1}^N \tag{6}$$

denotes the set of values of covariate j that belong to observations with the target class k, and $x_{hj}^{(k)}$ denotes the covariate values of the h'th observation in class k.
There exists a large number of quality criteria that are commonly used in unsupervised learning to evaluate clustering solutions. Also the idea of discriminative power can be interpreted as a classical test problem. The following describes three ways to measure DP, based on these principles.

1) ANOVA-approach: One classical way to test for differences in group means is the analysis of variance (ANOVA) [12]. Intuitively, the ANOVA expresses how much of the sample variance can be explained by the grouping. More concretely, the ANOVA tests whether there is a difference in the means of K groups based on its F-statistic.
Let $\bar{x}_j^{(k)} = \frac{1}{n_k}\sum_{h=1}^{n_k} x_{hj}^{(k)}$ denote the class mean of covariate j in target class k, where n_k is number of observations belonging to class k and $\bar{x}_j$ denotes the overall mean over all N observations. The according test statistic F_j measures the ratio of between-group variability and within-group variablity of covariate j via

$$F_j = \frac{(N-K)}{(K-1)} \frac{\sum_{k=1}^K n_k(\bar{x}_j^{(k)} - \bar{x}_j)^2}{\sum_{k=1}^K \sum_{h=1}^{n_k}(x_{hj}^{(k)} - \bar{x}_j^{(k)})^2}. \tag{7}$$

The value of the F-statistic is large in case that the distances between the groups are considerably higher than the distances within the groups. The higher the F-statistic, the higher the proportion of variance explained by the grouping, indicating significant differences in class means. We thus use the value of the F-statistic as one possibility for the measurement of discriminative power and determine the discount factor w_j^{DP} for the penalization in the subsequent step with $w_j^{(ANOVA)} = 1/F_j$. As $1/F_j$ can become quite large we use a logarithmic transform to attenuate the differences in DP between the genes and to avoid numerical instabilities.

2) Davies-Bouldin Index: The Davies-Bouldin index *DB* measure was developed for validating the clustering quality based on compactness and separation of the clusters [6]. As mentioned before, instead of evaluating a cluster solution, the K classes are evaluated. The DB index relates the compactness within the groups to the separation between the classes. The compactness of class k is measured root mean squared error of observations from class k to the class mean $\bar{x}_j^{(k)}$ of class k in covariate j, leading to

$$\Delta_j^{DB}(k) = \sqrt{\frac{1}{n_k}\sum_{h=1}^{n_k}(x_{hj}^{(k)} - \bar{x}_j^{(k)})^2},$$

which in the univariate case simplifies to the standard deviation of observations in group k. The separation between the groups k and l groups is measured via the Euclidian distance of their respective class means $\bar{x}_j^{(k)}$ and $\bar{x}_j^{(l)}$, which in the univariate case simplifies to

$$\delta_j^{DB}(k,l) = |\bar{x}_j^{(k)} - \bar{x}_j^{(l)}|.$$

The overall DB Index is then given as

$$DB_j = \frac{1}{K}\sum_{k=1}^{K}\max_{l\neq k}\left\{\frac{\Delta_j^{DB}(k) + \Delta_j^{DB}(l)}{\delta_j^{DB}(k,l)}\right\}, \qquad (8)$$

which compares each class to its closest class, as a more pessimistic measure. The better the groups are separated and compact, the lower the DB index becomes, and as a consequence, the less penalization this covariate should be subjected to. Therefore, the discount factor is taken as $w_j^{(DB)} = DB_j$.

3) Silhouette Index: The silhouette index S_j [24] considers the compactness and separation evaluated on the individual level. For the construction of the 'silhouette width' s_{ij} the closeness of observation i to all observations within its group $k = y_i$ is measured via

$$\Delta_j^{Sil}(i,k) = \frac{1}{(n_k - 1)}\sum_{h:y_h=k,h\neq i}|x_{ij} - x_{hj}^{(k)}|, \qquad (9)$$

which is similar to the compactness measure in the DB index. However, Δ_j^{Sil} takes the closeness to each individual observation into account, instead of measuring the deviation from the mean.

Seperation between the groups is measured via,

$$\delta_j^{Sil}(i,k) = \min_{l\neq k}\left\{\frac{1}{n_l}\sum_{h=1}^{n_l}|x_{ij} - x_{hj}^{(l)}|\right\}, \qquad (10)$$

which takes the minimum average distance to the members of any other class. The silhouette width s_{ij} combines compactness and separation which leads to

$$s_{ij} = \frac{\delta_j^{Sil}(i,k) - \Delta_j^{Sil}(i,k)}{\max\{\Delta_j^{Sil}(i,k), \delta_j^{Sil}(i,k)\}}. \qquad (11)$$

As a last step, the silhouette index S_j is calculated by averaging over the silhouette width s_{ij} of all N individuals,

$$S_j = \frac{1}{N}\sum_{i=1}^{N}s_{ij} \quad \in \quad [-1,1]. \qquad (12)$$

S_j which can be used as a global measure of clustering quality given the covariate j and the target classes.

The absolute silhouette index takes values close to 1, if all observations are compact within their groups and well separated to the other groups. The more the silhouette index S_j approaches 0, the less compact the observations are within their groups and the less separated among covariate j. In this case the groupings are not nicely decomposed, and therefore this covariate is considered as less decisive.

The higher the absolute value of the silhouette index of covariate j, the better the distinction of the two underlying groups. Covariates with a high absolute silhouette index should be penalized less, therefore we set $w_j^{(Sil)} = 1/|S_j|$.

IV. Empirical Comparison

In this section we first present the scRNA-sequencing benchmark data and test the performance of DP-Lasso with different choices for the DP against competing methods. For both the binary classification, described in Section IV-C and the multiclass classification, described in Section IV-D, we perform a 5−times repeated 10−fold cross validation. As the supervised model is based on the classes present in the training data, we can only predict the number of underlying classes that are part of the training data set, in contrast to unsupervised clustering models.

A. Single-cell RNA-sequencing data (ScRNA-Seq data)

Based on the paper of [28], we use the same single-cell RNA-sequencing datasets as [16]. As proposed by [28], we only include genes into our analysis with read counts higher than 1 transcript per million mapped read (TPM) in more than 25% of the considered cells. This leads to a differing number of covariates p in case of the binary classification and the multiclass classification task, as shown in Table I. For the choice of cell types, we use the same selection as [16]. In case of the binary response, two selected cell types will be analyzed (left side of Table I). In case of the multiclass classification task (right side of Table I), we analyze K cell populations. The underlying numbers of cells in case of the binary response ($K = 2$) are N_1 and N_2, and for the multiclass response ($K > 2$) the respective cell populations are denoted with $N_1, \cdots, N_K$.

In accordance to the paper of [16], we consider their proposed binary classification tasks. However, instead of their approach of all pairwise combinations, we use a multinomial model for the $K > 2$ cases, which means one model per dataset. In the following, the cell types of the analyzed single-cell RNA-sequencing datasets are described. The EMTAB2805 data of [5] contain the cell cycle stages *G1, S, G2M* of the mouse embryonic stem cell (mESC). For the dataset GSE45719 [7] we include the different states of transition of *mid blastocyst*,

TABLE I
BENCHMARK DATA, SHOWING THE NUMBER OF COVARIATES p, NUMBER OF OBSERVATIONS N, AND THE OBSERVATIONS PER CLASS N_1 VS. N_2 IN THE BINARY CLASSIFICATION TASK AND N_1 VS. N_2 VS. $\cdots$ VS. N_K IN THE MULTICLASS CLASSIFICATION TASK

	Binary Response				Multiclass Response			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
p	13,110	10,851	7,987	6,748	12,849	11,065	7,831	7,329
Subpopulation 1	*G1*	*mid blastocyst*	BMDC 1h LPS	NKT0	*G1*	*mid blastocyst*	BMDC 1h LPS	NKT0
N_1	96	60	96	45	96	60	96	45
Subpopulation 2	*G2M*	*16-cell stage embryo*	BMDC 4h LPS	NKT17	*G2M*	*16-cell stage embryo*	BMDC 4h LPS	NKT17
N_2	96	50	191	44	96	50	191	44
Subpopulation 3	-	-	-	-	*S*	*8-cell stage embryo*	*BMDC 6h LPS*	*NKT1*
N_3	-	-	-	-	96	37	191	46
Subpopulation 4	-	-	-	-		-	-	NKT2
N_4	-	-	-	-	-	-	-	68

8-cell stage embryo as well *16-cell stage embryo*. In case of the single-cell RNA-sequencing data of GSE48968 bone marrow-derived dendritic cells (BMDCs) were stimulated with three different pathogenic components, analyzing the different responses for the dataset [25]. We will analyze only the component Lipopolysaccharides (LPS) at different timepoints (*1h, 4h, 6h*) after incubation. The data set GSE74596 contains different types of Natural killer T (NKT) cells extracted from the thymus. The cell types *NKT1, NKT2 and NKT17* are subtypes of the helper T cells [9].

The objective is to determine a supervised model that can classify the different cell types, given the expression profiles in these datasets. Also, as a second objective it is important to find a sparse solution to focus on the most important genes.

B. Competing Methods

The L1-regularized regression is carried out with the R package *glmnet* [13]. The λ values are found via the internal 10-fold CV approach and chosen as the value λ leading to the smallest estimated generalization error. For adaptive Lasso, the covariate specific penalty weights are determined with ridge regression $w_j = 1/\hat{\beta}_j^{Ridge}$ due to the $p >> N$ situation.

We also compare our methods to the Elastic Net, as a baseline for good predictive performance. Elastic Net is fit using *glmnet* and $\alpha = 0.5$, leading to a equal mixture of L1 and L2-penalization (cf. Section II).

For DP-Lasso the ANOVA based DP weights are implemented with the R package *stats* [23]. The Silhouette Index is calculated with the R package *cluster* [18] and the Davies-Bouldin Index with the package *clusterSim* [34]. The final DP-Lasso model is again fit using the *glmnet* procedure, using the covariate specific penalty weights derived from the DP.

C. Binary classification

In this section the results for the experiments on binary classification tasks are presented and analysed.

1) Accuracy – Binary: Accuracy is measured in terms of the misclassfication rate, averaged over all folds. The results of the empirical comparison can be found in Table II. Elastic Net shows overall the lowest misclassifcation rate, however the difference to the DP-Lasso models and the normal Lasso is only marginal. The only exception is the adaptive Lasso, which performs clearly worse compared to the other methods. This is likely due to the strong correlation present in the data. The three proposed DP-Lasso model show only minor differences in terms of misclassification rate, with a slight advantage for DP-L$_{ANOVA}$. We conclude, that the accuracy of DP-Lasso is comparable to the competitors irrespective the choice of the discriminative power.

2) Number of Coefficients – Binary: If the primary objective is to identify biomarkers, it is very important to find sparse solutions, as the cost of follow up studies can be high. Next, we therefore analyse the number of covariates selected by each method, which is the number of non-zero coefficients left in the regularized model. Out of all methods, the Elastic Net (Enet) selects the highest number of covariates, which is expected, due to its part of L2-penalty.

All DP-Lasso models select significantly fewer covariates than the competing methods, on all binary classification tasks. Often the difference is quite large. For example on the GSE74596 dataset DP-L$_{Anova}$ selects only 4 covariates, whereas Lasso selects 18. An likely explanation is the over-shrinkage effect in Lasso regression, which takes in irrelevant predictors (cf. Section II). On the other hand, DP-L$_{Anova}$ is able to reduce the penalty on the important covariates and reaches a very sparse solution.

From the class of DP-Lasso models, DP-L$_{ANOVA}$ is the most selective and finds the sparsest solutions. However, DP-L$_{DB}$ and DP-L$_{Sil}$ also produce smaller model sizes compared to the competing methods on all binary classification tasks.

D. Multiclass Classification

DP-Lasso can also be applied for multiclass ($K > 2$) classification. Note, that in case of $K > 2$ and the multinomial-logit model $K - 1$ coefficient vectors β are fit for the different categories, whereas one category is used as reference category. DP is measured as before for each covariate, leading to an equal penalization for each of the outcome categories.

In contrast to the binary case, the adaptive Lasso uses a different penalization weight for each covariate and outcome

TABLE II
The misclassification rate for binary and multiclass classification on the four benchmark datasets. The best result on each dataset (lowest number) is marked in bold.

	Binary				Multiclass			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
Lasso	0.05 (0.006)	**0.01** (0.000)	**0.02** (0.003)	**0.00** (0.000)	**0.06** (0.010)	0.03 (0.009)	0.19 (0.100)	**0.01** (0.003)
Elastic Net	**0.04** (0.006)	**0.01** (0.000)	**0.02** (0.000)	**0.00** (0.000)	**0.06** (0.007)	**0.02** (0.005)	0.18 (0.008)	**0.01** (0.004)
adaptive Lasso	0.11 (0.008)	0.02 (0.000)	0.07 (0.007)	0.15 (0.031)	0.17 (0.006)	0.10 (0.013)	0.26 (0.010)	0.28 (0.015)
DP-L$_{ANOVA}$	0.05 (0.006)	**0.01** (0.000)	**0.02** (0.004)	**0.00** (0.000)	**0.06** (0.009)	0.11 (0.017)	**0.17** (0.009)	0.03 (0.006)
DP-L$_{DB}$	0.05 (0.009)	**0.01** (0.000)	**0.02** (0.004)	0.01 (0.001)	0.08 (0.007)	0.07 (0.016)	0.20 (0.014)	0.03 (0.006)
DP-L$_{Sil}$	**0.04** (0.006)	**0.01** (0.000)	0.04 (0.004)	**0.00** (0.006)	0.18 (0.018)	0.06 (0.008)	0.24 (0.011)	0.06 (0.013)

TABLE III
The number of selected coefficients for binary and multiclass classification on the four benchmark datasets. The best result (lowest number) on each dataset is marked in bold.

	Binary				Multiclass			
	EMTAB2805	GSE45719	GSE48968	GSE74596	EMTAB2805	GSE45719	GSE48968	GSE74596
Lasso	58 (1.9)	20 (0.4)	55 (0.9)	18 (0.6)	127 (3.5)	67 (1.0)	163 (5.5)	72(1.7)
Elastic Net	142 (1.8)	103 (1.1)	125 (1.2)	66 (0.5)	250 (13.1)	199 (1.5)	276 (10.2)	197 (1.9)
adaptive Lasso	38 (2.1)	13 (0.6)	48 (0.8)	27 (0.7)	65 (1.6)	36 (0.3)	84 (4.8)	52 (3.0)
DP-L$_{ANOVA}$	**17** (0.4)	**5** (0.1)	**19** (0.4)	**4** (0.2)	**45** (0.6)	**23** (1.2)	**70** (1.1)	**17** (0.5)
DP-L$_{DB}$	25 (0.9)	9 (0.1)	30 (0.3)	7 (0.1)	71 (1.3)	39 (0.8)	125 (1.6)	37 (0.3)
DP-L$_{Sil}$	22 (0.5)	9 (0.3)	36 (0.6)	8 (0.4)	181 (2.2)	32 (0.8)	172 (1.8)	90 (3.3)

category again resulting from the ridge estimator.

1) Accuracy – Multiclass: Accuracy is again measured as misclassification rate. The results can be found in Table II. Of all methods the Elastic Net shows the strongest predicitve performance, followed by the Lasso. the adaptive Lasso again performs clearly worse on all datasets in terms of accuracy. From the DP-Lasso models, DP-L$_{DB}$ is competitive on most datasets, and DP-L$_{ANOVA}$ remains competitive on three of the datasets and shows significantly worse performance on the GSE45719 data. DP-L$_{Sil}$ performs worse overall in the multinomial setting, but still notably better than the adaptive Lasso.

2) Number of Coefficients – Multiclass: In terms of model size, DP-L$_{ANOVA}$ again uniformly produces the sparsest solutions on all datasets. Lasso and Elastic Net keep around 3 to 10 times more non-zero coefficients in the model respectively. DP-L$_{DB}$ also produces relatively small models, on par with the adaptive Lasso, whereas DP-L$_{Sil}$ clearly struggles on the EMTAB2805, GSE48968 and GSE74596 datasets.

E. Empirical Results Summary

The empirical comparison on benchmark data indicates that DP-Lasso is able to maintain a high accuracy. At the same time DP-Lasso finds significantly smaller models, often by a factor of 3 to 10 compared to Lasso and Elastic Net. This is due to the incorporation of the DP into the penalization scheme, which helps to remove uninformative genes and instead focus on the relevant ones. To summarise, DP-Lasso and especially DP-L$_{ANOVA}$ produces significantly smaller model sizes, while being able to maintain accuracy on par with current state-of-the-art regularized regression approaches.

V. Simulation Study

In this section, we test our method on simulated data. The setup is as follows. $X_1, ..., X_{10}$ are drawn from a normal distribution $\mathcal{N}(-1, \sigma)$, for observations of class 1, and from $\mathcal{N}(1, \sigma)$ for observations of class 2. This reflects the assumption that relevant genes express differently between the target groups. All additional covariates $X_{11}, ..., X_p$ are drawn from $\mathcal{N}(0, \sigma)$ and can therefore be considered as irrelevant. We test the values $p \in \{100, 1000, 5000\}$ and $\sigma^2 \in \{1, 2, 3\}$ and draw $N = 100$ observations in each setting. With increasing σ the groups become more overlapping and we expect learning to become increasingly difficult. Note that the covariates are drawn independently, implying $X \sim \mathcal{N}_p(\mu, \sigma^2 \mathcal{I}_p)$, where $\mathcal{I}$ is the identity matrix, making it an ideal situation for all methods. Each experiment is repeated 10 times and the results averaged. As in this experiment the relevant covariates are known, we measure the methods capabilities to identify the decisive covariates. To this end, we measure the Precision as

$$\text{Precision} = \frac{||\hat{\beta}_{true}||^0}{||\hat{\beta}||^0}, \tag{13}$$

where $||\cdot||^0$ specifies the 0-norm, which counts up the non-zero entries and $\hat{\beta}_{true}$ denotes the first ten entries of the coefficient vector, which by design we know to be the correct effects. $\hat{\beta}$ denotes all coefficients obtained by the regularized model. This measure is useful as the number of potential covariates is

TABLE IV
THE PRECISION AND RECALL ON THE DIFFERENT SIMULATION SETTINGS, AVERAGED OVER 10 RUNS. RESULTS ARE PRESENTED AS PRECISION / RECALL. FOR EACH SETTING THE METHOD WITH THE HIGHEST PRECISION IS MARKED IN BOLD.

	$\sigma^2 = 1$			$\sigma^2 = 2$			$\sigma^2 = 3$		
	$p = 100$	$p = 1000$	$p = 5000$	$p = 100$	$p = 1000$	$p = 5000$	$p = 100$	$p = 1000$	$p = 5000$
Lasso	0.86 / 0.99	0.60 / 0.99	0.53 / 0.98	0.45 / 0.96	0.32 / 0.96	0.23 / 0.93	0.48 / 0.95	0.37 / 0.84	0.28 / 0.84
Elastic Net	0.55 / 1.00	0.27 / 1.00	0.20 / 1.00	0.29 / 1.00	0.15 / 1.00	0.10 / 0.98	0.37 / 0.99	0.20 / 0.96	0.14 / 0.91
adaptive Lasso	0.99 / 0.97	0.97 / 0.98	0.94 / 0.95	0.88 / 0.98	0.58 / 0.96	0.37 / 0.91	0.71 / 0.93	0.35 / 0.82	0.28 / 0.85
DP-L$_{ANOVA}$	**1.00** / 0.87	**1.00** / 0.92	**1.00** / 0.85	**0.99** / 0.95	**0.88** / 0.93	**0.80** / 0.91	**0.82** / 0.87	**0.50** / 0.83	**0.38** / 0.85
DP-L$_{DB}$	**1.00** / 0.95	**1.00** / 0.94	**1.00** / 0.92	0.92 / 0.98	0.77 / 0.94	0.50 / 0.91	0.71 / 0.94	0.35 / 0.85	0.28 / 0.84
DP-L$_{Sil}$	**1.00** / 0.94	**1.00** / 0.94	**1.00** / 0.91	0.96 / 0.98	0.76 / 0.93	0.67 / 0.90	0.63 / 0.88	0.41 / 0.79	0.31 / 0.81

high. However, If the model has a high Precision, the identified genes can be trusted.

Secondly, we measure the Recall

$$\text{Recall} = \frac{||\hat{\beta}_{true}||^0}{10}, \tag{14}$$

as the fraction of the relevant covariates that was discovered by the model.

The results are shown in Table IV. We can see that the DP-Lasso models show significantly higher Precision compared to Lasso and Elastic Net. The adaptive Lasso performs better than the Lasso in this ideal setting, in contrast to the results on the real data from the previous section. Overall DP-L$_{DB}$ and DP-L$_{ANOVA}$ show the highest Precision, even in very difficult data situations. For instance, with $N = 100, p = 5000, \sigma = 1$, DP-L$_{ANOVA}$, DP-L$_{DB}$ and DP-L$_{Sil}$ are able to maintain a 100% Precision and thus are very selective and able to find the correct covariates. DP-L$_{ANOVA}$ has the highest Precision in every setting.

It is also important to compare the Recall, as it reflects the fraction of true effects that are found by a model. Elastic Net shows the highest Recall, which is a result of the large number of coefficients that was kept in the model. On the other hand, all DP-Lasso models show a Recall which is typically slightly lower but still competitive with Lasso and adaptive Lasso. This again is due to very selective nature of DP-Lasso.

Overall, we conclude that the non-zero coefficients found by the DP-Lasso can be trusted more to reflect true mechanisms, compared to its competitors. At the same time DP-Lasso is capable to maintain a competitive Recall.

It is reassuring to note that on average the accuracy measured by the area under the curve AUC of the methods is very similar, with a slight edge for the DP-L$_{DB}$, DP-L$_{ANOVA}$ and the Elastic Net.

VI. CONCLUSION

With DP-Lasso, we propose a novel regularization based approach for covariate selection in the context of gene expression data. Incorporating univariate measures of discriminative power that are based on the principles of separation and compactness enriches the model with additional information. Our approach can also be interpreted as soft filtering: instead of removing genes a-priori, more promising genes are simply promoted, freeing the modeller from ad-hoc choices, such as selecting the correct number of genes to remove. In a boarder context we argue that soft filtering, instead of hard filtering, therefore also enhances reproducibility, as it reduces the 'researchers degrees of freedom' [26] involved in a study.

Empirically, we show that DP-Lasso shows accuracy on par with the popular methods Lasso and Elastic Net, while choosing significantly less genes. With a simulation study we confirm that DP-Lasso is capable of ignoring a large number of irrelevant predictors and instead focusses on the truly relevant ones – due to the double criteria of being relevant both univariatly and in the multivariate model. This selectiveness is very desirable in the context of gene expression data, as both the number of candidate genes is high and follow-up studies are costly. Therefore, a short but confident list of very promising genes, as given by the DP-Lasso model, is preferred in this context.

As currently the discriminative power is calculated univariately, it does not explicitly take the correlation structure ofthe covariates into account. An interesting direction for future work would therefore be to extend the DP-Lasso approach by taking the correlation structure between covariates into account and adjust the penalization accordingly, similar to the approach in [33].

In this article, we focussed on the application for genetic classification data, however DP-Lasso can also be applied in other domains. As long as the classes are expected to show differences in the univariate distribution of covariates, we expect DP-Lasso to deliver a good predictive performance coupled with a low number of selected covariates.

ACKNOWLEDGEMENTS

The first authors are very grateful for the support of the LMU mentoring program, connecting young researchers and providing mentors that give individual advice. In addition, we would like to thank Gerhard Tutz and Christian L. Müller for the very insightful and valuable discussion.

References

[1] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[2] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

[3] Nadia Bolshakova and Francisco Azuaje. Cluster validation techniques for genome expression data. *Signal Processing*, 83(4):825–833, 2003.

[4] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[5] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 2015.

[6] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 224–227, 1979.

[7] Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014.

[8] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1–13, 2006.

[9] Isaac Engel, Grégory Seumois, Lukas Chavez, Daniela Samaniego-Castruita, Brandie White, Ashu Chawla, Dennis Mock, Pandurangan Vijayanand, and Mitchell Kronenberg. Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nature Immunology*, 17(6):728–739, 2016.

[10] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer, 2007.

[11] Liang Fang and Cheng Su. *Statistical Methods in Biomarker and Early Clinical Development*. Springer, 2019.

[12] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in Statistics*, pages 66–70. Springer, 1992.

[13] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[14] Joyee Ghosh, Yingbo Li, and Robin Mitra. On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383, 2018.

[15] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of statistical learning: Data mining, inference, and prediction*. Springer, 2017.

[16] Beyrem Khalfaoui and Jean-Philippe Vert. DropLasso: A robust variant of Lasso for single cell RNA-seq data. *hal-01716704v2*, 2019.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[18] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2019. R package version 2.1.0.

[19] M Man, TS Nguyen, C Battioui, and G Mi. Predictive subgroup/biomarker identification and machine learning methods. In *Statistical Methods in Biomarker and Early Clinical Development*, pages 1–22. Springer, 2019.

[20] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[21] Masanori Oshi, Hideo Takahashi, Yoshihisa Tokumaru, Li Yan, Omar M Rashid, Ryusei Matsuyama, Itaru Endo, and Kazuaki Takabe. G2m cell cycle pathway score as a prognostic biomarker of metastasis in estrogen receptor (er)-positive breast cancer. *International Journal of Molecular Sciences*, 21(8):2921, 2020.

[22] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[24] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[25] Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

[26] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[27] Manabu Soda, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina, Hisashi Hatanaka, et al. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153):561–566, 2007.

[28] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, 2018.

[29] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):1–11, 2008.

[30] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

[31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[32] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[33] Gerhard Tutz and Jan Ulbricht. Penalized regression with correlation-based penalty. *Statistics and Computing*, 19(3):239–253, 2009.

[34] Marek Walesiak and Andrzej Dudek. The choice of variable normalization method in cluster analysis. In *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges*, pages 325–340. International Business Information Management Association (IBIMA), 2020.

[35] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[36] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.