Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Feature Selection Under Dimensionality Imbalance

J. Cabrera[a]* B. Emir,[b] Y. Cherkas,[c] and D. Alemeyahu,[b]

We describe an approach for combining and analyzing high dimensional genomic and low dimensional phenotypic data. The approach leverages a scheme of weights applied to the variables instead of observations and, hence, permits incorporation of the information provided by the low dimensional data source. This approach can be incorporated into commonly used downstream techniques, such as penalized regression. The approach is illustrated on a simulated lupus study involving genetic and clinical data.

Keywords: Model selection; penalized regression; dimension reduction, precision medicine

1. Introduction

Precision medicine has presented a new and interesting analytical paradigm with the incorporation of genomic data into standard analysis of clinical data from patients for prognosis of diseases ([Aramburu et al., 2015]). Genomic data are characterized by high dimensionality, involving millions of SNPs or tens of thousands of gene expressions, while clinical and demographics data are typically of low dimension, consisting of at most hundreds of variables. Consequently, the importance of clinical variables is overpowered by high dimensional genomic data when combining information from both data sources. To adequately capture crucial clinical information, it is important to ensure that these clinical features have a chance to be selected in the presence of high dimensional genomic data. Methods proposed in this manuscript have properties that can be used to model such situations ([Yang et al., 2010]).

[Amaratunga et al., 2007], [Amaratunga et al., 2008], [Amaratunga et al., 2014] and [Cabrera and Yu, 2007] proposed a scheme that assigns weights to individual features based on the strength of the association among the features and the response. For instance, the weight may be calculated as a function of p-values, obtained from a suitable test criterion and corrected for multiplicity. A key principle underlying the approach is that the weights are applied to the variables, instead of the observations. The multiplicity correction is crucial to minimize the possibility of identifying spurious associations. Thus, if the association of the variable with the response is not by chance, the weight assigned to the variable would also be high. In this connection, because the information of the outcome is used in the determination of the weights, there may be potential for overfitting. This issue will be addressed below in our simulation experiments.

[Amaratunga et al., 2014] also proposed the method of enriched Principal Component Analysis (ePCA) that uses the weights to calculate the principal components of high dimensional data. Earlier, [Amaratunga et al., 2007] considered the application of an enriched method to unsupervised clustering and Amaratunga et al (2008) proposed a supervised algorithm for enriched random forest.

As pointed out above, the primary purpose of combining genomic data and clinical data for identifying a combination of clinical predictors and important SNPs that would produce a good fit and a good predictive model. Due to the high dimensionality of the SNP data, it is often the case that clinical variables are not selected regardless of the variable selection method. Thus, important clinical variables may be omitted despite the fact they may have useful information

[a] *Department of Statistics, Rutgers, The State University of New Jersey, 501 Hill Center 110 Frelinghuysen Road, Piscataway, NJ 08854, USA*
[b] *Pfizer Inc, 235 East 42nd Street, New York, NY, 10017, USA*
[c] *Statistics and Decision Sciences, Janssen R&D, Springhouse, PA 19002*

* *Correspondence to: Javier Cabrera, Department of Statistics, Rutgers, The State University of New Jersey, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA E-mail: cabrera@stat.rutgers.edu*

to facilitate interpretation of the results. Indeed, this is known to be a major issue, as a result of which a few SNPs are selected first, and then combined later with clinical data.

In this manuscript, we provide a general overview of the analytical landscape and offer attractive options to simultaneously use genetic and clinical information in association studies. In section 2, the standard analytical approaches to combine clinical and genomic data are reviewed. In section 3, the proposed enriched approach for combining high and low dimensional data is described in detail and compared to the standard techniques. Finally, the proposed approaches are illustrated using simulations involving genomic and clinical data from lupus subjects.

2. Current Approaches

The standard analytical approach to combine data from clinical and genomic sources is a stepwise process, typically consisting of a univariate screening stage, followed by multivariate modeling. The goal in the first step is to identify predictors (such as SNPs or genes) that are important to carry forward. There are several different ways to identify such predictors as described in the next section. Once the set of predictors is selected, they are combined with clinical variables and used in the final model. One of the disadvantages of the customary stepwise process is that once the first step is completed and the selection is finalized, this selection may not be optimal for the second step.

2.1. Stepwise Screening

A typical stepwise screening approach involves application of either correlation or prognostic filters to remove genes or SNPs that are not related to the outcome of interest. For each SNP, simple univariate tests, such as Students t, chi-squared or logistic regression, are applied and the associated p-values are obtained. In most applications, suitable corrections for multiplicity are used to eliminate spurious correlations. Once the SNPs are selected, they are combined with the clinical variables and analyzed using suitable multivariate modeling techniques, such as penalized or standard regression models.

Despite its widespread use, there are several limitations of stepwise screening. One of them is the high computational burden for the considerable number of univariate tests that only assess one variable (SNP) at a time. It also ignores the associations among the SNPs and clinical variables in modeling. Most importantly, this approach is prone to identifying spurious correlations even when the correction for multiplicity is stringently applied.

2.2. Multivariable Modeling

Multivariate techniques in contrast are able to assess multiple variables simultaneously. The multivariate approaches are more computationally feasible and give the possibility to simultaneously combine all SNPs and clinical data and then apply a model such as penalized regression to select relevant variables. Examples of such models that perform well for high dimensional data include, least absolute shrinkage and selection operator (LASSO) ([Tibshirani, 1996], elastic net ([Zou and Hastie, 2005], or other regularized regression methods (see, e.g., [Hastie et al., 2009]). While this approach works relatively well under high-dimensionality compared to the use of standard regression models, it still suffers from several limitations. First, as shown in [Cai and Guo, 2017] and [Verzelen, 2012], variable selection algorithms such as the LASSO are mostly successful in discovering strong SNP signals of dimension $n/log(p)$, where n is the sample size and p is the number of predictors. For example, if there are 100 patients with one million SNPs, at most seven SNPs with strong signals can be detected using the LASSO. Moreover, with this strategy clinical data are not optimally utilized due to dimension disparity. As a consequence of the preponderance of spurious signals, clinical data may not have a chance to be represented in the final model.

3. Enriched Approaches

It is essential to formulate an approach that permits combining both low dimensional clinical and high dimensional genomic data directly, while minimizing the impact of the discrepancy in the dimensions of the two data sets. In this section, we propose an algorithm that addresses the issue through judicious construction and use of suitably defined weights. It may be noted that in most analytical approaches weights are applied to observations rather than variables. In contrast, the proposed approaches involve application of weights to variables, rather than data points ([Amaratunga et al., 2014]).

3.1. *Weight Construction*

[Benjamini and Hochberg, 1995] proposed a method that controls the expected proportion of discoveries (i.e., rejected null hypotheses) that are false, also dubbed the false discovery rate (FDR). A related quantity is the so-called q-value, which corresponds to the minimum FDR that can be attained when calling a given association significant. The q-values are calculated either assuming that p-values follow a uniform distribution between 0 and 1 or directly from the distribution of the order statistics of the p-values. [Benjamini and Yekutieli, 2001] provided the idea of FDR corrected p-values that was popularized by [Storey, 2002] under the name of q-values which also provides an alternative approach to FDR calculation. In the present context, weights can be calculated by either taking the reciprocal of the q-values or $-log(q)$. For our purpose here we define the vector of weights ω as the vector $-log(q_i)$ were i varies across all the variables in the dataset. Note again that the weights relate to the variables, and not to the observations. Obviously, these weights incorporate the strength of the univariate relationship between the outcome and the predictor after correcting for false discovery. High weights also indicate the presence of non-spurious relationships. This is shown by applying a random permutation to the outcome vector so the association of the permuted outcome and the SNPs is explained by chance. Therefore the q-values will all be large resulting in very similar low weights and no enrichment effect. On the other hand high weights indicate that those SNPs have a strong association with the outcome beyond what is expected by chance.

The reason for using q-values and not p-values is to avoid overfitting. Overfitting is a concern because, in order to calculate the weights, the enriched method makes use of the response and therefore is semi-supervised. However in the case that all the predictors were unrelated to the response the q-values will be all close to 1 and therefore the weights will be small but approximately equal. The resulting model will be approximately the same as when no weights are used. To illustrate this one could compare the observed weights to the values obtain when the response is permuted.

3.2. *Enriched Penalized Modeling*

As a direct approach for combining all p SNPs and m clinical variables, it is proposed that weights be calculated as mentioned previously, i.e., separately for all SNPs and clinical variables, as follows: $\omega_j^s = -log(q_j)$, for all SNPs $j = 1, \cdots, p$ and $\omega_k^c = -log(q_k)$, for all clinical variables $k = 1, \cdots, m$. We then allocate a predetermined distribution of the weights between the set of clinical variables and the set of SNPs. For example, this could be an equal allocation, with $\sum_{j=1}^{p} \left(\omega_j^s\right)^2 = \sum_{k=1}^{m} \left(\omega_k^c\right)^2 = 0.5$. In general, the allocation may be determined such that $\kappa = \sum_{j=1}^{p} \left(\omega_j^s\right)^2$ and $\sum_{k=1}^{m} \left(\omega_k^c\right)^2 = 1 - \kappa$. The combined analysis involving both the SNP and clinical variables may then be performed in a single model using suitable penalized regression approaches mentioned in Section 2.2. In this case, the reciprocals of variable weights κ and $1 - \kappa$ are included in the penalty (Equation 1). It is noted that this is equivalent to multiplying each standardized variable by its weight and using the standard algorithms (Equation 2)[Hastie et al., 2009].

$$\tilde{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p+m} x_{ij}\beta_j \right)^2 + \lambda \left(\sum_{j=1}^{p} \frac{1}{\omega_j^s} |\beta_j| + \sum_{j=1}^{m} \frac{1}{\omega_j^c} |\beta_{p+j}| \right) \right) \tag{1}$$

$$= \arg\min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \omega_j^s x_{ij} \beta_j^* - \sum_{j=1}^{m} \omega_j^c x_{i(p+j)} \beta_{p+j}^* \right)^2 + \lambda \sum_{j=1}^{p+m} |\beta_j^*| \right) \tag{2}$$

In order to implement the elastic net method the lasso penalty (Equation 2) is replaced by the elastic net penalty (Equation 3).

$$\tilde{\tilde{\beta}} \arg = \min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \omega_j^s x_{ij} \beta_j^* - \sum_{j=1}^{m} \omega_j^c x_{i(p+j)} \beta_{p+j}^* \right)^2 + \lambda \sum_{j=1}^{p+m} (\alpha {\beta_j^*}^2 + (1-\alpha)|\beta_j^*|) \right) \tag{3}$$

For choosing an optimal allocation of weights, it is proposed that a modeling scheme be followed that involves initially fitting the enriched regularized model for different values of κ. This could be done, for example, in increments of 0.1 on the interval $[0, 1]$. For each fit, the cross-validated residual sum of squares (CV_{MSE}) is calculated ([Hastie et al., 2009]). Then we plot CV_{MSE} versus κ , and choose the smallest value of κ where the CV_{MSE} nearly flattens. It is expected that the CV_{MSE} will be smallest when κ is equal to 1, which corresponds to giving 0 weight to the clinical variables. It is also likely that for smaller values of κ the corresponding CV_{MSE} would tend to be approximately the same. Therefore, we will choose a κ value smaller than 1 that gives some weight to the clinical variables, but sufficiently large that the corresponding CV_{MSE} is approximately the same as the optimal value, thereby having a final model which combines a subset of each data set.

4. Illustration of Approaches

To illustrate the proposed methods, we used data based on a clinical trial with active systemic lupus erythematosus (SLE) in which the outcome of interest was the occurrence of flares. Two groups of subjects, each of size 50, were generated, following the distribution of the clinical trial groups, one with flares and the other without flares. For each subject, we obtained $20,000$ SNPs, following the SNP multivariate distribution estimated from the original data, of which 5% had a mild signal above the noise level (i.e., $p < 0.01$ for the association) and 0.1% had a moderate-to-strong signal ($p < 0.0001$). In addition, eight clinical variables were generated (including age, gender, BMI, smoking, baseline SLE activity scores, anti-double-stranded DNA (anti-dsDNA) score, British Isles Lupus Assessment Group (BILAG) and Composite Lupus Assessment (BICLA), again reflecting the distributions of the clinical data. Age, gender, BMI, smoking and baseline SLE activity were selected to have association with the binary outcome variable i.e., presence or absence of flares. However the other variables were not associated to the response. The correlation structures were set to be the same as those observed in the clinical trial.

We coded the SNP values as categories $\{0, 1, 2\}$, as described earlier, and applied an enriched penalized analysis to the dataset. Each SNP was represented by two dummy variables corresponding to values 0 and 1. The analysis was performed using a weighted elastic net model, as implemented by the $glmnet$ procedure in R. Figure 1 shows a plot of CV_{MSE} vs various allocation of weights, κ. As shown in Figure 1(a), the values of CV_{MSE} did not decrease monotonically, but fluctuated up and down. This is due to the variability in selection of sub samples used for cross validation as implemented in $glmnet$. The fluctuations in the graph give an idea of the variability of CV_{MSE} and is useful to detect the region were it plateaus. The CV_{MSE} appeared to stabilize for values of κ greater than 0.3; therefore we chose $\kappa = 0.5$. It is noted that this value of κ appeared to give much higher weight to each clinical variables than it did to each SNP. As a result three clinical predictors remained in the final model (Figure 1(b)) as well as several SNPs (Figure 1(c)). On the other hand, assigning equal weights to all variables, which corresponds to a model with $\kappa = 0.999$, excluded all clinical variables.

5. Simulation Study

We performed a simulation study using the same SNP dataset as in the above example, with the 100 observations on 20,008 variables. The response was generated using the model fitted in the example with mixing proportion $\kappa = 0.5$. For the simulation, we computed the probabilities $P_i = P(Y = 1 \mid \mathbf{X_i})$, which were used to generate simulated responses from a Bernoulli $(P_i), (i = 1, \cdots, 100)$. Each of the three methods was then applied to the simulated data: (i) Standard LASSO, (ii) Principal Component Analysis (PCA) + LASSO, and (iii) Enriched LASSO using the log $q-$value weights. The simulation was then repeated 500 times.

The relative performances of the approaches were evaluated using MSE and the average bias squared, in both the raw and logit scale. The results, displayed in Table 1, show that the first two methods are very similar, as pointed by [Cai and Guo, 2017]. The LASSO performance is poor due to the high dimensionality of the SNP data relative to the sample size. However, this issue is addressed effectively by the enriched LASSO through the use of weights.

Similarly, Figures 2 and 3 show a comparison of the true probabilities and the average predicted probabilities over the 500 simulations. It is clear that the enriched method performs better than the other two approaches. In fact, PCA did not make any difference.

To assess whether the enriched LASSO does not overfit due to data snooping, we scrambled the 500 responses of the simulation by applying a random permutation. The enriched LASSO method did not find any signal and showed average fitted values very similar to the regular LASSO.

6. Concluding Remarks

This paper presents a novel framework for combining high dimensional genomic with low dimensional clinical data. The approach gives due importance to the clinical variables using data driven weights. Compared to conventional approaches, the proposed method gives the clinical variables a better chance to be included in the final model without much loss in optimality. This is particularly important in situations where it is desirable to maximize the information from phenotypic data in precision medicine.

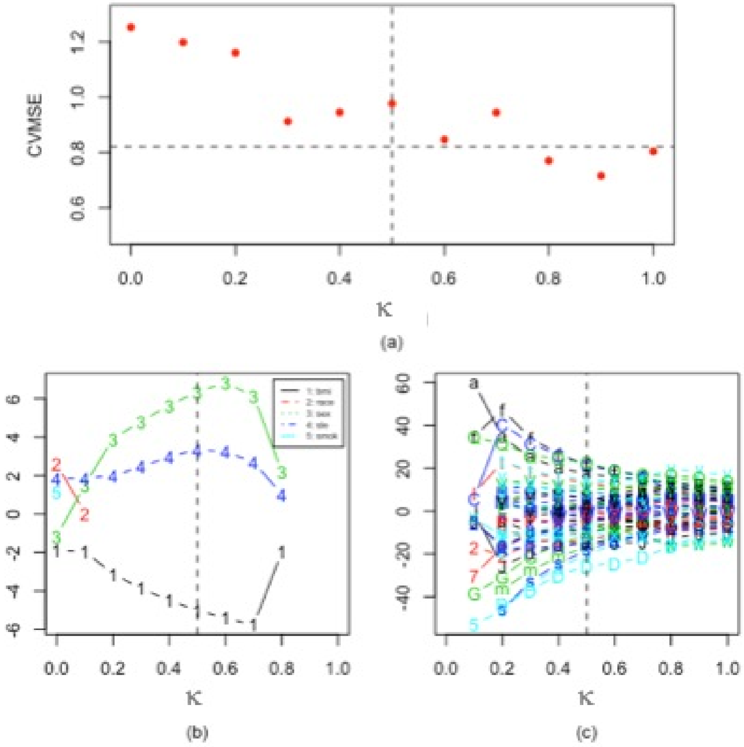

Figure 1. (a) Plot of CVMSE vs κ using a weighted elastic net algorithm. (b) Coefficients of clinical variables produced by the same fits as in (a) vs κ . (c) Coefficients of genomic variables produced by the same fits as in (a) vs κ

Table 1. Simulation Results: MSE and the average bias squared, in both the raw and logit scale for each model.

	Logit Scale		Raw Scale	
	Average Bias^2	MSE	Average Bias^2	MSE
Standard LASSO	3.62	4.047	0.065	0.082
PCA +LASSO	3.72	4.103	0.068	0.083
Enriched LASSO	0.803	2.784	0.017	0.057

The framework permits analysis of data including all variables in a regularized regression model or in a stepwise fashion. In the Enriched LASSO, the selected SNPs are combined with the clinical variables to produce a final model. Both approaches improve the chance of eliminating spurious signals in the genomic data and ensure incorporation of relevant clinical and genomic information in the final model.

The proposed methods can be implemented using R. Currently, the development of an R package is underway and will be submitted to CRAN in due course.

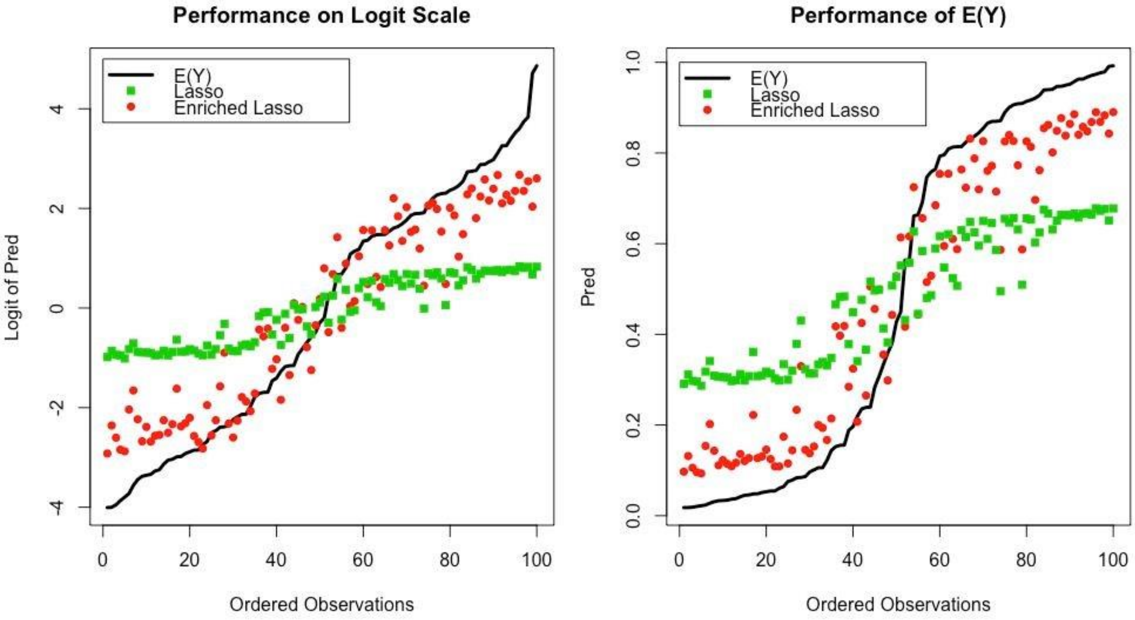

Figure 2. A comparison of the true predicted values to the average predicted values from of LASSO and Enriched LASSO over the 500 simulations

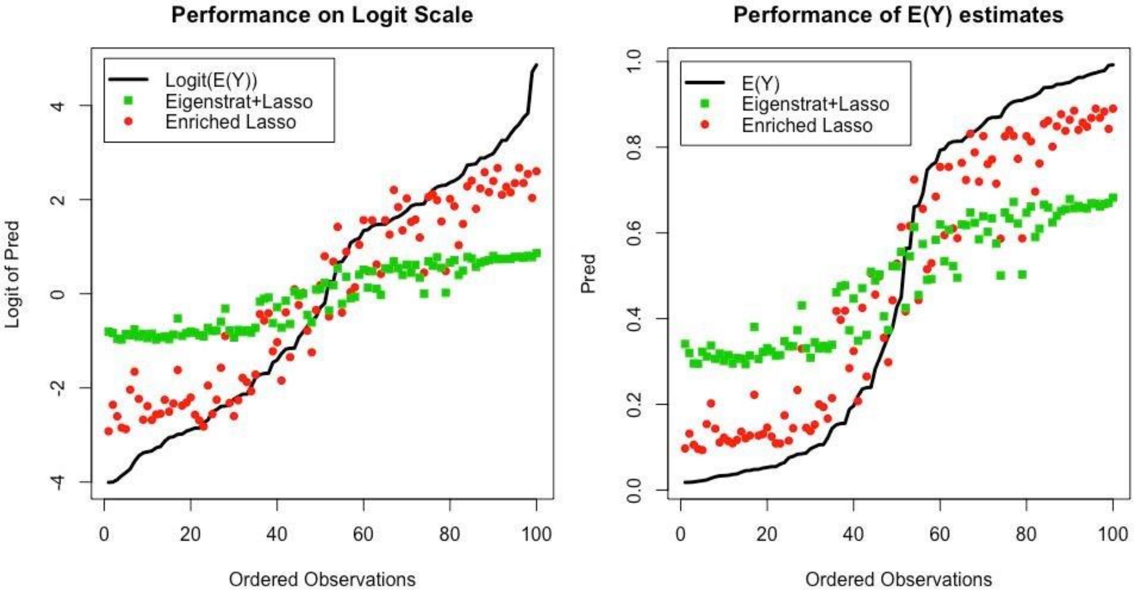

Figure 3. A comparison of the true predicted values to the average predicted values from of PCA + LASSO and Enriched LASSO over the 500 simulations

References

[Amaratunga et al., 2007]. Amaratunga, D., Cabrera, J., and Kovtum, V. (2007). Microarray learning with abc. *Biostatistics*, 9:128–136.

[Amaratunga et al., 2008]. Amaratunga, D., Cabrera, J., and Lee, Y. (2008). Enriched random forests. *Bioinformatics*, 24:2010–2014.

[Amaratunga et al., 2014]. Amaratunga, D., Cabrera, J., and Shkedy, Z. (2014). *Exploration and analysis of DNA microarray and protein array data*. John Wiley & Sons, New York, 2nd edition.

[Aramburu et al., 2015]. Aramburu, A., Zudaire, I., Pajares, M., Agorreta, J., Orta, A., Lozano, M., Grpide, A., Gmez-Romn, J., Martinez-Climent, J., Jassem, J., Skrzypski, M., Suraokar, M., Behrens, C., Wistuba, I., Pio, R., Rubio, A., and Montuenga, L. (2015). Combined clinical and genomic signatures for the prognosis of early stage non-small cell lung cancer based on gene copy number alterations. *BMC Genomics*, 16:752.

[Benjamini and Hochberg, 1995]. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *The Journal of the Royal Statistical Society, Series B, Statistical methodology*, (57):289–300.

[Benjamini and Yekutieli, 2001]. Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 24:1165–1188.

[Cabrera and Yu, 2007]. Cabrera, J. and Yu, C. (2007). Estimating the proportion of differentially expressed genes in comparative dna microarray experiments. *IMS Lecture Notes-Monograph Series*, 54:92–102.

[Cai and Guo, 2017]. Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, 45:615–646.

[Hastie et al., 2009]. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics, 2nd edition.

[Storey, 2002]. Storey, J. (2002). A direct approach to false discovery rates. *The Journal of the Royal Statistical Society, Series B, Statistical methodology*, 64:479 – 498.

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

[Verzelen, 2012]. Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

[Yang et al., 2010]. Yang, H., Yang, Y., Zhou, B., and Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5:296–308.

[Zou and Hastie, 2005]. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *The Journal of the Royal Statistical Society, Series B, Statistical methodology*, 67:301–320.