An Enriched Approach to Combining High-dimensional Genomic and Low-Dimensional Phenotypic Data

September 16, 2018

Abstract

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 EIGENSTRAT, penalized regression. The approach is illustrated on a simulated lupus study involving genetic and clinical data.

Keywords: EIGENSTRAT, Enriched EIGENSTRAT, 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. Ensemble 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. (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.

As pointed out above, the primary purpose of combining genomic data and clinical data is both for identifying important SNPs and for predicting response. 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 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. And finally, it doesn't have the ability to directly model ancestry information, which, as described below, could introduce potential bias in study results.

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. Finally, this approach also ignores important information such as ancestry, thereby introducing bias resulting from genetic heterogeneity.

2.3 EIGENSTRAT

Although human interactions face a lot of barriers (e.g., geographical, political, cultural, and religious, etc.), there is a remarkable degree of homogeneity among human beings, reflecting a common ancestry from Africa (Li et al. (2003)). However, there is a systematic difference in allele frequencies between subpopulations, possibly due to different ancestry, especially in the context of genome-wide association studies (Pritchard and Rosenberg (1999)). Analyses involving population stratification have the tendency to reduce the power of tests to detect true effects, and may also generate association findings that are spurious. Accordingly, it is essential to account for ancestry imbalance to minimize the reporting of spurious associations. EIGENSTRAT is a modeling attempt to correct for ancestry (see, e.g., Zhang et al. (2003); Patterson et al. (2006); and Carlson et al. (2006)). The procedure applies principal component analysis (PCA) to the SNPs coded as $0, 1$ or 2, with the values corresponding to homozygous, heterozygous, and wild type, respectively. After standardizing the data, i.e., subtracting the mean and dividing by the standard deviation, and recoding missing genotype as 0, PCA is applied to the matrix of standardized SNP values. As suggested by Patterson et al. (2006), typically the top k (where k is less than say 15) principal components (PCs) are related to ancestry and represent the strongest signal in the SNP data. The next few PCs should contain the signal related to the outcome, and the rest should be spurious or other non-specific signals. The top k PCs are then extracted and used to correct the SNPs for ancestry. In typical EIGENSTRAT applications the remaining $n - k$ PCs are inspected to determine whether there is an association with the response. If such an association exists, the relevant SNPs forming the PC are selected for further study to determine potential biomarkers.

2.4 Ancestry Corrected Modeling

Another application of the EIGENSTRAT approach is to perform regression analysis on ancestry-corrected data. More specifically, suppose there are n subjects and p SNPs. Let $\mathbf{S}$ be the $n \times p$ matrix of standardized SNPs, and $\mathbf{s_i}'$ be the transpose of $\mathbf{s_i}$, the ith row of $\mathbf{S}$. Let $\mathbf{l_i}$ be the corresponding $p \times 1$ PC loadings for $i = 1, \cdots, n$. It is noted that since $p > n$ there are only n PCs. The ancestry-corrected SNP vectors $\mathbf{s_i^\star}$ of size $p \times 1$ are given by:

$$\mathbf{s_i^\star = s_i' - \sum_{j=1}^{k} \left(l_j l_j'\right) s_i' \quad i = 1, \cdots, n} \tag{1}$$

Similarly, an analogous correction for ancestry is performed on the response $\mathbf{y}$ by projecting $\mathbf{y}$ onto the space spanned by the columns of $\mathbf{X}$, where $\mathbf{X} = \mathbf{SL}$, and $\mathbf{L}$ is the $p \times k$ matrix of factor loadings. Let η be a $k \times 1$ the vector of regression coefficients of $\mathbf{y}$ on $\mathbf{X}$. Then, $\mathbf{y}^\star$ is the residuals $\mathbf{y} - \hat{\mathbf{y}}$.

The stepwise approach discussed in Section 2.1 can also be applied to the ancestry-corrected $\mathbf{S}^\star$ and $\mathbf{y}^\star$.

Although this approach helps to address the issue of ancestral imbalance, it is associated with some of the drawbacks described earlier with the stepwise method, including the difficulty related to working with one SNP at a time, and the inability to take into account the association among SNPs and the clinical variables. Further, ancestry may be present in more than a few PCs and may not be completely orthogonal to the outcome variable. By removing k PCs from each SNP as in (Equation 1) we may weaken the association between the SNP and the response and therefor the subsequent analysis may be weaker. In the following, we will discuss a more general approach that addresses some of the issues raised above, including ancestry imbalance and incorporation of the correlation structure of the data.

3 Enriched Approaches

As described above, although EIGENSTRAT attempts to minimize spurious relationships emanating from ancestry imbalances, it is executed in a stepwise fashion and is associated with major underlying issues. It is, therefore, 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 two algorithms that address 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.

3.2 Enriched EIGENSTRAT

Once suitably defined weights are constructed, the EIGENSTRAT approach may be enriched by applying the weights to each SNP. More specifically, let $\mathbf{W}$ be the $p \times p$ diagonal matrix of ω, the p-dimensional vector of weights. Let $\mathbf{S}^\dagger = \mathbf{SW}$, where $\mathbf{S}$ is the $n \times p$ matrix of standardized SNP values. The enriched EIGENSTRAT approach consists of applying PCA to $\mathbf{S}^\dagger$. However, in order to find interesting SNPs, we go directly to the first few PCs, without the need to correct for ancestry. For the latter, any ancestry informations not related to the response will be down-weighted. In other words, the enriched EIGENSTRAT projects the SNPs into the first few important weighted PCs, whereas the standard EIGENSTRAT corrects for ancestry by projecting into the space orthogonal to the first few standard PCs.

Clearly, an attractive feature of this approach is that not only does it incorporate the correlation structure of the data and allow correction for ancestry imbalance, but also ensures that features from a high dimensional data source do not unduly overwhelm the importance of those from sources of smaller dimensions. Further, the incorporation of weights that are functions of q-values minimizes the potential for spurious correlations.

3.3 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 2). It is noted that this is equivalent to multiplying each standardized variable by its weight and using the standard algorithms.

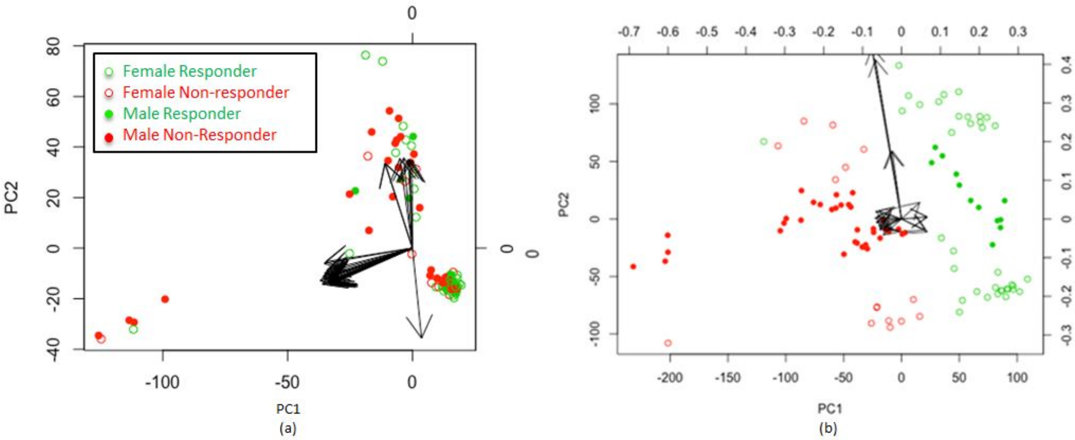
$$\begin{aligned}\tilde{\beta} &= \arg\min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p+m} x_{ij}\beta_j \right)^2 + \lambda \left(\sum_{j=1}^{p} \frac{1}{\omega_j^s} |\beta_j| + \sum_{j=1}^{m} \frac{1}{\omega_j^c} |\beta_{p+j}| \right) \right) \\ &= \arg\min_{\beta} \left(\sum_{i=1}^{N}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

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 sufficiently small that gives similar some weights to the clinical variables but sufficiently large that the corresponding CV_{MSE} is approximately the same as the optimal and genomic variables, 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 $\{0, 1, 2\}$, as described earlier, and applied first the EIGENSTRAT and enriched EIGENSTRAT approaches. The results are depicted in Figure 1,



(a)

(b)

Figure 1: EIGENSTRAT identified three clusters with significant overlaps between flare and non-flare patients. The enriched EIGENSTRAT method resulted in the separation of gender and flare and non-flare patients, along the first two principal components

using the first two PCs. The EIGENSTRAT approach (Figure 1 (a)) identified three clusters corresponding to three different ancestries, as shown in the oval shapes. However, significant overlaps were apparent between flare and non-flare patients. On the other hand, the enriched EIGENSTRAT method (Figure 1 (b)) resulted in the separation of flare and non-flare patients by the 1st PC and of males and females by the 2nd PC. In figure 1(b) the large arrows in the direction of the first PC represent the main SNPs associated with the response whereas, the arrows in the direction of the second PC represent a few SNPs from the X chromosome that are associated with gender. In the data, we observe that the reason the second PC is related to gender is that of the gender imbalance between flares and non-flares patients.

It is noted that one could use alternative representations of the SNP coding when the homozygous is rare. For example, two dummy variables, X_1 and X_2, may be used to denote each SNP, for heterozygous and homozygous, respectively, while keeping wild type as the reference. The corresponding weights for X_1 and X_2, may be constructed using the p-values from suitable association tests and calculating the respective q-values as described earlier. As depicted in Figure 2, the recoded SNP values resulted in an even more apparent and interpretable separation of not only flares/non-flares, but also males and females along the

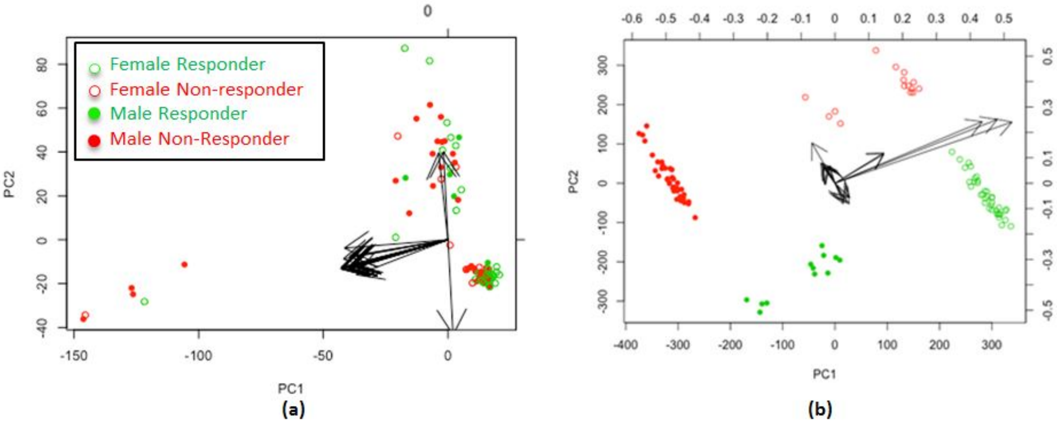


Figure 2: Coding SNP as two dummy variables to denote heterozygous and homozygous, respectively, while keeping wild type as the reference

first two principal components of the Enriched EIGENSTRAT. However, there appeared to be no substantial change in the results based on the EIGENSTRAT approach (Figure 2 (a)).

Finally, an enriched penalized analysis was applied to the dataset. The analysis was performed using a weighted elastic net model, as implemented by the *glmnet* procedure in R. Figure 3 shows a plot of CV_{MSE} vs various allocation of weights, κ. As shown in Figure 3(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 the clinical variables than it did to the SNPs. As a result three clinical predictors remained in the final model (Figure 3(b)) as well as several SNPs (Figure 3(c)). On the other hand, assigning equal weights to all variables, which corresponds to a model with $\kappa = 0.999$, excluded all clinical variables.

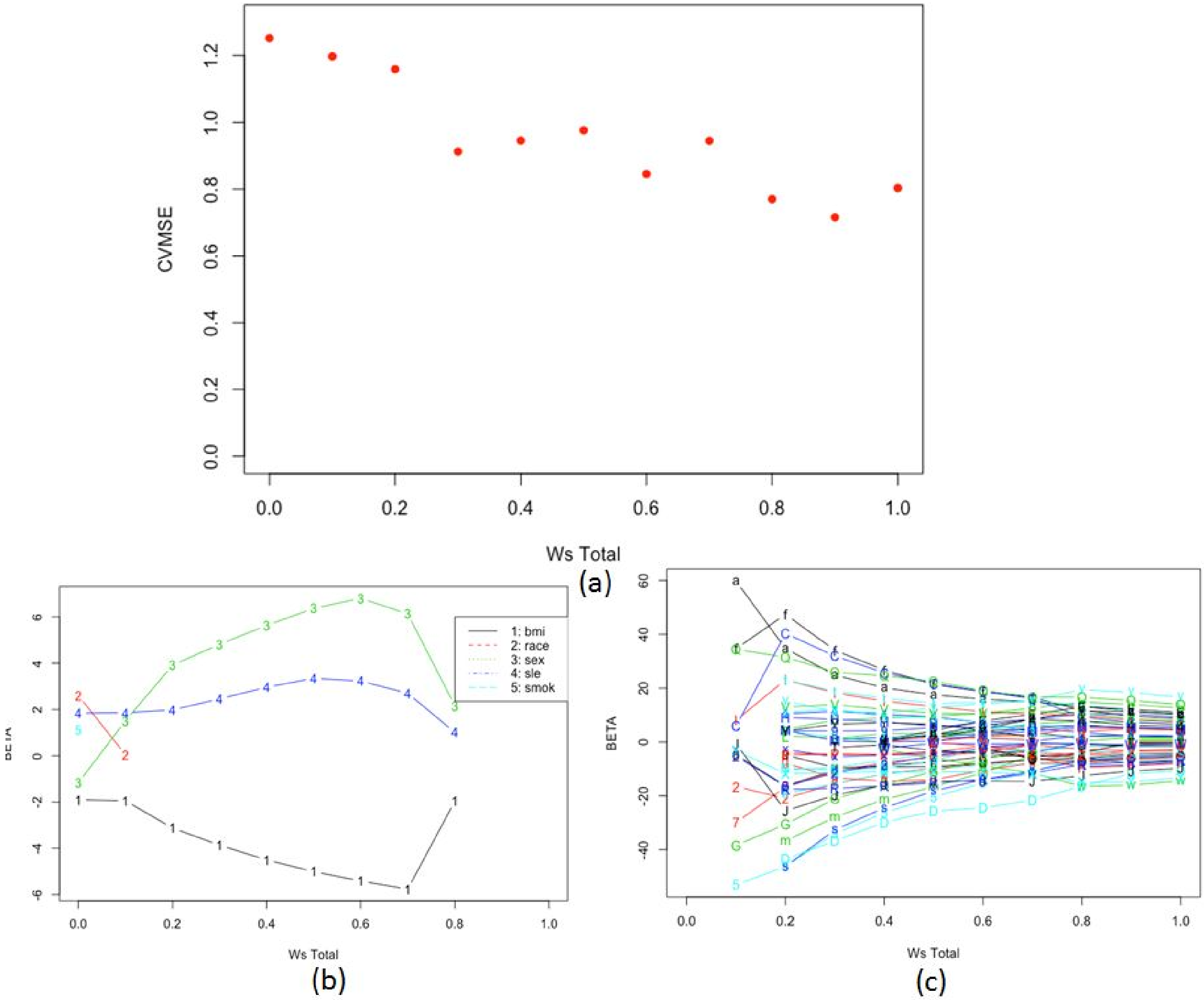


Figure 3: (a) Plot of CVMSE vs κ using a weighted elastic net algorithm. (b) Coefficients of clinical variables produced by the same fits as in (a) vs κ . (c) Coefficients of genomic variables produced by the same fits as in (a) vs κ

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) EIGENSTRAT + Lasso, and (iii) Enriched Lasso using the log $q-$value weights. To remove the ancestry signal, we subtracted the principal components that had an absolute correlation with the response of less than 0.2. 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 4 and 5 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, EIGENSTRAT 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

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
EIGENSTRAT +LASSO	3.72	4.103	0.068	0.083
Enriched LASSO	0.803	2.784	0.017	0.057

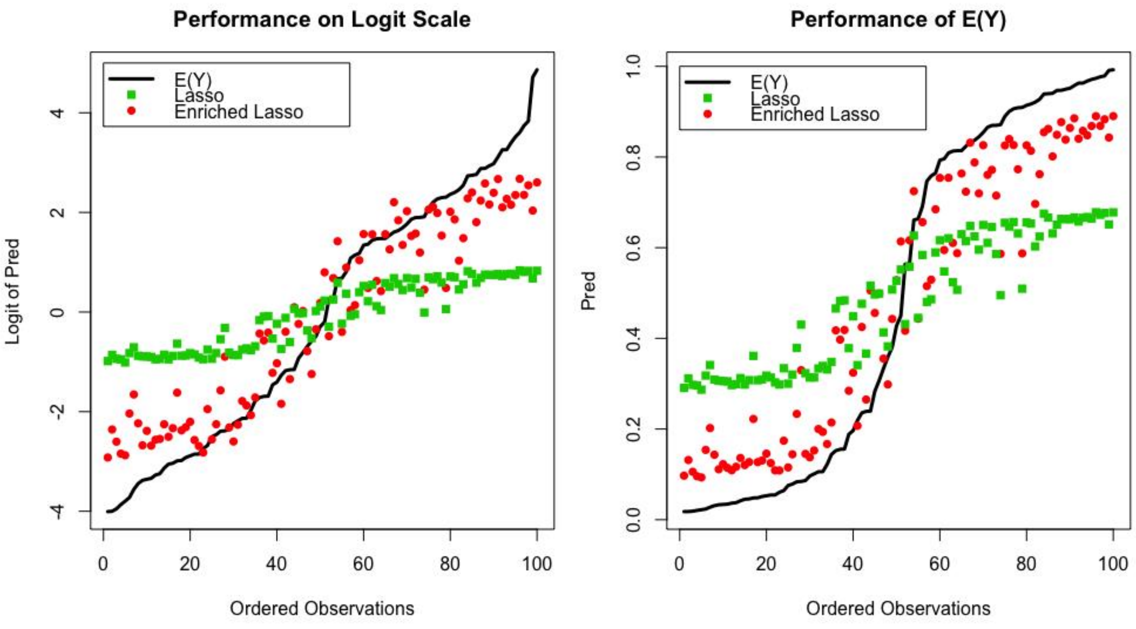


Figure 4: A comparison of the true predicted values to the average predicted values from of LASSO and Enriched LASSO over the 500 simulations

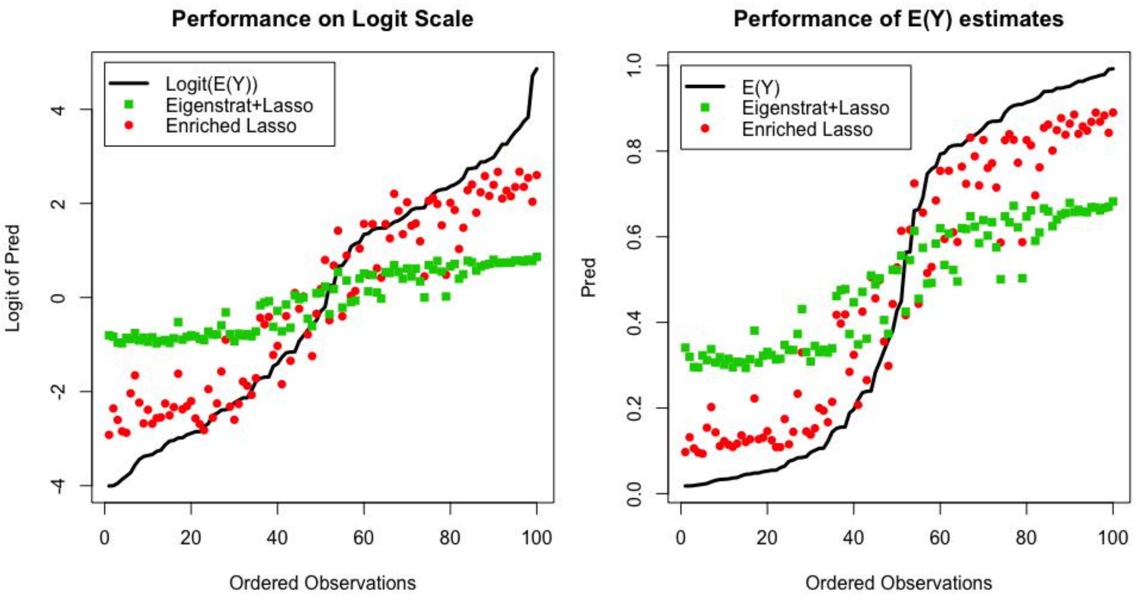


Figure 5: A comparison of the true predicted values to the average predicted values from of EIGENSTART + LASSO and Enriched LASSO over the 500 simulations

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.

The framework permits analysis of data including all variables in a regularized regression model or in a stepwise fashion. The latter involves use of a modified EIGENSTRAT procedure, called enriched EIGENSTRAT, in which a weighting scheme is incorporated in the usual EIGENSTRAT technique to select a subset of SNPs. In the final step, 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.

References

Amaratunga, D., Cabrera, J., and Lee, Y. (2007). Enriched random forests. *Bioinformatics*, 24:2010–2014.

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

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

Carlson, C., Eberle, M., Rieder, M., Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association. *Nature Genetics*, 38:904–909.

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

Li, J., Absher, D., Tang, H., Southwick, A., Casto, A., Ramachandran, S., Cann, H., Barsh, G., Feldman, M., Cavalli-Sforza, L., and Myers, R. (2003). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104.

Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigen analysis. *PLOS Genetics*, 2:12: e190.

Pritchard, J. and Rosenberg, N. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, 65:220–228.

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

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

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

Zhang, S., Zhu, X., and Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology*, 24:44–56.

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.