Linkage-Disequilibrium Regularized Support Vector Machines for Genome-Wide Association Studies

Mukund Sudarshan[1], **Lakshmi Subramanian**[1]
[1] Courant Institute of Mathematical Sciences, NYU
mukund@nyu.edu, lakshmi@cs.nyu.edu

Abstract

We exploit block-covariate structure in GWAS datasets to regularize linear classification models. Our algorithm first finds blocks of correlated SNPs, and adds a separate sparsity constraint for each one. When this regularization is added to the hinge-loss, we can achieve high test accuracy on a GWAS dataset, while still maintaining a very high significance level for each of the SNPs returned. Such an approach can potentially serve as a substitute for manual subsampling techniques.

1 Introduction

The genome is a great source for insight into previously unexplained diseases like [Tolstoi and Smith, 1999], [Ballard *et al.*, 2010], and [Peter and Seddon, 2010]. Using the genotypes and a phenotypic outcome for a sample of patients, researchers are able to identify relevant sections of the genome that inform diagnosis. Recent studies have shown that the rate of collection of genomic data is still increasing [Muir *et al.*, 2016]. Not only are sample sizes increasing, but dimensionality as well.

It is of great interest to perform whole genome analyses (WGA) in order to determine the causal single nucleotide polymorphisms (SNPs) for a particular outcome without manually subsampling the large genome. However, in many experiments, researchers focus only on a small portion of the genome [He and Lin, 2011], [Wu *et al.*, 2010], and [Fan and Lv, 2008]. This may be done to mitigate the effect of spurious correlations between certain SNPs and the outcome.

We hypothesize that by considering block structures in the Linkage Disequilibrium (LD) [Slatkin, 2008] matrix of the data, we can return a significantly smaller set of significant SNPs without a large drop in prediction accuracy. LD is a property observed in genomics where adjacent SNPs take on highly correlated values.

By penalizing *blocks* of SNPs rather than penalizing the SNPs on a genome-wide scale, we can select blocks that are predictive of the outcome, rather than individual SNPs. Nonetheless, there is value in locating individual SNPs within these blocks. We therefore need a model that can return SNPs that are potentially causal to a phenotype but are not alone in their block with respect to the outcome variable. Essentially, if a block on average is not predictive of the outcome, no SNP within the block should be selected.

Our main contribution is a two-step procedure in which we first determine the blocks in the LD matrix, then regularize an SVM model based on these blocks. Unlike some previous works like [Kim and Xing, 2012] and [Zhang *et al.*, 2012] in this area, we also show that block regularization is useful in the genomic context when combined with a max-margin classifier. We observe that despite a slightly larger set of SNPs being returned in the block-regularized case, the percentage of SNPs that are significant are better than or equal to the L_1-regularized models.

2 Related works

The objective of reducing the false discovery rate among SNPs has long been of interest. Many early methods like [He and Lin, 2011], [Wu *et al.*, 2010] and [Fan and Lv, 2008] use prior knowledge to screen variables before being used in a regression-based experiment. This is often highly subjective, and causes issues in reproducing these results [Abad-Grau *et al.*, 2012].

To return a sparse group of features, recent studies like [Waldmann *et al.*, 2013] and [Papachristou *et al.*, 2016] have employed some variant of LASSO and elastic-net [Zou and Hastie, 2005] regularization. While these are able to significantly reduce the number of features returned, this model formulation still doesn't consider groups of correlated variables.

Finally, to address this issue, studies like [Kim and Xing, 2012] and [Zhang *et al.*, 2012] have considered penalizing groups of variables jointly. These methods employ prior knowledge to determine groups of related SNPs and apply some form of the group-LASSO to a least-squares regression model.

We propose several extensions to this line of thinking. First, we isolate the problems of achieving low classification error, and selecting features that are highly significant, in order to focus on the latter. We explore a variety of penalty terms including a group-LASSO [Jacob *et al.*, 2009]. We also propose an SVM [Cortes and Vapnik, 1995] formulation with each regularization scheme and explain briefly how this can be optimized. Finally, we detail an algorithm to identify groups of correlated features as input to our regularization schemes.

3 Methods

We motivate our procedure with the following goal. While L_1-regularized models provide sparse solutions, they do not adequately deal with block covariate structure in the data. More specifically, we would not only like to identify relevant SNPs, but blocks of relevant SNPs. Rather than pick relevant SNPs across the entire feature space, it might be of interest to first treat each block as a single unit. Once we identify these blocks, we want to then identify those SNPs within the block that correlate most with a given phenotype.

We divide our procedure into two parts: block inference, and block regularization. We first approximate the number, and location of blocks using linkage-disequilibrium (LD). Once we have our blocks of correlated SNPs, we can define an objective that enforces sparsity both on the block level and on the individual SNP level.

Let X be the $n \times d$ data matrix where $X_{i,j} \in \{0, 1, 2\}$. Let Y be an $n \times 1$ vector where $Y_i \in \{0, 1\}$ or $Y_i \in \{-1, 1\}$ for classification tasks. In our experiments, we consider the least-squares regression (1) and the SVM (3) settings. We propose the usage of a max-margin classifier as it has been proven that SVMs converge to a separating hyperplane that is more robust to noise [Xu *et al.*, 2009].

$$\theta = \arg\min_{\theta} \lambda ||\theta||_1 + ||Y - X\theta||_2^2 \quad (1)$$

$$\theta = \arg\min_{\theta} \sum_{i=1}^{n} \left[1 - y_i \left(\sum_{j=1}^{d} X_{i,j}\theta_j \right) \right]_+ \quad (2)$$

$$\text{s.t.} ||\theta||_1 \leq s \quad (3)$$

To determine groups of contiguous SNPs, we first assume that the $d \times d$ linkage disequilibrium matrix Σ exhibits block structure. That is:

$$\Sigma_{i,j} = \begin{cases} \rho_k & \text{if } i, j \in G_k \\ \epsilon & \text{otherwise} \end{cases} \quad (4)$$

Where ρ_k is the linkage disequilibrium between every pair of SNPs $i, j : i \neq j$ which are in the same block G_k. If $X_i \in G_k, X_j \in G_{k'} : k \neq k'$, then we assume the linkage disequilibrium is some small quantity ϵ.

3.1 Block inference

We employ a constrained version of the reverse-delete minimum spanning tree algorithm (Algorithm 1) to identify l groups in the data. We simply compute pairwise linage-disequilibrium for all adjacent SNPs, then determine groups such that the total across-group linkage-disequilibrium is minimized as shown in (5).

$$\min \sum_{i=1}^{d} \sum_{j=1}^{d} \begin{cases} \text{LD}(s_i, s_j) & \text{if } \mathit{Group}(i) \neq \mathit{Group}(j) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We also don't need to hard-code a value for l, the number of groups. We instead maintain the percentage increase in the minimum linkage disequilibrium as we add iteratively add more groups. Once this value drops below a threshold τ, we stop adding groups.

Algorithm 1 Constrained reverse-delete clustering

```
Input: data $X_{n \times d}$, number of groups $l$
Initialize $\forall i \in \{1, 2, \ldots, l\}, c_i = 0$.
Initialize $v = \{0\}^{d-1}$
Initialize $w = \varnothing$
for $i = 1$ to $d - 1$ do
    $v_i \leftarrow \text{LD}(X_{:,i}, X_{:,i+1})$
end for
Sort $v$ in ascending order
Let $\omega$ be the sorted indices of $v$
Let $m$ be the median value in $v$
Let $\gamma$ be $\frac{|v_1 - m|}{m}$
$i \leftarrow 1$
while $\gamma < \tau$ and $i < d$ do
    $w \leftarrow w \cup \{\omega_i\}$
    $i \leftarrow i + 1$
    $\gamma \leftarrow \frac{|v_i - m|}{m}$
end while
$l \leftarrow i$
$w \leftarrow w \cup \{0, d\}$
Sort $w$ in ascending order
Let $\alpha = 1$
for $i = 1$ to $l$ do
    for $j = w_i$ to $w_{i+1}$ do
        $c_j \leftarrow \alpha$
    end for
    $\alpha \leftarrow \alpha + 1$
end for
```

3.2 Block regularization

Now that we have grouped SNPs into blocks of high linkage-disequilibirum, we can enforce our objectives – block-level and SNP-level sparsity – by penalizing each block independently, then adding an overall sparsity constraint. We begin by adding a separate regularization term for each block using block regularization [Jacob *et al.*, 2009]:

$$\theta = \arg \min_{\theta = \{\theta^{(1)}, \ldots, \theta^{(l)}\}} R(\theta) + \mathcal{L}(X, \theta) \quad (6)$$

$$R(\theta) = \lambda \sum_{k=1}^{l} \sqrt{n_k} ||\theta^{(k)}||_2^2 \quad (7)$$

$$n_k = |G_k| \quad (8)$$

Where $\mathcal{L}(X, \theta)$ is either a least squares or SVM objective. We notice that while sparsity within each group is enforced, overall sparsity is not. If there is a case where the number of groups is very high, (6) might devolve into a simple L_2-regularization. We therefore propose an additional constraint similar to [Vincent and Hansen, 2014] on the overall θ (9).

$$
\begin{aligned}
R(\theta) &= (1-\alpha)\lambda \sum_{k=1}^{l} \sqrt{n_k} ||\theta^{(k)}||_2^2 + \alpha\lambda ||\theta||_1 && (9)\\
&= (1-\alpha)\lambda \sum_{k=1}^{l} \sqrt{n_k} ||\theta^{(k)}||_2^2 + \alpha\lambda \sum_{k=1}^{l} ||\theta^{(k)}||_1 && (10)\\
&= \lambda \sum_{k=1}^{l} \Big((1-\alpha)\sqrt{n_k} ||\theta^{(k)}||_2^2 + \alpha ||\theta^{(k)}||_1 \Big) && (11)
\end{aligned}
$$

We use α to ensure that the combination of the block-level and the SNP-level regularization terms is convex. We can then rewrite $R(\theta)$ as (11), allowing us to use the elastic net argument [Zou and Hastie, 2005] to state that each term in the summation is convex.

Block-wise objective

Using our values for $R(\theta)$ and $\mathcal{L}(X, \theta)$, we notice that since $R(\theta)$ is convex, we can use any $\mathcal{L}(X, \theta)$ that is convex. We can perform a block-wise gradient descent by writing our objective function as the sum of the losses within each block:

$$
\begin{aligned}
\theta^{(j)} &= \arg\min_{\theta^{(j)}} \frac{1}{2n} \mathcal{L}^{(j)}(X^{(j)}, \theta^{(j)}) + R(\theta^{(j)}) && (12)\\
R(\theta^{(j)}) &= \lambda \Big((1-\alpha)\sqrt{n_j} ||\theta^{(j)}||_2^2 + \alpha ||\theta^{(j)}||_1 \Big) && (13)
\end{aligned}
$$

We use the same methodology as [Vincent and Hansen, 2014] to derive the block-specific loss functions:

$$
\begin{aligned}
&\mathcal{L}^{(j)}(X^{(j)}, \theta^{(j)}) && (14)\\
&= \begin{cases} ||r_{-j} - X^{(j)}\theta^{(j)}||_2^2 & \text{Sq Loss} \\ \sum_{i=1}^{n} \Big[1 - y_i r_{-j} - y_i X_i^{(j)} \theta^{(j)} \Big]_+ & \text{SVM} \end{cases} && (15)\\
r_{-j} &= \begin{cases} Y - \sum_{k \neq j} X^{(k)} \theta^{(k)} & \text{Sq Loss} \\ \sum_{k \neq j} X_i^{(k)} \theta^{(k)} & \text{SVM} \end{cases} && (16)
\end{aligned}
$$

For sake of brevity, we refer the reader to [Vincent and Hansen, 2014] for proof of convergence. Although they only detail the least squares case, the SVM objective is also convex [Cortes and Vapnik, 1995] so the same convergence properties hold.

4 Experiments

As our goal is to return a small set of causal SNPs, we present comparisons with only L_1-regularized models since L_2-regularized models are known to return less sparse sets of features. The three regularization schemes we consider are L_1, unconstrained block regularization (UBR) (7), and block regularization (BR) (9). Table 1 is a complete list of models we consider. Since we are focused more on feature selection than validation accuracy, we tune each model until maximum testing accuracy before reporting our summary statistics. We compare the proportion of all significant SNPs returned by each algorithm and the number of non-zero weight values from the runs with highest accuracy. We apply Bonferroni correction [Dunn, 1958] to our t-tests to account for multiple comparisons.

Table 1: Legend name to model lookup

Legend	Model
BRLog	Logistic regression with BR
UBRLog	Logistic regression with UBR
L1Log	Logistic regression with L_1
BRSVM	SVM with BR
UBRSVM	SVM with UBR
L1SVM	SVM with L_1

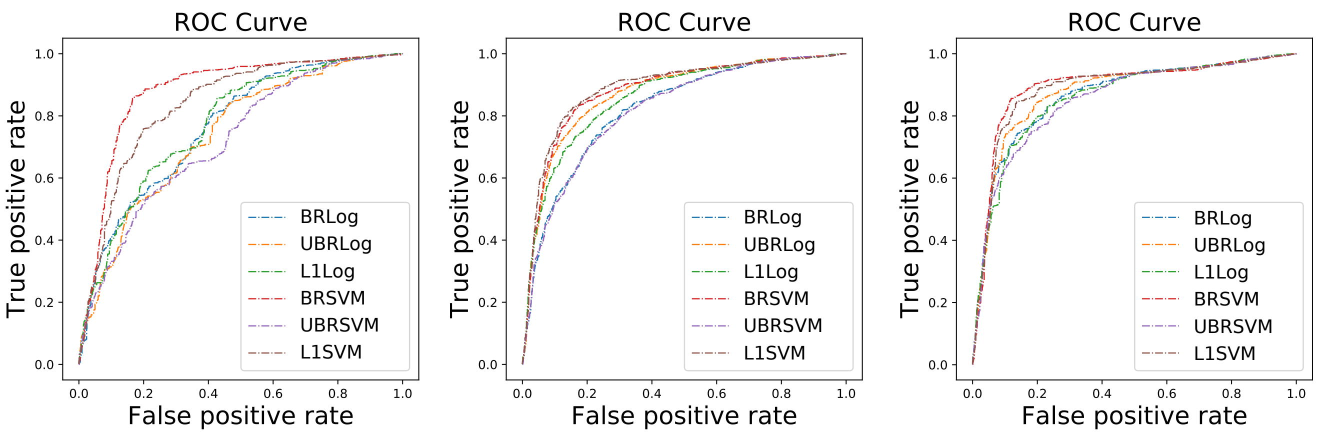
We use the program `ms` [Hudson, 2002] to generate genotypes that closely resemble human biology. We use the same procedure as [Kim and Xing, 2012] but with $\sim$ 20000 haplotypes or $\sim$ 10000 individuals, and $\sim$ 8000 SNPs. We choose $m : 1 \leq m \leq |\hat{G}_k|$ SNPs from each $\hat{G}_k$ where $\{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{k'}\} \subset \{G_1, G_2, \ldots, G_k\}$. These SNPs are combined linearly, then thresholded to generate a simulated phenotype. We also vary the recombination rate ρ between every pair of SNPs such that $\rho \in \{2 \times 10^{-i} : 6 \leq i \leq 8\}$. The higher the recombination rate, the weaker the block structure. As an additional source of noise, we randomly flip the labels of 10% of our data, bounding the maximum test accuracy achievable.

5 Results

In Figure 1 we see the ROC curves for our three settings. The data in Figure 1a has the highest level of linkage-disequilibrium and we see that the block-regularized SVM (BRSVM) performs best by a significant margin, followed by the L_1 regularized SVM (L1SVM). The other models perform similarly. When the linkage-disequilibrium decreases (Figure 1b), we notice that the BRSVM and L1SVM still perform well, but the other models are catching up. Finally, in Figure 1c where the linkage-disequilibrium is the lowest, thereby removing most of the block covariate structure in the data, the models mostly perform the same, but it is important to note that the BRSVM still performs on par with the rest.

We explore these results further in Table 2. We look specifically at the medium recombination rate setting in particular because the blocks are not as distinct as the low recombination rate setting, but are still present. The test accuracy for the BRSVM model is the highest by a noticeable margin. As we see in the ROC curves, the L1SVM is close behind with a test accuracy of 0.86. The unconstrained block-regularized SVM model is not too far behind with an accuracy of 0.84. We also notice that the BRLog model performs the worst in terms of test error, despite its loss converging to a stable minimum, while the UBRLog model performs better despite the larger number of features returned. We observe that the SVM models perform better than their counterparts in the cases of L_1 and BR.

In terms of non-zero features returned, the L_1 models return the fewest, followed by the BR, then the UBR model. This is due to a large number of groups being detected by Algorithm 1. This has no effect on the L_1 model, but greatly weakens the regularization effect on the UBR models.



(a) Low recombination rate

(b) Medium recombination rate

(c) High recombination rate

Figure 1: ROC curves

It is important to note that while the number of non-zero features returned by BRSVM is higher than the L1SVM, the percentage of significant SNPs is higher. Additionally, the BRSVM returns a smaller set of relevant features than the BRLog model, but has much higher test accuracy.

Table 2: Medium recombination rate

Model	Test accuracy	Sig. SNPs	Non-zero θs
BRLog	0.8110	0.714	42
UBRLog	0.8400	0.970	67
L1Log	0.8225	**1.000**	**24**
BRSVM	**0.8710**	**1.000**	40
UBRSVM	0.8155	0.800	85
L1SVM	0.8575	0.892	28

We continue the comparison between L1SVM and BRSVM in Figures 2 & 3. Each location along the x-axis represents each SNP we consider. The height of the point at each point represents the negative logarithm of the p-value. Blue circles represent returned SNPs that were significant, red diamonds represent returned SNPs that were insignificant. We see that the svmL1, which is the closest in terms of test accuracy to the svmBlock, has returned many more insignificant SNPs. This is important to note because despite returning more SNPs overall, the block-regularized model returns a lower number of insignificant SNPs than the L_1-regularized model.

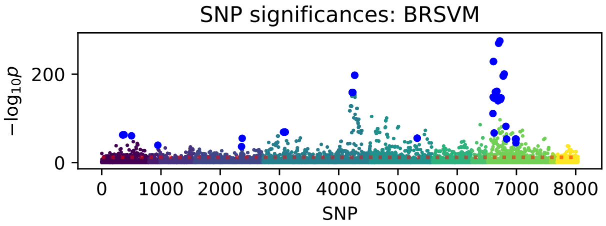


Figure 2: Manhattan plot for block-regularized SVM model

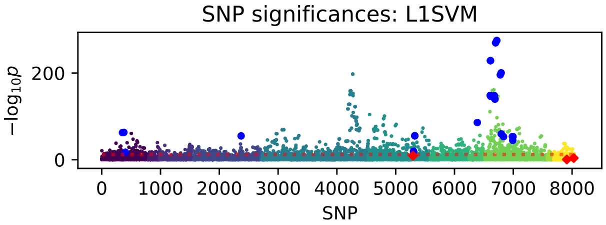


Figure 3: Manhattan plot for L_1-regularized SVM model

6 Concluding remarks

We have shown empirically that BRSVM exhibits superior performance in cases of block-covariate structure in the data. While the number of returned SNPs by the block-regularized models is higher, the percentage of significant SNPs is very high in the case of the SVM model. This is important because we are able to identify a larger number of causal SNPs that we might have missed out using a regularization scheme that doesn't consider linkage-disequilibirum. Even though each of the regularization schemes yield similar results in the case of very weak block-covariate structure, our proposed BRSVM model still performs well, implying that it can be used in settings of both low and high recombination rates.

References

[Abad-Grau *et al.*, 2012] Mara M Abad-Grau, Nuria Medina-Medina, Rosana Montes-Soldado, Fuencisla Matesanz, and Vineet Bafna. Sample reproducibility of genetic association using different multimarker tdts in genome-wide association studies: characterization and a new approach. *PloS one*, 7(2):e29613, 2012.

[Ballard *et al.*, 2010] David Ballard, Clara Abraham, Judy Cho, and Hongyu Zhao. Pathway analysis comparison using crohn's disease genome wide association studies. *BMC medical genomics*, 3:25, 2010.

[Cortes and Vapnik, 1995] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[Dunn, 1958] Olive Jean Dunn. Estimation of the means of dependent variables. *The Annals of Mathematical Statistics*, 29(4):1095–1111, 1958.

[Fan and Lv, 2008] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[He and Lin, 2011] Qianchuan He and Dan-Yu Lin. A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8, 2011.

[Hudson, 2002] Richard R Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–8, Feb 2002.

[Jacob *et al.*, 2009] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

[Kim and Xing, 2012] S. Kim and E. P. Xing. Feature Selection via Block-Regularized Regression. *ArXiv e-prints*, June 2012.

[Muir *et al.*, 2016] Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J. Spakowicz, Leonidas Salichos, Jing Zhang, George M. Weinstock, Farren Isaacs, Joel Rozowsky, and Mark Gerstein. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17(1):53, Mar 2016.

[Papachristou *et al.*, 2016] Charalampos Papachristou, Carole Ober, and Mark Abney. A lasso penalized regression approach for genome-wide association analyses using related individuals: application to the genetic analysis workshop 19 simulated data. *BMC Proceedings*, 10(7):53, Oct 2016.

[Peter and Seddon, 2010] Inga Peter and Johanna M Seddon. Genetic epidemiology: successes and challenges of genome-wide association studies using the example of age-related macular degeneration. *American journal of ophthalmology*, 150(4):450–452.e2, Oct 2010.

[Slatkin, 2008] M. Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–486, 2008.

[Tolstoi and Smith, 1999] L G Tolstoi and C L Smith. Human genome project and cystic fibrosis–a symbiotic relationship. *Journal of the American Dietetic Association*, 99(11):1421–7, Nov 1999.

[Vincent and Hansen, 2014] Martin Vincent and Niels Richard Hansen. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.

[Waldmann *et al.*, 2013] Patrik Waldmann, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4:270, 2013.

[Wu *et al.*, 2010] Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–85, Apr 2010.

[Xu *et al.*, 2009] Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, December 2009.

[Zhang *et al.*, 2012] Zhuo Zhang, Yanwu Xu, Jiang Liu, and Chee Keong Kwoh. Identify predictive snp groups in genome wide association study: A sparse learning approach. *Procedia Computer Science*, 11:107 – 114, 2012. Proceedings of the 3rd International Conference on Computational Systems-Biology and Bioinformatics.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.