Published in final edited form as:
Stat Anal Data Min. 2013 August 1; 6(4): . doi:10.1002/sam.11183.

Penalized Regression and Risk Prediction in Genome-Wide Association Studies

Erin Austin[1], **Wei Pan**[1], and **Xiaotong Shen**[2]

[1]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455

[2]School of Statistics, University of Minnesota, Minneapolis, MN 55455

Abstract

An important task in personalized medicine is to predict disease risk based on a person's genome, e.g. on a large number of single-nucleotide polymorphisms (SNPs). Genome-wide association studies (GWAS) make SNP and phenotype data available to researchers. A critical question for researchers is how to best predict disease risk. Penalized regression equipped with variable selection, such as LASSO and SCAD, is deemed to be promising in this setting. However, the sparsity assumption taken by the LASSO, SCAD and many other penalized regression techniques may not be applicable here: it is now hypothesized that many common diseases are associated with many SNPs with small to moderate effects. In this article, we use the GWAS data from the Wellcome Trust Case Control Consortium (WTCCC) to investigate the performance of various unpenalized and penalized regression approaches under true sparse or non-sparse models. We find that in general penalized regression outperformed unpenalized regression; SCAD, TLP and LASSO performed best for sparse models, while elastic net regression was the winner, followed by ridge, TLP and LASSO, for non-sparse models.

Keywords

AUC; GWAS; LASSO; Logistic regression; MLE; Ridge; SCAD; TLP; Elastic Net; SNP

1 INTRODUCTION

Genetic information has the potential to improve health outcomes by allowing an individual to tailor preventive care and treatment plans to his or her personalized medical needs. An important task in personalized medicine is using a person's genome to predict disease risk (and treatment response). A necessity for making accurate risk predictions based on individuals' genomes is obtaining data on their genetic variants and phenotypes. Genome-wide association studies (GWAS) provide such data to researchers. Now one critical question is how to best predict disease risk from a large number of genetic variants, such as single-nucleotide polymorphisms (SNPs). Penalized regression equipped with variable selection, such as LASSO (Tibshirani, 1996), is deemed to be promising in this setting. However, for some diseases the sparsity assumption used by penalized regression to facilitate variable selection may not hold, in which case it is not completely clear how to proceed: should we apply a penalized or unpenalized approach? how about other penalized methods that do not conduct variable selection, such as ridge regression (Hoerl and Kennard, 1970)? To answer these questions, our current research investigated the performance of an

Correspondence author: Wei Pan, Telephone: (612) 626-2705, Fax: (612) 626-0660, weip@biostat.umn.edu, Address: Division of Biostatistics, MMC 303, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455–0392, U.S.A.

unpenalized approach and several representative penalized regression approaches under various scenarios with sparse or non-sparse models.

GWAS identify risk SNPs by individually testing each SNP with a stringent significance level adjusting for multiple testing. Many SNPs discovered to be associated with disease have been validated (McCarthy et al., 2008). However, for many strongly heritable diseases, their risk cannot be adequately explained by only a small number of identified SNPs. For example, adding seven SNPs known to be associated with breast cancer to the National Cancer Institute's Breast Cancer Risk Assessment Tool increased the discriminatory accuracy of the tool by only a small amount as measured by the area under the receiver operating characteristic curve (AUC) (Gail, 2009). In related work Gail (2008) demonstrated that very large relative risks are needed for a single factor to meaningfully improve disease classification; therefore, estimation of the effect of many disease associated SNPs with small effects will require researchers to address the issue of candidate SNPs vastly outnumbering available case samples. Penalized regression with variable selection can address this issue. In another study the percent of phenotypic variance in the highly heritable trait height explained by SNPs increased from 5% to 45% when both genome-wide significant SNPs and many non-significant SNPs were considered simultaneously (Yang et al., 2010). Increasing the number of SNPs used may also impact risk prediction: the inclusion of many non-significant SNPs discriminated bipolar disorder, coronary heart disease, hypertension, and Crohn's disease to some degree better than when only fewer and more significant SNPs were included (Evans et al., 2009). Furthermore, there was evidence to support polygenic effects for many common diseases (Park et al., 2010). For example, the risk of schizophrenia seemed to be associated with hundreds to thousands of SNPs (The International Schizophrenia Consortium, 2009). It is now hypothesized that many common diseases are associated with many SNPs with small to moderate effects.

Two studies have confirmed the value in including up to thousands of SNPs when assessing disease risk (Kang et al., 2011; Wei et al., 2009). Importantly, both studies revealed that, while still noticeably better than random, logistic regression with maximum-likelihood estimation was suboptimal in utilizing large numbers of SNPs to classify disease status. A recent study concluded that utilizing penalized regression with variable selection, specifically LASSO, on a large number of SNPs in addition to those reaching the genome-wide significance level could improve prediction of Crohn's disease (Kooperberg et al., 2010). This disease is a form of inflammatory bowel disease affecting as many as 1.4 million Americans (About Crohn's Disease, 2009; Crohn's Disease, 2010). Patients with Crohn's disease have a chronic inflammation of the gastrointestinal tract that causes mild to severe symptoms such as abdominal pain, fever, and fatigue (Crohn's Disease, 2010). Gaya et al. (2006) presented evidence of the heritability of Crohn's disease. Two subsequent studies (WTCCC, 2007; Franke et al., 2010) identified six regions of chromosome 10 associated with Crohn's disease. To mimic real situations we use the real SNP data from chromosome 10 to generate simulated disease risks and disease phenotypes in order to assess the performance of various regression methods with respect to risk estimation and disease classification. Specifically, we consider four types of true models: (1) a sparse model with risk being determined by a small number of SNPs with large effect sizes, (2) a sparse model with a small number of SNPs with moderate effect sizes, (3) a non-sparse model with risk being determined by a large number of SNPs with moderate effects, and (4) a non-sparse model with an even larger number (> 1/3 of the sample size) of SNPs with small effect sizes. We consider both unpenalized and penalized regressions, the former based on maximum likelihood estimator (MLE) while the latter on (1) least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), (2) smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), (3) truncated L_1–penalty (TLP) (Shen et al., 2012), (4) ridge regression (Hoerl and Kennard, 1970) and (5) elastic net (Zou and Hastie, 2005). This study

is a follow-up on Kooperberg et al. (2010), in that we consider several new penalized regression methods and contrast the performance of the methods between sparse and non-sparse true models.

We also study the discrimination capabilities of the regression methods on two real data sets, Crohn's disease and bipolar disorder provided by the Wellcome Trust Case Control Consortium (WTCCC) (2007). It was confirmed that the best performer was dependent on the number and effect sizes of causal SNPs in the true model, and the inclusion of SNPs failing to meet the genome-wide significance level impacted the prediction accuracy.

2 METHODS

2.1 Data

We use the Crohn's disease and bipolar disorder case and control data provided by the WTCCC. The WTCCC has collected genotype data of about 500,000 SNPs for approximately 2,000 samples for each of seven diseases, such as type 1 diabetes, hypertension, bipolar disorder and Crohn's disease, and 3,000 controls (WTCCC, 2007). For simulations, we use the genotype data of 28501 SNPs on chromosome 10 for Crohn's disease cases and controls. For quality control purposes, per WTCCC recommendations, we remove some samples and retain 1748 Crohn's disease samples and 2938 control samples; we also exclude some SNPs as recommended. Next, we eliminate the SNPs with a minor allele frequency (MAF) less than 5%. Furthermore, to mimic practical situations while maintaining a reasonable size for repeated simulations, we test each SNP separately by a chi-squared test for its association with Crohn's disease, and remove those with p-values larger than 0.1. At the end, we have about 2300 SNPs left and use them throughout our simulations.

2.2 Model

Let $Y_i = 0$ or 1 be a binary disease indicator for subject $i = 1, \ldots, n$, and X_{ij} subject i's minor allele number (0,1, or 2) for SNP $j = 1, \ldots, m$. Our aim is to build a model to successfully estimate subject i's risk of disease, $P(Y_i = 1 \| x_i)$, based on his or her SNP data $x_i = (X_{i1}, \ldots, X_{im})^T$. As in standard practice for binary outcomes, we use a logistic regression model:

$$\text{logit}(P(Y_i=1|x_i))=\log(\frac{P(Y_i=1|x_i)}{1-P(Y_i=1|x_i)})=\beta_0+\sum_{k=1}^{p}X_{ik}\,\beta_k, \quad (1)$$

where β_0 and β_k are unknown regression coefficients to be estimated; $p \leq m$ indicates any user specified subset of the SNPs.

In *unpenalized* logistic regression with maximum-likelihood estimator (MLE), β_0 and $\beta = (\beta_1, \ldots, \beta_p)^T$ are estimated by maximizing the log-likelihood:

$$l(\beta_0,\beta)=\sum_{i=1}^{n}Y_j\,(\beta_0+x_i^T\beta)-\log[1+\exp(\beta_0+x_i^T\beta)]. \quad (2)$$

The MLE is asymptotically unbiased with fixed p as $n \to \infty$, but it may not be for a large p. One possible remedy is to introduce regularization or penalization on regression coefficients. The use of certain penalties, such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), shrinks many regression coefficient estimates to be 0, effectively selecting a subset of SNPs to be used for prediction. *Penalized* logistic regression provides

coefficient estimates for β_0 and β by maximizing a penalized log-likelihood (Friedman et al., 2008):

$$l(\beta_0, \beta) - \lambda P(\beta), \quad (3)$$

where $\lambda \geq 0$ is a tuning parameter controlling the extent of penalization imposed by penalty $P(\beta)$. LASSO regression uses

$$P(\beta) = \sum_{k=1}^{p} |\beta_k|, \quad (4)$$

which is convex and computationally convenient. However, LASSO estimates are biased and may not be consistent. To avoid these issues, Fan and Li (2001) proposed using the SCAD penalty $P(\beta, \lambda)$ replacing $\lambda P(\beta)$:

$$\frac{dP(\beta, \lambda)}{d\beta} = \sum_{k=1}^{p} \lambda \text{sign}(\beta_k) [I(|\beta_k| \leq \lambda) + \frac{(a\lambda - |\beta_k|)_+}{(a-1)\lambda} \cdot I(|\beta_k| > \lambda)] \quad (5)$$

for $a = 3.7$. While maintaining the capability of variable selection, the SCAD penalty does not introduce biased estimates for some larger coefficients. The truncated L_1-penalty (TLP) adaptively determines which larger coefficients will not be penalized by introducing a separate thresholding parameter $\tau > 0$ (Shen et al., 2012):

$$P(\beta) = \sum_{k=1}^{p} \min(|\beta_k|/\tau, 1). \quad (6)$$

If a coefficient $\beta_k > \tau$, it will not be further penalized. The TLP approaches the L_0-loss as $\tau \to 0^+$. A penalized method without the capability of variable selection is ridge regression (Hoerl et al., 1970) with penalty

$$P(\beta) = \sum_{k=1}^{p} \beta_k^2. \quad (7)$$

Certain true models might be best estimated using a hybrid penalty that simultaneously performs variable selection and continuous shrinkage (Zou and Hastie, 2005). In these settings elastic net penalized regression may be more suitable. Elastic net penalized regression has been shown to produce a sparse model with good prediction accuracy, possibly superior to LASSO, while simultaneously promoting the grouping of strongly correlated predictors (Zou and Hastie, 2005). Its penalty structure is a weighted combination of the LASSO and ridge penalties controlled by a user specified mixing parameter α, which is restricted to [0, 1]. The *naive* elastic net penalty (Zou and Hastie, 2005) is

$$P(\beta) = (1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1, \quad (8)$$

where α is selected to match the desired balance of variable selection and coefficient shrinkage. Zou and Hastie (2005) suggested that further gains may be possible from using a rescaled version of the elastic net penalty. However, Friedman et al. (2008) used the *naive* version of the penalty in R package glmnet they developed to perform elastic net penalized regression. Results presented here follow this convention and are not rescaled. For sparse true models (i.e. with few non-zero β_k's) with a large number of candidate predictors (i.e. a

large p), variable selection is often beneficial. However, for non-sparse models with many small non-zero $|\beta_k|$'s, variable selection will be difficult and may not result in good performance. On the other hand, since the ridge penalty has the grouping function (Zou and Hastie, 2005), ridge regression performs like model averaging. It is known that neither model selection nor model averaging can dominate the other, and each performs better under different situations (Yuan and Yang, 2005; Shen and Huang, 2006). In the current context, especially with non-sparse true models, it is not clear how LASSO, SCAD, and TLP compare to the ridge penalty for risk prediction, or if the elastic net penalty is superior, which is one of our aims.

3 SIMULATIONS

3.1 Simulation Set-ups

We use the real SNP data of the WTCCC control cohort to generate disease probabilities, $\pi_i = P(Y_i = 1)$. First, we randomly select p_1 causal SNPs (i.e. with corresponding $\beta_k \neq 0$). The true correlations for any two SNPs range from -0.8371 to 1 and approximately fit a symmetric unimodal distribution centered at 0. Table 1 provides summary statistics for all pairwise correlations for example sets of size p = 5, 10, 50, 100, 500, 1000 randomly selected SNPs. Table 1 demonstrates how the true models with various numbers of p SNPs contain a diverse range of minor, moderate or strong correlations among the SNPs.

We use p_1 = 10 for two sparse models, one with strong effects (i.e. large $|\beta_k|$'s) and the other with only moderate effects (i.e. smaller $|\beta_k|$'s); we also use p_1 = 300 and p_1 = 900 for two non-sparse models. Second, we set $\beta_0 = \log(0.05/0.95)$ to emulate diseases with low prevalence, and follow Wray et al. (2007) to create odds ratios (ORs, $\mathrm{OR}_k = \exp(\beta_k)$) of having disease for the p_1 causal SNPs. Specifically, we set $\mathrm{OR}_k = 1 + \varepsilon(\mathrm{OR}_0 - 1)$ with ε randomly generated from a standard exponential distribution $Exp(1)$ and OR_0 being the mean OR, which is 2.75 and 1.415 for the two sparse models and 1.17 and 1.125 for the two non-sparse models respectively. We also randomly choose the sign of each β_k to be positive or negative to reflect both risk and protective causal SNPs. Third, the disease probability π_i for each subject i = 1, …, 2938 in the WTCCC control cohort, is generated according to logistic regression model (1) with only chosen causal SNPs.

Finally, we use each π_i sequentially to generate disease status $Y_i \sim Bin(\pi_i)$; this step is repeated until we have n = 2000 cases and n = 2000 controls (while the other cases or controls are ignored) for each simulated dataset. One hundred datasets were generated under each of the four true models.

For each simulated dataset a randomly selected half of both the cases and controls is used as training set for building regression models, while the remaining half is the test set used for unbiased assessment of performance. The performance of each method is evaluated in two distinct settings. In the first setting we rank all SNPs by the p-values of their univariate association with disease. Starting with a few of the most significant SNPs, we fit and refit the logistic model for each method, sequentially adding more and more top ranked SNPs into the model (1) to be fit. The structure of this scenario informs when the inclusion of increasingly less significant SNPs improves or deteriorates the performance. Gail (2009) measured the impact of only seven SNPs on classification of one disease, breast cancer, finding a very minor effect. Although they were not directly studying prediction, Yang et al. (2010) identified one trait, height, whose heritability could be explained better with models that considered many non-significant SNPs. Our first modeling scenario generalizes this previous work to measure the impact of including more and more SNPs (by design including less significant SNPs) on a spectrum of models with less and less true sparsity. Thus, the results can inform about underlying genetic architectures for which penalized regression can

use additional SNPs to improve risk classification. The results presented in the following section for the unpenalized regression are from the usual MLE, while those for LASSO, SCAD and ridge use the tuning parameter λ selected via 10-fold cross-validation to have the smallest prediction error for any given number of candidate SNPs.

As exhibited in equations (6) and (8), the elastic net penalty depends on an additional parameter, α, and the TLP penalty requires specification of τ. Elastic net estimates are generated for each of a sequence of penalties defined by a uniformly spaced sequence of values for the mixing parameter, α. The elastic net regression models are fit starting with $\alpha = 0$, corresponding to ridge regression, and then with α increased by units of 0.10 until $\alpha = 1$, corresponding to LASSO regression. For the TLP we apply a range of τ values chosen to yield a series of models with minor to major coefficient shrinkage. To save computing time for tuning parameter selection for the simulated datasets, we use an independent tuning dataset of an equal size generated exactly like the training and test data set. The idea is similar to the CV except that we only need to fit a model once with the training data, then use the tuning data to calculate the prediction error and thus select λ and τ.

The second setting is designed to compare the performance of the methods with a large number of the candidate SNPs. In penalized regression the regularization parameter λ is systematically varied to generate a solution path of the regression coefficients, from which we identify a global maximum of some performance measurement to represent the best ever performance of the corresponding method.

For each method, the estimated β_0 and β_k from a training set are applied to the corresponding test set to obtain risk estimates, $\hat{\pi}_i$. The correlation of the $\hat{\pi}_i$ and the true π_i for the test samples is computed and used to compare the predictive performance of the methods.

This metric has been used in risk and outcome prediction for GWAS data (Wray et al., 2007; Lee et al., 2008). In addition, we also utilize the area under a receiver operating curve (AUC) for test samples to assess the discriminatory capabilities of the regression methods. The AUC is the gold standard metric that has been most consistently used in the GWAS literature. The use of AUC also permits direct comparison to previous related work. R package glmnet was used to fit the LASSO, ridge and elastic net penalized regression models. SCAD models were fit using the R package ncvreg. TLP models for the simulated data sets were fit using Feature Grouping and Selection Over and Undirected Graph (FGSG) software implemented in Matlab (Yang et al., 2012) while those for the real data were fit using our own implemented R function. Computational time necessitated using the FGSG software, which was much faster in fitting penalized linear regression models. It is known that linear regression models perform well for binary traits with GWAS data (Wu et al., 2010). We also compared the results from penalized logistic regression models fitted by the R function with those from linear models by the FGSG software for the first ten simulated datasets; their differences were within 0.031 in the correlation metric and within 0.01 in the AUC metric.

3.2 Main Results

We first investigate the effect of using an increasing number of top SNPs for risk prediction. Figure 1(a) presents the correlation between true risk and predicted risk, $Corr(\pi_i, \hat{\pi}_i)$. For each of the four true models, $Corr(\pi_i, \hat{\pi}_i)$ for each method is plotted as a gray curve against the number of the top SNPs used in the candidate model (before penalization) for each of the 100 simulated datasets, and the mean correlation curve over all 100 simulations is plotted as a dark red curve. The elastic net and TLP results are for the data-tuned values of α and τ respectively. In addition, vertical lines mark the number of the SNPs that would meet a Bonferroni adjusted genome-wide significance level at 0.05 when evaluated individually

using a chi-squared test. Examination of the curves beyond the vertical lines reveals situations in which better estimates of the disease risk can be obtained by considering more SNPs, including those failing to meet the genome-wide significance level. The horizontal lines mark the correlations obtained from the MLE of the true model (with exactly all the causal SNPs).

In the sparse model with strong effect sizes, all penalized methods predict risk nearly as well or better than the unpenalized method, as shown in Figure 1(b) where only the maximum correlation across various numbers of top SNPs from each simulated dataset is plotted. For the sparse model with weaker effects and both non-sparse models, all penalized methods by far surpass the MLEs, even ones based on the true models. Among the penalized methods, LASSO, SCAD, TLP and elastic net outperform ridge regression for sparse models, but the trend reverses for non-sparse models for all but the elastic net. As the number of causal SNPs increases or strength of effect decreases, the relative performance of the elastic net and TLP penalties improves. In fact, the elastic net outperforms all methods for the $p_1 = 300$ case, which is to be expected as it is a model balanced between extreme sparse and non-sparse models. The best performing elastic net models are at least as good in the non-sparse $p_1 = 900$ case as those of ridge regression, the best overall performer of the non-mixture penalties. Table 2 provides the mean values of the maximum performance metrics of each regression method for the datasets. The table allows quick comparisons of the various methods in all modeling scenarios. These results reinforce the importance of using a suitable penalty for a given problem, depending on whether the model sparsity assumption holds.

To quantify the impact of including more SNPs, we first examine the performance for the sparse models. The LASSO and SCAD, methods with a variable selection feature, are able to maintain near optimal performance even when the number of candidate SNPs far exceeds that of the true model. Further, the elastic net appears to improve on the LASSO. In contrast, both unpenalized and ridge regressions have their prediction accuracy worsened markedly with the inclusion of more SNPs. For non-sparse models containing many SNPs failing to meet the genome-wide significance level, LASSO, SCAD, and TLP are again able to deal with a large number of SNPs for better risk estimation than the MLE. TLP uses the additional SNPs noticeably better than LASSO and SCAD when the true number of causal SNPs grows. Ridge regression is able to surpass these three penalization methods. In all four models the elastic net performs comparably to the best of the other regressions. This is likely due to its being a hybrid of the sparse and non-sparse regression methods, and our method examined a range of α's corresponding to a range of models from those strongly favoring LASSO to those strongly favoring ridge regression. However, it is noteworthy that the elastic net was not bounded by the performances of LASSO and ridge regressions.

Next, the discriminatory abilities of the methods are assessed because correct classification of disease status is key to personalized medicine. The literature for the clinical application of disease assessments universally reported AUCs as the standard for comparing disease classification methods. Therefore, the current study will assess classification using this metric to enable comparisons to previous work. Figure 2 demonstrates the classification performance of the methods in terms of their AUCs. The main conclusions remain the same: SCAD, closely followed by LASSO, elastic net, and TLP, is the winner for the two sparse models, while elastic net and ridge regression beat other methods for the non-sparse model with $p_1 = 900$. However, for the non-sparse model with $p_1 = 300$, ridge regression performs worse than all other penalization methods. Elastic net performs best, followed by LASSO, TLP, and then SCAD. Overall, elastic net is either the top performer or close to the top for all true models, and every type of penalized regression always beats MLE.

The results presented in Figure 2 demonstrate the value of penalized regression in disease risk estimation and classification, especially in utilizing the information in less significant SNPs that may often go unused. A natural question is whether we can eliminate the need to rank SNPs marginally and examine all SNPs simultaneously. The below simulation results address this question. All the penalized methods start with a full model containing all available SNPs; by varying the tuning parameter λ monotonically, various models are fitted and their performance is assessed. Figure 3(a) provides curves for the correlations between true and predicted risk at any given value of λ for four of the penalized methods: LASSO, SCAD, ridge, and elastic net. Elastic net results for only models with $\alpha = 0.5$ are shown. Since one value of τ that provides a single intuitive interpretation across all four true models does not exist, TLP results as a regularization path in terms of λ would have limited comparability to the results from the other models in this setting and are not presented here. As before, the result for each simulation is represented by a gray curve, and the mean curve across all simulations is plotted as a dark red curve. For comparison, the horizontal lines mark the correlations obtained from maximum likelihood estimation using exactly the true causal SNPs. To facilitate plotting, for each penalized method, the value of λ is scaled by its maximum so that it falls inside the interval [0, 1].

As before, SCAD, LASSO and elastic net with $\alpha = 0.5$ outperform ridge regression for sparse models, while for both non-sparse models ridge regression is the best when judged by their optimal performance shown in Figure 3(b). Interestingly, LASSO outperforms SCAD in all situations, suggesting the robustness of LASSO to a large number of input variables. The performance of the elastic net with $\alpha = 0.5$ is between that of LASSO and ridge in all cases as expected. This elastic net's results are closer to the better of ridge and LASSO in all four models; however, the degree to which the best method outperforms the balanced elastic net ($\alpha = 0.5$) varies by true model. This provides strong evidence that matching the sparsity of the penalty to the model sparsity improves classification. Comparing with earlier results, we can conclude that simultaneous use of too many SNPs will deteriorate the performance of any penalized method, suggesting possible gain in performance by a preliminary screening of a large number of variables. Similar conclusions hold if AUC is used to measure the classification performance of the methods (Figure 4); however, LASSO, followed closely by the elastic net (α=0.5), is the overall winner, in particular it beats ridge regression even for the non-sparse model with $p_1 = 300$, indicating the necessity of variable selection for large p.

3.3 Other Results

Two of the penalties, SCAD and TLP, are non-convex. Thus, there may be multiple local maxima with respect to their corresponding penalized log-likelihood functions, leading to possibly different estimates with different starting values. To examine this issue the authors refit some SCAD and TLP models for the first 20 data sets. The refit models considered the top ranked 1000 SNPs and at a few fixed λ values (and a fixed $\tau = 0.1$ for TLP). Eight different sets of initial regression coefficient values were used as the starting values for SCAD and TLP: the estimated coefficient values with the true model, a vector of all zeros, and the coefficients estimated by LASSO at each of the six λ values: 0.01, 0.1, 1, 2.5, 5, and 10. This was done for the true models with 10 (strong) and 300 causal SNPs to represent one sparse true model and one non-sparse true model. The R package SIS was used to fit the SCAD models as it allowed user specified initial value sets. FGSG software was used to fit TLP models as before.

Figure 5 presents the findings. Each curve represents the average AUC at a given λ over the 20 data sets for each set of the starting values (with the solid one for the first set). The primary finding is that for the λ generating the best AUC given a set of initial coefficients,

all eight sets yield comparable AUCs. Results for many of the SCAD models could not be obtained when λ exceeded 0.1 due to numerical problems in the R package; the partial curves are still provided. Many AUC values were the same or within 0.01 for the SCAD scenarios, thus, given the scale the curves appear to overlap in the plots. Not surprisingly the AUC is impacted by the starting values used to find regression coefficient estimates. Importantly, the impact appears to be small near tuning parameter values yielding the top performing SCAD or TLP models.

Below is a short summary on computing time needed to fit each type of penalized regression models. We calculated the average CPU time for one value or one set of tuning parameters for each penalized regression method with 1000 candidate SNPs for the true models with 10 (strong) and 300 causal SNPs. For the 10 strong SNP scenario, SCAD used approximately 30 seconds to fit a model, TLP used 20 seconds, and the model fitting using glmnet ranged from 1.5 seconds for ridge regression to 7 seconds for LASSO. For the 300 SNP scenario, SCAD used approximately 44 seconds per model-fitting, TLP used 20 seconds, and glmnet ranged from 1.2 seconds for ridge regression to 5 seconds with LASSO.

4 EXAMPLES

The final part of our study examines the classification accuracy of the six regression methods on two WTCCC datasets for Crohn's disease and bipolar disorder. The training and test data are created by randomly dividing the WTCCC disease (case) and WTCCC control samples into two (almost) equally sized sets, one for training and one for test. We consider the 5000 most significant SNPs from all chromosomes as determined by a univariate chi-squared test on each SNP. The whole process, including randomly dividing the true cases and controls into training and test sets, and identifying the 5000 most significant SNPs, is repeated ten times. The results for each of these ten datasets are presented in the following plots. The number of the significant SNPs meeting the significance level of 0.05/373191 are plotted as vertical lines. Horizontal tick marks on the secondary y-axis represent the maximum AUC achieved by MLE with these significant SNPs.

4.1 Crohn's Disease

Current research has identified about 80 SNPs associated with Crohn's disease. Figure 6 shows that approximately only the top 50 SNPs are needed to obtain the best risk prediction for all the methods; however, this includes more than just those SNPs meeting the significance level of 0.05/373191. Interestingly, although TLP was the overall winner and all five penalized methods are better than the unpenalized one, the performance difference among the methods is small.

Figure 7 presents the results of the four penalized methods starting with all 5000 SNPs included in a candidate model. With such a large number of candidate SNPs, while the number of the truly predictive SNPs may be small, the ridge penalty is largely outperformed by LASSO and SCAD that are capable of variable selection. The ridge regression is similarly outperformed by an elastic net penalty that shifts part of the weight from the ridge penalty to the LASSO penalty.

4.2 Bipolar Disorder

Bipolar disorder is a condition in which people go back and forth between mania periods of a very good or irritable mood and depression (Bipolar disorder, 2011). Figure 8 presents the AUC results as the number of candidate SNPs was increased. Unlike Crohn's disease, penalized regression does not always outperform MLE. Elastic net penalized regression and TLP perform best, though again the performance difference among the methods is small. As

shown in 8(b), the inclusion of many SNPs failing to reach the genome-wide significance level does not diminish the discrimination strength of the penalized methods, and in fact ridge and elastic net regression and TLP better use these extra SNPs than both LASSO and SCAD to exceed or nearly exceed its performance achieved with only the few significant SNPs.

Next we include all SNPs in each penalized regression model and vary the tuning parameter λ (Figure 9). Again it seems that, with a large candidate model containing a large number of predictors, ridge regression performs less well than the other three penalized methods, perhaps due to the former's inability for variable selection. LASSO and elastic net with α= 0.5 are the winners.

5 DISCUSSION

The primary objective of our study was to provide insight into general categories of models for which penalized regression improved disease risk prediction and classification for GWAS data. More specifically, we investigated the performance of MLE, LASSO, SCAD, ridge, elastic net and TLP regression methods for four different true models. The four models were chosen to represent broad categories defined by sparsity and strength of SNPs associated with disease. Two sparse models were considered with strong or moderate association strengths of only 10 causal SNPs. Two non-sparse models included 300 and 900 causal SNPs with weak effects respectively. Overall, we confirmed the commonly held belief that penalized regressions based on the model sparsity assumption, such as LASSO, SCAD, TLP and elastic net weighted towards its LASSO component were most suitable for sparse true models. This was true for both risk prediction and discrimination. However, we did discover that when effect sizes were strong in a sparse model, MLE performed as well. An interesting result was about how the penalized regressions used the information (or lack of information) when many SNPs were considered, in particular SNPs that would not meet a strict genome-wide significance level. As a rule, if a various number of top SNPs ranked by their marginal association significance are allowed to enter into a model, the LASSO and SCAD regressions were able to detect and thus ignore many unassociated SNPs in sparse model settings, while ridge regression was able to outperform LASSO and SCAD for non-sparse models with many SNPs with only weak associations. This may be important going forward as non-sparse and polygenic models may hold for many common diseases and complex traits. For sparse models the TLP's performance was comparable to LASSO and SCAD, but it outperformed LASSO and SCAD, but not ridge, when the true model was non-sparse with many weakly associated SNPs. The elastic net demonstrated the value in both variable selection and continuous shrinkage features of a penalty as it was able to adapt to the true underlying model and yield the best or nearly the best performance of all penalties. It is noteworthy, though, that the elastic net did not uniformly outperform either TLP or SCAD, in particular the TLP performed best on the real Crohn's disease and bipolar disorder data in the modeling scenario where the number of input SNPs was varied.

We have focused on penalized regression methods, but Bayesian approaches (Guan and Stephens, 2011) are also potentially useful and worth further investigation, which however is beyond the scope of this paper.

The current statistical research on high-dimensional data has largely focused on sparse models, yielding many important and insightful results. Nonetheless, non-sparse models are also useful, as manifested by polygenic models for complex and common diseases. There are few theoretical studies on non-sparse models; an exception is the work of Cook et al. (2012) on dimension reduction. The main message of our study, certainly not new, is that different penalized methods may be more suitable depending on the underlying architecture

of the true model: for example if the model is sparse or non-sparse. Hopefully this will prompt more empirical and theoretical investigations for non-sparse models.

Acknowledgments

We thank the reviewers and editors for many helpful and constructive comments. This research was supported by NIH grants R01HL65462, R01HL105397 and R01GM081535. “This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113”

References

About Crohn’s Disease. 2009. Retrieved from http://www.ccfa.org/info/about/crohns

Bipolar disorder. 2011. Retrieved from http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001924/

Cook RD, Forzani L, Rothman AJ. Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. Annals of Statistics. 2012; 40:352–384.

Crohn’s Disease. 2010. Retrieved from http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001295/

Evans D, Visscher P, Wray N. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Human Molecular Genetics. 2009; 18:3525–3531. [PubMed: 19553258]

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of American Statistical Association. 2001; 96:1348–1360.

Franke A, McGovern D, Barrett J, Wang K, Radford-Smith G, Ahmad T, Lees C, Balschun T, Lee J, Roberts R, Anderson C, Bis J, Bumpstead S, Ellinghaus D, Festen E, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew C, Montgomery G, Prescott N, Raychaudhuri S, Rotter J, Schumm P, Sharma Y, Simms L, Taylor K, Whiteman D, Wijmenga C, Baldassano R, Colombel J, Cottone M, Stronati L, Denson T, De Vos M, D’Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery S, Halfvarson J, Verspaget H, Hugot J, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor D, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart A, Stokkers P, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan S, Brant S, Rioux J, D’Amato M, Weersma R, Kugathasan S, Griffiths A, Mansfield J, Vermeire S, Duerr R, Silverberg M, Satsangi J, Schreiber S, Cho J, Annese V, Hakonarson H, Daly M, Parkes M. Genome-wide meta-analysis increases to 71 the number of confirmed Crohns disease susceptibility loci. Nature Genetics. 2010; 42:1118–1125. [PubMed: 21102463]

Friedman J, Hastie T, Tibshirani Rl. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2008; 33:1–22. [PubMed: 20808728]

Gail M. Discriminatory Accuracy From Single-Nucleotide Polymorphisms in Models to Predict Breast Cancer Risk. Journal Natl Cancer Inst. 2008; 100:1037–1041.

Gail M. Value of Adding Single-Nucleotide Polymorphism Genotypes to a Breast Cancer Risk Model. Journal Natl Cancer Inst. 2009; 101:959–963.

Gaya D, Russell R, Nimmo E, Satsangi J. New genes in inflammatory bowel disease: lessons for complex diseases? Lancet. 2006; 367:1271–1284. [PubMed: 16631883]

Guan Y, Stephens M. Bayesian Variable Selection Regression for Genome-wide Association Studies, and other Large-Scale Problems. Annals of Applied Statistics. 2011; 5(3):1780–1815.

Hoerl A, Kennard R. Ridge regression: Biased estimation for non-orthogonal problem. Technometrics. 1970; 12:55–67.

Hoggart C, Clark T, Iorio M, Whittaker J, Balding D. New genes in inflammatory bowel disease: lessons for complex diseases? Genetic Epidemiology. 2008; 32:179–185. [PubMed: 18200594]

Kang J, Kugathasan S, Georges M, Zhao H, Cho JH. the NIDDK IBD Genetics Consortium. Improved risk prediction for Crohn’s disease with a multi-locus approach. Human Molecular Genetics. 2011; 20:2435–2442. [PubMed: 21427131]

Kooperberg C, LeBlanc M, Obenchain V. Risk Prediction Using Genome-Wide Association Studies. Genetic Epidemiology. 2010; 34:643–652. [PubMed: 20842684]

Kraft P, Hunter D. Genetic Risk Prediction - Are We There Yet? New England Journal of Medicine. 2009; 360:1701–1703. [PubMed: 19369656]

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM. Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. PLoS Genet. 2008; 4(10):e1000231. [PubMed: 18949033]

McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, Ioannidiis J, Hirschhorn J. Genome-wide Significance for Dense SNP and Resequencing Data. Nature Genetics. 2008; 9:356–369.

Park J, Wacholder S, Gail M, Peters U, Jacobs K, Chanock S, Chatterjee N. Estimating effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics. 2010; 42:570–575. [PubMed: 20562874]

Shen X, Huang H-C. Optimal model assessment, selection, and combination. JASA. 2006; 101:554–568.

Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. Journal of American Statistical Association. 2012; 107:223–232.

The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Association, Series B. 1996; 58:267–288.

Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N. Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. Journal Natl Cancer Inst. 2004; 96:434–442.

Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SF, Polychronakos C, Hakonarson H. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genetics. 2009; 5(10):e1000678. Epub 2009 Oct 9. [PubMed: 19816555]

Wray N, Goddard M, Visscher P. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007; 17:1520–1528. [PubMed: 17785532]

Wu J, Devlin B, Ringquist S, Trucco M, Roeder K. Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic Epidemiology. 2010; 34:275–285. [PubMed: 20088021]

Yang J, Benyamin B, McEvoy B, Gordon S, Henders A, Nyholt D, Madden P, Heath A, Martin N, Montgomery G, Goddard M, Visscher P. Commom SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010; 42:565–569. [PubMed: 20562875]

Yang S, Yuan L, Lai Y, Shen X, Wonka P, Ye J. Feature grouping and selection over an undirected graph. KDD. 2012

Yuan Z, Yang Y. Combining linear regression models: when and how? JASA. 2005; 100:1202–1214.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B. 2005; 76:301–320.

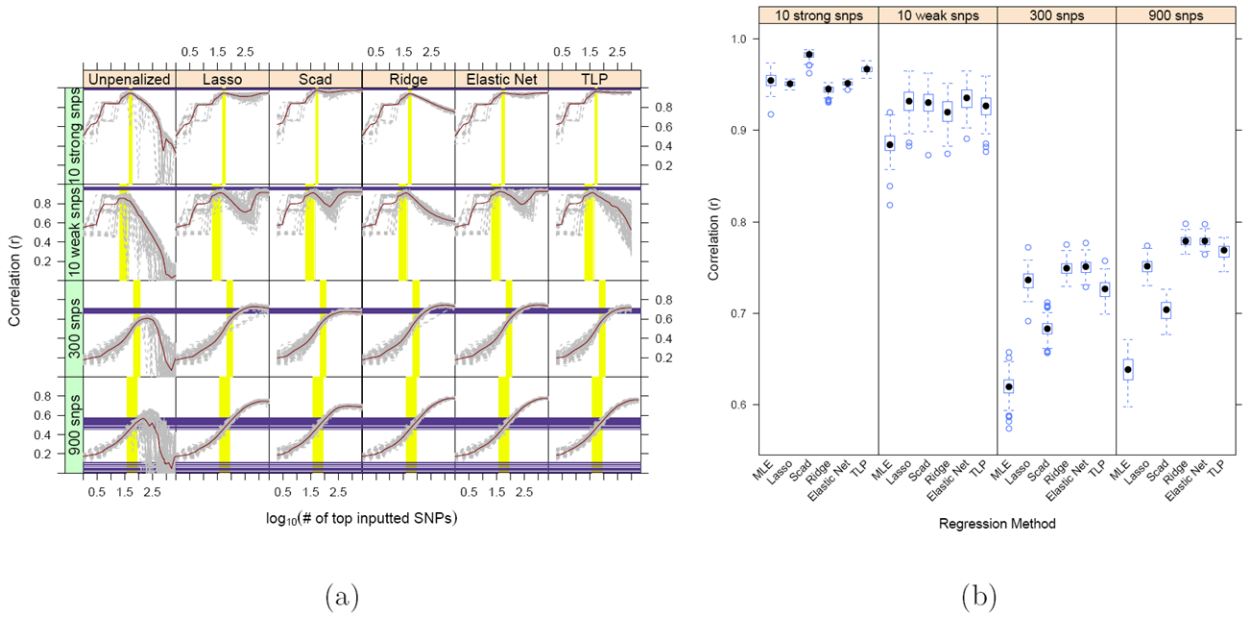


Figure 1.

Correlation of the true π_i and the $\hat{\pi_i}$ estimated with various numbers of top SNPs. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum correlation obtained for each simulated dataset across the number of top SNPs.

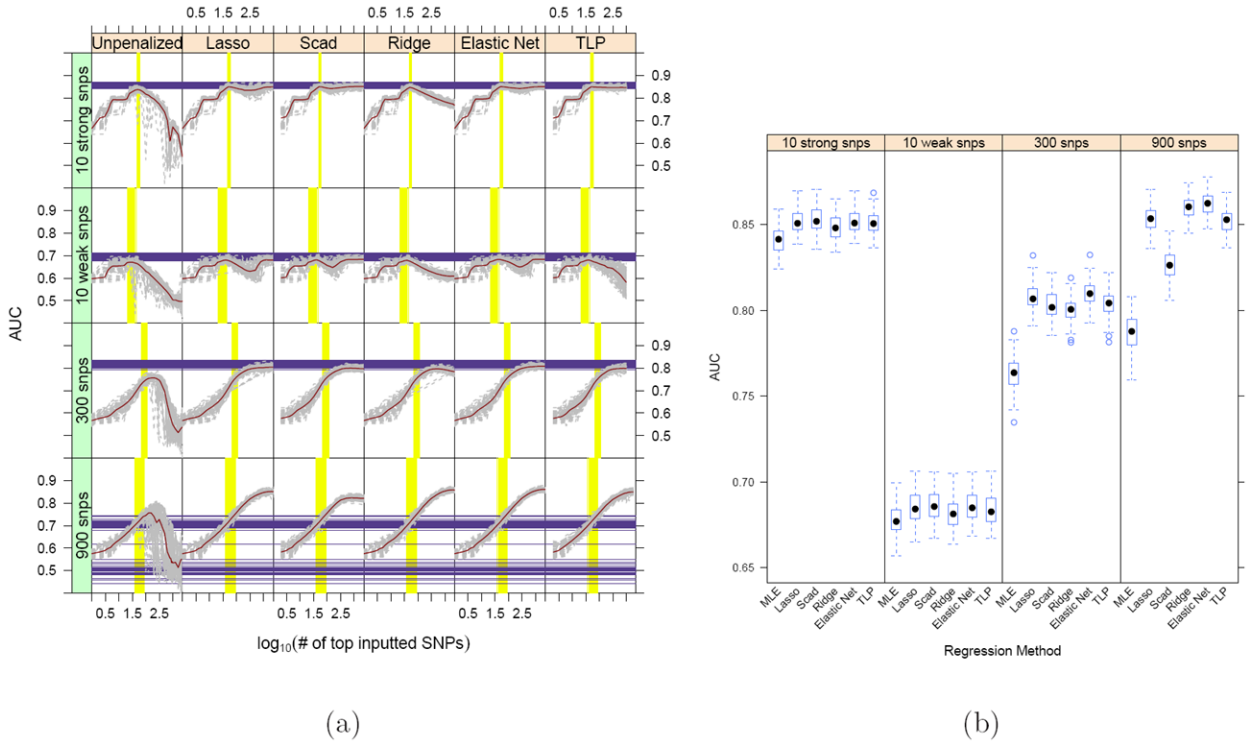


Figure 2.

AUC calculated for 100 simulated test datasets with various numbers of top SNPs. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum AUC obtained for each simulated dataset across the number of top SNPs

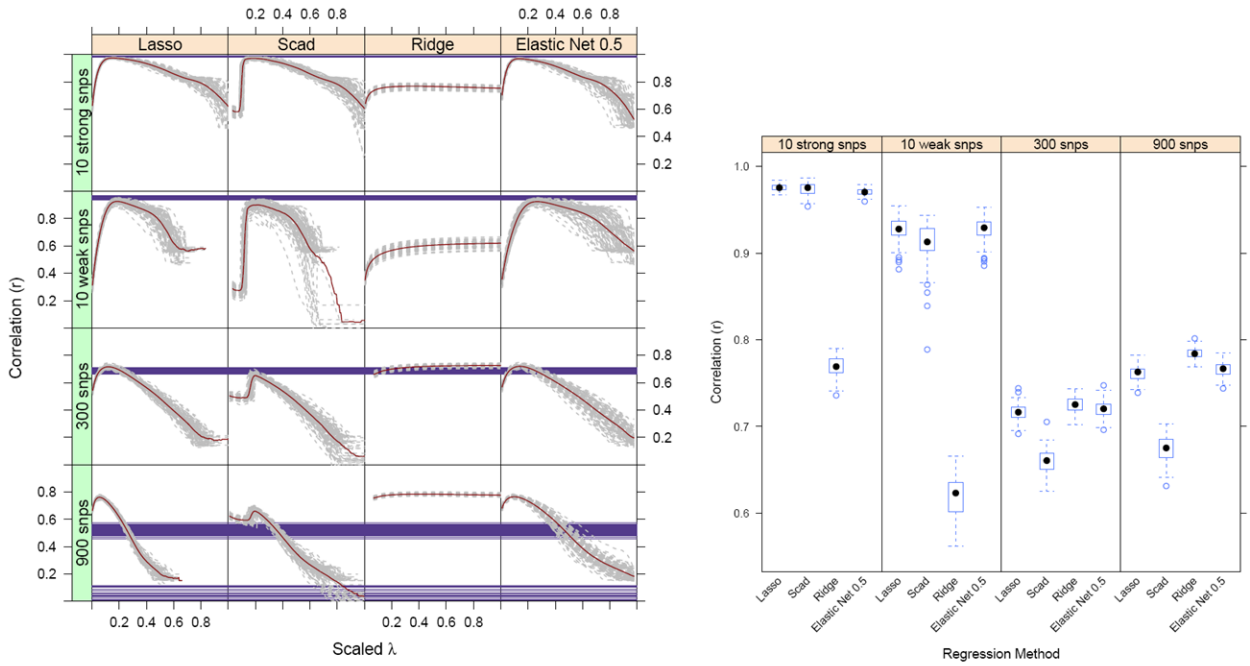
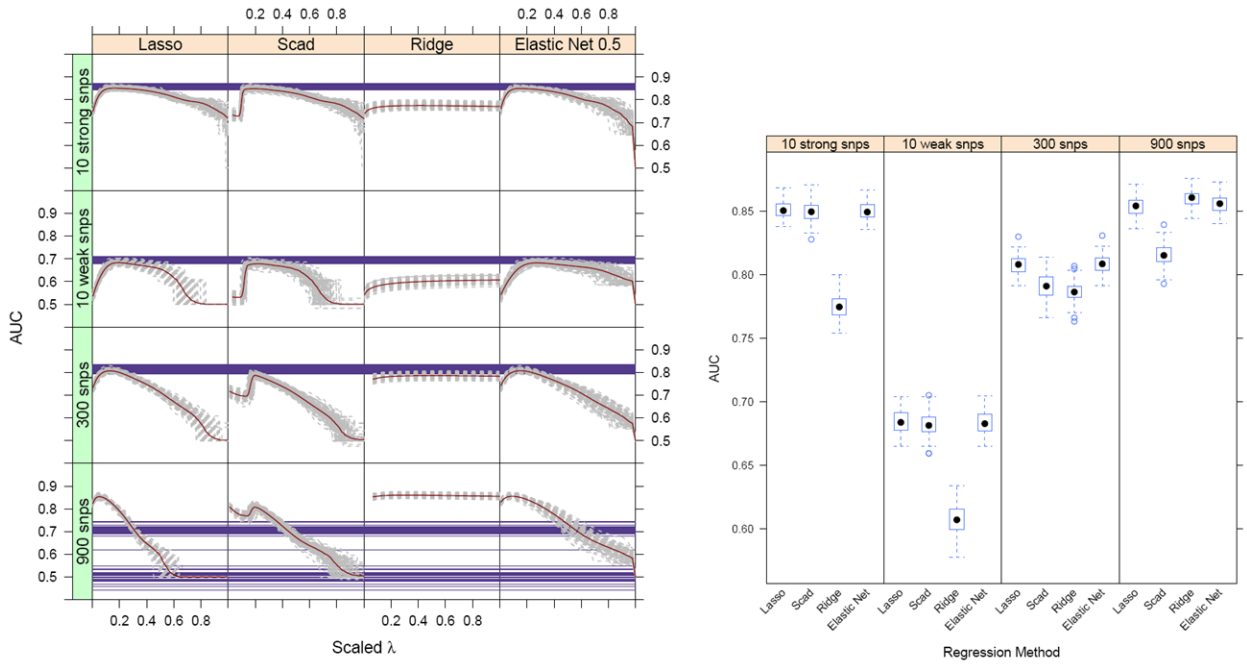


Figure 3.

Correlation of the true π_i and the $\hat{\pi_i}$ estimated from all SNPs with various values of regularization parameter λ. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum correlation obtained for each simulated dataset across the values of λ.



(a) (b)

Figure 4.

AUC calculated for 100 simulated test datasets from all SNPs with various values of λ. (a) Each panel displays the performance of a regression method (column) when estimating a true model (row). (b) Boxplots of the maximum AUC obtained for each simulated dataset across the values of λ.

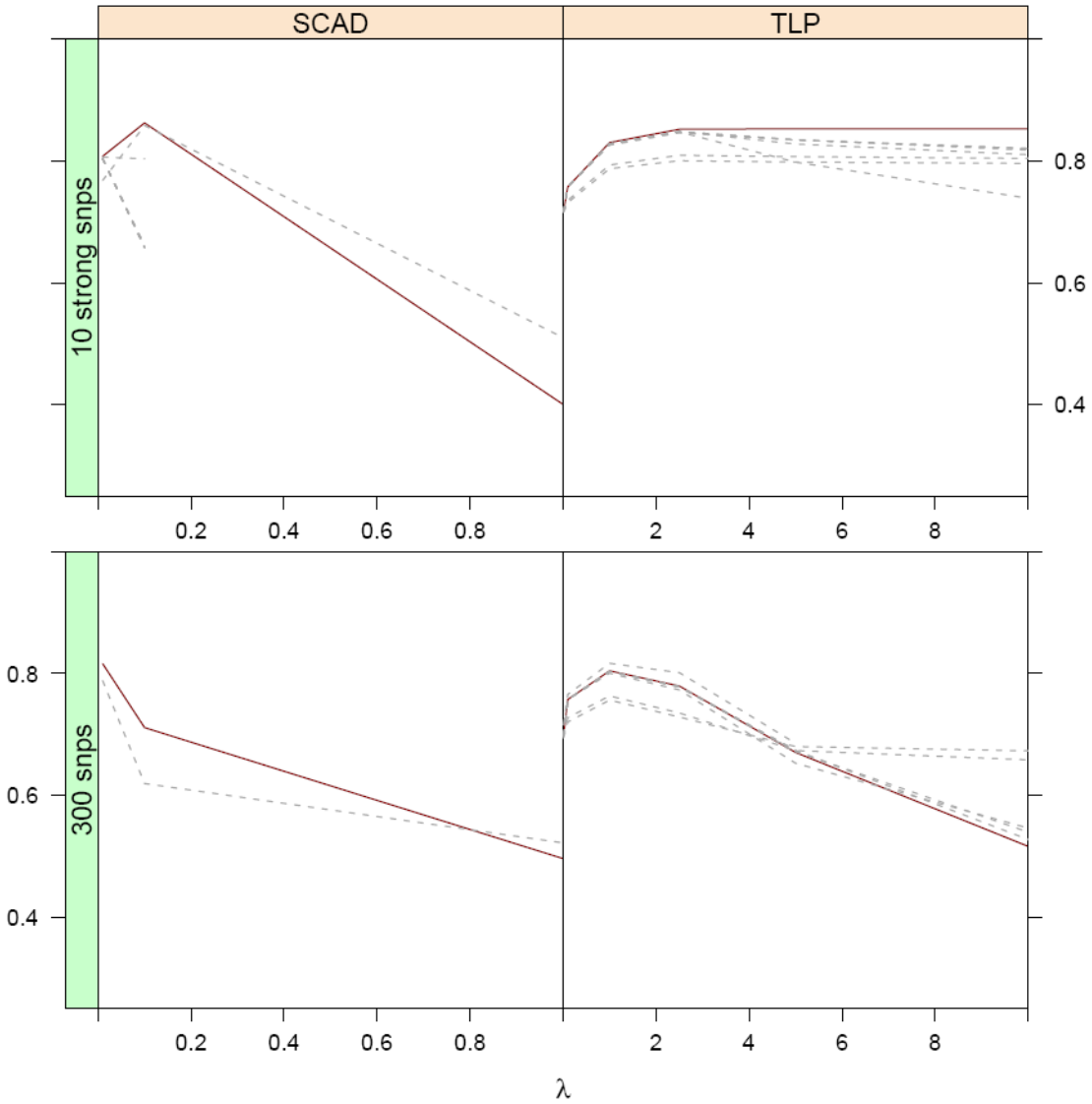


Figure 5.
Results of SCAD and TLP with various starting values.

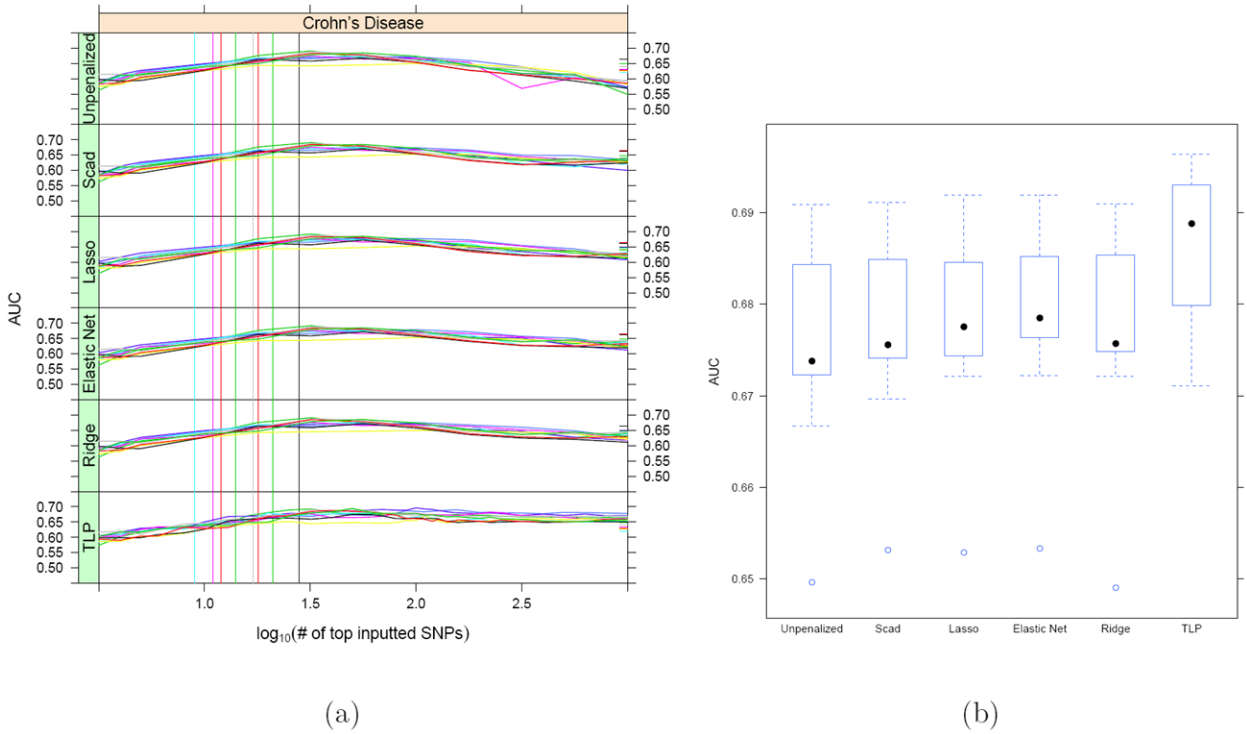


Figure 6.
AUC calculated for the Crohn's disease test datasets with various numbers of top SNPs. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the number of top SNPs.

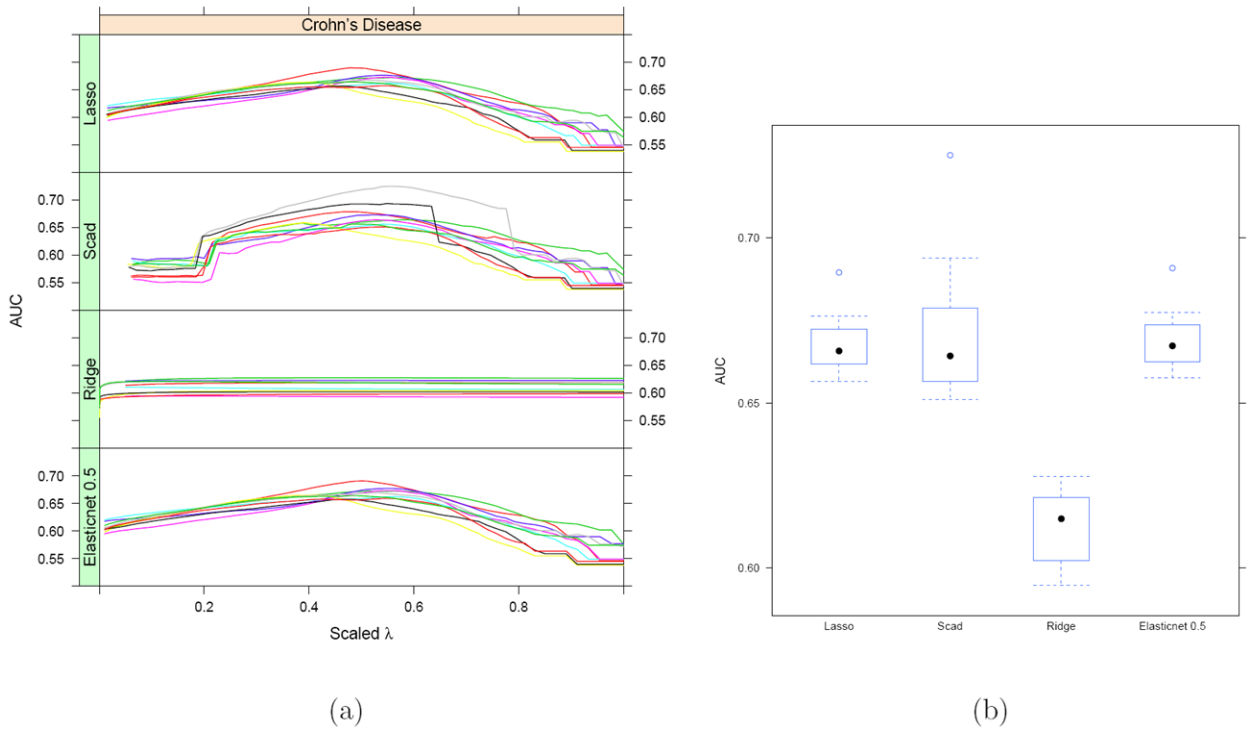


Figure 7.

AUC calculated for the Crohn's disease test datasets with all SNPs across the values of λ. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the values of λ.

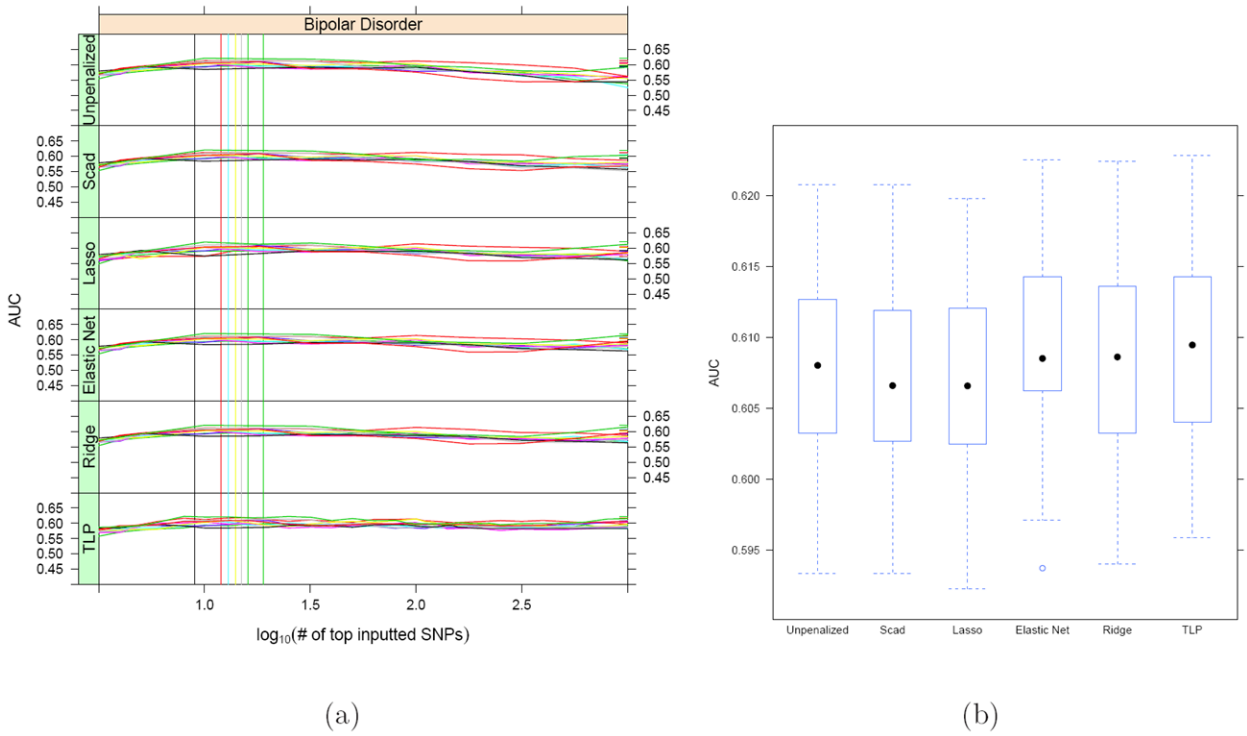


Figure 8.

AUC calculated for the bipolar disorder test datasets with various numbers of top SNPs. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each dataset across the number of top SNPs.

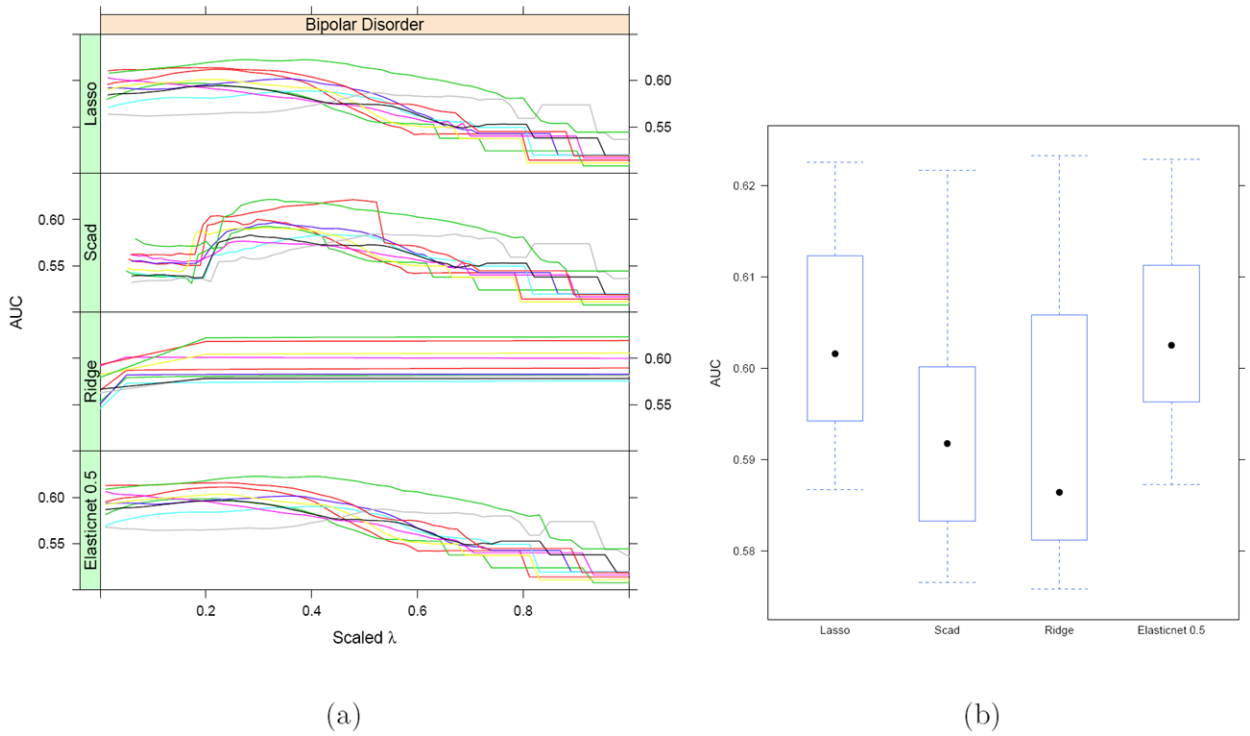


Figure 9.
AUC calculated for the bipolar disorder test datasets with all SNPs across various values of λ. (a) Each panel displays the performance of one regression method. (b) Boxplots of the maximum AUC obtained for each data across the values of λ.

Table 1

Summary Statistics for All Pairwise Correlations among the top *p* SNPs.

p	Minimum	1st Quartile	Median	3rd Quartile	Maximum
5	-0.096	-0.026	-0.003	0.002	0.019
10	-0.040	-0.013	-0.001	0.017	0.216
50	-0.136	-0.012	-0.001	0.011	0.994
100	-0.408	-0.013	0	0.012	0.999
500	-0.668	-0.013	0	0.013	1
1000	-0.835	-0.013	0	0.013	1

Table 2

Maximum mean correlations (Corr) and AUCs for each method under either varying or fixed numbers of input SNPs.

Model/Data	#SNPs	Metric	MLE	SCAD	LASSO	Elastic Net	Ridge	TLP
10 Strong SNPs	Varying	Corr	0.954	0.982	0.951	0.951	0.944	0.966
		AUC	0.841	0.853	0.852	0.852	0.849	0.851
	Fixed	Corr	-	0.974	0.975	0.970	0.769	-
		AUC	-	0.850	0.851	0.850	0.775	-
10 Weak SNPs	Varying	Corr	0.885	0.931	0.931	0.935	0.920	0.925
		AUC	0.678	0.686	0.685	0.686	0.682	0.684
	Fixed	Corr	-	0.912	0.928	0.927	0.620	-
		AUC	-	0.682	0.684	0.684	0.607	-
300 SNPs	Varying	Corr	0.619	0.683	0.735	0.750	0.749	0.726
		AUC	0.763	0.803	0.808	0.810	0.800	0.804
	Fixed	Corr	-	0.659	0.716	0.720	0.725	-
		AUC	-	0.791	0.808	0.808	0.786	-
900 SNPs	Varying	Corr	0.638	0.702	0.751	0.779	0.779	0.767
		AUC	0.787	0.827	0.854	0.862	0.860	0.852
	Fixed	Corr	-	0.674	0.761	0.766	0.784	-
		AUC	-	0.815	0.854	0.856	0.860	-
Crohn's Disease	Varying	AUC	0.675	0.677	0.678	0.678	0.677	0.686
	Fixed	AUC	-	0.672	0.668	0.660	0.612	-
Bipolar Disorder	Varying	AUC	0.607	0.606	0.606	0.609	0.608	0.609
	Fixed	AUC	-	0.595	0.602	0.603	0.594	-

Stat Anal Data Min. Author manuscript; available in PMC 2013 December 11.