University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2011 *Paper* 274

A Generalized Approach for Testing the Association of a Set of Predictors with an Outcome: A Gene Based Test

Benjamin A. Goldstein* Alan E. Hubbard†

Lisa F. Barcellos‡

*University of California - Berkeley, ben.goldstein@stanford.edu

†University of California, Berkeley

‡University of California, Berkeley

A Generalized Approach for Testing the Association of a Set of Predictors with an Outcome: A Gene Based Test

Benjamin A. Goldstein, Alan E. Hubbard, and Lisa F. Barcellos

Abstract

In many analyses, one has data on one level but desires to draw inference on another level. For example, in genetic association studies, one observes units of DNA referred to as SNPs, but wants to determine whether genes that are comprised of SNPs are associated with disease. While there are some available approaches for addressing this issue, they usually involve making parametric assumptions and are not easily generalizable. A statistical test is proposed for testing the association of a set of variables with an outcome of interest. No assumptions are made about the functional form relating the variables to the outcome. A general function is fit using any statistical learning algorithm, with the SuperLearner algorithm suggested. The parameter of interest is the cross-validated risk and this is compared to an expected risk. A Wald test is proposed using the influence curve of the cross-validated risk to obtain the variance. It is shown both theoretically and via simulation that the test maintains appropriate type I error control and is more powerful than parametric tests under more general alternatives. The test is applied to an MS candidate gene study. Three separate analyses are performed highlighting the flexibility of the approach.

1 Introduction

In many statistical problems one desires to relate a set of variables to an outcome. For example, it is typical in the social sciences to have data on race, income, education etc. and want to draw inference about the relationship between some outcome and socio-economic status (SES). SES itself is not observed but is instead a combination of the aforementioned variables (as well as others). Some approaches for answering such a question include F-tests and likelihood ratio tests, Fishers's method for combining p-values [Fisher 1948], and principal components regression. In addition to these generalized approaches, many discipline specific measures have been developed. For example, in psychology it is common to come up with "scores" on different survey instruments. While useful all of these methods suffer from two primary limitations. Firstly, they often rely on parametric modeling assumptions and secondly they often do not take into account the complex relationships of the variables.

Since the underlying relationship between a set of variables and an outcome is usually quite complex and unknown, ideally, instead of specifying a model relating the set of variables to the outcome one would be able to search for the best relationship. Typical statistical learning and prediction methodology is well suited for solving such problems. Statistical learning algorithms apply a basis function (or set of basis functions) to the data to find the best relationship to the outcome. While the best algorithm will depend on the true relationship between the predictor variables and the outcome, most are well suited for situations where the relationship is complex and/or of high dimension. Typically the focus is on trying to get the best estimate of the function relating the data to the outcome, but in recent years there has been a growing emphasis on variable importance (VI). However, most of the VI measures are ad-hoc and do not have sound statistical properties with a clear parameter of interest, though there is some work trying to formalize VI and attach statistic properties for more targeted analyses [van der Laan 2006]. Moreover, like with inferential statistics there is not a best means to relate a group of variables to an outcome.

The goal of this paper is to establish a statistical test for assessing the relationship between a set of variables and an outcome. Using tools from statistical learning, a general function is estimated. Others have assessed this relationship using a full permutation test (e.g. Radmacher et al. [2002]; Birkner et al. [2005]; Chaffee et al. [2010]), however, for all but the simplest algorithms, this is computationally infeasible. Instead, a simple to calculate statistic is proposed. The parameter to be evaluated is the risk between a predicted value and the observed value. This observed risk is compared to an expected risk via a Wald test. A rejection of the null hypothesis that the observed risk is less than the expected risk indicates that the prediction is better than would be expected by chance and that the set of predictors are related to the outcome.

The need for such a statistic can be motivated from two different perspectives: inferential statistics and statistical learning. From an inferential statistics perspective, such an approach can be used to test the association of a group of variables and an outcome (as motivated in the outset). Typically, the group of variables represent an unobserved

construct such as SES, a gene, or a stock index. From a statistical learning perspective, the current statistical test represents a means to test whether the prediction derived from a machine learning algorithm is better than what would be expected by chance. This is a question that is not typically asked within the machine learning literature as practitioners generally assume that the set of predictors is related to the outcome and a *significant* prediction can always be derived. While this is not an unreasonable assumption, with the growing ubiquity of prediction methodology, these algorithms are more often being applied to data that may not actually have predictive power (this is particularly true within genetic epidemiology). Therefore, the current method should have broad interest and applicability to both those who are primarily interested in inference as well as those that are primarily interested in prediction.

The paper is organized as follows. In section 2 some preliminaries on statistical learning and loss based estimation via cross-validation are presented. In section 3 the statistical test is proposed with discussion of how to estimate each of the parameters. The next section provides a brief overview of related literature. Section 5 shows some basic simulation results. Section 6 presents an application to genetic epidemiology data. Section 7 provides some concluding thoughts.

2 Preliminaries

We begin with the observed data $W_i = (Y_i, X_i) \sim P_W, i = 1, \ldots, n$. The Y_i are the outcome of interest and $Y_i \in \Re^1$. The Y_i can be continuous or binary. The X_i are the covariates, a $p-$dimensional vector, where $X_i \in \Re^p$. We relate X to Y by the functional transformation $f(\cdot)$:

$$\begin{aligned} E(Y|X) &= f(X) \\ P(Y = 1|X) &= f(X) \text{ for } Y \in \{0, 1\} \end{aligned} \tag{2.1}$$

In (2.1) we make no assumptions about the function form of $f(\cdot)$.

2.1 Statistical Learning

Statistical learning is concerned with estimating $f(\cdot)$ for the purposes of predicting future outcomes based on an observed covariate vector. All statistical learning algorithms provide a different means of Wrt to Paul's work it is definitely a more computationally friendly version b/c it doesn't require a complete permutation. I think it can also be viewed as an extension of yours and Sandrine's workestimating this function. Table 1 lists a range of different learning algorithms. Each algorithm applies a different type of basis function to the data. All algorithms also have a different set of tuning parameters (many have multiple). Changing the tuning parameters optimizes the algorithm for the specific data problem.

Learner	Type of Function	Tuning Parameters	Speed
Regression	Linear Relationship	Variables in Model	Fast
Lasso/Ridge Regression	Penalized Regression	Penalty	Moderate
Nearest Neighbors	Classification based on Proximity	Number Neighbors	Fast
CART	Tree	Tree Depth	Moderate
Splines	Piecewise Functions	Knots	Fast
Support Vector Machines	Transformation of Output Space	Transformation	Slow

Table 1: Different Learning Algorithms

Not all algorithms are appropriate for all data problems. For example, if there is a lot of additive structure in the data a linear algorithm will do much better than a tree based algorithm. Conversely, if there are many interactions, then a tree based algorithm would be preferable. Since the choice of best algorithm is dependent on the true underlying function, $f(\cdot)$, which is unknown, it is impossible to know which is best to use.

With this limitation in mind, as computational power has increased, there has been a growing use of ensemble based learners. Ensemble learning is a process of combining multiple learners (typically weak ones) together into one meta learner. There are many different types of ways to ensemble algorithms with the primary methods being: bagging, boosting, Bayesian model averaging and stacking. Ensemble algorithms differ in what their base learners consist of and how the algorithms are combined (i.e. the weights placed on each algorithm). For example the Random Forests algorithm [Breiman 2001] is an ensemble based algorithm where the base learners are unpruned CART trees and they are combined via bagging, a process of adding equal weight to all learners.

Different ensemble algorithms will have different strengths and weaknesses. In the present work, the goal is to use the algorithm that best estimates the underlying function $f(\cdot)$. The algorithm that has been found to be most adapted for this problem is the SuperLearner (SL) algorithm [van der Laan et al. 2007]. SL is an algorithm based on stacking [Wolpert 1992]. In stacking based algorithms a library of algorithms is applied to the data and each algorithm provides a predictions of the outcome via cross-validation (CV). The predictions for each observation are *stacked*, creating a matrix of predicted outcomes for each of the j learners. The true outcome is regressed onto the predicted values. The derived coefficients provide the weights for each algorithm.

In the implementation of SL, available in SuperLearner package in R, the regression performed is a non-negative least squares and the coefficients are scaled to sum to 1. Other authors [Breiman 1996; Ting and Witten 1997] have similarly found non-negative least squares to be the optimal majorizing function. While typical implementations of stacking involve using similar base learners with different tuning parameter settings, van der Laan

et al. advocate using a full library of different types of learners covering a range of basis functions. The authors were able to show that stacking satisfies certain oracle properties which we repeat here:

Oracle Inequalities: Let $d_o(\psi, \psi_0) = E_{P_X}\{L(X, \psi) - L(X, \psi_0)\}$ be the risk difference between the candidate estimate ψ and the true parameter value ψ_0. Also, suppose the $P\{(\hat{\Psi}_k(P_n) \in \Psi) : \forall k\} = 1$. Assume:

A1: $L(X, \psi)$ is uniformly bounded

A2: The variance of the ψ_0-centered loss function $(L(X, \psi) - L(X, \psi_0))$ can be uniformly bounded by its expectation uniformly in ψ.

then, for any $\lambda > 0$:

$$Ed_0(\Psi_{\hat{K}(P_n)}(P_{n,T(V)}), \psi_0) \leq (1 + 2\lambda)Ed_0(\Psi_{\tilde{K}(P_n)}(P_{n,T(V)}), \psi_0) + 2C(\lambda)\frac{1 + \log(K(n))}{np}$$

where p is the proportion of the observations in the validation sample and $C(\lambda)$ is a constant defined in van der Laan et al. [2006].

These results imply that the SL performs as well as the oracle selector in terms of expected risk difference and as long as the number of candidate learners $(K(n))$ is polynomial in sample size, the SL is the optimal learner. Moreover, if one of the candidate learner searches within a parametric model and that model contains the truth, then the SL attains an almost parametric rate of convergence $\log n/n$. This makes the SL an ideal learner when the true underlying function is unknown.

In practice, implementing the SL is fairly straightforward. The main tuning parameters include selecting the candidate library, the number of CV splits, and the majorizing function. The majorizing function, as mentioned, is typically non-negative least squares. While the larger the number of CV splits the better the estimate of the function, there is obviously a computational trade off. Typically 10-fold or 20 fold CV has been found to be appropriate. The most important aspect is the candidate library. Again, while one may say "the more the better" (up to a limit), in practice fitting each candidate can be highly computational and it is worth being judicious in the choice of candidate learners. Generally, then, it is best to apply SL with a library that spans a range of basis functions. One final note about the computation is that it is fairly straightforward to implement the SL either within a cloud environment or across nodes in a parallel environment.

2.2 Loss Based Estimation via Cross-Validation

Once the learning algorithm is fit to the data a prediction, $\hat{f}(X_i)$, is generated for each observation, Y_i. The goal in prediction is to minimize the risk over the training set:

$$argmin_f E[L(Y, \hat{f}(X)] \tag{2.2}$$

The "harder" $\hat{f}(\cdot)$ is fit to the data, the greater the potential for over-fitting, referred to as the *optimism*, and, consequentially under-estimating the risk [Hastie et al. 2009]. There are two main approaches for correcting for over-fitting. The first is by directly estimating the optimism and adding this to the estimated training error (e.g. AIC, BIC, MDL). The other is to directly estimate the test error (e.g. CV, Bootstrap methods). Direct methods are useful when the number of basis functions (effective number of parameters) are easily calculable (e.g. linear models, regularized regression). For more complex methods (including all ensembles), such calculations are intractable. CV methods provide the simplest approach to obtaining an honest estimate of $\hat{f}(X)$, and consequently the risk.

In all CV methods, the data are divided into a *training* and *validation* set. The estimator is computed (*trained*) on the training set and then tested (*validated*) on the remaining validation set. This process is iterated, allowing each observation to be part of the validation set, providing an unbiased estimate of the risk of the estimator [Dudoit and van der Laan 2005].

Using notation from Dudoit and van der Laan, we define a binary random vector, $B_n = (B_n(i) : i = 1, \ldots, n) \in \{0,1\}^n$, independent of the empirical distribution P_n. A realization, $B_n(i)$ represents an indicator of whether an observation is in the training or validation set:

$$B_n(i) = \begin{cases} 0, & i\text{th observation } X_i \text{ is in the } \textit{training} \text{ set,} \\ 1, & i\text{th observation } X_i \text{ is in the } \textit{validation} \text{ set,} \end{cases} \tag{2.3}$$

Let P^0_{n,B_n} and P^1_{n,B_n} denote the empirical distributions of the training and validation sets, respectively, and let the number and proportion of observations in the validation sets be denoted by $n_1 \equiv \sum_i B_n(i)$ and $p = p_n \equiv n_1/n$, respectively. Then a definition of the cross-validated risk estimator for $\psi_n = \hat{\Psi}(P_n)$ is

$$\begin{aligned} \hat{\theta}_{p_n,n} &\equiv E_{B_n} \Theta(\hat{\Psi}(P^0_{n,B_n}), P^1_{n,B_n}) \\ &= E_{B_n} \int L(x, \hat{\Psi}(P^0_{n,B_n})) dP^1_{n,B_n}(x) \\ &= E_{B_n} \frac{1}{n_1} \sum_{i:B_n(i)=1}^{n_1} L(X_i, \hat{\Psi}(P^0_{n,B_n})) \end{aligned} \tag{2.4}$$

where $\hat{\Psi}(P^0_{n,B_n})$ represents the estimator of the parameter ψ based on the training set.

The distribution of B_n determines the type of CV. The most common type of CV is *V-fold cross-validation*. In V-fold CV the learning set is randomly divided into V mutually exclusive sets of approximately equal size. Each set is then used in turn as a validation set (see figure 1). The distribution of B_n places mass $1/V$ on each of V binary vectors such that $\sum_i b^v_n(i) \approx n/V \ \forall \ v$ and $\sum_v b^v_n(i) = 1 \ \forall \ i$. Other forms of cross-validation

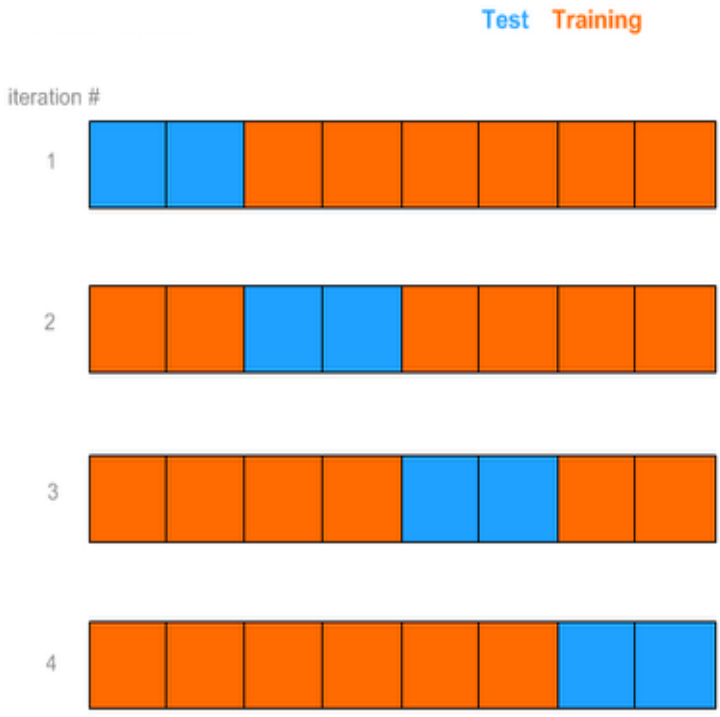

Figure 1: Illustration of 4-fold CV where $n = 8$ and $n_1 = 2$. In each cycle, 6 observations are used to train the learner and 2 are used to test or validate it. Courtesy of maxdama.com

include *Leave-one-out cross-validation*, *Monte Carlo cross-validation* and *Bootstrap-based cross-validation* [Dudoit and van der Laan 2005].

Dudoit and van der Laan showed that the risk estimate provided through CV is asymptotically linear with appropriate assumption (see Theorem 3 in their paper) and has influence curve (IC):

$$\begin{aligned} IC &\equiv L[Y, f(X)] - \theta \Rightarrow \\ \hat{\theta} &\cong \frac{1}{n}\sum_{i=1}^{n} L[Y_i, \hat{f}(X_i)] - \hat{\theta} \end{aligned} \tag{2.5}$$

The benefit of defining the IC of a parameter is that one can use the variance of the IC to obtain the variance of the estimator. Dudoit and van der Laan use this result to construct

confidence intervals for the risk estimate

$$\hat{\theta}_n \pm z_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}$$

where $\hat{\theta}_n$ is the estimated cross-validated risk and σ_n is the standard deviation of the cross-validated loss. Asymptotically the observed standard deviation, $\hat{\sigma}$, is an appropriate estimator. Bengio and Grandvalet [2004] showed that while there is no unbiased estimate for σ in finite samples, it does converge to the observed standard deviation fairly rapidly (see Section 4 for discussion).

3 The proposed test statistic

The parameter of interest, θ, is the risk from our prediction algorithm:

$$\begin{aligned} \theta &\equiv EL(Y, f(X)) \\ \hat{\theta} &\equiv EL(Y, \hat{f}(X)) \end{aligned} \tag{3.1}$$

based on the model defined in (2.1). While Dudoit and van der Laan [2005] focused on generating confidence intervals for the observed risk, $\hat{\theta}$, the present interest is in hypothesis testing. We can test the hypothesis:

$$\begin{aligned} H_o &: \theta \geq \theta^* \\ H_a &: \theta < \theta^* \end{aligned} \tag{3.2}$$

This is a one-sided hypothesis test of whether the observed risk, $\hat{\theta}_n$ is less than some expected risk under a null hypothesis, θ_n^*. We define

$$\theta^* \equiv E[L(Y, f(X))] s.t. Y \perp X \tag{3.3}$$

We can then use the same asymptotic linearity result and define a Wald-type statistic, with parameter ψ:

$$\begin{aligned} Z &= \frac{\hat{\psi}\sqrt{n}}{\sqrt{\hat{var}(IC(W;\psi)}} \\ &= \frac{\hat{\theta}_n - \hat{\theta}_n^*}{\sqrt{var(\hat{\theta}_n - \hat{\theta}_n^*)}} \sim N(0,1) \end{aligned} \tag{3.4}$$

This will be a one-tailed test as we are only interested in the case where $\hat{\theta}_n < \hat{\theta}_n^*$.

From (3.4) there are three values that need to be estimated:

- The estimated risk: $\hat{\theta}_n$
- The estimated risk under the null: $\hat{\theta}_n^*$
- The variance of the difference of the two: $var(\hat{\theta}_n - \hat{\theta}_n^*)$

3.1 The Observed Risk

The observed risk is the simplest value to estimate. Whereas any loss function can be used, a loss that has certain asymptotic properties will be needed to allow for the use of the IC to calculate the asymptotic variance. Based on the work of Dudoit and van der Laan [2005], this includes most any loss, with the notable exception of misclassification loss.

For mathematical simplicity, that will be seen later, squared error (ℓ_2) loss is used, with the estimated risk being:

$$\hat{\theta}_n \equiv E_n L(Y, \hat{f}(X)) \equiv E_n (Y - \hat{f}(X))^2 \tag{3.5}$$

3.2 The Null Risk

The null risk is the expected value of the loss between the observed outcome and the predicted outcome, when the set of covariates, X, is independent of the outcome, Y. The most direct way to estimate this is by permuting the X values and retraining the predictor. For all but the simplest prediction algorithms, though, this can be computationally infeasible. However it is possible to estimate this value. We need:

$$\theta(P^*) = E_{P^*}[Y - f(X)]^2$$

Assuming that f(X) is of fixed form, and the learning algorithm has an intercept:

$$Ef(X) = \mu_Y = EY$$

Therefore, we get:

$$\begin{aligned} &E_{P^*}[(Y - \mu_Y + \mu_Y - f(X))^2] \\ &= E_{P^*}(Y - \mu_Y)^2 + E_{P^*}(f(x) - \mu_Y)^2 + 2E_{P^*}\big((Y - \mu_Y)(\mu_Y - f(X)\big) \end{aligned}$$

So:

$$\begin{aligned} \theta &= var(Y) + var\big(f(X)\big) - 2cov\big(f(X), Y\big) \\ \theta^* &= var(Y) + var\big(f(X)\big) \end{aligned} \tag{3.6}$$

Thus a test of the null that $\theta = \theta^*$ is a test of the $cov\big(f(X), Y\big) = 0$.

3.3 The Variance of the Difference of Risks

The final consideration is defining the variance of the difference of $\hat{\theta}_n$ & $\hat{\theta}_n^*$. Using the property of ICs we have:

$$var(\hat{\theta} - \hat{\theta}^*) = \frac{var(IC[\hat{\theta}] - IC[\hat{\theta}^*])}{n} \tag{3.7}$$

Therefore we need to define the IC for $\hat{\theta}$ and $\hat{\theta}^*$. We note the parameter of interest can be expressed as:

$$\psi \equiv E[Y - f(X)]^2 - E[Y - EY]^2 - E[f(X) - E(f(X))]^2 \tag{3.8}$$

This is simply the difference of three loss functions. Dudoit and van der Laan [2005], provided the framework for calculating the influence curve for any general loss function, noted in (2.5). Therefore:

$$\begin{aligned} IC(W;\psi) &= IC_1 - IC_2 - IC_3 \\ &= \big(L[Y, f(X)] - \theta\big) - \big(L[Y, EY] - \theta_Y\big) - \big(L[f(X), Ef(X)] - \theta_{f(X)}\big) \end{aligned} \tag{3.9}$$

By substituting the appropriate loss function (in this case ℓ_2) and taking the variance of (3.9) one can calculate the inference for the test statistic. There are now all of the elements of the test statistic and referring back to (3.4) we can write:

$$\begin{aligned} Z &= \frac{\hat{\psi}\sqrt{n}}{\sqrt{\hat{var}(IC(W;\psi)}} \\ &= \frac{\hat{\theta}_n - \hat{\theta}_n^*}{\sqrt{var(\hat{\theta}_n - \hat{\theta}_n^*)}} \\ &= \frac{\sqrt{n}\Big[E_n(Y - \hat{f}(X))^2 - E_n(Y - \bar{Y})^2 - var(\hat{f}(X))\Big]}{\sqrt{var\Big(E_n[(Y_i - \hat{f}(X_i))^2 + \hat{\theta}] - E_n[(Y_i - \mu)^2 + \hat{\sigma}_Y^2] - E_n[(\hat{f}(X_i) - E\hat{f}(X_i))^2 + \hat{\sigma}^2_{\hat{f}(X)}]\Big)}} \end{aligned} \tag{3.10}$$

The statistical test as constructed will be unbiased for predicting on an independent sample. However, while asymptotically the proposed test statistic will approach $N(0,1)$ in practice the estimated variance a bit too small. This is due to the excess correlation induced by cross-validation (see [Bengio and Grandvalet 2004]). Simulations showed that the asymptotics do not fully kick in until extremely large samples ($n > 1 \times 10^7$). Grandvalet and Bengio [2006] proposed a finite sample correction (see section 4). Based on experimental results it was found that the variance in (3.10) underestimated the true variance by a factor of 2 (see figure 2 in section 5). This is obviously a less than ideal solution and current work involves trying to determine a more theoretical finite sample correction.

To calculate the statistic in R code:

```
sqLOSS <- function(x,y)(x-y)^2

predTestL2 <- function(pred,Y,LOSS = sqLOSS){
  n <-length(Y)
  ll <- LOSS(pred,Y)  ###\hat\theta
  lo <- LOSS(mean(Y),Y) - LOSS(pred,mean(pred)) ###\hat\theta^*
  LossD <- mean(ll - lo)
  varIC <- var(ll - lo)*2
  Z <- LossD*sqrt(n)/sqrt(varIC)
  return(Z)
}
```

3.4 A Permutation Based Test

It is also possible to construct a significance test via the permutation distribution. To test the independence of Y and the p-vector X one could permute the Y and continually retrain the predictor [Radmacher et al. 2002]. However, for the most part this is too computational. Alternatively, one can test the independence of Y and $\hat{f}(X)$. To do so:

1. Train the predictor to obtain $\hat{f}(X)$
2. Calculate $\hat{\theta} = \sum_i L(Y_i, \hat{f}(X_i))$
3. Permute the Y_i b times and calculate $\hat{\theta}^*_j = \sum_i L(Y_i^*, \hat{f}(X_i))$ for $j \in 1 \ldots b$
4. The permutation based p-value is $\frac{1}{b} \sum_j I\left[\hat{\theta} < \hat{\theta}^*_j\right]$

The choice of b will depend on the desired precision of the empirical p-value, with stronger associations requiring larger b. To calculate this in R code:

```
predTestPerm <- function(pred,Y, p = 1000, LOSS = sqLOSS){
  Yp <- replicate(p, sample(Y))
  mZ <- median(LOSS(pred,Y))
  Zp <- apply(Yp,2,function(x)median(LOSS(pred,x)))
  pval <- 1 - sum(mZ < Zp)/p
  return(pval)
}
```

4 Previous Work with Assessing Prediction

In a general sense, the proposed test can be considered as analogous to an F-test, used in linear regression. Both approaches aim to test the goodness-of-fit of a fitted function based

on the residual fit. The primary distinction is that while an F-test relies on the correctly specifying the parametric form of $f(\cdot)$, the proposed test can be seen as a semi-parametric alternative that does not depend on a specified functional form (see Figure 5 in Section 5).

There has been some related work assessing the significance of a prediction. The work of Dudoit and van der Laan [2005] focused on providing a theoretical basis for calculating the standard-error of the CV risk estimate for the purpose of constructing confidence intervals. Their work focused on the asymptotic properties, showing that it is both consistent and asymptotically linear.

Bengio and Grandvalet [2004] were also interested in constructing confidence intervals for the CV risk estimate, however they focused on finite samples. The authors showed both theoretically and via simulation that in finite samples it is not possible to get an unbiased estimate of the variance of the cross-validated risk estimate. They broke down the variance into three components:

(1) The variability of the prediction within each validation block

(2) The covariance between predictions within each block

(3) The covariance between predictions in different blocks

The first value is the quantity of interest, however, simply taking the empirical variance of $\hat{f}(X)$ is biased by the other two quantities. However, the authors showed, with modest sample sizes ($n > 100 - 500$) these two values go to 0, resulting in the desired value. The authors proposed a corrected test statistic based on an assumed maximum between block correlation of 0.7 [Grandvalet and Bengio 2006]. This correction is similar to the proposed correction in the current work. The primary limitation of their formulation is that they construct a t-test against a fixed value. However, as shown above, if one wants to test against an expected risk, it is necessary to estimate θ^* and therefore the variance of the estimate also needs to be considered.

Dietterich [1998] showed also that it is only in small samples one needs to be concerned with the variance estimates of the cross-validated risk. This work provides further finite sample justification for the work of Dudoit and van der Laan and much of the present discussion.

Other work includes Radmacher et al. [2002] who laid a general framework for assessing the prediction in micro-array studies. The authors advocated permuting and then refitting the learner, to assess the risk estimate via cross-validation. However, they were working with much simpler learners. Others have used this full permutation approach for testing genetic pathways [Birkner et al. 2005] and perfoming gene set tests [Chaffee et al. 2010]. Lusa et al. [2007] noted that simply calculating the odds-ratio on a two-by-two table led to inflated type I error. This was also noted by Lee [2007]. This observation conforms to the theory presented in Dudoit and van der Laan, which showed that misclassification loss is not an appropriate loss function to use for the present work.

5 Simulations

To examine the behavior of the test statistic a series of simulations were undertaken. It is re-noted that the Z-test is a one-tailed test where the more negative the test statistic, the more significant the association. The first simulation was aimed at examining the need for a correction of the test statistic. Figure 2 shows the true variance of $\hat{\theta} - \hat{\theta}^*$ compared to both the asymptotically calculated variance as well as the corrected variance. One-thousand simulations were performed using full terms regression as the only learner, and 10-fold CV (larger folds of CV were performed and did not impact the results). Both the sample size and the number of parameters in the model were varied. Results suggest, that the ratio between the true variance and the asymptotic variance is fairly consistent at 2. The absolute variance decreases as sample size increases and increases with the number of parameters. The great decrease in the absolute difference of the expected and estimated variance with large sample, suggests that the asymptotics are working, but the relative difference, indicates that the correction is of value.

Once determining an appropriate correction for the test statistic, the next series of simulations aimed to understand the statistic itself. The first simulation explored the null situation where $Y \sim N(0,1)$ was independent of $X \sim N(0,1) \in \Re^p$. One-thousand simulations were run, using 10 predictors, under three different sample sizes (100, $2,000$, $10,000$). Linear regression was the only model used to estimate $f(X)$. As the sample size increases the test statistic becomes more normally distributed. Moreover the correction becomes less important. For the permutation test, a Z-quantile for the empirical p-value was calculated (see Figure 3).

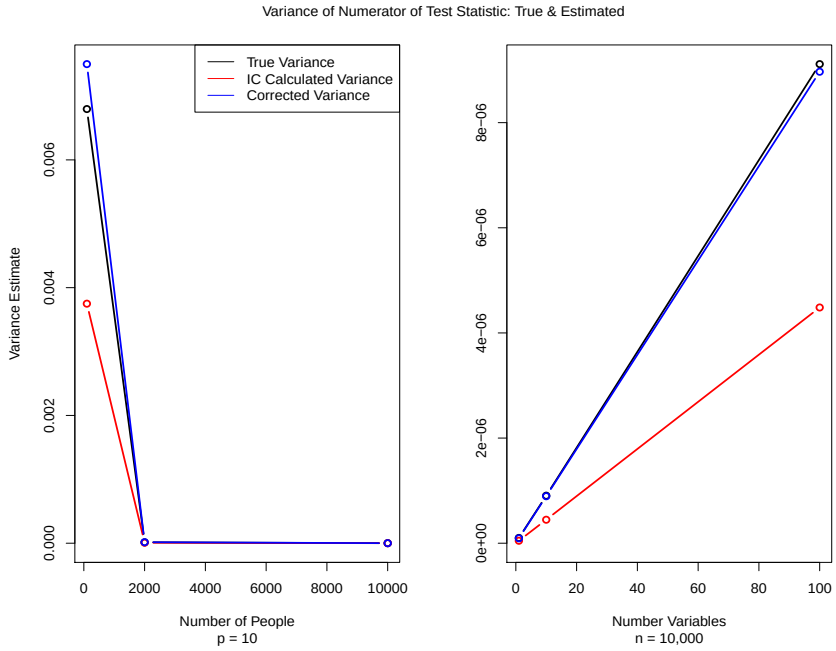

Figure 2: The true variance and estimated variance of $\hat{\theta}_n - \hat{\theta}_n^*$ across 1,000 simulations. In the left hand figure, the number of parameters is fixed at 10 and the number of people is varied from 100 to 10,000. In the right hand figure the number of people is fixed at 10,000 and the number of parameters is varied. Unsurprisingly the variance decreases with sample size and increases with the number of parameters. Of greater interest, the ratio between the asymptotically estimated variance and the true variance remains relatively constant at 2, though decreases in absolute terms with increasing sample size. Finally, the increase in variance due to the number of parameters appears to be linear in p.

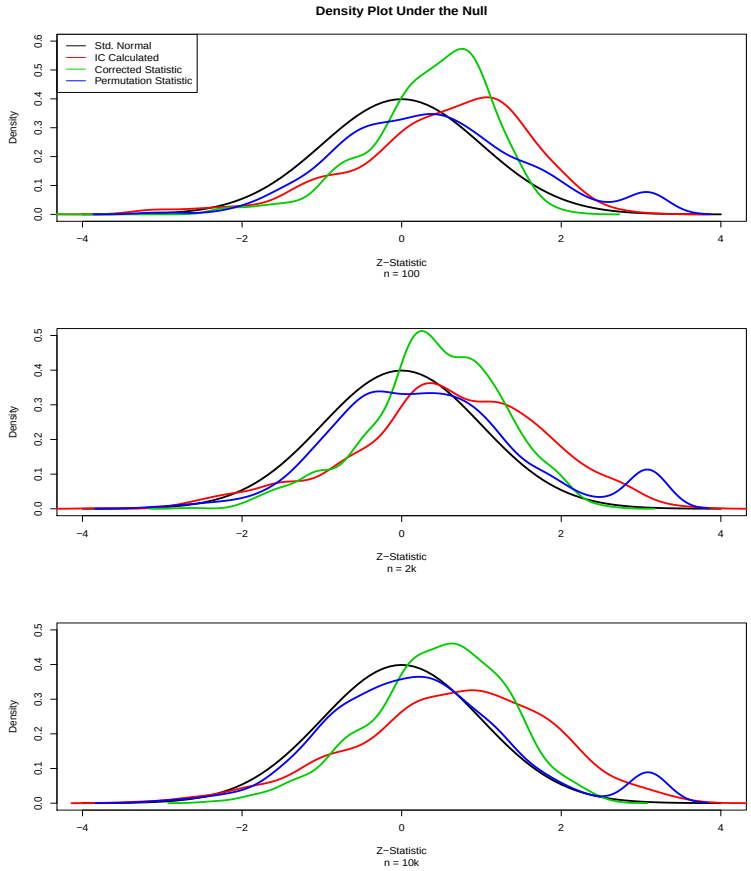

Figure 3: Behavior of test statistic with increasing sample size. One-thousand simulations performed with continuous X and $Y$and no association between them. 10 predictors are used. For larger n both the asymptotic and the corrected test statistics become more normal with variance 1. The uncorrected is slightly anti-conservative in the extreme (negative) tails highlighting the need for the correction at lower sample sizes. The permutation p-value have been transformed to a normal distribution for comparison and maintains appropriate error control.

Secondly, the test statistic was examined under an *alternative* scenario where there is a true relationship between Y & X. In this simulation, p = 10, and different sample sizes were used. $Y = X_1 + \epsilon$, with $Y \perp X_2 \dots X_{10}$. For simplicity, the corrected test statistic was calculated. Figure 4 shows the simulation results. While, the test has little power in low sample size (n = 100), the test gains power as the sample size increases, and maintains a $N(\mu, 1)$ distribution.

To illustrate the comparison to the F-test, two more simulations were undertaken, both under an alternative model. In the first, $Y = X_1 + \epsilon$ again. In the second, $Y = X_1 * X_2 + \epsilon$, an interaction between two of the X covariates but with no main effects. To calculate the F-statistic the full main effects model was fit. In scenario I, it is expected that the F-statistic should capture the Y, X_1 relationship. However, in simulation II, the model is now misspecified. To calculate the proposed test, a SuperLearner was fit, using a step wise algorithm and an intercept function.

The average p-values for each scenario are calculated and presented in figure 5. Both methods are able to detect the association when there is a main effect term in the model. However, once there is only an interaction, the F-test loses all of its power due to the misspecified model. The SuperLearner approach is able to flexibly search for the best model and consequently has ample power to detect the association. It is thus flexible against semi-parametric alternatives, giving this approach its power.

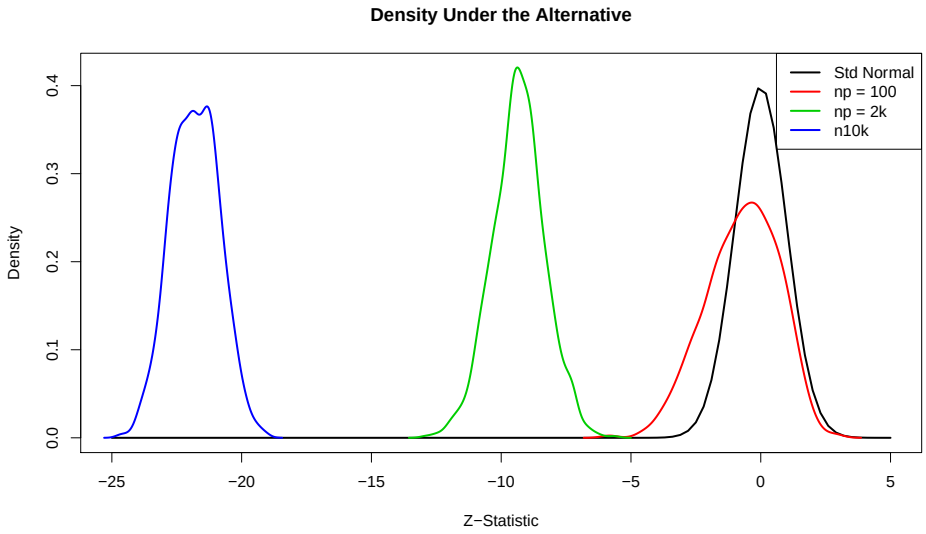

Figure 4: Distribution of the test-statistics when there is an association. Again linear regression is the only model used to estimate $f(X)$. The corrected test-statistic is calculated. As the sample size increases, the test statistic becomes both more significant (more negative) and approaches $N(\mu, 1)$.

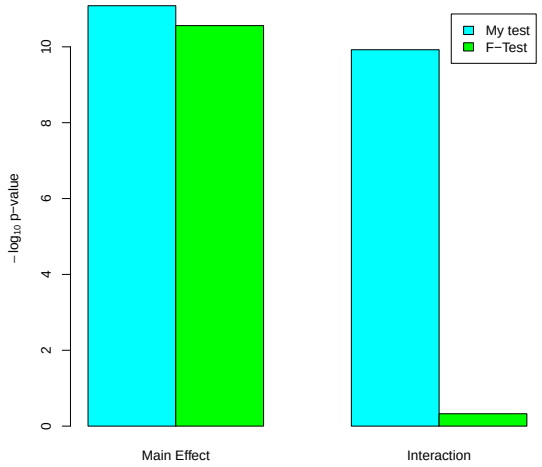

Figure 5: Average $-\log_{10}$ of the p-values for the two "alternative" simulation settings. In the first, there is a a true main effect term in the model and both the F-test and the proposed test are able to detect it. In the second, there is only an interaction, and the miss-specified regression model is not able to detect the association, while the more flexible SuperLearner based method is able to find the right model and detect the association.

6 Application to genetic data

A typical means of studying the genetic causes to diseases is via SNP association studies. This design involves recruiting 1000's of people with and without a disease of interest, referred to as cases and controls respectively. Each individual is typed on a set of single nucleotide polymorphisms (SNPs). Each SNP represents a single base pair of DNA where there is a degree of variation across the population (most DNA is fixed and does not vary). Genes are made up of 100s or 1000s of base pairs of DNA. Amongst these bases there will be dozens or 100s of SNPs. In Genome Wide Association (GWA) studies individuals may be typed on up to 1 million SNPs across the genome, aiming to capture ones common genetic variation (on the nucleotide level). In more focused candidate gene studies individuals may be typed on 10's of thousands of SNPs aimed at well characterizing specific genes. These studies represent an a hypothesis-free search across the genome for regions of interest to be followed up on.

For the most part GWAs have been successful at identifying SNPs, and be extension genes, associated with many common diseases [WTCCC 2007]. However, it is not presumed that any associated SNP is itself causal. An associated SNP may be correlated (referred to as in linkage disequilibrium [LD]) with the true causal variant, located within or near the same gene. Therefore all results need to be confirmed via replication. Moreover, since the studies are initially exploratory (until replication has occurred) the true unit of interest the gene in which the SNP lies. This has led towards the recognition of the need for gene based tests [Neale and Sham 2004].

In recent years there has been growth of gene based tests of association (see Beyene et al. [2009] for a recent review). These methods can roughly be broken down into (i) clustering and PCA based approaches and (ii) logistic modeling and combining marginal p-values. The method used in the popular and freely available software PLINK [Purcell et al. 2007], is a variation of Fisher's Method [Fisher 1948] that uses the permutation distribution to assess significance. However, most of these methods suffer from one primary limitation: they rely on marginal p-value (as calculated by a χ^2 test) to assess the gene based association. While marginal testing has been somewhat successful in detecting associations, and is (more importantly) computationally simple, it does not well capture complex associations. SNPs may interact or be involved in complex joint associations with other SNPs [Heidema et al. 2006]. Therefore, methods dependent on typical marginal tests may be ill suited for creating gene based tests.

One can consider the proposed to test as another means of performing a gene based test for association. In this setting, the observed units would be the SNPs and the unit of interest is the gene which they comprise. A prediction model relating the SNPs in a gene to disease status serves as a test for association for the entire gene. This point is illustrated first via simulation and then in application to a candidate gene study.

6.1 Simulation Study

A more comprehensive simulation study was undertaken to explore the use of the test statistic for candidate gene studies. Fifty simulations were performed. In each simulation 440 *genes* were simulated, consisting of 10 *SNPs*. Each of the 10 *SNPs* were independent and had a minor allele frequency of 0.3. In this sense a realistic genetic structure was not simulated, where one would expect complex correlation between SNPs and varying minor allele frequencies. However, to explore the performance of the method this was not necessary.

Of the 440 genes in each dataset, 40 (10%) of the genes were associated with the outcome. The goal of the simulation was specifically explore how the proposed test compares to standard methods when there is a complex association. Four different association models were used:

Additive: $P(D|SNPs_{Gene}) = \beta(X_1 + X_2 + X_3)$ with $X_i \in \{0, 1, 2\}$

Dominant: $P(D|SNPs_{Gene}) = \beta(X_1 + X_2 + X_3)$ with $X_i \in \{0, 1/2\}$

Recessive: $P(D|SNPs_{Gene}) = \beta(X_1 + X_2 + X_3)$ with $X_i \in \{0/1, 2\}$

Interaction: $P(D|SNPs_{Gene}) = \beta(X_1 * X_2)$ with $X_i \in \{0, 1, 2\}$

These represent genic style associations where multiple *SNPs* within a *gene* lead to an increase in the probability of disease.

For each dataset three measures of association were calculated. First, the marginal association for each of the $4,400$ *SNPs* was calculated via the allelic χ^2-test. This is the typical test for association in genetic epidemiology studies. It is a 1-df test that compares the frequency of the alleles between those with disease and those without. The second measure of association was the variation of Fisher's Method for combining p-values. To calculate the p-value, 10,000 permutation were performed. Finally, the current test was used to estimate the function

$$P(D|SNPs_{Gene}) = f(SNPs)$$

A SuperLearner was fit using a library of: k-Nearest Neighbors, a logistic regression step function, RandomForests, LASSO and an intercept function, using 10 fold CV to both fit and validate the function. The corrected parametric statistic was used to calculate the p-value.

For each method the false discovery rate (FDR) was controlled using the Benjamini-Hochberg (BH) [Benjamini and Hochberg 1995] procedure at a level of 5%. The average power and error-rate across the 50 simulations was calculated for each procedure. For the marginal testing, if one of the *SNPs* in the *gene* passed the significance threshold, then the entire gene was declared significant.

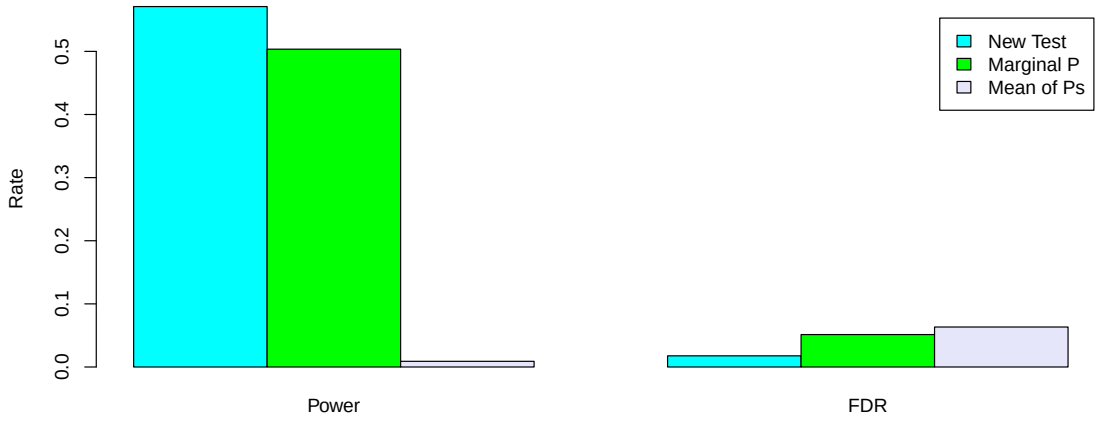

Figure 6: Bar plot comparing three methods for assessing the association of a *gene*. The proposed test has both the greatest power ($\sim 57\%$) lowest FDR ($\sim 1.8\%$) The FDR was controlled at a level of 5%. Marginal testing proved to be fairly successful, though the greater multiple testing burden incurred by the 10-fold increase in the number of tests makes it less powerful. Fisher's Method was not able to maintain a high significance level.

Figure 6 shows the power and FDR for the three methods. The proposed test has the most amount of power (57% vs. 50% & 1%) as well as the lowest error-rate. While marginal testing was fairly successful, the extra multiple testing burden incurred by the extra tests decreased its power. It should be noted, that in an actual association study, this burden would be even greater as most genes have many more than just 10 SNPs. Finally, Fisher's Method was least successful, this owes to the duel fact that it only looks at marginal associations and considers all tests simultaneously, while the proposed method, looks at the joint effects of only those tests of importance.

These results should not necessarily be interpreted that the proposed test is uniformly more powerful than these or other methods. The proposed statistic is an omnibus test, and only a few types of causal models were explored. For example, if only one *SNP* for a given *gene* were causal, then the marginal test would be most powerful. Likewise, if most of the *SNPs* for a given *gene* were associated than Fisher's Method would have more power. The value of the present statistic, though, is that it does not require one to specify a specific model, but search for the best fitting one, while still maintaining adequate power.

To illustrate this a second simulation was undertaken. As opposed to a complex as-

sociation, only one *SNP* was associated per *gene*. Moreover, the marginal p-value was simulated to be approximately 1×10^{-8}, the standard cut-off for genome wide significance. One thousand *genes* were simulated, and the median the p-value for the associated *SNP* was 8.5×10^{-9}. The median p-value on the gene-based test, was 1.6×10^{-4}. This shows that even when the causal mechanism favors the marginal test, there is still ample signal to find an association via this more flexible approach. For more complex associations, as well as for larger datasets with greater multiple testing burden, this difference will decrease.

6.2 Data Analysis

To examine how this methods works with real data, three analyses were performed using a data set derived from a candidate gene study. The first was a typical analysis to determine which genes were associated with disease. The second was aimed at exploring the association of a genetic pathway. The third is a more unique cluster based analysis.

Data for all analyses comes from a 2007 candidate gene study from the International Multiple Sclerosis Genetics Consortium. Multiple Sclerosis (MS) is an auto-immune disease known to have a strong heritable component based on epidemiological studies [Oksenberg and Barcellos 2005]. The major histocompatibility complex (MHC) region of chromosome 6 has long been known to be associated with MS, however few other genes have been definitively identified.

The goal of the 2007 was to follow-up on suspected MS genes. The data collection has been described previously [IMSGC 2010]. In brief, the data consisted of $1,379$ controls and $1,343$ cases. Data were collected on $52,801$ SNPs across 9552 genes. After data cleaning there were $46,057$ SNPs.

6.2.1 Candidate Gene Study

The first analysis was aimed at detecting which genes are associated with MS. All genes that had at least 8 SNPs in them and were not on chromosome X or within the MHC were selected. Chromosome X genes were dropped to avoid gender effects, while MHC genes were dropped since the goal was to detect genes that were not known to be associated with MS. This left $1,254$ genes comprising $25,362$ SNPs.

The three methods for association were calculated as in the simulation study: the generalized method, the mean of the p-values (Fisher's Method) and simple marginal testing. For the generalized method, a SuperLearner was fit to estimate the function using the same candidate learners as above, including also, Support Vector Machines and PolyClass. For the mean of p-values, $10,000$ permutations were used, performed in PLINK [Purcell et al. 2007]. For marginal testing, the minimum p-value for each gene was recorded.

After controlling for multiple testing using the BH-FDR, none of the measures of association provided an FDR below 10%, suggesting that the smallest observed p-values would

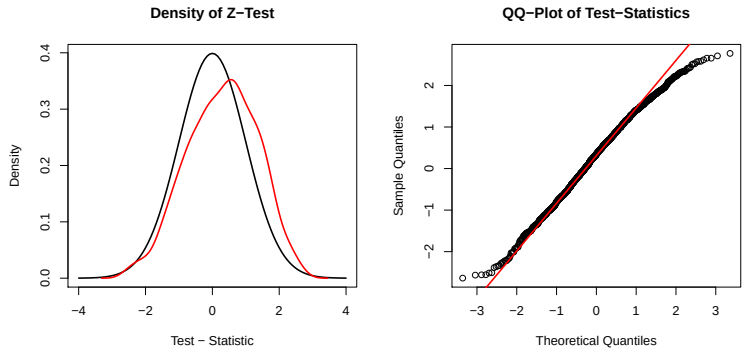

Figure 7: Distribution of test statistic for all genes. Left is a density plot while the right is a qq-plot. The plots suggest that there are no greater associations that would be expected by chance.

be expected simply by chance. Figure 7 shows the distribution of the test statistic across the 1254 genes, both as a density plot and qq-plot. This further reflects that even though small p-values were observed, these conform with what would be expected.

While it is not possible to reject the global null of no genes being associated, it is still of value to explore the most significant genes. Table 2 lists all the genes that had a p-value less than .01 on the proposed test. Also listed are their p-values for Fisher's Method as well as the smallest marginal p-value. From the table, it is clear that some genes had strong association regardless of the method used. These include STAT4, IL7 and CLEC16A. Unsurprisingly, IL7 and CLEC16A have previously been identified as MS genes [Hafler et al. 2007]. In this sense, the proposed method is able to replicate previous findings. However, the other genes, while showing some marginal association, would not be detected by typical methods. This suggests, that if truly associated, the mechanism is more likely more complex than a single SNP association. Also noteworthy, is that the size of the gene appears to be independent of the strength of association. Both small and large genes had strong associations.

Gene	Chr	Num SNPs	Gene P-val	Mean P-val	Min SNP P-val
STAT4	2	139	0.099	0.006	2.86×10^{-5}
EFNA5	5	10	0.099	0.386	0.056
MAP1B	5	13	0.005	0.379	0.055
IL7	8	73	0.009	4.0×10^{-4}	2.04×10^{-5}
RBM17	10	77	0.009	0.310	0.021
IRF7	11	8	0.006	0.044	3.49×10^{-4}
CLEC16A	16	14	0.005	1.0×10^{-4}	1.08×10^{-5}
MYO1D	17	9	0.004	0.165	0.075
APP	21	9	0.005	0.104	0.044

Table 2: All Genes that had a p-value less than .01 on the gene based test. The p-value based on Fisher's Method as well as the minimum marginal p-value is also shown. The genes highlighted in red probably would not have been detected by alternative methods. Other genes, such as IL7 & CLEC16A not surprisingly have previously been associated with MS.

6.2.2 Pathway Analysis

Using the same dataset, a second analysis was undertaken. Instead of looking at individual genes, whole genetic pathways were explored. Many genes have shared biological functions, referred to as genetic pathways. One such pathway of interest for MS is the DNA Repair pathway. The DNA repair pathway consists of four sub-pathways. Briggs et al. [2010] studied the pathways' relationship with MS, using a mixture of machine learning (Random Forests) and parametric modeling (Logistic Regression). The results suggested that the only important gene in the pathway is GTF2H4, which is located within MHC region of chromosome 6. However, the results were not definitive, and using the same data, the same pathways were reanalyzed using the current method.

In pathway analysis the observed unit is still individual SNPs, but the unit of interest is now a collection of genes. Methodologically, the approach is the same as analyzing a gene, except the number of SNPs are increased. A SuperLearner was fit using the same candidates as above. The four pathways were analyzed separately and the results are shown in Table 3. The only associated pathway is the NER pathway, which contains GTF2H4. Analyzing GTF2H4 independently, revealed an association very close to the full NER association. Finally, analyzing the NER pathway without GTF2H4 revealed no association. These results are a more definitive confirmation of the findings in Briggs et al. [2010] that GTF2H4 is the only important component of the DNA Repair pathways as it relates to MS.

Pathway	Num Genes (SNPs)	Z-Statistic	P-value
BER	22 (127)	0.12	0.55
HR	15 (124)	-0.29	0.38
NHEJ	9 (90)	-0.97	0.17
NER	26 (208)	-3.12	9.03×10^{-4}
NER (w/out GTF2H4)	197	-0.68	0.25
GTF2H4 only	11	-3.79	7.53×10^{-5}

Table 3: Results for the DNA Repair pathway analysis. Only the NER pathway shows any association with MS. However, upon further examination that association is based entirely on GTF2H4 confirming the results in Briggs et al. [2010].

6.2.3 Clustering the MHC

To illustrate the flexibility of this method to genetic data a very different analysis was performed. Instead of testing the association of a region, the predictions were used to cluster genes. As mentioned, the MHC region of chromosome 6 is well known to be highly associated with MS (as well as many other auto-immune diseases). While the gene based analysis revealed that almost every MHC gene had an association (often very strong ones), one question of interest is whether these associations are due to the strong and complex correlation (LD) in the MHC or are independent signals. While HLA-DRA is known to be associated with MS, recent studies have suggested that other genes may also be independently associated (e.g. Cree et al. [2010]).

In order to explore this question a novel approach was taken. The predictions ($\hat{f}(X)$) were calculated for each of the 78 MHC genes in the dataset. Then using the predictions of just those with disease ($n = 1343$) all of the genes were clustered using the partitioning around medoids (PAM) algorithm. PAM is similar to k-means clustering, with the primary difference being, instead of minimizing an average distance, a median distance is minimized, making the findings more robust. The MHC can be divided into three classes. Therefore the number of centers, k, was chosen to be 3. Since the goal was to capture the correlation among the predictions, a correlation based distance was used.

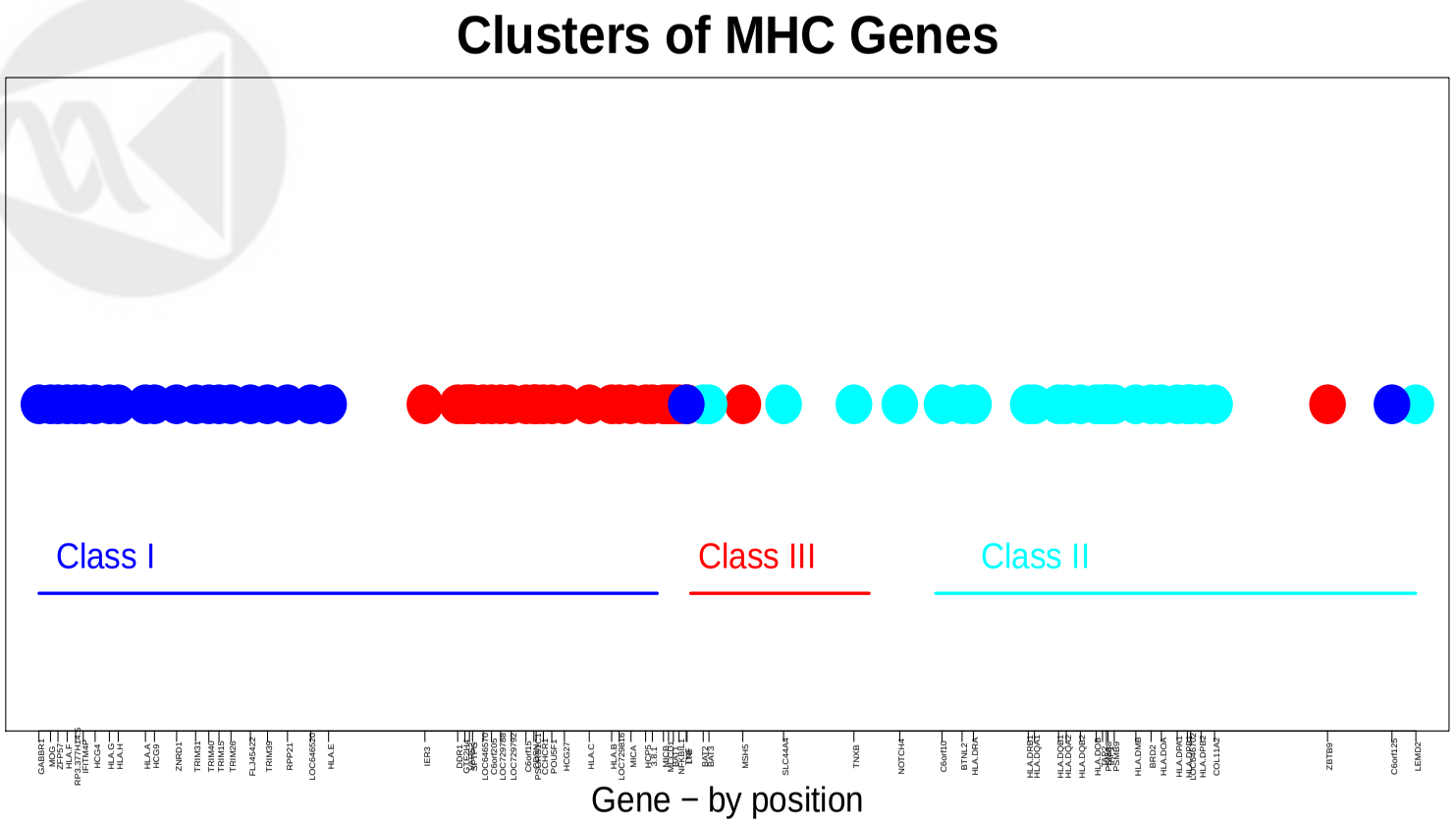

Figure 8: Clusters of the MHC genes based on the $\hat{f}(X)$ for cases. Genes are organized by position on Chromosome 6. The lines underneath indicate the boundaries for the three MHC classes. The three clusters of genes correspond very closely with their location within chromosome 6 and the cluster memberships correspond almost perfectly with the three MHC classes.

Figure 8 shows the clusters of the 78 MHC genes. The genes, are ordered by their position on Chromosome 6. The groupings of genes clearly correspond to their position, suggesting that the correlation between the $\hat{f}(X)$ is maintained through the position. Of greater interest, the cluster memberships correspond fairly well with the three MHC classes.

To explore the question of whether these clusters represented independent signals, a SuperLearner was trained using the SNPs in each of the three clusters, using the same procedures of above. The interest was not in whether each cluster is associated with MS (each clearly is), but whether the associations represent distinct signals. Table 4 shows the correlation matrix for the three clusters. While there is positive correlation between all three classes, Class I and Class II genes appear to be fairly independent of one another compared to Class III (which is physically located in between Class I & II). This suggests that there may be two separate associations within the MHC for MS. While this analysis does not represent a confirmation of independent signals, like previous analyses, it does correspond to previous analyses that they may exist.

	Class I	Class II	Class III
Class I	1.000	0.175	0.410
Class II	0.175	1.000	0.516
Class III	0.410	0.516	1.000

Table 4: Correlation matrix of the $\hat{f}(X)$ from the three MHC clusters, corresponding to the three MHC classes. Class I & II appear to be somewhat independent, suggesting that there may be two independent signals within the MHC for MS.

6.3 Thoughts on Application to Genetic Data

Three different applications to genetic data were illustrated, highlighting the flexibility of this approach. One particular challenge to this method is the ability to create a strong predictor using genetic data. The influence of ones genetics on disease (i.e. $P(\text{Disease}|\text{Genetics})$) is going to be relatively low. For example, while MS has a strong genetic component, its overall heritability is only estimated at 25% [Oksenberg and Barcellos 2005]. Therefore, if a gene is in fact causative of disease, one would not expect the $P(D|G)$ to differ much from average risk. Therefore the true risk (θ) will not be much less than the expected risk (θ^*) making detecting a significant association challenging.

Clayton [2009] looked at SNP data and noted that even highly associated SNPs have low predictive ability. However, Kooperberg et al. [2010] recently showed, that creating prediction models using SNPs that do not have strong marginal associations, does improve the models. This is essentially the approach undertaken here, using all SNPs within a region regardless of their marginal association. Even so, it is possible that the weak associations

detected in the candidate gene analysis, may not be indicative of a lack of effect, but simply a weak predictive ability of the SNPs in those genes.

7 Conclusion

The proposed method represents a powerful, flexible and semi-parametric approach to testing the relationship between a set of variables and an outcome. It has applicability both within the fields of prediction and machine learning as well as classical statistical inference. From a machine learning perspective, it represents a means of assessing whether the predictive ability of a set of predictors is better than what would be expected by chance. This represents an important but often unasked question of whether one should even attempt to construct a predictor from the given data. From an inferential perspective, it represents a means to assess whether a set of variables is related to an outcome. Traditional methodology is aimed at assessing the relationship between one covariate and one outcome. This allows one to look at multiple variables at once. This is particularly important in fields like genetics and the social sciences where one often has data on one level (e.g. SNPs, social variables) and wants to make inference on another level that aggregates the data (i.e. genes, SES). Such aggregation is often the only approach, as many times the level one wants to make inference on, represents a construct and not an actual observable variable.

The constructed test statistic is a Wald statistic, using influence curves to calculate the inference. A finite sample correction was necessary to appropriately scale the variance. While the current correction is somewhat adhoc it conforms with other current work in the area. An important area of further investigation is a formalized correction. For perspective, the test-statistic has been compared to the F-test used in linear regression. In this sense it can be thought of generalized goodness-of-fit test. Simulation results show the relationship to the F-test, but also how it is more powerful under a variety of alternatives.

A range of applications to genetic data were illustrated. The first involved testing the association of gene (and pathway) with disease. A more comprehensive simulation illustrated how this is more powerful than typical approaches under complex associations. While the candidate gene analysis was not able to reveal associations greater than what would be expected by chance, examination of results did reveal that genes that would not be identified by traditional approaches had strong associations. An analysis of genetic pathways was able to more strongly confirm results in a previous study that were merely speculated upon. A final analyses illustrated the variety of questions that could be addressed with this data. Using the same dataset, the goal was to find clusters of genes in the MHC. The clusters corresponded almost perfectly with biological understanding and the results provided insight that there may be two independent sources of association within the MHC for MS.

In all, this method represents an important contribution to a range of statistical applications, filling a need within both statistical learning and inferential statistics. It effectively

expands the range of questions that one can ask of their data and should represent an important tool for many analyses.

References

Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

J Beyene, D. Tritchler, J. L. Asimit, and J.S. Hamid. Gene- or region-based analysis of genome-wide association studies. *Genetic Epidemiology*, 33:s105–s110, 2009.

M. D. Birkner, A. E. Hubbard, and M. L. van der Laan. Data adaptive pathway testing. Technical Report 197, U.C. Berkeley Division of Biostatistics, November 2005.

L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

F.B. Briggs, B.A. Goldstein, J.L. McCauley, R.L. Zuvich, P.L. De Jager, J.D. Rioux, A.J. Ivinson, A. Compston, D.A. Hafler, S.L. Hauser, J.R. Oksenberg, S.J. Sawcer, M.A. Pericak-Vance, J.L. Haines, L.F. Barcellos, and International Multiple Sclerosis Genetics Consortium. Variation within dna repair pathway genes and risk of multiple sclerosis. *American Journal of Epidemiology*, 172:217–224, 2010.

P. Chaffee, A. E. Hubbard, and M. L. van der Laan. Permutation-based pathway testing using the super learner algorithm. Technical Report 263, U.C. Berkeley Division of Biostatistics, March 2010.

D.G. Clayton. Prediction and interaction in complex diseases. *Plos Genetics*, 5, 2009.

B. A. Cree, J. D. Rioux, J. L. McCauley, P. A. Gourraud, P. Goyette, J. McElroy, P. De Jager, A. Santaniello, T. J. Vyse, P. K. Gregersen, D. Mirel, D. A. Hafler, J. L. Haines, M. A. Pericak-Vance, A. Compston, S. J. Sawcer, J. R. Oksenberg, S. L. Hauser, IMAGEN, and IMSGC. A major histocompatibility Class I locus contributes to multiple sclerosis susceptibility independently from HLA-DRB1*15:01. *PLoS ONE*, 5:e11296, 2010.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2:131–154, 2005.

R.A. Fisher. Questions and answers no.14. *The American Statistician*, 2:30–31, 1948.

Y. Grandvalet and Y. Bengio. Hypothesis testing for cross-validation. Technical Report 1285, Departement dInformatique et Recherche Operationnelle, August 2006.

D. A. Hafler, A. Compston, S. Sawcer, E. S. Lander, M. J. Daly, P. L. De Jager, P. I. de Bakker, S. B. Gabriel, D. B. Mirel, A. J. Ivinson, M. A. Pericak-Vance, S. G. Gregory, J. D. Rioux, J. L. McCauley, J. L. Haines, L. F. Barcellos, B. Cree, J. R. Oksenberg, and S. L. Hauser. Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.*, 357:851–862, Aug 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, New York, 2 edition, 2009.

A. G. Heidema, J. M. Boer, N. Nagelkerke, E. C. Mariman, D. L. van der A, and E. J. Feskens. The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics*, 7:23, 2006.

IMSGC. Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies kif21b and tmem39a as susceptibility loci. *Human Molecular Genetics*, 19:953–962, 2010.

C. Kooperberg, M. Leblanc, and V. Obenchain. Risk prediction using genome-wide association studies. *Genetic Epidemiology*, pages 643–652, 2010.

S. Lee. Mistakes in validating the accuracy of a prediction classifier in high-dimensional but small-sample microarray data. *Statistical Methods in Medical Research*, 26:1102–1113, 2007.

L. Lusa, L.M. McShane, M.D. Radmacher, J.H. Shih, G.W. Wright, and R. Simon. Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. *Statistics in Medicine*, 26:1102–1113, 2007.

B.M. Neale and P.C. Sham. The future of association studies: gene-based analysis and replication. *American Journal of Human Genetics*, 75:353–362, 2004.

J. R. Oksenberg and L. F. Barcellos. Multiple sclerosis genetics: leaving no stone unturned. *Genes and Immunity*, 6:375–387, Aug 2005.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81: 559–575, Sep 2007.

M.D. Radmacher, L.M. McShane, and Simon R. A paradigm for class prediction using gene expression profiles. *Journal of Computational Biology*, 9:505–511, 2002.

K.M. Ting and I.H. Witten. Stacked generalization: when does it work? In *Procs. International Joint Conference on Artificial Intelligence*, pages 866–871, 1997.

M. J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions*, 3:373–395, 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Superlearner. *Statistical Applications in Genetics & Molecular Biology*, 6, 2007.

M.J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2, 2006.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.