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

Year 2008 *Paper* 233

Targeted Methods for Biomarker Discovery, the Search for a Standard

Catherine Tuglus* Mark J. van der Laan†

*Division of Biostatistics, University of California, Berkeley, ctuglus@berkeley.edu

†Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper233

Targeted Methods for Biomarker Discovery, the Search for a Standard

Catherine Tuglus and Mark J. van der Laan

Abstract

More often than not biomarker studies analyze large quantities of variables with complicated and generally unknown correlation structure. There are numerous statistical methods which attempt to unravel these variables and determine the underlying mechanism through identification of causally related biomarkers. Results from these methods are generally difficult to interpret and nearly impossible to compare across studies. The FDA has currently called for a standardization of methods and protocol for biomarker detection. In response, we propose targeted variable importance (tVIM) as a standardized method for biomarker discovery. Through the use of targeted Maximum Likelihood, tVIM provides double robust estimates of variable importance along with formal inference. These measures are biologically interpretable as a causal effect under specified conditions, allowing for reproducibility across populations. In this analysis we compare tVIM to four different measures of importance provided by three different statistical methods: univariate linear regression (LM), LASSO penalized multiple regression (Q), and two importance measures from randomForest (RF1 and RF2). Their performance is compared in simulation under conditions of increasing correlation. We are interested in their ability to distinguish "true" relevant biomarkers from correlated decoy biomarkers. The comparisons are based on the resulting ranked variable list for each method using the importance measures and p-values when available. In simulation, tVIM coupled with a data-adaptive model selection method outperforms linear regression, LASSO, and randomForest and is more resilient to increases in correlation. In application we apply all methods to the Golub et al 1999 Leukemia data and compare the resulting gene lists based on biological relevance. Both LM and tVIM are also applied to the van't Veer breast cancer data. We compare them based on the top 10 most important genes. From these results, tVIM appears to rank more biologically relevant genes at the top its list than the

other methods. Given extreme correlations, methods to reduce bias and provide realistic gene lists are also discussed.

1 Introduction

The increasing importance in biomarker detection has prompted the FDA to call for the development of a standard protocol for biomarker qualification. In a recently published FDA document (U.S, 2007) outlining "highly-targeted" research projects for the "Critical Path Initiative," number one of their six main areas is devoted to the development of biomarker analysis tools and disease models. The FDA outlines the need to develop a national standard for biomarker detection, quality, and usage attempting to create a level of quality assurance with respect to biological and more specifically biomarker research. Such a standard will provide a reliable protocol for the qualification of a biomarker's statistical significance and biological interpretation. It will identify specific "evidentiary standards" for biomarker usage in various areas of development and treatment. For instance in clinical trails, these standards would prohibit the use of unauthenticated biomarkers to determine treatment regime, providing more effective and reliable treatment decisions. Additionally the FDA hopes the creation of standard methods will allow for comparisons of results among studies (U.S, 2006).

The word biomarker embodies a wide variety of biological characteristics, including and not limited to gene expression, SNP pattern, copy number, and protein abundance. Due to its nature biomarker data is often high dimensional and highly correlated. In this paper, we will concentrate on biomarker discovery using gene expression measurements from microarray analysis. Gene expression data derived from microarray technologies are not only high dimensional and highly correlated, but they are also ladened with noise.

There are numerous statistical methods used for biomarker detection that aim to quantify the relationship between gene expression and a given biological outcome. Common statistical methods such as univariate linear regression focus on detecting a "significant" association between gene and outcome. Highly sensitive to increased correlation among the variables, univariate linear regression often leads to high false positive rate requiring the researcher to waste valuable resources on analyzing irrelevant genes. Multivariate regression is often used to control for confounding effects among highly correlated variables, however the need for model selection often limits its applicability for biomarker discovery. Prediction algorithms such as LASSO regression (Tibshirani, 1996) and randomForest (Breiman, 2001) are also commonly used for biomarker discovery, but will tend to produce bias estimates of importance when correlation among variables is high. These methods often lack formal inference and do not automatically provide p-values for their measures of importance.

Most biomarker discovery methods focus on measuring the association between gene expression and the biological outcome. However a significant association is often difficult to interpret and does not guarantee that the biomarker will be a suitable and reliable drug candidate or diagnostic surrogate. Ideally biomarker discovery analyses aim to identify genes that systematically effect your outcome through a biological pathway or mechanism, in other words genes causally related to the outcome of interest. Once these genes are identified, they can be further analyzed and eventually applied as potential drug targets or prognostic markers. Due to the complex nature of the human genome, this is not a straight forward task. Genes are often present in multiple pathways and can be highly correlated amongst themselves. For instance, a common approach is to estimate $E[Y|A] = \beta A$ using univariate linear regression and interpret the parameter β as a measure of importance for A. However, by ignoring potential confounders (W) (i.e. other genes in the pathway or system), this approach estimates a measure only interpretable as a causal effect in the case where there are no confounding factors (all W is kept constant at all values of A), and will identify genes which are merely correlated with the true causally related variables, resulting in a long ambiguous list of biomarkers.

In this paper we propose targeted variable importance measures (tVIM), as a new standard method for biomarker discovery. Using targeted maximum likelihood (tMLE) methodology as presented in van der Laan and Rubin (October 2006), tVIM targets the direct effect of a the variable of interest A, on outcome Y, adjusting for confounders W. More formally, given an observed data set defined as $O = (W*, Y) \sim P_0$, with covariate matrix W and outcome Y, consider the variable of interest A where $\{A, W) = W^*$. The tVIM of A at value a is defined as follows.

$$\Psi(P) = \mathbb{E}[\mathbb{E}_{\mathbb{W}}[Y|A = a, W] - \mathbb{E}[Y|A = 0, W]]$$

In biomarker analyses, A can represent a single biomarker or set of biomarkers, and W any biomarkers or variables controlled for in the overall density $Y|W^*$. In this paper we apply a semi-parametric regression version of tVIM. Models of this type have been presented in Robins et al. (1992); Robins and Rotnitzky (2001); Yu and van der Laan (September 2003).

One can estimate tVIM by updating an initial regression estimate $\mathbb{E}[Y|A, W]$ in a direction with targets the parameter of interest $\Psi(P)$ using tMLE methodology. This update is calculated using an estimate of $\mathbb{E}[A|W]$. This is often referred to as the "treatment mechanism," accounting for the effects of confounders W on "treatment" (or variable) A. Given correct

model specification for either $\mathbb{E}[Y|A,W]$ or $\mathbb{E}[A|W]$, the tVIM estimate is a consistent and asymptotically normal and linear estimate. The estimate is efficient when both models are correctly specified (a.k.a. "locally efficient"). This feature is referred to as "doubly robust." Improving the estimates of $\mathbb{E}[Y|A,W]$ and $\mathbb{E}[A|W]$ using data-adaptive or super learning algorithms will improve the overall consistency and efficiency of the estimate. The method of estimation is presented more throughly in section 4.3.

Targeted Variable Importance measures are biologically interpretable. In the case of a randomized trail, tVIM measures the causal effect of A on the outcome Y, where W would contain all variables which confound the effect of A on Y. In the case of observational data, one can still interpret $\Psi(P)$ as the causal effect one would have observed under experimental conditions which only control for the given set variables, W. By targeting the causal effect, the tVIM estimate can be obtained for any new population if the same condition distribution $Y|W$ holds and the new marginal distribution of W is known. If the casual variables are accounted for correctly, $Y|W$ will be consistent across populations because the causal mechanism is consistent across populations.

When the data is highly correlated, tVIM can be used in conjunction with a correlation cut-off to decrease bias in the estimate, and in practice provide a realistic ranking of variables. Given a correlation cut-off, tVIM will identify all causally related variables as well as all variables the data is unable to disentangle due to the high correlation structure. Determining the correct cut-off decreases bias in the estimate while still maintaining the optimal level of reproducibility. Targeted Variable Importance measures also have formal inference and multiple testing methods based on the influence curve which allows estimation of the overall joint distribution without resampling.

In this paper we compare tVIM to three other methods commonly used for determining variable importance in biomarker discovery analyses: univariate linear regression, LASSO regression with cross-validation based model-selection (Efron et al., 2004) - using R package *lars* (Efron and Hastie), and randomForest (Breiman, 2001) - using R package *randomForest* (Liaw and Wiener). Importance measures for univariate linear regression and LASSO regression are represented by the associated coefficient value. RandomForest provides two measure of importance based on the effect perturbing the variable of interest has on overall classification error and node splits. Table 9 in appendix A summarizes the merits of four methods in terms of variable importance identification.

The goal of biomarker discovery methods is to provide an accurate, generally ranked, list of biologically relevant genes. This list is often used to direct further biological or statistical analyses. Inaccuracies in this ranking would result in time consuming and expensive analyses on potentially unimportant biomarkers. Consequently we compare the four methods based on their ability to accurately produce this list. Often in practice, biomarkers are ranked in the order of increasing p-value, where markers with p-value below a particular cut-off are defined as "important." Alternatively if no p-values are provided, the biomarkers may be ranked by their importance measure, and the cut-off would be based on a required level of importance, a particular number limit for instance.

Simulated data is used to compare all four approaches under increasing correlation levels using a diagonal block correlation structure. The structure of the simulated data allows us to study the effects both correlated and uncorrelated variables have on the reported importance of the true variables using the four approaches. For each approach, the biomarkers will be ranked by the resulting importance measure and p-value (when available). The sensitivity and specificity of methods will be compared based on both p-value and rank-based cut-off values. Results will be summarized using ROC plots and in terms of power and false discovery rate. We will determine the ability of each approach to identify the true causally related variables and each variables true importance rank by comparing the length of list required to label each individual true variable and all true variables as "important." The ability of each method to determine the true level of importance will also be compared when appropriate.

We also apply and compare univariate regression, tVIM, and randomForest in application using the AML/ALL leukemia data provided in Golub et al. (1999). The performance of the methods are compared based on the biological relevance of the resulting gene lists. Targeted Variable Importance methods are applied Breast Cancer gene set provided by van 't Veer et al. (2002) as well. Results are discussed in terms of their biological relevance.

The paper is organized into 6 sections. After an introduction in the first section, the second section presents a general definition of variable importance. The third section gives background on the variable importance methods, specifically outlining the targeted estimation of tVIM. The fourth section presents a simulation study comparing variable importance measures. The fifth sections present applications of variable importance applied to the Golub et al. leukemia gene set and the vanVeer't et al. breast cancer gene set. The sixth provides discussion and future considerations.

2 Marginal Variable Importance

There are a variety of definitions of importance and multiple methods to measure each of them, making analyses difficult to compare and interpret. In this paper we prose a standard measure of importance which we refer to as the targeted variable importance (tVIM) . Analogous to the variable importance measure (VIM) measure in Bembom et al. (March 2008) which was presented for a binary A, we present the following measure of variable importance for a general A based on a semi-parametric model. This variable importance measure was first presented in van der Laan (2005). Models of this type have also been considered in Robins et al. (1992); Robins and Rotnitzky (2001); Yu and van der Laan (September 2003).

Given observed data $O \sim (W,Y)$, where W is the set of variables (i.e. genes), and Y is the outcome of interest, the marginal variable importance of a particular $A = W_j$ on outcome Y controlling for confounders $W^* = W_{-j}$ is defined generally as

$$\mu(a) = \mathbb{E}_{W^*}[m(a,W^*|\beta)]$$

for a user supplied model m, which models the effect

$$m(A = a,W^*|\beta) = \mathbb{E}_P[Y|A = a,W^*] - \mathbb{E}_P[Y|A = 0,W^*])$$

under the constraint $m(A = 0,W^*|\beta) = 0$ for all β and W^*. We define m(.) as the projection of the true effect onto a working model. The optimal estimator is developed under the assumption that m(.) is correct van der Laan (2005). Under randomization assumptions, this importance measure can be interpreted as a causal effect. If the working model (i.e. a semiparametric regression model) is misspecified, then it can be interpreted as a projection of the causal effect on a working model.

This method can be viewed as a semi-parametric variable importance measure. Given $\mathbb{E}[Y|A,W^*] = m(A,W^*|\beta) + g(W^*)$, where $g(W^*)$ is an unspecified function of W^*, we see that

$$\mathbb{E}_P[Y|A = a,W^*] - \mathbb{E}_P[Y|A = 0,W^*] = m(A = a,W^*|\beta) + g(W^*) - m(A = 0,W^*|\beta) - g(W^*) = m(A = a,W^*|\beta)$$

Given an estimator β_n of β, an estimate of this parameter of interest at a particular A=a is defined as
$\mu_n(a) = \frac{1}{n}\sum_{i=1}^{n}[m(a,W_i^*|\beta_n)]$
If we assume the model $m(A,W|\beta)$ is linear in A (i.e. $(m(A,W^*|\beta) = AW^*\beta)$), where A is continuous, the importance can be represented as the linear curve in terms of A, $\mathbb{E}_{W^*}[m(A = a,W^*|\beta)] = a\beta_{W^*}\mathbb{E}[W^*]$. Given a linear representation of m(.), the tVIM becomes a simple linear combination c^TA and formal inference can be estimate by applying the delta method. Further detail is provided in section 4.3 and Appendix B.

Compared to the tVIM method presented in Bembom et al. (March 2008), this model based approach to variable importance allows A to be continuous and can incorporate effect modification (i.e. $m(A,W^*|\beta = \beta_0 A + \beta_1 AW_1)$. This is especially relevant in clinical trials where the research is interested in finding genes (i.e. W_1) which modify the causal effect of a given a particular treatment (A) on overall disease response. Another benefit of tVIM for general A is the exclusion of inverse weighting making this measure more robust to experimental treatment assumption violations. Experimental treatment assumption violations are discussed further in section 4.3.3.

In this paper we focus on the simplest linear case $m(A,W^*|\beta) = A\beta_0$, where the marginal importance of A can be represented by single coefficient value β_0. This allows us to directly compare with alternative measures of importance obtained from univariate and multivariate regression methods.

Two alternative measures of variable importance often used in biomarker discovery analyses are derived from randomForest methodology (Breiman, 2001), which is a tree-based regression algorithm that exploits boosting to reduce the variance of a bias predictor fit. Though these measures lack causal interpretation they are often used in practice when the data is high dimensional and over-fitting is a risk. For the first measure, the relative importance of a variable is determined by the effect perturbing the variable values has on an "out-of-bag" error rate (Breiman, 2001). For the second measure, the affect of perturbation of the variable on node specific classification sensitivity is measured (Breiman, 2001). A more in-depth discussion of these parameters is presented in section 4.4. Although these measures are on a different scale than tVIM and have no formal inference, their ranked importance can be compared directly.

3 Background - Methods of Variable Importance

For purposes of discussing the methodology, variable importance methods will be presented in terms of a particular individual biomarker, $A = W_j$, with possible covariate set $W^* = W_{-j}$, and outcome Y. In practice the methods will be extended to all $A = W_j$ in W ($A = \{W_j : \forall j = 1...J\}$).

The five measures we consider in this study are listed below and then summarized in further detail.

1. LM: Marginal variable importance represented by the coefficient and p-value resulting from the univariate linear regression fit, $\mathbb{E}[Y|A]$
2. Q: Marginal Variable Importance represented by the coefficient of A in LASSO main term fit of $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$, where $W_s \subset W^*$ representing the subset of W^* found significant according to their univariate regression on Y. No p-values are provided (Efron and Hastie).
3. tVIM: Marginal Variable Importance measure obtained from applying targeted MLE to the initial density estimate provided by LASSO fit $Q(A, W_s)$. Coefficient of A is targeted directly. P-values are provided. The measure will be represented and compared in terms of the coefficient β_0 as presented in section 2
4. RF1: RandomForest importance measure based on "out-of-bag" error rate (Breiman, 2001; Liaw and Wiener) (no p-values provided)
5. RF2: RandomForest importance measure based on accuracy of node split (Breiman, 2001; Liaw and Wiener) (no p-values provided)

In practice the first three measures (LM, LASSO, tVIM) are computed sequentially, with each prior method informing the later. The last two methods are internally calculated in randomForest (Liaw and Wiener).

LASSO is applied using R package *lars* (Efron and Hastie), which does not provide formal inference. Given any estimate, bootstrap sampling may be used to provide standard error estimates and p-values. This is not done in this analysis. We choose to compare the methods based on their current merits and not on any additional processing. Also, in biomarker discovery there are thousands of genes and bootstrap sampling is computationally expensive. It may not be reasonable.

3.1 Univariate Linear Least Squares Regression

The standard univariate regression analysis approach estimates the coefficient β in the standard linear model $\mathbb{E}[Y|A] = \beta A$ using least squares regression minimizing

$$\mathcal{L}(O) = \sum_{i=1}^{n} (Y - \sum_{i=1}^{n} \beta A_i)^2$$

Given the observed data, we complete univariate linear regression analysis, estimating β in $\mathbb{E}[Y|A] = \beta A$ for all $A = \{W_j : \forall j = 1...J\}$. We refer to the associated coefficient values, β, as LM importance measures. The associated p-values, calculated using a standard t-test, are subjected to the Benjamini & Hochberg step-up FDR controlling procedure (Benjamini and Hochberg, 1995) to control for multiple testing. For many studies, the resulting FDR adjusted p-values (p_{LM_j}) are ranked and the set of genes with p-values less than 0.05 are considered significant and subjected to further analysis. This method does not account for any confounding and will often misclassify genes correlated with the "true" genes as significant. In most situations this importance measure can not be interpreted in as a causal effect.

3.2 Penalized Linear Least Squares Regression - LASSO

Often researchers refrain from controlling for confounding due to the high dimensionality of genetic data. The covariate set is too large for standard multivariate regression, and standard forward model selection methods often over-fit the data or are not a function of the variable of interest. In order to combat over-fitting but still allow a larger number of variables into the model, some variation of penalized regression is often used. Here we use LASSO regression (Tibshirani, 1996) which seeks to minimize the following,

$$\hat{\beta} = \arg\min_{\beta} = \sum_{i=1}^{n} (y - \beta_0 - \sum \beta_j X_{ij})^2$$

subject to the constraint that $\sum_j |\beta| <= s$ where $s > 0$. The threshold s is chosen using cross-validation using the R package lars(. . . ,method="lasso"). LASSO shrinks the values of the coefficients $\{\beta_j : \forall j = 1...J\}$ given the constraint, resulting in a less bias fit which is a function of all variables. There is no formal inference provided for this method. For computational ease, we decrease the dimensionality of $W^* = W_{-j}$, by allowing only univariate significant W^* (i.e. marginal LM adjusted p-value less than $\alpha < 0.05$). For a given A, we define the truncated covariate set as W_s. LASSO penalized regression is applied to covariate set $\{A, W_s\}$ resulting in an initial estimate of $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$. The resulting coefficients for each A are recorded as the LASSO (Q) importance measure. LASSO is implemented using *lars* R package and function (Efron and Hastie; Efron et al., 2004). This function does not provide any formal inference therefore p-values are not recorded, so we compare results based on the variable importance measure and its rank. Although this method does attempt to account for confounding to capture the causal effect, it still is a maximum likelihood method which focuses on estimating the overall distribution $\mathbb{E}[Y|A, W]$ and not the parameter of interest $\mathbb{E}[Y|A, W] - \mathbb{E}[Y|A = 0, W]$ resulting in a bias importance estimate. It also will only allow for n-1 non-zero coefficient values, making its applicability to high dimensional data limited.

3.3 Targeted Variable Importance measure - Using targeted MLE methodology

As previously stated, we define the targeted VIM measure as

$$\mu(a) = \mathbb{E}_{W^*}[m(a, W^*|\beta)] = \mathbb{E}_P[Y|A = a, W^*] - \mathbb{E}_P[Y|A = 0, W^*]$$

for a particular gene A=a, and covariate set W^*.

In this particular analysis we model the importance as $m(A, W^*|\beta) = A\beta(j)$, so that the marginal importance of A is represented by single coefficient value $\beta_n(j)$. We estimate $\beta_n(j)$ using targeted Maximum LIkelihood Estimation given a working model $m(A, W_s|\beta)$, and initial estimates for $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$ and $G(W_s) = \mathbb{E}(A|W_s)$, where W_s reflects a reduced set of W^* (van der Laan and Rubin, October 2006).

3.3.1 Targeted MLE

Targeted maximum likelihood estimation uses the general MLE framework and combines it with robust estimation using the efficient influence curve to provide a double robust and locally efficient estimator for the parameter of interest. Maximum Likelihood driven methods such as linear regression and LASSO focus on estimating the entire regression, $\mathbb{E}[Y|A, W^*]$. Estimating $\mu(a)$ using "plus and chug" MLE results in unnecessary bias in your predictor of interest because the estimate of $\mu(a)$ is based on the bias- variance trade-off for estimating $\mathbb{E}[Y|A, W]$. Targeted Maximum Likelihood (tMLE) methodology reduces the bias for the targeted parameter by maximizes the likelihood in a direction which corresponds to the best estimate of the targeted parameter (van der Laan and Rubin, October 2006), resulting in the doubly robust locally efficient estimate.

In biomarker discovery analyses tVIM is applied to all variables within the data matrix W, where W is a matrix of genes, SNPs, or other biological variables of interest. The method is outlined below for a single $A = W_j$, defining the possible covariate set as $W^* = W_{-j}$ and in practice is applied over all variables in W.

There are three initial components necessary for applying targeted Maximum Likelihood methodology to estimate tVIM.

1. Model $m(A, W^*|\beta)$ satisfying $m(0, W^* \mid \beta) = 0$ for all β and W^*. In this case it is defined as $m(A, W^*|\beta) = \beta A$
2. An initial regression estimate for $Q(A, W^*) = \mathbb{E}[Y|A, W^*]$ of the form $\mathbb{E}[Y|A, W^*] = m(A, W^*|\beta) + g(W^*)$, where $g(W^*)$ is estimated data-adaptively. We recommend using polymars (O'Connor; Kooperberg et al., 1997), lars (Efron and Hastie; Efron et al., 2004), or DSA (Sinisi and van der Laan, March 2004)
3. An estimate of the "treatment mechanism" $G(W^*) = \mathbb{E}[A|W^*]$, estimated data-adaptively.

Given these three components, tMLE can easily be applied in the following steps

1. Estimate the "clever covariate" which will allow us to update the initial regression in a direction which targets the parameter of interest. In this case the clever covariate is:

$$r(A,W^*) = \frac{d}{dB}m(A,W^*|\beta) - \mathbb{E}[\frac{d}{dB}m(A,W^*|\beta)|W]$$

which for this particular $m(A,W^*|\beta) = \beta A$ it simplifies to $r(A,W^*) = A - \mathbb{E}[A|W]$

2. Compute the fitted values for your initial estimate of $Q_n^0(A,W^*)$

3. Project Y onto $r(A,W)$ with $offset = Q_n^0(A,W^*)$, define the resulting coefficient as ε. This is done using standard software (lm() in R) setting the offset, and projecting onto the model $Y \sim r(A,W) + offset$. Note there is no intercept in your model, only the offset value.

4. update initial estimate $\beta_n^0 = \beta_n^0 + \varepsilon$ and overall density $Q_n^1(A,W^*) = Q_n^0(A,W^*) + \varepsilon r(A,W)$. These are now your single-step targeted estimates. Since this is a simple linear model, the single step solution is the final solution

5. Obtain standard error and inference for β can be obtained using the influence curve defined below, which corresponds to the double robust estimating function. This is possible because the tMLE solution also corresponds to the solution of the double robust estimating function (van der Laan, 2005). Given scale factor $c = \mathbb{E}[\frac{d}{d\beta}D(O|\beta_0,Q_0^1)]$, the empirical influence curve for a given $A = W_j$ in W.

$$IC(O)(j) = c^{-1}D(O|\beta_0,Q_0^1)$$

where,

$$D_h(p_0)(O) \equiv r(A,W)(Y - m(A,W|\beta_0) - Q_0(0,W))$$

The covariance of β_0 is asymptotically equivalent to the covariance of $IC(O)$. The empirical estimate of the covariance of β_n for a given $A = W_j$ in W is

$$\Sigma_n = \frac{1}{n}\sum \hat{IC(O)}\hat{IC(O)}^T$$

such that

$$\sqrt{n}(\beta_n - \beta_0) \sim N(0,\Sigma_n)$$

Covariance can also be estimated by bootstrap estimates of β, but this would requiring extra computational time. If $\mathbb{E}[A \mid W^*]$ is estimated consistently, then the variance estimates based on the influence curve are consistent or asymptotically conservative.

6. Using the estimated covariance, test the hypothesis $H_0 : \beta_n(j) = 0$, using a standard test statistic to obtain p-values.

$$T_n(j) = \frac{\sqrt{n}\beta_n}{\sqrt{\Sigma_n(j,j)}} \underset{n\to\infty}{\sim} Normal(0,1)$$

In Appendix B we outline the derivation of the targeted MLE methodology in further detail.

3.3.2 Inference and Testing

One of the great benefits of the tVIM is that formal inference is available. The covariance (Σ) for the set of tVIM measures β can be formally estimated using the empirical estimate of the conservative influence curve (see Appendix B and van der Laan and Robins (2003) for supporting theory and formal proof), or using a standard bootstrap resampling based covariance. Using the estimate $\hat{\Sigma}$, the significance of tVIM β can be determined using a standard t-test for an individual beta or a Wald test for a multivariate β. Also inference for linear combinations can be obtained by applying the delta method (see Appendix B). This provides formal p-values for the tVIMs which can be ranked and compared to their LM counterpart.

3.3.3 Experimental Treatment Assumptions and Consequences of its Violation

The Experimental Treatment Assumption (ETA) states that the probably of A given W^* must always be positive for all possible sets (a, W^*), $(P(A|W^*) > 0 \ \forall(a, W^*))$ (Wang et al., September 2006). In other words all values of A must be possible given any observed set of values W^*, and no W^* can be a perfect predictor of A. Given a nonparametric model, either case results in an unidentifiable variable importance measure. Such an estimate must be extrapolated leading to bias in the parameter estimate.

In reality it is difficult to assume a correct semi-parametric model. In the more realistic case, where we view our importance parameter as a projection onto a working semi-parametric model, violations of ETA can lead to instability in our parameter estimate. For instance given a situation where A is highly correlated with W, our initial regression estimate $Q(A,W)$ becomes highly sensitive. This instability in $Q(A,W)$ is reflected in the projection parameter. Therefore in practice it is important to avoid ETA violations in order to obtain unbiased importance estimates.

High correlation within a set of genes, W, poses a potential violation of ETA. However, if the "problem" variables (the variables highly correlated with the gene of interest A), are removed from the set of confounders (W^*), ETA violations can be avoided. Recently a method was proposed in Bembom et al. (March 2008), which defines an analytical formula for identifying these "problem" variables . In this paper, for simplicity we define a correlation cut-off, where all W whose correlation with A is greater than a particular correlation (ρ_{ETA}), are removed from the set of confounders for variable A prior to the application of tVIM method. However we recommend that future analyses use a data-adaptively selected correlation cut-off to determine the adjustment set W^* for each A.

Given that all causally related variables are detected and accounted for correctly in $Q(A,W)$ then the model form and parameter estimates in $Q(A,W)$ are reproducible among all populations with the same distribution of $Y|W$. As stated before this is based on the idea that the underlying causal mechanism is constant across populations with the same $Y|W$ distribution even if the marginal distribution for W shifts. We suggest that the tVIM measure β estimated through updating the causal parameter $Q(A,W)$ is also applicable to all populations with constant distributions for $Y|W$, making tVIM reproducible across populations. In this case, given a new marginal distribution of W and the estimated β, the formal tVIM measure (Ψ) can be calculated for a new population.

Once a correlation cut-off is applied, the method begins to detect non-causal variables that are highly correlated with the causal variables in the study population. This correlation structure does not necessarily hold in alternative populations. Therefore the more non-causal parameters detected the less reproducible the analysis. This is why data-adaptively selecting the correlation cut-off is important. By using a data-adaptively selected cut-off, we decrease bias while maintaining the highest level of reproducibility allowed by the current data.

3.4 randomForest

Though it does not estimate the same measure as LM, LASSO, or tVIM, randomForest is a commonly used algorithm in biomarker discovery analyses. It is applied directly to the full data matrix W using R package *randomForest* (Liaw and Wiener), which internally calculates two importance measures RF1 and RF2. However due to the nature of randomForest, there is no guarantee that all biomarkers will receive a measure of importance. Also no formal inference is available; therefore no p-values are recorded.

Random Forest (RF) is a tree-based algorithm developed by Breiman (2001). The algorithm selects a bootstrap sample of data, and then uses $\sim$2/3 of the data to fit a standard regression tree using forward recursive binary partitioning of the current covariate space, with the twist that at each node of the tree only a random sample of the variables are considered (Breiman, 2001). The prediction error on the final third of the bootstrap sample (out of bag error) is recorded. This is completed for a user specified number of trees (generally approx. 100). The algorithm assumes that all trees are independent draws from an identical distribution and assesses performance based on an overall percentage of misclassification (Breiman, 2001, 2003).

Within the random forest algorithm, Variable Importance measures are determined in an ad hoc fashion using the final RF set of trees. The first measure (RF1) is determined by running down the tree (predicting) data consisting of a perturbed version of the variable of interest and all other variables at original values. The change in the overall misclassification percentage compared to the original misclassification percentage is recorded. This is completed for 1000 perturbations of the variable. The average change in the misclassification becomes the RF1 variable importance measure reported by the algorithm (Breiman, 2003).

An additional measure of importance (RF2) relates the overall mean improvement in the gini criterion due to a particular gene. The gini criterion is used internally in random Forest to determine node splits (Breiman, 2003). For a given node in a given tree (using the selected variables) the gini criterion attempts to find the split which isolates the most prevalent category from all other categories. For a given set, S, of size n the gini criterion is calculated for categories $\{1,2\}$ as follows (Breiman et al., 1984).

$$gini(S) = a - p^2(Y=1|S)p^2(Y=2|S)$$

To determine the optimal split, the gini improvement (giniI) for a given split set $\{S_1,S_2\}$ with sizes (n_1,n_2) is calculated as follows (Breiman et al., 1984). The split with the most improvement (highest giniI) is chosen.

$$giniI(S_1,S_2|S) = gini(S) - \frac{n_1}{n}gini(S_1) - \frac{n_2}{n}gini(S_2)$$

Whenever the split is dependent on a particular gene, the giniI value is recorded. The sum of all these values normalized by the number of trees determines the RF2 importance measure for the particular gene. This importance value is also provided by the algorithm (Breiman, 2003).

The algorithm is applied to the data using the R package *randomForest* outlined in Breiman (2001) and Breiman (2003).

4 Simulation Study

4.1 Simulated Data

The full data is defined as $O = (W,Y) \sim P_0$, with covariate matrix W and outcome Y. Covariate matrix W consists of J=100 variables with n=300 observations simulated from a multivariate normal distribution with block diagonal correlation structure and mean vector created by randomly sampling mean values from $\{0.1, 0.2, ..., 9.9, 10.0, 10.1, ..., 50\}$, resulting in K=10 independent sets of variables, each correlated according to an exchangeable correlation structure with variance=1 and specified correlation ρ_{TRUE}. This forms a J by n matrix where each set of ten is correlated among themselves but independent from all other variables.

Outcome Y is simulated from a main effect linear model using one variable from each of the K sets. These K variables are designated as "true effects." The importance of a variable is determined by its coefficient value in simulation. Two sets of values are used: a constant value ($\{\beta_k = 4 : k = 1, ..., 10\}$) and an increasing set ($\{\beta_k = k : k = 1, ..., 10\}$). A normal error with mean zero and variance σ_Y is added as noise.

For both sets of coefficient values, simulations are run for $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ and $\sigma_Y = 1, 10, 20$. At $\sigma_Y = 1$ all methods perform very well, resulting in p-values much below zero. At $\sigma_Y = 20$ all methods became largely erratic and overcome by noise. Simulations at $\sigma_Y = 10$ had enough variation to highlight the different strengths of each method and are considered the most realistic noise scenario. For these reasons, only $\sigma_Y = 10$ results are presented in full.

4.2 Methods

Defining W as the biomarker set in question, importance measures according to the five methods outlined in section 4 (LM, LASSO(Q), tVIM, RF1, and RF2) are calculated for each individual biomarker, $A = W_j$ over all W ($A = \{W_j : \forall j = 1...J\}$), where J is the total number of biomarkers in set W. We also defined $W^* = W_{-j}$ as the set of possible covariates that are considered in any of the multivariate analyses. In this analysis, A is a single biomarker, however we can extend all methods analyze a set of biomarkers $\{A\}$.

Given any biomarker data set, we start by completing univariate regression analysis, estimating $E[Y|A]$ for all J biomarkers in set W. The associated coefficient values are recorded as the LM importance measures. Standard inference provides p-values which we record after adjusting for multiple testing. We use the Benjamini & Hochberg step-up FDR controlling procedure (Benjamini and Hochberg, 1995). This was applied using the mt.rawp2adjp() R function in package multtest() (Pollard et al., 2005).

In order to apply LASSO we decrease the dimensionality of $W^* = W_{-j}$, by allowing only univariate significant W^* (i.e. marginal LM adjusted p-value less than $\alpha = 0.05$). For a given A, we define the reduced covariate set as W_s. LASSO

penalized regression is applied to covariate set $\{A, W_s\}$ resulting in an initial estimate of $Q(A, W_s) = \mathbb{E}[Y|A, W_s]$. The coefficients for each A in each fit are recorded as the LASSO (Q) importance measures. The LARS implementation of LASSO (Efron and Hastie; Efron et al., 2004) does not provide any formal inference therefore p-values are not recorded.

The LASSO estimate of $Q(W_s)$ is used as the initial density estimate for tMLE procedure. We estimate $G(W) = \mathbb{E}[A|W]$ using LASSO as well citing that the additive main effect form of a LASSO derived model accurately reflects the correlation structure of the data giving us a correct estimate of $G(W)$. This guarantees under minimal ETA violations that we will obtain a consistent estimate due to the double robust nature of the tVIM measure (van der Laan and Rubin, October 2006).

We record the updated tVIM measure as well as its respective p-values are recorded. All p-values are adjusted for multiple testing using the Benjamini - Hochberg step-up FDR controlling procedure (Benjamini and Hochberg, 1995).

RandomForest is applied directly to the full data W, importance measures RF1 and RF2 are calculated internally.

4.3 Output

For each $\{\rho, \sigma_Y\}$ set, simulations of 100 are completed. Within each iteration 7 lists of 100 are saved, measures for each of 100 simulated biomarkers are recorded for each of the five methods, and p-values are recorded for univariate linear regression and tVIM. LASSO and both randomForest methods do not provide formal inference therefore no p-values are recorded. Measures and p-values are averaged over the 100 iterations. Each measure and p-value list is also translated into a list of ranks, which are also averaged over the 100 iterations. For list of measures, a rank of one is associated with the largest measure value. For list of p-values, a rank of one is associated with the lowest p-value. Each simulation run produces a set of 14 averaged lists. Sensitivity and Specificity calculations for each simulation are also determined for each individual iteration and averaged across the 100 iterations to produce the final estimates.

Although measure lists are produced for both randomForest importance measures in each simulation, only the measure rank list is used for comparison. RandomForest importance measures are not on the same scale as the univariate linear regression estimate, LASSO (Q) estimate, and tVIM estimate.

4.4 Measures of Performance

To summarize and compare results among the five methods, the following measures of performance are proposed.

4.4.1 Area Under the Curve (AUC)

The AUC value is a measure which summarizes the overall performance of a classifier. AUC (Area under the curve) values are derived from basic ROC curves which plot the true positive rate (Sensitivity) by the false positive rate (1-Specificity). These curves were originally used to determine the operating threshold for an receiver based on how well the receiver could detect a signal from noise (Flach, 2004). The plot is summarized in terms of its AUC. In pure noise conditions $AUC = 0.5$, which indicates at any threshold the false positive and true positive rate are equal (random classifier). When $AUC < 0.5$ the performance is worse than a random classifier. The more convex the resulting ROC curve, the better the classifier, or the higher the AUC the better the classifier. At AUC=1, it is considered a perfect classifier. AUC measures are often estimated using the trapezoid method. In this analysis we use function AUCi() from R package ROC which uses integrate() to calculate the AUC Carey.

AUC values are calculated and plotted versus correlation for each of the five methods using importance measure importance rank, and p-values when available. These measures are obtained by averaging across all 100 simulations.

4.4.2 Average Type I Error Rate and Power

When p-values are provided, the average type I error rate and power can be determined for any given cut-off α. Often in biomarker analyses a cut-off of $\alpha = 0.05$ is used. This cut-off provides a means to distinguish from significant and insignificant biomarkers. Here we plot the average Type I error rate and power for LM and tVIM methods versus correlation as a means to understand the impact increased correlation among biomarkers has on the importance results even when controlling for FDR from multiple testing.

4.4.3 Average Length of List

The length of the final list which contains all truly important variables is of great interest to biologists. Keeping this list short and accurate allows the biologist to spend money analyzing the top genes with confidence, knowing that the most important genes are at the top of the list. List length provides another interpretable representation of the false discovery rate (Type I error).

The required list length to find all 10 "true" variables is plotted versus correlation for all five measures and two p-value average ranked lists. The required length of list for $k = 1, ..., 10$ true variables for each available ranked list (rank by measure, rank by p-value) at each correlation level is also plotted.

4.4.4 Average Importance Value and Rank

Previous performance measures are focused on determining how well the methods rank the "true" variables with respect to all variables. Average importance measures showcase the ability of each method to not only distinguish "true" variables from decoys, but also properly determine the magnitude of importance accurately

The average importance value is plotted versus actual value for LM, Q (LASSO), and tVIM methods at each correlation level. This is only relevant for LM, Q, and tVIM, which are on the same scale as the simulated importance measures.

When the actual importance values are easily distinguished for the 10 truly dependent variables, such as when $\beta = 1, ..., 10$, we can distinguish the importance of the "true" variables relative to other "true" variables. When $\beta = 1, ..., 10$, average rank and importance should lie on the x=y line when plotting average rank or measure by "true" importance value.

The difference between the true measure/rank versus the estimated average measure/rank is summarized by calculating the mean squared deviation of the estimated values from the true values. These measures are plotted versus correlation providing a visual representation of the effect correlation among the covariates has on the overall accuracy of each method.

4.5 Results

Analysis found no appreciable difference in Sensitivity or Specificity when ranking by measure or p-value for LM and tVIM in these simulations, therefore results in terms of measure will be presently in more detail allowing us to include LASSO (Q) and RandomForest measures in all comparisons.

4.5.1 Area Under the Curve (AUC)

We compare the methods based on their AUC values at correlations, $\rho_{TRUE} = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ (Figure 1). LM, LASSO(Q), and tVIM all perform perfectly when correlation is zero, while randomForest suffers even at low correlations. However as the correlation is increased LM suffers, falling below 0.8 by $\rho = 0.5$, while tVIM performs marginally better than Q for all $\rho > 0.2$.

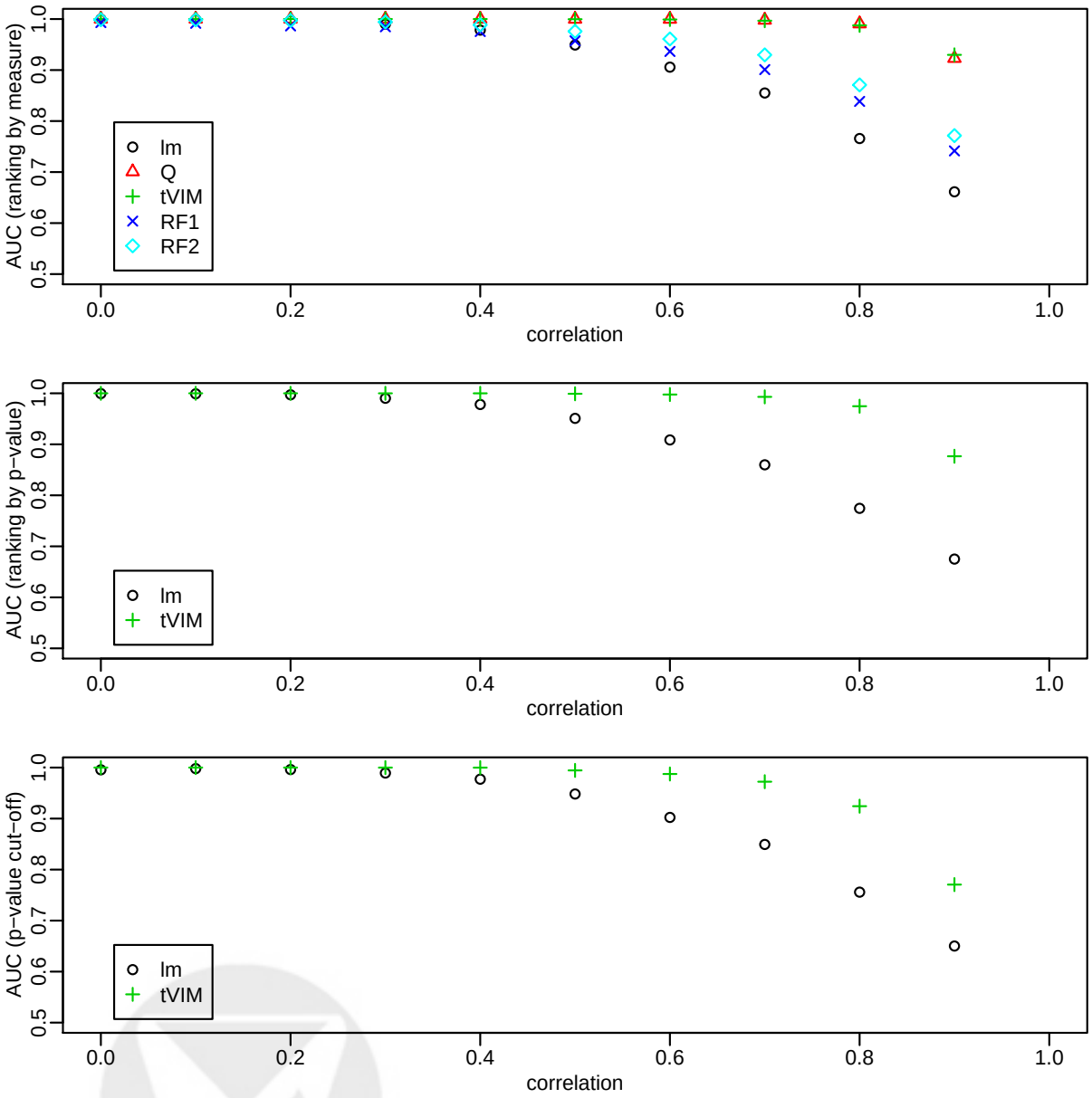

Figure 1: AUC value from ROC curves by $\rho = 0, ..., .9$ completed for (top) ranking by measure (middle) ranking by p-value, and (bottom) p-value cut-off. The later two only contain values for linear regression and tVIM ($\sigma_Y = 10$)Note: minimum AUC is 0.5, maximum and optimum is AUC=1. Simulation is done with $\sigma_Y = 10$ for n=300 with total number of variables at 100 of which 10 are truly related to the outcome. At zero correlation, LASSO (Q), tVIM, and LM perform perfectly with AUC=1. Plots are shown for constant $\beta = 4$, but results are comparable when $\beta = \{1, ..., 10\}$.

4.5.2 Average Type I Error Rate and Power

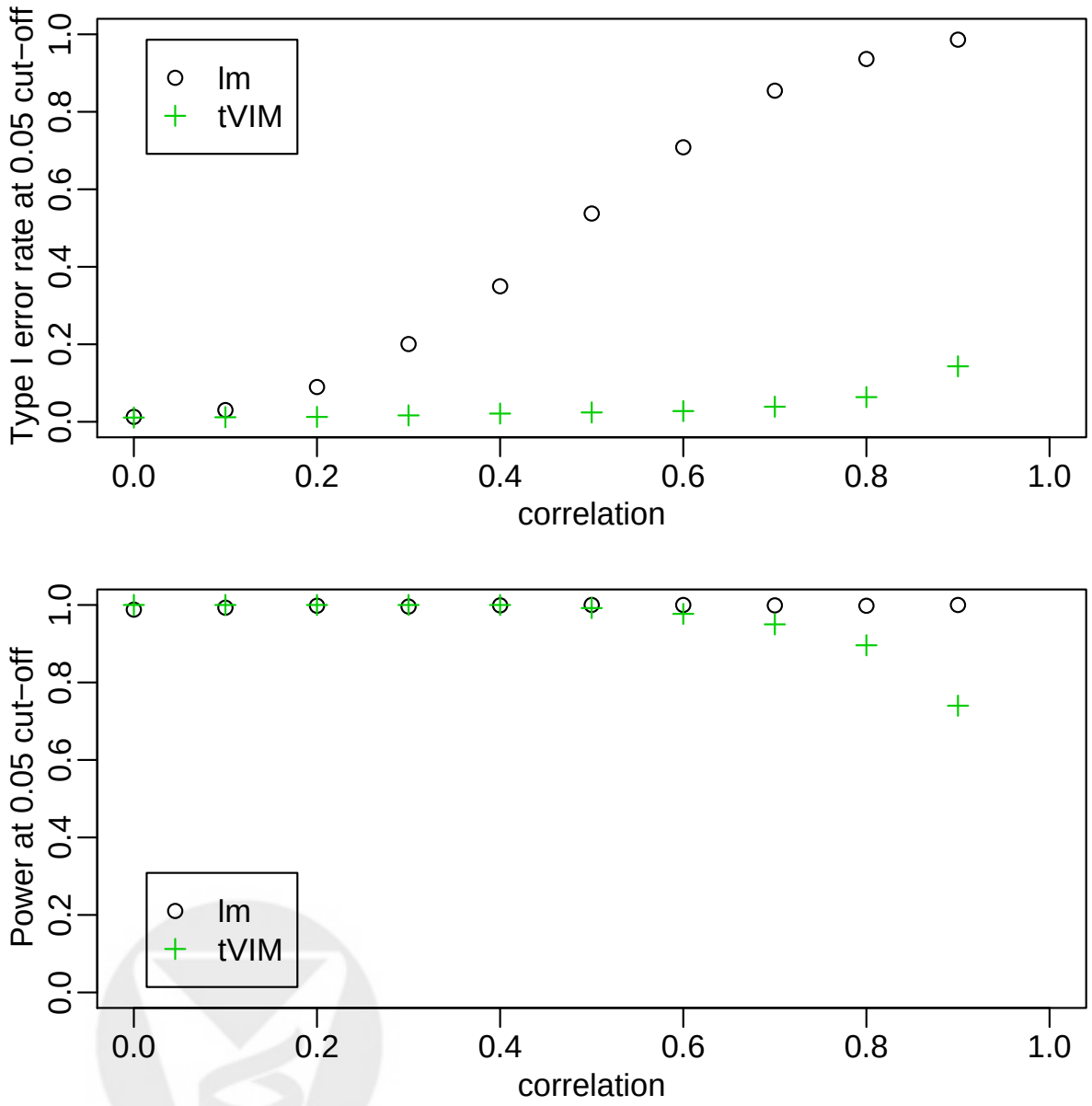

Figure 2: Controlling at $\alpha = 0.05$ (top) type I error rate and (bottom) power by $\rho = 0, ..., .9$. only contain values for linear regression and tVIM ($\sigma_Y = 10$). Plots are shown for constant β_{TRUE}, but results are comparable when $\beta_{TRUE} = \{1, .., 10\}$

Initial Simulations (not pictured) were performed using a naive estimate of Q which included all significant covariates in Q(A,W) and G(W). These simulations showed no Type I error but suffered a significant loss in power at higher correlation values due to over-fitting Q, where as marginal regression method showed no power loss but a large increase in type I error as correlation increased. To elevate this over-fitting we used LASSO penalized regression to estimate Q(A,W) and G(W) which effectively increased the power overall. A slight power loss was still apparent for tVIM at higher correlations

($\rho > 0.7$), but it is minimal when compared to the power loss of LM as correlation increases (Figure 2).

4.5.3 Average Length of List

Length of list is a direct reflection of Type I error rate. We see that overall tVIM performs best, though the improvement over LASSO is less clear when β_{TRUE} is constant. In general, tVIM has the shortest list and is less affected than any other methods by increases in correlation.

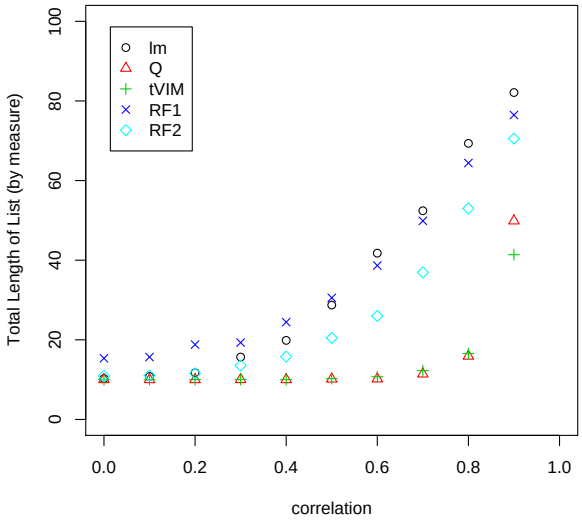

Figure 3: Total length of list required to have all ten true variables in the list by $\rho = 0, ..., .9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), targeted Variable Importance with LASSO (tVIM) and two randomForest based importance measures (RF1, RF2) are shown. Here β_{TRUE} is constant at 4.

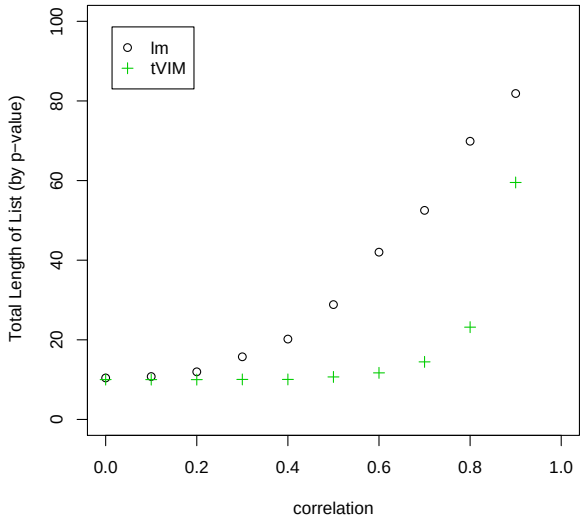

Figure 4: Total length of list required to get all ten true variables by $\rho = 0, ..., .9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and targeted Variable Importance with LASSO (tVIM) are shown. Here β_{TRUE} is constant at 4.

When β is constant at value four (Figure 3,4), list length for VImp was at its minimum until 0.6, reflecting the zero Type I error rate. Increases above the minimal list length were seen for both random forest and univariate regression methods at correlation values greater than 0.4, with random forest faring better at higher correlations. Note that univariate regression performs better than randomForest when correlations are less than 0.4. When ranking by p-value and similar trend for marginal regression was apparent. However tVIM did increase above the minimum list length at correlation 0.9, indicating an increase in Type I error rate.

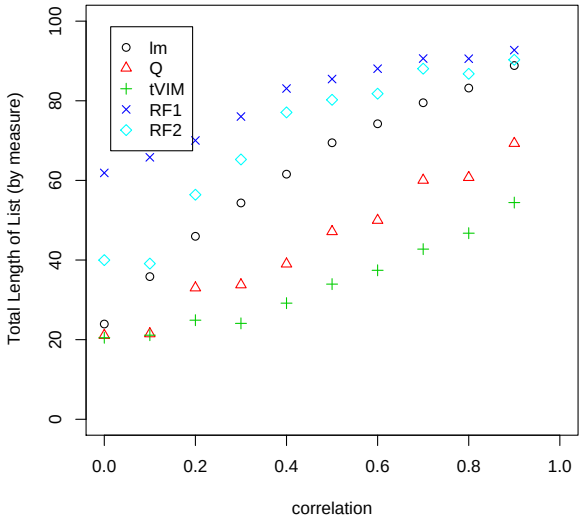

Figure 5: Total length of list required to have all ten true variables in the list by $\rho = 0, ..., .9$, ranking by importance measure. ($\sigma_Y = 10$) Results for univariate regression (LM), LASSO (Q), targeted Variable Importance with LASSO (tVIM) and two randomForest based importance measures (RF1, RF2) are shown. Here β_{TRUE} is set at $\{1, ..., 10\}$.

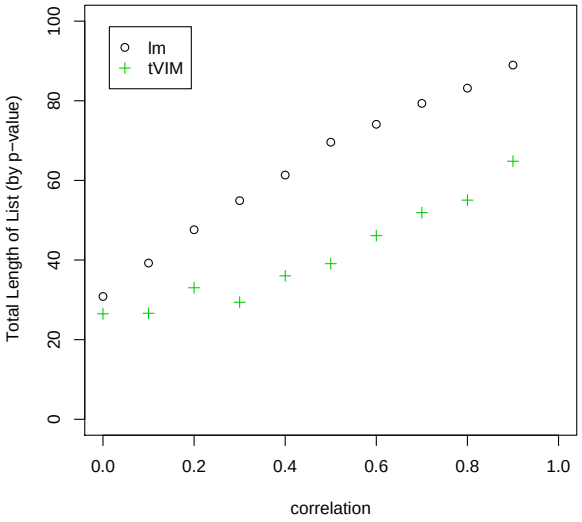

Figure 6: Total length of list required to get all ten true variables by $\rho = 0, ..., .9$, ranking by p-value. ($\sigma_Y = 10$) Results for univariate regression (LM), and targeted Variable Importance with LASSO (tVIM) are shown. Here β_{TRUE} is set at $\{1, ..., 10\}$.

In the case where β_{TRUE}= $\{1, ..., 10\}$ (Figures 5,6), the improvement of tVIM over LASSO is more pronounced, but detection of the first variable (with the lowest β value) is difficult for all methods. When ranking by measure or p-value, all methods have their lowest list length around 20 variables while the total number of variables expected is 10. When β was constant at value 4, the lowest list length was near its minimum at 10 (Figures 3,4). The shift in list length is due to the importance value for the variable associated with $\beta = 1$. At such a high noise level ($\sigma_Y = 10$), the lower importance values are more difficult to distinguish from the noise. This is apparent by comparing the average importance rank and average importance value for the variable with $\beta = 1$ (see figures 7,9). The rank is much higher than 10, but the value is close to one as it should be.

4.5.4 Average Importance Value

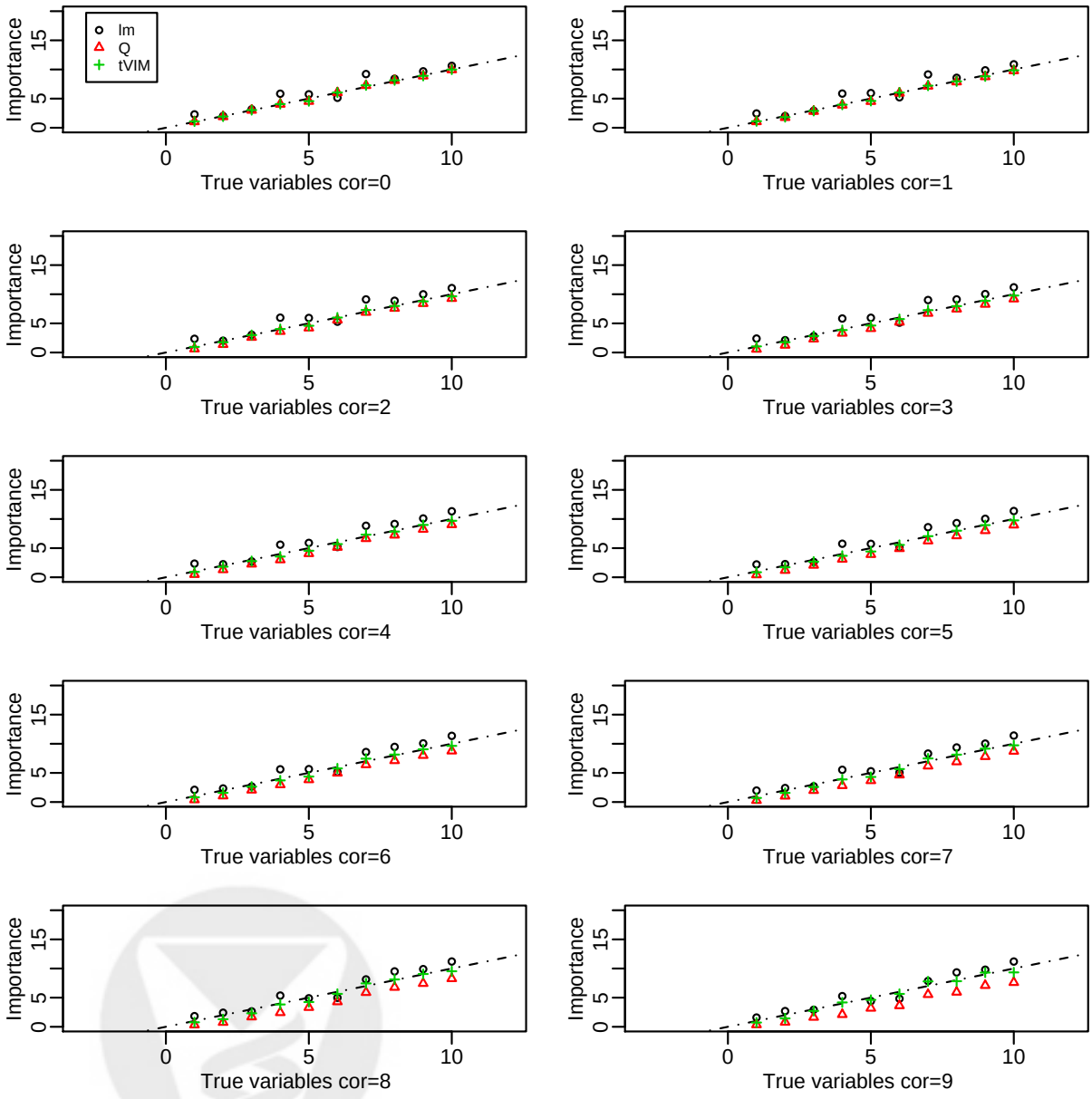

Figure 7: Average importance value for each of ten true variables with importance values = 1,...,10. Plots included for all $\rho = 0, \ldots, .9$. Only linear regression, LASSO, and tVIM are analyzed since RF values are not necessarily on the same scale as the true level of importance. ($\sigma_Y = 10$)

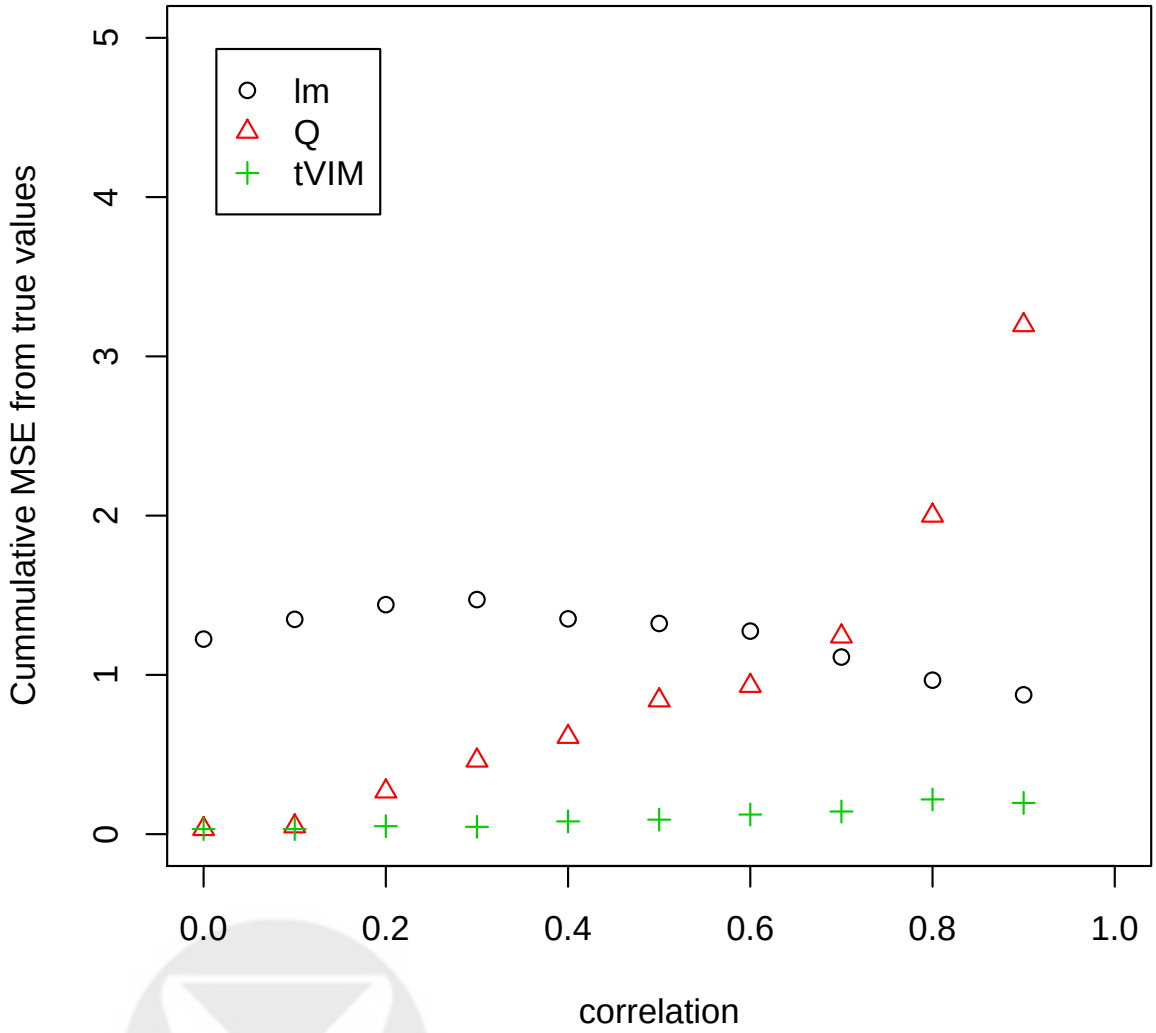

Figure 8: Mean square error difference between average importance values and true values at $\rho = 0, \ldots, .9$. This relates to Figure 7.

We see in Figure 7 that VImp method estimates the actual importance values accurately even at higher correlation values. After correlation of 0.2 LASSO (Q) measures began to deviate from the true values and become worse than LM estimates at 0.8 and above. This trend is more visible in the MSE plot (Figure 8). Linear regression does approximately the same across all correlations.

4.5.5 Average Importance Rank

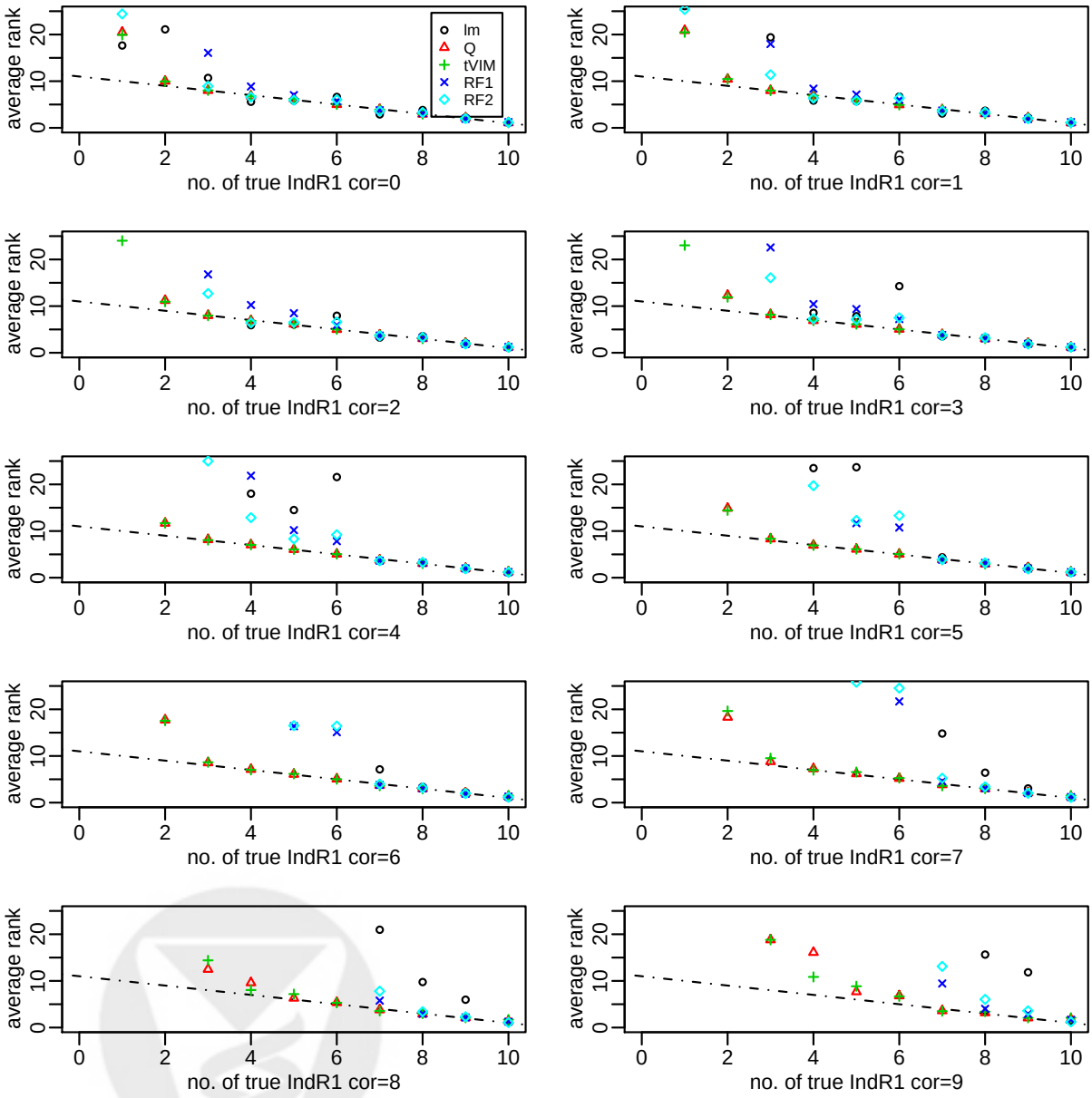

Figure 9: Average importance rank for each of ten true variables with actual ranks = 1,...,10. Plots included for all $\rho = 0, \ldots, .9$, ranking by measure. ($\sigma_Y = 10$). Comparing the methods by rank allows us to include RF1 and RF2 in the comparison even though their measures are on a different scale

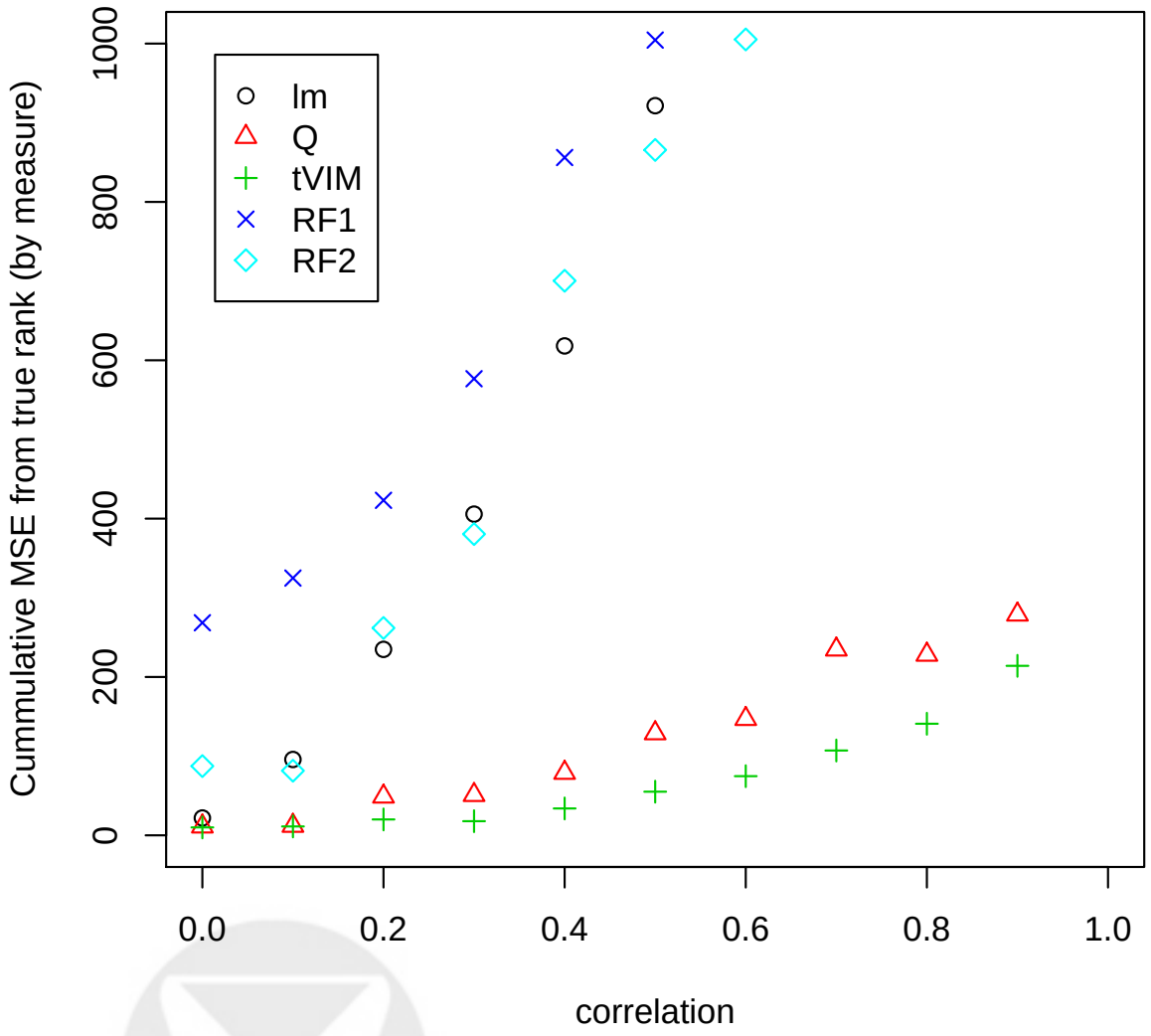

Figure 10: Mean square error difference between average importance ranks and true ranks at $\rho = 0, ..., .9$, when ranking by measure. This relates to Figure 9. Comparing the methods by rank allows us to include RF1 and RF2 in the comparison even though their measures are on a different scale.

In Figure 9, we can see clearly that the variables with $\beta = 1$ and 2 are harder for all methods to pinpoint and rank accurately, however tVIM performs the best overall correlation values as seen in the MSE plot (Figure 10).

4.6 Discussion - Simulations

From the simulation results, tVIM seems the most resilient to increases in correlation, providing the most accurate and concise lists of ranked variables. From AUC measures, tVIM and Q (the LASSO results) remain near one until correlation

increases above 0.5, while linear regression falls drastically once correlation is above 0. This decrease is due to increases in type I error rate which is seen very clearly in the length of list results.

Though LASSO, by accounting for confounding, is more resilient to type I error increases, tVIM still outperforms LASSO. Not only is ranking more accurate at higher correlations, but the measures are also more accurate. When using LASSO, the true importance values begin to decrease under high correlation. This is likely due to the regression increasing the coefficients of the variables highly correlated with variable A.

Importance measures from randomForest seem to perform worse than tVIM or LASSO in all comparisons. It's performance at low correlation is equivalent to LM, only improving on LM once correlations increase above 0.3. RF1 and RF2 consistently identify irrelevant variables as important even under low correlation conditions. RF2 in general outperforms RF1 as a measure of importance under increases in correlation.

These simulations address each methods ability to accurately identify the causally related genes as the correlation among variables increases. Though tVIM performs better than the three other methods, it is still sensitive to more extreme correlations (0.7-0.9). Our simulations show only a small increase in bias for the measure of the causally related variables at higher correlations. However, in practice, high correlation can adversely effect the tVIM estimate due to violation of the experimental treatment assumption. The increased length of the variable list when ranked by importance measure at correlation 0.8 and 0.9 indicates that even tVIM cannot distinguish the causally related variable among a group of variables when correlation is very high.

4.6.1 Highly correlated Variables - Violations of ETA

In cases of extreme correlation, it is generally impossible to distinguish the underlying causally related variables. In these cases, it is reasonable to label all potentially relevant variables as important, but it is also important to obtain the most accurate importance estimates for this list, reducing any bias that might be present due to ETA violations. We explored this case briefly in simulation.

In an effort to avoid bias due to ETA violation, a correlation cut-off was applied to subset W_s for each A before LASSO analysis. In this scenario, W_s is restricted to $W_i \in W_s$ where $cor(W_j, W_i) < \delta$, for various cut-offs $\delta = \{0.5, 0.75, 0.9, 1\}$. We applied this method to our simulated datasets from the previous section. Results showed that such a restriction resulted in the elimination of relevant W_i from the estimate of $\mathbb{E}[Y|A, W_s]$. In other words when A_d is a decoy variable highly correlated with a true variable W_t. Restrictions on the covariate set remove W_t from the possible covariate set for A_d, resulting in A_d having a higher more significant importance that it would have otherwise.

Applying a restriction of ρ_c will result the algorithm identifying all true variables as well as variables whose correlation with the true variables is higher than ρ_c. Once we select ρ_c, we are conceding that variables with correlations greater than ρ_c cannot be teased apart to determine the true underlying causal variable. By applying the correlation cut-off we are redefining our parameter. It is no longer the singular causal effect of A. Instead, we are estimating a correlation delta - controlled W adjusted variable importance, which admits that given the data, the true causal variable cannot be targeted when the data is highly correlated. Instead we measure an importance of A controlling for a newly defined subset of W. Given this new definition of the parameter, all importance variables according to the delta-controlled method include all causal variables and all variables whose correlation to a causal variable is greater than a particular delta cut-off. In practice each variable may have its own particular correlation cut-off.

We must select ρ_c high enough to reduce bias from ETA violation, but low enough to acquire all information on the causal effects allowed by the data, which maintains the greatest level of reproducibility. If ρ_c is higher than necessary, the list will contain decoy variables that could have been discounted using the available data. This would decrease the reproducibility of the measures in other populations. The relationship between the decoy variables and the causal variables (distribution of W) is not necessarily constant across populations while the causal mechanism (distribution of $Y|W$) can be assumed to be (i.e. the mechanisms of disease are consistent across all populations). Including decoy variables that could otherwise have been discounted adds unnecessary uncertainty when applying the final results to other populations.

We do not present simulation results using the correlation cut-off in full, however we do apply the correlation cut-off ($\rho_c = \{0.5, 0.75\}$) in application, where the truth is unknown and the data noisy. In application, we are interested in detecting all potentially relevant variables. Future applications should apply methods outlined in Bembom et al. (March 2008) which data-adaptively selects the correlation cut-off for each A, reducing the bias while detecting the most accurate gene set allowed by the data, maintaining reproducibility.

Table 1: Contrasting AML to ALL type Leukemia (ref)

Traits	AML	ALL
Effected population	Most common in adults	Most common in children
Biological characteristics	Identified with the myeloid line of white blood cells, which includes any leukocyte that is not a lymphocyte	Identified with abnormal lymphocytes, B-cells, T-cells, and natural killer (NK) cells
Clinical characteristics	Anemia, fatigue, weight loss, easy bruising, thrombocytopenia, and granulocytopenia with bacterial infections	Anemia, fatigue, weight loss, easy bruising, thrombocytopenia, and granulocytopenia with bacterial infections, bone pain, lymphadenopathy and hepatosplenomegaly
Relative overall survival rate	Low	Higher in both adults and childern
Additional characteristics		Can spread to the nervous system

5 Applications

5.1 Golub et al (1999)

5.1.1 Data

The dataset from Golub et al. (1999) has been used in many papers for methodological comparison, due to its relevance, limited gene set, and biological interpretability. One goal in the original study was to identify differentially expressed genes in patients with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In table 1 we summarize the basic differences between ALL and AML type leukemia.

Gene expression levels were measured using Affymetric oligonucleotide arrays with 6,817 human genes for n=38 patients (27 ALL, 11 AML). The gene expression set was pre-processed and reduced to 3,051 genes according to methods described in Dudoit et al. (2002). This dataset was obtained from the R package *multtest*, dataset golub (Pollard et al., 2005).

5.1.2 Analysis

This analysis mirrors the procedure implemented in the previous simulations. Univariate Linear regression is applied to all genes. We control for multiple testing by applying Benjamini & Hochberg step-up FDR controlling procedure (Benjamini and Hochberg, 1995) to the resulting univariate p-values. In this application we chose all genes with p-value less than 0.1 as the initial covariate set W^* for any given A. This set contains 550 genes initially. Unlike the previous simulations where the true is known and we are interested in measuring the performance of the method, in practice it is important to minimize bias due to ETA violations. A simple correlation cut-off of $\rho_c = \{0.5, 0.75\}$ is applied. In application where we do not known the truth and the data is especially noisy with a complex correlation structure, often effect of a gene can not be disentangled. Applying the correlation cut-off results in all potentially relevant genes labeled as important.

As in simulation we model the importance as $m(A, W_s|\beta) = \beta A$ for all A. For the initial Q we use a polynomial spline fit which allows for more complex structure of $g(W_s)$. We recommend using this or a similar data-adaptive algorithm such as DSA (Sinisi and van der Laan, March 2004) over LARS/Lasso in application, since in reality the structure of Q may have more than just additive main effects. We also estimate $G(W_s)$ using polymars (O'Connor). P-values are determined using standard t-tests.

In this as in many applications, the outcome is binary (ALL (Y=0) vs. AML (Y=1)), therefore we can interpret our tVIM measure as an estimate of excess risk.

$$m(A = a, W_s|\beta) = \mathbb{E}[Y|A = a, W_s] - \mathbb{E}[Y|A = 0, W_s] = \beta a$$

The model-based approach outlined in this paper must use standard gaussian regression for our estimate and update of $\mathbb{E}[Y|AW_s]$. However we believe the final list of ranked VIM measures and p-values are still relevant regardless. Future work

is focused on the development of a more generalized model-based VIM approach which will allow us to use generalized linear regression methods.

Updated tVIM measures and p-values are recorded and we adjust for multiple testing using Benjamini & Hochberg (1995) step-up FDR controlling procedure (Benjamini and Hochberg, 1995). We recommend selecting all genes with adjusted p-values less than or equal to an appropriate cut-off (we use a standard cut-off of 0.05), and then ranking this set of genes by their absolute tVIM measure to achieve the final importance ranking of genes. The same method is used to rank genes according to the LM measures and p-values. RF1 and RF2 importance measures are simply ranked. We compare the results in Tables 2-6.

5.1.3 Results

Using a p-value cut-off of 0.05, tVIM results in 272 significant genes at $\rho_c = 0.5$ and 225 significant genes at $\rho_c = 0.75$, while LM identifies 681 significant genes. We plot the number of significant genes for a range of p-value cut-offs in Figure 11. The overall trend showing LM to be much more conservative than tVIM which is what is expected.

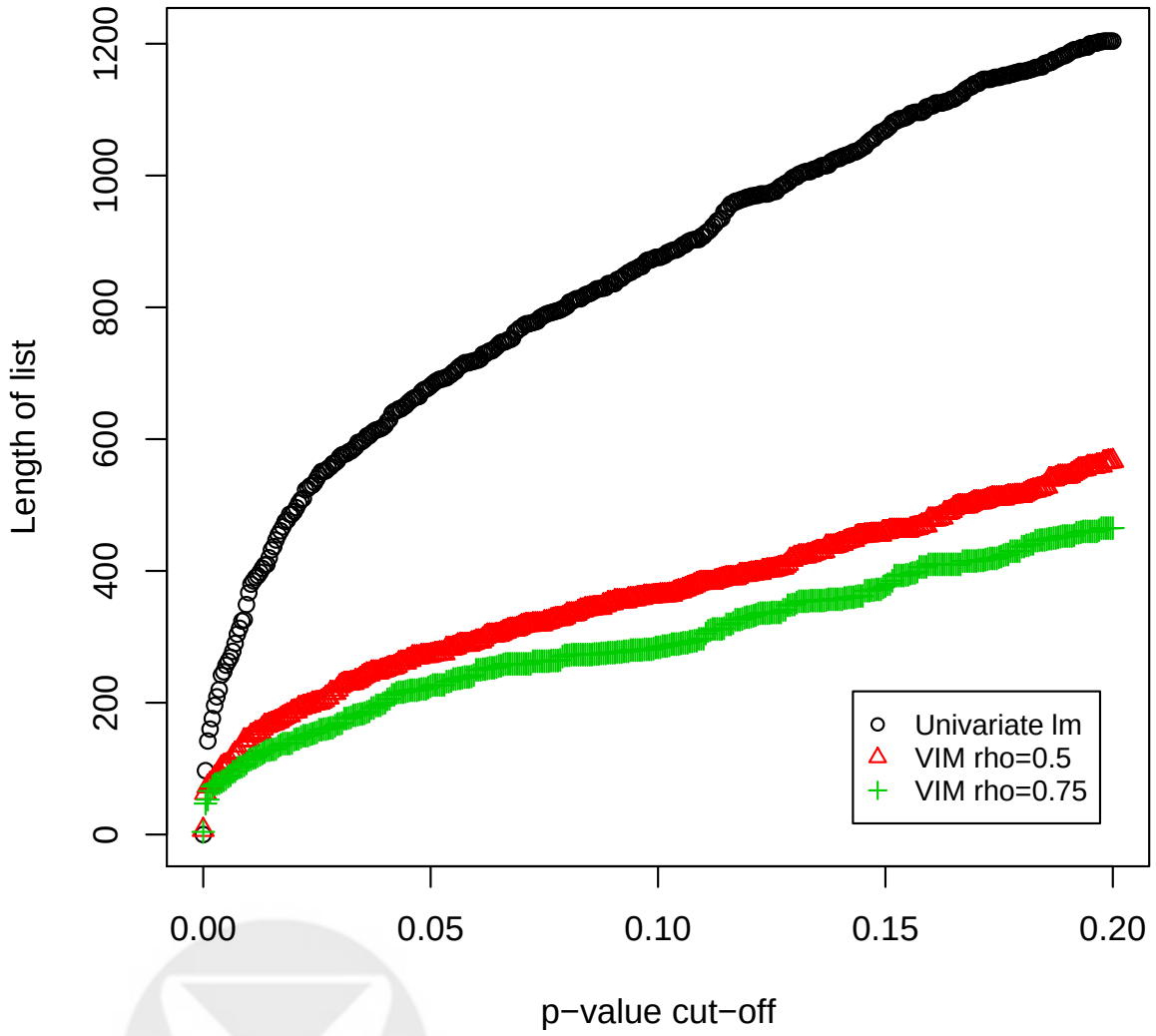

Figure 11: The number of significant genes (y-axis) given a p-value cut-off (x-axis) is plotted for LM and tVIM results for $\rho_c = 0.5, 0.75$

The top 10 genes according to their importance ranking for LM, RF1, RF2, and VIM ($\rho_c = 0.5, 0.75$) are shown below (Tables 2-6).

Table 2: Univariate Linear regression (LM): Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

Gene Name/Gene Symbol	Mapped IDs	LM	LM rankp	VIM rankp (0.75)	VIM rankp (0.5)	RF1 rank	RF2 rank
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891.at	0.258	1	13	17	6	3
CYSTATIN A	D88422.at	0.341	2	521	466	12	8
Zyxin	X95735.at	0.345	3	287	534	2	2
Macmarcks	HG1612-HT1612.at	-0.619	4	1041	1768	9	9
CD33 CD33 antigen (differentiation antigen)	M23197.at	0.517	5	906	28	26	22
C-myb gene extracted from Human (c-myb) gene, v-myb myeloblastosis viral oncogene homolog (avian)	U22376.cds2.s.at	-0.403	6	69	99	40	28
ELA2 Elastatse 2, neutrophil	M27783.s.at	0.334	7	104	1970	15	14
DF D component of complement (adipsin)	M84526.at	0.262	8	175	145	96	149
RETINOBLASTOMA BINDING PROTEIN P48	X74262.at	-0.431	9	291	266	57	31
Leukotriene C4 synthase (LTC4S) gene	U50136.rna1.at	0.725	10	146	2110	38	60

Table 3: RF1: Top 10 ranked genes according to their importance measures

Gene Name/Gene Symbol	Mapped IDs	RF1	RF1 rank	RF2 rank	tVIM rankp (0.75)	tVIM rankp (0.5)	LM rankp
FAH Fumarylacetoacetate	M55150.at	0.953	1	1	588	234	52
Zyxin	X95735.at	0.823	2	2	287	534	3
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	M31523.at	0.718	3	6	155	400	12
ADM Adrenomedullin	D14874.at	0.693	4	5	329	2136	57
PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta	M31166.at	0.691	5	33	33	201	28
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891.at	0.682	6	3	13	17	1
TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115.at	0.654	7	4	1	2	33
CCND3 Cyclin D3	M92287.at	0.621	8	10	481	924	19
Macmarcks	HG1612-HT1612.at	0.613	9	9	1041	1768	4
APLP2 Amyloid beta (A4) precursor-like protein 2	L09209.s.at	0.610	10	7	160	408	25

Table 4: RF2: Top 10 ranked genes according to their importance measures

Gene Name/Gene Symbol	Mapped IDs	RF2	RF2 rank	RF1 rank	tVIM rankp (0.75)	tVIM rankp (0.5)	LM rankp
FAH Fumarylacetoacetate	M55150.at	0.426	1	1	588	234	52
Zyxin	X95735.at	0.282	2	2	287	534	3
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891.at	0.218	3	6	13	17	1
TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115.at	0.208	4	7	1	2	33
ADM Adrenomedullin	D14874.at	0.200	5	4	329	2136	57
TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	M31523.at	0.186	6	3	155	400	12
APLP2 Amyloid beta (A4) precursor-like protein 2	L09209.s.at	0.183	7	10	160	408	25
CYSTATIN A	D88422.at	0.171	8	12	521	466	2
Macmarcks	HG1612-HT1612.at	0.164	9	9	1041	1768	4
CCND3 Cyclin D3	M92287.at	0.159	10	8	481	924	19

Table 5: tVIM using correlation cut-off of $\rho_c = 0.5$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

Gene Name/Gene Symbol	Mapped IDs	tVIM	tVIM rankp (0.5)	tVIM rankp (0.75)	LM rankp	RF1 rank	RF2 rank
TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115.at	-0.973	1	2	33	7	4
CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7	X70297.at	0.839	2	1	48	69	61
corneodesmosin	L20815.at	0.338	3	3	1875	1846	2004
BCL3 B-cell CLL/lymphoma 3	U05681.s.at	0.314	4	4	477	558	821
KTN1 kinectin 1 (kinesin receptor)	Z22551.at	-0.311	5	18	373	2967	118
CaM kinase II isoform mRNA	U81554.at	0.272	6	81	367	476	749
TCF7 transcription factor 7 (T-cell specific, HMG-box)	X59871.at	-0.159	7	6	569	635	887
PTTG1IP pituitary tumor-transforming 1 interacting protein	Z50022.at	0.310	8	5	2753	2674	483
MCL1 myeloid cell leukemia sequence 1 (BCL2-related)	L08246.at	0.293	9	2406	61	75	65
PI3K Phosphatidylinositol 3-kinase	Z46973.at	-0.172	10	113	734	772	1009

Table 6: tVIM using correlation cut-off of $\rho_c = 0.75$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

Gene Name/Gene Symbol	Mapped IDs	tVIM	tVIM rankp (0.75)	tVIM rankp (0.5)	LM rankp	RF1 rank	RF2 rank
CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7	X70297.at	1.260	1	2	48	69	61
TOP2B Topoisomerase (DNA) II beta (180kD)	Z15115.at	-0.946	2	1	33	7	4
corneodesmosin	L20815.at	0.327	3	3	1875	315	621
BCL3 B-cell CLL/lymphoma 3	U05681.s.at	0.181	4	4	477	316	622
Surface glycoprotein	Z50022.at	0.310	5	8	2753	317	474
TCF7 Transcription factor 7 (T-cell specific)	X59871.at	-0.175	6	7	569	318	623
CAT Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)	X04085.rna1.at	0.163	7	21	92	56	59
E2F4 transcription factor Dp-2 (E2F dimerization partner 2)	U18422.at	-0.256	8	42	1752	319	624
UGP2 Uridine diphosphoglucose pyrophosphorylase mRNA	U27460.at	-0.244	9	14	155	186	303
SELL Leukocyte adhesion protein beta subunit	M15395.at	0.183	10	43	340	84	316

When comparing the four lists it is difficult to determine which list is better. Especially when the lists include hundreds of genes. In this analysis we compare the top 10 of each list in an effort to compare their biological relevance. In any given list we include the top 10 genes of the particular method along with their ranks for all other methods. For many of the genes, these ranks vary greatly over the different methods. By consulting the literature, we hope to gain insight on the biologically validity of each list.

Among the top 10 genes according to LM results, CSTA, CD33, MYB, and ELA2 have all been associated with various types of cancer in the literature in previous quantitative analyses of association. CSTA has been proposed as a diagnostic and prognostic biomarker for cancer (Kos and Lah, 1998). CD33 antigen has been shown in vitro to induce apoptosis in AML cells (Vitale et al., 2001). MYB is the homolog of an avian viral oncogene (Clappier et al., 2007), and ELA2 has been related to acute promyelocytic leukemia (Lane and Ley, 2003).

Among the top 10 genes according to RF1 and RF2, all genes in RF1 were also in the top 10 for RF2 except CBX1, it was replaced by CSTA in the list for RF1. CSTA was also in the top 10 of LM. Out of the top 10 the following genes have been associated with various cancers: TCF3, TOP2B, CCND3, and CSTA. Chromosomal abnormalities in TCF3 have been linked to T-cell and B-cell ALL (Hunger, 1996). TOP2B is a current drug target having been linked to drug resistant cancers (Nebral et al., 2005; Kaufmann et al., 1998a). CCND3 is a cyclin D. In the absence of cyclin D's cells have shown increased resistance to oncogenic transformation in mouse models (Kozar et al., 2004).

There are marked differences and similarities between the tVIM based results using a correlation cut-off of 0.5 and 0.75. There are 5 genes that are common between the two lists, 4 of which have some cancer-related association : TOP2B, CHRNA7, BCL3, and TCF7. Directional relationships remain consistent between the two lists, but the magnitudes shift due to the different covariate sets. TOP2B, a current drug target (Nebral et al., 2005; Kaufmann et al., 1998a), was also identified by randomForest. BCL3 is a proto-oncogene biologically associated with B-cell ALL (Martin-Subero et al., 2007). TCF7 is a known biomarker for T-cell ALL, and is rarely expressed in AML cancer cells (Palomero et al., 2006). CHRNA7 has recently found to inform the role of nicotine in colon cancer (**?**). It is also important to note that CHRNA7 is highly correlated with CD33. Cancer relevant genes found only in table 5 ($\rho_c = 0.5$) are PTTG1IP, MCL1, PI3K, and

CAMK2G. PTTG1IP has been consistently found overly expressed in human tumors (Ramaswamy et al., 2003; Puri et al., 2001; Fujii et al., 2006; Zhu et al., 2006). MCL1 is related to BCL2 and is a negative regulator of apoptosis (Kaufmann et al., 1998b). PI3K is activated by cellular agents known to stimulate B and T cells (Fruman et al., 1999). CAMK2G has an active role in cell growth control and has tumor cell-specific variants (Tombes, 1997). Cancer relevant genes found only in table 6 ($\rho_c = 0.75$) are CAT and E2F4. CAT regulates BCL-2 and is often under-expressed in ALL tissues (Senturker et al., 1997; Komuro et al., 2005). E2F4 has an essential role in cell proliferation and cell fate decisions (Balciunaite et al., 2005) as well as activation of tumor suppressor proteins (Leone et al., 2001).

5.1.4 Discussion

The Golub et al 1999 AML/ALL dataset has been extensively used to demonstrate biomarker identification and classification methods due to its accessibility and its limited gene set (only 3,051 out of over 30,000 genes in the annotated human genome). Using simple univariate linear regression 681 genes were significant at the 0.05 level after adjusting for multiple-testing. However we know from general knowledge and our simulations, that univariate linear regression is highly sensitive to correlation among the variables, leading to large increases in type I error rate. Given this and a set of 681 genes, attempting to further analyze the lists to identify and biologically verify the relevant genes seems a nearly impossible and very expensive task.

Attempting to control type I error by adding additional covariates requires model selection methods that are geared towards prediction. RandomForest for instance is a prediction and classification method which includes a type of model selection. However understanding the resulting importance values is difficult. Given an importance value of 0.612 the relationship between the variable and the outcome is unclear - is it highly expressed in AML or ALL? We only know that that variable is more "important" than a variable with a value of 0.611. Also out of the top 10 lists for RF1 and RF2 (12 genes total), only four genes were found to be biologically associated with cancer and only 1 specifically relating to ALL/AML distinction, TCF3. Why TCF3 is rated second for RF1 and 6th for RF2 is unknown. In comparison, LM found four related to cancer, two of which specifically related to AML/ALL.

The tVIM measure provides directionality and is less sensitive to increases in correlation (see simulations). Given a importance measure of -0.175, we can conclude that this particular gene is up-regulated in ALL patient when compared to AML patients. This particular measure is for TCF7 using a correlation cut-off of 0.75. TCF7 is rarely expressed in AML and often highly expressed in ALL patients (esp. T-cell related). Out of the 6 cancer related genes in the top 10 list for 0.75 cut-off, 3 are biologically related to the AML/ALL distinction. When the cut-off is 0.5, there are 8 cancer related genes, 3 related to the AML/ALL distinction.

When comparing the four lists it is difficult to determine which list is better. By shifting through the literature, we had hoped to gain a sense of biologically validity associated with each list. Targeted VIM does have a greater number of cancer-related genes and a greater number of specifically AML/ALL related genes. However the increase over LM is small, and the comparison only includes the top 10 genes. Further support for tVIM is gained from the previous simulations where we demonstrated its resistance to increases in correlation and its control of type I error, while still being an interpretable and meaningful measure of importance. For all three AML/ALL related genes the directionality of the relationship is biologically correct.

5.2 van't Veer et al (2002)

The response to standard chemotherapy among breast cancer patients can drastically vary even among women with a common stage of breast cancer at initial diagnosis. Chemotherapy is a very long and difficult treatment process, and though it is known to reduce the occurrence of metastases in 70-80% of patients, for the remaining 30-20% there is little or no response. Knowing a priori a probability of response to treatment for a given patient would aid doctors in determining a more optimal and efficient treatment plan, reducing patient discomfort and the cost of expensive trail-and-error treatment regimes. This is reflective of the current trend towards the development of individualized or "patient-tailored" treatments.

The study in van 't Veer et al. (2002) attempts to develop a classifier predicting treatment response to adjuvant chemotherapy among breast cancer patients based on their pre-treatment (at diagnosis) genetic profile. Given that there are over 20,000 protein-coding genes in the human genome, developing a predictor requires first reducing the data to a set of relevant genes. Here we present an application of tVIM as a method to identify these genes. Unlike linear regression and other data mining

algorithms (randomForest, etc.), tVIM targets the causal effect instead of estimating only an association based on a predictive fit. We propose using tVIM to determine this subset of genes prior to the application of the prediction algorithm Super Learner (van der Laan et al., July 2007).

The initial dataset contains 98 patients with similar stages of breast cancer at the time they enter the study. All patients are exposed to adjuvant chemotherapy. It is unknown if any other treatment methods (i.e. radiation, surgery, etc.) are applied and to what extent. For the purposes of the van't Veer analysis, the patients are assumed to be part of the same treatment arm. We continue with that assumption. Of the 98 patients, 34 develop metastases within 5 years (bad responders, Y=1), while 44 remain disease free (good responders, Y=0) (van 't Veer et al., 2002).

5.2.1 Analysis

For computation considerations we reduced our dataset to genes whose raw p-values from univariate linear regression were less than or equal to 0.05 (2254 genes) or those which had a randomForest importance value greater than zero. We also did not include genes with more than 80% of their values missing. This left us with a total of 4446 genes. Its important to note that once we adjusted for multiple testing, there were no adjusted significant p-values at the 0.05 level among these genes. All missing data is imputed with the column mean (average gene expression over all patients). The maximum number of missing values for any gene was five.

We apply targeted VIM using correlation cutoffs of 0.5 and 0.75 as outlined fin Section 6.1.2. The covariate set prior to correlation cut-off included all genes among the 4446 whose raw univariate linear regression p-value was less than or equal to 0.01 (540 genes). Genes significant at the 0.05 level are used as input to Super Learner these results are outlined in Polley et al. 2008. Here we explore the relevance of the genes obtained using tVIM. We again rank the significant genes by their tVIM values.

5.2.2 Results

There were no statistically significant genes (at the 0.05 level) once the univariate linear regression p-values were adjusted for multiple testing, while for tVIM there were 197 and 204 genes when correlation cut-off was set at 0.5 and 0.75 respectively. We show the top 10 significant genes in table 7 and table 8.

Table 7: Targeted VIM using correlation cut-off of $\rho_c = 0.5$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

p-value	tVIM	GeneID	Description/Function
0.00E+00	6.455	GALNT14 (AA165698)	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 14
0.00E+00	6.164	AIP	aryl hydrocarbon receptor interacting protein
3.03E-05	3.517	LRTM1	leucine-rich repeats and transmembrane domains 1
2.33E-07	3.125	ZBTB22	zinc finger and BTB domain containing 22
6.94E-08	3.111	(AI524306)	unknown
0.00E+00	-2.843	FBXO41 (AA524093)	F-box protein 41
1.58E-06	-2.714	VAMP3	vesicle-associated membrane protein 3 (cellubrevin)
2.26E-02	-2.590	ERGIC1 (AI248720)	endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1
3.27E-03	2.564	CALCOCO1	sarcoma antigen nysar3
4.38E-02	2.546	NRG2	neuregulin 2

Table 8: Targeted VIM using correlation cut-off of $\rho_c = 0.75$: Top 10 ranked genes according to absolute importance measures among significant genes according to a p-value cut-off of 0.05.

p-value	tVIM	GeneID	Description/Function
0.00E+00	6.455	GALNT14 (AA165698)	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 14
0.00E+00	5.906	AIP	aryl hydrocarbon receptor interacting protein
0.00E+00	3.703	LRTM1	leucine-rich repeats and transmembrane domains 1
3.23E-08	3.609	(AI524306)	unknown
1.70E-08	3.331	ZBTB22	zinc finger and BTB domain containing 22
1.68E-06	-3.001	METTL1	methyltransferase like 1
6.94E-04	2.950	EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
9.88E-03	2.932	SH2D3C	SH2 domain containing 3C
0.00E+00	-2.843	FBXO41 (AA524093)	F-box protein 41
1.05E-04	2.719	CTLA4	cytotoxic T-lymphocyte-associated protein 4

There are 6 common genes between the two top 10 lists, of these 6 genes, 4 (GALNT4, AIP, ZBTB22, and FBXO41) have been associated with chemotherapy resistance. Beyond these 4, there are 3 other relevant genes in table 7 with correlation cut-off at 0.5 (VAMP3, CALCOCO1, and NRG2) and 3 others in table 8 with correlation cut-off at 0.75 (EIF4G1, SH2D3C, and CTLA4) making it 7 out of 10 relevant genes in both lists.

5.2.3 Discussion

GALNT14 is listed first (with the highest tVIM) in both tables and there is no variation in its estimate indicting no additional genes are removed from its covariate set when reducing the correlation cut-off from 0.75 to 0.5. GALNT14 has been recently acknowledge as an informative biomarker for Apo2/TRAIL - based cancer therapy. The Apo2/TRAIL - based cancer therapy falls into the class of apoptosis activating therapies - therapies which activate or enforce programmed cell death. Apoptosis regulates cell number in normal tissues. When apoptosis is no longer active, the tissue is considered malignant. Alternatively anthracycline, a common drug used in adjuvant chemotherapy, inhibits the topoisomerase II - alpha religation reaction leading to cytotoxic cell damage and death; while the taxane class drugs (also common in adjuvant chemotherapy) inhibits cell division (Cortes-Funes and Coronado, 2007). A major benefit of the Apo2/TRAIL ligand is that it preferentially induces apoptosis in cancer cells over normal cells (Fesik, 2005). A recent study, Wagner et al. (2007), has shown that GALNT14 levels determine the sensitivity of tumor cells to apoptosis induced by Apo2L/TRAIL ligand. Increased expression of GALNT14 increases tumor cell response to this ligand making it a beneficial biomarker for sensitivity to Apo2/TRAIL - based cancer therapy. Among the patients in this study, exposed to adjuvant chemotherapy, we find GALNT14 up-regulated among the "bad-responders." Given the results of Wagner et al. (2007) this could indicate that a Apo2/TRAIL - based cancer therapy may have been more beneficial for these patients. In addition to GALNT14, our results indicate that AIP, which is also known to reduce apoptosis (Vogel et al., 2007; Berwick et al., 2004), is up-regulated among "bad responders" and has the second highest VIM values in both lists.

Beyond the apoptosis-related genes, we also see various indictors of drug resistance. ZBTB22 binds to Cul3 forming a complex in the Ubiquitin system and elevated Cul3 has been identified as an indicator of drug resistance (Zhang et al., 2004). The over-expression of EIF4G1 has been directly identified as an indictor of chemotherapy resistance (Wagner et al., 2007). SH2D3C interacts with BCAR and partially responsible for resistance to anti-estrogen therapy in breast cancer cells (Near et al., 2007). Our results indicate that all three are elevated in bad responders. In addition, CALCOCO1 has been identified as a potential target for cancer vaccines (Lee et al., 2003). Anitbodies of CTLA-4 activate anti-tumor response in breast cancer cells. - drugs targeting this mechanism are in clinical trails (Korman, 2003). NRG2 interacts with the Erbb family (including the HER-2 receptor) and induces cell growth among breast cancer cells (**?**). All three again are found elevated in bad responders in our analysis. Also, FBXO41 has been found to be significant and important in numerous other biomarker discovery analyses, including ours, as an indicator of good prognosis (Alexe et al., 2006). Another interesting, though confusing result is the elevated expression of VAMP3 among "good responders." Past research has identified VAMP3 as an

indicator of drug resistance (NAKAMURA et al., 2005). It's possible that the specific chemotherapy treatment chosen was correct for patients with elevated VAMP3. Specifics are unknown.

6 Conclusion

In both simulation and application we see the necessity for a standard method. Results vary widely leading to long lists and confusion, which list to use? In this paper we propose using tVIM as a standard method for biomarker discovery. In simulation it has proven resilient to increases in correlation, controlling type I error. It also provides an interpretable and meaningful measure of importance, which given an appropriate study design is interpretable as a causal effect. In comparison, common univariate linear regression is highly susceptible to increases in type I due to increased correlation. And though LASSO/LARS provides some improvement, using tMLE to update its estimate increases the accuracy in importance measure and rank and provides the correct asymptotic inference.

By targeting the causal effect, the measures obtained by tVIM are less sensitive to changes in the covariate distribution and therefore more reproducible in any population given it has the same conditional distribution of $Y|W$. For instance, this allows tVIM measures to be generalizable across microarray platforms that may have different noise levels. This reproducibility is essential for any standardized method, increasing confidence in diagnostic and treatment decisions based on these measures. In other words, if the causal effect between gene A and the response is correctly estimated in a population, it will be applicable to other populations. If instead we attribute the effect to gene B which is highly correlated to the causal gene A in the first population the correlation between gene B and gene A is not necessarily consistent in the other populations making the measure effect inapplicable in those populations. For instance if people in the second population have a cold, and gene B is related to immune response. It's levels may be much higher and no longer correlated in the same degree with the level of gene A. Making inferences on the disease state from the level of gene B erroneous.

Through the dichotomy of tMLE and double robust estimating equations, formal asymptotic inference for tVIM is available. Confidence Intervals and p-values are estimated using a covariance estimate derived from the influence curve. Multiple testing procedures based on the overall joint distribution are also possible without resampling but instead sampling directly from a multivariate normal with covariance estimated from the influence curve (ref?).

In application, tVIM provides and interpretable measure with interpretable inference. In the analysis of the Golub 1999 AML/ALL dataset, linear regression results in a list of 681 genes. Among those genes, there are ones that are biologically relevant and possibly causally related, however determining the relevant from the irrelevant is an impossible task given regression alone. RandomForest results are even more ambiguous; while they may have a high importance value, the directionality and the meaning of the value is unclear. Comparing among the top 10 for each method, tVIM finds more relevant genes which are related to AML./ALL.

The resulting top 10 tables from the Breast Cancer analysis are even more promising. Not only do we identify genes biologically related to chemotherapy resistance, we also identify genes which indicate a possible mechanism of treatment for "poor responders" based on up-to-date biological information. The relevance of the gene list supports the use of tVIM not only for biomarker discovery but also as a pre-screening method for prediction.

Applying a correlation cut-off in practice reduces the bias in the tVIM estimate due to potential ETA violations. However, the difference between the lists for tVIM correlation cut-off 0.5 and 0.75 affirm the need for a method which identifies the proper cut-off for a given gene. Having too low of a cut-off neglects controlling for the appropriate genes to achieve an estimate of the causal effect, decreasing its reproducibility across populations. Having to high a cut-off leads to ETA violations which increase bias in our importance estimate. In Bembom et al. (March 2008), they propose a method which analytically choses the cut-off for each variable given a binary A, tailoring the controlling variable set for each A. This method has been recently extended for a general (i.e. continuous) A (Bembom et al., March 2008) and will be implemented in the future. We also will explore methods which help piece apart or at the very least elucidate the relationship among a group of heavily correlated variables in relation to a response.

Targeted Variable Importance (tVIM) is a robust, locally efficient, and interpretable measure of importance with formal inference. It is simple to implement and understand. It is adaptable most data types including binary variables, survival outcome, and longitudinal data (van der Laan and Rubin, October 2006). It's accuracy, reproducibility, practically, and flexibility make it an ideal standardized method for biomarker discovery.

A Comparison of methods for variable importance

	tVIM	Univariate Linear Regression	LASSO - Penalized Multiple Regression	randomForest
Targeted Parameter	$\Psi(A,W) = \mathbb{E}[Y\|A,W] - \mathbb{E}[Y\|A=0,W]$	$Q_{LM}(A) = \mathbb{E}[Y\|A]$	$Q(A,W) = \mathbb{E}[Y\|A,W]$	$Q_{RF}(A,W) = \mathbb{E}Y\|A,W]$
The Model	$Q(A,W) = m(A,W\|\beta) + g(W)$ (where $m(A=0,W\|\beta)=0$) and $G(W) = \mathbb{E}[A\|W]$, in which either g(W) or G(W) is correctly specified	$Q_{LM}(A) = \beta A$	$Q(A,W) = \beta_A A + \beta W$	$Q_{RF}(A,W) = f(A,W)$, where $f(.)$ is an average over multiple regression tree models
Measure of Importance	$\mathbb{E}[\Psi(A,W)]$, or given the simple model $m(A,W\|\beta) = \beta A$, β can be viewed as the measure of importance	β	β_A, the coefficient associated with the variable of interest, A	Change in node split accuracy or overall error rate under perturbation of variable A
Biological Interpretation	Causal Effect of A on Y, under experiment controlling for confounders, W	Marginal Association	Causal Effect of A on Y, under experiment controlling for confounders, W	Unknown
Formal Inference for the Importance measure	Yes, based on estimating function methodology, derived from the Influence Curve	Yes	None provided by statistical R package *lars*	No
Control of Type I error	Using standard MTPs as well as Joint MTP based on Influence Curve	Standard MTPs	Not possible using package *lars*	None
Model Selection	Data Adaptive Algorithm of choice given $m(A=0,W\|\beta)=0$	None	Shrinkage of coefficients through penalized regression	Average over multiple regression trees formed using cross-validation
Sensitivity of Estimate to Model Selection	Double Robust, requires proper specification of either $Q(A,W)$ or $G(W)$ for consistent estimate. Locally efficient if both are correct.	None	Importance Measure of A is often set to zero	Variable A not necessarily in enough regression trees to allow for accurate importance calculation
Sensitivity to high dimensional data	Provides importance measure for each variable, estimating $Q(A,W)$ with data-adaptive software given $m(A=0,W\|\beta)=0$, controls for type I error using MTP	Importance Measure for each variable, type I error controlled using MTP	Some importance measures may be zero, often all variables placed in single model and only $n-1$ coefficients may be non-zero (n= no. of observations), no type I error control	Importance not necessarily available for all A, all variables in single model
Sensitivity to high correlation	Adjusts for confounding variables in W for each A. Under high correlation, can produce a realistic importance measure using correlation cut-off to control for ETA bias	No adjustment for confounding, suffers from increase Type I error rate under high correlation	Variable A often assigned zero importance due to high correlation with another variable	Uses regression tree framework, often discounting variable A

Table 9:

B Appendix - tMLE

Targeted Maximum Likelihood (tMLE) methodology maximizes the likelihood in a direction which targets the parameter of interest using the appropriate bias-variance trade-off (van der Laan and Rubin, October 2006).

We defined

$$\mu(a) = \mathbb{E}_{W^*}[m(A=a,W^*|\beta)]$$

with the estimate at a particular A=a defined as

$$\mu(a) = \frac{1}{n}\sum_{i=1}^{n}[m(a,W_i^*|\beta)]$$

where m(.) models the effect

$$m(a,W^*|\beta) = \mathbb{E}_P[Y|A=a,W^*] - \mathbb{E}_P[Y|A=0,W^*])$$

When A is binary, IPTW and DR-IPTW (van der Laan, 2005; van der Laan and Rubin, October 2006) methods may be used to estimate $\mu(A)$ without model assumptions. When A is more general, it requires specification of a model $m(A,W|\beta(P))$ that satisfies $m(A=0,W|\beta(P))=0$, where the true $\beta_0 = \beta(P_0)$.

Targeted MLE methodology creates a path through the true density p^0, represented as the hardest sub-model $p^0(\varepsilon)$. The hardest submodel $p^0(\varepsilon)$ is selected to only vary $Q(p)(Y|A,W)$, with score equal to $D_h(p_n^0)$ at $\varepsilon = 0$. This sub-model is explicitly derived in van der Laan and Rubin (October 2006).

Where $D_h(p_n^0)$ is the efficient influence curve as defined according to the following theorem

Theorem 1 *(From (Yu and van der Laan, September 2003)) For parameter $p \to \beta(p)$ in model $M = \{p : \mathbb{E}_p(Y|A,W) - \mathbb{E}_p(Y|A=0,W) = m(0,W|\beta(p))$, satisfying $m(0,W|\beta)=0$ for all $\beta \in \mathbb{R}^d$ the orthogonal complement of the nuisance tangent space is*

$$T_{nuis}^{\perp}(p) = \{D_h(p) : h\}$$

where

$$D_h(p)(O) \equiv \{h(A,W) - \mathbb{E}_p(h(A,W)|W)\}(Y - m(A,W|\beta(p)) - \mathbb{E}_p(Y|A=0,W))$$

The efficient influence curve or canonical gradient is then defined as

$$D_{h_{opt}}(p)(O) = \{h_{opt}(A,W) - \mathbb{E}_p(h_{opt}(A,W)|W)\}(Y - m(A,W|\beta(p)) - \mathbb{E}_p(Y|A=0,W))$$

where

$$h_{opt} = \frac{1}{\sigma(A,W)}\left\{\frac{d}{d\beta}m(A,W|\beta) - \frac{\mathbb{E}\left[\frac{1}{\sigma(A,W)}\frac{d}{d\beta}m(A,W|\beta)|W\right]}{\mathbb{E}\left[\frac{1}{\sigma(A,W)}|W\right]}\right\}$$

and $Var(Y|AW) = \sigma(A,W)$. If we assume $Var(Y|A,W) = Var(Y|W)$, then a more practical form of h_{opt} is available

$$h_{opt}^* = \frac{1}{\sigma(A,W)}\left\{\frac{d}{d\beta}m(A,W|\beta) - \mathbb{E}\left[\frac{d}{d\beta}m(A,W|\beta)|W\right]\right\}$$

where $G(W) = \mathbb{E}(A=a|W)$ and $Q(a,W) = \mathbb{E}_P(Y|A=a,W)$ are nuisance parameters. The double robust nature of the estimating function gives $\mathbb{E}_{P_0}D(O|\beta_0,Q,G) = 0$, providing a consistent estimate of β, if either of the nuisance parameters ($G(W)$ and $Q(A,W)$) is specified correctly.

Note when $\rho > 0$, the experimental treatment assumption (ETA) (i.e. $P(W_j|W_{-j}) = P(W_j)$) no longer holds. However due to the nature of the simulated data where all variables are simulated from a multivariate normal, the dependency can be accurately modeled using a main term linear model due to the simple correlation structure (i.e. $\mathbb{E}(W_j|W_{-j}) = \beta_W W_{-j}$).

Given an initial estimate of the density $p_n^0 = Q^0(A,W)$, and defining the hardest sub-model $p^0(\varepsilon|p_n^0)$, $p^0(\varepsilon|p_n^0)$ is maximized with respect to ε, substituting in the new estimate ε_n, the updated density $p^1 = p^0(\varepsilon_n|p_n^0)$, is the new targeted density.

In some cases iteration is necessary (substituting the new density estimate as initial density estimate and solving again for ε). By maximizing $p^0(\varepsilon|p_n^0)$ for ε, tMLE maximizes the likelihood in the direction of the parameter of interest μ, making the final density estimate the solution to $P_n(D_h(O)) = 0$ as well.

Assuming a normal distribution for $Q(p_n^0)(Y|A,W)$ with mean $Q(p_n^0)(A,W) = \mathbb{E}_{p_n^0}(Y|A,W)$ and variance $\sigma^2(Q_n^0)(A,W)$, the hardest sub-model which updates the original $Q(p_n^0)(Y|A,W)$ in a direction which estimates the parameter of interest well, can be defined as

$$Q(p)(\varepsilon)(Y|A,W) = \frac{1}{\sigma(A,W)} f_0\left(\frac{Y - m(A,W|\beta_n^0(\varepsilon)) - Q_n^0(\varepsilon)(W)}{\sigma(A,W)}\right)$$

where f_0 represents the standard normal density with updated parameters $\beta_n^0(\varepsilon) = \beta_n^0(Q_n^0) + \varepsilon$ and $\theta_n^0(\varepsilon) = \theta_n^0(Q_n^0) + \varepsilon^T r(W)$ where

$$r(p_n^0)(W) = \frac{\mathbb{E}\left[\frac{1}{\sigma(A,W)} \frac{d}{d\beta} m(A,W|\beta)|W\right]}{\mathbb{E}\left[\frac{1}{\sigma(A,W)}|W\right]}$$

if we assume, with only some loss in efficiency that $\sigma(A,W) = \sigma(W)$, then the above reduces to

$$r^*(p_n^0)(W) = \mathbb{E}\left[\frac{d}{d\beta} m(A,W|\beta)|W\right]$$

The proper form of r(W) shown above is found by equating the score of $Q(p)(\varepsilon)$ in terms of ε at $\varepsilon = 0$ to the efficient influence curve $D_{h_{opt}}(p_n^0)$.

The likelihood for $Q(p)(\varepsilon)$ can now be maximized for ε using standard weighted least squares, updating the initial estimates of β and θ, providing the new targeted estimate of the overall density as well as the parameter of interest β. Given a linear model $m(A,W|\beta)$ in β, a closed form solution for ε does exist and standard weighted linear regression software packages such as lm() in R may be used.

B.1 Inference and Testing

Asymptotically tMLE is equivalent to solving

$$\mathbb{E}_P(D_h(O|\beta,Q,G)) = 0$$

where $D_h(O|\beta,Q,G)$ is the efficient influence curve, making formal inference dependent on the influence curve still applicable to TMLE derived estimates.

The covariance matrix for β can be estimated using the conservative influence curve, The conservative influence curve is defined as,

$$IC(O) = \frac{D(O|\beta_0,\Pi,\theta)}{\mathbb{E}[\frac{d}{d\beta} D(O|\beta_0,\Pi,\theta)]}$$

where

$$\sqrt{n}(\beta_n - \beta_0) \sim N(0,\Sigma_n)$$

asymptotically with covariance equal to

$$\Sigma_n = \frac{1}{n}\sum \hat{IC}(O)\hat{IC}(O)^T$$

Covariance can also be estimated by bootstrap estimates of β, but this would requiring extra computational time. In this study, we know that $G(W^*)$ is correct, therefore estimates based on the influence curve are consistent.

Testing $H_0 : \beta_0(j) = 0$, p-values can be determined using test statistic

$$T_n(j) = \frac{\sqrt{n}\beta_n(j)}{sqrt\Sigma_n(j,j)} \underset{n\to\infty}{\sim} N(0,1)$$

Testing for significance of the marginal variable importance curve when the effect of A is modified by W_i (i.e $\mathbb{E}[m(A = a, W = a|\beta)] = \beta_0 a + \beta a \mathbb{E}[W_i]$) is completed by testing the null hypothesis $H_0 : c^T \beta_0 = 0$, where c is the appropriate vector of A and W corresponding to m(.). Test statistic becomes $T_n(j) = \frac{\sqrt{n} c^T \beta_n(j)}{\sqrt{c^T \Sigma_n(j,j) c}}$ which is asymptotically distributed N(0,1).

References

Gabriela Alexe, Sorin Alexe, David E Axelrod, Tibérius O Bonates, Irina I Lozina, Michael Reiss, and Peter L Hammer. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research*, 8(4):R41, 2006.

E. Balciunaite, A. Spektor, N.H. Lents, H. Cam, H. Te Riele, A. Scime, M.A. Rudnicki, R. Young, and B.D. Dynlacht. Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells. *Mol Cell Biol.*, 25(18): 8166–78, 2005.

Oliver Bembom, Jeffrey W. Fessel, Robert W. Shafer, and Mark J. van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. Technical Report Working Paper 231, U.C. Berkeley Division of Biostatistics Working Paper Series, March 2008. URL `http://www.bepress.com/ucbbiostat/paper231`.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, 57:289–300, 1995.

Marianne Berwick, Giuseppe Matullo, Yan Shuang Song, Simonetta Guarrera, Gemma Dominguez, Irene Orlow, Mary Walker, and Paolo Vineis. Association between aryl hydrocarbon receptor genotype and survival in soft tissue sarcoma. *Journal of Clinical Oncology*, 22(19):3997–4001, 2004.

L. Breiman. Two-eyed algorithms and problems, 2003. ISSN 0302-9743.

L. Breiman, J. H. Freidman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL `http://dx.doi.org/10.1023/A:1010933404324`.

Vince Carey. Roc. R package.

E. Clappier, W. Cuccuini, A. Kalota, A. Crinquette, J.M. Cayuela, W.A. Dik, A.W. Langerak, B. Montpellier, B. Nadel, P. Walrafen, O. Delattre, A. Aurias, T. Leblanc, H. Dombret, A.M. Gewirtz, A. Baruchel, F. Sigaux, and J. Soulier. The c-myb locus is involved in chromosomal translocation and genomic duplications in human t-cell acute leukemia (t-all), the translocation defining a new t-all subtype in very young children. *Blood*, 110(4):1251–61, 2007.

Hernan Cortes-Funes and Cyntia Coronado. Role of anthracyclines in the era of targeted therapy. *Cardiovasc Toxicol*, 7: 56–60, 2007.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

Brad Efron and Trevor Hastie. lars. R package.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics (with discussion)*, 32(2):407–499, 2004.

Stephen W. Fesik. Promoting apoptosis as a strategy for cancer drug discovery. *Nature Reviews Cancer*, 5:876–885, 2005.

Peter A. Flach. Tutorial on "the many faces of roc analysis in machine learning". In *The Twenty-First International Conference on Machine Learning*, 2004.

D. A. Fruman, S. B. Snapper, C. M. Yballe, L. Davidson, J. Y. Yu, F. W. Alt, and L. C. Cantley. Impaired b cell development and proliferation in absence of phosphoinositide 3-kinase p85-alpha. *Science*, 283:393–397, 1999.

Tsutomu Fujii, Shuji Nomoto, Katsumi Koshikawa, Yasushi Yatabe, Osamu Teshigawara, Toshiaki Mori, Soichiro Inoue, Shin Takeda, and Akimasa Nakao. Overexpression of pituitary tumor transforming gene 1 in hcc is associated with angiogenesis and poor prognosis. *Hepatology*, 43:1267–1275, 2006.

T.R. Golub, D.K. Slonim, P Tamayo, C Huard, M Gaasenbeek, J.P. Mesirov, H Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(531-537), 1999.

Stephen P. Hunger. Chromosomal translocations involving the e2a gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood*, 87:1211–1224,, 1996.

S. H. Kaufmann, S. D. Gore, C. B. Miller, R. J. Jones, L. A. Zwelling, E. Schneider, P. J. Burke, and J. E. Karp. Topoisomerase ii and the response to antileukemic therapy. *LEUKEMIA LYMPHOMA*, 29(3-4):217–237, Apr 1998a. ISSN 1042-8194.

Scott H. Kaufmann, Judith E. Karp, Phyllis A. Svingen, Stan Krajewski, Philip J. Burke, Steven D. Gore, and John C. Reed. Elevated expression of the apoptotic regulator mcl-1 at the time of leukemic relapse. *Blood*, 91(3):991–1000, 1998b. URL http://bloodjournal.hematologylibrary.org/cgi/content/abstract/bloodjournal;91/3/991.

Iwao Komuro, Tomoyoshi Yasuda, Aikichi Iwamoto, and Kiyoko S. Akagawa. Catalase plays a critical role in the csf-independent survival of human macrophages via regulation of the expression of bcl-2 family. *J Biol Chem.*, 280(50): 41137–45, 2005.

Charles Kooperberg, Smarajit Bose, and Charles J. Ston. Polychotomous regression. *Journal of the American Statistical Association*, 92:117–127, 1997.

A Korman. Ctla-4 based therapy (mdx-010). *Breast Cancer Res*, 5(Suppl 1):63, 2003.

J. Kos and TT. Lah. Cysteine proteinases and their endogenous inhibitors: target proteins for prognosis, diagnosis and therapy in cancer (review). *Oncol Rep.*, 5(6):1349–61, 1998.

K. Kozar, M. A. Ciemerych, V. I. Rebel, H. Shigematsu, A. Zagozdzon, E. Sicinska, Y. Geng, Q. Yu, S. Bhattacharya, R. T. Bronson, K. Akashi, and P. Sicinski. Mouse development and cell proliferation in the absence of d-cyclins. *Cell*, 118: 477–491, 2004.

A. A. Lane and T. J. Ley. Neutrophil elastase cleaves pml-rar-alpha and is important for the development of acute promyelocytic leukemia in mice. *Cell*, 115(305-318), 2003.

S.Y. Lee, Y. Obata, M. Yoshida, E. Stockert, B. Williamson, A.A. Jungbluth, Y.T. Chen, L.J. Old, and M.J. Scanlan. Immunomic analysis of human sarcoma. *Proc Natl Acad Sci U S A.*, 100(5):2651–6, 2003.

G. Leone, R. Sears, E. Huang, R. Rempel, F. Nuckolls, C.-H. Park, P. Giangrande, L. Wu, H. I. Saavedra, S. J. Field, M. A. Thompson, H. Yang, Y. Fujiwara, M. E. Greenberg, S. Orkin, C. Smith, and J. R. Nevins. Myc requires distinct e2f activities to induce s phase and apoptosis. *Molec. Cell*, 8:105–113, 2001.

Andy Liaw and Matthew Wiener. randomforest. R package.

J I Martin-Subero, R Ibbotson, W Klapper, L Michaux, E Callet-Bauchu, F Berger, M J Calasanz, C De Wolf-Peeters, M J Dyer, P Felman, A Gardiner, R D Gascoyne, S Gesk, L Harder, D E Horsman, M Kneba, R Kuppers, A Majid, N Parry-Jones, M Ritgen, M Salido, F Sole, G Thiel, H-H Wacker, D Oscier, I Wlodarska, and R Siebert. A comprehensive genetic and histopathologic analysis identifies two subgroups of b-cell malignancies carrying a t(14;19)(q32;q13) or variant bcl3-translocation. *Leukemia*, 21(7):1532–1544, 2007. URL http://dx.doi.org/10.1038/sj.leu.2404695.

Yusuke NAKAMURA, Toyomasa KATAGIRI, and Shuichi NAKATSURU. (wo/2005/028676) method of diagnosing breast cancer. Patent, ONCOTHERAPY SCIENCE, INC., 2005. URL http://www.wipo.int/pctdb/en/wo.jsp?ELEMENT_SET=F&LANGUAGE=ENG&KEY=05%2F028676&IA=05%2F028676&DISPLAY=STATUS.

Richard I. Near, Yujun Zhang, Anthony Makkinje, Pierre Vanden Borre, and Adam Lerner. And-34/bcar3 differs from other nsp homologs in induction of anti-estrogen resistance, cyclin d1 promoter activation and altered breast cancer cell morphology. *J Cell Physiol.*, 212(3):655–65, 2007.

Karin Nebral, Helmut H. Schmidt, Oskar A. Haas, and Sabine Strehl. NUP98 Is Fused to Topoisomerase (DNA) IIbeta 180 kDa (TOP2B) in a Patient with Acute Myeloid Leukemia with a New t(3;11)(p24;p15). *Clin Cancer Res*, 11(18): 6489–6494, 2005. doi: 10.1158/1078-0432.CCR-05-0150. URL http://clincancerres.aacrjournals.org/cgi/content/abstract/11/18/6489.

Martin O'Connor. polymars. R package polspline.

Teresa Palomero, Duncan T. Odom, Jennifer O'Neil, Adolfo A. Ferrando, Adam Margolin, Donna S. Neuberg, Stuart S. Winter, Richard S. Larson, Wei Li, X. Shirley Liu, Richard A. Young, and A. Thomas Look. Transcriptional regulatory networks downstream of tal1/scl in t-cell acute lymphoblastic leukemia. *Blood*, 108(3):986–992, 2006.

K.S. Pollard, S. Dudoit, and M.J. van der Laan. *Multiple Testing Procedures: R multtest Package and Applications to Genomics in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Number 209-229. Springer (Statistics for Biology and Health Series), 2005.

R. Puri, A. Tousson, L. Chen, and Kakar S.S. Molecular cloning of pituitary tumor transforming gene 1 from ovarian tumors and its expression in tumors. *Cancer Lett.*, 163:131–139, 2001.

Sridhar Ramaswamy, Ken N. Ross, Eric S. Lander, and Todd R. Golub. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, 33:49–54, 2003.

J.M. Robins and A. Rotnitzky. Comment on the bickel and kwon article "inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

J.M. Robins, S.D. Mark, and W.K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(479-495), 1992.

Sema Senturker, Bensu Karahalil, Mine Inal, Hulya Yilmaz, Hamza Muslumanoglu, Gunduz Gedikoglu, and Miral Dizdaroglu. Oxidative dna base damage and antioxidant enzyme levels in childhood acute lymphoblastic leukemia. *FEBS Letters*, 416(3):286–290, 1997. URL http://www.sciencedirect.com/science/article/B6T36-3RD0S7F-2B/2/0ad89f9393a1d3afd2e750bac2bd3ab5.

Sandra E. Sinisi and Mark J. van der Laan. Loss-based cross-validated deletion/substitution/addition algorithms in estimation. Working paper 143, U.C. Berkeley Division of Biostatistics Working Paper Series, March 2004. URL http://www.bepress.com/ucbbiostat/paper143.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc B.*, 58(1):267–288, 1996.

G. W. Tombes, R. M.; Krystal. Identification of novel human tumor cell-specific camk-ii variants. *Biochim. Biophys. Acta*, 1355:281–292, 1997.

Critical Path Initiative Fact Sheet. U.S. Department of Health and Human Services and U.S. Food and Drug Administration, Janurary 2007. URL http://www.fda.gov/oc/initiatives/criticalpath/factsheet.html.

Critical Path Opportunities Report. U.S. Department of Health and Human Services and U.S. Food and Drug Administration, http://www.fda.gov/oc/initiatives/criticalpath/reports/opp$_r eport.pdf edition, March$2006.

Mark J. van der Laan. Statistical inference for variable importance. Technical Report Working Paper 188, U.C. Berkeley Division of Biostatistics Working Paper Series, 2005. URL http://www.bepress.com/ucbbiostat/paper188.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. Working paper 213, U.C. Berkeley Division of Biostatistics Working Paper Series, October 2006. URL http://www.bepress.com/ucbbiostat/paper213.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "super learner". Technical Report Working Paper 222, U.C. Berkeley Division of Biostatistics Working Paper Series, July 2007. URL http://www.bepress.com/ucbbiostat/paper222.

LJ van 't Veer, H Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M Mao, H.L. Peterse, K van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

C. Vitale, C. Romagnani, A. Puccetti, D. Olive, R. Costello, L. Chiossone, A. Pitto, A. Bacigalupo, L. Moretta, and M.C. Mingari. Surface expression and function of p75/airm-1 or cd33 in acute myeloid leukemias: engagement of cd33 induces apoptosis of leukemic cells. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5764–5769, 2001.

C.F. Vogel, W. Li, E. Sciullo, J. Newman, B. Hammock, J.R. Reader, J. Tuscano, and F. Matsumura. Pathogenesis of aryl hydrocarbon receptor-mediated development of lymphoma is associated with increased cyclooxygenase-2 expression. *Am J Pathol.*, 171(5):1538–48, 2007.

Klaus W Wagner, Elizabeth A Punnoose, Thomas Januario, David A Lawrence, Robert M Pitti, Kate Lancaster, Dori Lee, Melissa von Goetz, Sharon Fong Yee, Klara Totpal, Ling Huw, Viswanatham Katta, Guy Cavet, Sarah G Hymowitz, Lukas Amler, and Avi Ashkenazi. Death-receptor o-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand apo2l/trail. *Nat Med*, 13(9):1070–1077, 2007. URL http://dx.doi.org/10.1038/nm1627.

Yue Wang, Maya L. Petersen, David Bangsberg, and Mark J. van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Technical Report Working Paper 211, U.C. Berkeley Division of Biostatistics Working Paper Series., September 2006. URL http://www.bepress.com/ucbbiostat/paper211.

Zhuo Yu and Mark J. van der Laan. Measuring treatment effects using semiparametric models. Technical Report Working Paper 136, U.C. Berkeley Division of Biostatistics Working Paper Series, September 2003. URL http://www.bepress.com/ucbbiostat/paper136.

H.F. Zhang, A. Tomida, R. Koshimizu, Y. Ogiso, S. Lei, and T. Tsuruo. Cullin 3 promotes proteasomal degradation of the topoisomerase i-dna covalent complex. *Cancer Res.*, 64(3):1114–21, 2004.

X Zhu, Z Mao, Y Na, Y Guo, X Wang, and D Xin. Significance of pituitary tumor transforming gene 1 (pttg1) in prostate cancer. *Anticancer Res*, 26:1253–1259, 2006.