Chapter 22
Targeted Methods for Biomarker Discovery

Catherine Tuglus, Mark J. van der Laan

The use of biomarkers in disease diagnosis and treatment has grown rapidly in recent years, as microarray and sequencing technologies capable of detecting biological signatures have become more effective research tools. In an attempt to create a level of quality assurance with respect to biological and more specifically biomarker research, the FDA has called for the development of a standard protocol for biomarker qualification (Food and Drug Administration 2006). Such a protocol would define "evidentiary" standards for biomarker usage in areas of drug development and disease treatment and provide a standardized assessment of a biomarker's significance and biological interpretation. This is especially relevant for RCTs, where the protocol would prohibit the use of unauthenticated biomarkers to determine treatment regime, resulting in safer and more reliable treatment decisions (Food and Drug Administration 2006). Consequentially, identifying accurate and flexible analysis tools to assess biomarker importance is essential. In this chapter, we present a measure of variable importance based on a flexible semiparametric model as a standardized measure for biomarker importance. We estimate this measure with the TMLE.

Many biomarker discovery methods only measure the association between the marker 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. This is especially true with genomic data, where genes are often present in multiple pathways and can be highly correlated amongst themselves. Applying association-based methods to these data will often lead to a long and ambiguous listing of biomarkers, which can be expensive to analyze.

Ideally, biomarker discovery analyses should identify markers that systematically affect the outcome through a biological pathway or mechanism, in other words, markers causally related to the outcome of interest. Once these markers are identified, they can be further analyzed and eventually applied as potential drug targets or

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_22,

prognostic markers. Due to the complex nature of the human genome, this is not a straightforward task, and certain assumptions are required to identify a causal effect.

In general, causal effects are often difficult if not impossible to estimate correctly, especially based on high-dimensional and highly correlated genomic data structures. The required identifiability assumptions such as the time-ordering assumption, the randomization assumption, and the positivity assumption are often only fully realized in RCTs, making their utility in a standard protocol limited. However, measures that are causally interpretable in RCTs can still be biologically interpretable based on observational data as measures of importance.

Here, we present the typical representation of a causal effect as a potential measure of importance for a biomarker A:

$$\Psi(P_0)(a) = E_0[E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W)].$$

Given the observed data structure $O = (W, A, Y) \sim P_0$, this measure corresponds to the effect of a biomarker (A) on the outcome (Y), adjusting for confounders (W). Here, A can represent a single biomarker or set of biomarkers. This chapter will focus on the univariate case. This measure can be estimated in semiparametric models for P_0, and with formal inference, using the TMLE.

In this chapter, we present the TMLE of the variable importance measure (VIM) above under a semiparametric regression model, which can accommodate continuous treatment or exposure variables often seen in biomarker analyses. We will primarily focus on its application to biomarker discovery. However, this method also has important applications to clinical trial data when the treatment is binary or continuous, and when one wishes to test for possible effect modification by baseline variables, for instance, treatment modified by biomarkers measured at baseline.

We demonstrate the efficacy and functionality of this VIM and its TMLE in a simulation study. The simulations provide a performance assessment of our estimated measure under increasing levels of correlation of A with W. We show the accuracy with which the TMLE of the VIM can detect "true" variables from amongst increasingly correlated "decoy" variables. Additionally, we also evaluate the accuracy of three commonly used methods for biomarker discovery under the same conditions: univariate linear regression, lasso regression (Efron et al. 2004), and random forest (Breiman 1999, 2001a). We also apply the method in an application to a leukemia data set (Golub et al. 1999).

22.1 Semiparametric-Model-Based Variable Importance

Previous chapters have focused on the TMLE of the above VIM in a nonparametric model for variables A that are discrete; for instance, A might be an indicator for

receiving a particular treatment or exposure. However, particularly in the worlds of genomics and epidemiology, the variable of interest is often continuous. In this chapter, we present a semiparametric-regression-model-based measure of variable importance that is flexible enough to accommodate the typical data structures in genomics, epidemiology, and medical studies.

This VIM was proposed in van der Laan (2006) and estimated with the TMLE in van der Laan and Rubin (2006). These semiparametric regression models have been considered in Robins et al. (1992), Robins and Rotnitzky (2001), and Yu and van der Laan (2003). Under the semiparametric regression model, only the effect of A on the mean outcome of Y needs to be modeled with a parametric form, while the remainder of the conditional mean of the outcome Y remains unspecified. The semiparametric nature can accommodate both continuous and binary variables of interest as well as incorporate effect modification by W in a straightforward and interpretable manner.

We assume

$$E_0(Y \mid A, W) = m(A, V|\beta_0) + r(W),$$

for a specified parametric model $\{m(A, V \mid \beta) : \beta\}$ that satisfies $m(0, V \mid \beta) = 0$ for all β, and unspecified function $r(W)$. Here V is a user-supplied set of effect modifiers contained in the covariate vector W. This is equivalent to assuming $E_0(Y \mid A = a, W) - E_0(Y \mid A = 0, W) = m(a, V \mid \beta_0)$. For our purposes, we assume a linear form for $m(A, V \mid \beta_0)$, which puts this model in the class of partial linear regression models. Given this semiparametric form with user-supplied $m(.)$, the marginal variable importance of a particular A is defined generally as

$$\mu_0(a) = E_{W,0}(m(a, V \mid \beta_0)).$$

However, it is important to remember that the maximum likelihood estimator is developed under the assumption that $m()$ is correct.

Given an estimator β_n of β_0, an estimate of this parameter of interest at a particular $A = a$ is defined as

$$\mu_n(a) = \frac{1}{n} \sum_{i=1}^{n} (m(a, V_i \mid \beta_n)).$$

If we assume a linear model $m(A, V \mid \beta) = A\beta^\top V$, the variable importance can be represented as a linear curve at the level a for the biomarker given by $E_{W,0}(m(A = a, V \mid \beta_0)) = a\beta_0^\top E_0 V$. Thus, given this linear model, the VIM is identified by a simple linear combination (i.e., $\beta_0^\top E_0 V$), and formal statistical inference is obtained with a straightforward application of the delta method. Further details are provided in Tuglus and van der Laan (2008). Here, we focus on the simplest linear case $m(A, V \mid \beta) = A\beta$, where the marginal importance of A can be represented by single coefficient value β.

22.2 The TMLE

The data structure is $O = (W, A, Y) \sim P_0$ and the statistical model $\mathcal{M}$ consists of all probability distributions P for which $\bar{Q}(P)(A, W) = E_P(Y \mid A, W)$ is of the form $A\beta + E_P(Y \mid A = 0, W)$ for some β. The target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined as $\Psi(P) = \beta(P)$, the β-coefficient in front of A. This target parameter has an efficient influence curve at P that is given by $D^*(P)(O) = 1/\sigma^2(A, W)H^*(g(P))(A, W)(Y - \bar{Q}(P)(A, W))$, where

$$H^*(g(P))(A, W) = \frac{d}{d\beta} m(A, V \mid \beta) - \frac{E_P(\frac{d}{d\beta} m(A, V \mid \beta)/\sigma^2(A, W) \mid W)}{E_P(1/\sigma^2(A, W) \mid W)},$$

$g(P)(A \mid W) = P(A \mid W)$, and $\sigma^2(A, W)$ is the conditional variance of Y, given (A, W). It is of interest to note that σ^2 cancels out in the efficient influence curve *if* $\sigma^2(A, W) = \sigma^2(W)$ is only a function of W.

The TMLE requires selecting a loss function for $\bar{Q}$, $L(\bar{Q})$, and a submodel $\{\bar{Q}_g(\epsilon) : \epsilon\} \subset \mathcal{M}$ through $\bar{Q}$ at $\epsilon = 0$, so that the linear span of $d/d\epsilon\, L(\bar{Q}_g(\epsilon))$ at $\epsilon = 0$ includes this efficient influence curve $1/\sigma^2 H^*(g)(Y - \bar{Q}(A, W))$. We select the squared error loss function, $L(\bar{Q})(O) = (Y - \bar{Q}(A, W))^2/\sigma^2(A, W)$, and the univariate linear regression submodel, $\bar{Q}_g(\epsilon) = \bar{Q} + \epsilon H^*(g)$ with "clever covariate" $H^*(g)$. The TMLE is now defined as usual and the iterative TMLE algorithm converges in a single step.

For the sake of implementation, we restrict ourselves to the choice $\sigma^2 = 1$ (that is, we are estimating the nuisance parameter σ^2 with the trivial constant 1). For this choice $\sigma^2 = 1$, we have that the clever covariate $H^*(g) = (A - E_g(A \mid W))$ only depends on g through the conditional mean of A, given W. The consistency of the TMLE does not rely on σ^2 since the efficient influence curve is an unbiased estimating function in β for each choice of σ^2: the TMLE is consistent if either $\bar{Q}_0$ or g_0 is estimated consistently. However, as a consequence of not estimating σ^2, even if both $\bar{Q}_0$ and g_0 are consistently estimated, the TMLE will only be efficient if $\sigma_0^2(A, W)$ is only a function of W.

Implementation. In biomarker discovery analyses, one is interested in assessing the VIM for a whole collection of biomarkers. For each biomarker, one defines a corresponding adjustment set (e.g., all other biomarkers). We outline the TMLE implementation below for a single biomarker A and corresponding adjustment set W. There are three initial components necessary for applying TMLE to estimate the parameter of interest.

1. A model $m(A, V \mid \beta)$ satisfying $m(0, V \mid \beta) = 0$ for all β and V. In this case, it is defined as $m(A, V \mid \beta) = \beta A$.
2. An initial regression estimate of $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ of the form $\bar{Q}_n^0(A, W) = m(A, V \mid \beta_n^0) + r_n^0(W)$. The initial regression estimate of the

proper form may be obtained from semiparametric regression methods such as those of Hastie and Tibshirani (1990), among others, or by using methods like DSA, which allow the user to fix a portion of the regression model. However, we recommend a more flexible approach that allows one to use a wider range of data-adaptive software. This approach is outlined as follows. (i) Obtain an initial regression estimate of $\bar{Q}_0(A, W)$ of general form using data-adaptive machine learning algorithms such as the super learner, (ii) evaluate $r_n^0(W) = \bar{Q}_n^0(A = 0, W)$, and (iii) determine the least squares regression estimate β_n^0 for the linear regression working model $\bar{Q}_n^0(A, W) = m(A, W \mid \beta) + \alpha \bar{Q}_n^0(A = 0, W) + \text{error}$, treating $\bar{Q}_n^0(A = 0, W)$ as a covariate, and α and β as unknown coefficients.

3. An estimate of the conditional mean $\bar{g}_0(W) = E_0(A \mid W)$.

Given the obtained initial estimator $\bar{Q}_n^0$ of the correct form, the TMLE is now obtained as follows:

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

$$H^*(A, W) = \frac{d}{d\beta} m(A, V \mid \beta) - E\left(\frac{d}{d\beta} m(A, V \mid \beta) | W\right),$$

   which for this particular form $m(A, V \mid \beta) = \beta A$ simplifies to $H^*(\bar{g})(A, W) = A - \bar{g}(W)$, where $\bar{g}(W) = E_g(A \mid W)$. Let $\bar{g}_n$ be the estimator of the true conditonal mean $\bar{g}_0$ of A, given W.
2. Use least squares regression to regress the outcome Y onto the clever covariate $H^*(\bar{g}_n)(A, W)$ using $\bar{Q}_n^0(A, W)$ as offset, and define the resulting coefficient in front of the clever covariate as ϵ_n.
3. Update the initial estimate $\beta_n^1 = \beta_n^0 + \epsilon_n$, $r_n^1 = r_n^0 - \epsilon_n \bar{g}_n$, and thereby the corresponding regression $\bar{Q}_n^1(A, W) = \bar{Q}_n^0(A, W) + \epsilon_n H^*(\bar{g}_n)(A, W)$. Iteration of this updating procedure does not result in further updates of $\bar{Q}_n^1$. As a consequence, the TMLE of $\bar{Q}_0$ is given by $\bar{Q}_n^* = \bar{Q}_n^1$.

Statistical inference. The TMLE $\bar{Q}_n^* = (\beta_n^*, r_n^*)$ solves the estimating equation $0 = P_n D^*(\bar{Q}_n^*, \bar{g}_n)$, where $D^*(\bar{Q}_n^*, \bar{g}_n)(W, A, Y) = H^*(\bar{g}_n)(W, A)(Y - m(A, V \mid \beta_n^*) - r_n^*(W))$. We can represent the efficient influence curve $D^*(\bar{Q}_n^*, \bar{g}_n) = D(\beta_n^*, r_n^*, \bar{g}_n)$ as an estimating function for β. As a consequence, if one is willing to assume that $\bar{g}_n$ is consistent, then statistical inference can be based on the conservative influence curve $IC(O) = -c^{-1} D(\beta_0, r^*, \bar{g}_0)$ with scale factor $c = E_0\left(d/d\beta_0 D(\beta_0, r^*, \bar{g}_0)\right)$, and r^* represents the possibly misspecified limit of r_n^*.

The asymptotic covariance of $\sqrt{n}(\beta_n^* - \beta_0)$ can be estimated with the empirical estimate of the covariance matrix of this influence curve: $\Sigma_n = \frac{1}{n}\sum_{i=1}^n IC_n(O_i)IC_n(O_i)^\top$. For the sake of statistical inference we can use as working model $\sqrt{n}(\beta_n^* - \beta_0) \sim N(0, \Sigma_n)$. For example, one may test the null hypothesis $H_0 : \beta_0(j) = 0$, using a standard test statistic $T_n = (\sqrt{n}\beta_n^*(j))/\sqrt{\Sigma_n(j,j)}$, which is asymptotically $N(0,1)$ under the null hypothesis. Similarly, a multiple testing methodology can be applied based on the influence curves of the biomarker-specific variable importance estimator across a large collection of biomarkers. Statistical inference can also be based on the bootstrap, but in high-dimensional biomarker analyses it is important to have a computational friendly method available as well.

22.3 Variable Importance Methods

In this section we compare TMLE 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 2007), and random forest (Breiman 1999, 2001a) using R package randomForest (Liaw and Wiener 2002).

For each component of the covariate vector, using each of the methods, we assess the variable importance of this component, controlling for all other variables. For the univariate regression and TMLE methods that report p-values we may adjust for multiple testing using Benjamini–Hochberg step-up FDR-controlling procedure (Benjamini and Hochberg 1995) implemented with the mt.rawp2adjp() R function in package multtest (Ge and Dudoit 2002), and thereby classify the biomarkers as important or not accordingly. However, since the lasso method and random forest method do not allow for cutoffs based on valid p-values, we will focus on comparing the method-specific ranked lists, ranked by VIM or p-value when available.

Univariate linear regression (lm). Marginal variable importance is represented by the coefficient and p-value resulting from the univariate linear regression fit, $E_n(Y \mid A) = \beta_n A$. P-values are calculated using a standard t-test. This method does not account for any confounding and will often misclassify biomarkers correlated with the "true" biomarkers as significant.

Penalized regression (lasso). Marginal variable importance is represented by the coefficient of A in a lasso regression main term fit of $\bar{Q}_0(A, W_s) = E_0(Y \mid A, W_s)$, with $W_s \subset W$ representing the subset of W found significant according to their univariate regression on Y at p-value cutoff $\alpha = 0.05$. Lasso does not provide any formal statistical inference, therefore, p-values are not recorded. Lasso does attempt to account for confounding, but will only allow for main term linear regression fits with maximally $n-1$ nonzero coefficient values, making its applicability to high-dimensional data limited (Tibshirani 1996). Lasso is also a maximum likelihood method that focuses on estimating the overall regression $E_0(Y \mid A, W)$, and not the parameter of interest.

TMLE. The VIM is obtained by applying a TMLE to the initial regression estimator provided by the lasso fit of $\bar{Q}_0(A, W_s)$. We estimate $\bar{g}_0(W) = E_0(A \mid W)$ using lasso regression as well. *P*-values are calculated using a standard *t*-test.

Random forest (RF1, RF2). Random forest is a tree-based algorithm commonly used in biomarker discovery analyses, though it does not estimate the same measure as lm, lasso, or TMLE. Due to the nature of random forest, there is no guarantee that all biomarkers will receive a measure of importance. Also, as with lasso, no formal statistical inference is available. Two measures of importance, RF1 and RF2, are provided by the R function randomForest(), and we used the default setting with 500 trees. Random forest provides two measures of importance based on the perturbing effect the variable of interest has on overall classification error and node splits. The first, denoted RF1, is based on an"out-of-bag" error rate, and the second, RF2, is based on the accuracy of the node split (both with no *p*-values provided) (Breiman 1999, 2001a; Liaw and Wiener 2002).

22.4 Simulations

We simulate data to compare the four approaches for variable importance analysis under increasing correlation levels among the biomarkers, using a diagonal block correlation structure. The structure of the simulated data allows us to study the effects that both correlated and uncorrelated variables have on the reported importance of the true variables. 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 cutoff values, and will be summarized using ROC plots. We will also determine the ability of each approach to identify the true variables and each variables true importance rank by comparing the length of list required to label all true variables as "important."

The data structure is defined as $O = (W^*, Y) \sim P_0$, with a 100-dimensional covariate vector W^* and univariate outcome Y. The sample size is set at $n = 300$. The covariate vector W^* is 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 clusters of 10 variables, each variable having unit variance, and any pair of variables within the cluster has correlation ρ_{TRUE}. The outcome Y is simulated from a main effect linear model using one variable from each of the K clusters. These K variables are designated as "true variables." The importance of a variable is determined by its coefficient value in the linear regression of Y. 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^2 is added as noise. 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 $\sigma_Y^2 = 10$, using both sets of coefficient values.

For each setting of $\{\rho, \sigma_Y\}$ we simulated 100 data sets of size $n = 300$. The recorded importance measures and p-values are translated into a list of ranks, and the ranks are averaged over the 100 iterations. A rank of one is the largest importance value or smallest p-value. Sensitivity and specificity calculations for each simulation are also determined for each simulated data set and averaged across the 100 simulated data sets to produce the final estimates. Simulation results are summarized here in terms of area under the curve (AUC) and length of list.

AUC. The overall performance of a ranked list is often summarized in terms of the AUC, the area under the curve derived from the basic ROC curve, which plots the true positive rate (sensitivity) by the false positive rate (1-specificity) as a function of the cutoff for the list.Under pure noise conditions the AUC= 0.5, indicating that at any threshold the false positive and true positive rates are equal (random classifier). The more convex the curve becomes, the higher the AUC, and the better the ranked list, and a perfect ranked list will have AUC = 1. The calculated AUC values are plotted vs. correlation for each of the five methods using importance measure importance rank and p-values when available for correlations, $\rho_{TRUE} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. From Fig. 22.1 we can see that

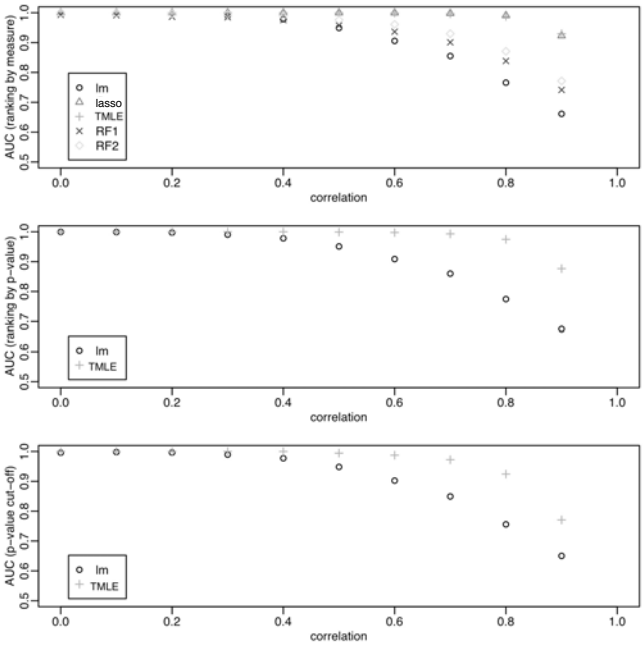


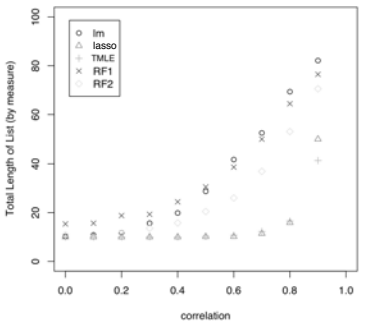
Fig. 22.1 AUC value from ROC curves by pairwise correlation $\rho = 0, ..., .9$, completed for ranking by measure (top), ranking by p-value (middle), and ranking by p-value using p-value cutoff, where $\sigma_Y = 10$, $n = 300$. Plots are shown for constant $\beta = 4$, but results are comparable when $\beta = \{1, ..., 10\}$

the TMLE performs well up to $\rho = 0.6$, performing only marginally better than lasso for $\rho > 0.2$, but with AUC visibly greater than random forest and lm as the correlation increases. As expected, lm is most susceptible to increases in correlation, performing perfectly when the correlation between biomarkers equals zero but failing consistently as the correlation increases, reaching below 0.8 by $\rho = 0.5$.

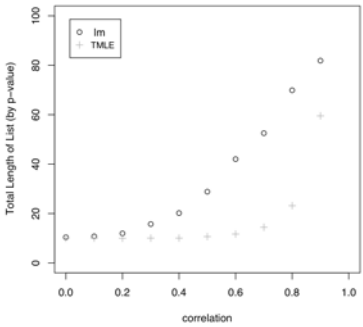
Average length of list. We can also compare the method-specific ranked lists of biomarkers based on the average cutoff for the list required to capture all "true" variables. Having a short cutoff allows the biologist to spend money analyzing the top genes with confidence, knowing that the most important genes are at the top of the list. The average required list length to find all ten "true" variables is plotted vs. correlation for all five measures and two p-value average ranked lists. These plots are shown for both constants $\beta_{true} = 4$ and $\beta_{true} = \{1 \ldots 10\}$. More detailed required length of lists for capturing the top k true variables, $k = 1, ..., 10$, for each available ranked list (rank by measure, rank by p-value) at each correlation level, as well as plots of the average rank and importance value, can be found in Tuglus and van der Laan (2008).

Length of list is a direct reflection of the type I error or false discovery rate associated with different cutoffs for the ranked lists of variables. Overall, the TMLE performs well up to correlations of 0.9, though the improvement over lasso is less clear when β_{TRUE} is constant (Figs. 22.2(a) and 22.2(b)). In the case where $\beta_{TRUE} = \{1, ..., 10\}$, (Figs.22. 2(c) and 22.2(d)), the improvement of TMLE 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. In contrast, when β was constant at 4, the lowest list length was near its minimum at 10. 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 $\beta = 1$ (Tuglus and van der Laan 2008). The rank is much higher than 10, but the value is close to one as it should be. In general, the TMLE has the shortest list and is less affected by increases in correlation between the biomarkers than any other methods.

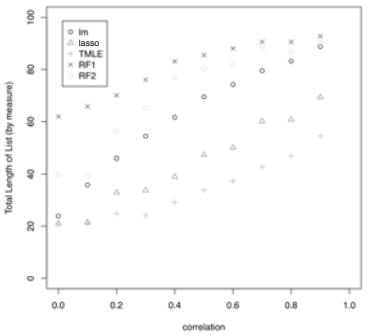
Though TMLE performs better than the three other methods, it is still sensitive to more extreme correlations (0.7–0.9). Our simulations show a small increase in bias for the measure of the true variables at higher correlations (see Tuglus and van der Laan 2008). However, in practice, high correlation can adversely affect the TMLE estimate due to violation of the positivity assumption. The increased length of the variable list when ranked by importance measure at correlations 0.8 and 0.9 indicates that the TMLE cannot distinguish the true variable from among a group of variables when the correlation is very high. Positivity violations or strong pairwise correlations can often be avoided if the "problem" variables (the variables highly correlated with the biomarker of interest A) are removed from the set of confounders (W). One simple method is to apply a correlation cutoff, where all W whose corre-



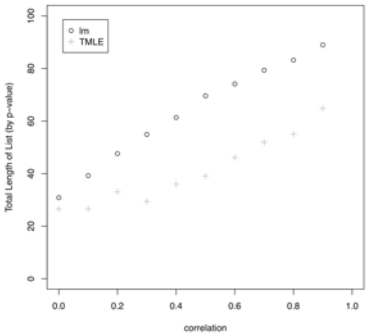
(a) VIM, β_{TRUE} constant



(b) p-value, β_{TRUE} constant



(c) VIM, β_{TRUE} variable



(d) p-value, β_{TRUE} variable

Fig. 22.2 Total length of list required to have all ten true variables in the list by $\rho = 0, ..., 0.9$ $\sigma_Y = 10, \beta_{TRUE} = 4$, (a) ranking by importance measure and (b) ranking by p-value. Then β_{TRUE} set at $\{1, ..., 10\}$ and plotted (c) ranking by importance measure and (d) ranking by p-value

lation with A is greater than a particular correlation (ρ_δ) are removed from the set of possible confounders for variable A prior to the application of the TMLE method.

The restriction of ρ_δ results in the algorithm's identifying all true variables as well as variables whose correlation with the true variables is higher than ρ_δ. Once we select ρ_δ, we are conceding that variables with correlations greater than ρ_δ cannot be teased apart to determine the true underlying (important) variable. By applying the correlation cutoff we are redefining our parameter. It is no longer the singular effect of A. Instead, we admit that, given the data, the true important variable cannot be targeted when the data are highly correlated and redefine our measure as a W_δ-adjusted importance where W_δ is a newly defined subset of W based on the correlation cutoff. Given this new definition of the parameter, important variables according to the W_δ-adjusted method include all important variables as well as all

variables whose correlation to an important variable is greater than a particular delta cutoff.

In the next section, we apply the correlation cutoff ($\rho_\delta = \{0.5, 0.75\}$) to a leukemia application, where the truth is unknown and the data are noisy. In practice, it is reasonable to label all potentially relevant variables as important when their effects cannot be disentangled. Setting a correlation cutoff explicitly specifies and acknowledges the method's threshold to detect the important variables among highly correlated confounders. We recommend that future applications use a larger set of ρ_δ values and provide importance measures and rankings for all variables given each ρ_δ, or data-adaptively select ρ_δ using the methods outlined in Bembom et al. (2008).

22.5 Leukemia Data Application

Biomarker data are generally high-dimensional and highly correlated; therefore, certain prescreening is necessary prior to performing a biomarker analysis. We are primarily concerned with screening the potential covariate set W. Reducing this set to relevant biomarkers can not only decrease computation time but, also result in better estimates from data-adaptive algorithms.

We want to reduce this set of covariates to include only potential confounders. Potential confounders for a given A are any W that are related to both Y and A. However, it can be time consuming to screen the confounder set for every A separately. We recommend screening only in terms of the association of the biomarkers with Y. This can be accomplished by discounting any components of W that are not significantly associated with Y based on simple univariate or bivariate (e.g., including A) regression, or a combination of results from multiple methods (e.g., all variables significant according to at least one of the following methods: linear regression, random forest, or lasso).

The above screening can also serve to reduce the number of variables for which we estimate variable importance (i.e., variables A). Removing these variables presumes they have insignificant importance. If the data are reduced based on the outcome Y, this reduction must be accounted for in any subsequent multiple testing procedures. An easy way to accomplish this is, after estimating the importance measure and calculating the associated p-values for a subset of the full variable set, automatically assign all prescreened variables (i.e., variables with no estimate) a p-value of one. Then apply the Benjamini–Hochberg step-up FDR-controlling multiple testing procedure (Benjamini and Hochberg 1995) to the full set of p-values as usual. This two-stage FDR multiple testing procedure still controls the FDR, and, if the prescreening has only discounted variables that would have had p-values greater than the cutoff, the procedure will also retain the type II control of the Benjamini–Hochberg step-up FDR-controlling procedure. See Tuglus and van der Laan (2009) for more details on the reasoning behind and performance of this procedure.

The data set from Golub et al. (1999) has been used in many papers for methodological comparison due to its relevance, limited gene set, and biological inter-

pretability. One goal in the original study was to identify differentially expressed genes in patients with acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix oligonucleotide arrays with 6,817 human genes for $n = 38$ patients (27 ALL, 11 AML). The gene expression set was preprocessed and reduced to 3,051 genes according to methods described in Dudoit et al. (2002).

This analysis mirrors the procedure implemented in the previous simulations. We first apply univariate linear regression to all genes and control for multiple testing using Benjamini–Hochberg step-up FDR controlling procedure. This resulting set contains 876 genes. To minimize bias due to positivity violations, a simple correlation cutoff of $\rho_c = \{0.5, 0.75\}$ is applied.

As in the simulation, we model the importance as $m(A, V \mid \beta) = \beta A$ for all A. For the initial $\bar{Q}_n^0(A, W)$ and $\bar{g}_n(W_s)$ we use a polynomial spline fit. We recommend using a data-adaptive algorithm such as super learner over lars/lasso in application, since in reality the structure of $E_0(Y \mid A, W)$ and $E_0(A \mid W)$ may have more than just additive main effects.

In this application, the outcome is binary, ALL ($Y = 0$) vs. AML ($Y = 1$); therefore we can interpret $\beta_0 a$ as the excess risk $P_0(Y = 1 \mid A = a) - P_0(Y = 1 \mid A = 0)$. The TMLE update presented in this chapter uses a linear regression working model, thereby not respecting the known probability bounds. For the sake of discovery, this limitation might not be that important. However, it is of interest to develop VIMs and the corresponding TMLE specifically designed for binary and bounded continuous outcomes.

The TMLEs of the VIMs and corresponding p-values are recorded and adjusted for multiple testing using Benjamini–Hochberg step-up FDR-controlling procedure. We selected all genes with adjusted p-values less than or equal to 0.05 and then ranked the selected set of genes by their absolute importance measures. The same method is used to rank genes according to the univariate regression measures, and p-values. RF1 and RF2 importance measures are simply ranked.

Using a p-value cutoff of 0.05, TMLE results in 272 significant genes at $\rho_c = 0.5$ and 225 significant genes at $\rho_c = 0.75$, while univariate regression identifies 681 significant genes. It is difficult to determine which list is better, especially when the lists include hundreds of genes. In this analysis, we compare the top then of each list in an effort to compare their biological relevance. In any given list, we include the top ten 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. The top ten genes according to their importance ranking for lm, RF1, RF2, and TMLE ($\rho_c = \{0.5, 0.75\}$) are shown in Tables 22.1–22.5.

Among the top ten genes according to the univariate regression results, CSTA, CD33, MYB, and ELA2 have all been associated with various types of cancer in the literature in previous quantitative analyses. 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 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 Ds, cells have shown increased resistance to oncogenic transformation in mouse models (Kozar et al. 2004).

There are marked differences and similarities between the TMLE-based results using a correlation cutoff of 0.5 and 0.75. There are five genes that are common between the two lists, four 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 random forest. 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 was recently found to inform the role of nicotine in colon cancer (Wong et al. 2007). It is also important to note that CHRNA7 is highly correlated with CD33. Cancer-relevant genes found only in Table 22.4 ($\rho_c = 0.5$) 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 and Krystal 1997). Cancer-relevant genes found only in Table 22.5 ($\rho_c = 0.75$) are CAT and E2F4. CAT regulates BCL-2 and is often underexpressed 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).

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 lm 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 toward prediction.

Random forest is a prediction and classification method, and the importance measures it provides are difficult to interpret. 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 the variable is more "important" than a vari-

Table 22.1 Top ten ranked genes according to absolute importance measures among significant genes according to a p-value cutoff of 0.05 using lm

Gene name/symbol	Mapped IDs	**lm**	**lm** rankp	**TMLE** rankp (0.75)	**TMLE** rankp (0.5)	**RF1** rank	**RF2** rank
CST3	M27891	0.258	1	13	17	6	3
CSTA	D88422	0.341	2	521	466	12	8
Zyxin	X95735	0.345	3	287	534	2	2
Macmarcks	HG1612-HT1612	−0.619	4	1041	1768	9	9
CD33	M23197	0.517	5	906	28	26	22
C-MYB	U22376.cds2.s	−0.403	6	69	99	40	28
ELA2	M27783.s	0.334	7	104	1970	15	14
DF	M84526	0.262	8	175	145	96	149
P48	X74262	−0.431	9	291	266	57	31
LTC4S	U50136.rna1	0.725	10	146	2110	38	60

Table 22.2 Top ten ranked genes according to their importance measures using RF1

Gene name/symbol	Mapped IDs	**RF1**	**RF1** rank	**RF2** rank	**TMLE** rankp (0.75)	**TMLE** rankp (0.5)	**lm** rankp
FAH	M55150	0.953	1	1	588	234	52
Zyxin	X95735	0.823	2	2	287	534	3
TCF3	M31523	0.718	3	6	155	400	12
ADM	D14874	0.693	4	5	329	2136	57
PTX3	M31166	0.691	5	33	33	201	28
CST3	M27891	0.682	6	3	13	17	1
TOP2B	Z15115	0.654	7	4	1	2	33
CCND3	M92287	0.621	8	10	481	924	19
Macmarcks	HG1612-HT1612	0.613	9	9	1041	1768	4
APLP2	L09209.s	0.610	10	7	160	408	25

Table 22.3 Top ten ranked genes according to their importance measures using RF2

Gene name/symbol	Mapped IDs	**RF2**	**RF2** rank	**RF1** rank	**TMLE** rankp (0.75)	**TMLE** rankp (0.5)	**lm** rankp
FAH	M55150	0.426	1	1	588	234	52
Zyxin	X95735	0.282	2	2	287	534	3
CST3	M27891	0.218	3	6	13	17	1
TOP2B	Z15115	0.208	4	7	1	2	33
ADM	D14874	0.200	5	4	329	2136	57
TCF3	M31523	0.186	6	3	155	400	12
APLP2	L09209.s	0.183	7	10	160	408	25
CSTA	D88422	0.171	8	12	521	466	2
Macmarcks	HG1612-HT1612	0.164	9	9	1041	1768	4
CCND3	M92287	0.159	10	8	481	924	19

Table 22.4 Top ten ranked genes according to absolute importance measures among significant genes according to a p-value cutoff of 0.05 using TMLE with correlation cutoff $\rho_c = 0.5$

Gene name/symbol	Mapped IDs	**TMLE**	**TMLE** rankp (0.5)	**TMLE** rankp (0.75)	**lm** rankp	**RF1** rank	**RF2** rank
TOP2B	Z15115	−0.973	1	2	33	7	4
CHRNA7	X70297	0.839	2	1	48	69	61
corneodesmosin	L20815	0.338	3	3	1875	1846	2004
BCL3	U05681.s	0.314	4	4	477	558	821
KTN1	Z22551	−0.311	5	18	373	2967	118
CaM	U81554	0.272	6	81	367	476	749
TCF7	X59871	−0.159	7	6	569	635	887
PTTG1IP	Z50022	0.310	8	5	2753	2674	483
MCL1	L08246	0.293	9	2406	61	75	65
PI3K	Z46973	−0.172	10	113	734	772	1009

Table 22.5 Top ten ranked genes according to absolute importance measures among significant genes according to a p-value cutoff of 0.05 using TMLE with correlation cutoff $\rho_c = 0.75$

Gene name/symbol	Mapped IDs	**TMLE**	**TMLE** rankp (0.75)	**TMLE** rankp (0.5)	**lm** rankp	**RF1** rank	**RF2** rank
CHRNA7	X70297	1.260	1	2	48	69	61
TOP2B	Z15115	−0.946	2	1	33	7	4
corneodesmosin	L20815	0.327	3	3	1875	315	621
BCL3	U05681.s	0.181	4	4	477	316	622
Surface glycoprotein	Z50022	0.310	5	8	2753	317	474
TCF7	X59871	−0.175	6	7	569	318	623
CAT	X04085.rna1	0.163	7	21	92	56	59
E2F4	U18422	−0.256	8	42	1752	319	624
UGP2	U27460	−0.244	9	14	155	186	303
SELL	M15395	0.183	10	43	340	84	316

able with a value of 0.611. Also, out of the top ten lists for RF1 and RF2 (12 genes total), only four genes were found to be biologically associated with cancer, and only one specifically relating to ALL/AML distinction, TCF3. Why TCF3 is rated second for RF1 and sixth for RF2 is unclear. In comparison, lm found four related to cancer, two of which specifically related to AML/ALL.

The TMLE measure provides directionality and is less sensitive to increases in correlation (Sect. 22.3). Given an importance measure of −0.175, we can conclude that this particular gene is up-regulated in ALL patients when compared to AML patients. This particular measure is for TCF7 using a correlation cutoff of 0.75. TCF7 is rarely expressed in AML and often highly expressed in ALL patients (especially T-cell related). Out of the six cancer-related genes in the top ten list for 0.75 cutoff, three are biologically related to the AML/ALL distinction. When the cutoff is 0.5, there are eight cancer-related genes, three related to the AML/ALL distinction. For

all three AML/ALL-related genes, the directionality of the relationship is biologically correct.

The TMLE results do 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 ten genes. Further support for TMLE 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.

22.6 Discussion

Variable importance results vary widely, leading to long lists and confusion. In this chapter, we proposed using the statistical analogs of causal effects as VIMs and using TMLE to statistically assess these effect measures. In simulation, this proved resilient to increases in correlation while controlling the 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, the commonly employed univariate linear regression is highly susceptible to increases in type I error due to increased correlation between the biomarkers. The utilization of machine learning algorithms, such as lasso/lars, to estimate these same target parameters is incomplete without the targeting carried out by the TMLE update, which removes bias and allows for statistical inference in terms of p-values, multiple testing, and confidence intervals.