# A Comparison of Node-Splitting Rules in Recursive Partitioning Analysis of Multivariate Quantitative Structure Activity Data

YUJUN WU, MARC G. GENTON, and LEONARD A. STEFANSKI

Tree-structured methods have been advocated to model nonlinear relationships in quantitative structure-activity relationship (QSAR) studies. One such algorithm, MultiSCAM, was developed to analyze QSAR data with multivariate continuous responses, classifying objects into similar categories by building a tree with recursive partitioning. Hotelling's $T^2$ test is the method used to determine splits in MultiSCAM. However, this test is not feasible when the dimension of an observation exceeds the number of observations, which often happens when growing trees. This problem is further exacerbated by the fact that missing values are common in QSAR data. We consider two alternatives, the pooled component test (PCT) and a simple ANOVA $F$ test. To compare the three node-splitting tests, we introduce a comprehensive simulation design that could be used by others to evaluate their methods of tree building. A drug discovery dataset is used to illustrate the methods.

**Key Words:** Classification tree; Hotelling's $T^2$; MultiSCAM; Pooled component test; QSAR.

## 1. Introduction

A key component of drug discovery is understanding how a chemical compound's biological activity, as measured against one or more proteins, relates to the chemical features of that compound. A common means of quantifying a chemical compound, and the method used for the type of data studied in this article, is to record the presence or absence of particular pairs of atoms in the molecule. In this case the feature vector for compound $i$ is a $q \times 1$ vector $X_i$ with a 1(0) in the $j$th position according to the presence(absence) of the $j$th atom pair in the compound. The number of atom pairs $q$ is large, often on the order of hundreds or thousands. Testing compound $i$ on $p$ proteins results in a $p \times 1$ response vector $Y_i$ of activity levels. Data containing information on $n$ compounds consists of pairs $\{Y_i, X_i\}$, $i = 1, \ldots, n$. Collecting the response and chemical descriptor vectors into matrices results in the response matrix $\mathbf{Y}_{n\times p}$ and chemical descriptor matrix $\mathbf{X}_{n\times q}$. The objective of statistical analysis is to determine those chemical descriptors (atom pairs) that explain variation in activity.

The matrix $\mathbf{X}$ is not a design matrix in the usual linear model sense. Biological activity often depends on the presence or absence of groups of atom pairs and thus the true design matrix is obtained by appending to the columns of $\mathbf{X}$ all of the multiway interaction columns formed by taking products of two or more columns of $\mathbf{X}$. Because of the prohibitive size of the full design matrix when $q$ is large, traditional linear models approaches are seldom used to analyze chemical structural features to biological activity. Instead tree-based methods have been advocated and shown to be successful in analyzing structure-activity data (Hawkins et al. 1997; Rusinko et al. 1999). Tree-based methods are semiparametric, able


Statistics in Biopharmaceutical Research
May 2009, Vol. 1, No. 2
DOI: 10.1198/sbr.2009.0020

to handle huge datasets, and often produce accurate predictions and classifications of cases; see Morgan and Sonquist (1963), Breiman et al. (1984), Segal (1992), and Zhang (1998).

Because the chemical descriptors are binary, QSAR data are often analyzed with trees constructed using recursive partitioning (RP) (Hawkins and Kass 1982). We consider one particular algorithm known as MultiSCAM (Keefer 2001), a multivariate counterpart of the univariate response algorithm called Statistical Classification of the Activity of Molecules (SCAM) studied by Rusinko et al. (1999). SCAM tree-building algorithms are simple and fast. Node splitting is accomplished via two-sample tests of means. For each $j$, the observations in a node are partitioned according to whether the components of the $j$th column of $\mathbf{X}$ equal 0 or 1, and the $p$ value, $p_j$, from a two-sample test of equality of means is calculated. The node is split if $p_{(1)} = \min\{p_1, \ldots, p_q\} < \alpha_0/q$ for some significance level $\alpha_0$; and the splitting descriptor is identified by that $j$ for which $p_j = p_{(1)}$. The root node contains all of the observations. Node-splitting occurs recursively until no nodes satisfy the criterion for splitting.

Operating characteristics of the algorithm are largely determined by the test used for determining splits. SCAM and MultiSCAM use Hotelling's $T^2$ (Hotelling 1931). This is problematic unless $p$ is small, as Hotelling's $T^2$ test is applicable only when the sample sizes $n_1$ and $n_2$ satisfy $n_1 + n_2 - 1 > p$; otherwise the pooled covariance matrix is singular. The latter occurs in tree building because node size $(n_1 + n_2)$ decreases with increasing tree size. Missing values, common in QSAR data, exacerbate the problem if only complete-cases are used. Alternatives to complete-case analyses are possible—imputation, for example—but have other difficulties that make them unattractive. For instance, imputing missing data effectively in the tree structure is an unsolved problem. In this article, we assume data missing completely at random, although in practice this assumption is likely violated to some extent.

We study the characteristics of trees constructed for QSAR data using two alternative testing methods that are applicable when $p \geq n_1 + n_2 - 1$, and thus are less affected by missing values. The first is the so-called *pooled component test* (PCT) developed by Wu, Genton, and Stefanski (2006). The PCT is a function of marginal sample means and variances and thus avoids singularity of the pooled sample covariance matrix when $p \geq n_1 + n_2 - 1$. Wu et al. (2006) derived an approximation to the sampling distribution of the PCT and showed that the test has good power. The second alternative test is a simple analysis of variance $F$-test (ANOVA $F$), derived under a working assumption that the random variation in activity levels is independent, experimental (or measurement) error resulting from the experiment that assessed activity.

We also develop models for simulating QSAR data from a classification tree and use them to compare the properties of trees formed using the three node-splitting tests (Hotelling's $T^2$, PCT, and ANOVA $F$), under cases with correlated and independent multivariate responses, and with and without missing values. Section 2 describes the three node-splitting tests. Section 3 presents results from a Monte Carlo study comparing the tests. Section 4 presents an application to real QSAR data. A summary and conclusions appear in Section 5.

## 2. Node-Splitting Tests

We briefly review the three node-splitting rules we study in the context of the SCAM algorithm. Partitioning the observations in a node via a column from $\mathbf{X}$ produces two $p$-variate samples $\{Y_{k,r}, r = 1, \ldots, n_k\}$, $k = 1, 2$.

### 2.1 Hotelling's $T^2$ Test

Hotelling's $T^2$ test statistic is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{Y}_1 - \overline{Y}_2)^T \mathbf{S}^{-1}(\overline{Y}_1 - \overline{Y}_2),$$

where $\overline{Y}_k = \sum_{r=1}^{n_k} Y_{k,r}/n_k$, $k = 1, 2$, and

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\left\{\sum_{r=1}^{n_1}(Y_{1,r} - \overline{Y}_1)(Y_{1,r} - \overline{Y}_1)^T + \sum_{r=1}^{n_2}(Y_{2,r} - \overline{Y}_2)(Y_{2,r} - \overline{Y}_2)^T\right\}.$$

When the data are iid samples from two normal populations with common covariance matrix, and $p < n_1 + n_2 - 1$, then $\tau T^2$ has an $F_{p,n_1+n_2-p-1}$ distribution under the null hypothesis of equal means where $\tau = \frac{n_1+n_2-p-1}{n_1+n_2-2}$. $T^2$ requires complete-case observations. When that is not the case, and only complete-case vectors are used in the construction of $T^2$, then sample sizes are reduced accordingly.

### 2.2 Pooled Component Test (PCT)

The PCT is designed to accommodate missing data. The pattern of missingness in sample $k$ is indicated by the $n_k \times p$ matrix $M_k$ with $(i, j)$th element $M_{k,ij} = 1$, if $Y_{k,ij}$ is not missing, and $= 0$ otherwise. The componentwise sample mean for variable $j$ in sample $k$ is $\overline{Y}^{(c)}_{k,j} = (\sum_{r=1}^{n_k} Y_{k,rj} M_{k,rj})/n_{k,j}$, and the pooled componentwise sample variance is

$$S_j^2 = \frac{\sum_{k=1}^{2}\sum_{r=1}^{n_k}(Y_{k,rj} - \overline{Y}^{(c)}_{k,j})^2 M_{k,rj}}{n_{1,j} + n_{2,j} - 2},$$

where $n_{k,j} = \sum_{r=1}^{n_k} M_{k,rj}$ is the number of nonmissing observations for variable $j$ in sample $k$. The PCT statistic $Q = \sum_{j=1}^{p} a_j Q_j / p$ where $Q_j = (\overline{Y}_{1,j}^{(c)} - \overline{Y}_{2,j}^{(c)})^2 / {S_j}^2$ and $a_j = (n_{1,j} n_{2,j})/(n_{1,j} + n_{2,j})$. For normally distributed data with common covariance matrix, the null distribution of $Q$ is well approximated by a scaled chi-squared distribution, say $b\chi_d^2$. Provided that $n_{1,j} + n_{2,j} > 6$, $j = 1, \ldots, p$, the scale parameter $b$ and degrees of freedom $d$ are estimated by matching the estimated mean and variance of $Q$ with the mean and variance of $b\chi_d^2$ (Wu et al. 2006). The approximating chi-squared distribution depends on the correlations among the multivariate responses and thus uses correlation information in the data even though the test statistic itself is based on component sample means and variances. Wu et al. (2006) showed that the PCT performs well for different covariance structures spanning a wide range of correlations.

### 2.3 ANOVA $F$ Test

Assuming that $Y_{k,ij}$ differs from its mean by only a random error arising from the experiment that produced $Y_{k,ij}$ leads to the model

$$Y_{k,ij} = \mu_{kj} + \epsilon_{kij}, \quad k = 1, 2; \quad i = 1, \ldots, n_k; \quad j = 1, \ldots, p, \tag{1}$$

where $\mu_{kj}$ is the mean response for variable $j$ in group $k$, and $\epsilon_{kij}$ are independent random errors. For this model the null hypothesis of no difference between group means corresponds to having $\mu_{1j} = \mu_{2j} = \mu_j$, $j = 1, \ldots, p$. This linear hypothesis is tested with the $F$-statistic

$$F = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_F)/(\mathrm{df}_R - \mathrm{df}_F)}{\mathrm{SSE}_F/\mathrm{df}_F},$$

on $\mathrm{df}_R - \mathrm{df}_F$ and $\mathrm{df}_F$ degrees of freedom where, allowing for missingness,

$$\mathrm{SSE}_F = \sum_{k=1}^{2}\sum_{j=1}^{p}\sum_{i=1}^{n_k}(Y_{k,ij} - \overline{Y}_{k,j}^{(c)})^2 M_{k,ij},$$

$$\mathrm{df}_F = \sum_{k=1}^{2}\sum_{j=1}^{p}(n_{k,j} - 1) I(n_{k,j} > 0),$$

$$\mathrm{SSE}_R = \sum_{j=1}^{p}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(Y_{k,ij} - \overline{Y}_{j}^{(c)})^2 M_{k,ij},$$

$$\mathrm{df}_R = \sum_{j=1}^{p}(n_{1,j} + n_{2,j} - 1) I(n_{1,j} + n_{2,j} > 0),$$

with $\overline{Y}_j^{(c)} = \sum_{k=1}^{2}\sum_{i=1}^{n_k} Y_{k,ij} M_{k,ij} / \sum_{k=1}^{2}\sum_{i=1}^{n_k} M_{k,ij}$, and $I(\cdot)$ is the indicator function.

Like the PCT, the ANOVA $F$ uses component-wise information for each response variable and thus avoids problems of high dimension and missing data. However, unlike the PCT, independence between the multivariate responses is assumed for the ANOVA $F$ test. The fact that small correlations will not severely impact the $F$ test method is evident in the QSAR data example in Section 4.

## 3. Monte Carlo Simulation Study

We introduced a comprehensive simulation design to compare the performance of the three node-splitting tests in building trees. The method comparisons are based upon several criteria indicating prediction accuracy and descriptor selection. The models developed for simulating tree data could be used by others in Monte Carlo studies evaluating methods of tree building.

### 3.1 Simulation Design

We designed two tree structures, a symmetric tree denoted by A, and a right-lopsided tree denoted by B, which imitate the shapes of the real trees in Section 4. Let $x_j$, $j = 1, \ldots, q$, with $q = 300$, denote the binary descriptors, and six of them, $x_5$, $x_{24}$, $x_{45}$, $x_{112}$, $x_{208}$, and $x_{244}$, were arbitrarily chosen as the splitting variables. The binary value of each descriptor was given by $I(V < 0.5)$, where $V$ is a Uniform(0, 1) random variable. Figures 1 and 2 illustrate the diagrams of the two trees, both of which have nine terminal nodes as indicated by rectangles. The variable in the circle at each internal node indicates the node splitting variable, with $x = 0$-valued subjects placed in the right-most path and the $x = 1$-valued subjects in the left-most path. We set sample size $n = 200$.

A linear model was used to generate the multivariate response data with the incorporation of underlying tree structure. In particular, for tree A, we generated the $p$-variate responses $Y_i$, $i = 1, \ldots, n$, by

$$\begin{aligned} Y_i = {} & \widetilde{x}_{5,i}\widetilde{x}_{24,i}\widetilde{x}_{112,i}\mu_1 + \widetilde{x}_{5,i}\widetilde{x}_{24,i}x_{112,i}\widetilde{x}_{244,i}\mu_2 \\ & + \widetilde{x}_{5,i}\widetilde{x}_{24,i}x_{112,i}x_{244,i}\mu_3 + \widetilde{x}_{5,i}x_{24,i}\mu_4 \\ & + x_{5,i}\widetilde{x}_{45,i}\widetilde{x}_{208,i}\mu_5 + x_{5,i}\widetilde{x}_{45,i}x_{208,i}\widetilde{x}_{244,i}\mu_6 \\ & + x_{5,i}\widetilde{x}_{45,i}x_{208,i}x_{244,i}\mu_7 \\ & + x_{5,i}x_{45,i}\widetilde{x}_{208,i}\mu_8 + x_{5,i}x_{45,i}x_{208,i}\mu_9 + \epsilon_i \end{aligned} \tag{2}$$

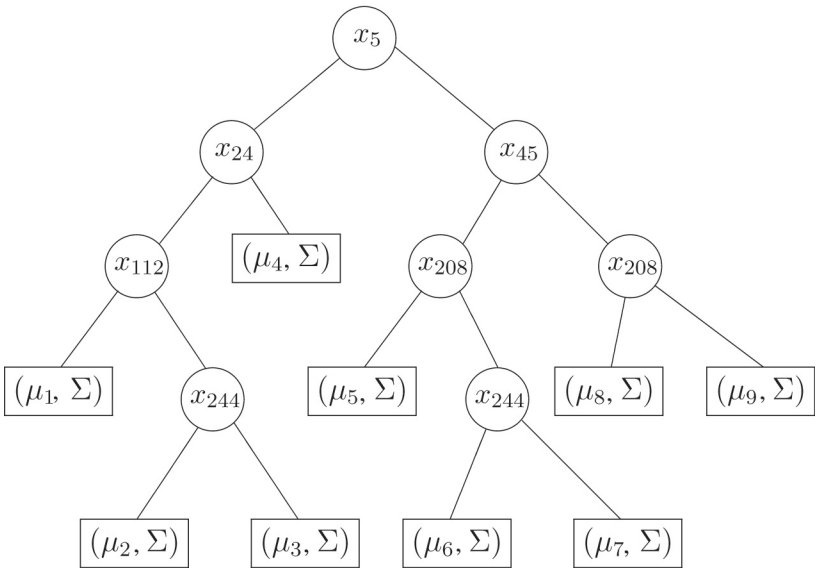



Figure 1. The symmetric tree A.

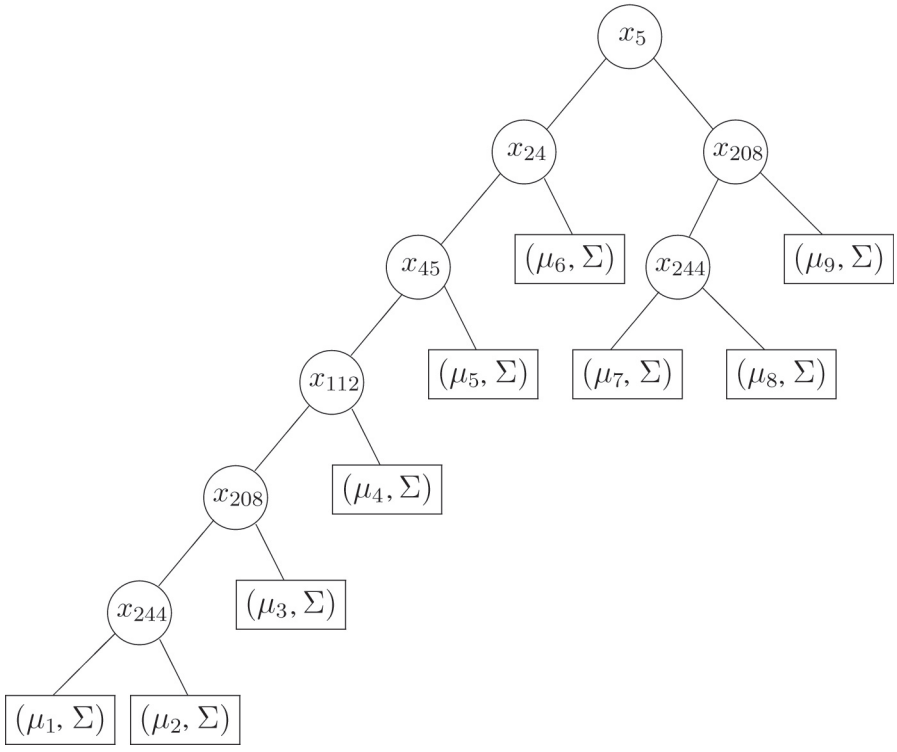



Figure 2. The lopsided tree B.

and, for tree B,

$$\begin{aligned} Y_i \;=\; & \widetilde{x}_{5,i}\widetilde{x}_{24,i}\widetilde{x}_{45,i}\widetilde{x}_{112,i}\widetilde{x}_{208,i}\widetilde{x}_{244,i}\,\mu_1 \\ & +\widetilde{x}_{5,i}\widetilde{x}_{24,i}\widetilde{x}_{45,i}\widetilde{x}_{112,i}\widetilde{x}_{208,i}x_{244,i}\,\mu_2 \\ & +\widetilde{x}_{5,i}\widetilde{x}_{24,i}\widetilde{x}_{45,i}\widetilde{x}_{112,i}x_{208,i}\,\mu_3 \\ & +\widetilde{x}_{5,i}\widetilde{x}_{24,i}\widetilde{x}_{45,i}x_{112,i}\,\mu_4 + \widetilde{x}_{5,i}\widetilde{x}_{24,i}x_{45,i}\,\mu_5 \\ & +\widetilde{x}_{5,i}x_{24,i}\,\mu_6 + x_{5,i}\widetilde{x}_{208,i}\widetilde{x}_{244,i}\,\mu_7 \\ & +x_{5,i}\widetilde{x}_{208,i}x_{244,i}\,\mu_8 \\ & +x_{5,i}x_{208,i}\,\mu_9 + \epsilon_i, \end{aligned} \tag{3}$$

where $\widetilde{x}_{j,i} = 1 - x_{j,i}$, and the random error $\epsilon_i$ has a $p$-variate normal distribution, $N_p(0, \Sigma)$, with $p = 10$. The fixed vectors $\mu_t$, $t = 1, \ldots, 9$, with a dimension of 10 were arbitrarily taken from the set of sample mean vectors for the terminal nodes of the real tree built by MultiSCAM in Section 4. In particular,

$$\begin{aligned} \mu_1 &= (5.09, 3.99, 3.92, 5.57, 5.78, 4.20, \\ &\quad 5.46, 4.33, 4.75, 5.64)^T, \\ \mu_2 &= (3.77, 4.39, 4.10, 5.19, 5.11, 4.89, \\ &\quad 5.41, 4.24, 5.42, 4.84)^T, \\ \mu_3 &= (3.76, 3.89, 5.31, 5.21, 5.79, 3.89, \\ &\quad 5.65, 5.36, 5.92, 5.57)^T, \\ \mu_4 &= (4.93, 7.26, 7.98, 5.09, 4.67, 4.21, \\ &\quad 5.16, 4.38, 5.94, 4.19)^T, \\ \mu_5 &= (3.83, 4.08, 3.94, 4.67, 4.67, 5.96, \\ &\quad 4.30, 4.15, 4.39, 4.00)^T, \\ \mu_6 &= (4.77, 5.25, 5.46, 5.69, 5.75, 5.12, \\ &\quad 5.16, 4.45, 5.19, 5.59)^T, \\ \mu_7 &= (3.85, 4.08, 6.89, 5.80, 5.49, 3.85, \\ &\quad 5.17, 4.30, 4.53, 5.64)^T, \\ \mu_8 &= (4.18, 4.00, 4.00, 4.69, 4.89, 3.89, \\ &\quad 5.93, 4.29, 5.11, 4.86)^T, \\ \mu_9 &= (5.72, 3.89, 4.37, 4.01, 4.94, 3.89, \\ &\quad 4.69, 4.41, 6.37, 5.02)^T. \end{aligned}$$

Each fixed item on the right side of model (2) (model (3)) indicates the path to a terminal node in tree A (tree B), and the data generated in the $t$th terminal node are distributed with a multivariate normal $N_p(\mu_t, \Sigma)$. We considered two covariance structures for $\Sigma$. One is the pooled sample covariance matrix for all the terminal nodes in the real tree, which has variances ranging from 0.10 to 0.97 and correlations ranging from 0.03 to 0.75 with an average of 0.23, identified as real covariance (RC); the other is the diagonal matrix with identical diagonal elements equal to 0.49, the average of the diagonal values of RC, identified as independence (InD). For each type of tree with different covariance structures, we generated 100 replications. Missing, or incomplete, data were generated by randomly deleting 3.5% of the data for each response variable, resulting in approximate 30% incomplete observations which have at least one element missing. We used $\alpha_0 = 0.05$ to determine significant split, using a Bonferroni adjustment as described in the introduction. We also generated data by doubling the variances in both of the InD and RC cases, with the correlation matrices kept the same. The results for the data with doubled errors are similar to those obtained for the original errors, and therefore are not reported.

We evaluated prediction accuracy of the built trees via model error (ME),

$$\mathrm{ME} = \frac{1}{n}\sum_{t=1}^{R}\sum_{i\in\text{node } t}(Y_{t,i}^{(m)} - \widehat{Y}_{t,i})^T\,\Sigma^{-1}(Y_{t,i}^{(m)} - \widehat{Y}_{t,i}),$$

where $Y_{t,i}^{(m)}$ is the mean vector of the $i$th observation in the terminal node $t$, $\widehat{Y}_{t,i}$ is the predicted value for the observation, which amounts to the sample mean in the node, and $R$ is the number of terminal nodes. A small value of ME is desirable. When the tree is correct, the expected value of ME should be equal to $9p/n$; see the proof in Appendix. Two other criteria assessing the selection of important descriptors are the inclusion and exclusion measures

$$\mathrm{InM} = \mathrm{card}(S \cap \hat{S})/\mathrm{card}(S)$$

and

$$\mathrm{ExM} = \mathrm{card}(\hat{S} - S),$$

where $S$ is the set of true splitting descriptors, that is, $S = \{x_5, x_{24}, x_{45}, x_{112}, x_{208}, x_{244}\}$, $\hat{S}$ is the set of the identified splitting descriptors in the built tree, and $\mathrm{card}(G)$ indicates the size of the set $G$. Larger InM indicates higher tendency of selecting informative descriptors; smaller ExM indicates higher tendency of excluding uninformative descriptors. The ideal results are InM = 1 and ExM = 0. In addition, we compared tree size, defined as the number of terminal nodes.

### 3.2 Simulation Results for Complete Data

Figures 3 and 4 display the boxplots of ME, tree size, InM, and ExM corresponding to different node-splitting tests for trees A and B, respectively. In each panel of the two figures, there are two clusters with the left showing the results for the InD case and the right showing the results for the RC case. For the complete data, Hotelling's $T^2$ test is expected to perform well, and hence serves as a benchmark.

We first concentrate on the results for the symmetric tree A. It turns out that Hotelling's $T^2$ test tends to build the trees having correct size, and excludes the

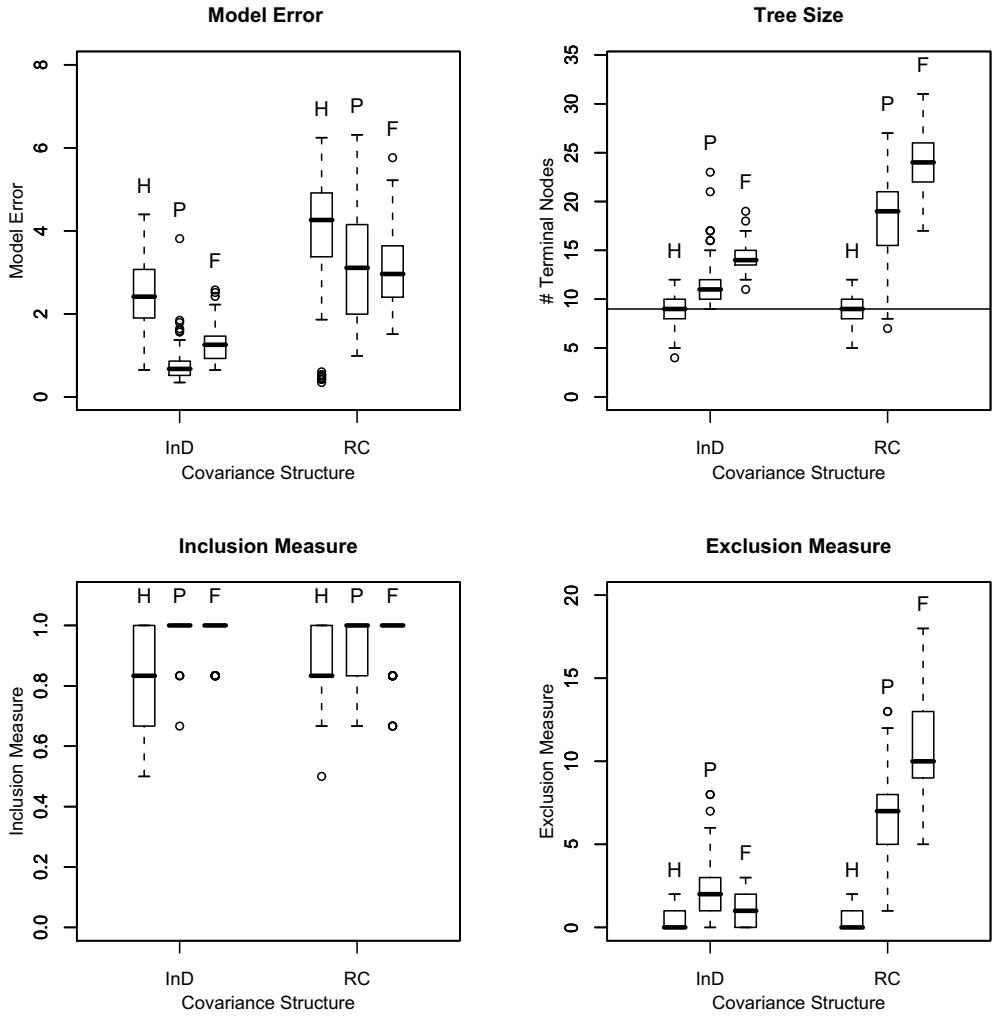
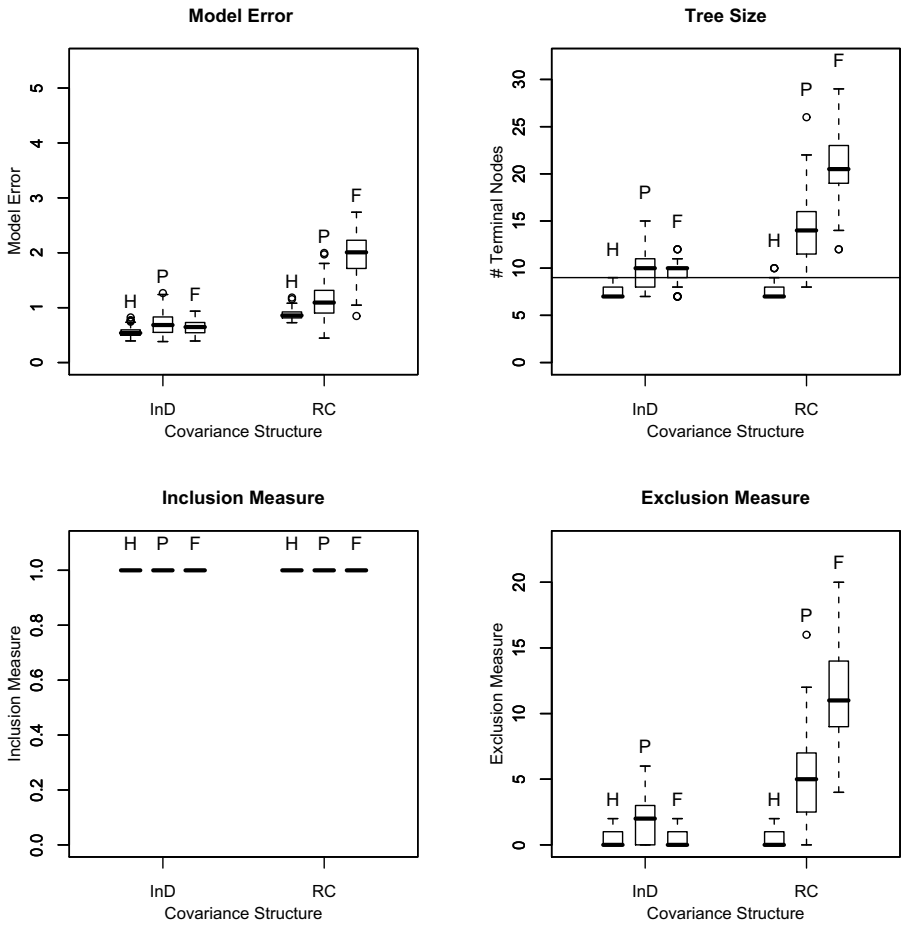



Figure 3. The boxplots of ME, tree size, InM, and ExM for Hotelling's $T^2$ test, PCT, and ANOVA $F$ test, for tree A when the data are complete. The black solid line in the up right graph corresponds to the underlying tree size 9.

uninformative descriptors, but performs not as well as the other two tests in terms of ME and more often misses selecting some important descriptors. For the InD case, both PCT and ANOVA $F$ test exhibit similar performance. However, for the RC case, PCT outperforms the ANOVA $F$ test by producing fewer overfitted trees and smaller ExM. The fact that the independence assumption is not satisfied could be one reason leading to the considerable overfitting of ANOVA $F$ test in this case. Although the overfitting for both PCT and ANOVA $F$ test leads to the inclusion of many noisy descriptors in the tree, that is, large ExM, it reduces the chance of missing the important descriptors, that is, large InM. It is of note that the path to each terminal node in the built tree provides a clue to the molecular structure associated with the activity. In practice, chemists often like to see nodes divided as much as possible as this provides additional insight into the structure of a chemical compound and its effect on activity.

With regards to the lopsided tree B, there is not much difference among the three tests in the situation of InD covariance structure. For the RC case, Figure 4 indicates that Hotelling's $T^2$ performs the best, whereas ANOVA $F$ test is the worst, in terms of all the criteria. In addition, Table 1 lists the results of average ME produced by all three tests for both trees A and B. In our design settings, for correct trees A and B, the expected value



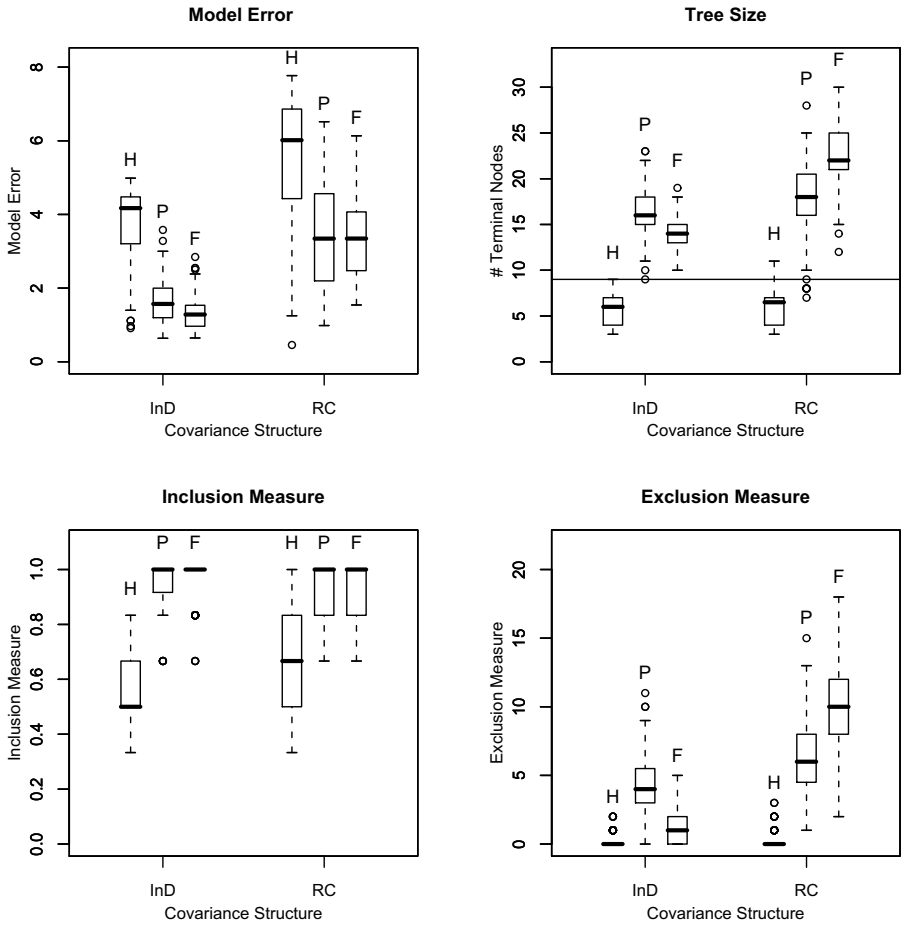

H : Hotelling's $T^2$ test P : PCT F : ANOVA $F$ test

InD : Independence RC : Real Covariance

Figure 4. The boxplots of ME, tree size, InM, and ExM for Hotelling's $T^2$ test, PCT, and ANOVA $F$ test, for tree B when the data are complete. The black solid line in the up right graph corresponds to the underlying tree size 9.

Table 1. The average ME (s.e.) of the built trees

| | | InD | | RC | |
|---|---|---|---|---|---|
| Tree | Tests | Complete | 30% incomplete | Complete | 30% incomplete |
| A | Hotelling | 2.48(0.09) | 3.82(0.10) | 3.97(0.14) | 5.36(0.19) |
| | PCT | 0.77(0.04) | 1.65(0.06) | 3.18(0.14) | 3.44(0.15) |
| | ANOVA $F$ | 1.25(0.04) | 1.34(0.05) | 3.07(0.10) | 3.32(0.10) |
| B | Hotelling | 0.55(0.01) | 0.93(0.07) | 0.88(0.01) | 1.50(0.10) |
| | PCT | 0.71(0.02) | 0.75(0.02) | 1.12(0.03) | 1.18(0.04) |
| | ANOVA $F$ | 0.65(0.01) | 0.66(0.01) | 1.97(0.04) | 1.98(0.04) |




H : Hotelling's $T^2$ test P : PCT F : ANOVA $F$ test

InD : Independence RC : Real Covariance

Figure 5. The boxplots of ME, tree size, InM, and ExM for Hotelling's $T^2$ test, PCT, and ANOVA $F$ test, for tree A when 30% data are incomplete. The black solid line in the up right graph corresponds to the underlying tree size 9.

Table 2. The simulation results for the null tree

| | InD | | RC | |
|---|---|---|---|---|
| Tests | ME (s.e.) | Size (s.e.) | ME (s.e.) | Size (s.e.) |
| Hotelling | 0.050(0.003) | 1.01(0.01) | 0.050(0.003) | 1.01(0.01) |
| PCT | 0.056(0.004) | 1.04(0.02) | 0.105(0.015) | 1.50(0.13) |
| ANOVA $F$ | 0.052(0.003) | 1.02(0.01) | 1.018(0.063) | 10.60(0.62) |

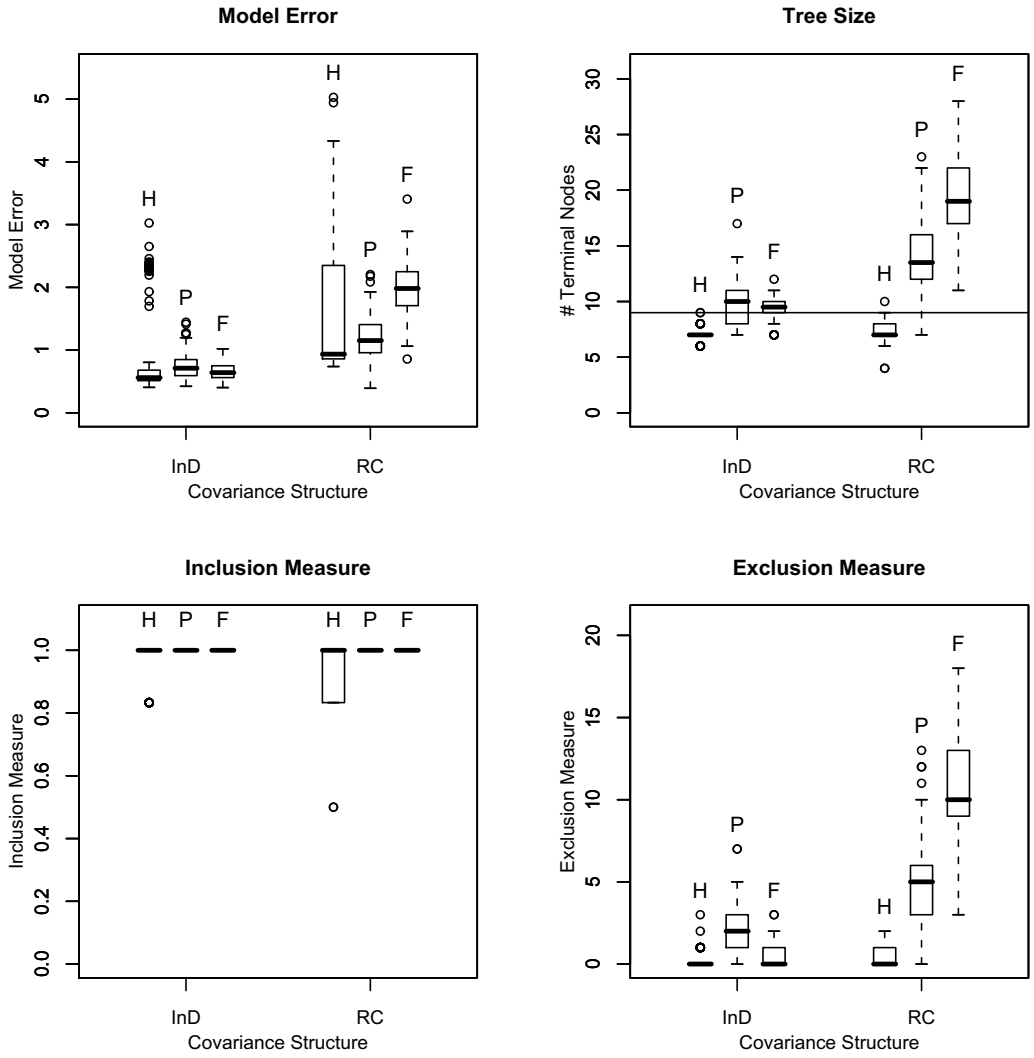



Figure 6. The boxplots of ME, tree size, InM, and ExM for Hotelling's $T^2$ test, PCT, and ANOVA $F$ test, for tree B when 30% data are incomplete. The black solid line in the up right graph corresponds to the underlying tree size 9.

Table 3. The tree results for the example

| | | | | Terminal node sizes | |
|---|---|---|---|---|---|
| Incomplete | Tests | Size | # Descriptors | Range | Mean |
| No | Hotelling | 31 | 30 | 1 ∼ 62 | 18.6 |
| | PCT | 75 | 69 | 1 ∼ 43 | 7.7 |
| | ANOVA $F$ | 63 | 59 | 1 ∼ 41 | 9.1 |
| 30% | Hotelling | 27 | 26 | 1 ∼ 60 | 21.3 |
| | PCT | 59 | 56 | 1 ∼ 40 | 9.8 |
| | ANOVA $F$ | 56 | 51 | 1 ∼ 40 | 10.3 |

of ME should be equal to 0.45 (=$9p/n = 90/200$). Our methods perform better for the lopsided tree B than for the symmetric tree A.

### 3.3 Simulation Results for Incomplete Data

The results for incomplete data appear in Figures 5 and 6 for trees A and B, respectively. As expected, the performance of the complete-case Hotelling's $T^2$ test is adversely affected by the missing information, and its prediction accuracy deteriorates. Ignoring the incomplete-case data worsens the problem of $p \geq n_1 + n_2 - 1$ that often renders the test unavailable to detect significance of splits. As a result, the tree tends to be underfitted, and more important descriptors are missed. The PCT and ANOVA $F$ tests performed relatively stable regardless of missingness. The average ME for the three tests for 30% incomplete data are compared with the results for the complete data in Table 1.

### 3.4 Simulation Results for Null Tree

We also conducted simulations to study the performance of the three tests for the null tree where all the compounds belong to the same group and there is only one terminal node. We kept the same data design settings and the binary descriptor matrices as before, except that we generated the responses by $Y_i = 4.5 + \epsilon_i$, for $i = 1, \ldots, n$. The method comparisons were for the complete data. Table 2 lists the results of average ME and tree size over 100 replications. In the InD case, all of the three tests perform similarly well with the false splitting errors very well controlled. In the RC case, the PCT is slightly inferior to Hotelling's $T^2$ test by introducing a few more false splitting errors, whereas the ANOVA $F$ test severely overfits the tree due to the violation of the independence assumption.

## 4. An Example from Drug Discovery

Here we present a real quantitative structure-activity dataset as an example to illustrate the methods. The data contain 1,024 binary descriptors and 576 chemical compounds with each one measured in 10 target proteins. The trees were built using the MultiSCAM algorithm with Hotelling's $T^2$ test, PCT, and ANOVA $F$ test. We used $\alpha_0 = 0.05$ to judge significance of the splits. Although there are often missing data in these types of datasets, here we started with a dataset with no missingness. To assess the performance with missing data, we randomly deleted 3.5% of each response data, resulting in approximate 30% incomplete-case observations, and we reconstructed the trees with the three tests again (Hotelling's $T^2$ test used only complete-case data).

The resulting dendrograms are presented in Figure 7. The figure indicates a competitive result for the PCT and ANOVA $F$ tests, both of which detect more significant splits with small sample sizes compared to Hotelling's $T^2$ test and hence lower probability of missing important descriptors, consistent with the simulation results. The missing information had small effects on the PCT and ANOVA $F$ tests, while Hotelling's $T^2$ test was rendered less effective by the loss of information when sample size becomes small. We note that the averages of the correlation elements within the pooled correlation matrices for the terminal nodes of the estimated trees are small, 0.23 for Hotelling's $T^2$ test, 0.12 for PCT, and 0.14 for ANOVA $F$ test, indicating that the real data are not highly correlated. This may explain the competitive performance of ANOVA $F$ test with PCT in this example.

More particular tree results are listed in Table 3, including tree size (the number of terminal nodes), the number of selected descriptors, and range and mean of the terminal node sizes. For the complete data, the Hotelling's $T^2$ and PCT trees had eight descriptors in common, and six of these eight also appeared in the ANOVA $F$ tree. In addition, the Hotelling's $T^2$ and ANOVA $F$ trees had nine common predictors; whereas the PCT and ANOVA $F$ trees had 24 predictors in common. For the 30% incomplete-case data, PCT retains 29% of the descriptors identified when it is used for the complete data, ANOVA $F$ test keeps 58%, while Hotelling's $T^2$ test only maintains 20%. Both the PCT and ANOVA $F$ tests identify more descriptors and the average terminal mode sizes are close with each other, but only half of that for Hotelling's $T^2$ test. The tendency to overfit the tree by PCT and ANOVA $F$ test may lead to some spurious splits. Ideally, it would be interesting to further explore the detected splits in terms of the underlying biology and chemistry. However, the data are proprietary and we cannot make this information available.

## 5. Conclusions

With a capability of handling nonlinear modeling, tree-structured methods are gaining popularity for analyzing large structure-activity data. MultiSCAM is one such algorithm, developed to uncover the structure-activity relationship for multivariate responses. The two-sample Hotelling's $T^2$ test faces the problem of $p \geq n_1 + n_2 - 1$ when node sizes are small, which is compounded by the existence of missing values, and hence lacks power in splitting the nodes. The two proposed alternatives, PCT and ANOVA $F$ test, overcome this problem. The simulation results indicate that both tests exhibit good performance in the selection of important descriptors and prediction. PCT is applicable to a wide range of covari-

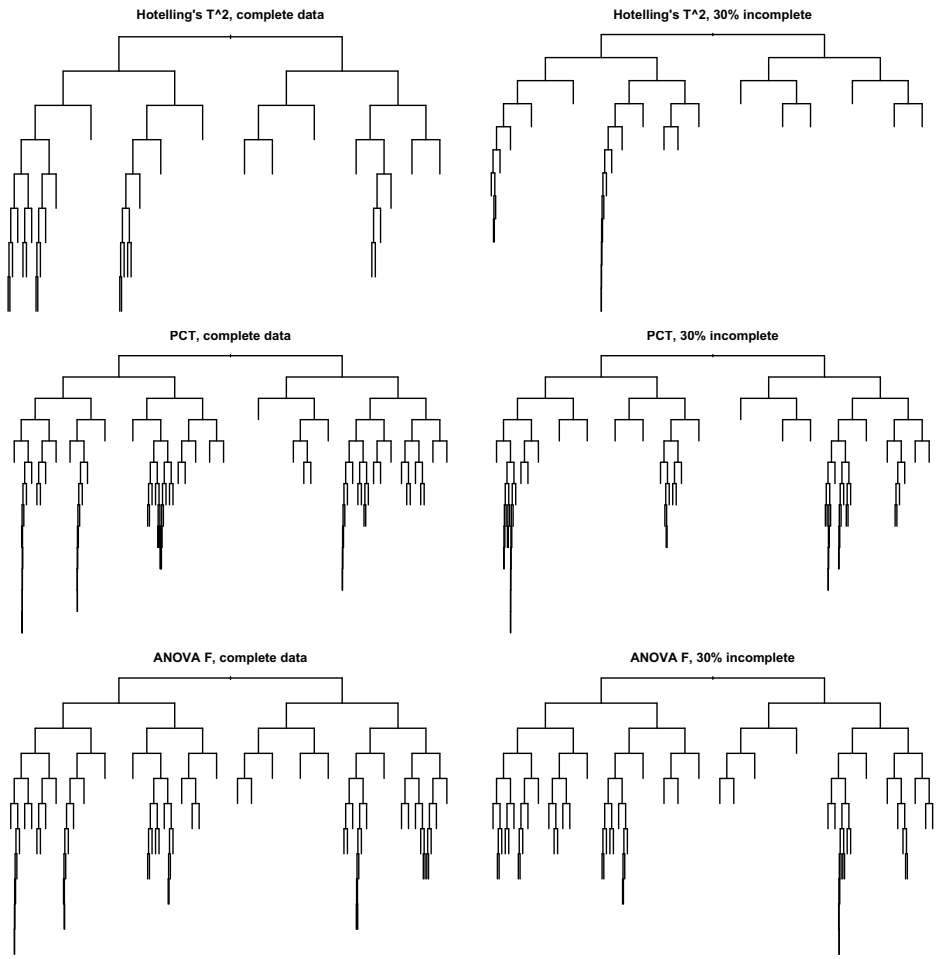



Figure 7. The tree dendrograms built with Hotelling's $T^2$, PCT and ANOVA $F$ test for both of the complete and incomplete (30%) drug discovery data.

ance structures. The ANOVA $F$ test splitting rule is derived under an assumption of independence and thus its performance degrades with increasing correlations; however, the impact is not severe for small correlations and thus it is a viable alternative in many applications.

Our results suggest the utility of hybrid tree-building algorithms that take advantage of the individual tests' strengths. The idea is to use Hotelling's $T^2$ test to split the nodes when sample sizes are sufficiently large. But as the tree grows and node sizes decrease, switch to either the PCT or ANOVA $F$ test depending on the amount of correlation detected in the data.

## Appendix: Expected Model Error

We now derive the expected model error for correct trees A and B. Both trees have nine terminal nodes. Denote the number of observations and the sample mean vector, which is also the predicted value for observations in the node, in terminal node $t$ by $n_t$ and $\overline{Y}_t$, respectively. When the tree is correct, then $R = 9$ is the true tree size and $Y_{t,i}^{(m)} = \mu_t$ for $i = 1, \ldots, n_t, t = 1, \ldots, R$. In this case,

$$\begin{aligned} E(\mathrm{ME}) &= \frac{1}{n}\sum_{t=1}^{R}\sum_{i\in\text{node } t} E\left\{(Y_{t,i}^{(m)} - \widehat{Y}_{t,i})^T \right. \\ &\qquad \left. \times\ \Sigma^{-1}(Y_{t,i}^{(m)} - \widehat{Y}_{t,i})\right\} \\ &= \frac{1}{n}\sum_{t=1}^{9} n_t E\{(\mu_t - \overline{Y}_t)^T \Sigma^{-1}(\mu_t - \overline{Y}_t)\} \\ &= \frac{1}{n}\sum_{t=1}^{9} n_t \mathrm{tr}(\Sigma^{-1}\Sigma/n_t) = \frac{9p}{n}. \end{aligned}$$

## Acknowledgments

We thank the editor, the associate editor, and two referees for several helpful comments that substantially improved the article. We also thank GlaxoSmithKline for financial support, and Chris Keefer and Chris Bizon in particular, for motivation and discussion of this research work.

*[Received November 2006. Revised November 2007.]*

## References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Boca Raton, FL: Chapman & Hall/CRC.

Hawkins, D. M., Young, S. S., and Rusinko, A. III. (1997), "Analysis of a Large Structure-Activity Data Set using Recursive Partitioning," *Quantitative Structure-Activity Relationships*, 16, 296–302.

Hawkins, D. M., and Kass, G.V. (1982), "Automatic Interaction Detection," in *Topics in Applied Multivariate Analysis*, ed. D. H. Hawkins, Cambridge: Cambridge University Press, pp. 269–302.

Hotelling, H. (1931), "The Generalization of Student's Ratio," *Annals of Mathematical Statistics*, 2, 360–378.

Keefer, C. E. (2001), "Use of Multivariate Data Mining Techniques in Pharmaceutical Systems Based Research," abstract of papers, *222nd ACS National Meeting*, Chicago.

Morgan, J. N., and Sonquist, J. A. (1963), "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58, 415–434.

Rusinko, A., III, Farmen, M. W., Lambert, C. G., Brown, P. L., and Young, S. S. (1999), "Analysis of a Large Structure/Biological Activity Data Set using Recursive Partitioning," *Journal of Chemical Information and Computer Sciences*, 39, 1017–1026.

Segal, M. R. (1992), "Tree-Structured Methods for Longitudinal Data," *Journal of the American Statistical Association*, 87, 407–418.

Wu, Y., Genton, M. G., and Stefanski, L. A. (2006), "A Multivariate Two-Sample Mean Test for Small Sample Size and Missing Data," *Biometrics*, 62, 877–885.

Zhang, H. (1998), "Classification Trees for Multiple Binary Responses," *Journal of the American Statistical Association*, 93, 180–193.

### About the Authors

Yujun Wu is Biostatistician, Biostatistics and Programming, Sanofi-Aventis, Bridgewater, NJ 08807 (E-mail: *yujun.wu@sanofi-aventis.com*). Marc G. Genton is Professor, University of Geneva, Department of Statistics, Texas A&M University, College Station, TX 77843-3143. Leonard A. Stefanski is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695.