Evaluation of Multilevel Decision Trees

Erwann Rogard* Andrew Gelman† Hao Lu+

October 21, 2003

Abstract

The evaluation of decision trees under uncertainty is difficult because of the required nested operations of maximizing and averaging. Pure maximizing (for deterministic decision trees) or pure averaging (for probability trees) are both relatively simple because the maximum of a maximum is a maximum, and the average of an average is an average. But when the two operators are mixed, no simplification is possible, and one must evaluate the maximizations and averagings in a nested fashion, following the structure of the tree. Nested evaluation requires large sample sizes (for data collection) or long computation times (for simulations).

An alternative to full nested evaluation is to perform a random sample of evaluations and use statistical methods to perform inference about the entire tree. We show that the most natural estimate is biased and consider three alternatives: normal-theory bias correction, the jackknife, and hierarchical Bayes inference. We explore the properties of these inferences through a simulation study and discuss general approaches to the problem.

Keywords: decision analysis, hierarchical Bayes, jackknife, nested computation

* Department of Statistics, Columbia University, New York, NY 10027
† Department of Statistics, Columbia University, New York, NY 10027, gelman@stat.columbia.edu, http://www.stat.columbia.edu/~gelman
+ Thales Corp., New York, NY
1. Introduction

1.1 The difficulty of evaluating decision trees

The standard paradigm for decision analysis under uncertainty is maximization of expected utility (see Luce and Raiffa, 1957, for a mathematical treatment and comparison to other axiomatic frameworks, and Clemen, 1996, for an applied introduction). A decision problem, or series of decision problems, is expressed as a tree with uncertainty nodes and decision nodes. The leaves of the tree are assigned utilities. (A leaf can itself represent a subtree, in which case the utility assigned to the leaf is the utility of that subtree.) At any decision node in the tree, the optimal decision is that which maximizes expected utility. The value of a tree is defined by averaging over uncertainty nodes and maximizing over decision nodes. The computation is most directly performed recursively, starting with the nodes adjacent to the leaves and working back to the root node.

The alternation between maximization and averaging makes the estimation of trees a qualitatively more difficult problem than, on the one hand, evaluating pure probability trees or, on the other, evaluating deterministic decision trees.

1.2 An idealized example

We illustrate with a simple idealized example. Suppose that there are two competing methods for teaching mathematics to elementary school children. Because of variation among teachers, students, local conditions, and so forth, method A is better in some school districts and method B is better in others. (Suppose that for practical reasons it is only possible to use one method in each school district, and performing "better" is defined as yielding higher average scores for the students on a particular standardized test.) Currently, the two methods are being applied with roughly equal frequency everywhere. A study is planned to evaluate the methods nationwide, with the goal of ultimately using, in each school district, the treatment that works there. Before this study is undertaken, however, it is desired to
estimate its potential consequences: (a) the proportion of school districts across the country for which each treatment is actually better, and (b) the potential average gain in test scores that would be obtained nationally if the locally-better method were applied in each district.

This problem has the structure of a decision tree (see Figure 1), in which the root node represents the school districts. This is an uncertainty node, in which the probability distribution over the school districts represents the desired national averages. For example, equal probabilities for the school districts yield an average that counts all districts equally; assigning each district a probability proportional to the number of students in the district yields an average that weights all students equally. The flexibility allowed in assigning probabilities corresponds to the different weightings that might be of interest.

Within each school district is a decision node representing the choice between the two teaching methods. Because the decisions will be made locally (each district chooses its own teaching method), the decision nodes are nested within the school districts.

Finally, within each decision branch is an uncertainty node representing the possible outcomes of the students in the district if they were given that particular teaching method. This last branch is indicated by a continuous sweep to indicate the (essentially) continuous nature of the outcome.

The tree can be evaluated by averaging within each decision branch, picking the best option within each school district, and then averaging over all districts in the country. It is the maximization step that presents a difficulty here. If we could simply average at all three steps, then evaluating the tree would be simple: just pick at random several districts, one decision at random within each district, and one student outcome at random within each district and decision option. The expectation of an expectation is an expectation, and hence this simple nonnested computation would produce an unbiased estimate of the value
of the tree.

However, with the maximization step, the inference is not so simple. But we would like to avoid a fully nested evaluation: even if we decided to use sampling at the outer level (that is, to work with a random sample of school districts), we would still need to gather potentially large samples of students for each of the two decisions at each of the sampled school districts.

1.3 Outline of the paper

An alternative to full nested evaluation is to perform a random sample of evaluations and use statistical methods to perform inference about the entire tree. In Section 2 of this paper we show that the most natural estimate is biased. Section 3 presents three alternatives: normal-theory bias correction, the jackknife, and hierarchical Bayes inference. In Section 4, we explore the properties of these inferences through a simulation study, and we conclude in Section 5 with suggestions for further work.

2 Evaluating decision trees using sampling

2.1 Model of the idealized example

We consider a finite sequence of decisions indexed by \( d \in D = \{1, \ldots, D\} \) with uncertain payoffs represented by the random variables \( y_d \) with means \( \mu_d \). Suppose we know nothing about the means, let alone their ordering. In order to determine the best decision \( d^* \) and its associated payoff \( \mu^* \) we sample one individual (or group of individuals) under each decision and compute simple estimates \( \hat{d} \) and \( \hat{\mu} \) as follows:

\[
\mu^* = \max_d \mu_d \quad \quad \hat{\mu} = \max_d y_d \\
\hat{d} = \arg \max_d y_d.
\]

We assume the means \( \{\mu_d\} \) can differ but that the distributions of the payoffs \( \{y_d\} \) overlap.
In the next section, we show how this key condition yields biased estimates, and under additional conditions we derive the biases analytically.

2.2 Bias of the simple estimate

2.2.1 For a given decision node

The condition that the distributions of \( \{y_d\} \) overlap is equivalent to asserting that the decision error, that is the indicator for whether a non-optimal decision is selected, is strictly greater than zero in expectation:

\[
\text{error} = \mathbb{E} \frac{y_d}{d 
eq d^*} > 0. \quad (1)
\]

It immediately follows that \( \hat{\mu} \) is biased upward strictly:

\[
\text{bias} = \mathbb{E} \left( \hat{\mu} - \mu^* \right) = \mathbb{E} \left( \left( \max_d y_d - y_{d^*} \right) (1_{d 
eq d^*}) \right) > 0. \quad (2)
\]

We evaluate (1) and (2), assuming the variables \( \{y_d\} \) are independent and normally distributed and, for convenience, with the number of decisions restricted to two, so that \( D = \{1, 2\} \):

\[
y_d - \mu_d \sim \text{iid } N(0, \sigma^2), \quad d \in D.
\]

We define the following parameter:

\[
q = \frac{1}{2} (\mu_1 - \mu_2), \quad \text{"decision effect"},
\]

and its simple estimate,

\[
\hat{q} = \frac{1}{2} (y_1 - y_2),
\]

which, in this case, has distribution \( N(\theta, \sigma_\theta^2) \), where \( \sigma_\theta^2 = \sigma^2 / 2 \).

We label \( \theta \) as the "decision effect" for the following reason: if we set decision 2 as the benchmark, then the net reward or "effect" of decision 1 is \( \theta \) (up to a constant of proportionality). The quantities \( \theta \) and \( \hat{\theta} \) are of interest because they completely determine error and bias:

\[
\text{error}(\hat{\theta}) = \mathbb{E}_{\sigma_\theta} \left( 1_{\text{sign}(\hat{\theta}) \neq \text{sign}(\theta)} \right)
\]

\[
\text{bias}(|\hat{\theta}|) = \mathbb{E}_{\sigma_\theta} \left( |\hat{\theta}| - |\theta| \right).
\]

The first equivalence is clear, and for the second it suffices to note that:
\[ \mu^* = \max_d \mu_d = (\mu_1 + \mu_2) / 2 + | \mu_1 - \mu_2 | / 2 \]
\[ \hat{\mu} = \max_d \hat{y}_d = (y_1 + y_2) / 2 + | y_1 - y_2 | / 2, \]
and that \((y_1 + y_2) / 2\) is an unbiased estimate of \((\mu_1 + \mu_2) / 2\).

Let \(\phi\) and \(\Phi\) denote the probability and cumulative density of the standard normal, and \(\gamma = | \theta | / \sigma_\theta\). The parameter \(1 / \gamma\) captures the degree of distribution overlap, so that the decision error and bias are increasing in \(1 / \gamma\).

\[
\text{error}(\hat{\theta}) = (1 - \Phi(\gamma)) \quad (3)
\]
\[
\text{bias}(| \hat{\theta} |) = \left| \theta \right| \left( \frac{2}{\gamma} \phi(\gamma) + 2 (\Phi(\gamma) - 1) \right)
= O(1 / \gamma) \text{ as } \gamma \to 0, \text{ for fixed } |\theta| \quad (4)
\]
\[
\text{var}(| \hat{\theta} |) = \theta^2 \left( 1 + \frac{1}{\gamma^2} - \left( \frac{2 \theta |\gamma|}{\gamma^2} + (2 \Phi[\gamma] - 1) \right)^2 \right)
= O(1 / \gamma^2), \text{ as } \gamma \to 0 \text{ for fixed } |\theta|. \quad (5)
\]

Figure 2a illustrates the first two relations above. For a given \(\gamma\), error is independent of \(| \theta |\) and bias is proportional to \(|\theta|\), so it suffices to vary \(\gamma\) (equivalently \(\sigma_\theta\)) to the described these relations.

[Place Figure 2 here]

2.2.2 Averaging over decision nodes

To make the tree complete it must have multiple nodes (or "strata") with \(\theta\) varying randomly across them. For tractability, the following assumption is made:

\(\theta \sim N(\mathbb{E}\theta, \sigma_\theta^2)\), and in particular: \(\mathbb{E}\theta = 0\).

Taking expectations of (1) and (2) with respect to the distribution of \(\theta\) gives:

\[
\mathbb{E}_{\sigma_\theta} \text{error}(\hat{\theta}) = \frac{1}{2} - \frac{\text{ArcTan}[\frac{\sigma_\theta}{\sigma_\theta}]}{\pi}
\]
\[
\mathbb{E}_{\sigma_\theta} \text{bias}(| \hat{\theta} |) = \sqrt{\frac{2}{\pi}} \sigma_\theta \left( \sqrt{1 + \left( \frac{\sigma_\theta}{\sigma_\theta} \right)^2} - 1 \right). \quad (7)
\]

The greater \(\sigma_\theta / \sigma_\theta\), the less likely an overlap so the smaller the error and bias, as reflected in the negative slopes for (6) and (7) (see Figure 2b).
As \( \sigma_\theta / \sigma_\check{\theta} \) approaches 0, the proportion of errors is worst, in this case as bad as deciding from a coin flip, yet the actual loss (max\( \mu_d \) - min\( \mu_d \)) = 2 | \( \theta \) | is zero since \( \sigma_\theta / \sigma_\check{\theta} = 0 \) implies that there is no preference for one decision over another. The actual loss, therefore, is maximized for a value of \( \sigma_\theta / \sigma_\check{\theta} \) greater than 0 so that the function is strictly concave.

3 Bias corrected estimates

3.1 Model

The preceding results provide a rationale for attempting to reduce the biases arising in using simple estimates. We will use the following model to construct and motivate three different bias correction methods.

Suppose there are \( J \) strata (which for simplicity have equal probability weights \( 1/J \)), within each of which are \( D \) decisions. For each combination of stratum and decision, we sample \( n \) observations, resulting in a total of \( JDn \) data points. For convenience we shall work with the case \( D = 2 \). Let \( y_{idj} \) denote the outcome for observation \( i \), decision \( d \) and stratum \( j \). Its location parameter, \( \mu_{dj} \), is the aggregation of a stratum payoff \( \beta_j \), a decision payoff \( \epsilon_{dj} \) nested within stratum \( j \), and a decision payoff common to all strata \( \alpha_d \). Using a natural notation, variability is controlled by \( \sigma_y \), \( \sigma_\epsilon \) and \( \sigma_\beta \), and the distributions are assumed to be normal.

\[
y_{idj} - \mu_{dj} \sim iid N(0, \sigma_y^2), \quad i \in \{1, 2, \ldots, n\}, \quad j \in \{1, \ldots, J\}, \quad d \in \{1, 2\} \\
\mu_{dj} = \alpha_d + \beta_j + \epsilon_{dj} \\
\beta_j \sim N(0, \sigma_\beta^2), \quad \epsilon_{dj} \sim N(0, \sigma_\epsilon^2).
\]

The sampling scheme implicit in this formulation is that any pair of observations corresponding to two different decisions must have been measured on different individuals, which is likely in an observational study.

To bridge Section 2.2 with this one, we use the same notation for the parameter of
interest—the decision effect—with subscript $j$ added to indicate local parameters.

$$
\theta_j = (\mu_1 - \mu_2) / 2 \quad \text{"decision effect in stratum } j\text{"}
$$

$$
\theta = (\alpha_1 - \alpha_2) / 2 \quad \text{"tree decision effect"}.
$$

Note that we clarified the meaning of "decision effect" in Section 2.2.1.

Our objective is to determine the proportion $P$ of strata for which the first decision is actually better, and the potential average gain $\Delta V$ that would be obtained for the whole tree, if the locally-better method were applied in each district. In terms of the parameters, this is:

$$
P = \mathbb{E}_{\theta_j > 0}
$$

$$
\Delta V = \mathbb{E} \max_d \mu_{dj} - \max_d \alpha_d = \mathbb{E} |\theta_j| - |\theta|.
$$

The relation $\theta_j = \theta + (\epsilon_1 - \epsilon_2) / 2$ implies:

$$
\theta_j \sim N(\theta, \sigma^2 / 2).
$$

In the remaining sections, we shall use $|\theta_j|$ unambiguously to denote the absolute value of the true parameter $\theta_j$. However, if $|.|$ is used in conjunction with a superscript or mark indicating an estimated parameter, say $|\hat{\theta}_j|$, then the latter means "estimate of $|\theta_j|$", not necessarily "the absolute value of $\hat{\theta}_j$", unless $\hat{\theta}_j$ was defined explicitly.

### 3.2 Simple estimates

We define the decision effects for stratum $j$ and for the entire tree as,

$$
\hat{\theta}_j = (\bar{y}_{1j} - \bar{y}_{2j}) / 2
$$

$$
\hat{\theta}^s = \Sigma_j \hat{\theta}_j / J.
$$

Using these definitions, the local and tree biases are,

$$
\text{bias}(|\hat{\theta}_j|) = \mathbb{E}_{\sigma^2_j}(|\hat{\theta}_j|) - |\theta_j|
$$

$$
\text{bias}(|\hat{\theta}^s|) = \mathbb{E}_{\sigma^2}(|\hat{\theta}^s|) - |\theta|.
$$
Combining local and tree estimates yields estimates for $P$ and $V$ (as defined in Section 1.2):
\[
\hat{P}^s = \sum_j 1_{\hat{\theta}_j > 0} / J \\
\Delta V^s = \sum_j |\hat{\theta}_j| / J - |\hat{\theta}^s|.
\]

According to the results of Section 2.2.1, for a given stratum and assuming $|\theta_j| = 1$, the variance ratio, $\gamma_j = |\theta_j| / \sigma_{\hat{\theta}^s_j}$ determines bias$|\hat{\theta}^s_j|$ (decreasing in $\gamma_j$). For an arbitrary $|\theta_j|$, bias is obtained by multiplying bias$|\hat{\theta}^s_j|$ under $|\theta_j| = 1$ by $|\theta_j|$. The average across strata, however, is unknown analytically (except for restrictive assumptions, as in 2.2.2) but will be determined the simulation study. For the tree decision problem, however, bias is known exactly by application of the results in Section 2.2.1.

In the next section we discuss the connection between $\sigma_{\theta_j}$, $\sigma_{\hat{\theta}^s_j}$ and the original parameters. However, we have already seen that $\sigma_{\beta}$ does not affect the bias because the $\theta_j$'s are independent of the $\beta_j$'s. The addition of a stratum-and-decision interaction term would make the bias dependent on $\sigma_{\beta}$ and a correlation parameter. The above choice of simple local estimate $\hat{\theta}^s_j$ seems the most obvious. However, since bias is an increasing function of the variance of the simple estimate, reducing the latter by using a weighted combination of the sample local effect $\hat{\theta}^s_j$ and the sample tree effect $\hat{\theta}^s$ would seem a viable strategy. As this source of bias is reduced, however, another arises because $\hat{\theta}^s_j$ and $\hat{\theta}^s$ are different in expectation, resulting in a net effect that could be worse or better depending on the weights attached to $\hat{\theta}^s_j$ and $\hat{\theta}^s$. More precisely, an optimal pooling would probably weight $\hat{\theta}^s$ more heavily as the distribution overlap increases, i.e. as $\sigma_{\gamma} / \sigma_\epsilon$ is higher. This is most naturally achieved by the Bayesian estimate but could potentially be implemented within the class of simple estimates.
3.3 Normal theory bias correction

Let $g(., .)$ be the bias function defined in Section 2.2, and recall that the arguments to this function are the absolute mean and variance of the sample estimate. In terms of the extended model, and given the true parameters, the local and tree biases are given as:

$$\text{bias}(|\hat{\theta}_j^s|) = g(|\theta_j|, \sigma_{\theta_j}^2), \quad \text{and} \quad \sigma_{\theta_j}^2 = \sigma_y^2 / 2n$$

$$\text{bias}(|\hat{\theta}_j^s|) = g(|\theta|, \sigma_{\theta}^2), \quad \text{and} \quad \sigma_{\theta}^2 = \frac{\sigma_y^2}{2J} + \frac{\sigma_e^2}{2nJ}.$$ 

In the absence of true parameters, the bias corrected estimates, $|\hat{\theta}_j^c|$ and $|\hat{\theta}^c|$, are defined by subtracting from the simple estimates $|\hat{\theta}_j^s|$, $|\hat{\theta}_j^s|$, the bias correcting function $g(., .)$ with appropriate arguments substituted by their sample estimates:

$$|\hat{\theta}_j^c| = |\hat{\theta}_j^s| - g(|\hat{\theta}_j^s|, \hat{\sigma}_{\theta_j}^2)$$

$$|\hat{\theta}^c| = |\hat{\theta}_j^s| - g(|\hat{\theta}_j^s|, \hat{\sigma}_{\theta}^2),$$

where $\hat{\sigma}_{\theta_j}^2$ and $\hat{\sigma}_{\theta}^2$ denote the estimates for $\sigma_{\theta_j}^2$ and $\sigma_{\theta}^2$, derived from estimates of $\sigma_y$ and $\sigma_e$. The analysis of variance in Table 1 shows how $\sigma_y$ and $\sigma_e$ are estimated.

[Place Table 1 here]

Reliance on sample estimates (even though $g(., .)$ is known exactly) is liable to the following difficulty: the greater $\hat{\sigma}_{\theta_j}^2$, the greater the bias using simple estimates, yet the less precise $|\hat{\theta}_j^s|$ as an estimate of $|\theta_j|$, and therefore the less reliable bias reduction. For similar reasons, the need for bias correction in estimating the tree effect $|\theta|$ will be (comparatively) small. From now on, therefore, we focus on estimating $|\theta_j|$.

The exact properties of bias corrected estimates are difficult to obtain, even by resorting to approximations. For a given stratum (the subscript $j$ is dropped for simplicity) assume $\hat{\sigma}_{\theta}^2 = \sigma_{\theta}^2$, which is reasonable if $J$ is large, such that $g(|\hat{\theta}_j^s|, \hat{\sigma}_{\theta}^2) = g(|\hat{\theta}_j^s|)$ and:

$$|\hat{\theta}_j^c| = |\hat{\theta}_j^s| - g(|\hat{\theta}_j^s|) \approx |\hat{\theta}_j^s| \left(1 - \nabla g(\mathbb{E}|\hat{\theta}_j^s|)\right) + \nabla g(\mathbb{E}|\hat{\theta}_j^s|)\mathbb{E}|\hat{\theta}_j^s|.$$
This implies:

$$\text{bias}(|\hat{\theta}_j^c|) \approx \text{bias}(|\hat{\theta}_j^s|).$$

This approximation will not be helpful for small \(n\). Similarly the approximation to variance does not lend itself to a straightforward analysis:

$$\text{Var}(|\hat{\theta}_j^c|) \approx \text{Var}(|\hat{\theta}_j^s|) \left(1 - \nabla g(\mathbf{E} |\hat{\theta}_j^s|)\right)^2.$$ 

Each term on the right hand side is complicated but known exactly:

$$\text{Var} |\hat{\theta}_j^s| = (\theta_j)^2 \left(1 + \left(\frac{\sigma_{\theta_j}}{\sigma_{\theta_j}}\right)^2 - \left(\frac{\sigma_{\theta_j}}{\sigma_{\theta_j}}\right)^2 \phi\left(\frac{\theta_j}{\sigma_{\theta_j}}\right) + \left(2 \Phi\left(\frac{\theta_j}{\sigma_{\theta_j}}\right) - 1\right)^2\right)$$

$$\mathbf{E} |\hat{\theta}_j^s| = g(|\theta_j|) + |\theta_j|$$

$$g(|x|) = |x| \left(\frac{x}{\sigma_{\theta_j}} \phi\left(\frac{x}{\sigma_{\theta_j}}\right) + 2 \left(\Phi\left(\frac{x}{\sigma_{\theta_j}}\right) - 1\right)\right)$$

$$\nabla g(|x|) = 2 \left(\Phi\left(\frac{x}{\sigma_{\theta_j}}\right) - 1\right).$$

The next step, taking the expectation across strata seems at best cumbersome. This underlines the need to derive the properties of such estimators empirically.

### 3.4 Jackknife estimation

A standard method for reducing bias relies on the jackknife (see, for example, Shao and Tu, 1996); if the initial bias is of order \(\frac{1}{n}\) the jackknife reduces it to order \(\frac{1}{n^2}\). Moreover, unlike the previous method, it requires no parametric assumptions. The methodology is as follows.

Let \(T = t(\{y_1, \ldots, y_n\})\) denote a statistic applied to an iid sample of size \(n\), \(T_{-i} = t(\{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\})\) the same statistic after excluding observation \(i\), and \(\bar{T} = \Sigma_i T_{-i} / n\). The jackknife estimator is defined as \(\hat{T} = nT - (n-1)\bar{T}\). In general \(T\) is an estimator for \(g(\theta)\), where \(\theta\) denotes some distribution parameter. It is common, therefore, that \(t(.) = g(f(.))\) where \(f(.)\) is an estimator of \(\theta\).

Presently the parameter of particular interest is \(|\theta_j|\), the within stratum decision effect. For simplicity we omit subscript \(j\). The simple estimate \(|\hat{\theta}^s|\) depends on two iid samples, \(\{y_{11}, \ldots, y_{1n}\}\) and \(\{y_{21}, \ldots, y_{2n}\}\), one for each decision. By combining these two
samples, an iid sample of pairs, \{(y_{11}, y_{21}), ..., (y_{1n}, y_{2n})\}, is created such that every observation appears only once and the order within each pair corresponds to the decision number. Letting \(y_i = (y_{i1} - y_{i2})/2\) yields \(\{y_1, ..., y_n\}\) which is iid, as required. We recognize that 

\[ T = |\tilde{\theta}| = g(f(y_1, ..., y_n)), g(.) = |.|/2 \]

and \(f(.)\) is the sample mean, so that a jackknife estimator \(|\tilde{\theta}|\) can be evaluated. The validity of this approach rests on the assumption of independence between the two original samples, a consequence of the sampling scheme (Section 3.1).

The jackknife above arbitrarily selects one among many equally valid permutations, suggesting it could be improved by combining all of them. To motivate this we adapt the one-sample example of Shao and Tu (1996, p.5) to two independent samples. Let \(\theta = (\mu_1, \mu_2)\), the true means in each sample, \(g(\theta) = (\mu_1 - \mu_2)^2\), and \(T = (\bar{y}_1 - \bar{y}_2)^2\). The resulting bias is positive:

\[
\text{bias}(T) = \mathbb{E}(\bar{y}_1 - \bar{y}_2)^2 - (\mu_1 - \mu_2)^2 = \frac{1}{n} (\sigma_1^2 + \sigma_2^2) \geq 0,
\]

where \(\sigma_d^2\) denotes the variance under distribution \(d\). Let \(T_{i,j} = g(y_{1i}, y_{2j})\) and \(b_{ij} = (n - 1) (T_{i,j} - T)\) denote a bias estimator. It is easily shown that \(\text{bias}(T) = \mathbb{E}(b_{ij}^*)\).

Any linear combination of the following form is therefore unbiased:

\[
\tilde{T} = T - \Sigma_{i,j} w_{i,j} b_{ij}^*, \quad \text{and} \quad \Sigma_{i,j} w_{i,j} = 1.
\]

The objective, therefore, is to minimize the variance of \(\tilde{T}\) under the above constraint. An equivalent formula is \(\tilde{T} = \Sigma_{i,j} w_{i,j} \tilde{T}_{i,j}\), where \(\tilde{T}_{i,j} = (n T - (n - 1) T_{i,j})\), so that the variance of \(\tilde{T}\) is \(w' V w\) where \(V\) denotes the \(n^2 \times n^2\) variance matrix of the \(\tilde{T}_{i,j}\)'s. Using the Lagrange multiplier method to solve the above problem yields a system of linear equations. Using an LU decomposition of the latter it is found that for all \((i, j)\), \(w_{i,j} = 1/n^2\), for \(n = 2\) and \(n = 3\). This result is free of distribution assumptions and depends only on the fact that there are at most three distinct values for the covariance terms, such that the matrix is made of blocks with identical patterns.

To validate this "two-sample jackknife" it would be necessary to generalize it to
arbitrary $n$ and bias function. However, for the present purpose i.e. $g(\theta) = |\mu_1 - \mu_2| / 2$, we notice the relationship with $g(\theta) = (\mu_1 - \mu_2)^2$ and verify empirically (not shown here) for $n = 3$, that, as expected, bias is identical to the "naive" jackknife above and its variance is lower. We therefore use the second method.

Properties of jackknife estimators are usually found by first approximating $T - g(\theta)$ by its second order Taylor expansion at $\theta$. The general result is that bias reduction is accompanied by variance increase, and mean squared error is not necessarily decreased. In the present case the Taylor approach is not straightforward so we use instead a simulation study (Section 4).

### 3.5 Bayesian estimation

For locally-better decisions, as mentioned in Section 3.2, a good property for an estimator would be to distribute increasing weight on data pooled across strata as $\sigma_y / \sigma_\epsilon$ increases. Here, we describe a hierarchical Bayes analysis to do this, using a flat prior distribution, $p(\alpha, \sigma_y, \sigma_\beta, \sigma_\epsilon) \propto 1$.

Thanks to the normal assumption, the posterior distribution of each parameter are known, allowing to sample from them directly. Following Gelman et al. (2003), we simulate multiple Gibbs sequences with overdispersed starting points, select the number of iterations $L$ on the basis of potential scale reduction $R$, and keep only the second half of the iterations. In this case, approximate convergence is reached after $L = 40$ iterations.

The local and tree decision effects at iteration $l$, $\hat{\theta}_j^l$ and $\hat{\theta}^l$ for $l = 1, ..., L$, are derived from the original (sampled) parameters. True values are estimated by the average:

$$\hat{\theta}_j^b = \frac{\Sigma_l \hat{\theta}_j^l}{L}$$
$$\hat{\theta}^b = \frac{\Sigma_l \hat{\theta}_j^l}{L}.$$

It is then straightforward to estimate $P$ and $V$, in a way identical to simple estimates:

$$\hat{P}^b = \frac{\Sigma_j \hat{\theta}_j^b}{J} - \hat{\theta}^b.$$
\[ \Delta \hat{V}^b = \Sigma_j | \hat{\theta}^b_j | / J - | \hat{\theta}^b |. \]

The posterior mean of the payoff for a given decision \( d \) and stratum \( j \), \( \mu_{dj} \), is expressed as a weighted average of the prior mean \( (\alpha_d + \beta_j) \) and the observed mean \( \bar{y}_{d,j} \), with weights proportional to the precisions, respectively \( 1/\sigma^2_e \) and \( n/\sigma^2_y \). If \( \sigma^2_e = 0 \), the model pools observations from all strata. Conversely, if \( \sigma^2_e = \infty \), the model reduces to the simple estimates, and inference is made in each stratum separately.

### 4 Simulation study

#### 4.1 Description

For a given set of parameters \( \alpha \), \( \sigma_e \), \( \sigma_y \), \( \sigma_\beta \), and \( n \), \( J \) fixed, two stages are involved in the sampling process: first the mean parameters (the \( \beta_j \)'s and \( \epsilon_{dj} \)'s) and secondly the observed data (the \( y_{idj} \)'s) must be sampled. The first stage is repeated, independently, for \( m = 1, \ldots, M \). For each \( m \), we draw \( K \) independent second-stage samples. To summarize:

\[ \beta^m_j \sim \text{iid } N(0, \sigma^2_\beta) \quad \epsilon^m_{dj} \sim \text{iid } N(0, \sigma^2_e) \quad m = 1, \ldots, M \]

\[ y^{m,k}_{idj} \sim \text{iid } N(\mu^m_{dj}, \sigma^2_y) \quad \mu^m_{dj} = \beta^m_j + \epsilon^m_{dj} \quad k = 1, \ldots, K. \]

From these, the quantities of interest are derived, specifically the true decision effects in each stratum \( (\theta_j) \), and the estimators for the absolute, and (if applicable) signed decision effects, \( | \hat{\theta}_j | \) and \( \hat{\theta}_j \). We examine the following performance measures:

\[
\text{bias}( | \hat{\theta}_j | ) = \mathbb{E}_{\sigma_e} \mathbb{E}_{\sigma_y} ( ( | \hat{\theta}_j | - | \theta_j | ) | \theta_j ) \\
\text{sd}( | \hat{\theta}_j | ) = ( \mathbb{E}_{\sigma_e} \mathbb{E}_{\sigma_y} ( ( | \hat{\theta}_j | - \mathbb{E}_{\sigma_y} | \hat{\theta}_j | )^2 | \theta_j ) )^{1/2} \\
\text{error}( \hat{\theta}_j ) = \mathbb{E}_{\sigma_e} \mathbb{E}_{\sigma_y} ( ( 1_{\text{sign}(\hat{\theta}_j) = \text{sign}(\theta_j)} | \theta_j ) \\
\text{loss}( \hat{\theta}_j ) = \mathbb{E}_{\sigma_e} \mathbb{E}_{\sigma_y} ( ( 1_{\text{sign}(\hat{\theta}_j) = \text{sign}(\theta_j)} 2 ( | \theta_j | ) ) | \theta_j ).
\]

Sample versions are obtained by replacing \( \mathbb{E}_{\sigma_e} \) and \( \mathbb{E}_{\sigma_y} \), respectively by \( \Sigma_{m=1}^M \Sigma_{j=1}^J \) and \( \Sigma_{k=1}^K \), and letting \( M, K \to \infty \). In our simulations, we set \( M = 20 \) and \( K = 20 \), and \( \sigma^2_y \) and \( \sigma^2_e \) are within 1±0.5 of their true values. For simple and jackknife estimates, it is not imperative to sum over \( m \); summing only over \( j \) and letting \( J \to \infty \) would suffice. However, bias corrected
estimators (through \( \delta \theta_j \)) and the Bayesian approach pool across all strata, such that their performance depends on \( J \).

Since the objective is to assess \( P \) and \( \Delta V \), it is tempting to estimate the biases \( \| \mathbb{E} \hat{P} - P \| \) and \( \| \mathbb{E} \Delta \hat{V} - \Delta V \| \) directly. This may lead to false conclusions. A small value for \( \| \mathbb{E} \hat{P} - P \| \) could arise even though error(\( \hat{\theta}_j \)) may be large, for example, if \( P = 1/2 \) and the sample classifier systematically picks the wrong decision.

4.2 Results

We repeat this process for each of a set of combinations of parameters. Specifically, we consider \( \theta = 0 \) and \( \theta = 0.1 \) and, for each, we increase \( \sigma_e \) and \( \sigma_y \) from 0.1 to 1, with the other parameters held constant. We set \( \sigma_B \) to some arbitrary value, which is irrelevant here as it affects neither error nor bias under the present model that contains no interaction term (see Section 3.2). Throughout, we set \( n = 3 \) and \( J = 20 \), corresponding to a small amount of data in each cluster and a moderate number of clusters.

The case \( \theta = \mathbb{E} \theta_j = 0 \), equivalently \( \alpha_A = \alpha_B = 1 \), corresponds to decisions that are equally good on average (although not necessarily in any given cluster), implying \( \Delta V = \mathbb{E} | \theta_j | \). The bias of all estimators is naturally increasing in \( \sigma_y^2 \) (twice the variance of the simple estimate), and decreasing in \( \sigma_e^2 \) (reducing overlap). The jackknife and bias corrected estimates have greater variance than the simple version. The standard deviation of the Bayesian estimator (\( \hat{\theta}^b_j \)) is the smallest because any increase in \( \sigma_y / \sigma_e \) is partially offset by greater data pooling. Overall, the Bayesian estimate is better in terms of MSE for all values of \( \sigma_y \) and \( \sigma_e \). These relationships are illustrated in Figure 3.

[Place Figure 3 here]

In Figure 4, the error and loss functions are nearly identical for Bayesian and simple estimators. Since \( \theta = 0 \), its posterior estimate \( \hat{\theta}^b \) is approximately 0. Since the within-stratum
posterior estimate \( \hat{\theta}^b \) is a weighted average of \( \hat{\theta}^s \) and \( \hat{\theta}^b \), it will differ from \( \hat{\theta}^s \) only by a constant of proportionality (which depends on \( \sigma_y / \sigma_\epsilon \)). The concave shape of loss, contrasting with the increasing pattern for error is consistent with our comments at the end of Section 2.2.2.

[Place Figure 4 here]

We repeat the analysis for \( \theta = \mathbb{E}\theta_j = 0.1 \), for which there is a global average difference between the two decisions. As verified by inspection of Figure 5, patterns for bias, variance and MSE are similar to those encountered in the first case. Figure 6 shows that contrary to the previous case, error and loss curves indicate that the Bayesian estimator is at least as good as the simple estimator, and notably better for small \( \sigma_\epsilon / \sigma_y \). Indeed, the within stratum posterior decision is no longer proportional to the simple version as it also incorporates a non-zero posterior tree-decision effect.

[Place Figure 5 and 6 here]

5. Discussion

We have seen how simple estimation is used to perform inference about an entire decision tree and the extent to which bias in estimating the best decisions' payoff and error from picking the best decision are unavoidable. Three alternatives have been proposed, normal-theory bias correction, the jackknife, and hierarchical Bayes inference, and their relative performance assessed in a simulation study. They all yield smaller biases than the simple estimate. However, the increase in variance for the jackknife and the bias corrected estimate is such that the Bayesian estimator performs better in terms of MSE across the entire spectrum of values for \( \sigma_\epsilon / \sigma_y \). Error and loss for the Bayesian estimator are as good as or better (for small \( \sigma_\epsilon / \sigma_y \)) than the simple versions.
Another advantage of the Bayesian method is the relative ease with which it can adapt to more complex models. For example, the tree structure can be refined in a number of ways. As hinted at previously, one could add a decision by stratum interaction. The methodology for simple and jackknife estimates would be unaffected. For the Bayesian method it would suffice to incorporate the new parameters into the model. Normal theory bias corrected estimates could also be implemented by relying on a modified decomposition of variance.

Payoffs were treated as continuous variables that can be physically measured. Often, there may be context where obtaining such measurements is costly, such that data is only available in the form of ordered categories. This situation can also arise, for example, in a survey in which agents have to rate their perceived payoff on a scale, thus implying some individual to individual variation in mapping the true payoffs to their subjective ratings. Such complexities could be accounted for, in the Bayesian framework, by adding latent variables, less obviously so by using simple estimates, let alone the other two alternatives.

The present study has focused on binary decisions but, realistically, agents often face multiple choices. While implementing normal theory bias correction would become a little cumbersome, the simple estimate and the other bias correcting methods would still be applicable. However, the procedures would be flawed if there existed an ordering among decisions such that the mean payoff would depend on the decision covariate through some unknown function. More precisely, the maximization step \( \hat{\mu}^* \) ignores the ordering among decisions and is altogether inapplicable if the decision domain \( \mathcal{D} \) is not finite, let alone uncountable. Instead, the function itself must be estimated in order to derive its maximum.

Finally, agents may be facing sequential problems. In this case, decisions are made in a specified order, contributing at each stage to the overall loss and at the same time reveal-
ing information about the system. For example, in a clinical trial where patients are recruited sequentially, the decision at each stage is to decide whether the cost of additional testing will be compensated by the incremental information resulting from it. In a sequential setting, there are further levels of nesting in the decision tree, and our procedures would need to be generalized to a more complex family of hierarchical models.

References


Luce, R.D., and Raiffa H. (1957), Games and Decisions, John Wiley & Sons.

Figure 1: Illustration of a decision tree. In the example given in the introduction, each stratum is a district, and decisions A and B are two different teaching programs.

<table>
<thead>
<tr>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>J  -  1</td>
<td>$SS_B = n \sum_j (\hat{\theta}_j^s - \hat{\theta}_j^q)^2$</td>
<td>$MS_B = SS_B / (J - 1)$</td>
<td>$n \frac{\sigma_\theta^2}{2} + \frac{\sigma_q^2}{2}$</td>
</tr>
<tr>
<td>J(n - 1)</td>
<td>$SS_W = \sum_{i,j}(\hat{\theta}<em>{ij}^s - \hat{\theta}</em>{ij}^q)^2$</td>
<td>$MS_W = SS_W / (J(n - 1))$</td>
<td>$\frac{\sigma_\theta^2}{2}$</td>
</tr>
</tbody>
</table>

Table 1: Analysis of variance for the simple estimate. In terms of the original variables, $\hat{\theta}_{ij}^s = (y_{i1j} - y_{i2j}) / 2$, $\hat{\theta}_j^q = (\bar{y}_{1j} - \bar{y}_{2j}) / 2$ and $\hat{\theta}_j^q = (\bar{y}_1 - \bar{y}_2) / 2$.

Figure 2: (left) Decision error and bias as a function of within stratum standard deviation $\sigma_\theta^2$. The parameter $\sigma_\theta^2$ determines the degree of overlap between two decisions, as reflected in the positive slopes; (right) $\theta$ is random across strata with variance $\sigma_\theta^2$, and $\sigma_\theta^2 = 1$. Overlap is controlled by $1 / \sigma_\theta^2$ resulting in negative slopes.
Figure 3: Bias, standard deviation, and root mean square error for four estimators of locally-best decisions, based on a simulation study in a hierarchical decision structure. We assume zero mean decision effect and vary within stratum variance, $\sigma_y$ (left), or the variance of the decision means, $\sigma_\epsilon$ (right), from 0 to 1.
Figure 4: Error and loss, for simple and Bayesian estimators of locally-best decisions, based on a simulation study in a hierarchical decision structure. We assume zero mean decision effect and vary within stratum variance, $\sigma_{y}$ (left), or the variance of the decision means, $\sigma_{\epsilon}$ (right), from 0 to 1. The close similarity of the Bayesian and simple estimates is a consequence of $\theta = 0$. Specifically, the posterior within stratum decision effect is proportional (up to sampling variation) to the simple version, so the same decision is picked in most strata. Even though error is highest for $\sigma_{\epsilon} = 0$, loss is zero for that value because in this case no decision is preferred to the other. In the other extreme, as $\sigma_{\epsilon} \to \infty$, errors vanish, but wrong decisions carry heavy losses, resulting in the concave shape.
Figure 5: Same as Figure 3 except the decision mean is greater than zero. This implies a reduced overlap between the two decision payoffs, resulting in lower bias.
Figure 6: Same as Figure 4 except the decision mean is greater than zero. This new assumption implies that the posterior decision effect within each stratum is no longer proportional to the simple version i.e. it also incorporates the tree-decision effect. As a result, the Bayesian estimator is at least as good as the Bayesian version, and notably better for small $\sigma_e / \sigma_y$ (visible on the right side).