Evaluation of multilevel decision trees*

Erwann Rogard† Andrew Gelman‡ Hao Lu§

September 8, 2005

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 maximization and averaging operations 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 two alternatives: the parametric bootstrap, and hierarchical Bayes inference. We explore the properties of these inferences through a simulation study.

AMS classification: 62C10

Keywords: decision analysis, hierarchical Bayes, nested computation.

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 action 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.

*To appear in the Journal of Statistical Planning and Inference. We thank the reviewers for helpful comments and the NSF for financial support.

†Department of Statistics, Columbia University, New York, NY 10027, er317@columbia.edu

‡Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027, gelman@stat.columbia.edu, http://www.stat.columbia.edu/∼gelman/

§Thales Corp., New York, NY, hlu@thales.com

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 children. Because of variation among teachers, students, local conditions, and so forth, method 1 is better in some school districts and method 2 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.) A study is planned to evaluate the methods nationwide, with the goal of ultimately using, in each school district, the treatment that works best. Before this study is undertaken, however, it is desired to estimate its potential consequences: (a) determine which treatment is best in each district (classification problem), and (b) estimate the test scores that would be obtained nationally if the locally-better method were applied in each district (estimation problem).

This problem has the structure of a decision tree (see Figure 1), in which circles and squares represent uncertainty and decision nodes, respectively. In our example the random root node is linked to all decision nodes, one for each district. In turn, each decision node is linked to two random nodes, one for each of the teaching methods (1 or 2), indicating the randomness of the student's scores. Finally, the continuous nature of the latter is shown by a continuous sweep. The random root node indicating that, for a given set of decisions, the value of the tree is obtained by averaging over the districts' payoffs, with weights (represented by the branches) that can be arbitrary (e.g. equal weights) or indeed reflect some truly random process (e.g. the proportion of the districts' student population relative to the national total). The value of the tree represents the average test score nationwide, or equivalently the expected test score of a student picked at random.

[Figure 1]

The optimal value of the tree can be evaluated by averaging within each decision node, picking the best option within that node, 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 non-nested 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 we formulate the problem. In Section 3, the simple and bootstrap estimators are presented. In Section 4, we consider the Bayesian approach.

2 Problem formulation

2.1 Notation and model

We now present our hierarchical decision problem formally:

$$y_{ijk} \sim N(\theta_{ij}, \sigma_y^2)$$

$$\theta_{ij} \sim N(\alpha_j, \sigma_\theta^2)$$

with the y's and θ's representing, data and expected data, respectively and $\{i, j, k\} \in \{1, \ldots, I\} \times \{1, \ldots, J\} \times \{1, \ldots, K\}$ indexing, respectively, stratum (e.g. district), action (e.g., teaching method), and sample observation. Our model has a hyper-parameter–parameter–data structure, which we denote $\xi = (\psi, \theta, y)$. Here, ψ represents the α's and the σ's, and θ is the vector of θ's. We further break θ into θ_i's to indicate parameters within the I strata.

To conform with a decision analysis framework, we would have to specify a utility function. Here, for simplicity, we set the utility associated with action j in stratum i to the expected outcome θ_{ij}. All parameters are assumed unknown. In particular, distributions of payoffs overlap under different actions. Otherwise the optimal action in each stratum is trivially obtained.

The sampling scheme implied above is that any pair of observations corresponding to two different decisions must have been measured on different individuals. In all, there are IJK data points.

2.2 Classification and estimation problems

We now make precise the classification and estimation problems encountered in the introduction. To do so, we define the value of the tree under generic actions $(\omega_1, ..., \omega_I)$ (by "generic", we mean with the action space unspecified) as

$$V(\omega_1, ..., \omega_I) = \sum_{i=1}^{I} p_i v(\omega_i),$$

where p_i indicates the weighting assigned to stratum i and $v(\omega_i)$ denotes the expected payoff in stratum i under action ω_i. We assume a uniform probability weighting, $p_i = 1/I$, but in general these probabilities could differ, for example being proportional to the number of students in each school district i.

In the *classification* problem, $\omega_i \equiv j_i \in \{1, \ldots, J\}$, and $v(j_i) = \theta_{ij_i}$, the expected payoff under action j_i. Note that we can alternatively define $v(.)$ as the indicator for whether or not the best action is picked i.e. $v(j_i) = 1_{[j_i = j_i^*]}$, where $j_i^* = \arg\max_j \theta_{ij}$. The objective is to maximize $V(.)$, which under perfect knowledge, yields the optimal tree value

$$V^* = V(j_1^*, \ldots, j_I^*) = \max_{j_1, \ldots, j_I} V(j_1, \ldots, j_I)$$

In the *estimation* problem, $\omega_i \equiv \theta_i \in \mathbb{R}$, with $v(\theta_i) = \theta_i$, and the objective is to equate the latter with $\theta_i^* = \max\{\theta_{i1}, \ldots, \theta_{iJ}\}$.

We consider the following local and aggregate losses:

$$loss(\omega_i|\theta_i) = |v(\omega_i) - v(\omega_i^*)|^p$$

$$loss(\omega|\theta) = \sum_{i=1}^{I} loss(\omega_i|\theta_i)/I$$

where $\omega = (\omega_1, \ldots, \omega_I)$. Defining the loss structure this way avoids situations where bad actions at the stratum level result in small aggregate losses, which would be awkward. We will give particular attention to bias: $bias(\omega_i|\theta_i) = v(\omega_i) - v(\omega_i^*)$.

A default choice for the estimation problem is $p = 2$. For the classification problem we set $p = 1$, which yields the following interpretation: $loss(.)$ is the shortfall in payoff from a particular action, relative to the optimal decision. If, however, we set $v(.)$ to the previously defined indicator function, our loss is the error, which we denote $error(.)$.

For each problem, our objective is to define a decision rule, i.e. a mapping, $\omega(.)$, from data, y, to the ω_i's, which minimizes the loss averaged over the parameter-data distribution (De Groot, 1970):

$$min_{\omega(.)} E(E(loss(\omega(y)|\theta)|\theta)).$$

In the classification problem, the solution is $j_i(y) = \arg\max_j E(\theta_{ij}|y)$, and in the estimation problem, it is $\theta_i(y) = E(\max_j \theta_{ij}|y)$.

3 Simple and bootstrap estimates

The preceding section set up the problem probabilistically, which naturally suggests a Bayesian decision process. However, for the sake of comparison, we consider the most natural empirical estimate, followed by a parametric bootstrap procedure that adjusts for the bias of the simple estimate.

3.1 Simple estimates

The most natural empirical estimate is obtained by separately estimating each θ_{ij} by its sample estimate, $\widehat{\theta}_{ij} = \overline{y}_{ij.}$ and then solving the classification and estimation problems, $\widehat{j}_i^* = \arg\max_j\{\widehat{\theta}_{ij}\}$, and $\widehat{\theta}_i^* = \arg\max_j\{\widehat{\theta}_{ij}\}$. In general, a nonlinear transformation of an unbiased estimator yields a biased estimator of the transformed parameter. In particular, since the max(.) function is convex and the distributions of data overlap under different actions, the expected bias is strictly positive for the estimation problem. For the classification problem, the expected error is strictly positive.

In order to study the properties of the estimators in more detail, we simplify the problem by assuming there are only $J = 2$ competing actions. Then

$$\theta_i^* = \max(\theta_{i1}, \theta_{i2}) = \overline{\theta}_{i.} + |\Delta\theta_{i.}|,$$

where $\overline{\theta}_{i.} = (\theta_{i1} + \theta_{i2})/2$, and $\Delta\theta_{i.} = (\theta_{i1} - \theta_{i2})/2$. The same formula holds for the sample versions. The sample estimators $\widehat{\overline{\theta}}_{i.}$ and $\widehat{\Delta\theta}_{i.}$ are conditionally independent, therefore

$$E(loss(\widehat{\theta}_i^*|\theta_i)|\theta_i) = E((\widehat{\overline{\theta}}_{i.} - \overline{\theta}_{i.})^2|\theta_i) + E((|\widehat{\Delta}\theta_{i.}| - |\Delta\theta_{i.}|)^2|\theta_i)$$

The same relationship holds for bias. The parameter $\overline{\theta}_{i.}$ is a linear function of the original parameter, and its estimator is unbiased. Little is lost from the standpoint of analysis, therefore, by ignoring the first term; i.e, we redefine $\theta_i^* = |\Delta\theta_{i.}|$ and likewise $\widehat{\theta}_i^* = |\widehat{\Delta}\theta_{i.}|$. The quantity $\widehat{\Delta}\theta_{i.}$ has conditional distribution $N(\Delta\theta_{i.}, \sigma_y^2/(2n))$, and marginal distribution $N(\Delta\alpha, \sigma_y^2/(2n) + \sigma_\theta^2/2)$. We use these characterizations to derive the next formulas.

Let us first consider the estimation problem:

$$E(bias(\widehat{\theta}_i^*|\theta_i)|\theta_i) = 2|\Delta\theta_{i.}|[\upsilon_y\phi(1/\upsilon_y) + (\Phi[1/\upsilon_y] - 1)] = O(\upsilon_y) \quad \text{as} \quad \upsilon_y \to \infty$$

$$E(loss(\widehat{\theta}_i^*|\theta_i)|\theta_i) = 2(\Delta\theta_{i.})^2[\upsilon_y^2 + 2(1 - (\phi(1/\upsilon_y) + \Phi(1/\upsilon_y)))] = O(\upsilon_y^2) \quad \text{as} \quad \upsilon_y \to \infty$$

where expectation is with respect to the distribution of data, $\upsilon_y = \frac{\sigma_y/\sqrt{2n}}{|\Delta\theta_{i.}|}$; $\phi(.)$ is the normal distribution's density; and $\Phi(.)$ is the cumulative density. Clearly, $\max_j \widehat{\theta}_{ij} \xrightarrow{p} \widehat{\theta}_{ij_{i^*}}$ as $\upsilon_y \to 0$, where υ_y controls the degree of overlap (specifically $1 - \Phi(1/\upsilon_y)$) between the two actions. Its relationship to bias, therefore, must be positive. The preceding convergence relation implies a convergence in variance when υ_y is decreased, equivalently, when $\Delta\theta_{i.}$ is increased. Conversely, a decrease in $\Delta\theta_{i.}$ is accompanied by a decrease in variance. The effect of expected bias and variance combined, therefore, is an increase in expected loss as a function of υ_y. Figure 2 illustrates the above relationships, for two values of $\Delta\theta_{i.}$, by varying σ_y (this is equivalent to varying υ). Also included is the Cramer-Rao bound, adjusted for the bias of the simple estimate.

[Figure 2]

The next steps are average over strata and over the distribution of the parameters. However, the following simplification (skipping the first step) $E(E(loss(\omega|\theta)|\theta)) = E(E(loss(\omega_i|\theta_i)|\theta_i))$, is a consequence of the assumption that the θ_i's are iid.

The marginal bias is $E(E(bias(\widehat{\theta}_i^*|\theta_i)|\theta_i)) = E(\widehat{\theta}_i^*) - E(\theta_i^*)$. Each term on the right hand side is the expectation of the maximum of a set of normal variables, an operation we have already encountered in evaluating the conditional bias. We can therefore obtain a closed form expression:

$$E(E(bias(\widehat{\theta}_i^*|\theta_i)|\theta_i)) = 2|\Delta\alpha|[(\upsilon_{y,\theta}\phi(1/\upsilon_{y,\theta})+\Phi(1/\upsilon_{y,\theta}))-(\upsilon_\theta\phi(1/\upsilon_\theta)+\Phi(1/\upsilon_\theta))] = O(1/\upsilon_\theta) \quad \text{as} \quad \upsilon_\theta \to \infty$$

where $\upsilon_\theta = \frac{\sigma_\theta/\sqrt{2}}{|\Delta\alpha|}$ and $\upsilon_{y,\theta} = \sqrt{\frac{\sigma_y^2/2n}{|\Delta\alpha|^2} + \upsilon_\theta^2}$.

Figure 3 illustrates the relations for bias and loss (the square root of mse) for two fixed values of $\alpha_2 - \alpha_1$ and varying σ_θ (this is equivalent, for each $\alpha_2 - \alpha_1$ value, to varying υ_θ). In the right panel, bias increases before it decreases. As υ_θ is increased, the chances of overlap diminish, so that both $E(\widehat{\theta}_i^*)$ and $E(\theta_i^*)$, which are both expectations of maxima, decrease, which determines the concave shape of bias.

Evaluating the loss (square root of the mse), requires the evaluation of the expectation of $\Phi(1/\upsilon_y)$, which we carried out numerically. The loss increases with υ_θ, with an upper limit that reflects the variance of the estimator (in agreement with our analysis of conditional variance), as the bias, as we have seen, converges to zero.

[Figure 3]

A similar analysis for the classification problem yields:

$$E(error(\widehat{a_i^*}|\theta_i)|\theta_i) = (1 - \Phi(1/\upsilon_y))$$

$$E(loss(\widehat{a_i^*}|\theta_i)|\theta_i) = |\Delta\theta_\cdot|(1 - \Phi(1/\upsilon_y)).$$

Only the first expression above is available in closed form, but the second is easily evaluated numerically. Graphically, these are shown together with the other estimators in the simulation study in Figure 5 and 7.

3.2 Bias correction using the parametric bootstrap

A high variance in the data, for example as a result of small sample sizes, implies, as we have seen, a greater bias. Before considering shrinkage estimators, we briefly develop the bootstrap method, which offers the flexibility to modify any characteristic of an estimator, in this case, bias.

Let F_0 denote the true unknown distribution, and F_1 the empirical distribution from a random sample, and F_2 the empirical distribution from a sample drawn with replacement from F_1. Let $g(F_0)$ denote the quantity of interest, and $g(F_1)$ its estimator. A common way to approach our problem is to find t such that: $E(g(F_1) - t(F_0) - g(F_0)|F_0) = 0$, equivalently, $t(F_0) = E(g(F_1) -$

$g(F_0)|F_0) = bias_E(F_0)$, where the E subscript means in expectation. Alternatively we may use the "multiplicative" approach: $E(g(F_1)(1 - t(F_0)) - g(F_0)|F_0) = 0$. As F_0 is unknown, and provided the relationship from F_1 to F_2 captures that of F_0 to F_1, it is natural approximate $t(F_0)$ by $t(F_1) = E(g(F_2) - g(F_1)|F_1) = bias_E(F_1)$. It is possible to iterate the bootstrap principle, with each iteration reducing the order of the error by a factor of at least $n^{-1/2}$ (Hall and Martin, 1998), which has to be traded off with increasing computational costs.

In our case, within a given stratum i, F_0 has a parametric characterization i.e. $F_0 = F_0(\theta_i, \sigma_y)$ and the formula for the expectation of bias, $bias_E(\Delta\theta_{i.}, \upsilon_y) = E(bias(\widehat{\theta}_i^*|\theta_i)|F_0)$ is known from our analysis of simple estimates. The bias corrected estimator, therefore, is $\widehat{\theta}_i^{*\,boot} = \widehat{\theta}_i^* - bias_E(\widehat{\Delta\theta}_{i.}, \widehat{\upsilon}_y)$, with $\widehat{\upsilon}_y = \upsilon_y(\widehat{\Delta\theta}_{i.}, \widehat{\sigma}_y) = \frac{\widehat{\sigma}_y/\sqrt{2n}}{\widehat{\Delta\theta}_{i.}}$. According to the assumed data collection process, $\widehat{\sigma}_y$ pools data from all I strata, and is independent of the $\widehat{\theta}_{ij}$'s, which are estimated in each stratum separately. We should expect, therefore, that σ_y is estimated relatively accurately in comparison to the $\Delta\theta_{i.}$'s.

So far, to obtain a convenient parametric formulation, we have implicitly assumed an arbitrary pairing of observations in each stratum, such that F_1 is the empirical distribution of $\{\{y_{i11}, y_{i21}\}, \ldots, \{y_{i1n}, y_{i2n}\}\}$, when in fact $\widehat{\Delta\theta}_{i.}$ combines data from two independent samples, specifically $F_1' = F_{11} \times F_{12}$, where F_{1j} is the empirical distribution from the sample for action j, $\{y_{ij1}, \ldots, y_{ijn}\}$. Resampling from F_1' rather than F_1 to obtain a bootstrap estimator would not modify the expected bias, but would reduce its variance.

4 Bayesian estimators

As noted in the problem formulation section, the appropriate paradigm for our problem is Bayesian (for a real world application of multilevel decision trees that motivated this work, see Lin et al., 1999). Moreover, our analysis of simple estimates has shown that their properties deteriorate, notably in terms of bias as the standard deviation to mean ratio, υ_y increases. One would expect, therefore, that shrinking an estimator towards its mean as the variance increases, thus reducing the variance of the modified estimator, would be a viable strategy. We evaluate a Bayesian approach under the assumption that the assumed normal model is correct, with the addition of a flat prior density on the hyperparameters $\psi = (\alpha_1, \alpha_2, \sigma_\theta, \sigma_y)$. This is a default choice compared to the estimators defined previously. In a particular application, especially with a small number of strata K, one might prefer a more informative prior distribution (Gelman, 2004).

For this model, computation is straightforward using the Gibbs sampler. We follow standard recommendations (Gelman et al., 2003) by simulating multiple Gibbs sequences with over-dispersed starting points, selecting the number of iterations on the basis of potential scale reduction $\widehat{R}$, and discarding the first half of the simulation. In this case, approximate convergence is reached after 40 iterations.

According to the results of Section 2, the solution to the estimation and classification problems

are

$$\widehat{\theta}_i^* = \widehat{E}(\theta_i^*|y)$$
$$\widehat{j}_i^* = \arg\max_j \widehat{E}(\theta_{ij}|y),$$

where $\widehat{E}(.)$ indicates averaging over Markov chain samples, which we index by $l \in S_{MC} = \{1, ..., L\}$. We obtain the optimal value of the tree by averaging over strata: $\widehat{V}^* = \sum_i \widehat{\theta}_i^*/\mathrm{I}$.

The posterior mean of the payoff for a given pair (i, j), θ_{ij}, is expressed as a weighted average of the prior mean α_a and the simple estimate $\overline{y}_{ij}$ with weights proportional to the precisions $1/\sigma_\theta^2$ and n/σ_y^2. If $\sigma_\theta = 0$, the model pools observations from all strata. Conversely, if $\sigma_\theta = \infty$, the estimator is identical to the simple estimate, and inferences is made in each stratum separately.

For each $l \in S_{MC}$, the distribution $p((\theta_{i1}, \theta_{i2})|\psi^{(l)}, y)$ is bivariate normal distribution, where ψ denotes the hyperparameters. Furthermore, we know, from our analysis of simple estimates, the exact formula for the expectation of the maximum of normal distributions. We can therefore, alternatively, define $\widehat{\theta}_i^{**} = \widehat{E}(E(\theta_i^*|\psi, y))$. This Rao-Blackwellized version is identical in expectation and often has better variance properties (Liu et al., 1994).

5 Empirical studies

In our simulation study, we explore the performance of the Bayes, bootstrap and simple estimators under various values of ψ. We set $I = 20$ and $K = 3$, corresponding to a moderate number of strata and a small amount of data in each stratum. We run 160 simulations for the parameters θ, and conditionally on each of the latter, sample 50 datasets y. The number of simulations is chosen so that the empirical estimates of ψ are approximately within $\pm 5\%$ of their true value 95% of the time. The number of simulations for the dataset is smaller, because the empirical version of $\sigma_y \in \psi$ pools data from all strata. In all, for each ψ, take into account, at the inner level, Gibbs sampling, there are 3 levels of nesting in our simulation study.

The Rao-Blackwell version of the Bayes estimator performed slightly better in terms of loss than the empirical version, so we report only the former.

Results are reported graphically in Figure 4, 5, 6 and 7. Even numbers illustrate the estimation problem, and odd numbers the classification problem. Each figure is split into four graphs. The top and bottom panels of the figures relating to the estimation problem show bias and loss (square root of mse), respectively. For the classification problem, the top and bottom panels show the error and loss respectively, as defined in Section 2. In all four figures, in the left and right panels, we vary σ_θ and σ_y, respectively, with the other hyperparameters held fixed. In the left panel, the true value of the tree increases with σ_θ, while it is constant in the right panel. The first two and the last two figures differ only in their assumed value for $\alpha_2 - \alpha_1$, specifically a low value (0.01) and high value (0.2), which induces a lower and higher value of the tree, respectively.

[Figure 4, 5, 6, and 7]

We should stress that it is marginal goodness of fit measures that we are reporting, that is the averages over the empirical distribution of parameters and data. For our analysis, it useful to rely on

$$E(E((\widehat{g(\theta)} - g(\theta))^2|\theta)) = E(Var(\widehat{g(\theta)}|\theta)) + E(E(bias(\widehat{g(\theta)}|g(\theta))^2|\theta).$$

In the estimation problem we recall that $g(\theta) = (g_1(\theta_1), \ldots, g_I(\theta_I))$, and $g_i(\theta_i)$ represents the max(.) between the two decisions in stratum i. The analysis, when σ_y is varied (right panels) is straightforward: as it is increased, $Var(\widehat{g(\theta)}|\theta)$ increases, both for the simple and bootstrap estimate. The Bayesian estimate, also has a tendency to increase, but to some degree, this is offset by the greater pooling across strata. The term $E(bias(\widehat{g(\theta)}|g(\theta))^2|\theta)$ is increasing in σ_y for all three estimators. As we have seen in Section 3.1, max(.) between two decisions in a given stratum, is estimated with increasing positive bias when within stratum data is used. The pooling effect of the Bayesian estimator, however, yields less extreme estimates for the quantities inside the max(.), thus for the max(.) itself. To appreciate this, we should note that if $g_i(.)$ was linear in θ_i, the expected bias would be zero in the case of the simple estimate, and therefore non-zero for the Bayesian estimator (although the overall performance, measured by mse would still be smaller for the latter, a result which dates back to the study of James-Stein estimators).

The analysis when σ_θ is increased (left panel) is as follows: we have seen, for the simple estimator, that bias converges to zero, and $E(Var(\widehat{g(\theta)}|\theta)) \rightarrow E(Var(\widehat{\theta}|\theta))$. The two effects combined, is reflected in mse increasing with upper limit $E(Var(\widehat{\theta}|\theta))$. As the degree of pooling is reduced, the Bayesian estimator converges to the simple estimate.

Overall, both the bias of the bootstrap and the Bayesian estimates are bounded above by that of the simple estimate. Moreover, the Bayesian estimator performs at least as well as the bootstrap in terms of bias (they are strikingly close relative to the simple estimate), but the former, unlike the second, incurs an increase in variance which leads to an overall deterioration of mse, compared with the the simple estimate.

The classification problem is to find $\arg\max_j \theta_{ij}$, that is, we are interested in $g_i(\theta_i) = \theta_i = (\theta_{i1}, \theta_{i2})$. Using the argument made previously, $g(\theta)$ is now estimated without bias using the simple estimate, and with some bias using the Bayesian estimate, but the marginal mse is greater for the former. By mapping θ onto the decision space, however, the superiority of the Bayesian estimator is not necessarily preserved, as shown in Figure 5, particularly for high values of σ_y (in the right panel), and mid–range values of σ_θ (left panel). In Figure 7, where the value of the tree is higher than in Figure 5, the Bayesian estimator always dominates the simple estimate, by a high margin, in the left panel, and is almost identical in the right panel. Therefore, if we multiply the differential in loss between the simple and the Bayesian estimator by the value of the tree, there is a substantial advantage, without any prior about the value of the tree, to use the Bayesian estimator.

6 Conclusion

Our paper is concerned with bias and variance of estimators of non–linear functions of the parameters, in the context of multilevel models, which arise commonly in medical or social studies, spanning distinct but related trials or various geographical regions, respectively. In the context of linear functions of the parameters, shrinkage trades more bias for less variance. In our context, shrinkage reduces both bias and variance. Hierarchical Bayesian inference can thus be helpful in evaluating trees as well as formulating the decision problem itself.

References

Bielza, C., Mueller, P. and Rios Insua, D. (1999). Decision analysis by augmented probability simulation. *Management Science* **45**, 995-1007.

Clemen, R. T. (1996). *Making Hard Decisions*, second edition. Belmont, Calif.: Duxbury Press.

De Groot, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

Gelman, A., Carlin, J. B., Stern H. S., and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd edition. London: CRC Press.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, to appear.

Hall, P., and Martin, A. (1988). On bootstrap resampling and iteration, *Biometrika* **75**, 661-671.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. (1999). Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation (with discussion). *Statistical Science* **14**, 305–337.

Liu, J. L., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes, *Biometrika* **8**1, 27-40

Luce, R. D., and Raiffa, H. (1957). *Games and Decisions.* New York: Wiley.

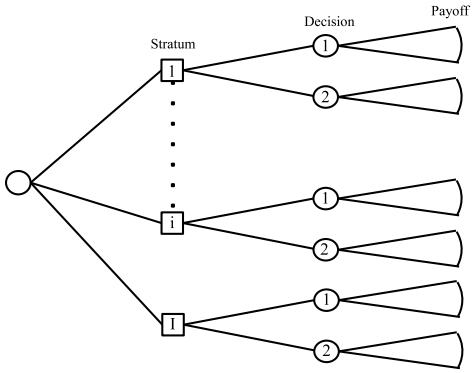

Figure 1: Illustration of a decision tree. Circles are uncertainty nodes and squares are decision nodes. Decisions are made in each district, between teaching method 1 or 2, each of which has a random payoff. To obtain the value of the tree, one has to average over the strata, which are weighted, for example, by their population size. This step is symbolized by a random root node.

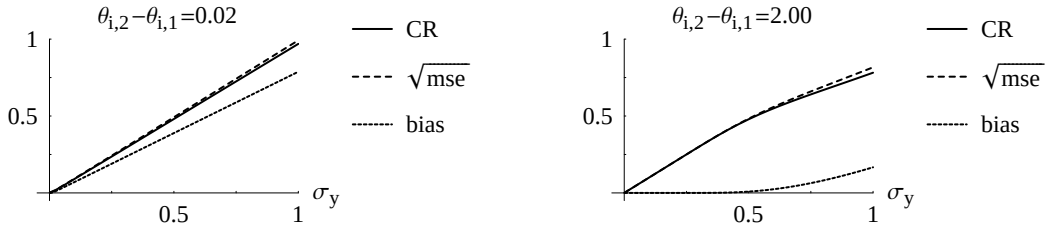

Figure 2: Conditioning on population parameters within stratum i, $(\theta_{i,1}, \theta_{i,2})$, Cramer-Rao bound, root mean square error and bias as a function of σ_y. In the left panel, the difference in expected payoffs between the two actions is small, and conversely in right panel. The Cramer-Rao bound is adjusted for bias, i.e. it is the best achievable mse given the bias of the simple estimate.

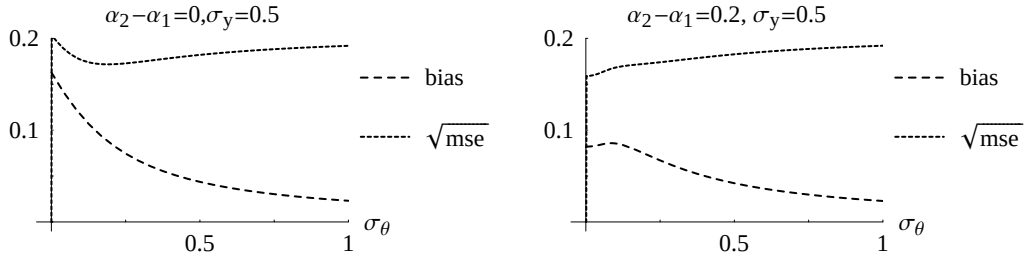

Figure 3: Bias and square root of mse, averaged over the parameter–data distribution. In both panels, we vary σ_θ, with σ_y held fixed. In the left panel, the marginal difference in payoff $\alpha_2 - \alpha_1$ is small, respectively large in the right panel.

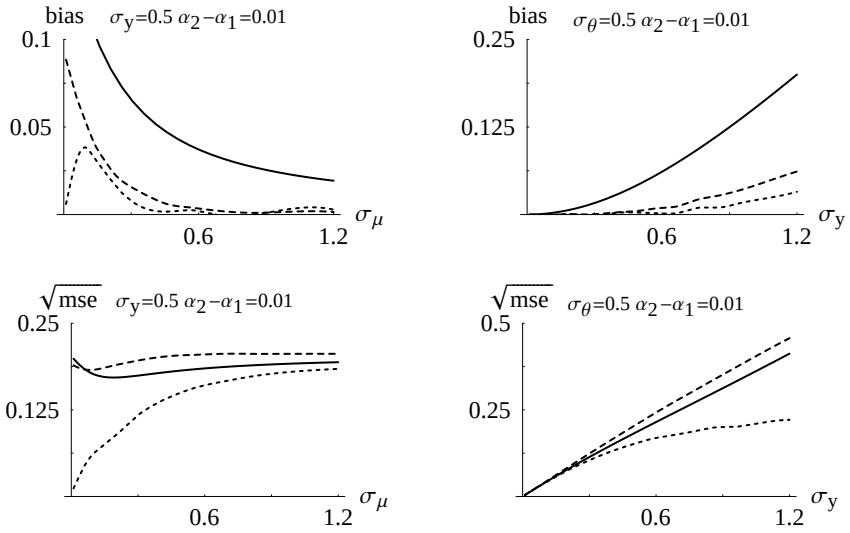

Figure 4: Bias and root mean square error, averaged over the distribution of parameters and data, for the simple, bootstrap and Bayes estimators, in a hierarchical decision structure. In the left panel we vary σ_y while σ_θ is held fixed. In the right panel, the varying and fixed hyperparameters are interchanged. The results for the first estimator are based on analytic or numerical evaluations. Results for the last two estimators are based on a simulation.

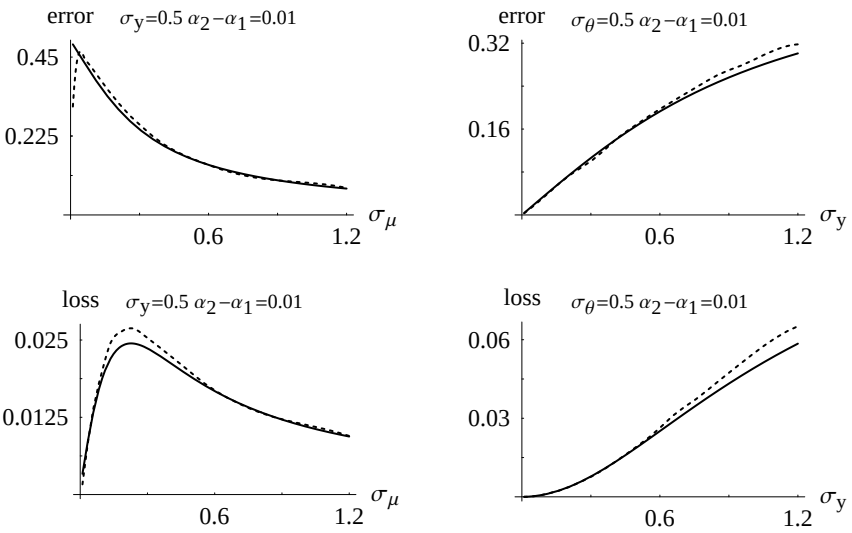

Figure 5: Error and loss, averaged over parameters and data, for the simple, bootstrap and Bayes estimators, in a hierarchical decision structure. In the left panel we vary σ_y while σ_θ is held fixed, and conversely in the right panel. The results for the first estimator are based on analytic or numerical evaluations. Those of the Bayes estimator are based on a simulation.

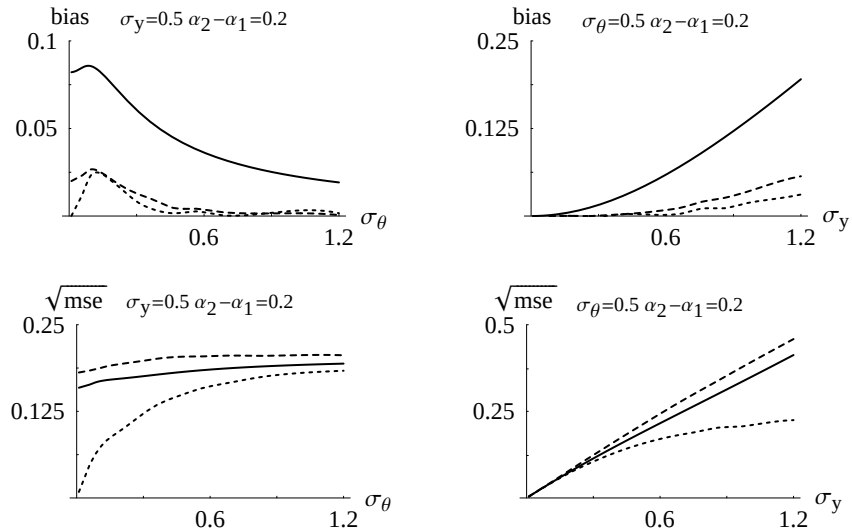

Figure 6: Same as Figure 4 except $\alpha_2 - \alpha_1$ has been increased from 0.01 to 0.2.

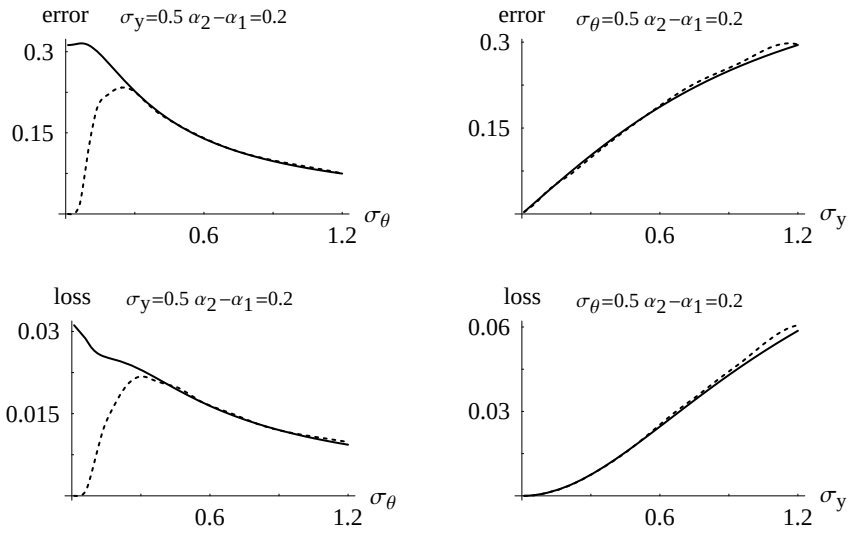

Figure 7: Same as Figure 5 except $\alpha_2 - \alpha_1$ has been increased from 0.01 to 0.2.