Submitted to the Journal of the ACM, August 2005.

Model Selection and Model Complexity: Identifying Truth Within A Space Saturated with Random Models

Paul Helman[1]

Abstract

A framework for the analysis of model selection issues is presented. The framework separates model selection into two dimensions: the model-complexity dimension and the model-space dimension. The model-complexity dimension pertains to how the complexity of a *single* model interacts with its scoring by standard evaluation measures. The model-space dimension pertains to the interpretation of the totality of evaluation scores obtained. Central to the analysis is the concept of evaluation *coherence*, a property which requires that a measure not produce misleading model evaluations. Of particular interest is whether model evaluation measures are misled by model complexity. Several common evaluation measures — apparent error rate, the BD metric, and MDL scoring — are analyzed, and each is found to lack complexity coherence. These results are used to consider arguments for and against the Occam razor paradigm as it pertains to overfit avoidance in model selection, and also to provide an abstract analysis of what the literature refers to as oversearch.

1. Introduction

The machine learning and statistics literature contains much analysis of how such factors as the complexity of models, the number of models evaluated, and the distributions of true models and relevant features affect model selection and error bound estimation. In this article, we propose that the questions are clarified when one makes explicit a separation of model selection factors into two dimensions: the model-complexity dimension and the model-space dimension. Intuitively, the model-complexity dimension pertains to how the complexity of a *single* model affects the distribution of its evaluation scores, while the model-space dimension pertains to how the characteristics of model space affect the interpretation of the totality of evaluation scores.

We postulate a pristine, limiting case set of assumptions which reflects an idealization of many high-dimensional applications (e.g., microarray, proteomic, and many other biomedically-inspired analyses) currently the subject of intense investigations. In such an environment, the number of features is virtually limitless, and most have no correlation with the class to be predicted. Our idealization facilitates the study of central issues, and the results are argued to provide insight into more realistic settings in which the assumptions are relaxed.

We develop a notion of measure coherence. Coherence means, roughly, that the model evaluation measure behaves in a rational way when used to compare models. Of particular interest is the question of whether measures exhibit an *a priori* bias for or against models of high complexity as compared to simple models. We study the question in the abstract, as well as by applying the analysis to standard data likelihood, to the apparent error rate evaluation measure (both with and without cross validation), to the Bayesian-Dirichlet (BD) metric, and to the minimum description length (MDL) scoring function.

We present both analytical and numerical results demonstrating lack of coherence for the error rate measure (with a bias toward more complex models), and for MDL and the BD metric (with a bias toward less complex models). We interpret these results in the context of such previous research as that presented in [1, 2, 13, 16, 17, 20, 23, 26, 30]. Our analysis is enabled by the separation of the model-complexity dimension from the model-space dimension: issues that often have been attributed to model space or to model search are now seen to be directly rooted in the non-coherence of the measure.

[1]Computer Science Department, University of New Mexico, Albuquerque, NM 87131

This work was supported in part by DARPA Contract N00014-03-1-0900 and by grants from the D.H.H.S. National Institutes of Health/National Cancer Institute (CA88361), the W.M. Keck Foundation, and National Tobacco Settlement funds to the State of New Mexico provided to UNM for Genomics and Bioinformatics.

In our final section, we briefly visit the model-space dimension. Here, we assume a coherent evaluation is used, and hence any issues that might arise are attributable solely to model space and search characteristics. Of primary interest here is the calculation of the *a priori* probability of selecting the true model M^* from amongst a large collection of random models. We calculate the *a priori* probability that M^* is selected when correct posterior evaluation is employed, and consider also the effect on the probability of selecting M^* of the number of models evaluated. This latter results is in contrast to oversearch results such as those of Quinlan and Cameron-Jones [20], demonstrating that when coherent evaluations are employed the oversearch phenomenon does not occur.

Critically, most of the conclusions reached are independent of distributional details of the actual true model M^*, and of how the true model is distributed in model space. In particular, we identify model selection biases that are not dependent on a predisposition for truth taking one form (e.g., simple) over another (e.g., complex).

The remainder of this article is organized as follows. Section 2 presents a brief review of some of the related literature on model complexity and model selection, and introduces our intuitive arguments for separating model selection issues along the two dimensions of model complexity and the model space. Beginning in Section 3, we explore by formal means the two dimensions. Section 3 defines measure coherence and views standard data likelihood in these terms. Section 4 demonstrates the non-coherence of apparent error rate, BD, and MDL. Section 5 considers model space issues, including an analysis of the probability of selecting the true model as a function of the number of models evaluated and the number of models in the space. Section 6 summarizes our conclusions and indicates several avenues for future work.

2. Model Complexity and the Model Space

2.1 Related Work

Schaffer [23] highlights that any selection bias toward simple models — such as overfit avoidance — is justified only by a prior judgement that simple is more likely. Much of what is presented here is in agreement with such results. However, the results presented here open another issue. While Schaffer's analysis is indisputable when a coherent evaluation is applied (such as model posterior or likelihood), we demonstrate here that as a result of properties inherent in many evaluation measures employed in practice, a bias may in fact exist for certain model complexities over others, independent of prior distributional assumptions on model complexity. That is, for some common model evaluation measures, complexity bias is present, independent of what the actual true model M^* can be, and with what probability.

The seminal work of Blumer, *et. al.* [2] relates generalization error bounds to model complexity. The more complex is a hypothesis that is used to encode the training labels, the weaker is the generalization error bound that is allocated to the hypothesis. There is a certain similarity between this result and the non-coherence of apparent error rate which we exhibit in Sections 4.1 and 4.2, and of the DL_{data} term of MDL scoring considered in Section 4.4, insomuch as we demonstrate that complex models that fit the training data well (and thus encode the training labels) are more likely than simple models that fit the training data equally well to be "false" models, and hence to not generalize to out-of-sample data. Our result, however, derives from the fact that many notions of "fit to the data" lead to non-coherent measures which exhibit an evaluation bias on *individual* models that is based on a model's complexity.

PAC learning [28] with its many contributors and extensions (for example, [9] and [15] discuss extensions related to several issues examined in the current work) is concerned primarily with bounding generalization error achievable by polynomial time learning algorithms, and relate the number of training instances that must be considered to the cardinality or the VC dimension [29] of the model space, or to the number of models evaluated. Such issues fit more closely into our model-space, rather than model-complexity, dimension, since our model-complexity dimension is concerned not with characteristics of the space, but rather with the interaction between the complexity of the individual models and the specifics of an evaluation measure. Even so, our model space focus differs from that of these other works in that we are concerned with the probability of selecting the true model M^* based on a coherent evaluation and number of models present, rather than on bounding the expected

error of the model selected by a polynomial time learning procedure. Indeed, our results speak to applications in which the number of training instances is far too small for error bounds to be meaningful, but where we still wish to know what search and evaluation procedure is best, and also to know when we are in a situation where the identification by any means of the true model is hopelessly improbable.

We note that Blum and Langford [1] recently presented an elegant framework for unifying PAC bounds with MDL model complexity.

Kearns *et. al.* [13] studies complexity classes of boolean "step" functions and experiments with different procedures for selecting the appropriate hypothesis complexity to match a target step function. The measure-of-fit criterion used to evaluate hypothesis functions against sample points is akin to an apparent error rate evaluation, and the need for such complexity adjustment is attributable to properties of the evaluation measure. Again, if true posterior (whose computation includes a prior distribution over the complexity of the target function to be selected) could be computed, there would be no grounds for an adjustment procedure of any sort. Since, however, true posterior is rarely computable, there is ample practical motivation for the development of such procedures. Empirical studies of specific procedures for pruning back decision trees motivated by overfit avoidance, similar to what is performed in CART [3], include Murphy and Pazzani [17] and Webb [30], who reach rather different conclusions. Many such studies employ evaluation criteria closely related to apparent error rate, reflecting criteria often used in practice. One ultimate goal of the current work is to utilize a formulation such as the evaluation ratio introduced in Section 3.1 toward developing quantitative means of trading model complexity against evaluation quality in the context of specific, non-coherent evaluation measures.

Quinlan and Cameron-Jones' oversearch analysis [20] claims that there is a point beyond which searching the space of classification trees appears to degrade generalization performance. Others [25] have pointed out that this result could be due to the choice of evaluation function rather than being rooted in oversearch. Indeed, our model space analysis, performed in its simple setting, proves that the more models evaluated by a coherent measure the better the chance of selecting the true model. That is, oversearching is not a phenomenon that occurs in this setting, and settings which extend it.

2.2 Two Dimensions of Analysis

Many researchers consider model selection while employing measures that are *a priori* biased in favor of models of one complexity over those of another complexity in a sense that is formalized in Section 3. Intuitively, the bias is that identical scores imply different actual posteriors for models differing only in their complexities. Critically, this bias does not depend on a prior bias for a model of one complexity being truth more often than a model of another complexity, nor on how many models of each complexity exist or are evaluated. Further, the bias depends only minimally on specifics of the true model itself. Hence, while such work as Schaffer's [23] and Webb's [30] correctly argue against the universality of the Occam razor paradigm for model selection by pointing out that overfit avoidance is a distribution-on-truth bias that is only sometimes appropriate, we demonstrate here that this conclusion hinges on the use of a coherent evaluation function. Apparent training-set error rate — cross validated or not — with its bias for complex models, and MDL and, in some contexts, the BD metric, with their bias for simple models, are classic examples of measures subject to an inherent complexity bias. We refer to this issue of complexity biased evaluations as the model-complexity dimension of model selection, and the biased evaluation measures themselves are said to lack coherence.

Orthogonal to the model-complexity dimension is the model-space dimension, which captures issues arising from the fact that model space generally contains an enormous number of false (e.g., random) models, and generally contains more complex models than simple models. While the two complexity dimensions are orthogonal, they often are blurred in the analysis, and the problems of a complexity biased evaluation are compounded by a space in which a disproportionate number of the models are complex. Issues arising from directed search — such as dependence between the models evaluated, and the fact that models found late in the search may be more suspect than models found early — are additional complicating factors.

An exact mathematical analysis, while quite intractable in many practically important settings, applied with

respect to a pristine (but, necessarily simplistic) set of model assumptions, serves to illuminate the issues. Model posterior is the definitive evaluation of a model, and, by definition, cannot exhibit bias of any type, including bias for or against models of certain complexities, unless such a bias is explicitly encoded in the priors. While this observation follows trivially from basic probability theory, practical application requires that we learn how to transfer its consequences to situations in which true posterior cannot feasibly be computed. Even so, one immediate implication of our analysis is that the afore mentioned issues such as model complexity biases often observed in the literature cannot arise when proper evaluations are used, yet many published analyses place fault inherently on search schemes or on the nature of learning itself, rather than identifying that these problems stem uniquely from imperfect model evaluation criteria. Once we put aside the distractions of biased evaluation or search, we can identify the fundamental limitations of model selection: when there are too many false models relative to the amount of data available, the probability that the true model M^* has the highest posterior, conditioned on that data, goes to zero. Unfortunately, this is a predicament that no amount of ingenuity in the design of evaluation procedures or search algorithms can rectify. The best we can do is quantify the uncertainty.

3. Model Complexity and Coherence

3.1 Coherence Properties of Measures

Model posterior $Pr\{M\ truth \mid d\}$ is the quintessential evaluation criterion and, when it can be computed and applied to model selection, many of the issues examined in this work are, by definition, accounted for. But, typically, model posterior cannot be computed exactly, and here we examine the consequences of other model evaluation measures, specifically highlighting the fact that many common measures are inappropriately affected by model complexity. We illustrate by means of a probability model of a general and intuitive form, and emphasize at the outset that the conclusions drawn in no way are peculiar to any particular realization of this probability model, such as the Bayesian networks and classification trees considered throughout this article.

The probability model postulates that there is a universe U of features, and some subset $PS \subseteq U$ of these features is statistically correlated with the class variable, whose value is to be predicted. Such a probability model can be viewed from the perspective of either a Bayesian network [6, 10, 18, 19] or of a classification tree [3, 4].

When viewed as a Bayesian network, the probability model takes the form of the parent set classification network developed in [11]. Given the values of the parents of the class label, which has no children in such networks, the class label is rendered statistically independent of the remaining features. Assuming that the features and class label are binary, there are 2^k states ps_i of the parent features (combinations of their binary values), and with each ps_i there is a conditional probability distribution for the class label, specifying $Pr\{C = 1 \mid ps_i\}$ and $Pr\{C = 2 \mid ps_i\}$ summing to 1.0. We allow for each ps_i, $Pr\{C = 1 \mid ps_i\}$ to be any value in $[0,1]$. When this conditional probability is 0.0 or 1.0, the class label is functionally determined by the parent state; otherwise, the relationship is probabilistic. One can also think of a probabilistic relationship as a functional one in which the value of the class label with some probability is altered by noise. For example, to capture a situation in which the class label is functionally determined to be 1 when the parent state is ps, but noise flips the label to 2 with probability 0.1, one would specify $Pr\{C = 1 \mid ps\} = 0.9$. Of course, probabilistic relationship with other semantics can be so modelled as well.

In addition to the Bayesian network realization, one can think of this probability model as a homogeneous classification tree, in which only a subset of the universe of features affects the classification. The tree is homogeneous in the sense that each root-to-leaf path contains the same sequence of features, and thus each combination of parent feature values in the Bayesian network parent set model corresponds to a path through the classification tree. The height of the tree is equal to the number k of parents in the Bayesian network model, and the number 2^k of leaves is equal to the number of states of the k parents. We will alternately view a model as a classification tree or a Bayesian network, depending on the analysis we wish to perform. Viewing models as classification trees facilitates comparisons with research (for example, [17, 20, 23, 30]) where apparent error rate and related measures on tree models are studied, while viewing models as Bayesian networks allows the natural inclusion in

the model of prior distributions over the distribution parameter θ and thus analysis of the *BD* metric in the terms of [10, 26], and also of MDL-based measures as considered in [6, 7, 14]. As [4] demonstrates, a classification tree in fact can be treated directly within the Bayesian framework as well.

In general, formal definitions will be presented in the Bayesian network terminology, with translations made to classification tree terminology when appropriate. As such, a model M has two components: $M =< G, \theta >$, where G is a network structure (that is, a directed acyclic graph, or DAG) and θ is a distribution consistent with the conditional independence assertions of G. In the context of this article, G is a parent set model, that is, each feature in some subset PS of the universe of features (the parents) is the tail of a directed edge into the class label, and G contains no other edges. Also in the context of this work, we take the consistency of θ with G to require that the parent set in G be a minimal set of features which renders the class label statistically independent in the distribution θ of the remaining features.

When we say a model $M^* =< G^*, \theta^* >$ is the true model, this means our observed data set d is generated in accordance with distribution θ^*.

3.1.1 Distribution Assumptions and the Model Selection Problem

All features and the class label are binary, assuming values in $\{1,2\}$. We assume that the features (which do not include the class label, which will not be referred to as a feature) are statistically independent and identically distributed (iid), each taking a value in $\{1,2\}$ with equal probability. Hence, all combinations of feature values are equally likely within a case $x_i \in d$.

We equate model complexity with the number of features on which the class label depends, i.e., the number of parents in the Bayesian network. $card_k$ denotes the set of all models with k parents. Members of $card_k$ thus have 2^k parent states, or *cells*, on which the class label is conditioned in the network, and this also is the number of parameters needed to describe the model. Thus, our measure of model complexity tracks with most model complexity conventions (for example, MDL measures, as applied to Bayesian networks [6, 7, 14]) typically considered. Note that the minimality assumption stated above implies that that sets $card_k$ of models are not nested, but, rather, are disjoint.

We postulate a model interaction that facilitates analysis and is the limiting case of many important applications having high dimensionality. We assume that model space is such that the parent sets of all models having nonzero prior probability are disjoint. That is, no pair of models that can be truth with nonzero prior probability share any features — the features of all models M other than M^* are uncorrelated with each other and with the class label. Such a model is said to be random. By the no correlation assumption, the event "M *not truth*" and "M *random*" are equivalent, and $(1 - Pr\{M\ truth\}) = Pr\{M\ not\ truth\} = Pr\{M\ random\}$. Also $Pr\{data\ d\ such\ that\ Pred(d) \mid M\ random\}$ and $Pr\{data\ d\ such\ that\ Pred(d) \mid M\ not\ truth\}$ are therefore logically equivalent for any predicate *Pred* and will be used interchangeably. We will also abbreviate the event "M *not truth*" to "*NOT* M".

The disjointness of features implies that, for each $card_k$, there is some (possibly enormous) number of disjoint subsets of k features such that only the models over each of these subsets have nonzero prior probability. Further, for distinct cardinalities k and k', the nonzero prior probability models in $card_k$ and $card_{k'}$ share no features. These assumptions approximate the limiting case of high dimensional model spaces in which almost all features are irrelevant and, importantly, any evaluation anomalies present under these assumptions will necessarily be present in a more intricate space that contains any subspace having the properties that are postulated here.

The disjointness assumption further implies that only one θ, which we often denote by $G(\theta)$, is associated with any network G, that is, for any graph G, at most one $M =< G, \theta >$ has nonzero prior probability. We will also write $M(\theta)$ depending on context. The associated θ in general is not known to the model evaluators, and how the model evaluators depend on $G(\theta)$, or on the distributions (e.g. Dirichlet priors) over the space Θ of possible θ which they assume, will be a focus of the analyses in the sections which follow.

When we say that the prior distribution of models is uniform, we mean that any nonzero-prior network G within this disjoint-feature subspace has equal probability of being selected as the true model. One can imagine

a process in which a network G^* is chosen uniformly at random from this subspace, and data d is generated in accordance with $G^*(\theta)$. The *model selection problem* is to compute $EV(M_i, d)$ for some collection of nonzero-prior models M_i and evaluation function EV, and decide how relatively likely is it that each M_i so evaluated is the chosen model $M^* =< G^*, G^*(\theta) >$ that generated the observations in d. Our interest in this article is in evaluation characteristics rather than search and, therefore, we abstract away the details of specific search algorithms by assuming that a model of a specified complexity is chosen for evaluation by an oracle with equal probability from among models with nonzero prior probability.

3.1.2 Measure Coherence

For a given scoring function $EV(M,d)$, we consider for a pair M, M' of models, the model posteriors conditioned on the models achieving a particular pair of scores on the observed data d, i.e.,

$$Pr\{M\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\ \}\ \text{and}$$
$$Pr\{M'\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\ \}$$

The conditioning events above each is the aggregation of data for which the given evaluations are obtained, rather than the data itself. By evaluating with scoring function EV, we replace knowledge of d with knowledge of the scores, i..e., we replace the data-specific posterior $Pr\{M\ truth \mid d\}$ with the above aggregation of data d_i based on a common EV score. The issue we wish to consider is, how well-behaved is the scoring function EV? For example, assuming that (nonzero) model priors $P(M)$ are all equal, if two models of differing complexities score the same, are their posteriors the same? That is, is it the case that

$$\begin{aligned} & Pr\{M\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v\ \} \\ = \quad & Pr\{M'\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v\ \} \end{aligned}$$

More generally, are models correctly ordered by the scores, or are scores inappropriately influenced, for example, by model complexity?

We can study the above posteriors by studying the *evaluation ratio*

$$\frac{Pr\{d\ such\ that\ EV(M,d) = v \mid M\ truth\}}{Pr\{d\ such\ that\ EV(M,d) = v \mid M'\ truth\ and\ EV(M',d) = v'\}} \tag{1}$$

When model priors $P(M_i)$ are equal, it follows from Bayes Theorem (e.g., see Theorem 3.1 below), that the posteriors are ordered as the ratios (1). That is, for every model pair M and M' and every pair of simultaneously achievable scores v and v',

$$\begin{aligned} & Pr\{M\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\} \\ < \quad & Pr\{M'\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\} \end{aligned}$$

if and only if

$$\frac{Pr\{d\ such\ that\ EV(M,d) = v \mid M\ truth\}}{Pr\{d\ such\ that\ EV(M,d) = v \mid M'\ truth\ and\ EV(M',d) = v'\}}$$
$$< \quad \frac{Pr\{d\ such\ that\ EV(M',d) = v' \mid M'\ truth\}}{Pr\{d\ such\ that\ EV(M',d) = v' \mid M\ truth\ and\ EV(M,d) = v\}}$$

Note that the strict inequality implies *[equality of the evaluation ratios]* iff *[equality of the posteriors]*. When model priors are not necessarily equal, we can generalize our results in terms of movement from the model priors, i.e., Bayes factor. We note further that the evaluation ratios (1), and hence model posteriors when model priors are equal, are ordered as above if and only if the ratio

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M' \text{ truth}\}}$$

is less than 1.0 (see Corollary 3.1). We choose to define coherence in terms of the evaluation ratio (1), rather than directly in terms such a likelihood ratio, because the evaluation ratio better reveals an interdependence between model scores that is central to much of the analysis which follows.

The following definition of coherence specifies that an evaluation measure EV is coherent if and only if EV orders models consistently with their evaluation ratios. EV maps M and d to the reals, and for some EV a small score is good while for others a large score is good. We use $v \prec v'$ to denote the total order on $\mathcal{R}$ in which v' is a better score than v. (When and only $v = v$, the pair is not ordered by $\prec$.)

Definition 3.1: EV is *coherent* if, for every pair of models M and M', and every pair of simultaneously achievable scores v and v', we have $v \prec v'$ if and only if

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v'\}}$$

$$< \quad \frac{Pr\{d \text{ such that } EV(M',d) = v' \mid M' \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}}$$

Since the scores are simultaneously achievable, the denominators are non-zero. Notice that, taking $v = v'$, this definition requires that

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v\}}$$

$$= \quad \frac{Pr\{d \text{ such that } EV(M',d) = v \mid M' \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v \mid M \text{ truth and } EV(M,d) = v\}}$$

for all model pairs M and M' and simultaneously achievable score v.

As noted above, when model priors are uniform, the ordering of the ratios determines the ordering of the posteriors.

Theorem 3.1: Assuming equal model priors, for any pair of models M and M' and any simultaneously achievable scores v and v'

$$Pr\{M \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}$$
$$< \quad Pr\{M' \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}$$

if and only if

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v'\}}$$

$$< \quad \frac{Pr\{d \text{ such that } EV(M',d) = v' \mid M' \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}}$$

Proof: By Bayes Theorem

$$Pr\{M \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}$$
$$< \quad Pr\{M' \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}$$

if and only if

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\} * P(M)}{Pr\{d \text{ such that } EV(M,d) = v \text{ and } E(M',d) = v'\}}$$
$$< \quad \frac{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M' \text{ truth}\} * P(M')}{Pr\{d \text{ such that } EV(M,d) = v \text{ and } E(M',d) = v'\}}$$

if and only if

$$Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\}$$
$$< \quad Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M' \text{ truth}\}$$

if and only if

$$Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\} \ * \ Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}$$
$$< \quad Pr\{d \text{ such that } EV(M',d) = v' \mid M' \text{ truth}\} \ * \ Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v'\}$$

if and only if

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v'\}}$$
$$< \quad \frac{Pr\{d \text{ such that } EV(M',d) = v' \mid M' \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}}$$

□

Corollary 3.1: Let

$$R = \frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid M' \text{ truth and } EV(M',d) = v'\}}$$

$$R' = \frac{Pr\{d \text{ such that } EV(M',d) = v' \mid M' \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}}$$

be a pair of evaluation ratios. Then

$$\frac{R}{R'} = \frac{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M' \text{ truth}\}}$$

and, if model priors $P(M) = P(M')$,

$$\frac{R}{R'} = \frac{Pr\{M \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}}{Pr\{M' \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}}$$

Proof: The results follow from application of the same standard identities used in the proof of Theorem 3.1. For example, write the numerator $Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}$ of R as

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M',d) = v' \mid M \text{ truth and } EV(M,d) = v\}}$$

□

The presence in the denominator of the evaluation ratio (1) of the joint conditioning event $[M' \text{ truth and } EV(M',d) = v']$ suggests a lack of independence of the scores. Indeed, knowledge of the true model's identity and of its evaluation score may influence the distribution of scores of a random model, even given our disjointness of model assumptions. This lack of independence is what necessitates consideration of the "joint" evaluation ratio (1) rather than simpler formulations, such as a "marginal" evaluation ratio

$$\frac{Pr\{d \text{ such that } EV(M,d) = v \mid M \text{ truth}\}}{Pr\{d \text{ such that } EV(M,d) = v \mid NOT\ M\}} \tag{2}$$

In general, marginal evaluations ratios tell us little about a measure. We can have non-coherence with respect to marginal ratios and be coherent; we can have coherence with respect to marginal ratios and still have the inconsistencies we are trying to avoid. We will see specific examples of such phenomena in later sections.

It is important to note that failure of coherence (in the $<$ direction) anytime $v = v'$ implies that there exist d such that models M and M' score the same on d, but M' has a higher posterior conditioned on full knowledge of d. This follows because if the posteriors of M and M' were the same for each d on which they score the same v, then the posterior conditioned on the set of d's in this intersection would necessarily be the same. Therefore, failure of coherence with respect to a pair of model scores is sufficient to imply a failure of coherence regardless of how many model scores are conditioned on, since there is at least one d on which M and M' achieve the same score yet M' has a higher posterior conditioned on full knowledge of this d. While the converse does not necessarily hold (the definition of coherence can be satisfied yet a non-coherence exists when additional model scores are conditioned on) our goal here generally is to exhibit the non-coherence of measures, requiring only the pairwise violation. Further, the pairwise definition is practically appropriate since it reflects how model selection procedures using an EV typically operate; rather than using full knowledge of d (e.g., synthesizing a collection of many scores), they perform pairwise comparisons of model scores. What's more, we will note in Section 3.2 that likelihood — the one measure we identify as being coherent — is in fact coherent under even full knowledge of d.

Coherence can fail due to any number of evaluation biases. Our primary interest here is the failure of coherence due to complexity biases, for example, a bias in which a single complex model scoring well has more of a chance

of being random (i.e., not truth) than does a single simple model scoring equally well. To study such issues, we must formalize a sense in which models can be said to differ only in their complexity.

Definition 3.2: Distributions θ and θ' associated with models M and M' are *homomorphic* if they assign the same set of conditional probabilities

$\{pr_{i1} \mid ps_i$ *is a parent cell and conditional class probability* $Pr\{C=1|ps_i\}$ *is assigned* $pr_{i1}\}$

and in the same proportions across the parent cells of the two models. (That is, the value pr_1 is assigned to $Pr\{C=1|ps_i\}$ for the same proportion of parent cells ps_i of M and M' by the corresponding θ and θ'.) Models $M =< G, \theta >$ and $M' =< G', \theta' >$ are *homomorphic* if θ and θ' are homomorphic. We say that homomorphic models differ only in their complexity.

We shall consider the following issue. Suppose models $M_k \in card_k$ and $M_{k'} \in card_{k'}$ $(k < k')$ differ only in their complexity. The question of complexity coherence is, for what evaluation measures are the evaluation ratios ill-behaved on such pairs M_k and $M_{k'}$ of models? Thus, while coherence requires consistency of ratios for all M and M' model pairs, *complexity coherence* requires consistency only in such cases where model complexity is the sole difference between models. In this sense, complexity coherence is a very minimal and reasonable requirement for a measure to obey.

A violation of complexity coherence results in

$$\frac{Pr\{d \text{ such that } EV(M_k,d)=v \mid M_k \text{ truth}\}}{Pr\{d \text{ such that } EV(M_k,d)=v \mid M_{k'} \text{ truth and } EV(M_{k'},d)=v'\}}$$
$$> \quad \frac{Pr\{d \text{ such that } EV(M_{k'},d)=v' \mid M_{k'} \text{ truth}\}}{Pr\{d \text{ such that } EV(M_{k'},d)=v' \mid M \text{ truth and } EV(M,d)=v\}}$$

for at least one pair of models M_k and $M_{k'}$ differing only in their complexity, and at least one pair of scores v and v' such that $v \prec v'$ or $v = v'$. It follows immediately from Theorem 3.1 that a lack of complexity coherence implies the posteriors are incorrectly ordered when model priors $P(M)$ are uniform.

Corollary 3.2: Suppose model prior $P(M)$ is uniform. If homomorphic M_k and $M_{k'}$ fail complexity coherence on the pair of values v and v', where $v \prec v'$ or $v = v'$, then

$$Pr\{M_k \text{ truth} \mid d \text{ such that } EV(M_k,d)=v \text{ and } EV(M_{k'},d)=v'\}$$
$$> \quad Pr\{M_{k'} \text{ truth} \mid d \text{ such that } EV(M_k,d)=v \text{ and } EV(M_{k'},d)=v'\}$$

When the direction of inequality of the ratios is consistent for differing complexities — say the more complex consistently has the lower ratio — a systematic bias is indicated for the (for example) more complex to score better when it is not truth relative to when it is truth. Consequently, when such a non-complexity coherent evaluation measure is used, we need to be more suspect of the score of a complex model, since it is more likely to be random and less likely to be truth, when scores of simple and complex models are equal or near.

We emphasize that since we are considering *single* models of each complexity, the potential issues identified are model-complexity issues, rather than of model-space issues, i.e., the potential evaluation anomalies do not derive from the fact that there are more complex models than simple models in model space, nor from any other properties of model space, nor from any search bias governing which models are evaluated. Further — and perhaps most importantly — the anomalies do not depend on any predisposition for truth taking the form of models of one complexity over those of another.

3.2 Model Posterior and Data Likelihood

Data likelihood is a universally valid measure that yields posterior consistent evaluations with respect to specified priors. Our primary motivations for considering here such a well-known concept as likelihood are to: (a) review its exact computation in a simple model-evaluation setting; (b) derive closed forms for the distribution of likelihood scores in this setting; (c) use the score distributions as illustration of the potential behaviors of a coherent measure, which will both motivate further our definition of coherence and serve as contrast with the non-coherent measure behaviors demonstrated in Section 4; and, (d) provide the necessary distributional machinery for considering the model space issues treated in Section 5.

In the previous section, we considered quantities of the form

$$Pr\{d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v' \mid M \text{ truth}\}$$

and

$$Pr\{M \text{ truth} \mid d \text{ such that } EV(M,d) = v \text{ and } EV(M',d) = v'\}$$

In these quantities, the predicted and conditioning event $[EV(M,d) = v \text{ and } EV(M',d) = v']$ is a set $\{d_i\}$ of observations, all of whose members yield particular, common evaluation scores (v and v') on the models M and M'. This differs from predicting or conditioning on a particular observation d, which gives rise to the familiar data likelihood $Pr\{d \mid M \text{ truth}\}$ or model posterior $Pr\{M \text{ truth} \mid d\}$.

Likelihood score $L(M,d) = Pr\{d \mid M \text{ truth}\}$ is a sufficient summary of d in the sense that likelihood scores correctly rank the models by posterior, conditioned on d, assuming model priors are equal. Further, knowing the likelihood score $L(M_i,d)$ of each M_i on d, or knowing $Pr\{d\}$, provides the normalization constant for transforming likelihood to posterior and is equivalent to knowing d exactly with respect to computing posterior $Pr\{M \mid d\}$, that is,

$$Pr\{M \mid d\} \;=\; Pr\{M \mid L(M_i,d) \text{ for each } M_i\} \;=\; Pr\{M \mid L(M,d),\, Pr\{d\}\}$$

In terms of our characterization of evaluation measures, the key property of data likelihood is:

Fact 3.1: When model priors $P(M)$ are uniform, the highest likelihood model evaluated on d has the highest posterior for being truth conditioned on d. Further, the ordering of models by likelihood evaluated on d is consistent with how probable the model is conditioned on d. That is, for all data d and models M and M':

$$Pr\{M \text{ truth} \mid d\} \;<\; Pr\{M' \text{ truth} \mid d\} \;iff\; L(M,d) \;<\; L(M',d)$$

Proof: Self-evident. Posterior is proportional to likelihood when priors $P(M)$ are uniform.

Fact 3.1 states a stronger property than what is required by our definition of coherence. The fact states that the posterior ordering is consistent with likelihood for each individual d, whereas the coherence of measure EV requires a consistent ordering only for aggregations of d according to EV scores. That is, it follows from Fact 3.1 that (when priors $P(M)$ are equal)

$$\begin{aligned} &Pr\{M \text{ truth} \mid d \text{ such that } L(M,d) = v \text{ and } L(M',d) = v'\} \\ &\quad< \quad Pr\{M' \text{ truth} \mid d \text{ such that } L(M,d) = v \text{ and } L(M',d) = v'\} \end{aligned}$$

iff $v < v'$. This consistency, of course, is equivalent to the evaluation ratio (1) condition for coherence (see Theorem 3.1).

We observe that these key properties of likelihood follow outside the treatment here of evaluation measures in general and, consequently, do not depend on any of our model assumptions, such as the disjoint feature assumptions specified in Section 3.1.

We observe also that for a non-coherent EV, we have, for some model pairs M and M' and scores v and v'

$$\begin{aligned} & Pr\{M\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\} \\ > \quad & Pr\{M'\ truth \mid d\ such\ that\ EV(M,d) = v\ and\ EV(M',d) = v'\} \end{aligned}$$

when $EV(M,d) \prec EV(M',d)$. Hence, for at least some d

$$Pr\{M\ truth \mid d\} \ > \ Pr\{M'\ truth \mid d\}$$

when $EV(M,d) \prec EV(M',d)$, and Fact 3.1 fails to hold for non-coherent EV.

3.3 The Distribution of Likelihood Scores and Coherent Behavior

In this section, we analyze how likelihood scores $L(M,d)$ are distributed for random and true models M. This will illustrate why, in general, the evaluation ratio (1) must be considered rather than, for example, simply the quantity $Pr\{d\ such\ that\ EV(M,d) = v \mid M\ random\}$ alone, or rather than the marginal ratio (2). Additionally, the distributions of likelihood scores derived here will be applied in Section 5 when we consider model space issues.

We return now to the assumptions of Section 3.1, and consider the computation of likelihood when a single, known $G(\theta)$ is associated with each network G, and write model $M =< G, G(\theta) >$. The more common situation in which the associated distribution $G(\theta)$ is unknown to the evaluation often is approached by specifying a Dirichlet prior over the space Θ of distributions, leading to the Bayesian-Dirichlet (BD) metric. Consequences of this approach are considered in Section 4.3.

The likelihood evaluation of any model $M =< G, G(\theta) >$ given data $d = \{x_1, x_2, ..., x_i, ..., x_N\}$ is given by

$$P\{d \mid M\} \ = \ \prod_i^N Pr\{x_i \mid M\}$$

Because the assumed structure of G (see assumptions in Section 3.1) asserts that some subset $PS \subseteq U$ of features renders the class node conditionally independent of the remaining features, and that all features are independent of each other and each assumes a binary value with equal probability, we may write

$$\prod_i^N Pr\{x_i \mid M) \ = \ \prod_i^N (Pr\{C(x_i) \mid ps(x_i)\} * (\frac{1}{2^{|U|}})),$$

where $ps(x_i)$ is the parent cell of M to which the i^{th} observation of d falls, $C(x_i)$ is the class (1 or 2) of this observation, and $Pr\{C(x_i) \mid ps(x_i)\}$ is the conditional probability assigned by $G(\theta)$ to this cell's class probability. The term $\frac{1}{2^{|U|}}$ is the probability of the feature value combination observed in x_i — since the features take on combinations of values with equal probability, this is a constant not depending on the particular feature value combination in x_i or on the model M being evaluated. Hence, we omit the term and write simply

$$L(M,d) \ = \ \prod_i^N Pr\{C(x_i) \mid ps(x_i)\} \tag{3}$$

Note that, in the context of such Bayesian networks, this data likelihood is equivalent to conditional class likelihood [8]. Also, [11] demonstrates that this formulation can be used to evaluate parent set Bayesian network classifiers, even when edges may exist between the features.

Consider a Θ space of distributions in which any given model's distribution θ assigns the classes in each of its cells either pr or $(1-pr)$, for some single $0.5 \leq pr \leq 1$. When this assumption of a binary θ (that is, a θ assigning one of two values as each conditional class probability) is relaxed, the distributions become the more complicated multinomial rather binomial distributions, but no fundamental change in formulation is required.

Under the assumption of a binary θ, the possible likelihood scores that a model $M =< G, G(\theta) >$ can achieve are

$$L(M,d) \;=\; P\{d \mid M\} \;=\; pr^H \;*\; (1-pr)^{(N-H)}$$

for $H \;=\; 0,1,...,N$. In particular, the score $pr^H \;*\; (1-pr)^{(N-H)}$ is achieved exactly when H of the N observations x_i of d take on the probability-pr class value b of the cell $ps(x_i)$ to which x_i falls. (That is, the value of the class $b \in \{1,2\}$ is such that $Pr\{C = b \mid ps(x_i)\} \;=\; pr$.)

In accordance with our general definition of coherence, which considers the ratio of the distributions of EV scores conditioned on true and random models, we consider these distributions for likelihood scores. In the case of likelihood, note that we are effectively considering the probability of a d such that the probability of d being generated is equal to some value v, conditioned on the model being true or random. That is, we are considering the collective probability, conditioned on M being true or random, of the set of d's with likelihood score $L(M,d) = v$.

Theorem 3.2: Let M be any model with associated θ that is such that θ assigns the classes in each of its cells either pr or $(1-pr)$, for some single $0.5 \leq pr \leq 1$, in any proportion. Then

$$\begin{aligned} &Pr\{d \;such\; that\; L(M,d) \;=\; pr^H \;*\; (1-pr)^{(N-H)} \mid M\; true\} \\ &= \;\; Binomial(N,H,pr) \\ &= \;\; \left(\begin{array}{c} N \\ H \end{array} \right) \;*\; (pr^H * (1-pr)^{(N-H)}) \end{aligned}$$

for $H \;=\; 0,1,...,N$, and the probability of $L(M,d)$ assuming any other score is 0.

Proof: For each observation $x \in d$ that M generates, x falls into either a pr cell or a $(1-pr)$ cell of θ. The probability that M generates an x falling into a pr cell of θ is pr, regardless of the parent cell to which x falls, i.e., if x falls to parent cell ps, and if $Pr\{C = b \mid ps\} \;=\; pr$, x will be of class b with probability pr and hence fall into the pr cell with probability pr. Consequently, the probability of M generating a d such that $Pr\{d \mid M\; true\} \;=\; pr^H \;*\; (1-pr)^{(N-H)}$ is distributed as the binomial distribution $Binomial(N,H,pr)$.

□

Note that the theorem asserts that, for any binary θ,

$$Pr\{d \;such\; that\; L(M,d) = v \mid M\; truth\} \;=\; \left(\begin{array}{c} N \\ H \end{array} \right) \;*\; Pr\{d \mid M\; truth\}$$

for any d achieving likelihood score $v \;=\; pr^H \;*\; (1-pr)^{(N-H)}$.

While Theorem 3.2 holds regardless of the proportion of ps for which θ assigns $Pr\{C = 1 \mid ps\}$ and $Pr\{C = 2 \mid ps\}$ the probability pr, the distribution for random models is considered first under the assumption of a symmetric θ, defined as follows.

Definition 3.3: Distribution θ is *symmetric* if pr is assigned as the conditional class probability to $Pr\{C = 1 \mid ps\}$ and $Pr\{C = 2 \mid ps\}$ with equal frequency, and hence the unconditional class probabilities are equal, i.e., $Pr\{C = 1\} \;=\; Pr\{C = 2\} \;=\; 0.5$.

Theorem 3.3: Let M be any model with associated θ that symmetrically assigns the classes in each of its cells either pr or $(1-pr)$, for some single $0.5 \leq pr \leq 1$. Then

$$Pr\{d \;such\; that\; L(M,d) \;=\; pr^H \;*\; (1-pr)^{(N-H)} \mid M\; random\}$$

$$
\begin{aligned}
&= Binomial(N,H,0.5) \\
&= \left(\begin{array}{c} N \\ H \end{array} \right) * (0.5^H * (1-0.5)^{(N-H)}) \\
&= \left(\begin{array}{c} N \\ H \end{array} \right) * 0.5^N
\end{aligned}
$$

for $H = 0,1,...,N$, and the probability of $L(M,d)$ assuming any other score is 0.

Proof: For each observation $x \in d$ that the true model M^* (different from M) generates, x falls into either a pr cell or a $(1-pr)$ cell of the θ associated with M. This depends on the parent cell ps of M to which x falls, and the class of case x. The class b of case x is determined by M^*'s generation of x, but by the disjointness assumption, M^* does not affect this ps of M, all of which are equally likely. Since θ is symmetric, it is equally likely that the ps to which x falls assigns pr to class b as it is that it assigns $(1-pr)$ to class b. Consequently, the probability of M^* generating a d such that $Pr\{d \mid M\ random\} = pr^H * (1-pr)^{(N-H)}$ is distributed as the binomial distribution $Binomial(N,H,0.5)$.

□

Note that the theorem asserts that, for any binary symmetric θ,

$$
Pr\{d\ such\ that\ L(M,d) = v \mid M\ random\} = \left(\begin{array}{c} N \\ H \end{array} \right) * Pr\{d \mid M\ random\}
$$

for any d achieving likelihood score $v = pr^H * (1-pr)^{(N-H)}$. See Example 3.1 below for the consequences on the distribution of scores of random models of relaxing the restriction to a symmetric θ.

We noted in the previous section that likelihood is a coherent measure, and that this result does not depend on any assumptions on θ; the examples which follow illustrate different forms that this coherent behavior can take. The evaluation ratio (1) used to characterize coherence takes a particularly simple form for likelihood when θ is symmetric, and we find this form useful in the analysis developed in Section 5. Theorem 3.4 establishes that when θ is symmetric, the distribution of likelihood scores of any model M_i (random or true) does not depend on knowledge of the scores achieved by any other model, and when M_i is random, the score of M_i also does not depend on the identity of the true model.

Theorem 3.4: If for each model M, there is a single pr such that θ symmetrically assigns the classes in M's parent cell either pr or $(1-pr)$, for some single $0.5 \leq pr \leq 1$, then
(A) For M different from the true model M^* and score $v = pr^H * (1-pr)^{(N-H)}$:

$$
\begin{aligned}
&Pr\{d\ such\ that\ L(M,d) = v \mid M^*\ truth,\ L(M_i,d)\ for\ all\ models\ M_i\ other\ than\ M\} \\
&\quad = Pr\{d\ such\ that\ L(M,d) = v \mid NOT\ M\} \\
&\quad = Binomial(N,H,0.5)
\end{aligned}
$$

(B) For the true model M^* and score $v = pr^H * (1-pr)^{(N-H)}$:

$$
\begin{aligned}
&Pr\{d\ such\ that\ L(M^*,d) = v \mid M^*\ truth,\ L(M_i,d)\ for\ all\ models\ M_i\ other\ than\ M^*\} \\
&\quad = Pr\{d\ such\ that\ L(M^*,d) = v \mid M^*\ truth\} \\
&\quad = Binomial(N,H,pr)
\end{aligned}
$$

Proof:
(A) The event $[NOT\ M]$ is equivalent to $[M\ random]$. Random M achieves score $v = pr^H * (1-pr)^{(N-H)}$ exactly

when H of the N cases of data d fall to the pr classes of the ps cells of M to which they fall. Without knowledge of to which ps cell of M a case x falls, even with knowledge of the class $C(x)$ of x, the probability of x falling to the pr class of its $ps(x)$ is 0.5, since θ is symmetric and all ps cells are equally likely, independent of knowledge of the values of all remaining features (other than those in M's parent set). Since the conditioning event

$$[M^* \; truth, \; L(M_i, d) \; for \; all \; models \; M_i \; other \; than \; M]$$

does not change the distribution of ps cells of M from equally likely (since parent sets are disjoint and their features are statistically independent), the probability remains $Binomial(N, H, 0.5)$.

(B) M^* achieves score $= pr^H * (1-pr)^{(N-H)}$ exactly when H of the N cases of data d fall to the pr class of the ps cells of M^* to which they fall. The score of any random model on d is statistically independent of the distribution of the classes in d since every θ is symmetric, and does not affect the ps distribution of M^*, since parent sets are disjoint and their features statistically independent. Since the conditioning event

$$[L(M_i, d) \; for \; all \; models \; M_i \; other \; than \; M^*]$$

does not change the distribution[2] of the classes in d or of the ps cells of M^* from all equally likely, the probability remains $Binomial(N, H, pr)$.

□

As will be seen for apparent error rate in Sections 4.1, and for the BD metric in Section 4.3, Theorem 3.4 does not hold for all measures. Further, even for likelihood, the result of Theorem 3.4 requires that θ be symmetric. That is, as the following example demonstrates, when θ is not symmetric, for $M \neq M^*$ we may fail to have either

$$Pr\{d \; such \; that \; L(M,d) = v \mid M^* \; truth, \; L(M^*,d) = v'\} = Pr\{d \; such \; that \; L(M,d) = v \mid M^* \; truth\}$$

or

$$Pr\{d \; such \; that \; L(M,d) = v \mid M^* \; truth\} = Pr\{d \; such \; that \; L(M,d) = v \mid NOT \; M\}$$

when θ is not symmetric. The lack of independence occurs because, without a symmetric θ, knowledge of the true model, or of its likelihood score, can affect the distribution of a random model's likelihood score, by leaking information of the relative class frequencies in d. In such a case, the behavior of the simple marginal evaluation ratio (2) can be misleading to an analysis of a measure's coherence. This is illustrated in the following example.

Example 3.1: Let $M_1, ..., M_q$ be members of $card_2$ (that is, each has a parent set of size 2), such that for each of these M_i, $Pr\{C = 1|ps_1\} = Pr\{C = 1|ps_2\} = Pr\{C = 1|ps_3\} = 1$ and $Pr\{C = 2|ps_4\} = 1$. Let the remaining model with nonzero prior probability M' be in $card_2$ as well, and suppose M' reverses the proportion of class probabilities: $Pr\{C = 2|ps_1\} = Pr\{C = 2|ps_2\} = Pr\{C = 2|ps_3\} = 1$ and $Pr\{C = 1|ps_4\} = 1$. All $q+1$ models respect the disjoint feature assumption. Note that for any of these $q+1$ models M,

$$Pr\{d \; such \; that \; L(M,d) = 1 \mid M \; truth\} = 1.0,$$

since each parent state of each model functionally determines the class label. Intuitively, each of the M_i models has a greater chance of scoring $L(M_i, d) = 1$ when it is random than does M' when it is random, because when

[2] Observe that the result would continue to hold even if we conditioned on full knowledge of every data value in the cases of d, including the class label but excluding the values of M^*'s parent set features; similarly, the result would continue to hold if we conditioned on the values of M^*'s parent features, but excluded each case's class label.

M_i is random, it is still extremely likely (assuming large enough q) that another M_j with the same allocation of conditional class probabilities generates the data. On the other hand, when M' is random, it is far less likely that it scores $L(M',d) = 1$, because when M' is random, some M_j with a reverse allocation of class conditional probabilities generates the data. This discrepancy at first may seem to suggest that likelihood is not coherent: two models M' and M_i attaining the same likelihood score don't appear to imply the same posterior. But this example illustrates only the danger of ignoring the inter-dependence of model scores by considering the behavior of the simple marginal ratio (2), or, similarly, the marginal posterior $Pr\{M|L(M,d) = v\}$.
To make the example concrete, consider data d consisting of a single observation. As observed above,

$Pr\{d \text{ such that } L(M_i,d) = 1 \mid M_i \text{ truth}\} = 1.0$
$Pr\{d \text{ such that } L(M',d) = 1 \mid M' \text{ truth}\} = 1.0$

since parent states functionally determine the class. Additionally, we compute:

$$\begin{aligned} & Pr\{d \text{ such that } L(M',d) = 1 \mid NOT\ M'\} \\ = \;& Pr\{d \text{ such that } L(M',d) = 1 \mid M_j \text{ truth, for some } 1 \le j \le q\} \\ = \;& Pr\{x_1 \text{ falls to one of the } C = 1 \ M_j \text{ cells}\} * Pr\{x_1 \text{ falls to the sinlge } C = 1 \ M' \text{ cell}\} \\ & + Pr\{x_1 \text{ falls to the single } C = 2 \ M_j \text{ cell}\} * Pr\{x_1 \text{ falls to one of the } C = 2 \ M' \text{ cells}\} \\ = \;& (3/4)*(1/4) \ + \ (1/4)*(3/4) \ = \ 6/16 \end{aligned}$$

$$\begin{aligned} & Pr\{d \text{ such that } L(M_i,d) = 1 \mid NOT\ M_i\} \\ \approx \;& Pr\{d \text{ such that } L(M_i,d) = 1 \mid M_j \text{ truth, for some } 1 \le j \le q, i \ne j\} \\ = \;& Pr\{x_1 \text{ falls to one of the } C = 1 \ M_j \text{ cells}\} * Pr\{x_1 \text{ falls to one of the } C = 1 \ M_i \text{ cells}\} \\ & + Pr\{x_1 \text{ falls to the single } C = 2 \ M_j \text{ cell}\} * Pr\{x_1 \text{ falls to the single } C = 2 \ M_i \text{ cell}\} \\ = \;& (3/4)*(3/4) \ + \ (1/4)*(1/4) \ = \ 10/16 \end{aligned}$$

(The approximation is valid for large q, in which case the event that $[M' \text{ is truth}]$, conditioned on $[NOT\ M_i]$, has a negligible contribution.)
Consequently,

$$\frac{Pr\{d \text{ such that } L(M_i,d) = 1 \mid M_i \text{ truth}\}}{Pr\{d \text{ such that } L(M_i,d) = 1 \mid NOT\ M_i\}} \approx \frac{16}{10}$$

while

$$\frac{Pr\{d \text{ such that } L(M',d) = 1 \mid M' \text{ truth}\}}{Pr\{d \text{ such that } L(M',d) = 1 \mid NOT\ M'\}} = \frac{16}{6}$$

This implies, when model priors are equal,

$Pr\{M_i \mid L(M_i,d) = 1\} \ < \ Pr\{M'|L(M',d) = 1\}$

and there is a temptation to conclude that likelihood is not coherent!
However, the pertinent issue to consider is, what can we conclude if, on a particular d, we observe that both M_i and M' achieve the same likelihood score of 1? That is, what can we conclude about the posteriors

$Pr\{M_i \text{ truth} \mid L(M_i,d) = 1 \text{ and } L(M',d) = 1\}$ *vs.* $Pr\{M' \text{ truth} \mid L(M_i,d) = 1 \text{ and } L(M',d) = 1\}$?

This leads us to the evaluation ratios (1) used in the definition of coherence, whose denominators in the current example are

$Pr\{d \text{ such that } L(M_i,d) = 1 \mid M' \text{ truth, } L(M',d) = 1\}$ and
$Pr\{d \text{ such that } L(M',d) = 1 \mid M_i \text{ truth, } L(M_i,d) = 1\}$.

$Pr\{d \text{ such that } L(M',d) = 1 \mid M_i \text{ truth}, L(M_i,d) = 1\}$ is the same as $Pr\{d \text{ such that } L(M',d) = 1 \mid NOT\ M'\}$ and, as computed above, equals $\frac{6}{16}$. But now the other denominator,

$Pr\{d \text{ such that } L(M_i,d) = 1 \mid M' \text{ truth}, L(M',d) = 1\}$,

clearly has this same value, i.e.,

$$\begin{aligned} & Pr\{d \text{ such that } L(M_i,d) = 1 \mid M' \text{ truth}, L(M',d) = 1\} \\ = \;& Pr\{x_1 \text{ falls to one of the } C = 2\ M' \text{ cells}\} * Pr\{x_1 \text{ falls to the single } C = 2\ M_i \text{ cell}\} \\ & + Pr\{x_1 \text{ falls to the single } C = 1\ M' \text{ cell}\} * Pr\{x_1 \text{ falls to one of the } C = 1\ M_i \text{ cell}\} \\ = \;& (3/4) * (1/4) \;+\; (1/4) * (3/4) \;=\; 6/16 \end{aligned}$$

Consequently, the evaluation ratios are the same and hence

$Pr\{M_i \text{ truth} \mid L(M_i,d) = 1 \text{ and } L(M',d) = 1\} \;=\; Pr\{M' \text{ truth} \mid L(M_i,d) = 1 \text{ and } L(M',d) = 1\}$

□

Examples 3.2 and 3.3 illustrate two types of scoring behavior for symmetric θ. In this case, when comparing the score ratios for M and M', applying Theorem 3.4 we simplify the denominator

$Pr\{d \text{ such that } L(M,d) = v \mid M' \text{ truth}, L(M',d) = v'\}$

to the equivalent

$Pr\{d \text{ such that } L(M,d) = v \mid NOT\ M\}$.

Example 3.2: If $M_k \in card_k$ and $M_{k'} \in card_{k'}$ are homomorphic (see Definition 3.2), each assigning to $Pr\{C = 1 \mid ps\}$ either pr or $(1 - pr)$ symmetrically, then each of the numerator and denominator of the coherence ratio is equal between M_k and $M_{k'}$ for every value v, that is,

$Pr\{d \text{ such that } L(M_k,d) = v \mid M_k \text{ truth}\} \;=\; Pr\{d \text{ such that } L(M_{k'},d) = v \mid M_{k'} \text{ truth}\}$

and

$Pr\{d \text{ such that } L(M_k,d) = v \mid M_k \text{ random}\} \;=\; Pr\{d \text{ such that } L(M_{k'},d) = v \mid M_{k'} \text{ random}\}$

This follows directly from the distributions specified by Theorems 3.2 and 3.3 above. From Theorem 3.4, we thus have also

$$\begin{aligned} & Pr\{d \text{ such that } L(M_k,d) = v \mid M_{k'} \text{ truth}, L(M_{k'},d) = v\} \\ = \;& Pr\{d \text{ such that } L(M_{k'},d) = v \mid M_k \text{ truth}, L(M_k,d) = v\} \end{aligned}$$

yielding a particularly simple form of coherent behavior.

□

Example 3.3: When the conditional probabilities assigned to class labels differ across models (and hence models are not homomorphic), a more complex interaction than in Example 3.2 is seen, even when the individual θ remain symmetric. Suppose $\theta_1 = M_1(\theta)$ assigns probabilities $pr_1 = 0.99$ and $(1 - pr_1) = 0.01$ to the classes conditioned on parent states ps, while $\theta_2 = M_2(\theta)$ assigns probabilities $pr_2 = 0.6$ and $(1 - pr_2) = 0.4$, with every θ symmetric. We compute distributions for likelihood score v_1 achieved by M_1 and scores v_{2a} and v_{2b} achieved by M_2, specified as:

$$\begin{aligned} v_1 &= ln(0.99^{89} * 0.01^{11}) = -51.551352\ (i.e., H_1 = 89) \\ v_{2a} &= ln(0.60^{98} * 0.40^{2}) = -51.893493\ (i.e., H_{2a} = 98) \\ v_{2b} &= ln(0.60^{99} * 0.40^{1}) = -51.488027\ (i.e., H_{2b} = 99) \end{aligned}$$

The computations are performed with respect to data d of size 100 observations, and log likelihoods (denoted $LL(M,d)$ below) and logs of probabilities are used to avoid numerical underflow.
Note that the scores v_1, v_{2a}, and v_{2b} are nearly identical, with $v_{2a} < v_1 < v_{2b}$. (The combinatorics do not yield a score appropriate for this illustration that is achievable exactly by both the models.) The following quantities are computed analytically:

$$ln(Pr\{d \text{ such that } LL(M_1,d) = v_1 | M_1 \text{ true}\}) \;=\; -18.967114$$
$$ln(Pr\{d \text{ such that } LL(M_1,d) = v_1 | M_1 \text{ random}\}) \;=\; -36.730480$$
$$ln(R(M_1,v_1)) = ln(\frac{Pr\{d \text{ such that } LL(M_1,d) = v_1 | M_1 \text{ true}\}}{Pr\{d \text{ such that } LL(M_1,d) = v_1 | M_1 \text{ random}\}}) \;=\; 17.76336$$

$$ln(Pr\{d \text{ such that } LL(M_2,d) = v_{2a} | M_2 \text{ true}\}) \;=\; -43.386350$$
$$ln(Pr\{d \text{ such that } LL(M_2,d) = v_{2a} | M_2 \text{ random}\}) \;=\; -60.807575$$
$$ln(R(M_2,v_{2a})) = ln(\frac{Pr\{d \text{ such that } LL(M_2,d) = v_{2a} | M_2 \text{ true}\}}{Pr\{d \text{ such that } LL(M_2,d) = v_{2a} | M_2 \text{ random}\}}) \;=\; 17.42125$$

$$ln(Pr\{d \text{ such that } LL(M_2,d) = v_{2b} | M_2 \text{ true}\}) \;=\; -46.882857$$
$$ln(Pr\{d \text{ such that } LL(M_2,d) = v_{2b} | M_2 \text{ random}\}) \;=\; -64.709548$$
$$ln(R(M_2,v_{2b})) = ln(\frac{Pr\{d \text{ such that } LL(M_2,d) = v_{2b} | M_2 \text{ true}\}}{Pr\{d \text{ such that } LL(M_2,d) = v_{2b} | M_2 \text{ random}\}}) \;=\; 17.826691$$

Observe how it is far more common for the model M_1 with conditional class probabilities $pr_1 = 0.99$ to achieve the score v_1 than it is for the model M_2 with conditional class probabilities $pr_2 = 0.60$ to achieve either of the nearly identical scores v_{2a} or v_{2b}. However, this tendency applies to the models both when random and true, and, consequently, the ratios (and hence the posteriors) are also nearly identical, and are ordered consistently with the scores, that is

$$R(M_2,v_{2a}) \;<\; R(M_1,v_1) \;<\; R(M_2,v_{2b})$$

To understand why the ratio (and hence the posteriors) will always be ordered as v and v' are ordered, observe that the binomial coefficient in the numerator and denominator of the ratio above for each v_j ($j = 1, 2a, 2b$) is identical, i,e, $\begin{pmatrix} N \\ H_j \end{pmatrix}$, and hence $Binomial(N,H_j,pr_j)/Binomial(N,H_j,0.5)$ reduces to the likelihood of the data given M_j divided by $0.5^{H_j} * 0.5^{(N-Hj)} \;=\; (0.5)^N$. The fact that the ratios are ordered as the scores follows from the fact that, after this cancellation of the $\begin{pmatrix} N \\ H_j \end{pmatrix}$ terms, the denominators are a common 0.5^N, and thus the order of the likelihood scores determines the order of the ratios. Even in cases when θ is not symmetric and the joint version does not reduce to the marginal ratio, the ratio for likelihood is coherent, though the analysis is considerably more complicated since it cannot be done in terms of the simpler marginal ratio.

□

An important implication of Example 3.3 is that the behavior of the quantity

$$Pr\{d \text{ such that } EV(M,d) \;=\; v \mid M \text{ random}\}$$

alone reveals little regarding model posteriors. That is, analyzing how likely it is for a random model with certain characteristics (e.g., peaked probabilities or high complexities) to achieve a good observed score by chance alone

(e.g., an observed score p-value) is of little relevance to the questions of model evaluation and model selection considered here.

4. Some Important Non-coherent Measures

4.1 Apparent Training Error

Apparent training error is one of the simplest, most intuitive, and widely used measures in application and theoretical analysis. This evaluation measure is most often applied to classification trees without reference to an associated distribution θ as, for example, in [3, 17, 20, 23, 30]. As was described in Section 3.1, our probability models can be viewed from the perspective of homogeneous classification trees, with the parent cells of M corresponding to the leaves of the tree. Equivalently, from the Bayesian network perspective, though M consists of a network structure G and a distribution θ, the analogous application of apparent training error does not depend on θ. In both modelings, the apparent error rate is a simple function of the model structure and the frequentist distribution this structure derives from the data d.

Definition 4.1: Let M be a model with parent cell ps, and suppose that for data d there are n and m members of each class falling to this cell. The smaller number $min(n,m)$ i.e., the minority count, is the *apparent error* for this ps (error is n, if $n = m$). The *apparent error rate* $ER(M,d)$ for the model M on data d is the sum over all of M's ps_i of these apparent errors.

In Section 4.2, apparent error rate is defined in the context of cross validation.

It is easy to see that ER is not a coherent measure, and fails on even the minimal requirement of complexity coherence, where the models evaluated differ only in their complexities. Consider, for example, a perfect score of $ER = 0$. As the number of parents k becomes arbitrarily large, and the size of data d is held fixed, the probability of more than one case x_i from d falling into the same parent cell approaches zero, regardless of whether the parents are correlated with the class or not, and regardless of the relative class counts in d. Since parent cells with fewer than two cases cannot incur observed errors, the total number of observed errors is zero with probability approaching 1.0, for both random and true models of high complexity, and thus the evaluation ratio (1) for achieving $ER = 0$ goes to 1 as k increases. However, when k is small, for any true model in which some conditional class probability $pr > 0.5$, the ratio for achieving $ER = 0$ is bounded away from 1 from above.

The result is most easily established with symmetric θ assigning a single pr as conditional class probabilities.

Theorem 4.1: Let $M_1 \in card_1$ and $M_k \in card_k$, with all θ symmetric and assigning a single $0.5 < pr \leq 1.0$. Then

$$\frac{Pr\{d \text{ such that } ER(M_1,d) = 0 \mid M_1 \text{ truth}\}}{Pr\{d \text{ such that } ER(M_1,d) = 0 \mid M_k \text{ truth and } ER(M_k,d) = 0\}} > (1+\gamma)$$

for some $\gamma > 0$, depending only on pr and $|d|$, and not k, while, as k increases,

$$\frac{Pr\{d \text{ such that } ER(M_k,d) = 0 \mid M_k \text{ truth}\}}{Pr\{d \text{ such that } ER(M_k,d) = 0 \mid M_1 \text{ truth and } ER(M_1,d) = 0\}}$$

becomes arbitrarily close to 1.0, for any fixed pr and $|d|$.

Proof: So long as $pr > 0.5$,

$$\frac{Pr\{d \text{ such that } ER(M_1,d) = 0 \mid M_1 \text{ truth}\}}{Pr\{d \text{ such that } ER(M_1,d) = 0 \mid NOT\ M_1\}}$$

is clearly greater than 1.0, and increases as size of d or pr increases. The dependence on $[M_k \; truth \; and \; ER(M_k,d) = 0]$ of the denominator

$$Pr\{d \; such \; that \; ER(M_1,d) = 0 \mid M_k \; truth \; and \; ER(M_k,d)\} = 0$$

of the evaluation ratio is diminished as k increases, since for k arbitrarily large and size of d fixed, $ER(M_k,d) = 0$ with probability approaching 1 and hence no information is gained from the event $[ER(M_k,d) = 0]$. Further, since all θ are symmetric, no information is obtained from the event $[M_k \; truth]$. Hence, for sufficiently large k,

$$Pr\{d \; such \; that \; ER(M_1,d) = 0 \mid M_k \; truth \; and \; ER(M_k,d) = 0\}$$

behaves as

$$Pr\{d \; such \; that \; ER(M_1,d) = 0 \mid NOT \; M_1\}$$

Each of the numerator and denominator of

$$\frac{Pr\{d \; such \; that \; ER(M_k,d) = 0 \mid M_k \; truth\}}{Pr\{d \; such \; that \; ER(M_k,d) = 0 \mid M_1 \; truth \; and \; ER(M_1,d) = 0\}}$$

is made arbitrarily close to 1.0 by increasing k so that none of the 2^k parent cells have more than a single data point with nonvanishing probability. (And note, as a concrete example of the nonequality of ratios if $pr = 1$ and $|d| = 2$, the M_1 ratio evaluates to approximately $1.0/(0.75) = 1.33$, while the second approaches 1.0 as k is increased.) □

Theorem 4.1 establishes that ER is not even complexity coherent, since the models considered are homomorphic. Consequently, when a complex and simple model score the same low score (e.g., 0 errors) on some data, the simpler model has a higher true posterior, assuming nothing beyond equal model priors. Critically, in contrast to the very correct objections such as those in [23, 30] against the existence of absolute complexity biases, the complexity bias for the non-coherent ER is distribution free – it is present provided only that θ has some pr different from 0.5 (and that nonuniform model priors don't push the posteriors in the opposite direction). The result, in particular, does not depend on a prior over model space favoring the selection for the true model M^* of simple models over complex models.

Example 4.1: Even when k is relatively small, the complexity bias can be quite pronounced. The results below show the behavior of the ratio for $k = 1$ and $k' = 5$ and the lowest error rates of 0-4. The θ are symmetric, assigning $Pr\{C = 1|ps\} = 0.8$ and $Pr\{C = 2|ps\} = 0.8$ equally often. The probabilities are estimated from a generation of 100 million d_i, each of size 10 observations. Repeated runs show the results to be stable.

$$Pr\{ER(M_1,d) = 0 \mid M_1 \; truth\} \; / \; Pr\{ER(M_1,d) = 0 \mid M_5 \; truth \; and \; ER(M_5,d) = 0\} \approx 22.082228$$
$$Pr\{ER(M_5,d) = 0 \mid M_5 \; truth\} \; / \; Pr\{ER(M_5,d) = 0 \mid M_1 \; truth \; and \; ER(M_1,d) = 0\} \approx 1.296588$$

$$Pr\{ER(M_1,d) = 1 \mid M_1 \; truth\} \; / \; Pr\{ER(M_1,d) = 1 \mid M_5 \; truth \; and \; ER(M_5,d) = 1\} \approx 9.461787$$
$$Pr\{ER(M_5,d) = 1 \mid M_5 \; truth\} \; / \; Pr\{ER(M_5,d) = 1 \mid M_1 \; truth \; and \; ER(M_1,d) = 1\} \approx 0.770556$$

$$Pr\{ER(M_1,d) = 2 \mid M_1 \; truth\} \; / \; Pr\{ER(M_1,d) = 2 \mid M_5 \; truth \; and \; ER(M_5,d) = 2\} \approx 2.682664$$
$$Pr\{ER(M_5,d) = 2 \mid M_5 \; truth\} \; / \; Pr\{ER(M_5,d) = 2 \mid M_1 \; truth \; and \; ER(M_1,d) = 2\} \approx 0.460982$$

$$Pr\{ER(M_1,d) = 3 \mid M_1 \; truth\} \; / \; Pr\{ER(M_1,d) = 3 \mid M_5 \; truth \; and \; ER(M_5,d) = 3\} \approx 0.647048$$
$$Pr\{ER(M_5,d) = 3 \mid M_5 \; truth\} \; / \; Pr\{ER(M_5,d) = 3 \mid M_1 \; truth \; and \; ER(M_1,d) = 3\} \approx 0.257815$$

$$Pr\{ER(M_1,d) = 4 \mid M_1 \; truth\} \; / \; Pr\{ER(M_1,d) = 4 \mid M_5 \; truth \; and \; ER(M_5,d) = 4\} \approx 0.133671$$
$$Pr\{ER(M_5,d) = 4 \mid M_5 \; truth\} \; / \; Pr\{ER(M_5,d) = 4 \mid M_1 \; truth \; and \; ER(M_1,d) = 4\} \approx 0.118234$$

Note, for example, on data d of size 10 for which M_1 and M_5 simultaneously incur 0 errors, the above results, combined with Corollary 3.1, allow us to conclude that M_1 is more probable than is M_5 by a factor of greater than 16, assuming equal model priors. Further, we see from the following that on data d of size 10 for which M_1 incurs 1 error and M_5 simultaneously incurs 0 errors, M_1 is more probable than is M_5 by a factor of greater than 4, assuming equal model priors. The model parameters are the same as above, and 100 million generations of d_i again are used to estimate the probabilities.

$Pr\{ER(M_1,d) = 1 \mid M_1\ truth\}\ /\ Pr\{ER(M_1,d) = 1 \mid M_5\ truth\ and\ ER(M_5,d) = 0\} \approx 6.347174$
$Pr\{ER(M_5,d) = 0 \mid M_5\ truth\}\ /\ Pr\{ER(M_5,d) = 0 \mid M_1\ truth\ and\ ER(M_1,d) = 1\} \approx 1.291417$

□

4.2 Cross Validation and the Apparent Error Rate Measure

It is well known that the apparent error rate of the best fitting of many models under-estimates the true out-of-sample error rate of this model. If many random models exist, at least some will fit the training data well, possibly with zero errors, if there are enough models relative to the size of the training data. The results of the previous section imply that individual complex models are more prone to this problem, since when they have low apparent error rates they are individually less likely than simple models with the same error rate to be the true model.

Leave one out cross validation (LOOCV) is a technique that yields low bias estimates of expected out-of-sample classification error, though with high variance. Consequently, we examined whether using the LOOCV error rate yields a complexity coherent measure in the context of the model selection problem under study. Our results indicate that this is not the case. Specifically, we consider a scenario in which a collection of potential models M are evaluated using cross validation error XER, defined below, and again conclude that when individual complex models score well on this measure they are individually less likely than simple models with the same score to be the true model. Note that in this model evaluation scenario, the structural specification (i.e., the identity of a model's parent set features, which, by definition, are binary and not to be binned) of each model M so evaluated is fixed in advance and is independent of the cross validation procedure. In particular, the only characteristic of the model M that is fold dependent is the parent cell class counts, which determine the classification of each held-out case.

The LOOCV error is the number of held-out cases that are evaluated incorrectly when the majority class rule with respect each fold's in-sample cases is used to evaluate the held-out case. Since LOOCV is deterministic in the sense that every case is a hold-out case exactly once, the LOOCV error of any model is determined by the model's parent set conditioning of the entire training set into ps cells. That is, we can determine exactly what the LOOCV error of a model M will be by examining the statistical breakdown on the entire training set of that model's parent cells ps. This leads to the following definition of LOOCV apparent error rate, which agrees with the number of errors incurred by M using the standard LOOCV procedure.

Definition 4.2: Suppose a model M induces on the entire training set d a ps_i cell class breakdown of $(n\,Class_1, m\,Class_2)$. Then the LOOCV error contributed by this cell ps_i on data d can be computed as:

a) If $n+1 < m$, then n. Every one of these n held out Class 1 cases is in the minority for its fold, and hence each incurs an error. Every one of these held out Class 2 cases is in the majority of its fold and is classified correctly.

b) If $n > m+1$, then m. Symmetric argument to case (a)

c) If $n = m$, then $(n+m)$. Every held out case is in the minority of its fold.

d) If $n+1 = m$, then $n + \frac{m}{2}$. Every one of the n held out Class 1 cases is in the minority for its fold, and hence each incurs an error. Every one of the held out Class 2 cases evaluates as a tie in its fold and hence incurs half an error.

e) If $n = m+1$, then $\frac{n}{2} + m$. Symmetric argument to case (d).

The total LOOCV error $XER(M,d)$ of model M incurred on data d is the sum of the LOOCV errors contributed by M's ps_i on d.

Note that in cases (a) and (b) XER gives the same error contribution as ER, while in case (c), (d), and (e) the contribution of XER is higher.

The results below for homomorphic models M_1 and M_2 of size 1 and 2 parent sets demonstrate that the evaluation ratio for the lowest achievable LOOVC error scores (0 and $\frac{1}{2}$) each is larger for the simpler model M_1 than for the more complex model M_2, demonstrating that LOOCV error rate is not a complexity coherent evaluation. This means that, when the simple model and the complex model score 0 LOOCV errors on the same data, and when each model scores $\frac{1}{2}$ error on the same data, the simpler is more likely to be truth, in each of these cases. That is, the model posteriors behave as

$$\begin{aligned} & Pr\{M_1 \; truth \mid d \; such \; that \; XER(M_1,d) = 0 \; and \; XER(M_2,d) = 0 \; \} \\ & \quad > \quad Pr\{M_2 \; truth \mid d \; such \; that \; XER(M_1,d) = 0 \; and \; XER(M_2,d) = 0 \; \} \end{aligned}$$

and

$$\begin{aligned} & Pr\{M_1 \; truth \mid d \; such \; that \; XER(M_1,d) = \frac{1}{2} \; and \; XER(M_2,d) = \frac{1}{2} \; \} \\ & \quad > \quad Pr\{M_2 \; truth \mid d \; such \; that \; XER(M_1,d) = \frac{1}{2} \; and \; XER(M_2,d) = \frac{1}{2} \; \} \end{aligned}$$

Example 4.2: Let $M_1 \in card_1$ and $M_2 \in card_2$, and the θ be symmetric, assigning $Pr\{C = 1|ps\} = 0.8$ and $Pr\{C = 2|ps\} = 0.8$ equally often. The following probabilities are estimated from a generation of 1 billion d_i, each of size 10 observations.[3] Repeated runs show the results to be stable.

$Pr\{XER(M_1,d) = 0 \mid M_1 \; truth\} \; / \; Pr\{XER(M_1,d) = 0 \mid M_2 \; truth \; and \; XER(M_2,d) = 0\} \approx 6.569512$
$Pr\{XER(M_2,d) = 0 \mid M_2 \; truth\} \; / \; Pr\{XER(M_2,d) = 0 \mid M_1 \; truth \; and \; XER(M_1,d) = 0\} \approx 5.110326$
$Pr\{XER(M_1,d) = \frac{1}{2} \mid M_1 \; truth\} \; / \; Pr\{XER(M_1,d) = \frac{1}{2} \mid M_2 \; truth \; and \; XER(M_2,d) = \frac{1}{2}\} \approx 7.629853$
$Pr\{XER(M_2,d) = \frac{1}{2} \mid M_2 \; truth\} \; / \; Pr\{XER(M_2,d) = \frac{1}{2} \mid M_1 \; truth \; and \; XER(M_1,d) = \frac{1}{2}\} \approx 0.516692$

We compared also ratios on data d of size 10 for which M_1 incurs $\frac{1}{2}$ error and M_2 simultaneously incurs 0 errors. As is seen below, assuming equal model priors, M_1 is more probable than is M_2 by a factor of greater than 16, despite M_2 being conditioned on the better XER score of no LOOVC errors, i.e,

$$\begin{aligned} & Pr\{M_1 \; truth \mid d \; such \; that \; XER(M_1,d) = \frac{1}{2} \; and \; XER(M_2,d) = 0 \; \} \\ & \quad > \quad Pr\{M_2 \; truth \mid d \; such \; that \; XER(M_1,d) = \frac{1}{2} \; and \; XER(M_2,d) = 0 \; \} \end{aligned}$$

In the following, the same model parameters as above are used, and again 1 billion generations of data d_i of 10 observations each is used to estimate the probabilities.

$Pr\{XER(M_1,d) = \frac{1}{2} \mid M_1 \; truth\} \; / \; Pr\{XER(M_1,d) = \frac{1}{2} \mid M_2 \; truth \; and \; XER(M_2,d) = 0\} \approx 9.418634$
$Pr\{XER(M_2,d) = 0 \mid M_2 \; truth\} \; / \; Pr\{XER(M_2,d) = 0 \mid M_1 \; truth \; and \; XER(M_1,d) = \frac{1}{2}\} \approx 0.606764$

[3] Note how the interaction of model scores results in the ratio

$Pr\{XER(M_2,d) = \frac{1}{2} \mid M_2 \; truth\} \; / \; Pr\{XER(M_2,d) = \frac{1}{2} \mid M_1 \; truth \; and \; XER(M_1,d) = \frac{1}{2}\}$

being less than 1.0. On data d for which M_1 scores $\frac{1}{2}$ error, it is relatively likely that the class counts are skewed, for example, M_1's parent cells yield a class-split pattern such as $ps_1 = (2 \; Class_1, \; 1 \; Class_2)$, $ps_2 = (7 \; Class_1, \; 0 \; Class_2)$, which in turn implies that on such d random M_2 incurs $\frac{1}{2}$ error with relatively high probability.

4.3 The BD Metric

In Section 3.2, we considered model posterior and likelihood as evaluation measures under a scenario in which each nonzero $P(M)$ model's network structure G had associated with it a single, *known* distribution $G(\theta)$. Recall how the computation of the likelihood of model $M =< G, G(\theta) >$ depends on knowledge of this $G(\theta)$:

$$L(M,d) \;=\; \prod_i^N Pr\{x_i \mid M\} \;=\; \prod_i^N (Pr\{C(x_i) \mid ps(x_i)\}),$$

where the conditional class probabilities $Pr\{C(x_i) \mid ps(x_i)\}$ are assigned by $G(\theta)$. (As in (3), we continue to omit from the likelihood expression the constant term $\frac{1}{2^{|U|}}$.)

The device of assuming that the associated $G(\theta)$ is fixed and known to the evaluation function, while allowing for a simple demonstration of how likelihood and model posterior can be computed and how they behave as model evaluators, does not reflect typical situations of interest. Typically, one does not know with any certainty what distribution θ might be associated with each network structure G, and quantifies this uncertainty by specifying to the evaluation function a prior distribution $g(\theta|G)$ (a distribution over distributions, or a so-called hyper-distribution) over the space Θ of possible θ. Under this modeling, each network structure G represents a generally infinite family of models $< G, \{\theta\} >$. The prior distribution $g(\theta|G)$ then combines with the data d to yield the posterior of a model structure G (without reference to a specific θ), obtained by integrating over the space Θ of the possible distributions θ. Since we continue to assume uniform priors $P(G)$ over model structures, it suffices to consider data likelihood

$$Pr\{d \mid G, g\} = \int Pr\{d \mid \theta\} \times g(\theta \mid G) d\theta$$

of model structure G, which is proportional to the posterior of the model structure. Notice that such a likelihood is the probability of data d, given the network structure G and prior $g(\theta|G)$, that is, this is the probability of data given a family $< G, \{\theta\} >$ of models and the distribution $g(\theta|G)$, rather than of a particular $M =< G, G(\theta) >$. Operationally, the likelihood evaluation of d given G is under the premise that if G is truth, a θ is associated with G according to the distribution $g(\theta|G)$, and it is this $G(\theta)$ that generates the observed d.

In this section, we consider the question, if models G are evaluated by such scoring functions, for what θ that might in actuality be associated with G, is the evaluation coherent, in particular, complexity coherent? Notationally, since the model evaluations are applied to a network structure G which represents a *family* $< G, \{\theta\} >$ of models, rather than to a single model $M =< G, \theta >$, G will be the argument to the evaluation measures, and we will speak of the complexity k of a model structure G_k, still defined in terms the number of parents of the class node.

In much of the Bayesian learning literature, the prior $g(\theta|G)$ takes the form of a Dirichlet distribution, which is a conjugate family of distributions for multinomial sampling, the latter being the distribution that governs the observed data. Heckerman [10] demonstrates also that a Dirichlet prior distribution is implied by a set of common assumptions (which includes parameter independence and likelihood equivalence). That the Dirichlet distribution is a conjugate family of distributions for multinomial sampling makes tractable the computation of data likelihood (and hence model structure posterior) whenever the prior $g(\theta|G)$ is Dirichlet. The resulting data likelihood is what is known as the *BD (Bayesian-Dirichlet) metric*[4]:

$$\begin{aligned} & Pr\{d \mid G,\, g\} \\ & \quad = \quad \int Pr\{d \mid \theta\} \times g(\theta \mid G) d\theta \end{aligned}$$

[4]Technically, the *BD* metric is more commonly defined in terms of the joint probability $Pr\{d,\, G\}$, which is simply the above expression multiplied by the network prior $P(G)$,

$$= \prod_n \prod_p \frac{\Gamma(\alpha_p)}{\Gamma(\alpha_p + N_p)} \prod_v \frac{\Gamma(\alpha_{pv} + N_{pv})}{\Gamma(\alpha_{pv})}, \tag{4}$$

where

Γ is the Gamma function;

n ranges over the nodes in G;

p ranges over values $< p >$ of the parent set of the node n fixed by the outermost $\prod$;

v ranges over the values of the node n fixed by the outermost $\prod$;

N_p is the number of observations in d falling to parent cell $< p >$ and N_{pv} is the number of observations in d falling to parent cell $< p >$ and having node n value v; and

α_p and α_{pv} are parameters of the Dirichlet prior distribution as is described in the following subsection.

Since we continue to assume that the networks we shall evaluate contain edges only from the members of the parent set to the class node, the *BD* value computed at each (non-class) feature node is the same for a fixed d, regardless of the particular structure G being evaluated. Hence, as we did for ordinary likelihood in Section 3.2, in the following we restrict the *BD* score (which is node-wise decomposable) of a network G to the score on the class node. That is, in the outer product of (4) we hold n fixed at the class node and obtain

$$BD(G,d) = \prod_p \frac{\Gamma(\alpha_p)}{\Gamma(\alpha_p + N_p)} \prod_v \frac{\Gamma(\alpha_{pv} + N_{pv})}{\Gamma(\alpha_{pv})}. \tag{5}$$

In fact, even when edges may exist between the features, [11] demonstrates that the classification power of a parent set network can be evaluated by restricting *BD* to the class node, obtaining, in the context of the parent set model, equivalence with conditional class likelihood [8].

As is standard, we write $BD(G,d)$ without explicit reference to the parameters α_p and α_{pv} of the Dirichlet prior $g(\theta|G)$. In the following, when we write the event $[G,\ g\ truth]$, we mean model structure G and Dirichlet prior $g(\theta|G)$ (with understood parameters α_p and α_{pv}) govern the generation of the data d. This expression (5) therefore is the likelihood of data d assuming a model structure G and its specific Dirichlet prior $g(\theta|G)$. Since *BD* is a likelihood, it is a coherent measure, as was indicated for likelihood in Section 3.2. In particular, when truth is a family consisting of a network structure G and an associated distribution $g(\theta|G)$ over Θ, coherence of the *BD* measure means that for all data d, model structure pairs G and G', and scores $v < v'$,

$$\frac{Pr\{d\ such\ that\ BD(G,d) = v \mid G,\ g\ truth\}}{Pr\{d\ such\ that\ BD(G,d) = v \mid G',\ g\ truth\ and\ BD(G',d) = v'\}}$$

$$< \frac{Pr\{d\ such\ that\ BD(G',d) = v' \mid G',\ g\ truth\}}{Pr\{d\ such\ that\ BD(G',d) = v' \mid G,\ g\ truth\ and\ BD(G,d) = v\}}$$

Assuming equal model structure priors $P(G)$, the equivalent consequence in terms of model posteriors is

$$Pr\{G\ truth,\ g\ truth \mid d\ such\ that\ BD(G,d) = v\ and\ BD(G',d) = v'\}$$
$$< Pr\{G'\ truth,\ g\ truth \mid d\ such\ that\ BD(G,d) = v\ and\ BD(G',d) = v'\}$$

for all data d, model structure pairs G and G', and scores $v < v'$.

In fact, the special properties of likelihood imply that this relationship holds for any single fixed d, not just for aggregations of d achieving the given BD scores. That is

$$\begin{aligned} & Pr\{G\ truth,\ g\ truth \mid d\ \} \\ < \quad & Pr\{G'\ truth,\ g\ truth \mid d\ \} \end{aligned}$$

for any d such that $BD(G,d) < BD(G',d)$

As was noted in the observation following Fact 3.1 in Section 3.2, these relationships are established from specific properties of likelihood, and do not depend on our assumptions, such as disjointness of model features. Indeed, in cases where the Dirichlet prior correctly captures how distributions θ are associated with true model structure G, the disjointness assumption on model features fails to hold, since in such cases the same model structure (defined over a single set of features) has a nonzero prior in association with an infinite number of θ.

Of central importance to the current analysis is the observation that the above evaluation ratio and posterior coherence translate into a lack of model evaluation bias in BD only if the Dirichlet prior $g(\theta|G)$ actually does reflect how d is generated when G is the true model structure. At best, a Dirichlet prior typically is employed as a surrogate, a device that allows reasonable prior assumptions to be captured while providing a closed form for the measure. A pertinent question to ask is, under what conditions, if any, is BD non-coherent when, in actuality, some specific (though unknown to the BD evaluation) θ is deterministically associated with each G? We are particularly interested in this behavior with respect to a complexity bias: for what homomorphic θ_k and $\theta_{k'}$ associated with model structures $G_k \in card_k$ and $G_{k'} \in card_{k'}$ does the BD score exhibit a complexity non-coherence? That is, when, if ever, do we have misordered evaluation ratios

$$\begin{aligned} & \frac{Pr\{d\ such\ that\ BD(G_k,d) = v \mid G_k,\ \theta_k\ truth\}}{Pr\{d\ such\ that\ BD(G_k,d) = v \mid G_{k'},\ \theta_{k'}\ truth\ and\ BD(G_{k'},d) = v'\}} \\ < \quad & \frac{Pr\{d\ such\ that\ BD(G_{k'},d) = v' \mid G_{k'},\ \theta_{k'}\ truth\}}{Pr\{d\ such\ that\ BD(G_{k'},d) = v' \mid G_k,\ \theta_k\ truth\ and\ BD(G_k,d) = v\}} \end{aligned}$$

and hence misordered posteriors

$$\begin{aligned} & Pr\{G_k,\ \theta_k\ truth \mid d\ such\ that\ BD(G_k,d) = v\ and\ BD(G_{k'},d) = v'\} \\ < \quad & Pr\{G_{k'},\ \theta_k\ truth \mid d\ such\ that\ BD(G_k,d) = v\ and\ BD(G_{k'},d) = v'\} \end{aligned}$$

for homomorphic θ_k and $\theta_{k'}$ and $v \leq v'$? Note well that in this formulation, BD is evaluated with respect to a Dirichlet prior, while G actually is always associated with some single $G(\theta)$ when G is truth.

4.3.1 Properties of the Dirichlet Distribution

The Dirichlet is a family of distributions parameterized by the values α_p and α_{pv} (where p and pv are as defined following (4) of the previous section), and we must identify which members our analysis will consider. The literature contains extensive study of attempts to specify noninformative and/or uniform priors, a goal that is known to be fraught with pitfalls [12, 27]. Often, one attempts to model an uninformative prior by specifying a Dirichlet with a uniform allocation of vanishing equivalent sample size α. In particular, Buntine [4] proposed modeling an noninformative Dirichlet prior by specifying $\alpha_{pv} = \frac{\alpha}{(number\ of\ parent\ cells) * (number\ of\ node\ values)}$, with $\alpha_p = \sum_v \alpha_{pv}$. Under our assumption of binary-valued nodes, this translates for $G_k \in card_k$ to $\alpha_{pv} = \frac{\alpha}{2^{k+1}}$ and $\alpha_p = \frac{\alpha}{2^k}$. Heckerman *et. al.* [10] notes that this is an instance of the BD_e metric which he terms BD_{ue}, for *uniform joint distribution and likelihood equivalent*. By specifying a vanishing equivalent sample size α, one attempts to

realize an uninformative (as well as uniform) prior, though as we shall see, unintended consequences for model evaluation arise.

Several interesting behaviors of the Dirichlet prior were observed previously by Steck [26], but in somewhat different contexts leading to different interpretations than here. An analysis reveals that the Dirichlet with vanishing α places virtually all density on θ which assign conditional class probabilities $Pr\{C = 1 \mid ps\}$ and $Pr\{C = 1 \mid ps\}$ near the extreme values of 0.0 and 1.0, while such a Dirichlet achieves it minima at the θ which assigns conditional class probabilities at the expectations $E(Pr\{C = 1 \mid ps\}) = 0.5$, assuming α_{pv} is uniform as in BD_{ue}. Consequently, models that fit the data with only class-pure cells have posterior density which dwarf that of any other model, since class-pure cells would be generated by the θ that are overwhelming most likely to be drawn from such a Dirichlet. Thus, by making α vanishingly small, one does not diminish the impact of the Dirichlet prior on the evaluation, but rather amplifies it.

Such a Dirichlet prior has the effect of favoring models, random or true, with class-pure cells, and, as α vanishes, provided that there are many features available relative to the size of d, this effect dominates model selection. As with apparent training error, a single complex model, random or true, has more of a chance of achieving class-pure cells than does a single simple model, random or true. However, there is a competing second order effect under Dirichlet, and this is that the number y of parent cells populated influences the score as well. If two models both have only class pure cells, the model with the fewer number of populated parent cells will score higher. Here, simple models, random or true, have the advantage among models with class-pure cells, being more likely to populate few cells, because there are fewer cells to populate.

We now derive the possible *BD* scores (which, recall, is defined by (5) and is restricted to the class node) a model can attain, regardless of being random or true, and regardless of the $G(\theta)$ associated with model structure G.

Lemma 4.1: Let *BD* be computed under uniform α_{pv} with vanishing α and let d be any data with N observations. Then for any $G_k \in card_k$, the achievable *BD* scores $BD(G_k, d)$ approach 0 and $\frac{1}{2^y}$, y an integer in the range $1 \leq y \leq min(2^k, N)$, as α goes to 0. Further, the score $\frac{1}{2^y}$ is approached only when exactly y of G_k's 2^k parent cells are nonempty in d, and each of these y nonempty parent cells has instances from only one of the classes.

Proof: The effect of Dirichlet with uniform α_{pv} with vanishing α is that for each ps, in any θ with non-vanishing density, the probability of $Pr\{C = 1|ps\}$ approaches either 0 or 1, with equal probability. Consequently, for a d in which any of the parent cells is not class pure, the *BD* score approaches 0, since the likelihood of G generating any non-pure cell when it is associated with such a θ approaches 0. For a d in which y of G_k's parent cells are non-empty, and each of these parent cells is pure, $BD(G_k, d)$ approaches $\frac{1}{2^y}$, since the likelihood of G_k with this Dirichlet over its θ generating data with a nonzero number of cases in the specific class cell (C=1 or C=2) of each of the y ps that d fills approaches the probability of Dirichlet selecting for G_k the θ which assigns to each ps $Pr\{C = 1|ps\}$ approaching 0 or 1 in agreement with which of the ps class cells is nonempty in d. For each of the y nonempty ps, this probability is 0.5 and independent of the other ps's; hence, the joint for the y nonempty ps_i having the correct $Pr\{C = 1|ps_i\}$ or $Pr\{C = 2|ps_i\}$ assigned a value approaching 1 is $\frac{1}{2^y}$. It thus follows that when in d y parent states are nonempty, $BD(G_k, d)$ approaches either $\frac{1}{2^y}$ or 0.

□

When considering the question of whether a random model, simple or complex, has a better chance of scoring well, there are several factors that must be taken into account. Simpler random models score $\frac{1}{2^y}$ for small y with higher probability than complex random models, but complex random models have higher probability of fitting the data with class pure cells, and thus achieving nonzero scores. A comparison of the expected scores of random complex and random simple models depends on the number of observations N in the data d, on the score (0 or $\frac{1}{2^y}$, $1 \leq y \leq min(2^k, N)$) in question, and on the complexities k and k' of the models, with several crossover points characterizing this relationship.

4.3.2 The Lack of Coherence of *BD* with Vanishing α

While the analysis of a random model's *BD* score distributions is both interesting and intricate, our primary focus is the behavior of the evaluation ratio for *BD* scores, where the distribution of the *BD* scores of the true model, and the inter-dependence of the *BD* scores of the models evaluated, are vital. Though the analysis depends on what the true model is, we nevertheless are able to make some general conclusions regarding *BD*'s lack of coherence.

We again associate with model structure G the simple, but natural symmetric θ considered in Section 3.2 in the context of likelihood. Of course, here the θ are unknown to the *BD* evaluation, which will continue to assume a uniform Dirichlet (i.e., BD_{ue} with vanishing equivalent sample size α). The theorems which follow demonstrate that, for these natural θ, simple model structures are inappropriately favored over the complex. That is, we demonstrate that simple model structures scoring the same *BD* score as complex model structures on the same d have lower posterior probabilities, a result that holds even under uniform priors on model complexity, the uniform BD_{ue} Dirichlet priors over Θ, and symmetric and homomorphic actual θ. Thus, even in the simplest and most homogeneous of model spaces, *BD* does not exhibit complexity coherence. Since more varied families of θ would often contain as subfamilies these trivial θ, the non-coherence of *BD* demonstrated here applies widely.

Theorem 4.2: Let *BD* be evaluated with respect to a Dirichlet prior with uniform α_{pv} with vanishing equivalent sample size α. Let G_k have $q = 2^k$ cells, and $G_{k'}$ have $q' = 2^{k'}$ cells, $1 \leq k < k'$. Let associated θ_k and $\theta_{k'}$ be symmetric, assigning conditional class probabilities $pr = 1.0$, and hence the models are homomorphic. For any nonzero scores $v = \frac{1}{2^y}$, $1 < y \leq q$, and $v' = \frac{1}{2^{y'}}$, $1 < y' \leq \frac{q'}{2}$,

$$\frac{Pr\{d \text{ such that } BD(G_k,d) \approx v \mid G_k,\ \theta_k \text{ truth}\}}{Pr\{d \text{ such that } BD(G_k,d) \approx v \mid G_{k'},\ \theta_{k'} \text{ truth and } BD(G_{k'},d) \approx v'\}}$$

$$= \frac{1}{\left(\frac{(\frac{q'}{2}-1)(\frac{q'}{2}-2)\ldots(\frac{q'}{2}-(y'-1))}{(q')^{y'-1}}\right)} + \varepsilon.$$

for ε positive and approaching 0 as $|d|$ increases.

Proof: The numerator is:

$$\begin{aligned}
&Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_k,\ \theta_k \text{ truth}\} \\
&\quad = \quad (\ Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_k,\ \theta_k \text{ truth and has } y \text{ ps touched in } d\} \\
&\qquad * \quad Pr\{G_k \text{ has } y \text{ ps touched in } d \mid G_k,\ \theta_k \text{ truth}\}\ ) \\
&\quad = \quad (\ Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_k,\ \theta_k \text{ truth and has } y \text{ ps touched in } d\} \\
&\qquad * \quad Pr\{|d| = N \text{ cases touch } y \text{ of } q = 2^k \text{ cells ps given uniform placement}\}\ ) \\
&\quad = \quad (Pr\{\text{each of the } y \text{ nonempty ps in } G_k \text{ is class pure} \mid G_k,\ \theta_k \text{ truth and has } y \text{ ps touched in } d\} \\
&\qquad * \quad Pr\{|d| = N \text{ cases touch } y \text{ of } q = 2^k \text{ cells ps given uniform placement}\}\ )
\end{aligned}$$

The denominator is:

$$Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_{k'},\ \theta_{k'} \text{ truth and } BD(G_{k'},d) \approx 1/2^{y'}\}$$

$$= \quad (\ Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_{k'},\ \theta_{k'} \text{ truth},\ BD(G_{k'},d) \approx 1/2^{y'} \text{ and } G_k \text{ has } y \text{ ps touched in } d\}$$
$$* \quad Pr\{G_k \text{ has } y \text{ ps touched in } d \mid G_{k'},\ \theta_{k'} \text{ truth},\ BD(G_{k'},d) \approx 1/2^{y'}\}\ )$$

$$= \quad (\ Pr\{d \text{ such that } BD(G_k,d) \approx 1/2^y \mid G_{k'}\ \theta_{k'} \text{ truth},\ BD(G_{k'},d) \approx 1/2^{y'} \text{ and } G_k \text{ has } y \text{ ps touched in } d\}$$
$$* \quad Pr\{|d| = N \text{ cases touch } y \text{ of } q = 2^k \text{ cells ps given uniform placement}\}\ )$$

$$= \quad (Pr\{\text{each of the } y \text{ nonempty ps in } G_k \text{ is class pure} \mid G_{k'},\ \theta_{k'} \text{ truth}, BD(G_{k'},d) \approx 1/2^{y'} \text{ and } G_k \text{ has } y \text{ ps touched in } d$$
$$* \quad Pr\{|d| = N \text{ cases touch } y \text{ of } q = 2^k \text{ cells ps given uniform placement}\}\ )$$

After cancellation in the numerator and denominator of the common factor[5]

$$Pr\{|d| = N \text{ cases touch } y \text{ of } q = 2^k \text{ cells ps given uniform placement}\}$$

we are left with

$$\frac{Pr\{\text{each of the } y \text{ nonempty ps in } G_k \text{ is class pure} \mid G_k,\ \theta_k \text{ truth and has } y \text{ ps touched in } d\}}{Pr\{\text{each of the } y \text{ nonempty ps in } G_k \text{ is class pure} \mid G_{k'},\ \theta_{k'} \text{ truth}, BD(G_{k'},d) \approx 1/2^{y'} \text{ and } G_k \text{ has } y \text{ ps touched in } d\}}$$

The numerator is 1 since $G_k,\ \theta_k$ truth means each G_k cell is pure, since θ_k assigns $pr = 1.0$.

Consider now the denominator. $G_{k'}$ achieves its score only if d touches y' cells of $G_{k'}$. Each of these y' cells ps induces a class distribution of either $Pr\{C = 1|ps\} = 1$ or $Pr\{C = 2|ps\} = 1$, since $\theta_{k'}$ assigns $pr = 1.0$. If these y' cells mix the class assigned $pr = 1$, for sufficiently large d, the probability that $BD(G_k,d) \approx \frac{1}{2^y}$ is ε (for $\varepsilon > 0$, vanishingly small), since G_k is random and d contains cases with a mix of classes (at least approximately $\frac{1}{y'}$ fraction of the cases in d are of the minority class). In the case that the y' touched cells of $G_{k'}$ are either all $Pr\{C = 1|ps\} = 1$ or $Pr\{C = 2|ps\} = 1$, $BD(G_k,d) \approx \frac{1}{2^y}$ with probability 1, since in this case d contains cases of only one of the classes. The probability that all y' touched cells of $G_{k'}$ are either all $Pr\{C = 1|ps\} = 1$ or $Pr\{C = 2|ps\} = 1$ is given by

$$\frac{(\frac{q'}{2}-1)(\frac{q'}{2}-2)...(\frac{q'}{2}-(y'-1))}{(q')^{y'-1}}$$

since we are assuming $\theta_{k'}$ is symmetric and hence there are $\frac{q'}{2}$ of each type of parent cell. That is, after the first of the y' cells is touched, this expression is the probability that the remaining $y'-1$ touched cells of $G_{k'}$ come from the remaining $(\frac{q'}{2}-1)$ same class parent cells of $G_{k'}$ as the first.

□

When $y' = 2$ (the $G_{k'}$ model score approaches $v' = \frac{1}{2^2}$), the denominator of the ratio for G_k for any nonzero scores $v = \frac{1}{2^y}$, $1 < y \leq q$, that G_k approaches reduces to

$$\frac{(\frac{q'}{2}-1)}{q'} + \varepsilon$$

At $q' = 2$ (i.e., $k' = 1$) this is ε, and increases monotonically in q', bounded above by $\frac{1}{2}$. More generally, we have the following.

Lemma 4.2: For $k' \geq 1$ (and thus $q' \geq 2$) and score $v' = \frac{1}{2^{y'}}$ that $G_{k'}$ approaches for y' in the range $1 < y' \leq \frac{q'}{2}$, the denominator

$$\frac{(\frac{q'}{2}-1)(\frac{q'}{2}-2)...(\frac{q'}{2}-(y'-1))}{(q')^{y'-1}} + \varepsilon$$

[5] Feller [5], page 102 Eq (2.4), gives an expression for this occupancy problem, but apparently no tractable closed from is known.

of the G_k ratio for any nonzero score $v = \frac{1}{2^y}$ that G_k approaches is increasing in q', from ε at $q' = 2$, approaching $\frac{1}{2^{y'-1}}$ as q' (and thus k') increases.

Consequently, we have our main result regarding the non-coherence of *BD* scores. If on any d for which a pair of models differing only in complexity approach a common score of $v = \frac{1}{2^y}$, the more complex model has a higher ratio and, assuming uniform model structure priors $P(G)$, a higher true posterior conditioned on such a d.

Theorem 4.3: Let *BD* be evaluated with respect to a Dirichlet prior with uniform α_{pv} with vanishing equivalent sample size α. Let G_k have $q = 2^k$ cells, and $G_{k'}$ have $q' = 2^{k'}$ cells, $1 \leq k < k'$. Let the associated θ_k and $\theta_{k'}$ be symmetric, assigning conditional class probabilities $pr = 1.0$, and hence the models are homomorphic. For any nonzero score $v = \frac{1}{2^y}$, $1 < y \leq \frac{q}{2}$, and d sufficiently large,

$$\frac{Pr\{d \text{ such that } BD(G_k,d) \approx v \mid G_k,\ \theta_k \text{ truth}\}}{Pr\{d \text{ such that } BD(G_k,d) \approx v \mid G_{k'},\ \theta_{k'} \text{ truth and } BD(G_{k'},d) \approx v\}}$$
$$< \quad \frac{Pr\{d \text{ such that } BD(G_{k'},d) \approx v \mid G_{k'},\ \theta_{k'} \text{ truth}\}}{Pr\{d \text{ such that } BD(G_{k'},d) \approx v \mid G_k,\ \theta_k \text{ truth and } BD(G_k,d) \approx v\}}$$

Proof: The result follows immediately from Theorem 4.2 and Lemma 4.2. That is, taking $y = y'$ and thus $v = v' = \frac{1}{2^y} = \frac{1}{2^{y'}}$, Theorem 4.2 implies the left-hand-side ratios is

$$\frac{1}{\left(\frac{(\frac{q'}{2}-1)(\frac{q'}{2}-2)\ldots(\frac{q'}{2}-(y-1))}{(q')^{y-1}}\right)} + \varepsilon_1$$

while the right-hand-side ratio is

$$\frac{1}{\left(\frac{(\frac{q}{2}-1)(\frac{q}{2}-2)\ldots(\frac{q}{2}-(y-1))}{q^{y-1}}\right)} + \varepsilon_2$$

Since $q' > q$, Lemma 4.2 implies the left-hand-side denominator is greater than the right-hand-side denominator, and hence the left-hand-side ratio is smaller than the right-hand-side ratio.

□

Assuming equal model priors, we have also that

$$Pr\{G_k,\ \theta_k \text{ truth} \mid d \text{ such that } BD(G_k,d) \approx v \text{ and } BD(G_{k'},d) \approx v\}$$
$$< \quad Pr\{G_{k'},\ \theta_{k'} \text{ truth} \mid d \text{ such that } BD(G_k,d) \approx v \text{ and } BD(G_{k'},d) \approx v\}$$

As noted in Section 3.1, this pairwise lack of coherence implies that there is at least one d in the intersection of the sets of data on which G_k and $G_{k'}$ each scores v such that the posteriors are ordered as above, conditioned on this d. That is, from the pairwise inconsistency, we know there must exist at least one d such that the posteriors conditioned fully on this d is inconsistent with the *BD* scores on this d.

Example 4.3 The following numerical results show how pronounced the non-coherence can be, demonstrating that complex models ($G_5 \in card_5$) scoring the same good *BD* score ($v = \frac{1}{2^y}$, for small y) on a common d as simple models ($G_3 \in card_3$) have significantly higher ratios, and hence higher posteriors on such d. While Theorem 4.3 describes the behavior for d sufficiently large, the following results show a pronounced complexity bias even when d is size only 10 cases. We compute results for symmetric θ assigning $pr = 0.8$, as well as the $pr = 1.0$

case covered by the theorem. The probabilities below are estimated from generation of 100 million d for each of the two illustrations ($pr = 1.0$ and $pr = 0.8$), and we report ratios for the best scores $v = \frac{1}{2^y}$ (i.e., the lowest y) for which either there is at least one d_i generated such that

$$BD(G_3, d_i) \approx \tfrac{1}{2^y} \textit{ and } BD(G_5, d_i) \approx \tfrac{1}{2^y} \textit{ when } G_3 \textit{ truth}$$

or there is at least one d_j generated such that

$$BD(G_3, d_j) \approx \tfrac{1}{2^y} \textit{ and } BD(G_5, d_j) \approx \tfrac{1}{2^y} \textit{ when } G_5 \textit{ truth}$$

$$pr = 1.0$$

$$Pr\{BD(G_3,d) \approx 1/2^4 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^4 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^4\} \approx 6.159124$$
$$Pr\{BD(G_5,d) \approx 1/2^4 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^4 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^4\} \approx 45.166910$$
$$Pr\{BD(G_3,d) \approx 1/2^5 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^5 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^5\} \approx 11.181197$$
$$Pr\{BD(G_5,d) \approx 1/2^5 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^5 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^5\} \approx 27.578595$$
$$Pr\{BD(G_3,d) \approx 1/2^6 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^6 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^6\} \approx 10.036237$$
$$Pr\{BD(G_5,d) \approx 1/2^6 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^6 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^6\} \approx 17.942872$$
$$Pr\{BD(G_3,d) \approx 1/2^7 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^7 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^7\} \approx 6.793580$$
$$Pr\{BD(G_5,d) \approx 1/2^7 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^7 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^7\} \approx 9.601065$$

$$pr = 0.8$$

$$Pr\{BD(G_3,d) \approx 1/2^5 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^5 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^5\} \approx 1.893928$$
$$Pr\{BD(G_5,d) \approx 1/2^5 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^5 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^5\} \approx 3.248872$$
$$Pr\{BD(G_3,d) \approx 1/2^6 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^6 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^6\} \approx 2.339007$$
$$Pr\{BD(G_5,d) \approx 1/2^6 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^6 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^6\} \approx 3.255562$$
$$Pr\{BD(G_3,d) \approx 1/2^7 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^7 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^7\} \approx 2.142916$$
$$Pr\{BD(G_5,d) \approx 1/2^7 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^7 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^7\} \approx 2.626725$$
$$Pr\{BD(G_3,d) \approx 1/2^8 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^8 \mid G_5 \textit{ truth and } BD(G_5,d) \approx 1/2^8\} \approx 1.769336$$
$$Pr\{BD(G_5,d) \approx 1/2^8 \mid G_5 \textit{ truth}\} \;/\; Pr\{BD(G_5,d) \approx 1/2^8 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^8\} \approx 1.961687$$

While Theorem 4.3 and the numerical results above demonstrate a non-coherence for equal *BD* scores, Theorem 4.2 further implies that for symmetric θ assigning $pr = 1.0$ there exist $k < k'$ and $y < y'$ such that the G_k score approaches $\frac{1}{2^y}$ and the $G_{k'}$ score approaches $\frac{1}{2^{y'}}$ (i.e., G_k achieves a *BD* score strictly better than $G_{k'}$) on some d, yet $G_{k'}$ has a far larger ratio and hence a far larger posterior (assuming equal model priors) than does G_k. For example, comparing G_1 (and hence $q = 2$) with G_3 (and hence $q' = 8$) when the G_1 score approaches $\frac{1}{2^2}$ and the G_3 score approaches $\frac{1}{2^3}$, we compute from Theorem 4.2 that G_1's ratio is is approximately $\frac{1}{\frac{(3*2)}{64}}$ while G_3's ratio grows arbitrarily large with $|d|$.

Even for d of size only 10, we obtain from a generation of 100 million such d_i and a symmetric θ assigning $pr = 1.0$ the following results.

$$Pr\{BD(G_1,d) \approx 1/2^2 \mid G_1 \textit{ truth}\} \;/\; Pr\{BD(G_1,d) \approx 1/2^2 \mid G_3 \textit{ truth and } BD(G_3,d) \approx 1/2^3\} \approx 6.927846$$
$$Pr\{BD(G_3,d) \approx 1/2^3 \mid G_3 \textit{ truth}\} \;/\; Pr\{BD(G_3,d) \approx 1/2^3 \mid G_1 \textit{ truth and } BD(G_1,d) \approx 1/2^2\} \approx 170.369936$$

Consequently, assuming equal model structure priors, the model posteriors behave as

$$\begin{aligned} &Pr\{G_1,\ \theta_1 \textit{ truth} \mid d \textit{ such that } BD(G_1,d) \approx v \textit{ and } BD(G_3,d) \approx v'\} \\ &\quad < \quad Pr\{G_3,\ \theta_3 \textit{ truth} \mid d \textit{ such that } BD(G_1,d) \approx v \textit{ and } BD(G_3,d) \approx v'\} \end{aligned}$$

for this pair $v = \frac{1}{2^2}$, $v' = \frac{1}{2^3}$ of scores, with the posterior of the more complex G_3 exceeding that of G_1 by a factor of greater than 24, despite being conditioned on the inferior *BD* score.

4.4 MDL Measures

Model evaluation measures derived from the MDL principle [21] attempt to balance a model's "fit" to the data with the model's complexity. Since the measure of fit — denoted below as MDL_{data} — correlates closely with the apparent error rate measures analyzed in Sections 4.1 and 4.2, the complexity penalty terms of the MDL measure — denoted below as MDL_{graph} and MDL_{table} – are well motivated. As the following analysis demonstrates, however, the standard forms that this complexity penalty take are not sophisticated enough to ensure a coherent evaluation in all situations.

Several similar versions of MDL have been proposed in the context of Bayesian networks and Bayesian network classifiers. Since all share the same basic form, slight modifications to the following examples suffice to exhibit similar non-coherence for the most commonly employed variants. The general form of the MDL score applied to a Bayesian network $M =< G, \theta >$ is

$$MDL(M,d) = DL_{graph}(M) + DL_{table}(M,d) + DL_{data}(M,d)$$

We perform the evaluations below with the specific MDL realization used by Friedman and Goldszmidt [7]. Applied to our parent set network structure (where the only edges are from a subset of the features to the class node, and all features and the class node are binary), the three terms of the score are given by:

$$DL_{graph}(M) = log_2 F \ + \ log_2(\left(\begin{array}{c} F \\ k \end{array} \right))$$

$$DL_{table}(M,d) = \tfrac{1}{2} \ * \ 2^k \ * \ log_2 N$$

$$DL_{data}(M,d) = \ -\textstyle\sum_{pv} N_{pv} \ * \ log_2(\frac{N_{pv}}{N_p})$$

where F is the number of features (not including the class node, which is not available to be its own parent), k is the number of parents in M of the class node, $N = |d|$ is the number of observations in d, and N_p and N_{pv} are defined following (4) in the context of the *BD* metric. As we did for *BD* and simple likelihood, the above restricts the MDL score to the score on the class node, since, again, the score on feature nodes does not vary across different parent set networks. Note that $MDL(M,d)$ does not depend on $M(\theta)$, nor on a prior distribution over Θ.

The very similar MDL formulation of Lam and Bacchus [14] replaces the term DL_{graph} with $k \ * \ log_2(F)$. The details differ only slightly, and, since in both versions the DL_{graph} term grows when either F or k grows, the same non-coherence demonstrated in the following example is seen in either formulation.

Example 4.3: We specify problem parameters so that a complex model has a higher (worse) MDL score than a simpler model, yet the complex model has a higher evaluation ratio, and thus a higher posterior conditioned on these scores, assuming uniform model priors. Let there be $F = 1000$ features, and data d of size $N = 10$ observations. We compare homomorphic models of complexities $k = 1$ and $k' = 2$, with symmetric θ assigning $pr = 0.7$. We compute as follows MDL score components DL_{graph} and DL_{table}, which do not depend on the data (beyond its size N):

$M_1 \in card_1$: $DL_{graph} = 19.931569, \ \ DL_{table} = 3.321928, \ \ DL_{graph} + DL_{table} = 23.253497$

$M_2 \in card_2$: $DL_{graph} = 28.895909, \ \ DL_{table} = 6.643856, \ \ DL_{graph} + DL_{table} = 35.539765$

When a $card_2$ model M_2 fits the data perfectly (only class-pure parent cells), $DL_{data} = 0$ and hence $MDL(M_2,d) = 35.539765$. A $card_1$ model M_1 incurs a DL_{data} score of 9.709506 and hence $MDL(M_1,d) = 32.963003$ when it fits the data with an even class-split pattern, such as $ps_1 = (3\ Class_1, \ 2\ Class_2), \ ps_2 = (2\ Class_1, \ 3\ Class_2)$, in its two parent cells. Letting $v = 32.963003$ and $v' = 35.539765$, we compare the ratio on data d for which M_1 achieves v and M_2 simultaneously achieves v'. Despite the fact that M_1's score is better, probabilities estimated from 10 million generated d_i of 10 observations each show the ratio for M_2 to be 10 times higher than the ratio

for M_1. Specifically, the estimates produced from the 10 million d_i are:

$$\frac{Pr\{MDL(M_1,d) = v \mid M_1\ truth\}}{Pr\{MDL(M_1,d) = v \mid M_2\ truth,\ MDL(M_2,d) = v'\}} \approx \frac{0.04827380}{0.07393976} = 0.652880$$
$$\frac{Pr\{MDL(M_2,d) = v' \mid M_2\ truth\}}{Pr\{MDL(M_2,d) = v' \mid M_1\ truth,\ MDL(M_1,d) = v\}} \approx \frac{0.05013270}{0.007449175} = 6.729966$$

The values are stable over several sets of 10 million runs each. The implication is that, when these models score as specified (i.e. $MDL(M_1,d) < MDL(M_2,d)$, so M_1 scores better), M_2 is greater than 10 times more likely to be the true model, assuming equal model priors, i.e., for these $v < v'$ (M_1 achieves the better MDL score v)

$$Pr\{M_1\ truth \mid MDL(M_1,d) = v\ and\ MDL(M_2,d) = v'\}$$
$$< \quad Pr\{M_2\ truth \mid MDL(M_1,d) = v\ and\ MDL(M_2,d) = v'\}$$

□

One might consider whether a complexity coherent variant of MDL scoring could be devised. Note first that the DL_{data} term by itself exhibits a non-coherence favoring complex models that is quite similar to the bias seen in apparent error rate. For example, using the above parameters, except that the scores compared are for M_1 and M_2 each achieving $DL_{data} = 0$ (perfect fit to the data), we find that M_1 has a ratio greater than three times that of M_2 (254.926454 vs. 72.113131 where probabilities are again estimated from a generation of 10 million d, each with 10 observations). Thus, some term compensating for this complexity bias in DL_{data} is required if the measure is to be complexity coherent. Consider, then, the general form

$MDL_Y = pen(k) + DL_{data}$,

with $pen(k)$ increasing without bound in k, as does each of DL_{table} and DL_{graph}. Consider a generalization of the original example in which M_1 again fits the data with an even class-split pattern, yielding an MDL_Y score of $v = pen(1) + 9.709506$. Consider $M_{k'}$ which fit the data perfectly, for increasing k', yielding $DL_{data} = 0$ and thus an MDL_Y score of $v' = pen(k')$. Note that, for sufficiently large k', the term $pen(k')$ results in $v' > v$. However, the ratio for M_1 is approximately 0.652880 when $k' = 2$ and decreases as k' increases (this poor DL_{data} score is more likely when M_1 is random than truth), while for $M_{k'}$, as k' increases, the ratio for its perfect DL_{data} score approaches 1.0 from above. Consequently, no such *pen* term can result in a coherent measure. It therefore appears that DL_{data} would need to be combined with a rather sophisticated complexity adjustment term — one that depends on the models' DL_{data} scores as well as on the models' complexities — if a complexity coherent variant of MDL is to be obtained.

One conclusion for MDL in its standard form, and for BD_{ue} with vanishing equivalent sample size, is that if these measures work well in practice, it may be explained by the fact that often applications possess a bias for the true model to be simple.

5. Model Space Issues

5.1 Overview

The second dimension that must be accounted for in model evaluation is the model space. Even assuming that a coherent evaluation measure EV is employed, there remain important issues that depend on characteristics of the model space. Of interest here is how does the number of models in the space, or the number of models evaluated, affect the interpretation of a coherent scoring function.

For concreteness, we assume here that likelihood $L(M,d)$ is the evaluation function, under the scenario of Section 3.2 that $M(\theta)$ is known to L. Note that, by assuming a coherent scoring function, the structural characteristics of the models evaluated (e.g., the complexity of the models) become irrelevant to selection, beyond

what is captured in model priors $P(M)$. If $P(M)$ is uniform on the model space, then the number of models of each complexity class $card_k$ evaluated is irrelevant, and the total number of models in the space (or the number of models examined) is the only factor to consider. Even if there are far more $card_{k'}$ models than $card_k$ models (for example, when $k' > k$) evaluated, and even if the score of the best of the $card_{k'}$ models is better by only a minuscule amount than that of the best of the $card_k$ models, there is no reason to prefer the best $card_k$ model (i.e., overfit or complexity avoidance is not justified). The coherence of the evaluation function and uniformity of $P(M)$ imply this. In particular, the appropriateness of any complexity-related score adjustment stems from evaluation non-coherence rather than from properties of model space or of search.

Three results follow.

a) *A priori*, the true model has the highest distribution of likelihood scores, i.e., higher than any single random model, before data is observed.

b) The probability that we can identify correctly the true model M^* decreases as the number Q of models in the space increases relative to the amount of data. This is captured in the collective *a priori* distributions of likelihood scores for models in the space, and reflects on generalization error, though not on selection criteria. While this result seems self-evident, it is valuable to quantify the $|d|$ vs. Q tradeoff so as to realize when we are in a hopeless situation, that no amount of ingenuity, such as sophisticated search or re-sampling, can remedy.

c) In a space with Q models, if we evaluate $W \leq Q$ models, the probability of identifying the true model M^* monotonically increases as W increases. The strength of M^* — the certainty of the classification given the parent cell — determines the rate of increase. While this might not appear surprising, Quinlan and Cameron-Jones [20] and others have seemingly observed a contradictory non-nonmonotonic behavior referred to as oversearching. There to our knowledge has not been developed an abstract analysis of the oversearch phenomenon that is not obfuscated by details of a specific search strategy, or by potential shortcomings of specific evaluation measures.

5.1 Assumptions

We continue under the assumptions specified in Section 3.1 which are summarized here for convenience.

a) Feature sets are disjoint. We continue to consider model-space issues when model interaction is trivial, so as to factor out confounding effects in an attempt to gain insight into the intrinsic principles.

b) Every model M is associated with an $M(\theta)$ which assigns conditional class probabilities of pr and $1 - pr$ symmetrically to the two classes, so the unconditional class probabilities are equal. As observed, the assignment of a single, fixed $0.5 \leq pr \leq 1.0$ is sufficient to capture, for example, a functional parent cell - class relationship, with a noise process flipping the class label with a single, fixed probability.

c) Our focus continues to be on evaluation characteristics rather than on details of search. Thus, we continue to assume that a model of a specified cardinality is chosen by an oracle for evaluation with equal probability from among models with nonzero prior probability.

5.2 Distribution of Likelihood Scores

As is recapped in Fact 3.1 of Section 3.1, and Theorem 3.1 linking likelihood to posterior, given any d, the model with the highest likelihood on d is most likely to be truth, assuming uniform model priors. While this governs model posterior after d has been observed (i.e., posterior conditioned on d), Theorems 3.2 and 3.3 derive the *a priori* distribution of likelihood scores:

$$Pr\{d \text{ such that } L(M,d) \;=\; pr^H \;*\; (1-pr)^{(N-H)} \mid M \text{ true}\} \;=\; \left(\begin{array}{c} N \\ H \end{array}\right) \;*\; (pr^H * (1-pr)^{(N-H)})$$

$$Pr\{d \text{ such that } L(M,d) \;=\; pr^H \;*\; (1-pr)^{(N-H)} \mid M \text{ random}\} \;=\; \left(\begin{array}{c} N \\ H \end{array}\right) \;*\; 0.5^N$$

for $0 \leq H \leq N$, where $N = |d|$. Further, Theorem 3.4 establishes that (when θ is symmetric, as is assumed here) knowledge of the true model's identity, or of any model's likelihood score $L(M,d)$, does not affect the distribution of another model's score $L(M',d)$.

It follows from the *a priori* score distributions for random and true models that true model M^* has its scores distributed strictly higher than any other single model in model space, provided only that the *pr* assigned by $M^*(\theta)$ is other than 0.5.

Theorem 5.1: The cumulative distribution function (CDF) for the true model's likelihood scores, except for equality at the upper extreme score, is everywhere below the CDF for any random model's scores, provided that the true model's $pr \neq 0.5$. That is,

$$\begin{aligned} & Pr\{d \text{ such that } L(M^*,d) \;\leq\; pr^H \;*\; (1-pr)^{(N-H)} \mid M^* \text{ true}\} \\ < \quad & Pr\{d \text{ such that } L(M,d) \;\leq\; pr^H \;*\; (1-pr)^{(N-H)} \mid M \text{ random}\} \end{aligned}$$

for all $0 \leq H < N$, with equality at $H = N$.

Proof: Since $pr \neq 0.5$, $pr > 0.5$ since, by our convention, $pr \geq (1-pr)$. Consider first the relationship between

$pr^H * (1-pr)^{N-H}$ and 0.5^N

when $pr > 0.5$. At $H = 0$ we have

$(1-pr)^N < (0.5)^N$

while at $H = N$ we have

$pr^N > 0.5^N$

Since $pr^H * (1-pr)^{N-H}$ is increasing in H, it follows that there exists some T, $0 < T \leq N$, such that for all $0 \leq q < T$

$pr^q * (1-pr)^{(N-q)} \leq (0.5)^N$

and for all $T \leq k \leq N$,

$pr^k * (1-pr)^{(N-k)} > (0.5)^N$

It follows from Theorems 3.2 and 3.3 that

$$Pr\{d \text{ such that } L(M^*,d) \;\leq\; pr^H \;*\; (1-pr)^{(N-H)} \mid M^* \text{ true}\} \;=\; \sum_{k=0}^{H} \left(\begin{array}{c} N \\ k \end{array}\right) \;*\; (pr^k * (1-pr)^{(N-k)})$$

$$Pr\{d \text{ such that } L(M,d) \;\leq\; pr^H \;*\; (1-pr)^{(N-H)} \mid M \text{ random}\} \;=\; \sum_{k=0}^{H} \left(\begin{array}{c} N \\ k \end{array}\right) \;*\; 0.5^N$$

If $0 \leq H < T$ (where T is as identified above), then every term of the true model's summation is less than or equal to the corresponding term of the random model's summation, and at the first term ($k = 0$) of the summations, the random model's term is strictly larger. If $T \leq H < N$, re-write the expansion of the two CDF's as

$$1.0 \;-\; \sum_{k=H+1}^{N} \left(\begin{array}{c} N \\ k \end{array}\right) \;*\; (pr^k * (1-pr)^{(N-k)}) \text{ and}$$

$$1.0 \;-\; \sum_{k=H+1}^{N} \left(\begin{array}{c} N \\ k \end{array}\right) \;*\; 0.5^N$$

Each of these terms in the summation for the true model is larger than the corresponding term in the summation for the random model, hence establishing the relationship between the CDF's for all $0 \leq H < N$. At $H = N$, both CDF's evaluate to 1.0.

□

Observe that differential in the CDF's at any point $0 \leq H < N$ increases as pr approaches 1.0, and is independent of model complexity.

5.3 How Probable is the Selection of the True Model?

This section considers how probable are we to select the true model if, after observing the data d, we select the model with the highest posterior conditioned on d, which, of course, is the criterion that makes the selection of the true model most probable after observing the data d. That is, we consider how likely is it, before observing the data d, that the true model M^* (which is to generate the data d) will have the highest posterior conditioned on that data d and thus be correctly selected as being the true model. Since model priors are assumed equal, we analyze this question in terms of likelihood scores $L(M,d)$.

We first assume that the true model is among the W that we evaluate, which would be the case, for example, if we evaluated all the models in the model space. The next section utilizes this result to address the question of how many models should we evaluate.

Suppose W models, the true model M^* plus $W-1$ random models, are evaluated. We select as our guess for truth the highest scoring model. If h models, $1 \leq h \leq W$, tie for the highest evaluation, assume each of these h models has an equal $\frac{1}{h}$ chance of being selected by some tie breaking procedure — by our previous analysis, these h models are equi-probable and cannot be distinguished. We wish to compute the *a priori* probability (before data is observed) of this procedure resulting in the selection of the true model M^*. That is, we wish to compute

$$
\begin{aligned}
& S(W) \\
&= Pr\{M^* \ selected \ from \ the \ highest \ scoring \ models \ evaluated \mid W-1 \ random \ models \ and \ M^* \ are \ evaluated\} \\
&= \sum_{r=0}^{W-1} \frac{Pr\{M^* \ and \ r \ of \ W-1 \ random \ models \ score \ highest\}}{(r+1)}
\end{aligned}
$$

Theorem 5.2:

$$
\begin{aligned}
& S(W) \\
&= Pr\{M^* \ selected \ from \ the \ highest \ scoring \ models \ evaluated \mid W-1 \ random \ models \ and \ M^* \ are \ evaluated\} \\
&= \frac{1}{W} * \sum_{k=0}^{N} \frac{Pr\{M^* \ scores \ v_k\}}{a_k} * ((b_{k+1})^W - (b_k)^W)
\end{aligned}
$$

where:

$v_k = pr^k * (1-pr)^{N-k}$

$a_k = Pr\{random \ model \ M \ scores \ v_k\}$

$b_k = Pr\{random \ model \ M \ scores \ less \ than \ v_k\}$

Proof: Recall that θ symmetric implies model scores are independent (Theorem 3.4). N is the number of cases in d, and represents the exponent of the highest possible likelihood score, i.e., $0 \leq k \leq N$, again assuming $pr > 0.5$. Thus

$$
\begin{aligned}
& S(W) \\
&= \textstyle\sum_{k=0}^{N} [P\{M^* \ scores \ v_k\} \\
&\quad * \textstyle\sum_{r=0}^{W-1} (\frac{1}{r+1} * Pr\{some \ r \ of \ W-1 \ random \ models \ score \ v_k \ and \ none \ of \\
&\qquad\qquad the \ remaing \ (W-1-r) \ random \ models \ score \ higher \ than \ v_{k-1}\})
\end{aligned}
$$

]

Note $b_0 = 0$ and $b_{N+1} = 1$. Hence

$$\begin{aligned} & S(W) \\ &= \sum_{k=0}^{n} (Pr\{M^* \text{ scores } v_k\} * \sum_{r=0}^{W-1} \frac{\left(\begin{array}{c} W-1 \\ r \end{array} \right) * [{a_k}^r * {b_k}^{(W-1-r)}]}{r+1}) \end{aligned}$$

Since $\frac{\left(\begin{array}{c} W-1 \\ r \end{array} \right)}{r+1} = \frac{1}{W} * \left(\begin{array}{c} W \\ r+1 \end{array} \right)$, we can write

$$\begin{aligned} & \sum_{r=0}^{W-1} \frac{\left(\begin{array}{c} W-1 \\ r \end{array} \right) * [{a_k}^r * {b_k}^{(W-1-r)}]}{r+1} \\ &= \frac{1}{W} * \sum_{r=0}^{W-1} (\left(\begin{array}{c} W \\ r+1 \end{array} \right) * [{a_k}^r * {b_k}^{(W-1-r)}]) \end{aligned}$$

Changing the limits of summation we obtain

$$\begin{aligned} & \frac{1}{W} * \sum_{r=0}^{W-1} (\left(\begin{array}{c} W \\ r+1 \end{array} \right) * [{a_k}^r * {b_k}^{(W-1-r)}]) \\ &= \frac{1}{W} * (\frac{1}{a_k} * \sum_{r=0}^{W} (\left(\begin{array}{c} W \\ r \end{array} \right) * [{a_k}^r * {b_k}^{(W-r)}]) - \frac{1}{a_k} * (\left(\begin{array}{c} W \\ 0 \end{array} \right) * [{a_k}^0 * {b_k}^W])) \\ &= \frac{1}{W * a_k} * [(a_k + b_k)^W - {b_k}^W], \text{ by the Binomial Theorem.} \end{aligned}$$

From the definitions of a_k and b_k, we have for $0 \leq k \leq N$

$$\begin{aligned} & (a_k + b_k) \\ &= Pr\{\text{random model } M \text{ scores } v_k\} + Pr\{\text{random model } M \text{ scores less than } v_k\} \\ &= Pr\{\text{random model } M \text{ scores } v_k \text{ or less}\} \\ &= b_{k+1} \end{aligned}$$

Hence, we may write

$$\begin{aligned} & \frac{1}{W * a_k} * [(a_k + b_k)^W - {b_k}^W] \\ &= \frac{1}{W * a_k} * [{b_{k+1}}^W - {b_k}^W] \end{aligned}$$

and thus

$$S(W)$$

$$= \frac{1}{W} * \sum_{k=0}^{N} \frac{Pr\{M^* \text{ scores } v_k\}}{a_k} * ((b_{k+1})^W - (b_k)^W)$$

□

Note that with all parameters but W held fixed, the probability $S(W)$ is decreasing as W increases, since M^* is assumed to be among the W models evaluated, regardless of how large this number W is.

We first apply the expression for $S(W)$ to the situation where $Q = W$, meaning that we evaluate all models in the space, and we wish to analyze, for various values of the parameters $|d|$ and pr, how the probability of selecting M^* decreases as the size $W = Q$ of the model space increases. Figures 1 (a)-(d) plot $log_{10}(W)$ vs. $log_{10}(S(W))$, where, recall,

$$S(W) = Pr\{M^* \text{ selected from the highest scoring models evaluated} \mid W-1 \text{ random models and } M^* \text{ are evaluated}\}$$

The four plots are for true models M^* with pr=0.6, 0.7, 0.8, and 0.9, respectively. Each plot displays a curve for each of the three sizes of d, N=20, 60, and 100 observations. For example, we see from Figure 1(b) that when true model M^* assigns $pr = 0.7$ and the model space contains $Q = 10^6$ models, and when data is to contain $N = 60$ observations, the *a priori* probability that M^* will be correctly identified then is $0.0387 = log_{10}(-1.412362)$, assuming all models are to be evaluated. Note that since the number of models in the space often is exponential in the number of features, in many high dimensional applications, the data relative to the size of the model space generally is far sparser than what the plotted values specify. These plots thus indicate the upper limits of when there is reasonable probability of correctly identifying the true model M^*.

5.4 How Many Models Should We Evaluate?

When we evaluate W out of Q models, the probability that M^* is among the W evaluated is $\frac{W}{Q}$, assuming models are chosen uniformly for evaluation. While we can select M^* as the true model only if it is among the W we evaluate, the probability, given that M^* is among the W we evaluate, that M^* also is among the r highest models evaluated and then is selected rather than one of the $r-1$ random models ($0 \leq r \leq W$) with the same maximal score decreases as W increases. How do these competing forces trade-off against each other?

Let $PSelect(W) = Pr\{M^* \text{ is selected from among the } W \text{ we evaluate}\}$. Then

$PSelect(W)$

$= Pr\{M^* \text{ among the } W \text{ we evaluate}\}$
$\quad * Pr\{M^* \text{ selected from the highest scoring models evaluated} \mid M^* \text{ among the } W \text{ we evaluate}\}$

$= Pr\{M^* \text{ among the } W \text{ we evaluate}\}$
$\quad * Pr\{M^* \text{ selected from the highest scoring models evaluated} \mid W-1 \text{ random models and } M^* \text{ are evaluated}\}$

$= \frac{W}{Q} * S(W)$,

where $S(W)$ is as defined and derived in the previous section.

Theorem 5.3: If all θ are symmetric and assign $pr \neq 0.5$, then
$PSelect(W) = \frac{W}{Q} * S(W)$ is strictly increasing in W, that is, the more models evaluated the higher the probability that the true model M^* is selected.

Proof: Since $pr \neq 0.5$, $pr > 0.5$ since, by our convention, $pr \geq (1-pr)$. For $W > 0$, let $Incl(W) = \frac{W}{Q}$ and thus $Pselect(W) = Incl(W) * S(W)$. The strategy is to show

$$\frac{PSelect(W)}{PSelect(W-1)} = \frac{Incl(W) * S(W)}{Incl(W-1) * S(W-1)} > 1, \text{ for } W > 1.$$

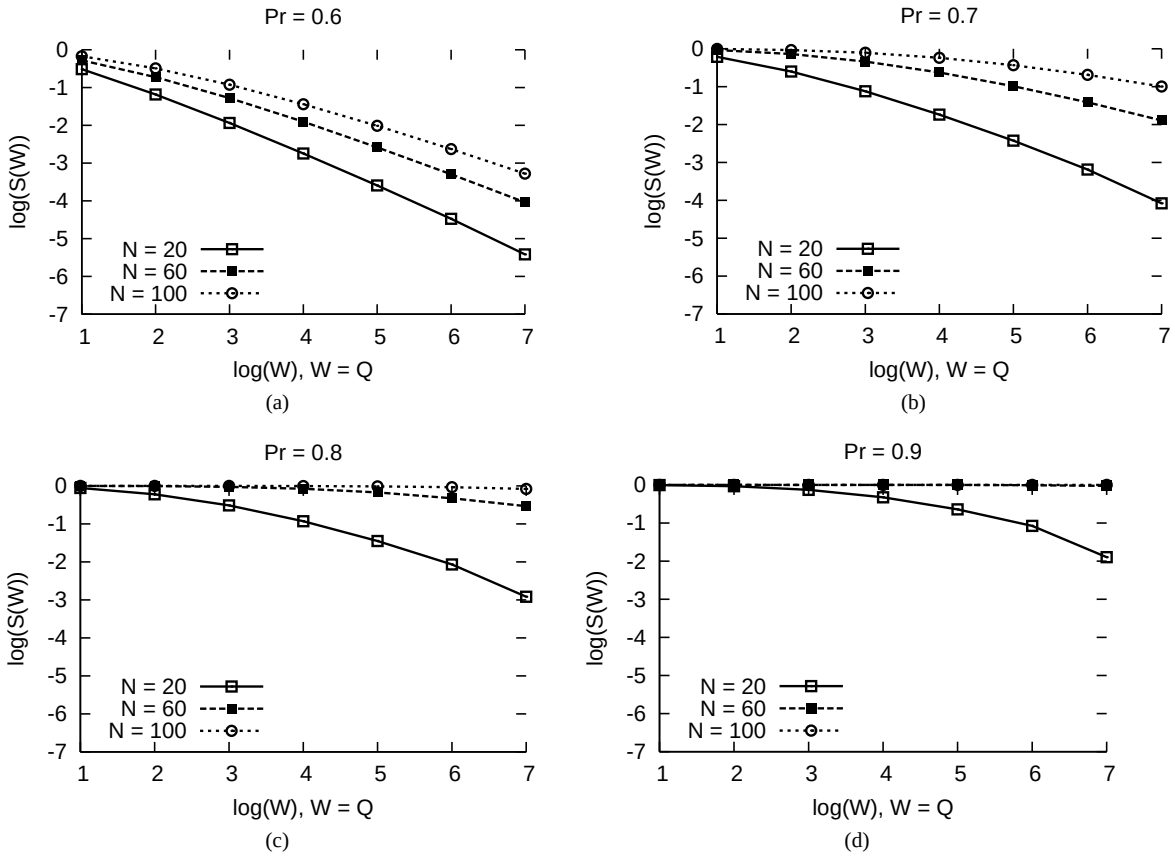

Figure 1: Effect of number of models Q in space on probability $S(W)$ of selecting M^*, assuming all models are evaluated (i.e., Q=W), for various numbers N of observations and model strengths pr. Plots are log_{10} vs. log_{10} .

$\frac{Incl(W)}{Incl(W-1)} = \frac{\frac{W}{Q}}{\frac{W-1}{Q}} = \frac{W}{W-1}$ and, from Theorem 5.2, the first factor of
$\frac{S(W)}{S(W-1)}$ is $\frac{\frac{1}{W}}{\frac{1}{W-1}} = \frac{W-1}{W}$, so the result follows iff

$$\frac{\sum_{k=0}^{N}(\frac{Pr\{M^* \text{ scores } v_k\}}{a_k} * [(b_{k+1})^W - (b_k)^W])}{\sum_{k=0}^{N}(\frac{Pr\{M^* \text{ scores } v_k\}}{a_k} * [(b_{k+1})^{W-1} - (b_k)^{W-1}])} > 1.0, \ for\ W > 1.$$

It follows from Theorems 3.2 and 3.3 that, for each k,

$$\frac{Pr\{M^* \text{ scores } v_k\}}{a_k} = \frac{\left(\begin{array}{c} N \\ k \end{array} \right) * (pr^k * (1-pr)^{N-k})}{\left(\begin{array}{c} N \\ k \end{array} \right) * (0.5)^N}$$

Hence, after cancellation of the $\left(\begin{array}{c} N \\ k \end{array} \right)$ terms and the constant $(0.5)^N$, the result follows iff

$$\frac{\sum_{k=0}^{N}(pr^k * (1-pr)^{N-k}) * [(b_{k+1})^W - (b_k)^W]}{\sum_{k=0}^{N}(pr^k * (1-pr)^{N-k}) * [(b_{k+1})^{W-1} - (b_k)^{W-1}]} > 1.0, \ for\ W > 1.$$

Consider the sum for a given exponent m:

$$\sum_{k=0}^{N}(pr^k * (1-pr)^{N-k}) * [(b_{k+1})^m - (b_k)^m])$$

Association of the terms yields the telescoping sum

$$\begin{array}{l} -(pr^0 * (1-pr)^N * {b_0}^m) \\ +(\sum_{k=1}^{N}(pr^{(k-1)} * (1-pr)^{N-(k-1)} \ - \ pr^k * (1-pr)^{(N-k)}) * {b_k}^m) \\ +(pr^N * (1-pr)^0 * {b_{N+1}}^m) \end{array}$$

$b_0 = 0$, so the first term $-(pr^0 * (1-pr)^N * {b_0}^m)$ is 0.
$b_{N+1} = 1$, so the final term $(pr^N * (1-pr)^0 * {b_{N+1}}^m) = pr^N$ and does not depend on the exponent m.
Each of the middle terms

$(pr^{(k-1)} * (1-pr)^{N-(k-1)} \ - \ pr^k * (1-pr)^{(N-k)}) * {b_k}^m$

is less than 0 provided $pr > 0.5$ (and is 0 at $pr = 0.5$). Since $b_k < 1$ ($k \leq N$), increasing the exponent m diminishes ${b_k}^m$ and hence increases the total sum. Therefore,

$$\frac{\sum_{k=0}^{N}(pr^k * (1-pr)^{N-k}) * [(b_{k+1})^W - (b_k)^W]}{\sum_{k=0}^{N}(pr^k * (1-pr)^{N-k}) * [(b_{k+1})^{W-1} - (b_k)^{W-1}]} > 1, \ for\ W > 1.$$

establishing the theorem.

□

The Quinlan Cameron-Jones [20] oversearch result thus cannot materialize in this scenario of a coherent measure. We conjecture that the monotonicity result continues to hold in many other scenarios, provided that a coherent evaluation is used. Note that Theorem 5.3 implies also that the number of models of each complexity that are evaluated is irrelevant, since a coherent evaluation measure is unaffected by model complexity.

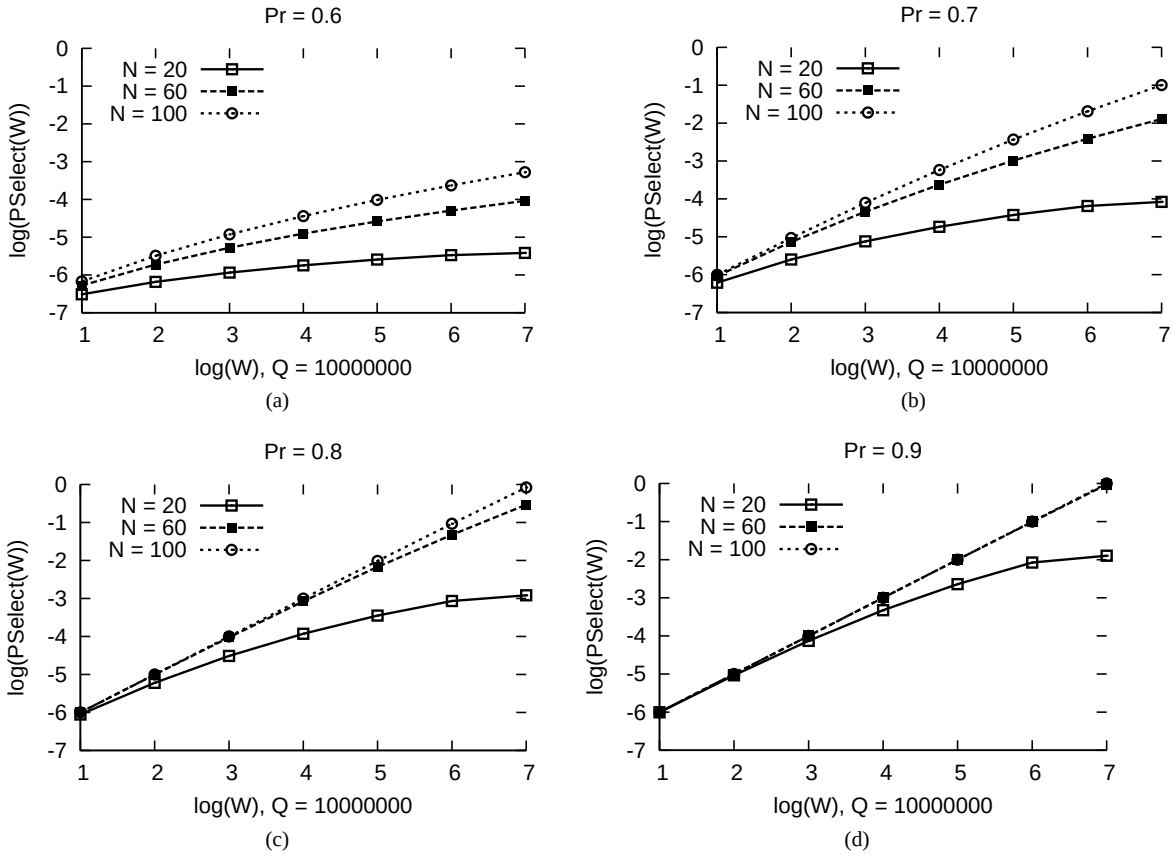

Figure 2: Effect of number of models W evaluated on probability $PSelect(W)$ of selecting M^*, assuming a model space of size Q=10^7, for various numbers N of observations and model strengths pr. Plots are log_{10} vs. log_{10} .

Figures 2 (a)-(d) plot $log_{10}(W)$ vs. $log_{10}(PSelect(W))$, where, recall,

$$PSelect(W) = Pr\{M^* \textit{ is selected from among the } W \textit{ we evaluate}\}$$

Here, the size Q of the model space is held fixed at 10^7, and pr and N are varied as in Figure 1. The plots indicate that for strong models (pr near 1.0) and relatively large data sets there is a nearly linear increase in $Pselect(W)$ as W increases. For weaker models and relatively sparser data, the probability of selecting the true model grows sublinearly as the number W of models evaluated increases.

6. Conclusions and Future Work

We have defined a notion of coherence for a model evaluation measure. A violation of coherence implies that a measure's evaluations may result in inconsistent model selection with respect to actual model posterior. We study in particular violations of complexity coherence, where models differing only in their complexities experience a non-coherent evaluation. We demonstrate that the common evaluation measures apparent error rate (cross validated and not), the BD_{ue} metric with vanishing equivalent sample size, and standard MDL scoring are not complexity coherent.

Our results are in general agreement with Schaffer [23]. If a coherent evaluation measure is used, overfit avoidance is justified only if there is an *a priori* preference for simple models to be true more often than complex models. However, one of our central tenets is that if the non-coherent apparent error rate is used, there is a bias for complex models to score well, and a complexity adjustment in such cases is appropriate, independent of distributional assumptions on model complexity.

The model space results presented in Section 5 demonstrate that when a coherent measure such as data likelihood is used, the oversearch phenomenon described by Quinlan and Cameron-Jones [20] cannot occur. The more models evaluated by a coherent measure, the higher the probability that the true model will be selected, regardless of the complexities of the models evaluated. This result suggests that the evaluation criteria utilized in experiments where oversearch has been observed may be non-coherent. However, in work such as [20], Laplace error is used to evaluate individual rules, not complete classification trees, each of whose leaves corresponds to individual rules, and it is not immediately clear how to assess the coherence of the resulting evaluation procedure in the context of the model selection problem.

There also is a relationship between our results and Occam razor and PAC learning generalization results. A hypothesis that encodes the training labels with (for example) zero errors is akin to a probability model incurring zero apparent training errors, and, similarly, to achieving a perfect DL_{data} term of MDL. Since our results show each of these evaluation measures to be non-coherent, both theories produce weaker generalization bounds (in our terms, less chance of the model being true) when models of increasing complexity are considered. However, it should be noted that the Occam razor and PAC results derive from the dependence between generalization bounds and the number of models of each complexity which exist, whereas our results derive from the interaction between the complexity of a *single* model and noncoherent evaluation measures.

An area for future research is to investigate how the evaluation ratio can be used as a correction factor for non-coherent measures, and how the correction correlates with existing factors, such as those supplied by structural risk minimization [29], the Akaike Information Criterion (AIC) [22], and the Bayesian Information Criterion (BIC) [24]. As we observed at the conclusion of Section 3.3, a p-value-like correction based solely on score distributions of random models is potentially misleading, but utilizing fully the evaluation ratio appears promising.

We are exploring the effect of relaxing some of our model assumptions. Allowing multiple class conditional probabilities pr_i simply moves the distributions from binomials to the more complicated multinomials, but in most cases does not alter the results in any fundamental way. On the other hand, relaxing the disjoint and uncorrelated feature set assumption makes many of the analyses considerably more intricate. Non-truth models are no longer random (though, in a large space, "almost all" models other than M^* would remain random), since features shared or correlated with M^*'s features would correlate with the class label. In such a model space, the interaction with search becomes important. How would the application of directed search (e.g., greedy or beam search) interact with evaluation measures and our current conclusions regarding the model space and the *a priori* probability of selecting the true model M^*?

We are exploring also modifications to some of the non-coherent evaluation measures considered here. In addition to the modifications to MDL discussed at the conclusion of Section 4.4, the *BD* metric can be studied under other values for the Dirichlet parameters. For example, Steck [26] considers the behavior of the Dirichlet distribution for a range of equivalent sample sizes α. Also, the non-likelihood equivalent $K2$ metric [10], in which all parameters α_{pv} are assigned the value 1.0, can be considered. While it is not immediately clear how the resulting evaluation measures will behave with respect to complexity coherence, it is clear that as the equivalent sample size grows, biases for distributions θ of one form over another will increase in strength, presenting potential evaluation anomalies of their own.

Acknowledgements. The author wishes to thank Vikas Hamine and Haixia Jia for their many useful comments and suggestions on early drafts of this work. The author additionally thanks Vikas Hamine for rendering the plots of Figures 1 and 2.

References

[1] Blum, A., and Langford, J. 2003. PAC-MDL bounds. *Proceedings of the 16th Annual Conference on Computational Learning Theory, COLT '03*.

[2] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. 1987. Occam's razor. *Information Processing Letters* 24, 377-380.

[3] Breiman, L., Friedman, J., Olshen, R., and Stone, C. 1984. *Classification and Regression Trees.* Wadsworth and Brooks, Pacific Grove, CA.

[4] Buntine, W. 1992. Learning classification trees. *Statistics and Computing* 2, 63–73.

[5] Feller, W. 1968. *An Introduction to Probability Theory and its Applications, Vol. I, 3rd Edition.* John Wiley & Sons, New York.

[6] Friedman, N., Geiger, D., and Goldszmidt, M. 1997. Bayesian network classifiers. *Machine Learning* 29, 131–163.

[7] Friedman, N., and Goldszmidt, M. 1996. Learning Bayesian networks with local structure. In *Proceedings 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, 211–219, Morgan Kaufmann.

[8] Grossman, D., and Domingos, P. 2004. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proceedings 21st International Conference on Machine Learning*, 361–368 .

[9] Haussler, D. 1990. Probably approximately correct learning. In *Proceedings of the 8th National Conference on Artificial Intelligence* 90, 1101-1108.

[10] Heckerman, D., Geiger, D., and Chickering, D. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20, 197–243.

[11] Helman, P., Veroff, R., Atlas, SR., and Willman, C. 2004 A Bayesian network classification methodology for gene expression data. *Journal of Computational Biology* 11, 581-615.

[12] Kass, R., and Wasserman, L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 91(431), 1343–1370.

[13] Kearns, M., Mansour, Y., Ng, A., and Ron, D. 1997. An experimental and theoretical comparison of model selection methods. *Machine Learning* 27(1), 7-50.

[14] Lam, W., and Bacchus, F. 1994. Learning Bayesian belief networks: an approach based on the MDL principle, *Computational Intelligence* 10, 269–293.

[15] Langford, J., and Blum, A. 2003. Microchoice bounds and self bounding learning algorithms. *Machine Learning* 51(2), 165-179.

[16] MacKay, D. 1995. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* 6, 469–505.

[17] Murphy, P., and Pazzani, M. 1994. Exploring the decision forest: an empirical investigation of Occam's razor in decision tree induction. *Journal of Artificial Intelligence Research* 1, 257–275.

[18] Pearl, J. 1988. *Probabilistic reasoning for intelligent systems.* Morgan Kaufmann, San Francisco.

[19] Pearl, J., and Verma, T. 1991. A theory of inferred causation. In *Knowledge Representation and Reasoning: Proc. 2nd International Conference*, 411–452, Morgan Kaufmann.

[20] Quinlan, J., and Cameron-Jones, R. 1995. Oversearching and layered search in empirical learning. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1019-1024.

[21] Rissen, J. 1978. Modeling by shortest data description. *Automatica* 14, 465–471.

[22] Sakamoto, T., Ishiguro, M., and Kitagawa, G. 1986. *Akaike Information Criterion Statistics*, D. Reidel, Holland.

[23] Schaffer, C. 1993. Overfitting avoidance as bias. *Machine Learning* 10, 153-178.

[24] Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

[25] Segal, R. 1996. An analysis of oversearch. Unpublished manuscript.

[26] Steck, H., and Jaakkola, T. 2002. On the Dirichlet prior and Bayesian regularization. In *Advances in Neural Information Processing Systems 15*.

[27] Syversveen, A. 1998. Noninformative Bayesian priors, interpretation and problems with construction and applications. *Preprint No,3/98, http://www.math.ntnu.no/preprint/statistics/1998*.

[28] Valiant, L. 1984. A theory of the learnable. *Communications of the ACM* 27(11), 1134–1142.

[29] Vapnik, V. 1998. *Statistical Learning Theory.* Wiley-Interscience, New York.

[30] Webb, G. 1996. Further experimental evidence against the utility of Occam's Razor. *Journal of Artificial Intelligence Research* 4, 397–417.