Bayes Factors, Relations to Minimum Description Length, and Overlapping Model Classes

Richard Shiffrin, Suyog Chandramouli, Peter Grünwald

Abstract

This article presents a non-technical perspective on two prominent methods for analyzing experimental data in order to select among model classes. Each class consists of model instances; each instance predicts a unique distribution of data outcomes. One method is Bayesian Model Selection (BMS), instantiated with the Bayes factor. The other is based on the Minimum Description Length principle (MDL), instantiated by a variant of Normalized Maximum Likelihood (NML): the variant is termed NML* and takes prior probabilities into account. The methods are closely related. The Bayes factor is a ratio of two values: V_1 for model class M_1, and V_2 for M_2. Each V_j is the *sum* over the instances of $M_{j,}$ of the joint probabilities (prior times likelihood) for the observed data, normalized by a sum of such sums for all possible data outcomes. NML* is qualitatively similar: The value it assigns to each class is the *maximum* over the instances in M_i of the joint probability for the observed data normalized by a sum of such maxima for all possible data outcomes. The similarity of BMS to NML* is particularly close when model classes do not have instances that overlap, a way of comparing model classes that we advocate generally. These observations and suggestions are illustrated throughout with use of a simple example borrowed from Heck, Wagenmakers, and Morey (2015) in which the instances predict a binomial distribution of number of success in N trials. The model classes posit the binomial probability of success to lie in various regions of the interval [0,1]. We illustrate the theory and the example not with equations but with tables coupled with simple arithmetic. Using the binomial example we carry out comparisons of BMS and NML* that do and do not involve model classes that overlap, and do and do not have uniform priors. When the classes do not overlap BMS and NML* produce qualitatively similar results.

Bayes Factors, Relations to Minimum Description Length, and Overlapping Model Classes

Introduction

This article is a largely non-technical perspective on methods to choose among candidate models purported to explain observed data. There are numerous goals of modeling, such as prediction of more data collected under the same conditions, generalization to similar but different situations, elegance, simplicity, approximating unknown truth, gaining understanding of complex situations, maximizing utility based on applications of the model, and others. No method can simultaneously accomplish all these goals. This article will cover the two methods that have emerged as 'best compromises' in satisfying many of these goals. We will not try to lay out the precise way in which the methods satisfy the different goals, because that is a highly technical matter well beyond the scope of this paper, but it will become clear that they are aimed at providing a good fit to observed data with as simple a model as possible, while taking prior knowledge into account.

One method is Bayesian Model Selection (BMS), instantiated in the form of Bayes factors. The other is the Minimum Description Length principle (MDL), instantiated by a variant of Normalized Maximum Likelihood (NML) that takes prior probabilities into account; we term this NML*. NML and NML* are the same for uniform priors. We show a surprisingly simple relation of BMS to NML* that helps explain why the two methods often give qualitatively similar results. One situation in which the two methods differ, apparently to the detriment of NML, occurs when the model classes under comparison overlap. We suggest that in most such cases it is better to change the model comparison so that the classes do not overlap; generally this may be accomplished by deleting the shared model instances from the larger class.

Throughout the article we illustrate the observations and suggestions with a simple example borrowed from Heck, Wagenmakers, and Morey (2015). They compare Bayes factors (BF) and NML for Bernoulli model classes inferring the probability of success, θ, for flat priors (i.e. uniform priors): One model class posits θ to lie in the range [0,1], which we term M_3, and the other posits θ to lie in a restricted range [0,z], which we term M_1: the instances in the range [0,z] are common to the two classes. Certain results seem to favor the use of BF over NML*. We compare their findings with those arising when the comparison of model classes is altered to compare θ in [0,z] vs θ in (z,1], the latter we term M_2. We also consider a variant of their example with geometric priors. In all cases that compare M_1 vs. M_2, BMS and NML* produce qualitatively similar results.

Terminology, definitions, and an exposition using tables

We assume that all measurements and all probabilities in this article are discretized into suitably small intervals. Using discretized measures to approximate continuous distributions not only simplifies our exposition but also accords with actual practice, and matches computational approaches used in model selection. In addition, discretization implicitly recognizes that all models are approximations to reality, reflects the fact that no measurements are infinitely precise, and limits inference to practical levels of precision. Among other things this assumption implies that all model instances in all model classes, the probabilities assigned to those, and all experimental outcomes are countable in number. The more typical treatments are based on continuous distributions and integrals of those. In the present treatment all integrals are replaced by sums.

We use x_i to denote the potential outcomes of an experiment, in terms of the variables recorded. Thus if a study measures number of successes in N trials, the potential outcomes x_i are the integral numbers from 0 to N. We use y to denote the observed outcome of a study (y will match one of the x_i).

It is important in this article to distinguish a model instance, defined as a member of a model class with the values of all parameters specified, from a model class, usually characterized as a set of instances with the same functional form but with parameter values that are not (yet) specified: M_j denotes a given model class. This distinction is critical because all inference is carried out on the basis of model instances: A model class does not predict any particular distribution of outcomes other than the collection of distributions or data outcomes predicted by its instances. (In other writings model classes are sometimes called 'models' and model instances called 'hypotheses'. Our terminology is chosen to reduce confusion as far as possible).

We could describe the k-th model instance of model class M_j as $M_j(\theta_k)$, where θ_k denotes the set of values assigned to the parameters. However, this is cumbersome terminology and it will be clear in context if we denote a model instance simply by θ_k. Thus a class of linear regression models could be written as $u = a\text{w} + b + \varepsilon$ where, u and w are dependent and explanatory variables, ε is representative of Gaussian noise ($\varepsilon \sim N(0, \sigma^2)$) with mean of zero and unspecified variance σ^2 , and a and b are parameters whose values also remain to be specified. A model instance, θ_k in this linear class might be $u = 3\text{w} + 2 + N(0,1)$ This instance predicts a distribution that is Gaussian with mean 3w + 2 and variance 1.

In the examples we will be using in this article, the model classes will be sets of model instances; each model instance is a unique binomial distribution, with the one parameter, denoting the probability of success operating in N independent trials. If we will call this probability of success parameter θ, then a distribution for a given θ represents a model instance θ_k in our model space. The probability of n successes in N trials will, for a given probability of success θ be:

$$p(n) = \binom{N}{n} (\theta)^n (1-\theta)^{N-n} \qquad (1)$$

Our aim in this article is a non-technical exposition that will be clear to readers unfamiliar with model selection techniques and of varying types of background knowledge. For this reason we illustrate the methods and results with tables and keep equations to a minimum (the equations corresponding to the tables will be given in appendices).

The general situation prior to carrying out analysis of experimental data is depicted in Table 1a: Along the top row of the table are the possible data outcomes (x_i). In the examples in this article these outcomes will be the number of successes, n, in N trials (see Table 1b). The rows below the top row list the distributions associated with the different model instances; the sum of probabilities in each row is 1.0. Thus p_{ki} gives the probability of outcome x_i conditional on model instance θ_k generating the data. The column to the right of the table lists the model instances, θ_k, that are found in two classes under consideration; each instance is listed to the right of the distribution it predicts. There is a one-to-one correspondence of model instances with their associated distributions. To the right of each instance is the prior probability of that instance, based on all knowledge prior to the current experiment; the sum of prior probabilities down the column is 1.0.

It is important to note that Tables 1a to 1d list all model instances in all classes. BMS in the form of Bayes Factors and MDL in the form of NML* typically treat just one model class at a time. In this case only the instances in a single class are considered, and the prior (and posterior) probabilities would be normalized to add to 1.0 within each class. We will later show tables for the individual classes--they can be obtained from the present ones by truncating the table and normalizing the resultant priors. However showing all instances in all classes together allows one to see not only the prior probabilities of instances relative to each other within class, but also the prior probabilities of instances relative to each other between classes, and hence the prior probabilities of the classes. When all instances in all classes are considered together, as in Table 1a, we refer to the priors as *unconditional*; when the instances are restricted to a single class we refer to the priors as *conditional* (conditioned on the class).

A few model classes are illustrated on the right side of the table: Two of the classes, M_1 and M_2, do not overlap, as indicated by the medium size brackets; for convenience all the instances in M_1 are listed above those in M_2. The Heck et al. (2015) article analyzes a case when two classes overlap, in that one class is a proper subset of the other. We represent that situation in the table with model class M_3 that includes all the instances, as shown by the largest bracket to the far right.

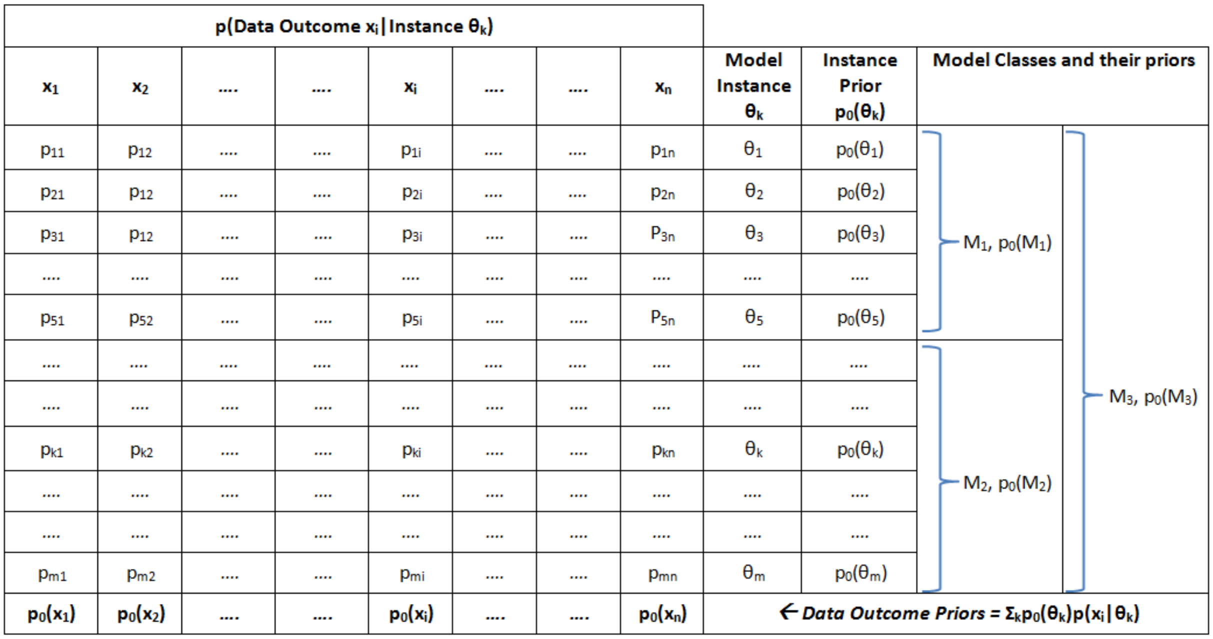
The assignment of prior probabilities to model instances is of course subjective and has engendered many different approaches over the years, among other things depending on one's inference goals. This issue will be discussed later in this article. A natural way to assign probabilities (prior or posterior) to model classes is to sum the unconditional probabilities (prior or posterior) for the instances in that class (see Appendix 1). This idea works well for classes that do not overlap, but is problematic when the two classes share instances. How to handle class probabilities when the classes do and do not overlap will be taken up in the body of this article.[1]

Footnote 1: Each model instance predicts a unique and different data distribution. Normally there would be a one-to-one mapping of instances to data distributions but due to discretization there could be several instances, even ones with slightly different functional form, that predict the same interval of data distributions. That possibility will not arise in the examples examined in this article: Each instance in the left column will be adjacent to the row showing a binomial distribution based on that value of θ.

Table 1b gives a version of Table 1a for the binomial example, with ten trials having number of possible successes ranging from 0 to 10, with eleven model instances positing $\theta = 0, .1, .2, ..., 1.0$, with a uniform prior for all instances, with model class M_1 positing values of θ to be one of [0, .1, .2], model class M_2 having the remaining values of θ, and model class M_3 positing θ to have all eleven model instances.

In Tables 1a and 1b the row at the bottom of the table gives the prior probabilities of the different data outcomes: $p_0(x_i) = \Sigma_k(p_0[\theta_k])p(x_i|\theta_k)$, i.e. the sum of the prior column times the likelihood column for outcome x_i. This prior is due to assigning probabilities to all instances in all classes. This result, and also the calculation of posterior probabilities, is even easier to see in an alternate form of these tables: The entries in Tables 1c and 1d give the joint probabilities of a given outcome and the prior probability of a given instance. The entries in the body of Table 1c (1d) are obtained from Table 1a (1b) by taking the product of the each entry in the column of prior probabilities times each entry in each column in the body of the 1a (1b) table.[2] As a result, in Table 1c (1d), the sum of all interior entries adds to 1.0, the column sums give the prior probabilities of the data outcomes and the row sums give the prior probabilities of the model instances. The circled entries in Tables 1c and 1d indicate the maximum value in each column (i.e. for each data outcome, the maximum joint probability of a model instance and the outcome across all candidate model instances. These joint probabilities depend on instance priors: Each is the instance prior times the likelihood of an outcome given that instance. These maxima will be used when describing NML*; ignore them for now.

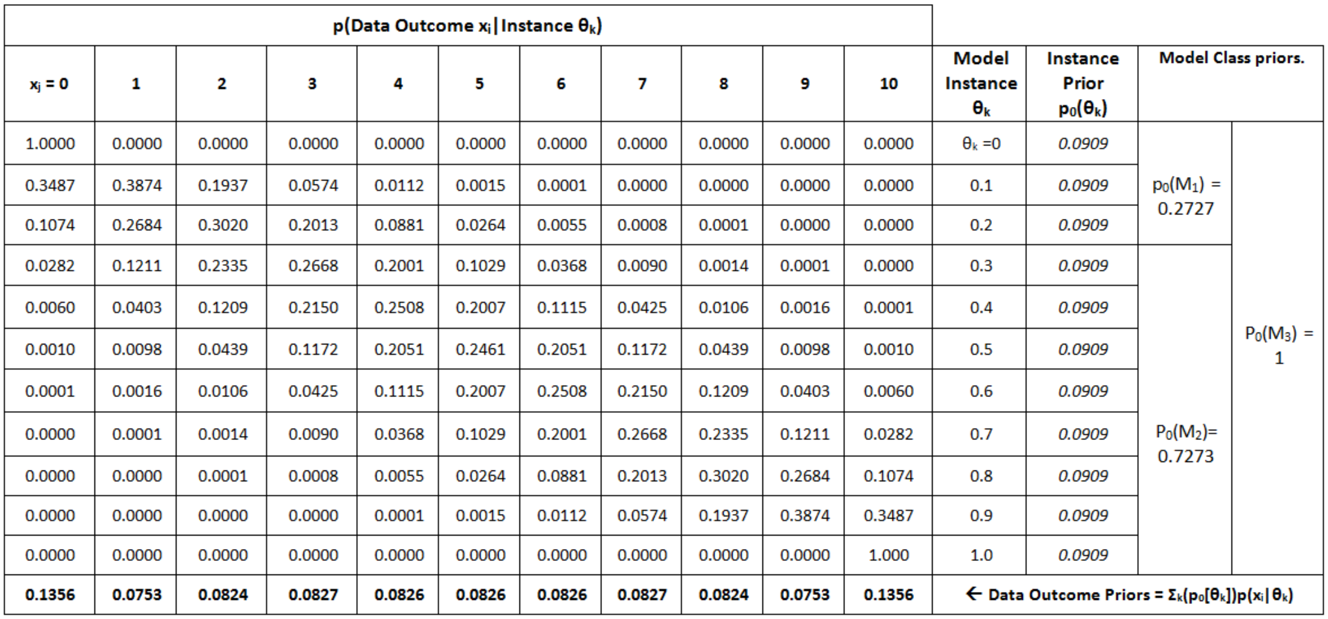
Table 1a - Unconditional Priors of Model Instances and likelihood of outcomes for each. (General case before analyzing experimental data)



p(Data Outcome x_i|Instance θ_k)											
x_1	x_2	....	....	x_i	....	....	x_n	Model Instance θ_k	Instance Prior $p_0(\theta_k)$	Model Classes and their priors	
p_{11}	p_{12}	....	....	p_{1i}	....	....	p_{1n}	θ_1	$p_0(\theta_1)$	M_1, $p_0(M_1)$	M_3, $p_0(M_3)$
p_{21}	p_{12}	....	....	p_{2i}	....	....	p_{2n}	θ_2	$p_0(\theta_2)$		
p_{31}	p_{12}	....	....	p_{3i}	....	....	P_{3n}	θ_3	$p_0(\theta_3)$		
....	....	....	....	....	....	....	....	....	....		
p_{51}	p_{52}	....	....	p_{5i}	....	....	P_{5n}	θ_5	$p_0(\theta_5)$		
....	....	....	....	....	....	....	....	....	....	M_2, $p_0(M_2)$	
....	....	....	....	....	....	....	....	....	....		
p_{k1}	p_{k2}	....	....	p_{ki}	....	....	p_{kn}	θ_k	$p_0(\theta_k)$		
....	....	....	....	....	....	....	....	....	....		
....	....	....	....	....	....	....	....	....	....		
p_{m1}	p_{m2}	....	....	p_{mi}	....	....	p_{mn}	θ_m	$p_0(\theta_m)$		
$p_0(x_1)$	$p_0(x_2)$	....	....	$p_0(x_i)$	....	....	$p_0(x_n)$	← *Data Outcome Priors* = $\Sigma_k p_0(\theta_k)p(x_i\|\theta_k)$			

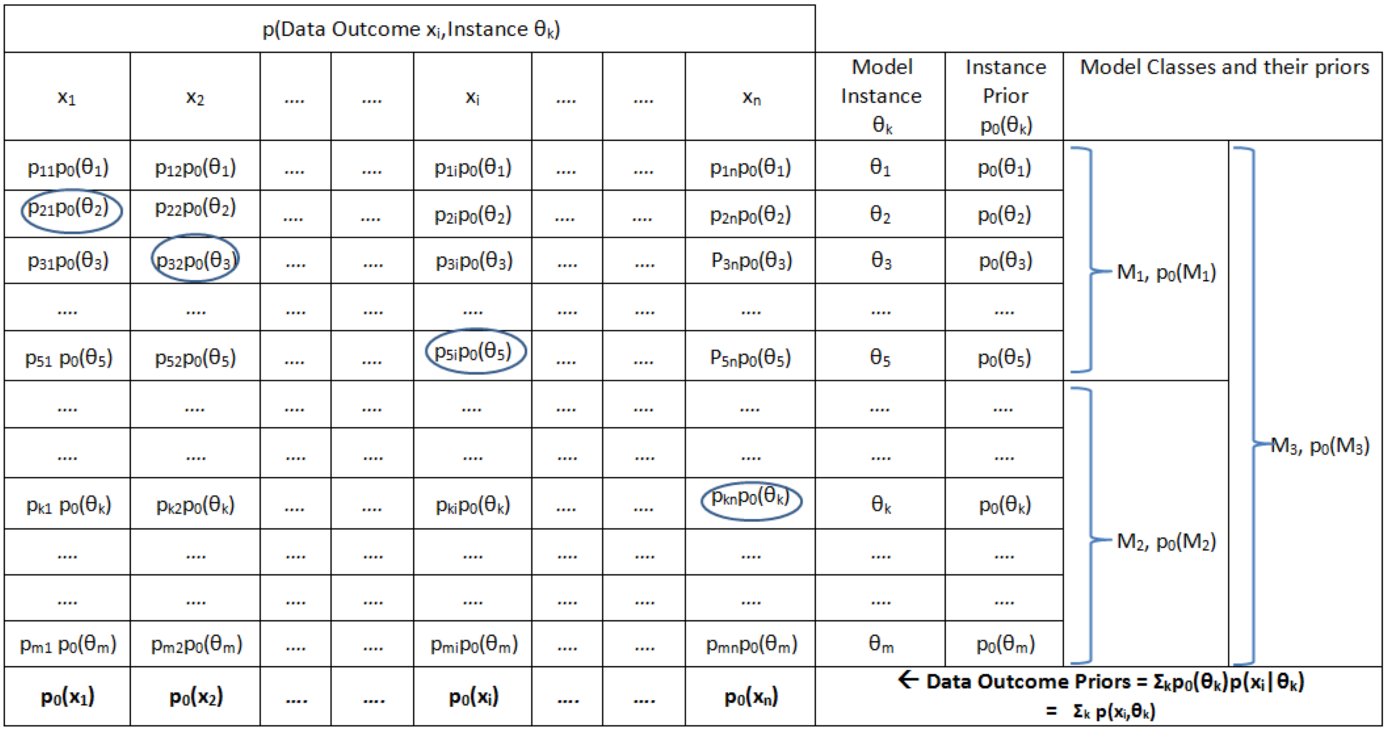
Footnote 2: The entries in the table could be written in an equivalent form that makes the joint probability more evident: $p_{ki}p_0(\theta_k) = p_0(x_i,\theta_k)$.

Table 1b - Unconditional Priors of Model Instances and likelihood of outcomes for each. (Binomial example with N=10 before analyzing experimental data)



$p(\text{Data Outcome } x_i \mid \text{Instance } \theta_k)$														
$x_j = 0$	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Prior $p_0(\theta_k)$	Model Class priors.	
1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	$\theta_k = 0$	*0.0909*	$p_0(M_1) =$ 0.2727	$P_0(M_3) =$ 1
0.3487	0.3874	0.1937	0.0574	0.0112	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000	0.1	*0.0909*		
0.1074	0.2684	0.3020	0.2013	0.0881	0.0264	0.0055	0.0008	0.0001	0.0000	0.0000	0.2	*0.0909*		
0.0282	0.1211	0.2335	0.2668	0.2001	0.1029	0.0368	0.0090	0.0014	0.0001	0.0000	0.3	*0.0909*	$P_0(M_2)=$ 0.7273	
0.0060	0.0403	0.1209	0.2150	0.2508	0.2007	0.1115	0.0425	0.0106	0.0016	0.0001	0.4	*0.0909*		
0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010	0.5	*0.0909*		
0.0001	0.0016	0.0106	0.0425	0.1115	0.2007	0.2508	0.2150	0.1209	0.0403	0.0060	0.6	*0.0909*		
0.0000	0.0001	0.0014	0.0090	0.0368	0.1029	0.2001	0.2668	0.2335	0.1211	0.0282	0.7	*0.0909*		
0.0000	0.0000	0.0001	0.0008	0.0055	0.0264	0.0881	0.2013	0.3020	0.2684	0.1074	0.8	*0.0909*		
0.0000	0.0000	0.0000	0.0000	0.0001	0.0015	0.0112	0.0574	0.1937	0.3874	0.3487	0.9	*0.0909*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.000	1.0	*0.0909*		
0.1356	**0.0753**	**0.0824**	**0.0827**	**0.0826**	**0.0826**	**0.0826**	**0.0827**	**0.0824**	**0.0753**	**0.1356**	**← Data Outcome Priors = $\Sigma_k(p_0[\theta_k])p(x_i \mid \theta_k)$**			

Table 1c - Joint Probability of Model Instances and Data Outcomes (General case before analyzing experimental data)

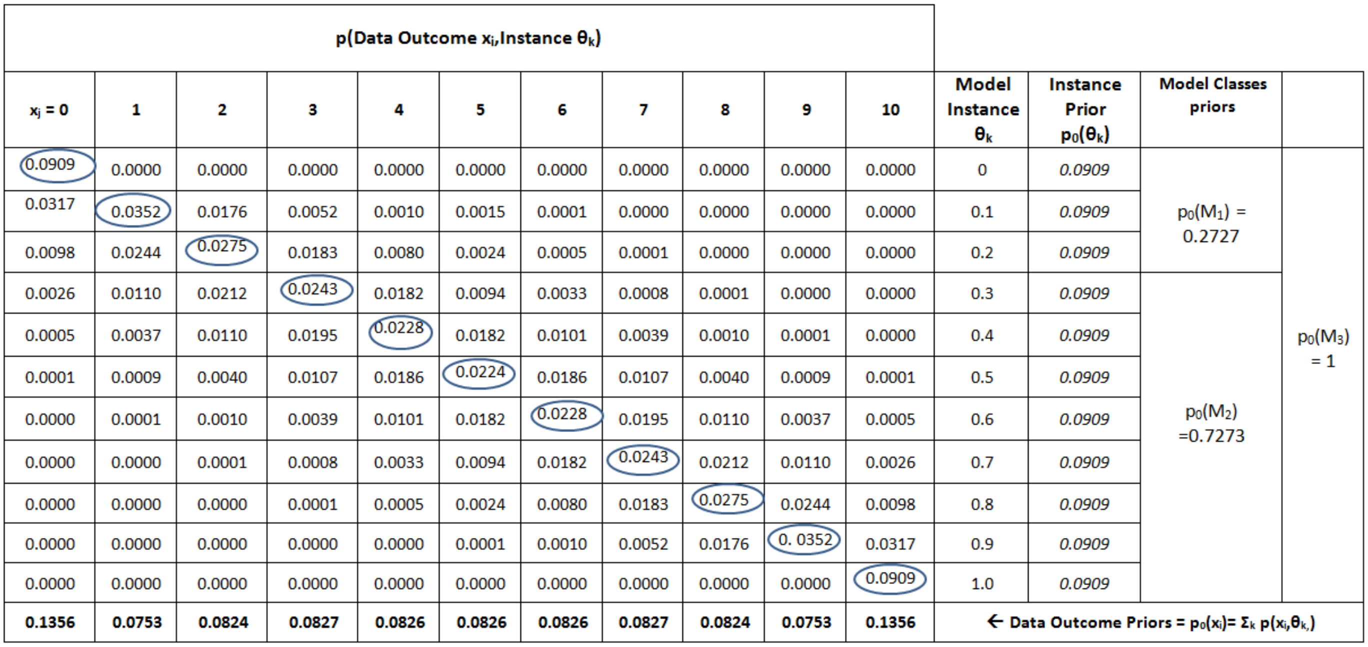


$p(\text{Data Outcome } x_i, \text{Instance } \theta_k)$											
x_1	x_2	....	....	x_i	....	....	x_n	Model Instance θ_k	Instance Prior $p_0(\theta_k)$	Model Classes and their priors	
$p_{11}p_0(\theta_1)$	$p_{12}p_0(\theta_1)$	....	....	$p_{1i}p_0(\theta_1)$	....	....	$p_{1n}p_0(\theta_1)$	θ_1	$p_0(\theta_1)$	M_1, $p_0(M_1)$	M_3, $p_0(M_3)$
$p_{21}p_0(\theta_2)$	$p_{22}p_0(\theta_2)$	....	....	$p_{2i}p_0(\theta_2)$	....	....	$p_{2n}p_0(\theta_2)$	θ_2	$p_0(\theta_2)$		
$p_{31}p_0(\theta_3)$	$p_{32}p_0(\theta_3)$	....	....	$p_{3i}p_0(\theta_3)$	....	....	$P_{3n}p_0(\theta_3)$	θ_3	$p_0(\theta_3)$		
....	....	....	....	....	....	....	....	....	....		
$p_{51}\ p_0(\theta_5)$	$p_{52}p_0(\theta_5)$	....	....	$p_{5i}p_0(\theta_5)$	....	....	$P_{5n}p_0(\theta_5)$	θ_5	$p_0(\theta_5)$		
....	....	....	....	....	....	....	....	....	....	M_2, $p_0(M_2)$	
....	....	....	....	....	....	....	....	....	....		
$p_{k1}\ p_0(\theta_k)$	$p_{k2}p_0(\theta_k)$	....	....	$p_{ki}p_0(\theta_k)$	....	....	$p_{kn}p_0(\theta_k)$	θ_k	$p_0(\theta_k)$		
....	....	....	....	....	....	....	....	....	....		
....	....	....	....	....	....	....	....	....	....		
$p_{m1}\ p_0(\theta_m)$	$p_{m2}p_0(\theta_m)$	....	....	$p_{mi}p_0(\theta_m)$	....	....	$p_{mn}p_0(\theta_m)$	θ_m	$p_0(\theta_m)$		
$p_0(x_1)$	**$p_0(x_2)$**	**....**	**....**	**$p_0(x_i)$**	**....**	**....**	**$p_0(x_n)$**	**← Data Outcome Priors = $\Sigma_k p_0(\theta_k)p(x_i \mid \theta_k)$ $= \Sigma_k\ p(x_i, \theta_k)$**			

Applications of Bayes Theorem to Model Instances

Bayes Theorem can be used to modify Tables 1a-1d to show the results of Bayesian inference, based on the observed data outcome y. These tables of posteriors could serve as tables of priors for a continuation/replication of the same study. We show only posterior versions of Tables 1a and 1b: Tables 2a and 2b differ from Tables 1a and 1b only in the change from prior probabilities to posterior probabilities conditional on observing outcome y, as indicated by the column on the right side, and in the new probabilities of various outcomes, as indicated in the row at the bottom. For the binomial example (Table 2b) we assume three successes have been observed. The posterior gives the adjusted probability based both on the prior probability of the instance and the likelihood of the observed data y, the likelihood being conditional on the assumption that the observed data y was sampled from the distribution associated with the instance in question. For observed data outcome y, the posterior probability of θ_k, $p(\theta_k|y)$, is just $p(y|\theta_k)p_0(\theta_k)/p_0(y) = p_{yk}p_0(\theta_k)/p_0(y)$. I.e. in Table 1a, take the entry in the row for θ_k and the column for outcome y, multiply by the prior probability of θ_k and divide by the prior probability for outcome y given below that column; in Table 1b we do the same for three successes (the fourth column). The same posteriors can of course be obtained from Tables 1c and 1d by taking an entry in column y and dividing by the column sum below. Either way, this is just a restatement of Bayes Theorem (see Appendix 1). The posterior probabilities we see are those for all model instances in all classes considered together.

Table 1d - Joint Probability of Model Instances and Data Outcomes
(Binomial example with N=10 before analyzing experimental data)



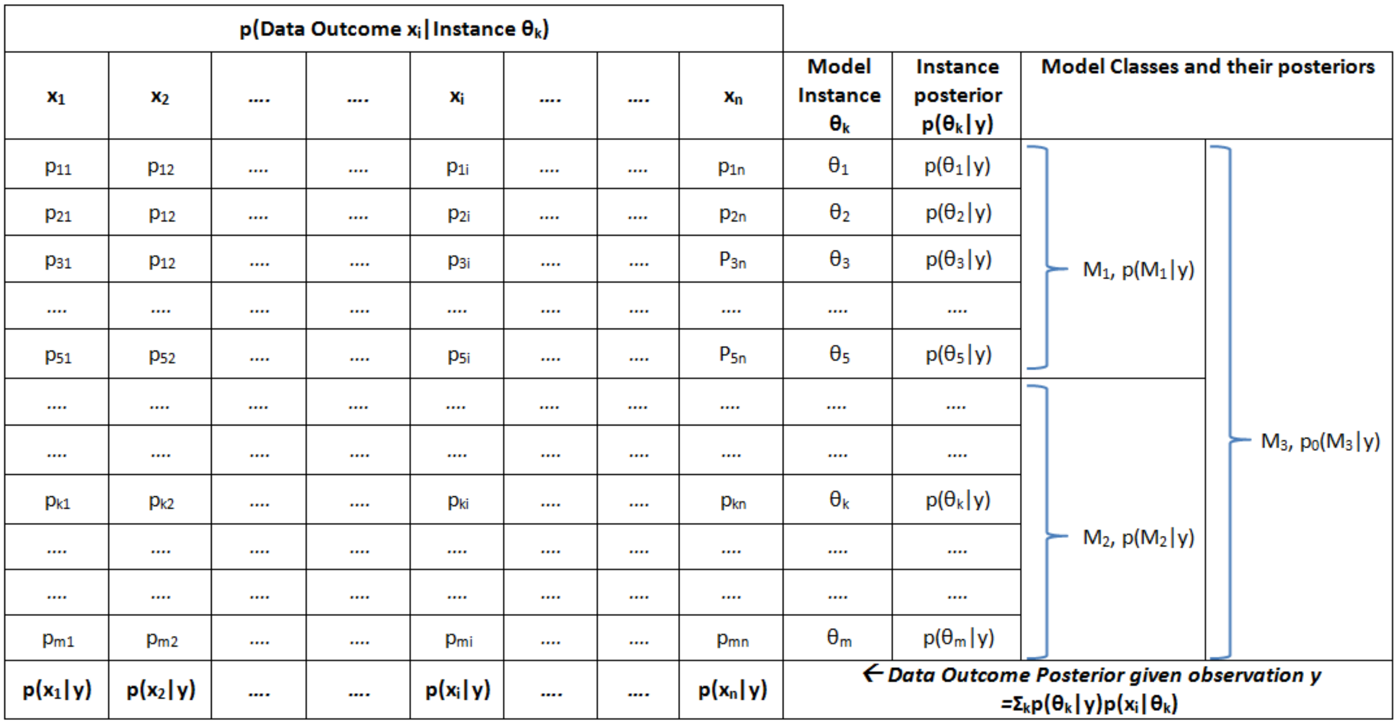
p(Data Outcome x_i,Instance θ_k)														
x_j = 0	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Prior $p_0(\theta_k)$	Model Classes priors	
0.0909	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	*0.0909*	$p_0(M_1)$ = 0.2727	$p_0(M_3)$ = 1
0.0317	0.0352	0.0176	0.0052	0.0010	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000	0.1	*0.0909*		
0.0098	0.0244	0.0275	0.0183	0.0080	0.0024	0.0005	0.0001	0.0000	0.0000	0.0000	0.2	*0.0909*		
0.0026	0.0110	0.0212	0.0243	0.0182	0.0094	0.0033	0.0008	0.0001	0.0000	0.0000	0.3	*0.0909*	$p_0(M_2)$ =0.7273	
0.0005	0.0037	0.0110	0.0195	0.0228	0.0182	0.0101	0.0039	0.0010	0.0001	0.0000	0.4	*0.0909*		
0.0001	0.0009	0.0040	0.0107	0.0186	0.0224	0.0186	0.0107	0.0040	0.0009	0.0001	0.5	*0.0909*		
0.0000	0.0001	0.0010	0.0039	0.0101	0.0182	0.0228	0.0195	0.0110	0.0037	0.0005	0.6	*0.0909*		
0.0000	0.0000	0.0001	0.0008	0.0033	0.0094	0.0182	0.0243	0.0212	0.0110	0.0026	0.7	*0.0909*		
0.0000	0.0000	0.0000	0.0001	0.0005	0.0024	0.0080	0.0183	0.0275	0.0244	0.0098	0.8	*0.0909*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0010	0.0052	0.0176	0. 0352	0.0317	0.9	*0.0909*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0909	1.0	*0.0909*		
0.1356	**0.0753**	**0.0824**	**0.0827**	**0.0826**	**0.0826**	**0.0826**	**0.0827**	**0.0824**	**0.0753**	**0.1356**	**← Data Outcome Priors = $p_0(x_i)$= Σ_k $p(x_i,\theta_k)$**			

A later section of the article discusses in more detail BMS for model class comparisons. Here we note only that the posterior probabilities for instances can be used to make inferences about the model classes containing those instances, at least for model classes that do not overlap. One must be careful to consider the relation of instance probabilities to class probabilities: E.g. equal probabilities assigned to all instances when two classes differ in size imply that the prior probabilities of the two classes differ. This is seen in Table 1b where there are more instances in Class M_2. If we wanted to make the two classes equal in probability we could leave the instance probabilities uniform within class, but then would have to make the unconditional instance probabilities lower in the larger class. Doing this introduces a preference for smaller (simpler) classes, and is one basis for Bayesian Model Selection.

Geometric and Uniform Priors in the Binomial Example

There are many reasons to include prior knowledge in inference, and many cases where one would want to do so, both for BMS and MDL, although there are other reasons why one most often sees in practice the use of uninformative, uniform, or transformation invariant priors. Thus, in the framework of our binomial example, we analyze both a uniform prior and a geometric prior.[3]

Table 2a - Posterior Probabilities of Model Instances given observed Data (General case after analyzing experimental data)

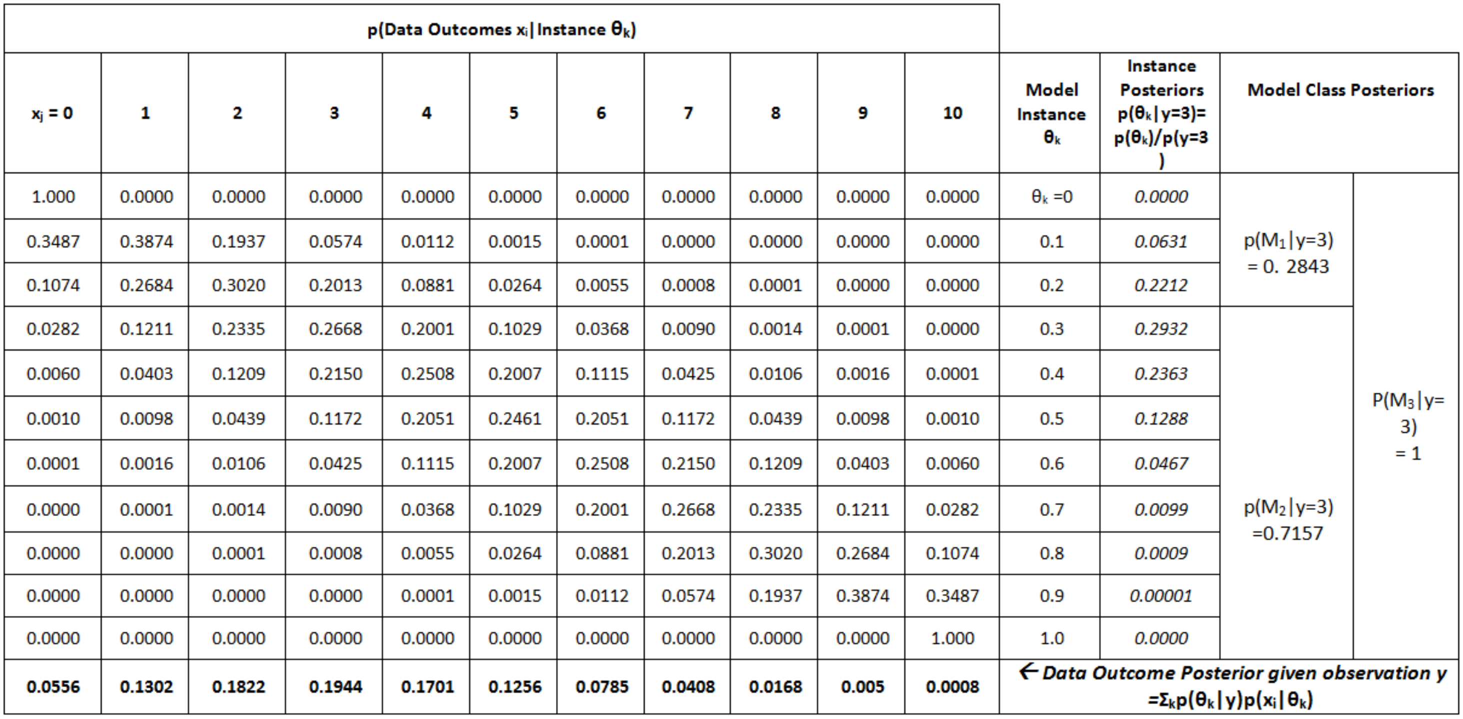


p(Data Outcome x_i | Instance θ_k)											
x_1	x_2	....	....	x_i	....	....	x_n	Model Instance θ_k	Instance posterior $p(\theta_k \| y)$	Model Classes and their posteriors	
p_{11}	p_{12}	....	....	p_{1i}	....	....	p_{1n}	θ_1	$p(\theta_1 \| y)$	M_1, $p(M_1 \| y)$	M_3, $p_0(M_3 \| y)$
p_{21}	p_{12}	....	....	p_{2i}	....	....	p_{2n}	θ_2	$p(\theta_2 \| y)$		
p_{31}	p_{12}	....	....	p_{3i}	....	....	P_{3n}	θ_3	$p(\theta_3 \| y)$		
....	....	....	....	....	....	....	....	....	....		
p_{51}	p_{52}	....	....	p_{5i}	....	....	P_{5n}	θ_5	$p(\theta_5 \| y)$		
....	....	....	....	....	....	....	....	....	....	M_2, $p(M_2 \| y)$	
....	....	....	....	....	....	....	....	....	....		
p_{k1}	p_{k2}	....	....	p_{ki}	....	....	p_{kn}	θ_k	$p(\theta_k \| y)$		
....	....	....	....	....	....	....	....	....	....		
....	....	....	....	....	....	....	....	....	....		
p_{m1}	p_{m2}	....	....	p_{mi}	....	....	p_{mn}	θ_m	$p(\theta_m \| y)$		
$p(x_1 \| y)$	$p(x_2 \| y)$	....	....	$p(x_i \| y)$	....	....	$p(x_n \| y)$	← *Data Outcome Posterior given observation y* $=\Sigma_k p(\theta_k \| y) p(x_i \| \theta_k)$			

Footnote 3: *MDL implemented as simple NML is too simple an approach to be an adequate model comparison metric for many reasons in addition to incorporation of priors, as discussed by Grünwald, 2007.*

In typical applications of BMS it is not the model instances that are assumed equally likely but the model classes: The instances then are assumed equally likely within each class, but have values differing between class so that sum of priors is the same for the two classes. Thus if the two classes have N_1 and N_2 instances, unconditional instance priors of $1/(2N_1)$ and $1/(2N_2)$ respectively will make the class prior probabilities equal. A prior that is uniform within class and equal for the classes being compared is termed

Table 2b - Posterior Probabilities of Model Instances given observed Data (Binomial example after an experiment with 3 successes observed in 10 trials)



p(Data Outcomes x_i | Instance θ_k)														
$x_j = 0$	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Posteriors $p(\theta_k \| y=3)= p(\theta_k)/p(y=3)$	Model Class Posteriors	
1.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	$\theta_k = 0$	*0.0000*	$p(M_1 \| y=3)$ = 0. 2843	$P(M_3 \| y=3)$ = 1
0.3487	0.3874	0.1937	0.0574	0.0112	0.0015	0.0001	0.0000	0.0000	0.0000	0.0000	0.1	*0.0631*		
0.1074	0.2684	0.3020	0.2013	0.0881	0.0264	0.0055	0.0008	0.0001	0.0000	0.0000	0.2	*0.2212*		
0.0282	0.1211	0.2335	0.2668	0.2001	0.1029	0.0368	0.0090	0.0014	0.0001	0.0000	0.3	*0.2932*	$p(M_2 \| y=3)$ =0.7157	
0.0060	0.0403	0.1209	0.2150	0.2508	0.2007	0.1115	0.0425	0.0106	0.0016	0.0001	0.4	*0.2363*		
0.0010	0.0098	0.0439	0.1172	0.2051	0.2461	0.2051	0.1172	0.0439	0.0098	0.0010	0.5	*0.1288*		
0.0001	0.0016	0.0106	0.0425	0.1115	0.2007	0.2508	0.2150	0.1209	0.0403	0.0060	0.6	*0.0467*		
0.0000	0.0001	0.0014	0.0090	0.0368	0.1029	0.2001	0.2668	0.2335	0.1211	0.0282	0.7	*0.0099*		
0.0000	0.0000	0.0001	0.0008	0.0055	0.0264	0.0881	0.2013	0.3020	0.2684	0.1074	0.8	*0.0009*		
0.0000	0.0000	0.0000	0.0000	0.0001	0.0015	0.0112	0.0574	0.1937	0.3874	0.3487	0.9	*0.00001*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.000	1.0	*0.0000*		
0.0556	**0.1302**	**0.1822**	**0.1944**	**0.1701**	**0.1256**	**0.0785**	**0.0408**	**0.0168**	**0.005**	**0.0008**	***← Data Outcome Posterior given observation y*** **$=\Sigma_k p(\theta_k \| y) p(x_i \| \theta_k)$**			

U, and used in the worked examples. In our examples we consider also a geometric prior for model instances that favors low values of θ, termed GL. In GL each (discrete) value of θ is a proportion 0.82 of the next lower, and all are normalized to sum to 1.0. Here we do not assume equal class priors, for reasons given later, and the prior for a class is simply the sum of the (unconditional and geometrically varying) priors for the instances in that class. For the case of 11 (binomial) model instances, the two priors are shown in the panels of Figure 1: For U the figure plots U for the case when there are three instances in class 1: θ = 0, .1, .2.

Bayesian Model Selection

Given some observed data, y, we have described how Bayes theorem can be applied to Tables 1a-1d to produce posterior probabilities for the instances in the model classes, and thereby for the classes themselves, as depicted in Tables 2a and 2b. In general it is undesirable to prefer model classes so large they can predict any possible outcome, and desirable to favor a model class that predicts a small range of outcomes when that range includes the observed data. It is typical to assign equal probability to classes, thereby requiring smaller instance probabilities for larger classes and thereby introducing a preference for the smaller class. The main result is given here in Eq. 2. When assessing two model classes it is typical to do this with the odds (both prior odds and posterior odds) favoring one over the other. E.g. the posterior odds for M_1 over M_2 is simply $p(M_1)/p(M_2)$. Appendix 1 give the derivations and shows that Bayes Theorem can be applied to produce $p(M_1|y)$, $p(M_2|y)$, and their ratio:

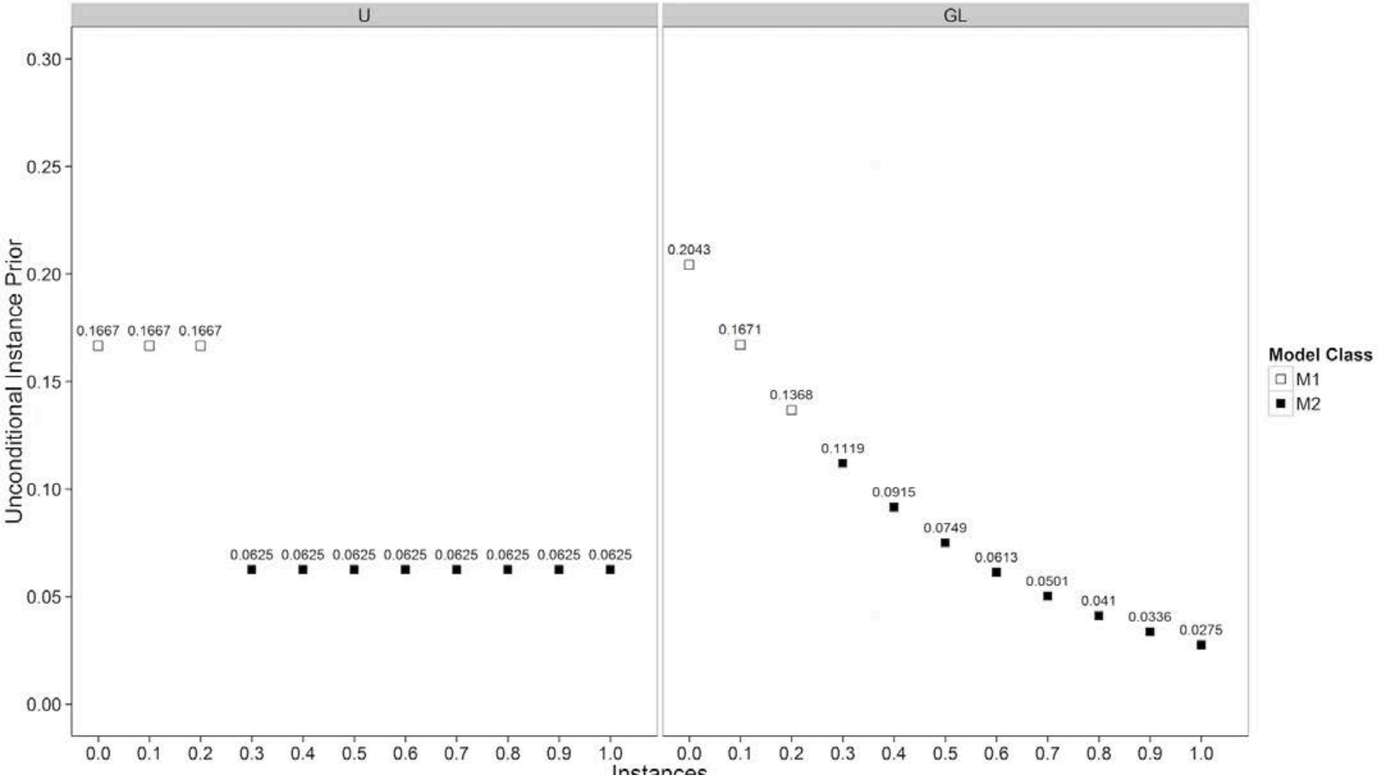


Figure 1: Left panel: unconditional instance prior probabilities when uniform within class, and classes have equal prior probabilities. Right panel: Unconditional instance prior probabilities according to a geometric distribution (classes do not have equal prior probabilities).

$$\frac{p(M_1|y)}{p(M_2|y)} = \left\{\frac{\sum_w p(y|\theta_w, M_1)p_0\,(\theta_w|M_1)}{\sum_w p(y|\theta_w, M_2)p_0\,(\theta_w|M_2)}\right\} \quad * \quad \left\{\frac{p_0(M_1)}{p_0(M_2)}\right\} \qquad (2)$$

Eq. 2 separates the expression into two factors. The ratio in the left hand brackets is termed the Bayes Factor. The numerator and denominator of the Bayes Factor are denoted V_1 and V_2 for convenience in the following. V_1 and V_2 are each based on *conditional* priors that add to 1.0 within class. The ratio in the right hand brackets is the prior ratio for the model classes, which we will term R_0 for short in the following. R_0 is a ratio of *unconditional* summed priors. E.g. $p_0(M_1) = \Sigma_w p_0(\theta_w, M_1)$; $p_0(M_2) = \Sigma_w p_0(\theta_w, M_2)$.

It is traditional and conventional to assess classes with the Bayes Factor alone, either ignoring the ratio of class priors, or treating it as equal to 1.0. Doing this introduces a preference for smaller model classes, a preference sometimes referred to as the Bayesian Occam's Razor. It is easiest to see how this preference arises by considering the case of uniform priors within class. Assume that the class M_1 with N_1 instances is equal in probability to Class M_2 with N_2 instances. Thus the M_1 instances each have unconditional prior probability $(.5/N_1)$ and the M_2 instances each have unconditional prior probability $(.5/N_2)$. The conditional prior probabilities are respectively $1/N_1$, and $1/N_2$. Eq. 2 then becomes:

$$\frac{p(M_1|y)}{p(M_2|y)} = \left\{\frac{\sum_w p(y|\theta_w, M_1)(\frac{1}{N_1})}{\sum_w p(y|\theta_w, M_2)(\frac{1}{N_2})}\right\} = \left\{\left[\frac{N_2}{N_1}\right]\frac{\sum_w p(y|\theta_w, M_1)}{\sum_w p(y|\theta_w, M_2)}\right\} \quad (3)$$

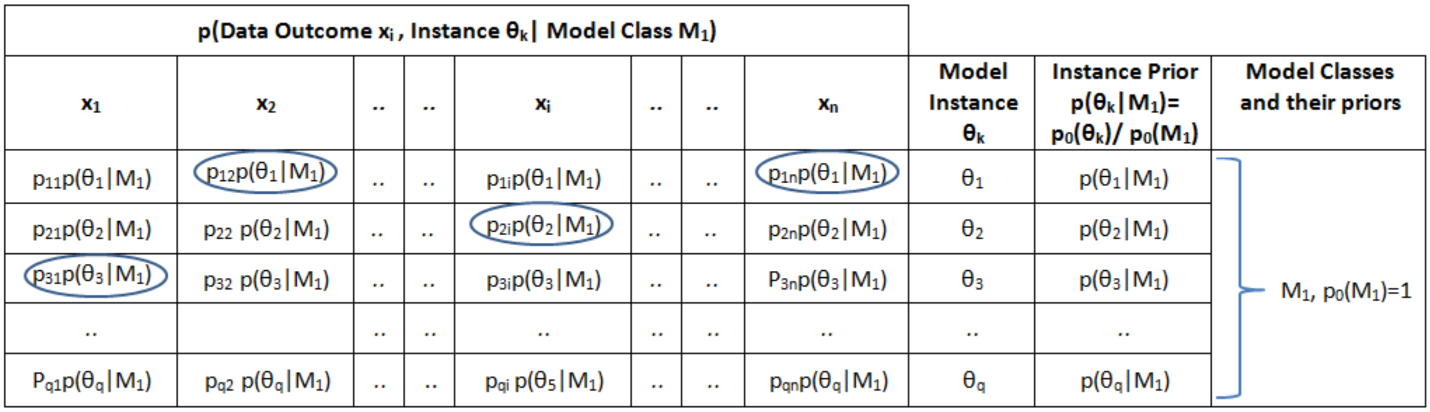
The preference for smaller classes is thus seen in the ratio N_2/N_1, which multiplies the ratio of summed likelihoods within each class. If on the other hand all instances in all classes have the same prior value, then the larger class would have a larger sum, in effect cancelling the preference for smaller classes. Assuming equal instance priors in all classes is equivalent to the use of Bayes factor with flat priors within class and class priors proportional to class size (e.g. Mulder, 2014). It could be argued that class priors proportional to class size is as plausible as equal class priors. In our view both are implausible, being arbitrary conventions. If one has prior knowledge, something almost always the case, then one should use that knowledge to inform all the choices of priors.

Because prior knowledge generally arises in situations differing from the present experimental setting, it is unusual to be able to specify it in anything more precise than rough qualitative terms. Yet it is in most situations necessary to take prior knowledge into account to at least some degree in order to produce valid inference. For example consider the situation illustrated with Tables 1b and 1d: inferring of the value of θ when M_1 includes θ values of 0, .1, and .2, and M_2 includes the remaining eight instances. Uniform priors within class and equal class probabilities entail that the unconditional prior probabilities for θ have a discontinuity at the class boundary: The priors change from 1/6 to 1/16. Prior knowledge may or may not make this reasonable.

In practice, uniform priors within class, coupled with a model class ratio of 1.0, are often assumed (as illustrated in Figure 1), partly to eliminate different views on the values of the priors, partly to facilitate communication of results, and partly to instantiate a preference for simpler model classes. The degree of simplicity preference introduced by the use of uniform priors coupled with equal class probabilities seems to work well in practice in many settings. Sometimes a uniform assumption is maintained for instances within class, but the ratio of class priors is set to some other value than 1. It should be obvious that the choice of value for $p_0(M_1)/p_0(M_2)$ will then directly affect the preference for simplicity. Soon we analyze the binomial example and show that non-uniform prior knowledge for θ requires careful handling.

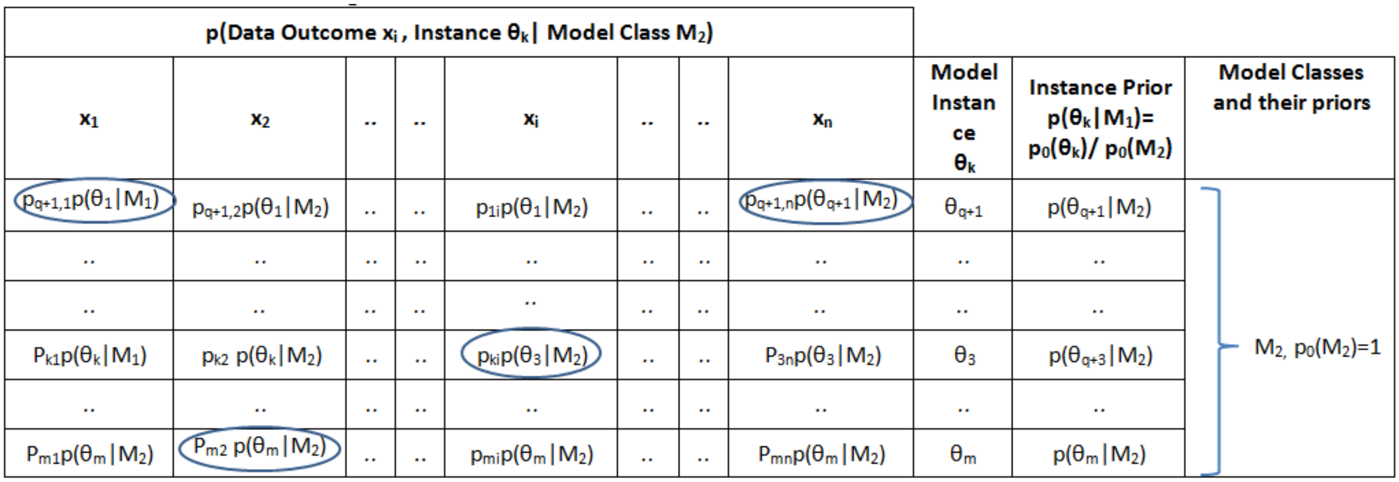
Because the Bayes Factor incorporates a bias for simplicity it is sometimes used to compare classes where one class contains another (e.g. M_1 vs. M_3). Doing so introduces some problems (as we shall see later; also see Mulder, Hoijtink, & Klugkist, 2010). However, we shall argue for a different way to deal with model classes that overlap, an approach that retains a preference for simpler model classes, but in which the shared instances in the smaller class are deleted from the larger class prior to class comparison. In the case of our example, that analysis would be the comparison of M_1 vs M_2.

Table 3a - Conditional Priors of Model Instances in M1 (General Case)



$p(\text{Data Outcome } x_i\text{, Instance } \theta_k \mid \text{Model Class } M_1)$											
x_1	x_2	..	..	x_i	..	..	x_n	Model Instance θ_k	Instance Prior $p(\theta_k \mid M_1)=$ $p_0(\theta_k)/ p_0(M_1)$	Model Classes and their priors
$p_{11}p(\theta_1 \mid M_1)$	$p_{12}p(\theta_1 \mid M_1)$	..	..	$p_{1i}p(\theta_1 \mid M_1)$	..	..	$p_{1n}p(\theta_1 \mid M_1)$	θ_1	$p(\theta_1 \mid M_1)$	M_1, $p_0(M_1)=1$
$p_{21}p(\theta_2 \mid M_1)$	$p_{22}\ p(\theta_2 \mid M_1)$	..	..	$p_{2i}p(\theta_2 \mid M_1)$	..	..	$p_{2n}p(\theta_2 \mid M_1)$	θ_2	$p(\theta_2 \mid M_1)$	
$p_{31}p(\theta_3 \mid M_1)$	$p_{32}\ p(\theta_3 \mid M_1)$	..	..	$p_{3i}p(\theta_3 \mid M_1)$	..	..	$P_{3n}p(\theta_3 \mid M_1)$	θ_3	$p(\theta_3 \mid M_1)$	
..		..	..	..	..	..	..	..	..	
$P_{q1}p(\theta_q \mid M_1)$	$p_{q2}\ p(\theta_q \mid M_1)$	..	..	$p_{qi}\ p(\theta_5 \mid M_1)$	..	..	$p_{qn}p(\theta_q \mid M_1)$	θ_q	$p(\theta_q \mid M_1)$	

Table 3b - Conditional Priors of Model Instances in M2 (General Case)



$p(\text{Data Outcome } x_i\text{, Instance } \theta_k \mid \text{Model Class } M_2)$										
x_1	x_2	..	..	x_i	..	..	x_n	Model Instance θ_k	Instance Prior $p(\theta_k \mid M_1)=$ $p_0(\theta_k)/ p_0(M_2)$	Model Classes and their priors
$p_{q+1,1}p(\theta_1 \mid M_1)$	$p_{q+1,2}p(\theta_1 \mid M_2)$	..	..	$p_{1i}p(\theta_1 \mid M_2)$	..	..	$p_{q+1,n}p(\theta_{q+1} \mid M_2)$	θ_{q+1}	$p(\theta_{q+1} \mid M_2)$	M_2, $p_0(M_2)=1$
..	..	..	..	..	..	..	..	..	..	
..	..	..	..	..	..	..	..	..	..	
$P_{k1}p(\theta_k \mid M_1)$	$p_{k2}\ p(\theta_k \mid M_2)$	..	..	$p_{ki}p(\theta_3 \mid M_2)$	..	..	$P_{3n}p(\theta_3 \mid M_2)$	θ_3	$p(\theta_{q+3} \mid M_2)$	
..	..	..	..	..	..	..	..	..	..	
$P_{m1}p(\theta_m \mid M_2)$	$P_{m2}\ p(\theta_m \mid M_2)$	..	..	$p_{mi}p(\theta_m \mid M_2)$	..	..	$P_{mn}p(\theta_m \mid M_2)$	θ_m	$p(\theta_m \mid M_2)$	

Minimum Description Length and Normalized Maximum Likelihood

A chief aim of this article is the comparison of BMS to another major method for model comparison, one based on the Minimum Description Length principle (MDL; see Grünwald, 2007). At the core of MDL is the idea that one can characterize the regularities in data by a code that compresses the description to an optimal degree. Thus the MDL approach begins with a preference for simplicity. We say no more here about MDL but give a somewhat more detailed overview in Appendix 2.

The examples we use allow comparison of BMS and Bayes Factors to a particular way of implementing MDL that we term NML*. NML* incorporates priors, thereby extending the simplest version of implementing MDL known as NML. When prior probabilities are flat NML* reduces to NML. NML* (and hence also NML) can be described in the context of the tables we have introduced. NML* (as well as

Bayes Factors) starts by considering only one class at a time. A value V* is calculated for each class. This is illustrated with the tables that are specific to each class alone--Table 1c for class M_3, and Tables 3a and 3b for classes M_1 and M_2.

For any of these three tables the NML* score we term V* is calculated as follows: Take the maximum joint probability in the column for the observed data, y, and divide it (i.e. normalize) by the sum of such maxima for all data outcomes. The maximum in each column of each table is circled. The result is V*, the NML* score for a given class. For any two classes (say M_1 and M_3), the ratio of the V* scores (say, V1*/V3*) is interpreted as the analogue of the Bayes Factor.

Non-uniform Priors, Within-Class and Between Class.

The typical application of BMS uses uniform priors within class and equal priors for the two classes. As seen in Eq. 3 this produces a simplicity preference in proportion to the class sizes. When one has prior information that some instances and classes are more probable than others, and decides to incorporate that information into BMS, considerable care is needed in deciding how to represent that information within and between classes, and the decisions made affect the degree of preference for simpler models.

The issues are easiest to present with the help of an example, a version of the binomial inference example we will be analyzing extensively in this article. Suppose one desires to draw inferences about the values of θ operating in N trials. To keep things simple suppose that there are 10 trials in the experiment and that there are two model classes, M_1 positing θ to lie in [0, .5] and M_2 positing θ to lie in (.5, 1].

Let us now suppose that a variety of studies in other paradigms at least vaguely similar to the present one; suppose these prior studies strongly suggest that the value of θ in the present study is more likely to the degree that it is lower. This judgment about θ is of course not precise due to the difference in the prior settings and paradigms. Nonetheless we decide to represent this knowledge as best we can with a prior over instances and classes that starts high at low values of θ and decreases geometrically as θ increases, as illustrated by GL in Figure 1. Tables 4a, 4b, and 4c illustrate the situation; they correspond to 3a, 3b, and 1c, but for 10 trials, within-class priors geometrically favoring small values of θ, and z = .5.

Using a geometric prior ***within*** each model class seems relatively uncontroversial, even if the knowledge is imprecise. Thus in the calculation of the Bayes Factor or the NML* ratio, one can use the geometric prior normalized to add to 1.0 within each class, as shown in Tables 4a, 4b, 4c. But what should one assume about the prior probability of the model classes? Suppose we let the classes have equal probability. Assume now that the experiment produced five successes. It is easy to show that the V_1 and V_2 values that produce the Bayes Factor are .066 and .087, so that the Bayes Factor (V_1/V_2) is 0.76. Combining this with equal class priors produces a preference for M_2: p(M_1) = .432. Similarly the V_1* and V_2* values that are used to calculate the NML* ratios are .45 and .118, so the NML* ratio is 0.38, also expressing a preference for M_2. This result is due of course to the fact that the conditional prior within M_1 is highest at θ = 0, but the conditional prior within M_2 is highest at θ = .6, and the observed data proportion is closer to .6. On the other hand, this preference for M_2 flies in the face of our prior knowledge that low values of θ are more likely. What has happened is that the priors within class reflect our prior knowledge but not the priors between classes.

Table 4a - Conditional Priors of Model Instances in M_1
(Binomial example with N=10, Geometric priors favoring Low theta)

p(Data Outcome x_i , Instance θ_k| Model Class M_1)													
x_j = 0	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Prior $p(\theta_k\|M_1)=p_0(\theta_k)/p_0(M_1)$	Model Classes and their priors
0.2597	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	*0.2598*	M_1, $p_0(M_1)$=1
0.0741	0.0823	0.0412	0.0122	0.0024	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.1	*0.2125*	
0.0187	0.0467	0.0525	0.0350	0.0153	0.0046	0.0010	0.0001	0.0000	0.0000	0.0000	0.2	*0.1739*	
0.0040	0.0172	0.0332	0.0380	0.0285	0.0146	0.0052	0.0013	0.0002	0.0000	0.0000	0.3	*0.1423*	
0.0007	0.0047	0.0141	0.0250	0. 0292	0.0234	0.0130	0.0049	0.0012	0.0002	0.0000	0.4	*0.1163*	
0.0001	0.0009	0.0042	0.0112	0.0195	0.0234	0.0195	0.0112	0.0042	0.0009	0.0001	0.5	*0.0952*	
0.3573	**0.1518**	**0.1452**	**0.1214**	**0.0949**	**0.0663**	**0.0387**	**0.0175**	**0.0056**	**0.0011**	**0.0001**	**← V**		
0.4985	**0.1580**	**0.1008**	**0.0729**	**0.0560**	**0.0449**	**0.0374**	**0.0215**	**0.0081**	**0.0017**	**0.0002**	**← V***		

Table 4b - Conditional Priors of Model Instances in M_2
(Binomial example with N=10, Geometric priors favoring Low theta)

p(Data Outcome x_i , Instance θ_k| Model Class M_2)													
x_j = 0	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Prior $p(\theta_k\|M_2)=p_0(\theta_k)/p_0(M_2)$	Model Classes and their priors
0.0000	0.0005	0.0030	0.0122	0.0320	0.0576	0.0720	0.0617	0.0347	0.0116	0.0017	0.6	0.2870	M_2, $p_0(M_2)$=1
0.0000	0.0000	0.0003	0.0021	0.0086	0.0242	0.0470	0.0627	0.0548	0.0284	0.0066	0.7	0.2347	
0.0000	0.0000	0.0000	0.0002	0.0011	0.0051	0.0169	0.0387	0.0580	0.0516	0.0206	0.8	0.1922	
0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0018	0.0090	0.0305	0.0609	0.0548	0.9	0.1572	
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1286	1.0	0.1286	
0.0000	**0.0005**	**0.0033**	**0.0145**	**0.0417**	**0.0871**	**0.1377**	**0.1723**	**0.1780**	**0.1525**	**0.2123**	**← V**		
0.0000	**0.0010**	**0.0061**	**0.0250**	**0.0656**	**0.1181**	**0.1477**	**0.1286**	**0.1190**	**0.1249**	**0.2638**	**← V***		

Table 4c - Conditional Priors of Model Instances in M_3
(Binomial example with N=10, Geometric priors favoring Low theta)

P_0(Data Outcomes x_i, Instance θ_k)														
$x_j = 0$	1	2	3	4	5	6	7	8	9	10	Model Instance θ_k	Instance Prior $p(\theta_k \mid M_3)= p_0(\theta_k)/ p_0(M_3)$	Model Classes and their priors	
0.2043	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	*0.2043*	$P_0(M_1)$= 0.7865	$P_0(M_3)$ = 1
0.0583	0. 0648	0.0324	0.0096	0.0019	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.1	*0.1671*		
0.0147	0.0367	0.0413	0.0275	0.0120	0.0036	0.0008	0.0001	0.0000	0.0000	0.0000	0.2	*0.1368*		
0.0032	0.0135	0.0261	0.0299	0. 0224	0.0115	0.0041	0.0010	0.0002	0.0000	0.0000	0.3	*0.1119*		
0.0006	0.0037	0.0111	0.0197	0.0230	0.0184	0.0102	0.0039	0.0010	0.0001	0.0000	0.4	*0.0915*		
0.0001	0.0007	0.0033	0.0088	0.0154	0.0184	0. 0154	0.0088	0.0033	0.0007	0.0001	0.5	*0.0749*		
0.0000	0.0001	0.0007	0.0026	0.0068	0.0123	0.0154	0.0132	0.0074	0.0025	0.0004	0.6	*0.0613*	$P_0(M_2)$= 0.2135	
0.0000	0.0000	0.0001	0.0005	0.0018	0.0052	0.0100	0.0134	0.0117	0.0061	0.0014	0.7	*0.0501*		
0.0000	0.0000	0.0000	0.0000	0.0002	0.0011	0.0036	0.0083	0.0124	0.0110	0.0044	0.8	*0.0410*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.004	0.0019	0.0065	0.0130	0. 0117	0.9	*0.0336*		
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0275	1.0	*0.0275*		
0.2810	**0.1195**	**0.1149**	**0.0985**	**0.0835**	**0.0708**	**0.0598**	**0.0505**	**0.0424**	**0.0334**	**0.454**	**← V**			
0.4410	**0.1398**	**0.0892**	**0.0645**	**0.0496**	**0.0398**	**0.0332**	**0.0289**	**0.0267**	**0.0281**	**0.0593**	**← V***			

It is possible to imagine situations in which these priors would make sense-- e.g. two factories making biased coins, each factory favoring low values of θ but for different ranges of θ; a coin is sampled randomly from one of the factories. We consider a case like this more a matter of probabilistic inference than model selection (as discussed in Appendix 4). In a model selection situation like the present scenario it is hard to imagine why there would be a sharp increase in the prior for θ at the class boundary. It seems more sensible to assign class priors simply from the sum of unconditional priors within each class.

Thus in the worked binomial examples to follow we use what can be thought of as two extreme approaches: 1) Equal class priors combined with uniform priors within class, which corresponds to the use to Bayes Factors and NML* ratios, and a strong belief in the equal likelihood of classes, and 2) geometric priors assigned to all instances in all classes, but then class priors assigned simply by adding the unconditional instance priors within each class, corresponding to a strong belief in the priors assigned to instances. More precisely this second approach sets R_0, the class prior ratio in Eq. 2, to the default value $p_0(M_1)/p_0(M_2) = \Sigma_w p_0(\theta_w,M_1)/\Sigma_w p_0(\theta_w,M_2)$. As we will discuss in a later section there are various ways to construe priors with and between classes, and to justify them, and the decisions affect the degree to which smaller classes are preferred.

When assuming GL priors, we have seen by taking the ratios of V* values in Tables 4a, 4b, and 4c that the same issues arise for NML*. If we wish to take into account the prior knowledge that M_1 is much more likely than M_2, then we should multiply the NML* ratio by an appropriate value. To make comparison with BMS arguably more valid, the NML* analyses for the GL prior cases in the next section use the default value, $p_0(M_1)/p_0(M_2)$ as the multiplier.

BMS and NML* for the Binomial Example

We now show model selection results for the binomial example, for the Bayes Factor and NML* (NML * reduces to NML for flat priors, so we shall refer to NML* only in the following). The Bayes Factor and the ratio of NML* scores give the odds for one model class over the other. For comparisons of M_1 against M_2 or M_3 the odds are $p(M_1)/[1-p(M_1)]$ because the probabilities of the two classes must add to 1.0. It helps us show the results to display them as $p(M_1)$ rather than as a ratio, and that convention is used in the following figures. Figures 2 - 7 give the main results. Figures 2-4 give results for uniform priors and Bayes factors or NML* ratios with a class prior ratio of 1.0. Figures 5-7 give results for the geometric prior, GL, when the class prior ratio is set to the default value R_0. It is important to emphasize that the U and GL cases reflect different applications of Eq. 2: The right hand bracket, the class prior ratio, is 1.0 for U and is the default value, R_0, for GL. Each group of three figures shows results for z =.2, z = .5, and z = .8 which specify M_1 and M_2 ($\theta\in[0,z]$, and $\theta\in[z,1]$), in that order. There are two functions shown in each panel of each figure, one for BMS (triangles and lighter line) and one for NML* (diamonds and darker lines). Because we show results for $p(M_1)$ we refer in the following to BMS rather than the Bayes Factor as the approach that generates the triangles and lighter lines.

The left hand side of each figure gives the results for the comparison of M_1 vs. M_2 and the right hand side for M_1 vs. M_3. Each panel in each figure gives $p(M_1)$ corresponding to each potential x_i/N. The results are given for proportion rather than x_i, because the different panels show results for different values of N (shown in the grey bars on the right side and in the upper right of each panel). All the results in the figures are calculated using 1001 instances each 0.001 apart ranging from 0.000 to 1.000. The outcomes are virtually invariant when the number of instances is large in comparison to N and is akin to performing continuous operations. It should be noted that the analyses reported in Heck et al. (2015) are a subset of those we report here: they only analyzed a uniform prior, and what we term M_1 vs. M_3.

There are many results in these figures. Some are quite general: BMS and NML* give quantitatively different results throughout, something unsurprising given the difference in the way the ratios are calculated. For the comparison of M_1 vs. M_2 the two methods are generally or at least arguably in qualitative alignment.

It is generally the case, for comparisons of M_1 vs. M_2 and of M_1 vs. M_3, and for both uniform and geometric priors, that NML* predicts $p(M_1)$ to be below BMS when the data proportion is lesser than a value near z and above BMS when the data proportion is larger than a value near z. To say this another way, if the proportion of observed successes in in the range roughly $[0,z]$ then NML* gives a lower posterior

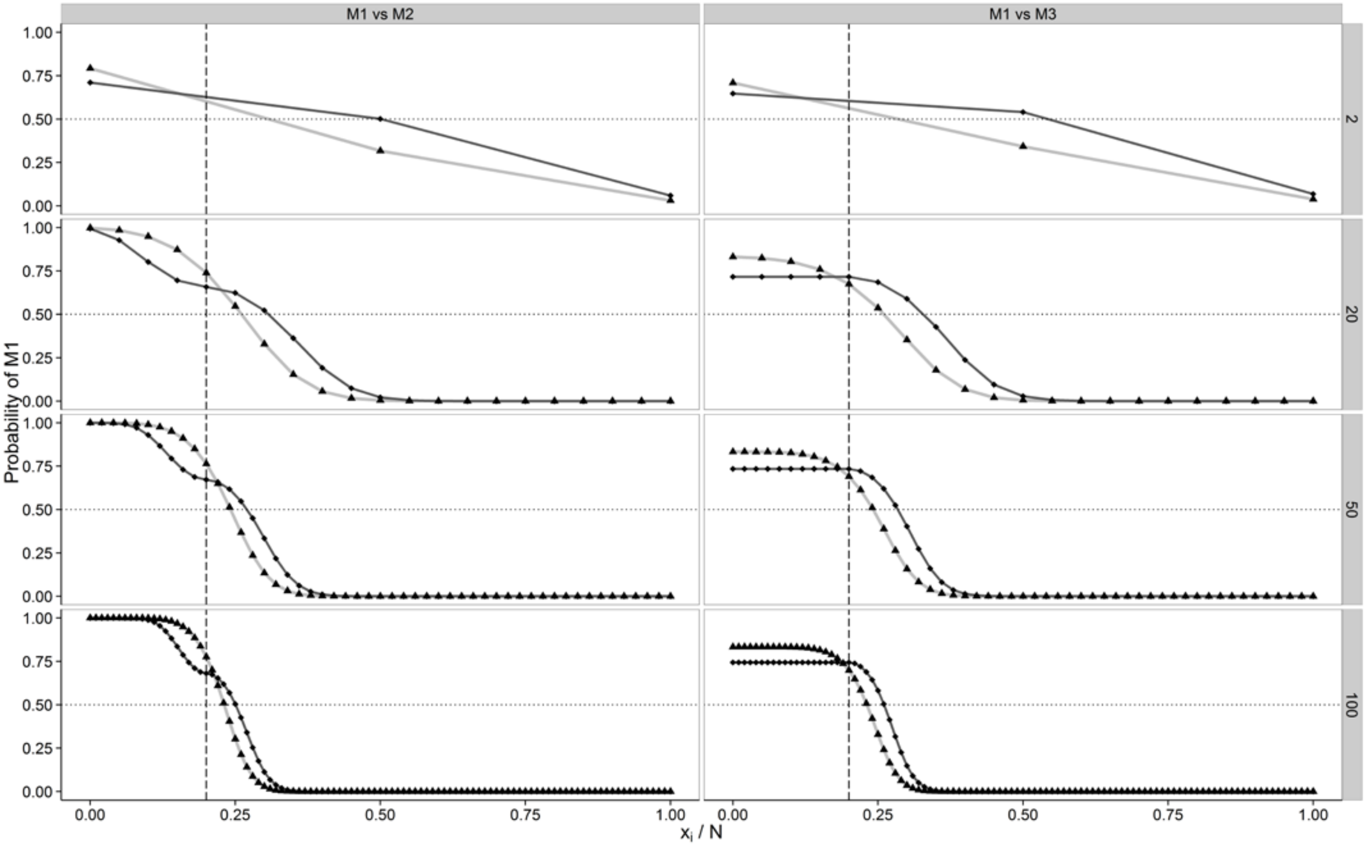


Figure 2: Model class comparison, shown as probability of Class M_1, for uniform within class priors, equal class priors, and z = .2. Triangles show results for BMS, circles for NML*.

probability to M_1 than does BMS; the reverse occurs when the observed proportion of successes is in the range roughly $(z,1]$. Thus when the data is consistent with the predictions of a model class, NML* favors that model class less than BMS. This could be due to the reliance of NML* upon a maximum joint probability, a possibility that could be clarified by the characterization in the next section. Whether this result would hold up for other model classes, including those with greater complexity, is an open question.

For both methods, p(M_1) calculated for the M_1 vs. M_2 comparison asymptotes at 1 for small proportions of successes and at 0 for high proportions of successes. This is as it should be: data most consistent with a model class and least consistent with the other should maximize the probability of the favored class. This result holds only partially for the M_1 vs. M_3 comparison: p(M_1) approaches 0 as the proportion of success rises toward 1. However, the same result does not hold for the M_1 vs. M_3 comparison when the proportion of success drops toward 0: instead p(M_1) stabilizes at an intermediate value for small proportions of success (see Mulder et al, 2010). We view this result as undesirable -- no matter how much data is collected and how few successes are found, one cannot infer with certainty that M_1 is the preferred model. This asymmetric and undesirable result is due to the comparison of model classes when one contains the other. We believe this is one reason to compare model classes that do not share instances, a case that will be made in detail in a later section.

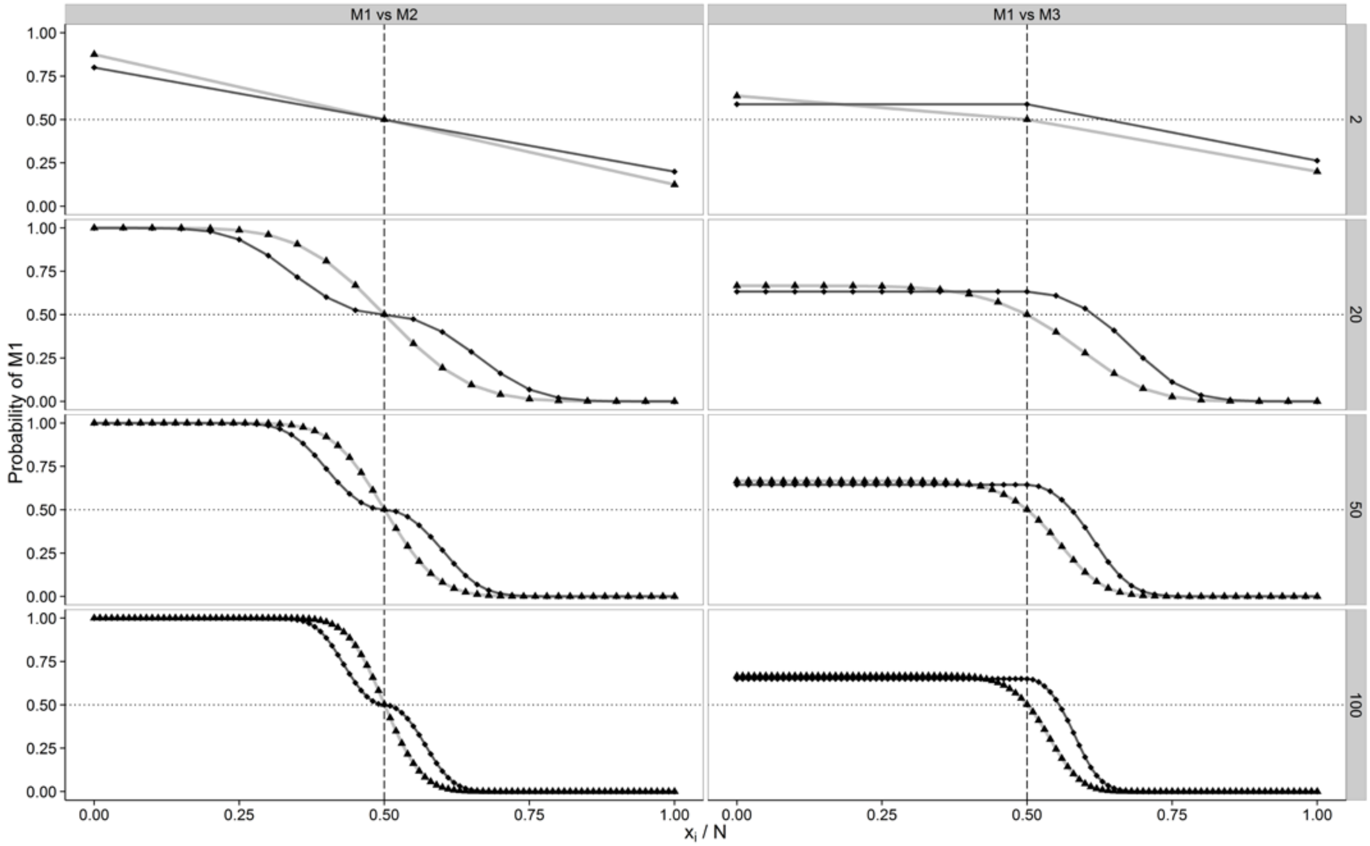


Figure 3: Model class comparison, shown as probability of Class M_1, for uniform within class priors, equal class priors, and z = .5. Triangles show results for BMS, circles for NML*.

The comparison of M_1 vs. M_3 produces a related but more extreme problem that holds only for NML*: When the proportion of successes is less than *z* p(M_1) is constant and does not change as the proportion changes. It is clear why this occurs: NML* for each model class is a max for the observed data divided by a sum of such maxes for all data outcomes. The numerator for each class is the same for each outcome below *z* because the two classes overlap in the range containing the max. The divisor for each class does not depend on the observed outcome. Thus p(M_1) in this range is just the ratio of the divisors (multiplied by R_0 for GL). We noted in the previous paragraph that we would like to favor low values of θ increasingly as the observed proportions of successes drops toward zero, so the NML* results for the comparison of M_1 vs. M_3 are especially troublesome. This problem is not present when comparing M_1 vs. M_2.

It is of interest how BMS and NML differ in the way they make adjustments for complexity. In general this is a complex issue. One way to look at this issue for our example is to focus on observed data for which the maximum probability of the data (maximum across all model instances) is the same for the two model classes being compared. For M_1 vs. M_2 such data lies at the boundary of the two classes:

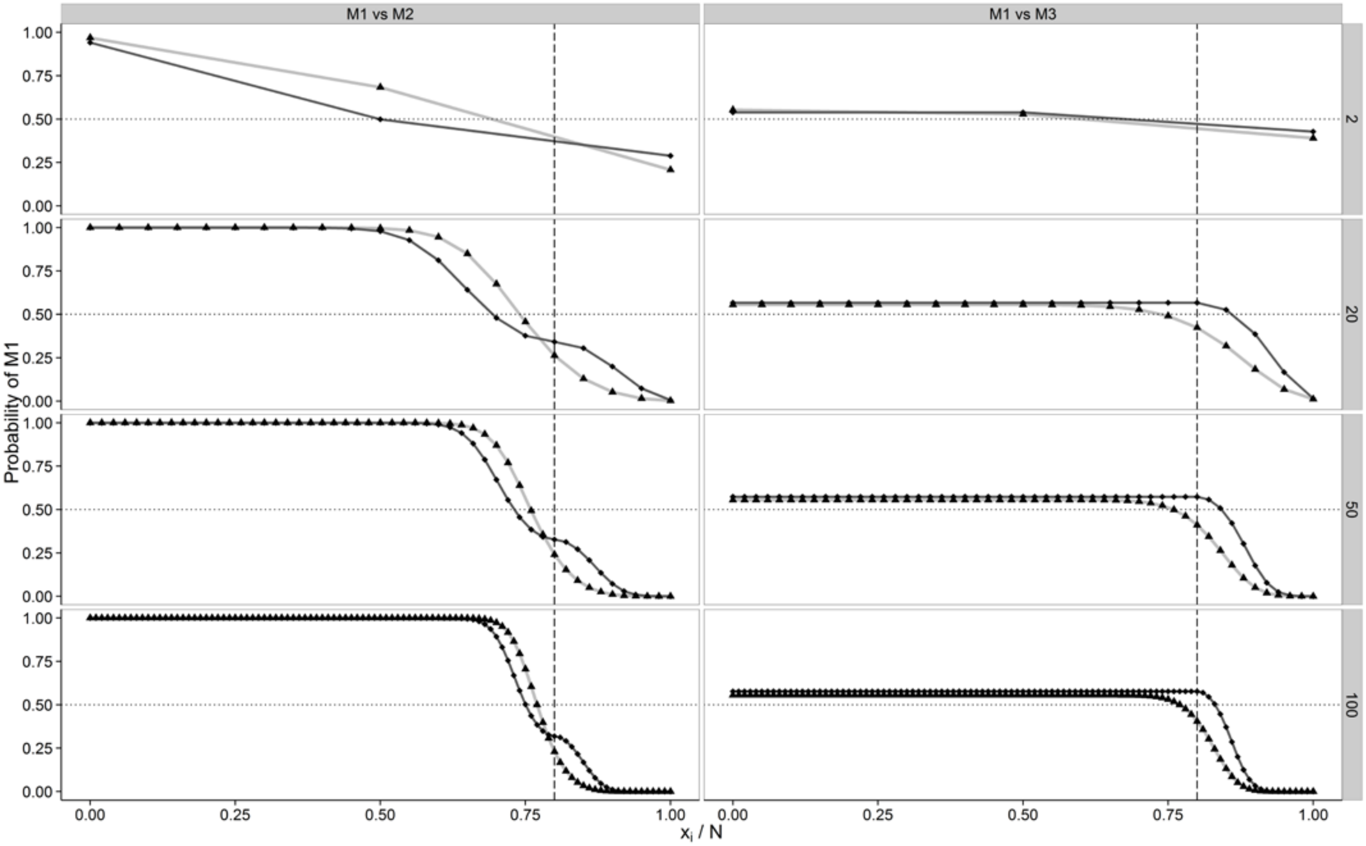


Figure 4: Model class comparison, shown as probability of Class M_1, for uniform within class priors, equal class priors, and z = .8. Triangles show results for BMS, circles for NML*.

proportions of successes .2, .5, and .8 for these same values of z. M_1 and M_3 share the instances in [0,z], so data proportions matching instances in this entire region are given the same maximum likelihood by the two models. Nonetheless it is illustrative to compare BMS and NML* for the same data proportions, .2, .5, and .8, again for the same values of z. These results, given in Figures 2-7 at the vertical dashed lines, are similar for all N. Thus for convenience we table the 36 results for N = 100 in Table 5.

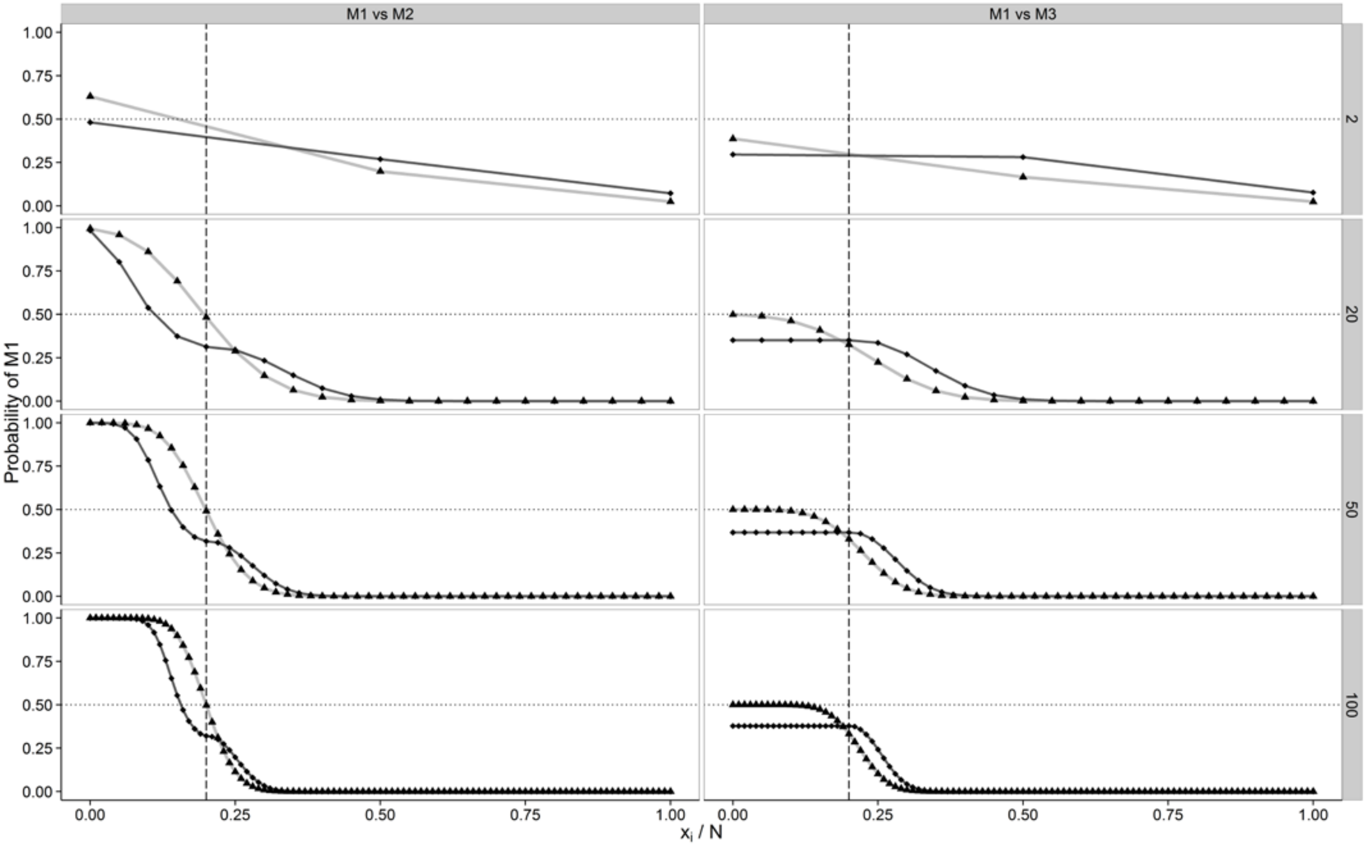


Figure 5: Model class comparison, shown as probability of Class *M*$_1$, for geometric unconditional prior probabilities for all instances. Class prior probabilities are the sum of the instance probabilities in the class. z = .2. Triangles show results for BMS, circles for NML*.

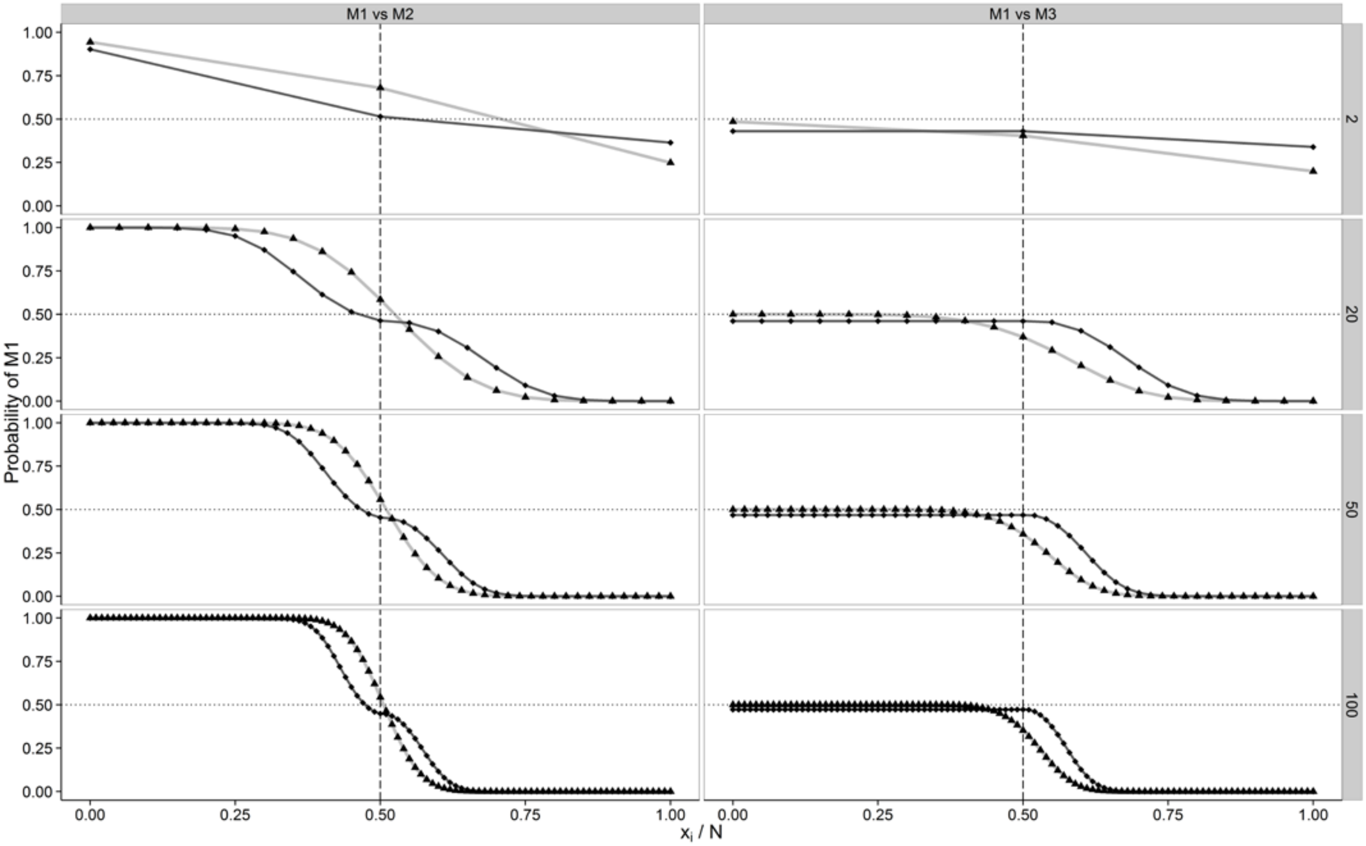


Figure 6: Model class comparison, shown as probability of Class *M_1*, for geometric unconditional prior probabilities for all instances. Class prior probabilities are the sum of the instance probabilities in the class. z = .5. Triangles show results for BMS, circles for NML*.

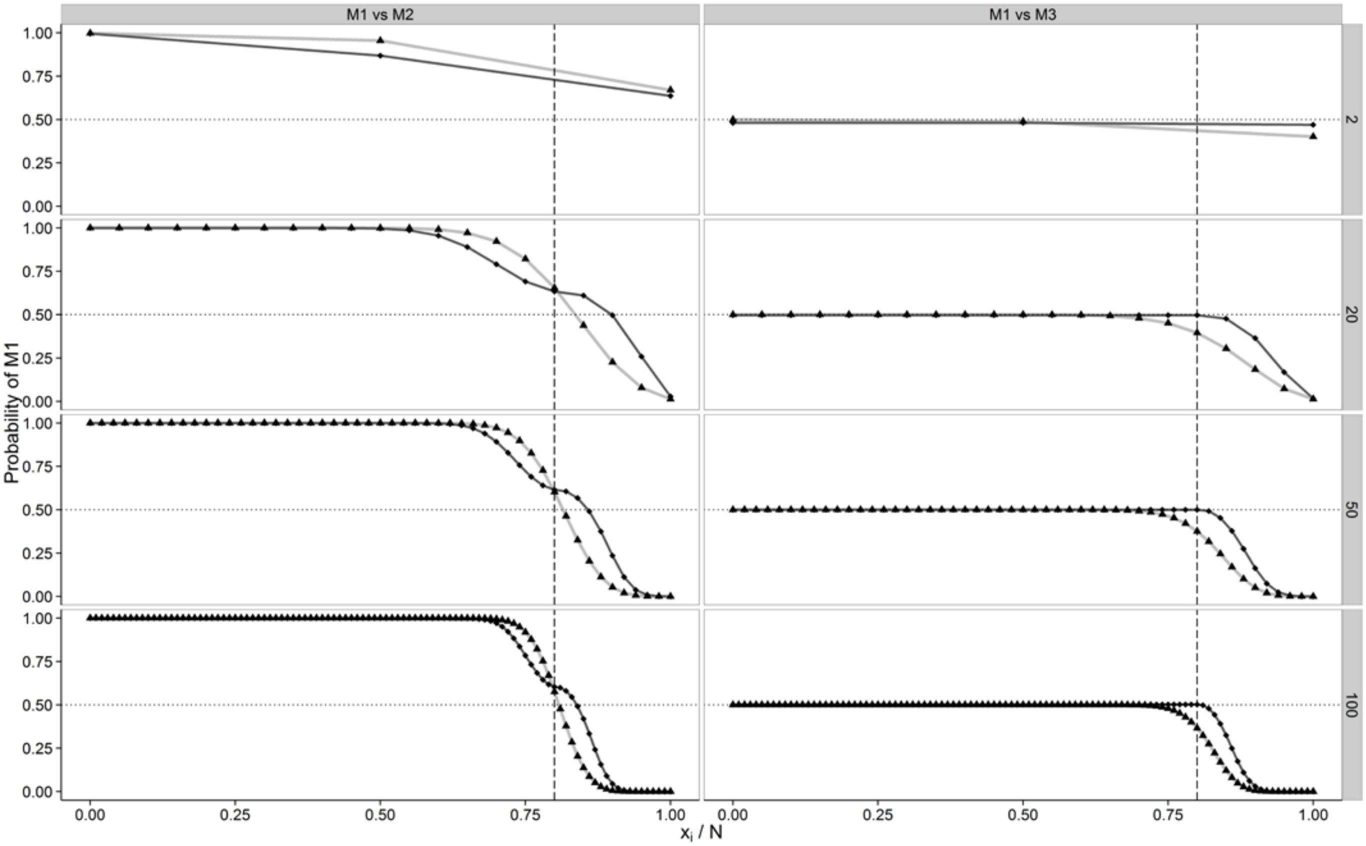
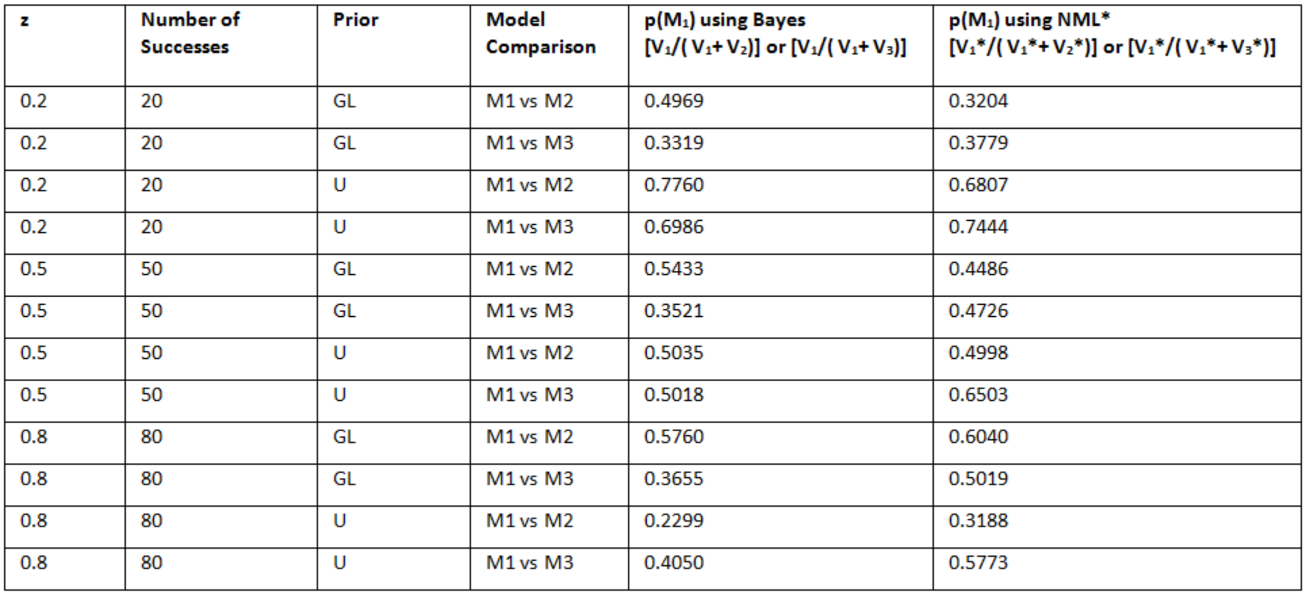


Figure 7: Model class comparison, shown as probability of Class *M*1, for geometric unconditional prior probabilities for all instances. Class prior probabilities are the sum of the instance probabilities in the class. z = .8. Triangles show results for BMS, circles for NML*.

Table 5 Preference for M_1 in Binomial Example at the boundary of M_1 and M_2 (θ=z) for different z-values and different Priors



z	Number of Successes	Prior	Model Comparison	p(M_1) using Bayes [V_1/(V_1+ V_2)] or [V_1/(V_1+ V_3)]	p(M_1) using NML* [V_1*/(V_1*+ V_2*)] or [V_1*/(V_1*+ V_3*)]
0.2	20	GL	M1 vs M2	0.4969	0.3204
0.2	20	GL	M1 vs M3	0.3319	0.3779
0.2	20	U	M1 vs M2	0.7760	0.6807
0.2	20	U	M1 vs M3	0.6986	0.7444
0.5	50	GL	M1 vs M2	0.5433	0.4486
0.5	50	GL	M1 vs M3	0.3521	0.4726
0.5	50	U	M1 vs M2	0.5035	0.4998
0.5	50	U	M1 vs M3	0.5018	0.6503
0.8	80	GL	M1 vs M2	0.5760	0.6040
0.8	80	GL	M1 vs M3	0.3655	0.5019
0.8	80	U	M1 vs M2	0.2299	0.3188
0.8	80	U	M1 vs M3	0.4050	0.5773

BMS results in Figures 2 - 4 assume uniform priors within class and equal priors between class, so incorporate a simplicity preference equal to the ratio of the class sizes. BMS results in Figures 5 - 7 assume geometric priors and do not assume equal class priors and instead let the class prior ratio, R_0, be the default value. The use of the default value removes the simplicity preference. This point is hard to see for geometric priors but is evident for uniform priors: Assume that the N_1 + N_2 instances in both classes all have the same probability. The Bayes Factor and the NML* ratio would not change. However the use of a default value for R_0 amounts to multiplying the Bayes Factor by [N_1/(N_1+N_2)]/[N2/(N_1+N_2)] = N_1/N_2; this factor would cancel the N_2/N_1 simplicity preference shown in Eq. 3. The result is that the p(M_1) then drops accordingly if N_1 is smaller than N_2. E.g. consider the results for N = 100, z = .2, and N_1 = 20 successes. For both BMS and NML* the Table 5 preferences for M_1 drop considerably: For M_1 vs. M_3, p(M_1) for NML* drops from .744 to .522, and for BMS , p(M_1) drops from .699 to .465. For M_1 vs. M_2, , p(M_1) for NML* drops from .681 to .444, and for BMS , p(M_1) drops from .766 to .565. Let us keep this in mind when interpreting the results for the geometric priors.

Returning to the Table 5 results, the findings are clearest for U and the comparison of M_1 vs. M_2: At 50 successes and z= .5, p(M_1) is .5 for both BF and NML*, as it should be, For 20 successes and z = .2, p(M_1) rises above .5 for both BF and NML*, but does so more so for BF--thus in this case the complexity penalty is somewhat larger for BMS. The results for 80 successes and z = .8 are symmetrically reversed, of course, so again the complexity penalty is somewhat higher for BMS.

Even for uniform priors, the results for M_1 vs. M_3 are more complex: BMS and NML* both prefer M_1 for 20 successes and z = .2, slightly more so for NML*, so NML* has a slightly higher complexity penalty. For 50 successes and z = .5, BMS has p(M_1) at .5, but NML* has p(M_1) well above .5, showing a greater complexity penalty for NML*. For 80 successes and z = .8, BMS has p(M_1) below .5 but NML* has p(M_1) higher than .5, again showing a greater complexity penalty for NML*.

The results for the geometric priors should not reflect a simplicity preference for BMS, so the comparisons with NML* (based on a simplicity principle) could in principle be useful: For M_1 vs. M_2: BMS has p(M_1) about .5 for 20 successes; p(M_1) rises a little as the number of successes moves to 50 and z = .5 and then rises slightly more as number of successes rises to 80 and z = .8. Thus these cases represent slightly differing ways likelihoods and number of instances are balancing priors favoring low values of θ. NML* has p(M_1) below .5 for 20 successes and z = .2, slightly below .5 for 50 successes and z =.5, and slightly above .5 for 80 successes and z = .8. It is arguable that these results show a complexity preference relative to BMS, but it would be a mistake to try to draw any general conclusions from this complicated pattern, especially since several factors are being balanced to produce the results.

The results for M_1 vs. M_3 for the geometric case show a general preference for M_3 for both BMS and NML*, but more so for BMS. Many would regard this preference for the larger class as an undesirable inference, and could be used as an argument supporting our suggestion that shared instance be deleted from model class comparisons. The results could also be used as an argument against using the default value for R_0, and imposing an additional multiplier that reflects an additional simplicity preference. We leave this possibility open for future research. Beyond these points we would not want to draw any conclusions about relative complexity penalties for BMS and NML* from the M_1 vs. M_3 comparisons.

Although quantitative comparisons of BMS and NML* are of interest to some researchers, others may prefer to focus on certain qualitative differences, such as cases when BMS predicts p(M_1) to be above .5 and NML* predicts p(M_1) to be below .5 (light dots in Figure 8--call this Case 1), or the reverse (dark triangles in Figure 8--call this Case 2). These results are present in Figures 2 - 7, but more systematically illustrated in Figure 8. The upper panels show results for M_1 vs. M_2 and the lower panels for M_1 vs. M_3. There are six panels in each row: the three on the left give results for prior U, the three on the right for prior GL. Within each group of three the results are for z = .2, .5, and .8 from left to right. In each panel the vertical axis shows the value of N and the horizontal axis the proportion of successes.

The uniform prior is easiest to interpret: For M_1 vs. M_2 the upper panels show symmetric results: At high N the cases occur when the proportion lies near z, but as N drops the cases occur more and more shifted toward .5. This is likely a statistical artifact: E.g. for N = 1, the methods will agree; for N =2, the methods will agree except for a proportion of .5 (one success, one failure), and so forth. Otherwise the results confirm what was seen in Figures 2 - 4, that p(M_1) is higher for NML* than BMS for proportions of successes greater than a value near z and lower for NML* than BMS for proportions below that value of z. The comparison of M_1 vs. M_3 shows the bias seen in Figures 5 - 7: All cases are ones in which NML* favors M_1 and BMS does not. Presumably this reflects the differences in the way BMS and NML* deal with the cases of shared instances.

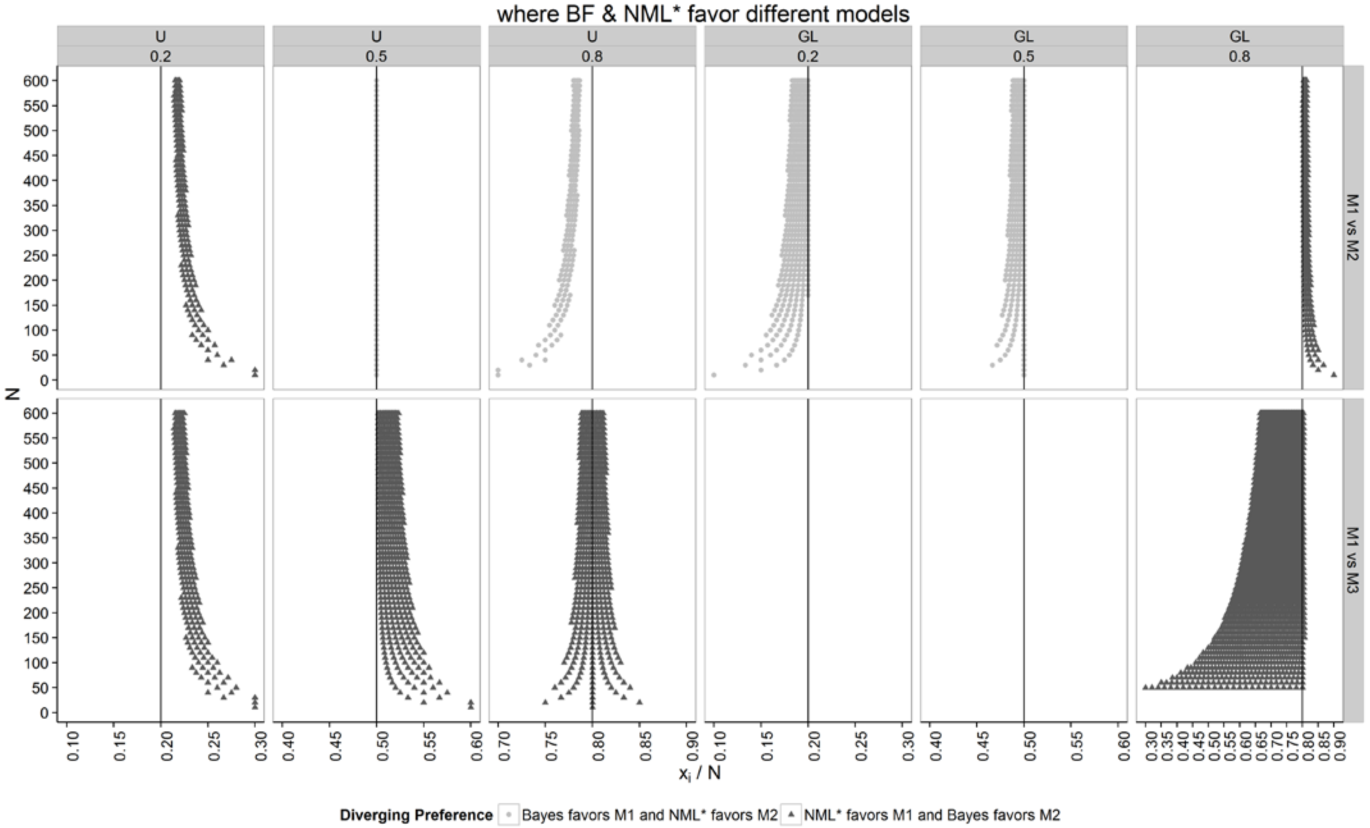


Figure 8: Data proportions for different numbers of total observations for which BMS and NML* prefer different models. Left panel: Uniform priors, z = .2, .5, .8 from left to right; right panel: Geometric priors, z = .2, .5, .8 from left to right. Top panels: M_1 vs. M_2; lower panels: M_1 vs. M_3. Lighter symbols indicate cases in which BMS prefers M_1 and NML* prefers M_2; the reverse holds for darker symbols.

The GL findings are complex: Cases of disagreement for M_1 vs. M_2 for z = .2 and z = .5 all are ones showing p(M_1) to be higher for BMS than NML*, but then these reverse for z = .8. For the comparisons of M_1 vs. M_3 ithe only cases of disagreement occur for z = .8 and all show p(M_1) to be higher for NML* than BMS. These results can be explained by detailed quantitative examination of the two methods, but do not seem to provide any general rules that are likely to extend to other model selection scenarios.

Our general conclusion from these analyses is that it is probably best to delete shared instances from model class comparisons, and then BMS and NML* *differ quantitatively* but generally give similar results *qualitatively*. Insights into why this might be the case are taken up next.

Relation of BMS to NML*

The analyses of the binomial examples demonstrate qualitative similarity of BMS and NML* for all comparisons of M_1 vs. M_2. There is an alternative but equivalent characterization of the Bayes Factor that helps explain why this might be the case. This characterization is illustrated in Tables 3a, 3b, and 1c. These give one table for each model class, each table based on prior probabilities that add to 1.0 -- 3a and 3b are just 1c truncated, with normalized priors. The Bayes Factor is the ratio of the V values for each model class (i.e. table) being compared. The V value for a given table is just the column sum for the observed data. However the sum of such column sums for all columns (i.e. the entire table) is 1.0, because the priors are conditioned on the model class. This means we can equivalently state the Bayes Factor as a ratio of V values, where each V value is the ***sum*** of conditional joint probabilities for the observed data divided by the ***sum of such sums*** for all possible data outcomes (the divisor equaling 1.0). The NML* criterion is similar, except we replace V by V*, where each V* value is a ***maximum*** conditional joint probability for the observed data divided by the ***sum of such maxima*** for all possible data outcomes.[4]

Footnote 4: The ratio of V terms is identical whether one uses conditional or unconditional joint probabilities: because one is just a constant time the other.*

This correspondence is even more general. Each term in the Bayes Factor is a sum divided by a sum of sums: One can divide the numerator and denominator of this term by any constant without changing its value. If the constant is the number of instances in the class, then the Bayes Factor can be stated equivalently as a ratio of normalized averages (i.e. normalized means).

When classes do not overlap, these characterizations of the Bayes Factor and NML* provide insights into the reasons one might expect qualitatively similar but quantitatively differing outcomes: Quantitative difference would be expected because identical results would require that every distribution in the table is symmetric (the mean and maximum align): This case would be most unlikely. On the other hand, although the max of a distribution and the mean (or sum) will often differ by a large amount, and although such differences are likely to vary across the different distributions in Table 5, the normalization by a sum of like quantities tends to bring the two expressions into rough alignment. We have not tried to determine general rules governing the degree of alignment because we suspect finding such rules would be a difficult enterprise best left for future research.

When classes do overlap, our tables also provide insight into the qualitative differences that then appear when comparing the Bayes factor to NML*. The derivation of BMS with the results shown in Eq. 2 is based on the assumption that shared instances are given the *same prior probability* in each class they inhabit. As a result Eq. 2 cannot be used to compare models that share instances--e.g. in cases where one class contains another (viz. M_1 vs. M_3) Eq. 2 would guarantee a higher score for the larger model.[5] However, using only the Bayes Factor (implicitly or explicitly dropping a direct linkage to the class priors) means one calculates the numerator and denominator for each class separately, and instances common to both classes will not have the same priors. E.g. in our example M_3 contains M_1; if their sizes are defined in terms of number of instances as N_3 and N_1 and flat priors are assumed, the instances in M_1 would have higher prior probabilities than those same instances in M_3 by a ratio N_3/N_1. This approach works well in most cases because larger classes are then penalized for having more instances, even when those instances are shared. Of course one must then let the ratio of class priors be specified independently, and/or ignored, and there are various justifications for doing this (see Mulder et al., 2010, for a closely related discussion).

Footnote 5: Similar problems arise whenever classes share instances, although we omit details.

Our example with flat priors makes it easy to see that a similar fix does not work well for NML*. Figures 5, 6, and 7 show that all observed data (number of successes) having a maximum joint probability in [0, z] produces the same NML* (or NML) score. This result is for most observers a qualitatively poor inference, and occurs despite NML* being calculated for each class separately. Thus cases with shared instances seem better handled by Bayes Factors than NML*. In the next section we suggest elimination of shared instances before model class comparison--our examples comparing M_1 vs. M_3 show that doing so removes this qualitative advantage for Bayes Factors.[6]

Footnote 6: We leave to future research the possibility that MDL/NML can be brought into better alignment with the Bayes Factor for overlapping classes by allowing different priors for instances in the two classes, or making other sorts of additional assumptions. This is likely to be highly technical issue.

An Alternative Approach to Model Comparison when Classes Share Instances

In our view the simpler and more coherent approach to model comparison in most cases when classes overlap is the one we have utilized in our binomial example, one that entails removal of shared instances from the model classes being compared. This is most easily and sensibly accomplished by subtracting the shared instances from the larger class prior to comparison. This approach can be justified for both BMS and NML* and the fact that it produced qualitatively similar results for BMS and NML* for the binomial example is possibly one argument for its adoption.

Using another version of our worked example, we can give another reason for preferring a model comparison without shared instances, one predicated upon deleting shared instances from the larger class. Consider the common case where we are interested in testing a model class with θ in [0,1] against a model class with θ in [.49,.51]. Call this Class Comparison I. In essence we are asking if θ is at chance, and there is arguably a conceptual difference between that question and simply calculating a posterior for θ. Contrast this situation with one where we compare model class θ in [0,1] with a class positing θ in $[0,.49] \cup [.51,1]$. The latter class contains all θ except those in [.49,.51]. Call this Class Comparison II. Class Comparisons I and II seem to be asking the same question: in both comparisons the only difference is the region [.49,.51]. However, the complexity penalty imposed by traditional BMS and the Bayes Factor varies considerably depending on which comparison is carried out. Some might argue that these comparisons are different, so different complexity penalties ought to apply. We suggest that such a position mistakes inference tools for the scientific and statistical goals of comparison-- goals ought to take precedence. Thus we believe the correct and more sensible comparison in both cases is θ in [.49,.51] vs. $\theta \in [0,.49] \cup [.51,1]$. This comparison is one derived from the other cases by subtracting the smaller model from the larger. Once we do this, then the class comparison is the same as posterior inference, in the sense that the class posterior can be viewed as the sum of instance posteriors, and we still maintain a complexity penalty for the resultant larger class.

At first glance it appears that our recommended approach is at odds with the oldest and most common form of model comparison: Comparing a model class with a given set of unspecified parameters to a restricted class in which some of those parameters are given specified values (Klugkist, Kato, & Hoijtink, 2005; Barlow, Bartholomew, Bremner, & Brunk, 1972). However this inconsistency in almost all cases is an illusion: Because the larger class is usually larger by order(s) of magnitude, the BMS or NML* score for the larger class will not be altered by subtracting the insignificant proportion of shared instances. The use of the term 'larger' here is a somewhat loose stand-in for a more technical description stating that for the smaller class the sum of the joint probabilities of shared instances and possible data outcomes (either conditional or unconditional joint probabilities) is small in relation to the sum for all instances in the larger class. Thus when the shared instances are order(s) of magnitude less than the non-shared instances, ***one can delete from the larger class all shared instances, and let the smaller class have all the shared instances***. We think it clear that class comparisons will not be altered significantly by this change in characterization of the larger class. This may not be entirely obvious in the context of our examples because for simplicity we discretize instances into just 11 cases--if we restricted consideration to a single instance for a restricted model, we would be deleting one of 11 cases and that could have a noticeable effect on inferences concerning the larger class. However, if the discretization were more reasonably based on hundreds or thousands of instances then deleting one instance from many neighbor instances that all predict approximately the same probability of the observed data (or any possible observed data) could hardly produce a noticeable change in inferences concerning the larger class. Thus after deletion of shared instances from the larger class, the resultant class comparisons can therefore be handled by the analyses we have been proposing (i.e. the expression in Eq. 2).

This argument is couched in discrete terms but holds generally. For example, consider the prototypical, standard case where priors over the larger model are represented as densities over a parameter space and the embedded smaller models have prior measure zero. They can thus be removed from the larger model without having any effect on the instance or class posteriors, since posterior densities are only defined up to sets of measure zero.

The observation that deletion of shared instance leaves inference unchanged of course holds only when the difference in size of the classes is substantial. In the M_1 vs M_3 comparisons we consider in this article, [0,*z*] is a significant fraction of [0,1]--we have seen that subtraction of the shared instances from the larger class changes the model selection results. We have argued that those changes are desirable and reasonable. To repeat just one reason, take the case where z =.5, and suppose that the two classes are a priori equally probable, and within each class no θ value is considered more probable than any other. Suppose that the observed data is consistent with both model classes, in the sense that the proportion of observed successes in N trials is less than .5. Most observers would want the smaller class to gain support, do so especially for small proportions of successes, and do so more for a fixed proportion as N increases. However, the Bayes Factor has a penalty for complexity due to the distribution of priors across a double size interval, and for z = .5, the penalty is 2:1. As a result, for small proportions of successes the BF for the model class [0,.5] never rises above 2/1 as N increases. A case can be made that one should compare classes for which a large enough data collection will (or at least may) produce a definitive answer (in statistical terms, one wants to only consider *identifiable* properties of the phenomenon one is modeling). Reconsidering the purpose of this comparison, a case can be made that one should be asking instead whether θ is below or above .5. If one compares [0,.5] to [.5,1] then everything works as it should. For example the ratio favoring [0,.5] rises without bound as N increases, as we showed in the worked examples. It should be noted that this concern applies equally to NML*. As the worked examples show, the NML* ratio for the shared instances case also asymptotes at a small ratio as N increases, one only slightly higher than 2:1, but rises without bound for the comparison without shared instances.

One can take a slightly more technical MDL perspective (see Appendix 2): An underlying goal of MDL is the attempt to find a good code for the data, where one codes data by first coding the M_j Model

Class index j (this takes 1 bit for two classes) and then coding the data given M_j (this takes - log[$P_{NML*}(x^n)$] bits). Now suppose one model class contains another. In our example, M_3 contains M_1, and the codelength for the data given M_3 is shorter no matter what is the data, an obviously undesirable result. Thus satisfying the underlying goal of MDL (data compression, not truth finding) is better accomplished by removing the overlap and replacing M_3 with M_2.

From a more general perspective, we believe that there should usually be a qualitative rather than quantitative underlying conceptual basis for class comparisons. We have noted that it is typical to compare parameterized models to 'smaller' models defined by restrictions of some of the parameters, and have suggested such cases are easy to handle by subtracting the lower dimensional model instances from the larger. Ease of handling aside, when one compares a restricted parameter class to a more complete model, one typically has in mind a somewhat independent role for the parameter(s) that are restricted, so it makes sense to ask whether it (or they) play a significant role. When we compare M_1 to M_3 this is not the case. We might be asking if θ is less than chance or greater than chance, or we might be asking whether it is either of those or in a region close to chance. But these classes would not overlap. Alternatively we could want to know the value of θ, in which case a posterior inference about θ would be appropriate, not class comparison. Even when one tests a series of embedded descriptive models for which the parameters have no particular interpretation (e.g. what degree polynomial best fits some data set), that comparison involves models with different dimensionality, and therefore can be handled without shared instances, as we already explained..

Overview of the Proposal to Delete Shared Instances

Deleting shared instances from the larger class prior to model comparison is non-traditional but not as radical as it might appear. In most cases, such as comparisons of a class to a smaller one with certain parameter values specified, the larger class is enough larger that the deletion does not alter the model comparison. In the few cases where the classes are reasonably close in size we think a good case can be made that deletion produces a comparison that better conforms to the goals of inference. Thus a model class with a parameter ranging in (a,2a) might be compared to one with that parameter ranging in (a,3a); with uniform priors the Bayes factor favoring (a,2a) over (a,3a) would be twice that favoring (a,2a) over (2a,3a). Although there are reasons to prefer simpler models this difference in the Bayes factor is of modest size. More important, it is not clear what question is being addressed when the two model classes are the same except for a doubling of the range of one parameter: There is no qualitative difference in these classes. It seems rather that the question being addressed is one of two possibilities: 1) parameter estimation--where does this parameter lie? This is a question best answered by the Bayesian posterior predicted distribution; 2) does the parameter lie above or below a critical value (in this example, 2a)? This is a question best answered by deleting shared instances prior to comparison, and comparing (a,2a) to (2a,3a). Another consideration arises when the size or parameter range of the smaller class approaches that of the larger class. The traditional method does not distinguish the two classes, but deletion before comparison answers the question the researcher is probably trying to ask: Does the data support the hypothesis that the model instances lie in the region of difference between the two classes. Finally, and to some researchers the an important consideration, deleting shared instances prior to comparison brings BMS and MDL into qualitative alignment.

Situations Justifying Retention of Shared Instances

There are settings where it is natural to use the probabilistic rules of Bayes theorem to discriminate model classes that have the same instances but differ in the priors assigned to them. We regard these situations as ones better characterized as probabilistic inference rather than model selection. These situations are discussed in Appendix 4.

Proper Management of Priors

For both BMS and MDL, and any other potential basis for inference and induction, incorporating prior knowledge is an important consideration, particularly when highly relevant knowledge exists. When an apparently well-designed study claims a demonstration of ESP, we doubt the conclusion not just because a theoretical justification is absent but because the prior probability is low. In most scientific model selection problems such relevant knowledge does exist. It may be vague and hard to quantify, but nonetheless will usually play an important role in inference. Most researchers are aware this is the case through experience with their new graduate students: A student reports 'strange' results from a new study, results claimed to be accurate because they have been checked for errors in the program and analysis. The scientist nonetheless 'knows' the results are almost certainly in error, and instructs the student to recheck with greater care. Subsequent investigation almost always reveals the student has made some sort of error. Even though the study may be new, the scientist's prior knowledge allows generally accurate assessment of the likelihood of error.

There are several reasons that have led to the general use of a default procedure by which priors are assumed uniform, uninformative or uninfluenced by changes of scale within class, and class priors are set equal (e.g. Bartlema et al., 2014). Some reasons involve the need to 'standardize' inference in publications and scientific communication; others involve the difficulty of quantifying prior knowledge. One we believe especially important is the desirability of insuring a preference for simpler (in our discretization approach smaller) model classes, a principal that is in effect the starting point for the MDL approach. One might even want to justify the use of the use of Bayes Factor with class priors of 1.0 on the basis that NML* is calculated in qualitatively similar fashion and is better justified (see Grünwald, 2007, page 543).

However it should be kept in mind that any prior, for example uniform, does specify a particular belief about the model instances and classes, and is certainly not neutral (for example, Jeffreys' prior for the Bernoulli model says that a priori, a strongly biased coin is more likely than a fair one). While using a conventionally chosen form of prior is sometimes useful it is not a substitute for actual knowledge (in the MDL approach, the 'luckiness' interpretation discussed in Chapter 17 of Grünwald, 2007, helps to understand why one often gets away with it). The importance of priors has been voiced many times (most recently by Wolf Vanpaemel at the 2014 Mathematical Psychology Meetings- Bartlema, Lee, Wetzels, & Vanpaemel, 2014). In general we believe it better to incorporate one's knowledge than ignore it, especially when ignoring prior knowledge amounts to employing a prior that might well be more unlikely than numerous alternatives. In addition, when knowledge is vague it might cause one to choose priors that do not strongly favor particular classes or instances, but that is not the same as pretending one has no knowledge. These remarks ought to apply to both BMS and MDL.

It has often been noted that the model classes and instances in any actual application are always wrong, but useful (e.g. Box, 1976). The utility arises in part because the classes and instances provide a good approximation to the state of the world that produced the distribution from which the observed data is sampled. However we always have a good deal of knowledge about the likely form of the true distribution, even when we are less sure about the relative likelihoods of the model instances (and their associated distributions). For example, if we are inferring something about θ (as in our examples), and measuring the number of successes, n, in 10 trials, the possible true distributions of n are not equally probable: We would not expect a distribution for n that would assign high probabilities to prime numbers (1,2,3,5,7) and low probabilities to other n's (0,4,6,8,9,10).

However one ends up assigning priors, it should always be done consistently. For example within a given model class two instances that predict the same distribution of data outcomes are the same, and need to have the same prior. We have proposed that class comparisons in most cases be carried out for classes that do not share instances, but what makes two instances the same should be determined by the similarity of the distributions they predict, and not the nominal functional form. E.g. two model classes are same if they have parameters x and y related by y = f(x), so that every distribution predicted by y is predicted identically by f(x). In accord with our suggestion, these instances should be considered shared, eliminated from class comparisons, and hence should not be assigned different priors.

Note that MDL, in the various forms proposed by Grünwald (2007), not only allows prior knowledge to be incorporated in the method, but allows various other factors such as outcome utilities to play a role as well. BMS incorporates prior knowledge, uses these to calculate posteriors and then starts with those posteriors as a basis for incorporating other selection factors such as utilities. We have adopted this latter approach in specifying NML*, incorporating only the priors in the specification, but presumably allowing other factors to play a role thereafter.

The argument that prior knowledge ought to be incorporated in inference does leave an important unanswered question: Once one uses non-uniform priors, what multiplier should be used for the Bayes Factor and for the NML* ratio? That ratio is one for the usual BMS procedure when model classes are assumed equally likely, and such a ratio implicitly engenders a higher unconditional prior for instances in the smaller class, and hence a simplicity preference. But for many or most cases with non-uniform priors, like the GL prior used in our example, the class priors ought not to be 1.0: Prior knowledge will tell us one class is more likely than another. In our example we used a default value for the multiplier, based on the sum of unconditional instance priors within the class. However this is an extreme view that places a good deal of trust in the validity of the prior assigned to instances. In both our example and in most model selection situations the prior is usually imprecise, even when estimated with some confidence qualitatively. Thus one might not want to assign class priors by summing these imprecise instance priors.

Perhaps most troubling is the rationale for switching approaches (as we have done in the examples in this article): a) When 'ignoring' prior knowledge one assumes equal class priors and uniform conditional priors within class. b) When incorporating prior knowledge one represents such knowledge by setting non-uniform unconditional priors for instances and assigns priors to classes by summing the unconditional priors for the class instances. In a) there is a simplicity preference in proportion to the class sizes (see Eq. 3) and in b) there is no additionally imposed simplicity preference. However, consider our example with a geometric prior, GL: Suppose the parameter of the geometric is gradually altered so the priors across instances become increasingly uniform: At what point would it be appropriate to switch from approach b) to a)? One possible way to deal with this issue would have the following steps, to be used in all settings: 1) Assign unconditional instance priors based on prior knowledge. 2) Assign default class priors, R_0, by summing the unconditional instance priors in each class. 3) Calculate Eq. 2 using both factors. 4) Multiply the result in 3) by another factor meant to impose a simplicity preference. Perhaps this last factor could be the same in all settings. One choice for such a factor would be N_2/N_1 since that is the one implicit in the standard BMS approach. For example in the case of the geometric prior in our examples, one might multiply the result from Eq. 2 by N_2/N_1. The idea is to impose a simplicity preference to the same degree in all settings. This suggestion would have to be explored in future research. In any event these examples highlight both the importance and the difficulty of selecting appropriate within-class and between class priors.

Priors, Posteriors, Complexity, and Precision of Observed Data

Subtle issues arise due to the power and precision of measurement. Two model instances (from say two different model classes) might have different functional forms, but predict very similar distributions of outcomes. We are recommending that shared instances be removed from model class comparisons, so it is essential that we define in consistent fashion what makes instances the same. This issue is discussed in Appendix 5.

Goals of Inference

The methods we have described in this article are considered present state-of-the-art not because they simultaneously satisfy every goal of inference, something probably impossible for any method of model selection. They do represent a good compromise that emphasizes a balance of good fit to data with simplicity. When one has limited and noisy data, an overly complex model class will have instances that appear highly likely because they fit not only the underlying true generating processes but also the noise in the data. A new data set will have a new sample of noise and hence prediction will suffer. Thus the present methods in emphasizing simpler model classes are also designed well to choose models that predict well for exact replications. There are many issues concerning ability to predict that go beyond the scope of this article (see Romeijn et al, 2012 for some discussion of these). We note only that our emphasis upon the one-to-one relation of model instances to data distributions does imply a certain reliance upon prediction: It is the data that we want to predict in the future. Of course we use models to make the prediction, and one thereby seems to put a good deal of trust in the validity of the models in question. An alternative approach we have advocated extends BMS (and NML*) by inferring the probability that model instances and classes provide the best approximation to the unknown but true generating distribution (see Shiffrin & Chandramouli, in press; Chandramouli & Shiffrin, in press, b, this issue); there is a possibility that this approach could be more robust in generalizing to similar but different paradigms.

There are of course many other goals of model selection, some of which involve utilities associated with choices based on the results of inference, and some methods are justified by maximizing utilities. This is a highly technical subject not appropriate for this article, but one extensive treatment and analysis can be found in Grünwald (2007); there is a large literature on this subject that we cannot mention. As we have said, it is possible to take the results of probabilistic inference in the form of posteriors, and use the instance and class probabilities as part of a subsequent attempt to maximize utility.

Although we use quantitative methods to select among model classes, we usually try to distinguish model classes that differ in some qualitative fashion. Some model classes that differ in parameter range restrictions should be considered qualitatively different: An example occurs when the restricted model has a parameter with a specified value. Another example occurs when the classes differ in the value being above or below some specified value. Both are easily handled by subtracting shared instance from the larger class. It is of course possible to specify model classes that differ only in the range of some parameter, both ranges being significantly large. We considered such cases in the examples analyzed earlier in the paper. It is hard to imagine scenarios in which such model classes would be considered enough different qualitatively to make model selection appropriate. To us it seems that such cases are better analyzed to infer the likely value of the parameter in question.

Summary

This article explains in non-technical manner two chief methods for deciding which model classes better explain observed data. A model class is a collection of model instances. An instance is characterized by the distribution of experimental data outcomes it predicts--there is a one-to-one mapping between instances and associated distributions. We show such distributions in the form of tables and explain the model selection methods by simple arithmetic applied to the entries in the tables.

The two chief methods for model selection we consider are Bayesian Model Selection (BMS) and applications of the Minimum Description Length Principle (MDL). MDL is usually approximated by Normalized Maximum Likelihood (NML), but NML is too simple to produce general methods for model selection, in part because it has no role for prior knowledge. We therefore present and analyze a variant of NML termed NML* that incorporates prior knowledge. Prior knowledge in both BMS and NML* is represented by prior probabilities assigned to instances (and hence the associated distributions) prior to analysis of the observed data. Any method to choose among models must balance the ability of the instances in a model class to fit the observed data, the instance likelihoods, against the ability of those instances to fit any possible data outcome, the model class complexity. MDL and NML* methods are rooted in the idea that a model class is preferred if it explains the data simply. BMS uses probability in the form of Bayes Theorem and the preference for simplicity arises from different prior probabilities assigned to instances in different model classes. Such differing priors arise when Bayes Factors are used to compare models. The Bayes Factor is a ratio of terms, one for each model. Each term is a sum over the class instances of the product of likelihood and prior. The priors in a class sum to one so the larger class has on average smaller priors.

BMS and NML* appear quite different at first glance, but there is an underlying similarity: Each term in the Bayes Factor is a ratio -- the sum (or mean) likelihood times prior for the observed data divided by the sum of such terms for all possible data outcomes. NML* is also a ratio of two terms: Each term is again a ratio -- the maximum likelihood times prior for the observed data divided by the sum of such maxima for all possible data outcomes. Thus BMS is based on a normalized sum (or mean) of joint probabilities and NML* is based on a normalized maximum of the same joint probabilities. As a result BMS and NML* might be expected to differ quantitatively but align in most cases qualitatively, both in the way they select models and the way they balance fit and complexity.

We demonstrate that this is the case with detailed analysis of a binomial example: The data, an observed proportion of successes, are used to make inferences about the probability of success, θ; different model classes consist of different regions of the interval [0,1] that are posited to explain the data. When these intervals are disjoint, so that the two model classes do not overlap (e.g. [0,z] and (z,1]) BMS and NML* align qualitatively; when the two model classes overlap (e.g. [0,z] and [0,1]) BMS and NML* exhibit qualitatively differing patterns. Thus deletion of shared instances from the larger of the two classes has the virtue of bringing BMS and MDL into qualitative alignment. However, as we argued at length, there are several additional reasons justifying such deletion in scientific model selection: When the smaller class is very much smaller, as occurs when one compares models having unrestricted parameters to models with some of those parameters having specified values, deletion of shared instances will not cause any change in the traditionally calculated result. When the two classes are very similar in size, differing only in a small range of some parameter, the model selection question is really whether the parameter lies in that small interval, and deletion prior to model selection provides such an answer. Sometimes two classes differ modestly in size, for example when the range of some parameter in one class is double that range in another class. In this situation the models do not differ qualitatively, and we suggest that we should not be carrying out model comparison but rather estimation of that parameter value. When the two models do differ qualitatively, as when the two classes differ in some parameter having a value above or below some point value, then one really ought to be comparing the classes specified to lie on either side of this value, which is accomplished by deletion of shared instances.

References

P. Bartlett, P. Grünwald, P. Harremoes, F. Hedayati, W. Kotlowski. (2013) Horizon-Independent Optimal Prediction with Log-Loss in Exponential Families, JMLR Workshop and Conference Proceedings 30, *COLT (Conference on Learning Theory) 2013*: 662-69,

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). Statistical Inference under. *Order Restrictions*, John Wiley, New York.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. Journal of Mathematical Psychology.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356),

B. Clarke and A.P. Dawid (Unpublished manuscript, 1999). Online prediction with experts under a log-scoring rule.

Grünwald, P. (2005). A tutorial introduction to the minimum description length principle. *Advances in minimum description length: Theory and applications,* 23-81.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Heck, D. W., Wagenmakers, E. J., & Morey, R. D. (2015). Testing order constraints: Qualitative differences between Bayes factors and Normalized Maximum Likelihood. *Statistics & Probability Letters*, *105*(4), *157-162*.

Kakade, S., Seeger, M., & Foster, D. (2006). Worst-case bounds for Gaussian process models. In *Proceedings of the 19th Annual Conference on Neural Information Processing Systems* (No. EPFL-CONF-161315).

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57-69.

Kruschke, J. (2014). Doing Bayesian data analysis, Second Edition: A Tutorial with R, JAGS, and Stan. Academic Press.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*,*140*(4), 887-906.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465-471.

Rissanen, J. J. (1996). Fisher information and stochastic complexity.*Information Theory, IEEE Transactions on*, *42*(1), 40-47.

Shiffrin, R.M., & Chandramouli, S. H. (in press). Model Selection, Data Distributions and Reproducibility. In H. Atmanspacher and S. Maasen (Eds.) Reproducibility: Principles, Problems, and Practices. John Wiley, New York.

Chandramouli, S. H, Shiffrin, R.M. (in press). Extending Bayesian Induction. *Journal of Mathematical Psychology.*

De Rooij, S., & Grünwald, P. D. (2011). *Luckiness and regret in Minimum Description Length inference. Handbook of the Philosophy of Science, Volume 7: Philosophy of Statistics* (edited by Prasanta S. Bandyopadhyay and Malcolm Forster). Elsevier Science Publishers, 865-900

Roos, T., Myllymäki, P., & Tirri, H. (2005, January). On the behavior of MDL denoising. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)* (pp. 309-316).

Zhang, J. (2011).Model selection with informative normalized maximal likelihood: Data prior and model prior (book chapter). In Dzhafarov, E.N and Perry, L. (Eds) *Descriptive and Normative Approaches to Human Behavior,* World Scientific, New Jersey (pp. 303-319).

Appendix 1: Bayesian Inference

Assume that *M₁* and *M₂* share no instances that predict very similar distributions of data outcomes, up to the precision of the ability to measure in some experimental setting. Bayes Theorem makes posterior inference concerning instances and posterior inference about classes consistent with each other:

By Class: We use Bayes Theorem to calculate the posterior probabilities of model class M_1 (for two classes the probability of model class M_2 is 1.0 minus that of M_1). The instances of M_1 are denoted (θ, M_1) and the instances of M_2 are denoted (θ, M_2). The posterior probability of M_1 is:

$$(\mathbf{A1}) \qquad \boldsymbol{p(M_1|y) = \frac{p(y, M_1)}{p(y)} = \frac{\sum_w p(y, \theta_w, M_1)}{\mathbf{p(y)}}}$$

$$= \boldsymbol{\frac{\sum_w p(y|\, \theta_w, M_1) p_0(\theta_w, M_1)}{\mathbf{p(y)}}}$$

$$= \boldsymbol{\frac{\sum_w p(y|\, \theta_w, M_1) p_0(\theta_w|M_1) p_0(M_1)}{\mathbf{p(y)}}}$$

Where the denominator is:

$$\textbf{(A2)} \qquad \boldsymbol{p(y) = \sum_w p(y|\, \theta_w, M_1) p_0(\theta_w|M_1) p_0(M_1) + \sum_w p(y|\, \theta_w, M_2) p_0(\theta_w|M_2) p_0(M_2)}$$

A similar equation can of course be written for $p(M_2|y)$. We take the ratio of these posterior probabilities to connect with the usual way of characterizing classes by the Bayes Factor:

$$\boldsymbol{\frac{p(M_1|y)}{p(M_2|y)} = \frac{\sum_w p(y|\, \theta_w, M_1) p_0(\theta_w|M_1) p_0(M_1)}{\sum_w p(y|\, \theta_w, M_2) p_0(\theta_w|M_2) p_0(M_2)}} =$$

$$(\boldsymbol{A3}) \qquad \boldsymbol{\left\{\frac{\sum_w [p(y|\theta_w, M_1) p_0(\theta_w|M_1)]}{\sum_w [p(y|\theta_w, M_2) p_0(\theta_w|M_2)]}\right\} \quad \left\{\frac{p_0(M_1)}{p_0(M_2)}\right\}}$$

The left hand ratio in Eq. A3 is termed the Bayes Factor, and the rightmost is considered the prior odds for the model classes, albeit this is the discrete formulation (see Kass and Raftery, 1995). If we start with instances rather than classes, and again use Bayes Theorem to calculate the posterior probability of θ_k given y observed, and then define the class posterior as the sum of its instance posteriors, we end up with the same equations (due to our assumptions that the classes do not share instances).

Relating BF and NML*

The numerator of the Bayes Factor can be written as:

$$(\mathbf{A4}) \quad \boldsymbol{BF_{numerator}} = \frac{\sum_w [\boldsymbol{p(y|\theta_w, M_1)p_0(\theta_w|M_1)}]}{\sum_i \sum_w [\boldsymbol{p(x_i|\theta_w, M_1)p_0(\theta_w|M_1)}]}$$

In the denominator of Eq. A4 we are summing over all possible data outcomes. The denominator sums to 1.0 because some outcome must be observed. Therefore this divisor can be inserted in the numerator of the Bayes Factor in Eq. A3 without changing its value. We can write a similar expression for the denominator of the Bayes Factor. Hence the Bayes Factor is seen as a ratio across the two classes: Each term in the ratio is the *sum* of the joint probabilities for the instances and the observed data (conditional on the model class in question) normalized by a sum of like quantities for all data outcomes. This characterization can be generalized: One can multiply the numerator and denominator of A4 by any constant without changing the result. If the constant is 1/n_w, the number of instances in the class, then one can restate the result by saying each term is a normalized *average* joint probability.

Thus we get the relation to NML*, which can also be stated as a ratio of terms for each class. Each term is the ***maximum*** joint probability for the observed data normalized by a sum of such maxima over all data outcomes. The equivalent statement for the Bayes Factor says that each term is the ***sum (or average)*** joint probability for the observed data normalized by a sum of such sums (or averages) over all data outcomes.

Appendix 2: MDL (The Minimum Description Length Principle)

The NML instantiation of MDL was introduced by Rissanen (1996). This informal description leaves out an enormous amount of important information that is covered in detail by Grünwald (2007); see Grünwald (2005) for a shorter (but somewhat dated) tutorial-style review, and De Rooij and Grünwald (2011) for a modern, short, informal introduction. for a modern, short, informal introduction. It would take us too far afield to include more than a slight introduction to this literature, but a brief summary is useful. Like Bayesian inference, MDL is a general theory of inductive inference. Unlike Bayesian inference, its roots are information-theoretic: the central idea is that any regularity present in the data can be used to compress that data, and hence, one may recast the quest for patterns in the data as a quest for good codes, that substantially compress the data. When applied to model selection and averaging, one views model classes as defining candidates for codes that compress the data, and one should prefer a class such that the corresponding code compresses the available data to the greatest degree. MDL is typically – like BMS – applied to probabilistic model classes,

i.e. model classes whose instances are probability distributions. To make MDL model selection well-defined for such model classes one thus needs a mapping from sets of probability distributions to codes; different instantiations of MDL correspond to different such mappings. For a single model instance – equivalently a model class with just one instance – there is only one reasonable choice for such mappings: By the Kraft inequality – a central result in information theory – we know that:

For every probability distribution P that one can define on a space of outcomes X, there exists a code such that for every x in X, x is encoded with length L(x|P) = - log P(x) bits, rounded up to the nearest integer (the "|P"notation indicates that the code, and hence its lengths, depend on P: if were to start with a different P, we would construct a different length). *(Kraft Inequality)*

Here the logarithm is taken to base 2 and we ignore rounding issues; note that P(x) cannot exceed one so L(x) cannot be negative. By `code' we mean a code for lossless data compression, i.e. given a codeword, one can uniquely determine the outcome x that gave rise to it. As a very simple example, suppose that X consists of 8 outcomes and P is uniform, P(x) = 1/8 for all x in X. Then we could encode the outcomes as 000, 001, 010, …, 111, needing 3 = - log (1/8) bits per outcome. The Kraft inequality extends this idea to non-uniform distributions, leading to non-fixed length codes.

By another central result in information theory – the *information inequality* - the code with lengths – log P(x) would be the optimal code to use if data were sampled from P (in the sense that it minimizes the *expected* codelength under P, among all of the many codes that one could design for the data). Moreover, the better P fits the actually observed data, the smaller the associated codelength. For these two reasons, the code with these lengths is the only natural candidate code that one can associate with a model instance P: it effectively translates goodness-of-fit (in terms of log-likelihood) into codelength.

However, our goal is to associate not just a model instance, but a complete model class with a single code. The most straightforward candidate (suggested first by Rissanen (1978)) for finite model classes *M*, containing instances θ_k, is to use a *two-part code*, in which one codes the data by first coding the element θ of the model class that best fits the data (assigns highest probability) using a fixed-length code (so that if there were four elements in *M*, |*M*| = 4, one would need two bits to encode the model instance: the four instances would be encoded as 00, 01,10, 11). Next, in the second stage, one encodes the data using the code corresponding to the distribution indexed by θ. In this way, one associates model *M* with a code with lengths:

$$L(x|M) = \log|M| - \log \max_k [p(x|\theta_k, M)]$$

MDL model selection between models M_1 and M_2 would then proceed by picking the model minimizing the total codelength $L(x| M_j) = \log | M_j| - \log \max_k [p(x|\theta_k, M_j)]$. Note that this automatically induces an (at this point still rather crude) penalty for model complexity – the larger the model, the more the likelihood according to the maximum likelihood distribution within the model gets discounted.

Over the 1980s and 1990s it was gradually realized that the two-part code is not the best code to use: there exist codes which compress some outcomes x more, and no outcome x less than the two-part codes, and these should be preferred. A general theory of what codes are sensible to use was developed – among these are the two-part code (reasonable but far from optimal) and the so-called Bayes mixture code (already much better); and it was also established that among these, one code stands out as being the optimal one. This is the NML (Normalized Maximum Likelihood) code, the code that minimizes the worst-case coding regret, a notion that would take us too far astray to explain here. However, we do note that the Kraft inequality above has a converse: for every code C that allows lossless data compression, coding each x with length L(x|C) , there is a probability distribution P such that for all x in the outcome space X, -log P(x) = L(x|c). We can thus think of NML model selection as selecting the model M_j such that $-\log P(x| M_j): = L(x| M_j)$ is minimized, where $L(x| M_j)$ is the codelength of x when encoded with the NML code corresponding to M_j and $P(.|M_j)$ is the distribution corresponding to this code. Thus, we can also think of NML associating each model class with a *distribution* $P(. | M_j)$ and selecting the model maximizing $P(x| M_j)$. Here $P(x| M_j)$ is the *NML distribution*, which is defined further below; note that in this formulation, we clearly see the similarities to BMS, which also can be thought of as picking the model class maximizing P(x| M_j), but now P(.|M_j) is instantiated as the Bayes marginal distribution.

Unfortunately, for model classes containing infinitely many distributions defined over infinite sample spaces (including such basic cases as the Gaussian location family) the NML is usually not well-defined; moreover, the basic definition does not allow us to incorporate prior knowledge, even if we have it. For these reasons, several authors have suggested augmenting basic NML. Various proposals were unified by Grünwald (2007, pg. 311) by introducing the idea of a *luckiness function*, which addresses both problems with the original NML; the original NML is recovered when the luckiness function is taken to be uniform (i.e. constant over model instances). The luckiness function plays a role somewhat akin to the role of priors in BMS, and in our particular setting with finite model classes, there is indeed a one-to-one correspondence between luckiness functions and priors. Thus, the augmented NML model selection, termed NML*, associates a given model M_0 and luckiness function $p_0(\theta_k|M_0)$ with a new distribution over possible data realizations,

$$P_{nml} * (y|M_0) \ = \frac{\max_k[p(y|\theta_k, M_0)p_0(\theta_k|M_0)]}{\sum_i \max_k[P(x_i|\theta_k, M_0)p_0(\theta_k|\ M_0)]} \qquad (3)$$

This equation formally defines the result described with use of Table 5. Eq. 3 is used to obtain a value for each model under consideration. The amount of evidence in the observed data y for model M_1 over M_2 is measured as the ratio $P_{nml} * (y|M_1)/P_{nml} * (y|M_2)$. As stated earlier, if the prior over model instances is uniform, then both numerator and denominator in Eq. 3 share a constant prior probability that cancels from the equation – the resulting equation is then the original NML.

To elaborate slightly concerning 'luckiness', Grünwald (2007) introduced this to incorporate a form of subjective judgment into MDL, generalizing previous approaches by Kakade, Seeger and Foster (2006), Roos, Myllymäki, and Tirri (2005), Zhang (2011, based on an earlier unpublished manuscript from 2005) and Clarke and Dawid (1999). This approach was originally called Luckiness Normalized Maximum Likelihood-2, since it was just one of a variety of proposals; but it has turned out to be the most fruitful one and is indeed, becoming the standard, see e.g. Bartlett et al. (2013). Avoiding the cumbersome original terminology, we have simply denoted it NML*.

It is worth noting a general asymptotic equality of BMS and NML*: We know from Grünwald (2007) that for every *fixed* two statistical models that are sufficiently regular, for any sequence of trials such that the maximum likelihood estimator remains bounded away from the boundaries of the parameter space in both models for all large N, model selection by NML* and model selection by BMS will eventually, that is, for all large N, select the same models. This holds in particular for our Bernoulli example for the choices of M_1, and M_2: e.g. M_1 positing θ to lie in [0, .5] and M_2 positing θ to lie in (.5, 1]. However, this is an asymptotic result – as we see in this paper, for small samples there will generally be quantitative differences and sometimes be qualitative difference in the model selected. Moreover, note that qualitative disagreement between NML* and BMS occurs for data with maximum likelihood estimators (frequency of successes) close to the boundary between M_1 and M_2 – exactly the type of data for which the asymptotic result does not say much

Appendix 3: Equivalence Classes

Assume we have priors for all instances in all classes, and that the priors for classes are the sum of priors for each class's instances. The present study might not have the power to discriminate some of the data distributions predicted by the instances in the classes. If the distribution for any instance in the smaller class cannot be discriminated from one or more instances in the larger class, those larger class instances are deleted prior to model comparison, following our recommendation that shared instances be removed. Now consider the instances within each class. We can group the instances θ_i into similarity classes S(k) such that discrimination cannot be done within class, on the basis of the present observed data: $P(x|\theta_i)$ is the same value for all θ_i in S(k). Note that $\Sigma_i p_o(\theta_i,M_0)] = p_o(S(k))$ for $\theta_i \in S(k)$. Then we can rewrite the following term from Eq. A1:

$$(\mathbf{A5}) \qquad \frac{\sum_w \boldsymbol{p(y|\,\theta_w, M_1)p_0(\theta_w, M_1)}}{\mathbf{p(y)}}$$

$$= \frac{\sum_w \sum_k \{\boldsymbol{p[y|S(k), M_0]\; p_0(S(k), M_1)}\}}{\mathbf{p(y)}}$$

Thus all inference can be carried out on the basis of the similarity classes defined by the precision of the present data, where the prior knowledge that allows discrimination within similarity class is represented by a sum of such priors across each such class, and the relative values within each similarity class remain unchanged.

Appendix 4: Situations Requiring Retention of Shared Instances

An example discussed by Kruschke (2014, pgs. 268-273) has factories A and B making biased coins, with evidence from past sampling producing posteriors for $p_A(s)$ characterized as a beta distribution (beta(A)) with (say) a low mean and the posterior for $p_B(s)$ characterized as a beta distribution (beta(B)) with a high mean. One coin is sampled from A or B (say with probability .5) flipped N times with n successes, and one wants to determine the probability it was sampled from A. This can be conceived as model selection between A and B, and BMS (in the form of Bayes Factors because the class priors are equal) applied to determine the desired probabilities. This is a valid approach, but we would argue first that the situation is better considered probabilistic inference than scientific model selection, and second that there are alternative ways to characterize the situation and reach the same conclusions, without using shared instances.

First one might argue that this is not model selection because the state of the world that generates data is known, and not in question: Two factories with known characteristics. If we were asking whether one or two factories were producing observed data, that would be a question of scientific model selection, but in that case the two-factory model class would be much larger and the one-factory model instances could be deleted from the larger class without changing results.

Second, we could re-characterize the two-factory situation so that the issue of shared instances does not arise: We could say that there exists a single model class, not two: The instances of this one class are characterized by a pair of beta distribution parameters (beta(A), beta(B)). We can calculate posterior instance probabilities (for each pair of beta parameters) based on the observed data, and then use those posterior probabilities to calculate the probability that the coin was sampled from A. Thus it is equally valid to we view this problem as one of deduction, rather than induction, reaching the same answer with the same calculations. In the two-factory example, as N increases the posterior probability of a given instance (value of θ) will rise to 1.0, but the posterior odds of A vs. B will only converge to the ratio of the beta densities at that value of θ. We could want to classify inference as 'model selection' when increasing amounts of data cause the posterior probability of a class to rise to 1.0. In a different variant of Kruschke's example we might choose A or B randomly and then sample N different coins from it, tossing each coin once, observing n successes. Now as N increases the posterior probability of A or B will converge because n/N will come to match the mean of beta(A) or beta(B). But in this case the classes are the instances: There are just two of both: Factory A and Factory B. The issue of shared instances does not arise. The preference for convergence of posteriors would admittedly not be a convincing argument to everyone, so we simply conclude that in cases like these there are alternative characterizations of the inference problem, with or without shared instances, each of which would lead to the same result.

To clarify these types of situations further, note that they are analogous to situations often arising in hierarchical modeling. For example, consider subject differences: In some task we believe the N subjects have their individual parameters (k each) sampled from a master distribution. In modeling this situation we could define an instance to be a particular set of kN parameter values assigned to all N subjects, where the parameters for a given subject differ from those for another. We could collect data from a random subject and ask which subject is likely to be the origin of the data (deduction) or could ask a variety of model selection questions (induction): E.g. Are all subjects (more) likely to have equal parameter values? In this latter example, the number of instances is far larger for the unrestricted model, so it makes no difference whether or not we subtract the shared instance (equal parameters) from the larger model. On the other hand we could suppose there are just four subjects and ask whether each subject has a parameter in the range [0,.9] or in the range [0,1]. Now the overlap of instances is considerable (roughly (.9)**4 = .65) and we argue that it is now important to remove the shared instances from the larger model, thereby changing the model comparison question to one more appropriate.

Appendix 5: Priors, Posteriors, Complexity, and Precision of Observed Data

We suggest that distributions that cannot be distinguished on the basis of the experimental data be treated as identical; when such instances lie in different classes they should be treated as shared. E.g. consider an example in which one wants to infer the y intercept at x = 0 of a zero slope linear model on the basis of observations at various x values that are equal to the true intercept plus 0,1 gaussian noise. Model Class A says the intercept lies in [0,1]. Model Class B says the intercept lies in [0,.5]. Now consider two other model classes C and D that posit a zero slope sine wave with some fixed phase but very low amplitude (suppose that the sine looks flat and linear to the eye when graphed across the x values used for data collection). Model Class C posits a y intercept at x = 0 that lies in [0,1], and Model Class D posits a y intercept lying in [0,.5]. On what basis would one want to distinguish model classes A from C and B from D? The likelihoods of any reasonable amount of data would be the same, so any differences in inference would be due solely to the priors assigned to the models, and there would be no point in using inference machinery. Thus although not standard practice, we suggest that instances be considered the same if they predict distributions that cannot be discriminated on the basis of the current data (whatever the functional form of the instance or class). Obviously this would require some criterion for discriminability (perhaps based on some sort of Kullback-Leibler comparison) but the precise choice is probably not critical, because we are interested in preferring one model over another only if there is a fairly large statistical advantage.

In theory, and perhaps also in practice, defining discriminability of data distributions is necessary in order to define complexity of a model class. We have counted instances as a measure of class size, and this is sensible if all data distributions lie in a very large multidimensional space tiled by volumes just noticeably different. Thus in the binomial example an assumption of equal priors for all values of θ in increments of .01 is implicitly an assumption that the associated data distributions are equally discriminable. But of course in almost all settings this will not be the case--variability is much higher near θ = .5 than for θ values near 0 and 1, so discriminability will be lower near .5. One might therefore in theory want to assign equal priors to instances whose width along the θ axis is scaled to variability. If that were to be done, it would of course require assessment of all sources of variability not just that due to the binomial sampling. The bottom line is that the assignment of priors to instances and classes needs to take the discriminability of data distributions into account.

The idea that complexity be defined not by the potential discriminability of instances and classes but the actual discriminability determined by the current data seems to remove (some of) the bias for simplicity inherent in model selection. However, suppose two model classes differ in their complexity of parameterization (say 23 parameters vs. 2 parameters) but every instance in one has an instance in the other that matches up to the level of precision in the present data. If the simpler model makes the workings of the model easier to understand this provides a reason for preferring the simpler model even though there is no difference in the number of data distributions predicted by the two models. However the reverse could also be the case: We believe that the latter basis for a simplicity bias should become important only when the amount of data in the current experiment is large enough to produce discriminable predictions.

A counter argument would say that prior studies may have contained far more data, enough to discriminate the model instances and classes, making it entirely appropriate to treat the larger class as more complex. This might well be the case, but in this situation, the present study does not and cannot change the inference about the portions of the model classes that produce distributions that cannot be discriminated on the basis of the present data--our beliefs about these instances that cannot be discriminated rest entirely on the priors. This point and a generalization are given in Appendix 3.

These arguments should not be taken to imply that one should always favor model classes with fewer parameters, given that both classes have predicted distributions that cannot be discriminated by the experimental data. It is quite possible to have a model class with few parameters (such as a recurrent net) whose parameter values produce differences that are difficult to understand, whereas an alternative model with more parameters may enhance understanding by assigning different parameters to different processes. The point here should be clear: The focus should be fundamentally on the goals of inference, not the tools.