Adrian Smith Volume

Merlise Clyde and Edwin S. Iversen
Department of Statistical Science, Duke University, Durham NC 27708-0251 U.S.A.

CLARENDON PRESS • OXFORD
2012

Abstract

Consideration of multiple models is routine statistical practice. With computational advances over the past decade, there has been increased interest in methods for making inferences based on combining models. Examples include boosting, bagging, stacking, and Bayesian Model Averaging (BMA), which often lead to improved performance over methods based on selecting a single model. Bernardo and Smith have described three Bayesian frameworks for model selection known as the M-closed, M-complete, and M-open perspectives. The standard formulation of Bayesian Model Averaging arises as an optimal solution for combining models in the M-closed perspective where one believes that the "true" model is included in the list of models under consideration. In the M-complete and M-open perspectives the "true" model is outside the space of models to be combined, so that model averaging using posterior model probabilities is no longer applicable. Using a decision theoretic approach, we present optimal Bayesian solutions for combining models in these frameworks. We illustrate the methodology with an example of combining models representing two distinct classes, prospective classification trees and retrospective multivariate discriminant models applied to gene expression data from advanced stage serous ovarian cancers. The goals of this analysis are two-fold: identifying molecular tumor characteristics associated with prognosis and determining if long-term survival can be predicted by features intrinsic to the molecular biology of the tumor.

Key words: BMA, cross-validation M-closed, M-complete, M-open, model comparison

1

BAYESIAN MODEL AVERAGING IN THE M-OPEN FRAMEWORK

1.1 Introduction

Consideration of multiple models is ubiquitous in statistical practice. In Chapter 6, Bernardo & Smith (9) describe three distinct settings for the model comparison problem, denoted as $\mathcal{M}$–*closed* , $\mathcal{M}$–*complete* , and $\mathcal{M}$–*open* which have far reaching consequences for how models should be compared, selected or combined.

The predominant perspective is the $\mathcal{M}$–*closed* view, where one entertains a collection of models $\mathcal{M} = \{\mathcal{M}_j, j = 1, \ldots J\}$, with the belief that one of the models in $\{\mathcal{M}_j\}$ is the "true" generating model for the data, but that the true generating model is unknown. In this framework, a Bayesian would use probabilities $p(\mathcal{M}_j)$ to represent one's subjective (or objective) prior beliefs about the "truth" of model $\mathcal{M}_j$. These beliefs combined with any prior beliefs about parameters within models are updated via Bayes theorem to obtain a joint posterior distribution for models and model specific parameters. The Bayesian paradigm provides a comprehensive framework for accounting for both parameter and model uncertainty, leading to the well known Bayesian Model Averaging solution. For additional references, history, and examples we refer the reader to review articles by (24). In conjunction with a decision theoretic approach, this joint posterior distribution can be used to construct optimal decision rules for selecting the "best" model, make inferences about parameters or predict future observations under selected utility functions. For additional references for the Bayesian approach to model choice we refer the reader to review articles by (17; 12; 16; 24) for history and examples.

In reality, the true process generating the data may be too complex to be used in practice or even to articulate as a probabilistic model, which leads to the $\mathcal{M}$–*complete* and $\mathcal{M}$–*open* perspectives discussed in (9). In both of these formulations, the true generating model $\mathcal{M}_T$ is not included in the collection of models $\mathcal{M}$, rather the models in $\mathcal{M}$ are viewed as potential proxies available for comparison or model selection. With the belief that the true model is $\mathcal{M}_T$, assigning prior probabilities to models in $\mathcal{M}$ no longer makes sense, as the model is no longer part of the unknown specification of the data generating process. In the $\mathcal{M}$–*complete* specification, while one can specify $p(\mathbf{Y} \mid \boldsymbol{\theta}, \mathcal{M}_T)$, one may still wish to select a proxy model in $\mathcal{M}$ because of its attractive simplicity or ease of communication of results with others or computational tractability. In the more realistic $\mathcal{M}$–*open* alternative, the list of models in $\mathcal{M}$ are also to be used in place of $\mathcal{M}_T$, however, there is no explicit specification of a belief model $p(\mathbf{Y} \mid \mathcal{M}_T)$.

Under the $\mathcal{M}$–*open* perspective, (9; 26) motivate the role of cross-validation to evaluate expected utility, leading to intrinsic Bayes factors for the model choice problem. Rather than model selection, our goal in this paper is optimal combination of multiple proxy models in the $\mathcal{M}$–*open* framework. In the next section, we review the standard $\mathcal{M}$–*closed* Bayesian Model Averaging approach and decision-theoretic methods for producing inferences and decisions. We then review model selection from the $\mathcal{M}$–*complete* and $\mathcal{M}$–*open* perspectives, before formulating a Bayesian solution to model averaging in the $\mathcal{M}$–*open* perspective. We construct optimal weights for MOMA: $\mathcal{M}$–*open* Model Averaging using a decision-theoretic framework, where models are treated as part of the "action space" rather than unknown states of nature. We illustrate MOMA using "incompatible" retrospective and prospective models for data from a case-control study and demonstrate that MOMA gives better predictive accuracy than using any of the proxy models. We conclude with open questions and future directions.

1.2 $\mathcal{M}$–*closed* Model Averaging

In the standard setup the joint distribution of the data and all unknowns may be described hierarchically, with $p(\mathbf{Y}|\boldsymbol{\theta}_j, \mathcal{M}_j)$ specifying the distribution of the data $\mathbf{Y} = (Y_1, \ldots Y_n)^T$ given model specific parameters $\boldsymbol{\theta}_j$ in model $\mathcal{M}_j$, $p(\boldsymbol{\theta}_j \mid \mathcal{M}_j)$ reflecting prior uncertainty in the model specific parameters. A Bayesian would assign a prior probability, $p(\mathcal{M}_j)$, representing their belief (subjective or objective) that each model $\mathcal{M}_j$ is the true model.

In turn, posterior model uncertainty is represented by the posterior probabilities of models obtained via Bayes theorem

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_{j=1}^{J} p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)} \tag{1.1}$$

where

$$p(\mathbf{Y} \mid \mathcal{M}_j) = \int p(\mathbf{Y} \mid \boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j \mid \mathcal{M}_j)d\boldsymbol{\theta}_j, \tag{1.2}$$

is the marginal likelihood of $\mathcal{M}_j$. Given observed data $\mathbf{Y}$, the posterior probability of each model $p(\mathcal{M}_j|\mathbf{Y})$ represents a posterior measure that model $\mathcal{M}_j$ generated the data. The joint posterior distribution of $\boldsymbol{\theta}_j, \mathcal{M}_j$, $p(\boldsymbol{\theta}_j|\mathbf{Y}, \mathcal{M}_j)p(\mathcal{M}_j|\mathbf{Y})$, provides a complete post-data representation of parameter and model uncertainty that can be used for a variety of inferences and decisions. For example, the distribution of a future observation Y^*. Under the hierarchical model for the data, the Bayesian predictive distribution of Y^* is a mixture model

$$p(Y^* \mid \mathbf{Y}) = \sum_j p(Y^* \mid \mathcal{M}_j, \mathbf{Y})p(\mathcal{M}_j \mid \mathbf{Y}), \tag{1.3}$$

with components in the mixture the conditional predictive distributions

$$p(Y^* \mid \mathcal{M}_j, \mathbf{Y}) = \int p(Y^* \mid \boldsymbol{\theta}_j, \mathcal{M}_j)p(\boldsymbol{\theta}_j \mid \mathcal{M}_j, \mathbf{Y})d\boldsymbol{\theta}_j \tag{1.4}$$

and mixing weights given by the posterior probabilities of models from equation (1.1).

1.2.1 *Optimal Decisions*

The joint posterior distribution of $\mathcal{M}_j$ and $\boldsymbol{\theta}_j$ provides a complete summary of one's beliefs after seeing the data. Combined with a decision-theoretic framework, this posterior permits making inferences or decisions that optimize one's utility. More formally, let $u(\omega, a)$ be a mapping from $A \times \Omega$ to $\mathbb{R}$ that reflects the utility of taking action a when the unknown state of nature is ω. Commonly used utility functions are negative quadratic loss for estimation or prediction or proper scoring rules if ω is a distribution. For a Bayesian, the optional action to take is the one that maximizes the posterior expected utility

$$a^* = \arg\sup_{a \in A} \int_\Omega u(\omega, a) p(\omega \mid \mathbf{Y}) d\omega \tag{1.5}$$

where $p(\omega \mid \mathbf{Y})$ is the posterior (predictive) distribution of ω given the data $\mathbf{Y}$.

Consider the decision problem of prediction under quadratic loss

$$u(Y^*, a) = -(Y^* - a)^2$$

where Y^* is the unknown "state of nature", a is a possible action in action space $A = \mathbb{R}$ and u is the utility of taking action a when the future value is Y^*. Under quadratic loss for prediction, the optimal action for the point prediction of Y^* is $a^* = E(Y^* \mid \mathbf{Y})$, the (posterior predictive) mean of Y^* given Y, which under the $\mathcal{M}$-closed perspective, can be expressed as

$$E(Y^* \mid \mathbf{Y}) = \sum_{j=1}^{J} \mathsf{E}(Y^* \mid \mathcal{M}_j, \mathbf{Y}) p(\mathcal{M}_j \mid \mathbf{Y}) = \sum_{j=1}^{J} p(\mathcal{M}_j \mid \mathbf{Y}) \hat{Y}^*_{\mathcal{M}_j} \tag{1.6}$$

where $\hat{Y}^*_{\mathcal{M}_j}$ is the posterior mean under model $\mathcal{M}_j$. This is the well known Bayesian Model Averaging solution, where the prediction is a weighted average of the model specific predictions $\hat{Y}^*_{\mathcal{M}_j}$ with weights that are given by the posterior model probabilities. Such model averaging or mixing procedures have been developed and advocated for by (27), (20), (17), (32) and (15), and are now widespread.

If the goal is to find the single model that leads to the best prediction under quadratic loss, then the set of actions consist of selecting a model and reporting the prediction under that model. The solution given by (9), Sec 6.1 is the model that minimizes $(\hat{\mathbf{Y}}^*_{\mathcal{M}_j} - \mathsf{E}_{Y^*}(Y^* \mid \mathbf{Y}))^2$; the single model whose predictions are closest to the BMA solution. For the case of 2 models, this is the highest probability model, but in general the model closest to the BMA solution may not correspond to the highest probability model nor the median probability model of (1) except in special circumstances. While closed form expressions are generally unavailable, one can determine the best model by evaluating the distances between the models under consideration. When there is high correlation among the predictors this is preferable to the median probability model. (33) use a log scoring rule for selecting the model that is closest to model averaging solution in terms of predictive densities.

1.2.2 *BMA is not a Panacea*

In problems where there are a large number of predictors (p), such as genome wide association studies, one might consider model averaging where the candidate models are each based on a single predictor, rather than approximating model averaging by stochastic search over the 2^p potential models. To illustrate that model averaging in such a setting may fail, consider the simplified setting with just two models in $\mathcal{M}$

$$\mathcal{M}_1 : \mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{e} \tag{1.7}$$

$$\mathcal{M}_2 : \mathbf{Y} = \mathbf{X}_2\beta_2 + \mathbf{e} \tag{1.8}$$

leading to $\hat{\mathbf{Y}}^* = p(\mathcal{M}_1 \mid \mathbf{Y})\mathbf{X}_1\hat{\beta}_1 + p(\mathcal{M}_2 \mid \mathbf{Y})\mathbf{X}_2\hat{\beta}_2$. This seems appealing as the model averaging solution does contain all potential predictors, even though the full model was not included in the list of candidate models. If the "true" model does in fact contain both predictors $\mathbf{Y} = \mathbf{X}_1\beta_{1T} + \mathbf{X}_2\beta_{2T} + \mathbf{e}$ then, under standard regularity conditions, the BMA model weights converge to 1 for the model that is "closest" to true model (in terms of Kullback-Leibler divergence); BMA only uses predictions from that model; and in the limit BMA is not consistent if $\mathcal{M}_T \notin \mathcal{M}$. The obvious solution is to add the full model to the list of models under consideration for this toy example. However, if the true model is some complex nonlinear function, one may need to use a richer set of basis vectors, such as in over-complete representations, for model averaging to lead to consistent results (38). For ease of exposition, one may still wish to use simple proxy models when the true model is not included in $\mathcal{M}$, which leads to the $\mathcal{M}$*–closed* and $\mathcal{M}$*–open* perspectives.

1.3 Model Comparison without the True Model

When the true model is not in $\mathcal{M}$, we consider two cases

$\mathcal{M}$*–complete* **:** we know the true model, $\mathcal{M}_T$, and $p(Y^* \mid \mathbf{Y}) = p^C(Y^* \mid \mathcal{M}_T, \mathbf{Y})$ is available. We may, however, wish to use the models in $\mathcal{M}$ because of ease in communication of results, tractability of computations, reasonable proxies, etc.

$\mathcal{M}$*–open* **:** we know that the true model is NOT in $\mathcal{M}$, but we cannot specify $p(Y^* \mid \mathbf{Y}) = p^o(Y^* \mid \mathcal{M}_T, \mathbf{Y})$ because it is too difficult, we lack time to do so, or do not have the expertise, computational intractability, etc.

For the model comparison problem in the $\mathcal{M}$*–complete* setting, one simply finds the model in $\mathcal{M}$ which maximizes the expected utility, where now the expectation is with respect to the predictive distribution $p^c(Y^* \mid \mathcal{M}_T, \mathbf{Y})$. For the $\mathcal{M}$–*open* case, one again finds the optimal model and action $a^*(\mathbf{Y}, \mathcal{M}_j)$ under model $\mathcal{M}_j$ that maximizes expected utility,

$$\int u(y^*, a^*(\mathbf{Y}, \mathcal{M}_j))p^o(y^* \mid \mathcal{M}_T, \mathbf{Y})\, dy. \tag{1.9}$$

As the predictive distribution is not available in the $\mathcal{M}$–*open* setting (9, 26) argue that for exchangeable data and large n that the expected utility can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n} u(y_i, a^*(\mathbf{Y}_{(i)}, \mathcal{M}_j)) \tag{1.10}$$

based on partitioning $\mathbf{Y}^T = (y_i, \mathbf{Y}_{(i)}^T)$ into n partitions of the data where $\mathbf{Y}_{(i)}$ denotes the data without the ith observation and serves as a proxy for the observed data and y_i as a proxy for the future value Y^*. Randomly selecting from K of these partitions, they suggest a law of large numbers argument to justify that as $n, K \to \infty$

$$\left| \int u(Y^*, a(\mathbf{Y}, \mathcal{M}_j))p^o(Y^* \mid \mathbf{Y}, \mathcal{M}_T)\; dY^* - \frac{1}{K}\sum_{k=1}^{K} u(y_k, a(Y_{(-k)}, \mathcal{M}_j)) \right| \to 0,$$

thereby justifying the use of cross-validation to approximate expected utility.

Walker & Gutiérrez-Peña (34) in the discussion of (26) provide an alternative justification for the above as an approximation based in a Bayesian nonparametric model. If we assume the data are exchangeable, coming from an unknown distribution F, then one may place a nonparametric prior on F, such as a Dirichlet process,

$$F \sim DP(\alpha_0, F_o)$$

with α_0 a scale or prior weight parameter and F_0 a parametric distribution that is the location parameter such that $E(F) = F_0$ (18). Given a sample of size n, the posterior of F is again $DP(\alpha_n, F_n)$ where $\alpha_n = n + \alpha_0$, $F_n = (n\hat{F}_n + \alpha_0 F_0)/(n + \alpha_0)$ and $\hat{F}_n$ is the empirical distribution of the data. Using the nonparametric prior, the posterior predictive distribution for a new observation Y^* is F_n and the expected utility is expressed as

$$\int u(y^*, a^*(\mathbf{Y}, \mathcal{M}_j))dF_n(y^*) = \frac{n}{n+\alpha_0}\int u(y^*, a^*(\mathbf{Y}, \mathcal{M}_j))d\hat{F}_n(y^*) + \tag{1.11}$$

$$\frac{\alpha_0}{n+\alpha_0} u(y^*, a^*(\mathbf{Y}, \mathcal{M}_j))dF_o(y^*). \tag{1.12}$$

(23) take F_0 to be centered at the $\mathcal{M}$–*closed* predictive distribution, so that as $\alpha_0 \to \infty$ one recovers the $\mathcal{M}$–*closed* solution, while as $\alpha_0 \to 0$

$$\int u(y^*, a^*(\mathbf{Y}, \mathcal{M}_j))dF_n(y^*) \to \frac{1}{n}\sum_{i=1}^{n} u(y_i, a^*(\mathbf{Y}, \mathcal{M}_j)) \tag{1.13}$$

leads to their $\mathcal{M}$–*open* solution. The main difference between (1.13) and (1.10) is that the optional action under model $\mathcal{M}_j$ uses all data in constructing the distributions that go into $a^*(\mathbf{Y}, \mathcal{M}_j)$ in (1.13), while (26) use $a^*(\mathbf{Y}_{(i)}, \mathcal{M}_j)$ based on the training data in (1.10). Equation (1.13) provides internal rather than external validation as in (1.10).

In the case of prediction under quadratic loss under the limiting DP model, we would minimize over $\mathcal{M}$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - E(Y_i \mid \boldsymbol{M}_j, \mathbf{Y}))^2. \tag{1.14}$$

In the case of linear models $E[\mathbf{Y}] = \mathbf{X}_M \beta_{\mathcal{M}}$ with non-informative priors $p(\beta_{\mathcal{M}}) \propto 1$ assigned to parameters in $\boldsymbol{M}_j$, $E(Y_i \mid \boldsymbol{M}_j, \mathbf{Y}) = \mathbf{x}_i^T \hat{\beta}_{\mathcal{M}}$ where $\hat{\beta}_{\mathcal{M}}$ is the ordinary least squares estimate. The criterion would lead to picking the model with the smallest residual sum of squares regardless of model dimension (or highest R^2), which is leads to poor predictive performance. In contrast, the CV-approach of Bernardo & Smith chooses the model that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - E(Y_i \mid \boldsymbol{M}_j, \mathbf{Y}_{(i)}))^2 \tag{1.15}$$

over $\mathcal{M}$. This captures how well model $\mathcal{M}_j$ predicts, on average, a left out observation given the remaining cases (21). While we do not advocate non-informative priors in this setting, this highlights a potential problem of using the (limiting) DP prior.

Under the log scoring rule

$$\int \log(p(y \mid \mathcal{M}_j, \mathbf{Y}))p(y \mid \mathbf{Y})dy \tag{1.16}$$

(9, 26) propose the following approximation to the expected utility

$$\frac{1}{K}\sum_{k=1}^{k}\log(p(y_k \mid \mathbf{Y}_{(k)}, \mathcal{M}_j)) \tag{1.17}$$

which may be rearranged to form a criterion that implies that one would prefer model $\mathcal{M}_i$ to model M_0 if

$$\prod_{k=1}^{K}\left[\frac{p(y_k \mid \mathbf{Y}_{(k)}, \mathcal{M}_j)}{p(y_k \mid \mathbf{Y}_{(k)}, \mathcal{M}_0)}\right]^{1/K} > 1. \tag{1.18}$$

This corresponds to the Geometric Intrinsic Bayes factor criterion of (5, 4, 6) where $\mathbf{Y}_{(k)}$ represents a minimal training sample. This is related to the expression in (36) where $\mathbf{Y}$ replaces $\mathbf{Y}_{(k)}$.

Winkler (35) in the discussion of (26) raises the question of “Why there is so much focus on model choice? If there is no ‘true’ model, why do we have to choose a single model?”. As George Box is often quoted “Essentially all models are wrong, but some are useful.” Why should we restrict attention to just one of the proxy models if all are potentially useful? Rather selecting a model, we examine optimal weighted averages.

1.4 Combining Models in the $\mathcal{M}$*–open* Setting

In the $\mathcal{M}$*–closed* setting models are essentially an expansion of the "parameter" space so that the unknown state of nature Ω is comprised of models and model specific parameters. The optimal weights for prediction $\sum w_j E(Y^* \mid \mathbf{Y}, \mathcal{M}_j)$ are the posterior probabilities of models, which are proportional to the prior model probabilities times the marginal distributions of the data $p(\mathbf{Y} \mid |M_j)$. In both the $\mathcal{M}$*–complete* and $\mathcal{M}$*–open* viewpoints, the assignment of prior probabilities $\{p(\mathcal{M}_i), i \in \mathcal{M}\}$ no longer makes sense as a measure of our degree of belief in model $\mathcal{M}_j$ if we really believe that $\mathcal{M}_T \notin \mathcal{M}$. Thus the standard BMA solution to combining models using weights that are posterior model probabilities is not applicable.

In the decision theoretic framework for the $\mathcal{M}$*–closed* or $\mathcal{M}$*–open* perspectives, models are not part of the state of unknowns Ω, but may be part of the decision or action space. Under this alternative viewpoint model weights w_j are solely part of the action space A, so that optimal weights to combine predictions or predictive distributions from the collection of proxy models in $\mathcal{M}$ becomes a decision problem.

1.4.1 *Combining Models as a Decision Problem*

Let $\{w_j, j \in J \; w_j \in \Omega\}$ denote the weights $\mathbf{w}$ and consider decision rules of the form $a(\mathbf{Y}, \mathbf{w}) = \sum_j w_j E(y^* \mid \mathbf{Y}, \mathcal{M}_j)$ in the case of prediction or $a(\mathbf{Y}, \mathbf{w}) = \sum_j w_j p(y^* \mid \mathbf{Y}, \mathcal{M}_j)$ for predictive densities. For quadratic loss, the expected utility is

$$E_{Y^*}[u(y^*, a(\mathbf{Y}, w)) \mid \mathbf{Y}] = -\int \|y^* - \sum_j w_j \hat{y}^*_{\mathcal{M}_j}\|^2 p(y^* \mid \mathbf{Y}, \mathcal{M}_t) \; dy^*$$

while for the log scoring rule,

$$E_{Y^*}[u(y^*, a(\mathbf{Y}, \mathbf{w})) \mid \mathbf{Y}] = \int \log(\sum_j w_j p(y^* \mid \mathcal{M}_j, \mathbf{Y})) p(y^* \mid \mathbf{Y}, \mathcal{M}_t) \, dy^*.$$

In the $\mathcal{M}$*–complete* perspective, since we have $\mathcal{M}_T$, we can in principle solve the optimization problem.

In the $\mathcal{M}$*–open* formulation, as before, partition $\mathbf{Y}^T = (y_k^T, \mathbf{Y}_{(k)}^T)$ into future and training data vectors of size $n - m$ and m respectively. Randomly select K from these n choose m partitions to construct the approximate criterion

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \frac{1}{K} \sum_{k=1}^{K} u(y_k, a(Y_{(-k)}, w)). \tag{1.19}$$

For the problem of prediction the problem may be stated as,

$$\text{Solve} \quad \hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \; -\frac{1}{K}\sum_{k=1}^{K} \|y_k - \sum_{\mathcal{M}_j \in \mathcal{M}} w_j \hat{Y}_{(k),\mathcal{M}_j}\|^2 \tag{1.20}$$

$$\text{subject to} \tag{1.21}$$

$$\sum_{j=1}^{J} w_j = 1 \tag{1.22}$$

$$w_j \geq 0 \quad \forall j \in \{1, \ldots, J\} \tag{1.23}$$

where the optimal solution may be found using a quadratic programming algorithm. This has an equivalent representation using Lagrangians:

$$-\frac{1}{K}\sum_{k=1}^{K} \|y_k - \sum_{j=1}^{J} w_j \hat{Y}_{(k),\mathcal{M}_j}\|^2 - \lambda_0(\sum_{j=1}^{J} w_j - 1) + \sum_{j=1}^{J} \lambda_j w_j .$$

The two constraint functions ensure that the weights have the same support as posterior model probabilities in BMA (sum to one and non-negativity), however, the weights do not have an interpretation as posterior probabilities. While the expression above looks like a log likelihood from a Gaussian distribution, it is not the predictive log likelihood.

Similarly, under the log scoring rule,

$$\frac{1}{K}\sum_{k=1}^{K} \log\left(\sum_{j=1}^{J} w_j p(y_k \mid \mathbf{Y}_{(k),\mathcal{M}_j}, \mathcal{M}_j)\right) - \lambda_0(\sum_{j=1}^{J} w_j - 1) + \sum_{j=1}^{J} \lambda_j w_j$$

the expected utility takes a form similar to a likelihood from a mixture model. This relationship permits an iterative solution for the weights as in estimation of mixture models, where starting with initial weights $\hat{w}_j^{(0)}$, we update the weights

$$\hat{w}_j^{(t)} \equiv \frac{1}{K}\sum_{k} \frac{\hat{w}_j^{(t-1)} p(y_k \mid \mathbf{Y}_{(k)}, \mathcal{M}_j)}{\sum_j \hat{w}_j^{(t-1)} p(y_k \mid \mathbf{Y}_{(k)}, \mathcal{M}_j)} \tag{1.24}$$

until convergence.

Remark on Solution: In the $\mathcal{M}$–*closed* setting the model weights may be obtained from Bayes factors for comparing any model to a base model, $P(\mathcal{M}_j \mid \mathbf{Y}) = w_j B(\mathcal{M}_j : \mathcal{M}_0)/\sum_j w_j B(\mathcal{M}_j : \mathcal{M}_0)$ where w_j is the prior probability of $\mathcal{M}_j$. In the $\mathcal{M}$–*open* framework, the geometric intrinsic Bayes factor (GIBF) leads to a model selection criterion under the log scoring rule (5), however, the optimal model weights in (1.24) for combining models under the log scoring rule are not equivalent to the renormalized GIBF nor the closely related arithmetic intrinsic Bayes Factors.

1.4.2 *Restrictions on Weights*

As predictions may have support on $\mathbb{R}^K$, we may impose a range of restrictions on the weights by choice of λ_j. With $\lambda_0 = \lambda_1 = \ldots = \lambda_J = 0$ we obtain arbitrary weights ($w_j \in \mathbb{R}$). Setting $\lambda_1 = \ldots = \lambda_J = 0$ enforces the weights to sum to one, but allows positive and negative weights, while with all $\lambda_j >> 0$ the weights are constrained to be non-negative and sum to one. Because the weights are non-negative, the last penalty resembles a "lasso" or L_1 penalty, which permits weights to be zero. For the log-scoring rule we must have both sets of constraints for the MOMA density estimate to be a valid density.

To illustrate the behaviour of the solutions under quadratic loss, we focus on the case where $m = n - 1$ and let $\hat{e} = [\hat{\epsilon}_{kj}] = [y_k - \hat{y}_{(-k)\mathcal{M}_j}]$ denote the $n \times J$ matrix of predicted residuals for predicting y_k under model $\mathcal{M}_j$ using data $\mathbf{Y}_{(k)}$.

Remark 1: With the sum to one constraint alone, $\hat{\mathbf{w}} \propto (\hat{\mathbf{e}}^T \hat{\mathbf{e}})^{-1}\mathbf{1}$. If residuals from models are uncorrelated, then weights are proportional to the inverse of the Predicted REsidual Sum of Squares for model $\mathcal{M}_j$, $\text{PRESS}_j = \sum_k \hat{\epsilon}^2_{kj}$. With non-informative priors in linear models, $\text{PRESS}_j = \sum e^2_{kj}/(1 - h_{kk})$ where e_{jk} is the ordinary residual for case k under model j and h_{kk} is the leverage of case k, putting a premium on models that are able to fit points with high leverage. In the more general case of correlated predicted residuals across models, the weights are adjusted for the other models.

Remark 2: With highly correlated residuals under similar proxies, weights with just the sum to one constraint may be negative and highly unstable. The non-negativity constraint induces a lasso like L_1 penalty, which stabilizes weights and may drive the optimal weights to zero for redundant components.

Remark 3: The solution to (1.20) is in fact equivalent to the frequentist method of stacking (11; 10), although the Bayesian may prefer to include predictions under more robust prior distributions than using the non-informative prior distribution which leads to the least squares predictions.

1.5 Airplane Failures Examples

We compare the Dirichlet Process $\mathcal{M}$*–open* model averaging procedure of (37) and MOMA, which uses the predictive reuse approximation to the predictive distribution of future observations. The data are based on $n = 30$ intervals between air conditioning failure times for plane 7912 (30). Walker et al. (37) consider two models: an exponential

$$p_E(y) = \theta \exp(\theta y) \tag{1.25}$$

$$\theta \sim G(a, b) \tag{1.26}$$

and a log-normal model

$$p_L(y) = (2\pi\sigma^2)^{-1/2} \frac{1}{y} \exp\left[-\frac{1}{2}\left(\frac{\log(y) - \mu}{\sigma}\right)^2\right] \tag{1.27}$$

$$\mu \mid \sigma^2 \sim N(\mu_0, n_0\sigma^2) \tag{1.28}$$

$$(\sigma^2)^{-1} \sim G(\eta_0/2, \sigma_0^2\eta_0/2.) \tag{1.29}$$

Using non-informative improper prior distributions, traditional model averaging leads to indeterminate Bayes factors due to potentially arbitrary constants in the improper priors. That is not a problem with the $\mathcal{M}$–*open* approach. Using non-informative priors $a = b = 0$ and $\mu_0 = n_0 = \sigma_0 = 0, \eta_0 = -1/2$, to obtain the predictive densities for a future observation under each model, the MOMA weight for the exponential model is 0.435 using equation (1.24). In contrast the optimal weight from (37) is 0.315 for the exponential model. The predictive distributions for both approaches are illustrated in Figure 1.1. While the distributions are very similar in the tails, the main difference is near the mode.

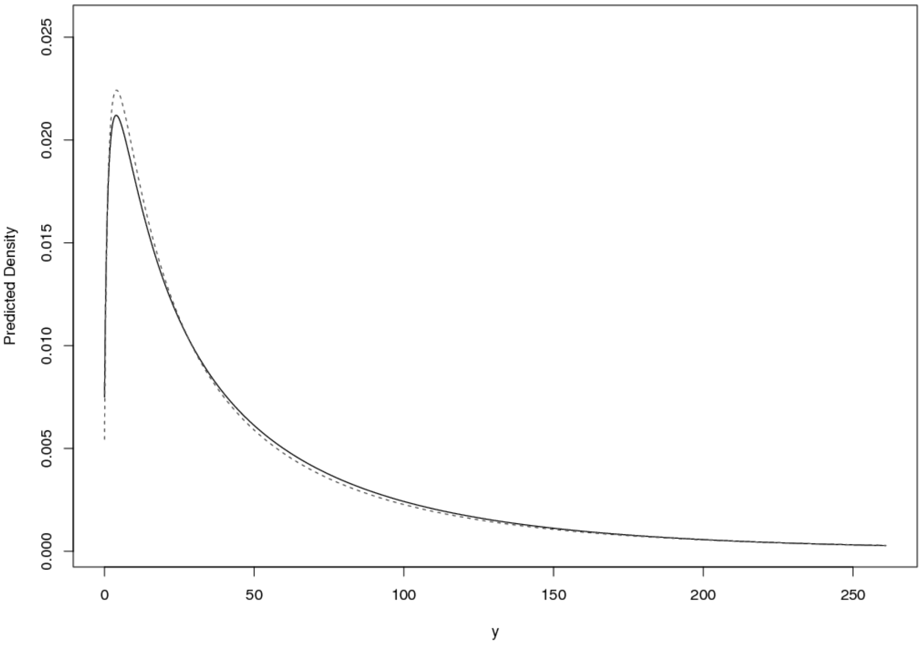

FIG. 1.1. Predicted distributions under MOMA (solid line) and the DP process estimate of (37) (dashed line).

To compare the two methods, we generated 30 observations from a $G(2,1)$ distribution and used Monte Carlo integration to evaluate the utility $\int \log \hat{p}(y \mid \mathbf{Y}) p_G(y) dy$ under the true Gamma model. Figure 1.2 illustrates one realization, where the utility under the MOMA estimate is -1.65 compared to -1.68 for the DP method. The estimate of the mixing weight for the exponential predictive distribution is 0.22 under MOMA while it is 2.2×10^{-13}, virtually zero, under the DP mixture. The differences between the two approaches is most pronounced in small samples, while the methods yield very similar results for larger sample

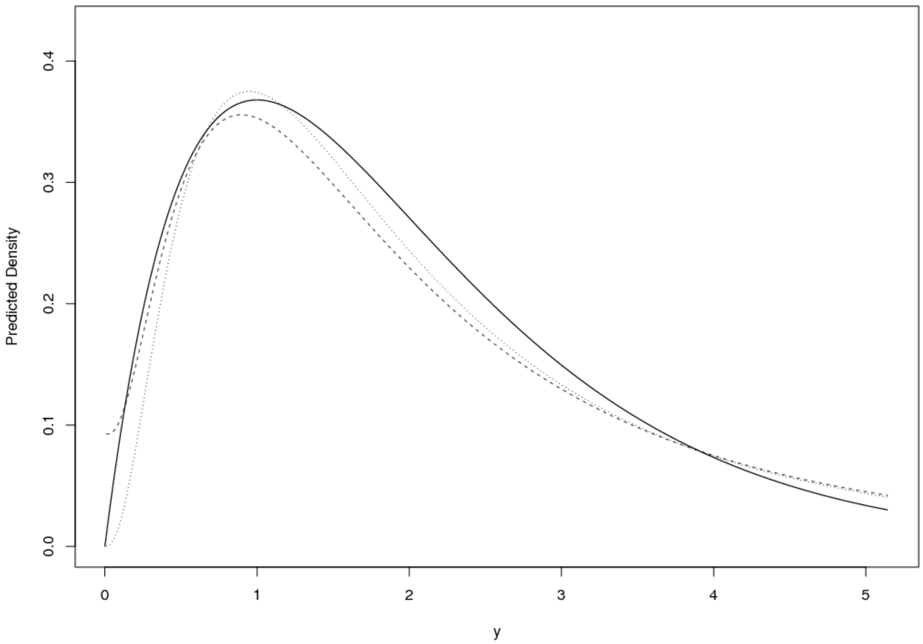

FIG. 1.2. Predicted distributions under MOMA (long dash) and the DP process estimate of (37) (short dashed line), with the true Gamma(2,1) distribution (solid line).

sizes.

1.6 Ovarian Cancer Example

Berchuck et al. (3) develop a model to predict binary survival status (short-term: < 3 years versus long-term: > 7 year) among patients diagnosed with advanced stage serous ovarian cancer using gene expression data from the primary tumor. The data consist of a retrospective sample with $n = 30$ short-term survivors, $n = 24$ long-term survivors and eleven early stage (I/II) cases. Expression was measured for $22,283$ targets using the Affymetrix U133a microarray. In addition to the tumor phenotype, six variables of clinical relevance (age, post-treatment CA125 levels, etc) were also collected for each women.

Based on the retrospective sampling design, the likelihood would be proportional to the joint distribution of the $22,283$ expression and 6 clinical variables given survival status. The sample size and the dimensionality of the problem preclude undertaking serious joint modeling, and instead several proxy models are used to develop predictive models. We consider three classes of models:

Clinical Trees (5 variants) Prospective Bayesian classification and regression tree models using only the six clinical variables;

Expression Trees (4 variants) Prospective Bayesian Classification and regression tree models using only expression data;

Expression LDA (4 variants) Retrospective discriminant models built using expression data given survival status.

Bayesian CART models using both clinical variables and expression variables lead to models that included only clinical variables, so this combination is excluded from the analysis as it would be redundant.

The clinical and expression tree models use a prospective likelihood and are based on Bayesian model averaged predictions from the Bayesian CART method of (29) with the different variants corresponding to different choices of the hyperparameter settings. The LDA discriminant model is based on classification with a retrospective model described in (25); predictions under this model are also model averaged. Because the data are retrospectively sampled, the prospective tree models provide simple proxies for prediction. The LDA method is a simple classification model that attempts to construct predictions assuming either conditional independence (labeled "P1") or a sparse dependence structure (labeled "P2") among the genes included in the analysis (either 100 or 200), and cannot reasonably be viewed as the true model. Traditional model averaging is not suitable for combining these predictions as there is no probability model that encompasses the two approaches employed in this example. Instead, we consider constructing optimal weights for MOMA estimates of the probability of being a long-term survivor under the quadratic loss criterion. This treats the 54 vectors of survival status, clinical data and expression data as being exchangeable in approximating the expected utility. A more realistic assumption would be that of partial exchangeability; given disease status the vectors of clinical and expression data are exchangeable. Population proportions of disease status could then be used for over or undersampling the across the disease groups to construct the predictive distribution.

Figure 1.3 illustrates the effects of the constraints on the weights. Predictions from the expression trees are quite similar to each other leading to high correlation in the predicted residuals (Figure 1.4, middle block). This is also true in the case of the clinical trees and the LDA models, although less so, while correlations between residuals from models in different classes are weakly correlated. High positive correlations lead to large positive and negative weights within a class of models which in effect cancel the contribution from these models to the overall prediction. Under the additional non-negativity constraint, all of the weights of the expression trees are 0, with the resulting MOMA predictions based predominantly on a subset of one clinical trees and two LDA models.

1.6.1 *Validation Experiment*

We evaluated the sensitivity of predictions to the form of constraint (sum-to-zero *versus* sum-to-zero and non-negativity) applied to the weight vector using

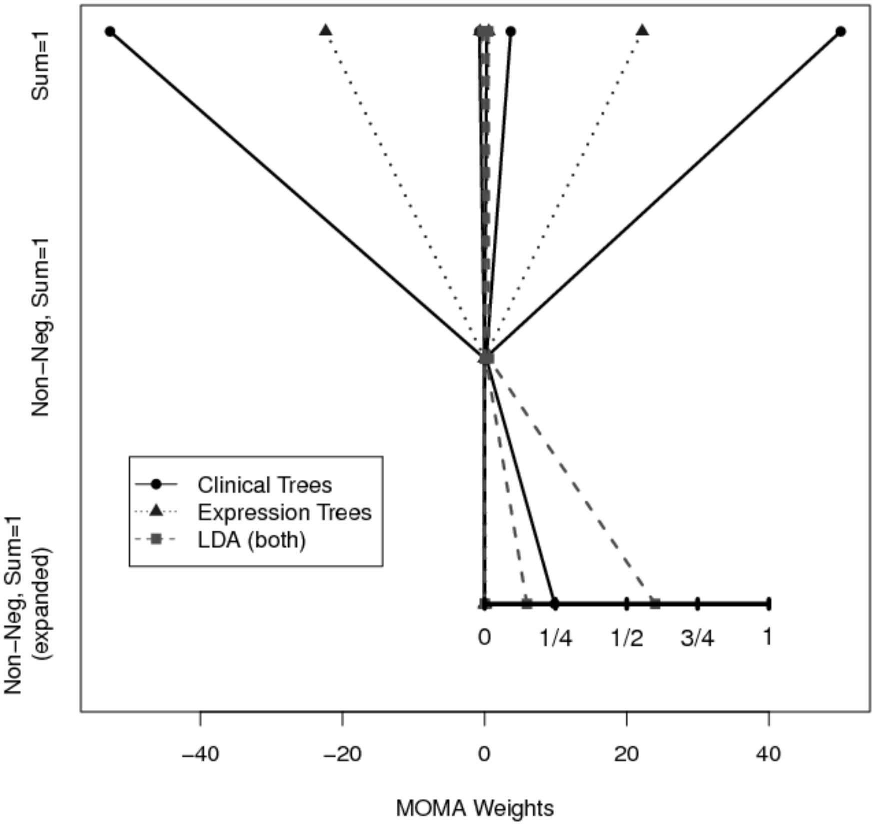

FIG. 1.3. MOMA weights under the (top) sum to one constraint and (middle and bottom) the sum to one and non-negativity constraint. Solid with circles denotes the Clinical trees, dotted line with triangles the expression trees and dashes with squares the LDA models.

5-fold cross-validation. We randomly split the data into a training set $\mathbf{Y}$ and a validation set $\mathbf{Y}^V$. Using the training data we obtained model weights $\hat{\mathbf{w}}$, using the Monte Carlo approximation to the expected utility in (1.20). We then constructed the MOMA estimates of the probability of long term survival $\hat{p}_j = \sum_i \hat{w}_i \hat{Y}^*_{\mathcal{M}_i}(\mathbf{Y})$ for the left out validation samples and classified an individual as a long-term survivor if $\hat{p}_j \geq 1/2$. Finally, we computed the classification accuracy for the validation set, then repeated the procedure for the remaining partitions of the data. Tables 1.1 and 1.2 illustrate the variability of the weights across

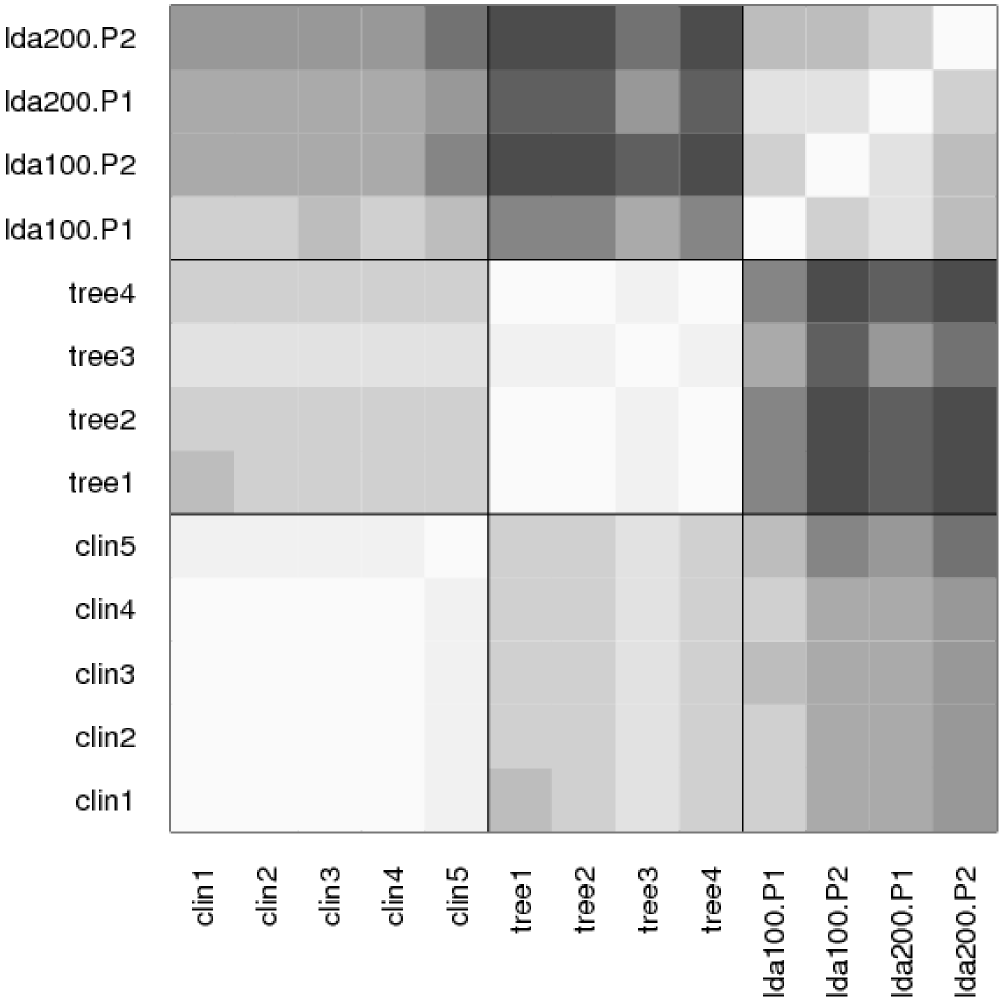

FIG. 1.4. Correlation of predicted residuals between the clinical trees (bottom 5), expression trees (middle 4) and LDA models (top 4); the range is 0.9327 - 0.9999 for clinical trees, 0.9257-0.9999 for expression trees, and 0.7257 - 0.8424 for LDA models.

different training sets.

Overall the accuracy of the MOMA with non-negative weights and the sum to zero constraint is better than any of the individual models. The expression data, through the LDA models, appear to improve the predictions over the clinical trees alone. A followup study by (2) confirmed association of the top genes from

	set1	set2	set3	set4	set5
clin1	53.08	−4.43	−0.01	−24.41	15.94
clin2	−79.92	−5.16	0.90	0.80	−4.63
clin3	−1.25	−0.24	−0.90	−0.01	5.35
clin4	27.36	10.14	−0.33	23.73	−17.24
clin5	1.13	0.27	0.27	0.36	0.55
tree1	−0.05	−0.55	−2.92	0.03	27.93
tree2	−0.12	−0.07	−3.21	−0.62	0.63
tree3	0.51	0.53	0.15	0.48	−3.35
tree4	−0.28	0.22	6.26	−0.04	−24.10
lda100.P1	−0.40	0.04	−0.01	0.02	−0.11
lda100.P2	0.44	−0.02	0.53	−0.06	−0.07
lda200.P1	0.30	0.17	−0.32	0.09	−0.03
lda200.P2	0.21	0.08	0.60	0.63	0.12
Accuracy	0.64	0.64	0.46	0.73	0.60

Table 1.1. MOMA weights and prediction accuracy for the ovarian cancer data for 5 randomly selected training and validation sets using the sum to one constraint.

	set1	set2	set3	set4	set5
clin1	0.00	0.07	0.00	0.00	0.00
clin2	0.00	0.00	0.00	0.00	0.00
clin3	0.00	0.00	0.00	0.00	0.00
clin4	0.00	0.11	0.00	0.00	0.00
clin5	0.30	0.17	0.07	0.41	0.00
tree1	0.00	0.00	0.00	0.00	0.77
tree2	0.00	0.00	0.00	0.00	0.21
tree3	0.23	0.44	0.21	0.00	0.01
tree4	0.00	0.00	0.00	0.00	0.01
lda100.P1	0.00	0.00	0.00	0.00	0.00
lda100.P2	0.22	0.00	0.30	0.00	0.00
lda200.P1	0.00	0.00	0.00	0.00	0.00
lda200.P2	0.26	0.21	0.41	0.58	0.00
Accuracy	0.82	0.73	0.55	0.73	0.60

Table 1.2. MOMA weights and prediction accuracy for the ovarian cancer data for 5 randomly selected training and validation sets using the non-negativity and sum to one constraints.

the LDA models with long-term survival status.

1.7 Discussion

Expanding on the original work of (9, 26) we have presented a method for model averaging in the $\mathcal{M}$–*open* setting using sample re-use methods to approximate the predictive distribution of future observations (19; 20). The solution of the mixing weights depends on the choice of utility functions as well as any constraints that are incorporated in the problem. The non-negativity constraint, which is natural, behaves like a lasso penalty and forces weights to be zero so that redundant predictions are not added. This is a striking difference to the traditional BMA solution where models with similar predictions often have similar marginal likelihoods, with the result that posterior mass is “diluted” over similar models (14; 22). MOMA using the log scoring rule is closely related to Ensemble BMA (31), which uses maximum likelihood to estimate the weights of a weighted average of bias corrected forecasts. The method of (37) uses a Dirichlet Process prior on the unknown distribution of the data to obtain the predictive distribution and leads to similar solutions to MOMA, although their method in the Gaussian case utilizes ordinary residuals, which may favor more complex models. The solutions in MOMA do not take into account the complexity of the models, however, the use of predicted residuals provides some penalization for poor out-of-sample prediction. If model complexity is an important criteria this can be incorporated as an additional constraint in the procedure.

An open question regards the selection of partitions of the data for MOMA. Rather than partition $\mathbf{Y}$ into $\{y_k, \mathbf{Y}_{(k)}\}$ where y_k is a single observation, (28) considered partitions $\{\mathbf{Y}_{-S}, \mathbf{Y}_S\}$ where $\mathbf{Y}_S$ is now a vector. In a simple but illustrative example, (28) showed that using the maximal training sample Y_S of dimension $n-1$ could lead to inconsistent selection of the true model, while use of the minimal training sample (dimension oone) had more desirable asymptotic properties. As MOMA and the DP method may be viewed as using training samples of size $n-1$ and n, asymptotic properties of model averaging in this framework is an area that needs additional reearch.

Unlike traditional BMA, MOMA provides a formal mechanism for combining predictions from models where the likelihoods are not commensurate. We have focused on linear combinations of the predictions and predictive distributions as motivated by traditional BMA. However, there is a rich literature on the related problem of combining expert opinions in risk analysis (with beliefs represented as distributions) (13). Rather than using the linear opinion pool (dating back to Laplace), alternative methods such as logarithmic pooling or Bayesian approaches may be useful in this context.

1.8 Acknowledgment

This work was supported by National Institute of Health grant 1-R01-HL-090559 and grant National Science Foundation DMS-1106891.

Bibliography

[1] Barbieri, Maria Maddalena and Berger, James O. (2004, June). Optimal predictive model selection. *Annals of Statistics*, **32**(3), 870–897.

[2] Berchuck, A., Iversen, E.S¿, Luo, J., Clarke, J.P., Levine, H. Horne D.A., Boyd, J., Alonso, M.A., Secord, A.A., Bernardini, M.Q., Barnett, J.C., Boren, T., Murphy, S.K., Dressman, H.K., Marks, J.R., and Lancaster, J.M. (2009). Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. *Clinical Cancer Research*, **15**, 2448–2455.

[3] Berchuck, A., Jr., E.S. Iversen, Lancaster, J.M., Pittman, J., Luo, J., Lee, P., Murphy, S., Dressman, H.K., Febbo, P.G., West, M., Nevins, J.R., and Marks, J.R. (2005). Patterns of gene expression that characterize long–term survival in advanced stage serous ovarian cancers. *Clinical Cancer Research*, **11**(10), 3686–3696.

[4] Berger, James O. and Pericchi, Luis R. (1996). The intrinsic Bayes factor for linear models. See (7), pp. 25–44.

[5] Berger, James O. and Pericchi, Luis R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

[6] Berger, James O. and Pericchi, Luis R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (ed. P. Lahiri), Volume 38 of *Lecture Notes in Statistics*, pp. 135–193. Institute of Mathematical Statistics, Hayward, CA.

[7] Bernardo, José Miguel, Berger, James O., Dawid, A. Phillip, and Smith, Adrian F.M. (ed.) (1996). *Bayesian Statistics 5*, Oxford, UK. Oxford Univ. Press.

[8] Bernardo, José Miguel, Berger, James O., Dawid, A. Phillip, and Smith, Adrian F.M. (ed.) (1999). *Bayesian Statistics 6*, Oxford, UK. Oxford Univ. Press.

[9] Bernardo, José M. and Smith, Adrian F. M. (1994). *Bayesian Theory.* Wiley, New York, NY.

[10] Breiman, Leo (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, **24**, 2350–2383.

[11] Breiman, Leo (1996). Stacked regressions. *Machine Learning*, **24**, 49–64.

[12] Chipman, Hugh A., George, Edward I., and McCulloch, Robert E. (2001). The practical implementation of Bayesian model selection. In *Model Selection* (ed. P. Lahiri), Volume 38 of *Lecture Notes in Statistics*, pp. 65–134. Institute of Mathematical Statistics, Hayward, CA.

[13] Clemen, Robert T. and Winkler, Robert L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, **19**(2), 187–203.

[14] Clyde, Merlise (1999). Bayesian model averaging and model search strategies (with discussion). See (8), pp. 157–185.

[15] Clyde, Merlise, DeSimone, Heather, and Parmigiani, Giovanni (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, **91**, 1197–1208.

[16] Clyde, Merlise and George, Edward I. (2004, May). Model uncertainty. *Statistical Science*, **19**(1), 81–94.

[17] Draper, David (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 45–70.

[18] Ferguson, Thomas S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**(4), pp. 615–629.

[19] Geisser, Seymour (1975). A predictive sample reuse method with application. *jasa*, **70**, 320–328.

[20] Geisser, Seymour (1993). *Predictive Inference: An Introduction.* Chapman & Hall, New York, NY.

[21] Geisser, Seymour and Eddy, William F. (1979). A predictive approach to model selection (Corr: V75 p. 765). *Journal of the American Statistical Association*, **74**, 153–160.

[22] George, Edward I. (1999). Discussion of "Model averaging and model search strategies" by M. Clyde. See (8).

[23] Gutirrez-Pea, E. and Walker, S.G. (2001). A bayesian predictive approach to model selection. *Journal of Statistical Planning and Inference*, **93**(1-2), 259 – 276.

[24] Hoeting, Jennifer A., Madigan, David, Raftery, Adrian E., and Volinsky, Chris T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, **14**(4), 382–401. Corrected version at `http://www.stat.washington.edu/www/research/online/hoeting1999.pdf`.

[25] Iversen, Edwin S. and Luo, Rosy J. (2003). Molecular and genetic modeling of disease risk. In *Proceedings of the American Statistical Association, Risk Section*, Alexandria, VA: American Statistical Association, pp. CD–ROM.

[26] Key, Jane T., Pericchi, Luis R., and Smith, Adrian F. M. (1999). Bayesian model choice: What and why? See (8), pp. 343–370.

[27] Leamer, Edward E. (1978). *Specification searches: Ad hoc inference with nonexperimental data.* Wiley, New York, NY.

[28] Mukhopadhyay, Nitai, Ghosh, Jayanta K., and Berger, James O. (2005). Some bayesian predictive approaches to model selection. *Statistics & Probability Letters*, **73**, 369–379.

[29] Pittman, Jennifer, Huang, Erich, Nevins, Joseph, Wang, Quanli, and West, Mike (2004). Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. *Biostatistics*, **5**(4), 587–601.

[30] Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375–383.

[31] Raftery, Adrian E., Gneiting, Tilmann, Balabdaoui, Fadoua, and Polakowski, Michael (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

[32] Raftery, Adrian E., Madigan, David, and Volinsky, Chris T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. See (7), pp. 323–349.

[33] San Martini, A. and Spezzaferri, Fulvio (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society, Series B*, **46**, 296–303.

[34] Walker, Stephen G. and Gutiérrez-Pena, Eduardo (1999). Discussion of Bayesian model choice: What and why? See (8), pp. 367.

[35] Walker, Stephen G. and Gutiérrez-Pena, Eduardo (1999). Discussion of Bayesian model choice: What and why? See (8), pp. 367.

[36] Walker, Stephen G. and Gutiérrez-Pena, Eduardo (1999). Robustifying Bayesian procedures. See (8), pp. 685–710.

[37] Walker, Stephen G., Guti/'errez-Pena, Eduardo, and Muliere, Peietro (2001). A decision theoretic approach to model averaging. *The Statistician*, **50**, 31–39.

[38] Wolpert, Robert L., Clyde, Merlise A., and Tu, Chong (2011). Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. *Annals of Statistics*, **39**, 1916–1962.