

Selection of Prior Weights for Weighted Model Averaging

Paul H. Garthwaite* and Emmanuel Mubwandarikwa*

Abstract. This paper addresses the task of choosing prior weights for models that are to be used for weighted model averaging. Models that are very similar to each other should usually be given smaller weights than models that are quite distinct. Otherwise, the importance of a model in the weighted average could be increased by augmenting the set of models with duplicates of the model or virtual duplicates of it. Similarly, the importance of a particular model feature (a certain covariate, say) could be exaggerated by including many models with that feature. Ways of forming a correlation matrix that reflects the similarity between models are suggested. Then weighting schemes are proposed that assign prior weights to models on the basis of this matrix. The weighting schemes have a dilution property: smaller weights are given to models that are more highly correlated. Other desirable properties in a weighting scheme are identified and we examine the extent to which these properties are held by the proposed methods. Also, the weighting schemes are applied to real data sets and prior weights, posterior weights and Bayesian model averages are determined. For these data sets, empirical Bayes methods were used to form the correlation matrices that yield the prior weights. Predictive variances are examined as empirical Bayes methods can result in unrealistically small variances.

Keywords: Bayesian model averaging, dilution priors, prior weights.

1 Introduction

The purpose of this paper is to propose methods of choosing prior weights for Bayesian model averaging and to examine the methods critically. Suppose models

* Department of Mathematics and Statistics, The Open University, Milton Keynes, UK, <mailto:p.h.garthwaite@open.ac.uk>

M_1, \dots, M_k each predict the value that an uncertain quantity Y will take. Rather than using model selection and choosing one of the M_i to make the prediction, an alternative is to form a weighted average $\sum w_i \hat{y}_i$, where the w_i are weights and \hat{y}_i is the prediction given by M_i . The Bayesian paradigm chooses the w_i as follows. Let p_i denote the prior weight attached to M_i . (If it were assumed that exactly one of the models is correct, then p_i would be the probability that M_i is correct.) If no data are available, then the weights are chosen by setting w_i equal to p_i for $i = 1, \dots, k$. If there are data, D say, then for $i = 1, \dots, k$,

$$w_i = \frac{p_i f_i(D | M_i)}{\sum_{j=1}^k p_j f_j(D | M_j)}, \quad (1)$$

where $f_i(D | M_i)$ is the probability or probability density function for the data if M_i were the correct model. In some circumstances, the influence of a prior distribution decreases as more data are gathered, so that the choice of prior distribution becomes less important as the quantity of data increases. Here, however, p_1, \dots, p_k have a multiplicative effect on the w_i so their importance does not dissipate as data are gathered. The concern of this paper is with weighting schemes for choosing them.

We should stress that we are not concerned with calculating Bayes factors [ratios of the form $f_i(D | M_i)/f_j(D | M_j)$]. These are needed to calculate the posterior weights (w_i) in equation (1), but determining them is a different task and it has been addressed by others (e.g. Aitkin, 1991; O'Hagan, 1995; Berger and Pericchi, 1996; Raftery et al., 1997).

The task of specifying prior weights for models has attracted some attention. One plausible choice is to give each model equal weight, putting $w_i = 1/k$ for each i . This is sometimes called the uniform prior. However, what if some models are very similar to each other while others are quite distinct? In regression analysis, for example, suppose that most models contain virtually the same set of explanatory variables while a few models contain very distinct choices of variables. Giving the models equal weight perhaps gives too much weight to the popular set of

explanatory variables. Similarly, if a set of models is expanded by adding a model that is virtually the same as one of the models already in the set, should the weight of the duplicated model be reduced? These issues were first raised by George (1999) in the discussion of a paper by Clyde (1999). He argued that if a set of models is increased by adding virtual duplicates of one model then the weight given to the original model should be divided between that model and its duplicates. He referred to this division of a prior weight as “dilution”.

Clyde agreed the importance of dilution and gave the following helpful illustration in her rejoinder to the discussion (Clyde, 1999, p. 180).

“It is not clear that the independent uniform prior on the model space is sensible. ... Suppose we start off with one [explanatory] variable, X_1 , and consider two models ($\{1\}$, $\{1, X_1\}$) where $\{1\}$ represents the model with just an intercept. Assign both models equal prior probabilities 0.5. Now consider adding a second variable X_2 that is highly correlated (or even perfectly correlated) with X_1 , with possible models ($\{1\}$, $\{1, X_1\}$, $\{1, X_2\}$, $\{1, X_1, X_2\}$) and uniform prior probabilities (0.25, 0.25, 0.25, 0.25). The total prior probability mass of the last three models is 0.75, while, if X_2 is really a proxy for X_1 , the mass should be closer to 0.5, as these three models are approximately equivalent (or exactly with perfect collinearity), and should have the same weight as in the original model space with just X_1 . The uniform prior has inflated the importance of X_1 while a more sensible prior should dilute the 0.5 mass over the three equivalent models, in order to be consistent with the first prior distribution. This inflation to the prior distribution carries over to the posterior distribution and hence has an effect on both model selection and model averaging.”

Other methods of determining prior weights have focused on variable selection problems in regression, where models differ in the explanatory variables that

they contain. One approach specifies the prior weight for M_i as

$$p_i = \prod_j \pi_j^{\delta_{ij}} (1 - \pi_j)^{1 - \delta_{ij}}, \quad (2)$$

where π_j is the prior probability that variable j should be included in the regression model, and δ_{ij} is an indicator variable that shows whether or not variable j is included in model M_i (see, for example, Hoeting et al., 1999). The values chosen for the δ_{ij} could vary if prior information about the importance of each variable is available. However, more commonly the same value is chosen for all δ_{ij} ; the resulting independent Bernoulli prior has proved useful in practice (Clyde and George, 2004). Chipman (1996) extends these priors to handle structured dependence between variables, giving smaller prior weights to models that are unlikely. For example, a small weight would be given to any model that contained an interaction term between two variables without containing either variable as a main effect, or to a model that contained higher order polynomial terms without the corresponding lower order terms. These weighting schemes do not explicitly address model duplication and so they suffer similar problems to those noted by Clyde (1999) and George (1999) for the uniform prior.

In work presented in conferences, George (2001, 2003) suggests various dilution priors that give reduced weight to models that are very similar. However, none of the methods are based on the correlation between models, which is perhaps the most obvious means of judging the similarity between models. Here we suppose there is a $k \times k$ correlation matrix that measures the degree of similarity between models M_1, \dots, M_k . It will be convenient to refer to the (i, j) element of the correlation matrix as *the correlation between models M_i and M_j* . We examine three different methods of choosing weights for the models on the basis of the correlation matrix. The methods are new, as a correlation matrix has not been used for this purpose before. They aim to avoid the difficulties noted by George (1999), Clyde (1999) and Chipman et al. (2001) by diluting the weights given to models that are very similar.

We examine features of the methods, focusing on the following properties that are desirable in a weighting scheme.

1. Models that are highly correlated with other models should be given smaller weights than those that have low correlation with other models.
2. Suppose a new model is added to the set of models and the new model is identical to a model or models already in the set. Then the weight previously given to the identical model(s) should be divided between those models and the new model, while the weights given to the other models should be unchanged.
3. When a new model is added to the set of models, weights of models already in the set will generally change, but none of them should increase.

We will refer to these properties as (1) *the dilution property*, (2) *the strong dilution property*, and (3) *the monotonicity property*.

In Section 2 we suggest approaches to forming correlation matrices that reflect the similarity between models. Three schemes for choosing weights are proposed in Section 3 and in Section 4 we examine their properties and give examples. Although it is difficult to construct weighting schemes that have the strong dilution property, two of them have this property. In Section 5 the weighting schemes are applied to two sets of real data. Prior weights and posterior weights are calculated and the predictive variances of Bayesian model averages are examined. Concluding comments are given in Section 6.

2 Forming Correlation Matrices

In order to apply weighting schemes proposed here, it must be possible to form correlation matrices between models. In this section ways this may be done are suggested. In principle, prior weights should reflect prior knowledge and could be based on expert opinion. The expert might be asked to specify the prior weights directly, but appropriate weights depend on the joint correlation matrix of all the

models. Hence, specifying the weights directly is a difficult task as it requires all the models to be considered simultaneously. In contrast, the task of specifying a correlation matrix can be broken down into simpler tasks, which may make it an easier task to perform. Also, if each model gives predictions for a response variable, then the correlation matrix might be defined as the correlation between predictions at, say, the sample means of the explanatory variables. The task of quantifying opinion about a set of predictions has been used to quantify opinion about a regression model and methods derived for that purpose, such as the method of Kadane et al. (1980), could readily be adapted to elicit the correlation matrix between models. After determining a correlation matrix to represent an expert's opinion, it would be sound practice to check that the expert is content that the prior weights derived from the matrix are a reasonable representation of his/her opinions.

Prior distributions are often chosen using some mechanical method, rather than basing them on expert opinion or background knowledge. In that spirit, the following strategy may be used to form prior weights, assuming sample data are available.

1. Determine the posterior distribution for each model, under the assumption that it is the true model.
2. Predict the response at each data point in the sample using each model.
3. Calculate the correlation matrix between the predictions of the models and take this as the correlation matrix between models.
4. Use the correlation matrix to assign prior weights to the models.

A drawback of this strategy is that the weighting schemes we propose penalize models for being highly correlated, and the predictions of good models will be correlated because the models are predicting the same quantity.

To counter this drawback, one possibility is to base the correlation matrix on deviations between models, as follows. In the above algorithm, expand step 2 to also calculate the average prediction of models at each data point and, for each model, determine the deviations between its predictions and the model averages. Then, in step 3 calculate the correlation matrix between the *deviations* of the models (rather than their predictions) and take this as the correlation matrix between models. In Section 5, where methods are applied to real data sets, the results from basing correlation matrices on both predictions and deviations are examined. We do not consider using residuals (as an alternative to deviations) because correlations between the residuals of different models are very sensitive to outliers. However, the choice of the correlation matrix remains a key open issue; no doubt there are alternative methods for its construction that others will prefer.

A criticism of the above type of strategy is that prior weights should really be specified before data are examined. Otherwise, there is a danger of ‘using the data twice’, since it will influence both the prior distribution and the likelihood. However, the prior weights are principally determined by the inter-relationships between models (represented by the correlation matrix), whereas the overall likelihood is obtained by determining separate likelihoods individually for each model. Thus the prior weights and the likelihood seem to reflect different features of the data that are only loosely related, so that overuse of the data may not occur to any great extent. As Cuzick (1991, p.135) notes, good empirical Bayes methods use the data twice and they are successful “. . . because the two uses are in a sense orthogonal.”

Examination of predictive variances can reveal if data are overused. Increasing the quantity of data typically results in more accurate predictions. Similarly, if data are ‘used twice’ or overused, predictive variances will reduce – spurious accuracy is a common potential problem with poor empirical Bayes methods. For example, Xu (2007, p.520) comments, “Because E-Bayes [an empirical Bayes method] is empirical, the data are overused, and as such the variances of the

estimated regression coefficients are anticonservative (smaller than they should be)”. In the examples that are examined in Section 5, we calculate predictive variances when prior weights are determined using our weighting schemes and compare them with the predictive variances obtained when all models are given equal prior weight. Two of our weighting schemes give variances that are virtually the same or marginally larger than the policy of equal weights, suggesting that they do not overuse the data and may be viewed as reasonable empirical Bayes methods.

Weighting schemes proposed by George (2001, 2003) are designed for regression problems. They assign prior weights to models solely on the basis of the explanatory variables used in each model. This avoids any problems of overuse of data. In principle, we could follow a similar approach and form a correlation matrix between models that is based on the explanatory variables the models contain. (An appropriate means of forming the correlation matrix would need to be devised.) However, we have not followed this approach because we believe that the similarity between models is determined not only by the explanatory variables that they contain but also by which explanatory variables affect the response. Suppose the explanatory variables in two models are independent apart from a single X -variable that is only contained in one of them. A tenable position is that the models are distinct automatically if this X -variable is not highly correlated with the other X -variables. However, we feel that the models should be regarded as distinct only if this X -variable is not highly correlated with the other X -variables *and its regression coefficient differs markedly from zero*. We can envisage situations where there are various plausible models but a scientist favors one of them in particular. Concerned that the favored model, though basically correct, may be missing some useful explanatory variable, the scientist might also propose variations of the favored model that differ from it by each including an additional explanatory variable. These additional variables may be almost uncorrelated with the explanatory variables in the favored model but, if they have little

effect on the response, we would argue that the additional models are essentially the same as the favored model. In such circumstances we advocate dilution of the weight given to the favored model, giving much of its weight to the additional models. It follows that we would rather not determine prior weights solely on the basis of the variables in models.

3 Weighting Schemes Based on the Correlation Matrix

3.1 Minimum Variance (MV) Weighting Scheme

Let $\hat{Y}_1, \dots, \hat{Y}_k$ denote predictions given by models, M_1, \dots, M_k . If predictions depend upon explanatory variables, we suppose that values have been specified for these variables and that predictions are conditional on these values. We let \mathbf{R} denote a (known) correlation matrix for the models and assume $\text{var}(\hat{Y}_1, \dots, \hat{Y}_k)' = c\mathbf{R}$, where c is a (possibly unknown) positive scalar that may depend on the values of the explanatory variables. We assume that \mathbf{R} is a positive definite matrix.

Weighted model averaging would form a linear combination of these predictions, $\sum_{i=1}^k w_i \hat{Y}_i$ with $\sum_{i=1}^k w_i = 1$ (cf. equation (1)), where the w_i weights are based on both the prior weights and Bayes factors. In principle, the prior weights should not depend on the Bayes factors or the quantity of sample data. Thus a good choice of prior weights should be good for any quantity of sample data. Our first weighting scheme takes this a stage further by seeking prior weights that are in some sense optimal as the quantity of data shrinks towards nothing while \mathbf{R} and $\hat{Y}_1, \dots, \hat{Y}_k$ are unchanged.

In the limit, when there are no data, the prior and posterior weights will be equal. Thus the first weighting scheme chooses prior weights p_1, \dots, p_k that in some sense optimize $\sum_{i=1}^k p_i \hat{Y}_i$ with $\sum_{i=1}^k p_i = 1$. As an optimality criteria, this weighting scheme aims to minimize the variance, $\text{var}(\sum_{i=1}^k p_i \hat{Y}_i)$, and we refer to it as the *MV (minimum variance) weighting scheme*. Putting $\mathbf{p} = (p_1, \dots, p_k)'$ gives $\text{var}(\sum_{i=1}^k p_i \hat{Y}_i) = c\mathbf{p}'\mathbf{R}\mathbf{p}$, so $\mathbf{p}'\mathbf{R}\mathbf{p}$ must be minimized subject to $\mathbf{p}'\mathbf{1} = 1$, where

$\mathbf{1}$ is a $k \times 1$ vector of ones. Hence, $\phi = \mathbf{p}'\mathbf{R}\mathbf{p} + \lambda(\mathbf{p}'\mathbf{1} - 1)$ must be minimized, where λ is a Lagrange multiplier. Equating $d\phi/d\mathbf{p}$ to zero gives $\mathbf{R}\mathbf{p} = \lambda\mathbf{1}$. This is a minimum (not a maximum) as $d^2\phi/d\mathbf{p}^2 = \mathbf{R}$ and \mathbf{R} is positive definite. Hence,

$$\mathbf{p} = \lambda\mathbf{R}^{-1}\mathbf{1}. \quad (3)$$

Let q_{ij} denote the (i, j) element of \mathbf{R}^{-1} and let $p_1^{MV}, \dots, p_k^{MV}$ denote the weights of the MV weighting scheme. From equation (3),

$$p_i^{MV} = \sum_{j=1}^k q_{ji} / \sum_{i=1}^k \sum_{j=1}^k q_{ji} \quad (4)$$

for $i = 1, \dots, k$. Although $\hat{Y}_1, \dots, \hat{Y}_k$ were used to motivate this weighting scheme, the weights are determined from the correlation matrix \mathbf{R} and do not require the existence of $\hat{Y}_1, \dots, \hat{Y}_k$.

The same weighting scheme arises from maximum likelihood if $\hat{Y}_1, \dots, \hat{Y}_k$ follow a multivariate normal distribution,

$$(\hat{Y}_1, \dots, \hat{Y}_k) \sim N(Y\mathbf{1}, c\mathbf{R}). \quad (5)$$

The expected value of each \hat{Y}_i is the scalar Y and we assume that this is the quantity to be estimated. From, for example, Mardia et al. (1979, equation 4.2.9), the maximum likelihood estimate of Y is

$$\hat{Y}_{ML} = \frac{\mathbf{1}'\mathbf{R}^{-1}(\hat{Y}_1, \dots, \hat{Y}_k)'}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}}. \quad (6)$$

It follows that

$$\hat{Y}_{ML} = \sum_{i=1}^k p_i^{MV} \hat{Y}_i, \quad (7)$$

with p_i^{MV} defined in equation (4), so the maximum likelihood estimate of the weights is identical to the MV estimate.

Thus the MV weighting scheme can be derived from at least two optimality criteria. Nevertheless, examples in Section 4.1 show that the MV method has flaws as a means of choosing prior weights.

3.2 Capped Eigenvalue (CE) Weighting Scheme

The other two weighting schemes that we propose are somewhat *ad hoc* in that they are not derived from any optimality criteria. However, examples given in later sections suggest that they work well in practice. The first of these weighting schemes is based upon the spectral decomposition of the correlation matrix \mathbf{R} :

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}', \quad (8)$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues and the columns of \mathbf{Q} are standardized eigenvectors. As each diagonal element of \mathbf{R} equals 1, eigenvalues greater than 1 indicate duplication of information in the models. (If the information in one model were not held in any other model, then \mathbf{R} would be a diagonal matrix with eigenvalues of 1.) The *capped eigenvalue (CE) method* forms a matrix $\mathbf{\Lambda}^*$ by changing any elements of $\mathbf{\Lambda}$ that exceed 1 to 1. Then the diagonal elements of $\mathbf{R}^* = \mathbf{Q}\mathbf{\Lambda}^*\mathbf{Q}'$ are normalized to sum to one and taken as the prior weights. Thus, if r_{ii}^* denotes the (i, i) element of $\mathbf{R}^* = \mathbf{Q}\mathbf{\Lambda}^*\mathbf{Q}'$, then the capped eigenvalue method gives M_i a prior weight of

$$p_i^{CE} = r_{ii}^* / \sum_{j=1}^k r_{jj}^* \quad \text{for } i = 1, \dots, k. \quad (9)$$

Motivation for this weighting scheme stemmed from the case where \mathbf{R} is a block-diagonal matrix. To illustrate, suppose

$$\mathbf{R} = \left(\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & \approx 1 & \approx 1 \\ 0 & 0 & \approx 1 & 1 & \approx 1 \\ 0 & 0 & \approx 1 & \approx 1 & 1 \end{array} \right),$$

where the off-diagonal elements in the second block are just less than 1 so that \mathbf{R} is positive-definite. The three models that give the second block are effectively duplicates of each other, as their correlations are almost 1. Common sense suggests that the combined weight given to them should equal the weight given to each of

the models in the first block. That is, the weights given to the five models should be $1/3, 1/3, 1/9, 1/9, 1/9$. The eigenvalues of \mathbf{R} approximately equal 3, 1, 1, 0, 0 (where the eigenvalue of 3 results from the second block), so the diagonal elements of $\mathbf{\Lambda}^*$ are approximately 1, 1, 1, 0, 0. Suppose we regard a model as containing one unit of information. Then $\text{trace}(\mathbf{\Lambda}) = \text{trace}(\mathbf{R})$ is a raw sum of the information in each of the models. Reducing the first eigenvalue from 3 to 1 accounts appropriately for the duplication in the second block: the information given by the three models in that block is only equivalent to the information given by one model. It is readily calculated that the diagonal elements of $\mathbf{R}^* = \mathbf{Q}\mathbf{\Lambda}^*\mathbf{Q}'$ are 1, 1, $1/3, 1/3, 1/3$, approximately, so comparing \mathbf{R} and \mathbf{R}^* , the only diagonal elements that have changed are those that relate to the second block. From these diagonal elements, the CE weighting scheme gives weights to the five models of $1/3, 1/3, 1/9, 1/9$ and $1/9$, consistent with common sense. Thus with block diagonal matrices the CE weighting scheme gives appropriate weights to the two extremes: those cases where models are independent (as with the first two models) and those where models are duplicated (as with the last three models).

More generally, $\text{trace}(\mathbf{\Lambda}^*)$ might be considered a measure of the quantity of effective information, where ‘effective’ means that duplicated information is counted only once. Hence, since $\text{trace}(\mathbf{\Lambda}^*) = \text{trace}(\mathbf{R}^*)$, basing weights on $\text{trace}(\mathbf{R}^*)$ has merit. Moreover, models that are highly correlated will load onto the same eigenvalues, causing those eigenvalues to exceed 1. Conversely, reducing those eigenvalues to 1 will have most impact on the models that load highly onto them. Thus the appropriate diagonal elements of \mathbf{R}^* are reduced most and basing weights on the diagonal of \mathbf{R}^* penalizes models for being correlated, which is the desired effect.

3.3 Cos-Square (CS) Weighting Scheme

To motivate this weighting scheme, suppose each model gives predictions at n design points. Let $\mathbf{y}_1, \dots, \mathbf{y}_k$ denote the vectors of predictions for the k models.

Each \mathbf{y}_i is standardized to give vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ that have means of zero and variances of one. Now $\mathbf{u}_1, \dots, \mathbf{u}_k$ may be considered as vectors in n -space. They are not orthogonal (unless $\mathbf{y}_1, \dots, \mathbf{y}_k$ are uncorrelated) and we ask, *What is the minimum amount by which we could adjust $\mathbf{u}_1, \dots, \mathbf{u}_k$ so that they form an orthogonal set?* If \mathbf{y}_i and \mathbf{y}_j are highly correlated, then \mathbf{u}_i and \mathbf{u}_j will point in similar directions, so \mathbf{u}_i and/or \mathbf{u}_j will need to be adjusted substantially if they are to become orthogonal. In contrast, if \mathbf{y}_i is almost uncorrelated with all the other \mathbf{y} -vectors, then \mathbf{u}_i will be almost orthogonal to the other \mathbf{u}_j ($j \neq i$), and \mathbf{u}_i will be adjusted very little. With this weighting scheme, the greater the adjustment to \mathbf{u}_i , the smaller the prior weight we will give to M_i , so highly correlated models are penalized. Suppose, \mathbf{u}_i is adjusted to become \mathbf{u}_i^* . Then $\mathbf{u}_i' \mathbf{u}_i^*$ is the cosine of the angle between \mathbf{u}_i and \mathbf{u}_i^* , and the smaller the adjustment, the larger $\mathbf{u}_i' \mathbf{u}_i^*$. The following *Cos-Square (CS) weighting scheme* will be shown to have good properties. It maximizes $\sum_{i=1}^k (\mathbf{u}_i' \mathbf{u}_i^*)^2$ and gives each M_i a weight proportional to $(\mathbf{u}_i' \mathbf{u}_i^*)^2$.

1. Put $\mathbf{U}\mathbf{A} = \mathbf{U}^* = (\mathbf{u}_1^*, \dots, \mathbf{u}_k^*)$. Choose \mathbf{A} so that $\sum_{i=1}^k (\mathbf{u}_i' \mathbf{u}_i^*)^2$ is maximized subject to $\mathbf{u}_1^*, \dots, \mathbf{u}_k^*$ being a set of standardized orthogonal vectors.
2. For $i = 1, \dots, k$, give model M_i the weight $p_i^{CS} = (\mathbf{u}_i' \mathbf{u}_i^*)^2 / \sum_{j=1}^k (\mathbf{u}_j' \mathbf{u}_j^*)^2$.

Our intention was to base weights on the correlation matrix, \mathbf{R} , so we must show that this weighting scheme may be expressed in terms of \mathbf{R} . In the present context, \mathbf{R} is the correlation matrix of $\mathbf{y}_1, \dots, \mathbf{y}_k$. Thus $\mathbf{R} = \mathbf{U}'\mathbf{U}$, where $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_k)$. The weighting scheme chooses \mathbf{A} so that the diagonal elements of $\mathbf{U}'\mathbf{U}^*$, when squared and added, form as large a total as possible, under the constraint that $(\mathbf{U}^*)'\mathbf{U}^*$ is the $k \times k$ identity matrix. Now, $\mathbf{U}'\mathbf{U}^* = \mathbf{U}'\mathbf{U}\mathbf{A} = \mathbf{R}\mathbf{A}$ and $(\mathbf{U}^*)'\mathbf{U}^* = \mathbf{A}'\mathbf{R}\mathbf{A}$. Hence, the optimal choice of \mathbf{A} depends only on \mathbf{R} . Also, $\mathbf{u}_i' \mathbf{u}_i^*$ is the i th diagonal element of $\mathbf{R}\mathbf{A}$ so $p_1^{CS}, \dots, p_k^{CS}$ are proportional to the squares of the diagonal elements of $\mathbf{R}\mathbf{A}$. An algorithm to obtain $p_1^{CS}, \dots, p_k^{CS}$ for a correlation matrix \mathbf{R} is given in an appendix.

4 Properties and Examples

4.1 The Dilution Property

Depending on the set of correlations, it may be unclear which models are more correlated than others, making it difficult to evaluate the extent to which a weighting scheme has the dilution property. Here we use examples in which there is a clear ordering.

In the first two examples, there are four models and the correlation matrices are \mathbf{R}_1 and \mathbf{R}_2 :

$$\mathbf{R}_1 = \begin{matrix} & \begin{matrix} M_1 & M_2 & M_3 & M_4 \end{matrix} \\ \begin{pmatrix} 1 & 0.4 & 0.3 & 0.1 \\ 0.4 & 1 & 0.2 & 0.1 \\ 0.3 & 0.2 & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 1 \end{pmatrix}, & \mathbf{R}_2 = \begin{matrix} & \begin{matrix} M_1 & M_2 & M_3 & M_4 \end{matrix} \\ \begin{pmatrix} 1 & 0.9 & 0.6 & 0.3 \\ 0.9 & 1 & 0.5 & 0.3 \\ 0.6 & 0.5 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{pmatrix} \end{matrix}$$

To indicate notation, in \mathbf{R}_1 the correlation between model M_1 and M_2 is 0.4, between M_1 and M_3 is 0.3, and so on. The correlations are quite low in \mathbf{R}_1 and quite high in \mathbf{R}_2 . It is obvious that for both \mathbf{R}_1 and \mathbf{R}_2

$$M_1 \succ M_2 \succ M_3 \succ M_4, \tag{10}$$

where $M_i \succ M_j$ means that M_i is more highly correlated with other models than is M_j . Table 1 shows the weights given to each model by the different weighting schemes for these two correlation matrices. All three weighting schemes give prior weights to models in the reverse order to the sequence in (9), with M_1 always receiving the smallest weight and M_4 the largest. Also, the differential in weights is greater with \mathbf{R}_2 than with \mathbf{R}_1 (i.e. the weights are more spread out with \mathbf{R}_2), reflecting the higher correlations in \mathbf{R}_2 . Hence, all three weighting schemes display the dilution property.

Table 1 also shows that the CE and CS methods give fairly similar weights to models while the weights from the MV method are more spread out. A disturbing

Weighting Scheme	\mathbf{R}_1				\mathbf{R}_2			
	M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4
MV Method	0.189	0.232	0.256	0.323	0	0.304	0.304	0.391
CE Method	0.226	0.237	0.251	0.286	0.192	0.206	0.262	0.340
CS Method	0.243	0.246	0.251	0.258	0.198	0.220	0.278	0.303

Table 1: Weights given to models M_1, \dots, M_4 by the MV, CE and CS weighting schemes. Correlations between the models are quite low with the correlation matrix \mathbf{R}_1 and quite high with \mathbf{R}_2 .

feature of the MV weighting scheme is that it sometimes gives models a prior weight of 0, as illustrated by the weight given to M_1 for \mathbf{R}_2 . A model with this prior weight will also have a posterior weight of 0, regardless of the data. Hence, a prior weight of 0 is only appropriate for models that are known *a priori* to have no predictive benefit.

Another feature of the MV weighting scheme that is even more disturbing (though more easily rectified) is that it can give a model a negative weight. For example, if the correlation matrix is

$$\mathbf{R}_3 = \begin{pmatrix} & M_1 & M_2 & M_3 \\ & 1 & 0.6 & 0.0 \\ & 0.6 & 1 & 0.6 \\ & 0.0 & 0.6 & 1 \end{pmatrix},$$

then the MV weights given to the three models are 0.667, -0.333 and 0.667, with M_2 receiving negative weight. The MV method could be modified to constrain weights to be non-negative, but the modification would not stop weights from being 0 (as in Table 1). Forcing weights to be positive, by insisting that they exceed some arbitrary positive threshold, begs the question of what value to choose for the threshold – the choice matters, as the MV method will commonly set some weights equal to the threshold.

4.2 The Strong Dilution Property

The MV weighting scheme obviously has the strong dilution property. Suppose a model M^* is added to the set of models and it is identical to one or more other models already in the set. As M^* provides no extra information, its addition does not reduce the variance of the minimum variance estimator or change the value of $\sum_{i=1}^k p_i \hat{Y}_i$. Hence, the weights previously given to models identical to M^* are divided arbitrarily between M^* and those models, while the weights given to the remaining models are unchanged.

Garthwaite et al. (2008) show that the strong dilution property is also held by the CS weighting scheme. The CE weighting scheme does not have the strong dilution property. This is most easily shown through an example. Suppose we take the correlation matrix \mathbf{R}_2 (considered earlier) and add a fifth model, M_5 , that is identical to M_4 . The resulting correlation matrix is:

$$\mathbf{R}_2^+ = \begin{pmatrix} & M_1 & M_2 & M_3 & M_4 & M_5 \\ & 1 & 0.9 & 0.6 & 0.3 & 0.3 \\ & 0.9 & 1 & 0.5 & 0.3 & 0.3 \\ & 0.6 & 0.5 & 1 & 0.3 & 0.3 \\ & 0.3 & 0.3 & 0.3 & 1 & 1.0 \\ & 0.3 & 0.3 & 0.3 & 1.0 & 1 \end{pmatrix}.$$

The CE weighting scheme gives M_1, \dots, M_5 the weights 0.182, 0.194, 0.246, 0.189 and 0.189, respectively. Comparison with Table 1 shows that M_1 , M_2 and M_3 now have slightly lower weights. M_4 had a weight of 0.340, which is a little less than the combined weight now given to M_4 and M_5 ($0.189 + 0.189 = 0.378$). For \mathbf{R}_2^+ , the MV method gives weights of 0, 0.304, 0.304, α_1 and $0.391 - \alpha_1$, so the weight previously given to M_4 is divided between M_4 and M_5 . Similarly, the CS method gives weights of 0.198, 0.220, 0.278, α_2 and $0.303 - \alpha_2$, so again the weight previously given to M_4 is divided between M_4 and M_5 .

Extending this example, suppose model M_2 is also replicated and that it is

replicated three times. With both the MV and CS methods, the weights given to M_1 , M_2 , M_4 and M_5 are unchanged while M_2 and its replicates share the weight formally given to M_2 . The way that the weight of M_2 is divided between M_2 and its copies is arbitrary except that weights must be non-negative with the CS method. With the CE method, M_2 and its replicates receive equal weights that total 0.303, much greater than the weight of 0.194 that was given to M_2 before its replication. Other models receive correspondingly less weight, with M_1 having the biggest reduction (from 0.182 to 0.101) because of its high correlation with M_2 .

4.3 The Monotonicity Property

The monotonicity property holds if no model ever has its weight increased when another model is added to the set of models. Through examples it is shown that none of the three weighting schemes have this property. Specifically, for each weighting scheme we give an example where one of the models gains weight when the set of models is increased by adding a model M_0 . First, consider the following correlation matrices:

$$\mathbf{R}_4 = \begin{pmatrix} & M_1 & M_2 & M_3 \\ & 1 & 0.55 & 0.0 \\ & 0.55 & 1 & 0.55 \\ & 0.0 & 0.55 & 1 \end{pmatrix}, \quad \mathbf{R}_4^+ = \begin{pmatrix} & M_0 & M_1 & M_2 & M_3 \\ & 1 & 0.4 & 0.0 & 0.4 \\ & 0.4 & 1 & 0.55 & 0.0 \\ & 0.0 & 0.55 & 1 & 0.55 \\ & 0.4 & 0.0 & 0.55 & 1 \end{pmatrix}.$$

The MV weighting scheme gives model M_2 a weight of -0.125 for \mathbf{R}_4 and a weight of +0.129 for \mathbf{R}_4^+ . As the only difference between the correlation matrices is that the model M_0 has been added to \mathbf{R}_4 to form \mathbf{R}_4^+ , adding a model has increased the weight given to M_2 . Hence the MV weighting scheme does not have the monotonicity property.

Let \mathbf{I}_h denote the $h \times h$ identity matrix and let $\mathbf{0}_h$ denote a $h \times 1$ vector of zeros. For the CE weighting scheme we take the correlation matrices \mathbf{R}_4 and \mathbf{R}_4^+

and add a further h models $M_4 \dots M_{3+h}$ that are each uncorrelated with all other models. The resulting correlation matrices are:

$$\mathbf{R}_5 = \begin{pmatrix} M_1 & M_2 & M_3 & \dots & M_{3+h} \\ 1 & 0.55 & 0.0 & & \mathbf{0}'_h \\ 0.55 & 1 & 0.55 & & \mathbf{0}'_h \\ 0.0 & 0.55 & 1 & & \mathbf{0}'_h \\ \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & & \mathbf{I}_h \end{pmatrix}, \quad \mathbf{R}_5^+ = \begin{pmatrix} M_0 & M_1 & M_2 & M_3 & \dots & M_{3+h} \\ 1 & 0.4 & 0.0 & 0.4 & & \mathbf{0}'_h \\ 0.4 & 1 & 0.55 & 0.0 & & \mathbf{0}'_h \\ 0.0 & 0.55 & 1 & 0.55 & & \mathbf{0}'_h \\ 0.4 & 0.0 & 0.55 & 1 & & \mathbf{0}'_h \\ \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & & \mathbf{I}_h \end{pmatrix}.$$

The weights given to M_2 by the CE weighting scheme are $0.611/(h+2.22)$ for \mathbf{R}_5 and $0.685/(h+3.04)$ for \mathbf{R}_5^+ . If $h \geq 5$, then $0.611/(h+2.22) < 0.685/(h+3.04)$ and the weight given to M_2 has increased by adding M_0 to the models M_1, \dots, M_{3+h} . Hence the CE weighting scheme does not have the monotonicity property either.

To show that the CS method does not have the monotonicity property we use the correlation matrices,

$$\mathbf{R}_6 = \begin{pmatrix} M_1 & M_2 & M_3 & \dots & M_{3+h} \\ 1 & 0.6 & 0.0 & & \mathbf{0}'_h \\ 0.6 & 1 & 0.7 & & \mathbf{0}'_h \\ 0.0 & 0.7 & 1 & & \mathbf{0}'_h \\ \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & & \mathbf{I}_h \end{pmatrix}, \quad \mathbf{R}_6^+ = \begin{pmatrix} M_0 & M_1 & M_2 & M_3 & \dots & M_{3+h} \\ 1 & 0.0 & 0.7 & 0.6 & & \mathbf{0}'_h \\ 0.0 & 1 & 0.6 & 0.0 & & \mathbf{0}'_h \\ 0.7 & 0.6 & 1 & 0.7 & & \mathbf{0}'_h \\ 0.6 & 0.0 & 0.7 & 1 & & \mathbf{0}'_h \\ \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & \mathbf{0}_h & & \mathbf{I}_h \end{pmatrix}.$$

The CS method gives M_1 a weight of $0.8854/(h+2.376)$ for \mathbf{R}_6 and a weight of $0.9076/(h+2.916)$ for \mathbf{R}_6^+ . If $h \geq 20$, then $0.8854/(h+2.376) < 0.9076/(h+2.916)$ and the weight given to M_1 is increased by adding M_0 to the set of models. Hence the CS weighting scheme also does not have the monotonicity property.

It should be mentioned that it was relatively easy to find examples that showed the MV method does not have the monotonicity property. Examples that show the CE and CS methods do not have this property were noticeably more difficult to find.

5. Examples with Real Data

The Boston house-price data of Harrison and Rubinfeld (1978) were used to examine the weighting schemes with real data. The data have been widely used to examine a variety of regression methods. There are 506 observations and thirteen explanatory variables. The quantity to be predicted is the median value of owner-occupied houses in a district. Here, the logarithm of this quantity was taken as the dependent variable. Eight linear regression models were fitted that used different subsets of the explanatory variables and simple transformations of them. (The transformations were the natural logarithm and the square of a variable.) The variables included in each model are shown by asterisks in Table 2.

The correlation between predictions given by the different models could be used as the correlation matrix on which to base weights. However, as noted earlier, it can be argued that models *should* be correlated because they all predict the same quantity, so penalizing those that are highly correlated might be considered inappropriate. As an alternative, we also determined the average prediction for each observation and calculate the deviations between a model's predictions and these averages. The correlations between these deviations was taken as a second correlation matrix for models.

The following matrix shows the two sets of correlations.

$$\begin{array}{cccccccc}
 M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 & M_8 \\
 \left(\begin{array}{cccccccc}
 1 & 0.9998 & 0.9366 & 0.9544 & 0.9541 & 0.721 & 0.892 & 0.99999 \\
 0.9946 & 1 & 0.9368 & 0.9546 & 0.9544 & 0.721 & 0.892 & 0.9998 \\
 -0.067 & -0.064 & 1 & 0.9677 & 0.9680 & 0.639 & 0.756 & 0.9365 \\
 -0.186 & -0.184 & 0.583 & 1 & 0.9997 & 0.664 & 0.851 & 0.9544 \\
 -0.183 & -0.180 & 0.587 & 0.9945 & 1 & 0.662 & 0.851 & 0.9541 \\
 -0.598 & -0.600 & -0.354 & -0.587 & -0.596 & 1 & 0.736 & 0.721 \\
 -0.220 & -0.219 & -0.795 & -0.417 & -0.413 & 0.249 & 1 & 0.892 \\
 0.9997 & 0.9920 & -0.067 & -0.185 & -0.183 & -0.597 & -0.220 & 1
 \end{array} \right) .
 \end{array}$$

Variable	Model							
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Intercept	*	*	*	*	*	*	*	*
Rm							*	
Rm ²	*	*		*	*			*
Age	*	*						*
Dis						*		
ln(Dis)	*	*	*					*
Rad			*					
ln(Rad)	*	*						*
Tax	*	*		*	*			*
PTRatio	*	*		*	*		*	*
B	*	*					*	*
ln(LStat)	*	*	*	*	*			*
Crim	*	*				*	*	*
Zn	*	*						*
Indus	*	*				*		*
ln(Indus)			*					
Chas	*	*						*
NOx		*	*					*
NOx ²	*			*				*

Table 2: The explanatory variables in eight linear regression models fitted to the Boston housing data. An asterisk indicates the variable was included in the model.

Values above the main diagonal of the matrix are based on predictions and those below the diagonal are based on the deviations between predictions and average predictions. The former are mostly above 0.9 while the latter vary more, with many of them negative. There is an extremely strong correlation between M_1 and M_8 . These two models are also highly correlated with M_2 , but to a lesser degree. The other models that are very highly correlated are M_4 and M_5 .

Method	Model							
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
<i>Using predictions</i>								
<i>Without M_8</i>								
MV weight	-0.120	-0.765	1.428	-0.513	-0.180	0.359	0.791	
CE weight	0.102	0.101	0.134	0.109	0.109	0.276	0.169	
CS weight	0.067	0.064	0.194	0.071	0.066	0.303	0.235	
<i>With M_8</i>								
MV weight	-7.22	-0.954	1.450	-0.533	-0.179	0.357	0.800	7.279
CE weight	0.085	0.085	0.122	0.098	0.098	0.269	0.158	0.085
CS weight	0.038	0.055	0.194	0.071	0.066	0.303	0.234	0.039
<i>Using deviations</i>								
<i>Without M_8</i>								
MV weight	0.172	0.068	0.133	0.065	0.128	0.262	0.173	
CE weight	0.126	0.126	0.163	0.122	0.122	0.144	0.196	
CS weight	0.111	0.132	0.188	0.124	0.107	0.140	0.198	
<i>With M_8</i>								
MV weight	2.601	-0.351	0.136	0.213	-0.021	0.259	0.176	-2.014
CE weight	0.090	0.091	0.162	0.120	0.119	0.133	0.194	0.090
CS weight	0.032	0.123	0.186	0.125	0.104	0.137	0.197	0.095

Table 3: Weights given by the MV, CE and CS methods to regression models fitted to the Boston house-price data. The table shows the effect of adding model M_8 to the set of models. Correlation matrices were based either on the correlations between predictions or between deviations from the model average.

The MV, CE and CS weighting schemes were used to derive weights from both the correlation matrix based on predictions and, separately, the correlation matrix based on deviations. So as to examine the effect of adding a model to a set of models, weights were first determined when the set of models was M_1, \dots, M_7 ,

and then for M_1, \dots, M_8 . Results are presented in Table 3. All three methods display the dilution property with the least correlated models, M_3 , M_6 and M_7 being given the highest weights. The table also shows that the MV method can give negative weights in a real example. The CE and CS methods give reasonably similar weights, especially when the correlation matrix is based on deviations from the average.

When the eighth model was added to the set of seven models, the combined weight given to M_1 , M_2 and M_8 by the CS method was about the same as the weight it previously gave to M_1 and M_2 . This was anticipated as the CS method has the strong dilution property. In contrast, the CE method gave M_1 , M_2 and M_8 a combined weight that is noticeably bigger than the weight it gave to M_1 and M_2 before the inclusion of M_8 . (The difference is 25% when correlations are based on predictions.) A result that will hold more generally is that the CE method gave highly correlated models very similar weights whereas the CS method differentiated between them, even when all correlations were close to 1. This can be seen clearly in the weights given to M_1 , M_2 and M_8 . Although they do not have the monotonicity property, both the CE and CS methods gave weights to models that were no greater after inclusion of M_8 than before its inclusion.

Table 3 also shows that it makes a difference whether the correlation matrix is based on predictions or deviations. For example, M_6 receives much lower weight when it is based on deviations. This happens because M_6 gives very different predictions to other models (its predictions are poorer). Examination of Table 2 shows that the models containing similar sets of variables are M_1 , M_2 and M_8 , and also M_4 and M_5 . To the writers these similarities are better reflected by the correlations between deviations from average predictions, rather than the predictions themselves. Hence we advocate basing the correlation matrix on deviations from average predictions.

We also examined the influence of the prior weights on Bayesian model averages. Our primary aim is to see how predictive variances vary with the choice

of prior weighting method. If a method is overusing the data or using it twice, it would be expected to yield smaller predictive variances than when equal prior weights are used.

For Bayesian model averaging, prior distributions for model parameters must be specified and these distributions must not be improper. We use a form of prior distribution for regression models that Raftery, Madigan and Hoeting (1997) designed for Bayesian model averaging. The priors are proper and reasonably flat over the range of parameter values that could plausibly arise. Likelihoods were calculated for each model and then scaled to sum to 1. M_1 , M_2 and M_8 fitted the data much better than the other five models: their scaled likelihoods were 0.526, 0.356 and 0.118, respectively, while those of the other models were 0.000. Seven sets of prior weights are of interest: equal weights and the sets of weights for M_1, \dots, M_8 given in Table 3. With the MV method, some prior weights were negative. These were set equal to 0 and the other weights were re-scaled to add to 1. Unfortunately, this resulted in the MV method giving only one model (either M_1 or M_8) a weight greater than 0.001.

Each set of prior weights was taken in turn and combined with likelihoods to obtain posterior weights, using equation (1). Posterior distributions for each model were determined and used to predict the log house price for each observation in the data set. Predictive variances were then calculated. Let \hat{Y}_{ij} denote the prediction given by M_i for the j th observation and let $\text{var}(\hat{Y}_{ij})$ denote the associated predictive variance. If w_1, \dots, w_8 are the posterior weights of the eight models, then the predictive variance given by Bayesian model averaging is

$$\text{var}(\hat{Y}_j^*) = \sum_{i=1}^8 w_i \text{var}(\hat{Y}_{ij}) + \sum_{i=1}^8 w_i (\hat{Y}_{ij} - \hat{Y}_j^*)^2, \quad (11)$$

where $\hat{Y}_j^* = \sum_{i=1}^8 w_i \hat{Y}_{ij}$. The average of these variances over the data set ($\sum_{j=1}^{504} \text{var}(\hat{Y}_j^*)/504$) was determined for each method of forming prior weights. For equal prior weights this average was 0.03354. Average predictive variances for the other prior weighting methods are given in Table 4. The table also gives the

Basis of correlation matrix	Prior Weighting Method		
	MV	CE	CS
Predictions	0.03355(37.4)	0.03354(81.5)	0.03354(97.2)
Deviations	0.03349(0.0)	0.03354(97.4)	0.03356(98.6)

Table 4: Average predictive variances from Bayesian model averaging for the Boston house-price data. Figures in brackets are the percentage of observations for which the prior weighting method gave a larger predictive variance than that given by equal prior weights.

percentage of observations for which a predictive variance was larger when using one of our weighting methods than when using equal prior weights. Differences in average predictive variance are very slight because M_1 , M_2 and M_8 give very similar predictions and other models have posterior weights that are virtually 0. However the directions of differences have clear traits. For the great majority of observations the CE and CS methods gave larger predictive weights than equal prior weights, suggesting that these methods do not overuse the data. Little can be inferred from the results for the MV method as its results reflect the predictions of a single model, rather than a model average.

In the last analysis, model averages were essentially based on three models or fewer. Another data set was used to examine predictive variances when a larger number of models play a significant role in forming averages. The data set gives crime rates in 47 U.S. states in 1960 and the values of 15 explanatory variables that are candidate predictors of crime rate. It has been widely analyzed and is given in full by Vandaele (1978). The data were used by Raftery, Madigan and Hoeting (1997, hereafter RMH) to test two methods of Bayesian model averaging. They called these methods MC^3 and *Occam's razor*. RMH list the ten models

Interval	MC^3 models						Occam's razor models					
	Predictions			Deviations			Predictions			Deviations		
	MV	CE	CS	MV	CE	CS	MV	CE	CS	MV	CE	CS
0–0.019	7	0	0	2	0	1	14	6	4	14	7	8
0.020–0.049	0	1	3	0	1	1	3	7	11	3	7	5
0.050–0.099	0	4	1	3	6	3	2	8	6	2	7	5
0.100–0.199	0	5	5	4	2	4	1	1	1	1	1	4
0.200–0.499	3	0	1	1	1	1	2	0	0	2	0	0

Table 5: Frequency distributions of posterior model weights for the crime rate data, using predictions or deviations from average prediction to derive the correlation matrix.

that MC^3 identified as ‘best’. Occam’s razor selected 22 models as useful and these are all listed by RMH. With both methods RMH only considered using prior weights that are equal.

The same methods that we used for the Boston housing data were applied to the crime rate data, first using the ten models given by MC^3 and then using the 22 models given by Occam’s razor. Posterior weights were calculated and Bayesian model averages determined. The frequency distributions of the posterior weights are given in Table 5 and show that, with the CE and CS methods, several models make a significant contribution to each model average. Predictive variances were calculated at each of the 47 data points. For equal prior weights, the average predictive variances was 0.04349 for MC^3 and 0.04785 for Occam’s razor. Average predictive variances for the other weighting methods are given in Table 6 and are generally slightly larger. The one exception occurs when the MV method led to a model average that effectively involved only 3 models (c.f. Table 5). Table 6 also gives the percentages of observations for which a predictive variance was larger when using one of our weighting methods than when using equal prior weights. The table gives no suggestion that the CE and CS methods overuse the data through using it to form prior weights: average predictive variances are

Basis of correlation matrix	Prior Weighting Method		
	MV	CE	CS
<i>MC³ models</i>			
Predictions	0.04119(6.4)	0.04383(95.7)	0.04539(100.0)
Deviations	0.04418(85.1)	0.04352(38.3)	0.04362(46.8)
<i>Occam's razor models</i>			
Predictions	0.04824(70.2)	0.04838(100.0)	0.04959(97.9)
Deviations	0.04824(51.1)	0.04824(74.5)	0.04814(66.0)

Table 6: Average predictive variances from Bayesian model averaging for the crime rate data. Figures in brackets are the percentage of observations for which the prior weighting method gave a larger predictive variance than that given by equal prior weights.

greater with these methods than with equal prior weights and the same is true of individual predictive variances at the majority of data points.

6 Concluding Comments

Three weighting schemes have been considered here, the MV, CE and CS methods. The MV method does not seem viable as it can yield prior weights that are zero or negative. This was unexpected as it is derived from optimality criteria that are useful in other circumstances. In contrast, the CE and CS methods are novel but they appear to have given sensible prior weights in all the examples we examined. When some models are highly correlated, these methods seem a better way of assigning prior weights than the common approach of giving all models equal weight.

Choosing between the CE and CS methods is not simple. The CS method might be preferred because it has the strong dilution property, but another clear

difference between the two methods is in the weights they give to models that are very highly correlated. As shown with the Boston house-price data, the CE method gives very similar weights to models that are almost identical, while the CS method can give them noticeably different weights, depending on their correlations with other models. Which of these qualities is to be preferred may depend on the application and the number of models under consideration.

As noted earlier, Chipman (1996) suggests ways of choosing prior probabilities for models to reflect their model structure. For example, his method would give a small probability to an implausible model that contained an interaction term for two variables without containing either variable as a main effect. Allocating weights to models on the basis of model structure or the correlations between models should not be alternatives: weighting schemes are needed that simultaneously take account of both aspects. A promising possibility is to modify the CS method. Suppose t_1, \dots, t_k are the prior probabilities given to M_1, \dots, M_k purely on the basis of their model structures. Then rather than maximise $\sum_{i=1}^k (\mathbf{u}'_i \mathbf{u}_i^*)^2$ as for the unmodified CS method (cf. Section 3.3), $\sum_{i=1}^k t_i (\mathbf{u}'_i \mathbf{u}_i^*)^2$ might instead be maximized. The t_i might also be chosen to reflect the number of variables in a model, so as to favor smaller models.

Further research is also needed into ways of forming correlation matrices to reflect the similarity between models. For the CE and CS methods, it seems reasonable to use empirical Bayes methods for this task – with the two data sets we analyzed, predictive variances were not reduced through using the data to form prior weights. We believe this result is likely to hold for most methods of forming dilution priors, as giving more weight to less similar models and less weight to more similar models is akin to putting more weight in the tails of a distribution, and hence likely to lead to larger variances. The result is important, as it opens the door to using data when forming dilution priors.

Supplemental Material: An Algorithm to Calculate Weights for the Cos-Square Weighting Scheme

The following algorithm determines the weights $p_1^{CS}, \dots, p_k^{CS}$ corresponding to a $k \times k$ correlation matrix \mathbf{R} . Theory underlying the method is given in Garthwaite et al. (2008).

1. Set \mathbf{D}_1 equal to the $k \times k$ identity matrix and put $i = 1$.
2. At the i th iteration, perform a spectral decomposition of $\mathbf{D}_i \mathbf{R} \mathbf{D}_i$, giving

$$\mathbf{D}_i \mathbf{R} \mathbf{D}_i = \mathbf{Q}_i \mathbf{\Lambda}_i \mathbf{Q}_i',$$

where $\mathbf{\Lambda}_i$ is a diagonal matrix whose non-zero elements are the eigenvalues of $\mathbf{D}_i \mathbf{R} \mathbf{D}_i$ and the columns of \mathbf{Q}_i are the corresponding eigenvectors. [For $j = 1, \dots, k$, the j th column of \mathbf{Q}_i is the eigenvector corresponding to eigenvalue in the (j, j) element of $\mathbf{\Lambda}_i$.]

3. Put $\mathbf{E}_i = \mathbf{D}_i^{-1} \mathbf{Q}_i \mathbf{\Lambda}_i^{1/2} \mathbf{Q}_i'$, where $\mathbf{\Lambda}_i^{1/2}$ is a diagonal matrix and $\mathbf{\Lambda}_i^{1/2} \mathbf{\Lambda}_i^{1/2} = \mathbf{\Lambda}_i$.
4. Set \mathbf{D}_{i+1} equal to a diagonal matrix, with diagonal equal to the diagonal of \mathbf{E}_i .
5. Return to step (2) until convergence, when $\mathbf{D}_{i+1} \approx \mathbf{D}_i$.
6. Let p_1, \dots, p_k denote the diagonal elements of \mathbf{D}^* , where \mathbf{D}^* is the value of \mathbf{D}_i at convergence. For $i = 1, \dots, k$ put $p_i^{CS} = p_i^2 / \sum_{j=1}^k p_j^2$.

References

- Aitkin, M. (1991). "Posterior Bayes factors (with discussion)." *Journal of the Royal Statistical Society, Series B*, 53: 111–142.
- Berger, J. O. and Pericchi, L. R. (1996). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122.

- Chipman, H. A. (1996). “Bayesian variable selection with related predictors.” *Canadian Journal of Statistics*, 24: 17–36.
- Chipman, H. A., George, E.I. and McCulloch, R. E. (2001). “The practical implementation of Bayesian model selection (with discussion).” In Lahiri, P. (ed.), *Model Selection*, 65–134. IMS: Beachwood, OH.
- Clyde, M. (1999). Rejoinder to the discussion of “Bayesian model averaging and model search strategies” by M. Clyde. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 179–185. Oxford University Press: Oxford.
- Clyde, M. and George, E. I. (2004). “Model uncertainty.” *Statistical Science*, 149: 81–94.
- Cuzick, J. (1991). Discussion of “Posterior Bayes factors” by M. Aitkin. *Journal of the Royal Statistical Society, Series B*, 53: 135–136.
- Garthwaite, P. H., Critchley, F. and Mubwandarikwa, E. (2008). “Orthogonalisation of vectors with minimal adjustment.” Technical Report 8/15, Department of Mathematics and Statistics, Open University.
http://statistics.open.ac.uk/TechnicalReports/tech_report1.pdf
- George, E. I. (1999). “Sampling considerations for model averaging and model search.” Invited discussion of “Bayesian model averaging and model search strategies” by M. Clyde. In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 175–177. Oxford University Press: Oxford.
- George, E. I. (2001). “Dilution priors for model uncertainty.” Presentation at the MSRI Workshop on Nonlinear Estimation and Classification,
<http://www.msri.org/publications/ln/msri/2001/nle/george/1/index.html>.
- George, E. I. (2003). “Dilution priors for model uncertainty.” Presentation at the IMS-ISBA Conference in Puerto Rico.

- Harrison, D. and Rubinfeld, D. L. (1978). “Hedonic housing prices and demand for clean air.” *Journal of Environmental Economics and Management*, 5: 81–102.
- Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). “Bayesian model averaging: A tutorial.” *Statistical Science*, 14: 382–401.
- Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S. and Peters, S. C. (1980). “Interactive elicitation of opinion for a normal linear model.” *Journal of the American Statistical Association*, 75: 845–854.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press: London.
- O’Hagan, A. (1995). “Fractional Bayes factors for model comparison (with discussion).” *Journal of the Royal Statistical Society, Series B*, 57: 99–138.
- Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). “Bayesian model averaging for linear regression models.” *Journal of the American Statistical Association*, 92: 179–191.
- Xu, S. (2007). “An empirical Bayes method for estimating epistatic effects of quantitative trait loci.” *Biometrics*, 63: 513–521.
- Vandaele, W. (1978). “Participation in illegitimate activities; Ehrlich revisited.” In Blumstein, A., Cohen, J. and Nagin, D. (eds.), *Deterrence and Incapacitation*, 270–335. National Academy of Science Press: Washington, DC.

Acknowledgments

This work was supported by a PhD studentship from the Open University. It benefitted from the constructive comments of a referee, an Associate Editor and the Editor. It was completed while Paul Garthwaite was a visiting academic at the University of New South Wales.