# On evidence weighted mixture classification

Richard M. Everson, Wojtek J. Krzanowski, Trevor C. Bailey
and Jonathan E. Fieldsend

School of Engineering, Computer Science and Mathematics
University of Exeter
Exeter, UK.

## Abstract

Calculation of the marginal likelihood or evidence is a problem central to model selection and model averaging in a Bayesian framework. Many sampling methods, especially (Reversible Jump) Markov chain Monte Carlo techniques, have been devised to avoid explicit calculation of the evidence, but they are limited to models with a common parameterisation. It is desirable to extend model averaging to models with disparate architectures and parameterisations. In this paper we present a straightforward general computational scheme for calculating the evidence, applicable to any model for which samples can be drawn from the posterior distribution of parameters conditioned on the data. The scheme is demonstrated on a simple feature subset selection example.

## 1 Introduction

Model comparison, model selection and model averaging depend upon the availability of a measure of the quality of a model given some data. Within the Bayesian paradigm the relevant quantity is $p(D \,|\, \mathcal{M})$ which is known as the model *evidence* [MacKay, 1995] or *marginal likelihood* [Kass and Raftery, 1995] and which measures the probability of the data $D$ conditioned on the model $\mathcal{M}$. The evidence appears as the denominator in Bayes' rule for the posterior probability of the model parameters $\boldsymbol{\theta}$ given observed data $D$:

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{M}) = \frac{p(D \,|\, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \,|\, \mathcal{M})}{p(D \,|\, \mathcal{M})}. \tag{1}$$

Specification of the model defines the likelihood $p(D \,|\, \boldsymbol{\theta}, \mathcal{M})$ while priors over the parameters $p(\boldsymbol{\theta} \,|\, \mathcal{M})$ are typically assigned to embody subjective expectations. The evidence, however, may be regarded as merely a normalising factor, ensuring that the posterior density integrates to unity, and is unimportant for comparing posterior probabilities of parameters for a particular model, $\mathcal{M}$. Indeed, analytic calculation of the evidence is intractable in all but the simplest cases.

Nonetheless, the model evidence is essential for comparing models. As MacKay [2003, page 379] remarks,

> *The normalising constant $Z$ $[= p(D \,|\, \mathcal{M})]$ is often the most important number in the problem, and I think every effort should be devoted to calculating it.*

The purpose of this paper is to present a straightforward computational scheme for calculating the evidence using samples drawn from the parameter posterior distribution, and to show how it can be used in practice.

In the remainder of this section we discuss Bayesian model averaging in more detail, before describing our computational scheme in section 2. We demonstrate the efficacy of the scheme in section 3 and conclude with a discussion in section 4.

## 1.1 Model averaging

Although the marginal likelihood is useful for comparing unsupervised models, for definiteness we concentrate on the supervised learning of the mapping from features $\mathbf{x}$ to targets $y$. We assume that a model $\mathcal{M}$, which depends on some parameters, $\boldsymbol{\theta}$, gives the probability $p(y \,|\, \mathbf{x}, \boldsymbol{\theta}, \mathcal{M})$ of the target having value $y$ given the feature $\mathbf{x}$. For example, if $\mathcal{M}$ is the family of linear discriminants with link function $\phi$, so that $y = \phi(\sum_i w_i x_i + w_0)$, then $p(y \,|\, \mathbf{x}, \boldsymbol{\theta}, \mathcal{M})$ is the probability that $\mathbf{x}$ belongs to one of two classes and $\boldsymbol{\theta} = \mathbf{w}$ is a vector of weights and the bias, $w_0$. As a second example, if $\mathcal{M}$ is the family of radial basis function regressors, then $\boldsymbol{\theta}$ is the collection of basis functions centres and variances together with the interconnection weights. Traditional methods fix $\boldsymbol{\theta}$ at a 'best' value, usually chosen by minimising an error function (or, equivalently, maximising a likelihood or penalised likelihood) based on some training data $D$. These 'best' parameters are then used in the model to make predictions on previously unseen data. The Bayesian model-averaging approaches instead average over the set of parameters $\boldsymbol{\theta}$. See Clyde and George [2004] for a recent review of developments in Bayesian model averaging. The averages are taken with respect to $p(\boldsymbol{\theta} \,|\, D, \mathcal{M})$, the posterior probability of the parameters having observed the data:

$$p(y \,|\, \mathbf{x}, D, \mathcal{M}) = \int p(y \,|\, \mathbf{x}, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \,|\, D, \mathcal{M}) \, d\boldsymbol{\theta}. \tag{2}$$

This averaging process provides a number of advantages over the 'best' value approach. First, weighted averaging over models is a natural response to the incorrect specification of the model and avoids the brittleness associated with a single highly-tuned (but probably incorrect) model. Secondly, the posterior distribution of models allows the estimation of confidence measures in the resultant prediction or classification. Thirdly, Bayesian averaging is the optimal decision policy with respect to quadratic loss.

Analytic expressions for the posterior probabilities are available only in rare, simple situations [see, for example, Bernardo and Smith, 1994]. Using Bayes' rule (1) in equation (2) shows that the difficulty lies in calculating the evidence, which is the normalising factor $p(D \,|\, \mathcal{M})$ in (1):

$$p(D \,|\, \mathcal{M}) = \int p(D \,|\, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \,|\, \mathcal{M}) \, d\boldsymbol{\theta}. \tag{3}$$

Recourse is frequently made to Markov chain Monte Carlo (MCMC) methods to approximate (2). The key to the MCMC process is that samples $\boldsymbol{\theta}^{(k)}$ may be drawn from the posterior $p(\boldsymbol{\theta} \,|\, D, \mathcal{M})$, even when the posterior is only known up to a constant normalising factor. With $K$ samples the averaging integral (2) may then be approximated as

$$p(y \,|\, \mathbf{x}, D, \mathcal{M}) \approx \frac{1}{K} \sum_{k=1}^{K} p(y \,|\, \mathbf{x}, \boldsymbol{\theta}^{(k)}, \mathcal{M}). \tag{4}$$

With the burgeoning computational resources available, MCMC methods have become well established (see Liu [2001] for a recent survey). In addition, the Reversible Jump extension to MCMC [RJMCMC: Green, 1995; Denison et al., 2002] permits integration over parameter sets whose dimensionality is not fixed. For example, the number of hidden units in a neural network [Andrieu et al., 2001], the number of neighbours in a k-nearest neighbour (k-NN) model [Holmes and Adams, 2002] or the number of inputs for feature selection [e.g., Carlin and Chib, 1995; Sykacek, 2000; Vehtari and Lampinen, 2001].

However, state of the art (RJ)MCMC currently permits averaging over the parameters of models with a common parameterisation, $\boldsymbol{\theta}$, which we call a *family*, denoted by $\mathcal{M}$; thus one may average over all k-NN models or all linear logistic regressors, but not over both families together. A logical development of this idea would be to extend model averaging beyond single model families to cover several architectural families of model.

Extending the RJMCMC formalism to cover changes in dimension associated with jumps between families is a superficially attractive avenue. However, formulating an overarching model to include several disparate models is difficult. Even if a grand model could be constructed, coming up with efficient proposal densities to make transitions between the disparate families within the model is likely to very hard, and the derivation of the attendant Jacobian of the transformation $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}'$ for inter-family transitions will be difficult and error prone. Moreover, one would have to construct special proposals and calculate Jacobians between every pair of families comprising the grand model.

If the evidence $p(D \,|\, \mathcal{M}_i)$ for each model family $\mathcal{M}_i$ were available then the model averaging could be extended to cover families of models:

$$p(y \,|\, \mathbf{x}, D) = \sum_i p(y \,|\, \mathbf{x}, D, \mathcal{M}_i) p(\mathcal{M}_i \,|\, D) \tag{5}$$

$$\propto \sum_i p(y \,|\, \mathbf{x}, D, \mathcal{M}_i) p(D \,|\, \mathcal{M}_i) p(\mathcal{M}_i). \tag{6}$$

where $p(\mathcal{M}_i)$ are priors over each of the model families. With the evidences on hand it is simple to perform this averaging: with a target of $K$ samples overall, we run Markov chains for each family $\mathcal{M}_i$ and average the predictions from $K \times p(D \,|\, \mathcal{M}_i) p(\mathcal{M}_i)$ independent samples from the chain for model family $\mathcal{M}_i$.

Alternatively, instead of averaging over all families the evidence may be used to select the model family for which there is greatest evidence and inferences can then be conducted by (RJ)MCMC within this family.

## 2 Calculating Evidence

Various methods have been proposed for direct MCMC calculation of evidences [Kass and Raftery, 1995; Bos, 2002]. The simplest, but least efficient, is to sample from the prior $p(\boldsymbol{\theta} \,|\, \mathcal{M})$ and then to estimate (3) by averaging the likelihood $p(D \,|\, \boldsymbol{\theta}, \mathcal{M})$ over the sampled values. The harmonic mean of the likelihood values using samples from the posterior $p(\boldsymbol{\theta} \,|\, D, \mathcal{M})$ is much more efficient, but suffers from a certain amount of instability [Kass and Raftery, 1995], although modifications to avoid the instability have been proposed by Newton and Raftery [1994] and Gelfand and Dey [1994].

Our approach to calculating the evidence is essentially that of Aitken [1991]. The key is to partition the available training data, $D$, into an evidence set, $D_E$, and a

training set, $D_T$. Then the evidence for model $\mathcal{M}$ conditioned on the training set, is

$$p(D_E \,|\, D_T, \mathcal{M}) = \int p(D_E \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, D_T, \mathcal{M})\, d\boldsymbol{\theta} \tag{7}$$

which can be calculated by drawing samples from $p(\boldsymbol{\theta} \,|\, D_T, \mathcal{M})$ in the usual way and using them to average the likelihood of the evidence partition, $p(D_E \,|\, \boldsymbol{\theta})$. The result, $p(D_E \,|\, D_T, \mathcal{M})$, is the likelihood of the evidence data given the training data under the model $\mathcal{M}$. It is analogous to the innovations probability $p(x_t \,|\, D_{t-1}, \mathcal{M})$ in hidden Markov models, where $x_t$ is an observation at time $t$ and $D_{t-1}$ denotes all the data observed up to and including time $t-1$ [e.g. Rabiner, 1989].

However, the above method of calculating the evidence suffers from the drawback that each of $D_T$ and $D_E$ form only a portion of the available data $D$, and in this there is a direct analogy with the problems encountered in data-based assessment of classification rules. In such assessment we split the available data into a portion for training the classifier and a portion for assessing its performance, while in the present case the 'training' is the estimation of the posterior probabilities $p(\boldsymbol{\theta} \,|\, D_T, \mathcal{M})$ and the 'assessment' is the calculation of $p(D_E \,|\, \boldsymbol{\theta})$ for averaging. Increasing the $D_E/D_T$ ratio will improve estimation of $p(D_E \,|\, \boldsymbol{\theta})$ but at the expense of poorer estimation of $p(\boldsymbol{\theta} \,|\, D_T, \mathcal{M})$, and vice-versa. This may not be a problem if very large amounts of training data are available, but in general we would like to optimise data usage in the estimation process. To do this we can draw on techniques from data-based classification assessment.

In order to maximise the use of the information in the data, rather than splitting the dataset into just a single pair of evidence and training partitions, we use a scheme similar to $G$-fold cross validation. The data set $D$ is divided into $G$ equally-sized disjoint subsets, $D_g$, so that $D = \cup_{g=1}^{G} D_g$, and $D_{-g}$ denotes all the data not in partition $g$, $D_{-g} = D \setminus D_g$. Then $p(D_g \,|\, D_{-g}, \mathcal{M})$ is calculated using (7) and a straightforward plug-in estimate of the total evidence is:

$$p(D \,|\, \mathcal{M}) \approx \prod_{g=1}^{G} p(D_g \,|\, D_{-g}, \mathcal{M}). \tag{8}$$

As we now demonstrate, (8) provides a viable and effective approach for estimating the model evidence.

# 3 Illustration

We illustrate this approach on variable selection in a simple synthetic linear regression problem, for which analytic results are available [Lindley and Smith, 1972; Bernardo and Smith, 1994].

We have simulated data in which the dependent variable $y$ is generated as a linear combination of four-dimensional data $\mathbf{x}$ with Gaussian-distributed observational noise:

$$y = w_0 + \sum_{j=1}^{4} w_j x_j + \epsilon. \tag{9}$$

The $x_j$ themselves are Gaussian distributed with zero mean and unit variance. In half the data $y$ is a linear combination of $x_2$ and $x_3$ plus noise of unit variance: $y = x_2 - 4x_3 + \epsilon$; in the other half of the data the predictive variables are $x_1$ and $x_3$ and the noise has variance $0.2^2$: $y = x_1 - x_3/2 + \epsilon$. $x_4$ is irrelevant for prediction.

| $\mathcal{M}_i$ | $\log p(D_E \mid \mathcal{M}_i)$ | $\log p(D_E \mid D_T, \mathcal{M}_i)$ |
|---|---|---|
| (4) | -9518.8 | -9497.7 |
| (2, 4) | -9508.4 | -9481.0 |
| (2) | -9500.9 | -9481.1 |
| (1, 4) | -9494.9 | -9467.4 |
| (1) | -9487.2 | -9467.4 |
| (1, 2, 4) | -9482.8 | -9449.1 |
| (1, 2) | -9475.2 | -9449.2 |
| (3, 4) | -8959.5 | -8930.6 |
| (3) | -8951.8 | -8930.3 |
| (2, 3, 4) | -8943.9 | -8908.5 |
| (2, 3) | -8936.3 | -8908.2 |
| (1, 3, 4) | -8924.9 | -8889.8 |
| (1, 3) | -8917.0 | -8889.5 |
| (1, 2, 3, 4) | -8907.0 | -8865.4 |
| (1, 2, 3) | -8899.2 | -8865.2 |

Table 1: Estimates of $p(D_E \mid D_T, \mathcal{M}_i)$ and 'true' evidence $p(D_E \mid M_i)$ for each of 15 models $\mathcal{M}_i$ corresponding to different feature combinations.

We regard the linear regression model with each of the possible combinations of input variables as a separate model so that, excluding the model with no inputs, there are 15 possible models $\mathcal{M}_i$. If $\mathbf{w}$ denotes the vector of coefficients and the bias $w_0$ for the relevant model, then we choose conjugate, Normal-Inverse-Gamma (NIG) priors over $\mathbf{w}$ and the noise variance $\sigma^2$; thus

$$p(\mathbf{w}, \sigma^2) = p(\mathbf{w} \mid \sigma^2) p(\sigma^2) \tag{10}$$

$$= N(\mathbf{w} \mid \mathbf{m}, \sigma^2 \mathbf{V}) IG(\sigma^2 \mid \alpha, \beta), \tag{11}$$

where $\mathbf{V} = v\mathbf{I}$ with $v = 10^4$, and where $IG$ denotes the inverse-gamma density:

$$IG(\sigma^2 \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-2(\alpha+1)} e^{-\beta/\sigma^2}, \tag{12}$$

with hyperparameters $\alpha$ and $\beta$. Note that in order to simplify the calculation of the marginal likelihood, we set a Normal prior with mean zero over all the coefficients including the bias $w_0$, rather than the more common improper uniform prior over $w_0$.

Standard theory [see, for example, Bernardo and Smith, 1994] shows that the posterior distribution for $\mathbf{w}$ and $\sigma^2$ is also NIG with parameters:

$$\mathbf{m}^* = (\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{13}$$

$$\mathbf{V}^* = (\mathbf{V}^{-1} + \mathbf{X}^T\mathbf{X})^{-1} \tag{14}$$

$$\alpha^* = \alpha + N/2 \tag{15}$$

$$\beta^* = \beta + [\mathbf{y}^T\mathbf{y} - (\mathbf{m}^*)^T(\mathbf{V}^*)^{-1}\mathbf{m}^*]/2. \tag{16}$$

Here $\mathbf{X}$ is the matrix of the $N$ observations $\mathbf{x}_n$ arranged as rows and $\mathbf{y}$ is the vector of the $y_n$.

The 'true' evidence of data $D = \{\mathbf{X}, \mathbf{y}\}$ is calculated by evaluating the marginal likelihood:

$$p(D \mid \mathcal{M}_i) = (2\pi)^{-N/2} \frac{|\mathbf{V}^*|^{1/2}}{|\mathbf{V}|^{1/2}} \frac{(\beta^*)^{-\alpha^*}}{\beta^{-\alpha}} \frac{\Gamma(\alpha^*)}{\Gamma(\alpha)}. \tag{17}$$

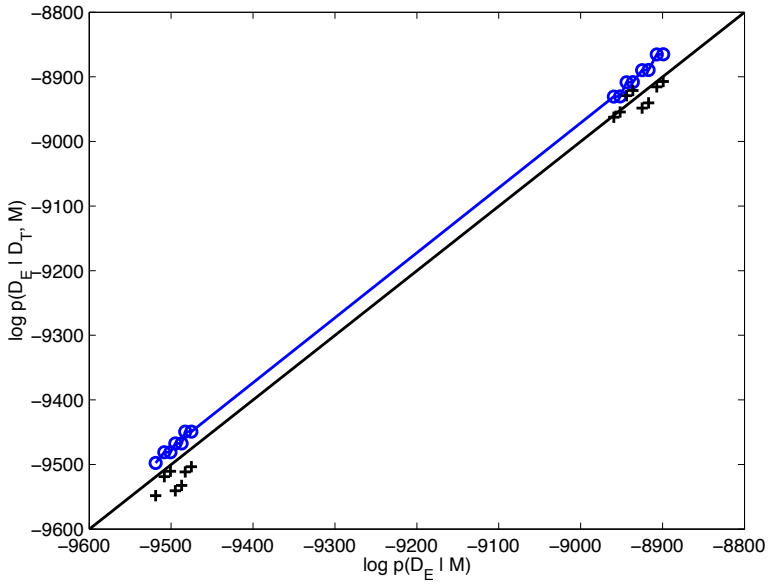


Figure 1: Joined circles: Estimates of $p(D_E \,|\, D_T, \mathcal{M}_i)$ plotted versus $p(D_E \,|\, M_i)$ for each of 15 models corresponding to different feature combinations. Crosses: $p(D_T \,|\, \mathcal{M}_i)$.

As an initial illustration we estimate $p(D_E \,|\, D_T, \mathcal{M}_i)$ for independently generated datasets $D_E$ and $D_T$, each comprised of $N = 4000$ observations. For each of the 15 models corresponding to combinations of the input features, $p(D_E \,|\, D_T, \mathcal{M}_i)$ was calculated by drawing samples $\{\mathbf{w}^{(k)}, \sigma^{(k)}\}$ from the posterior density $p(\mathbf{w}, \sigma \,|\, D_T, \mathcal{M}_i)$; these were used to approximate (7):

$$p(D_E \,|\, D_T, M_i) \approx \frac{1}{K} \sum_{k=1}^{K} p(D_E \,|\, \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i). \tag{18}$$

where the likelihood for the 'evidence partition', $D_E = \{\mathbf{X}, \mathbf{y}\}$ is

$$p(D_E \,|\, \mathbf{w}, \sigma) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{ -\frac{(\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w})}{2\sigma^2} \right\}. \tag{19}$$

Table 1 and Figure 1 compare the estimated evidence $p(D_E \,|\, D_T, \mathcal{M}_i)$ with the true evidence $p(D_E \,|\, \mathcal{M}_i)$ calculated using (17) for the 15 models corresponding to combinations of the input features. The models are ordered by the true evidence. There is most evidence for the model using the relevant predictors $(x_1, x_2, x_3)$ and the models fall into two distinct groups: those with high evidence include $x_3$, which is predictive for both halves of the data, as an input feature; and those not including $x_3$ with low evidence.

As Table 1 and Figure 1 show there is good agreement between the estimated evidence and the true value. The approximation to the evidence correctly orders the models according to $p(D_E \,|\, \mathcal{M}_i)$. Furthermore, the numerical values of the approximations are close to the true values. As a crude measure of the variability that might be expected in $p(D_E \,|\, \mathcal{M}_i)$, the Figure 1 also shows the 'true' evidence $p(D_T \,|\, \mathcal{M}_i)$ for the training partition, which was generated in an identical manner

| $\mathcal{M}_i$ | $\log p(D \mid \mathcal{M}_i)$ | $\log \prod_g p(D_g \mid D_{-g}, \mathcal{M}_i)$ |
|---|---|---|
| (4) | -9548.2 | -9526.5 |
| (1, 4) | -9540.7 | -9511.9 |
| (1) | -9532.5 | -9511.3 |
| (2, 4) | -9518.8 | -9490.0 |
| (1, 2, 4) | -9511.6 | -9475.8 |
| (2) | -9510.6 | -9489.3 |
| (1, 2) | -9503.4 | -9475.1 |
| (3, 4) | -8962.5 | -8934.6 |
| (3) | -8954.4 | -8934.0 |
| (1, 3, 4) | -8948.5 | -8913.2 |
| (1, 3) | -8940.3 | -8912.6 |
| (2, 3, 4) | -8929.1 | -8894.1 |
| (2, 3) | -8921.0 | -8893.6 |
| (1, 2, 3, 4) | -8915.5 | -8873.2 |
| (1, 2, 3) | -8907.3 | -8872.6 |

Table 2: The 'true' evidence $p(D \mid M_i)$ and estimates of $p(D \mid \mathcal{M}_i)$ from $\prod_g p(D_g \mid D_{-g}, \mathcal{M})$ for each of 15 models $\mathcal{M}_i$ corresponding to different feature combinations.

to $D_E$ and has the same number of observations as $D_E$. This indicates that the approximation to the evidence is within the variability that might arise from the observation of a particular dataset. However, the reason for the approximations being systematically larger than the true values is not at present known.

It is also interesting to note that the 'staircase' appearance of Figure 1 is due to the fact that $p(D_E \mid D_T, \mathcal{M}_i)$ for a model which includes the irrelevant variable $x_4$ is only very slightly lower than the evidence for the model excluding $x_4$.

The calculation discussed above partitioned the data into a set $D_T$ for estimating the model parameters, while the approximation was evaluated on the remaining data, $D_E$. We now examine the quality of the 'cross validation' approximation (8). We used a total of $N = 4000$ observations from the linear regression data and partitioned them into $G = 5$ subsets $D_g$. The data in the complement of $D_g$ were used to run a Markov chain to draw samples $\{\mathbf{w}^{(k)}, \sigma^{(k)}\}$ from $p(\mathbf{w}, \sigma \mid D_{-g}, \mathcal{M}_i)$; the samples were then used to average the likelihood $p(D_g \mid \mathbf{w}, \sigma)$ in a similar manner to (18). The averages from the 5 separate chains were combined using (8) to estimate $p(D \mid \mathcal{M}_i)$.

Comparisons of the true evidence and the approximation for each of the 15 feature combinations are shown in Table 2 and Figure 2. Again, except for one model, the approximation correctly orders the models according to $p(D \mid \mathcal{M}_i)$, and the numerical values of the approximations are close to the true values.

We remark that some care must be taken with the averaging of the conditional likelihoods $p(D_g \mid \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)$. In this problem, as Figures 3 and 4 show, $\log p(D_g \mid \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)$ is approximately normally distributed. This presents two difficulties. First, samples in the extreme right-hand tail of the distribution have a large influence on the value of $\langle p(D_g \mid \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i) \rangle$. Since these samples are rare it is necessary to run the Markov chain for long enough to faithfully represent this tail of the distribution. In the experiments reported here we used $10^5$ samples for averaging; independence between samples was ensured by using only every seventh sample in a chain of length $7 \times 10^5$, following a burn-in period. Secondly, the wide range of values of $\log p(D_g \mid \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)$ means that care must be taken to pre-

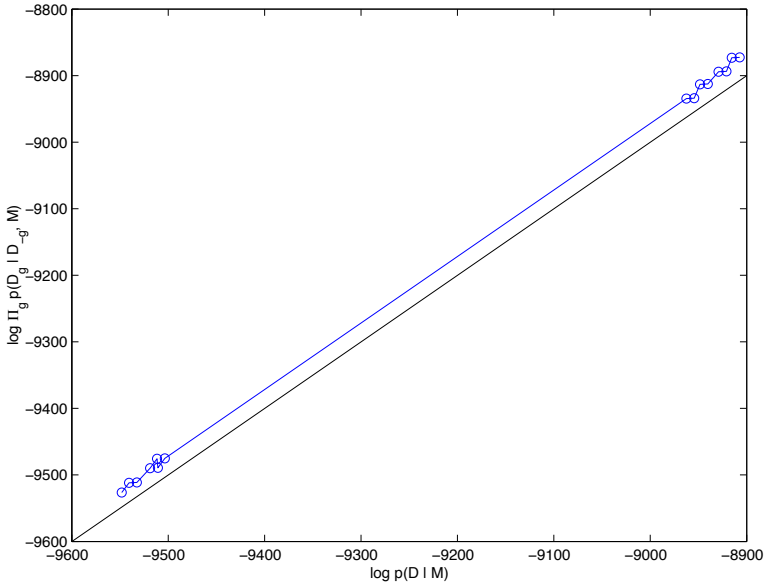


Figure 2: Estimates of $p(D \mid D_T, \mathcal{M}_i)$ plotted versus $p(D \mid M_i)$ for each of 15 models corresponding to different feature combinations.

vent numerical overflow and underflow when averaging the 'un-logged' probabilities.

The normal nature of the distribution of $\log p(D_g \mid \boldsymbol{\theta}, \mathcal{M}_i)$ in this particular problem means that the integral (7) could be finessed with a semi-analytic approximation. However, calculations for other models (e.g., the probabilistic k-NN model of Holmes and Adams [2002]) indicate that the distribution of $p(D_g \mid \boldsymbol{\theta})$ may have a quite different shape, necessitating straightforward Monte Carlo integration of (7) as performed here. On the other hand the distribution of $p(D_g \mid \boldsymbol{\theta})$ for the probabilistic k-NN model is skewed towards large likelihood values thus reducing the number of samples needed to obtain a good approximation to the average.

# 4 Discussion

In this paper we have proposed a straightforward scheme for numerically approximating the evidence of a model or family of models. Calculations on a simple, but non-trivial, linear regression problem indicate that the scheme yields reasonably accurate approximations. Implementation of the approximation scheme requires only that samples can be drawn from the posterior distribution of the model parameters, together with calculation of model likelihoods. It is thus widely applicable to the many model families for which (RJ)MCMC methods have been developed.

In order to make full use of the information available in a dataset, we combined estimates from partitions of the data using the 'plug-in' estimate (8), which gives reasonably accurate results. An alternative method of combining evidence estimates from partitions $D_g$ and $D_{-g}$ is to approximate the integral (7) as a sum:

$$p(D_g \mid \mathcal{M}) = \sum_{g=1}^{G} p(D_g \mid D_{-g}, \mathcal{M}) p(D_{-g} \mid \mathcal{M}). \tag{20}$$

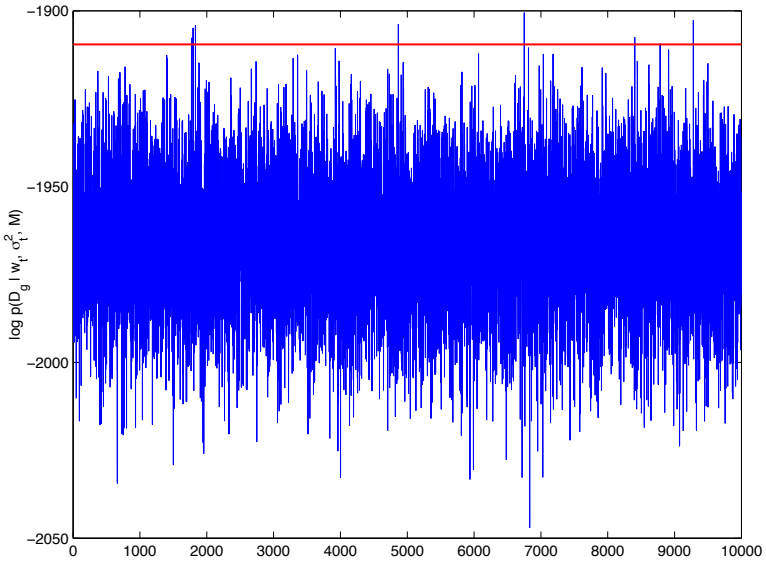


Figure 3: Samples $\log p(D_g \,|\, \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)$ used for averaging $p(D_g \,|\, \boldsymbol{\theta}, \mathcal{M}_i)$. The horizontal line indicates $\log\langle p(D_g \,|\, \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)\rangle$.

Methods which treat (20) as a system of linear equations to be solved for $p(D_g \,|\, \mathcal{M})$ and $p(D_{-g} \,|\, \mathcal{M})$ are a focus of further work.

We mention that in the nomenclature of statistical physics, the evidence is known as the partition function $Z$ and the free energy of an ensemble is $F = -\log(Z)$. Neal [1993] reviews a number of numerical methods—importance sampling, thermodynamic integration, and distribution overlap—and has proposed an annealed importance sampling method [Neal, 1998] for estimating the difference in free energy for a pair of systems. However, these methods estimate the free energy *difference,* and in general the systems must be quite similar and require a common parameterisation.

Variational methods [e.g., Jaakkola et al., 1998] also provide an approximation to the evidence although they are of course approximations and good variational approximations are not available for all models. An alternative method for calculating the evidence is via a Laplace approximation to (3). In practice, variational methods and Laplace approximations can be difficult to apply and the quality of the approximation they provide is unknown.

The evidence calculation scheme proposed here opens up a viable method for Bayesian model averaging over architecturally distinct families of model. Thus current work is focusing on averaging over families of classifier, such as linear logistic regressors (a family of global models), radial basis function classifiers (semi-local models) and probabilistic k-nearest neighbour classifiers (local models). Averaging within each family may be performed by Reversible Jump MCMC integration.

This method requires the running of $G \geq 2$ Markov chains for each model or model family over which averaging is to be performed, even if it subsequently turns out that some of the models have negligible evidence. It is therefore not a replacement for RJMCMC methods, the beauty of which lies in their ability to sample models in proportion to the evidence for the model, even if the space of possible models is very large or infinite. We therefore view the principal application of this

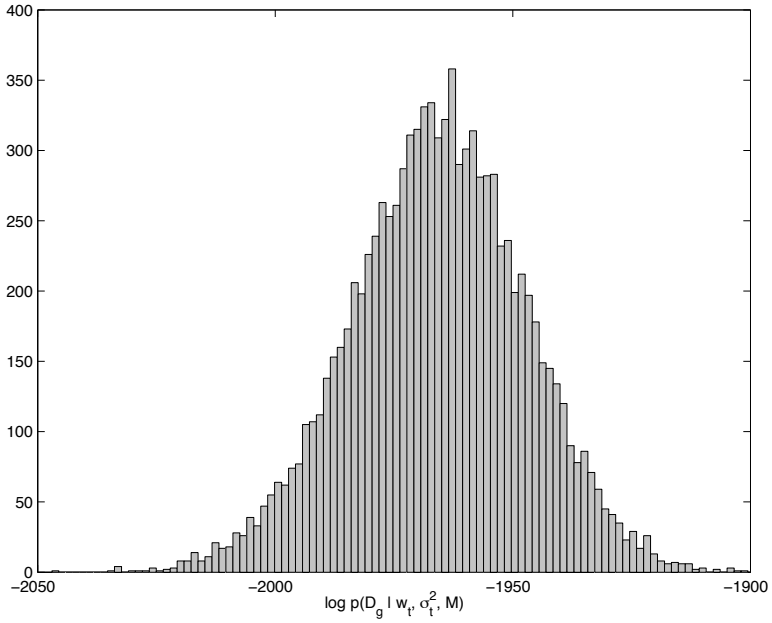


Figure 4: Histogram of the samples $\log p(D_g \mid \mathbf{w}^{(k)}, \sigma^{(k)}, \mathcal{M}_i)$ shown in Figure 3.

method to be the extension of model averaging or model selection to a (relatively small) range of architecturally distinct model families.

Nonetheless, this method might be used for model averaging within a model family in which RJMCMC is possible, but technically difficult. For example, Richardson and Green [1997] showed how to construct reversible jump Markov chains for Gaussian mixture models in one dimension, in which the reversible jumps are made over the number of Gaussian kernels in the mixture model. Extending their technique to mixture models in more than one dimension is theoretically straightforward, but finding efficient proposal densities and calculating the relevant Jacobians is daunting. On the other hand MCMC for each fixed number of mixture kernels can be performed by straightforward Gibbs sampling; if the evidence for each number of kernels were assessed with this technique, samples from each model could then easily combined in the correct proportions.

**Acknowledgements**

This work was in part supported by EPSRC grant GR/R24357/01.

**References**

M. Aitken. Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B*, 53:111–142, 1991.

C. Andrieu, J. de Freitas, and A. Doucet. Robust full Bayesian learning for radial basis networks. *Neural Computation*, 13(10):2359–2407, 2001.

J.N. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 1994.

C.S. Bos. A comparison of marginal likelihood computation methods. Technical Report TI2002-084/3, Vrije Universiteit, Amsterdam, 2002.

B.P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo. *J. Royal Statistical Society*, B57:473–484, 1995.

M. Clyde and E.I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

D.G.T. Denison, C.C. Holmes, B.K. Mallick, and A.F.M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, 2002.

A.E. Gelfand and D.K. Dey. Bayesian model choice. *Journal of the Royal Statistical Society, Series B*, 56:501–514, 1994.

P.J. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(711-732), 1995.

C.C. Holmes and N.M. Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal Royal Statistical Society B*, 64:1–12, 2002.

T.S. Jaakkola, M.I. Jordan, Z. Ghahramani, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistician*, 90:773–795, 1995.

D.V. Lindley and A.F.M. Smith. Bayes estimates for the linear model (with discussion). *J. Royal Statistical Society B*, 34:1–41, 1972.

J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer, 2001.

D.J.C. MacKay. Probable neural networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 4:448–472, 1995.

D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

R.M. Neal. Annealed importance sampling. Technical Report 9805, Dept. of Statistics, University of Toronto, 1998.

M.A. Newton and A.E. Raftery. Approximate Bayesian inference by weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE Trans. Sig. Processing*, 77(2):257–286, 1989.

S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.

P. Sykacek. On input selection with reversible jump Markov chain Monte Carlo Sampling. In S.A. Solla, T.K. Leen, and K.-R. M uller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 638–644, 2000.

A. Vehtari and J. Lampinen. Bayesian input variable selection using cross-validation predictive densities and reversible jump MCMC. Technical Report 951-22-5644-4, Helsinki University of Technology, 2001.