Bayesian Learning using Automatic Relevance Determination Prior with an Application to Earthquake Early Warning

Chang Kook Oh [1] James L. Beck [2] and Masumi Yamada [3]

Abstract

A novel method of Bayesian learning with automatic relevance determination prior is presented that provides a powerful approach to problems of classification based on data features, for example, classifying soil liquefaction potential based on soil and seismic shaking parameters, automatically classifying the damage states of a structure after severe loading based on features of its dynamic response, and real-time classification of earthquakes based on seismic signals. After introduction of the theory, the method is illustrated by applying it to an earthquake record dataset from nine earthquakes to build an efficient real-time algorithm for near-source versus far-source classification of incoming seismic ground motion signals. This classification is needed in the development of early warning systems for large earthquakes. It is shown that the proposed methodology is promising since it provides a classifier with higher correct classification rates and better generalization performance than a previous Bayesian learning method with a fixed prior distribution that was applied to the same classification problem.

Key words: Bayesian learning, Automatic relevance determination prior, Pattern recognition and classification, Seismic early warning

Introduction

Classification is a sub-topic of machine learning which can be defined as 'the act of taking in raw data and taking an action based on the category of the data' (Duda et al. 2000). By using a given training dataset, a separating boundary is identified that separates different-class data in the feature space, then the category to which new data belongs is decided by using that separating boundary.

This classification is performed in three phases:

- Phase I (Feature Extraction Phase) : This phase distills a small number of features from a large set of data that are thought to characterize each class of interest in the data.
- Phase II (Training Phase) : This phase identifies a separating boundary based on extracted features that are most relevant to the data classification, usually using some form of regularization.
- Phase III (Prediction Phase) : In this phase, a prediction is made using the separating boundary from the previous phase to decide to which class new data belongs.

Bayesian methods for classification problems have the advantage that they

[1] Ph. D., Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA. 91125, ockoogi@gmail.com

[2] Professor, Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA. 91125, jimbeck@caltech.edu

[3] Assistant Professor, Earthquake Hazards Division, Kyoto University, Gokasyo, Uji, 611-0011, Japan, masumi@eqh.dpri.kyoto-u.ac.jp

make probabilistic predictions (rather than giving only a possibly misleading yes/no answer) for the class that corresponds to a given feature vector (Bishop 2006). These predictions are based on a rigorous Bayesian learning procedure that rests on the axioms of probability. The essential ingredients are a set of predictive probability models involving a parameterized separating boundary function and a probability model (the prior distribution) over this set. The prior can be pragmatically chosen by the user to regularize the ill-conditioned problem of identifying a boundary that separates the classes in the feature vector space. In the absence of such regularization, the training phase will be usually lead to "over-fitting" of the data, so that generalization beyond the training data in the prediction phase will perform poorly.

In this paper, the novel method of *Bayesian learning with automatic relevance determination (ARD) prior* is presented and illustrated for an interesting classification problem in earthquake early warning systems (Yamada et al. 2007) because of its exceptional regularization ability (Mackay 1994; Oh and Beck 2006; Tipping 2004). The presented Bayesian approach is useful for other problems of data-based classification in earthquake engineering and structural health monitoring, such as liquefaction for sandy soil sites based on soil properties and ground shaking intensity, classifying damage states based on sensor data (Oh and Beck 2006), and so on. In the application presented here, the Bayesian learning method with ARD prior provides an algorithm for probabilistic predictions of whether the seismic ground motion signal that is transmitted from a seismic sensor network corresponds to near-source or far-source ground motion with respect to the causative fault. This information is important for an early warning system when it is automatically estimating the location and magnitude of the earthquake in real time (Yamada et al. 2007).

Since an earthquake is a sudden event that comes without much warning, there is increasing research interest in automated seismic early warning systems that can take rapid actions to mitigate damage and loss before the onset of the damaging ground shaking at a facility (Allen and Kanamori 2003; Cua 2005; Grasso and Beck 2007). Seismic early warning is based on the principle that an automated and reliable system may allow time for taking mitigation measures because the speed of the most damaging S-waves (about 3.5 km/s) is slower than that of electrically transmitted signals from the seismic network sensors (about $300,000$ km/s) that detect the onset of the event.

A recently-developed method for an early warning system, called the Virtual Seismologist (VS) method (Cua 2005) can estimate the location of the epicenter and the magnitude within a few seconds after the detection of the P-waves near the causative fault. This VS method, however, currently works for moderate earthquakes of magnitude less than about 6.5 because it assumes a point-source model for the rupture (Cua 2005). To construct a seismic early warning system dealing with larger earthquakes, knowledge of the fault geometry is essential and an important ingredient in establishing the extent of the rupturing fault is to be able to estimate whether the station is close to the fault (near-source) or at some distance (far-source) based on the waveform data available at the given station (Yamada et al. 2007).

The earthquake dataset and the extracted features are described in the next section and then the training and predicting phase with the Bayesian learning procedure is described. The results obtained by the proposed method for near-source (NS) versus far-source (FS) classification are presented and compared with those from a recent related study (Yamada et al. 2007) with the conclusions followed at the end.

Feature Extraction for Training Data

We chose the dataset used previously by Yamada et al. (2007). It consists of 695 strong-motion records from 9 earthquakes of magnitude greater than 6.0: Imperial Valley (1979), Loma Prieta (1989), Landers (1992), Northridge (1994), Hyogoken-Nanbu (1995), Izmit (1999), Chi-Chi (1999), Denali (2002) and Niigataken-Chuetsu(2004). Records are categorized as near-source (NS) if the corresponding station is less than 10 km from the fault rupture and far-source (FS) otherwise. Only stations with fault distances less than 200 km are included since otherwise the ground motion amplitudes are small, resulting in a low signal-to-noise ratio. The precise number of NS and FS records for each earthquake is listed in Table 1. For each baseline-corrected time history in the dataset, the values of peak jerk, acceleration, velocity and displacement in the horizontal and vertical directions were extracted by taking numerical derivatives or integrals when necessary (Yamada et al. 2007). We note that *jerk* is defined as the rate of acceleration change and so it is computed as the derivative of acceleration with respect to time. Motions with higher-frequency content such as acceleration and jerk are more informative about the fault distance, since the amplitudes of these motions decay more rapidly than those of lower-frequency motions such as displacements and velocities (Hanks and McGuire 1981). For the two horizontal components of each record, the square root of the sum of squares of the peak quantities were used. Since the peak amplitudes are utilized for classification, the peak of the S-wave needs to have arrived at a given station before predictions with the Bayesian classifier can be made.

This data processing leads to the eight extracted features listed in Table 2 for

each of the 695 records. These features are combined into a vector $\underline{x} \in \mathbb{R}^8$:

$$\underline{x} = [\log_{10} H_j,\ \log_{10} Z_j,\ \log_{10} H_a,\ \log_{10} Z_a,\ \log_{10} H_v,\ \log_{10} Z_v,\ \log_{10} H_d,\ \log_{10} Z_d]^T$$

where H and Z mean the peak horizontal and vertical components and j, a, v and d stand for jerk, acceleration, velocity and displacement, respectively. The dataset of feature vectors is the same as that used in Yamada et al. (2007) where a Bayesian classification scheme was applied that used a fixed prior.

Bayesian Learning and Prediction

Let $\mathcal{D}_N = \{(\underline{x}_n, y_n) : n = 1, ..., N\} = (\mathbf{X}, \underline{y})$ denote the data with features (predictor variables) $\underline{x}_n \in \mathbb{R}^m$ and labels $y_n \in \{0, 1\}$ ($y_n = 0$ for far-source data, $y_n = 1$ for near-source data, $N = 695$ and $m = 8$ in our application).

Suppose that the function characterizing the separating boundary between the two classes is taken as a linear combination of features $\underline{x} = [x_1, ..., x_m]^T$ with unknown coefficients $\underline{\theta} = [\theta_0,\ \theta_1,\ ...,\ \theta_m]^T \in \mathbb{R}^{m+1}$:

$$f(\underline{x}|\underline{\theta}) = \sum_{j=1}^{m} \theta_j x_j + \theta_0 \tag{1}$$

The separating boundary function $f(\underline{x}|\underline{\theta})$ is also called the (linear) discriminant function. We note in passing that the method presented here also works if the x_j in (1) are replaced by nonlinear functions $g_j(\underline{x})$. For a known parameter vector $\underline{\theta}$, the separating boundary between the different classes (NS and FS in our application) is defined as $f(\underline{x}|\underline{\theta}) = 0$ and probabilistic predictions of the class label $y \in \{0, 1\}$ corresponding to extracted features $\underline{x}$ will be based on the probability model:

$$P(y|\underline{x},\underline{\theta}) = \begin{cases} \phi(f(\underline{x}|\underline{\theta})), & \text{if } y = 1 \\ 1 - \phi(f(\underline{x}|\underline{\theta})), & \text{if } y = 0 \end{cases}$$
$$= \phi(f(\underline{x}|\underline{\theta}))^y \{1 - \phi(f(\underline{x}|\underline{\theta}))\}^{1-y} \tag{2}$$

where $\phi(\cdot) \in [0,1]$ is the monotonically increasing sigmoid function on $\mathbb{R}$ defined by $\phi(x) = 1/(1+e^{-x})$ so $\lim_{x\to\infty}\phi(x) = 1$, $\lim_{x\to-\infty}\phi(x) = 0$ and $\phi(x) + \phi(-x) = 1$ (See Figure 1). Thus, when $f(\underline{x}|\underline{\theta})$ is large and positive, the probability is near 1 that $\underline{x}$ corresponds to an instance of the $y = 1$ class, while when $f(\underline{x}|\underline{\theta})$ is negative with large magnitude, $\underline{x}$ corresponds to an instance of the $y = 0$ class with probability near 1. Note that the boundary $f(\underline{x}|\underline{\theta}) = 0$ corresponds to a probability of 0.5 for both classes and it is invarient to a scaling of f; however, this scaling is important because it controls how rapidly the probability of a class approaches its asymptotic values of 0 and 1 as the feature vector $\underline{x}$ is moved away from the boundary.

Bayesian Learning

Since (1) is just a model for the separating boundary, there are no true values of $\underline{\theta}$ to be "estimated" but we can learn about how plausible its various values are by Bayesian updating using the data $\mathcal{D}_N$.

From Bayes' Theorem:

$$p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha}) = \frac{P(\mathcal{D}_N|\underline{\theta})\; p(\underline{\theta}|\underline{\alpha})}{P(\mathcal{D}_N|\underline{\alpha})} \tag{3}$$

where $p(\underline{\theta}|\mathcal{D}_N,\underline{\alpha})$, $P(\mathcal{D}_N|\underline{\theta})$, $p(\underline{\theta}|\underline{\alpha})$ and $P(\mathcal{D}_N|\underline{\alpha})$ represent the posterior, likelihood, prior and evidence, respectively. The hyperparameters $\underline{\alpha}$ define the

ARD prior as explained shortly.

The likelihood $P(\mathcal{D}_N|\underline{\theta})$ measures how well the predictive probability model defined by $\underline{\theta}$ predicts the actual data:

$$\begin{aligned} P(\mathcal{D}_N|\underline{\theta}) &= \prod_{n=1}^{N} P(y_n|\underline{x}_n, \underline{\theta}) \\ &= \prod_{n=1}^{N} \phi(f(\underline{x}_n|\underline{\theta}))^{y_n} \{1 - \phi(f(\underline{x}_n|\underline{\theta}))\}^{1-y_n} \end{aligned} \tag{4}$$

The prior $p(\underline{\theta}|\underline{\alpha})$ provides a means of regularizing the learning process. A novel feature of this work is the introduction of the ARD prior (Mackay 1994; Tipping 2004), which is simply a Gaussian PDF with mean $\underline{0}$ and covariance matrix $\mathbf{A}(\underline{\alpha})^{-1} = \text{diag}\{\alpha_0^{-1}, \alpha_1^{-1}, ..., \alpha_m^{-1}\}$:

$$p(\underline{\theta}|\underline{\alpha}) = (2\pi)^{-\frac{m+1}{2}} |\mathbf{A}(\underline{\alpha})|^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2}\underline{\theta}^T \mathbf{A}(\underline{\alpha})\underline{\theta}\Big\} \tag{5}$$

The previous study by Yamada et al. (2007) adopted a fixed and non-informative prior that assigned the same value for all α_i, i.e., $\alpha_i = 100^{-2}$, $i = 0, .., m$, while the ARD prior uses an independent α_i for each parameter θ_i and these independent α_is are estimated during the learning process. The ARD prior combined with Bayesian model class selection plays an important role in selecting the significant features by utilizing only a small number of relevant features and automatically pruning the remaining features, instead of considering all possible model classes, one after another, as in Yamada et al. (2007).

The hyperparameter $\underline{\alpha} \in \mathbb{R}_+^{m+1}$ can be viewed as defining a model class $\mathcal{M}(\underline{\alpha})$ consisting of the set of predictive probability models $\{P(y|\underline{x}, \underline{\theta}) : \underline{\theta} \in \mathbb{R}^{m+1}\}$ along with the above prior PDF $p(\underline{\theta}|\underline{\alpha})$ over this set. We will then use model class selection based on the evidence $P(\mathcal{D}_N|\underline{\alpha})$ for $\mathcal{M}(\underline{\alpha})$ to select the most

probable model class $\mathcal{M}(\underline{\hat{\alpha}})$ based on data $\mathcal{D}_N$ (Beck and Yuen 2004; Mackay 1992). It was shown by Tipping (2004) that the ARD prior suppresses ill-conditioning by discouraging strong correlations between terms in (1) that are not supported by the data; in fact, it may happen that some $\hat{\alpha}_j \to \infty$ during the optimization to find $\mathcal{M}(\underline{\hat{\alpha}})$ which completely suppresses the corresponding terms in (1) (i.e., $\theta_j = 0$ since for $\mathcal{M}(\underline{\hat{\alpha}})$, θ_j has a Gaussian prior with zero mean and vanishing variance).

The next step is to construct a Gaussian approximation of the posterior $p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$ using Laplace's asymptotic approximation (Beck and Katafygiotis 1998; Mackay 1992). This is achieved by making a quadratic approximation of the log-posterior around the most probable value, $\underline{\hat{\theta}}$, given by maximization of the posterior PDF. This produces a Gaussian distribution with mean $\underline{\hat{\theta}}$ and covariance matrix $\hat{\Sigma}$ which is the inverse of the negative of the Hessian matrix of the log-posterior.

The detailed procedure for the Laplace approximation is as follows (Oh 2007): (1) For a given value of $\underline{\alpha}$, the log-posterior from (3), (4) and (5) is (ignoring irrelevant additive terms that depend only on $\underline{\alpha}$):

$$\begin{aligned}\ln[p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})] &= \sum_{n=1}^{N} \ln[P(y_n|\underline{\theta}, \underline{x}_n)] + \ln[p(\underline{\theta}|\underline{\alpha})] \\ &= \sum_{n=1}^{N} \Big[y_n \cdot \ln \phi_n(\underline{\theta}) + (1 - y_n) \cdot \ln\{1 - \phi_n(\underline{\theta})\} \Big] \\ &\quad -\frac{1}{2}\underline{\theta}^T \mathbf{A}(\underline{\alpha})\underline{\theta} \end{aligned} \tag{6}$$

where $\mathbf{A}(\underline{\alpha}) = \text{diag}\{\alpha_0, \alpha_1, ..., \alpha_N\}$ and $\phi_n(\underline{\theta}) = \phi(f(\underline{x}_n|\underline{\theta}))$. By using an iterative procedure based on a second-order Newton method (or any other optimization method), the most probable values $\underline{\hat{\theta}}(\underline{\alpha})$ are estimated by maximizing

$\ln[p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})]$.

(2) The inverse covariance matrix is $\hat{\mathbf{\Sigma}}^{-1}(\underline{\alpha}) = -\nabla_{\underline{\theta}}\nabla_{\underline{\theta}} \ln p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha})$ evaluated at $\hat{\underline{\theta}}(\underline{\alpha})$ and the resulting Gaussian approximation of the posterior distribution is:

$$p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}) \cong (2\pi)^{-(m+1)/2}|\hat{\mathbf{\Sigma}}|^{-1/2} \ \exp\Big\{-\frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T\hat{\mathbf{\Sigma}}^{-1}(\underline{\theta} - \hat{\underline{\theta}})\Big\} \tag{7}$$

where

$\hat{\mathbf{\Sigma}}(\underline{\alpha}) = (\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} + \mathbf{A})^{-1} \in \mathbb{R}^{(m+1)\times(m+1)}$: covariance matrix for $\underline{\theta}$, given $\underline{\alpha}$

$\hat{\underline{\theta}}(\underline{\alpha}) = \hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\mathbf{B}\hat{\underline{y}}(\underline{\alpha})$: the most probable value of parameter $\underline{\theta}$, given $\underline{\alpha}$

$\hat{\underline{y}}(\underline{\alpha}) = \mathbf{\Phi}\hat{\underline{\theta}} + \mathbf{B}^{-1}(\underline{y} - \phi(\mathbf{\Phi}\hat{\underline{\theta}})) \in \mathbb{R}^N$

$\mathbf{B}(\underline{\alpha}) = \text{diag}\{\beta_1, ..., \beta_N\} \in \mathbb{R}^{N\times N}$ with $\beta_n(\underline{\alpha}) = \phi_n(\hat{\underline{\theta}})(1 - \phi_n(\hat{\underline{\theta}}))$

$\mathbf{\Phi} = [\underline{\tau}_1, ..., \underline{\tau}_N]^T \in \mathbb{R}^{N\times(m+1)}$

$\underline{\tau}_n = \underline{\tau}(\underline{x}_n) = [1, \ \underline{x}_n^T]^T \in \mathbb{R}^{m+1}$.

The posterior in (7) contains all that is known about the parameters $\underline{\theta}$ based on the assumed model class $\mathcal{M}(\underline{\alpha})$ and the data $\mathcal{D}_N$.

Bayesian Model Class Selection when using ARD Prior

In the next step, Bayesian model class selection is used to select the most probable hyperparameter $\hat{\underline{\alpha}} \in \mathbb{R}_+^{m+1}$. The most probable model class $\mathcal{M}(\hat{\underline{\alpha}})$ based on data $\mathcal{D}_N$ is given by finding $\hat{\underline{\alpha}}$ that maximizes the probability $p(\underline{\alpha}|\mathcal{D}_N)d\underline{\alpha} \propto P(\mathcal{D}_N|\underline{\alpha})p(\underline{\alpha})d\underline{\alpha}$ for model class $\mathcal{M}(\underline{\alpha})$. If a uniform prior on $\underline{\alpha}$ is considered, then it is equivalent to the maximization of the evidence $P(\mathcal{D}_N|\underline{\alpha})$, which is equivalent to the maximization of $\ln P(\mathcal{D}_N|\underline{\alpha})$ given by:

$$\mathcal{L}(\underline{\alpha}) = \ln P(\mathcal{D}_N|\underline{\alpha})$$

$$
\begin{aligned}
&= \ln \int_{-\infty}^{\infty} P(\mathcal{D}_N, \underline{\theta}|\underline{\alpha}) d\underline{\theta} \\
&= \ln \int_{-\infty}^{\infty} P(\mathcal{D}_N|\underline{\theta}) p(\underline{\theta}|\underline{\alpha}) d\underline{\theta} \\
&\cong -\frac{1}{2}\Big[N \ln 2\pi + \ln|\mathbf{B}^{-1} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T| \ + \underline{y}^T(\mathbf{B}^{-1} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T)^{-1}\underline{y}\Big] \\
&= -\frac{1}{2}\Big[N \ln 2\pi + \ln|\mathbf{C}| + \underline{y}^T\mathbf{C}^{-1}\underline{y}\Big] \qquad (8)
\end{aligned}
$$

where Laplace's asymptotic approximation is used on the integral in (8) expanding about $\hat{\underline{\theta}}(\underline{\alpha})$, the maximum of the integrand, $\mathbf{C} = \mathbf{B}^{-1} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^T$ and $\mathbf{A}(\underline{\alpha})$, $\mathbf{B}(\underline{\alpha})$ and $\mathbf{\Phi}$ are defined as before (see Faul and Tipping 2002).

The maximization of $\mathcal{L}(\underline{\alpha})$ is performed using an iterative procedure as follows. $\mathcal{L}(\underline{\alpha})$ can be re-written by isolating the terms containing α_i:

$$
\begin{aligned}
\mathcal{L}(\underline{\alpha}) &= -\frac{1}{2}\Big[N \ln 2\pi + \ln|\mathbf{C}_{-i}| + \underline{y}^T\mathbf{C}_{-i}^{-1}\underline{y} - \ln\alpha_i \ + \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) \\
&\quad - \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \\
&= \mathcal{L}(\alpha_{-i}) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i) \ + \frac{(\underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y})^2}{\alpha_i + \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i}\Big] \qquad (9)
\end{aligned}
$$

where $\mathbf{C}_{-i}$ is the covariance matrix $\mathbf{C}$ with the components of $\underline{\tau}_i$ removed and so $\mathbf{C}_{-i}$ does not depend on α_i, only on the other components of $\underline{\alpha}$. By setting the derivative of (9) with respect to α_i to zero, the value that maximizes $\mathcal{L}(\underline{\alpha})$ is found to be

$$
\hat{\alpha}_i = \begin{cases} \infty, & \text{if } Q_i^2 \le S_i \\ \frac{S_i^2}{Q_i^2 - S_i}, & \text{if } Q_i^2 > S_i \end{cases} \qquad (10)
$$

where $Q_i = \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{y}$ and $S_i = \underline{\tau}_i^T\mathbf{C}_{-i}^{-1}\underline{\tau}_i$ (Faul and Tipping 2002).

Starting with an initial estimate of $\underline{\hat{\alpha}}$, $\hat{\alpha}_i$ is iteratively calculated from (10) for

each $i = 0, ..., m$, always utilizing the latest estimates for the α_j to evaluate $\mathbf{C}(\underline{\alpha})$, and this procedure is continued until it converges to $\hat{\underline{\alpha}}$. In this process, some of the α_i may become infinite, resulting in a pruning of the corresponding components of $\underline{\tau}_i$ since $\hat{\alpha}_i \to \infty \Rightarrow \hat{\theta}_i \to 0$ and $\hat{\Sigma}_{ii} \to 0$, so $\theta_i \to 0$ from (7). Thus, only the components that have $\hat{\alpha}_i$ finite are used in determining the separating boundary, so that the maximization of the evidence with respect to $\underline{\alpha}$ automatically determines which terms in $f(\underline{x}|\underline{\theta})$ in (1) are relevant for classification.

Prediction Phase

Based on the results from the previous subsection, prediction is performed as follows. Let $\tilde{y}$ denote the unknown label for new feature, $\underline{\tilde{x}}$, then the desired probability is given by:

$$\begin{aligned} P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) &= \int P(\tilde{y}, \underline{\theta}, \underline{\alpha}|\underline{\tilde{x}}, \mathcal{D}_N) d\underline{\theta} d\underline{\alpha} \\ &= \int P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N, \underline{\theta}) p(\underline{\theta}|\mathcal{D}_N, \underline{\alpha}) p(\underline{\alpha}|\mathcal{D}_N) d\underline{\theta} d\underline{\alpha} \end{aligned}$$

Using Laplace's approximation twice:

$$P(\tilde{y}|\underline{\tilde{x}}, \mathcal{D}_N) \cong \int P(\tilde{y}|\underline{\tilde{x}}, \underline{\theta}) p(\underline{\theta}|\mathcal{D}_N, \hat{\underline{\alpha}}) d\underline{\theta} \cong P(\tilde{y}|\underline{\tilde{x}}, \hat{\underline{\theta}}(\hat{\underline{\alpha}})) \tag{11}$$

where $\hat{\underline{\theta}}$, $\hat{\underline{\alpha}}$ are the most probable values for $\underline{\theta}$, $\underline{\alpha}$ based on data $\mathcal{D}_N$ determined as in (7) and (10), respectively, and $P(\tilde{y}|\underline{\tilde{x}}, \hat{\underline{\theta}}(\hat{\underline{\alpha}}))$ is given by (2). Notice that the predictive probability in (11) is controlled by the optimal boundary function $f(\underline{\tilde{x}}|\hat{\underline{\theta}}(\hat{\underline{\alpha}}))$ given by (1).

Near-field versus Far-field Classification Results

Function for Separating Boundary

In a previous study that used a fixed prior (instead of the ARD prior), the three-parameter model given in (12) was found to give the optimal separating boundary function based on the earthquake dataset described in Table 1 (Yamada et al. 2007):

$$\mathcal{M}_1 : \ f(\underline{x}|\hat{\underline{\theta}}) = 6.046 \log_{10} Z_a + 7.885 \log_{10} H_v - 27.091 \tag{12}$$

This corresponds to a model class, denoted $\mathcal{M}_1$ here, that was selected by finding the most probable model class among 255 ($=2^8 - 1$) models consisting of all possible combinations of the 8 features in Table 2 and using a fixed Gaussian prior $p(\underline{\theta}|\mathcal{M})$ for each model class $\mathcal{M}$. The misclassification rates for $\mathcal{M}_1$ are 22.00% and 2.02% for the NS and FS data, respectively.

Since $\mathcal{M}_1$ was estimated by using a *constant* standard deviation of 100 ($= \alpha_i^{-1/2}$) for the Gaussian prior for each θ_i, the proposed method of Bayesian learning with the ARD prior is first applied to a model class with the same features as in (12) but using an *independent* variance α_i for each θ_i ($i = 0, 1, 2$) in the prior. The procedure described in the previous section is applied to the earthquake dataset and the optimal boundary function for this model class $\mathcal{M}_2$ is given in (13): the corresponding misclassification rates are 23.00% and 2.02% for NS and FS data, respectively. The corresponding prior variances are given later.

$$\mathcal{M}_2 : \ f(\underline{x}|\hat{\underline{\theta}}) = 6.129 \log_{10} Z_a + 7.484 \log_{10} H_v - 26.588 \tag{13}$$

Based on the misclassification rates, it could be concluded that the difference in performance between the two three-parameter models (12) and (13) is negligible. However, it is shown later that $\mathcal{M}_1$ is much less probable that $\mathcal{M}_2$ based on the data.

Finally, the proposed methodology of Bayesian learning with the ARD prior is applied to a model containing all 8 features in Table 2. It produces a five-parameter model class $\mathcal{M}_3$ whose optimal separating boundary function is:

$$\begin{aligned} \mathcal{M}_3: \ f(\underline{x}|\hat{\underline{\theta}}) &= 2.055 \log_{10} H_j + 5.350 \log_{10} Z_a + 4.630 \log_{10} H_v \\ &\quad + 1.972 \log_{10} H_d - 30.982 \end{aligned} \tag{14}$$

Note that for $\mathcal{M}_2$, the Bayesian learning algorithm is restricted to have no more than $\log_{10} Z_a$ and $\log_{10} H_v$, the features that are used for $\mathcal{M}_1$, while $\mathcal{M}_3$ selects 4 features from a potential of 8 by automatically pruning the other features. The corresponding misclassification rates for $\mathcal{M}_3$ are 18.00% and 1.85% for NS and FS data, respectively, significantly smaller than those for $\mathcal{M}_1$ and $\mathcal{M}_2$.

The coefficients for the optimal separating boundaries, the prior variances and the corresponding classification results for each model class are summarized in Tables 3, 4 and 5, respectively. The performance of these three model classes are next examined by leave-one-out cross-validation and then by calculating their evidence based on the earthquake data $\mathcal{D}_N$.

Leave-One-Out Cross-Validation

Table 5 shows the classification results for models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ using all 695 records in the earthquake dataset, both for training and predicting the

labels. As shown in this table, $\mathcal{M}_3$ outperforms the two other models on the basis of smaller misclassification rates. For another check on the performance of these three models for predicting the class, leave-one-out cross-validation (LOOCV) is performed.

LOOCV, as the name implies, takes one data point at a time from the whole dataset and then a prediction is made based on the optimal separating boundary determined from the remaining data. This procedure is repeated until each data point has been compared with the corresponding prediction (taken here as the class with the higher predictive probability, that is, the class with probability exceeding 0.5). Actually, LOOCV is equivalent to K-fold cross-validation where K(= 695 here) is equal to the number of data in the original dataset. Note that LOOCV is commonly used in Tikhonov regularization to select the regularizing parameter, but this is handled automatically in the Bayesian approach presented here.

The results of LOOCV for each model class are presented in Table 6. Based on the misclassification rate, which is the ratio of the number of misclassified data to the total number of data, classification model $\mathcal{M}_3$ shows a better performance.

Posterior Probability of Each Model Class

In this section the posterior probability of each model class in the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3\}$ is computed based on the dataset $\mathcal{D}_N$ of 695 records by using Bayes' Theorem:

$$P(\mathcal{M}_i|\mathcal{D}_N,\mathcal{M}) = \frac{P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})}{P(\mathcal{D}_N|\mathcal{M})}$$
$$= \frac{P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})}{\sum_{i=1}^{I} P(\mathcal{D}_N|\mathcal{M}_i)P(\mathcal{M}_i|\mathcal{M})} \tag{15}$$

where $P(\mathcal{D}_N|\mathcal{M}_i)$ is the evidence for $\mathcal{M}_i$, $P(\mathcal{M}_i|\mathcal{M})$ is the prior reflecting the initial choice of the probability of each model class in set $\mathcal{M}$ and the denominator $P(\mathcal{D}_N|\mathcal{M})$ is a normalizing constant. Assigning equal prior probability to each model class, the posterior probability of each model class is proportional to its evidence

$$P(\mathcal{M}_i|\mathcal{D}_N,\mathcal{M}) \propto P(\mathcal{D}_N|\mathcal{M}_i) \tag{16}$$

Using the Theorem of Total Probability, the evidence is calculated from:

$$P(\mathcal{D}_N|\mathcal{M}_i) = \int P(\mathcal{D}_N|\underline{\theta}_i,\mathcal{M}_i)p(\underline{\theta}_i|\mathcal{M}_i)d\underline{\theta}_i \tag{17}$$

This is the average value of the likelihood weighted by the corresponding prior probability over all possible values of the parameters $\underline{\theta}_i$. For a large number of data, an asymptotic approximation can be applied to the integral in (17) (Beck and Yuen 2004):

$$P(\mathcal{D}_N|\mathcal{M}_i) \cong P(\mathcal{D}_N|\hat{\underline{\theta}}_i,\mathcal{M}_i)\frac{(2\pi)^{N_i/2}p(\hat{\underline{\theta}}_i|\mathcal{M}_i)}{\sqrt{|H(\hat{\underline{\theta}}_i)|}} \tag{18}$$

where $\hat{\underline{\theta}}_i$ is the most probable value of $\underline{\theta}_i$ and N_i is the number of parameters in model class $\mathcal{M}_i$. The first factor in (18) is the likelihood and the remaining factors together are the *Ockham factor*. This Ockham factor penalizes more complex models. The Hessian matrix $H(\underline{\theta}_i)$ in (18) is given by the same expression as for $\hat{\Sigma}^{-1}(\underline{\alpha})$ after (7) where each variance α_i^{-1} is given in Table 4. The posterior probabilities for each of $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are presented in Table

7, which shows that $\mathcal{M}_3$ is much more probable than $\mathcal{M}_1$ and $\mathcal{M}_2$ based on the dataset $\mathcal{D}_N$.

There is a refined information - theoretic interpretation (Beck and Yuen 2004; Muto and Beck 2008) of the log evidence that shows that it consists of the difference between a datafit term (the posterior mean of the log likelihood function for the model class) and a relative entropy term (Shannon 1948) which quantifies the amount of information extracted from the data by the model class. It is the latter term that prevents over-fitting to the data and which leads to an automatic Principle of Model Parsimony (Beck and Yuen 2004) when Bayesian updating is performed over a set of model classes, as done here. This information - theoretic interpretation is evident from the asymptotic approximation (18) for large N which shows that the log evidence is approximated by the sum of the log likelihood of the most probable model in the model class and the log Ockham factor, which is an asymptotic approximation for the negative of the relative entropy. This is how it was originally discovered (Beck and Yuen 2004) but more recently it has been proved for the general case (Muto and Beck 2008).

Effect of Prior

As we stated, the likelihood in (18) calculated for a more complex model is usually larger than that for a simpler one, since a more complex model gives a better fit to the data (e.g. see Table 7). Therefore, if a model is selected that maximizes the likelihood alone, it tends to prefer the more complex model and may lead to an over-fitting problem. In the Bayesian learning method, this problem is inherently avoided by employing a prior distribution where

the standard deviation of the prior controls the trade-off between the datafit error and model complexity (Bishop 2006). This trade-off occurs because the posterior probability of a model class depends on the evidence for the model class, which can be expressed as the product of a datafit factor and an Ockham factor, as explained in the previous sub-section.

It is interesting that in the application here $\mathcal{M}_1$ is a 'simpler' model than $\mathcal{M}_3$ (it has fewer parameters) and yet Table 7 shows that its Ockham factor is much smaller than that of $\mathcal{M}_3$, so the 'simpler' model is penalized more than the more complex one. This is a caution that one cannot simply count the number of uncertain parameters N_i in a model class $\mathcal{M}_i$ to judge its complexity. For the same reason, one must be cautious in using simple model selection criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978), since they replace the Ockham factor in (18) with $\exp(-N_i)$ and $\exp(-\frac{1}{2}N_i \ln N)$, respectively.

The reason for the lower Ockham factor for $\mathcal{M}_1$ is that it has much larger prior standard deviations than $\mathcal{M}_3$, so the change in entropy from the very broad prior PDF of $\mathcal{M}_1$ to its narrow posterior PDF is very large, and as a consequence the relative entropy term in the information-theoretic interpretation mentioned in the previous sub-section is larger for $\mathcal{M}_1$ than for $\mathcal{M}_3$. This interpretation shows that the correct measure of 'complexity' for a model class is the amount of information that it extracts from the data, so in this sense, $\mathcal{M}_1$ is actually more complex than $\mathcal{M}_3$.

Concluding Remarks

A novel method of Bayesian learning with the automatic relevance determination (ARD) prior is presented and applied to classify earthquake ground motion data into near-source and far-source. The extracted features correspond to the $\log_{10}$ values of peak jerk, acceleration, velocity and displacement in the horizontal and vertical directions and these are used with Bayesian learning to establish a separating boundary in the feature space. The ARD prior plays an important role by promoting sparsity when selecting the important features (i.e. by utilizing only a small number of relevant features after automatically pruning the remaining features).

The discussion in the previous sub-section and the results presented for the near-source/far-source classification problem demonstrate that broad prior PDFs should be used with caution when defining a model class. An important advantage of using the ARD prior is that model class selection automatically chooses an appropriate prior that does not overly penalize complexity; it provides a balance between the datafit of the model class and its complexity in terms of the amount of information that it extracts from the data.

The selected most probable separating boundary for classification of seismic signals into near-source and far-source is:

$$\begin{aligned} f(\underline{x}_i|\hat{\underline{\theta}}) = {} & 2.055 \log_{10} H_j + 5.350 \log_{10} Z_a + 4.630 \log_{10} H_v \\ & + 1.972 \log_{10} H_d - 30.982 \end{aligned} \tag{19}$$

where H_j, Z_a, H_v and H_d are the horizontal jerk, vertical acceleration, and horizontal velocity and displacement, respectively, of the ground motion record. Based on (19), the probability for new data with features $\tilde{\underline{x}}$ to be classified as

near-source ($\tilde{y} = 1$) or far-source ($\tilde{y} = 0$) is:

$$P(\tilde{y} = 1|\underline{\tilde{x}}, \underline{\hat{\theta}}) = \frac{1}{1 + \exp(-f(\underline{\tilde{x}}|\underline{\hat{\theta}}))} \tag{20}$$

$$P(\tilde{y} = 0|\underline{\tilde{x}}, \underline{\hat{\theta}}) = 1 - P(\tilde{y} = 1|\underline{\tilde{x}}, \underline{\hat{\theta}}) \tag{21}$$

The proposed method is readily applied to real-time analysis of recorded seismic ground motions for near-source and far-source classification since the only calculations involved are those implied by (19) to (21).

In view of the results so far achieved, it can be concluded that it is beneficial to use the proposed Bayesian learning with the ARD prior because it leads to:

- higher correct classification rates (equivalent to a lower misclassification rate) (see Table 5)
- better generalization performance as demonstrated by the leave-one-out cross-validation results (see Table 6)
- the most probable model class based on the calculated posterior probability (see Table 7)

Additional studies are underway in performance-based earthquake engineering in order to apply the method to develop component fragility functions for multiple engineering demand parameters and multiple damage states.

References

Akaike, H. (1974). "A new look at the statistical identification model." *IEEE Trans. Autom. Control*, 19, 716-723.

Allen, R. M., and Kanamori, H. (2003). "The potential for earthquake early warning in Southern California." *Science*, 300, 786-789.

Beck, J. L., and Katafygiotis, L. S. (1998). "Updating models and their uncertainties: Bayesian statistical framework." *J. Eng. Mech.*, 124, 455-461.

Beck, J. L., and Yuen, K. V. (2004). "Model selection using response measurements: a Bayesian probabilistic approach." *J. Eng. Mech.*, 130, 192-203.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, Springer, New York.

Cua, G. B. (2005). "Creating the virtual seismologist: developments in ground motion characterization and seismic early warning." Ph. D. Thesis in Civil Engineering, California Institute of Technology.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*, Wiley-interscience, New York.

Faul, A. C., and Tipping, M. E. (2002). "Analysis of sparse Bayesian learning." *Advances in Neural Information Processing Systems*, 14, 383-389.

Grasso, V. F., Beck, J. L., and Manfredi, G. (2007). "Automated decision procedure for earthquake early warning." *Engineering Structures*, 29, 3455-3463.

Hanks, T. C., and McGuire, R. K. (1981). "The character of high-frequency strong ground motion." *Bull. Seism. Soc. Am.*, 71(6), 2071-2095.

Hartzell, S., and Heaton, T. (1983). "Inversion of strong ground motion and teleseismic waveform data for the fault rupture history of the 1979 Imperial Valley, California, earthquake." *Bull. Seism. Soc. Am.*, 73, 1553-1583.

Honda, R., Aoi, S., Morikawa, N., Sekiguchi, H., Kunugi, T., and Fugiwara, H. (2005). "Ground motion and rupture process of the 2004 mid Niigata prefecture earthquake obtained from strong motion data of K-NET and KiK-net." *Earth Planets Space*, 57, 527-532.

Ji, C., Helmberger, D. V., Wald, D. J., and Ma, K. F. (2003). "Slip history and dynamic implication of 1999 Chi-Chi earthquake." *J. Geophys. Res.*, 108(B9).

Mackay, D. J. C. (1992). "The evidence framework applied to classification networks." *Neural Computation*, 4, 720-736.

Mackay, D. J. C. (1994). "Bayesian non-linear modelling for the prediction competition." *ASHRAE Transactions*, 100, 2, 1053-1062.

Muto, M., and Beck, J. L. (2008). "Bayesian updating and model class selection for hysteretic structural models using stochastic simulation." *J. Vibr. Control*, 14, 7-34.

Oh, C. K., and Beck, J. L. (2006). "Sparse Bayesian learning for structural health monitoring." *Proceedings of the 4th World Conference on Structural Control and Monitoring*, San Diego, CA.

Oh, C. K. (2007). "Bayesian learning for earthquake engineering applications and structural health monitoring." Ph. D. Thesis in Civil Engineering, California Institute of Technology.

Schwarz, G. (1978). "Estimating the dimension of a model." *Ann. Stat.*, 6(2), 461-464.

Sekiguchi, H., and Iwata, T. (2002). "Rupture process of the 1999 Kocaeli, Turkey, earthquake estimated from strong-motion wave forms." *Bull. Seism. Soc. Am.*, 92, 300-311.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27, 379-423 and 623-656.

Tipping, M. E. (2004). "Bayesian inference: An introduction ot principles and practive in Machine Learning." *Advanced Lectures on Machine Learning*, 41-62, Springer.

Tsuboi, S., Komatitsch, D., Ji, C., and Tromp, J. (2003). "Broadband modeling of the 2002 Denali fault earthquake on the Earth Simulator." *Phys. Earth Planet. Interiors*, 139, 305-312.

Vapnik, V. N. (1998). *Statistical learning theory*, Wiley, New York.

Wald, D. J., Heaton, T., and Helmberger, D. V. (1991). "Rupture model of the 1989 Loma Prieta earthquake from the inversion of strong motion and broadband teleseismic data." *Bull. Seism. Soc. Am.*, 81, 1540-1572.

Wald, D. J., and Heaton, T. (1994). "Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake." *Bull. Seism. Soc. Am.*, 84, 668-691.

Wald, D. J., Heaton, T., and Hudnut, K. W. (1996). "A dislocation model of the 1994 Northridge, California, earthquake determined from strong-motion, GPS, and leveling-line data." *Bull. Seism. Soc. Am.*, 86, 49-70.

Wald, D. J. (1996). "Slip history of the 1995 Kobe, Japan, earthquake determined from strong motion, teleseismic, and geodetic data." *J. Phys. Earth*, 44, 489-503.

Yamada, M., Heaton, T., and Beck, J. L. (2007). "Real-time estimation of fault rupture extent using near-source versus far-source classification." *Bull. Seism. Soc. Am.*, 97, 1890-1910.

Fig.1. Shape of Sigmoid Function.

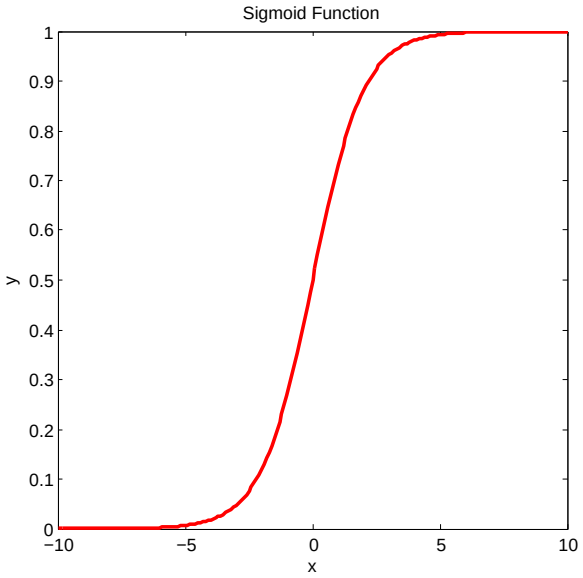

Table 1. Number of Near-source and Far-source Records in Earthquake Dataset Used for Classification (from Yamada et al. 2007).

Earthquake	M_w[a]	NS	FS	Total	Fault Model[b]
Imperial Valley (1979)	6.5	14	20	34	Hartzell and Heaton (1983)
Loma Prieta (1989)	6.9	8	39	47	Wald et al. (1991)
Landers (1992)	7.3	1	112	113	Wald and Heaton (1994)
Northridge (1994)	6.6	17	138	155	Wald et al. (1996)
Hyogoken-Nanbu (1995)	6.9	4	14	18	Wald (1996)
Izmit (1999)	7.6	4	13	17	Sekiguchi and Iwata (2002)
Chi-Chi (1999)	7.6	42	172	214	Ji et al. (2003)
Denali (2002)	7.8	1	29	30	Tsuboi et al. (2003)
Niigataken-Chuetsu (2004)	6.6	9	58	67	Honda et al. (2005)
Total		100	595	695	

[a] Moment magnitude (M_w) is cited from Havard CMT solution.

[b] To classify near-source and far-source station, listed fault models are utilized.

Table 2. Eight Extracted Features (from Yamada et al. 2007).

Ground Motion Feature	Unit
Horizontal Peak Ground Jerk (H_j)	(cm/s^3)
Vertical Peak Ground Jerk (Z_j)	(cm/s^3)
Horizontal Peak Ground Acceleration (H_a)	(cm/s^2)
Vertical Peak Ground Acceleration (Z_a)	(cm/s^2)
Horizontal Peak Ground Velocity (H_v)	(cm/s)
Vertical Peak Ground Velocity (Z_v)	(cm/s)
Horizontal Peak Ground Displacement (H_d)	(cm)
Vertical Peak Ground Displacement (Z_d)	(cm)

Table 3. Coefficients for Optimal Separating Boundary Function for Each Model Class.

$\mathcal{M}$	N_i[a]	1	H_j	Z_j	H_a	Z_a	H_v	Z_v	H_d	Z_d
$\mathcal{M}_1$	3	-27.091	-[b]	-	-	6.046	7.885	-	-	-
$\mathcal{M}_2$	3	-26.588	-	-	-	6.129	7.484	-	-	-
$\mathcal{M}_3$	5	-30.982	2.055	0[c]	0	5.350	4.623	0	1.972	0

[a] N_i is the number of parameters used for each model.

[b] '-' means the corresponding parameters are not considered for each model.

[c] '0' means the corresponding parameters are automatically pruned during training.

Table 4. Prior Covariance Matrix for Each Model Class.

$\mathcal{M}$	Prior Covariance Matrix
$\mathcal{M}_1$	diag[a]$(100^2, 100^2, 100^2)$
$\mathcal{M}_2$	diag$(26.76^2, 6.19^2, 7.57^2)$
$\mathcal{M}_3$	diag$(31.23^2, 2.25^2, 5.48^2, 4.97^2, 2.20^2)$

[a] 'diag' means diagonal matrix with the diagonal elements following.

Table 5. Classification Results for Earthquake Database Using Three Different Model Classes.

	Actual Class	Predicted Class		Total Observations
		Near-source	Far-source	
$\mathcal{M}_1$	Near-source	78(78.00%)	22(22.00%)	100
	Far-source	12(2.02%)	583(97.98%)	595
	Total Predictions	90	605	695
$\mathcal{M}_2$	Near-source	77(77.00%)	23(23.00%)	100
	Far-source	12(2.02%)	583(97.98%)	595
	Total Predictions	89	606	695
$\mathcal{M}_3$	Near-source	82(82.00%)	**18(18.00%)**	100
	Far-source	**11(1.85%)**	584(98.15%)	595
	Total Predictions	93	602	695

Table 6. Misclassification Rates Based on Leave-One-Out Cross-Validation.

Model	Prediction Error
$\mathcal{M}_1$	36/695 (5.18%)
$\mathcal{M}_2$	37/695 (5.32%)
$\mathcal{M}_3$	**31/695 (4.46%)**

Table 7. Posterior Probability Calculation for Bayesian Model Class Selection.

$\mathcal{M}$	ln Ockham[a]	ln Likelihood[a]	ln Evidence[a]	Probability[b]
$\mathcal{M}_1$	-15	-81	-96	0.00
$\mathcal{M}_2$	-10	-79	-89	0.11
$\mathcal{M}_3$	**-12**	**-75**	**-87**	**0.89**

[a] These values are natural logarithms of the Ockham factor, likelihood and evidence, respectively.

[b] Probability is calculated from the evidence on the basis that the $\mathcal{M}_i$ $(i = 1, 2, 3)$ are equally probable a priori.