Statistical Science
2004, Vol. 19, No. 1, 128–139
DOI 10.1214/088342304000000099

Bayesian Methods for Neural Networks and Related Models

D. M. Titterington

Abstract. Models such as feed-forward neural networks and certain other structures investigated in the computer science literature are not amenable to closed-form Bayesian analysis. The paper reviews the various approaches taken to overcome this difficulty, involving the use of Gaussian approximations, Markov chain Monte Carlo simulation routines and a class of non-Gaussian but "deterministic" approximations called variational approximations.

Key words and phrases: Bayesian methods, Bayesian model choice, feed-forward neural network, graphical model, Laplace approximation, machine learning, Markov chain Monte Carlo, variational approximation.

1. INTRODUCTION

This contribution will review the impact that Bayesian analysis has had on the class of models known as (artificial) neural networks. It will also draw attention to Bayesian methodology that has been publicized mainly in the computer science and neural computing literature; some of this work is not directly related to neural networks per se, but is of interest also in the context of other nonsimple types of model, many of which come under the general description of graphical models.

That there are substantial links between statistical modelling and neural networks has become well documented in recent years, through review papers such as those by Cheng and Titterington (1994) and Ripley (1994) and in monographs such as Bishop (1995) and Ripley (1996). The interface is also emphasized in collections such as Kay and Titterington (1999), the impact of the work of researchers from beyond the mainstream statistical community on graphical-model methodology is a major feature of Jordan (1999), more bridges with the machine-learning world are built in the monograph by Hastie, Tibshirani and Friedman (2001) and the increasing cross-fertilization is apparent across a wide range of statistical, neural-network and machine-learning journals.

D. M. Titterington is Professor, Department of Statistics, University of Glasgow, Glasgow, G12 8QQ, Scotland, UK (e-mail: mike@stats.gla.ac.uk).

2. BAYESIAN ANALYSIS OF FEED-FORWARD NEURAL NETWORKS

2.1 The Feed-Forward Neural Network

The feed-forward neural network, otherwise known as the multilayer perceptron, can be thought of as a particular type of nonlinear regression or classification model, in which a set of p input variables $\mathbf{x} = (x_1, \ldots, x_p)$ is related to an output (response) variable y; although in practice y might be multivariate, we shall consider for simplicity only the scalar case. The most familiar version of this structure leads to a response function of the form

$$g\left\{w_{00} + \sum_{j=1}^{m} w_{0j} f\left(w_{j0} + \sum_{k=1}^{p} w_{jk} x_k\right)\right\},$$

where the $\mathbf{w} := \{w_{jk}\}$ are called (*connection*) *weights*. For each k the w_{jk}, for $k = 1, \ldots, p$, correspond to connections from the input variables to the jth of a layer of m *hidden nodes*, and the w_{0j}, for $j = 1, \ldots, m$, correspond to connections from the hidden nodes to the output node. The function $g(\cdot)$ is the so-called *activation function* at the output node and $f(\cdot)$ is the common activation function at each of the hidden nodes. The output from the jth hidden node is

$$z_j = f\left(w_{j0} + \sum_{k=1}^{p} w_{jk} x_k\right).$$

The activation function f is usually taken to be sigmoidal, and therefore nonlinear, the most common

choices being the logistic sigmoid, for which $f(u) = 1/(1+e^{-u})$, giving $0 \le z \le 1$, and the "tanh" function, giving $-1 \le z \le 1$. When the response variable is continuous, the most common version of the model takes the activation function g to be the identity function, so that the output is a linear combination of the outputs from the hidden nodes, and a (Gaussian) white noise error term ε with variance σ^2 is added that makes it perfectly recognizable as a nonlinear regression model:

$$y = g\left\{ w_{00} + \sum_{j=1}^{m} w_{0j} f\left(w_{j0} + \sum_{k=1}^{p} w_{jk} x_k \right) \right\} + \varepsilon.$$

Versions in which y is a categorical variable lead to nonlinear models for discrimination or classification. For example, if y is a 0–1 binary variable, then the model can represent a nonlinear logistic regression model in which

$$\begin{aligned} &\text{logit}\{\text{pr}(y=1)\} \\ &\quad = w_{00} + \sum_{j=1}^{m} w_{0j} f\left(w_{j0} + \sum_{k=1}^{p} w_{jk} x_k \right). \end{aligned}$$

Although these models are strictly speaking parametric, their usage is more commonly nonparametric in spirit; the more hidden nodes are included, possibly arranged in more than one layer, the more flexible is the model at representing a regression surface.

2.2 Bayesian Posterior Inference

In general, the model defines a probability density of the form $p(y|\mathbf{w})$, where the $\mathbf{w}$ are parameters, possibly also including σ^2 as appropriate. Clearly there should be a mention of $\mathbf{x}$ behind the conditioning sign; this will be understood throughout. In practice a (training) dataset D of n realizations of $(y, \mathbf{x})$ is available, providing a likelihood function $p(D|\mathbf{w})$ based on the above $p(y|\mathbf{w})$, and the standard Bayesian paradigm can proceed once a prior $p(\mathbf{w})$ has been specified.

We shall concentrate on the regression problem, in which y is real-valued and the additive noise is Gaussian. In early work it was assumed that, a priori, the weights were all independently $N(0, \alpha^{-1})$, for some positive hyperparameter α, so that, a posteriori,

$$p(\mathbf{w}|D) \propto \exp\left[-\frac{1}{2} \left\{ \frac{E(D,\mathbf{w})}{\sigma^2} + \alpha \sum_{j,k} w_{jk}^2 \right\} \right], \tag{1}$$

in which $E(D, \mathbf{w})$, the "error" function, is the sum of squared residuals corresponding to the model based on $\mathbf{w}$.

For fixed σ^2 and α the maximum a posteriori (MAP) estimate of the weights is therefore the minimizer of

$$\frac{E(D,\mathbf{w})}{\sigma^2} + \alpha \sum_{j,k} w_{jk}^2.$$

Pre-Bayesian "training" of neural networks involved finding $\mathbf{w}$ to minimize $E(D, \mathbf{w})$, equivalent in the probabilistic interpretation to maximum likelihood. As with other highly parameterized or ill-posed problems, this led to overfitting (too much variance, relative to bias), and one strategy for dealing with this was to add on a quadratic penalty function, $\alpha \sum w_{jk}^2$, yielding a version of nonlinear ridge regression with an estimate for $\mathbf{w}$ equivalent to the Bayesian MAP. The constant α was called the *weight-decay* constant and was chosen by one of a number of methods, including cross-validation. Of course, the situation is complicated because the nonlinearity of the model precludes any explicit formula for the MAP for $\mathbf{w}$, nor is there any guarantee that $p(\mathbf{w}|D)$, or for maximum likelihood $-E(D, \mathbf{w})$, has a unique local maximum; in general quite the opposite is the case, with multiple maxima, strategies for model averaging and so on.

The complicated nature of $E(D, \mathbf{w})$ also prevents the posterior density $p(\mathbf{w}|D)$ from having a convenient closed form. A commonly used approximation is to approximate $\log p(\mathbf{w}|D)$ by a quadratic in $\mathbf{w}$, yielding a Gaussian approximation for $p(\mathbf{w}|D)$, especially when trying to calculate integrals as, for example, in obtaining predictive distributions by the Laplace method. If a number of local maxima of $p(\mathbf{w}|D)$ are identified, then a corresponding set of local Gaussian approximations can be combined with weights to provide a Gaussian mixture approximation for $p(\mathbf{w}|D)$. For a proper Bayesian analysis the unknown α and σ^2, the latter of which is largely regarded as a nuisance parameter, have to be dealt with. For σ^2 a noninformative conjugate prior is typically chosen and σ^2 is (easily) integrated out from the resulting joint posterior for $\mathbf{w}$ and σ^2. So far as α is concerned, Buntine and Weigend (1991) integrated it out. The alternative approach is to plug in an empirical Bayes (Type II maximum likelihood, Berger, 1985) estimate obtained by maximizing

$$p(D|\alpha) := \int p(D|\mathbf{w}) p(\mathbf{w}|\alpha)\, d\alpha;$$

this is the key to MacKay's *evidence framework* (MacKay, 1992a–c), discussed later in Section 2.3, and is now perceived to be the method of choice, as indicated there. (A similar approach should perhaps be

taken when dealing with σ^2.) Neal (1996, page 20) points out that the MacKay method turns out to provide more effective predictions.

More flexible priors on $\mathbf{w}$ can be obtained by including more than one hyperparameter. The most extreme situation would be to allow each weight to have its own hyperparameter, in which case

$$p(\mathbf{w}|\alpha) \propto \exp\left\{-\tfrac{1}{2}\sum_{j,k}\alpha_{jk}w_{jk}^2\right\},$$

where now $\alpha = \{\alpha_{jk}\}$; see, for example, Tipping (2001). Although this creates a large number of hyperparameters, it often turns out that many of the empirical Bayes estimates of the α_{jk}'s become very large. The consequence of this is that the posterior densities for the corresponding weights are highly concentrated around zero, the weights effectively disappear from the model and the network can be "pruned" by deleting the associated apparently unnecessary connections in the graph.

This behavior is revealed by the following very simple example. Consider a two-parameter multiple linear regression problem in which the data satisfy

$$\mathbf{y} = \mathbf{a}_1 w_1 + \mathbf{a}_2 w_2 + \varepsilon,$$

where $\mathbf{y}$ represents a vector of observations, w_1 and w_2 are scalar parameters, $\mathbf{a}_1$ and $\mathbf{a}_2$ are column vectors of covariates and now $\varepsilon \sim N(0, I)$ is a vector of "errors." Suppose also that $\mathbf{a}_1^T\mathbf{a}_2 = 0$ and that (w_1, w_2) have prior density proportional to $\sqrt{(\alpha_1\alpha_2)}\exp\{-\frac{1}{2}(\alpha_1 w_1^2 + \alpha_2 w_2^2)\}$. Then simple manipulation shows that the plug-in estimate of α_j, for $j = 1, 2$, is

$$\hat{\alpha}_j = (\mathbf{a}_j^T\mathbf{a}_j)^2/\{(\mathbf{y}^T\mathbf{a}_j)^2 - \mathbf{a}_j^T\mathbf{a}_j\}.$$

To find out what happens with large samples, replace $(\mathbf{y}^T\mathbf{a}_j)^2$ by its expectation, which is easily shown to be $w_j^2(\mathbf{a}_j^T\mathbf{a}_j)^2 + (\mathbf{a}_j^T\mathbf{a}_j)$, so that, approximately,

$$\hat{\alpha}_j = 1/w_j^2;$$

a small true w_j will lead to a large value of the hyperparameter.

In Neal and MacKay's technique of automatic relevance determination (ARD) (MacKay, 1995) a single hyperparameter is associated with all weights corresponding to connections from a particular input variable, so that any pruning that results amounts to the omission of covariates that are "irrelevant" to the prediction problem.

The natural alternative to developing analytical approximations to the complicated $p(\mathbf{w}|D)$ is of course to use modern simulation techniques. If conjugate gamma and inverse gamma priors are chosen for α and σ^2, respectively, then Gibbs-sampling steps are available for updating α and σ^2, but the complexity of $E(D, \mathbf{w})$ prevents this for $\mathbf{w}$. Instead, Neal (1996) promotes and develops the hybrid Monte Carlo technique of Duane, Kennedy, Pendelton and Roweth (1987), favoring this over other possible approaches such as Metropolis. Suppose we write, for fixed values of α and σ^2,

$$p(\mathbf{w}|D) \propto \exp\{-F(\mathbf{w})\},$$

where $F(\mathbf{w})$ is referred to as an *energy function* in statistical physics terminology. The above expression is also that of the full conditional density of $\mathbf{w}$ given D, α and σ^2, from which we wish to sample. For each element of $\mathbf{w}$ we introduce a *momentum variable*, all of which together form a collection $\mathbf{v}$, say, and a *kinetic energy*

$$K(\mathbf{v}) := \sum_{j,k} v_{jk}^2/(2m_{jk})$$

is defined, where the $\{m_{jk}\}$ are open to choice and used to define a *Hamiltonian function*

$$H(\mathbf{w}, \mathbf{v}) := F(\mathbf{w}) + K(\mathbf{v}).$$

The sampling procedure then is split into stages of simulating Hamiltonian dynamics in discretized fictitious time. The aim here is to sample values of $\mathbf{w}$ and $\mathbf{v}$ with the value of $H(\mathbf{w}, \mathbf{v})$ fixed. However, this is only possible with continuous-time simulation, which is not practicable. The discrete-time procedure leads to sampled values $(\mathbf{w}^*, \mathbf{v}^*)$, say, and is followed by a simple Metropolis step that either retains the initial $(\mathbf{w}, \mathbf{v})$ or accepts $(\mathbf{w}^*, -\mathbf{v}^*)$, the latter to occur with probability

$$\min\left\{1, \exp[-\{H(\mathbf{w}^*, -\mathbf{v}^*) - H(\mathbf{w}, \mathbf{v})\}]\right\}.$$

The Metropolis step corrects the bias in H created by the time-discretization of the Hamiltonian dynamics. Neal (1996) discusses in detail issues such as the discretization used in the Hamiltonian dynamics; decisions have to be made about the step size in the discretization of fictitious time and the number L of time steps to be used before undertaking the Metropolis step. The case of $L = 1$ corresponds to Langevin Monte Carlo, and Neal (1996) illustrates how allowing L to be greater than 1 enables better exploration of the distribution of interest.

2.3 Bayesian Model Choice

As with any class of complicated models, model selection is an important issue for feed-forward networks. Which input variables are important? Which connections can be deleted? (We have already seen in Section 2.2 that ARD represents one approach to this.) How many hidden variables are necessary? Should more than one layer of hidden nodes be used? In the literature, the non-Bayesian methods that have been used include "the usual suspects," such as cross-validation and Akaike's AIC. There is now the opportunity to try out the recent ideas from Spiegelhalter, Best, Carlin and van der Linde (2002); it is interesting that some of the notions in that paper have in fact been stimulated by material from the machine-learning literature (Moody, 1992; Ripley, 1995).

The other Bayesian approach to the comparison of models is of course to use Bayes factors, and to base intermodel inference on the relative values of

$$p(M_r|D) \propto p(D|M_r)p(M_r),$$

for $m = 1, \ldots, R$, where $M_1, \ldots, M_R$ are R competing models, the $p(M_r)$'s are the models' "prior" probabilities, and for each r

$$p(D|M_r) = \int p(D|\mathbf{w})p(\mathbf{w})\,d\mathbf{w},$$

the marginal likelihood of the data, where strictly speaking M_r should appear behind the conditioning signs on the right-hand side. When $p(\mathbf{w})$ involves hyperparameters α, the above integration leads to $p(D|M_r, \alpha)$. In his *evidence framework*, an argument based on approximating the relevant integral amounts to MacKay (1992a–c) substituting the α that maximizes $p(D|M_r, \alpha)$ to obtain a well-defined $p(D|M_r)$ for use in model choice, and this approach has been adopted by many of the writers in the machine-learning literature. If the models are believed a priori to be equally probable, then the support for model M_r relative to model M_s should be based on the Bayes factor $p(D|M_r)/p(D|M_s)$. MacKay (1992a) also notes that, if $p(\mathbf{w}|D)$ is highly peaked around the MAP, $\mathbf{w}_{\text{MAP}}$, with "width" Δ_{post}, and if the prior is relatively flat, with "width" Δ_{prior}, then, approximately,

$$p(D|M_r) = p(D|w_{\text{MAP}})\frac{\Delta_{\text{post}}}{\Delta_{\text{prior}}},$$

the ratio of the Δ's representing an "Ockham factor"; more highly parameterized models will be automatically penalized in the model-selection exercise because their Ockham factors will be small.

The above discussion is quite general. MacKay (1992b, c) has used it with the corresponding Gaussian approximations in the context of feed-forward networks. Others have used it in other contexts, as reported in Section 2.4. Instead of plugging in an α, an alternative approach would appear to be to average it out with respect to a hyperprior, as proposed by Buntine and Weigend (1991), who then construct a Gaussian approximation so far as $\mathbf{w}$ is concerned. MacKay (1999) compares the two approaches at a general level, showing among other things that the former approach is more robust in the context of ill-posed problems and indeed that the latter method can lead one into serious difficulties, such as yielding nonsensical posterior modes.

Paige and Butler (2001) consider model choice for feed-forward networks with one hidden layer, the tanh activation function, and one "skip layer," which leads to an extra term in the response function that is linear in the covariates. They note that Bayes factors alone are not enough for selecting a parsimonious model, it being better to base choice on modal estimates used in the Laplace approximation used to calculate the model probabilities. Lee (2001) considers feed-forward networks for classification and uses a Bayesian random search to do Markov chain Monte Carlo (MCMC) over the total space of all models, with a Metropolis step used to move between models. Incidentally, Lee (2003) remarks that proper priors are hard to come by for feed-forward networks. Instead he advocates the use of a flat prior on "restricted" parameter space. Vila, Wagner and Neveu (2000) base model choice for neural networks on predictive power.

Neal (1996) notes that the choice of the number of hidden units in a feed-forward network can be sidestepped by allowing the number to be infinite, in that the model corresponding to this limiting case turns out to be a Gaussian process model, which can be parameterized and analyzed in its own right; see also Williams (1998). As a result Neal's view is that, if a neural network model is to be used, then in practice one should simply incorporate as many hidden units as can be dealt with computationally, rather than embark on a complicated search over the space of numbers of hidden units.

2.4 Further Discussion of Application of the Bayesian Paradigm

In this section we briefly list the variety of applications of the standard Bayesian paradigm to the

feed-forward network and other models. Inevitably approximations, deterministic or simulation-based, are necessary, and hyperparameters have had to be dealt with.

The publications of MacKay and Neal have been highly influential in the computer science literature, and Bishop (1995, Chapter 10), Thodberg (1996), Husmeier, Penny and Roberts (1999) and Lampinen and Vehtari (2001) provide reviews of that framework. Thodberg (1996) emphasizes the possible use of "committees" of network models, in which the predictions from different models are averaged with weights proportional to the "evidences"; for further discussion of committees and ARD, see Penny and Roberts (1999). Husmeier, Penny and Roberts (1999) report an empirical study in which ARD fails, Husmeier (2000) applies the approach to a version of density estimation and Wright (1999) considers the case where the inputs are subject to uncertainty.

A number of papers apply MacKay's paradigm to particular application areas: Gencay and Qi (2001) apply the approach to problems in finance; Medeiros, Veiga and Pedreira (2001) model exchange rates; Zhang et al. (2001) model wavelet series by feed-forward networks in forecasting futures trading; Edwards et al. (1999) apply the approach to the papermaking industry, showing in particular that overfitting can still occur in spite of the influence of the Ockham factor; and Vivarelli and Williams (2001) consider problems in image classification.

Müller and Ríos-Insua (1998) and Holmes and Mallick (1998) investigate up-to-date Monte Carlo ways of selecting the complexity of neural networks; both papers apply the reversible jump MCMC method of Green (1995), the former to multilayer perceptrons and the latter to radial basis function (RBF) networks, in which the regression function is a linear combination of basis functions (such as Gaussian density kernels) for which the locations and number have to be selected. Andrieu, de Freitas and Doucet (2001) also apply reversible jump MCMC to radial basis function networks, providing a convergence theorem for the reversible jump procedure; see also Andrieu, de Freitas and Doucet (2000) and Konishi, Ando and Imoto (2004). de Freitas, Niranjan, Gee and Doucet (2000) apply sequential Monte Carlo methods to both feed-forward networks and variable dimension RBF networks; some parameters can be integrated out and particle filtering is carried out on the rest, using importance sampling and MCMC steps on each particle to avoid sample depletion.

There are other network-like models in the literature, and Bayesian methods have featured in their recent development. Kwok (1999, 2000) applies MacKay's evidence framework to the case of support vector machines, in which the regression function is a linear combination of kernels, one centered on each data-point and with weights chosen to optimize a particular penalized, nondifferentiable loss function. Typically many of the optimal weights turn out to be zero, thereby providing a parsimonious fitted model; the data-points corresponding to the nonzero weights identify the so-called support vectors (see also Van Gestel et al., 2002). Chu, Keerthi and Ong (2001) carry out the same program of work using a modified, differentiable loss function. Tipping (2000) introduces the relevance vector machine as a version of the support vector machine with a direct probabilistic interpretation immediately open to Bayesian analysis. He uses the evidence framework with a hyperparameter for each weight. In empirical work typically very many weights go to zero, providing very parsimonious fitted models, although he admits that the training phase is complex; see also Chen, Gunn and Harris (2001). The hierarchical mixture of experts model (Jordan and Jacobs, 1994) is a mixture model with added structure, in particular incorporating covariates; Bayesian analysis of this model, including the use of MCMC methodology, is covered by Peng, Jacobs and Tanner (1996) and Jacobs, Peng and Tanner (1997). Luttrell (1994) and Utsugi (1997) apply Bayesian methods to Kohonen's self-organizing map (SOM), the latter paper giving an interpretation of the SOM prescription as an approximate MAP estimator. Utsugi (1998) applies MacKay's procedures to the SOM. Bayesian treatments are provided of the so-called generative topographic mapping by Bishop, Svensén and Williams (1998) and of the type of learning rule used in the Hopfield network by Sommer and Dayan (1998).

The phrase "Bayesian neural networks" in the title of a paper generally means "Bayesian approach to feed-forward networks." There is of course a whole class of models known as Bayesian networks, also known as belief networks, causal networks and influence diagrams, and representable as directed acyclic graphs (DAGs). The various articles on the Bayesian analysis of such models include Heckerman (1999), Geiger, Heckerman and Meek (1999) and references therein.

3. THE USE OF VARIATIONAL APPROXIMATIONS IN BAYESIAN INFERENCE

3.1 Fundamentals of the Variational Approach

As in much of the recent mainstream statistical literature about Bayesian methods, the complexity of the models and the lack of computationally simple exact posterior and predictive distributions have been central features in the context of these neural and related networks. Apart from the use of the (analytical) Laplace approximation, the statistical literature has been dominated by the development of simulation-based approximations to the underlying distributions or functionals thereof, based on MCMC algorithms, particle filters and so on. MCMC methods have the advantage that, all being well, the resulting approximations become asymptotically "correct," if enough simulations are done and if the underlying model does represent the truth. The more complicated the model, however, especially if there are many parameters and/or hyperparameters, the more expensive become these Monte Carlo approaches, in terms of time and storage. Issues here include monitoring the convergence of Monte Carlo schemes. As an alternative to the use of these stochastically generated approximations, the computer-science literature has created a body of work about deterministic, so-called variational approximations.

At a very general level, suppose that a model includes observed items D and unobserved, or missing or latent or hidden, items $\mathbf{u}$. Suppose also that we are interested in $p(\mathbf{u}|D)$, that the latter is very complicated and that we are prepared to use instead a more amenable approximation $q(\mathbf{u})$; obviously, q will depend on D but we omit explicit mention of that, for clarity. We shall base our choice of an optimal q on the Kullback–Leibler directed divergence,

$$KL(q, p_D) := E_q \log(q/p) = \int q(\mathbf{u}) \log\left\{\frac{q(\mathbf{u})}{p(\mathbf{u}|D)}\right\} d\mathbf{u},$$

where the integral becomes a summation if $\mathbf{u}$ is discrete; p_D denotes the density $p(\cdot|D)$. Clearly, the optimal solution is to take $q(\mathbf{u}) = p(\mathbf{u}|D)$ but, to make q amenable, we must impose some simplifying structure on q and optimize subject to the consequent constraints. One might be highly prescriptive and assume that q is, say, Gaussian, so that one is left with the task of optimizing a mean vector and covariance matrix, but extra flexibility is obtained if, to some extent, the choice of the distributional types making up q also falls out of the optimization, thereby justifying the use of the word "variational," with further, lower-level optimization of the relevant "variational" parameters. We shall see that often the type of structure imposed on q corresponds to its taking a *factorized* form, with a view to facilitating the expectation operation in $KL(q, p_D)$, in other words assuming some sort of independence structure among the unobserved items $\mathbf{u}$. Ideally, the forms of the constituent factors will be chosen optimally.

The same solution can be derived from a different motivation, namely that of obtaining a lower bound approximation to the marginal probability (density) $p(D)$ of D. This follows because

$$\log p(D) = \int q(\mathbf{u}) \log\left\{\frac{p(D, \mathbf{u})}{q(\mathbf{u})}\right\} d\mathbf{u} + KL(q, p_D),$$

as can be seen by combining the two terms on the right-hand side. The properties of the Kullback–Leibler divergence then both provide the desired lower bound, in that then

$$\log p(D) \geq \int q(\mathbf{u}) \log\left\{\frac{p(D, \mathbf{u})}{q(\mathbf{u})}\right\} d\mathbf{u}, \tag{2}$$

and demonstrate that a q that minimizes $KL(q, p_D)$ provides the best lower bound for $\log p(D)$. The right-hand side of (2) is known in statistical physics as the *free energy* associated with q.

This source of variational approximations has been used in both likelihood and Bayesian contexts, in both cases with D representing the observed data. If in incomplete-data problems $\mathbf{u}$ represents missing values or latent variables, as for example in mixture models, hidden Markov chain models or factor-analysis models, then the method provides a lower bound for the observed-data loglikelihood $\log p(D|\mathbf{w})$ for any *fixed* value of the parameter vector $\mathbf{w}$; for details of this use of variational approximations in likelihood analysis, see, for example, Jordan, Ghahramani, Jaakkola and Saul (1999), which includes references to many contributions by Jordan and others, and Hall, Humphreys and Titterington (2002). Jaakkola (2001) motivates variational approximations, but not through the Kullback–Leibler measure; see Gibbs and MacKay (2000) for an application. An application relevant to the other motivation for the approximations, namely obtaining through $KL(q, p_D)$ a workable approximation to the conditional density of missing values given observed data, is in algorithms for finding maximum likelihood estimates in incomplete-data problems. In particular the E-step of the EM algorithm requires one to average with respect to a $p_D(\mathbf{u})$, and approximations

of the type considered here have indeed been found useful in this context (Zhang, 1992, 1993; Archer and Titterington, 2002; Ghahramani and Jordan, 1997; Humphreys and Titterington, 2000b). The method has strong connections to the so-called mean-field approximations in statistical physics, particularly in cases where the selected q is an independence model, with each factor parameterized by the corresponding marginal mean. Much could be said about these and more refined approximations in non-Bayesian contexts, and work is in progress to review them elsewhere, but we restrict attention in this paper to directly Bayesian applications and invite the reader to browse among the papers collected in Opper and Saad (2001). However, we do point out that Neal and Hinton (1999) have used (2) as the basis of a new rationale for the EM algorithm and that there are links back to Csiszár and Tusnády (1984), as hinted at in my discussion of Meng and van Dyk (1997).

3.2 Applications in Bayesian Inference

Although in the more directly Bayesian implementation of the method one could simply adapt the above methodology to obtain an approximation to the posterior density at a particular value of the parameters $\mathbf{w}$, it is more natural to take the unobservables $\mathbf{u}$ to contain both hidden or latent variables $\mathbf{z}$, say, and unknown aspects of the model. The latter will normally constitute the unknown parameters $\mathbf{w}$, but there might be other features too, such as hyperparameters and, say, the number k of components to be included in a mixture model fitted to the data D. The motivation based on $KL(q, p_D)$ is therefore to obtain a best approximation, with a prescribed structure to the joint distribution of $\mathbf{w}$ and $\mathbf{z}$ given D, whereafter appropriate marginalization leads to an approximation to $p(\mathbf{w}|D)$ itself. Of course, if the approximating q is specified to represent independence between $\mathbf{w}$ and $\mathbf{z}$, then this marginalization will be trivial. The resulting optimal q then yields, as a result of the other motivation, a lower bound on the marginal probability of the observed data, that is, the evidence which might subsequently be used to approximate Bayes factors if model selection is an issue.

The Bayesian version of inequality (2) is

$$\log p(D) \geq \int q(\mathbf{z}, \mathbf{w}) \log \left\{ \frac{p(D, \mathbf{z}, \mathbf{w})}{q(\mathbf{z}, \mathbf{w})} \right\} d\mathbf{z}\, d\mathbf{w}.$$

The first stage in choosing a special structure for q is to assume that it factorizes into a factor associated with $\mathbf{z}$ and one for $\mathbf{w}$, so that $q(\mathbf{z}, \mathbf{w}) = q_{\mathbf{z}}(\mathbf{z}) q_{\mathbf{w}}(\mathbf{w})$, in which $q_{\mathbf{z}}$ and $q_{\mathbf{w}}$ are themselves densities. This gives

$$\log p(D) \geq \int q_{\mathbf{w}}(\mathbf{w})\, d\mathbf{w} \left[\int q_{\mathbf{z}}(\mathbf{z}) \log \left\{ \frac{p(D, \mathbf{z}|\mathbf{w})}{q_{\mathbf{z}}(\mathbf{z})} \right\} d\mathbf{z} + \log \left\{ \frac{p(\mathbf{w})}{q_{\mathbf{w}}(\mathbf{w})} \right\} \right].$$

Consider, as do Ghahramani and Beal (2001) (see also Sato, 2001) the case where the complete-data distribution belongs to the exponential family, so that

$$p(D, \mathbf{z}|\mathbf{w}) = f(D, \mathbf{z}) g(\mathbf{w})^s \exp\{\theta(\mathbf{w})^T \mathbf{t}(D, \mathbf{z})\},$$

where $\theta(\mathbf{w})$ are the natural parameters and $\mathbf{t}(D, \mathbf{z})$ are the complete-data sufficient statistics. Suppose also that

$$p(\mathbf{w}) \propto g(\mathbf{w})^\alpha \exp\{\theta(\mathbf{w})^T \beta\},$$

where α and β are hyperparameters, is the appropriate conjugate prior. Then the following theorem (Ghahramani and Beal, 2001) follows directly from the properties of the Kullback–Leibler directed divergence.

THEOREM. *The optimal factorized approximation to $q(\mathbf{w}, \mathbf{z})$ has factors of the form*

$$q_{\mathbf{w}}(\mathbf{w}) \propto g(\mathbf{w})^{\tilde{\alpha}} \exp\{\theta(\mathbf{w})^T \tilde{\beta}\},$$

where $\tilde{\alpha} = \alpha + s$ and $\tilde{\beta} = \beta + E_{q_{\mathbf{z}}} \mathbf{t}(D, \mathbf{z})$, and

$$q_{\mathbf{z}}(\mathbf{z}) \propto f(D, \mathbf{z}) \exp\left[E_{q_{\mathbf{w}}} \{\theta(\mathbf{w})\}^T \mathbf{t}(D, \mathbf{z}) \right].$$

Thus $q_{\mathbf{w}}(\mathbf{w})$ belongs to the complete-data conjugate family. Since the theorem does not decouple $q_{\mathbf{w}}$ and $q_{\mathbf{z}}$, the optimal factorization and the relevant hyperparameters cannot be calculated without iterative numerical methods; the usual practice is to calculate updates for the factors $q_{\mathbf{z}}$ and $q_{\mathbf{w}}$ alternately, in the spirit respectively of the E- and M-steps of the EM algorithm.

An important special case is when the data represent n independent and identically distributed observations, $D = (y_1, \ldots, y_n)$, in which case $s = n$ and the sufficient statistics take the form

$$\mathbf{t}(D, \mathbf{z}) = \sum_i \mathbf{t}'(y_i, z_i).$$

In this case q_z factorizes into $q_z(z) = \prod_i q_{z_i}(z_i)$, for which the optimal solutions satisfy

$$q_{z_i}(z_i) \propto f'(y_i, z_i) \exp\left[E_{q_{\mathbf{w}}} \{\theta(\mathbf{w})\}^T \mathbf{t}'(y_i, z_i) \right].$$

Here f' is such that $f(D, \mathbf{z}) = \prod_i f'(y_i, z_i)$.

In many of the applications of this approach, the missing $\mathbf{z}$ are discrete variables, often indicators, so

the relevant integrations become summations. The key advantage of the variational approach is that complex multiple integrals or summations are reduced to simple integrals or summations. However, implementation of the methods is not completely straightforward and there is a basic drawback as we now show with what is arguably the simplest version of the problem.

Humphreys and Titterington (2000a) consider the very simple case of a mixture of two known densities, with unknown mixing weight w. For any observation, the missing quantity is the mixture-component membership indicator, a Bernoulli variable, so that the conjugate prior is a Beta density. The same is true therefore of the variational approximation q_w to the posterior distribution of the mixing weight, and the q_{z_i} gives predictive probabilities of component membership for the ith observation. Humphreys and Titterington (2000a) show that computation of the optimal hyperparameters required to minimize the relevant Kullback–Leibler divergence involves iterative calculations that require the repeated evaluation of the digamma function. The example also reveals that the approach "fails" in a crucial way, in that the true posterior and the variational approximation are essentially different; the former is a complicated mixture of Beta densities whereas the latter is a pure Beta. Humphreys and Titterington (2000a) also note that, if a Beta approximation is required, then recursive methods already exist for which the hyperparameters are calculable trivially (Titterington, Smith and Makov, 1985, Chapter 6), although they admit that the resulting inference is dependent on the order in which the observations are processed.

This variational treatment has been applied in a variety of contexts, including feed-forward neural networks (Hinton and van Camp, 1993; Barber and Bishop, 1998), hidden Markov chains (MacKay, 1997), Gaussian mixtures with the inclusion of a prior distribution on the number of mixture components (Attias, 1999b), more general graphical models (Attias, 2000), mixtures of experts models (Waterhouse, MacKay and Robinson, 1996), principal components analysis (Bishop, 1999), mixtures of factor analyzers (Ghahramani and Beal, 2000), independent component analysis (Miskin and MacKay, 2000; Choudrey and Roberts, 2001; Chan, Lee and Sejnowski, 2003), independent factor analysis (Attias, 1999a), support vector machines (Seeger, 2000) and relevance vector machines (Bishop and Tipping, 2000, who incorporate hyperparameters as well as parameters into $\mathbf{u}$). Humphreys and Titterington (2001) develop recursive (on-line) treatments of some structured mixtures and hidden Markov models. Sato (2001) considers on-line versions for the case where the complete data are modelled by an exponential family distribution. de Freitas, Højen-Sørensen, Jordan and Russell (2001) combine the deterministic and stochastic approaches to approximation by suggesting the use of variational approximations as proposal distributions in Metropolis versions of MCMC. Corduneanu and Bishop (2001) applied the approach to the analysis of finite Gaussian mixtures, subsuming the component membership indicators of the observations and the parameters of the Gaussian component densities within the unobserved items u but treating the mixing weights π, say, as hyperparameters. The variational method was used to obtain a lower bound to the marginal loglikelihood $\log\{p(D|\pi)\}$, which was then maximized with respect to π by an EM-type iteration. It was initially assumed that the model contained a large number of components, with the idea that, in the spirit of automatic relevance determination, redundant components would drop out and a model with an appropriate number of components would result. Ueda and Ghahramani (2002) also treat Gaussian mixtures. Bishop, Spiegelhalter and Winn (2003) describe a software implementation of variational inference for Bayesian networks.

If we do not assume an exponential family model for the complete data together with the relevant conjugate prior, then a convenient variational approximation does not fall out. If we return to the general framework in which the missing items are represented by $\mathbf{u} = \{u_k\}$ and if we choose a factorized form $q = \prod_k q_{u_k}(u_k)$ for q, then the optimal solution is to take

$$q_{u_k}(u_k) = \frac{\exp[E_{\setminus k} \log\{p(D, \mathbf{u})\}]}{\int \exp[E_{\setminus k} \log\{p(D, \mathbf{u})\}]\, du_k}, \tag{3}$$

where $E_{\setminus k}$ denotes expectation with respect to the joint, factorized density

$$\prod_{l \neq k} q_{u_l}(u_l).$$

However, except in simple cases such as those involving exponential family models with conjugate priors, the right-hand side of (3) does not have a closed-form expression. In some cases perhaps, as Ghahramani and Beal (2001) point out, it might be possible to approximate a complete-data model of interest adequately by an exponential-family model. In Bishop and Tipping's (2000) variational treatment of relevance vector machines as applied to classification problems, they deal with the lack of a convenient fully conjugate structure

by employing a further variational approximation introduced by Jaakkola and Jordan (2000) in their approach to the Bayesian analysis of logistic regression models.

So far as comparison with stochastic "approximation" by MCMC is concerned, in the finite-sample case the MCMC method in principle is better, in that, provided enough effort is put into the simulation, an arbitrarily good approximation to the true posterior density can be obtained. This clearly will not happen with the deterministic variational approach, as the simple mixture example shows. It is an issue of current work to consider what happens in the case of large samples; in circumstances where the true posterior converges to a Gaussian distribution, does the variational approximation converge to the same Gaussian? For some comments in this direction see Attias (1999a, b).

4. CONCLUDING REMARKS

The Bayesian work on neural networks and similar models so far has involved both the application of the standard Bayesian paradigm with the help of "traditional" approximation ideas such Laplace's approximation and MCMC, and the development of new approximation ideas such as the variational methods. The area is one of considerable current activity and statisticians are strongly encouraged both to explore the relevant journals and conference proceedings in computer science and to play active roles in future developments.

So far as the variational approach is concerned, it should be fruitful to investigate potentially more refined approximations than those discussed above. Bishop, Lawrence, Jaakkola and Jordan (1998) use mixture approximations; Humphreys and Titterington (2000b) use truncated Bahadur expansions; more generally Bethe–Kikuchi methods involve less fully factorized approximations and should lead to improvements. So far these methods have not been developed in the standard Bayesian scenario, but much has been done in other statistical contexts; see, for example, Yedidia, Freeman and Weiss (2001) and Tanaka, Inoue and Titterington (2003). Drawbacks are that they do not provide guaranteed bounds and that they will be much more computationally intensive than the Kullback–Leibler-based approximations described above. Other approximations that do provide sharper bounds could evolve from ideas in Leisink and Kappen (2001), but again so far the development has not covered the Bayesian approach. In addition, the credentials of the variational approximations should continue to be investigated. Current work with Bo Wang indicates that in some problems, such as Bayesian inference for mixing weights in mixture distributions, the variational posterior mode is "consistent," whereas in other problems, such as some simple nonlinear state-space models, there is a lack of consistency. Finally, current work with Clare McGrory is investigating the use of variational approximations in the calculation of Spiegelhalter et al.'s (2002) DIC in problems where exact calculation of the criterion is not feasible.

ACKNOWLEDGMENTS

This work was supported by a grant from the U.K. Science and Engineering Research Council. I am very grateful for feedback from Chris Bishop, Mike Jordan, David MacKay and Radford Neal about an earlier version of this paper.

REFERENCES

ANDRIEU, C., DE FREITAS, J. F. G. and DOUCET, A. (2000). Robust full Bayesian methods for neural networks. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) **12** 379–385. MIT Press.

ANDRIEU, C., DE FREITAS, N. and DOUCET, A. (2001). Robust full Bayesian methods for radial basis networks. *Neural Computation* **13** 2359–2407.

ARCHER, G. E. B. and TITTERINGTON, D. M. (2002). Parameter estimation for hidden Markov chains. *J. Statist. Plann. Inference* **108** 365–390.

ATTIAS, H. (1999a). Independent factor analysis. *Neural Computation* **11** 803–851.

ATTIAS, H. (1999b). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. Uncertainty in Artificial Intelligence* 21–30. Morgan Kaufmann, San Mateo, CA.

ATTIAS, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) **12** 209–215. MIT Press.

BARBER, D. and BISHOP, C. M. (1998). Variational learning in Bayesian neural networks. In *Neural Networks and Machine Learning* (C. M. Bishop, ed.) 215–237. Springer, New York.

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York.

BISHOP, C. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon, Oxford.

BISHOP, C. M. (1999). Variational principal components. In *Proc. 9th Internat. Conf. Artificial Neural Networks* **1** 509–514. Institution of Electrical Engineers, London.

BISHOP, C. M., LAWRENCE, N., JAAKKOLA, T. and JORDAN, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. In *Advances in Neural Information Processing Systems* (M. I. Jordan, M. J. Kearns and S. A. Solla, eds.) **10** 416–422. MIT Press.

BISHOP, C. M., SPIEGELHALTER, D. J. and WINN, J. (2003). VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems* (S. Becker, S. Thrun and K. Obermayer, eds.) **15** 793–800. MIT Press.

BISHOP, C. M., SVENSÉN, M. and WILLIAMS, C. K. I. (1998). Developments of the generative topographic mapping. *Neurocomputing* **21** 203–224.

BISHOP, C. M. and TIPPING, M. E. (2000). Variational relevance vector machines. In *Proc. 16th Conf. Uncertainty in Artificial Intelligence* (C. Boutilier and M. Goldszmidt, eds.) 46–53. Morgan Kaufmann, San Mateo, CA.

BUNTINE, W. and WEIGEND, A. (1991). Bayesian backpropagation. *Complex Systems* **5** 603–643.

CHAN, K., LEE, T.-W. and SEJNOWSKI, T. J. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation* **15** 1991–2011.

CHEN, S., GUNN, S. R. and HARRIS, C. J. (2001). The relevance vector machine technique for channel equalization application. *IEEE Trans. Neural Networks* **12** 1529–1532.

CHENG, B. and TITTERINGTON, D. M. (1994). Neural networks: A review from a statistical perspective (with discussion). *Statist. Sci.* **9** 2–54.

CHOUDREY, R. and ROBERTS, S. J. (2001). Variational Bayesian independent component analysis with flexible sources. Technical Report PARG-01-03, Dept. Engineering Science, Univ. Oxford.

CHU, W., KEERTHI, S. S. and ONG, C. J. (2001). Bayesian inference in support vector regression. Technical Report CD-01-15, Natl. Univ. Singapore.

CORDUNEANU, A. and BISHOP, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Proc. 8th Internat. Conf. Artificial Intelligence and Statistics* (T. Richardson and T. Jaakkola, eds.) 27–34. Morgan Kaufmann, San Mateo, CA.

CSISZÁR, I. and TUSNÁDY, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions* (suppl. issue) **1** 205–237.

DE FREITAS, N., HØJEN-SØRENSEN, P., JORDAN, M. I. and RUSSELL, S. (2001). Variational MCMC. In *Proc. 18th Conf. Uncertainty in Artificial Intelligence* (J. Breese and D. Koller, eds.) 120–127. Morgan Kauffman, San Mateo, CA.

DE FREITAS, J. F. G., NIRANJAN, M., GEE, A. H. and DOUCET, A. (2000). Sequential Monte Carlo methods to train neural network models. *Neural Computation* **12** 955–993.

DUANE, S., KENNEDY, A. D., PENDELTON, B. J. and ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B* **195** 216–222.

EDWARDS, P. J., MURRAY, A. F., PAPADOPOULOS, G., WALLACE, A. R., BARNARD, J. and SMITH, G. (1999). The application of neural networks to the papermaking industry. *IEEE Trans. Neural Networks* **10** 1456–1464.

GEIGER, D., HECKERMAN, D. and MEEK, C. (1999). Asymptotic model selection for directed networks with hidden variables. In *Learning in Graphical Models* (M. Jordan, ed.) 461–477. MIT Press.

GENCAY, R. and QI, M. (2001). Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping and bagging. *IEEE Trans. Neural Networks* **12** 726–734.

GHAHRAMANI, Z. and BEAL, M. (2000). Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) **12** 449–455. MIT Press.

GHAHRAMANI, Z. and BEAL, M. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich and V. Tresp, eds.) **13** 507–513. MIT Press.

GHAHRAMANI, Z. and JORDAN, M. I. (1997). Factorial hidden Markov models. *Machine Learning* **29** 245–273.

GIBBS, M. N. and MACKAY, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Trans. Neural Networks* **11** 1458–1464.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

HALL, P., HUMPHREYS, K. and TITTERINGTON, D. M. (2002). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 549–564.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining, Inference and Prediction.* Springer, New York.

HECKERMAN, D. (1999). A tutorial on learning with Bayesian networks. In *Learning in Graphical Models* (M. Jordan, ed.) 301–354. MIT Press.

HINTON, G. E. and VAN CAMP, D. (1993). Keeping neural networks simple by minimizing the description length of the weights. In *Proc. 6th ACM Conf. Computational Learning Theory* 5–13. ACM Press, New York.

HOLMES, C. C. and MALLICK, B. K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation* **10** 1217–1233.

HUMPHREYS, K. and TITTERINGTON, D. M. (2000a). Approximate Bayesian inference for simple mixtures. In *Proc. Computational Statistics 2000* (J. G. Bethlehem and P. G. M. van der Heijden, eds.) 331–336. Physica-Verlag, Heidelberg.

HUMPHREYS, K. and TITTERINGTON, D. M. (2000b). Improving the mean-field approximation in belief networks using Bahadur's reparameterisation of the multivariate binary distribution. *Neural Processing Lett.* **12** 183–197.

HUMPHREYS, K. and TITTERINGTON, D. M. (2001). Some examples of recursive variational approximations. In *Advanced Mean Field Methods*: *Theory and Practice* (M. Opper and D. Saad, eds.) 179–195. MIT Press.

HUSMEIER, D. (2000). The Bayesian evidence scheme for regularizing probability-density estimating neural networks. *Neural Computation* **12** 2685–2717.

HUSMEIER, D., PENNY, W. D. and ROBERTS, S. J. (1999). An empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers. *Neural Networks* **12** 677–705.

JAAKKOLA, T. (2001). Tutorial on variational approximation methods. In *Advanced Mean Field Methods*: *Theory and Practice* (M. Opper and D. Saad, eds.) 129–159. MIT Press.

JAAKKOLA, T. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statist. Comput.* **10** 25–37.

JACOBS, R. A., PENG, F. and TANNER, M. A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks* **10** 231–241.

JORDAN, M. I., ed. (1999). *Learning in Graphical Models*. MIT Press.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. In *Learning in Graphical Models* (M. Jordan, ed.) 105–162. MIT Press.

JORDAN, M. I. and JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6** 181–214.

KAY, J. W. and TITTERINGTON, D. M., eds. (1999). *Statistics and Neural Networks*: *Advances at the Interface*. Oxford Univ. Press.

KONISHI, S., ANDO, T. and IMOTO, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* **91** 27–43.

KWOK, J. T.-Y. (1999). Moderating the outputs of support vector machine classifiers. *IEEE Trans. Neural Networks* **10** 1018–1031.

KWOK, J. T.-Y. (2000). The evidence framework applied to support vector machines. *IEEE Trans. Neural Networks* **11** 1162–1173.

LAMPINEN, J. and VEHTARI, A. (2001). Bayesian approach for neural networks—review and case studies. *Neural Networks* **14** 257–274.

LEE, H. K. H. (2001). Model selection for neural network classification. *J. Classification* **18** 227–243.

LEE, H. K. H. (2003). A noninformative prior for neural networks. *Machine Learning* **50** 197–212.

LEISINK, M. A. R. and KAPPEN, H. J. (2001). A tighter bound for graphical models. *Neural Computation* **13** 2149–2171.

LUTTRELL, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation* **6** 767–794.

MACKAY, D. J. C. (1992a). Bayesian interpolation. *Neural Computation* **4** 415–447.

MACKAY, D. J. C. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Computation* **4** 448–472.

MACKAY, D. J. C. (1992c). The evidence framework applied to classification networks. *Neural Computation* **4** 720–736.

MACKAY, D. J. C. (1995). Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network*: *Computation in Neural Systems* **6** 469–505.

MACKAY, D. J. C. (1997). Ensemble learning for hidden Markov models. Technical report, Cavendish Lab., Univ. Cambridge.

MACKAY, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation* **11** 1035–1068.

MEDEIROS, M. C., VEIGA, A. and PEDREIRA, C. E. (2001). Modeling exchange rates: Smooth transitions, neural networks and linear models. *IEEE Trans. Neural Networks* **12** 755–764.

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 511–567.

MISKIN, J. W. and MACKAY, D. J. C. (2000). Ensemble learning for blind image separation and deconvolution. In *Advances in Independent Component Analysis* (M. Girolami, ed.) 123–141. Springer, New York.

MOODY, J. E. (1992). The *effective* number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) **4** 847–854. Morgan Kaufmann, San Mateo, CA.

MÜLLER, P. and RÍOS-INSUA, D. (1998). Issues in Bayesian analysis of neural network models. *Neural Computation* **10** 749–770.

NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer, New York.

NEAL, R. M. and HINTON, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (M. Jordan, ed.) 355–368. MIT Press.

OPPER, M. and SAAD, D., eds. (2001). *Advanced Mean Field Methods*: *Theory and Practice*. MIT Press.

PAIGE, R. L. and BUTLER, R. W. (2001). Bayesian inference in neural networks. *Biometrika* **88** 623–641.

PENG, F., JACOBS, R. A. and TANNER, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91** 953–960.

PENNY, W. D. and ROBERTS, S. J. (1999). Bayesian neural networks for classification: How useful is the evidence framework? *Neural Networks* **12** 877–892.

RIPLEY, B. D. (1994). Neural networks and related methods for classification (with discussion). *J. Roy. Statist. Soc. Ser. B* **56** 409–456.

RIPLEY, B. D. (1995). Choosing network complexity. In *Probabilistic Reasoning and Bayesian Belief Networks* (A. Gammerman, ed.) 97–108. Waller, Henley-on-Thames, UK.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press.

SATO, M. (2001). Online model selection based on the variational Bayes. *Neural Computation* **13** 1649–1681.

SEEGER, M. (2000). Bayesian model selection for support vector machines, Gaussian processes, and other kernel classifiers. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) **12** 603–609. MIT Press.

SOMMER, F. T. and DAYAN, P. (1998). Bayesian retrieval in associative memories with storage errors. *IEEE Trans. Neural Networks* **9** 705–713.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639.

TANAKA, K., INOUE, J. and TITTERINGTON, D. M. (2003). Probabilistic image processing by means of Bethe approximation for the Q-Ising model. *J. Phys. A* **36** 11,023–11,035.

THODBERG, H. H. (1996). A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Trans. Neural Networks* **7** 56–72.

TIPPING, M. E. (2000). The relevance vector machine. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen and K.-R. Müller, eds.) **12** 652–658. MIT Press.

TIPPING, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Res.* **1** 211–244.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

UEDA, N. and GHAHRAMANI, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks* **15** 1223–1241.

UTSUGI, A. (1997). Hyperparameter selection for self-organizing maps. *Neural Computation* **9** 623–635.

UTSUGI, A. (1998). Density estimation by mixture models with smoothing priors. *Neural Computation* **10** 2115–2135.

VAN GESTEL, T., SUYKENS, J. A. K., LANCKRIET, G., LAMBRECHTS, A., DE MOOR, B. and VANDEWALLE, J. (2002). Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Computation* **14** 1115–1147.

VILA, J.-P., WAGNER, V. and NEVEU, P. (2000). Bayesian nonlinear model selection and neural networks: A conjugate prior approach. *IEEE Trans. Neural Networks* **11** 265–278.

VIVARELLI, F. and WILLIAMS, C. K. I. (2001). Comparing Bayesian neural network algorithms for classifying segmented outdoor images. *Neural Networks* **14** 427–437.

WATERHOUSE, S., MACKAY, D. and ROBINSON, T. (1996). Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems* (M. C. Mozer, D. S. Touretzky and M. E. Hasselmo, eds.) **8** 351–357. MIT Press.

WILLIAMS, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation* **10** 1203–1216.

WRIGHT, W. A. (1999). Bayesian approach to neural-network modeling with input uncertainty. *IEEE Trans. Neural Networks* **10** 1261–1270.

YEDIDIA, J., FREEMAN, W. T. and WEISS, Y. (2001). Bethe free energy, Kikuchi approximations and belief propagation algorithms. Technical Report MERL TR 2001-16, Mitsubishi Electric Research Laboratories, Cambridge, MA.

ZHANG, B.-L., COGGINS, R., JABRI, M. A., DERSCH, D. and FLOWER, B. (2001). Multiresolution forecasting for futures trading using wavelet decompositions. *IEEE Trans. Neural Networks* **12** 765–775.

ZHANG, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Trans. Signal Process.* **40** 2570–2583.

ZHANG, J. (1993). The mean field theory in EM procedures for blind Markov random field image restoration. *IEEE Trans. Image Process.* **2** 27–40.