Pattern Recognition Letters 31 (2010) 1618–1626

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Expectation Propagation for microarray data classification

Daniel Hernández-Lobato [a,*], José Miguel Hernández-Lobato [b], Alberto Suárez [b]

[a] Department of Computing Science and Engineering, Université Catholique de Louvain, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium
[b] Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/ Francisco Tomás y Valiente, 11, Madrid 28049, Spain

ARTICLE INFO

Article history:
Received 5 June 2009
Available online 13 May 2010
Communicated by T. Vasilakos

Keywords:
Microarray data
Bayesian inference
Expectation Propagation

ABSTRACT

Microarray experiments are a very promising tool for early diagnosis and disease treatment. The datasets obtained in these experiments typically consist of a small number of instances and a large number of covariates, most of which are irrelevant for discrimination. These characteristics pose severe difficulties for standard learning algorithms. A Bayesian approach can be useful to overcome these problems and produce more accurate and robust predictions. However, exact Bayesian inference is computationally costly and in many cases infeasible. In practice, some form of approximation has to be made. In this paper we consider a Bayesian linear model for microarray data classification based on a prior distribution that favors sparsity in the model coefficients. Expectation Propagation (EP) is then used to perform approximate inference as an alternative to computationally more expensive methods, such as Markov Chain Monte Carlo (MCMC) sampling. The model considered is evaluated on 15 microarray datasets and compared with other state-of-the-art classification algorithms. These experiments show that the Bayesian model trained with EP performs well on the datasets investigated and is also useful to identify relevant genes for subsequent analysis.

1. Introduction

Microarray chips based on c-DNA hybridization generate large amounts of data by simultaneously measuring the expression level of several thousands of genes. However, the number of samples in which these biomedical experiments are performed is typically very small (Ramaswamy et al., 2003; Bourquin et al., 2006). These characteristics make the analysis of microarray data a challenging task that requires specialized statistical algorithms. A common application of microarray experiments is the diagnose of diseases on the basis of the gene expression level measured for an individual (Dudoit and Fridlyand, 2003). There is extensive empirical evidence that only a reduced number of genes are actually relevant for classification (Guyon et al., 2002; Tibshirani et al., 2002; Dudoit et al., 2002; Lee et al., 2005). Thus, most microarray classification algorithms use some form of feature selection to identify subsets of genes that are relevant for prediction. The selection is usually implemented by enforcing sparsity in the solutions. Assuming that the model prediction can be made in terms of a linear combination of normalized gene expression levels, some coefficients in the combination are driven to zero in the course of learning. The corresponding genes are eventually discarded in the classification process. The relevant non-zero coefficients can be determined by minimizing an estimate of the generalization error (Guyon et al., 2002; Tibshirani et al., 2002; Díaz-Uriarte and Alvarez de Andrés, 2006) or by maximum posterior estimation with a sparse regularization term (Li et al., 2002; Krishnapuram et al., 2004; Cawley and Talbot, 2006). Irrespective of the method used, the selection process can be unreliable because of the reduced amount of data available from each microarray experiment (Dougherty, 2001). In particular, Li et al. (2002) and Cawley and Talbot (2006) find that small modifications in the training dataset lead to substantial variations in the set of features that are selected. Bayesian models with prior distributions that encourage sparsity can be useful to overcome this problem, because they compute posterior probability distributions for the model coefficients rather than point estimates (MacKay, 2003; Bishop, 2006). Unlike approaches in which the posterior probability is maximized, these models do not directly generate sparse solutions. For a finite amount of data, the estimates of the model coefficients are uncertain and, in consequence, the posterior probability of a coefficient being exactly zero is generally zero. Nonetheless, Bayesian models with priors that encourage sparsity allow the separation of the model parameters into two groups: a first group where the coefficients are close to zero with high posterior probability, and a second one, in which some coefficients have a significant probability of being different from zero (Seeger, 2008). The genes in this last group are the ones that are relevant for prediction.

In this paper we consider a Bayesian model for microarray data classification based on the *spike and slab* sparse prior distribution (George and McCulloch, 1997). This prior introduces a binary latent

* Corresponding author. Tel.: +32 10 47 2415; fax: +32 10 45 0345.
E-mail addresses: daniel.hernandez-lobato@uclouvain.be (D. Hernández-Lobato), josemiguel.hernandez@uam.es (J.M. Hernández-Lobato), alberto.suarez@uam.es (A. Suárez).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.05.007

variable for each gene that indicates whether its expression level should be used for classification. Bayes' theorem is then employed to compute an estimate of the probability that each of these variables is active, based on the evidence given by the available data. This estimate discriminates among different subsets of genes and hence, can be useful for identifying relevant genes. In the model considered exact inference is infeasible and approximate techniques need to be used. However, sparse prior distributions often lead to complicated posteriors that can be extremely multi-modal (Seeger, 2008). This, combined with the large dimensionality of the data obtained in microarray experiments, makes approximate inference a very difficult task. MCMC sampling methods are standard techniques that can be used to address this problem (Lee et al., 2003; Zhou et al., 2004; Bae and Mallick, 2004). These are based on sampling from a Markov chain whose stationary distribution is the posterior distribution of the model parameters. These samples are then used to compute probability estimates or to approximate the predictive distribution of the model for new data. A drawback of this approach is that obtaining accurate estimates of the posterior distribution generally requires the simulation of very long Markov chains.

As an alternative of MCMC sampling methods, we propose to use Expectation Propagation (EP) (Minka, 2001b), which is a fast approximate inference algorithm. EP approximates the posterior distribution of the model parameters by the product of factors that belong to the exponential family of probability distributions. These factors are iteratively updated until they converge to a fixed point. The update rules are derived from moment matching constraints. A drawback of EP is that poor approximations can be obtained when the posterior distribution is multi-modal (Bishop, 2006). In spite of this possible limitation, the performance of the Bayesian model trained with EP is comparable with other microarray classification techniques in the datasets analyzed. Furthermore, genes whose coefficients have a high a probability of being different from zero according to the approximation obtained for the posterior are good candidates for subsequent analysis.

2. Approach

The goal of this investigation is to learn a decision function that discriminates on the basis of gene expression measurements between tissues belonging to different classes (e.g. tumor and normal samples). For this purpose, a set of n d-dimensional input examples $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and the corresponding target class labels $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{-1, 1\}$ are available. Typically, in microarray measurements the number of observations n is small and the number of attributes d is large. The model considered assumes that there is a monotonic relation between the probability of the target value y_i and the value of a linear combination of the gene expression measurements $\boldsymbol{x}_i$. Assuming that the gene expression measurements $\boldsymbol{X}$ are contaminated by Gaussian noise, the classification rule is

$$y_i = \begin{cases} 1 & \text{if } \boldsymbol{w}^t \boldsymbol{x}_i + \epsilon_i \geqslant 0, \\ -1 & \text{otherwise}, \end{cases} \tag{1}$$

where ϵ_i follows a standard Gaussian distribution and $\boldsymbol{x}_i$ includes, besides the gene expression levels, a constant bias component. An alternative is to assume that the noise term follows the logistic distribution (Krishnapuram et al., 2004), which is less sensitive to outliers. This functional form for the noise distribution has not been considered in this study because it makes Bayesian inference more complicated (Bishop, 2006). For this model the likelihood function for a vector of weights $\boldsymbol{w}$ given $\boldsymbol{y}$ and $\boldsymbol{X}$ is

$$\mathcal{P}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X}) = \prod_{i=1}^{n} \mathcal{P}(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} \Phi(y_i \boldsymbol{w}^t \boldsymbol{x}_i), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative probability function of the standard Gaussian distribution. To simplify the notation, in the rest of the article, $\boldsymbol{x}_i$ is used to denote the vector of attributes scaled by the corresponding class label y_i. Furthermore, since $\boldsymbol{X}$ always appears as a conditioning variable, it is assumed implicit in the expressions of the probabilities.

To complete the Bayesian description we select a prior distribution for $\boldsymbol{w}$. Because in microarray datasets only a small subset of the genes are actually relevant for discrimination, it seems natural to make use of a prior distribution that encourages sparsity. In particular, we employ the *spike and slab* prior introduced by George and McCulloch (1997) and later used by Lee et al. (2003) to make Bayesian inference in microarray classification problems. In this prior all components of the vector $\boldsymbol{w}$ are assumed to be independent. Binary latent variables γ_i are introduced to indicate whether the expression level of the ith gene is used for classification ($\gamma_i = 1$) or not ($\gamma_i = 0$). Given $\boldsymbol{\gamma}$, the prior for $\boldsymbol{w}$ is

$$\mathcal{P}(\boldsymbol{w}|\boldsymbol{\gamma}) = \prod_{i=1}^{d} \mathcal{N}(w_i|0, \sigma_1^2)^{\gamma_i} \mathcal{N}(w_i|0, \sigma_0^2)^{1-\gamma_i}, \tag{3}$$

where $\mathcal{N}(w_i|0, \sigma_1^2)$ denotes a Gaussian density with zero mean and σ_1^2 variance evaluated at w_i. To enforce sparsity, the standard deviation of the *slab* σ_1 is set to 1 and the distribution of the *spike* is assumed to be a delta function centered at the origin ($\sigma_0 \to 0^+$). Finally, independent *Bernoulli* priors are assumed for the components of $\boldsymbol{\gamma}$

$$\mathcal{P}(\boldsymbol{\gamma}) = \prod_{i=1}^{d} \rho^{\gamma_i} (1-\rho)^{1-\gamma_i}. \tag{4}$$

This functional form for the prior assumes that the gene expression levels have probability ρ of being included in the classification model.

Besides the *spike and slab* prior, other prior distributions that could have been considered to encourage sparsity in this model are the Laplace prior (Williams, 1995; Tibshirani, 1996; Roth, 2002), the Student-t prior (Tipping, 2001), and in general, any distribution of the form $\propto \exp(-|\cdot|^{\alpha})$, with $\alpha < 1$ (Seeger, 2008). An advantage of the *spike and slab* prior over these alternatives is that the degree of sparsity can be directly specified by ρ. This parameter is a probability value that determines the fraction of latent variables that are expected to be active before any data has been observed. In the Laplace and the Student-t distributions the degree of sparsity depends on a parameter that has data-dependent scale. Therefore, to achieve a specified degree of sparsity the value of the parameter needs to be tuned for each problem separately. Another advantage of the *spike and slab* prior distribution is that it performs a selective shrinkage of the model coefficients (Ishwaran and Rao, 2005). That is, when the degree of sparsity is increased, the estimates corresponding to coefficients that should be zero become smaller, while the absolute values of coefficients that should be different from zero remain large. By contrast, priors such as the Laplace distribution perform a continuous shrinkage towards zero of all the model coefficients, including those coefficients that should actually be different from zero (Tibshirani, 1996; Zou and Hastie, 2005).

In the model considered the joint posterior distribution for $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ is computed using Bayes' theorem

$$\mathcal{P}(\boldsymbol{\gamma}, \boldsymbol{w}|\boldsymbol{y}) = \frac{\mathcal{P}(\boldsymbol{y}|\boldsymbol{w}) \mathcal{P}(\boldsymbol{w}|\boldsymbol{\gamma}) \mathcal{P}(\boldsymbol{\gamma})}{\mathcal{P}(\boldsymbol{y})}, \tag{5}$$

where $\mathcal{P}(\boldsymbol{y})$ is a normalization constant that can be used to perform model selection (MacKay, 2003; Bishop, 2006).

A test instance $\boldsymbol{x}^{\text{test}}$ is classified using the predictive distribution for its target class $y^{\text{test}} \in \{1, -1\}$, namely

$$\mathcal{P}(y^{\text{test}}|\boldsymbol{x}^{\text{test}},\boldsymbol{y}) = \int \mathcal{P}(y^{\text{test}}|\boldsymbol{x}^{\text{test}},\boldsymbol{w})\mathcal{P}(\boldsymbol{\gamma},\boldsymbol{w}|\boldsymbol{y})d\boldsymbol{\gamma}d\boldsymbol{w}. \quad (6)$$

This probability value quantifies the uncertainty in the prediction.

Finally, the genes that contribute the most to the classification process can be identified using the posterior distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{y}$

$$\mathcal{P}(\boldsymbol{\gamma}|\boldsymbol{y}) = \int \mathcal{P}(\boldsymbol{\gamma},\boldsymbol{w}|\boldsymbol{y})d\boldsymbol{w}. \quad (7)$$

Unfortunately, the exact evaluation of (5)–(7) is too costly to be practicable and one has to resort to approximation techniques. An approach based MCMC sampling was proposed in (Lee et al., 2003). However, this algorithm requires long simulations of the Markov chain. Inspired by the success of EP (Minka, 2001b) in a related problem, which involves the analysis of time-dependent gene expression data (Hernández-Lobato et al., 2008), we propose to apply this efficient algorithm to perform approximate inference.

3. Expectation Propagation

This section presents a review of the EP algorithm in its general form. To perform inference, EP approximates the posterior distribution of the model parameters by a distribution that facilitates the subsequent computations (Minka, 2001b). Given a set of i.i.d. input variables $\boldsymbol{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ with corresponding class labels $\boldsymbol{y} = \{y_1,\ldots,y_n\}$, the joint distribution of the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{y}$ can be expressed as a product of terms

$$\mathcal{P}(\boldsymbol{\theta},\boldsymbol{y}) = \prod_{i=1}^{n}\mathcal{P}(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})\mathcal{P}(\boldsymbol{\theta}) = \prod_{i=1}^{n+1} t_i(\boldsymbol{\theta}), \quad (8)$$

where $t_{n+1}(\boldsymbol{\theta}) = \mathcal{P}(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$ and $t_i(\boldsymbol{\theta}) = \mathcal{P}(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})$ for $i = 1,\ldots,n$. EP approximates (8) by a product of simpler terms $\tilde{t}_i$

$$\prod_{i=1}^{n+1} t_i(\boldsymbol{\theta}) \approx \prod_{i=1}^{n+1}\tilde{t}_i(\boldsymbol{\theta}) = \mathcal{Q}(\boldsymbol{\theta}), \quad (9)$$

where the approximate terms $\tilde{t}_i$ are restricted to belong to the same family $\mathcal{F}$ of exponential distributions, except that they need not be normalized. Because of the closure property of exponential distributions under the product operation, $\mathcal{Q}$ also belongs to $\mathcal{F}$ and can be readily normalized by integrating over $\boldsymbol{\theta}$. In each iteration EP updates the term $\tilde{t}_i$ so that

$$\mathcal{Q}(\boldsymbol{\theta}) = \tilde{t}_i(\boldsymbol{\theta})\prod_{j\neq i}\tilde{t}_j(\boldsymbol{\theta}) = \tilde{t}_i(\boldsymbol{\theta})\mathcal{Q}^{\backslash i}(\boldsymbol{\theta}) \quad (10)$$

is as close as possible to

$$t_i(\boldsymbol{\theta})\prod_{j\neq i}\tilde{t}_j(\boldsymbol{\theta}) = t_i(\boldsymbol{\theta})\mathcal{Q}^{\backslash i}(\boldsymbol{\theta}) \quad (11)$$

in terms of the Kullback–Leibler (KL) divergence.

The EP algorithm involves the following steps:

1. Initialize the terms $\tilde{t}_i$ and $\mathcal{Q}$ to be uniform.
2. Repeat until all $\tilde{t}_i$ converge:
   (a) Select a $\tilde{t}_i$ to refine and compute $\mathcal{Q}^{\backslash i}$ dividing $\mathcal{Q}$ by $\tilde{t}_i$.
   (b) Update the term $\tilde{t}_i$, so that the KL divergence between $t_i\mathcal{Q}^{\backslash i}$ and $\tilde{t}_i\mathcal{Q}^{\backslash i}$ is minimized.
   (c) Re-compute $\mathcal{Q}$ as the product of the new $\tilde{t}_i$ and $\mathcal{Q}^{\backslash i}$.

The optimization problem of step (b) is convex and has a single optimum (Bishop, 2006; MacKay, 2003). Because both $\tilde{t}_i$ and $\mathcal{Q}$ belong to the same exponential family $\mathcal{F}$, this optimum is found by matching the sufficient statistics between $\tilde{t}_i\mathcal{Q}^{\backslash i}$ and $t_i\mathcal{Q}^{\backslash i}$. Once $\mathcal{Q}$ is normalized, it can be used as an estimate of $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{y})$ and its normalization constant can be used to approximate $\mathcal{P}(\boldsymbol{y})$. Even though the convergence of the EP algorithm is not guaranteed, empirically it is seen to converge except when the approximating family is a poor model for the exact posterior (Minka, 2001a). EP has shown a good overall performance when compared to other approximate inference techniques like Monte Carlo methods and variational inference (Minka, 2001b). Nevertheless, because of the moment matching procedure of the algorithm, EP can lead to poor approximations when the posterior is multi-modal (Bishop, 2006).

4. EP for the *spike and slab* model

In this section we describe the application of the EP algorithm to the Bayesian model for microarray data analysis introduced in Section 2. First, the posterior in (5) is approximated by a factorized exponential distribution

$$\mathcal{P}(\boldsymbol{\gamma},\boldsymbol{w}|\boldsymbol{y}) \approx \prod_{j=1}^{d} p_j^{\gamma_j}(1-p_j)^{1-\gamma_j}\mathcal{N}(w_j|\mu_j,v_j), \quad (12)$$

where the vectors $\boldsymbol{p}$, $\boldsymbol{\mu}$ and $\boldsymbol{v}$ are free parameters. In this approximation, the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ and their components are assumed to be independent, which makes the approximate inference process faster. Approximations that assume correlations between $\boldsymbol{\gamma}$ and $\boldsymbol{w}$ and their components can be more accurate, but also increase the computational cost of EP. The numerator of the right-hand-side of (5) is factorized as the product of $n + d + 1$ terms t_i; namely n terms for the likelihood, (2), d terms for the prior for $\boldsymbol{w}$, (3), and one term for the prior for $\boldsymbol{\gamma}$ (4). Thus, the EP approximation is computed as the product of $n + d + 1$ terms $\tilde{t}_i$, all of which are of the form given by (12), except for normalization

$$\tilde{t}_i(\boldsymbol{\gamma},\boldsymbol{w}) = s_i\prod_{j=1}^{d} a_{ij}^{\gamma_j}b_{ij}^{1-\gamma_j}\exp\left(-\frac{1}{2v_{ij}}(w_j-m_{ij})^2\right), \quad (13)$$

where s_i is a constant that ensures that the integral of $\tilde{t}_i\mathcal{Q}^{\backslash i}$ is the same as the integral of $t_i\mathcal{Q}^{\backslash i}$ and the vectors $\boldsymbol{a}_i$, $\boldsymbol{b}_i$, $\boldsymbol{m}_i$ and $\boldsymbol{v}_i$ are free parameters.

The minimization of the KL divergence between $\tilde{t}_i\mathcal{Q}^{\backslash i}$ and $t_i\mathcal{Q}^{\backslash i}$ leads to the following constraints for the expected values of the model parameters

$$\mathbb{E}_{\tilde{t}_i\mathcal{Q}^{\backslash i}}[\boldsymbol{w}] = \mathbb{E}_{t_i\mathcal{Q}^{\backslash i}}[\boldsymbol{w}], \quad (14)$$

$$\mathbb{E}_{\tilde{t}_i\mathcal{Q}^{\backslash i}}[\boldsymbol{w}\circ\boldsymbol{w}] = \mathbb{E}_{t_i\mathcal{Q}^{\backslash i}}[\boldsymbol{w}\circ\boldsymbol{w}], \quad (15)$$

$$\mathbb{E}_{\tilde{t}_i\mathcal{Q}^{\backslash i}}[\boldsymbol{\gamma}] = \mathbb{E}_{t_i\mathcal{Q}^{\backslash i}}[\boldsymbol{\gamma}], \quad (16)$$

where the operator $\circ$ indicates the Hadamard (element-wise) product.

Matching the expected values (14) and (15) for each likelihood term t_i, $i = 1,2,\ldots,n$ gives the following update rules for (12) and the corresponding approximate term $\tilde{t}_i$

$$\boldsymbol{\mu}^{\text{new}} = \boldsymbol{\mu}^{\text{old}} + \alpha_i\boldsymbol{v}^{\text{old}}\circ\boldsymbol{x}_i, \quad (17)$$

$$\boldsymbol{v}^{\text{new}} = \boldsymbol{v}^{\text{old}} - \frac{\alpha_i(\boldsymbol{x}_i^t\boldsymbol{\mu}^{\text{new}}+\alpha_i)}{\boldsymbol{x}_i^t(\boldsymbol{v}^{\text{old}}\circ\boldsymbol{x}_i)+1}(\boldsymbol{v}^{\text{old}}\circ\boldsymbol{x}_i)\circ(\boldsymbol{v}^{\text{old}}\circ\boldsymbol{x}_i), \quad (18)$$

$$\boldsymbol{v}_i^{\text{new}} = \left((\boldsymbol{v}^{\text{new}})^{-1} - (\boldsymbol{v}^{\text{old}})^{-1}\right)^{-1}, \quad (19)$$

$$\boldsymbol{m}_i^{\text{new}} = \boldsymbol{\mu}^{\text{old}} + \alpha_i\boldsymbol{v}_i^{\text{new}}\circ\boldsymbol{x}_i + \alpha_i\boldsymbol{v}^{\text{old}}\circ\boldsymbol{x}_i, \quad (20)$$

$$s_i = \Phi(z)\prod_{j=1}^{d}\sqrt{\frac{v_{ij}^{\text{new}}+v_j^{\text{old}}}{v_{ij}^{\text{new}}}}\exp\left(\sum_{j=1}^{d}\frac{(m_{ij}^{\text{new}}-\mu_j^{\text{old}})^2}{2(v_j^{\text{old}}+v_{ij}^{\text{new}})}\right), \quad (21)$$

where

$$\boldsymbol{v}^{\text{old}} = (\boldsymbol{v}^{-1} - \boldsymbol{v}_i^{-1})^{-1}, \tag{22}$$

$$\boldsymbol{\mu}^{\text{old}} = \boldsymbol{\mu} + \boldsymbol{v}^{\text{old}} \circ \boldsymbol{v}_i^{-1} \circ (\boldsymbol{\mu} - \boldsymbol{m}_i), \tag{23}$$

$$\alpha_i = \frac{1}{\sqrt{\boldsymbol{x}_i^t(\boldsymbol{v}^{\text{old}} \circ \boldsymbol{x}_i) + 1}} \frac{\mathcal{N}(z|0,1)}{\Phi(z)}, \tag{24}$$

$$z = \frac{\boldsymbol{x}_i^t \boldsymbol{\mu}^{\text{old}}}{\sqrt{\boldsymbol{x}_i^t(\boldsymbol{v}^{\text{old}} \circ \boldsymbol{x}_i) + 1}}, \tag{25}$$

and the inverse of a vector is defined as a new vector whose components are the inverse of the components of the original vector (for further details see (Minka, 2001b)). Note that the likelihood terms $\{t_i : i = 1,2,\ldots,n\}$ are independent of $\boldsymbol{\gamma}$. Therefore, the vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ corresponding to $\{\tilde{t}_i : i = 1,2,\ldots,n\}$ are left unchanged. These update rules might fail because of a negative value in some component of $\boldsymbol{v}^{\text{old}}$. To deal with this situation, whenever a component of $\boldsymbol{v}^{\text{old}}$ becomes negative, the corresponding likelihood term is ignored until the next iteration of the algorithm (Minka, 2001b).

Matching the expected values (14)–(16) for each term of the prior t_{n+i}, $i = 1,2,\ldots,d$ gives the following update rules for (12) and the corresponding approximate term $\tilde{t}_{n+i}$

$$\mu_i^{\text{new}} = \mu_i^{\text{old}} + c_1 v_i^{\text{old}}, \tag{26}$$

$$v_i^{\text{new}} = v_i^{\text{old}} - c_3 (v_i^{\text{old}})^2, \tag{27}$$

$$p_i^{\text{new}} = \frac{p_i^{\text{old}} \mathcal{G}_1}{p_i^{\text{old}} \mathcal{G}_1 + (1 - p_i^{\text{old}}) \mathcal{G}_0}, \tag{28}$$

$$v_{(n+i)i}^{\text{new}} = c_3^{-1} - v_i^{\text{old}}, \tag{29}$$

$$m_{(n+i)i}^{\text{new}} = \mu_i^{\text{old}} + c_1 \left(v_{(n+i)i}^{\text{new}} + v_i^{\text{old}} \right), \tag{30}$$

$$a_{(n+i)i}^{\text{new}} = p_i^{\text{new}} / p_i^{\text{old}}, \tag{31}$$

$$b_{(n+i)i}^{\text{new}} = (1 - p_i^{\text{new}}) / (1 - p_i^{\text{old}}), \tag{32}$$

$$s_i = Z \sqrt{\frac{v_i^{\text{old}} + v_{ii}^{\text{new}}}{v_{ii}^{\text{new}}}} \exp\left(\frac{1}{2} \frac{c_1^2}{c_3} \right), \tag{33}$$

where

$$v_i^{\text{old}} = (v_i^{-1} - v_{(n+i)i}^{-1})^{-1}, \tag{34}$$

$$\mu_i^{\text{old}} = \mu_i + v_i^{\text{old}} v_{(n+i)i}^{-1} (\mu_i - m_{(n+i)i}), \tag{35}$$

$$Z = p_i^{\text{old}} \mathcal{G}_1 + (1 - p_i^{\text{old}}) \mathcal{G}_0, \tag{36}$$

$$c_1 = Z^{-1} \left(p_i^{\text{old}} \mathcal{G}_1 \frac{-\mu_i^{\text{old}}}{v_i^{\text{old}} + \sigma_1^2} + (1 - p_i^{\text{old}}) \mathcal{G}_0 \frac{-\mu_i^{\text{old}}}{v_i^{\text{old}} + \sigma_0^2} \right), \tag{37}$$

$$c_2 = Z^{-1} \frac{1}{2} \left(p_i^{\text{old}} \mathcal{G}_1 \left(\frac{\mu_i^{\text{old}^2}}{(v_i^{\text{old}} + \sigma_1^2)^2} - \frac{1}{v_i^{\text{old}} + \sigma_1^2} \right) + (1 - p_i^{\text{old}}) \mathcal{G}_0 \left(\frac{\mu_i^{\text{old}^2}}{(v_i^{\text{old}} + \sigma_0^2)^2} - \frac{1}{v_i^{\text{old}} + \sigma_0^2} \right) \right), \tag{38}$$

$$c_3 = c_1^2 - 2c_2, \tag{39}$$

$$\mathcal{G}_0 = \mathcal{N}(0|\mu_i^{\text{old}}, v_i^{\text{old}} + \sigma_0^2), \tag{40}$$

$$\mathcal{G}_1 = \mathcal{N}(0|\mu_i^{\text{old}}, v_i^{\text{old}} + \sigma_1^2), \tag{41}$$

$$p_i^{\text{old}} = \frac{p_i / a_{(n+i)i}}{p_i / a_{(n+i)i} + (1 - p_i) / b_{(n+i)i}}. \tag{42}$$

When processing each of these terms, only one component of the vectors $\boldsymbol{p}$, $\boldsymbol{\mu}$, $\boldsymbol{v}$, $\boldsymbol{a}_{n+i}$, $\boldsymbol{b}_{n+i}$, $\boldsymbol{m}_{n+i}$ and $\boldsymbol{v}_{n+i}$ is updated. Thus, the components $j \neq i$ of the vectors $\boldsymbol{a}_{n+i}$, $\boldsymbol{b}_{n+i}$, $\boldsymbol{m}_{n+i}$ and $\boldsymbol{v}_{n+i}$ are not changed. Recall that all approximate terms $\tilde{t}_i$ are initialized to be uniform at the beginning of EP (i.e. $a_{ij} = b_{ij} = 1$, $m_{ij} = 0$, $v_{ij} = \infty$, $\forall i,j$).

If p_i is initialized to ρ $\forall i$, it is not necessary to process the term t_{n+d+1}, which corresponds to the prior for $\boldsymbol{\gamma}$ (Minka, 2001b). The update of the terms $\{\tilde{t}_i : i = 1,2,\ldots,n\}$ corresponding to the likelihood takes $\mathcal{O}(nd)$ steps. However, the terms $\{\tilde{t}_{n+i} : i = 1,2,\ldots,d\}$ corresponding to the prior can be updated in only $\mathcal{O}(d)$ steps because of the factorized form of the approximation. Hence, the total cost of one iteration of EP is $\mathcal{O}(nd)$. By contrast, if efficient matrix factorizations are employed (George and McCulloch, 1997), the cost of the MCMC sampling algorithm proposed by Lee et al. (2003) is on average $\mathcal{O}(\rho^2 d^3 k)$, where k is the number of samples drawn from the posterior distribution, usually several thousands and hence, of the same order as d.

Once the EP algorithm has converged, the normalization constant of $\mathcal{Q}$ is

$$\mathcal{P}(\boldsymbol{y}) \approx C (2\pi)^{\frac{d}{2}} \prod_{j=1}^{d} \sqrt{v_j} \exp(B/2) \prod_{i=1}^{n+d} s_i, \tag{43}$$

where

$$B = \boldsymbol{\mu}^t (\boldsymbol{v}^{-1} \circ \boldsymbol{\mu}) - \sum_{i=1}^{n} \boldsymbol{m}_i^t (\boldsymbol{v}_i^{-1} \circ \boldsymbol{m}_i) - \sum_{i=1}^{d} \frac{m_{(n+i)i}^2}{v_{(n+i)i}}, \tag{44}$$

$$C = \prod_{i=1}^{d} \left(a_{(n+i)i} \rho + b_{(n+i)i} (1 - \rho) \right). \tag{45}$$

To derive these equations, we have used the fact that several components of the vectors $\boldsymbol{m}_i$, $\boldsymbol{v}_i$, $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ remain constant during EP and, in consequence, do not contribute to (43).

Finally, the optimal classification rule for an arbitrary instance $\boldsymbol{x}^{\text{test}}$ can be approximated in terms of (12) as

$$\mathcal{P}(y^{\text{test}}|\boldsymbol{x}^{\text{test}}, \boldsymbol{y}) \approx \int \mathcal{P}(y^{\text{test}}|\boldsymbol{x}^{\text{test}}, \boldsymbol{w}) \prod_{j=1}^{d} \mathcal{N}(w_j|\mu_j, v_j) d\boldsymbol{w}, = \Phi \left(\frac{(\boldsymbol{x}^{\text{test}})^t \boldsymbol{\mu}}{\sqrt{(\boldsymbol{x}^{\text{test}})^t (\boldsymbol{v} \circ \boldsymbol{x}^{\text{test}}) + 1}} \right), \tag{46}$$

and to identify the genes that contribute the most to the classification process, the EP approximation of (7) can be used

$$\mathcal{P}(\boldsymbol{\gamma}|\boldsymbol{y}) \approx \prod_{j=1}^{d} p_j^{\gamma_i} (1 - p_j)^{1 - \gamma_j}. \tag{47}$$

5. Experiments

The performance of the Bayesian model trained with EP is assessed on 15 publicly available microarray classification problems. The characteristics and the sources of the datasets are shown in Table 1. Non-binary classification problems are binarized as follows: in *Lymphoma* we discriminate between 42 samples of diffuse large B-cell lymphoma and 20 samples of follicular lymphoma and

Table 1
Main characteristics of the microarray datasets used.

Dataset	Genes	Patients	Original paper
Adenocarcinoma	9868	76	Ramaswamy et al. (2003)
Brain A	5597	20	Pomeroy et al. (2002)
Brain B	2275	34	Pomeroy et al. (2002)
Brain C	4452	60	Pomeroy et al. (2002)
Breast ER	5313	49	West et al. (2001)
Breast LN	5313	49	West et al. (2001)
Colon	2000	62	Alon et al. (1999)
Down Syndrome	4656	63	Bourquin et al. (2006)
Leukemia	3571	72	Golub et al. (1999)
Lymphoma	4026	62	Alizadeh et al. (2000)
Metastasis	4869	77	van 't Veer et al. (2002)
Mutation	3226	22	Hedenfalk et al. (2001)
Ovarian	1536	54	Schummer et al. (1999)
Prostate	6033	102	Singh et al. (2002)
SRBCT	2308	43	Khan et al. (2001)

chronic lymphocytic leukemia; in *SRBCT* we discriminate between 23 Ewing family of tumors and 20 rhabdomyosarcomas; in *Metastasis* we discriminate between patients that developed metastasis within 5 years and those who did not; finally, in *Brain A* we discriminate between 10 malignant gliomas and 10 atypical teratoid/rhabdoid tumors. The datasets are preprocessed as follows: *Adenocarcinoma* and *Metastasis* are preprocessed as in (Díaz-Uriarte and Alvarez de Andrés, 2006); *Colon* is preprocessed as in (Guyon et al., 2002); *Breast ER*, *Breast LN*, *Brain A*, *Leukemia*, *Lymphoma*, *Prostate* and *SRBCT* are preprocessed as in (Dettling and Bühlmann, 2002); *Mutation* is log-transformed and normalized across genes and samples; *Brain B*, *Brain C* and *Down Syndrome* are preprocessed following the protocol described in the supplementary material of (Pomeroy et al., 2002). In particular, these datasets are thresholded, filtered by variation, log-transformed and normalized across genes and samples. The floor of the threshold step is set to 20 for *Brain B* and *Down Syndrome* and to 100 for *Brain C*. The ceiling is set to 16,000 in these three datasets. Filtering by variation is performed by removing genes with $\max/\min \leqslant 5$ or $(\max - \min) \leqslant 500$. The *Ovarian* dataset does not require any further preprocessing (Li et al., 2002).

The generalization performance of the Bayesian model trained with EP is compared with several classification systems: (a) a Bayesian model trained with EP that uses a standard Gaussian prior (i.e. $\rho = 1$). This model is included to determine whether the use of a *spike and slab* prior, which favors sparse models, is beneficial; (b) the Bayesian model trained with MCMC sampling (Lee et al., 2003). This method has been included to determine whether the EP algorithm, which can fail when the posterior distribution is multi-modal (Bishop, 2006), is sufficiently accurate; (c) three methods that have shown a good performance in several reviews of classification methods applied to microarray data (Dudoit et al., 2002; Dettling, 2004; Lee et al., 2005) but which do not include any gene selection procedure: the Support Vector Machine (SVM) with a linear kernel (Vapnik, 1995), the k-nearest neighbor classifier (KNN) (Hastie et al., 2001) and the diagonal linear discriminant analysis (DLDA) (Dudoit et al., 2002); (d) five methods that involve variable selection: the recursive feature elimination (RFE) algorithm introduced in (Guyon et al., 2002), the nearest shrunken centroids (NSC) method analyzed in (Tibshirani et al., 2002), the random forest (RF) classifier investigated in (Díaz-Uriarte and Alvarez de Andrés, 2006), the relevance vector machine (RVM) (Li et al., 2002) and the generalized LASSO (least absolute selection and shrinkage operator) (Roth, 2002).

The first three methods that incorporate a feature selection mechanism (RFE, NSC and RF) employ an estimate of the generalization error to determine which genes are relevant for classification. In the case of the RFE algorithm, a sequence of linear SVMs are trained on the data using a set of explanatory variables. This set initially contains all the genes whose expression levels are measured in the experiment. At each iteration, this set is reduced by eliminating the explanatory variable with the lowest estimated relevance. The absolute value of the coefficient w_j of the jth gene, $j = 1,2,\ldots,d$, in the resulting hyper-plane is used an estimate of the sensitivity of the SVM to the expression level of that gene. Therefore, at each iteration, the gene j^* whose associated coefficient w_{j^*} has the smallest absolute value is removed from the set of explanatory variables. The number of explanatory genes is determined by stopping the iterative process at the minimum of a cross-validation estimate of the generalization error. Finally, the SVM that corresponds to this minimum is used for prediction. The NSC method computes the centroid associated to each class label. The different components of these centroids are then shrunk towards zero. The amount of shrinkage is determined by minimizing the cross-validation error. Given an unlabeled test instance, the associated class label is computed in terms of the nearest shrunken centroid. The RF classifier builds an initial ensemble of random-trees generated on different bootstrap samples of the training data (Breiman, 2001). Then, iteratively, the expression levels of each gene are randomly permuted across the different data samples. The relative importance of each gene is determined by the increase in an estimate of the generalization error with respect to the original configuration, in which the expression levels are left unchanged. Out-of-bag data are used to compute this estimate (Breiman, 1996). At this point, a fixed fraction of the explanatory genes are removed sequentially according to their relative importance and a new ensemble of random-trees is built on the new data. This process is repeated until a critical number of explanatory genes remain. This number is determined by minimizing an estimate of the generalization error computed with the out-of-bag data. The final predictor is the resulting ensemble.

The last two techniques, the RVM and the generalized LASSO, can be viewed as Bayesian models that use prior distributions that favor sparsity to perform feature selection. Specifically, the RVM is a Bayesian linear model that assumes a factorized Gaussian prior distribution for the model coefficients (Tipping, 2001). The hyper-parameters of this prior α_j ($j = 1,\ldots,d$) correspond to the inverse of the variance (precision parameter) for each of the coefficients. Maximizing the marginal likelihood makes some of these hyper-parameters tend to infinity, which forces that the associated coefficients of the decision hyper-plane go to zero. This approach can be shown to be equivalent to maximizing the posterior probability (MAP), when a non-factorized sparsifying prior distribution is assumed (Wipf and Nagarajan, 2008). The generalized LASSO is a linear model that minimizes a logistic regression loss with a penalty term that is proportional to the sum of the absolute values of the model coefficients (Roth, 2002). This type of regularization favors sparsity; as a result of the minimization, some of the model coefficients are driven to zero. The generalized LASSO can be shown to be equivalent to MAP estimation in a Bayesian model that assumes a factorized Laplace prior for the model coefficients with a common scale parameter (Tibshirani, 1996). The common scale parameter is typically estimated by cross-validation.

To evaluate the methods described, the data are randomly partitioned into two disjoint sets for training and testing, respectively. The test sets need to be sufficiently large, so that there is sufficient empirical evidence to discriminate among the performance of different classification systems. Following (Dudoit et al., 2002), two thirds of the instances available are used for training and the remaining one third for testing. To reduce the variability of the results, this splitting process is repeated 50 times and the test error estimates are averaged over these different train-test partitions of the data. The gene expression levels are normalized so that they have zero mean and unit standard deviation on the training set. The hyper-parameters of the different methods are obtained by cross-validation using only the training set, which avoids any selection bias (Ambroise and McLachlan, 2002). In the RFE algorithm, half of the variables are removed at each step until less than 500 variables remain, similarly as in (Rakotomamonjy, 2003). Then, variables are removed one at a time. The C parameter of the SVM is set to 100 as suggested by Guyon et al. (2002). In the k-NN classifier the Euclidean distance is used. The parameter k is selected from the range of odd values $k = 1,3,\ldots,19$. In the RF approach the default settings s.e. = 1, mtryFactor = 1 and ntree = 5000 are used. The out-of-bag error estimate is employed for variable elimination as in (Díaz-Uriarte and Alvarez de Andrés, 2006). In the RVM algorithm we use the regression likelihood function and fix $\sigma = 1$ as suggested by Li et al. (2002). In the EP algorithm, the parameter ρ of the prior for γ is fixed so that, on average, 32 components of the vector are set to one, similarly as in (Lee et al., 2003). The hyper-parameters σ_0^2 and σ_1^2 of the *spike and slab* prior are set to zero and one respectively. The same values for the priors

and the parameters are used in the MCMC method. The Markov chain is implemented with Gibbs sampling (Lee et al., 2003). A fast updating algorithm based on efficient matrix factorizations (Gill et al., 1974) is used as suggested by George and McCulloch (1997). The posterior distribution is approximated using 5000 samples after a burn-in period of 1000 samples. Exploratory experiments indicate that this number of samples seems to be sufficient and that only marginal improvements are obtained in longer runs. In any case, the number of samples that can be generated in practice is limited because of the high computational costs. In the NSC method we use the default settings and employ a non-balanced cross-validation procedure to find the optimal threshold values. In the NSC implementation of (Tibshirani et al., 2002) this cross-validation procedure is assumed to be balanced by default (i.e. the test sets contain equal number of samples from each class). Finally, to optimize the objective function in the generalized LASSO we use the R package *glmnet* (Friedman et al., 2009). The regularization parameter of this model is selected by minimizing the cross-validation error from a grid of 100 candidate values extracted from the regularization path (Friedman et al. (2009)). These are the default settings of the *glmnet* package.

Table 2 reports the estimated prediction error of the methods investigated in the different datasets. The last two rows of this table display the average test error and the average rank of the classifiers. For each problem, the method that has the lowest error is highlighted in boldface. The overall performance of the Bayesian model with *spike and slab* priors, trained using EP is fairly good. This method obtains the lowest average prediction error and the highest average rank in the datasets investigated. Using the *spike and slab* prior instead of a standard Gaussian prior improves the performance of the method. In most problems EP also outperforms the MCMC method based on Gibbs sampling (Lee et al., 2003). This indicates that EP provides a sufficiently accurate approximation to the posterior probability at a reduced computational cost. This is an unexpected result because EP may have difficulties when the posterior distribution is multi-modal, as one would expect when a *spike and slab* prior is used (Seeger, 2008). The good performance of EP may indicate that, in the problems investigated, the posterior distribution has a single dominant mode. Among the other benchmark classifiers, the SVM with linear kernel has a good overall accuracy and obtains the best results in several problems. The predictors obtained by k-NN and DLDA are clearly inferior to those of the SVM. However, these predictors are remarkably accurate in a few datasets. The performance of the generalized LASSO is very good in some of the datasets investigated but it is severely impaired in others, e.g. *Brain A*, *Brain B* and *Mutation*. The poor behavior of this method in these datasets can be explained by the rather small number of training samples available: The LASSO can only select at most n variables for prediction, where n is the total number of training samples (Efron et al., 2004). Finally, RFE and NSC have the best performance among the remaining techniques that use feature selection. Nevertheless, RFE is worse than the SVM in several problems, which is in agreement with the findings of (Ambroise and McLachlan, 2002).

Additional experiments are carried out to compare the computational cost of the EP algorithm and the costs of the other four methods that have the best overall prediction performance in the problems investigated. Namely, the Bayesian model with MCMC sampling, the SVM, the RFE algorithm and the NSC method. The cost of training the Bayesian model with the EP algorithm is $\mathcal{O}(nd)$, where n is the number of training samples and d is the number of genes. In this model, inference with MCMC sampling has an average cost of $\mathcal{O}(\rho^2 d^3 k)$, where ρ is the prior probability of $\gamma_i = 1$, with $i = 1, \ldots, d$, and k is the number of samples generated from the Markov chain. Since in our setting $d \gg n$, the cost of training the SVM is $\mathcal{O}(n^2 d)$, which is the cost of computing the Gram matrix using a linear kernel. Finally, the training cost for the NSC method is $\mathcal{O}(nd)$ and the training cost of the RFE algorithm is approximately $\mathcal{O}(((2 - 2^{-p})d + d^2 2^{-2p})n^2)$, where $p = \lceil \log_2(d) - \log_2(500) \rceil$, since we remove half of the variables each time until less than 500 variables remain. To measure the time needed for training these different predictors we carry out experiments in a representative subset of the classification problems listed in Table 1: *Ovarian*, *Colon*, *Leukemia*, *Down Syndrome*, *Prostate* and *Adenocarcinoma*. On each of these datasets we train each method using all the instances available. This process is repeated ten times and the average training time of each method is recorded. The computer that is used for the calculation of these estimates is an AMD Opteron 252/2.6 GHz. To guarantee a fair comparison of the different methods a version of the EP algorithm in C has been used. The MCMC sampling approach and the NSC method have been coded in Fortran. The SVM and the RFE methods use the fast SVM implementation provided in the *LIBSVM*, which is written in C++ (Chang and Lin, 2001). The results of these experiments are displayed in Table 3. This table confirms that the EP algorithm is much faster than the MCMC sampling approach. In all the datasets investigated the EP algorithm only requires a few seconds for training the Bayesian model. By contrast, the MCMC approach requires several thou-

Table 2
Classification errors of each method for the different microarray datasets. The last two rows display the average error and the average rank of the classification systems over all the datasets, respectively.

Dataset	$EP_{\rho=1}$	EP	MCMC	SVM	KNN	DLDA	RFE	RVM	NSC	RF	LASSO
Adenocarcinoma	38.0 ± 10.1	15.2 ± 5.5	15.0 ± 6.0	**13.8 ± 6.0**	17.0 ± 4.7	30.6 ± 8.0	19.2 ± 6.6	22.5 ± 8.1	16.2 ± 5.5	18.7 ± 7.0	16.6 ± 5.4
Brain A	5.4 ± 9.5	4.9 ± 9.4	4.9 ± 9.4	2.3 ± 5.3	**0.6 ± 2.8**	6.9 ± 10.1	5.7 ± 10.4	14.6 ± 15.7	18.3 ± 21.4	18.0 ± 14.4	15.7 ± 21.3
Brain B	28.4 ± 15.7	21.3 ± 15.7	24.7 ± 14.7	**15.1 ± 10.0**	24.4 ± 13.4	17.5 ± 12.2	20.0 ± 11.8	24.2 ± 12.8	18.4 ± 12.9	23.8 ± 12.4	26.4 ± 9.6
Brain C	39.9 ± 10.0	37.7 ± 10.7	39.2 ± 10.5	**33.8 ± 8.8**	40.0 ± 8.0	38.6 ± 8.9	38.3 ± 9.2	45.1 ± 9.1	38.5 ± 8.3	42.4 ± 11.1	39.8 ± 10.4
Breast ER	12.2 ± 7.4	**12.1 ± 7.3**	12.4 ± 7.4	12.5 ± 8.1	18.9 ± 9.6	13.6 ± 8.3	17.1 ± 9.8	16.4 ± 10.4	15.6 ± 10.2	20.2 ± 9.2	16.5 ± 9.6
Breast LN	38.5 ± 12.1	33.4 ± 10.8	35.6 ± 12.6	41.1 ± 11.3	44.0 ± 10.2	39.4 ± 11.9	33.6 ± 12.3	27.0 ± 9.7	30.4 ± 16.0	34.9 ± 12.0	**26.2 ± 10.6**
Colon	17.3 ± 7.3	16.3 ± 7.2	17.2 ± 7.6	17.6 ± 8.5	28.8 ± 10.4	16.4 ± 7.3	19.0 ± 7.7	18.8 ± 7.5	**13.9 ± 7.3**	23.4 ± 8.7	19.7 ± 8.4
Down Syndrome	9.3 ± 7.0	8.2 ± 7.2	8.4 ± 6.8	6.1 ± 5.2	15.3 ± 7.4	11.3 ± 6.7	5.6 ± 5.1	7.9 ± 4.7	5.5 ± 6.1	6.8 ± 5.9	**5.2 ± 4.3**
Leukemia	6.0 ± 4.8	4.2 ± 3.5	5.4 ± 4.2	**2.0 ± 2.4**	5.7 ± 4.0	2.8 ± 3.2	3.2 ± 3.8	6.1 ± 4.7	3.9 ± 4.4	5.8 ± 5.3	6.8 ± 5.5
Lymphoma	4.2 ± 3.3	4.0 ± 3.4	4.2 ± 3.3	**0.5 ± 1.4**	2.0 ± 2.6	1.6 ± 2.5	2.8 ± 3.2	3.0 ± 2.7	3.3 ± 3.6	5.6 ± 5.5	1.6 ± 3.3
Metastasis	36.2 ± 7.4	36.0 ± 7.4	36.4 ± 7.6	38.8 ± 9.1	42.2 ± 8.0	35.6 ± 8.2	38.3 ± 10.0	38.5 ± 8.0	38.2 ± 8.8	38.5 ± 8.5	**35.2 ± 10.4**
Mutation	13.7 ± 12.6	**12.0 ± 12.4**	12.9 ± 12.3	24.0 ± 13.7	24.0 ± 12.1	23.1 ± 13.5	23.1 ± 15.0	32.6 ± 16.3	27.4 ± 18.4	38.3 ± 14.6	34.3 ± 16.1
Ovarian	9.7 ± 5.8	**9.3 ± 5.7**	9.8 ± 6.2	9.7 ± 4.2	15.8 ± 7.8	9.9 ± 5.6	12.3 ± 7.0	12.4 ± 7.2	11.0 ± 5.8	12.1 ± 7.9	11.8 ± 6.5
Prostate	9.0 ± 4.4	9.2 ± 6.1	**7.8 ± 4.2**	9.4 ± 4.0	20.0 ± 7.1	37.6 ± 12.4	8.8 ± 4.6	10.4 ± 4.9	10.8 ± 5.2	8.9 ± 4.7	9.5 ± 4.4
SRBCT	4.7 ± 5.7	4.0 ± 5.0	3.9 ± 5.0	3.9 ± 5.2	21.6 ± 9.4	7.7 ± 7.2	**3.1 ± 4.1**	5.4 ± 6.0	3.7 ± 5.8	8.4 ± 6.6	5.4 ± 6.7
Avg. Error	18.2 ± 13.8	**15.2 ± 11.7**	15.8 ± 12.3	15.4 ± 13.4	21.3 ± 13.4	19.5 ± 13.6	16.7 ± 12.4	19.0 ± 12.6	17.0 ± 11.8	20.4 ± 12.9	18.0 ± 12.0
Avg. Rank	6.5 ± 3.0	**3.7 ± 2.2**	4.9 ± 2.6	4.0 ± 3.2	8.5 ± 3.1	5.8 ± 3.3	5.1 ± 2.6	7.7 ± 2.2	5.1 ± 2.8	8.3 ± 2.5	6.4 ± 3.4

Table 3
Average training time (in seconds) for the five methods that have the best overall performance. Results are displayed for a representative subset of the microarray classification problems. The problems are ordered according to the total number of genes.

Dataset	Genes	Patients	Average training time				
			EP	MCMC	SVM	RFE	NSC
Ovarian	1536	54	0.44	3427.29	0.20	123.66	0.07
Colon	2000	62	1.10	4625.56	0.30	219.01	0.11
Leukemia	3571	72	2.07	10197.48	0.68	190.51	0.26
Down Syndrome	4656	63	2.37	10374.56	0.76	91.66	0.27
Prostate	6033	102	8.44	21866.55	1.72	209.33	0.62
Adenocarcinoma	9868	76	16.01	26182.53	2.05	142.59	0.99

sands. The largest difference in the execution time of these two methods is found in the *Adenocarcinoma* dataset. In this dataset the EP algorithm only requires on average ≈16 s for training while the MCMC sampling approach requires more than 7 h. Training a single SVM is also very fast. In general, faster than the EP algorithm. Comparing the results obtained, NSC is the fastest model to train. On average, it never requires more than 1 s. Except for the RFE algorithm, the training cost of all methods increase with the number of genes. Training an RFE model is particularly expensive because it involves training several SVMs. The highest cost of this method corresponds to the final stage of the search for subsets of relevant variables: The size of the set of explanatory variables is iteratively reduced by retaining only half of variables in the previous step, until a subset of size 500 or less remains. Beyond this point the algorithm attempts to remove variables one at a time. This explains the larger cost of training RFE in the *Colon* problem (500 variables in the final stage), and the lower cost in the *Down Syndrome* problem (291 variables in the final stage).

Besides being faster and more accurate than the MCMC sampling approach, the EP algorithm is also useful to identify relevant genes for subsequent analysis. To illustrate this point, we use EP to train a Bayesian model assuming *spike and slab* priors on the *Leukemia* dataset using all the instances available. We then select the 16 genes that are ranked as most relevant for the classification process according to the probability vector $\boldsymbol{p}$ of the posterior approximation for $\boldsymbol{\gamma}$. Fig. 1 displays a heat map of these genes for each patient in the dataset. The columns correspond to different patients. The rows represent the selected genes. The ID of the genes is indicated on the right-hand-side of the table. The posterior probability of the corresponding gene appears on the left-hand-side of the table. Each cell in the map represents the expression of one gene in one patient and its color depicts the intensity (normalized expression level) from blue (large negative) to yellow (large positive). Patients are grouped according to the assigned class label. Genes are ordered according to their average level of expression in the majority class: upper rows correspond to high average nor-

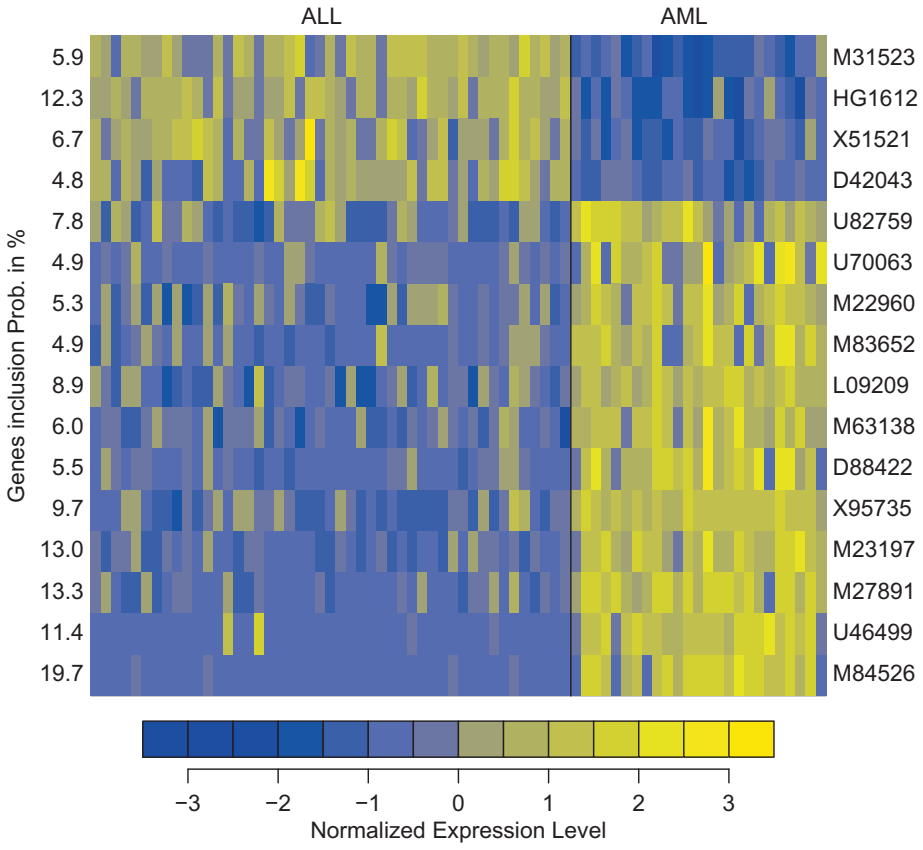

Fig. 1. Normalized expression level of the 16 top-ranked genes in the *Leukemia* dataset using the EP algorithm. Each row shows the expression level of a gene for each patient. Patients are grouped according to their class (*AML* and *ALL*). Genes are sorted according to the mean expression level within the majority class.

malized expression levels in the majority class. This ordering of patients and genes uncovers a clear clustering pattern. There is a first group of genes (M31523 to D42043) whose normalized expression levels are consistently high for *AML* patients and low for *ALL* patients. The remaining genes exhibit the opposite pattern. The relevance of these genes for discriminating between the two class labels is apparent in this graph. Further evidence of the relevance of these genes comes from the scientific literature: The *M84526* genes are listed among the most relevant features in (Bø and Jonassen, 2002; Lee et al., 2003) and are deemed useful for classification in (Golub et al., 1999; Krishnapuram et al., 2004); *M23197* appears as a relevant gene in (Golub et al., 1999; Li et al., 2002; Lee et al., 2003; Wang et al., 2005); *HG1612* is also selected for further analysis in (Guyon et al., 2002; Bø and Jonassen, 2002; Lee et al., 2003; Krishnapuram et al., 2004); *U46499* appears as an important gene in (Li et al., 2002; Bø and Jonassen, 2002; Lee et al., 2003; Wang et al., 2005); *X95735* is ranked first in (Guyon et al., 2002; Wang et al., 2005,) fourth in (Lee et al., 2003), and is described as relevant in (Golub et al., 1999; Li et al., 2002).

6. Conclusions

Microarray data classification problems often have a large high-dimensional attribute space and a very limited number of samples. In this work, we consider a Bayesian model to address the difficulties that arise in these non-standard learning conditions. A *spike and slab* prior distribution is used to favor the selection of sparse models (George and McCulloch, 1997). Sparsity is enforced by introducing one binary latent variable per gene. This latent variable indicates whether the corresponding expression level is used for prediction or not. Exact inference is infeasible because of the large dimensionality of the problem. Therefore, approximate techniques have to be employed. Because sparse prior distributions often lead to posterior distributions that are multi-modal, spreading the probability mass among the different modes, accurate approximate inference can be difficult (Seeger, 2008). The parameters of the Bayesian model are determined using Expectation Propagation (EP), which is an efficient algorithm for approximate inference (Minka, 2001b). The EP algorithm assumes that the posterior distribution is a product of simple terms that belong to the exponential family. These terms are iteratively updated until convergence using rules derived from moment matching conditions. Because the posterior approximation is assumed to factorize, the time-complexity of the algorithm is only $\mathcal{O}(nd)$, where n is the number of instances and d is the number of attributes (genes) per instance. This cost is typically much lower than alternative inference algorithms, such as MCMC sampling (Lee et al., 2003; Zhou et al., 2004; Bae and Mallick, 2004). Empirical evaluation on a set of 15 microarray datasets shows that EP is competitive with other microarray classification techniques. The posterior approximation obtained by EP is sufficiently accurate to provide a good overall performance at a reduced computational cost. Furthermore, the Bayesian approach considered is also useful for the identification of genes that are relevant for classification. A detailed investigation in the *Leukemia* dataset (Golub et al., 1999) shows that genes with the highest rank in the posterior approximation have expression levels that are either high for *AML* patients and low for *ALL* patients and the other way round. Therefore, these genes are also good candidates for subsequent analysis. Most of these genes are listed as relevant for classification in the literature (Golub et al., 1999; Guyon et al., 2002; Li et al., 2002; Bø and Jonassen, 2002; Lee et al., 2003; Krishnapuram et al., 2004; Wang et al., 2005). A shortcoming of the model considered is that it only applies to binary classification problems. Nevertheless, a multi-category extension may be possible following the analysis of (Zhou et al., 2006).

Acknowledgment

All authors acknowledge support from the Spanish Ministerio de Ciencia e Innovación under project TIN2007-66862-C02-02.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403 (6769), 503–511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96 (12), 6745–6750.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 99 (10), 6562–6566.

Bae, K., Mallick, B.K., 2004. Gene selection using a two-level hierarchical Bayesian model. Bioinformatics 20 (18), 3423–3430.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

Bø, T., Jonassen, I., 2002. New feature subset selection procedures for classification of expression profiles. Genome Biol. 3 (4), 1–11.

Bourquin, J.-P., Subramanian, A., Langebrake, C., Reinhardt, D., Bernard, O., Ballerini, P., Baruchel, A., 2006. Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. Proc. Natl. Acad. Sci. USA 103, 3339–3344.

Breiman, L., 1996. Out-of-bag estimation. Tech. Rep., Statistics Department, University of California.

Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.

Cawley, G.C., Talbot, N.L.C., 2006. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics 22 (19), 2348–2355.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at <http://www.csie.ntu.edu.tw/cjlin/libsvm>.

Dettling, M., 2004. Bagboosting for tumor classification with gene expression data. Bioinformatics 20 (18), 3583–3593.

Dettling, M., Bühlmann, P., 2002. Supervised clustering of genes. Genome Biol. 3, 1–15.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7 (1), 3.

Dougherty, E.R., 2001. Small sample issues for microarray-based classification. Comp. Funct. Genomics 2 (1), 28–34.

Dudoit, S., Fridlyand, J., 2003. Statistical Analysis of Gene Expression Microarray Data. CRC Press (Chapter 3).

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist., 407–451.

Friedman, J.H., Hastie, T., Tibshirani, R., 2009. Regularization paths for generalized linear models via coordinate descent. J. Statist. Software 33 (1), 1–22.

George, E.I., McCulloch, R.E., 1997. Approaches for Bayesian variable selection. Statist. Sinica 7 (2), 339–373.

Gill, P.E., Golub, G.H., Murray, W., Saunders, M.A., 1974. Methods for modifying matrix factorizations. Math. Comput. 28 (126), 505–535.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531–537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learn. 46 (1–3), 389–422.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning. Springer.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A., Trent, J., 2001. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. 344 (8), 539–548.

Hernández-Lobato, J.M., Dijkstra, T., Heskes, T., 2008. Regulator discovery from gene expression time series of malaria parasites: A hierarchical approach. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, vol. 20. The MIT Press, pp. 649–656.

Ishwaran, H., Rao, J., 2005. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann. Statist. 33 (2), 730–773.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7 (6), 673–679.

Krishnapuram, B., Carin, L., Hartemink, A.J., 2004. Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data. J. Comput. Biol. 11 (2–3), 227–242.

Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., Mallick, B.K., 2003. Gene selection: A Bayesian variable selection approach. Bioinformatics 19 (1), 90–97.

Lee, J.W., Lee, J.B., Park, M., Song, S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Comput. Statist. Data Anal. 48 (4), 869–885.

Li, Y., Campbell, C., Tipping, M., 2002. Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18 (10), 1332–1339.

MacKay, D.J.C., 2003. Information Theory, Inference & Learning Algorithms. Cambridge University Press, Cambridge, UK.

Minka, T., 2001a. Expectation propagation for approximate Bayesian inference. In: Breese, J.S., Koller, D. (Eds.), Proc. 17th Annual Conf. on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, pp. 362–369.

Minka, T., 2001b. A family of algorithms for approximate Bayesian inference. Ph.D. Thesis, Massachusetts Institute of Technology.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415 (6870), 436–442.

Rakotomamonjy, A., 2003. Variable selection using svm based criteria. J. Machine Learn. Res. 3, 1357–1370.

Ramaswamy, S., Ross, K.N., Lander, E.S., Golub, T.R., 2003. A molecular signature of metastasis in primary solid tumors. Nat. Genet. 33, 49–54.

Roth, V., 2002. The generalized LASSO: A wrapper approach to gene selection for microarray data. Tech. rep., University of Bonn, Computer Science III, Roemerstr. 164, D-53117 Bonn, Germany.

Schummer, M., Ng, W.V., Bumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y., Hood, L., 1999. Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 238 (2), 375–385.

Seeger, M.W., 2008. Bayesian inference and optimal design for the sparse linear model. J. Machine Learn. Res. 9, 759–813.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B – Methodol. 58 (1), 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA 99 (10), 6567–6572.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Machine Learn. Res. 1, 211–244.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 (6871), 530–536.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc..

Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F., Mewes, H.W., 2005. Gene selection from microarray data for cancer classification – A machine learning approach. Comput. Biol. Chem. 29 (1), 37–46.

West, M., Blanchette, C., Dressman, H., 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. 98 (20), 11462–11467.

Williams, P.M., 1995. Bayesian regularization and pruning using a Laplace prior. Neural Comput. 7 (1), 117–143.

Wipf, D., Nagarajan, S., 2008. A new view of automatic relevance determination. In: Advances in Neural Information Processing Systems, vol. 20.

Zhou, X., Liu, K.-Y., Wong, S.T., 2004. Cancer classification and prediction using logistic regression with Bayesian gene selection. J. Biomed. Inform. 37 (4), 249–259.

Zhou, X., Wang, X., Dougherty, E., 2006. Multi-class cancer classification using multinomial probit regression with Bayesian gene selection. IEE Proc. Systems Biol. 153 (2), 70–78.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Statist. Soc. Ser. B 67, 301–320.