J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 219–234

Bayesian classification of tumours by using gene expression data

Bani K. Mallick,

Texas A&M University, College Station, USA

Debashis Ghosh

University of Michigan, Ann Arbor, USA

and Malay Ghosh

University of Florida, Gainesville, USA

[Received June 2002. Final revision June 2004]

Summary. Precise classification of tumours is critical for the diagnosis and treatment of cancer. Diagnostic pathology has traditionally relied on macroscopic and microscopic histology and tumour morphology as the basis for the classification of tumours. Current classification frameworks, however, cannot discriminate between tumours with similar histopathologic features, which vary in clinical course and in response to treatment. In recent years, there has been a move towards the use of complementary deoxyribonucleic acid microarrays for the classification of tumours. These high throughput assays provide relative messenger ribonucleic acid expression measurements simultaneously for thousands of genes. A key statistical task is to perform classification via different expression patterns. Gene expression profiles may offer more information than classical morphology and may provide an alternative to classical tumour diagnosis schemes. The paper considers several Bayesian classification methods based on reproducing kernel Hilbert spaces for the analysis of microarray data. We consider the logistic likelihood as well as likelihoods related to support vector machine models. It is shown through simulation and examples that support vector machine models with multiple shrinkage parameters produce fewer misclassification errors than several existing classical methods as well as Bayesian methods based on the logistic likelihood or those involving only one shrinkage parameter.

Keywords: Gibbs sampling; Markov chain Monte Carlo methods; Metropolis–Hastings algorithm; Microarrays; Reproducing kernel Hilbert space; Shrinkage parameters; Support vector machines

1. Introduction

Precise classification of tumours is of critical importance to the diagnosis and treatment of cancer. Targeting specific therapies to pathogenetically distinct types of tumour is important for the treatment of cancer because it maximizes efficacy and minimizes toxicity (Golub *et al.*, 1999). Diagnostic pathology has traditionally relied on macroscopic and microscopic histology and tumour morphology as the basis for the classification of tumours. Current frameworks, however, cannot discriminate between tumours with similar histopathologic features, which vary in clinical course and in response to treatment. There is increasing interest in changing the

Address for correspondence: Bani K. Mallick, Department of Statistics, 3143 TAMU, Texas A&M University, College Station, TX 77843-3143, USA.
E-mail: bmallick@stat.tamu.edu

 1369–7412/05/67219

basis of tumour classification from morphologic to molecular, using microarrays which provide expression measurements for thousands of genes simultaneously (Schena *et al.*, 1995; DeRisi *et al.*, 1997), a key goal being to perform classification via different expression patterns. Several studies using microarrays to profile colon, breast and other tumours have demonstrated the potential power of expression profiling for classification (Alon *et al.*, 1999; Hedenfalk *et al.*, 2001). Gene expression profiles may offer more information than and provide an alternative to morphology-based tumour classification systems. We focus on the classification of microarray data.

In such analyses, there is a set of observations that contains vectors of gene expression data as well as the labels of the corresponding tissues. Golub *et al.* (1999) used supervised learning methods and derived discriminant decision rules by using the magnitude and threshold of prediction strength. However, they did not provide the procedure for selecting a cut-off value, which is an essential ingredient for their approach. Heuristic rules for selection of the threshold can be used, but this introduces an unavoidable subjectivity. Moler *et al.* (2000) proposed a naïve Bayes model and Xiong *et al.* (2000) conducted linear discriminant analysis for the classification of tumours. Brown *et al.* (2000) used a support vector machine (SVM) to classify genes rather than samples. Dudoit *et al.* (2002) compared several discriminant methods for the classification of tumours.

The main difficulty with microarray data analysis is that the sample size n is small compared with the number of genes p. This is known as the 'large p, small n' problem (West, 2003). In this situation, dimension reduction is needed to reduce the high dimensional gene space. Most existing approaches perform a preliminary selection of genes based on some criterion and use only 5–10% of the genes for classification. In our approach, we can utilize all the genes rather than eliminating most of them on the basis of a crude criterion.

In this paper we construct Bayesian binary classification models for prediction based on a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950; Parzen, 1970) approach. The methods are quite general and, in particular, can be used for the classification of tumours. One nice property of RKHS methods is that they allow us to project the prediction problem into a space which is of dimension $n \ll p$. Usually RKHSs have been used in a decision theory framework with no explicit underlying probabilistic model. Consequently, it is not possible to assess the uncertainty either of the classifier itself or of predictions that are based on it.

Our goal is to present a full probabilistic model-based approach to RKHS-based classification. First we shall consider the logistic classifier in this framework and then extend it to SVM classifiers (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002). As with other regularization methods, there are smoothing or regularization parameters which need to be tuned for efficient classification. One popular approach is to use generalized approximate cross-validation (Wahba *et al.*, 2002) to tune the smoothing parameters. In this paper we take a different approach, by developing a hierarchical model where the unknown smoothing parameter will be interpreted as a shrinkage parameter (Denison *et al.*, 2002). We shall assign a prior distribution to it and obtain its posterior distribution via the Bayesian paradigm. In this way, we obtain not only the point predictors but also the associated measures of uncertainty. Furthermore, we shall extend the model to incorporate multiple smoothing parameters, leading to significant improvements in prediction for the examples that are considered.

Before proceeding further, we review briefly related work in Bayesian learning and compare our proposal with some existing methods. Tipping (2000, 2001) and Bishop and Tipping (2000) introduced relevance vector machines (RVMs) in place of SVMs. Their objective, like ours, was to obtain predictive distributions for future observations rather than just point predictors. In the classification context, they began with a likelihood based on binary data with a logit

link, as in Section 4.1 of this paper. The logits were assumed to have a regression structure with a fairly general basis function including the basis function that is considered here. Then a normal distribution was assigned to the vector of regression coefficients (which they called 'weights'). Finally the Bayesian procedure was implemented by finding the posterior modes of these regression coefficients, or some approximations thereof. Figueiredo (2002) took a similar approach but used the probit instead of the logit link. He also used Jeffreys's prior instead of the usual Gaussian prior for regression coefficients. Zhu and Hastie (2002) proposed a frequentist approach using only a subset of the regression vectors. They referred to the resulting procedure as an import vector machine. They used iterative reweighted least squares as the fitting method.

The present approach, though similar in spirit, is operationally quite different from these approaches. First, the logits or the probits are not deterministic functions of the basis vectors but include in addition a random error to account for any unexplained sources of variation. For classification models with binary data it is well known that conjugate priors do not exist for the regression coefficients and hence the computation becomes more difficult. By adopting a Gaussian residual effect, many of the conditional distributions for the model parameters are of standard form, which greatly aids the computations. Also, rather than estimating the hyperparameters, we assign distributions to them, thus accounting for uncertainty due to the estimation of hyperparameters. Finally, a key feature of our method is the treatment of model uncertainty through a prior distribution over the kernel parameter.

RVMs introduce sparseness in the model by considering heavy-tailed priors such as double-exponential priors for the regression coefficients (Bishop and Tipping, 2000; Figueiredo, 2002). This opportunity exists also for the SVM that is considered in this paper, even though the binary probabilities are then modelled differently. In fact, in our examples, with a Bayesian hierarchical set-up the SVM shows more sparseness than does the logistic model. Several researchers exploited this sparseness property to select significant genes (Roth, 2002; Lee *et al.*, 2003). Our main emphasis, however, is to obtain predictive distributions for future observations to be used for classification rather than direct estimation of the parameters.

The idea of multiple smoothing parameters that is used in this paper has also been addressed elsewhere. In the machine learning literature; this is known as automatic relevance determination (Mackay, 1996; Neal, 1996). An advantage of using multiple parameters is that it enables us to detect the varying influence of different regression coefficients for prediction or classification.

Section 2 of this paper introduces the RKHS-based classification method. The hierarchical classification model is introduced in Section 3. Section 4 provides the various likelihoods for the logistic and the SVM classification models. Implementation of the Bayesian method is discussed in Section 5. Section 6 discusses prediction and choice of model. Section 7 contains the examples. Section 8 contains some simulation results. Finally, some concluding remarks are made in Section 9.

2. Classification method based on reproducing kernel Hilbert spaces

For a binary classification problem, we have a training set $\{y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$, where y_i is the response variable indicating the class to which the ith observation belongs and $\mathbf{x}_i$ is the vector of covariates of size p. The objective is to predict the posterior probability of belonging to one of the classes given a set of new covariates, based on the training data. Usually the response is coded as $y_i = 1$ for class 1 and $y_i = 0$ (or $y_i = -1$) for the other class. We utilize the training data $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ to fit a model $p(\mathbf{y}|\mathbf{x})$ and use it to obtain $P(y_* = 1|\mathbf{y}, \mathbf{x}_*)$ for a future observation y_* with covariate $\mathbf{x}_*$.

For our problem, we have binary responses as $y_i = 1$ indicates that the tumour sample i is from class 1 and $y_i = 0$ (or $y_i = -1$) indicates that it belongs to class 2, for $i = 1, \ldots, n$. Gene expression data on p genes for n tumour samples are summarized by an $n \times p$ matrix, so x_{ij} is the measurement of the expression level of the jth gene for the ith sample ($i = 1, \ldots, n$; $j = 1, \ldots, p$). It is assumed that the expression levels x_{ij} represent rigorously processed data that have undergone image processing as well as within- and between-slide normalization.

To develop a general model for classification, we need to specify a probability model for $p(\mathbf{y}|\mathbf{x})$ where $\mathbf{x}$ is high dimensional. To simplify the structure, we introduce the latent variables $\mathbf{z} = (z_1, \ldots, z_n)$ and assume that $p(\mathbf{y}|\mathbf{z}) = \Pi_{i=1}^n p(y_i|z_i)$, i.e. the y_i are conditionally independent given the z_i. In the next stage, the latent variables z_i are modelled as $z_i = f(\mathbf{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, where f is not necessarily a linear function, and ε_i, the random residual effects, are independent and identically distributed $N(0, \sigma^2)$. The use of a residual component is consistent with the belief that there may be unexplained sources of variation in the data. By adopting this Gaussian residual effect, many of the conditional distributions for the model parameters are of standard forms, which greatly aids the computations. To develop the complete model, we need to specify $p(\mathbf{y}|\mathbf{z})$ and f.

In the machine learning literature, most of the binary classification procedures emerged from a loss-function-based approach. In the same spirit, we model $p(\mathbf{y}|\mathbf{z})$ on the basis of a loss function $l(\mathbf{y}, \mathbf{z})$, which measures the loss for reporting $\mathbf{z}$ when the truth is $\mathbf{y}$. Mathematically, minimizing this loss function is equivalent to maximizing $-l(\mathbf{y}, \mathbf{z})$, where $\exp\{-l(\mathbf{y}, \mathbf{z})\}$ is proportional to the likelihood function. This duality between 'likelihood' and 'loss', particularly viewing the loss as the negative of the log-likelihood, is referred to in the Bayesian literature as a 'logarithmic scoring rule' (see, for example, Bernardo (1979), page 688). Specific choices of the loss functions and the corresponding likelihood functions are discussed in Section 4.

To model the high dimensional function $f(\mathbf{x})$, we adopt the RKHS approach. A Hilbert space H is a collection of functions on a set T with an associated inner product $\langle g_1, g_2 \rangle$ and norm $\|g_1\| = \langle g_1, g_1 \rangle^{1/2}$ for $g_1, g_2 \in H$. An RKHS H with reproducing kernel K (usually denoted as H_K) is a Hilbert space having an associated function K on $T \times T$ with the properties

(a) $K(\cdot, \mathbf{x}) \in H$ and
(b) $\langle K(\cdot, \mathbf{x}), g(\cdot) \rangle = g(\mathbf{x})$ for all $\mathbf{x} \in T$ and for every g in H.

Here $K(\cdot, \mathbf{x})$ and $g(\cdot)$ are functions that are defined on T with values at $\mathbf{x}^* \in T$ equal to $K(\mathbf{x}^*, \mathbf{x})$ and $g(\mathbf{x}^*)$ respectively. The reproducing kernel function provides the fundamental building-blocks of H as a result of the following lemma from Parzen (1970).

Lemma 1. If K is a reproducing kernel for the Hilbert space H, then $\{K(\cdot, \mathbf{x})\}$ span H.

To prove the lemma it suffices to prove that the only function g in H orthogonal to $K(\cdot, \mathbf{x})$ is the zero function, but this is obvious, since by the reproducing property $\langle g(\cdot), K(\cdot, \mathbf{x}) \rangle = 0$ for every $\mathbf{x} \in T$ implies that $g(\mathbf{x}) = 0$ for all $\mathbf{x}$.

Lemma 1 has the consequence that functions of the form $g_N(\cdot) = \Sigma_{j=1}^N \beta_j K(\cdot, \mathbf{x}_j)$, where $\mathbf{x}_j \in T$ for each $j = 1, \ldots, N$, are dense in H. More precisely, for any $g \in H_K$, there are choices of N and $\beta_1, \ldots, \beta_N$ such that a g_N can be constructed to approximate g to any desired level of accuracy. Thus, the reproducing kernel functions are the natural choice for basis expansion modelling in an RKHS setting.

In the present problem, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the observed covariate values and $\mathbf{z}$ the latent responses, and we take the unknown function $f \in H_K$ where the choice of K is discussed below. To find the optimal f based on $\mathbf{z}$ and $\mathbf{x}$, we minimize $\Sigma_{i=1}^n \{z_i - f(\mathbf{x}_i)\}^2 + \|f\|^2$ with respect to f. Arguing as in chapter 1 of Wahba (1990), this minimizer must admit the representation

$$f(\cdot) = \beta_0 + \sum_{j=1}^{n} \beta_j \, K(\cdot, \mathbf{x}_j). \tag{1}$$

This reduces the optimization problem to a finite dimension n which is not large for gene expression data. Also, inference about f boils down to inference about $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_n)^{\mathrm{T}}$.

With the present Bayesian formulation we need to assign a prior to $\boldsymbol{\beta}$. We shall provide a flexible and computationally convenient hierarchical prior for $\boldsymbol{\beta}$ in the next section. In addition we shall allow the kernel functions to depend on some unknown parameters to enrich the class of kernels and express them as $K(\cdot, \cdot|\boldsymbol{\theta})$. Hence, K becomes a function of an unknown parameter $\boldsymbol{\theta}$, but this dependence will be implicit through the remainder of the paper for notational simplicity.

Different choices of the reproducing kernel K generate different function spaces. Two classical choices are

(a) the Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\theta\}$ and
(b) the polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^{\theta}$.

Here $\mathbf{a} \cdot \mathbf{b}$ denotes the inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$. Both these kernels contain a single parameter θ.

3. Hierarchical classification model

We can construct a hierarchical model for classification as

$$p(y_i|z_i) \propto \exp\{-l(y_i, z_i)\}, \qquad i = 1, \ldots, n, \tag{2}$$

where the $y_1, y_2, \ldots, y_n$ are conditionally independent given $z_1, z_2, \ldots, z_n$ and l is any specific choice of the loss function as explained in the previous section. We relate z_i to $f(\mathbf{x}_i)$ by $z_i = f(\mathbf{x}_i) + \varepsilon_i$, where the ε_i are residual random effects.

As explained in the previous section, we express f as

$$f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{n} \beta_j \, K(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) \tag{3}$$

where K is a positive definite function of the covariates (inputs) $\mathbf{x}$ and we allow some unknown parameters $\boldsymbol{\theta}$ to enrich the class of kernels.

The random latent variable z_i is thus modelled as

$$z_i = \beta_0 + \sum_{j=1}^{n} \beta_j \, K(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) + \varepsilon_i = \mathbf{K}_i' \boldsymbol{\beta} + \varepsilon_i, \tag{4}$$

where the ε_i are independent and identically distributed $N(0, \sigma^2)$ variables, and

$$\mathbf{K}_i' = (1, K(\mathbf{x}_i, \mathbf{x}_1|\boldsymbol{\theta}), \ldots, K(\mathbf{x}_i, \mathbf{x}_n|\boldsymbol{\theta})), \qquad i = 1, \ldots, n.$$

To complete the hierarchical model, we need to assign priors to the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ^2. We assign to $\boldsymbol{\beta}$ the Gaussian prior with mean $\mathbf{0}$ and variance $\sigma^2 \mathbf{D}_*^{-1}$, where $\mathbf{D}_* \equiv \mathrm{diag}(\lambda_1, \lambda, \ldots, \lambda)$ is an $(n+1) \times (n+1)$ diagonal matrix, λ_1 being fixed at a small value, but λ is unknown. This amounts to a large variance for the intercept term. We shall assign a proper uniform prior to $\boldsymbol{\theta}$, an inverse gamma prior to σ^2 and a gamma prior to λ. A gamma(α, ξ) distribution for a random variable, say U, has probability density function proportional to $\exp(-\xi u) u^{\alpha - 1}$, and the reciprocal of U will then be said to have an IG(α, ξ) distribution. Our model is thus given by

$$p(y_i|z_i) \propto \exp\{-l(y_i, z_i)\},$$

$$z_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \overset{\text{ind}}{\sim} N_1(z_i|\mathbf{K}_i'\boldsymbol{\beta}, \sigma^2), \tag{5}$$

$$\boldsymbol{\beta}, \sigma^2 \sim N_{n+1}(\boldsymbol{\beta}|\mathbf{0}, \sigma^2\mathbf{D}_*^{-1})\,\mathrm{IG}(\sigma^2|\gamma_1, \gamma_2), \tag{6}$$

$$\boldsymbol{\theta} \sim \prod_{q=1}^{p} U(a_{q1}, a_{q2}),$$

$$\lambda \sim \text{gamma}(m, c), \tag{7}$$

where $U(a_{q1}, a_{q2})$ is the uniform probability density function over (a_{q1}, a_{q2}).

We can extend this model by using multiple smoothing parameters so that the prior for $(\boldsymbol{\beta}, \sigma^2)$ is

$$\boldsymbol{\beta}, \sigma^2 \sim N_{n+1}(\boldsymbol{\beta}|0, \sigma^2\mathbf{D}^{-1})\,\mathrm{IG}(\sigma^2|\gamma_1, \gamma_2), \tag{8}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_{n+1}$. Once again λ_1 is fixed at a small value, but all other λs are unknown. We assign independent gamma(m, c) priors to them. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{n+1})^{\mathrm{T}}$.

To avoid the problem of specifying the hyperparameters m and c of $\boldsymbol{\lambda}$, we can use Jeffreys's independence prior $p(\boldsymbol{\lambda}) \propto \Pi_{i=1}^{n+1} \lambda_i^{-1}$. This is a limiting form of the gamma prior when both m and c go to 0. Figueiredo (2002) observed that this type of prior promoted sparseness, thus reducing the effective number of parameters in the posterior. Sparse models are preferable as they predict accurately using fewer parameters.

4. Likelihoods of reproducing kernel Hilbert space models

We now consider several possible expressions for $l(y_i, z_i)$ in expression (2).

4.1. Logistic classification model

If we code the responses y_i as 0 or 1 according to the classes, then the probability function is $p(y_i|z_i) = p_i^{y_i}(z_i)\{1 - p_i(z_i)\}^{1-y_i}$, where $p_i(z_i) = \exp(z_i)/\{1 + \exp(z_i)\}$. Then the log-likelihood is $\Sigma_{i=1}^n y_i z_i - \Sigma_{i=1}^n \log\{1 + \exp(z_i)\}$. We can use this log-likelihood function and the Bayesian model given in expressions (5)–(7) or expressions (5), (7) and (8) for prediction purposes. In the probit classification model, the set-up is similar, except that $p_i(z_i) = \Phi(z_i)$, where Φ denotes the standard normal distribution function.

4.2. Support vector machine model

We now describe the SVM classification method; for more details, the reader is referred to Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002) and Herbrich (2002). We code the class labels as $y_i = 1$ or $y_i = -1$. The idea behind SVMs is to find a linear hyperplane that separates the observations with $y = 1$ from those with $y = -1$ that has the largest minimal distance from any of the training examples. This largest minimal distance is known as the *margin*. As shown by Wahba (1999) or Pontil *et al.* (2000), this optimization problem amounts to finding $\boldsymbol{\beta}$ which minimizes $\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\,\Sigma_{i=1}^n [1 - y_i\, f(\mathbf{x}_i)]_+$, where $[a]_+ = a$ if $a > 0$ and $[a]_+ = 0$ otherwise, $C \geqslant 0$ is a penalty term and f is defined in equation (3). The problem can be solved by using non-linear programming methods. Fast algorithms for computing SVM classifiers can be found in chapter 7 of Cristianini and Shawe-Taylor (2000).

In a Bayesian formulation, this optimization problem is equivalent to finding the posterior mode of $\boldsymbol{\beta}$, where the likelihood is given by $\exp[-\Sigma_{i=1}^n \{1 - y_i\, f(\mathbf{x}_i)\}_+]$, and $\boldsymbol{\beta}$ has the

$N(\mathbf{0}, C\mathbf{I}_{n+1})$ prior. However, in our formulation with latent variables $\mathbf{z}$, we begin instead with the density

$$p(\mathbf{y}|\mathbf{z}) \propto \exp\Big(-\sum_{i=1}^{n}[1-y_i z_i]_+\Big) \tag{9}$$

and assume independent $N\{f(\mathbf{x}_i), \sigma^2\}$ priors for the z_i. The rest of the prior is the same as that given in expressions (6) and (7) or expressions (7) and (8).

If we use the density in expression (9), the normalizing constant may involve $\mathbf{z}$. Following Sollich (2001), we may bypass this problem by assuming a distribution for $\mathbf{z}$ such that the normalizing constant cancels out. If the normalized likelihood is

$$p(\mathbf{y}|\mathbf{z}) = \exp\Big(-\sum_{i=1}^{n}[1-y_i z_i]_+\Big)\Big/c(\mathbf{z}),$$

where $c(\cdot)$ is the normalizing constant, then, choosing $p(\mathbf{z}) \propto Q(\mathbf{z})\, c(\mathbf{z})$, the joint distribution turns out to be

$$p(\mathbf{y}, \mathbf{z}) \propto \exp\Big(-\sum_{i=1}^{n}[1-y_i z_i]_+\Big)\, Q(\mathbf{z}), \tag{10}$$

as the $c(\cdot)$ cancels from the expression. We shall take $Q(\mathbf{z})$ as the product of independent normal probability density functions with means $f(\mathbf{x}_i)$ and common variance σ^2. This method will be referred to as the Bayesian support vector machine (BSVM) classification.

The above procedure, which is analogous to that in Sollich (2001), makes the Bayesian analysis quite similar to the usual SVM analysis, but the prior on $\mathbf{z}$ is rather artificial and is intended mainly to cancel out a normalizing constant. The other option is to use a Bayesian approach to this problem by evaluating the normalizing constant properly and using it in the likelihood. Then the probability model (cf. Sollich (2001)) is

$$p(y_i|z_i) = \begin{cases} \{1+\exp(-2y_i z_i)\}^{-1} & \text{for } |z_i| \leqslant 1, \\ (1+\exp[-y_i\{z_i + \operatorname{sgn}(z_i)\}])^{-1} & \text{otherwise,} \end{cases} \tag{11}$$

where $\operatorname{sgn}(u) = 1, 0, -1$ according to whether u is greater than, equal to or less than 0.

The probability density function that is given in expression (11) will also be used to perform a Bayesian analysis. The resulting approach will be referred to as complete SVM (CSVM) classification and will be compared with the BSVM.

5. Bayesian analysis

For classification problems with binary data and a logistic likelihood, conjugate priors do not exist for the regression coefficients. Hence, without the tailored proposal densities that are needed for the implementation of the Metropolis–Hastings accept–reject algorithm, mixing in the Markov chain Monte Carlo sampler can be poor as updates are rarely accepted. The construction of good proposals depends on both the model and the data. The introduction of the latent variables z_i simplifies the computation (Holmes and Held, 2003), as we now show.

From the Bayes theorem,

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \sigma^2, \boldsymbol{\lambda}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\lambda})\; p(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \sigma^2). \tag{12}$$

This distribution is complex, and implementation of the Bayesian procedure requires Markov chain Monte Carlo sampling techniques, and, in particular, Gibbs sampling (Gelfand

and Smith, 1990) and Metropolis–Hastings algorithms (Metropolis *et al.*, 1953). The Gibbs sampler generates posterior samples by using conditional densities of the model parameters which we describe below.

First, we note again that, conditional on $\mathbf{z}$, all the other parameters are independent of $\mathbf{y}$ and furthermore the distributions follow from standard results for the Bayes linear model. This allows us to adopt conjugate priors for (β, σ^2) and hence to perform simulations as well as to marginalize over some of the parameter space.

5.1. Conditional distributions

The prior distributions given in expressions (6) and (7) of Section 2 are conjugate for β and σ^2, whose posterior density conditional on $\mathbf{z}, \boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is normal–inverse gamma,

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = N_{n+1}(\boldsymbol{\beta} | \tilde{\mathbf{m}}, \sigma^2 \tilde{\mathbf{V}}) \, \mathrm{IG}(\sigma^2 | \tilde{\gamma}_1, \tilde{\gamma}_2), \tag{13}$$

where $\tilde{\mathbf{m}} = (\mathbf{K}_0' \mathbf{K}_0 + \mathbf{D})^{-1} (\mathbf{K}_0' \mathbf{z})$, $\tilde{\mathbf{V}} = (\mathbf{K}_0' \mathbf{K}_0 + \mathbf{D})^{-1}$, $\tilde{\gamma}_1 = \gamma_1 + n/2$ and $\tilde{\gamma}_2 = \gamma_2 + \frac{1}{2}(\mathbf{z}'\mathbf{z} - \tilde{\mathbf{m}}' \tilde{\mathbf{V}} \tilde{\mathbf{m}})$. Here $\mathbf{K}_0' = (\mathbf{K}_1, \ldots, \mathbf{K}_n)$, and we recall that $\mathbf{K}_i = (K(\mathbf{x}_i, \mathbf{x}_1), \ldots, K(\mathbf{x}_i, \mathbf{x}_n))^{\mathrm{T}}$.

The conditional distribution for the precision parameter λ_i given the coefficient β_i is gamma and is given by

$$p(\lambda_i | \beta_i) = \mathrm{gamma}\left(m + \frac{1}{2}, c + \frac{1}{2\sigma^2} \beta_i^2 \right), \qquad i = 2, \ldots, n+1. \tag{14}$$

Finally, the full conditional density for z_i is

$$p(z_i | z_{-i}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\lambda}) \propto \exp\left[-l(y_i, z_i) - \frac{1}{2\sigma^2} \left\{ z_i - \sum_{j=1}^{n} \beta_j \, K(\mathbf{x}_i, \mathbf{x}_j) \right\}^2 \right].$$

Similarly, the full conditionals are found when $\lambda_2 = \ldots = \lambda_{n+1} = \lambda$ from expressions (7) and (8).

5.2. Posterior sampling of the parameters

We make use of the distributions that are given in Sections 4 and 5 through a Gibbs sampler that iterates through the following steps:

(a) update $\mathbf{z}$;
(b) update $\mathbf{K}, \boldsymbol{\beta}$ and σ^2;
(c) update $\boldsymbol{\lambda}$.

For the update to $\mathbf{z}$, we propose to update each z_i in turn conditionally on the rest, i.e. we update $z_i | \mathbf{z}_{-i}, \mathbf{y}, \mathbf{K}, \sigma^2, \boldsymbol{\beta}$ $(i = 1, \ldots, n)$, where $\mathbf{z}_{-i}$ indicates the $\mathbf{z}$-vector with the ith element removed.

The conditional distribution of z_i does not have an explicit form; we thus resort to the Metropolis–Hastings procedure with a proposal density $T(z_i^* | z_i)$ that generates moves from the current state z_i to a new state z_i^*. The proposed updates are then accepted with probabilities

$$\alpha = \min\left\{ 1, \frac{p(y_i | z_i^*) \; p(z_i^* | \mathbf{z}_{-i}, \mathbf{K}) \; T(z_i | z_i^*)}{p(y_i | z_i) \; p(z_i | \mathbf{z}_{-i}, \mathbf{K}) \; T(z_i^* | z_i)} \right\}; \tag{15}$$

otherwise the current state is retained.

We obtain $p(y_i | z_i)$ from expression (9) and

$$p(z_i | \mathbf{z}_{-i}, \mathbf{K}) \propto \exp\{ -(z_i - \mathbf{K}_i \boldsymbol{\beta})^2 / 2\sigma^2 \}.$$

It is convenient to take the proposal distribution $T(z_i^*|z_i)$ to be a symmetric distribution (say Gaussian) with mean equal to the old value z_i and a prespecified standard deviation.

An update of $\mathbf{K}$ is equivalent to that of $\boldsymbol{\theta}$ and we need a Metropolis–Hastings algorithm to perform it. Now we need the marginal distribution of $\boldsymbol{\theta}$ conditional on $\mathbf{z}$. We can write

$$p(\boldsymbol{\theta}|z) \propto p(z|\boldsymbol{\theta})\ p(\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}^*$ denote the proposed change to the parameter. Then we accept this change with acceptance probability

$$\alpha = \min\left\{1, \frac{p(z|\boldsymbol{\theta}^*)}{p(z|\boldsymbol{\theta})}\right\}. \tag{16}$$

The ratio of the marginal likelihoods is given by

$$\frac{p(z|\boldsymbol{\theta}^*)}{p(z|\boldsymbol{\theta})} = \frac{|\tilde{\mathbf{V}}^*|^{1/2}}{|\tilde{\mathbf{V}}|^{1/2}} \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_2^*}\right)^{\tilde{\gamma}_1}, \tag{17}$$

where $\tilde{\mathbf{V}}^*$ and $\tilde{\gamma}_2^*$ are similar to $\tilde{\mathbf{V}}$ and $\tilde{\gamma}_2$ with $\boldsymbol{\theta}^*$ replacing $\boldsymbol{\theta}$. Updating $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\lambda}$ is straightforward as they are generated from standard distributions.

6. Prediction and choice of model

For a new sample with gene expression $\mathbf{x}_{\text{new}}$, the posterior predictive probability that its tissue type, denoted by y_{new}, is cancerous is

$$p(y_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{y}) = \int p(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \boldsymbol{\phi}, \mathbf{y})\ p(\boldsymbol{\phi}|\mathbf{y})\ \mathrm{d}\boldsymbol{\phi}, \tag{18}$$

where $\boldsymbol{\phi}$ is the vector of all the model parameters. Assuming conditional independence of the responses, this integral reduces to

$$\int p(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \boldsymbol{\phi})\ p(\boldsymbol{\phi}|\mathbf{y})\ \mathrm{d}\boldsymbol{\phi}. \tag{19}$$

The associated measure of uncertainty is $p(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{y})\{1 - p(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{y})\}$. The integral in expression (19) can be approximated by the Monte Carlo estimate

$$\sum_{i=1}^{M} p(y_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \boldsymbol{\phi}^{(i)})/M, \tag{20}$$

where $\boldsymbol{\phi}^{(i)} (i = 1, \ldots, M)$ are the Markov chain Monte Carlo posterior samples of the parameter $\boldsymbol{\phi}$.

To select from the various models, we shall generally use the misclassification error. When a test set is provided, we first obtain the posterior distributions of the parameters (training the model) based on the training data and use them to classify the test samples. For a new observation from the test set, say $y_{i,\text{tst}}$, we shall obtain the probability $p(y_{i,\text{tst}} = 1|\mathbf{y}_{\text{trn}}, \mathbf{x}_{\text{tst}})$ by using an equation that is similar to expression (19) and approximate it by its Monte Carlo estimate as in equation (20). When this estimated probability exceeds 0.5, the new observation is classified as 1. Otherwise, it is classified as 0 or -1, depending on whether we use the logistic or the SVM likelihood.

If no test set is available, we use a leave-one-out cross-validation approach. We shall exploit the technique that is described in Gelfand (1996) to simplify our computation. For the

cross-validation predictive density, in general, writing $\mathbf{y}_{-i}$ as the vector of y_js minus y_i,

$$p(y_i|\mathbf{y}_{-i}) = \frac{p(\mathbf{y})}{p(\mathbf{y}_{-i})} = \left\{ \int p(y_i|\mathbf{y}_{-i}, \phi)^{-1} \, p(\phi|\mathbf{y}) \, \mathrm{d}\phi \right\}^{-1}. \tag{21}$$

Monte Carlo integration yields

$$\hat{p}(y_i|\mathbf{y}_{-i}) = M \Big/ \sum_{j=1}^{M} p(y_i|\mathbf{y}_{-i}, \phi^{(j)})^{-1},$$

where $\phi^{(j)}$, $j = 1, \ldots, M$, are the Markov chain Monte Carlo posterior samples of the parameter vector ϕ. This simple expression is due to the fact that y_is are conditionally independent given ϕ_is. If we wish to make draws from $p(y_i|\mathbf{y}_{-i,\mathrm{trn}})$, then we need to use importance sampling (Gelfand, 1996).

7. Examples

We illustrate the methodology with several examples. For all examples, six models were fitted:

(a) logistic regression with a single penalty parameter;
(b) logistic regression with multiple penalty parameters;
(c) BSVM classification with a single penalty parameter;
(d) BSVM classification with multiple penalty parameters;
(e) CSVM classification with a single penalty parameter;
(f) CSVM classification with multiple penalty parameters.

We have used the SVM MATLAB toolbox to obtain the classical SVM results (see `http://eewww.eng.ohio-state.edu/~maj/osu`$_{\mathrm{svm}}$`/`). The tuning parameters are chosen by using an iterative solving method for the quadratic programming formulation of the SVMs which is known as sequential minimization optimization. We obtained the RVM (Tipping, 2000) MATLAB code from `http://research.microsoft.com/mlp/RVM/relevance.htm`.

Throughout the examples, we selected γ_1 and γ_2 to give a tight inverse gamma prior for σ^2 with mean 0.1. For $\boldsymbol{\lambda}$ we chose m and c so that the mean of the gamma distribution was small, say 10^{-3}, but with a large variance; a_{q1} and a_{q2}, the prior parameters of $\boldsymbol{\theta}$, are chosen by using the $\mathbf{x}$ in such a way that computation of the kernel function does not overflow or underflow. We performed the data analysis with both the Gaussian and the polynomial kernels $\mathbf{K}$ as introduced in Section 3, and the results showed very little difference. The results that are reported here are based on Gaussian kernels.

In all the examples we used a burn-in of 5000 samples, after which every 100th sample was retained in the next 50000 samples. The convergence and mixing of the chain were checked by using two independent chains and the methods that were described in Gelman (1996).

7.1. Bench-mark comparisons

We utilize artificially generated data in two dimensions to compare our models with other popular models. In these artificial data both class 1 and class 2 were generated from mixtures of two Gaussian distributions by Ripley (1996), with the classes overlapping to the extent that the Bayes error is around 8%.

In addition, we also analysed three well-known bench-mark data sets, compared results with several state of the art techniques and present the results in Table 1. The first two data sets are Pima Indians diabetes and *Leptograpsus* crabs (Ripley, 1996). The third data set is the

Table 1. Modal classification error rates and 95% credible intervals (in parentheses) for the bench-mark data sets

Method	*Error rates for the following data sets:*		
	Ripley's	*Pima*	*Crabs*
Logistic (single)	13.0 (11,17)	21.4 (20.1,24.3)	5 (4,6)
Logistic (multiple)	9.2 (9,12)	19.4 (18.9,21.4)	2 (1,3)
BSVM (single)	12.4 (11.1,16.8)	21 (20,23.9)	4 (2,5)
BSVM (multiple)	8.8 (8.4,11.6)	18.9 (18.3,20.6)	1 (0,4)
CSVM (single)	12.7 (10.8,16.7)	21.3 (19.9,24.1)	4 (2,5)
CSVM (multiple)	9.1 (8.9,12)	19.2 (18.9,21.6)	2 (1,4)
RVM	9.3	19.6	2
Variational RVM	9.2	19.6	—†
Jeffreys prior	9.6	18.5	0
Neural networks	—†	22.5	3.0
Classical SVM	13.2	21.2	4

†Not applicable.

Wisconsin breast cancer data which contain 10 basic features to classify two types of cancer: malignant and benign. We split the data randomly into training and testing partitions of sizes 300 and 269, and we report average results over 10 partitions.

In addition to the methods that are listed at the beginning of Section 7, we have performed analyses with variational RVMs (Bishop and Tipping, 2000), Bayesian neural networks (Williams and Barber, 1998) and the analysis using the Jeffreys prior as described in Figueiredo (2002). The results are given in Table 1. All our multiple shrinkage parameter models perform nearly as well as the best available alternatives.

7.2. Leukaemia data

The leukaemia data set was described in Golub *et al.* (1999). Bone marrow or peripheral blood samples were taken from 72 patients with either acute myeloid leukaemia or acute lymphoblastic leukaemia. Following the experimental set-up of Golub *et al.* (1999), the data are split into training and test sets. The former consists of 38 samples, of which 27 are acute lymphoblastic and 11 are acute myeloid leukaemia cases; the latter consists of 34 samples, 20 acute lymphoblastic and 14 acute myeloid leukaemia cases. The data set contains expression levels for 7129 human genes produced by Affymetrix high density oligonucleotide microarrays.

Golub *et al.* (1999) constructed a predictor by using their weighted voting scheme on the training samples and classified correctly on all samples for which a prediction is made, 29 of the 34, declining to predict for the other five. We have provided our results in Table 2 with the modal or most frequent number of misclassification errors (the modal values) as well as the error bounds (the maximum and minimum number of misclassifications).

Table 2 shows that the results that are produced by the multiple shrinkage parameter models are superior to the single-precision models as well as the classical SVM models. Though all the multiple shrinkage parameter models performed well, the best performer among these appears to be the BSVM model.

The use of RKHSs leads to a reduction in the dimension of the model, but the dimension can still be as high as the sample size. In the Bayesian hierarchical modelling framework, owing to shrinkage priors, we obtain sparsity automatically (Tipping, 2000). The effective number of

Table 2. Modal classification error rates and 95% credible intervals for the leukaemia data

Model	*Modal misclassification error*	*Error bound*
Logistic (single)	4	(2,6)
Logistic (multiple)	2	(1,4)
BSVM (single)	4	(3,7)
BSVM (multiple)	1	(0,3)
CSVM (single)	5	(3,8)
CSVM (multiple)	2	(1,6)
Classical SVM	4	
RVM	2	

parameters is the degrees of freedom DF of the model, which can be calculated as the trace of $\mathbf{K}(\mathbf{K}'\mathbf{K} + \mathbf{D}^{-1})^{-1}\mathbf{K}'$ (Hastie and Tibshirani (1990), page 52). Owing to the presence of the unknown parameter θ in the expression of $\mathbf{K}$, this θ induces a posterior distribution for DF (rather than a fixed value). The posterior distributions of DF for all the three multiple shrinkage parameter models were very similar.

7.3. Hereditary breast cancer data

Hedenfalk *et al.* (2001) studied gene expression in hereditary and sporadic breast cancers. Studying such cancers will allow physicians to understand the difference between the cancers from mutations in the BRCA1 and the BRCA2 breast cancer genes. In their study, Hedenfalk *et al.* (2001) examined 22 breast tumour samples from 21 breast cancer patients; all the patients except one were women and 15 of the women had hereditary breast cancer, seven had tumours with BRCA1 and eight had tumours with BRCA2. In the analysis of a complementary deoxyribonucleic acid microarray, 3226 genes were used for each breast tumour sample. We use our methods to classify BRCA1 *versus* the other (BRCA2 and sporadic). As a test data set is not available, we have used a full leave-one-out cross-validation test and use the number of misclassifications to compare the various approaches. We present our results in Table 3.

Table 3. Modal classification error rates and 95% credible intervals for the breast cancer data

Model	*Modal cross-validation error*†	*Error bound*
Logistic (single)	5	(4,8)
Logistic (multiple)	2	(2,4)
BSVM (single)	4	(3,7)
BSVM (multiple)	0	(0,3)
CSVM (single)	5	(3,8)
CSVM (multiple)	2	(1,4)
Feed-forward neural networks	2	
Probabilistic neural networks ($r = 0.01$)	3	
k nearest neighbour ($k = 1$)	4	
SVM	4	
Perceptron	5	

†Number of misclassified samples.

We have compared our cross-validation results with other popular classification algorithms including feed-forward neural networks (Williams and Barber, 1998), k nearest neighbours (Fix and Hodges, 1951), classical SVMs (Vapnik, 2000), perceptrons (Rosenblatt, 1962) and probabilistic neural networks (Specht, 1990) in Table 2. All these methods have used expression values of only 51 genes as used in Hedenfalk *et al.* (2001). All the multiple shrinkage parameter models have performed better than any other methods, with SVM performing the best.

7.4. Simulation study

To simulate a realistic data set for comparing the successful multiple shrinkage BSVM, SVM and CSVM models, we used the leukaemia data as a prototype. As realistic values of the parameters θ and $\boldsymbol{\beta}$ we used the posterior means from the original analysis of the data. Then we followed the structure of our models and performed two sets of simulations to generate the responses Y, one using the logistic model and the other using the CSVM model. We replicate each of the simulations 25 times, generating 25 different data sets. Then we analyse these training data sets by using the logistic, BSVM and CSVM models and obtain the average misclassifications in the test data for the three models. The average misclassifications should be lowest if we use the true model, but we want to see how the other models perform in this situation.

When the data are actually generated from a logistic model, the average number of misclassifications in the test data by using the logistic, BSVM and CSVM models are respectively 2.5, 2.7 and 3.2. Similarly, when the data are actually generated from a logistic CSVM model, the average number of misclassifications in the test data by using the logistic, BSVM and CSVM models are respectively 3.8, 2.2 and 2.1. Though none of the data were originally generated from the BSVM model (as it has no normalized distribution), in both cases it is very near the correct (best) model in terms of the average misclassification error.

7.5. Analysis with Jeffreys's prior

As discussed in Section 3, sparseness can be promoted by using the Jeffreys prior (Figueiredo, 2002). We reanalysed the two data sets by using the multiple shrinkage parameter models and Jeffreys's prior. The modal number of misclassification and average DF (within parentheses) results are presented in Table 4. In terms of misclassification, Jeffreys's prior, in general, is doing worse than the Gaussian prior models but it has smaller DF.

Table 4. Analysis with Jeffreys's prior: average number of misclassifications and DF (within parentheses)

Model	*Number of misclassifications for the following data sets:*	
	Leukaemia	*Breast cancer*
Logistic	2 (6.1)	3 (7.2)
BSVM	2 (5.2)	2 (5.9)
CSVM	3 (5.6)	3 (6.4)

8. Discussion

We have proposed an RKHS-based classification method for microarray data. It is shown that these models in a Bayesian hierarchical set-up with priors over the shrinkage (smoothing) parameters performed better than other popular classification methods. Also, multiple shrinkage parameter models always appear to be superior to single-parameter shrinkage models. With multiple shrinkage parameters, the regular BSVM model emerges as the winner in all the examples with the CSVM finishing a close second all the time. However, the CSVM provides a more formal probabilistic motivation for the use of SVMs and is more satisfactory from a Bayesian angle.

We point out also that, although SVMs have been very popular in the machine learning community, one problem with their use in practice in a non-Bayesian framework is the inability to quantify prediction error. By using the Bayesian framework, we can calculate the uncertainty that is associated with the predictions. Although Sollich (2001) also viewed SVMs from a Bayesian perspective, his approach did not include priors for the hyperparameters and also did not accommodate any potential error in the model specification.

One of the advantages of SVMs is that their performance does not deteriorate with high input dimension. When the number of parameters exceeds the number of observations, in contrast with other machine learning methods, SVMs do not require an additional projection to the sample space, and then the application of a classification algorithm; the dimension reduction is built automatically into SVM methodology. Preliminary selection of highly informative genes can reduce the noise in the data and thus improve the predictive misclassification rate, with tighter bounds.

Use of the probit model with the introduction of latent variables (Albert and Chib, 1993) rather than a logistic model accelerates the computation significantly. We tried all the examples with the probit model and the results are almost identical to the logistic model results.

Acknowledgements

The first author's research was partially supported by National Science Foundation grant DMS-0203215 and National Cancer Institute grant CA-57030. The second author's research was partially supported by Munn idea grant, a prostate 'Specialized programs of research excellence' seed grant from the University of Michigan and grant 1R01GM72007-01 from the National Science Foundation and National Institute of General Medical Sciences. The third author's research was partially supported by National Institutes of Health grant R01-85414. The authors are indebted to Randy Eubank for many constructive suggestions that led to a much improved exposition. The authors are grateful to Emanuel Parzen and Raymond Carroll for helpful conversations. They also gratefully acknowledge the Joint Editor, the Associate Editor and two referees for their constructive suggestions that led to significant improvements in the presentation of the paper.

References

Albert, J. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

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

Aronszajn, N. (1950) Theory of reproducing kernels. *Trans. Am. Math. Soc.*, **68**, 337–404.

Bernardo, J. M. (1979) Expected information as expected utility. *Ann. Statist.*, **7**, 686–690.

Bishop, C. and Tipping, M. (2000) Variational relevance vector machines. In *Proc. 16th Conf. Uncertainty and Artificial Intelligence* (eds C. Boutilier and M. Goldszmidt), pp. 46–53. San Francisco: Morgan Kauffman.
Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S. and Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natn. Acad. Sci. USA*, **97**, 262–267.
Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines*. Cambridge: Cambridge University Press.
Denison, D., Holmes, C., Mallick, B. and Smith, A. F. M. (2002) *Bayesian Methods for Nonlinear Classification and Regression*. London: Wiley.
DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–685.
Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Statist. Ass.*, **97**, 77–87.
Figueiredo, M. (2002) Adaptive sparseness using Jeffreys prior. In *Neural Information Processing Systems 14* (eds T. Dietterich, S. Becker and Z. Ghahramani), pp. 697–704. Cambridge: MIT Press.
Fix, E. and Hodges, J. L. (1951) Discriminatory analysis-nonparametric discrimination: consistency properties. US Air Force School of Aviation Medicine, Randolf Field.
Gelfand, A. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–158. London: Chapman and Hall.
Gelfand, A. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.
Gelman, A. (1996) Inference and monitoring convergence. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 131–140. London: Chapman and Hall.
Golub, T. R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. and Trent, J. (2001) Gene expression profiles in hereditary breast cancer. *New Engl. J. Med.*, **344**, 539–548.
Herbrich, R. (2002) *Learning Kernel Classifiers*. Cambridge: MIT Press.
Holmes, C. and Held, L. (2003) Bayesian auxiliary variable models for binary and polychotomous regression. *Technical Report*. Imperial College, London.
Kimeldorf, G. and Wahba, G. (1971) Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, **33**, 82–95.
Lee, K. E., Sha, N., Dougherty, E., Vannucci, M. and Mallick, B. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.
MacKay, D. (1996) Bayesian non-linear modelling for the 1993 energy prediction competition. In *Maximum Entropy and Bayesian Methods* (ed. G. Heidbreder), pp. 221–234. Dordrecht: Kluwer.
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Moler, E. J., Chow, M. L. and Mian, I. S. (2000) Analysis of molecular profile data using generative and discriminative methods. *Physiol. Genet.*, **4**, 109–126.
Neal, R. (1996) *Bayesian Learning for Neural Networks*. New York: Springer.
Parzen, E. (1970) Statistical inference on time series by rkhs methods. In *Proc. 12th Bienn. Sem.* (ed. R. Pyke), pp. 1–37. Montreal: Canadian Mathematical Congress.
Pontil, M., Evgeniou, T. and Poggio, T. (2000) Regularization networks and support vector machines. *Adv. Comput. Math.*, **13**, 1–50.
Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
Rosenblatt F. (1962) *Principles of Neurodynamics*. New York: Spartan.
Roth, V. (2002) The generalized LASSO: a wrapper approach to gene selection for microarray data. Department of Computer Science, University of Bonn, Bonn.
Schena, M., Shalon, D., Davis, R. and Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
Schölkopf, B. and Smola, A. (2002) *Learning with Kernels*. Cambridge: MIT Press.
Sollich, P. (2001) Bayesian methods for support vector machines: evidence and predictive class probabilities. *Mach. Learn.*, **46**, 21–52.
Specht, D. F. (1990) Probabilistic neural networks. *Neur. Netwrks*, **3**, 109–118.
Tipping, M. (2000) The relevance vector machine. In *Neural Information Processing Systems 12* (eds S. Solla, T. Leen and K. Muller), pp. 652–658. Cambridge: MIT Press.
Tipping, M. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.
Vapnik, V. N. (2000) *The Nature of Statistical Learning Theory*, 2nd edn. New York: Springer.
Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wahba, G. (1999) Support vector machines, reproducing kernel hilbert spaces and the randomized GACV. In *Advances in Kernel Methods* (eds B. Schölkopf, C. Burges and A. Smola), pp. 69–88. Cambridge: MIT Press.

Wahba, G., Lin, Y., Lee, Y. and Zhang, H. (2002) Optimal properties and adaptive tuning of standard and non-standard support vector machines. In *Nonlinear Estimation and Classification* (eds D. Denison, M. Hansen, C. Holmes, B. Mallick and B. Yu), pp. 125–143. New York: Springer.

West, M. (2003) Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics 7* (eds J. M. Bernardo, M. Bayarri, A. P. Dawid, J. Berger, D. Heckerman, A. F. M. Smith and M. West), pp. 723–732. Oxford: Oxford University Press.

Williams, C. and Barber, D. (1998) Bayesian classification with Gaussian priors. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1342–1351.

Xiong, M., Jin, L., Li, W. and Boerwinkle, E. (2000) Computational methods for gene expression-based tumor classification. *Biotechniques*, **29**, 1264–1270.

Zhu, J. and Hastie, T. (2002) Kernel logistic regression and the import vector machine. In *Neural Information Processing Systems 14* (eds T. Dietterich, S. Becker and Z. Ghahramani), pp. 1081–1088. Cambridge: MIT Press.