Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1271–1282
Pattern Recognition Letters
www.elsevier.com/locate/patrec

On Bayesian classification with Laplace priors

Ata Kabán

School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK

Received 27 February 2006; received in revised form 20 November 2006
Available online 28 February 2007

Communicated by M. Singh

Abstract

We present a new classification approach, using a variational Bayesian estimation of probit regression with Laplace priors. Laplace priors have been previously used extensively as a sparsity-inducing mechanism to perform feature selection simultaneously with classification or regression. However, contrarily to the 'myth' of sparse Bayesian learning with Laplace priors, we find that the sparsity effect is due to a property of the maximum a posteriori (MAP) parameter estimates only. The Bayesian estimates, in turn, induce a posterior weighting rather than a hard selection of features, and has different advantageous properties: (1) It provides better estimates of the prediction uncertainty; (2) it is able to retain correlated features favouring generalisation; (3) it is more stable with respect to the hyperparameter choice and (4) it produces a weight-based ranking of the features, suited for interpretation. We analyse the behaviour of the Bayesian estimate in comparison with its MAP counterpart, as well as other related models, (a) through a graphical interpretation of the associated shrinkage and (b) by controlled numerical simulations in a range of testing conditions. The results pinpoint the situations when the advantages of Bayesian estimates are feasible to exploit. Finally, we demonstrate the working of our method in a gene expression classification task.

Keywords: Laplace prior; Variational Bayes; Sparsity; Shrinkage effect; Predictive features; Microarray gene expressions

1. Introduction

The Laplacian density has been widely used as a sparsity-inducing prior in various contexts (Shevade and Keerthy, 2003; Figueiredo, 2003; Cawley and Talbot, 2006; Moulin and Liu, 1999). However, since all these works employ the MAP estimates (even if referred to as Bayesian in a 'broad' sense), it should also be interesting to inspect the full posterior and examine the working of a Bayesian estimate.

In (Ju et al., 2002), a Gibbs sampling approach has been developed, to provide Bayesian estimates. However, the results were contradictory and largely inconclusive. Indeed, the computationally demanding sampling procedure has not been particularly suitable to extensive experimental studies. Another recent attempt (Park, unpublished), using Gibbs sampling, presents some results for the overdetermined linear regression case only.

Apart from these two unpublished works, we know of no further studies and no conclusive analyses to understand the seemingly pronounced difference in the behaviour of MAP and Bayesian estimates for this model. This is what we address in this paper, with a focus on under-determined classification problems. We develop a practical variational Bayesian algorithm which allows us to conduct a comprehensive experimental validation.

The remainder of the paper is organised as follows. After reviewing the model of probit regression with Laplace priors, Section 2 gives insights into the exact Bayesian analysis of this model in the univariate case and develops a practical variational Bayesian estimation algorithm for the multivariate case. Section 3 shows that 'sparse Bayesian learning with Laplace priors' should really be termed as 'sparse MAP learning with Laplace priors' and

E-mail address: A.Kaban@cs.bham.ac.uk

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.02.010

explains how the sparsity effect is actually due to a property of maximum a posteriori (MAP) parameter estimates. Section 4 completes the hyperparameter inference in the variational Bayesian framework. Section 5 discusses relationships with other models through a graphical interpretation of the associated shrinkage. Section 6 presents extensive experimental demonstration analysing the behaviour of the Bayesian estimate in comparison with its MAP counterpart, as well as other related models. The Bayesian estimate does not provide sparse estimates. In turn, (1) it provides better estimates of the prediction uncertainty; (2) it is able to retain correlated features favouring generalisation; (3) it is more stable with respect to the hyperparameter choice and (4) it produces a weight-based ranking of the features, suited for interpretation. Finally, Section 7 concludes the paper.

2. Probit regression with Laplace priors

Consider a training set of the form $\{(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_N, z_N)\}$, where $\boldsymbol{x}_n$ are T-dimensional input points, N is the number of observations and z_n are labels in $\{-1, 1\}$. We will refer to all input points as $\boldsymbol{x} \in \mathscr{R}^{N\times T}$ and associated labels as $\boldsymbol{z} \in \mathscr{R}^{N\times 1}$. In two-class classification, the task is to learn a mapping from the inputs to the targets; which is able to predict the target values of previously unseen points that follow the same distribution as the training data. In probabilistic terms, such a mapping is specified by a likelihood model together with prior distributions on the parameters of the likelihood.

The simplest form of likelihood model is the linear regression likelihood, where the target values are continuous valued. Denoting by $\boldsymbol{y} = (y_1, \ldots y_N)^{\mathrm{T}} \in \mathscr{R}^{N\times 1}$ the continuous targets, this is the following:

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w} \sim \mathscr{N}(\boldsymbol{y}|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \tag{1}$$

Here, $\mathscr{N}()$ denotes the normal density with σ^2 being the variance parameter. Further, $\boldsymbol{w} \in \mathscr{R}^{\mathrm{T}}$ is the parameter of the likelihood model and $\boldsymbol{I} \in \mathscr{R}^{N\times N}$ is the identity matrix. The dot product $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$ also includes a bias term, which can be handled by concatenating a feature of ones to $\boldsymbol{x}$, provided that care is taken that the bias term needs not be regularised.

A technically convenient way to obtain a likelihood model suitable to classification is to employ the probit link. It is common to fix $\sigma = 1$ (Figueiredo, 2003), and the variables $\boldsymbol{y}$ are now seen as latent variables of the additional probit likelihood:

$$z_n|y_n \sim P(z_n = 1|y_n) = \Phi(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}) \quad \forall n = 1, \ldots, N \tag{2}$$

where z_n are binary class labels in $\{-1, 1\}$ and $\Phi(y) = \int_{-\infty}^{y} \mathscr{N}(u|0, 1)\,\mathrm{d}u$ is the cumulative density function of a standard Gaussian density.

The prior on $\boldsymbol{w}$ is chosen to be a Laplace density

$$\boldsymbol{w} \sim \frac{\sqrt{\lambda}}{2}\mathrm{e}^{-\sqrt{\lambda}|\boldsymbol{w}|}$$

The Laplace density is heavy tailed and peaked at zero, hence expressing the prior belief that the distribution of $\boldsymbol{w}$—equivalently, the distribution of feature relevances w.r.t. to the target is strongly peaked around zero. In addition, the Laplace density is log-convex, which conveniently ensures the convexity of the posterior density.

In regression, the use of the Laplacian prior is known as the LASSO (Tibshirani, 1996; Efron et al., 2004). A probabilistic re-interpretation was given in (Figueiredo, 2003), by rewriting the Laplace density in a hierarchical manner

$$w_t|\tau_t \sim \mathscr{N}(w_t|0, \tau_t) \tag{3}$$

$$\tau_t|\lambda \sim Ga(\tau_t|1, \lambda/2) = \frac{\lambda}{2}\mathrm{e}^{-\lambda\tau_t/2} \tag{4}$$

Here, the variances τ_t of the Gaussian are hidden variables and $Ga()$ is the gamma distribution (in the present instantiation, an exponential (Bernardo and Smith, 1994)). Integrating over τ_t recovers the Laplacian marginal prior density.

In (Figueiredo, 2003), an expectation maximisation (EM) procedure was derived for this model, which iteratively computes the maximum a posteriori (MAP) estimates of $\boldsymbol{w}$.

2.1. Estimation methods

In models having prior distributions, two main estimation techniques are available to use.

Bayesian estimation makes no difference between parameters and latent random variables. This presents several advantages, such as finding a full distribution of the parameters, avoiding overfitting and assessing model order selection in the Bayesian framework (Bernardo and Smith, 1994; MacKay, 2003). After observing the data, the posterior density is computed with the use of Bayes theorem, e.g. in case of linear regression this is the following:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{\int \mathrm{d}\boldsymbol{w} p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}$$

Often the integral in the denominator (also known as the marginal likelihood or the evidence) is not analytically computable and approximations must be employed.

The *Maximum a posteriori (MAP) estimation* method avoids the intractable integral of the Bayesian approach, by computing the mode of the posterior only, and using that as a point estimate.

$$\operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}|\boldsymbol{y}) = \operatorname*{argmax}_{\boldsymbol{w}} p(\boldsymbol{w}, \boldsymbol{y})$$

The mode can be computed without computing the full posterior distribution. It is however well known that the price is a tradeoff of some information loss against computational efficiency, and in general the MAP method is also known to be more prone to overfitting in small sample size problems (MacKay, 2003).

We now turn to developing a Bayesian analysis of the probit regression model with Laplace priors.

2.2. Bayesian analysis: insights

Observe that in the 1D case, the convolution of a Gaussian likelihood with a Laplace prior is analytically computable (by elementary integration or using symbolic computation packages). Fig. 1 depicts the posterior mean and one posterior standard deviation on both sides, against fixed equidistant values of $y_{1,\ldots,N}$ in the range between -3 and 3, and having set all $x_{1:N}$ to 1. Equivalently, the same plot can be regarded as the multivariate posterior mean vector of $\boldsymbol{w}$ against the 'true' Maximum Likelihood vector value of $\boldsymbol{w}$, when the input set $\boldsymbol{x}_{1:N}$ is fixed to the identity matrix.

Fig. 1 reveals the rather interesting insight that the Laplace prior induces a nonlinear shrinkage (feature weighting) effect rather than a hard thresholding (feature selection). The smaller the values w_t, the larger the shrinkage incurred, however, in no interval is w mapped to a value of zero. This is in contrast with the known sparsity promoting property of the MAP estimator and this will be discussed in some detail later. We can also observe from Fig. 1, that the posterior variances are such that the Bayesian credible interval shrinks gradually for small values of w_t, although it always remains non-zero. In other words, small values are 'less important' but are not discarded.

It is now of interest to know what are the practical implications of this posterior, and whether the Bayesian estimate would be preferable to the MAP estimate. Before we can conduct empirical studies on multivariate data to address these questions, we need to derive a practical algorithm for obtaining Bayesian estimates. Since apart from the univariate case exact inference is no longer possible, in the next section we derive a variational solution.

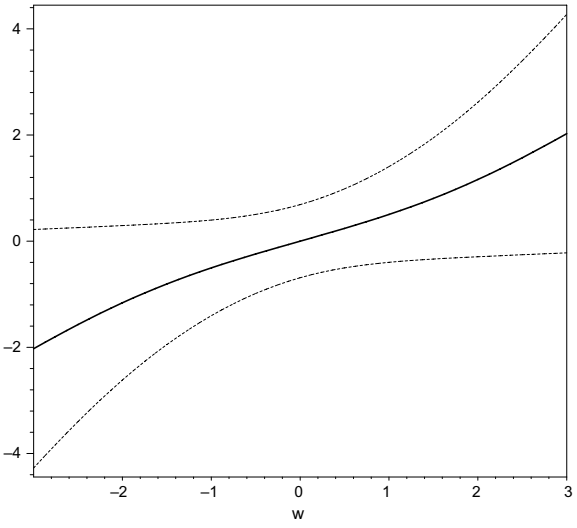

Fig. 1. Posterior mean and one standard deviation on both sides (vertical axis), against the maximum likelihood values of $\boldsymbol{w}$ (horizontal axis), when $\boldsymbol{x}$ is set to the identity matrix and $\lambda = 1$.

2.3. A variational Bayesian solution

Employing Jensen's inequality, it is straightforward to lower bound the log probability of $\boldsymbol{y}|\boldsymbol{x}$:

$$\log p(\boldsymbol{y}|\boldsymbol{x}) = \log \int \mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\tau} p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x}) p(\boldsymbol{w}|\boldsymbol{\tau}) \prod_t p(\tau_t)$$
$$\geqslant \int \mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\tau} q(\boldsymbol{w},\boldsymbol{\tau}) \log \frac{p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x}) p(\boldsymbol{w}|\boldsymbol{\tau}) \prod_t p(\tau_t)}{q(\boldsymbol{w},\boldsymbol{\tau})}$$

where $q(\boldsymbol{w},\boldsymbol{\tau}) = q(\boldsymbol{w})\prod_t q(\tau_t)$ is the variational posterior sought in a factorial form (Bernardo and Smith, 1994; MacKay, 2003; Bishop et al., 2000). This decouples into computing the approximate posteriors for each parameter separately, as follows. Denoting by $\boldsymbol{\Lambda}$ the diagonal matrix with elements $1/\tau_t$, we have

$$q(\boldsymbol{w}) \propto \exp \int \mathrm{d}\boldsymbol{\tau} q(\boldsymbol{\tau}) \log \mathcal{N}(\boldsymbol{y}|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}, \boldsymbol{I}/\sigma^2) \mathcal{N}(\boldsymbol{w}|0,\boldsymbol{\Lambda})$$
$$= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$$

where

$$\boldsymbol{\mu}_{\boldsymbol{w}} = \sigma^{-2}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{x}\boldsymbol{y} = \langle \boldsymbol{w} \rangle \tag{5}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{w}} = \{\langle \boldsymbol{\Lambda} \rangle + \sigma^{-2}\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\}^{-1} = \langle \boldsymbol{w}\boldsymbol{w}^{\mathrm{T}} \rangle - \langle \boldsymbol{w} \rangle \langle \boldsymbol{w} \rangle^{\mathrm{T}} \tag{6}$$

and $\langle \cdot \rangle = E[\cdot]$ denotes the expectation operator, taken w.r.t. to the variational posterior. The matrix inversion is computed using the well-known Sherman–Morrison–Woodbury formula, and $\langle \boldsymbol{\Lambda} \rangle = \mathrm{diag}_t(\langle 1/\tau_t \rangle)$. Further, we have

$$q(\tau_t) \propto \exp \int \mathrm{d}w_t q(w_t) \log \mathcal{N}(w_t|0,\tau_t) Ga\left(\tau_t \middle| 1, \frac{\lambda}{2}\right)$$
$$\propto \mathcal{N}\left(\sqrt{\langle w_t^2 \rangle}|0,\tau_t\right) \exp(0.5\lambda\tau_t) \tag{7}$$

The normalisation constant of this density is

$$\int \mathrm{d}\tau_t \mathcal{N}\left(\sqrt{\langle w_t^2 \rangle}\middle|0,\tau_t\right) Ga\left(\tau_t \middle| 1, \frac{\lambda}{2}\right) = \frac{\sqrt{\lambda}}{2} \exp\left\{-\sqrt{\lambda\langle w_t^2 \rangle}\right\}$$

and the expectation required in (6) is computed as

$$\langle 1/\tau_t \rangle = \int \mathrm{d}\tau_t \frac{1}{\tau_t} q(\tau_t) = \sqrt{\frac{\lambda}{\langle w_t^2 \rangle}} \tag{8}$$

Formally, the difference between this algorithm and the algorithm derived in (Figueiredo, 2003) for the model (1) and (3)–(4) is in Eqs. (7) and (8), where the posterior variances $\langle w_t^2 \rangle$ appear (instead of $\langle w_t \rangle^2$ of Figueiredo (2003)). In other words, in (Figueiredo, 2003) $q(\boldsymbol{w})$ was assumed to be a delta function around the mode of the true posterior[1] which amounts to computing the MAP estimate of $\boldsymbol{w}$. Here, in turn, $q(\boldsymbol{w})$ is a full approximating Gaussian.

Further, for probit-classification, $\boldsymbol{y}$ is an intermediate latent variable, so in addition to the above, $\boldsymbol{y}$ needs to be

[1] Although the mode equals the mean in the case of Gaussian posteriors, the mode and mean of the posterior corresponding to the overall hierarchical (Laplace) prior differ.

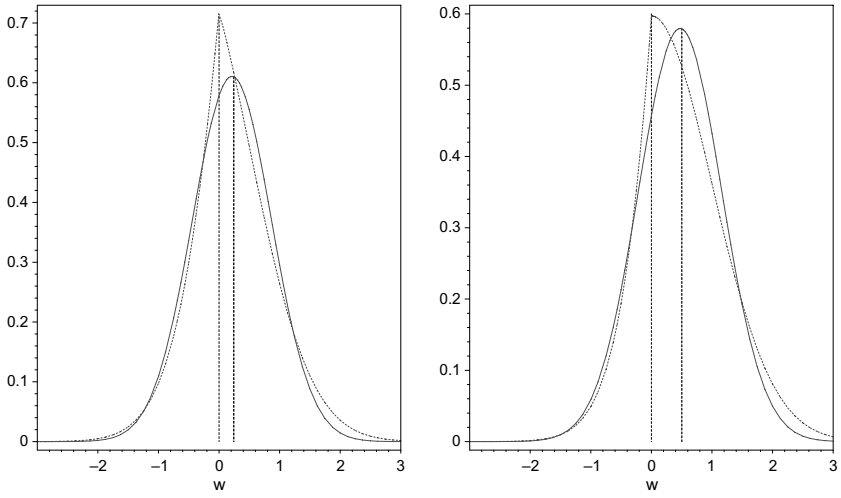

Fig. 2. True (dotted line) versus approximate (continuous line) posteriors in the univariate case, at $y = 0.5$ (left) and $y = 1$ (right). $\lambda = 1$ was fixed throughout.

integrated out from $q(\boldsymbol{w})$, and this amounts to replacing $\boldsymbol{y}$ in (5) by its expectation $\langle \boldsymbol{y} \rangle$ w.r.t. the additional variational posterior $q(\boldsymbol{y})$. The latter, as in (Figueiredo, 2003), is a product of truncated Gaussians, whose expectation is then

$$\langle \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \langle \boldsymbol{w} \rangle + z \frac{\mathcal{N}(\boldsymbol{x}^{\mathrm{T}} \langle \boldsymbol{w} \rangle | 0, 1)}{\Phi(z \boldsymbol{x}^{\mathrm{T}} \langle \boldsymbol{w} \rangle)} \tag{9}$$

2.4. Inspecting the posterior approximation

Using the feasible univariate case, it is of interest to inspect the true posterior induced by the Laplace prior versus the approximation $q(w)$. These are shown in Fig. 2 for two fixed values of y. The mean of the best approximating Gaussian $q(w)$ is indistinguishable from the mean of the true posterior, and the approximate posterior does capture the probability mass quite well. However, it is most apparent that on both of these plots the mass of the true posterior density differs from its mode and while the mode is at zero for both target values, the mean is non-zero. The next section shows that there is a measurable interval, where the mode of the true posterior stays at zero.

3. The maximum a posteriori method: whence the sparsity?

Sparse prediction machines have been quite popular, however, with few exceptions (Tibshirani, 1996; Efron et al., 2004), there has been little explanation on how and why do sparse solutions emerge. Based on earlier works in the area of signal denoising (Moulin and Liu, 1999; Goutte and Hansen, 1997), in this section we provide details showing how the sparsification occurs in the MAP estimates. For the ease of exposition, the formulation is given for regression. Similar reasoning applies for the probit likelihood.

For the analysis pursued here it is more convenient to work with the Laplace prior directly, rather than its hierarchical formulation. Assume that one component w_t is to be estimated while keeping the others fixed. The MAP solution is the maximum argument of the log probability of the complete data, i.e.

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmax}} - \sum_n \frac{(y_n - \boldsymbol{x}_n^{\mathrm{T}} \boldsymbol{w})^2}{2\sigma^2} - \sqrt{\lambda} |\boldsymbol{w}| \tag{10}$$

This is not differentiable at zero, therefore the unique maximum may be achieved either at zero or at a stationary point on the strictly negative or strictly positive domain.

Consider the positive case first, $w_t > 0$. Taking derivatives on the positive domain and equating to zero, we obtain

$$\begin{aligned} 0 &= \frac{\partial \log p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{x})}{\partial w_t} \\ &= \sum_n \left(y_n - \sum_{t' \neq t} w_{t'} x_{t'n} - w_t x_{tn} \right) x_{tn} / \sigma^2 - \sqrt{\lambda} \\ &= \sum_n \left(y_n - \sum_{t' \neq t} w_{t'} x_{t'n} \right) x_{tn} / \sigma^2 - w_t \sum_n x_{tn}^2 / \sigma^2 - \sqrt{\lambda} \\ w_t &= \frac{\sum_n (y_n - \sum_{t' \neq t} w_{t'} x_{t'n}) x_{tn}}{\sum_n x_{tn}^2} - \sigma^2 \frac{\sqrt{\lambda}}{\sum_n x_{tn}^2} \end{aligned}$$

Note the first term above is in fact the ML solution of w_t, which will be denoted by w_t^{ML}. Note further that, since we assumed $w_t > 0$, the above equation admits a solution only if $w_t^{\mathrm{ML}} > \frac{\sqrt{\lambda}\sigma^2}{\sum_n x_{tn}^2}$.

Analogously, on the strictly negative domain, $w_t < 0$, we get that a solution exists only if $w_t^{\mathrm{ML}} < -\frac{\sqrt{\lambda}\sigma^2}{\sum_n x_{tn}^2}$. Thus, in all other cases (i.e. when the ML solution lies between these two thresholds), the solution must be exactly zero. Indeed, Moulin and Liu (1999) proves in a denoising context that for any log-prior that is non-differentiable at zero, there is a non-zero neighbourhood around zero where the MAP solution is zero. In summary, for the unique maximum argument w_t^* we have that

$$w_t^* = \begin{cases} w_t^{\mathrm{ML}} - \epsilon_t \mathrm{sign}(w_t^{\mathrm{ML}}), & \text{when } |w_t^{\mathrm{ML}}| > \epsilon_t \\ 0, & \text{when } |w_t^{\mathrm{ML}}| \leqslant \epsilon_t \end{cases} \quad (11)$$

where $\epsilon_t = \frac{\sqrt{\lambda}\sigma^2}{\sum_n x_{tn}^2}$. The obtained threshold levels are quite intuitive: The level of threshold is inversely proportional to the sample variance and the number of training examples and directly proportional to the noise variance. The lower the sample variance of a feature, the more likely it is negligible. Also, a smaller sample size, and a larger observation noise implies a higher overall threshold level. In addition, λ may be varied to further control the sparsity level.

4. Hyperparameter inference

Irrespective of the estimation method used, the hyperparameter λ controls the strength of shrinkage globally. Cross-validation over a grid of values is often the method of choice, for the MAP-estimated model or in the case of non-probabilistic formulations (Tibshirani, 1996; Shevade and Keerthy, 2003). The Bayesian framework also offers more efficient alternatives, such as the maximum likelihood II (MLII), known also as the evidence maximisation procedure, or the specification of a hyper-prior on λ. We follow the latter option here. Since problems have been noted with vague (non-informative) hyper-priors or a MLII procedure when the sample size is too small (Qi et al., 2004; Park, unpublished), we specify a Gamma hyper-prior

$$\lambda \sim Ga(\lambda|\alpha, \beta) \quad (12)$$

with $\alpha = \beta = 1$ far enough from zero to avoid problems of vague priors in small sample size conditions. In other words, λ can vary cf. an exponential hyper-prior with mean of 1.

Now, the algorithm derived in Section 2.3 needs to be extended to accommodate λ as a random variable, by computing the variational posterior $q(\lambda)$ and the expectation $\langle\lambda\rangle$, and then replacing λ by $\langle\lambda\rangle$ throughout in Section 2.3. These additions are the following.

$$q(\lambda) \propto \exp \int \mathrm{d}\tau_t q(\tau_t) \sum_t \log\{p(\tau_t|\lambda)p(\lambda|\alpha,\beta)\}$$
$$= Ga\left(\lambda|\alpha + T, \beta + \frac{1}{2}\sum_t \langle\tau_t\rangle\right)$$

which yields

$$\langle\lambda\rangle = \frac{2(\alpha+T)}{2\beta + \sum_t \langle\tau_t\rangle} \quad (13)$$

Further, the posterior expectation additionally required for computing (13) is evaluated as

$$\langle\tau_t\rangle = \int \mathrm{d}\tau_t \tau_t q(\tau_t) = \frac{1 + \sqrt{\langle w_t^2\rangle\langle\lambda\rangle}}{\langle\lambda\rangle} = \frac{1}{\langle\lambda\rangle} + \sqrt{\frac{\langle w_t^2\rangle}{\langle\lambda\rangle}}$$
$$= \frac{1}{\langle\lambda\rangle} + \langle 1/\tau_t\rangle^{-1} \quad (14)$$

The algorithm is then to iterate the updating of all required posterior statistics until convergence.

4.1. An extension: feature-specific hyperparameters

Similarly to generalised LASSO (Roth, 2004), we may also consider separate λ parameters for each feature and place independent Gamma priors on each.

$$\lambda_t \sim Ga(\lambda_t|\alpha, \beta) \quad (15)$$

Of course, it should be noted that in this case the overall prior on w_t is no longer a Laplacian, but a convolution of a Laplacian with a Gamma. This is no longer log-convex and may be expected to behave similarly to a Student t prior.

Then we have

$$q(\lambda_t) = Ga\left(\lambda_t|\alpha + 1, \beta + \frac{1}{2}\langle\tau_t\rangle\right)$$

and so,

$$\langle\lambda_t\rangle = \frac{2(\alpha+1)}{2\beta + \langle\tau_t\rangle}; \quad \langle\tau_t\rangle = \frac{1}{\langle\lambda_t\rangle} + \langle 1/\tau_t\rangle^{-1}$$

It is outside the scope of this paper to study this extension in great detail but we find it useful for making a connection to the RVM (Tipping, 2001; Bishop et al., 2000; Li et al., 2002). As will be seen in the experiments shortly, if $\alpha = \beta = 10^{-6}$ (non-informative priors) are employed in (15), then a very similar behaviour to that of the RVM is obtained.

5. Related methods

To eliminate the parameter λ, Figueiredo (2003) proposes to use Jeffreys improper prior for the variables τ_t. In our experiments with high-dimensional and scarce data sets (detailed in a later section) we found this leads to an exaggerated sparsification, which did not turn out to be beneficial.

A slightly different but related model, known as the Relevance Vector Machine (RVM) (Tipping, 2001) is based on the notion of automatic relevance determination (Neal, 1996). Essentially, the prior adopted in RVM is an independent Student density and this is seen as an approximation to the Laplace (Bishop et al., 2000).

An obvious difficulty with attempting to study the relationship between various methods on the modelling level is that it would be difficult to disentangle the effects of the prior specification from those of the employed posterior approximations. To get round of this problem, we will characterise the joint effects of these two factors together, on the algorithmic level. Analogously to the so-called shrinkage functions, frequently used in the statistical regression literature to illustrate the behaviour of various methods, we obtain and visualise the posterior shrinkage (estimated posterior mean and posterior variance against the true values of $\boldsymbol{w}$) induced by the various algorithms

considered in this paper. This provides a graphical interpretation that is easy and intuitive to follow.

The posterior estimates of $\boldsymbol{w}$ against the true values, as obtained from simulations, when $\boldsymbol{x} = \boldsymbol{I}$ is fixed, are comparatively shown in Fig. 3, for several methods and parameter settings. The solid lines depict the posterior means obtained from the variational solution with Laplace prior (as derived in Section 2.3). The larger the λ value is, the greater the shrinkage effect. However, for any fixed λ, the shape of the posterior mean shrinkage is smoothly nonlinear, so that no components are completely discarded. Comparatively, the dash-dot curves show the MAP solutions. We see that for any fixed λ value, there is an interval around the origin, where the $\boldsymbol{w}$ component gets mapped to exactly zero. On the same plot, for comparison, the dotted line shows the estimates of $\boldsymbol{w}$ obtained with Jeffreys prior as proposed in (Figueiredo, 2003). This generates a rather large interval where weights are mapped to zero and there is no parameter to adjust the length of this interval. Finally the posterior mean estimates obtained from RVM (Tipping, 2001) are also plotted, and as expected, the posterior mean shrinkage obtained with variational Bayes estimation (Bishop et al., 2000) and MLII estimation (Tipping, 2001) are practically indistinguishable for the RVM. The input-specific modelling discussed in Section 4.1 produces identically looking estimates too. This indicates that both variants of RVM, as well as the method in Section 4.1 may be expected to behave very similarly and they all exhibit a severe feature down-weighting effect, rather close to thresholding.

A more refined intuition is provided by additionally plotting the posterior variances, and these plots are shown in Fig. 4. Again the two RVM variants and the component-wise Laplace + Gamma are indistinguishable, both in terms of posterior means and posterior variances. The posterior variances shrink dramatically within an interval around the origin, which means that some of the features

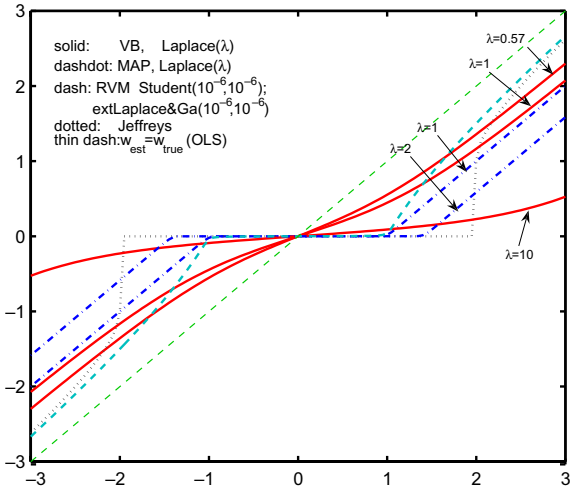

Fig. 3. Comparative plot of the estimated $\boldsymbol{w}$ against 'true' $\boldsymbol{w}_{\mathrm{ML}}$ values, when $\boldsymbol{x}$ is set to identity matrix.

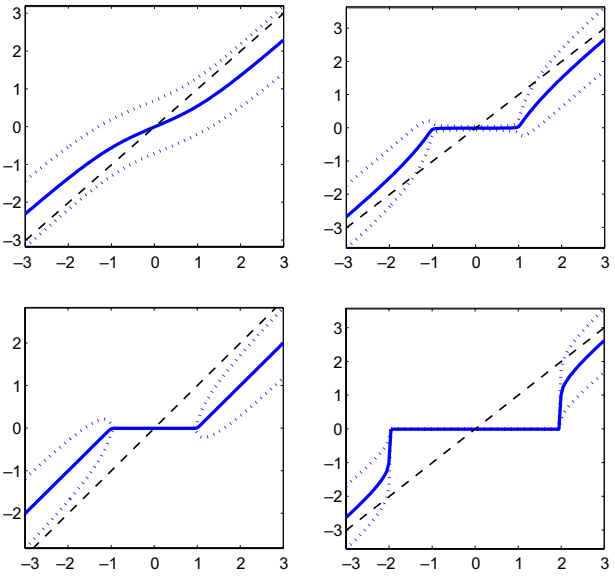

Fig. 4. Posterior mean ± one posterior standard deviation of $\boldsymbol{w}$ plotted against $\boldsymbol{w}_{\mathrm{ML}}$ when $\boldsymbol{x}$ is fixed to identity matrix and $\lambda = 1$. Top left: Laplace prior, VB solution (cf. Sec 2.3); Bottom left: Laplace prior, MAP solution (cf. Figueiredo, 2003); Top right: RVM (Tipping, 2001; Bishop et al., 2000); Bottom right: Gauss and Jeffreys hierarchical prior with MAP solution (cf. Figueiredo, 2003). For the latter two, the posterior variances do not play any role in the model estimation but were computed afterwards, for the purpose of displaying. The dashed line represents $w_{\mathrm{estimate}} = w_{\mathrm{true}}$ (the OLS solution).

will be severely down-weighted. The variational estimate of the log-convex model (left top corner), however, exhibits a 'mild' feature down-weighting (shrinkage) effect compared to all other methods. The posterior mean is indistinguishable from the true one (Fig. 1), the posterior variance being somewhat affected by the variational approximation. Still, it should be stressed there is a genuine difference from a ridge or Bayesian ridge regression model, which is immediately visible from the nonlinear shape of the shrinkage, and also recalling that τ_t are input feature specific.

6. Class prediction

6.1. Predictive distributions

Having estimated the model, let us denote a test input point by $\boldsymbol{x}^*$. The predictive distribution of the target for this test point, y^* is computed as the following:

$$p(y^*|\boldsymbol{x}^*) = \int \mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\tau}s\lambda p(y^*|\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{w})q(\boldsymbol{w})q(\boldsymbol{\tau})q(\lambda)$$

$$= \int \mathrm{d}\boldsymbol{\tau}\mathrm{d}\lambda \mathcal{N}(y^*|\boldsymbol{x}^{*\mathrm{T}}\boldsymbol{\mu}_{\boldsymbol{w}}, \boldsymbol{x}^*\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{x}^{*\mathrm{T}} + \sigma^2)q(\boldsymbol{\tau})q(\lambda)$$

where $\boldsymbol{\mu}_{\boldsymbol{w}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ are conditioned on the higher level variables. The commonly used approximation to the above integral is adopted in the reported experiments:

$$p(y^*|\boldsymbol{x}^*) \approx \mathcal{N}(\mu_{y^*}, v_{y^*}) \tag{16}$$

where

$$\mu_{y^*} = \boldsymbol{x}^* \sigma^{-2} \{ \langle \boldsymbol{\Lambda} \rangle + \sigma^{-2} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x} \}^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y} \tag{17}$$

$$v_{y^*} = \boldsymbol{x}^* (\langle \boldsymbol{\Lambda} \rangle + \sigma^{-2} \boldsymbol{x} \boldsymbol{x}^{\mathrm{T}})^{-1} \boldsymbol{x}^{*\mathrm{T}} + \sigma^2 \boldsymbol{I} \tag{18}$$

Further for probit regression, we have

$$P(z = 1 | \boldsymbol{x}^*) = \int \mathrm{d}y^* P(z = 1 | y^*) p(y^* | \boldsymbol{x}^*) = \Phi\left(\frac{\mu_{y^*}}{\sqrt{v_{y^*}}} \right) \tag{19}$$

6.2. Numerical simulations

Here we empirically study the behaviour and the prediction performance of the methods in the under-determined case, i.e. when the number of features exceeds the number of examples. Such settings are of interest from the perspective of gene expression classification applications, since the number of measured genes typically exceeds the number of samples. In addition, it is the under-determined case when the prior distribution may be expected to have a substantial influence on the estimate.

This study was partly motivated by a recent controversy regarding the relative fraction of genes that are in a causal relationship with a certain medical condition. While it has been commonly believed that a large fraction of genes are not needed for predicting the target condition, recent studies (Zou and Hastie, 2005; Qi et al., 2004) challenge this assumption. Although obtained on different grounds, these works seem to indicate that an intensive use of feature selection or feature thresholding may not necessarily be the optimal choice in terms of diagnosis prediction performance and this finding motivates further research.

6.2.1. I.i.d. features

The first set of experiments is concerned with synthetic data having 200-dimensional i.i.d. features. The training and test sets were generated using $\boldsymbol{w} = [3, \ldots, 3, 0, \ldots, 0]^{\mathrm{T}}$, where the number of non-zero components (equivalently, the number of relevant features) was varied in the set $\{5, 10, 30, 50, 100\}$, to provide a range of testing conditions. A data matrix $\boldsymbol{x}$ was drawn from independent zero-mean and unit variance Gaussians, and $\boldsymbol{y}$ drawn from a Gaussian with mean $\boldsymbol{X}\boldsymbol{w}$ and unit variance. These $\boldsymbol{y}$ were then thresholded at zero to provide the class labels $\boldsymbol{z}$. The training set size was varied in $\{30, 60\}$. For each training set, a test set of 3000 points was also generated from the same model as the associated training set.

Fig. 5 presents the test set classification errors averaged over 30 data sets for each of the experimental setting described above. For Laplace-MAP, we did an internal 5-fold cross-validation within the training phase in order to determine a suitable value for λ, from a grid of candidate vales. Since the square root of λ is required in the algorithms, the grid of values that we included in this search was $\{0.1, 0.5, 1, 4\}$. For Laplace-VB, we investigate two versions, one using a fixed value of $\lambda = 1$ throughout, and the other inferring λ from the data as described in Section 4. The method using Jeffreys prior (Figueiredo, 2003), the RVM (Tipping, 2001) and our component-wise Laplace + Gamma extension discussed in Section 4.1 were also included in the comparison. As expected, the latter two turned out to perform identically.

6.2.1.1. Misclassification results. A first observation from the comparative results is that for scarce data, the method with Jeffreys prior (Figueiredo, 2003), is inadequate. With insufficient training data, this method tends to switch all or nearly all weights to zero, leading to a random classification. Naturally, this would unlikely be the case if data was abundant and in addition, it is also possible that a different algorithmic implementation would lead to a different behaviour. However, because we are concerned with under-determined data settings, this finding warns for extra care.

It is also apparent from the plots that as expected, all methods improve with increasing the training set (from 30 to 60 points), and the rate of the improvements is typically greater in the case of the 'sparse' methods (Jeffreys, Laplace-MAP, RVM, Laplace and Gamma (extension)). This is because the influence of the data is greater than that of the prior specification for those methods.

Interestingly, for Laplace-VB with fixed $\lambda = 1$ and Laplace-VB with adaptively inferred λ, the misclassification

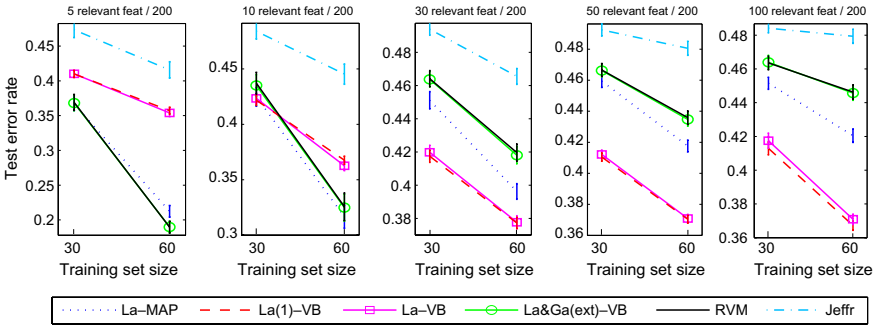

Fig. 5. Test classification errors in terms of mean and one standard error on both sides, computed over 30 generated data sets for each experiment. Each data set contained 200 i.i.d. features, of which the fraction of relevant features is varied in $\{5, 10, 30, 50, 100\}$, as indicated in the title line of each plot. For each of these four settings, the test results with a sample size of 30 and 60 points are shown on each plot. The test performance was measured on an independent test set of 3000 points, generated from the same distribution as the corresponding training set.

rates obtained are statistically equal. This suggests that Laplace-VB is quite stable against the specification of λ.

The comparison of the Laplace-MAP method against Laplace-VB is also interesting. As we know, Laplace priors have been previously used extensively as a sparsity-inducing mechanism and often referred to as 'sparse Bayesian learning'. However we have just seen that this really refers to sparse Bayesian MAP learning, since the sparsity effect is only a property of the MAP parameter estimates in models that employ a Laplace prior. The Bayesian estimate, in turn, induces a posterior weighting rather than a hard selection of features. The results confirm that this implies significant differences in the behaviour of VB versus MAP and we now seek to clarify these differences and characterise the situations in which one is preferable to use over the other.

Concerning misclassification error rates, it is clear from Fig. 5 that Laplace-MAP is superior when the fraction of relevant features is small, whereas Laplace-VB is superior when a moderate or larger fraction of features is relevant w.r.t. to the target. The RVM and Laplace-MAP performance is fairly similar, the latter having slightly more possibility to adapt to moderate levels of sparsity. In the light of the shrinkage effects shown in the previous subsection, these observations should not be really surprising, even though they might have been difficult to foresee otherwise. The hope is, of course, that the understanding gained from this study will facilitate the understanding of real data sets, and we may get a clue for the relative fraction of relevant features they may contain.

6.2.1.2. Brier score results. In certain applications, e.g. in a medical context, the uncertainty estimates are also important. It is not the same to what confidence a prediction is given. Therefore, beyond misclassification rates, it is enlightening to inspect error measures that incorporate the uncertainty information. The Brier score is such a measure, previously proposed for the evaluation of gene expression classification results (Yeung et al., 2005). This is simply the mean square error between the prediction (a number between 0 and 1) and the binary (0 or 1) target, i.e. $1/N\sum_n(z_n - p(z_n = 1|\boldsymbol{x}_n))^2$. The Brier scores for the same set of synthetic data experiments are shown in Fig. 6.

These results highlight an important point, namely a weakness incurred by 'sparse' approaches in not representing uncertainties appropriately. The VB method in turn is able to consistently improve over MAP in terms of Brier scores—in all those test cases in which the Laplace prior is an appropriate description of the distribution of feature relevances. Moreover, the benefit of the hyperparameter inference now becomes evident: The adaptive version does provide additional improvements in terms of predictive uncertainties and this is nicely seen in the obtained Brier scores.

These results also pinpoint the cases when the advantage of Laplace-VB in terms of providing better estimates of the prediction uncertainty is feasible to exploit, and just as importantly, when it is not. As we see, the latter concerns the cases when the relative fraction of irrelevant features is excessively high, so that a log-convex density is no longer a good enough description. This latter point is a negative finding (in the most positive sense) given the wide use of phrases like 'sparse Bayesian learning with Laplace priors'.

6.2.2. Correlated features

Most real-world data sets contain some redundancy in their feature sets. Before turning to real data sets, next we demonstrate experiments on simulated data having correlated features. Now we generated the data $\boldsymbol{X}$ as the concatenation of 10 groups of 5 correlated features each, as follows. A $10 \times N$ data set $\tilde{\boldsymbol{X}}$ was first generated from a i.i.d. standard Gaussian, then each group of 5 correlated features of $\boldsymbol{X}$ was derived from one of the features of $\tilde{\boldsymbol{X}}$ by adding Gaussian noise with zero-mean and variance of 0.1. One hundred and fifty noise features were also concatenated to $\boldsymbol{X}$, so we have 200-dimensional data of N samples. N was varied in $\{30, 60\}$, as before. The true generating $\boldsymbol{w}$ was then set to $(3, \ldots, 3, 0, \ldots, 0)^{\mathrm{T}}$ where the number of non-zero components is 50 and the number of zero components is 150. The target labels were obtained from thresholding $\mathcal{N}(\boldsymbol{X}\boldsymbol{w}, \boldsymbol{I})$ at zero, as before.

The results are shown in Fig. 7, summarised from 30 repeated experiments for both training set sizes. The left hand plot sows the misclassification rates. The Jeffrey prior approach is again quasi-random, RVM and Laplace-MAP perform equally, and Laplace-VB with both fixed hyperpa-

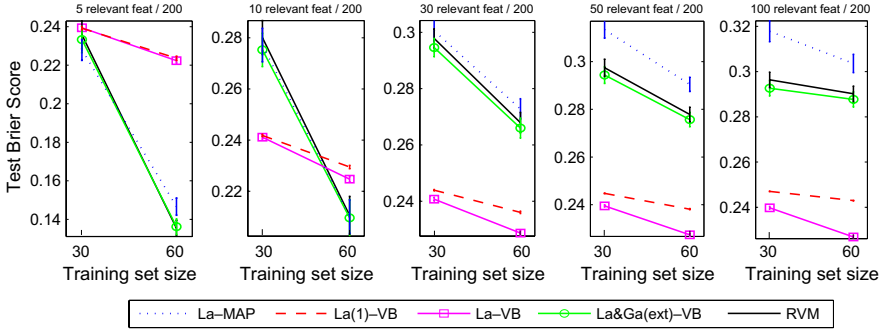

Fig. 6. Test Brier scores computed from the experiment described in the previous figure.

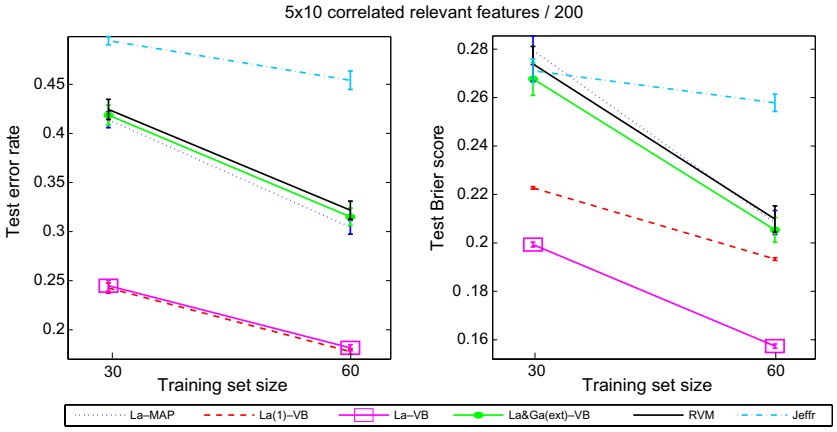

Fig. 7. Test classification errors (left) and brier scores (right) from experiments on 200-D data sets having 10 groups of 5 correlated features each, and 150 i.i.d noise features. The test performance was measured on an independent test set of 3000 points, generated from the same distribution as the corresponding training set. Each result represents the average and one standard error over 30 generated train + test sets from the same model.

rameter ($\lambda = 1$) and adaptively inferred hyperparameter produce significantly lower error rates. The two variants of Laplace-VB are again equal in terms of misclassification error rates.

On the right-hand plot of Fig. 7 we see the Brier scores for the same experiments. The picture is similar, again Laplace-VB is superior and the version with hyperparameter inference is the overall winner.

Observe also that the actual values of the errors in both measures are much lower than those obtained on the data with 50 i.i.d. features. The Laplace-VB model has an additional advantage over the sparse approaches in settings with correlated features. This stems from the fact that the latter tend to discard the redundant features and this is detrimental to their generalisation performance. This was also pointed out in (Qi et al., 2004). The following toy example will illustrate the issue.

Fig. 8 shows the decision boundaries obtained by the different methods under consideration, for a small 2D generated data set. The data set has two correlated features, therefore one of them is redundant. Apart from Laplace-VB, all sparsity-inducing methods discard one of these features and come up with a separation boundary based on a single feature only. However, clearly this is highly suboptimal from the generalisation point of view.

6.3. Colon cancer prediction

The colon data set (Alon et al., 1999) contains expression levels of 2000 genes from 40 tumour and 22 normal colon tissues. It is a widely studied benchmark data set for gene expression classification algorithms, and so it permits a comprehensive comparison with previous results.

We perform bootstrap repeats, randomly sampling 50 points for training and testing on the remaining 12 points. We have chosen this splitting proportion because it is the most frequently used by other authors on this data set (see e.g. Li et al., 2002; Qi et al., 2004; Shevade and Keerthy, 2003) so that more meaningful comparisons can be made.

In each repeat, we use an identical pre-processing to that employed in (Chu et al., 2005), which is as follows. Each gene (feature) of the normalised (zero-mean and unit variance) training data is tested with the Wilcoxon rank sum test, at significance level $p = 0.01$. This procedure is closely related to the Significance Analysis of Microarrays (Tusher et al., 2001). The genes found non-differentially expressed at this stage are then discarded both from training and testing. We performed 100 independent bootstrap repeats (train-test splits) with Laplace-MAP (each time determining the value of λ by internal 5-fold cross-validation) and 500 independent bootstrap repeats with the Laplace-VB method (using the hyperparameter inference described in Section 4). Table 1 summarises the results. We also experimented with the full 2000-genes data set, and results will be discussed later in this section.

Table 1 details the number of false positives, the number of false negatives, the accuracy (error rates) both as the number of miss-classifications and as a percentage, and in addition it gives the area under the ROC curve (Fawcett

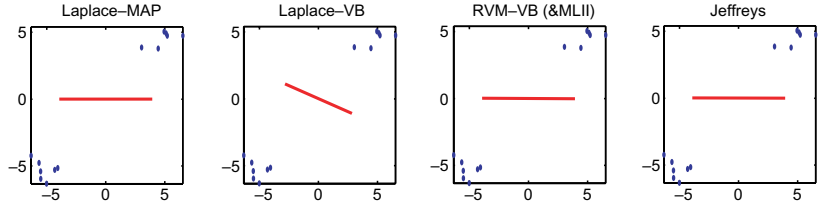

Fig. 8. The decision boundary as obtained with the different methods on a synthetic example.

Table 1
Classification results on the Colon data set: average and one standard error (in brackets), over bootstrap repeats randomly splitting the data into 50 training and 12 test points

PROBIT	La-MAP 100 rep	La(λ)-VB 500 rep
False +ves (#)	0.660 (0.074)	0.678 (0.034)
False −ves (#)	1.410 (0.108)	1.270 (0.042)
Error rate (#)	2.070 (0.122)	1.948 (0.050)
Error rate (%)	17.250 (1.014)	16.233 (0.414)
AUROC × 100	87.976 (1.059)	88.225 (0.482)

et al., 2004) (AUROC). The latter is a measure that takes both prediction uncertainties (in the case of probit-classification) and class imbalances (unequally represented classes) into account.

The results of the method employing Jeffreys prior (Figueiredo, 2003) is not given in the table, since it resulted in no better than a random classifier due to its tendency to turn off all features when the sample size is too small. RVM (with MLII) has been previously used in (Li et al., 2002), and their results are cited for comparison in the sequel.

6.3.1. Comparison with previous results for Colon: discussion

There are a number of results available from previous studies of the colon data set that we review here for comparison. The best 10-fold cross-validation result recently reported in (Chu et al., 2005) for the Colon data set (identically preprocessed) has been $16.19 \pm 13.65\%$ and selected 26 genes, using a variant of a non-convex model (probit Gaussian process with Gamma priors on the length scale parameters) estimated by MLII and combined with a sophisticated heuristic gene ranking scheme involving data resampling. The rest of the quoted results use identical experimental protocol as ours and utilise the full 2000 genes Colon data set. Sparse logistic regression with data resampling heuristic to aid stability was used in (Shevade and Keerthy, 2003) and produced 17.7% miss-classifications. Both of these are comparable to our result, despite we do not use any data resampling. The best results of Qi et al. (2004), using a technically involved method they call predictive-ARD-EP with logistic likelihood was 1.63 ± 0.11 miss-classifications, with 156.76 ± 11.86 genes in average. Another method they devised, called evidence-ARD-EP gave 2.54 ± 0.13 miss-classifications in average, selecting 7.92 ± 0.14 genes. Earlier results of Li et al. (2002) have been $2.90 \pm 0.13\%$ miss-classifications in average selecting 8.15 ± 0.13 genes, using RVM with MLII with some speed-up heuristics. Results obtained by support vector methods (Guyon et al., 2002) were 2.84 ± 0.14 misclassification with 4.25 ± 0.12 genes, and 2.68 ± 0.15 misclassification with 14.45 ± 5.35 genes, using SVM with recursive feature elimination and SVM with Fisher score respectively.

All these results are comparable to ours. The Laplace-VB procedure is just as simple to run as Laplace-MAP, it does not require data resampling heuristics and it also computes the hyperparameters without having to resort to expensive cross-validation.

In the light of our experiments and analysis, Laplace-VB is suitable when the feature importance is uneven and can also deal with correlated features. However, it is not suitable in cases when a too large fraction of features is completely irrelevant. We also did experiments on the full 2000 genes data set and found the accuracy of the variational procedure does drop (we obtained $\text{AUROC} \times 100 = 82.688 \pm 1.493$ from 100 bootstrap repeats) below that of the MAP-Laplace ($\text{AUROC} \times 100 = 87.753 \pm 1.436$) and the RVM. This suggests the number of genes that have an effect on the target must be significantly less than 2000 but larger than a handful few of the order of ten, in accordance with recent results (Qi et al., 2004).

It is yet an open problem to find a more flexible prior suitable to describing more realistic distributions of feature relevance. Nevertheless, it is clear that a thorough and systematic understanding of the behaviour of the methods is fundamental for making progress and this is what we primarily attempted in this paper. The synthetic examples were designed to showcase the differences in behaviour, which gives improved understanding of the difficulties faced in under-determined problems, the reasons behind results obtained on real data and can guide the ways of improving them or interpreting them. From a practical point of view, features with zero importance can easily filtered out at a pre-processing stage and we found this beneficial both in terms of accuracy and for saving computation time.

6.4. Interpretability and stability

Methods that select a small subset of the original features are often thought of as favouring interpretability. Although this is desirable for a biologist, the cautionary notes flashed out in (Diaz-Uriate, 2005) cannot be stressed enough and are once more highlighted in our results. Different technical approaches lead to different parameters and consequently different feature rankings, even when their predictive capabilities are comparable.

Here we use all 62 samples of the Colon data set to estimate $\boldsymbol{w}$ and we ask the question whether the magnitudes associated with the genes would be interpretable in the sense of an importance ranking. The ultimate answer to this question is down to the domain experts, however to gain insights into the meaningfulness of the larger number of genes brought up by Laplace-VB, we evaluate leave-one-out validation errors on nested subsets of genes by progressively adding genes according to $|\langle w_t \rangle|$. Note these numbers are validation errors only—they could be used e.g. to inform a post-processing procedure but not as an absolute indication of prediction performance, since $\boldsymbol{w}$ was estimated from all the data. Fig. 9 shows these results comparatively for the methods under consideration. We included several choices of λ in these plots because this gives us

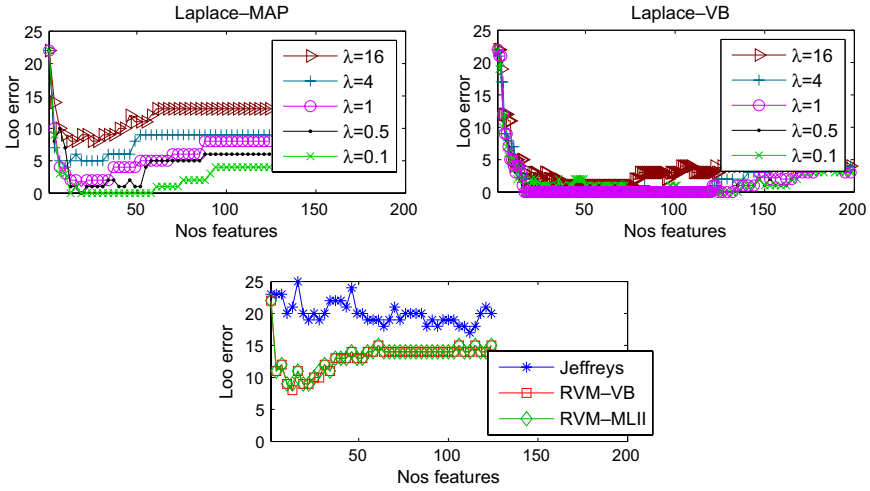

Fig. 9. Validation errors corresponding to subsets of genes ranked by the magnitudes of $|\langle \boldsymbol{w} \rangle|$.

another opportunity to show the extent of sensitivity to this choice. We see that Laplace-MAP is more sensitive to the misspecification of λ whereas Laplace-VB is fairly stable. This was also seen in previous experiments from the fact that the somewhat arbitrarily fixed value of $\lambda = 1$ and the inferred posterior expectation $\langle \lambda \rangle$ produced very similar misclassification results.

In terms of interpretability of the higher weighted genes, as measured by the leave-one-out validation errors, clearly, Laplace-MAP is also more sensitive to the exact number of genes included, even with the optimally selected λ. Laplace-VB in turn displays stability in this respect as well.

Finally, the last plot of Fig. 9 shows the l-o-o errors for the two RVM variants and the method with Jeffreys prior. We see that RVM-MLII and RVM-VB are equal, and are less stable than Laplace-VB and the parameter-free Jeffreys prior, is again quasi-random.

We also computed l-o-o validation errors in the same way on the full 2000 genes data set and interestingly, the results have been qualitatively similar. This suggests that further experiments may be conducted to investigate the incorporation of a post-processing stage into Laplace-VB, which may improve its performance when the fraction of irrelevant features is too high.

7. Conclusions

We presented a variational Bayesian analysis of probit regression with Laplace priors and investigated this model for high-dimensional (under-determined) classification, including its potential use in diagnosis prediction problems from microarray gene expressions. We explained how and why with the use of Laplace priors sparsity is induced in the MAP parameter estimates but not in the Bayesian estimates. We discussed and demonstrated the advantageous properties of Bayesian estimates, which include better uncertainty estimates, stability with respect to the hyperparameter choice and the ability to retain correlated features. Extensive numerical experiments and detailed analysis was provided to understand the somewhat complementary nature of the behaviour of Laplace-VB versus Laplace-MAP and other related methods for the first time and to pinpoint the situations when the expected advantages are exploitable in practice. From a practical point of view, a pre-filtering of irrelevant features was found to be beneficial to obtain a better match of the distribution of feature relevances with the model density and this also reduces subsequent computations.

There are a number of avenues for further research. The interplay between the practical desire for obtaining sparse solutions and manner of obtaining them as a byproduct of an approximate inference method due to model intractability is a rather interesting issue, which should deserve further study. One potentially fruitful direction would be to investigate a more detailed modelling of the origins of uncertainty factors.

Acknowledgements

Stimulating discussions with Wenbin Wei, Francesco Falciani, and support from the Welcome Trust VIP Award (Project 10835) and a Paul and Yuanbi Ramsay research award are gratefully acknowledged. Thanks to Lehel Csató for assistance with Maple.

References

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

Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, Chichester, UK.

Bishop, C.M., Tipping, M.E., 2000. Variational relevance vector machines. In: Proc. Uncertainty in Artificial Intelligence.

Cawley, G.C., Talbot, N.L.C., 2006. Gene selection in cancer classification using sparse logistic regression with Bayesian regularisation. Bioinformatics.

Chu, W., Ghahramani, Z., Falciani, F., Wild, D.L., 2005. Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics.

Diaz-Uriate, R., 2005. Supervised methods with genomic data: A review and cautionary view. Data Analysis and Visualisation in Genomics and Proteomics. Springer-Verlag.

Efron, B., Hastie, T., Jonstone, I., Tibshirani, R., 2004. Least Angle Regression. Ann. Statist. 32, 407–499.

Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. Technical Report, HP Laboratories, Palo Alto, CA 94304, USA, April.

Figueiredo, M.A.T., 2003. Adaptive sparseness for supervised learning. IEEE Trans. Pattern Anal. Machine Intell. 25 (9).

Goutte, C., Hansen, L.K., 1997. Regularisation with a pruning prior. Neural Networks 10 (6), 1053–1059.

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

Ju, W.-H., Madigan, D., Scott, S., 2002. On Bayesian learning of sparse classifiers. (unpublished) <http://www.stat.rutgers.edu/~madigan/PAPERS/sparse3.pdf>.

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

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

Moulin, P., Liu, J., 1999. Analysis of multiresolution image denoising schemes using generalised Gaussian and complexity priors. IEEE Trans. Inform. Theory 45, 909–919.

Neal, R.M., 1996. Bayesian learning for neural networks. In: Lecture Notes in Statistics. Springer.

Park, T., Casella, G. The Bayesian Lasso. (unpublished) <http://www.stat.ufl.edu/~casella/Papers/bayeslasso.pdf>.

Qi, Y., Minka, T.P., Picard, R.W., Ghahramani, Z., 2004. Predictive automatic relevance determination by expectation propagation. In: Proc. Internat. Conf. on Machine Learning.

Roth, V., 2004. The generalised LASSO. IEEE Trans. Neural Networks 15 (1), 16–29.

Shevade, S.K., Keerthy, S.S., 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19 (17), 2246–2253.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. Roy. Statist. Soc. Ser. B 58, 267–288.

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

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionising radiation response. Proc. Natl. Acad. Sci. USA 98, 5116–5121.

Yeung, K.Y., Bumgarner, R.E., Raftery, A.E., 2005. Bayesian model averaging: Development of an improved multi-class gene selection and classification tool for microarray data. Bioinformatics 21 (10), 2394–2402.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Statist. Soc. B 67 (2), 301–320.