LETTER Communicated by Manfred Opper

Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors

Mark Girolami
girolami@dcs.gla.ac.uk
Simon Rogers
srogers@dcs.gla.ac.uk
Department of Computing Science, University of Glasgow, Glasgow, Scotland

It is well known in the statistics literature that augmenting binary and polychotomous response models with gaussian latent variables enables exact Bayesian analysis via Gibbs sampling from the parameter posterior. By adopting such a data augmentation strategy, dispensing with priors over regression coefficients in favor of gaussian process (GP) priors over functions, and employing variational approximations to the full posterior, we obtain efficient computational methods for GP classification in the multiclass setting.[1] The model augmentation with additional latent variables ensures full a posteriori class coupling while retaining the simple a priori independent GP covariance structure from which sparse approximations, such as multiclass informative vector machines (IVM), emerge in a natural and straightforward manner. This is the first time that a fully variational Bayesian treatment for multiclass GP classification has been developed without having to resort to additional explicit approximations to the nongaussian likelihood term. Empirical comparisons with exact analysis use Markov Chain Monte Carlo (MCMC) and Laplace approximations illustrate the utility of the variational approximation as a computationally economic alternative to full MCMC and it is shown to be more accurate than the Laplace approximation.

1 Introduction

Albert and Chib (1993) were the first to show that by augmenting binary and multinomial probit regression models with a set of continuous latent variables y_k, corresponding to the kth response value where $y_k = m_k + \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$ and $m_k = \sum_j \beta_{kj} x_j$, an exact Bayesian analysis can be performed by Gibbs sampling from the parameter posterior. As an example, consider binary probit regression on target variables $t_n \in \{0, 1\}$;

[1] Matlab code to allow replication of the reported results is available online at http://www.dcs.gla.ac.uk/people/personal/girolami/pubs_2005/VBGP/index.htm.

the probit likelihood for the nth data sample taking unit value ($t_n = 1$) is $P(t_n = 1|\mathbf{x}_n, \boldsymbol{\beta}) = \Phi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n)$, where Φ is the standardized normal cumulative distribution function (CDF). This can be obtained by the following marginalization, $\int P(t_n = 1, y_n|\mathbf{x}_n, \boldsymbol{\beta})dy_n = \int P(t_n = 1|y_n)p(y_n|\mathbf{x}_n, \boldsymbol{\beta})dy_n$, and by definition $P(t_n = 1|y_n) = \delta(y_n > 0)$, we see that the required marginal is simply the normalizing constant of a left truncated univariate gaussian so that $P(t_n = 1|\boldsymbol{x}_n, \boldsymbol{\beta}) = \int \delta(y_n > 0)\mathcal{N}_{y_n}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n, 1)dy_n = \Phi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n)$. The key observation here is that working with the joint distribution $P(t_n = 1, y_n|\mathbf{x}_n, \boldsymbol{\beta}) = \delta(y_n > 0)\mathcal{N}_{y_n}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n, 1)$ provides a straightforward means of Gibbs sampling from the parameter posterior, which would not be the case if the marginal term, $\Phi(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_n)$, was employed in defining the joint distribution over data and parameters.

This data augmentation strategy can be adopted in developing efficient methods to obtain binary and multiclass gaussian process (GP) (Williams & Rasmussen, 1996) classifiers, as will be presented in this article. With the exception of Neal (1998), where a full Markov chain Monte Carlo (MCMC) treatment to GP-based classification is provided, all other approaches have focused on methods to approximate the problematic form of the posterior,[2] which allow analytic marginalization to proceed. Laplace approximations to the posterior were developed in Williams and Barber (1998), and lower- and upper-bound quadratic likelihood approximations were considered in Gibbs and MacKay (2000). Variational approximations for binary classification were developed in Seeger (2000) where a logit likelihood was considered and mean field approximations were applied to probit likelihood terms in Opper and Winther (2000) and Csato, Fokue, Opper, Schottky, and Winther (2000), respectively. Additionally, incremental (Quinonero-Candela & Winther, 2003) or sparse approximations based on assumed density filtering (ADF; Csato & Opper, 2002), informative Vector Machines (IVM, Lawrence, Seeger, and Herbrich, 2003), and expectation propagation (EP; Minka, 2001; Kim, 2005) have been proposed. With the exceptions of Williams and Barber (1998), Gibbs and MacKay (2000), Seeger and Jordan (2004), and Kim (2005) the focus of most recent work has largely been on the binary GP classification problem. Seeger and Jordan (2004) developed a multiclass generalization of the IVM employing a multinomial-logit softmax likelihood. However, considerable representational effort is required to ensure that the scaling of computation and storage required of the proposed method matches that of the original IVM with linear scaling in the number of classes. In contrast, by adopting the probabilistic representation of Albert and Chib (1993), we will see that GP-based K-class classification and efficient sparse approximations (IVM generalizations with scaling linear in the number of classes) can be realized by optimizing a strict lower

[2] The likelihood is nonlinear in the parameters due to either the logistic or probit link functions required in the classification setting.

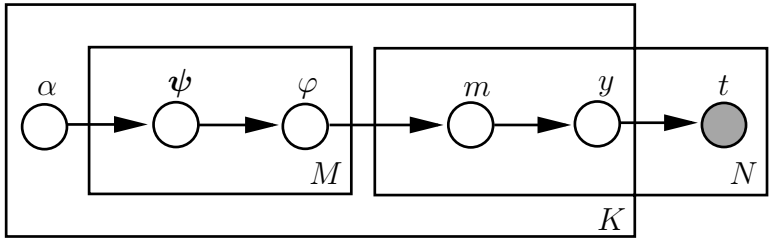


Figure 1: Graphical representation of the conditional dependencies within the general multinomial probit regression model with gaussian process priors.

bound of the marginal likelihood of a multinomial probit regression model that requires the solution of K computationally independent GP regression problems while still operating jointly (statistically) on the data. We will also show that the accuracy of this approximation is comparable to that obtained via MCMC.

The following section introduces the multinomial-probit regression model with GP priors.

2 Multinomial Probit Regression

Define the data matrix as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$, which has dimension $N \times D$ and the $N \times 1$–dimensional vector of associated target values as $\mathbf{t}$ where each element $t_n \in \{1, \ldots, K\}$. The $N \times K$ matrix of GP random variables m_{nk} is denoted by $\mathbf{M}$. We represent the $N \times 1$ dimensional columns of $\mathbf{M}$ by $\mathbf{m}_k$ and the corresponding $K \times 1$–dimensional rows by $\mathbf{m}_n$. The $N \times K$ matrix of auxiliary variables y_{nk} is represented as $\mathbf{Y}$, where the $N \times 1$–dimensional columns are denoted by $\mathbf{y}_k$ and the corresponding $K \times 1$–dimensional rows as $\mathbf{y}_n$. The $M \times 1$ vector of covariance kernel hyperparameters for each class is denoted by $\boldsymbol{\varphi}_k$, and associated hyperparameters $\boldsymbol{\psi}_k$ and $\boldsymbol{\alpha}_k$ complete the model.[3]

The graphical representation of the conditional dependency structure in the auxiliary variable multinomial probit regression model with GP priors in the most general case is shown in Figure 1.

3 Prior Probabilities

From the graphical model in Figure 1 a priori we can assume class-specific GP independence and define model priors such that $\mathbf{m}_k|\mathbf{X}, \boldsymbol{\varphi}_k \sim GP(\boldsymbol{\varphi}_k) = \mathcal{N}_{\mathbf{m}_k}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varphi}_k})$, where the matrix $\mathbf{C}_{\boldsymbol{\varphi}_k}$ of dimension $N \times N$ defines

[3] This is the most general setting; however, it is more common to employ a single and shared GP covariance function across classes.

the class-specific GP covariance.[4] Typical examples of such GP covariance functions are radial basis–style functions such that the i, jth element of each $\mathbf{C}_{\varphi_k}$ is defined as $\exp\{-\sum_{d=1}^{M} \varphi_{kd}(x_{id} - x_{jd})^2\}$, where in this case $M = D$; however, there are many other forms of covariance functions that may be employed within the GP function prior (see, e.g., MacKay, 2003).

As in Albert and Chib (1993), we employ a standardized normal noise model such that the prior on the auxilliary variables is $y_{nk}|m_{nk} \sim \mathcal{N}_{y_{nk}}(m_{nk}, 1)$ to ensure appropriate matching with the probit function. Rather than having this variance fixed, it could also be made an additional free parameter of the model and therefore would yield a scaled probit function. For the presentation here, we restrict ourselves to the standardized model and consider extensions to a scaled probit model as possible further work. The relationship between the additional latent variables $\mathbf{y}_n$ (denoting the nth row of $\mathbf{Y}$) and the targets t_n as defined in multinomial probit regression (Albert and Chib, 1993) is adopted here:

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \le k \le K} \{y_{nk}\}. \tag{3.1}$$

This has the effect of dividing $\mathbb{R}^K$ ($\mathbf{y}$ space) into K nonoverlapping K-dimensional cones $\mathcal{C}_k = \{\mathbf{y} : y_k > y_i, k \neq i\}$ where $\mathbb{R}^K = \cup_k \mathcal{C}_k$, and so each $P(t_n = i|\mathbf{y}_n)$ can be represented as $\delta(y_{ni} > y_{nk} \; \forall \, k \neq i)$. We then see that similar to the binary case, where the probit function emerges from explicitly marginalizing the auxiliary variable, the multinomial probit takes the form given below (details are given in appendix A),

$$\begin{aligned} P(t_n = i|\mathbf{m}_n) &= \int \delta(y_{ni} > y_{nk} \; \forall \, k \neq i) \prod_{j=1}^{K} p(y_{nj}|m_{nj}) \, d\mathbf{y} \\ &= \int_{\mathcal{C}_i} \prod_{j=1}^{K} p(y_{nj}|m_{nj}) d\mathbf{y} = E_{p(u)} \left\{ \prod_{j \neq i} \Phi(u + m_{ni} - m_{nj}) \right\}, \end{aligned}$$

where the random variable u is standardized normal, $p(u) = \mathcal{N}(0, 1)$. A hierarchic prior on the covariance function hyperparameters is employed such that each hyperparameter has, for example, an independent exponential distribution $\varphi_{kd} \sim \text{Exp}(\psi_{kd})$ and a gamma distribution is placed on the mean values of the exponential $\psi_{kd} \sim \Gamma(\sigma_k, \tau_k)$, thus forming a conjugate pair. Of course, as detailed in Girolami and Rogers (2005), a

[4] The model can be defined by employing $K - 1$ GP functions and an alternative truncation of the gaussian over the variables y_{nk}; however, for the multiclass case, we define a GP for each class.

more general form of covariance function can be employed that will allow the integration of heterogeneous types of data, which takes the form of a weighted combination of base covariance functions. The associated hyper-hyperparameters $\boldsymbol{\alpha} = \{\sigma_{k=1,\ldots,K}, \tau_{k=1,\ldots,K}\}$ can be estimated via type II maximum likelihood or set to reflect some prior knowledge of the data. Alternatively, vague priors can be employed such that, for example, each $\sigma_k = \tau_k = 10^{-6}$. Defining the parameter set as $\boldsymbol{\Theta} = \{\mathbf{Y}, \mathbf{M}\}$ and the hyperparameters as $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_{k=1,\ldots,K}, \boldsymbol{\psi}_{k=1,\ldots,K}\}$, the joint likelihood takes the form

$$p(\mathbf{t}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\mathbf{X}, \boldsymbol{\alpha}) = \prod_{n=1}^{N} \left\{ \sum_{i=1}^{K} \delta(y_{ni} > y_{nk} \; \forall \, k \neq i) \delta(t_n = i) \right\} \times \prod_{k=1}^{K} p(y_{nk}|m_{nk}) p(\mathbf{m}_k|\mathbf{X}, \boldsymbol{\varphi}_k) p(\boldsymbol{\varphi}_k|\boldsymbol{\psi}_k) p(\boldsymbol{\psi}_k|\boldsymbol{\alpha}_k). \quad (3.2)$$

4 Gaussian Process Multiclass Classification

We now consider both exact and approximate Bayesian inference for GP classification with multiple classes employing the multinomial-probit regression model.

4.1 Exact Bayesian Inference: The Gibbs Sampler. The representation of the joint likelihood (see equation 3.2) is particularly convenient in that samples can be drawn from the full posterior over the model parameters (given the hyperparameter values) $p(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$ using a Gibbs sampler in a very straightforward manner with scaling per sample of $\mathcal{O}(KN^3)$. Full details of the Gibbs sampler are provided in appendix D, and this sampler will be employed in the experimental section.

4.2 Approximate Bayesian Inference: The Laplace Approximation. The Laplace approximation of the posterior over GP variables, $p(\mathbf{M}|\mathbf{t}, \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$ (where $\mathbf{Y}$ is marginalized) requires finding the mode of the unnormalized posterior. Approximate Bayesian inference for GP classification with multiple classes employing a multinomial logit (softmax) likelihood has been developed previously Williams and Barber (1998). Due to the form of the multinomial logit likelihood, a Newton iteration to obtain the posterior mode will scale at best as $\mathcal{O}(KN^3)$. Employing the multinomial probit likelihood, we find that each Newton step will scale as $\mathcal{O}(K^3N^3)$. Details are provided in appendix E.

4.3 Approximate Bayesian Inference: A Variational and Sparse Approximation. Employing a variational Bayes approximation (Beal, 2003; Jordan, Ghahramani, Jaakkola, & Saul, 1999; MacKay, 2003) by using an approximating ensemble of factored posteriors such that $p(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha}) \approx$

$\prod_{i=1} Q(\mathbf{\Theta}_i) = Q(\mathbf{Y})Q(\mathbf{M})$ for multinomial probit regression is more appealing from a computational perspective as a sparse representation, with scaling $\mathcal{O}(KNS^2)$ (where S is the subset of samples entering the model and $S \ll N$), can be obtained in a straightforward manner, as will be shown in the following sections. The lower bound[5] (see, e.g., Beal, 2003; Jordan et al., 1999; MacKay, 2003) on the marginal likelihood $\log p(\mathbf{t}|\mathbf{X}, \mathbf{\Phi}, \boldsymbol{\alpha}) \geq E_{Q(\mathbf{\Theta})}\left\{\log p(\mathbf{t}, \mathbf{\Theta}|\mathbf{X}, \mathbf{\Phi}, \boldsymbol{\alpha})\right\} - E_{Q(\mathbf{\Theta})}\left\{\log Q(\mathbf{\Theta})\right\}$ is minimized by distributions that take an unnormalized form of $Q(\mathbf{\Theta}_i) \propto \exp\left(E_{Q(\mathbf{\Theta}\setminus\mathbf{\Theta}_i)}\{\log P(\mathbf{t}, \mathbf{\Theta}|\mathbf{X}, \mathbf{\Phi}, \boldsymbol{\alpha})\}\right)$ where $Q(\mathbf{\Theta}\setminus\mathbf{\Theta}_i)$ denotes the ensemble distribution with the ith component of $\mathbf{\Theta}$ removed. Details of the required posterior components are given in appendix A.

The approximate posterior over the GP random variables takes a factored form such that

$$Q(\mathrm{M}) = \prod_{k=1}^{K} Q(\mathbf{m}_k) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{m}_k}(\widetilde{\mathbf{m}}_k, \Sigma_k), \tag{4.1}$$

where the shorthand tilde notation denotes posterior expectation, $\widetilde{f(a)} = E_{Q(a)}\{f(a)\}$, and so the required posterior mean for each k is given as $\widetilde{\mathbf{m}}_k = \mathbf{\Sigma}_k\widetilde{\mathbf{y}}_k$, where $\mathbf{\Sigma}_k = \mathbf{C}_{\boldsymbol{\varphi}_k}(\mathbf{I} + \mathbf{C}_{\boldsymbol{\varphi}_k})^{-1}$ (see appendix A for full details).

We will see that each row, $\widetilde{\mathbf{y}}_n$, of $\widetilde{\mathbf{Y}}$ will have posterior correlation structure induced, ensuring that the appropriate class-conditional posterior dependencies will be induced in $\widetilde{\mathbf{M}}$. It should be stressed here that while there are K a posteriori independent GP processes, the associated K-dimensional posterior means for each of N data samples induces posterior dependencies between each of the K columns of $\widetilde{\mathbf{M}}$ due to the posterior coupling over each of the auxiliary variables $\mathbf{y}_n$. We will see that this structure is particularly convenient in obtaining sparse approximations (Lawrence et al., 2003) for the multiclass GP in particular.

Due to the multinomial probit definition of the dependency between each element of $\mathbf{y}_n$ and t_n (see equation 3.1), the posterior for the auxiliary variables follows as

$$Q(\mathbf{Y}) = \prod_{n=1}^{N} Q(\mathbf{y}_n) = \prod_{n=1}^{N} \mathcal{N}_{\mathbf{y}_n}^{t_n}(\widetilde{\mathbf{m}}_n, \mathbf{I}) \tag{4.2}$$

where $\mathcal{N}_{\mathbf{y}_n}^{t_n}(\widetilde{\mathbf{m}}_n, \mathbf{I})$ denotes a conic truncation of a multivariate gaussian such that if $t_n = i$ where $i \in \{1, \ldots, K\}$, then the ith dimension has the largest

[5] The bound follows from the application of Jensen's inequality, for example, $\log p(\mathbf{t}|\mathbf{X}) = \log \int \frac{p(\mathbf{t}, \Theta|\mathbf{X})}{Q(\Theta)} Q(\Theta) d\Theta \geq \int Q(\Theta) \log \frac{p(\mathbf{t}, \Theta|\mathbf{X})}{Q(\Theta)} d\Theta$.

value. The required posterior expectations $\widetilde{y}_{nk}$ for all $k \neq i$ and $\widetilde{y}_{ni}$ follow as

$$\widetilde{y}_{nk} = \widetilde{m}_{nk} - \frac{E_{p(u)}\left\{\mathcal{N}_u(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1)\Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nk})\Phi_u^{n,i,k}\right\}} \tag{4.3}$$

$$\widetilde{y}_{ni} = \widetilde{m}_{ni} - \left(\sum_{j \neq i} \widetilde{y}_{nj} - \widetilde{m}_{nj}\right), \tag{4.4}$$

where $\Phi_u^{n,i,k} = \prod_{j \neq i,k} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj})$, and $p(u) = \mathcal{N}_u(0, 1)$. The expectations with respect to $p(u)$, which appear in equation 4.3, can be obtained by quadrature or straightforward sampling methods.

If we also consider the set of hyperparameters, $\mathbf{\Phi}$, in this variational treatment, then the approximate posterior for the covariance kernel hyperparameters takes the form of

$$Q(\boldsymbol{\varphi}_k) \propto \mathcal{N}_{\widetilde{\mathbf{m}}_k}(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varphi}_k}) \prod_{d=1}^{M} \mathrm{Exp}(\varphi_{kd} | \widetilde{\psi}_{kd}),$$

and the required posterior expectations can be estimated employing importance sampling. Expectations can be approximated by drawing S samples such that each $\varphi_{kd}^s \sim \mathrm{Exp}(\widetilde{\psi}_{kd})$, and so

$$\widetilde{f(\boldsymbol{\varphi}_k)} \approx \sum_{s=1}^{S} f(\boldsymbol{\varphi}_k^s) w(\boldsymbol{\varphi}_k^s) \;\text{ where }\; w(\boldsymbol{\varphi}_k^s) = \frac{\mathcal{N}_{\widetilde{\mathbf{m}}_k}\left(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varphi}_k^s}\right)}{\sum_{s'=1}^{S} \mathcal{N}_{\widetilde{\mathbf{m}}_k}\left(\mathbf{0}, \mathbf{C}_{\boldsymbol{\varphi}_k^{s'}}\right)}. \tag{4.5}$$

This form of importance sampling within a variational Bayes procedure has been employed previously in Lawrence, Milo, Niranjan, Rashbass, and Soullier (2004). Clearly the scaling of the above estimator per sample is similar to that required in the gradient-based methods that search for optima of the marginal likelihood as employed in GP regression and classification (e.g., MacKay, 2003).

Finally, we have that each $Q(\psi_{kd}) = \Gamma_{\psi_{kd}}(\sigma_k + 1, \tau_k + \widetilde{\varphi}_{kd})$, and the associated posterior mean is simply $\widetilde{\psi}_{kd} = (\sigma_k + 1)/(\tau_k + \widetilde{\varphi}_{kd})$.

4.4 Summarizing Variational Multiclass GP Classification. We can summarize what has been presented by the following iterations that, in the general case, for all k and d, will optimize the bound on the marginal likelihood (explicit expressions for the bound are provided in appendix C):

$$\widetilde{\mathbf{m}}_k \leftarrow \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}(\mathbf{I} + \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k})^{-1}(\widetilde{\mathbf{m}}_k + \mathbf{p}_k) \tag{4.6}$$

$$\widetilde{\boldsymbol{\varphi}}_k \leftarrow \sum_s \boldsymbol{\varphi}_k^s w(\boldsymbol{\varphi}_k^s) \tag{4.7}$$

$$\widetilde{\psi}_{kd} \leftarrow \frac{\sigma_k + 1}{\tau_k + \widetilde{\varphi}_{kd}}, \tag{4.8}$$

where each $\varphi^s_{kd} \sim \text{Exp}(\widetilde{\psi}_{kd})$, $w(\boldsymbol{\varphi}^s_k)$ is defined as previously and $\mathbf{p}_k$ is the kth column of the $N \times K$ matrix $\mathbf{P}$ whose elements p_{nk} are defined by the right-most terms in equations 4.3 and 4.4: for $t_n = i$, then for all $k \neq i$, $p_{nk} = -\frac{E_{p(u)}\{\mathcal{N}_u(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1)\Phi_u^{n,i,k}\}}{E_{p(u)}\left\{\Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nk})\Phi_u^{n,i,k}\right\}}$ and $p_{ni} = -\sum_{j \neq i} p_{nj}$.

These iterations can be viewed as obtaining K one-against-all binary classifiers: however, most important, they are not statistically independent of each other but are a posteriori coupled via the posterior mean estimates of each of the auxiliary variables $\mathbf{y}_n$. The computational scaling will be linear in the number of classes and cubic in the number of data points $\mathcal{O}(K N^3)$. It is worth noting that if the covariance function hyperparameters are fixed, then the costly matrix inversion requires being computed only once. The Laplace approximation will require a matrix inversion for each Newton step when finding the mode of the posterior (Williams & Barber, 1998).

4.4.1 Binary Classification. Previous variational treatments of GP-based binary classification include Seeger (2000), Opper and Winther (2000), Gibbs and MacKay (2000), Csato and Opper (2002), and Csato et al. (2000). It is, however, interesting to note in passing that for binary classification, the outer plate in Figure 1 is removed, and further simplification follows as only $K - 1$, that is, one set of posterior mean values requires being estimated, and so the posterior expectations $\widetilde{\mathbf{m}} = \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}}(\mathbf{I} + \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}})^{-1}\widetilde{\mathbf{y}}$ now operate on $N \times$ 1–dimensional vectors $\widetilde{\mathbf{m}}$ and $\widetilde{\mathbf{y}}$. The posterior $Q(\mathbf{y})$ is now a product of truncated univariate gaussians, so the expectation for the latent variables y_n has an exact analytic form. For a unit variance gaussian truncated below zero if $t_n = 1$ and above zero if $t_n = -1$, the required posterior mean $\widetilde{\mathbf{y}}$ has elements that can be obtained by the following analytic expression derived from straightforward results for corrections to the mean of a gaussian due to truncation $\widetilde{y}_n = \widetilde{m}_n + t_n \mathcal{N}_{\widetilde{m}_n}(0, 1)/\ \Phi(t_n \widetilde{m}_n)$.[6] So the following iteration will guarantee an increase in the bound of the marginal likelihood,

$$\widetilde{\mathbf{m}} \leftarrow \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}}(\mathbf{I} + \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}})^{-1}(\widetilde{\mathbf{m}} + \mathbf{p}), \tag{4.9}$$

where each element of the $N \times 1$ vector $\mathbf{p}$ is defined as $p_n = t_n \mathcal{N}_{\widetilde{m}_n}(0, 1)/\Phi(t_n \widetilde{m}_n)$.

4.5 Variational Predictive Distributions. The predictive distribution, $P(t_{new} = k | \mathbf{x}_{new}, \mathbf{X}, \mathbf{t})$, for a new sample $\mathbf{x}_{new}$ follows from results for

[6] For $t = +1$, then $\widetilde{y} = \int_0^{+\infty} y \mathcal{N}_y(\widetilde{m}, 1)/\{1 - \Phi(-\widetilde{m})\} dy = \widetilde{m} + \mathcal{N}_{\widetilde{m}}(0, 1)/\Phi(\widetilde{m})$, and for $t = -1$, then $\widetilde{y} = \int_{-\infty}^0 y \mathcal{N}_y(\widetilde{m}, 1)/\Phi(-\widetilde{m}) dy = \widetilde{m} - \mathcal{N}_{\widetilde{m}}(0, 1)/\Phi(-\widetilde{m})$.

standard GP regression.[7] The $N \times 1$ vector $\mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{new}$ contains the covariance function values between the new point and those contained in $\mathbf{X}$, and $c_{\widetilde{\boldsymbol{\varphi}}_k}^{new}$ denotes the covariance function value for the new point and itself. So the GP posterior $p(\mathbf{m}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})$ is a product of K gaussians each with mean and variance

$$\widetilde{m}_k^{new} = \widetilde{\mathbf{y}}_k^{\mathrm{T}} \left(\mathbf{I} + \mathbf{C}_{\widetilde{\varphi}_k}\right)^{-1} \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{new}$$
$$\widetilde{\sigma}_{k,new}^2 = c_{\widetilde{\boldsymbol{\varphi}}_k}^{new} - (\mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{new})^{\mathrm{T}} \left(\mathbf{I} + \mathbf{C}_{\widetilde{\varphi}_k}\right)^{-1} \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{new}$$

using the following shorthand, $\widetilde{v}_k^{new} = \sqrt{1 + \widetilde{\sigma}_{k,new}^2}$, then it is straightforward (details in appendix B) to obtain the predictive distribution over possible target values as

$$P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = E_{p(u)} \left\{ \prod_{j \neq k} \Phi \left(\frac{1}{\widetilde{v}_j^{new}} \left[u\widetilde{v}_k^{new} + \widetilde{m}_k^{new} - \widetilde{m}_j^{new} \right] \right) \right\},$$

where, as before, $u \sim \mathcal{N}_u(0, 1)$. The expectation can be obtained numerically employing sample estimates from a standardized gaussian. For the binary case, the standard result follows:

$$\begin{aligned} P(t_{new} = 1|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) &= \int \delta(y^{new} > 0) \mathcal{N}_{y^{new}}(\widetilde{m}^{new}, \widetilde{v}^{new})\, dy^{new} \\ &= 1 - \Phi\left(-\frac{\widetilde{m}^{new}}{\widetilde{v}^{new}}\right) = \Phi\left(\frac{\widetilde{m}^{new}}{\widetilde{v}^{new}}\right). \end{aligned}$$

5 Sparse Variational Multiclass GP Classification

The dominant $\mathcal{O}(N^3)$ scaling of the matrix inversion required in the posterior mean updates in GP regression has been the motivation behind a large body of literature focusing on reducing this cost via reduced rank approximations (Williams & Seeger, 2001) and sparse online learning (Csato & Opper, 2002; Quinonero-Candela & Winther, 2003) where assumed density filtering (ADF) forms the basis of online learning and sparse approximations for GPs. Likewise in Lawrence et al. (2003) the informative vector machine (IVM) (refer to Lawrence, Platt, & Jordan, 2005, for comprehensive details) is proposed, which employs informative point selection criteria (Seeger, Williams, & Lawrence, 2003) and ADF updating of the approximations of

[7] Conditioning on $\widetilde{\mathbf{Y}}$, $\widetilde{\boldsymbol{\varphi}}$, $\widetilde{\boldsymbol{\psi}}$, and $\boldsymbol{\alpha}$ is implicit.

the GP posterior parameters. Only binary classification has been considered in Lawrence et al. (2003), Csato and Opper (2002), and Quinonero-Candela and Winther (2003), and it is clear from Seeger and Jordan (2004) that extension of ADF-based approximations such as IVM to the multiclass problem is not at all straightforward when a multinomial-logit softmax likelihood is adopted. However, we now see that sparse GP-based classification for multiple classes (multiclass IVM) emerges as a simple by-product of online ADF approximations to the parameters of each $Q(\mathbf{m}_k)$ (multivariate gaussian). The ADF approximations when adding the nth data sample, selected at the lth of S iterations, for each of the K GP posteriors, $Q(\mathbf{m}_k)$, follow simply from details in Lawrence et al. (2005) as given below:

$$\Sigma_{k,n} \leftarrow \mathbf{C}^n_{\boldsymbol{\varphi}_k} - \mathbf{M}_k^{\mathrm{T}}\mathbf{M}_{k,n} \tag{5.1}$$

$$\mathbf{s}_k \leftarrow \mathbf{s}_k - \frac{1}{1+s_{kn}}\mathrm{diag}\left(\Sigma_{k,n}\Sigma_{k,n}^{\mathrm{T}}\right) \tag{5.2}$$

$$\mathbf{M}_k^l \leftarrow \frac{1}{\sqrt{1+s_{kn}}}\Sigma_{k,n}^{\mathrm{T}} \tag{5.3}$$

$$\widetilde{\mathbf{m}}_k \leftarrow \widetilde{\mathbf{m}}_k + \frac{\widetilde{y}_{nk} - \widetilde{m}_{nk}}{1+s_{kn}}\Sigma_{k,n}. \tag{5.4}$$

Each $\widetilde{y}_{nk} - \widetilde{m}_{nk} = p_{nk}$ as defined in section 4.4 and can be obtained from the current stored approximate values of each $\widetilde{m}_{n1}, \ldots, \widetilde{m}_{nK}$ via equations 4.3 and 4.4, $\Sigma_{k,n}$, an $N \times 1$ vector, is the nth column of the current estimate of each Σ_k; likewise, $\mathbf{C}^n_{\boldsymbol{\varphi}_k}$ is the nth column of each GP covariance matrix. All elements of each $\mathbf{M}_k$ and $\mathbf{m}_k$ are initialized to zero, while each $\mathbf{s}_k$ has initial unit values. Of course, there is no requirement to explicitly store each $N \times N$–dimensional matrix $\boldsymbol{\Sigma}_k$; only the $S \times N$ matrices $\mathbf{M}_k$ and $N \times 1$ vectors $\mathbf{s}_k$ require storage and maintenance. We denote indexing into the lth row of each $\mathbf{M}_k$ by $\mathbf{M}_k^l$, and the nth element of each $\mathbf{s}_k$ by s_{kn}, which is the estimated posterior variance.

The efficient Cholesky factor updating as detailed in Lawrence et al. (2005) will ensure that for N data samples, K distinct GP priors, and a maximum of S samples included in the model where $S \ll N$, then at most $\mathcal{O}(KSN)$ storage and $\mathcal{O}(KNS^2)$ compute scaling will be realized.

As an alternative to the entropic scoring heuristic of Seeger et al. (2003) and Lawrence et al. (2003), we suggest that an appropriate criterion for point inclusion assessment will be the posterior predictive probability of a target value given the current model parameters for points that are currently not included in the model that is, $P(t_m|\mathbf{x}_m, \{\mathbf{m}_k\}, \{\boldsymbol{\Sigma}_k\})$, where the subscript m indexes such points. From the results of the previous section, this is equal

to $Pr(\mathbf{y}_m \in \mathcal{C}_{t_m=k})$, which is expressed as

$$E_{p(u)}\left\{\prod_{j\neq k}\Phi\left(\frac{1}{\nu_{jm}}\left[u\nu_{km}+\widetilde{m}_{mk}-\widetilde{m}_{mj}\right]\right)\right\}, \tag{5.5}$$

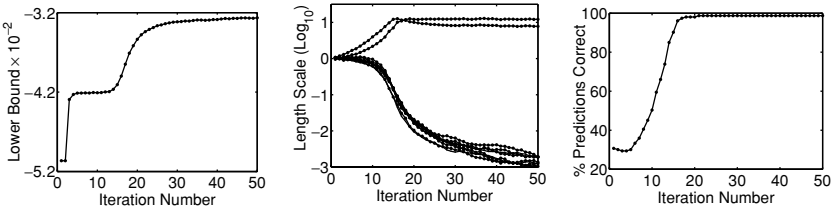
where k is the value of t_m, $\nu_{jm}=\sqrt{1+s_{jm}}$, and so the data point with the smallest posterior target probability should be selected for inclusion. This scoring criterion requires no additional storage overhead, as all $\widetilde{\mathbf{m}}_k$ and $\mathbf{s}_k$ are already available and it can be computed for all m not currently in the model in, at most, $\mathcal{O}(KN)$ time.[8] Intuitively, points in regions of low target posterior certainty, that is, class boundaries, will be the most influential in updating the approximation of the target posteriors. And so the inclusion of points with the most uncertain target posteriors will yield the largest possible translation of each updated $\mathbf{m}_k$ into the interior of their respective cones $\mathcal{C}_k$. Experiments in the following section will demonstrate the effectiveness of this multiclass IVM.

6 Experiments

6.1 Illustrative Multiclass Toy Example. Ten-dimensional data vectors, $\mathbf{x}$, were generated such that if $t=1$, then $0.5 > x_1^2 + x_2^2 > 0.1$; for $t=2$, then $1.0 > x_1^2 + x_2^2 > 0.6$; and for $t=3$, then $[x_1, x_2]^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, 0.01\mathbf{I})$ where $\mathbf{I}$ denotes an identity matrix of appropriate dimension. Finally $x_3, \ldots, x_{10}$ are all distributed as $\mathcal{N}(0,1)$. Both of the first two dimensions are required to define the three class labels, with the remaining eight dimensions being irrelevant to the classification task. Each of the three target values was sampled uniformly, thus creating a balance of samples drawn from the three target classes.

Two hundred forty draws were made from the above distribution, and the sample was used in the proposed variational inference routine, with a further 4620 points being used to compute a 0-1 loss class prediction error. A common radial basis covariance function of the form $\exp\{-\sum_d \varphi_d |x_{id} - x_{jd}|^2\}$ was employed, and vague hyperparameters, $\sigma = \tau = 10^{-3}$ were placed on the length-scale hyperparameters $\psi_1, \ldots, \psi_{10}$. The posterior expectations of the auxiliary variables $\widetilde{\mathbf{y}}$ were obtained from equations 4.3 and 4.4 where the gaussian integrals were computed using 1000 samples drawn from $p(u) = \mathcal{N}(0,1)$. The variational importance sampler employed 500 samples drawn from each $\mathrm{Exp}(\widetilde{\psi}_d)$ in estimating the corresponding posterior means $\widetilde{\varphi}_d$ for the covariance function parameters. Each $\mathbf{M}$ and $\mathbf{Y}$ was initialized randomly, and $\boldsymbol{\varphi}$ had unit initial values. In this

[8] Assuming constant time to approximate the expectation.



(a) (b) (c)

Figure 2: (a) Convergence of the lower bound on the marginal likelihood for the toy data set considered. (b) Evolution of estimated posterior means for the inverse squared length scale parameters (precision parameters) in the RBF covariance function. (c) Evolution of out-of-sample predictive performance on the toy data set.

example, the variational iterations ran for 50 steps, where each step corresponds to the sequential posterior mean updates of equations 4.6 to 4.8. The value of the variational lower bound was monitored during each step, and as would be expected, a steady convergence in the improvement of the bound can be observed in Figure 2a.

The development of the estimated posterior mean values for the covariance function parameters $\widetilde{\varphi}_d$, Figure 2b, shows automatic relevance detection (ARD) in progress (Neal, 1998), where the eight irrelevant features are effectively removed from the model.

From Figure 2c we can see that the development of the predictive performance (out of sample) follows that of the lower bound (see Figure 2a), achieving a predictive performance of 99.37% at convergence. As a comparison to our multiclass GP classifier, we use a directed acyclic graph (DAG) SVM (Platt, Cristianni, & Shawe-Taylor, 2000) (assuming equal class distributions the scaling[9] is $\mathcal{O}\left(N^3K^{-1}\right)$) on this example. By employing the values of the posterior mean values of the covariance function length scale parameters (one for each of the 10 dimensions) estimated by the proposed variational procedure in the RBF kernel of the DAG SVM, a predictive performance of 99.33% is obtained. So on this data set, the proposed GP classifier has comparable performance, under 0-1 loss, to the DAG SVM. However, the estimation of the covariance function parameters is a natural part of the approximate Bayesian inference routines employed in GP classification. There is no natural method of obtaining estimates of the 10 kernel parameters for the SVM without resorting to cross validation (CV), which, in the case of a single parameter, is feasible but rapidly becomes infeasible as the number of parameters increases.

[9] This assumes the use of standard quadratic optimization routines.

6.2 Comparing Laplace and Variational Approximations to Exact then Inference via Gibbs Sampling. This section provides a brief empirical comparison of the variational approximation, developed in previous sections, to a full MCMC treatment employing the Gibbs sampler detailed in appendix D. In addition, a Laplace approximation is considered in this short comparative study.

Variational approximations provide a strict lower bound on the marginal likelihood, and this bound is one of the approximation's attractive characteristics. However, it is less well understood how much the parameters obtained from such approximations differ from those obtained using exact methods. Preliminary analysis of the asymptotic properties of variational estimators is provided in Wang and Titterington (2004). A recent experimental study of EP and Laplace approximations to binary GP classifiers has been undertaken by Kuss and Rasmussen (2005), and it is motivating to consider a similar comparison for the variational approximation in the multiple-class setting. Kuss and Rasmussen observed that the marginal and predictive likelihoods, computed over a wide range of covariance kernel hyperparameter values, were less well preserved by the Laplace approximation than the EP approximation when compared to that obtained by MCMC. We then consider the predictive likelihood obtained by the Gibbs sampler and compare this to the variational and Laplace approximations of the GP-based classifiers.

The toy data set from the previous section is employed, and as in Kuss and Rasmussen (2005), a covariance kernel of the form $s \exp\{-\varphi \sum_d \|x_{id} - x_{jd}\|^2\}$ is adopted. Both s & φ are varied in the range (log scale) -1 to $+5$ and at each pair of hyperparameter values, a multinomial probit GP classifier is induced using (1) MCMC via the Gibbs sampler, (2) the proposed variational approximation, and (3) a Laplace approximation of the probit model. For the Gibbs sampler, after a burn-in of 2000 samples, the following 1000 samples were used for inference purposes, and the predictive likelihood (probability of target values in the test set) and test error (0-1 error loss) were estimated from the 1000 post-burn-in samples as detailed in appendix D.

We first consider a binary classification problem by merging classes 2 and 3 of the toy data set into one class. The first thing to note from Figure 3 is that the predictive likelihood response under the variational approximation preserves, to a rather good degree, the predictive likelihood response obtained when using Gibbs sampling across the range of hyperparameter values. However, the Laplace approximation does not do as good a job in replicating the levels of the response profile obtained using MCMC over the range of hyperparameter values considered. This finding is consistent with the results of Kuss and Rasmussen (2005).

The Laplace approximation to the multinomial probit model has $\mathcal{O}(K^3N^3)$ scaling (see appendix E), which limits its application to situations where the number of classes is small. For this reason, in the following experiments we instead consider the multinomial logit Laplace approximation

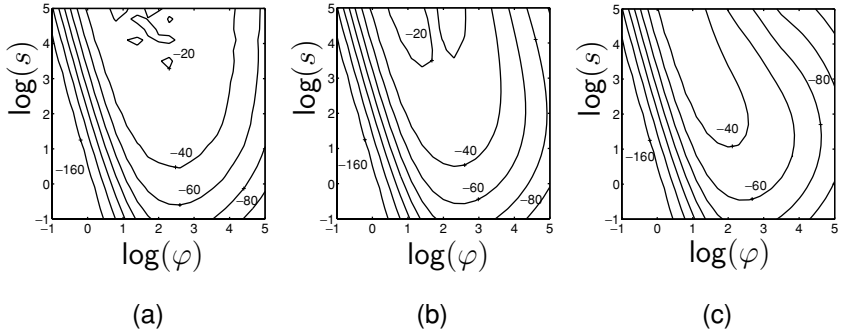


Figure 3: Isocontours of predictive likelihood for binary classification problem: (a) Gibbs sampler, (b) variational approximation, (c) Laplace approximation.

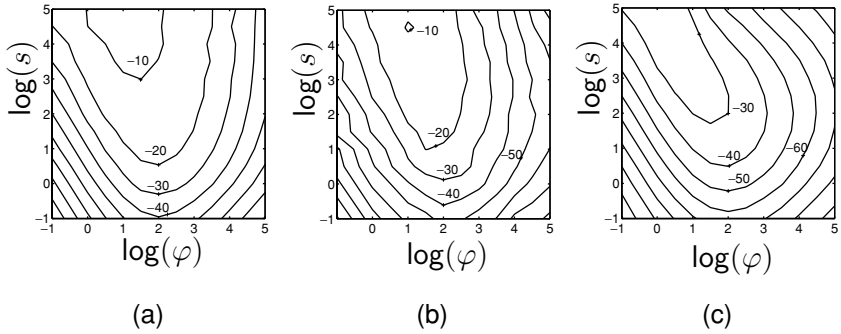


Figure 4: Isocontours of predictive likelihood for multiclass classification problem: (a) Gibbs sampler, (b) variational approximation, (c) Laplace approximation.

(Williams & Barber, 1998). In Figure 4 the isocontours of predictive likelihood for the toy data set in the multiclass setting under various hyperparameter settings are provided.

As with the binary case, the variational multinomial probit approximation provides predictive likelihood response levels that are good representations of those obtained from the Gibbs sampler. The Laplace approximation for the multinomial logit suffers from the same distortion of the contours as does the Laplace approximation for the binary probit; in addition, the information in the predictions is lower. We note, as in Kuss and Rasmussen (2005), that for $s = 1$ ($\log s = 0$), the Laplace approximation compares reasonably with results from both MCMC and variational approximations.

In the following experiment, four standard multiclass data sets (Iris, Thyroid, Wine, and Forensic Glass) from the UCI Machine Learning Data

Table 1: Results of Comparison of Gibbs Sampler, Variational, and Laplace Approximations When Applied to Several UCI Data Sets.

	Laplace	Variational	Gibbs Sampler
Toy Data			
Marginal likelihood	-169.27 ± 4.27	-232.00 ± 17.13	-94.07 ± 11.26
Predictive error	3.97 ± 2.00	3.65 ± 1.95	3.49 ± 1.69
Predictive likelihood	-98.90 ± 8.22	$\mathbf{-72.27 \pm 9.25}$	$\mathbf{-73.44 \pm 7.67}$
Iris			
Marginal likelihood	-143.87 ± 1.04	-202.98 ± 1.37	-45.27 ± 6.17
Predictive error	3.88 ± 2.00	4.08 ± 2.16	4.08 ± 2.16
Predictive likelihood	-10.43 ± 1.12	$\mathbf{-7.35 \pm 1.27}$	$\mathbf{-7.26 \pm 1.40}$
Thyroid			
Marginal likelihood	-158.18 ± 1.94	-246.24 ± 1.63	-68.82 ± 8.29
Predictive error	4.73 ± 2.36	3.86 ± 2.04	3.94 ± 2.02
Predictive likelihood	-19.01 ± 2.55	$\mathbf{-14.62 \pm 2.70}$	$\mathbf{-14.47 \pm 2.39}$
Wine			
Marginal likelihood	-152.22 ± 1.29	-253.90 ± 1.52	-68.65 ± 6.19
Predictive error	2.95 ± 2.16	2.65 ± 1.87	2.78 ± 2.07
Predictive likelihood	-14.57 ± 1.29	$\mathbf{-10.16 \pm 1.47}$	$\mathbf{-10.47 \pm 1.41}$
Forensic Glass			
Marginal likelihood	-275.11 ± 2.87	-776.79 ± 5.75	-268.21 ± 5.46
Predictive error	36.54 ± 4.74	32.79 ± 4.57	34.00 ± 4.62
Predictive likelihood	-90.38 ± 3.25	$\mathbf{-77.60 \pm 3.91}$	$\mathbf{-79.86 \pm 4.80}$

Note: Best results for Predictive likelihood are highlighted in bold.

Repository,[10] along with the toy data previously described, are used. For each data set a random 60% training and 40% testing split was used to assess the performance of each of the classification methods being considered, and 50 random splits of each data set were used. For the toy data set 50, random train and test sets were generated. The hyperparameters for an RBF covariance function taking the form of $\exp\{-\sum_d \varphi_d \|x_{id} - x_{jd}\|^2\}$ were estimated employing the variational importance sampler, and these were then fixed and employed in all the classification methods considered. The marginal likelihood for the Gibbs sampler was estimated by using 1000 samples from the GP prior. For each data set and each method (multinomial logit Laplace approximation, variational approximation, and Gibbs sampler), the marginal likelihood, (lower bound in the case of the variational approximation), predictive error (0-1 loss), and predictive likelihood were measured. The results, given as the mean and standard deviation over the 50 data splits, are listed in Table 1.

The predictive likelihood obtained from the multinomial logit Laplace approximation is consistently, across all data sets, lower than that of the

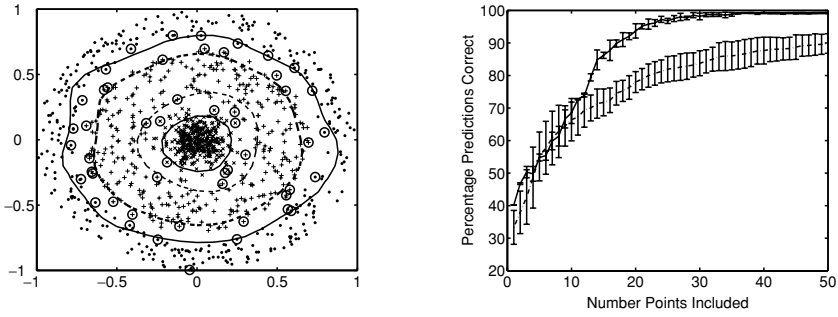
[10] Available online at http://www.ics.uci.edu/~mlearn/MPRepository.html.

variational approximation and the Gibbs sampler. This indicates that the predictions from the Laplace approximation are less informative about the target values than both other methods considered. In addition, the variational approximation yields predictive distributions that are as informative as those provided by the Gibbs sampler; however, the 0-1 prediction errors obtained across all methods do not differ as significantly. Kuss and then Rasmussen (2005) made a similar observation for the binary GP classification problem when Laplace and EP approximations were compared to MCMC. It will be interesting to further compare EP and variational approximations in this setting.

We have observed that the predictions obtained from the variational approximation are in close agreement with those of MCMC, while the Laplace approximation suffers from some inaccuracy. This has also been reported for the binary classification setting in Kuss and Rasmussen (2005).

6.3 Multiclass Sparse Approximation. A further 1000 samples were drawn from the toy data generating process already described, and these were used to illustrate the sparse GP multiclass classifier in operation. The posterior mean values of the shared covariance kernel parameters estimated in the previous example were employed here, and so the covariance kernel parameters were not estimated. The predictive posterior scoring criterion proposed in section 6 was employed in selecting points for inclusion in the overall model. To assess how effective this criterion is, random sampling was also employed to compare the rates of convergence of both inclusion strategies in terms of predictive 0-1 loss on a held-out test set of 2385 samples. A maximum of $S = 50$ samples was to be included in the model defining a 95% sparsity level.

In Figure 5a the first two dimensions of the 1000 samples are plotted with the three different target classes denoted by symbols. The isocontours of constant target posterior probability at a level of one-third (the decision boundaries) for each of the three classes are shown by the solid and dashed lines. What is interesting is that the 50 included points (circled) all sit close to, or on, the corresponding decision boundaries as would be expected given the selection criteria proposed. These can be considered as a probabilistic analog to the support vectors of an SVM. The rates of 0-1 error convergence using both random and informative point sampling are shown in Figure 5b. The procedure was repeated 20 times using the same data samples, and the error bars show one standard deviation over these repeats. It is clear that on this example at least, random sampling has the slowest convergence, and the informative point inclusion strategy achieves less than 1% predictive error after the inclusion of only 30 data points. Of course we should bridle our enthusiasm by recalling that the estimated covariance kernel parameters are already supplied. Nevertheless, multiclass IVM makes Bayesian GP inference on large-scale problems with multiple classes feasible, as will be demonstrated in the following example.

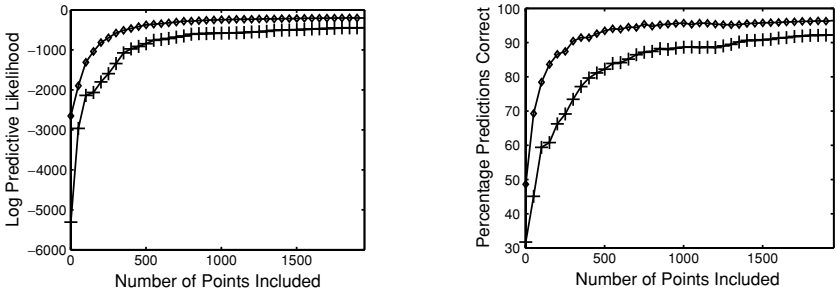


(a) (b)

Figure 5: (a) Scatter plot of the first two dimensions of the 1000 available data sample. Each class is denoted by $\times, +, \bullet$ and the decision boundaries denoted by the contours of target posterior probability equal to one-third are plotted by solid and dashed line. The 50 points selected based on the proposed criterion are circled, and it is clear that these sit close to the decision boundaries. (b) The averaged predictive performance (percentage predictions correct) over 20 random starts (dashed line denotes random sampling and solid line denotes informative sampling) is shown, with the slowest converging plot characterizing what is achieved under a random sampling strategy.

6.4 Large-Scale Example of Sparse GP Multiclass Classification. The Isolet data set[11] comprises of 6238 examples of letters from the alphabet spoken in isolation by 30 individual speakers, and each letter is represented by 617 features. An independent collection of 1559 spoken letters is available for classification test purposes. The best reported test performance over all 26 classes of letter was 3.27% error achieved using 30-bit error correcting codes with an artificial neural network. Here we employ a single RBF covariance kernel with a common inverse length scale of 0.001 (further fine tuning is of course possible), and a maximum of 2000 points from the available 6238 are to be employed in the sparse multiclass GP classifier. As in the previous example, data are standardized; both random and informative sampling strategies were employed, with the results given in Figure 6 illustrating the superior convergence of an informative sampling strategy. After including 2000 of the available 6238 samples in the model, under the informative sampling strategy, a test error rate of 3.52% is achieved. We are unaware of any multiclass GP classification method that has been applied to such a large-scale problem in terms of both data samples available and the number of classes.

[11] The data set is available online from http://www.ics.uci.edu/mlearn/databases/isolet.



(a) (b)

Figure 6: (a) The predictive likelihood computed on held-out data for both random sampling (solid line with + markers) and informative sampling (solid line with ♦ markers). The predictive likelihood is computed once every 50 inclusion steps. (b) The predictive performance (percentage predictions correct) achieved for both random sampling (solid line with + markers) and informative sampling (solid line with ♦ markers)

Qi, Minka, Picard, and Ghahramani (2004) have presented an empirical study of ARD when employed to select basis functions in relevance vector machine (RVM) (Tipping, 2000) classifiers. It was observed that a reliance on the marginal likelihood alone as a criterion for model identification ran the risk of overfitting the available data sample by producing an overly sparse representation. The authors then employ an approximation to the leave-one-out error, which emerges from the EP iterations, to counteract this problem. For Bayesian methods that rely on optimizing in-sample marginal likelihood (or an appropriate bound), great care has to be taken when setting the convergence tolerance, which determines when the optimization routine should halt. However, in the experiments we have conducted, this phenomenon did not appear to be such a problem with the exception of one data set, discussed in the following section.

6.5 Comparison with Multiclass SVM. To briefly compare the performance of the proposed approach to multiclass classification with a number of multiclass SVM methods, we consider the recent study of Duan and Keerthi (2005). In that work, four forms of multiclass classifier were considered: WTAS (one-versus-all SVM method with winner-takes-all class selection, MWVS (one-versus-one SVM with a maximum votes class selection strategy, PWCK (one-versus-one SVM with probabilistic outputs employing pairwise coupling; see Duan & Keerthi, 2005), for details, and PWCK (kernel logistic regression with pairwise coupling of binary outputs). Five multiclass data sets from the UCI Machine Learning Data Repository were employed: ABE (16 dimensions and 3 classes), a subset of the Letters data

Table 2: SVM and Variational Bayes GP Multiclass Classification Comparison.

	WTAS	MWVS	PWCP	PWCK	VBGPM	VBGPS
SEG	9.4 ± 0.5	$\mathbf{7.9 \pm 1.2}$	$\mathbf{7.9 \pm 1.2}$	$\mathbf{7.5 \pm 1.2}$	$*\mathbf{7.8 \pm 1.5}$	11.5 ± 1.2
DNA	10.2 ± 1.3	9.9 ± 0.9	$\mathbf{8.9 \pm 0.8}$	9.7 ± 0.7	74.0 ± 0.3	13.3 ± 1.3
ABE	$\mathbf{1.9 \pm 0.8}$	$\mathbf{1.9 \pm 0.6}$	$\mathbf{1.8 \pm 0.6}$	$\mathbf{1.8 \pm 0.6}$	$*\mathbf{1.8 \pm 0.8}$	2.4 ± 0.8
WAV	17.2 ± 1.4	17.8 ± 1.4	$\mathbf{16.4 \pm 1.4}$	$\mathbf{15.6 \pm 1.1}$	25.2 ± 1.2	$*\mathbf{15.6 \pm 0.7}$
SAT	$\mathbf{11.1 \pm 0.6}$	$\mathbf{11.0 \pm 0.7}$	$\mathbf{10.9 \pm 0.4}$	11.2 ± 0.6	12.0 ± 0.4	12.1 ± 0.4

Note: The asterisk highlights the cases where the proposed GP-based multiclass classifiers were part of the best-performing set.

set using the letters A, B, and E; DNA (180 dimensions and 3 classes); SAT (36 dimensions and 6 classes)—Satellite Image; SEG (18 dimensions and 7 classes)—Image Segmentation; and WAV (21 dimensions and 3 classes)—Waveform. For each of these, Duan and Keerthi (2005) created 20 random partitions into training and test sets for three different sizes of training set, ranging from small to large. Here we consider only the smallest training set sizes.

In Duan and Keerthi (2005) thorough and extensive cross validation was employed to select the length-scale parameters (single) of the gaussian kernel and the associated regularization parameters used in each of the SVMs. The proposed importance sampler is employed to obtain the posterior mean estimates for both single and multiple length scales—VBGPS (variational bayes gaussian process classification) for the single-length scale and VBGPM (variational bayes gaussian process classification), for Multiple-length scales—for a common GP covariance shared across all classes. We monitor the bound on the marginal and consider that convergence has been achieved when less than a 1% increase in the bound is observed for all data sets except for ABE where a 10% convergence criterion was employed due to a degree of overfitting being observed after this point. In all experiments, data were standardized to have zero mean and unit variance.

The percentage test errors averaged over each of the 20 data splits (mean $\pm$ standard deviation) are reported in Table 2. For each data set the classifiers that obtained the lowest prediction error and whose performances were indistinguishable from each other at the 1% significance level using a paired t-test are highlighted in bold. An asterisk highlights the cases where the proposed GP-based multiclass classifiers were part of the best-performing set. We see that in three of the five data sets, performance equal to the best-performing SVMs is achieved by one of the GP-based classifiers without recourse to any cross validation or in-sample tuning with comparable performance being achieved for SAT and DNA. The performance of VBGPM is particularly poor on DNA, and this is possibly due to the large number (180) of binary features.

7 Conclusion and Discussion

The main novelty of this work has been to adopt the data augmentation strategy employed in obtaining an exact Bayesian analysis of binary and multinomial probit regression models for GP-based multiclass (of which binary is a specific case) classification. While a full Gibbs sampler can be straightforwardly obtained from the joint likelihood of the model, approximate inference employing a factored form for the posterior is appealing from the point of view of computational effort and efficiency. The variational Bayes procedures developed provide simple iterations due to the inherent decoupling effect of the auxiliary variable between the GP components related to each class. The scaling is still dominated by an $\mathcal{O}(N^3)$ term due to the matrix inversion required in obtaining the posterior mean for the GP variables and the repeated computing of multivariate gaussians required for the weights in the importance sampler. However, with the simple decoupled form of the posterior updates, we have shown that ADF-based online and sparse estimation yields a full multiclass IVM that has linear scaling in the number of classes and the number of available data points, and this is achieved in a straightfoward manner. An empirical comparison with full MCMC suggests that the variational approximation proposed is superior to a Laplace approximation. Further ongoing work includes an investigation into the possible equivalences between EP and variational-based approximate inference for the multiclass GP classification problem as well as developing a variational treatment to GP-based ordinal regression (Chu & Ghahramani, 2005).

Appendix A

A.1 $Q(\mathbf{M})$. We employ the shorthand $Q(\boldsymbol{\varphi}) = \prod_k Q(\boldsymbol{\varphi}_k)$ in the following. Consider the $Q(\mathbf{M})$ component of the approximate posterior. We have

$$\begin{aligned} Q(\mathbf{M}) &\propto \exp\left\{ E_{Q(\mathbf{Y})Q(\boldsymbol{\varphi})} \left(\sum_n \sum_k \log p(y_{nk}|m_{nk}) + \log p(\mathbf{m}_k|\boldsymbol{\varphi}_k) \right) \right\} \\ &\propto \exp\left\{ E_{Q(\mathbf{Y})Q(\boldsymbol{\varphi})} \left(\sum_k \log \mathcal{N}_{\mathbf{y}_k}(\mathbf{m}_k, \mathbf{I}) + \log \mathcal{N}_{\mathbf{m}_k}(\mathbf{0}|\mathbf{C}_{\boldsymbol{\varphi}_k}) \right) \right\} \\ &\propto \prod_k \mathcal{N}_{\tilde{\mathbf{y}}_k}(\mathbf{m}_k, \mathbf{I}) \mathcal{N}_{\mathbf{m}_k}\left(\mathbf{0}, \left(\widetilde{\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}} \right)^{-1} \right), \end{aligned}$$

and so we have

$$Q(\mathbf{M}) = \prod_{k=1}^{K} Q(\mathbf{m}_k) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{m}_k}(\tilde{\mathbf{m}}_k, \Sigma_k),$$

where $\Sigma_k = \left(\mathbf{I} + \widetilde{\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}}\right)^{-1}$ and $\widetilde{\mathbf{m}}_k = \Sigma_k \widetilde{\mathbf{y}}_k$. Now each element of $\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}$ is a nonlinear function of $\boldsymbol{\varphi}_k$, and so, if considered appropriate, a first-order approximation can be made to the expectation of the matrix inverse such that $\widetilde{\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}} \approx \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{-1}$, in which case $\Sigma_k = \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k}(\mathbf{I} + \mathbf{C}_{\widetilde{\boldsymbol{\varphi}}_k})^{-1}$.

A.2 $Q(\mathbf{Y})$

$$
\begin{aligned}
Q(\mathbf{Y}) &\propto \exp\left\{ E_{Q(\mathbf{M})}\left(\sum\nolimits_n \log p(t_n|\mathbf{y}_n) + \log p(\mathbf{y}_n|\mathbf{m}_n) \right) \right\} \\
&\propto \exp\left\{ \sum\nolimits_n \log p(t_n|\mathbf{y}_n) + \log \mathcal{N}_{\mathbf{y}_n}(\widetilde{\mathbf{m}}_n|\mathbf{I}) \right\} \\
&\propto \prod\nolimits_n \mathcal{N}_{\mathbf{y}_n}(\widetilde{\mathbf{m}}_n, \mathbf{I})\, \delta(y_{ni} > y_{nk}\ \forall\, k \neq i)\delta(t_n = i).
\end{aligned}
$$

Each $\mathbf{y}_n$ is then distributed as a truncated multivariate gaussian such that for $t_n = i$, the ith dimension of $\mathbf{y}_n$ is always the largest, and so we have

$$
Q(\mathbf{Y}) = \prod_{n=1}^{N} Q(\mathbf{y}_n) = \prod_{n=1}^{N} \mathcal{N}_{\mathbf{y}_n}^{t_n}(\widetilde{\mathbf{m}}_n, \mathbf{I}),
$$

where $\mathcal{N}_{\mathbf{y}_n}^{t_n}(.,.)$ denotes a K-dimensional gaussian truncated such that the dimension indicated by the value of t_n is always the largest.

The posterior expectation of each $\mathbf{y}_n$ is now required. Note that

$$
Q(\mathbf{y}_n) = \mathcal{Z}_n^{-1} \prod_k \mathcal{N}_{y_{nk}}(\widetilde{m}_{nk}, 1),
$$

where $\mathcal{Z}_n = Pr(\mathbf{y}_n \in \mathcal{C})$ and $\mathcal{C} = \{\mathbf{y}_n : y_{nj} < y_{ni},\ j \neq i\}$. Now

$$
\begin{aligned}
\mathcal{Z}_n &= Pr(\mathbf{y}_n \in \mathcal{C}) \\
&= \int_{-\infty}^{+\infty} \mathcal{N}_{y_{ni}}(\widetilde{m}_{ni}, 1) \prod_{j \neq i} \int_{-\infty}^{y_{ni}} \mathcal{N}_{y_{nj}}(\widetilde{m}_{nj}, 1) dy_{ni} dy_{nj} \\
&= E_{p(u)} \left\{ \prod_{j \neq i} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj}) \right\},
\end{aligned}
$$

where u is a standardized gaussian random variable such that $p(u) = \mathcal{N}_u(0, 1)$. For all $k \neq i$, the posterior expectation follows as

$$
\widetilde{y}_{nk} = \mathcal{Z}_n^{-1} \int_{-\infty}^{+\infty} y_{nk} \prod_{j=1}^{K} \mathcal{N}_{y_{nj}}(\widetilde{m}_{nj}, 1) dy_{nj}
$$

$$= \mathcal{Z}_n^{-1} \int_{-\infty}^{+\infty} \int_{-\infty}^{y_{ni}} y_{nk} \mathcal{N}_{y_{nk}}(\widetilde{m}_{nk}, 1) \prod_{j \neq i,k} \mathcal{N}_{y_{ni}}(\widetilde{m}_{ni}, 1) \Phi(y_{ni} - \widetilde{m}_{nj}) dy_{ni} dy_{nk}$$

$$= \widetilde{m}_{nk} - \mathcal{Z}_n^{-1} E_{p(u)} \left\{ \mathcal{N}_u(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1) \prod_{j \neq i,k} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj}) \right\}.$$

The required expectation for the ith component follows as

$$\widetilde{y}_{ni} = \mathcal{Z}_n^{-1} \int_{-\infty}^{+\infty} y_{ni} \mathcal{N}_{y_{ni}}(\widetilde{m}_{ni}, 1) \prod_{j \neq i} \Phi(y_{ni} - \widetilde{m}_{nj}) dy_{ni}$$

$$= \widetilde{m}_{ni} + \mathcal{Z}_n^{-1} E_{p(u)} \left\{ u \prod_{j \neq i} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj}) \right\}$$

$$= \widetilde{m}_{ni} + \sum_{k \neq i} (\widetilde{m}_{nk} - \widetilde{y}_{nk}).$$

The final expression in the above follows from noting that for a random variable $u \sim \mathcal{N}(0, 1)$ and any differentiable function $g(u)$, $E\{ug(u)\} = E\{g'(u)\}$, in which case

$$E_{p(u)} \left\{ u \prod_{j \neq i} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj}) \right\}$$

$$= \sum_{k \neq i} E_{p(u)} \left\{ \mathcal{N}_u(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1) \prod_{j \neq i} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj}) \right\}.$$

A.3 $Q(\boldsymbol{\varphi}_k)$. For each k, we obtain the posterior component

$$Q(\boldsymbol{\varphi}_k) \propto \exp\left\{ E_{Q(\mathbf{m}_k)Q(\boldsymbol{\psi}_k)} \left(\log p(\mathbf{m}_k | \boldsymbol{\varphi}_k) + \log p(\boldsymbol{\varphi}_k | \boldsymbol{\psi}_k) \right) \right\}$$

$$= \mathcal{Z}_k \mathcal{N}_{\widetilde{\mathbf{m}}_k}(\mathbf{0} | \mathbf{C}_{\boldsymbol{\varphi}_k}) \prod_d \mathrm{Exp}_{\varphi_{kd}}(\widetilde{\psi}_{kd}),$$

where $\mathcal{Z}_k$ is the corresponding normalizing constant for each posterior, which is unobtainable in closed form. As such, the required expectations can be obtained by importance sampling.

A.4 $Q(\boldsymbol{\psi}_k)$. The final posterior component required is

$$Q(\boldsymbol{\psi}_k) \propto \exp\left\{ E_{Q(\boldsymbol{\varphi}_k)} \left(\log p(\boldsymbol{\varphi}_k | \boldsymbol{\psi}_k) + \log p(\boldsymbol{\psi}_k | \boldsymbol{\alpha}_k) \right) \right\}$$

$$\propto \prod_d \mathrm{Exp}_{\widetilde{\varphi}_{kd}}(\psi_{kd})\Gamma_{\psi_{kd}}(\sigma_k, \tau_k)$$
$$= \prod_d \Gamma_{\psi_{kd}}(\sigma_k + 1, \tau_k + \widetilde{\varphi}_{kd}),$$

and the required posterior mean values follow as $\widetilde{\psi}_{kd} = \frac{\sigma_k+1}{\tau_k+\widetilde{\varphi}_{kd}}$.

Appendix B

The predictive distribution for a new point $\mathbf{x}_{new}$ can be obtained by first marginalizing the associated GP random variables such that

$$p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \int p(\mathbf{y}^{new}|\mathbf{m}^{new})\, p(\mathbf{m}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) d\mathbf{m}^{new}$$
$$= \prod_{k=1}^{K} \int \mathcal{N}_{m_k^{new}}(y_k^{new}, 1)\mathcal{N}_{m_k^{new}}(\widetilde{m}_k^{new}, \widetilde{\sigma}_k^{new}) d m_k^{new}$$
$$= \prod_{k=1}^{K} \mathcal{N}_{y_k^{new}}(\widetilde{m}_k^{new}, \widetilde{v}_k^{new}),$$

where the shorthand $\widetilde{v}_k^{new} = \sqrt{1 + \widetilde{\sigma}_{k,new}^2}$ is employed. Now that we have the predictive posterior for the auxiliary variable $\mathbf{y}^{new}$, the appropriate conic truncation of this spherical gaussian yields the required distribution $P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})$ as follows. Using the following shorthand, $P(t_{new} = k|\mathbf{y}^{new}) = \delta(y_k^{new} > y_i^{new}\ \forall\ i \neq k)\delta(t_{new} = k) \equiv \delta_{k,new}$, then

$$P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \int P(t_{new} = k|\mathbf{y}^{new})\, p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) d\mathbf{y}^{new}$$
$$= \int_{\mathcal{C}_k} p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) d\mathbf{y}^{new}$$
$$= \int \delta_{k,new} \prod_{k=1}^{K} \mathcal{N}_{y_k^{new}}(\widetilde{m}_k^{new}, \widetilde{v}_k^{new})\, d y_k^{new}$$
$$= E_{p(u)} \left\{ \prod_{j \neq k} \Phi\left(\frac{1}{\widetilde{v}_j^{new}} \left[u\widetilde{v}_k^{new} + \widetilde{m}_k^{new} - \widetilde{m}_j^{new} \right] \right) \right\}.$$

This is the probability that the auxiliary variable $\mathbf{y}^{new}$ is in the cone $\mathcal{C}_k$, so

$$\sum_{k=1}^{K} P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \sum_{k=1}^{K} \int_{\mathcal{C}_k} p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) d\mathbf{y}^{new}$$
$$= \int_{\mathbb{R}^K} p(\mathbf{y}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) d\mathbf{y}^{new} = 1,$$

thus yielding a properly normalized posterior distribution over classes $1, \ldots, K$.

Appendix C

The variational bound conditioned on the current values of $\boldsymbol{\varphi}_k, \boldsymbol{\psi}_k, \boldsymbol{\alpha}_k$ (assuming these are fixed values) can be obtained in the following manner using the expansion of the relevant components of the lower bound:

$$\sum_k \sum_n E_{Q(\mathbf{M})Q(\mathbf{Y})} \left\{ \log p(y_{nk}|m_{nk}) \right\} \tag{C.1}$$

$$+ \sum_k E_{Q(\mathbf{M})} \left\{ \log p(\mathbf{m}_k|\mathbf{X}, \boldsymbol{\varphi}_k) \right\} \tag{C.2}$$

$$- \sum_k E_{Q(\mathbf{m}_k)} \{ \log Q(\mathbf{m}_k) \} \tag{C.3}$$

$$- \sum_n E_{Q(\mathbf{y}_n)} \{ \log Q(\mathbf{y}_n) \}. \tag{C.4}$$

Expanding each component in turn obtains

$$- \frac{1}{2} \sum_k \sum_n \left\{ \widetilde{y^2}_{nk} + \widetilde{m^2}_{nk} - 2\widetilde{y}_{nk}\widetilde{m}_{nk} \right\} - \frac{NK}{2} \log 2\pi \tag{C.5}$$

$$- \frac{1}{2} \sum_k \log |\mathbf{C}_{\boldsymbol{\varphi}_k}| - \frac{1}{2} \sum_k \widetilde{\mathbf{m}}_k^{\mathrm{T}} \mathbf{C}_{\boldsymbol{\varphi}_k}^{-1} \widetilde{\mathbf{m}}_k$$

$$- \frac{1}{2} \sum_k trace \left\{ \mathbf{C}_{\boldsymbol{\varphi}_k}^{-1} \Sigma_k \right\} - \frac{NK}{2} \log 2\pi \tag{C.6}$$

$$- \frac{NK}{2} - \frac{NK}{2} \log 2\pi - \frac{1}{2} \sum_k \log |\Sigma_k| \tag{C.7}$$

$$- \frac{1}{2} \sum_k \sum_n \left\{ \widetilde{y^2}_{nk} + \widetilde{m}_{nk}^2 - 2\widetilde{y}_{nk}\widetilde{m}_{nk} \right\} - \sum_n \log \mathcal{Z}_n - \frac{N}{2} \log 2\pi. \tag{C.8}$$

Combining and manipulating equations C.5 to C.8 gives the following expression for the lower bound,

$$-\frac{NK}{2}\log 2\pi + \frac{N}{2}\log 2\pi + \frac{NK}{2} - \frac{1}{2}\sum_k trace\{\Sigma_k\}$$

$$-\frac{1}{2}\sum_k \widetilde{\mathbf{m}}_k^{\mathrm{T}}\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}\widetilde{\mathbf{m}}_k - \frac{1}{2}\sum_k trace\left\{\mathbf{C}_{\boldsymbol{\varphi}_k}^{-1}\Sigma_k\right\}$$

$$-\frac{1}{2}\sum_k \log|\mathbf{C}_{\boldsymbol{\varphi}_k}| + \frac{1}{2}\sum_k \log|\Sigma_k| + \sum_n \log \mathcal{Z}_n,$$

where each $\mathcal{Z}_n = E_{p(u)}\left\{\prod_{j\neq i}\Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj})\right\}$.

Appendix D

Details of the Gibbs sampler required to obtain samples from the posterior $p(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$ now follow. From the definition of the joint likelihood (see equation 3.2) it is straightforward to see that the conditional distribution for each $\mathbf{y}_n|\mathbf{m}_n$ will be a truncated gaussian defined in the cone $\mathcal{C}_{t_n}$, centered at $\mathbf{m}_n$ with identity covariance, and denoted by $\mathcal{N}_{\mathbf{y}}^{t_n}(\mathbf{m}_n, \mathbf{I})$. The distribution for each $\mathbf{m}_k|\mathbf{y}_k$ is multivariate gaussian with covariance $\Sigma_k = \mathbf{C}_{\boldsymbol{\varphi}_k}(\mathbf{I} + \mathbf{C}_{\boldsymbol{\varphi}_k})^{-1}$ and mean $\Sigma_k\mathbf{y}_k$. Thus, the Gibbs sampler, for each n and k, takes the simple form

$$\mathbf{y}_n^{(i)}|\mathbf{m}_n^{(i-1)} \sim \mathcal{N}_{\mathbf{y}}^{t_n}(\mathbf{m}_n^{(i-1)}, \mathbf{I})$$

$$\mathbf{m}_k^{(i+1)}|\mathbf{y}_k^{(i)} \sim \mathcal{N}_{\mathbf{m}}(\Sigma_k\mathbf{y}_k^{(i)}, \Sigma_k),$$

where the superscript (i) denotes the ith sample drawn. The dominant scaling will be $\mathcal{O}(KN^3)$ per sample draw. With the multinomial probit likelihood for a new data point defined as

$$P(t_{new} = k|\mathbf{m}^{new}) = E_{p(u)}\left\{\prod_{j\neq k}\Phi(u + m_k^{new} - m_j^{new})\right\},$$

the predictive distribution[12] is then obtained from

$$P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t}) = \int P(t_{new} = k|\mathbf{m}^{new})\, p(\mathbf{m}^{new}|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})\, d\mathbf{m}^{new}.$$

A Monte Carlo estimate of the above required marginal posterior expectation can be obtained by drawing samples from the full posterior

[12] Conditioning on $\boldsymbol{\Phi}$ and $\boldsymbol{\alpha}$ is implicit.

distribution, $p(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$, using the above sampler. Then for each $\boldsymbol{\Theta}^{(i)}$ sampled, an additional set of samples $m_k^{new,s}$ is drawn, such that for each k, $m_k^{new,s}|\mathbf{y}_k^{(i)} \sim \mathcal{N}_m(\mu_k^{new,i}, \sigma_{k,new}^2)$, where $\mu_k^{new,i} = (\mathbf{y}_k^{(i)})^{\mathrm{T}} \left(\mathbf{I} + \mathbf{C}_{\varphi_k}\right)^{-1} \mathbf{C}_{\boldsymbol{\varphi}_k}^{new}$, and the associated variance is $\sigma_{k,new}^2 = c_{\boldsymbol{\varphi}_k}^{new} - (\mathbf{C}_{\boldsymbol{\varphi}_k}^{new})^{\mathrm{T}} \left(\mathbf{I} + \mathbf{C}_{\varphi_k}\right)^{-1} \mathbf{C}_{\boldsymbol{\varphi}_k}^{new}$. The approximate predictive distribution can then be obtained by the following Monte Carlo estimate:

$$\frac{1}{N_{samps}} \sum_{s=1}^{N_{samps}} E_{p(u)} \left\{ \prod_{j \neq k} \Phi(u + m_k^{new,s} - m_j^{new,s}) \right\}.$$

An additional Metropolis-Hastings subsampler can be employed within the above Gibbs sampler to draw samples from the posterior $p(\boldsymbol{\Theta}, \boldsymbol{\Phi}|\mathbf{t}, \mathbf{X}, \boldsymbol{\alpha})$ if the covariance function hyperparameters are to be integrated out.

Appendix E

The Laplace approximation requires the Hessian matrix of second-order derivatives of the joint log likelihood with respect to each $\mathbf{m}_n$. The derivatives of the noise component, $\log P(t_n = k|\mathbf{m}_n) = \log E_{p(u)} \times \{\prod_{j \neq k} \Phi(u + \widetilde{m}_{nk} - \widetilde{m}_{nj})\}$, follow, where we denote expectation with respect to a gaussian truncated in the cone $\mathcal{C}_k$ as $E_{\mathcal{N}_{\mathbf{y}}^k}\{\cdot\}$:

$$\begin{aligned} \frac{\partial}{\partial m_{ni}} \log P(t_n = k|\mathbf{m}_n) &= \frac{1}{P(t_n = k|\mathbf{m}_n)} \int_{\mathcal{C}_k} (y_{ni} - m_{ni}) \mathcal{N}_{\mathbf{y}_n}(\mathbf{m}, \mathbf{I}) d\mathbf{y} \\ &= E_{\mathcal{N}_{\mathbf{y}}^k}\{y_{ni}\} - m_{ni} \end{aligned}$$

and

$$\frac{\partial^2}{\partial m_{nj} \partial m_{ni}} \log P(t_n = k|\mathbf{m}_n) = E_{\mathcal{N}_{\mathbf{y}}^k}\{y_{ni} y_{nj}\} - E_{\mathcal{N}_{\mathbf{y}}^k}\{y_{ni}\} E_{\mathcal{N}_{\mathbf{y}}^k}\{y_{nj}\} - \delta_{ij}.$$

This then defines an $NK \times NK$–dimensional Hessian matrix that, unlike the Hessian of the multinomial logit counterpart, cannot be decomposed into a diagonal plus multiplicative form (refer to Williams & Barber, 1998, for details), due to the cross-diagonal elements $E_{\mathcal{N}_{\mathbf{y}}^k}\{y_{ni} y_{nj}\}$, and so the required matrix inversions of the Newton step and those required to obtain the predictive covariance will operate on a full $NK \times NK$ matrix.

Acknowledgments

This work is supported by Engineering and Physical Sciences Research Council grants GR/R55184/02 & EP/C010620/1. We are grateful to

Chris Williams, Jim Kay, and Joaquin Quiñonero-Candela for motivating discussions regarding this work. In addition, the comments and suggestions made by the anonymous reviewers helped to improve the manuscript significantly.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistcial Association, 88*(422), 669–679.

Beal, M. (2003). *Variational algorithms for approximate bayesian inference*. Unpublished doctoral dissertation, University College London.

Chu, W., & Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research, 6*, 1019–1041.

Csato, L., Fokue, E., Opper, M., Schottky, B., & Winther, O. (2000). Efficient approaches to gaussian process classification. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 252–257). Cambridge, MA: MIT Press.

Csato, L., & Opper, M. (2002). Sparse online gaussian processes. *Neural Computation, 14*, 641–668.

Duan, K., & Keerthi, S. (2005). Which is the best multi-class SVM method? An empirical study. In N. C. Oza, R. Polikar, J. Kittler, & F. Roli (Eds.), *Proceedings of the Sixth International Workshop on Multiple Classifier Systems* (pp. 278–285). Seaside, CA.

Gibbs, M. N., & MacKay, D. J. C. (2000). Variational gaussian process classifiers. *IEEE Transactions on Neural Networks, 11*(6), 1458–1464.

Girolami, M., & Rogers, S. (2005). Hierarchic Bayesian models for kernel learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 241–248). New York: ACM.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*, 183–233.

Kim, H. C. (2005). *Bayesian and ensemble kernel classifiers*. Unpublished doctoral dissertation, Pohang University of Science and Technology. Available online at http://home.postech.ac.kr/∼grass/publication/.

Kuss, M., & Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Research, 6*, 1679–1704.

Lawrence, N. D., Milo, M., Niranjan, M., Rashbass, P., & Soullier, S. (2004). Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics, 20*(4), 518–526.

Lawrence, N. D, Platt, J. C., & Jordan, M. I. (2005). Extensions of the informative vector machine. In J. Winkler, N. D. Lawrence, & M. Niranjan (Eds.), *Proceedings of the Sheffield Machine Learning Workshop*. Berlin: Springer-Verlag.

Lawrence, N. D., Seeger, M., & Herbrich, R. (2003). Fast sparse gaussian process methods: The informative vector machine. In S. Thrun, S. Becker, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15*. Cambridge, MA: MIT Press.

MacKay, D. J. C (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Neal, R. (1998). Regression and classification using gaussian process priors. In A. P. Dawid, M. Bernardo, J. O. Berger, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (pp. 475–501). New York: Oxford University Press.

Opper, M., & Winther, O. (2000). Gaussian processes for classification: Mean field algorithms. *Neural Computation, 12*, 2655–2684.

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 547–553). Cambridge, MA: MIT Press.

Qi, Y., Minka, T. P., Picard, R. W., & Ghahramani, Z. (2004). Predictive automatic relevance determination by expectation propagation. In R. Greiner & D. Schuurmans (Eds.), *Proceedings of the Twenty-First International Conference on Machine Learning*. New York: ACM.

Quinonero-Candela, J., & Winther, O. (2003). Incremental gaussian processes. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15*. Cambridge, MA: MIT Press.

Seeger, M. (2000). Bayesian model selection for support vector machines, gaussian processes and other kernel classifiers. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing Systems, 12* (pp. 603–609). Cambridge, MA: MIT Press.

Seeger, M., & Jordan, M. I. (2004). *Sparse gaussian Process classification with multiple classes* (Tech. Rep. 661). Berkeley: Department of Statistics, University of California.

Seeger, M., Williams, C. K. I., & Lawrence, N. D. (2003). Fast forward selection to speed up sparse gaussian process regression. In C. M. Bishop, & B. J. Frey (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Np: Society for Artificial Intelligence and Statistics.

Tipping, M. (2000). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211–244.

Wang, B., & Titterington, D. M. (2004). *Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values* (Tech. Rep. No. 04–02). Glasgow: Department of Statistics, University of Glasgow.

Williams, C. K. I., & Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(12), 1342–1351.

Williams, C. K. I., & Rasmussen, C. E. (1996). Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural processing systems, 8* (pp. 598–604). Cambridge, MA: MIT Press.

Williams, C. K. I., & Seeger, M. (2001). Using the Nystrom method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 682–688). Cambridge, MA: MIT Press.

Received July 1, 2005; accepted November 8, 2005.