Pseudo-Marginal Bayesian Multiple-Class Multiple-Kernel Learning for Neuroimaging Data

Andrew D. O'Harney*, Andre Marquand†‡, Katya Rubia‡, Kaylita Chantiluke‡,
Anna Smith‡, Ana Cubillo‡, Camilla Blain‡ and Maurizio Filippone*
*School of Computing Science, University of Glasgow, Glasgow, UK
†Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
‡Institute of Psychiatry, King's College London, London, UK

***Abstract*—In clinical neuroimaging applications where subjects belong to one of multiple classes of disease states and multiple imaging sources are available, the aim is to achieve accurate classification while assessing the importance of the sources in the classification task. This work proposes the use of fully Bayesian multiple-class multiple-kernel learning based on Gaussian Processes, as it offers flexible classification capabilities and a sound quantification of uncertainty in parameter estimates and predictions. The exact inference of parameters and accurate quantification of uncertainty in Gaussian Process models, however, poses a computationally challenging problem. This paper proposes the application of advanced inference techniques based on Markov chain Monte Carlo and unbiased estimates of the marginal likelihood, and demonstrates their ability to accurately and efficiently carry out inference in their application on synthetic data and real clinical neuroimaging data. The results in this paper are important as they further work in the direction of achieving computationally feasible fully Bayesian models for a wide range of real world applications.**

I. Introduction

Kernel based classifiers, such as support vector machines (SVMs) [1] and Gaussian process (GP) classifiers [2], represent a successful class of nonlinear models for predictive classification. Important extensions to these include multiple-kernel learning (MKL) [3] and multiple-class (MC) classification variants. Such variants can determine an instance as belonging to one of multiple classes when considering multiple representations of the same data; this is termed the multiple-class multiple-kernel learning (MC-MKL) problem.

Several formulations of the MC-MKL problem have been proposed in the literature [3]. While most of these algorithms have been successfully employed to tackle a number of challenging problems, there are generally difficulties in incorporating any form of uncertainty in predictions, or in the assessment of the relative importance of different kernels. This can be problematic in applications, such as the ones in clinical neuroimaging that are considered here, where little data is available and a sound quantification of uncertainty is of primary interest. From this point of view, approaches to the MC-MKL problem based on a probabilistic formulation of Gaussian Processes (GPs) have previously been proposed [4]. MC-MKL classifiers based on GPs [2] offer a reliable way to statistically assess the importance of kernels and quantify uncertainty in predictions, while retaining the features that make SVM classifiers successful. Namely, the flexibility to construct non-linear classification boundaries between classes. In neuroimaging applications, however, given the high dimensionality of the problem and the fact that the mapping of the weights in the input space is only exact in the linear case, kernels are particularly appealing as they offer an efficient way to represent similarity between very high dimensional images.

Inferring the parameters (kernel weightings) of such GP classifiers, which is key to reliably quantifying uncertainty, is very challenging. As such, a number of approximate schemes have been proposed that allow for the development of tractable classification models, but can severely affect performance and the ability to reliably quantify uncertainty. This paper applies inference techniques based on Markov chain Monte Carlo (MCMC) methods [5], as they offer asymptotic guarantees of convergence to exact inference, and make it possible to achieve results up to a given precision level in a Monte Carlo sense [6]. Previous MCMC approaches applied to MC-MKL classification using GPs have been proposed in [4], but they are characterized by slow convergence speed and low efficiency. In particular, this is due to the fact that the marginal likelihood of GP classifiers cannot be expressed in closed form. In order to circumvent such an intractability, this paper proposes to extend the Pseudo-Marginal (PM) MCMC approach in [7], where the marginal likelihood is replaced by an unbiased estimate in the Hastings ratio, to the MC-MKL setting. An unbiased estimate of the marginal likelihood is constructed using importance sampling with an importance distribution obtained by the Laplace approximation [2]. By employing this MCMC approach, it is possible to produce samples from the correct posterior distribution over parameters, without the need to actually compute the marginal likelihood exactly.

By allowing for the effective sampling of the posterior over parameters, statistical models that suitably account for uncertainty can be constructed. This development has relevance within a number of fields aiming to tackle MC-MKL problems, and in this paper we demonstrate the applicability of the proposed approach to clinical neuroimaging examples. Within this context an important and emerging objective is to be able to accurately diagnose and predict the future clinical outcome of patients based on a number of imaging modalities (here represented using kernels). This is a challenging problem because neurological and psychiatric diseases are often characterised by overlapping symptom profiles and individual variations in disease progression that make them difficult to discriminate. Thus, a reliable measurement of classification uncertainty is required.

The remainder of this paper is structured as follows.

Sec. II describes GPs within the MC-MKL framework. Sec. III describes the proposed PM MCMC approach. Sec. IV reports the experimental work, and Sec. V concludes the paper.

II. MC-MKL Gaussian Processes

A GP is a set of random variables, any finite set of which is jointly Gaussian [2]. Denote by $f(\mathbf{x})$ the realization of one of these variables at input $\mathbf{x}$. The GP is fully specified by a mean function, say $\mu(\mathbf{x})$, and a covariance function $k(\mathbf{x}, \mathbf{x}')$. The covariance function can be considered as the equivalent of the kernel function in kernel machines. Due to the properties of GPs, given a set of n input vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the corresponding variables $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^{\mathrm{T}}$ are jointly Gaussian with mean $(\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^{\mathrm{T}}$ and covariance matrix with entries $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Our approach to multiple-class (MC) classification using GPs follows [2]. Let $\{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ be a set of n pairs, where $\mathbf{x}_i$ is an input vector and $\mathbf{y}_i$ its associated class. In the case of MC problems, each $\mathbf{y}_i = (y_i^1, \ldots, y_i^C)^{\mathrm{T}}$ represents the assignment of one input to one of C classes; if an instance i belongs to class c, then the c^{th} element of $\mathbf{y}_i$ is 1 and 0 everywhere else. The model assumes that class labels are conditionally independent given a set of class specific latent variables $\mathbf{f} = (f_1^1, \ldots, f_n^1, f_1^2, \ldots, f_n^2, f_1^C, \ldots, f_n^C)^{\mathrm{T}}$. Then, the probability π_i^c that instance i belongs to class c is given by the soft-max transformation:

$$\pi_i^c = \frac{\exp(f_i^c)}{\sum_{c'} \exp(f_i^{c'})} \tag{1}$$

The likelihood function is multinomial with probabilities given by π_i^c. We assume that each of the C class-specific sets of latent variables has a GP prior $\mathcal{N}(\mathbf{f}^c|\mathbf{0}, K^c)$. This assumption implies a GP prior on all latent variables $\mathcal{N}(\mathbf{f}|\mathbf{0}, K)$, with K block diagonal and with K^i on the i^{th} diagonal block. In all, the model can be thought of as hierarchical, where class labels $\mathbf{y}$ are conditioned on the latent variables $\mathbf{f}$, which in turn are conditioned on $\boldsymbol{\theta}$.

A further extension of the classification framework using GPs is represented by the incorporation of different sources in the learning process. Assume that a set of d input sources are available and that, from these, a set of d covariance matrices S_h can be derived. To asses the importance of a given data source in the classification task, we propose to model the covariance associated to the GP prior of each class-specific set of latent variables as a linear combination of the data sources:

$$K^c = \sum_{h=1}^{d} \theta_{ch} S_h \tag{2}$$

In this way, each θ represents a kernel weighting in the classification process, and can thus be used as a measure of how important a given input source is in determining a class.

A. Adaptation of covariance parameters and predictions

For some new input, say $\mathbf{x}_*$, the purpose of the classifier is to predict its label $\mathbf{y}_*$. One standard and widely used approach is to maximize the marginal likelihood with respect to $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left[p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\boldsymbol{\theta}) d\mathbf{f} \right] \tag{3}$$

and approximate the predictive distribution for $\mathbf{y}_*$ as follows:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{y}, \hat{\boldsymbol{\theta}}) = \int p(\mathbf{y}_*|\mathbf{f}_*) p(\mathbf{f}_*|\mathbf{f}, \hat{\boldsymbol{\theta}}) p(\mathbf{f}|\mathbf{y}, \hat{\boldsymbol{\theta}}) d\mathbf{f}_* d\mathbf{f} \tag{4}$$

Unfortunately, the integral with respect to $\mathbf{f}$ in eq. 3 is not analytically tractable. A number of schemes have been proposed within the context of GP models to solve this problem, for example the Laplace Approximation (LA) [2] that we will employ here. The LA scheme for GPs [2] attempts to approximate the posterior over the latent variables $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$ as a Gaussian. This can be achieved by inspection of the second order Taylor expansion around the mode of the logarithm of the posterior over $\mathbf{f}$. The mode can be found using Newton's iterative method; defining the logarithm of the posterior as Ψ, Newton's iterations are based on the following update:

$$\mathbf{f}_{\mathrm{new}} = \mathbf{f} - (\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\mathbf{f}))^{-1} \nabla_{\mathbf{f}} \Psi(\mathbf{f}) \tag{5}$$

starting from $\mathbf{f} = \mathbf{0}$ until convergence.

Let Π be the matrix obtained by stacking by row the C matrices $\mathrm{diag}(\boldsymbol{\pi}^c)$ each of size $n \times n$. Also, let $\boldsymbol{\pi}$ and $\mathbf{y}$ be the vectors obtained by concatenating all the $\boldsymbol{\pi}^c$ and $\mathbf{y}^c$, respectively. The gradient is $\nabla_{\mathbf{f}} \Psi(\mathbf{f}) = -K^{-1}\mathbf{f} + \mathbf{y} - \boldsymbol{\pi}$, while the negative Hessian reads:

$$-\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\mathbf{f}) = K^{-1} + \mathrm{diag}(\boldsymbol{\pi}) - \Pi\Pi^{\mathrm{T}} \tag{6}$$

In the last equation we added an illustration of the structure of the matrices, assuming three classes; grey areas indicate a nonzero value in the corresponding matrix. Newton's iterations require the inversion of the negative Hessian and its multiplication by the gradient. Approaching this calculation naively would lead to a prohibitive computational cost (scaling with the cube of the number of classes as well as with the cube of the number of training data) and the storage of a $(nc) \times (nc)$ matrix. By exploiting the structure of the Hessian, however, it is possible to reduce the computational cost making it linear in the number of classes, and storing at most $n \times n$ matrices. Full details can be found in [2]. Denoting by $q(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \hat{\Sigma})$ the approximating Gaussian we have $\hat{\Sigma}^{-1} = -\nabla_{\mathbf{f}} \nabla_{\mathbf{f}} \Psi(\hat{\mathbf{f}})$.

This approach has two primary drawbacks: (i) parameters are optimized and not inferred, and (ii) optimization is based on a possibly inaccurate approximation to $p(\mathbf{y}|\boldsymbol{\theta})$. In particular, (i) is problematic for interpreting the parameters in a given application and because no uncertainty in parameter estimates is carried forward to predictions, while (ii) may lead to wrong conclusions in the interpretation of model parameters.

B. Bayesian inference of parameters and predictions

A way to avoid the aforementioned limitations would be to carry out a fully Bayesian treatment of the problem where all latent variables and parameters are integrated out, giving a predictive distribution on $\mathbf{y}_*$:

$$p(\mathbf{y}_*|\mathbf{y}) = \int p(\mathbf{y}_*|\mathbf{f}_*) p(\mathbf{f}_*|\mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{f}_* d\mathbf{f} d\boldsymbol{\theta} \tag{7}$$

where we dropped the conditioning on the $\mathbf{x}$'s for simplicity of notation. However, this is problematic from an analytic perspective, given that it requires integrating over the intractable

posterior distribution over $\mathbf{f}$ and $\boldsymbol{\theta}$. A tractable way to solve eq. 7 is through MCMC methods. Denote by $\mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)}$ the i^{th} of N samples drawn from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$. It is then possible to approximate the predictive distribution by means of a Monte Carlo integration:

$$p(\mathbf{y}_*|\mathbf{y}) \simeq \frac{1}{N}\sum_{i=1}^{N}\int p(\mathbf{y}_*|\mathbf{f}_*)p(\mathbf{f}_*|\mathbf{f}^{(i)}, \boldsymbol{\theta}^{(i)})d\mathbf{f}_* \qquad (8)$$

where the remaining integral can be easily approximated using a further Monte Carlo integration. In the limit of long runs, MCMC simulation is known to asymptotically converge to the exact predictive distribution [5].

While MCMC methods allow for a tractable solution of eq. 7, the structure of the MC-MKL model means that there is high coupling between parameters and latent variables. In particular it is not possible to easily sample $\mathbf{f}$ and $\boldsymbol{\theta}$ jointly, so it is necessary to resort to sample schemes where $\mathbf{f}$ and $\boldsymbol{\theta}$ are sampled sequentially from $p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$ and from $p(\boldsymbol{\theta}|\mathbf{f})$. However, this latter sampling scheme is known to be extremely inefficient, as the coupling between parameters and latent variables yields a poor exploration of the posterior distribution over $\boldsymbol{\theta}$. While the sampling from $p(\mathbf{f}|\boldsymbol{\theta}, \mathbf{y})$ can be done very efficiently [8], the bottleneck is represented by the sampling from $p(\boldsymbol{\theta}|\mathbf{f})$, as this ties the posterior distribution over parameters to the current values of the latent variables. A number of approaches have been proposed to tackle this problem; prominent methods to decoupling parameters and latent variables within MCMC approaches include reparameterization techniques [8]. Such techniques mitigate the coupling effect, but low sampling efficiency and slow convergence speed are still problematic. Recently, a different approach to solving the coupling effect has been proposed in [7]. In the next section we illustrate this technique applied to MC-MKL problems using GPs.

III. Pseudo-Marginal MC-MKL with GPs

In this section we focus on the sampling from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. Before illustrating the idea behind the proposed approach, we briefly present the Metropolis-Hastings (MH) algorithm [5] used in the present work. Assuming that $p(\mathbf{y}|\boldsymbol{\theta})$ is available in closed form and that a prior $p(\boldsymbol{\theta})$ is assigned to the parameters, the MH algorithm proceeds as follows. The algorithm is initialized randomly from $\boldsymbol{\theta}$. Then, a new set of parameters $\boldsymbol{\theta}$ is proposed, say from $\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})$, and this is accepted or rejected with probability based on the Hastings ratio $[p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')]/[p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})]$. This process is repeated for several iterations. The first iterations (termed the 'burn-in' period) are usually discarded and the rest are used to estimate the predictive probability and to analyze the uncertainty in the estimate of parameters. Given that $p(\mathbf{y}|\boldsymbol{\theta})$ is intractable, this procedure is clearly not viable, and we propose the use of the PM MCMC approach which avoids computing $p(\mathbf{y}|\boldsymbol{\theta})$ exactly and still achieves samples from the exact $p(\boldsymbol{\theta}|\mathbf{y})$.

The PM MCMC approach is based on the results in [9], that show that using an unbiased estimate of the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ in the MH algorithm is enough to ensure the algorithm samples from the exact posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. Denoting by $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ such an unbiased estimate of the marginal likelihood, the acceptance criterion for the MH algorithm becomes:

$$\min\left\{1, \frac{\tilde{p}(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')\pi(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\tilde{p}(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \qquad (9)$$

This effectively gets around the problem of coupling of $\mathbf{f}$ and $\boldsymbol{\theta}$, as latent variables are approximately integrated out of the model, while retaining an exact MCMC scheme. The way we propose to obtain an unbiased estimate of the marginal likelihood is based on importance sampling [5]. Importance sampling is conducted by drawing N_{imp} samples from the approximating distribution $q(\mathbf{f}|\boldsymbol{\theta})$, and estimating:

$$\tilde{p}(\mathbf{y}|\boldsymbol{\theta}) \simeq \frac{1}{N_{\text{imp}}}\sum_{i=1}^{N_{\text{imp}}}\frac{p(\mathbf{y}|\mathbf{f}^{(i)})p(\mathbf{f}^{(i)}|\boldsymbol{\theta})}{q(\mathbf{f}^{(i)}|\boldsymbol{\theta})} \qquad (10)$$

Note that the variance of the importance sampling estimator can affect the efficiency of the proposed PM MCMC approach. In particular, it can happen that one estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ is so large that it makes it difficult for the MCMC approach to accept any new proposal $\boldsymbol{\theta}'$. This effect adds to the difficulty in exploring a potentially large parameter space, so it is important to employ approximations to reduce variance in the estimate of the marginal likelihood. Note also that the only significant extra computational burden, compared to employing an MCMC approach based on approximate marginal likelihoods, is in the drawing of latent variables from $q(\mathbf{f}|\boldsymbol{\theta})$, and this can be done efficiently by exploiting the structure of the precision matrix of $q(\mathbf{f}|\boldsymbol{\theta})$ as discussed next.

A. Sketch of the implementation

The proposed approach requires the sampling from $q(\mathbf{f}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \hat{\Sigma})$. A naive implementation would factorize $\hat{\Sigma}$, or its inverse, leading to the need to store a $(nc) \times (nc)$ matrix, which can be prohibitive even for a reasonably sized data set with only a few classes. In this section we avoid this by sequentially drawing the latent variables pertaining to each class as follows:

$$\mathbf{f}^1 \qquad \mathbf{f}^2|\mathbf{f}^1 \qquad \mathbf{f}^3|\mathbf{f}^2, \mathbf{f}^1 \qquad \dots \qquad \mathbf{f}^C|\mathbf{f}^{C-1}, \dots, \mathbf{f}^1$$

By employing standard identities for marginal and conditional distributions of jointly Gaussian variables, we can approach the drawing of the latent variables by storing only $n \times n$ matrices. Note also that the order in which variables are sampled does not matter. Define $\Lambda = K^{-1} + \text{diag}(\boldsymbol{\pi}) - \text{III}^{\text{T}}$ as the precision of $q(\mathbf{f}|\boldsymbol{\theta})$ at the mode $\hat{\mathbf{f}}$. Denote by $\Lambda_{[i,j]}$ the block i, j of the matrix Λ, and by $\Lambda_{[i:j,m:n]}$ the concatenation of blocks from i to j column-wise and from m to n row-wise.

1) Sampling $\mathbf{f}^1$: Using standard Gaussian identities, the distribution of $\mathbf{f}^1$ is Gaussian with mean $\hat{\mathbf{f}}^1$ and covariance

$$\Lambda_{[1,1]} - \Lambda_{[1,2:C]}\Lambda^{-1}_{[2:C,2:C]}\Lambda_{[2:C,1]} \qquad (11)$$

In the equation we highlighted the blocks of Λ involved in the calculation, assuming a three class problem. The calculation of Σ_{mar} requires the inversion of a potentially large matrix $\Lambda^{-1}_{[2:C,2:C]}$. Following the same derivation of the Newton iterations [2], this can be done by storing only $n \times n$ matrices.

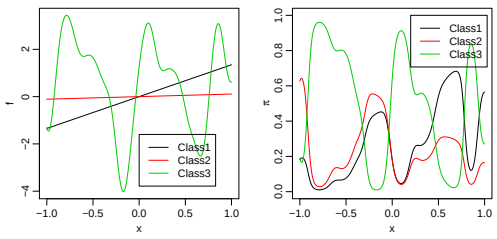

Fig. 1. Left: Latent functions drawn from the multiple-class GP model used to generate the synthetic data set. Right: soft-max transformation of the latent functions to obtain the probabilities of each class.

2) Sampling $\mathbf{f}^r|\mathbf{f}^{r-1},\ldots,\mathbf{f}^1$*:* Define $\mathbf{f}^{[1:(r-1)]}$ as the concatenation of the $r-1$ previously sampled latent variables. The sampling from this conditional distribution can be done by noticing that $\mathbf{f}^r$ is Gaussian with covariance $\Lambda_{[r,r]}$ and mean

$$\hat{\mathbf{f}}^r - \Lambda_{[r,r]}^{-1}\Lambda_{[r,1:(r-1)]}(\mathbf{f}^{[1:(r-1)]} - \hat{\mathbf{f}}^{[1:(r-1)]}) \qquad (12)$$

This requires the factorization of $\Lambda_{[r,r]}$, which is of size $n \times n$.

IV. Experimental results

In this section we present the application of our methodology to one synthetic data set and two real data sets. All experiments rely on the MH algorithm applied to the logarithm of $\boldsymbol{\theta}$. In order to ensure an efficient MCMC sampling scheme we followed common practice and allowed for an adaptive phase where we tuned the MH proposal variance to obtain an acceptance rate between 20% and 30%. To assess the efficiency of our sampler we used the Effective Sample Size (ESS) [10] on 10,000 samples gathered after the adaptive phase for ten parallel chains. The ESS metric is based on autocorrelation, and is used to assess the independence between samples generated in the MCMC sequence. A high ESS relative to the total number of samples shows good efficiency in the scheme, whereas a low ESS shows high correlation between samples, meaning that the sampler is not efficiently exploring the space.

To measure the convergence of a chain on the posterior distribution we employed the Gelman and Rubin shrink factor [11], $\hat{R}$. $\hat{R}$ gives an indication on the convergence of MCMC algorithms (convergence is achieved as $\hat{R}$ approaches 1), and in our experiments it was estimated based on ten parallel chains. We report $\hat{R}$ across 1000, 2000 and 10,000 iterations for the least efficient variable in each case. We also show the average acceptance rate across all chains, as it gives indication of a poor approximation of the marginal likelihood using importance sampling as discussed previously.

A. Synthetic Data

A synthetic three-class data set consisting of 150 instances, equally divided into three classes, was constructed using the combination of two kernels with weights $\theta_{11} = \theta_{21} = \theta_{32} = 2$ and 0 for all other class/kernel combination. The first kernel is derived from an RBF covariance function, while the second was derived from a linear one. Fig. 1 shows the latent functions drawn from the model and their transformation using the soft-max function used to generate the observed labels $\mathbf{y}$. For this analysis we used a $\mathrm{Gamma}(1,1)$ prior on kernel weights and an adaptive phase of 2000 iterations.

TABLE I. Efficiency and convergence measured over 10,000 samples on the synthetic data set.

N_{imp}	ESS		$\hat{R}$			% Acc
	Min (σ)	Max (σ)	@10^3	@$2\cdot10^3$	@10^4	Rate
450	172(33)	734(40)	1.05	1.11	1.01	23.3
100	164(19)	724(85)	1.09	1.02	1.01	23.1
10	166(33)	693(103)	1.13	1.02	1.01	22.3

The posterior distribution over kernel weightings for the synthetic data are shown in fig. 2. The results demonstrate the ability of our method to characterize the posterior distribution over the kernel weighting. Despite the large posterior variance, there is an apparent distinction between kernel weights in the first two classes and the third class, consistently with the kernel parameters that were used to construct the data. Table I shows the efficiency achieved by our method across the synthetic data when varying the number of importance samples. Remarkably, we note that decreasing the number of importance samples has little effect on the efficiency, and that near optimal acceptance rates are achieved across all importance sampling schemes.

B. Parkinsonian Data

The second data set consists of structural magnetic resonance imaging (MRI) data acquired from 62 subjects who were either healthy controls (14 subjects) or patients with one of three akinetic-rigid neurological disorders (multiple system atrophy (MSA, 18 subjects), progressive supranuclear palsy (PSP, 16 subjects), or idiopathic Parkinson's disease (IPD, 14 subjects)). For each subject a T1-weighted structural image was acquired using a spoiled gradient recalled imaging sequence (SPGR). These images were preprocessed using the SPM8 software package (`www.fil.ion.ucl.ac.uk/spm`), which consisted of segmenting the images into different tissue types using the "new segment" routine, then normalized to a standard space using the diffeomorphic anatomical registration using exponential lie algebra (DARTEL) toolbox [12]. Normalised grey matter images were smoothed with an isotropic 6mm smoothing kernel before being parcellated anatomically into six target regions of interest (brainstem, cerebellum, caudate, middle occipital gyrus, putamen, and one for all other brain regions). A linear (dot product) kernel was computed from all voxels in each of the regions, yielding a total of six kernels. Full details on diagnostic criteria, MRI sequence parameters and data preprocessing can be found elsewhere [4], [13]. We use the previous results in [4] as a benchmark for the present work, which used the same data and preprocessing to evaluate different MCMC approaches to classify between the different disorders and to assess the importance of different brain regions in the classification process.

It is noted that in early disease stages it can be difficult to differentiate between the disorders present in this data set, and that misdiagnosis is common [4], [14]. Finding sensitive and specific neurobiological markers of disease type may help with early diagnosis and ultimately improve outcomes for patients, since the treatments currently available are not equally effective in all disorders. As such, providing accurate estimates of predictive confidence are of the utmost importance in this application. An additional requirement is the ability to accurately assess the usefulness of each brain region for

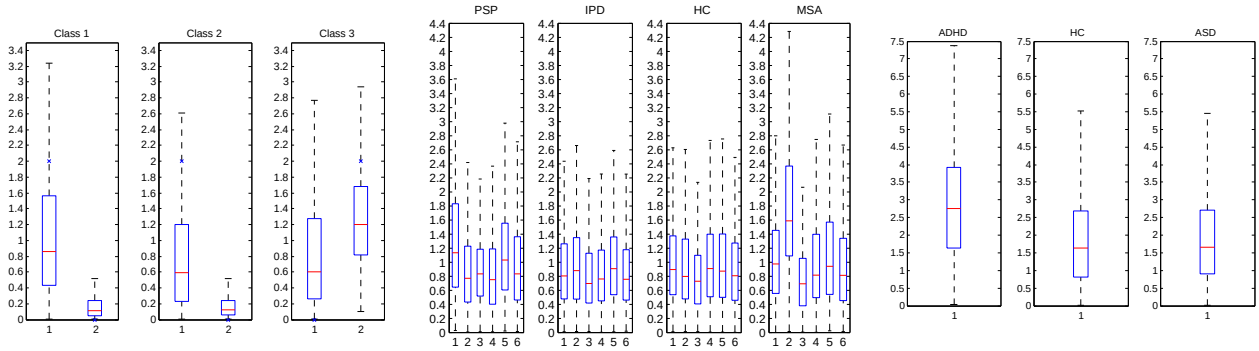

Fig. 2. Left: Posterior distribution of parameters in the synthetic data example. Center: Posterior weight estimates for region data. Regions: (1)Brainstem, (2)Cerebellum, (3)Caudate, (4)Middle occipital gyrus (5)Putamen (6)Other regions. Right: Posterior weight estimates for ADHD and ASD.

TABLE II. EFFICIENCY AND CONVERGENCE ON PARKINSONIAN DATA.

N_{imp}	ESS		$\hat{R}$			% Acc
	Min (σ)	Max (σ)	@10^3	@$2 \cdot 10^3$	@10^4	Rate
450	36.4(11.0)	118.4(13.0)	1.8	1.14	1.07	10.0
100	25.0(14.7)	106.4(42.0)	2.42	1.62	1.40	8.1
10	11.06(6.0)	51.3(14.3)	1.89	1.60	1.24	7.3

TABLE III. RESULTS ON ADHD AND ASD DATA.

Method	N_{imp}	ESS		$\hat{R}$		% Acc
		Min (σ)	Max (σ)	@10^3	@10^4	Rate
PM	450	436(111)	677(130)	1.15	1.00	30.9
	100	378(79)	725(141)	1.05	1.00	31.0
	10	410(56)	637(111)	1.08	1.00	29.6
AA	–	134(16)	166(13)	1.09	1.01	33.2

classification, which provides valuable information about the differential pathology underlying the disorders.

1) Posterior Distribution: Following the same experimental procedure as in [4], we use a $\text{Gamma}(2, 2)$ prior distribution in our MCMC sampler and allow for an adaptive phase of 5000 iterations. The results of our method in providing a distribution over kernel weightings for the Parkinsonian data are shown in fig. 2. We find that the results are comparable with those achieved in [4]. Like the reference results, we find that there is high variance on each estimate due to the small data set and weak prior in use. However, as discussed further in [4], the results do show some specific features that show a good correspondence with the known pathology of the diseases.

2) Efficiency and convergence: The efficiency of our approach on the brain region data is shown in Table II, and refers to $15,000$ samples. Results show that the mean ESS across all variables is quite low, being around 0.24%. In all, results provide an indication that the MH sampler may not be well suited to the problems investigated here. We propose that the large parameter space plays a preponderant role in limiting the efficiency that can be achieved by the MH algorithm. Furthermore, the LA may also provide a poor estimation of the distribution, thus leading to inefficiency. Despite the sampler achieving a very low percentage of independence, we notice that convergence is still achieved within the first 2000 samples. In particular, we highlight that our methodology achieved comparable convergence than the worst variable reported in the comparative test done in [4]. This shows that, despite poor efficiency in the sampling of the parameter, and relatively few samples in total, our approach is still able to quickly converge.

C. ADHD and ASD data

The third data set consists of structural MRI data derived from 77 adolescent subjects (aged 10-18) who were either healthy controls (29 subjects) or patients with either attention deficit/hyperactivity disorder (ADHD, 29 subjects) or autism spectrum disorder (ASD, 19 subjects). For each subject a T1 weighted SPGR structural image was acquired, which were preprocessed using SPM8. Similar to the previous application, images were segmented into different tissue types using new segment, normalised to a standard space using DARTEL, modulated to preserve total grey matter density then smoothed with an 8mm smoothing kernel prior to analysis. For these data, a single linear kernel was constructed from the normalised, smoothed whole-brain grey matter MRI data. Full details on the subject recruitment, MRI data acquisition and data preprocessing can be found in [15]. In [15] the LA is used to perform multi-class classification of the disorders, and will once again serve as a benchmark for the present work. The motivation for this application is to quantify the capability of MRI for discriminating between the disorders and healthy controls and to find biological markers predictive of disease state that can complement the behaviourally derived clinical diagnoses. Our analysis of the MCMC chains is based on the $\text{Gamma}(1, 1)$ prior and 10,000 samples gathered after 5,000 iterations where the MH algorithm was adapted.

The distribution over kernel weightings for the ADHD and ASD data are shown in fig. 2. Overall, we find that our method provides the ability to effectively assess the weighting distribution. Much like the previous two examples however, there is little discriminative ability provided by kernel weightings, particularly between HC and ASD patients. Regardless, the method is still useful in developing a predictive model for discrimination between the three classes.

1) Efficiency and Convergence: The efficiency and convergence achieved by our approach on the ADHD and ASD data are shown in Table III. The results show that the results are only marginally affected by the number of importance samples, and convergence is fast. Table III also reports a comparison of the ESS and of the $\hat{R}$ statistics with reparameterization techniques, and in particular with the Ancillary-Augmentation (AA) reparameterization as in [4]. The analysis of these

results indicate that the PM MCMC approach achieves higher efficiency with respect to reparameterization techniques, thus suggesting that the proposed approach is effective in breaking the correlation between latent variables and hyper-parameters.

2) Classification performance: Before concluding this section, we motivate the fully Bayesian treatment of MC-MKL problems by reporting the classification performance achieved by the proposed method. Uncertainty in model parameters is accounted for in the predictions, as given by eq. 8, and class labels are assigned based on the maximum of the predictive distribution across the three classes. The confusion matrix of the classifier, in a leave-one-out setting, results in:

		Predicted		
		ADHD	HC	ASD
Actual	ADHD	22	6	1
	HC	9	18	2
	ASD	2	1	16

From the confusion matrix we derive a balanced accuracy of 74.0%, which is higher than the comparable result of 68.2% that can be achieved when employing a multiple-class GP classifier using the LA [15]. The sensitivity of our classification over the three classes is 75.9%, 62.1%, and 84.2%, and the positive predictive values are 66.7%, 72.0%, and 84.2%.

V. Conclusion

In this paper, we showed the application of advanced MCMC methods to the multiple-class multiple-kernel learning (MC-MKL) problem using Gaussian Processes (GPs). We also demonstrated this application on real neuroimaging data. The results on the ADHD and ASD data show that the proposed fully probabilistic MC-MKL with GPs outperforms alternate approximate methods in terms of classification. This is consistent with the results obtained in other applications of fully Bayesian classification approaches [4], [7]. This demonstrates the issues in carrying out approximate inference as opposed to the proposed exact approach. Another result is in the capability of the proposed approach of providing posterior distributions over parameters that weight the contribution of different imaging sources to the classification problem. This is important for interpretation purposes. From the methodological perspective, we demonstrated that it is possible to devise an exact MCMC scheme, building upon deterministic approximations, requiring a comparable computational cost. The results suggest that the proposed MCMC approach yields chains that exhibit good convergence properties in relatively few samples. Furthermore, in some cases, good approximations to the marginal likelihood can be achieved using relatively few importance samples.

Overall, the results are important in furthering work in the direction of achieving feasible fully Bayesian inference across a wide range of applications. Furthermore, our results show some inefficiency when the sampling is carried out in high-dimensional parameter spaces; this is due to the inefficient exploration of the random walk characterizing the MH algorithm. As such, future work may investigate the use of alternative sampling methodologies, such as Hybrid Monte Carlo [5], which can produce less correlated samples, or the use of annealed importance sampling to reduce the variance of the estimator of the marginal likelihood as in [16]. Also, it would be interesting to investigate the use of alternative approximation schemes such as the recently proposed Expectation Propagation algorithm for multiple-class classification using GPs [17].

Acknowledgment

AFM gratefully acknowledges support from the King's College London Centre of Excellence in Medical Engineering, funded by the Wellcome Trust and EPSRC under Grant No. WT088641/Z/09/Z. We would also like to thank the NIHR Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, for their on-going support of the Centre for Neuroimaging Sciences.

References

[1] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[3] M. Gonen and E. Alpaydin, "Multiple Kernel Learning Algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[4] M. Filippone, A. F. Marquand, C. R. V. Blain, S. R. Williams, J. Mourão-Miranda, and M. Girolami, "Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities," *Annals of Applied Statistics*, vol. 6, no. 4, pp. 1883–1905, 2013.

[5] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993.

[6] J. M. Flegal, M. Haran, and G. L. Jones, "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" *Statistical Science*, vol. 23, no. 2, pp. 250–260, 2007.

[7] M. Filippone and M. Girolami, "Pseudo-marginal Bayesian inference for Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[8] M. Filippone, M. Zhong, and M. Girolami, "A comparative evaluation of stochastic-based inference methods for Gaussian process models," *Machine Learning*, vol. 93, no. 1, pp. 93–114, 2013.

[9] C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, vol. 37, no. 2, pp. 697–725, 2009.

[10] C. J. Geyer, "Practical Markov chain Monte Carlo," *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992.

[11] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992.

[12] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95 – 113, 2007.

[13] A. F. Marquand, M. Filippone, J. Ashburner, M. Girolami, J. Mourao-Miranda, G. J. Barker, S. C. R. Williams, P. N. Leigh, and C. R. V. Blain, "Automated, high accuracy classification of parkinsonian disorders: A pattern recognition approach," *PLoS ONE*, 2013.

[14] I. Litvan, K. P. Bhatia, D. J. Burn, C. G. Goetz, A. E. Lang, I. McKeith, N. Quinn, K. D. Sethi, C. Shults, and G. K. Wenning, "SIC Task Force Appraisal of Clinical Diagnostic Criteria for Parkinsonian Disorders," *Movement Disorders*, vol. 18, no. 5, pp. 467–486, 2003.

[15] L. Lim, A. Marquand, A. A. Cubillo, A. B. Smith, K. Chantiluke, A. Simmons, M. Mehta, and K. Rubia, "Disorder-specific predictive classification of adolescents with attention deficit hyperactivity disorder (adhd) relative to autism using structural magnetic resonance imaging," *PLoS ONE*, vol. 8, no. 5, p. e63660, 05 2013.

[16] M. Filippone, "Bayesian inference for Gaussian process classifiers with annealing and exact-approximate MCMC," 2013, arXiv:1311.7320.

[17] J. Riihimäki, P. Jylänki, and A. Vehtari, "Nested expectation propagation for Gaussian process classification," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 75–109, 2013.