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

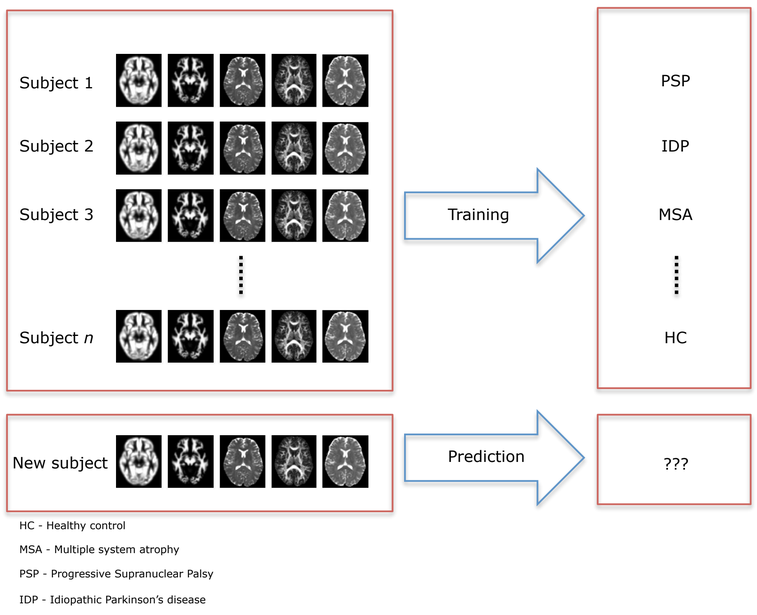
Andrew D. O'Harney[1], Andre Marquand[2,3], Katya Rubia[2], Kaylita Chantiluke[2], Anna Smith[2], Ana Cubillo[2], Camilla Blain[2], Maurizio Filippone[1]

1 - University of Glasgow, UK. 2 - King's College London, UK. 3 - Radboud University, The Netherlands.

Contact email address: maurizio.filippone@glasgow.ac.uk

Scope of this work

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.

Data

- Structural magnetic resonance imaging (MRI) data
  - T1-weighted structural imaging
  - Preprocessing using the SPM8 software package (`www.fil.ion.ucl.ac.uk/spm`)

Parkinsonian data

- Segmentation parcellating anatomically into six target regions of interest (brainstem, cerebellum, caudate, middle occipital gyrus, putamen, and one for all other brain regions)
- 62 subjects (healthy controls + patients with one of three akinetic-rigid neurological disorders)
  - 14 subjects healthy controls
  - 18 subjects multiple system atrophy (MSA)
  - 16 subjects progressive supranuclear palsy (PSP)
  - 14 subjects idiopathic Parkinson's disease (IPD)

ADHD and ASD data

- 77 adolescent subjects (aged 10-18)
  - 29 subjects healthy controls
  - 29 subjects with either attention deficit/hyperactivity disorder (ADHD)
  - 19 subjects with autism spectrum disorder (ASD)

Methods

Classification approach

- Multiple class
- Kernel-based - pairwise similarity between subjects
- Multiple kernel - multiple input data
- Probabilistic - we need sound quantification of uncertainty

Multi-Class Multiple Kernel classifier using Gaussian Processes

References

[1] M. Filippone, A. F. Marquand, C. R. V. Blain, S. C. R. Williams, J. Mourão-Miranda, M. Girolami. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *Annals of Applied Statistics*, 6(4):1883–1905, 2012.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] M. Filippone, "Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC," in ICPR 2014.

Classification Model

Data

- $\{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ set of n input/class pairs
- $\mathbf{y}_i = (y_i^1, \ldots, y_i^C)^{\mathrm{T}}$, with $y_i^c = 1$ if $\mathbf{x}_i$ belongs to cth class

Model

- Class labels conditionally independent given a set of class specific latent variables

$$\mathbf{f} = (f_1^1, \ldots, f_n^1, f_1^2, \ldots, f_n^2, f_1^C, \ldots, f_n^C)^{\mathrm{T}}$$

- Probability π_i^c that instance i belongs to class c is given by

$$\pi_i^c = \frac{\exp(f_i^c)}{\sum_s \exp(f_i^s)}$$

- Multinomial likelihood with probabilities π_i^c.

Latent Variables

- Class-specific sets of latent variables are assigned Gaussian Process (GP) priors $\mathcal{N}(\mathbf{f}^c|\mathbf{0}, K^c)$.
- Assume that d input sources are given, from which d covariance matrices S_h can be derived. Define

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

θ's can be used to asses importance of input sources in determining classes.

Laplace Approximation

Gaussian distribution $q(\mathbf{f}|\boldsymbol{\theta})$ approximating $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$

- Define $\Psi = \log[p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})]$
- Locate the mode of Ψ using Newton's iterations

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

- Match the covariance of the target and the approximating distributions at the mode of Ψ

Implementation storing at most $n \times n$ matrices

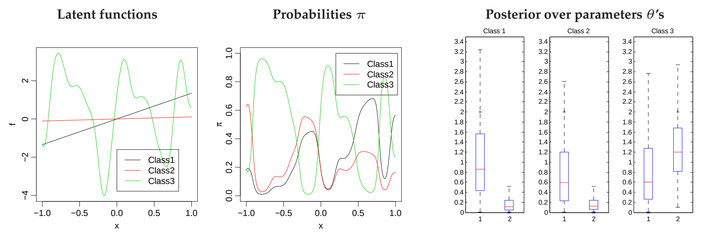
- Gradient $\nabla_{\mathbf{f}} \Psi(\mathbf{f}) = -K^{-1}\mathbf{f} + \mathbf{y} - \boldsymbol{\pi}$, and negative Hessian

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

where Π is obtained by stacking by row the matrices $\mathrm{diag}(\boldsymbol{\pi}^c)$

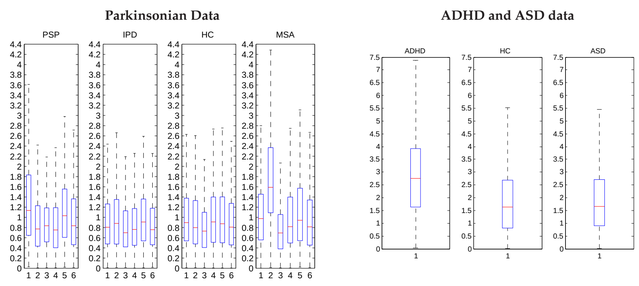
Synthetic data

- 150 data divided in three classes
- Two kernels: one RBF and one linear
- $\theta_{11} = \theta_{21} = \theta_{32} = 2$ and 0 for all others



Results

Posterior distributions



Sampling Efficiency and Convergence

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

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

Prediction Accuracy ADHD and ASD data

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

Balanced accuracy of 74.0% higher than 68.2% obtained by multiple-class GP using the Laplace Approximation

Inference and predictions

- Predictions of label $\mathbf{y}_*$ for new input $\mathbf{x}_*$

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

- Approximate integral using Markov chain Monte Carlo
  - Draw N samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$.
  - Approximate the predictive distribution by

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

- **Asymptotically exact**

Challenges

- Not possible to draw samples from $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$ directly
- Need to resort to Gibbs sampling types of schemes
- Sampling $\boldsymbol{\theta}|\mathbf{f}$ makes the MCMC approach very inefficient (sampling $\mathbf{f}|\boldsymbol{\theta}, \mathbf{y}$ is relatively easy instead)

Proposal

- Sample $\boldsymbol{\theta}|\mathbf{y}$ using the Metropolis-Hastings algorithm:
  - initialize the algorithm randomly from $\boldsymbol{\theta}$
  - propose a new set of parameters $\boldsymbol{\theta}'$ from $\pi(\boldsymbol{\theta}'|\boldsymbol{\theta})$
  - accept proposal with probability

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

- $\ldots$ but $p(\mathbf{y}|\boldsymbol{\theta})$ is analytically intractable!
- Nevertheless, we can get around this problem by using just an unbiased estimate $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ of $p(\mathbf{y}|\boldsymbol{\theta})$

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

- **This will yield samples from the correct** $p(\boldsymbol{\theta}|\mathbf{y})$
- We propose to estimate an $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ using importance sampling

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

where N_{imp} samples are drawn from a Gaussian distribution $q(\mathbf{f}|\boldsymbol{\theta})$ approximating $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$.

- Variance of the estimator affects MCMC efficiency

Drawing from $q(\mathbf{f}|\boldsymbol{\theta})$

Implementation requiring the storage of $n \times n$ matrices only

- Sequentially draw latent variables pertaining to each class

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

- Define precision of $q(\mathbf{f}|\boldsymbol{\theta})$ as $\Lambda = K^{-1} + \mathrm{diag}(\boldsymbol{\pi}) - \Pi\Pi^{\mathrm{T}}$
- Inverse covariance of $\mathbf{f}^1$ is

$$\Lambda_{[1,1]} - \Lambda_{[1,2:C]} \Lambda_{[2:C,2:C]}^{-1} \Lambda_{[2:C,1]} \tag{1}$$

$\Lambda_{[2:C,2:C]}$ can be inverted storing at most $n \times n$ matrices

- Mean of $\mathbf{f}^r|\mathbf{f}^{r-1}, \ldots, \mathbf{f}^1$ is

$$\hat{\mathbf{f}}^r - \Lambda_{[r,r]}^{-1} \Lambda_{[r,1:(r-1)]} (\mathbf{f}^{[1:(r-1)]} - \hat{\mathbf{f}}^{[1:(r-1)]}) \tag{2}$$

Conclusions and ongoing work

- We proposed the use of Gaussian Process classification for multiple-class multiple-kernel (MC-MKL) learning for neuroimaging data where:
  - Quantification of uncertainty in predictions is of primary interest
  - Assessment of importance of different sources of information is key to design future experiments
- We demonstrated that exact quantification of uncertainty in parameter estimates leads to better predictions compared to approximate methods
- We proposed a practical way to carry out exact quantification of uncertainty based on advanced Markov chain Monte Carlo methods
- A large variance for $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$ can severely reduce the efficiency of the proposed method
- We are studying alternatives to importance sampling and to the Laplace Approximation to reduce the variance of $\tilde{p}(\mathbf{y}|\boldsymbol{\theta})$