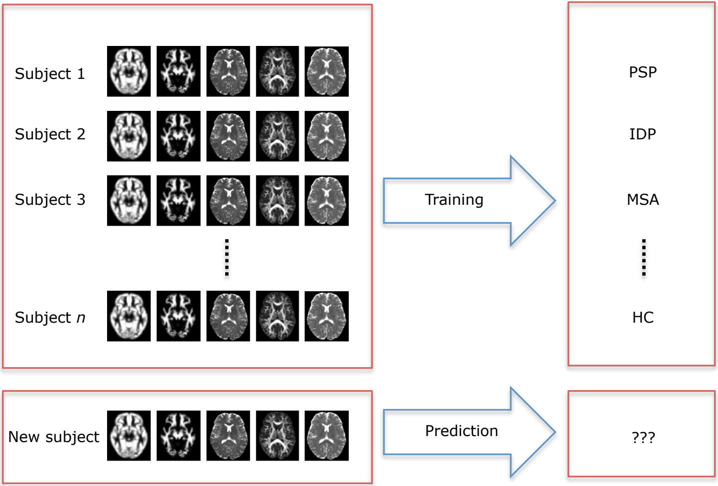
Exact-Approximate Bayesian Inference for Gaussian Processes

Maurizio Filippone

School of Computing Science
University of Glasgow
maurizio.filippone@glasgow.ac.uk

January 31st, 2014
Università di Torino

Motivating Application



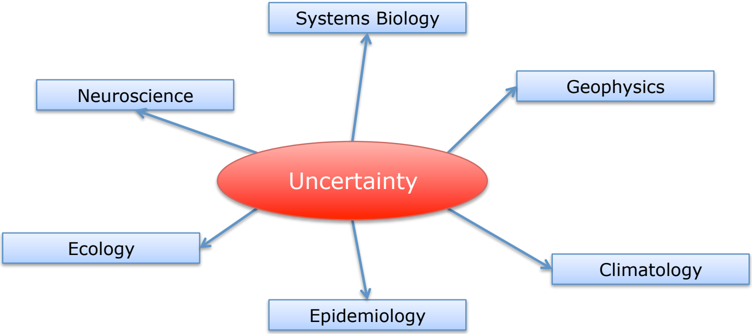
HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

Relevance of the problem



Measuring Uncertainty

- Data viewed as random variables

- Probabilities as degrees of belief

Models

- Mapping input to labels

$$\text{label} = \text{function}(\text{input}, \text{par}, \text{noise})$$

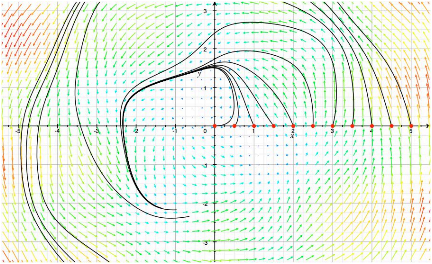
Models

- Mapping input to labels

$$\text{label} = \text{function}(\text{input}, \text{par}, \text{noise})$$

- function can describe a physical system

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ x \end{bmatrix} + \begin{bmatrix} \cos(\text{par}_1 x) \\ \sin(\text{par}_2 y) \end{bmatrix} - \text{par}_3 \begin{bmatrix} x \\ y \end{bmatrix}$$

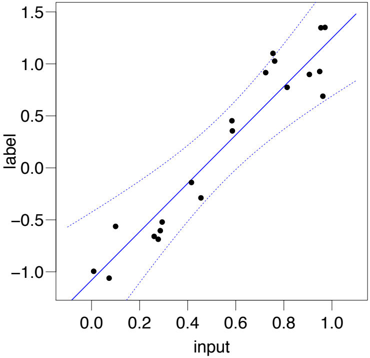


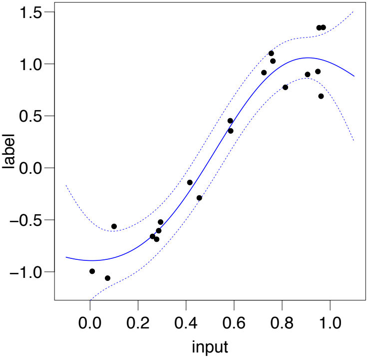
Models

- Mapping input to labels

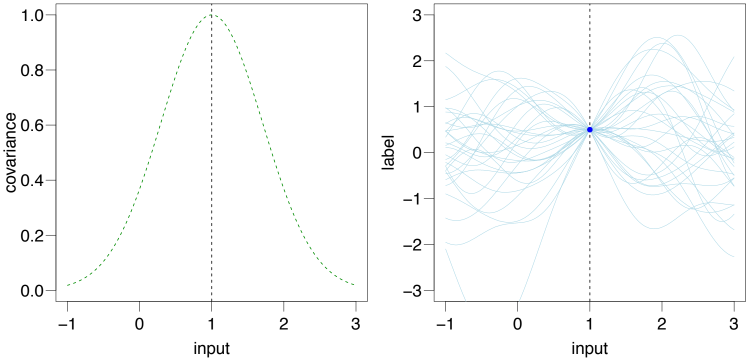
$$\text{label} = \text{function}(\text{input}, \text{par}, \text{noise})$$

- We may have no clue so we need assumptions on function

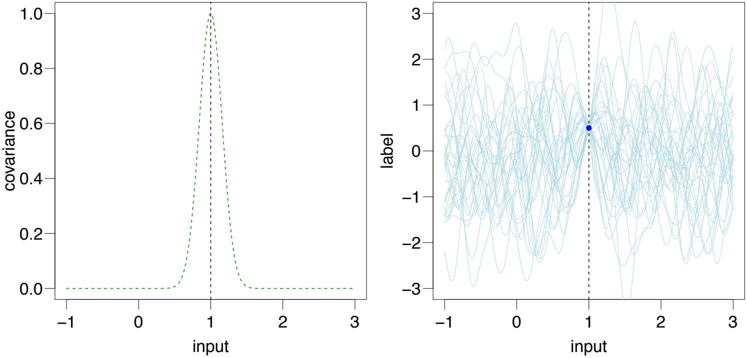




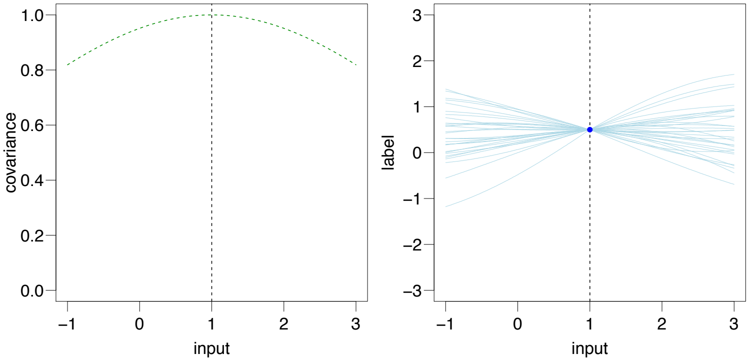
- Gaussians with distance dependent covariance



- Gaussians with distance dependent covariance



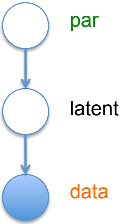
- Gaussians with distance dependent covariance



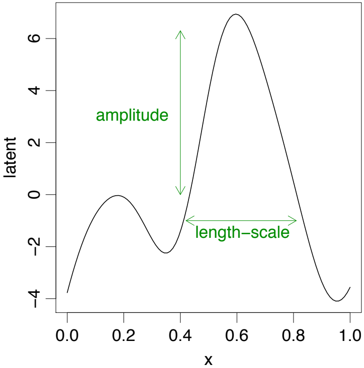
- Class of hierarchical models

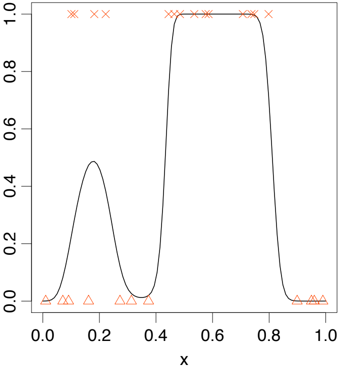
$$p(\text{data}|\text{latent}) \qquad p(\text{latent}|\text{par}) \qquad p(\text{par})$$

- $p(\text{latent}|\text{par}) = \text{Gaussian Process}$

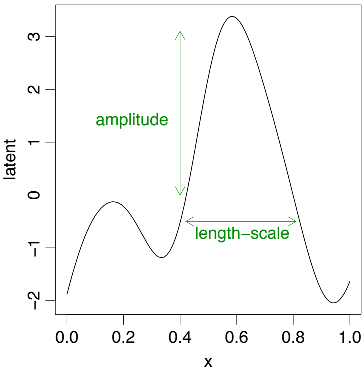


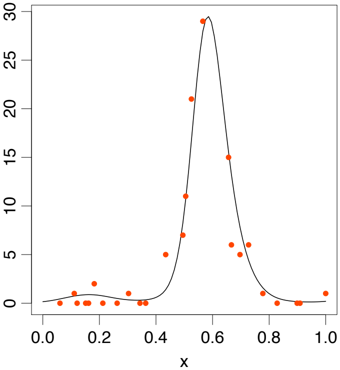
Gaussian Process Models - Classification example





Gaussian Process Models - Count data example



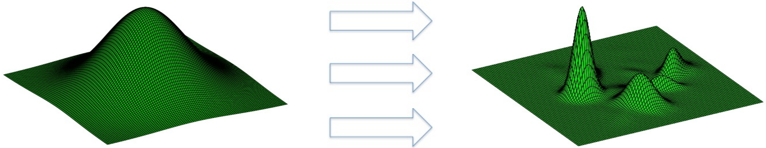


Bayesian Inference

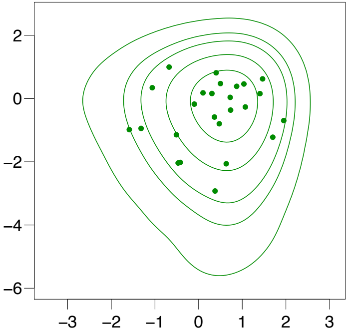
- Inference using Bayes theorem:

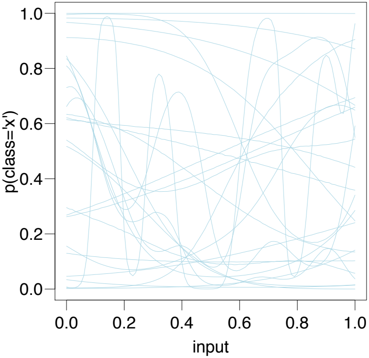
$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$

$p(\text{par})$ $\qquad$ $p(\text{par}|\text{data})$

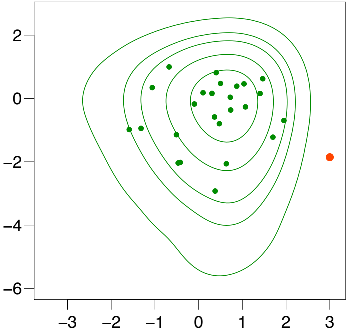


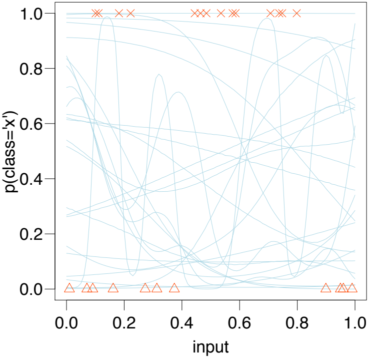
Bayesian Inference - Prior



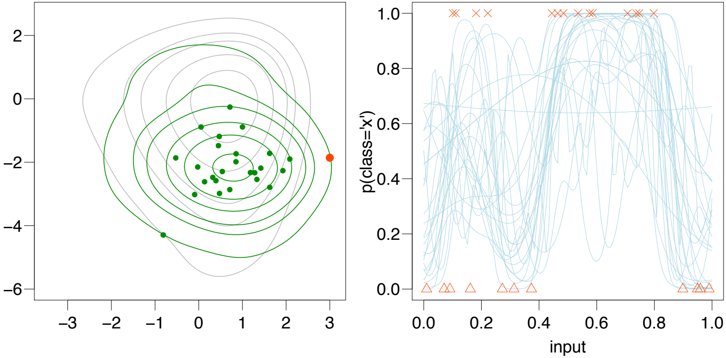


Bayesian Inference - Data

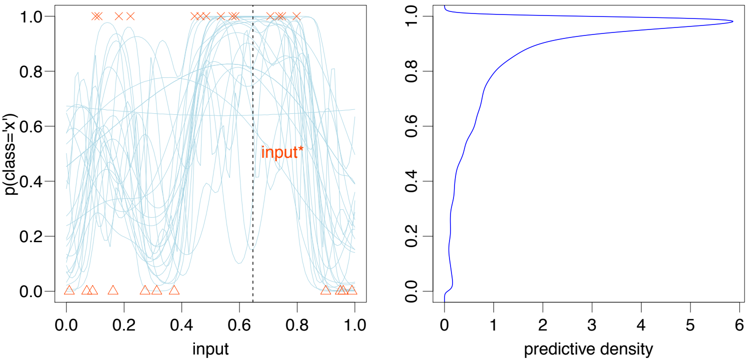




Bayesian Inference - Posterior



Bayesian Inference and Predictions



Bayesian Inference and Predictions

- Predictions for new data

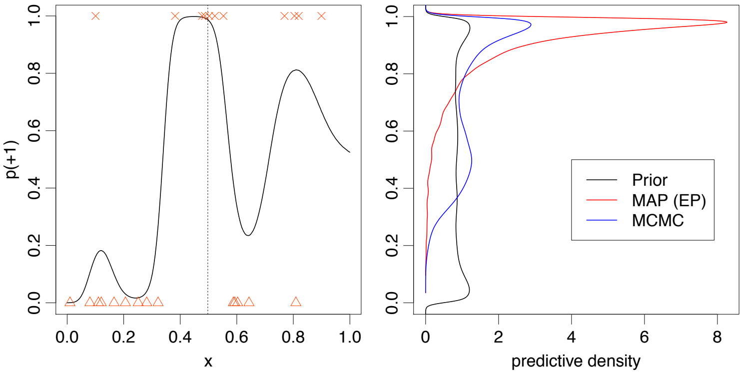
$$p(\text{label}_*|\text{label}) = \int p(\text{label}_*|\text{par})p(\text{par}|\text{label})\, d\text{par}$$

- Monte Carlo integration:

$$\int p(\text{label}_*|\text{par})p(\text{par}|\text{label})\, d\text{par} \simeq \frac{1}{N}\sum_{i=1}^{N} p(\text{label}_*|\text{par}_i)$$

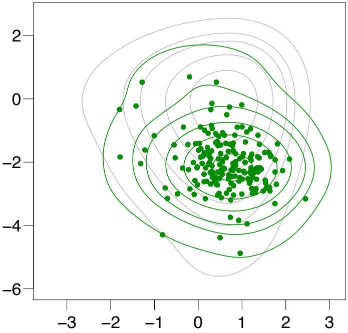
with par_i drawn from $p(\text{par}|\text{data})$

Bayesian Inference vs Parameter Optimization



Bayesian Inference and Predictions

- Draw samples according to the posterior density



Markov chain Monte Carlo

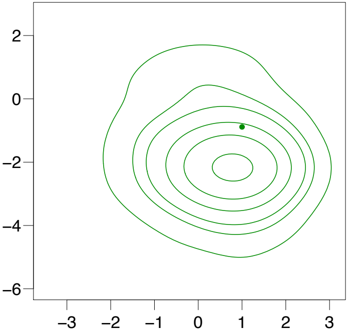
- Bayesian inference

$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$

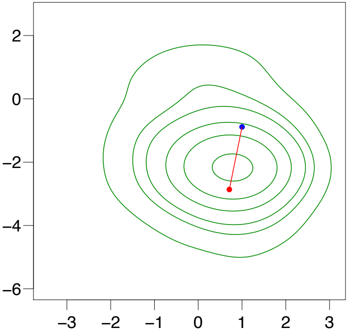
- Random walk sampler - accept a proposal with probability

$$\min\left(1, \frac{p(\text{par}'|\text{data})}{p(\text{par}|\text{data})}\right)$$

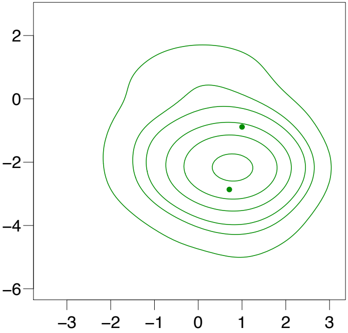
- Explore the parameter space according to the density



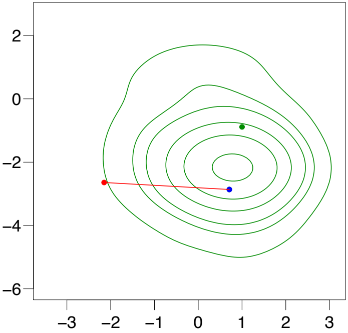
- Explore the parameter space according to the density



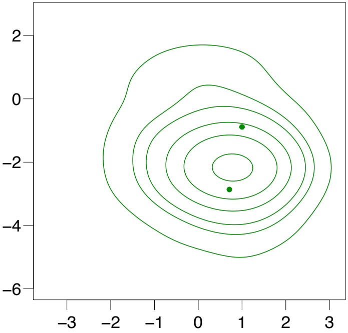
- Explore the parameter space according to the density



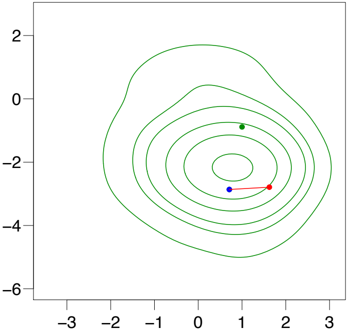
- Explore the parameter space according to the density



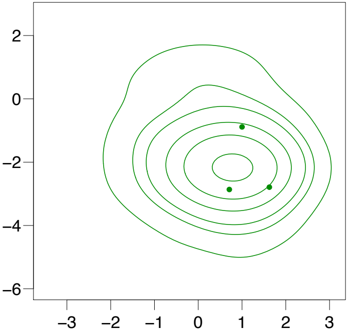
- Explore the parameter space according to the density



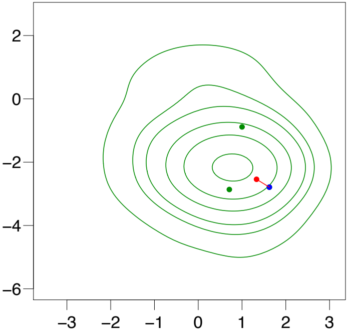
- Explore the parameter space according to the density



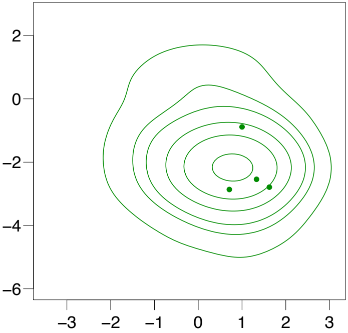
- Explore the parameter space according to the density



- Explore the parameter space according to the density

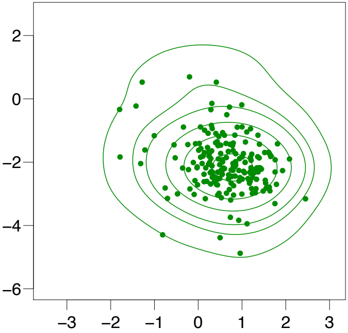


- Explore the parameter space according to the density



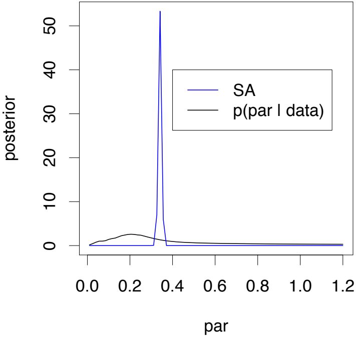
Markov chain Monte Carlo

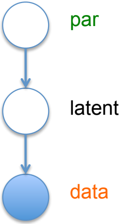
- Explore the parameter space according to the density



Obvious iterative scheme (aka Sufficient Augmentation (SA) scheme). Alternate between:

- Drawing from $p(\mathrm{latent}|\mathrm{par}, \mathrm{data})$
- Drawing from $p(\mathrm{par}|\mathrm{latent})$ (**bad idea** - see figure)





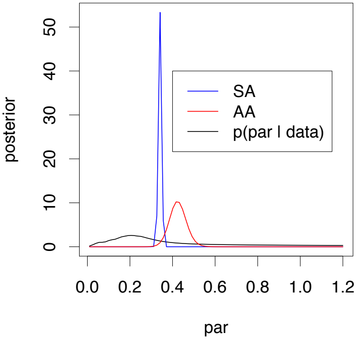
- No exact Gibbs steps - need to employ Metropolis within Gibbs steps - waste of computations when rejecting
- Updates of par cost $O(n^3)$
- par can be large dimensional (e.g., Automatic Relevance Determination (ARD) covariance function)
- There are n latent variables (as many as the number of observations)

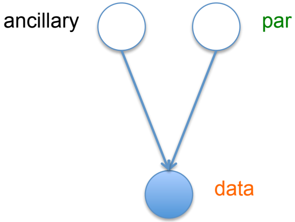
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization:

$$K = LL^{\mathrm{T}} \qquad \text{ancillary} = L^{-1}\,\text{latent}$$

- Replace sampling of par with $p(\text{par}|\text{ancillary}, \text{data})$





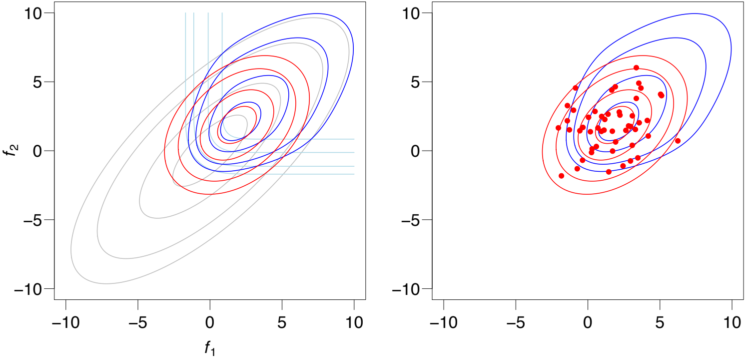
- Replace posterior by unbiased estimate

$$\min\left(1, \frac{\tilde{p}(\mathrm{par}'|\mathrm{data})}{\tilde{p}(\mathrm{par}|\mathrm{data})}\right)$$

- Achieved by using an unbiased estimate of $\tilde{p}(\mathrm{data}|\mathrm{par})$

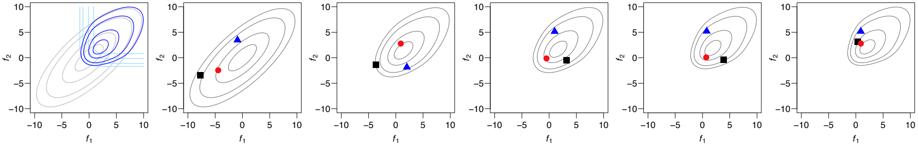
Importance Sampling estimator

- Approximate posterior over latent variables
- $\tilde{p}(\text{data}|\text{par})$ based on the approximation



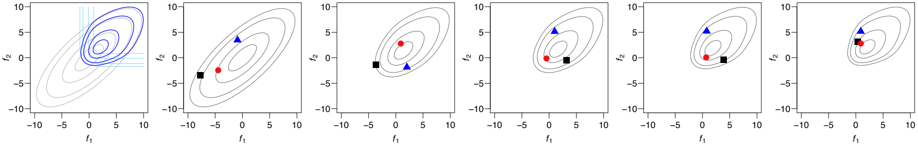
Annealed Importance Sampling estimator

- Annealing from the prior

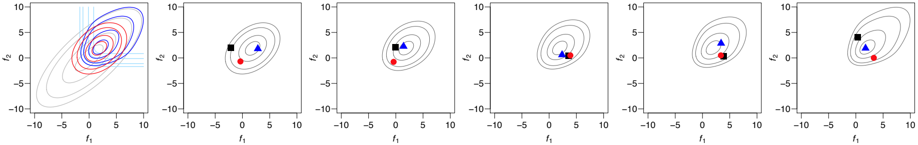


Annealed Importance Sampling estimator

- Annealing from the prior

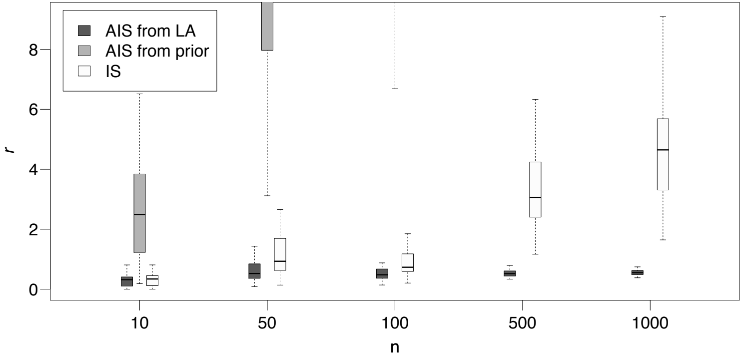


- Annealing from an approximation

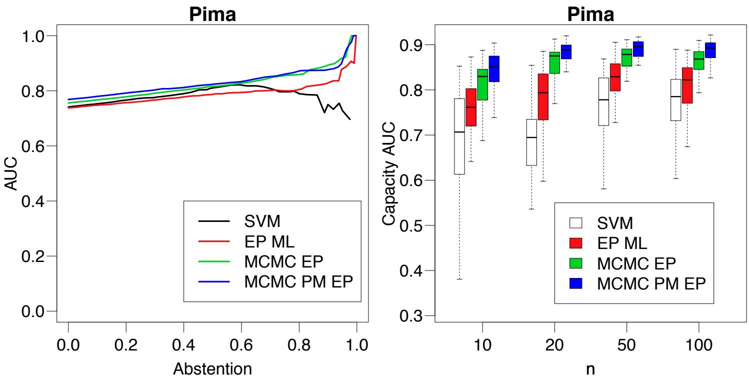


Comparison between AIS with IS

- Analysis of the variance of the AIS and IS estimators



Some Results



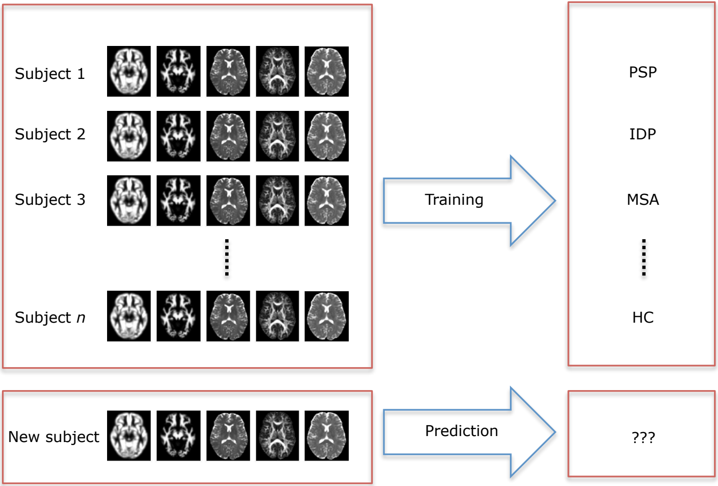
Isotropic covariance

N_{imp}	Glass $n = 214,\ d = 9$ IS	Glass $n = 214,\ d = 9$ AIS	Thyroid $n = 215,\ d = 5$ IS	Thyroid $n = 215,\ d = 5$ AIS	Breast $n = 682,\ d = 9$ IS	Breast $n = 682,\ d = 9$ AIS
1	2.8(1.6)	5.2(1.9)	1.1(1.0)	3.2(2.3)	17.9(2.4)	28.0(2.7)
10	10.4(3.1)	11.4(5.3)	4.1(3.8)	6.4(3.9)	30.5(4.1)	36.4(3.5)

Isotropic covariance

N_{imp}	Pima $n = 768,\ d = 8$ IS	Pima $n = 768,\ d = 8$ AIS	Banknote $n = 1372,\ d = 4$ IS	Banknote $n = 1372,\ d = 4$ AIS	USPS $n = 1540,\ d = 256$ IS	USPS $n = 1540,\ d = 256$ AIS
1	24.8(1.4)	29.3(2.6)	1.1(0.6)	3.2(3.9)	0.6(0.6)	1.1(1.0)
10	30.8(2.6)	30.8(1.7)	4.7(1.0)	12.6(4.1)	0.6(0.5)	2.0(0.4)

Motivating Application



HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

Multiclass classification with multiple sources

- Multiclass classification based on GPs

$$p(\text{disease} = c|\text{sources}) = \text{unknown function}$$

- unknown function modeled using GPs

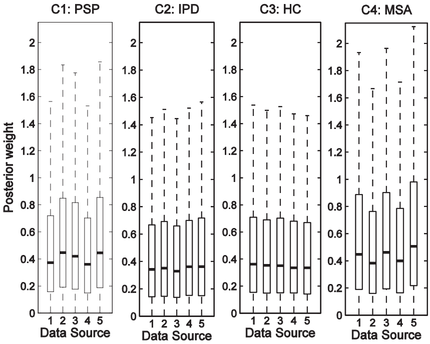
Multiclass classification with multiple sources

- Covariance based on source-dependent covariances S_k

$$\sum_{k=1}^{K} w_{ck} S_k(subject_i, subject_j)$$

Parkinson syndromes data - multi source

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418

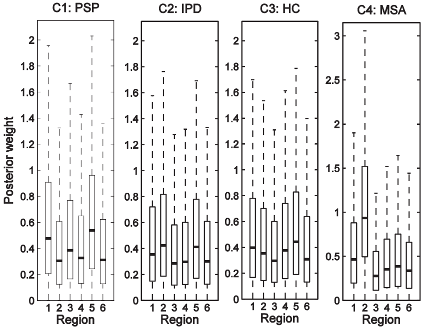


Analysis of brain regions

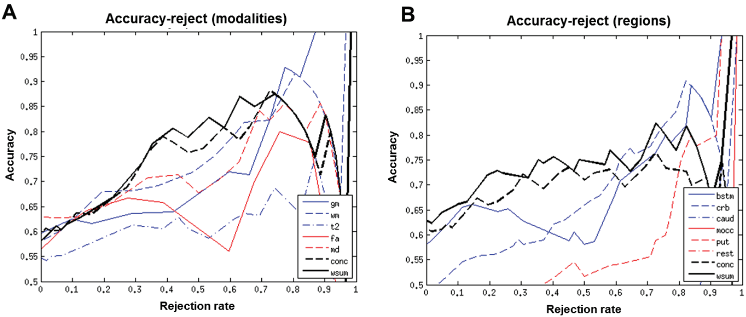
1. brainstem
2. bilateral cerebellum
3. bilateral caudate
4. bilateral middle occipital gyrus
5. bilateral putamen
6. all other regions

Multiclass classification with multiple regions

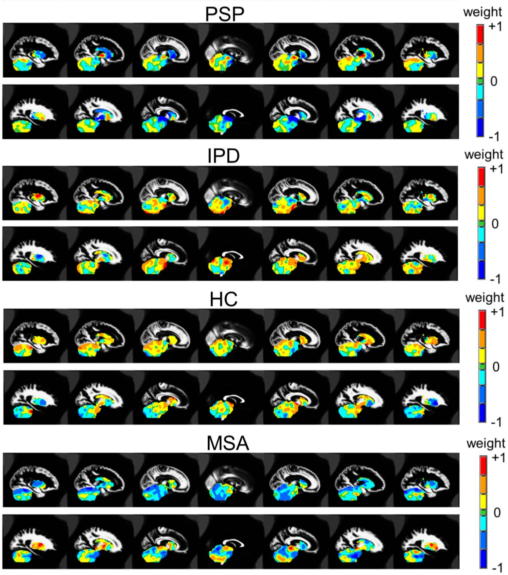
Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



Reject Option



Brain Maps



Conclusions and ongoing work

- Bayesian inference offers a powerful methodology to accurately quantifying uncertainty

Conclusions and ongoing work

- Bayesian inference offers a powerful methodology to accurately quantifying uncertainty
- Gaussian Processes yield flexible and interpretable nonparametric models

Conclusions and ongoing work

- Bayesian inference offers a powerful methodology to accurately quantifying uncertainty
- Gaussian Processes yield flexible and interpretable nonparametric models
- Challenges in making Bayesian computations for Gaussian Processes scalable

Conclusions and ongoing work

- Bayesian inference offers a powerful methodology to accurately quantifying uncertainty
- Gaussian Processes yield flexible and interpretable nonparametric models
- Challenges in making Bayesian computations for Gaussian Processes scalable
- Research at the interface between Physics, Computing, Statistics and Mathematics

Acknowledgements

- Dr Andre F. Marquand (King's College London)
- Prof Mark Girolami (University of Warwick)
- Dr Guido Sanguinetti (University of Edinburgh)
- Dr Alessandro Vinciarelli (University of Glasgow)

[1] A. Gelman et al., Bayesian data analysis. Chapman and Hall, 1995.

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

[3] M. Filippone et al. **Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities**. *Annals of Applied Statistics, 6(4):1883-1905, 2012.*

[4] A. F. Marquand et al. **Automated, high accuracy classification of parkinsonian disorders: a pattern recognition approach**. *PLoS ONE, 2013.*

[5] M. Filippone et al. **A comparative evaluation of stochastic-based inference methods for Gaussian process models**. *Machine Learning, 93(1):93-114, 2013.*

[6] M. Filippone and M. Girolami. **Exact-approximate Bayesian inference for Gaussian processes**, 2013, arXiv:1310.0740.

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

Acknowledgements

Flickr images (Creative Commons license)

- "roll those babies!!" by torbakhopper
- "digital-drugs-binaural-beats" by digitalbob8