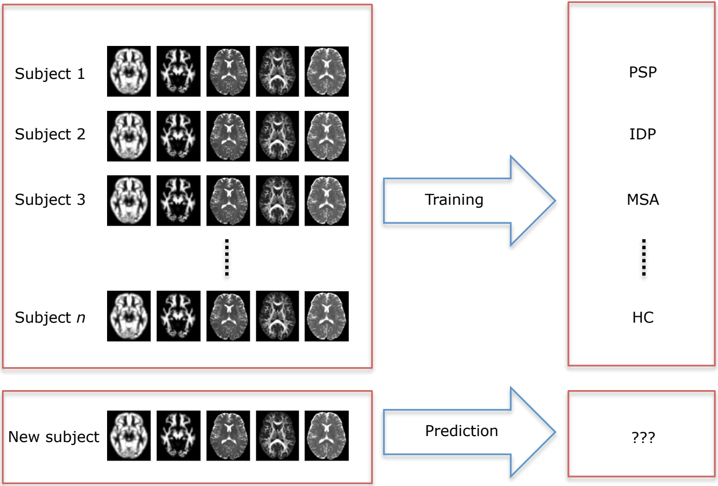
Pseudo-Marginal Bayesian Inference for Gaussian Processes

Maurizio Filippone

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

October 16th, 2014

Motivating Application



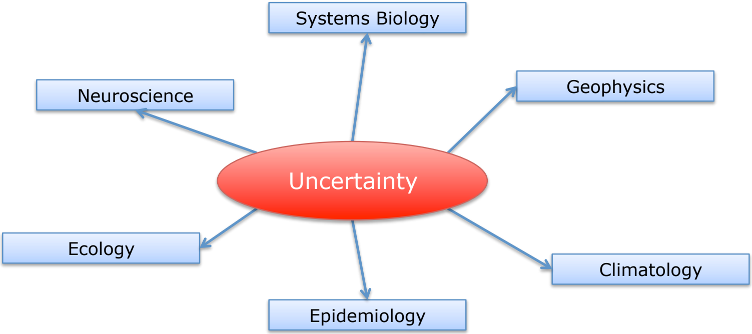
HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

Relevance of the problem



Relevance of the problem

Probabilistic Modeling

- How do we estimate model parameters?
- How do we assess that a model is preferable over another?
- How do we incorporate knowledge by experts?
- How can we attach confidence intervals to our predictions and parameter estimates?

Probabilistic modeling offers an answer to these questions

Measuring Uncertainty

- Data viewed as random variables


- Probabilities as degrees of belief

Models

- Mapping input to labels

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

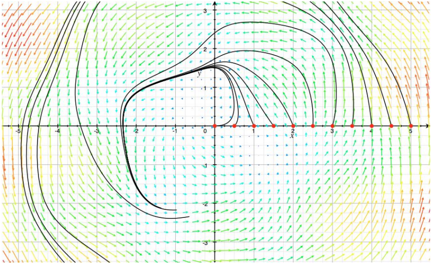
Models

- Mapping input to labels

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

- function can describe a physical system

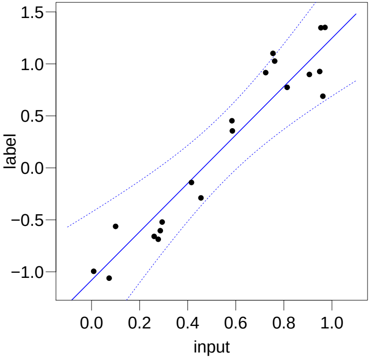
$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ x \end{bmatrix} + \begin{bmatrix} \cos(\text{par}_1 x) \\ \sin(\text{par}_2 y) \end{bmatrix} - \text{par}_3 \begin{bmatrix} x \\ y \end{bmatrix}$$

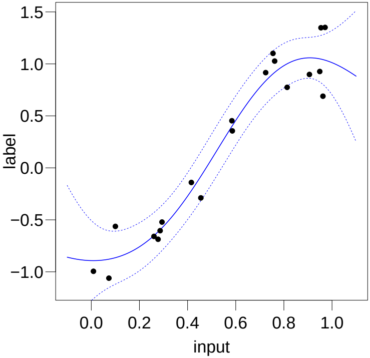


- Mapping input to labels

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

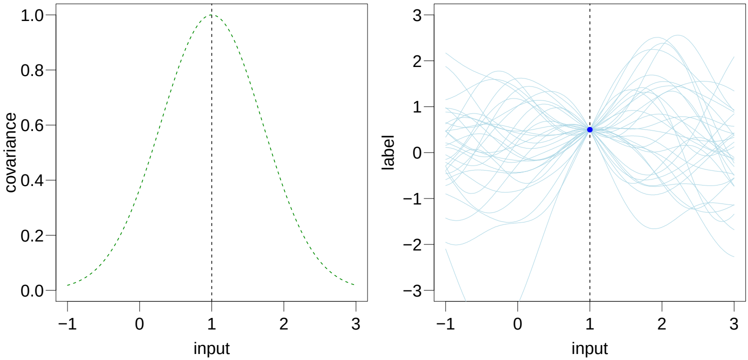
- We may have no clue about function - we need assumptions





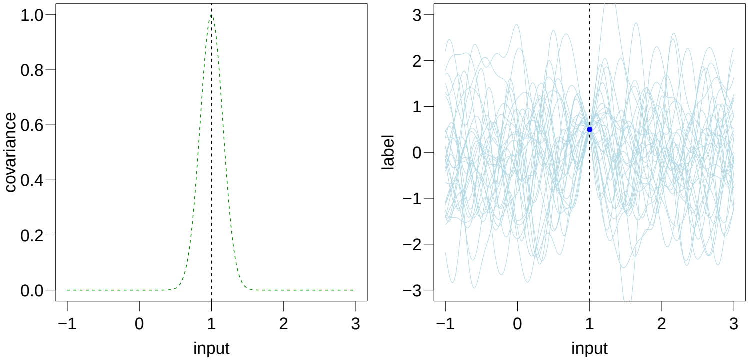
Gaussian Processes

- Gaussians with distance dependent covariance

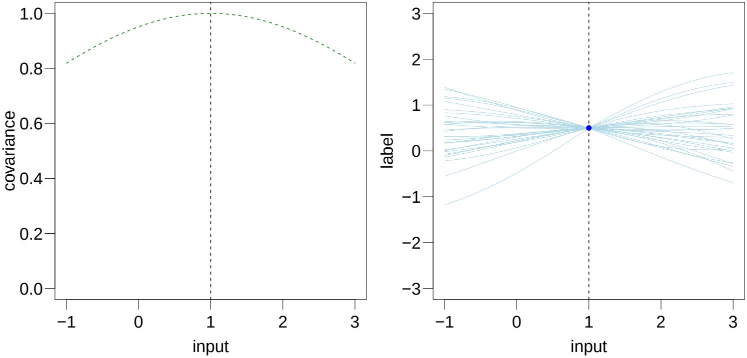


Gaussian Processes

- Gaussians with distance dependent covariance



- Gaussians with distance dependent covariance

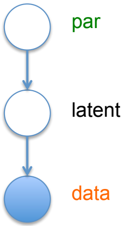


Gaussian Process Models

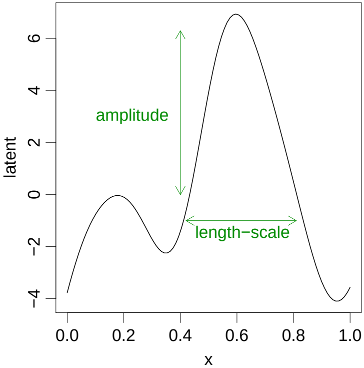
- Class of hierarchical models

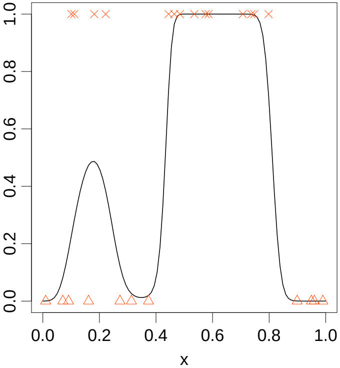
$$p(\text{data}|\text{latent}) \qquad p(\text{latent}|\text{par}) \qquad p(\text{par})$$

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

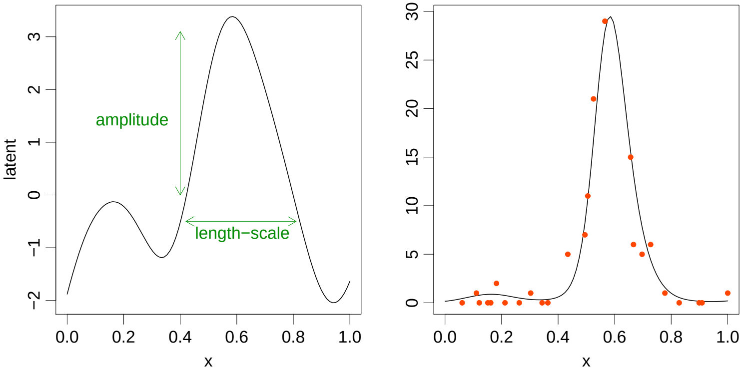


Gaussian Process Models - Classification example





Gaussian Process Models - Count data example

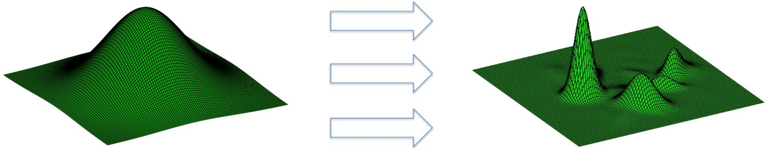


Bayesian Inference

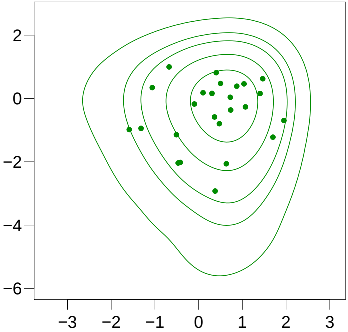
- Inference using Bayes theorem:

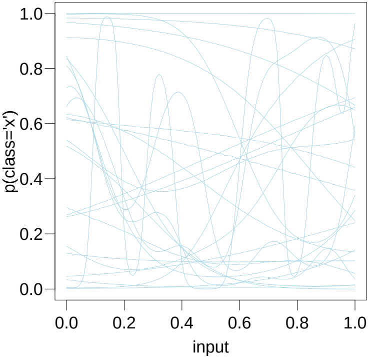
$$p(\mathrm{par}|\mathrm{data}) = \frac{p(\mathrm{data}|\mathrm{par})p(\mathrm{par})}{\int p(\mathrm{data}|\mathrm{par})p(\mathrm{par})d\mathrm{par}}$$

$p(\mathrm{par})$ $p(\mathrm{par}|\mathrm{data})$

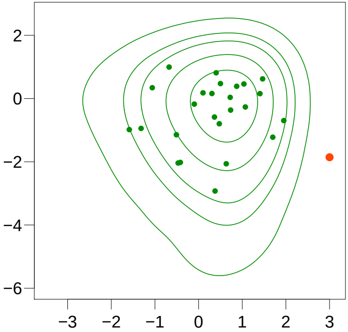


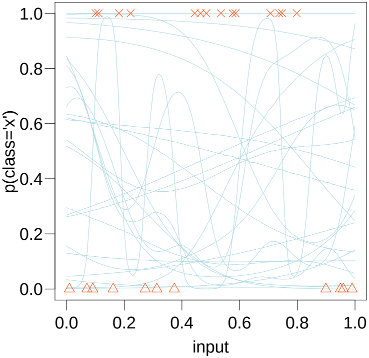
Bayesian Inference - Prior



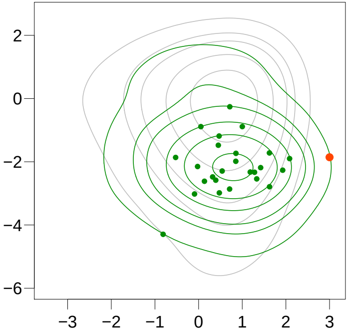


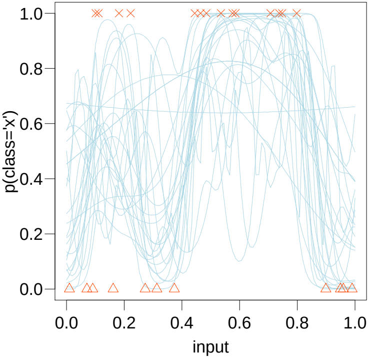
Bayesian Inference - Data



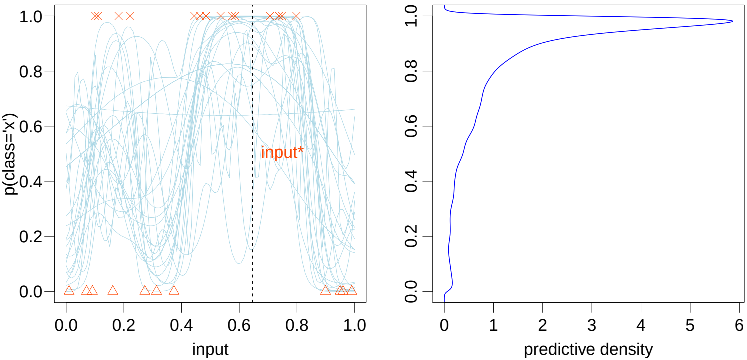


Bayesian Inference - Posterior





Bayesian Inference and Predictions



Bayesian Inference and Predictions

- Predictions for new data

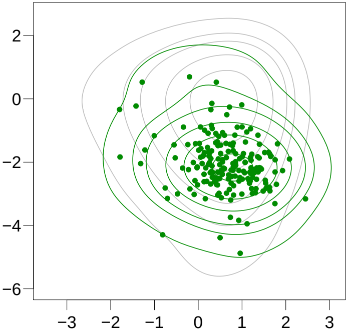
$$p(\mathrm{label}_*|\mathrm{label}) = \int p(\mathrm{label}_*|\mathrm{par})p(\mathrm{par}|\mathrm{label})\, d\mathrm{par}$$

- Monte Carlo integration:

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

with $\mathrm{par}^{(i)}$ drawn from $p(\mathrm{par}|\mathrm{data})$

- Draw samples according to the posterior density



Markov chain Monte Carlo

- Bayesian inference

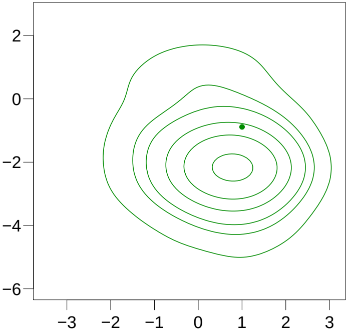
$$p(\mathrm{par}|\mathrm{data}) = \frac{p(\mathrm{data}|\mathrm{par})p(\mathrm{par})}{\int p(\mathrm{data}|\mathrm{par})p(\mathrm{par})d\mathrm{par}}$$

- Random walk sampler - accept a proposal with probability

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

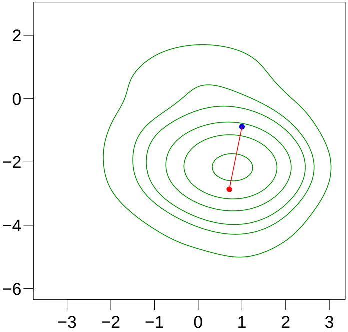
Markov chain Monte Carlo

- Explore the parameter space according to the density

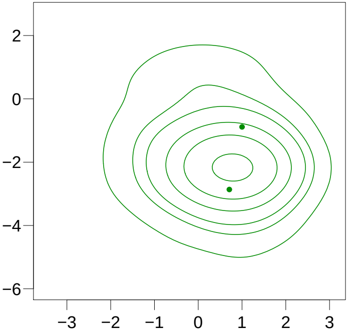


Markov chain Monte Carlo

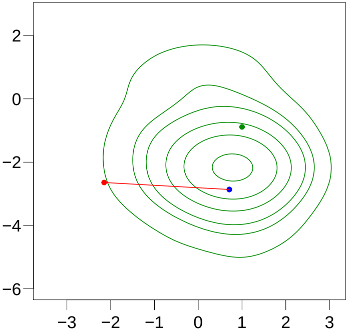
- Explore the parameter space according to the density



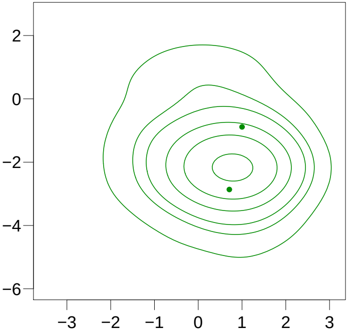
- Explore the parameter space according to the density



- Explore the parameter space according to the density

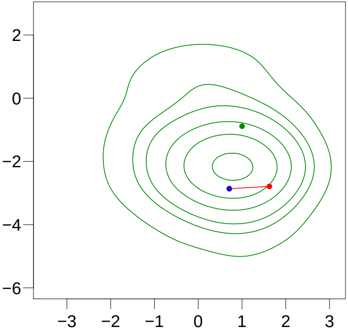


- Explore the parameter space according to the density

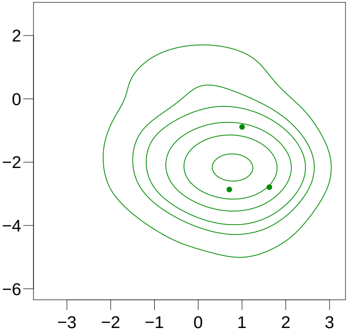


Markov chain Monte Carlo

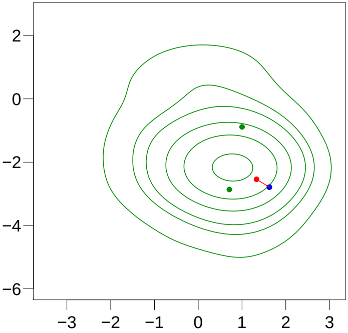
- Explore the parameter space according to the density



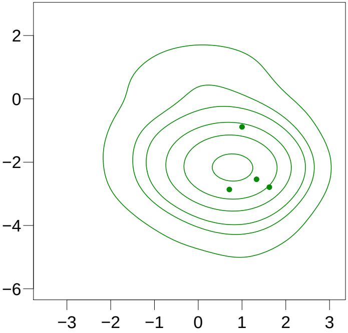
- Explore the parameter space according to the density



- Explore the parameter space according to the density

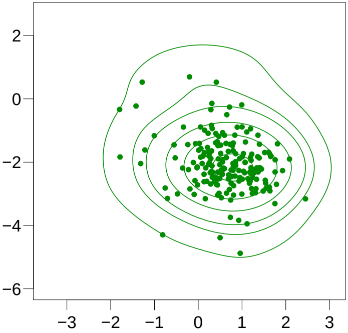


- Explore the parameter space according to the density



Markov chain Monte Carlo

- Explore the parameter space according to the density



How can we draw from $p(\text{par}|\text{data})$ in GPMs?

Marginal likelihood

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

is unavailable analytically. Options:

- Approximate $p(\text{data}|\text{par})$ within MCMC
- Sample from $p(\text{par}, \text{latent}|\text{data})$
- Pseudo-Marginal MCMC

Gaussian Approximations for marginal likelihood

Gaussian approximation to $p(\mathrm{latent}|\mathrm{data}, \mathrm{par})$

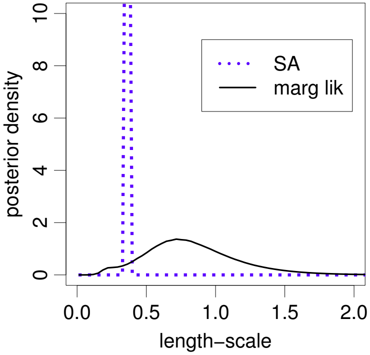
- Laplace Approximation
- Expectation Propagation
- Variational Bayes
- ...

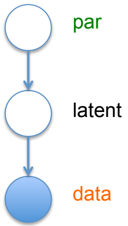
Challenges in MCMC for GPMs - Cost & exploration

- 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)

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

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



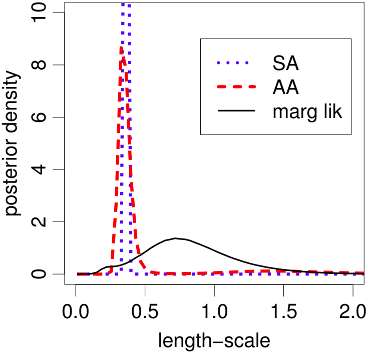


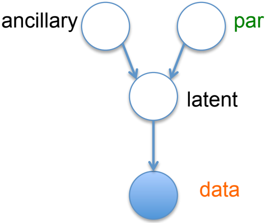
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization:

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

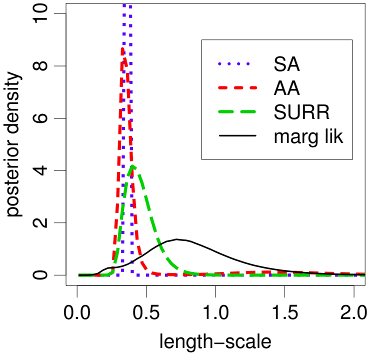
- Replace sampling of par with $p(\text{par}|\text{ancillary}, \text{data})$





Surrogate data model (SURR):

- Introduce set of auxiliary variables informed by the posterior over latent



$$\text{surrogate} = f(\text{latent}, \text{par})$$

Pseudo-Marginal MCMC

- Replace posterior by unbiased estimate

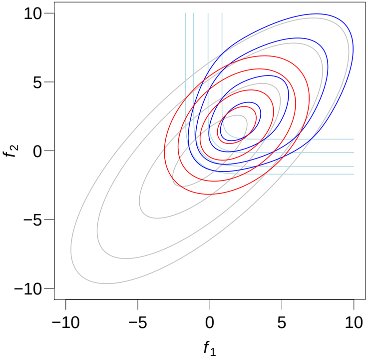
$$\min\left(1, \frac{\tilde{p}(\mathrm{par}'|\mathrm{data})}{\tilde{p}(\mathrm{par}|\mathrm{data})}\right)$$

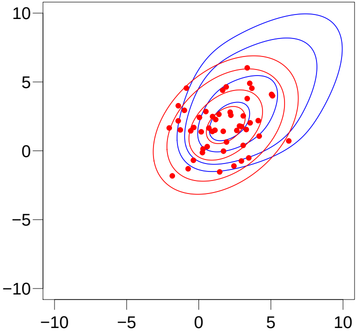
- Achieved by using an unbiased estimate of $\tilde{p}(\mathrm{data}|\mathrm{par})$

Importance Sampling estimator

- Approximate posterior over latent variables using $q(\text{latent})$
- Then

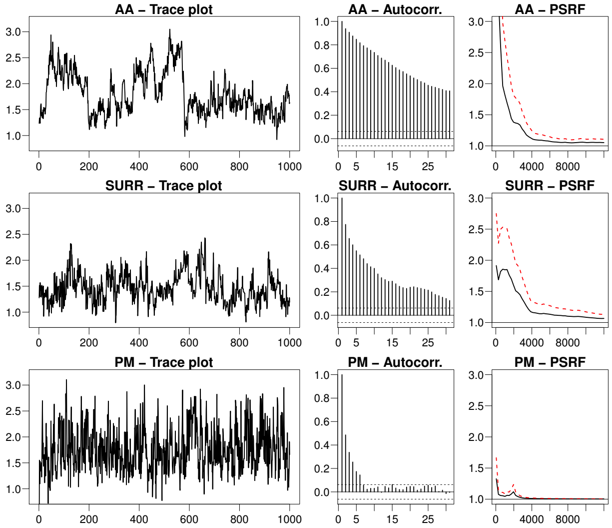
$$\tilde{p}(\text{data}|\text{par}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\text{data}|\text{latent}^{(i)})p(\text{latent}^{(i)}|\text{par})}{q(\text{latent}^{(i)})}$$





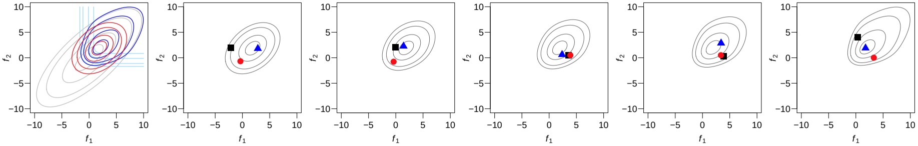
Convergence speed and efficiency

Abalone data set (two classes) $n = 2835$ - inference of length-scale



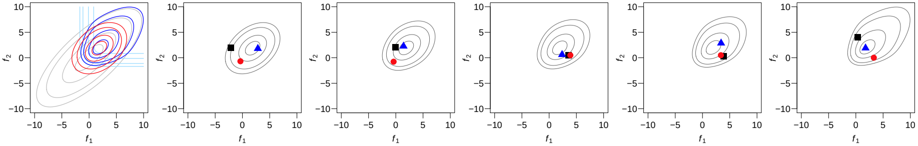
Annealed Importance Sampling estimator

- Annealing from an approximation

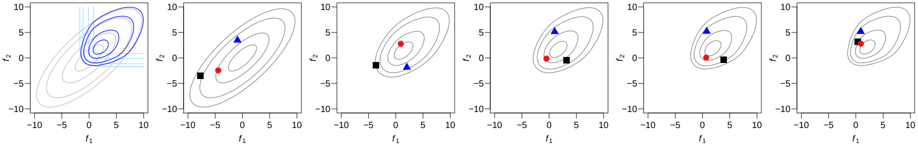


Annealed Importance Sampling estimator

- Annealing from an approximation

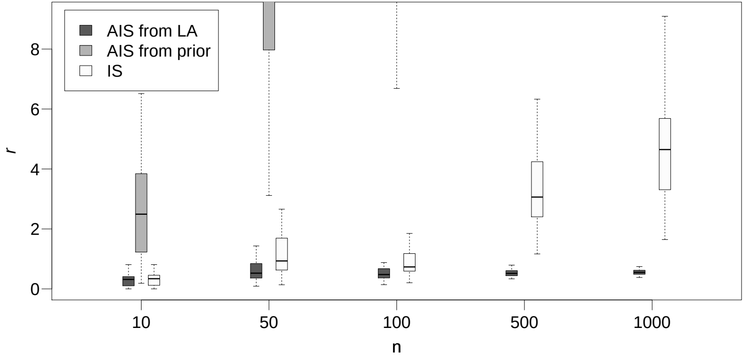


- Annealing from the prior

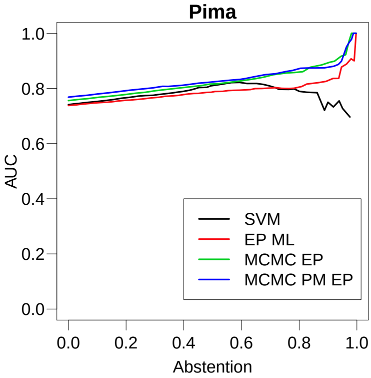


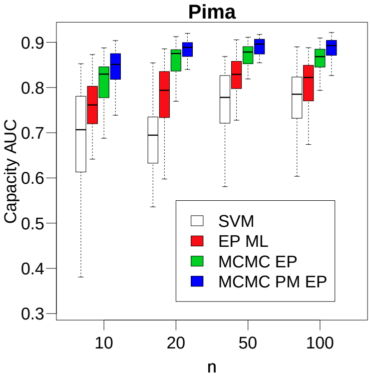
Analysis of the variance of the AIS and IS estimators

- r is the variance of the $\log_{10}$ marginal likelihood

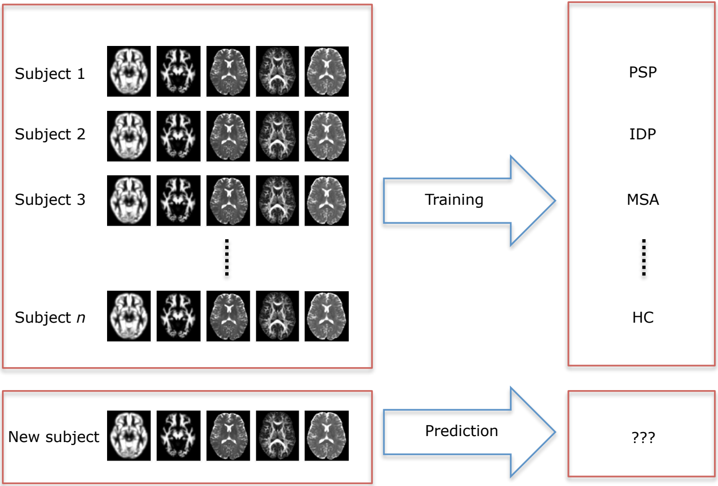


Some Results





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

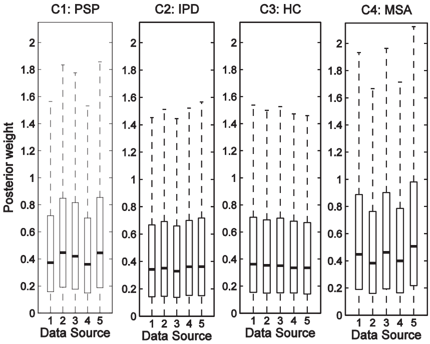
Multiclass classification with multiple sources

- 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



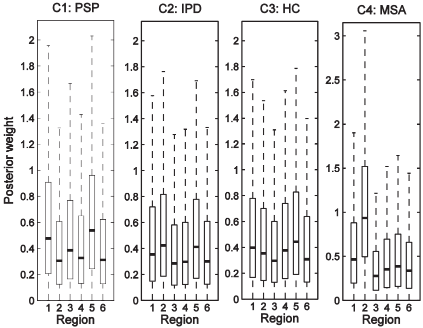
Multiclass classification with multiple regions

Analysis of brain regions

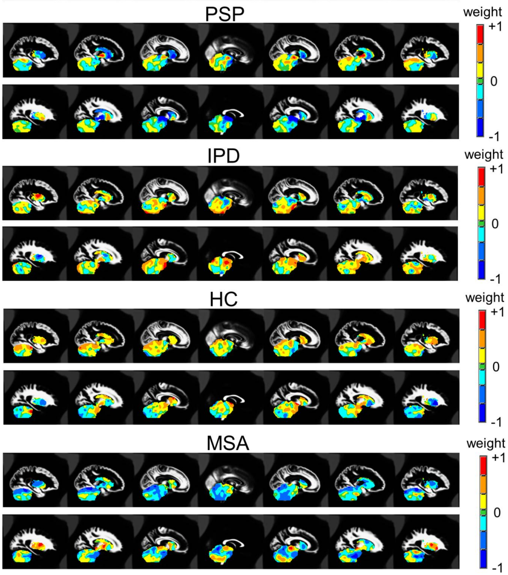
1. brainstem
2. bilateral cerebellum
3. bilateral caudate
4. bilateral middle occipital gyrus
5. bilateral putamen
6. all other regions

Multiclass classification with multiple regions

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



Brain Maps



Conclusions and ongoing work

- Gaussian Processes yield flexible and interpretable nonparametric models

Conclusions and ongoing work

- Gaussian Processes yield flexible and interpretable nonparametric models
- Bayesian inference to accurately quantifying uncertainty in such models

Conclusions and ongoing work

- Gaussian Processes yield flexible and interpretable nonparametric models
- Bayesian inference to accurately quantifying uncertainty in such models
- Pseudo-Marginal MCMC offers a practical way to carry out exact Bayesian computations

Conclusions and ongoing work

- Gaussian Processes yield flexible and interpretable nonparametric models
- Bayesian inference to accurately quantifying uncertainty in such models
- Pseudo-Marginal MCMC offers a practical way to carry out exact Bayesian computations
- How to make exact Bayesian computations for Gaussian Processes scalable?

Acknowledgements

- Dr Andre F. Marquand (Radboud University)
- Prof Mark Girolami (University of Warwick)
- Dr Guido Sanguinetti (University of Edinburgh)
- Dr Alessandro Vinciarelli (University of Glasgow)

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

[2] M. Filippone. **Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC**, In *ICPR, 2014.*

[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.*