



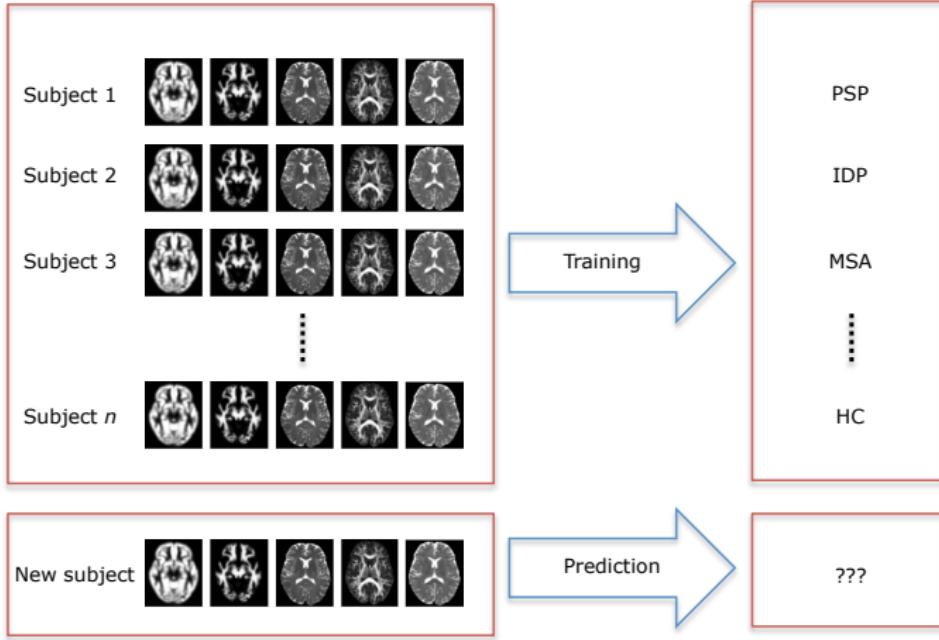
Unbiased computations for MCMC-based inference of Gaussian process covariance parameters

Maurizio Filippone

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

Apr 23rd, 2015

Motivating Application



HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

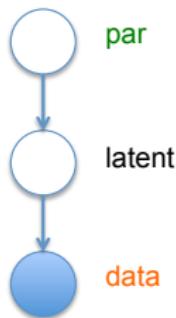
IDP - Idiopathic Parkinson's disease

Gaussian Process Models

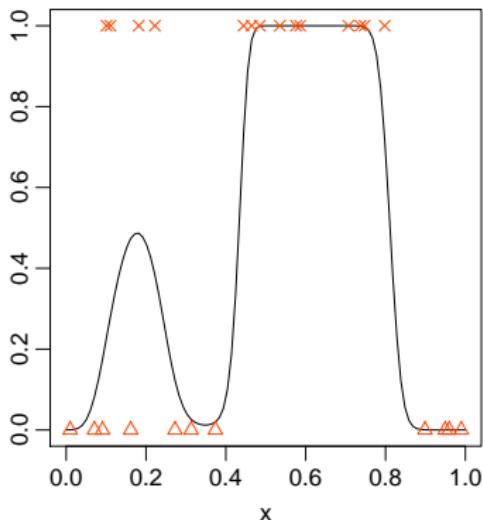
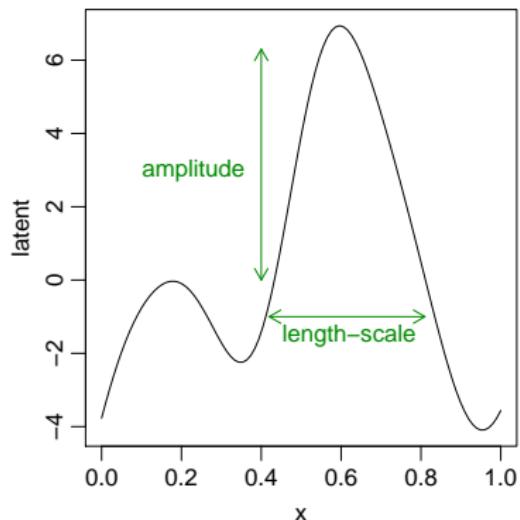
- Class of hierarchical models

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

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



Gaussian Process Models - Classification example



Multiclass classification with multiple sources

- Multiclass classification based on GPs

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

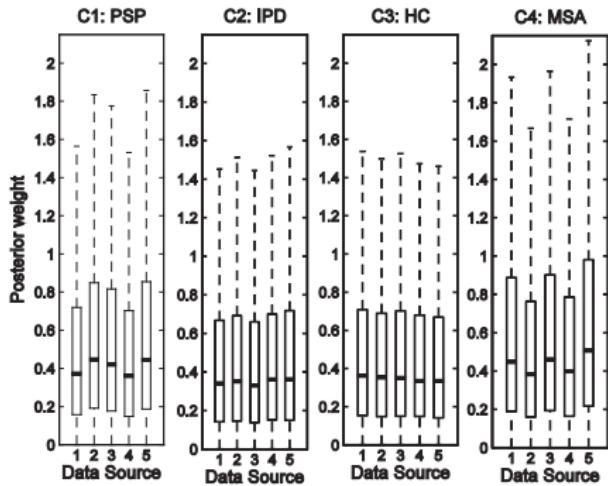
- unknown function modeled using GPs
- Covariance based on source-dependent covariances S_k

$$\sum_{k=1}^K w_{ck} S_k(\text{subject}_i, \text{subject}_j)$$

Filippone et al., AoAS, 2012

Parkinson syndromes data - multi source

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418



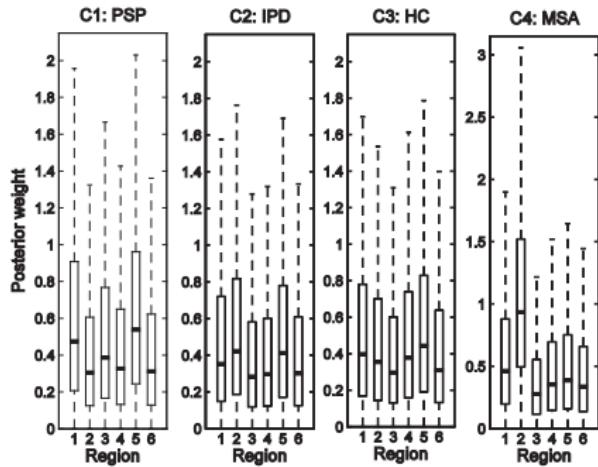
Filippone et al., AoAS, 2012

Analysis of brain regions

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

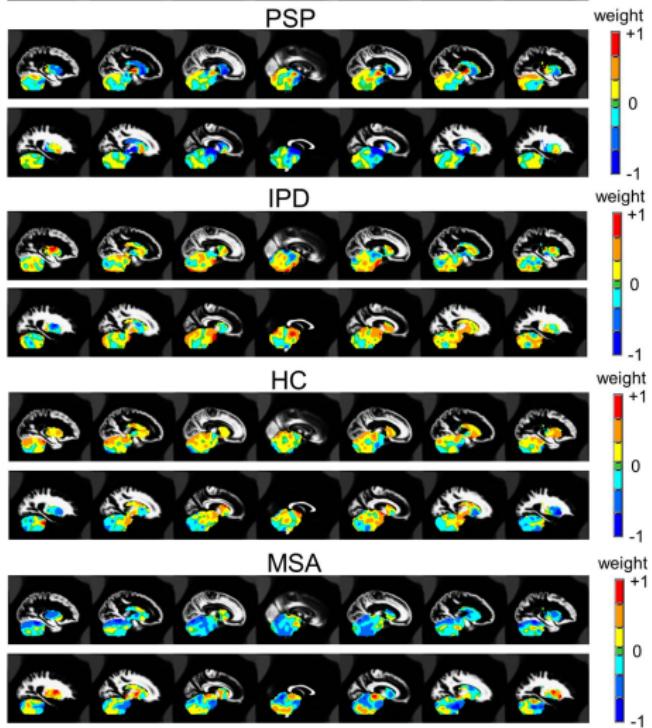
Multiclass classification with multiple regions

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229

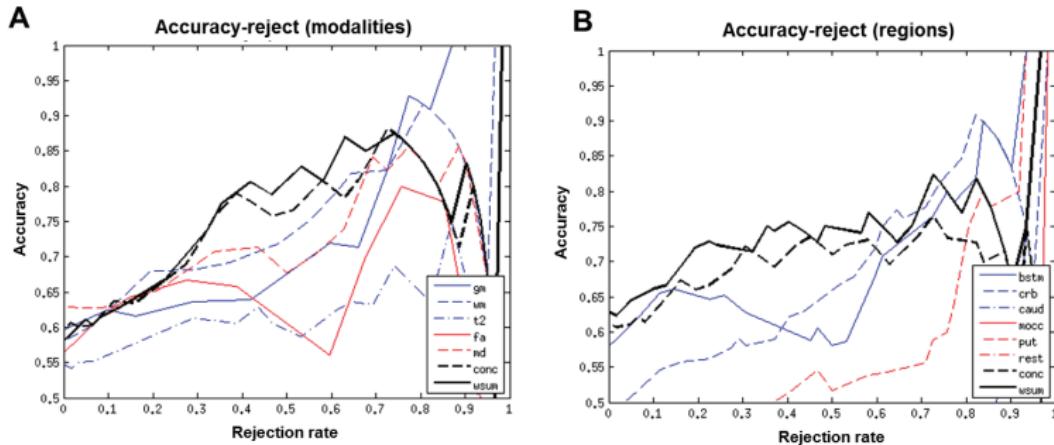


Filippone et al., AoAS, 2012

Brain Maps

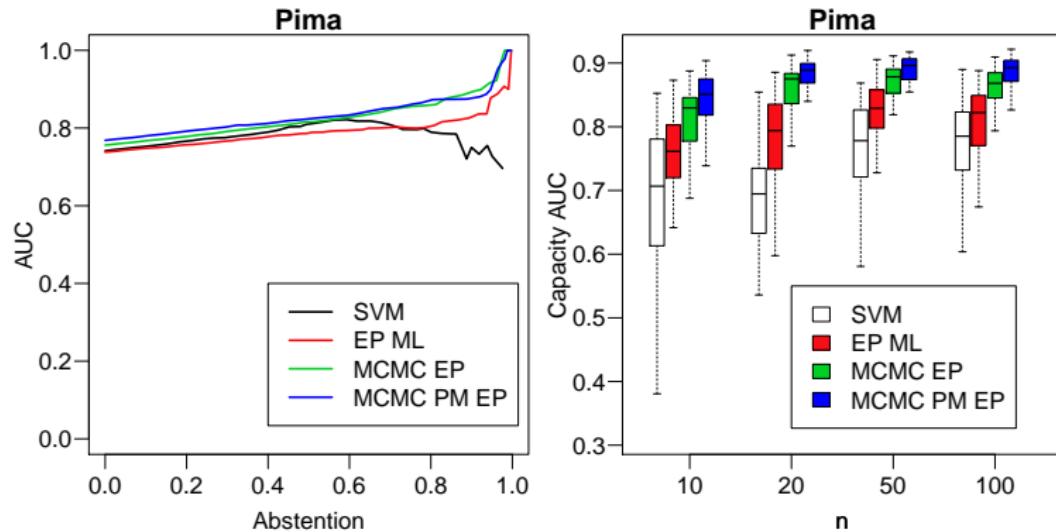


Reject Option



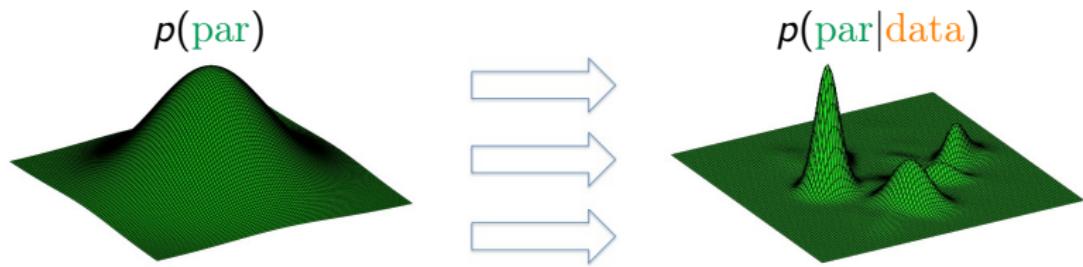
Filippone et al., AoAS, 2012

Importance of characterizing posterior over par

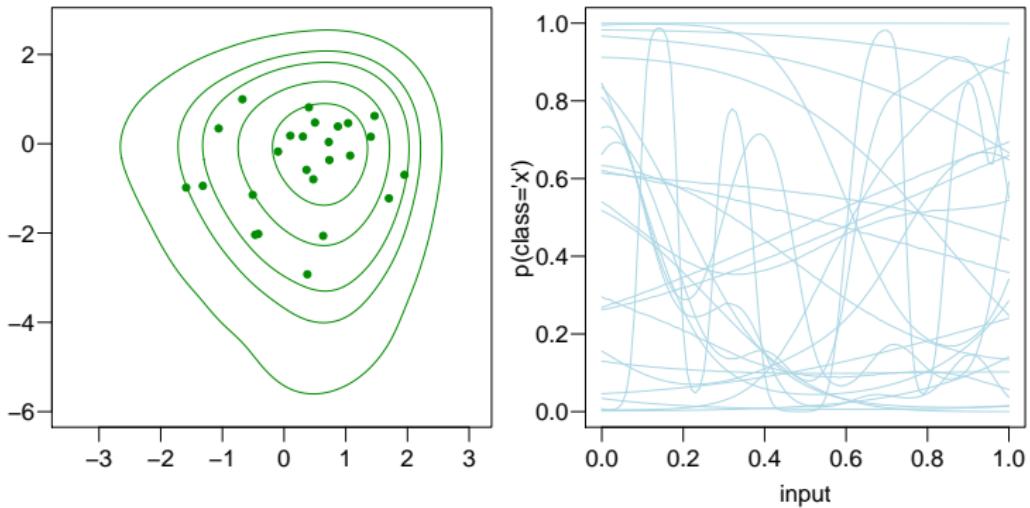


- Inference using Bayes theorem:

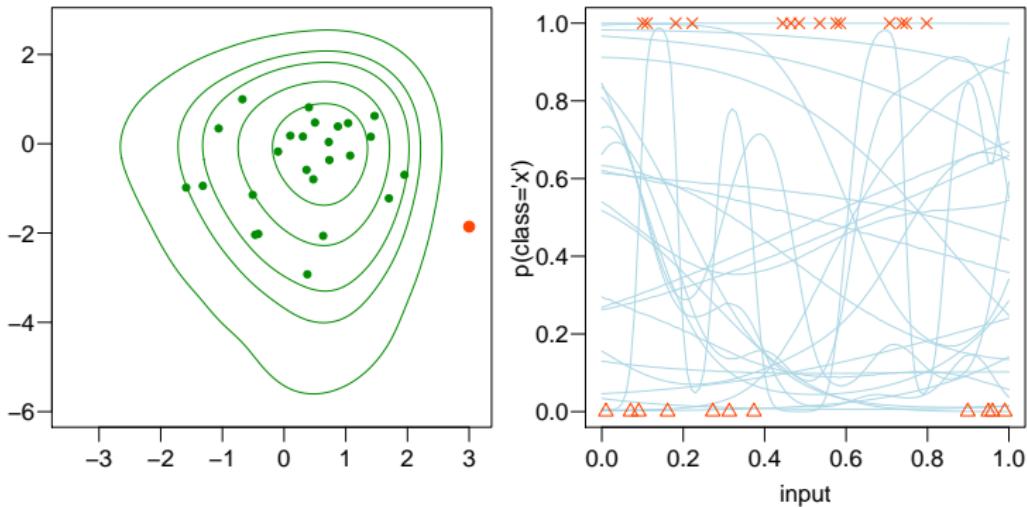
$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$



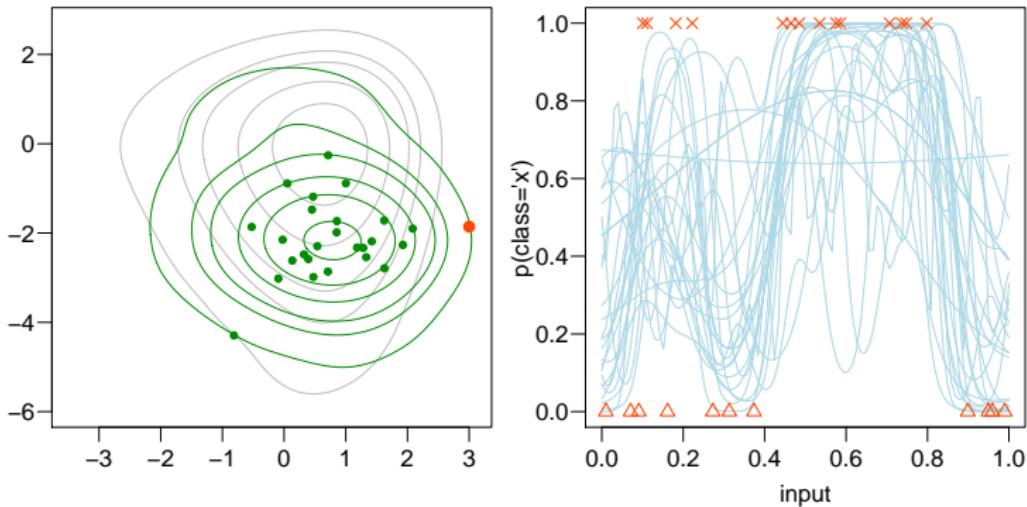
Bayesian Inference - Prior



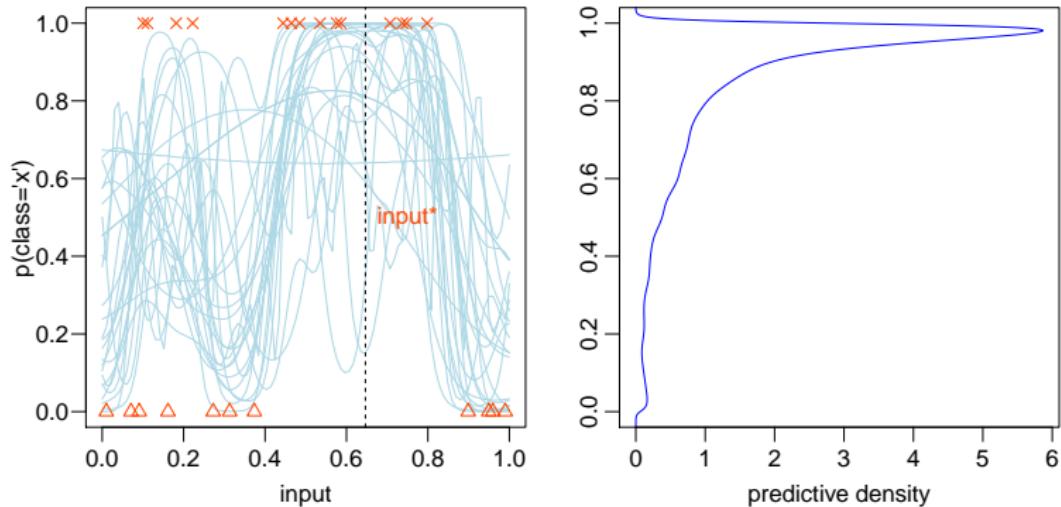
Bayesian Inference - Data



Bayesian Inference - Posterior



Bayesian Inference and Predictions



- Predictions for new data

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

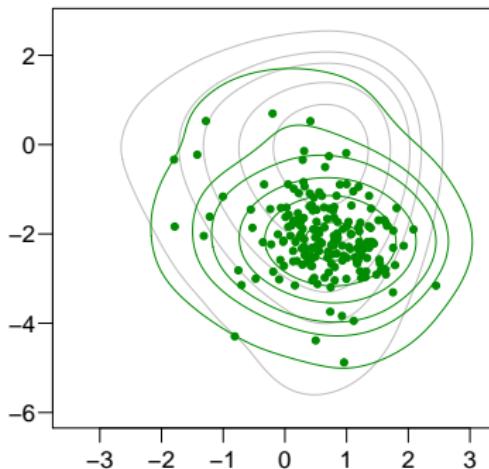
- Monte Carlo integration:

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

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

Bayesian Inference and Predictions

- Draw samples according to the posterior density



- 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

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

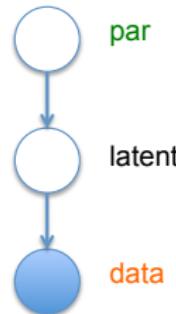
Acceptance probability : $\min \left(1, \frac{p(\text{data}|\text{par}')p(\text{par}')}{p(\text{data}|\text{par})p(\text{par})} \right)$

Metropolis et al., *JoCP*, 1953 - Hastings, *Biometrika*, 1970

Challenges

- `par` can be large dimensional (ARD/factor covariances)
 - No exact Gibbs steps
 - $p(\text{par}|\text{data})$ can be multi-modal
 - How to check convergence?

- `par` can be large dimensional (ARD/factor covariances)
- No exact Gibbs steps
- $p(\text{par}|\text{data})$ can be multi-modal
- How to check convergence?
- $p(\text{data}|\text{par})$ might be expensive to compute
- $p(\text{data}|\text{par})$ might not even be computable!



- Marginal likelihood

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

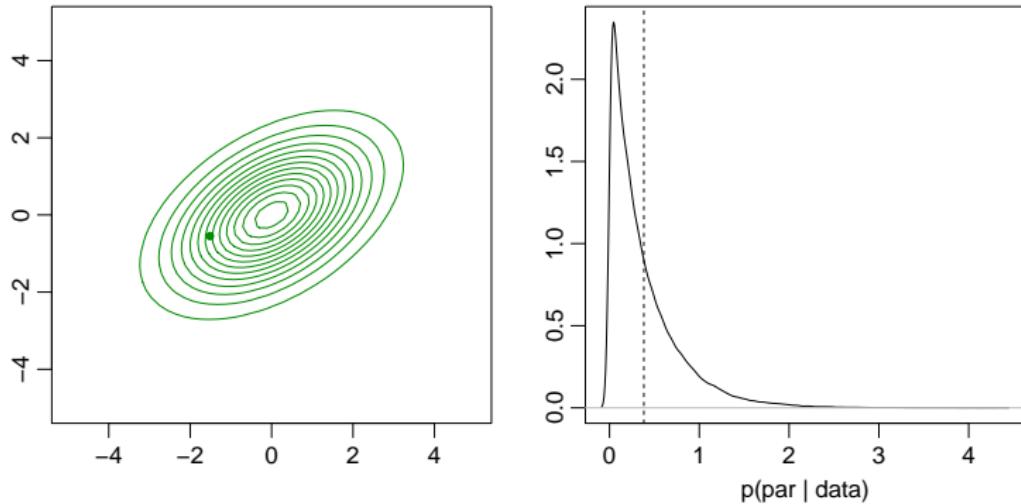
- Can only be computed if $p(\text{data}|\text{latent})$ is Gaussian
- ... even then

$$\log[p(\text{data}|\text{par})] = -\frac{1}{2} \log |K| - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} + \text{const.}$$

with K $n \times n$ dense matrix!

“Noisy” Markov chain Monte Carlo

$$\mathbb{E} \{ \tilde{p}(\text{data} | \text{par}) \} = p(\text{data} | \text{par})$$



Andrieu and Roberts, AoS, 2009 - Filippone and Girolami, IEEE-TPAMI, 2014

“Noisy” Markov chain Monte Carlo

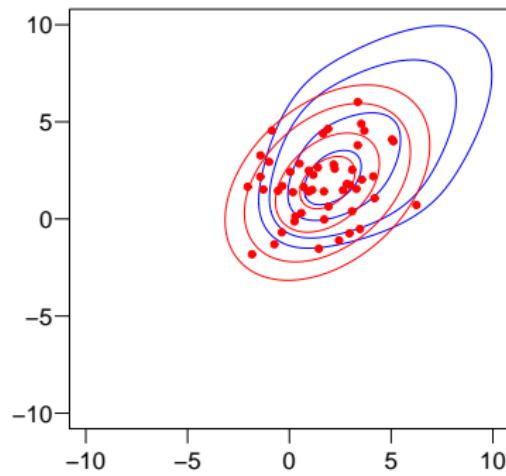
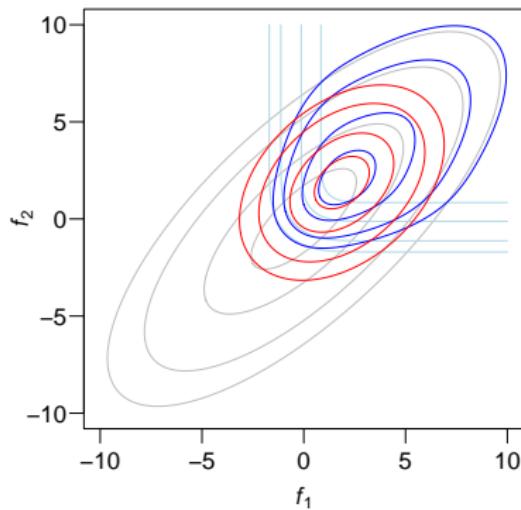
Acceptance probability : $\min \left(1, \frac{\tilde{p}(\text{data}|\text{par}')p(\text{par}')}{\tilde{p}(\text{data}|\text{par})p(\text{par})} \right)$

Andrieu and Roberts, *AoS*, 2009 - Filippone and Girolami, *IEEE-TPAMI*, 2014

Importance Sampling estimator

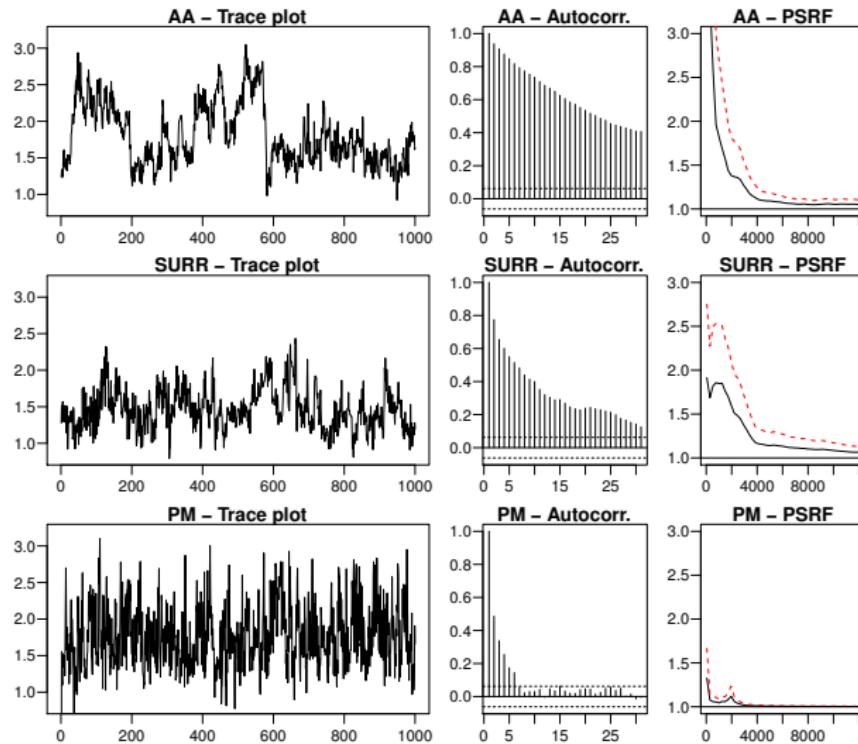
- Approximate posterior over latent variables using $q(\text{latent})$
- Then

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



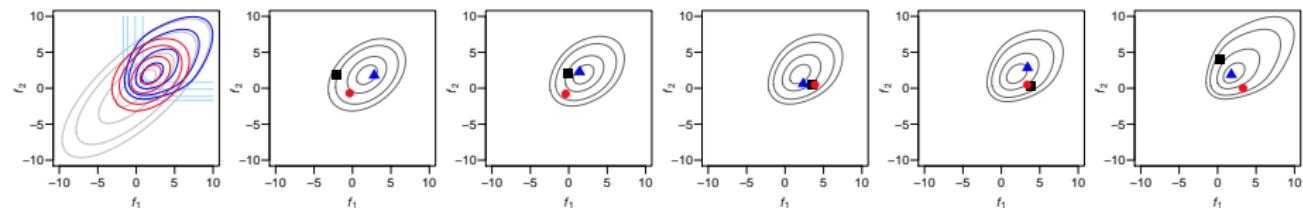
Convergence speed and efficiency

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



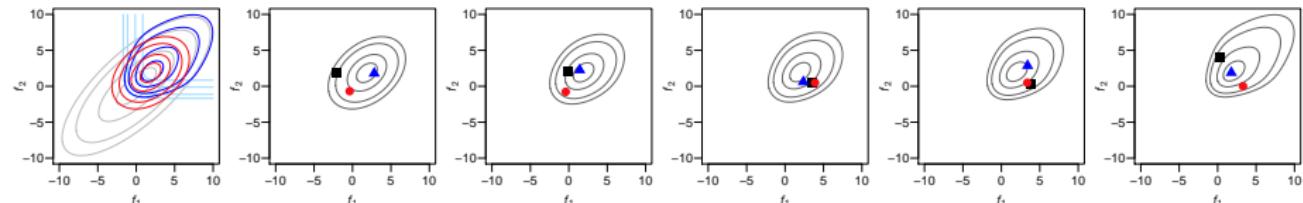
Annealed Importance Sampling estimator

- Annealing from an approximation

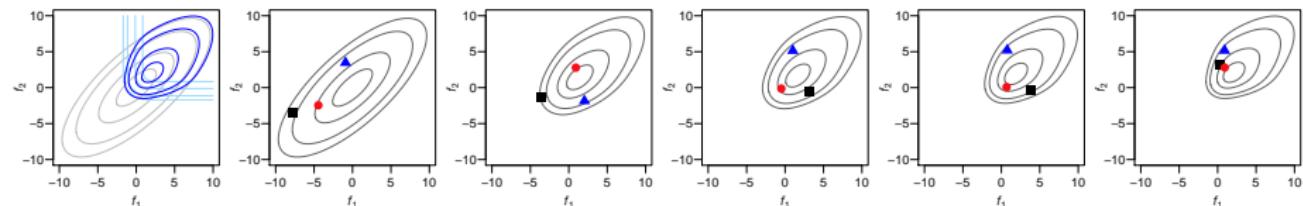


Annealed Importance Sampling estimator

- Annealing from an approximation



- Annealing from the prior

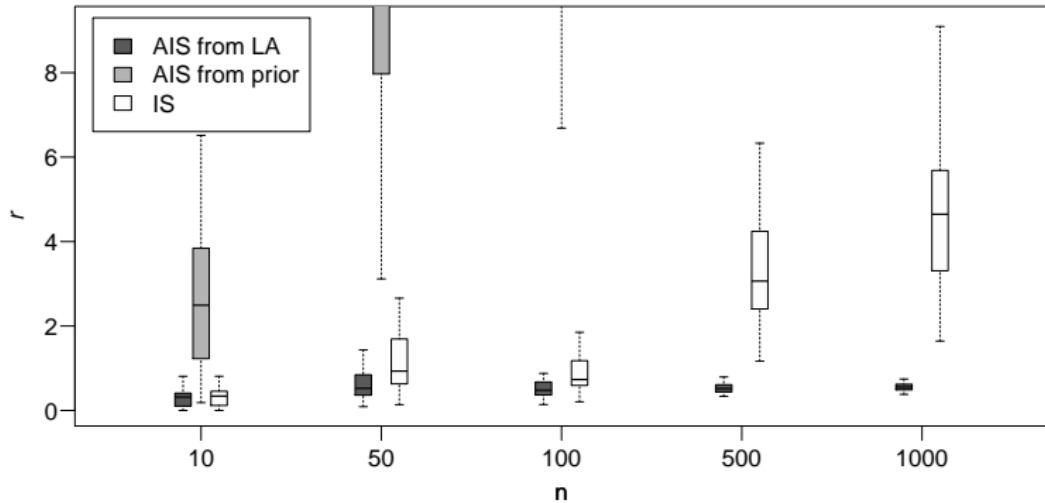


Filippone, ICPR, 2014

Comparison between AIS with IS

Analysis of the variance of the AIS and IS estimators

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

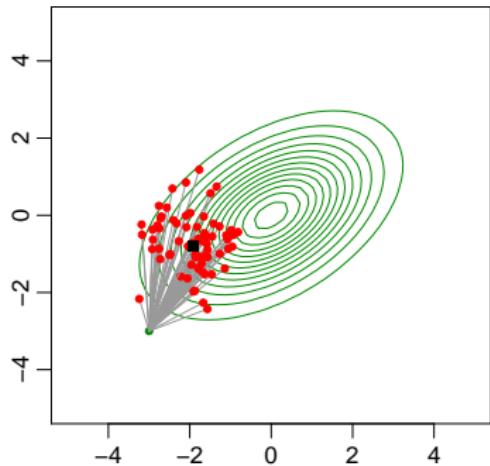
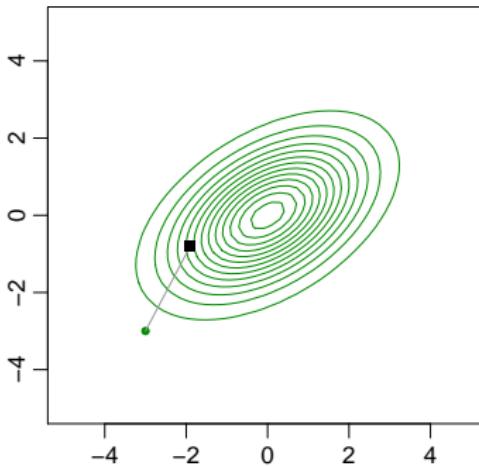


Gradient ascent

$$\text{par}' = \text{par} + \frac{\alpha}{2} \nabla_{\text{par}} \log[p(\text{data}|\text{par})p(\text{par})]$$

Stochastic Gradient ascent

$$E \left\{ \widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})] \right\} = \nabla_{\text{par}} \log[p(\text{data}|\text{par})]$$



Robbins and Monro, AoMS, 1951

Stochastic Gradient ascent

$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})p(\text{par})] \quad \alpha_t \rightarrow 0$$

Robbins and Monro, AoMS, 1951

Stochastic Gradient-based Markov chain Monte Carlo

$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})p(\text{par})] + \eta_t \quad \eta_t \sim \mathcal{N}(0, \alpha_t)$$

Stochastic Gradients in GP regression

- Marginal likelihood

$$\log[p(\text{data}|\text{par})] = -\frac{1}{2} \log |K| - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} + \text{const.}$$

- Derivatives wrt par

$$\frac{\partial \log[p(\text{data}|\text{par})]}{\partial \text{par}_i} = -\frac{1}{2} \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \text{par}_i} \right) + \frac{1}{2} \mathbf{y}^T K^{-1} \frac{\partial K}{\partial \text{par}_i} K^{-1} \mathbf{y}$$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\text{Tr} \left(K^{-1} \frac{\partial K}{\partial \text{par}_i} \right) = \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \text{par}_i} E[\mathbf{r}\mathbf{r}^T] \right) = E \left[\mathbf{r}^T K^{-1} \frac{\partial K}{\partial \text{par}_i} \mathbf{r} \right]$$

with $E[\mathbf{r}\mathbf{r}^T] = I$ - e.g., r_j drawn from $\{-1, 1\}$ with $p = 1/2$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

$$\text{Tr} \left(K^{-1} \frac{\partial K}{\partial \text{par}_i} \right) = \text{Tr} \left(K^{-1} \frac{\partial K}{\partial \text{par}_i} E[\mathbf{r}\mathbf{r}^T] \right) = E \left[\mathbf{r}^T K^{-1} \frac{\partial K}{\partial \text{par}_i} \mathbf{r} \right]$$

with $E[\mathbf{r}\mathbf{r}^T] = I$ - e.g., r_j drawn from $\{-1, 1\}$ with $p = 1/2$

- Stochastic gradient

$$-\frac{1}{2N_r} \sum_{i=1}^{N_r} \mathbf{r}^{(i)T} K^{-1} \frac{\partial K}{\partial \theta_i} \mathbf{r}^{(i)} + \mathbf{y}^T K^{-1} \frac{\partial K}{\partial \theta_i} K^{-1} \mathbf{y}$$

- Only linear systems!

Solving linear systems

- Linear systems:

$$Ks = b$$

- Can be solved using conjugate gradient:

$$s = \arg \min_x \left(\frac{1}{2} x^T K x - x^T b \right)$$

- Iterative update $s = s_0 + \delta_1 + \dots + \delta_T$
- Requires only Kv multiplications! $O(n^2)$ time
- No need to store K ! $O(n)$ space

- Accelerate the solution of dense linear systems
- ... returning an unbiased estimate of the solution

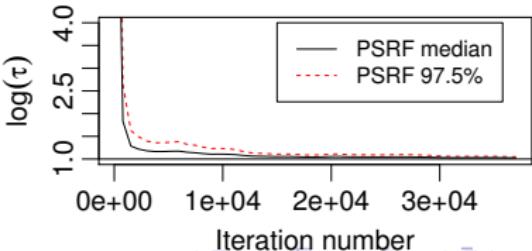
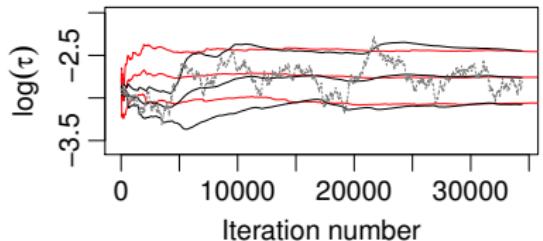
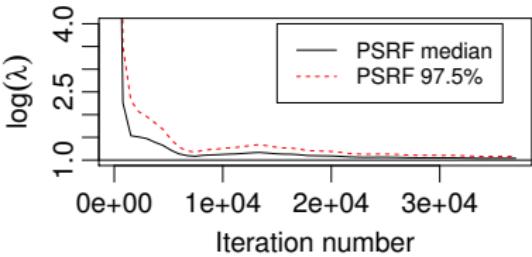
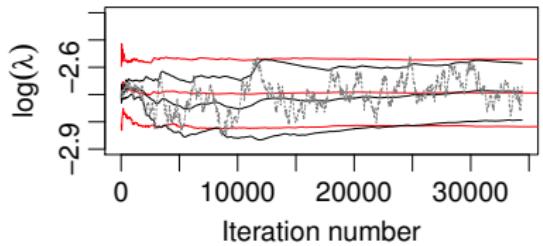
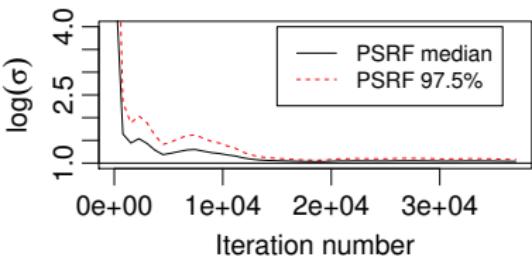
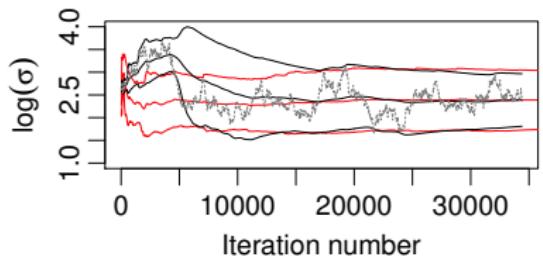
- Full CG solution:

$$\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \dots + \boldsymbol{\delta}_l + \boldsymbol{\delta}_{l+1} \dots + \boldsymbol{\delta}_T$$

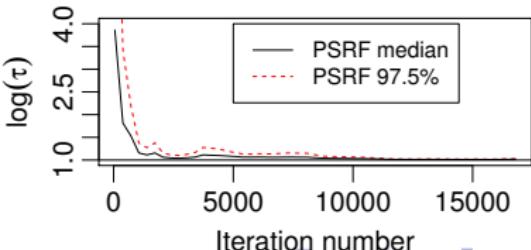
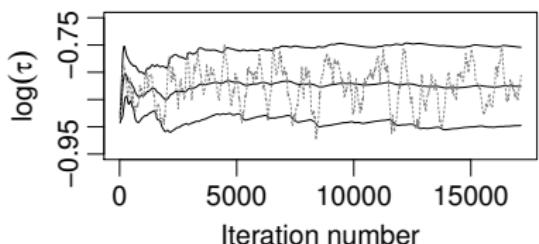
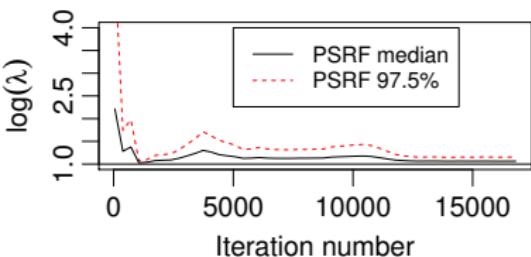
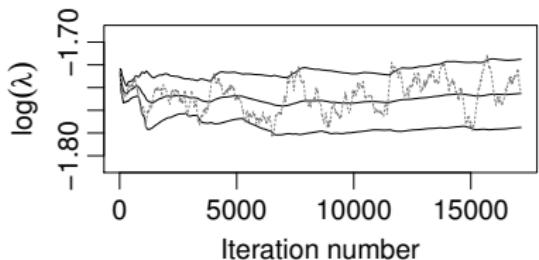
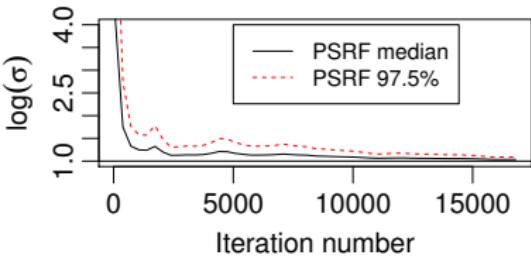
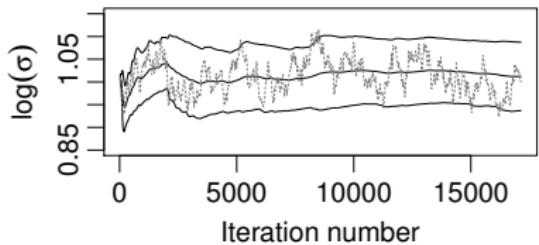
- ULISSE:

- Initialize $\tilde{\mathbf{s}} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \dots + \boldsymbol{\delta}_l$
- With probability p_1 set $\tilde{\mathbf{s}} = \tilde{\mathbf{s}} + \boldsymbol{\delta}_{l+1}/p_1$ otherwise stop
- With probability p_2 set $\tilde{\mathbf{s}} = \tilde{\mathbf{s}} + \boldsymbol{\delta}_{l+2}/(p_1 p_2)$ otherwise stop
- With probability p_3 set $\tilde{\mathbf{s}} = \tilde{\mathbf{s}} + \boldsymbol{\delta}_{l+3}/(p_1 p_2 p_3)$ otherwise stop
- ...
- $\tilde{\mathbf{s}}$ is an unbiased estimate of \mathbf{s} !

Comparison with MCMC - Concrete dataset - $n \approx 1K$



Larger n - Census dataset - $n \approx 23K$



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
 - “Noisy” MCMC offers a practical and scalable way to carry out exact Bayesian computations for GPs

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 R. Engler. Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE), In *ICML*, 2015.
- [2] M. Filippone and M. Girolami. Pseudo-Marginal Bayesian inference for Gaussian processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214-2226, 2014.
- [3] M. Filippone. Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC, In *ICPR*, 2014.
- [4] 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.
- [5] A. F. Marquand et al. Automated, high accuracy classification of Parkinsonian disorders: a pattern recognition approach. *PLoS ONE*, 2013.
- [6] M. Filippone et al. A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, 93(1):93-114, 2013.