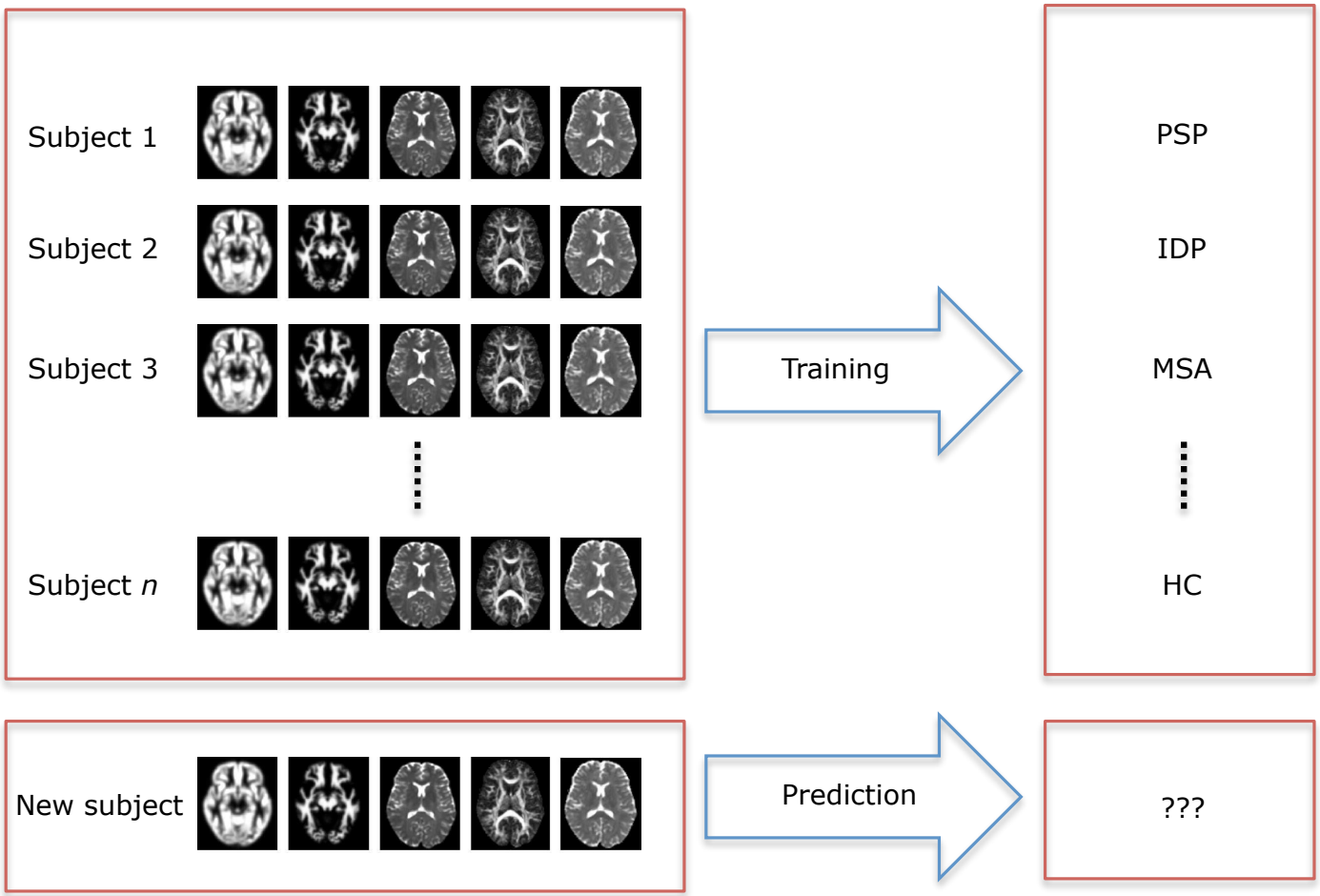
Unbiased computations for tractable and scalable learning of Gaussian processes

Maurizio Filippone

Department of Data Science, EURECOM
maurizio.filippone@eurecom.fr

May 23rd, 2016

Motivating Application



HC - Healthy control

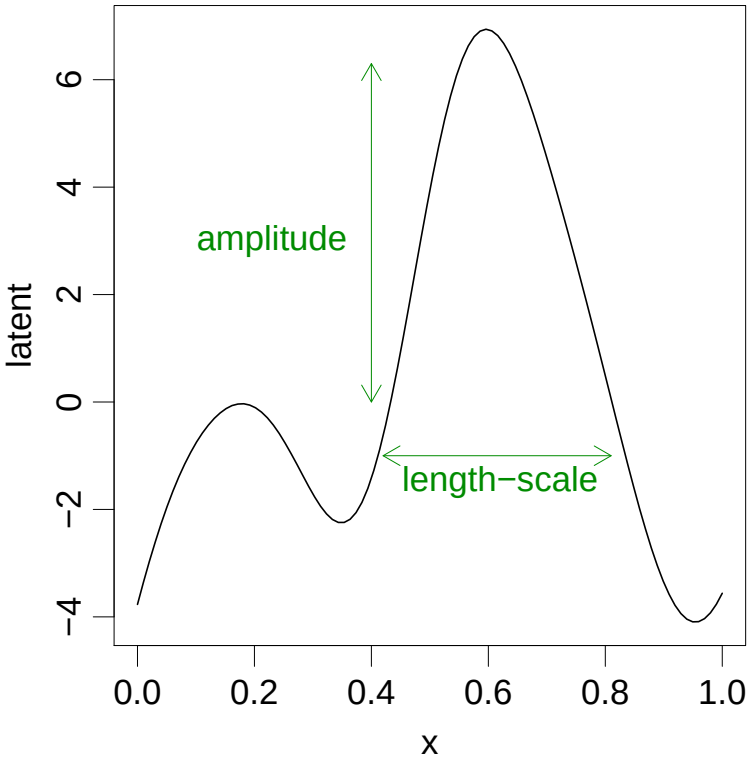
MSA - Multiple system atrophy

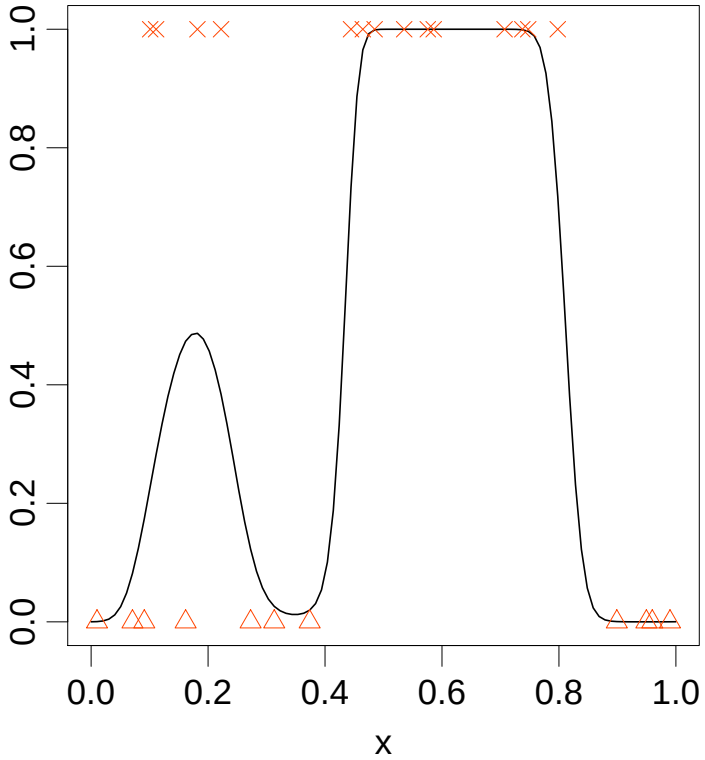
PSP - Progressive Supranuclear Palsy

IDP - Idiopathic Parkinson's disease

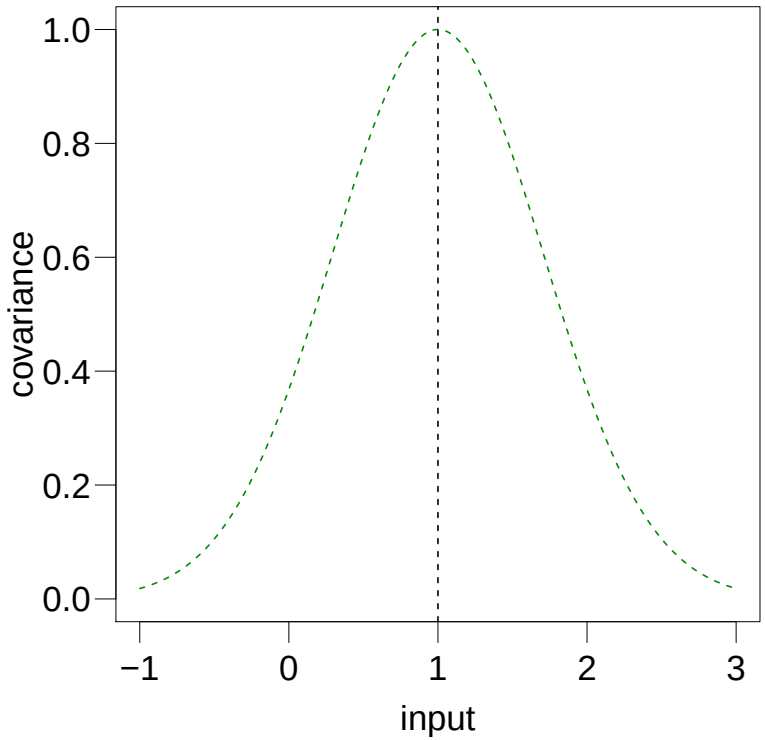
Filippone, Marquand et al., *AoAS*, 2012

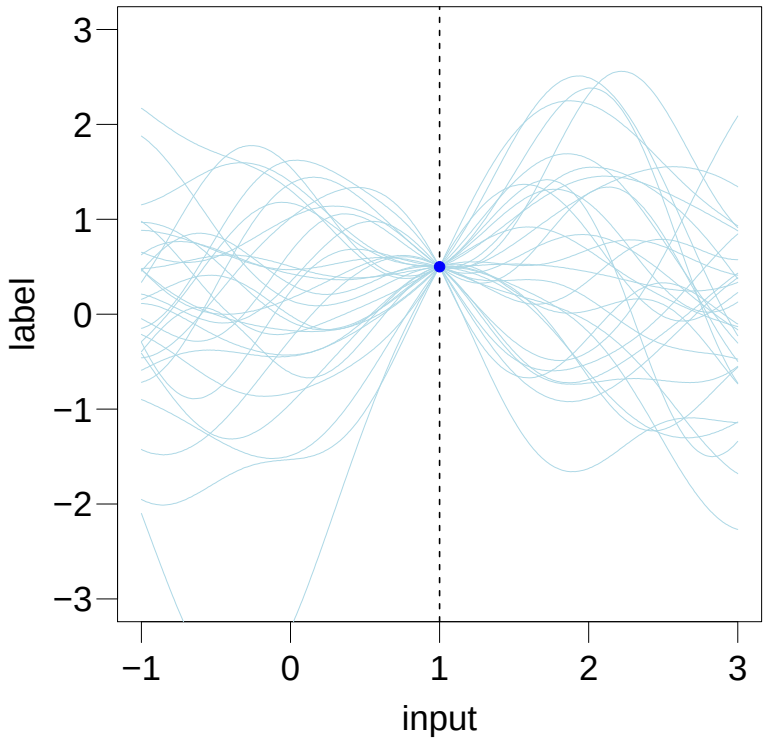
Gaussian Process Models - Classification example





- Gaussians with distance dependent covariance





Multiclass classification with multiple sources

- Multiclass classification based on GPs

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

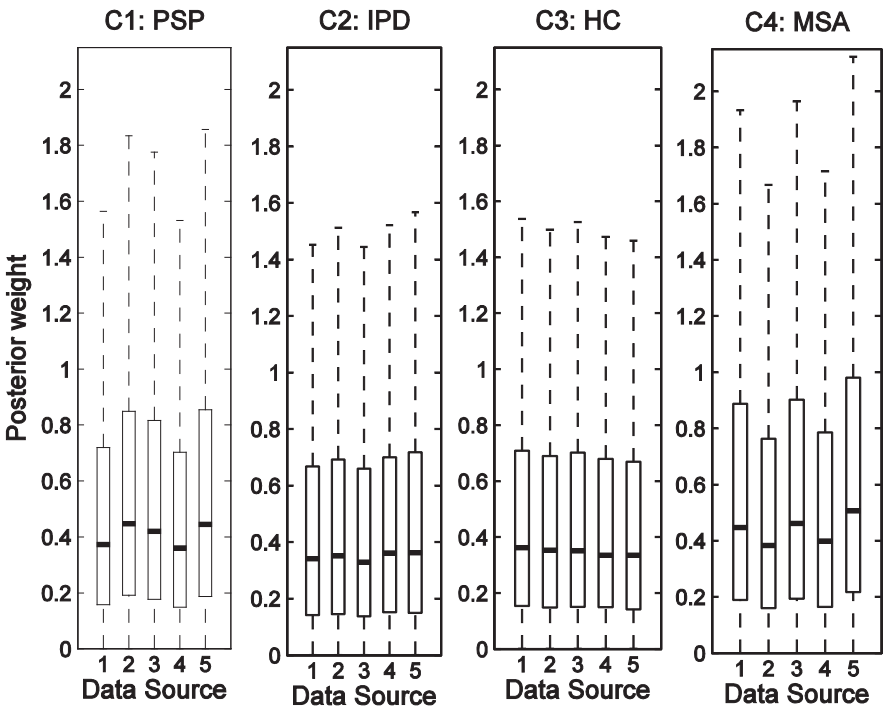
- unknown function modeled using GPs
- Covariance based on source-dependent covariances S_k

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

Filippone, Marquand et al., *AoAS*, 2012

Parkinsonian disorders data - multiple sources classification

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418



Filippone, Marquand et al., *AoAS*, 2012

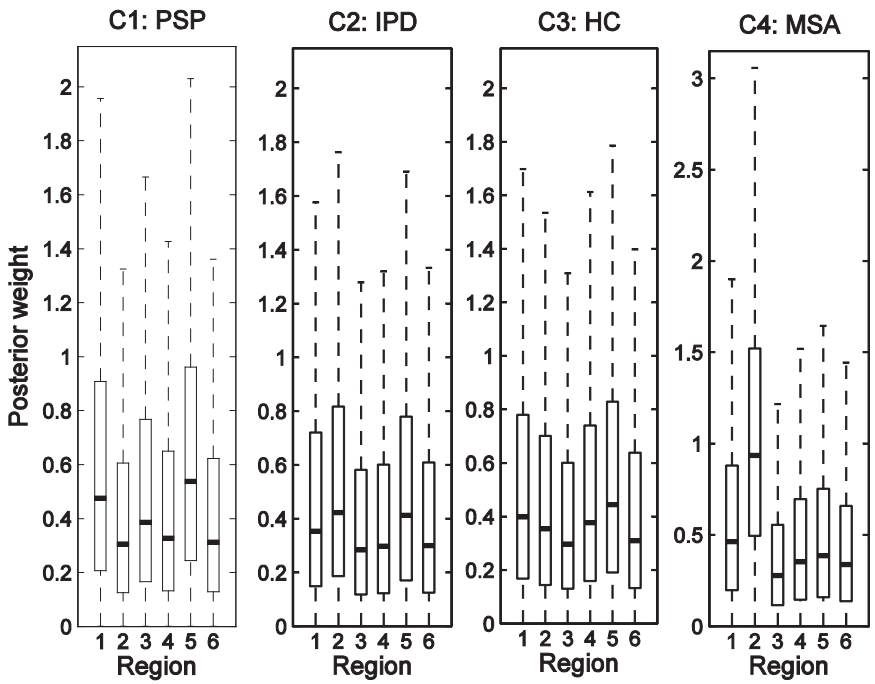
Multiclass classification with multiple regions

Analysis of brain regions

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

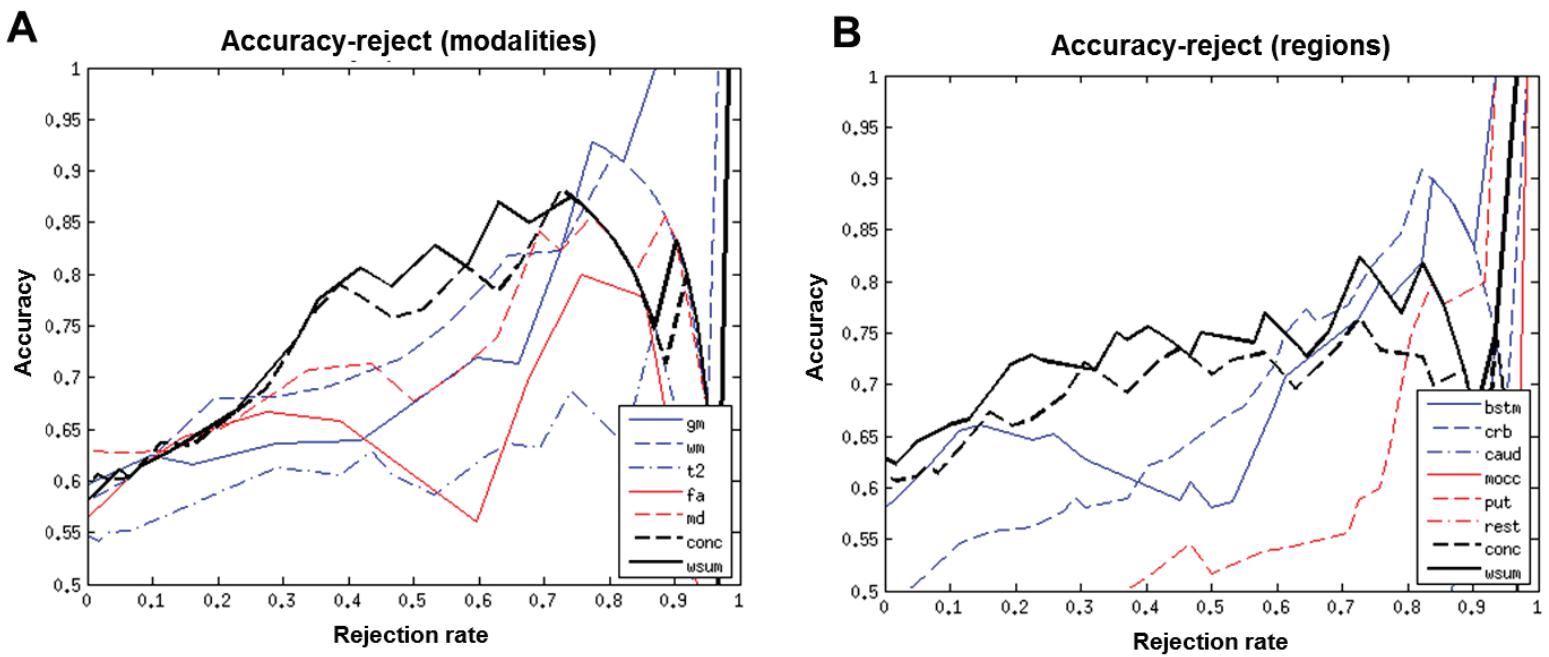
Parkinsonian disorders data - multiple regions classification

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



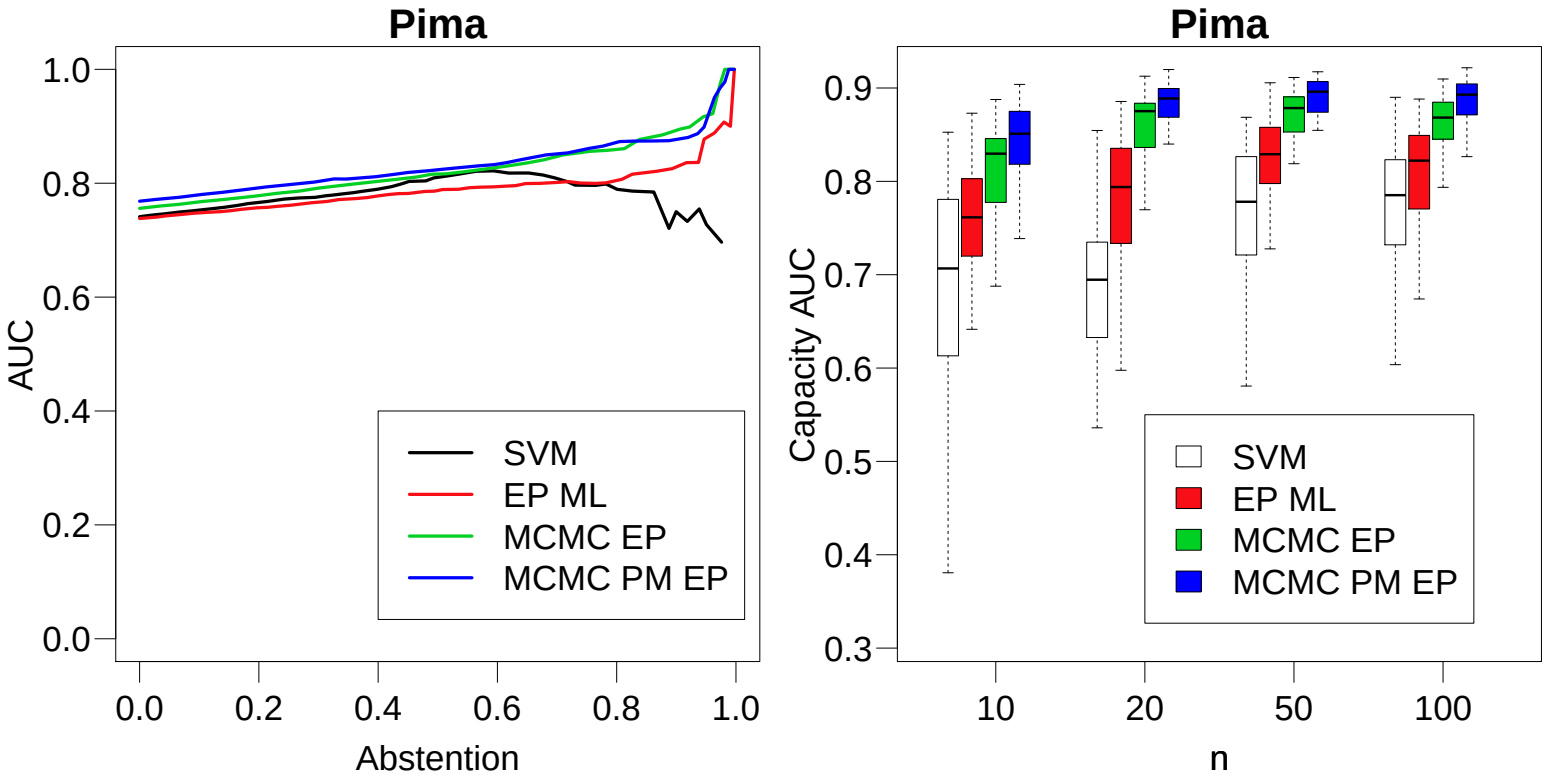
Filippone, Marquand et al., *AoAS*, 2012

Reject Option



Filippone, Marquand et al., *AoAS*, 2012

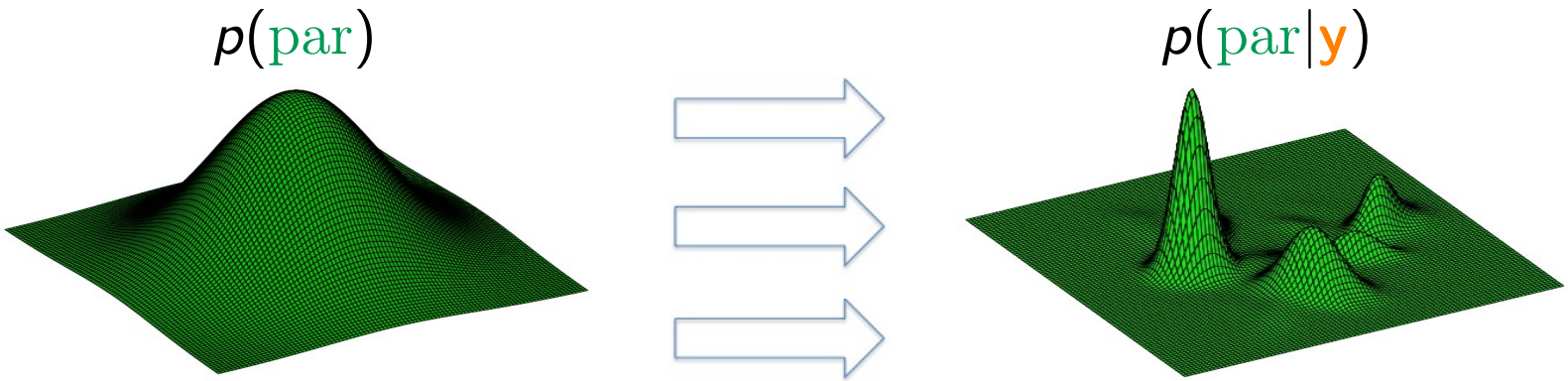
Does being Bayesian buy you anything?



Filippone and Girolami, *IEEE-TPAMI*, 2014

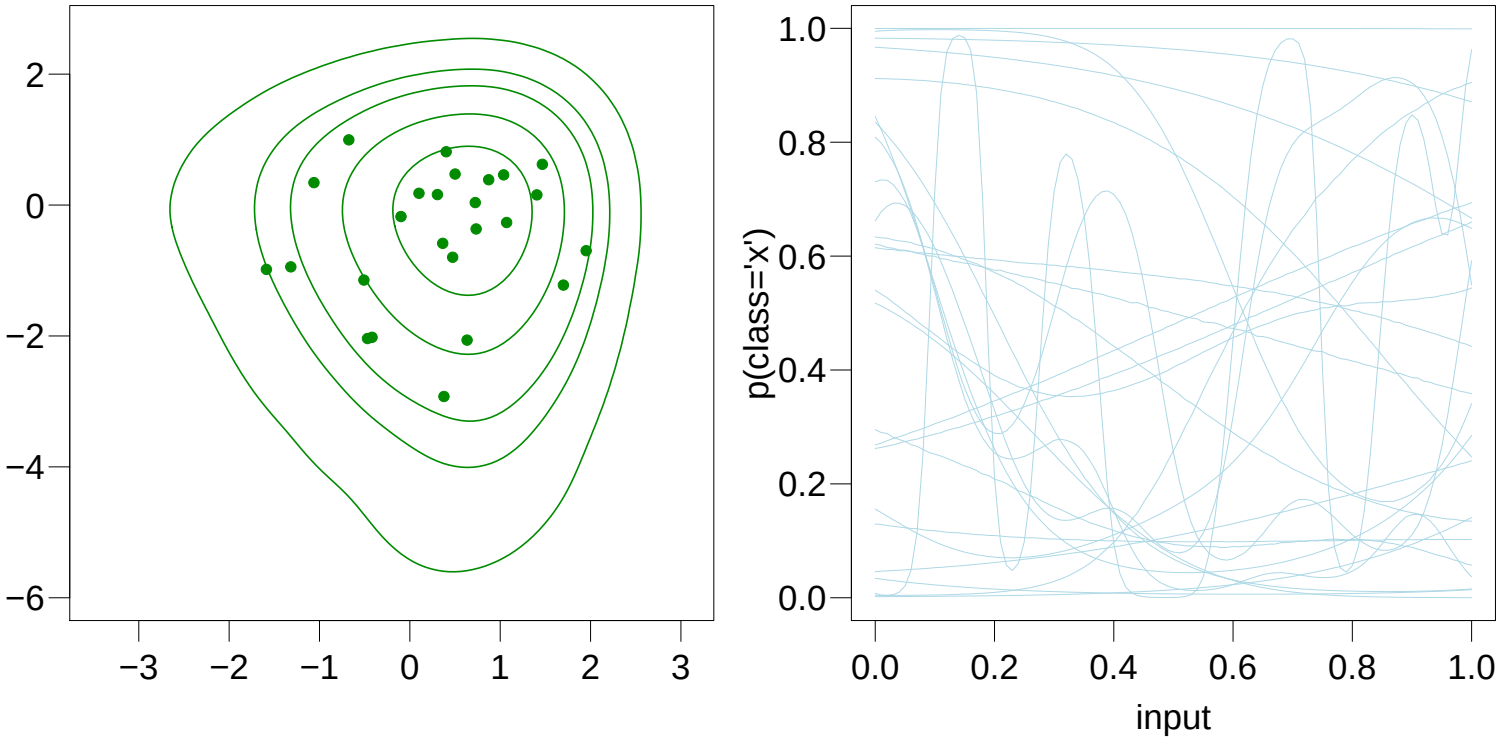
Bayesian Inference

- Inputs $= X$ Labels $= y$
- $K = K(X, \text{par})$

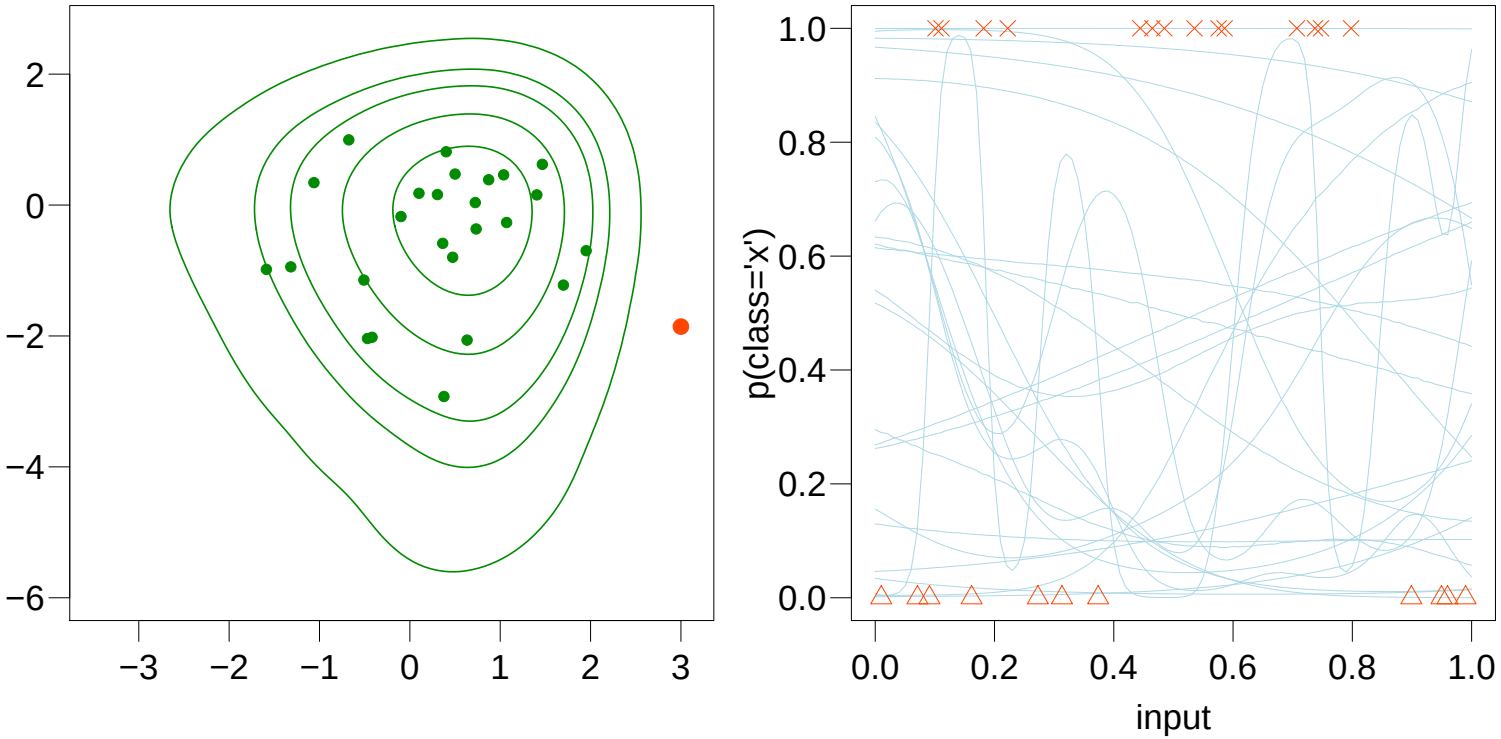


$$p(\text{par}|\mathbf{y}) = \frac{p(\mathbf{y}|\text{par})p(\text{par})}{\int p(\mathbf{y}|\text{par})p(\text{par})d\text{par}}$$

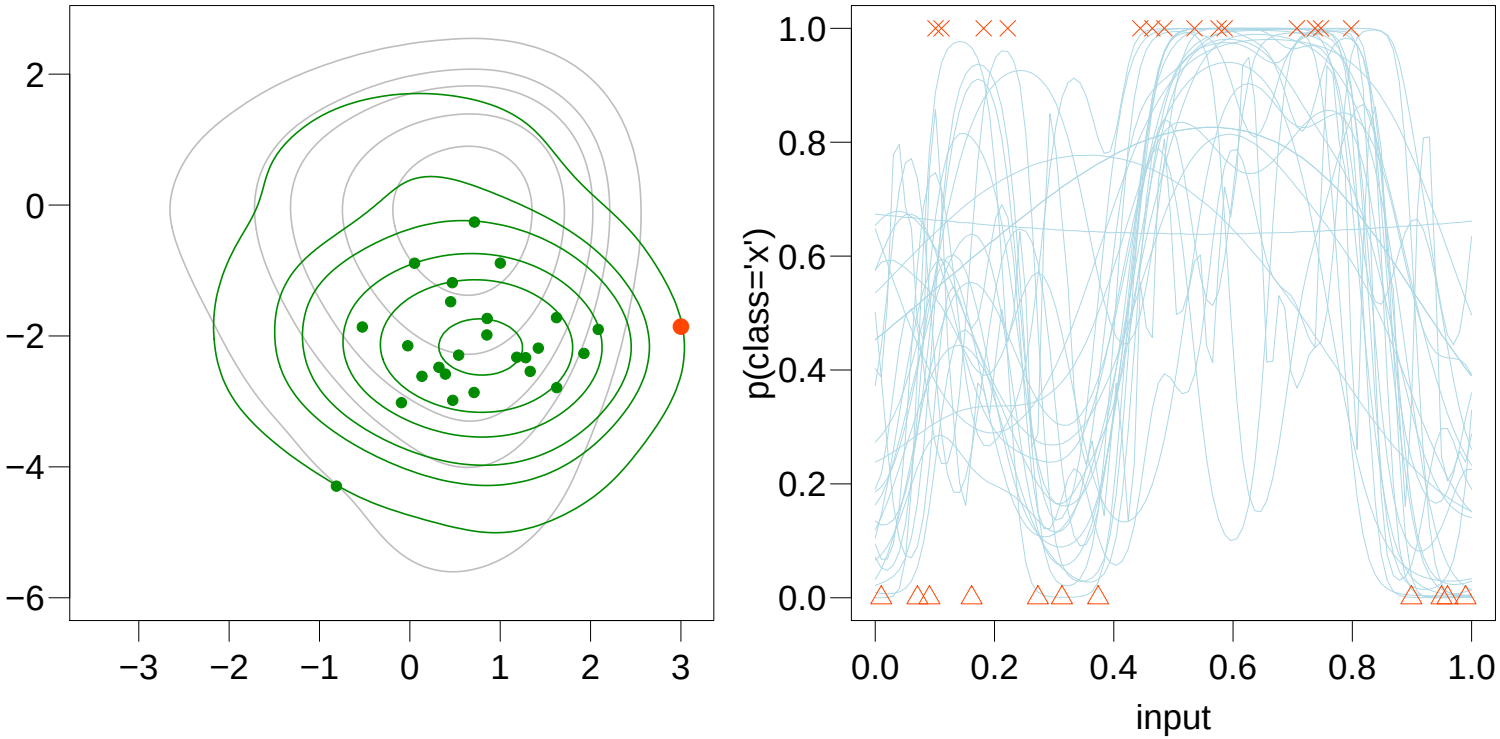
Bayesian Inference - Prior



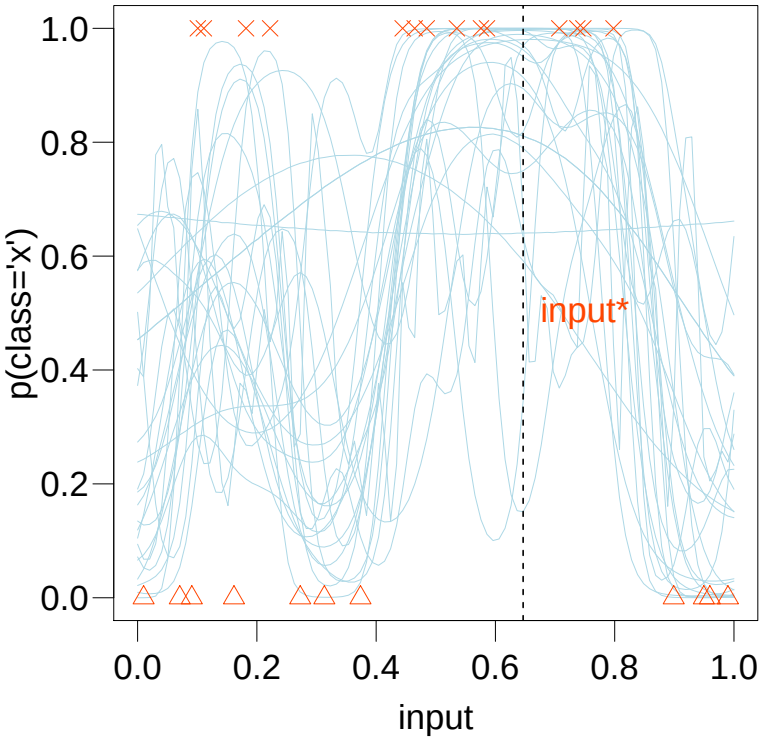
Bayesian Inference - Data

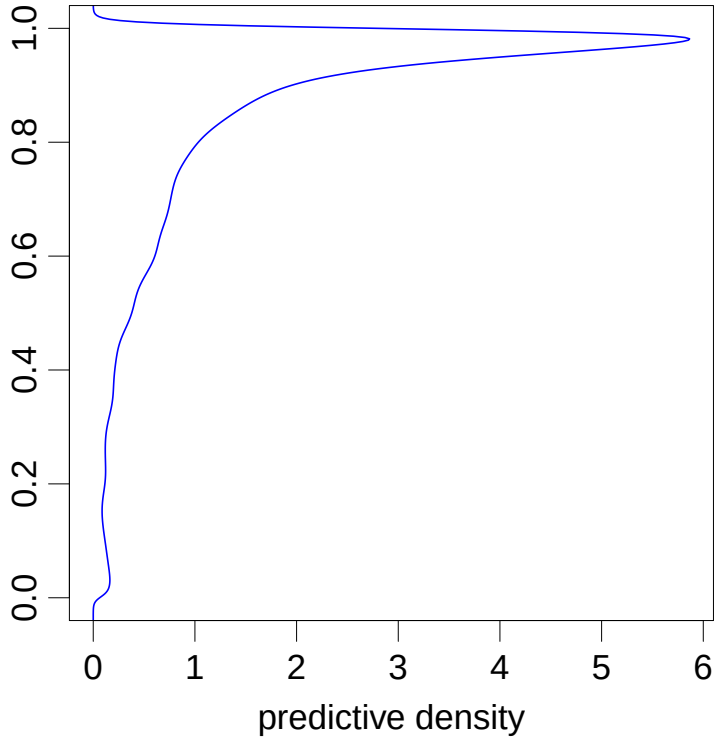


Bayesian Inference - Posterior



Bayesian Inference and Predictions





Bayesian Inference and Predictions

- Predictions for new data

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

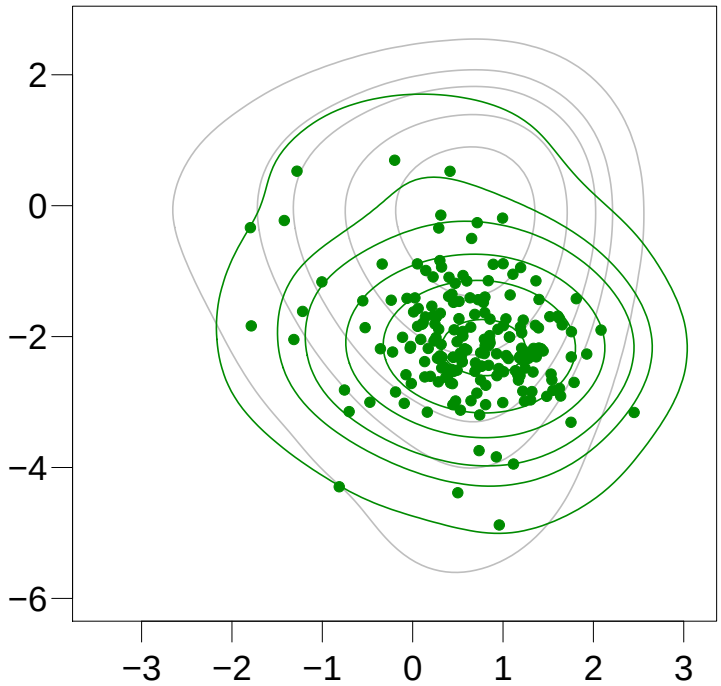
- Monte Carlo integration:

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

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

Bayesian Inference and Predictions

- Draw samples according to the posterior density



Markov chain Monte Carlo

- Bayesian inference

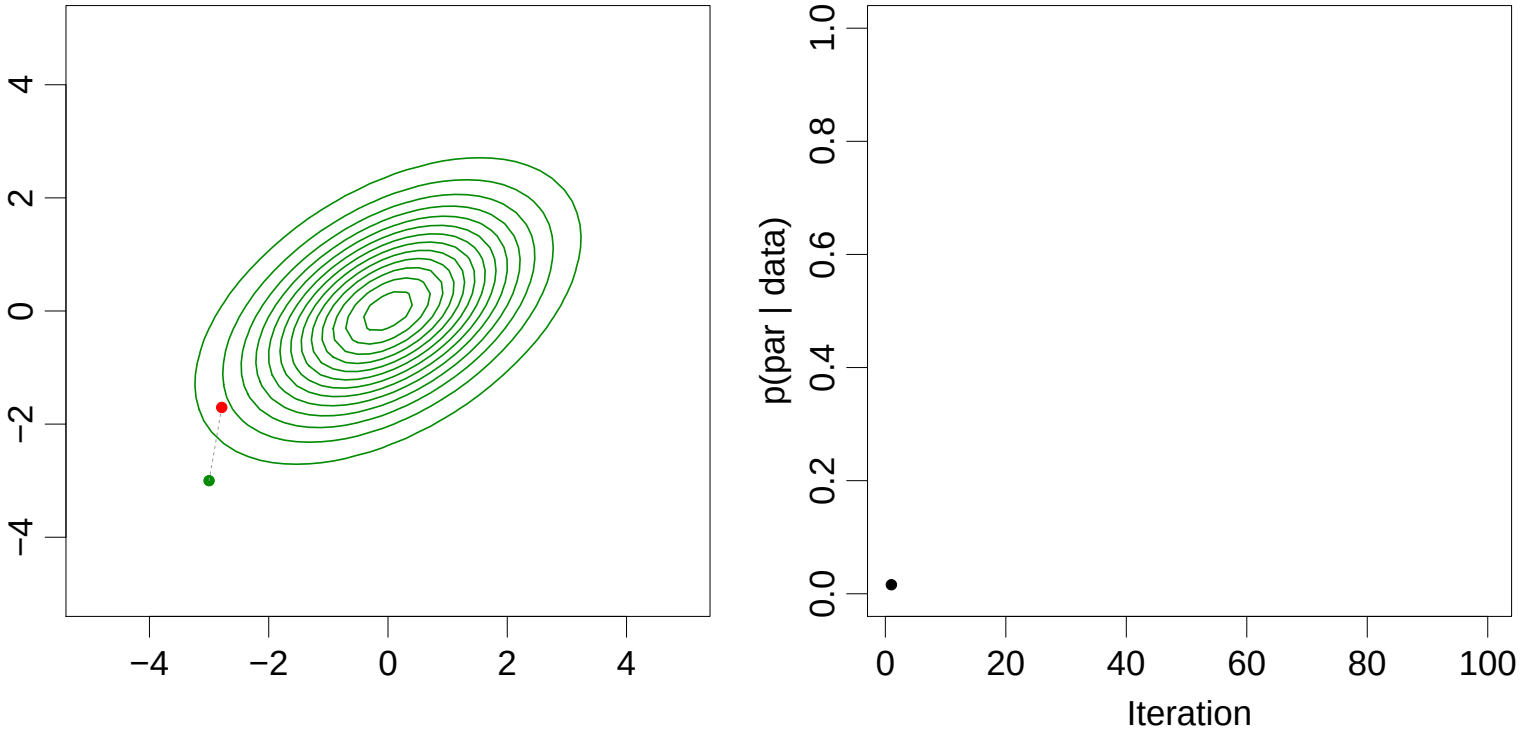
$$p(\mathrm{par}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathrm{par})p(\mathrm{par})}{\int p(\mathrm{data}|\mathrm{par})p(\mathrm{par})d\mathrm{par}}$$

- Random walk sampler - accept a proposal with probability

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

Markov chain Monte Carlo - Random walk example

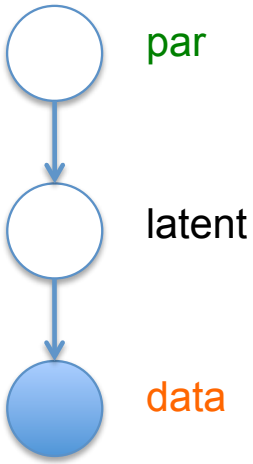
$$\text{Acceptance probability : } \min\left(1, \frac{p(\mathsf{y}|\text{par}')p(\text{par}')}{p(\mathsf{y}|\text{par})p(\text{par})}\right)$$



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

Challenges in Gaussian processes

- $p(\mathrm{y}|\mathrm{par})$ might be expensive to compute
- $p(\mathrm{y}|\mathrm{par})$ might not even be computable!



Marginal likelihood in GP models

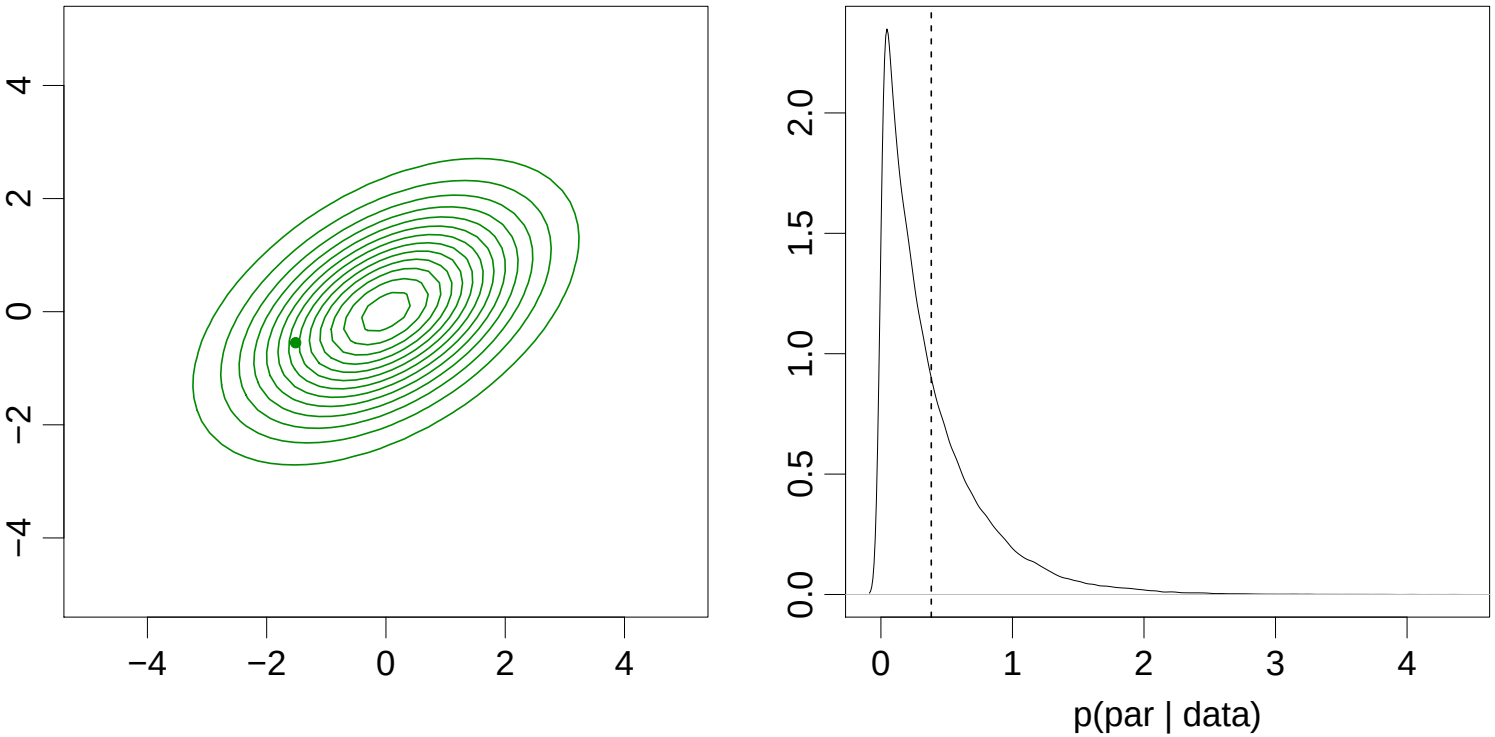
- Marginal likelihood

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

  can only be computed if $p(\mathbf{y}|\text{latent})$ is Gaussian
- What if $p(\mathbf{y}|\text{latent})$ is **not** Gaussian?

"Noisy" Markov chain Monte Carlo

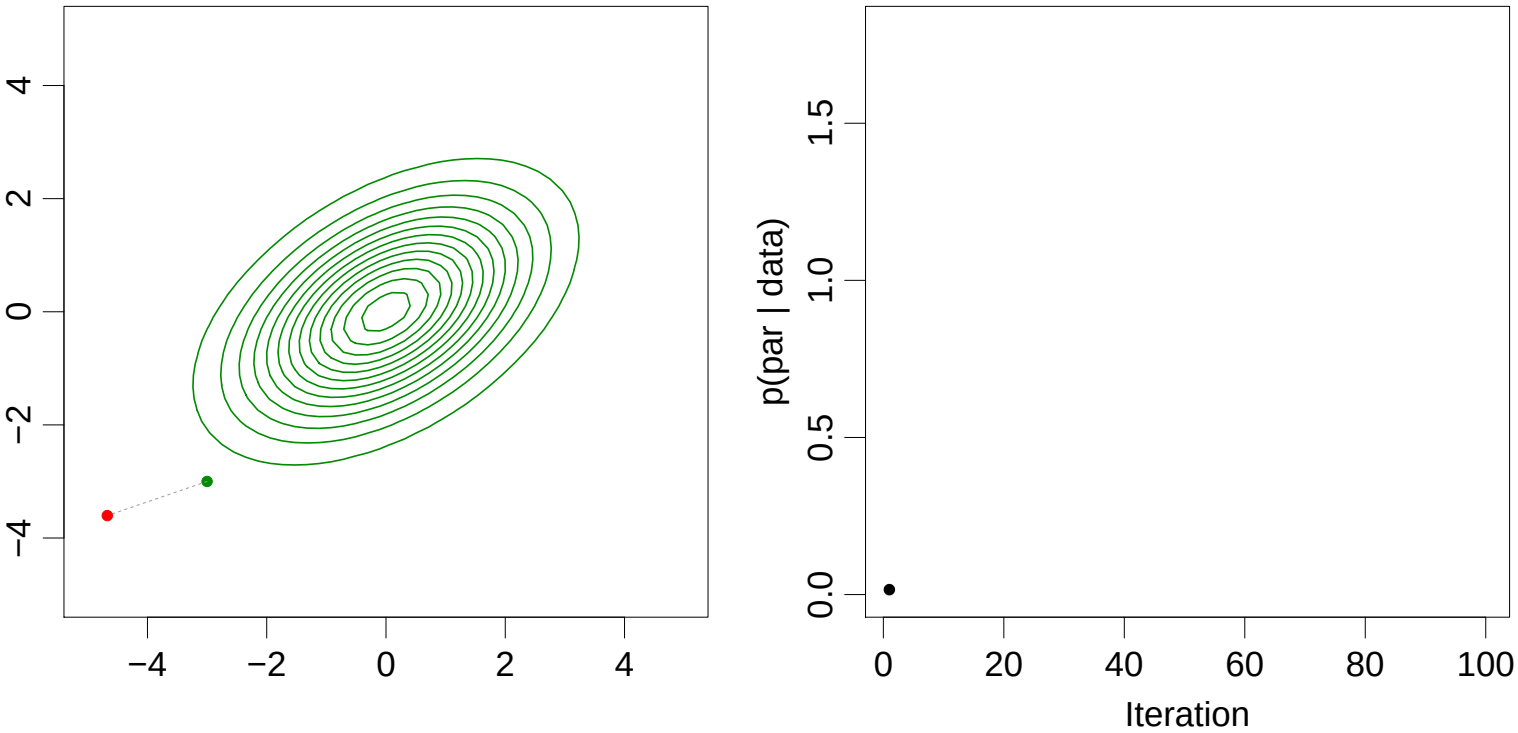
$$\mathrm{E}\left\{\tilde{p}(\mathbf{y}|\mathrm{par})\right\} = p(\mathbf{y}|\mathrm{par})$$



Andrieu and Roberts, *AoS*, 2009

"Noisy" Markov chain Monte Carlo

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

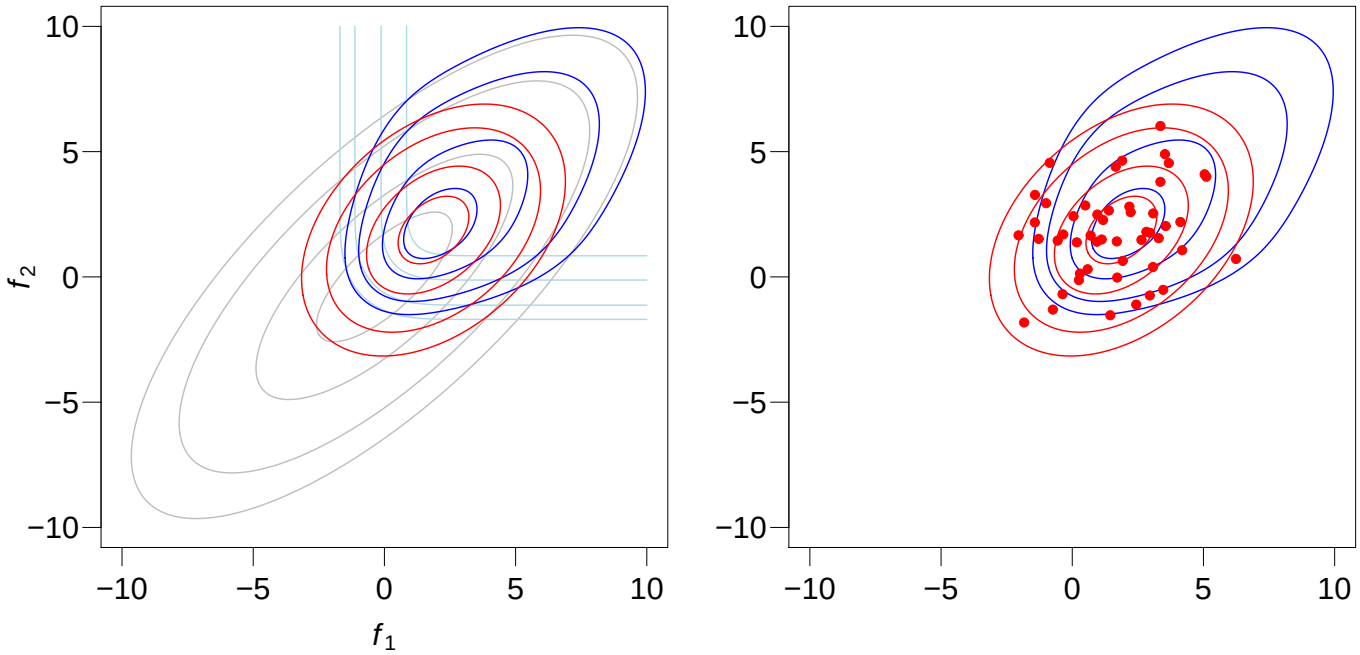


Andrieu and Roberts, *AoS*, 2009

Importance Sampling estimator

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

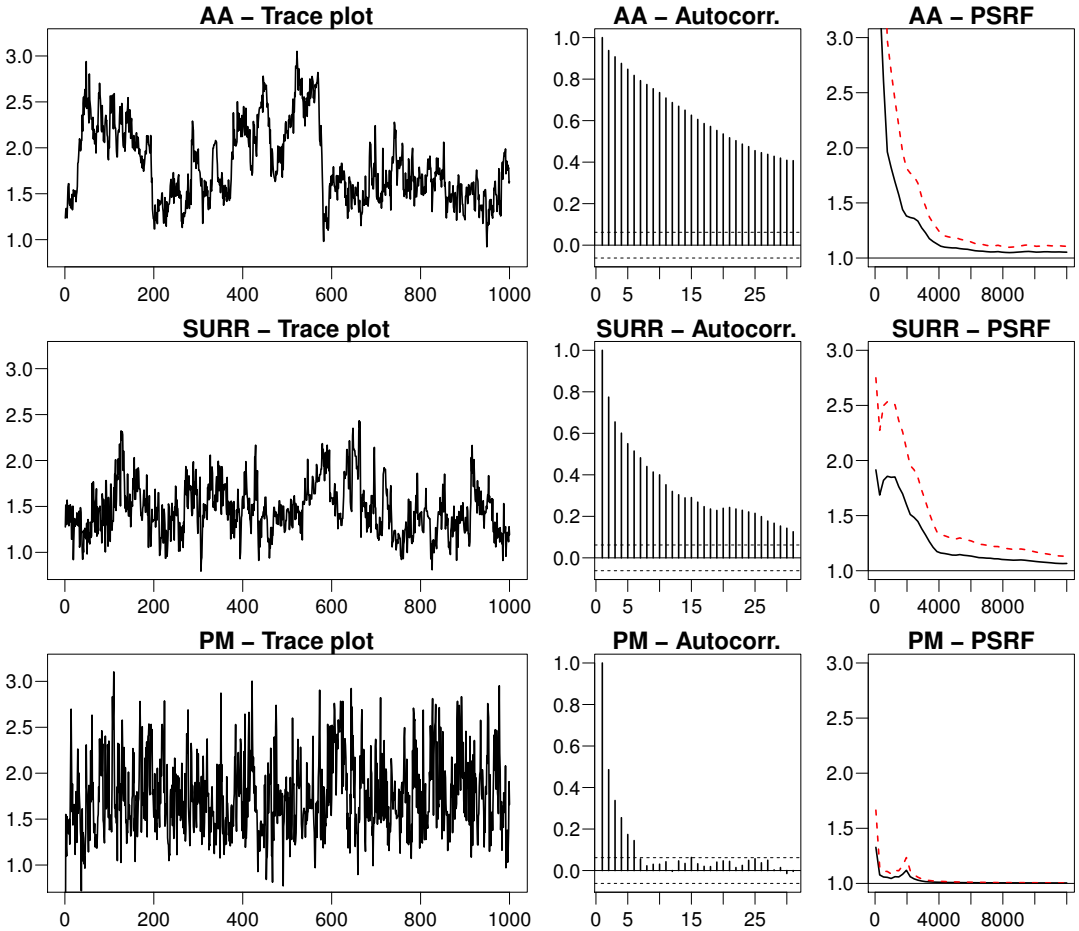
$$\tilde{p}(\mathbf{y}|\text{par}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(\mathbf{y}|\text{latent}^{(i)})p(\text{latent}^{(i)}|\text{par})}{q(\text{latent}^{(i)})}$$



Filippone and Girolami, *IEEE-TPAMI*, 2014 - Filippone, *ICPR*, 2014

Convergence speed and efficiency

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



Filippone and Girolami, *IEEE-TPAMI*, 2014 - Filippone, *ICPR*, 2014

Marginal likelihood in GP models

- Marginal likelihood

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

  can only be computed if $p(\mathbf{y}|\text{latent})$ is Gaussian

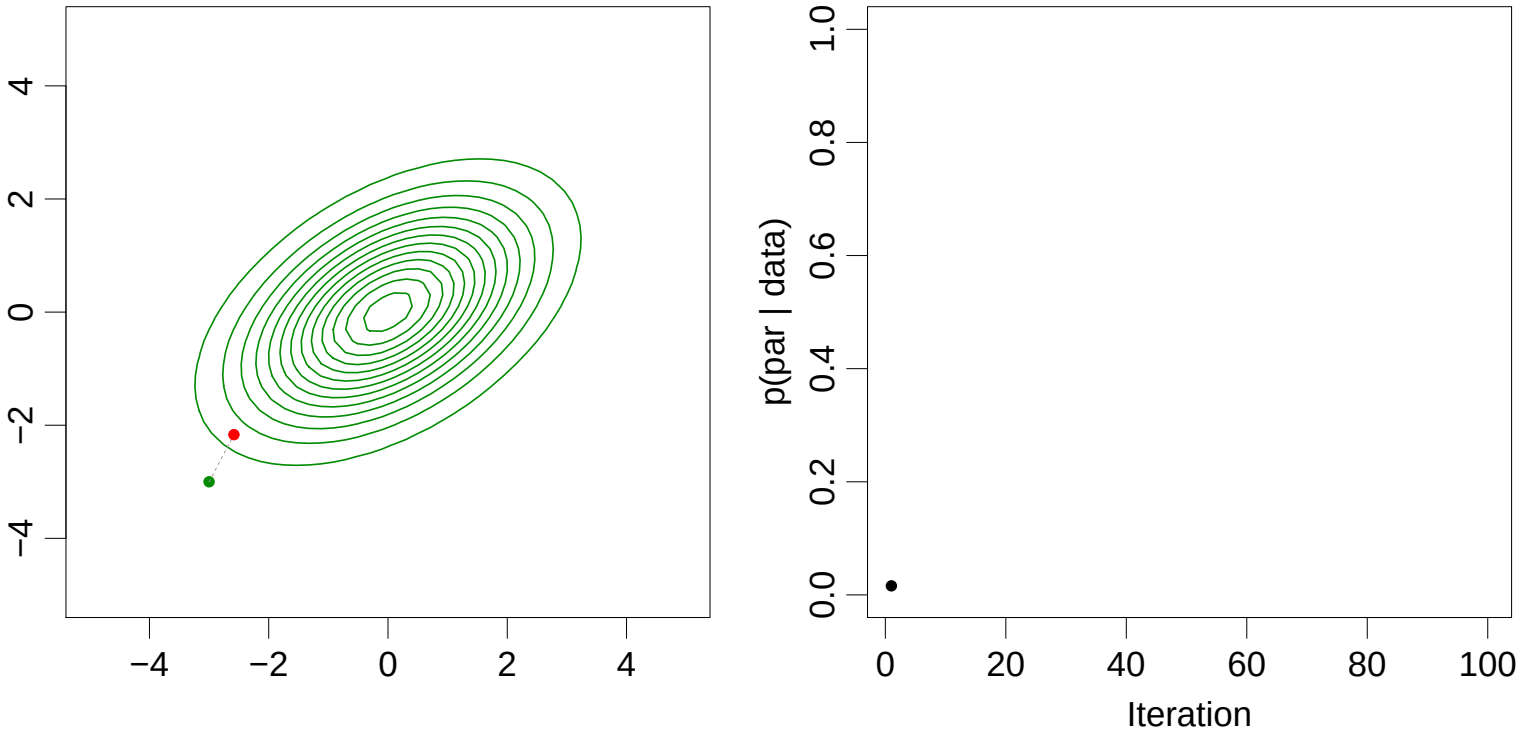
- ... even then

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

  where $K = K(X, \text{par})$ is a $n \times n$ dense matrix!

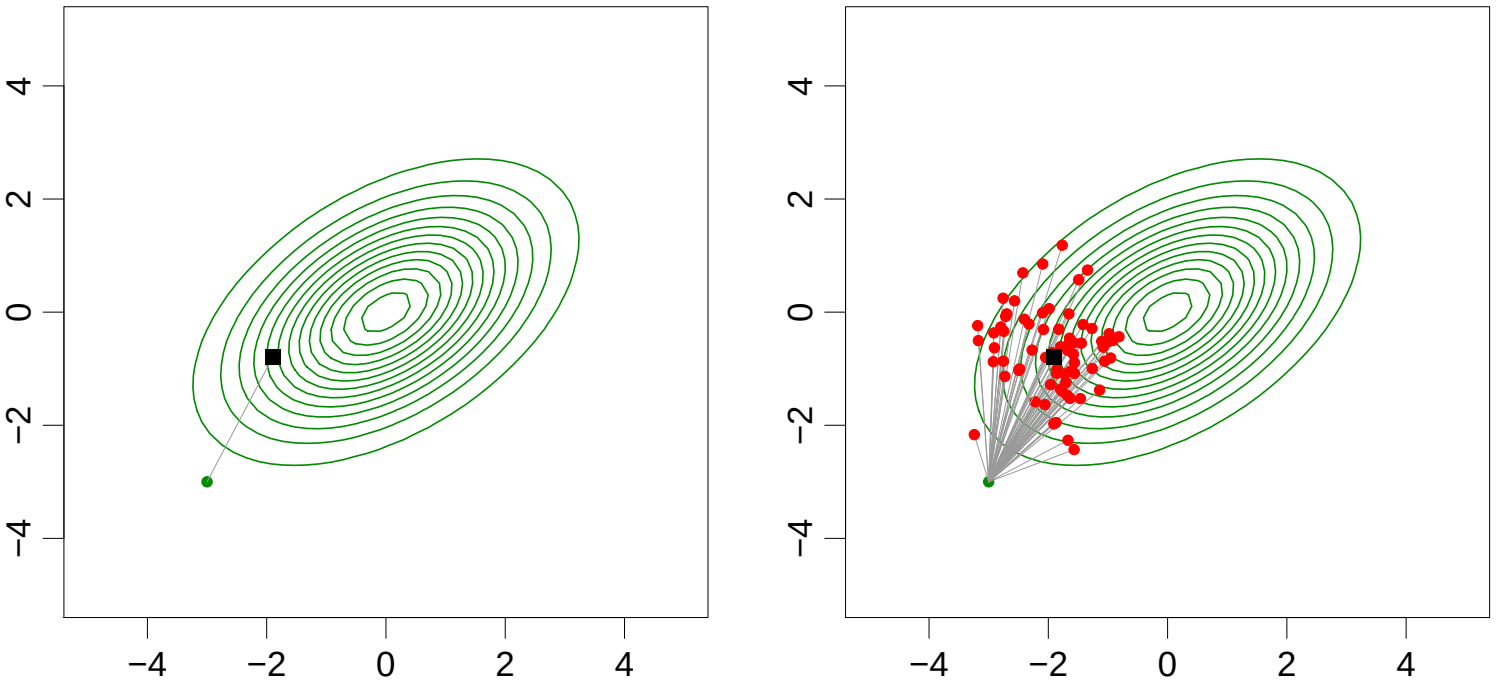
Gradient ascent

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



Stochastic Gradient ascent

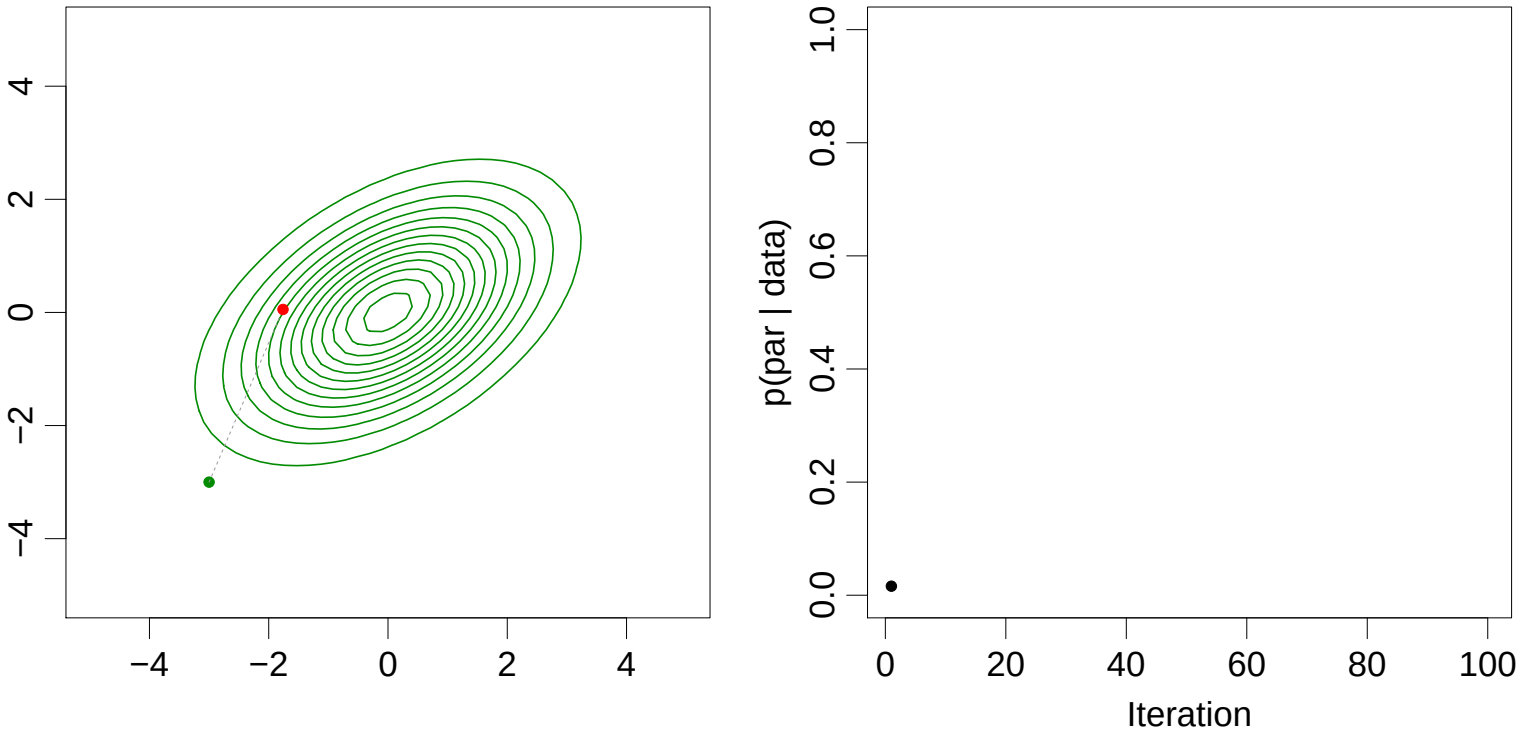
$$\mathrm{E}\left\{\widetilde{\nabla_{\mathrm{par}}} \log[p(\mathsf{y}|\mathrm{par})]\right\} = \nabla_{\mathrm{par}} \log[p(\mathsf{y}|\mathrm{par})]$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient ascent

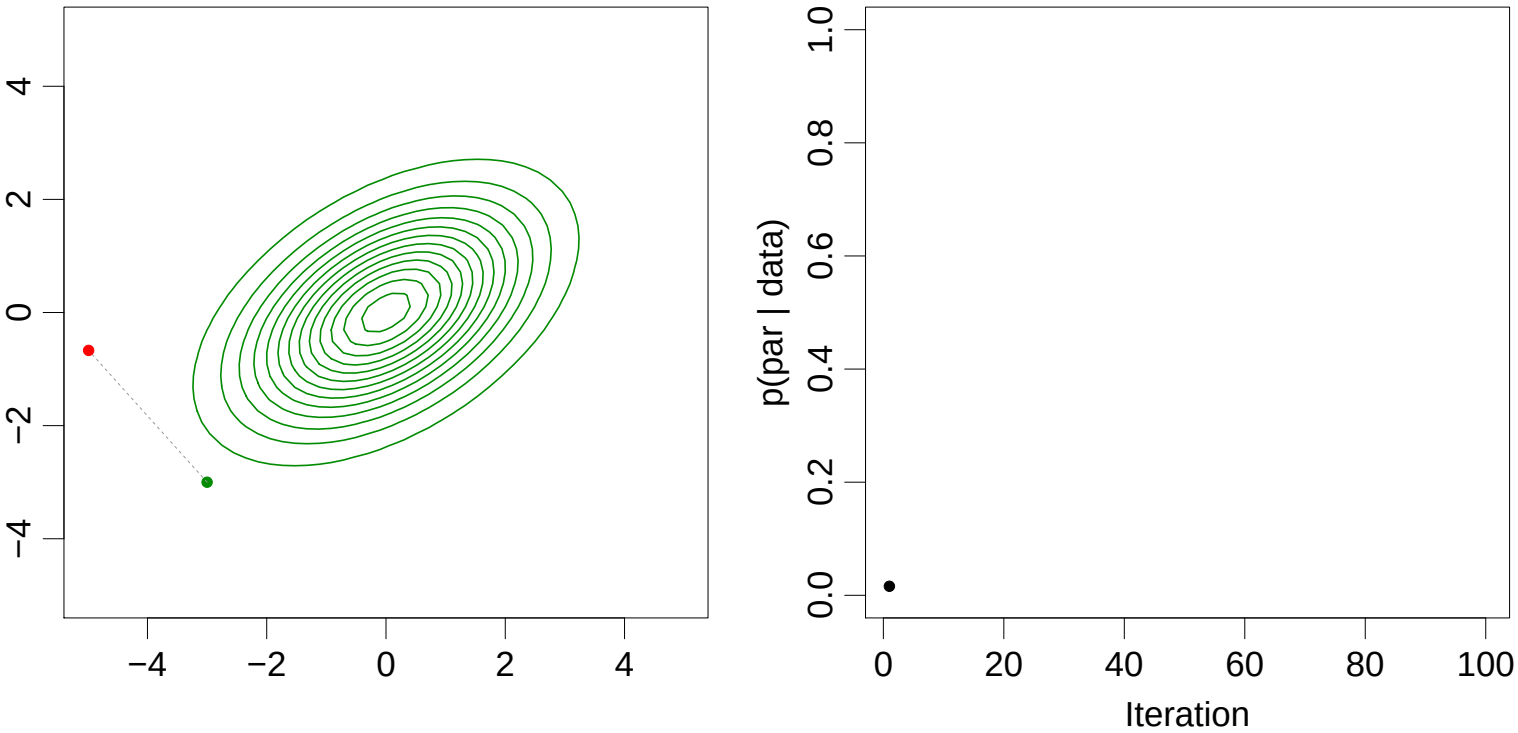
$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\mathsf{y}|\text{par})p(\text{par})] \qquad \alpha_t \to 0$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient Langevin Dynamics (SGLD) algorithm

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



Welling and Teh, *ICML*, 2011 - Filippone and Engler, *ICML*, 2015

Stochastic Gradients in GP regression

- Marginal likelihood

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

- Derivatives wrt par

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

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

  with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$

- Stochastic gradient

$$-\frac{1}{2N_{\mathbf{r}}}\sum_{i=1}^{N_{\mathbf{r}}}\mathbf{r}^{(i)\mathrm{T}}K^{-1}\frac{\partial K}{\partial \mathrm{par}_i}\mathbf{r}^{(i)} + \frac{1}{2}\mathbf{y}^{\mathrm{T}}K^{-1}\frac{\partial K}{\partial \mathrm{par}_i}K^{-1}\mathbf{y}$$

- Linear systems only!

Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Solving linear systems

- Linear systems:

$$K\mathbf{s} = \mathbf{b}$$

- Can be solved using conjugate gradient:

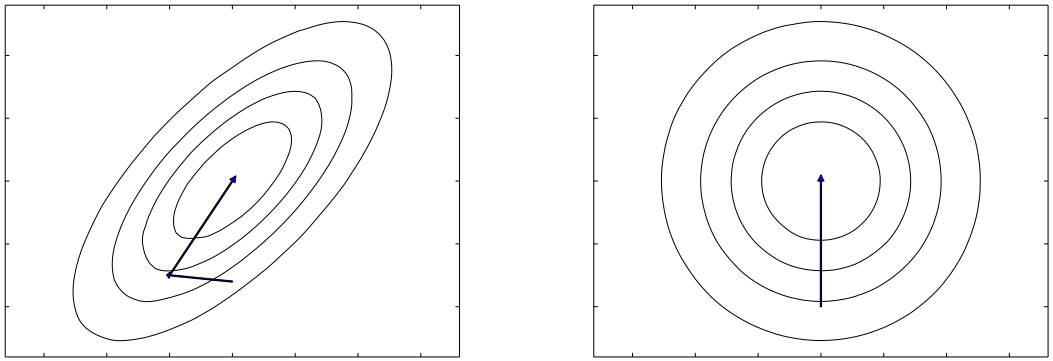
$$\mathbf{s} = \arg\min_{\mathbf{x}} \left(\frac{1}{2}\mathbf{x}^{\mathrm{T}} K\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{b} \right)$$

- Iterative update $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \ldots + \boldsymbol{\delta}_T$
- Requires only $K\mathbf{v}$ multiplications! $O(n^2)$ time
- No need to store K! $O(n)$ space

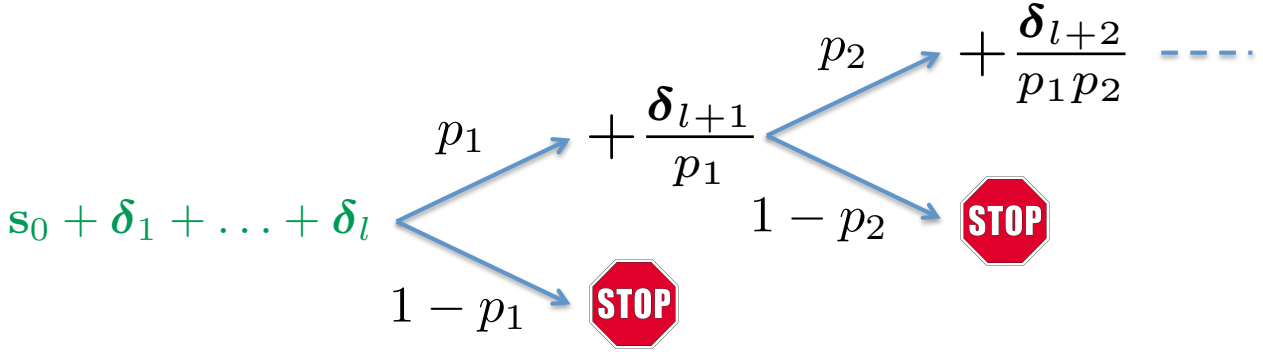
Filippone and Engler, *ICML*, 2015 - Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Solving linear systems - some nice twists

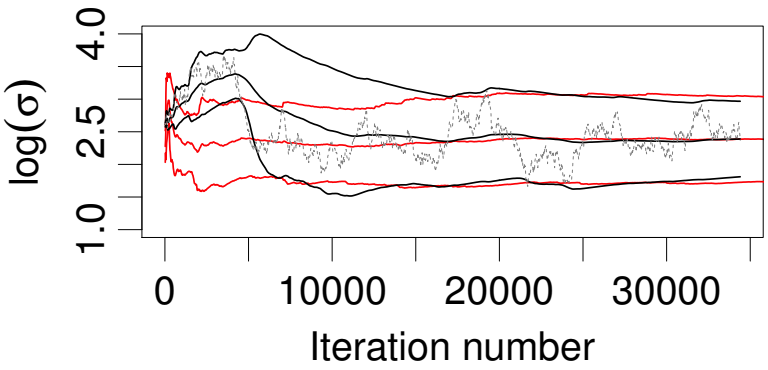
- Preconditioning

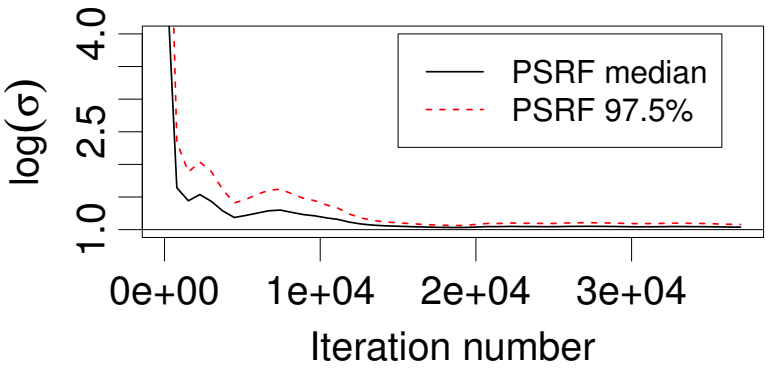


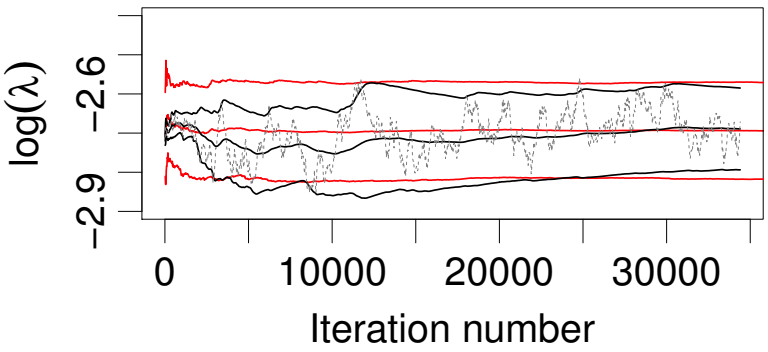
- Unbiased LInear Systems SolvEr (ULISSE)

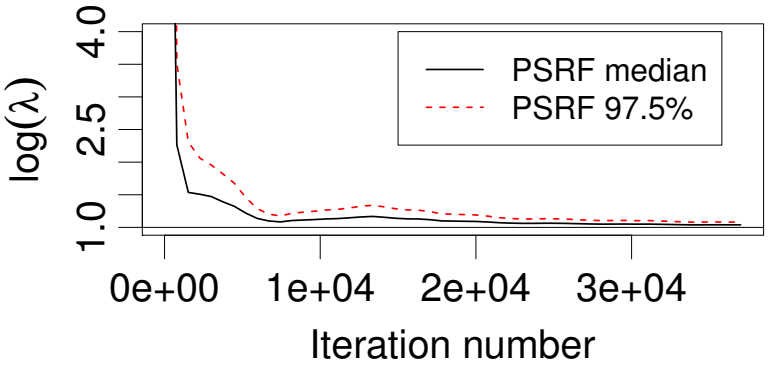


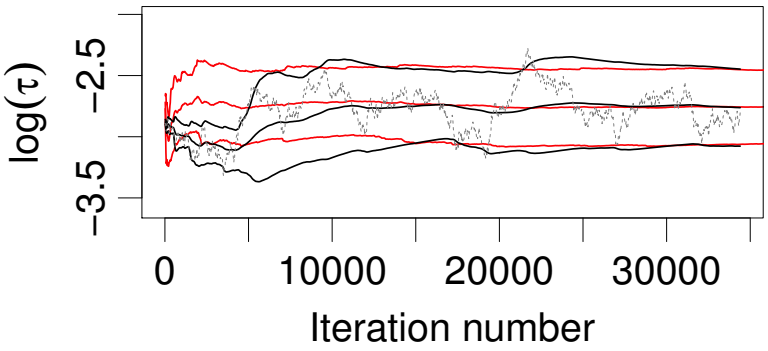
Comparison with MCMC - Concrete dataset - $n \approx 1K$

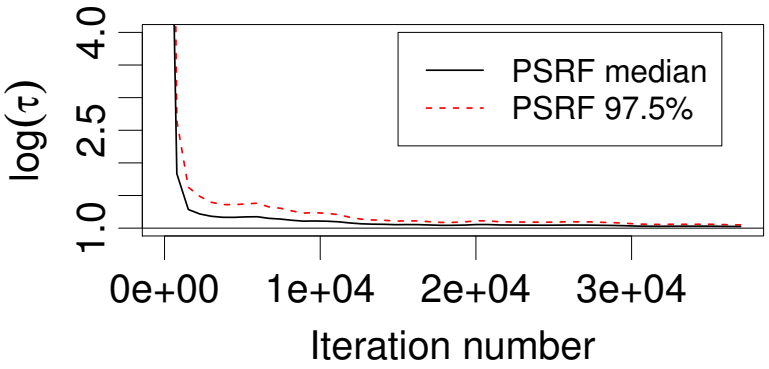




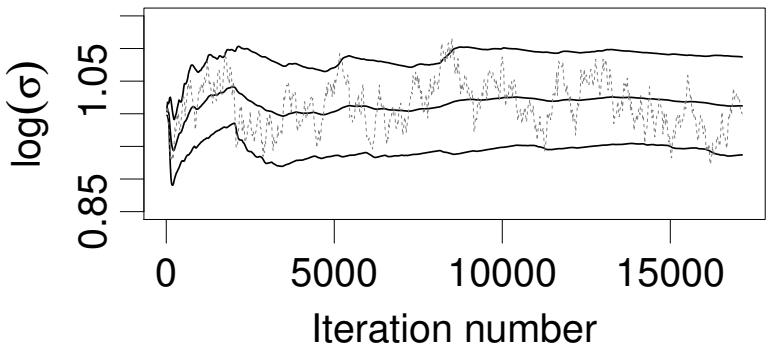


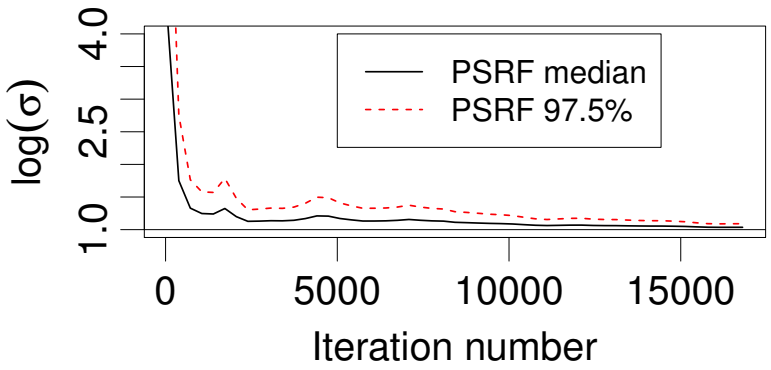


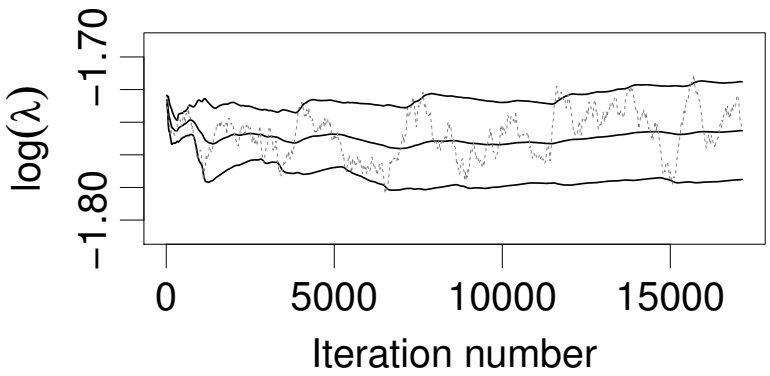


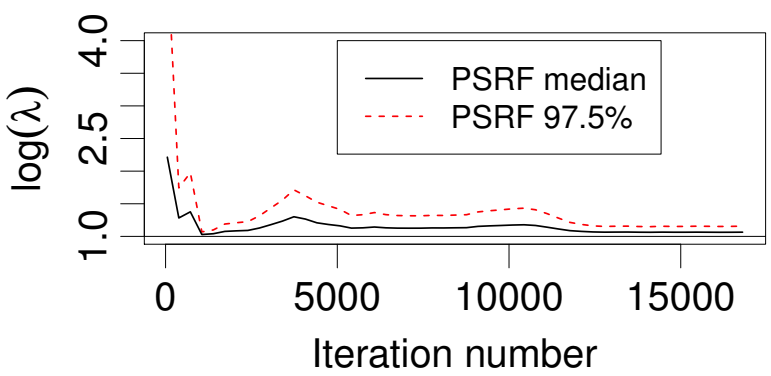


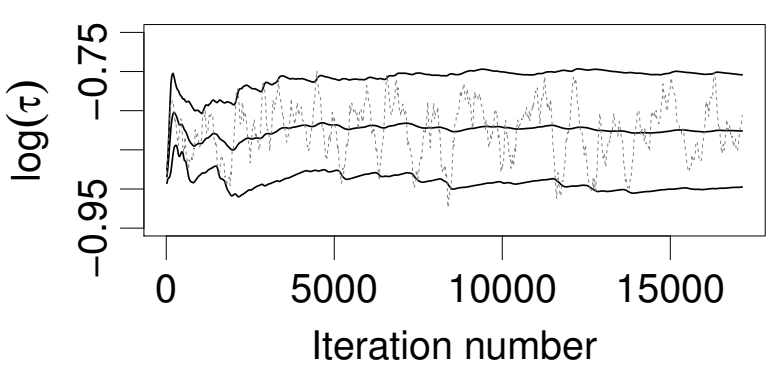
Larger n - Census dataset - $n \approx 23K$

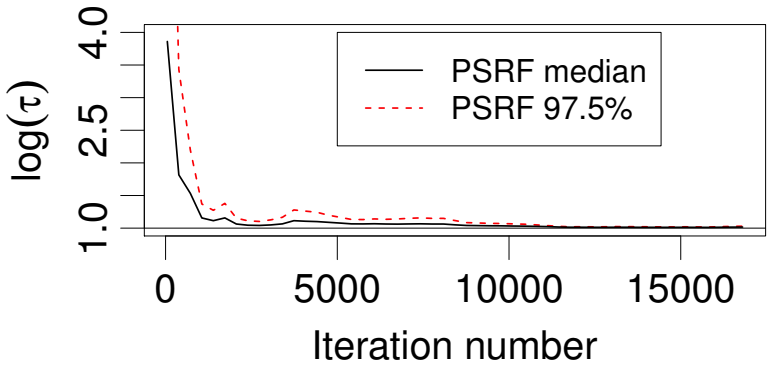






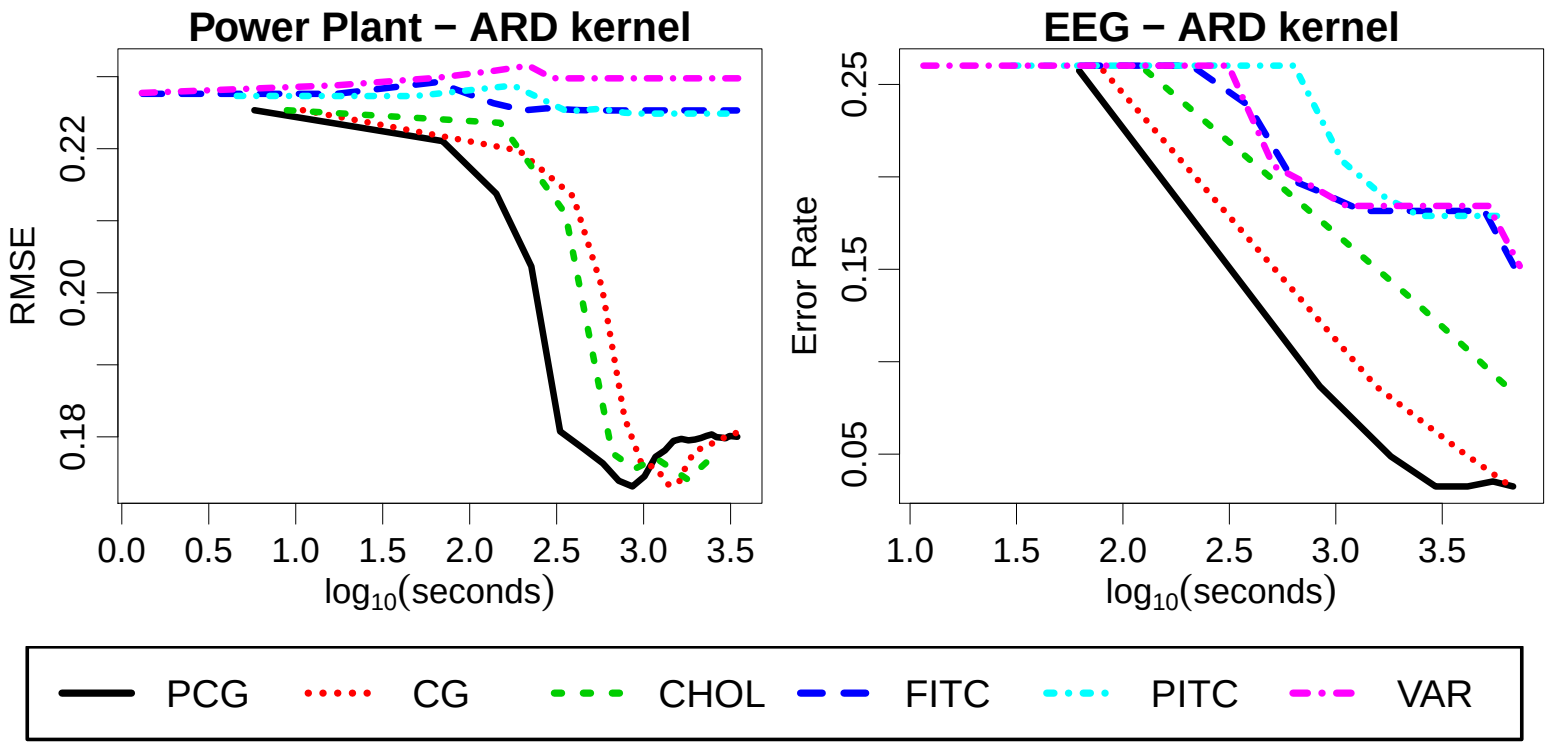






Preconditioning Kernel Matrices

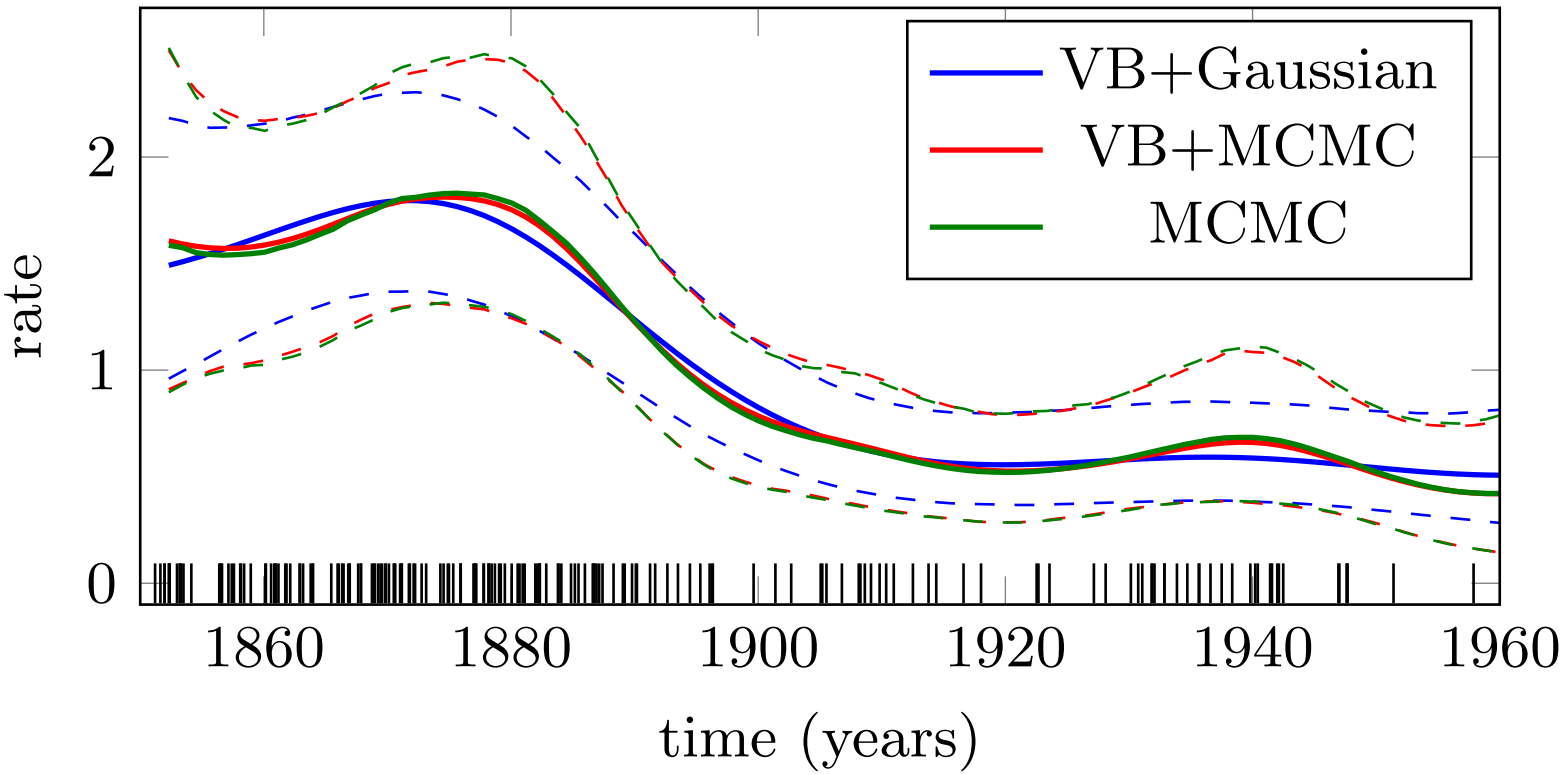
- Stochastic Gradient Optimization



Cutajar, Osborne, Cunningham, Filippone, *ICML*, 2016

Marrying Sparse Variational and MCMC learning for GPs

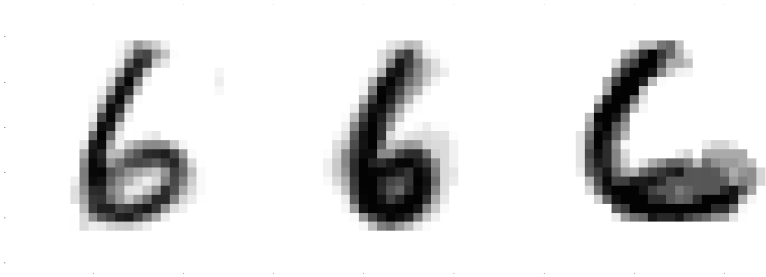
- Coal mining disaster data
- Analysis with 30 inducing points



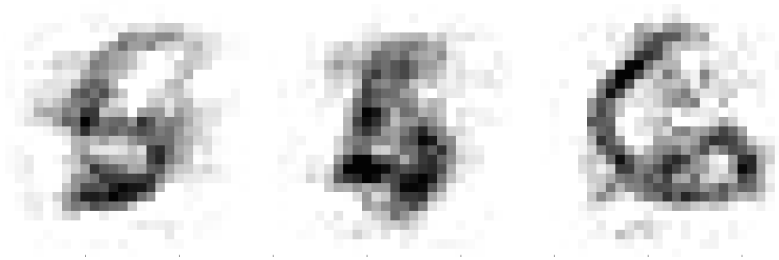
Hensman, Matthews, Filippone, Ghahramani, *NIPS*, 2015

Marrying Sparse Variational and MCMC learning for GPs

- MNIST data
- Three inducing points before optimizing their position



- . . . and after



- Overall accuracy 98.04% with 500 inducing points

Hensman, Matthews, Filippone, Ghahramani, *NIPS*, 2015

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intensive

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intensive
- Introducing approximations can severely affect quantification of uncertainty and reduce accuracy

Conclusions and ongoing work

- GPs form the basis of flexible and interpretable nonparametric statistical models
- GP learning is generally computationally intensive
- Introducing approximations can severely affect quantification of uncertainty and reduce accuracy
- Approximate **unbiased** computations offer practical and scalable ways to carry out **exact** GP learning

Acknowledgments & References


Andre Marquand
Donders


Guido Sanguinetti
Edinburgh


James Hensman
Lancaster


Mark Girolami
Warwick


Alessandro Vinciarelli
Glasgow


Dirk Husmeier
Glasgow

[1] K. Cutajar, M. A. Osborne, J. P. Cunningham, and M. Filippone. **Preconditioning kernel matrices**, *ICML 2016.*

[2] J. Hensman, A. G. de G. Matthews, M. Filippone, and Z. Ghahramani. **MCMC for variationally sparse Gaussian processes**, *NIPS 2015.*

[3] M. Filippone and R. Engler. **Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE)**, *ICML 2015.*

[4] M. Filippone and M. Girolami. **Pseudo-Marginal Bayesian inference for Gaussian processes**, *IEEE T-PAMI, 2014.*

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

[6] M. Filippone et al. **Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities**. *Annals of Applied Statistics, 2012.*

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

[8] M. Filippone et al. **A comparative evaluation of stochastic-based inference methods for Gaussian process models**. *Machine Learning, 2013.*