University of Glasgow

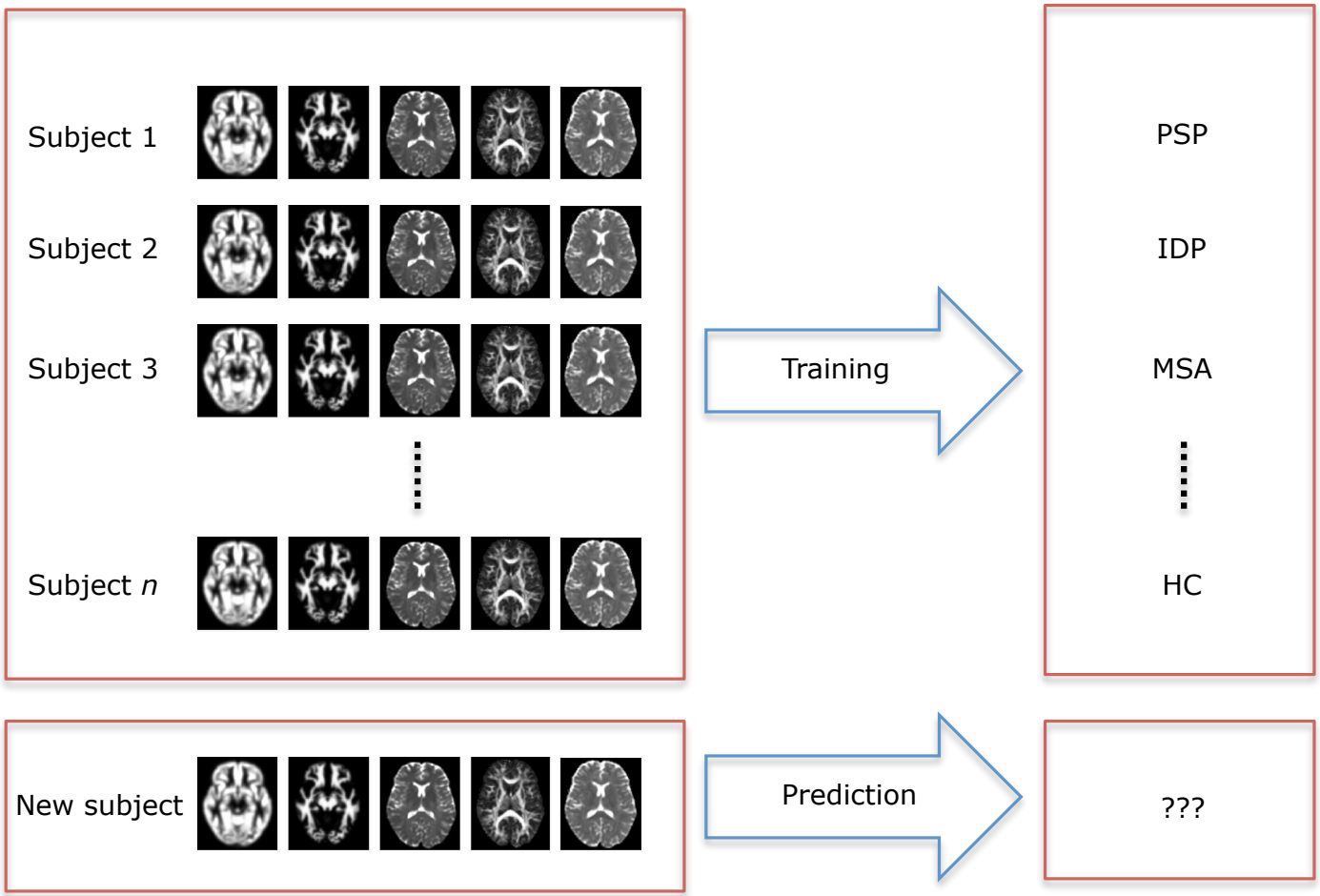
Unbiased computations for MCMC-based inference of Gaussian process covariance parameters

Maurizio Filippone

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

May 14th, 2015

Motivating Application



HC - Healthy control

MSA - Multiple system atrophy

PSP - Progressive Supranuclear Palsy

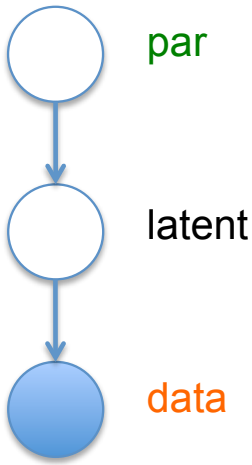
IDP - Idiopathic Parkinson's disease

Filippone et al., *AoAS*, 2012

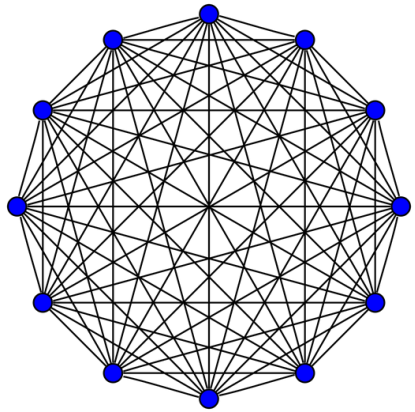
Gaussian Process Models

- Class of hierarchical models

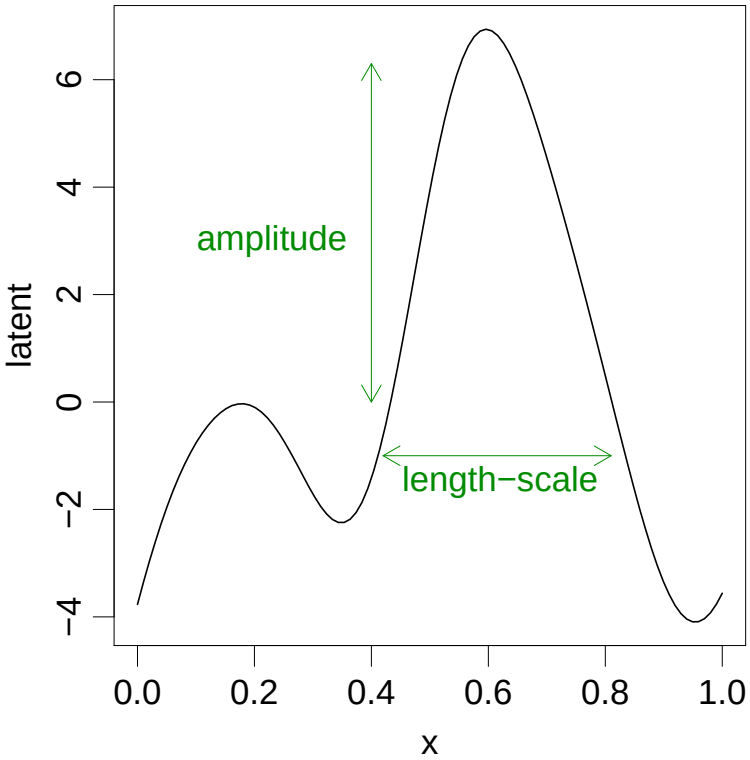
$$p(\text{data}|\text{latent}) \qquad p(\text{latent}|\text{par}) \qquad p(\text{par})$$

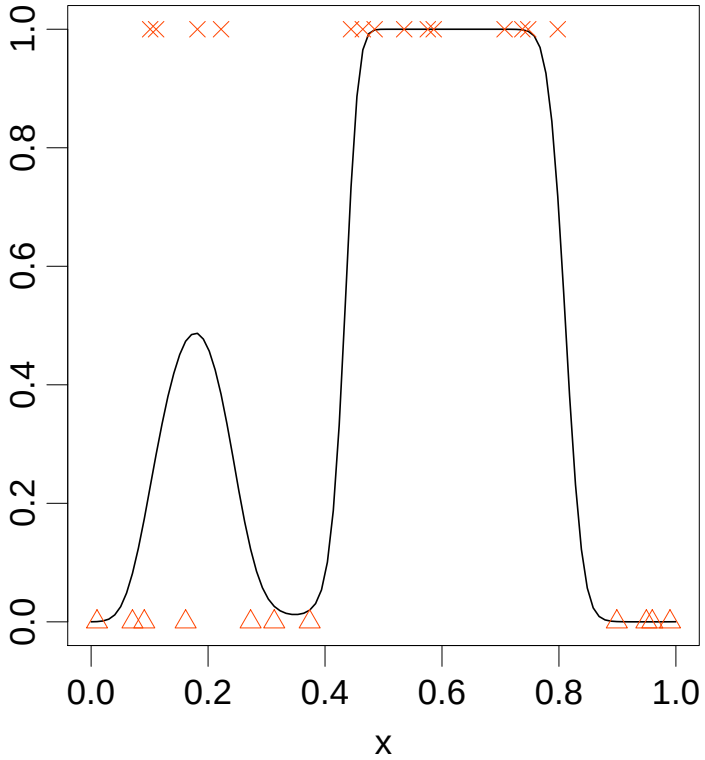


$$p(\text{latent}|\text{par}) = \text{Gaussian Process}$$



Gaussian Process Models - Classification example





Multiclass classification with multiple sources

- Multiclass classification based on GPs

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

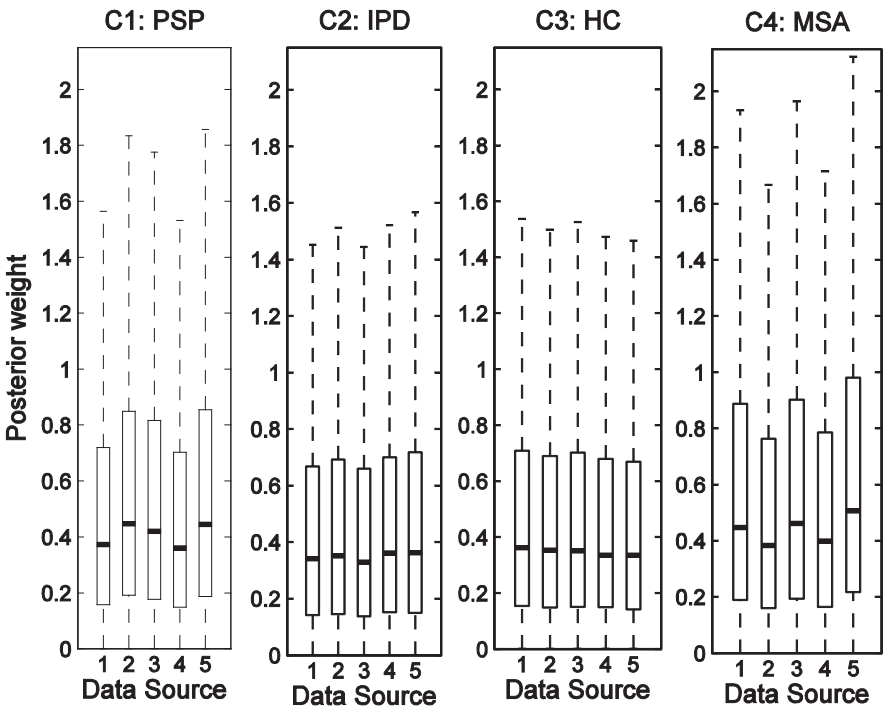
- unknown function modeled using GPs
- Covariance based on source-dependent covariances S_k

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

Filippone et al., *AoAS*, 2012

Parkinson syndromes data - multi source

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418



Filippone et al., *AoAS*, 2012

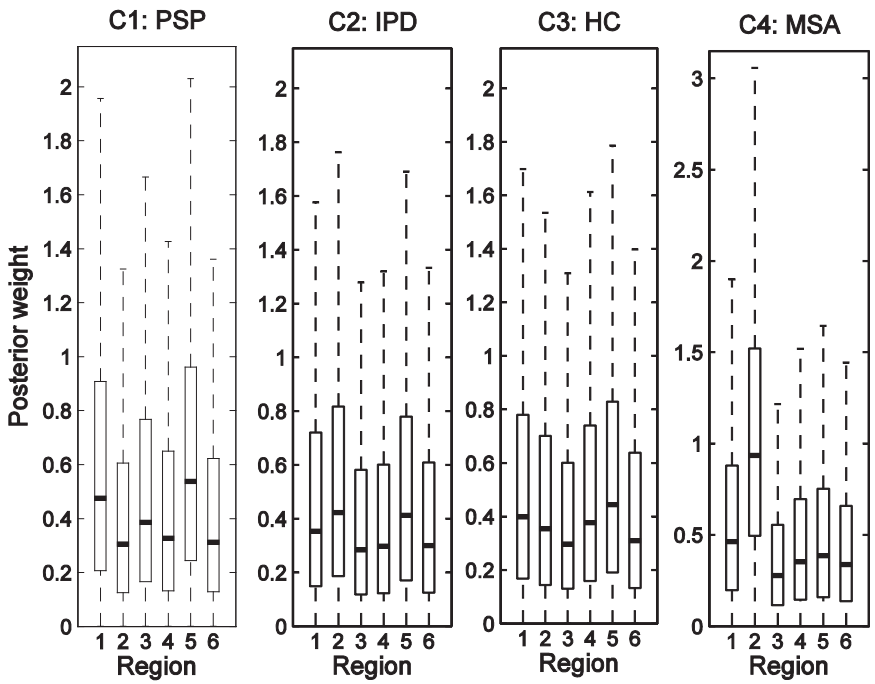
Multiclass classification with multiple regions

Analysis of brain regions

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

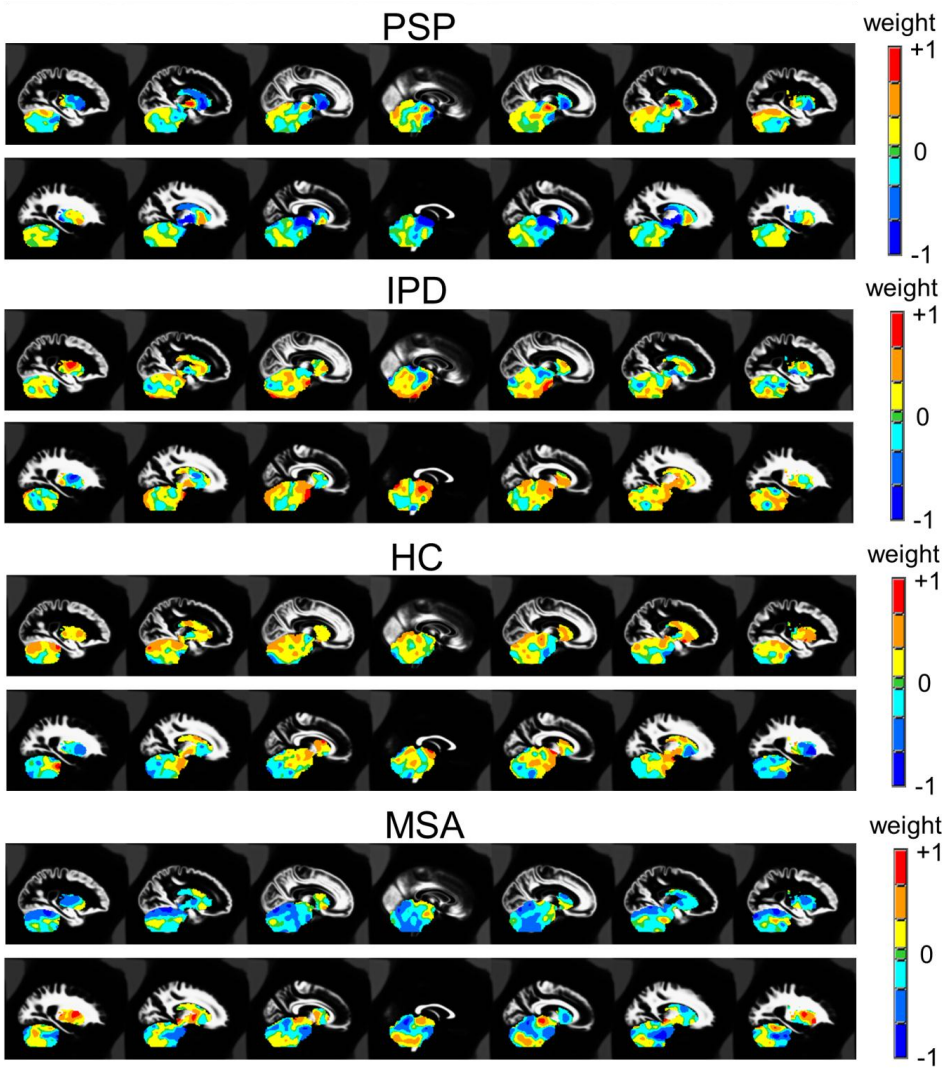
Multiclass classification with multiple regions

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229



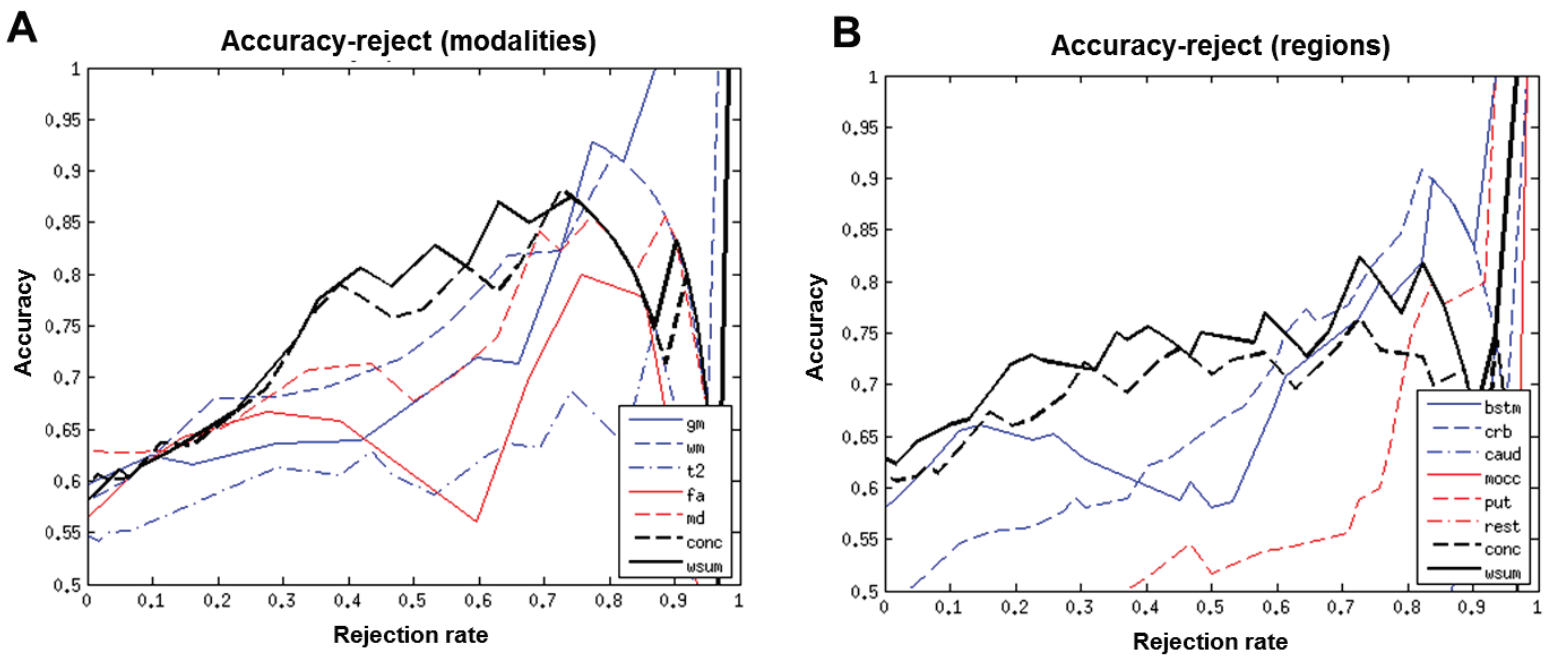
Filippone et al., *AoAS*, 2012

Brain Maps



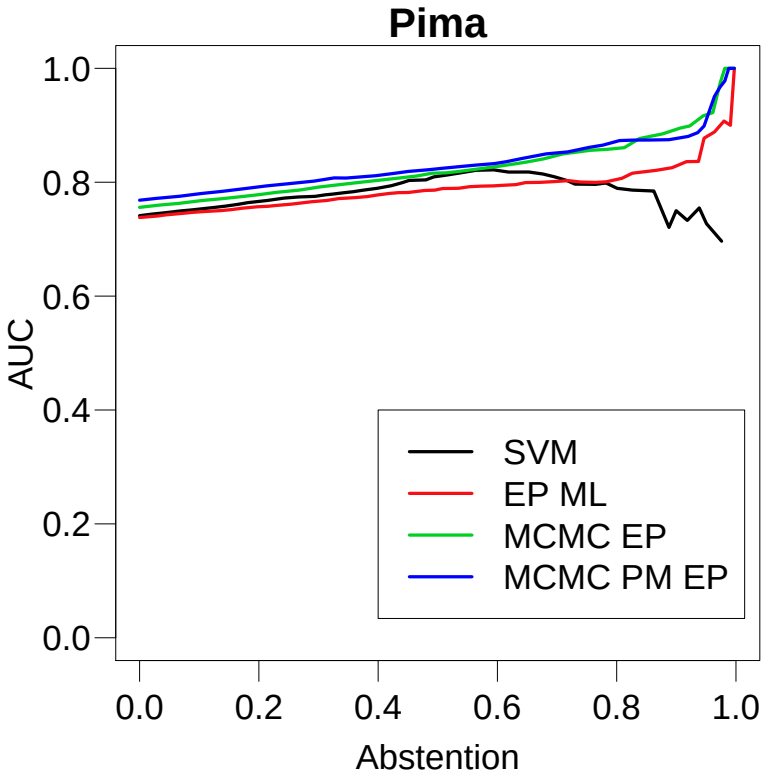
Marquand, Filippone et al., *PLOS ONE*, 2013

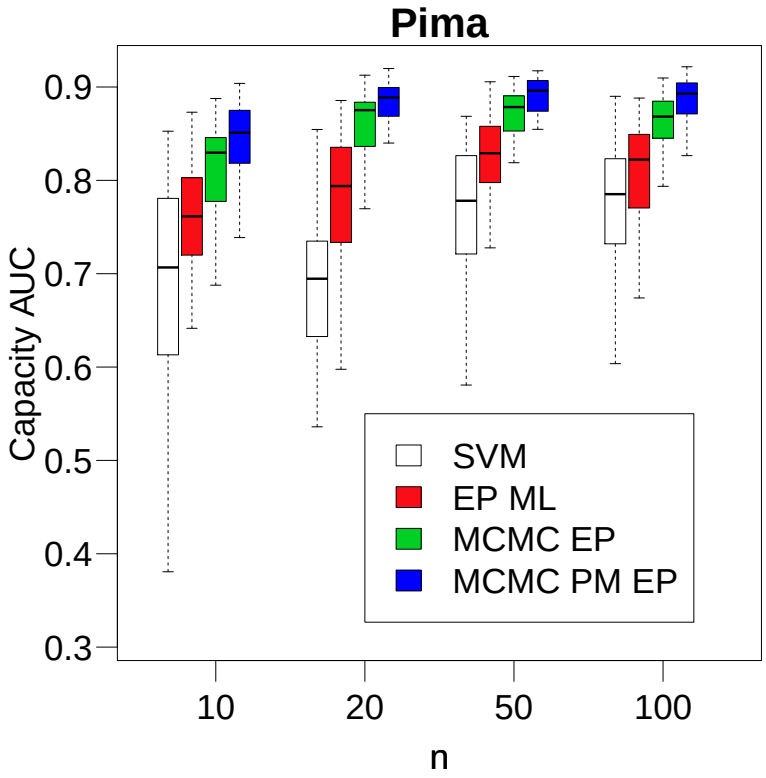
Reject Option



Filippone et al., *AoAS*, 2012

Importance of characterizing posterior over par



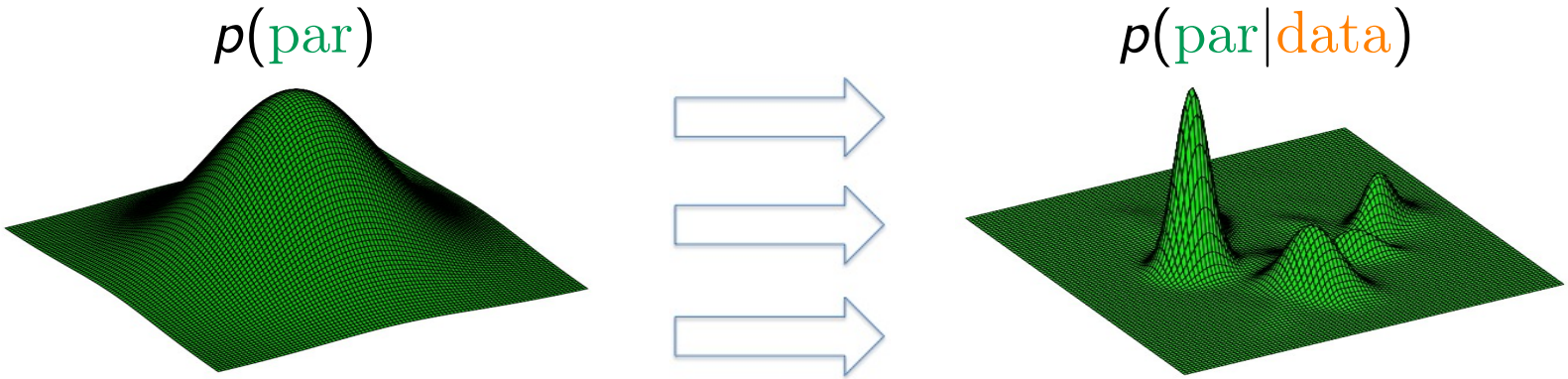


Filippone and Girolami, *IEEE-TPAMI*, 2014

Bayesian Inference

- Inference using Bayes theorem:

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



Markov chain Monte Carlo

- Bayesian inference

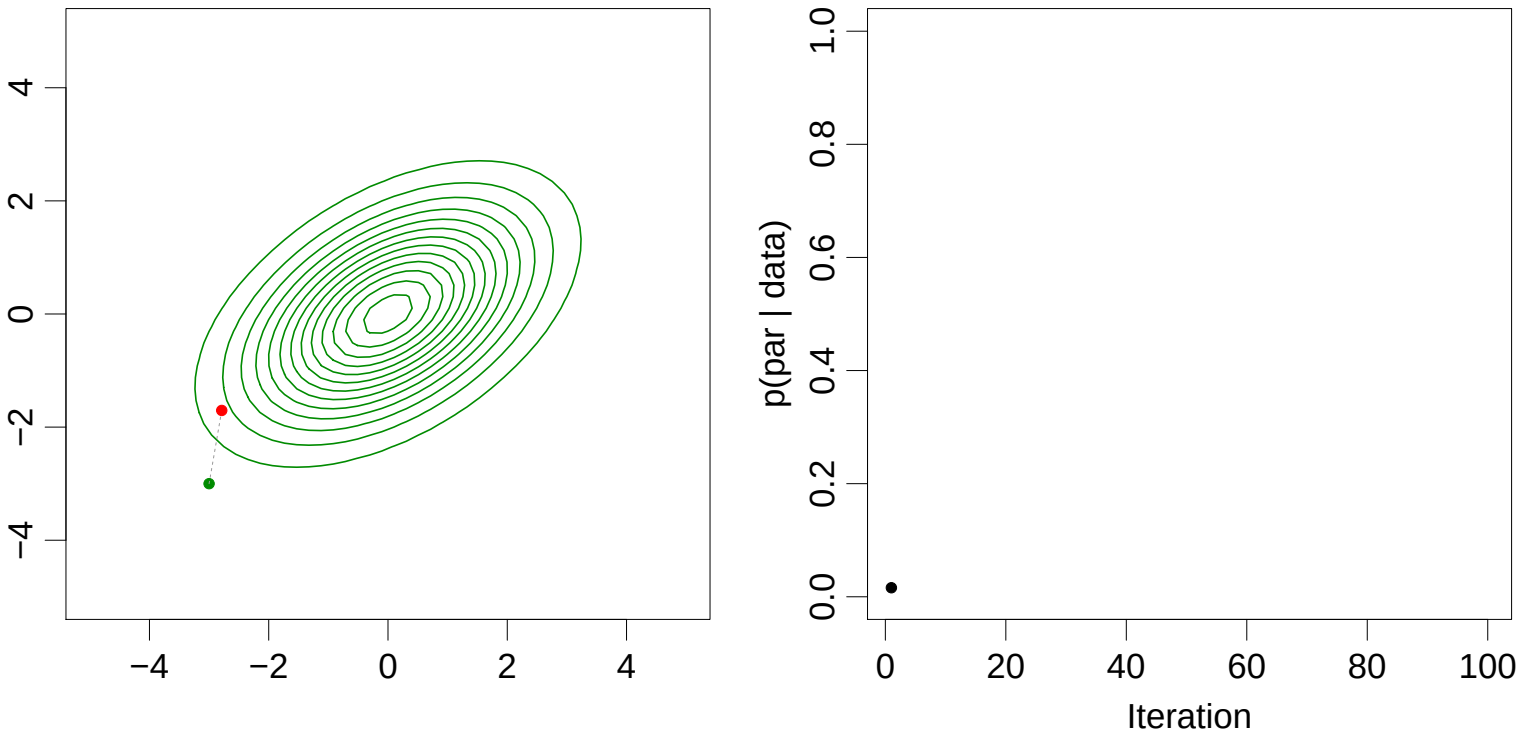
$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$

- Random walk sampler - accept a proposal with probability

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

Markov chain Monte Carlo

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



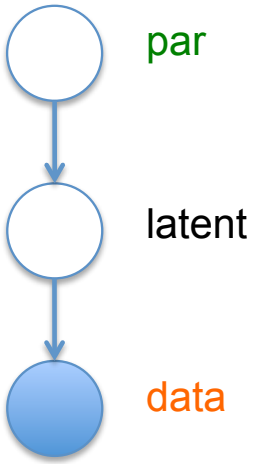
Metropolis et al., *JoCP*, 1953 - Hastings, *Biometrika*, 1970

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?

Challenges

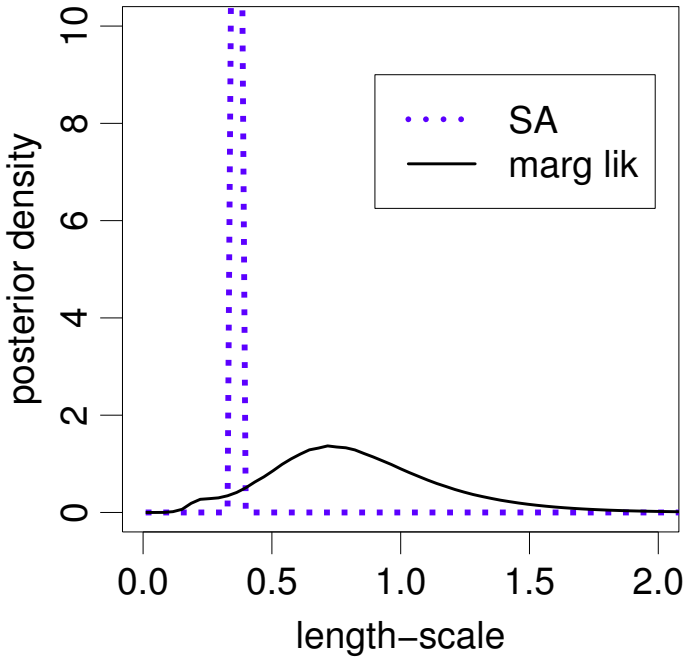
- par can be large dimensional (ARD/factor covariances)
- No exact Gibbs steps
- $p(\text{par}|\text{data})$ can be multi-modal
- How to check convergence?
- $p(\text{data}|\text{par})$ might be expensive to compute
- $p(\text{data}|\text{par})$ might not even be computable!

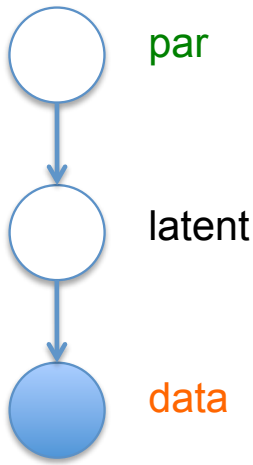


Challenges in MCMC for GPMs - Structure

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**





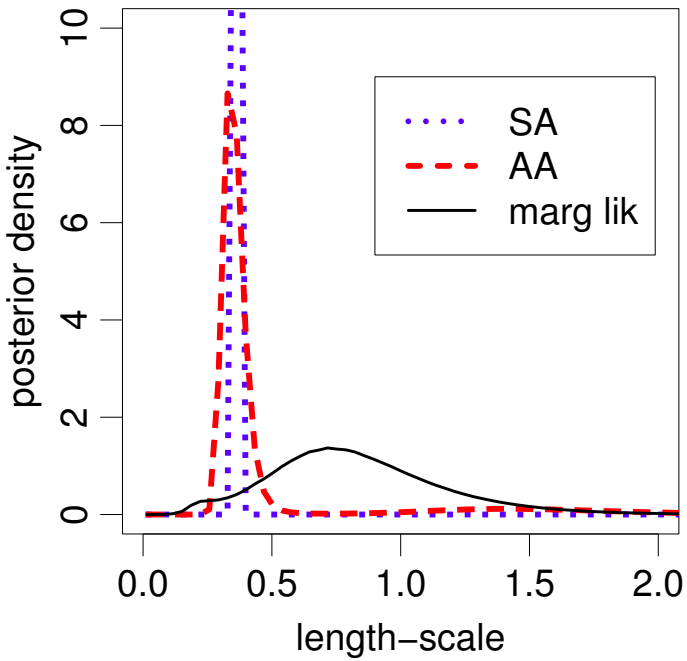
Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

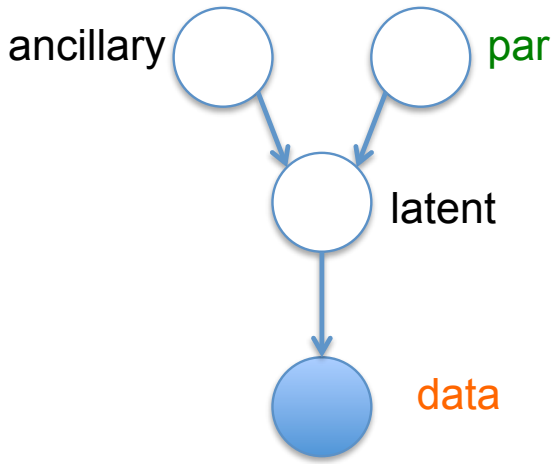
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization:

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

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



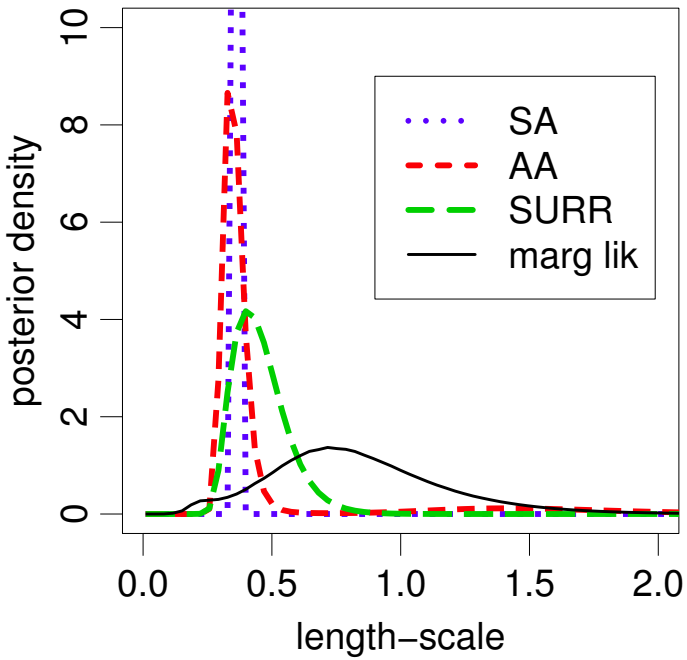


Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

Mitigating coupling effect through reparameterization

Surrogate data model (SURR):

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



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

Murray and Adams, *NIPS*, 2010 - Filippone et al., *Mach. Learn.*, 2013.

Marginal likelihood in GPMs

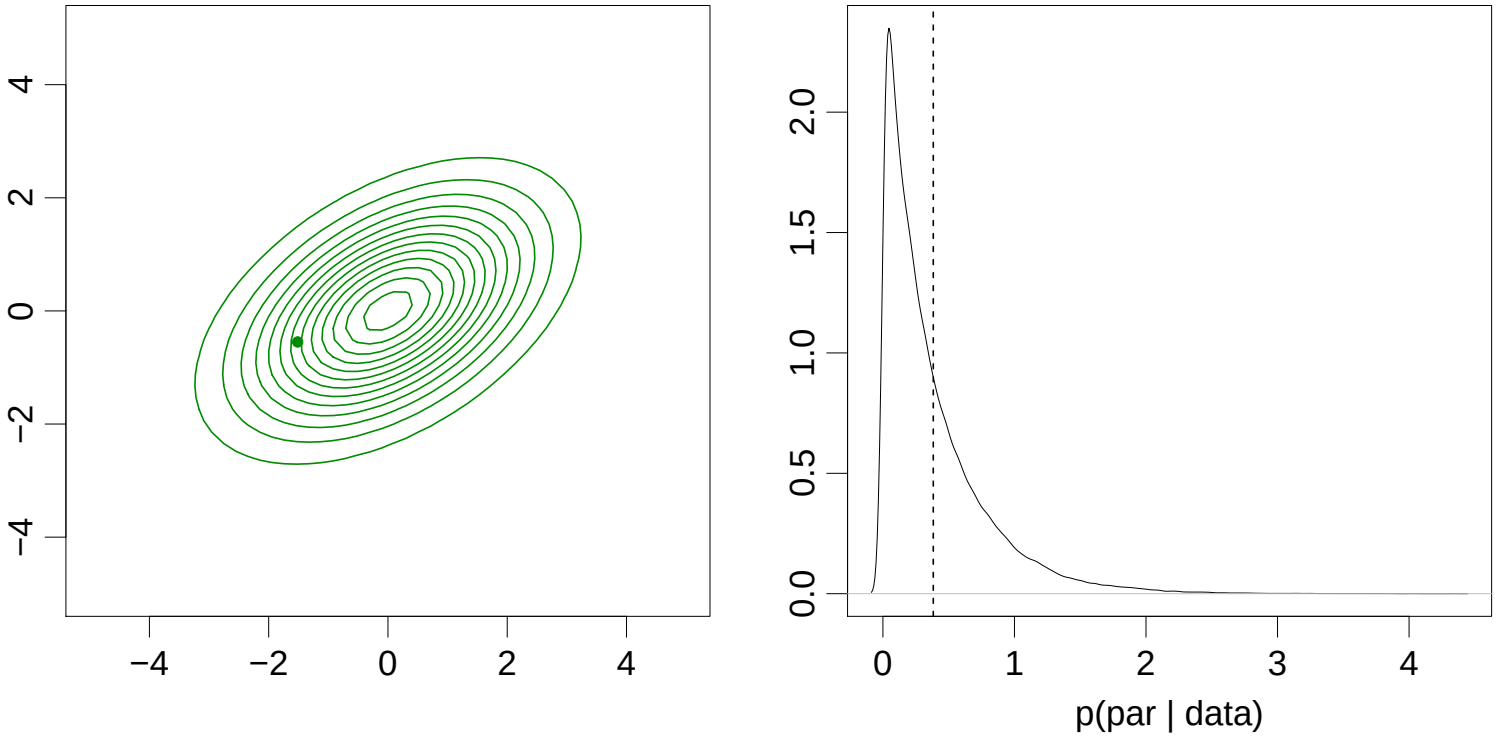
- Marginal likelihood

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

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

"Noisy" Markov chain Monte Carlo

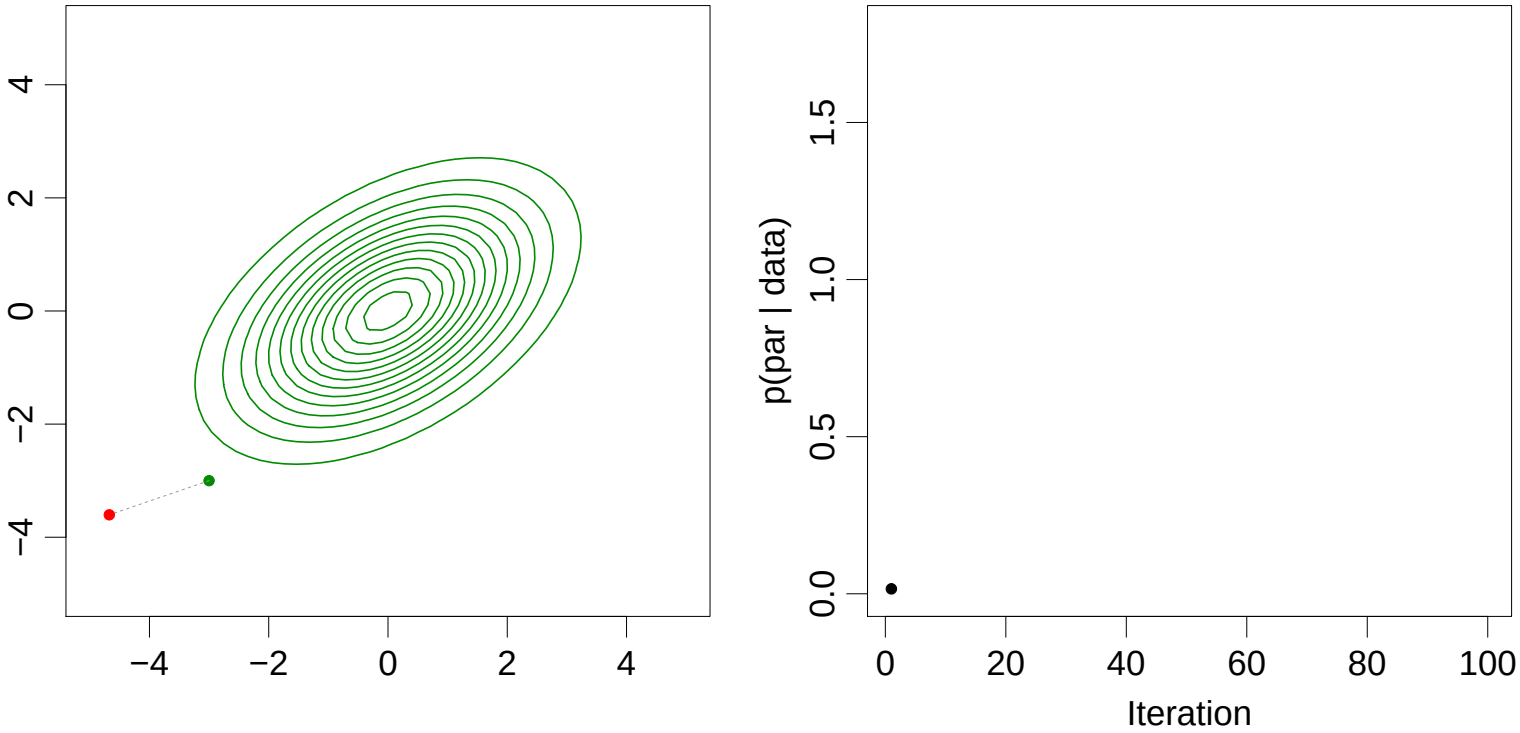
$$\mathrm{E}\left\{\tilde{p}(\text{data}|\text{par})\right\} = p(\text{data}|\text{par})$$



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

"Noisy" Markov chain Monte Carlo

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

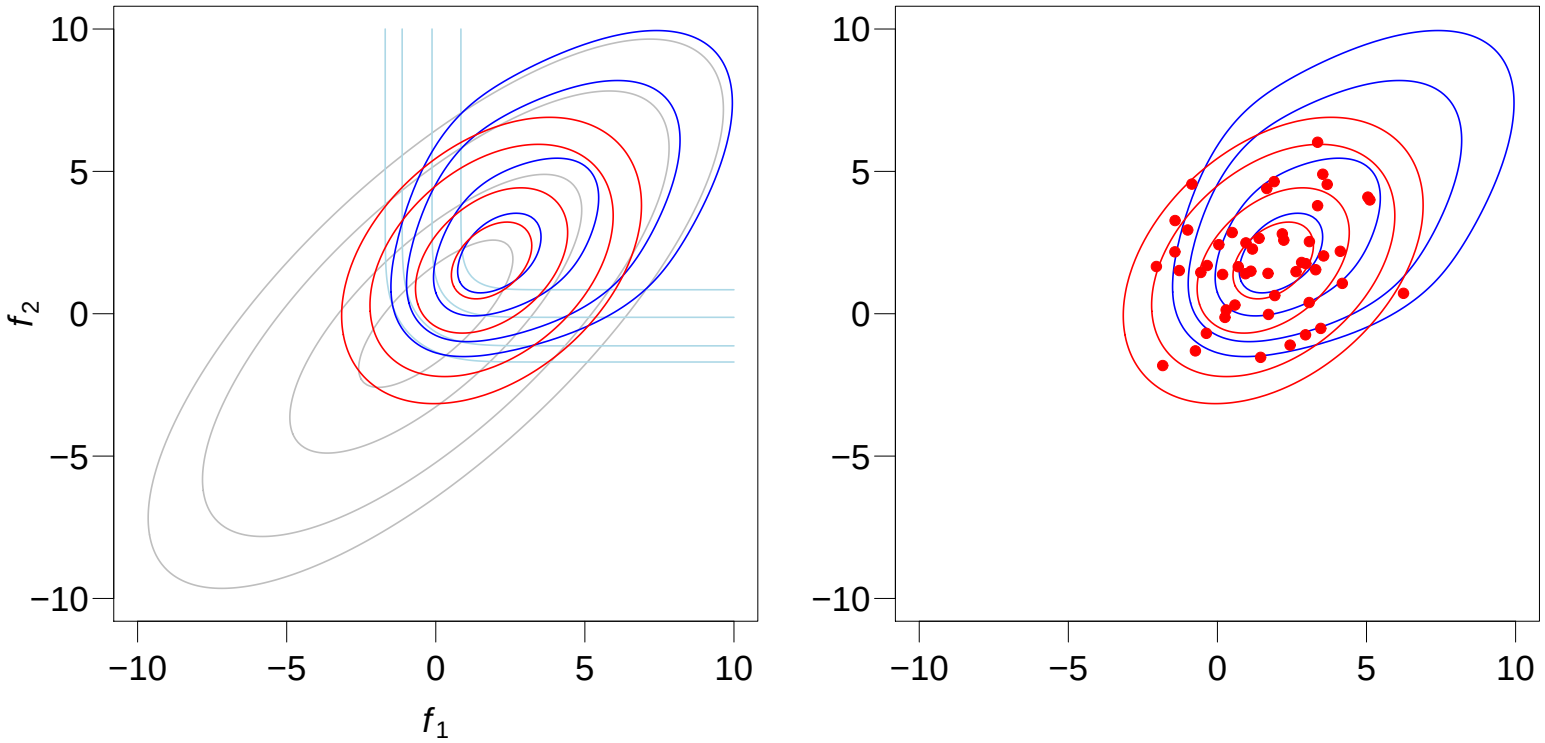


Andrieu and Roberts, *AoS*, 2009 - Filippone and Girolami, *IEEE-TPAMI*, 2014

Importance Sampling estimator

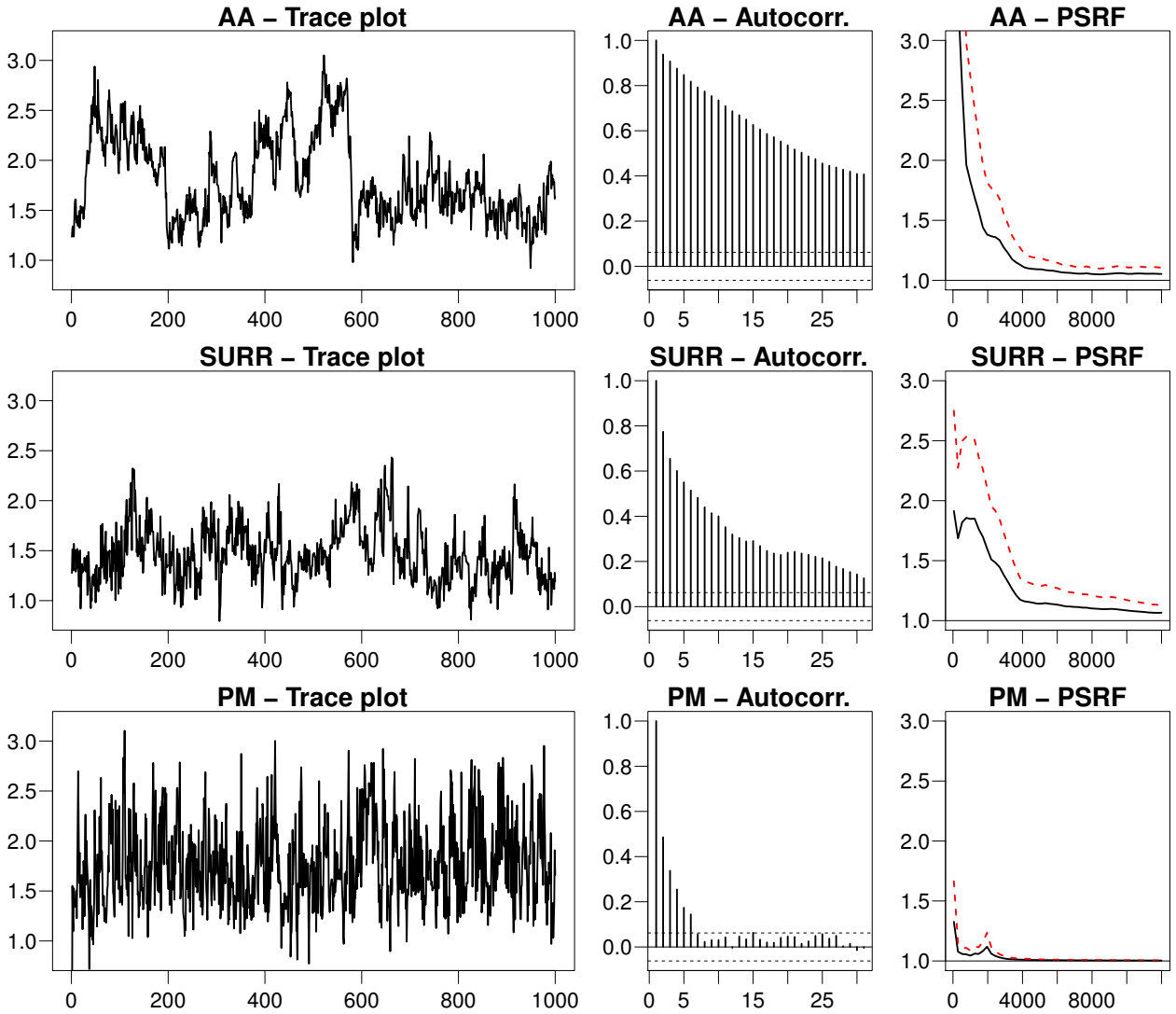
- Approximate posterior over latent variables using $q(\text{latent})$
- Then

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



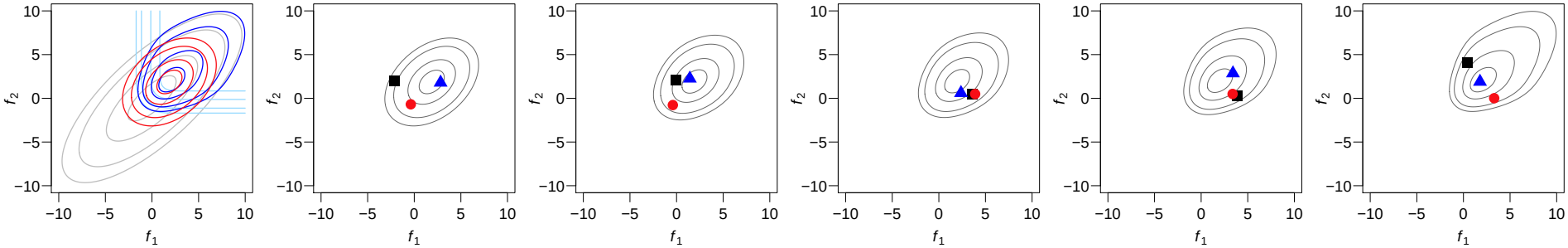
Convergence speed and efficiency

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



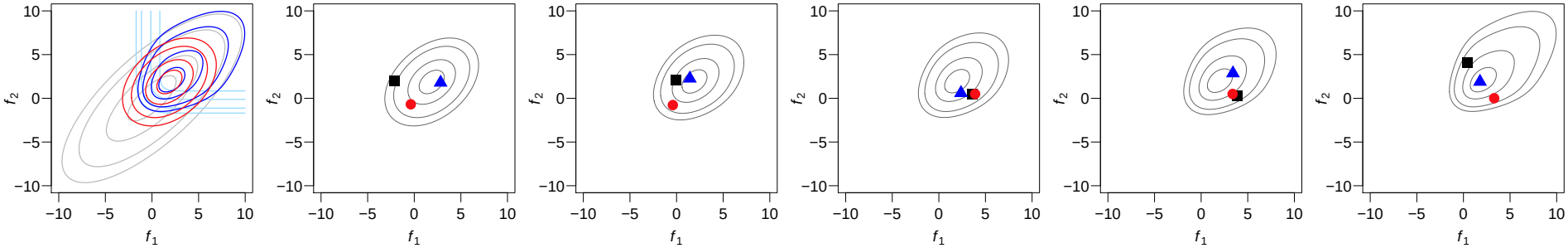
Annealed Importance Sampling estimator

- Annealing from an approximation

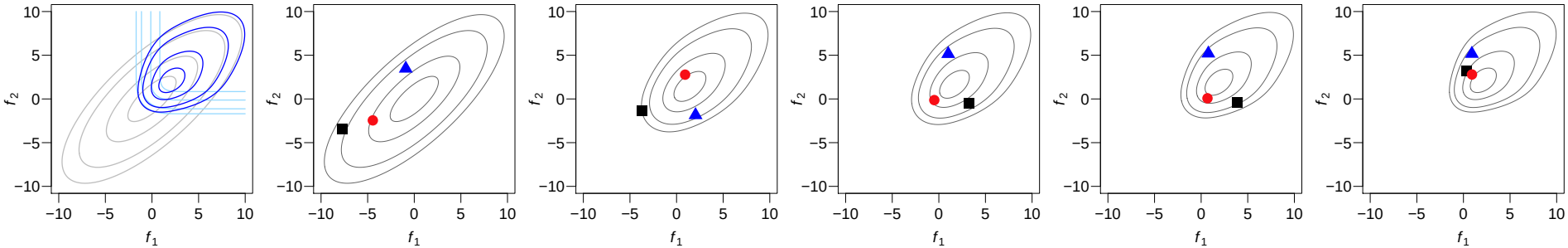


Annealed Importance Sampling estimator

- Annealing from an approximation



- Annealing from the prior

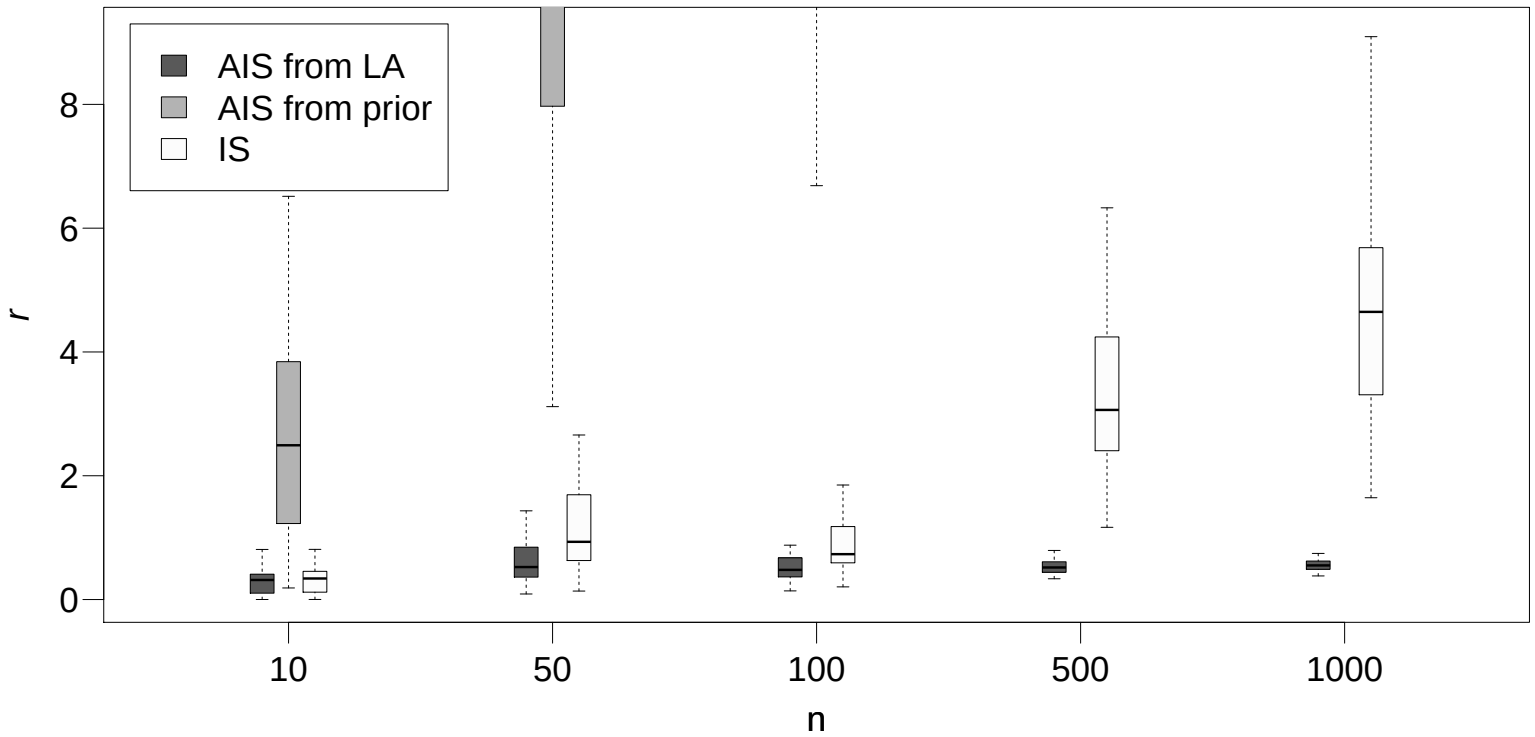


Filippone, *ICPR*, 2014

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



Filippone, *ICPR*, 2014

Marginal likelihood in GPMs

- 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

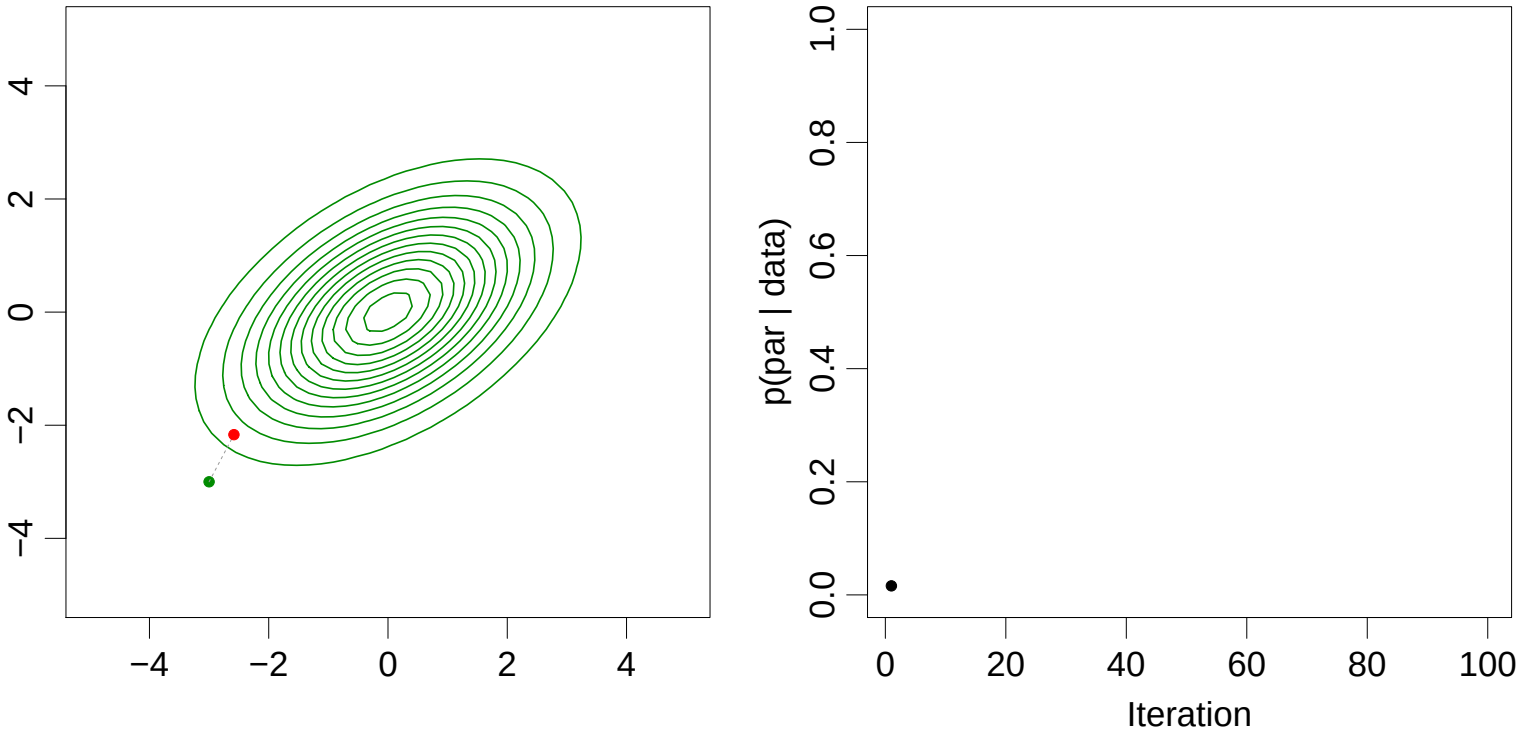
- ... even then

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

  where K is a $n \times n$ dense matrix!

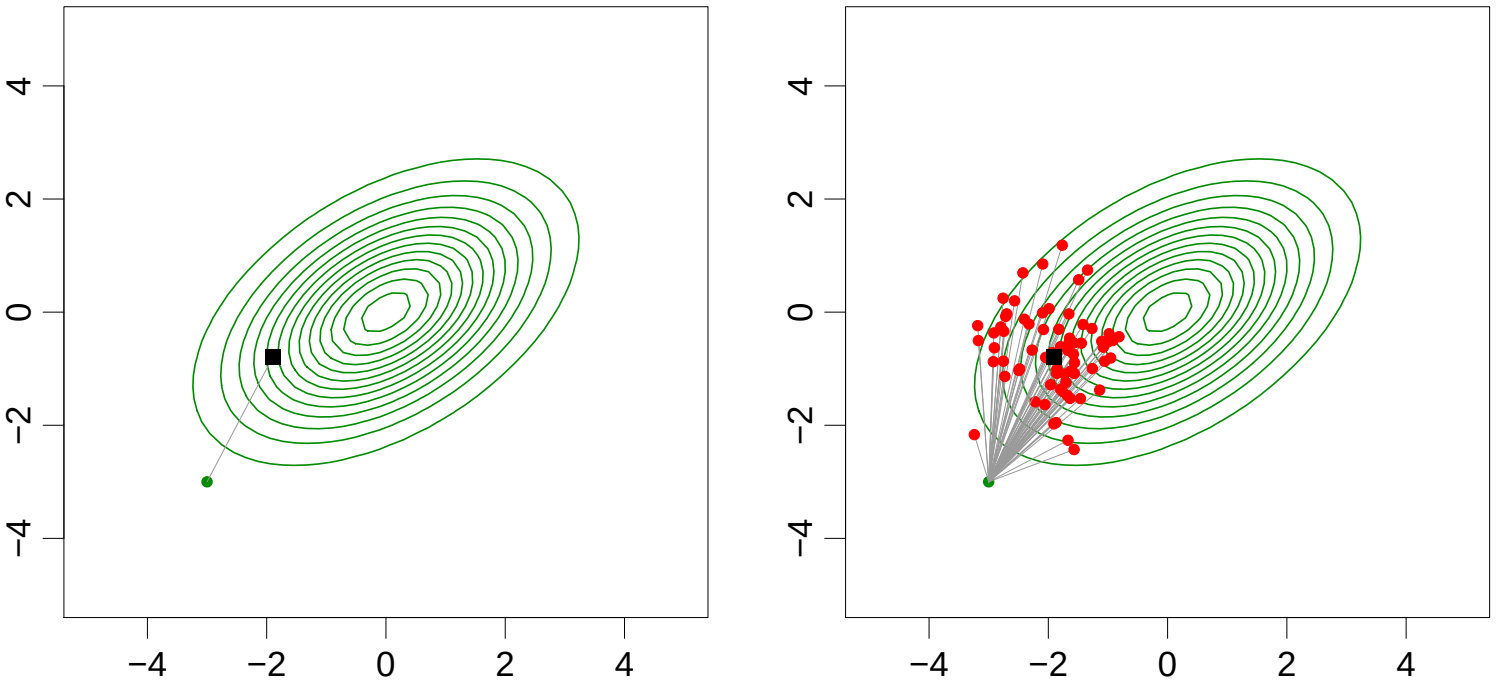
Gradient ascent

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



Stochastic Gradient ascent

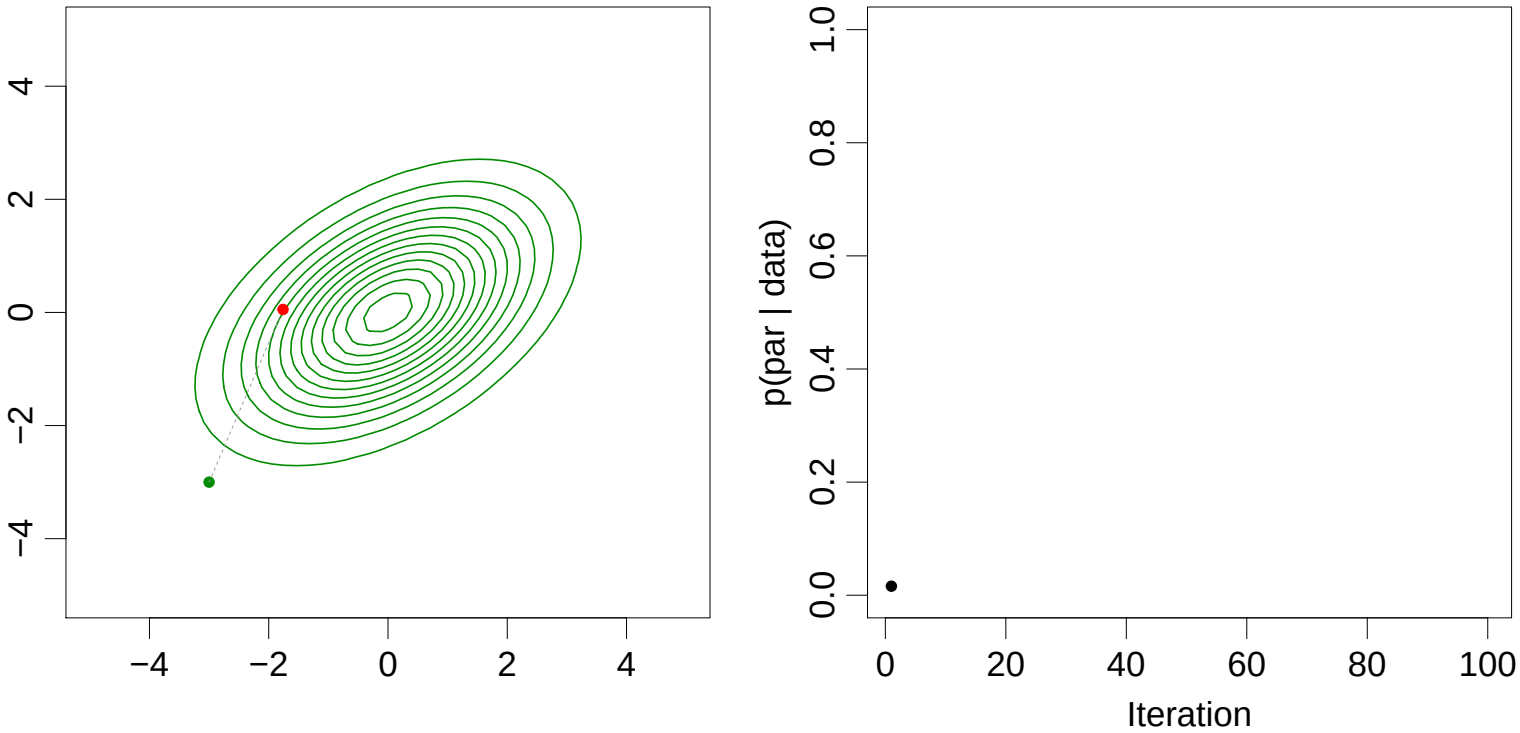
$$\mathrm{E}\left\{\widetilde{\nabla_{\mathrm{par}}}\log[p(\mathrm{data}|\mathrm{par})]\right\} = \nabla_{\mathrm{par}}\log[p(\mathrm{data}|\mathrm{par})]$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient ascent

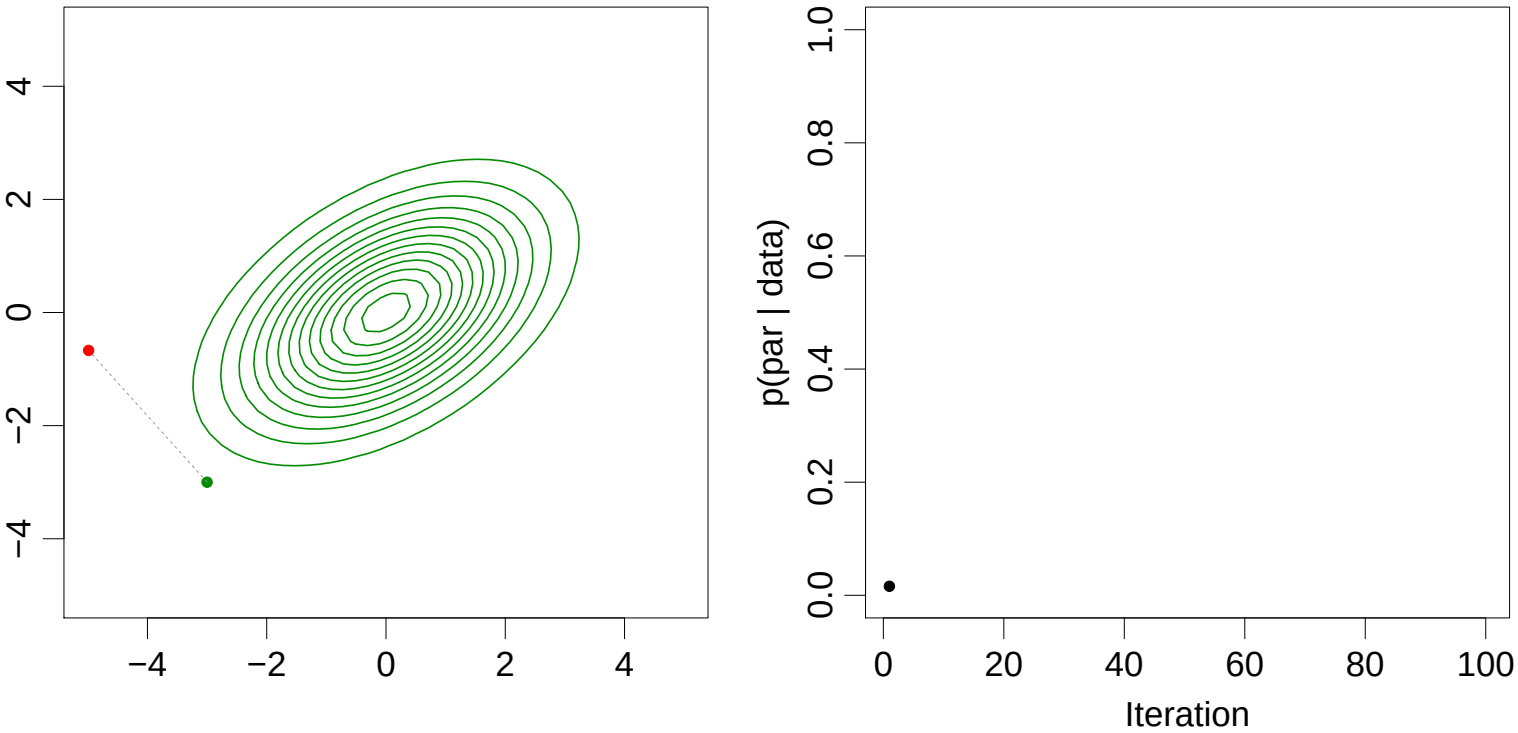
$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})p(\text{par})] \qquad \alpha_t \to 0$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient-based Markov chain Monte Carlo

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



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

Stochastic Gradients in GP regression

- Marginal likelihood

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

- Derivatives wrt par

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

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

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

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

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

- Stochastic gradient

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

- Only linear systems!

Filippone and Engler, *ICML*, 2015

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

The Unbiased LInear System SolvEr - ULISSE

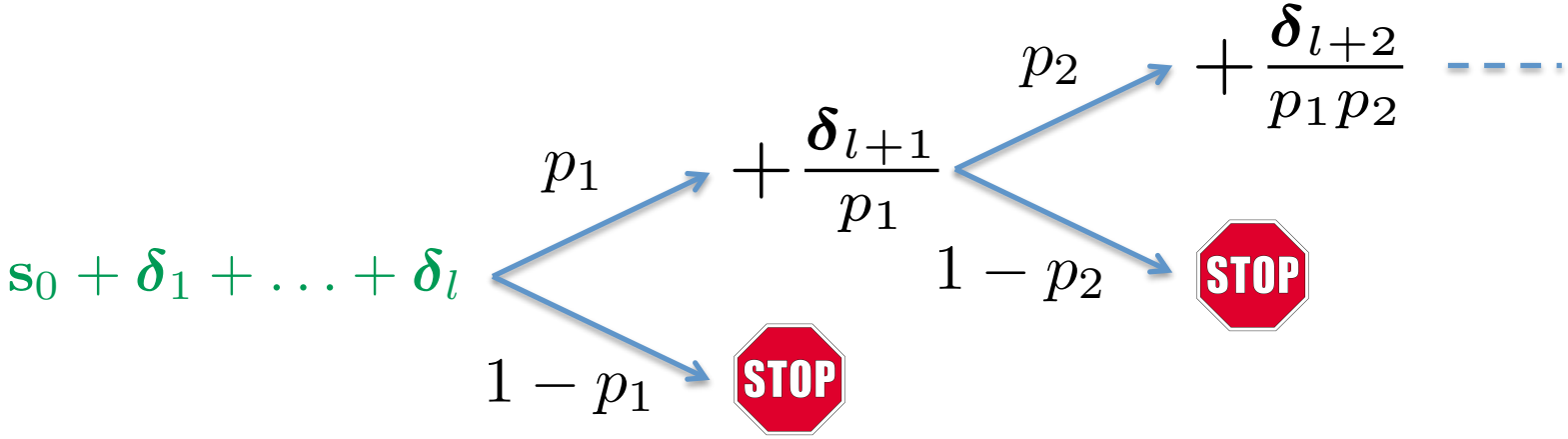
- Accelerate the solution of dense linear systems
- . . . returning an unbiased estimate of the solution

The Unbiased LInear System SolvEr - ULISSE

- Full CG solution:

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

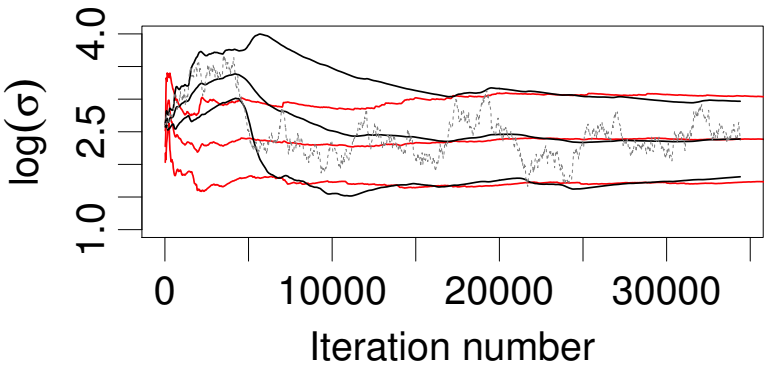
- ULISSE:

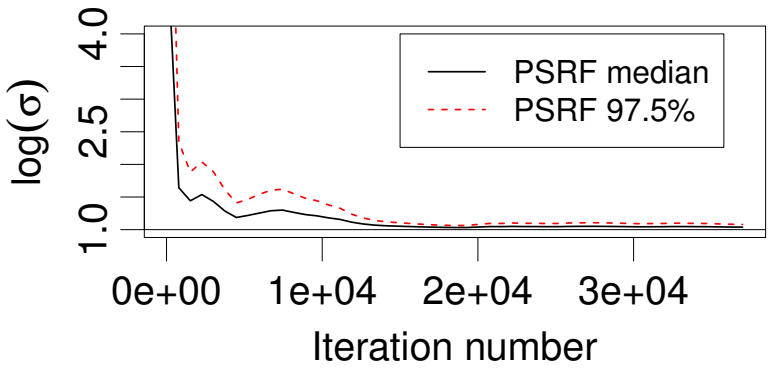


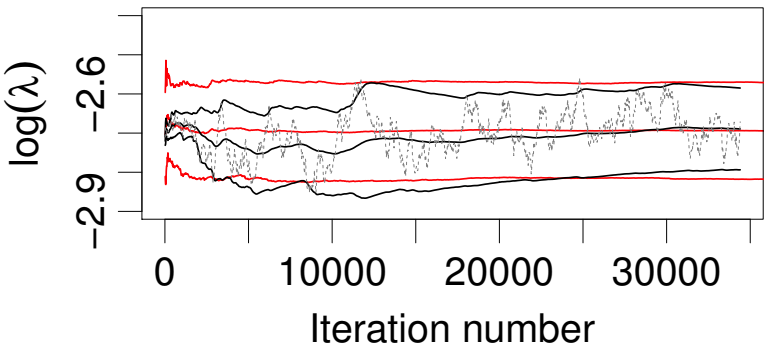
- Final solution is an unbiased estimate of $\mathbf{s}$!

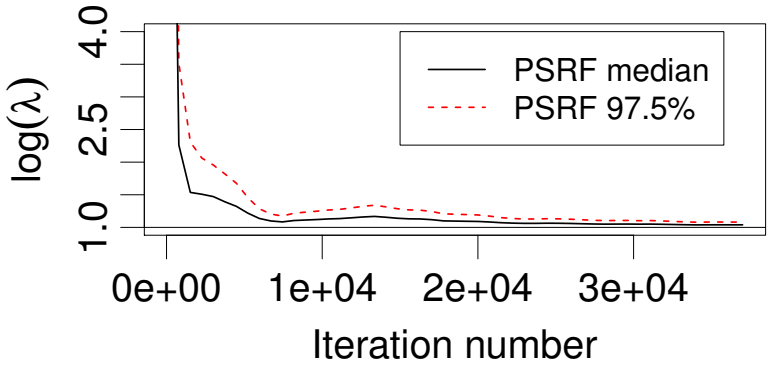
Filippone and Engler, *ICML*, 2015

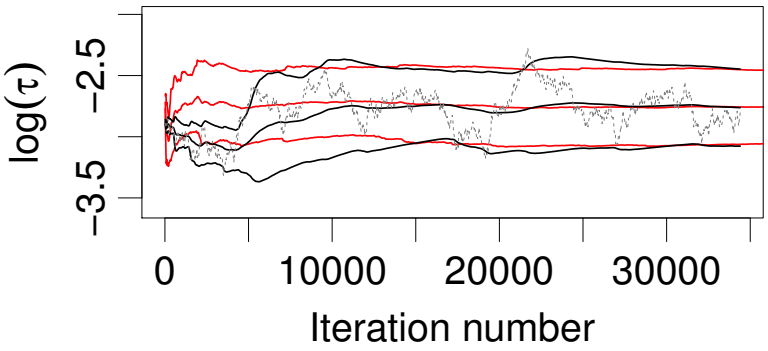
Comparison with MCMC - Concrete dataset - $n \approx 1K$

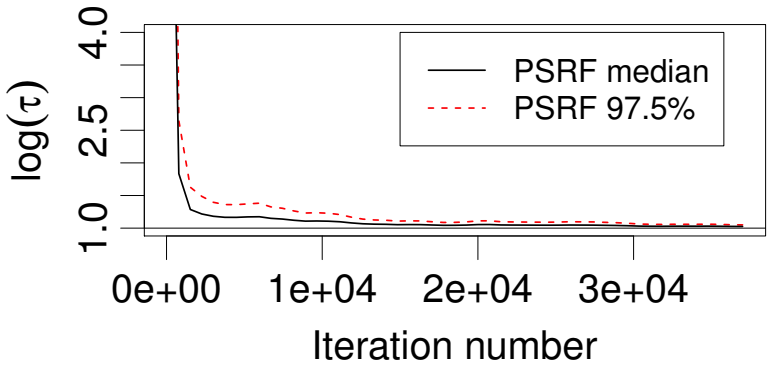




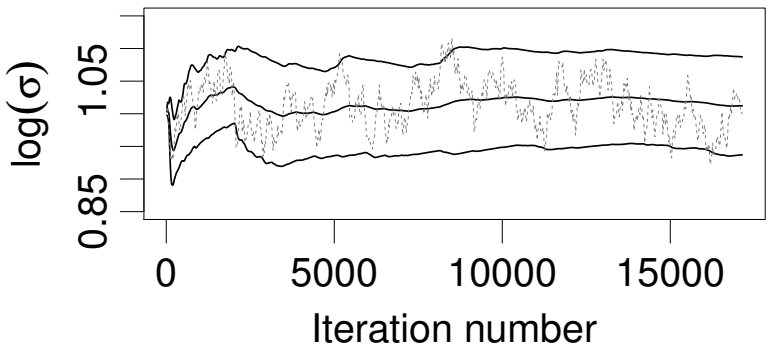


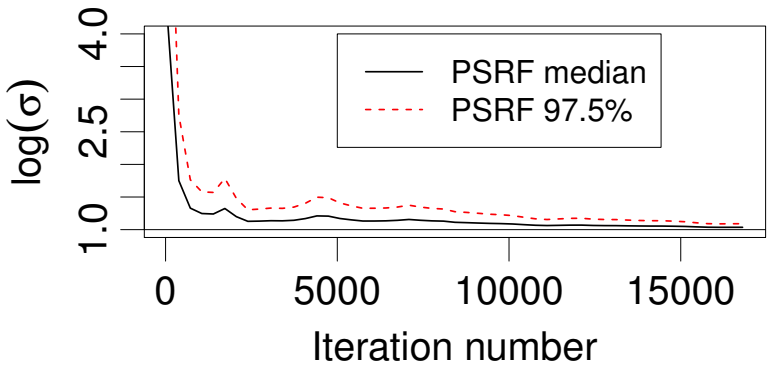


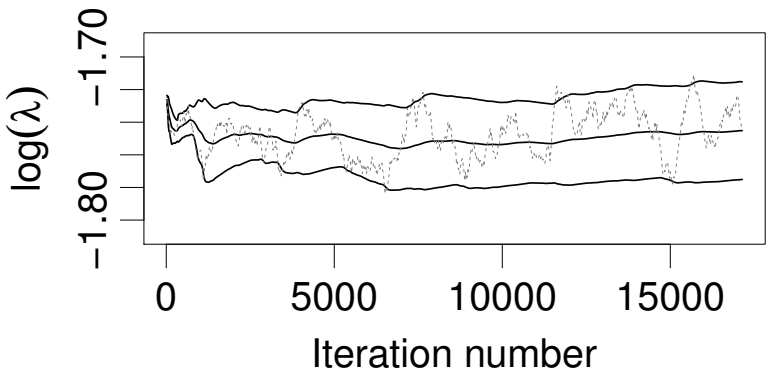


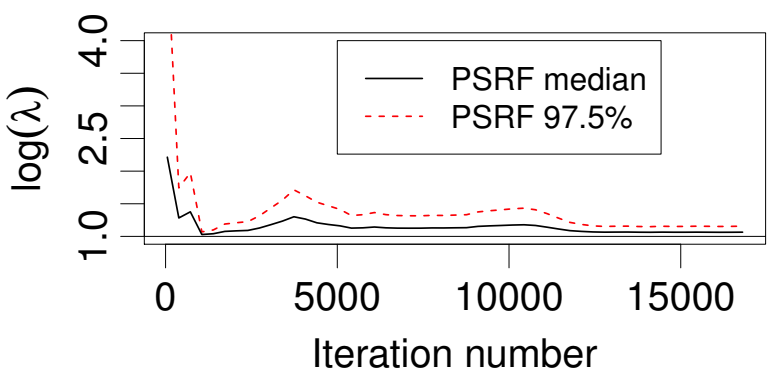


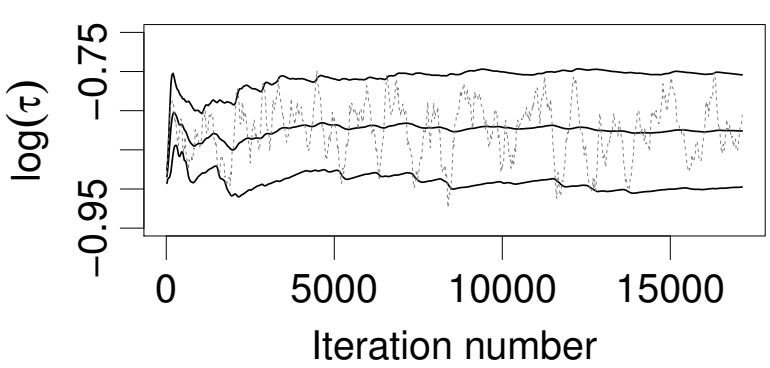
Larger n - Census dataset - $n \approx 23K$

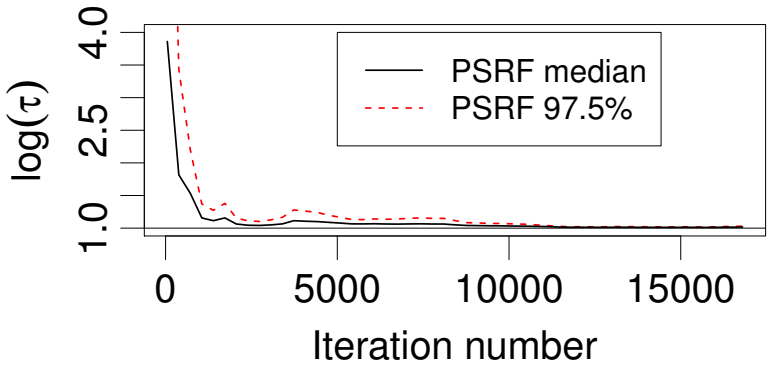












Conclusions and ongoing work

- 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 & References


Andre Marquand
Radboud


Guido Sanguinetti
Edinburgh


James Hensman
Sheffield


Mark Girolami
Warwick


Alessandro Vinciarelli
Glasgow


Dirk Husmeier
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 *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, July 6-11, 2015, 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 *Proceedings of the 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014, pages 614-619. IEEE, 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, 8(7):e69237+, 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.*