Scaling Bayesian inference for Gaussian processes using the Stochastic Gradient Langevin Dynamics algorithm

Maurizio Filippone

EURECOM, Sophia Antipolis, France
&
University of Glasgow, Glasgow, UK

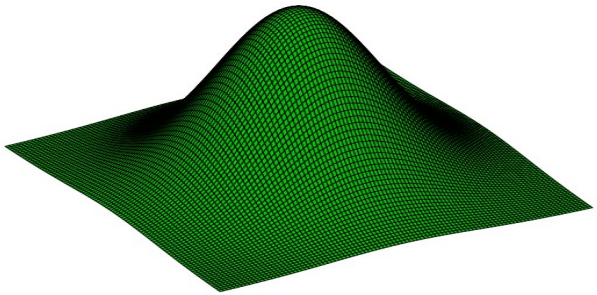
maurizio.filippone@glasgow.ac.uk

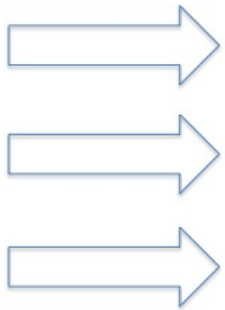
June 15th, 2015

Bayesian Inference

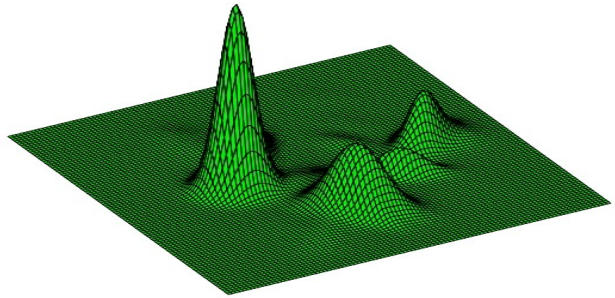
$$p(\text{par}|\text{data}) = \frac{p(\text{data}|\text{par})p(\text{par})}{\int p(\text{data}|\text{par})p(\text{par})d\text{par}}$$

$p(\text{par})$



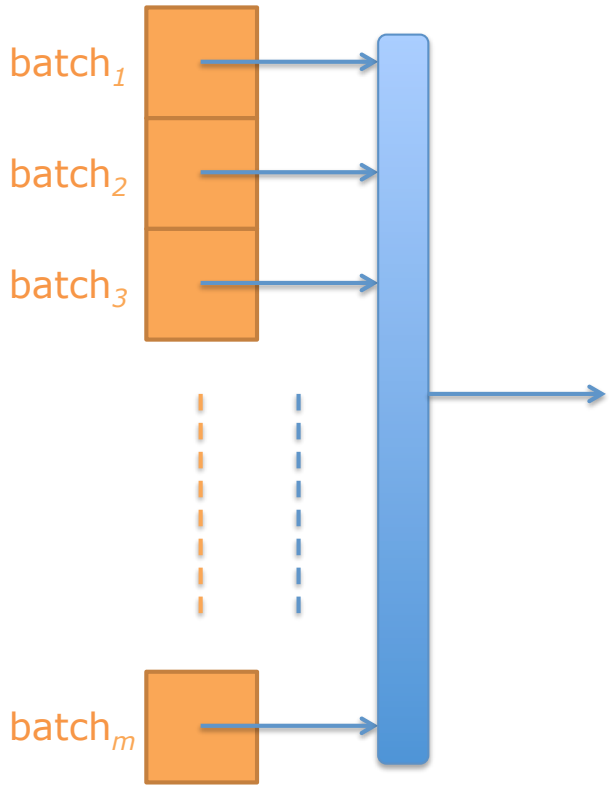


$p(\text{par}|\text{data})$



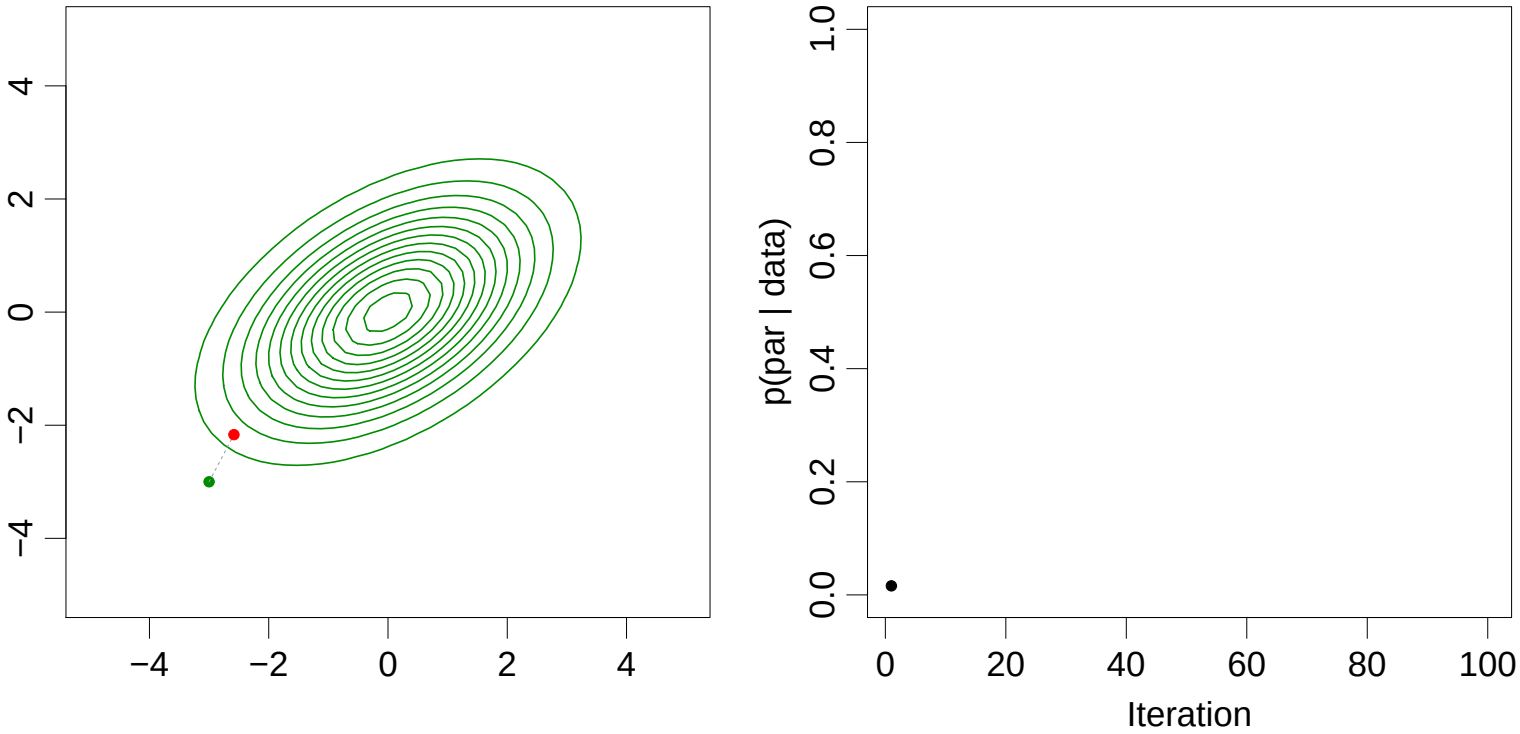
Scaling Inference using Mini-Batches

$$p(\text{data}|\text{par}) = \prod_{i=1}^{n} p(\text{data}_i|\text{par})$$



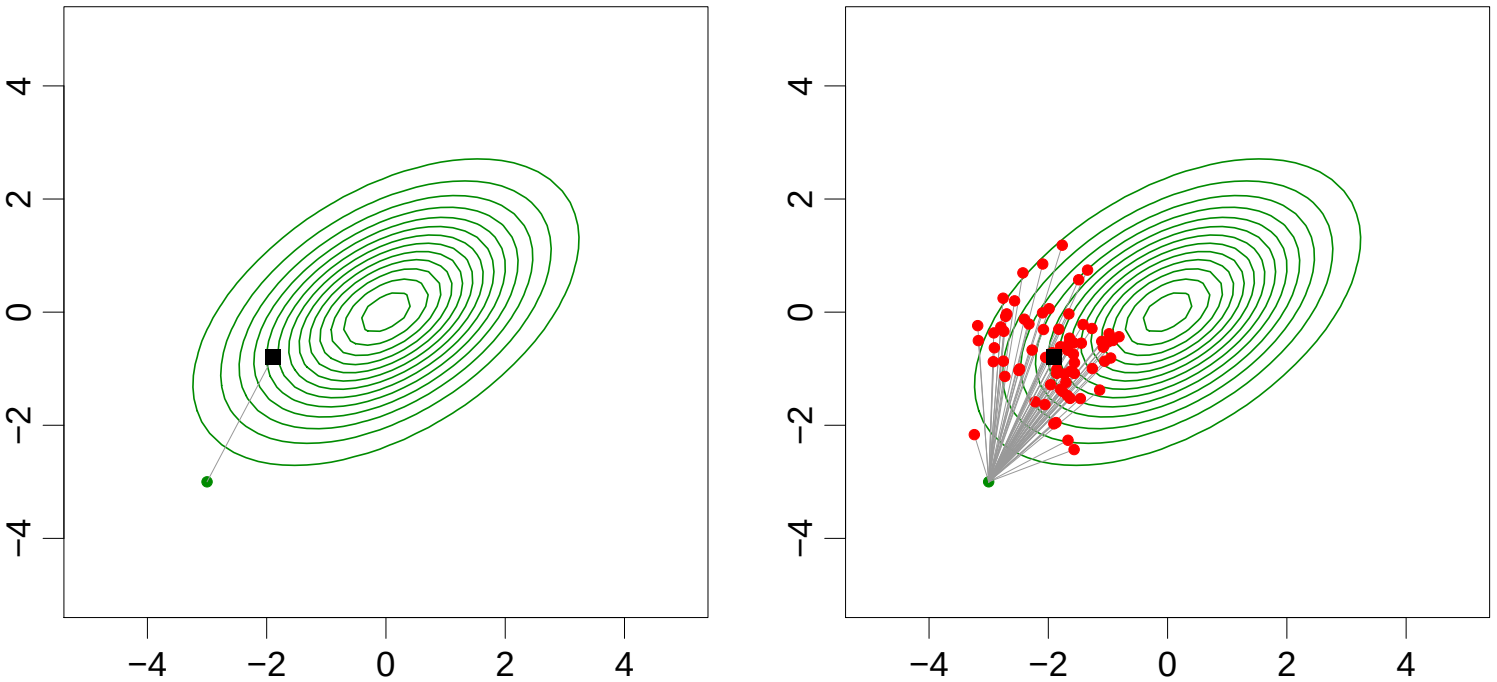
Gradient ascent

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



Stochastic Gradient ascent

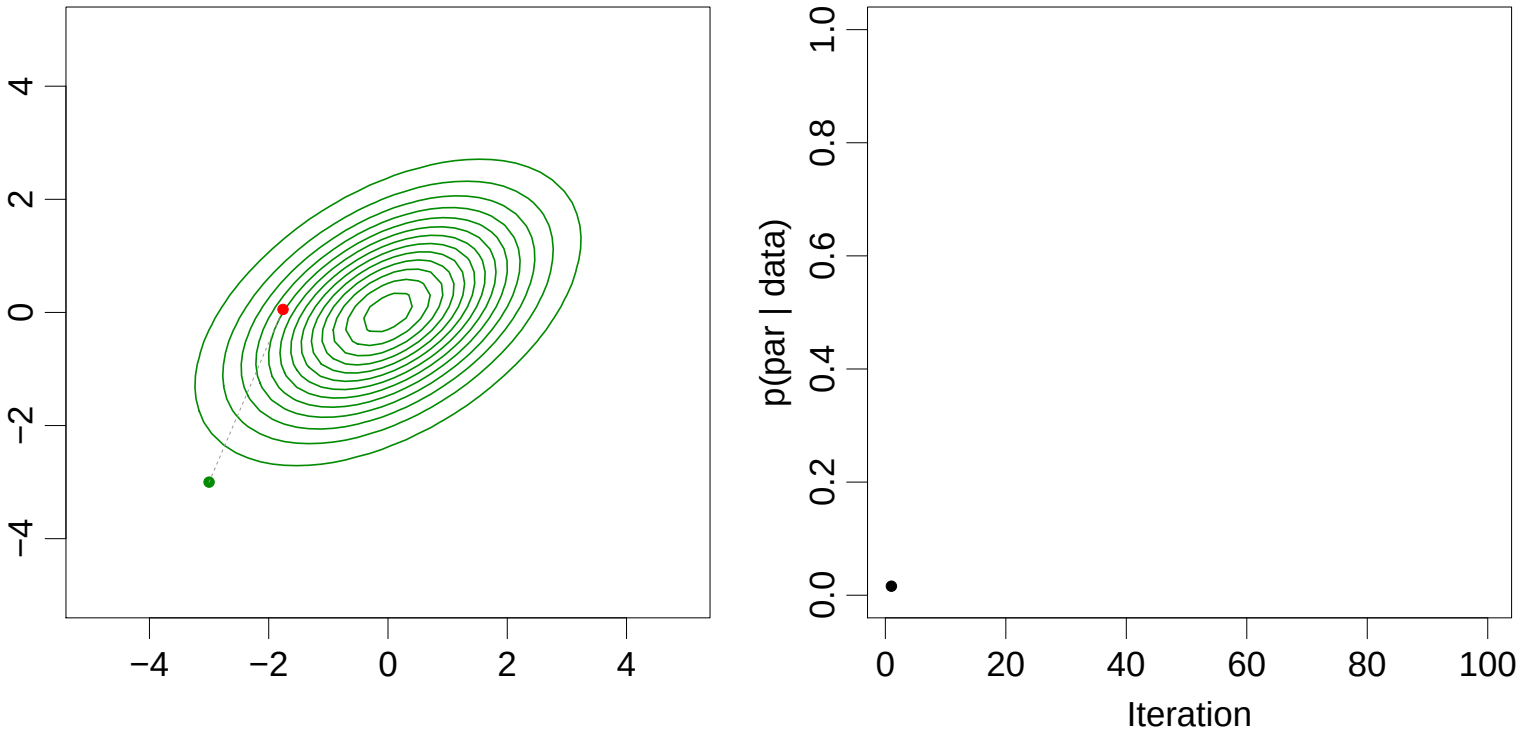
$$\mathrm{E}\left\{\widetilde{\nabla_{\mathrm{par}}}\log[p(\mathrm{data}|\mathrm{par})]\right\} = \nabla_{\mathrm{par}}\log[p(\mathrm{data}|\mathrm{par})]$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient ascent

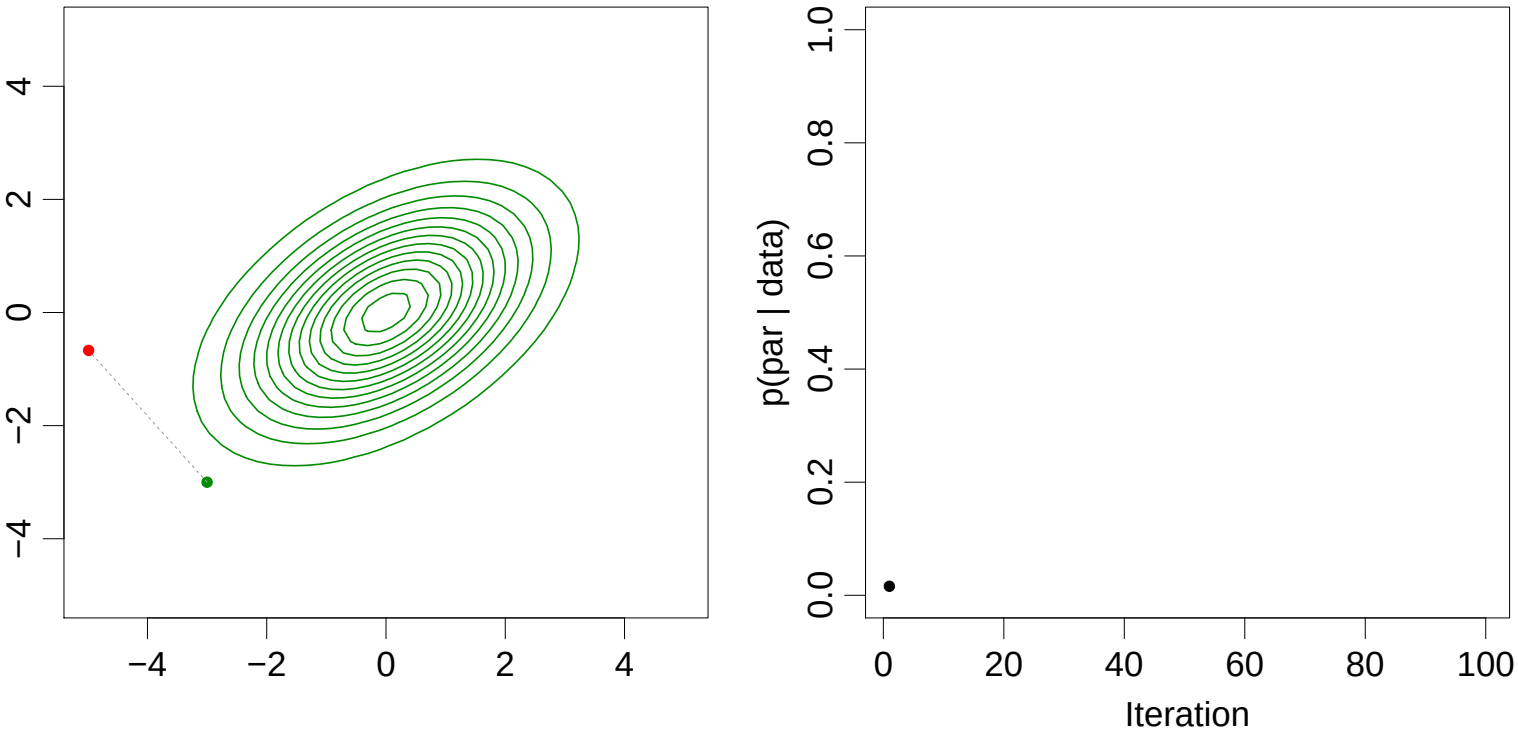
$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})p(\text{par})] \qquad \alpha_t \to 0$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient Langevin Dynamics algorithm

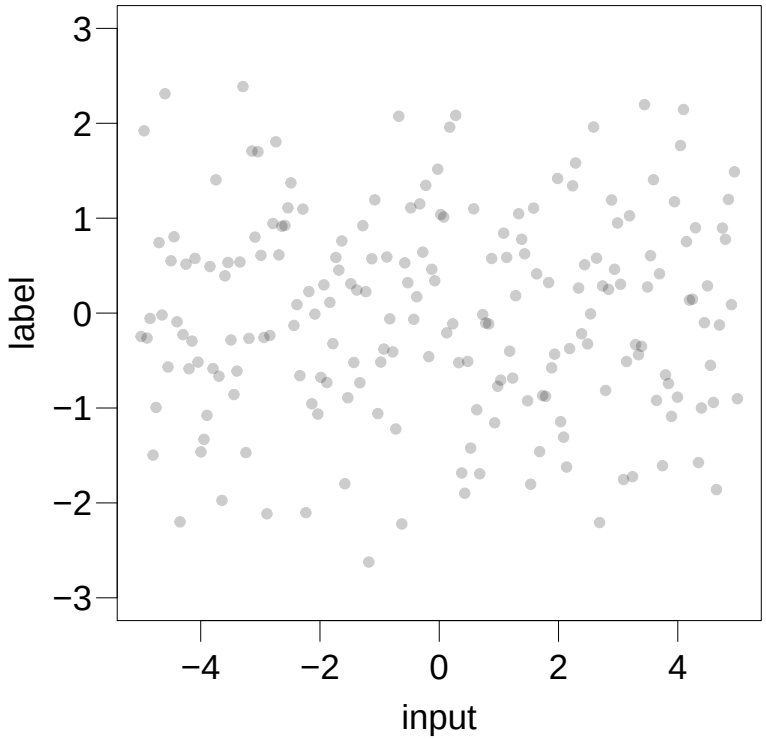
$$\text{par}' = \text{par} + \frac{\alpha_t}{2}\widetilde{\nabla_{\text{par}}} \log[p(\text{data}|\text{par})p(\text{par})] + \eta_t \qquad \eta_t \sim \mathcal{N}(0, \alpha_t)$$

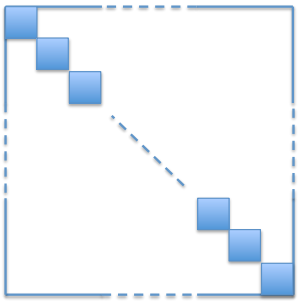


Welling and Teh, *ICML*, 2011

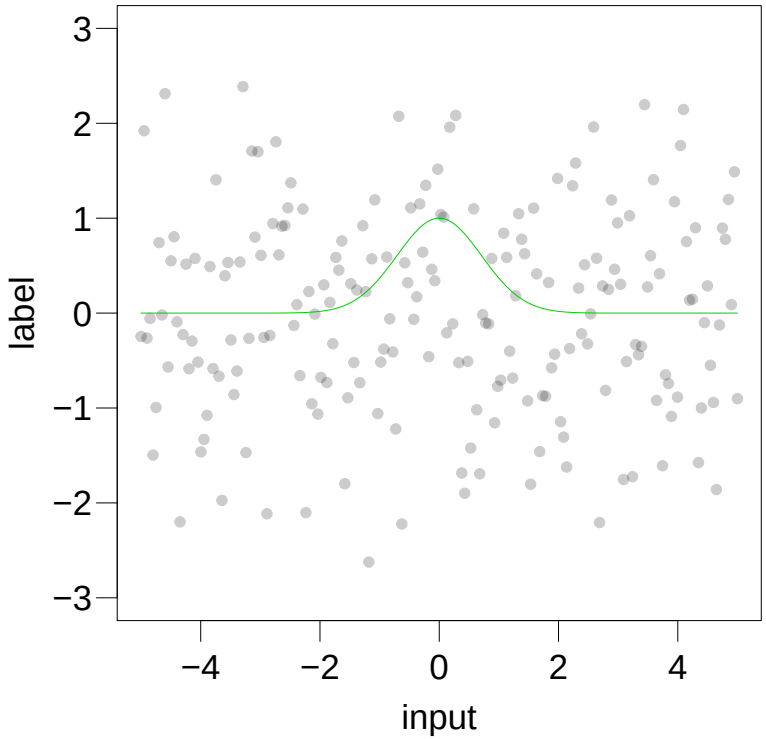
What if $p(\text{data}|\text{par})$ does NOT factorize into $\prod_{i=1}^{n} p(\text{data}_i|\text{par})$?

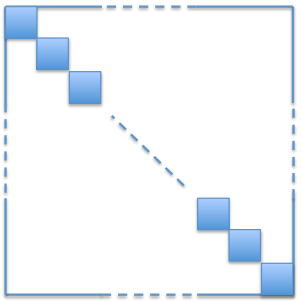
Gaussian Processes



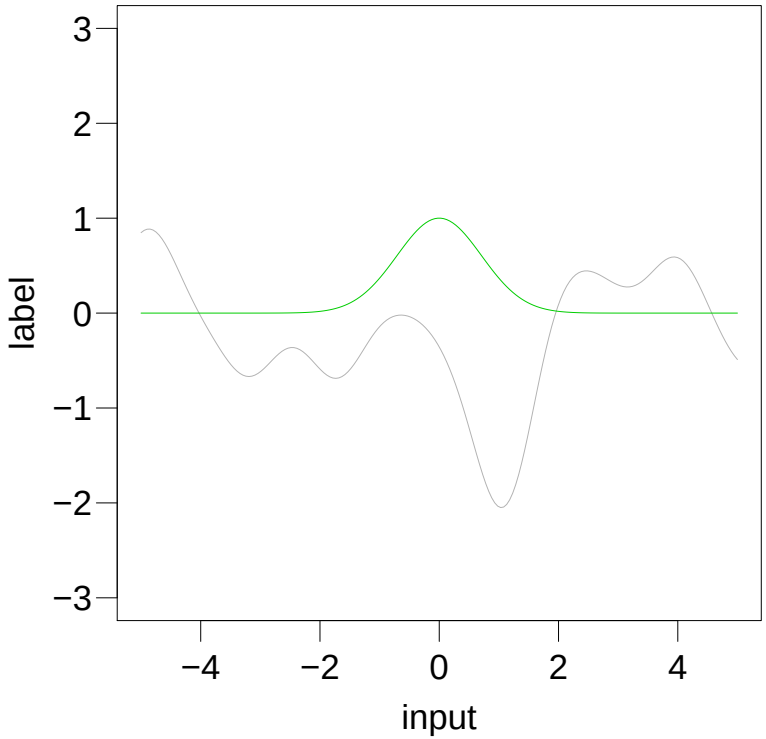
$K =$ 

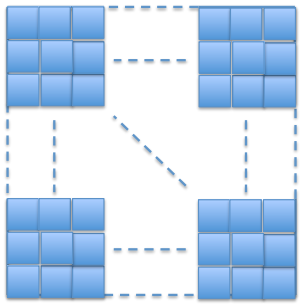
Gaussian Processes



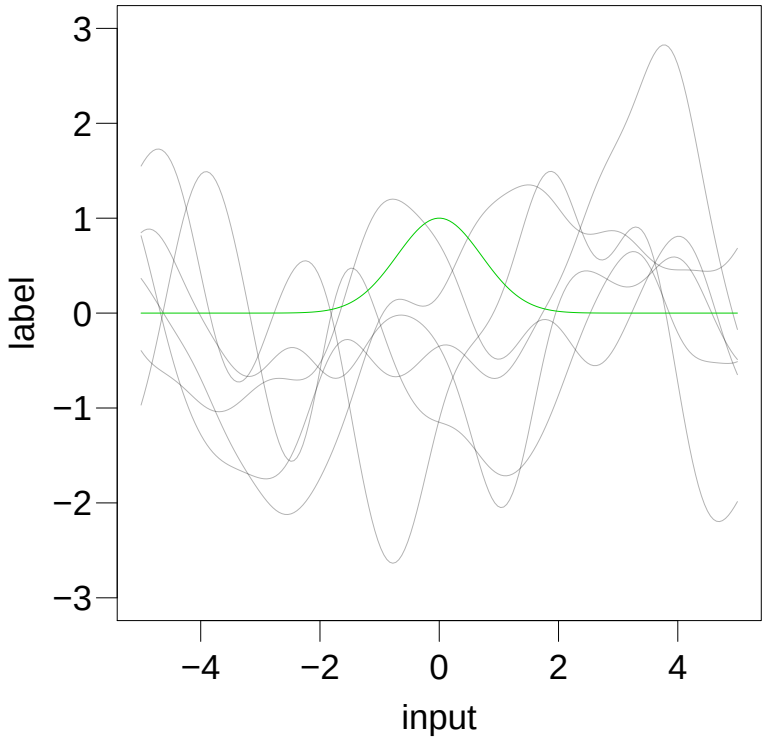
$K =$ 

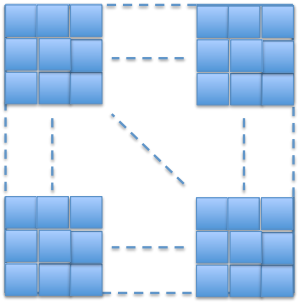
Gaussian Processes



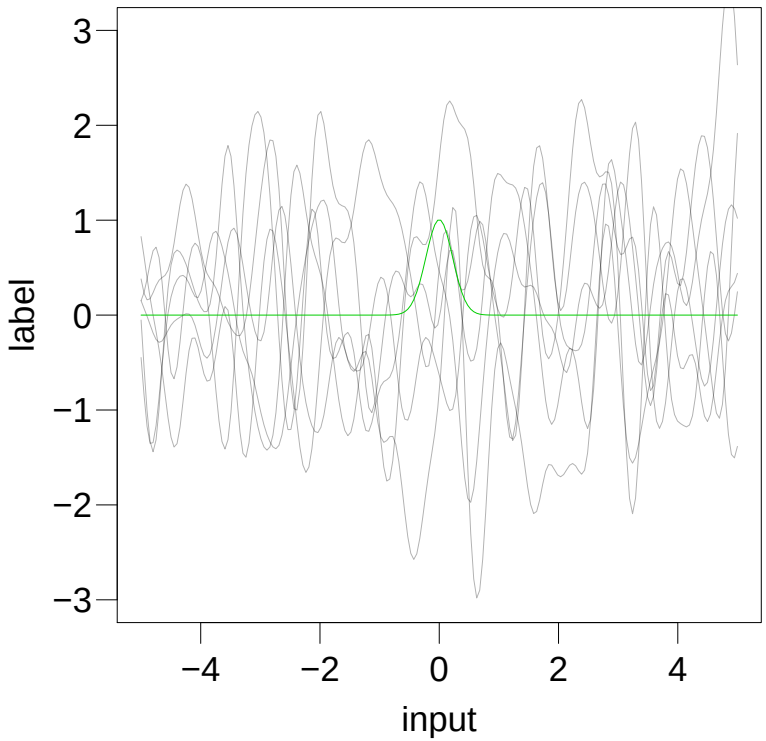
$K =$ 

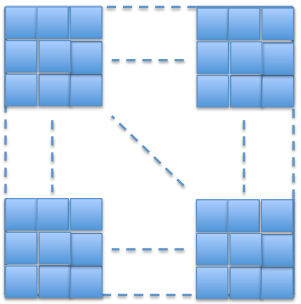
Gaussian Processes



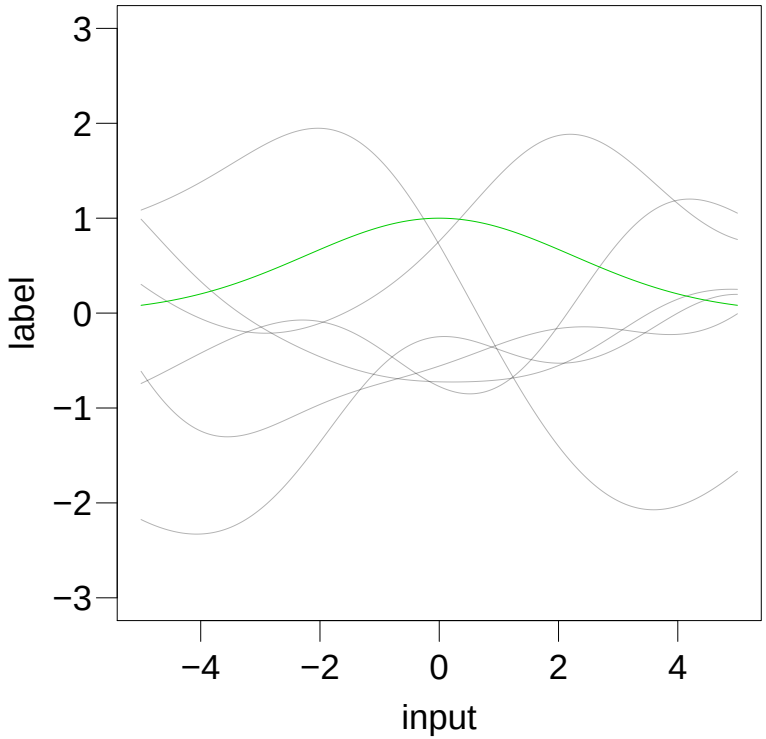
$K =$ 

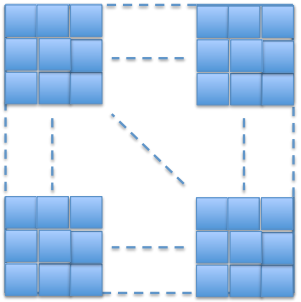
Gaussian Processes



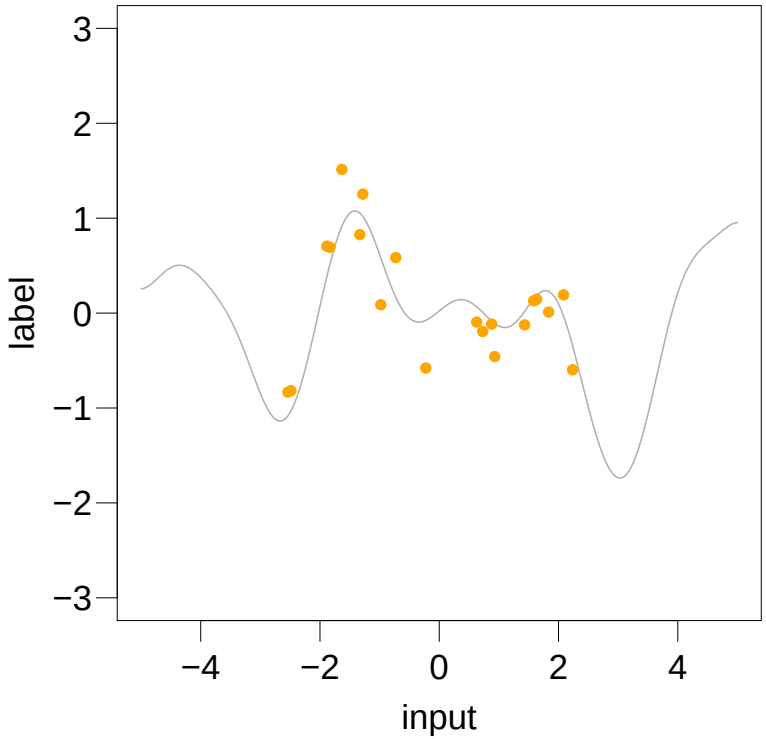
$K =$ 

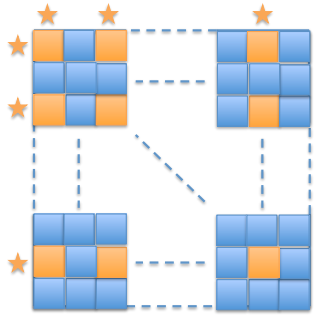
Gaussian Processes



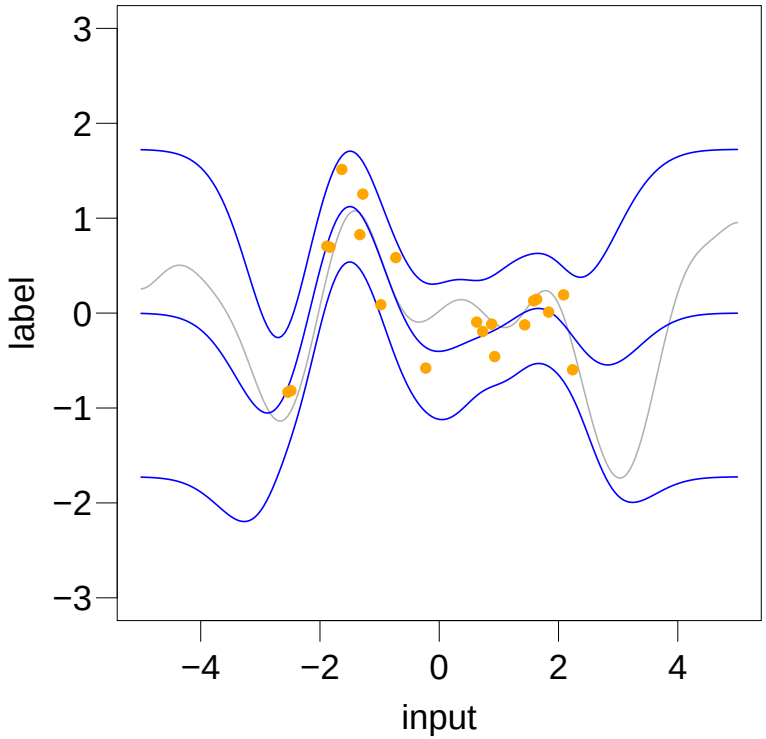
$K =$ 

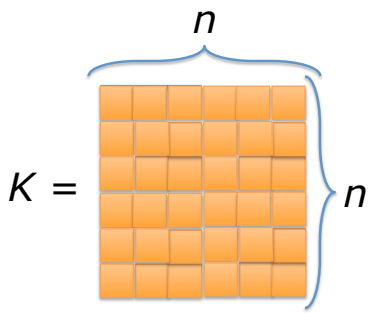
Gaussian Processes



$K =$ 

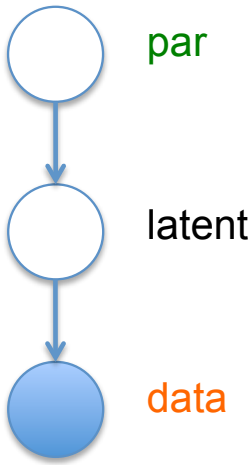
Gaussian Processes



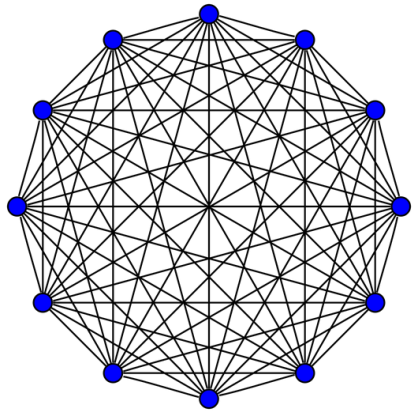


- Gaussian Process models

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



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



Non-factorizing likelihood

- 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 $K = K(\text{par})$ is an $n \times n$ dense matrix!

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

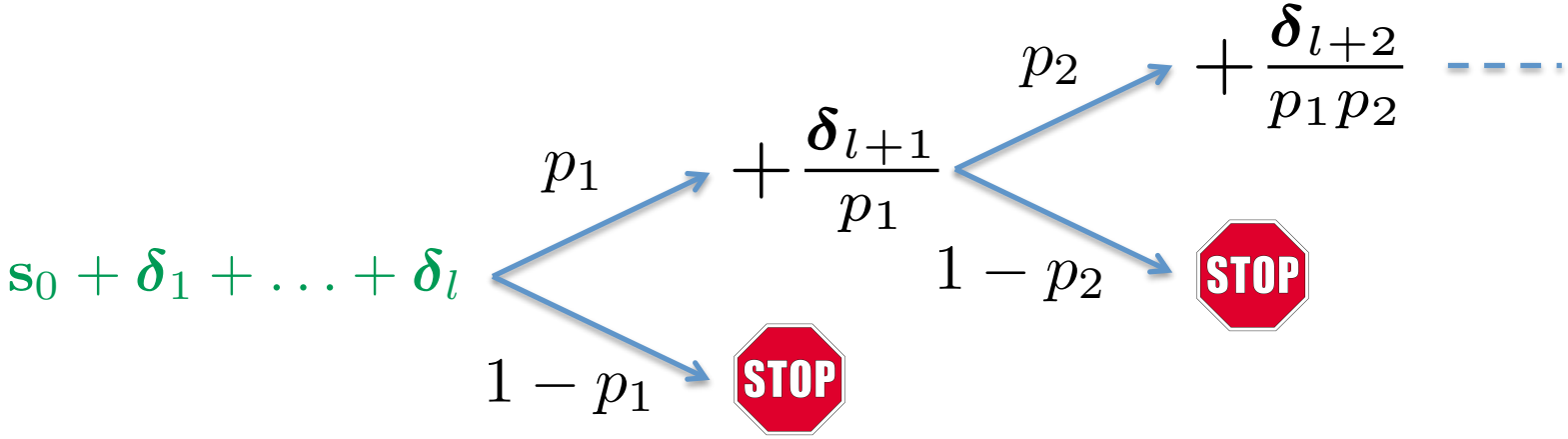
- Accelerate the solution of dense linear systems
- . . . returning an unbiased estimate of the solution

The Unbiased LInear System SolvEr - ULISSE

- Full CG solution:

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

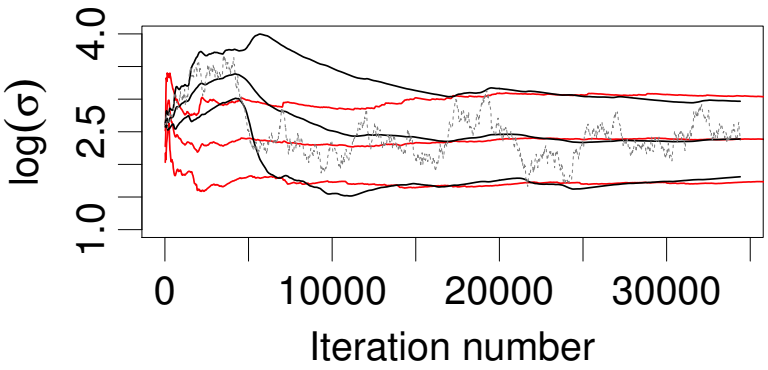
- ULISSE:

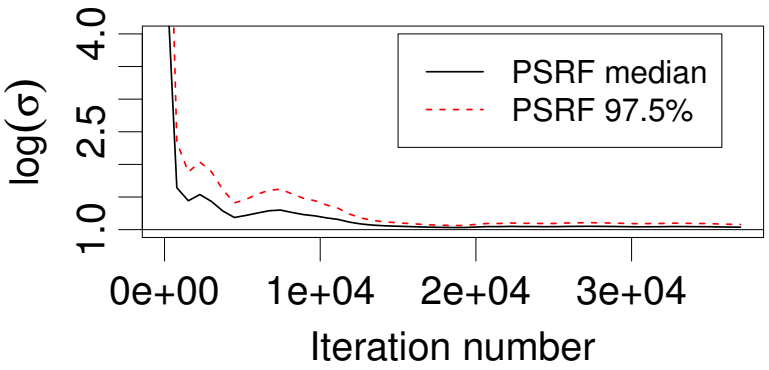


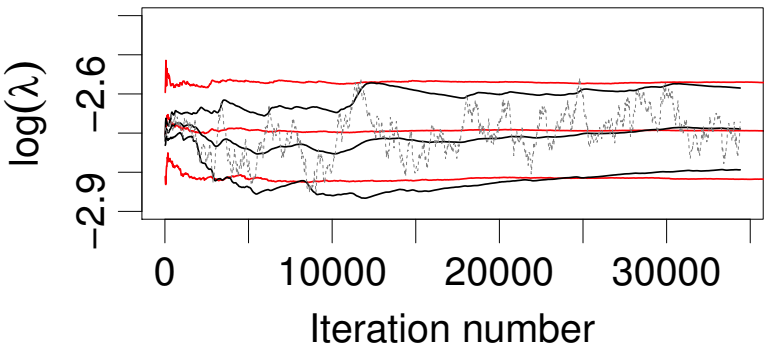
- Final solution is an unbiased estimate of $\mathbf{s}$!

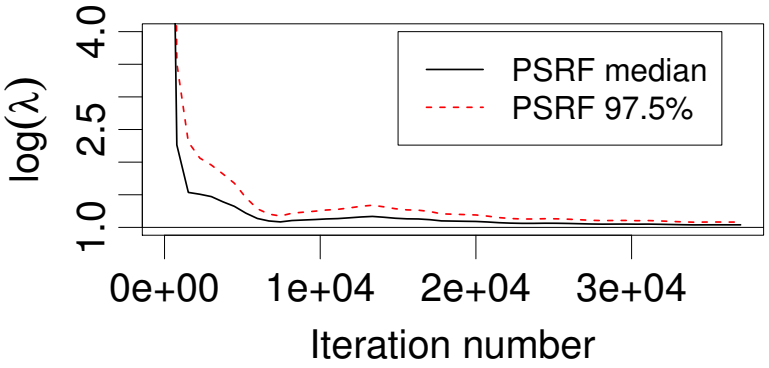
Filippone and Engler, *ICML*, 2015

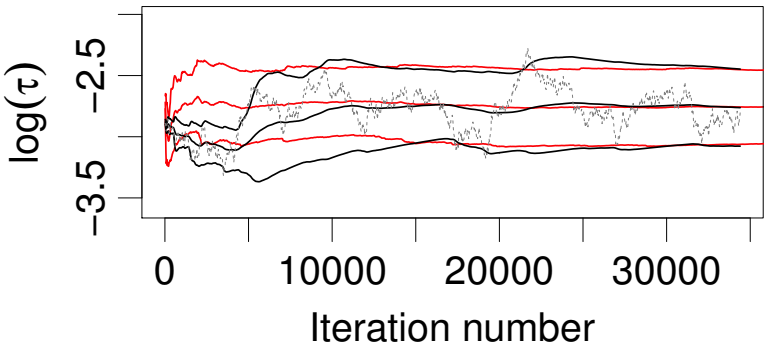
Comparison with MCMC - Concrete dataset - $n \approx 1K$

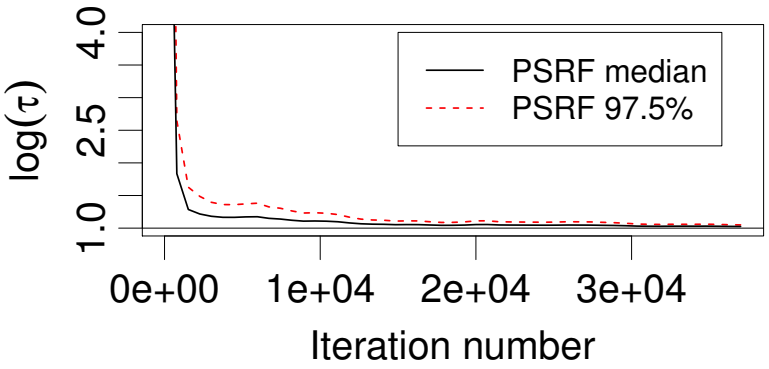




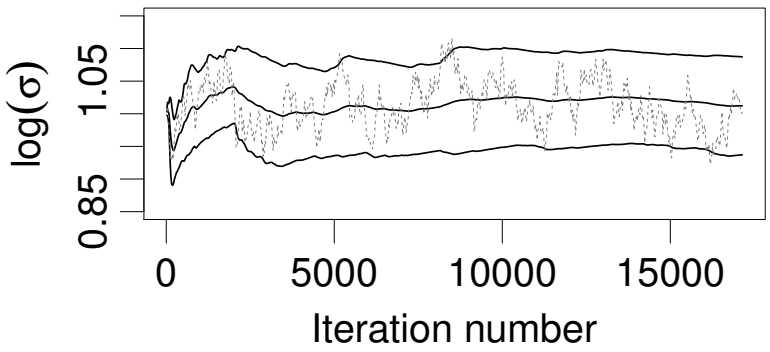


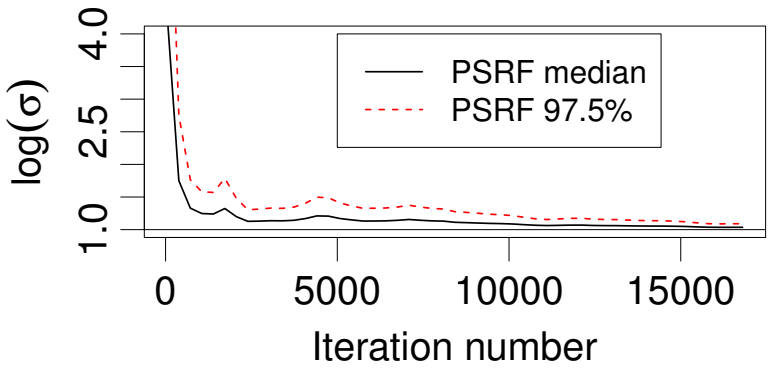


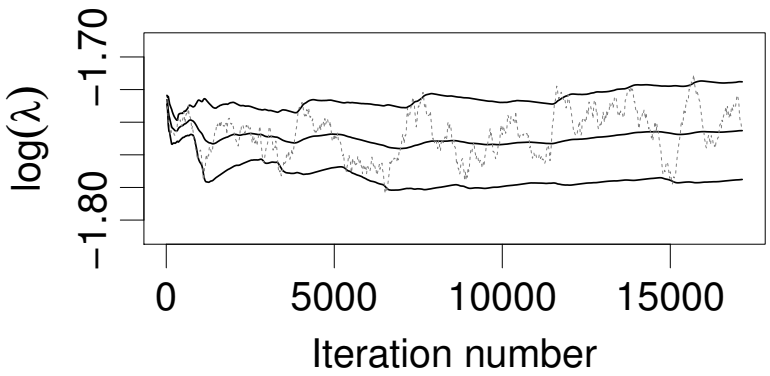


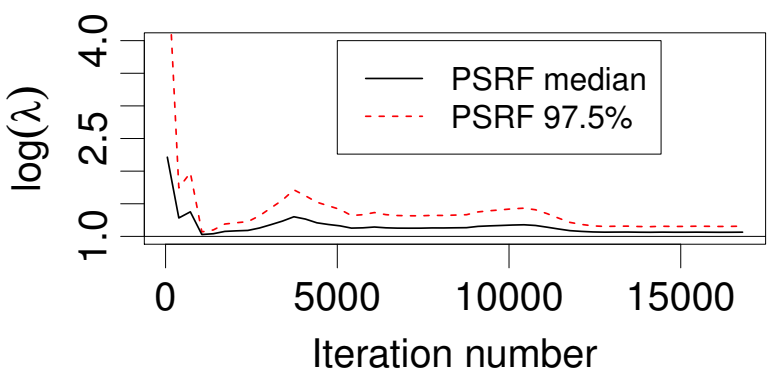


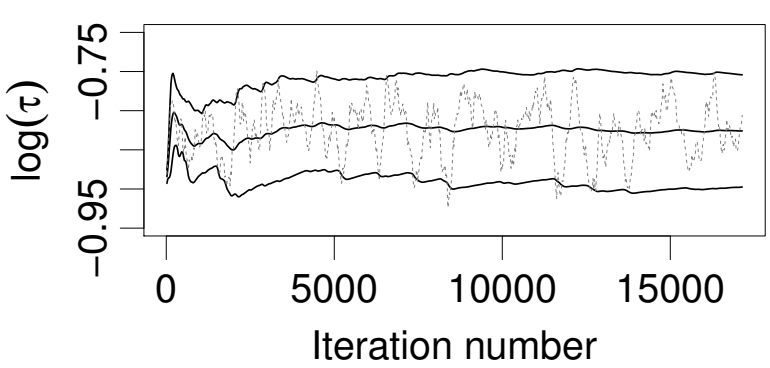
Larger n - Census dataset - $n \approx 23K$

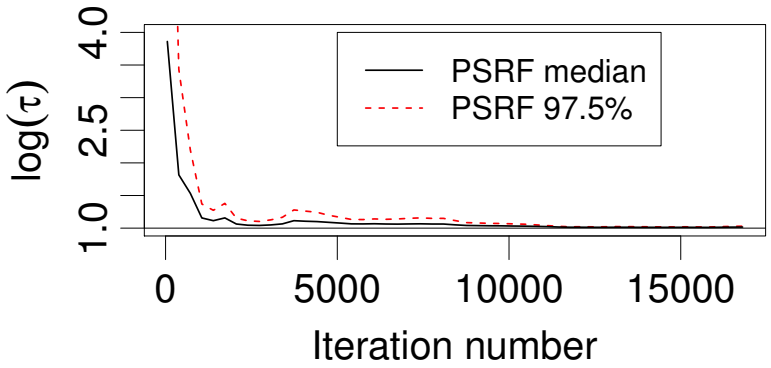












Conclusions and ongoing work

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