

Scaling inference for Gaussian processes using stochastic linear algebra techniques

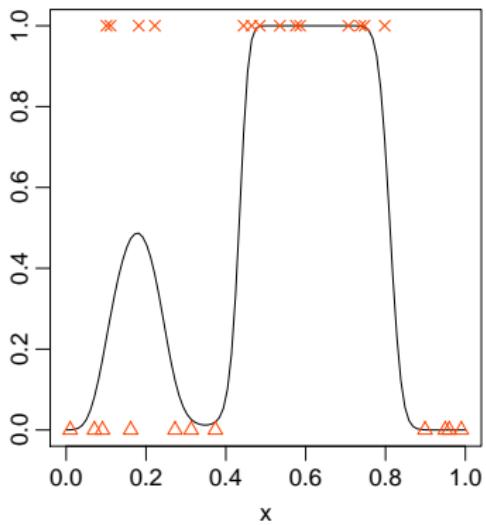
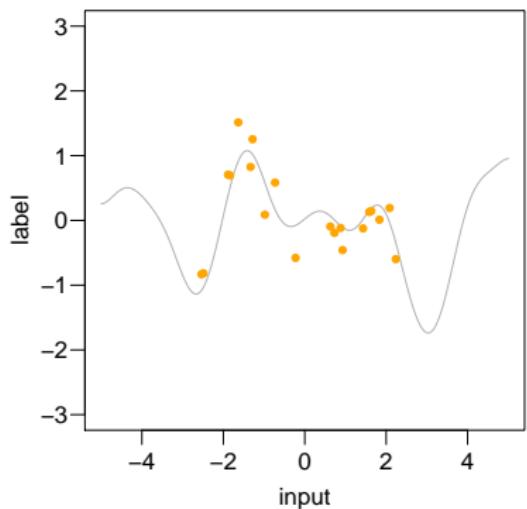
Maurizio Filippone

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

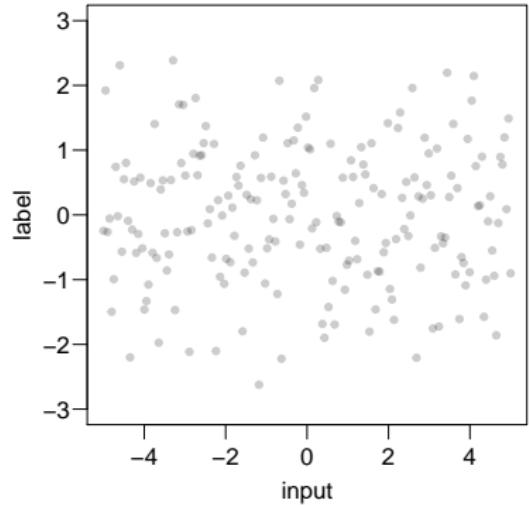
maurizio.filippone@glasgow.ac.uk

June 23rd, 2015

Data modeling examples



Gaussian Processes - priors over functions

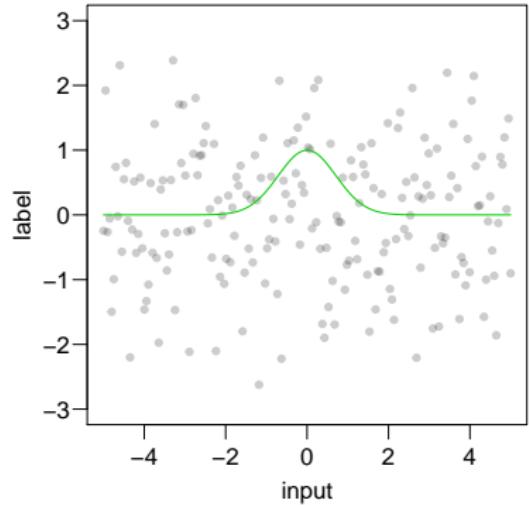


$$K = \begin{matrix} & \text{---} \\ \text{---} & \end{matrix}$$

A diagram illustrating a covariance matrix K . It is represented as a square grid of blue squares. The main diagonal consists of solid blue squares. Above the main diagonal, there are dashed blue squares in the first two positions. Below the main diagonal, there are dashed blue squares in the last two positions. The rest of the grid is white, representing zero entries. This structure indicates a sparse covariance matrix with specific long-range dependencies.

$$K =$$

Gaussian Processes - priors over functions

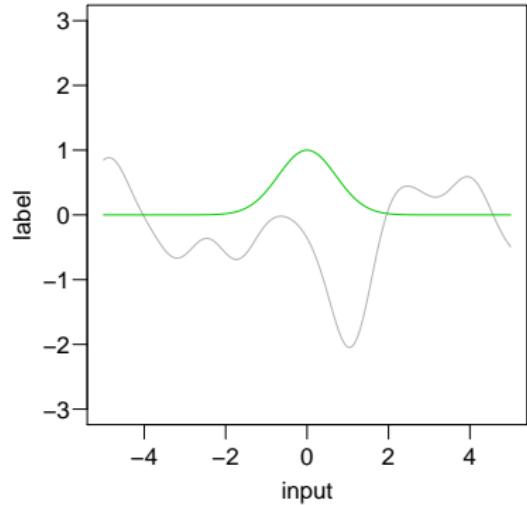


$$K = \begin{matrix} & \text{---} \\ \text{---} & \end{matrix}$$

A diagram illustrating a covariance matrix K . It shows a square grid with blue squares at the diagonal positions and dashed lines connecting them, representing a positive definite matrix structure.

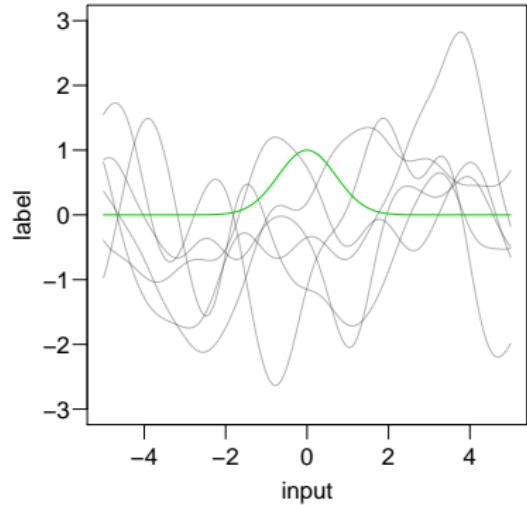
$$K =$$

Gaussian Processes - priors over functions



$$K = \begin{matrix} & \begin{matrix} \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \end{matrix} \\ \begin{matrix} \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \end{matrix} & \begin{matrix} \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} \end{matrix} \end{matrix}$$

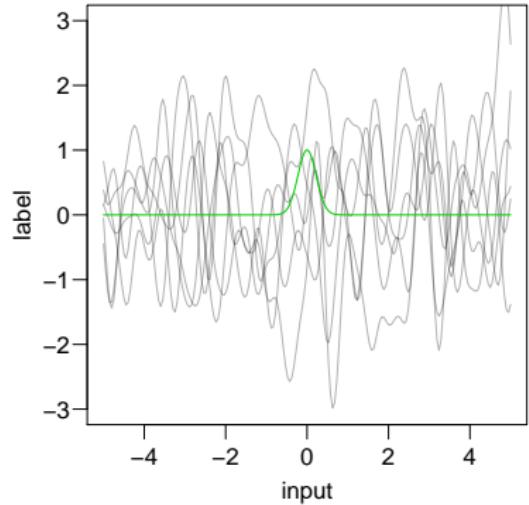
Gaussian Processes - priors over functions



$$K = \begin{matrix} & \begin{matrix} \text{---} & \end{matrix} \\ \begin{matrix} \text{---} & \end{matrix} & \begin{matrix} \text{---} & \end{matrix} \end{matrix}$$

A diagram illustrating the covariance matrix K of a Gaussian process. It consists of four 3x3 grids of blue squares, arranged in a 2x2 pattern. Dashed lines connect the centers of the grids, forming a larger 4x4 grid structure. This represents the covariance between different points in the input space.

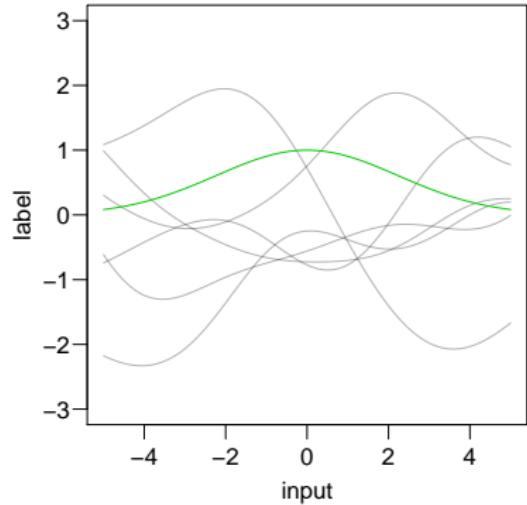
Gaussian Processes - priors over functions



$$K = \begin{matrix} \text{---} & \text{---} \\ \text{---} & \end{matrix}$$

A diagram illustrating a covariance matrix K . It consists of four 3x3 grids of blue squares arranged in a 2x2 pattern. Dashed lines connect the centers of the grids, forming a larger 4x4 square. This represents the covariance structure between multiple input points.

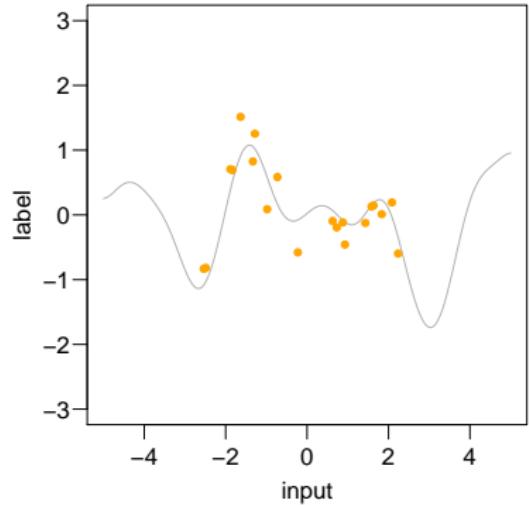
Gaussian Processes - priors over functions



$$K = \begin{matrix} \text{---} & \text{---} \\ \text{---} & \text{---} \end{matrix}$$

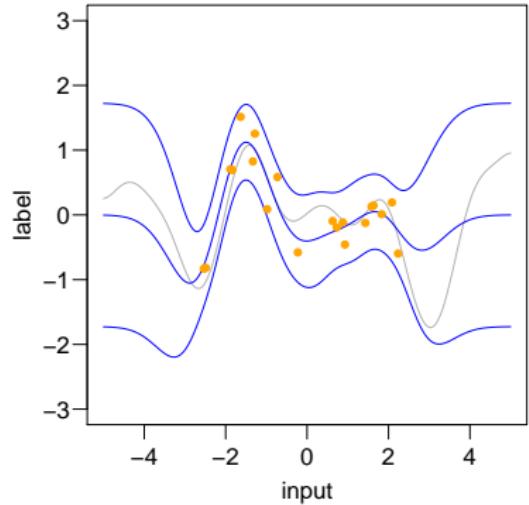
$K =$

Gaussian Processes - priors over functions



$$K = \begin{matrix} & \star & \star \\ \star & \begin{matrix} \text{blue} & \text{orange} \\ \text{blue} & \text{orange} \end{matrix} & \begin{matrix} \text{blue} & \text{orange} \\ \text{blue} & \text{orange} \end{matrix} \\ \star & \begin{matrix} \text{blue} & \text{orange} \\ \text{blue} & \text{orange} \end{matrix} & \begin{matrix} \text{blue} & \text{orange} \\ \text{blue} & \text{orange} \end{matrix} \end{matrix}$$

Gaussian Processes - priors over functions



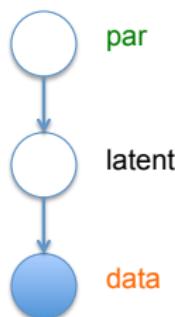
$$K = \begin{matrix} & n \\ \text{---} & \text{---} \\ & n \end{matrix}$$

A diagram illustrating a covariance matrix K . It is represented as a square grid of orange squares, indicating a $n \times n$ matrix where n is the number of data points. Brackets above and to the right of the grid both have the label n , representing the dimension of the matrix.

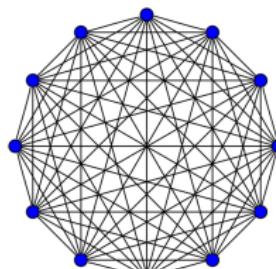
Gaussian Process models

- Gaussian Process models

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



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



Marginal 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

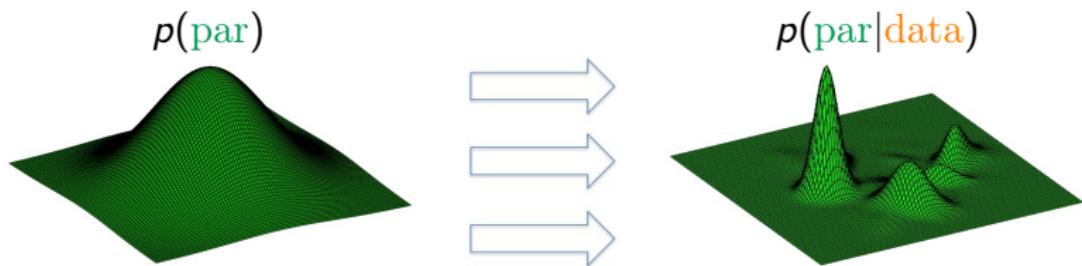
- ... even then

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

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

Bayesian Inference

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



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

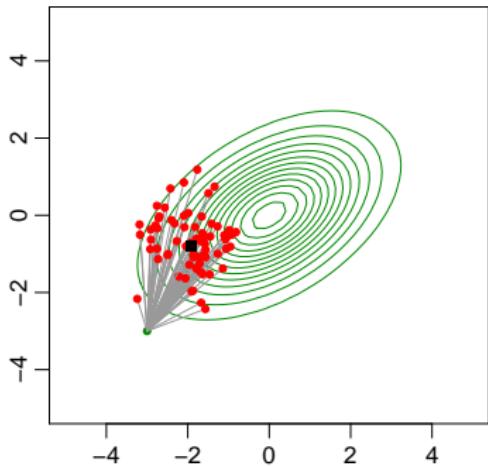
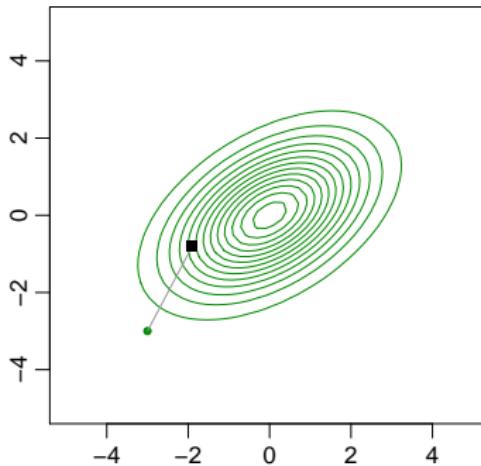
Metropolis et al., *JoCP*, 1953 - Hastings, *Biometrika*, 1970

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 Langevin Dynamics algorithm

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

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:

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

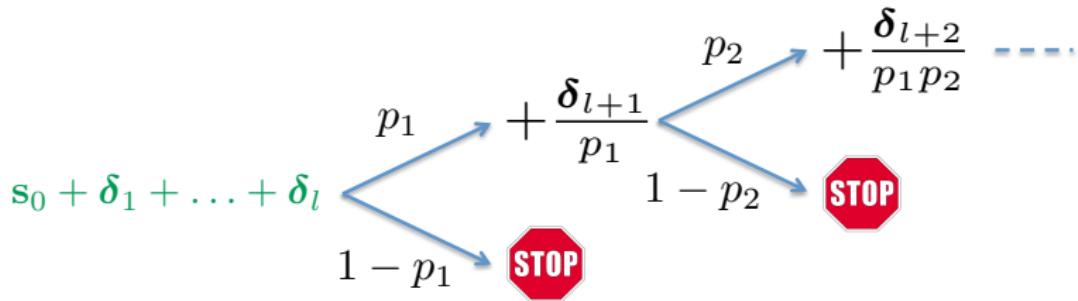
- Iterative update $s = s_0 + \delta_1 + \dots + \delta_T$
- Requires only Kv multiplications! $O(n^2)$ time
- No need to store K ! $O(n)$ space

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

- Full CG solution:

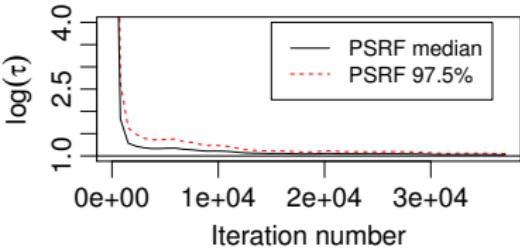
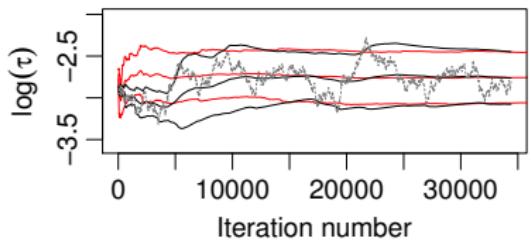
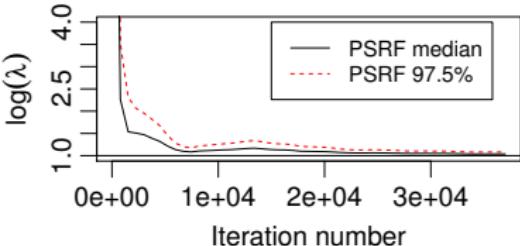
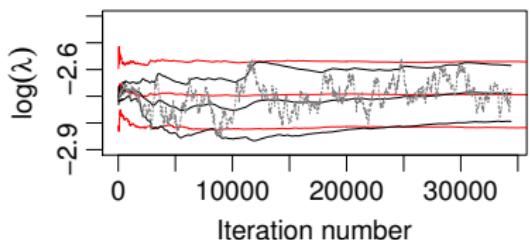
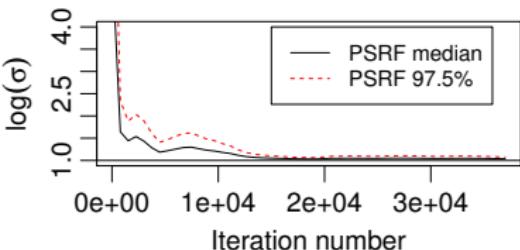
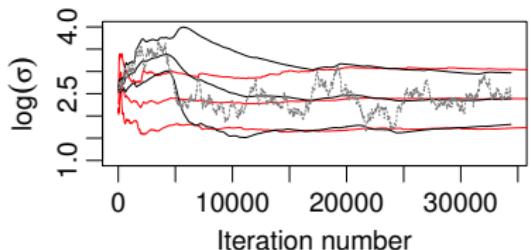
$$\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\delta}_1 + \dots + \boldsymbol{\delta}_l + \boldsymbol{\delta}_{l+1} \dots + \boldsymbol{\delta}_T$$

- ULISSE:

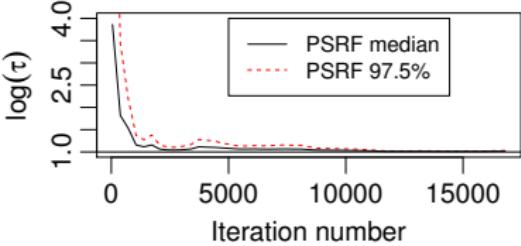
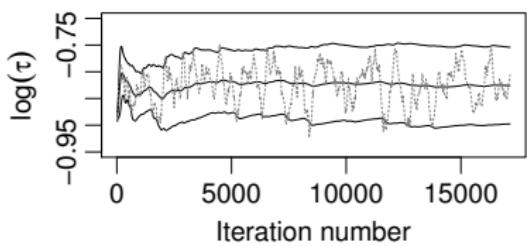
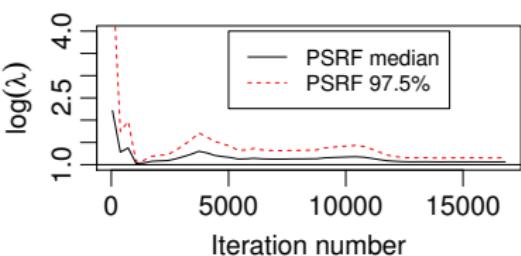
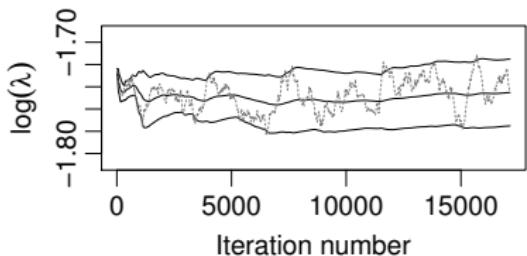
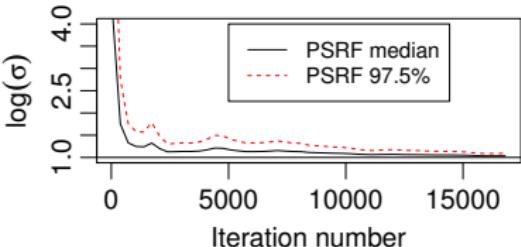
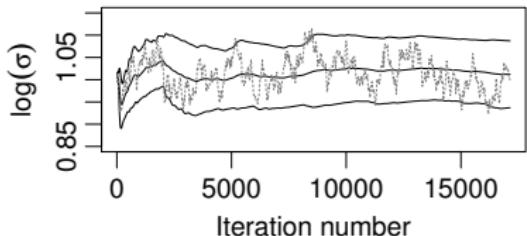


- Final solution is an unbiased estimate of \mathbf{s} !

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
- 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.