Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE)

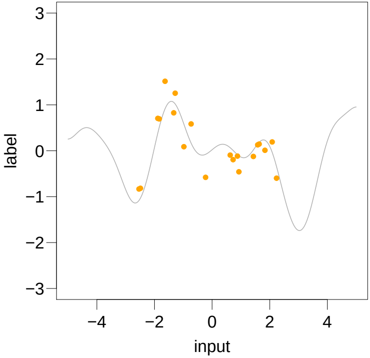
Maurizio Filippone

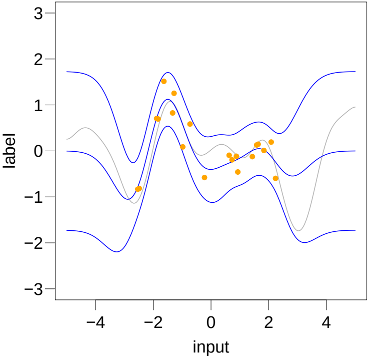
EURECOM, Sophia Antipolis, France &
University of Glasgow, Glasgow, UK

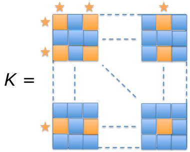
Maurizio.Filippone@eurecom.fr

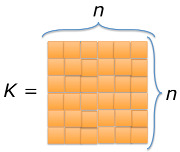
ICML 2015, Lille
July 9th, 2015

Gaussian Processes





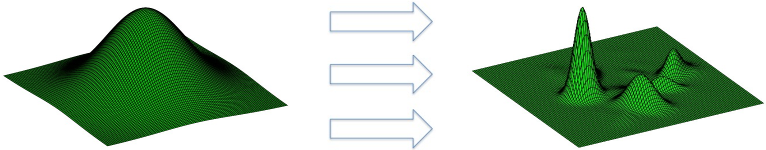




Bayesian Inference

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

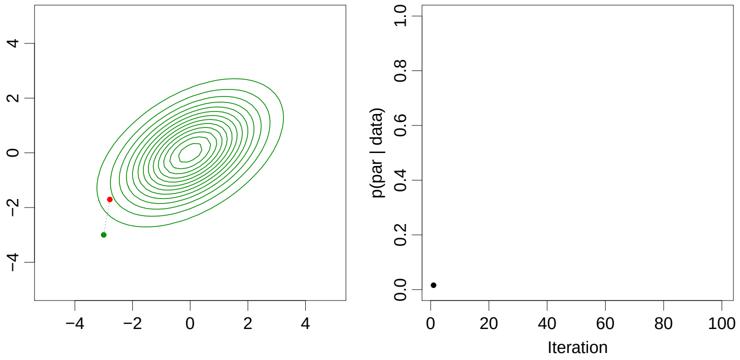
$p(\mathrm{par})$ $p(\mathrm{par}|\mathbf{y})$



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

Markov chain Monte Carlo - Random walk example

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



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

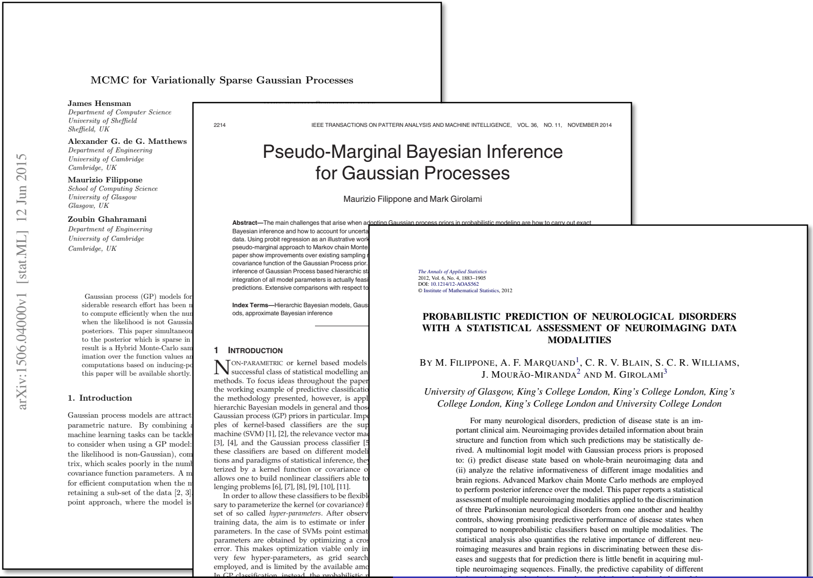
Marginal likelihood

- Gaussian likelihood case

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

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

Should we care about inference of covariance parameters?



MCMC for Variationally Sparse Gaussian Processes

James Hensman
Department of Computer Science
University of Sheffield
Sheffield, UK

Alexander G. de G. Matthews
Department of Engineering
University of Cambridge
Cambridge, UK

Maurizio Filippone
School of Computing Science
University of Glasgow
Glasgow, UK

Zoubin Ghahramani
Department of Engineering
University of Cambridge
Cambridge, UK

arXiv:1506.04000v1 [stat.ML] 12 Jun 2015

Gaussian process (GP) models for
siderable research effort has been m
to compute efficiently when the num
when the likelihood is not Gaussia
posteriors. This paper simultaneou
to the posterior which is sparse in
result is a Hybrid Monte-Carlo sam
imation over the function values an
computations based on inducing-po
this paper will be available shortly.

1. Introduction

Gaussian process models are attract
parametric nature. By combining a
machine learning tasks can be tackle
to consider when using a GP model:
the likelihood is non-Gaussian), com
trix, which scales poorly in the numb
covariance function parameters. A m
for efficient computation when the n
retaining a sub-set of the data [2, 3]
point approach, where the model is

2214 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 11, NOVEMBER 2014

Pseudo-Marginal Bayesian Inference for Gaussian Processes

Maurizio Filippone and Mark Girolami

Abstract—The main challenges that arise when adopting Gaussian process priors in probabilistic modeling are how to carry out exact
Bayesian inference and how to account for uncerta
data. Using probit regression as an illustrative work
pseudo-marginal approach to Markov chain Monte
paper show improvements over existing sampling
covariance function of the Gaussian Process prior.
inference of Gaussian Process based hierarchic st
integration of all model parameters is actually feasi
predictions. Extensive comparisons with respect to

Index Terms—Hierarchic Bayesian models, Gaus
ods, approximate Bayesian inference

1 INTRODUCTION

NON-PARAMETRIC or kernel based models
successful class of statistical modelling an
methods. To focus ideas throughout the paper
the working example of predictive classificatio
the methodology presented, however, is appl
hierarchic Bayesian models in general and thos
Gaussian process (GP) priors in particular. Impo
ples of kernel-based classifiers are the sup
machine (SVM) [1], [2], the relevance vector mac
[3], [4], and the Gaussian process classifier [5
these classifiers are based on different modeli
tions and paradigms of statistical inference, they
terized by a kernel function or covariance o
allows one to build nonlinear classifiers able to
lenging problems [6], [7], [8], [9], [10], [11].

In order to allow these classifiers to be flexibl
sary to parameterize the kernel (or covariance) f
set of so called *hyper-parameters*. After observ
training data, the aim is to estimate or infer
parameters. In the case of SVMs point estimat
parameters are obtained by optimizing a cros
error. This makes optimization viable only in
very few hyper-parameters, as grid search
employed, and is limited by the available amo
In GP classification, instead, the probabilistic n

The Annals of Applied Statistics
2012, Vol. 6, No. 4, 1883–1905
DOI: 10.1214/12-AOAS562
© Institute of Mathematical Statistics, 2012

PROBABILISTIC PREDICTION OF NEUROLOGICAL DISORDERS WITH A STATISTICAL ASSESSMENT OF NEUROIMAGING DATA MODALITIES

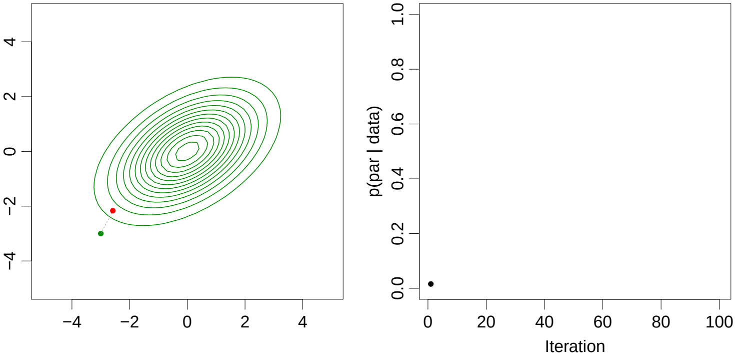
BY M. FILIPPONE, A. F. MARQUAND[1], C. R. V. BLAIN, S. C. R. WILLIAMS, J. MOURÃO-MIRANDA[2] AND M. GIROLAMI[3]

University of Glasgow, King's College London, King's College London, King's College London, King's College London and University College London

For many neurological disorders, prediction of disease state is an important clinical aim. Neuroimaging provides detailed information about brain structure and function from which such predictions may be statistically derived. A multinomial logit model with Gaussian process priors is proposed to: (i) predict disease state based on whole-brain neuroimaging data and (ii) analyze the relative informativeness of different image modalities and brain regions. Advanced Markov chain Monte Carlo methods are employed to perform posterior inference over the model. This paper reports a statistical assessment of multiple neuroimaging modalities applied to the discrimination of three Parkinsonian neurological disorders from one another and healthy controls, showing promising predictive performance of disease states when compared to nonprobabilistic classifiers based on multiple modalities. The statistical analysis also quantifies the relative importance of different neuroimaging measures and brain regions in discriminating between these diseases and suggests that for prediction there is little benefit in acquiring multiple neuroimaging sequences. Finally, the predictive capability of different

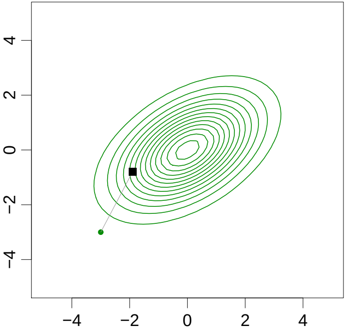
Gradient ascent

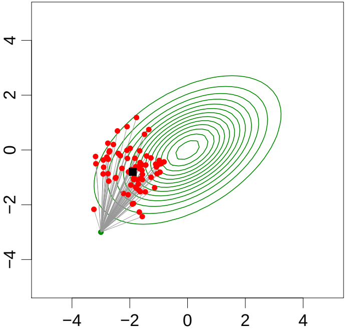
$$\mathrm{par}' = \mathrm{par} + \frac{\alpha}{2} \nabla_{\mathrm{par}} \log[p(\mathbf{y}|\mathrm{par})p(\mathrm{par})]$$



Stochastic Gradient ascent

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

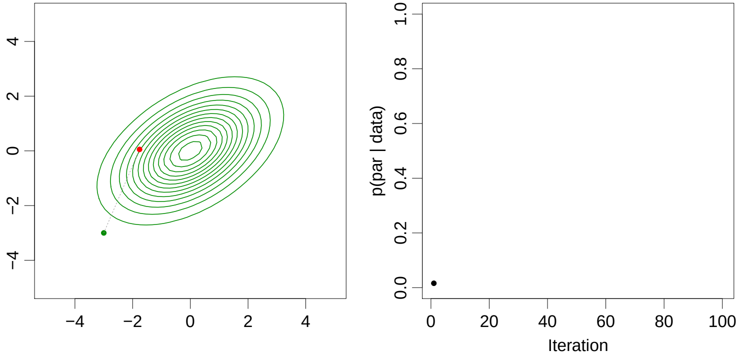




Robbins and Monro, *AoMS*, 1951

Stochastic Gradient ascent

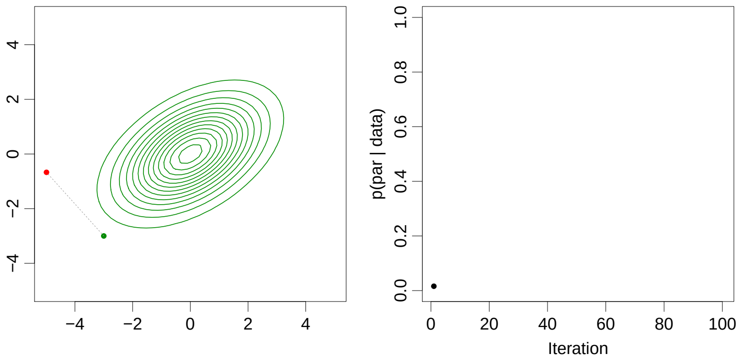
$$\text{par}' = \text{par} + \frac{\alpha_t}{2} \widetilde{\nabla_{\text{par}}} \log[p(\mathbf{y}|\text{par})p(\text{par})] \qquad \alpha_t \to 0$$



Robbins and Monro, *AoMS*, 1951

Stochastic Gradient Langevin Dynamics (SGLD) algorithm

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



Welling and Teh, *ICML*, 2011

Stochastic Gradients in SGLD

- Traditionally, in SGLD stochastic gradients

$$\widetilde{\nabla_{\mathrm{par}}} \log[p(\mathbf{y}|\mathrm{par})p(\mathrm{par})]$$

  are computed based on mini-batches of data
- In GPs the likelihood DOES NOT factorize
- What can we do?

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

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

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

- For example r_j drawn from $\{-1, 1\}$ with $p = 1/2$

Stochastic Gradients in GP regression

- Stochastic estimate of the trace

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

  with $\mathrm{E}[\mathbf{r}\mathbf{r}^{\mathrm{T}}] = I$
- For example r_j drawn from $\{-1, 1\}$ with $p = 1/2$
- Stochastic gradient

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

- Linear systems only!

Filippone and Engler, *ICML*, 2015

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

- Accelerate the solution of dense linear systems
- … returning an unbiased estimate of the solution
- Basic idea - unbiased estimator for generic sums $a + b$:

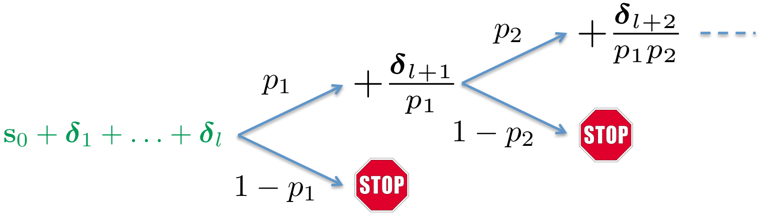
$$\begin{tikzcd} & +\frac{b}{p} \\ a \arrow[ur, "p"] \arrow[dr, "1-p"'] & \\ & +0 \end{tikzcd}$$

The Unbiased LInear System SolvEr - ULISSE

- Full CG solution:

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

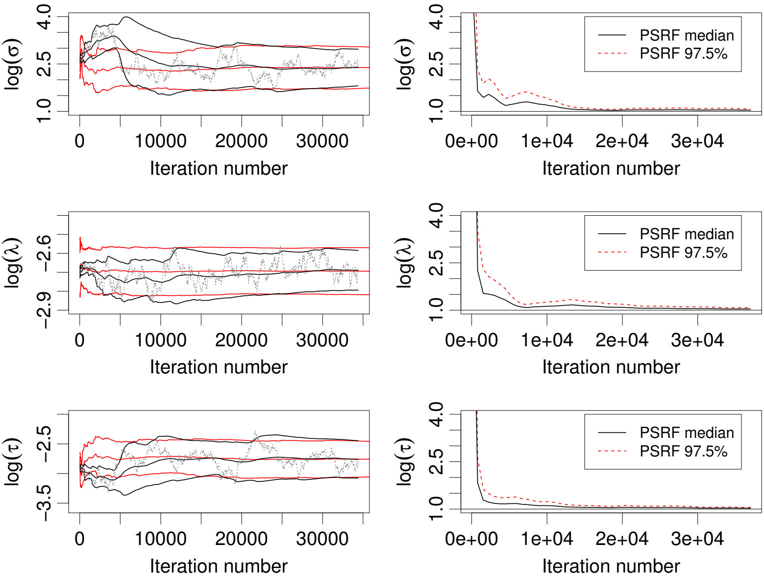
- ULISSE:



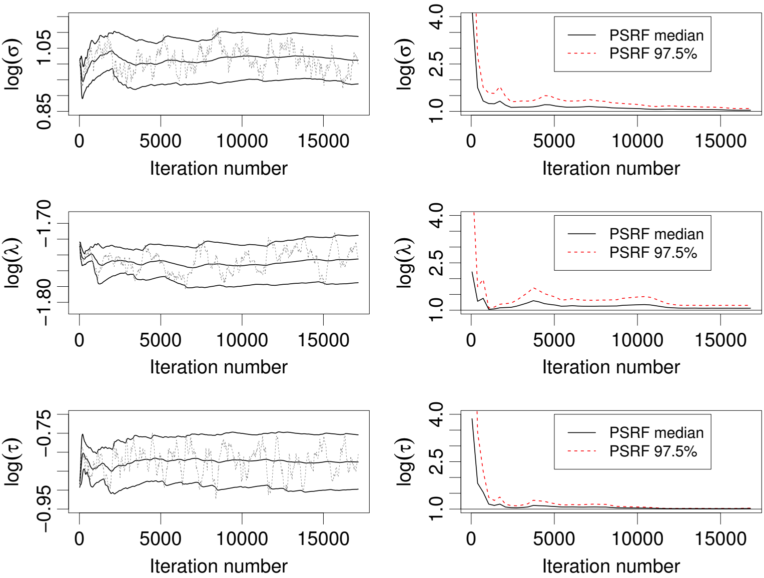
- Final solution is an unbiased estimate of s!

Filippone and Engler, *ICML*, 2015

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



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



Conclusions and ongoing work

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

Conclusions and ongoing work

- "Noisy" MCMC offers a practical and scalable way to carry out "exact" Bayesian computations for GPs
- Novel adaptation of SGLD when the likelihood does not factorize
- Novel linear solver ULISSE to speed up computations of stochastic gradients

Conclusions and ongoing work

- "Noisy" MCMC offers a practical and scalable way to carry out "exact" Bayesian computations for GPs
- Novel adaptation of SGLD when the likelihood does not factorize
- Novel linear solver ULISSE to speed up computations of stochastic gradients
- General likelihoods?
- Preconditioners?

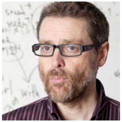
Acknowledgements & References

- Joint work with Raphael Engler

Andre Marquand
Radboud

Guido Sanguinetti
Edinburgh

James Hensman
Sheffield



Mark Girolami
Warwick

Alessandro Vinciarelli
Glasgow

Dirk Husmeier
Glasgow

[1] M. Filippone, A. F. Marquand, C. R. V. Blain, S. C. R. Williams, J. Mourão-Miranda, and M. Girolami. **Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities**. *Annals of Applied Statistics, 6(4):1883-1905, 2012.*

[2] M. Filippone, M. Zhong, and M. Girolami. **A comparative evaluation of stochastic-based inference methods for Gaussian process models**. *Machine Learning, 93(1):93-114, 2013.*

[3] 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.*

[4] M. Filippone. **Bayesian inference for Gaussian process classifiers with annealing and pseudo-marginal MCMC**, In *ICPR 2014, pages 614-619. IEEE, 2014.*

[5] J. Hensman, A. G. de G. Matthews, M. Filippone, and Z. Ghahramani. **MCMC for variationally sparse Gaussian processes**, arXiv:1506.04000, 2015.