A Comparative Evaluation of Stochastic-based Inference Methods for Gaussian Process Models

Maurizio Filippone

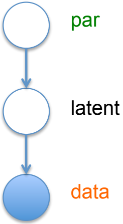
School of Computing Science
University of Glasgow
maurizio.filippone@glasgow.ac.uk

September 25th, 2013

Outline of the talk

1. Gaussian Process Models
2. Inference in GPMs using MCMC
3. Results
4. Conclusions

Gaussian Process Models - GPMs



par

latent

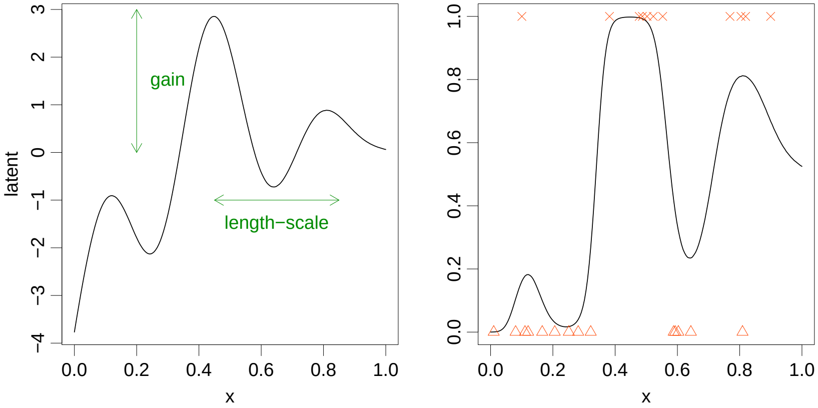
data

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

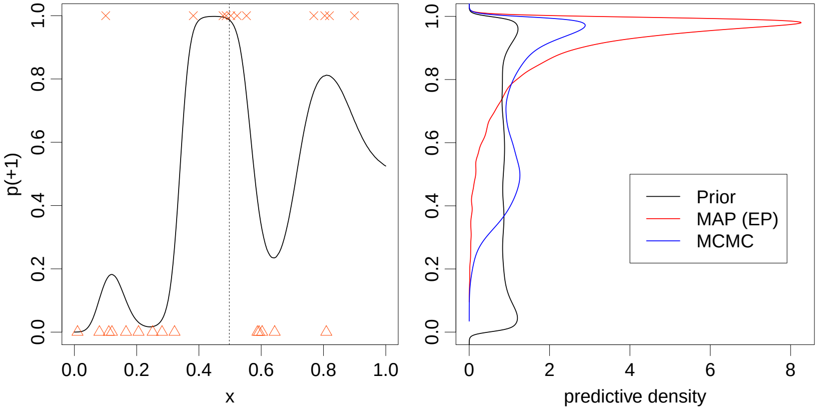
$$p(\text{latent}|\text{par}) = \text{GP}(\mu(\text{par}), K(\text{par}))$$

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

GPM - Probit regression example



GPM - Probit regression example - MAP vs MCMC



Approximate inference

- Approximate marginal likelihood might be inaccurate; no way to quantify degree of inaccuracy
- Quadrature can't be employed if par is large dimensional

MCMC offers a way to do “exact” inference in these cases

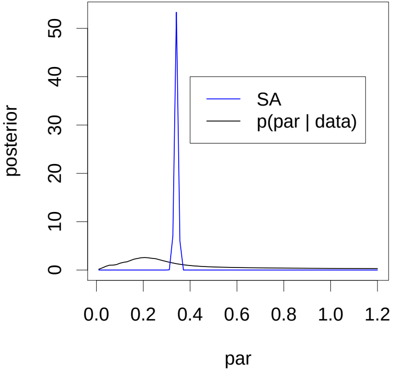
Goal of this work

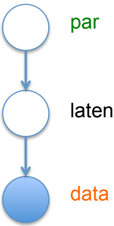
- Study best ways to draw samples $p(\mathrm{latent}, \mathrm{par}|\mathrm{data})$

Challenges in MCMC for GPMs - Structure

Obvious iterative scheme (aka Sufficient Augmentation (SA) scheme). Alternate between:

- Drawing from $p(\mathrm{latent}|\mathrm{par}, \mathrm{data})$
- Drawing from $p(\mathrm{par}|\mathrm{latent})$ (**bad idea** - see figure)





Challenges in MCMC for GPMs - Cost & exploration

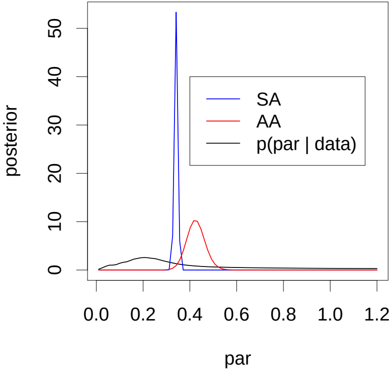
- No exact Gibbs steps - need to employ Metropolis within Gibbs steps - waste of computations when rejecting
- Updates of par cost $O(n^3)$
- par can be large dimensional (e.g., Automatic Relevance Determination (ARD) covariance function)
- There are n latent variables (as many as the number of observations)

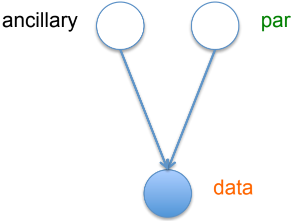
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization:

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

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





Other strategies

- SURR - Surrogate method in Murray and Adams (2010): reparameterization using cleverly constructed auxiliary variables
- KHR - Joint sampler by Knorr-Held and Rue (2002): propose new par' and $\mathrm{latent}'|\mathrm{par}', \mathrm{data}$ and then **jointly** Accept/Reject
- ASIS - Interweave AA and SA as in Yu and Meng (2011)

Transition operators for sampling latent variables

Scaled Metropolis-Hastings (MH) proposals - Neal (1999)

- MH v1:

$$\mathrm{latent}' = \mathrm{latent} + \alpha\, z$$

- MH v2:

$$\mathrm{latent}' = \sqrt{1-\alpha^2}\,\mathrm{latent} + \alpha\, z$$

with

$$z \sim \mathcal{N}(0, K)$$

Transition operators for sampling latent variables

Scaled Hybrid Monte Carlo (HMC) with mass matrix M. Negative Hessian of the log-posterior is $K^{-1} + \Delta(\mathbf{f})$, with Δ diagonal.

- HMC v1 - Christensen et al. (2005)

$$M = K^{-1} + \Delta(\mathbf{0})$$

- HMC v2 - proposed in this work

$$M = K^{-1}$$

Both conveniently implemented by deriving HMC specifying M^{-1}.

Transition operators for sampling latent variables

- Manifold MCMC - Girolami and Calderhead (2011)
  Simplified Manifold MALA (SMMALA) gradient and curvature information
- ELL-SS - Murray et al. (2010)
  Elliptical Slice Sampling adaptation of slice sampling to the sampling of latent variables in GPMs

Transition operators for sampling par

- Metropolis-Hastings (MH) random walk
- Hybrid Monte Carlo (HMC) gradient information
- Simplified version of Manifold MALA (SMMALA) gradient and curvature information

Convergence analysis and efficiency of MCMC algorithms

- Models employ a RBF ARD covariance
- Convergence speed measured using $\hat{R}$ statistics. To visualize convergence between 1000 and 20000 iterations

▯ < 1.1 < ▂ < 1.3 < ▅ < 2 < ▮

- Efficiency measured through (the minimum across variables) Effective Sample Size (ESS)

Results

Table: Comparison of transition operators to sample $\mathrm{latent}|\mathrm{par}, \mathrm{data}$ for data generated from logistic regression GPMs. T is the number of MCMC iterations

	$n = 400$				
	$d = 2$		$d = 10$		
	ESS	$\hat{R}$	ESS	$\hat{R}$	$\#O(n^3)$
MH v1	22 (7)		8 (1)		1
MH v2	67 (17)		30 (2)		1
SMMALA	457 (212)		48 (5)		$T + 1$
ELL-SS	104 (25)		50 (2)		1
HMC v1	1352 (380)		2962 (155)		3
HMC v2	1566 (342)		2995 (129)		1

Results

Table: SA - Comparison of transition operators to sample $\mathrm{par}|\mathrm{latent}$. In HMC $\bar{\lambda}$ is the average number of leapfrog steps.

	$n = 400$				
	$d = 2$		$d = 10$		
	ESS	$\hat{R}$	ESS	$\hat{R}$	$\#O(n^3)$
MH	2124 (125)		77 (33)		T
HMC	12556 (661)		293 (137)		$T(\bar{\lambda} + 1)$
SMMALA	10241 (2672)		47 (17)		$T(d + 2)$

Results

Table: AA - Comparison of transition operators to sample par|data, ancillary for data generated from logistic regression GPMs.

	$n = 400$				
	$d = 2$		$d = 10$		
	ESS	$\hat{R}$	ESS	$\hat{R}$	$\#O(n^3)$
MH	512 (177)	▯▯▯▯▯	56 (11)	▯▯▯▯▯	T
HMC	2666 (973)	▯▯▯▯▯	223 (39)	▄▄▯▯▯	$T(d\bar{\lambda} + 1)$
SMMALA	6877 (1584)	▯▯▯▯▯	47 (21)	▃▯▯▯▯	$T(d + 1)$

Results

Table: Comparison of different strategies to sample $\mathrm{latent}, \mathrm{par}|\mathrm{data}$ in four UCI data sets modeled using logistic regression GPMs.

	Pima		Wisconsin		SPECT		Ionosphere	
	$n = 768, d = 8$		$n = 683, d = 9$		$n = 80, d = 22$		$n = 351, d = 34$	
	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$	ESS	$\hat{R}$
AA	34 (4)		42 (15)		99 (18)		12 (5)	
ASIS	35 (8)		47 (11)		215 (23)		24 (8)	
KHR	153 (14)		20 (10)		101 (16)		2 (2)	
SA	5 (2)		7 (3)		97 (12)		11 (7)	
SURR	76 (10)		25 (14)		84 (14)		9 (4)	

Conclusions

- Sampling from the posterior of latent variables can be efficiently done in a number of ways (scaled HMC, ELL-SS)
- AA scheme with par sampled using the MH algorithm seems a good compromise between efficiency and cost
- Sampling efficiency is sometimes less than 1% for the best sampling strategy
- Fully automated MCMC for GPMs still an open problem