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

M. Filippone[1], M. Zhong[2], M. Girolami[3]

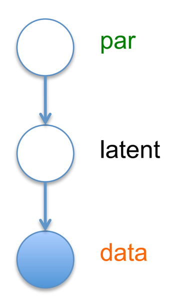
1 - University of Glasgow, UK. email: maurizio.filippone@glasgow.ac.uk 2 - University of Edinburgh, UK. 3 - University College London, UK.

Scope of this work

Gaussian Process models (GPMs) are extensively used in data analysis given their flexible modeling capabilities and interpretability. The fully Bayesian treatment of GP models is analytically intractable, and therefore it is necessary to resort to approximations. This work focuses on Markov chain Monte Carlo (MCMC) inference techniques. The hierarchical structure of GPMs and the large dimensionality of parameter and latent variable spaces pose serious challenges to the development of efficient MCMC methods for GPMs. The work employs strategies based on efficient parameterizations and efficient proposal mechanisms and compares them on simulated and real data on the basis of convergence speed, sampling efficiency, and computational cost.

Gaussian Process Models - GPMs

- GPMs:

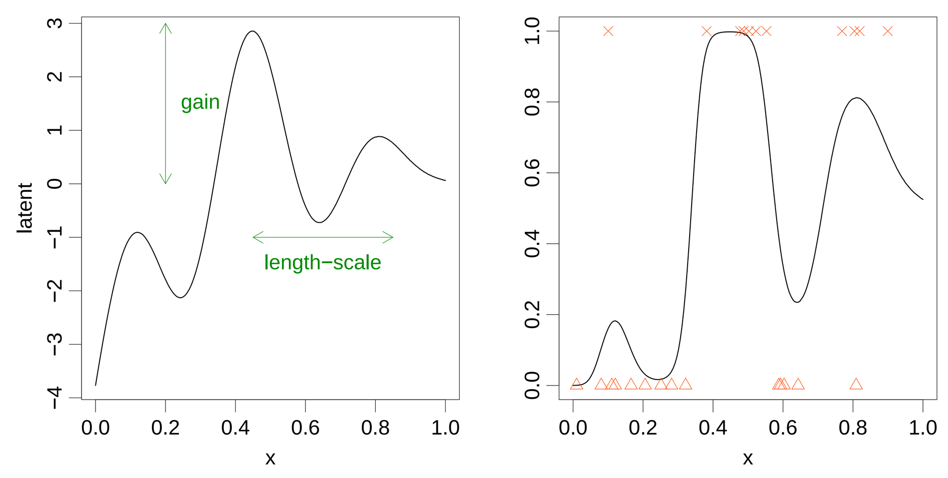


$p(\text{par})$

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

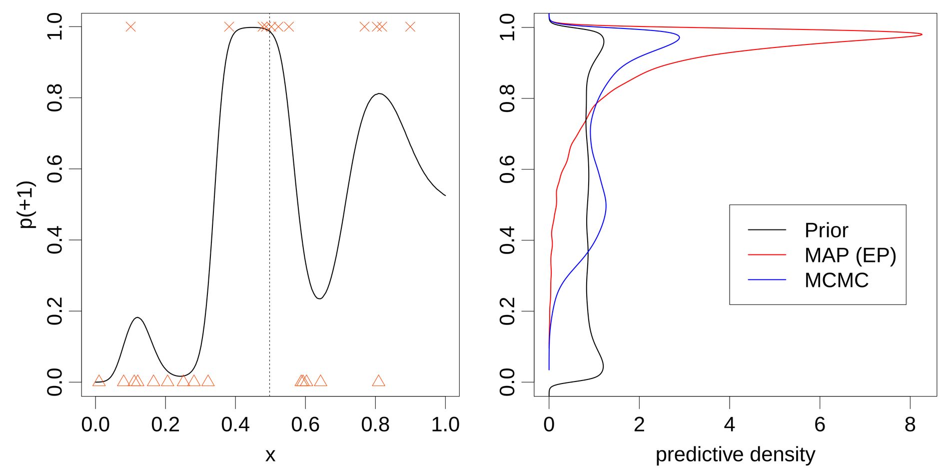
$p(\text{data}|\text{latent})$

- GPM - Probit regression example:



Why MCMC?

- Comparison MCMC vs Maximum a Posteriori (MAP) using EP



MAP solution **does not** account for the **uncertainty** in par! Also:

- Approximate marginal likelihood might be inaccurate
- Quadrature can't be employed if par is large dimensional

References

[1] O. F. Christensen, G. O. Roberts, and M. Sköld. Robust Markov chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 15(1):1–17, Mar. 2006.

[2] M. Filippone, A. F. Marquand, C. R. V. Blain, S. C. R. Williams, J. Mourão-Miranda, M. Girolami. Probabilistic Prediction of Neurological Disorders with a Statistical Assessment of Neuroimaging Data Modalities. *Annals of Applied Statistics*, 6(4):1883–1905, 2012.

[3] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

[4] L. Knorr-Held and H. Rue. On Block Updating in Markov Random Field Models for Disease Mapping. *Scandinavian Journal of Statistics*, 29(4):597–614, 2002.

[5] I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In *NIPS*, 1732–1740, 2010.

[6] I. Murray, R. P. Adams, and D. J. C. MacKay. Elliptical slice sampling. *JMLR - Proceedings Track*, 9:541–548, 2010.

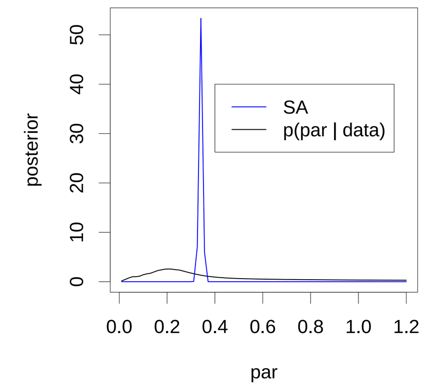
[7] R. M. Neal. Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics*, 6:475–501, 1999.

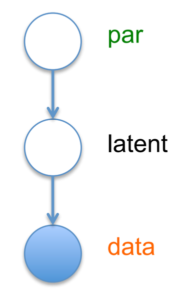
[8] Y. Yu and X.-L. Meng. To Center or Not to Center: That Is Not the Question–An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

Challenges in employing MCMC for inference in GPMs

Obvious choice: Sufficient Augmentation (SA) scheme - alternate between:

- Drawing from $p(\text{latent}|\text{par}, \text{data})$
- Drawing from $p(\text{par}|\text{latent})$ **bad idea!**





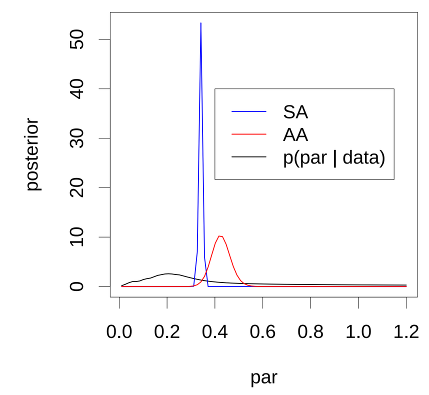
- Need to employ Metropolis within Gibbs steps - waste of computations when rejecting
- Updates of par cost $O(n^3)$
- par can be large dimensional, e.g., Automatic Relevance Determination (ARD) covariance
- There are n latent variables (as many as the number of observations)

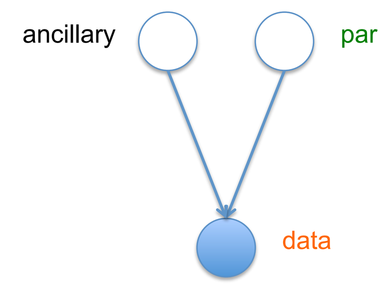
Mitigating coupling effect through reparameterization

Ancillary Augmentation (AA) scheme - reparametrization using GP prior covariance K:

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

- 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 $\text{latent}'|\text{par}', \text{data}$ and then **jointly** Accept/Reject
- ASIS - Interweave AA and SA as in Yu and Meng (2011)

Transition operators for latent variables

- Scaled Metropolis-Hastings (MH) proposals - Neal (1999) - draw $z \sim \mathcal{N}(0, K)$
  - MH v1:
  $$\text{latent}' = \text{latent} + \alpha\, z$$
  - MH v2:
  $$\text{latent}' = \sqrt{1-\alpha^2}\,\text{latent} + \alpha\, z$$
- 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 are conveniently implemented by deriving HMC specifying M^{-1}
- Manifold MCMC - Girolami and Calderhead (2011)
  Simplified Manifold MALA (SMMALA) gradient and curvature information
- Elliptical Slice Sampling (ELL-SS) (Murray et al. 2010) is an adaptation of slice sampling to sample latent variables in GPMs

Transition operators for parameters par

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

Results

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

$$< 1.1 < \quad < 1.3 < \quad < 2 <$$

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

Table 1: Comparison of transition operators to sample $\text{latent}|\text{par}, \text{data}$ 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

Table 2: SA - Comparison of transition operators to sample $\text{par}|\text{latent}$. 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)		$T(d+2)$

Table 3: AA - Comparison of transition operators to sample $\text{par}|\text{data}, \text{ancillary}$ 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)		$T(d+1)$

Table 4: Comparison of different strategies to sample $\text{latent}, \text{par}|\text{data}$ in four UCI data sets modeled using logistic regression GPMs.

	Pima $n = 768, d = 8$		Wisconsin $n = 683, d = 9$		SPECT $n = 80, d = 22$		Ionosphere $n = 351, d = 34$	
	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