Posterior Inference in Latent Gaussian Models Using Manifold MCMC Methods

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

1 - University of Glasgow, UK. email: maurizio@dcs.gla.ac.uk 2 - University College London, UK. email: girolami@stats.ucl.ac.uk

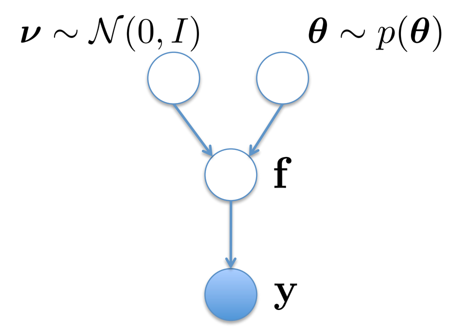
Scope of this work

In this work, we study the inference problem in latent Gaussian models. Efficiently sampling from the posterior distribution of the latent process and hyper-parameters is complex because of their strong coupling. We consider a set of recently proposed MCMC methods based on the natural geometry of the underlying statistical model to achieve efficient sampling.

Latent Gaussian Models - LGMs

- Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n covariates $\mathbf{x}_i \in \mathbb{R}^d$, associated with observed responses $\mathbf{y} = y_1, \ldots, y_n$.
- Let k be the covariance function parameterized by hyperparameters $\boldsymbol{\theta}$
- Consider the following general form of latent Gaussian models to model the generative process of the observed $\mathbf{y}$ (given X).

$p(\boldsymbol{\theta})$	prior $\boldsymbol{\theta}$
$K = LL^T$	covariance matrix
$p(\boldsymbol{\nu}) \sim \mathcal{N}(0, I)$	whitened latent
$\mathbf{f} = L\boldsymbol{\nu}$	transformation
$\Downarrow$	
$p(\mathbf{f}\vert\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}\vert\mathbf{0}, K)$	prior latent $\mathbf{f}$
$p(\mathbf{y}\vert\mathbf{f}) = \mathcal{E}(\mathbf{y}\vert\boldsymbol{\zeta}(\mathbf{f}))$	likelihood



- Distribution of the observed random variables $\mathbf{y}$: exponential family $\mathcal{E}$ with natural parameters given by a transformation of the latent variables $\boldsymbol{\zeta}(\mathbf{f})$.

$$\mathcal{E}(\mathbf{y}|\boldsymbol{\zeta}(\mathbf{f})) = h(\mathbf{y})g(\boldsymbol{\zeta}(\mathbf{f}))\exp(\boldsymbol{\zeta}(\mathbf{f})^{\mathrm{T}}\mathbf{u}(\mathbf{y}))$$

- The inference problem amounts in obtaining the posterior distribution over the parameters and use it to compute the predictive distribution

$$p(y_*|\mathbf{y}) = \int\int\int p(y_*|f_*)p(f_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f},\boldsymbol{\theta}|\mathbf{y})df_*d\mathbf{f}d\boldsymbol{\theta}$$

which is intractable. In this work we apply Markov Chains Monte Carlo (MCMC) methods to integrate out $\mathbf{f}$ and $\boldsymbol{\theta}$.

Manifold MCMC

- Manifold MCMC methods make use of the natural geometry of the underlying statistical model to achieve efficient sampling.
- Key quantities in information geometry are the Fisher Information (FI) and the connection. Consider a statistical model $S = \{p(\mathbf{y}|\boldsymbol{\psi})|\boldsymbol{\psi} \in \Psi\}$ for the observed variables $\mathbf{y}$ with parameters $\boldsymbol{\psi}$. Under quite general conditions , S can be considered a C^{∞} manifold, and is called statistical manifold. Let $\mathcal{L} = \log[p(\mathbf{y}|\boldsymbol{\psi})]$; the FI matrix G of S at $\boldsymbol{\psi}$ is defined as:

$$G(\boldsymbol{\psi}) = \mathrm{E}_{p(\mathbf{y}|\boldsymbol{\psi})}\left[(\nabla_{\boldsymbol{\psi}}\mathcal{L})(\nabla_{\boldsymbol{\psi}}\mathcal{L})^{\mathrm{T}}\right]$$

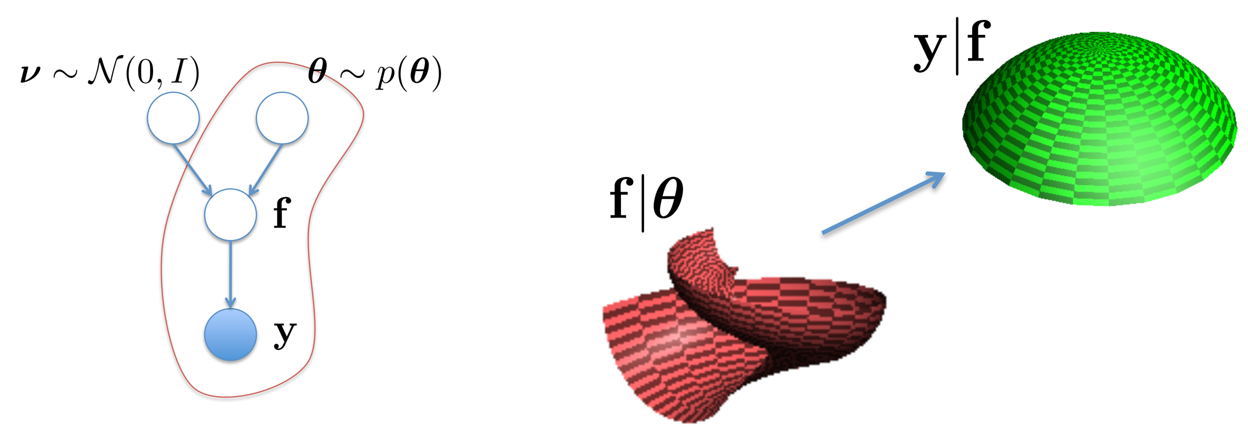
and is the natural metric on S. Christoffel symbols characterize connections on curved manifolds:

$$\Gamma^i_{kl} = \frac{1}{2}\sum_m \left(\frac{\partial g_{mk}}{\partial \psi_l} + \frac{\partial g_{ml}}{\partial \psi_k} - \frac{\partial g_{kl}}{\partial \psi_m}\right)$$

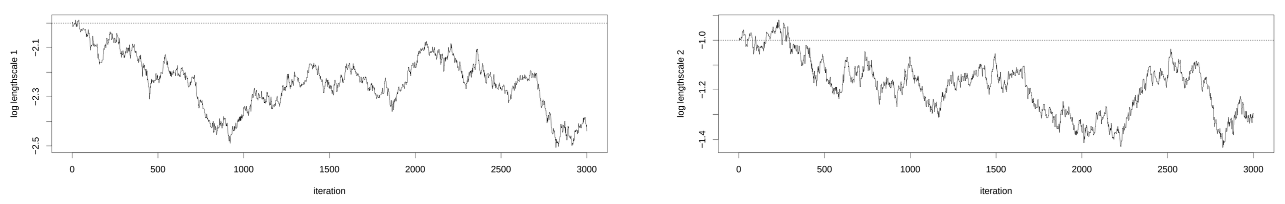
Centered vs Non-centered parameterization

Centered Parameterization

- Consider the sampling of $p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y})$. Efficiently sampling from this posterior distribution is complex because of the strong coupling between $\mathbf{f}$ and $\boldsymbol{\theta}$.

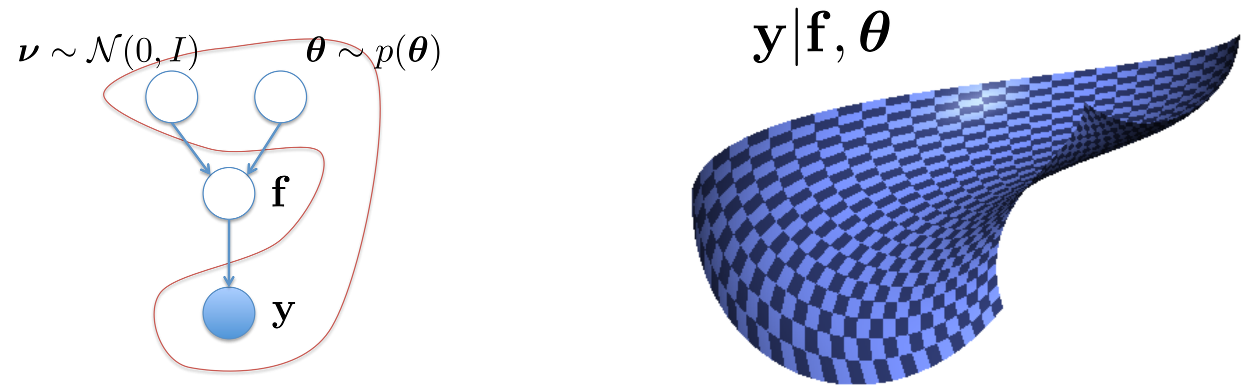


Poor mixing

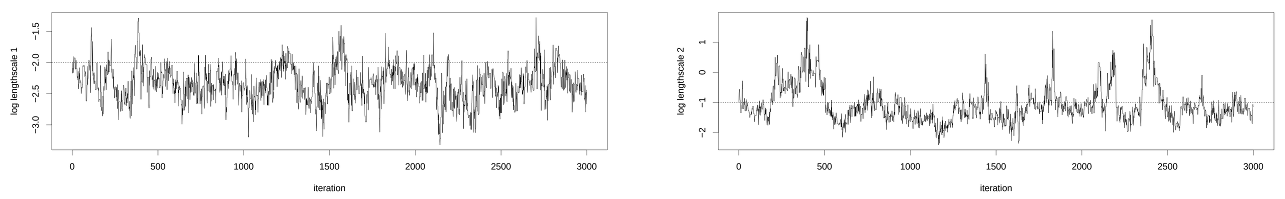


Non-Centered Parameterization

- If we focus on $p(\boldsymbol{\nu}, \boldsymbol{\theta}|\mathbf{y})$, instead, we can remove the prior correlation.



Better mixing!



Manifold methods for inference in LGMs

- We consider sampling from the posterior distribution of $p(\boldsymbol{\nu}, \boldsymbol{\theta}|\mathbf{y})$. The expression of the log-joint density reads:

$$\mathcal{L} = \log[h(\mathbf{y})] + \log[g(\boldsymbol{\zeta})] + \boldsymbol{\zeta}^{\mathrm{T}}\mathbf{u}(\mathbf{y}) \qquad \boldsymbol{\zeta} = \boldsymbol{\zeta}(L\boldsymbol{\nu}) = \boldsymbol{\zeta}(\mathbf{f})$$

- Define

$$R = \mathrm{E}_{\mathbf{y}}[\boldsymbol{\rho}\boldsymbol{\rho}^{\mathrm{T}}] \qquad \boldsymbol{\rho} = \boldsymbol{\rho}(\boldsymbol{\theta}, \boldsymbol{\nu}, \mathbf{y}) = \nabla_{\mathbf{f}}\boldsymbol{\zeta}(\mathbf{f})\left(\frac{\nabla_{\boldsymbol{\zeta}}g(\boldsymbol{\zeta})}{g(\boldsymbol{\zeta})} + \mathbf{u}(\mathbf{y})\right)$$

- The FI results in

$$G_{\boldsymbol{\nu},\boldsymbol{\nu}} = L^{\mathrm{T}}RL + I \qquad G_{\boldsymbol{\nu},\theta_i} = L^{\mathrm{T}}R\frac{\partial L}{\partial \theta_i}\boldsymbol{\nu}$$

$$G_{\theta_i,\theta_j} = \boldsymbol{\nu}^{\mathrm{T}}\frac{\partial L^{\mathrm{T}}}{\partial \theta_i}R\frac{\partial L}{\partial \theta_j}\boldsymbol{\nu} - \frac{\partial^2 \log[p(\boldsymbol{\theta})]}{\partial \theta_i \partial \theta_j} \qquad G = \begin{pmatrix} G_{\boldsymbol{\nu},\boldsymbol{\nu}} & G_{\boldsymbol{\nu},\boldsymbol{\theta}} \\ G_{\boldsymbol{\theta},\boldsymbol{\nu}} & G_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix}$$

Conclusions and Future work

- non-centered reparameterization improves sampling efficiency; however, it provides a method to decouple prior correlations, not posterior correlations
- results for few data show that manifold methods applied to the non-centered parameterization do not improve mixing; preliminary results on larger data sets show that manifold methods improve mixing
- can we use information geometry to decouple posterior correlations?

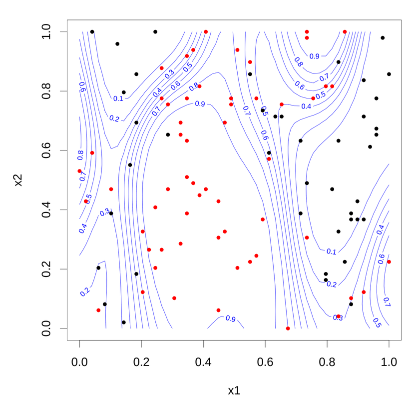
References

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

[2] M. Girolami and B. Calderhead. Riemann Manifold Langevin and Hamiltonian Monte Carlo. *Journal Of The Royal Statistical Society Series B* (with discussion), to appear.

[3] I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*. 2010.

Results - Logistic regression with Gaussian Process prior



- bivariate logistic regression with Gaussian Process prior
- results are for $n = 100$ data points
- effective sample size (ESS) is computed on 10000 samples from the posterior obtained after 1000 burn-in samples
- HMC achieves the best ESS but is the most expensive ($\nabla\mathcal{L}$ is expensive)