Bayesian inference in latent variable models and applications

Maurizio Filippone

Department of Statistical Science
University College London
maurizio@stats.ucl.ac.uk

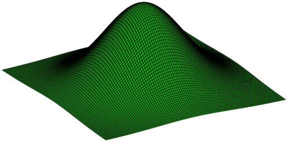
Joint work with V. Stathopoulos, Z. Mingjun, M. Girolami

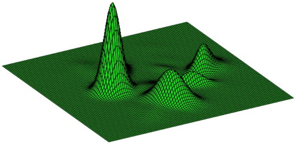
June 9th, 2011

Inference and model selection

- Parameters and data are viewed as random variables

$$p(\mathrm{Par}) \qquad \Rightarrow \qquad p(\mathrm{Par}|\mathrm{Data})$$



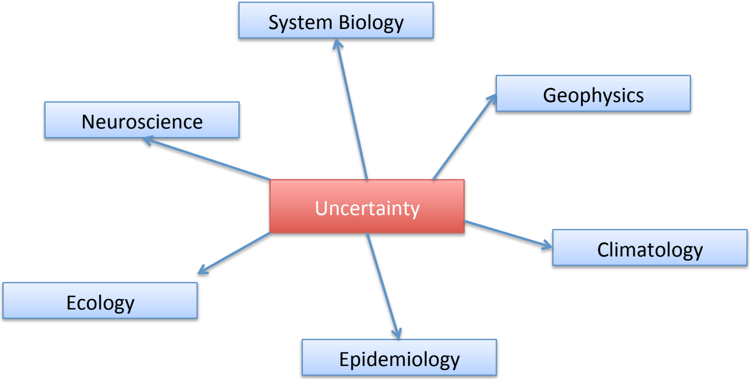


- Inference - Bayes theorem:

$$p(\mathrm{Par}|\mathrm{Data}) = \frac{p(\mathrm{Data}|\mathrm{Par})p(\mathrm{Par})}{\int p(\mathrm{Data}|\mathrm{Par})p(\mathrm{Par})d\mathrm{Par}}$$

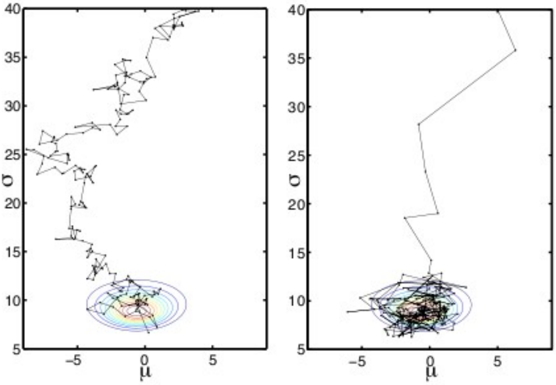
- Denominator: model evidence used for model comparison
- Usually analytically intractable!

Relevance of the problem



Markov chain Monte Carlo

- Explore the parameter space according to the density
- Set up a Markov chain with $p(\mathrm{Par}|\mathrm{Data})$ as invariant distribution



Markov chain Monte Carlo

Proposals can be based on:

- Random walk

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \varepsilon \qquad \varepsilon \sim \mathcal{N}(\varepsilon|\mathbf{0}, \Sigma)$$

- Langevin diffusion or Hamiltonian mechanics where the log-likelihood is viewed as a potential energy and a mass matrix allows different scalings across dimensions

How do we systematically tune the parameters of the proposal?

Manifold sampling

- Statistical model $S = \{p(\mathbf{y}|\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}$
- S can be considered a C^{∞} manifold (statistical manifold)
- Let $\mathcal{L} = \log[p(\mathbf{y}|\boldsymbol{\theta})]$
- Fisher Information (FI) - natural metric on S:

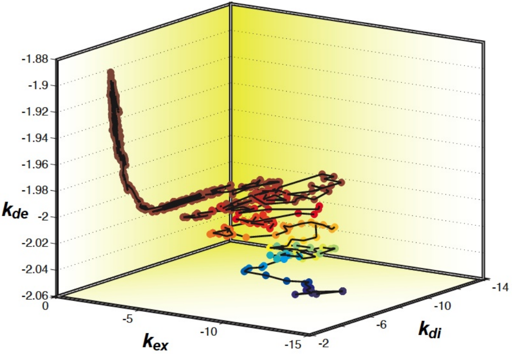
$$G(\boldsymbol{\theta}) = \mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\left[(\nabla_{\boldsymbol{\theta}}\mathcal{L})(\nabla_{\boldsymbol{\theta}}\mathcal{L})^{\mathrm{T}}\right] = -\mathrm{E}_{p(\mathbf{y}|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\mathcal{L}]$$

- Christoffel symbols characterize connections on curved manifolds:

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

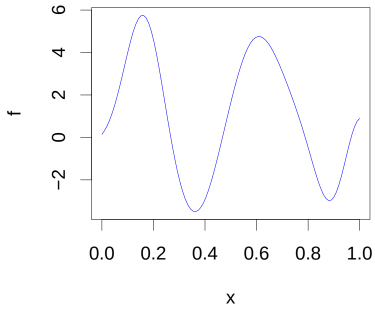
Manifold sampling

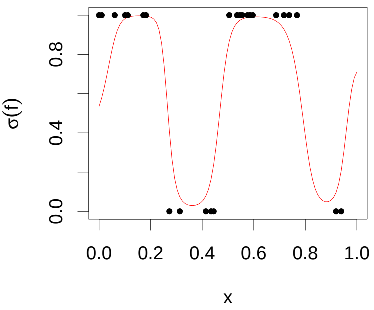
- Random walk, diffusion or Hamiltonian dynamic on the statistical manifold (Girolami and Calderhead 2010)



What about latent variable models?

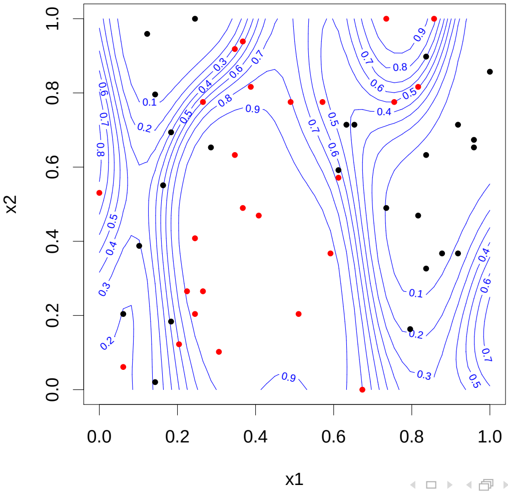
Example: Logistic regression





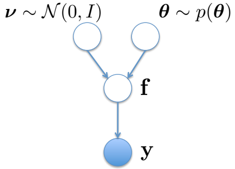
What about latent variable models?

Example: Logistic regression



Latent Gaussian Models - (LGMs)

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



Squared exponential covariance function

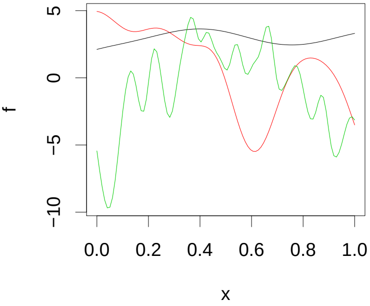
$$k(\mathbf{x}_i, \mathbf{x}_j|\theta) = \alpha \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} A(\mathbf{x}_i - \mathbf{x}_j)\right]$$

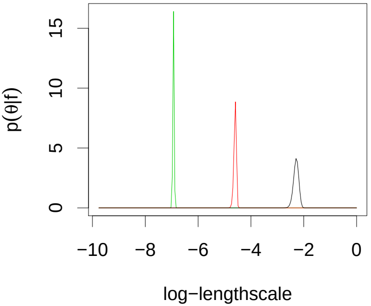
LGMs - Other examples

- Log-Gaussian Cox model (Møller et al. 1998)
- Gaussian copula process volatility model (Wilson and Ghahramani 2010)
- Gaussian processes for ordinal regression (Chu and Ghahramani 2005)

Model structure and efficient sampling

- The structure of the model poses a serious challenge to MCMC methods
- sampling $p(\mathbf{f}|\theta, \mathbf{y})$ and $p(\theta|\mathbf{f}, \mathbf{y})$ would be extremely inefficient



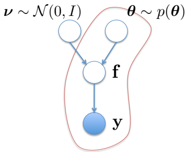


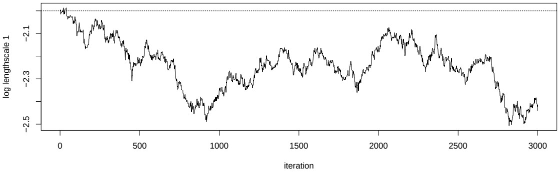
Additional Challenges

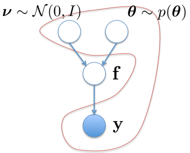
- computation of the likelihood is in $O(n^3)$ (same complexity for approximate methods)
- conditional distributions $p(\mathbf{f}|\theta, \mathbf{y})$ and $p(\theta|\mathbf{f}, \mathbf{y})$ are such that Gibbs sampler updates require a Metropolis acceptance step

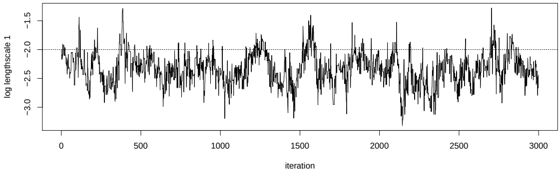
Model structure and efficient sampling

Centered vs non-centered parametrizations (Papaspiliopoulos et al. 2007)



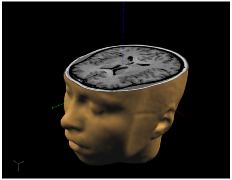






An application

- Infer subject's cognitive state from fMRI data
- Discriminate between cognitive states



- fully Bayesian non-linear discriminative method

Data

- Experiments reported here are with a single subject listening passively to vocal and non-vocal stimuli
- Preprocessing: time correction, spatial smoothing, masking, normalization, and voxel reduction (*t*-test)
- We have 200 samples with 4,436 covariates
- classes: 1 vocal and 0 non-vocal stimuli

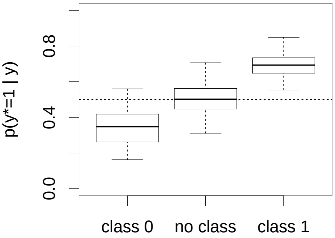
Results - Experimental setting

- classifier based on GP (GPC) (same cost for the two classes)
  - Gibbs sampler:
  - $\mathbf{f}|\theta, \mathbf{y}$ using manifold methods
  - $\theta|\mathbf{f}, \mathbf{y}$ using non-centered parametrization (i.e., $\theta|\nu, \mathbf{y}$)
- Support Vector Machines (SVM)
  - tested with both linear and radial basis function kernel
  - parameters (C and kernel bandwidth) were optimized using 10-fold cross validation
- GPC and non-linear SVMs use isotropic covariance/kernel functions

Results - Classification accuracy

Classification result using 4-fold validation

Method	Accuracy (std err)
SVM (lin)	75.5% (5.9%)
SVM (rbf)	76% (1.4%)
GPC	78.5% (3.8%)



- we can use the predictive distribution for finer decision rules
- by doing so we achieve 92.8% accuracy on 90 samples

Conclusions and ongoing work

- Recent advances in MCMC allow to approach the fully Bayesian treatment of several models commonly used in statistics
- Benefits of a fully Bayesian treatment in the descriptive power of the model (as demonstrated in the fMRI application)
- We are applying these ideas to mixture models and latent Dirichlet allocation models

Acknowledgements

This research is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/E052029/1.