On the Fully Bayesian Treatment of Latent Gaussian Models using Stochastic Simulations

Maurizio Filippone

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

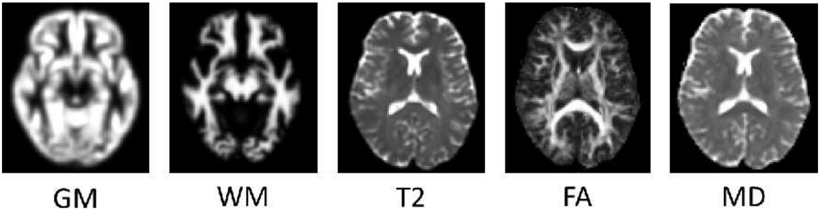
April 2nd, 2012

Outline of the talk

1. Motivating application
2. Latent Gaussian Models
3. Inference in Latent Gaussian Models
4. Results on Neuroimaging data

Parkinson syndromes data

- 62 subjects
- Early stage prediction of development of
  - Parkinson Syndromes
  - Multiple System Atrophy
  - Progressive Supranuclear Palsy
- Given neuroimages



Probabilistic Modeling

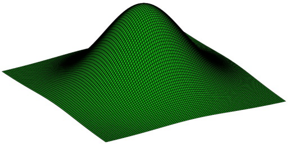
- which sources carry the most discriminative information?
- how do we assess which regions of the brain are responsible for the three diseases?
- how do we compare different models?
- how do we incorporate knowledge by experts?
- how can we attach confidence intervals to our predictions and parameter estimates?

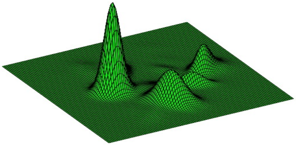
Probabilistic modeling offers an answer to these questions

Inference and model selection

- Parameters and data are viewed as random variables

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



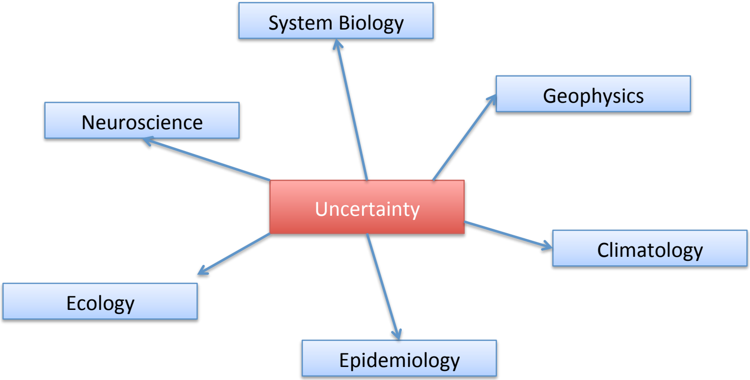


- Inference - Bayes theorem:

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

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

Relevance of the problem



Inference and predictions

- Predictions for new data Data_*

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

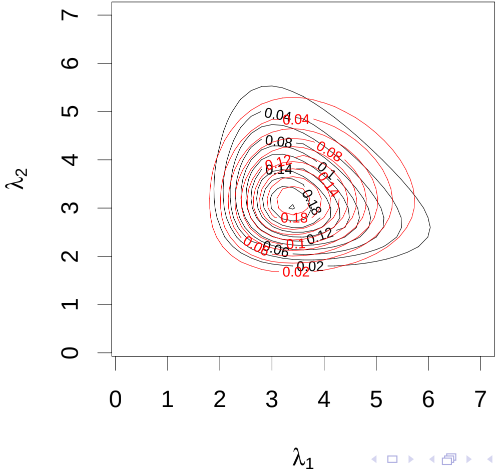
- requires the posterior distribution $p(\mathrm{Par} | \mathrm{Data})$

Approximate integration

- Approximations to solve analytically intractable integrals:
    - Deterministic: Laplace approximation, Variational Approximations (Expectation Propagation)
    - Stochastic: Markov chain Monte Carlo (MCMC)

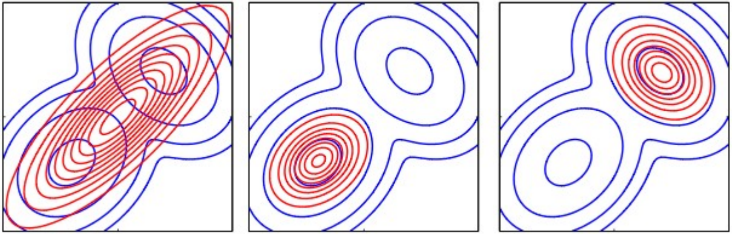
Deterministic approximations

- Variational Approximations (Expectation Propagation)



Deterministic approximations

- Multimodalities can be a problem



Stochastic approximations - Monte Carlo integration

- Predictions for new data Data_* is an expectation

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

- Monte Carlo approximation:

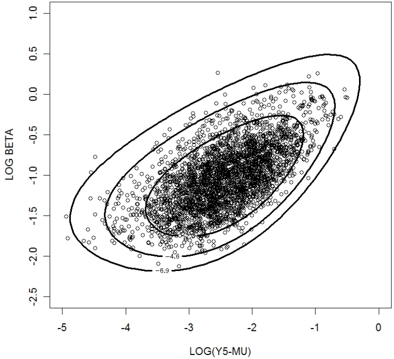
$$\mathrm{E}[f(x)] = \int f(x)p(x)dx \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

with x_i drawn from $p(x)$

- the variance of $\mathrm{E}[f(x)] \to 0$ in $O(1/N)$
- this requires independence of the x_i

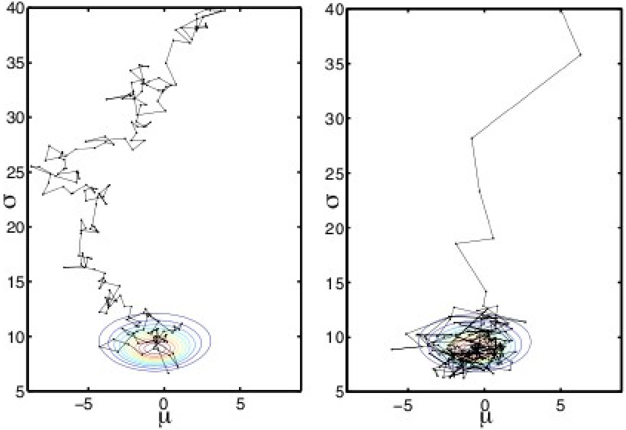
Stochastic approximations - MCMC

- Explore the parameter space according to the density



Stochastic approximations - MCMC

- Exploration of the space according to the density



Stochastic approximations - MCMC

- MCMC needs the density up to a normalization constant
- Random walk sampler - accept a proposal with probability

$$\min\left(1, \frac{p(\mathrm{Par}'|\mathrm{Data})}{p(\mathrm{Par}|\mathrm{Data})}\right)$$

- by Bayes' theorem

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

- Therefore:

$$\min\left(1, \frac{p(\mathrm{Data}|\mathrm{Par}')p(\mathrm{Par}')}{p(\mathrm{Data}|\mathrm{Par})p(\mathrm{Par})}\right)$$

Stochastic approximations - Monte Carlo integration

Random walk can be inefficient, so why not use

- gradient information - Hybrid Monte Carlo (Neal 1995), Langevin diffusion (Roberts and Stramer 2002)
- curvature information (Fisher Information) - Manifold methods (Girolami and Calderhead 2011)

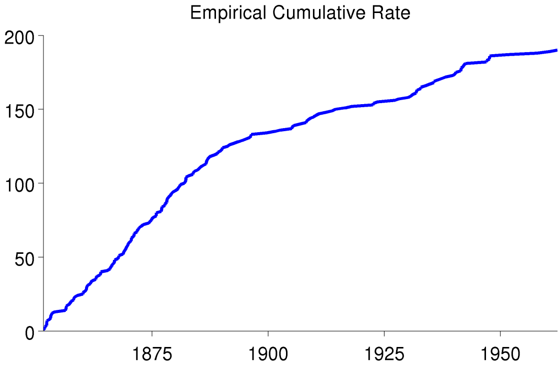
A class of hierarchical models

- Models can have more complex structures
- For example:

$$p(\text{Data}|\text{latent state}) \quad p(\text{latent state}|\text{Par}) \quad p(\text{Par})$$

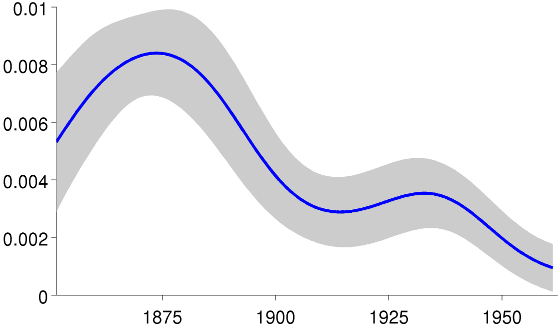
Coal mine disasters data

191 accidents between 1851 and 1962



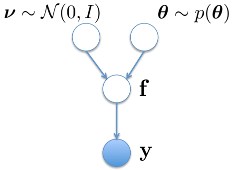
Coal mine disasters data

191 accidents between 1851 and 1962



Latent Gaussian Models - (LGM)

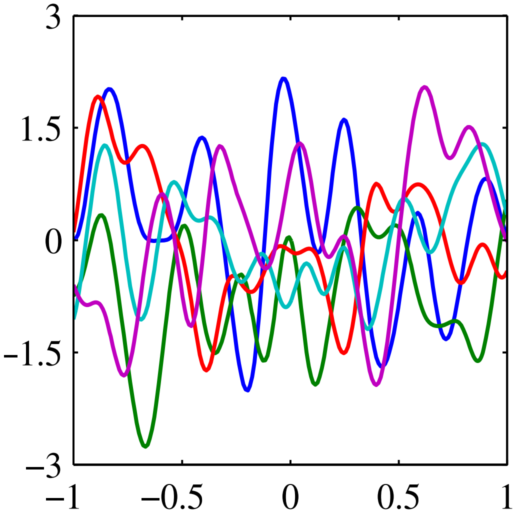
$p(\boldsymbol{\theta})$	prior $\boldsymbol{\theta}$
$p(\mathbf{f}\vert\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}\vert\mathbf{0}, K)$	prior latent $\mathbf{f}$
$p(\mathbf{y}\vert\mathbf{f}) = \mathcal{E}(\mathbf{y}\vert\zeta(\mathbf{f}))$	likelihood



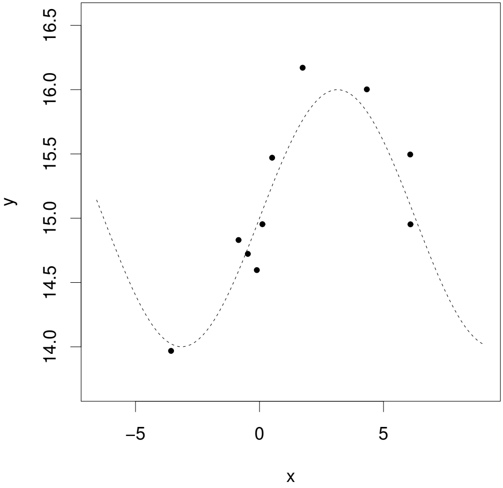
Squared exponential covariance function

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

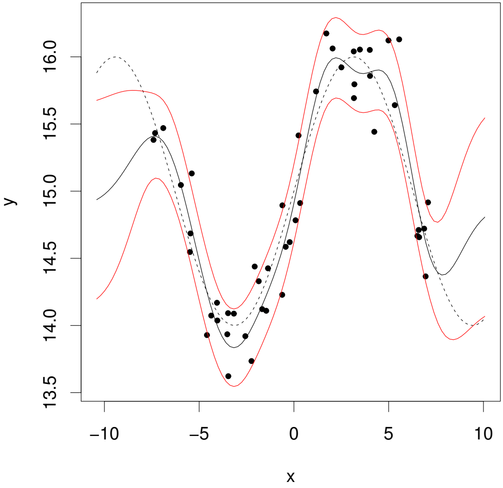
Example - Regression with Gaussian processes



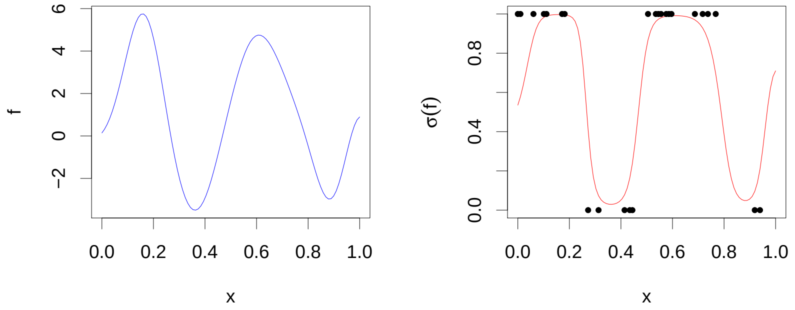
Example - Regression with Gaussian processes



Example - Regression with Gaussian processes



LGM - Logistic regression example



Latent Gaussian models - 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)

Challenges

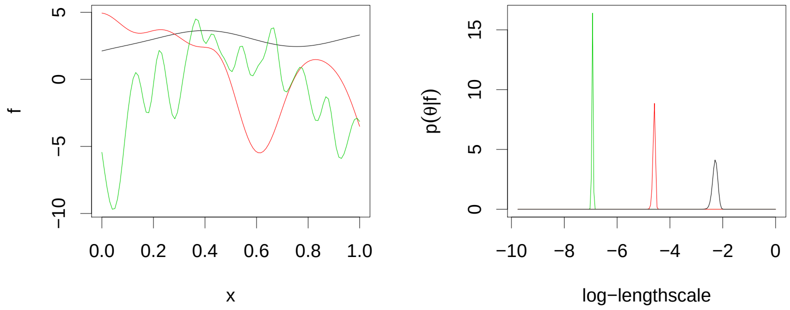
- computation of the likelihood is in $O(n^3)$ (same complexity for approximate methods) as the likelihood contains:

$$\log |K| \qquad \mathbf{f}^{\mathrm{T}} K^{-1} \mathbf{f}$$

- 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

The structure of the model poses a serious challenge to MCMC methods for efficiently sampling from posterior distributions



Reparametrization for MCMC in hierarchical models

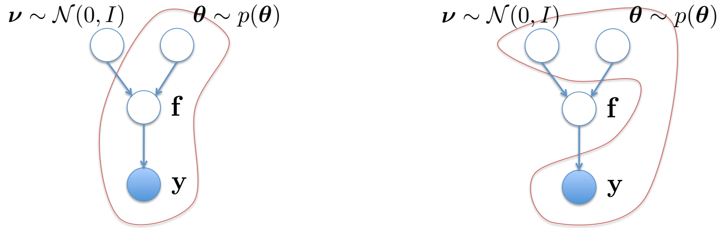
- Efficiency of parametrizations in strong/weak data limits (Papaspiliopoulos et al. 2007, Murray and Adams 2010)
- Yu and Meng (2011) - combining different parametrizations to boost MCMC efficiency

ASIS for LGMs

- Reparametrization for LGMs:

$$\mathbf{f}|\mathbf{y}, \boldsymbol{\theta} \quad \longrightarrow \quad \boldsymbol{\theta}|\mathbf{y}, \mathbf{f} \quad \longrightarrow \quad \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\nu}$$

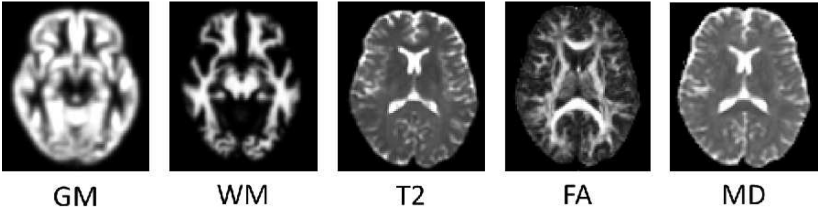
- Schematic view



Parkinson syndromes data

Multiclass classification with multiple sources

- Parkinson Syndromes
- Multiple System Atrophy
- Progressive Supranuclear Palsy
- Healthy controls



Results - Parkinson syndromes data

- latent variables $f_c(x)$ with GP prior with covariance

$$\mathrm{cov}(f_c(x_1), f_c(x_2)) = \sum_{s=1}^{q} w_{cs} C_s(x_1, x_2)$$

- Multinomial likelihood

$$p(\mathrm{disease} = c | \mathrm{Sources}) = \frac{\exp(f_c(x))}{\sum_{r=1}^{m} \exp(f_r(x))}$$

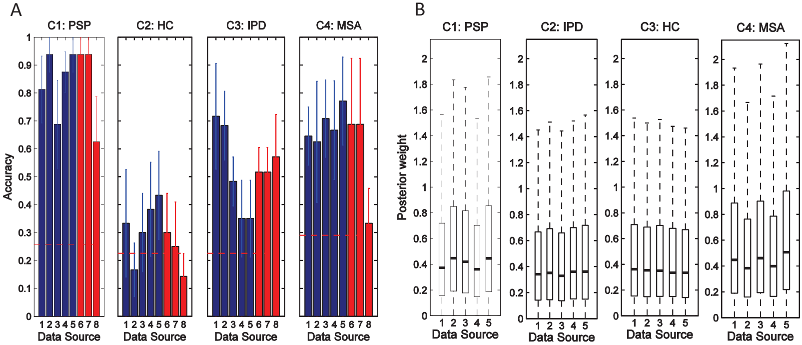
- this problem is aka Multiple Kernel Learning

Parkinson syndromes data

Table: Predictive accuracy (multi-source classifier)

Method	Accuracy
GP classifier	0.598
SimpleMKL	0.418

Parkinson syndromes data - multi source



Parkinson syndromes data

Analysis of brain regions

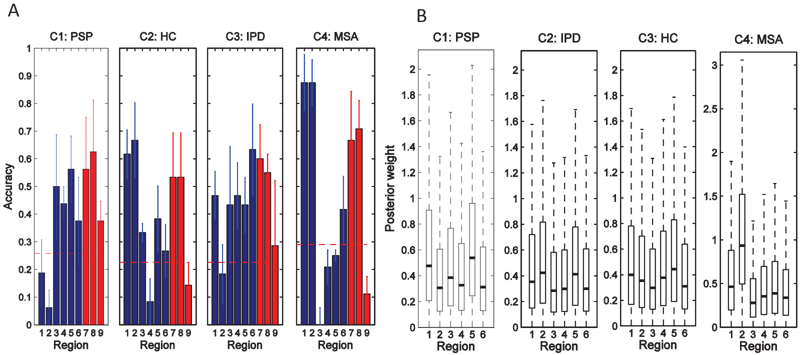
- for this analysis we used only the GM data modality
- we used an anatomical template as in Shattuck et al. 2008 to parcellate the GM images into:
  - brainstem
  - bilateral cerebellum
  - bilateral caudate
  - bilateral middle occipital gyrus
  - bilateral putamen
  - all other regions

Parkinson syndromes data

Table: Predictive accuracy (multi-region classifier)

Method	Accuracy
GP classifier	0.614
SimpleMKL	0.229

Parkinson syndromes data - multi region



Conclusions and ongoing work

- benefits of a fully Bayesian treatment in the descriptive power of the model
- recent advances in inference in hierarchical models allow to make a step forward into fully Bayesian inference in LGMs
- complexity is in $O(n^3)$ - same for deterministic approximations

References

[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. To appear.*

[2] M. Filippone, M. Zhong, and M. Girolami. **On the fully Bayesian treatment of latent Gaussian models using stochastic simulations**. Technical Report TR-2012-329, School of Computing Science, University of Glasgow, February 2012.

References

Thank you!

Questions?

Vocal/Non vocal 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 (number of voxels remaining after the t-test)
- classes: 1 vocal and 0 non-vocal stimuli

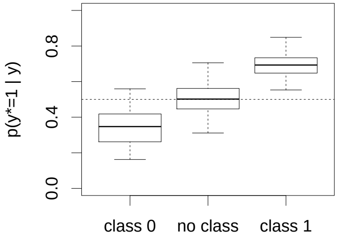
Results - Experimental setting

- binary Logistic Regression with GP prior
- 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