Bayesian Data Integration with Gaussian Process Priors: Combining Classifiers

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

UNIVERSITY
of
GLASGOW

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification

Overview

UNIVERSITY *of* GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification
- Regression with Gaussian process priors over functions

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification
- Regression with Gaussian process priors over functions
- Classification with Gaussian processes

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification
- Regression with Gaussian process priors over functions
- Classification with Gaussian processes
- Enabling Variational inference via multinomial-probit likelihood

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification
- Regression with Gaussian process priors over functions
- Classification with Gaussian processes
- Enabling Variational inference via multinomial-probit likelihood
- Data integration with composite covariance functions

Overview

UNIVERSITY
of
GLASGOW

- Motivation for Data Integration in Classification setting
- Adopting Bayesian inference for non-parametric classification
- Regression with Gaussian process priors over functions
- Classification with Gaussian processes
- Enabling Variational inference via multinomial-probit likelihood
- Data integration with composite covariance functions
- Experiments, conclusions & ongoing work

Data Integration

UNIVERSITY *of* GLASGOW

- Classifier combination schemes observed to outperform single best classifier

Data Integration

UNIVERSITY of GLASGOW

- Classifier combination schemes observed to outperform single best classifier
- Availability of multiple independent feature representations and structured heterogeneous data

Data Integration

UNIVERSITY *of* GLASGOW

- Classifier combination schemes observed to outperform single best classifier
- Availability of multiple independent feature representations and structured heterogeneous data
- Integrating & combining diverse sources of data in classification setting - empirical evidence suggests enhanced performance over use of single best data source

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)

Data Integration

UNIVERSITY
of
GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes

Data Integration

UNIVERSITY
of
GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations
- 64 PCA coefficients

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations
- 64 PCA coefficients
- 240 pixel averages in 2 x 3 windows

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations
- 64 PCA coefficients
- 240 pixel averages in 2 x 3 windows
- 47 Zernike moments

Data Integration

UNIVERSITY *of* GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations
- 64 PCA coefficients
- 240 pixel averages in 2 x 3 windows
- 47 Zernike moments
- 6 morphological features

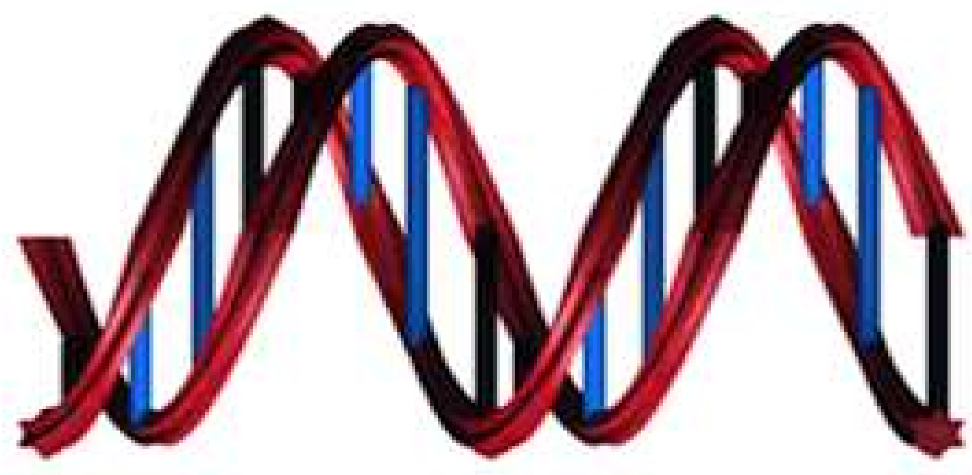
Data Integration

UNIVERSITY
of
GLASGOW

- Classification of handwritten digits (Duin *et al*)
- Each digit represented by six independent feature sets
- 76 Fourier coefficients of the character shapes
- 216 profile correlations
- 64 PCA coefficients
- 240 pixel averages in 2 x 3 windows
- 47 Zernike moments
- 6 morphological features
- Possible (not advisable) to embed within common feature space

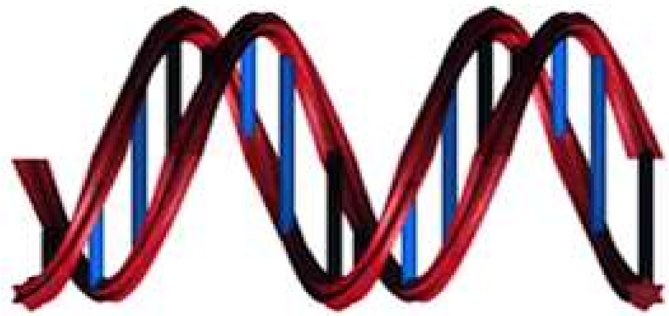
Data Integration

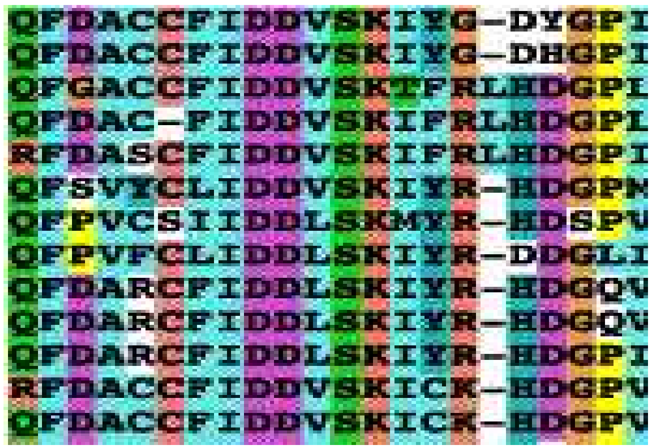
- Multiple heterogeneous representations of a gene



Data Integration

- Amino Acid sequence and sequence specific features

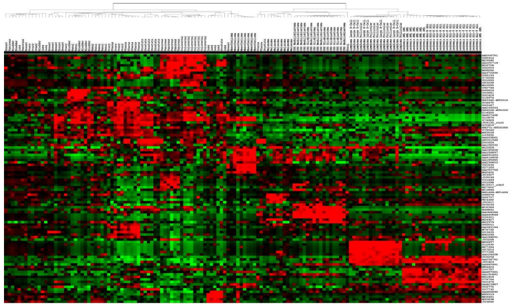


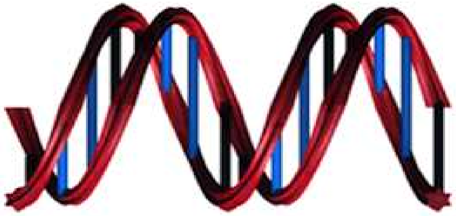


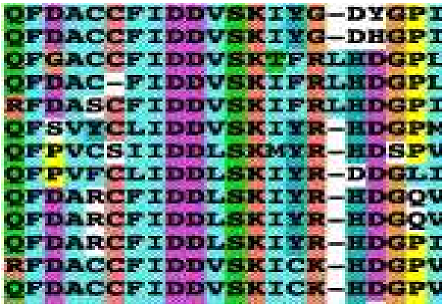
Data Integration

UNIVERSITY of GLASGOW

- Measurements of mRNA from gene in various cellular conditions

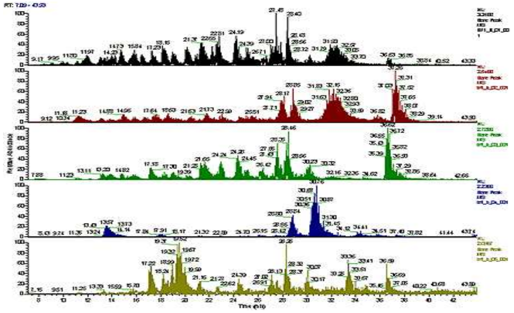


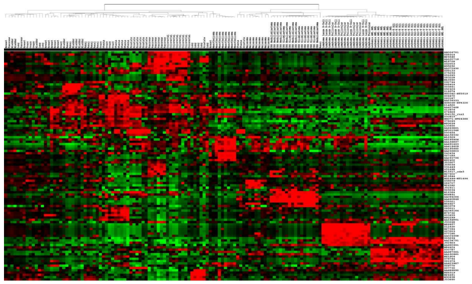


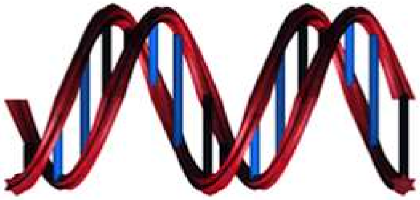


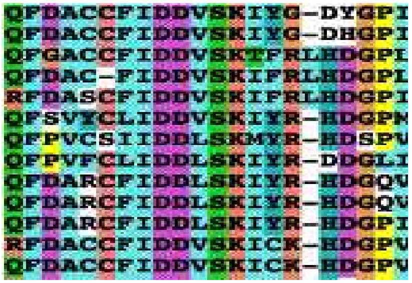
Data Integration

- Profile of peptides for protein gene codes



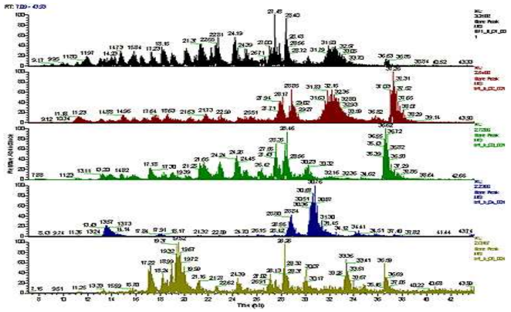
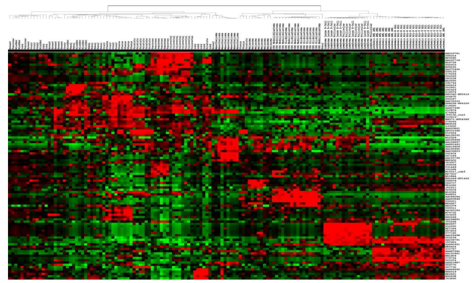
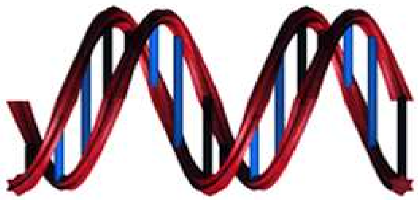


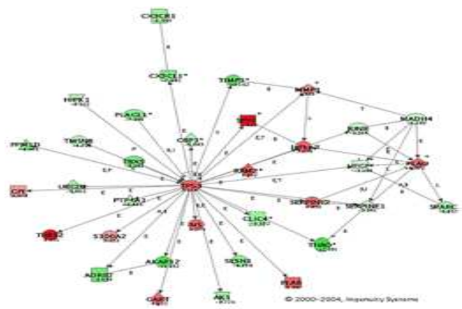




Data Integration

- Network of gene interactions

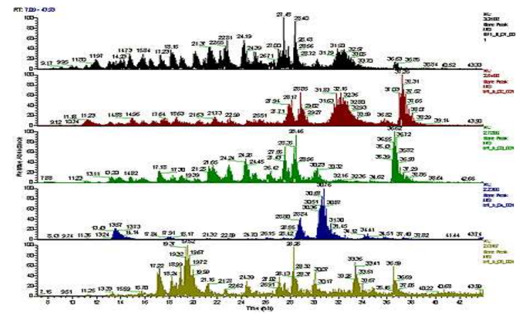




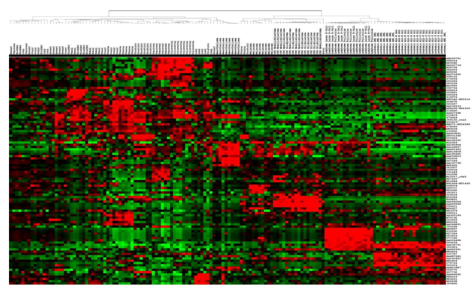



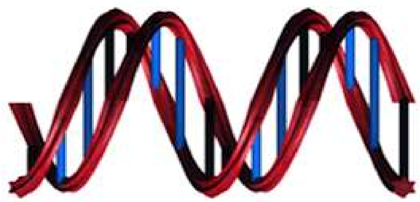
Data Integration

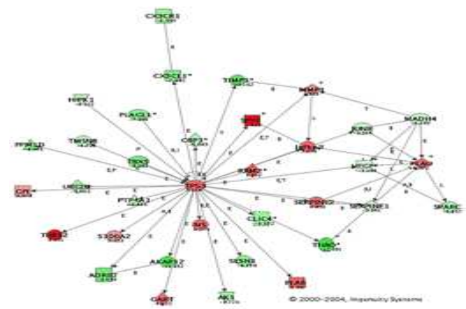
UNIVERSITY
of
GLASGOW

- Network of gene interactions









- Multiple heterogeneous data representations available for exploitation in classification problems

Data Integration

- Combination schemes for probabilistic classifiers studied by Kittler *et al*

Data Integration

UNIVERSITY *of* GLASGOW

- Combination schemes for probabilistic classifiers studied by Kittler *et al*
- For each of $\mathcal{J}$ data & feature representations of object X obtain class posterior probabilities

$$P(t = C|\mathcal{F}_1(X)) \cdots P(t = C|\mathcal{F}_{\mathcal{J}}(X))$$

from each of $\mathcal{J}$ independent classifiers

Data Integration

UNIVERSITY of GLASGOW

- Combination schemes for probabilistic classifiers studied by Kittler *et al*
- For each of $\mathcal{J}$ data & feature representations of object X obtain class posterior probabilities

$$P(t = C|\mathcal{F}_1(X)) \cdots P(t = C|\mathcal{F}_{\mathcal{J}}(X))$$

from each of $\mathcal{J}$ independent classifiers

- Employ individual posteriors to approximate joint probability

$$P(t = C|\mathcal{F}_1(X) \cdots \mathcal{F}_{\mathcal{J}}(X))$$

Data Integration

UNIVERSITY *of* GLASGOW

- Obtain sum and product combination rules

Data Integration

UNIVERSITY of GLASGOW

- Obtain sum and product combination rules
- Product combination

$$P(t = C|\mathcal{F}_1(X) \cdots \mathcal{F}_{\mathcal{J}}(X)) \approx \frac{\prod_j P(t = C|\mathcal{F}_j(X))}{\sum_{C'} \prod_{j'} P(t = C'|\mathcal{F}_{j'}(X))}$$

Data Integration

UNIVERSITY of GLASGOW

- Obtain sum and product combination rules
- Product combination

$$P(t = C|\mathcal{F}_1(X) \cdots \mathcal{F}_{\mathcal{J}}(X)) \approx \frac{\prod_j P(t = C|\mathcal{F}_j(X))}{\sum_{C'} \prod_{j'} P(t = C'|\mathcal{F}_{j'}(X))}$$

- Sum combination

$$P(t = C|\mathcal{F}_1(X) \cdots \mathcal{F}_{\mathcal{J}}(X)) \approx \frac{1}{\mathcal{J}} \sum_j P(t = C|\mathcal{F}_j(X))$$

Data Integration

UNIVERSITY of GLASGOW

- Obtain sum and product combination rules
- Product combination

$$P(t = C|\mathcal{F}_1(X)\cdots\mathcal{F}_{\mathcal{J}}(X)) \approx \frac{\prod_j P(t = C|\mathcal{F}_j(X))}{\sum_{C'}\prod_{j'} P(t = C'|\mathcal{F}_{j'}(X))}$$

- Sum combination

$$P(t = C|\mathcal{F}_1(X)\cdots\mathcal{F}_{\mathcal{J}}(X)) \approx \frac{1}{\mathcal{J}}\sum_j P(t = C|\mathcal{F}_j(X))$$

- Empirically observed to perform well on certain problems. Classifiers induced independently however desirable to induce joint classifier with statistical inference operating on all data jointly.

Data Integration

UNIVERSITY *of* GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.

Data Integration

UNIVERSITY of GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.
- Define kernel specific to each data-type and create linear combination $\mathcal{K}(X_m, X_n) = \sum_j \gamma_j \mathcal{K}_j(\mathcal{F}_j(X_m), \mathcal{F}_j(X_n))$, then employ in SVM

Data Integration

UNIVERSITY *of* GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.
- Define kernel specific to each data-type and create linear combination $\mathcal{K}(X_m, X_n) = \sum_j \gamma_j \mathcal{K}_j(\mathcal{F}_j(X_m), \mathcal{F}_j(X_n))$, then employ in SVM
- Objects with multiple representations
  - Proteins, *Lanckriet et al, 2004* - SDP & SVM

Data Integration

UNIVERSITY *of* GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.
- Define kernel specific to each data-type and create linear combination $\mathcal{K}(X_m, X_n) = \sum_j \gamma_j \mathcal{K}_j(\mathcal{F}_j(X_m), \mathcal{F}_j(X_n))$, then employ in SVM
- Objects with multiple representations
  - Proteins, *Lanckriet et al, 2004* - SDP & SVM
  - Protein-Protein interactions, *Ben-Hur, Noble, 2005*

Data Integration

UNIVERSITY of GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.
- Define kernel specific to each data-type and create linear combination $\mathcal{K}(X_m, X_n) = \sum_j \gamma_j \mathcal{K}_j(\mathcal{F}_j(X_m), \mathcal{F}_j(X_n))$, then employ in SVM
- Objects with multiple representations
  - Proteins, *Lanckriet et al, 2004* - SDP & SVM
  - Protein-Protein interactions, *Ben-Hur, Noble, 2005*
  - Enzyme Networks, *Yamanishi, Vert, Kanehisa, 2005*

Data Integration

UNIVERSITY
of
GLASGOW

- Kernel based non-parametric classification e.g. Support Vector Machines provide appropriate embeddings for heterogeneous representations of objects.
- Define kernel specific to each data-type and create linear combination $\mathcal{K}(X_m, X_n) = \sum_j \gamma_j \mathcal{K}_j(\mathcal{F}_j(X_m), \mathcal{F}_j(X_n))$, then employ in SVM
- Objects with multiple representations
  - Proteins, *Lanckriet et al, 2004* - SDP & SVM
  - Protein-Protein interactions, *Ben-Hur, Noble, 2005*
  - Enzyme Networks, *Yamanishi, Vert, Kanehisa, 2005*
- Learning kernel weights γ_j employing Semi-Definite programming for SVM classification enables heterogeneous data integration

Bayesian Classification

UNIVERSITY of GLASGOW

- SVM defines a binary classifier, multiple classes requires heuristic multiple one-vs-one, one-vs-rest combinations of binary output coding or DAG's

Bayesian Classification

- SVM defines a binary classifier, multiple classes requires heuristic multiple one-vs-one, one-vs-rest combinations of binary output coding or DAG's
- SVM non-probabilistic though some form of probabilistic semantics can be obtained post-hoc

Bayesian Classification

UNIVERSITY *of* GLASGOW

- SVM defines a binary classifier, multiple classes requires heuristic multiple one-vs-one, one-vs-rest combinations of binary output coding or DAG's
- SVM non-probabilistic though some form of probabilistic semantics can be obtained post-hoc
- Inference over feature or data relevance spawns numerous methods e.g. SDP for kernel combinations (only in binary case)

Bayesian Classification

UNIVERSITY *of* GLASGOW

- SVM defines a binary classifier, multiple classes requires heuristic multiple one-vs-one, one-vs-rest combinations of binary output coding or DAG's
- SVM non-probabilistic though some form of probabilistic semantics can be obtained post-hoc
- Inference over feature or data relevance spawns numerous methods e.g. SDP for kernel combinations (only in binary case)
- Strength of non-parametric classification of SVM - kernel method enables heterogeneous data integration - wish to combine non-parametrics with probabilistic semantics

Bayesian Classification

UNIVERSITY of GLASGOW

- Probabilistic inference over class membership of objects desirable in many applications

Bayesian Classification

UNIVERSITY
of
GLASGOW

- Probabilistic inference over class membership of objects desirable in many applications
- For high-dimensional and structured heterogeneous data it may be required to provide an additional level of inference

Bayesian Classification

- Probabilistic inference over class membership of objects desirable in many applications
- For high-dimensional and structured heterogeneous data it may be required to provide an additional level of inference
- Success of non-parametric methods of classification (SVM) in many diverse applications

Bayesian Classification

UNIVERSITY of GLASGOW

- Probabilistic inference over class membership of objects desirable in many applications
- For high-dimensional and structured heterogeneous data it may be required to provide an additional level of inference
- Success of non-parametric methods of classification (SVM) in many diverse applications
- Adopting Gaussian Process priors provides consistent probabilistic framework for Bayesian inference for general non-parametric classification problems (multiple classes, feature weighting, data integration, kernel combinations) without recourse to ad-hockery

Gaussian Processes

UNIVERSITY
of
GLASGOW

- GP defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$

Gaussian Processes

UNIVERSITY of GLASGOW

- GP defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$
- Stochastic process defined by mean, $\mu(\mathbf{x}) = E\{f(\mathbf{x})\}$, and covariance $C(\mathbf{x}_i, \mathbf{x}_j) = E\{f(\mathbf{x}_i)f(\mathbf{x}_j)\}$ functions

Gaussian Processes

UNIVERSITY *of* GLASGOW

- GP defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$
- Stochastic process defined by mean, $\mu(\mathbf{x}) = E\{f(\mathbf{x})\}$, and covariance $C(\mathbf{x}_i, \mathbf{x}_j) = E\{f(\mathbf{x}_i)f(\mathbf{x}_j)\}$ functions
- $p(f)$ is a GP if for any finite subset of $\mathcal{X}$ the marginal distribution $p(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N))$ is multivariate Gaussian

Gaussian Processes

UNIVERSITY *of* GLASGOW

- GP defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$
- Stochastic process defined by mean, $\mu(\mathbf{x}) = E\{f(\mathbf{x})\}$, and covariance $C(\mathbf{x}_i, \mathbf{x}_j) = E\{f(\mathbf{x}_i)f(\mathbf{x}_j)\}$ functions
- $p(f)$ is a GP if for any finite subset of $\mathcal{X}$ the marginal distribution $p(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N))$ is multivariate Gaussian
- For N samples $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset \mathcal{X}$ then $\mathbf{f} = \{f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)\} \sim GP(\boldsymbol{\mu}, \mathbf{C}) = \mathcal{N}_{\mathbf{f}}(\boldsymbol{\mu}, \mathbf{C})$

Gaussian Processes

UNIVERSITY *of* GLASGOW

- GP defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}$
- Stochastic process defined by mean, $\mu(\mathbf{x}) = E\{f(\mathbf{x})\}$, and covariance $C(\mathbf{x}_i, \mathbf{x}_j) = E\{f(\mathbf{x}_i)f(\mathbf{x}_j)\}$ functions
- $p(f)$ is a GP if for any finite subset of $\mathcal{X}$ the marginal distribution $p(f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N))$ is multivariate Gaussian
- For N samples $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \subset \mathcal{X}$ then $\mathbf{f} = \{f(\mathbf{x}_1), \cdots, f(\mathbf{x}_N)\} \sim GP(\boldsymbol{\mu}, \mathbf{C}) = \mathcal{N}_{\mathbf{f}}(\boldsymbol{\mu}, \mathbf{C})$
- GP prior encodes knowledge or assumptions on functional class ('smooth', 'rough')

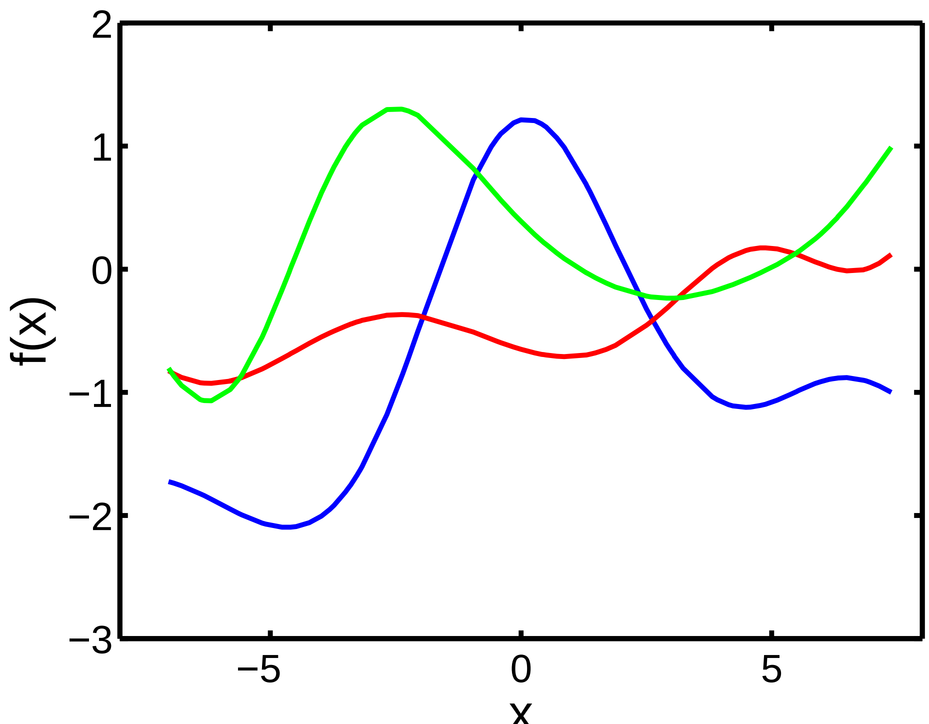
Gaussian Processes

UNIVERSITY *of* GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi|x_i - x_j|^2\}$

Gaussian Processes

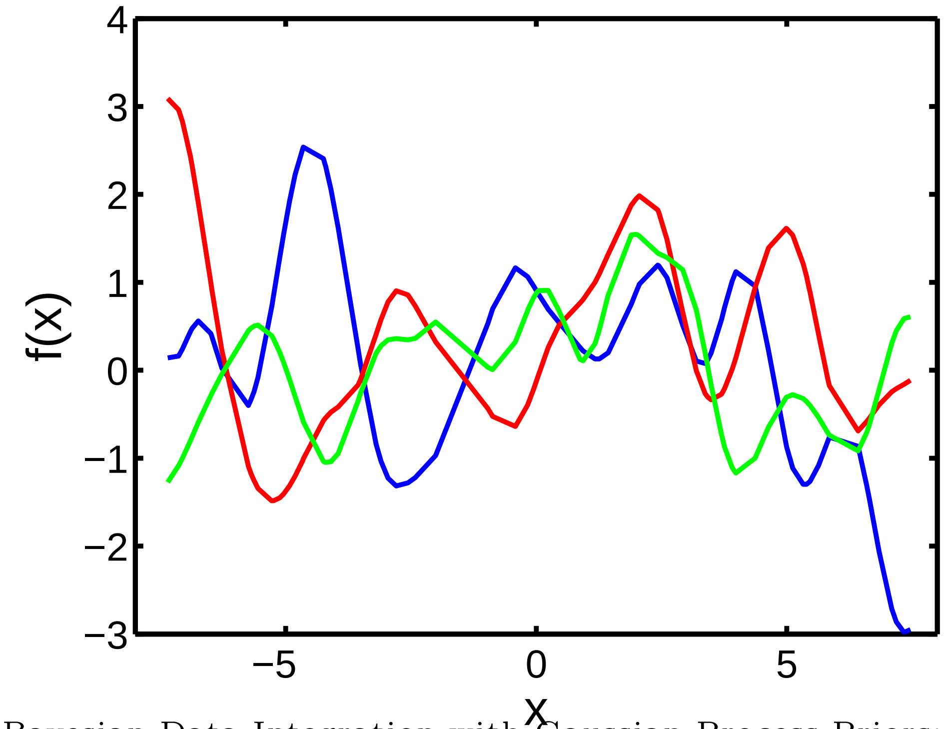
- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi |x_i - x_j|^2\}$
- GP prior sampled at 30 points on -8 to +8, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\theta = 1.0, \varphi = 0.1$,



Gaussian Processes

UNIVERSITY *of* GLASGOW

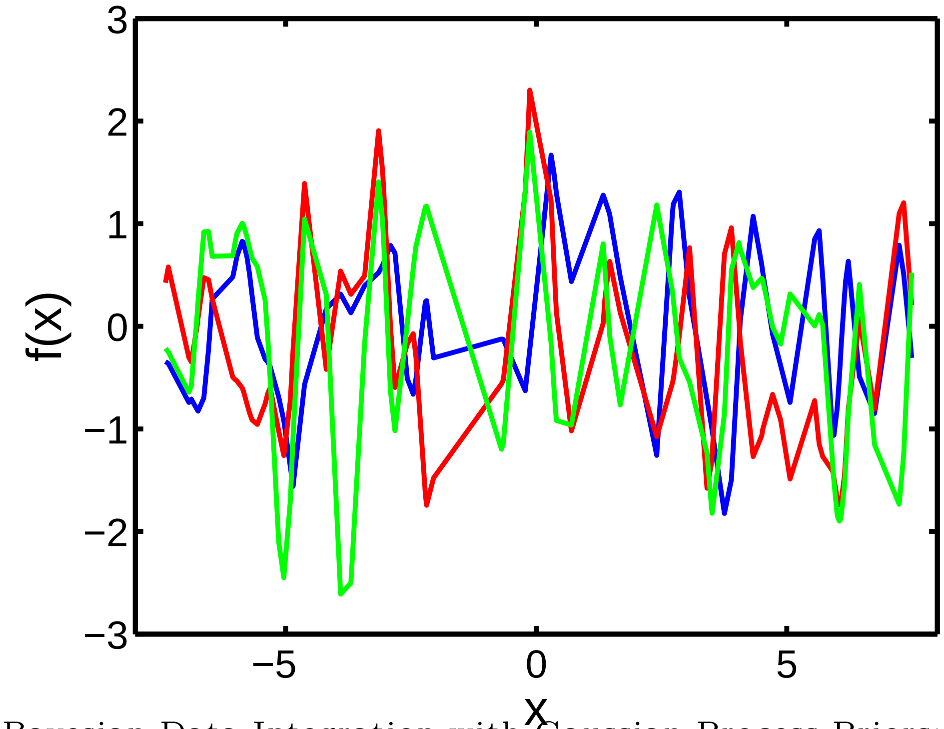
- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi |x_i - x_j|^2\}$
- GP prior sampled at 30 points on -8 to +8, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\theta = 1.0, \varphi = 1.0$,



Gaussian Processes

UNIVERSITY *of* GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi |x_i - x_j|^2\}$
- GP prior sampled at 30 points on -8 to +8, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, where $\theta = 1.0, \varphi = 10.0$,



Gaussian Processes

UNIVERSITY of GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi|x_i - x_j|^2\}$
- Radial Basis kernel is ∞ differentiable

Gaussian Processes

UNIVERSITY *of* GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi|x_i - x_j|^2\}$
- Radial Basis kernel is ∞ differentiable
- Width of kernel, φ, controls spectral decay rate of process, high decay rate $\Rightarrow$smooth process

Gaussian Processes

UNIVERSITY of GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi|x_i - x_j|^2\}$
- Radial Basis kernel is ∞ differentiable
- Width of kernel, φ, controls spectral decay rate of process, high decay rate $\Rightarrow$smooth process
- Classes of covariance functions to represent prior assumptions

Gaussian Processes

UNIVERSITY of GLASGOW

- Choose covariance function to define prior over function space e.g. $C(x_i, x_j) = \theta \exp\{-\varphi|x_i - x_j|^2\}$
- Radial Basis kernel is ∞ differentiable
- Width of kernel, φ, controls spectral decay rate of process, high decay rate $\Rightarrow$smooth process
- Classes of covariance functions to represent prior assumptions
- Consider simple regression problem as an example

GP Regression

UNIVERSITY *of* GLASGOW

- Consider simple function $f(x) = \frac{\sin(x)}{x}$

GP Regression

UNIVERSITY *of* GLASGOW

- Consider simple function $f(x) = \frac{\sin(x)}{x}$
- Observations (i.i.d) $\mathcal{D} = \{(x_1, t_1), \cdots, (x_N, t_N)\} = (\mathbf{x}, \mathbf{t})$ where $t_n = f(x_n) + \epsilon_n$, and assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$

GP Regression

UNIVERSITY *of* GLASGOW

- Consider simple function $f(x) = \frac{\sin(x)}{x}$
- Observations (i.i.d) $\mathcal{D} = \{(x_1, t_1), \cdots, (x_N, t_N)\} = (\mathbf{x}, \mathbf{t})$ where $t_n = f(x_n) + \epsilon_n$, and assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Place GP prior on functions $\mathbf{f} \in \mathbb{R}^N$,

$$\mathbf{f}|\mathbf{x}, \varphi, \theta \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

GP Regression

UNIVERSITY *of* GLASGOW

- Consider simple function $f(x) = \frac{\sin(x)}{x}$
- Observations (i.i.d) $\mathcal{D} = \{(x_1, t_1), \cdots, (x_N, t_N)\} = (\mathbf{x}, \mathbf{t})$ where $t_n = f(x_n) + \epsilon_n$, and assume $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$
- Place GP prior on functions $\mathbf{f} \in \mathbb{R}^N$,

$$\mathbf{f} | \mathbf{x}, \varphi, \theta \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

Likelihood

$$\mathbf{t} | \mathbf{f}, \sigma \sim \prod_n \mathcal{N}_{t_n}(f_n, \sigma^2) = \mathcal{N}_{\mathbf{t}}(\mathbf{f}, \sigma^2 \mathbf{I})$$

GP Regression

UNIVERSITY *of* GLASGOW

Posterior over functions

$$p(\mathbf{f}|\mathbf{x},\mathbf{t},\varphi,\theta,\sigma) \;=\; \frac{p(\mathbf{t}|\mathbf{f},\sigma)p(\mathbf{f}|\mathbf{x},\varphi,\theta)}{p(\mathbf{t}|\mathbf{x},\varphi,\theta,\sigma)}$$

GP Regression

UNIVERSITY *of* GLASGOW

Posterior over functions

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{x},\mathbf{t},\varphi,\theta,\sigma) &= \frac{p(\mathbf{t}|\mathbf{f},\sigma)p(\mathbf{f}|\mathbf{x},\varphi,\theta)}{p(\mathbf{t}|\mathbf{x},\varphi,\theta,\sigma)} \\
&= \frac{\mathcal{N}_{\mathbf{t}}(\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}_{\mathbf{f}}(\mathbf{0},\mathbf{C})}{\int \mathcal{N}_{\mathbf{t}}(\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}_{\mathbf{f}}(\mathbf{0},\mathbf{C})d\mathbf{f}}
\end{aligned}
$$

GP Regression

UNIVERSITY *of* GLASGOW

Posterior over functions

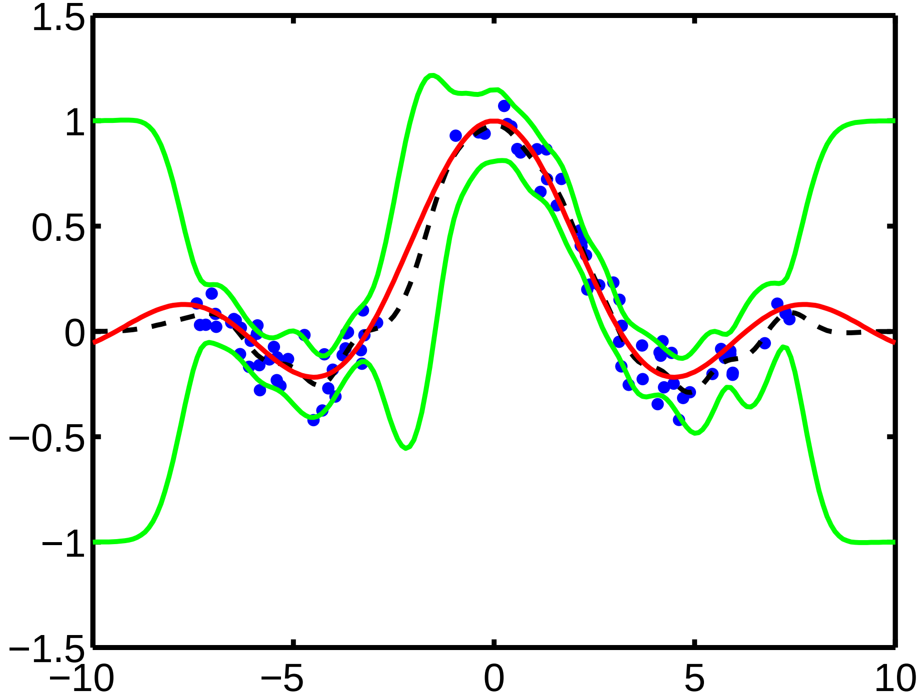
$$
\begin{aligned}
p(\mathbf{f}|\mathbf{x},\mathbf{t},\varphi,\theta,\sigma) &= \frac{p(\mathbf{t}|\mathbf{f},\sigma)p(\mathbf{f}|\mathbf{x},\varphi,\theta)}{p(\mathbf{t}|\mathbf{x},\varphi,\theta,\sigma)} \\
&= \frac{\mathcal{N}_{\mathbf{t}}(\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}_{\mathbf{f}}(\mathbf{0},\mathbf{C})}{\int \mathcal{N}_{\mathbf{t}}(\mathbf{f},\sigma^2\mathbf{I})\mathcal{N}_{\mathbf{f}}(\mathbf{0},\mathbf{C})d\mathbf{f}} \\
&= \mathcal{N}_{\mathbf{f}}(\boldsymbol{\mu},\boldsymbol{\Sigma})
\end{aligned}
$$

where $\boldsymbol{\Sigma} = \sigma^2\mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I})^{-1}$ and $\boldsymbol{\mu} = \mathbf{C}(\mathbf{C}+\sigma^2\mathbf{I})^{-1}\mathbf{t}$.

Predictive distribution over *new* data samples are also Gaussian

GP Regression

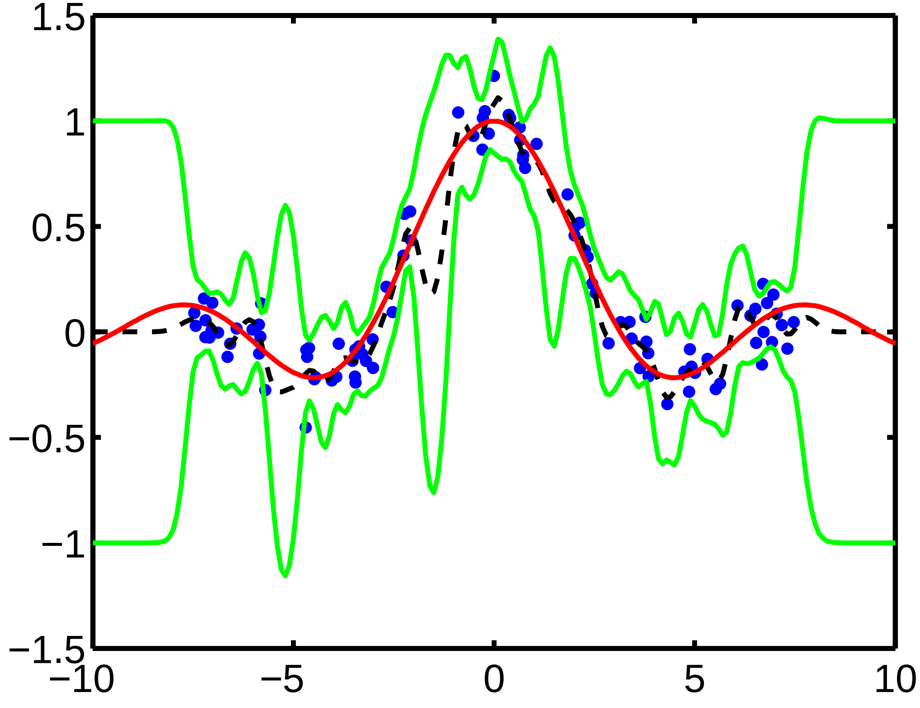
Noise level $\sigma^2 = 0.1$, 100 samples, $\theta = 1$, $\varphi = 1$



GP Regression

UNIVERSITY of GLASGOW

Noise level $\sigma^2 = 0.1$, 100 samples, $\theta = 1$, $\varphi = 5$



GP Regression

UNIVERSITY *of* GLASGOW

- For regression with Normal errors analytic Bayesian inference is possible

GP Regression

UNIVERSITY
of
GLASGOW

- For regression with Normal errors analytic Bayesian inference is possible
- Inference over covariance parameters requires either MCMC or type II ML

GP Regression

UNIVERSITY
of
GLASGOW

- For regression with Normal errors analytic Bayesian inference is possible
- Inference over covariance parameters requires either MCMC or type II ML
- As marginal likelihood $p(\mathbf{t}|\mathbf{x}, \varphi, \theta, \sigma) = \mathcal{N}_{\mathbf{t}}(\mathbf{0}, \mathbf{C} + \sigma^2 \mathbf{I})$

GP Regression

- For regression with Normal errors analytic Bayesian inference is possible
- Inference over covariance parameters requires either MCMC or type II ML
- As marginal likelihood $p(\mathbf{t}|\mathbf{x}, \varphi, \theta, \sigma) = \mathcal{N}_{\mathbf{t}}(\mathbf{0}, \mathbf{C} + \sigma^2 \mathbf{I})$
- Optimisation to obtain type-II estimates of hyper-parameters φ, θ, σ (evidence maximisation) i.e.

$$\hat{\varphi}, \hat{\theta}, \hat{\sigma} = \underset{\varphi,\theta,\sigma}{\operatorname{argmax}} \log \mathcal{N}_{\mathbf{t}}(\mathbf{0}, \mathbf{C} + \sigma^2 \mathbf{I})$$

GP Classification

- Classification setting data discrete $t \in \{1, \cdots, K\}$

GP Classification

- Classification setting data discrete $t \in \{1, \cdots, K\}$
- Assume one GP prior per class (overcomplete representation) and *a priori* inter-class GP independence

$$p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \varphi_1, \cdots, \varphi_K, \mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{f}_k}(\mathbf{0}, \mathbf{C}_k)$$

GP Classification

UNIVERSITY *of* GLASGOW

- Classification setting data discrete $t \in \{1, \cdots, K\}$
- Assume one GP prior per class (overcomplete representation) and *a priori* inter-class GP independence

$$p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \varphi_1, \cdots, \varphi_K, \mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{f}_k}(\mathbf{0}, \mathbf{C}_k)$$

- Likelihood follows as multinomial over targets

$$p(\mathbf{t}|\mathbf{f}_1, \cdots, \mathbf{f}_K, \boldsymbol{\theta}) \propto \prod_n \prod_k q_k(\mathbf{f}_n)^{\delta(t_n, k)}$$

GP Classification

UNIVERSITY of GLASGOW

- Classification setting data discrete $t \in \{1, \cdots, K\}$
- Assume one GP prior per class (overcomplete representation) and *a priori* inter-class GP independence

$$p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \varphi_1, \cdots, \varphi_K, \mathbf{X}) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{f}_k}(\mathbf{0}, \mathbf{C}_k)$$

- Likelihood follows as multinomial over targets

$$p(\mathbf{t} | \mathbf{f}_1, \cdots, \mathbf{f}_K, \boldsymbol{\theta}) \propto \prod_n \prod_k q_k(\mathbf{f}_n)^{\delta(t_n, k)}$$

- Where usual multinomial-logit definition is

$$q_k(\mathbf{f}_n) = \frac{\exp(f_{nk})}{\sum_{k'} \exp(f_{nk'})}$$

GP Classification

- Nice analytic inference not possible in classification setting

GP Classification

UNIVERSITY *of* GLASGOW

- Nice analytic inference not possible in classification setting
- Simulate samples from posterior using MCMC

GP Classification

UNIVERSITY *of* GLASGOW

- Nice analytic inference not possible in classification setting
- Simulate samples from posterior using MCMC
- Good approximations often desirable

GP Classification

- Nice analytic inference not possible in classification setting
- Simulate samples from posterior using MCMC
- Good approximations often desirable
- Laplace approximation for GP classification previously proposed by Williams & Barber, 1998.

GP Classification

UNIVERSITY *of* GLASGOW

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \mathbf{\Phi})$ with a Gaussian

GP Classification

UNIVERSITY *of* GLASGOW

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \boldsymbol{\Phi})$ with a Gaussian
- Gaussian centered at maximum of posterior density i.e. $\mathbf{f}_+^{MAP}$ where $\mathbf{f}_+ \equiv vec[\mathbf{f}_1, \cdots, \mathbf{f}_K]$ $(NK \times 1)$

GP Classification

UNIVERSITY of GLASGOW

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \mathbf{\Phi})$ with a Gaussian
- Gaussian centered at maximum of posterior density i.e. $\mathbf{f}_+^{MAP}$ where $\mathbf{f}_+ \equiv vec[\mathbf{f}_1, \cdots, \mathbf{f}_K]$ $(NK \times 1)$
- $\mathbf{\Sigma} = -\nabla_{\mathbf{f}_+} \nabla_{\mathbf{f}_+} \log p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{\Phi}) = \left(\mathbf{K}^{-1} - \mathbf{W}\right)^{-1}$

GP Classification

UNIVERSITY *of* GLASGOW

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \mathbf{\Phi})$ with a Gaussian
- Gaussian centered at maximum of posterior density i.e. $\mathbf{f}_+^{MAP}$ where $\mathbf{f}_+ \equiv vec[\mathbf{f}_1, \cdots, \mathbf{f}_K]$ $(NK \times 1)$
- $\mathbf{\Sigma} = -\nabla_{\mathbf{f}_+} \nabla_{\mathbf{f}_+} \log p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{\Phi}) = \left(\mathbf{K}^{-1} - \mathbf{W}\right)^{-1}$

$$\mathbf{K} = \begin{pmatrix} \mathbf{C}_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{C}_K \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{K1} & \cdots & \mathbf{W}_{KK} \end{pmatrix}$$

GP Classification

UNIVERSITY *of* GLASGOW

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \mathbf{\Phi})$ with a Gaussian
- Gaussian centered at maximum of posterior density i.e. $\mathbf{f}_+^{MAP}$ where $\mathbf{f}_+ \equiv vec[\mathbf{f}_1, \cdots, \mathbf{f}_K]$ $(NK \times 1)$
- $\mathbf{\Sigma} = -\nabla_{\mathbf{f}_+} \nabla_{\mathbf{f}_+} \log p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{\Phi}) = \left(\mathbf{K}^{-1} - \mathbf{W}\right)^{-1}$

$$\mathbf{K} = \begin{pmatrix} \mathbf{C}_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{C}_K \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{K1} & \cdots & \mathbf{W}_{KK} \end{pmatrix}$$

- Where each $(\mathbf{W}_{ij})_n = \frac{\partial^2}{\partial f_{nj} \partial f_{ni}} \log p(t_n | \mathbf{f}_1, \cdots, \mathbf{f}_K)$

GP Classification

- Approximate $p(\mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{t}, \mathbf{\Phi})$ with a Gaussian
- Gaussian centered at maximum of posterior density i.e. $\mathbf{f}_+^{MAP}$ where $\mathbf{f}_+ \equiv vec[\mathbf{f}_1, \cdots, \mathbf{f}_K]$ $(NK \times 1)$
- $\mathbf{\Sigma} = -\nabla_{\mathbf{f}_+} \nabla_{\mathbf{f}_+} \log p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K | \mathbf{X}, \mathbf{\Phi}) = \left(\mathbf{K}^{-1} - \mathbf{W}\right)^{-1}$

$$\mathbf{K} = \begin{pmatrix} \mathbf{C}_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{C}_K \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{K1} & \cdots & \mathbf{W}_{KK} \end{pmatrix}$$

- Where each $(\mathbf{W}_{ij})_n = \frac{\partial^2}{\partial f_{nj} \partial f_{ni}} \log p(t_n | \mathbf{f}_1, \cdots, \mathbf{f}_K)$
- Newton iterations to obtain mode $\mathbf{f}_+^{MAP}$

GP Classification

UNIVERSITY
of
GLASGOW

- Weakness with Laplace approximation

GP Classification

- Weakness with Laplace approximation
- Mode of high-dimensional Gaussian may not represent mass

GP Classification

UNIVERSITY *of* GLASGOW

- Weakness with Laplace approximation
- Mode of high-dimensional Gaussian may not represent mass
- Gaussian approximation to posterior in large sample limit - small samples available

GP Classification

UNIVERSITY *of* GLASGOW

- Weakness with Laplace approximation
- Mode of high-dimensional Gaussian may not represent mass
- Gaussian approximation to posterior in large sample limit - small samples available
- Variational methods with mean field approximations possibly more accurate alternative

Variational Approximations

- Approximate posterior over sets of variables, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots \boldsymbol{\theta}_M\}$ with a factored ensemble

$$P(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}) \approx \mathcal{Q}(\boldsymbol{\Theta}) = \prod_{i=1}^{M} Q(\boldsymbol{\theta}_i)$$

Variational Approximations

UNIVERSITY *of* GLASGOW

- Approximate posterior over sets of variables, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots \boldsymbol{\theta}_M\}$ with a factored ensemble

$$P(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}) \approx \mathcal{Q}(\boldsymbol{\Theta}) = \prod_{i=1}^{M} Q(\boldsymbol{\theta}_i)$$

- Optimise bound on marginal density (Jensen inequality)

$$\log P(\mathbf{t}|\mathbf{X}) \geq E_{\mathcal{Q}(\boldsymbol{\Theta})}\{\log P(\mathbf{t}, \boldsymbol{\Theta}|\mathbf{X})\} - E_{\mathcal{Q}(\boldsymbol{\Theta})}\{\log \mathcal{Q}(\boldsymbol{\Theta})\}$$

Variational Approximations

UNIVERSITY *of* GLASGOW

- Approximate posterior over sets of variables, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \cdots \boldsymbol{\theta}_M\}$ with a factored ensemble

$$P(\boldsymbol{\Theta}|\mathbf{t}, \mathbf{X}) \approx \mathcal{Q}(\boldsymbol{\Theta}) = \prod_{i=1}^{M} Q(\boldsymbol{\theta}_i)$$

- Optimise bound on marginal density (Jensen inequality)

$$\log P(\mathbf{t}|\mathbf{X}) \geq E_{\mathcal{Q}(\boldsymbol{\Theta})}\{\log P(\mathbf{t}, \boldsymbol{\Theta}|\mathbf{X})\} - E_{\mathcal{Q}(\boldsymbol{\Theta})}\{\log \mathcal{Q}(\boldsymbol{\Theta})\}$$

- To obtain optimal form of components of approximate posterior

$$Q(\boldsymbol{\theta}_i) \propto \exp\left(E_{\mathcal{Q}(\boldsymbol{\Theta}_{-i})}\{\log P(\mathbf{t}, \boldsymbol{\Theta}|\mathbf{X})\}\right)$$

Variational Approximations

UNIVERSITY of GLASGOW

- Each component of approximate posterior requires expectations w.r.t all other posterior components

Variational Approximations

UNIVERSITY
of
GLASGOW

- Each component of approximate posterior requires expectations w.r.t all other posterior components
- As multinomial-logit not in exponential family no closed form representations for approximate posteriors available

Variational Approximations

UNIVERSITY *of* GLASGOW

- Each component of approximate posterior requires expectations w.r.t all other posterior components
- As multinomial-logit not in exponential family no closed form representations for approximate posteriors available
- Gibbs & MacKay (1998) make additional specific approximations to the multinomial-logit - undesireable

Variational Approximations

UNIVERSITY *of* GLASGOW

- Each component of approximate posterior requires expectations w.r.t all other posterior components
- As multinomial-logit not in exponential family no closed form representations for approximate posteriors available
- Gibbs & MacKay (1998) make additional specific approximations to the multinomial-logit - undesireable
- Variational approximations for multinomial-logit likelihood inappropriate - Stuck with Laplace Approximation

Variational Approximations

UNIVERSITY *of* GLASGOW

- Each component of approximate posterior requires expectations w.r.t all other posterior components
- As multinomial-logit not in exponential family no closed form representations for approximate posteriors available
- Gibbs & MacKay (1998) make additional specific approximations to the multinomial-logit - undesireable
- Variational approximations for multinomial-logit likelihood inappropriate - Stuck with Laplace Approximation
- However progress can be made with variational approximations by considering alternative likelihood terms to the multinomial-logit

Data Augmentation Trick

UNIVERSITY of GLASGOW

- Consider Probit function $p(t_n = 1|f_n) = \Phi(f_n)$, by introducing the auxiliary variable $y_n \sim \mathcal{N}_y(f_n, 1)$ then

$$\int P(t_n = 1, y_n|f_n)dy_n = \int P(t_n = 1|y_n)p(y_n|f_n)dy_n$$

Data Augmentation Trick

UNIVERSITY *of* GLASGOW

- Consider Probit function $p(t_n = 1|f_n) = \Phi(f_n)$, by introducing the auxiliary variable $y_n \sim \mathcal{N}_y(f_n, 1)$ then

$$\int P(t_n = 1, y_n|f_n)dy_n = \int P(t_n = 1|y_n)p(y_n|f_n)dy_n$$

- By definition $P(t_n = 1|y_n) = \delta(y_n > 0)$ then the marginal is the normalizing constant of a left truncated univariate Gaussian

$$P(t_n = 1|f_n) = \int \delta(y_n > 0)\mathcal{N}_{y_n}(f_n, 1)dy_n$$

Data Augmentation Trick

UNIVERSITY of GLASGOW

- Consider Probit function $p(t_n = 1|f_n) = \Phi(f_n)$, by introducing the auxiliary variable $y_n \sim \mathcal{N}_y(f_n, 1)$ then

$$\int P(t_n = 1, y_n|f_n)dy_n = \int P(t_n = 1|y_n)p(y_n|f_n)dy_n$$

- By definition $P(t_n = 1|y_n) = \delta(y_n > 0)$ then the marginal is the normalizing constant of a left truncated univariate Gaussian

$$P(t_n = 1|f_n) = \int \delta(y_n > 0)\mathcal{N}_{y_n}(f_n, 1)dy_n$$

- Now have a Gaussian in joint distribution which allows us to make progress

Multinomial Probit

UNIVERSITY *of* GLASGOW

- Case for multiple classes slightly more involved as now auxiliary variable is a K-dim vector

Multinomial Probit

UNIVERSITY *of* GLASGOW

- Case for multiple classes slightly more involved as now auxiliary variable is a K-dim vector
- For 1 from K classes then

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \leq k \leq K} \{y_{nk}\}$$

Multinomial Probit

UNIVERSITY *of* GLASGOW

- Case for multiple classes slightly more involved as now auxiliary variable is a K-dim vector
- For 1 from K classes then

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \leq k \leq K} \{y_{nk}\}$$

- This has the effect of dividing $\mathbb{R}^K$ ($\mathbf{y}$ space) into K non-overlapping K-dimensional cones $\mathcal{C}_k = \{\mathbf{y} : y_k > y_i, k \neq i\}$ where $\mathbb{R}^K = \cup_k \mathcal{C}_k$

Multinomial Probit

UNIVERSITY *of* GLASGOW

- Case for multiple classes slightly more involved as now auxiliary variable is a K-dim vector
- For 1 from K classes then

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \leq k \leq K} \{y_{nk}\}$$

- This has the effect of dividing $\mathbb{R}^K$ ($\mathbf{y}$ space) into K non-overlapping K-dimensional cones $\mathcal{C}_k = \{\mathbf{y} : y_k > y_i, k \neq i\}$ where $\mathbb{R}^K = \cup_k \mathcal{C}_k$
- So each

$$P(t_n = i|\mathbf{y}_n) = \delta(y_{ni} > y_{nk} \,\forall\, k \neq i)\delta(t_n = i)$$

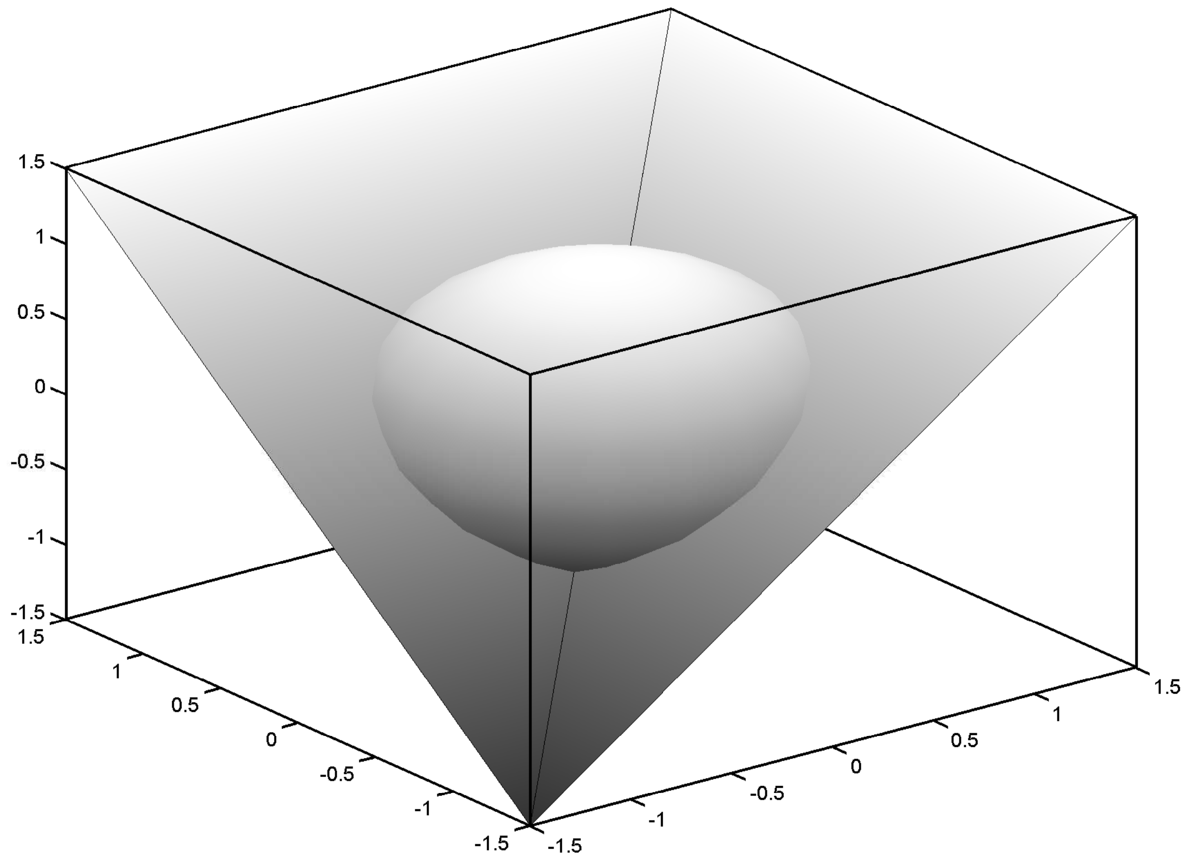
Multinomial Probit

UNIVERSITY
of
GLASGOW

Conic truncation of $\mathbb{R}^3$

Multinomial Probit

Conic truncation of $\mathbb{R}^3$



Multinomial Probit

Multinomial-Probit Likelihood follows as

$$P(t_n = i | f_{n1}, \cdots, f_{nK}) =$$

$$\int \delta(y_{ni} > y_{nk} \; \forall \; k \neq i) \prod_{j=1}^{K} p(y_{nj}|f_{nj}) d\mathbf{y} =$$

$$\int_{\mathcal{C}_i} \prod_{j=1}^{K} p(y_{nj}|f_{nj}) d\mathbf{y} = E_{p(u)} \left\{ \prod_{j \neq i} \Phi(u + f_{ni} - f_{nj}) \right\}$$

Joint Likelihood

UNIVERSITY of GLASGOW

- Augmented joint distribution, $p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K, \mathbf{y}_1, \cdots, \mathbf{y}_K | \mathbf{X}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_K)$, given as

$$= \prod_{n=1}^{N} \left\{ \sum_{i=1}^{K} \delta(y_{ni} > y_{nk} \; \forall \; k \neq i) \delta(t_n = i) \right\} \times$$

$$\prod_{k=1}^{K} p(y_{nk} | f_{nk}) p(\mathbf{f}_k | \mathbf{X}, \boldsymbol{\varphi}_k)$$

Joint Likelihood

UNIVERSITY *of* GLASGOW

- Augmented joint distribution, $p(\mathbf{t}, \mathbf{f}_1, \cdots, \mathbf{f}_K, \mathbf{y}_1, \cdots, \mathbf{y}_K | \mathbf{X}, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_K)$, given as

$$= \prod_{n=1}^{N} \left\{ \sum_{i=1}^{K} \delta(y_{ni} > y_{nk} \; \forall \; k \neq i) \delta(t_n = i) \right\} \times$$

$$\prod_{k=1}^{K} p(y_{nk} | f_{nk}) p(\mathbf{f}_k | \mathbf{X}, \boldsymbol{\varphi}_k)$$

- Now obtain approximate posteriors $Q(\mathbf{f}_1, \cdots, \mathbf{f}_K)$ & $Q(\mathbf{y}_1, \cdots, \mathbf{y}_K)$

Approximate Posteriors

UNIVERSITY of GLASGOW

The approximate posteriors are

$$Q(\mathbf{f}_1, \cdots, \mathbf{f}_K) = \prod_{k=1}^{K} Q(\mathbf{f}_k) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{f}_k}(\boldsymbol{\Sigma}_k \widetilde{\mathbf{y}}_k, \boldsymbol{\Sigma}_k)$$

where $\boldsymbol{\Sigma}_k = \mathbf{C}_k \left(\mathbf{I} + \mathbf{C}_k\right)^{-1}$ and $\widetilde{f(a)} = E_{Q(a)}\{f(a)\}$ denotes posterior expectation

Approximate Posteriors

UNIVERSITY *of* GLASGOW

The approximate posteriors are

$$Q(\mathbf{f}_1, \cdots, \mathbf{f}_K) = \prod_{k=1}^{K} Q(\mathbf{f}_k) = \prod_{k=1}^{K} \mathcal{N}_{\mathbf{f}_k}(\mathbf{\Sigma}_k \widetilde{\mathbf{y}}_k, \mathbf{\Sigma}_k)$$

where $\mathbf{\Sigma}_k = \mathbf{C}_k \left(\mathbf{I} + \mathbf{C}_k\right)^{-1}$ and $\widetilde{f(a)} = E_{Q(a)}\{f(a)\}$ denotes posterior expectation

$$Q(\mathbf{y}_1, \cdots, \mathbf{y}_K) = \prod_{n=1}^{N} \mathcal{N}_{\mathbf{y}_n}^{t_n}(\widetilde{\mathbf{f}}_n, \mathbf{I})$$

Conic truncations of a multivariate Gaussians such that if $t_n = i$ where $i \in \{1, \cdots, K\}$ then the i'th dimension has the largest value.

GP Classification

UNIVERSITY *of* GLASGOW

- The required posterior expectations $\widetilde{y}_{nk}$ for all $k \neq i$ and $\widetilde{y}_{ni}$ follow as

$$
\begin{aligned}
\widetilde{y}_{nk} &= \widetilde{f}_{nk} - \frac{E_{p(u)}\left\{\mathcal{N}_u(\widetilde{f}_{nk} - \widetilde{f}_{ni}, 1)\Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi(u + \widetilde{f}_{ni} - \widetilde{f}_{nk})\Phi_u^{n,i,k}\right\}} \\
\widetilde{y}_{ni} &= \widetilde{f}_{ni} - \left(\sum_{j \neq i} \widetilde{y}_{nj} - \widetilde{f}_{nj}\right)
\end{aligned}
$$

where $\Phi_u^{n,i,k} = \prod_{j \neq i,k} \Phi(u + \widetilde{f}_{ni} - \widetilde{f}_{nj})$, and $p(u) = \mathcal{N}_u(0, 1)$.

GP Classification

UNIVERSITY of GLASGOW

- Posterior mean for auxilliary variables fully defined by GP posterior means (row vs columnwise)

GP Classification

UNIVERSITY
of
GLASGOW

- Posterior mean for auxilliary variables fully defined by GP posterior means (row vs columnwise)
- Posterior mean estimates for each set of GP variables

$$\widetilde{\mathbf{f}}_k \leftarrow \mathbf{C}_k(\mathbf{I} + \mathbf{C}_k)^{-1}(\widetilde{\mathbf{f}}_k + \mathbf{p}_k)$$

GP Classification

UNIVERSITY *of* GLASGOW

- Posterior mean for auxilliary variables fully defined by GP posterior means (row vs columnwise)
- Posterior mean estimates for each set of GP variables

$$\widetilde{\mathbf{f}}_k \leftarrow \mathbf{C}_k(\mathbf{I} + \mathbf{C}_k)^{-1}(\widetilde{\mathbf{f}}_k + \mathbf{p}_k)$$

- Where $\mathbf{p}_k$ is the k^{th} column of the $N \times K$ matrix $\mathbf{P}$ whose elements p_{nk} are defined as follows:- for $t_n = i$ then for all $k \neq i$ $p_{nk} = -\frac{E_{p(u)}\left\{\mathcal{N}_u(\widetilde{f}_{nk} - \widetilde{f}_{ni}, 1)\Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi(u + \widetilde{f}_{ni} - \widetilde{f}_{nk})\Phi_u^{n,i,k}\right\}}$ and $p_{ni} = -\sum_{j \neq i} p_{nj}$.

GP Classification

UNIVERSITY *of* GLASGOW

- Posterior mean for auxilliary variables fully defined by GP posterior means (row vs columnwise)
- Posterior mean estimates for each set of GP variables

$$\widetilde{\mathbf{f}}_k \leftarrow \mathbf{C}_k(\mathbf{I}+\mathbf{C}_k)^{-1}(\widetilde{\mathbf{f}}_k + \mathbf{p}_k)$$

- Where $\mathbf{p}_k$ is the k^{th} column of the $N \times K$ matrix $\mathbf{P}$ whose elements p_{nk} are defined as follows:- for $t_n = i$ then for all $k \neq i$ $p_{nk} = -\frac{E_{p(u)}\left\{\mathcal{N}_u(\widetilde{f}_{nk} - \widetilde{f}_{ni}, 1)\Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi(u + \widetilde{f}_{ni} - \widetilde{f}_{nk})\Phi_u^{n,i,k}\right\}}$ and $p_{ni} = -\sum_{j \neq i} p_{nj}$.
- Scaling $\mathcal{O}(KN^3)$ worst case (Laplace $\mathcal{O}(K^3N^3)$)

GP Classification

UNIVERSITY of GLASGOW

- Variational Bayesian treatment of hyper-parameters also feasible - employ importance sampling to obtain posterior mean estimates

GP Classification

- Variational Bayesian treatment of hyper-parameters also feasible - employ importance sampling to obtain posterior mean estimates
- Predictive likelihood, $P(t_{new} = k|\mathbf{x}_{new}, \mathbf{X}, \mathbf{t})$, follows as

$$E_{p(u)}\left\{\prod_{j\neq k}\Phi\left(\frac{1}{\widetilde{\nu}_j^{new}}\left[u\widetilde{\nu}_k^{new} + \widetilde{f}_k^{new} - \widetilde{f}_j^{new}\right]\right)\right\}$$

where each $\widetilde{\nu}_k^{new} = \sqrt{1 + \widetilde{\sigma^2_{k,new}}}$

Comparison with MCMC

- How good is the VB approximation?

Comparison with MCMC

UNIVERSITY *of* GLASGOW

- How good is the VB approximation?
- Assume gold standard obtained from MCMC (straightforward Gibbs sampler)

Comparison with MCMC

UNIVERSITY of GLASGOW

- How good is the VB approximation?
- Assume gold standard obtained from MCMC (straightforward Gibbs sampler)
- Take predictive likelihood on independent held-out sample to be measure of goodness

Comparison with MCMC

- How good is the VB approximation?
- Assume gold standard obtained from MCMC (straightforward Gibbs sampler)
- Take predictive likelihood on independent held-out sample to be measure of goodness
- How much information do the predictive probabilities provide regarding the predicted classes

Comparison with MCMC

UNIVERSITY *of* GLASGOW

- How good is the VB approximation?
- Assume gold standard obtained from MCMC (straightforward Gibbs sampler)
- Take predictive likelihood on independent held-out sample to be measure of goodness
- How much information do the predictive probabilities provide regarding the predicted classes
- 0-1 error rate - blunt instrument, marginal likelihood - very difficult to reliably estimate

Experiments

- Employ 3-Class data set for *training* & *testing* - UCI Wine

Experiments

- Employ 3-Class data set for *training* & *testing* - UCI Wine
- Obtain MCMC, VB & Laplace based GP classifiers

Experiments

UNIVERSITY *of* GLASGOW

- Employ 3-Class data set for *training* & *testing* - UCI Wine
- Obtain MCMC, VB & Laplace based GP classifiers
- Record predictive likelihood on *test* set

Experiments

UNIVERSITY *of* GLASGOW

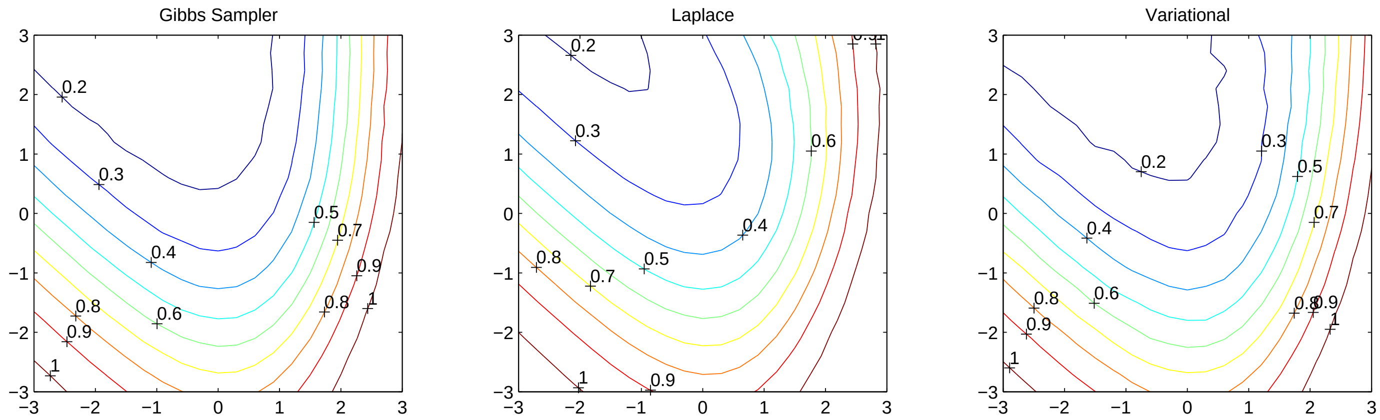
- Employ 3-Class data set for *training* & *testing* - UCI Wine
- Obtain MCMC, VB & Laplace based GP classifiers
- Record predictive likelihood on *test* set
- Employ single covariance function across all classes $\theta \exp\{-\varphi|\mathbf{x}_i - \mathbf{x}_j|^2\}$ (Kuss & Rassmussen, 2005)

Experiments

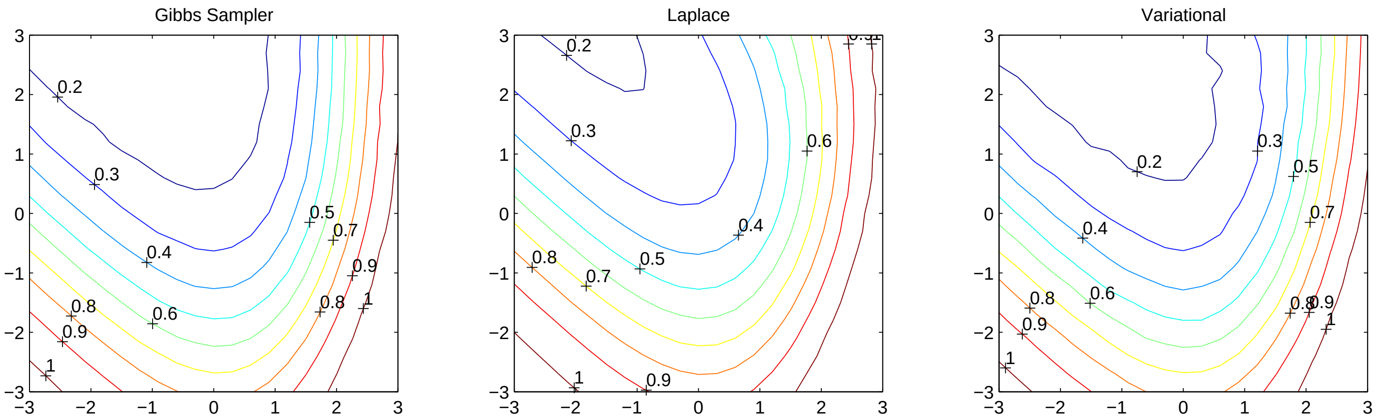
UNIVERSITY of GLASGOW

- Employ 3-Class data set for *training* & *testing* - UCI Wine
- Obtain MCMC, VB & Laplace based GP classifiers
- Record predictive likelihood on *test* set
- Employ single covariance function across all classes $\theta \exp\{-\varphi|\mathbf{x}_i - \mathbf{x}_j|^2\}$ (Kuss & Rassmussen, 2005)
- Evaluate predictive performance over a 21 $\times$ 21 grid of hyper-parameter values

Comparison with MCMC



Comparison with MCMC



On a number of datasets it is observed that the systematic predictive likelihood response is better preserved by the Variational approximation

Experiments

Toy-Data	Laplace	Variational	Gibbs Sampler
Marginal Likelihood	-169.27 $\pm$ 4.27	-232.00 $\pm$ 17.13	-94.07 $\pm$ 11.26
Predictive Error	3.97 $\pm$ 2.00	3.65 $\pm$ 1.95	3.49 $\pm$ 1.69
Predictive Likelihood	-98.90 $\pm$ 8.22	**-72.27 $\pm$ 9.25**	**-73.44 $\pm$ 7.67**
Iris	Laplace	Variational	Gibbs Sampler
Marginal Likelihood	-143.87 $\pm$ 1.17	-202.98 $\pm$ 1.37	-45.27 $\pm$ 6.17
Predictive Error	4.12 $\pm$ 2.14	4.08 $\pm$ 2.16	4.08 $\pm$ 2.16
Predictive Likelihood	-10.41 $\pm$ 1.28	**-7.35 $\pm$ 1.27**	**-7.26 $\pm$ 1.40**
Thyroid	Laplace	Variational	Gibbs Sampler
Marginal Likelihood	-158.52 $\pm$ 1.83	-246.24 $\pm$ 1.63	-68.82 $\pm$ 8.29
Predictive Error	4.08 $\pm$ 2.26	3.86 $\pm$ 2.04	3.94 $\pm$ 2.02
Predictive Likelihood	-18.75 $\pm$ 2.47	**-14.62 $\pm$ 2.70**	**-14.47 $\pm$ 2.39**

Experiments

UNIVERSITY *of* GLASGOW

Wine	Laplace	Variational	Gibbs Sampler
Marginal Likelihood	-152.22 $\pm$ 1.29	-253.90 $\pm$ 1.52	-68.65 $\pm$ 6.19
Predictive Error	3.08 $\pm$ 2.16	2.65 $\pm$ 1.87	2.78 $\pm$ 2.07
Predictive Likelihood	-14.61 $\pm$ 1.29	**-10.16 $\pm$ 1.47**	**-10.47 $\pm$ 1.41**
Forensic Glass	Laplace	Variational	Gibbs Sampler
Marginal Likelihood	-275.11 $\pm$ 2.87	-776.79 $\pm$ 5.75	-268.21 $\pm$ 5.46
Predictive Error	36.54 $\pm$ 4.74	32.79 $\pm$ 4.57	34.00 $\pm$ 4.62
Predictive Likelihood	-90.38 $\pm$ 3.25	**-77.60 $\pm$ 3.91**	**-79.86 $\pm$ 4.80**

Experiments

- Consider inference over parameters and hyper-parameters

Experiments

UNIVERSITY of GLASGOW

- Consider inference over parameters and hyper-parameters
- MCMC requires Metropolis-Hastings sub-sampler to obtain hyper-parameter samples within overall Gibbs sampler

Experiments

UNIVERSITY of GLASGOW

- Consider inference over parameters and hyper-parameters
- MCMC requires Metropolis-Hastings sub-sampler to obtain hyper-parameter samples within overall Gibbs sampler
- VB employs importance sampler to obtain posterior-mean estimates for hyper-parameters

Experiments

UNIVERSITY of GLASGOW

- Consider inference over parameters and hyper-parameters
- MCMC requires Metropolis-Hastings sub-sampler to obtain hyper-parameter samples within overall Gibbs sampler
- VB employs importance sampler to obtain posterior-mean estimates for hyper-parameters
- Employ toy-data from Neal (1998), two features required to define classes with two additional redundant features included

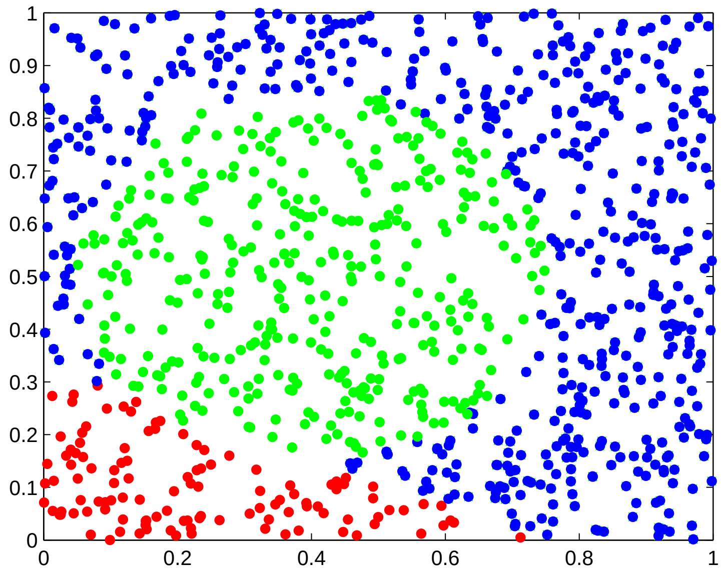
Experiments

UNIVERSITY *of* GLASGOW

- Distribution of two relevant features defining class partitioning

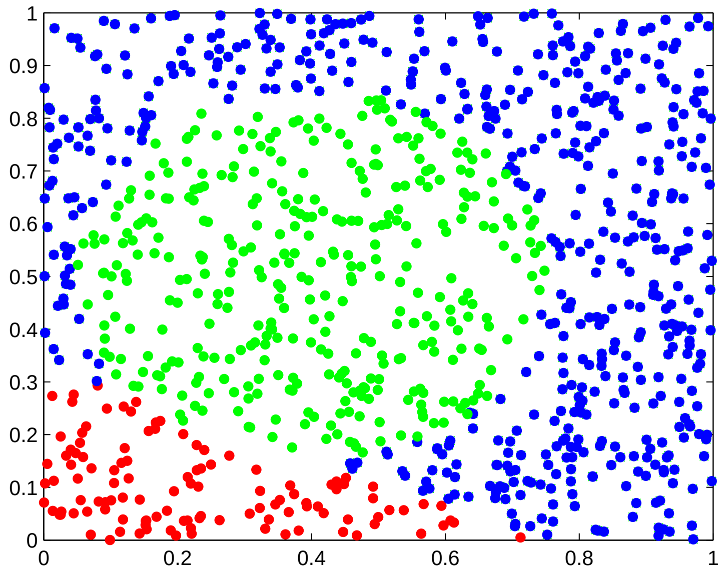
Experiments

- Distribution of two relevant features defining class partitioning



Experiments

- Distribution of two relevant features defining class partitioning



- Two additional redundant features included

Experiments

- Compare MCMC with Variational Approximation

Experiments

- Compare MCMC with Variational Approximation
- Measure predictive likelihood achieved under both schemes, employ RBF covariance function $C(x_i, x_j) = \exp\{-\sum_d \varphi_d |x_{id} - x_{jd}|^2\}$

Experiments

UNIVERSITY of GLASGOW

- Compare MCMC with Variational Approximation
- Measure predictive likelihood achieved under both schemes, employ RBF covariance function $C(x_i, x_j) = \exp\{-\sum_d \varphi_d |x_{id} - x_{jd}|^2\}$
- Gibbs sampler, after 5,000 sample burn-in for each posterior sample an additional 100 samples per test point drawn from predictive priors to obtain MC estimate of predictive likelihood

Experiments

UNIVERSITY *of* GLASGOW

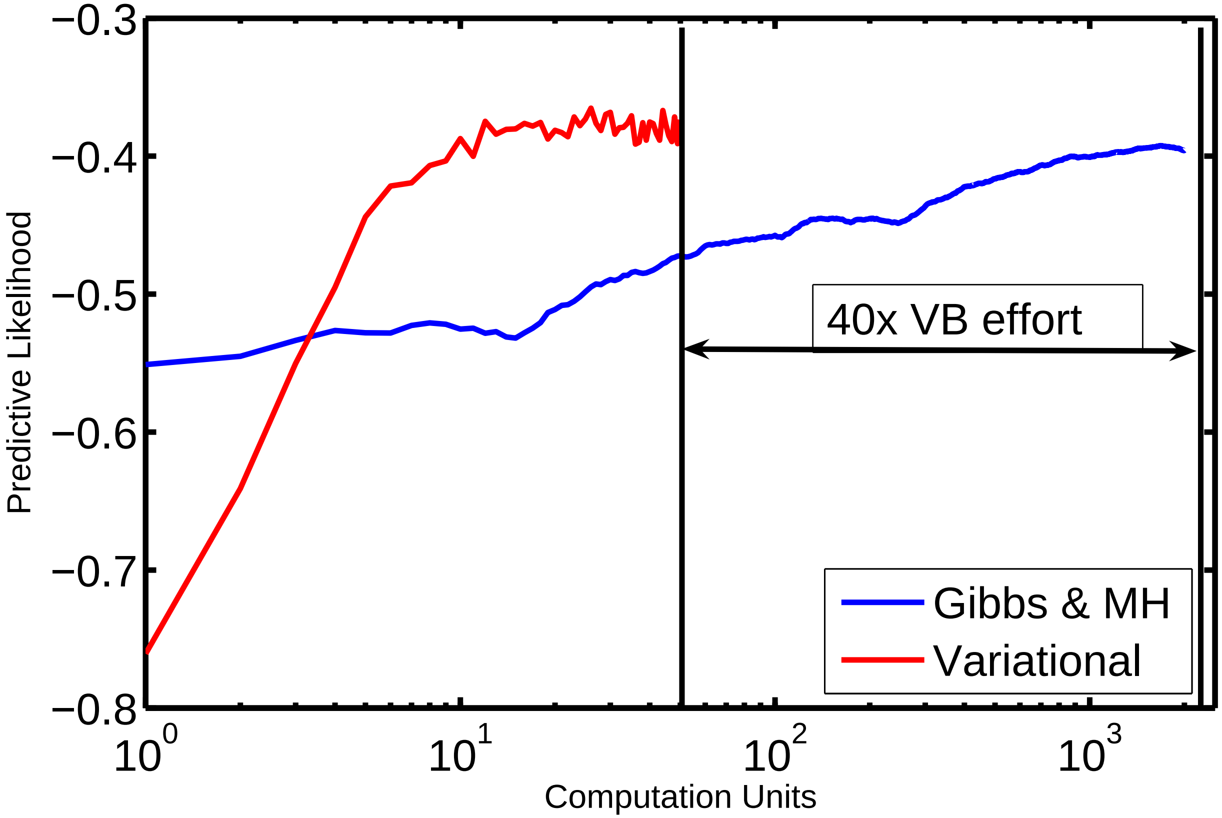
- Compare MCMC with Variational Approximation
- Measure predictive likelihood achieved under both schemes, employ RBF covariance function $C(x_i, x_j) = \exp\{-\sum_d \varphi_d |x_{id} - x_{jd}|^2\}$
- Gibbs sampler, after 5,000 sample burn-in for each posterior sample an additional 100 samples per test point drawn from predictive priors to obtain MC estimate of predictive likelihood
- Gibbs sampler, for each posterior sample drawn MH requires 2,000 sample burn-in before single hyper-parameter sample drawn

Experiments

UNIVERSITY *of* GLASGOW

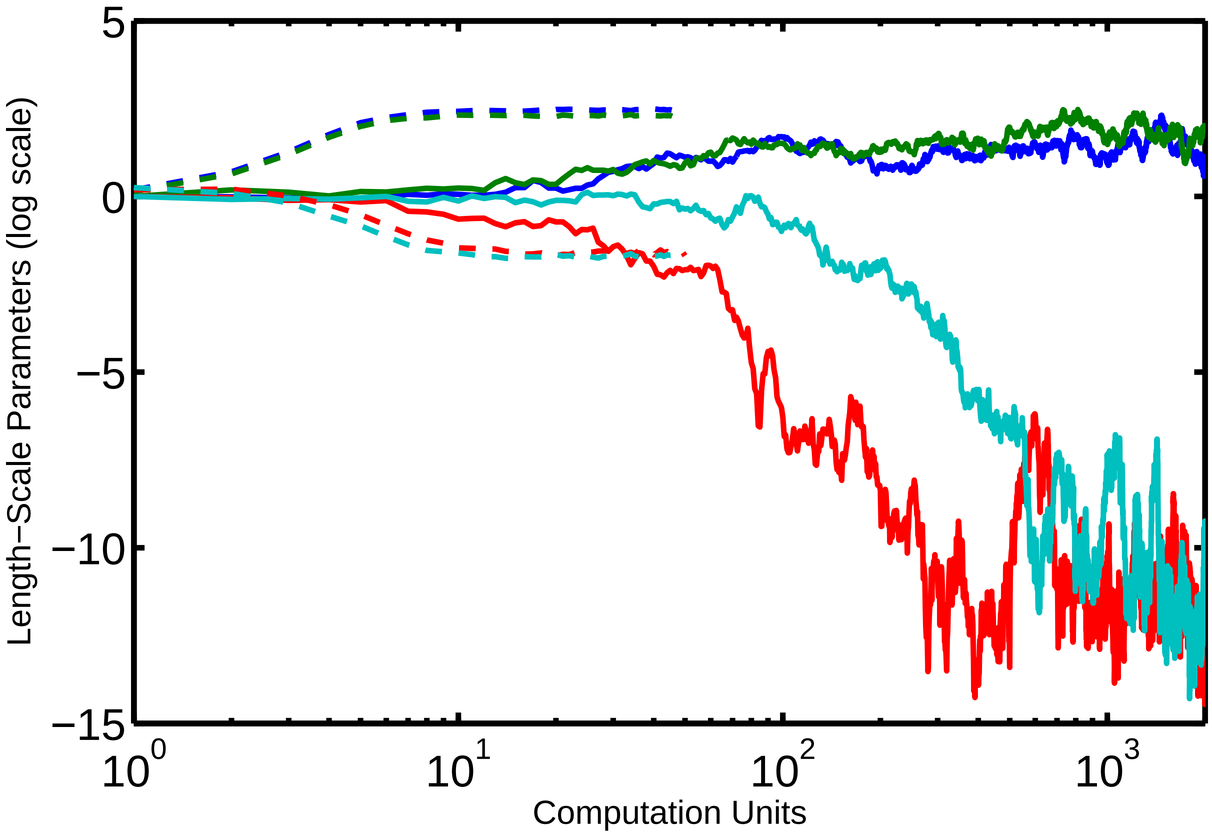
- Compare MCMC with Variational Approximation
- Measure predictive likelihood achieved under both schemes, employ RBF covariance function $C(x_i, x_j) = \exp\{-\sum_d \varphi_d |x_{id} - x_{jd}|^2\}$
- Gibbs sampler, after 5,000 sample burn-in for each posterior sample an additional 100 samples per test point drawn from predictive priors to obtain MC estimate of predictive likelihood
- Gibbs sampler, for each posterior sample drawn MH requires 2,000 sample burn-in before single hyper-parameter sample drawn
- Variational approximation, 2,000 samples drawn from hyper-parameter prior to estimate posterior mean

Experiments



Experiments

UNIVERSITY of GLASGOW



Composite Covariance

UNIVERSITY of GLASGOW

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$.

Composite Covariance

UNIVERSITY *of* GLASGOW

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is $f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}_j)$.

Composite Covariance

UNIVERSITY of GLASGOW

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is $f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}_j)$.
- Where each $f_j(\mathbf{x}_j) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}_j)$ and $C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$

Composite Covariance

UNIVERSITY *of* GLASGOW

- Distinct feature representation of X, $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$.
- A linear model is employed in this new space such that the overall nonlinear transformation is $f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}_j)$.
- Where each $f_j(\mathbf{x}_j) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}_j)$ and $C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$
- Then $f(X) \sim GP\left(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{\mathcal{J}}, \beta_1 \cdots \beta_{\mathcal{J}}\right)$ where now the overall mean and covariance functions follow as $\sum_{j=1}^{\mathcal{J}} \beta_j m_j(\mathbf{x}_j)$ and $\sum_{j=1}^{\mathcal{J}} \beta_j^2 C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$

Composite Covariance

UNIVERSITY of GLASGOW

- Protein fold recognition problem - predict 27 SCOP folds of proteins with low sequence similarity

Composite Covariance

UNIVERSITY
of
GLASGOW

- Protein fold recognition problem - predict 27 SCOP folds of proteins with low sequence similarity
- Problem first considered in Ding & Dubchak, 2000, employing 6 parameter datasets

Composite Covariance

UNIVERSITY *of* GLASGOW

- Protein fold recognition problem - predict 27 SCOP folds of proteins with low sequence similarity
- Problem first considered in Ding & Dubchak, 2000, employing 6 parameter datasets
- One vs One combination of SVM's followed by heuristic voting combination

Composite Covariance

- Protein fold recognition problem - predict 27 SCOP folds of proteins with low sequence similarity
- Problem first considered in Ding & Dubchak, 2000, employing 6 parameter datasets
- One vs One combination of SVM's followed by heuristic voting combination
- On independent test set of proteins 43.5% correct predictions achieved

Composite Covariance

UNIVERSITY *of* GLASGOW

- Protein fold recognition problem - predict 27 SCOP folds of proteins with low sequence similarity
- Problem first considered in Ding & Dubchak, 2000, employing 6 parameter datasets
- One vs One combination of SVM's followed by heuristic voting combination
- On independent test set of proteins 43.5% correct predictions achieved
- Manual investigation of different combinations of datasets showed possible increase to 56.5% (62% published July 2006)

Composite Covariance

UNIVERSITY *of* GLASGOW

- Six datasets (AAC, SS, H, P, Pz, V) of D&D employed also include one random *noise* dataset

Composite Covariance

- Six datasets (AAC, SS, H, P, Pz, V) of D&D employed also include one random *noise* dataset
- Seven Gram matrices (RBF and inner-products) available, define Dirichlet prior on $\beta_1^2, \cdots, \beta_{\mathcal{J}}^2$ & Gamma on Dirichlet mean

Composite Covariance

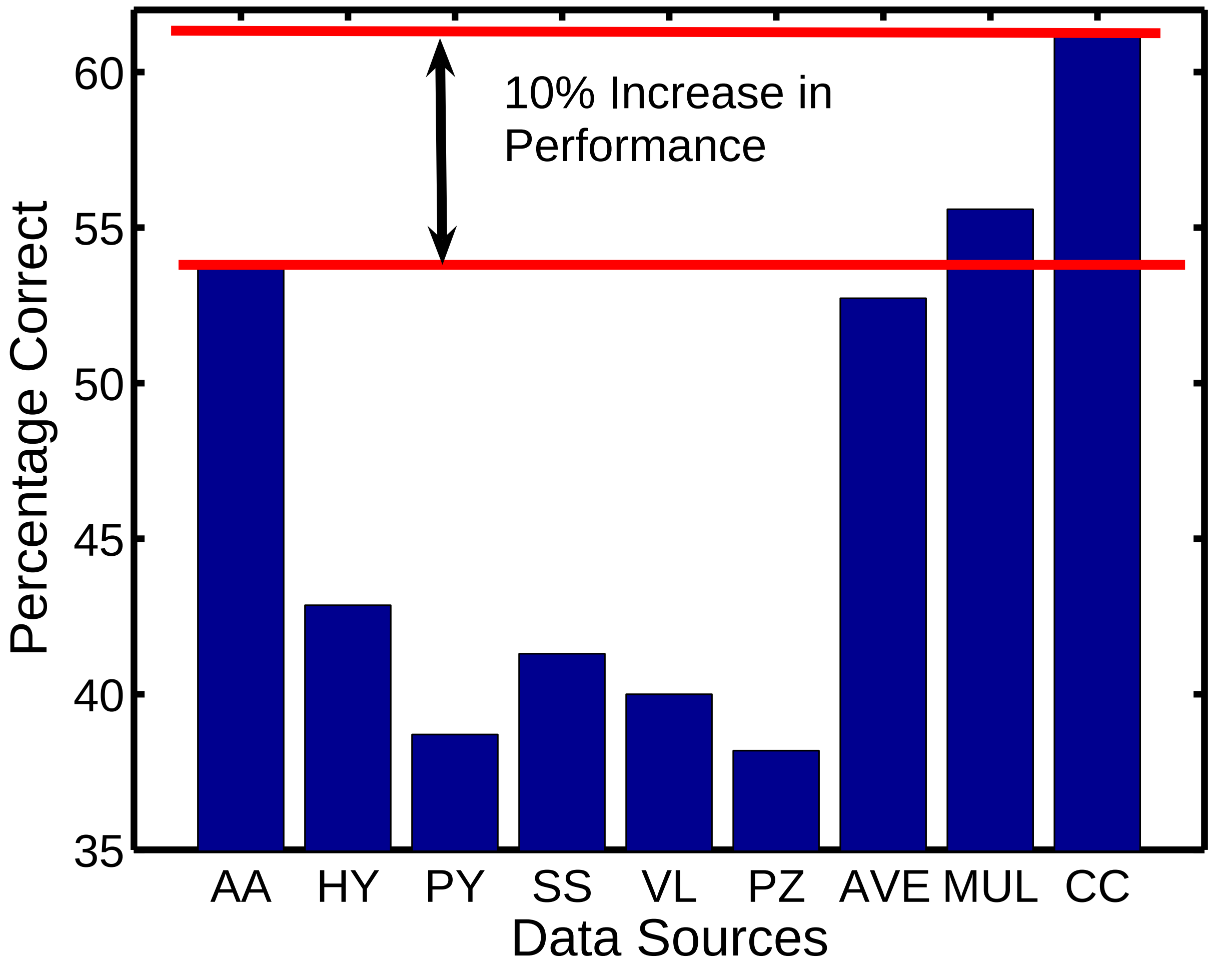
UNIVERSITY of GLASGOW

- Six datasets (AAC, SS, H, P, Pz, V) of D&D employed also include one random *noise* dataset
- Seven Gram matrices (RBF and inner-products) available, define Dirichlet prior on $\beta_1^2, \cdots, \beta_{\mathcal{J}}^2$ & Gamma on Dirichlet mean
- Run variational Bayes routine with multinomial-probit over all 27 classes

Composite Covariance

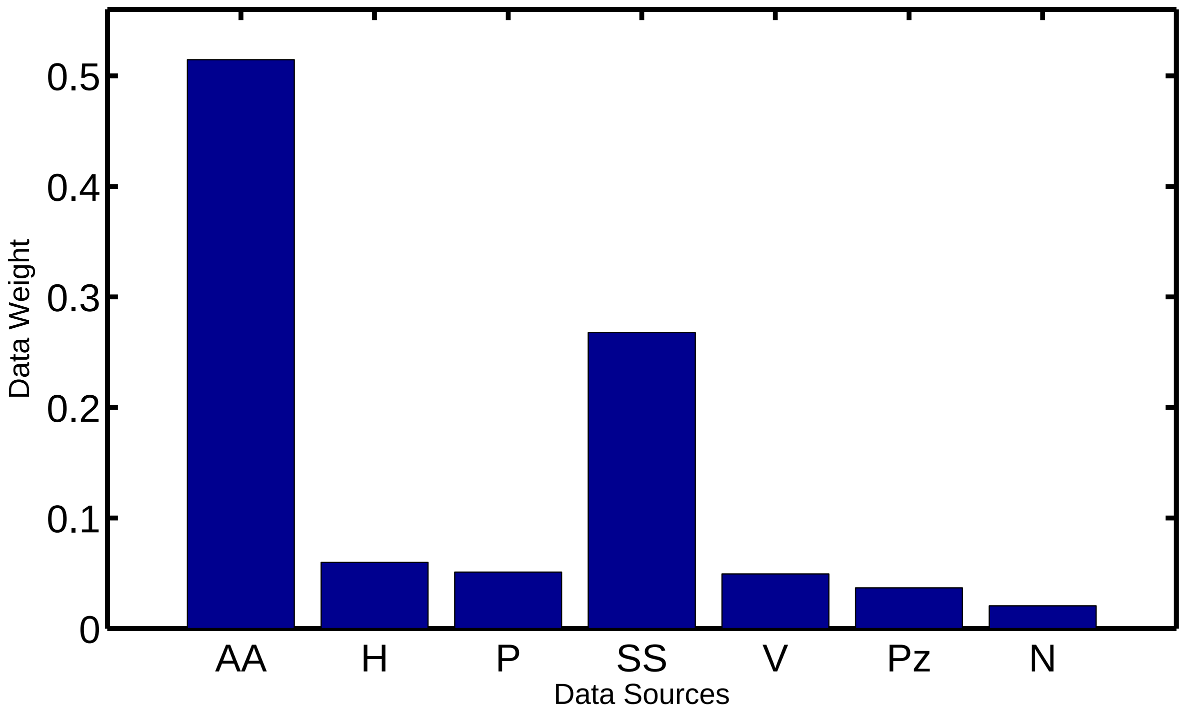
- Six datasets (AAC, SS, H, P, Pz, V) of D&D employed also include one random *noise* dataset
- Seven Gram matrices (RBF and inner-products) available, define Dirichlet prior on $\beta_1^2, \cdots, \beta_{\mathcal{J}}^2$ & Gamma on Dirichlet mean
- Run variational Bayes routine with multinomial-probit over all 27 classes
- Consider achievable performance over each individual dataset and combination *learned* plus product and Sum posterior combinations

Composite Covariance



Composite Covariance

UNIVERSITY
of
GLASGOW



Composite Covariance

- Recognition of handwritten digits '0' to '9'

Composite Covariance

UNIVERSITY *of* GLASGOW

- Recognition of handwritten digits '0' to '9'
- Four representations based on Zernike moments (47), Karhunen-Loeve coefficients (64), pixel averages (240), Fourier coefficients (76)

Composite Covariance

UNIVERSITY
of
GLASGOW

- Recognition of handwritten digits '0' to '9'
- Four representations based on Zernike moments (47), Karhunen-Loeve coefficients (64), pixel averages (240), Fourier coefficients (76)
- Previously employed in Tax *et al* comparing Sum & Product combinations of classifiers

Composite Covariance

UNIVERSITY *of* GLASGOW

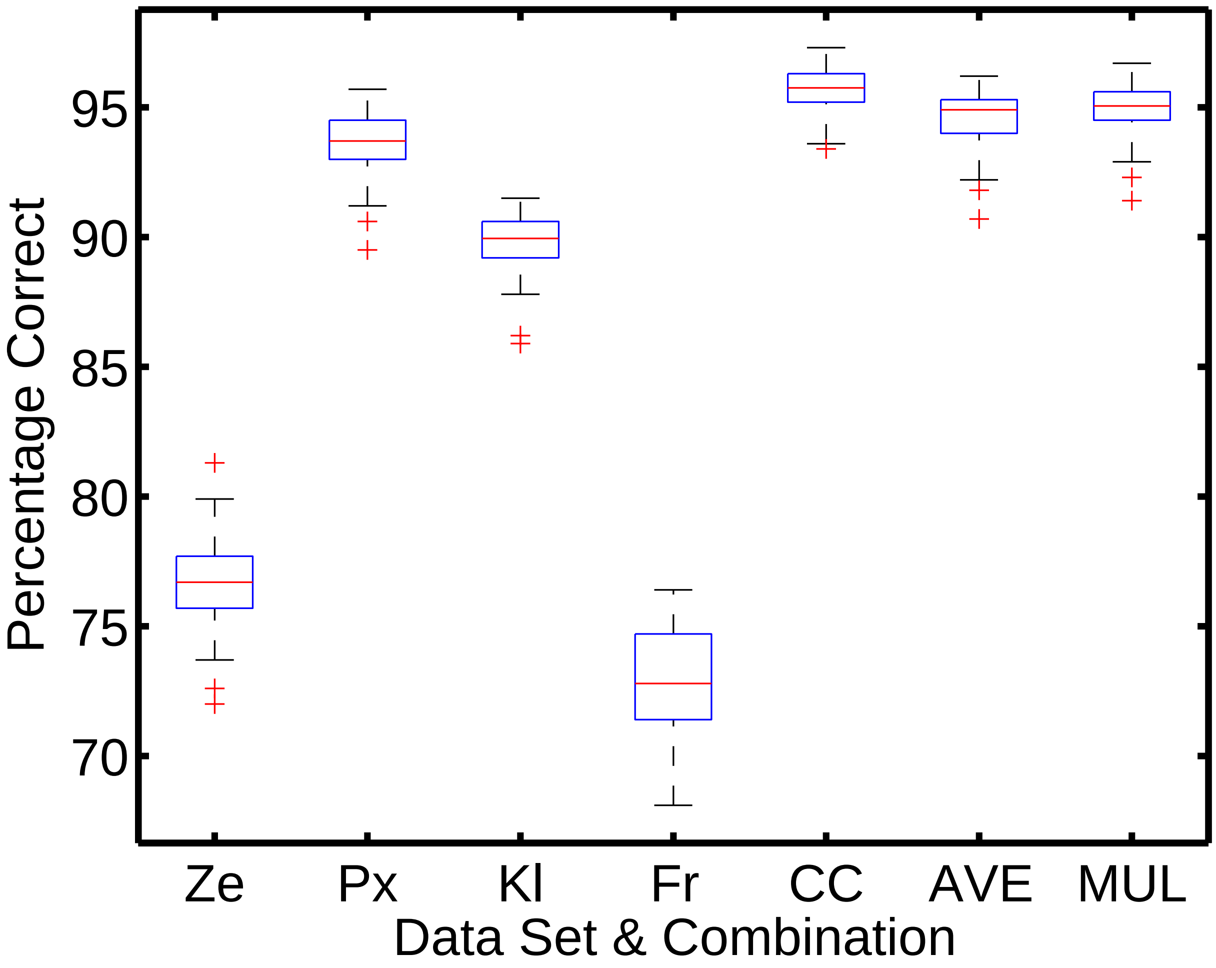
- Recognition of handwritten digits '0' to '9'
- Four representations based on Zernike moments (47), Karhunen-Loeve coefficients (64), pixel averages (240), Fourier coefficients (76)
- Previously employed in Tax *et al* comparing Sum & Product combinations of classifiers
- In sample size of 200 characters, test size 1800 characters

Composite Covariance

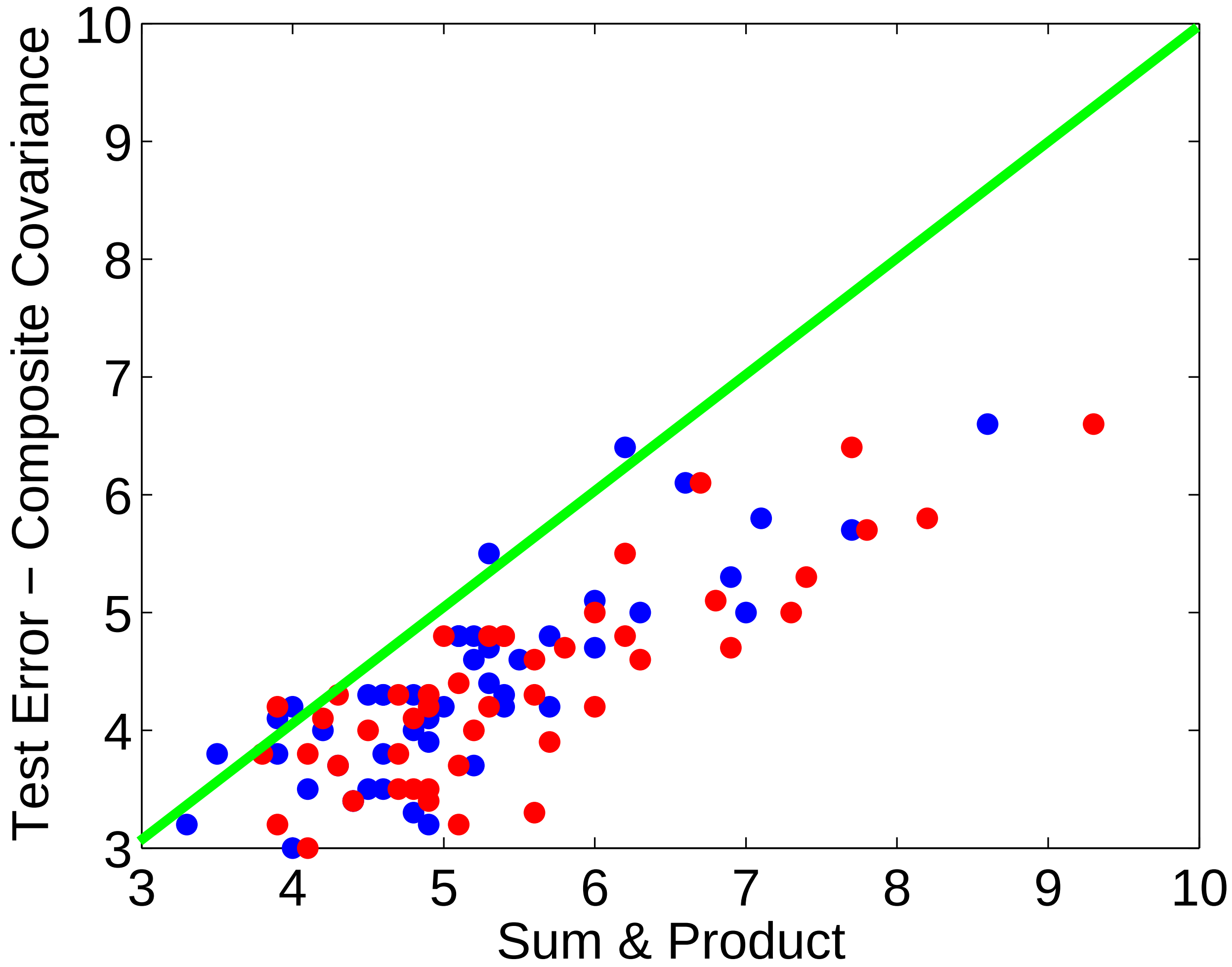
UNIVERSITY *of* GLASGOW

- Recognition of handwritten digits '0' to '9'
- Four representations based on Zernike moments (47), Karhunen-Loeve coefficients (64), pixel averages (240), Fourier coefficients (76)
- Previously employed in Tax *et al* comparing Sum & Product combinations of classifiers
- In sample size of 200 characters, test size 1800 characters
- Repeated train & test split resampling to compare single and combination schemes

Composite Covariance



Composite Covariance



Conclusions

- Integration of data within classification setting

Conclusions

UNIVERSITY
of
GLASGOW

- Integration of data within classification setting
- Bayesian perspective adopted & non-parametric classification achieved with GP's

Conclusions

UNIVERSITY
of
GLASGOW

- Integration of data within classification setting
- Bayesian perspective adopted & non-parametric classification achieved with GP's
- Efficient approximate inference methods developed for general multi-class setting

Conclusions

UNIVERSITY of GLASGOW

- Integration of data within classification setting
- Bayesian perspective adopted & non-parametric classification achieved with GP's
- Efficient approximate inference methods developed for general multi-class setting
- Inferring linear combination of covariance functions to integrate possibly heterogeneous feature representations

Conclusions

UNIVERSITY *of* GLASGOW

- Integration of data within classification setting
- Bayesian perspective adopted & non-parametric classification achieved with GP's
- Efficient approximate inference methods developed for general multi-class setting
- Inferring linear combination of covariance functions to integrate possibly heterogeneous feature representations
- Shown to provide superior predictive classification than standard Sum & Product combination rules

Conclusions

UNIVERSITY *of* GLASGOW

- Integration of data within classification setting
- Bayesian perspective adopted & non-parametric classification achieved with GP's
- Efficient approximate inference methods developed for general multi-class setting
- Inferring linear combination of covariance functions to integrate possibly heterogeneous feature representations
- Shown to provide superior predictive classification than standard Sum & Product combination rules
- Achieved state-of-art performance on difficult protein-fold prediction problem without recourse to heavy engineering and tuning of classifier settings.

Conclusions

UNIVERSITY
of
GLASGOW

- Bayesian classification over multiple classes employing GPs - analytically intractable

Conclusions

UNIVERSITY of GLASGOW

- Bayesian classification over multiple classes employing GPs - analytically intractable
- Approximations as alternatives to full MCMC - limited to Laplace

Conclusions

UNIVERSITY
of
GLASGOW

- Bayesian classification over multiple classes employing GPs - analytically intractable
- Approximations as alternatives to full MCMC - limited to Laplace
- Multinomial-logit likelihood inappropriate for variational approximations

Conclusions

- Bayesian classification over multiple classes employing GPs - analytically intractable
- Approximations as alternatives to full MCMC - limited to Laplace
- Multinomial-logit likelihood inappropriate for variational approximations
- Exploiting data augmentation trick (Albert & Chib, 1993) multinomial-probit likelihood provides *nice* solution to GP multi-class problem

Conclusions

UNIVERSITY
of
GLASGOW

- Bayesian classification over multiple classes employing GPs - analytically intractable
- Approximations as alternatives to full MCMC - limited to Laplace
- Multinomial-logit likelihood inappropriate for variational approximations
- Exploiting data augmentation trick (Albert & Chib, 1993) multinomial-probit likelihood provides *nice* solution to GP multi-class problem
- Statistical coupling of GP variables via posterior means maintains simple *a priori* factored structure *a posteriori*

Conclusions

- Computational scaling favourable, linear in number of classes, cubic in number of samples

Conclusions

UNIVERSITY *of* GLASGOW

- Computational scaling favourable, linear in number of classes, cubic in number of samples
- Online Bayesian estimation reduces to linear scaling in number of samples and classes

Conclusions

UNIVERSITY
of
GLASGOW

- Computational scaling favourable, linear in number of classes, cubic in number of samples
- Online Bayesian estimation reduces to linear scaling in number of samples and classes
- Empirical comparison with MCMC indicates predictive likelihood response, over range of hyper-parameters, better preserved under VB than Laplace approximation

Conclusions

UNIVERSITY
of
GLASGOW

- Computational scaling favourable, linear in number of classes, cubic in number of samples
- Online Bayesian estimation reduces to linear scaling in number of samples and classes
- Empirical comparison with MCMC indicates predictive likelihood response, over range of hyper-parameters, better preserved under VB than Laplace approximation
- Variational approximation provides computationally economic alternative to MCMC

Conclusions

UNIVERSITY *of* GLASGOW

- Computational scaling favourable, linear in number of classes, cubic in number of samples
- Online Bayesian estimation reduces to linear scaling in number of samples and classes
- Empirical comparison with MCMC indicates predictive likelihood response, over range of hyper-parameters, better preserved under VB than Laplace approximation
- Variational approximation provides computationally economic alternative to MCMC
- Integrating heterogeneous data via *kernel learning* available for free

Acknowledgments & Ref

- EPSRC Grants GR/R55184/02 & EP/C010620/1, MRC Discipline Hopping Award

Acknowledgments & Ref

UNIVERSITY
of
GLASGOW

- EPSRC Grants GR/R55184/02 & EP/C010620/1, MRC Discipline Hopping Award

- Girolami, M., Rogers, S., Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Computation, MIT Press. Vol. 18, Nos. 8, pp 1790-1817.

Acknowledgments & Ref

UNIVERSITY
of
GLASGOW

- EPSRC Grants GR/R55184/02 & EP/C010620/1, MRC Discipline Hopping Award
- Girolami, M., Rogers, S., Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Computation, MIT Press. Vol. 18, Nos. 8, pp 1790-1817.
- Girolami, M. Zhong, M., Data Integration for Classification Problems Employing Gaussian Process Priors, to appear, Advances in Neural Information Processing Systems 19, 2007.

Acknowledgments & Ref

UNIVERSITY *of* GLASGOW

- EPSRC Grants GR/R55184/02 & EP/C010620/1, MRC Discipline Hopping Award

- Girolami, M., Rogers, S., Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Computation, MIT Press. Vol. 18, Nos. 8, pp 1790-1817.
- Girolami, M. Zhong, M., Data Integration for Classification Problems Employing Gaussian Process Priors, to appear, Advances in Neural Information Processing Systems 19, 2007.
- `www.dcs.gla.ac.uk/people/personal/girolami/pubs_2005/VBGP/index.htm`