UNIVERSITY OF CAMBRIDGE

Latent Gaussian Processes for Distribution Estimation of Multivariate Categorical Data

Yarin Gal • Yutian Chen • Zoubin Ghahramani

yg279@cam.ac.uk

Distribution Estimation

UNIVERSITY OF CAMBRIDGE

- Distribution estimation of categorical data $\{y_n \mid n = 1, ..., N\}$ with $y_n \in \{1, ..., K\}$ is easy.
- But learning a distribution

$$P(\mathbf{y})$$

from vectors of categorical data,

$$\{\mathbf{y}_n = (y_{n1}, ..., y_{nD}) \mid n = 1, ..., N\}$$

with

$$y_{nd} \in \{1, ..., K\},$$

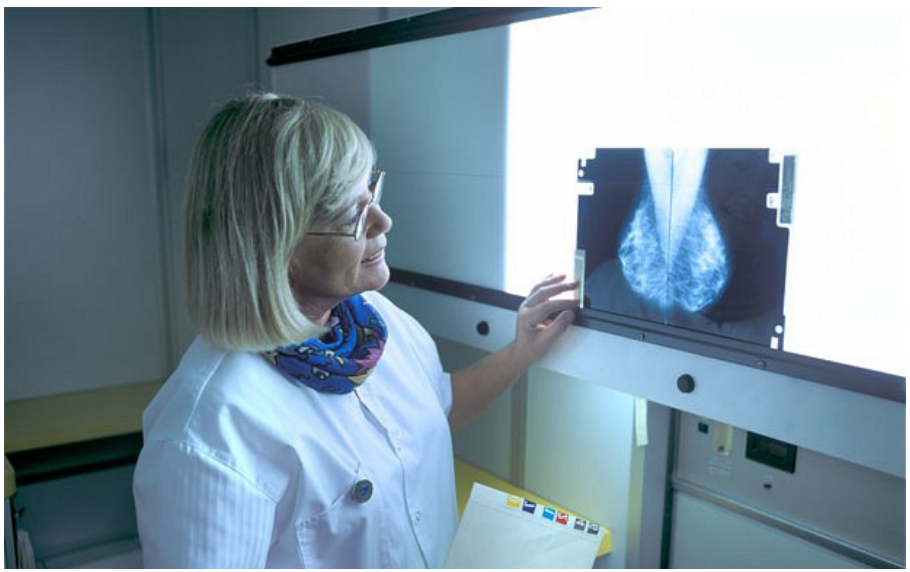
is much more difficult.

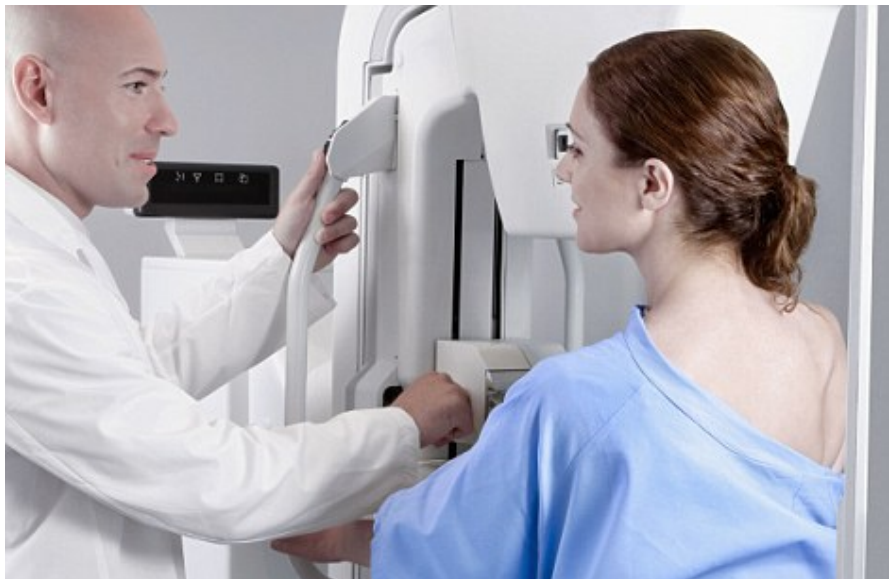
Multivariate Categorical Data

UNIVERSITY OF CAMBRIDGE

Multivariate Categorical Data

UNIVERSITY OF CAMBRIDGE





Example: Wisconsin Breast Cancer

UNIVERSITY OF CAMBRIDGE

Sample code	Thickness	Unif. Cell Size	Unif. Cell Shape	Marginal Adhesion	Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1000025	5	1	1	1	2	1	3	1	1	Benign
1002945	5	4	4	5	7	10	3	2	1	Benign
1015425	3	1	1	1	2	2	3	1	1	Benign
1016277	6	8	8	1	3	4	3	7	1	Benign
1017023	4	1	1	3	2	1	3	1	1	Benign
1017122	8	10	10	8	7	10	9	7	1	Malignant
1018099	1	1	1	1	2	10	3	1	1	Benign
1018561	2	1	2	1	2	1	3	1	1	Benign
1033078	2	1	1	1	2	1	1	1	5	Benign
1033078	4	2	1	1	2	1	2	1	1	Benign
1035283	1	1	1	1	1	1	3	1	1	Benign
1036172	2	1	1	1	2	1	2	1	1	Benign
1041801	5	3	3	3	2	3	4	4	1	Malignant
1043999	1	1	1	1	2	3	3	1	1	Benign
1044572	8	7	5	10	7	9	5	5	4	Malignant
1047630	7	4	6	4	6	1	4	3	1	Malignant
1048672	4	1	1	1	2	1	2	1	1	Benign
1049815	4	1	1	1	2	1	3	1	1	Benign
1050670	10	7	7	6	4	10	4	1	2	Malignant
1050718	6	1	1	1	2	1	3	1	1	Benign
1054590	7	3	2	10	5	10	5	4	4	Malignant
1054593	10	5	5	3	6	7	7	10	1	Malignant
1056784	3	1	1	1	2	1	2	1	1	Benign
1057013	8	4	5	1	2	?	7	3	1	Malignant
1059552	1	1	1	1	2	1	3	1	1	Benign
1065726	5	2	3	4	2	7	3	6	1	Malignant
1066373	3	2	1	1	1	1	2	1	1	Benign
1066979	5	1	1	1	2	1	2	1	1	Benign
1067444	2	1	1	1	2	1	2	1	1	Benign
1070935	1	1	3	1	2	1	1	1	1	Benign
1070935	3	1	1	1	1	1	2	1	1	Benign
1071760	2	1	1	1	2	1	3	1	1	Benign
1072179	10	7	7	3	8	5	7	4	3	Malignant

683 patients, 2×10^9 possible configurations

Categorical Latent Gaussian Process

UNIVERSITY OF CAMBRIDGE

We define our model as

Continuous latent space of patients $\mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_x^2 I)$

Dist. over K dim. functions generating weights $\mathbf{f}_d \sim \mathrm{GP}(\cdot)$

$y_{nd} \overset{\text{iid}}{\sim} \mathrm{Softmax}(\mathbf{f}_d(\mathbf{x}_n))$ test result

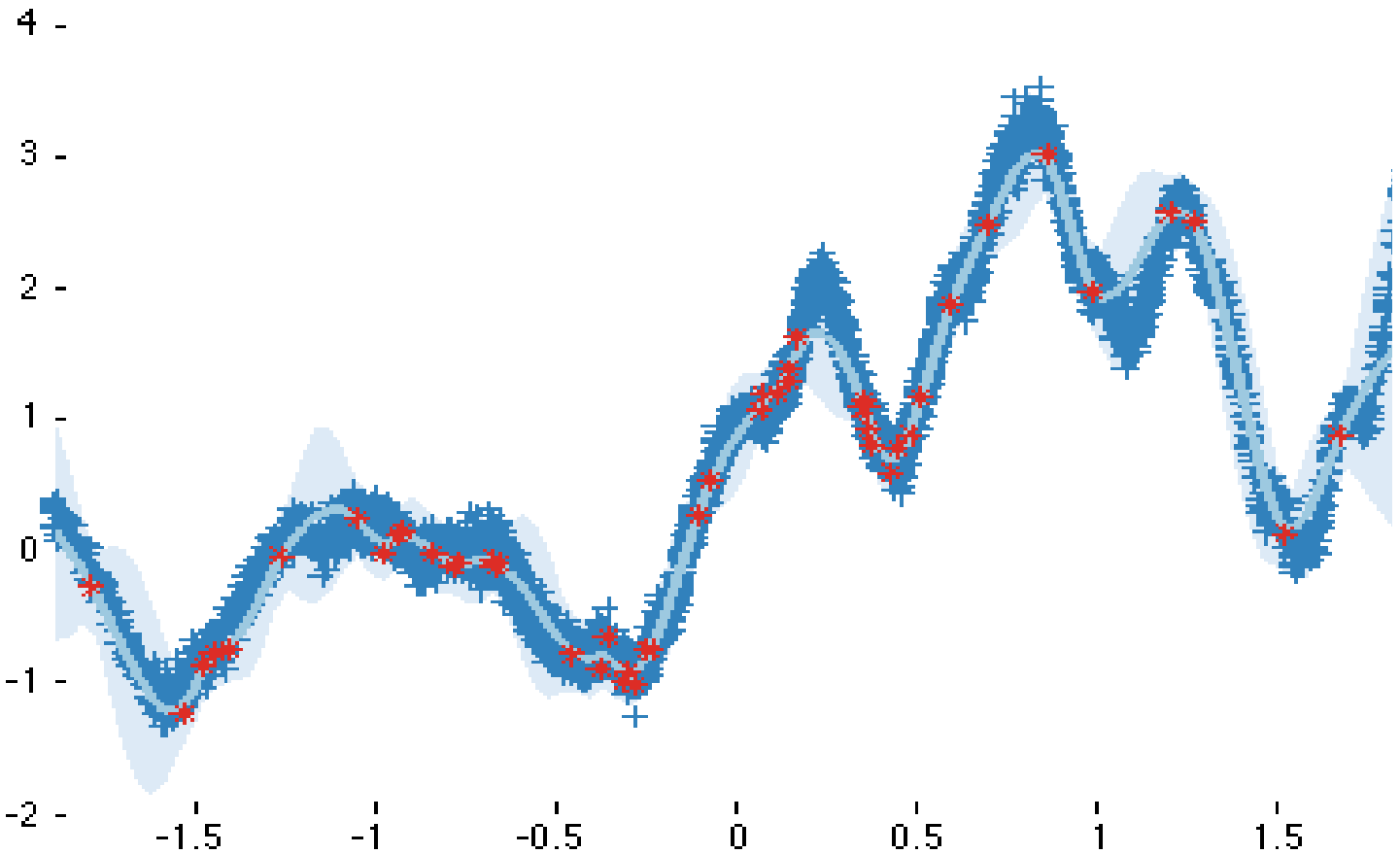
$\mathbf{y}_n = (y_{n1}, \ldots, y_{nD})$ medical assessment

N patients $\mathbf{x}_n$, D categorical tests, each with K possible test results

Sparse Gaussian Processes

UNIVERSITY OF CAMBRIDGE

- A small number of inducing points support a distribution over functions



- In dark blue are the data points, in red are the inducing points

Sparse Gaussian Processes

UNIVERSITY OF CAMBRIDGE

- Given sufficient statistics $\{Z_m, u_m\}_{m=1}^M$,
  - Define K_{MM} as the kernel evaluated at points (Z_m),
  - Define $K_{Mn} = K_{nM}^T$ as the kernel evaluated between points (Z_m) and point $\mathbf{x}_n$,
- A sparse Gaussian process is defined through a conditional Gaussian distribution by

$$f_n \sim \mathcal{N}(\mathbf{a}_n^T \mathbf{u}, b_n)$$

with

$$\mathbf{a}_n = \mathbf{K}_{MM}^{-1} \mathbf{K}_{Mn}, \quad b_n = K_{nn} - \mathbf{K}_{nM} \mathbf{K}_{MM}^{-1} \mathbf{K}_{Mn},$$

- or equivalently,

$$f_n = \mathbf{a}_n^T \mathbf{u} + \sqrt{b_n} \varepsilon_n^{(f)} \qquad \varepsilon_n^{(f)} \sim \mathcal{N}(0, 1).$$

Generative Model

UNIVERSITY OF CAMBRIDGE

- The resulting generative model, with kernel $\mathbf{K}(\cdot,\cdot)$, is

$$
\begin{aligned}
\mathbf{x}_n &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_x^2 I) && \text{patient } n \\
(\mathbf{u}_{dk})_{k=1}^{K} &\sim \mathrm{GP}(0, \mathbf{K}((Z_m)_{m=1}^{M})) && \text{inducing points} \\
f_{ndk} | \mathbf{x}_n, \mathbf{u}_{dk} &\sim \mathcal{N}(\mathbf{a}_n^T \mathbf{u}_{dk}, b_n) && \text{weight of test result } k \\
y_{nd} &\overset{\text{iid}}{\sim} \mathrm{Softmax}(f_{nd1}, \ldots, f_{ndK}) && \text{examination } d \\
\mathbf{y}_n &= (y_{n1}, \ldots, y_{nD}) && \text{medical assessment}
\end{aligned}
$$

- We want to find the posterior over $\mathbf{X}$ and $\mathbf{U}$:

$$p(\mathbf{X}, \mathbf{U} | \mathbf{Y}).$$

- Our marginal log-likelihood is intractable.
- We lower bound the log evidence with a variational approximate posterior $q(\mathbf{X}, \mathbf{F}, \mathbf{U}) = q(\mathbf{X})q(\mathbf{U})p(\mathbf{F}|\mathbf{X}, \mathbf{U})$, with

$$q(\mathbf{X}) = \prod_{n=1}^{N} \prod_{i=1}^{Q} \mathcal{N}(x_{ni}|m_{ni}, s_{ni}^2),$$

$$q(\mathbf{U}) = \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}(\mathbf{u}_{dk}|\boldsymbol{\mu}_{dk}, L_d L_d^T)$$

- Such that...

$$q(\mathbf{X}, \mathbf{F}, \mathbf{U}) \approx p(\mathbf{X}, \mathbf{F}, \mathbf{U}|\mathbf{Y}).$$

Re-parametrising the Model

UNIVERSITY OF CAMBRIDGE

This is equivalent to...

$$x_{ni} = m_{ni} + s_{ni}\varepsilon_{ni}^{(x)} \qquad \varepsilon_{ni}^{(x)} \sim \mathcal{N}(0,1)$$
$$\mathbf{u}_{dk} = \boldsymbol{\mu}_{dk} + \mathbf{L}_d\boldsymbol{\varepsilon}_{dk}^{(u)} \qquad \boldsymbol{\varepsilon}_{dk}^{(u)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$$
$$f_{ndk} = \mathbf{a}_n^T\mathbf{u}_{dk} + \sqrt{b_n}\varepsilon_{ndk}^{(f)} \qquad \varepsilon_{ndk}^{(f)} \sim \mathcal{N}(0,1)$$

Then,

$$\log p(\mathbf{Y}) \geq -\text{KL divergence terms}$$
$$+\sum_{n=1}^{N}\sum_{d=1}^{D}\mathbb{E}_{\varepsilon_n^{(x)},\varepsilon_d^{(u)},\varepsilon_{nd}^{(f)}}\Bigg[$$
$$\log \text{Softmax}\left(\mathbf{y}_{nd}\middle|\mathbf{f}_{nd}\left(\varepsilon_{nd}^{(f)}, \mathbf{U}_d(\varepsilon_d^{(u)}), \mathbf{x}_n(\varepsilon_n^{(x)})\right)\right)\Bigg].$$

Monte Carlo integration

UNIVERSITY OF CAMBRIDGE

- Monte Carlo integration approximates the likelihood obtaining noisy gradients:

$$\mathbb{E}_{\varepsilon_n^{(x)},\varepsilon_d^{(u)},\varepsilon_{nd}^{(f)}} \log \text{Softmax}\left(\cdot\right)$$
$$\approx \frac{1}{T}\sum_{i=1}^{T} \log \text{Softmax}\left(\mathbf{y}_{nd}\middle|\mathbf{f}_{nd}\left(\varepsilon_{nd,i}^{(f)}, \mathbf{U}_d(\varepsilon_{d,i}^{(u)}), \mathbf{x}_n(\varepsilon_{n,i}^{(x)})\right)\right)$$

  with $\varepsilon_{\cdot,i}^{(x)}, \varepsilon_{\cdot,i}^{(u)}, \varepsilon_{\cdot,i}^{(f)} \sim \mathcal{N}(0,1)$, independent of the parameters.
- Adaptive learning-rate stochastic optimisation is used to optimise the noisy objective wrt the parameters of the variational distribution.

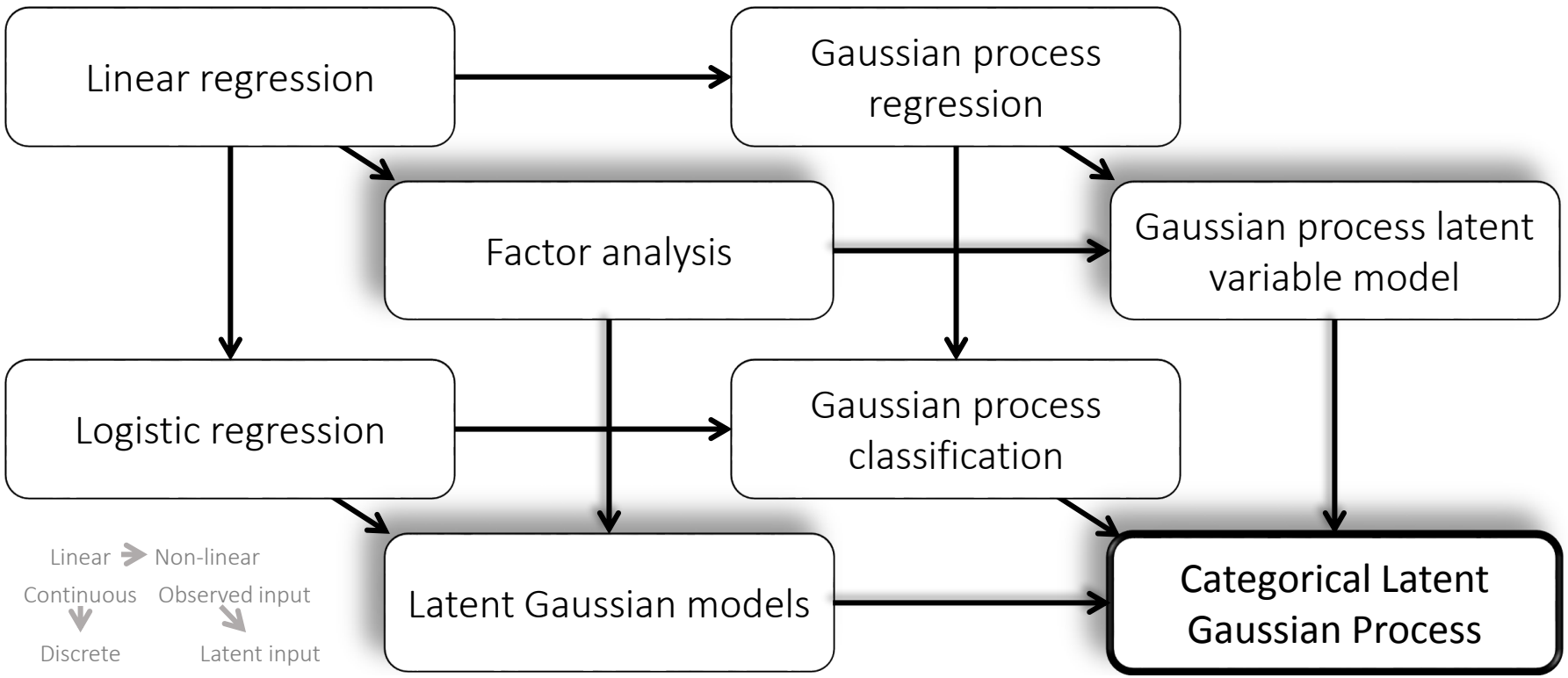
Categorical Latent Gaussian Process

UNIVERSITY OF CAMBRIDGE

- Symbolic differentiation gives simple code:

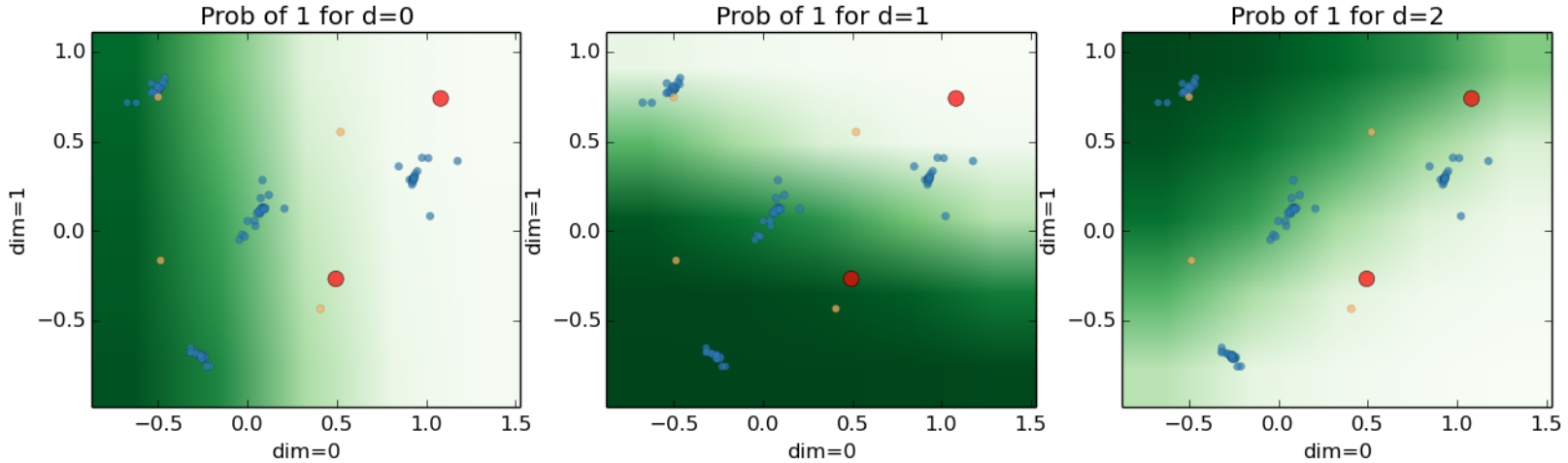
```
import theano.tensor as T
X = m + s * randn(N, Q)
U = mu + L.dot(randn(M, K))
Kmm, Kmn, Knn = RBF(sf2, l, Z), RBF(sf2, l, Z, X), RBFnn(sf2, l, X)
KmmInv = sT.matrix_inverse(Kmm)
A = KmmInv.dot(Kmn)
B = Knn - T.sum(Kmn * KmmInv.dot(Kmn), 0)
F = A.T.dot(U) + B[:,None]**0.5 * randn(N, K)
S = T.nnet.softmax(F)
KL_U, KL_X = get_KL_U(), get_KL_X()
LS = T.sum(T.log(T.sum(Y * S, 1))) - KL_U - KL_X
LS_func = theano.function(['''inputs'''], LS)
dLS_dm = theano.function(['''inputs'''], T.grad(LS, m)) # and others
# ... and run RMS-PROP
```

Relations to Existing Models

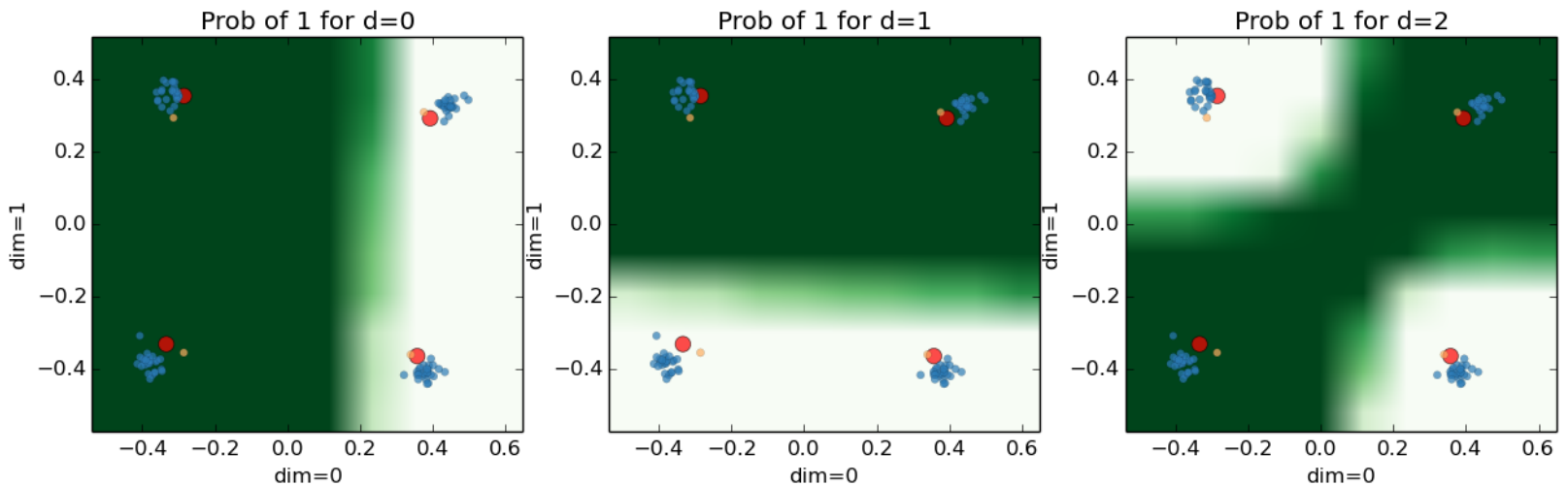


Multi-modal Distributions

UNIVERSITY OF CAMBRIDGE



(a) (Linear) Latent Gaussian Model



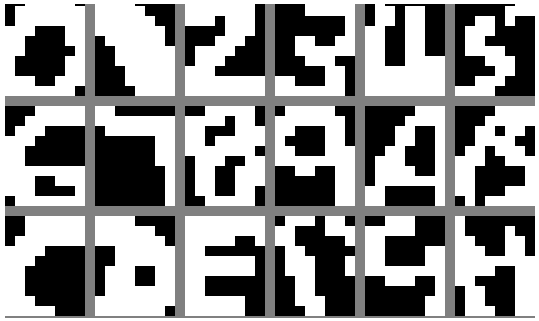
(b) Categorical Latent Gaussian Process

$p(\mathbf{y}_d = 1|\mathbf{x})$ **as a function of x, for** $d = 0, 1, 2$

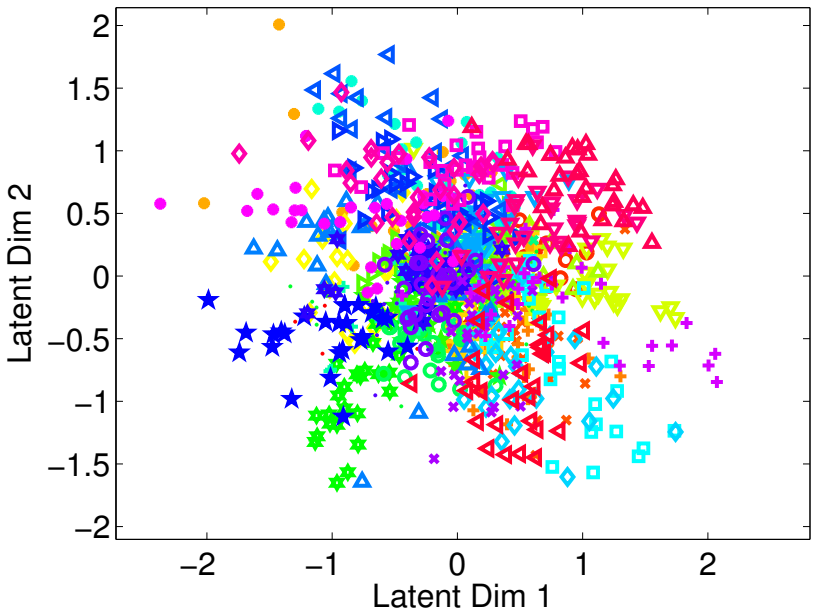
(first, second, and third digits in the XOR triplet left to right)

Data Visualisation

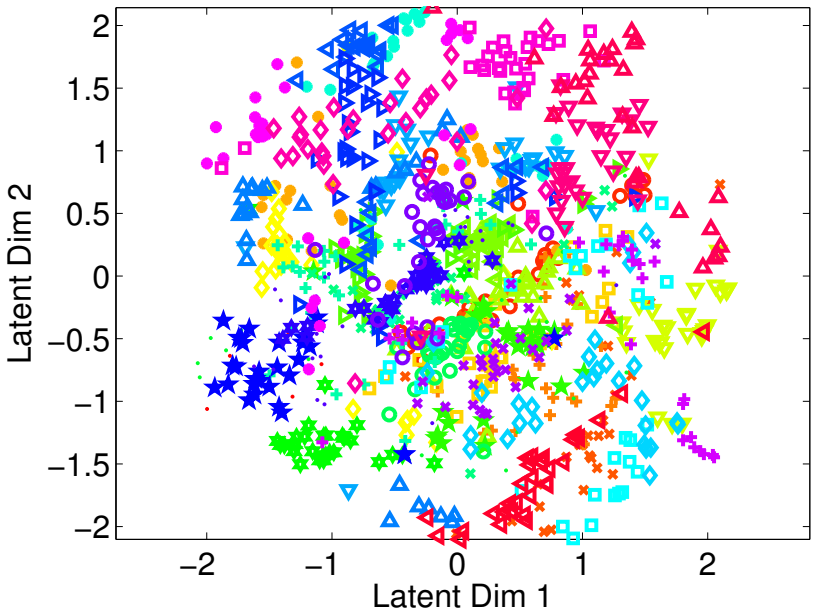
UNIVERSITY OF CAMBRIDGE



(a) Example alphadigits



(b) (Linear) Latent Gaussian Model latent space



(c) Categorical Latent Gaussian Process latent space

Data Imputation

UNIVERSITY OF CAMBRIDGE

Terror Warning Effects on Political Attitude (START Terrorism Data Archive - 17 categorical variables with 5-6 values)

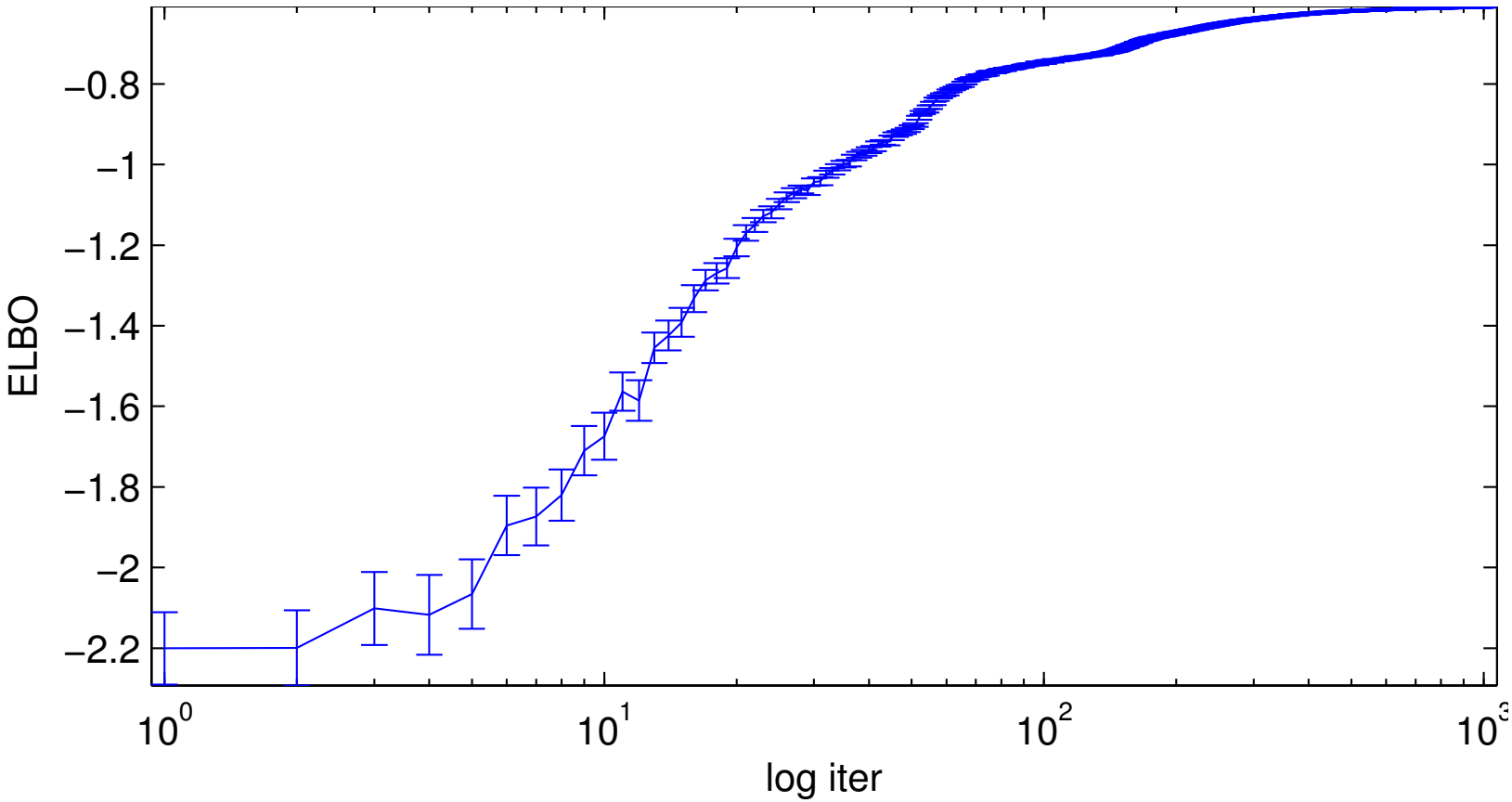
Uniform	Multinomial	(Linear) Latent Gaussian Model	**Categorical Latent Gaussian Process**
2.50	2.20	3.07	**2.11**

Wisconsin Breast cancer (9 categorical variables with 10 values)

Uniform	Multinomial	(Linear) Latent Gaussian Model	**Categorical Latent Gaussian Process**
8.68	4.41	3.57 ± 0.208	$\mathbf{2.86 \pm 0.119}$

Test perplexity predicting randomly missing values

Inference Robustness



Lower bound and MC standard deviation per iteration (on log scale) for the Alphadigits dataset.

Scaling Up!

UNIVERSITY OF CAMBRIDGE

- Why not scale the model up?
  - We recently showed that sparse Gaussian processes can be scaled well to millions of data points [Gal, van der Wilk, Rasmussen, 2014]
  - Mini-batch GP optimisation [Hensman el at. 2014]
- Recognition models with sparse Gaussian processes
  - Integrate over unobserved variables
  - Handle missing data
- Mixed data
  - Using link functions alternative to the Softmax
  - Continuous variables, Positives, Ordinals

Currently running experiments!

Scaling Up!

UNIVERSITY OF CAMBRIDGE

- Why not scale the model up?
  - We recently showed that sparse Gaussian processes can be scaled well to millions of data points [Gal, van der Wilk, Rasmussen, 2014]
  - Mini-batch GP optimisation [Hensman el at. 2014]
- Recognition models with sparse Gaussian processes
  - Integrate over unobserved variables
  - Handle missing data
- Mixed data
  - Using link functions alternative to the Softmax
  - Continuous variables, Positives, Ordinals

Currently running experiments!

Scaling Up!

UNIVERSITY OF CAMBRIDGE

- Why not scale the model up?
  - We recently showed that sparse Gaussian processes can be scaled well to millions of data points [Gal, van der Wilk, Rasmussen, 2014]
  - Mini-batch GP optimisation [Hensman el at. 2014]
- Recognition models with sparse Gaussian processes
  - Integrate over unobserved variables
  - Handle missing data
- Mixed data
  - Using link functions alternative to the Softmax
  - Continuous variables, Positives, Ordinals

Currently running experiments!

Thank you

UNIVERSITY OF CAMBRIDGE