VARIATIONAL INFERENCE

in Gaussian process models

James Hensman

Approximate Inference workshop, NIPS 2015

Lancaster Univeristy

Alex Matthews
Cambridge Univeristy

Nicolo Fusi
Microsoft Research

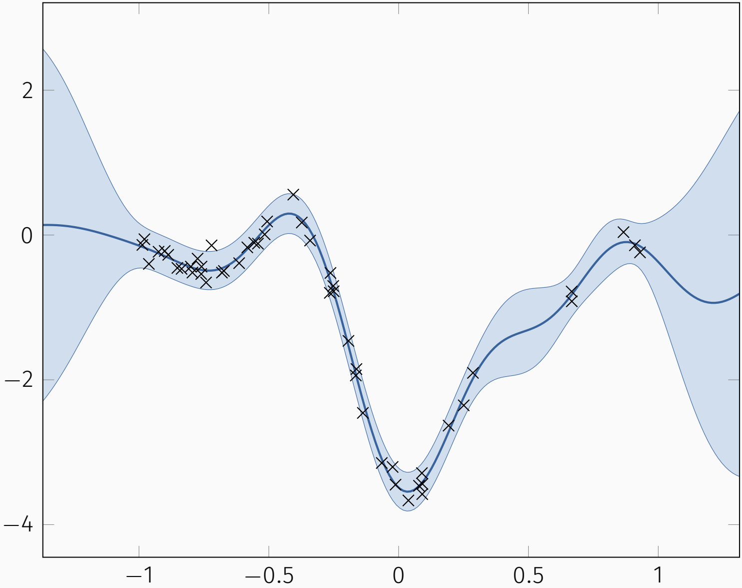
Maurizio Filippone
Eurecom

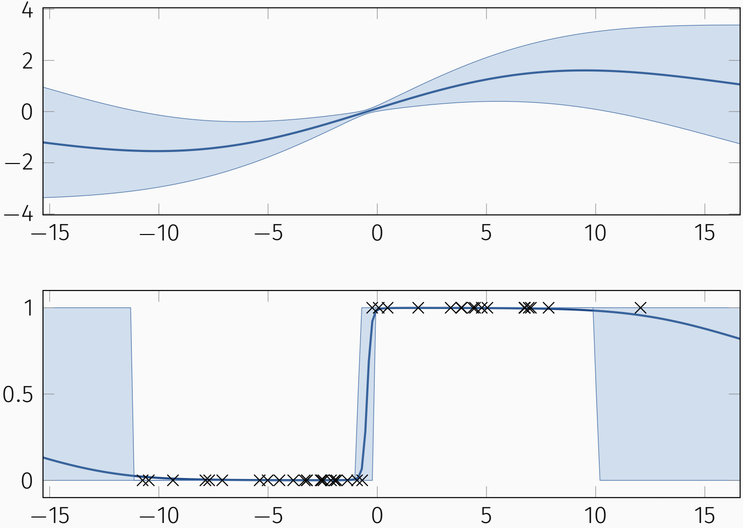
Rich Turner
Cambridge Univeristy

Neil D. Lawrence
Sheffield Univeristy

Zoubin Ghahramani
Cambridge Univeristy

WHAT CAN GAUSSIAN PROCESSES DO?




4
2
0
−2
−4
−15
−10
−5
0
5
10
15
1
0.5
0
−15
−10
−5
0
5
10
15

A unified view of variational GP aproximations

- Deals with non-Gaussian posterior
- Deals with $\mathcal{O}(n^3)$ complexity (sparse)
- The variational distribution contains a (conditionally) Gaussian process

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

with:

$$\mathrm{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$$

$$y_i | f_i \sim \mathrm{Po}(y_i \mid e^{f_i}) \qquad \text{or} \qquad \mathrm{Bin}(y_i \mid \sigma(f_i)) \qquad \text{or} \ldots$$

DEALING WITH NON-CONJUGACY

- Local variational bounds (classification only) [1]
- Expectation Propagation [2]
- For classification, EP > VB [3]
- Variational methods need only 2N parameters [4]
- VB methods can be fast too! [5]
- VB can be applied to lots of different likelihoods [6]

[1] MN Gibbs, DJC MacKay - Variational Gaussian process classifiers - IEEE TNN 2000

[2] Minka, T. P. A family of algorithms for approximate Bayesian inference. Doctoral dissertation, MIT - 2001

[3] H Nickisch, CE Rasmussen - Approximations for binary Gaussian process classification - JMLR 2008

[4] M. Opper and C. Archambeau – The variational Gaussian approximation revisited - Neural comp. 2009

[5] E Khan, S Mohamed, KP Murphy - Fast Bayesian inference for non-conjugate Gaussian process regression- NIPS 2012

[6] Nguyen and Bonilla – Automated variational inference for Gaussian process models - NIPS 2014

DEALING WITH $O(n^3)$ COMPLEXITY

- Subset-of-data methods[7] [8] hence 'sparse'.
- Pseudo-inputs introduced [9]
- A unifying view brings several ideas together [10]
- Variational approach [11] makes for better placement of pseudo/inducing points
- Variational approach can be optimized with SVI [12]

[7] AJ Smola, P Bartlett - Sparse greedy Gaussian process regression - NIPS 2001

[8] M Seeger, C Williams - Fast forward selection to speed up sparse Gaussian process regression - AISTATS 2003

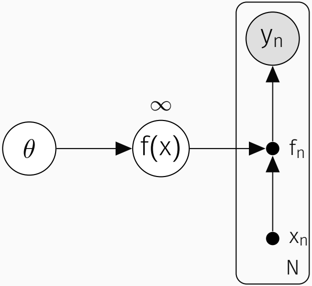
[9] E Snelson, Z Ghahramani - Sparse Gaussian processes using pseudo-inputs - NIPS 2005

[10] J Quiñonero-Candela, CE Rasmussen - A unifying view of sparse approximate Gaussian process regression - JMLR 2005

[11] M. Titsias - Variational learning of inducing variables in sparse Gaussian processes - AISTATS 2009

[12] J. Hensman, N. Fusi and N. Lawrence - Gaussian Processes for Big Data - UAI 2013

A GRAPHICAL MODEL FOR GAUSSIAN PROCESSES

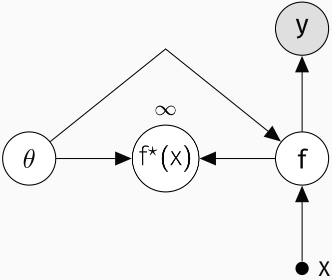


$$\theta \sim p(\theta)$$

$$f(x) \sim \mathcal{GP}(0, k(x, x'; \theta))$$

$$\mathbf{f} = [f(x_1), f(x_2) \ldots f(x_n)]^\top$$

$$y_n \sim p(y_n \mid f(x_n))$$

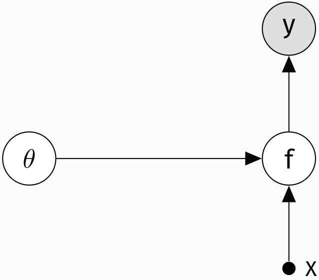


$$\theta \sim p(\theta)$$

$$\mathbf{f}|\theta \sim \mathcal{N}(0, \mathbf{K})$$

$$y_n \sim p(y_n \mid f(x_n))$$

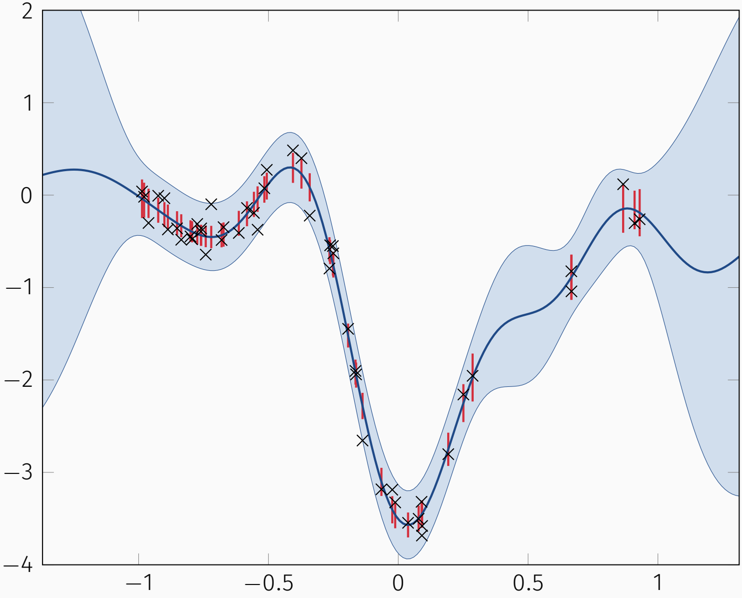
$$f^\star(x)|\mathbf{f}, \theta \sim \mathcal{GP}\big(\mathbf{a}(x)^\top \mathbf{f}, b(x, x')\big)$$



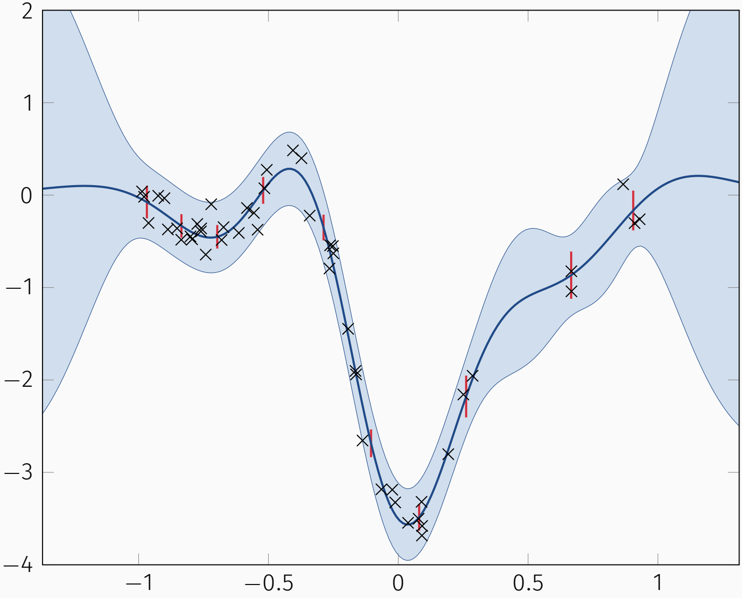
$$\theta \sim p(\theta)$$

$$\mathbf{f}|\theta \sim \mathcal{N}(0, \mathrm{K})$$

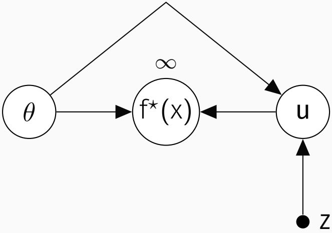
$$y_n \sim p(y_n \mid f(x_n))$$



2
1
0
−1
−2
−3
−4
−1
−0.5
0
0.5
1



2
1
0
−1
−2
−3
−4
−1
−0.5
0
0.5
1



$$\theta, \mathbf{u} \sim q(\theta, \mathbf{u})$$

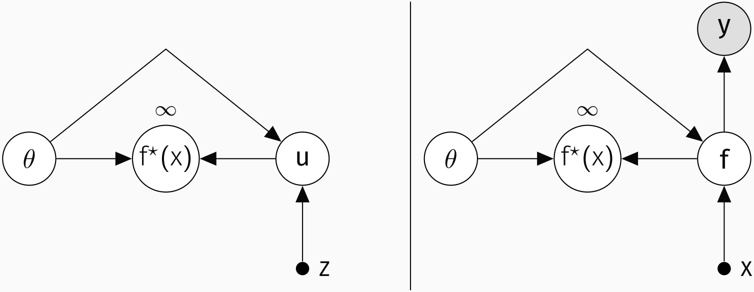
$$f^\star(x) \sim \mathcal{GP}\big(\mathbf{a}'(x)^\top \mathbf{u}, b'(x)\big)$$

KL DIVERGENCE BETWEEN GAUSSIAN PROCESSES?

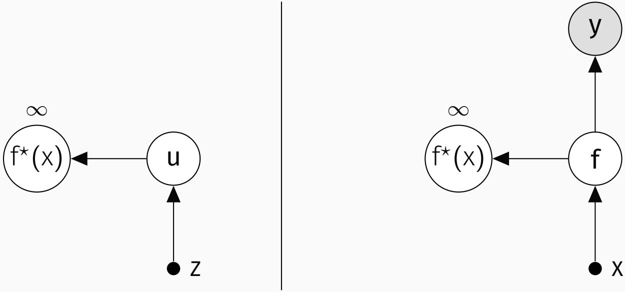
Intuitive version: $\mathbf{f}^\star$ is a really long vector containing all points of interest.

Rigorous version: Matthews et al.[13]

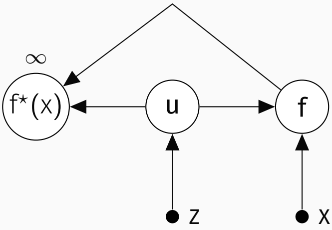
[13] On Sparse variational methods and the Kullback-Leibler divergence between stochastic processes http://arxiv.org/abs/1504.07027

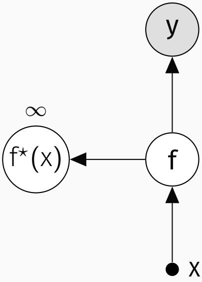

∞
θ
f⋆(x)
u
z
y
∞
θ
f⋆(x)
f
x

Let's ignore θ for now

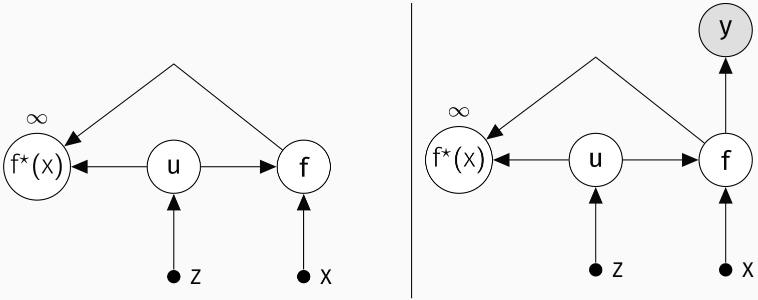


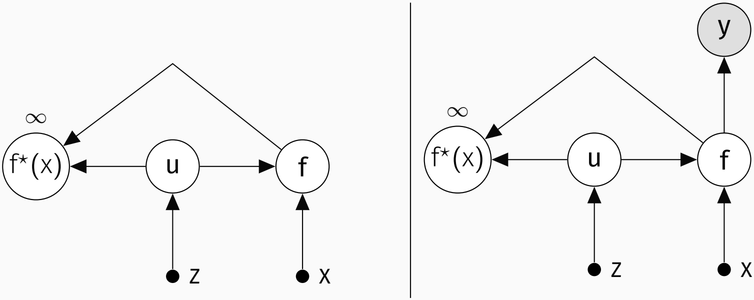
Where are the **f** in the approximation?



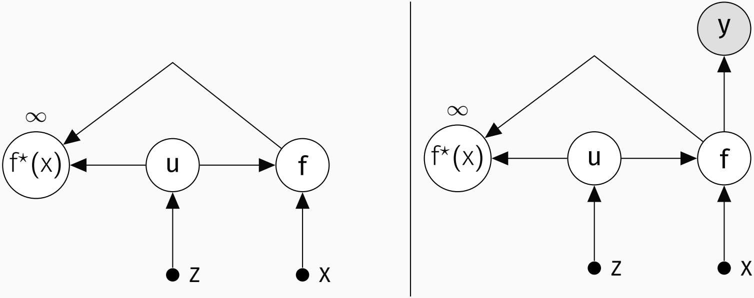


Where are the **u** in the model?

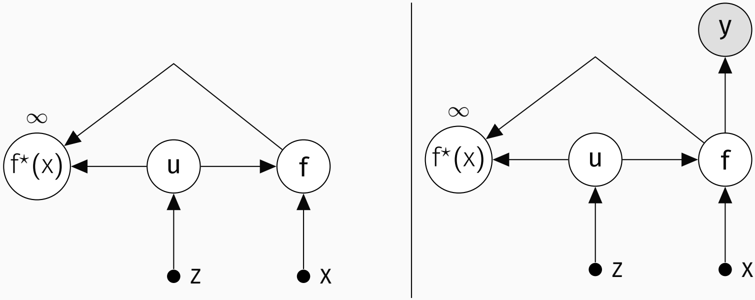




$$\text{ELBO} = \mathbb{E}_{q(f^\star, \mathbf{f}, \mathbf{u}, \theta)}\left[\log \frac{p(\mathbf{y}\,|\,\mathbf{f})p(\mathbf{f}\,|\,\mathbf{u},\theta)p(f^\star\,|\,\mathbf{f},\mathbf{u},\theta)p(\mathbf{u}\,|\,\theta)p(\theta)}{q(\mathbf{f}\,|\,\mathbf{u},\theta)q(f^\star\,|\,\mathbf{f},\mathbf{u},\theta)q(\mathbf{u}\,|\,\theta)q(\theta)}\right]$$



$$\text{ELBO} = \mathbb{E}_{q(\cancel{f^\star}, \mathbf{f}, \mathbf{u}, \theta)}\left[\log \frac{p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{u}, \theta)\, \cancel{p(f^\star \mid \mathbf{f}, \mathbf{u}, \theta)}\, p(\mathbf{u} \mid \theta)\, p(\theta)}{q(\mathbf{f} \mid \mathbf{u}, \theta)\, \cancel{q(f^\star \mid \mathbf{f}, \mathbf{u}, \theta)}\, q(\mathbf{u} \mid \theta)\, q(\theta)}\right]$$



$$\mathrm{ELBO} = \mathbb{E}_{q(\cancel{f^\star}, \mathbf{f}, \mathbf{u}, \theta)} \left[\log \frac{p(\mathbf{y} \mid \mathbf{f}) \cancel{p(\mathbf{f} \mid \mathbf{u}, \theta)} \cancel{p(f^\star \mid \mathbf{f}, \mathbf{u}, \theta)} p(\mathbf{u} \mid \theta) p(\theta)}{\cancel{q(\mathbf{f} \mid \mathbf{u}, \theta)} \cancel{q(f^\star \mid \mathbf{f}, \mathbf{u}, \theta)} q(\mathbf{u} \mid \theta) q(\theta)} \right]$$

Strategy 1: Gaussian [14]

Let $q(\mathbf{u}, \theta) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{L}\mathbf{L}^\top)\delta(\theta - \hat{\theta})$

Optimize wrt $\mathbf{m}, \mathbf{L}, \hat{\theta}$ (and $\mathbf{Z}$!)

Strategy 2: Free-form[15]

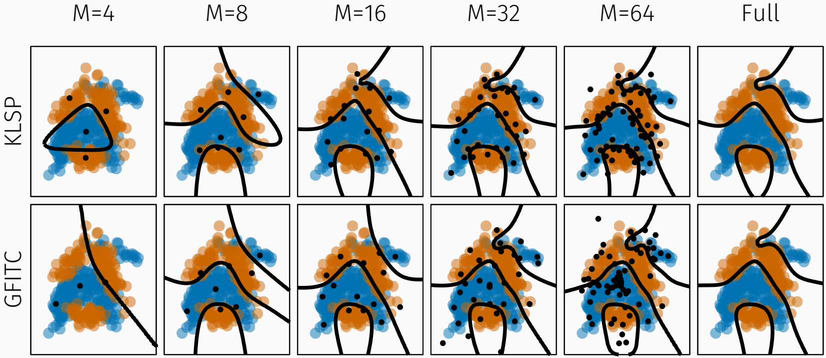
Given the limited size of $\mathbf{Z}$ (and thus $\mathbf{u}$), write down the optimal, intractable, form for $q(\mathbf{u}, \theta)$, and sample from it using HMC.

[14] J Hensman, A Matthews, Z Ghahramani - Scalable Variational Gaussian Process Classification - AISTATS 2015

[15] J Hensman, AGG Matthews, M Filippone - MCMC for Variationally Sparse Gaussian Processes - NIPS 2015

The objective function (which minimizes the KL between the q-process and the p-process) is

$$\mathcal{L} = \sum_i \mathbb{E}_{q(f_i)}[\log p(y_i|f_i)] - \text{KL}[q(u)||p(u)]$$


M=4
M=8
M=16
M=32
M=64
Full
KLSP
GFITC

HIGH DIMENSIONAL PROBLEMS

Left: three k-means centers used to initialize the inducing point positions. Center: the positions of the same inducing points after optimization. Right: difference.

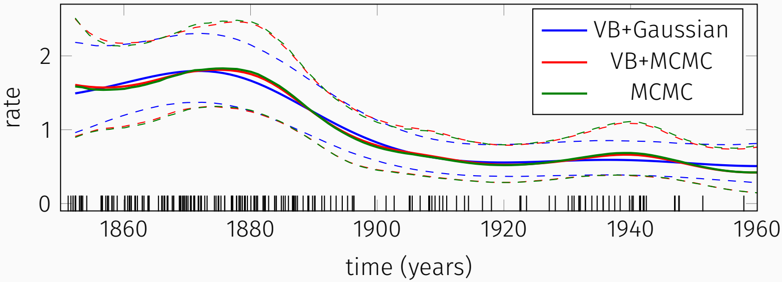
Data: N=60,000, D=784

Accuracy: 98.04%

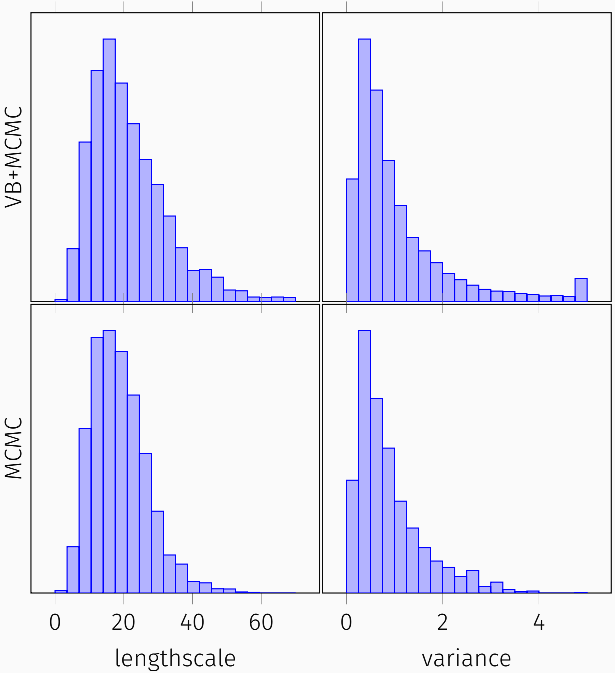
The ‘perfect’ distribution $\hat{q}(\mathbf{u}, \theta)$ which minimises the KL divergence (with no further restrictions) is

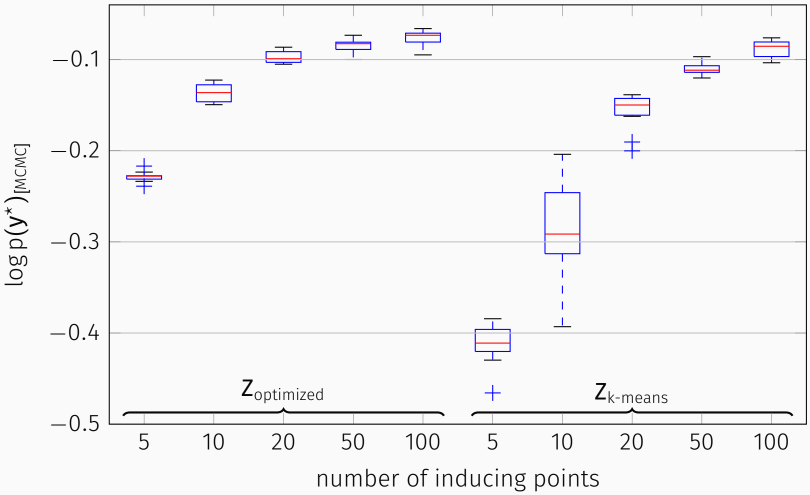
$$\log \hat{q}(\mathbf{u}, \theta) = \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{f})] + \log p(\mathbf{u}, \theta) + \text{const.}$$

Sampling $\hat{q}$ costs $\mathcal{O}(NM^2)$.



The posterior of the rates for the coal mining disaster data.


VB+MCMC
MCMC
0
20
40
60
lengthscale
0
2
4
variance


−0.1
−0.2
−0.3
−0.4
−0.5
$\log p(\mathbf{y}^\star)_{[\text{MCMC}]}$
$Z_{\text{optimized}}$
$Z_{\text{k-means}}$
5
10
20
50
100
5
10
20
50
100
number of inducing points

- Exact inference (Gaussian likelihood, $Z = X$)
- Subset-of-data methods (e.g. IVM [16])
- Inter-domain approximations [17]
- Black box likelihoods [18]
- Log Gaussian Cox processes [19]

[16] Lawrence, Seeger and Herbrich - The Informative Vector Machine - NIPS 2003

[17] Alvarez, Rosasco and Lawrence - Kernels for vecotr valued functions, a review - foundationa and trends in ML 2011

[18] Dezfouli and Bonilla - Gaussian Process Models with Black-Box Likelihoods - NIPS 2015

[19] Lloyd et al - Variational Inference for Gaussian Process Modulated Poisson Processes - ICML 2015

Thanks for listening