Automated Probabilistic Reasoning Via Variational Inference

Edwin V. Bonilla

January 15th, 2018

Bayesian Reasoning — Challenges

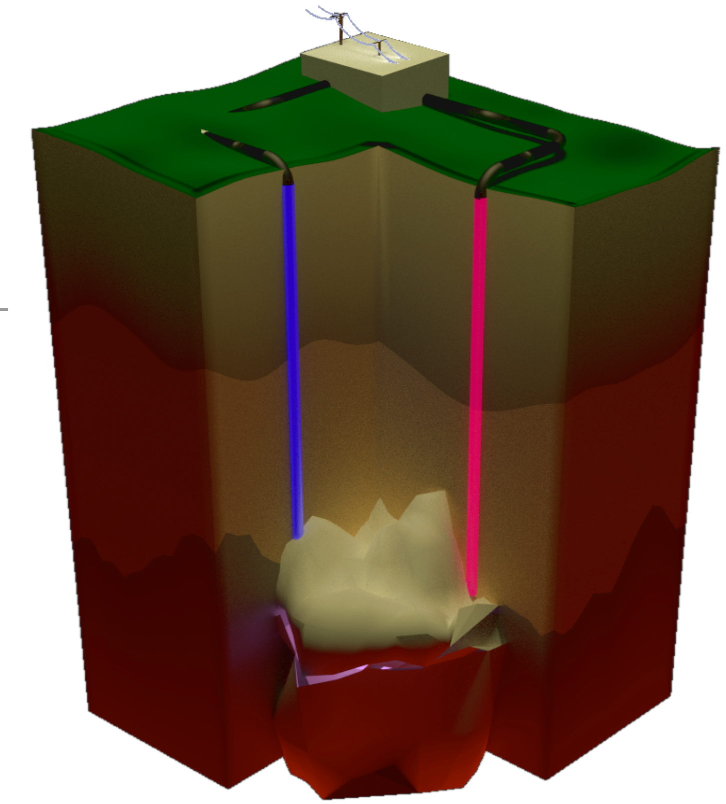
Prior over geology and rock properties

Forward model's likelihood

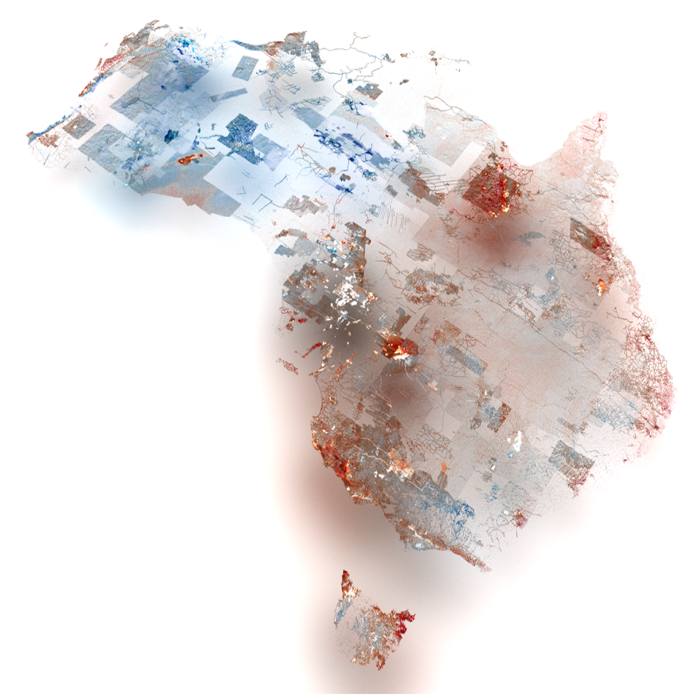
$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{f})p(\mathbf{y}|\mathbf{f})}{\int p(\mathbf{f})p(\mathbf{y}|\mathbf{f})d\mathbf{f}}$$

← Hard bit

Posterior geological model



$20 Million



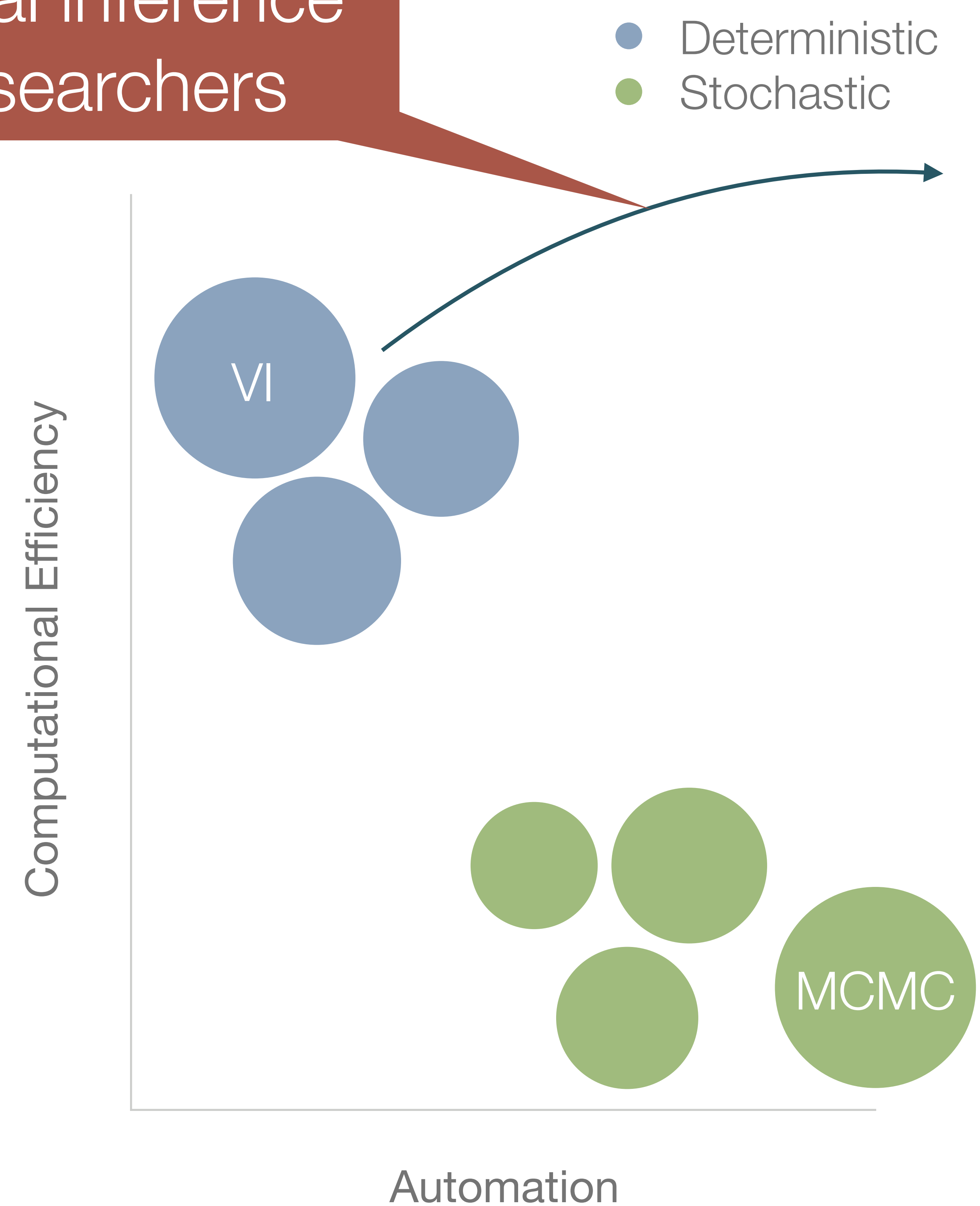
Surveys and explorations

- General likelihood models (non-linear fwd models)
- Large datasets

Goal: Build generic yet practical inference tools for practitioners and researchers

Automated Probabilistic Reasoning

- Other dimensions
  - Accuracy
  - Convergence



Gaussian Process Priors

Notoriously unscalable!

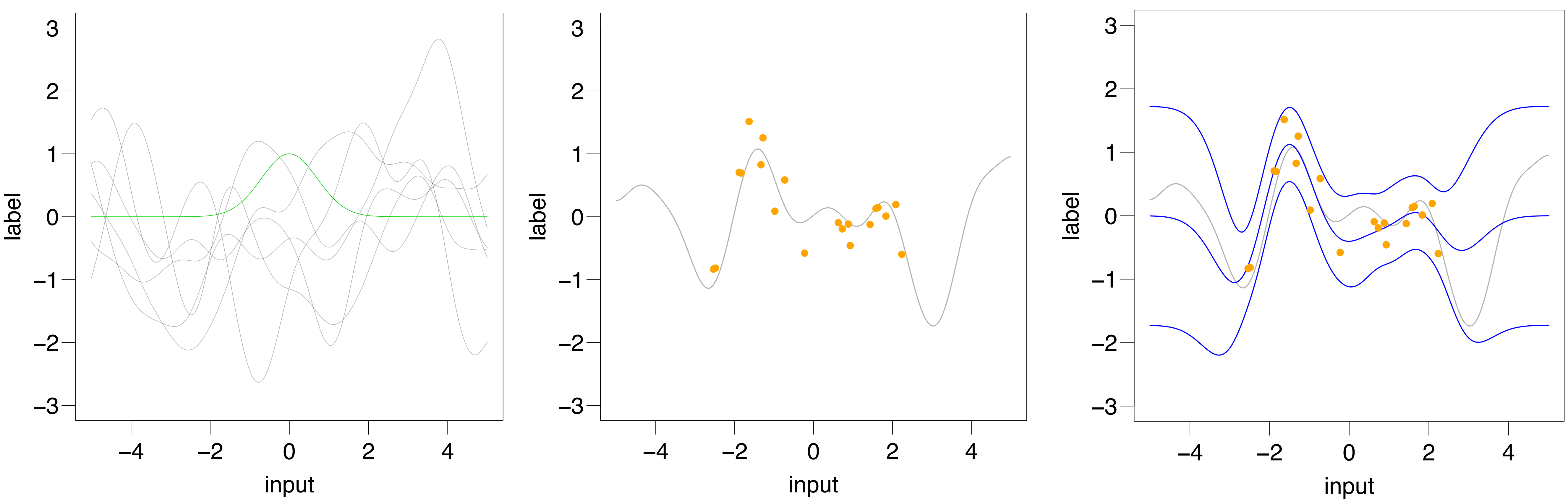
- Distribution over functions
- f($\mathbf{x}_1$), f($\mathbf{x}_2$), … f($\mathbf{x}_N$) follow a joint Gaussian distribution
  - Example: density field

Mean function

Covariance function

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}))$$

Hyper-parameters



Latent Gaussian Process Models (LGPMs)

- Supervised learning: $\mathcal{D} = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^{N}$

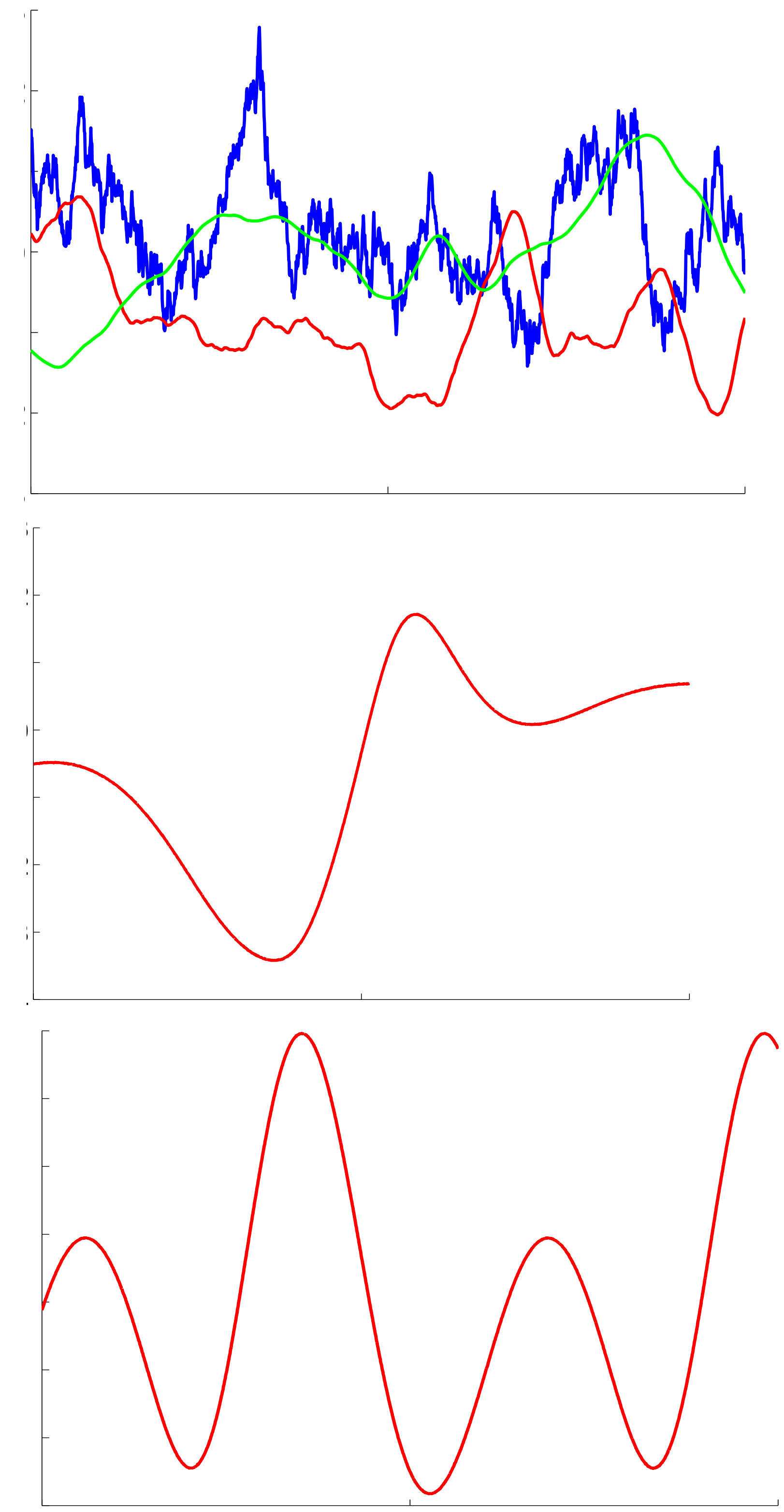
- Factorised prior over Q latent functions

$$f_j \sim \mathcal{GP}\left(0, \kappa_j(\cdot, \cdot)\right)$$

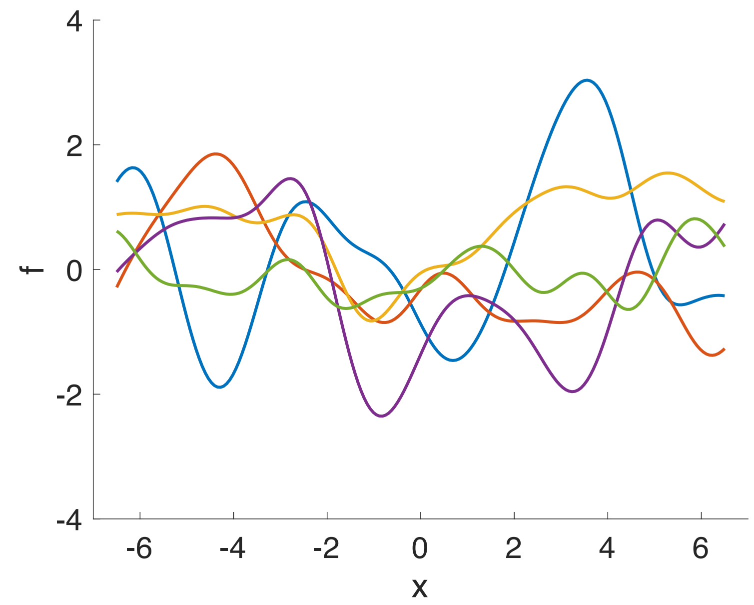
$$p(\mathbf{f}|\boldsymbol{\theta}) = \prod_{j=1}^{Q} \mathcal{N}(\mathbf{f}_{\cdot j}; \mathbf{0}, \mathbf{K}_j)$$

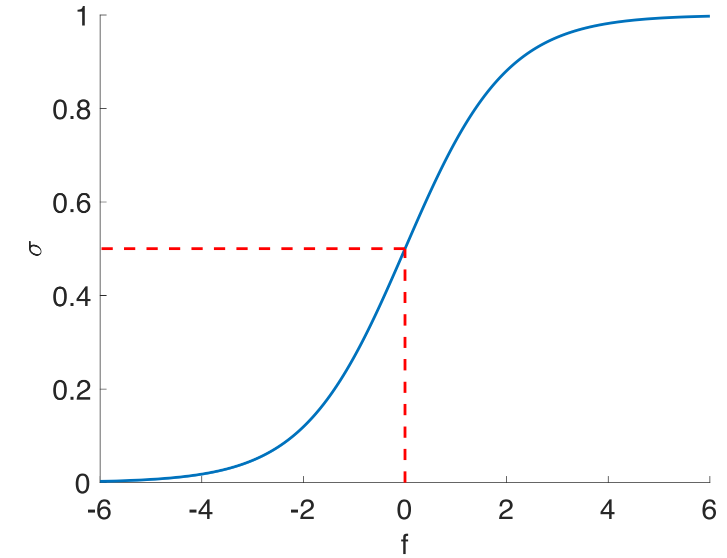
- Factorised likelihood over observations

$$p(\mathbf{y}|\mathbf{f}, \phi) = \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{f}_{n\cdot}, \phi)$$

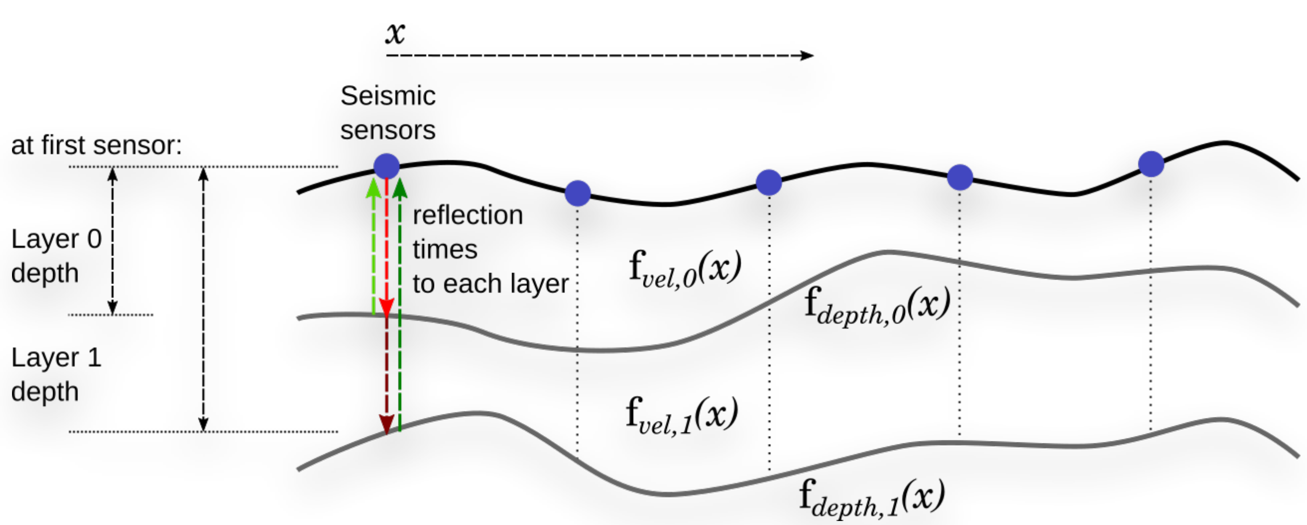


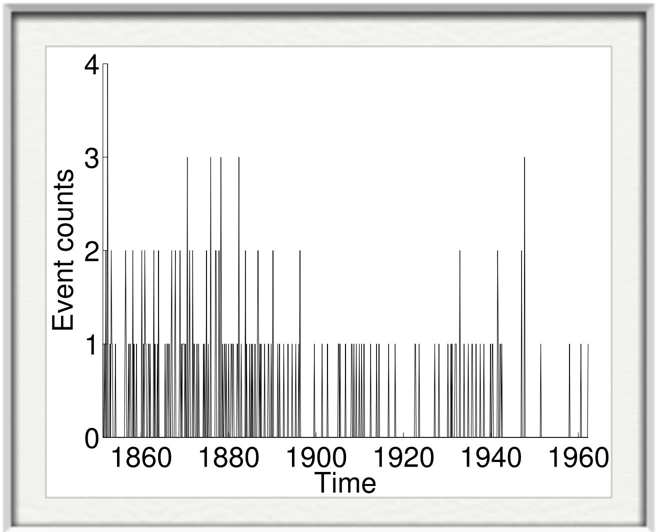
Examples of LGPMs





- Multi-class classification
  - Q classes, softmax likelihood
- Multi-output regression
- Inversion problems
- LGCPs for count data
- Access to ‘black-box’ likelihood

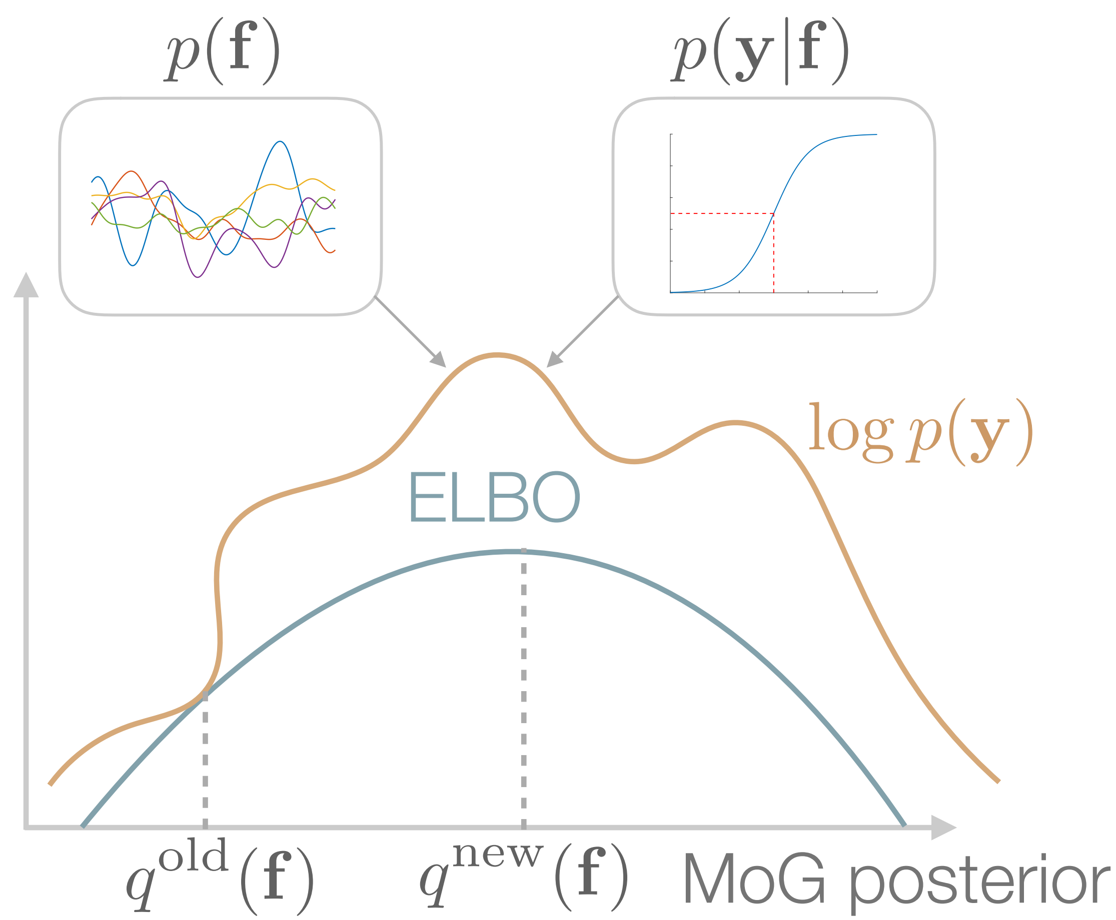




Solution 1.0: Automated VI for LGPMs

Nguyen and Bonilla (NIPS, 2014)

- ELBO = -KL + ELL
- KL
  - Analytical lower bound
  - Exact gradients
- ELL
  - Samples from univariate Gaussians
  - No explicit gradients required

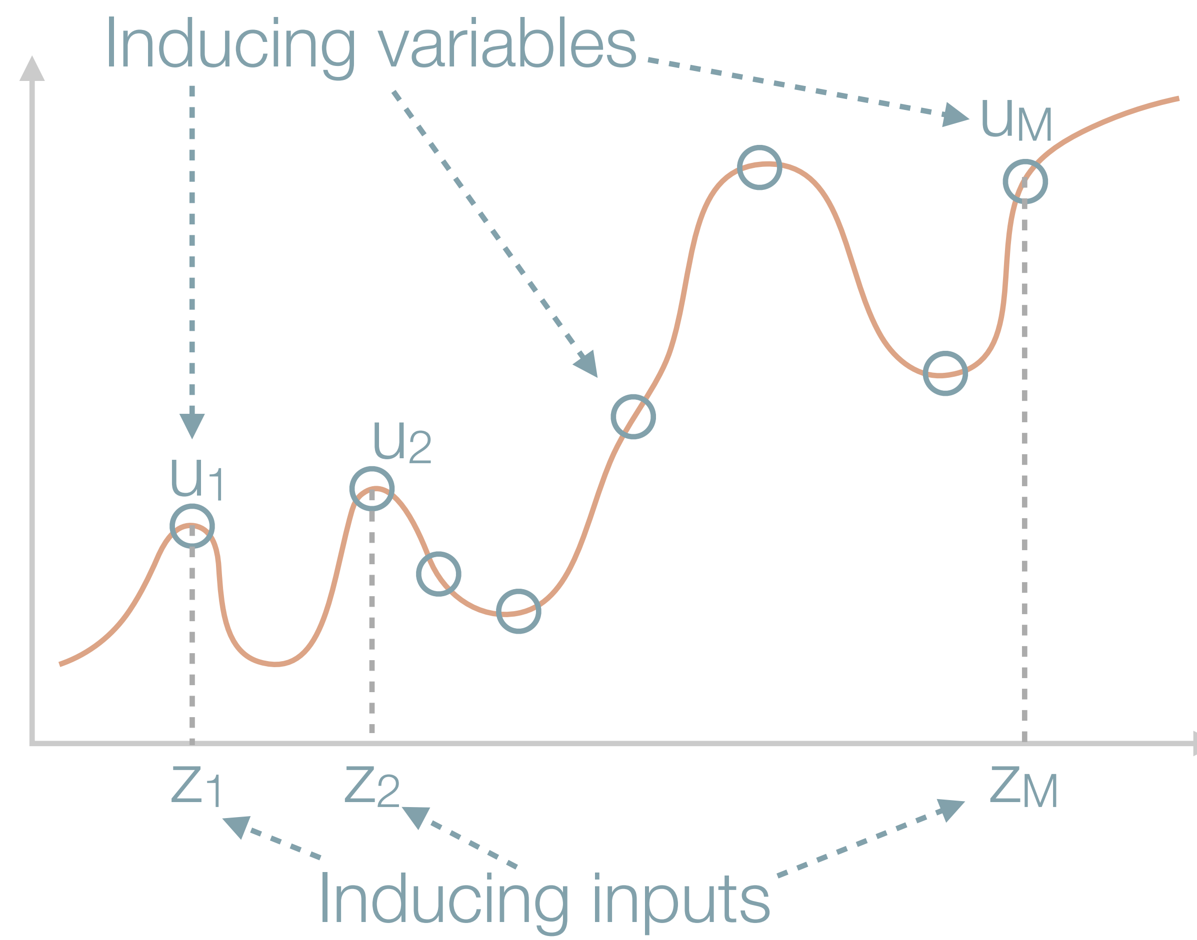


- ⋆ Efficient parameterisation
- ⋆ As good as hand-coded solutions
- ⋆ Orders of magnitude faster than MCMC

Solution 1.1: SAVIGP

Dezfouli and Bonilla (NIPS, 2015)

- Scalability through “sparsification”: M<< N
- Statistical efficiency
- Efficient parameterisation
- Control variates

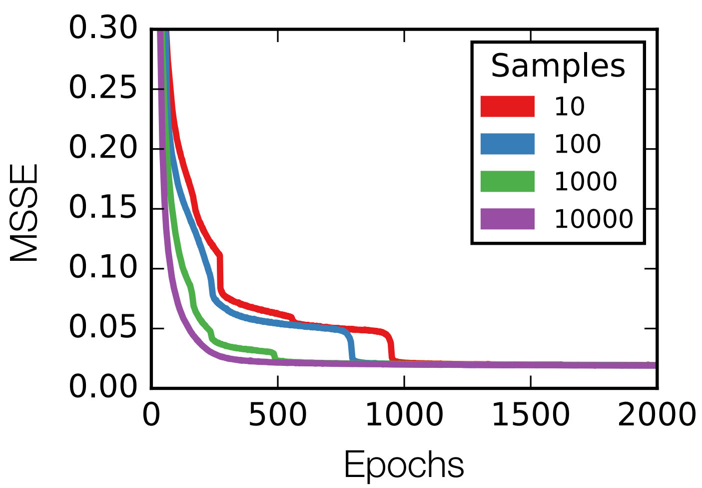


Large-scale inference for GP priors and general black-box likelihoods

Solution 1.2: AutoGP

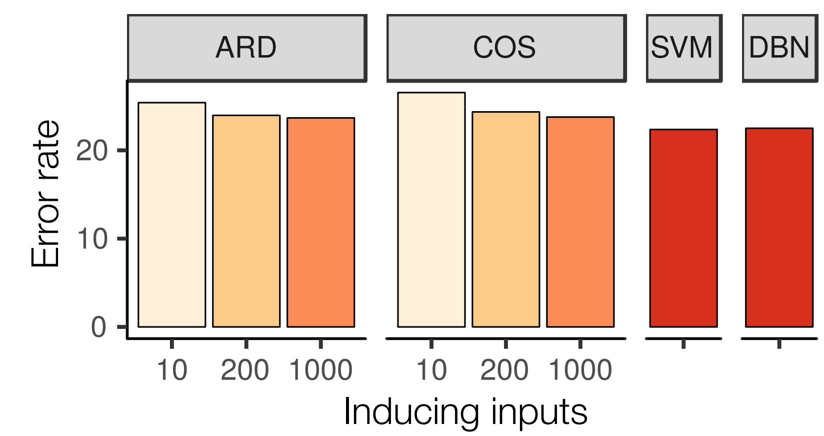
Krauth, Bonilla, Cutajar and Filippone (UAI, 2017)

- ⭑ Breaks error-barrier on MNIST for GP models
- ⭑ Unprecedented scale



Scalability & efficient computation

Low-variance gradient estimates



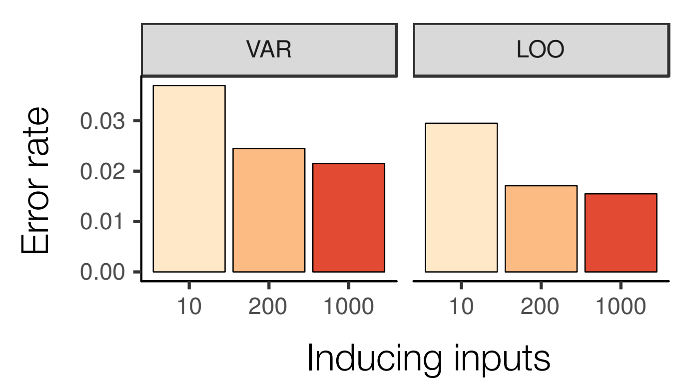
Well-targeted objective functions

Leave-one-out hyper-parameter learning

The holy trinity of machine learning

Representational power

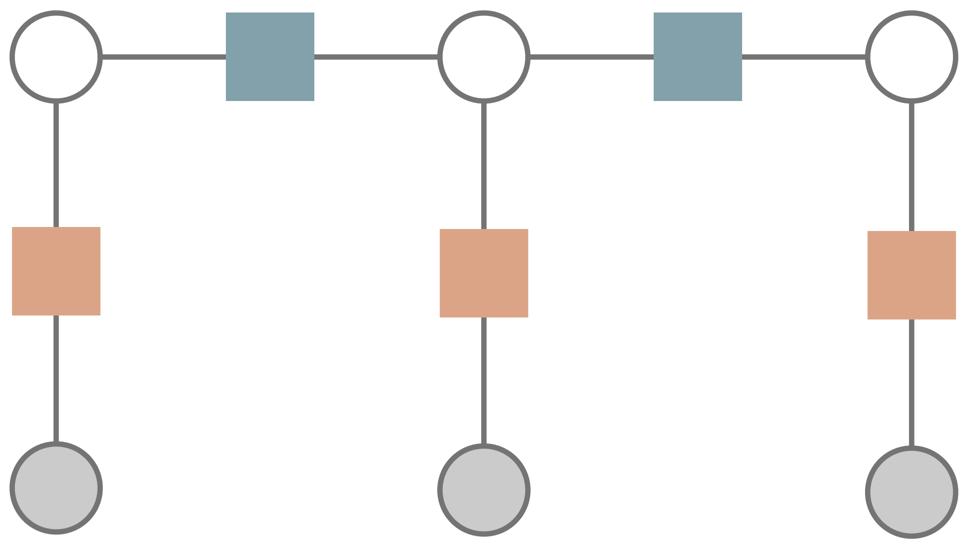
Flexible kernels



Generalisations to More Complex Settings

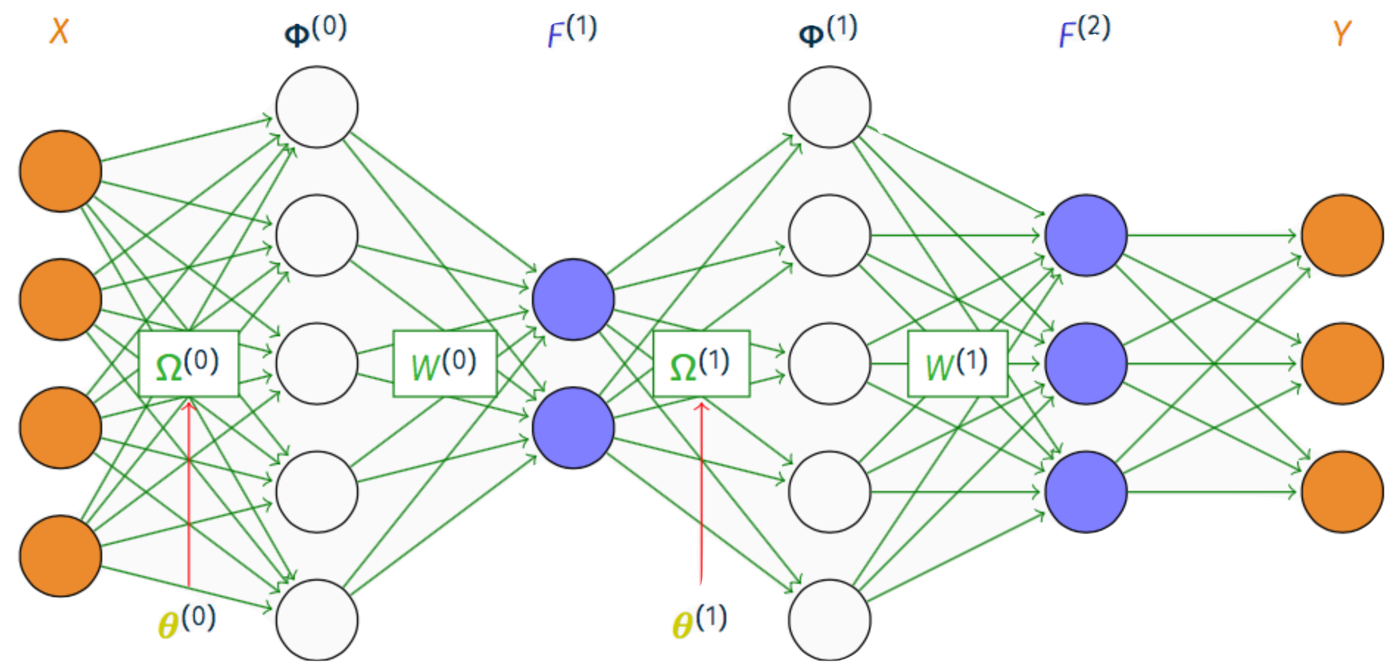
Structured prediction

(Galliani, Dezfouli, Bonilla and Quadrianto, AISTATS, 2017)

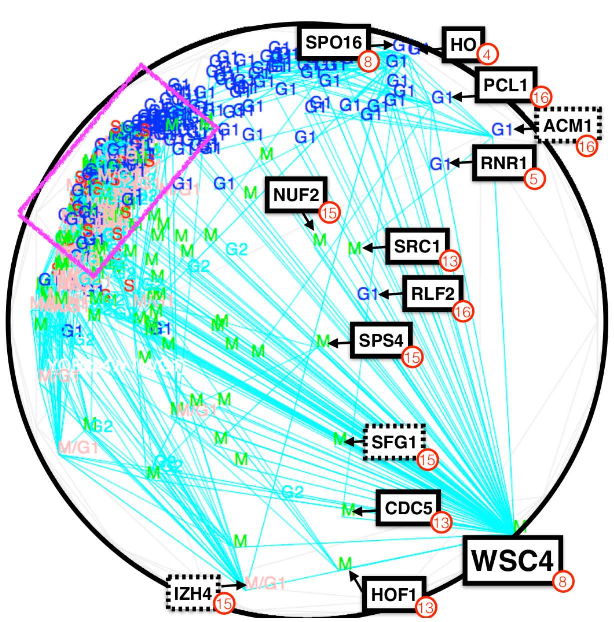


Deep Gaussian processes

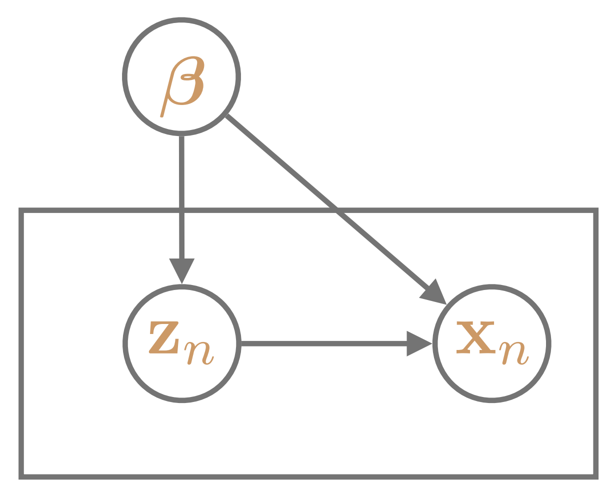
(Cutajar, Bonilla, Michardi and Filippone, ICML, 2017)



Network Structure Discovery



Implicit models



$$\mathbf{z}_n = g_z(\boldsymbol{\beta})$$

$$\mathbf{x}_n = g_x(\mathbf{z}_n; \boldsymbol{\beta})$$

$$q(\mathbf{Z}, \boldsymbol{\beta} | \mathbf{X}) = g(\mathbf{Z}, \boldsymbol{\beta})$$

Summary & Conclusions

- General framework for GP priors and non-linear likelihoods
- Scalable automated variational inference
- AutoGP
- Generalisations