Variational Bayesian Approximation of Inverse Problems using Sparse Precision Matrices

Jan Povala* **Ieva Kazlauskaite*** Eky Febrianto Fehmi Cirak
Mark Girolami

November 30, 2022

Published at Journal of Computer Methods in Applied Mechanics and Engineering (CMAME) 2022

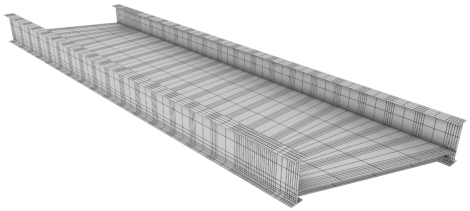
Motivating example: MX3D bridge

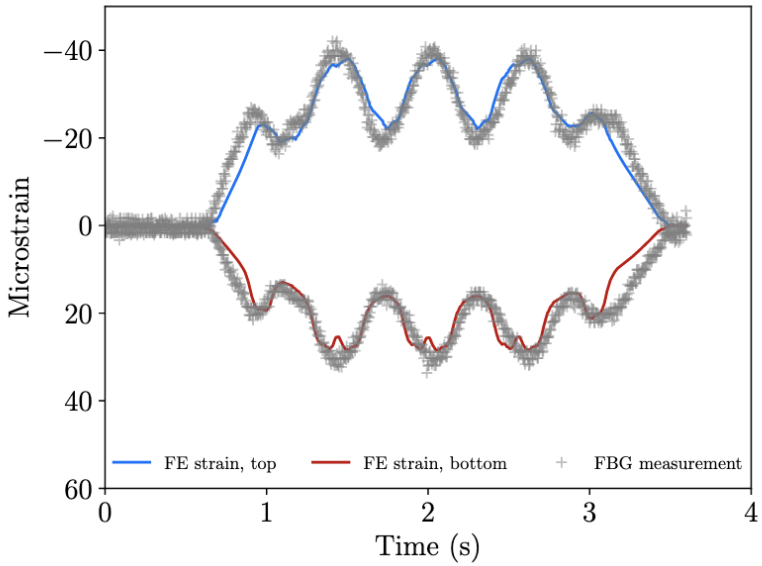
The Turing Institute video

Motivating example II

Febrianto, Butler, Girolami & Cirak (2021)

A railway bridge in Staffordshire which has been instrumented with fibre optic sensors (top left), its digital twin (top right), sample sensor measurements (bottom):





Using the Model and the Data

1. Using the measurements of the state of the system to infer the properties of the system.
   - Inverse problem

2. Using the model to answer 'what if' scenarios: How would the system behave under different conditions?
   - Forward problem

Rest of the Talk

1. Bayesian Inverse problem formulation.
2. Variational Bayes as an alternative to Markov Chain Monte Carlo methods.
3. Leveraging problem structure to specify the approximating family of distributions.
4. Results on elliptic PDEs.
5. Bimodal example.
6. Conclusions.

General Inverse Problem (Stuart 2010)

- Objective: Find $\kappa \in \mathcal{K}$, the parameters to a model, given $y \in \mathcal{Y}$, a noisy observation of the solution of a PDE.

General Inverse Problem (Stuart 2010)

- Objective: Find $\kappa \in \mathcal{K}$, the parameters to a model, given $y \in \mathcal{Y}$, a noisy observation of the solution of a PDE.
- For a suitable space $\mathcal{U}$, let $\mathcal{A}\colon \mathcal{K} \to \mathcal{U}$ be a possibly non-linear solution operator of a PDE. For a particular $\kappa \in \mathcal{K}$, the solution $u \in \mathcal{U}$ is

$$u = \mathcal{A}(\kappa). \tag{1}$$

General Inverse Problem (Stuart 2010)

- Objective: Find $\kappa \in \mathcal{K}$, the parameters to a model, given $y \in \mathcal{Y}$, a noisy observation of the solution of a PDE.
- For a suitable space $\mathcal{U}$, let $\mathcal{A}\colon \mathcal{K} \to \mathcal{U}$ be a possibly non-linear solution operator of a PDE. For a particular $\kappa \in \mathcal{K}$, the solution $u \in \mathcal{U}$ is

$$u = \mathcal{A}(\kappa). \tag{1}$$

- To obtain observation y, we define an additive observational noise $\eta \in \mathcal{Y}$. This gives

$$y = \mathcal{A}(\kappa) + \eta. \tag{2}$$

General Inverse Problem (Stuart 2010)

- Objective: Find $\kappa \in \mathcal{K}$, the parameters to a model, given $y \in \mathcal{Y}$, a noisy observation of the solution of a PDE.
- For a suitable space $\mathcal{U}$, let $\mathcal{A}\colon \mathcal{K} \to \mathcal{U}$ be a possibly non-linear solution operator of a PDE. For a particular $\kappa \in \mathcal{K}$, the solution $u \in \mathcal{U}$ is

$$u = \mathcal{A}(\kappa). \tag{1}$$

- To obtain observation y, we define an additive observational noise $\eta \in \mathcal{Y}$. This gives

$$y = \mathcal{A}(\kappa) + \eta. \tag{2}$$

- This is an ill-posed problem: there may be **no solution**, it may **not** be **unique**, and it may depend **sensitively** on y.

General Inverse Problem (Stuart 2010)

- Objective: Find $\kappa \in \mathcal{K}$, the parameters to a model, given $y \in \mathcal{Y}$, a noisy observation of the solution of a PDE.
- For a suitable space $\mathcal{U}$, let $\mathcal{A}\colon \mathcal{K} \to \mathcal{U}$ be a possibly non-linear solution operator of a PDE. For a particular $\kappa \in \mathcal{K}$, the solution $u \in \mathcal{U}$ is

$$u = \mathcal{A}(\kappa). \tag{1}$$

- To obtain observation y, we define an additive observational noise $\eta \in \mathcal{Y}$. This gives

$$y = \mathcal{A}(\kappa) + \eta. \tag{2}$$

- This is an ill-posed problem: there may be **no solution**, it may **not** be **unique**, and it may depend **sensitively** on y.
- Assumptions about κ are implemented via **regularisation**. Bayesian approach is one of the popular approaches:
  1. Describe prior knowledge of κ in terms of a prior probability on $\mathcal{K}$.
  2. Use Bayes' rule to calculate the posterior probability for κ given y.

Example: Poisson Problem

- We consider a classical elliptic Poisson problem:

$$\begin{aligned} -\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) &= f(\boldsymbol{x}), & \boldsymbol{x} &\in \Omega \subset \mathbb{R}^d, \\ u(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \partial\Omega, \end{aligned} \tag{3}$$

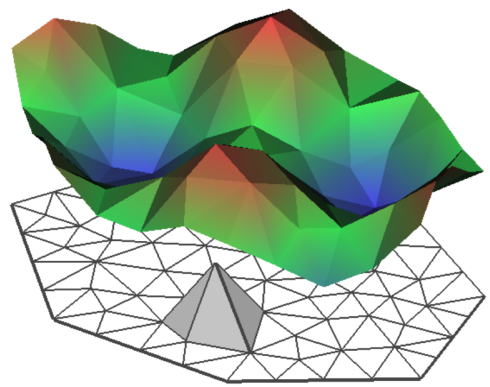
  where $\kappa(\boldsymbol{x}) \in \mathbb{R}$ is the log-diffusion coefficient, $u(\boldsymbol{x}) \in \mathbb{R}$ is the unknown, and $f(\boldsymbol{x}) \in \mathbb{R}$ is a deterministic forcing term.

Example: Poisson Problem

- We consider a classical elliptic Poisson problem:

$$\begin{aligned} -\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) &= f(\boldsymbol{x}), & \boldsymbol{x} &\in \Omega \subset \mathbb{R}^d, \\ u(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \partial\Omega, \end{aligned} \tag{3}$$

  where $\kappa(\boldsymbol{x}) \in \mathbb{R}$ is the log-diffusion coefficient, $u(\boldsymbol{x}) \in \mathbb{R}$ is the unknown, and $f(\boldsymbol{x}) \in \mathbb{R}$ is a deterministic forcing term.
- Disretise the weak form of the problem with a standard **finite element approach**.

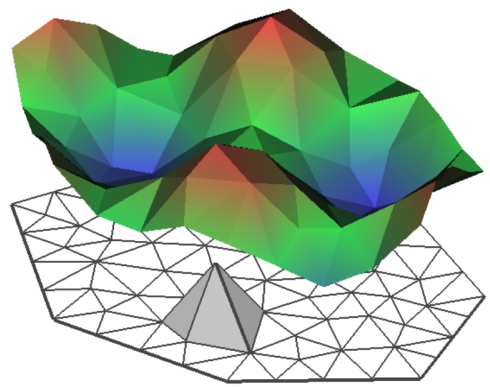


Example: Poisson Problem

- We consider a classical elliptic Poisson problem:

$$\begin{aligned} -\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) &= f(\boldsymbol{x}), & \boldsymbol{x} &\in \Omega \subset \mathbb{R}^d, \\ u(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \partial\Omega, \end{aligned} \tag{3}$$

  where $\kappa(\boldsymbol{x}) \in \mathbb{R}$ is the log-diffusion coefficient, $u(\boldsymbol{x}) \in \mathbb{R}$ is the unknown, and $f(\boldsymbol{x}) \in \mathbb{R}$ is a deterministic forcing term.
- Disretise the weak form of the problem with a standard **finite element approach**.



- Combining these terms, we obtain a linear system:

$$\mathbf{A}(\boldsymbol{\kappa})\mathbf{u} = \mathbf{f}\,, \tag{4}$$

  where $\mathbf{A}(\boldsymbol{\kappa}) \in \mathbb{R}^{n_u \times n_u}$ is the stiffness matrix, $\boldsymbol{\kappa} \in \mathbb{R}^{n_{\boldsymbol{\kappa}}}$ is the log-diffusion vector, $\mathbf{f} \in \mathbb{R}^{n_u}$ is the nodal source vector.

Example: Prior and Likelihood

- Placing a zero-mean Gaussian process prior on κ gives

$$\log p(\kappa) \sim \mathcal{GP}\big(0, k_\psi(\cdot,\cdot)\big). \tag{5}$$

- Finally, the likelihood is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = p(\boldsymbol{y} \mid u(\boldsymbol{\kappa})) = \mathcal{N}(\mathbf{A}(\boldsymbol{\kappa})^{-1}\mathbf{f}, \sigma_y^2\mathbf{I})\,. \tag{6}$$

- We consider an elliptic PDE of the form:

$$-\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}),$$

- Using FEM, we obtain a linear system:

$$\mathbf{A}(\boldsymbol{\kappa})u = \mathbf{f}\,,$$

- The likelihood is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = p(\boldsymbol{y} \mid u(\boldsymbol{\kappa})) = \mathcal{N}(\mathbf{A}(\boldsymbol{\kappa})^{-1}\mathbf{f}, \sigma_y^2\mathbf{I})\,.$$

- The prior is:

$$\log p(\boldsymbol{\kappa}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_\psi(x, x)).$$

- We consider an elliptic PDE of the form:

$$-\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}),$$

- Using FEM, we obtain a linear system:

$$\mathbf{A}(\boldsymbol{\kappa})u = \mathbf{f}\,,$$

- The likelihood is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = p(\boldsymbol{y} \mid u(\boldsymbol{\kappa})) = \mathcal{N}(\mathbf{A}(\boldsymbol{\kappa})^{-1}\mathbf{f}, \sigma_y^2\mathbf{I})\,.$$

- The prior is:

$$\log p(\boldsymbol{\kappa}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_\psi(x, x)).$$

We now develop the variational inference scheme for this problem.

Stochastic Variational Bayesian Inference

We posit a family of distributions $\mathcal{D}_q$ from which we choose the minimiser of the Kullback-Leibler divergence:

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa}) \in \mathcal{D}_q}{\arg\min}\ \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa} \mid \boldsymbol{y})). \tag{7}$$

Stochastic Variational Bayesian Inference

We posit a family of distributions $\mathcal{D}_q$ from which we choose the minimiser of the Kullback-Leibler divergence:

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa}) \in \mathcal{D}_q}{\arg\min}\ \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa} \mid \boldsymbol{y})). \tag{7}$$

This is equivalent to maximising the evidence lower bound (ELBO):

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa}) \in \mathcal{D}_q}{\arg\max}\ \mathbb{E}_q\left[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})\right] - \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa})). \tag{8}$$

Stochastic Variational Bayesian Inference

We posit a family of distributions $\mathcal{D}_q$ from which we choose the minimiser of the Kullback-Leibler divergence:

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa}) \in \mathcal{D}_q}{\arg\min}\ \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa} \mid \boldsymbol{y})). \tag{7}$$

This is equivalent to maximising the evidence lower bound (ELBO):

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa}) \in \mathcal{D}_q}{\arg\max}\ \mathbb{E}_q\left[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})\right] - \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa})). \tag{8}$$

Generally, $\mathbb{E}_q\left[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})\right]$ is not available in closed form. Instead, we use a Monte Carlo approximation with N_{SVI} samples from $q(\boldsymbol{\kappa})$ as follows

$$\mathbb{E}_q\left[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})\right] \approx N_{\mathsf{SVI}}^{-1} \sum_{i=1}^{N_{\mathsf{SVI}}} \log p(\boldsymbol{y} \mid \boldsymbol{\kappa}^{(i)}), \tag{9}$$

where $\boldsymbol{\kappa}^{(i)}$ is the i-th sample from $q(\boldsymbol{\kappa})$.

Multivariate Gaussian Parametrisations I

Different parametrisations of $q(\boldsymbol{\kappa})$ for the family of multivariate Gaussian distributions:

1. Diagonal covariance matrix, known as mean-field VB [MFVB]

$$q(\boldsymbol{\kappa}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{D})$$

2. Fully-specified Cholesky factor (full-covariance VB [FCVB])

$$q(\boldsymbol{\kappa}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^{\top})$$

3. Cholesky factor of the sparse precision matrix [PMVB]

$$q(\boldsymbol{\kappa}) \sim \mathcal{N}(\boldsymbol{\mu}, (\boldsymbol{L}_Q\boldsymbol{L}_Q^{\top})^{-1})$$

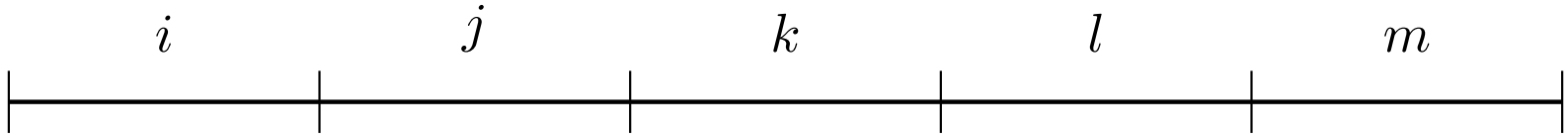
Precision Matrices and Conditional Independence

- Specify the multivariate Gaussian through the precision matrix: $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$.
- The elements of the precision matrix reflect **conditional independence**:

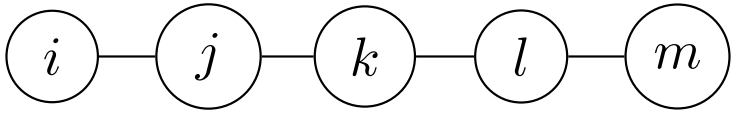
$$p(\kappa_i,\, \kappa_j \mid \boldsymbol{\kappa}_{-\{i,j\}}) = p(\kappa_i \mid \boldsymbol{\kappa}_{-\{i,j\}}) p(\kappa_j \mid \boldsymbol{\kappa}_{-\{i,j\}}) \quad \Leftrightarrow \quad Q_{ij} = 0\,. \tag{10}$$

- For more details, see Rue & Held (2005).

Leveraging Sparsity 1D Example

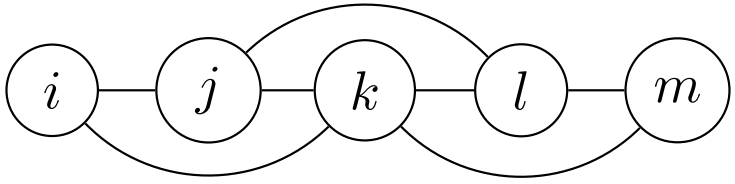


(a) Labelling of the five elements.



$$\begin{array}{c} \\ i \\ j \\ k \\ l \\ m \end{array} \begin{array}{c} \begin{array}{ccccc} i & j & k & l & m \end{array} \\ \begin{pmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix} \end{array}$$

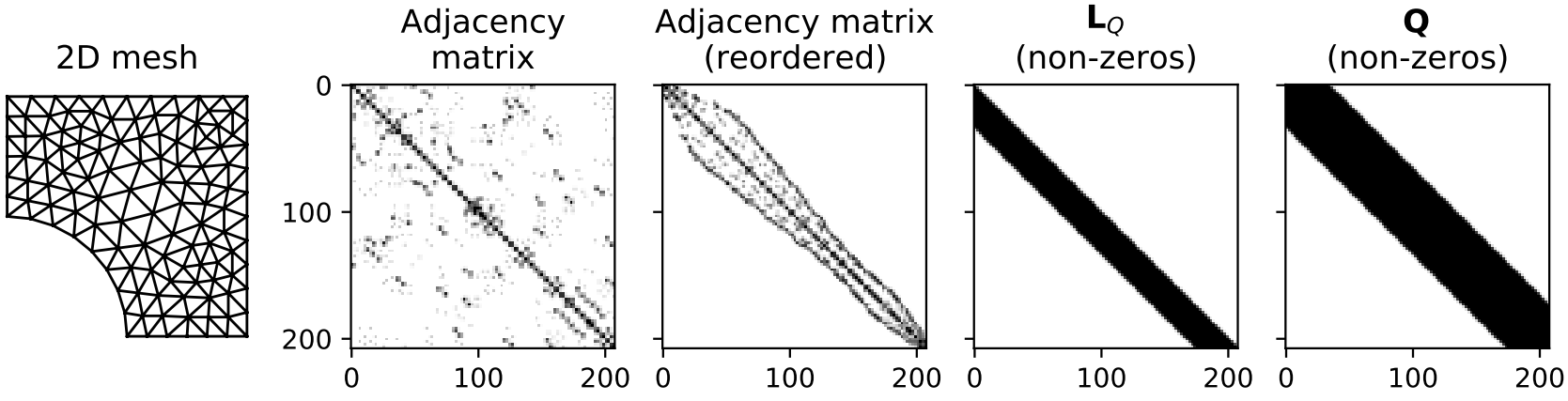
(b) 1-neighbourhood structure



$$\begin{array}{c} \\ i \\ j \\ k \\ l \\ m \end{array} \begin{array}{c} \begin{array}{ccccc} i & j & k & l & m \end{array} \\ \begin{pmatrix} \times & \times & \times & & \\ \times & \times & \times & \times & \\ \times & \times & \times & \times & \times \\ & \times & \times & \times & \times \\ & & \times & \times & \times \end{pmatrix} \end{array}$$

(c) 2-neighbourhood structure

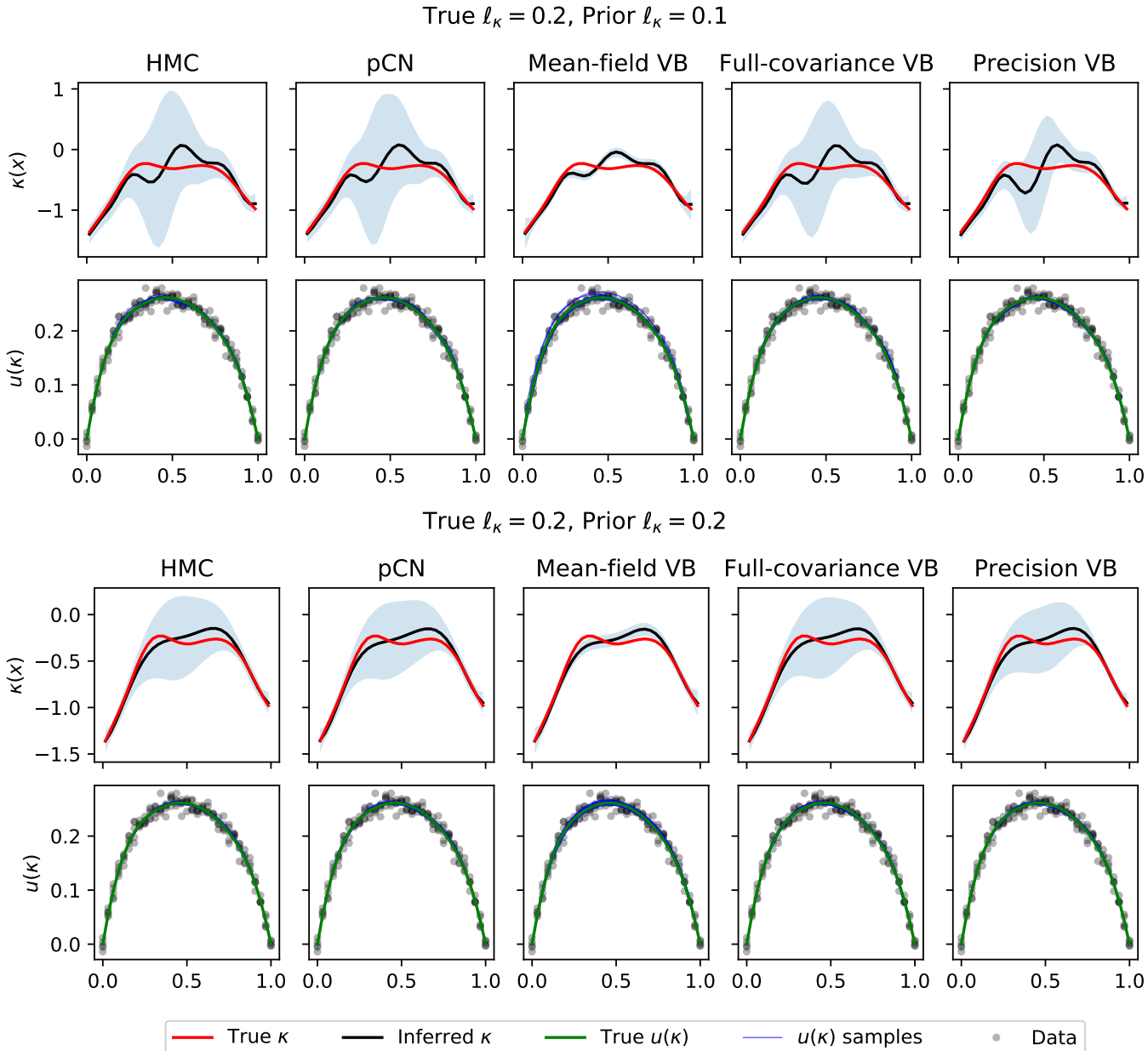
Leveraging Sparsity 2D Example



Experiments – 1D Poisson Problem

- Objective: infer the posterior distribution of $\boldsymbol{\kappa}$ given measurements of $\boldsymbol{y}$.
- Sensors are placed on the discretisation nodes and the sensor noise is $\sigma_y = 0.01$. Five readings are taken.
- Constant forcing, $f(x) = 1$ is assumed and we impose Dirichlet boundary conditions as $u(0) = u(1) = 0$.
- Data generated by assuming κ is a sample from a unit-variance zero-mean GP with lengthscale $\ell_\kappa = 0.2$.

Results – Uncertainty Estimates



Results – Uncertainty Propagation

We assess the propagation of uncertainty using log of total flux through the left boundary:

$$r(\kappa) = \log \int_{\Gamma_b} e^{\kappa(s)} \nabla u(s) \cdot \boldsymbol{n} \, \mathrm{d}s, \tag{11}$$

where $\boldsymbol{n}$ is a unit vector normal to the boundary.

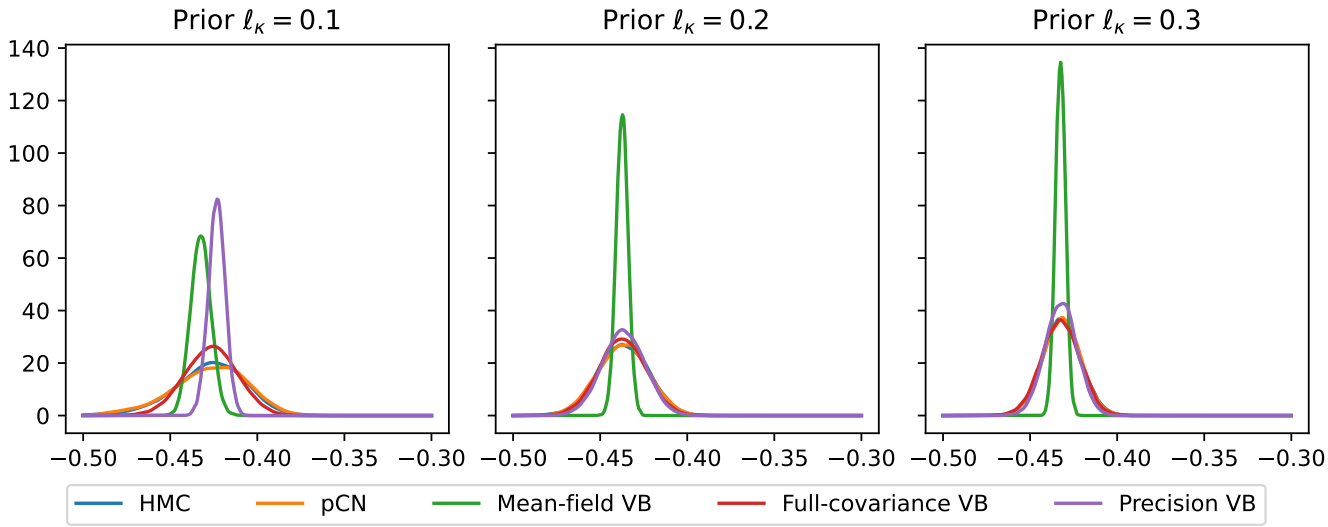
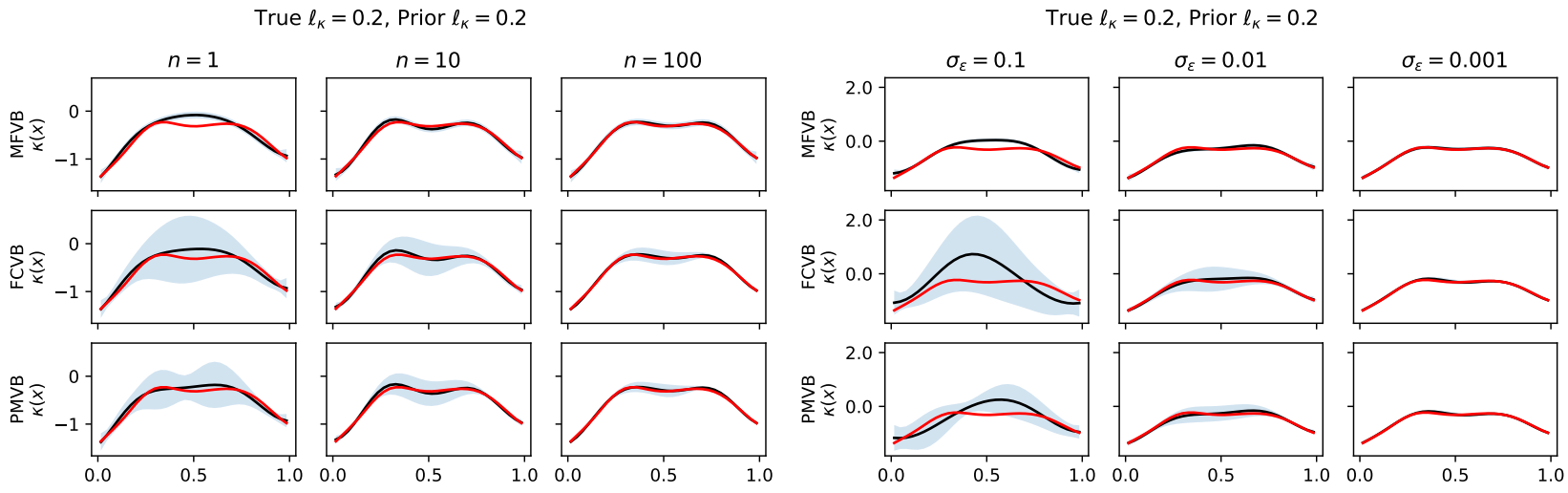


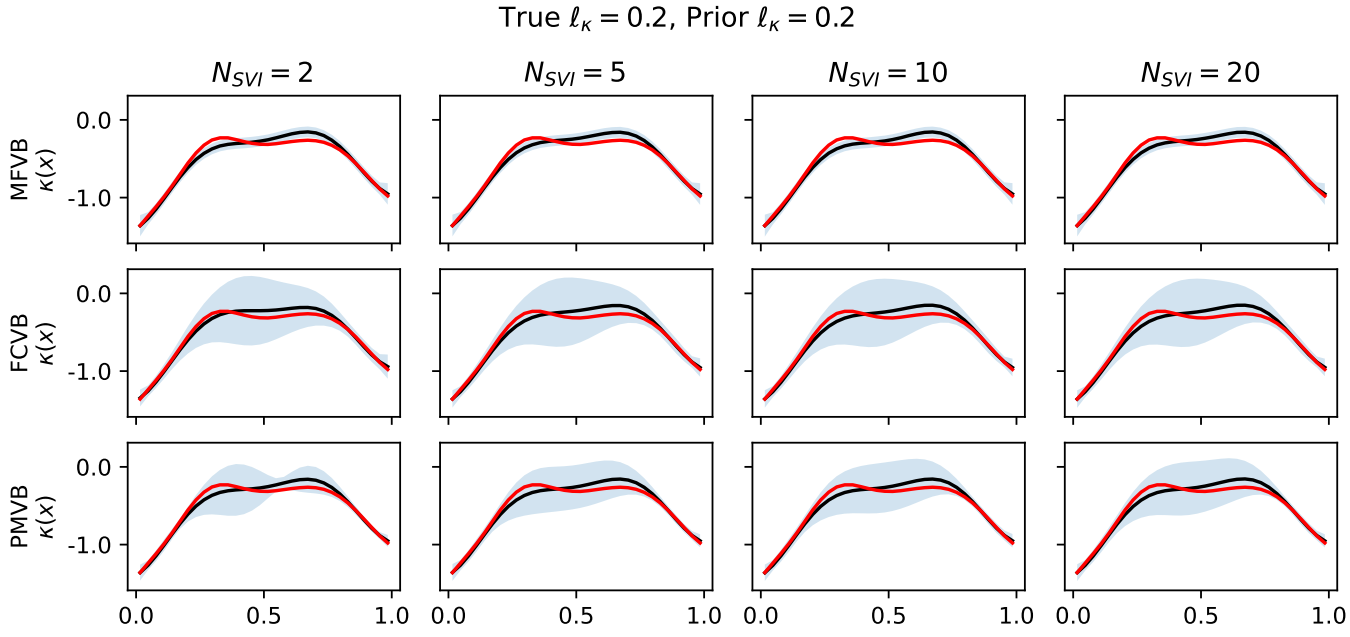
Figure: Log of the boundary flux at the left boundary node ($x = 0$) for the 1D Poisson example. For PMVB, the precision matrix bandwidth of 10 is used.

Results – Data Quality



- Mean estimates for all methods get closer to the true κ with improved information.
- FCVB and PMVB uncertainty estimates get narrower with increasing number of observations and with decreasing observational noise.

Results – Number of Samples for VB



- Figure shows the posterior estimates for different number of Monte Carlo samples in the estimation of ELBO,
- On a qualitative level, a small number of samples is sufficient to obtain a good estimate.

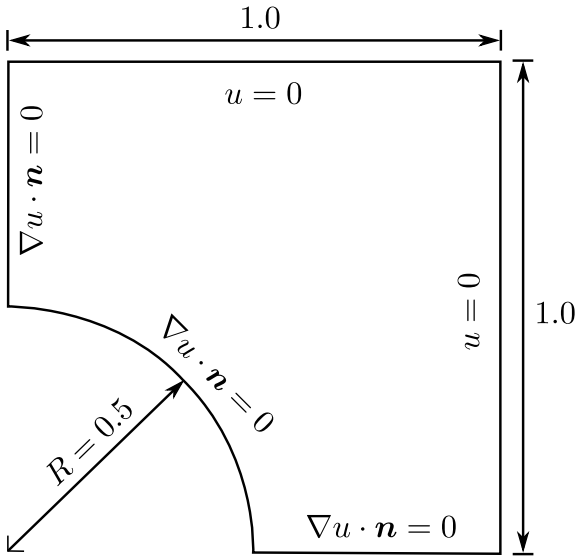
Results – Computation Time

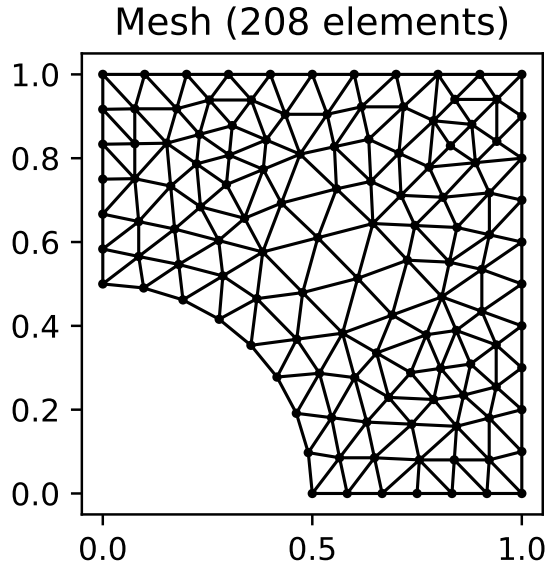
true ℓ_κ	prior ℓ_κ	Time (hours)			
		HMC	MFVB	FCVB	PMVB
0.1	0.1	15.2	1.1	3.6	2.1
	0.2	11.1	0.7	2.7	2.1
	0.3	7.2	0.6	2.3	2.0
0.2	0.1	15.2	0.6	2.2	1.8
	0.2	10.4	0.6	2.3	2.0
	0.3	7.0	0.5	1.7	1.8

Table: Run-times in hours for the Poisson 1D problem. For VB methods, $N_{\mathsf{SVI}} = 3$.

Experiments – 2D Poisson Problem

- Same setup as before, but now with the sensor noise $\sigma_y = 0.001$, and different boundary conditions.
- Dirichlet boundary conditions $u(x, y) = 0$ when $x = 1$ or $y = 1$. Neumann boundary conditions on the rest of the boundary





2D Example – Results I

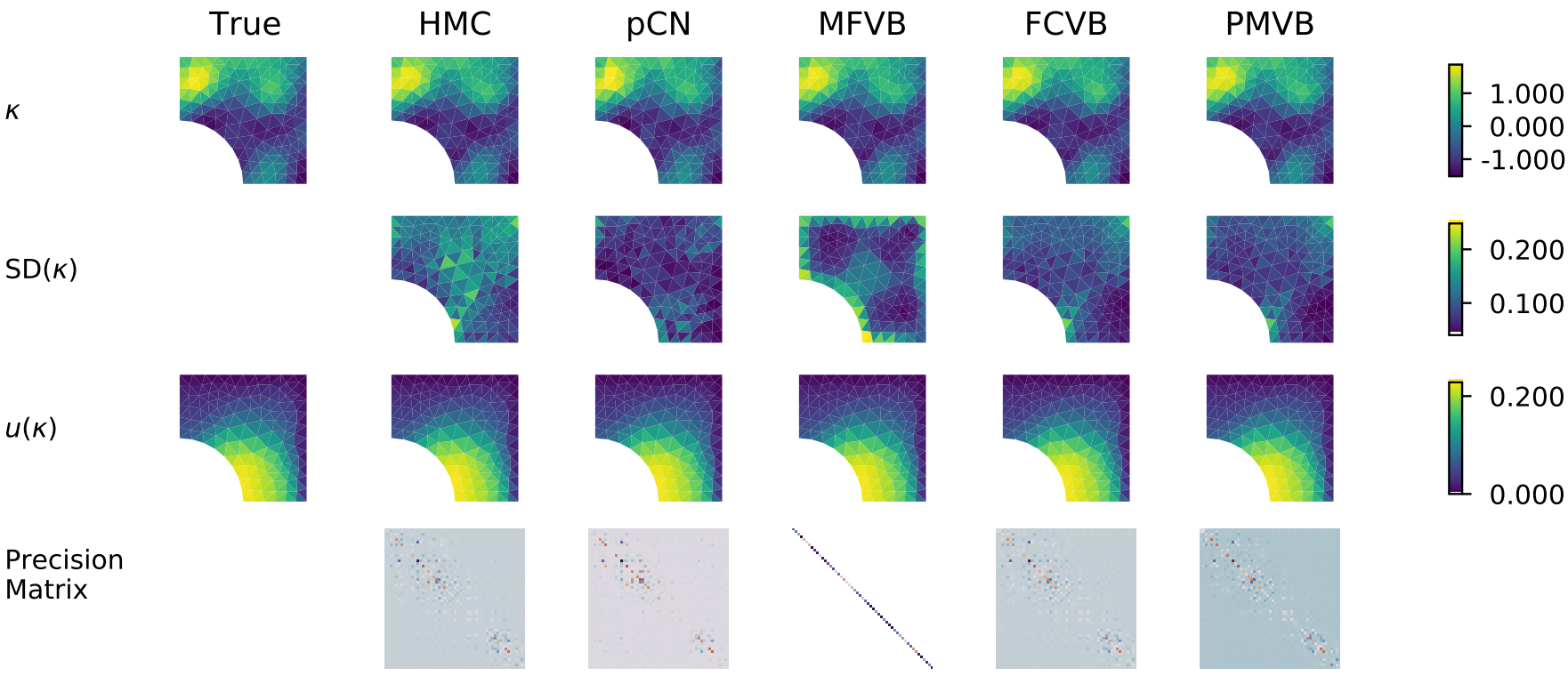


Figure: Posterior mean and standard deviation for $\boldsymbol{\kappa}$ and the corresponding u for 2D Poisson example with prior length-scale $\ell_\kappa = 0.1$. The bottom row shows the structure of the precision matrix for each inference scheme.

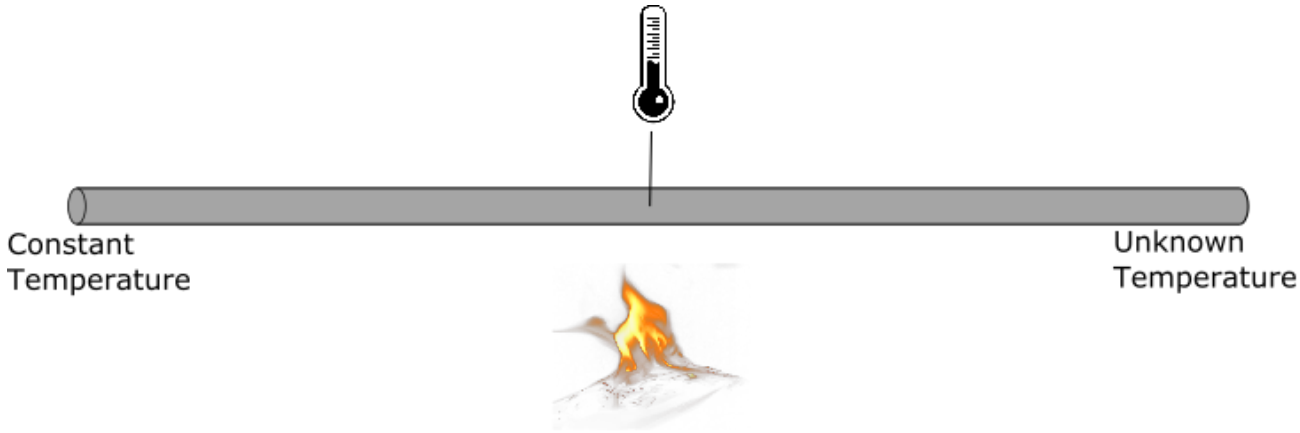
2D Example – Computation Cost

true ℓ_κ	prior ℓ_κ	Time (hours)			
		HMC	MFVB	FCVB	PMVB
0.1	0.1	240.6	6.4	29.6	28.1
	0.2	295.5	6.6	32.6	28.9
	0.3	242.0	7.3	27.3	30.6
0.2	0.1	242.7	6.2	34.3	27.2
	0.2	264.3	7.4	33.7	34.0
	0.3	221.9	7.8	31.3	34.0

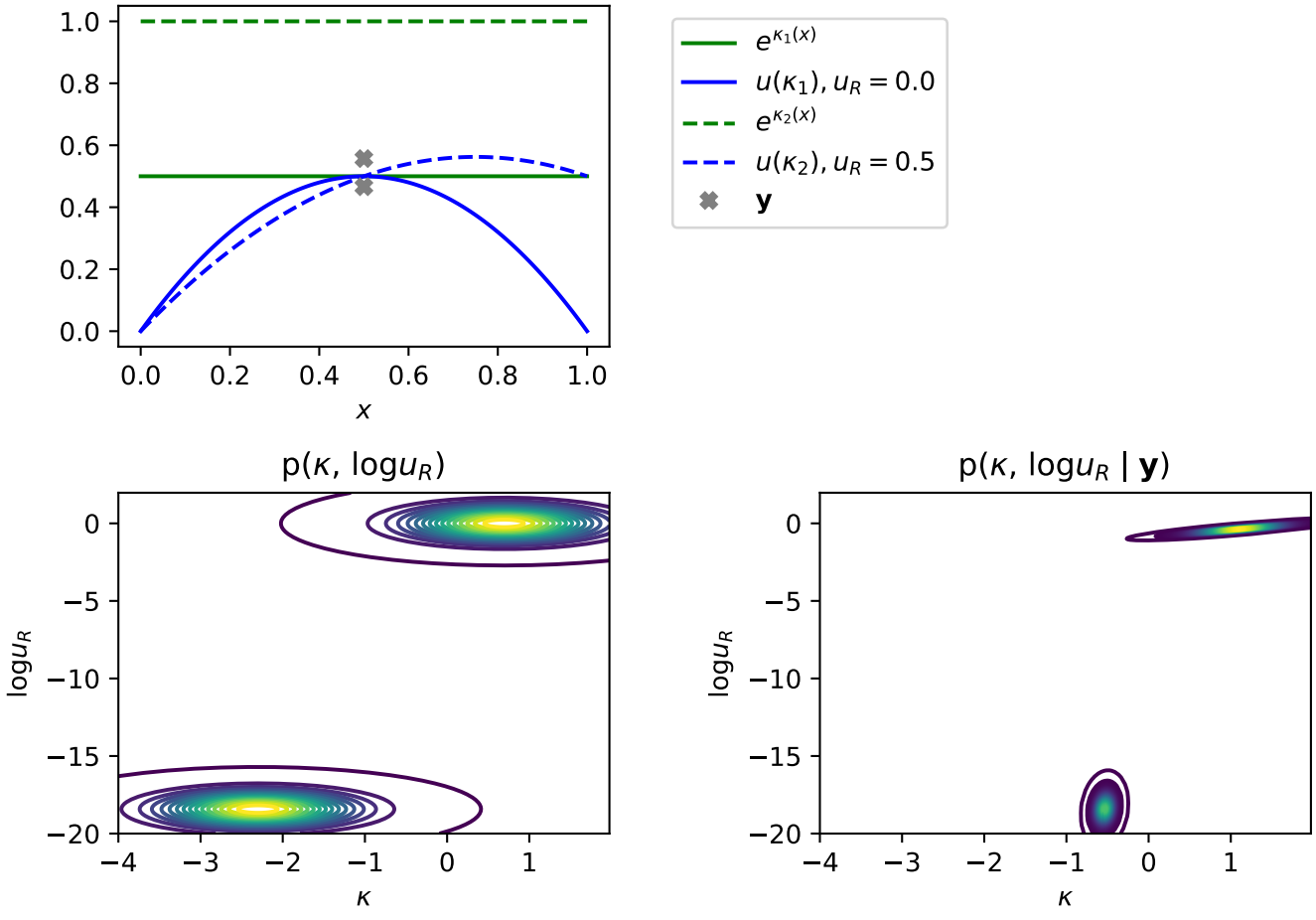
Table: Run-times for different inference schemes in seconds. The number of Monte Carlo samples is $N_{\mathsf{SVI}} = 5$ for all MFVB, FCVB, and PMVB.

Heat Example

- Metal rod with *unknown* conductivity properties (constant throughout).
- A uniform heat source throughout the rod.
- Temperature on the RHS is fixed and *unknown*, and on the LHS is fixed and known (BC).
- We obtain two readings of temperature in the centre of the rod.
- What can we say about heat conductivity, and the temperature on the RHS (unknown BC)?



VB Can Handle Multi-modal Posteriors



Implementation

- FEM C++ code written by CSMLab, led by Prof Fehmi Cirak.
- Our Tensorflow (Python) code performs the inference and interfaces with the FEM module.
- To maximise the ELBO in (8), we use the ADAM algorithm (Kingma & Ba 2015). ADAM is a member of a larger class of stochastic gradient decent optimisation methods for maximising non-convex cost functions.
- Available on Github.

Conclusions

- the mean of the variational posterior provides an accurate point estimate irrespective of the choice of the parametrisation,
- VB with a full-covariance or precision matrix structure adequately estimates posterior uncertainty compared to HMC and pCN,
- sparse precision matrix parameterisation leverages the structure of the problem to balance computational complexity and the ability to capture dependencies in the posterior distribution,
- VB provides a good estimate for mean and variance in a time that is at least an order of magnitude faster than HMC or pCN,
- the VB estimates may be used effectively in downstream tasks to estimate various quantities of interest.

Outlook:

- Leverage sparse linear algebra routines and different optimisation schemes.
- Consider stochastic forcing.

Bibliography I

Febrianto, E., Butler, L., Girolami, M. & Cirak, F. (2021), 'Digital twinning of self-sensing structures using the statistical finite element method'.

Kingma, D. P. & Ba, J. (2015), Adam: A method for stochastic optimization, *in* Y. Bengio & Y. LeCun, eds, 'International Conference on Learning Representations', San Diego, CA, USA.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, number 104 *in* 'Monographs on Statistics and Applied Probability', Chapman & Hall/CRC, Boca Raton.

Stuart, A. M. (2010), 'Inverse problems: A Bayesian perspective', *Acta Numerica* **19**, 451–559.