Uncertainty Quantification and Experimental Design for Large-Scale Linear Inverse Problems under Gaussian Process Priors, with Applications to Geophysics.

Cedric Travelletti

August 25, 2021

Joint work with D. Ginsbourger (UniBe) and N. Linde (UniL)

Motivating Example: Stromboli Volcano

Want to learn the interior structure of the Stromboli volcano.

- Only allowed to measure gravitational field on the surface

Question

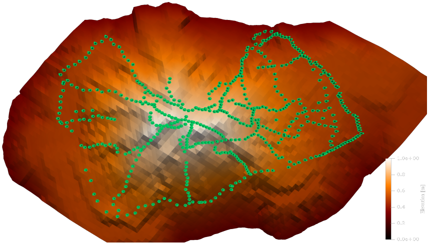
Where should I collect data to get the best possible reconstruction?

Section 1

Problem Setup

Problem Setup

- **Want to recover:** unknown density field $\rho : D \to \mathbb{R}$
- measurement sites $s_1, \ldots, s_n \in S$ on the surface
- **Available data:** surface gravity field $\{G_{s_i}[\rho]\}_{i=1,\ldots,n}$.



Measurement (forward) operator

$$G_{s_i}[\rho] = \int_D \rho(x) \frac{x^{(3)} - s_i^{(3)}}{\|x - s_i\|^3} dx$$

Solve problem in a Bayesian way by putting a GP prior on subsurface density field.

- Available observations (gravity) are linear forms of the field.
- Use conditional distribution to approximate unknown field ρ.

(Discrete) Bayesian Inversion

Discretize inversion domain into cubic cells of fixed side length.

- discretize GP prior on m cells: $\mathfrak{D} = \{x_1, ..., x_m\}$.
- Prior mean $\mu_0 = (\mu(x_i))_{i=1,\dots,m}$, covariance matrix $K_{ij} = k(x_i, x_j)$.

Posterior is gaussian with mean vector and covariance matrix

$$\tilde{\mu} = \mu_0 + KG^T\left(GKG^T + \tau^2 I\right)^{-1}(\mathbf{y} - G\mu_0)$$

$$\tilde{K} = K - KG^T\left(GKG^T + \tau^2 I\right)^{-1} GK$$

Get posterior by updating mean vector and covariance matrix.

Challenges and Limitations

Conditioning equations involve matrices that are impossible to store for (moderately) fine-grained inversion.

Challenges and Limitations

Conditioning equations involve matrices that are impossible to store for (moderately) fine-grained inversion.

$$\tilde{\mu} = \mu_0 + KG^T\left(GKG^T + \tau^2 I\right)^{-1}(\mathbf{y} - G\mu_0)$$
$$\tilde{K} = K - KG^T\left(GKG^T + \tau^2 I\right)^{-1}GK$$

- Covariance matrix quadratic in number of inversion cells ($\mathcal{O}(m^2)$ storage).
- Forward operator is *dense*: each datapoint influenced by **all** cells in discretization $\implies$ No Sparsity.
- SPECIFIC TO INVERSE PROBLEMS.

Impossible to store covariance matrices for *real-world sized* problems.

Challenges and Limitations: Stromboli Memory Needs

$$\tilde{\mu} = \mu_0 + KG^T \left(GKG^T + \tau^2 I\right)^{-1} (\mathbf{y} - G\mu_0)$$

$$\tilde{K} = K - KG^T \left(GKG^T + \tau^2 I\right)^{-1} GK$$

Matrix	# Elements	Storage
K	$2.9 * 10^{10}$	115 GB
KG, G	$9.2 * 10^{7}$	369 MB
$(\cdots)^{-1}$	$2.9 * 10^{5}$	1.2 MB
μ	$1.7 * 10^{5}$	0.7 MB

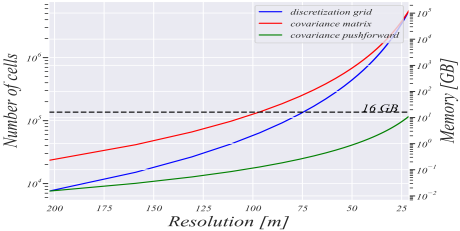


Figure: Grid and matrices size vs resolution on Stromboli example.

Compared to GP-regression, different dimensions at play.
(specific to inversion).

Summary of Challenges for large GP-based Inversion

- Larger-than-memory covariance matrices
- Difficult to exploit sparsity when observation operators are dense (inversion-specific).
- Sequential updates impossible (cannot store intermediate covariances).

Section 2

Solving the Memory Bottleneck

Implicit Representation and Update of Covariance

Use an (almost-) matrix-free implicit representation of the posterior covariance matrix.

Implicit Representation and Update of Covariance

Use an (almost-) matrix-free implicit representation of the posterior covariance matrix.

Basic Idea

Most/all posterior covariance information can be extracted by computing products of the posterior covariance matrix with other matrices.

- Store *cooking recipe* for multiplication with posterior covariance.
- *Recipe* is much lighter than full matrix.
- Can be updated quickly if new data becomes available.

Implicit Representation and Update of Covariance

Use an (almost-) matrix-free implicit representation of the posterior covariance matrix.

Basic Idea

Most/all posterior covariance information can be extracted by computing products of the posterior covariance matrix with other matrices.

- Store *cooking recipe* for multiplication with posterior covariance.
- *Recipe* is much lighter than full matrix.
- Can be updated quickly if new data becomes available.

Implicit representation allows:

- Fast inclusion of new data (update of posterior).
- Transfer of computations to multiple computational units and/or GPUs (only small multiplications involved).

Implicit Representation of Covariance Matrix

Observation

Posterior covariance information may be extracted via products with *tall and thin* matrices:

$$\tilde{K}A,\ A \in \mathbb{R}^{m \times p},\ p \ll m$$

Only need to maintain a multiplication routine.

Implicit Representation: Sequential Setting

Consider sequential data assimilation setting.

- Measurements $G_1, ..., G_n$.
- Covariance after inclusion of first n batches: $K^{(n)}$.
- Do not compute $K^{(n)}$, only maintain a right-multiplication routine.

$$\mathit{CovMul}_n : A \mapsto K^{(n)} A$$

Implicit Representation: Sequential Setting

Consider sequential data assimilation setting.

- Measurements $G_1, ..., G_n$.
- Covariance after inclusion of first n batches: $K^{(n)}$.
- Do not compute $K^{(n)}$, only maintain a right-multiplication routine.

$$\mathit{CovMul}_n : A \mapsto K^{(n)} A$$

- Update this *implicit* representation at every new data inclusion.

$$K^{(n)} A = K^{(0)} A - \sum_{i=1}^{n} \bar{K}_i R_i^{-1} \bar{K}_i^T A$$

$$\bar{K}_i := K^{(i-1)} G_i^T,$$

$$R_i^{-1} := \left(G_i K^{(i-1)} G_i^T + \tau^2 I \right)^{-1}.$$

Implicit Representation: Computational Cost

For reconstruction of full posterior covariance matrix, only need to store two matrices at each data ingestion stage:

- $m \times d_n$ matrix $K_{n-1}G_n^T$
- $d_n \times d_n$ matrix $\left(G_n K_{n-1} n G_n^T + \Delta\right)^{-1}$.

With d_n number of datapoints at stage n and Δ data noise covariance.

- Only involves right-multiplication with previous stage covariance.
- Can hence be defined recursively.
- Substantial memory saving compared to storing full covariance.

Example Memory Footprint: Stromboli

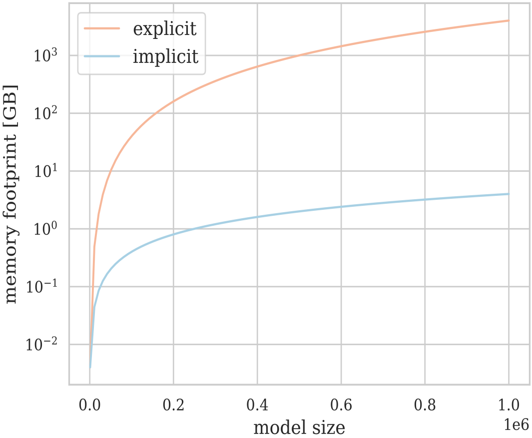


Figure: Memory footprint of posterior covariance as a function of model size.

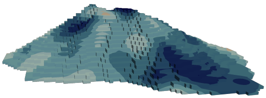
Section 3

Optimal Design

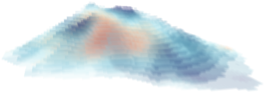
Example: Optimal Design on Stromboli Volcano

Use posterior covariance information to guide future data acquisition.

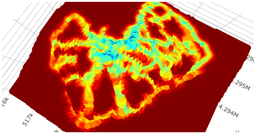
- Observation locations selected according to pre-defined objective.



(a) simulated surface gravity



(b) reconstructed internal mass density



(c) residual variance

- New observation locations selected sequentially.
- Locations selected based on how observation there would change covariance.
- Requires fast access/update of posterior covariance.

Possible Objective: Excursion Set Reconstruction

Question

What if want to learn regions of high density instead of full matter distribution?

- Target is excursion set $\Gamma = \{x \in D : Z_x > t\}$.
- Observations close to excursion regions more informative.
- $\implies$ Find criterion for how informative (about the excursion set) a given observation location s is.

(weighted) IVR criterion

$$\mathrm{IVR}(s) := \int_D \left(\mathrm{Var}_n \left[Z_x\right] - \mathrm{Var}_n \left[Z_x \mid G_{(s)}\right]\right) \mathbb{P}_n \left[x \in \Gamma\right] dx$$

Where n data ingestion steps have already been performed and variances and excursion probabilities are computed under the current (stage n) conditional law of the field.

IVR and Implicit Update Framework

$$\mathrm{IVR}(G) \cong \sum_{i=1}^{m} \left(K G^T \left(G K G^T + \tau^2 I \right)^{-1} G K \right)_{ii}$$

- Requires multiplication with current covariance matrix K.
- In a sequential and large scale setting, K cannot be stored.
- Computation of IVR without update formulae requires inverting concatenated dataset.
- Update formulae allow to compute only contribution of new observation G.
- No need to re-invert whole dataset.
- Cost scales cubically in the size of the new observations (comparted to cubic in dataset size for direct approach).

Test Case: Stromboli

- Want fine grained reconstruction (50m resolution) of excursion set.
- High resolution $\implies$ big grid ($\sim$ 150k cells) $\implies$ big cov matrix ($\sim$ 150GB).

Optimal observation plan by myopic optimization of wIVR criterion.

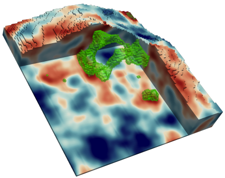


Figure: True density and excursion set (generated from model).

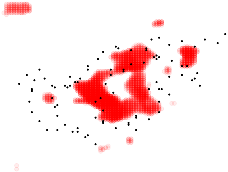


Figure: Proposed data collection plan, wIVR long-range strategy, total budget of 90 observations.

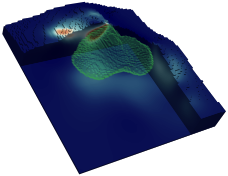


Figure: Estimated excursion set (Vorob'ev Expectation) and coverage function.

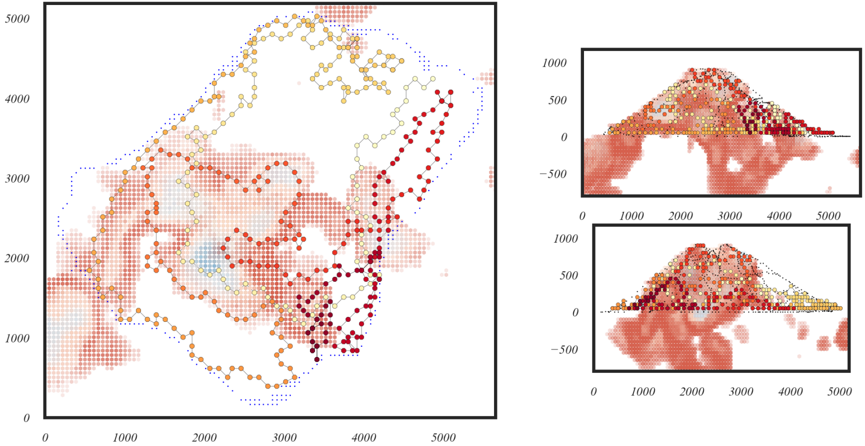


Figure: Projection of true excursion set. Island boundary denoted in blue, observation location from previous field campaign denoted by black dots.

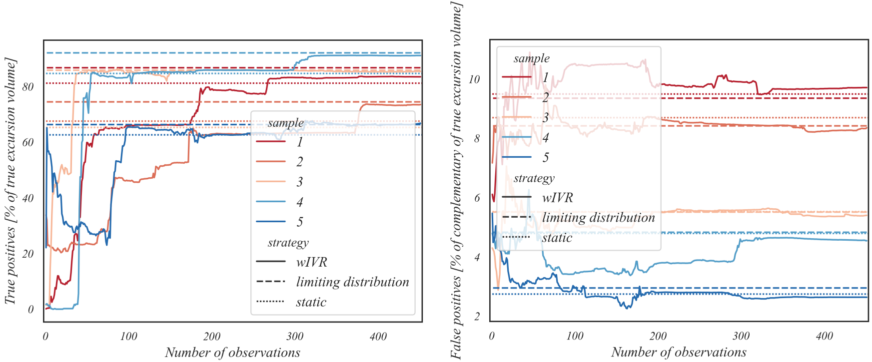
Diagnostics: Limiting Distribution

Since implicit representation allows fast inclusion of new datapoints, can study the *limiting distribution*.

Limiting Distribution

- Posterior distribution after data collected at all accessible locations.
- Gives sense of *minimal residual uncertainty* (inherent to this type of data).

Results



(a) True positives (b) False positives

Figure: Evolution of true and false positives for the *small* scenario as a function of the number of observations.

UQ on Excursion Volume

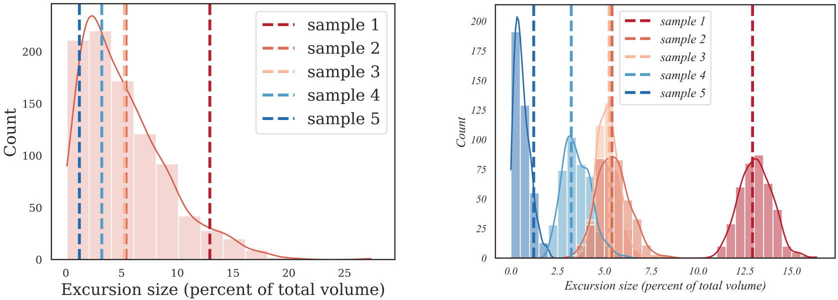


Figure: Prior (left) and empirical posterior (right) distribution (after 450 observations) of the excursion volume for each ground truth. True volumes are denoted by vertical lines.