Universal Inversion: Extending Universal Kriging to Include Trends in Bayesian Inverse Problems

Cedric Travelletti

June 6, 2022

Joint work with D. Ginsbourger (UniBe) and N. Linde (UniL)
Section 1

Introduction
Problem Setup

Want to learn an unknown function

\[ f : D \rightarrow \mathbb{R} \]
Problem Setup

Want to learn an unknown function

\[ f : D \rightarrow \mathbb{R} \]

given some data \( f(x_i) \).
Gaussian Process Regression

Can be done in a Bayesian way by assuming $f$ is a realization of a Gaussian process prior $Z \sim \text{Gp}(m_0, k)$.

![Figure: Posterior mean (blue).](image)

Approximate $f$ by posterior of $Z$ conditional on the data $Z_{x_i} = f(x_i)$. 
What if do not have point data $f(x_i)$ but more general data:

$$y_i = \ell_i(f)$$

for some linear functionals $\ell_i$. 
What if do not have point data $f(x_i)$ but more general data:

$$y_i = \ell_i(f)$$

for some linear functionals $\ell_i$.

**Examples**

- **Derivative observations** $\ell_i(f) = f'(x_i)$
- **Integral data**: $\ell(f) = \int_D f(x)dx$
- **Fourier coefficients** $\ell_k(f) = \int_D e^{-2\pi ikx} f(x)dx$
- **Kernel operators** $\ell_s(f) = \int_D f(x)g(x, s)dx$, for some function $g$. 
Linear operator data arise everywhere in science:

**Remote Sensing**
- Observe reflected light
- Recover land properties

**Tomography**
- Observe transmitted X-Ray intensity
- Recover material properties

**Geoscience**
- Observe gravitational field
- Recover rock density

Broadly known as (linear) **Inverse Problems**.
GPs can easily handle linear operator data.

\[ \ell_k(f) = \int_D e^{-2\pi i k x} f(x) dx, \quad k = 1, 5, 7, 10 \]

\[ \Rightarrow \text{can use GP priors in inverse problems (Bayesian Inversion)} \]
Problem for today:

Can we scale the GP + linear operator framework to "real-world" problems?
Example Real-World Problem: Gravimetric Inversion

Recover interior of Stromboli volcano from surface gravity.

- **density**
- **observed gravity**
Properties:

- **Linear operator data.**
- **"Large-scale"**: large 3 dimensional inversion grid.
- **Sequential** assimilation of new data important in practice.
Gravity field $G(s_i, \rho)$ at site $s_i$ generated by underground density $\rho$ can be written as linear operator:

$$y_i = G(s_i, \rho) = \int_D \rho(x) g(x, s_i) dx$$
Gravity field $G(s_i, \rho)$ at site $s_i$ generated by underground density $\rho$ can be written as linear operator:

$$y_i = G(s_i, \rho) = \int_D \rho(x) g(x, s_i) dx$$

- Traditionally solved by discretizing on a grid $x = (x_1, \ldots, x_m)$.
- Observation model

$$y = G(\rho(x_1), \ldots, \rho(x_m))^T$$

- Discretized observation operator $G \in \mathbb{R}^{n \times m}$ for $n$ observations.
- Posterior described by:
  - mean vector $\tilde{m} = (\tilde{m}_{x_1}, \ldots, \tilde{m}_{x_m})^T$
  - covariance matrix $\tilde{K} = \left( \tilde{k}(x_i, x_j) \right)_{i,j=1,\ldots,m}$
\[ \tilde{m} = m_0 + KG^T (GKG^T + \tau^2 I)^{-1} (y - Gm_0) \]
\[ \tilde{K} = K - KG^T (GKG^T + \tau^2 I)^{-1} GK \]

Linear operator data (can) involve all grid points at the same time.
Figure: Grid and matrices size vs resolution on Stromboli example.
Section 2

Implicit Covariance Representation for Large-Scale Inversion
Solving **practical** difficulties of Bayesian inversion

- Covariance matrix too big? \(\Rightarrow\) Don’t store it, nor build it.

**Implicit Representation**

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

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

\(\Rightarrow\) Only need to maintain a multiplication routine.

---

Rigorous introduction of such an implicit representation requires us to understand which linear operators are allowed for the conditional law to be well defined.

- Conditioning is a bit "taboo" in the GP community.
- Usually done by considering the finite-dimensional distributions $Z_{x_1}, \ldots, Z_{x_m}$
- What if we have linear operator data?

Disintegrations of Gaussian measures offer sound theoretical framework.
For Gaussian measures on a separable Banach space $X$, conditioning wrt. linear operator data well-defined for bounded observation operators

$$G : X \rightarrow Y$$

into a separable Banach space $Y$. 
For Gaussian measures, conditioning done by disintegration:

**Theorem**

Let $X, Y$ be real separable Banach spaces and $\mu$ be a Gaussian measure on the Borel $\sigma$-algebra $\mathcal{B}(X)$ with mean element $m_\mu \in X$ and covariance operator $C_\mu : X^* \to X$. Let also $G : X \to Y$ be a bounded linear operator.

Then there exists a continuous affine map $\tilde{m}_\mu : Y \to X$, a symmetric positive operator $\tilde{C}_\mu : X^* \to X$ and a disintegration $(\mu|_{G=y})_{y \in Y}$ of $\mu$ with respect to $G$ such that for each $y \in Y$ the measure $\mu|_{G=y}$ is Gaussian with mean element $\tilde{m}_\mu(y)$ and covariance operator $\tilde{C}_\mu$. The mean element also satisfies $G\tilde{m}_\mu(y) = y$ for all $y \in Y_0 := Gm_\mu + GC_\mu G^* (Y^*)$. 
So far have rigorous conditioning wrt. linear operators for Gaussian measures.

**Question**

Under what conditions are GPs and Gaussian measures related?
Gaussian Processes ⇐⇒ Gaussian Measures

Under which conditions does a GP with trajectories in a given Banach space $X$ induce a Gaussian measure on $X$?

- For $X = C(D)$ Banach space of continuous functions on a compact metric space $D$ ✓.
- For $X = \mathcal{H}$ reproducing kernel Hilbert space ✓.
Under which conditions does a GP with trajectories in a given Banach space $X$ induce a Gaussian measure on $X$?

- For $X = C(D)$ Banach space of continuous functions on a compact metric space $D$.
- For $X = \mathcal{H}$ reproducing kernel Hilbert space.

**Lemma**

Let $Z$ be a Gaussian process with covariance kernel $k$ and assume that $k$ is bounded on the diagonal: $k(x, x) < C^2$, all $x \in D$. Then there exists $0 < \theta \leq 1$ such that $Z$ induces a Gaussian measure on $\mathcal{H}^\theta_k$. 
Morality:

- Language of disintegrations of measures provides rigorous formulation of conditioning wrt. linear operators.
- Observation operator has to be a bounded operator into a separable Banach space.

Need conditions for GP to induce Gaussian measure.

- Banach space of continuous functions on compat metric space.
- Reproducing kernel Hilbert space.
- Under mild conditions, sample paths of a GP lie in an RKHS that is ”slightly larger” that the RKHS of its covariance kernel.
Some References


Consider sequential data assimilation setting.

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

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

- 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: Advantages

- Drastically reduced memory footprint.

**Figure**: Memory footprint of posterior covariance vs grid size.

- Fast inclusion of new data.
- Update done in small chunks $\implies$ can send to GPU.
Inversion results (Matern 3/2 kernel), hyperparameters trained with MLE on field data are in agreement with field knowledge.

**Figure:** Posterior mean $[kg/m^3]$.  

**Figure:** Posterior mean $[kg/m^3]$.  

Cedric Travelletti (UniBe)
Section 3

Universal Inversion
GP regression with trends known as universal kriging.
GP regression with trends known as **universal kriging**.

**Goal:** Extend to inverse problems to get "universal inversion".
Include expert knowledge in the inversion through trends.

Model known geological structures such as:
- Layers
- Chimneys
- Depth-dependence
- ...

**Figure:** Radial trend (chimney).
Prior with trend

Assume prior is sum of trend + fluctuations described by Gaussian process

\[ Z_x = F_x \beta + \eta_x , \]

- \( \eta \) is a (centred) GP with kernel \( k \)
- \( F_x \) is a vector of basis functions \((F_x)_i = f_i(x)\)

Put a Gaussian prior on the trend coefficients

\[ \beta \sim \mathcal{N}(\mu, \Sigma) . \]
Theorem

Conditionally on linear operator data $Y = GZ_w + \epsilon$, the posterior of the trend coefficients is Gaussian with mean and covariance given by:

$$
\mathbb{E} [\beta | y] = \mu + \Sigma F_w^T G^T Q^{-1}_y (y - GF_w \mu)
$$

$$
\text{Cov} (\beta, \beta | y) = \Sigma - \Sigma F_w^T G^T Q^{-1}_y GF_w \Sigma,
$$

assuming that the matrix $Q_y := G (F_w \Sigma F_w^T + K_{ww}) G^T + \sigma^2 \epsilon I$ is invertible.

Conditionally on the data, the distribution of $Z$ is also that of a GP, with mean and covariance function given by:

$$
m_{Z|y}(x) = F_x \mu + (F_x \Sigma F_w^T + K_{xw}) G^T Q^{-1}_y (y - GF_w \mu)
$$

$$
k_{Z|y}(x, x') = K_{xx'} + F_x \Sigma F_{x'}^T - (F_x \Sigma F_w^T + K_{xw}) G^T Q^{-1}_y G (F_w \Sigma F_{x'}^T + K_{wx'})
$$
Again consider dataset consisting of two batches of data $\mathbf{Y} = (\mathbf{Y}_i, \mathbf{Y}_{-i})$.

Want to compute residual when we predict $\mathbf{Y}_i$ using $\mathbf{Y}_{-i}$

$$\mathbf{Y}_i - \hat{\mathbf{Y}}_i^{(-i)}$$

- Want fast formula for CV residual (avoid full recomputation of predictor).
- Formula should be valid for any subset of indices $i$ (k-fold).
By generalizing [GS] all the information we need is contained in the augmented matrix:

\[
\tilde{K} = \begin{pmatrix}
GKG^T & GF \\
F^TG^T & 0
\end{pmatrix}.
\]

We partition it as:

\[
\tilde{K} = \begin{pmatrix}
\tilde{K}_{ii} & \tilde{K}_{i-i} \\
\tilde{K}_{-ii} & \tilde{K}_{-i-i}
\end{pmatrix} = \begin{pmatrix}
G_{i\cdot}K_{G_{i\cdot}}^T & G_{i\cdot}K_{G_{-i\cdot}}^T & G_{i\cdot}F \\
G_{-i\cdot}K_{G_{i\cdot}}^T & G_{-i\cdot}K_{G_{-i\cdot}}^T & G_{-i\cdot}F \\
F^T G_{i\cdot}^T & F^T G_{-i\cdot}^T & 0
\end{pmatrix}.
\]

CV residuals can then be computed by extracting subblocks of the inverse.
Upper left block of the inverse give us (inverse) covariance of residuals:

\[
\tilde{K}_{ii}^{-1} = \text{Cov} \left( \hat{Y}_i^{(-i)}, \hat{Y}_i^{(-i)} \right)^{-1},
\]

where \( \tilde{K}_{ii}^{-1} \) denotes the \( ii \) sub-block of \( \tilde{K}^{-1} \). Can get the residuals in the same way:

\[
\tilde{K}_{ii}^{-1} \left( \tilde{K}^{-1} [:, 1 : n] \boldsymbol{Y} \right)_i = \boldsymbol{Y}_i - \hat{Y}_i^{(-i)}. 
\]
Cross-validation can be used for model selection (future developments).

- Leave-one out residuals (Constant model)
  Mean leave-one out residual $-1.223$ [mGal].

- Cylindrical trend
  Mean leave-one out residual $-0.2281$ [mGal].
Future Work

- Include expert-defined trends.
- Model selection via cross-validation (penalize complexity).
- Beyond leave-one-out (k-fold).
Conclusion

- **Implementation**
  - GPs priors can handle large-scale inversion via implicit representation of the posterior covariance.
  - Implicit representation allows fast updates and can be run on GPU.

- **Theory**
  - Disintegrations of Gaussian measures provide rigorous treatment of conditioning under linear operator data.
  - Mild conditions on the GP guarantee process $\Longleftrightarrow$ measure equivalence.

- **Modelling**
  - Universal kriging can be extended to inverse problems.
  - Allows inclusion of field know-how.
  - Fast cross-validation formulae unlock path to model selection.
Implementation

- GPs priors can handle large-scale inversion via implicit representation of the posterior covariance.
- Implicit representation allows fast updates and can be run on GPU.

Theory

- Disintegrations of Gaussian measures provide rigorous treatment of conditioning under linear operator data.
- Mild conditions on the GP guarantee process $\iff$ measure equivalence.

Modelling

- Universal kriging can be extended to inverse problems.
- Allows inclusion of field know-how.
- Fast cross-validation formulae unlock path to model selection.
Implementation
- GPs priors can handle large-scale inversion via implicit representation of the posterior covariance.
- Implicit representation allows fast updates and can be run on GPU.

Theory
- Disintegrations of Gaussian measures provide rigorous treatment of conditioning under linear operator data.
- Mild conditions on the GP guarantee process $\Longleftrightarrow$ measure equivalence.

Modelling
- Universal kriging can be extended to inverse problems.
- Allows inclusion of field know-how.
- Fast cross-validation formulae unlock path to model selection.
David Ginsbourger and Cedric Schärer, *Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances*. 