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)

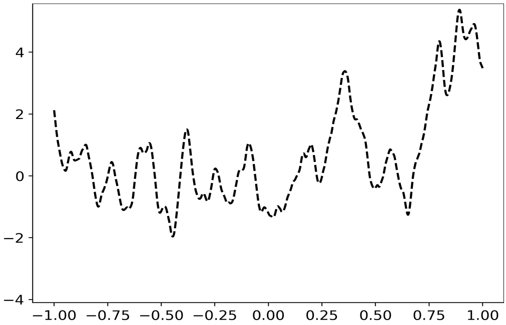
Section 1

Introduction

Problem Setup

Want to learn an unknown function

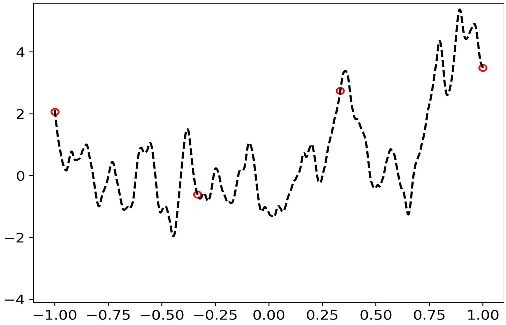
$$f : D \to \mathbb{R}$$



Problem Setup

Want to learn an unknown function

$$f : D \to \mathbb{R}$$



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 $Z \sim \mathrm{Gp}(m_0, k)$.

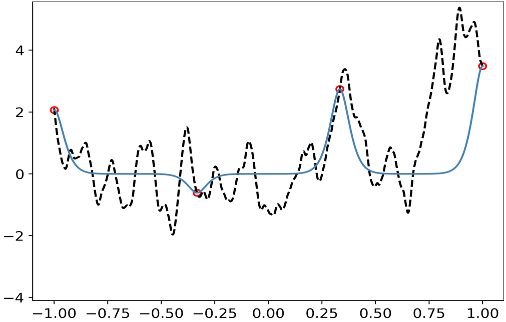


Figure: Posterior mean (blue).

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 ℓ_i.

What if do not have point data $f(x_i)$ but more general data:

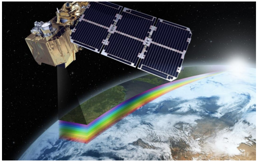
$$y_i = \ell_i(f)$$

for some linear functionals ℓ_i.

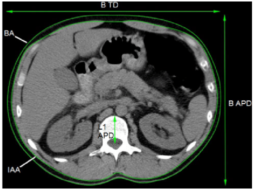
Examples

- 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.

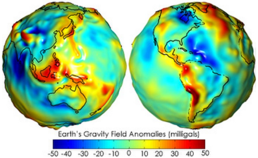
Linear operator data arise everywhere in science:



Remote Sensing



Tomography



Geoscience

- Observe reflected light
- Recover land properties

- Observe transmitted X-Ray intensity
- Recover material properties

- Observe gravitational field
- Recover rock density

Broadly known as (linear) **Inverse Problems**.

Inverse Problems and GP

GPs can easily handle linear operator data.

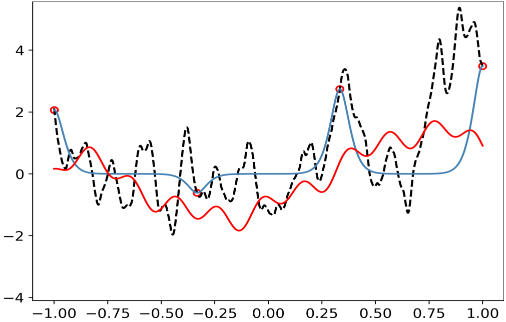


Figure: Posterior mean (red) for Fourier data

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

.

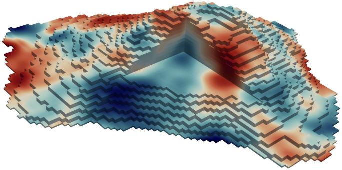
$\implies$ can use GP priors in inverse problems (Bayesian Inversion)

Problem for today:

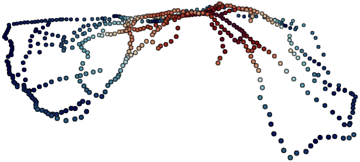
Can we scale the GP + linear operator framework to "real-world" problems?

Example Real-World Problem: Gravimetric Inversion

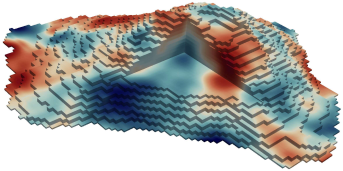
Recover interior of Stromboli volcano from surface gravity.



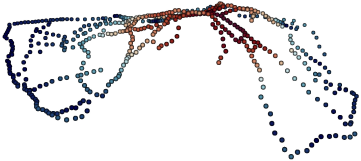
density



observed gravity



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 ρ 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 ρ 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 $\boldsymbol{x} = (x_1, ..., x_m)$.
- Observation model

$$\boldsymbol{y} = \boldsymbol{G}(\rho(x_1), ..., \rho(x_m))^T$$

.

- Discretized observation operator $\boldsymbol{G} \in \mathbb{R}^{n \times m}$ for n observations.
- Posterior described by:
  - mean vector $\tilde{\boldsymbol{m}} = (\tilde{m}_{x_1}, ..., \tilde{m}_{x_m})^T$
  - covariance matrix $\tilde{\boldsymbol{K}} = \left(\tilde{k}(x_i, x_j)\right)_{i,j=1,...,m}$

The Problem with Large, Sequential Bayesian Inversion

$$\tilde{\boldsymbol{m}} = \boldsymbol{m}_0 + \boldsymbol{K}\boldsymbol{G}^T \left(\boldsymbol{G}\boldsymbol{K}\boldsymbol{G}^T + \tau^2 \boldsymbol{I}\right)^{-1} (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{m}_0)$$
$$\tilde{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{K}\boldsymbol{G}^T \left(\boldsymbol{G}\boldsymbol{K}\boldsymbol{G}^T + \tau^2 \boldsymbol{I}\right)^{-1} \boldsymbol{G}\boldsymbol{K}$$

Linear operator data (can) involve **all** grid points at the same time.

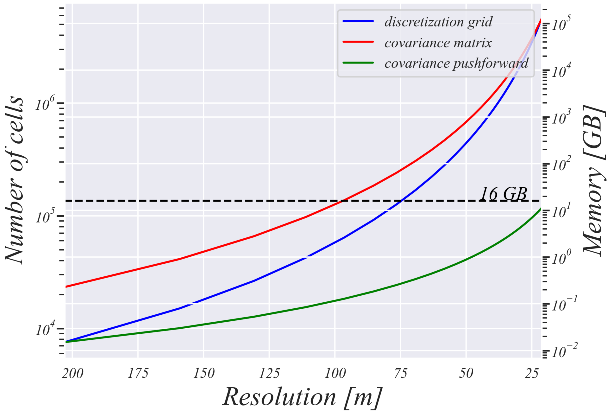


Figure: Grid and matrices size vs resolution on Stromboli example.

Section 2

Implicit Covariance Representation for Large-Scale Inversion

Solving **practical** difficulties of Bayesian inversion

- Covariance matrix too big? $\Longrightarrow$ Don't store it, nor build it.

Implicit Representation

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

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

$\Longrightarrow$ Only need to maintain a multiplication routine.

Travelletti, C., Ginsbourger, D. and Linde, N. (2022). *Uncertainty Quantification and Experimental Design for Large-Scale Linear Inverse Problems under Gaussian Process Priors* *https://arxiv.org/abs/2109.03457*.

A Theoretical Detour

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

$$Z_{x_1}, ..., Z_{x_m}$$

- 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

$$G : X \to Y$$

into a separable Banach space Y.

For Gaussian measures, conditioning done by **disintegration**:

Theorem

Let X, Y be real separable Banach spaces and μ be a Gaussian measure on the Borel σ-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 $\left(\mu_{|G=y}\right)_{y \in Y}$ of μ 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 $\iff$ 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 ✓.

Gaussian Processes $\iff$ 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 ✓.

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 $\mathcal{H}_k^\theta$.

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

Rajput, B. S. and Cambanis, S. (1972). *Gaussian processes and Gaussian measures* Annals of Mathematical Statistics 43, 1944–1952.

Tarieladze, V. and Vakhania, N. (2007). *Disintegration of Gaussian measures and average-case optimal algorithms* Journal of Complexity 23(4), 851–866.

Steinwart, I. (2019). *Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties* Potential Analysis 51(3), 361–395.

Travelletti, C. and Ginsbourger, D. (2022). *Disintegration of Gaussian Measures for Sequential Bayesian Learning with Linear Operator Data (To appear on Arxiv.).*

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.

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

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

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

Implicit Representation: Advantages

- Drastically reduced memory footprint.

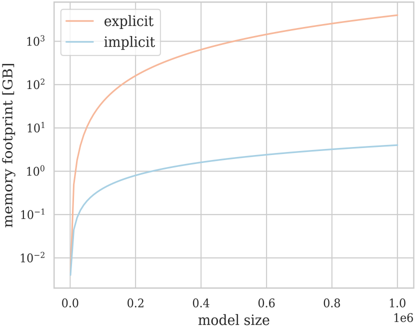


Figure: Memory footprint of posterior covariance vs grid size.

- 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.

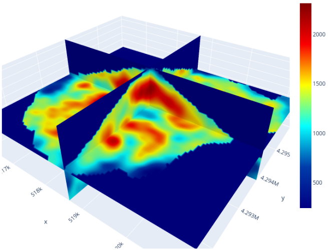


Figure: Posterior mean $[kg/m^3]$.

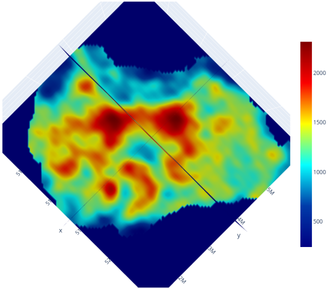
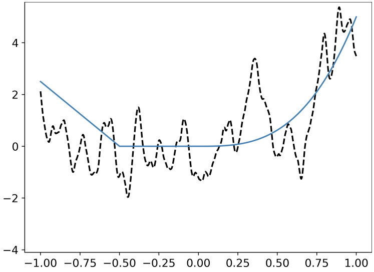


Figure: Posterior mean $[kg/m^3]$.

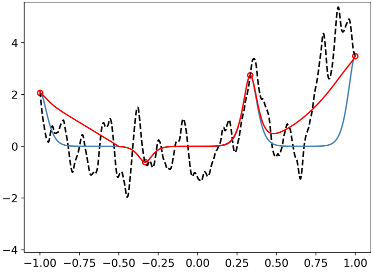
Section 3

Universal Inversion

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

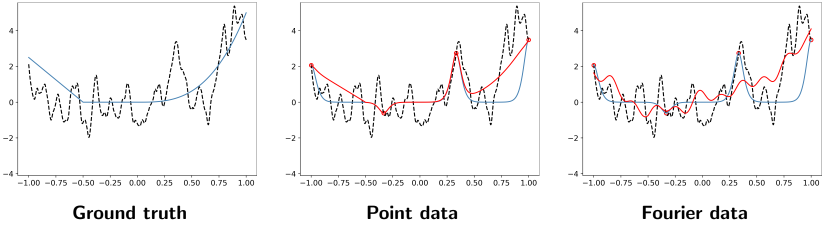


Ground truth



Point data

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 stuctures such as:

- Layers
- Chimneys
- Depth-dependence
- ...

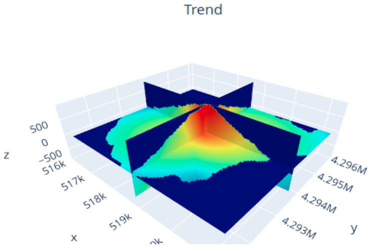


Figure: Radial trend (chimney).

Prior with trend

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

$$Z_x = F_x \boldsymbol{\beta} + \eta_x,$$

- η 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

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

Theorem

Conditionally on linear operator data $\boldsymbol{Y} = GZ_{\boldsymbol{w}} + \boldsymbol{\epsilon}$, *the posterior of the trend coefficients is Gaussian with mean and covariance given by:*

$$\mathbb{E}\left[\boldsymbol{\beta}|\boldsymbol{y}\right] = \boldsymbol{\mu} + \Sigma \mathcal{F}_{\boldsymbol{w}}^T G^T Q_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y} - G\mathcal{F}_{\boldsymbol{w}}\boldsymbol{\mu}\right)$$
$$\mathrm{Cov}\left(\boldsymbol{\beta}, \boldsymbol{\beta}|\boldsymbol{y}\right) = \Sigma - \Sigma \mathcal{F}_{\boldsymbol{w}}^T G^T Q_{\boldsymbol{y}}^{-1} G \mathcal{F}_{\boldsymbol{w}} \Sigma,$$

assuming that the matrix $Q_{\boldsymbol{y}} := G\left(\mathcal{F}_{\boldsymbol{w}}\Sigma\mathcal{F}_{\boldsymbol{w}}^T + K_{\boldsymbol{w}\boldsymbol{w}}\right)G^T + \sigma_\epsilon^2 I$ *is invertible.*

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

$$m_{Z|\boldsymbol{y}}\left(\boldsymbol{x}\right) = \mathcal{F}_{\boldsymbol{x}}\boldsymbol{\mu} + \left(\mathcal{F}_{\boldsymbol{x}}\Sigma\mathcal{F}_{\boldsymbol{w}}^T + K_{\boldsymbol{x}\boldsymbol{w}}\right)G^T Q_{\boldsymbol{y}}^{-1}\left(\boldsymbol{y} - G\mathcal{F}_{\boldsymbol{w}}\boldsymbol{\mu}\right)$$
$$k_{Z|\boldsymbol{y}}\left(\boldsymbol{x}, \boldsymbol{x'}\right) = K_{\boldsymbol{x}\boldsymbol{x'}} + \mathcal{F}_{\boldsymbol{x}}\Sigma\mathcal{F}_{\boldsymbol{x'}}^T - \left(\mathcal{F}_{\boldsymbol{x}}\Sigma\mathcal{F}_{\boldsymbol{w}}^T + K_{\boldsymbol{x}\boldsymbol{w}}\right)G^T Q_{\boldsymbol{y}}^{-1} G \left(\mathcal{F}_{\boldsymbol{w}}\Sigma\mathcal{F}_{\boldsymbol{x'}}^T + K_{\boldsymbol{w}\boldsymbol{x'}}\right)$$

Cross Validation

Again consider dataset consisting of two batches of data $\boldsymbol{Y} = (\boldsymbol{Y_i}, \boldsymbol{Y_{-i}})$.

Cross Validation

Want to compute residual when we predict $\boldsymbol{Y_i}$ using $\boldsymbol{Y_{-i}}$

$$\boldsymbol{Y_i} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{(-\boldsymbol{i})}$$

- Want fast formula for CV residual (avoid full recomputation of predictor).
- Formula should be valid for any subset of indices $\boldsymbol{i}$ (k-fold).

Fast k-fold Cross-Validation Formulae

By generalizing [GS] all the information we need is contained in the augmented matrix:

$$\tilde{K} = \begin{pmatrix} GKG^T & GF \\ F^T G^T & \mathbf{0} \end{pmatrix}.$$

We partition it as:

$$\tilde{K} = \begin{pmatrix} \tilde{K}_{ii} & \tilde{K}_{i-i} \\ \tilde{K}_{-ii} & \tilde{K}_{-i-i} \end{pmatrix} = \begin{pmatrix} G_{\boldsymbol{i}\bullet} K G^T_{\bullet\boldsymbol{i}} & G_{\boldsymbol{i}\bullet} K G^T_{\bullet-\boldsymbol{i}} & G_{\boldsymbol{i}\bullet} F \\ G_{-\boldsymbol{i}\bullet} K G^T_{\bullet\boldsymbol{i}} & G_{-\boldsymbol{i}\bullet} K G^T_{\bullet-\boldsymbol{i}} & G_{-\boldsymbol{i}\bullet} F \\ F^T G^T_{\bullet\boldsymbol{i}} & F^T G_{\bullet-\boldsymbol{i}} & \mathbf{0} \end{pmatrix}.$$

CV residuals can the be computed by extracting subblocks of the inverse.

Fast k-fold Cross-Validation Formulae

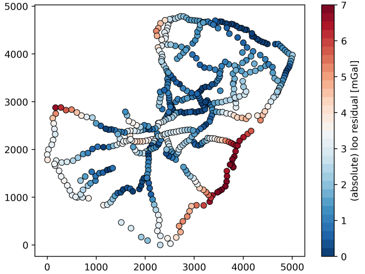
Upper left block of the inverse give us (inverse) covariance of residuals:

$$\tilde{K}_{\boldsymbol{ii}}^{-1} = \mathrm{Cov}\left(\hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{(-\boldsymbol{i})}, \hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{(-\boldsymbol{i})}\right)^{-1},$$

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

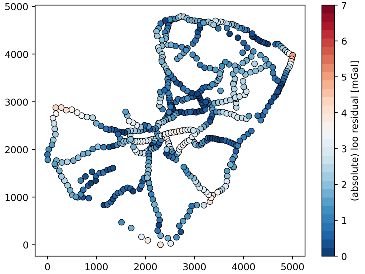
$$\tilde{K}_{\boldsymbol{ii}}^{-1}\left(\tilde{K}^{-1}\left[:, 1:n\right]\boldsymbol{Y}\right)_{\boldsymbol{i}} = \boldsymbol{Y}_{\boldsymbol{i}} - \hat{\boldsymbol{Y}}_{\boldsymbol{i}}^{(-\boldsymbol{i})}.$$

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 rigorours 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.

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 rigorours 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.

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 rigorours 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.

References I

David Ginsbourger and Cedric Schärer, *Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances*.