Universal Inversion: a Framework for Infusing Expert Knowledge in Bayesian Inverse Problems

Cedric Travelletti

February 28, 2023

Joint work with D. Ginsbourger, A. Gautier and N. Linde (UniL)

Section 1

Introduction: Inverse Problems and Stromboli Volcano

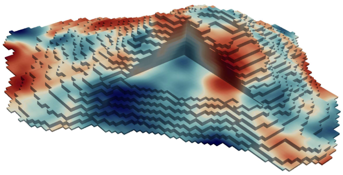
Motivating Example: Stromboli Volcano

Want to learn the interior structure of the Stromboli volcano.

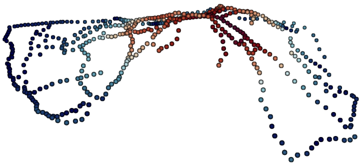
- Only allowed to measure gravitational field on the surface

Example Real-World Problem: Gravimetric Inversion

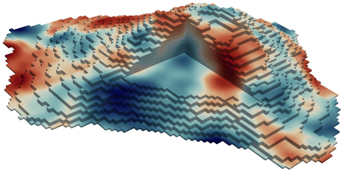
Want to recover interior density field $\rho : D \to \mathbb{R}$ from surface gravity.



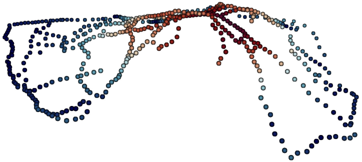
density



observed gravity



density



observed gravity

Properties:

- **Linear operator data**.
- **"Large-scale"**: large 3 dimensional inversion grid.
- **Sequential** assimilation of new data and **optimal design of experiment** 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

$$\boldsymbol{y} = \boldsymbol{G}(\rho(w_1), ..., \rho(w_r))^T =: \boldsymbol{G}\rho_{\boldsymbol{W}}$$

.

- Discretized observation operator $\boldsymbol{G} \in \mathbb{R}^{n \times r}$ for n observations.

Gaussian Process Regression with Linear Operator Data

Assume ρ is a realization of a GP prior $Z \sim \mathrm{GP}(m, k)$.

- Data model is $\boldsymbol{Y} = \boldsymbol{G} Z_{\boldsymbol{W}} + \boldsymbol{\epsilon}$.
- Approximate ρ by posterior, conditional on the observed data $\boldsymbol{Y} = \boldsymbol{y}$.
- Posterior is Gaussian, fully characterized by:
  - mean vector $\tilde{\boldsymbol{m}} = (\tilde{m}_{x_1}, ..., \tilde{m}_{x_m})^T$
  - covariance matrix $\tilde{\boldsymbol{K}} = \left(\tilde{k}(x_i, x_j)\right)_{i,j=1,...,m}$

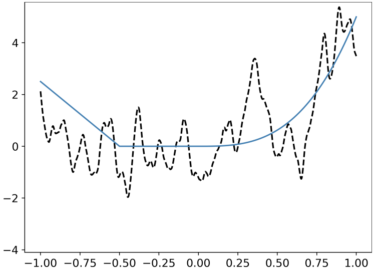
$$\tilde{\boldsymbol{m}} = \boldsymbol{m} + \boldsymbol{K}\boldsymbol{G}^T \left(\boldsymbol{G}\boldsymbol{K}\boldsymbol{G}^T + \boldsymbol{\Delta}\right)^{-1} (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{m})$$
$$\tilde{\boldsymbol{K}} = \boldsymbol{K} - \boldsymbol{K}\boldsymbol{G}^T \left(\boldsymbol{G}\boldsymbol{K}\boldsymbol{G}^T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

Hard to implement in practice [TGL21]

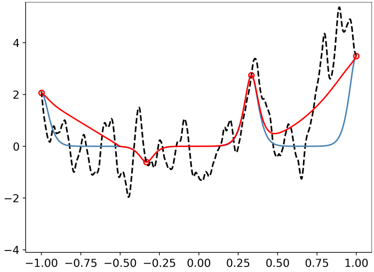
Section 2

Bayesian Inversion with Trends

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

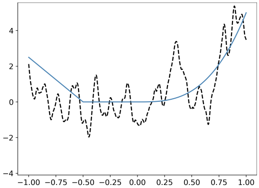


Ground truth

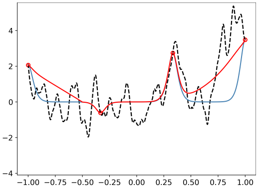


Point data

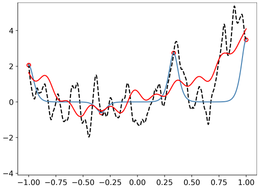
GP regression with trends known as **universal kriging**.



Ground truth



Point data



Fourier data

Goal: Extend to inverse problems to get **"universal inversion"**.

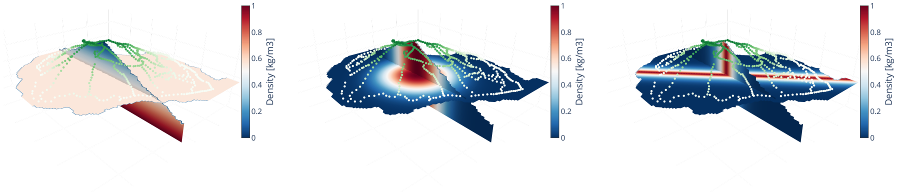
Trends allow inclusion of **expert knowledge**.

Model known geological stuctures such as:

- **Depth-dependence:** resulting from way volcanoes aggregate mass.
- **Layers**
- **Chimneys:** higher densities in central lava conduit.
- **Fault lines:** some volcanoes separate along fault through which high density magma infiltrates.

target = trend + fluctuations

Trends modeled by specifying basis functions:



(a) depth-dependence (b) chimney (c) fault-line

Figure: Trend basis functions used in the experiments. Solid balls denote the locations of the gravimetry observations (Stromboli dataset).

Prior with trend

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

$$Z_x = \sum_{i=1}^{p} \beta_i f_i(x) + \eta_x,$$

- η is a (centred) GP with kernel k,
- basis functions $f_i : D \to \mathbb{R}$,
- trend coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$.

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:*

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

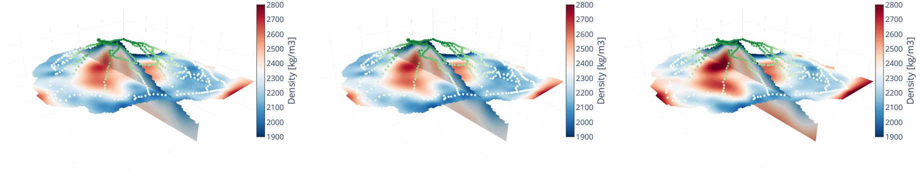
assuming that the matrix $Q_{\boldsymbol{y}} := G\left(F_{\boldsymbol{W}}\Sigma F_{\boldsymbol{W}}^T + K_{\boldsymbol{W}\boldsymbol{W}}\right)G^T + \sigma_\epsilon^2 I$ *is invertible.*

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

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

Volcano Inversion with Trend

Inversion with trends gives consistent results (data fit within noise std).



(a) constant (b) chimney (c) fault-line

Figure: Posterior mean for the various trend models (Stromboli data).

But gives rise to new questions.

Universal Inversion brings new challenges:

Questions

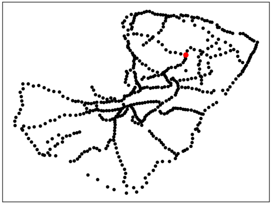
1. Can we compare the performance of different trend models?
2. How can we "rank" different trend models? (penalize for complexity)
3. Can we perform measurements to "optimally" discriminate models?

Cross-Validation: Leave-One-Out

Idea

Given a data model $\boldsymbol{Y} = GZ_{\boldsymbol{W}}$ and observed value $\boldsymbol{y}$:

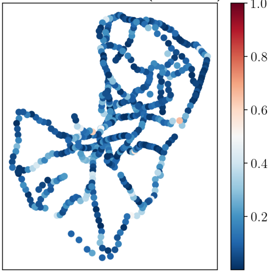
1. Remove one data point $\boldsymbol{y}_i$.
2. Fit model on remaining data $\boldsymbol{y}^{(-i)}$.
3. Compute mean of fitted model $\hat{Z}^{(-i)} := \mathbb{E}[Z|\boldsymbol{Y}^{(-i)} = \boldsymbol{y}^{(-i)}]$
4. Predict missing data using mean, compute residual $e_i := \boldsymbol{y}_i - G\hat{Z}_{\boldsymbol{W}}^{(-i)}$.



$\sum_{i=1}^{n} ||\boldsymbol{y}_i - G\hat{Z}_{\boldsymbol{W}}^{(-i)}||^2$ gives measure of model generalisation error.

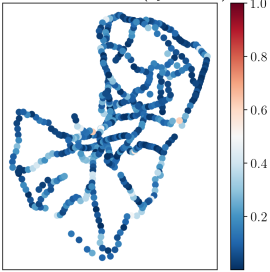
Leave-One Out (LOO) Cross-Validation on Stromboli

Leave-one-out residual (**constant**)



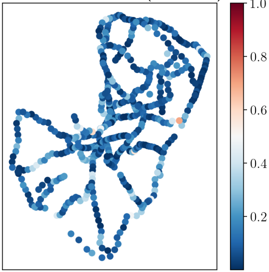
Leave-one-out rmse: 0.1590 [mGal]

Leave-one-out residual (**cylindrical**)



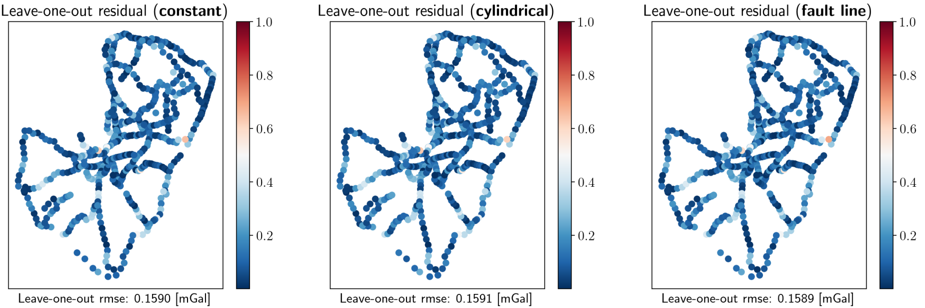
Leave-one-out rmse: 0.1591 [mGal]

Leave-one-out residual (**fault line**)



Leave-one-out rmse: 0.1589 [mGal]

Leave-One Out (LOO) Cross-Validation on Stromboli



Gravimetric observations highly correlated

$\Longrightarrow$ LOO not informative.

Beyond LOO: Leave-k-out

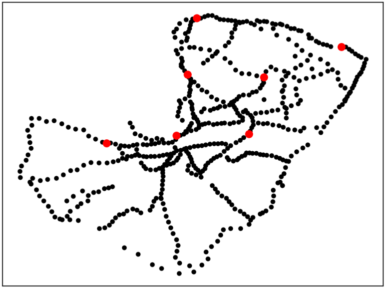
Can also consider subsets of size k: separate data in two batches:

$$\boldsymbol{y} = (\boldsymbol{y_i}, \boldsymbol{y_{-i}}),$$

where $\boldsymbol{y_i}$ has size k.

Want to compute residual when we predict $\boldsymbol{y_i}$ using $\boldsymbol{y_{-i}}$

$$\boldsymbol{e}_i := \boldsymbol{y_i} - G\hat{Z}_{\boldsymbol{W}}^{(-\boldsymbol{i})}$$



Problems

- Have to recompute full posterior at each CV pass.
- Full leave-k-out infeasible (combinatorial explosion).

- Need to select "representative" subsets.
- Need efficient formulae for computing residuals.

Fast k-fold Cross-Validation Formulae

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

$$\tilde{K} = \begin{pmatrix} G K_{\boldsymbol{W}\boldsymbol{W}} G^T & G F_{\boldsymbol{W}} \\ F_{\boldsymbol{W}}^T G^T & \mathbf{0} \end{pmatrix}.$$

We partition it as:

$$\tilde{K} = \begin{pmatrix} \tilde{K}_{\boldsymbol{i}\boldsymbol{i}} & \tilde{K}_{\boldsymbol{i}-\boldsymbol{i}} \\ \tilde{K}_{-\boldsymbol{i}\boldsymbol{i}} & \tilde{K}_{-\boldsymbol{i}-\boldsymbol{i}} \end{pmatrix} = \begin{pmatrix} G_{\boldsymbol{i}\bullet} K_{\boldsymbol{W}\boldsymbol{W}} G^T_{\bullet\boldsymbol{i}} & G_{\boldsymbol{i}\bullet} K_{\boldsymbol{W}\boldsymbol{W}} G^T_{\bullet-\boldsymbol{i}} & G_{\boldsymbol{i}\bullet} F_{\boldsymbol{W}} \\ G_{-\boldsymbol{i}\bullet} K_{\boldsymbol{W}\boldsymbol{W}} G^T_{\bullet\boldsymbol{i}} & G_{-\boldsymbol{i}\bullet} K_{\boldsymbol{W}\boldsymbol{W}} G^T_{\bullet-\boldsymbol{i}} & G_{-\boldsymbol{i}\bullet} F_{\boldsymbol{W}} \\ F^T_{\boldsymbol{W}} G^T_{\bullet\boldsymbol{i}} & F^T_{\boldsymbol{W}} G_{\bullet-\boldsymbol{i}} & \mathbf{0} \end{pmatrix},$$

where $\tilde{K}_{\boldsymbol{i}\boldsymbol{i}} := G_{\boldsymbol{i}\bullet} K_{\boldsymbol{W}\boldsymbol{W}} G^T_{\bullet\boldsymbol{i}}$.

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

Fast k-fold Cross-Validation Formulae

Residuals can be computed by extracting and inverting upper left block of the inverse:

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

where $\tilde{K}_{\boldsymbol{ii}}^{-1}$ denotes the $\boldsymbol{ii}$ sub-block of $\tilde{K}^{-1}$ and $\boldsymbol{y}$ has dimension n.

Drasting computational savings by avoiding recomputation of full posterior.

Application: k-fold CV to detect trends in Stromboli

Instead of considering all data subsets of a given size, consider k different subsets (folds).

- Remove one fold and predict using remaining ones.

Choosing Folds

How can we choose the k subsets?

- Start with spatial clustering (heuristic).

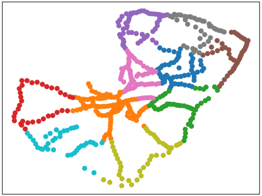
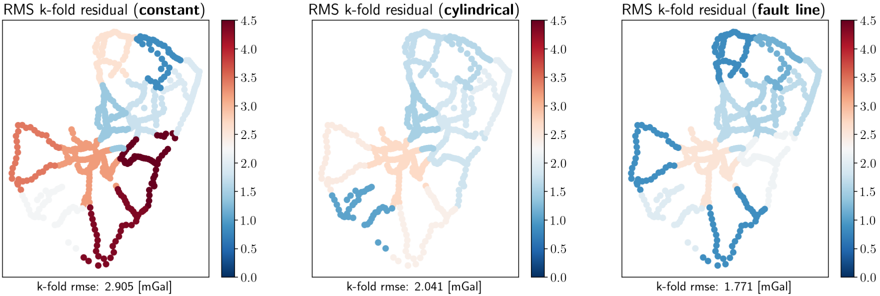


Figure: Folds from kMeans clustering, $k = 10$.

K-fold Cross-Validation: Stromboli (experimental)

kMean CV favors **fault line** model.



Question?

Is this representative of generalisation error?

Beyond CV: Distribution of the Residuals

Can consider the residuals as random variables

$$\boldsymbol{E_i} := G_{\boldsymbol{i}} Z_{\boldsymbol{W}} - G_{\boldsymbol{i}} \hat{Z}_{\boldsymbol{W}}^{(-\boldsymbol{i})},$$

where $\hat{Z}^{(-\boldsymbol{i})}$ is the UK predictor based on evaluation of $G_{-\boldsymbol{i}} Z$.

Theorem

(under the model) The residuals are jointly Gaussian distributed, centred and with covariance

$$\operatorname{Cov}\left(E_{\boldsymbol{i}}, E_{\boldsymbol{j}}\right) = \left(\tilde{K}^{-1}_{\boldsymbol{ii}}\right)^{-1} \tilde{K}^{-1}_{\boldsymbol{ij}} \left(\tilde{K}^{-1}_{\boldsymbol{jj}}\right)^{-1}$$

What Next?

How can we leverage knowledge of theoretical distribution of residuals?

Choose folds to mimick generalisation error?

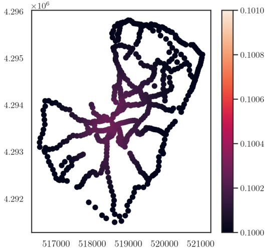


Figure: Theoretical CV variances (constant trend).

Gaussian hypothesis too "smooth" (few gravimetric observations kill all variance).

References I

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

Cédric Travelletti, David Ginsbourger, and Niklas Linde, *Uncertainty quantification and experimental design for large-scale linear inverse problems under gaussian process priors*, arXiv 2109.03457 (to appear in SIAM JUQ), 2021.