On Gaussian Process multiple-fold cross-validation

David Ginsbourger (University of Bern)

Based on joint works with Athénaïs Gautier, Cedric Schärer, Cédric Travelletti.

DG, AG and CT acknowledge support from the Swiss National Science Foundation project number 178858.

Gaussian Process Summer School 2024
University of Manchester

September 11th 2024

Outline

Outline

FAQ

What is the talk about?

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework?

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework? Yes and no. Linear (ridge) regression, RKHS regularization, etc. enjoy this approach in full or in part.

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework? Yes and no. Linear (ridge) regression, RKHS regularization, etc. enjoy this approach in full or in part.

What if I am not familar with GP models, cross-validation, and all that?

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework? Yes and no. Linear (ridge) regression, RKHS regularization, etc. enjoy this approach in full or in part.

What if I am not familar with GP models, cross-validation, and all that? Take a deep breath, relax, we will briefly recall useful basics before take off!

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework? Yes and no. Linear (ridge) regression, RKHS regularization, etc. enjoy this approach in full or in part.

What if I am not familar with GP models, cross-validation, and all that? Take a deep breath, relax, we will briefly recall useful basics before take off!

And if I am already an expert of GPs. . . and all that?

FAQ

What is the talk about? A fast approach to calculate cross-validation residuals & their distribution in the framework of Gaussian Process models.

Do the results only apply to this framework? Yes and no. Linear (ridge) regression, RKHS regularization, etc. enjoy this approach in full or in part.

What if I am not familar with GP models, cross-validation, and all that? Take a deep breath, relax, we will briefly recall useful basics before take off!

And if I am already an expert of GPs. . . and all that? You might still enjoy colorful block inversions and multiple-fold cross-validation on the Stromboli!

What is cross-validation (CV)?

CV has developed as an important approach in model selection, see

S. Arlot, A. Celisse (2010).
A survey of cross-validation procedures for model selection
Statist. Surv. 4: 40-79.

and references therein. Several variations of CV (hold-out, leave-one-out, leave-multiple-out, multiple-fold) have been use in various contexts.

What is cross-validation (CV)?

CV has developed as an important approach in model selection, see

S. Arlot, A. Celisse (2010).
A survey of cross-validation procedures for model selection
Statist. Surv. 4: 40-79.

and references therein. Several variations of CV (hold-out, leave-one-out, leave-multiple-out, multiple-fold) have been use in various contexts.

The essence of CV is to leave part of the available data / training set away (a "fold"), perform predictions at the left out "points" based on the remaining data, and compare predicted versus left out responses.

Of what use is CV for GP modelling?

In GP modelling, cross-validation (CV) has been used for

- diagnosing models without requiring external/validation data,
- estimating hyperparameters (via criteria building on CV outputs),
- and also, for guiding sequential design strategies

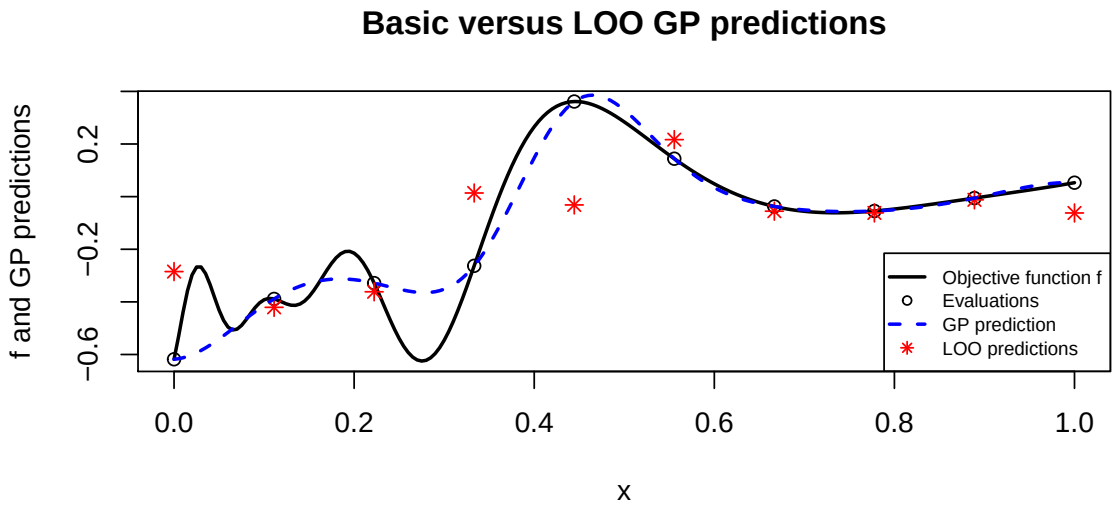
Of what use is CV for GP modelling?

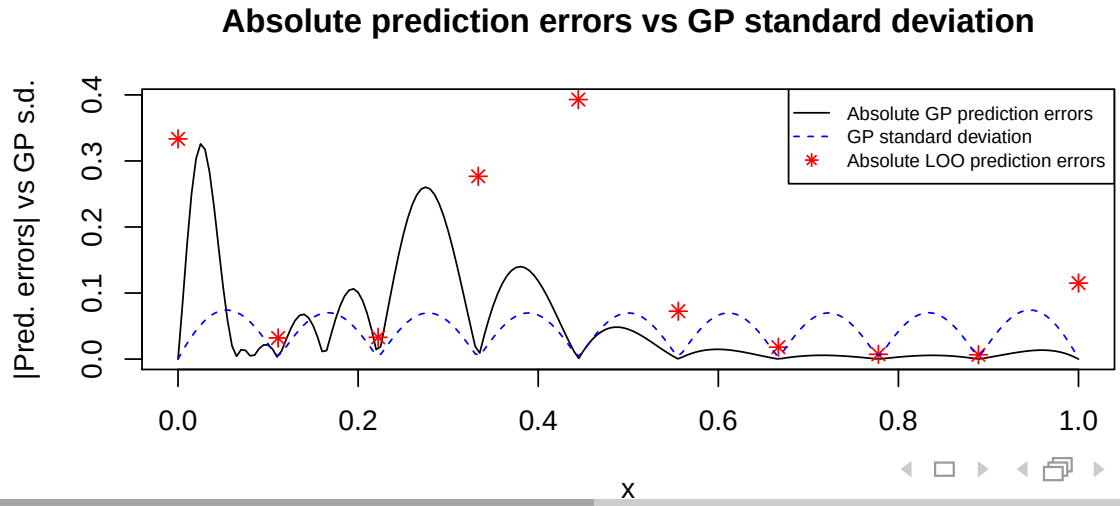
In GP modelling, cross-validation (CV) has been used for

- diagnosing models without requiring external/validation data,
- estimating hyperparameters (via criteria building on CV outputs),
- and also, for guiding sequential design strategies

The most commonly implemented CV approach in GP modelling is doubtlessly the so-called Leave-One-Out Cross-Validation (LOO-CV).

LOO-CV based on regularly spaced points





Gaussian Process model(s)

Let D be a set.

A real-valued Gaussian Process $\xi = (\xi(\mathbf{x}))_{\mathbf{x}\in D}$ indexed by D is a collection of random variables over a common probability space such that, for any $n \in \mathbb{N}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, $(\xi(\mathbf{x}_1), \ldots, \xi(\mathbf{x}_n))$ has a Gaussian (joint) distribution.

Gaussian Process model(s)

Let D be a set.

A real-valued Gaussian Process $\xi = (\xi(\mathbf{x}))_{\mathbf{x}\in D}$ indexed by D is a collection of random variables over a common probability space such that, for any $n \in \mathbb{N}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$, $(\xi(\mathbf{x}_1), \ldots, \xi(\mathbf{x}_n))$ has a Gaussian (joint) distribution.

Considered observation model

For $n \in \mathbb{N}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ considered as fixed, one observes for $1 \leq i \leq n$

$$\begin{aligned} Z_i &= \xi(\mathbf{x}_i) + \varepsilon_i \\ &= \sum_{j=1}^{p} \beta_j f_j(\mathbf{x}_i) + \eta(\mathbf{x}_i) + \varepsilon_i \end{aligned}$$

where η is a centred real-valued GP with covariance kernel k, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \sim \mathcal{N}(\mathbf{0}, \Sigma_\varepsilon)$ independently of η.

Considered (BLU) predictors in a nutshell

In linear prediction (with β unknown, or a null trend), a predictor of the form

$$\widehat{\xi}(\mathbf{x}) = \boldsymbol{\lambda}(\mathbf{x})^\top \mathbf{Z} \quad (\mathbf{x} \in D)$$

is sought, where $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$.

The BLUP at $\mathbf{x}$ is obtained by determining $\boldsymbol{\lambda}(\mathbf{x})$ so as to minimize

$$\mathrm{Var}[\xi(\mathbf{x}) - \widehat{\xi}(\mathbf{x})] = k(\mathbf{x},\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^\top (K + \Sigma_\varepsilon)\boldsymbol{\lambda}(\mathbf{x}) - 2\mathbf{k}(\mathbf{x})^\top K \boldsymbol{\lambda}(\mathbf{x})$$

s.t. $F^\top \boldsymbol{\lambda}(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, where

- $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j \in \{1,\ldots,n\}}$,
- $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_i))_{i \in \{1,\ldots,n\}}$,
- $F = (f_j(\mathbf{x}_i))_{1 \leq i \leq n, 1 \leq j \leq p} \in \mathbb{R}^{n \times p}$,
- and $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))^\top \in \mathbb{R}^p$.

Universal Kriging equations ($\boldsymbol{\beta}$ unknown)

Assuming that $K + \Sigma_\varepsilon$ and $\mathcal{I}_{\boldsymbol{\beta}} = F^\top (K + \Sigma_\varepsilon)^{-1} F$ are invertible, solving for $\boldsymbol{\lambda}(\mathbf{x})$ delivers the Universal Kriging predictor, that can ultimately be written as

$$\widehat{\xi}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}} + \mathbf{k}(\mathbf{x})^\top (K + \Sigma_\varepsilon)^{-1} (\mathbf{Z} - F\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}} = \mathcal{I}_{\boldsymbol{\beta}}^{-1} F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{Z}$.

Universal Kriging equations ($\boldsymbol{\beta}$ unknown)

Assuming that $K + \Sigma_\varepsilon$ and $\mathcal{I}_{\boldsymbol{\beta}} = F^\top (K + \Sigma_\varepsilon)^{-1} F$ are invertible, solving for $\boldsymbol{\lambda}(\mathbf{x})$ delivers the Universal Kriging predictor, that can ultimately be written as

$$\widehat{\xi}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}} + \mathbf{k}(\mathbf{x})^\top (K + \Sigma_\varepsilon)^{-1} (\mathbf{Z} - F\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}} = \mathcal{I}_{\boldsymbol{\beta}}^{-1} F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{Z}$.

Furthermore the residual covariance writes, for arbitrary $\mathbf{x}, \mathbf{x}' \in D$:

$$\begin{aligned}\mathrm{Cov}[\xi(\mathbf{x}) - \widehat{\xi}(\mathbf{x}), \xi(\mathbf{x}') - \widehat{\xi}(\mathbf{x}')] &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}') \\ &+ (\mathbf{f}(\mathbf{x}) - F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}))^\top \mathcal{I}_{\boldsymbol{\beta}}^{-1} (\mathbf{f}(\mathbf{x}') - F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}')).\end{aligned}$$

In particular, $\mathrm{Var}[\xi(\mathbf{x}) - \widehat{\xi}(\mathbf{x})]$ gives the prediction variance.

Universal Kriging equations ($\boldsymbol{\beta}$ unknown)

Assuming that $K + \Sigma_\varepsilon$ and $\mathcal{I}_{\boldsymbol{\beta}} = F^\top (K + \Sigma_\varepsilon)^{-1} F$ are invertible, solving for $\boldsymbol{\lambda}(\mathbf{x})$ delivers the Universal Kriging predictor, that can ultimately be written as

$$\widehat{\xi}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \widehat{\boldsymbol{\beta}} + \mathbf{k}(\mathbf{x})^\top (K + \Sigma_\varepsilon)^{-1} (\mathbf{Z} - F\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}} = \mathcal{I}_{\boldsymbol{\beta}}^{-1} F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{Z}$.

Furthermore the residual covariance writes, for arbitrary $\mathbf{x}, \mathbf{x}' \in D$:

$$\begin{aligned}\mathrm{Cov}[\xi(\mathbf{x}) - \widehat{\xi}(\mathbf{x}), \xi(\mathbf{x}') - \widehat{\xi}(\mathbf{x}')] &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}') \\ &+ (\mathbf{f}(\mathbf{x}) - F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}))^\top \mathcal{I}_{\boldsymbol{\beta}}^{-1} (\mathbf{f}(\mathbf{x}') - F^\top (K + \Sigma_\varepsilon)^{-1} \mathbf{k}(\mathbf{x}')).\end{aligned}$$

In particular, $\mathrm{Var}[\xi(\mathbf{x}) - \widehat{\xi}(\mathbf{x})]$ gives the prediction variance.

NB: GLS is a special case ($k \equiv 0$). Simple Kriging assumes a known trend.

More on CV notation and the settings of the example

Throughout the presentation, we denote by **i** a non-void vector of ordered indices from $\{1, \ldots, n\}$ and by $\mathcal{S}$ the set of all such vectors.

More on CV notation and the settings of the example

Throughout the presentation, we denote by **i** a non-void vector of ordered indices from $\{1, \ldots, n\}$ and by $\mathcal{S}$ the set of all such vectors.

We further denote by $\mathbf{Z}[\mathbf{i}]$ the corresponding subvector extracted from $\mathbf{Z}$ and by $\widehat{Z}^{(-\mathbf{i})}[\mathbf{i}]$ the GP prediction of $\mathbf{Z}[\mathbf{i}]$ based on the remaining components of $\mathbf{Z}$.

More on CV notation and the settings of the example

Throughout the presentation, we denote by $\mathbf{i}$ a non-void vector of ordered indices from $\{1, \ldots, n\}$ and by $\mathcal{S}$ the set of all such vectors.

We further denote by $\mathbf{Z}[\mathbf{i}]$ the corresponding subvector extracted from $\mathbf{Z}$ and by $\widehat{Z}^{(-\mathbf{i})}[\mathbf{i}]$ the GP prediction of $\mathbf{Z}[\mathbf{i}]$ based on the remaining components of $\mathbf{Z}$.

$\mathbf{E_i} = \mathbf{Z}[\mathbf{i}] - \widehat{Z}^{(-\mathbf{i})}[\mathbf{i}]$ denotes the corresponding (CV) residual.

More on CV notation and the settings of the example

Throughout the presentation, we denote by $\mathbf{i}$ a non-void vector of ordered indices from $\{1, \ldots, n\}$ and by $\mathcal{S}$ the set of all such vectors.

We further denote by $\mathbf{Z}[\mathbf{i}]$ the corresponding subvector extracted from $\mathbf{Z}$ and by $\widehat{Z}^{(-\mathbf{i})}[\mathbf{i}]$ the GP prediction of $\mathbf{Z}[\mathbf{i}]$ based on the remaining components of $\mathbf{Z}$.

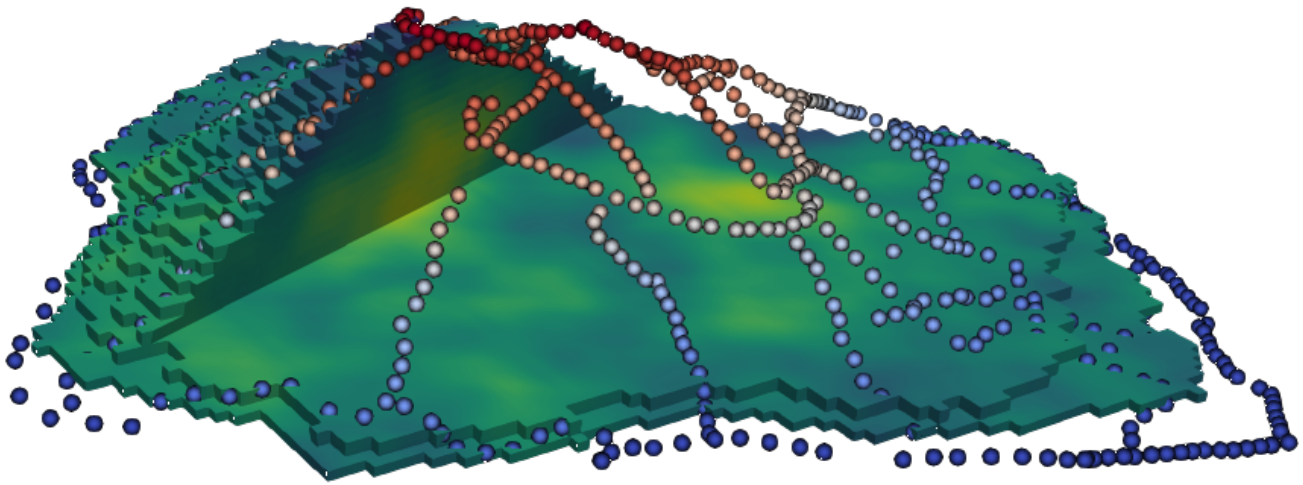
$\mathbf{E_i} = \mathbf{Z}[\mathbf{i}] - \widehat{Z}^{(-\mathbf{i})}[\mathbf{i}]$ denotes the corresponding (CV) residual.

In LOO-CV, the considered index vectors are the singletons $i \in \{1, \ldots, n\}$.

In the previous example, the GP model is applied to a deterministic function and there is no observation noise ($\Sigma_\varepsilon = \mathbf{0}$).

The trend is an estimated constant (Ordinary Kriging). The kernel is of the form $k(\mathbf{x}, \mathbf{x}') = \sigma^2 r(\mathbf{x} - \mathbf{x}')$ with r a Matérn correlation function $\nu = 5/2$.

Another example: Gravimetric inversion on Stromboli



Broader goals: reconstruct the mass density inside Stromboli from gravimetric measurements on its surface. We use a GP model under integral observations (collaboration with Prof. Niklas Linde, University of Lausanne).

In Simple Kriging settings, Bachoc presented several results (building upon an approach pioneered by Dubrule in various settings) highlighting the instrumental role of $Q = \Sigma^{-1}$ in calculating LOO-CV residuals.

F. Bachoc (2013).

Cross validation and maximum likelihood estimation of hyperparameters of gaussian processes with model misspecification.

Computational Statistics and Data Analysis, 66:55-69.

O. Dubrule (1983).

Cross validation of kriging in a unique neighborhood.

Journal of the International Association of Mathematical Geology 15, 687-699.

In Simple Kriging settings, Bachoc presented several results (building upon an approach pioneered by Dubrule in various settings) highlighting the instrumental role of $Q = \Sigma^{-1}$ in calculating LOO-CV residuals.

F. Bachoc (2013).

Cross validation and maximum likelihood estimation of hyperparameters of gaussian processes with model misspecification.

Computational Statistics and Data Analysis, 66:55-69.

O. Dubrule (1983).

Cross validation of kriging in a unique neighborhood.

Journal of the International Association of Mathematical Geology 15, 687-699.

The LOO residuals $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)^\top$ can be written in compact form:

$$\mathbf{E} = \mathrm{diag}((Q[1], \ldots, Q[n])^{-1})Q\mathbf{Z}.$$

In Simple Kriging settings, Bachoc presented several results (building upon an approach pioneered by Dubrule in various settings) highlighting the instrumental role of $Q = \Sigma^{-1}$ in calculating LOO-CV residuals.

F. Bachoc (2013).

Cross validation and maximum likelihood estimation of hyperparameters of gaussian processes with model misspecification.

Computational Statistics and Data Analysis, 66:55-69.

O. Dubrule (1983).

Cross validation of kriging in a unique neighborhood.

Journal of the International Association of Mathematical Geology 15, 687-699.

The LOO residuals $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)^\top$ can be written in compact form:

$$\mathbf{E} = \mathrm{diag}((Q[1], \ldots, Q[n])^{-1})Q\mathbf{Z}.$$

In turn, this formulas also provide a means efficiently calculate standardized residuals (commonly used within qq-plots for diagnostics) via $Q\mathbf{Z}$.

In Simple Kriging settings, Bachoc presented several results (building upon an approach pioneered by Dubrule in various settings) highlighting the instrumental role of $Q = \Sigma^{-1}$ in calculating LOO-CV residuals.

F. Bachoc (2013).

Cross validation and maximum likelihood estimation of hyperparameters of gaussian processes with model misspecification.

Computational Statistics and Data Analysis, 66:55-69.

O. Dubrule (1983).

Cross validation of kriging in a unique neighborhood.

Journal of the International Association of Mathematical Geology 15, 687-699.

The LOO residuals $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)^\top$ can be written in compact form:

$$\mathbf{E} = \mathrm{diag}((\mathrm{Q}[1], \ldots, \mathrm{Q}[\mathrm{n}])^{-1})\mathrm{Q}\mathbf{Z}.$$

In turn, this formulas also provide a means efficiently calculate standardized residuals (commonly used within qq-plots for diagnostics) via $Q\mathbf{Z}$.

In the context where $\Sigma = K = \sigma^2 R$, it has been used to estimate σ^2 by setting $||Q\mathbf{Z}||^2 = n$, leading to $\widehat{\sigma^2}_{\mathrm{LOO}} = \frac{1}{n}\mathbf{Z}^\top R^{-1}(\mathrm{diag}(R^{-1}))^{-1}R^{-1}\mathbf{Z}$.

Computational speed-ups of fast versus “naive” LOO

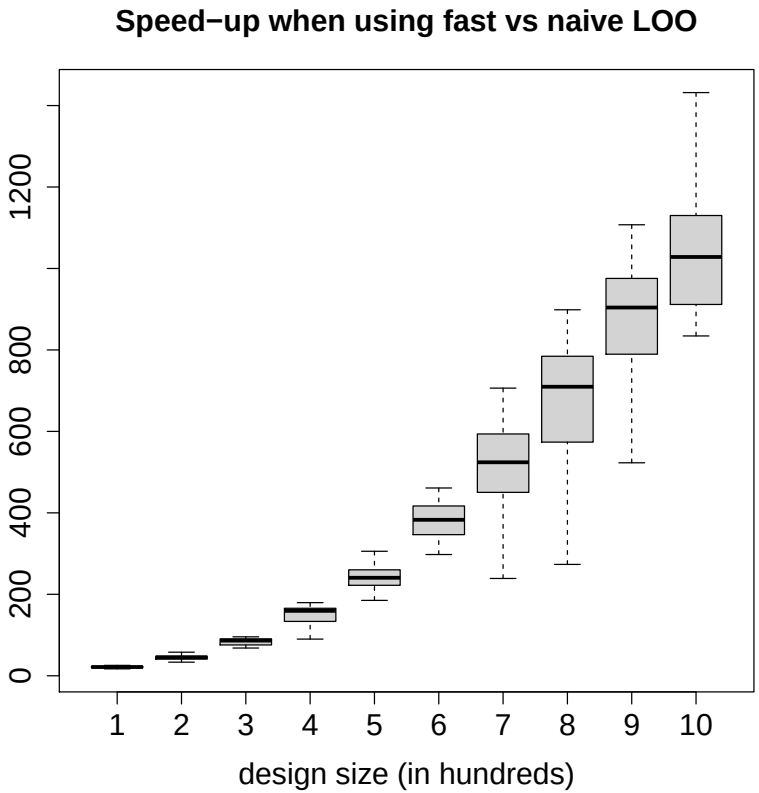
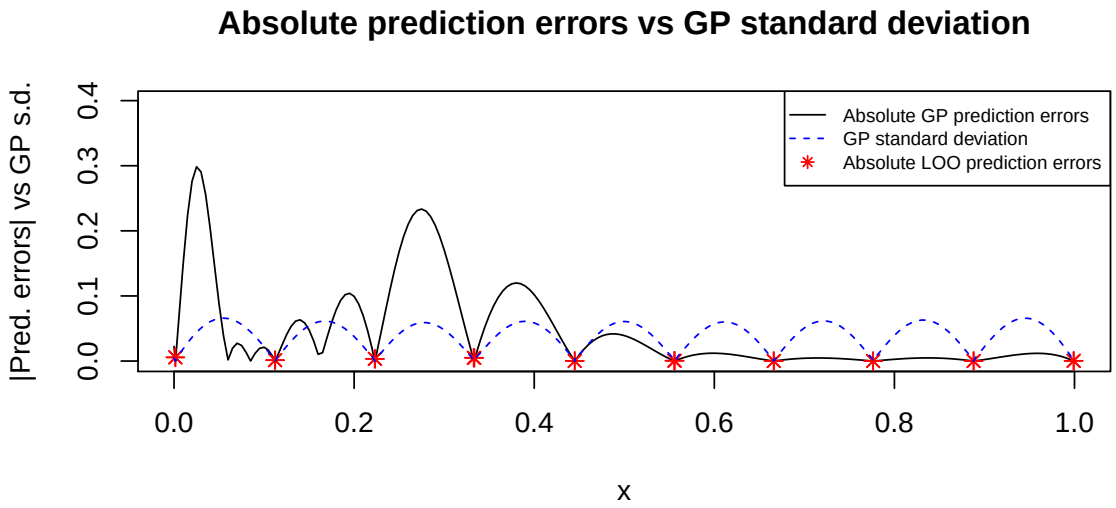
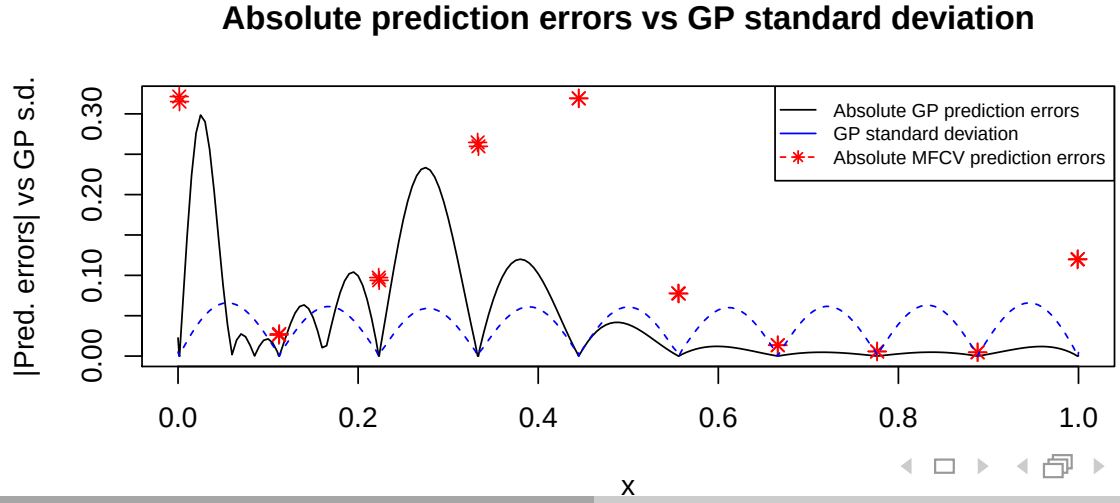


Figure: Speed-up (ratio between times required to run the naive and fast methods) measured for LOO on 10 regular designs, with 100 to 1000 points equidistributed on $[0, 1]$, where each speed-up measure is repeated 50 times.

LOO- vs MF-CV when the design has clustered points





A few challenges and outlined contributions

The previously presented efficient LOO formula have been generalized to CV with **arbitrary folds** (both in Simple and Universal Kriging frameworks), see

D. Ginsbourger and C. Schärer (2024).

Fast calculation of Gaussian Process multiple-fold cross-validation residuals and their covariances.
Journal of Computational and Graphical Statistics, 1-14.

In the next section we will review some of the main results of this paper.

A few challenges and outlined contributions

The previously presented efficient LOO formula have been generalized to CV with **arbitrary folds** (both in Simple and Universal Kriging frameworks), see

D. Ginsbourger and C. Schärer (2024).

Fast calculation of Gaussian Process multiple-fold cross-validation residuals and their covariances.
Journal of Computational and Graphical Statistics, 1-14.

In the next section we will review some of the main results of this paper.

Finally we will show how they can be applied on an inverse problem context from volcano geophysics, and leveraged to investigate the influence of fold design on parameter estimation by minimization of the norm of CV residuals.

Outline

Theorem

For any $\mathbf{i} \in \mathcal{S}$*,the Simple Kriging residual* $\mathbf{E_i} = \mathbf{Z}[\mathbf{i}] - \widehat{\mathbf{Z}^{(-\mathbf{i})}}[\mathbf{i}]$ *obtained when predicting at locations indexed by* $\mathbf{i}$ *based on observations at* $-\mathbf{i}$ *writes*

$$\mathbf{E_i} = (Q[\mathbf{i}])^{-1}(Q\mathbf{Z})[\mathbf{i}].$$

Consequently, for any $q > 1$ *and* $\mathbf{i}_1, \ldots, \mathbf{i}_q \in \mathcal{S}$*, the* $\mathbf{E}_{\mathbf{i}_j}$ $(1 \leq j \leq q)$ *are jointly Gaussian, centred, and with covariance structure given by*

$$\mathrm{Cov}(\mathbf{E_i}, \mathbf{E_j}) = (Q[\mathbf{i}])^{-1} Q[\mathbf{i}, \mathbf{j}](Q[\mathbf{j}])^{-1} \quad (\mathbf{i}, \mathbf{j} \in \mathcal{S}).$$

Theorem

For any $\mathbf{i} \in \mathcal{S}$*,the Simple Kriging residual* $\mathbf{E}_{\mathbf{i}} = \mathbf{Z}[\mathbf{i}] - \widehat{\mathbf{Z}^{(-\mathbf{i})}}[\mathbf{i}]$ *obtained when predicting at locations indexed by* $\mathbf{i}$ *based on observations at* $-\mathbf{i}$ *writes*

$$\mathbf{E}_{\mathbf{i}} = (Q[\mathbf{i}])^{-1}(Q\mathbf{Z})[\mathbf{i}].$$

Consequently, for any $q > 1$ *and* $\mathbf{i}_1, \ldots, \mathbf{i}_q \in \mathcal{S}$*, the* $\mathbf{E}_{\mathbf{i}_j}$ $(1 \leq j \leq q)$ *are jointly Gaussian, centred, and with covariance structure given by*

$$\mathrm{Cov}(\mathbf{E}_{\mathbf{i}}, \mathbf{E}_{\mathbf{j}}) = (Q[\mathbf{i}])^{-1} Q[\mathbf{i}, \mathbf{j}](Q[\mathbf{j}])^{-1} \quad (\mathbf{i}, \mathbf{j} \in \mathcal{S}).$$

In particular, for the case of an ensemble of folds $\mathfrak{I} = (\mathbf{i}_1, \ldots, \mathbf{i}_q)$ *such that concatenation of* $\mathbf{i}_1, \ldots, \mathbf{i}_q$ *gives* $(1, \ldots, n)$*, then*

$$\mathrm{Cov}(\mathbf{E}_{\mathfrak{I}}) = B_{\mathfrak{I}} Q B_{\mathfrak{I}},$$

where $B_{\mathfrak{I}} = \mathrm{blockdiag}\left((Q[\mathbf{i}_1])^{-1}, \ldots, (Q[\mathbf{i}_q])^{-1}\right)$ *.*

Block inversion in colours

Theorem (Block matrix inversion via Schur complement: a classic!)

Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ *be a real matrix with* A, B, C, D *of conformable dimensions.*

Assuming that D *and* $A - BD^{-1}C$ *are invertible, then so is* M *with*

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}.$$

Block inversion in colours

Theorem (Block matrix inversion via Schur complement: a classic!)

Let $M = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$ *be a real matrix with* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ *of conformable dimensions.*

Assuming that $\boldsymbol{D}$ *and* $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ *are invertible, then so is* M *with*

$$M^{-1} = \begin{pmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}.$$

NB: if $\mathbf{A}$ and $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ are invertible, we also have

$$M^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}.$$

Block inversion in colours

Theorem (Block matrix inversion via Schur complement: a classic!)

Let $M = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}$ *be a real matrix with* $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ *of conformable dimensions.*

Assuming that $\boldsymbol{D}$ *and* $\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ *are invertible, then so is* M *with*

$$M^{-1} = \begin{pmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}.$$

NB: if $\mathbf{A}$ and $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ are invertible, we also have

$$M^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}.$$

Identifying blocks give useful formulas!

Revisiting (Simple) Kriging in "transductive" settings

Before proving the theorem, let us revisit Simple Kriging when observation and prediction points are all from $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$.

WLOG, let $\mathbf{i}_o = (1, \ldots, m)$, $\mathbf{j}_o = (m+1, \ldots, n)$ where $m \leq n-1$.

Here we want to predict $\mathbf{Z}[\mathbf{i}_0] = (Z_1, \ldots, Z_m)^\top$ from $\mathbf{Z}[\mathbf{j}_0] = (Z_{m+1}, \ldots, Z_n)^\top$.

Revisiting (Simple) Kriging in "transductive" settings

Before proving the theorem, let us revisit Simple Kriging when observation and prediction points are all from $\mathbf{x}_1, \ldots, \mathbf{x}_n \in D$.

WLOG, let $\mathbf{i}_o = (1, \ldots, m)$, $\mathbf{j}_o = (m+1, \ldots, n)$ where $m \leq n-1$.

Here we want to predict $\mathbf{Z}[\mathbf{i}_0] = (Z_1, \ldots, Z_m)^\top$ from $\mathbf{Z}[\mathbf{j}_0] = (Z_{m+1}, \ldots, Z_n)^\top$.

The corresponding (Simple) Kriging predictor is known to be given by

$$\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0] = \mathrm{Cov}(\mathbf{Z}[\mathbf{i}_0], \mathbf{Z}[\mathbf{j}_0])\, \mathrm{Cov}(\mathbf{Z}[\mathbf{j}_0])^{-1} \mathbf{Z}[\mathbf{j}_0]$$

with associated residual covariance

$$\begin{aligned}\mathrm{Cov}(\mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]) \\ = \mathrm{Cov}(\mathbf{Z}[\mathbf{i}_0]) - \mathrm{Cov}(\mathbf{Z}[\mathbf{i}_0], \mathbf{Z}[\mathbf{j}_0])\, \mathrm{Cov}(\mathbf{Z}[\mathbf{j}_0])^{-1}\ \mathrm{Cov}(\mathbf{Z}[\mathbf{j}_0], \mathbf{Z}[\mathbf{i}_0])\end{aligned}$$

Revisiting (Simple) Kriging in transductive settings

Denoting $Q = \Sigma^{-1} = (K + \Sigma_\varepsilon)^{-1}$, we have by block matrix inversion

$$Q[\mathbf{i}_o] = (\Sigma[\mathbf{i}_o] - \Sigma[\mathbf{i}_o, \mathbf{j}_o]\Sigma[\mathbf{j}_o]^{-1}\Sigma[\mathbf{j}_o, \mathbf{i}_o])^{-1}$$

$$Q = \begin{pmatrix} \Sigma^{-1}[\mathbf{i}_o] & -\Sigma^{-1}[\mathbf{i}_o]\Sigma[\mathbf{i}_o, \mathbf{j}_o]\Sigma[\mathbf{j}_o]^{-1} \\ -\Sigma[\mathbf{j}_o]^{-1}\Sigma[\mathbf{j}_o, \mathbf{i}_o]\Sigma^{-1}[\mathbf{i}_o] & \text{not represented} \end{pmatrix},$$

whereof

$$\mathrm{Cov}\left(\mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]\right) = Q[\mathbf{i}_o]^{-1}$$

and

$$\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0] = -(Q[\mathbf{i}_o])^{-1} Q[\mathbf{i}_o, \mathbf{j}_o]\mathbf{Z}_{\mathbf{j}_o}$$

Revisiting (Simple) Kriging in transductive settings

Denoting $Q = \Sigma^{-1} = (K + \Sigma_\varepsilon)^{-1}$, we have by block matrix inversion

$$Q[\mathbf{i}_o] = (\Sigma[\mathbf{i}_o] - \Sigma[\mathbf{i}_o, \mathbf{j}_o]\Sigma[\mathbf{j}_o]^{-1}\Sigma[\mathbf{j}_o, \mathbf{i}_o])^{-1}$$

$$Q = \begin{pmatrix} \Sigma^{-1}[\mathbf{i}_o] & -\Sigma^{-1}[\mathbf{i}_o]\Sigma[\mathbf{i}_o, \mathbf{j}_o]\Sigma[\mathbf{j}_o]^{-1} \\ -\Sigma[\mathbf{j}_o]^{-1}\Sigma[\mathbf{j}_o, \mathbf{i}_o]\Sigma^{-1}[\mathbf{i}_o] & \text{not represented} \end{pmatrix},$$

whereof

$$\mathrm{Cov}\left(\mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]\right) = Q[\mathbf{i}_o]^{-1}$$

and

$$\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0] = -(Q[\mathbf{i}_o])^{-1} Q[\mathbf{i}_o, \mathbf{j}_o]\mathbf{Z}_{\mathbf{j}_o}$$

In other words, both quantities can be calculated based on blocks of Q.

Keys towards the proof (1/2)

The proof relies on block matrix inversion results.

We have indeed for arbitrary indices such as the inverses involved do exist (See. e.g., Horn and Johnson),

$$M^{-1}[\mathbf{i}] = (M[\mathbf{i}] - M[\mathbf{i}, -\mathbf{i}]M[-\mathbf{i}]^{-1}M[-\mathbf{i}, \mathbf{i}])^{-1}$$

and, more generally,

$$\begin{aligned} M^{-1}[\mathbf{i}, \mathbf{j}] &= -(M[\mathbf{i}] - M[\mathbf{i}, \mathbf{j}]M[-\mathbf{i}]^{-1}M[\mathbf{j}, \mathbf{i}])^{-1}M[\mathbf{i}, \mathbf{j}]M[\mathbf{j}]^{-1} \\ &= -M[\mathbf{j}]^{-1}M[\mathbf{j}, \mathbf{i}](M[\mathbf{i}] - M[\mathbf{i}, \mathbf{j}]M[\mathbf{j}]^{-1}M[\mathbf{j}, \mathbf{i}])^{-1}. \end{aligned}$$

Keys towards the proof (1/2)

The proof relies on block matrix inversion results.

We have indeed for arbitrary indices such as the inverses involved do exist (See. e.g., Horn and Johnson),

$$M^{-1}[\mathbf{i}] = (M[\mathbf{i}] - M[\mathbf{i}, -\mathbf{i}]M[-\mathbf{i}]^{-1}M[-\mathbf{i}, \mathbf{i}])^{-1}$$

and, more generally,

$$\begin{aligned} M^{-1}[\mathbf{i}, \mathbf{j}] &= -(M[\mathbf{i}] - M[\mathbf{i}, \mathbf{j}]M[-\mathbf{i}]^{-1}M[\mathbf{j}, \mathbf{i}])^{-1}M[\mathbf{i}, \mathbf{j}]M[\mathbf{j}]^{-1} \\ &= -M[\mathbf{j}]^{-1}M[\mathbf{j}, \mathbf{i}](M[\mathbf{i}] - M[\mathbf{i}, \mathbf{j}]M[\mathbf{j}]^{-1}M[\mathbf{j}, \mathbf{i}])^{-1}. \end{aligned}$$

From there one gets that $\mathbf{E_i} = (Q[\mathbf{i}])^{-1}(Q\mathbf{Z})[\mathbf{i}]$.

Keys towards the proof (2/2)

In order to highlight the joint Gaussianity and the covariance structure at once, let us further define

$$\Delta_{\mathbf{i}} = I_n[\mathbf{i}, (1, \ldots, n)]$$

to be the $\#\mathbf{i} \times n$ "subsetting" matrix.

Keys towards the proof (2/2)

In order to highlight the joint Gaussianity and the covariance structure at once, let us further define

$$\Delta_{\mathbf{i}} = I_n[\mathbf{i}, (1, \ldots, n)]$$

to be the $\#\mathbf{i} \times n$ "subsetting" matrix. We then have that for any $\mathbf{i} \in \mathcal{S}$,

$$\mathbf{E}_{\mathbf{i}} = (Q[\mathbf{i}])^{-1} \Delta_{\mathbf{i}} Q \mathbf{Z},$$

so that concatenating any finite number $q \geq 1$ of random vectors $\mathbf{E}_{\mathbf{i}_1}, \ldots, \mathbf{E}_{\mathbf{i}_q}$ leads to a Gaussian vector by left multiplication of $\mathbf{Z}$ by a deterministic matrix.

Keys towards the proof (2/2)

In order to highlight the joint Gaussianity and the covariance structure at once, let us further define

$$\Delta_{\mathbf{i}} = I_n[\mathbf{i}, (1, \ldots, n)]$$

to be the $\#\mathbf{i} \times n$ "subsetting" matrix. We then have that for any $\mathbf{i} \in \mathcal{S}$,

$$\mathbf{E}_{\mathbf{i}} = (Q[\mathbf{i}])^{-1} \Delta_{\mathbf{i}} Q \mathbf{Z},$$

so that concatenating any finite number $q \geq 1$ of random vectors $\mathbf{E}_{\mathbf{i}_1}, \ldots, \mathbf{E}_{\mathbf{i}_q}$ leads to a Gaussian vector by left multiplication of $\mathbf{Z}$ by a deterministic matrix.

The special case presented at the end of the theorem corresponds to a situation where the stacked $\Delta_{\mathbf{i}}$'s form the identity matrix (with size $n \times n$).

A remark following the theorem

For arbitary $\mathfrak{I}$ (without imposing ordering between $\mathbf{i}_j$'s or that they form a partition) we obtain a similar result yet without the above simplification, i.e.

$$\mathrm{Cov}(\mathbf{E}_{\mathfrak{I}}) = D_{\mathfrak{I}}\Delta_{\mathfrak{I}}Q\Delta_{\mathfrak{I}}^{T}B_{\mathfrak{I}} \text{ with } \Delta_{\mathfrak{I}} = (\Delta_{\mathbf{i}_1}^{\top}, \ldots, \Delta_{\mathbf{i}_q}^{\top})^{\top}.$$

A remark following the theorem

For arbitary $\mathfrak{I}$ (without imposing ordering between $\mathbf{i}_j$'s or that they form a partition) we obtain a similar result yet without the above simplification, i.e.

$$\text{Cov}(\mathbf{E}_{\mathfrak{I}}) = D_{\mathfrak{I}} \Delta_{\mathfrak{I}} Q \Delta_{\mathfrak{I}}^T B_{\mathfrak{I}} \text{ with } \Delta_{\mathfrak{I}} = (\Delta_{\mathbf{i}_1}^\top, \ldots, \Delta_{\mathbf{i}_q}^\top)^\top.$$

N.B.: an extreme case would be to consider all possible non-empty subsets of $\{1, \ldots, n\}$, leading to $q = 2^n - 1$ and $n2^{n-1}$ lines for $\Delta_{\mathfrak{I}}$.

About speed-ups

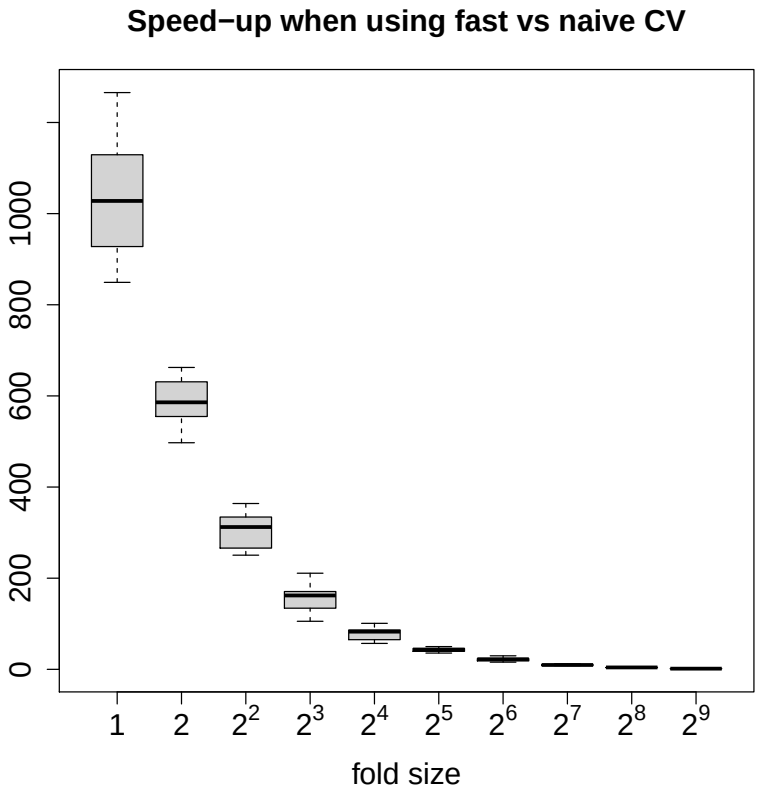
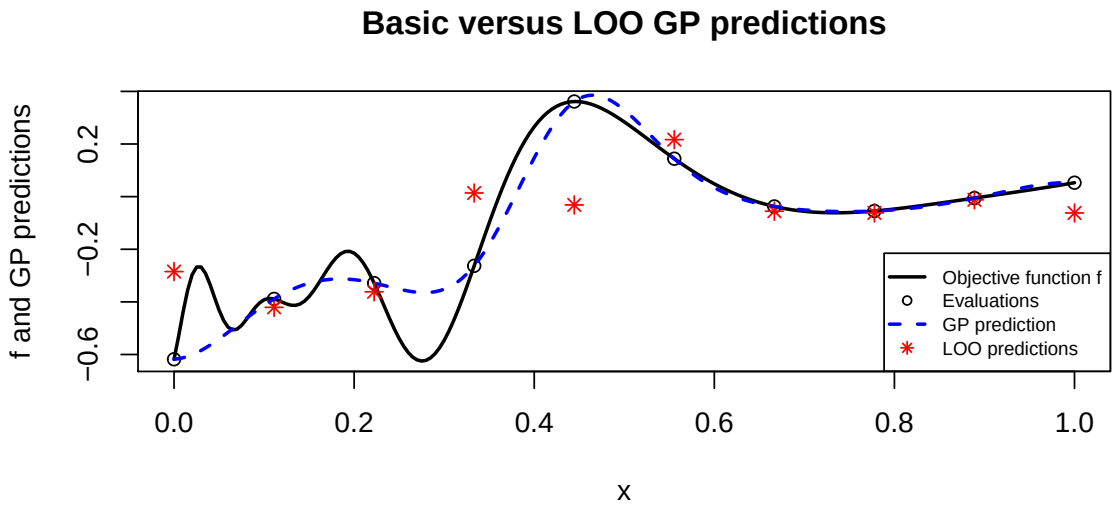
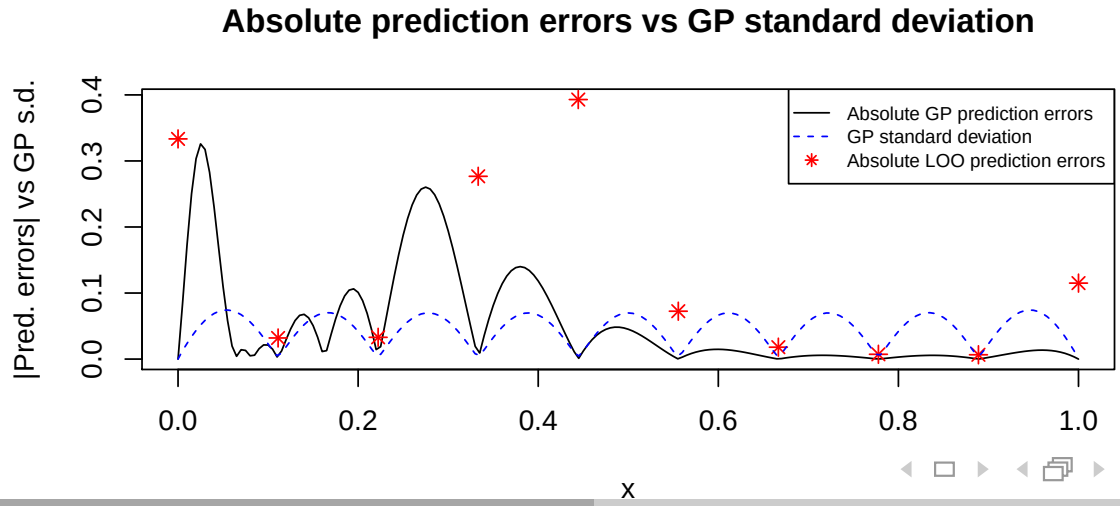


Figure: Speed-up (ratio between times required to run the naive and fast methods) measured for q-fold CV, where q decreases from 1024 to 2 and 50 seeds are used that affect here both model fitting and the folds.

Back to the first example





About the correlation between LOO residuals

We are here in the case where $q = n$ and the $\mathbf{i}_j$'s are set to (j) $(1 \leq j \leq n)$.

One recovers fast leave-one-out cross-validation formulae, and we obtain as a by-product the covariance matrix of leave-one-out residuals

$$\mathrm{diag}(\mathrm{Q}[1]^{-1}, \ldots, \mathrm{Q}[n]^{-1})\ \mathrm{Q}\ \mathrm{diag}(\mathrm{Q}[1]^{-1}, \ldots, \mathrm{Q}[n]^{-1})$$

About the correlation between LOO residuals

We are here in the case where $q = n$ and the $\mathbf{i}_j$'s are set to (j) $(1 \leq j \leq n)$.

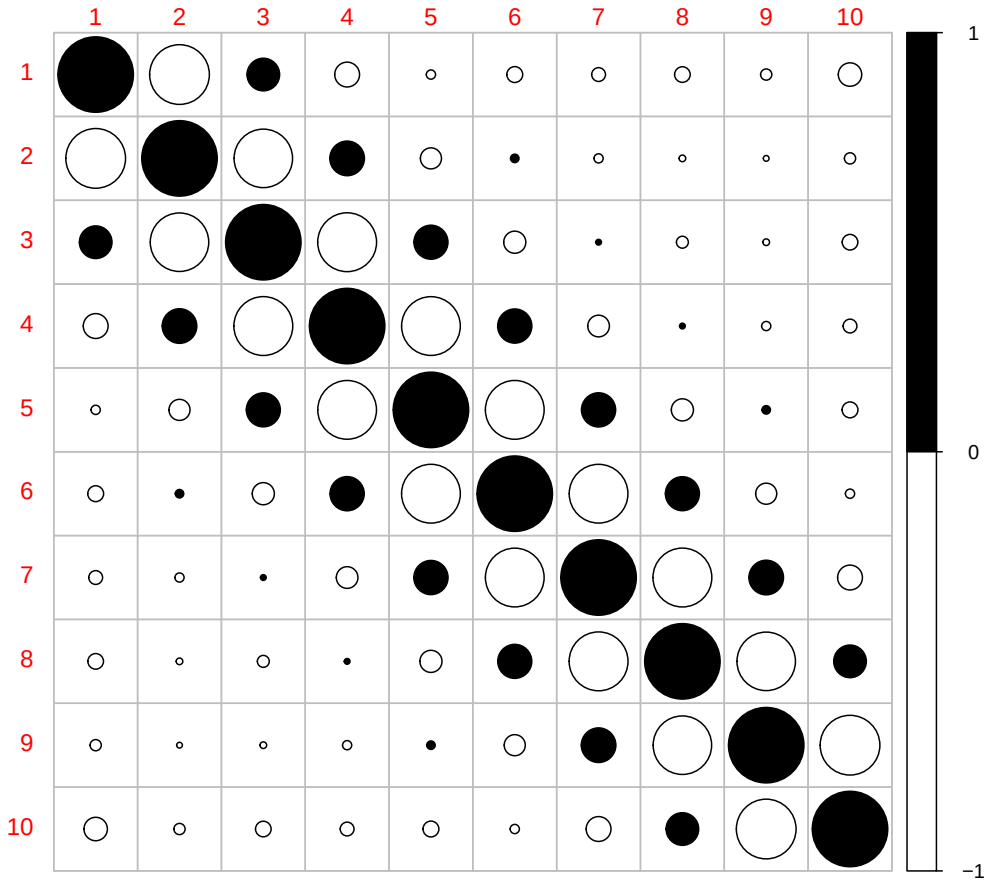
One recovers fast leave-one-out cross-validation formulae, and we obtain as a by-product the covariance matrix of leave-one-out residuals

$$\mathrm{diag}(\mathrm{Q}[1]^{-1}, \ldots, \mathrm{Q}[\mathrm{n}]^{-1})\ \mathrm{Q}\ \mathrm{diag}(\mathrm{Q}[1]^{-1}, \ldots, \mathrm{Q}[\mathrm{n}]^{-1})$$

leading to the following formula for the correlation matrix of LOO residuals

$$\mathrm{diag}(\mathrm{Q}[1]^{-1/2}, \ldots, \mathrm{Q}[\mathrm{n}]^{-1/2})\ \mathrm{Q}\ \mathrm{diag}(\mathrm{Q}[1]^{-1/2}, \ldots, \mathrm{Q}[\mathrm{n}]^{-1/2})$$

Correlation of LOO-CV residuals on the first example



Some consequence of CV residuals being correlated

It is not appropriate to consider “standardized” LOO (or further CV) residuals separately when building diagnostics such as QQ-plots

$\Rightarrow$ A decorrelating operation seems in order!

Some consequence of CV residuals being correlated

It is not appropriate to consider "standardized" LOO (or further CV) residuals separately when building diagnostics such as QQ-plots

$$\Rightarrow \text{A decorrelating operation seems in order!}$$

Assuming multiple-fold settings from the second part of the main theorem, any matrix $A \in \mathbb{R}^{n \times n}$ such that $AB_{\mathfrak{J}}QB_{\mathfrak{J}}A^{\top} = I_n$ does the job.

Some consequence of CV residuals being correlated

It is not appropriate to consider "standardized" LOO (or further CV) residuals separately when building diagnostics such as QQ-plots

$$\Rightarrow \text{A decorrelating operation seems in order!}$$

Assuming multiple-fold settings from the second part of the main theorem, any matrix $A \in \mathbb{R}^{n \times n}$ such that $AB_{\mathfrak{J}}QB_{\mathfrak{J}}A^{\top} = I_n$ does the job.

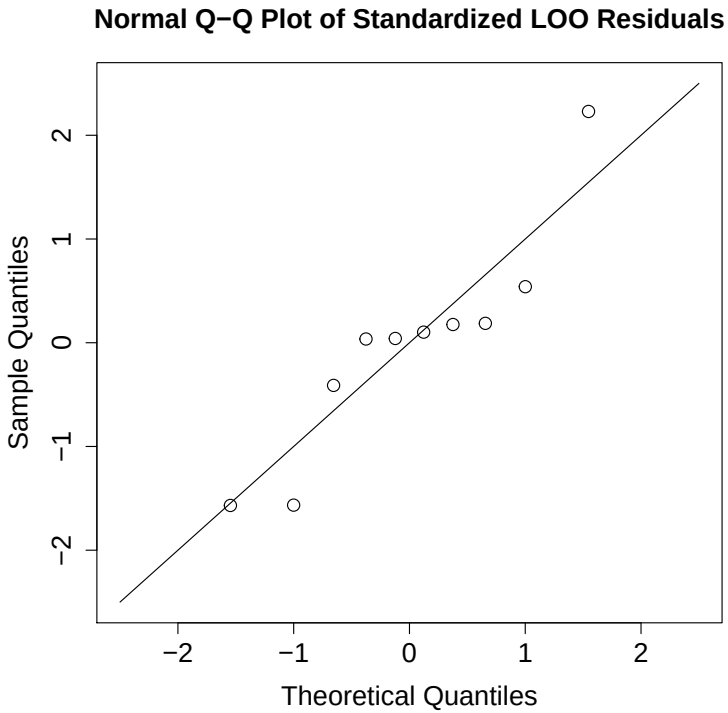
More specifically, with $A_{\mathfrak{J}} = \Sigma^{1/2} D_{\mathfrak{J}}^{-1}$, one gets indeed

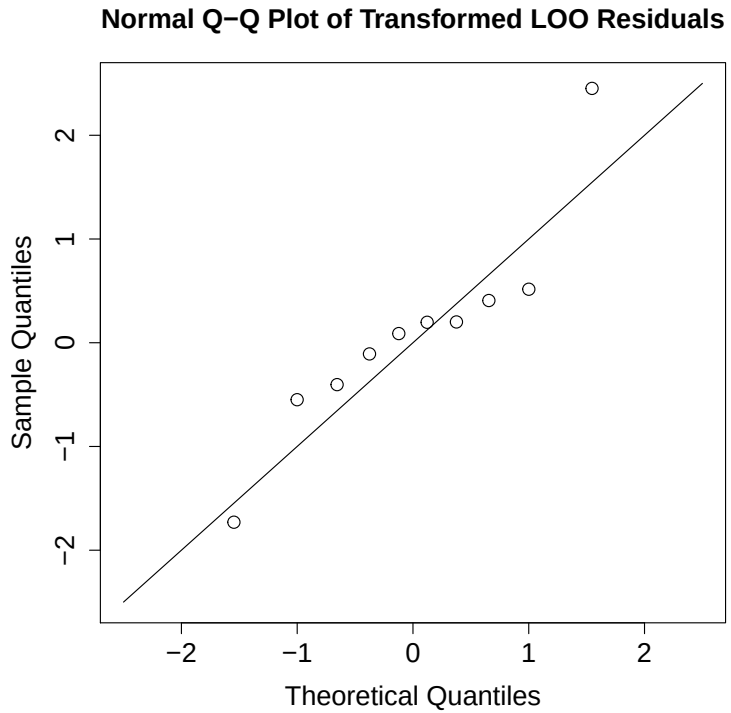
$$A_{\mathfrak{J}}\mathbf{E}_{\mathfrak{J}} = \Sigma^{-1/2}\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_n).$$

Hence the hypothesis of a correct model can be questioned using standard means relying on such a pivotal multivariate Gaussian distributed quantity.

Standardized vs transformed CV residuals (example)

Back to our first example, we obtain the following comparison between merely "standardized" against properly "transformed" LOO residuals:





About the estimation of σ^2 (1/2)

The leave-one-out-based estimator of σ^2 investigated in Bachoc 2013 reads

$$\widehat{\sigma}^2_{\mathrm{LOO}} = \frac{1}{n}\mathbf{Z}R^{-1}(\mathrm{diag}(R^{-1}))^{-1}R^{-1}\mathbf{Z},$$

and originates from the idea (traced back by Bachoc to Cressie 1993) that

$$C^{(1)}_{\mathrm{LOO}}(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{E}_i)^2}{\mathrm{Var}(\mathbf{E}_i)},$$

should take a value close to one.

About the estimation of σ^2 (2/2)

Yet, in order to correct for the covariance between LOO residuals, one may revise the criterion $C^{(1)}_{\text{LOO}}(\sigma^2)$ into

$$\begin{aligned} C^{(1)}_{\widetilde{\text{LOO}}}(\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{Z}[i] - \widehat{\mathbf{Z}^{(-i)}}[i])(BQB)^{ij}(\mathbf{Z}[j] - \widehat{\mathbf{Z}^{(-j)}}[j]) \\ &= \frac{1}{n\sigma^2}\mathbf{E}^\top \operatorname{diag}(R^{-1})R\operatorname{diag}(R^{-1})\mathbf{E} \\ &= \frac{1}{n\sigma^2}\mathbf{Z}^\top R^{-1}\mathbf{Z}, \end{aligned}$$

so that setting this modified criterion to 1 would plainly result in

$$\widehat{\sigma}^2_{\widetilde{\text{LOO}}} = \frac{1}{n}\mathbf{Z}R^{-1}\mathbf{Z} = \widehat{\sigma}^2_{\text{MLE}}.$$

Outline

On MF-CV-estimation of further kernel parameters

We now focus on by-products of fast multiple-fold CV for the estimation of θ and tackle in particular the following research questions/challenges:

- Closed-form formula for the ℓ^2 norm^2 of CV errors in function of R_θ
- Application to a Bayesian inverse problem from volcano geophysics
- Numerical study of resulting θ estimators depending on fold design

Fast square norm of MFCV residuals in closed form

We now consider the fast/closed-form calculation of a multiple-fold CV criterion for θ estimation, namely

$$C_{\mathrm{CV}}(\theta;\mathfrak{I}) = \sum_{j=1}^{q} ||\mathbf{Z}[\mathbf{i}_j] - \widehat{\mathbf{Z}^{(-\mathbf{i}_j)}[\mathbf{i}_j]}(\theta)||^2 = \sum_{j=1}^{q} ||\mathbf{E}_{\mathbf{i}_j}(\theta)||^2 = ||\mathbf{E}_{\mathfrak{I}}(\theta)||^2.$$

Fast square norm of MFCV residuals in closed form

We now consider the fast/closed-form calculation of a multiple-fold CV criterion for θ estimation, namely

$$C_{\mathrm{CV}}(\theta;\mathfrak{I}) = \sum_{j=1}^{q} ||\mathbf{Z}[\mathbf{i}_j] - \widehat{\mathbf{Z}^{(-\mathbf{i}_j)}[\mathbf{i}_j]}(\theta)||^2 = \sum_{j=1}^{q} ||\mathbf{E}_{\mathbf{i}_j}(\theta)||^2 = ||\mathbf{E}_{\mathfrak{I}}(\theta)||^2.$$

Building up upon the main theorem, we obtain (case $\Sigma = \sigma^2 R$) that

$$C_{\mathrm{CV}}(\theta;\mathfrak{I}) = \mathbf{Z}^\top R_\theta^{-1} \,\mathrm{blockdiag}\left((R_\theta^{-1}[\mathbf{i}_1])^{-2}, \ldots, (R_\theta^{-1}[\mathbf{i}_q])^{-2}\right) R_\theta^{-1}\mathbf{Z}.$$

Fast square norm of MFCV residuals in closed form

We now consider the fast/closed-form calculation of a multiple-fold CV criterion for θ estimation, namely

$$C_{\mathrm{CV}}(\theta;\mathfrak{I}) = \sum_{j=1}^{q} ||\mathbf{Z}[\mathbf{i}_j] - \widehat{\mathbf{Z}^{(-\mathbf{i}_j)}[\mathbf{i}_j]}(\theta)||^2 = \sum_{j=1}^{q} ||\mathbf{E}_{\mathbf{i}_j}(\theta)||^2 = ||\mathbf{E}_{\mathfrak{I}}(\theta)||^2.$$

Building up upon the main theorem, we obtain (case $\Sigma = \sigma^2 R$) that

$$C_{\mathrm{CV}}(\theta;\mathfrak{I}) = \mathbf{Z}^\top R_\theta^{-1} \,\mathrm{blockdiag}\left((R_\theta^{-1}[\mathbf{i}_1])^{-2}, \ldots, (R_\theta^{-1}[\mathbf{i}_q])^{-2}\right) R_\theta^{-1}\mathbf{Z}.$$

Note that this criterion derives from a rather basic scoring approach. More general approaches have been considered (not covered here).

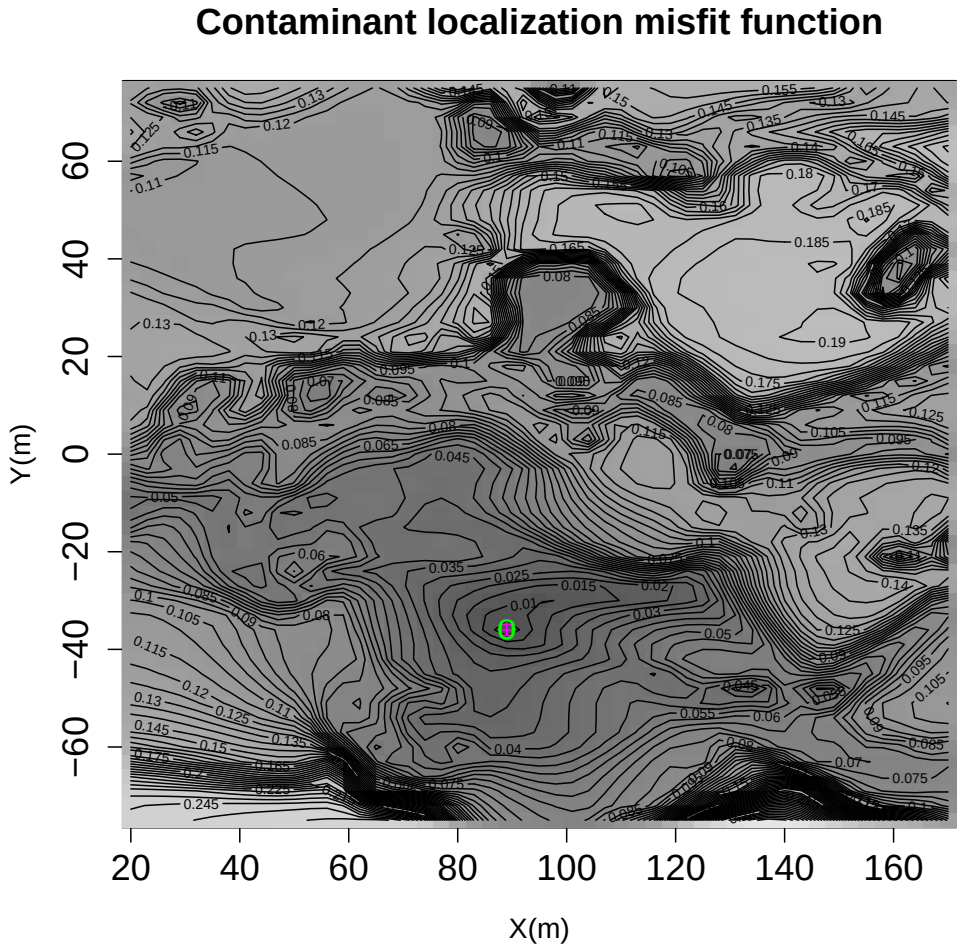


Figure: Contaminant localization test function designed by summing misfits between given concentrations at monitoring wells and corresponding simulation results when varying the candidate source localization.

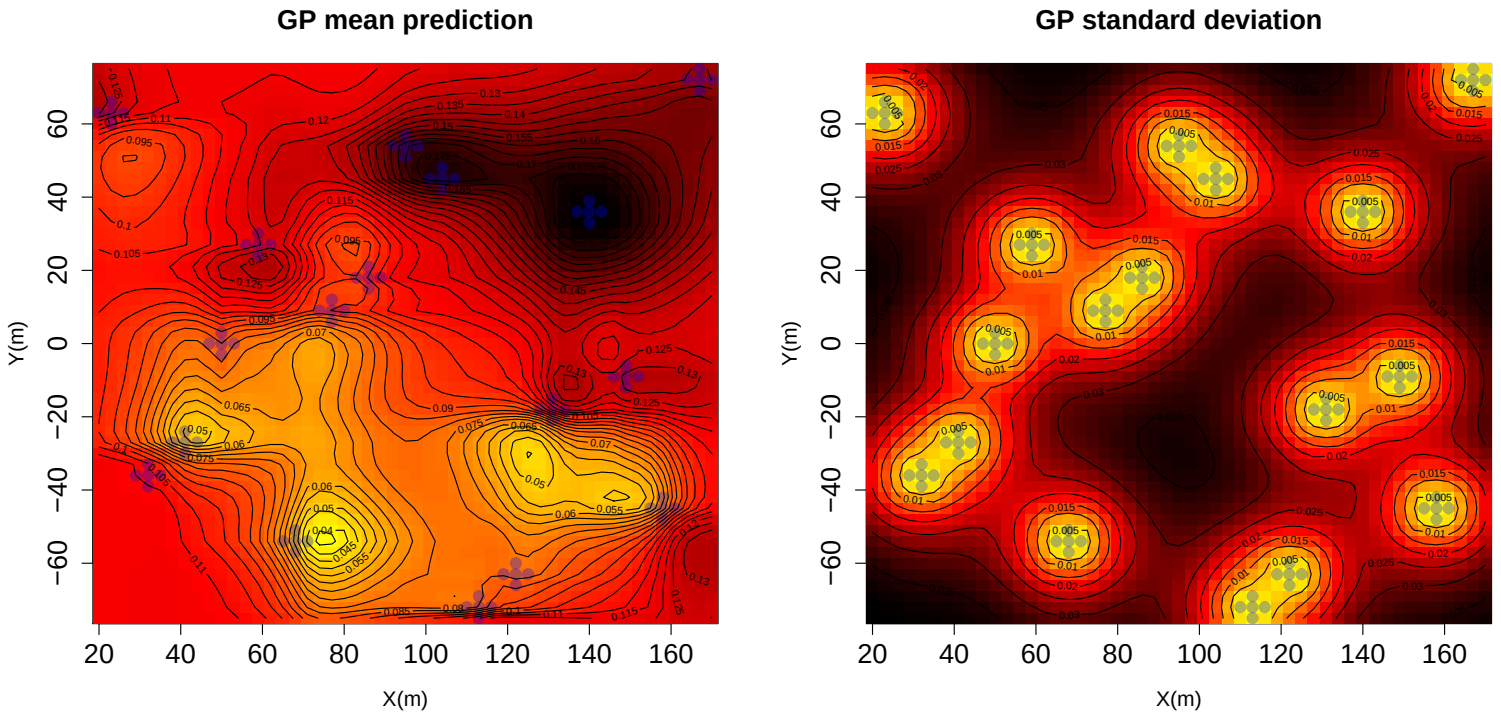


Figure: Gaussian Process prediction mean and standard deviation on the contaminant localization test function with 25 clover-shape 5-element observation clusters and covariance parameters estimated by MLE.

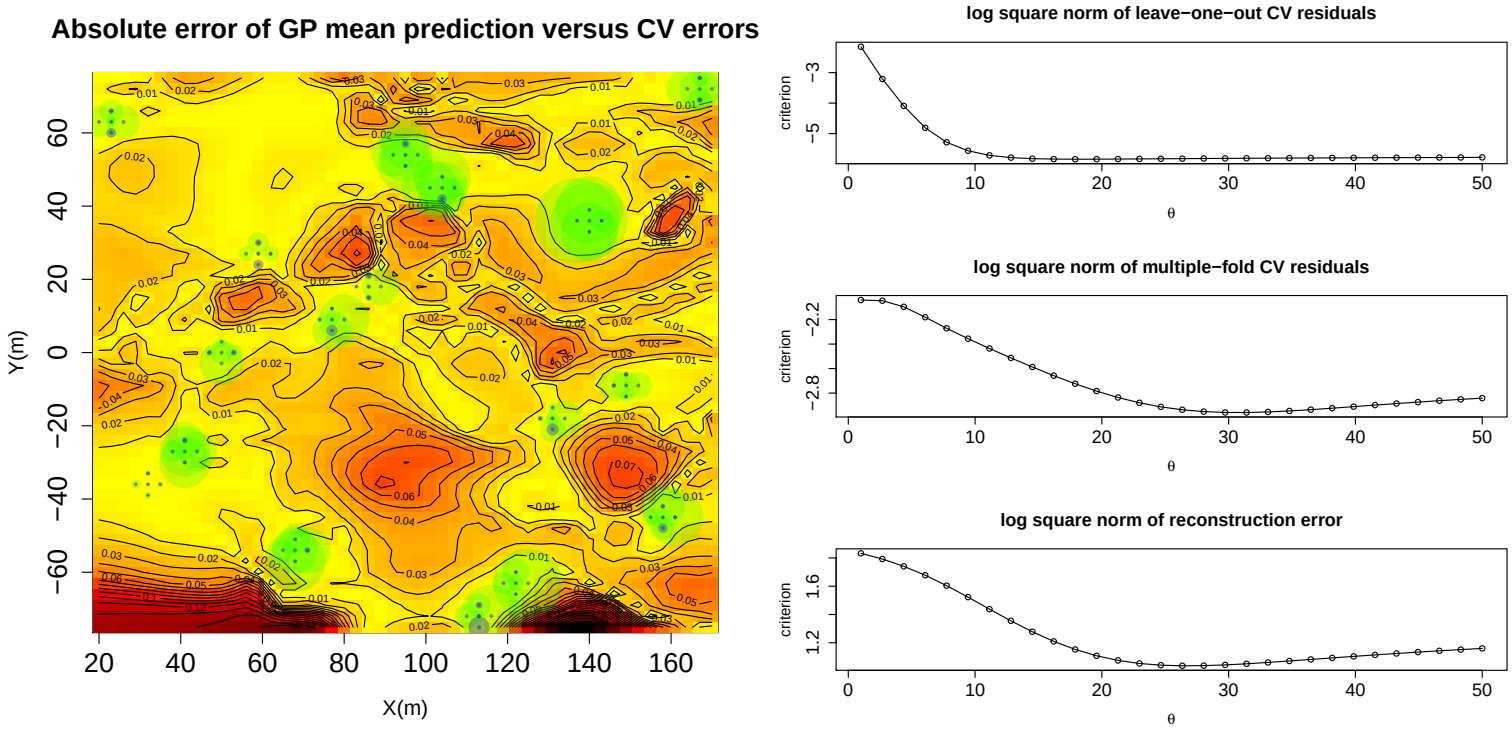


Figure: Left: absolute prediction errors (heatmap) versus CV residuals (disks of radii proportional to absolute residuals, blue for LOO and green for MFCV). Right: log square norm of LOO residuals (top), of MFCV residuals (center), and of reconstruction error (bottom) as a function of the range hyperparameter.

Cross-validating gravimetry on the Stromboli

We will now present an application test case where multiple-fold CV is used for the estimation of a correlation parameter θ of the input field on a Bayesian inverse problem where observations are gravimetry measurements.

Joint work with Athénaïs Gautier and Cédric Travelletti.

Cross-validating gravimetry on the Stromboli

We will now present an application test case where multiple-fold CV is used for the estimation of a correlation parameter θ of the input field on a Bayesian inverse problem where observations are gravimetry measurements.

Joint work with Athénaïs Gautier and Cédric Travelletti.

For more detail about the underlying inverse problem and the GP model, see

C. Travelletti, D. Ginsbourger, and N. Linde (2023)
Uncertainty Quantification and Experimental Design for Large-Scale Linear Inverse Problems under Gaussian Process Priors
SIAM/ASA Journal on Uncertainty Quantification 11(1)

Gravimetric inversion on Stromboli: first simulation

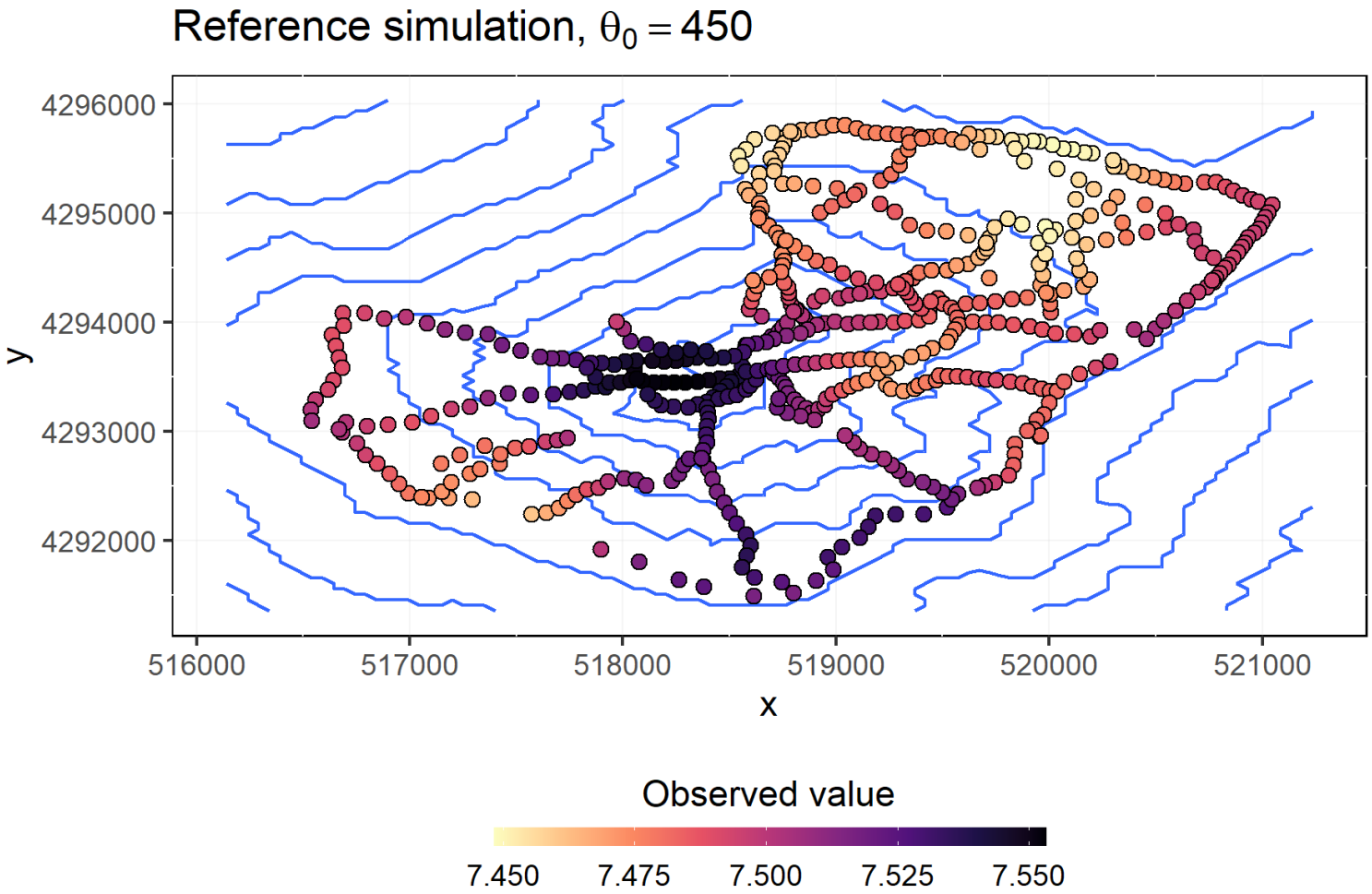


Figure: Simulated gravimetry measurements (generated with $\theta_o = 450$)

CV on the first simulation example: clustered folds

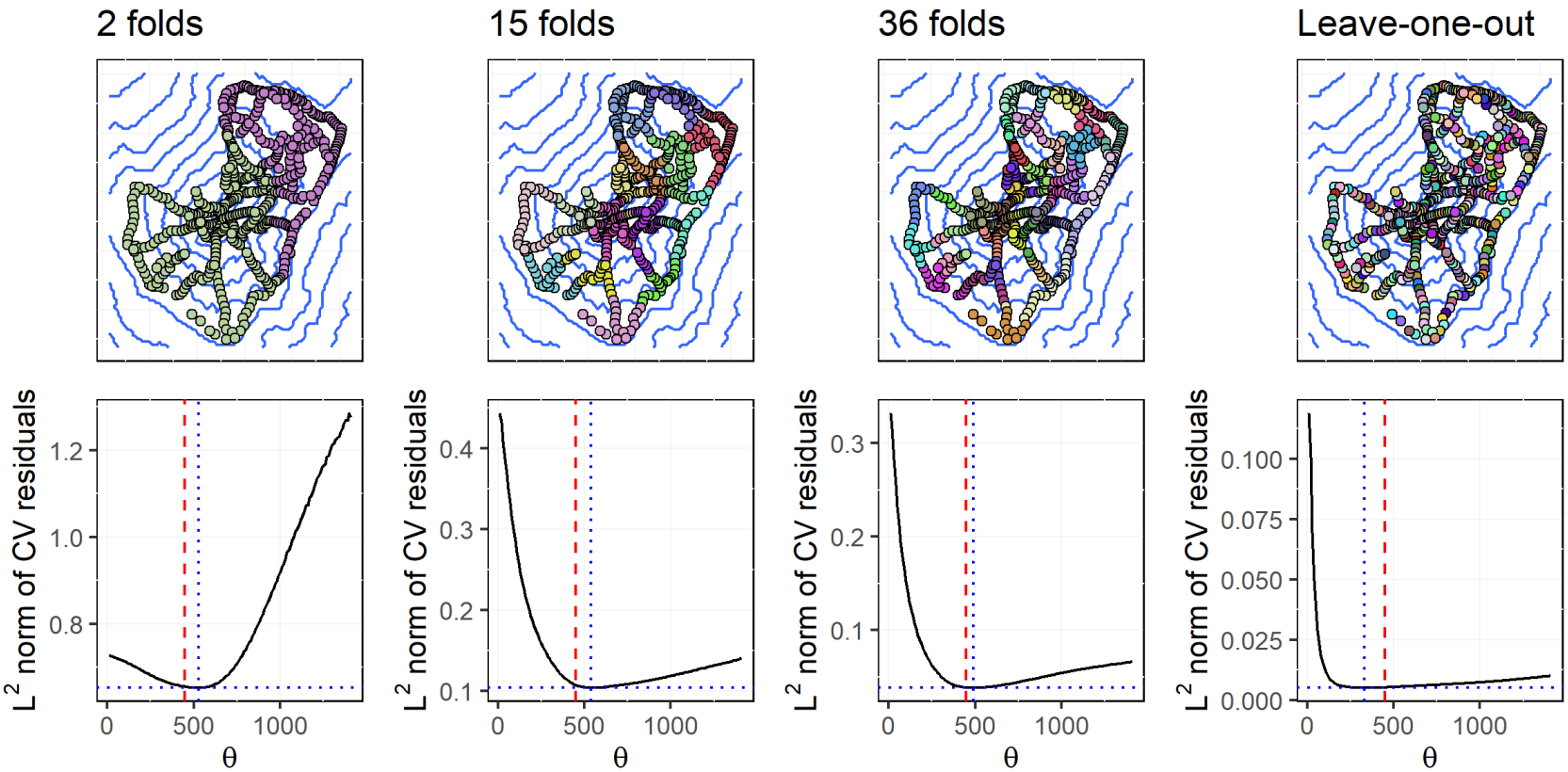


Figure: L^2 norm of CV residuals for various fold designs resulting from clustering.

CV on the first simulation example: random folds

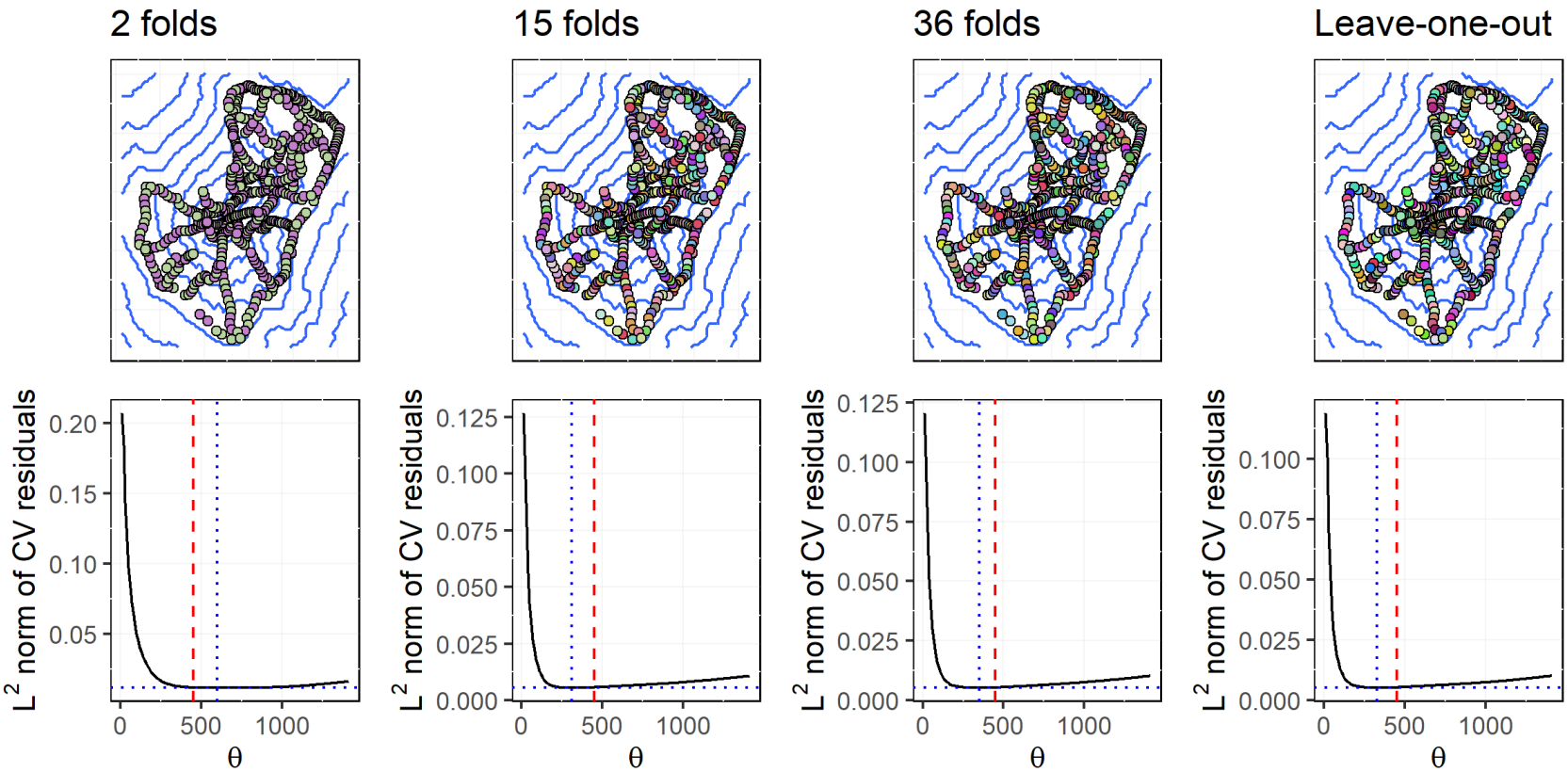


Figure: L^2 norm of CV residuals for various fold designs resulting from randomization.

Simulation study results (clustered folds)

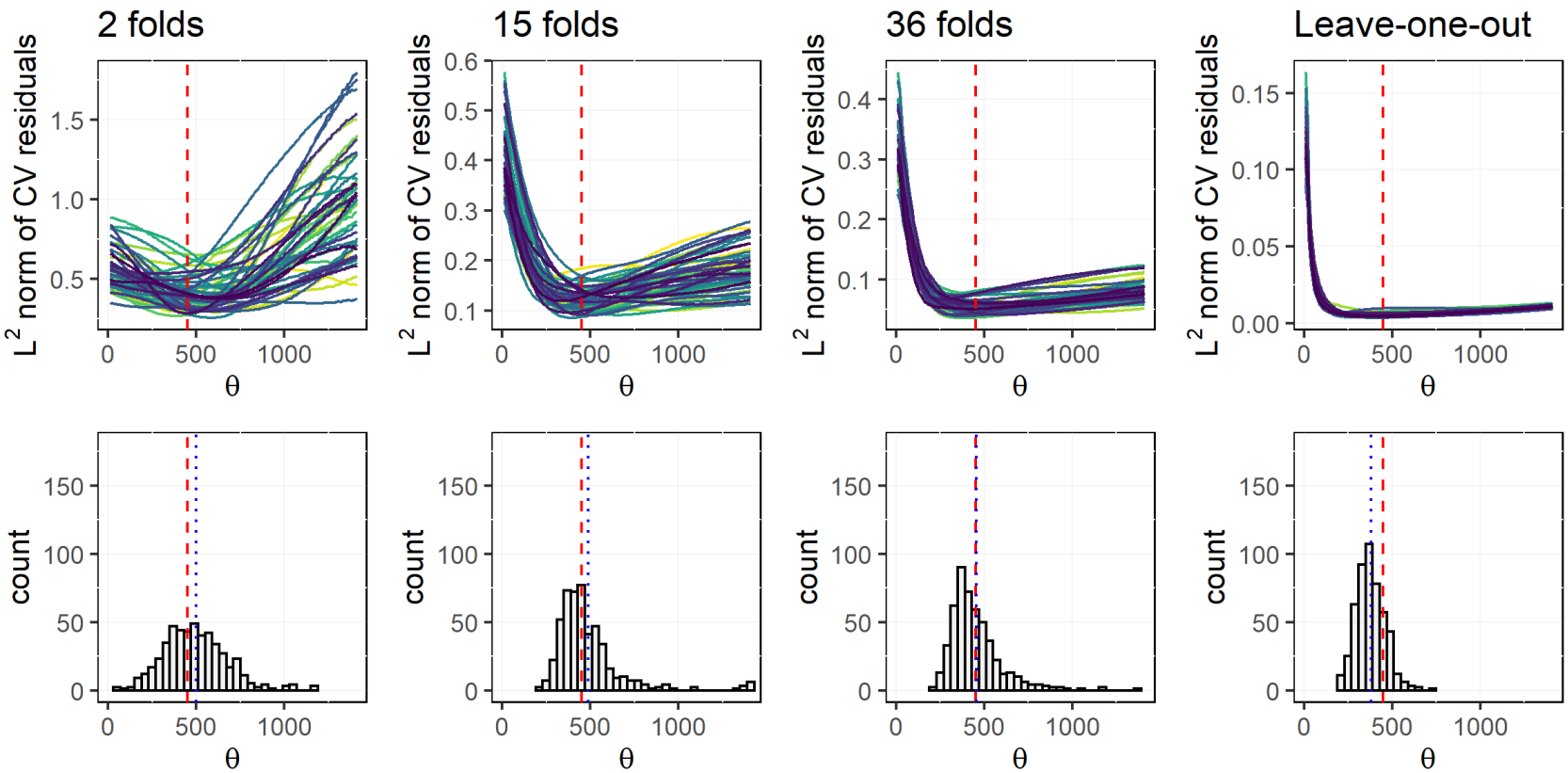


Figure: L^2 norm of residuals for 500 simulations (50 curves displayed), for various fold designs resulting from clustering.

Simulation study results (randomized folds)

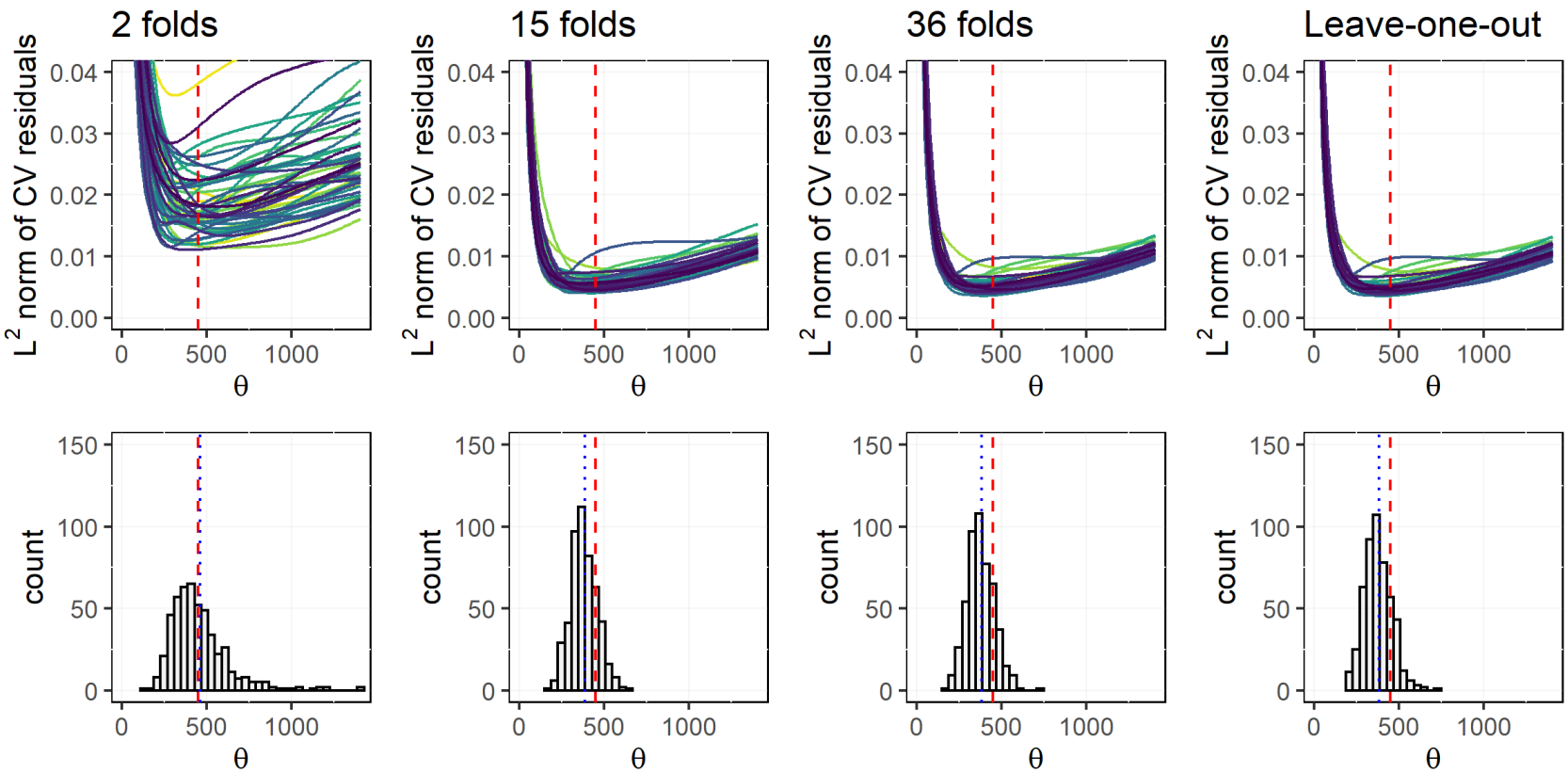


Figure: L^2 norm of residuals for 500 simulations (50 curves displayed), for various fold designs resulting from randomization.

Bias

Folds	$\theta_0 = 150$		$\theta_0 = 450$		$\theta_0 = 750$	
	Clusters	Random	Clusters	Random	Clusters	Random
2	4.28	7.4	49.18	11.72	55.64	-91.8
4	2.66	9.5	55.16	-25.92	35.04	-174.24
5	1.52	13	42.8	-44	23.64	-182.68
8	3.16	6.56	46.02	-57.3	36.1	-204.42
15	3.96	9.4	37.36	-61.9	7.2	-213.48
25	4.6	10.54	46.98	-65.82	-7.18	-213.9
36	5.42	9.2	4.2	-65.96	-61.08	-219.54
54	5.44	10.1	5.02	-62.82	-94.74	-218.66
60	4.54	11.48	-26.36	-64.4	-152.26	-223.1
90	5	8.32	-28.58	-64.94	-166.64	-222.12
108	5.26	10.68	-41.14	-66.1	-180.1	-221.52
180	4.34	9.36	-53.46	-66.26	-198.22	-222.78
271	4.96	10.28	-76.7	-67.18	-245.44	-222.3
LOO	**9.98**		**-66.9**		**-222.9**	
MLE	5.36		23.92		41.56	

Estimation standard deviation

Folds	$\theta_0 = 150$		$\theta_0 = 450$		$\theta_0 = 750$	
	Clusters	Random	Clusters	Random	Clusters	Random
2	59.59	38.6	178.1	173.82	290.13	202.37
4	39.05	42.2	184.26	122.75	243.53	124.41
5	35.7	46.61	172.35	98	232.24	113.03
8	34.43	39.25	191.34	86.28	245.44	104.24
15	27.1	35.92	185.26	76.64	250.37	96.91
25	25.24	39.45	217.25	80.84	259.84	100.6
36	23.19	37.16	143.57	78.02	214.31	97.3
54	21.8	38.76	152.64	85.42	205.78	97.6
60	21.45	45.26	93.59	82.8	120.9	97
90	21.53	36.86	106.32	81.42	118.77	97.24
108	23.01	40.74	79.38	79.46	92.1	97.54
180	22.18	36.84	66.26	79.48	86.71	96.35
271	26.25	39.58	56.96	79.32	69.44	96.39
LOO	**38.84**		**79.89**		**97.1**	
MLE	5.49		11.25		19.25	

Root mean square error

Folds	$\theta_0 = 150$		$\theta_0 = 450$		$\theta_0 = 750$	
	Clusters	Random	Clusters	Random	Clusters	Random
2	59.74	39.3	184.76	174.21	295.42	222.22
4	39.14	43.25	192.34	125.46	246.03	214.1
5	35.73	48.39	177.58	107.42	233.44	214.82
8	34.57	39.79	196.8	103.58	248.08	229.46
15	27.38	37.13	188.99	98.51	250.47	234.45
25	25.65	40.84	222.27	104.25	259.94	236.38
36	23.82	38.28	143.63	102.17	222.84	240.13
54	22.47	40.06	152.72	106.04	226.54	239.45
60	21.93	46.69	97.23	104.9	194.42	243.28
90	22.1	37.78	110.1	104.15	204.63	242.47
108	23.61	42.12	89.41	103.36	202.28	242.04
180	22.6	38.01	85.13	103.48	216.36	242.72
271	26.72	40.89	95.54	103.95	255.07	242.3
LOO	**40.1**		**104.2**		**243.13**	
MLE	7.67		26.43		45.8	

Ongoing work and selected perspectives

- CV-based selection of trends in Bayesian Inversion (Stromboli!)
  → See Cédric Travelletti's PhD thesis (2023)
- Investigating further scoring rules in the context of MFCV
- Formalizing and developing "fold design" (criteria, algorithms, etc.)
- Exploring generalizations beyond the considered classes of models.

Ongoing work and selected perspectives

- CV-based selection of trends in Bayesian Inversion (Stromboli!)
  → See Cédric Travelletti's PhD thesis (2023)
- Investigating further scoring rules in the context of MFCV
- Formalizing and developing "fold design" (criteria, algorithms, etc.)
- Exploring generalizations beyond the considered classes of models.

Thank you very much for your attention!

A few more references

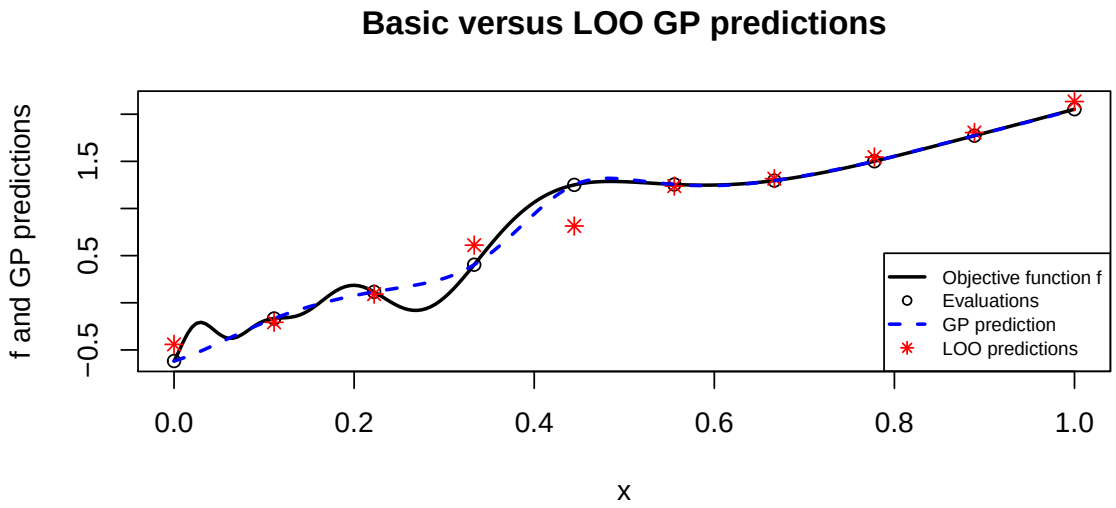
P. Zhang (1993).
Model Selection Via Multifold Cross Validation
Annals of Statistics, 21(1): 299-313.

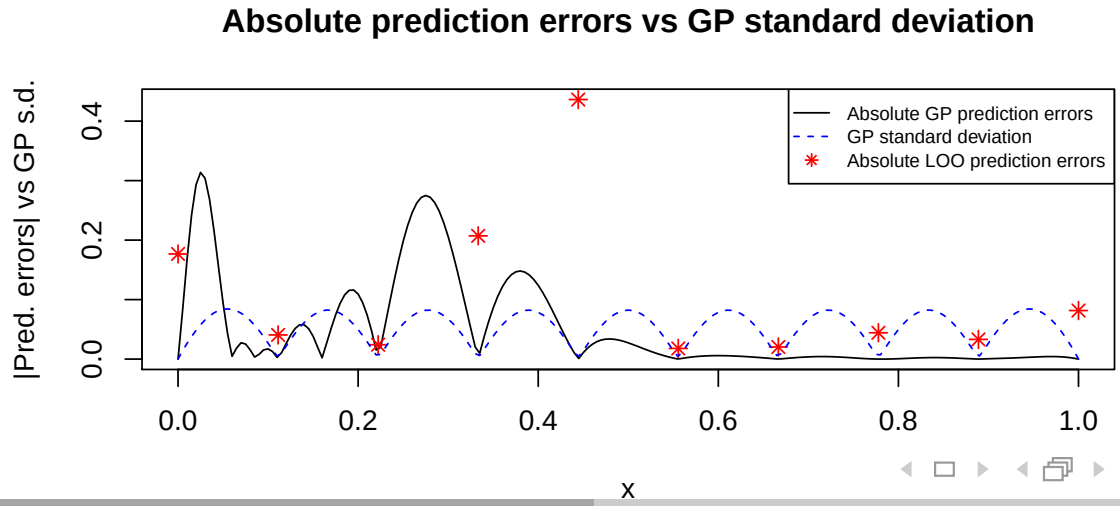
S. An, W. Liu, S. Venkatesh (2007).
Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression.
Pattern Recognition, 40(8), 2154-2162.

A. Rabinowicz and S. Rosset (2020).
Cross-validation for correlated data.
Journal of the American Statistical Association, 117:538, 718-731

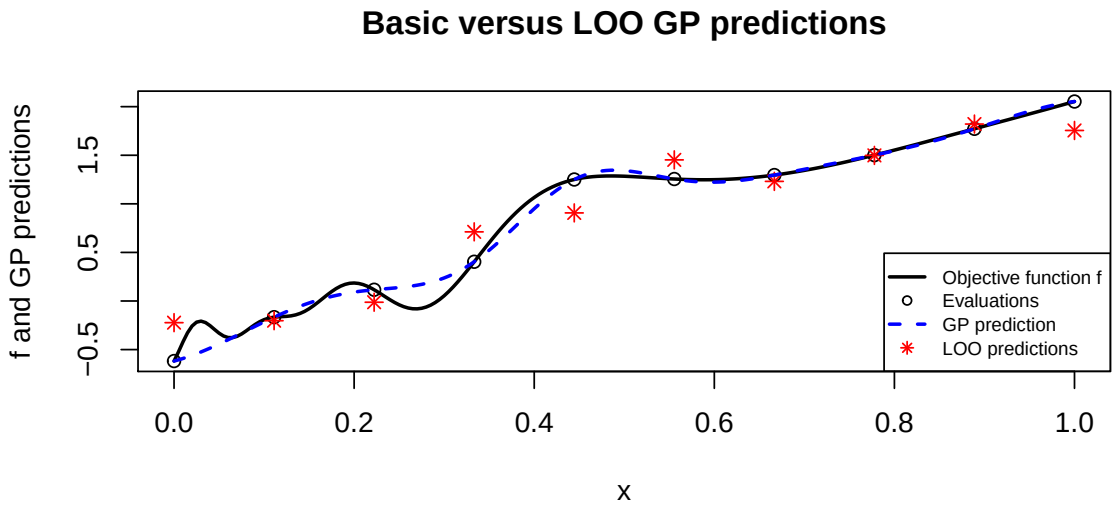
S. Bates and T. Hastie and R. Tibshirani (2023).
Cross-validation: what does it estimate and how well does it do it?
Journal of the American Statistical Association.

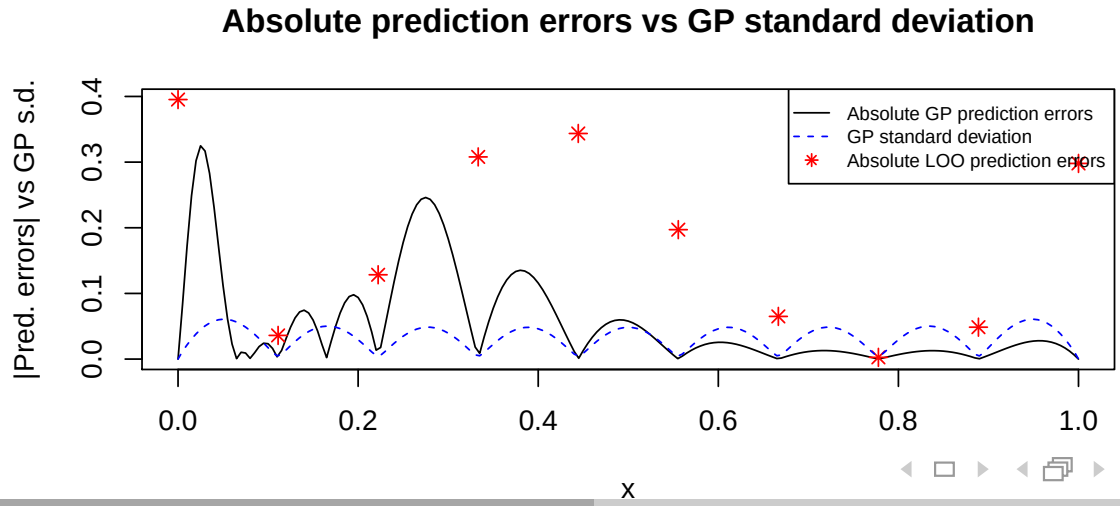
Back to first example, with a trend (UK)





Back to first example, with a trend (OK)





Revisiting (Universal) Kriging in transductive settings

In Universal Kriging, similar results do apply, yet by augmenting Σ with the design matrix F in the sense of the following matrix M:

$$M = \begin{pmatrix} \Sigma & F \\ F^\top & 0 \end{pmatrix},$$

Denoting now by $\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]$ the Universal Kriging predictor of $\mathbf{Z}[\mathbf{i}_o]$ based on $\mathbf{Z}[\mathbf{j}_o]$, one can then show that

$$\mathrm{Cov}\left(\mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]\right) = M^{-1}[\mathbf{i}_o]^{-1}$$

and

$$\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0] = -(M^{-1}[\mathbf{i}_o])^{-1} M^{-1}[\mathbf{i}_o, \mathbf{j}_o]\mathbf{Z}_{\mathbf{j}_o}$$

Revisiting (Universal) Kriging in transductive settings

In Universal Kriging, similar results do apply, yet by augmenting Σ with the design matrix F in the sense of the following matrix M:

$$M = \begin{pmatrix} \Sigma & F \\ F^\top & 0 \end{pmatrix},$$

Denoting now by $\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]$ the Universal Kriging predictor of $\mathbf{Z}[\mathbf{i}_o]$ based on $\mathbf{Z}[\mathbf{j}_o]$, one can then show that

$$\mathrm{Cov}\left(\mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]\right) = M^{-1}[\mathbf{i}_o]^{-1}$$

and

$$\widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0] = -(M^{-1}[\mathbf{i}_o])^{-1} M^{-1}[\mathbf{i}_o, \mathbf{j}_o]\mathbf{Z}_{\mathbf{j}_o}$$

This time, both quantities can be calculated based on blocks of M^{-1}. Note that $M^{-1}[\mathbf{i}_o]^{-1} = \Sigma^{-1} - \Sigma^{-1}F(F\Sigma^{-1}F^\top)^{-1}F^\top\Sigma^{-1}$.

Cross-validation residuals from precision matrices

Denote $\mathbf{E}_{\mathbf{i}_o} = \mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]$ (either way).

Simple Kriging case ($Q = \Sigma^{-1}$)

$$\mathbf{E}_{\mathbf{i}_o} = (Q[\mathbf{i}_o])^{-1}(Q\mathbf{Z})[\mathbf{i}_o]$$
$$\mathrm{Cov}\left(\mathbf{E}_{\mathbf{i}_o}\right) = (Q[\mathbf{i}_o])^{-1}$$

Cross-validation residuals from precision matrices

Denote $\mathbf{E}_{\mathbf{i}_o} = \mathbf{Z}[\mathbf{i}_0] - \widehat{\mathbf{Z}}^{(-\mathbf{i}_0)}[\mathbf{i}_0]$ (either way).

Simple Kriging case ($Q = \Sigma^{-1}$)

$$\begin{aligned} \mathbf{E}_{\mathbf{i}_o} &= (Q[\mathbf{i}_o])^{-1}(Q\mathbf{Z})[\mathbf{i}_o] \\ \mathrm{Cov}\left(\mathbf{E}_{\mathbf{i}_o}\right) &= (Q[\mathbf{i}_o])^{-1} \end{aligned}$$

Universal Kriging case ($\widetilde{Q} = Q - QF(FQF^\top)F^\top Q$)

$$\begin{aligned} \mathbf{E}_{\mathbf{i}_o} &= (\widetilde{Q}[\mathbf{i}_o])^{-1}(\widetilde{Q}\mathbf{Z})[\mathbf{i}_o] \\ \mathrm{Cov}\left(\mathbf{E}_{\mathbf{i}_o}\right) &= (\widetilde{Q}[\mathbf{i}_o])^{-1} \end{aligned}$$

Theorem

For any $\mathbf{i} \in \mathcal{S}$*,the Universal Kriging residual* $\mathbf{E_i} = \mathbf{Z[i]} - \widehat{\mathbf{Z}^{(-\mathbf{i})}}[\mathbf{i}]$ *obtained when predicting at locations indexed by* $\mathbf{i}$ *based on observations at* $-\mathbf{i}$ *writes*

$$\mathbf{E_i} = (\widetilde{Q}[\mathbf{i}])^{-1}(\widetilde{Q}\mathbf{Z})[\mathbf{i}].$$

Consequently, for any $q > 1$ *and* $\mathbf{i}_1, \ldots, \mathbf{i}_q \in \mathcal{S}$*, the* $\mathbf{E}_{\mathbf{i}_j}$ $(1 \leq j \leq q)$ *are jointly Gaussian, centred, and with covariance structure given by*

$$\mathrm{Cov}(\mathbf{E_i}, \mathbf{E_j}) = (\widetilde{Q}[\mathbf{i}])^{-1}\widetilde{Q}[\mathbf{i}, \mathbf{j}](\widetilde{Q}[\mathbf{j}])^{-1} \quad (\mathbf{i}, \mathbf{j} \in \mathcal{S}).$$

Theorem

For any $\mathbf{i} \in \mathcal{S}$*,the Universal Kriging residual* $\mathbf{E}_\mathbf{i} = \mathbf{Z}[\mathbf{i}] - \widehat{\mathbf{Z}^{(-\mathbf{i})}}[\mathbf{i}]$ *obtained when predicting at locations indexed by* $\mathbf{i}$ *based on observations at* $-\mathbf{i}$ *writes*

$$\mathbf{E}_\mathbf{i} = (\widetilde{Q}[\mathbf{i}])^{-1}(\widetilde{Q}\mathbf{Z})[\mathbf{i}].$$

Consequently, for any $q > 1$ *and* $\mathbf{i}_1, \ldots, \mathbf{i}_q \in \mathcal{S}$*, the* $\mathbf{E}_{\mathbf{i}_j}$ $(1 \leq j \leq q)$ *are jointly Gaussian, centred, and with covariance structure given by*

$$\mathrm{Cov}(\mathbf{E}_\mathbf{i}, \mathbf{E}_\mathbf{j}) = (\widetilde{Q}[\mathbf{i}])^{-1}\widetilde{Q}[\mathbf{i},\mathbf{j}](\widetilde{Q}[\mathbf{j}])^{-1} \quad (\mathbf{i},\mathbf{j} \in \mathcal{S}).$$

In particular, for the case of an ensemble of folds $\mathfrak{I} = (\mathbf{i}_1, \ldots, \mathbf{i}_q)$ *such that concatenation of* $\mathbf{i}_1, \ldots, \mathbf{i}_q$ *gives* $(1, \ldots, n)$*, then*

$$\mathrm{Cov}(\mathbf{E}_\mathfrak{I}) = \widetilde{B}_\mathfrak{I}\widetilde{Q}\widetilde{B}_\mathfrak{I},$$

where $\widetilde{B}_\mathfrak{I} = \mathrm{blockdiag}\left((\widetilde{Q}[\mathbf{i}_1])^{-1}, \ldots, (\widetilde{Q}[\mathbf{i}_q])^{-1}\right)$.