On Gaussian Process multiple-fold cross-validation

David Ginsbourger

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.

MASCOT NUM 2021 meeting

April 27th 2021

Outline

Outline

Cross-validation for Gaussian Process models

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

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

Cross-validation for Gaussian Process models

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

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

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

Cross-validation for Gaussian Process models

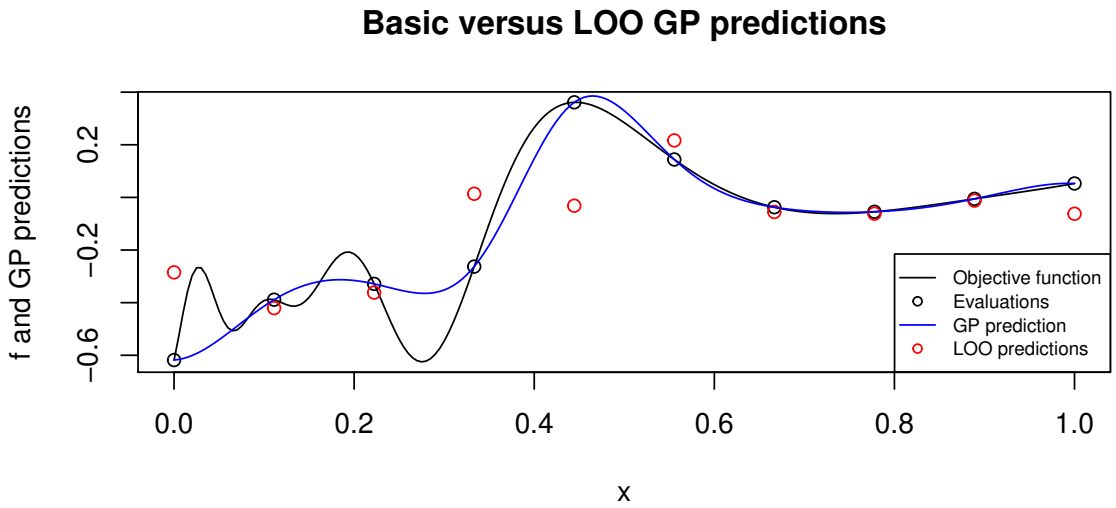
In GP modelling, cross-validation (CV) has been used for

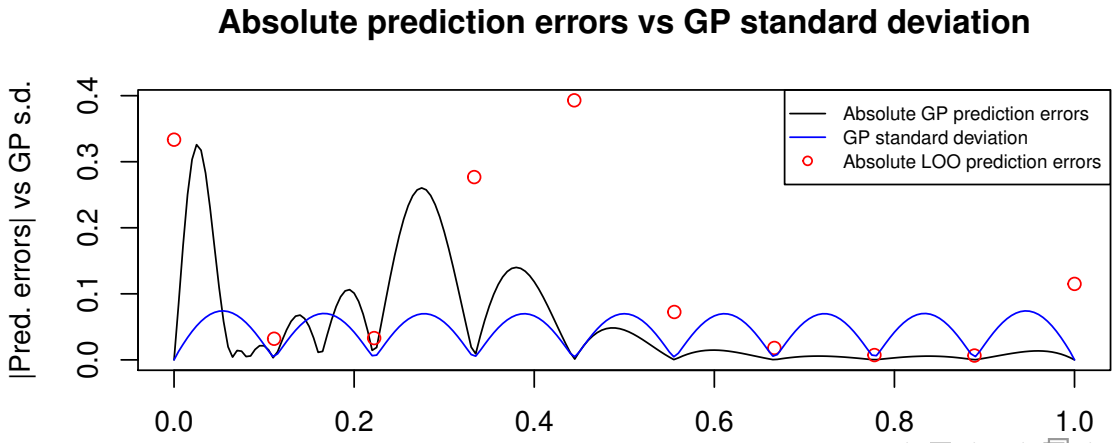
- validating models without requiring external/validation data,
- estimating hyperparameters (via criteria building on CV outputs),
- and also, for guiding sequential design strategies

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

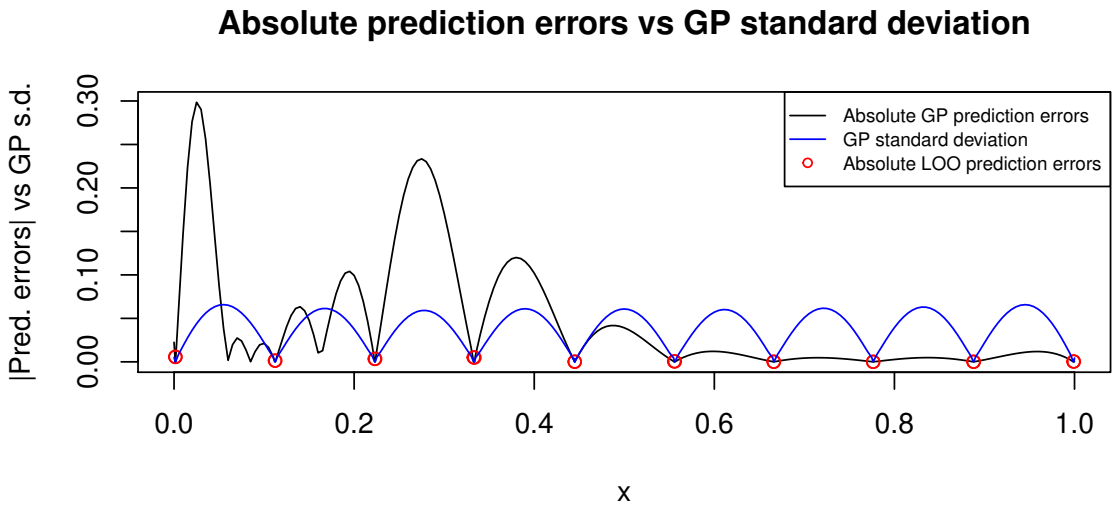
This operation is conducted multiple times in sequence, i.e. for multiple folds, and then diagnostics/criteria are calculated based on the corresponding set of residual vectors and related model outcomes.

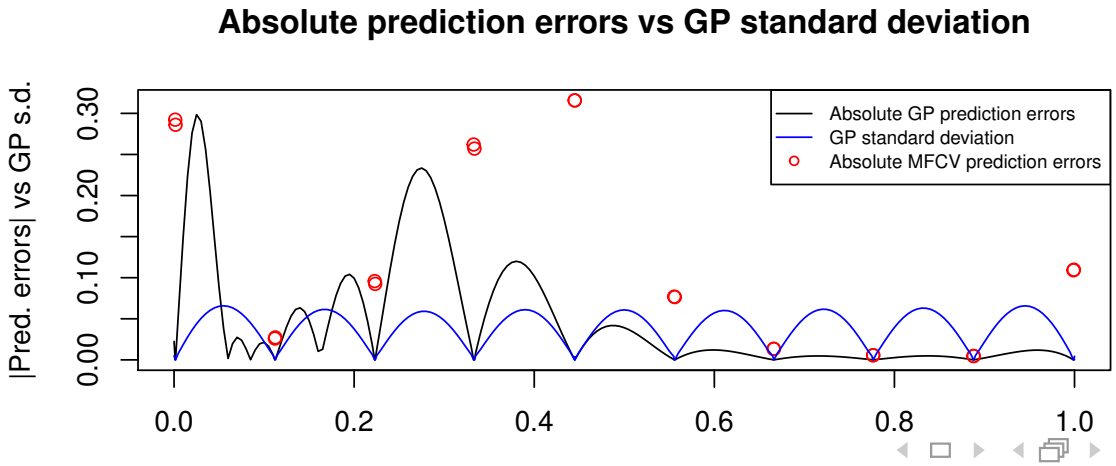
LOO-CV based on regularly spaced points





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





On notation and settings

To keep things fairly general and simple, let us remark that things boil down here to predicting subvectors of a squared integrable (often, Gaussian) random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ given complementary subvectors.

Denote $\mathbf{Z}$'s mean by $\boldsymbol{\mu}$, its covariance matrix by $K = \sigma^2 R_\theta$, and further

On notation and settings

To keep things fairly general and simple, let us remark that things boil down here to predicting subvectors of a squared integrable (often, Gaussian) random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ given complementary subvectors.

Denote $\mathbf{Z}$'s mean by $\boldsymbol{\mu}$, its covariance matrix by $K = \sigma^2 R_\theta$, and further

- by $\mathbf{Z_i}$ the subvector of $\mathbf{Z}$ associated with $\mathbf{i} \in \mathcal{S}$ (ordered index vectors),
- by $\widehat{\mathbf{Z}}_\mathbf{i}$ the vector of GP/Kriging mean values at $i \in \mathbf{i}$ based on the vector of responses "at" the remaining indices, denoted $\mathbf{Z}_{-\mathbf{i}}$, and
- by $\mathbf{E_i} = \mathbf{Z_i} - \widehat{\mathbf{Z}}_\mathbf{i}$ the residual obtained when predicting at indices contained in $\mathbf{i}$ based on observations at the remaining indices.

It was shown –Cf. notably Dubrule 1983, Bachoc 2013– that the inverse of K (or related matrices) are instrumental in efficiently calculating CV residuals.

It was shown –Cf. notably Dubrule 1983, Bachoc 2013– that the inverse of K (or related matrices) are instrumental in efficiently calculating CV residuals.

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

Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ *be a real* $n \times n$ *matrix with* A, B, C, D *of meaningful dimensions. Assuming that* D *and* $A - BD^{-1}C$ *are invertible, then so is* M *with*

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

It was shown –Cf. notably Dubrule 1983, Bachoc 2013– that the inverse of K (or related matrices) are instrumental in efficiently calculating CV residuals.

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

Let $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ *be a real* $n \times n$ *matrix with* A, B, C, D *of meaningful dimensions. Assuming that* D *and* $A - BD^{-1}C$ *are invertible, then so is* M *with*

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

Let us briefly recall well-known results for the Leave-One-Out (LOO) case, where index vectors are singletons simplified into $i \in \{1, \ldots, n\}$.

Denoting by K^{ij} the (i,j) coefficient of the inverse covariance matrix K^{-1}, the vector of concatenated LOO residuals can indeed be written in compact form:

$$\mathbf{E} = \mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})\mathrm{K}^{-1}\mathbf{Z}.$$

Denoting by K^{ij} the (i,j) coefficient of the inverse covariance matrix K^{-1}, the vector of concatenated LOO residuals can indeed be written in compact form:

$$\mathbf{E} = \mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})\mathrm{K}^{-1}\mathbf{Z}.$$

In turn, the squared norm of LOO residuals can be elegantly expressed as

$$\boxed{\|\mathbf{E}\|^2 = \mathbf{Z}'K^{-1}\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-2})\mathrm{K}^{-1}\mathbf{Z}},$$

a formula that has been useful for covariance parameter estimation, see

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.

Computational speed-ups of fast versus “naive” LOO

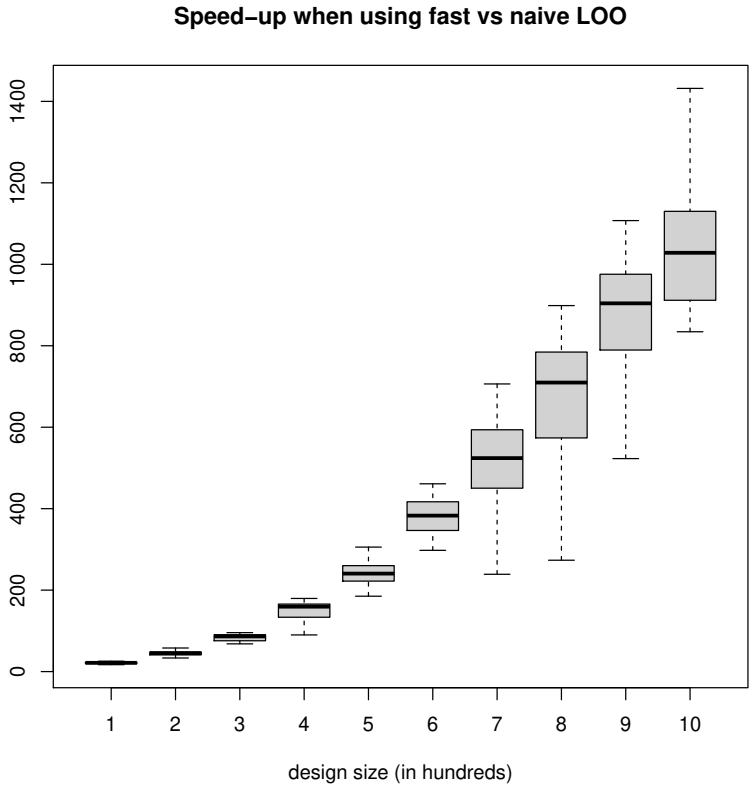


Figure: Speed-up (ratio between times required to run the naive and fast methods) measured for LOO on 10 regular designs, with 100 to 1000 points equidistributed on $[0, 1]$, where each speed-up measure is repeated 50 times.

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 (2021).

Fast calculation of Gaussian Process multiple-fold cross-validation residuals and their covariances.
arXiv:2101.03108,

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 (2021).

Fast calculation of Gaussian Process multiple-fold cross-validation residuals and their covariances.
arXiv:2101.03108,

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* $\mathbf{E_i} = \mathbf{Z_i} - \widehat{\mathbf{Z}}_\mathbf{i}$ *obtained when predicting at locations indexed by* $\mathbf{i}$ *based on observations at* $-\mathbf{i}$ *writes*

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

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*

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

Theorem

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

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

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*

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

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*

$$\mathrm{Cov}(\mathbf{E}_\mathfrak{I}) = D_\mathfrak{I} K^{-1} D_\mathfrak{I},$$

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

Keys towards the proof (1/2)

The proof relies on block matrix inversion results.

Reformulating indeed a textbook result presented in Horn and Johnson, we have indeed for arbitrary indices such as the inverses involved do exist,

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

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.

Reformulating indeed a textbook result presented in Horn and Johnson, we have indeed for arbitrary indices such as the inverses involved do exist,

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

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

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 $\mathbf{i} \in \mathcal{S}$,

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

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 $\mathbf{i} \in \mathcal{S}$,

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

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.

$$\mathrm{Cov}(\mathbf{E}_{\mathfrak{I}}) = D_{\mathfrak{I}} \Delta_{\mathfrak{I}} K^{-1} \Delta_{\mathfrak{I}}^T D_{\mathfrak{I}} \text{ with } \Delta_{\mathfrak{I}} = (\Delta_{\mathbf{i}_1}^\top, \ldots, \Delta_{\mathbf{i}_q}^\top)^\top.$$

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.

$$\mathrm{Cov}(\mathbf{E}_{\mathfrak{I}}) = D_{\mathfrak{I}}\Delta_{\mathfrak{I}}K^{-1}\Delta_{\mathfrak{I}}^{T}D_{\mathfrak{I}} \text{ with } \Delta_{\mathfrak{I}} = (\Delta_{\mathbf{i}_1}^{\top}, \ldots, \Delta_{\mathbf{i}_q}^{\top})^{\top}.$$

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

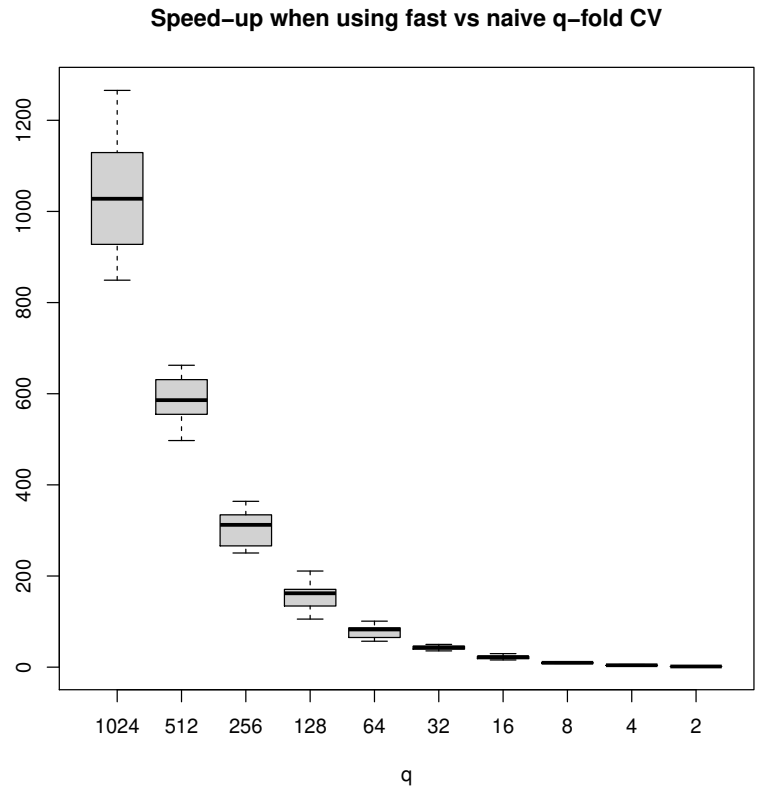
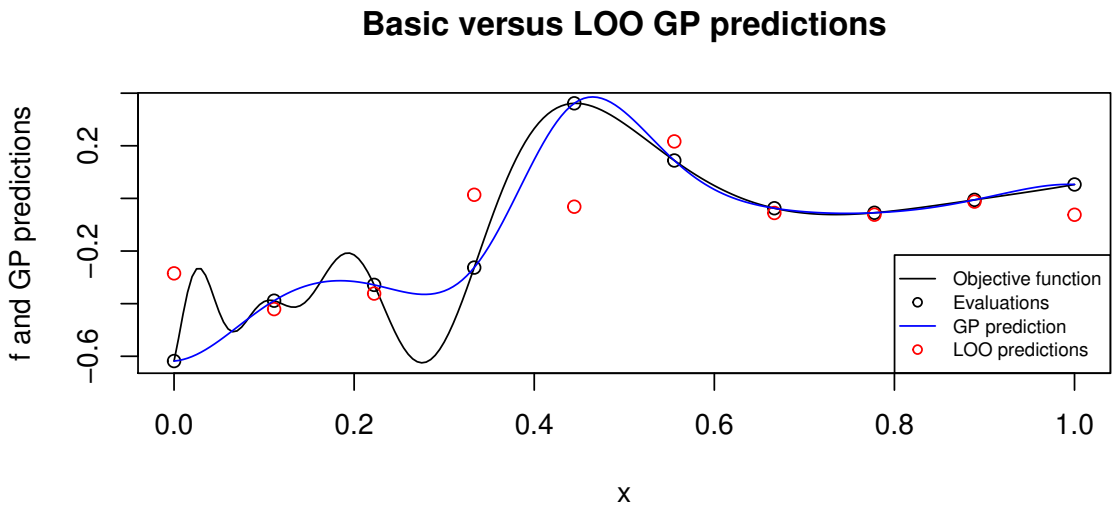
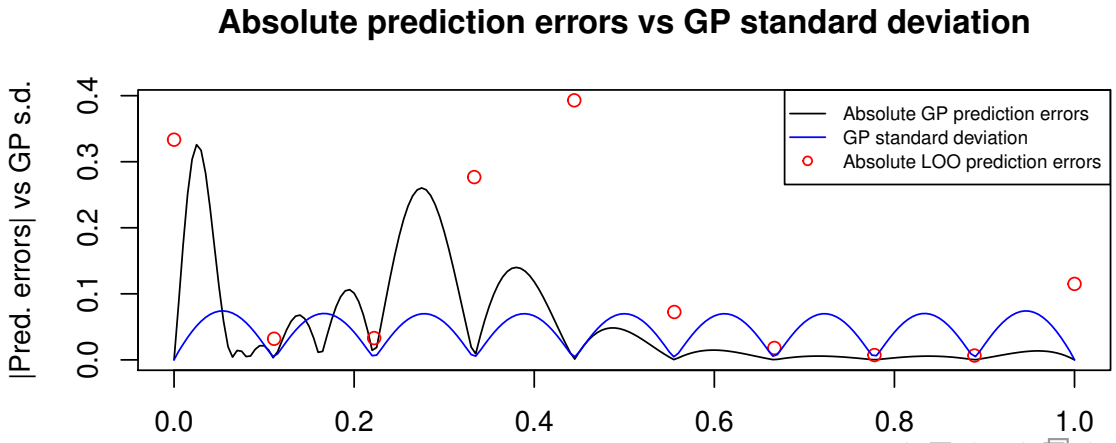


Figure: Speed-up (ratio between times required to run the naive and fast methods) measured for q-fold CV, where q decreases from 1024 to 2 and 50 seeds are used that affect here both model fitting and the folds.

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

$$\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})\mathrm{K}^{-1}\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})$$

About the correlation between LOO residuals

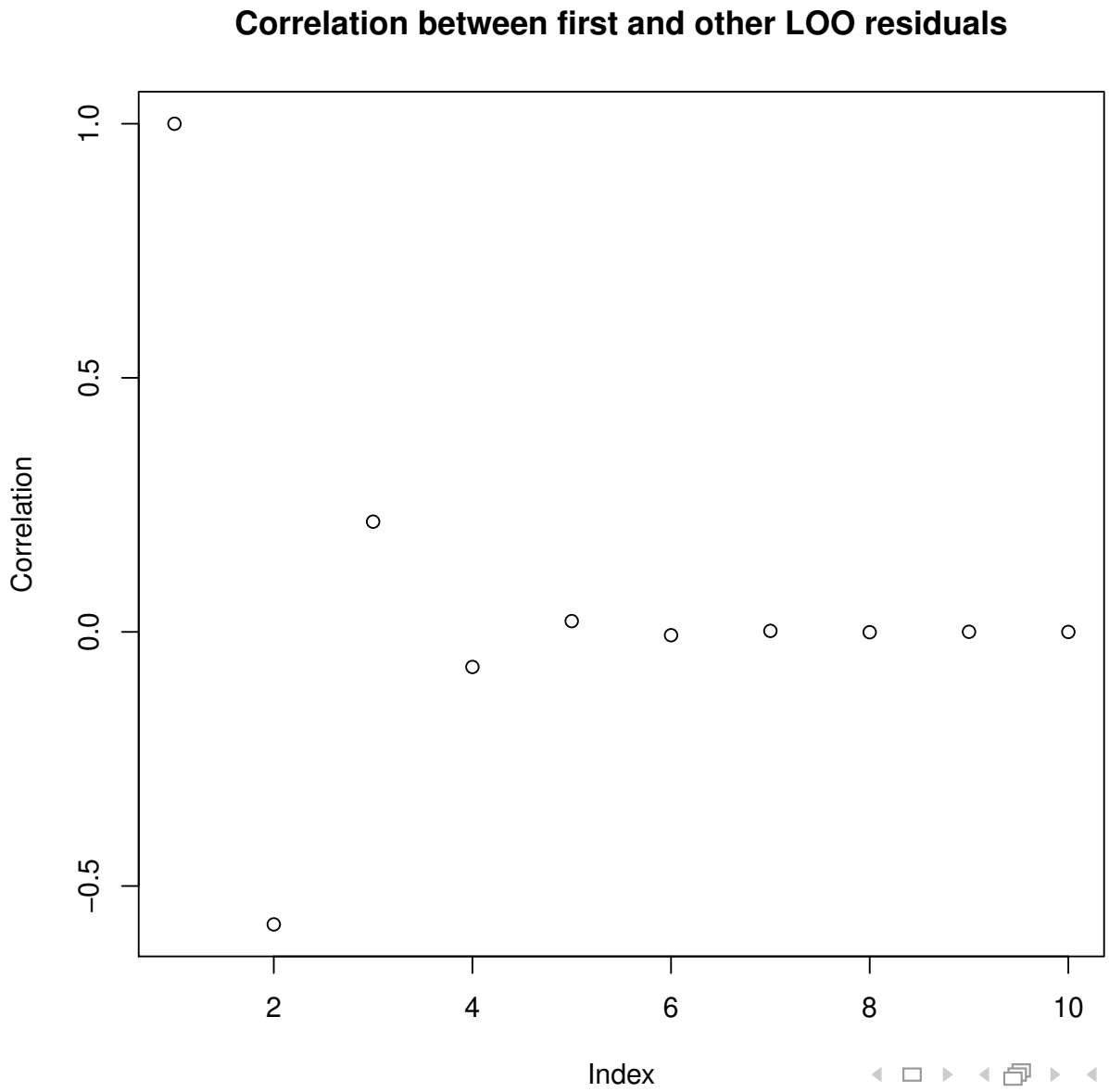
We are here in the case where $q = n$ and the $\mathbf{i}_j$'s are set to (j) $(1 \leq j \leq n)$.

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

$$\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})\mathrm{K}^{-1}\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{-1})$$

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

$$\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{+1/2})\mathrm{K}^{-1}\mathrm{diag}((\mathrm{K}^{\mathrm{ii}})^{+1/2})$$


Correlation between first and other LOO residuals
Correlation
1.0
0.5
0.0
−0.5
2
4
6
8
10
Index

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 $AD_{\mathfrak{J}}K^{-1}D_{\mathfrak{J}}A^{\top} = I_n$ 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 $AD_{\mathfrak{J}}K^{-1}D_{\mathfrak{J}}A^{\top} = I_n$ does the job.

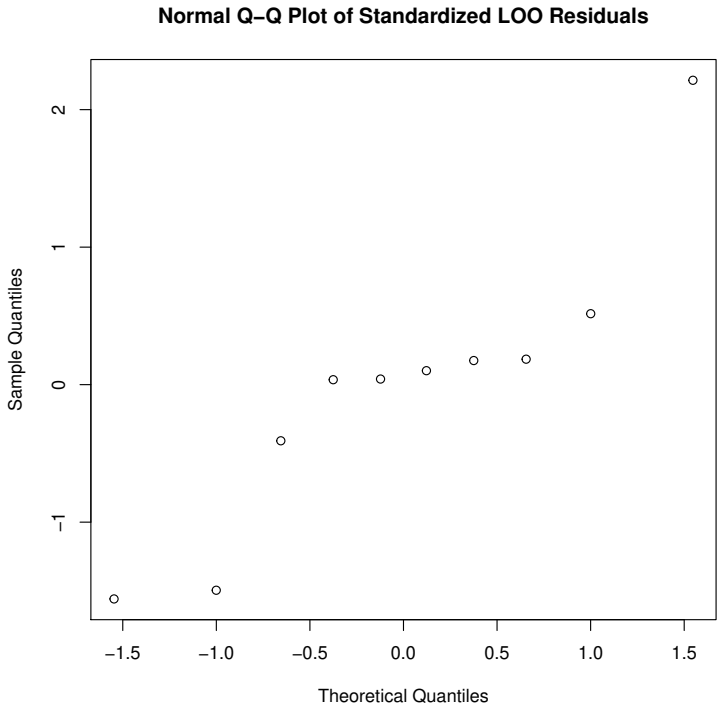
More specifically, with $A_{\mathfrak{J}} = K^{1/2}D_{\mathfrak{J}}^{-1}$, one gets indeed

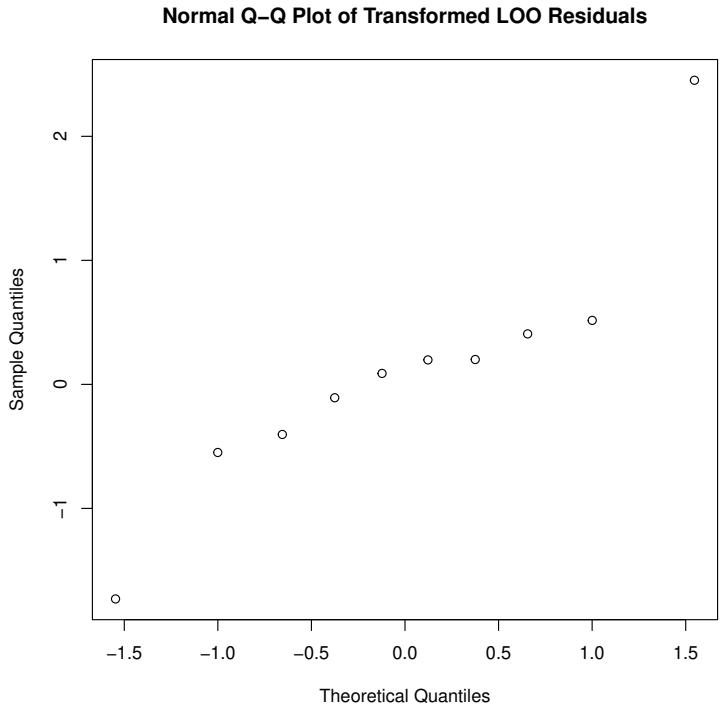
$$A_{\mathfrak{J}}\mathbf{E}_{\mathfrak{J}} = K^{-1/2}\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_n).$$

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

$$C^{(1)}_{\mathrm{LOO}}(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{Z}_i - \widehat{\mathbf{Z}}_i)^2}{\sigma^2 c^2_{-i}},$$

should take a value close to one, where $c^2_{-i} = (s^2_{-i})/\sigma^2$.

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

$$C^{(1)}_{\mathrm{LOO}}(\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{Z}_i - \widehat{\mathbf{Z}}_i)^2}{\sigma^2 c^2_{-i}},$$

should take a value close to one, where $c^2_{-i} = (s^2_{-i})/\sigma^2$.

This leads to $\widehat{\sigma}^2_{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{Z}_i - \widehat{\mathbf{Z}}_i)^2}{c^2_{-i}}$ and ultimately to the estimator above.

About the estimation of σ^2 (2/2)

Yet we claim that in order to remove undesirable effects due to the covariance between LOO residuals, it is natural to revise the criterion $C^{(1)}_{\mathrm{LOO}}(\sigma^2)$ into

$$\begin{aligned} C^{(1)}_{\widetilde{\mathrm{LOO}}}(\sigma^2) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{Z}_i - \widehat{\mathbf{Z}}_i)(DK^{-1}D)^{ij}(\mathbf{Z}_j - \widehat{\mathbf{Z}}_j) \\ &= \frac{1}{n\sigma^2}\mathbf{E}'\,\mathrm{diag}(R^{-1})R\,\mathrm{diag}(R^{-1})\mathbf{E} \\ &= \frac{1}{n\sigma^2}\mathbf{Z}'R^{-1}\mathbf{Z}, \end{aligned}$$

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

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

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

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

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

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

Building up upon the main theorem, we obtain 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

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

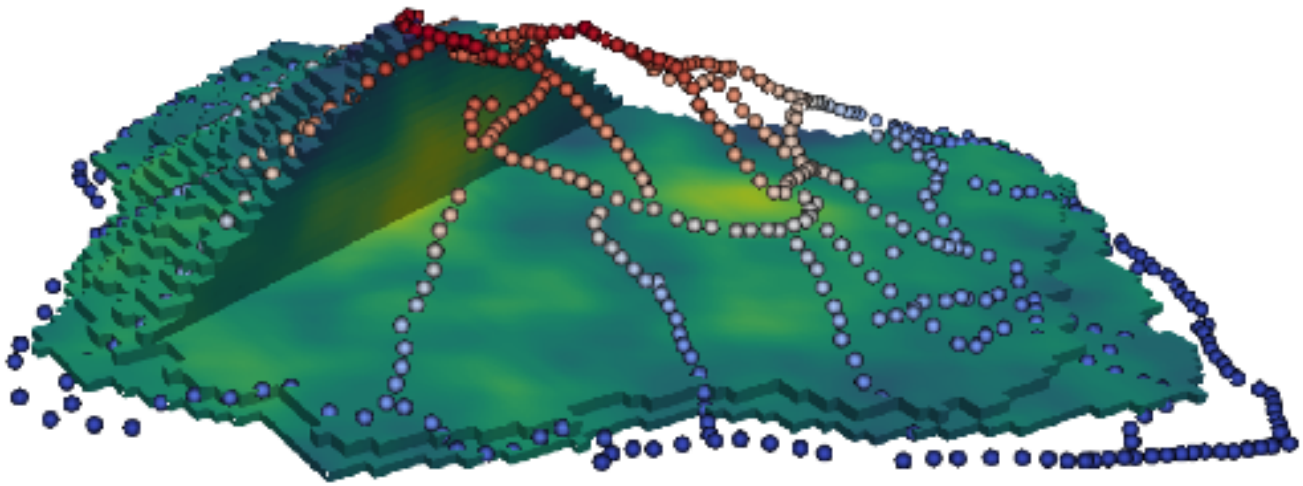
Building up upon the main theorem, we obtain that

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

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.

Gravimetric inversion on Stromboli: illustration



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).

Gravimetric inversion on Stromboli: first simulation

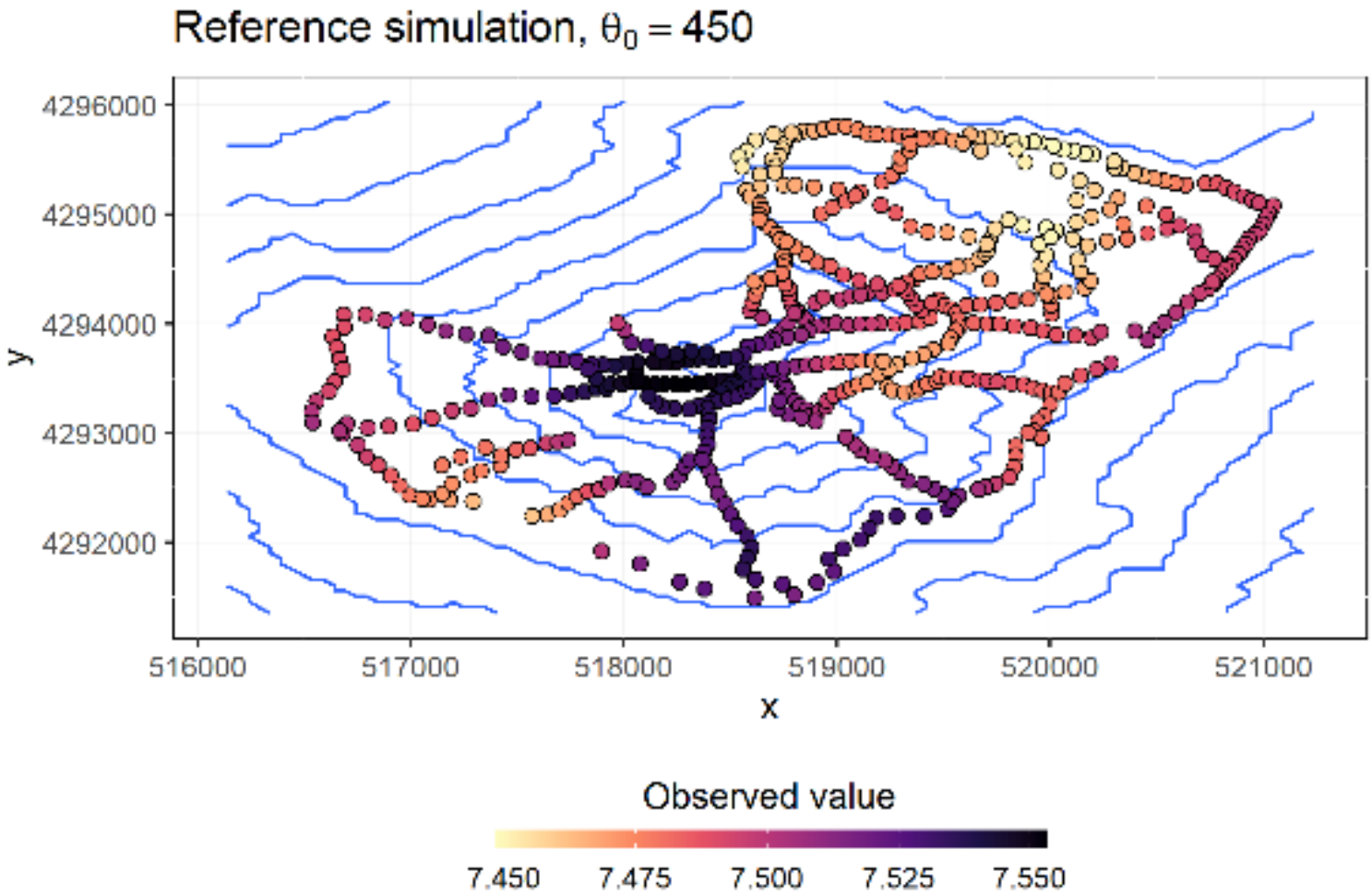


Figure: Simulated gravimetry measurements (generated with $\theta_o = 450$)

CV on the first simulation example: clustered folds

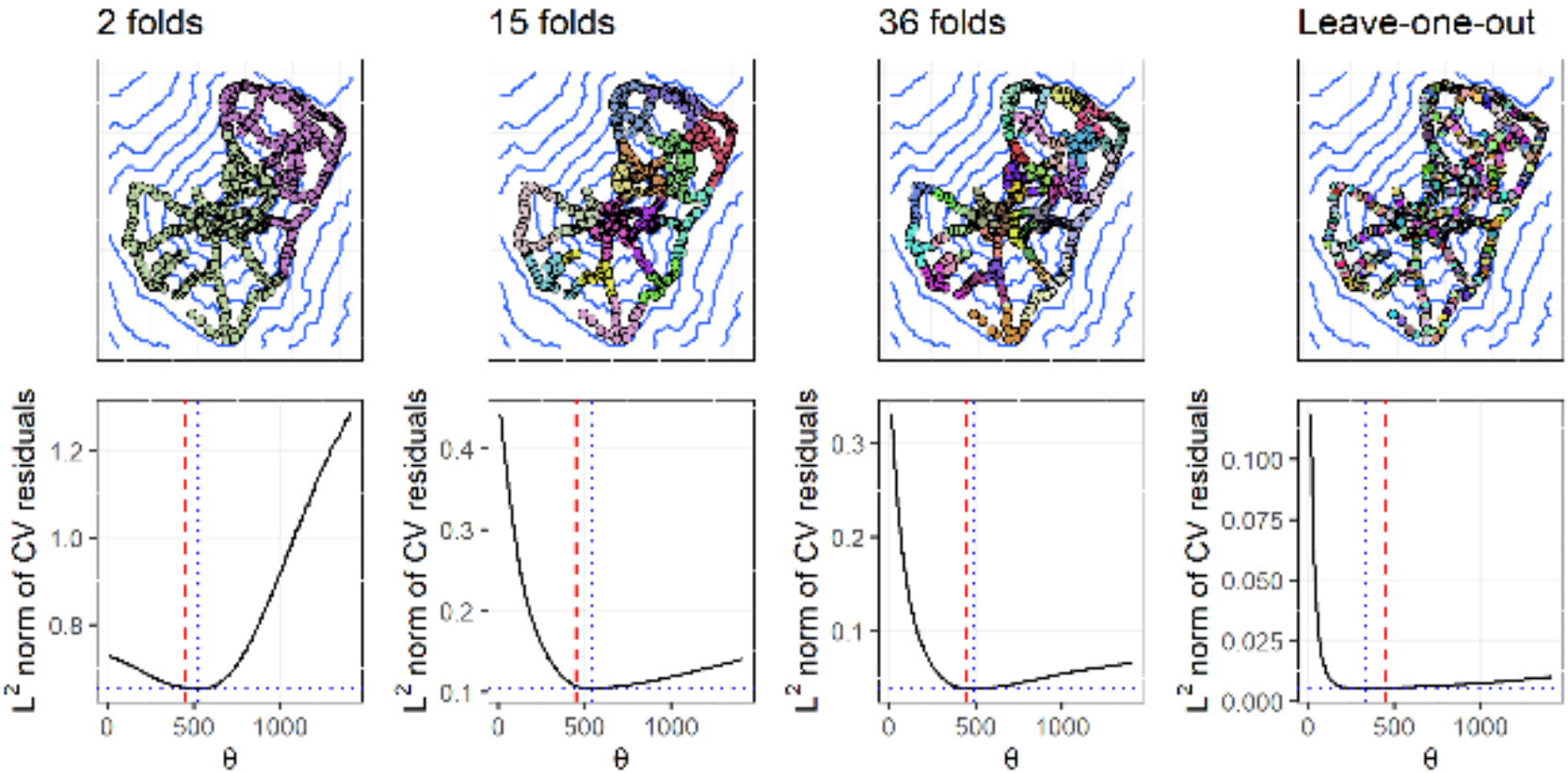


Figure: L^2 norm of CV residuals for various fold designs resulting from clustering.

CV on the first simulation example: random folds

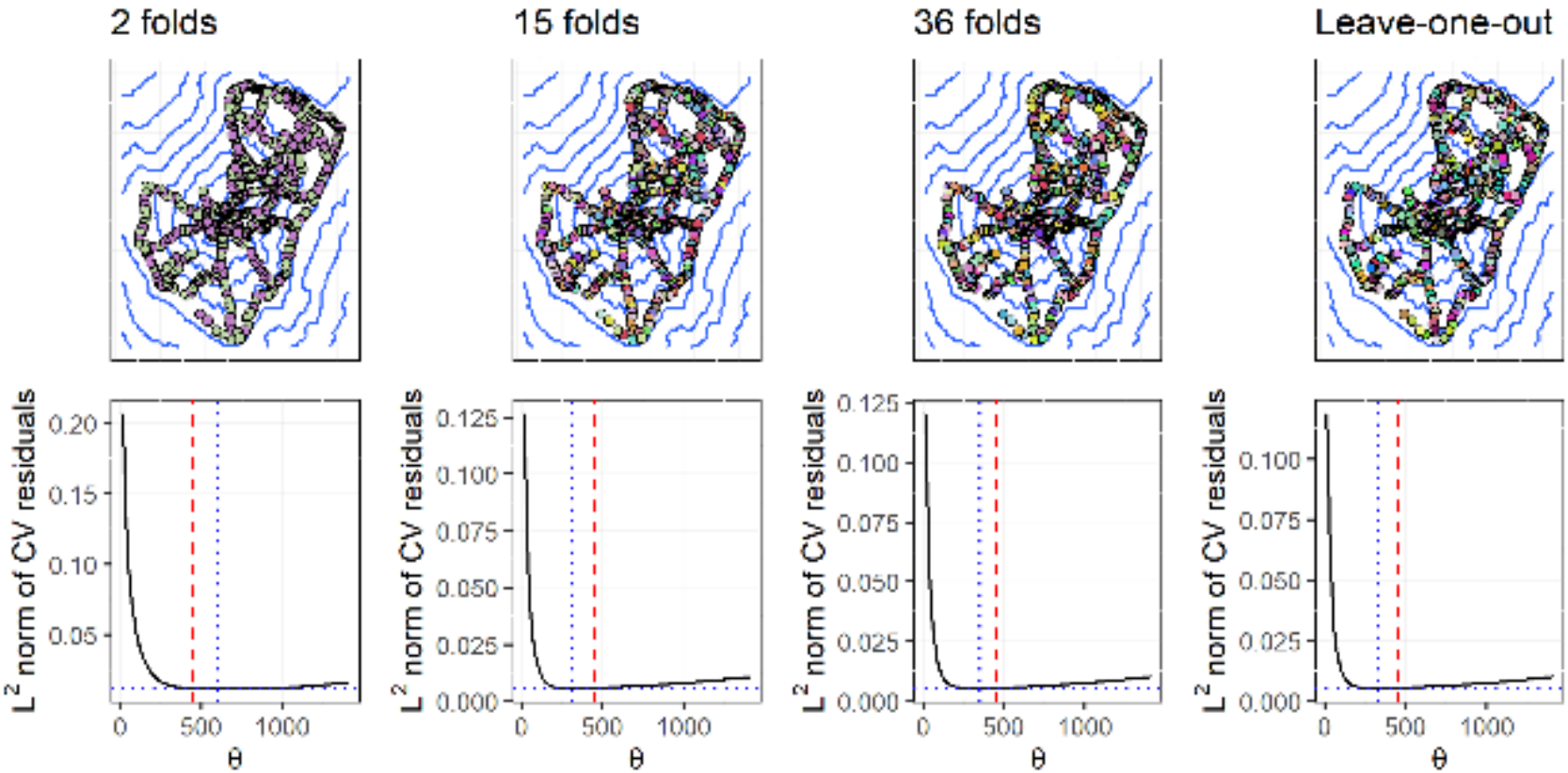


Figure: L^2 norm of CV residuals for various fold designs resulting from randomization.

Simulation study results (clustered folds)

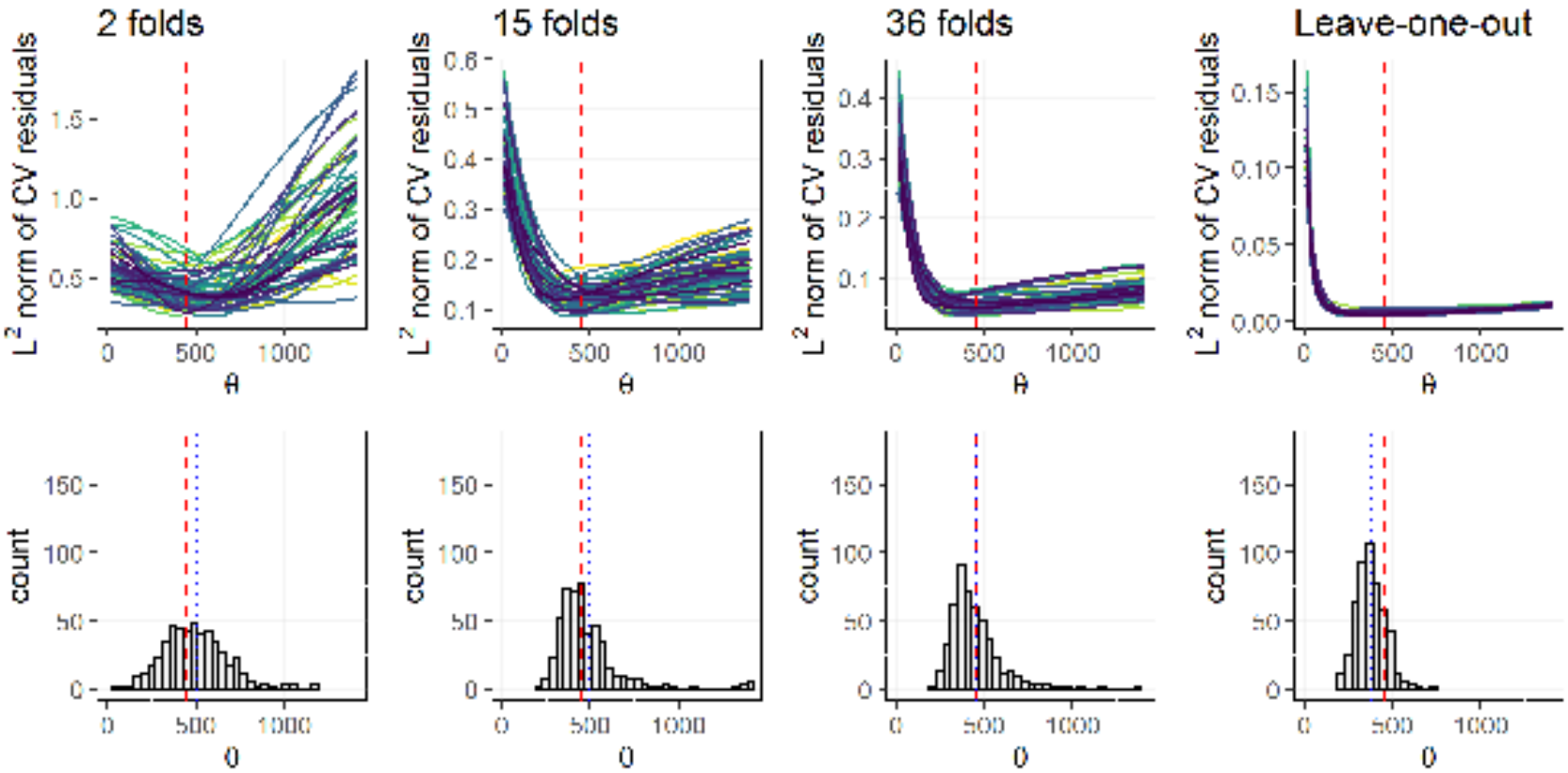


Figure: L^2 norm of residuals for 500 simulations (50 curves displayed), for various fold designs resulting from clustering.

Simulation study results (randomized folds)

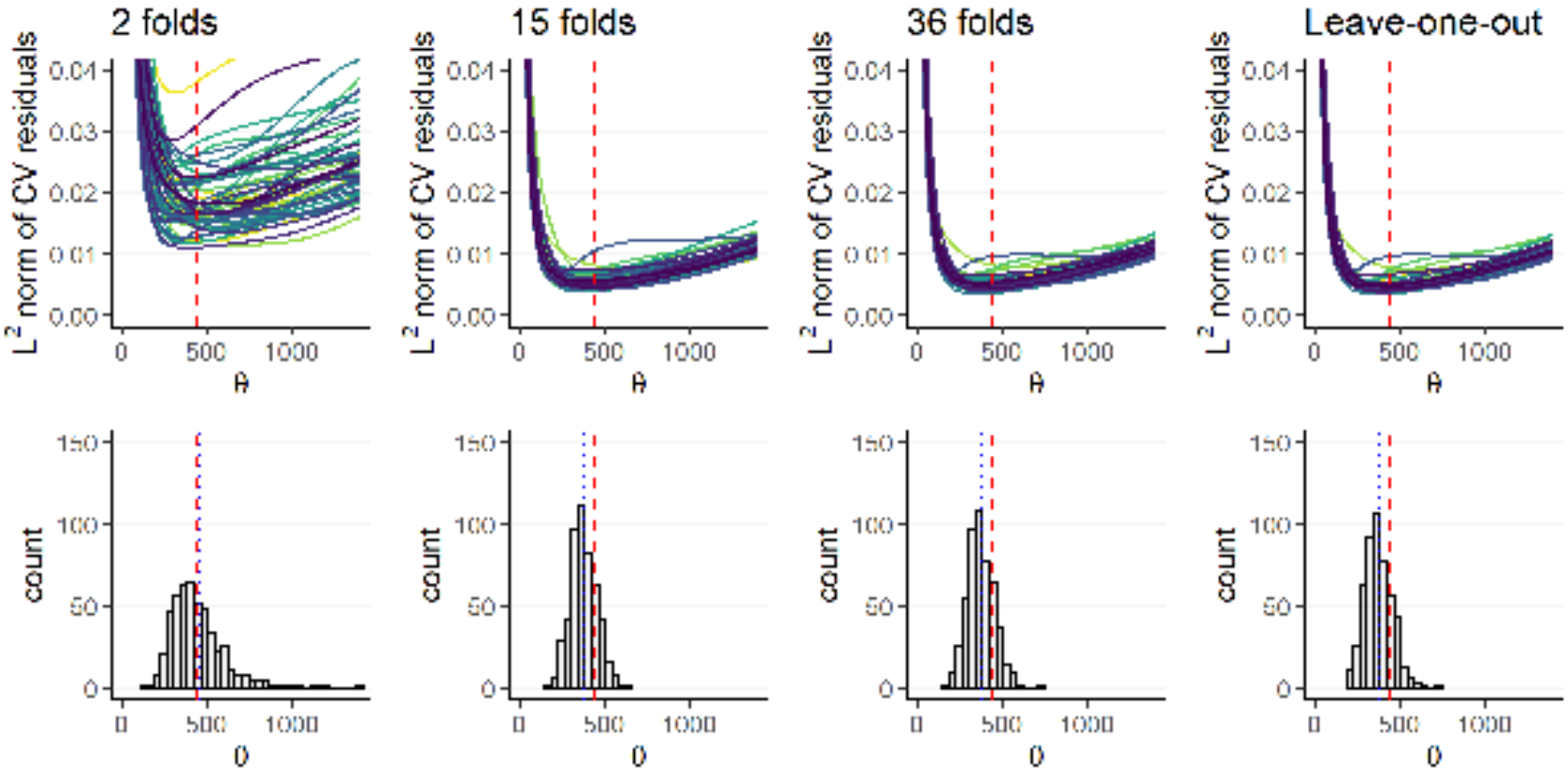


Figure: L^2 norm of residuals for 500 simulations (50 curves displayed), for various fold designs resulting from randomization.

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	