Em algorithm for generalized ridge regression with spatial covariates

Valérie Monbet, Said Obakrim, Nicolas Raillard, Pierre Ailliot

EM ALGORITHM FOR GENERALIZED RIDGE REGRESSION WITH SPATIAL COVARIATES

Said Obakrim
IRMAR
Université de Rennes 1
said.obakrim@univ-rennes1.fr

Pierre Ailliot
LMBA
Université de Bretagne Occidentale
pierre.ailliot@univ-brest.fr

Valérie Monbet
IRMAR
Université de Rennes 1
valerie.monbet@univ-rennes1.fr

Nicolas Raillard
LCSM
Ifremer
nicolas.raillard@ifremer.fr

***K*eywords** Generalized Ridge, EM algorithm, Spatial covariates, Matérn, Conditional Autoregressive

ABSTRACT

The generalized Ridge penalty is a powerful tool for dealing with overfitting and for high-dimensional regressions. The generalized Ridge regression can be derived as the mean of a posterior distribution with a Normal prior and a given covariance matrix. The covariance matrix controls the structure of the coefficients, which depends on the particular application. For example, it is appropriate to assume that the coefficients have a spatial structure in spatial applications. This study proposes an expectation-maximization algorithm for estimating generalized Ridge parameters whose covariance structure depends on specific parameters. We focus on three cases: diagonal (when the covariance matrix is diagonal with constant elements), Matérn, and conditional autoregressive covariances. A simulation study is conducted to evaluate the performance of the proposed method, and then the method is applied to predict ocean wave heights using wind conditions.

1 Introduction

Consider an experiment where we have the data $\{y, X\}$, of n observations of a continuous variable Y and $n \times d$ matrix of covariates X. Suppose that Y is related to X via a linear model

$$Y = X\beta + \epsilon, \tag{1}$$

where β are model coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the model error. We suppose that the intercept is either included in β (so that the first column of X is a vector of 1) or that Y and X are centered. The least squares estimates are the best linear unbiased estimates of the parameters β. However, in the case of multicollinearity or high-dimensionality, penalized linear regression methods, like Ridge regression, are needed to control the variance. Ridge estimator of the problem (1) is

$$\hat{\beta}_\lambda^{Ridge} = \arg\min_\beta -\ell(\beta, \sigma^2) + \lambda\|\beta\|^2 \tag{2}$$

where λ is the regularization parameter and $\ell(\beta, \sigma^2)$ is the log-likelihood of the model (1). High values of λ permit to reduce the variance and increase the bias of the model. A good model should have a trade-off between variance and bias (Hastie, Tibshirani, Friedman and Friedman, 2009). In order to find a trade-off between bias and variance, the hyperparameter λ needs to be selected.

Boonstra, Mukherjee and Taylor (2015) classified methods for selecting λ into goodness-of-fit-based and likelihood-based methods. Goodness-of-fit-based methods define a goodness of fit criterion (such as the mean squared error) and minimize it in terms of λ. The most common goodness-of-fit-based method is the k-fold cross-validation which

consists of partitioning observations into k groups and estimating β k times for each λ leaving out one group. For each λ, a goodness of fit score is calculated, and λ with the maximum score value is chosen. The typical choice of k is 5 and 10, while setting $k = n$ leads to leave-one-out cross-validation (LOOCV). LOOCV leads to a better estimation of λ; however, it is computationally expensive given that it requires fitting the model n times (Patil, Wei, Rinaldo and Tibshirani, 2021). Generalized cross-validation (GCV) (Golub, Heath and Wahba, 1979) is an approximation of LOOCV that does not require fitting n models. GCV uses a weighted version of the predicted residual error sum of squares (PRESS) statistic (Allen, 1974) as a goodness of fit criterion. One of the problems with goodness-of-fit-based methods is the selection of the grid of λ, which influences the estimation.

Assuming that $Y|\beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, Ridge regression can be derived as the mean of a posterior distribution with the prior $\beta \sim \mathcal{N}(0_d, \sigma^2\lambda^{-1} I_d)$ (van Wieringen, 2015) and as in Bayesian hierarchical linear regression, likelihood-based methods maximize the likelihood with respect to σ^2 and λ using for instance an iterative method (Lee and Nelder, 1996). Unlike goodness-of-fit-based methods, the advantage of likelihood-based approaches is, on the one hand, that they do not require grid selection for the regularization parameters. On the other hand, likelihood-based methods can be generalized to consider any form of prior for the coefficients β. In some applications, the regression coefficients can be penalized differently, or a joint penalization of the coefficients is required. For example, in spatial statistics, where predictors have a spatial structure, it is reasonable to suppose that coefficients have a spatial structure. To do that, the generalized Ridge (van Wieringen, 2015) can be used. Generalized Ridge extends the equation (2) by replacing the term $\lambda\|\beta\|^2$ to $\beta^T\Delta\beta$, where Δ is called the penalty matrix. In general, Δ depends on some regularization parameters (see, e.g., Goeman (2008) and Hemmerle (1975)); however, when the number of the regularization parameters is greater than 1, goodness-of-fit-based methods struggle with the problem of combinatorial explosion. Generalized Ridge in the hierarchical linear model framework, is equivalent to suppose that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$ where Σ_θ is a covariance matrix that depends on some parameters θ. Note that Σ_θ corresponds to the inverse of the penalty matrix Δ. The classical Ridge is a special case of this model when the covariance matrix Σ_θ is diagonal, and θ is the usual regularization parameter λ.

Considering β as a hidden variable, Bishop and Nasrabadi (2006) proposed an expectation-maximization (EM) algorithm to find the maximum likelihood estimation (MLE) of parameters of a Bayesian linear regression model. The EM algorithm (Dempster, Laird and Rubin, 1977) is a method for estimating the parameters of a model with hidden variables. The EM algorithm alternates between two steps: the expectation and maximization steps. The E-step calculates the conditional expectation of the log-likelihood given the observations and current parameters. In the M-step, the parameters are estimated by maximizing the conditional expectation of the log-likelihood calculated in the E-step. In this study, we extend the algorithm in Bishop and Nasrabadi (2006) and propose an EM algorithm to estimate the parameters of hierarchical linear regression when $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. At first, we study the case where Σ_θ is diagonal with constant elements, which corresponds to the classical Ridge in equation (2) and the problem studied by (Bishop and Nasrabadi, 2006). Then, we consider the case where the coefficients β have a spatial structure, especially when Σ_θ is the Matérn or the conditional autoregressive (CAR) covariance. A simulation study is done to assess the performance of the method. Then, the proposed method is applied to oceanographic data where the response variable represents a wave parameter in a location in the Bay of Biscay, and X represents wind conditions over the North Atlantic (Obakrim, Ailliot, Monbet and Raillard, 2022).

This paper is organized as follows. The proposed method and its special cases are presented in Section 2. Then, a simulation study is conducted in Section 3 to assess the performance of the proposed method. In section 4, we apply the methodology to oceanography data. Finally, this study is concluded in Section 5.

2 Proposed method

As stated in the introduction, Ridge regression can be viewed as a hierarchical linear model where $\beta \sim \mathcal{N}(0_d, \sigma^2\lambda^{-1} I_d)$. When there is a structure on the coefficients, it is unreasonable to consider all possible covariance functions as possible candidates for β. Therefore, we suppose that the covariance of β depends on some parameters θ, so that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$. This motivates using the EM algorithm to find the maximum likelihood estimation of the parameters, where the model parameters are then $\Theta = (\sigma^2, \theta)$. The proposed method is described in this section, and three special cases of the covariance Σ_θ (the diagonal, Matérn, and CAR) are studied.

2.1 EM algorithm for generalized Ridge

Consider the linear model (1) and assume that β is a latent variable that follows a normal distribution. We define the regression model hierarchically as

$$\begin{aligned} &\beta \sim \mathcal{N}(0_d, \Sigma_\theta) \\ &Y \mid \beta, \Theta \sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{3}$$

where $\Theta = (\sigma^2, \theta)$. Note that for simplicity, we assume that the mean of β is zero. The EM algorithm for the case where β has a non-zero mean will be presented in the Appendix.

Given a sample $y = (y_1, ..., y_n)$, the complete log-likelihood is expressed as

$$\begin{aligned} \ln p(y, \beta; \Theta) &= \ln p(y \mid \beta; \sigma^2) + \ln p(\beta; \theta) \\ &= -\frac{1}{2}\left(d\ln(2\pi) + \ln(|\Sigma_\theta|) + \beta^T \Sigma_\theta^{-1}\beta + n\ln(2\pi) + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta\|^2 \right) \end{aligned} \tag{4}$$

Maximum likelihood estimation consists of maximizing (4) with respect to the parameters Θ. This is usually done with the Expectation-Maximization algorithm in the latent variable context. The EM algorithm alternates between the E-step and M-step. In the E-step, the expectation $Q(\Theta|\Theta^{(t)})$ of the complete likelihood with respect to the posterior distribution of the latent variable β and the parameters $\Theta^{(t)}$ from the previous iteration t is calculated. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

The E-step and M-step are defined as follows

- E-step:

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}(\ln p(y, \beta; \Theta) \mid y, \Theta^{(t)}). \tag{5}$$

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2} X^T X)^{-1} \\ \mu_{\beta|y} = (X^T X + \sigma^2 \Sigma_\theta^{-1})^{-1} X^T y. \end{cases} \tag{6}$$

Note that $\mu_{\beta|y}$ defined in (6) is a generalized Ridge estimator (see e.g. van Wieringen (2015)) solution of the optimization problem

$$\mu_{\beta|y} = \arg\min_\beta \frac{\|y - X\beta\|^2}{\sigma^2} + \beta^T \Sigma_\theta^{-1} \beta \tag{7}$$

Therefore,

$$Q(\Theta|\Theta^{(t)}) = -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) + \ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y, \Theta^{(t)}) \right) + C \tag{8}$$

where C is a constant and

$$\begin{cases} \mathbb{E}(\beta\beta^T | y; \Theta^{(t)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2 | y; \Theta^{(t)}) = \|y\|^2 - 2y^T X \mu_{\beta|y} + \mathrm{Tr}(X^T X \mathbb{E}(\beta\beta^T | y; \Theta^{(t)})) \end{cases} \tag{9}$$

- M-step:

The maximization step computes

$$\Theta^{(t+1)} = \arg\max_\Theta Q(\Theta|\Theta^{(t)}) \tag{10}$$

which leads to the following updates of the parameters σ^2 and θ

$$\begin{aligned} \sigma^{2,(t+1)} &= \frac{1}{n}(\|y\|^2 - 2y^T X \mu_{\beta|y} + \mathrm{Tr}(X^T X \mathbb{E}(\beta\beta^T | y; \Theta^{(t)}))) \\ \theta^{(t+1)} &= \arg\max_\theta \ \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) \end{aligned} \tag{11}$$

2.2 Special cases

The M-step in equation (11) requires the maximization of $Q(\Theta|\Theta^{(t)})$ over the parameters of the covariance Σ_θ. In this study, we will explore three cases. First, we consider the case where Σ_θ is diagonal. Then, the case where β has a spatial structure, especially when the parametric covariance is the Matérn covariance function. Finally, we consider the conditional autoregressive model (CAR).

2.2.1 Diagonal case

In the classical Ridge, the covariance matrix of the coefficients β is supposed to be diagonal such that

$$\Sigma_\theta = \sigma_\beta^2 \mathbf{I_d}. \tag{12}$$

The M-step of the covariance in (11) becomes

$$\sigma_\beta^{2,(t+1)} = \arg\max_{\sigma_\beta^2} \; -d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})). \tag{13}$$

Setting the derivatives with respect to σ_β^2 to zero, we obtain the M-step

$$\sigma_\beta^{2,(t+1)} = \frac{\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)}))}{d}. \tag{14}$$

Note that $\frac{1}{\sigma_\beta^2}$ corresponds to the regularization parameter λ in equation (1). As stated in the introduction, Ridge regression requires the selection of the regularization parameter. Therefore, the EM algorithm can be an alternative to cross-validation for estimating Ridge coefficients along with the regularization parameter. A comparison of the two methods (cross-validation and EM algorithm) is given in the Appendix.

2.2.2 Spatial covariance functions

In spatial statistics applications, one may assume that β has a spatial structure. One way to do that is to assume that β has a parametric covariance function. There are many choices of covariance functions that are widely used for Gaussian processes and kriging (Schulz, Speekenbrink and Krause, 2018). In this study, we focus on the stationary Matérn covariance, which has the form

$$K(h; \phi, \kappa) = \frac{\sigma_\beta^2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{\phi}\right)^\kappa K_\kappa \left(\frac{h}{\phi}\right) \tag{15}$$

where h is the distance between two points, Γ is the Gamma function, and K_κ is the modified Bessel function (Abramowitz, Stegun and Romer, 1988). The Matérn function is parameterized by the variance parameter σ_β^2, the range parameter ϕ, and the smoothness parameter κ. The range parameter ϕ controls the decay rate with distance, with larger values of ϕ corresponding to more strongly correlated variables, and the smoothness parameter κ controls the mean-square differentiability of the spatial process.

The M-step of the covariance of β in (12) becomes

$$(\sigma_\beta^{2,(t+1)}, \theta^{(t+1)}) = \arg\max_{\sigma_\beta^2, \theta} \; \ln(|R_\theta^{-1}|) - d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)})) \tag{16}$$

where R_θ is the Matérn correlation and $\theta = (\phi, \kappa)$. Since the variance parameter is constant and following Bachoc (2013), the optimization of the variance parameter σ_β^2 can be carried out separately with the correlation parameters ϕ and κ. Therefore,

$$\begin{aligned} \sigma_\beta^{2,(t+1)} &= \frac{\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)}))}{d} \\ \theta^{(t+1)} &= \arg\max_{\theta} \; \ln(|R_\theta^{-1}|) - d\ln(\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)}))). \end{aligned} \tag{17}$$

The solution to the optimization problem in equation (17) cannot be done analytically; therefore, numerical optimization algorithms are used. This study uses the quasi-Newton method L-BFGS-B to optimize the parameters. Given the difficulties in estimating Matérn parameters (Kaufman and Shaby, 2013), we a priori fix the smoothness parameter as $\frac{3}{2}$, which gives the classical $\frac{3}{2}$-Matérn covariance function.

2.2.3 Conditional autoregressive model

The M-step in equation (10) requires the inversion of the covariance matrix, which can be challenging for large matrices. This problem is wildly discussed in Gaussian processes literature (Ambikasaran, Foreman-Mackey, Greengard, Hogg and O'Neil, 2015; Storkey, 1999). Therefore, it can be numerically advantageous to parameterize the precision matrix (inverse of the covariance matrix) instead of the covariance matrix. This is motivated by the fact that the precision matrix $P_\theta = \Sigma_\theta^{-1}$ can be approximated by a sparse matrix (Tajbakhsh, Aybat and Del Castillo, 2020). In fact, the

off-diagonal elements of the precision matrix correspond to the conditional covariance between two variables given the remaining variables. Therefore, conditionally independent variables have zero values in the precision matrix.

Gaussian Markov random fields (GMFs) are wildly used in spatial statistics (Cressie and Wikle, 2015). GMFs models have a Markov property making them computationally and theoretically suitable (Rue, 2001). Furthermore, (Rue and Tjelmeland, 2002) demonstrated that a GMF model can approximate a Gaussian field with a Matérn correlation function and other families of correlation functions. Conditional autoregressive (CAR) models are classes of GMFs with well-defined joint Gaussian distribution (Cressie and Kapat, 2008). This subsection will study cases where the coefficients β have the CAR model property. The joint distribution of a CAR is expressed as

$$\beta \sim \mathcal{N}(0, \tau^2(I_d - \alpha H)^{-1}\Phi). \tag{18}$$

The distribution of β depends on unknown parameters α and τ^2, and many types of CAR models depend on the choice of the matrix H and Φ. Following (Besag, York and Mollié, 1991), in this study, we consider the Weighted CAR (WCAR) model where

$$\Phi = \text{diag}(|N_1|^{-1}, ..., |N_d|^{-1}) \tag{19}$$

where $|N_i|$ is the number of neighbors of location i and $H = \left(\frac{a_{ij}}{|N_i|}\right)_{d\times d}; i, j = 1, ..., d$, where a_{ij} is the (i, j) element of the adjacency matrix $A = (a_{ij})_{d\times d}$, where $a_{ij} = a_{ji} = 1$ if and only if location i and j are neighbors and otherwise $a_{ij} = 0$. Putting $P_\theta = \tau^{-2}(I_d - \alpha H)\Phi^{-1}$, the second part of the M-step in the equation (11) becomes

$$\theta^{(t+1)} = \arg\max_\theta \ \ln(|P_\theta|) - \text{Tr}(P_\theta \mathbb{E}(\beta\beta^T \mid y, \Phi^{(t)})) \tag{20}$$

where $\theta = (\tau^2, \alpha)$.

As for the Matérn covariance, the solution to the optimization problem (20) cannot be done analytically, and the numerical optimization algorithm L-BFGS-B is used. Note that the optimization of the variance parameter τ^2 can also be carried out separately with the parameter α.

Remark that this leads to a spatial extension of the fused Ridge method proposed in (Goeman, 2008). When $\alpha = 1$, we obtain

$$\frac{1}{\tau^2}\beta^T\Phi^{-1}(I_d - \alpha H)\beta = \frac{1}{2\tau^2}\sum_{(i,j)|a_{ij}=1}(\beta_i - \beta_j)^2. \tag{21}$$

This shows that any spatial coefficient variations will be penalized when solving (7). In this case, replacing the L2 norm with the L1 norm leads to the fused LASSO method proposed in (Tibshirani, Saunders, Rosset, Zhu and Knight, 2005). However, the matrix $(I_p - \alpha H)$ is semi-positive definite when $\alpha = 1$ and thus Σ_θ is degenerate. Hereafter we impose the constraints $|\alpha| < 1$ to ensure that the precision matrix is positive definite. Another strategy would consist of adding a regular Ridge penalty (e.g., the discussion in van Wieringen (2015)).

3 Simulation study

In this section, a simulation study is conducted to assess the performance of the proposed method for estimating model parameters for the three cases: diagonal, Matérn, and CAR.

3.1 Setup

This study focuses on using the proposed method for spatial applications. Therefore, we consider a 15×15 regular spatial grid in a square domain $[1, 15]^2$ where each location j has a covariate x_j. We generate $X = (x_{ij})_{n\times d}$ of n independent and identically distributed observations from a multivariate normal distribution with zero mean and a Matérn covariance with some arbitrary parameters $(\sigma_x^2, \phi_x, \kappa_x) = (6, 2, 3/2)$. Then, the coefficients β, kept the same for all observations, are simulated using either the diagonal, Matèrn, or CAR case. Finally, for a given σ^2, Y is simulated from the normal distribution according to equation (3).

The parameters chosen for each case are:

- Diagonal: $\sigma^2 = 36$ and $\sigma_\beta^2 = 7$
- Matérn: $\sigma^2 = 36$, $\sigma_\beta^2 = 0.1$ and $\phi = 4$
- CAR: $\sigma^2 = 36$, $\tau^2 = 1$ and $\alpha = 0.9$

The parameters are chosen so that the results of the three methods are comparable. For the CAR model, we consider four neighbors to construct the adjacency matrix, and we chose $\alpha = 0.9$ to sufficiently smooth the resulting coefficients.

The EM algorithm is initialized with an arbitrary set of parameters, and the E-step and M-step are repeated until no further improvement can be made to the likelihood value or to limit the computational cost until a maximum number of iterations is reached. The computation time for one iteration on an i5-7500 CPU and 16Go computer is 0.16, 3, and 1.8 seconds for diagonal, Matérn, and CAR, respectively.

3.2 Results

At first, one simulation is done for each case (diagonal, Matérn, and CAR) with $n = 800$. The parameters are estimated using the EM algorithm presented in the previous section. Figure 1 shows the first simulation results. Left panels correspond to the true β, and right panels correspond to the estimated β using the EM algorithm. For all the cases, the EM algorithm does well in estimating the parameters, especially the variance σ^2.

To assess the influence of the sample size on the estimations, for each case, we perform 100 independent random simulations for each sample size varying from 50 to 850. For each simulation, the EM algorithm is used to estimate the parameters. Figure 2 shows the normalized root mean square error $NRMSE_\beta$ and $NRMSE_y$ for the three cases where

$$\begin{aligned} NRMSE_\beta &= \frac{\sqrt{\frac{1}{d}\sum_j^d (\beta_j - \hat{\beta}_j)^2}}{\hat{\sigma_\beta}} \\ NRMSE_y &= \frac{\sqrt{\frac{1}{n'}\sum_i^{n'} (y_i - \hat{y}_i)^2}}{\hat{\sigma}_y} \end{aligned} \tag{22}$$

where $\hat{\beta}_j$ and $\hat{y}_i$ are the estimated β_j and y_i and $\hat{\sigma_\beta}$ and $\hat{\sigma}_y$ are the sample standard deviation of β and y, respectively. $NRMSE_y$ is calculated in a test set (which is not used in the estimation) of size $n' = \frac{n}{2}$. For the three cases, $NRMSE_\beta$ and $NRMSE_y$ decrease as the sample size increases.

To evaluate the parameter estimates, we compare the EM estimates with the maximum likelihood estimates of the parameters, hereafter referred to as MLE, knowing the true β. More precisely, the MLE estimates are defined as

$$\Theta_{\text{MLE}} = \arg\max_{\Theta} -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \beta_{\text{true}}^T \Sigma_\theta^{-1} \beta_{\text{true}} + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta_{\text{true}}\|^2\right) + C \tag{23}$$

where β_{true} is the true β simulated for each case with the parameters given in section 3.1. Along with the sample size, we are also interested in how the estimates behave when varying the dimension of X, d, and the variance parameter σ^2. Note that in practice, Θ_{MLE} cannot be found directly, given that the true β is not observed (latent variable). Therefore, we expect the EM algorithm to provide less accurate estimates than MLE. However, we expect that by varying the sample size, the dimension, and the variance σ^2, the estimations asymptotically will be close to MLE estimates.

Figures 3, 4 and 5 show boxplots of EM (red) and MLE (blue) estimates for the diagonal, Matérn and CAR cases as a function of sample size, dimension d, and variance σ^2. For the diagonal case, the estimate of σ^2 seems to converge to the true value of the parameter (blue line) when the sample size n increases as it does in the usual linear regression model. Note that the estimate of the spatial variance $sigma_\beta^2$ does not seem to converge to the true value of the parameter as the sample size increases, but when n is large enough, EM and MLE seem to provide similar results. This is not unexpected since both methods are based on a single sample of the d-dimensional field β. As expected, the dimension d also affects the estimate of the parameter σ_β^2, which converges towards the true value as d increases; however, no significant change is observed for σ^2 when d increases. The effect of the variance σ^2 on the estimation of σ_β^2 is small, and we observe that for σ^2 larger than 100, the EM and MLE tend to underestimate σ_β^2. Similar behavior can be observed for the Matérn case: the variance parameter σ^2 seems to converge towards the actual value with increasing sample size. However, there is no significant change in the other parameters (the variance σ_β and the range ϕ). The dimension d mainly influences the parameters σ_β and ϕ, which describe the spatial structure of the d-dimensional field β, and as d increases, the estimates converge to the actual values. As for the diagonal case, the EM algorithm underestimates the parameters σ_β and ϕ when the variance σ^2 increases. Finally, for the CAR case, the sample size influences the parameters σ^2 and τ^2, but only slightly the correlation parameter α, which is mainly influenced by the dimension d. The variance σ^2 has a significant influence on τ^2, but only a small one on α. To summarize:

- The sample size n mainly influences the estimation of the variance of the residuals σ^2

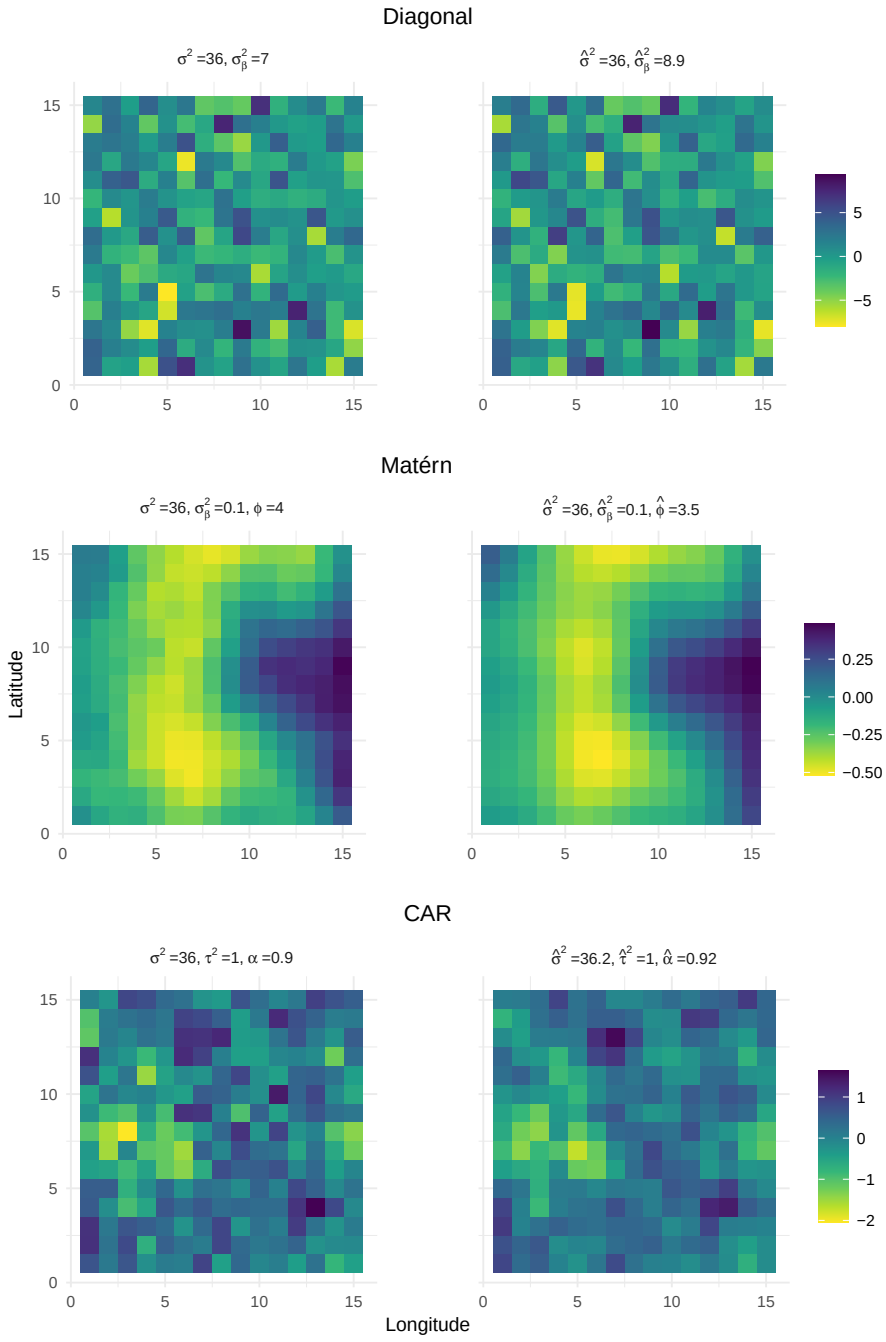

Figure 1: Simulation results for the three cases (diagonal, CAR, and Matérn). The left panels correspond to the true β coefficients with the true parameters given in section 3.1, and the right panels correspond to the β estimated when the sample size $n = 800$.

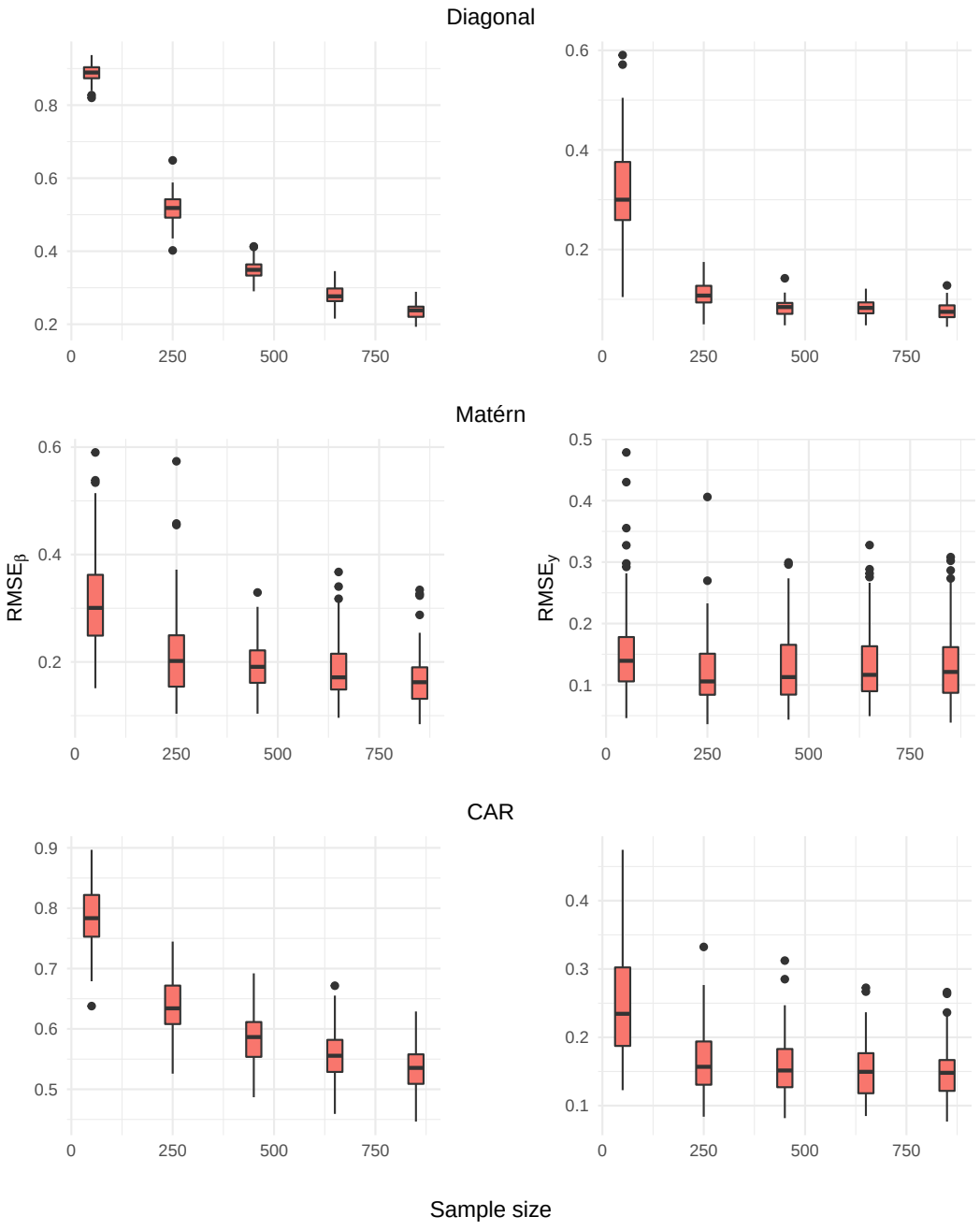

Figure 2: Results of $RMSE_\beta$ (left panels) and $RMSE_y$ (right panels) for the diagonal, CAR, and Matérn case as a function of the sample size varying from 50 to 850.

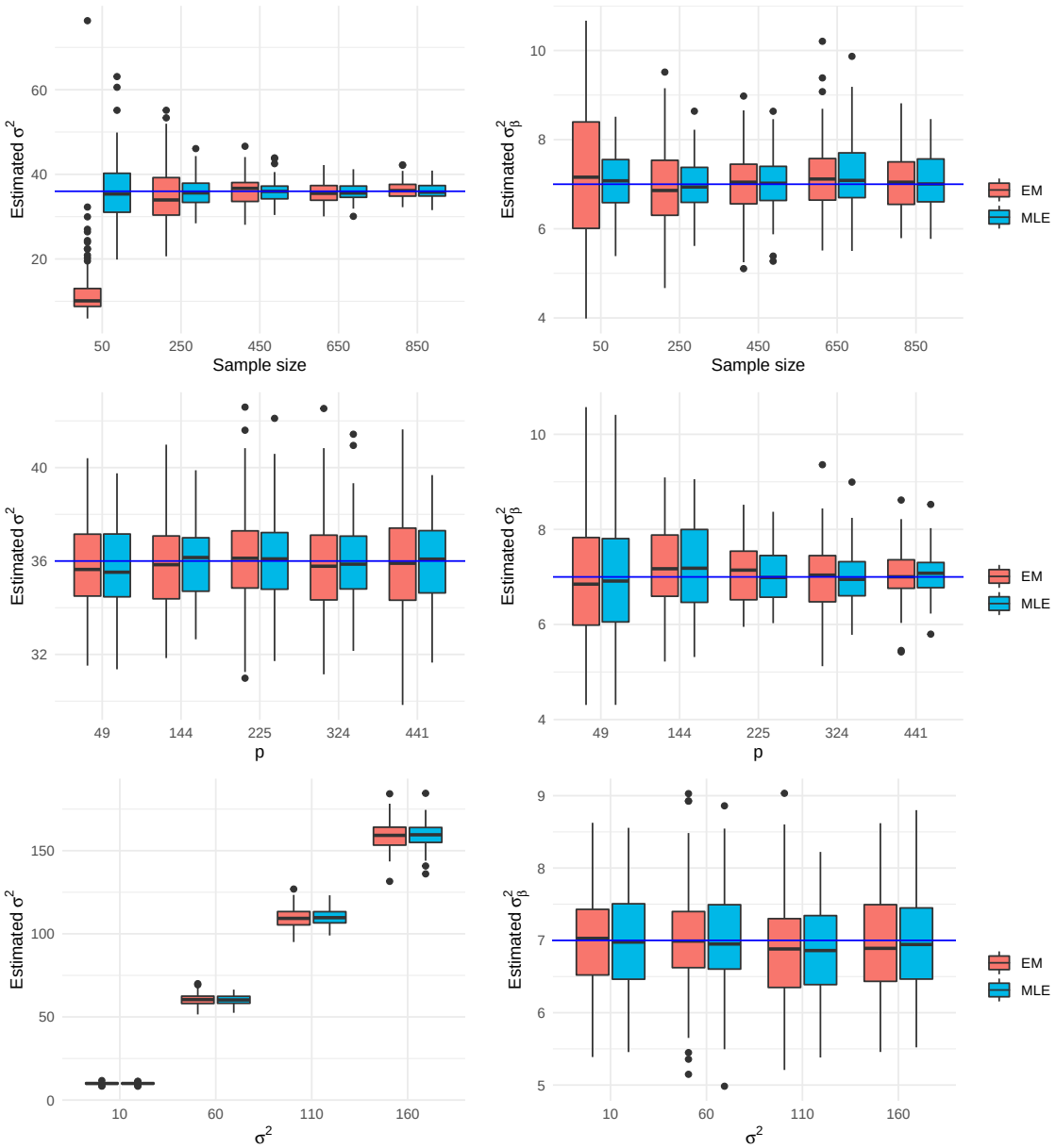

Figure 3: Estimated parameters in the case where the covariance of β is diagonal as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2 and σ^2_β, which are equal to 36 and 7, respectively.

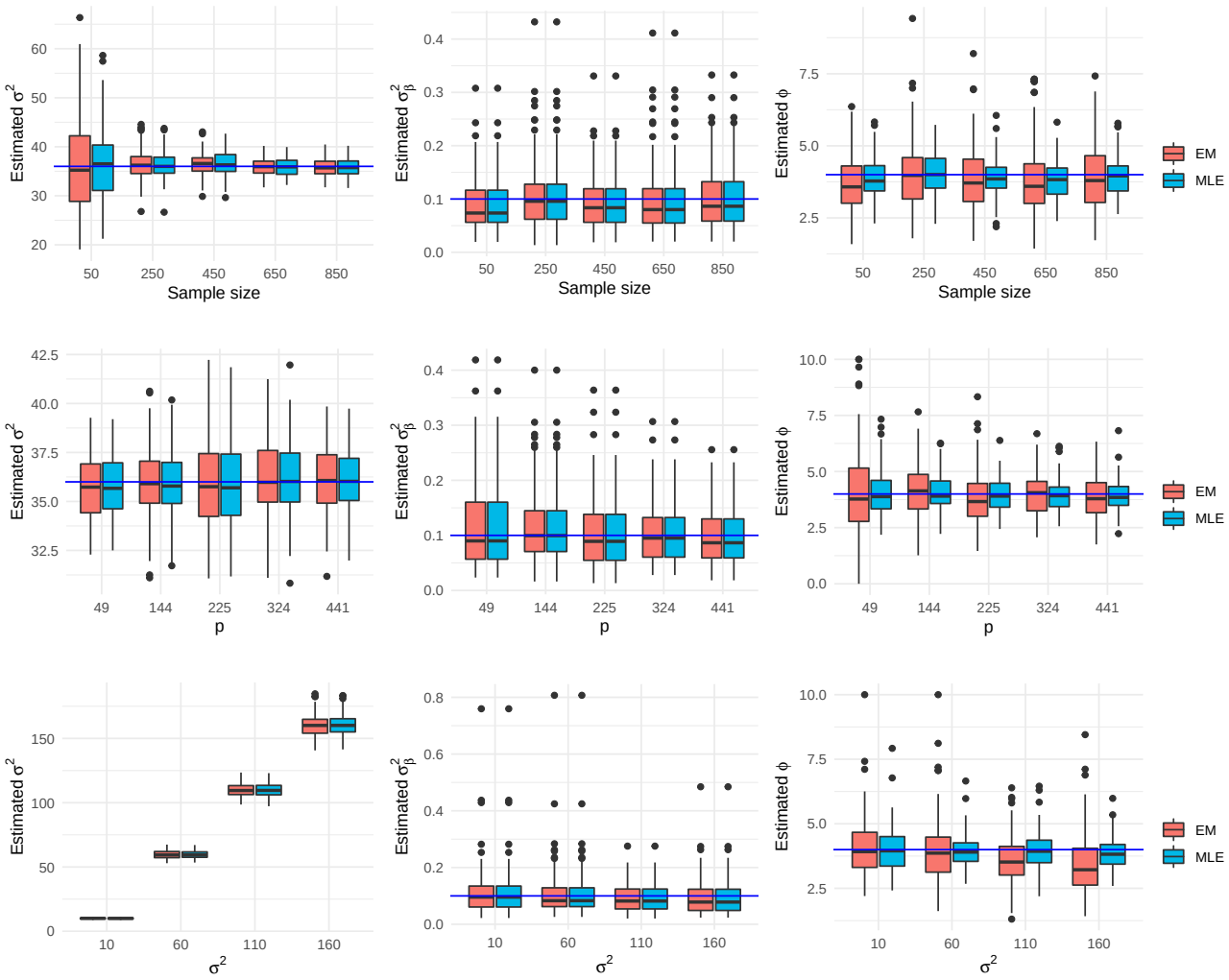

Figure 4: Estimated parameters in the case where the covariance of β is the Matérn as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ^2_β, and ϕ, which are equal to 36, 0.1, and 4, respectively.

- The parameters which describe the spatial structure of β are mainly influenced by the dimension d
- As the variance σ^2 increases, EM underestimates the parameter σ^2_β of the diagonal and Matérn case, and the range parameter ϕ
- EM estimates are close to MLE estimates in most cases when the sample size and the dimension d are large enough and the variance σ^2 is small

Another interesting aspect that needs to be studied is when the coefficients β are simulated using one covariance and estimated using another covariance model. To do that, we perform 100 independent simulations of β using the Matérn covariance function, and we estimate the parameters using the three cases: diagonal, CAR, and Matérn. Figure 6 shows the results of $NRMSE_\beta$ and $NRMSE_y$ of the experiment. It is clear that using the Matérn covariance for the estimation gives better results in terms of $NRMSE_\beta$. Not surprisingly, the diagonal case is the worst model for estimating the coefficients. However, in terms of $NRMSE_y$, there is a small difference between the three methods.

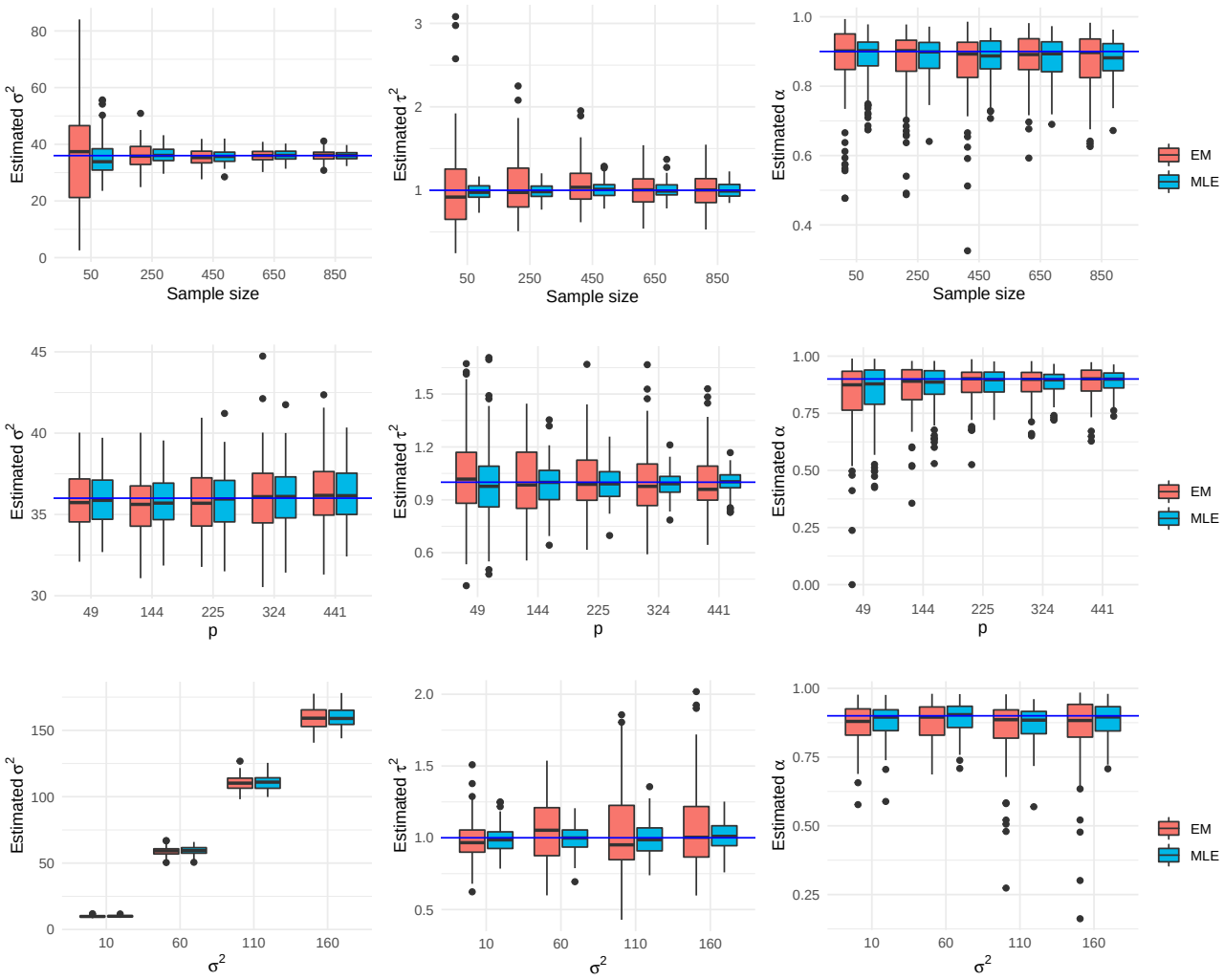

Figure 5: Estimated parameters in the case where the covariance of β is the CAR as a function of the sample size, the dimension of X, d, and the variance σ^2. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ^2_β, and α, which are equal to 36, 1, and 0.9, respectively.

4 Application

The proposed method is applied to the problem of predicting the significant wave height (H_s) at a location in the Bay of Biscay using wind conditions over the North Atlantic (figure 7), where the significant wave height is the average height of the highest third of the waves, a key measure of wave height that provides information about wave energy. The data used for H_s comes from the Homere hindcast database (Boudière, Maisondieu, Ardhuin, Accensi, Pineau-Guillou and Lepesqueur, 2013), and the wind data comes from Climate Forecast System Reanalysis (CFSR) (Saha, Moorthi, Pan, Wu, Wang, Nadiga, Tripp, Kistler, Woollen, Behringer et al., 2010). The wind data are pre-processed before being used as a predictor (see (Obakrim et al., 2022) for the pre-processing procedure). We consider 23 years of H_s and wind data from 1994 to 2016 with a temporal resolution of 3 hours.

The regression problem is of the form

$$H_s(t) = \sum_{j=1}^{d} X_j(t)\beta_j + \epsilon(t) \quad t = 1, ..., n \tag{24}$$

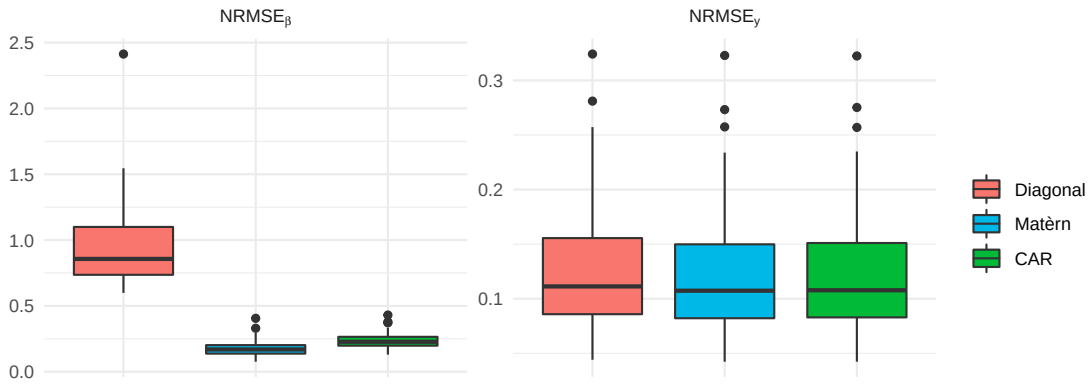

Figure 6: Results of the estimations when the true beta is simulated from Matérn with the parameters $\sigma^2 = 36$, $\sigma^2_\beta = 0.1$ and $\phi = 4$ and sample size $n = 800$. The left panel correspond to $NRMSE_\beta$ and the right one for $NRMSE_y$.

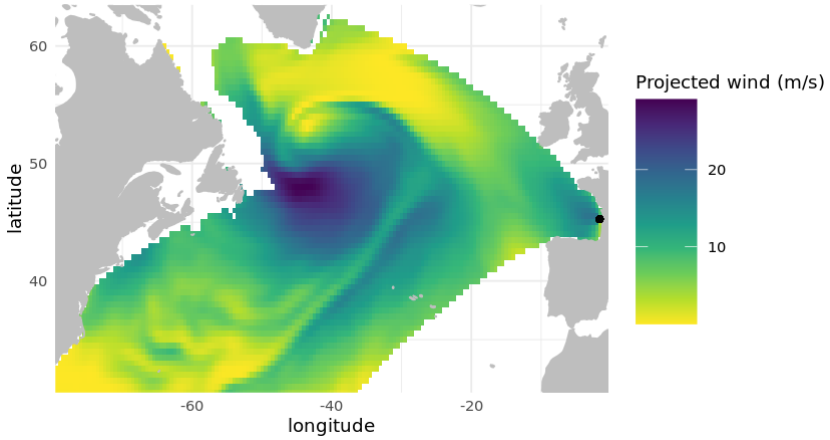

Figure 7: CFSR projected wind in the North Atlantic in 1994-01-01 00h:00. The black point represents the target point.

where $X_j(t)$ is the predictor at time t and location j defined as

$$X_j(t; t_j, \alpha_j) = \frac{1}{2\alpha_j+1} \sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \quad t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n \tag{25}$$

where W_j is the projected wind (figure 7) defined as

$$W_j = U_j \, cos\left(\frac{1}{2}(b_j - \theta_j)\right) \tag{26}$$

U_j is the wind speed, b_j is the great circle bearing, and θ_j is the wind direction at location j. α_j controls the length of the time window, and t_j is the mean travel time of waves which are estimated using the maximum correlation between H_s and the predictor

$$(\hat{t}_j, \hat{\alpha}_j) = arg \max_{t_j, \alpha_j} \big(corr(H_s, X_j^g(t_j, \alpha_j))\big). \tag{27}$$

Let $X = X_1, ..., X_d$ be the predictor which has the size 67088×5651. Since the predictor has a spatial structure. It is reasonable to assume that the coefficients β also have a spatial structure so that nearby locations have close contributions to the waves at the target point. This assumption is equivalent to suppose that $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. For the covariance Σ_θ,

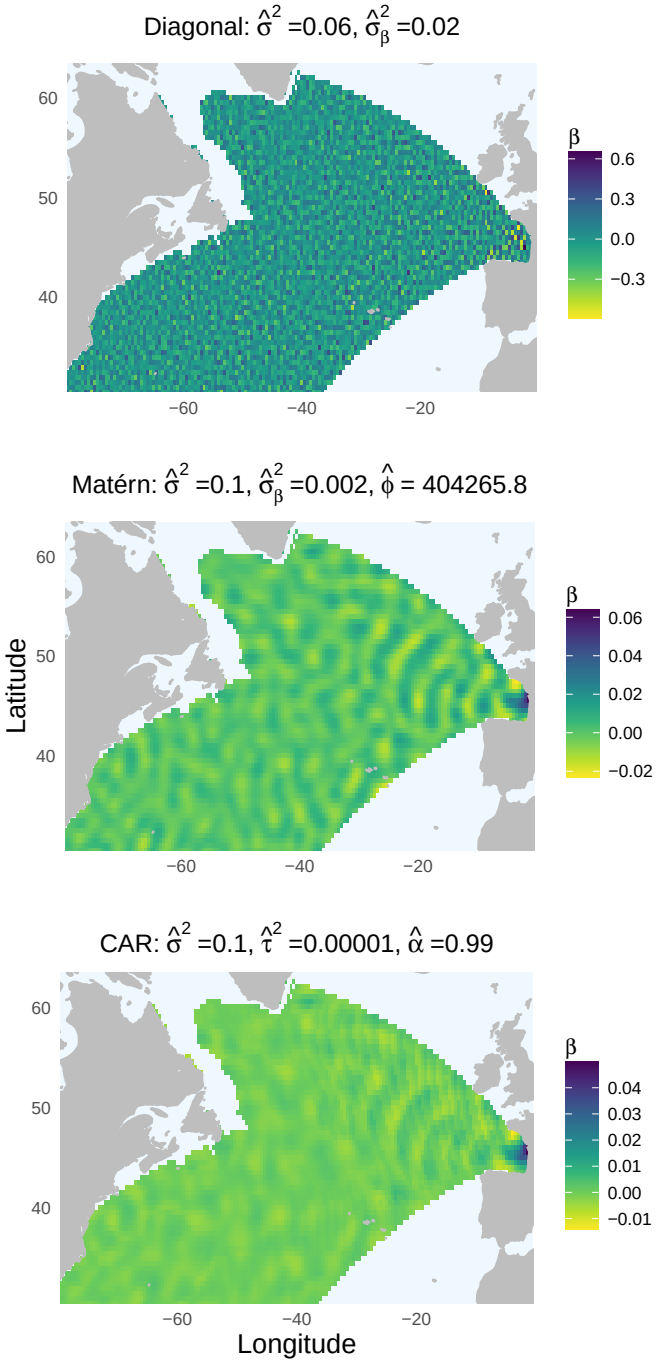

Figure 8: The coefficients β estimated using the EM algorithm with diagonal, Matérn, and CAR covariance.

Method	r	RMSE(m)	bias(m)
Diagonal	0.941	0.414	-0.0004
Matérn	0.956	0.354	-0.04
CAR	0.957	0.352	-0.06

Table 1: Quantitative comparison of the diagonal, Matérn, and CAR methods in the validation set using the correlation (r), root mean square error (RMSE), and bias.

we will consider the cases of Matérn and CAR. For comparison, we also consider the diagonal case even though it does not consider any structure between coefficients.

The model's parameters (equation 24) are estimated using data from 1994 to 2013, and the model is evaluated in terms of correlation, RMSE, and bias, using a validation set from 2014 to 2016. Figure 8 shows the results of estimating β and the covariance parameters using the EM algorithm when the covariance structure is assumed to be diagonal, Matérn and CAR. Not surprisingly, the coefficients estimated with the diagonal covariance show no physical spatial structure. Therefore, the assumption that close locations have close coefficients cannot be taken into account using the diagonal case. This motivates using the Matérn and CAR covariances. The Matérn and CAR covariances give the smoothest coefficients with a clear spatial structure. In addition, locations close to the target point have larger coefficients. Therefore, the obtained coefficients are more physically interpretable and take into account our assumption about the covariance. Note that the CAR method is less expensive numerically than the Matérn, which involves inverting the covariance matrix at each iteration of the optimization algorithm used in the M-step.

Table 1 shows the results of the quantitative comparison between the three methods for predicting significant wave height in the validation set using correlation (r), root mean square error (RMSE), and bias. In terms of correlation and RMSE, the diagonal method is the less accurate method. Therefore, adding the spatial structure in the covariance is advantageous in predicting the significant wave height. The CAR and Matérn methods lead to close results regarding r, RMSE, and bias.

5 Conclusions

This study proposed an EM algorithm for estimating generalized Ridge regression with spatial covariates. We have studied three cases: the diagonal, Matérn, and the CAR case. A simulation study is carried out to evaluate the performance of the algorithms, and the EM algorithm successfully estimates the parameters in all cases. We have studied the influence of the sample size, dimension of X, and the variance σ^2 on the estimation. The sample size mainly influences the variance parameter σ^2. The range parameter of the Matérn and correlation parameter of the CAR are mainly influenced by dimension d.

The proposed method is applied to the problem of downscaling the significant wave height in the Bay of Biscay using wind conditions over the North Atlantic. The Matérn method gives smooth coefficients with a clear spatial structure; however, the CAR method slightly outperforms the Matérn method in terms of RMSE. The Matérn covariance is clearly a better choice for spatial applications. However, estimating the parameters requires the inversion of the covariance matrix at each iteration of the optimization method in the M-step, which may be a computational bottleneck in many applications. To address this issue, instead of parameterizing the covariance matrix, one can parameterize the precision matrix directly as we did with the CAR method.

Appendices

A Comparison between cross-validation and EM

As stated in section 2, the EM algorithm can be used as an alternative for cross-validation for estimating Ridge regression. In this section, we perform a simulation study to compare the two approaches and use the same simulation procedure discussed in section 3.1. Given the same covariates X (presented in section 3.1) we perform 50 independent random samples of coefficients β using the diagonal method (with parameters $\sigma^2 = 36$ and $\sigma^2_\beta = 7$). For each simulation, we estimate the coefficients using the EM algorithm and the cross-validation method. Figure 9 shows the box plot of $NRMSE_\beta$ and $NRMSE_y$. The EM algorithm outperforms cross-validation in estimating the coefficients β and predicting y.

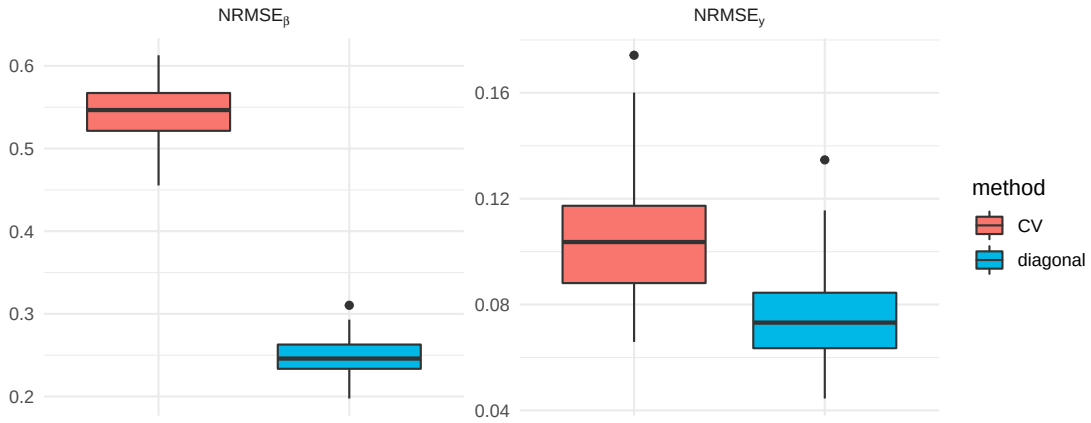

Figure 9: Results of estimating Ridge regression with the EM algorithm and 10-fold cross-validation in the Gaussian case.

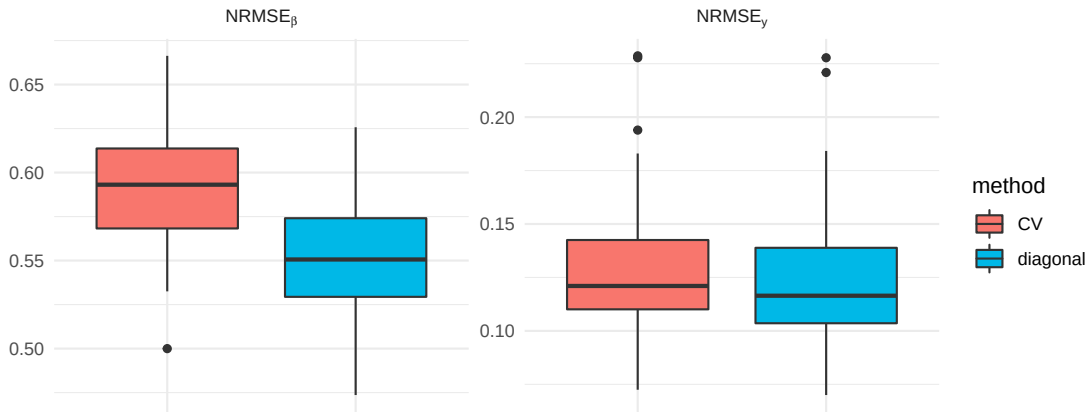

Figure 10: Results of estimating Ridge regression with the EM algorithm and 10-fold cross-validation in the non-Gaussian case.

The comparison we performed here is for the Gaussian case; therefore, it is straightforward that the EM algorithm will outperform cross-validation. To see how the two approaches behave in the non-Gaussian case, we simulate the response variable Y using the model

$$Y = X\beta + \epsilon, \quad \text{where } \epsilon \sim U(2, 30) \tag{A.1}$$

Where $U(2, 30)$ is the uniform distribution on the interval $[2, 30]$. Figure 10 shows the estimation results using the EM algorithm and cross-validation. The EM algorithm still outperforms cross-validation in both $NRMSE_\beta$ and $NRMSE_y$; however, the difference between the two methods here is small than in the Gaussian case.

B The case where β has a non-zero mean

In this section, we consider the case where β has a non-zero mean as defined by the hierarchically model

$$\begin{aligned} \beta &\sim \mathcal{N}(\mu_\xi, \Sigma_\theta) \\ Y \mid \beta, \Theta &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{B.1}$$

where $\Theta = (\sigma^2, \mu_\xi, \theta)$.

The complete log-likelihood is expressed as

$$\begin{aligned}\ln p(y,\beta;\Theta) &= \ln p(y \mid \beta;\sigma^2) + \ln p(\beta;\theta) \\ &= -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \beta^T\Sigma_\theta^{-1}\beta - 2\beta^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta\|^2\right) + C\end{aligned} \tag{B.2}$$

Where C is a constant. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

- E-step:

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}(\ln p(y,\beta;\Theta) \mid y,\Theta^{(t)}). \tag{B.3}$$

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2}X^TX)^{-1} \\ \mu_{\beta|y} = \Sigma_{\beta|y}(\Sigma_\theta^{-1}\mu_\xi + \frac{1}{\sigma^2}X^Ty). \end{cases} \tag{B.4}$$

Therefore,

$$Q(\Theta|\Theta^{(t)}) = -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y,\Theta^{(t)})) - 2\mu_{\beta|y}^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi + n\ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y,\Theta^{(t)})\right) + C \tag{B.5}$$

where

$$\begin{cases} \mathbb{E}(\beta\beta^T|y;\Theta^{(t)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2|y;\Theta^{(t)}) = \|y\|^2 - 2y^TX\mu_{\beta|y} + \mathrm{Tr}(X^TX\mathbb{E}(\beta\beta^T|y;\Theta^{(t)})) \end{cases} \tag{B.6}$$

- M-step:

The maximization step computes

$$\Theta^{(t+1)} = \arg\max_\Theta Q(\Theta|\Theta^{(t)}) \tag{B.7}$$

which leads to the following updates of the parameters

$$\begin{aligned}\sigma^{2,(t+1)} &= \frac{1}{n}(\|y\|^2 - 2y^TX\mu_{\beta|y} + \mathrm{Tr}(X^TX\mathbb{E}(\beta\beta^T|y;\Theta^{(t)}))) \\ (\xi^{(t+1)},\theta^{(t+1)}) &= \arg\max_{\xi,\theta}\ \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y,\Theta^{(t)})) + 2\mu_{\beta|y}^T\Sigma_\theta^{-1}\mu_\xi^{(t)} - \mu_{\xi^{(t)}}^T\Sigma_\theta^{-1}\mu_\xi^{(t)}\end{aligned} \tag{B.8}$$

References

Abramowitz, M., Stegun, I.A., Romer, R.H., 1988. Handbook of mathematical functions with formulas, graphs, and mathematical tables.

Allen, D.M., 1974. The relationship between variable selection and data agumentation and a method for prediction. technometrics 16, 125–127.

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D.W., O'Neil, M., 2015. Fast direct methods for gaussian processes. IEEE transactions on pattern analysis and machine intelligence 38, 252–265.

Bachoc, F., 2013. Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments. Ph.D. thesis. Université Paris-Diderot-Paris VII.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics 43, 1–20.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. volume 4. Springer.

Boonstra, P.S., Mukherjee, B., Taylor, J.M., 2015. A small-sample choice of the tuning parameter in ridge regression. Statistica Sinica 25, 1185.

Boudière, E., Maisondieu, C., Ardhuin, F., Accensi, M., Pineau-Guillou, L., Lepesqueur, J., 2013. A suitable metocean hindcast database for the design of marine energy converters. International Journal of Marine Energy 3, e40–e52.

Cressie, N., Kapat, P., 2008. Some diagnostics for markov random fields. Journal of computational and graphical statistics 17, 726–749.

Cressie, N., Wikle, C.K., 2015. Statistics for spatio-temporal data. John Wiley & Sons.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–22.

Goeman, J.J., 2008. Autocorrelated logistic ridge regression for prediction based on proteomics spectra. Statistical Applications in Genetics and Molecular Biology 7.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. volume 2. Springer.

Hemmerle, W.J., 1975. An explicit solution for generalized ridge regression. Technometrics 17, 309–314.

Kaufman, C., Shaby, B.A., 2013. The role of the range parameter for estimation and prediction in geostatistics. Biometrika 100, 473–484.

Lee, Y., Nelder, J.A., 1996. Hierarchical generalized linear models. Journal of the Royal Statistical Society: Series B (Methodological) 58, 619–656.

Obakrim, S., Ailliot, P., Monbet, V., Raillard, N., 2022. Statistical modeling of the space-time relation between wind and significant wave height .

Patil, P., Wei, Y., Rinaldo, A., Tibshirani, R., 2021. Uniform consistency of cross-validation estimators for high-dimensional ridge regression, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 3178–3186.

Rue, H., 2001. Fast sampling of gaussian markov random fields. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63, 325–338.

Rue, H., Tjelmeland, H., 2002. Fitting gaussian markov random fields to gaussian fields. Scandinavian journal of Statistics 29, 31–49.

Saha, S., Moorthi, S., Pan, H.L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al., 2010. The ncep climate forecast system reanalysis. Bulletin of the American Meteorological Society 91, 1015–1058.

Schulz, E., Speekenbrink, M., Krause, A., 2018. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. Journal of Mathematical Psychology 85, 1–16.

Storkey, A.J., 1999. Truncated covariance matrices and toeplitz methods in gaussian processes, in: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99.(Conf. Publ. No. 470), IET. pp. 55–60.

Tajbakhsh, S.D., Aybat, N.S., Del Castillo, E., 2020. On the theoretical guarantees for parameter estimation of gaussian random field models: A sparse precision matrix approach. Journal of Machine Learning Research 21, 1–41.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 91–108.

van Wieringen, W.N., 2015. Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169 .