Efficient Bayesian linear models for a large number of observations

Hassan Maatouk, Didier Rullière, Xavier Bay

Efficient Bayesian linear models for a large number of observations

Hassan Maatouk, Didier Rullière and Xavier Bay

Abstract Bayesian linear models are widely used as efficient approaches for nonparametric function estimation. In this paper, we present a Bayesian method for generating finite-dimensional linear models that can handle large datasets. This method is based on an efficient Markov chain Monte Carlo algorithm. The advantage of this approach is that sampling is performed before conditioning, rather than after. This enables the use of efficient samplers when the prior covariance matrix exhibits special properties, such as being Toeplitz, block-Toeplitz, or sparse. Numerical examples are provided to illustrate the performance of the proposed method.

1 Introduction

Gaussian processes (GPs) are commonly used as effective prior distributions over function spaces. They are frequently employed in tasks like regression and classification in machine learning [16, 19]. Over the years, GPs have gained popularity as a Bayesian tool, finding applications across diverse fields such as geosciences, physics, biology, chemistry, engineering, finance, and machine learning [1, 3, 4, 18].

In this paper, Bayesian linear models with Gaussian random weights for a large number of observations are considered. In this context, *direct* approaches based on Cholesky factorization and eigendecomposition [10] become computationally

Hassan Maatouk
LAMPS, Université de Perpignan via Domitia, 52 av. Paul Alduy, 66860 Cedex 9 Perpignan, France, e-mail: hassan.maatouk@univ-perp.fr

Didier Rullière
Mines Saint-Étienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Saint-Étienne, F-42023, France e-mail: drulliere@emse.fr

Xavier Bay
Mines Saint-Étienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Saint-Étienne, F-42023, France e-mail: bay@emse.fr

prohibitive when handling large datasets, as their computational complexity grows cubically with the number of observations. To address this problem, we propose the use of a Markov Chain Monte Carlo (MCMC) method. This approach builds upon the efficient Elliptical Slice Sampling (ESS) algorithm [15] and requires only the evaluation of the likelihood function at each MCMC iteration. The main advantage of this method is that sampling is performed before conditioning rather than after. Consequently, if the prior covariance (precision) matrix admits special structures, such as Toeplitz or block-Toeplitz (i.e., when evaluating stationary product kernels on a regularly spaced grid) [20, 21], highly efficient samplers can be employed such as the Fast Fourier Transform (FFT) [20] and the fast large-scale approaches developed in [12, 13, 14]. In the numerical examples of this paper, a comparison is conducted between Cholesky factorization and the highly efficient LS.KLE sampler developed in [13] for generating the prior.

The paper is organized as follows: in Sect. 2, GP regression is briefly reviewed. Section 3 is divided into two parts. The first part revisits *direct* approaches for generating linear models, while the second part develops a highly efficient MCMC algorithm tailored for large datasets. Section 4 is dedicated to numerical examples.

2 Gaussian process regression review

In this paper, we consider the following regression problem with Gaussian noise, where n vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ and n responses $y_i \in \mathbb{R}$ are observed from the model

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

We assume that $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, is an additive i.i.d. zero-mean Gaussian noise with a constant variance of σ^2. Here, $f : \mathcal{X} \subset \mathbb{R}^d \longrightarrow \mathbb{R}$ represents an unknown function that generates the observed target values $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$. The sequence $\{(\boldsymbol{x}_i, y_i)\}$ represents the training samples. The set $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and without loss of generality, we suppose that $\mathcal{X}$ is the unit hypercube.

A GP is a stochastic process, i.e., a collection of random variables, such that every finite subset of these variables follows a multivariate normal (MVN) distribution. A GP is fully specified by its mean function $\mu(\cdot)$ and covariance function $k(\cdot, \cdot)$, where the covariance function plays a crucial role in controlling the smoothness of the GP sample paths. If we denote this GP by Z, then we can write

$$Z \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot)),$$

where the mean and covariance functions are defined as:

$$\begin{aligned} &\mu(\boldsymbol{x}) = \mathrm{E}[Z(\boldsymbol{x})], \quad \forall \boldsymbol{x} \in \mathcal{X}; \\ &k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}\left(Z(\boldsymbol{x}), Z(\boldsymbol{x}')\right), \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}. \end{aligned}$$

Additionally, $Z(\cdot) = \mu(\cdot) + Y(\cdot)$, where Y is a zero-mean GP, i.e., $Y \sim \mathcal{GP}(0, k)$. In the regression framework (1), GPs are widely recognized as efficient prior distributions over function spaces [19]. Conditionally on data, we have

$$\{Y(\cdot)|Y(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\} \sim \mathcal{GP}\left(\widetilde{\mu}(\cdot), \widetilde{k}(\cdot,\cdot)\right),$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_n)$ and the conditional mean function $\widetilde{\mu}$ and covariance function $\widetilde{k}$ are given by

$$\begin{aligned} \widetilde{\mu}(\boldsymbol{x}) &= \mathrm{E}\left[Y(\boldsymbol{x})|\boldsymbol{y}\right] = k(\boldsymbol{x}, \mathbb{X})^\top \left(k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n\right)^{-1} \boldsymbol{y}; \\ \widetilde{k}(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \mathbb{X})^\top (k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n)^{-1} k(\boldsymbol{x}', \mathbb{X}); \end{aligned} \tag{2}$$

with I_n the $n \times n$ identity matrix, $\mathbb{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ the design matrix, $k(\mathbb{X}, \mathbb{X})$ the covariance matrix of the Gaussian vector $Y(\mathbb{X}) \in \mathbb{R}^n$ and $k(\boldsymbol{x}, \mathbb{X})$ the vector of covariance between $Y(\boldsymbol{x})$ and $Y(\mathbb{X})$, i.e., $k(\boldsymbol{x}, \mathbb{X}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top$.

3 Finite-dimensional Bayesian linear models

In this section we assume that the parent GP $\{Y(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{X}}$ is approximated by the following Bayesian linear model with Gaussian random weights, called the *weight-space* view of GPs [19, Sect. 2.1.1]

$$\breve{Y}(\boldsymbol{x}) = \Phi(\boldsymbol{x})^\top \boldsymbol{\eta}, \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where $\boldsymbol{\eta} \in \mathbb{R}^m$ is a zero-mean Gaussian vector with positive-definite covariance matrix $\tau^2 \boldsymbol{K}$, with τ^2 the *signal variance* parameter, and Φ is a sequence of deterministic basis functions such that $\Phi(\boldsymbol{x}) \in \mathbb{R}^m$, for any $\boldsymbol{x} \in \mathcal{X}$. For simplicity of notations, we denote by $\boldsymbol{\Phi} = \Phi(\mathbb{X}) \in \mathbb{R}^{n \times m}$, where $\mathbb{X}$ is the design matrix. Therefore, the set of noisy data $\{\breve{Y}(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ can be written in matrix form as follows:

$$\boldsymbol{\Phi\eta} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is an independent zero-mean Gaussian noise vector, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathrm{I}_n)$.

3.1 Direct sampling approaches

In this section, we briefly review the *direct* classical approaches for generating the posterior distribution $\{\boldsymbol{\eta}|\boldsymbol{\Phi\eta} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$. On the one hand, according to [19, Sect. 2.1.1] and Bayes's rule, we have the following posterior distribution:

$$p(\boldsymbol{\eta}|\boldsymbol{\Phi},\boldsymbol{y}) := \frac{p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\boldsymbol{y}|\boldsymbol{\Phi})}, \tag{3}$$

where $p(\boldsymbol{y}|\boldsymbol{\Phi})$ is the normalizing constant, also known as the *marginal likelihood*. It is independent of $\boldsymbol{\eta}$ and given by

$$p(\boldsymbol{y}|\boldsymbol{\Phi}) := \int_{\mathbb{R}^m} p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\eta})p(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}.$$

By developing the likelihood and prior in (3), we obtain

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{\Phi},\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\eta}]^\top[\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\eta}]\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\eta}^\top\boldsymbol{K}^{-1}\boldsymbol{\eta}\right) \\ &\propto \exp\left(-\frac{1}{2}[\boldsymbol{\eta}-\breve{\boldsymbol{\mu}}]^\top\left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}/\sigma^2+\boldsymbol{K}^{-1}/\tau^2\right][\boldsymbol{\eta}-\breve{\boldsymbol{\mu}}]\right). \end{aligned} \tag{4}$$

Consequently, the posterior distribution $\{\boldsymbol{\eta}|\boldsymbol{\Phi}\boldsymbol{\eta}+\boldsymbol{\epsilon}=\boldsymbol{y}\} \sim \mathcal{N}(\breve{\boldsymbol{\mu}},\breve{\boldsymbol{K}})$, where

$$\begin{cases} \breve{\boldsymbol{\mu}} = \breve{\boldsymbol{K}}\boldsymbol{\Phi}^\top\boldsymbol{y}/\sigma^2; \\ \breve{\boldsymbol{K}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}/\sigma^2+\boldsymbol{K}^{-1}/\tau^2\right)^{-1}. \end{cases} \tag{5}$$

On the other hand, based on the predictive equations in (2), we have $\{\boldsymbol{\eta}|\boldsymbol{\Phi}\boldsymbol{\eta}+\boldsymbol{\epsilon}=\boldsymbol{y}\} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{K}})$, where [10, Proposition 1]

$$\begin{cases} \widetilde{\boldsymbol{\mu}} &= \tau^2(\boldsymbol{\Phi}\boldsymbol{K})^\top(\tau^2\boldsymbol{\Phi}\boldsymbol{K}\boldsymbol{\Phi}^\top+\sigma^2\mathrm{I}_n)^{-1}\boldsymbol{y}; \\ \widetilde{\boldsymbol{K}} &= \tau^2\boldsymbol{K}-\tau^4(\boldsymbol{\Phi}\boldsymbol{K})^\top(\tau^2\boldsymbol{\Phi}\boldsymbol{K}\boldsymbol{\Phi}^\top+\sigma^2\mathrm{I}_n)^{-1}\boldsymbol{\Phi}\boldsymbol{K}. \end{cases} \tag{6}$$

Comments on the two *direct* approaches in (5) **and** (6)

The two approaches in Equations (5) and (6) are equivalent. However, their computational complexities are different. In Equation (6), the matrix inversion depends on the number of training samples n, while in the approach in (5), it depends on the dimension m of the Gaussian vector $\boldsymbol{\eta}$. As a result, for a fixed m, the approach in (5) is more efficient for large datasets n. It is worth noting that the two approaches (5) and (6) are based on sampling after conditioning rather than before. Therefore, if the prior covariance matrix $\boldsymbol{K}$ exhibits a particular structure such as Toeplitz or block-Toeplitz [20, 21] or banded and sparse [6], this property will be lost in the sampling step. Finally, let us mention that generating a Gaussian vector can be achieved using Cholesky factorization or eigendecomposition [10], with the computational complexity growing cubically as a function of the dimension m [7].

In the following section, we present an alternative approach for generating samples from the posterior distribution $\{\boldsymbol{\eta}|\boldsymbol{\Phi}\boldsymbol{\eta}+\boldsymbol{\epsilon}=\boldsymbol{y}\}$, where sampling is performed before conditioning rather than after. Consequently, this approach is well-suited for

handling high-dimensional spaces (when the prior covariance matrix $\boldsymbol{K}$ exhibits special properties) and large datasets.

3.2 Alternative sampling approaches

Due to its computational complexity, which grows cubically, *direct* approaches become infeasible for generating the posterior distribution $\{\boldsymbol{\eta}|\boldsymbol{\Phi}\boldsymbol{\eta}+\boldsymbol{\epsilon}=\boldsymbol{y}\}$ when m is significant large (i.e., $m \gg 1,000$), regardless of whether the dataset is small or large. In this section, we present a different approach to address this issue. Precisely, we explain how MCMC approaches can handle the problem of generating the posterior distribution $\{\boldsymbol{\eta}|\boldsymbol{\Phi},\boldsymbol{y}\}$ when both m and n are large. The posterior probability density function (pdf) in (4) is proportional to the product of a likelihood function and a zero-mean Gaussian prior:

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{\Phi},\boldsymbol{y}) &\propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\eta}]^\top[\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\eta}]\right)}_{\text{likelihood}} \underbrace{\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\eta}^\top\boldsymbol{K}^{-1}\boldsymbol{\eta}\right)}_{\text{Gaussian prior}} \\ &= L(\boldsymbol{\eta})\mathcal{N}(\boldsymbol{\eta};\boldsymbol{0},\tau^2\boldsymbol{K}). \end{aligned}$$

The logarithm of the likelihood function $L(\boldsymbol{\eta})$ can be expressed as follows:

$$\log[L(\boldsymbol{\eta})] = -\frac{1}{2\sigma^2}\|y-\boldsymbol{\Phi}\boldsymbol{\eta}\|^2. \tag{7}$$

The logarithm function in (7), which has a computational complexity of order $O(nN)$ will be evaluated at each MCMC iteration. In this context, sampling from (4) can be performed using Metropolis-Hastings (MH) proposals [5, 16]:

$$\boldsymbol{\eta}' = \rho\boldsymbol{\zeta} + \sqrt{1-\rho^2}\boldsymbol{\eta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0},\tau^2\boldsymbol{K}), \tag{8}$$

where $\rho \in [-1,1]$ is a step-size parameter, $\boldsymbol{\eta}$ is the current state, and $\boldsymbol{\eta}'$ is the proposal state. Recall that the MH acceptance ratio, $\alpha = \min\{1, L(\boldsymbol{\eta}')/L(\boldsymbol{\eta})\}$ depends solely on the likelihood ratio and is independent of ρ. Furthermore, this method is straightforward to implement and can be readily applied to a wide range of models with Gaussian priors.

The ESS relies on the parametrization $\rho = \sin(\theta)$ in (8), offering an adaptive and automated approach to tuning the step-size parameter ρ, which ensures acceptance at every step. As a result, the MH proposal in (8) is reformulated as follows:

$$\boldsymbol{\eta}' = \sin(\theta)\boldsymbol{\zeta} + \cos(\theta)\boldsymbol{\eta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{0},\tau^2\boldsymbol{K}),$$

where the angle θ is uniformly generated from a $[\theta_{\min},\theta_{\max}]$ interval which is shrunk exponentially fast until an acceptable state is reached. For a given value of θ, a uniform

random number is generated and compared with the likelihood ratio $L(\boldsymbol{\eta}')/L(\boldsymbol{\eta})$. If the proposal $\boldsymbol{\eta}'$ is rejected, one shrinks the bracket of θ, and continues this process until acceptance. Detailed guidelines for shrinking the bracket are provided in [15].

Comments on the MCMC approach

Unlike *direct* methods, the key advantage of the MCMC approach described in this section is its ability to avoid matrix inversion. Moreover, the sampling process is carried out prior to conditioning, rather than after. Consequently, highly efficient samplers can be employed when the prior covariance matrix $\boldsymbol{K}$ exhibits special structures, such as Toeplitz or block-Toeplitz [20]. In Sect. 4, we investigate the performance of the developed approach denoted linGP-ESS in terms of computational running time when using a highly efficient sampler for the prior. Furthermore, the MCMC method introduced in this section is capable of addressing more complex posterior inference (complex likelihood function), such as when additional shape constraints are required [11]. However, this sampling method remains an approximation of the posterior distribution. Moreover, it requires evaluating the log-likelihood function (7) at each MCMC iteration.

4 Numerical performance

In this section, we investigate the performance of the developed approach for large datasets. To this end, we consider the finite-dimensional Bayesian linear model with Gaussian random weights proposed by [9]. This approach has gained significant attention for its ability to incorporate a wide range of shape constraints across the entire domain, as well as for its strong theoretical foundation [2, 8]. For simplicity, we recall this approach in the one-dimensional case. Let $\{t_j\}$ denote a sequence of $m \geqslant 2$ equally spaced knots on $\mathcal{X}$, i.e., $0 = t_1 < \ldots < t_m = 1$. Then,

$$\check{Y}(x) := \sum_{j=1}^{m} \eta_j \phi_j(x) = \Phi(x)^\top \boldsymbol{\eta}, \quad x \in \mathcal{X}, \tag{9}$$

where $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_m]^\top$ is a zero-mean Gaussian vector with positive-definite covariance matrix $\tau^2 \boldsymbol{K} \in \mathbb{R}^{m \times m}$, and $\Phi(\cdot) = [\phi_1(\cdot), \ldots, \phi_m(\cdot)]^\top \in \mathbb{R}^m$, with ϕ_j the compactly supported basis function associated to the knot t_j. Thus, we have for any $j \in \{2, \ldots, m-1\}$

$$\phi_j(x) := \begin{cases} 1 - \frac{|x - t_j|}{t_{j+1} - t_j} & \text{if } x \in [t_{j-1}, t_{j+1}]; \\ 0 & \text{otherwise.} \end{cases}$$

Additionally, we define $\phi_1(x) := 1 - (m-1)|x|$, if $x \in [t_1, t_2]$, and zero otherwise. Similarly, for $j = m$, we define $\phi_m(x) := 1 - (m-1)|x - 1|$, if $x \in [t_{m-1}, t_m]$,

and zero otherwise. It is worth noting that, in this case, if a stationary covariance kernel, such as, the Matérn family of covariance functions, is used, then the prior covariance matrix $\boldsymbol{K}$ exhibits a Toeplitz structure [20, 21]. This property is explored in the numerical examples presented in this section.

Now, we consider the statistical problem of recovering an unknown function $f : \mathcal{X} \longrightarrow \mathbb{R}$ from the training samples $\{(x_i, y_i)\}_{i=1}^n$. The following target function proposed by [17] is considered

$$f(x) = \frac{1}{[1 + (10x)^4]} + \frac{1}{2} \exp\left\{-100(x - 0.5)^2\right\}, \quad x \in \mathcal{X}. \tag{10}$$

This function is known to be nonnegative on $\mathcal{X}$, but this constraint is not incorporated into the model developed in this section, as it falls outside the scope of the paper. This function is particularly relevant in our case because it exhibits significant variation. Consequently, a substantial number of observations are required to achieve a *satisfactory* approximation. The training samples are obtained as follows: the covariates $\{x_i\}$ are sampled uniformly over $\mathcal{X}$, and the observed values $\{y_i\}$ were obtained using model (1) with the target function f defined in (10) and a standard deviation of $\sigma = 0.1$. The positive-definite covariance matrix $\tau^2\boldsymbol{K}$ of the Gaussian vector $\boldsymbol{\eta}$ is obtained using the stationary Matérn family of covariance functions

$$k(x, x') = \tau^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right)^{\nu} B_{\nu} \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right), \tag{11}$$

for any $x, x' \in \mathcal{X}$, where $\Gamma(\cdot)$ is the Gamma function and $B_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind of order ν. It is worth noting that a process with the Matérn kernel of order ν admits sample paths that are $\lceil \nu - 1 \rceil$ times differentiable [19, Sect. 4.2.1].

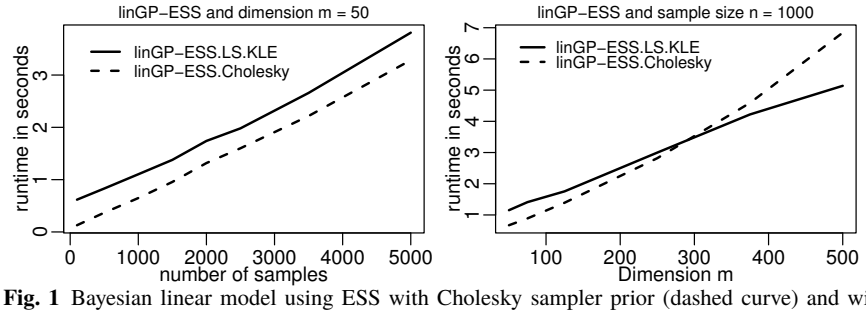

Fig. 1 Bayesian linear model using ESS with Cholesky sampler prior (dashed curve) and with LS.KLE sampler prior (solid curve). The computational running time of generating 1,000 posterior sample paths as a function of the number of samples (left) and of the dimension m (right)

In Fig. 1, the Matérn covariance function with a smoothness parameter $\nu = 5/2$ is employed (11), where the length-scale parameter ℓ is chosen such that the correlation at the maximum possible separation between the covariates equals 0.05. We illustrate

the computational runtime required to generate 1,000 posterior sample paths as a function of the number of samples (left) and the dimensionality (right). The dashed curve represents the runtime in seconds for the developed approach when using the Cholesky sampler to generate the prior, while the solid curve represents the runtime when using the LS.KLE sampler [13]. As expected, the runtime of the developed approach increases linearly with both the number of samples, n, and the dimension, m. Moreover, the Cholesky factorization outperforms the LS.KLE sampler for low values of m, whereas the LS.KLE sampler is more efficient for high values of m.

5 Conclusion

In this paper, Bayesian linear regression models for a large number of observations are considered. Direct approaches based on Cholesky factorization and eigendecomposition become computationally prohibitive when applied to large datasets. To address this issue, we present an efficient Markov Chain Monte Carlo (MCMC) algorithm based on Elliptical Slice Sampling (ESS), which requires only the evaluation of the likelihood function at each MCMC iteration. The main advantage of this approach is that sampling is performed before conditioning rather than after. This enables the use of highly efficient methods in the sampling procedure when the prior covariance matrix exhibits special structures, such as Toeplitz, banded and sparse. The effectiveness of this approach is demonstrated using synthetic data in the context of nonparameteric function estimation.

References

1. S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.
2. X. Bay, L. Grammont, and H. Maatouk. Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electron. J. Stat.*, 10(1):1580–1595, 2016.
3. C. Blanchet-Scalliet, B. Demory, T. Gonon, and C. Helbert. Gaussian process regression on nested spaces. *SIAM/ASA JUQ*, 11(2):426–451, 2023.
4. C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.
5. S.L. Cotter, G.O. Roberts, A.M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Stat. Sci.*, 28(3):424–446, 2013.
6. N. Durrande, V. Adam, L. Bordeaux, S. Eleftheriadis, and J. Hensman. Banded matrix operators for Gaussian Markov models in the automatic differentiation era. In *The 22nd International Conference on AISTATS*, volume 89, pages 2780–2789. PMLR, 2019.
7. G. Golub and C.F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, 1996.
8. L. Grammont, H. Maatouk, and X. Bay. Equivalence between constrained optimal smoothing and Bayesian estimation. *J. Nonparametric Stat.*, 0(0):1–22, 2024.
9. H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. *Math. Geosci.*, 49(5):557–582, 2017.

10. H. Maatouk, X. Bay, and D. Rullière. A note on simulating hyperplane-truncated multivariate normal distributions. *Stat. Probabil. Lett.*, 191:109650, 2022.
11. H. Maatouk, D. Rullière, and X. Bay. Efficient constrained Gaussian process approximation using elliptical slice sampling. working paper or preprint, March 2024.
12. H. Maatouk, D. Rullière, and X. Bay. Large scale Gaussian processes with Matheron's update rule and Karhunen-Loève expansion. In Aicke Hinrichs, Peter Kritzer, and Friedrich Pillichshammer, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 469–487, Cham, 2024. Springer International Publishing.
13. H. Maatouk, D. Rullière, and X. Bay. Sampling large hyperplane-truncated multivariate normal distributions. *Comput. Stat.*, 39(4):1779–1806, 2024.
14. H. Maatouk, D. Rullière, and X. Bay. Large-scale constrained Gaussian processes for shape-restricted function estimation. *Stat. Comput.*, 35(7), 2025.
15. I. Murray, R. Adams, and D. MacKay. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on AISTATS*, pages 541–548, 2010.
16. R.M. Neal. Regression and classification using Gaussian process priors. *Bayesian Statistics*, 6:475–501, 1999.
17. A. Pensoneault, X. Yang, and X. Zhu. Nonnegativity-enforced Gaussian process regression. *TAML*, 10(3):182–187, 2020.
18. S. Petit. *Improved Gaussian process modeling: Application to Bayesian optimization*. PhD thesis, Université Paris-Saclay, 2022.
19. C.K. Williams and C.E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
20. A.T.A. Wood and G. Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Stat.*, 3(4):409–432, 1994.
21. D.L. Zimmerman. Computationally exploitable structure of covariance matrices and generalized convariance matrices in spatial models. *J. Stat. Comput. Simul.*, 32(1-2):1–15, 1989.