Bayesian linear models for large datasets: Markov chain Monte Carlo or Matheron's update rule

Hassan Maatouk[*]

(*) LAMPS, Université de Perpignan Via Domitia, 52 av. Paul Alduy,
66860 Cedex 9 Perpignan, France

hassan.maatouk@univ-perp.fr

JS2O 2-4 avril 2025

Université Perpignan Via Domitia
CRÉATRICE D'AVENIRS DEPUIS 1350

LAMPS

Outline I

1. Gaussian process regression review
   - Gaussian processes
   - Gaussian process regression
2. Bayesian linear models with large datasets
   - Introduction
   - First approach : Markov chain Monte Carlo (MCMC)
   - Second approach : Matheron's update rule (MUR)
3. Performance illustrations
   - Illustrative examples 1D
   - Real-world application : Diamond data
   - Illustrative examples 2D
4. Conclusion

Table of Contents I

Table of Contents I

Gaussian processes

The statistical problem of recovering an unknown function f from the training samples $\{(\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ with Gaussian noise is considered :

$$y_i = f(\boldsymbol{x}_i) + \underbrace{\epsilon_i}_{\text{noise}}, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n. \tag{1}$$

- f is an unknown function representing a physical phenomenon :

$$f \ : \ \mathcal{X} \subset \mathbb{R}^d \longrightarrow \mathbb{R}.$$

- The data $\boldsymbol{y} = f(\mathbb{X}) + \boldsymbol{\epsilon} \in \mathbb{R}^n$, with $\mathbb{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top \in \mathbb{R}^n$.

Gaussian processes

The statistical problem of recovering an unknown function f from the training samples $\{(\boldsymbol{x}_i, y_i\}_{i=1}^n$ with Gaussian noise is considered :

$$y_i = f(\boldsymbol{x}_i) + \underbrace{\epsilon_i}_{\text{noise}}, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n. \tag{1}$$

- f is an unknown function representing a physical phenomenon :
$$f\ :\ \mathcal{X} \subset \mathbb{R}^d \longrightarrow \mathbb{R}.$$
- The data $\boldsymbol{y} = f(\mathbb{X}) + \boldsymbol{\epsilon} \in \mathbb{R}^n$, with $\mathbb{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top \in \mathbb{R}^{n \times d}$ and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top \in \mathbb{R}^n$.

Definition (Gaussian processes)

Let $(Y(\boldsymbol{x}))_{\boldsymbol{x} \in \mathcal{X}}$ be a stochastic process on $\mathcal{X}$ such that $\mathrm{E}[Y(\boldsymbol{x})^2] < +\infty$. Then

- *mean function* : $\mu(\boldsymbol{x}) = \mathrm{E}[Y(\boldsymbol{x})]$;
- *covariance function* : $k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}'))$.

*A **Gaussian process (GP)** Y, i.e., $Y \sim \mathcal{GP}(\mu, k)$, is a stochastic process s.t.*

$$[Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n)]^\top \sim \overbrace{\mathcal{N}(\underbrace{\boldsymbol{\mu}}_{\textit{vector}}, \underbrace{\boldsymbol{K}}_{\textit{matrix}})}^{\textit{Gaussian vector}}.$$

Simulation of Gaussian processes

TABLE: Some popular covariance functions $k(x, x')$ used in Machine Learning.

Name	Expression	Class
Squared exponential (SE)	$\exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$	$\mathcal{C}^\infty$
Matérn $\nu = 5/2$	$\left(1 + \frac{\sqrt{5}\lvert x-x'\rvert}{\ell} + \frac{5(x-x')^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}\lvert x-x'\rvert}{\ell}\right)$	$\mathcal{C}^2$
Matérn $\nu = 3/2$	$\left(1 + \frac{\sqrt{3}\lvert x-x'\rvert}{\ell}\right)\exp\left(-\frac{\sqrt{3}\lvert x-x'\rvert}{\ell}\right)$	$\mathcal{C}^1$
Exponential	$\exp\left(-\frac{\lvert x-x'\rvert}{\ell}\right)$	$\mathcal{C}^0$

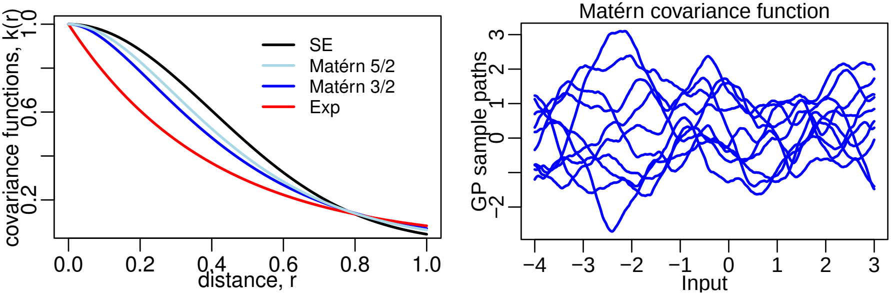

FIGURE: Some popular covariance functions (left) and GP sample paths using the Matérn $\nu = 3/2$ kernel (right).

Simulation of Gaussian processes

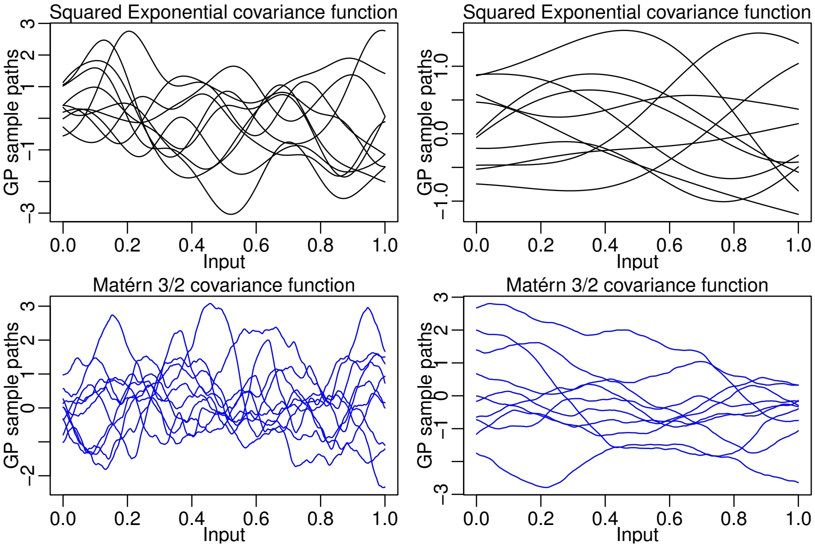

FIGURE: Ten zero-mean GP sample paths using the SE covariance function (top) and the Matérn covariance (bottom). The length-scale parameter ℓ if fixed at 0.1 (left) and at 0.4 (right).

Table of Contents I

Gaussian process regression (GPR)

Gaussian process regression

- GPR is based on assuming a GP prior on the underlying function f.
- If $Y \sim \mathcal{GP}(0, k)$, then

$$\{Y|\boldsymbol{y}\} \sim \mathcal{GP}(\check{\mu}, \check{k}),$$

where the conditional mean $\check{\mu}$ and covariance function $\check{k}$ are given as follows :

$$\text{prediction} \rightarrow \quad \check{\mu}(\boldsymbol{x}) = \mathrm{E}\left[Y(\boldsymbol{x})|\boldsymbol{y}\right] = k(\boldsymbol{x}, \mathbb{X})^\top \left(k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n\right)^{-1} \boldsymbol{y};$$

$$\text{CI} \rightarrow \quad \check{k}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \mathbb{X})^\top (k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n)^{-1} k(\boldsymbol{x}', \mathbb{X}).$$

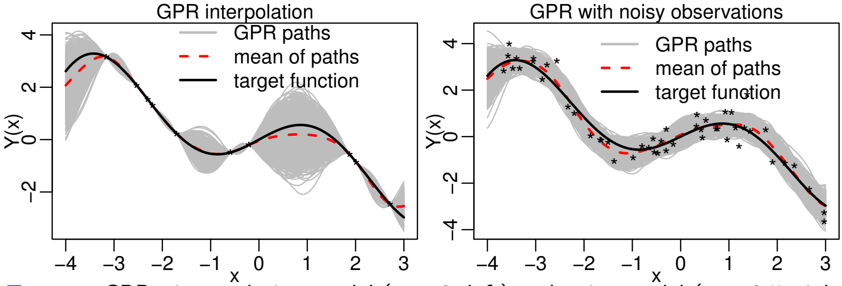

FIGURE: GPR : interpolation model ($\sigma = 0$, left) and noisy model ($\sigma = 0.5$, right).

Gaussian process regression (GPR)

Gaussian process regression

- GPR is based on assuming a GP prior on the underlying function f.
- If $Y \sim \mathcal{GP}(0, k)$, then

$$\{Y|\boldsymbol{y}\} \sim \mathcal{GP}(\check{\mu}, \check{k}),$$

where the conditional mean $\check{\mu}$ and covariance function $\check{k}$ are given as follows :

prediction $\rightarrow$ $$\check{\mu}(\boldsymbol{x}) = \mathrm{E}\left[Y(\boldsymbol{x})|\boldsymbol{y}\right] = k(\boldsymbol{x}, \mathbb{X})^\top \left(k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n\right)^{-1} \boldsymbol{y};$$

CI $\rightarrow$ $$\check{k}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \mathbb{X})^\top (k(\mathbb{X}, \mathbb{X}) + \sigma^2 \mathrm{I}_n)^{-1} k(\boldsymbol{x}', \mathbb{X}).$$

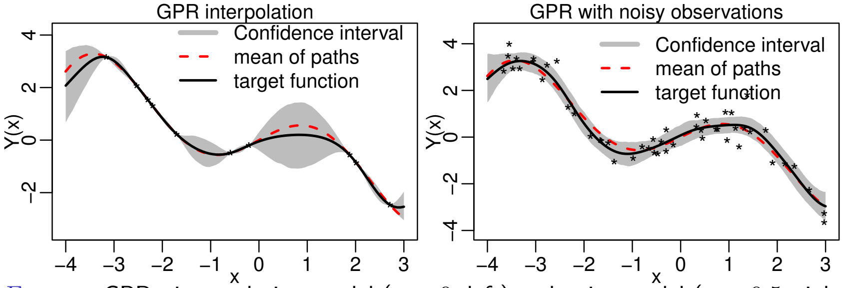

FIGURE: GPR : interpolation model ($\sigma = 0$, left) and noisy model ($\sigma = 0.5$, right).

Table of Contents I

Table of Contents I

Bayesian linear models with large datasets

Definition (Finite-dimensional Bayesian linear models)

$(Y(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$ is approximated by a finite-dimensional Bayesian linear model :

$$Y(\boldsymbol{x}) \approx \sum_{j=1}^{N} \xi_j \phi_j(\boldsymbol{x}) = \phi(\boldsymbol{x})\boldsymbol{\xi} := Y^N(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{K})$ with $\tau^2 \boldsymbol{K}$ a positive-definite covariance matrix and $\phi(\cdot)$ is a deterministic function. Examples : KLE, Bernstein polynomials, B-splines, Hermite polynomials.

Bayesian linear models with large datasets

Definition (Finite-dimensional Bayesian linear models)

$(Y(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$ *is approximated by a finite-dimensional Bayesian linear model :*

$$Y(\boldsymbol{x}) \approx \sum_{j=1}^{N} \xi_j \phi_j(\boldsymbol{x}) = \phi(\boldsymbol{x})\boldsymbol{\xi} := Y^N(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \tau^2\boldsymbol{K})$ with $\tau^2\boldsymbol{K}$ a positive-definite covariance matrix and $\phi(\cdot)$ is a deterministic function. Examples : KLE, Bernstein polynomials, B-splines, Hermite polynomials.

Bayesian linear models with large datasets

- **Data** : The dataset $\{Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ can be written in matrix form as follows :

$$\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

  where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is a zero-mean Gaussian noise vector with covariance matrix $\sigma^2\mathbf{I}_n$, and $\boldsymbol{X} := \phi(\mathbb{X}) \in \mathbb{R}^{n\times N}$.

- **Posterior distribution** : Conditionally on the observations $\boldsymbol{y}$

$$\{\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{y}\} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{K}}\right), \quad \text{where,} \tag{3}$$

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

Bayesian linear models with large datasets

Definition (Finite-dimensional Bayesian linear models)

$(Y(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$ *is approximated by a finite-dimensional Bayesian linear model :*

$$Y(\boldsymbol{x}) \approx \sum_{j=1}^{N} \xi_j \phi_j(\boldsymbol{x}) = \phi(\boldsymbol{x})\boldsymbol{\xi} := Y^N(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{K})$ with $\tau^2\boldsymbol{K}$ a positive-definite covariance matrix and $\phi(\cdot)$ is a deterministic function. Examples : KLE, Bernstein polynomials, B-splines, Hermite polynomials.

Bayesian linear models with large datasets

- **Data** : The dataset $\{Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ can be written in matrix form as follows :

$$\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

  where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is a zero-mean Gaussian noise vector with covariance matrix $\sigma^2\mathbf{I}_n$, and $\boldsymbol{X} := \phi(\mathbb{X}) \in \mathbb{R}^{n\times N}$.
- **Posterior distribution** : Conditionally on the observations $\boldsymbol{y}$

$$\{\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{y}\} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{K}}\right), \quad \text{where,} \tag{3}$$

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

- The predictive equations in (4) require a matrix inversion of dimension $n \times n$, making this approach infeasible for a large number of observations n.

Table of Contents I

Markov chain Monte Carlo (MCMC)

- According to [Williams and Rasmussen, 2006, Sect. 2.1.1] and Bayes' rule,

$$p(\boldsymbol{\xi}|\boldsymbol{X},\boldsymbol{y}) := \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\xi})p(\boldsymbol{\xi})}{p(\boldsymbol{y}|\boldsymbol{X})}, \tag{5}$$

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

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int_{\mathbb{R}^N} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\xi})p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}.$$

- By developing the likelihood $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{y})$ and the prior $p(\boldsymbol{\xi})$ in (5), we obtain

$$\begin{aligned} p(\boldsymbol{\xi}|\boldsymbol{X},\boldsymbol{y}) &\propto \exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}]^\top[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}]\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top\boldsymbol{K}^{-1}\boldsymbol{\xi}\right) \\ &\propto \exp\left(-\frac{1}{2}[\boldsymbol{\xi}-\boldsymbol{\mu}]^\top\boldsymbol{\Sigma}^{-1}[\boldsymbol{\xi}-\boldsymbol{\mu}]\right), \end{aligned} \tag{6}$$

  where

$$\begin{cases} \boldsymbol{\mu} = \left[\boldsymbol{X}^\top\boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2\right]^{-1}\boldsymbol{X}^\top\boldsymbol{y}/\sigma^2; \\ \boldsymbol{\Sigma} = \left[\boldsymbol{X}^\top\boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2\right]^{-1}. \end{cases} \tag{7}$$

Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo

The posterior pdf in (6) is proportional to the product of a likelihood function and a zero-mean Gaussian :

$$
\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{X},\boldsymbol{y}) &\propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}]^\top[\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}]\right)}_{L(\boldsymbol{\xi}):\ \text{likelihood function}}\underbrace{\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top\boldsymbol{K}^{-1}\boldsymbol{\xi}\right)}_{\text{Gaussian prior}} \\
&\propto L(\boldsymbol{\xi})\mathcal{N}(\boldsymbol{\xi};\mathbf{0},\tau^2\boldsymbol{K}). \qquad (8)
\end{aligned}
$$

The logarithm function in (8), which has a computational complexity of order $\mathcal{O}(nN)$ will be evaluated at each MCMC iteration.

- Sampling : Metropolis-Hastings (MH) proposals [Neal, 1999] :

$$\boldsymbol{\xi}' = \rho\boldsymbol{\nu} + \sqrt{1-\rho^2}\boldsymbol{\xi}, \quad \boldsymbol{\nu}\sim\mathcal{N}(\mathbf{0},\tau^2\boldsymbol{K}), \qquad (9)$$

  where $\rho\in[-1,1]$ is a step-size parameter, $\boldsymbol{\xi}$ is the current state, and $\boldsymbol{\xi}'$ is the proposal state.

Recall that the MH acceptance ratio, $\alpha = \min\left\{1, L(\boldsymbol{\xi}')/L(\boldsymbol{\xi})\right\}$ depends solely on the likelihood ratio and is independent of ρ.

Table of Contents I

Matheron's update rule

Proposition (Matheron's update rule (MUR))

Let $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \tau^2 \boldsymbol{K})$*. Suppose that* $\boldsymbol{X} \in \mathbb{R}^{n \times N}$ *is a given matrix of rank* n*, and* $\boldsymbol{y} \in \mathbb{R}^n$ *is an output vector representing the data i.e.,* $\boldsymbol{X}\boldsymbol{\xi} = \boldsymbol{y}$*. Then*

$$\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \underbrace{\boldsymbol{\xi}}_{\text{prior}} + \underbrace{(\boldsymbol{X}\boldsymbol{K})^\top \left(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi})}_{\text{update}}. \tag{10}$$

Additionally, we have

$$\phi(\cdot)\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \phi(\cdot) \left[\boldsymbol{\xi} + (\boldsymbol{X}\boldsymbol{K})^\top \left(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi})\right],$$

where $\phi(\cdot)$ *is the basis vector appearing in the Bayesian linear model.*

Matheron's update rule

Proposition (Matheron's update rule (MUR))

Let $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \tau^2 \boldsymbol{K})$. *Suppose that* $\boldsymbol{X} \in \mathbb{R}^{n \times N}$ *is a given matrix of rank* n, *and* $\boldsymbol{y} \in \mathbb{R}^n$ *is an output vector representing the data i.e.,* $\boldsymbol{X}\boldsymbol{\xi} = \boldsymbol{y}$. *Then*

$$\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \underbrace{\boldsymbol{\xi}}_{\text{prior}} + \underbrace{(\boldsymbol{X}\boldsymbol{K})^\top \left(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi})}_{\text{update}}. \tag{10}$$

Additionally, we have

$$\phi(\cdot)\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \phi(\cdot) \left[\boldsymbol{\xi} + (\boldsymbol{X}\boldsymbol{K})^\top \left(\boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi})\right],$$

where $\phi(\cdot)$ *is the basis vector appearing in the Bayesian linear model.*

MUR with Noisy observations

More generally, if the data are observed with independent Gaussian noise $\{\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then

$$\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \boldsymbol{\xi} + \tau^2 (\boldsymbol{X}\boldsymbol{K})^\top \underbrace{\left(\tau^2 \boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top + \sigma^2 \mathbf{I}_n\right)}_{n \times n \text{ matrix}}{}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi} - \boldsymbol{\epsilon}). \tag{11}$$

Matheron's update rule

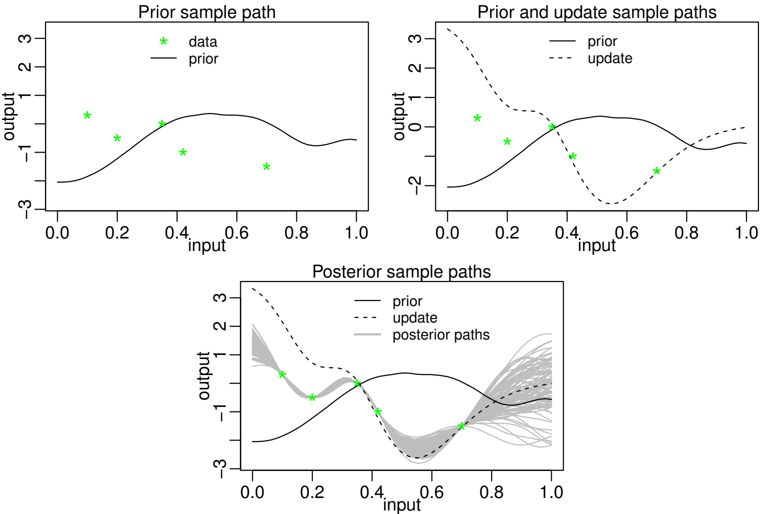

Figure: Visual representation of the MUR (*noise-free* case). **Top left** : a single path of the prior together with the data (green stars). **Top right** : the corresponding update sample path, derived from (10) is displayed as black dashed curve. **Bottom** : posterior paths (gray solid curves) are obtained by combining the prior and the update as per (10).

Matheron's update rule for large datasets

Proposition (MUR for large datasets [Maatouk et al., 2025])

Under the same settings as Proposition 1, we have

$$\{\boldsymbol{\xi} | \boldsymbol{X}, \boldsymbol{y}\} \stackrel{d}{=} \underbrace{\boldsymbol{\xi}}_{\text{prior}} + \underbrace{(\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi} - \boldsymbol{\epsilon})/\sigma^2}_{\text{update}},$$

where $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi} - \boldsymbol{\epsilon}$ *represents the residual and* σ^2 *is the variance of the noise.*

Comments

- The update part requires a matrix inversion of dimension $N \times N$ instead of $n \times n$.
- As for the MCMC approaches, sampling is performed before conditioning rather than after.
- Unlike MCMC approaches, we do not need to evaluate a likelihood function at each iteration.

Matheron's update rule for large datasets

Before proving Proposition 2, let us present the following Lemma.

Lemma

Consider three random vectors $\boldsymbol{V}_1 \in \mathbb{R}^N$, $\boldsymbol{V}_2 \in \mathbb{R}^n$ *and* $\boldsymbol{V}_3 \in \mathbb{R}^N$ *s.t.*

$$\boldsymbol{V}_1 \overset{d}{=} f(\boldsymbol{V}_2) + \boldsymbol{V}_3,$$

where f *is a measurable function of* $\boldsymbol{V}_2$ *and where* $\boldsymbol{V}_2$ *is independent of* $\boldsymbol{V}_3$. *Then,*

$$\{\boldsymbol{V}_1 | \boldsymbol{V}_2 = \boldsymbol{\theta}\} \overset{d}{=} f(\boldsymbol{\theta}) + \boldsymbol{V}_3,$$

for any $\boldsymbol{\theta} \in \mathbb{R}^n$.

Démonstration.

The proof is provided in [Wilson et al., 2021, Lemma 2]. □

Matheron's update rule for large datasets

Proof of Proposition 2 Maatouk, Rullière and Bay (2025).

From the equivalent between the two *direct* approaches Equations (4) and (7), we have

$$(\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)\boldsymbol{X}^\top/\sigma^2 = \tau^2 \boldsymbol{K}\boldsymbol{X}^\top(\tau^2 \boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top + \sigma^2 \mathbf{I}_n)^{-1}. \qquad (12)$$

Let $\boldsymbol{V}_3 := \boldsymbol{\xi} - (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})/\sigma^2$. Additionally, we have

$$\mathrm{E}[\boldsymbol{\xi}|\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}] = (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})/\sigma^2.$$

Thus, we can write :

$$\begin{aligned} \boldsymbol{\xi} &= \mathrm{E}[\boldsymbol{\xi}|\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}] + (\boldsymbol{\xi} - \mathrm{E}[\boldsymbol{\xi}|\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}]) \\ &= (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})/\sigma^2 + \boldsymbol{V}_3. \end{aligned}$$

Let $\boldsymbol{V}_1 = \boldsymbol{\xi}$ and $\boldsymbol{V}_2 = \boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}$. Since $\boldsymbol{V}_2$ and $\boldsymbol{V}_3$ are jointly Gaussian but uncorrelated, it follows that they are independent. $\square$

Matheron's update rule for large datasets

Rest of the proof.

Indeed,

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{V}_2, \boldsymbol{V}_3) &= \mathrm{Cov}(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon}, \boldsymbol{\xi} - (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})/\sigma^2) \\
&= \tau^2 \boldsymbol{X}\boldsymbol{K} - \mathrm{Var}(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})[(\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top/\sigma^2]^\top \\
&= \tau^2 \boldsymbol{X}\boldsymbol{K} - (\tau^2 \boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top + \sigma^2 \mathbf{I}_n)(\tau^2 \boldsymbol{X}\boldsymbol{K}\boldsymbol{X}^\top + \sigma^2 \mathbf{I}_n)^{-1}\tau^2 \boldsymbol{X}\boldsymbol{K} \\
&= \tau^2 \boldsymbol{X}\boldsymbol{K} - \tau^2 \boldsymbol{X}\boldsymbol{K} = \mathbf{0}_{n,N},
\end{aligned}$$

where $\mathbf{0}_{n,N}$ is the $n \times N$ zero matrix. The second-to-last line is done using Equation (12). Setting $f(\boldsymbol{V}_2) = (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top \boldsymbol{V}_2/\sigma^2$ and using Lemma 1, we obtain

$$\begin{aligned}
\{\boldsymbol{\xi}|\boldsymbol{X}, \boldsymbol{y}\} &\overset{d}{=} (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top \boldsymbol{y}/\sigma^2 + \boldsymbol{\xi} - \\
&\quad\ (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon})/\sigma^2 \\
&\overset{d}{=} \boldsymbol{\xi} + (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi} - \boldsymbol{\epsilon})/\sigma^2.
\end{aligned}$$

Hence, the claim follows. □

Table of Contents I

Table of Contents I

Illustrative examples 1D

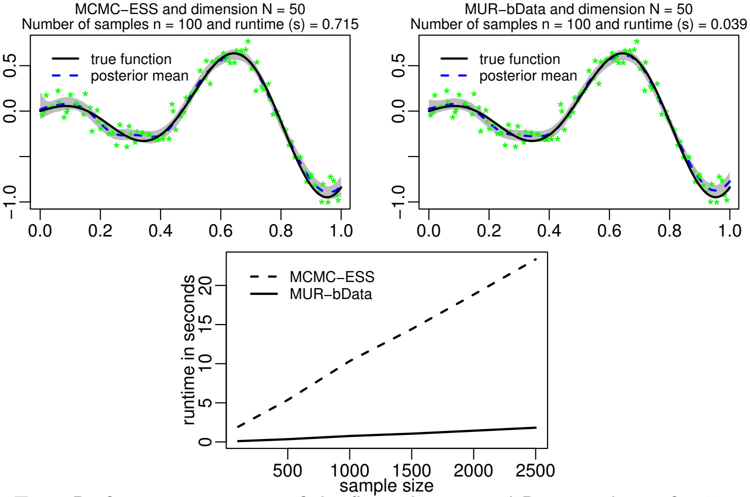

FIGURE: **Top** : Performance accuracy of the finite-dimensional Bayesian linear for $N = 50$. The number of samples (green stars) is fixed at $n = 100$. The gray shaded area represents the 95% confidence interval based on **6,000** sample paths. The highly efficient MCMC approach is employed in the left panel, while the proposed MUR for big data is applied in the right panel. **Bottom** : Runtime in seconds for generating **15,000** posterior sample paths as a function of the number of samples for the two competing approaches.

Illustrative examples 1D (big data and extreme case)

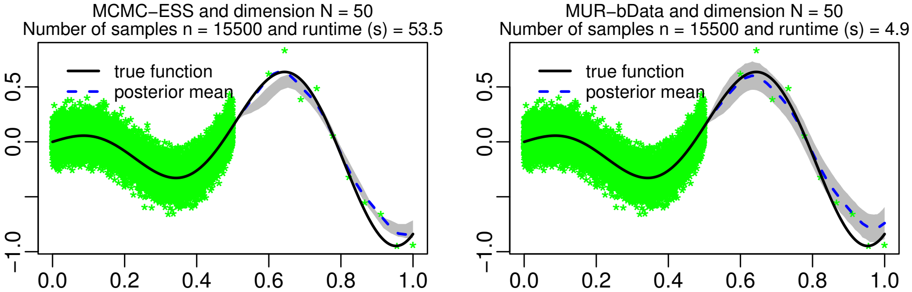

FIGURE: Same settings as Fig. 5 except the number of samples n which is fixed at **15,500** instead of **100**. There are **15490** observations available in the first half of the domain, while only **10** are available in the second. Unlike MUR-bData, the 95% confidence interval (gray shaded area) of the MCMC-ESS approach does not closely follow the posterior mean (blue dashed curve).

Table of Contents I

Real-world diamond data

Now, we apply the two strategies developed in this presentation to real-world diamond data. This dataset consists of the prices in US dollars (326$-18,823$) of $n = 53,940$ diamonds as a function of their carat, i.e., weight of the diamond (0.2-5.01).

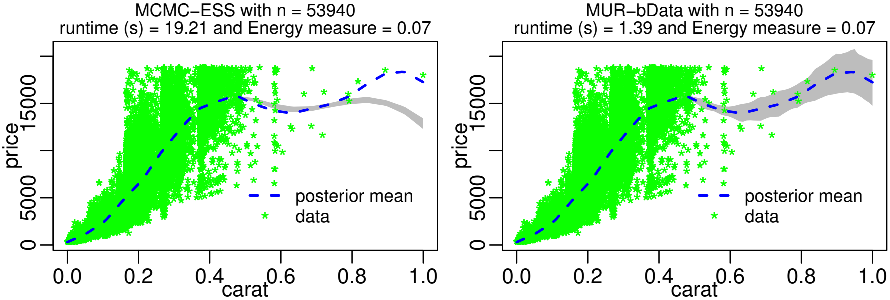

FIGURE: Accuracy estimation of the price of $n = 53,940$ diamonds as a function of carat. The two developed strategies are employed MCMC-ESS (left panel) and MUR-bData (right panel). The computational running time of generating 1,000 sample paths and the energy measure criterion are displayed in the main of each panel.

Table of Contents I

Illustrative examples 2D

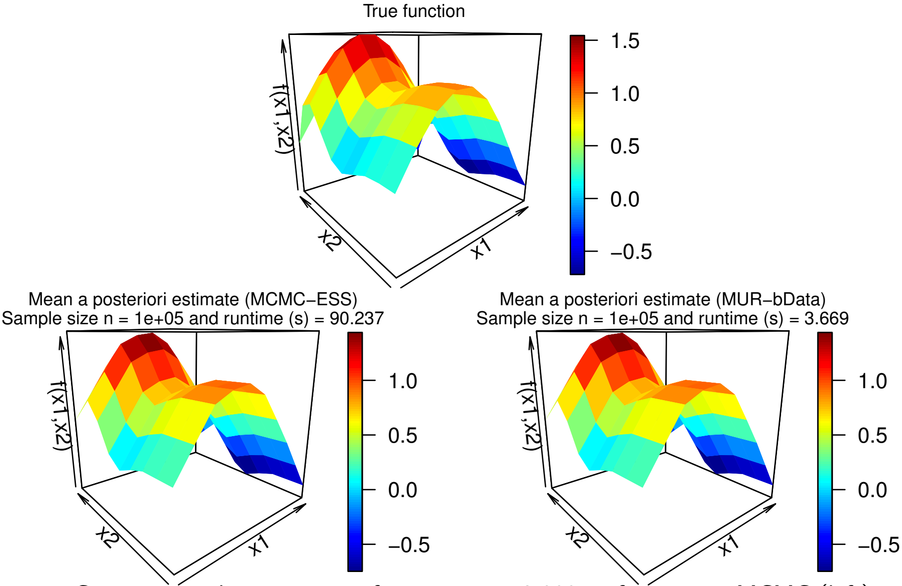

FIGURE: Computational running time for generating 2,000 surfaces using MCMC (left) and MUR-bData (right) on large datasets. The number of observations is $n = 100,000$.

Table of Contents I

Conclusion

Characteristic	MCMC-ESS	MUR-bData
	Approximate approach	Exact approach
	Sampling before conditioning	Sampling before conditioning
	Fast	Faster
	Flexible approach	Less flexible approach
	Iterative simulation	Iterative or direct simulation

TABLE: Comparison between MCMC-ESS and MUR-bData approaches.

References I

Maatouk, H., Rullière, D., and Bay, X. (2025).
Bayesian linear models for large datasets : Markov chain Monte Carlo or Matheron's update rule.
Under review in MCQMC2024.

Neal, R. (1999).
Regression and classification using Gaussian process priors.
Bayesian Statistics, 6 :475–501.

Williams, C. K. and Rasmussen, C. E. (2006).
Gaussian processes for machine learning, volume 2.
MIT press Cambridge, MA.

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. (2021).
Pathwise conditioning of Gaussian processes.
JMLR, 22(105) :1–47.

QUESTIONS ?

Thank you for your attention.