Université de Liège, Faculty of Applied Sciences

MECA 0010-1 -Uncertainty Quantification and Stochastic Modeling

Assignement 1: project report

Instructor
ARNST Maarten
Assistant
LECLERE Nicolas

BRANDOIT Julien
MONTAGNINO Clémence

Academic year 2024-2025

Introduction

The estimation of unknown parameters from indirect observations is a fundamental challenge in many scientific and engineering disciplines. In this assignment, such a *probabilistic inverse problem* is addressed where the goal is to estimate the epicentral coordinates of a seismic source based on the arrival times of seismic waves recorded at six stations. This problem is approached using a *Bayesian framework*, which allows us to incorporate uncertainties in the data and model them probabilistically.

The observed data consist of seismic wave arrival times at six known station locations, with the relationship between the arrival times and the epicenter governed by the travel time of seismic waves. Given the inherent uncertainties in measuring these arrival times, the data are treated as noisy observations. The objective is to estimate the epicenter coordinates (x, y) that best explain these noisy observations.

To perform this estimation, the *posterior distribution* of the epicentral coordinates is explored, which represents the updated belief about the epicenter location after accounting for the observed data and prior knowledge. To efficiently sample from this potentially complex posterior distribution, the *Random-Walk Metropolis-Hastings* [6] algorithm is employed, a well-established method within the class of *Markov Chain Monte Carlo* (MCMC) techniques.

In this analysis, two different ways to estimate the epicenter will be examined: The *Maximum A Posteriori* (MAP) estimator, which provides the most likely epicenter coordinates by identifying the peak of the posterior distribution. A *credible region* that quantifies the uncertainty surrounding the estimate, defining a region in which the true epicenter is expected to lie with high probability.

In this study, a MCMC method will be utilized to build the estimators, even though the problem at hand is relatively simple. While simpler, brute-force methods could be applied to this specific case, MCMC sampling was chosen to explore its application and advantages. This approach not only serves the current needs but also prepares for tackling more complex inverse problems where MCMC techniques become essential. By comparing the Maximum A Posteriori (MAP) estimator, which offers a single point estimate of the epicenter, with the credible region that captures the uncertainty surrounding the estimate, the aim is to illustrate the strengths and limitations of both methods. This exploration will enhance the understanding of how MCMC can be effectively used in seismic source estimation and other challenging scenarios.

Problem formulation

In this report, the aim is to estimate the epicentral coordinates of a seismic source $\boldsymbol{\theta} = (x, y)$ using Bayesian inference. It has been shown that, under certain assumptions, the posterior distribution can be expressed as:

$$\pi(x, y|\boldsymbol{x}^{obs}) \propto \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\sqrt{(x_j - x)^2 + (y_j - y)^2}/v - t_{obs,j})^2}{2\sigma^2}\right), \tag{1}$$

where N is the number of observations, v is the assumed constant and known wave speed, and σ is the uncertainty level - also referred to as the standard deviation in this report -

which will be the focus of the discussion..

The analysis focuses on obtaining point estimates through the Maximum A Posteriori (MAP) estimator :

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{x,y} \pi(x, y|\boldsymbol{x}^{obs}), \tag{2}$$

as well as quantifying uncertainty via a credible region C_α. This region is defined as :

$$C_\alpha = \{(x, y) : \pi(x, y|\boldsymbol{x}^{obs}) \geq \tau_\alpha\}, \tag{3}$$

where τ_α is determined such that :

$$\int_{C_\alpha} \pi(x, y|\boldsymbol{x}^{obs})\, dx\, dy = \alpha. \tag{4}$$

This credible region C_α can be interpreted as the smallest area in the parameter space (x, y) that contains α of the posterior probability mass. Thus, it provides a Bayesian measure of uncertainty around the estimated epicentral coordinates. In other words, given the observed data $\boldsymbol{x}^{obs}$, there is a probability α that the true epicenter lies within this region. Such a credible region is sometimes called HDR (highest density region) [2]. There are other credibility regions that could be studied [7].

Method

Random-Walk Metropolis-Hastings (RWMH) Algorithm Overview : The Random-Walk Metropolis-Hastings (RWMH) algorithm is a Markov Chain Monte Carlo (MCMC) method used to sample from a target posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x}^{obs})$, where $\boldsymbol{\theta} = (x, y)$ represents the unknown epicentral coordinates of the seismic source. The RWMH algorithm builds a Markov chain by proposing a new state $\boldsymbol{\theta}^*$ based on the current state $\boldsymbol{\theta}^{(t)}$, with acceptance probability:

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x}^{obs})}{\pi(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}^{obs})} \frac{q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right),$$

where $q(\cdot|\cdot)$ is the proposal distribution. If the proposal $\boldsymbol{\theta}^*$ is accepted, the chain moves to $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$; otherwise, it remains at $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

In this approach, a Gaussian transition matrix is used, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)}) = \mathcal{N}(\boldsymbol{\theta}^{(t)}, \frac{2.4^2}{2}[\Sigma])$, where $[\Sigma]$ is the covariance matrix parameter. This Gaussian choice results in a symmetric scenario, as $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$. This symmetry simplifies the acceptance probability calculation, reducing it to:

$$\alpha = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{x}^{obs})}{\pi(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}^{obs})}\right).$$

Covariance Matrix Selection and Scale Calibration : Selecting an appropriate covariance matrix $[\Sigma]$ is crucial for the RWMH algorithm because it balances exploration

and exploitation, impacting the acceptance rate of proposals [8]. A well-chosen $[\Sigma]$ enhances convergence speed and sampling efficiency, leading to more accurate posterior estimates. A two-step approach is employed to determine $[\Sigma]$:

1. *Initial Simulation with Identity Covariance*: An initial short run of the MCMC algorithm is started using $[\Sigma] = \mathbf{1}$, an identity covariance matrix, to approximate the Maximum A Posteriori (MAP) location and gather initial samples.

2. *Covariance Matrix Calibration*: Using the approximate MAP, the Hessian matrix of the log-posterior is computed via central finite differences, yielding a local curvature estimate [12]. This development is valid if the distribution is assumed to be Gaussian in the vicinity of the point in question. In our case, this assumption is assumed to be not completely erroneous. By inverting the Hessian, an estimate of the posterior covariance is obtained, and its diagonal is used as the scaled proposal covariance matrix for subsequent simulations: $[\Sigma] \approx \text{diag}(\Sigma_{xx}, \Sigma_{yy})$. One has the following :

$$\mathcal{H}(\boldsymbol{\theta}) \approx [\Sigma]^{-1} \tag{5}$$

This scaling improves the proposal's efficiency by matching the local spread of the posterior. The use of curvature at the MAP location is based on the idea that this region is where the chain is likely to spend many iterations. Methods that use variable covariance also exist and can be very effective [5].

Finite Difference Method for Hessian Estimation : For the Hessian estimation at the approximate MAP location, a central finite difference method is used [12, 1], which provides more accurate curvature information compared to forward or backward finite differences. This method allows to capture the behavior of the posterior in both the x and y directions more precisely. A step of $h = 10^{-3}$ is used which is clearly small compared to the scale of the problem. One has :

$$\mathcal{H}_{xx}(\boldsymbol{\theta}) \approx -\frac{\log \pi(x+h, y|\boldsymbol{x}^{obs}) + 2\log \pi(x, y|\boldsymbol{x}^{obs}) + \log \pi(x-h, y|\boldsymbol{x}^{obs})}{h^2} \tag{6}$$

$$\mathcal{H}_{yy}(\boldsymbol{\theta}) \approx -\frac{\log \pi(x, y+h|\boldsymbol{x}^{obs}) + 2\log \pi(x, y|\boldsymbol{x}^{obs}) + \log \pi(x, y-h|\boldsymbol{x}^{obs})}{h^2} \tag{7}$$

Burn-In Period and Initialization Strategy : The MCMC chain is initialized from a random point within a prior-defined region, specifically $[0, 20] \times [0, 15]$. To address the transient behavior at the start of the chain, a burn-in period is used, during which initial samples are discarded. This approach reduces the influence of the starting position, which is probably in a region of low probability density [3]. For subsequent runs, the chain is initialized at the approximate MAP, as this is a region of high probability density, reducing the need for additional burn-in. A burn-in of $t_{burn-in} = 1000$ steps is used which seems clearly sufficient. One has the following output from the RWMH chain :

$$\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{N} \longrightarrow \{\boldsymbol{\theta}^{(t)}\}_{t=t_{\text{burn-in}}}^{N} \tag{8}$$

Convergence Assessment using Gelman-Rubin Diagnostic : To ensure the MCMC chain has converged, the Gelman-Rubin convergence diagnostic is used [4], $\hat{R}$, computed separately for each parameter, x and y. The diagnostic compares the variance within

chains to the variance between chains, with values close to 1 indicating convergence. This procedure confirms that the chains adequately sample from the posterior distribution. A very often used threshold value is $\hat{R} \leq 1.1$ [10] even if some more critical applications use $\hat{R} \leq 1.01$.

Maximum A Posteriori (MAP) Estimation : To approximate the MAP estimator, the sample from the MCMC output that maximizes the posterior probability density (up to a multiplicative constant since the arg max operation is unaffected by a constant factor) is identified. This sample provides the MAP estimate $\hat{\boldsymbol{\theta}}_{MAP}$, representing the most probable epicenter coordinates based on the observed data. One has :

$$\hat{\boldsymbol{\theta}}_{MAP} \approx \arg \max_{\boldsymbol{\theta}^* \in \{\boldsymbol{\theta}^{(t)}\}_{t=t_{\text{burn-in}}}^{N}} \pi(\boldsymbol{\theta}^* | \boldsymbol{x}^{obs}) \tag{9}$$

Construction of Credible Region : To quantify the uncertainty around the MAP estimate, a credible region is constructed based on the posterior distribution:

1. *Rectangle Approximation using Marginal Quantiles*: A preliminary credible region is defined by computing marginal credible intervals for each parameter, x and y, using quantiles from the MCMC samples. This approach yields a rectangular credible region. Note that there is no guaranty on the probability of the mass enclosed in this rectangular region, as only a marginal probability is imposed - for each parameter separately. One has the following, where 1D serves as a reminder that the region is built marginally :

$$C_\alpha^{\texttt{1D}} = \left[x_{(1-\alpha)/2}, x_{(1+\alpha)/2}\right] \times \left[y_{(1-\alpha)/2}, y_{(1+\alpha)/2}\right] \tag{10}$$

where $x_{(k)}$ denotes the k-th ordered statistic of the MCMC samples for the parameter x, and similarly for y.

2. *2D Histogram-Based Credible Region*: For a more nuanced credible region, a 2D histogram with n bins is created, partitioning the parameter space into a grid. The number of bins is an hyperparameter of the method, $n = 50$ is used. By counting the number of samples in each bin, the density of the patch is estimated and a 2D quantile region C_α is constructed. Indeed, the bins can be ordered by approximated density and this order is used to build a *not* a priori rectangular region. This region provides a flexible and accurate measure of uncertainty by including areas with high posterior mass, aligning with the highest density region (HDR) approach. 'Mathematically, this method approximates equation (4) by converting the integral into a sum over bins. The criterion in equation (3) (the value of τ_α) is then approximated by ordering bins according to their estimated density. Instead of evaluating point by point, the region is composed of contiguous patches. Other methods exist, such as using kernel density estimation [11, 9] - which has the advantage of being non-parametric - to estimate the density map from the MCMC samples and then identify the iso-probability lines that enable us to construct an HDR.

Results

The posterior distribution for different uncertainty levels

In this simplified 2D problem, the posterior distribution (up to a multiplicative constant) can directly be visualized, which provides a unique perspective on the spatial uncertainty

of the epicenter. This simplification allows to examine how varying levels of uncertainty affect the posterior distribution in a straightforward, interpretable way.

For this analysis, two different values of the standard deviation are considered, $\sigma = 0.1$ and $\sigma = 0.25$. Figure 1 displays the posterior probability density function (PDF) for each σ value, illustrating how the spread of the distribution changes with increasing uncertainty. When $\sigma = 0.1$, the posterior is more concentrated, indicating tighter confidence in the epicenter location. In contrast, for $\sigma = 0.25$, the posterior distribution is more diffuse, capturing greater uncertainty in the epicenter position due to the larger assumed observational error.

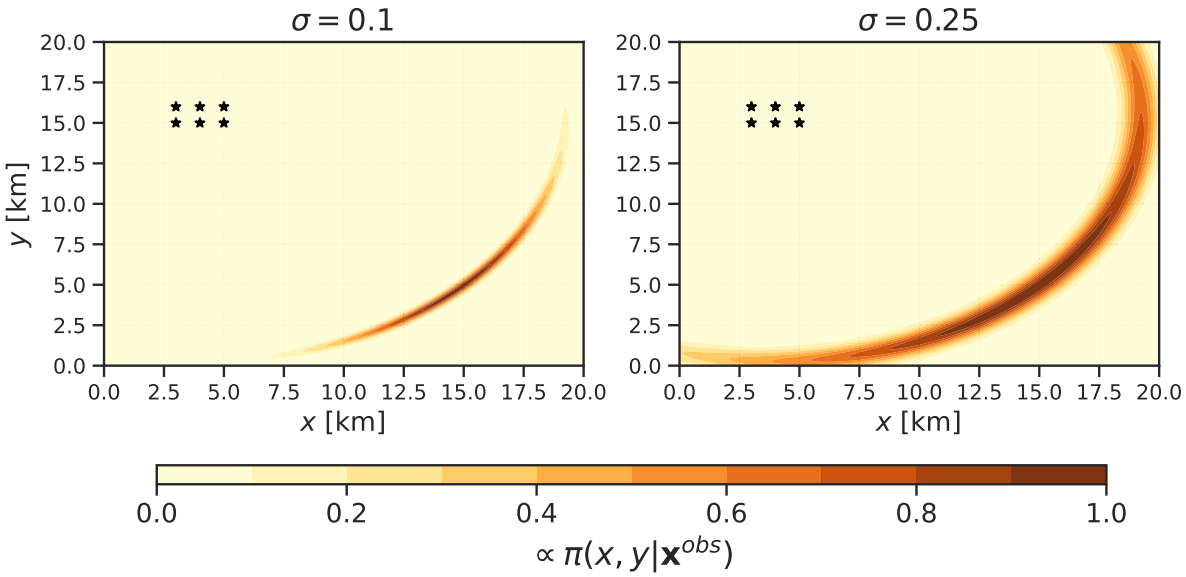


Figure 1: Posterior probability density function (*not normalized*) for different uncertainty levels σ

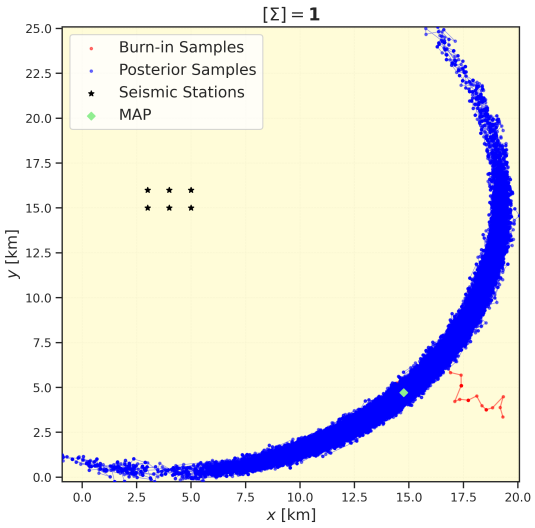
The two successive chains of Random-Walk Metropolis-Hasting

The method for estimating parameters via RWMH chains is a two-stage process. A first chain of length $N = 100000$, with a burn-in period of $t_{burn-in} = 1000$ samples, of random start sampled from the rectangular region $[0, 20] \times [0, 15]$ and based on a unit covariance matrix is used to estimate two things: 1. A first appoximative MAP position of the probability distribution. 2. A diagonal approximation of the covariance matrix at this point via the inverse of the finite-difference Hessian matrix. A second chain, of longer length $N = 1000000$, without burn-in, initiated from the first MAP approximation and with the covariance matrix calibrated from the first chain, is used to construct the final estimators. These approximations are obtained for the two uncertainty values:

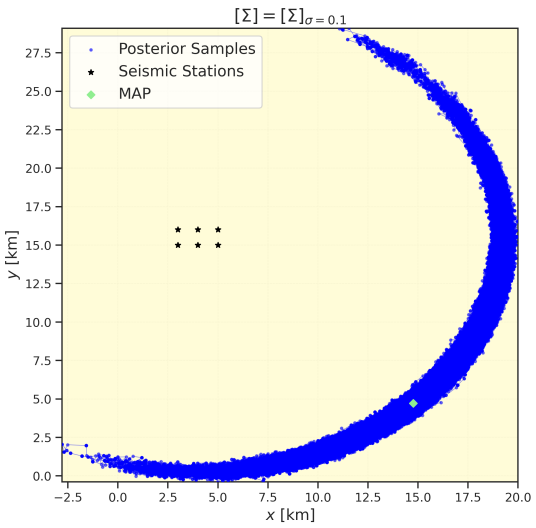
$$\hat{\boldsymbol{\theta}}_{\text{MAP},\sigma=0.1} = \begin{bmatrix} 14.73 \\ 4.69 \end{bmatrix} (km) \quad \text{and} \quad \hat{\boldsymbol{\theta}}_{\text{MAP},\sigma=0.25} = \begin{bmatrix} 14.75 \\ 4.7 \end{bmatrix} (km)$$

$$[\Sigma]_{\sigma=0.1} = \begin{bmatrix} 0.084 & 0 \\ 0 & 0.083 \end{bmatrix} (km^2) \quad \text{and} \quad [\Sigma]_{\sigma=0.25} = \begin{bmatrix} 0.52 & 0 \\ 0 & 0.52 \end{bmatrix} (km^2)$$

The MCMC results from $\sigma = 0.1$ for these two different matrices are displayed in Figure 2. There are no interesting differences in the chains from $\sigma = 0.25$ except that this one covers a larger area, similar to Fig. 1.



(a) Chain from the first RWMH run with identity covariance. $N = 100000$, $t_{burn-in} = 1000$, $\boldsymbol{\theta}^{(0)} \sim \text{Uniform}\left((0, 20); (0, 15)\right)$



(b) Chain from the second RWMH run with the calibrated covariance. $N = 1000000$, $t_{burn-in} = 0$, $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_{MAP,\text{run 1}}$

Figure 2: Results from the RWMH algorithm's two-stage process illustrated for $\sigma = 0.1$

The MAP is clearly unaffected by the σ value - and this can be confirmed by the analytical expression of $\pi(x, y|\boldsymbol{x}^{obs})$ where σ plays only a scaling-factor role. The $[\Sigma]$ matrices are affected by the uncertainty and are in both case smaller than the original guest $[\Sigma] = \mathbf{1}$.

Convergence analysis using the Gelman-Rubin metric

The figure shows how the $\hat{R}$ statistics evolve as the number of samples increases, starting from 1 sample up to 25 000 (after removing the burn-in). It is evident that $\hat{R}$ values approach the convergence threshold of 1.1 as the number of samples increases, indicating the stabilization of the chains. Notably, the Gelman-Rubin statistics for both components remained below the threshold, confirming satisfactory convergence for all chains given a sufficient number of samples. The convergence plots reveal that while both σ values led to effective convergence, the chains with $\sigma = 0.25$ exhibited a slightly slower convergence rate. This difference is not very large, however, and convergence can be said to be similar for both uncertainty values. Confidence in the results presented in this report is supported by the use of chains that are considerably longer than the minimum sample size required for convergence.

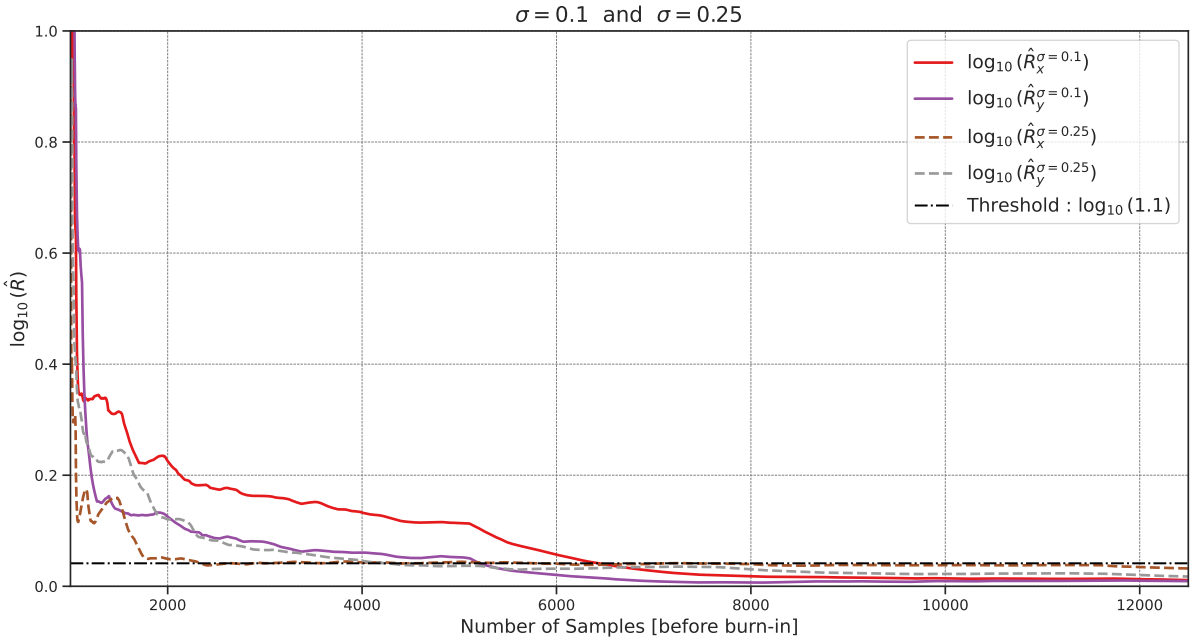
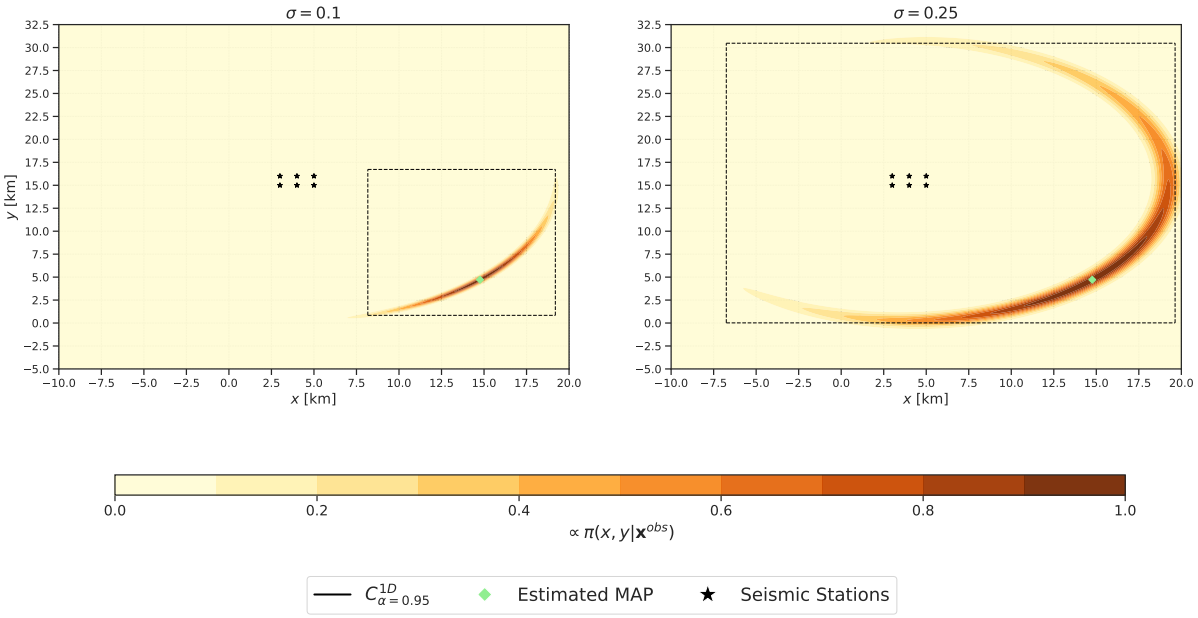


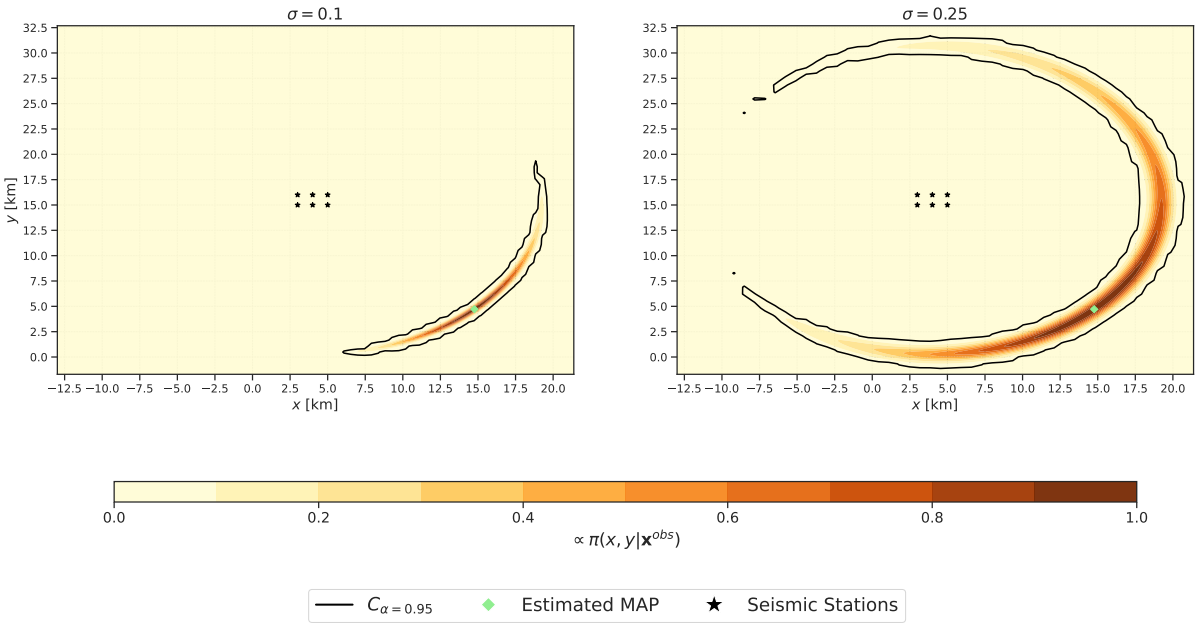
Figure 3: Convergence of the Gelman-Rubin $\hat{R}$ metrics for $\sigma = 0.1$ and $\sigma = 0.25$. *10 chains of max length* $N = 25000$, $t_{burn-in} = 1000$, $\boldsymbol{\theta}^{(0)} \sim$ *Uniform* $((0, 20); (0, 15))$ *were used. No interesting differences are observed after* 12000 *samples so the graph is cropped*

Epicenter estimations from the Random-Walk Metropolis-Hasting samples

This *"toy" problem* was chosen as an opportunity to explore the construction of estimators using only samples from an MCMC chain. Following the procedures described in the *Methods* section, three estimators for the earthquake epicenter were constructed: the Maximum à Posteriori estimator, $\boldsymbol{\theta}MAP$; the credible region $C\alpha^{1D}$, a prior rectangular region aligned with the axes and determined using Marginal Quantiles; and the credible region C_α, a more complex shape derived from a 2D histogram of the samples. The two credible regions were constructed from the $N = 1,000,000$ sample chains generated in the second step of the RWMH procedure, using a confidence level of $\alpha = 0.95$. The $C^{1D}_{\alpha=0.95}$ and $C_{\alpha=0.95}$ credible regions are respectively shown in Fig. 4a and Fig. 4b. One can clearly see that the uncertainty level has an effect on the size of the credible regions, but not on the position of the MAP. The larger the uncertainty, the larger the credible regions (their areas in the parameters space). On can also see that the histogram-based credible regions follow more closely the true level contours of $\pi(x, y|\boldsymbol{x}^{obs})$ than the rectangular ones.



(a) $C^{1D}_{\alpha=0.95}$ - Rectangle Approximation using Marginal Quantiles



(b) $C_{\alpha=0.95}$ - 2D Histogram-Based Credible Region

Figure 4: Estimated credible regions for both $\sigma = 0.1$ and $\sigma = 0.25$

Discussion

Estimators for the seismic epicenter were successfully constructed solely from the MCMC chain samples using a covariance matrix calibration method. These estimators were developed following an analysis of chain convergence based on Gelman-Rubin statistics to ensure a sufficiently large sample size. The estimators vary in nature and convey fundamentally different information about the epicenter, which in turn affects their interpretation and practical application.

The credible region is particularly valuable as it quantifies the uncertainty associated with each pair of values. As previously discussed, this region can be computed using two distinct approaches. In the rectangular approximation, the probability mass within the region remains unspecified, while the 2D histogram-based method explicitly defines it. The latter ensures that only areas with high probability density are included. However, since the credible region encompasses multiple points rather than a single estimate, it can complicate interpretation and subsequent analyses derived from epicenter estimation.

To mitigate the complexities associated with the credible region, the Maximum A Posteriori (MAP) estimate can be utilized. As a point estimate, the MAP identifies the most probable pair of values, making it significantly easier to manipulate in practice. However, it is important to note that the MAP offers no measure of uncertainty, which is a notable limitation.

The importance of the burn-in period is well illustrated in the initial step of the RWMH procedure shown in Fig. 2a. Here, the initial MCMC estimates fall outside the region of interest, necessitating the exclusion of these early samples to enhance accuracy and ensure proper convergence of the chains.

BComparison of the covariance matrices from the initial run (identity) and the calibrated second run reveals that the initial estimate is larger than the calibrated matrices. Nevertheless, this difference is not excessive, and the two results presented in Fig. 2a and Fig. 2b show no fundamental disparities. This finding aligns with existing literature on the importance of covariance matrix selection [8]. It is essential to recognize that the covariance matrix is not fixed; rather, it should not be excessively large or small relative to the problem scale. An appropriate covariance choice promotes good convergence (in comparison to a matrix that is too large) and facilitates rapid convergence (compared to one that is too small). Issues frequently arise when sampling from multimodal distributions.

The covariance matrix is also influenced by the level of uncertainty. Increased uncertainty leads to a flatter distribution, resulting in a larger covariance. This variability in the covariance matrix enables the algorithm to explore a more widely spread distribution with larger steps.

Utilizing a calibrated covariance matrix may explain why the convergence rates between the two values of σ are not significantly different. Although a more spread-out distribution necessitates larger steps in parameter space, it allows for a similar exploration rate. This assumption holds true in contexts where the overall shape of the distribution, as well as its unimodal nature, remains approximately unaffected by σ, with only the scale being affected.

Finally, the advantages of rectangular regions based on marginal distributions with those derived from histograms can be compared. While rectangular regions are much simpler to construct, they have significant drawbacks, including the inability to impose a joint probability mass and the excessive restrictions imposed by rectangular shapes aligned with the axes, especially in more complex problems like the one at hand. Nonetheless, these regions are still influenced by uncertainty, and their size could serve as a measure of this uncertainty.

In contrast, 2D histogram-based regions are much more representative, although more complex to construct. They also encode a measure of uncertainty through their size and allow for a stricter (although still approximate) imposition of the encompassed probability mass. However, this method has the disadvantage of being parametric, which may

necessitate exploring the influence of the number of bins used. This approach permits a wider variety of region shapes, even though they are still constrained by rectangular patches. Exploring methods based on Kernel Density Estimation (KDE) [11, 9] could address both of these issues.

In conclusion, although Markov Chain Monte Carlo (MCMC) methods were used to construct estimators, it is worth noting that more precise methods could have been considered for this specific problem, given its low dimensionality and relative simplicity. For example, a gradient ascent method could be employed to determine the Maximum A Posteriori (MAP), and contour extraction could provide a more accurate estimation for the credible region.

References

[1] Neculai Andrei. "Diagonal Approximation of the Hessian by Finite Differences for Unconstrained Optimization". In: *Journal of Optimization Theory and Applications* (2020). DOI: 10.1007/s10957-020-01676-z.

[2] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons, 1973. ISBN: 9780471574286. DOI: 10.1002/9781118033197.

[3] Andrew Gelman. *Bayesian Data Analysis*. 2nd. Boca Raton, Fla.: Chapman & Hall / CRC, 2004. ISBN: 978-1584883883.

[4] Andrew Gelman and Donald B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–472. DOI: 10.1214/ss/1177011136.

[5] Heikki Haario, Eero Saksman, and Johanna Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.2 (2001), pp. 223 –242.

[6] W. K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1 (1970), pp. 97–109.

[7] E. T. Jaynes. "Confidence Intervals vs Bayesian Intervals". In: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Ed. by W. L. Harper and C. A. Hooker. Dordrecht: D. Reidel, 1976.

[8] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. USA: Oxford University Press, 1999. ISBN: 978-0198517979.

[9] E. Parzen. "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3 (1962), pp. 1065–1076. DOI: 10.1214/aoms/1177704472.

[10] Roger D. Peng. *Monitoring Convergence*. https://bookdown.org/rdpeng/advstatcomp/monitoring-convergence.html. Accessed: 2024-10-31. 2020.

[11] M. Rosenblatt. "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3 (1956), pp. 832–837. DOI: 10.1214/aoms/1177728190.

[12] Ka-Veng Yuen. *Bayesian Methods for Structural Dynamics and Civil Engineering*. Wiley, 2010. DOI: 10.1002/9780470824566.