

Chapter 5

Bayesian inference and kinematic source inversion

*“La probabilité est relative en partie à cette ignorance,
en partie à nos connaissances.”*

— Pierre-Simon Laplace, 1840

Contents

| | | | | | |
|------------|---|---|---|---|------------|
| 5.1 | Introduction | . | . | . | 141 |
| 5.2 | Bayesian inference | . | . | . | 143 |
| 5.2.1 | Bayes Theorem | . | . | . | 143 |
| 5.2.2 | Key ingredients of Bayesian statistical inference | . | . | . | 144 |
| 5.3 | Markov Chain Monte Carlo and Hamiltonian Monte Carlo | . | . | . | 145 |
| 5.3.1 | Markov Chain Monte Carlo (MCMC) | . | . | . | 145 |
| 5.3.2 | Hamiltonian Monte Carlo (HMC) | . | . | . | 148 |
| 5.3.3 | MCMC vs HMC: encouraging examples | . | . | . | 153 |
| 5.4 | HMC and the Kinematic Source Inversion Problem | . | . | . | 159 |
| 5.4.1 | Resetting the kinematic source inversion problem | . | . | . | 160 |
| 5.4.2 | HMC kinematic source inversion examples | . | . | . | 162 |
| 5.5 | Conclusion | . | . | . | 175 |

5.1 Introduction

As I show in the previous chapters, the kinematic source inversion problem consists of providing a reconstruction of the time and space history of a seismic rupture which explains, to some degree of accuracy, a given set of observations (seismograms, accelerograms or geodetic data). However, the solution of the ill-posed kinematic inverse problem is non-unique. For instance, it is possible to appreciate the large variability between the kinematic models proposed to explain the rupture of the 2010 ($M_w 8.8$) Maule, Chile, earthquake in Figure 1.10. In addition, according to current acquisition systems surrounding active faults, this problem is highly under-determined, in spite of its rather simple

formulation as a linear inverse problem. Regarding all these limitations, it is possible to say that there are as many acceptable kinematic models for a given earthquake as authors working on it. However, accurate kinematic models are fundamental to enhance our knowledge of the seismic cycle as well as to improve surface ground motion prediction. Consequently, because it is impossible (up to now) to give an accurate and unique solution correctly representing a real earthquake source, it is necessary to assess the uncertainties associated to any proposed model.

Optimistically, it can be seen that for some recent earthquakes that present a large set of observations with a significant good azimuthal distribution (reducing the under-determination of the problem) the proposed models provided by different authors are not so different from each other (see for instance the results proposed for the 2016 Kumamoto earthquake from Asano and Iwata, 2016; Uchide et al., 2016; Hao et al., 2017; Sánchez-Reyes et al., 2018). Agreement between solutions might mean that we are all observing solutions in the vicinity of the true solution. Due to uncertainties on the modeling, fault geometry, description of the medium and data accuracy, a closer estimation to the true solution is prevented. However, similar results might be inside the same cloud of probable scenarios, or in other words, they belong to the same probability distribution. Therefore, providing a description of this vicinity of probable solutions seems to be much more suitable to tackle this problem than providing a unique (inaccurate) rupture model for a given earthquake.

Some authors have focused their work to study all the possible sources of uncertainties that are part of this problem. For instance, regarding our poor knowledge of the velocity-density structures and the impact that this limited knowledge has in our inferences of the source parameters, I can mention the work from Yagi and Fukahata (2011) and Hallo and Gallović (2016). Another important source of uncertainties is related to all the simplistic geometrical assumptions that are used to describe the fault surfaces where the rupture occurs. Usually, simple models with one, two or even three different fault segments with prescribed geometries (strike and dip angles) are used to represent the complex surfaces where the faulting takes place. Recently, Ragon et al. (2018) presented a framework which allows to account for the uncertainties of fault geometry into the static source inversion. However, the extension of this strategy to the kinematic inversion problem has not yet been explored. A third source of uncertainties is related to the different possible shapes that the slip-rate functions can take (under non-linear formulations). The work from Razafindrakoto and Mai (2014) presents an investigation regarding the impact of the uncertainties of the assumed source time functions, as well as the unknown velocity-density structure, into the kinematic source inversion problem. Finally, all the available data used to study earthquakes has a significant level of noise that might not be completely removed by the digital filters, which might be another source of uncertainties that can impact our results without mentioning the limited coverage that current acquisition systems provide.

Duputel et al. (2012) and Minson et al. (2013) are some of the pioneering works related to the uncertainty assessment of static and kinematic models in the earthquake source imaging field. To my knowledge, these works are the first to explore in a completely Bayesian framework this problem. Several other works such as Duputel et al. (2014) or Hallo and Gallović (2016) continue the development of this research field, establishing ways to account for the estimated uncertainties related to the unknown velocity structure into the source inversion strategies. However, most of the works related to this topic still relies on non-linear formulations of the forward problem: the source history is described by a reduced number of parameters such as rise time, rupture velocity, rake angle and slip amplitude. These parameters are not linearly related to the observed seismograms. And, even though such parameters drastically diminishes the dimensionality of the model space where the exploration has to be performed, this parametrization implies drastic assumptions on the rupture history.

In this chapter, I present a Monte Carlo strategy to tackle the kinematic source inversion problem under a linear time-domain formulation. This strategy is based on Hamiltonian dynamics, which can be used to efficiently describe, under a Bayesian framework, the high dimensional probability distribution of solutions to this inverse problem. The development here presented could be used to tackle the following two goals: 1) To make an attempt to obtain the solution to the kinematic source inversion problem together with the associated uncertainties or 2) to assess the uncertainties around a given solution which can come from either a Bayesian or a deterministic approach. In this chapter I present a preliminary investigation focused on the second goal. Under this approach, I provide two synthetic examples under an idealized source-receiver geometry to illustrate the advantages and performance of this promising technique.

Outline Consequently, this chapter is organized as follows:

- In Section 5.2, I mention basic concepts related to Bayesian inference. In this section, I introduce concepts such as the prior, likelihood and data evidence probability distribution functions and how these concepts can be used to infer characteristics of a target function that needs to be explored.
- In Section 5.3, I present two different strategies to perform the exploration of a posterior probability functions: 1) a standard Markov Chain Monte Carlo strategy and 2) a Hamiltonian Monte Carlo approach. Once the fundamental concepts of both strategies are presented, I give two encouraging examples that illustrate some of the advantages of the HMC approach over the standard MCMC.
- In Section 5.4, I reformulate the inverse problem that I present in Chapters 3 and 4 to be able to work with the HMC approach. In this section, I demonstrate the suitability of the HMC strategy to tackle the kinematic source inversion problem applying it to two synthetic case: 1) a simple spike test and 2) a more complex bilateral rupture.
- Finally, in Section 5.5 I discuss some conclusions and perspectives for this new strategy as well as its limitations.

5.2 Bayesian inference

Let me start with a non-exhaustive description of how the Bayesian statistical inference works and what are the fundamental ingredients necessary to infer probable values of the unknown parameters which solve an inverse problem. This section does not offer a rigorous mathematical formulation of the concepts behind Bayesian inference. To have a deeper insight on this topic and its application to geophysical problems, I invite the reader to consult works as Olofsson and Andersson (2005); Sen and Stoffa (2013) and Tarantola (2005).

5.2.1 Bayes Theorem

Let me first state Bayes theorem, which is the mathematical expression that lies behind all what is presented in this chapter.

To do that, let us define two events that are observed to occur independently. The first event is denoted as event **A**, while the second is **B**. The probabilities of observing each of these two events

are defined as $P(\mathbf{A})$ and $P(\mathbf{B})$. Such probabilities are also known as the *marginal probabilities*. I can now define the conditional probability of observing event \mathbf{A} given that event \mathbf{B} happened as $P(\mathbf{A}|\mathbf{B})$. This conditional probability is also known as the *likelihood* of \mathbf{A} given \mathbf{B} . Here, I denote the joint probability of both events occurring as $P(\mathbf{A} \cup \mathbf{B})$. This joint probability of both events occurring is estimated as the product of the marginal probability of \mathbf{B} and the likelihood of \mathbf{A} given \mathbf{B} , or vice versa. In other words,

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}). \quad (5.1)$$

Thus, the relationship between the two conditional probabilities can be expressed as

$$P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})}. \quad (5.2)$$

In terms of a general inverse problem, such as $\underline{\mathcal{G}}\underline{m} = \underline{d}$, where \underline{d} is the vector of observations, \underline{m} is the vector of unknowns and $\underline{\mathcal{G}}$ is the mapping operator from the space of unknowns to the space of observables, equation (5.2) can also be written as,

$$P(\underline{m}|\underline{d}) = \frac{P(\underline{d}|\underline{m})P(\underline{m})}{P(\underline{d})}, \quad (5.3)$$

which means that the probability of a certain model \underline{m} given the observed data \underline{d} can be obtained as the product between the marginal probability of the model (i.e. *prior probability of the model*) and the likelihood of the data divided by the marginal probability of the data (i.e. *evidence of the data*). In terms of the kinematic source inversion problem, $P(\underline{m}|\underline{d})$ is the probability of a reconstructed slip-rate time-space history $\underline{m} = \underline{V}(\xi, \tau) = [V_\phi(\xi, \tau), V_\delta(\xi, \tau)]^T$ given a certain set of data (e.g. seismograms) $\underline{d} = \underline{v}(\underline{x}, t) = [v_x(\underline{x}, t), v_y(\underline{x}, t), v_z(\underline{x}, t)]^T$.

I would like to highlight that the main goal of equation (5.3), contrary to the form $\underline{m} = \underline{\mathcal{G}}^{-1}\underline{d}$, is not to find the most suitable \underline{m} explaining \underline{d} . What equation (5.3) aims at doing is to provide an estimate of the probability of \underline{m} given the observed data \underline{d} , and nothing more. Consequently, because the most suitable \underline{m} remains unknown, it is our duty to smartly use this estimated probability $P(\underline{m}|\underline{d})$ to infer not only the most suitable solution but also the uncertainties related to it.

5.2.2 Key ingredients of Bayesian statistical inference

As it was mentioned above, Bayes theorem relies on three fundamental probabilities to estimate the suitability and uncertainties of a given model \underline{m} . These three ingredients are: 1) $P(\underline{d}|\underline{m})$ the likelihood of the data given a certain model, 2) $P(\underline{m})$ the prior model probability and 3) $P(\underline{d})$ the evidence of the data. It can be seen in equation (5.3) that the evidence of the data acts as a normalization factor. This term is the one in charge of limiting $P(\underline{m}|\underline{d})$ to a range between 0 (unlikely) and 1 (likely). Therefore, the term related to the evidence of the data can be estimated as,

$$P(\underline{d}) = \int_{\underline{m}} P(\underline{m}|\underline{d})P(\underline{m})d\underline{m} \quad (5.4)$$

However, most of the real and complex problems tackled in the real world have large dimensions. These real world problems pose some big obstacles to estimate this data evidence term. For instance, one important reason why $P(\underline{d})$ is often computationally intractable is the fact that for its computation it is required to integrate over the whole model space. Whether a closed form for the prior or the likelihood *pdfs* are known or not, this integration makes the computation of the data evidence prohibitive. Therefore, the evidence of the data is commonly not estimated in practice.

Fortunately, the evidence of the data is only a normalization factor. Consequently, even if this term is not estimated, $P(\underline{m}|\underline{d})$ remains proportional to the product of the prior probability $P(\underline{m})$ and the likelihood $P(\underline{d}|\underline{m})$,

$$P(\underline{m}|\underline{d}) \propto P(\underline{d}|\underline{m})P(\underline{m}). \quad (5.5)$$

This proportionality is the feature used to infer characteristics of the probability distribution function $P(\underline{m}|\underline{d})$, which is the one we are interested in.

Now that the evidence of the data is discarded, it is possible to focus our attention to the other remaining two terms: the prior probability $P(\underline{m})$ and the likelihood $P(\underline{d}|\underline{m})$. These two terms, together with another one, known as the *proposal probability distribution function*, are the principal ingredients of any Bayesian inference technique. Let me then define these terms.

Prior probability distribution function $P(\underline{m})$: In Bayesian statistical inference, this function is the one in charge of reflecting the degree of certainty that we have of a particular model \underline{m} based on the prior knowledge that we have about the possible values that this model can take. This degree of certainty is taken from a probability distribution function that is built based on the available prior knowledge. It is important to notice that the probability that is associated to a certain model \underline{m} by this probability distribution function does not take into account any evidence (such as seismograms in the source reconstruction problem) that can help to support or disapprove our level of confidence on that model. In terms of the kinematic source inversion problem, this function is used to estimate the probability that a particular rupture model based on any prior knowledge in our hands, such as the shape of expected slip-rate functions, rupture velocity, maximum slip, etc.

Likelihood probability distribution function $P(\underline{d}|\underline{m})$: Contrary to the prior probability distribution function, the likelihood is defined as a particular function in charge of measuring the probability to observe a certain dataset \underline{d} given a known model \underline{m} . The associated model \underline{m} in this probability function is believed to well represent the frequency distribution of the population of all the possible models that explain, to some extent, the observed evidence. Once again, in terms of the kinematic source inversion problem, the likelihood is a function that estimates the probability of a given synthetic set of seismograms, associated to a given rupture model, to correctly represent the observed wavefields.

These ingredients, the prior and the likelihood probability distribution functions are at the core of the Bayesian statistical inference field. Let me now introduce two different techniques that have been proposed to infer characteristics of $P(\underline{m}|\underline{d})$ based on the concepts above mentioned.

5.3 Markov Chain Monte Carlo and Hamiltonian Monte Carlo

5.3.1 Markov Chain Monte Carlo (MCMC)

The Markov Chain Monte Carlo methods are a family of computational statistical techniques that employ two concepts: 1) Markov Chains and 2) Monte Carlo simulations to infer, or to estimate, characteristics of complex functions that, without using these numerical strategies, would be extremely difficult to assess. In order to understand the main idea behind these methods, let me explain briefly each of the concepts that form part of these techniques.

Monte Carlo methods: These techniques are direct sampling methods. These techniques are also known with the term *stochastic simulations*, as the sampling of a given function is randomly selected from all the available set of values that the function takes. The principal idea behind these techniques is that characteristics and other estimations of a given function can be obtained based on a repetitive massive random sampling of it.

Markov Chain: Following (Brooks et al., 2011), a sequence $\underline{m}^{(1)}, \underline{m}^{(2)}, \dots, \underline{m}^{(n)}$, of random elements of some set is said to be a *Markov Chain*, if the conditional distribution of $\underline{m}^{(n+1)}$ given $\underline{m}^{(1)}, \dots, \underline{m}^{(n)}$, depends on $\underline{m}^{(n)}$ only, which is denoted as $P(\underline{m}^{(n)}|\underline{m}^{(n-1)})$. The set in which every $\underline{m}^{(i)}$ takes place is known as the *state space* of the Markov Chain.

In the Bayesian statistical inference field there are several techniques that could be used to make an attempt to retrieve a correct description of $P(\underline{m}|d)$. One big family of these techniques are related to the Monte Carlo method as well as to the concept of Markov Chains. In order to further detail these techniques and their application to the kinematic source inversion problem, it is first necessary to introduce another important ingredient known as the *proposal probability distribution function*. Most of the Bayesian strategies that incorporate the Monte Carlo technique into their algorithms rely on a systematic evaluation of the conditional probability function $P(\underline{m}|d)$ of a large number of models (thousands or millions of them). Therefore, for those techniques it is very important to define a way, or a function, to select each of these models to be qualified from the enormous population of all the possible models. This function is known as the *proposal probability distribution function*.

Proposal probability distribution function $P(\underline{m}^{(k)}|\underline{m}^{(k-1)})$: This probability function acts as a conventional transition operator that allows to go from one initial state $\underline{m}^{(k-1)}$ to a next state $\underline{m}^{(k)}$ according to a given probability. Commonly, the new proposed state depends only on the previous state. This characteristic of the proposal distribution is also known as a *memoryless* transition. In terms of the kinematic source problem, this proposal function can be seen as an engine that is able to provide different rupture models (that have to respect some physical constraints) that might depend or not on a previous estimated rupture model.

Markov Chain Monte Carlo methods (MCMC): These strategies combine the Monte Carlo random sampling techniques together with the Markov Chain concepts to perform a more efficient, compared to simple Monte Carlo methods, exploration of a given function. The idea is to construct a particular Markov Chain of K states, $\underline{m}^{(1)}, \dots, \underline{m}^{(K)}$, using a repetitive random sampling Monte Carlo method described by the proposal probability distribution function. In order to accept a new proposal state $\underline{m}^{(k*)}$ to form part of the state space, an acceptance/rejection criterion has to be passed by the new proposed state. As the number of random samples grows, $K \rightarrow \infty$, the state space constructed by the MCMC technique is expected to converge towards the sampled function (feature known as *stationarity* of the Markov Chain). Consequently, the built Markov Chain from these strategies can be used to infer characteristics of the *ppd* which is the probability distribution we want to describe.

It is important to highlight that, with the massive use of the ingredients, $P(\underline{m})$, $P(d|\underline{m})$, and $P(\underline{m}^{(k)}|\underline{m}^{(k-1)})$, the MCMC methods are able to determine (approximately and fulfilling some requirements) the shape of $P(\underline{m}|d)$, which is the main goal. When this distribution is sampled enough, the resulting *posterior probability density (ppd)*, offers a correct representation or a map of the probable solutions to a given problem and how well they are distributed inside the space of probable solutions. For instance, in terms of the source reconstruction problem, the shape of this function will allow us to

evaluate the most probable value of the slip-rate function at a position on the fault and at a given time, $\underline{V}(\xi, \tau)$, taking into account the observed seismograms, $\underline{v}(\underline{x}, t)$ and any prior information. Moreover, not only the most probable scenario could be obtained, but also a description of its vicinity, which allows to assess the uncertainties of the most suitable rupture model.

Certainly, many more things can be said about the Monte Carlo simulations, Markov Chains, and the MCMC strategies. The purpose of this chapter, however, is not to detail these strategies. The reader is invited to consult Brooks et al. (2011), and references therein, to have an exhaustive explanation of many of the different properties, and derived techniques related to the MCMC methods.

One of the principal ingredients of the MCMC methods, besides the probability distribution functions that were already mentioned in the previous section (i.e. prior, likelihood and proposal probability distribution functions), is the acceptance/rejection criterion. This criterion has been a complete field of study. Metropolis et al. (1953); Hastings (1970) and Green (1995) are three of the most important publications discussing this criterion, as well the impact that it has on the MCMC strategies, and its generalization to facilitate a correct sampling of the target functions. In this work, however, I keep simple the discussion about this criterion. Therefore, the reader has to know that the acceptance/rejection criterion in this work is always referred to a Metropolis-Hastings (MH) generalization. Which means that, in order to accept a new proposed state $\underline{m}^{(k*)}$ to be part of the Markov Chain as $\underline{m}^{(k)}$, the ratio of probabilities r , estimated as

$$r = \frac{P(\underline{m}^{(k*)} | \underline{d}) P(\underline{m}^{(k-1)} | \underline{m}^{(k*)})}{P(\underline{m}^{(k-1)} | \underline{d}) P(\underline{m}^{(k*)} | \underline{m}^{(k-1)})} \quad (5.6)$$

has to be compared to a number u , that is sampled from a uniformly distributed function $u = \text{Unif}(0, 1)$, and the decision to accept or reject the proposal state is based on that comparison as follows,

$$\begin{cases} \underline{m}^{(k)} = \underline{m}^{(k*)} & \text{if } u < r, \\ \underline{m}^{(k)} = \underline{m}^{(k-1)} & \text{if } u \geq r, \end{cases} \quad (\text{accept}) \quad (5.7)$$

This acceptance/rejection criterion has the particular characteristic that, for every proposal whose probability $P(\underline{m}^{(k*)} | \underline{d})$ is larger than $P(\underline{m}^{(k-1)} | \underline{d})$ (being $r > 1$), the proposal $\underline{m}^{(k*)}$ is always accepted. Therefore, an efficient exploration towards the regions with denser probabilities is ensured. In addition, other proposals whose probabilities $P(\underline{m}^{(k*)} | \underline{d})$ are less than $P(\underline{m}^{(k-1)} | \underline{d})$ can still be accepted, as long as $u < r$. Such acceptance/rejection criterion is supposed to alleviate the problems that a MCMC exploration can have when dealing with rough functions, in which the chain can be prone to get trapped in certain regions (preventing a complete exploration of the target function).

Based on all the above mentioned: prior, likelihood and proposal distribution functions, as well as Monte Carlo simulations, Markov Chains and the acceptance/rejection criterion, a standard MCMC algorithm based on the Metropolis-Hastings criterion can be set as follows,

Algorithm 3: Standard Monte Carlo Markov Chain algorithm based on MH acceptance criterion.

input : Expressions for the estimation of the prior, likelihood and proposal probability distribution functions;

- 1: Set $k = 0$
- 2: Get first state, $\underline{m}^{(k)}$, from the prior probability functions $\underline{m}^{(k)} \sim P(\underline{m})$
- 3: **while** $k < K$ **do**

 - 3.1: set $k = k + 1$;
 - 3.2: Draw a new state $\underline{m}^{(k*)}$ using the proposal probability functions $\underline{m}^{(k*)} \sim P(\underline{m}^{(k*)} | \underline{m}^{(k-1)})$;
 - 3.3: Estimate the ratio $r = \frac{P(\underline{m}^{(k*)} | d)P(\underline{m}^{(k-1)} | \underline{m}^{(k*)})}{P(\underline{m}^{(k-1)} | d)P(\underline{m}^{(k*)} | \underline{m}^{(k-1)})}$ (Metropolis-Hastings algorithm);
 - 3.4: Draw a random number u , where $u = \text{Unif}(0, 1)$;
 - 3.5: **if** $u < r$ **then**

 - 3.5.1 accept the proposal state $\underline{m}^{(k*)}$ and set $\underline{m}^{(k)} = \underline{m}^{(k*)}$;

 - 3.5: **else**

 - 3.5.1 reject and set $\underline{m}^{(k)} = \underline{m}^{(k-1)}$;

Numerous problems can be tackled using the standard MCMC Algorithm here presented. However, this powerful strategy reduces exponentially its efficiency as the dimensionality of the problems tackled increases. This feature is well-known in the related research fields, and its basically due to two things: 1) the exponential growth of the space where the exploration has to be performed and 2) the random walk behavior along high dimensional spaces. To improve the efficiency of the space exploration several different strategies, based on the simplest MCMC algorithm, have been proposed to mitigate these effects. In this chapter, I focus my work on a specific technique based on Hamiltonian dynamics that, in recent years, has called the attention of many researchers working with high-dimensional problems, such as the kinematic source inversion problem that I want to tackle.

5.3.2 Hamiltonian Monte Carlo (HMC)

In the last decades, more sophisticated Monte Carlo strategies have been specifically developed in order to tackle high-dimensional problems. All these strategies are based on an efficient exploration of the high-dimensional space instead of a simple random-walk. Some of these methods, such as the optimal directional Gibbs sampler, relies on a pre-computation of optimal and independent directions where to explore (Christen et al., 2017). Other strategies, as the Langevin and Hamiltonian Monte Carlo rely on physical principles to generate in a more efficient way a new state of the random walk that allows a faster sampling of the target *ppd*.

Standard MCMC algorithms based on Metropolis-Hastings acceptance/rejection criterion are well-known to be hampered by their random-walk behaviors, specially for high-dimensional problems. One way to accelerate the space exploration is to add auxiliary variables that can help the chain to move more rapidly though the target distribution. An interesting family of methods that apply this augmented strategy are the Hamiltonian Monte Carlo methods. These techniques rely on the addition of a mo-

mentum term to each component of the target space. Then, both variables are updated together using Hamiltonian dynamics and the MH criterion. In this sense, Hamiltonian dynamics are used in order to get a new proposed state while MH criterion is in charge of accepting or rejecting the proposal. Contrary to the random-walk of the MCMC strategies, the HMC techniques take benefit from a different sampling rule. Such sampling rule, instead of being controlled by the proposal probability function, is based on a physical system known as Hamiltonian dynamics. This algorithm is designed to favor jumps inside the space according to a distance and a direction, which are controlled by the auxiliary variables that are related to the gradient of the target density function being explored. Thus, this different sampling algorithm allows the chain to move rapidly through the space, which allows a faster convergence towards the high density regions of the target distribution. In the following sections of this chapter, I present the fundamental concepts behind the HMC as well as the application of this technique to the kinematic source inversion problem.

Hamiltonian dynamics

Hamiltonian dynamics are a very useful tool to describe the motion of a particle through a system, which is shaped by the misfit landscape of the model space. In this section, I explain how it is possible to use the concepts related to Hamiltonian dynamics to assess the uncertainties of a general inverse problem. Let me then start describing the system that links the position and momentum of a particle and how this system can be linked to the probability distribution function that we want to explore.

Consider first a given particle position \underline{q} , which represent the model parameters (e.g. the time-space slip-rate history), and its momentum \underline{p} . The vector \underline{p} has the same dimension as \underline{q} , and each of its elements represent an auxiliary variable that is used to perturb the corresponding position (i.e. slip-rate value) of the particle. The position and the momentum have an associated potential $U(\underline{q})$ and kinetic $K(\underline{p})$ energies respectively. One can see that, neglecting other forces or interactions in the system, the total energy of the system, hereafter referred as Hamiltonian, at a given time instant t is the sum of these two energies,

$$H(\underline{q}, \underline{p}) = U(\underline{q}) + K(\underline{p}). \quad (5.8)$$

Because no external forces are acting in the system, such as friction, the Hamiltonian or total energy $H(\underline{q}, \underline{p})$ must remain constant during the evolution of the system. For instance, if there is an increment of the potential energy, the kinetic energy has to be reduced in order to keep constant the total energy of the system. The Hamiltonian equations are a set of differential equations describing this balance between potential and kinetic energies during the time evolution of the system,

$$\frac{\partial q_i}{\partial t} = \frac{\partial H}{\partial p_i} = \frac{\partial K(\underline{p})}{\partial p_i} \quad (5.9)$$

$$\frac{\partial p_i}{\partial t} = -\frac{\partial H}{\partial q_i} = -\frac{\partial U(\underline{q})}{\partial q_i} \quad (5.10)$$

The ensemble of equations (5.9) and (5.10) form a symplectic structure. This set of Ordinary Differential Equations (ODEs) make H to be constant when a proper integration in time is performed. From equations (5.9) and (5.10), it can be seen that having expressions for $\partial U(\underline{q})/\partial q_i$ and $\partial K(\underline{p})/\partial p_i$, and knowing the initial conditions of the system $[\underline{q}, \underline{p}]_{t=0}$, it is possible to predict the position and momentum of an object at any time instant by simulating the evolution of this dynamic system.

Let me now establish the link between the total energy $H(\underline{q}, \underline{p})$, which is preserved during integration thanks to the symplectic structure, and the probability distribution function $P(\underline{m}|\underline{d})$ that we want to describe. The energy $H(\underline{q}, \underline{p})$ can be related to $P(\underline{m}|\underline{d})$ through what is known as the *canonical distribution*,

$$\begin{aligned} P(H) &= \frac{1}{Z} \exp(-H(\underline{q}, \underline{p})) = \frac{1}{Z} \exp(-(U(\underline{q}) + K(\underline{p}))) \\ &\propto \exp(-U(\underline{q})) \exp(-K(\underline{p})) \propto P(\underline{q})P(\underline{p}) \end{aligned} \quad (5.11)$$

where $P(\underline{q})$ and $P(\underline{p})$ are independent probability distributions and Z is a normalization constant.

From equation (5.11), it is possible to see that the joint canonical distribution $P(H(\underline{q}, \underline{p}))$ can be factorized into the two independent terms $P(\underline{q})$ (for the position) and $P(\underline{p})$ (for the momentum). As I shall discuss further on, the position, which is related to values of parameters and its related potential energy can be directly linked to the misfit function of an inverse problem.

On the other hand, the momentum, and its associated kinetic energy, can be linked to the perturbations of the model parameters that are being investigated. The independence of both terms is an important property. Thanks to this feature, samples from the joint distribution $P(H(\underline{q}, \underline{p}))$ can be taken and, by ignoring the term related to the momentum (driving the model perturbation), inferences of the shape of the distribution associated to the position (misfit function of the inverse problem) can be done.

Potential energy: Assuming that a direct link between \underline{q} and the unknown model parameters exists $\underline{q} = \underline{m}$, the potential energy $U(\underline{q})$, from equation (5.8), has to be a particular function that can help us to infer the shape of the probability distribution that we are looking for. Therefore, we can define this function in terms of the prior probability and the likelihood as

$$U(\underline{q}) = -\log(P(\underline{q})) = -\log(P(\underline{m}|\underline{d})) \propto -\log(P(\underline{d}|\underline{m})P(\underline{m})). \quad (5.12)$$

Equation (5.12) describes a function of $P(\underline{q})$ that, when negated and run through the exponential function can provide us, at least, a proportional estimation of the target distribution. Using this potential energy function, a straightforward physical interpretation of this function can be observed: a large potential energy, associated to a given model \underline{m} , corresponds to very unlikely model parameter values, while a small potential energy is translated into a high probable scenario.

Kinetic energy: Because the probability distribution for $P(\underline{q})$ is independent of the distribution $P(\underline{p})$, we can choose any distribution from which to sample the momentum variables. In this field, several authors (e.g Neal et al., 2011; Betancourt, 2017; Fichtner et al., 2018) suggest to use a zero-mean normal distribution with covariance matrix $\underline{\underline{M}}^{-1}$. The matrix $\underline{\underline{M}}$ is interpreted in the Hamiltonian dynamics as the mass matrix. This matrix must have a symmetric and positive definite structure. Therefore, a simple quadratic kinetic energy function can be defined in terms of the particle momentum as follows,

$$K(\underline{p}) = \frac{\underline{p}^T \underline{\underline{M}}^{-1} \underline{p}}{2}, \quad (5.13)$$

in such a way that

$$P(\underline{p}) \propto \exp(-K(\underline{p})), \quad (5.14)$$

It is worth to mention that it is the structure of the mass matrix which is in charge of controlling the independence (diagonal) or the coupling (non-diagonal) between the momentum of different particles.

The correlation between these momentum variables can be crucial when tackling problems such as the kinematic source inversion problem where we know there is an important trade-off between parameters. In addition, the form of the kinetic energy is chosen to be quadratic due to the fact that this shape allows to use the Hamiltonian dynamics as a transition probability function (proposal engine) with an important feature known as *reversibility*. A reversible transition probability function implies that the constructed Markov Chain, based on this transition function, will also be reversible and stationary. These features are fundamental to infer any characteristic of the target probability function from the Markov Chain.

I would like to point out the following. Because the particle position \underline{q} is related to the model parameters of an inverse problem, the momentum \underline{p} can be seen as the vector of auxiliary variables that ease a faster transition of a Markov Chain towards the regions with denser probabilities. In other words, we have double the number of unknowns, but \underline{p} will serve us only to perturb \underline{q} according to the information of the local estimation of the gradient of the misfit function.

Integration of Hamiltonian system: The main idea of the HMC strategy is to use the Hamiltonian system to accelerate the model space exploration of the Markov Chain. This is performed by using the time evolution of the system as the engine to draw a new sample, that, then should be accepted or rejected according to the MH criterion. Therefore, it is crucial to make evolve in time the Hamiltonian system. To do this, the authors working in this domain suggest to use symplectic integrators¹. These numerical methods have, among some very useful characteristics, the property to keep (nearly) constant the total energy equal to the Hamiltonian expression. In fact, even though symplectic integrators are associated to Hamiltonian systems, a perfect energy conservation is not computationally mandatory or desired for a good exploration of a target function (e.g. see the detailed explanation provided by Neal et al., 2011). The simplest symplectic integration scheme, and the one I use through all this chapter, is the Leap-Frog integrator. Using this integration scheme, the system described by equations (5.9) and (5.10) is integrated in time in the following way,

$$\begin{aligned} p_i(t + \delta/2) &= p_i(t) - (\delta/2) \frac{\partial U(\underline{q}(t))}{\partial q_i} \\ q_i(t + \delta) &= q_i(t) + \delta \frac{p_i(t + \delta/2)}{m_i} \\ p_i(t + \delta) &= p_i(t + \delta/2) - (\delta/2) \frac{\partial U(\underline{q}(t + \delta))}{\partial q_i}. \end{aligned} \quad (5.15)$$

Before putting in practice the HMC, it is important also to highlight that many other strategies can be coupled with it, among them tempering trajectories (see Neal et al., 2011, and references therein). In addition, HMC can incorporate constraints in the space of solutions that is to be explored. Following Neal et al. (2011), these constraints can be included if the potential energy $U(\underline{q})$ is separated into different parts. For instance, if the parameters \underline{q} are to be found between prescribed lower and upper bounds, a new potential energy can be defined as,

$$U(\underline{q}) = U_*(\underline{q}) + C_r[m_{solution}](\Delta q_i), \quad (5.16)$$

¹A symplectic integrator is a numerical integration scheme suitable for Hamiltonian systems. Symplectic integrators form the subclass of geometric integrators which, by definition, are canonical transformations. They are widely used in nonlinear dynamics, molecular dynamics, discrete element methods, accelerator physics, etc.

where $U_*(\underline{q})$ is the previous potential function described in (5.12), and $C_r[\underline{m}_{solution}](\Delta q_i)$ can be any barrier function. For instance, one can use a function of the following form,

$$C_r[\underline{m}_{solution}](\Delta q_i) = \begin{cases} r^{r+1}(l_i - \Delta q_i)^r & \text{if } \Delta q_i < l_i \\ 0 & \text{if } l_i \leq q_i \leq u_i \\ r^{r+1}(\Delta q_i - u_i)^r & \text{if } \Delta q_i > u_i \end{cases}, \quad (5.17)$$

with

$$\Delta q_i = q_i - m_{isolution}. \quad (5.18)$$

In this case, l_i and u_i represent the lower and upper bounds of every q_i . In other words, they act as control quantities, while Δq_i is the distance from which the HMC is allowed to explore far from the given solution $\underline{m}_{solution}$. It is important to recall that the position is linked to the parameters of the current model which probability is being evaluated $\underline{q} = \underline{m}$. In equation (5.17), one can see that $\lim_{r \rightarrow \infty} C_r[\underline{m}_{solution}](\Delta q_i)$ is equal to zero for any $l_i \leq \Delta q_i \leq u_i$ and infinity for any $\Delta q_i > u_i$ or $\Delta q_i < l_i$. Moreover, for $r > 1$, $U(\underline{q})$ is differentiable, so we can still use it to define Hamiltonian dynamics. The new total Hamiltonian is then defined as,

$$H(\underline{q}, \underline{p}) = U_*(\underline{q})/2 + [C_r[\underline{m}_{solution}](\Delta q_i) + K(\underline{p})] + U_*(\underline{q})/2. \quad (5.19)$$

Before continuing, it is important to recall that the main goal of using this approach is to be able to assess the uncertainties associated to a given kinematic rupture model. Therefore, the function $C_r[\underline{m}_{solution}](\Delta q_i)$ plays a very important role regarding the allowed amount of perturbation to the solution $\underline{m}_{solution}$ to be considered during the HMC exploration.

The changes that the definition of this new Hamiltonian implies for the symplectic integration are further detailed in Neal et al. (2011). The important thing to keep in mind is that, function $C_r[\underline{m}_{solution}](\Delta q_i)$ acts as a barrier function. This very steep barrier has to be climbed if the value of Δq_i passes the threshold u_i or l_i . Because the barrier is too steep as $r \rightarrow \infty$, in practice while the system is being integrated, once the Δq_i passes the prescribed bounds, we fall down the the same q_i we had before passing the constrained boundary but with a negated momentum p_i which will prevent us to move again towards the barrier. This constrained Hamiltonian system is of great importance if we have in mind that in some cases, instead of making the attempt to find a solution to the kinematic source inversion problem together with the associated uncertainties, we would prefer to assess the uncertainties of a given solution in a restricted region around that solution. Therefore upper and lower limits for the slip-rate functions can be set using this constrained version of the HMC.

To end this section, as well as it is done in Section 5.3.1, the steps to perform a Hamiltonian Monte Carlo exploration are provided in the following algorithm.

Algorithm 4: Standard Hamiltonian Monte Carlo (HMC) algorithm.

```

input : Form of gradient of the potential function, the integration step  $\delta$ , and (optionally) the
        number of integration steps  $L$ , and the mass matrix  $\underline{M}$  used to propose the momentum;
        Set  $k = 0$ ;
        Get first state  $\underline{m}^{(k)}$ , from the prior probability function  $\underline{m}^{(k)} \sim P(\underline{m})$ ;
        while  $k < K$  do
            4.1: set  $k = k + 1$ ;
            4.2: sample a momentum variable from the momentum distribution  $\underline{p}^{(0)} \sim P(\underline{p})$ ;
            4.3: set  $\underline{q}^{(0)} = \underline{m}^{(k-1)}$ ;
            4.4: define a number of integration steps  $L$ 
            4.5: run Leap-Frog algorithm starting at  $(\underline{q}^{(0)}, \underline{p}^{(0)})$  for  $L$  steps and step size  $\delta$  to get proposed
                  state  $(\underline{q}^{(L)}, \underline{p}^{(L)})$ ;
            4.6: Estimate the ratio  $r = \frac{P(\underline{q}^{(L)}, \underline{p}^{(L)})}{P(\underline{q}^{(0)}, \underline{p}^{(0)})} \propto \exp(-U(\underline{q}^{(L)}) - K(\underline{p}^{(L)}) + U(\underline{q}^{(0)}) + K(\underline{p}^{(0)}))$ ;
            4.7: Draw a random number  $u$ , where  $u = \text{Unif}(0, 1)$ ;
            4.8: if  $u < r$  (Metropolis-Hastings acceptance criterion) then
                4.8.1: accept the proposal state  $\underline{q}^{(L)}$  and set  $\underline{m}^{(k)} = \underline{q}^{(L)}$ 
                else
                4.8.2: reject and set  $\underline{m}^{(k)} = \underline{m}^{(k-1)}$ ;

```

5.3.3 MCMC vs HMC: encouraging examples

In this section, I present two different examples where some of the advantages of the Hamiltonian Monte Carlo method are illustrated. For both examples, I compare the results obtained from the HMC approach to the ones coming from a standard MCMC strategy. These examples shall encourage us to use HMC to tackle high-dimensional problems such as the kinematic source inversion problem, that is discussed in the following section.

A 2D problem in seismology: Source spectrum estimation

The first example in this section is related to the inference of the two parameters that are used to describe the displacement source spectrum from Brune's model for a real data set. The inference of these two parameters and their uncertainties is carried out using two strategies: 1) a standard MCMC strategy as the one described in Section 5.3.1 and in Algorithm 3 and 2) the HMC approach as described in Section 5.3.2 and in Algorithm 4. The principal objective of this example is to show the applicability of HMC to seismological problems as well as to illustrate its faster convergence and exploration of the model space compared to a standard MCMC approach.

Problem setup: One parameter that has been used for long time to characterize the size of earthquakes and other important features related to these events (such as the stress-drop) is the corner frequency (f_c) of the source spectrum. One of the mathematical models that is commonly used in seismology to represent theoretically the displacement source spectrum is Brune's model (Brune, 1970),

$$u(f) = \frac{\Omega_0}{1 + (f/f_c)^n}, \quad (5.20)$$

or in logarithmic form,

$$\log(u(f)) = \log(\Omega_0) - \log(1 + (f/f_c)^n), \quad (5.21)$$

where the amplitude of the displacement spectrum u at a given frequency f can be determined based on the long-period amplitude (proportional to the seismic moment) Ω_0 , the corner frequency f_c and the high-frequency falloff rate n (i.e., $n = 2$ for ω^2 models). Ω_0 and f_c are very important source attributes that can be related to the stress-drop of an earthquake (Madariaga, 1976; Abercrombie, 1995; Shearer et al., 2006). Therefore, it is important to infer these attributes as well as their uncertainties.

For this first exercise, the goal is to retrieve the corner frequency f_c and the long-period amplitude Ω_0 , as well as the related uncertainties, of a real recording. For that, I use as dataset the raw (in gain) recording of a foreshock from the 2011 Tohoku earthquake. This event (according to the JMA Catalog) occurred the 10th of March 2011 at 6:23:59.75 (GMT Time) at 38.1722° North and 143.0448° East at a depth of approximately 9.3 km. I use the vertical component of receiver KNIH from the Hi-net, that recorded the particle velocity for that earthquake at approximately 200 km from the epicenter. I do not process the recording to account for any instrument correction. The recording was base-line corrected. The signal was cut to account only for the first 40 seconds, and the Fourier amplitude spectrum is estimated from a tapered version of that signal. This spectrum is then smoothed using the Konno-Ohmachi smoothing function (Konno and Ohmachi, 1998). The resulting spectrum is shown in Figure 5.1a (solid black line). The goal in this exercise is to compare the performance of the HMC against a standard MCMC.

In order to apply the standard MCMC strategy as detailed in Algorithm 3, the prior, likelihood and proposal probability distribution functions have to be defined. To do that, the likelihood is designed as a normal distribution function with a standard deviation $\sigma_{likeli} = 2$. Remember that the observed and predicted spectra are compared in a logarithmic scale. Therefore, the standard deviation is related to the logarithmic values of the spectra. On the other hand, I decide to use for each parameter (separately) a non-informative uniform distribution. For the corner frequency the minimum and maximum values of the uniform distribution are -2.3 and 1.6, which correspond to $f_{c_{min}} = \exp(-2.3) = 0.1$ Hz and $f_{c_{max}} = \exp(2.6) = 13.4$ Hz. For the long-period amplitude Ω_0 , the minimum and maximum values of the uniform distribution are set to 1 and 13 (logarithm of the amplitude in gain), which corresponds to $\Omega_{0_{min}} = \exp(1) = 2.7183$ and $\Omega_{0_{max}} = \exp(15) = 3.26 \times 10^6$. As it can be noticed, the exploration of the space is performed in a logarithmic space. Therefore, I define a 2D normal distribution centered at the current state and with a diagonal covariance matrix, $\text{diag}(\underline{\Sigma}_{proposal}) = [0.01, 0.01]^T$, as the proposal distribution function in the logarithmic space. For the MCMC strategy, the total number of iterations is fixed to 2000 iterations. Finally, the first guess is $f_c = 10$ Hz and $\Omega_0 = 1 \times 10^4$ (gain) (blue dashed line in Figure 5.1a). This initial guess is also used by the HMC approach that is detailed in the next paragraph.

In order to apply the HMC strategy to this problem, several things need to be established. The first is to define the potential function that is needed by HMC. For this problem, I decide to define it as the negative logarithm of an exponential whose argument is the L2 norm between the logarithm of

the observed, $y(f)$, and estimated, $u(f)$, amplitude spectra. By defining it in such way we obtain the following potential function,

$$\begin{aligned} U(\underline{q}) &= -\log \left(\exp \left(-\frac{1}{2} \left(\log(y(f)) - \log(u(f, [f_c, \Omega_0])) \right)^2 \right) \right) \\ &= \frac{1}{2} \left(\log(y(f)) - \log(u(f, [f_c, \Omega_0])) \right)^2, \end{aligned} \quad (5.22)$$

whose partial derivatives with respect to Ω_0 and f_c , that are also required by the HMC strategy, are

$$\begin{aligned} \frac{\partial U(\underline{q})}{\partial \Omega_0} &= -\frac{\log(y) + \log((f/f_c)^n + 1) - \log(\Omega_0)}{\Omega_0} \\ \frac{\partial U(\underline{q})}{\partial f_c} &= -\frac{n(f/f_c)^{(n-1)}(\log(y) + \log((f/f_c)^n + 1) - \log(\Omega_0))f}{((f/f_c)^n + 1)f_c^2}. \end{aligned}$$

Another important parameter to set for the HMC strategy is the mass matrix $\underline{\underline{M}}$. After some initial tests, and regarding the different variability that Ω_0 can have (ranging from 1×10^3 to 1×10^9 (gain)) compared to f_c (ranging from 1×10^{-1} to $1 \times 10^{1.7}$ Hz), I decide to define a smaller mass for Ω_0 than for f_c , so that Ω_0 can vary more freely,

$$\underline{\underline{M}} = \begin{bmatrix} 1 \times 10^{-4} & 0 \\ 0 & 1 \end{bmatrix}.$$

Finally, for the HMC it is necessary to define the time integration step ϵ and the number of steps of integration for the Leap-Frog scheme. For the integration step, I set a small value that prevents the divergence of the integration of this system ($\delta = 0.1$). For the number of steps, I decide to use a generator of random integer numbers (ranging from 1 to 40). Defining L as a random number of integration steps can be seen as another type of perturbation to the parameter exploration. A total number of 2000 iterations is also fixed for the HMC.

The results from this first exercise are summarized in Figure 5.1. According to the data misfit shown in Figure 5.1c and the comparison in Figure 5.1a, both strategies provide a correct estimation of both source attributes. Similar means and covariance matrices are obtained from both strategies. According to the MCMC and HMC explorations, the parameters follow a lognormal probability distribution function. The MCMC reports the mean values as $f_{c,mean}^{MCMC} = \exp(1.1953) = 3.3046$ Hz, $\Omega_{0,mean}^{MCMC} = \exp(13.720) = 9.0886 \times 10^5$ (gain), while the HMC estimates $f_{c,mean}^{HMC} = \exp(1.1931) = 3.2974$ Hz, $\Omega_{0,mean}^{HMC} = \exp(13.709) = 8.9938 \times 10^5$ (gain). However, the covariance matrices of the two lognormal distributions that are found, showed that the results from the standard MCMC strategy are more dispersive than the ones from the HMC: feature that is consistent with the behavior observed in Figure 5.1c.

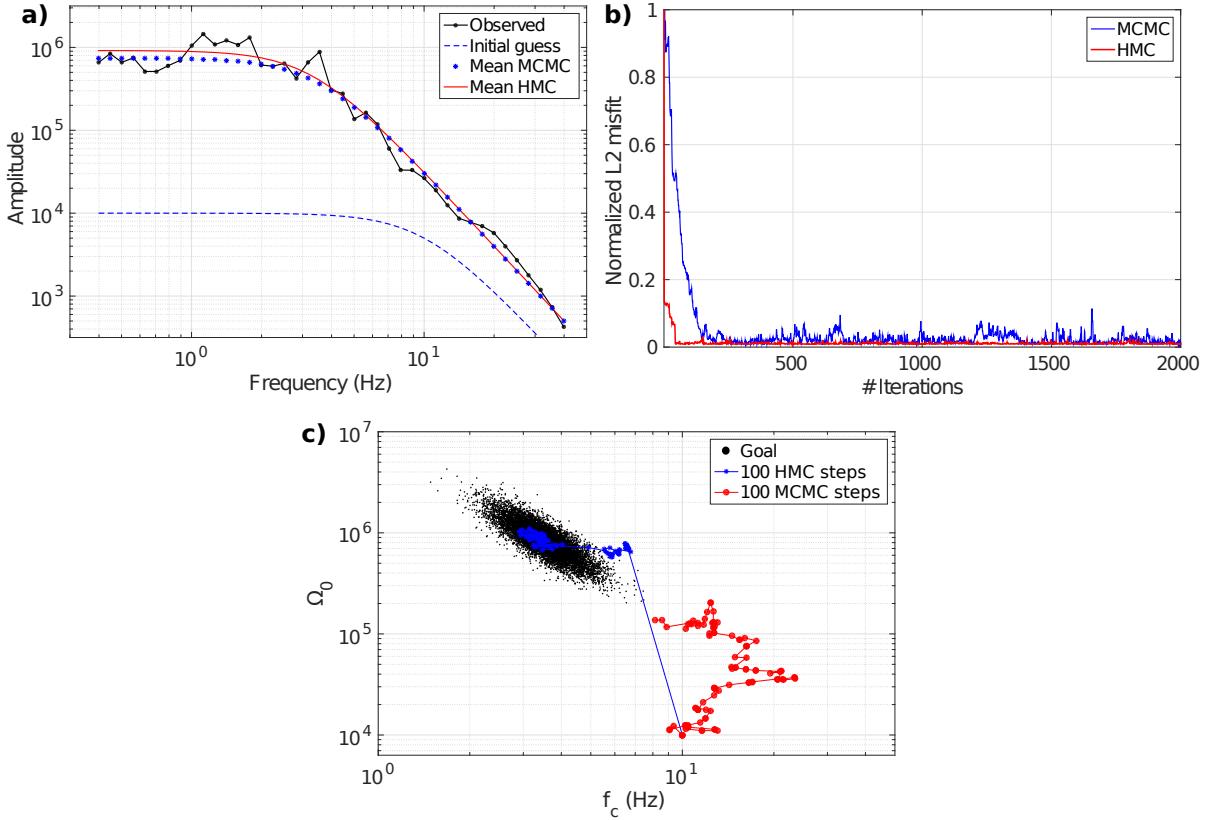


Figure 5.1: Summary of the results from Brune's model exercise. a) compares the observed displacement spectrum (black line) with the initial guess (dashed blue line) and the two spectra estimated based on the mean values of Ω_0 and f_c that were obtained through the MCMC (blue dotted line) and the HMC (red solid line) strategies. b) illustrates the evolution of the normalized data misfit through the iterations of the two approaches. It can be noticed that the burn in phase of the HMC is approximately ten times faster than the MCMC for this 2D example. For this exercise the decay is set as $n = 3$. c) shows the evolution of the two Markov Chains obtained. The first 100 samples of the chains are plotted for both strategies (MCMC in red while HMC is in blue). Notice that in less than 100 steps HMC arrives to the region of denser probability and performs its exploration, while the MCMC has not yet arrived to this region of interest.

$$\text{cov}(\Omega_0, f_c)^{MCMC} = \begin{bmatrix} 0.0426 & -0.0148 \\ -0.0148 & 0.0107 \end{bmatrix}, \quad \text{cov}(\Omega_0, f_c)^{HMC} = \begin{bmatrix} 0.0322 & -0.0125 \\ -0.0125 & 0.0079 \end{bmatrix}.$$

One of the important results to highlight from this exercise is the faster convergence of the HMC to the region where the most probable scenarios are located in the model space. This faster exploration is represented in Figure 5.1b, where the first 100 members of the constructed Markov chains (MCMC and HMC) are plotted for both strategies. In that figure one can see that, while the MCMC (blue line) has not yet arrived to the dense probable region, the HMC (red line) is already exploring it. This feature can be also noticed in Figure 5.1c, where the normalized misfit function of the HMC (red line) strategy converges much faster than the one for the MCMC (blue line).

A multimodal case

Another well-known difficulty for a standard MCMC (and other techniques) is to deal with multimodal functions. These type of functions mean that, in the space of probable solutions, there might be more than one region with dense probabilities. In other words, more than one solution is likely. The purpose of this second exercise is to compare the performance of the exploration of the model space using the HMC and the standard MCMC in a complex multimodal case. To do that, I define the following multimodal probability distribution as the target function to be described,

$$P(\underline{m}|\underline{d}) = \frac{0.7 \exp(-(x_1^2 + y_1^2)) + 1.3 \exp(-(x_2^2 + y_2^2)) + 2 \exp(-(x_3^2 + y_3^2))}{11.8679} \quad (5.23)$$

with

$$\begin{aligned} x_1 &= (x - 2)/0.5; & y_2 &= (y - 1.8)/0.7; \\ y_1 &= y/2; & x_3 &= (x + 7)/1.2; \\ x_2 &= x/1.8; & y_3 &= (y + 7)/0.6; \end{aligned}$$

In this exercise, three exponential functions are summed. Therefore, for this problem there are three regions with dense probabilities: 1) at $(x_1 = 2, y_1 = 0)$, 2) at $(x_2 = 0, y_2 = 1.8)$ and 3) at $(x_3 = -7, y_3 = -7)$, as it can be seen in Figure 5.2a.

For the standard MCMC, the setting is simple. I use as proposal function a 2D normal distribution function which mean vector is the current state of the Markov Chain and the covariance matrix is taken as the identity. The estimation of $P(\underline{m}|\underline{d})$ for the current and proposed state is performed as a simple evaluation of equation (5.23). The initial guess for both strategies is $[x, y] = [0, -8]$. The total number of iterations is set to 200,000, in order to ensure a complete exploration of the space of solutions.

In order to apply the HMC strategy, it is necessary to first define a potential function. This function is taken as $U(\underline{q}) = -\log(P(\underline{m}|\underline{d}))$, where $P(\underline{m}|\underline{d})$ is given by equation (5.23). The gradient of the potential energy function, also required by the HMC, is analytical,

$$\begin{aligned} \frac{\partial U(\underline{q})}{\partial x} &= -2 \left(\frac{25x + 175}{18} \right) \exp \left(-\left(\frac{5x + 35}{6} \right)^2 - \left(\frac{5y + 35}{3} \right)^2 \right) \\ &\quad - \frac{7(8x - 16)}{10} \exp \left(-(2x - 4)^2 - \left(\frac{y}{2} \right)^2 \right) - \frac{65x}{81} \exp \left(-\left(\frac{10y - 18}{7} \right)^2 - \frac{25x^2}{81} \right), \end{aligned} \quad (5.24)$$

$$\begin{aligned} \frac{\partial U(\underline{q})}{\partial y} &= -2 \left(\frac{50y + 350}{9} \right) \exp \left(-\left(\frac{5x + 35}{6} \right)^2 - \left(\frac{5y + 35}{3} \right)^2 \right) \\ &\quad - \frac{13}{10} \left(\frac{200y - 360}{49} \right) \exp \left(-\left(\frac{10y - 18}{7} \right)^2 - \frac{25x^2}{81} \right) - \frac{7y}{20} \exp \left(-(2x - 4)^2 - \left(\frac{y}{2} \right)^2 \right). \end{aligned} \quad (5.25)$$

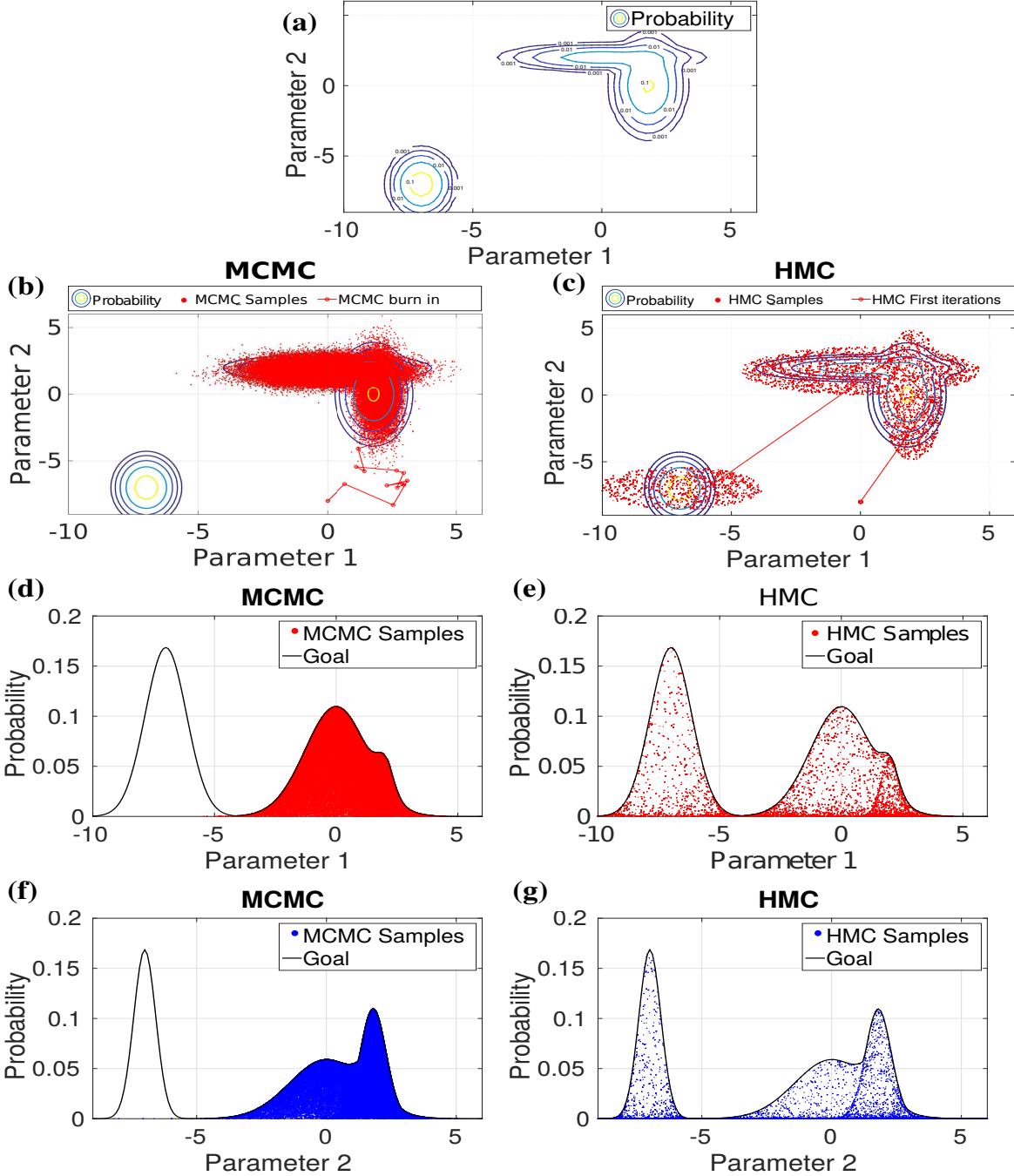


Figure 5.2: Summary of the results from the exercise that explores a multimodal 2D probability function. (a) Illustrates a 2D map of the multimodal probability function (sum of three exponential functions). (b) and (c) show the evolution of the Markov Chains that are obtained using the MCMC (200,000 samples red dots in (b)) and the HMC (30,000 samples red dots in (c)) strategies. Notice that in the first 30 samples of the HMC exploration (red line in (c)) the HMC was able to explore the three denser regions of probability, while the MCMC (red line in (b)) has not yet arrived to the region where it gets stuck. (d) and (f) illustrate in a profile view the exploration that the standard MCMC strategy does after 200,000 iterations. (e) and (g) are the same as (d) and (f) but for the HMC strategy after 30,000 iterations.

For this exercise, the integration step and number of steps for the Leap-Frog scheme are set to $\delta = 0.1$ and $L = 20$. These values are not necessarily optimal, since I did not perform an exhaustive exploration of how to set these parameters for this example. The total number of iterations for the HMC is set to 30,000. The mass matrix is taken as the identity and the probability distribution from which the momentum parameters are sampled is a 2D normal distribution with zero mean. The results from this exercise are summarized in Figure 5.2.

The results from this second exercise illustrate another advantage of the HMC strategy over the standard MCMC approach. As it can be seen in Figure 5.2, HMC is able to explore the three regions with dense probabilities. On the contrary, the MCMC is not able to get away from the first region where it arrived after its burn in phase. Several test were performed, and the results were always the same: either the third region at $(x_3 = -7, y_3 = -7)$ is sampled but not the other two, or the first two are sampled but the third is not. The MCMC approach is not able to move across the three regions because one of them is too far from the other two. As it can be seen in Figure 5.2, MCMC is able to explore the other two regions, because the distance between them is not so large as the one separating them from the third. Certainly, there are other MCMC strategies that could tackle this problem in a better way. Another possibility to overcome this problem using the same MCMC approach would be to set a larger perturbation in the proposal function instead of using the identity as the proposal covariance matrix. The fact is that, without a complex tuning of the parameters involved in the HMC strategy (e.g. step size, number of steps, mass matrix), the HMC strategy is able to visit the three regions and to provide a correct description of the probability function we are interested in even after very few iterations (30 iterations as shown in Figure 5.2c).

Before applying the HMC technique to the kinematic source inversion problem I would like to point out that the application of this method to the two exercises previously presented belong to the most complex approach to each of this problems: the attempt is to retrieve the solution to an inverse problem together with the associated uncertainties. Once again, another approach would be to only investigate the uncertainties associated to a given solution of an inverse problem. Even though the results from these first two exercises demonstrate that the HMC technique is good at tackling problems under the difficult approach, it is important to point out that in more complex problems (e.g. the source inversion problem) the second approach should be preferred.

5.4 HMC and the Kinematic Source Inversion Problem

Contrary to the goal of Chapters 3 and 4, the purpose of this chapter is not only to provide a rupture model that explains the observations, but to assess also the uncertainties associated to that solution. Therefore, instead of only providing one single rupture model that explains, to some extent, the observed wavefields; my goal is to provide a description of the *posterior probability density (ppd)* of a set of parameters, describing the seismic rupture, given the observed evidence (e.g. seismograms, static displacements, etc.). As I mentioned in the beginning of this chapter, the HMC technique can be used to achieve two different goals: 1) obtain a solution to the kinematic source inversion problem and its related uncertainties or 2) assess the uncertainties in the vicinity of a given solution. The preliminary investigation presented in this section is mainly focused on the second goal. Under that approach, it is important to mention that the given solution around which we can describe the related uncertainties can come from the results of the deterministic approach that is presented in the previous chapters, as well as from other strategies. The answer to the first goal, which seems more challenging, is also an interesting approach to be explored.

As it was discussed in the previous chapters, the linear or non-linear formulations of the kinematic

source inversion problem involves a large number of parameters (tens of thousands). Therefore, if we want to keep the linearity of the problem - feature that must be promoted for several reasons (e.g. linear propagation of errors, quadratic convex shape of the misfit function) - we should be aware of the difficulty that this feature implies for the uncertainty assessment. As mentioned before, one of the recent strategies that has been proved to be suitable to tackle such large dimension problems is the Hamiltonian Monte Carlo strategy. Therefore, in this section I investigate the suitability of the HMC strategy to assess the uncertainties of the kinematic source inversion problem under the linear time-domain formulation that was presented in the previous chapters.

5.4.1 Resetting the kinematic source inversion problem

Because there is no closed form of the *ppd* for the source inversion problem, I expect to construct a large enough Markov Chain which stationary distribution tends to the desired *ppd* of the model parameters for a given solution. Therefore, my goal is to describe $P(\underline{m}|\underline{d})$, which is the target *ppd* of the vector of model parameters (\underline{m}) given the observations (\underline{d}). In this case, the model parameters represent the time-space slip-rate history $\underline{V}(\xi, \tau)$ which is discussed in Chapters 3 and 4, while the observations are the particle velocity recordings $\underline{v}(x, t)$. In the same sense, the likelihood function from equation (5.2), $P(\underline{d}|\underline{m})$, is the conditional probability related to the forward problem. In addition, $P(\underline{m})$ is the marginal prior probability that takes into account all the prior information that we can have about the expected source model. It is important to mention that $P(\underline{m})$ is the term in charge of injecting the physics into the linear kinematic source inversion formulation, as it is done by the regularization term used in the previous chapters. In fact, as discussed by Figueiredo (2003); Aster et al. (2004); Tarantola (2005) and several other authors, considering a likelihood function which includes (besides the data misfit term) a L2 model regularization term, as the one used in previous chapters, is equivalent to assume an independent data misfit likelihood function being multiplied by a normal prior distribution. Therefore, I focus my investigation more on the description of the likelihood function, assuming the following,

$$P(\underline{m}|\underline{d}) \propto \underbrace{P^*(\underline{d}|\underline{m})}_{\text{data likelihood term}} \quad \underbrace{P^*(\underline{m})}_{\text{prior model term}} \propto \underbrace{P(\underline{d}|\underline{m})}_{\text{data + L2 regularization}} \quad (5.26)$$

where $P^*(\underline{d}|\underline{m})$ and $P^*(\underline{m})$ are the independent likelihood and prior *pdfs* respectively, while $P(\underline{d}|\underline{m})$ is the likelihood *pdf* which already incorporates the prior information as a L2 model regularization term.

Because my goal is to use the HMC technique to assess the uncertainties associated to a given solution of the kinematic source inversion problem, it is necessary to first define a potential energy function. In terms of the source inversion problem, this potential energy is set as a negative logarithm of an exponential function whose argument is the regularized misfit function. This misfit function is in charge of measuring the difference between the observed and the synthetic seismograms (in a least-squares sense). The regularization term of that misfit function injects any available prior information that can be used to constrain the problem. The time-space slip-rate history on the prescribed fault geometry represents the model parameters $\underline{m} = \underline{V}(\xi, \tau)$, while the target *ppd* is the set of physically plausible rupture scenarios surrounding a given solution which explain, to some extent, the observations $\underline{d} = \underline{v}(x, t)$. Therefore, the potential energy can be defined as

$$U(q) = -\log(P(\underline{m}|\underline{d})) \propto -\log(P(\underline{d}|\underline{m})) = -\log(\exp(-C(\underline{V}))) = C(\underline{V}). \quad (5.27)$$

From equation (5.27), it can be seen that the potential energy of the Hamiltonian system, to be used to explore the space of solutions, is exactly the same as the expression of the misfit function $C(\underline{V})$ that is

used in Chapter 3 and 4. Therefore, under the HMC approach,

$$\begin{aligned}
 U(\underline{q}) &= -\log(P(\underline{m}|\underline{d})) \\
 &\propto -\log(P(\underline{d}|\underline{m})) \\
 &\propto -\log\left(\exp\left(-\frac{1}{2}\sum_{\underline{x}}\int_t \Delta\underline{d}[\underline{V}]^T \underline{\underline{W}}_d^T \underline{\underline{W}}_d \Delta\underline{d}[\underline{V}] dt\right)\right) \\
 &\quad -\log\left(\exp\left(-\frac{\epsilon}{2}\int_{\xi}\int_{\tau} \Delta\underline{m}[\underline{V}]^T \underline{\underline{W}}_m^T \underline{\underline{W}}_m \Delta\underline{m}[\underline{V}] d\tau d\xi\right)\right)
 \end{aligned} \tag{5.28}$$

where it is possible to identify from equation (5.28), the data misfit term (equation (3.11)) and the model misfit term (equation (3.13)), which are exhaustively discussed in Section 3.2.2. As a brief summary, $\Delta\underline{d}[\underline{V}]$ represents the residuals between the observed, $\underline{u}(\underline{x}, t) = [u_x(\underline{x}, t), u_y(\underline{x}, t), u_z(\underline{x}, t)]^T$ and the calculated seismograms, $\underline{v}[\underline{V}](\underline{x}, t) = [v_x[\underline{V}](\underline{x}, t), v_y[\underline{V}](\underline{x}, t), v_z[\underline{V}](\underline{x}, t)]^T$, where $[\underline{V}]$ is used to denote the dependence of $\Delta\underline{d}$ on a given rupture model \underline{V} , while $\Delta\underline{m}[\underline{V}]$ represents the model residual between a current rupture model $\underline{V}(\xi, \tau)$ and the prior model $\hat{\underline{V}}(\xi, \tau)$. In addition, $\underline{\underline{W}}_d$ and $\underline{\underline{W}}_m$ are used to denote the data and model weighting matrices respectively. Notice that, both matrix multiplications can be directly linked to the covariance of the probability functions, being $\underline{\underline{W}}_m^T \underline{\underline{W}}_m = \underline{\underline{C}}_m^{-1}$ and $\underline{\underline{W}}_d^T \underline{\underline{W}}_d = \underline{\underline{C}}_d^{-1}$.

Equations (5.27) and (5.28) imply that the gradient of the chosen potential energy is also the same as the gradient of the misfit function (equation (3.18)) that is presented in Section 3.2.2,

$$\begin{aligned}
 (\nabla_{q_i} U(\underline{q}))_k &= \mathcal{G}_k[\underline{V}](\xi, \tau) \\
 &= \sum_{\underline{x}} \int_{t_1}^{t_2} \underline{\underline{W}}_d^T \underline{\underline{W}}_d \Delta d_n[\underline{V}](\underline{x}, t) \mathcal{T}_{ni}(\xi, \tau - t; \underline{x}, 0) \mathcal{P}_{ik}^T dt \\
 &\quad + \epsilon [\underline{\underline{W}}_m^T \underline{\underline{W}}_m \Delta m[\underline{V}](\xi, \tau)]_k,
 \end{aligned} \tag{5.29}$$

$n, i \in [x, y, z]$ and $k \in [\phi, \delta]$ or $k \in [\lambda]$,

where I use the same index and Einstein summation convention as for equations (3.7), (3.16) and (3.18). As in equation (3.18), the first term in equation (5.29) is the data term (or likelihood *pdf*), which is the sum of the projections of the unexplained residues at all the receivers onto the fault plane through the stress-state tensor, while the second term incorporates the prior model information (prior *pdf*). In this case where we are interested in describing the *ppd* around a given solution, it is the prior model which takes the role of that given solution, which can come from the deterministic approach previously presented.

From equations (5.27) and (5.29) we can see the advantage of choosing the potential energy described in equation (5.28). It is possible to use all the development presented in the previous chapters to assess the uncertainties of the proposed rupture models using the HMC strategy.

A standard MCMC scheme is based on three principal steps. However, the definition of the proposal transition function, which is required by that scheme, appears to be not a easy engine to build in order to assess the uncertainties associated to the source reconstruction problem (under the linear formulation where thousands of unknowns have to be perturbed simultaneously). Certainly, this proposal transition function can be designed in terms of perturbations of physical parameters of the expected shape of the slip-rate time histories, rupture velocity, rake angle, etc. However, somehow, this model perturbation bring us back to the non-linear formulation of the problem. On the contrary, the HMC strategy has the

ability to move rapidly through the high-dimensional model space based on perturbations to the model that take into account the gradient of the misfit function, that is computed with respect to every single time and space slip-rate sample describing the rupture history. I shall recall the reader that, thanks to the precomputed stress-state tensors that are presented in Chapter 2, the computation of the gradient is a rather efficient operation.

As mentioned in Section 5.3, the HMC also requires to define a mass matrix. The inverse of this matrix, as pointed out by several authors (Neal et al., 2011; Betancourt, 2017), and as it is briefly described in the first example from Section 5.3.3, acts as the covariance matrix of the *pdf* used to draw the auxiliary momentum variables that serve us to perturb the unknown parameters. Therefore, the definition of its structure is directly linked to the speed in which the exploration of the desired *ppd* is performed. Some authors, for instance Fichtner et al. (2018), propose that an optimal mass matrix for linear forward problems with Gaussian priors (least-squares misfit functions) can be the inverse posterior covariance matrix, $\underline{M} = \underline{C}_{\text{posterior}}^{-1}$. In other words, an optimal choice of the inverse of the mass matrix is the Hessian. However, for high-dimensional problems, the access to the exact inverse of the covariance matrix, can be intractable. Nevertheless, strategies such as lBFGS can offer an approximation estimate of the Hessian through an iterative optimization process (Nocedal and Wright, 2006). Furthermore, knowing beforehand the posterior covariance matrix turns meaningless the purpose of performing the sampling. Other studies, such as Biswas and Sen (2017) or Sen and Biswas (2017), assume as mass matrix the identity, and leave without tuning this matrix that could help improving the efficiency of the exploration. This rather simple assumption might be due to the complexity that the tuning of this matrix can represent.

In this work, I propose a block diagonal mass matrix (symmetric positive definite), whose structure is based on the estimated gradient of the initial guess and a 2D (e.g. time and space) correlation distance between parameters. The value of each element in the matrix has to be assigned according to the absolute value of the gradient with respect to that parameter and its correlation with its neighbors. It is very important to point out that, in this sense, there are two possible scenarios that depend on the quality of the illumination of the seismic source given the monitoring receiver acquisition configuration. The first is related to good or perfect illuminations (not realistic scenarios). In this scenario, the elements of the mass matrix that correspond to parameters with a large absolute value of the gradient, have to take small values, which will promote large perturbations and a faster exploration. On the contrary, very large values will be given for the elements of the matrix that correspond to small absolute values of the gradient (i.e. the exploration is freezed for these parameters). The second case, and the more common one, is when there is lack of illumination. In that case, the mass needs to be large for the parameters with relative high gradient values, while small mass values have to be given to the parameters with relative low gradient values. This important change of the definition of the mass matrix when there is lack of illumination is due to the fact that we are interested in exploring more rapidly the places where the gradient is low and where the uncertainties are expected to be large. Finally, by defining the mass matrix in terms of the first initial gradient can be seen as a prescribed constrain to the exploration: depending on the values of the first gradient some parameters are permanently forced to be largely perturbed, while others remain freezed.

5.4.2 HMC kinematic source inversion examples

The two examples presented in this section are focused on the understanding of the HMC exploration in a synthetic configuration with simple geometry and assuming perfect illumination. More complex

ruptures with realistic source-receiver geometries remain as perspectives of my research work and are not part of what is presented here.

A simple line source well illuminated

For the following two exercises, the source-receiver geometry is composed by a line source located at a depth of $z = 2$ km. The line source is represented by 21 spatial nodes distributed every 0.5 km along the x -direction, from $x_{min} = 12$ km to $x_{max} = 22$ km, with a constant location along y -direction. 420 receivers monitoring this line source are deployed surrounding it. Every fault spatial node is surrounded by 20 receivers (following a circular array). The distance between each spatial fault node and the 20 receivers around it is of 2 km. This type of idealized tube-like source-receiver geometry is illustrated in Figure 5.3. The velocity-density structure is a homogeneous half-space with $\alpha = 4.8$ km/s, $\beta = 2.6$ km/s and $\rho = 2.3$ g/cm³, as the P and S wave velocities and density respectively.

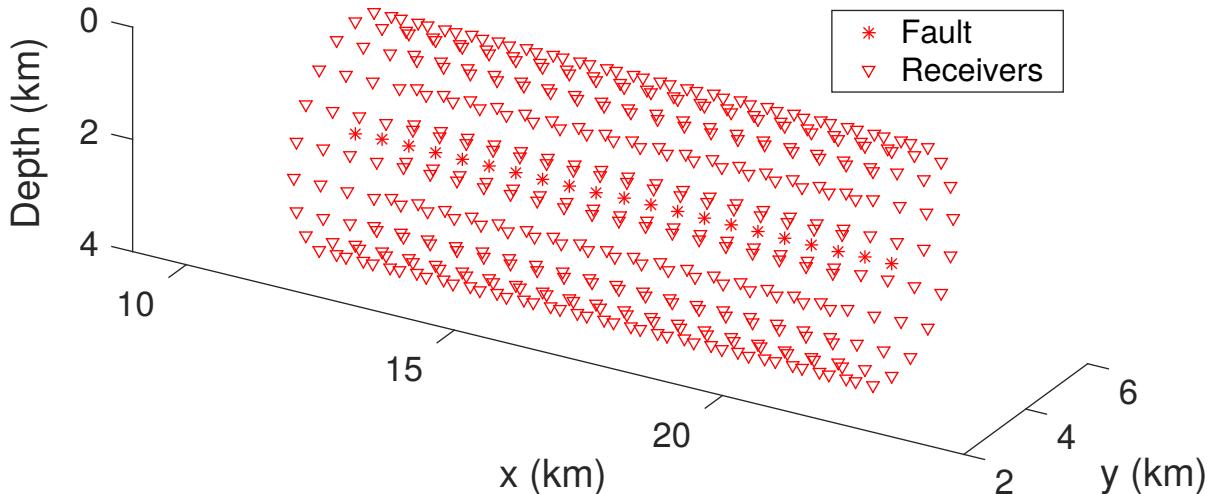


Figure 5.3: Idealistic tube-like source-receiver geometry assumed for the first two examples presented in Section 5.4.2. 21 spatial nodes (asterisks) separated every 0.5 km along the x -direction (from $x_{min} = 12$ to $x_{max} = 22$ km) at a constant depth of $z = 2$ km and at $y = 4$ km formed the line source to be monitored. 420 receivers (triangles) surrounding the line source represented the acquisition system. Every spatial node is surrounded by 20 receivers equally separated from the fault node by 2 km (forming circular arrays around the nodes).

A spike test

The goal of this first exercise is to assess the uncertainties of a simple spike test. For that purpose, I assume that, at the center of the line source (fault node 11 at $x = 17$ km) a sudden slip-rate impulse takes place. Therefore, $V(\xi, \tau) = 1$ m/s at $\xi = [17, 4, 2]$ km and at $\tau = 1$ s. This sudden slip-rate impulse is illustrated in Figure 5.4a. The focal mechanism is fixed to strike 90°, dip 90° and rake 90° for all the nodes on the fault. The time-space slip history is fixed to last 10 s and to have a sampling rate of $\Delta t = 0.25$ s. Therefore, the total number of unknowns (dimensions where to search) is 41 [time samples] \times 21 [fault nodes] = 861 [unknowns]. The exact solution to this problem requires that the unknown #415 is equal to 1.0 m/s while all the other remaining unknowns are equal to zero. Using all this information, the 3 components (E-W, N-S and U-D) at the 420 receivers are estimated and then

they are assumed as the available observations. Notice that these "observed" recordings are noise-free and that the velocity-density structure is perfectly known.

In this example, I do not investigate uncertainties related to the source geometry, or the influence that noisy data can have on the source reconstruction. My main focus is related to assess the uncertainties inside this fixed time-space-amplitude slip-rate grid. Because, the illumination is perfect for this case, and the source is rather simple, I do not use any prior model information (the model term in equation (5.27) is equal to zero) as well as the hyper-parameter ϵ . Therefore, in this case, the desired ppd is assumed to be proportional to the data likelihood. Knowing the form of the potential energy (equation (5.28)) and its gradient (equation (5.29)), the only missing terms that need to be defined before running the HMC exploration are the integration step δ , the number of steps L and the mass matrix.

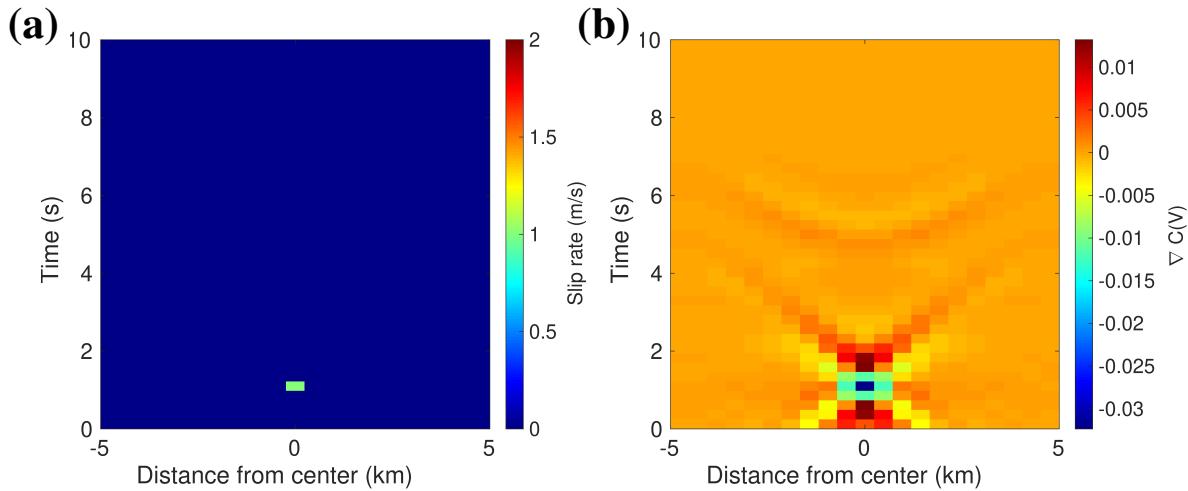


Figure 5.4: Illustration of the spike test model and gradient of the misfit function at first iteration. (a) illustrates a slip-rate sudden impulse ($V(\xi, \tau) = 1$ m/s) at the center of the source array ($\xi = [17, 4, 2]$ km) from Figure 5.3 at $\tau = 1$ s. (b) illustrates the gradient of the misfit function at the first iteration. The misfit is computed taking into account the 3 components of the 420 receivers. The initial guess is a rupture model with no slip-rate values anywhere. It can be seen in (b), that the gradient is pointing out where the slip-rate from the first guess has to be incremented to better explain the data (yellow region in (b)). The elongated artifacts are due to the low frequency approximation of the source.

Due to the rather simple source time-space history to be reconstructed, δ , L and \underline{M} do not required a very sophisticated tuning to ensure a correct behavior of the exploration and convergence to the desired ppd . The integration step is chosen to be $\delta = 0.1$, which ensures a correct integration of the Hamiltonian system. The number of integration steps is randomly selected for each HMC iteration from a uniformly distributed probability distribution of integer numbers ranging from 1 to 100 steps. The first guess, slip-rate values at the first HMC iteration, is zero slip-rate values everywhere.

According to this first guess, the gradient of the misfit function is computed (see Figure 5.4b). Then, the mass matrix is determined based on the gradient from the first guess. Thanks to the information provided by the gradient, I decide only to perform an exploration of the parameter #415 (out of the 861 unknowns), where the gradient presents the maximum absolute value. Therefore, because I decide not to perturb other parameters, the mass matrix is designed as a diagonal matrix, with very large values for the elements that are not the #415 and the unit for the parameter #415. In other words, this mass matrix allows the HMC strategy to explore the dimension related to the parameter #415 and the other unknowns remain freezed (decision based on the initial guess). The way this mass matrix is designed

can be understood in the scope that was previously defined: assess the uncertainties related to a given solution. In this case, thanks to the gradient from the first guess, we know that the most probable solution is to have an important slip-rate value at unknown #415 and the chosen mass matrix will help us to determine the *ppd* associated to that solution. It is important to mention that this simple design is due to the shape of the source to be reconstructed as well as the good illumination, circumstances that are usually not common in realistic problems.

The results from this first exercise are summarized in Figure 5.5. Thanks to the restrictions imposed by the mass matrix (which is designed based on the gradient), the evolution of the Hamiltonian system (illustrated in Figure 5.5a) is well-behaved and the particle trajectories resemble concentric circles which center is the expected slip-rate value ($q = 1, p = 0$). It is exactly at this point where the probability is maximum as the slip-rate value takes the corresponding value of the solution and the momentum is equal to zero (no need to perturb more the parameter).

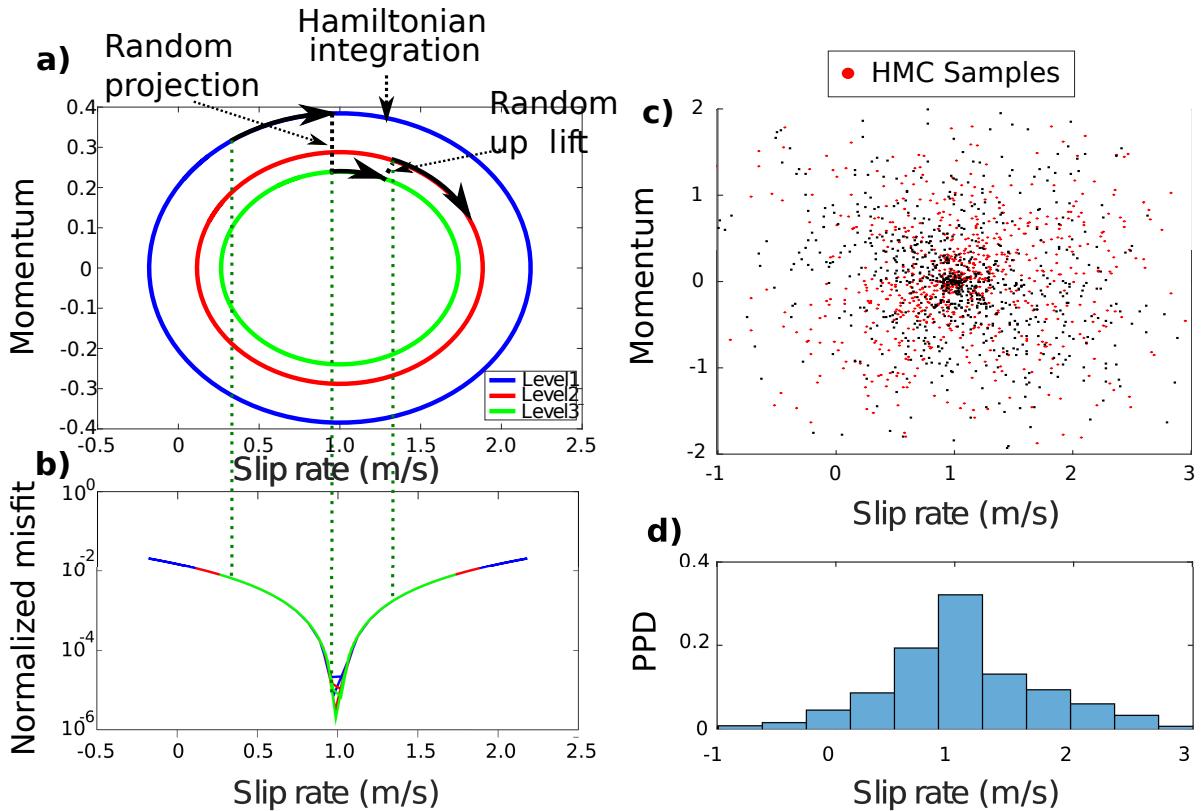


Figure 5.5: Summary of the results from the HMC exploration related to the spike test. a) illustrates the random projection and uplift produced by the auxiliary momentum variable. In a) the duration of the Hamiltonian integration is represented by the length of the arc circles along the contour lines representing different energy levels. b) shows how the trajectories of the three levels of energy from a) map the misfit function (convex quadratic). c) illustrates the samples forming the constructed Markov Chain after 1,000 iterations. In c) the black dots represent the beginning and the end of the Hamiltonian trajectories that are accepted to form part of the Markov Chain after passing the acceptance/rejection criterion. d) represents the *ppd* inferred from the HMC exploration for the slip-rate values of parameter #415. Notice that, ignoring the part related to the momentum variable from c) we can infer the *ppd* of the slip-rate which is shown in d).

It is very interesting to see how the perturbation introduced by the auxiliary variable allows to map the geometry of the misfit function even when random projections and uplifts to different energy levels are done (Figure 5.5a,b). The energy levels (solid colored lines in Figure 5.5a) represent the constant value of the Hamiltonian (total energy of the system) for a given value of the slip-rate and momentum for the particle #415. During the evolution of the system, which starts at some initial slip-rate value with a random momentum, the Hamiltonian remains constant and, as a consequence, the balance between the slip-rate and the momentum creates the circular trajectory around the solution. After, 1,000 iterations of the HMC exploration (keeping only the starting and ending points of the Hamiltonian trajectories) together with the MH criterion we obtain a Gaussian-shape *ppd* that can be seen in Figure 5.5c. This Gaussian-shape is in agreement with the quadratic convex shape of the misfit function that we expected and that is shown in Figure 5.5b.

It is important to highlight that in reality the HMC requires very few iterations to solve this simple problem (there is only 1 unknown). For instance, after 500 HMC iterations I obtain already a mean equal to 1.0006 m/s with a $\sigma = 0.64$ m/s for this problem. The 1,000 iterations illustrated in Figure 5.5 were carried out only for illustration and research purposes.

The results from this simple exercise help us to validate the reformulation of the problem that I presented in the previous chapters and to get familiar with the behavior of the HMC exploration for this problem. This spike test serve us only for two main purposes: 1) to validate the formulation and numerical methods and 2) to assess the amplitude uncertainty of a well-located point source. Let me then move on to a more complex exercise.

A bilateral rupture

This exercise is several steps further in complexity from the previous exercise. In the previous example the ambiguity related to time and space was discarded thanks to the information provided by the gradient at the first iteration (based on a good illumination). In this second exercise I keep the same source-receiver configuration but instead of assuming a punctual impulsive source as the target, I assume a small bilateral rupture. This rupture starts at the center of the line source and propagates to both sides of the line source. The total duration of this source is 2.5 seconds. Only three out of the 21 spatial nodes (the central one and two nodes beside) exhibit significant slip-rate values (see Figure 5.6a). The slip-rate time history of each of the three fault nodes involved in this rupture are exactly the same: each time history is just delayed according to a rupture velocity of 2 km/s. The representation of this synthetic bilateral rupture is shown in Figure 5.6a. In total, this rupture represents 8 [time samples] \times 3 [spatial nodes] = 24 [unknowns] out of the total 861 dimensions of the model.

Let me recall the two different approaches where the HMC technique can be used: 1) to obtain a solution and the associated uncertainties or 2) to assess the uncertainties of a given solution. In this exercise, I propose four different configurations where these two different approaches can be illustrated. The first configuration assumes a simple diagonal mass matrix whose elements in the diagonal are either large ($1e10$) or small (equal to 1). The elements which values are small are the ones related to the 24 unknowns that are easy to identified thanks to the values of the first gradient (due to the good illumination and the noise-free data). The other parameters are set to have very large mass values. For the other two exercises, the inverse of the mass matrix (covariance of the momentum probability distribution function) is designed based on a 2D anisotropic Gaussian-shape correlation function (as described in equation (4.1)), which block diagonal structure allows to take into account an expected time and space coherence of the rupture model (see Figure 5.8c). Therefore, for these two HMC explorations the 24 auxiliary momentum variables related to the 24 unknown are drawn from a probability

distribution which covariance matrix is linked to that mass matrix. For the other remaining dimensions the mass is assumed to be very large ($1e10$). The integration step is set as $\delta = 0.1$ and the number of integrations steps is selected from a uniform distribution of integer numbers ranging from 1 to 50 for the four HMC runs. The initial guess for the four runs is the same, a rupture model with zero slip-rate values everywhere. The total number of HMC iterations is set as 2700 for the all explorations. Notice that the first two explorations belong to the first approach: a solution to the problem and the associated uncertainties have to be determined.

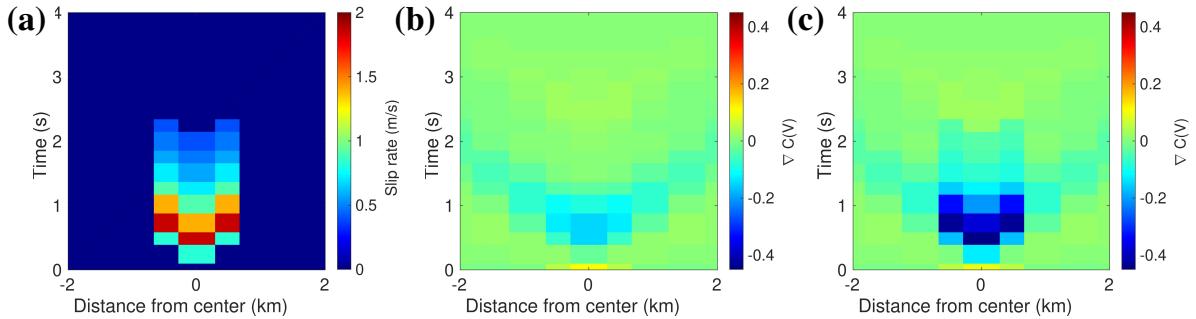


Figure 5.6: Illustration of the a) synthetic bilateral rupture model, and the gradient of the misfit function at the first iteration b) without considering prior information and c) considering it. The gradients here shown are computed taking into account the 3 components of the 420 receivers. The initial guess is a rupture model with no slip-rate values anywhere. The model used as a prior model (affecting the gradient) is shown in Figure 5.8b. It can be seen in (b), that the gradient is pointing out where the slip-rate from the first guess has to be incremented to better explain the data. In c) the shape of the gradient is improved and its shape is much closer to the expected rupture model.

On the contrary, the third and fourth configurations belong to the second approach: the assessment of the uncertainties have to be determined based on a given solution. For the third HMC exploration the regularization information is incorporated. Therefore, the prior model (a given solution), which is close to the target (compare Figures 5.6a and 5.8b) is considered. This prior model (which is very close to the target) is chosen in order to evaluate the effect of including accurate prior information into the HMC exploration. The prior model is just a small modification of the target model, Figure 5.7, compares the two different slip-rate functions from the target and the prior models. In realistic cases, however, the prior model might be far from the true solution, which can prevent us from looking close to the vicinity of the target solution. The hyper-parameter ϵ from the regularization term is set in such a way that the prior information is equally balanced with the data misfit term. This third case, however, considers the identity matrix as the mass matrix of the auxiliary momentum variables. Therefore, we are interested in looking at the impact of the prior information even when the mass matrix is poorly tuned. Finally, the fourth case considers exactly the same configuration as the third, but instead of considering an identity mass matrix it uses the time-spatial coherence matrix mentioned above (Figure 5.8c).

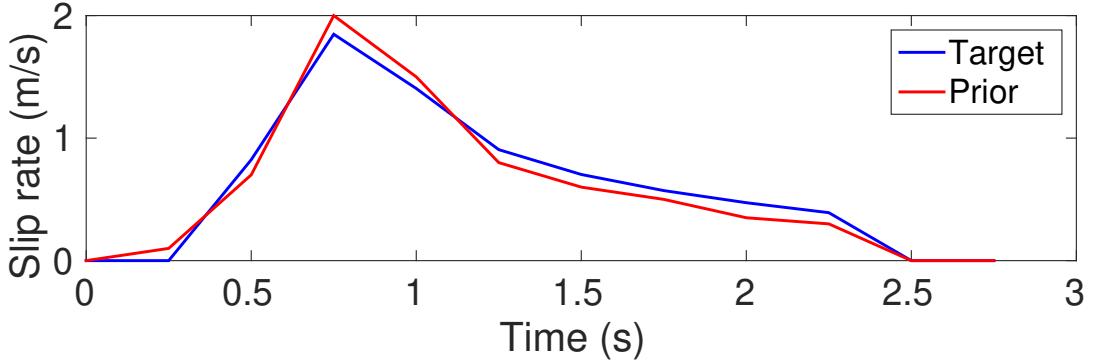


Figure 5.7: Comparison of the slip-rate time functions of the target (blue line) and the prior (red line) models respectively.

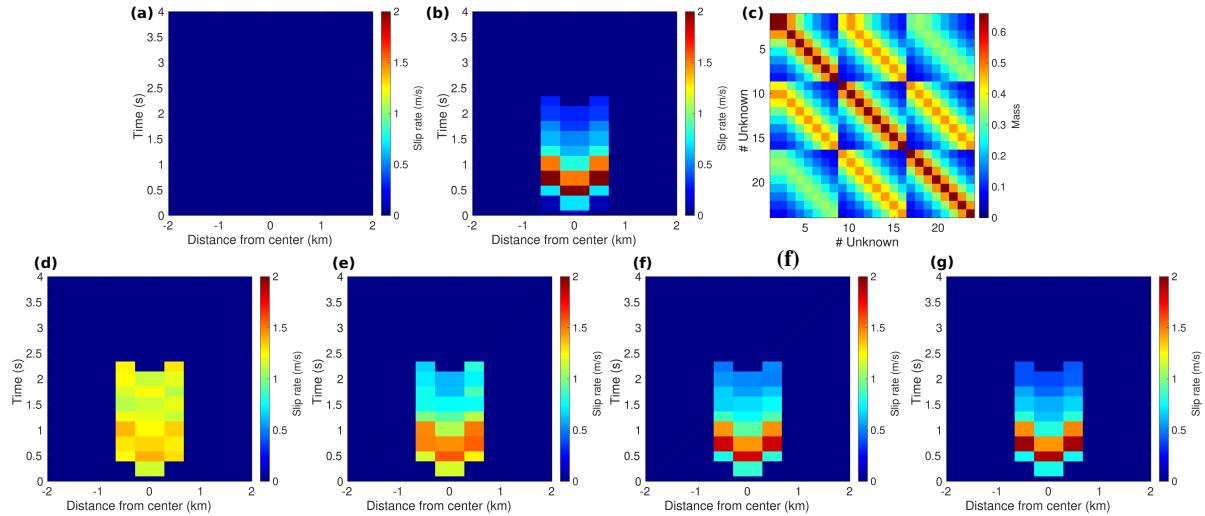


Figure 5.8: Summary of the resulting mean rupture models obtained from the four different HMC explorations. (a) initial guess used for the four explorations. (b) prior model used only for the third and fourth exercises ((f) and (g)). (c) mass matrix used for the second and fourth ((e) and (g)) explorations to draw the auxiliary momentum variables. On the bottom, a comparison between the resulting mean rupture models obtained from the four explorations: (d) without taking into account prior information and with the identity as the mass matrix, (e) only using the mass matrix showed in (c) and no prior information, (f) using the prior information from (b) and with an identity mass matrix, and (g) using the mass matrix shown in (c) plus the prior model illustrated in (b). Notice that by tuning the mass matrix to account for time-space coherence the resulting mean model (e) improves significantly. Results from (f) where the prior information is considered are also quite improved even if the mass matrix considered is diagonal.

Notice that these four different explorations are also helpful to see the influence of the mass matrix and the prior information into the HMC exploration. In addition, these four HMC explorations incorporate upper and lower limits (constraints as described in equation (5.17), which force the exploration to

be in a range from -1 to 2.5 m/s. Finally, as in the previous exercise, I use the synthetic bilateral rupture to compute the exact seismograms that are then used as the noise-free observations at the 3 components of the 420 receivers around the line source.

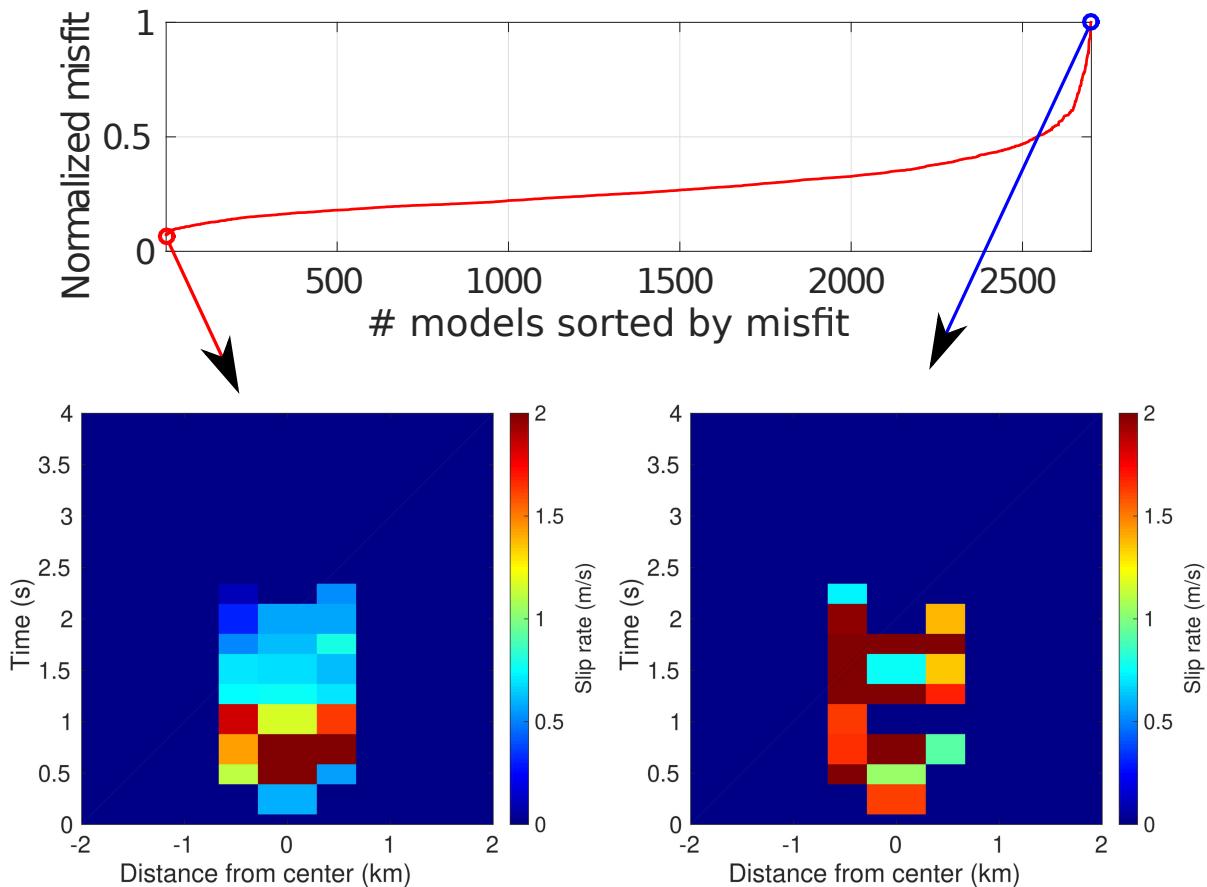


Figure 5.9: a) Misfit of the 2,700 models explored by the HMC strategy sorted from the minimum (most likely scenario) to the maximum (most unlikely). In a) the red and blue circles represent the misfit of the minimum and maximum misfit scenarios respectively. b) and c) illustrate the rupture models with the minimum and maximum misfits (according to the red and blue arrows). These plots correspond to the third HMC exploration, which considers prior information and a block diagonal mass matrix to account for time-space coherence.

On the contrary, the results from the second and third exercises are much better than the first case. It is evident that, for the second case the enforced spatial coherence (due to the non-diagonal mass matrix) allows to account for a coherent perturbation of the parameters. Therefore, the HMC exploration is able to focus only on the models that respect this time-space coherence and the resulting *ppd* exhibits a Gaussian-shape (see Figure 5.10b). Notice also that the mean model from that second exploration is less smooth than the one from the first case (compare Figures 5.8e with 5.8d). Finally, the best case is the third one, which includes the prior information besides the tuning of the mass matrix. For that case, the exploration is driven closer to the real solution. The mean model from this last case is the closest to the target solution (see the mean model from Figure 5.8f). The *ppd* is the less similar to a uniform distribution and its Gaussian-shape is better defined (compare Figures 5.10a,b,c).

The results from the four runs above described are summarized from Figures 5.9 to 5.13. The first run, which does not consider any prior information or time-space coherence (enforced by the mass matrix) provides a mean model very smooth and far from the real target model to be reconstructed (Figure 5.8d). Looking at the 24-dimensional *ppd* that is found by this run, it is clear that, when no prior information is incorporated into the potential energy (misfit function) and a simple diagonal mass matrix is assumed, it is not enough to determine correctly the expected solution and the estimated *ppd* behaves as a uniform distribution (see the uniform behavior of Figure 5.10a). The shape of the posterior distribution shows a small evidence of where the most probable might be, but in general, exhibits a uniform-shape *ppd*. This resulting *ppd* means that any slip-rate value ranging from -1 to 2.5 m/s randomly organized in this 24-dimensional model are equally accepted to explain, to some extent, the available data. It is important to highlight that this is a striking result if we recall that this exercise is carried out in a unrealistic perfect illumination configuration. Therefore, to get simultaneously an accurate approximation to the solution to the problem together with the associated *ppd* without using any information or tuned mass matrix (first approach) appears to be very difficult even in perfect illumination conditions.

Regarding the evolution of the Hamiltonian system, I can say the following. The individual trajectories (perturbations of the slip-rate parameters) are completely chaotic if the mass matrix is assumed as diagonal (no correlation between parameters) and/or if no prior information is used. For the case that assumes no prior information and the identity as the mass matrix, each auxiliary variable is independent from the others. Therefore, the parameters are randomly perturbed in all possible directions without any link between them. The fact is that, this random independent perturbation is not a correct assumption as we might expect that parameters close in time and space have to act similarly. The consequence that can be seen from this random independent perturbation is that the chaotic behavior of the particle proposes scenarios where all the parameters interact in such a way that their responses sum or cancel the effect of the others with the only goal to fit the data. These interactions provide scenarios that are very far from the target solution but that produce seismograms that are not completely far from the observations (belonging to the null space). As a consequence, most of these scenarios are accepted, fact that is reflected by the uniform-like shape of the *ppd* obtained from the first exercise (Figure 5.10a).

It is important to notice that, when the prior information is considered (Figures 5.8f,g), the resulting mean model remains close to the assumed prior model, regardless of the shape of the mass matrix (either diagonal or non-diagonal). Therefore, the prior model is able to enforce the time and space coherence that is not taken into account by the diagonal matrix. However, for this exercises, which might not be realistic, the prior model is very close from the target solution.

The Figure 5.11, 5.12 and 5.13 form the corner plot of the 24-dimensional *ppd* found by the fourth HMC exploration. In those figures, it is clear that thanks to the prior information and enforced coherence between parameters, the shape of the *ppd* tends to a Gaussian-shape. As a final illustration, in Figure 5.9 I provide a plot of the misfit values for all the 2700 models explored by the third HMC exploration. The values of the misfit are sorted in an incremental fashion. Below that plot, I provide the models with the minimum and maximum value of misfit from the the whole population of the 2700 models.

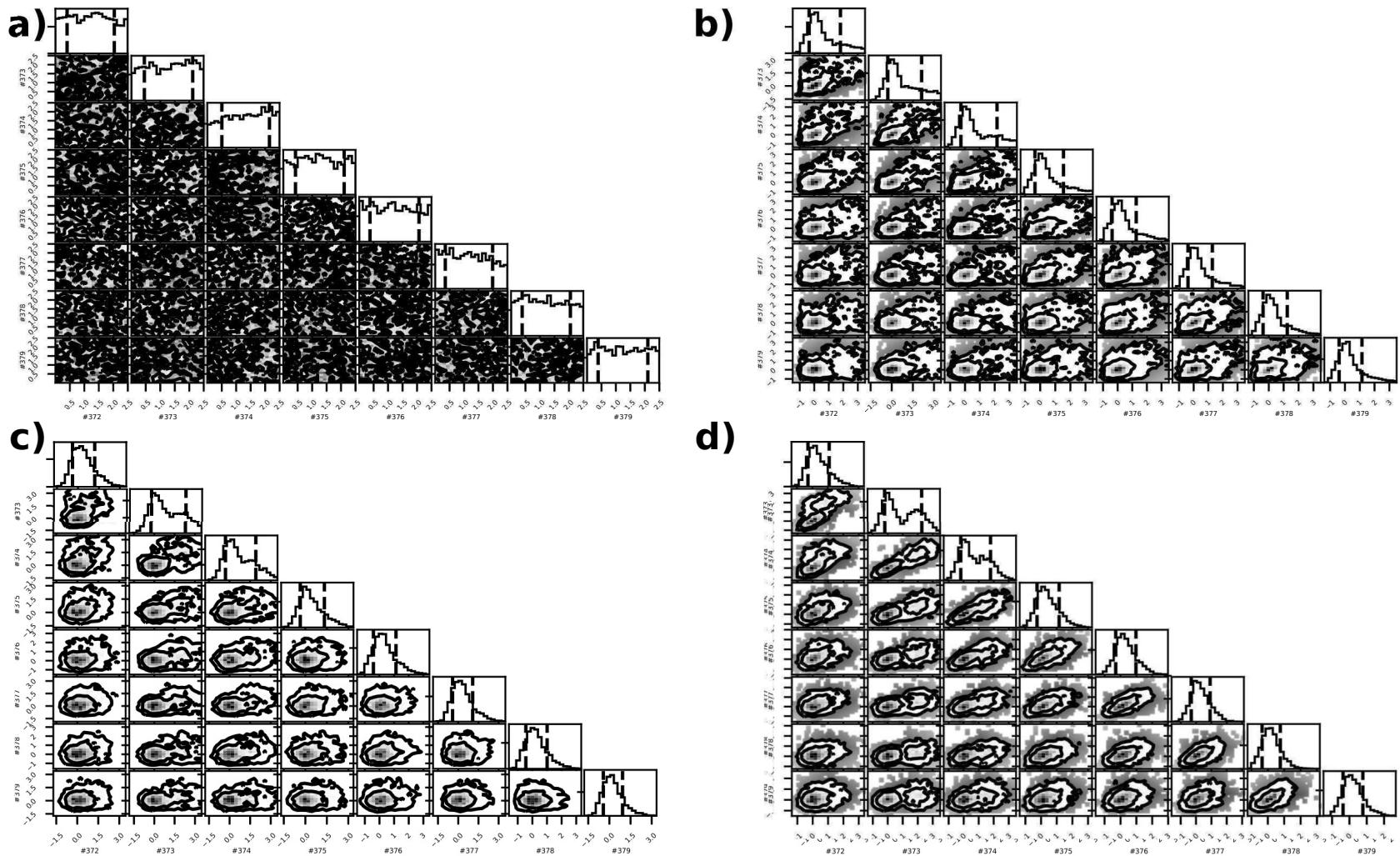


Figure 5.10: Comparison of the first 7 dimensions of the obtained *ppds* from the four HMC explorations after 2,700 iterations. a) No regularization and diagonal mass matrix, b) using a non-diagonal mass matrix and no prior information and c) accounting for prior information and a diagonal mass matrix and d) using a non-diagonal mass matrix and accounting for prior information. Only the first 7 dimensions out of the 24 explored are presented in this figure. The 24 dimension of d) are illustrated in Figures 5.11 5.12 and 5.13. In (a), (b) (c) and (d) quantiles 0.16 and 0.84 are represented by the vertical dashed lines in the 1D plots, while the levels of confidence 0.68 and 0.95 are shown as the contour lines in the 2D plots. The seven dimensions here illustrated represent the slip-rate time history of the node at the left of the center of the line source.

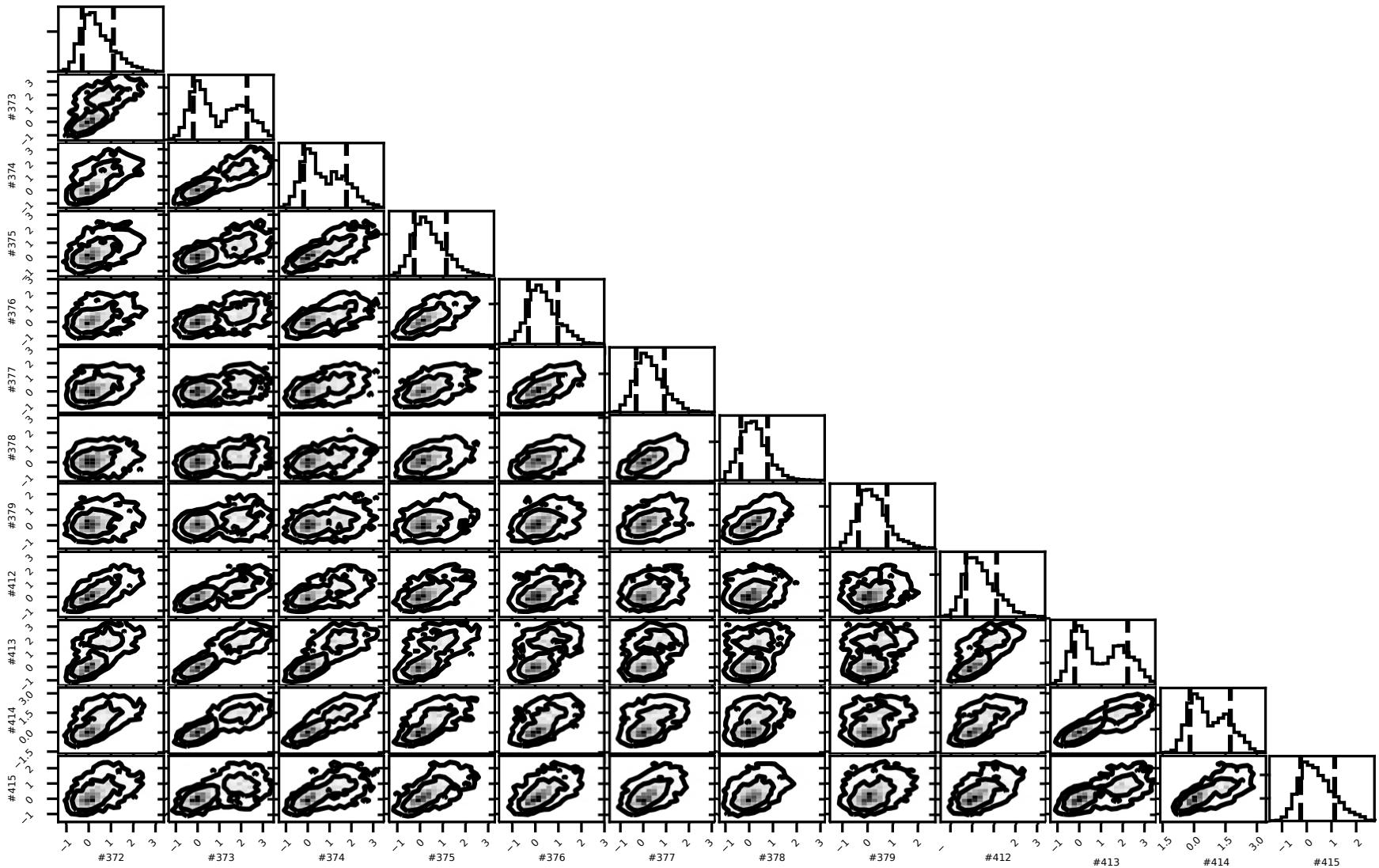


Figure 5.11: Corner plot illustrating the 24-dimensional *ppd* obtained for the bilateral rupture exercise from the third HMC exploration, which accounts for time-space coherence (i.e. non-diagonal mass matrix) and prior information. In this plot, quantiles 0.16 and 0.84 are represented by the vertical dashed lines in the 1D plots, while the levels of confidence 0.68 and 0.95 are shown as the contour lines in the 2D plots. The 24 dimensions here illustrated represent the slip-rate time history of the 3 spatial nodes located at the center of the line source. This *ppd* is obtained after 2,700 HMC iterations.

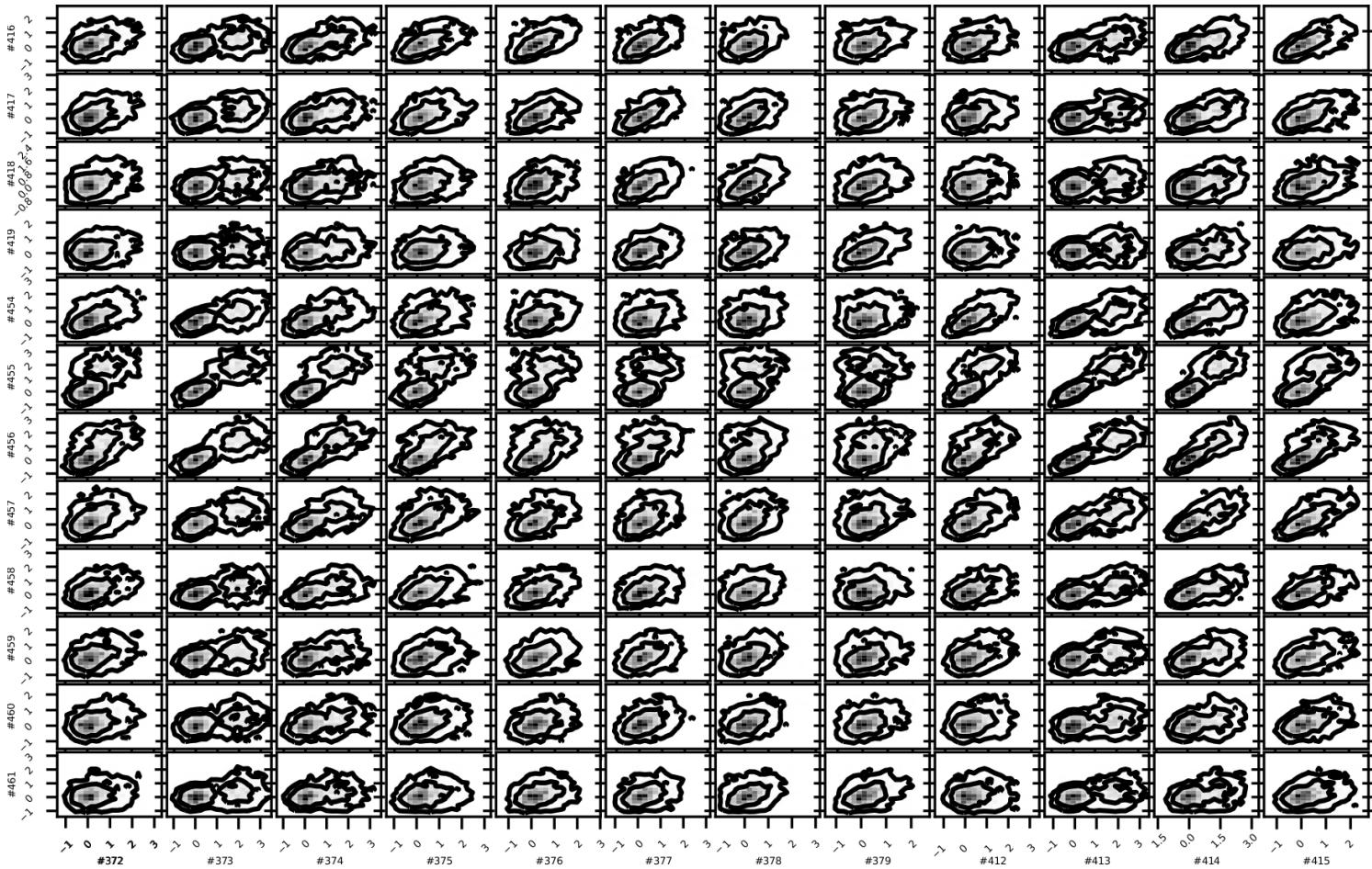


Figure 5.12: Continuation of Figure 5.11

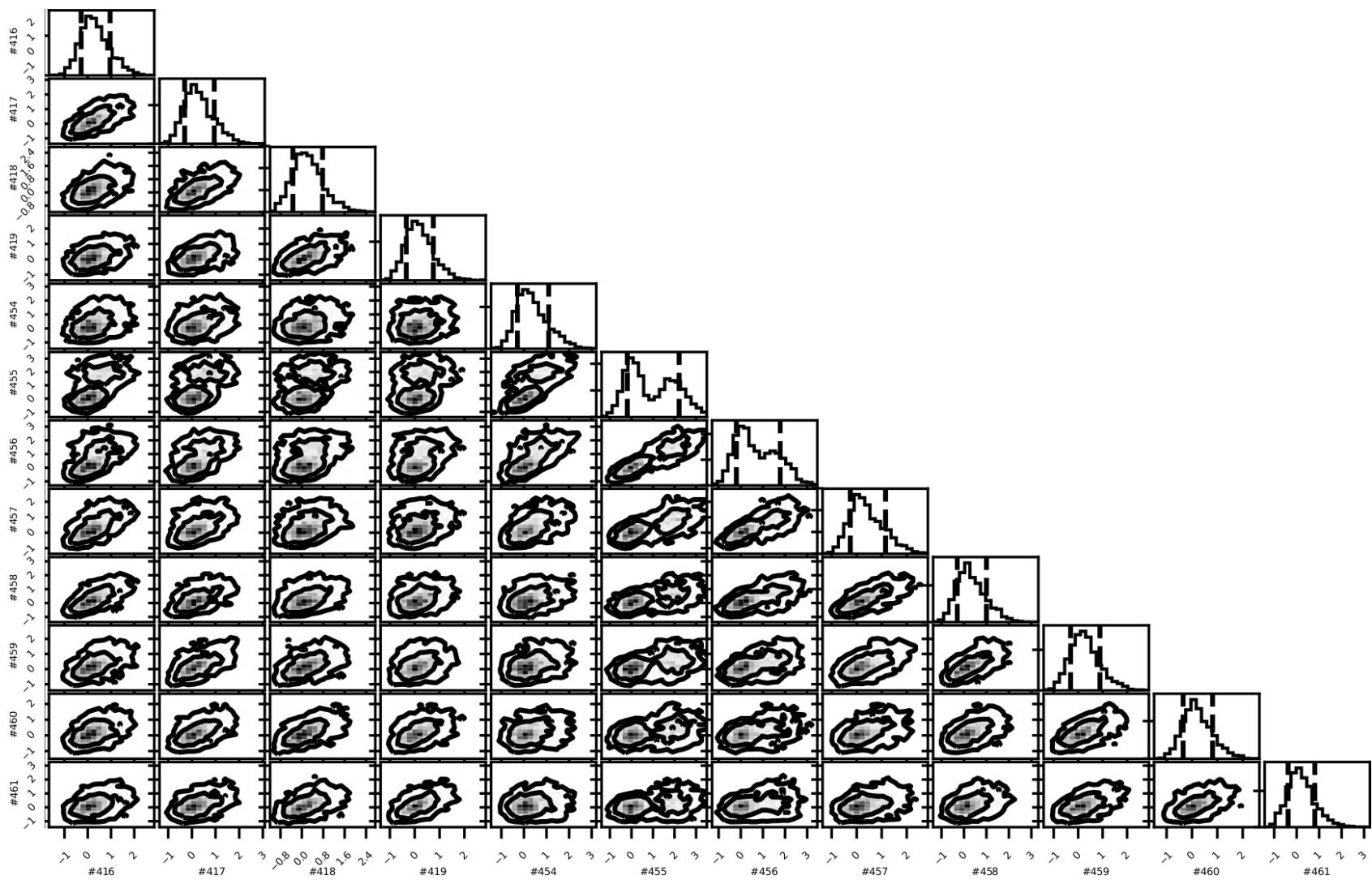


Figure 5.13: Continuation of Figure 5.11 and 5.12.

5.5 Conclusion

In this chapter, I present the basic theory of Bayesian inference as well as the fundamental concepts behind a standard Markov Chain Monte Carlo (MCMC) strategy and the Hamiltonian Monte Carlo (HMC) approach. I provide simple examples where some of the advantages of the HMC strategy are illustrated. Then, I apply the HMC strategy to the kinematic source inversion problem taking into account all the previous development presented in Chapters 3 and 4.

The HMC technique can be applied to the kinematic source reconstruction problem in order to achieve two different goals: 1) to obtain a solution to the problem together with the related uncertainties or 2) to assess the uncertainties given a prescribed solution. In this chapter, the preliminary investigation presented is focused on the second goal. The investigation related to the first goal remains as a very interesting and challenging perspective of this work. Taking into account that the HMC technique is able to describe multimodal functions (like in the second exercise of Section 5.3), its application to the source inversion problem under a non-linear formulation seems as a very interesting approach to be explored. Under the linear formulation of the problem, the HMC exploration around a given solution appears to be more logical and consistent: once a solution is found the HMC strategy serves as an efficient variational technique to describe the closest vicinity. The importance of the linear formulation under this approach relies on the fact that a small perturbation of the slip-rate time space history maps directly into small changes in the estimated seismograms. This feature is not ensured under the non-linear formulation of the problem. Therefore, a more general exploration (first approach) would be more pertinent if a non-linear formulation of the problem is to be used.

According to this preliminary investigation, the HMC approach seems to be pertinent and compatible to perform the exploration of the space of solutions around a given solution under the linear time-domain formulation of the source reconstruction problem. However, the tuning of some fundamental parameters of the technique appears to be crucial for a correct exploration. In the hierarchy of importance, I should point out that the incorporation of prior information (injecting the physics to the problem) as well as the tuning of the mass matrix (controlling the time-space coherence) seem to be the most relevant. It is clear that, when dealing with problems such as the source inversion where a significant trade-off between parameters is present, these ingredients play a very important role. While the prior information allows us not to explore very far from the desired physical scenarios, the mass matrix (used to draw the auxiliary momentum variables) allows to account for the existing interaction between parameters. Under the linear time-domain formulation of the kinematic source inversion problem we know that the number of unknowns is quite large, however, these parameters have to be somehow linked physically. It is the task of this matrix to control the level of perturbation of all the parameters together taking into account this physical interaction. Certainly, a more detailed investigation of a correct way to choose the mass matrix has to be done for this type of problems, the identity matrix is completely far from the optimal to tackle high correlated parameters. In a second level of importance, I would mention the randomness pattern of the integration step δ and the number of steps L , which seem to be less important than the mass matrix, at least for the exercises explored.

In addition, it is important to highlight the crucial role played by the prior information. The chaotic behavior that the HMC exploration exhibited in the non-regularized case which used a diagonal mass matrix of the simple bilateral exercise is certainly not only due to the incorrect choice of the mass matrix. In fact, this behavior is also related to the lack of physics of the problem (even when a good illumination is available). From the results of the two exercises where the prior information is incorporated (based on the model regularization term), it can be seen that the chaotic behavior is mitigated and the exploration starts to focus its attention in the regions surrounding the most probable scenarios. This can be interpreted in the following way: the misfit function of this problem is convex and quadratic, however, without prior information its shape in the high-dimensional space of solutions is very flat at its minimum (many scenarios are possible and equivalent). The prior information, that injects to the problem the necessary physics to be honored, allows us to focus our attention only at a determined region of interest of that immense space of solution. Then, including the prior information into the problem allows us to pass from a misfit function with a very flat U-shape to a more defined V-shape. Even though this effect of the prior model can be risky, this is certainly less risky than assuming a different (reduced) parameter design of the slip-rate function, which can consider few parameters such as starting time, rise time, and final slip. In that sense, if the assumed prior model is quite far from the true solution, a closer look to the vicinity where the true solution

lies will be far from the region where we are performing our exploration.

Finally, I would like to add that the application to the HMC to much more complex cases with realistic source-receivers configurations and with a significant increment of the number of unknowns (dimensions) remains as a very interesting perspective. In particular, the progressive time-domain strategy, that is presented in Chapters 3 and 4, seems to be a very interesting technique to be coupled with the HMC strategy. The coupling of both strategies would allow the HMC exploration to progressively determine the regions of interest in the large dimension space of solutions, in a similar way to the importance sampling technique or the particle filter strategy.