Geophysical Journal International

Geophys. J. Int. (2016) **205,** 1710–1728
Advance Access publication 2016 March 28
GJI Seismology

doi: 10.1093/gji/ggw108

Solving probabilistic inverse problems rapidly with prior samples

Paul Käufl, Andrew P. Valentine, Ralph W. de Wit and Jeannot Trampert

Department of Earth Sciences, PO Box 80.115*,* 3508 *TC, Utrecht, The Netherlands. E-mail:* *jeannot@geo.uu.nl*

Accepted 2016 March 21. Received 2016 March 20; in original form 2015 April 23

SUMMARY

Owing to the increasing availability of computational resources, in recent years the probabilistic solution of non-linear, geophysical inverse problems by means of sampling methods has become increasingly feasible. Nevertheless, we still face situations in which a Monte Carlo approach is not practical. This is particularly true in cases where the evaluation of the forward problem is computationally intensive or where inversions have to be carried out repeatedly or in a timely manner, as in natural hazards monitoring tasks such as earthquake early warning. Here, we present an alternative to Monte Carlo sampling, in which inferences are entirely based on a set of prior samples—that is, samples that have been obtained independent of a particular observed datum. This has the advantage that the computationally expensive sampling stage becomes separated from the inversion stage, and the set of prior samples—once obtained—can be reused for repeated evaluations of the inverse mapping without additional computational effort. This property is useful if the problem is such that repeated inversions of independent data have to be carried out. We formulate the inverse problem in a Bayesian framework and present a practical way to make posterior inferences based on a set of prior samples. We compare the prior sampling based approach to a Markov Chain Monte Carlo approach that samples from the posterior probability distribution. We show results for both a toy example, and a realistic seismological source parameter estimation problem. We find that the posterior uncertainty estimates obtained based on prior sampling can be considered conservative estimates of the uncertainties obtained by directly sampling from the posterior distribution.

Key words: Neural networks, fuzzy logic; Inverse theory; Probability distributions; Earthquake source observations; Early warning.

1 INTRODUCTION

In geophysics, we are concerned with a wide variety of inverse problems. Examples include imaging the Earth's interior and determination of earthquake source parameters using observed seismic waves; the determination of the Earth's density structure from gravity measurements; or the determination of magnetic susceptibility from measurements of the Earth's magnetic field. In all these problems, we are able to predict a set of observables exactly or approximately given a certain set of model parameters and parametrization, which we call the 'forward problem'. Given an observation, the associated 'inverse problem' is to find parameter values compatible with all information available on the problem. We thereby distinguish between prior information, which is available before a particular observation has been made, and posterior information, which is the result of combining our prior knowledge with the information provided by the observation (Tarantola 2005).

Due to the need to deal with observational uncertainties, uncertainties attached to any approximations in the forward problem and the inherent non-uniqueness of many inverse problems, a probabilistic treatment is advantageous. In a probabilistic approach, all information is represented by probability density functions (pdfs), and the solution to the inverse problem is thus given by the posterior pdf. However, owing to non-linearity and, often, the complication that the forward problem itself lacks a closed-form solution, we cannot typically find closed-form expressions for the posterior quantities of interest such as the maximum likelihood point and covariance matrix. In these cases, we have to resort to testing models at random to capture the information about the relation between observable data and model parameters (often referred to as 'sampling' this relationship).

Broadly speaking, sampling can be approached from two directions. Techniques such as Markov Chain Monte Carlo (MCMC) methods (Sambridge & Mosegaard 2002) rely on 'posterior' sampling, whereby samples are targeted towards explaining some specific set of observations. These methods have been studied extensively, and have found wide application throughout the physical sciences. However, they are typically computationally expensive: tens or hundreds of thousands of models may need to be tested to obtain meaningful results. In practice, this also makes them time-intensive: given that sampling cannot commence until observations are available, posterior sampling is poorly adapted to applications such as earthquake early warning (EEW), where results must be obtained in the shortest possible time.

Posterior sampling is also ill-suited to situations where a particular type of inverse problem must be solved repeatedly—typically, where analysis must be conducted at different points in space and/or time. This requirement is quite common in geophysics: examples include routine determination of earthquake locations and source mechanisms, or inversion for local crustal properties at many locations. In such circumstances, the physical processes linking model and observables are common to all the inversions—but since posterior sampling is tuned towards explaining one data set, it is not generally possible to 'recycle' any information gained in one case when analysing another. Thus, the computational costs for using posterior sampling repeatedly can be immense.

These challenges motivate the development of the second approach to sampling—the subject of this paper—which we refer to as 'prior sampling'. In this framework, all samples are evaluated prior to consideration of any particular set of observations, informed by any prior knowledge that might be available; in time-sensitive contexts, such as the example of EEW, sampling may be conducted before an event has even occurred. Then, using this set of samples, the probabilistic inverse problem may be solved extremely rapidly once data are available, and the same set of samples can be efficiently re-used in conjunction with numerous observations. This repeatability may also be a useful property in exploratory studies, making it straightforward to investigate which aspects of a data set convey useful information: for example, a single set of samples may be post-processed in a variety of different ways, and then applied to inversion of synthetic data.

By design, all samples tested during posterior sampling contribute directly to the solution of the inverse problem being considered. In contrast, prior sampling may entail evaluating many samples that turn out to lie far from the observations made in any particular case. Such samples are therefore 'wasted' from the perspective of any single inversion—and where only one data set needs to be analysed, posterior approaches should generally be adopted unless it is desirable to exploit particular properties of the prior sampling framework. However, if a sufficient number of distinct inversions are to be performed, all prior samples can be expected to be 'useful' in some cases, and the approach represents an efficient use of computational resources overall.

Viewing the same issue from another direction: if we have many samples that lie close to a given observation, we are in a much stronger position to make robust inferences than if we must extrapolate from distant samples. The quality of results from prior sampling is therefore intrinsically linked to the density of samples used. As is well known, the number of samples required to maintain a given density grows exponentially as the dimension of the model space (i.e. the number of free model parameters) increases (e.g. Curtis & Lomax 2001; MacKay 2003). Given that there are inevitable practical restrictions on the number of samples that can be generated for a particular problem, this may limit the size of problem for which a prior sampling approach is viable. The implementation described in this paper is designed to return the prior probability distribution in cases where no useful inference can be made from the samples available. If insufficient samples are available for the complexity of the problem, the system may not provide informative results. Equally, it should not yield actively misleading outputs. However, we have yet to fully explore the practicalities of applying prior sampling in high dimensions. As with any technique, it is incumbent upon potential users to satisfy themselves that performance and accuracy are appropriate to the problem at hand.

Just as 'posterior sampling' encompasses a wide range of different algorithms and techniques, many different detailed implementations of 'prior sampling' may be possible. As we will show in this paper, inference using prior sampling is based upon probability density estimation, and this can be tackled in a variety of ways. Here, we adopt a non-linear, neural-network based tool, called a Mixture Density Network (MDN; Bishop 1995). This exploits an assumption that the underlying pdf is smooth and continuous to enable interpolation between samples. Once prior samples have been obtained and the MDN has been constructed, the mapping from data to posterior pdf can be evaluated within milliseconds on a standard desktop computer. In particular, note that access to the large data set constituting the 'prior samples' is only required during the network construction phase: the MDNs assimilate this information, and can then be run operationally on machines of modest specification.

This paper is organized as follows: after discussing a few theoretical considerations, we compare MDN estimates obtained from prior samples to reference solutions obtained by a Metropolis–Hastings (MH) MCMC sampling algorithm (Hastings 1970). First, we consider a multimodal non-linear toy problem, for which an analytical reference solution can be calculated. We then investigate the case of a realistic, seismological point-source inversion problem, taken from Käufl *et al.* (2014, 2015).

2 TWO ROUTES TO POSTERIOR INFERENCE

In a (geophysical) inference problem, we are typically dealing with three sets of variables. First, we make measurements, collectively represented by the observed data vector $\mathbf{d}_0 \in \mathbb{D}$. Second, we describe the physical reality by means of a model, which is parametrized by the model parameters $\mathbf{m} \in \mathbb{M}$. Third, we assume that we can make predictions $\mathbf{d} \in \mathbb{D}$, by evaluating the forward operator $g(\mathbf{m})$ for any given model $\mathbf{m}$. We assume that observations $\mathbf{d}_0 \in \mathbb{D}$ are subject to noise and can be related to predictions $\mathbf{d}$ by an observational noise process. Here, $\mathbb{D}$ and $\mathbb{M}$ are normed vector spaces, which are referred to as the data and model space, respectively. The goal of the inference process is then to identify a set of model parameter vectors that can explain a given observation $\mathbf{d}_0$, perhaps subject to additional assumptions, such as observational and modeling uncertainty estimates. In the following, we treat the variables $\mathbf{m}$, $\mathbf{d}$ and $\mathbf{d}_0$ as random variables and express their relations and any prior assumptions by pdfs.

First, we may have independent prior information on the model parameters, represented by the pdf $\rho(\mathbf{m}|A)$, where the symbol A is used to represent our assumptions about the model prior, such as the shape and parameters of the pdf and the particular parametrization of the physical model. Furthermore, we use $\rho(\mathbf{d}, \mathbf{d}_0|B)$ to denote the joint prior distribution over predictions $\mathbf{d}$ and observations $\mathbf{d}_0$, subject to some set of additional assumptions B, typically describing the observational noise process. Finally, we describe the information contained in the forward operator—that is, the correlation between model parameters and predictions—by the joint pdf $\Theta(\mathbf{m}, \mathbf{d}|C)$, where C denotes the assumptions made on the theoretical relation, such as the choice of a particular forward operator $g(\mathbf{m})$ and potential modeling uncertainties.

Therefore, the complete knowledge about the system is given by the probabilistic conjunction of all individual pieces of information (Tarantola 2005)

$$\sigma(\mathbf{m}, \mathbf{d}, \mathbf{d}_0|A, B, C) = k\frac{\rho(\mathbf{m}|A)\rho(\mathbf{d}, \mathbf{d}_0|B)\Theta(\mathbf{m}, \mathbf{d}|C)}{\mu(\mathbf{m})\mu(\mathbf{d})\mu(\mathbf{d}_0)}, \tag{1}$$

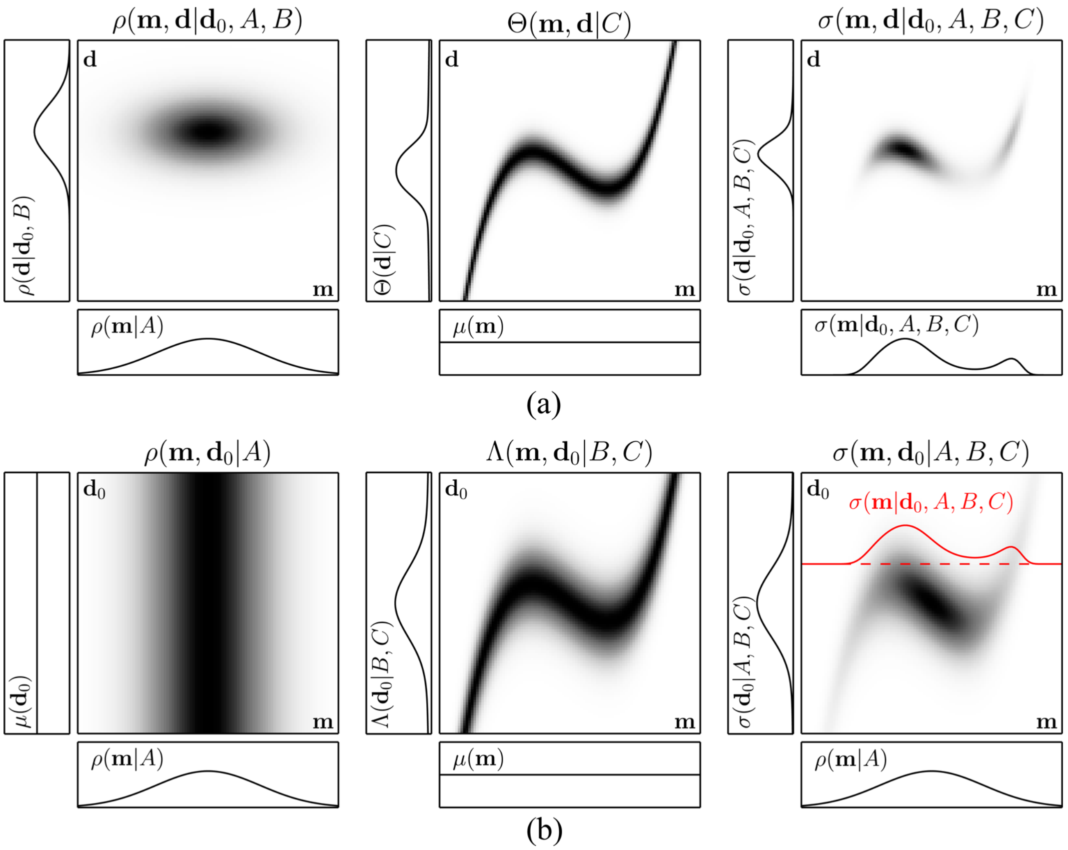

$\rho(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B)$
$\Theta(\mathbf{m}, \mathbf{d}|C)$
$\sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C)$
$\rho(\mathbf{d}|\mathbf{d}_0, B)$
$\Theta(\mathbf{d}|C)$
$\sigma(\mathbf{d}|\mathbf{d}_0, A, B, C)$
d
m
$\rho(\mathbf{m}|A)$
$\mu(\mathbf{m})$
$\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$
(a)
$\rho(\mathbf{m}, \mathbf{d}_0|A)$
$\Lambda(\mathbf{m}, \mathbf{d}_0|B, C)$
$\sigma(\mathbf{m}, \mathbf{d}_0|A, B, C)$
$\mu(\mathbf{d}_0)$
$\Lambda(\mathbf{d}_0|B, C)$
$\sigma(\mathbf{d}_0|A, B, C)$
$\mathbf{d}_0$
$\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$
m
$\rho(\mathbf{m}|A)$
$\mu(\mathbf{m})$
$\rho(\mathbf{m}|A)$
(b)

Figure 1. Two ways to express the posterior pdf $\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$ of model parameters $\mathbf{m}$ given observation $\mathbf{d}_0$ and subject to the sets of assumptions A, B and C, about the prior distribution of model parameters and parametrization, observational errors and the forward model, respectively. Figure after Tarantola (2005). (a) The fact that we have observed $\mathbf{d}_0$ enters the inference process in the form of prior information described by the pdf $\rho(\mathbf{d}|\mathbf{d}_0, B)$ (left-hand panel). Theoretical predictions are represented by $\Theta(\mathbf{m}, \mathbf{d}|C)$, taking into account any assumptions about the relation between data and model parameters (middle panel). The posterior pdf is given by the marginal pdf $\int_{\mathbb{D}} \sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C)\mathrm{d}\mathbf{d}$ (right-hand panel). (b) We assume that the observation $\mathbf{d}_0$ is available only at a later time and is thus treated as an unknown variable with prior distribution $\mu(\mathbf{d}_0)$ (left-hand panel). The pdf $\Lambda(\mathbf{m}, \mathbf{d}_0|B, C)$ carries the combined assumptions on theoretical modeling and observational uncertainties (middle panel). The posterior pdf is given by the conditional pdf $\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C)$ (red solid line, right-hand panel).

where k is a normalization constant and $\mu(\mathbf{m})$, $\mu(\mathbf{d})$ and $\mu(\mathbf{d}_0)$ are the homogeneous distributions of the model and data space, respectively, which are constant in the case of linear vector spaces. Given a particular observation $\mathbf{d}_0 \in \mathbb{D}$, the probabilistic solution to the inverse problem is given by the *posterior* distribution over the model space $\sigma(\mathbf{m}|\mathbf{d}_0)$. In what follows, we will demonstrate that the posterior pdf $\sigma(\mathbf{m}|\mathbf{d}_0)$ can be expressed in two different ways, which are distinguished by the moment at which the information that we have made a particular observation $\mathbf{d}_0$ enters the inference process, that is,—in probabilistic terms—when the conditioning operation on $\mathbf{d}_0$ is applied. In one case, the fact that we have made a particular observation $\mathbf{d}_0$ enters the inference as prior information, whereas in the other we treat the observation itself as a random variable. Fig. 1 provides a visual comparison of the two approaches, where (for illustrative purposes) we have assumed prior distributions, observational uncertainties and modeling errors to be described by Gaussian distributions.

First, we focus on the case depicted in Fig. 1(a). Here we assume that the observation $\mathbf{d}_0$ has been obtained before we start the inference process and we can therefore consider $\mathbf{d}_0$ as fixed, in which case we can write for eq. (1)

$$\sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C) = \tilde{k}\,\frac{\rho(\mathbf{m}|A)\rho(\mathbf{d}|\mathbf{d}_0, B)\Theta(\mathbf{m}, \mathbf{d}|C)}{\mu(\mathbf{m})\mu(\mathbf{d})}, \qquad (2)$$

where $\tilde{k}$ is a normalization constant. Note that we have defined $\rho(\mathbf{d}, \mathbf{d}_0) = \rho(\mathbf{d}|\mathbf{d}_0)\mu(\mathbf{d}_0)$, that is, we choose the homogeneous distribution as a marginal distribution for $\mathbf{d}_0$, since the prior distribution over the data variables should not restrict the possible observations $\mathbf{d}_0$ in any way.

The posterior pdf can now be obtained as the marginal distribution over the predicted data variables, that is,

$$\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C) = \int_{\mathbb{D}} \sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0, A, B, C)\mathrm{d}\mathbf{d}. \qquad (3)$$

In the special case that the forward problem can be assumed to be exact we have

$$\Theta(\mathbf{m}, \mathbf{d}|C) = \delta(\mathbf{d} - g(\mathbf{m}))\mu(\mathbf{m}), \qquad (4)$$

and eq. (3) can be written in the form

$$\sigma(\mathbf{m}|\mathbf{d}_0, A, B, C) = \frac{\rho(\mathbf{m}|A)L(\mathbf{m}|\mathbf{d}_0, B, C)}{\sigma(\mathbf{d}_0|A, B, C)}, \qquad (5)$$

where

$$L(\mathbf{m}|\mathbf{d}_0, B, C) = \rho(g(\mathbf{m})|\mathbf{d}_0, B) \qquad (6)$$

is referred to as the *likelihood function* and where the normalization constant $\sigma(\mathbf{d}_0|A, B, C) = \int \rho(\mathbf{m}|A)L(\mathbf{m}|\mathbf{d}_0, B, C)\mathrm{d}\mathbf{m}$ is the Bayesian evidence.

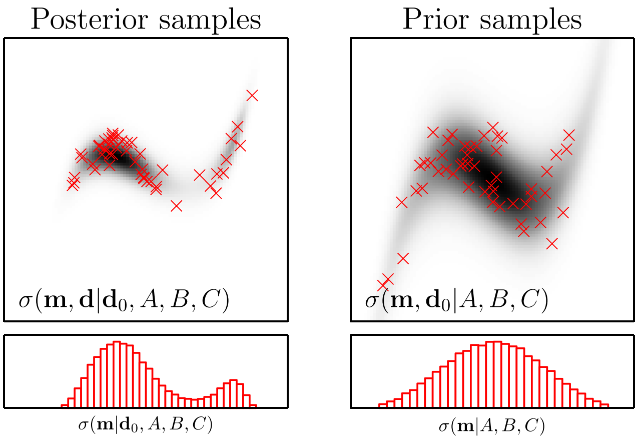


Figure 2. Two approaches to sampling the relationship shown in Fig. 1. A set of posterior samples $\mathcal{D}_{\text{post}}$ (left), whose density follows the distribution $\sigma(\mathbf{m},\mathbf{d}|\mathbf{d}_0,A,B,C)$ and a set of prior samples $\mathcal{D}_{\text{prior}}$ (right) following $\sigma(\mathbf{m},\mathbf{d}_0|A,B,C)$.

We now turn to the case depicted in Fig. 1(b). Here, we assume that $\mathbf{d}_0$ will only become available at a later time during the inference process. Therefore, we treat $\mathbf{d}_0$ as a random variable, whose value is unknown, and do not perform the conditioning operation on $\mathbf{d}_0$ yet. Instead, we marginalize over the predicted data variables $\mathbf{d}$ first and write using eq. (1)

$$\begin{aligned}\sigma(\mathbf{m},\mathbf{d}_0|A,B,C) &= \int_{\mathbb{D}} \sigma(\mathbf{m},\mathbf{d},\mathbf{d}_0|A,B,C)\mathrm{d}\mathbf{d}\\ &= k\frac{\rho(\mathbf{m}|A)}{\mu(\mathbf{m})\mu(\mathbf{d}_0)}\int_{\mathbb{D}}\frac{\rho(\mathbf{d},\mathbf{d}_0|B)\Theta(\mathbf{m},\mathbf{d}|C)}{\mu(\mathbf{d})}\mathrm{d}\mathbf{d}\\ &= k\frac{\rho(\mathbf{m}|A)}{\mu(\mathbf{m})\mu(\mathbf{d}_0)}\Lambda(\mathbf{m},\mathbf{d}_0|B,C), \qquad (7)\end{aligned}$$

where we have defined

$$\Lambda(\mathbf{m},\mathbf{d}_0|B,C) = \int_{\mathbb{D}}\frac{\rho(\mathbf{d},\mathbf{d}_0|B)\Theta(\mathbf{m},\mathbf{d}|C)}{\mu(\mathbf{d})}\mathrm{d}\mathbf{d}. \qquad (8)$$

Once a particular observation $\mathbf{d}_0$ becomes available, we condition eq. (7) on $\mathbf{d}_0$ in order to find the posterior

$$\sigma(\mathbf{m}|\mathbf{d}_0,A,B,C) = \frac{\sigma(\mathbf{m},\mathbf{d}_0|A,B,C)}{\sigma(\mathbf{d}_0|A,B,C)} \qquad (9)$$

as before.

In particular, under the assumption of Gaussian observational and modeling uncertainties with covariances $\mathbf{C}_d$ and $\mathbf{C}_g$, respectively, and with the spaces $\mathbb{M}$ and $\mathbb{D}$ being linear, we have

$$\Lambda(\mathbf{m},\mathbf{d}_0|\mathbf{C}_D) \propto \exp\left\{-\frac{1}{2}\left[\mathbf{d}_0 - g(\mathbf{m})\right]^T\mathbf{C}_D^{-1}\left[\mathbf{d}_0 - g(\mathbf{m})\right]\right\}, \qquad (10)$$

with the combined observational and theoretical noise covariance matrix $\mathbf{C}_D = \mathbf{C}_d + \mathbf{C}_g$ (see Tarantola 2005, section 6.21).

In practise, due to non-Gaussian distributions and the nonlinearity of $g(\mathbf{m})$, we often cannot obtain closed-form expressions for eqs (3) or (9) and we have to resort to methods based on generating random samples. As we now discuss, practical ways for making probabilistic inferences involve generating samples from either eq. (2) or eq. (7). Note that we drop the explicit conditioning on the assumptions A, B and C from now on.

2.1 Obtaining samples from the posterior pdf

If we can find a set of samples $\mathcal{D}_{\text{post}} = \{(\mathbf{m},\mathbf{d})_i\}$ distributed according to the pdf $\sigma(\mathbf{m},\mathbf{d}|\mathbf{d}_0)$ (eq. 2), we can use this set to directly make posterior inferences, for example, by plotting histograms (*cf.* Fig. 2, left-hand panel) or evaluating sample estimates of other posterior quantities of interest, such as means, standard deviations or covariance matrices. Since the sampling distribution is conditioned on the observation $\mathbf{d}_0$, we call the set $\mathcal{D}_{\text{post}}$ a set of *posterior samples*.

Obtaining samples from the posterior is challenging in many cases, since most of the posterior probability mass may be contained in only small regions of the model space. Once the regions of high posterior probability have been found, samples have to be generated such that the sampling density follows the posterior pdf. While there are a number of approaches (see e.g. MacKay 2003, for an overview), generally MCMC methods have proven to be very efficient for solving both problems, particularly if the dimensionality of the sampling space is high. MCMC algorithms are able to directly produce samples of the posterior pdf by constructing a Markov Chain that converges to the posterior pdf. In many implementations, this is done by performing a random walk in the model space and accepting or rejecting proposed samples based on an acceptance criterion that depends on the likelihood of the proposed sample. A disadvantage of that method is that samples are typically strongly correlated. This correlation can be dealt with by thinning the chains afterwards, that is, by only keeping every nth sample.

For comparisons presented in this paper, we adopt an MH algorithm (Hastings 1970), which is described in detail in Appendix A. This choice is mainly based on its simplicity and ease of implementation, and MH is sufficient for the comparisons we wish to perform. Nevertheless, a wide variety of advanced algorithms are available to the practitioner, with properties that are intended to be superior to plain MH in particular cases. Notable recent developments include trans-dimensional algorithms—which incorporate the problem of model selection into the inversion process (e.g. Sambridge *et al.* 2006; Bodin *et al.* 2012)—and iterative algorithms, also referred to as sequential Monte Carlo. These also aim to make probabilistic inversions feasible in time-limited situations, by progressively updating the posterior pdf as new data become available (e.g. Dettmer *et al.* 2011).

2.2 Making inferences with prior samples using probability density estimation

Generating samples from the posterior as in the previous section requires the knowledge about the observation $\mathbf{d}_0$ to be available at the time of sampling. However, it can be advantageous to perform the sampling stage, which is often computationally expensive, before a particular observation has been made. Furthermore, it is desirable to reuse the obtained set of samples for the inversion of several independent observations, which cannot generally be done if the set of samples is targeted at one specific observation. For example, a regional seismograph network may be used for the automatic characterization of local earthquakes on a routine basis. Each time an earthquake is observed, the inversion that is triggered is subject to essentially the same prior information and the same physical laws governing wave propagation. Although the observed data vector $\mathbf{d}_0$ changes every time an earthquake is observed, the Earth model used for calculating the synthetic seismograms and the known areas of regional seismicity are likely to stay the same between inversions.

In such a case, it is preferable to be able to re-use the same set of samples, and not have to re-generate them for each new earthquake.

Here, we aim to generate the set of samples $\mathcal{D}_{\text{prior}}$ distributed according to the pdf $\sigma(\mathbf{m}, \mathbf{d}_0)$ (eq. 7) rather than $\sigma(\mathbf{m}, \mathbf{d}|\mathbf{d}_0)$ (eq. 2). Since no information on $\mathbf{d}_0$ has been used for the generation of the samples we call $\mathcal{D}_{\text{prior}}$ a set of *prior samples*. If we choose the form of the distributions $\rho(\mathbf{d}, \mathbf{d}_0)$ and $\Theta(\mathbf{m}, \mathbf{d})$ such that their convolution $\Lambda(\mathbf{m}, \mathbf{d}_0)$ takes a well-defined form, it is straightforward to obtain $\mathcal{D}_{\text{prior}}$ by drawing a set of models from the prior distribution $\rho(\mathbf{m})$, solving the deterministic forward problem $g(\mathbf{m})$ for each and adding a noise component according to the combined theoretical and observational noise distribution. For example, in the case of Gaussian observational and modeling errors (eq. 10) with covariances $\mathbf{C}_g$ and $\mathbf{C}_d$, respectively, we have

$$\mathcal{D}_{\text{prior}} = \left\{(\mathbf{m}, g(\mathbf{m}) + \boldsymbol{\epsilon}_D)_i\right\} \tag{11}$$

where $\boldsymbol{\epsilon}_D$ is drawn from a normal distribution with zero mean and covariance $\mathbf{C}_g + \mathbf{C}_d$. Note that while it may be difficult to estimate $\mathbf{C}_g$ and $\mathbf{C}_d$, they do not have to be known independently under the Gaussian assumption, since they only enter the inference via the perturbation $\boldsymbol{\epsilon}$.

The different nature of the sets $\mathcal{D}_{\text{post}}$ and $\mathcal{D}_{\text{prior}}$ is depicted in Fig. 2. Note how the set $\mathcal{D}_{\text{post}}$ can directly be used to make inferences by plotting histograms, whereas by definition the distribution of the samples in $\mathcal{D}_{\text{prior}}$ does not carry information on the observation $\mathbf{d}_0$. Thus, in order to make probabilistic inferences based on the set $\mathcal{D}_{\text{prior}}$ additional steps are required.

In order to be able to evaluate the posterior pdf $\sigma(\mathbf{m}|\mathbf{d}_0)$ repeatedly for new observations $\mathbf{d}_0$, we will find an (approximate) representation of the conditional pdf (9) as a function of $\mathbf{d}_0$ based on the set $\mathcal{D}_{\text{prior}}$ denoted by

$$\tilde{p}(\mathbf{m}|\mathbf{d}_0; W) \approx \sigma(\mathbf{m}|\mathbf{d}_0). \tag{12}$$

The set of parameters W thereby controls the approximation and is determined—or *learned*—from the set $\mathcal{D}_{\text{prior}}$. This approach is often referred to as probability density estimation, and can be tackled using a wide variety of strategies. Here, in order to be able to make inferences repeatedly and efficiently for a potentially large number of independent observations, we adopt an approach based on machine learning—the MDN (Bishop 1995). MDNs are a general tool for conditional probability density estimation, and are based on neural networks, which may be viewed as general non-linear function approximators (Hornik *et al.* 1989). The resulting approximation can be evaluated very quickly with light-to-moderate demands on computational power and memory. Note, however, that most of our arguments are in fact independent of the specific tool used for estimating $\tilde{p}(\mathbf{m}|\mathbf{d}_0; W)$ and are in principle also applicable to other Bayesian machine-learning and regression methods.

An MDN forms a global parametric model of a conditional pdf and is based on a feed-forward neural network. As such it is capable of representing multimodal conditional pdfs, which might obey a highly non-linear relation between observations and model parameters over a wide range. Since their introduction by Bishop (1994), MDNs have been widely applied in many different areas, such as non-linear control problems (Herzallah & Lowe 2004), speech recognition (Richmond 2007), the reconstruction of spectral reflection curves (Ribés & Schmitt 2003), finance (Schittenkopf & Dorffner 2001), surf height prediction (Carney *et al.* 2005) and the inversion of satellite scatterometer data (Cornford *et al.* 1999). In particular, they have recently been used to efficiently solve seismological inverse problems such as the inversion of surface wave data for elastic Earth structure (Meier *et al.* 2007a,b), the determination of petrophysical parameters from seismic data sets (Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012), the determination of 1-D seismic Earth structure from body-wave traveltimes (de Wit *et al.* 2013), the inversion of normal mode observations for the Earth's radial elastic and anelastic structure (de Wit *et al.* 2014) and rapid probabilistic earthquake parameter estimation (Käufl *et al.* 2014, 2015). It is therefore of interest to understand how inferences made using MDNs within a prior sampling framework compare to solutions obtained by posterior sampling using MCMC algorithms.

If no further information on the shape of $\tilde{p}(\mathbf{m}|\mathbf{d}_0)$ is available, a general approach is to describe $\tilde{p}(\mathbf{m}|\mathbf{d}_0)$ as a sum of a fixed number of Gaussian kernels—a Gaussian Mixture Model (GMM). It can be shown that with a sufficient number of Gaussian kernels, any probability density can be approximated by a GMM to arbitrary accuracy (McLachlan & Basford 1988).

Since $\mathbb{M}$ may be high dimensional and eq. (12) thus difficult to present and interpret, as well as being difficult to estimate from the finite set of samples, we focus on the marginal pdfs

$$\tilde{p}(m_k|\mathbf{d}_0) = \int \tilde{p}(\mathbf{m}|\mathbf{d}_0)\mathrm{d}m_{i\neq k}, \tag{13}$$

where $\mathrm{d}m_{i\neq k} = \mathrm{d}m_1 \ldots \mathrm{d}m_{k-1}\mathrm{d}m_{k+1} \ldots \mathrm{d}m_c$ and c is the number of model space dimensions. Note that it is possible to generalize the method to higher dimensional pdfs as done in, for example, de Wit *et al.* (2013), and Käufl *et al.* (2014), although we do not consider this further in this paper. In particular, we have not investigated whether the results presented here generalize to the higher dimensional case. It is important to recognize that marginalization can mask subtleties in the underlying probability distribution, such as correlations or trade-offs between parameters. This must be borne in mind when results are interpreted and used.

We denote the MDN approximation of the marginal posterior pdf (13) by

$$p_{\text{MDN}}(m_k|\mathbf{d}_0, \mathbf{w}) = \sum_{i=1}^{M} \alpha_i(\mathbf{d}_0; \mathbf{w})\phi_i(m_k|\mathbf{d}_0), \tag{14}$$

where M is the number of kernels, $\alpha_i(\mathbf{d}_0; \mathbf{w})$ are mixture coefficients that sum to one, and

$$\phi_i(m_k|\mathbf{d}_0) = \frac{1}{\sqrt{2\pi}\sigma_i(\mathbf{d}_0; \mathbf{w})} \exp\left\{-\frac{[m_k - \mu_i(\mathbf{d}_0; \mathbf{w})]^2}{2\sigma_i(\mathbf{d}_0; \mathbf{w})^2}\right\} \tag{15}$$

are Gaussian kernels with mean $\mu_i(\mathbf{d}_0; \mathbf{w})$ and standard deviation $\sigma_i(\mathbf{d}_0; \mathbf{w})$. The parameters $\alpha_i(\mathbf{d}_0; \mathbf{w})$, $\mu_i(\mathbf{d}_0; \mathbf{w})$ and $\sigma_i(\mathbf{d}_0; \mathbf{w})$ are functions of $\mathbf{d}_0$ and are parametrized by a feed-forward neural network with parameters $\mathbf{w}$ (e.g. Bishop 1995). Typical values for M might lie in the range 3–10, depending on problem complexity; as we mention below, an extension to the method allows M to be treated as a random variable that is then marginalized out.

The network parameters $\mathbf{w}$ are determined during a training stage by maximizing the likelihood of a training set $\mathcal{D}_{\text{tr}} \subset \mathcal{D}_{\text{prior}}$. This is done by iteratively minimizing the loss function

$$E[\mathcal{D}_{\text{tr}}] = -\sum_n \ln p\left[(m_k)_n|\mathbf{d}_n, \mathbf{w}\right], \tag{16}$$

where the sum runs over all members of $\mathcal{D}_{\text{tr}}$. A typical training set for a problem with a modest number of free parameters might contain anything from 10^3 to 10^6 examples: obviously, larger training sets will generally allow for more detailed results, but be more computationally expensive to obtain.

In the examples shown below, we minimize (16) by means of the L-BFGS quasi-Newton method (Nocedal 1980), where the partial derivatives of (16) are calculated efficiently using error backpropagation (Rumelhart *et al.* 1986). In order to avoid overfitting to the details in the particular training set—that is, details that would not be present in all possible training sets obtained in the same fashion—the error of a second independent validation set $\mathcal{D}_{\mathrm{val}} \subset \mathcal{D}_{\mathrm{prior}}$, $\mathcal{D}_{\mathrm{tr}} \cap \mathcal{D}_{\mathrm{val}} = \emptyset$ is monitored during training and the optimal set of parameters is subsequently given by

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ E[\mathcal{D}_{\mathrm{val}}], \tag{17}$$

a procedure commonly referred to as early-stopping. However, it has been shown (e.g. Hjorth & Nabney 2000) that such a maximum likelihood approach can lead to biased results, particularly because the optimization problem given by eq. (16) can be highly non-linear and is in general non-convex, leading to multiple, possible equally likely solutions for the weight vector $\mathbf{w}^*$. A better approach is therefore to remove the explicit dependence of eq. (14) on the network parameters $\mathbf{w}$ by writing

$$p_{\mathrm{MDN}}(m_k|\mathbf{d}_0) = \int p(m_k|\mathbf{d}_0, \mathbf{w})p(\mathbf{w}|\mathcal{D}_{\mathrm{test}})d\mathbf{w}, \tag{18}$$

where $p(\mathbf{w}|\mathcal{D}_{\mathrm{test}})$ is the posterior probability of the parameter vector $\mathbf{w}$ given a third independent set of prior samples $\mathcal{D}_{\mathrm{test}} \subset \mathcal{D}_{\mathrm{prior}}$ with $\mathcal{D}_{\mathrm{test}} \cap \mathcal{D}_{\mathrm{tr}} = \emptyset$, $\mathcal{D}_{\mathrm{test}} \cap \mathcal{D}_{\mathrm{val}} = \emptyset$. However, the integral in (18) is hard to evaluate, since an expression for $p(\mathbf{w}|\mathcal{D}_{\mathrm{test}})$ cannot in general be obtained and the weight space is typically very high-dimensional, prohibiting numerical integration. While several solutions for this problem have been suggested (for an overview, see e.g. MacKay 1996), here we adopt a pragmatic approach and approximate the integral in (18) by the finite sum (Käufl *et al.* 2014)

$$p_{\mathrm{MDN}}(m_k|\mathbf{d}_0) = \sum_{i=1}^{C} \frac{\omega_i}{\sum_j \omega_j} p(m_k|\mathbf{d}_0, \mathbf{w}_i^*), \tag{19}$$

running over a number of C independently obtained MDNs, trained using different random weight initializations and training sets as described above, and $\mathbf{w}_i^*$ denotes the set of weights of the ith MDN. Each individual contribution to such an ensemble of MDNs is weighted by a factor of

$$\omega_i = \exp\left\{-\frac{E[\mathcal{D}_{\mathrm{test}}, \mathbf{w}_i^*]}{N}\right\}, \tag{20}$$

where $N = |\mathcal{D}_{\mathrm{test}}|$. Note that this approach could easily be extended to averaging over members with different numbers of mixture components M in order not to depend on a particular choice for M which may be hard to justify. However, for simplicity, and given that our experiments indicate our choice of M is not limiting in the cases discussed here, we do not adopt such a strategy. Again, the optimal number of committee members to use will be application-dependent, and a higher value for C will generally improve results (but at increased computational cost). As a very rough guide, we suggest that a reasonable starting point for C may lie in the range 10–20.

3 A 'CURSED' TOY PROBLEM

First, we demonstrate the framework and illuminate some of its properties by means of a probability density estimation toy problem. Suppose we wish to predict the location of a point in c-dimensional space, given its distance from the origin. The model parameters $\mathbf{m}$ are thus the coordinates of the point and the 'observable' datum d is taken to be its distance from the origin, contaminated by random noise. The 'forward problem' in this example is thus

$$g(\mathbf{m}) = ||\mathbf{m}||_2, \tag{21}$$

where $||\cdot||_2$ refers to the L2-norm. For the model parameters, we assume a uniform prior distribution in a c-dimensional cube, that is $\mathbf{m} \in \mathbb{M} = [-1, 1]^c$ and $\rho(\mathbf{m}) = \mathrm{const}$. We assume the forward theory to be exact and the observations subject to additive random Gaussian noise. Given an 'observed' value for $d_0 \in \mathbb{R}$, we are now interested in evaluating the posterior $\sigma(\mathbf{m}|d_0)$. We can do this analytically in the special case $d_0 = 0$ and we obtain approximate solutions by sampling directly from the posterior and by estimating the conditional posterior pdf from a set of prior samples.

In the special case that $d_0 = 0$, using eq. (5) with the Gaussian likelihood $L(\mathbf{m}) = \exp\left[-(d_0 - g(\mathbf{m}))^2/(2\sigma_d^2)\right]/(\sqrt{2\pi}\sigma_d)$ and the posterior becomes (see Appendix B)

$$\sigma(\mathbf{m}|d_0 = 0) = \frac{1}{(2\pi\sigma_d^2)^{c/2}} \exp\left[-\frac{\sum_{k=1}^{c} m_k^2}{2\sigma_d^2}\right], \tag{22}$$

with marginal distributions

$$p(m_k|d_0 = 0) = \frac{1}{\sqrt{2\pi}\sigma_d} \exp\left[-\frac{m_k^2}{2\sigma_d^2}\right]. \tag{23}$$

For $d_0 > 0$, the distribution becomes non-Gaussian with multimodal marginal distributions and we cannot easily derive a closed-form solution. See Fig. 3 for the case $c = 2$, where we have obtained $p(m_k|d_0 = 0.7)$ by numerical integration.

We obtain a set of prior samples $\mathcal{D}_{\mathrm{prior}} = \{(\mathbf{m}_i, d_i)\}$ by drawing $\{\mathbf{m}_i\}$ uniformly and assigning

$$d_i = ||\mathbf{m}_i||_2 + \epsilon_i, \tag{24}$$

where $\epsilon_i \in \mathbb{R}$ is a random number drawn from the normal distribution $\mathcal{N}(0, \sigma_d)$. Thus, the samples approximately lie on the submanifold $d = ||\mathbf{m}||_2$ of the joint data-model space. Fig. 3 shows slices through the space at $d = 0$ and $d = 0.7$ for $c = 2$ and $\sigma_d = 0.1$, where areas of higher sampling density are shaded in darker colours. Note that the distribution of samples in the data space is far from uniform despite the uniform prior distribution of model parameters, due to the non-linearity of the forward problem. Therefore, in the following, we evaluate the posterior at two different d_0, corresponding to areas of relatively low and high sampling density in the joint data-model space.

We construct the approximation $p_{\mathrm{MDN}}(m_k|d_0)$ from the set $\mathcal{D}_{\mathrm{prior}}$ using MDNs, as described in the previous section, for different choices of $c \geq 2$. Experiments indicated that it is sufficient to use $M = 3$ mixture components for the individual MDNs, to accurately capture the bimodal behaviour of the pdf: taking a higher value for M does not lead to markedly different results, but does increase computational costs. The influence of the parameter M on the predicted pdfs is discussed in more detail below.

We subsequently evaluate the trained MDNs at $d_0 = 0$ and $d_0 > 0$, respectively. In the case $d_0 = 0$, we compare the results to the analytical solution (23) and for $d_0 > 0$ to an approximate solution obtained by sampling directly from the posterior using a MH sampling algorithm. See Appendix A for implementation details. The analytical and MCMC solutions hereby serve as a benchmark for the MDN estimate. We ran each chain for a fixed number of iterations and while we did not employ any formal convergence criteria, we verified by visual inspection that the posterior did not change significantly if further iterations were performed.

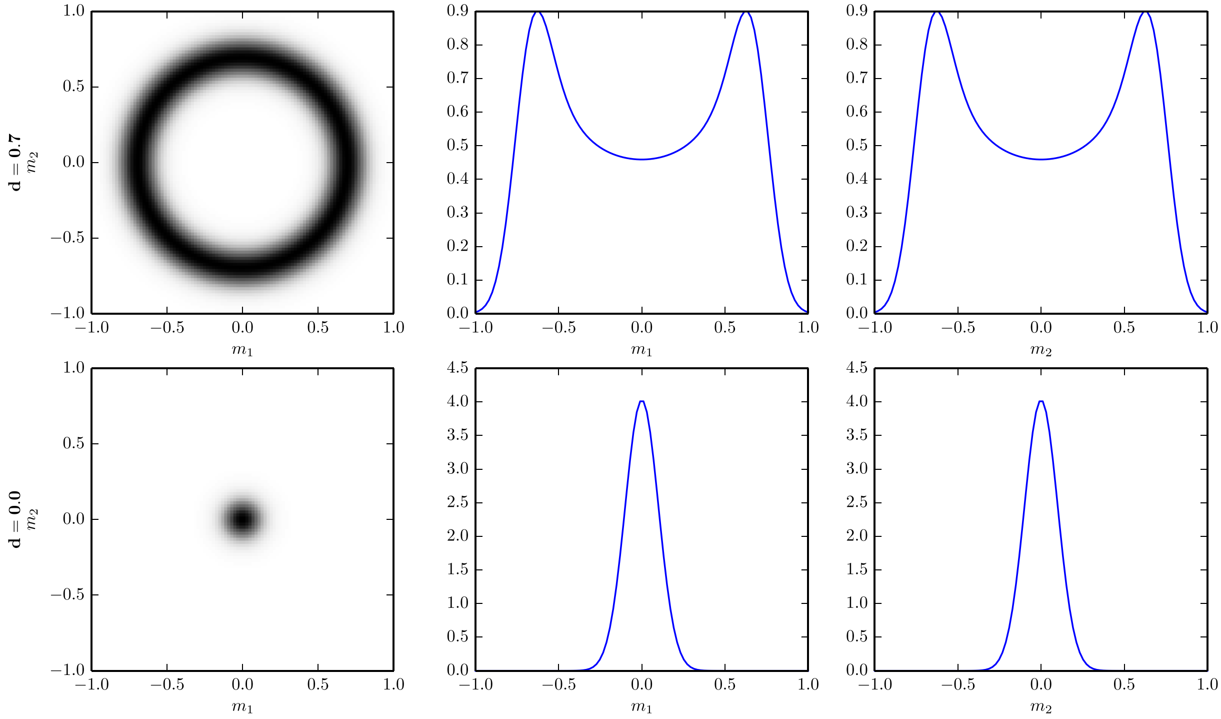


Figure 3. Top row, left: posterior $p(\mathbf{m}|d_0 = 0.7)$ for the case $n = 2$. Middle and right: marginals $p(m_1|d_0 = 0.7)$ and $p(m_2|d_0 = 0.7)$, respectively. Bottom row: the same, but for $d_0 = 0.0$. Darker colours correspond to higher probabilities.

3.1 Discussion

Fig. 4 shows the MDN estimates $p_{\mathrm{MDN}}(m_1|d_0)$ for $c \in \{2, 5, 10\}$. In the cases $c \le 5$, each MDN ensemble comprises 10 members and in the case $c = 10$, 15 members that have been trained on sets of 5000 prior samples. The variation in the individual ensemble members is caused by the different random initializations of the network weights and the differing noise components ϵ_i, respectively, that are re-drawn for the training of each member. Note that while two runs of the MCMC sampler were required for each choice of c—one for $d_0 = 0$ and one for $d_0 = 0.7$—only one MDN ensemble has been trained and subsequently evaluated twice for different values of d_0.

We see from Fig. 4 that the MDN approximation is good in regions where there is plenty of training data available, but it deviates further from the desired distribution, the more it has to extrapolate into regions with little or no training data. It appears that the output of an MDN ensemble gives a more conservative estimate of the true posterior pdf—in the sense that it is broader than the desired distribution—if insufficient training data are available. In general terms, this follows intuitively from the less-targeted nature of prior sampling. However, this is not a rigorously proven property; indeed, the point-source inversion example below demonstrates that while it holds in most cases, the test set also contains a small number of counter examples.

The effect can be understood as follows: by making inferences using a fixed set of samples $\mathcal{D}$, rather than allowing $\mathbf{m}$ to vary continuously, we are effectively replacing the exact forward problem $g(\mathbf{m})$ with the piecewise constant approximation

$$\tilde{g}(\mathbf{m}) = g\left(\operatorname*{argmin}_{\mathbf{m_i}\in\mathcal{D}} ||\mathbf{m} - \mathbf{m}_i||\right). \qquad (25)$$

This replacement introduces the discretization error

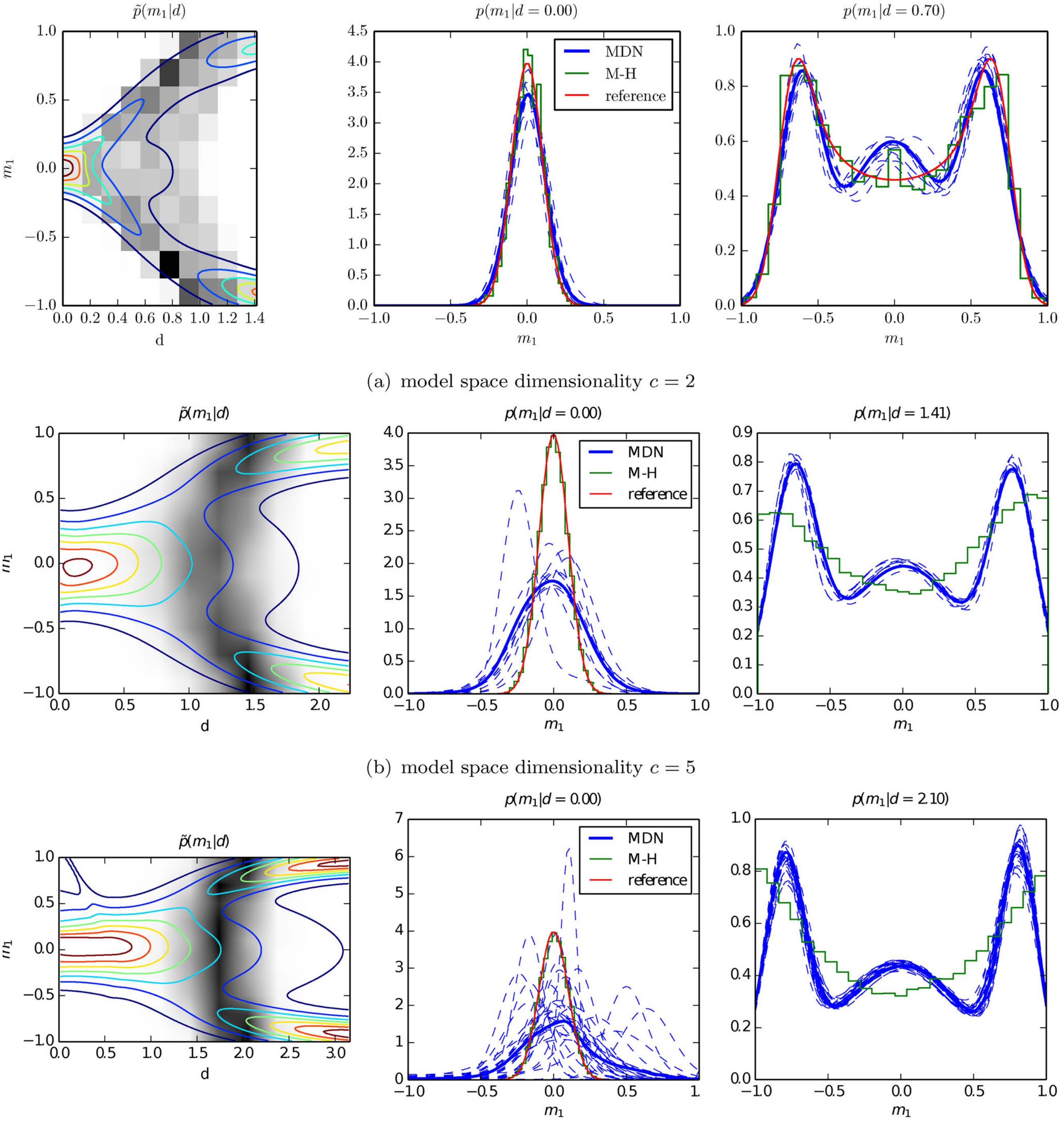
$$\boldsymbol{\epsilon}_g(\mathbf{m}) = g(\mathbf{m}) - \tilde{g}(\mathbf{m}). \qquad (26)$$

If we make the assumption that $g(\mathbf{m})$ is sufficiently linear in the vicinity of any given prior sample $\mathbf{m}_i$, we can express the probabilistic correlation between model parameters $\mathbf{m}$ and predictions $\mathbf{d}$ under the influence of the discretization error by

$$\tilde{\Theta}(\mathbf{d}|\mathbf{m}) = (2\pi)^{-\frac{k}{2}} |\mathbf{C}_g(\mathbf{m})|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2}\left[\mathbf{d} - \tilde{g}(\mathbf{m})\right]^T \mathbf{C}_g(\mathbf{m})^{-1}\left[\mathbf{d} - \tilde{g}(\mathbf{m})\right]\right\} \qquad (27)$$

where the covariance matrix $\mathbf{C}_g(\mathbf{m})$ represents the uncertainty induced by the discretization error in the vicinity of the point $\mathbf{m}$. If observational errors are assumed to be Gaussian (as in this toy problem) and if we replace the exact forward relation $\Theta(\mathbf{d}|\mathbf{m})$ with $\tilde{\Theta}(\mathbf{d}|\mathbf{m})$ in eq. (8), then from eq. (10) we see that the combined observational and theoretical error covariances simply sum up to give the combined covariance matrix $\mathbf{C}_D(\mathbf{m}) = \mathbf{C}_d + \mathbf{C}_g(\mathbf{m})$, in the vicinity of the point $\mathbf{m}$. As a consequence, a training set generated under the assumption of a perfect forward operator, but with an insufficient sampling density (in the sense that the discretization error will be larger than the observational errors) cannot be distinguished from a training set where a theoretical Gaussian modeling error with covariance $\mathbf{C}_g(\mathbf{m})$ is assumed, thus imposing a lower bound on the posterior uncertainties.

When using a technique such as the MDN we are implicitly estimating the covariance $\mathbf{C}_D(\mathbf{m})$ from the set of prior samples $\mathcal{D}_{\mathrm{prior}}$ during the training process. In practice, this is aided by several measures. First, each ensemble member is initialized in such a way that it will output the marginal prior distribution over the model space, that is initially we have $p_{\mathrm{MDN}}(m_k|\mathbf{d}_0) = \rho(m_k)$. During the training procedure, the output distribution is then incrementally refined to resemble the distribution of the training data. Second, by evaluating the error of an independent validation set (see eq. 17) we stop the training process if the performance on the validation set would deteriorate by a further refinement of the pdf learnt thus



(a) model space dimensionality $c = 2$

(b) model space dimensionality $c = 5$

(c) model space dimensionality $c = 10$

Figure 4. The left-hand panel in each row (a)–(c) shows the conditional pdf $p_{\mathrm{MDN}}(m_1|d_0)$ (coloured contour lines) approximated by an ensemble of MDNs and the density of the training data (shaded areas, darker colours refer to higher sampling density). The middle panel shows the MDN ensemble's (solid blue) and the individual ensemble members' prediction (dashed blue), a set of samples from the posterior obtained using a Metropolis–Hastings sampler (green histogram) and the analytical solution (solid red) for the distribution $p(m_1|d_0 = 0)$. The right-hand panel shows the case $p(m_1|d_0 = 0.7)$. The reference (red line) in the right-hand panel of row (a) refers to a solution obtained using numerical integration, rather than an analytical solution. In higher dimensions $c > 2$ this is not feasible, however, and therefore no objective reference is available.

far. Thereby, we effectively avoid underestimating the covariance $\mathbf{C}_D(\mathbf{m})$. This is related to the well-known bias-variance trade-off in regression problems (Bishop 1995). Finally, when moving into undersampled regions of the joint data-model space, the answers given by different ensemble members will typically show an increasing variability, since the functional form of the mapping from data into model space is less well constrained by the training set in those regions. By averaging over several independently trained ensemble members and within the bounds on flexibility imposed by factors such as the neural network architecture and the number of network

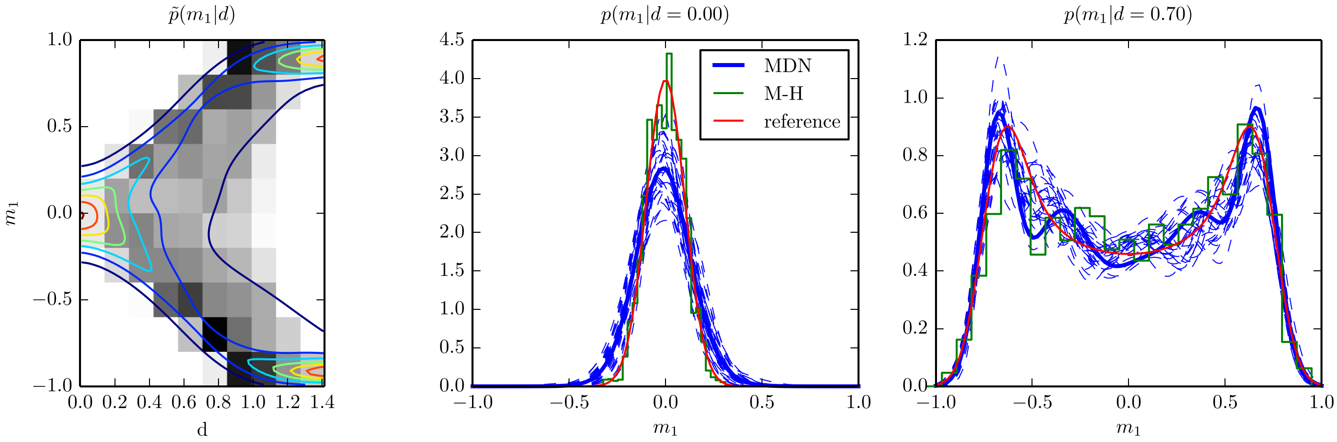


Figure 5. Same as Fig. 4(a), but each ensemble member uses five Gaussian kernels. The increased flexibility is used to follow the target distribution more closely. Note that more ensemble members are required to stabilize the distribution, due to their increased variations.

parameters, we are approximately marginalizing over the space of MDNs compatible with the training data (see eq. 19). In Fig. 4 (central column), the increasing variability of the ensemble members (dashed blue lines) becomes apparent, when the dimensionality c is increased and therefore the sampling density in the vicinity of $d_0 = 0$ decreased. In the case of a sufficiently high sampling density (Fig. 4 (a) central, and (a) to (c) right-hand panel), however, all members provide essentially the same answer. It is clear, however, that $g(\mathbf{m})$ is required to be sufficiently smooth for this procedure to work. A local roughness of the function $g(\mathbf{m})$ would be unlikely to be detected by either the training or the validation set, if $g(\mathbf{m})$ varies strongly on a length scale much shorter than the typical distance of the samples.

Another factor influencing the quality of the MDN approximation to the posterior is the number of mixture components M. For the generation of Fig. 5, we set $M = 5$. The target distribution is matched more closely, while at the same time the variability of the ensemble members increases due to the increased number of degrees of freedom. This can be counteracted by in turn increasing the number of ensemble members. The parameter M thus acts as a regularization parameter controlling the bias-variance trade-off of the individual ensemble members. Alternatively, as already mentioned, we could easily extend the approach to encompass members with a varying number of mixture components, effectively marginalizing over different choices of M. Rather than fixing M to one particular value we would then have to define a suitable prior distribution from which to draw M.

Finally, a particular aspect of this toy problem, related to the well-known 'curse of dimensionality' (e.g. Curtis & Lomax 2001)—a problem from which most high-dimensional real world inverse problems suffer—is worth highlighting. When increasing the dimensionality of the model space, the bulk of the prior samples in the joint data-model space get concentrated 'far away' from the location of the target model $(\mathbf{m}_0, d_0) = (\mathbf{0}, 0)$, despite it being at the centre of the model space and therefore well covered by all marginal prior distributions $\rho(m_k)$. Intuitively, we expect the MDN approximation to become more accurate as more samples are being added to the training set. However, the nature of this toy problem is such that even a very large number of prior samples would not be likely to improve the approximation near $d_0 = 0$. This can be seen from the expected value of the likelihood of the training set as a function of the model space dimensionality c

$$E\left[L(\mathbf{m}|d_0 = 0)\right] = \int L(\mathbf{m}|d_0 = 0)\rho(\mathbf{m})\mathrm{d}\mathbf{m}$$
$$= \left(\frac{1}{2}\right)^c \left(2\pi\sigma_d^2\right)^{\frac{c-1}{2}}, \qquad (28)$$

which is plotted in Fig. 6 (light blue line). Also shown are averages taken over a set of random samples of varying size. Note that while the sample average of (28) over the training set will asymptotically approach the theoretical relation as more samples are being added, the average likelihood is bounded from above by eq. (28). This suggests that the approximation cannot be improved upon significantly by increasing the number of prior samples, since the majority of samples will continue to fall into regions of low likelihood, which will subsequently dominate the contribution to the error in the neural network training stage (see eq. 16). The effect is particularly severe in the context of this toy example, which has been designed so that the probability density $\sigma(d_0 = 0)$ rapidly decreases with increasing c. However, many real world inverse problems suffer from similar effects and we cannot in general expect that more samples will significantly improve the prediction accuracy of the MDN approximation.

This reveals a fundamental difference between an approach based on prior samples following $\sigma(\mathbf{m}, \mathbf{d}_0)$ and an approach that generates samples from the posterior $\sigma(\mathbf{m}|\mathbf{d}_0) = \sigma(\mathbf{m}, \mathbf{d}_0)/\sigma(\mathbf{d}_0)$. When basing our inference on prior samples we naturally do not take the denominator $\sigma(\mathbf{d}_0)$—the probability of making the particular observation $\mathbf{d}_0$—into account (see also the right-hand panel in Fig. 1b). This probability may be very small for a given $\mathbf{d}_0$ compared to other possible outcomes and therefore the set of prior samples may be very unlikely to contain any samples close to a given observation $\mathbf{d}_0$, even if a large number of prior samples are being used. Therefore, we expect that a prior sampling based estimate of a posterior quantity deviates further from an estimate obtained by posterior sampling, the smaller $\sigma(\mathbf{d}_0)$ becomes. Factors that influence $\sigma(\mathbf{d}_0)$ are the non-linearity of the forward operator $g(\mathbf{m})$, the choice of prior model parameter distributions and the parametrization itself. In the case of this toy example, we could easily improve upon the situation either by requiring the model parameters to be correlated, thereby reducing the intrinsic dimensionality of the problem or by

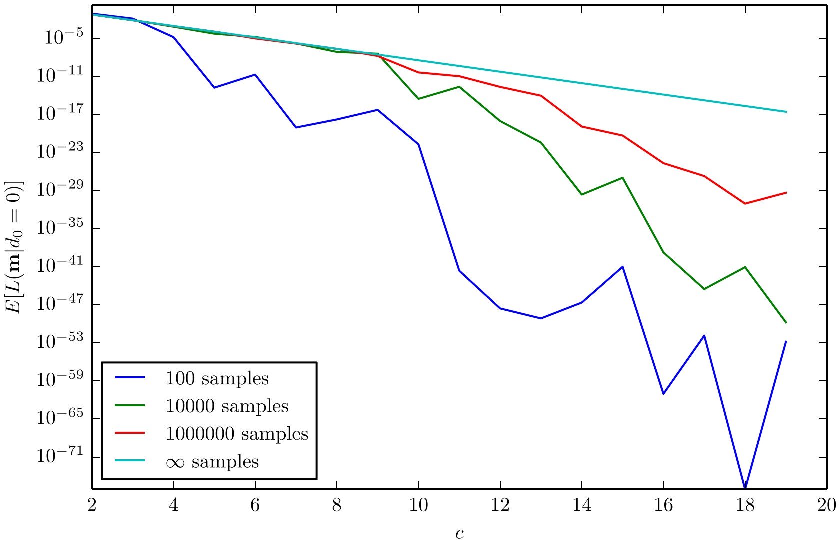


Figure 6. Expectation value of the likelihood $L(\mathbf{m}|d_0 = 0)$ plotted on a log scale as a function of the number of dimensions c for a uniform prior distribution (light blue line). Sample average of the likelihood for three data sets drawn randomly from the prior consisting of 10^2 (dark blue), 10^4 (green) and 10^6 (red) samples.

parametrizing the problem in terms of spherical coordinates centred at the origin. In this case, all the variation in the data could be explained by a single degree of freedom. This suggests that the choice of prior distribution and model parametrization plays a crucial role, even more so than in approaches based on sampling from the posterior.

3.2 Conclusions

From this toy problem, we have learned that an approach based on estimating the conditional posterior probability density as a function of the observation $\mathbf{d}_0$ from a set of prior samples can in principle work, although we have to deal with several intrinsic limitations. Nevertheless, there are many cases in which such an approach will have advantages over traditional deterministic techniques or expensive Monte Carlo sampling. This is mainly due to the fact that the set $\mathcal{D}_{\text{prior}}$ can be re-used for repeated inversions of new observations, allowing conservative uncertainty estimates can be obtained in a computationally efficient manner. In the remainder of this paper, we will apply the method to a realistic seismological point-source parameter estimation problem and investigate whether the underlying assumptions are satisfied and if the MDN approximation thus forms a meaningful approximation to the posterior pdf.

4 A SEISMOLOGICAL EXAMPLE—INVERSION OF COSEISMIC DISPLACEMENT OBSERVATIONS FOR POINT-SOURCE PARAMETERS

To first order an earthquake can be described by a moment-tensor point source (Burridge & Knopoff 1964) and a typical non-linear inverse problem in seismology is the joint determination of point-source parameters such as epicentral location, hypocentre depth, magnitude and source mechanism from observed seismic or geodetic data. Such inversions are carried out in an automated manner on a routine basis after seismic events are detected using data from regional and global seismograph networks (e.g. Ekström *et al.* 2012). Often this inverse problem is solved using an iterative least-squares approach (Dziewonski *et al.* 1981), sometimes under additional constraints to suppress spurious non-double-couple components (e.g. Liu *et al.* 2004). The small number of parameters, in the order of a few to tens of parameters, is suited to a Bayesian treatment using sampling based methods (e.g. Stähler & Sigloch 2014). In particular for EEW or disaster response, however, it is important to rapidly determine source parameters after first observations have been made—a situation in which repeated computation of the forward problem, as required by MCMC sampling or iterative gradient-based approaches, are prohibitive. This is especially true if the forward problem is computationally expensive, for example, if realistic 3-D heterogeneous Earth models are to be used. Therefore, real-time source estimates are typically either obtained using empirical attenuation relations to rapidly determine the magnitude (an overview is given by Kanamori 2005) or by using a database of pre-computed Green's functions to which observations can subsequently be compared in order to determine the source mechanism (e.g. Lee *et al.* 2010). The former approach requires large, high-quality databases of observed earthquake records and corresponding source estimates, which often lack a significant amount of large events, limiting their applicability to large earthquakes. The latter approach is limited by the requirement to keep a potentially very large waveform database in the memory of a computer and to develop efficient algorithms to compare them to the observed waveform data, which can often only be done in big data centres using large-scale computing facilities.

It appears that MDN ensembles could nicely bridge the gap between empirical regression methods and approaches based on wave propagation modeling, by incorporating MDN ensembles to interpolate between a set of pre-computed samples and quickly output a probabilistic prediction on the source parameters of interest. Käufl *et al.* (2014) investigate the feasibility of such an approach for the rapid inversion of static coseismic displacement observations. Käufl *et al.* (2015) extend the approach to the real-time inversion of waveform data. A neural network thereby forms a highly compact and rapidly evaluable representation of a pre-computed Green's function database. For comparison: a state-of-the-art regional real-time moment-tensor monitoring system operational in Taiwan (Lee *et al.* 2013) uses a 1-D Green's function database consisting of more than 60,000 Green's functions, stored on a 32-CPU cluster, on which a

Table 1. Point-source parametrization and prior distributions.

Parameter	Prior distribution*	Description
κ	$\mathcal{U}(0, 2\pi)$	Strike (periodic)
σ	$\mathcal{U}(-\frac{\pi}{2}, \frac{\pi}{2})$	Rake
h	$\mathcal{U}(0, 1)$	cos (dip)
M_w	$\mathcal{U}(5.0, 8.0)$	Moment magnitude
lat	$\mathcal{U}(32.0, 33.5)$	Centroid latitude [°]
lon	$\mathcal{U}(-117.0, -114.5)$	Centroid longitude [°]
depth	$\mathcal{U}(1.5, 20)$	Centroid depth [km]

* $\mathcal{U}(a, b)$ denotes a uniform distribution on the interval $[a, b]$.

search is performed in parallel, in order to be able to perform an inversion every 2 s. More detailed Earth models and the ability to accurately model high-frequency data are likely to further increase these requirements in the future. On the other hand, the trained neural network ensembles used in Käufl *et al.* (2015) typically comprise 10^4–10^6 parameters—for each source parameter to be determined. The set of network ensembles required to perform real-time inversions thus easily fits the memory of a standard desktop computer.

We adopt the parametrization and setup introduced in Käufl *et al.* (2014) to investigate to what extent the MDN predictions are compatible with a Monte Carlo solution. Earthquakes are thereby described by a set of seven independent source parameters with uniform prior distributions as given in Table 1. The parameters κ, σ and h govern the orientation, M_w the magnitude of the source. The parameters lat, lon and depth determine the spatial location within the Earth. For further details on the implementation (see Käufl *et al.* 2014, 2015).

4.1 Results

In the following, we compare two estimates of the marginal posterior pdfs $\sigma(m_k|\mathbf{d}_0)$ for the probabilistic point-source estimation problem by means of a synthetic test. We generate a test set $\mathcal{D}_{\text{test}} = \{(\mathbf{m}, \mathbf{d})_i\}$ by drawing 50 samples from the uniform prior distribution independently. We calculate synthetic observations by solving the forward problem $\mathbf{d}_0 = g(\mathbf{m}) + \boldsymbol{\epsilon}_d$, where $\boldsymbol{\epsilon}_d$ is a noise vector distributed according to a normal distribution with covariance $\mathbf{C}_d$ and $g(\mathbf{m})$ is calculated using a deterministic code (O'Toole & Woodhouse 2011) which simulates wave propagation in a 1-D layered elastic medium. We use the MCMC sampling procedure described in Appendix A to obtain a reference solution for the 50 test set examples.

Subsequently, we also present the noisy, synthetic data vectors $\{\mathbf{d}_i\}$ to a set of MDN ensembles trained using an independent training set $\mathcal{D}_{\text{tr}}$ containing 100 000 samples. Throughout this section, we denote the MDN approximation to the posterior according to eq. (19) by $p_{\text{MDN}}(m_k|\mathbf{d})$ and the reference solution obtained by MCMC sampling as described above by $p_{\text{MH}}(m_k|\mathbf{d})$.

We found that the MCMC sampler did not always converge within a fixed 25 000 iterations, chosen to limit the computational cost. We identify these cases by means of a goodness-of-fit test. As explained in detail in Appendix A, for each test set example $(\mathbf{m}, \mathbf{d})_i$, we have H (thinned and burn-in–corrected) Markov chains $(\mathbf{m}_n)_h^{(i)}$ of length N. We denote by

$$(\chi_n^2)_h^{(i)} = \frac{\left[\mathbf{d}_i - g\left(\mathbf{m}_{n,h}^{(i)}\right)\right]^T \mathbf{C}_d^{-1} \left[\mathbf{d}_i - g\left(\mathbf{m}_{n,h}^{(i)}\right)\right]}{\nu} \tag{29}$$

the reduced misfit of the nth sample in the hth chain for test set example i, where $\nu = D - M - 1$ is the number of effective degrees of freedom. Moreover, we calculate the average reduced misfit

$$\left(\overline{\chi^2}\right)_h^{(i)} = \sum_{n=1}^{N} (\chi_n^2)_h^{(i)} . \tag{30}$$

If a Markov chain has converged, the reduced misfit (29) follows a χ^2 distribution with mean 1.0. We therefore discard test set example i, if

$$\left[\left(\overline{\chi^2}\right)_h^{(i)} - 1.0\right] > 0.5 \tag{31}$$

for all H chains. This criterion serves as a pragmatic way of excluding the examples for which the MCMC sampler did not produce models with sufficiently high likelihood, while keeping chains that are reasonably close to an optimal solution, even if they have not yet reached convergence. We are not interested in the former, since they do not provide a meaningful reference for comparison with the MDN estimates. Based on this criterion, 9 of the 50 test set examples were excluded. Potentially, more advanced search algorithms could have provided faster convergence and more robust results (e.g. Geyer 1991; Sambridge 1999; Skilling 2006; Dosso & Dettmer 2011; Dettmer & Dosso 2012). However, such attempts are outside of the scope of this paper.

We denote the probabilities assigned to the intervals $[t_k; t_k + \delta_k]$ and $[t_k - \delta_k; t_k]$ left and right of the target value t_k, respectively, by

$$P^+ = \int_{t_k}^{t_k+\delta_k} p(m_k|\mathbf{d})\mathrm{d}m_k \tag{32}$$

and by

$$P^- = \int_{t_k-\delta_k}^{t_k} p(m_k|\mathbf{d})\mathrm{d}m_k, \tag{33}$$

where δ_k is chosen to be 5 per cent of the prior range for the kth parameter. The quantity

$$P(|t_k - m_k| \le \delta_k) = P^+ + P^- \tag{34}$$

gives the total probability assigned to a region of width $2\delta_k$ around the target value. Note that for parameter κ we let the integrals (32) and (33) wrap around at the boundaries of the interval $[0; 2\pi]$ to take into account the 2π-periodicity. We subsequently use (34) to assess if the estimated pdf assigns probability to the 'right' region in model space. If the posterior equals the prior—that is nothing could be learned from the data—we have $P(|t_k - m_k| \le \delta_k) = 0.1$. The quantities P^+ and P^-, respectively, can be used to identify any potential local bias—that is a case in which more probability is assigned to either side of the target value—in the vicinity of the target value.

Furthermore, we use the Kullback–Leibler divergence

$$D_{\text{KL}}\left[p(m_k|\mathbf{d})\right] = \int \ln\left(\frac{p(m_k|\mathbf{d})}{\rho(m_k)}\right) p(m_k|\mathbf{d})dm_k, \tag{35}$$

to estimate the relative change in uncertainty when moving from the prior $\rho(m_k)$ to the posterior pdf $p(m_k|\mathbf{d})$. A higher D_{KL} value indicates that the posterior distribution is narrower relative to the prior pdf, whereas a value of zero would indicate that the two distributions are identical and nothing has been learned upon seeing the data $\mathbf{d}$. In order to compare MDN and MCMC estimates, we calculate the information gain difference

$$\Delta D_{\text{KL}} = D_{\text{KL}}\left[p_{\text{MH}}(m_k|\mathbf{d})\right] - D_{\text{KL}}\left[p_{\text{MDN}}(m_k|\mathbf{d})\right], \tag{36}$$

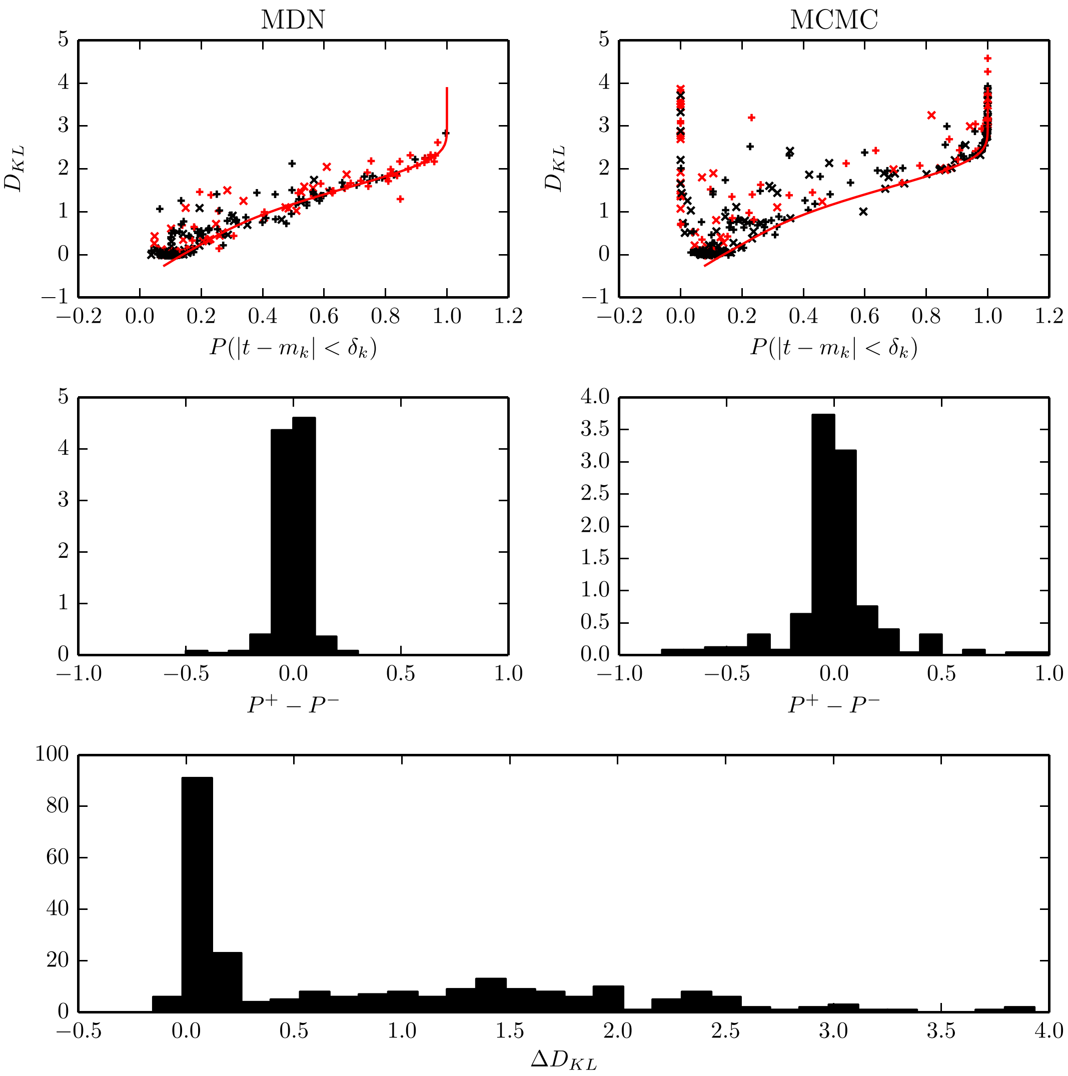


Figure 7. Comparison of 350 1-D marginal distributions obtained by MCMC sampling of the posterior, and by training MDNs on prior samples, respectively. The red line corresponds to a Gaussian pdf located at the centre of the model space with decreasing standard deviation for reference. Dots marked in red correspond to the 14 examples for which the MCMC sampling procedure did not produce samples with a sufficiently high likelihood. The corresponding test set examples have therefore been excluded in the histograms in row two and three (see the main text). The 'x'-symbols correspond to parameters governing source orientation, '+'-symbols correspond to source location, depth and magnitude. Top: accuracy of the estimated 1-D pdfs. The horizontal axis shows the total probability $P(|t - m_k| < \delta_k)$ assigned to an interval of width $2\delta_k$ around the target value. The vertical axis shows the information gain D_{KL} with respect to the uniform prior. Middle: estimation of the bias in the vicinity of the target value. Shown are histograms of the difference between the probability $P^-(m_k > (t - \delta))$ and $P^+(m_k < (t + \delta))$. Bottom: distribution of the information gain difference ΔD_{KL} between MDN and MCMC estimates.

which is a measure of the relative difference between the MDN and the MCMC solutions. Again, in the case that the posterior equals the prior, that is $p_{\mathrm{MH}} = p_{\mathrm{MDN}}$, $\Delta D_{\mathrm{KL}} = 0$. Note that if $\Delta D_{\mathrm{KL}} \geq 0$ we can consider p_{MDN} to be a conservative estimate of the reference density p_{MH}, in the sense that p_{MH} is more informative than p_{MDN}, since it narrows down our prior belief to a larger extent than p_{MDN}.

For each of the seven source parameters m_k, we obtain $p_{\mathrm{MDN}}(m_{k,i}|\mathbf{d}_i)$ and $p_{\mathrm{MH}}(m_{k,i}|\mathbf{d}_i)$ for all test set examples. Note that in order to obtain $p_{\mathrm{MH}}(m_{k,i}|\mathbf{d}_i)$, we have to run 5 Markov chains of length 25 000 for each of the 50 examples, that is we need to solve the forward problem a total of 6 250 000 times. In order to obtain the MDN approximation, we only generate a total of 100 000 prior samples and subsequently train 7 MDN ensembles with 15 members each. Depending on the computational demands of the forward problem and how accurate the MDN approximation has to be, the total computational cost for generating the training set and subsequently training the MDNs in order to perform 50 inversions can be significantly lower.

We calculate $P_{\mathrm{MH}}(|t_k - m_k| \leq \delta_k)$, $p_{\mathrm{MDN}}(|t_k - m_k| \leq \delta_k)$, $D_{\mathrm{KL}}[p_{\mathrm{MH}}(m|\mathbf{d})|p(m)]$ and $D_{\mathrm{KL}}[p_{\mathrm{MDN}}(m|\mathbf{d})|p(m)]$, respectively, for the 7 marginal distributions in the 50 test cases and plot the resulting 350 probabilities versus the corresponding D_{KL} values in Fig. 7 (top row). In the cases where $D_{\mathrm{KL}} \approx 0$, we expect $P(|t_k - m_k| \leq \delta_k) \approx 0.1$. Similarly, if D_{KL} is large, indicating that there are regions to which the posterior gives preference, we expect $P(|t_k - m_k| \leq \delta_k)$ to also be large, indicating that the regions of high probability are indeed assigned to the vicinity of the target value. For reference, the behaviour of a Gaussian distribution with decreasing width, located at the target value at the centre of the model space is plotted as red line in Fig. 7 (top). The saturation $P(|t_k - m_k| \leq \delta_k)$ for large D_{KL} values is due to the fact that eventually the entire posterior

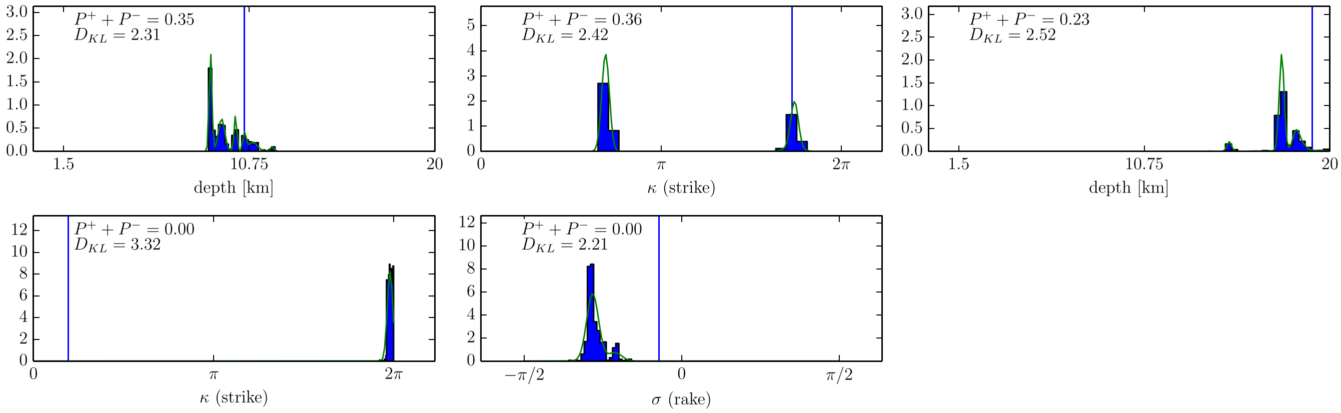


Figure 8. Examples of 1-D marginal pdfs (green curve) for different parameters estimated from posterior samples obtained by MCMC sampling (blue histograms). The target value is shown as vertical blue line. The five examples correspond to outliers in Fig. 7 (top right), for which the ratio of D_{KL} to $P(|t - m_k| < \delta_k)$ deviates from the value expected for a unimodal Gaussian centred around the target value (red line in Fig. 7). Causes for non-convergence are multimodality, periodicity and missed out modes (see the main text).

probability mass will be assigned to the finite interval $[t_k - \delta_k; t_k + \delta_k]$. Despite removal of the non-converging examples, we observe a number of outliers in Fig. 7 (top, right). We verified by visual inspection that these are not problematic and give a few examples in Fig. 8. Most of the outliers correspond to the parameters governing the source orientation (x-symbols in Fig. 7), which typically show strong multimodal behaviour due to symmetries in $g(\mathbf{m})$. It appears that the sampler occasionally gets stuck in one of the modes which happens not to be the one corresponding to the target value, but which still gives an acceptable data fit. This could potentially be resolved by either running more chains or increasing the number of MCMC iterations. Moreover, the mixing of the Markov chains in the case of parameter κ could potentially be improved by allowing the parameter to wrap around at the boundaries of the model space (*cf.* Fig. 8, panels one, three and five). Fig. 7 (second row) also reveals that neither the MCMC nor the MDN estimates show a significant local bias, as indicated by the difference $P^+ - P^-$. A value of zero indicates that the probability assigned to an interval of width δ_k left of the target value equals the probability assigned to the according interval right of the target value.

Finally, we plot a histogram of ΔD_{KL} in Fig. 7 (third row). The fact that $\Delta D_{\mathrm{KL}} \gtrsim 0$ in almost all cases confirms that p_{MDN} is less informative than p_{MH} and leads to the conclusion that the MDN approximation can be considered a conservative estimate of the posterior probability density. A comparison between marginal pdfs obtained by means of the two methods is given in Fig. 9 for three test set examples. The example plotted in the first row of Fig. 9 corresponds to a relatively small earthquake, which has a poor signal-to-noise ratio at almost all receivers. Therefore, neither the reference pdf $p_{\mathrm{MH}}(m_k|\mathbf{d})$ (green curve), nor the MDN estimate $p_{\mathrm{MDN}}(m_k|\mathbf{d})$ (black curve) can constrain any of the parameters except magnitude M_w. Note also that the difference between the two distributions, as quantified by ΔD_{KL}, is negligible. In the second case (second row of Fig. 9), most parameters can be constrained rather well by the MCMC estimate. To all seven parameters, the MDN estimates assign a larger posterior uncertainty than the reference pdf as expected. Finally, the third example (third row) shows the most negative ΔD_{KL} value (for parameter 'lon') observed across the test set. Note, however, that this is a singular outlier as can be seen from the histogram of ΔD_{KL} values in Fig. 7.

5 DISCUSSION

5.1 Applicability of the method to high-dimensional problems

Throughout this paper, we have only considered relatively low-dimensional problems, and it is reasonable to ask how performance scales with the dimensionality of the sampling space. Since the method is solely based on samples that are obtained before any measurement was made, we cannot—as in MCMC methods—exploit a random walk to explore the model space. Instead, our method bears some resemblance to importance sampling, assuming the sampling distribution is chosen to be the prior distribution. Therefore, we expect the method to show similar scaling behaviour to importance sampling—which, admittedly, is known to scale badly with the number of model space dimensions (MacKay 2003). It is clear that additional work is required to explore how prior sampling should be implemented in high-dimensional problems; as with importance sampling it is difficult to know in advance the number of samples required in order for the algorithm to converge to the true posterior pdf. However, within the framework proposed here, this can be tested in any particular case, by monitoring the test set performance.

In addition, we have the advantage that the MDNs are initialized so that they output a representation of the prior pdf regardless of inputs. The training procedure is then designed so that they only 'learn' information from the training set, $\mathcal{D}_{\mathrm{tr}}$, if it also improves the system's ability to explain the independent data set, $\mathcal{D}_{\mathrm{test}}$. This helps to prevent the MDNs from focusing on learning the particular characteristics of individual samples, and promotes generalization. If insufficient samples are available for a given problem, MDN updates computed from $\mathcal{D}_{\mathrm{tr}}$ will not generally improve predictions for $\mathcal{D}_{\mathrm{test}}$, and hence the MDN will continue to output the prior pdf. Applying the system to observations may not then yield informative results, but they should at least not be misleading. Note that obtaining a 'null' result of this form does not necessarily imply that a given observation contains no useful information: a larger training set, or the use of a more targeted inversion technique, may allow more confident inferences to be drawn.

Furthermore, even if prior sampling requires a significantly larger number of samples than posterior approaches, its repeatability may continue to count in its favour. Assume that a number

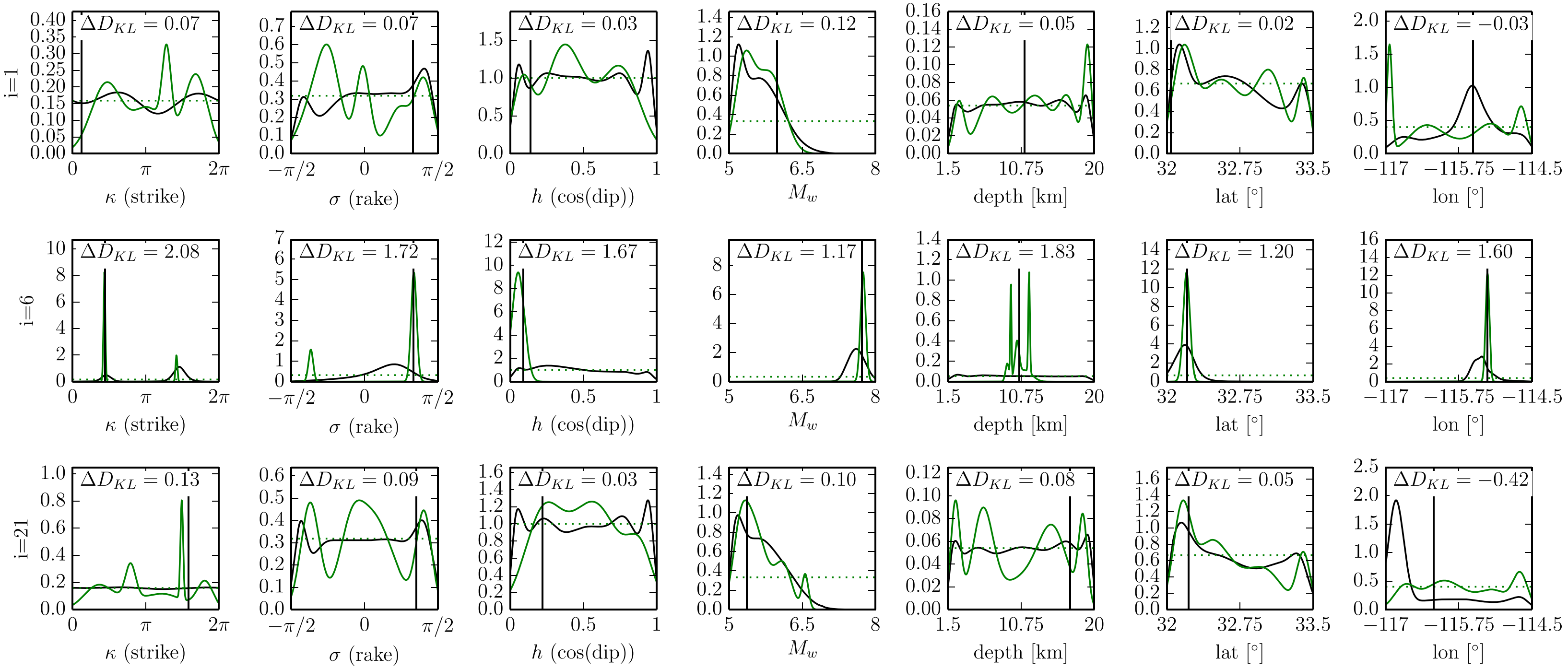


Figure 9. Posterior pdfs approximated by seven MDN ensembles (black curves) and estimated using posterior samples obtained by MCMC sampling (green curves) for the test set examples 1 (top), 6 (middle) and 21 (bottom). Example 21 (lon) shows the most negative ΔD_{KL} value of all test set examples and is also visible as an outlier in Fig. 7, third row.

of N_{prior} prior samples are required to achieve acceptable performance on an independent test set. Now, suppose that n inversions are to be performed independently (e.g. this could be the number of events that are expected to occur in a given magnitude range over the lifetime of an earthquake monitoring system). If we furthermore assume that the problem can be solved using a posterior sampling method, which requires N_{post} evaluations of the forward problem, then prior sampling becomes advantageous once $n > N_{\text{prior}}/N_{\text{post}}$.

5.2 Model selection and noise estimation

A challenging aspect of any inverse problem is the choice of an appropriate parametrization. In particular, it is hard to *a priori* know to what detail a given data set can constrain the model. Trans-dimensional approaches address this issue (Sambridge *et al.* 2006) by treating the number of model parameters as an unknown of the inversion that is determined along with the data. It is an interesting question how a trans-dimensional scheme could be realized using a technique based on prior sampling. In theory, a training set need not be limited to contain only samples obtained using one fixed parametrization and we could potentially treat, for example, the number of layers in an Earth model, or the number of slip patches of a finite-source inversion as a variable. This would however increase the dimensionality of the sampling space and further experimentation is required to assess the feasibility of such an approach.

A related problem is that of data noise estimation. According to eq. (24) generating the set $\mathcal{D}_{\text{prior}}$ requires a means to generate noise vectors ϵ_i that follow the combined observational and modeling noise distribution. Although we can often obtain a rather accurate estimate of the observational noise distribution, describing noise caused by systematic mismatch between observations and predictions due to simplifications of the underlying theory is very difficult if not impossible in many cases. It may therefore be desirable to treat both data and modeling noise estimates as unknowns and marginalize over different noise levels. Again, this could potentially be addressed by constructing the training set accordingly, but we have not yet investigated this point in any detail.

6 CONCLUSIONS

We have shown how posterior inferences can be made using a set of samples that were obtained before the observation to be inverted is available. The approach separates the sampling from the inversion stage, which has the advantage that no expensive calculations need to be carried out at the time of inversion. The fact that we base our inferences on a fixed set of prior samples introduces an additional contribution to the posterior uncertainty, which we in turn estimate by fitting a parametric pdf to the set of samples. Moreover, we have seen that the presented probabilistic algorithm based on MDNs—a flexible tool for conditional probability density estimation—can provide an unbiased and conservative estimate of the marginal posterior pdf $\sigma(m_k|\mathbf{d}_0)$ in the case of a realistic point-source parameter estimation problem. From this and a toy problem, however, it has also become clear that such an approach ultimately suffers from a number of limitations that may be hard or even impossible to overcome in practice. The fundamental difference between sampling from $\sigma(\mathbf{m}, \mathbf{d}_0)$, rather than $\sigma(\mathbf{m}|\mathbf{d}_0) = \frac{\sigma(\mathbf{m},\mathbf{d}_0)}{\sigma(\mathbf{d}_0)}$ lies in the omission of the normalization constant $\sigma(\mathbf{d}_0)$—the unconditional probability of making a particular observation $\mathbf{d}_0$ or 'marginal likelihood'. The probability of observing a particular datum $\mathbf{d}_0$ compared to other possible outcomes may be very small and therefore also the density of the set of samples drawn from the joint distribution $\sigma(\mathbf{m}, \mathbf{d}_0)$ will be low in the vicinity of $\mathbf{d}_0$. Even increasing the number of prior samples by orders of magnitude cannot necessarily mitigate this problem, as we have seen from the simple toy problem presented in Section 3. Methods based on posterior sampling naturally take this into account, since the sampling distribution is implicitly conditioned on the observation $\mathbf{d}_0$.

Nevertheless, the presented approach may enable us to find a useful approximate probabilistic answer for problems which could hitherto only be solved deterministically or not at all with given computational resources in a timely manner. The approach is therefore appealing in cases where an expensive inverse problem has to be solved either repeatedly with new observations under similar prior constraints and governing physics or subject to tight temporal constraints. Moreover, although not discussed at length in this paper, the approach enables us to reuse the same set of samples to answer multiple questions, including those that may not have been posed prior to sampling. In particular, it allows us to flexibly define new parameters or observables based on the existing variables without the need for resampling. Instead, we merely have to train a new set of neural network ensembles on the new derived input and target variables. As an example, consider coordinate changes, averages of multiple parameters or other linear and non-linear transformations (see de Wit *et al.* (2014) for an example). In contrast, with methods that obtain posterior samples in the vicinity of a given observation and with a given parametrization, such as MCMC, resampling is generally required if the definition of the misfit functional is changed, or a different set of (derived) parameters is to be determined.

Finally, pattern recognition approaches, in particular neural networks, have proven to be very robust with respect to noise and unmodeled signal in the observations (e.g. Böse *et al.* 2012; Käufl *et al.* 2014). This can be understood by recognizing that an approach based on prior sampling does not involve the minimization of a misfit functional, which can easily lead to overfitting the model to certain details in the particular observation if these have not been taken into account by the forward relation or the noise model. With the presented approach, on the other hand, the relative weight of the individual measurements is determined solely based on the training set, which—if designed carefully—does not contain any unwanted bias. Therefore inversions are less prone to outliers, that is, unmodeled noise and signal in the data—a property that may be particularly useful for real-time monitoring tasks, such as EEW, in which no manual data quality control can be performed.

ACKNOWLEDGEMENTS

We thank Suzanne Atkins for her comments on a draft of the manuscript and Jan Dettmer for his detailed review of the paper that greatly helped to improve the manuscript. Moreover, we thank an anonymous reviewer for his likewise insightful comments. PK, AV and RDW are funded by The Netherlands Organisation for Scientific Research (NWO) under the grant ALW Top-subsidy 854.10.002.

REFERENCES

Bilmes, J.A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, Int. Comput. Sci. Inst., Berkeley, Tech. Rep, TR-97-021.

Bishop, C.M., 1994. Mixture density networks, Tech. Rep., Aston Univ., Birmingham.

Bishop, C.M., 1995. *Neural Networks for Pattern Recognition,* vol. 92, Oxford University Press.

Bodin, T., Sambridge, M., Rawlinson, N. & Arroucau, P., 2012. Transdimensional tomography with unknown data noise, *Geophys. J. Int.,* **189**(3), 1536–1556.

Böse, M., Heaton, T.H. & Hauksson, E., 2012. Rapid estimation of earthquake source and ground-motion parameters for earthquake early warning using data from a single three-component broadband or strong-motion sensor, *Bull. seism. Soc. Am.,* **102**(2), 738–750.

Burridge, R. & Knopoff, L., 1964. Body force equivalents for seismic dislocations, *Bull. seism. Soc. Am.,* **54**(6), 1875–1888.

Carney, M., Cunningham, P., Dowling, J. & Lee, C., 2005. Predicting probability distributions for surf height using an ensemble of mixture density networks, in *Proceedings of the 22nd International Conference on Machine Learning (ICML'05),* pp. 113–120, ACM, New York.

Cornford, D., Nabney, I.T. & Bishop, C.M., 1999. Neural network-based wind vector retrieval from satellite scatterometer data, *Neural Comput. Appl.,* **8**(3), 206–217.

Curtis, A. & Lomax, A., 2001. Tutorial prior information, sampling distributions, and the curse of dimensionality, *Geophysics,* **66**(2), 372–378.

Dettmer, J. & Dosso, S.E., 2012. Trans-dimensional matched-field inversion with hierarchical error models and interacting Markov chains, *J. acoust. Soc. Am.,* **132,** 2239–2250.

Dettmer, J., Dosso, S.E. & Holland, C.W., 2011. Sequential trans-dimensional Monte Carlo for range-dependent geoacoustic inversion, *J. acoust. Soc. Am.,* **129**(4), 1794–1806.

de Wit, R.W., Valentine, A.P. & Trampert, J., 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks, *Geophys. J. Int.,* **195**(1), 408–422.

de Wit, R.W., Käufl, P., Valentine, A.P. & Trampert, J., 2014. Bayesian inversion of free oscillations for Earth's radial (an)elastic structure, *Phys. Earth planet. Inter.,* **237,** 1–17.

Dosso, S.E. & Dettmer, J., 2011. Bayesian matched-field geoacoustic inversion, *Inverse Probl.,* **27**(5), 055009, doi:10.1088/0266-5611/27/5/055009.

Dziewonski, A.M., Chou, T.-A. & Woodhouse, J.H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.,* **86**(B4), 2825–2852.

Ekström, G., Nettles, M. & Dziewonski, A.M., 2012. The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200–201,** 1–9.

Geyer, C., 1991. Markov chain Monte Carlo maximum likelihood, in *Proceedings of the 23rd Symposium on the Interface,* pp. 156–163, Interface Foundation of North America, Fairfax Station, VA.

Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* **57**(1), 97–109.

Herzallah, R. & Lowe, D., 2004. A mixture density network approach to modelling and exploiting uncertainty in nonlinear control problems, *Eng. Appl. Artif. Intell.,* **17,** 145–158.

Hjorth, L. & Nabney, I., 2000. Bayesian training of mixture density networks, in *Proceedings of the International Joint Conference on Neural Networks,* vol. 4, pp. 455–460, IEEE.

Hornik, K., Stinchcombe, M. & White, H., 1989. Multilayer feedforward networks are universal approximators, *Neural Netw.*, **2**(5), 359–366.

Kanamori, H., 2005. Real-time seismology and earthquake damage mitigation, *Annu. Rev. Earth Planet. Sci.,* **33**(1), 195–214.

Käufl, P., Valentine, A.P., O'Toole, T.B. & Trampert, J., 2014. A framework for fast probabilistic centroid-moment-tensor determination–inversion of regional static displacement measurements, *Geophys. J. Int.,* **196**(3), 1676–1693.

Käufl, P., Valentine, A.P., de Wit, R.W. & Trampert, J., 2015. Robust and fast probabilistic source parameter estimation from near-field displacement waveforms using pattern recognition, *Bull. seism. Soc. Am.,* **105**(4), 2299–2312.

Lee, S.-J., Huang, B.-S., Liang, W.-T. & Chen, K.-C., 2010. Grid-based moment tensor inversion technique by using 3-D Green's functions database: a demonstration of the 23 October 2004 Taipei earthquake, *Terr. Atmos. Sci.,* **21**(3), 503–514.

Lee, S.-J. *et al.*, 2013. Towards real-time regional earthquake simulation I: Real-time moment tensor monitoring (RMT) for regional events in Taiwan, *Geophys. J. Int.,* **196**(1), 432–446.

Liu, Q., Polet, J., Komatitsch, D. & Tromp, J., 2004. Spectral-element moment tensor inversions for earthquakes in southern California, *Bull. seism. Soc. Am.,* **94**(5), 1748–1761.

MacKay, D.J.C., 2003. *Information Theory, Inference and Learning Algorithms,* Cambridge Univ. Press.

MacKay, D.J.C., 1996. Hyperparameters: optimize, or integrate out?, in *Maximum Entropy and Bayesian Methods,* vol. 62 of Fundamental Theories of Physics, pp. 43–59, ed. Heidbreder, G.R., Springer, The Netherlands.

McLachlan, G.J. & Basford, K.E., 1988. Mixture models. Inference and applications to clustering, in *Statistics: Textbooks and Monographs,* vol.1, Dekker, New York, 1988.

Meier, U., Curtis, A. & Trampert, J., 2007a. Fully nonlinear inversion of fundamental mode surface waves for a global crustal model, *Geophys. Res. Lett.,* **34**(16), 1–6.

Meier, U., Curtis, A. & Trampert, J., 2007b. Global crustal thickness from neural network inversion of surface wave data, *Geophys. J. Int.,* **169**(2), 706–722.

Nocedal, J., 1980. Updating Quasi-Newton matrices with limited storage, *Math.Comput.,* **35**(151), 773.

O'Toole, T.B. & Woodhouse, J.H., 2011. Numerically stable computation of complete synthetic seismograms including the static displacement in plane layered media, *Geophys. J. Int.,* **187**(3), 1516–1536.

Ribés, A. & Schmitt, F., 2003. A fully automatic method for the reconstruction of spectral reflectance curves by using mixture density networks, *Pattern Recog. Lett.,* **24,** 1691–1701.

Richmond, K., 2007. Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion, in *Advances in Nonlinear Speech Processing,* pp. 263–272, Springer.

Rumelhart, D., Hinton, G. & Williams, R., 1986. Learning representations by back-propagating errors, *Nature,* **323,** 533–536.

Sambridge, M., 1999. Geophysical inversion with a neighbourhood algorithm – II. Appraising the ensemble, *Geophys. J. Int.,* **138**(3), 727–746.

Sambridge, M. & Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems, *Rev. Geophys.,* **40**(3), 1–29.

Sambridge, M., Gallagher, K., Jackson, A. & Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence, *Geophys. J. Int.,* **167**(2), 528–542.

Schittenkopf, C. & Dorffner, G., 2001. Risk-neutral density extraction from option prices: improved pricing with mixture density networks, *IEEE Trans. Neural Netw.,* **12,** 716–725.

Shahraeeni, M.S. & Curtis, A., 2011. Fast probabilistic nonlinear petrophysical inversion, *Geophysics,* **76**(2), E45–E58.

Shahraeeni, M.S., Curtis, A. & Chao, G., 2012. Fast probabilistic petrophysical mapping of reservoirs from 3d seismic data, *Geophysics,* **77**(3), O1–O19.

Skilling, J., 2006. Nested sampling for general Bayesian computation, *Bayesian Analy.,* **1**(4), 833–860.

Stähler, S.C. & Sigloch, K., 2014. Fully probabilistic seismic source inversion—Part 1: Efficient parameterisation, *Solid Earth,* **5**(2), 1055–1069.

Tarantola, A., 2005. *Inverse Problem Theory,* vol. 4, SIAM.

APPENDIX A: MARKOV CHAIN MONTE CARLO SAMPLING OF THE POSTERIOR

Algorithm 1 Metropolis–Hastings algorithm used to generate samples from the posterior distribution (eq. 5).

```
Draw $\mathbf{m}^{(1)}$ from the uniform prior distribution $p(\mathbf{m})$
for $i = 1$ to $N$ do
    $\mathbf{m}' \leftarrow \mathbf{m}^{(i)}$
    for $k = 1$ to $M$ do
        Update $m'_k$ with a value drawn from $q_k\left(m_k | m_k^{(i)}\right)$
        $a \leftarrow p^*(\mathbf{m}')/p^*(\mathbf{m}^{(i)})$
        if $u \sim \mathcal{U}(0, 1) \leq a$ then
            $\mathbf{m}^{(i)} \leftarrow \mathbf{m}'$
        end if
    end for
    $\mathbf{m}^{(i+1)} \leftarrow \mathbf{m}^{(i)}$
    potentially tune parameters of $q_k\left(m_k | m_k^{(i)}\right)$ (see the main text).
end for
```

The overdetermined, but non-linear and possibly non-unique inverse problem of inferring the centroid location and moment tensor of a source from observations in a layered medium cannot be solved analytically. In order to obtain a reference solution for comparison with the MDN approximations, we generate samples from the posterior distribution (eq. 3) by MCMC sampling. We assume that observational uncertainties are described by an additive Gaussian noise process with covariance $\mathbf{C}_d$ and consider the forward operator to be exact, that is the posterior is given by eq. (5) with the likelihood function

$$L(\mathbf{m}|\mathbf{d}_0) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{C}_d|}} \times \exp\left[-\frac{1}{2}\left(\mathbf{d} - g(\mathbf{m})\right)^T \mathbf{C}_d^{-1}\left(\mathbf{d} - g(\mathbf{m})\right)\right], \qquad \text{(A1)}$$

where D is the dimensionality of the data space, which in the case of the source inversion problem in this paper is given by $D = 102$. The covariance matrix $\mathbf{C}_d$ is in this case chosen to be diagonal with covariances resembling typical noise levels for static displacement measurements ($(0.004\,\mathrm{m})^2$ for the horizontal, and $(0.01\,\mathrm{m})^2$ for the vertical components; *cf.* Käufl *et al.* 2014). Moreover, throughout this paper, we work with uniform prior distributions and have

$$\rho(\mathbf{m}) = \text{const.} \qquad \text{(A2)}$$

Therefore, the term $\rho(\mathbf{m})/\sigma(\mathbf{d}_0)$ in eq. (5) is independent of $\mathbf{m}$ and obtaining samples from the posterior distribution is thus equivalent to obtaining samples from (see e.g. MacKay 2003)

$$p^*(\mathbf{m}) \propto L(\mathbf{m}|\mathbf{d}_0). \qquad \text{(A3)}$$

In the case of the source inversion problem, we use a Metropolis–Hastings algorithm with normal proposal distributions $q_k(m_k|m_k^{(i)})$ for each component of $\mathbf{m}$. That is, each component of $\mathbf{m}$ is updated sequentially and updates are uncorrelated between parameters. Initial test runs indicated that chains are converging more quickly to regions of high likelihood with this setup, instead of block-updating all components of $\mathbf{m}$ at once at every iteration.[1] The proposal distributions $q_k(m_k|m_k^{(i)})$ are Gaussian with mean $m_k^{(i)}$ and standard deviation σ_k, which is initially set to 0.1 for all parameters. Note that all parameters are rescaled to the range $[-1; 1]$. Every 500th iteration we update σ_k based on the acceptance rate of the chain so far. Updates become less frequent and of decreasing magnitude as the chains converge.

Independent posterior samples are obtained from each chain by removing the initial 5000 samples (‘burn-in period’), in which most of the chains appear to be non-stationary. Furthermore, the samples are decorrelated by taking into account only every 500th sample (‘thinning’). These measures are common practice and the particular values are chosen based on visual inspection of a number of test chains. See Fig. A1 for an example. We denote by X_h the set of samples $\{(\mathbf{m}, \mathbf{d})_i\}$ in the hth chain. We run $H = 5$ independent chains starting from different random starting values for of $\mathbf{m}$ for $N = 25000$ iterations each. We preferred multiple shorter chains over one long chain, since we found that the posterior often exhibits strong multimodal behaviour where a single chain often becomes trapped in one of the modes for a very long time. The fact that multiple independent chains do not converge to the same solution is typically taken as an indication for the non-convergence of all chains and that more iterations are required or the mixing of the chains has to be improved. However, since we are dealing with purely synthetic examples with known target values and noise properties, we can easily identify chains that give misleading results. Synthetic examples in which the sampler did not produce models with a sufficiently high likelihood are later excluded from the interpretation. The algorithm is summarized in listing 1.

In order to incorporate the information contained in all Markov chains into the analysis and since we are only interested in marginal posterior distributions, we generate—independently, for each dimension of $\mathbb{M}$—a set of scalar posterior samples from all available chains by resampling according to a piecewise constant approximation of the marginal likelihood

$$p^*(m_k) = \int p(\mathbf{d}|\mathbf{m})\, \mathrm{d}m_{i \neq k}, \qquad \text{(A4)}$$

as follows. First, we collect the samples of all thinned chains in the joint set

$$X = \bigcup_{h=1}^{H} X_h. \qquad \text{(A5)}$$

Second, we calculate histograms of the samples and assign to the lth bin in the kth dimension of $\mathbf{m}$ the weight

$$w_l^{(k)} = \sum_{\mathbf{m} \in M_l} p^*(\mathbf{m})/|M_l|, \qquad \text{(A6)}$$

where $M_l = \{\mathbf{m}|m_k \in [a_l, b_l[\}$, a_l and b_l are the bin boundaries and $|M_l|$ is the number of samples falling into the lth bin.

Subsequently, we generate a set of random numbers

$$Y_l^{(k)} = \{y_i \sim \mathcal{U}(a_l, b_l)\}, \qquad \text{(A7)}$$

in such a way that $|Y_l^{(k)}| \propto w_l$. The joint set of samples $Y^{(k)} = \bigcup_l Y_l$ is now distributed according to a piecewise constant approximation to the marginal likelihood $p^*(m_k)$.

Finally, for comparison with the MDN results, we fit one-dimensional GMMs with six mixture components to the set of samples $Y^{(k)}$ for each dimension k individually using the expectation–maximization (EM) algorithm (see Bishop 1995; Bilmes 1998). An EM fit for the seven source parameters of a synthetic test set example is shown in Fig. A2.

[1] Note that for the toy problem in Section 3 all parameters are updated at once, instead.

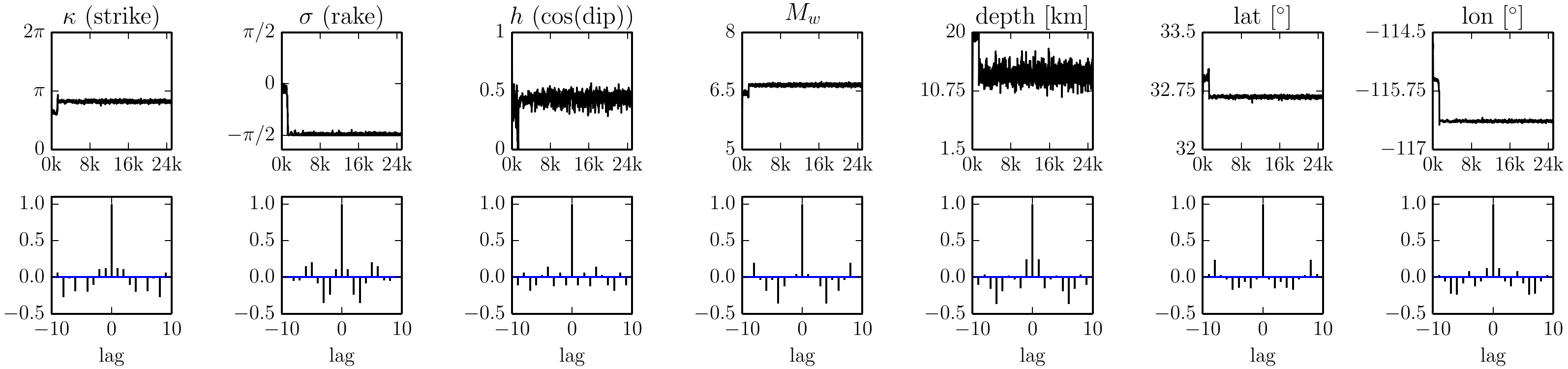


Figure A1. A Markov Chain obtained by Algorithm 1 for the static source inversion problem. The first row shows the raw traces, the second row the autocorrelation function after removal of the burn-in period and thinning with a factor of 500.

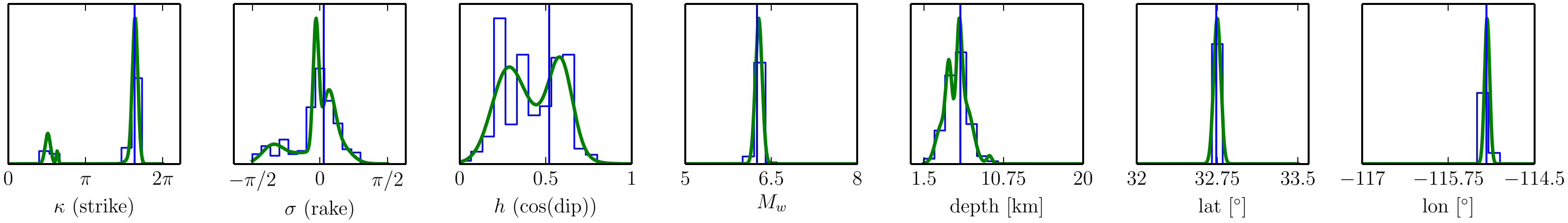


Figure A2. EM fit of 1-D marginal GMMs (green curve) to a set of samples obtained by MCMC after resampling according to eqs (A5) to (A7) (blue histograms) for a synthetic test set example. The vertical lines denote the target value positions for each parameter.

APPENDIX B: DERIVATION OF INTEGRALS

The posterior probability density

$$\sigma(\mathbf{m}|d_0 = 0) = \frac{L(\mathbf{m}|d_0 = 0)\rho(\mathbf{m})}{\sigma(d_0 = 0)} \tag{B1}$$

for the toy problem in Section 3 can be derived as follows. With the prior distribution

$$\rho(\mathbf{m}) = \begin{cases} \left(\frac{1}{2}\right)^c & \text{if } -1 \leq c \leq 1 \\ 0 & \text{else} \end{cases}, \tag{B2}$$

and the likelihood

$$L(\mathbf{m}|d_0) = \frac{\exp\left[-(d_0 - g(\mathbf{m}))^2/(2\sigma_d^2)\right]}{\sqrt{2\pi}\sigma_d}, \tag{B3}$$

where $g(\mathbf{m}) = ||\mathbf{m}||_2$, we have

$$\sigma(d_0 = 0) = \int_{-\infty}^{\infty} L(\mathbf{m}|d_0 = 0)p(\mathbf{m})\mathrm{d}m_1 \ldots \mathrm{d}m_c \tag{B4}$$

$$= \left(\frac{1}{2}\right)^c \frac{1}{\sqrt{2\pi}\sigma_d} \int_{-1}^{1} \exp\left[-\frac{1}{2\sigma_d^2}\sum_k m_k^2\right] \tag{B5}$$

$$\approx \left(\frac{1}{2}\right)^c \frac{1}{\sqrt{2\pi}\sigma_d} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma_d^2}\sum_k m_k^2\right] \tag{B6}$$

$$= \left(\frac{1}{2}\right)^c \left(2\pi\sigma_d^2\right)^{\frac{c-1}{2}} \tag{B7}$$

and thus

$$\sigma(\mathbf{m}|d_0 = 0) = \frac{1}{(2\pi\sigma_d^2)^{c/2}} \exp\left[-\frac{\sum_{k=1}^{c} m_k^2}{2\sigma_d^2}\right]. \tag{B8}$$