Varying prior information in Bayesian inversion

2014 Inverse Problems 30 065002

(http://iopscience.iop.org/0266-5611/30/6/065002)

IOP Publishing Inverse Problems

Inverse Problems **30** (2014) 065002 (20pp) doi:10.1088/0266-5611/30/6/065002

Varying prior information in Bayesian inversion

Matthew Walker and Andrew Curtis

School of GeoSciences, Grant Institute, University of Edinburgh, King's Buildings, Edinburgh, UK

E-mail: matt.walker@ed.ac.uk

Received 15 September 2013, revised 18 March 2014
Accepted for publication 18 March 2014
Published 19 May 2014

Abstract

Bayes' rule is used to combine likelihood and prior probability distributions. The former represents knowledge derived from new data, the latter represents pre-existing knowledge; the Bayesian combination is the so-called posterior distribution, representing the resultant new state of knowledge. While varying the likelihood due to differing data observations is common, there are also situations where the prior distribution must be changed or replaced repeatedly. For example, in mixture density neural network (MDN) inversion, using current methods the neural network employed for inversion needs to be retrained every time prior information changes. We develop a method of *prior replacement* to vary the prior without re-training the network. Thus the efficiency of MDN inversions can be increased, typically by orders of magnitude when applied to geophysical problems. We demonstrate this for the inversion of seismic attributes in a synthetic subsurface geological reservoir model. We also present results which suggest that *prior replacement* can be used to control the statistical properties (such as variance) of the final estimate of the posterior in more general (e.g., Monte Carlo based) inverse problem solutions.

Keywords: neural network inversion, Bayesian inversion, prior replacement, mixture density network

S Online supplementary data available from stacks.iop.org/IP/30/065002/mmedia

(Some figures may appear in colour only in the online journal)

1. Introduction

Bayesian statistics is based on the idea that propositions can be assigned some uncertainty which can be represented by a probability. Thus the Bayesian interpretation of probability is as a measure of the state of knowledge about a proposition (Jaynes 1986). In Bayesian inversion the previous (or so-called prior) probability can be updated in the light of new data which

provides information about the proposition (Gelman *et al* 2003, p 6), and such information is combined using Bayes' rule (Ulrych *et al* 2001). In this paper we subvert the above usual order of application of Bayes' rule in Bayesian inversion: instead we take a probability already created using Bayes' rule, and remove the prior probability, replacing it with a different prior probability. We call this process *prior replacement*.

Consider estimating a model parameter vector, $\mathbf{m}$, which by some deterministic relationship (a forward model) is related to a data vector, $\mathbf{d}$. Then the solution to the Bayesian problem of inferring $\mathbf{m}$ given measurements $\mathbf{d}$ is obtained from Bayes' rule (Duijndam 1988, Scales and Tenorio 2001, Ulrych *et al* 2001) as

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}. \tag{1}$$

All quantities in equation (1) are probability mass functions (PMFs) if $\mathbf{m}$ is discrete, or probability density functions (PDFs) if $\mathbf{m}$ is continuous. Equation (1) gives the so-called posterior distribution—the probability of the model parameters $\mathbf{m}$ given the data $\mathbf{d}$, by combining information from the new data represented by the likelihood $p(\mathbf{d}|\mathbf{m})$, and the prior information $p(\mathbf{m})$ (information that was already known about the parameters prior to the new data acquisition). The final term on the right hand side of equation (1) is $p(\mathbf{d})$ which is known as the evidence. For our purposes we regard this as a normalizing constant since it does not depend on parameters $\mathbf{m}$ (Sambridge *et al* 2006). It ensures that the right hand side of equation (1) is a valid PDF or PMF. Similarly, $p(\mathbf{d}|\mathbf{m})$ is often interpreted as an unnormalized probability distribution over $\mathbf{m}$, since $\mathbf{d}$ is assumed to be measured and therefore known.

Suppose that we have obtained a posterior distribution for the inversion of some data $\mathbf{d}$ with some prior information. If for some reason we wanted to change the prior used in this inversion then we can make a simple calculation: roughly speaking, we divide the posterior distribution in equation (1) by the existing prior, $p(\mathbf{m})$ and multiply by the new prior distribution. Thus we *replace* the prior in equation (1) with the new prior. We have only found two explicit treatments of this operation in the literature, both in reference to statistical classification models—that is, probabilistic classification of objects into discrete classes based on associated data (Michie *et al* 1994). Bishop (1995, p 223) uses prior replacement to modify the outputs of a Bayesian classification neural network (NN), and Bailer-Jones and Smith (2010) use the term 'prior replacement' to describe the operation for discrete classification problems. However, neither work discusses how it may be applied to continuous model parameters, nor any potential uses for the operation in a wider context.

In this paper we develop prior replacement as a general operation which can be applied to discrete or continuous valued model parameter posterior distributions. Furthermore, we show that it has a specific practical application in mixture density neural network (MDN) inversion. A NN can be viewed as a flexible model, mapping a set of inputs to a set of outputs (Roth and Tarantola 1994). In MDN inversion, values for a NN's parameters can be found at relatively high computational expense, by a process referred to as *training* (Rumelhart *et al* 1986, Johansson *et al* 1991, Bishop 1995); this causes the MDN to emulate the mapping from a data vector to the corresponding posterior probability distribution over the model parameters. The NN can then determine the posterior corresponding to any data vector extremely rapidly and efficiently. However, the resulting posterior is only valid for the particular pre-specified prior distribution that was used during training. When using such methods, if we wish to change the prior information we would have to re-train our NN. We refer to this methodology as the *prior-specific training* method, since the MDN is trained for a specific prior distribution.

Because of its efficiency and ability to emulate arbitrarily complex mappings, NN inversion is used extensively in many areas of geophysics to estimate subsurface parameters, by solving between thousands and billions of individual inverse problems. However, early work

focussed on using NNs which solved deterministic rather than Bayesian inverse problems (Van der Baan and Jutten 2000). For example, it was used to invert well-data (Liu and Liu 1998), full waveform seismic data (Roth and Tarantola 1994) and resistivity data (El-Qady and Ushijima 2001). The first to solve Bayesian inverse problems using NNs in the geosciences were Devilee *et al* (1999) who used a NN to predict the parameters of a histogram describing the posterior probability of the Earth's crustal thicknesses across Eurasia, given surface wave velocity data. This was followed by the work of Meier *et al* (2007a, 2007b, 2009) who successfully applied MDN inversion (developed originally in the machine learning community, see e.g., Bishop (1994)) to predict a parametrized version of the posterior PDF describing a global crustal seismological model (Meier *et al* 2007a, 2007b) and global variations of mantle seismic velocities, temperatures and water content (Meier *et al* 2009), given surface wave velocities and prior information on rock physical relations. At the same time a related methodology, the so-called Bayesian NN, was developed for use in inverting German Continental Deep Drilling Program borehole data (Maiti *et al* 2007, Maiti and Tiwari 2010). More recently, MDN inversion has been developed for exploration geophysical applications by Shahraeeni and Curtis (2011), who increased the resolving power of the outputs of the NN by increasing the flexibility of the kernels which they describe. This improved inversion methodology was used to efficiently invert seismic attribute data for subsurface reservoir parameters (Shahraeeni and Curtis 2011, Shahraeeni *et al* 2012).

The latter application motivates the current work. In that application a single NN is used repeatedly to invert many individual data distributed spatially across a 3D grid of cells spanning a subsurface hydrocarbon reservoir. Each grid cell is populated with seismic impedance estimates which are treated as data. The aim is to invert the impedance estimates for rock physical and fluid parameters in each cell, and there may be up to billions of cells in real problems. If prior information changes spatially (and it usually does), we would normally have to use prior-specific training to obtain correct posterior estimates at each position, thus forfeiting some (if not all) of the efficiency and speed gained by the use of NN or MDN inversion methods. However, if we make use of prior replacement we may increase efficiency. The purpose of this paper is to investigate this efficiency gain, and the quality of the solution obtained. Although our primary use of prior replacement is in conjunction with NN inversion, we will also discuss its possible use in more general Bayesian inversion settings. The latter may result in a wider class of applications in future.

We first describe prior replacement in detail, and its application to MDN inversion. We then give a numerical example of its application both to a single inversion in isolation and a reservoir-scale inversion. Finally we discuss the implications of our results with respect to both NN inversion and Bayesian inversion in general. We also discuss the effect of prior replacement on the quality of the final posterior estimate obtained. The discussion of quality is supported by auxiliary results presented in the appendices (supplied as supplementary online material, available from stacks.iop.org/IP/30/065002/mmedia) for a simple example Bayesian inverse problem. These results also suggest that prior replacement may be used as a variance reduction technique similar to importance sampling (indeed, prior replacement seems to outperform importance sampling in this respect for the simple problem presented therein).

2. Prior replacement

2.1. Probabilistic development

We now write out the Bayesian solution to an inverse problem in two different situations. Both situations involve an inverse problem with the same forward function, thus the likelihood distribution is identical in both. However, in the first, so-called 'old' situation there is a

different prior probability distribution to that of the second 'new' situation. We denote these with 'old' and 'new' subscripts. It follows from Bayes' theorem that the posterior must also vary. Accordingly the evidence may also vary, which can be seen if we write it in the integral form in the denominator of Bayes' theorem for the two situations:

$$p_{\text{old}}(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p_{\text{old}}(\mathbf{m})}{p_{\text{old}}(\mathbf{d})} = \frac{p(\mathbf{d}|\mathbf{m})p_{\text{old}}(\mathbf{m})}{\int_{\mathbf{m}} p(\mathbf{d}|\mathbf{m})p_{\text{old}}(\mathbf{m})\,\mathrm{d}\mathbf{m}} \tag{2}$$

and

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p_{\text{new}}(\mathbf{m})}{p_{\text{new}}(\mathbf{d})} = \frac{p(\mathbf{d}|\mathbf{m})p_{\text{new}}(\mathbf{m})}{\int_{\mathbf{m}} p(\mathbf{d}|\mathbf{m})p_{\text{new}}(\mathbf{m})\,\mathrm{d}\mathbf{m}} \tag{3}$$

where the integral forms are simply evaluations of the relevant normalizing constants in each case. We can therefore see that $p_{\text{new}}(\mathbf{m}|\mathbf{d})$ can be written in terms of $p_{\text{old}}(\mathbf{m}|\mathbf{d})$ (and vice versa) by

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = p_{\text{old}}(\mathbf{m}|\mathbf{d})\frac{p_{\text{new}}(\mathbf{m})}{p_{\text{old}}(\mathbf{m})}\frac{p_{\text{old}}(\mathbf{d})}{p_{\text{new}}(\mathbf{d})}. \tag{4}$$

In the context of inversion, we are usually supplied with a fixed data vector $\mathbf{d}$. Hence, in both new and old situations we assume that the data observed is the same. The evidence is dependent upon the form of the prior so may vary between the two situations. Nevertheless it is still independent of the value of the model vector. Therefore, for later convenience we set $p_{\text{new}}(\mathbf{d})/p_{\text{old}}(\mathbf{d}) = k$, such that

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k}\frac{p_{\text{new}}(\mathbf{m})}{p_{\text{old}}(\mathbf{m})}p_{\text{old}}(\mathbf{m}|\mathbf{d}). \tag{5}$$

Equation (5) now has a form which allows us to evaluate the new posterior distribution from the old one, assuming that we know both the old and the new prior, $p_{\text{old}}(\mathbf{m})$ and $p_{\text{new}}(\mathbf{m})$ respectively, and that we can evaluate the scale factor k. The latter can be shown to be a normalizing constant: since from the definition of PMFs and PDFs we have that $\int_{-\infty}^{+\infty} p_{\text{new}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m} = 1$, so integrating over both sides of equation (5) yields

$$k = \int_{-\infty}^{+\infty} \frac{p_{\text{new}}(\mathbf{m})}{p_{\text{old}}(\mathbf{m})}p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m}. \tag{6}$$

Equation (5) shows the main operation involved in prior replacement. This will yield a valid result only under certain conditions. One can interpret equation (5) as trying to correct for a prior that is incorrect. The old posterior is divided by the old prior in an attempt to remove its effects. If the old prior had regions of zero probability then this will result in undefined values (0/0) where the old prior and posterior are simultaneously zero in the model space. We can interpret this as follows: when the old prior was initially applied and the old posterior obtained, we lost all information about the likelihood in those regions, and we cannot regain such information by changing the prior. Thus we are forced to assume that these undefined regions still have zero probability if we wish to continue. We implement this through our new prior: it is a condition that this must have zero probability where the old prior had zero probability, hence the new posterior will have zero probability in such areas too. We refer to this as the *support condition* below.

2.2. Mixture density neural network inversion

Any posterior PDF like that in equation (1) can be approximated by the sum of K normalized multivariate Gaussians each weighted by a constant (Bishop 1994, 1995, McLachlan and Peel 2004)

$$p(\mathbf{m}|\mathbf{d}) = \sum_{i=1}^{K} \alpha_i \phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{7}$$

where $\{\alpha_i | i \in 1, 2, \ldots, K\}$ are normalizing weights which obey $\sum_{i=1}^{K} \alpha_i = 1$, and $\phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a normalized multivariate Gaussian function of $\mathbf{m}$ with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ (where normalized implies that $\int_{-\infty}^{+\infty} \phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\, \mathrm{d}\mathbf{m} = 1$). This approximation of a PDF by a series of weighted, normalized Gaussians is referred to henceforth as a Gaussian mixture model (GMM).

In NN inversion using a mixture density network (MDN), a NN is determined that can predict values of α_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ in the mixture model which approximate the correct posterior (the left hand side of equation (7)) for any given value of $\mathbf{d}$. The parameters of a NN with such properties are estimated by a training process using samples from the distribution $p(\mathbf{m}, \mathbf{d}) = p(\mathbf{m})p(\mathbf{d}|\mathbf{m})$. Such samples are obtained by first sampling from the model space using the prior distribution $p(\mathbf{m})$, then obtaining the corresponding samples of $\mathbf{d}$ from the probabilistic forward model $p(\mathbf{d}|\mathbf{m})$ which is assumed to be known. The process of training is usually treated as a non-linear optimization for the parameters of the NN which maximize the likelihood of the training samples. For a full description of the training process of a MDN, see Bishop (1995, pp 140–161) for isotropic Gaussian kernels (ϕ), or Shahraeeni and Curtis (2011) who extended the method to anisotropic Gaussian kernels.

2.3. Prior replacement in neural network inversion

We can directly apply the prior replacement equations (1)–(6) to the results of the MDN inversion. If we equate the old posterior that appears in these equations to the mixture model output of the MDN then

$$p_{\mathrm{old}}(\mathbf{m}|\mathbf{d}) = \sum_{i=1}^{K} \alpha_i \phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{8}$$

for some set of weights α_i. Substitution of equation (8) into equations (5) and (6) permits us to write

$$p_{\mathrm{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k} \frac{p_{\mathrm{new}}(\mathbf{m})}{p_{\mathrm{old}}(\mathbf{m})} \sum_{i=1}^{K} \alpha_i \phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{9}$$

and

$$k = \int_{-\infty}^{+\infty} \frac{p_{\mathrm{new}}(\mathbf{m})}{p_{\mathrm{old}}(\mathbf{m})} \sum_{i=1}^{K} \alpha_i \phi(\mathbf{m}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\, \mathrm{d}\mathbf{m}. \tag{10}$$

Thus, equations (9) and (10) provide a method of performing prior replacement for the output of a MDN (i.e., for a GMM). As with the general equations for prior replacement (equations (5) and (6)), these equations only have well defined results for $p_{\mathrm{old}}(\mathbf{m})$ and $p_{\mathrm{new}}(\mathbf{m})$ distributions that satisfy the support condition as described earlier. However, an added complication arises because in equations (9) and (10) we use a GMM approximation to the posterior, $p_{\mathrm{old}}(\mathbf{m}|\mathbf{d})$. This GMM approximation is non-zero everywhere (except in the impractical case of Gaussian kernels with zero variance); the real $p_{\mathrm{old}}(\mathbf{m}|\mathbf{d})$ may not be non-zero everywhere, hence the non-zero nature of the GMM is an artefact of the approximation. Therefore $p_{\mathrm{new}}(\mathbf{m}|\mathbf{d})$ should still be zero wherever $p_{\mathrm{old}}(\mathbf{m})$ is zero (from equation (5)). Since we know that the GMM approximation is in error in this case, we should therefore still apply a new prior $p_{\mathrm{new}}(\mathbf{m})$ which has zero probability where the old prior has zero probability. In other words, the support condition still holds in this instance.

In appendix A, the prior replacement operations for MDN's are developed in more detail for certain analytical forms of the priors (old and new). We show that if the new prior is Gaussian or Uniform, and the old prior is Uniform, that equations (9) and (10) can be

written as truncated GMMs (we will later make use of these derivations). However due to this truncation they cannot be integrated analytically (Drezner 1992), so numerical integration techniques must be used to determine the normalizing constant. By contrast in appendix A we also show that, if both old and new prior distributions are Gaussian, equations (9) and (10) are themselves GMMs, and as such analytical integration can be used to solve them. We will not use these derivations in our examples below, but we include them since they potentially permit the prior replacement operation to be performed extremely rapidly. They are also of interest mathematically since they involve the division of Gaussians: this operation is non-trivial compared to the multiplication of Gaussians, and is only possible under certain conditions on the old and new priors. Whilst Gaussian multiplication is widespread in the literature (Tarantola 2002, Buland and Omre 2003, Petersen and Pedersen 2006), we have found little reference to such a 'Gaussian division' operation elsewhere.

3. Testing prior replacement in a MDN inversion

3.1. Inversion of the shaly-sand model

We compared the accuracy and computational efficiency of prior-specific training to prior replacement for a synthetic rock-physics inverse problem, for the case of a (i) Uniform, and (ii) Gaussian new prior. To do this we used a variant of a well-known rock physics model, the Yin–Marion shaly-sand model (Marion 1990, Yin *et al* 1993), which has been used previously as the forward model in MDN inversion (Shahraeeni 2011, p 16). This model predicts seismic wave impedances of S-waves (d_1) and P-waves (d_2), given the clay content by volume (m_1) and the sandstone matrix porosity (m_2) of a rock comprising a mixture of sandstone and shale. The impedances d_1 and d_2 at each point in the subsurface can be estimated from geophysical surveys; thus we construct an inverse problem for m_1 and m_2 to be solved at each such point.

An exact specification of the forward model is provided in appendix B. It is a deterministic model but we included a random element by adding random Gaussian noise to its output. Thus the probabilistic forward relation can be written

$$p(\mathbf{d}|\mathbf{m}) = \frac{1}{\sqrt{(2\pi)^2|\mathbf{\Sigma_d}|}} \exp(-(\mathbf{d}-\mathbf{f}(\mathbf{m}))^T \mathbf{\Sigma_d}^{-1}(\mathbf{d}-\mathbf{f}(\mathbf{m}))) \tag{11}$$

where $\mathbf{f}(\mathbf{m})$ represents the Yin–Marion shaly-sand model, $\mathbf{m} = [m_1, m_2]$ is the vector of model parameters, and $\mathbf{d} = [d_1, d_2]$ is the data vector of impedances at any point. $\mathbf{\Sigma_d}$ is a diagonal covariance matrix describing the (uncorrelated) random noise applied to the data, and represents our degree of uncertainty in the model's prediction of the data.

As explained above, the probabilistic forward function was used in conjunction with a prior to generate samples from $p(\mathbf{d}, \mathbf{m})$ which allowed us to train MDNs. In prior-specific training, samples are made directly from the new prior. For prior replacement, sampling is initially made from a Uniform old prior $p_{\text{old}}(\mathbf{m})$ which was chosen to be as broad as possible in the context of the model space, i.e.,

$$p_{\text{old}}(\mathbf{m}) = p_{\text{old}}(m_1, m_2) = \begin{cases} 0 & \text{for } m_i \notin [0,1],\ i = 1,2 \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

This old prior was then replaced by the new prior in each case. Note that all possible $p_{\text{new}}(\mathbf{m})$ PDFs are contained within the bounds of the Uniform distribution in equation (12), as is required by the support condition described above.

This test uses an entirely synthetic inversion: the data inverted by the MDN was also generated using the probabilistic forward function. The same data point, $\mathbf{d}$, was used in both cases. It was chosen arbitrarily, since we simply use it to demonstrate the method. In each

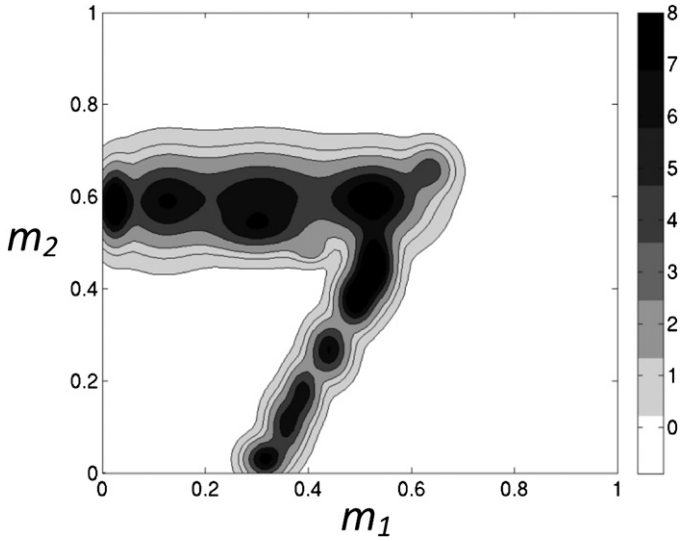

Figure 1. (a) The old posterior obtained from the output of a neural network (MDN) trained with samples made from the broad old prior defined in equation (12). Prior replacement was applied to this PDF to emplace a Uniform and a Gaussian new prior in figures 2 and 3, respectively.

of cases (i) and (ii) the appropriate prior replacement equations in appendix A were solved. The particular procedures for each case are described below. In order to make the comparison fair between the results of prior replacement and prior-specific training, an equal number of kernels were used: $K = 20$ in equations (7) through (10) for all MDN's trained in the following examples. Since the data point was the same in both cases, the same old posterior PDF was used for prior replacement of the Uniform and Gaussian priors. This PDF is shown in figure 1.

A Markov-chain Monte-Carlo (McMC) solution was obtained for reference in each case. This PDF was generated by taking $>10^4$ samples from the appropriate posterior and then estimating the densities. Because a large number of samples were taken, we can effectively consider this as the true posterior PDF. This is supported by the fact that the magnitude of autocorrelation between samples within the Markov-chain, in both cases, was typically much less than 0.01 at lags greater than 15 samples. The time taken to make the samples from the posterior using McMC, and the time taken in fitting the density to these, is far in excess of the time required by the MDN's to return a posterior estimate. However, we do not seek to compare the efficiency of MDN inversion to McMC methods (the advantages in terms of efficiency have already been demonstrated by Shahraeeni *et al* (2012) and references therein): we only use the McMC results for a comparison of solution quality.

(i) Uniform new prior. In order to perform prior replacement in this instance, equations (A.15) and (A.14) were evaluated. Numerical integration techniques were used to calculate the normalizing constant in equation (A.15). Figure 2(a) shows the new Uniform prior (that is, the prior which we want to apply). Figure 2(b) shows the McMC solution for $p_{\text{new}}(\mathbf{m}|\mathbf{d})$. Figure 2(c) shows the estimate of $p_{\text{new}}(\mathbf{m}|\mathbf{d})$ obtained using prior-specific training of a MDN with the Uniform new prior. Figure 2(d) shows the estimate of $p_{\text{new}}(\mathbf{m}|\mathbf{d})$ obtained by using prior replacement to replace the old prior implicit within the old posterior in figure 1 by the Uniform new prior in figure 2(a).

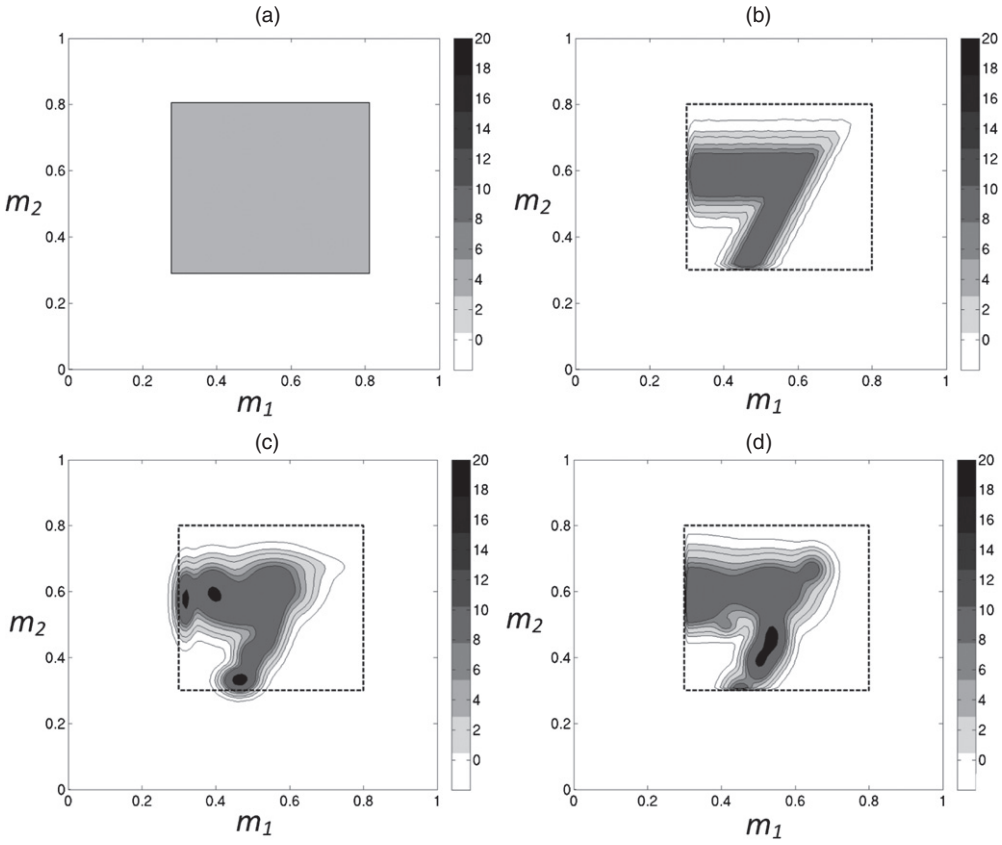

Figure 2. (a) The Uniform new prior PDF. (b) The posterior PDF obtained by McMC sampling in the case of the Uniform new prior. This can be viewed as the 'true' posterior PDF for comparison. (c) The new posterior PDF obtained from the output of a neural network (MDN) trained with samples generated directly from the new prior, i.e., prior-specific training. (d) The new posterior PDF obtained by removing the old prior from the old posterior in figure 1, and applying the new prior by prior replacement. In (b)–(d) the non-zero extent of the new prior is plotted with a stippled line. Prior-specific training has resulted in density appearing outside these bounds.

(ii) Gaussian new prior. In order to perform prior replacement in this case, equations (A.19) through (A.23) were evaluated. Numerical integration techniques were used to calculate the normalizing constant in equation (A.23). Figure 3(a) shows the new Gaussian prior (that is, the prior we wish to apply). Figure 3(b) shows the McMC solution for $p_{\text{new}}(\mathbf{m}|\mathbf{d})$. Figure 3(c) shows the estimate of $p_{\text{new}}(\mathbf{m}|\mathbf{d})$ obtained using prior-specific training of a MDN with the Gaussian new prior. Figure 3(d) shows the estimate of $p_{\text{new}}(\mathbf{m}|\mathbf{d})$ obtained by using prior replacement to replace the old prior implicit within the old posterior in figure 1 by the Gaussian new prior in figure 3(a).

3.2. Application to reservoir-scale inversion

Results (i) and (ii) above for the inversion of a single datum show that although variations exist, prior replacement and prior-specific training give comparable results. Thus prior replacement is shown to work in practice using MDN's. Furthermore, we may conclude that the prior replacement method would always be faster than prior-specific training if many such inversions

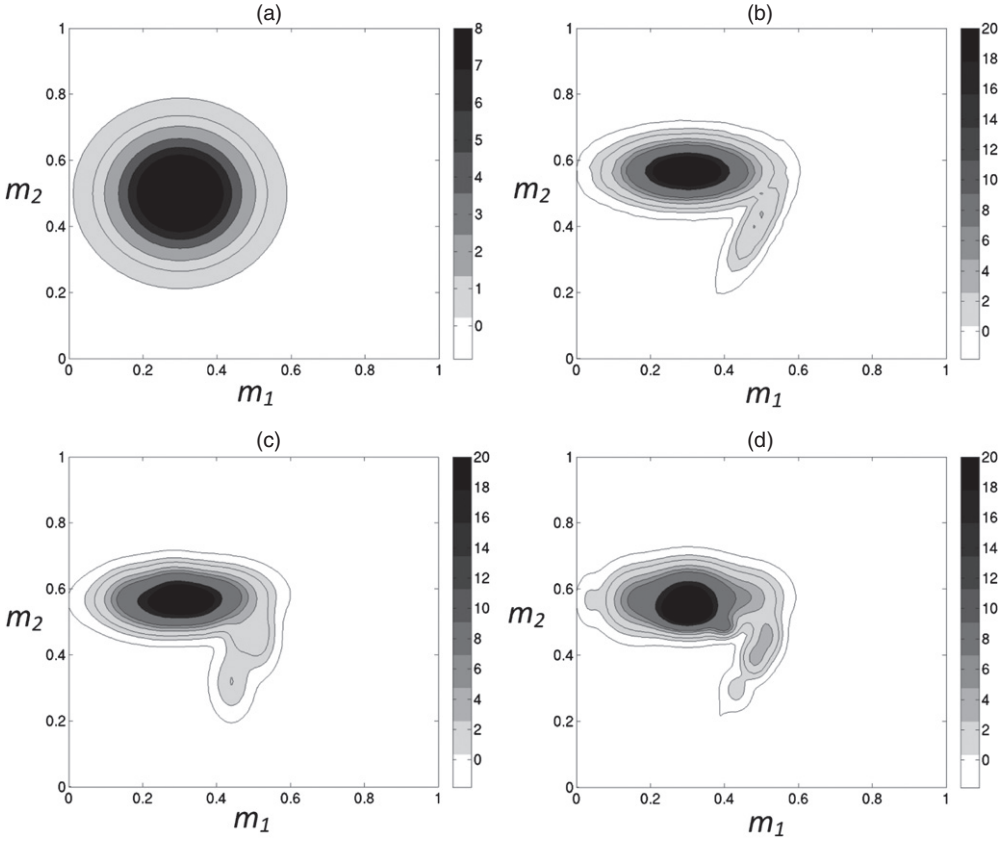

Figure 3. (a) The Gaussian new prior PDF. (b) The posterior PDF obtained by McMC sampling in the case of the Gaussian new prior. This can be viewed as the 'true' posterior PDF for comparison. (c) The new posterior PDF estimate obtained from the output of a neural network (MDN) trained with samples generated directly from the new prior, i.e., prior-specific training. (d) The new posterior PDF estimate obtained by removing the old prior from the old posterior in figure 1, and applying the new prior by prior replacement.

were performed, and if prior information varies between inversions. This can be understood by considering the computation times in the examples given: for prior replacement it took $\sim 10^2$ seconds to train the MDN using the old prior, then $\sim 10^{-3}$ seconds to run the MDN to obtain outputs for any given datum. Using prior replacement to construct the posterior for a new prior PDF for both the Uniform new prior (solving (A.15) and (A.14)) and Gaussian new prior (solving (A.23) and (A.22)) took $\sim 10^{-2}$ seconds. The total cost of prior replacement is therefore $\sim 10^2 + p \times (10^{-3} + 10^{-2})$ seconds, where p is the number of times prior information changes. For prior-specific training it also took $\sim 10^2$ seconds to train the MDN and again $\sim 10^{-3}$ seconds to run the MDN to obtain the outputs for a given datum. However, a new MDN has to be trained each time the prior changes so the total cost of prior-specific training is $\sim p \times (10^2 + 10^{-3})$ seconds. Therefore it is clear that if we were to apply both methods to the inversion of a large amount of data with varying priors then prior replacement could be orders of magnitude faster than prior-specific training.

An example application that occurs in practice is solving a petrophysical inverse problem similar to the above, but with one such problem defined in each cell of a discretized subsurface

(e.g., hydrocarbon) reservoir model. Then p would be equal to the number of cells in the model, which is typically over $\sim 10^5$ for 2D surveys and can approach $\sim 10^9$ for 3D surveys (Buland and Omre 2003, Shahraeeni and Curtis 2011, Shahraeeni *et al* 2012). To demonstrate the usefulness of the above conclusions in this case we carried out an inversion test on a 2D synthetic reservoir model using prior replacement. We created a model comprising a 50×50 grid of cells populated with the clay content by volume (m_1) and sandstone matrix porosity (m_2) parameters. Synthetic seismic impedance data were created using the forward model described in equation (11). It was assumed that wells were present within the reservoir, down which m_1 and m_2 were known exactly. This well data was used to generate the (varying) prior information across the reservoir model in a realistic way (as commonly performed in industrial geophysics): Gaussian prior distributions were determined at each cell by kriging (a form of interpolation, see e.g., Olea (1999, pp 7–17)) the known model parameters at the wells to each unknown cell using an appropriate covariance function and mean. The kriging estimate and variance were used as the Gaussian prior's mean and variance, respectively, in each cell. Inversion was carried out initially at each cell using a MDN trained with the broad old prior in equation (12), then the Gaussian priors were applied using prior replacement at each cell individually. Figure 4 depicts the model, the kriging-derived priors, and the inversion results. The inversion took $\sim$200 seconds using the prior replacement method. Given that the grid contains $50 \times 50 = 2500$ cells, an equivalent result using prior-specific training would take $\sim 10^5$ seconds. Thus, even in this simple test, the prior replacement method provided a reasonable Bayesian posterior solution with a factor 10^3 gain in computational efficiency over previous methods.

The fundamental reason for the spatial variability in the prior distributions in this example is that a geological model has been applied. This model is simple: it states that there should be some degree of continuity, or in other words spatial correlation, of the geological variables. Thus the assumption of continuity yields information (decreasing with distance from the wells) about the variables surrounding the well bores. In practical problems, more detailed prior information may be available about the subsurface geology. For example one may wish to invoke more complex information about heterogeneity than simple continuity, such as the distribution of faulting or sequence stratigraphic features (Remy *et al* 2009). This would change the Gaussian mean and variance, but would not alter the above conclusions about efficacy and efficiency.

4. Discussion

4.1. Numerical efficiency

We have shown that prior replacement can be useful for efficiently obtaining MDN inversion results with varying prior information. However, there is a significant computation required in the prior replacement method which is absent in prior-specific training. This is the normalization step (equation (A.15) or (A.23)), which must take place during every inversion for which the prior changes. While in the case of the results above it does not seem to slow the inversion greatly, as the number of dimensions of the model space grows, non-analytic integration will become significantly more costly. Using more advanced semi-analytical integration techniques for Gaussians (Drezner 1992) may reduce this cost to some extent (we used only numerical integration here). We might also consider using only Gaussian priors for both training an MDN, and for use in the prior replacement methodology. As shown in appendix A.5 this allows the normalizing constant to be calculated analytically. However, this puts constraints on the form of the priors that may be non-physical. For example, assuming

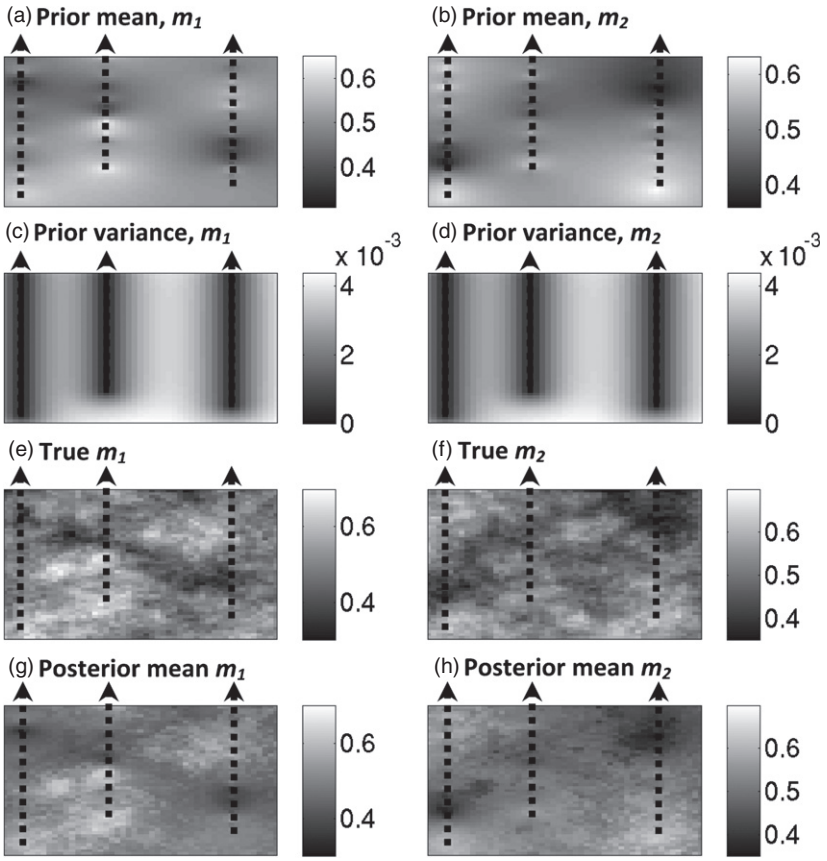

Figure 4. Synthetic inversion of seismic attributes for reservoir parameters for rock-physical parameters, clay content by volume (m_1) and sandstone matrix porosity (m_2), on a 2D grid model. Gaussian prior PDFs over m_1 and m_2 were determined at each cell by kriging the known values of those parameters from well trajectories (marked with stippled lines): the kriging mean and variance were used as the prior Gaussian mean and variance at each unknown cell. Then prior replacement, using these prior distributions, was applied to the old posterior in figure 1, to produce individual mixture density network (MDN) inversion results at each model cell. (a)–(b) and (c)–(d) show the prior means and variances in each cell, respectively, obtained by kriging the well data for m_1 and m_2. (e)–(f) The true rock physics parameters used to generate the synthetic data for each cell using equation (11). (g)–(h) The posterior mean for m_1 and m_2 determined using prior replacement in MDN inversion (note that these maps are smoother than the true model because we show the mean model estimator). The entire inversion method took $\sim$200 s using prior replacement. An equivalent result using prior-specific training would take $\sim 10^5$ s.

non-truncated Gaussians means assuming that the model space has non-zero probability everywhere; this might not be appropriate if we have hard constraints on model parameter values (e.g., in Shahraeeni and Curtis (2011), as above, porosity must lie between 0 and 1).

Nevertheless, if we are able to perform efficient analytical normalization (whether using the results derived in appendix A.5 or some alternative parametrization of posterior and priors) then prior replacement may be used for general Bayesian inverse problems (i.e., not MDN inversion) of much higher model space dimension. This could be very useful for problems where no closed form solution exists for the inverse. For example in subsurface reservoir

studies, flow data measured at wells are often used to infer the permeability structure of the subsurface. Due to the sparsity of data in time and space the problem is ill-posed. Furthermore, the forward physics which is used to assess the likelihood of any particular model must be solved numerically at great computational cost using flow simulation. Thus, if McMC methods are used to obtain an estimate of the posterior distribution over the subsurface permeability structure then it will be extremely computationally expensive. Due to the subjective nature of subsurface geological interpretation, however, prior information may change dramatically throughout the operational lifetime of a subsurface reservoir. In this scenario the ability to change the prior distribution, a utility which prior replacement provides, may lead to hugely increased efficiency. This would be possible, given the discussion above, since GMMs of the posterior distribution are often used in practice for such problems (Gu and Oliver 2005).

It should also be noted that normalization is not mandatory. If we do not require the absolute value of the probability, for example if we only wish to find the maximum-*a posteriori* estimator or wish simply to sample from the GMM, then the normalization step is not required and the new method becomes faster still. Furthermore, normalization is unlikely to be an issue in problems which employ NN inversion since the model space dimensionality is limited (typically to less than 10) by the amount of training data which may be processed in network training (Vapnik *et al* 1994).

4.2. Quality of the posterior estimate

Prior replacement always returns a distribution which is consistent with the final (i.e., the new) prior that is applied. This is not necessarily the case for prior-specific training because it fits the posterior distribution using Gaussians of finite size, and hence for example will always position some density outside of the bounds of a Uniform prior. This failure is clear in the results of prior-specific training in figure 2(c) where non-zero contours of the posterior lie in the zero probability regions of the new prior. By contrast, figure 2(d) shows that when prior replacement is used, no density is emplaced outside of the bounds of the new prior since the multiplication of prior and likelihood is explicit. Thus we envisage that prior replacement could be used in the future to ensure that the 'hard' bounds of a prior are enforced in the final posterior estimate.

Figure 3(c) shows a poor quality result using prior-specific training. Here the diagonally orientated lobe of low probability observed in the true posterior in figure 3(b) is poorly resolved in figure 3(c). The prior replacement result in figure 3(d) resolves this feature better. This phenomena may be attributed to the data used to train the MDN in each case. Specifically, in prior replacement samples are spread more equally across the model space due to the broader old prior that is used. As such, the variance of the posterior distribution may be better reproduced. By contrast in prior-specific training, sampling was concentrated around a peak in the posterior induced by the new, more informative, prior. Thus we might expect that the regions of high probability and hence the mean of the posterior would be better reproduced in this case. Indeed, it does appear that the high probability lobe in figure 3(c) compares more favourably in shape to that in figure 3(b) than does the lobe in figure 3(d). Thus, it appears that some aspects of the posterior estimates may be improved by prior replacement (compared to prior-specific training), whereas other aspects appear to be more poorly estimated. Thus, again we envisage that prior replacement could be used in the future to enhance the results of MDN inversion, where prior-specific training gives inadequate results. For example, it may be desirable that the posterior is better resolved within a certain region of the model space, thus we might use prior replacement to ensure that the training data contains more samples from this important region by using an appropriate old prior.

Clearly, a more sophisticated analysis of the quality of the results is necessary if the effect of prior replacement on the posterior estimate is to be understood in greater depth. To this end we have performed an empirical analysis of the effect of prior replacement on an inverse problem where the posterior is modelled by a single Gaussian kernel in appendices (included as supplementary online material, available from stacks.iop.org/IP/30/065002/mmedia). The results support our hypothesis that the effect of prior replacement on the quality of the posterior estimate is due to the distribution of samples used to estimate the old posterior (i.e., the form of the old prior). They also show that the effect is comparable, but not identical, to that of the Monte-Carlo technique of importance sampling (see e.g., Bishop 2006 pp 532–536), which suggests that at least an intuitive understanding of the effects of prior replacement may be borrowed from that method. The results in appendices (included as supplementary online material, available from stacks.iop.org/IP/30/065002/mmedia) also suggest that prior replacement could be used to manipulate the quality of the posterior estimate for general Bayesian inverse problems. For example, one may wish to better constrain the variance of the posterior in a Bayesian inverse problem solved using McMC. Then, similarly to those results obtained in MDN inversion in figure 3, this could be achieved by initially assuming a broad old prior and then, using prior replacement, emplacing the appropriate PDF as the new prior. However, more work is required to formalize such an operation.

There are a number of additional sources of error in the methodology which we have not yet described explicitly. The first of these arises from the fact that the NN which is used to emulate the mapping between data and model space has a number of parameters which must be defined manually. The most important of these is the number of weights in the network, which controls the complexity of the mapping. Allowing too much complexity may lead to over-fitting, whilst the opposite may lead to bias (a poor fit to training data). Also, the GMM itself is an imperfect model of the posterior since it has a finite number of kernels. Furthermore, NN training is performed using optimization which may be subject to local convergence effects. Thus careful effort must be made to validate the NN model before combining it with prior replacement. In general, one should be aware that it is much more difficult to predict the accuracy of the resulting posterior probabilities obtained using network inversion (especially coupled with prior replacement) than those obtained using McMC.

5. Conclusion

We have derived expressions which allow the analytical computation of Bayesian posterior probability distributions with a variety of prior distributions using the method of prior replacement, particularly for GMMs. This procedure involves inverting for an ‘old’ posterior, determined by a likelihood PDF and old prior PDF, and then analytically replacing the old prior with a ‘new’ prior. We have shown that prior replacement can be a useful method for varying the prior distribution within the result of MDN inversion. This avoids the computationally expensive step of MDN re-training at every instance that prior information changes (i.e., the MDN only has to be trained once). Prior replacement will then return a correct posterior provided the new prior distribution is non-zero only within the non-zero region of the old prior. We have also shown that prior replacement can be used as a tool to improve the results of MDN inversion in terms of certain statistical characteristics of the posterior distribution.

Acknowledgments

We would like to thank TOTAL E&P UK for supporting this work and two anonymous reviewers for their helpful suggestions and comments.

Appendix A. Prior replacement in mixture density network inversion

A.1. Preliminaries

We define two domains M_{old} and M_{new} which correspond to the non-zero regions of $p_{\text{old}}(\mathbf{m})$ and $p_{\text{new}}(\mathbf{m})$, respectively. As described in the main text $p_{\text{new}}(\mathbf{m})$ must be zero everywhere that $p_{\text{old}}(\mathbf{m})$ is zero, thus

$$M_{\text{new}} \subseteq M_{\text{old}}. \tag{A.1}$$

In general the priors are referred to as $p_{\text{new}}(\mathbf{m})$ and $p_{\text{old}}(\mathbf{m})$. However, we will employ Uniform distributions frequently so it is useful to define a Uniform distribution for both of these now, to aid the analysis in the following sections. We define a boxcar-like function δ, which has the properties

$$\delta(\mathbf{m}; M) = \begin{cases} 0 & \text{for} \quad \mathbf{m} \notin M \\ 1 & \text{for} \quad \mathbf{m} \in M \end{cases} \tag{A.2}$$

where $\mathbf{m}$ is the model vector and M is a region of the space of possible $\mathbf{m}$'s. Thus we define Uniform new and old priors for later use:

$$u_{\text{old}}(\mathbf{m}) = c_{\text{old}}\delta(\mathbf{m}; M_{\text{old}}) \tag{A.3}$$

$$u_{\text{new}}(\mathbf{m}) = c_{\text{new}}\delta(\mathbf{m}; M_{\text{new}}) \tag{A.4}$$

where the constants c_{old} and c_{new} are probability densities, whose exact values are related to the volumes of M_{old} and M_{new} (but are not important here).

A.2. Calculating the posterior PDF with a Uniform old prior

If $M_{\text{new}} \subseteq M_{\text{old}}$ is true and $p_{\text{old}}(\mathbf{m}) = u_{\text{old}}(\mathbf{m})$ then equation (9) can be simplified because $p_{\text{old}}(\mathbf{m})$ is constant over the volume in which $p_{\text{new}}(\mathbf{m}) \neq 0$. Substituting (A.3) into equation (5) we obtain

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k}\frac{p_{\text{new}}(\mathbf{m})}{p_{\text{old}}(\mathbf{m})}p_{\text{old}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k}\frac{p_{\text{new}}(\mathbf{m})}{c_{\text{old}}\delta(\mathbf{m}; M_{\text{old}})}p_{\text{old}}(\mathbf{m}|\mathbf{d}). \tag{A.5}$$

Given that $M_{\text{new}} \subseteq M_{\text{old}}$, $p_{\text{new}}(\mathbf{m})$ has zero probability density throughout the extent of the region of zero probability density of $u_{\text{old}}(\mathbf{m})$. Therefore, if we stipulate that $\mathbf{m} \in M_{\text{new}}$, the box-car function is unnecessary and may be removed from (A.5) thus:

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k}\frac{p_{\text{new}}(\mathbf{m})}{c_{\text{old}}}p_{\text{old}}(\mathbf{m}|\mathbf{d}), \quad m \in M_{\text{new}}. \tag{A.6}$$

Similarly, substituting (A.3) into equation (6) and again stipulating that $\mathbf{m} \in M_{\text{new}}$ allows the boxcar function to be removed and the limits of integration to be set to M_{new} thus

$$k = \int_{-\infty}^{+\infty} \frac{1}{k}\frac{p_{\text{new}}(\mathbf{m})}{c_{\text{old}}\delta(\mathbf{m}; M_{\text{old}})}p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m} \tag{A.7}$$

$$= \int_{M_{\text{new}}} \frac{p_{\text{new}}(\mathbf{m})}{c_{\text{old}}}p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m}. \tag{A.8}$$

Combining (A.6) and (A.8) and cancelling the constants we obtain the equation

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k'}p_{\text{new}}(\mathbf{m})p_{\text{old}}(\mathbf{m}|\mathbf{d}) \tag{A.9}$$

where the normalizing constant is

$$k' = \int_{M_{\text{new}}} p_{\text{new}}(\mathbf{m})p_{\text{old}}(\mathbf{m}|\mathbf{d}). \tag{A.10}$$

It should be noted that the change in the limit of integration in (A.8) may not be trivial if the dimensionality of the model space is high and/or the Uniform distribution has complicated bounds.

A.3. Calculating the posterior with a Uniform old prior and Uniform new prior

Equations (A.9) and (A.10) can be used under the conditions that $M_{\text{new}} \subseteq M_{\text{old}}$ and the old prior is Uniform, $p_{\text{old}}(\mathbf{m}) = u_{\text{old}}(\mathbf{m})$. If also the new prior is Uniform, $p_{\text{new}}(\mathbf{m}) = u_{\text{new}}(\mathbf{m})$, then the result is simpler. Combining (A.4), (A.9) and (A.10) we obtain

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{c_{\text{new}}\delta(\mathbf{m}; M_{\text{new}}, p_{\text{old}})p_{\text{old}}(\mathbf{m}|\mathbf{d})}{c_{\text{new}}\int_{M_{\text{new}}}\delta(\mathbf{m}; M_{\text{new}})p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m}}, \tag{A.11}$$

as before $\mathbf{m} \in M_{\text{new}}$ so the boxcar functions may be removed. Doing this and cancelling constants gives

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{\int_{M_{\text{new}}} p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m}} p_{\text{old}}(\mathbf{m}|\mathbf{d}) \tag{A.12}$$

$$= \frac{1}{k''} p_{\text{old}}(\mathbf{m}|\mathbf{d}), \quad m \in M_{\text{new}} \tag{A.13}$$

where we have recognized that we have now a normalizing constant in the denominator which we denote with k''. Substituting equation (8) into (A.13) yields

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k''}\sum_{i=1}^{N}\alpha_i\phi(\mathbf{m}; \mu_i, \mathbf{\Sigma}_i), \quad m \in M_{\text{new}} \tag{A.14}$$

where

$$k'' = \int_{M_{\text{new}}} p_{\text{old}}(\mathbf{m}|\mathbf{d})\,\mathrm{d}\mathbf{m} = \int_{M_{\text{new}}}\sum_{i=1}^{N}\alpha_i\phi(\mathbf{m}; \mu_{\mathbf{i}}, \mathbf{\Sigma}_i)\,\mathrm{d}\mathbf{m}. \tag{A.15}$$

Evaluation of the normalizing constant k'' requires only the integration of the series of Gaussians (the GMM) in (A.15) over the non-zero region of M_{new}. This implies the need to evaluate a definite integral of a multivariate normal distribution. Whilst this does not have an analytic expression (Drezner 1992), it has been widely studied due to its importance in probability theory. Many algorithms exist for its evaluation (Drezner and Wesolowsky 1990, Genz and Bretz 1999, 2002, Genz 2004), apart from simple numerical integration techniques (Riley *et al* 2006, pp 1000–1009).

A.4. Calculating the posterior with Uniform old prior and Gaussian new prior

If $p_{\text{old}}(\mathbf{m})$ is Uniform and $p_{\text{new}}(\mathbf{m})$ is a Gaussian then we can use (A.10) to evaluate the normalizing factor in (A.9), and hence find the new posterior. We must explicitly state that this new prior obeys $M_{\text{new}} \subseteq M_{\text{old}}$, that is that its non-zero extent is limited to that of the old prior. Thus, we define the new prior as a truncated Gaussian—the product of a Gaussian and the boxcar-type function defined in (A.4):

$$p_{\text{new}}(\mathbf{m}) = c\phi(\mathbf{m}; \mu_{\text{new}}, \mathbf{\Sigma}_{\text{new}})\delta(\mathbf{m}; M_{\text{new}}) \tag{A.16}$$

where c is a normalizing constant. We again use the notation $\phi(\mathbf{m}; \mu, \mathbf{\Sigma})$ to denote a normalized Gaussian function as a function of $\mathbf{m}$ with mean vector μ and covariance matrix $\mathbf{\Sigma}$. The subscript new indicates that we refer to parameters belonging to the new prior, p_{new}. Substituting (A.16) into (A.9), the c constant disappears henceforth (since it exists in both the numerator and denominator), then stipulating that $\mathbf{m} \in M_{\text{new}}$ allows us to write

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k'}\phi(\mathbf{m}; \mu_{\text{new}}, \mathbf{\Sigma}_{\text{new}})\sum_{i=1}^{N}\alpha_i\phi(\mathbf{m}; \mu_i, \mathbf{\Sigma}_i), \quad m \in M_{\text{new}}. \tag{A.17}$$

Similarly for the normalizing factor we can substitute (A.16) into (A.10) and since $\mathbf{m} \in M_{\text{new}}$, remove the boxcar function:

$$k' = \int_{M_{\text{new}}} \phi(\mathbf{m};\, \mu_{\text{new}},\, \boldsymbol{\Sigma}_{\text{new}}) \sum_{i=1}^{N} \alpha_i \phi(\mathbf{m};\, \mu_i,\, \boldsymbol{\Sigma}_i)\, \mathrm{d}\mathbf{m}. \tag{A.18}$$

In order to simplify (A.18) and subsequently to evaluate (A.17) we use the result that the product of two Gaussians is an un-normalized Gaussian (Ahrendt 2005). This allows us to obtain an analytical expression for a series of single Gaussians within each of these equations. We can combine the Gaussians as such (Ahrendt 2005)

$$\sum_{i=1}^{N} \alpha_i \phi(\mathbf{m};\, \mu_i,\, \boldsymbol{\Sigma}_i) \phi(\mathbf{m};\, \mu_{\text{new}},\, \boldsymbol{\Sigma}_{\text{new}}) = \sum_{i=1}^{N} \alpha_i R_i \phi(\mathbf{m};\, \mu_i{}',\, \boldsymbol{\Sigma}_i{}') \tag{A.19}$$

where the mean and covariance parameters are now given by

$$\mu'_i = \left(\boldsymbol{\Sigma}'_i \boldsymbol{\Sigma}_{\text{new}}{}^{-1} \mu_{\text{new}}\right) + \left(\boldsymbol{\Sigma}'_i \boldsymbol{\Sigma}_i{}^{-1} \mu_i\right) \qquad \text{and} \qquad \boldsymbol{\Sigma}'_i = \left(\boldsymbol{\Sigma}_{\text{new}}{}^{-1} + \boldsymbol{\Sigma}_i{}^{-1}\right)^{-1}, \tag{A.20}$$

and the constant R_i is given by

$$R_i = |2\pi\,(\boldsymbol{\Sigma}_{\text{new}} + \boldsymbol{\Sigma}_i)|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\mu_{\text{new}} - \mu_i)^T (\boldsymbol{\Sigma}_{\text{new}} + \boldsymbol{\Sigma}_i)^{-1} (\mu_{\text{new}} - \mu_i)\right]. \tag{A.21}$$

Upon substitution of the Gaussian product given in (A.19), (A.17) becomes

$$p_{\text{new}}(\mathbf{m}|\mathbf{d}) = \frac{1}{k'} \sum_{i=1}^{N} \alpha_i R_i \phi(\mathbf{m};\, \mu_i{}',\, \boldsymbol{\Sigma}_i{}') \tag{A.22}$$

and (A.18) becomes

$$k' = \int_{M_{\text{new}}} \sum_{i=1}^{N} \alpha_i R_i \phi(\mathbf{m};\, \mu_i{}',\, \boldsymbol{\Sigma}_i{}')\, \mathrm{d}\mathbf{m}. \tag{A.23}$$

Equation (A.23) can be evaluated by integration over the truncated Gaussians as in the previous section. Once this is substituted into (A.22) the full posterior can be calculated.

A.5. Calculating the posterior with both old and new Gaussian priors

The special case of having both a Gaussian old prior $p_{\text{old}}(\mathbf{m})$, and a Gaussian new prior $p_{\text{new}}(\mathbf{m})$, is interesting since this may permit the normalization constant to be calculated analytically in equations (9) and (10). To see this we explicitly expand the priors in terms of Gaussian kernels. In contrast to the previous section, we express the new and old priors as full Gaussians so we do not need to truncate either prior as they both span the infinite model space. Therefore

$$p_{\text{old}}(\mathbf{m}) = \phi(\mathbf{m};\, \mu_{\text{old}},\, \boldsymbol{\Sigma}_{\text{old}}), \tag{A.24}$$

and

$$p_{\text{new}}(\mathbf{m}) = \phi(\mathbf{m};\, \mu_{\text{new}},\, \boldsymbol{\Sigma}_{\text{new}}). \tag{A.25}$$

Since both priors are Gaussian we substitute (A.24) and (A.25) into equation (10),

$$k = \int_{-\infty}^{+\infty} \frac{\phi(\mathbf{m};\, \mu_{\text{new}},\, \boldsymbol{\Sigma}_{\text{new}}) \sum_{i=1}^{N} \alpha_i \phi(\mathbf{m};\, \mu_i,\, \boldsymbol{\Sigma}_i)}{\phi(\mathbf{m};\, \mu_{\text{old}},\, \boldsymbol{\Sigma}_{\text{old}})} \mathrm{d}\mathbf{m}. \tag{A.26}$$

As previously, the Gaussians can be combined in some way to make the calculation simpler. There are two ways of combining the Gaussians in (A.26). We could divide the GMM by the old prior and then multiply by the new prior, or we could divide the new prior by the old prior and then multiply by the GMM. We discuss the latter here as it is much simpler because it

involves only the division of two single Gaussians rather than involving the series of Gaussians in the division (since this is more complicated than the multiplication of two Gaussians, as discussed below).

The multiplication of one Gaussian by another is always Gaussian (Bromiley 2003), therefore if we can ensure that the division of the new prior by the old prior is Gaussian then the whole operation will always yield a Gaussian. However, the division of one Gaussian by another does not always yield a Gaussian. This can be seen by first writing out the expression for a multivariate Gaussian

$$\phi(\mathbf{m};\, \mu,\, \mathbf{\Sigma}) = |2\pi\, \mathbf{\Sigma}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{m}-\mu)^T\, \mathbf{\Sigma}^{-1}(\mathbf{m}-\mu)}. \tag{A.27}$$

For the expression in (A.27) to behave as a Gaussian the covariance matrix must be positive definite (Rue and Held 2005). Then, since the inverse of a positive definite matrix is positive definite, the condition

$$\mathbf{m}^T\, \mathbf{\Sigma}^{-1} \mathbf{m} > 0 \quad \forall\, \mathbf{m} \in \mathbb{R}^d \tag{A.28}$$

must be true for a valid Gaussian. We can write the division of the new by the old prior in (A.26) as a product but with the covariance matrix of the old prior multiplied by -1,

$$k = \int_{-\infty}^{+\infty} \phi(\mathbf{m};\, \mu_{\text{new}},\, \mathbf{\Sigma}_{\text{new}}) \phi(\mathbf{m};\, \mu_{\text{old}},\, -\mathbf{\Sigma}_{\text{old}}) \sum_{i=1}^{N} \alpha_i \phi(\mathbf{m};\, \mu_i,\, \mathbf{\Sigma}_i)\, \mathrm{d}\mathbf{m} \tag{A.29}$$

and the Gaussian division can be written in the form of a single Gaussian as

$$k = \int_{-\infty}^{+\infty} \phi(\mathbf{m};\, \mu',\, \mathbf{\Sigma}') \sum_{i=1}^{N} \alpha_i \phi(\mathbf{m};\, \mu_i,\, \mathbf{\Sigma}_i)\, \mathrm{d}\mathbf{m}. \tag{A.30}$$

Then the equations for the mean vector, covariance matrix and normalization constant (given in (A.20) and (A.21)) for the product of two Gaussians are valid, with this modification to the old covariance matrix. Thus for the result of this Gaussian division, from (A.20) we have

$$\mathbf{\Sigma}' = \left(\mathbf{\Sigma}_{\text{new}}^{-1} - \mathbf{\Sigma}_{\text{old}}{}^{-1}\right)^{-1}, \tag{A.31}$$

and

$$\mu' = \left(\mathbf{\Sigma}' \mathbf{\Sigma}_{\text{new}}^{-1} \mu_{\text{new}}\right) - \left(\mathbf{\Sigma}' \mathbf{\Sigma}_{\text{old}}^{-1} \mu_{\text{old}}\right). \tag{A.32}$$

Clearly, $\mathbf{\Sigma}'$ must be positive definite for the Gaussian division to yield a valid Gaussian. In other words, the condition (A.28) must apply to $\mathbf{\Sigma}'$. Thus substituting (A.31) into (A.28) yields

$$\mathbf{m}^T\, \mathbf{\Sigma}'^{-1} \mathbf{m} = \mathbf{m}^T \left(\mathbf{\Sigma}_{\text{new}}^{-1} - \mathbf{\Sigma}_{\text{old}}^{-1}\right) \mathbf{m} > 0 \quad \forall\, \mathbf{m} \in \mathbb{R}^d \tag{A.33}$$

which may be rewritten to give the condition as

$$\mathbf{m}^T\, \mathbf{\Sigma}_{\text{new}}^{-1} \mathbf{m} - \mathbf{m}^T\, \mathbf{\Sigma}_{\text{old}}^{-1} \mathbf{m} > 0 \quad \forall\, \mathbf{m} \in \mathbb{R}^d. \tag{A.34}$$

If both the old and new priors are valid Gaussians then their covariance matrices are positive definite and obey (A.28). Thus (A.34) cannot be true for all possible $\mathbf{\Sigma}_{\text{new}}$ and $\mathbf{\Sigma}_{\text{old}}$. In order to ensure that (A.33) holds we could design the new and old priors specifically by manipulating their eigen-decompositions, for example (but we will not discuss such possibilities here). Usefully, if (A.33) is true, (A.32) will always give a valid (i.e., real) mean vector for the resulting Gaussian. Therefore, the values of the mean vectors of the old and new priors do not effect whether the division of these two Gaussians yields another Gaussian or not, and so the means of the old and new priors may have any value.

Appendix B. The forward-model: the Yin–Marion shaly-sand model

The forward petrophysical model which we use is the Yin–Marion shaly-sand model (Marion 1990, Yin *et al* 1993, Avseth *et al* 2005). In this model two distinct domains are defined for sand-shale mixtures: sandstones with a secondary shale component, called shaly-sands, and shales with secondary sand component, called sandy-shales. In the former domain clay particles are assumed to be within the pore space of a sandstone frame. Increasing shale content fills this pore space, decreasing porosity linearly. Thus in this case the porosity varies according to

$$\phi = \phi_s - C(1 - \phi_{\text{sh}}), \quad \forall\, C < \phi_s \tag{B.1}$$

where C is the shale volume fraction, ϕ_s is porosity of the clean sandstone frame and ϕ_{sh} is the intrinsic porosity of the shale. In the other domain, the sandy-shale domain, the shale volume fraction is greater than the porosity of the clean sandstone frame. In this case the rock is no longer considered to consist of a sandstone frame with a pore space, but instead it is considered to be shale with sand inclusions. There is no sandstone porosity, only isolated grains, and the only porosity which exists is within the intrinsic pore space of the shale. The total porosity is then:

$$\phi = C\phi_{\text{sh}}, \quad \forall\, C \geqslant \phi_s. \tag{B.2}$$

The volume fractions of the components (i.e., shale, sand and pore fluid) predicted by these equations can then be treated in a number of different ways to predict the S-wave impedance (d_1) and P-wave impedance (d_2) of the bulk rock. To do this, we chose to use the upper Hashin–Shtrikman bound for the mixture in the shaly-sand case and the lower bound in the sandy-shale case (following Avseth *et al* 2005) to approximately simulate the two different assumed micro-geometries of the domains (see Mavko *et al* (2009), for an explanation of the micro-geometry implied by these bounds). The densities can be calculated with the volume fractions and the known densities of the constituents.

We assumed a constant mineralogy of the shale and sand components in this model. We also assumed that the pore-filling water was pure water. Thus the values for the elastic moduli and densities of these constituent materials are taken from examples in the literature (e.g., Mavko *et al* 2009). Furthermore the intrinsic porosity of shale is kept constant. Thus there are two model parameters which could vary: the intrinsic sandstone porosity (ϕ_s) and the clay volume of the rock (C). Thus, we write the petrophysical model parameters vector as $\mathbf{m} = [m_1, m_2]$, where we have used for convenience in the main text the notation $m_1 = C$ and $m_2 = \phi_s$.

We symbolically write the Yin–Marion shaly-sand model described above as $\mathbf{f}(\mathbf{m})$. We use it to predict S-wave impedance (d_1) and P-wave impedance (d_2) given volume clay content (m_1) and sandstone matrix porosity (m_2). It is a deterministic model but we included a random element by adding random Gaussian noise ($\mathbf{e}$) to its output. Thus the forward model is written

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) + \mathbf{e}, \quad \mathbf{e} \sim \phi(\mathbf{e}; \mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{d}}), \quad \boldsymbol{\Sigma}_{\mathbf{d}} = \begin{bmatrix} \sigma_P^2 & 0 \\ 0 & \sigma_S^2 \end{bmatrix} \tag{B.3}$$

where $\mathbf{f}(\mathbf{m})$ represents the Yin–Marion shaly-sand model, $\phi()$ has its usual meaning as a Gaussian function, $\mathbf{m} = [m_1, m_2]$ is the vector of model parameters and $\mathbf{d} = [d_1, d_2]$ is the data vector of impedances. The random Gaussian noise is specified by the standard deviations in the data covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{d}}$: $\sigma_P = 1.5 \times 10^4\ \text{s}^{-1}\text{m}^{-2}\text{kg}$ and $\sigma_S = 1.0 \times 10^4\ \text{s}^{-1}\text{m}^{-2}\text{kg}$. Since the noise is Gaussian, an appropriate PDF can be constructed as in equation (11).

References

Ahrendt P 2005 The multivariate Gaussian probability distribution *Technical report* IMM2005-03312 Technical University of Denmark

Avseth P, Mukerji T and Mavko G 2005 *Quantitative Seismic Interpretation, Applying Rock Physics Tool to Reduce Interpretation Risk* (Cambridge: Cambridge University Press)

Bailer-Jones C and Smith K 2011 Combining probabilities *Technical report* GAIA-C8-TN-MPIA-CBJ-053 Max Planck Institute for Astronomy, Heidelberg

Bishop C M 1994 Mixture density networks *Technical report* NCRG/4288 Aston University

Bishop C M 1995 *Neural Networks for Pattern Recognition* (Oxford: Clarendon)

Bishop C M 2006 *Pattern Recognition and Machine Learning* (New York: Springer)

Bromiley P A 2003 Products and convolutions of Gaussian distributions *TINA Internal Report* 2003-003 University of Manchester

Buland A and Omre H 2003 Bayesian linearized AVO inversion *Geophysics* **68** 185–98

Daniels M J 1999 A prior for the variance in hierarchical models *Can. J. Stat.* **27** 567–78

Devilee R, Curtis A and Roy-Chowdhury K 1999 An efficient, probabilistic neural network approach to solving inverse problems: inverting surface wave velocities for Eurasian crustal thickness *J. Geophys. Res.* **104** 28841–57

Dong Y, Forster B and Milne A 1997 Segmentation of radar imagery using Gaussian Markov random field model *Int. J. Remote Sensing* **20** 1617–39

Drezner Z 1992 Computation of the multivariate normal integral *ACM Trans. Math. Softw.* **18** 470–80

Drezner Z and Wesolowsky G O 1990 On the computation of the bivariate normal integral *J. Stat. Comput. Simul.* **35** 101–7

Duijndam A 1988 Bayesian estimation in seismic inversion: Part I. Principles *Geophys. Prospect.* **36** 878–98

Eidsvik J, Finley A O, Banerjee S and Rue H 2012 Approximate Bayesian inference for large spatial datasets using predictive process models *Comput. Stat. Data Anal.* **56** 1362–80

El-Qady G and Ushijima K 2001 Inversion of DC resistivity data using neural networks *Geophys. Prospect.* **49** 417–30

Gelman A, Carlin J B, Stern H S and Rubin D B 1995 *Bayesian Data Analysis* (London: Chapman and Hall)

Genz A 2004 Numerical computation of rectangular bivariate and trivariate normal and t probabilities *Stat. Comput.* **14** 251–60

Genz A and Bretz F 1999 Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts *J. Stat. Comput. Simul.* **63** 103–17

Genz A and Bretz F 2002 Comparison of methods for the computation of multivariate t probabilities *J. Comput. Graph. Stat.* **11** 950–71

Gu Y and Oliver D 2005 History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter *SPE J.* **10** 217–24

Hershey J R and Olsen P A 2007 Approximating the Kullback Leibler divergence between Gaussian mixture models *ICASSP'07: IEEE Int. Conf. Acoustics, Speech and Signal Processing* vol 4 (Piscataway, NJ: IEEE) pp IV–317–20

Hobert J P and Casella G 1996 The effect of improper priors on Gibbs sampling in hierarchical linear mixed models *J. Am. Stat. Assoc.* **91** 1461–73

Jaynes E T 1986 Bayesian methods: general background *Maximum Entropy and Bayesian Methods in Applied Statistics* (Cambridge: Cambridge University Press) pp 1–25

Jeffreys H 1961 *Theory of Probability* (Oxford: Clarendon)

Johansson E M, Dowla F U and Goodman D M 1991 Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method *Int. J. Neural Syst.* **2** 291–301

Kullback S and Leibler R A 1951 On information and sufficiency *Ann. Math. Stat.* **22** 79–86

Liu Z and Liu J 1998 Seismic-controlled nonlinear extrapolation of well parameters using neural networks *Geophysics* **63** 2035–41

Lupton R H 1993 *Statistics in Theory and Practice* (Princeton, NJ: Princeton University Press)

Maiti S, Krishna Tiwari R and Kümpel H J 2007 Neural network modelling and classification of lithofacies using well log data: a case study from KTB borehole site *Geophys. J. Int.* **169** 733–46

Maiti S and Tiwari R K 2010 Automatic discriminations among geophysical signals via the Bayesian neural networks approach *Geophysics* **75** E67–78

Marion D P 1990 Acoustical, mechanical and transport properties of sediments and granular materials *PhD Thesis* Stanford University, Department of Geophysics
Mavko G, Mukerji T and Dvorkin J 2009 *The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media* (Cambridge: Cambridge University Press)
McLachlan G and Peel D 2000 *Finite Mixture Models* (New York: Wiley)
Meier U, Curtis A and Trampert J 2007a Fully nonlinear inversion of fundamental mode surface waves for a global crustal model *Geophys. Res. Lett.* **34** L16304
Meier U, Curtis A and Trampert J 2007b Global crustal thickness from neural network inversion of surface wave data *Geophys. J. Int.* **169** 706–22
Meier U, Trampert J and Curtis A 2009 Global variations of temperature and water content in the mantle transition zone from higher mode surface waves *Earth Planet. Sci. Lett.* **282** 91–101
Michie D, Spiegelhalter D J and Taylor C C 1994 *Machine Learning, Neural and Statistical Classification* (New York: Ellis Horwood)
Mosegaard K and Sambridge M 2002 Monte Carlo analysis of inverse problems *Inverse Problems* **18** R29
Oehlert G W 1992 A note on the delta method *Am. Stat.* **46** 27–9
Olea R 1999 *Geostatistics for Engineers and Earth Scientists* (Boston, MA: Kluwer)
Petersen K B and Pedersen M S 2006 *The Matrix Cookbook* (Technical University of Denmark)
Remy N, Boucher A and Wu J 2009 *Applied Geostatistics with SGeMS: A User's Guide* (Cambridge: Cambridge University Press)
Riley K F, Hobson M P and Bence S J 2006 *Mathematical Methods for Physics and Engineering* (Cambridge: Cambridge University Press)
Roth G and Tarantola A 1994 Neural networks and inversion of seismic data *J. Geophys. Res.* **99** 6753–68
Rue H and Held L 2005 *Gaussian Markov Random Fields: Theory and Applications* (London: Chapman and Hall)
Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors *Nature* **323** 533–6
Sambridge M, Gallagher K, Jackson A and Rickwood P 2006 Trans-dimensional inverse problems, model comparison and the evidence *Geophys. J. Int.* **167** 528–42
Scales J A and Tenorio L 2001 Prior information and uncertainty in inverse problems *Geophysics* **66** 389–97
Shahraeeni M S 2011 Inversion of seismic attributes for petrophysical parameters and rock facies *PhD Thesis* The University of Edinburgh
Shahraeeni M S and Curtis A 2011 Fast probabilistic nonlinear petrophysical inversion *Geophysics* **76** E45–58
Shahraeeni M S, Curtis A and Chao G 2012 Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data *Geophysics* **77** O1–19
Sun D, Tsutakawa R K and He Z 2001 Propriety of posteriors with improper priors in hierarchical linear mixed models *Stat. Sin.* **11** 77–95
Sun Y, Li B and Genton M G 2012 Geostatistics for large datasets *Advances and Challenges in Space-time Modelling of Natural Events* ed E Porcu *et al* (Berlin: Springer) pp 55–77
Tarantola A 1987 *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation* (Amsterdam: Elsevier)
Ulrych T J, Sacchi M D and Woodbury A 2001 A Bayes tour of inversion: a tutorial *Geophysics* **66** 55–69
Van der Baan M and Jutten C 2000 Neural networks in geophysical applications *Geophysics* **65** 1032–47
Van der Vaart A W 1998 *Asymptotic Statistics* (Cambridge: Cambridge University Press)
Vapnik V, Levin E and Le Cun Y 1994 Measuring the VC-dimension of a learning machine *Neural Comput.* **6** 851–76
Yin H, Nur A and Mavko G 1993 Critical porosity A physical boundary in poroelasticity *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.* **30** 805–8
Zhang R, Czado C and Sigloch K 2013 A Bayesian linear model for the high-dimensional inverse problem of seismic tomography *Ann. Appl. Stat.* **7** 1111–38