Geophysical Journal International

Geophys. J. Int. (2018) **214,** 845–875
Advance Access publication 2018 April 26
GJI Seismology

doi: 10.1093/gji/ggy163

Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies

Muhammad Atif Nawaz and Andrew Curtis

School of Geosciences, Grant Institute, University of Edinburgh, Edinburgh, United Kingdom. E-mail: Muhammad.AtifNawaz@ed.ac.uk; Andrew.Curtis@ed.ac.uk

Accepted 2018 April 20. Received 2018 March 27; in original form 2017 December 19

SUMMARY

We introduce a new Bayesian inversion method that estimates the spatial distribution of geological facies from attributes of seismic data, by showing how the usual probabilistic inverse problem can be solved using an optimization framework while still providing full probabilistic results. Our mathematical model consists of seismic attributes as observed data, which are assumed to have been generated by the geological facies. The method infers the post-inversion (posterior) probability density of the facies plus some other unknown model parameters, from both the seismic attributes and geological prior information. Most previous research in this domain is based on the *localized likelihoods* assumption, whereby the seismic attributes at a location are assumed to depend on the facies only at that location. Such an assumption is unrealistic because of imperfect seismic data acquisition and processing, and fundamental limitations of seismic imaging methods. In this paper, we relax this assumption: we allow probabilistic dependence between seismic attributes at a location and the facies in any neighbourhood of that location through a spatial filter. We term such likelihoods *quasi-localized*.

Exact Bayesian inference is impractical because it requires normalization of the posterior distribution which is intractable for large models and must be approximated. Stochastic sampling (e.g. by using Markov chain Monte Carlo) is the most commonly used approximate inference method but it is computationally expensive and detection of its convergence is often subjective and unreliable. We use the *variational Bayes* method which is a more efficient alternative that offers reliable detection of convergence. It achieves this by replacing the intractable posterior distribution by a tractable approximation. Inference can then be performed using the approximate distribution in an optimization framework, thus circumventing the need for sampling, while still providing probabilistic results. We show in a noisy synthetic example that the new method recovered the coefficients of the spatial filter with reasonable accuracy, and recovered the correct facies distribution. We also show that our method is robust against weak prior information and non-localized likelihoods, and that it outperforms previous methods which require likelihoods to be localized. Our method is computationally efficient, and is expected to be applicable to 3-D models of realistic size on modern computers without incurring any significant computational limitations.

Key words: Image processing; Spatial analysis; Numerical approximations and analysis; Neural networks, fuzzy logic; Inverse theory; Probability distributions.

1 INTRODUCTION

Geological heterogeneity plays a key role in reservoir characterization and fluid-flow prediction in all subsurface reservoirs, and in the quantification of concomitant reservoir development and economic risk. The spatial distributions of geological facies can be estimated using a variety of information such as seismic and borehole data together with prior geological knowledge. Seismic data provides 2- or 3-D subsurface coverage but is limited in resolution, usually to heterogeneity on length scales greater than tens or hundreds of metres. Borehole data, on the other hand, exhibit far higher resolution along the 1-D borehole trajectory, but boreholes are usually sparsely distributed and therefore provide poor 3-D spatial coverage. These differences in spatial coverage and resolution provide different types and degrees of information or uncertainty about the geological facies. An appropriate data inversion or inference method must therefore combine a variety

of measurements with different coverage and resolution, and all available prior information, and must assess the true resultant state of information and uncertainty about the subsurface.

Bayesian inference offers a convenient mathematical framework to achieve this. Information about parameters of interest is described by probability distributions. The distribution that describes uncertainty in model parameters given only the prior information is called the *prior distribution*, while the distribution describing the total resultant state of information given all the above data and information is called the *posterior distribution*. Estimating properties of the latter is the goal in probabilistic inverse problems.

The computation and digital storage of complete joint posterior probability distributions over a large number of parameters using Bayesian inversion is commonly intractable given available computing power. Probability distributions in high-dimensional spaces are therefore generally explored using Markov chain Monte Carlo (McMC) sampling—a suite of methods that theoretically produce a set of samples of parameter values which converge in density to that of the true posterior probability distribution as the number of samples tends to infinity. McMC methods therefore obtain a numerical approximation of the true posterior distribution using a finite number of samples, with a theoretical guarantee of asymptotic convergence only as sampling extends to infinity.

Monte Carlo-based inversion methods are computationally expensive in most models of practical interest because as the number of parameters gets large, one can only expect that the distribution of samples would converge after generating an infeasible number of samples—often referred to as the curse of dimensionality (Curtis & Lomax 2001). Further, detecting convergence of any McMC method to the posterior distribution requires the use of subjectively selected criteria. As a result, any estimate of the posterior that is obtained from any specific, fixed, finite set of Monte Carlo samples may be biased by that particular set of samples depending on the criteria used for the detection of convergence. We refer to such a bias as *convergence-related bias*. Posterior distributions can sometimes be assumed to take factorizable forms that divide high dimensional problems into lower dimensional problems alleviating some of the difficulties in McMC sampling (e.g. Sambridge & Mosegaard 2002; Gallagher *et al.* 2009; Rimstad *et al.* 2012), but even for well-designed formulations of the posterior distributions the curse of dimensionality remains a barrier to rapid and accurate sampling-based estimation.

Although McMC is a very general method for probabilistic inference and has been used to solve a wide variety of inverse problems, more efficient inference methods can be designed for certain classes of spatial problems based on variational principles. Variational Bayesian inference is an alternative to McMC-based inference that provides proxy posterior distributions which are close to the true posteriors in some specific sense. They are also more computationally tractable, and may be estimated analytically or numerically within an optimization framework. In this paper, we introduce a Bayesian inversion method under the variational approximation, which we refer to as variational Bayesian inversion (VBI).

This paper is organized as follows. We first provide some background information and a short literature review on the inversion of spatial distributions of geological facies from geophysical (seismic and borehole) data, and introduce our model. Then we discuss the Bayesian inversion approach and introduce the mathematical formulation of the VBI method to jointly estimate spatial distribution of facies and model parameters. Within the variational Bayesian framework we discuss the expectation maximization (EM) algorithm which iteratively estimates the unknown variables and model parameters alternately in the E-step and the M-step, respectively. In the E-step, we propose the use of a *message passing* algorithm [such as loopy belief propagation (LBP)] and provide a mathematical justification of that in the light of previous research in Statistical Physics and Artificial Intelligence. The M-step maximizes the likelihoods of the input seismic attributes by optimizing the model parameters. After providing mathematical details of our algorithm, we discuss its computational complexity. Then we provide a synthetic test example where we invert for three geological facies (shale, brine-sand and gas-sand) in a deltaic environment from multiple seismic attributes computed using rock physics relationships. The synthetic test example is followed by a discussion, and finally the conclusions of this research.

2 BACKGROUND

Many different methods have been developed to invert for rock properties from seismic data constrained by borehole data. Since various seismic measurements such as seismic amplitudes and velocities are directly affected by the elastic properties of rocks, several deterministic and probabilistic methods have been developed to invert seismic data for the elastic properties of rocks (Buland & Omre 2003; Bosch *et al.* 2010; Grana & Mukerji 2015). Some methods have also been developed that aim to invert seismic data for petrophysical rock properties, exploiting implicit and empirical correlations between the petrophysical and elastic properties of rocks (Bachrach 2006; Grana & Della Rossa 2010; Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012). In this paper, we invert for geological facies (lithology and fluid types of rocks) from the spatial distribution of seismic attributes.

Bayesian inversion of seismic data for geological facies usually involves cluster analysis for discrimination between various rock and fluid types, and inference for posterior distributions is generally carried out using a stochastic approach based on Monte Carlo simulations (Doyen *et al.* 1989; Mukerji *et al.* 2001; Grana *et al.* 2012; Wang *et al.* 2016). Cluster analysis methods attempt to group various attributes of interest within some generally multidimensional feature- (or attribute-) space, based on some distance measure between each data point from the centres of identified groups of data which are commonly referred to as clusters. The probabilities that each data point belongs to each of the identified clusters can then be estimated by assigning a probability distribution to each of the clusters, for example, using a multivariate Gaussian distribution in a Gaussian mixture model (GMM) which is a term used to describe a sum of Gaussians, for example, Grana *et al.* (2016). A typical problem with such cluster analysis methods is that they estimate probabilities with high entropy (uncertainty) for those data

points that fall equidistant from cluster centres, or in cases where there is a significant overlap between different cluster distributions. For this reason, cluster analysis may only be used in Bayesian inversion to represent the information obtained directly from the data (referred to as the likelihood in Bayesian theory) about the probabilities that each data point belongs to a specific cluster. Prior information must then be introduced to produce more geologically reasonable estimates of the posterior distribution. Also, cluster analysis methods generally assume that the data points are independent and identically distributed (*i.i.d*) and therefore do not acknowledge spatial correlations present in the data.

GMMs for spatial data have been developed in which the spatial nature of data points influences the prior distribution, while estimating the parameters of a Gaussian mixture distribution using a Bayesian approach (Zhao *et al.* 2016). We present a method of Bayesian inversion for geological facies that jointly estimates the posterior marginal distributions of facies in each model cell, and the parameters of the Gaussian mixture distribution of seismic attributes. We introduced a spatial Gaussian mixture distribution in our method that estimates the likelihood of observing (or estimating) seismic attributes at a location given the geological facies in the neighbouring locations. We refer to such likelihoods as *quasi-localized likelihoods*. In our model, the parameters of the spatial Gaussian mixture distribution are spatially constrained through both the *a priori* distribution of facies, and their quasi-localized likelihoods.

In the following, we refer to the coordinate system that presents the seismic attributes with respect to their geographical locations and characterizes the spatial distribution of facies given by the prior information as the *model space*, and the coordinate system that is used to cross-plot seismic attributes and allows analysis of their mutual correlation irrespective of their spatial locations as the *attribute space*. Our method uses a variational form of EM algorithm which iteratively estimates the posterior marginal distributions of facies in the model space during the E-step, and updates the parameters of the GMM in the attribute space during the M-step.

A key requirement in seismic inversion is to capture the probabilistic spatial distribution of facies and heterogeneity of properties that is consistent with the true earth, and inject this information into the inversion process in the form of prior information. This can be achieved by using a Markov random field (MRF), a structured set of probabilistic relationships among various parameters of interest, under the assumption that given the facies in the neighbourhood of any location in the model, the facies at that location are conditionally independent of the facies in the rest of the model [the so called (first-order) *Markov assumption*]. In other words, facies at any location are assumed to have a direct dependence only on the facies in the neighbouring locations. Such a model is simple enough that it allows rigorous and efficient probabilistic inference by leveraging the conditional independence structure of the model, yet sophisticated enough to represent complex spatial patterns of facies in the form of prior information. The classes of problems that we consider here are those that can be represented with sufficient accuracy by this type of model.

Many different methods for probabilistic seismic inversion have been developed that make use of the Markov assumption. Larsen *et al.* (2006) used a 1-D Markov chain prior model for inversion of lithology–fluid classes along vertical profiles through a reservoir zone. Ulvmoen & Omre (2010) and Ulvmoen *et al.* (2010) used a *profile* MRF (see Eddy 1998) which induces correlations between 1-D Markov chains across two or three dimensions to model lateral coupling of lithology–fluid classes as commonly found in geological strata. Rimstad *et al.* (2012) used an MRF prior model to invert seismic AVO data for lithology/fluid classes, elastic properties and porosity. A common feature of these methods is that they perform inference from a joint posterior distribution across all or a subset of parameters in the model that is mathematically intractable: even if one could compute the joint posterior distribution, it would require prohibitive amounts of computer memory capacity to store it digitally. The joint posterior distribution in high dimensional models must therefore be explored through Monte Carlo-based sampling methods, or mathematically approximated, for example, by introducing conditional independence assumptions. The former approach generally uses McMC sampling which, as discussed earlier, is slow and suffers from convergence-related bias. The latter approach has seen recent development for models in which conditional independence may be introduced among variables of interest without sacrificing any significant probabilistic dependencies (Walker & Curtis 2014; Nawaz & Curtis 2017). Spatial models are excellent examples of such models since their random variables are expected to be strongly correlated only in a certain region around any given location. In fact, most McMC-based methods also introduce conditional independence assumptions in spatial models in order to improve their computational efficiency, yet such methods remain slow to converge. Hence, computationally efficient alternatives are desired in many large-scale problems.

Walker & Curtis (2014) developed a method for facies inversion that is based on an exact sampling approach as a more efficient alternative to McMC. The advantage of their method is that every sample is an independent sample of the posterior distribution (which is not true of McMC methods). However, their method requires computation and digital storage of the full joint distribution which may not be practical for real scale seismic volumes, or that it is approximated by the distribution across a model subspace. Nawaz & Curtis (2017) developed an alternative method for Bayesian inversion of facies based on a 2-D hidden Markov model (2-D-HMM) that entirely obviates the need for a sampling-based inference approach. Their method computes marginal posterior distributions of facies analytically and is therefore computationally more efficient than previous methods, and requires far less computer memory. They also presented a *copula* function-based method that performs exact sampling from the marginal posterior distributions in order to circumvent the need for digital storage of the full joint posterior distribution. However, a key assumption in all of these methods is that the likelihoods are localized; in other words, given the facies in any model cell, the seismic attributes in that cell are assumed to be conditionally independent of the facies and attributes in the rest of the model. This assumption was implicit or explicit in most of the previous research described above (e.g. Larsen *et al.* 2006; Ulvmoen & Omre 2010; Ulvmoen *et al.* 2010; Walker & Curtis 2014; Nawaz & Curtis 2017). Geophysical data derived from seismic imaging can be strongly correlated spatially due to mixing of information across different spatial locations (sometimes referred to as blurring or smearing). This occurs due to errors in the seismic velocity model which cause mislocation of seismic attributes, Fresnel zone smearing, migration errors

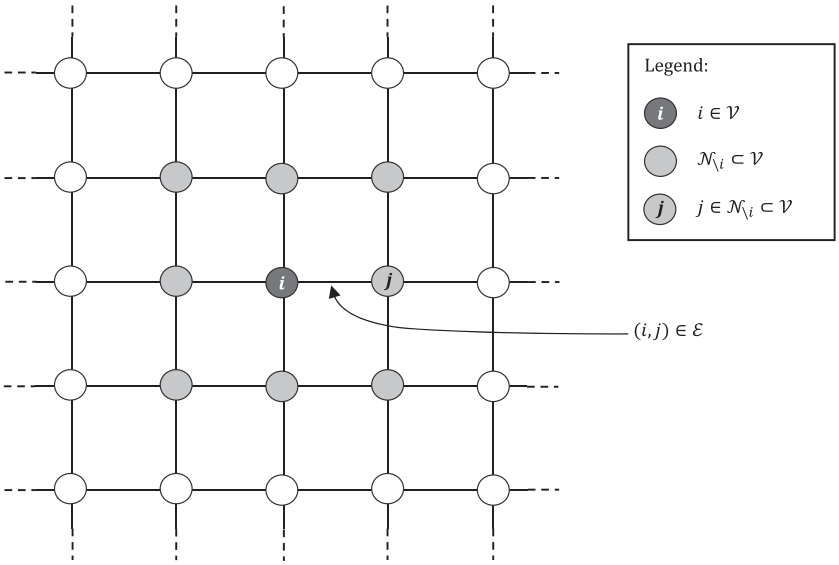


Figure 1. A graphical representation of a Markov random field (MRF) where circles represent vertices $\mathcal{V}$ and the connecting lines represent edges $\mathcal{E}$ in the graph. The central dark-grey circle represents a vertex $\boldsymbol{i}$ in consideration and the light-grey circles around it form the Markov blanket $\mathcal{N}_{\setminus \boldsymbol{i}}$ of $\boldsymbol{i}$. The dotted lines show possible extension of edges and the graph that is not shown in the figure. This graph contains only pairwise cliques, that is, cliques that contain just two vertices that share an edge (as used in this paper). A more complex MRF may also involve diagonal edges, thus containing cliques of size 3 or more.

due to the limited apertures of seismic arrays and a variety of other factors. There is therefore an important need to find methods that relax this localized likelihoods assumption.

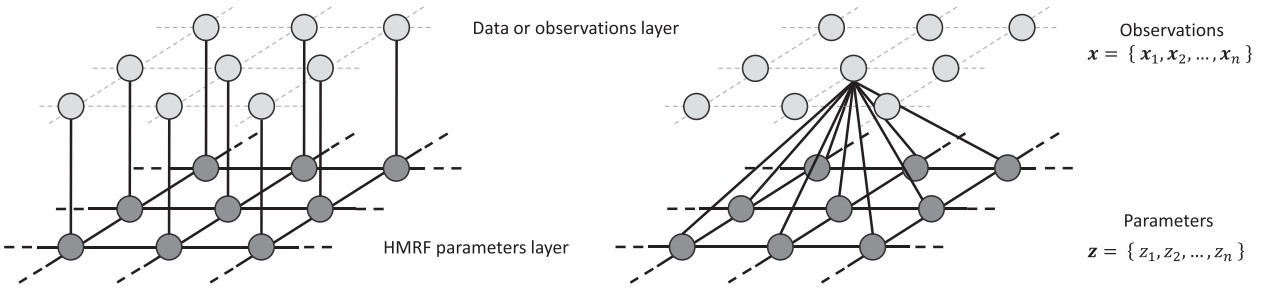
In this paper, we present a VBI method as a more efficient alternative to McMC-based methods to solve spatial inverse problems, and which also relaxes the localized likelihoods assumption. Section 4.2 provides a description of how the localized likelihoods assumption is relaxed. Examples of previous research on Bayesian inversion methods in which likelihoods are not (fully) localized include Lindberg & Omre (2014, 2015), Grana *et al.* (2017) and Lindberg *et al.* (2015): they used a convolved two-level, 1-D hidden Markov model for inversion of categorical variables (such as lithology–fluid classes) represented as the bottom hidden-layer of the model, continuous system response variables (such as reflection coefficients) represented as the middle hidden-layer and the measured convolved data represented in the observation layer. The advantages of our new approach are that it is multidimensional and that it allows for joint estimation of model parameters and the spatial distribution of geological facies. As a result, the model parameters are chosen such that they provide best estimates of the spatial distribution of geological facies that are consistent with the prior information and which are also constrained by the seismic attribute data.

Before proceeding, we define the notation used in this paper. We use a linear index denoted by lower case letters such as i and j to define the locations (or cells) in our model, or equivalently vertices in the underlying graph. Sets are represented with italic, regular (non-boldface) capital (English or Greek) letters, for example, $\mathcal{V}$ and $\mathcal{G}$. We use the term vector for a 1-D row or column matrix. We use boldface font with lower case (English or Greek) letters for vectors, for example, $\boldsymbol{r}$ or $\boldsymbol{\beta}$, and upper case letters for matrices, for example, $\boldsymbol{R}$. A subscript used with such letters connotes meanings indicated in the text. The identity matrix is represented as $\boldsymbol{I}$. A superscript T stands for transpose of a vector or matrix. Bracketed superscripts indicate an estimate of a quantity at the iteration number specified in brackets during the course of an iterative update, for example, $\theta^{(t)}$ represents an estimate of some quantity θ after t iterations of an iterative algorithm. A hat, or caret, over a parameter (or random variable) denotes its estimator, for example, $\hat{\theta}$ represents an estimator of θ. Other commonly used statistical and set theoretic notations include: '$\sim$' for a random variable which reads '*is distributed as*', '$\setminus$' for set difference, '$\cup$' for the union of two sets, '$\cap$' for the intersection of two sets, and '$|\cdot|$' for cardinality (or number of elements) of a set.

3 MODEL

We use a so-called hidden Markov random field (HMRF) as the underlying graph behind our method. This defines an MRF over latent (or unobserved) variables. An MRF is an undirected graphical model $\mathbb{G}(\mathcal{V}, \mathcal{E})$ which defines the topology of some physical space (Fig. 1), where $\mathcal{V} = \{1, \ldots, n\}$ is a set of vertices (also called nodes), and $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V} \wedge i \neq j\}$ is the set of undirected edges (or connections between vertices) in the graph where $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. The edges in an undirected graphical model have no orientation and represent unordered pairs, that is, an edge $(i, j) \in \mathcal{E}$ is identical to the edge $(j, i) \in \mathcal{E}$. A path in the graph is defined by an ordered sequence of vertices in $\mathcal{V}$ such that any two consecutive vertices in this sequence share an edge from $\mathcal{E}$. For any disjoint sets $A, B, C \subset \mathcal{V}$, set C is said to separate A and B if every path from any vertex in A to any vertex in B passes through C.

Every vertex $i \in \mathcal{V}$ is associated with a set $\mathcal{N}_{\setminus i} \subset \mathcal{V} \setminus \{i\}$ of neighbouring vertices; these share an edge in $\mathcal{E}$ from the vertex $i \in \mathcal{V}$, and are referred to as the *neighbourhood* of i. So $j \in \mathcal{N}_{\setminus i}$ if and only if $(i, j) \in \mathcal{E}$. By definition, the neighbouring relationship must satisfy two



(a) A typical HMRF Model

(b) HMRF Model as used in this paper

Figure 2. A graphical depiction of a hidden Markov random field (HMRF) with two layers where the upper layer consists of observed variables $\boldsymbol{x}$ represented by light-grey circles and the lower layer consists of latent variables $\boldsymbol{z}$ represented by dark-grey circles. The solid black lines represent the edges between different latent vertices and between latent and observed vertices in the model whereas dotted grey lines in the upper layer are only guidelines included for clarity in order to portray the relative positions of observed vertices in the model grid. The grid is shown in two dimensions with a 3×3 square matrix of vertices for illustration purpose only. The actual grid may be higher dimensional and much larger in size. (a) A typical HMRF model where an observed variable $\boldsymbol{x_i}$ at a location $\boldsymbol{i}$ is directly connected only to the latent variable $\boldsymbol{z_i}$ at that location. (b) A variant of the HMRF model used in this paper where each observed variable $\boldsymbol{x_i}$ at location $\boldsymbol{i}$ is connected to all latent variables $\boldsymbol{z}_{\mathcal{N}_i}$ where $\mathcal{N}_i$ refers to the neighbourhood of $\boldsymbol{i}$ (with $\boldsymbol{i}$ inclusive). The edges between latent and observed variables are shown only for one observed variable for clarity, but all observed vertices in the model are assumed to be connected to latent variables in a similar fashion.

properties: a vertex cannot be a neighbour of itself, that is, $i \notin \mathcal{N}_{\setminus i}$ (as is emphasized by the subscript '$\setminus i$'), and the neighbouring relationship is commutative, that is, $i \in \mathcal{N}_{\setminus j} \Rightarrow j \in \mathcal{N}_{\setminus i}$. The neighbourhood $\mathcal{N}_{\setminus i}$ of a vertex i in an MRF is also sometimes referred to as the *Markov blanket* of i. We often need to consider the set $\mathcal{N}_{\setminus i} \cup \{i\}$ which is used in the rest of this document, so in order to reduce the notational clutter we denote it with $\mathcal{N}_i$, and also refer to it as the neighbourhood of i while the subscript clearly indicates whether the vertex i is included in the set or not. A *neighbourhood system* $\mathcal{N}$ in graph $\mathbb{G}(\mathcal{V}, \mathcal{E})$ is defined as

$$\mathcal{N} = \left\{\mathcal{N}_{\setminus i} \subset \mathcal{V} \setminus \{i\} : \forall i \in \mathcal{V}\right\}. \tag{1}$$

A *clique* $\mathfrak{c} \subseteq \mathcal{V}$ of a graph is any subset of its vertices which are fully connected, that is, every vertex in $\mathfrak{c}$ shares an edge from $\mathcal{E}$ with every other vertex in $\mathfrak{c}$. The *maximal cliques* of a graph are defined to be its cliques that are as large as they can be given edges $\mathcal{E}$. So $\mathfrak{c}$ is a maximal clique of $\mathbb{G}$ if it fails to remain a clique when any additional vertex from $\mathcal{V} \setminus \mathfrak{c}$ is added to $\mathfrak{c}$. The set of all of the cliques in $\mathbb{G}$ is represented by $\mathcal{C}$ and the set of all of the maximal cliques in $\mathbb{G}$ is represented by $\mathcal{C}_{\max}$.

Each vertex in our graphical model represents either an observed or a latent (unobserved) random variable. An HMRF graphical model may be visualized as consisting of two layers, where the upper layer contains the observed variables $\boldsymbol{x}$ and the lower layer contains the latent variables $\boldsymbol{z}$ (Fig. 2). In our model definition, each observed variable $\boldsymbol{x}_i$ in the upper layer represents seismic attributes that can be computed from the latent variables $\boldsymbol{z}_{\mathcal{N}_i}$ in the lower layer which represent the geological facies (litho-fluid type) within the neighbourhood of a location i. Since the true spatial distribution of geological facies is unknown, $\boldsymbol{z}$ is referred to as latent variables whereas the seismic attributes are referred to as observed variables. Each z_i may take any of the k values from a discrete set $\mathcal{G}$ of pre-defined geological facies, where $k = |\mathcal{G}|$. The observed variables $\boldsymbol{x}_i$ are assumed to have continuous values and are denoted with a boldface font because these may be vectors containing multiple seismic attributes. While the seismic attributes are actually inferred from seismic data, they are referred to as 'observed data' or 'observations' henceforth in order to explicitly distinguish them from the latent variables such as geological facies.

The geological facies and seismic attributes at each of the $n = |\mathcal{V}|$ locations in the model are represented as $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^T$ and $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$, respectively. Then $\boldsymbol{z}$ forms a MRF over the graph $\mathbb{G}$ if it satisfies two properties: the *positivity property* according to which the joint probability of the random variables $\boldsymbol{z}$ is strictly positive, that is, $\mathcal{P}(\boldsymbol{z}) > 0$, for all possible configurations of $\boldsymbol{z}$, and the *Markovian property* which requires that given the facies in the neighbourhood of a vertex i, the facies z_i at i become conditionally independent of the facies in the rest of the model, that is, $\mathcal{P}(z_i \mid \boldsymbol{z}_{\setminus i}) = \mathcal{P}(z_i \mid \boldsymbol{z}_{\mathcal{N}_{\setminus i}})$, where $\boldsymbol{z}_{\setminus i} \equiv \boldsymbol{z} \setminus \{z_i\}$ and $\boldsymbol{z}_{\mathcal{N}_{\setminus i}} \equiv \{z_j : j \in \mathcal{N}_{\setminus i}\}$. The Markovian property implies that for any disjoint subsets $A, B, C \subset \mathcal{V}$ such that C separates A from B in $\mathbb{G}$, we have $\boldsymbol{z}_A$ that is conditionally independent of $\boldsymbol{z}_B$ given $\boldsymbol{z}_C$, that is, $(\boldsymbol{z}_A \perp \boldsymbol{z}_B \mid \boldsymbol{z}_C)$, where $\boldsymbol{z}_A = \{z_i : i \in A\}$ and similarly defining $\boldsymbol{z}_B$ and $\boldsymbol{z}_C$. Therefore, according to the Markovian property, any two latent vertices in an MRF are conditionally independent if all paths between them pass through the observed vertices. Accordingly in our model definition, given the geological facies at the neighbouring locations of any vertex, the knowledge of facies in the rest of the model has no influence on the facies at the vertex in question.

The latter property is achieved by assuming that the observed variables in our model are mutually conditionally independent given the latent variables. A typical MRF assumes that given the latent variable z_i at a location i, the corresponding observed variable $\boldsymbol{x}_i$ at that location is also conditionally independent of all of the other latent variables $\boldsymbol{z}_{\setminus i}$ in the model, which is commonly referred to as the 'localized likelihoods' assumption (Fig. 2a). We define a variant of a typical HMRF model where each observed variable $\boldsymbol{x}_i$ at a location i is directly connected to the latent variables $\boldsymbol{z}_{\mathcal{N}_i}$ in the neighbourhood $\mathcal{N}_i$ of i (with i inclusive). This means that given the latent variable z_i at i, the corresponding observed variable $\boldsymbol{x}_i$ is not assumed to be conditionally independent of other latent variables $\boldsymbol{z}_{\setminus i}$ in the model. This relaxes

the assumption of localized likelihoods (Fig. 2b). We develop this concept further in Section 4.2 to introduce the concept of quasi-localized likelihoods—an assumption which is a less stringent than the localized likelihoods assumption.

The MRF model has its origin in statistical physics where it was introduced to model the energy states of a large number of mutually interacting particles which exhibit a stochastic behaviour, but where their mutual interactions obey some natural rules. For example, a natural system commonly prefers lower energy states and it continuously updates the local energy states of the particles that compose the system until the system attains the lowest energy state. Local energy states of particles depend only on their interactions with neighbouring particles. Such a behaviour can be modelled with an MRF called an *Ising* or *Pott's model* as this provides a mathematical specification of any joint distribution over a large number of particles by exploiting the conditional independence among most of the (non-neighbouring) particles. We use our MRF model to parameterize the prior information on geological facies as embodied within a training image, since heterogeneity typically observed in geology may be assumed to be globally random while the facies in neighbouring locations are more likely to be similar than those in the distant cells.

3.1 Gibbs distribution

A mathematically tractable specification of a joint probability distribution over an MRF is provided by the Hammersley–Clifford theorem (Hammersley & Clifford 1971) proved by Besag (1974). It states that any joint distribution over an MRF may be expressed as a *Gibbs distribution* which takes the form

$$\mathcal{P}\ (\boldsymbol{z}) = \frac{1}{\mathcal{Z}} \exp\left\{-\frac{1}{T}\sum_{\mathfrak{c}\in\mathcal{C}} E_{\mathfrak{c}}\,(\boldsymbol{z}_{\mathfrak{c}})\right\}, \tag{2}$$

where $\mathcal{C}$ represents the set of cliques in the graph, $E_{\mathfrak{c}}(\boldsymbol{z}_{\mathfrak{c}})$ represents the *energy function* of the local configurations $\boldsymbol{z}_{\mathfrak{c}} \subseteq \boldsymbol{z}$ of each clique $\mathfrak{c}$ in the graph $\mathbb{G}$ such that low energy states correspond to high probability configurations, T is a parameter called *temperature*, and $\mathcal{Z}$ is a constant known as the *partition function* that ensures normalization of the joint distribution to be a valid probability function and is given by the sum of the numerator over all possible configurations of $\boldsymbol{z}$, that is

$$\mathcal{Z} = \sum_{\boldsymbol{z}} \exp\left\{-\frac{1}{T}\sum_{\mathfrak{c}\in\mathcal{C}} E_{\mathfrak{c}}\,(\boldsymbol{z}_{\mathfrak{c}})\right\}. \tag{3}$$

In an MRF, the energy states of a system are conventionally expressed in the form of potential functions over cliques, called *clique potentials* $\psi_{\mathfrak{c}}(\boldsymbol{z}_{\mathfrak{c}})$, given by

$$\psi_{\mathfrak{c}}\ (\boldsymbol{z}_{\mathfrak{c}}) = \exp\left\{-\frac{E_{\mathfrak{c}}\,(\boldsymbol{z}_{\mathfrak{c}})}{T}\right\} \tag{4}$$

such that the joint distribution over an MRF may be expressed as a product of clique potentials

$$\mathcal{P}\ (\boldsymbol{z}) = \frac{1}{\mathcal{Z}} \prod_{\mathfrak{c}\in\mathcal{C}} \psi_{\mathfrak{c}}\,(\boldsymbol{z}_{\mathfrak{c}}). \tag{5}$$

The clique potentials $\psi_{\mathfrak{c}}(\boldsymbol{z}_{\mathfrak{c}})$ are real-valued positive functions of local configurations $\boldsymbol{z}_{\mathfrak{c}} \subseteq \boldsymbol{z}$ for each clique $\mathfrak{c}$ in the graph $\mathbb{G}$. The clique potentials may either be defined over pairwise cliques, that is edges from $\mathcal{E}$ in the graph. In this case, the model is termed as a *pairwise MRF*. The pairwise clique potentials are functions of two neighbouring variables expressed as $\psi_{ij}(z_i, z_j)$ such that $(i, j) \in \mathcal{E}$, which model the affinity or relative compatibility of two neighbouring random variables in a pairwise MRF. The pairwise clique potentials are also referred to as *edge potentials* for obvious reasons, and are estimated using some form of parameter learning. It is important to note that an MRF that defines clique potentials over higher order cliques in the graph is more expressive than a pairwise MRF (Koller & Friedman 2009) in the sense that higher order cliques can capture more complex conditional independence relationships among vertices in an MRF. Higher order cliques involve probabilistic dependence between more than two locations (vertices in a graph) at a time, and therefore can model more complex patterns of facies distribution in space (e.g. map-view of meandering channels in a deltaic environment). This concept has been used, for example, in the development of multiple-point statistics (MPS) methods in geostatistics (Mariethoz & Caers 2014). Nevertheless, here we use a pairwise MRF to model dependencies within the hidden variables $\boldsymbol{z}$ (i.e. geological facies at different locations) for simplicity in the derivations that follow.

4 BAYESIAN INVERSION IN A MARKOV RANDOM FIELD

We perform inference on latent variables $\boldsymbol{z}$, which represent geological facies, conditioned on the observed variables $\boldsymbol{x}$ which represent seismic attributes, by defining their joint distribution $\mathcal{P}(\boldsymbol{x},\ \boldsymbol{z})$. The joint distribution can be obtained from the product of data likelihood $\mathcal{P}(\boldsymbol{x}|\boldsymbol{z})$, which encodes how likely it is to observe $\boldsymbol{x}$ given any particular configuration of latent variables $\boldsymbol{z}$, and the prior model distribution $\mathcal{P}(\boldsymbol{z})$. Their product is then proportional to the joint distribution $\mathcal{P}(\boldsymbol{x},\ \boldsymbol{z})$, and the posterior distribution $\mathcal{P}\ (\boldsymbol{z}|\boldsymbol{x}) = \mathcal{P}(\boldsymbol{x},\ \boldsymbol{z})/\mathcal{P}(\boldsymbol{x})$. Thus, computing the posterior distribution requires the normalization constant, $\mathcal{P}(\boldsymbol{x})$, to be known.

Contrary to the localized likelihoods assumption, herein we assume that given the geological facies at any location and its neighbouring locations, the seismic attributes observed at that location are conditionally independent of attributes at other locations, and of the geological facies in the rest of the model. We refer to this assumption as quasi-localized likelihoods, as in this case the attributes at any location may have direct probabilistic dependence on *the* facies in neighbouring locations, yet such likelihoods may not be regarded as fully non-localized unless the neighbourhood of every location is defined to be the entire set of other locations in $\mathcal{V}$ (i.e. $\mathcal{V}$ is fully connected). Thus, fully non-localized likelihoods are a subclass of quasi-localized likelihoods, but the methods introduced herein are not efficient for that case so we do not consider it further. Localized likelihoods are also a subclass of quasi-localized likelihoods, and thus our assumed model is significantly more general than the norm. This concept is further discussed in Section 4.2.

4.1 Prior

A priori knowledge about the spatial distribution of facies is often available based on previously acquired data and from prior experience of geoscientists on the local and regional geology. Such prior information may be quantitatively embodied within geological process modelling, and may be presented in the form of training images. We aim to reconstruct the spatial distribution of facies in a Bayesian framework by combining the data likelihoods with this *a priori* information. How best to describe and incorporate geological prior information in Bayesian inverse problems is still under research. Depending on the inversion algorithm and on the type and complexity of the *a priori* information, different methods exist that mathematically transform *a priori* information into probability distributions. However, when prior information on facies distribution only involves correlations between similar facies in neighbouring locations, the strength of such correlations may be encoded in parameters which depend on the relative locations of the neighbours.

A convenient way to embody *a priori* knowledge about the spatial distributions of facies is through a training image. A training image is a conceptual depiction of typical patterns of geological features that are expected to exist in the subsurface based on the subjective opinion of geoscientists, or on other objective geological measurements of facies distributions. Mathematically, a training image embodies statistics of facies heterogeneity over a lattice of model cells (or vertices). These statistics may then be extracted during the inversion process as and when desired. Thus, a training image also serves as a compact embodiment of joint and conditional probability distributions over spatial variables which would otherwise require a comparatively large amount of computer memory for their digital storage. Another conceptual advantage of using a training image is that it restricts the expected spatial patterns of facies to a limited set of geologically plausible patterns as depicted in the image. It is, however, important to note that a training image only provides contextual information about the local geology, and not location-specific information as is supplied by the data in the form of likelihoods.

We represent prior information on the spatial distributions of facies as a joint distribution $\mathcal{P}(\boldsymbol{z})$ of facies over a pairwise MRF given by

$$\mathcal{P}\ (\boldsymbol{z}) = \frac{1}{\mathcal{Z}} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}\left(z_i, z_j\right). \tag{6}$$

The pairwise potential functions $\psi_{ij}(z_i, z_j)$ in the above equation are estimated by scanning the training image and building histograms for various combinations of facies z_i and z_j over pixels with offset distance and direction depending on the neighbourhood structure. For example, a histogram is built by counting the occurrence of any two facies in laterally or vertically adjacent pixels in the training image. These counts are then normalized by the total number of possible combinations of facies within the same configuration of pixels across the training image to give prior probabilities. This assigns zero probability to configurations of facies that are geologically implausible, such as brine directly over gas. The prior probability of occurrence of facies z_i at a location i given the facies in its neighbourhood $\mathcal{N}_{\setminus i}$ is therefore given by

$$\mathcal{P}\left(z_i | \boldsymbol{z}_{\mathcal{N}_{\setminus i}}\right) \propto \prod_{j\in\mathcal{N}_{\setminus i}} \psi_{ij}\left(z_i, z_j\right). \tag{7}$$

4.2 Likelihood

Rock properties and the spatial distribution of facies are generally obtained from inversion of seismic data constrained by borehole data and prior geological information about the reservoir. There is always residual uncertainty in the estimation of geological facies from the observed data at a given location. The uncertainty is either due to the presence of noise in the data, or due to intrinsic uncertainty in the relationship between facies and the observed seismic attributes, or both. We assume that multiple seismic attributes are associated to each model cell, which are noisy due to imperfect seismic acquisition and processing, and are blurred by the band-limited nature of seismic waves and the limited aperture of seismic acquisition systems. Such spatial blurring of seismic attributes results in significant superposition of attribute values that are otherwise attributable to different facies. Consequently, the seismic attribute response of facies at each location is corrupted by the attribute response of facies at neighbouring locations. This implies that the seismic data or attributes computed at a given location must be related to the facies at that location according to some probability distribution that also involves facies in the neighbouring locations. Thus, data are quasi-localized.

The probability of observing specific (usually measured) data given a set of model parameters is called the data likelihood. The likelihood may be interpreted to be a function of model parameters given the observed data. The data in our model represent the seismic attributes, and

the model parameters represent the spatial distribution of geological facies. Much of the previous research in inversion of seismic attributes for geological facies uses the localized likelihoods assumption (e.g. Larsen *et al.* 2006; Ulvmoen & Omre 2010; Ulvmoen *et al.* 2010; Walker & Curtis 2014; Nawaz & Curtis 2017), but such an assumption is not valid for seismic attributes as discussed above: any residual non-localized effects that are not fully accounted for during seismic data processing (Fresnel zone smearing, seismic velocity errors causing mislocation of attributes, migration errors due to limited apertures of seismic arrays, etc.) propagate into the computed seismic attributes. The region of the model around the location of observation used to estimate likelihoods must therefore be expanded from one vertex (or cell) to include a certain neighbourhood around that vertex. A more robust inversion method is then required that acknowledges the non-localized nature of seismic data and incorporates spatial correlations present in the data to capture the true spatial distribution of facies.

In this paper, the spatial correlations of facies may be reconstructed from the non-localized facies-attributes relations, and in addition from correlations in the prior information. We estimate the likelihood of seismic attributes at a location i given the facies in the neighbourhood $\mathcal{N}_i$ of that location in order to account for the blurring effect of the band-limited seismic data. We refer to this assumption as *quasi-localized likelihoods* since the dependence of seismic attributes on facies in neighbouring locations may not be regarded as fully non-localized likelihoods (unless the neighbourhood spans the entire domain).

All facies classification methods assume that the variation in rock properties within a facies is smaller than variations between different facies. Any ambiguity in classification due to overlap of rock properties among multiple facies might be able to be resolved by introducing the spatial context of each data point. This can be done by conditioning each data point on its spatial neighbours based on the information contained in spatial priors or quasi-localized likelihoods. For example, if spatial neighbours of a point are more likely to belong to a particular facies, then that point is generally more likely to belong to the same facies based on typical spatial priors and quasi-localized likelihoods. In this respect, quasi-localized likelihoods reduce the entropy of (the degree of variation in) classification compared to the localized likelihoods which offer no spatial context for the classification task.

Even though the likelihoods are not assumed to be localized, we still assume that the seismic attributes at a location are conditionally independent of *attributes* observed at all other locations given the facies model. This implies that any spatial correlations in the observations of seismic attributes are assumed to be a direct consequence of spatial distribution of facies within the neighbourhood structure, and not due to correlations in the data measurements that are independent of the geology (for example, due to correlated random or systematic noise).

We consider a set $\mathcal{G} = \{1, \ldots, K\}$ of discrete variables representing geological facies which refers to well-defined and distinct rock and fluid types. Each facies $k \in \mathcal{G}$ is defined in terms of its expected attributes $\boldsymbol{\mu}_k$ and the corresponding covariance matrix $\boldsymbol{\Sigma}_k$ that represents intrafacies variations. Let $\boldsymbol{R}_{\mathcal{N}_i} = (\boldsymbol{r}_j : j \in \mathcal{N}_i)$ be a $p \times q$ matrix of expected local facies responses for p seismic attributes at each of the q locations in the neighbourhood $j \in \mathcal{N}_i$ such that $|\mathcal{N}_i| = q$ is fixed and is independent of location i in the graph, and $\boldsymbol{r}_j$ represents local facies responses at each location given by some mapping from the latent variable z_j to the domain of observed variables $\boldsymbol{x}$. To make this more concrete, define $\boldsymbol{r}_j$ as the expectation of a set of superposed Gaussian distributions $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k \in \mathcal{G}$ for each facies, weighted by some estimate of the posterior marginal distribution $\hat{\mathcal{P}}_j(z_j)$ of the facies at each location j:

$$\boldsymbol{r}_j = \mathbb{E}\left(\sum_{k\in\mathcal{G}} \hat{\mathcal{P}}_j\left(z_j = k\right) N\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right) = \sum_{k\in\mathcal{G}} \hat{\mathcal{P}}_j\left(z_j = k\right)\boldsymbol{\mu}_k. \tag{8}$$

The seismic attributes $\boldsymbol{x}_i$ computed at a location i are assumed to be a weighted linear combination of facies responses $\boldsymbol{R}_{\mathcal{N}_i}$ in neighbourhood $\mathcal{N}_i$ such that

$$\boldsymbol{x}_i = \sum_{j\in\mathcal{N}_i} \beta_j \boldsymbol{r}_j + \boldsymbol{\varepsilon}_i = \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{9}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of p seismic attributes, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ vector of errors which are assumed to be jointly distributed according to a Normal distribution $N(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. The seismic attributes are assumed to have been pre-standardized to have unit variance while keeping the covariances intact, so that the definition of regressors $\boldsymbol{R}_{\mathcal{N}_i}$ allows us to interpret $\boldsymbol{\beta}$ as a weighting kernel over all of the attributes observed at multiple locations in the neighbourhood of location i (Fig. 3). The attributes can be de-standardized later to their original means and variances for display and interpretation purposes. Now define the set of model parameters as $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, $k \in \mathcal{G}$. So, given the expected facies responses $\boldsymbol{R}_{\mathcal{N}_i}$ in the neighbourhood of i, the seismic attributes $\boldsymbol{x}_i$ are normally distributed with mean $\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}$. The quasi-localized likelihood of seismic attributes $\boldsymbol{x}_i$ computed at i given the geological facies $\boldsymbol{z}_{\mathcal{N}_i} \equiv \{z_j : j \in \mathcal{N}_i\} \subseteq \boldsymbol{z}$ in the neighbourhood $\mathcal{N}_i$ of location i is therefore given by

$$\mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta\right) = \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}, \Theta\right) = N\left(\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}\right). \tag{10}$$

We can show that the likelihood of observing seismic attributes $\boldsymbol{x}_i$ at a location i given the geological facies $\boldsymbol{z}_{\mathcal{N}_{\setminus i}}$ in the neighbourhood $\mathcal{N}_{\setminus i}$ of i and the model parameters Θ is given by

$$\begin{aligned}\mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_{\setminus i}}, \Theta\right) &= \sum_{z_i} \mathcal{P}\left(\boldsymbol{x}_i, z_i|\boldsymbol{z}_{\mathcal{N}_{\setminus i}}, \Theta\right) \\ &= \sum_{z_i} \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta\right)\ \mathcal{P}\left(z_i|\boldsymbol{z}_{\mathcal{N}_{\setminus i}}, \Theta\right),\end{aligned} \tag{11}$$

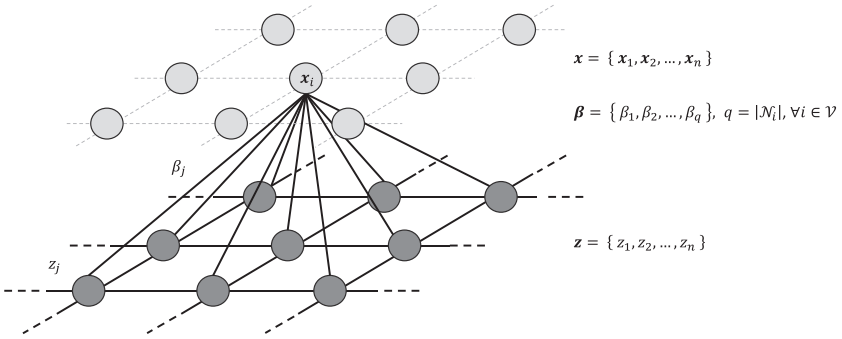


Figure 3. A graphical depiction of a hidden Markov random field (HMRF) as in Fig. 2(b), where each edge between an observed variable $\boldsymbol{x_i}$ at a location $\boldsymbol{i}$ and the latent variables $\boldsymbol{z_j}$, $\boldsymbol{j} \in \mathcal{N}_i$ within the neighbourhood $\mathcal{N}_i$ of $\boldsymbol{i}$ is associated with a weight parameter $\boldsymbol{\beta_j}$ which may be interpreted as the strength of the connection between the two variables in the definition of quasi-localized likelihoods.

which represents a *spatial* Gaussian mixture distribution with components given by the quasi-localized likelihoods $\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta)$ in eq. (10), each of which is scaled with the *spatial priors* $\mathcal{P}(z_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta)$ given by the MRF prior model—eq. (7).

Under the assumption of conditional independence of seismic attributes $\boldsymbol{x}$ given the facies $\boldsymbol{z}$ and the model parameters Θ, the likelihood $\mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \Theta)$ of observed seismic attributes $\boldsymbol{x}$ given a particular facies model $\boldsymbol{z}$ is given by

$$\mathcal{P}\ (\boldsymbol{x}|\boldsymbol{z}, \Theta) = \prod_i \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta\right) = \prod_i N\left(\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta},\ \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_i}\right). \tag{12}$$

We can write eq. (9) for all of the n cells in the model as

$$\boldsymbol{x} = \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ is a $np \times 1$ vector of p seismic attributes in each of the n cells, $\boldsymbol{R} = (\boldsymbol{R}_{\mathcal{N}_1}, \boldsymbol{R}_{\mathcal{N}_2}, \ldots, \boldsymbol{R}_{\mathcal{N}_n})^T$ is a $np \times q$ matrix of facies responses at q neighbours of each of the n cells, and $\boldsymbol{\varepsilon}$ is a $np \times 1$ vector of errors which are assumed to be uncorrelated with the covariates $\boldsymbol{R}$ and are jointly distributed according to a Normal distribution $N(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. Therefore, given the facies responses $\boldsymbol{R}$, the seismic attributes are normally distributed with mean $\boldsymbol{R\beta}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, that is, $\boldsymbol{x}|\boldsymbol{z} \sim N(\boldsymbol{R\beta},\ \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$. Thus, the log-likelihood $\mathcal{L}(\Theta; \boldsymbol{x}|\boldsymbol{z})$ of seismic attributes $\boldsymbol{x}$ given the facies $\boldsymbol{z}$ as a function of model parameters Θ may be written as

$$\begin{aligned}
\mathcal{L}(\Theta; \boldsymbol{x}|\boldsymbol{z}) &\equiv \log \mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \Theta) \\
&= \sum_i \log \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta\right) \\
&\qquad\qquad \text{[using the conditional independence assumption over } \boldsymbol{x}\text{]} \\
&= \sum_i \log \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}, \Theta\right) \\
&\qquad\qquad \text{[using equation (12)]} \\
&= \sum_i \log \left\{(2\pi)^{-n/2}|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}|^{-1/2} \exp\left(\frac{-1}{2}\left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)\right)\right\} \\
&\qquad\qquad \text{[using definition of a Gaussian distribution]} \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}| - \frac{1}{2}\sum_i \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right) \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{R\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}(\boldsymbol{x} - \boldsymbol{R\beta}).
\end{aligned} \tag{14}$$

4.3 Posterior distribution

In a so called *generative model*, the seismic attributes $\boldsymbol{x}$ are assumed to have been generated by the unobserved facies $\boldsymbol{z}$ according to a probability distribution $\mathcal{P}(\boldsymbol{x}|\boldsymbol{z},\ \Theta)$, where Θ is the set of parameters that defines the probability distribution and models the dependencies between the facies and the observed seismic attributes—eqs (11) and (14). The posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ of facies $\boldsymbol{z}$ given the seismic attributes $\boldsymbol{x}$ and parameters Θ is then given in terms of their joint distribution using *Bayes' theorem*

$$\mathcal{P}\ (\boldsymbol{z}|\boldsymbol{x}, \Theta) = \frac{\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)}{\mathcal{P}(\boldsymbol{x}|\Theta)} = \frac{\mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \Theta)\mathcal{P}(\boldsymbol{z})}{\mathcal{P}(\boldsymbol{x}|\Theta)}, \tag{15}$$

where the denominator represents the marginal likelihood of observed variables $\boldsymbol{x}$ given the model parameters Θ and plays the same normalizing role as the partition function $\mathcal{Z}$ in eq. (2). Given the conditional independence assumption, the quasi-localized likelihood

$\mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \Theta)$ given by eq. (12) can be written as

$$\mathcal{P}\ (\boldsymbol{x}|\boldsymbol{z}, \Theta) = \prod_{i\in\mathcal{V}} \mathcal{P}\left(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta\right) = \prod_{i\in\mathcal{V}} \varphi_i\left(\boldsymbol{x}_i, \boldsymbol{z}_{\mathcal{N}_i}|\Theta\right), \tag{16}$$

where $\varphi_i\ (\boldsymbol{x}_i, \boldsymbol{z}_{\mathcal{N}_i}|\Theta) = \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta)$ represents a potential function of $\boldsymbol{x}_i$ and $\boldsymbol{z}_{\mathcal{N}_i}$ that is called the *vertex potential* in an MRF model. It represents the physical dependency between observables and facies in the model, including errors in the data. With the prior $\mathcal{P}(\boldsymbol{z})$ given by eq. (6) and the likelihood $\mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \Theta)$ given by eq. (16), the posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ may be written as

$$\mathcal{P}\ (\boldsymbol{z}|\boldsymbol{x}, \Theta) = \frac{\mathcal{P}\,(\boldsymbol{x}, \boldsymbol{z}|\Theta)}{\mathcal{P}\,(\boldsymbol{x}|\Theta)} = \frac{1}{\mathcal{Z}'} \prod_{i\in\mathcal{V}} \varphi_i\left(\boldsymbol{x}_i, \boldsymbol{z}_{\mathcal{N}_i}\right) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}\left(z_i, z_j\right), \tag{17}$$

where constant $\mathcal{P}(\boldsymbol{x}|\Theta)$ has been absorbed in $\mathcal{Z}'$. This demonstrates that although we only assumed that the prior distribut ion $\mathcal{P}(\boldsymbol{z})$ on the latent variables $\boldsymbol{z}$ is an MRF, the posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ and the joint distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ then also turn out to be MRFs as a consequence of the conditional independence assumption on the observed variables $\boldsymbol{x}$. Note that without such an assumption the joint distribution would not be tractable, making inference impossible for models of practical interest. The above formulation is quintessentially the *generative approach* as it models the posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ via the joint distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)$, as opposed to the *discriminative approach* that directly models the posterior distribution.

Vertex potentials $\varphi_i(\boldsymbol{x}_i, \boldsymbol{z}_{\mathcal{N}_i}|\Theta)$ are estimated from the data using a rock physics model of the relationship between facies and seismic attributes. The edge potentials $\psi_{ij}(z_i, z_j)$, on the other hand, are estimated only from the prior information expressed in the form of a training image. This means that any spatial correlations in the data are only used in the reconstruction of the spatial distribution of facies through the likelihood function. The form of the probability distribution in eq. (17) suggests that this model is an undirected alternative to a 2-D-HMM (Nawaz & Curtis 2017). A hidden Markov model (HMM) is a directed graphical probabilistic model (with directed edges) with *causal* probabilistic influence. In other words the direction of flow of probabilistic influence in a directed model is constrained to the directions of edges. Although such causality has no direct physical interpretation in a spatial context, this allows for analytical computation of posterior probabilities. An MRF (or HMRF), on the other hand, is a more natural representation of spatial phenomena, but it does not allow analytical computation of posterior probabilities because of the intractable normalizing constant $\mathcal{Z}'$. For this reason, inference is typically performed using sampling-based methods such as McMC.

In view of the high computational cost of McMC methods discussed earlier, we seek alternative numerical inference methods. In this paper, we use *variational Bayes* method which is an attractive alternative to McMC-based methods because of its computational efficiency. The variational Bayes method is generally used to find a factorizable distribution as a proxy for the true posterior distribution which is not always factorizable. We would like to emphasize here that besides its computational efficiency, the variational Bayes method for inference is a natural choice in our model due to the fact that the posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ as given by eq. (17) is fully factorized as a consequence of the conditional independence assumption over the observed variables $\boldsymbol{x}$.

5 VARIATIONAL BAYESIAN INFERENCE

The major challenge in Bayesian inversion is calculation of the partition function $\mathcal{P}(\boldsymbol{x}|\Theta)$ in eq. (15). This requires summation over a prohibitively large number of configurations (other than in toy problems). We therefore seek a more tractable approximation. A common approach uses stochastic sampling, such as McMC, which is based on the concept that the probability distribution emerges as the number of samples tends to infinity. Variational Bayesian inference offers a more efficient alternative. It approximates a complex posterior distribution by a simpler, so called *auxiliary* distribution with which it is relatively easier to work. Such an approximation is commonly referred to as the *variational approximation*. We exploit the conditional independence assumption over observed variables $\boldsymbol{x}$ in our MRF model in order to perform inference on the latent variables $\boldsymbol{z}$ conditioned on $\boldsymbol{x}$, by reducing their joint distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ to a factorizable form as given in eq. (17). Such factorization connotes conditional independence among various sets of variables in the model. Such an approximation is inherent in our choice of the MRF model as the conditional independence among the cliques is a characteristic of an MRF. Below we show how factorization allows the Bayesian inference problem to be transformed into a constrained optimization problem under the variational approximation.

As a function of the model parameters Θ, the joint distribution $\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ over observed and latent variables $\boldsymbol{x}$ and $\boldsymbol{z}$ in eq. (15) is called the complete-data likelihood and the normalization constant $\mathcal{P}(\boldsymbol{x}|\Theta)$ is called the incomplete-data likelihood (as it does not involve the latent variables $\boldsymbol{z}$), the partition function or the model evidence. Estimation of the partition function requires marginalization of the complete-data likelihood over all possible configurations of the latent variables $\boldsymbol{z}$, that is

$$\mathcal{P}\ (\boldsymbol{x}|\Theta) = \sum_{\boldsymbol{z}} \mathcal{P}\,(\boldsymbol{x}, \boldsymbol{z}|\Theta)\,. \tag{18}$$

Rather than trying to estimate $\mathcal{P}(\boldsymbol{x}|\Theta)$ as a general function of Θ which is intractable, we first try to estimate Θ from the observations $\boldsymbol{x}$; once Θ has been fixed, estimating $\mathcal{P}(\boldsymbol{x}|\Theta)$ is a more tractable problem. The parameters Θ can be estimated from the observations $\boldsymbol{x}$ using the *maximum-likelihood* (ML) method that aims to find the parameters by setting $\Theta = \hat{\Theta}_{\mathrm{ML}}$ that maximizes the joint likelihood, or equivalently

the logarithm of joint likelihood of $\boldsymbol{x}$ and $\boldsymbol{z}$ as a function of the model parameters Θ:

$$\hat{\Theta}_{\mathrm{ML}} = \underset{\Theta}{\operatorname{argmax}} \{\log \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)\} \equiv \underset{\Theta}{\operatorname{argmax}} \{\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})\}. \tag{19}$$

If the facies $\boldsymbol{z}$ were observed, $\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})$ defines the complete log-likelihood as a function of the model parameters Θ. However, since $\boldsymbol{z}$ is a latent variable, it must be marginalized out resulting in the marginal log-likelihood of the observed variables, henceforth referred to as incomplete log-likelihood, that can be written as a function of parameters Θ as

$$\mathcal{L}(\Theta; \boldsymbol{x}) \equiv \log \mathcal{P}(\boldsymbol{x}|\Theta) = \log \sum_{\boldsymbol{z}} \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta). \tag{20}$$

The estimation of Θ is hard in this case as the presence of latent variables in the model may introduce dependencies between the parameters. In order to address these difficulties, we use the variational Bayesian approach (Beal 2003) that employs an auxiliary variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ of the latent variables $\boldsymbol{z}$ from a family $\mathcal{Q}$ of distributions that are more easily manipulated (typically expressible in a factorized form). The expected complete log-likelihood under $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x}) \in \mathbb{Q}$ may then be defined as a function of the model parameters Θ as

$$\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})] \equiv \sum_{\boldsymbol{z}} \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x}) \log \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta), \tag{21}$$

which is linear in the complete log-likelihood and is equally factorizable. The notation $\mathbb{E}_{\mathcal{Q}}[\cdot]$ represents expectation of the argument with respect to the auxiliary distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$. As we show below, this allows estimation of posterior marginal distributions and the maximum a posteriori (MAP) solution to the Bayesian inverse problem through inference on $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ rather than $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$. Since there is no ambiguity in the arguments of $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ as it does not explicitly depend on Θ, we often denote it just as $\mathcal{Q}$.

The expected complete log-likelihood $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$ acts as a lower bound on the incomplete log-likelihood $\mathcal{L}(\Theta; \boldsymbol{x})$ as can be seen by

$$\begin{aligned}
\mathcal{L}(\Theta; \boldsymbol{x}) &= \log \sum_{\boldsymbol{z}} \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta) \\
&= \log \mathbb{E}_{\mathcal{Q}}\left[\frac{\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)}{\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&\geq \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)] - \mathbb{E}_{\mathcal{Q}}[\log \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})] \qquad \text{[using Jensen's inequality]} \qquad (22) \\
&= \mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})] + \mathcal{S}_{\mathcal{Q}}(\boldsymbol{z}) \qquad (23) \\
&\equiv \mathcal{F}(\mathcal{Q}, \Theta), \qquad (24)
\end{aligned}$$

where $\mathcal{S}_{\mathcal{Q}}(\boldsymbol{z}) = -\mathbb{E}_{\mathcal{Q}}[\log \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})]$ is the *entropy* of the distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ and the functional $\mathcal{F}(\mathcal{Q}, \Theta)$ is called the *variational free energy* or simply *free energy*. The first term in eq. (23), $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$, represents the expectation of the complete log-likelihood $\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})$ with respect to the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ as defined in eq. (21). If we interpret $-\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})$ as the energy function of the MRF then $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$ represents negative of the so called expected energy under $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$, and $\mathcal{F}(\mathcal{Q}, \Theta)$ corresponds to the negative of Gibbs free energy in statistical physics (Feynman 1972).

Variational inference methods aim to estimate the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ of the latent variables $\boldsymbol{z}$ that maximizes the free energy functional $\mathcal{F}(\mathcal{Q}, \Theta)$ since this is guaranteed to increase the allowable values of the incomplete log-likelihood $\mathcal{L}(\Theta; \boldsymbol{x})$ by eq. (22). Since $\mathcal{F}(\mathcal{Q}, \Theta)$ is a lower bound of $\mathcal{L}(\Theta; \mathrm{x})$, the variational Bayesian method allows us to cast the inference problem into a constrained optimization problem. We attempt to maximize $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to both $\mathcal{Q}$ and Θ rather than directly estimating $\mathcal{L}(\Theta; \mathrm{x})$ which is intractable in most high-dimensional problems. Also by definition

$$\begin{aligned}
\mathcal{F}(\mathcal{Q}, \Theta) &= \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)] - \mathbb{E}_{\mathcal{Q}}[\log \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})] \\
&= \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)] + \mathbb{E}_{\mathcal{Q}}[\log \mathcal{P}(\boldsymbol{x}|\Theta)] - \mathbb{E}_{\mathcal{Q}}[\log \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})] \\
&= \mathbb{E}_{\mathcal{Q}}\left[\log \frac{\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)}{\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})}\right] + \log \mathcal{P}(\boldsymbol{x}|\Theta) \qquad \text{[as } \log \mathcal{P}(\boldsymbol{x}|\Theta) \text{ is independent of } \mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})\text{]} \\
&= -KL(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)) + \mathcal{L}(\Theta; \boldsymbol{x}), \qquad (25)
\end{aligned}$$

where $KL(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta))$ is the Kullback–Leibler (KL) divergence (or relative-entropy) between the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ and the true posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$. Since $\mathcal{L}(\Theta; \boldsymbol{x})$ is independent of $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$, maximizing $\mathcal{F}(\mathcal{Q}, \Theta)$ is equivalent to minimizing the relative-entropy $KL(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta))$ (Fig. 4). The KL divergence takes a minimum value of zero when the two distributions that it compares are identical. Therefore, by maximizing the free energy $\mathcal{F}(\mathcal{Q}, \Theta)$ for a given set of parameters Θ the VBI effectively estimates $\mathcal{Q}$ that best approximates the posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$.

As $\mathcal{P}(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ factorizes over the cliques in an MRF by definition, it follows from eq. (23) that $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$ can be computed efficiently, but estimation of $\mathcal{S}_{\mathcal{Q}}(\boldsymbol{z})$, and hence $\mathcal{F}(\mathcal{Q}, \Theta)$, is still computationally expensive. In order to overcome this difficulty, we use a variational form of the EM algorithm (Dempster *et al.* 1977; Beal 2003) which attempts to approximate $\mathcal{F}(\mathcal{Q}, \Theta)$ in an iterative fashion such that the lower bound $\mathcal{F}(\mathcal{Q}, \Theta)$ is increased while decreasing $KL(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta))$ for a given set of parameters in each iteration. The EM algorithm involves two steps in each iteration: the so-called E-step and the M-step, which aim to alternately maximize the free-energy $\mathcal{F}(\mathcal{Q}, \Theta)$ with

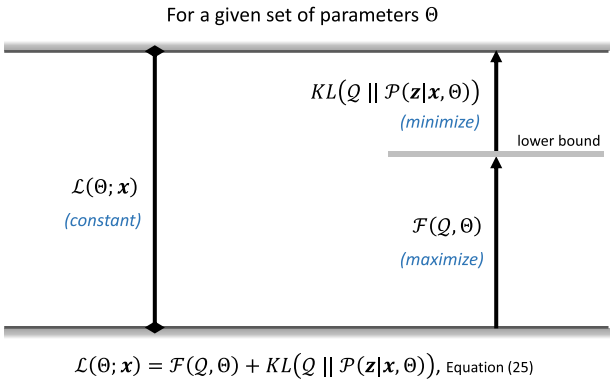


Figure 4. A schematic illustration of minimizing the relative-entropy $\boldsymbol{KL}(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta}))$ between the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ and the true posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta})$ for a fixed set of parameters $\boldsymbol{\Theta}$. Since the log-likelihood $\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{x})$ of observed variables $\boldsymbol{x}$ is a constant for fixed $\boldsymbol{\Theta}$, maximizing the variational free energy $\mathcal{F}(\mathcal{Q}, \boldsymbol{\Theta})$ with respect to $\mathcal{Q}$ corresponds to minimizing the relative-entropy $\boldsymbol{KL}(\mathcal{Q} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta}))$ between $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ and $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta})$.

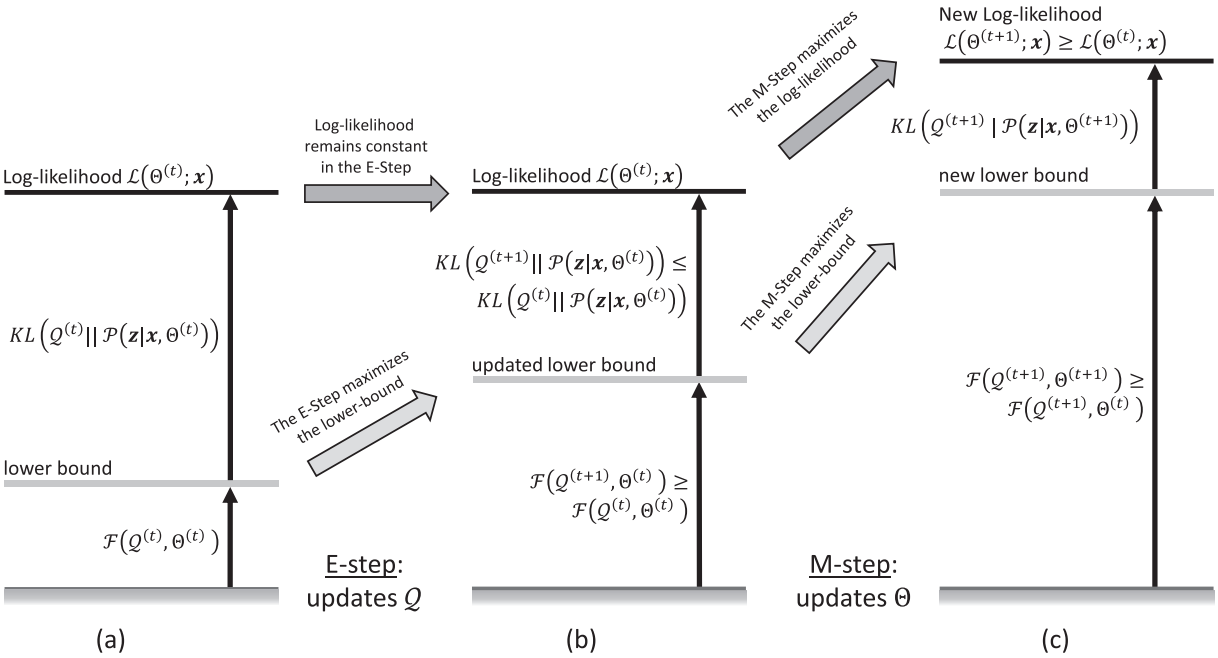


Figure 5. A schematic illustration of the variational EM algorithm. (a) After $\boldsymbol{t}$ iterations of the EM algorithm, we have estimates of the variational distribution as $\mathcal{Q}^{(t)}$ and the parameters $\boldsymbol{\Theta}^{(t)}$. (b) The E-step then maximizes the lower bound (free energy functional) $\mathcal{F}(\mathcal{Q}^{(t)}, \boldsymbol{\Theta}^{(t)})$ of the incomplete log-likelihood $\mathcal{L}(\boldsymbol{\Theta}^{(t)}; \boldsymbol{x})$ for fixed $\boldsymbol{\Theta}^{(t)}$ yielding updated estimate of the variational distribution $\mathcal{Q}^{(t+1)}$ and the updated lower bound $\mathcal{F}(\mathcal{Q}^{(t+1)}, \boldsymbol{\Theta}^{(t)}) \geq \mathcal{F}(\mathcal{Q}^{(t)}, \boldsymbol{\Theta}^{(t)})$. The incomplete log-likelihood $\mathcal{L}(\boldsymbol{\Theta}^{(t)}; \boldsymbol{x})$ remains constant during the E-step, and as a consequence the relative-entropy $\boldsymbol{KL}(\mathcal{Q}^{(t)} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta}^{(t)}))$ between the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ and the true posterior distribution $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta})$ is minimized to $\boldsymbol{KL}(\mathcal{Q}^{(t+1)} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta}^{(t)}))$ for fixed $\boldsymbol{\Theta}^{(t)}$—see Fig. 4. (c) The M-step estimates a new set of parameters $\boldsymbol{\Theta}^{(t+1)}$ by maximizing the incomplete log-likelihood to $\mathcal{L}(\boldsymbol{\Theta}^{(t+1)}; \boldsymbol{x}) \geq \mathcal{L}(\boldsymbol{\Theta}^{(t)}; \boldsymbol{x})$ thereby maximizing the lower bound to $\mathcal{F}(\mathcal{Q}^{(t+1)}, \boldsymbol{\Theta}^{(t+1)}) \geq \mathcal{F}(\mathcal{Q}^{(t+1)}, \boldsymbol{\Theta}^{(t)})$ to yield a new estimate of the relative-entropy as $\boldsymbol{KL}(\mathcal{Q}^{(t+1)} \,||\, \mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\Theta}^{(t+1)}))$. This is subsequently minimized in the E-step of the next iteration of the EM algorithm which iterates until convergence.

respect to $\mathcal{Q}$ and Θ, respectively. In concept, the E-step estimates the posterior distribution of facies $\boldsymbol{z}$ (which factorizes in an MRF as shown by eq. 2) in the model space for a given estimate of parameters Θ, whereas the M-step updates the current estimate of model parameters Θ in the attribute space by maximizing their likelihood for the current estimate of the posterior distribution of facies (Fig. 5). Alternate E- and M-steps therefore improve the estimates of $\mathcal{Q}$ and Θ such that the log-likelihood is guaranteed not to decrease in any iteration. With a suitable initialization, the EM algorithm is guaranteed to converge to a local optimum within a relatively small number of iterations (Balakrishnan *et al.* 2017).

5.1 E-step—estimation of $\mathcal{Q}$ from the current estimate of model parameters Θ

In the E-step of iteration t, the variational distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ over the latent variables $\boldsymbol{z}$ is estimated from the current estimate of the model parameters $\Theta^{(t)}$ by maximizing the free-energy $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to $\mathcal{Q}$. The E-step may therefore be written as

$$\mathcal{Q}^{(t+1)} = \underset{\mathcal{Q}}{\operatorname{argmax}} \left\{\mathcal{F}\left(\mathcal{Q}, \Theta^{(t)}\right)\right\}, \tag{26}$$

where the bracketed superscripts refer to the iteration number. Since $\mathcal{F}(\mathcal{Q}, \Theta)$ is a lower bound of $\mathcal{L}(\Theta; \boldsymbol{x})$, maximizing the lower bound $\mathcal{F}(\mathcal{Q}, \Theta^{(t)})$ of the incomplete log-likelihood $\mathcal{L}(\Theta^{(t)}; \boldsymbol{x})$ with respect to $\mathcal{Q}$ results in $\mathcal{Q}^{(t+1)}$ equal to the estimate of the posterior distribution $\hat{\mathcal{P}}(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t)})$. This can be proved by setting $\mathcal{Q}$ equal to $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t)})$ in the inequality (24).

Since computing free energy $\mathcal{F}(\mathcal{Q}, \Theta)$ is computationally hard, we seek more efficient approximate alternatives. The distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ in a pairwise MRF is specified by approximate marginal distributions $b_i(z_i)$ over the vertices, and $b_{ij}(z_i, z_j)$ over the edges in the graphical model as defined below. The free energy $-\mathcal{F}(\mathcal{Q}, \Theta)$ can then be approximated for pairwise MRFs by the *Bethe's free energy* $\hat{\mathcal{F}}_B$ (also called *Kikuchi free energy* for general MRFs) given by

$$\hat{\mathcal{F}}_B = \sum_{(i,j)\in\mathcal{E}} \sum_{(z_i,\, z_j)} b_{ij}\left(z_i,\, z_j\right) \log\left(\frac{b_{ij}\left(z_i,\, z_j\right)}{\varphi_i\left(z_i\right)\, \varphi_j\left(z_j\right)\, \psi_{ij}\left(z_i,\, z_j\right)}\right) - \sum_{i\in\mathcal{V}} \left(|\mathcal{N}_{\setminus i}| - 1\right) \sum_{z_i} b_i\left(z_i\right) \log\left(\frac{b_i\left(z_i\right)}{\varphi_i\left(z_i\right)}\right), \tag{27}$$

where $|\mathcal{N}_{\setminus i}|$ represents the neighbourhood cardinality of i, that is, the number of vertices that are neighbours of i (Bethe 1935; Yedidia *et al.* 2001a,b). The Bethe's free energy only approximates the entropy term $\mathcal{S}_{\mathcal{Q}}(\boldsymbol{z})$ in eq. (23) which is hard to compute; the expectation term $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$ remains exact. The approximate marginal distributions $b_i(z_i)$ and $b_{ij}(z_i, z_j)$ over vertices and edges in the graph are commonly referred to as *pseudo-marginals* or *beliefs*. The above expression for Bethe's free energy $\hat{\mathcal{F}}_B$ is a direct consequence of a re-parametrization of the posterior distribution from the original parameters in terms of potential functions $\varphi_i(z_i)$ and $\psi_{ij}(z_i, z_j)$, to the new parameters in terms of beliefs $b_i(z_i)$ and $b_{ij}(z_i, z_j)$ under the so called *admissibility constraints*:

$$\mathcal{P}\left(\boldsymbol{z}|\boldsymbol{x}\right) = \frac{1}{\mathcal{Z}} \prod_i \varphi_i\left(\boldsymbol{x}_i,\, \boldsymbol{z}_{\mathcal{N}_i}\right) \prod_{(i,j)\in\mathcal{E}} \psi_{ij}\left(z_i,\, z_j\right) \propto \prod_{(i,j)\in\mathcal{E}} b_{ij}\left(z_i,\, z_j\right) / \prod_i b_i(z_i)^{|\mathcal{N}_{\setminus i}|-1}. \tag{28}$$

The Bethe's free energy $\hat{\mathcal{F}}_B$ is exactly equal to the free energy $\mathcal{F}(\mathcal{Q}, \Theta)$ for an acyclic MRF (Koller & Friedman 2009) and can be computed efficiently as it only involves summation over the vertices and edges in a pairwise MRF. Yedidia *et al.* (2001a,b) showed that the stationary points of Bethe's free-energy correspond to the fixed points of an iterative message-passing algorithm, the so called *belief propagation* (BP) algorithm introduced by Pearl (1982).

BP performs approximate inference in graphical models by estimating marginal distributions of unobserved variables conditioned on any observed variables by passing messages over edges in the graph. A message $\mathfrak{m}_{j\to i}(z_i)$ from the vertex j to the vertex i is a real function with domain z_i, the set of values that can be taken by an unobserved vertex i, and represents probabilistic influence of a vertex j on the vertex i. In other words, a message $\mathfrak{m}_{j\to i}(z_i)$ encodes 'belief' of a vertex j about the state z_i of an unobserved vertex i. The beliefs $b_i(z_i)$ and $b_{ij}(z_i, z_j)$ can be expressed in terms of messages as

$$b_i\left(z_i\right) \propto \varphi_i\left(z_i\right) \prod_{j\in\mathcal{N}_{\setminus i}} \mathfrak{m}_{j\to i}\left(z_i\right) \tag{29}$$

$$b_{ij}\left(z_i, z_j\right) \propto \varphi_i\left(z_i\right) \varphi_j\left(z_j\right) \psi_{ij}\left(z_i, z_j\right) \prod_{h\in\mathcal{N}_{\setminus i}\setminus\{j\}} \mathfrak{m}_{h\to i}\left(z_i\right) \prod_{h\in\mathcal{N}_{\setminus j}\setminus\{i\}} \mathfrak{m}_{h\to j}\left(z_j\right). \tag{30}$$

Combining these equations yields the belief propagation equation (Pearl 1982)

$$\mathfrak{m}_{j\to i}\left(z_i\right) \propto \sum_{z_j} \varphi_j\left(z_j\right) \psi_{ij}\left(z_i, z_j\right) \prod_{h\in\mathcal{N}_{\setminus j}\setminus\{i\}} \mathfrak{m}_{h\to j}\left(z_j\right), \tag{31}$$

which presents a schedule for message passing and shows how a vertex encodes messages that it receives from its neighbours except the target vertex, and passes the encoded messages to its target neighbouring vertex. The schedule starts with a vertex j receiving messages $\mathfrak{m}_{h\to j}(z_j)$ from each of its neighbours $h \in \mathcal{N}_{\setminus j}\setminus\{i\}$ except its target vertex i. Fig. 6 shows a typical illustration of the schedule of messages received by a given vertex from its neighbours other than the target vertex, and the message it sends to its target neighbouring vertex. The received messages are multiplied together for each of the possible values of z_j and then scaled with the vertex and edge potentials $\varphi_j(z_j)$ and $\psi_{ij}(z_i, z_j)$ for a given value of the state z_i of i. The resulting scaled products of messages are then summed over all of the possible values of z_j and then forwarded by the vertex j to the vertex i encoding the belief of j regarding the state of i being equal to z_i. The observed vertices in a hidden MRF also send messages to their neighbouring latent vertices; however, they cannot receive any messages as their values are fixed.

Eq. (31) is often referred to as the sum-product equation for obvious reasons and may be solved using the BP algorithm. The BP algorithm is an exact inference method for tree-structured graphs in which case it can be shown to converge to the true marginal distributions in a number of iterations equal to the diameter of the tree (Koller & Friedman 2009). In cyclic graphs, a variant of BP known as the LBP can be used which is an approximate inference method. LBP is not guaranteed to converge, however, it has been shown empirically to converge in most cases (Pearl 1982, 1988; Murphy *et al.* 1999). We discuss this point further in Section 9, but the LBP algorithm has seen wide

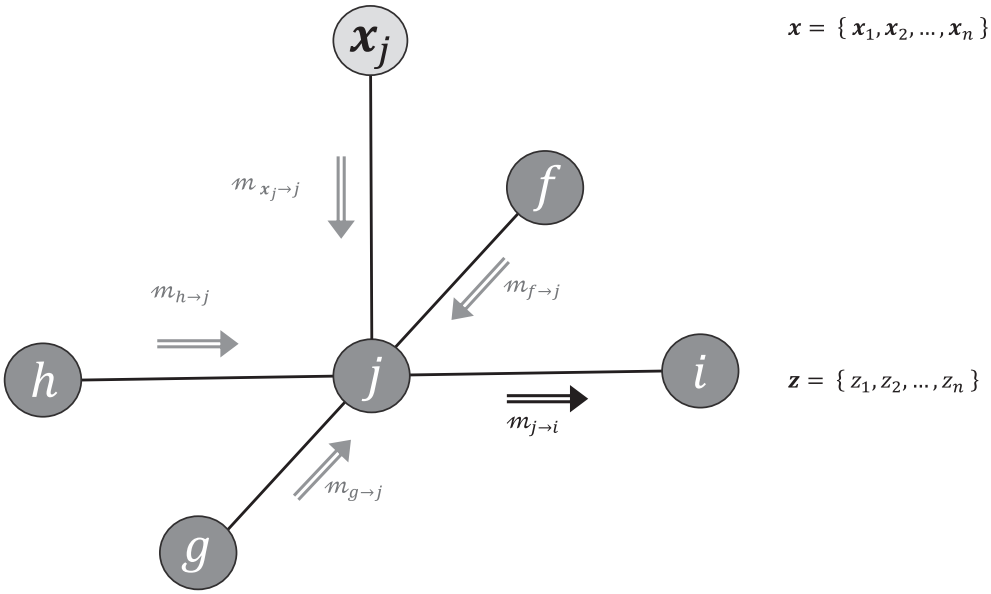


Figure 6. A schematic illustration of message passing. The light-grey circle represents an observed vertex (or variable), dark-grey circles represent latent vertices and the solid lines connecting these circles represent edges in the graphical model. Double-lined arrows represent messages flowing between vertices as labelled. The vertex $\boldsymbol{j}$ receives messages $\boldsymbol{m}_{\cdot \to j}$ from all of its neighbours (including the observed vertex $\boldsymbol{x}_j$) except the vertex $\boldsymbol{i}$ which is the current target for a message from $\boldsymbol{j}$. The messages received by $\boldsymbol{j}$ are combined together and encoded into a message $\boldsymbol{m}_{j \to i}$ according to eq. (31). The encoded message $\boldsymbol{m}_{j \to i}$ is then forwarded by $\boldsymbol{j}$ to $\boldsymbol{i}$. Only latent vertices can receive messages. Observed vertices can only send messages to their neighbouring latent vertices, and cannot receive any messages as their values are fixed. Propagation of messages in this manner between all vertices in a graph constitutes what is commonly known as the belief propagation (BP) algorithm.

applicability and success in various fields of research, for example in statistics (e.g. Pearl 1988; Yasuda *et al.* 2015), digital signal and image processing (e.g. Sudderth & Freeman 2008), artificial intelligence (e.g. Tatikonda & Jordan 2002) and biology (e.g. Sinoquet & Mourad 2014). The LBP is a variant of BP in which messages are passed iteratively until convergence is detected or until a maximum number of iterations is exceeded. Convergence may be detected if less than a pre-defined threshold number of vertices are updated, or all vertices are updated by an amount less than a predefined tolerance.

Messages are generally initialized with unity or with random numbers greater than a positive tolerance, and then updated according to a pre-defined message schedule using eq. (31). After the messages have converged based on some convergence detection criteria, the vertex beliefs are updated according to eq. (29) to give approximate marginal posterior distributions. Despite that the vertex potentials $\varphi_i(x_i)$ and the edge potentials $\psi_{ij}(z_i, z_j)$ need not be exact probabilities, their marginalization and normalization ensures numerical stability of the LBP algorithm. Also, since the LBP involves several iterative multiplications of potential functions at each vertex, the LBP algorithm is usually run in the logarithmic domain in order to avoid numerical underflow.

The LBP algorithm may also be used to perform MAP inference which computes the most likely configuration, rather than the approximate marginal posterior distributions. MAP inference minimizes the error probability that the most likely configuration, also known as the MAP estimate, does not coincide with the true one. This can be achieved by replacing the summation in the sum-product eq. (31) with the max function yielding the corresponding max-product equation as

$$\mathfrak{m}_{j \to i}(z_i) \propto \max_{z_j} \left\{ \varphi_j(z_j)\, \psi_{ij}(z_i, z_j) \prod_{h \in \mathcal{N}_{\setminus j} \setminus \{i\}} \mathfrak{m}_{h \to j}(z_j) \right\}. \tag{32}$$

The LBP algorithm on an MRF is summarized in Algorithm 1. If Algorithm 1 converges, the beliefs $b_i(z_i)$ and $b_{ij}(z_i, z_j)$ are updated using eqs (29) and (30). The variational distribution $\mathcal{Q}^{(t+1)}$ at the end of the E-step of $(t+1)$th iteration of the EM algorithm is then approximated to $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x})$ using eq. (28).

5.2 M-step—estimation of model parameters Θ from the current estimate of $\mathcal{Q}$

In the M-step, the current estimate of the variational distribution $\mathcal{Q}^{(t+1)}$ obtained from the E-step is used to compute the updated set of model parameters $\Theta^{(t+1)}$ that maximize the free-energy $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to Θ. The M-step may therefore be written as follows:

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} \;\; \mathcal{F}\left(\mathcal{Q}^{(t+1)}, \Theta\right) = \underset{\Theta}{\operatorname{argmax}} \;\; \mathbb{E}_{\mathcal{Q}^{(t+1)}}\left[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})\right], \tag{33}$$

which follows from the fact that $\mathcal{S}_{\mathcal{Q}}(\boldsymbol{z})$ in eq. (23) is independent of the model parameters Θ. Thus, maximizing $\mathcal{F}(\mathcal{Q}, \Theta)$ with respect to Θ only requires that $\mathbb{E}_{\mathcal{Q}}[\mathcal{L}(\Theta; \boldsymbol{x}, \boldsymbol{z})]$ be maximized with respect to Θ. Accordingly, it turns out that the M-step may only require a few statistics

Algorithm 1: Loopy belief propagation over an undirected graphical model $\mathbb{G}(\mathcal{V},\mathcal{E})$ with accuracy ε and for a maximum number of iterations L. Comments follow the hash signs '#' till the end of each line.

```
Set sum_product ← true # or false for max-product
Initialize messages m_{j→i}^(0)(z_i)
Set l ← 1 # LBP iteration number
while l ≤ L
   Set δ ← 0
   for each i ∈ V
      for each j ∈ N_{\i} ⊂ V
         if sum_product # for sum-product algorithm
            Compute m_{j→i}^(l)(z_i) using equation (31)
         else # for max-product algorithm
            Compute m_{j→i}^(l)(z_i) using equation (32)
         end if
         Set δ ← max(δ, m_{j→i}^(l)(z_i) − m_{j→i}^(l−1)(z_i))
      end for j
   end for i
   if δ < ε
      Update beliefs using equations (29) and (30)
      print 'Converged!'
      exit
   end if
   Set l ← l + 1
end while
print 'Not converged!'
```

of the latent variables $\boldsymbol{z}$ computed in the E-step, instead of the full distribution $\mathcal{Q}^{(t+1)}(\boldsymbol{z}|\boldsymbol{x})$. Expanding eq. (33) in terms of the incomplete log-likelihood and substitution from eq. (14) gives

$$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}} \sum_i \sum_{z_i} b_i{}^{(t+1)}(z_i) \left(-\frac{n^2}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}| - \frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}{}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right) \right). \tag{34}$$

The solution to the above equation can be obtained with and without the assumption of *homoscedasticity* whereby the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is assumed to be scalar such that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} = \sigma^2 \boldsymbol{I}$. With this assumption, maximizing log-likelihood under the constraints $\sum_{z_i} b_i(z_i) = 1$ is equivalent to minimizing the residual sum-of-squares

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right), \tag{35}$$

which gives the ordinary least squares (OLS) solution

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{x}. \tag{36}$$

The OLS solution is also the unbiased maximum-likelihood solution if $\boldsymbol{R}$ is a full-rank matrix, otherwise one may seek the regularized least squares (RLS) solution given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{RLS}} = \left(\boldsymbol{R}^T\boldsymbol{R} + k\boldsymbol{I}\right)^{-1}\boldsymbol{R}^T\boldsymbol{x}, \tag{37}$$

where k is the control parameter which governs the relative strength of regularization (damping) applied. Similarly, the maximum-likelihood solution of eq. (34) with respect to $\boldsymbol{\Sigma_\varepsilon} = \sigma^2 \boldsymbol{I}$ is given by

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\hat{\boldsymbol{\beta}}\right), \tag{38}$$

but this is a biased estimator; the bias-corrected estimate (Rencher 2002) is given by

$$\hat{\sigma}^2 = \frac{1}{n-q-1} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\hat{\boldsymbol{\beta}}\right), \tag{39}$$

where we recall that $q = |\mathcal{N}_i|$ is the neighbourhood cardinality, which is assumed to be a constant for each location i in our graphical model. In the general case of heteroscedasticity whereby the covariance matrix is non-scalar, maximizing log-likelihood is equivalent to minimizing the residual weighted sum-of-squares

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\{ n \log|\boldsymbol{\Sigma_\varepsilon}| + \sum_i \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma_\varepsilon}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}\right) \right\}. \tag{40}$$

With $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} = (\hat{\sigma}_{kl})\,, \quad k, l \in \{1, \ldots, p\}$, where $\hat{\sigma}_{kl}$ is given by the OLS solution in eq. (39), the generalized least squares (GLS) solution is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\boldsymbol{R}^T \left(\boldsymbol{I}_n \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}\right)^{-1} \boldsymbol{R}\right)^{-1} \boldsymbol{R}^T \left(\boldsymbol{I}_n \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}\right)^{-1} \boldsymbol{x}, \tag{41}$$

where $\otimes$ represents the Kronecker product defined between two matrices $\boldsymbol{A} = [a_{mn}]$ and $\boldsymbol{B} = [b_{pq}]$ as a $(mp \times nq)$ matrix with elements

$$(\boldsymbol{A} \otimes \boldsymbol{B})_{i,j} = a_{\lfloor (i-1)/p \rfloor + 1, \lfloor (j-1)/q \rfloor + 1}\, b_{i - \lfloor (i-1)/p \rfloor p,\, j - \lfloor (j-1)/q \rfloor q},$$

where $\lfloor \cdot \rfloor$ represents the *floor* function which returns the greatest integer less than or equal to its argument.

The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k, \quad k \in \mathcal{G}$ of the Gaussian mixture distribution are iteratively updated by weighted averages of the seismic attributes $\boldsymbol{x}_i$ at each location i with respect to the estimated posterior marginal distributions $\hat{\mathcal{P}}_i(z_i|\boldsymbol{x}, \Theta)$ as estimated in the E-step to honor the spatial dependence among z_i's, as

$$\boldsymbol{\mu}_k{}^{(t+1)} = \frac{\sum_{i=1}^n \hat{\mathcal{P}}_i\left(z_i|\boldsymbol{x}, \Theta^{(t)}\right) \boldsymbol{x}_i}{\sum_{i=1}^n \hat{\mathcal{P}}_i\left(z_i|\boldsymbol{x}, \Theta^{(t)}\right)} \tag{42}$$

and

$$\boldsymbol{\Sigma}_k{}^{(t+1)} = \frac{\sum_{i=1}^n \hat{\mathcal{P}}_i\left(z_i|\boldsymbol{x}, \Theta^{(t)}\right) \cdot \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k{}^{(t+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k{}^{(t+1)}\right)^T}{\sum_{i=1}^n \hat{\mathcal{P}}_i\left(z_i|\boldsymbol{x}, \Theta^{(t)}\right)}, \tag{43}$$

where $\hat{\mathcal{P}}_i(z_i|\boldsymbol{x}, \Theta^{(t)})$ is approximated by the vertex beliefs $b_i{}^{(t)}(z_i)$ estimated from the LBP algorithm in the E-step of tth iteration.

In summary, at the end of $(t+1)$th iteration the E-Step of the EM algorithm yields the free energy $\mathcal{F}(\mathcal{Q}^{(t+1)}, \Theta^{(t)})$ equal to $\mathcal{L}(\Theta^{(t)}; \boldsymbol{x})$ which is the upper bound of $\mathcal{F}(\mathcal{Q}, \Theta^{(t)})$, and the M-step maximizes $\mathcal{F}(\mathcal{Q}^{(t+1)}, \Theta^{(t)})$ with respect to Θ. Therefore, the E-step improves the estimate of the posterior distribution of facies $\hat{\mathcal{P}}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$ in the model space while the M-step improves the estimates of model parameters Θ in the attribute space, such that the combined E–M steps are guaranteed not to decrease the incomplete log likelihood $\mathcal{L}(\Theta; \boldsymbol{x})$ during any iteration of the EM algorithm.

6 COMPUTATIONAL COST

The computational cost of this algorithm is defined by the cost of the LBP algorithm in the E-step and the solution of the linear problem in the M-step. The LBP is an iterative procedure and its computational cost depends on the number of iterations. The cost of E-step is therefore given by

$$C_E \leq n * K^2 * \max\ |\mathcal{N}| * L, \tag{44}$$

where $n = |\mathcal{V}|$ is the number of locations (vertices in the graph), $K = |\mathcal{G}|$ is the number of facies considered, max $|\mathcal{N}|$ represents the maximum neighbourhood cardinality (the maximum number of neighbouring vertices $\mathcal{N}_{\setminus i}$ of any vertex $i \in \mathcal{V}$ in the graph), and L is the total number of iterations in the LBP algorithm. Although there are cases when belief propagation does not converge (as in the case of repulsive potential functions, Koller & Friedman 2009), we consider the number of iterations assuming that the algorithm does converge. If the belief propagation algorithm converges, the required number of iterations depends on the desired accuracy in Algorithm 1, the initial values of beliefs, model size and complexity. Initial beliefs close to a local optimum result in a smaller number of iterations. A good choice for initial beliefs are the localized likelihoods (Walker & Curtis 2014; Nawaz & Curtis 2017), which are then updated based on the priors by the belief propagation algorithm.

Starting with reasonable initial beliefs, the belief propagation algorithm requires 10s to 100s of iterations in most cases, depending on the model size and complexity. Different strategies may be adopted to limit computational demands in models of large sizes and complexity.

For example, the regions in the graph in which beliefs do not change significantly in some pre-defined number of previous iterations may be eliminated from future iterations, thus effectively reducing the size of the graph.

The computational cost of solving the linear equations in the M-step is given by

$$C_M = n * p * (\max\ |\mathcal{N}|)^2, \tag{45}$$

where p is the number of input seismic attributes at each location. Since the E- and M-steps run alternately, the computational cost of the two steps is a sum of the computational cost of each step. The total computational cost of the EM algorithm is therefore

$$C_{\text{total}} \leq (C_{\text{E}} + C_{\text{M}}) * t_{\max}, \tag{46}$$

where $t_{\max}$ is the total number of EM iterations.

Convergence of the EM algorithm is fast and guaranteed provided that the belief propagation algorithm in the E-step converges. The important considerations for computational cost are the number of facies ($K = |\mathcal{G}|$) considered and the maximum size ($\max |\mathcal{N}|$) of the neighbourhood structure in the graph, as the above expressions (44) and (45) are quadratic in these variables, respectively. The size of the neighbourhood structure defines the extent of spatial correlations in seismic attributes that is incorporated within the likelihood function. The maximum size of the neighbourhood structure must therefore not be excessively large in order to avoid prohibitive computational costs. For this reason, the likelihoods in our method cannot be solved in fully non-localized form in realistic problems, hence the term *quasi-localized*. Our method is therefore based on a trade-off between computational tractability and the extent of spatial correlations incorporated from the data. All other parameters in the above expressions (44) and (45) are linear and therefore do not cause serious computational implications.

In previous research (Rimstad & Omre 2010; Walker & Curtis 2014; Nawaz & Curtis 2017), the size of the space of geological facies $K = |\mathcal{G}|$ was critical as the computational cost of those algorithms increases exponentially with K. It was therefore required that K is chosen to be as small as possible. As a consequence, that research considered typically small numbers of facies [e.g. Walker & Curtis (2014) and Nawaz & Curtis (2017) inverted for the same three classes as in the current synthetic test—see Section 7, while Rimstad & Omre (2010) inverted for four classes]. If the range of distinct facies classes is large, one was required to limit the number of facies classes, for example by merging or nesting relatively closely related facies (e.g. limestone and dolomite or shale and marl) within one another. By contrast, the computation cost of our algorithm herein is only quadratic in K. Thus, our algorithm may be able to operate with a larger number of facies compared to previous algorithms without requiring merging or nesting of similar facies before incurring serious computational limitations.

Since both the LBP and the EM algorithms are iterative, the current method is more demanding of computational power and is not as parallelizable as the HMM-based method of Nawaz & Curtis (2017) which is mainly analytical and can easily be parallelized as it does not require iterative convergence. However, the cost of the current method confers the additional advantages of parameter optimization as part of the inversion scheme, and the fact that it does not require the assumption of localized likelihoods as is required in the algorithms of Walker & Curtis (2014) and Nawaz & Curtis (2017).

For large models which require parallelization of the algorithm in order to improve computational speed, each iteration of the LBP algorithm in the E-step may be parallelized over the vertices of the graph (the *for* loop in line 6 of Algorithm 1). A key consideration concerning the convergence and computational performance of the LBP algorithm is message scheduling. Although synchronous scheduling may be desired where all of the messages are updated at once for higher computation efficiency, an asynchronous schedule is optimal both for convergence and performance. Koller & Friedman (2009) suggested a residual belief propagation schedule which dynamically detects convergence in different parts of the graph and schedules messages in the parts where beliefs disagree most strongly. Also, the solution of the linear problem in the M-step may be parallelized to improve performance (e.g. Koc & Piedra 1991).

The memory required by the current algorithm is similar to that of Nawaz & Curtis (2017) and is far less than that required to store the partial conditional distributions in the method of Walker & Curtis (2014). Although below we demonstrate the method on a 2-D model, our method can be applied to practical 3-D models with reasonably sized neighbourhood structures without requiring any modifications or extensions, and without facing severe computational limitations on standard modern computers: this follows from the fact that our method is based on a linear indexing of cells, and models of any dimensionality can be unwrapped to conform to such linear indexing.

7 SYNTHETIC TEST

In order to test our algorithm, and in particular to benchmark it against previous research, we generated synthetic seismic attribute data similar to that used in Walker & Curtis (2014) and Nawaz & Curtis (2017). The synthetic example is based on two vertical cross-sections extracted from a 3-D geological process model that contains channel-filled and overbank sand deposits with background shale. The channel sands are mostly filled with brine with some of the channels containing gas such that the two fluids obey gravitational ordering (gas above brine, all else being equal). The lithofacies considered for discrimination are therefore given by the sample space

$$\mathcal{G} = \{\ \text{shale},\ \text{brine}-\text{sand},\ \text{gas}-\text{sand}\ \}\ .$$

We used one of the vertical cross-sections with dimensions of 200 × 200 cells as a given training image (Fig. 7a), and another with dimensions of 100 × 100 cells as the target model (Fig. 7b) representing the unknown true earth. Since both the training image and the target cross-section were extracted from the same geological process model, they are assumed to contain statistically similar patterns and conditional distributions of facies.

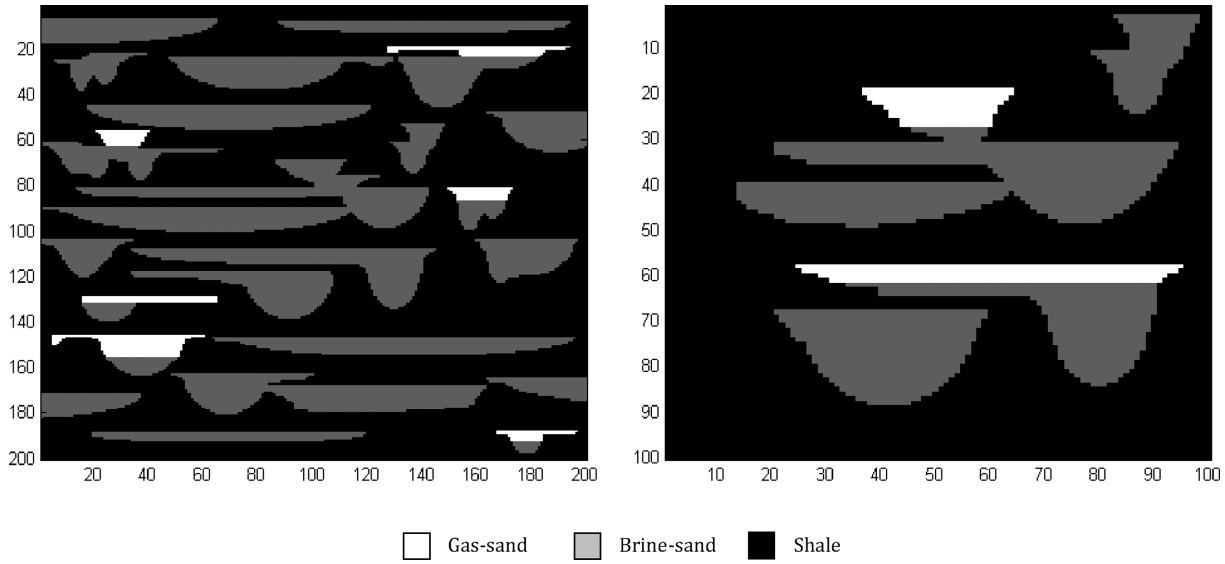


Figure 7. The training image (TI) and the target image extracted as 2-D cross-sections from a 3-D geological process model containing channels with filled and overbank sand deposits and shale in the background. The sand is filled with brine or gas, which obey gravitational ordering of the two fluids. (a) The training image: this represents a conceptual depiction of typical forms of expected geological structures and spatial distributions of facies. It lacks any location-specific information about the real geology of the target image. It was scanned with a 3 × 3 template to obtain spatial conditional prior probability distributions of facies (for pairs of locations at a time) to obtain prior information related to the spatial continuity and association of various geological facies. (b) The target image: this represents the true geological model which is the target for spatial facies inversion. It is assumed to contain statistically similar spatial patterns and conditional distributions of facies as the training image.

The training image encodes the spatial conditional prior distributions of facies graphically. These can be extracted by scanning it with a template of cells whose shape and size are defined by the neighbourhood structure $\mathcal{N}_i$. In our example, the prior information was extracted from the training image in terms of prior probabilities $\mathcal{P}(z_i|z_j \in \boldsymbol{z}_{\mathcal{N}_i})$ constructed from histograms of various facies configurations that occur in the image, where $\boldsymbol{z}_{\mathcal{N}_i} = \{z_j \in \boldsymbol{z} : j \neq i \ \wedge \ (i, j) \in \mathcal{E}\}$ for various configurations of i and j that define their relative positions within a square 3 × 3 cell neighbourhood structure. The prior information encoded in $\mathcal{P}(z_i|z_j \in \boldsymbol{z}_{\mathcal{N}_i})$ represents the probability of observing facies $z_i \in \mathcal{G}$ at the vertex $i \in \mathcal{V}$ given that facies $z_j \in \mathcal{G}$ exists at vertex $j \in \mathcal{N}_i$ for every $(i, j) \in \mathcal{E}$. These are the so called two-point statistics from the training image. These prior probabilities encapsulate the spatial conditional distributions of facies under the assumption that they are stationary over the entire model space.

As discussed in Section 6, the computational cost of the M-step (using the linear problem) increases as the square of the maximum neighbourhood cardinality in the graph. The computational cost of the E-step (the LBP algorithm) only increases linearly with the maximum neighbourhood cardinality. The size of the neighbourhood template could therefore be reasonably increased from 3 × 3 model cells (or graph vertices) without incurring any serious computational limitations. Even if a 3 × 3 neighbourhood template is too small to reproduce more complex geological patterns (such as complex aerial meandering of channels in a deltaic environment), it is shown below to reproduce the cross-sectional patterns of channels in our model reasonably well.

Synthetic *P*- and *S*-wave impedance profiles were generated from the target cross-section to represent the corresponding real data-derived seismic attributes. These synthetic seismic attributes were then inverted using our algorithm to estimate marginal posterior distributions of geological facies with the aim to reproduce the target cross-section. Synthetic attributes $\boldsymbol{x}'_i$ were first generated independently in each model cell i from a Gaussian mixture distribution using the Yin–Marion shaly-sand model (Marion 1990; Yin *et al.* 1993; Avseth *et al.* 2005). The Yin–Marion model is defined by the petrophysical parameters $\boldsymbol{m}_k = [V_{\text{clay}}, \ \varphi_{\text{sand}}, S_{\text{w}}]_k$, where V_{clay} represents the volume of clay, φ_{sand} represents the sand matrix porosity, S_{w} represents the water saturation (such that the gas saturation is given by $1 - S_{\text{w}}$) and the subscript k refers to the facies. The conditional probability $\mathcal{P}(\boldsymbol{x}'_i \mid z_i)$ of local facies responses $\boldsymbol{x}'_i$ given the geological facies z_i at location i is then given by

$$\mathcal{P}\left(\boldsymbol{x}'_i \mid z_i\right) = \int \int \int_{l}^{u} \mathcal{P}\left(\boldsymbol{x}'_i \mid \boldsymbol{m}_k\right) \ \mathcal{P}\left(\boldsymbol{m}_k \mid z_i\right) \ d\boldsymbol{m}_k, \tag{47}$$

where $\boldsymbol{l}$ and $\boldsymbol{u}$ (with boldface vector notation), respectively, represent the lower and the upper vector bounds for each of the three petrophysical parameters in $\boldsymbol{m}_k$, and are given in Table 1.

The conditional distribution $\mathcal{P}(\boldsymbol{m}_k \mid z_i)$ in eq. (47) represents the probabilistic relationship between the petrophysical parameters $\boldsymbol{m}_k$ and the facies z_i in each cell of the target cross-section, and was set to a Uniform distribution within the pre-defined bounds $\boldsymbol{l}$ and $\boldsymbol{u}$ on $\boldsymbol{m}_k$ (Table 1). The conditional distribution $\mathcal{P}(\boldsymbol{x}'_i \mid \boldsymbol{m}_k)$ in eq. (47) represents the statistical rock physics model and was set to a Normal distribution

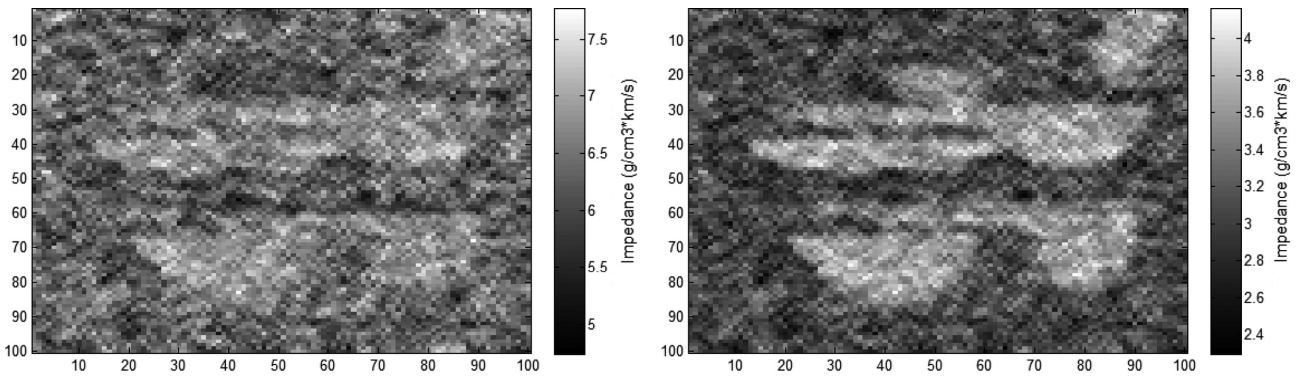
Table 1. Lower and upper bounds used to define Uniform distributions $\boldsymbol{P(m_k \mid z_i)}$ over petrophysical parameters $\boldsymbol{m_k} = [\boldsymbol{V_{\text{clay}}}, \boldsymbol{\varphi_{\text{sand}}}, \boldsymbol{S_w}]_k$.

Lithology–fluid class	Clay content by volume ($\boldsymbol{V_{\text{clay}}}$)	Sandstone matrix porosity ($\boldsymbol{\varphi_{\text{sand}}}$)	Water saturation ($\boldsymbol{S_w}$)
Shale	[0.50, 0.90]	[0.10, 0.40]	[1.00, 1.00]
Brine-sand	[0.00, 0.20]	[0.20, 0.40]	[0.40, 1.00]
Gas-sand	[0.10, 0.40]	[0.20, 0.40]	[0.00, 0.30]

Table 2. Covariance matrices of seismic attributes (*P*- and *S*-wave impedances) for the three facies considered.

Lithology–fluid Class	Covariance matrix for seismic attributes: *P*- and *S*-wave impedances
Shale	$\begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 0.5 \end{bmatrix}$
Brine-sand	$\begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}$
Gas-sand	$\begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.5 \end{bmatrix}$

The diagonal entries in the above matrices are variances of *P*- and *S*-wave impedances, whereas the cross-diagonal entries are the covariances of *P*- and *S*-wave impedances.



(a) P-wave impedance (b) S-wave impedance

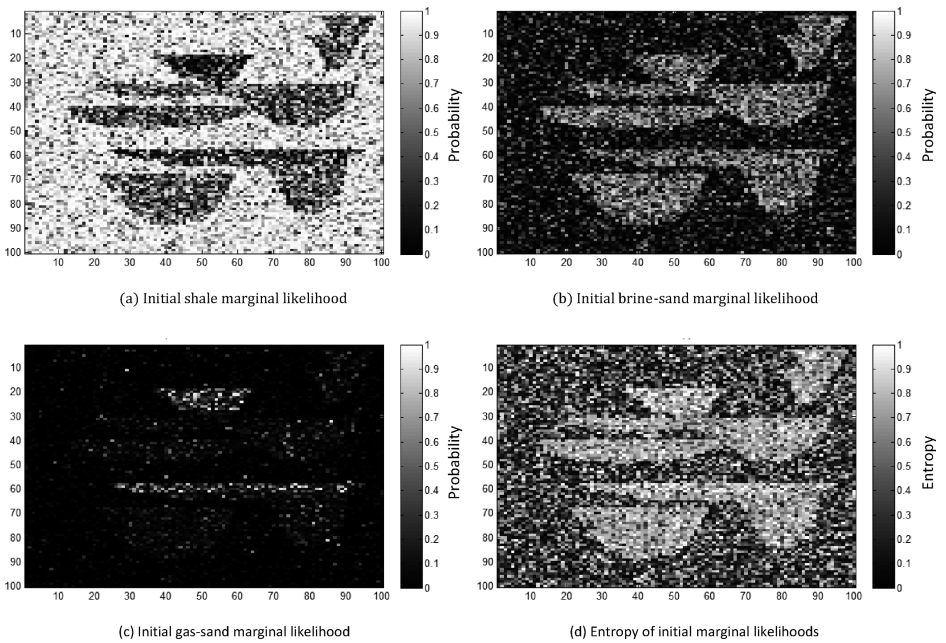
Figure 8. Synthetic (a) *P*- and (b) *S*-wave impedance attributes first sampled independently in each cell of the target cross-section in Fig. 7(b) using a probabilistic forward model based on a Gaussian distribution per facies with mean computed from the Yin–Marion shaly-sand rock physics model (Marion 1990; Yin *et al.* 1993; Avseth *et al.* 2005) and covariance matrix given in Table 2. The impedance sections thus obtained are then spatially filtered using the 5×5 banana-shaped kernel in eq. (48) to mimic blurring caused by non-localized effects of seismic data processing.

$N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k = f(\boldsymbol{m}_k)$ is given by the Yin–Marion shaly-sand model and $\boldsymbol{\Sigma}_k$ is the covariance matrix per facies given in Table 2 for the three facies considered. Collocated synthetic seismic attributes, *P*- and *S*-wave impedances $\boldsymbol{x}_i$, were then generated in each cell i from local facies responses $\boldsymbol{x}'_{\mathcal{N}_i}$ within the neighbourhood $\mathcal{N}_i$ of i in order to model the non-localized blurring effect of seismic imaging. This is achieved by using a Gaussian probabilistic forward model $\mathcal{P}(\boldsymbol{x}_i|\boldsymbol{z}_{\mathcal{N}_i}, \Theta) = N(\boldsymbol{R}_{\mathcal{N}_i}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$ from the design matrix $\boldsymbol{R}_{\mathcal{N}_i}$ of expected local responses of facies in each cell in $\mathcal{N}_i$ and the spatial filter $\boldsymbol{\beta}$ (see Section 4.2) which was chosen as a 5×5 banana-shaped kernel to represent the kind of blurring that may take place during seismic migration using a reference velocity model that is slightly too slow:

$$\beta = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0.125 & 0.125 & 0.125 & 0 \\ 0.0625 & 0.125 & 0 & 0.125 & 0.0625 \end{bmatrix} \tag{48}$$

and the resulting *P*- and *S*-wave impedances are shown in Fig. 8.

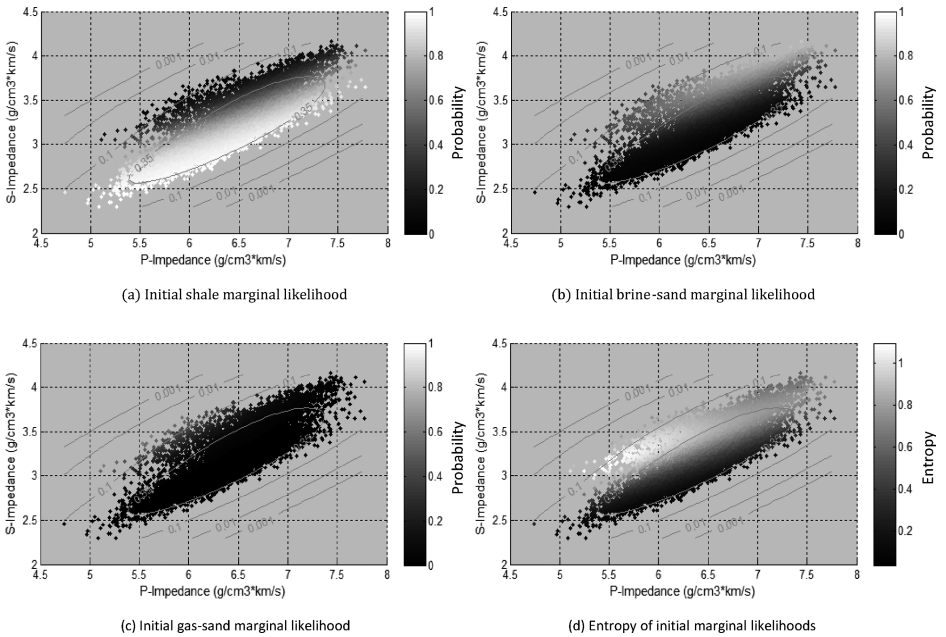
The seismic attributes and the model parameters were then inverted with the aim to reproduce the target cross-section. The initial estimates of parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k \in \mathcal{G}$ of the Gaussian mixture distribution were obtained using a mixture density neural-network (MDN; Meier *et al.* 2007a,b, 2009; Shahraeeni & Curtis 2011; Shahraeeni *et al.* 2012) based clustering of seismic attributes. In a real problem, estimates of these parameters may also be obtained from prior information based on well-logs or other data sources. The spatial filter $\boldsymbol{\beta}$ was initialized to a centred-spike with amplitude equal to 1 at the central element of the kernel while the rest of the elements were all set to 0. The initialization of $\boldsymbol{\beta}$ as a centred-spike effectively results in estimation of localized likelihoods $\mathcal{P}(\boldsymbol{x}_i \mid z_i)$ as a starting point before the parameters Θ (and hence $\boldsymbol{\beta}$) are updated during the M-step of the EM algorithm. The localized likelihoods were estimated from a GMM with components $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k \in \mathcal{G}$. Since the localized likelihoods are estimated only from the seismic attributes observed at the location of estimation, they are susceptible to noise in the data (Figs 9 and 10) and therefore do not abide by the geological plausibility rules of various facies configurations (such as gravitational ordering of fluids) and the conditional spatial distributions of facies depicted in the training image.



(a) Initial shale marginal likelihood (b) Initial brine-sand marginal likelihood

(c) Initial gas-sand marginal likelihood (d) Entropy of initial marginal likelihoods

Figure 9. Model space plots of initial cell-wise marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the initial estimates of parameters, and (d) entropy as a measure of uncertainty in the model under the initial likelihoods.



(a) Initial shale marginal likelihood (b) Initial brine-sand marginal likelihood

(c) Initial gas-sand marginal likelihood (d) Entropy of initial marginal likelihoods

Figure 10. Attribute space plots of initial cell-wise marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the initial estimates of parameters, and (d) entropy as a measure of uncertainty of the model under the initial likelihoods. Equidistant contours represent the initial Gaussian mixture distribution for the three facies.

The E-step of the EM algorithm estimates marginal posterior distributions in the model space using the LBP algorithm with the initial estimate of Θ in the first iteration. Contrary to the general practice of initializing the LBP algorithm with random or constant beliefs, we initialized it with the localized likelihoods estimated using a GMM with components $N(\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k),\quad k \in \mathcal{G}$ obtained from the Yin–Marion rock physics model. Such initialization of vertex beliefs with the estimated localized likelihoods allowed faster convergence. The parameters Θ were then updated in the attribute space during the M-step using the current estimate of posterior marginal distributions $\mathcal{P}_j(z_j)$ obtained from the E-step, as follows. The filter coefficients $\boldsymbol{\beta}$ were estimated using eq. (41) with the expected facies responses $\boldsymbol{r}_j$ at each location j computed using eq. (8). The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k,\ k \in \mathcal{G}$ were updated using eqs (42) and (43). The parameters updated during the M-step were then used in the E-step of the subsequent iteration until convergence. On convergence, the EM algorithm resulted in estimates of quasi-localized likelihoods that show a higher quality of facies discrimination (Figs 11 and 12), and the estimates of marginal posterior distributions $\mathcal{P}(\boldsymbol{z}_i|\boldsymbol{x},\ \hat{\Theta})$ for facies $\boldsymbol{z}$ in each model cell i (Figs 13 and 14) given the observed seismic attributes $\boldsymbol{x}$ and the final estimate

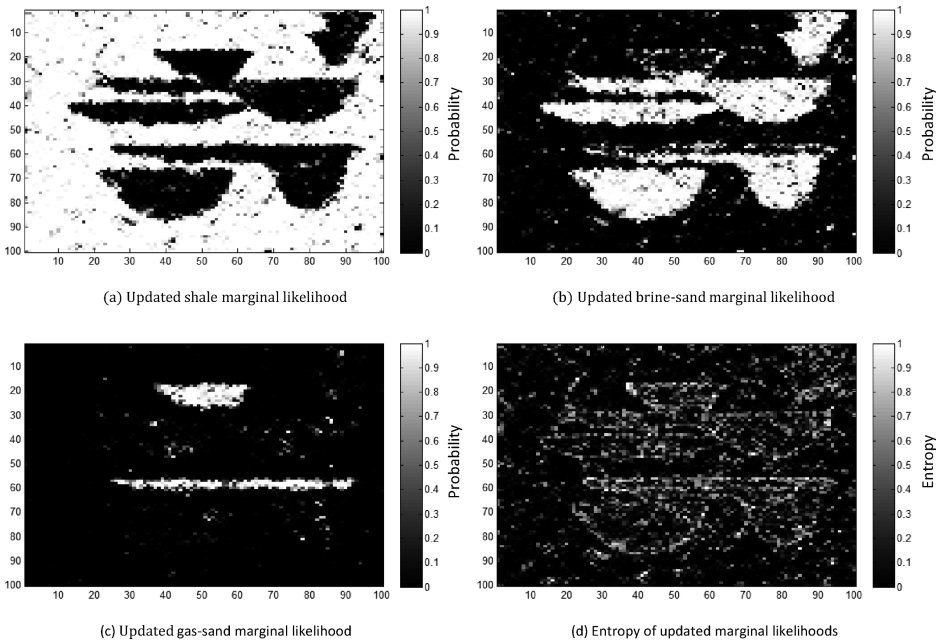


Figure 11. Model space plots of updated cell-wise quasi-localized marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the updated model parameters after running the EM algorithm and (d) normalized entropy as a measure of model uncertainty under the updated likelihoods.

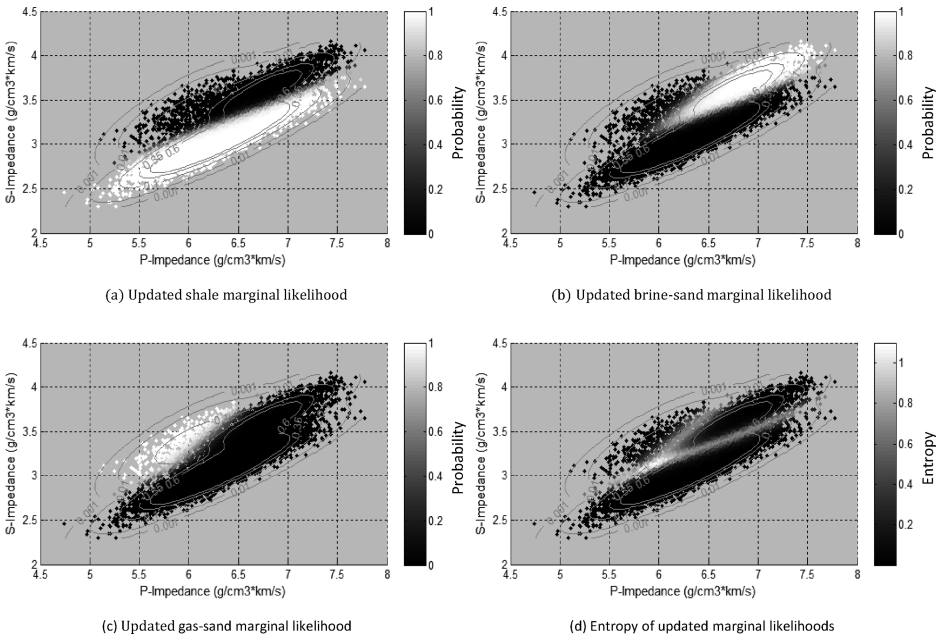


Figure 12. Attribute space plots of updated cell-wise quasi-localized marginal likelihoods of (a) shale, (b) brine-sand and (c) gas-sand computed from the updated model parameters after running the EM algorithm and (d) normalized entropy as a measure of the model uncertainty under the updated likelihoods. Equidistant contours represent the Gaussian mixture distribution for the three components (facies).

$\hat{\Theta}$ of parameters Θ, by incorporating both the prior information $\mathcal{P}(\boldsymbol{z})$ elicited from the training image (Fig. 7a) and the final estimates of non-localized likelihoods $\mathcal{P}(\boldsymbol{x}|\boldsymbol{z}, \hat{\Theta})$ (Figs 11a–c).

Fig. 15 shows the MAP estimate of the geological facies obtained from the max-product eq. (32) based LBP using the parameters updated by the EM algorithm. The MAP estimate matches quite reasonably with the 'true' geology (Fig. 7b). Figs 9,fig10 fig11 fig12 fig13 fig14.fig 9-14(d) show entropy of distributions in each of the corresponding figures (a–c) as a measure of uncertainty under the respective distributions. It is evident from these figures that the entropy reduces significantly starting with the entropy of the localized likelihoods in Figs 9(d) and 10(d) to the entropy in the marginal posterior distributions in Figs 13(d) and 14(d).

Although the prior information $\mathcal{P}(\boldsymbol{z})$ was formulated from the training image as spatial distributions between just two neighbouring locations at a time (so called 2-point statistics, or pairwise cliques), the approximate posterior distributions $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \hat{\Theta})$ estimated by LBP

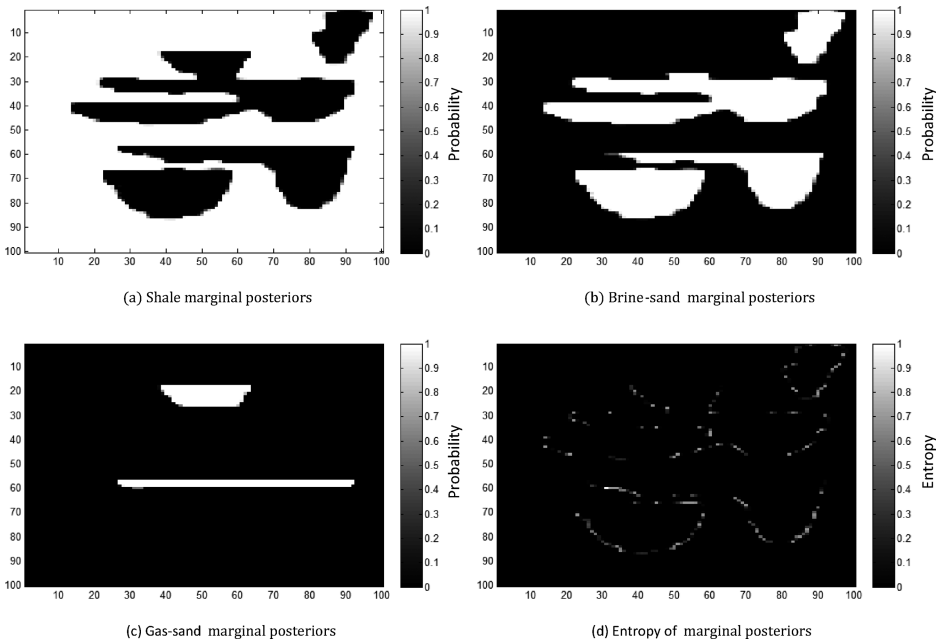


Figure 13. Model space plots of cell-wise marginal posterior distributions of (a) shale, (b) brine-sand and (c) gas-sand and (d) entropy as a measure of model uncertainty under the marginal posterior distributions.

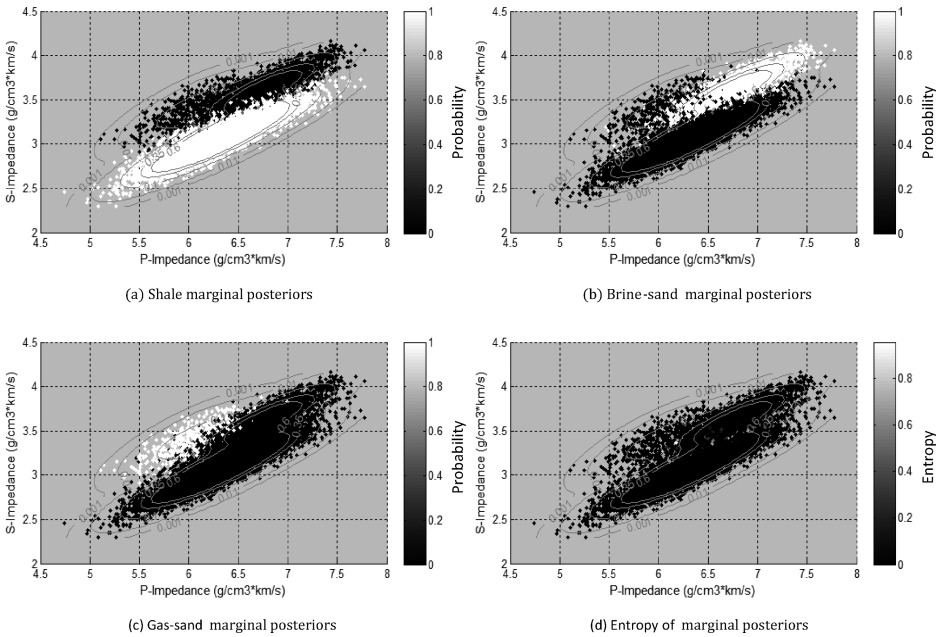


Figure 14. Attribute space plots of cell-wise marginal posterior distributions of (a) shale, (b) brine-sand and (c) gas-sand and (d) normalized entropy as a measure of model uncertainty under the marginal posterior distributions. Equidistant contours represent probability distributions of individual components of Gaussian mixture.

algorithm are reasonably close to the desired target distributions $\mathcal{P}(\boldsymbol{z}|\boldsymbol{x}, \Theta)$. This suggests that Bayesian inversion using non-localized likelihoods requires much less prior information about the conditional spatial distributions of facies to yield reliable estimates of posterior marginal distributions of facies. By contrast, the previous research (Walker & Curtis 2014; Nawaz & Curtis 2017) based on localized likelihoods used prior information extracted using larger templates in the form of joint distributions of facies over multiple points at a time from the same training image (for geological patterns of the same complexity). This is evident from Figs 9 and 11 since the localized likelihoods used in the first iteration of the EM algorithm are much noisier than the quasi-localized likelihoods estimated using parameters updated in the M-step. Our current algorithm can, however, be modified to incorporate the prior information from cliques of size greater than two. Such a modification is expected to allow the reconstruction of richer features observed in more complex geologies (we leave it for future research).

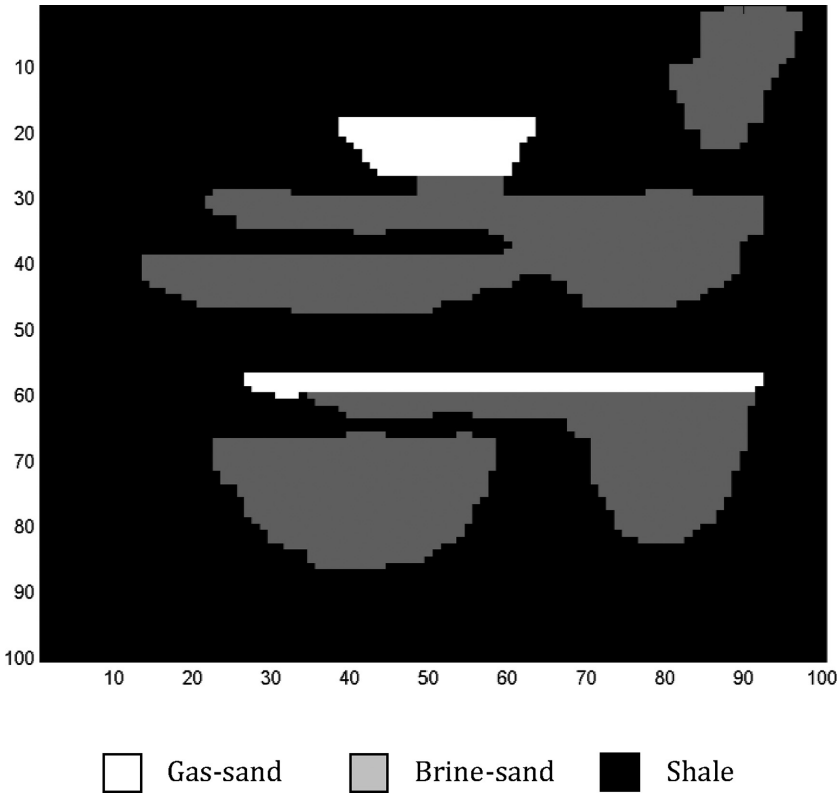


Figure 15. Model space plot of the inverted MAP estimate of facies in each of the model cell using the variational Bayesian inversion (VBI) showing a reasonable reconstruction of the target model (Fig. 7b).

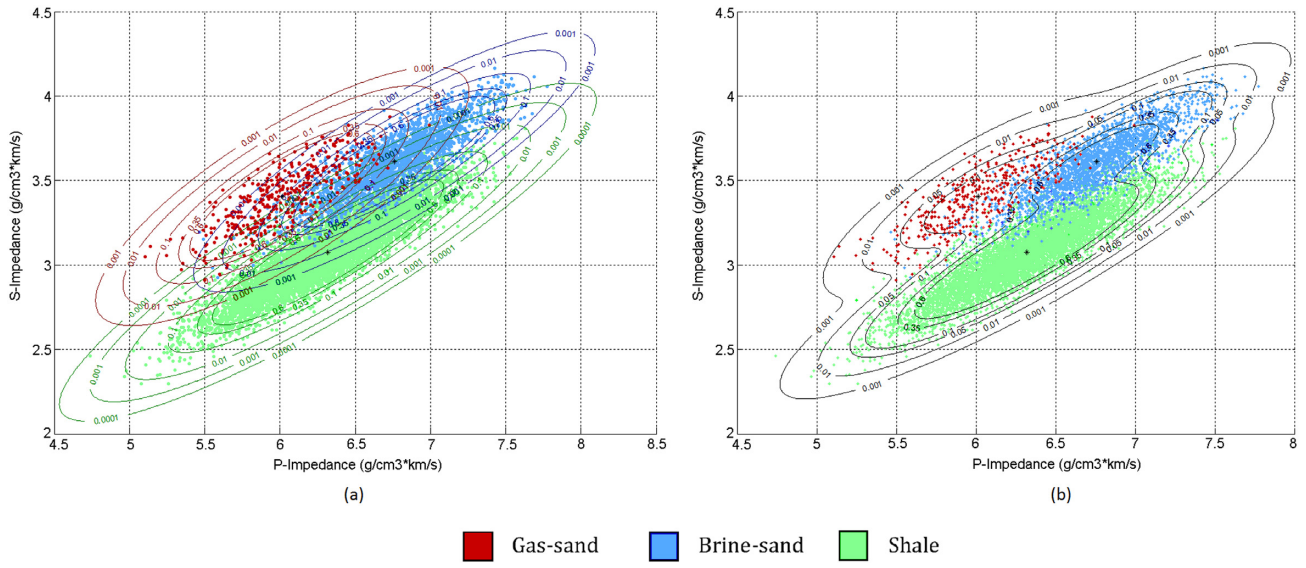


Figure 16. Attribute space plots of (a) components of the Gaussian mixture model and the seismic attributes colour-coded with the facies of maximum marginal posterior distributions, (b) The Gaussian mixture distribution obtained from a weighted sum of the Gaussian components per facies as displayed in (a).

It is also noteworthy that the marginal posterior distributions are updated in the model space during the E-step of the EM algorithm such that the spatial conditional distributions of various facies comply with those encapsulated in the training image. As a consequence of this, the model parameters are updated in the attribute space to reflect the inter-mixing of attributes (and overlap of their distributions) that are generated by different facies (Gaussian components) — see Fig. 16.

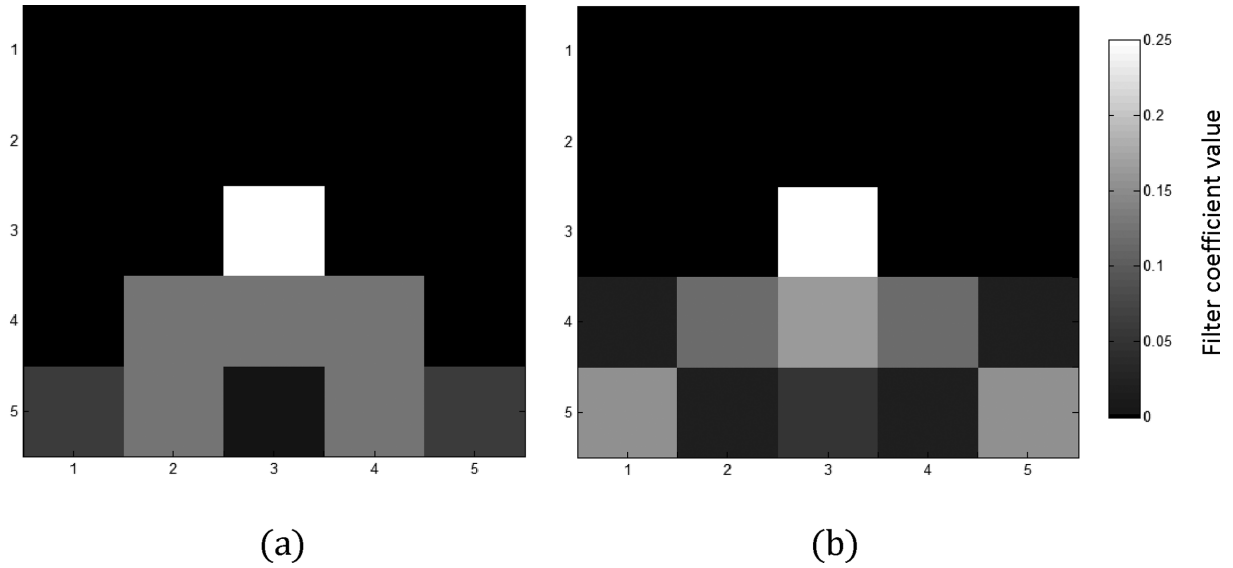


Figure 17. Comparison of (a) the spatial filter $\boldsymbol{\beta}$ used to blur the synthetic attributes and (b) the recovered spatial filter $\hat{\boldsymbol{\beta}}$. The amplitudes are scaled to a maximum value of 0.25 in both the plots.

The coefficients of estimated spatial filter $\hat{\boldsymbol{\beta}}$ were estimated from the M-step of the last iteration of the EM algorithm under the constraint that the resulting matrix is laterally symmetric (symmetric across columns). The estimated coefficients are shown below in the matrix form

$$\hat{\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 \\ 0.018 & 0.118 & 0.162 & 0.118 & 0.018 \\ 0.153 & 0.016 & 0.049 & 0.016 & 0.153 \end{bmatrix}. \tag{49}$$

Fig. 17 shows a comparison of the spatial filter $\boldsymbol{\beta}$ that was used to blur the seismic attributes and the estimated spatial filter $\hat{\boldsymbol{\beta}}$, both scaled to a maximum amplitude value of 0.25, showing that while not perfect, a reasonable estimate of the spatial blurring can be obtained.

8 COMPARISON WITH INVERSION METHODS THAT USE THE LOCALIZED LIKELIHOODS ASSUMPTION

The previously published methods of facies inversion from seismic attributes (e.g. Larsen *et al.* 2006; Ulvmoen & Omre 2010; Ulvmoen *et al.* 2010; Walker & Curtis 2014; Nawaz & Curtis 2017) assume that any spatial correlations present in the data (seismic attributes) are a direct consequence of, and therefore can be completely described by, the spatial distribution of facies as encoded in the prior information, that is, geological patterns depicted in the training image: this is the so called localized likelihoods assumption. In effect, these methods may not account for any spatial correlations present in the data due to other effects unrelated to the geology, such as those due to spatial blurring caused by processing-related artefacts and limited resolution of seismic data. Also, such methods do not make effective use of any spatial correlations in the data that are related to the local geology. We hypothesize that these methods have been successful to-date mainly because they rely on the prior information to reconstruct the spatial distribution of facies, and not the likelihoods. This hypothesis suggests that in the case that the prior information is limited (e.g. using small neighbourhood templates to scan the training image) or is inconsistent with the true geology (e.g. if geological patterns in the training image are not rich enough or are different from those present in the true subsurface), the localized likelihoods-based inversion methods may not reconstruct the spatial distribution of facies successfully. The quasi-localized likelihoods, on the other hand, complement the prior information by incorporating the spatial correlations present in the data within some neighbourhood of each location in the model. Our method based on quasi-localized likelihoods is therefore expected to be more robust against insufficient or incorrect prior information.

In order to test our hypothesis we used our synthetic test data (from Section 7) to compare our quasi-localized likelihoods-based algorithm with the previous algorithm of Nawaz & Curtis (2017) which is based on a 2-D-HMM and provides virtually analytic posterior marginal solutions using localized likelihoods. Henceforth this is referred to as the 'localized likelihoods-based method'. The comparison is made in terms of the quality of inverted posterior marginal distributions per facies when the data is spatially blurred (i.e. when the seismic attributes for various facies overlap significantly in the attribute space) and the amount of prior information is either limited or inconsistent with the true geology.

Fig. 18 shows such a comparison with respect to the amount of prior information used. The prior information on the spatial distribution of facies is extracted from the training image by scanning it with a 3×3 template and then supplied to the localized likelihoods-based

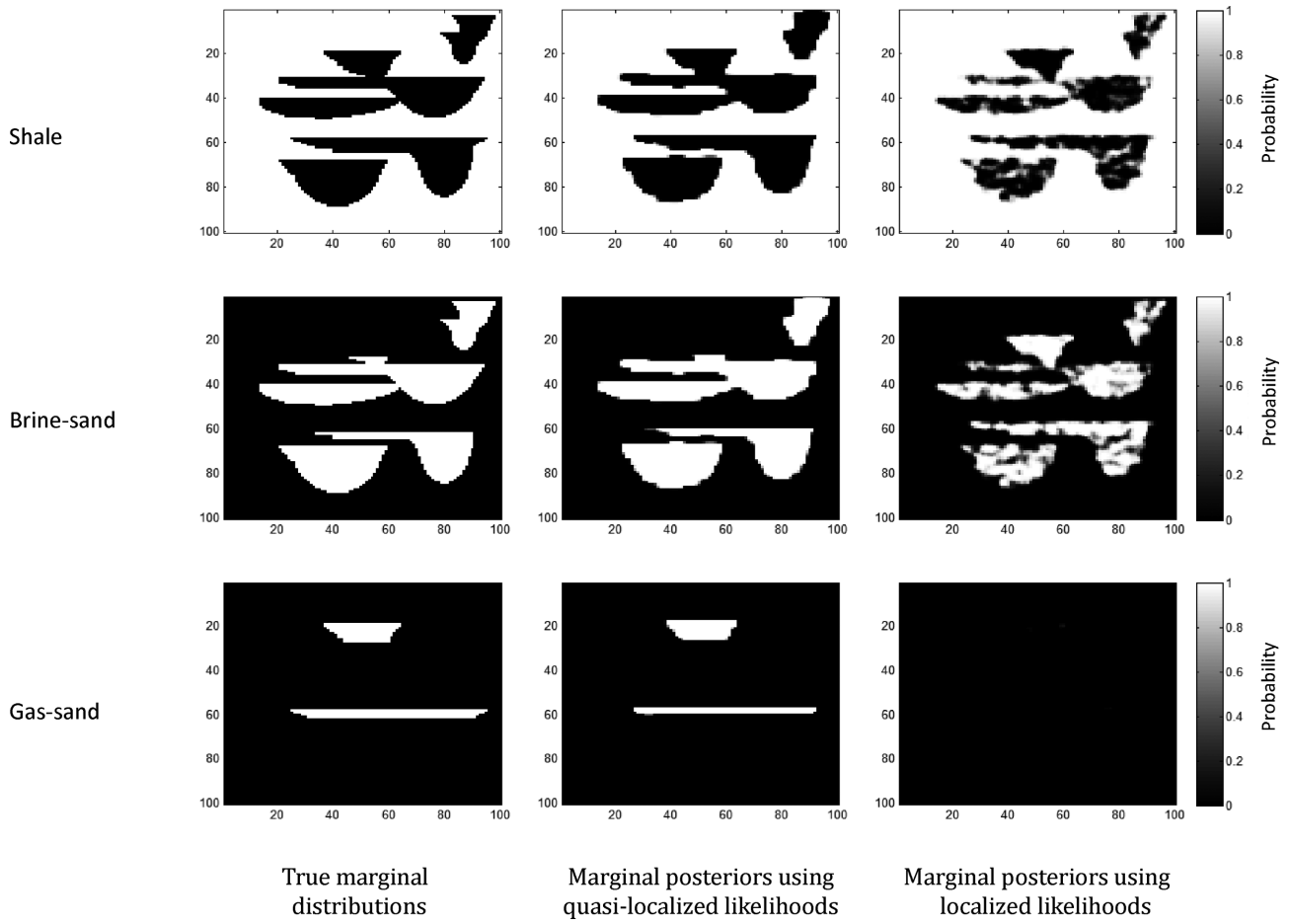


Figure 18. Model space plots of the inverted cell-wise marginal distributions per facies—shale, brine-sand and gas-sand in the order from the top to the bottom row: (left column) true marginal distributions in the synthetic model as in Fig. 7(b), (middle column) that obtained using our current method which is based on the quasi-localized likelihoods, and (right column) that obtained using the method of Nawaz & Curtis (2017) which solves the problem using the localized likelihoods assumption.

method of Nawaz & Curtis (2017). This corresponds to a clique size of 9, that is the prior information is encoded as a joint distribution over neighbouring vertices in a square matrix with three rows and three columns. In comparison, since our current method uses only pairwise cliques, it requires the prior information to be formulated as spatial distributions between just two neighbouring locations at a time. Even though our current methods uses much less prior information, it reconstructs the marginal posterior distributions quite reasonably as it uses the quasi-localized likelihoods which is a less stringent assumption compared to the localized likelihoods.

The results using the localized likelihoods-based method (right column in Fig. 18) show that this method could not discriminate between brine-sand and gas-sand and indeed failed to detect any gas-sand. Also, the reconstruction of the spatial distribution of shale and brine-sand is not as good as in our current method (middle column in Fig. 18). In this case, we found that if we increased the size of the prior template to 5×7, the localized likelihoods-based method can reconstruct the posterior marginal distributions just as reasonably as with quasi-localized likelihoods. This explains the previous success of methods that assumed localized likelihoods: they can work well with significantly non-localized data, but only if the prior information supplied is sufficiently strong to overcome the erroneous assumption.

Next, we generated synthetic seismic attributes as described in Section 7 except that the 'true' geology now contains dipping sand lenses (with no overbank deposits), in a hypothetical scenario where the stratum is tilted after lithification (Fig. 19). The same training image is used as in Fig. 7(a) with sand channels and overbank deposits with a background shale in an assumed horizontal stratum (i.e. without tilting). This allowed us to make a comparison between the two methods when the prior information supplied in the form of a training image is inconsistent with the true geology (Fig. 20). In this case, the prior information on the spatial distribution of facies is supplied to the localized likelihoods-based method of Nawaz & Curtis (2017) by using a 5×3 template. This corresponds to a clique size of 15, that is, the prior information is encoded as a joint distribution over neighbouring vertices in a rectangular matrix with five rows and three columns. The prior information for our current quasi-localized likelihoods-based method still comprises the joint distribution over just two neighbouring vertices. In this case, the localized likelihoods-based method fails to discriminate between shale and brine-sand, though the reconstruction of posterior marginal distributions of gas-sand is somewhat reasonable (right column in Fig. 20). Our current method based on quasi-localized likelihoods, however, reconstructs the posterior marginal distributions of all of the three facies quite well (middle column in Fig. 20) and therefore proves to be significantly more robust against incorrect prior information than localized likelihoods-based methods.

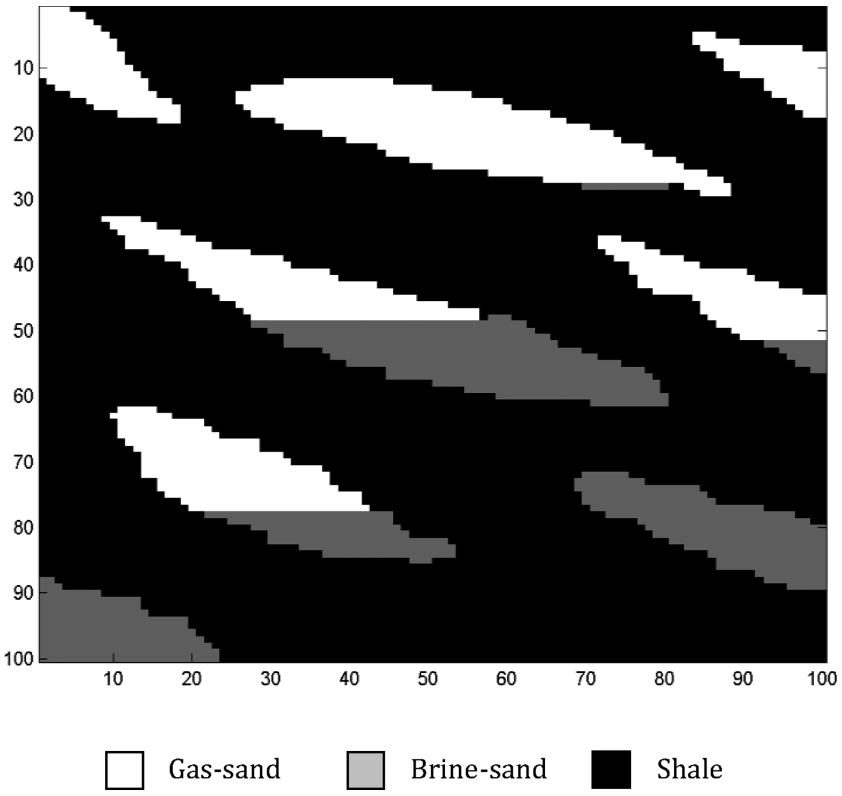


Figure 19. The target image representing the 'true' geological model consisting of dipping sand lenses (with no overbank deposits), in a hypothetical scenario where the stratum is tilted after lithification. This is the target for spatial facies inversion in the case that the prior information presented in the form of training image in Fig. 7(a) is inconsistent with this 'true' geological image.

The above comparisons show that the inversion methods based on the localized likelihoods assumption require well-informed priors: that is, the priors must be sufficiently informative to overcome errors due to the incorrect localized assumption, and must be consistent with the true geology. This means that the geological patterns depicted in a training image must be rich—diverse enough to include any possible facies patterns expected to be present in the subsurface. In other words, in order to use localized likelihoods-based methods we should have sufficient prior information about local geology to overcome errors caused by erroneous likelihoods. Our current method based on quasi-localized likelihoods, on the other hand, is expected to perform better even in the case that we do not have sufficient prior information, or that our prior information is only partially consistent with the true geology.

9 DISCUSSION

A major motivation of this research was to remove the localized-likelihoods assumption in the Bayesian inversion method of previous research by Larsen *et al.* (2006), Ulvmoen & Omre (2010), Ulvmoen *et al.* (2010), Walker & Curtis (2014) and Nawaz & Curtis (2017). The implication of this assumption is that the seismic attributes are assumed to be perfectly localized by processing, and by correcting the seismic data for any non-localized effects of wave propagation such as attenuation, Fresnel zone smearing, etc. (Nawaz & Curtis 2017). Such perfect localization can only be dreamed of because seismic data is band-limited in nature and seismic data processing involves uncertain models and various approximations in theory (e.g. ray theory as an approximation to wave theory; smooth and erroneous velocity models used for migration) and algorithms. As an example, seismic migration is never guaranteed to collapse the reflection Fresnel zone to a point in space, resulting in blurring or smearing in the image.

Localized likelihoods can only help to constrain the local presence or otherwise of facies and do not make use of any spatial correlations present in the data which may provide useful information about the geological heterogeneity and spatial conditional distributions of facies. By contrast, the current method derives the local presence of facies only from the likelihoods while the spatial distributions of facies are derived from both the prior information and the likelihoods. This suggests that the amount of prior information required for the reliable reconstruction of spatial distribution of facies that must be combined with localized likelihoods is significantly higher than with the quasi-localized likelihoods: prior information was expressed as a joint distribution of facies using a 3×3 template (9 cells) in the synthetic example of Walker & Curtis (2014), and from a 7×1 partition element in the synthetic example of Nawaz & Curtis (2017), both of which used localized likelihoods. In our synthetic example, the prior information is expressed as a joint distribution over just two cells in the form of pairwise edge potentials, and still the quality of resulting posterior distributions is noticeably better as compared to those of Walker & Curtis (2014) and similar to those of Nawaz & Curtis (2017). Further, it can be shown that the quality of facies discrimination with quasi-localized likelihoods

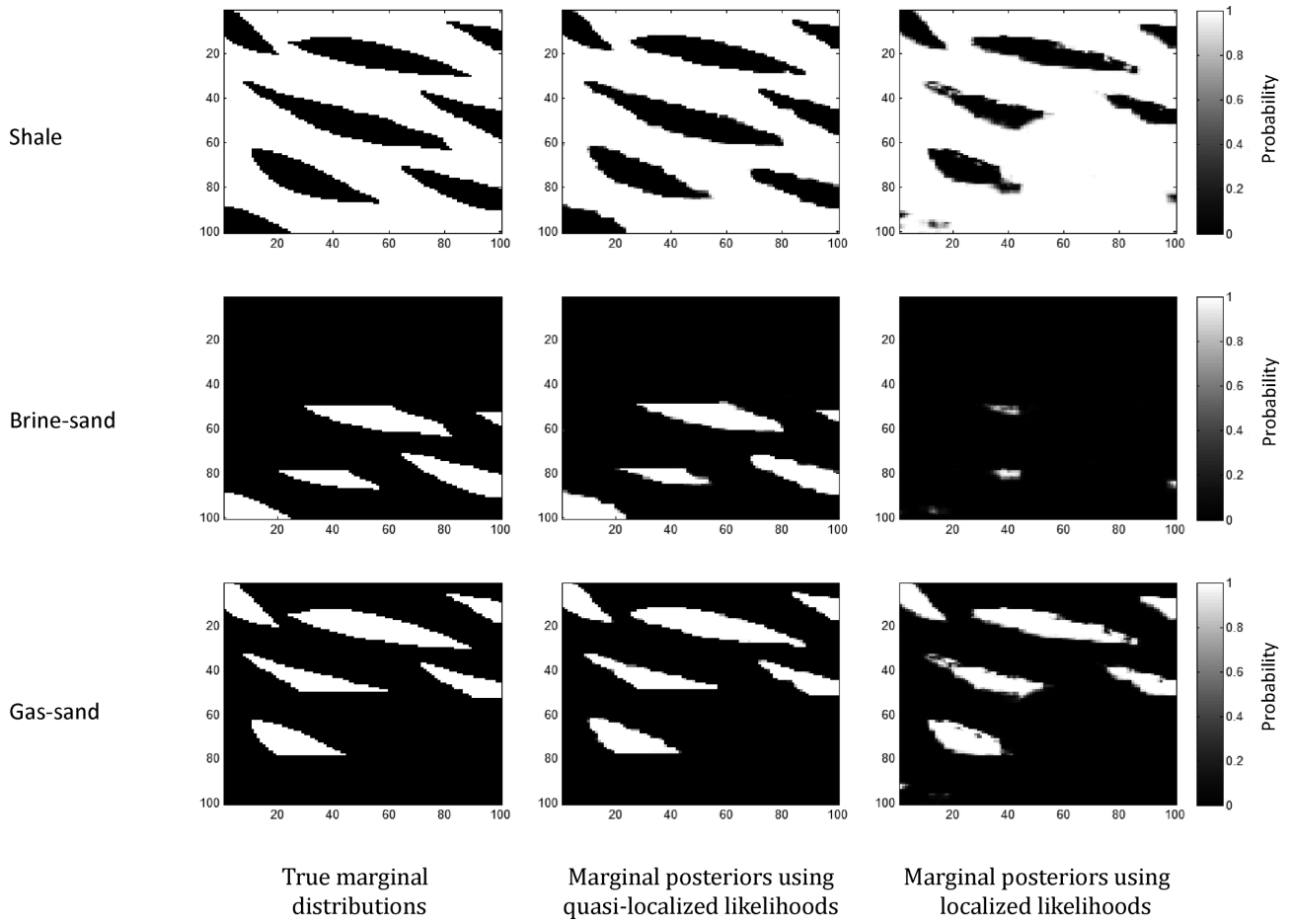


Figure 20. Model space plots of the inverted cell-wise marginal distributions per facies—shale, brine-sand and gas-sand in the order from the top to the bottom row: (left column) true marginal distributions in the synthetic model as in Fig. 7(b), (middle column) that obtained using our method as presented in this paper which is based on the quasi-localized likelihoods, and (right column) that obtained using the method of Nawaz & Curtis (2017) which is based on the localized likelihoods assumption.

(Fig. 11) is higher than with localized likelihoods (Fig. 9) since the seismic attributes contain spatial correlations and are not independent. The prior information further improves the discrimination and the spatial distribution of facies when combined with the quasi-localized likelihoods, for example ensuring that geologically implausible configurations (e.g. brine-sand directly overlaying gas-sand in some areas of Fig. 9) are disregarded in the computation of marginal posterior distributions (Fig. 13). Although we have not tested this explicitly, it is to be expected that the older methods that use localized likelihoods from other authors cited herein will have similar short-comings to those of our previous paper.

The localized likelihoods assumption was used in the previous research by Walker & Curtis (2014) and Nawaz & Curtis (2017) in order to address the computational intractability of mathematical inference in models with non-localized likelihoods. The current method evades such computational intractability by retaining the conditional independence assumption of seismic attributes given the facies, and resorting to an iterative optimization-based approximation (the EM algorithm) rather than an analytical approach as in Nawaz & Curtis (2017) for estimation of marginal posterior distributions of facies. This paper introduces the concept of quasi-localized likelihoods as a step towards methods that incorporate fully non-localized likelihoods, which is clearly a topic for future research.

A major challenge with any inference or parameter estimation based on the LBP algorithm is that there is no theoretical guarantee about the convergence of the LBP algorithm. This contrasts with McMC-based methods which are theoretically guaranteed to converge asymptotically. The empirical evidence, on the other hand, is very strong that LBP converges in most cases. The convergence of LBP depends on the topology of the graph as well as on the strength of potential functions. Strong and appropriate potential functions encode strong Lagrangian constraints that drive this algorithm towards faster and improved convergence. Koller & Friedman (2009) discussed many different possible reasons of non-convergence of LBP and their suggested remedies. In the situations when LBP fails to converge, it is observed that the non-convergence is either local or is due to oscillations in the beliefs. Koller & Friedman (2009) suggested using a dampening of the difference between two subsequent updates of beliefs as a remedy for oscillatory beliefs. If non-convergence is local, most of the beliefs converge except just a few. Averaged beliefs over a number of iterations may be used in case of local non-convergence. In either case, we recommend a careful examination of the problem before applying any such remedy. These problems may be caused by conflicts between the vertex and edge potentials which are likely to be caused by the presence of noise in the data, or by problematic parameters learnt during

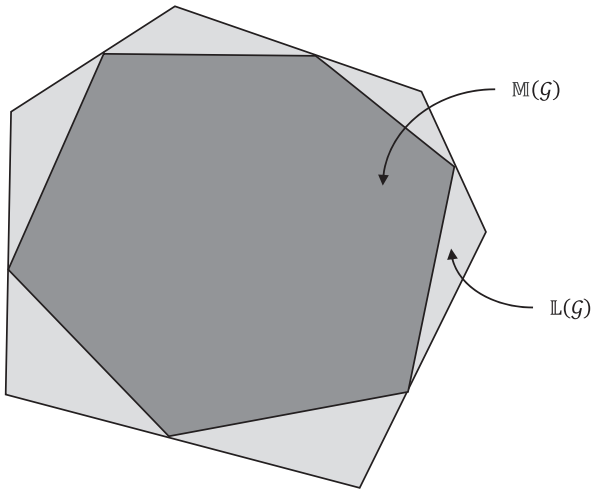


Figure 21. A graphical depiction of a marginal polytope $\mathbb{M}(\mathcal{G})$ for a graph $\mathcal{G}$ as a subset of local-consistency polytope $\mathbb{L}(\mathcal{G})$.

the M-step. Inappropriate parameters may lead to weak likelihoods. In such cases, either the input attributes must be properly conditioned or the parameters must be constrained through the use of an appropriate kernel function. Interested readers are recommended to consult Mooij & Kappen (2007) for a detailed account on the sufficient conditions for convergence of the LBP algorithm. Nevertheless, in contrast to McMC-based methods where it is impossible to detect convergence objectively, non-convergence in LBP is always detectable (see Algorithm 1).

An important consideration regarding LBP is that it may converge to a stationary point of Bethe's free energy other than the global minimum. This can be illustrated as follows. If $\mathcal{G}$ represents the finite state space of geological facies and $|\mathcal{G}|$ represents its size (i.e. the number of geological facies considered), the set $\mathbb{M}(\mathcal{G})$ of realizable marginals τ of some distribution $\mathcal{Q}(\boldsymbol{z}|\boldsymbol{x})$ is defined as

$$\mathbb{M}\left(\mathcal{G}\right) \equiv \left\{\tau \in \mathbb{R}^{d} : \exists\, \mathcal{Q}\left(\boldsymbol{z}|\boldsymbol{x}\right),\ \tau_{i}\left(z_{i}\right), \tau_{ij}\left(z_{i}, z_{j}\right)\ \forall\, i, j \in \mathcal{V}\ \wedge\ (i, j) \in \mathcal{E}\right\}, \tag{50}$$

where $d\ = |\mathcal{G}|\ (|\mathcal{V}| + |\mathcal{E}| \cdot |\mathcal{G}|)$. The set $\mathbb{M}(\mathcal{G})$ is commonly referred to as the *marginal polytope* of the graph $\mathbb{G}$ (Wainwright & Jordan 2008). The beliefs $b_i(z_i)$ and $b_{ij}(z_i,\ z_j)$ in eq. (27) are constrained through the Lagrangian to be locally consistent in order to ensure that they represent proper marginal distributions. A set $\mathbb{L}(\mathcal{G})$ of locally consistent beliefs is defined as

$$\mathbb{L}\left(\mathcal{G}\right) \equiv \left\{ b \in \mathbb{R}^{d} :\ \begin{array}{c} b_{i}\left(z_{i}\right) \geq 0,\ \ \sum_{z_{i}} b_{i}\left(z_{i}\right) = 1, \forall\, i \in \mathcal{V} \\ \sum_{z_{j}} b_{ij}\left(z_{i}, z_{j}\right) = b_{i}\left(z_{i}\right),\ \forall\ (i, j) \in \mathcal{E} \end{array} \right\}, \tag{51}$$

which represents a polytope, commonly referred to as the *local-consistency polytope*, defined by $\mathcal{O}(|\mathcal{V}| + |\mathcal{E}|)$ constraints. Any arbitrary set of locally consistent beliefs $b_i(z_i)$ and $b_{ij}(z_i,\ z_j)$ may not be jointly realizable by some distribution $\mathcal{Q}$; global consistency is not guaranteed to ensure the existence of a joint distribution $\mathcal{Q}$ that corresponds to these beliefs. Its converse, however, is true: all jointly realizable beliefs are locally consistent. This implies that $\mathbb{M}(\mathcal{G}) \subseteq \mathbb{L}(\mathcal{G})$, where equality holds only for tree-structured graphs (Fig. 21). As a consequence of this, even if LBP converges, it may converge to different stationary points of the Bethe's free-energy other than the global minimum. This problem may be addressed by heuristic initialization and multiple restarts as suggested by Koller & Friedman (2009).

Another important consideration in probabilistic inversion is the ability to draw stochastic realizations (samples) from the posterior distribution. Methods that are based on McMC algorithm (e.g. Larsen *et al.* 2006; Ulvmoen & Omre 2010; Ulvmoen *et al.* 2010; Hammer & Tjelmeland 2011; Rimstad & Omre 2013; Lindberg & Omre, 2014, 2015) are computationally demanding as they generate samples of the model from full joint posterior distributions and estimate marginal posterior distributions from these realizations. However, the comparative advantage of such methods is that they provide samples that may be used to perform any desired inference that cannot be performed directly from the marginal posterior distributions alone. The recursive algorithms of Walker & Curtis (2014) and Nawaz & Curtis (2017) provide alternatives to McMC-based methods, and they also allow drawing samples from computed probabilities. Nawaz & Curtis (2017) presented a method to draw samples from the marginal posterior distributions using a copula function which encapsulates conditional spatial distributions of facies as stored in the prior information presented in the training image. Since their method assumes localized likelihoods, the spatial distributions of facies were only incorporated from the prior information, and not from the data. A key feature of our method is that it focuses on estimating the posterior marginal distributions (as in Nawaz & Curtis 2017) instead of the full joint posterior distribution as in Walker & Curtis (2014), and the previous research that uses the McMC algorithm because the latter is intractable for large-scale models. Such an approach in its current form sacrifices the ability to draw samples in favour of computational efficiency and provides the most commonly desired marginal posterior distributions of facies. Even though our current algorithm estimates marginal posterior distributions, samples cannot be drawn from them directly using the copula function-based approach of Nawaz & Curtis (2017) as the samples drawn in this manner would not incorporate the spatial correlations observed in the data, which would make relaxation of the localized likelihoods assumption in

our method pointless. A modification to the copula function-based sampling is therefore needed that also incorporates the spatial correlations observed in the data—another topic for future research.

In comparison to the McMC method that has been used to solve all sorts of inverse problems in various fields of research because of its general applicability, the variational Bayes method requires analytical derivations or numerical algorithms that are specific to the problem in question. The variational form of the EM algorithm as used in this paper is expected to offer a significant step towards generalization of the VBI for solving problems which specifically involve a spatial grid of observed data that are collocated with the unknown model parameters. Our method can be extended further to invert for continuous variables (such as rock properties from seismic waveform data) in spatial inverse problems. We leave such an extension for future research.

Since our method uses a pairwise MRF as the spatial model for the distribution of facies, we anticipate that it may not be so capable of reconstructing complex spatial patterns of geological facies (e.g. those found in aerial view of intersecting sand channels in a deltaic environment). Multipoint statistics-based simulation (Strebelle 2001; Caers & Zhang 2004; Arpat 2005; Journel & Zhang 2006; Mariethoz & Caers 2014) and related stochastic inversion methods have been developed for such cases (Haas & Dubrule 1994; Francis 2005; Nunes *et al.* 2017). We have not tested the current method for such a case. However, a general MRF with higher order cliques would be required in this case which is an extension of our current model, and we leave this for future research.

10 CONCLUSIONS

We presented a Bayesian method for inversion of geological facies from seismic attributes under the variational approximation as a computationally efficient alternative to the commonly used McMC-based methods. In addition, our method also allows for reliable detection of convergence, in contrast to the McMC-based spatial inversion methods which are known to have difficulties with detection of convergence. Geological facies and seismic attributes are considered as latent and observed variables, respectively, in Bayesian inversion. The prior information is presented in the form of a training image that encodes the spatial distribution of facies according to some geological process model or subjective opinion of a geoscientist. The prior spatial distribution of facies is modelled as an MRF. The likelihoods are assumed to have a Gaussian distribution with expectations at a location given by a linear combination of local facies responses within the neighbourhood of that location. We termed the likelihoods estimated in this manner as quasi-localized likelihoods which refer to a relaxation of the assumption of localized likelihoods as was generally used in previous research. The seismic attributes are assumed to be conditionally independent given the geological facies and are assumed to be distributed as a Gaussian mixture distribution with number of components given by the number of facies considered. We also showed that the quasi-localized likelihoods define a spatial Gaussian mixture distribution for seismic attributes observed at a location given the facies at the neighbouring locations, whose parameters are estimated while acknowledging the spatial nature of seismic attributes. Although the prior distribution is modelled as an MRF, we showed that by virtue of the conditional independence assumption on seismic attributes, the joint and hence the posterior distribution of facies given seismic attributes also represents an MRF.

We used a Bayesian approach to jointly estimate the posterior marginal distributions of facies in the model space, and of the model parameters in attribute space, using a variational form of EM algorithm. The EM algorithm performs inference on facies and model parameters in an iterative fashion by alternately estimating the approximate marginal posterior distributions of facies from the current estimate of model parameters in the so-called E-step, and maximizing the expectation of log-likelihood of the model parameters from the current estimate of marginal posterior distributions of facies in the so-called M-step. We used the LBP algorithm to estimate marginal posterior distributions of facies in the E-step, and solved the linear model to obtain estimates of the spatial filter that maximizes the likelihood of observing seismic attributes given the current estimate of posterior marginal distributions of facies in the M-step of the EM algorithm. The EM algorithm is guaranteed not to decrease the log-likelihood of the observed variables (seismic attributes) given current estimate of model parameters at any iteration. Thus, in essence, the VBI performs inference on the latent variables by solving a constrained optimization problem.

We compared our method with the previous methods of facies inversion from seismic attributes that are based on the assumption of localized likelihoods using a synthetic data example. It shows that our current method requires far less prior information to reconstruct an accurate estimate of the true marginal posterior distributions of facies in the subsurface as compared to our previous inversion method (Nawaz & Curtis, 2017) that is based on the localized likelihoods assumption. Also we showed that our current method is more robust against incorrect prior information.

ACKNOWLEDGEMENTS

We are thankful to TOTAL UK for their sponsorship of this research. We would also like to show our gratitude to Mohammed Shahraeeni and Constantin Gerea of TOTAL UK for providing useful suggestions and sharing their expertise during the course of this research. We are also grateful to Prof. Klaus Mosegaard and an anonymous reviewer for their valuable comments on an earlier version of this manuscript.

REFERENCES

Arpat, G.B., 2005. *Sequential simulation with patterns, PhD thesis,* Stanford University.

Avseth, P., Mukerji, T. & Mavko, G., 2005. *Quantitative Seismic Interpretation,* Vol. **1,** Cambridge Univ. Press.

Bachrach, R., 2006. Joint estimation of porosity and saturation using stochastic rock-physics modelling, *Geophysics,* **71**(5), O53–O63.

Balakrishnan, S., Wainwright, M.J. & Yu, B., 2017. Statistical guarantees for the EM algorithm: from population to sample-based analysis, *Ann. Stat.,* **45**(1), 77–120.

Beal, M.J. 2003, Variational algorithms for approximate Bayesian inference. *PhD thesis,* Gatsby Computational Neuroscience Unit, University College London.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems, *J. R. Stat. Soc. B,* **36**(2), 192–236.

Bethe, H. 1935. Statistical theory of superlattices, *Proc. R. Soc. A,* **150**(871), 552–575 .

Bosch, M., Mukerji, T. & Gonzalez, E.F., 2010. Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: a review, *Geophysics,* **75**(5), 75A165–75A176.

Buland, A. & Omre, H., 2003. Bayesian linearized AVO inversion, *Geophysics,* **68**(1), 185–198.

Caers, J. & Zhang, T., 2004. Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models, in *Integration of Outcrop and Modern Analogs in Reservoir Modeling (AAPG Memoir),* pp. 384–394, eds. Grammer, G.M., Harris, P.M. & Eberli, G.P., American Association of Petroleum Geologists.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions and the curse of dimensionality, *Geophysics,* **66,** 372–378

Dempster, A.P., Laird, N.M. & Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. B,* **39,** 1–38.

Doyen, P.M., Guidish, T.M. & de Buyl, M.H., 1989. Monte Carlo simulation of lithology from seismic data in a Channel-Sand reservoir, *Soc. Pet. Eng.,* SPE Annual Technical Conference and Exhibition, 8-11 October, San Antonio, Texas, SPE paper # 19588.

Eddy, S.R., 1998. Profile hidden markov models, *Bioinformatics*. **14,** 755–763.

Feynman, R.P., 1972. *Statistical Mechanics: A Set of Lectures,* W. A. Benjamin.

Francis, A., 2005. Limitations of deterministic and advantages of stochastic inversion, *CSEG Recorder,* **30**(2), 5–11.

Gallagher, K., Charvin, K., Nielsen, S., Sambridge, M. & Stephenson, J., 2009. Markov chain Monte Carlo (McMC) sampling methods to determine optimal models, model resolution and model choice for Earth science problems, *Mar. Pet. Geol.,* **26**(4), 525–535.

Grana, D. & Della Rossa, E., 2010. Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion, *Geophysics,* **75**(3), O21–O37.

Grana, D., Fjeldstad, T. & Omre, H., 2017. Bayesian Gaussian mixture linear inversion for geophysical inverse problems. *Math. Geosci.,* **49,** 493–515,

Grana, D., Lang, X. & Wu, W., 2016. Statistical facies classification from multiple seismic attributes: comparison between Bayesian classification and expectation–maximization method and application in petrophysical inversion, *Geophys. Prospect.,* **65**(2), 544–562.

Grana, D. & Mukerji, T., 2015. Bayesian inversion of time-lapse seismic data for the estimation of static reservoir properties and dynamic property changes, *Geophys. Prospect.,* **63**(3), 637–655.

Grana, D., Mukerji, T., Dvorkin, J. & Mavko, G., 2012. Stochastic inversion of facies from seismic data based on sequential simulations and probability perturbation method, *Geophysics,* **77**(4), M53–M72.

Haas, A. & Dubrule, O., 1994. Geostatistical inversion—a sequential method of stochastic reservoir modeling constrained by seismic data, *First Break,* **12,** 561–569.

Hammer, H.L. & Tjelmeland, H., 2011. Approximate Forward–Backward algorithm for a switching linear Gaussian model. *Comput. Stat. Data Anal.* **55**(1), 154–167, doi:10.1016/j.csda.2010.06.008.

Hammersley, J.M. & Clifford, P., 1971. Markov fields on finite graphs and lattices, unpublished work.

Journel, A. & Zhang, T., 2006. The necessity of a multiple-point prior model, *Math. Geol.,* **38**(5), 591–610.

Koc, C.K. & Piedra, R.M., 1991. A parallel algorithm for exact solution of linear equations, In *Proceedings of International Conference on Parallel Processing,* Vol. **III,** Charles, S., ed., CRC Press, Boca Raton, FL, pp. 1–8.

Koller, D. & Friedman, N., 2009. *Probabilistic Graphical Models: Principles and Techniques* , MIT Press.

Larsen, A.L., Ulvmoen, M., Omre, H. & Buland, A., 2006. Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model, *Geophysics,* **71**(5), R69–R78.

Lindberg, D. & Omre, H., 2014. Blind categorical deconvolution in two level hidden Markov models, *IEEE Trans. Geosci. Remote Sens.,* **52**(11), 7435–7447.

Lindberg, D. & Omre, H., 2015. Inference of the transition matrix in convolved hidden Markov models and the generalized Baum-Welch algorithm, *IEEE Trans. Geosci. Remote Sens.,* **53**(12), 6443–6456.

Lindberg, D.V., Rimstad, E. & Omre, H., 2015. Inversion of well logs into facies accounting for spatial dependencies and convolution effects, *J. Pet. Sci. Eng.,* **134,** 237–246.

Mariethoz, G. & Caers, J., 2014. *Multiple-point Geostatistics: Stochastic Modeling with Training Images,* Wiley-Blackwell.

Marion, D.P., 1990. Acoustical, mechanical, and transport properties of sediments and granular materials, *PhD thesis,* Department of Geophysics, Stanford University.

Meier, U., Curtis, A. & Trampert, J., 2007a. Fully nonlinear inversion of fundamental mode surface waves for a global crustal model, *Geophys. Res. Lett.,* **34,** L16304, doi:10.1029/2007GL030989.

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

Meier, U., Trampert, J. & 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.

Mooij, J.M. & Kappen, J.H., 2007. Sufficient conditions for convergence of the sum-product algorithm, *IEEE Trans. Inf. Theory,* **53**(12), 4422–4437.

Mukerji, T., Jørstad, A., Avseth, P., Mavko, G. & Granli, J.R., 2001. Mapping lithofacies and pore-fluid probabilities in a North Sea reservoir: seismic inversions and statistical rock physics, *Geophysics,* **66,** 988–1001.

Murphy, K.P., Weiss, Y. & Jordan, M.I., 1999. Loopy belief propagation for approximate inference: an empirical study, *Proceedings of the fifteenth conference on uncertainty in artificial intelligence,* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 467–475.

Nawaz, M.A. & Curtis, A., 2017. Bayesian inversion of seismic attributes for geological facies using a hidden Markov model, *Geophys. J. Int.,* **208,** 1184–1200.

Nunes, R., Soares, A., Azevedo, L. & Pereira, P., 2017. Geostatistical seismic inversion with direct sequential simulation and co-simulation with multi-local distribution functions, *Math. Geosci.,* **59**(5), 583–601.

Pearl, J., 1982. Reverend Bayes on inference engines: a distributed hierarchical approach, *Proceedings of the Second National Conference on Artificial Intelligence (AAAI'82),* AAAI Press, Menlo Park, CA, pp. 133–136.

Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* 2nd edn, Morgan Kaufmann Publishers Inc.

Rencher, A.C., 2002. *Methods of Multivariate Analysis,* 2nd edn, Wiley-Interscience.

Rimstad, K., Avseth, P. & Omre, H., 2012. Hierarchical Bayesian lithology/fluid prediction: a North Sea case study, *Geophysics,* **77**(2), B69–B85.

Rimstad, K. & Omre, H., 2010. Impact of rock-physics depth trends and Markov random fields on hierarchical Bayesian lithology/fluid prediction, *Geophysics,* **75**(4), R93–R108.

Rimstad, K. & Omre, H., 2013. Approximate posterior distributions for convolutional two-level hidden Markov models, *Comput. Stat. Data Anal.,* **58**(1), 187–200.

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

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.

Sinoquet, C. & Mourad, R., 2014. *Probabilistic Graphical Models for Genetics, Genomics and Postgenomics,* Oxford Univ. Press, pp. 480.

Strebelle, S., 2001. Conditional simulation of complex geological structures using multiple-point statistics, *Math. Geol.,* **34**(1), 1–21.

Sudderth, E. & Freeman, W., 2008. Signal and image processing with belief propagation, *IEEE Signal Proc. Mag.,* **25**(2), 114–141.

Tatikonda, S.C. & Jordan, M.I., 2002. Loopy belief propagation and Gibbs measures, in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI'02),* eds Darwiche, A. & Friedman, N., Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 493–500.

Ulvmoen, M. & Omre, H., 2010. Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations, Part 1 — methodology, *Geophysics,* **75**(2), R21–R35.

Ulvmoen, M., Omre, H. & Buland, A., 2010. Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations: Part 2 — real case study, *Geophysics,* **75**(2), B73–B82.

Wainwright, M.J. & Jordan, M.I., 2008. Graphical models, exponential families, and variational inference, *Found. Trends Mach. Learn.,* **1**(1–2), 1–305.

Walker, M. & Curtis, A., 2014. Spatial Bayesian inversion with localized likelihoods: an exact sampling alternative to McMC, *J. Geophys. Res. Solid Earth,* **119,** 5741–5761.

Wang, H., Wellmann, J.F., Li, Z., Wang, X. & Liang, R.Y., 2016. A segmentation approach for stochastic geological modeling using hidden Markov random fields, *Math. Geosci.* **49**(2), 145–177.

Yasuda, M., Kataoka, S. & Tanaka, K., 2015. Statistical analysis of loopy belief propagation in random fields, *Phys. Rev.,* **92,** 042120, doi:10.1103/PhysRevE.92.042120.

Yedidia, J.S., Freeman, W.T. & Weiss, Y., 2001a. Bethe free energy, Kikuchi approximations and belief propagation algorithms, Technical report, Mitsubishi Electric Res. Labs, TR-2001-16.

Yedidia, J.S., Freeman, W.T. & Weiss, Y., 2001b. Understanding belief propagation and its generalizations, Technical report, Mitsubishi Electric Res. Labs., TR-2001-15.

Yin, H., Nur, A. & Mavko, G., 1993. Critical porosity a physical boundary in poroelasticity, *Int. J. Rock Mech. Min. Sci. Geomech. Abstr.,* **30**(7), 805–808.

Zhao, B., Zhong, Y., Ma, A. & Zhang, L., 2016. A spatial Gaussian mixture model for optical remote sensing image clustering, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.,* **9,** 5748–5759.