JGR Solid Earth

RESEARCH ARTICLE
10.1029/2018JB016652

Rapid Discriminative Variational Bayesian Inversion of Geophysical Data for the Spatial Distribution of Geological Properties

M. A. Nawaz[1] and **A. Curtis**[1,2]

[1]School of Geosciences, Grant Institute, University of Edinburgh, Edinburgh, UK, [2]Exploration and Environmental Geophysics Group, ETH Zürich, Zürich, Switzerland

Key Points:
- Our fully nonlinear Bayesian inversion method requires no assumptions of localization or conditional independence of data
- The discriminative approach ensures that the method can be applied using supervised machine learning
- The method performs fully nonlinear probabilistic inference while avoiding Markov chain Monte Carlo sampling

Supporting Information:
- Supporting Information S1

Correspondence to:
M. A. Nawaz,
muhammad.atifnawaz@ed.ac.uk

Citation:
Nawaz, M. A., & Curtis, A. (2019). Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties. *Journal of Geophysical Research: Solid Earth*, *124*, 5867–5887. https://doi.org/10.1029/2018JB016652

Received 8 SEP 2018
Accepted 29 APR 2019
Accepted article online 3 MAY 2019
Published online 27 JUN 2019

Abstract We present a new, fully probabilistic and nonlinear inversion method to estimate the spatial distribution of geological properties (depositional facies, diagenetic rock types, or other rock properties) from geophysical data (e.g., seismic data). Contrary to the conventional generative approach that models solution probabilities via the likelihood of observed data, our method uses a discriminative approach that directly models the posterior distribution of the geological properties given the data. This reduces the modeling effort significantly and allows machine learning algorithms such as neural networks to be deployed to solve large geophysical inference problems. We show that our method honors spatial distributions of geological parameters supplied as prior information about local geology and can be trained using supervised learning to be robust against noise present in the data as long as we can provide statistical characteristics of the noise. Exact Bayesian inference is almost always infeasible in practice because it requires normalization of the posterior distribution; this is intractable for large models and must therefore be approximated. Most existing probabilistic inversion methods use stochastic sampling (e.g., Markov chain Monte Carlo, McMC) for approximate inference. However, McMC involves the use of subjective criteria to detect convergence. We use the variational Bayes method to transform probabilistic inference into numerical optimization. This is a more efficient, deterministic alternative to McMC-based inference for suitably structured problems. Our method thus avoids extensive sampling during inference, yet provides fully probabilistic Bayesian results, and is therefore scalable to higher dimensional problems.

Plain Language Summary We present a new method for the estimation of geological properties such as type and physical properties of rocks, from geophysical measurements such as seismic data. Most existing methods assume that the geophysical data have been perfectly localized to produce data at each point in space (e.g., through tomography or imaging) and that the data are free of correlated noise or errors. Although neither requirement is met in reality, existing methods use these assumptions to make solutions computationally tractable. Our method removes both of these assumptions and is still computationally tractable for suitably structured problems (a class of problems that can be decomposed into interlinked subproblems). We achieve this by abandoning the usual approach of modeling the likelihood—a measure of how probable it is that the observed data were generated by any given geological model. Instead, our method models the geological parameters from the observed data directly, using examples of the direct data-model relationship. This reduces the required computational resources significantly for large-scale problems. To further improve computational efficiency, our method avoids extensive use of Monte Carlo sampling, and instead uses numerical optimization to estimate the desired geological properties and their fully probabilistic uncertainties.

1. Introduction

This paper presents a new way to invert geophysical (or any spatial) data for models of the probabilistic distribution of subsurface geological parameters such as rock types (facies) or their physical properties. In the natural world, geological parameters at neighboring locations are more likely to be similar than those at distant locations. The spatial context in geology thus induces a higher degree of correlation between data observed at neighboring locations compared to data observed at distant locations. Such probabilistic dependence between geology and data across space is an example of geological prior information, which is incorporated into solutions in order to improve the quality of geological predictions made from geophysical data.

Bayesian inversion uses *Bayes'* rule to combine the data *likelihood*, which contains information from the current data set, with *prior information* about expected spatial distributions of geological properties, both presented in the form of probability distributions (Mosegaard & Tarantola, 1995; Tarantola & Valette, 1982). The resulting distribution describes uncertainty in the estimated model given all of the available prior information and data and is called the *posterior distribution*: this is the complete solution of the inverse problem of inferring a geological model from the geophysical data.

The computation of joint posterior probability distributions over a large number of parameters using Bayesian inversion is computationally intractable. Probability distributions are therefore generally explored through stochastic sampling, usually using the *Markov chain Monte Carlo* (McMC) method, (e.g., Mosegaard & Sambridge, 2002; Mosegaard & Tarantola, 1995; Sambridge & Mosegaard, 2002). McMC generates chains of samples that are distributed according to the posterior distribution as the number of samples tends to infinity. However, in most applications successive samples are highly correlated, which severely reduces the information content of any finite sample set compared to a similarly sized set of independent samples. Hence, alternative methods have been developed, which avoid McMC-based sampling. For example, Walker and Curtis (2014a) developed a facies inversion method using *exact sampling* as an efficient alternative to McMC sampling. In that method every sample is an independent sample from the posterior probability distribution leading to far greater information content in any fixed set of samples. Using alternative strategies for such spatial models, Nawaz and Curtis (2017, 2018) developed inversion methods which avoid sampling entirely by computing the posterior distribution using numerical optimization. Our current method follows the latter philosophy and allows probabilistic inversion to be performed while avoiding McMC. The method uses supervised learning, so may require some type of Monte Carlo sampling to generate an example database for training purposes, but typically, this is a far lower dimensional and more computationally tractable sampling process than that required to use general McMC methods to solve the entire inference problem.

In order to appreciate the significance of our advance, it is first necessary to understand the set of assumptions that are commonly made in spatial statistical inference problems. Bayesian inversion is usually performed by defining a joint probability distribution over all of the observed as well as the unobserved (called *hidden*) variables. Modeling the joint distribution over hidden as well as observed variables is commonly referred to as *generative modeling*, since given the joint distribution over all of the variables we can use it to *generate* new synthetic data corresponding to known model parameter values. This is the standard method in most previous geophysical applications of probabilistic inverse theory.

To limit the analytical and computational complexity of modeling the joint distribution over all of the observed and hidden variables, previous research in geostatistical inversion assumed that the likelihoods are localized (or quasi-localized) and that the observed data are independent and identically distributed (e.g., Caers et al., 2006; Grana, 2018; Hoffman & Caers, 2007; Larsen et al., 2006; Nawaz & Curtis, 2017; Shahraeeni & Curtis, 2011; Shahraeeni et al., 2012; Ulvmoen & Omre, 2010; Walker & Curtis, 2014a). The *localized-likelihood* assumption models the probabilistic relationship between the data observed at a location and the model parameters at the same location independently from other locations. Another typical assumption in spatial inversion using soft conditioning data (such as seismic data) is that such data are spatially smooth, and therefore that smooth spatial patterns of geological parameters may be inferred directly by using such data without the need to perform spatial inference (Caers & Ma, 2002; Grana, 2018; Shahraeeni et al., 2012; Shahraeeni & Curtis, 2011). However, this approach is more susceptible to noise present in the data. Examples of previous research in which the localized likelihood assumption has been relaxed in 1-D Bayesian inversion methods are Lindberg and Omre (2014, 2015) and Grana et al. (2017). Nawaz and Curtis (2018) relaxed the localized likelihood assumption by introducing multidimensional quasi-localized likelihoods (QLLs), which relate observed data at a location to the model parameters in any finite neighborhood of that location.

Another commonly used assumption in inversion is the *conditional independence of data*, whereby data are assumed to be independent of each other given the model parameters. In other words, no correlations in data noise across space (or time on temporal grids) are accounted for: the data are assumed to be mutually uncorrelated, apart from their interdependence due to correlated geological parameters. For example, the McMC-based method of Larsen et al. (2006) performs a trace-by-trace inversion of recorded seismic waveforms

(traces) and assumes that each trace is conditionally independent of others, given the geology. Walker and Curtis (2014a) and Nawaz and Curtis (2017, 2018) perform a multidimensional inversion (with 2-D examples) but still assume conditional independence of data given the geology.

Although the assumptions of localized likelihoods and conditional independence of data allow simpler mathematical treatment of the inverse problem and more efficient computation of its solution, they come at the cost of introducing two major limitations in modeling. First, these assumptions underestimate long-range correlations present in the data because long-range correlations may only enter the solution explicitly through the geological prior distribution. Second, solutions ignore any correlated noise present in the data, which may therefore percolate erroneously into the inversion results; for example, we may have a residual data acquisition foot-print in images, improper focusing in the imaging process due to model errors, or residual multiples and surface wave noise in seismic images. Since such effects commonly impact almost all geophysical surveys, any acquired data may contain long-range correlations due both to the reflected signal from geological layers, and to noise resulting from inaccuracies in data processing, or the acquisition footprint (Chopra & Larsen, 2000). Accounting for long-range correlations is therefore vital for the realistic reconstruction of complex geological patterns and thus for reliable subsurface modeling.

In this paper we remove the assumptions of localized likelihoods and conditional independence of data that are used in most of the previous research on this topic. Using *nonlocalized likelihoods* in solving a spatial inverse problem requires coupling of the model and data spaces such that all of the model parameters may be conditioned (depend) on any of the data, irrespective of the locations of observations. Conversely, our current method also allows data from anywhere in the model to be related to the model parameters at any location if there exists such a logical or conceptual association.

Exact computation of fully nonlocalized likelihoods is intractable in most models of practical interest. To address this problem while also avoiding McMC, we propose a Bayesian inversion method that directly models the posterior distribution without requiring that the joint distribution over all of the variables is specified. This approach is called *discriminative modeling*. Although classification of data using a discriminative model is a common choice in the machine learning community, Bayesian inversion using a generative model is the standard approach in large-scale geophysical problems. In this paper, we therefore introduce a new approach to Bayesian inversion in geophysical problems based on a discriminative model in which prior and likelihoods are implicitly incorporated such that the posterior distribution is modeled without explicit mathematical modeling of the joint distribution over the observed and hidden variables (Figure 1). We refer to Bayesian inversion using a discriminative model as *discriminative Bayesian inversion* or simply *discriminative inversion*.

Bayesian inversion is reviewed in section 2, where the conventional approach to Bayesian inversion based on a generative model is first discussed in more detail in section 2.1 to explain why it is difficult to remove assumptions of localized likelihoods and conditional independence of data. The discriminative inversion approach is then introduced in section 2.2 as a tractable alternative to the generative approach in large and complex models. A model for the posterior distribution is proposed in section 2.3. Then a mathematical formulation of the VB method is introduced in section 3 to perform inference, and an approximate inference method is derived in section 3.1, and an associated method for parameter estimation is presented in section 3.2. The computational complexity of this method is discussed in section 4. After providing mathematical details of the method, a synthetic test example is provided in section 5 where this method is applied to invert multiple seismic attributes for geological facies (shale, brine-sand, and gas-sand) in the presence of strongly correlated noise. Finally, the implications of the method are discussed in section 6 and conclusions of this research are provided in section 7.

Before proceeding, we define the notational schema used in this paper. We use a linear index denoted by lower case letters such as i for locations (cells) in the geological model or for training examples used in supervised learning. Sets are represented with italic, regular (nonboldface) capital (English or Greek) letters, for example, $\mathscr{V}$ and $\mathscr{G}$. We use boldface font with lower case (English or Greek) letters for vectors, for example, $\boldsymbol{m}$ or $\boldsymbol{d}$, and upper case letters for matrices, for example, $\mathbf{H}$. The identity matrix is represented as $\mathbf{I}$. A superscript T stands for the transpose of a vector or matrix. Bracketed superscripts indicate an index over training examples used in supervised learning, for example, $\boldsymbol{m}^{(i)}$ represents the ith instance of a quantity $\boldsymbol{m}$. Other

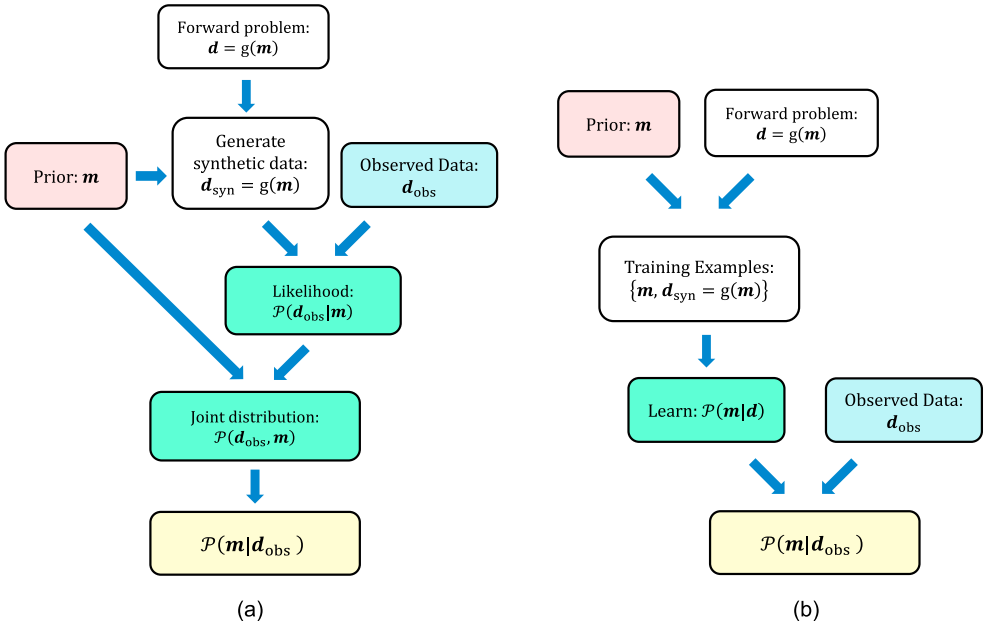

Figure 1. Flow chart comparison of (a) the conventional method of geophysical probabilistic inversion using a generative model and (b) the discriminative probabilistic inversion method introduced here. Colors match related steps between (a) and (b).

commonly used statistical and set theoretic notations include "~" for a random variable, which reads "is distributed as"; "⊂" reads "is a proper subset of"; "∈" reads "is a member of"; "|·|" is for cardinality (or number of elements) of a set; and "←" denotes the "assignment" (or the "update") operation, which means that "the value of the variable on the left is update from its old value using the expression on the right-hand-side of the left-arrow."

2. Bayesian Inversion

The probabilistic inverse problem that we solve is to infer the unknown geological model parameters $\boldsymbol{m}$ from the observed geophysical data or its attributes $\boldsymbol{d}$. We use Bayesian inversion, which is a probabilistic paradigm for solving an inverse problem. It combines uncertainty in modeling the observed data $\boldsymbol{d}$ from model $\boldsymbol{m}$ by solving a forward problem, and the uncertainty in $\boldsymbol{m}$ reflected by prior information about the true model that is independent of $\boldsymbol{d}$. The uncertainty in $\boldsymbol{d}$ for a given model $\boldsymbol{m}$ is encoded in the likelihood function, which is defined as the conditional probability $\mathscr{P}(\boldsymbol{d}|\boldsymbol{m})$ of observing $\boldsymbol{d}$ given that $\boldsymbol{m}$ is the true model, taking into account observational uncertainties in measuring the data. The uncertainty in $\boldsymbol{m}$ based on prior information alone is encoded in the form of a prior probability distribution $\mathscr{P}(\boldsymbol{m})$. The likelihood and the prior probabilities form the essence of Bayesian inversion as they may be combined to yield the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ using *Bayes' theorem*, expressed as

$$\mathscr{P}(\boldsymbol{m}|\boldsymbol{d}) = \frac{\mathscr{P}(\boldsymbol{d}, \boldsymbol{m})}{\mathscr{P}(\boldsymbol{d})} = \frac{\mathscr{P}(\boldsymbol{d}|\boldsymbol{m})\mathscr{P}(\boldsymbol{m})}{\mathscr{P}(\boldsymbol{d})} \tag{1}$$

The denominator $\mathscr{P}(\boldsymbol{d})$ represents the *marginal likelihood* of the observed data $\boldsymbol{d}$ (also called *model evidence* or simply the *evidence*). It acts as a normalization constant and is given by

$$\mathscr{P}(\boldsymbol{d}) = \int_{\boldsymbol{m}} \mathscr{P}(\boldsymbol{d}, \boldsymbol{m}) d\boldsymbol{m} = \int_{\boldsymbol{m}} \mathscr{P}(\boldsymbol{d}|\boldsymbol{m})\mathscr{P}(\boldsymbol{m}) d\boldsymbol{m} \tag{2}$$

2.1. Bayesian Inversion Using a Generative Model

We see from equations (1) and (2) that the prior $\mathscr{P}(\boldsymbol{m})$ and the likelihood $\mathscr{P}(\boldsymbol{d}|\boldsymbol{m})$ completely specify the posterior distribution through the joint distribution $\mathscr{P}(\boldsymbol{d},\boldsymbol{m})$. A model that describes the probability of $\boldsymbol{m}$ given $\boldsymbol{d}$ in terms of their joint distribution is commonly referred to as a *generative model*. Thus, a generative model explicitly expresses the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ in terms of the data likelihood $\mathscr{P}(\boldsymbol{d}|\boldsymbol{m})$ and the prior model distribution $\mathscr{P}(\boldsymbol{m})$ using equation (1).

Explicit specification of priors and likelihoods in Bayes' theorem has an intuitive meaning: the data $\boldsymbol{d}$ are assumed to have been generated by the unknown model $\boldsymbol{m}$ according to a prespecified probability distribution $\mathscr{P}(\boldsymbol{d}|\boldsymbol{m})$, while the probability $\mathscr{P}(\boldsymbol{m})$ of $\boldsymbol{m}$ is known a priori. It is for this reason that explicit modeling of the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ in terms of the joint distribution $\mathscr{P}(\boldsymbol{d},\boldsymbol{m}) = \mathscr{P}(\boldsymbol{d}|\boldsymbol{m})\mathscr{P}(\boldsymbol{m})$ is known as *generative modeling*, since given the joint distribution over all of the hidden as well as observed variables we can artificially *generate* more data from it.

Estimation of a joint distribution over all of the variables offers a full description of a probabilistic system; it allows marginalization and conditioning over any subset of variables in order to perform inference, sampling, and prediction. For this reason, generative modeling seems to be an attractive approach. However, the joint distribution over observed and hidden variables is generally too complex to be modeled accurately. In addition, since the generative approach requires modeling a joint distribution over all of the variables that comprise a system, it may turn out to be an inefficient approach in situations where our objective is to solve a specific problem rather than to characterize the entire system. For instance, in geophysics our objective is usually only to compute the conditional distribution of $\boldsymbol{m}$ given $\boldsymbol{d}$; we can achieve this by manipulating the probabilistic relationships among various dependent variables mathematically, without modeling the full joint distribution over both $\boldsymbol{m}$ and $\boldsymbol{d}$. In a dense system where every variable depends on a large number of other variables, this task is practically as daunting as estimating the full joint distribution over all of the variables. However, many problems of practical interest regarding spatial phenomena involve sparse systems (Besag, 1974). Examples include cases where parameter dependencies can be modeled as a *Markov random field* (MRF; see Nawaz & Curtis, 2018) in which marginalization can be performed efficiently using dynamic programming (Denardo, 2003) or some approximate methods that do not require estimation of the full joint distribution (Koller & Friedman, 2009). In such a case, estimating the joint distribution over all of the variables may be regarded as a cumbersome and unnecessary intermediate step that requires immense modeling efforts and intense computational power.

2.2. Bayesian Inversion Using a Discriminative Model

The above concerns regarding generative modeling leads us to explore the alternative *discriminative modeling* approach. This directly estimates the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ as a nonlinear mathematical function, without modeling the joint distribution $\mathscr{P}(\boldsymbol{m},\boldsymbol{d})$ over $\boldsymbol{d}$ and $\boldsymbol{m}$ as an intermediate step. In this manner, discriminative modeling alleviates some of the effort required to model any complex dependencies among variables through their full joint distribution and proves to be parsimonious in the use of computational resources.

Since discriminative modeling does not require estimation of the joint distribution over hidden as well as the observed variables, we can deploy the modeling effort and computational resources to incorporate additional sophistication in the model without tremendously increasing the computational cost of the overall method. Based on this notion, we propose a *discriminative Bayesian inversion* method that uses nonlocalized likelihoods and accounts for correlations observed in the data, without making any conditional independence assumptions about the observed variables. Inversion methods that are based on a generative model are computationally too demanding to allow for such sophistication in the model. Below we present a discriminative analogue of a MRF, which we use as a model for the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ that implicitly incorporates spatial priors and nonlocalized likelihoods.

2.3. Posterior Model

A commonly used probabilistic model for expressing prior information about the spatial distributions of geological properties is a MRF (Arpat & Caers, 2007; Nawaz & Curtis, 2018; Rimstad & Omre, 2010; Ulvmoen & Omre, 2010). A MRF is a model of the joint distribution of hidden variables $\mathscr{P}(\boldsymbol{m})$ in a model that decomposes the entire set of hidden variables $\boldsymbol{m}$ into subsets called *cliques*, denoted by $\mathscr{C}$, each of which

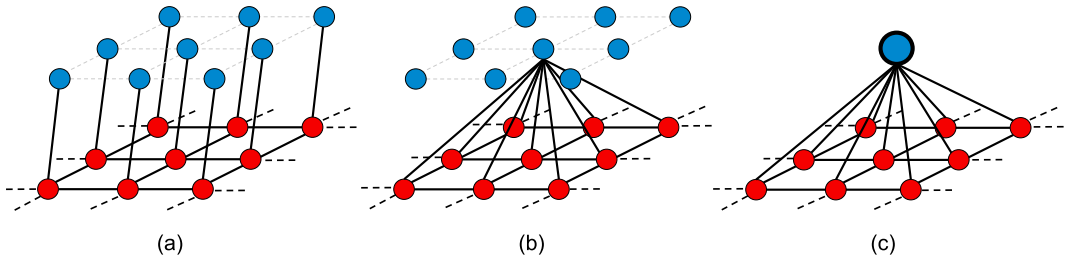

Figure 2. Schematic graphical representation of (a and b) hidden Markov random fields (HMRF), and (c) a conditional random field. A red circle represents hidden variables $\boldsymbol{m}_i$ at each location i in the model, blue circles represent observed variables $\boldsymbol{d}_i$, larger blue circle with a thick border in (c) represents all of the observed data $\boldsymbol{d}$, and solid black lines connecting circles represent direct probabilistic dependence between the connected variables. The dotted lines only represent the location grid and not the probabilistic dependence. The HMRF in (a) assumes localized likelihoods, while the HMRF in (b) assumes quasi-localized likelihoods (Nawaz & Curtis, 2018). Both HMRFs assume conditional independence of data $\boldsymbol{d}$ given model parameters $\boldsymbol{m}$. The conditional random field in (c) makes no such assumptions. This figure shows only pairwise cliques represented by pairs of connected hidden variables (red circles). In general, cliques may represent higher-order dependence among more than two variables and hence may extend beyond the 3×3 grid of pairwise connected variables shown here in red.

contains only mutually dependent variables. A clique is called a maximal *maximal clique*, denoted by $\hat{c}$, if it is defined in terms of the maximum number of mutually dependent variables such that it would cease to be a clique if any other variable is added to it. The set of all maximal cliques in a model is represented as $\hat{\mathscr{C}}$. According to the *Hammersley-Clifford theorem* (Besag, 1974), a MRF can express the joint distribution $\mathscr{P}(\boldsymbol{m})$ of hidden variables $\boldsymbol{m}$ as a *Gibbs distribution* in terms of a product of positive valued functions $\psi(\boldsymbol{m}_{\hat{c}})$ (also called *potential functions*) over the maximal cliques $\hat{c}$ in a model, and this can be written exactly as

$$\mathscr{P}(\boldsymbol{m}) = \frac{1}{\mathscr{Z}} \prod_{\hat{c} \in \hat{\mathscr{C}}} \psi(\boldsymbol{m}_{\hat{c}}) \tag{3}$$

where $\mathscr{Z}$ is a normalizing constant given by

$$\mathscr{Z} = \int_{\boldsymbol{m}} \prod_{\hat{c} \in \hat{\mathscr{C}}} \psi(\boldsymbol{m}_{\hat{c}}) d\boldsymbol{m} \tag{4}$$

A *hidden Markov random field* (HMRF) is a variant of a MRF that models the joint distribution of some unobserved (hidden) variables $\boldsymbol{m}$, each of which, say, $\boldsymbol{m}_i$ at a location i, is conditioned on some observed data $\boldsymbol{d}_i$ (Figure 2a). A *conditional random field* (CRF; Lafferty et al., 2001) is essentially a HMRF defined over hidden variables $\boldsymbol{m}$, each of which may be conditioned to all of the observed data $\boldsymbol{d}$. A schematic comparison of a HMRF and a CRF is shown in Figure 2 in the form of a graphical model where variables are represented as nodes (circles) and probabilistic dependence among them is represented as edges (links between nodes).

A CRF models the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d})$ in terms of some strictly positive functions $\psi(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$, also called potential functions, defined over the domain of model parameters $\boldsymbol{m}_{\hat{c}}$ within a maximal clique and data $\boldsymbol{d}$. The logarithm of potential functions are typically expressed as a linear combination of some, generally nonlinear, prespecified vector of feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$ of $\boldsymbol{m}_{\hat{c}}$ and $\boldsymbol{d}$ with relative weights $\boldsymbol{w}$ such that the posterior distribution $\mathscr{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ of $\boldsymbol{m}$ given $\boldsymbol{d}$ parameterized by $\boldsymbol{w}$ may be written as

$$\mathscr{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w}) = \frac{1}{\mathscr{Z}(\boldsymbol{d};\boldsymbol{w})} \prod_{\hat{c} \in \hat{\mathscr{C}}} \psi(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}) = \frac{1}{\mathscr{Z}(\boldsymbol{d};\boldsymbol{w})} \exp\left(\sum_{\hat{c} \in \hat{\mathscr{C}}} \boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})\right) \tag{5}$$

where the denominator $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$ is the evidence that acts as a normalization constant and is given by

$$\mathcal{Z}(\boldsymbol{d};\boldsymbol{w}) = \int_{\boldsymbol{m}} \exp\left(\sum_{\hat{c}\in\hat{\mathcal{C}}} \boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})\right) d\boldsymbol{m} \tag{6}$$

which is a function of the observed data $\boldsymbol{d}$ and is parametrized by $\boldsymbol{w}$.

The feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d})$ are assumed to encode sufficient statistics of the desired distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$, and their eloquent specification is therefore crucial for accurate modeling of the true posterior distribution. This requires that the feature functions are chosen such that they capture the data-model relationship adequately. For example, feature functions may be defined as a measure of how likely are some features in the data given the spatial distributions of geological properties within a maximal clique. The feature functions thus implicitly model the spatial priors over $\boldsymbol{m}_{\hat{c}}$ and the nonlocalized likelihoods that define the probabilistic relationship between $\boldsymbol{m}_{\hat{c}}$ and $\boldsymbol{d}$. Since conditioning can be performed over the entire set of observed variables $\boldsymbol{d}$, no localization of likelihoods is required in this model. In a discriminative framework, this is what allows the direct modeling of the posterior distribution, which is otherwise intractable if no conditional independence is assumed over the observed variables (e.g., in a HMRF).

Once the feature functions have been defined, the next step is to devise efficient methods to estimate parameters $\boldsymbol{w}$ in equations (5) and (6), and for spatial inference. Spatial inference involves estimating the normalization constant $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$, the marginal posterior distributions over cliques and individual variables in the model, and any posterior statistics of interest such as the most likely overall model $\boldsymbol{m}^*$ of $\boldsymbol{m}$ given $\boldsymbol{d}$, such that $\boldsymbol{m}^* = \text{argmax}_{\boldsymbol{m}}\{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})\}$. Parameter estimation can be performed in a supervised manner by using training examples of model $\boldsymbol{m}$ and the corresponding data $\boldsymbol{d}$ in order to obtain an estimate of the parameters $\boldsymbol{w}$ that best describe the distribution of $\boldsymbol{m}$ given $\boldsymbol{d}$ under the posterior model in equation (5). We discuss inference and parameter estimation methods in a CRF model in sections 3.1 and 3.2, respectively.

3. Variational Bayesian Inference

In the light of computational challenges in McMC-based inference, we use the VB method, which allows computationally tractable approximate Bayesian inference in large-scale models (Koller & Friedman, 2009), and is an efficient and prominent alternative to McMC in decomposable models such as the CRF used in this paper. It uses numerical optimization in order to estimate the probabilities and statistics of interest, by transforming probabilistic inference and parameter estimation problems into a variational optimization framework.

For given data $\boldsymbol{d}$ we want to maximize $\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$ as a function of $\boldsymbol{w}$, which is intractable. VB defines a lower bound on the log-evidence $\mathcal{L}(\boldsymbol{w};\boldsymbol{d}) \equiv \log\mathcal{Z}(\boldsymbol{d};\boldsymbol{w})$, which is maximized with respect to $\boldsymbol{w}$ as a surrogate for maximization of the generally intractable log-evidence. In effect, VB approximates a generally intractable joint posterior distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d},\boldsymbol{w})$ with an auxiliary distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$, where $\mathbb{Q}$ is a family of tractable distributions and $\boldsymbol{w}$ is the set of model parameters. The variational distribution $\mathcal{Q}$ is chosen from $\mathbb{Q}$ such that it minimizes some distance measure between distributions $\mathcal{Q}$ and the desired $\mathcal{P}$: the measure is commonly chosen to be the *KL-divergence* (also called *relative-entropy*) $KL(\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})\|\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w}))$, or simply $KL(\mathcal{Q}\|\mathcal{P})$, given by

$$KL(\mathcal{Q}\|\mathcal{P}) = \mathbb{E}_{\mathcal{Q}}\left[\log\frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})}\right] = \int_{\boldsymbol{m}} \mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \log\frac{\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})}{\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})} d\boldsymbol{m} \geq 0 \tag{7}$$

where $\mathbb{E}_{\mathcal{Q}}$ represents the expectation with respect to the distribution $\mathcal{Q}$. Equality to zero holds in equation (7) when $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) = \mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$.

In order to estimate $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as an approximation to the desired $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$, we express the log evidence $\mathcal{L}(\boldsymbol{w};\boldsymbol{d})$ in terms of $KL(\mathcal{Q}\|\mathcal{P})$ (see the supporting information for a mathematical derivation, or, e.g., Nawaz & Curtis, 2018) as

$$\mathscr{L}(\boldsymbol{w};\boldsymbol{d}) = \mathscr{F}(\mathbb{Q},\boldsymbol{w}) + KL(\mathbb{Q}\|\mathscr{P}) \tag{8}$$

where $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$ is known as the *variational free energy*, or simply *free energy*, and is given by

$$\mathscr{F}(\mathbb{Q},\boldsymbol{w}) = \mathbb{E}_{\mathbb{Q}}\left(\sum_{\hat{c}\in\hat{\mathscr{C}}} \log \psi(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})\right) + \mathscr{S}(\mathbb{Q}) \tag{9}$$

where $\mathscr{S}(\mathbb{Q}) = -\int_{\boldsymbol{m}} \mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) \log \mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) d\boldsymbol{m}$ is the entropy of the variational distribution $\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d})$ as a function of data $\boldsymbol{d}$. The nonnegativity of the relative-entropy $\mathrm{KL}(\mathbb{Q}\|\mathscr{P})$ ensures that

$$\mathscr{L}(\boldsymbol{w};\boldsymbol{d}) \geq \mathscr{F}(\mathbb{Q},\boldsymbol{w}) \tag{10}$$

from equation (8), where the inequality follows from the fact that $KL(\mathbb{Q}\|\mathscr{P}) \geq 0$ in equation (7). The free energy $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$ given by equation (9) is a concave function of $\mathbb{Q}$ (Cover & Thomas, 1991); thus, for any $\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) \neq \mathscr{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ the variational approximation is guaranteed to be a lower bound on the desired log-evidence, $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$. In other words, if we manage to find an auxiliary distribution $\mathbb{Q}$, which equals $\mathscr{P}$, then we also ensure that free energy $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$ equals the log evidence $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$.

The VB method therefore casts inference into an optimization framework by defining the objective function to be the variational free energy $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$, which needs to be maximized. Since $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$ is constant for a given model, maximizing the free energy $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$ for fixed $\boldsymbol{w}$ is equivalent to minimizing $\mathrm{KL}(\mathbb{Q}\|\mathscr{P})$ by equation (8). Thus, by maximizing the free energy $\mathscr{F}(\mathbb{Q},\boldsymbol{w})$ as a function of both $\mathbb{Q}$ and $\boldsymbol{w}$, we can estimate the desired quantity $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$. As an intrinsic outcome of this optimization, we obtain the approximate posterior distribution $\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) \approx \mathscr{P}(\boldsymbol{m}|\boldsymbol{d},\boldsymbol{w})$. Although $\mathscr{Z}(\boldsymbol{d};\boldsymbol{w})$ remains intractable, its evaluation is not required in the variational optimization as it is independent of $\mathbb{Q}$.

The expectation term of the free energy in equation (9) involves expectations over individual cliques with respect to $\mathbb{Q}$ and is therefore easy to compute for cliques of reasonable size, provided that the choice of the family of possible $\mathbb{Q}$s allows efficient inference. The entropy term, on the other hand, involves expectations over all possible realizations of $\boldsymbol{m}$ and does not necessarily factorize. Thus, the computational complexity of the entropy term depends on the properties of $\mathbb{Q}$. This entails some approximation to overcome the computational complexity of $\mathscr{S}(\mathbb{Q})$.

3.1. Mean Field Approximation

Within the VB framework, various approximate inference methods have been proposed to address the intractability in large-scale probabilistic graphical models or models involving variables with dense dependencies. Nawaz and Curtis (2018) used Bethe's approximation (Bethe, 1935; Yedidia et al. 2001a, 2001b) in the *Loopy-Belief Propagation* method for a pairwise graphical model to estimate the marginal posterior distribution of model parameters under the QLL assumption. Here we use the *mean-field* (MF) approximation (Koller & Friedman, 2009; Opper & Saad, 2001) as discussed below.

In models with no cyclic dependencies among variables, dynamic programming can be used to perform exact inference by exploiting the conditional independence between most variables (Denardo, 2003). In graphs with cycles (or loops), the MF method makes variational inference viable. The MF approximation is based on numerical optimization and assumes some type of independence over the hidden variables $\boldsymbol{m}$. In the context of a CRF, this independence is assumed to be conditioned to the observed variables.

A naive MF approximation (Jaakkola, 1997; Koller & Friedman, 2009) assumes that all of the hidden variables $\boldsymbol{m}_i, i \in \mathscr{V}$ are independent of each other, that is

$$\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) \cong \prod_{i\in\mathscr{V}} \mathbb{Q}_i(\boldsymbol{m}_i|\boldsymbol{d}) \tag{11}$$

Such a fully factorized distribution may not capture the information in a general multivariate distribution $\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d})$. We obtain a MF approximation by taking $\mathbb{Q}$ to be a family of factorizable distributions such that the auxiliary distribution $\mathbb{Q}(\boldsymbol{m}|\boldsymbol{d}) \in \mathbb{Q}$ factorizes into marginal distributions $\mathbb{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ over some proper subcliques c of the maximal cliques $\hat{c}$ in the model, with some prespecified order $|c| = q$, such that

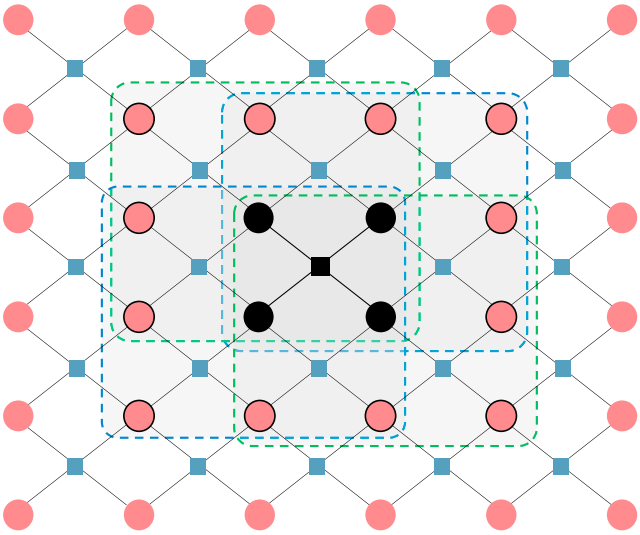

Figure 3. Graphical illustration of the mean field updates. Circles represent vertices in the graph (or the hidden variables $\boldsymbol{m}$); squares which connect vertices through edges (lines) represent clique configurations $\boldsymbol{m}_c$ (also called factors) over approximating cliques c with size 2×2 vertices. Consider an approximating clique c with 2×2 vertices in the center (shown in black color) for the mean-field update of the approximate marginal distribution $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$. Assume that the maximal cliques $\hat{c}$ in the graph have a size of 3×3 vertices. Four of the (3×3) maximal cliques, which share the approximating clique c, are shown as rounded rectangles with dashed boundaries. For the model parameters $\boldsymbol{m}_c$ in c, the summation in equation (13) runs over the set of maximal cliques that share c to compute the conditional expectation over the factors $\boldsymbol{m}_{\hat{c}}$ given $\boldsymbol{m}_c$.

$$\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d}) \cong \prod_{c \subset \hat{c} \in \hat{\mathcal{C}}} \mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \tag{12}$$

We refer to this approximation as the *higher-order MF approximation*. Note that the above equation degenerates to the naive MF approximation given by equation (11) for $|c| = 1$. The approximate marginal posterior distributions $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ over subcliques c may be obtained by maximizing $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ as a function of $\mathcal{Q}$ for a given set of parameters $\boldsymbol{w}$ (see the supporting information for a mathematical derivation), which gives

$$\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d}) \leftarrow \frac{1}{\mathcal{Z}_c(\boldsymbol{d})} \exp\Big\{ \sum_{\hat{c} \in \hat{\mathcal{C}} : c \subset \hat{c}} \mathbb{E}_{\mathcal{Q}_{\setminus c}} \left[\boldsymbol{w}^T \boldsymbol{f}(\boldsymbol{m}_{\hat{c}}, \boldsymbol{d}) | \boldsymbol{m}_c \right] \Big\} \tag{13}$$

where $\mathcal{Q}_{\setminus c}$ represents the per-clique marginals of $\mathcal{Q}$ except for the clique c. Thus, marginal distribution $\mathcal{Q}_c$ over each approximating clique c is updated by using the expression on the right-hand side of the left-arrow. The subscript $\hat{c} \in \hat{\mathcal{C}} : c \subset \hat{c}$ of summation in the above expression reads "for all $\hat{c}$ in $\hat{\mathcal{C}}$ such that c is a proper subset of $\hat{c}$." In simple words, the summation in this expression runs over each maximal clique $\hat{c}$ in the model that contains the approximating clique c that is being updated.

The system of $|\mathcal{C}|$ nonlinear update equations (13) collectively represents the *higher-order MF equations*, which may be solved successively in an iterative manner. Since each update has a closed form solution, the free energy $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ increases monotonically in each iteration; convergence is therefore guaranteed. (However, there are some caveats about convergence that are discussed in section 6, which must not be ignored). The factorized joint distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ can therefore be evaluated by summation of terms, which are defined over a relatively small number of variables (small compared to the exponential number of terms over all of the variables for an unfactorizable distribution). As a consequence, the computational cost depends mainly on the size of the factors (approximating cliques) and not on the structure of the spatial dependencies. This allows tractable approximate inference in graphs with complex structures where exact inference would require exponential time.

Although the form of updates is different, the MF update algorithm resembles message passing over a cluster graph, for example, cluster belief propagation (Koller & Friedman, 2009), where clusters refer to higher-order cliques and messages represent approximate marginal distributions over cliques. Figure 3 shows a graphical illustration of the MF update of $\mathcal{Q}_c(\boldsymbol{m}_c|\boldsymbol{d})$ with an example where the approximating clique c has a size of 2×2 vertices, while the maximal cliques $\hat{c}$ in the graph have a size of 3×3 vertices. Unlike Bethe's approximation, the MF approximation does not approximate the objective (the energy functional); it only approximates the restricted optimization space $\mathbb{Q}$ of distributions. In general, any solution that results from the mean field updates is a locally optimal solution but is nonunique because $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ is nonconcave in the approximate marginal distributions $\mathcal{Q}_c$. The quality of the higher-order MF approximation depends on the difference in the order of maximal cliques $\hat{c}$ and the approximating cliques c: the smaller the difference $|\hat{c}| - |c|$, the better the approximation.

3.2. Parameter Estimation

The CRF parameters $\boldsymbol{w}$ in equation (5) can be estimated by using the *regularized maximum conditional-likelihood* method that searches for the parameters that maximize the conditional log likelihood of the model for a given training data set (Sutton & McCallum, 2012). In other words, in parameter estimation we aim to find a set of parameters $\boldsymbol{w}$ that makes the approximate posterior distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as close to the true distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d}; \boldsymbol{w})$ as possible. This method requires computation of the gradient of the log likelihood $\mathcal{L}(\boldsymbol{w}; \boldsymbol{d})$, which is intractable and cannot be computed exactly. For this reason, we also use the MF approximation to estimate the log likelihood.

The true joint distribution $\mathscr{P}(\boldsymbol{m},\boldsymbol{d})$ over the entire model is unknown, but we assume that we have a training data set that consists of *independent and identically distributed* samples from the true distribution over maximal cliques—prespecified subsets of the model. We assume that the training data $D=\{\boldsymbol{m}^{(i)},\boldsymbol{d}^{(i)}:i=1,\ldots,N\}$ contain N pairs of local configurations of the hidden variables $\boldsymbol{m}^{(i)}=\left\{\boldsymbol{m}_{\hat{c}}{}^{(i)}:\forall\hat{c}\in\mathscr{C}\right\}$ over maximal cliques $\hat{c}$ and the corresponding input data $\boldsymbol{d}^{(i)}$, where the bracketed superscript (i) indicates an index over the training instance. The input data $\boldsymbol{d}^{(i)}$ are not required to have the same topology as that of a clique template; however, there should exist some conceptual or logical association between $\boldsymbol{d}^{(i)}$ and $\boldsymbol{m}^{(i)}$. For example, the training data could be prepared from some real data that are manually interpreted and classified by experts, or built from stochastic simulation of geological properties and corresponding data using a variety of Earth models (or training images) of expected geology.

The conditional log likelihood $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})=\log\mathscr{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$ is then given by

$$\mathscr{L}(\boldsymbol{w};\boldsymbol{d})=\sum_{i=1}^{N}\left[\log\mathscr{P}\left(\boldsymbol{m}^{(i)}|\boldsymbol{d}^{(i)},\boldsymbol{w}\right)\right]-\lambda|\boldsymbol{w}|^{2} \tag{14}$$

$$=\sum_{i=1}^{N}\left[\sum_{\hat{c}\in\mathscr{C}}\boldsymbol{w}^{T}\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}}{}^{(i)},\boldsymbol{d}^{(i)}\right)-\log\mathscr{Z}\left(\boldsymbol{d}^{(i)},\boldsymbol{w}\right)\right]-\lambda|\boldsymbol{w}_{p}|^{2} \tag{15}$$

where we used equation (5) in the second equality, and $\lambda>0$ is a regularization parameter, which controls the strength of regularization. The conditional log likelihood $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$ in the above equation cannot be maximized analytically. We therefore use gradient-based nonlinear numerical optimization. The gradient of the conditional log likelihood in equation (15) is given by

$$\nabla_{\boldsymbol{w}}\mathscr{L}(\boldsymbol{w};\boldsymbol{d})=\sum_{i=1}^{N}\left[\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}}{}^{(i)},\boldsymbol{d}^{(i)}\right)-\mathbb{E}_{\boldsymbol{m}\sim\mathscr{P}\left(\boldsymbol{m}|\boldsymbol{d}^{(i)},\boldsymbol{w}\right)}\left[\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}},\boldsymbol{d}^{(i)}\right)\right]\right]-2\lambda\boldsymbol{w} \tag{16}$$

The zero-gradient conditions thus require that the feature functions $\boldsymbol{f}(\boldsymbol{m}_{\hat{c}},\boldsymbol{d})$ have the same expectations under the model (the CRF) and the empirical (training) distributions. We therefore approximate the expected features using the MF inference method as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{m}_{\hat{c}}\sim\mathscr{P}\left(\boldsymbol{m}_{\hat{c}}|\boldsymbol{d}^{(i)},\mathbf{w}\right)}\left[\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}},\boldsymbol{d}^{(i)}\right)\right]&\cong\mathbb{E}_{\boldsymbol{m}_{c}\sim\mathscr{Q}_{\boldsymbol{w}}\left(\boldsymbol{m}_{c}|\boldsymbol{d}^{(i)}\right)}\left[\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}},\boldsymbol{d}^{(i)}\right)\right]\\&=\sum_{\boldsymbol{m}_{c}}\mathscr{Q}_{\boldsymbol{w}}\left(\boldsymbol{m}_{c}|\boldsymbol{d}^{(i)}\right)\boldsymbol{f}\left(\boldsymbol{m}_{\hat{c}},\boldsymbol{d}^{(i)}\right)\end{aligned} \tag{17}$$

where $\mathscr{Q}_{\boldsymbol{w}}\left(\boldsymbol{m}_{c}|\boldsymbol{d}^{(i)}\right)$ refers to the marginals of $\mathscr{Q}_{\boldsymbol{w}}\left(\boldsymbol{m}|\boldsymbol{d}^{(i)}\right)$ under the approximation $\mathscr{Q}_{\boldsymbol{w}}\left(\boldsymbol{m}|\boldsymbol{d}^{(i)}\right)\cong\mathscr{P}\left(\boldsymbol{m}|\boldsymbol{d}^{(i)},\boldsymbol{w}\right)$, that is, for a given set of parameters $\boldsymbol{w}$. Since we assume that all of the variables ($\boldsymbol{m}^{(i)}$ and $\boldsymbol{d}^{(i)}$) are observed in the training data, the log likelihood $\mathscr{L}(\boldsymbol{w};\boldsymbol{d})$ is a concave function. Therefore, any local maximum is indeed a global maximum. The log likelihood can therefore be maximized using gradient ascent optimization as long as we can compute the gradient exactly; however, this is known to be too slow to converge (Yuan, 2010). Newton or quasi-Newton type methods such as L-BFGS (Dennis & Schnabel, 1996) use local curvature of the objective function to achieve faster convergence; however, these methods require computation and inversion of the Hessian matrix $\mathbf{H}$ given by

$$\mathbf{H}(\mathscr{L}(\boldsymbol{w};\boldsymbol{d}))=-\sum_{i=1}^{N}\left(\mathrm{Cov}_{\boldsymbol{m}\sim\mathscr{P}\left(\boldsymbol{m}|\boldsymbol{d}^{(i)},\boldsymbol{w}\right)}\left[\boldsymbol{f}\left(\boldsymbol{m},\boldsymbol{d}^{(i)}\right)\right]\right)-2\lambda\mathbf{I} \tag{18}$$

A key challenge in performing probabilistic inversion with nonlocalized likelihoods is that the inverse problem is highly nonlinear because inference for the posterior distribution requires some estimates of the model parameters (CRF weights $\boldsymbol{w}$), whereas estimation of the model parameters requires some estimates of the posterior distribution. This paradox may be solved by first performing inference with randomly initialized parameters $\boldsymbol{w}$ to approximate the marginal posterior distributions, and then updating the parameters by using these approximate posterior distributions. Then inference and parameter estimation are carried out in an iterative fashion until both the model parameters and estimated marginal posterior distributions converge to within a predefined tolerance.

4. Computational Cost

The computational cost of this method can be divided into the three main components of the algorithm: learning the feature functions, MF inference, and CRF parameter estimation. Feature functions are a rather general concept, and their learning cost depends on the complexity of the task and on the exact method used for learning. For example, the computational complexity of learning feature functions using a multilayer perceptron neural network is at most quadratic in the total number of neurons in the network.

The overall cost C_{MF} of the MF algorithm, expressed in terms of the maximum number of floating point operations required, is given by

$$C_{MF} \leq |\mathscr{C}| \times \max |\boldsymbol{m}_c| \times \max |\mathscr{N}_c| \times L_{\mathrm{MF}} \tag{19}$$

where c is the clique that defines the mean field approximation in equation (13), $|\mathscr{C}|$ is the total number of cliques c in the model, $\max |\boldsymbol{m}_c|$ is the maximum dimensionality of model parameters in a clique c, $\max |\mathscr{N}_c|$ is the maximum number of maximal cliques $\hat{c}$ that contain c as a subset in the model, which is also referred to as the *neighborhood cardinality* of the model, and L_{MF} is the total number of MF iterations.

Similarly, the cost C_{PE} of parameter estimation for the CRF model with the L-BFGS method is given by

$$C_{\mathrm{PE}} \leq (|\mathscr{C}| \times \max |\boldsymbol{m}_c| \times \max |\mathscr{N}_c|) \times N \times n_{\boldsymbol{w}}^2 \times L_{\mathrm{PE}} \tag{20}$$

where N is the number of training examples, $n_{\boldsymbol{w}}$ is the number of weights in the CRF model, and L_{PE} is the total number iterations required for the L-BFGS algorithm to converge.

Equations (19) and (20) show that the factors which control the cost of this method are the number $n_{\boldsymbol{w}}$ of CRF parameters, the dimensionality of model parameters in a clique $|\boldsymbol{m}_c|$, and the size of neighbourhood cardinality $|\mathscr{N}_c|$. The latter two factors themselves depend on the clique size $|c|$ of the approximating distribution $\mathscr{Q}(\boldsymbol{m}|\boldsymbol{d})$ and the maximal clique size $|\hat{c}|$ in the graph. If the clique size is too small, it may not be able to capture the expected complexity in the model parameters, and subsequent inference may not be able to model the true spatial distribution of model parameters. If the clique size is too large it may induce unnecessary model complexity and require more computational power for learning than is actually needed for a given problem. A trade-off is thus required between geological complexity that is to be modeled and the required computational resources. Nevertheless, the above cost is expected to be far lower than would be required to solve the same problem using Monte Carlo methods for the class of problems, which involve non-localized likelihoods and which make no conditional independence assumptions on the observed data in high dimensions.

5. Synthetic Test

Removing the assumption of localized likelihoods and conditional independence of data means that our method should be able to account for any correlations present in the data due to spatial blurring of data or due to correlated noise, as long as we can model some salient characteristics (or features) such as the spatial correlation of noise. In order to test this, and to benchmark the current method against previous research, we used the same test Earth model as used in the previous research of Walker and Curtis (2014a) and Nawaz and Curtis (2017, 2018). Here for the first time, we demonstrate that the new method is capable of inverting seismic attributes for facies with reasonable accuracy even in the presence of strongly correlated noise.

The synthetic example is based on two independent vertical cross sections extracted from a 3-D geological process model that contains channel-filled and overbank sand deposits within a background of shale (Nawaz & Curtis, 2018; Walker & Curtis, 2014a). 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 litho-facies considered for discrimination are therefore given by the sample space

$$\mathscr{G} = \{ \text{shale}, \text{brine-sand}, \text{gas-sand} \} \tag{21}$$

We used one of the vertical cross sections with dimensions of 200×200 cells as a training image (Figure 4a), and another with dimensions of 100×100 cells as the target geological model (Figure 4b), which we refer to

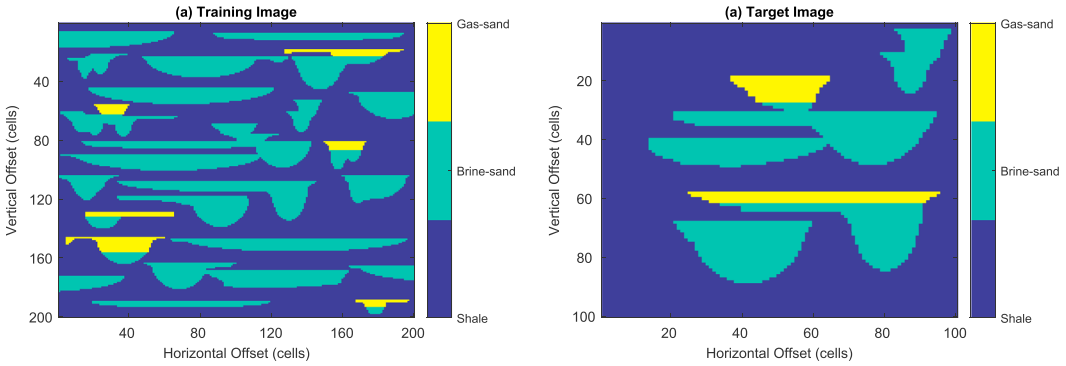

Figure 4. (a) The training image and (b) the target image. Both were extracted as independent 2-D cross sections from a 3-D geological process model containing channels with filled and overbank sand deposits within a background of shale. The sand is filled with brine or gas, which obey gravitational ordering of the two fluids. Note the change in spatial scale between the two images (distance units are arbitrary).

as the "true geology." 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. 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 maximal cliques in the graphical model.

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 $\boldsymbol{d}'_i$ 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). Further details and parameters of the rock physics model used to generate synthetic data may be found in Nawaz and Curtis (2018). Correlated random noise was then introduced in the simulated seismic attributes in the form of NW-SE oriented random streaks of amplitudes by convolving the simulated attribute sections with a NW-SE oriented filter, in order to generate collocated synthetic seismic attributes (*P* and *S* wave impedances) as the noisy input $\boldsymbol{d}$ for our method (Figures 5a and 5b). The aim is to train our algorithm to disregard the correlated noise and reproduce the true distribution of facies. We refer to the resulting synthetic attributes as the "true data" as these were then inverted with our method with the aim to reproduce the "target geology" (Figure 4b).

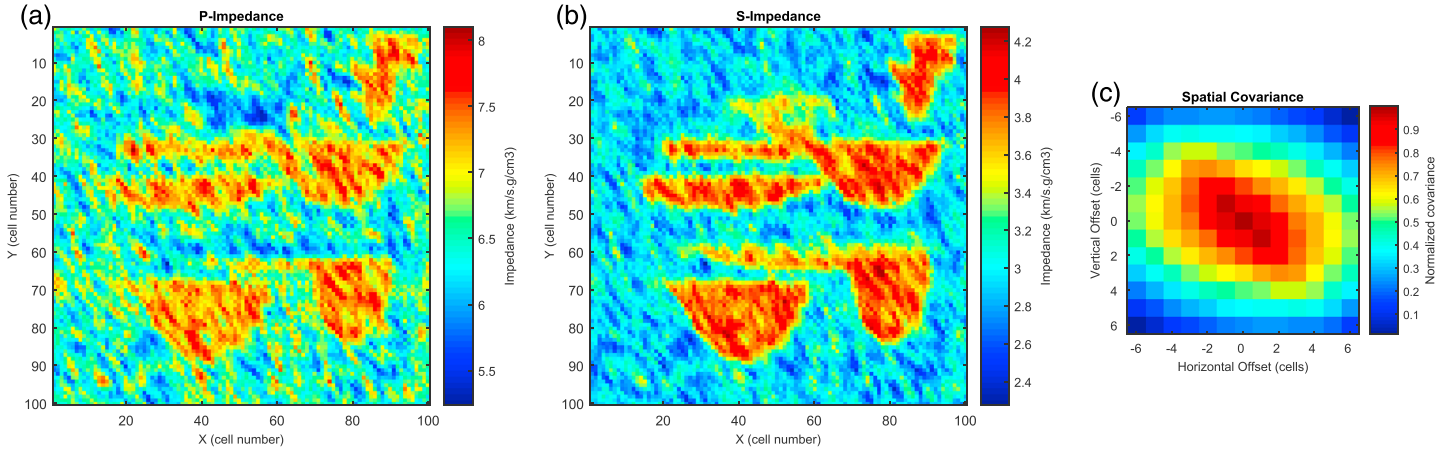

Figure 5. Synthetic (a) *P* wave and (b) *S* wave impedance attributes used as input for the synthetic test. (c) Spatial covariance matrix computed from the synthetic attributes (*P* and *S* wave impedances) cross sections in panels a and b, for a maximum vertical and lateral offset of 6 × 6 cells.

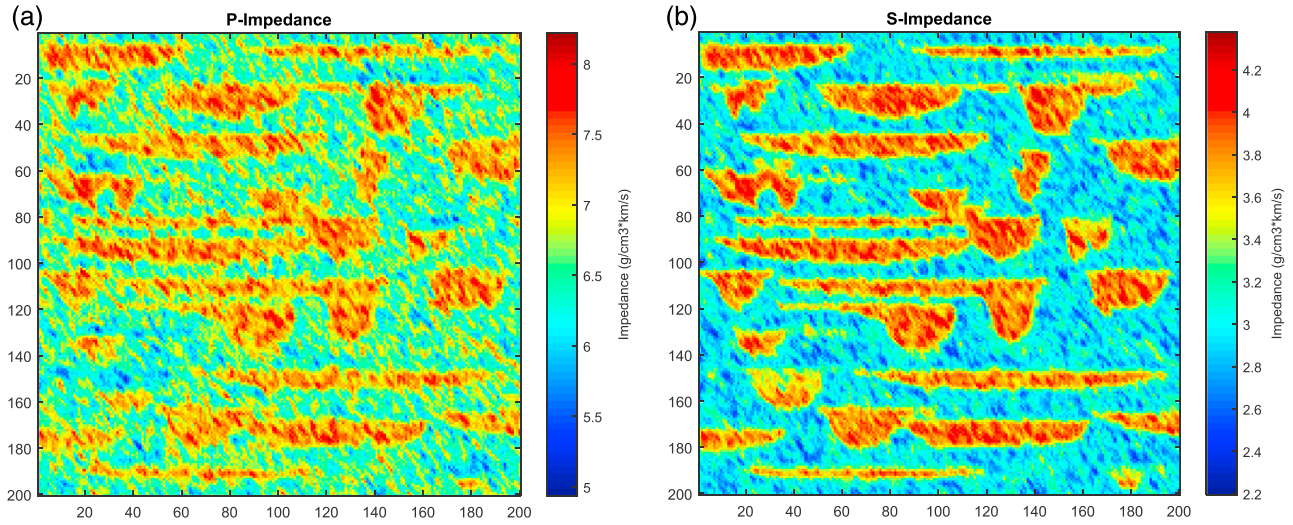

Figure 6. Simulated (a) *P* and (b) *S* wave impedance sections generated by convolving stochastically simulated attributes from the training image (Figure 4a) with the spatial correlation matrix in Figure 5c in order to mimic the correlated noise observed in the input attribute sections (Figures 5a and 5b). These simulated sections are used to generate stochastic examples for training the neural network in order to learn feature functions.

The spatial covariance matrix was computed from these synthetic attributes (the input data), which provides an estimate of the spatial variability of impedances in the presence of strongly correlated noise. The computed covariance matrix was then tapered to retain the maximum amplitudes along the main diagonal, while the off-diagonal correlations were suppressed to yield a filter that can introduce similar correlated noise in the simulated examples that we used later for supervised learning. The normalized spatial covariance matrix is shown in Figure 5c, which shows strong correlations in the NW-SE direction similar to the orientation of noise streaks in the data. Such an approach where noise is estimated from the observations under the assumption of stationarity is commonly referred to as the *empirical Bayes* method.

Prior information was extracted from the training image in terms of prior probabilities $\mathcal{P}(\boldsymbol{m}_{\hat{c}})$ constructed from histograms of various facies configurations that occur in the image. We chose two clique templates each with a size of 9×9 model cells to relate facies patterns in a clique with the corresponding *P* wave and *S* wave impedances, respectively. The size of the clique template was chosen based on the size and shape of features observed to be present in the attributes.

Next we prepared examples of seismic attributes and the desired facies patterns. Since the attributes that are used as the data to test our method are synthetically generated, we use the term "simulated" (rather than "synthetic") for the attribute sections used to build stochastic examples for training a neural network to learn feature functions. Simulated attributes were generated using the rock physics model described above from facies patterns present in the training image (Figure 4a). In order to introduce correlated noise in the simulated attributes, these were cross correlated with the tapered form of the spatial covariance matrix estimated from the 'true data' shown in Figure 5c. The resulting noisy sections of *P* and *S* wave impedances simulated from the facies present in the training image are shown in Figure 6.

An example database was then prepared for supervised learning in the form of two sets of facies patterns extracted from the training image (Figure 4a) within the prespecified clique templates, and the corresponding cells in the simulated attributes sections (Figure 6). In the context of supervised learning, we refer to the facies patterns in the example database as the "target facies," and the corresponding simulated attributes as "simulated features." The simulated features were extracted from each of the simulated attribute sections (Figure 6) using windows of the same size as the clique templates (9×9 cells). In this example, the size of training features was chosen to be the same as that of the clique templates (9×9 cells), which adequately captured the salient characteristics of data and correlated noise with respect to the corresponding facies configurations. A total of 5,000 examples were stochastically "chosen from" facies configurations within the predefined clique templates and the corresponding features (simulated *P* and *S* impedances) in the example

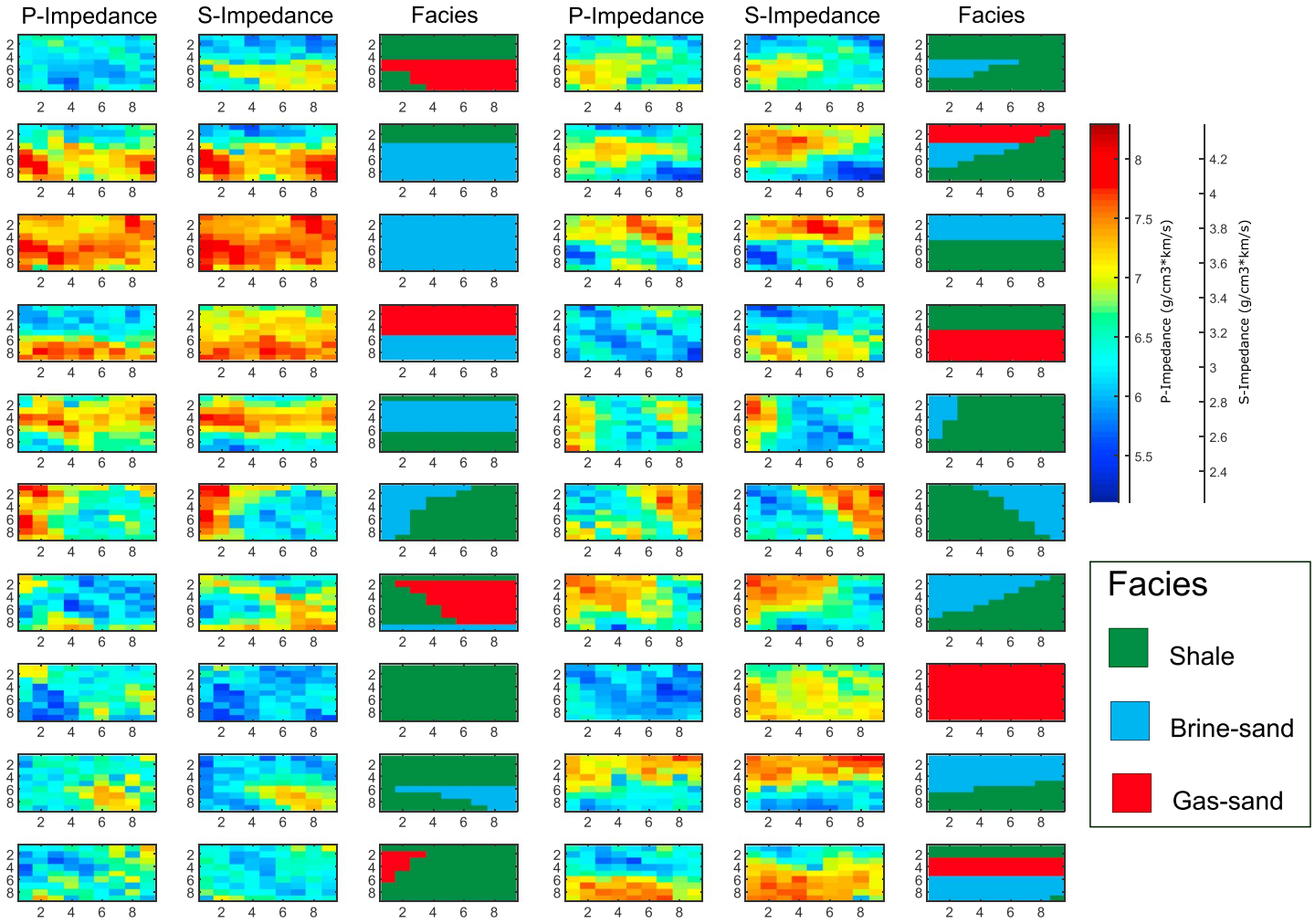

Figure 7. Examples of simulated *P* and *S* wave impedances and corresponding facies patterns in a window of size 9×9 model cells. These examples were used to train a neural network in order to learn feature functions.

database. Figure 7 shows a few such examples with training features from each of the clique templates and the corresponding target facies.

Feature functions were then defined for each of the clique templates as a vector of indicator variables corresponding to the facies in each cell of the clique template. Each of the indicator variables is set to 1 for the facies present in the target pattern and 0 for all other facies patterns. Separate neural networks were then trained with the training features (e.g., the *P* impedance and *S* impedance columns in Figure 7) as input and the corresponding feature functions (e.g., the indicator representation of the facies columns in Figure 7) as the desired output for each of the clique templates. In this manner the outputs of a trained neural network may be interpreted as a measure of how likely is a facies configuration for a given input feature. After training the neural networks on "stochastically selected examples", features were extracted from the "true data" corresponding to each of the clique templates, and the associated feature functions were computed using the trained neural network.

After computing the feature functions, we initialized the CRF weights $\boldsymbol{w}$ randomly and approximate inference was performed using the MF update equations (13) to obtain expected feature functions under the auxiliary distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ as an approximation to the expected feature functions under the model distribution $\mathcal{P}(\boldsymbol{m}|\boldsymbol{d};\boldsymbol{w})$. These posterior distributions were then used to update the CRF weights using the quasi-Newton optimization method L-BFGS. Since estimation of the posterior distributions requires the CRF weights to be known, and estimation of the CRF weights requires the posterior distributions to be

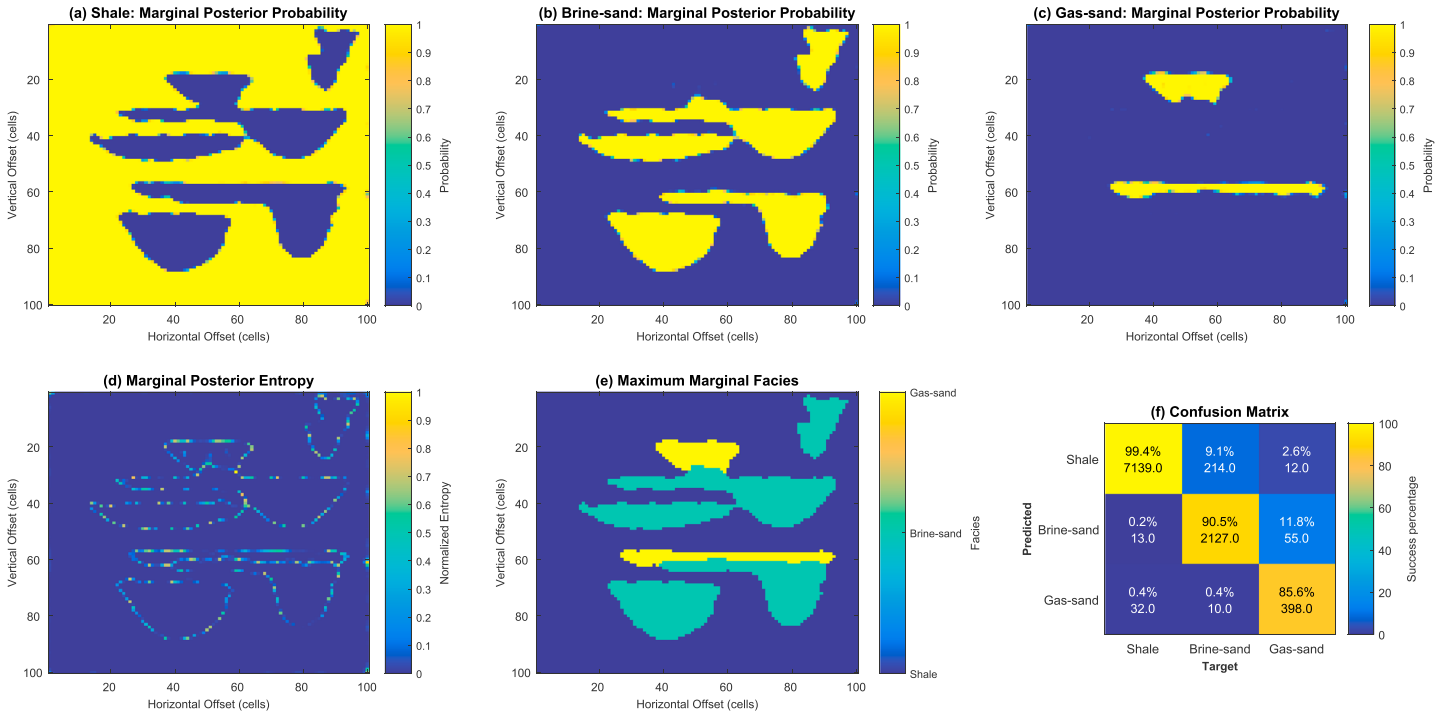

Figure 8. (a–c) Approximate marginal posterior distributions for the three facies (shale, brine-sand, and gas-sand) obtained after mean-field approximation with optimized conditional random field parameters. (d) Entropy of the approximate marginal posterior distributions. (e) Facies with maximum marginal distribution. Note that this is not a Maximum-a-Posteriori estimate (i.e., it is not a realization). (f) Confusion matrix showing the success rate of predictions versus targets for the three facies.

known, each of these were alternately updated in an iterative fashion until both converged within a prespecified tolerance. The final estimates of the marginal posterior distributions in each cell are shown in Figures 8a–8c, and the entropy (a measure of uncertainty) is shown in Figure 8d. The map of the facies that has the maximum of the marginals in each cell, shown in Figure 8e, shows reasonable reconstruction of the target geology (Figure 4b) given that the input attributes contain strongly correlated noise. The quality of prediction is quantified in terms of success rate computed as a percentage of cells with predicted facies for each of the three facies in the model. This is shown by the *confusion matrix* in Figure 8f.

The quality of prediction is very good as the overall accuracy rate is 97%. The major errors lie in false prediction of shale when the true facies was brine-sand, and false prediction of brine-sand when the true facies was gas-sand. Errors are mostly found at the transitions between different facies where entropy is at its highest (see Figure 8d). The high accuracy of prediction resulted from the fact that the noise follows a linear (NW-SE) trend (Figure 5c) that is different from the trend of geological correlations and that the prior information extracted from the training image is a good representation of the "true" geology. Either of these may not be guaranteed in real-data problems. Therefore, the accuracy rate may not be as good in practical situations and it depends on the quality of geological prior information and our ability to discriminate noise correlations from expected geological correlations. Nevertheless, high prediction accuracy in this synthetic example does show that the method is reliable provided the required inputs are available with reasonable accuracy.

5.1. Summary of the Method as Applied Above

The following is a stepwise summary of the overall method used in this synthetic example:

1. Identify features of the data and collect data correlation statistics.
2. Perform forward simulation of data that corresponds to the training image incorporating the correlation statistics.
3. Define clique templates, and feature functions that relate data features in a clique template to facies patterns in a maximal clique.

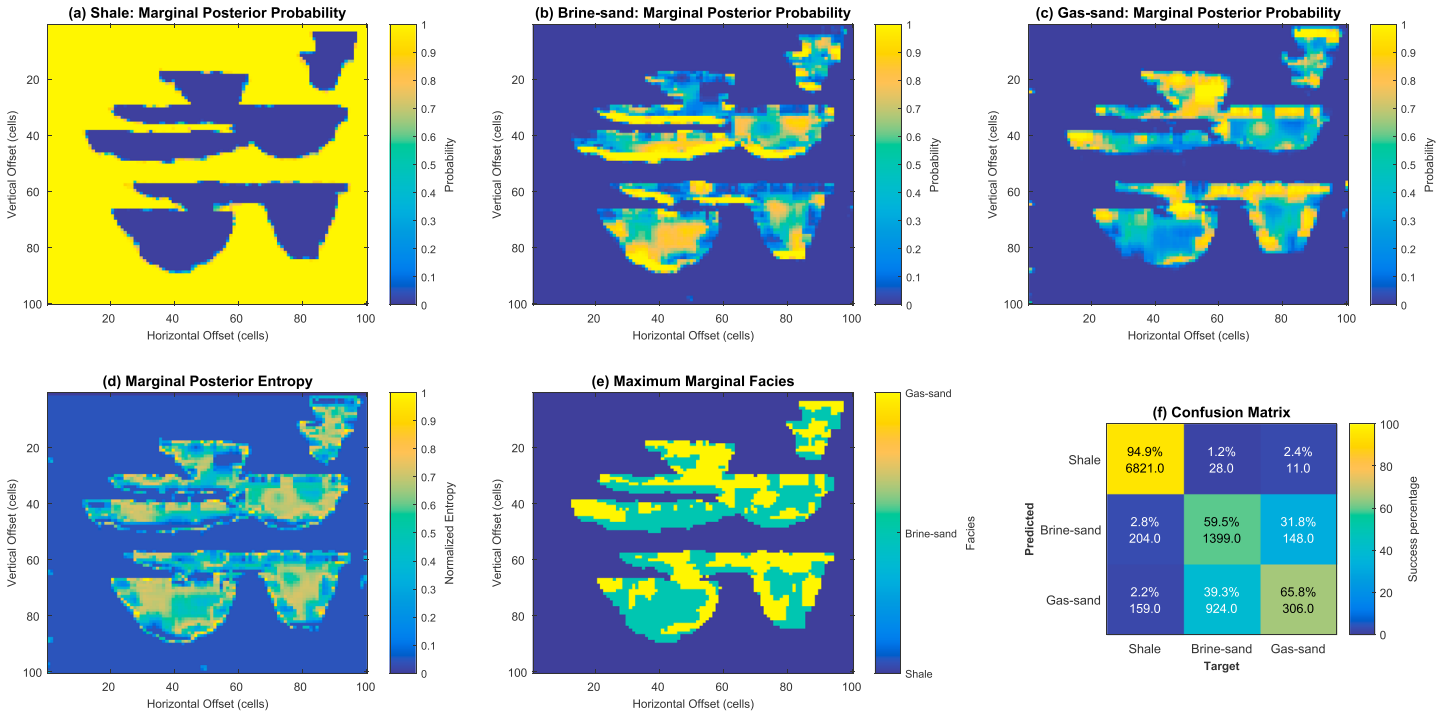

Figure 9. (a–c) Approximate posterior marginal distributions for the three facies (shale, brine-sand, and gas-sand) obtained using the quasi-localized likelihoods based facies inversion method of Nawaz & Curtis, 2018. (d) Entropy of the approximate marginal posterior distributions. (e) Facies with maximum marginal distribution. Note that this is not a Maximum-a-Posteriori estimate (i.e., it is not a realization). (f) Confusion matrix showing the success rate of predictions versus targets for the three facies.

4. Train a machine learning model (e.g., a neural network) on training examples extracted from the training image and its associated simulated data, to learn feature functions from the data.
5. Define a CRF model using equation (5) with feature functions as the basis functions and initialize CRF weights $\boldsymbol{w}$ randomly.
6. Perform MF inference using equations (13) to estimate approximate posterior distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ from the current estimate of CRF weights $\boldsymbol{w}$.
7. Update CRF weights $\boldsymbol{w}$ using a nonlinear optimization method (e.g., L-BFGS) with the gradient of the conditional log likelihood in equation (16) computed from the current estimate of approximate posterior distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$.
8. Repeat steps 6 and 7 until the approximate posterior distribution $\mathcal{Q}(\boldsymbol{m}|\boldsymbol{d})$ and the CRF weights $\boldsymbol{w}$ converge to within a predefined tolerance.

5.2. Comparison With Previous Research

For a comparison we applied our previous method of facies inversion using "Quasi-Localized Likelihoods (QLLs)" (Nawaz & Curtis, 2018) to the data with strongly correlated noise as shown in Figure 5. Nawaz and Curtis (2018) already showed that the QLL-based method performs significantly better than localized methods in this problem. In order to make a fair comparison between the two methods, we modified the algorithm of Nawaz and Curtis (2018) to use higher-order cliques of size 9×9 instead of just pairwise cliques. The results from our previous method are shown in Figure 9: these exhibit good discrimination between shale and sand (Figure 9a), while the discrimination between brine-sand and gas-sand is poor (Figures 9b and 9c). The entropy (or uncertainty) of prediction in the QLL-based method (Figure 9d) is significantly higher than the entropy estimated in our current method (Figure 8d). This is also evident from the map of most probable facies in each cell, which shows significant errors in discrimination between brine-sand and gas-sand compared to the 'true' facies map shown in Figure 4b. A quantitative comparison between the two methods is shown in Figures 8f and 9f. This shows that shale is correctly predicted in 94.9% of the cells, which is slightly lower than 99.4% in our current method), whereas brine-sand and gas-sand are correctly predicted in only 59.5% and 65.8% of the cells compared to 90.5% and 85.6% in our current method,

respectively. The latter occurs because although spatial inference is performed in the QLL-based method in order to reproduce geologically plausible patterns of facies (as depicted in the training image in Figure 4a), it could not handle correlated noise in the data. Thus, our current method shows substantial improvement over the approach of Nawaz and Curtis (2018) in the presence of strongly correlated noise. Here we recall that most previously existing methods of facies inversion assume that any correlations present in the data are a direct consequence of correlations in the geology—the so called *conditional independence* assumption (on data). Our current method, on the other hand, provides a new mathematical framework for probabilistic inference that incorporates complex features in the data that should be acknowledged or accounted for during the inversion process and is capable of providing reliable results (Figure 8) even in the presence of strongly correlated noise.

6. Discussion

Both generative and discriminative modeling require reasonable knowledge of the underlying relationship between model parameters and the data. This relationship is often presented in the form of mathematical or computational functions in generative modeling and is presented as (often simulated) training examples from which mathematical functions, here referred to as feature functions, may be derived in the discriminative approach. The advantage of the discriminative approach is that it learns the inverse of the underlying forward model, and the inverse may be arbitrarily complex and nonlinear, may represent nonuniqueness in that inverse relation, and may represent the true model-data relationship (given suitable training examples from the real relationship) rather than a synthetic approximation to that relationship. Consequently, discriminative modeling may learn more complexity in a problem with less effort than is required to produce an accurate generative model for the same problem.

As an example, we showed with a synthetic example in section 5 that we only needed to model and learn some statistical characteristics of correlated noise present in the data in order to properly account for it during inversion of the noisy data. Note here that the noise was introduced by convolving a filter with the noise-free attributes; that is, noise is correlated with the signal and not additive. Noise in acquired data due to acquisition footprint, nonuniform source directivity, or multiple scattering of energy in the subsurface are examples of such a case where noise is convolved with the signal. Applying the generative modeling approach to such an example requires reliable prediction of the correlated noise. Formulating the joint distribution over noisy data and the desired model parameters in a generative approach can be hard as it would require reliable prediction of the noise along with the signal for any given set of model parameters. The discriminative approach simplifies this by not attempting to model the noise; only statistical characteristics of noise are needed in order to discriminate between signal and noise.

Generating and learning from training examples may be a tedious task; however, the effort spent preparing training examples and learning the inverse mapping (from data to model parameters) often depends mostly on the complexity of the problem, and not so much on the size of the problem in cases where the problem can be decomposed (factorized) into smaller subcomponents. This means that the same training examples that are prepared for inversion of a small seismic section may be used to invert a large 3-D seismic volume provided that the assumption of stationarity (that the same training examples are equally appropriate everywhere in the volume) is valid. In other words, the expensive part of our method (the learning stage) operates at a scale that is greatly reduced compared to the full problem, allowing the method to scale to far larger problems.

The feature functions must be defined such that they effectively capture complex relationships between the geological model and the data. If feature functions are not designed to capture the desired features of the data and model parameters adequately, it may introduce significant inaccuracies in the solution. Various machine learning methods have been proposed to achieve this task, for example, random forests (Ho, 1995), support vector machines (SVM, Cortes & Vapnik, 1995), and deep neural networks (DNN, e.g., Hinton et al., 2006) such as convolutional neural networks (Fukushima, 1980). The definition of feature functions and the decision about which method is used to learn these depends mainly on the type and complexity of the features that are to be modeled and requires an interpretative approach. The general approach presented here allows any feature learning method or a combination thereof to be employed under the assumption that the training examples represent the data-model relationship reasonably well and that the accuracy of feature functions learnt from the training examples is acceptable.

Training examples can be created in at least two ways: feature vectors could be extracted from real data and manually classified by experts to provide the corresponding geological parameters or pairs of feature vectors and their classes could be created by stochastically generating synthetic data for a variety of Earth models (or training images) of expected geological features. The former approach is a type of expert elicitation in which statistical information is elicited from experts based on their subjective opinion about the extracted data features (Macrae et al., 2016; Polson & Curtis, 2010, 2015; Walker & Curtis, 2014b). The latter approach uses a generative framework where data are modeled from the spatial distribution of geological properties obtained, for example, by using geological process modeling. Task-specific features in the data must be captured in the training examples to define feature functions. Although the overall inversion still uses a discriminative framework for learning the posterior distribution of geological properties across the entire model given all of the data, it may thus be decomposed into smaller generative models, each of which only models the distribution of geological properties within a maximal clique (or a clique template) and a specific associated data feature.

Feature functions do not require the data to be defined in the same domain as the geological model, so geological properties in each clique may potentially be related to features in all of the data. For example, the geology may be spatial and the seismic data may be in space-time domain. If the desired data features are prohibitively large to be stored in computer memory and subsequently analyzed, their size may be reduced by using dimensionality reduction techniques such as principal component analysis (Pearson, 1901). Since data features of arbitrary shape and size may be used to define feature functions, the likelihoods encoded by the feature functions are fully nonlocalized. Additionally, defining feature functions requires no assumption about conditional independence of data and hence can model any arbitrary correlations in the data. Complex feature functions may be learnt, and the definition of the posterior distribution in equation (5) shows that any number of feature functions can be included in the design. Also, feature functions may be defined to suit the problem at hand. For example, a separate feature function may be defined for features extracted from each of the seismic attributes, and additional feature functions may be defined to model any correlated noise present in the data in order to account for it in the inversion process. Thus, our method is reasonably general and may be applied to a variety of problems and many types of data.

A principal motivation of the current research was to remove two commonly used assumptions in probabilistic inversion: the localized likelihood assumptions and the conditional independence assumption on data. This is achieved in the posterior probability model since the feature functions implicitly encode the prior distribution and the nonlocalized likelihoods. Extending any Bayesian inversion method that uses localized or QLLs to fully nonlocalized tomography or full waveform inversion problems is not straightforward. Our method does not require the data to be defined on a spatial grid that is the same as the geological model. Therefore, we may hope to extend this method to seismic tomography and full waveform inversion type problems in future.

The proposed inversion method combines the machine learning-based discriminative classification with spatial inference to solve the probabilistic inverse problem of determining the spatial distribution of geological properties from geophysical data. Spatial inference corrects inaccuracies and reduces uncertainties in feature functions by constraining the spatial distribution of geological properties at neighboring locations to be consistent with both the spatial priors and the nonlocalized likelihoods. The dimensionality of geological properties in a large clique template may be too high. This is addressed by the MF approximation. Also, as discussed earlier, exact inference is not possible for large cyclic graphical models, so MF inference was used as a tractable approximation. The naive MF method is quite limited as it assumes independence of individual vertices; the quality of such an approximation is governed by the density (as opposed to sparsity), scale, and strength of neglected interactions among various variables of interest. The higher-order mean field approximation defined in this paper attempts to ameliorate the loss due to neglecting significant interactions among variables as it assumes independence of nonmaximal cliques in the graph: if the size of such cliques is sufficiently large to capture the expected spatial distributions of geological properties, MF inference proves to be an efficient and reliable approximation in models where posterior distribution is factorizable (e.g., in a MRF).

Any solution of the MF equations is a stationary point and is not guaranteed to be an optimum. However, in practice, a MF solution is empirically known to converge to local optima in most scenarios because it is

highly unlikely for a solution to get stuck at an unstable stationary point (e.g., a saddle point; Koller & Friedman, 2009). Also, it is important to note that the locally optimal solution obtained from the MF updates is not guaranteed to be the same as the globally best factorized approximation $\mathcal{Q}$. This is because the solution depends on the initial CRF weights $\boldsymbol{w}$ and on the ordering of MF updates, both of which should usually be chosen randomly. In our experience, as long as the approximating cliques are large enough to capture the expected spatial patterns of geological properties, the MF algorithm converges to a consistent solution. In principle, the MF equations may be solved within a global optimization framework such as simulated annealing for global optimization of the free energy functional $\mathcal{F}(\mathcal{Q}, \boldsymbol{w})$ in equation (9), although we found that there was no need to do so in examples that we have tested.

In the light of above discussion, the quality of solutions from our method is determined by the choice of feature functions and their accuracy, that is, how well they relate the data and corresponding model parameters, the amount of prior information injected (defined by the maximal clique size), and how close the size of approximating cliques is to the maximal clique size. The latter factor mainly governs the computational cost of the method and essentially defines a trade-off between accuracy and computational efficiency. The MF inference that we deploy offers a more computationally efficient method compared to McMC; however, it is worthwhile to note that McMC is a general method that is in principle applicable to any inverse problem, while mean field inference offers a reasonable approximation only in models where the true posterior distribution is factorizable (e.g., a MRF). A MRF model is used in this paper because it is the most widely used model in spatial statistics (in particular geostatistics), even in most McMC-based geostatistical inversion methods (e.g., Luo & Tjelmeland, 2017; Rimstad & Omre, 2010; Ulvmoen & Omre, 2010). A fair comparison of accuracy and computational cost of McMC versus mean field inference requires such comparison to be made with respect to a given problem, that is, under the same set of assumptions. We leave such a comparison as a topic for future research.

7. Conclusions

We introduced a discriminative approach to Bayesian inversion of geophysical data for geological model parameters (discrete facies or continuous rock properties). This method models the posterior distribution of model parameters given the observed data directly using a CRF, as opposed to the commonly used generative-modeling-based Bayesian inversion that models the posterior distribution through the joint distribution of model parameters and observed data. For problems that are decomposable into interlinked subproblems as described herein, the presented discriminative approach thus circumvents the prohibitive amount of computational time and digital storage commonly required by the joint distribution and allows tractable inversion in complex problems for which the conventional generative approach becomes intractable. This allowed us to add more sophistication to our model and to remove the commonly used assumptions of localized likelihoods and conditional independence of data, without incurring significant computational limitations. Our proposed method incorporates spatial prior information and nonlocalized likelihoods and is therefore capable of modeling complex correlations in both data and geology.

Exact spatial inference is intractable in high-dimensional models. For this reason, approximate inference is generally performed using random sampling, for example, using McMC method. However, McMC can be computationally expensive and its convergence is neither guaranteed nor detectable in high-dimensional problems. We avoided the use of stochastic sampling and introduced a higher-order MF method for approximate inference within the variational Bayesian framework. Convergence to a local (and potentially global) optimum is guaranteed in this method. The MF inference may be performed within a global optimization framework such as simulated annealing or genetic algorithms to encourage global convergence. In a synthetic example, we demonstrated that this method is capable of inverting seismic attributes for facies with reasonable accuracy even in the presence of strongly correlated noise.

Acknowledgments
We thank TOTAL UK for their sponsorship of this research. We would also like to express 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 Klaus Mosegaard, an anonymous reviewer, and the Editors for their comments and constructive criticism on an earlier version of this manuscript. The authors report no conflicts of interest. This paper presents scientific advancement that is supported only by synthetic data. No real data were used in this research.

References

Arpat, G. B., & Caers, J. (2007). Conditional simulation with patterns. *Mathematical Geology*, *39*(2), 177–203. https://doi.org/10.1007/s11004-006-9075-3

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B: Methodological*, *36*(2), 192–225. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Bethe, H. (1935). Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, *150*(871), 552–575.
Caers, J., Hoffman, T., Strebelle, S., & Wen, X. H. (2006). Probabilistic integration of geologic scenarios, seismic, and production data—A West Africa turbidite reservoir case study. *The Leading Edge*, *25*(3), 240–244. https://doi.org/10.1190/1.2184087
Caers, J., & Ma, X. (2002). Modeling conditional distributions of facies from seismic using neural nets. *Mathematical Geology*, *34*(2), 143–167. https://doi.org/10.1023/A:1014460101588
Chopra, S., & Larsen, G. (2000). Acquisition footprint—Its detection and removal. *Canadian Society of Exploration Geophysicists Recorder*, *25*(8), 16–20.
Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
Cover, T., & Thomas, J. (1991). *Elements of Information Theory*. New York: John Wiley. https://doi.org/10.1002/0471200611
Denardo, E. V. (2003). *Dynamic Programming: Models and Applications*. Mineola, NY: Dover Publications.
Dennis, J. E., & Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, *SIAM Classics in Applied Mathematics*, (Vol. 16). Philadelphia: SIAM.
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 93–202.
Grana, D. (2018). Joint facies and reservoir properties inversion. *Geophysics*, *83*(3), M15–M24. https://doi.org/10.1190/geo2017-0670.1
Grana, D., Fjeldstad, T., & Omre, H. (2017). Bayesian Gaussian mixture linear inversion for geophysical inverse problems. *Mathematical Geoscience*, *49*(4), 493–515. https://doi.org/10.1007/s11004-016-9671-9
Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.
Ho, T. K., (1995). Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14-16 August 1995. pp. 278-282.
Hoffman, B. T., & Caers, J. (2007). History matching by jointly perturbing local facies proportions and their spatial distribution: Application to a North Sea Reservoir. *Journal of Petroleum Science and Technology*, *57*(3-4), 257–272.
Jaakkola, T. S. (1997). Variational methods for inference and learning in graphical models. PhD thesis, Massachusetts Institute of Technology (MIT).
Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press.
Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th ICML 2001, 282-289.
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. https://doi.org/10.1190/1.2245469
Lindberg, D., & Omre, H. (2014). Blind categorical deconvolution in two level hidden Markov models. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(11), 7435–7447. https://doi.org/10.1109/TGRS.2014.2312484
Lindberg, D., & Omre, H. (2015). Inference of the transition matrix in convolved hidden Markov models and the generalized Baum-Welch algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(12), 6443–6456. https://doi.org/10.1109/TGRS.2015.2440415
Luo, X., & Tjelmeland, H. (2017). Prior specification for binary Markov mesh models. *Statistics and Computing*, *29*(2), 367–389. https://doi.org/10.1007/s11222-018-9813-7
Macrae, E. J., Bond, C. E., Shipton, Z. K., & Lunn, R. J. (2016). Increasing the quality of seismic interpretation. *Interpretation*, *4*(3), T395–T402. https://doi.org/10.1190/INT-2015-0218.1
Marion, D. P. (1990). Acoustical, mechanical, and transport properties of sediments and granular materials, (PhD thesis). Stanford University, Department of Geophysics.
Mosegaard, K., & Sambridge, M. (2002). Monte Carlo analysis of inverse problems. *Inverse Problems*, *18*(3), R29–R54. https://doi.org/10.1088/0266-5611/18/3/201
Mosegaard, K., & Tarantola, A. (1995). Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research*, *100*(B7), 12,431–12,447. https://doi.org/10.1029/94JB03097
Nawaz, M. A., & Curtis, A. (2017). Bayesian inversion of seismic attributes for geological facies using a hidden Markov model. *Geophysical Journal International*, *208*(2), 1184–1200. https://doi.org/10.1093/gji/ggw411
Nawaz, M. A., & Curtis, A. (2018). Variational Bayesian inversion of seismically derived non-localized rock properties for the spatial distribution of geological facies. *Geophysical Journal International*, *214*(2), 845–875. https://doi.org/10.1093/gji/ggy163
Opper, M., & Saad, D. (Eds) (2001). *Advanced Mean Field Methods: Theory and Practice*, *Neural Information Processing Series*, (p. 273). Cambridge, Massachusetts, London, England: The MIT Press.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space, Philosophical Magazine, Series 6, vol. 2, no. 11, pp. 559-572.
Polson, D., & Curtis, A. (2010). Dynamics of uncertainty in geological interpretation. *Journal of the Geological Society*, *167*(1), 5–10. https://doi.org/10.1144/0016-76492009-055
Polson, D., & Curtis, A. (2015). Assessing individual influence on group decisions in geological carbon capture and storage problems. In P. Diviacco, P. Fox, C. Pshenichny, & A. Leadbetter (Eds.), *Collaborative Knowledge in Scientific Research Network* (Chap. 4, pp. 55–75). IGI Books. https://doi.org/10.4018/978-1-4666-6567-5.ch004
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. https://doi.org/10.1190/1.3463475
Sambridge, M., & Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, *40*(3), 1009. https://doi.org/10.1029/2000RG000089
Shahraeeni, M. S., & Curtis, A. (2011). Fast probabilistic nonlinear petrophysical inversion. *Geophysics*, *76*(2), E45–E58. https://doi.org/10.1190/1.3540628
Shahraeeni, M. S., Curtis, A., & Chao, G. (2012). Fast probabilistic petrophysical mapping of reservoirs from 3D seismic data. *Geophysics*, *77*(3), O1–O19. https://doi.org/10.1190/geo2011-0340.1
Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, *4*(4), 267–373. https://doi.org/10.1561/2200000013
Tarantola, A., & Valette, B. (1982). Inverse problems = quest for information. *Journal of Geophysics*, *50*(3), 150–170.
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. https://doi.org/10.1190/1.3294570
Walker, M., & Curtis, A. (2014a). Spatial Bayesian inversion with localized likelihoods: an exact sampling alternative to McMC. *Journal of Geophysical Research: Solid Earth*, *119*, 5741–5761. https://doi.org/10.1002/2014JB011010

Walker, M., & Curtis, A. (2014b). Expert elicitation of geological spatial statistics using genetic algorithms. *Geophysical Journal International*, *198*, 342–356. https://doi.org/10.1093/gji/ggu132

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. *International Journal of Rock Mechanics and Mining Science and Geomechanics Abstracts*, *30*(7), 805–808. https://doi.org/10.1016/0148-9062(93)90026-A

Yuan, Y. (2010). Gradient methods for large scale convex quadratic functions. In Y. Wang, A. G. Yagola, & C. Yang (Eds.), *Optimization and Regularization for Computational Inverse Problems and Applications* (Chap. 7, pp. 141–155). Beijing: Higher Education Press.