Math Geosci (2013) 45:383–409
DOI 10.1007/s11004-013-9456-3

Construction of Binary Multi-grid Markov Random Field Prior Models from Training Images

Håkon Toftaker · Håkon Tjelmeland

Received: 29 May 2012 / Accepted: 22 March 2013 / Published online: 15 May 2013

Abstract Bayesian modeling requires the specification of prior and likelihood models. In reservoir characterization, it is common practice to estimate the prior from a training image. This paper considers a multi-grid approach for the construction of prior models for binary variables. On each grid level we adopt a Markov random field (MRF) conditioned on values in previous levels. Parameter estimation in MRFs is complicated by a computationally intractable normalizing constant. To cope with this problem, we generate a partially ordered Markov model (POMM) approximation to the MRF and use this in the model fitting procedure. Approximate unconditional simulation from the fitted model can easily be done by again adopting the POMM approximation to the fitted MRF. Approximate conditional simulation, for a given and easy to compute likelihood function, can also be performed either by the Metropolis–Hastings algorithm based on an approximation to the fitted MRF or by constructing a new POMM approximation to this approximate conditional distribution. The proposed methods are illustrated using three frequently used binary training images.

Keywords Markov random field · Forward–backward algorithm · Multi-grid · Facies modeling · Maximum likelihood

1 Introduction

Following the seminal paper of Geman and Geman (1984), Markov random fields (MRFs) are presently used frequently as prior distributions in image analysis (Hurn et al. 2003; Li 2009; Winkler 2003). MRF is typically used only as a token prior. In the binary case, for example, the autologistic model (Besag 1974) is frequently

H. Toftaker (✉) · H. Tjelmeland
Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
e-mail: toftaker@math.ntnu.no

adopted as a prior despite that this model is not reflecting the large scale properties of the phenomenon under study. Such simple MRF priors are favored for several reasons. First, efficient Markov chain Monte Carlo simulation from the corresponding posterior distribution is usually possible and easy to implement. Second, the information content in the data is typically sufficient to remove from the posterior unrealistic large scale properties present in the prior. Thereby, it is only the small scale properties of the prior that significantly influence the posterior. The last, but perhaps most important, reason for residing with a token prior is that it is computationally difficult to construct a more realistic prior distribution. This is because discrete MRFs have a computationally intractable normalizing constant, which makes, for example, maximum likelihood estimation problematic. However, the literature includes examples (Descombes et al. 1995; Tjelmeland and Besag 1998) where the maximum likelihood estimator is found by adopting the Markov chain Monte Carlo maximum likelihood procedure (Geyer and Thompson 1995).

Spatial models for discrete variables are also important when modeling the spatial distribution of rock types in petroleum reservoirs (Eidsvik et al. 2004; Gonzalez et al. 2008; Strebelle 2002; Ulvmoen and Omre 2010). Discrete MRFs have so far attained less popularity in this type of application, mainly because the information content in the available data is often not sufficient to remove from the posterior unrealistic properties of a token prior. Available data are typically well and seismic data. Well data are exact observations of rock types in a few nodes. Seismic data is heavily blurred and has a much lower signal to noise ratio than in most image analysis applications. When using a Bayesian model formulation it therefore becomes essential to adopt a prior that honestly represents the available prior knowledge about the phenomenon under study, including the large scale properties. To compensate for the lack of realistic MRF priors for the reservoir characterization application, less formal prior formulations have won popularity. These are often termed multi-point statistics (Chatterjee et al. 2012; Journel and Zhang 2006; Strebelle 2002; Zhang et al. 2012). The prior model is defined from a training image, which is believed to be representative for the spatial phenomenon under study. This can either be a hand drawn image by a geologist, a realization from some other stochastic model, or based on outcrop data from an area believed to have a similar geological origin as the area of interest. Letting $t = (t(1), \ldots, t(n))$ denote a permutation of the integers from 1 to n, the multi-point statistics model for the joint distribution of n variables $x_1, \ldots, x_n$ is defined as a mixture distribution

$$p(x) = \sum_t p_t(x), \tag{1}$$

where $x = (x_1, \ldots, x_n)$ is the vector of the n variables, the mixture component for a permutation t is $p_t(x) = p(x_{t(1)})p(x_{t(2)}|x_{t(1)}) \cdot \cdots \cdot p(x_{t(n)}|x_{t(1)}, \ldots, x_{t(n-1)})$, and the sum is over all possible permutations t. Each of the factors $p(x_{t(i)}|x_{t(1)}, \ldots, x_{t(i-1)})$ is estimated from the training image. Clearly, the number of probabilities $p(x_{t(i)}|x_{t(1)}, \ldots, x_{t(i-1)})$ that need to be estimated is formidable, so to make the task somewhat more manageable many of them are set equal by adopting Markov assumptions. However, the resulting number of parameters that needs to be estimated from the training image is still very large. We note in passing that each of the mixture components in the multi-point statistics models is an instance of a partially or-

dered Markov model (Cressie and Davidson 1998). Many of the various multiple-point statistics models that have been proposed are quite successful in reproducing the characteristics of a wide variety of training images. As such, they are reasonable prior models. However, when conditioning on available data it is neither possible to handle the posterior distribution analytically, nor to simulate from it. To see the problem, let z denote a vector of available data and let $\psi(z|x)$ denote the corresponding likelihood function. From the Bayes theorem, we get the conditional distribution corresponding to Eq. (1)

$$p(x|z) = \frac{p(x)\psi(z|x)}{\sum_{\widetilde{x}} p(\widetilde{x})\psi(z|\widetilde{x})} = \sum_t \left[\frac{\psi(z|x)}{\sum_{\widetilde{x}} p(\widetilde{x})\psi(z|\widetilde{x})} \; p_t(x) \right]. \tag{2}$$

Accordingly, the conditional distribution corresponding to one mixture component $p_t(x)$ is

$$p_t(x|z) = \frac{p_t(x)\psi(z|x)}{\sum_{\widetilde{x}} p_t(\widetilde{x})\psi(z|\widetilde{x})}. \tag{3}$$

Solving the last expression with respect to $p_t(x)$ and inserting the result into Eq. (2), we get

$$p(x|z) = \sum_t w_t(z)\, p_t(x|z), \quad \text{where } w_t(z) = \frac{\sum_{\widetilde{x}} p_t(\widetilde{x})\psi(z|\widetilde{x})}{\sum_{\widetilde{x}} p(\widetilde{x})\psi(z|\widetilde{x})}. \tag{4}$$

Thus, $p(x|z)$ is mixture of $p_t(x|z)$ over all permutations t, with a weight $w_t(z)$ associated with $p_t(x|z)$. The problem is that the weights are computationally intractable, which in turn makes it computationally infeasible to sample from $p(x|z)$. The standard multi-point statistics solution to this problem is to modify the posterior expression to get a distribution from which it is easier to sample. The details of this depends on the type of data available. For example, if only exact values in a few nodes are observed, the most popular strategy is to restrict oneself to permutations that start with the observed nodes. This is equivalent to using Eq. (4) with equal weights $w_t(z)$ for all permutations t which have the observed nodes first, and to put $w_t(z)$ equal to zero for all other permutations. This simulation strategy is in fact often quite successful, at least as far as it is possible to evaluate from visual inspection of generated realizations. When more complicated likelihoods are of interest, it becomes more difficult to prescribe how to modify the posterior distribution, and in these cases one often sees clear visual artifacts when inspecting the generated multi-point statistics realizations.

An alternative to the multi-point statistics model is to use a partially ordered Markov model, or POMM (Cressie and Davidson 1998). A POMM is simply the $p_t(x)$ used in Eq. (1), but for a fixed permutation t. In contrast to what is used for multi-point statistics models, POMM typically adopts parametric formulas for the factors $p(x_{t(1)})$, $p(x_{t(2)}|x_{t(1)})$, $\ldots$, $p(x_{t(n)}|x_{t(1)}, \ldots, x_{t(n-1)})$. The main advantage of the POMM formulation relative to multi-point statistics models is that explicit formulas for the prior distribution are available for the POMM. Parameter estimation can thereby easily be done by adopting, for example, the maximum likelihood strategy. Moreover, conditional simulation for POMM is possible via a Markov chain Monte Carlo algorithm. The main problem with the POMM formulation is the fixed

permutation order, which typically generates artifacts in realizations from the model unless the parametric model is carefully chosen. Stien and Kolbjørnsen (2011) define a parametric POMM where this does not seem to be a problem. In the present paper, the objective is to define a prior distribution that is both able to represent the properties of typical training images in use, and for which posterior simulation via Markov chain Monte Carlo is possible. For simplicity, the attention here is limited to binary fields, but this strategy can be generalized to a situation with more than two possible values. A POMM is adopted as prior distribution, but with the POMM specified indirectly as an approximation to an MRF. Thereby, we avoid the artifacts that often occur when the POMM is explicitly specified as in the multi-point statistics mixture components. To approximate an MRF with a POMM, the strategy of Tjelmeland and Austad (2012) is adopted. For this approximation procedure to be reasonably accurate, the neighborhood size of the MRF must be reasonably small. To be able to represent both the large and small scale properties of the training image using MRFs with a small neighborhood size, it was necessary to adopt a multi-grid approach. The resulting model is thereby a product of POMMs, which is again itself a POMM.

The paper is organized as follows. In Sects. 2 and 3, the definition and some basic properties of POMMs and MRFs, are reviewed. In Sect. 4, how to find a POMM approximation to a given MRF is discussed. Thereafter, how such a POMM approximation can be used in an optimization algorithm to find the maximum likelihood estimator of the MRF for a given training image is described. In Sect. 5, the multi-grid MRF is defined and the POMM approximation is adapted to this situation. In this section, how to define a POMM approximation to a conditional multi-grid model is also discussed. Finally, Sect. 6 presents simulation examples and evaluations of the proposed procedures, and Sect. 7 provides concluding remarks.

2 Binary Partially Ordered Markov Models (POMM)

A complete introduction to POMMs can be found in Cressie and Davidson (1998). In the following only the basic concepts necessary to understand the POMM approximation to binary MRFs are introduced. Here, it is assumed that we have an $n \times m$ rectangular lattice and let $S = \{(i, j), i = 1, \ldots, n, j = 1, \ldots, m\}$ be the set of lattice nodes. To each node $(i, j) \in S$, we associate a so-called adjacent lower neighborhood $N_{ij} \subseteq S \setminus \{(i, j)\}$. These adjacent lower neighborhoods are required so that there exists a complete ordering of the lattice nodes from 1 to mn so that all nodes included in N_{ij} are ordered before node (i, j). Note that this requirement implies that at least one node $(i, j) \in S$ has $N_{ij} = \emptyset$. It should be noted that the total ordering does not need to be unique and is not a part of the POMM specification. To each node $(i, j) \in S$ of the lattice, we associate a binary variable $x_{ij} \in \{0, 1\}$ and let $x = (x_{ij}, (i, j) \in S) \in \Omega = \{0, 1\}^{mn}$. In the rest of this paper, the standard notations $x_A = (x_{ij}, (i, j) \in A)$ for $A \subseteq S$, $x_{-A} = x_{S \setminus A}$ and $x_{-(i,j)} = x_{-\{(i,j)\}}$ are also used. Letting θ denote a vector of model parameters, the joint distribution of the POMM is

$$p_\theta(x) = \prod_{(i,j) \in S} p_\theta(x_{(i,j)} | x_{N_{ij}}). \tag{5}$$

To evaluate the likelihood $p_\theta(x)$ for a given image x is straight forward from Eq. (5). In particular, the normalizing constants of the conditional distributions are readily available as these are distributions for binary variables. To sample from $p_\theta(x)$ is also easily done by simulating each x_{ij} in turn following a complete ordering as discussed above.

3 Binary Markov Random Fields

General introductions to MRFs can be found in Besag (1974), Kindermann and Snell (1980) and Cressie (1993). Here, an introduction is given to binary MRFs defined on a rectangular lattice. As in the previous section, we assume that we have an $n \times m$ rectangular lattice and denote the set of lattice nodes by $S = \{(i, j), i = 1, \ldots, n, j = 1, \ldots, m\}$. To each node $(i, j) \in S$ a set of neighbor nodes $\partial(i, j) \subseteq S \setminus \{(i, j)\}$ is associated, where the neighborhood is required to be symmetric in that $(i, j) \in \partial(r, s) \Leftrightarrow (r, s) \in \partial(i, j)$ for any distinct pairs $(i, j), (r, s) \in S$. Following common practice (i, j) and (r, s) are declared neighbors whenever $(r, s) \in \partial(i, j)$. A set $C \subseteq S$ is said to be clique if $(r, s) \in \partial(i, j)$ for all distinct pairs $(i, j), (r, s) \in C$. We let $\mathcal{C}$ denote the set of all cliques. As in the above a binary variable $x_{ij} \in \{0, 1\}$ is associated to each $(i, j) \in S$ and we let $x = (x_{ij}, (i, j) \in S) \in \Omega = \{0, 1\}^{mn}$. Again letting θ denote a vector of model parameters, x is then said to be a binary MRF with respect to the given neighborhood system if the joint distribution $p_\theta(x) > 0$ for all $x \in \Omega$, and the full conditionals fulfill the Markov assumption

$$p_\theta(x_{ij}|x_{-(i,j)}) = p_\theta(x_{ij}|x_{\partial(i,j)}), \tag{6}$$

for all $x \in \Omega$. The positivity condition $p_\theta(x) > 0$ ensures that there exists an energy function $U_\theta(x)$ so that the joint distribution $p_\theta(x)$ can be expressed as

$$p_\theta(x) = c(\theta) \exp\{-U_\theta(x)\}, \tag{7}$$

where $c(\theta)$ is a normalizing constant. As indicated in the notation, the normalizing constant $c(\theta)$ will be a function of θ. The Hammersley–Clifford theorem (Besag 1974; Clifford 1990) states that given the Markov property in Eq. (6) the most general form the energy function can take is

$$U_\theta(x) = \sum_{C \in \mathcal{C}} V_C(x_C, \theta), \tag{8}$$

where the potential function $V_C(x_C, \theta) \in (-\infty, \infty)$ is an arbitrary function of x_C and θ.

Simulation from a given MRF $p_\theta(x)$, both unconditionally and conditioned on observed data, is relatively straight forward by the Metropolis–Hastings algorithm, see for example the references given above. Estimation of the parameter vector θ from one or more training images is computationally a lot more problematic. The main reason for this is the computationally intractable normalizing constant $c(\theta)$. Clearly,

$$c(\theta) = \left[\sum_{x \in \Omega} \exp\{-U_\theta(x)\}\right]^{-1}, \tag{9}$$

but the number of terms in this sum is typically much too large to be used to compute $c(\theta)$. Thereby, for example, numerical maximization of the likelihood function to find the maximum likelihood estimator (MLE) for θ is not directly possible. It is, however, possible to find an approximate MLE from a set of Markov chain Monte Carlo samples (Geyer and Thompson 1995).

4 Forward–Backward Algorithm and the POMM Approximation

In this section, an exact forward–backward algorithm (Künsch 2001; Pettitt et al. 2003; Scott 2002) for a binary MRF $p_\theta(x)$ is described. As this exact procedure is practical only for MRFs defined on small lattices and with small neighborhoods, the corresponding approximate algorithm of Tjelmeland and Austad (2012), which produces a POMM approximation $\widetilde{p}_\theta(x)$ to the MRF is reviewed. Finally, how realizations from $\widetilde{p}_\theta(x)$ can be used in an optimization algorithm to find the maximum likelihood estimate of θ for a given image x is discussed.

4.1 The Exact Forward–Backward Algorithm

Bartolucci and Besag (2002), Friel and Rue (2007) and Friel et al. (2009) define forward–backward algorithms that can be run for binary MRFs whenever both the neighborhoods and one of the lattice dimensions are sufficiently small. These forward–backward algorithms are based on an ordering of the lattice nodes from 1 to mn. Let $\rho(i,j)$ denote the number assigned to node (i,j) and let $\rho^{-1}(\cdot)$ be the corresponding inverse mapping so that $k=\rho(i,j) \Leftrightarrow (i,j)=\rho^{-1}(k)$. For example, one may use the lexicographical ordering where $\rho(i,j)=(i-1)n+j$. The forward part of the forward–backward algorithm sequentially computes

$$p_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}}) \quad \text{for } k=2,\ldots,mn, \tag{10}$$

by summing out $x_{\rho^{-1}(k)}$, $k=1,\ldots,mn-1$ in turn. It should be noted that the Markov property of the original $p_\theta(x)$ induces a Markov property also in Eq. (10). In particular $x_{\rho^{-1}(k)}$ is connected only to x_{A_k} for a subset $A_k \subseteq \{\rho^{-1}(l),\, l=k+1,\ldots,mn\}$, so that Eq. (10) can be decomposed into a product

$$p_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}}) = g_\theta(x_{\rho^{-1}(k)}, x_{A_k}) h_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}). \tag{11}$$

When summing out $x_{\rho^{-1}(k)}$ from Eq. (10) the $h_\theta(\cdot)$ factor can be put outside the summation sign. Thereby, the computational complexity of this step of the algorithm becomes $2^{|A_k|+1}$, where $|A_k|$ is the number of elements in the set A_k. From Eq. (10), the conditional distribution

$$p_\theta(x_{\rho^{-1}(k)} | x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}) = \frac{p_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})}{p_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}})} \propto g_\theta(x_{\rho^{-1}(k)}, x_{A_k}) \tag{12}$$

for $k=1,\ldots,mn-1$ is readily available. Thus, we have the decomposition

$$p_\theta(x) = p_\theta(x_{\rho^{-1}(mn)}) \prod_{k=1}^{mn-1} p_\theta(x_{\rho^{-1}(k)} | x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}). \tag{13}$$

It should be noted that Eq. (13) is now expressed as a POMM where the lower adjacent neighborhood to node (i, j) is $A_{\rho(i,j)}$. In particular, computation of the likelihood $p_\theta(x)$ for a given image x is straight forward and simulation from $p_\theta(x)$ is easily done by sampling $x_{\rho^{-1}(mn)}, x_{\rho^{-1}(mn-1)}, \ldots, x_{\rho^{-1}(1)}$ in turn.

4.2 The Approximation

As mentioned above, the exact forward–backward algorithm is practical only for MRFs with small neighborhoods on small lattices, as otherwise most of the sets $A_1, \ldots, A_{mn-1}$ become so large that the algorithm is infeasible. Tjelmeland and Austad (2012) define an approximate forward–backward algorithm that is possible to run also for larger neighborhoods and lattice sizes. The approximate algorithm follows the same structure as the exact one. First one defines $\widetilde{p}_\theta(x) = p_\theta(x)$ and sequentially for $l = 2, \ldots, mn$ computes approximations $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})$ to Eq. (10). To compute $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}})$ from $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})$, one uses the same type of decomposition as in Eq. (11). Assuming $x_{\rho^{-1}(k)}$ in $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})$ is connected only to $x_{\widetilde{A}_k}$ for $\widetilde{A}_k \subseteq A_k$, an exact marginalization is performed whenever $|\widetilde{A}_k| \leq \kappa$, where κ is an input parameter to the algorithm. Thus, when $|\widetilde{A}_k| \leq \kappa$, we have

$$\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}) = \sum_{x_{\rho^{-1}(k)}=0}^{1} \widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}}). \tag{14}$$

If $|\widetilde{A}_k| > \kappa$, an approximation is introduced to reduce the computational complexity of the marginalization operation. First, a sum of squares approximation for $\ln \widehat{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})$ to $\ln \widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})$ is defined where x_k in $\widehat{p}_\theta(\cdot)$ is restricted to be connected only to a set $\widehat{A}_k$ of κ other variables. The details of this approximation procedure are described in Tjelmeland and Austad (2012). Thereby, we get a decomposition of $\widehat{p}_\theta(x_{\rho^{-1}(l),\, l=k,\ldots,mn})$ corresponding to Eq. (11)

$$\widehat{p}_\theta(x_{\rho^{-1}(l),\, l=k,\ldots,mn}) = \widehat{g}_\theta(x_{\rho^{-1}(k)}, x_{\widehat{A}_k})\widehat{h}_\theta(x_{\rho^{-1}(l),\, l=k+1,\ldots,mn}). \tag{15}$$

Thereafter, $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}})$ is defined as in Eq. (14), but with $\widetilde{p}_\theta(x_{\rho^{-1}(l),\, l=k,\ldots,mn})$ substituted by the new approximation $\widehat{p}_\theta(x_{\rho^{-1}(l),\, l=k,\ldots,mn})$. From the approximate distributions $\widetilde{p}(x_{\rho^{-1}(l),\, l=k,\ldots,mn})$ an approximate joint distribution $\widetilde{p}_\theta(x)$ is defined by following the same structure as in Eqs. (12) and (13). Thus,

$$\widetilde{p}_\theta(x) = \widetilde{p}_\theta(x_{\rho^{-1}(mn)}) \prod_{k=1}^{mn-1} \widetilde{p}_\theta(x_{\rho^{-1}(k)} | x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}), \tag{16}$$

where

$$\begin{aligned} &\widetilde{p}_\theta(x_{\rho^{-1}(k)} | x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}}) \\ &\quad = \frac{\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}})}{\widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k+1,\ldots,mn\}})} \propto \widetilde{p}_\theta(x_{\{\rho^{-1}(l),\, l=k,\ldots,mn\}}) \end{aligned} \tag{17}$$

for $k = 1, \ldots, mn - 1$. As for the exact decomposition in Eq. (13), $\widetilde{p}_\theta(x)$ is here expressed as a POMM, where the lower adjacent neighborhood for node (i, j) is $\widetilde{A}_{\rho^{-1}(i,j)}$. Both sampling from Eq. (16) and computation of the likelihood for a given image x is thereby straight forward. For an observed image x, it may also be tempting to try to find an approximation to the maximum likelihood estimator by optimizing numerically $\widetilde{p}_\theta(x)$ with respect to θ. However, this may become problematic as the approximation $\widetilde{p}_\theta(x)$ is not continuous or differentiable as a function of θ. Thereby, such a numerical optimization procedure may quickly become stuck in a local maximum induced by the approximation. Below the maximization of $p_\theta(x)$, and in particular how the POMM approximation can be used to bypass the problem with the computationally intractable normalizing constant $c(\theta)$ in $p_\theta(x)$, is considered.

4.3 Maximum Likelihood Estimation by Importance Sampling

To cope with the computationally intractable normalizing constant $c(\theta)$ in $p_\theta(x)$, importance sampling is used. The general strategy is outlined in Geyer and Thompson (1995) and a more detailed algorithm is given in Tjelmeland (1996), both in a situation where Markov chain Monte Carlo is used to generate dependent samples from the distribution in question. As independent samples from the POMM approximation $\widetilde{p}_\theta(x)$ can be generated, the situation considered here is somewhat simpler than in the two references just cited. To simplify notation, let $\varphi_\theta(x) = \exp\{-U_\theta(x)\}$ so that $p_\theta(x) = c(\theta)\varphi_\theta(x)$. Using that $c(\theta)$ is given by Eq. (9), we get for a fixed parameter vector θ^0

$$\begin{aligned}\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)} &= \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \sum_{z\in\Omega} \varphi_\theta(z) = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \sum_{z\in\Omega} \left[\frac{\varphi_\theta(z)}{\widetilde{p}_{\theta^0}(z)} \widetilde{p}_{\theta^0}(z)\right] \\ &= \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \mathrm{E}\left[\frac{\varphi_\theta(z)}{\widetilde{p}_{\theta^0}(z)}\right], \end{aligned} \tag{18}$$

where the expectation is given with respect to $z \sim \widetilde{p}_{\theta^0}(\cdot)$. Thereby, for any value of the parameter vector θ, an unbiased estimate of $\widetilde{p}_{\theta^0}(x)/p_\theta(x)$ is

$$\widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)}} = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \cdot \frac{1}{R} \sum_{r=1}^{R} \frac{\varphi_\theta(z^r)}{\widetilde{p}_{\theta^0}(z^r)}, \tag{19}$$

where $z^1, \ldots, z^R$ are independent realizations from $\widetilde{p}_{\theta^0}(\cdot)$. One should note that the reason for considering $\widetilde{p}_{\theta^0}(x)/p_\theta(x)$ and not the inverse quantity, is that no unbiased estimate of the inverse quantity is available. Having Eq. (19) available, it is tempting to find an approximate maximum likelihood estimate for θ by numerically minimizing Eq. (19) with respect to θ. However, this is not a recommendable procedure because for parameter vectors θ far away from the fixed θ^0, the Monte Carlo variance of Eq. (19) may become very large. The numerical optimization of Eq. (19) should therefore be stopped whenever the (estimated) variance of the decrease obtained by doing the next step of the optimization algorithm becomes too large compared to the estimated decrease itself. The θ^0 should then be redefined to take the value of θ at this point in the optimization procedure and a new POMM approximation should be

constructed with new realizations $z_1, \ldots, z_R$ generated to obtain a new estimate of Eq. (19) with lower variance close to the current value of θ. An unbiased estimate of the decrease of the function $\widetilde{p}_{\theta^0}(x)/p_\theta(x)$ when going from θ to θ' is

$$\frac{\widehat{\widetilde{p}_{\theta^0}(x)}}{p_\theta(x)} - \frac{\widehat{\widetilde{p}_{\theta^0}(x)}}{p_{\theta'}(x)} = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)}\frac{\varphi_\theta(z^r)}{\varphi_\theta(x)} - \frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)}\frac{\varphi_{\theta'}(z^r)}{\varphi_{\theta'}(x)}\right], \tag{20}$$

and the corresponding empirical variance of each term in this sum is

$$\widehat{\sigma}^2(\theta,\theta') = \frac{1}{R-1}\sum_{r=1}^{R}\left[\frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)}\frac{\varphi_\theta(z^r)}{\varphi_\theta(x)} - \frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)}\frac{\varphi_{\theta'}(z^r)}{\varphi_{\theta'}(x)} - \left(\frac{\widehat{\widetilde{p}_{\theta^0}(x)}}{p_\theta(x)} - \frac{\widehat{\widetilde{p}_{\theta^0}(x)}}{p_{\theta'}(x)}\right)\right]^2. \tag{21}$$

Assuming the estimate in Eq. (20) to be negative, the optimization procedure should then be stopped whenever the absolute value of the estimated decrease is larger than some given multiple, γ say, of $\sqrt{\widehat{\sigma}^2(\theta,\theta')/R}$. Our experience is that it is beneficial to start out with a large value for γ and then gradually decrease this value as one approaches the maximum likelihood estimate. To save computation time, it is also natural to start with a small number of realizations R and later increase this number. Suggestions for a detailed procedure for how to change γ and R can be found in Tjelmeland (1996). One should note that a requirement for the above optimization algorithm to be successful in finding the maximum likelihood estimate is that the POMM approximation is accurate enough to give a sufficiently small variance $\widehat{\sigma}^2(\theta,\theta')$ at least when θ and θ' is close to θ^0, as otherwise the optimization procedure will become stuck. In practice, the above optimization algorithm can thereby only be used for MRFs with reasonably small neighborhoods and, as also discussed in Sect. 1, an MRF with a small neighborhood is typically not able to represent both the small and large scale properties of frequently used training images. To cope with this complication, we next introduce a multi-grid version of MRFs and adapt the POMM approximation to such a situation. Whenever a numerical optimization algorithm is run, there is a risk of getting trapped in a local optimum. To test for this in the simulation examples in Sect. 6, for each optimization problem multiple optimizations with different starting values are run. These optimizations for all cases resulted in the same optimum, thus confirming that this potential complication turned out not to be relevant for these examples. However, there is clearly no guarantee that the same will happen for other training images.

5 Multi-grid MRF and POMM Approximation

In this section, a general multi-grid MRF is defined and adapted to the POMM approximation defined in Sect. 4.2, and applied to this situation. Thereafter, how to construct a POMM approximation to a corresponding conditional distribution is discussed, before the parameter estimation procedure discussed in Sect. 4.3 is adapted to the multi-grid MRF situation.

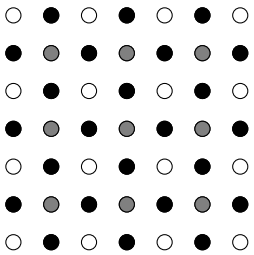

Fig. 1 Illustration of the splitting of node set S for a 7×7 lattice into three sub-lattices S_1, S_2, and S_3. The *white nodes* are in S_1, the *gray nodes* in S_2, and the *black nodes* in S_3

5.1 Multi-grid MRF

In the multi-grid approach, the nodes in our rectangular lattice S are split into a series of an odd number, T say, of sub-lattices $S_1, \ldots, S_T$. Figure 1 illustrates this process when $T = 3$. The first sub-lattice, S_1, is a rectangular lattice of dimensions $n_1 \times m_1$ say, where $n_1 < n$ and $m_1 < m$. The next sub-lattice, S_2, is an $(n_1 - 1) \times (m_1 - 1)$ rectangular lattice where the nodes in S_2 are placed between the nodes in S_1, as illustrated in Fig. 1. The nodes in the sub-lattice S_3 are placed between the nodes in $S_1 \cup S_2$, again illustrated in the same figure. One should note that the nodes in S_3 do not form a rectangular lattice, but if we look at the nodes at a 45° angle they are still organized into rows and columns. If $T \geq 5$, the sub-lattice S_4 is a $2(n_1 - 1) \times 2(m_1 - 1)$ rectangular lattice where the nodes are placed between the nodes in $S_1 \cup S_2 \cup S_3$, corresponding to how the nodes in S_2 are placed between the nodes in S_1. The nodes in S_5 are placed between the nodes in $S_1 \cup S_2 \cup S_3 \cup S_4$ corresponding to how the nodes in S_3 are placed between the nodes in $S_1 \cup S_2$. This structure is then continued up to sub-lattice S_T. The number of nodes in the various sub-lattices become $|S_1| = n_1 m_1$, $|S_t| = 2^{t-2}(n_1 - 1)(m_1 - 1)$ when t is even, and $|S_t| = 2^{t-2} n_1 m_1 + (2^{(t-3)/2} - 2^{t-2})(n_1 + m_1) + 2^{t-2} - 2^{(t-3)/2+1}$ when $t > 1$ is odd. The joint distribution for $x = (x_{(i,j)}, (i, j) \in S)$ is specified by the marginal distribution for x_{S_1} and, for each $t = 2, \ldots, T$, the conditional distribution for x_{S_t} given $x_{S_{1:t-1}}$, where $S_{1:t-1} = S_1 \cup \cdots \cup S_{t-1}$. A separate parameter vector is adopted for each of these T distributions, denoted by $\theta_1, \ldots, \theta_T$, respectively. Thereby, we have

$$p_\theta(x) = p_{\theta_1}(x_{S_1}) \prod_{t=2}^{T} p_{\theta_t}(x_{S_t} | x_{S_{1:t-1}}), \tag{22}$$

where $\theta = (\theta_1, \ldots, \theta_T)$. For the marginal distribution $p_{\theta_1}(x_{S_1})$ an MRF exactly as discussed in Sect. 3 is adopted, whereas for each of $p_{\theta_t}(x_{S_t} | x_{S_{1:t-1}})$ an MRF where the conditioning variables are included as covariates is adopted. It should be noted that the normalizing constant in the model $p_{\theta_t}(x_{S_t} | x_{S_{1:t-1}})$ becomes a function of not only the parameter vector θ_t, but also the conditioning variables $x_{S_{1:t-1}}$. In Sect. 6.1, the neighborhood structure and parametric form of the potential functions used in the simulation examples are specified. Now the focus is on how to apply the POMM approximation to $p_{\theta_1}(x_{S_1})$ and each of $p_{\theta_t}(x_{S_t} | x_{S_{1:t-1}})$.

5.2 POMM Approximations for the Multi-grid MRF

To get a POMM approximation to the multi-grid MRF defined above, the approximation scheme discussed in Sect. 4 can be adopted to both $p_{\theta_1}(x_{S_1})$ and $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}), t = 2, \ldots, T$. The $p_{\theta_1}(x_{S_1})$ is an MRF exactly as discussed in Sect. 3, so the approximation scheme defined in Sect. 4 can be directly applied. For $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$, at least two possibilities exist for how to cope with the conditioning variables. One may either find a POMM approximation for specific values of the conditioning variables, or one may construct a general POMM approximation as a function of $x_{S_{1:t-1}}$. In the following, details of the two alternatives are discussed in turn.

5.2.1 A First POMM Approximation for $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$

If the purpose of computing the POMM approximation is to evaluate (approximately) the likelihood function in order to find, for example, the maximum likelihood estimator for θ based on a given training image, observed values are available for $x_{S_{1:t-1}}$ and one may insert these values in $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$. Thereafter, the POMM approximation defined in Sect. 4 can be directly applied. This approximation is denoted by $\widehat{p}_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ and the corresponding approximation of the joint distribution by $\widehat{p}_{\theta}(x)$. The strategy of inserting values for the conditioning variables $x_{S_{1:t-1}}$ can also be used if the goal is to simulate unconditionally (and approximately) from $p_{\theta}(x)$. It is then natural to simulate each of x_{S_t} for $t = 1, \ldots, T$ in turn. Thus, when x_{S_t} is to be simulated, values for $x_{S_{1:t-1}}$ have already been simulated and can thereby be inserted in $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$. Thereafter, the POMM approximation can be established and values for x_{S_t} can be simulated by a backward pass. One should note, however, that if it is of interest to generate several realizations from $p_{\theta}(x)$, this implies that new POMM approximations must be established for each of $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ for each realization. Moreover, it is not possible to use this POMM approximation scheme to efficiently generate conditional realizations of x given some components in x. The second approximation scheme for $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$, which can also be used for conditional simulation is discussed next.

5.2.2 A Second POMM Approximation for $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$

To see how to define a POMM approximation for $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ without inserting specific values for the conditioning variables, first recall that the model is specified via an energy function $U_{\theta_t}(\cdot)$, so corresponding to Eq. (7) we have

$$p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = c(\theta_t, x_{S_{1:t-1}}) \exp\bigl\{-U_{\theta_t}(x_{S_{1:t}})\bigr\}, \tag{23}$$

where $c(\theta_t, x_{S_{1:t-1}})$ is the computationally intractable normalizing constant, now a function of both the parameter vector θ_t and the conditioning variables $x_{S_{1:t-1}}$. To see how to cope with this intractable normalizing constant consider first the following distribution for x_{S_t},

$$f_{\theta_t}(x_{S_{1:t}}) \propto \exp\bigl\{-U_{\theta_t}(x_{S_{1:t}})\bigr\}, \tag{24}$$

noting that the corresponding conditional distribution for x_{S_t} given $x_{S_{1:t-1}}$, and marginal distribution for $x_{S_{1:t-1}}$ becomes

$$f_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = p_\theta(x_{S_t}|x_{S_{1:t-1}}) \quad \text{and} \quad f_{\theta_t}(x_{S_{1:t-1}}) = \frac{1}{c(\theta_t, x_{S_{1:t-1}})}, \tag{25}$$

respectively. As $f_{\theta_t}(x_{S_{1:t}})$ is an MRF, the approximation scheme defined in Sect. 4 can be directly applied to this distribution. Adopting a node order rule $\rho(\cdot,\cdot)$ where the nodes in S_t are assigned numbers from 1 to $|S_t|$, and stopping the summation procedure when the first $|S_t|$ (approximate) summations are finished results in approximations to the two distributions in Eq. (25). As detailed in Sect. 4.2, the approximation to the conditional distribution is given as a product of univariate conditional distributions, that is

$$\widetilde{p}_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = \prod_{k=1}^{|S_t|} \widetilde{f}_{\theta_t}(x_{\rho^{-1}(k)}|x_{\rho^{-1}(l)}, l = k+1, \ldots, |S_{1:t}|), \tag{26}$$

whereas the approximation to $f_{\theta_t}(x_{S_{1:t-1}})$, which we denote by

$$\widetilde{f}_{\theta_t}(x_{S_{1:t-1}}) = \frac{1}{\widetilde{c}(\theta_t, x_{S_{1:t-1}})}, \tag{27}$$

has no special form. Performing the procedure discussed above for each $t = 2, \ldots, T$ and combining the results two alternative approximations of $p_\theta(x)$ are obtained. Replacing the computationally intractable normalizing constants for each of $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ with the corresponding approximation given in Eq. (27) results in the following approximation

$$\widetilde{p}_\theta(x) \propto \exp\{-U_{\theta_1}(x_{S_1})\} \prod_{t=2}^{T} \widetilde{c}(\theta_t, x_{S_{1:t-1}}) \exp\{-U_{\theta_t}(x_{S_{1:t}})\}, \tag{28}$$

whereas by combining the approximations in Eq. (26), we obtain the approximation

$$p^\star_\theta(x) = \widetilde{p}_{\theta_1}(x_{S_1}) \prod_{t=2}^{T} \widetilde{p}_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}), \tag{29}$$

where $\widetilde{p}_{\theta_1}(x_{S_1})$ is the POMM approximation to $p_{\theta_1}(x_{S_1})$. The latter approximation, $p^\star_\theta(x)$, is given as a product of univariate conditional distributions and is then by definition a POMM. Thereby, unconditional realizations from $p^\star_\theta(x)$ can be generated very efficiently once the POMM approximation is established. This is in contrast to the situation for the first POMM approximation discussed above, where the generation of each realization requires a number of new POMM approximations to be established. The approximation $\widetilde{p}_\theta(x)$ is not a POMM and direct simulation from the distribution is not possible. However, up to a normalizing constant an explicit formula for the distribution is available, and thus a Metropolis–Hastings algorithm can be used to generate samples from the distribution. It is also interesting to note that $p^\star_\theta(x)$ can be obtained as a POMM approximation to $\widetilde{p}_\theta(x)$ by adopting the approximation scheme discussed in Sect. 4 if letting the nodes in S_T be numbered from 1 to $|S_T|$, the nodes in S_{T-1} be numbered from $|S_T|+1$ to $|S_T \cup S_{T-1}|$ and so on, and letting the nodes within each S_t be numbered in the same order as used

when constructing Eq. (27). It is therefore reasonable to consider $\widetilde{p}_\theta(x)$ to be a better approximation than $p^\star_\theta(x)$.

5.3 Conditional Simulation

Let $p_\theta(x)$ be a multi-grid MRF for x as defined above and let $\widetilde{p}_\theta(x)$ and $p^\star_\theta(x)$ be the corresponding approximations defined by Eqs. (28) and (29), respectively. Further let z denote a vector of observed quantities which is related to x via a likelihood function $\psi(z|x)$. In the following, we assume the likelihood $\psi(z|x)$ to be known and easy to compute. For example, z may contain exact observations of some elements in x, or z may be of the same dimension as x and the components of z may contain conditionally independent noisy observations of each component of x. The resulting conditional distribution $p_\theta(x|z)$ is clearly not computationally feasible, but in the following how to define and simulate from approximations to this conditional distribution is discussed. For each of the two approximations $\widetilde{p}_\theta(x)$ and $p^\star_\theta(x)$ to $p_\theta(x)$, there are corresponding approximations to $p_\theta(x|z) \propto p_\theta(x)\psi(z|x)$, namely $\widetilde{p}_\theta(x|z) \propto \widetilde{p}_\theta(x)\psi(z|x)$ and $p^\star_\theta(x|z) \propto p^\star_\theta(x)\psi(z|x)$. Direct simulation is not possible from either of these but up to a normalizing constant explicit formulas are available for both so simulation can be done with a suitable Metropolis–Hastings algorithm. As discussed above, $\widetilde{p}_\theta(x)$ is the better approximation to $p_\theta(x)$, so it is reasonable to also assume that $\widetilde{p}_\theta(x|z)$ is the better approximation to $p_\theta(x|z)$. Moreover, as the computational complexity of the Metropolis–Hastings algorithms of the two approximate conditional distributions are essentially the same, it is recommended to use $\widetilde{p}_\theta(x|z)$ as the approximation to $p_\theta(x|z)$. An alternative to using the Metropolis–Hastings algorithm to simulate from $\widetilde{p}_\theta(x|z)$ is to establish a corresponding POMM approximation. The approximate distribution $\widetilde{p}_\theta(x|z)$ is an MRF, so it can be fed into the approximation scheme in Sect. 4. Independent realizations can thereafter be efficiently generated from the resulting POMM approximation. For this last POMM approximation, we find it reasonable to use the ordering of the nodes defined in the end of Sect. 5.2. In particular, this produces an internal consistency in the approximations as it implies that if there is no data (i.e. z is empty) the resulting POMM approximation becomes $p^\star_\theta(x)$.

5.4 Parameter Estimation by Maximum Likelihood

Let x be a given training image to which we want to fit a multi-grid MRF. The maximum likelihood principle is adopted to estimate $\theta = (\theta_1, \ldots, \theta_T)$ and thus the likelihood function in Eq. (22) must be maximized with respect to θ. As the multi-grid MRF is specified with a separate parameter vector θ_t to each of the T MRF components, the maximization can be done with respect to each θ_t separately. Moreover, as each model component is an MRF the optimization procedure specified in Sect. 4.3 can be directly applied. Note that for the parameter estimation procedure the POMM approximation only needs to be available for the values of the conditional variables that appear in the training image. In the estimation of θ_t for $t > 1$, it is therefore natural to use the first POMM approximation discussed in Sect. 5.2.

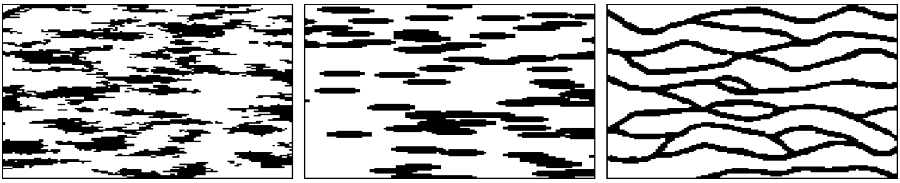

Fig. 2 The three training images used in the evaluation of our approximate model fitting procedures

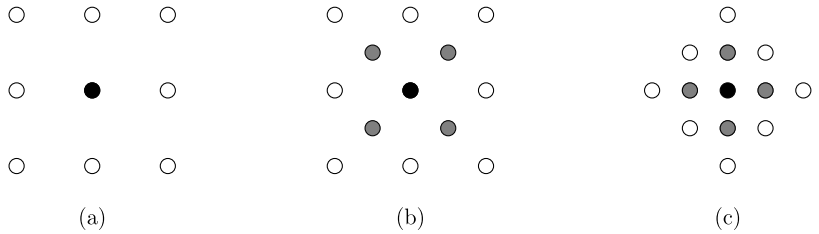

Fig. 3 The neighborhoods used in the multi-grid MRF for nodes not located on the boundary of the lattice. (**a**) Neighborhood for sub-lattice S_1. (**b**) Neighborhood for S_t when t is even. (**c**) Neighborhood for S_t when $t > 1$ is odd

6 Examples

To evaluate the performance of the approximation scheme it is applied to the three 121×121 training images shown in Fig. 2. For all three training images, we let S_1 be a 16×16 lattice and use $T = 7$ sub-lattices. The total lattice S then becomes 121×121. In the following, details are given of the parametric form for the multi-grid MRF used in the examples. Thereafter, we define the node ordering used for the approximations within each level, and finally present numerical examples.

6.1 Parametric Multi-grid MRF Used in the Simulation Examples

In this section, the exact neighborhood system and parametric energy functions used in the numerical examples presented below are defined. Large neighborhoods and an energy function with many parameters clearly give flexible models that can be fitted to a large variety of training images. However, the computational cost of the fitting and simulation process grows quickly with the neighborhood size and the dimension of the parameter vector. It is also important not to include too many parameters so as to avoid overfitting. Lastly, the multi-grid MRF structure defined above reduces the need for a large neighborhood system and many parameters in each level of the model. In the MRF for x_{S_1} we use a second-order neighborhood system. Then each interior node has eight neighbors as illustrated in Fig. 3(a). The white colored nodes are neighbors of the black node. The number of neighbors for the nodes on the boundary of the lattice is correspondingly reduced, so that the four corner nodes have only

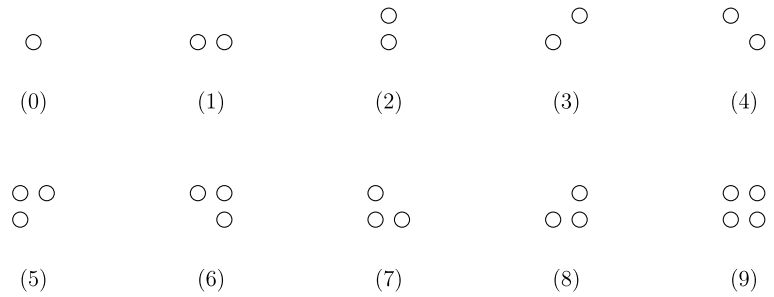

Fig. 4 Clique types used in the definition of $p_{\theta_1}(x_{S_1})$

three neighbors and other boundary nodes have five neighbors. With this neighborhood system, we get cliques of up to four nodes, and assuming translation invariant potential functions, we then have ten clique types to consider; see Fig. 4. For each of these clique types, a corresponding parameter is associated ($\theta_{1,k}$ for clique type (k) in Fig. 4) and for clique type (k) we adopt the potential function

$$V_C(x_C, \theta_1) = \theta_{1,k} \prod_{(i,j)\in C} x_{ij}, \tag{30}$$

where θ_1 is a vector of the model parameters. Thus, the potential for a clique is equal to the value of the associated parameter if all nodes in the clique have value one, and the potential is zero otherwise. Without loss of generality, one of the ten parameters can be set as equal to zero, thus for the rest of this paper we fix $\theta_{1,0} = 0$ and are left with the parameter vector $\theta_1 = (\theta_{1,1}, \ldots, \theta_{1,9})$ that has to be estimated from the training image.

In the conditional MRF for x_{S_t} when t is even a second-order neighborhood model is again adopted, but in addition each node $(i, j) \in S_t$ is associated with a set B_{ij} that contains the four nodes in $S_{1:t-1}$ that are located closest to (i, j). An illustration is given in Fig. 3(b), where the gray nodes are the four nodes in B_{ij}. For the energy function, the following parametric form

$$U_{\theta_t}(x_{S_t}, x_{S_{1:t-1}}) = U^1_{\theta_{t,1:9}}(x_{S_t}) + \sum_{(i,j)\in S_t} U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}}), \tag{31}$$

is adopted, where the parameter vector θ_t has nineteen elements and is split into $\theta_{t,1:9}$ and $\theta_{t,10:19}$. These contain the first nine and the remaining elements of θ_t, respectively. For $U^1_{\theta_{t,1:9}}(x_{S_t})$ exactly the same parametric form as for $U_{\theta_1}(x_{S_1})$ is adopted. For the specification of $U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}})$, a similar strategy as for that of the energy function for x_{S_1} is adopted, but we include only terms corresponding to one and two elements in B_{ij}. More precisely, $U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}})$ is a sum of ten terms, one for each of the node sets shown in Fig. 5, where node (i, j) is shown in black and the nodes in B_{ij} are shown in gray. The potential function corresponding to node set (k) in Fig. 5 is

$$V_C(x_{ij}, x_C, \theta_{t,10:19}) = \theta_{tk}\, x_{ij} \prod_{(r,s)\in C} x_{rs}, \tag{32}$$

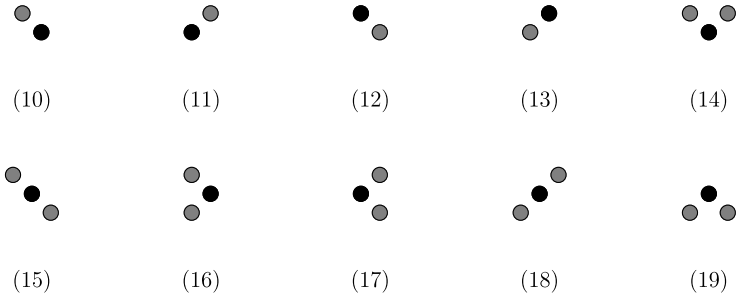

Fig. 5 Sets, numbered from ten to nineteen, used to define the potential functions building up the energy function $U^2_{\theta_t,10:19}(x_{ij}, x_{B_{ij}})$

where C is the set of gray nodes in the figure. One should note that with Eq. (31) the conditioning variables $x_{S_{1:t-1}}$ only affect the first order effects, corresponding to clique type (0) in Fig. 4. It is clearly possible to generalize the model definition to allow the conditioning variables to modify also the pairwise, triple, and quadruple interactions, but we have chosen not to do so here because this will result in a dramatic increase in the number of parameters.

When $t > 1$ is odd, the nodes in S_t are organized in a lattice that is rotated 45° relative to the lattices making up S_1 and S_t for t even (Fig. 1). We define the energy function for the conditional MRF for x_{S_t} when $t > 1$ is odd in the same way as we did for t even, except that all cliques and sets B_{ij} are rotated 45° clockwise; the resulting neighborhood is shown in Fig. 3(c). The total number of components in the parameter vector of the multi-grid MRF becomes $19T - 10$.

6.2 Numbering of Nodes Used in the Simulation Examples

To fully define the POMM approximation used in the simulation examples, it remains to define the node numbering of the approximate forward–backward algorithm. The nodes in each of S_1 and S_t when t is even constitute rectangular lattices and the lexicographical ordering of the nodes is used. As mentioned above, the nodes in S_t when $t > 1$ is odd can be seen as nodes in a lattice that is rotated 45° relative to a rectangular lattice; to explain our numbering here refer to Fig. 1. The black nodes (i.e. S_3) in this 7×7 lattice are numbered in the order (6, 1), (7, 2), (4, 1), (5, 2), (6, 3), (7, 4), (2, 1), and so on.

6.3 Computational Parameters Used in the Simulation Examples

In the computation of the maximum likelihood estimator, the values of the parameters γ and R defined in Sect. 4.3 must be specified. This begins with $\gamma = 3.5$ and $R = 100$. If the relative decrease of the estimated likelihood is less than 0.05, we let $\gamma := \gamma/2$ and $R := R * 2$ and if the decrease is greater than 0.95 we assign $\gamma := \gamma * 2$ and $R := R/2$. The optimization is stopped when $R = 3200$ and the decrease is less than 0.05. In the computations, the POMM approximations depend on the value of κ. In all examples presented, $\kappa = 12$ is used, which we think is a reasonable trade off

between approximation quality and computational complexity. The same examples for $\kappa = 14$ have also been run, without detecting significant differences in the fitted models.

6.4 Simulation Examples

In this section, the results of the model fitting procedure for the three training images in Fig. 2 are presented. How well the features of the training images are reproduced by the models, and the quality of the different approximations introduced above are investigated. First, a comment is made on the efficiency of the likelihood optimization. Second, a look at realizations from the fitted $\widehat{p}(x|\theta)$ is presented. Such a simple visual inspection gives a good indication of the quality of the model, but to get a more accurate measure of this, the descriptive statistics introduced in Stien and Kolbjørnsen (2011) are also used. Third, realizations from $p^{\star}_{\theta}(x)$ are presented. This distribution is investigated in the same way and compared with the previous approximation. As $p^{\star}_{\theta}(x)$ is a POMM, the resulting lower adjacent neighborhood is also studied. Lastly, the approximations to the conditional distribution $p_{\theta}(x|z)$ are explored.

The optimization of the likelihood is done by estimating the likelihood by importance sampling. When estimating θ_t, independent samples from the POMM approximation of $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ are needed. Finding this POMM approximation is computationally the most demanding part of the algorithm. The number of POMM approximations we need to compute varies quite a lot. Starting with a parameter vector of only zeros, the maximum number of POMM approximations we need to compute to reach the MLE was approximately one hundred.

Each of the training images have some distinct features that put the model fitting procedure to the test. In training image (a), the shape of the black objects are very irregular, and thus the best results are expected for this training image. Image (b), by comparison, has very regular black objects which are mostly convex. It may not be possible to reproduce this very regular shape in realizations from the fitted model. Training image (c) has objects which extend from one side of the lattice to the other and will be the hardest test for our fitting procedure. For each of the fitted models, realizations are generated from $\widehat{p}_{\theta}(x)$ and we judge the quality of the model by visual inspection. The left, middle, and right columns of Fig. 6 show three realizations from $\widehat{p}_{\theta}(x)$ fitted to the training image in Fig. 2(a), (b) and (c), respectively. It seems like the fitted models for training images (a) and (b) reproduce the main features of the corresponding training images. The three realizations from the model fitted to training image (a) are quite difficult to distinguish from the training image, except that the realizations contain a much larger number of small black and white objects. White objects within black ones occur only a very small number of times in the training image. Note that the irregular nature of the objects means that this model is the most difficult to judge by visual inspection. In the realizations from the fitted model to training image (b), it is fair to say that the objects are slightly less regular versions of the objects in the training image. However, in applications such as reservoir characterization, this training image is not necessarily a more realistic description of the real phenomenon than the fitted model. The realizations from the model fitted to training image (c) contain channels similar to those in the training image, but most of them

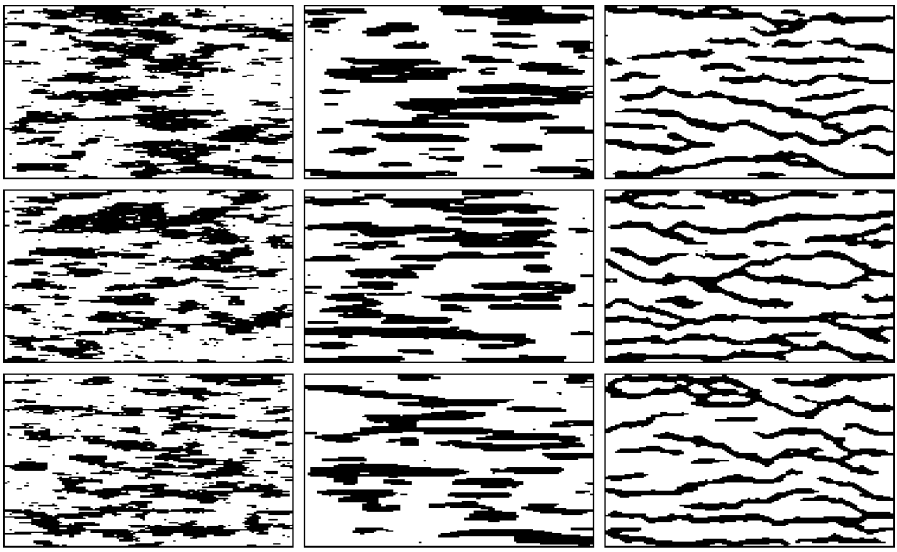

Fig. 6 For each training image, three realizations from the fitted $\widehat{p}_\theta(x)$

do not extend all the way across the image. In petroleum reservoirs, the continuity of the structures is very important and, therefore, this model would fail at modeling a key feature of the reservoir.

To get a better impression, the descriptive statistics are also studied. The statistics considered are, for each of the two colors, area fraction (Fraction), average ratio of area and circumference of an object (Area/Circ), average area of an object (Area), number of objects in an image (# Objects), average extension of an object in x- (x-ext) and y-directions (y-ext), and average circumference of an object (Circ). We compute the statistics from 100 realizations from $\widehat{p}_\theta(x)$ and standardize these by dividing by the corresponding value from the training image. Box and Whisker plots of the results are shown in Fig. 7. The value of the training image, which is one, is indicated by a vertical solid line. Plots corresponding to black and white objects are found in the left and right columns, respectively, whereas the upper, middle and lower rows in the figure correspond to the training images in Fig. 2(a), (b), and (c), respectively. The standardized statistics for the fitted model for training image (a) all have a median value close to one, but only four of the boxes cover this value. This discrepancy is mostly due to the fact that there are too many small objects of either color. If one omits objects smaller than four nodes (figures for this not shown) all of the boxes cover one. In the fitted model for training image (b), either the boxes or whiskers cover one for all the statistics considered. The biggest disparity is again in the number of white objects, but note that the variance of this statistic is large. For the fitted model to training image (c), most of the statistics of the training image are not reproduced by $\widehat{p}_\theta(x)$. Surprisingly, this suggests that training image (b) is the image that fits best with our fitted model. In petroleum applications, the ratio of different classes is a very important statistic. In all the fitted models, the ratio of black and white is well reproduced. To also assess the performance for $p^\star_\theta(x)$, we repeat the same simulation

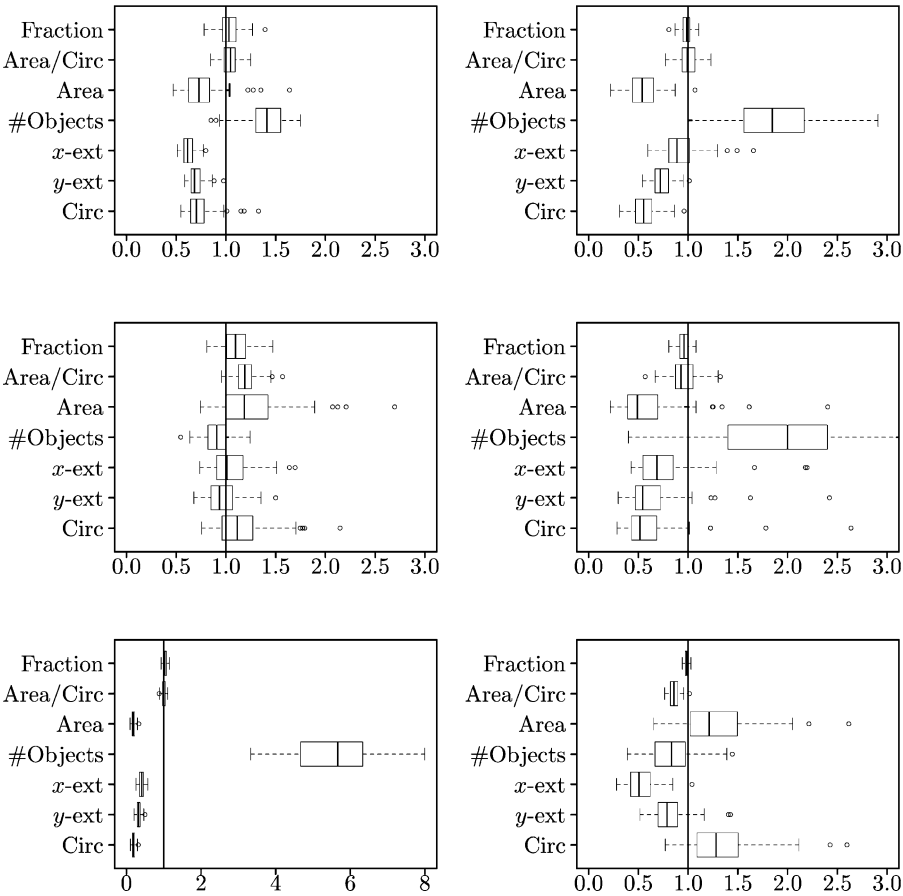

Fig. 7 Box and Whisker plots of the standardized descriptive statistics for unconditional simulations from the fitted model $\widehat{p}_\theta(x)$

exercise for this POMM approximation. The results of this are found in Figs. 12 and 13 in Appendix. These figures are organized similar to Figs. 6 and 7, respectively. The Box and Whisker plots are very similar for $p^\star_\theta(x)$ and $\widehat{p}_\theta(x)$, but many of the boxes for $p^\star_\theta(x)$ have moved slightly away from one relative to the situation for $\widehat{p}_\theta(x)$. For instance, the number of white objects in the fitted models for training image (b) now has a median that is larger than two. There are also some statistics which have moved closer to one, for example, the number of black objects in the fitted model to training image (b).

It is quite difficult intuitively to understand the nature of the fitted POMM, $p^\star_\theta(x)$. Therefore, Fig. 8 also shows the resulting lower adjacent neighborhood N_{ij} of one node $(i, j) \in S_4$ well away from the borders of the lattice. In the figure, node (i, j) is shown in black with a circle surrounding it. Nodes in N_{ij} are also shown in black, whereas the nodes in $\{\rho^{-1}(l), l = \rho(i, j) + 1, \ldots, mn\} \setminus N_{ij}$ are shown in gray. The upper, middle, and lower plots correspond to the training image in Fig. 2(a), (b), and (c), respectively. We see that below node (i, j), N_{ij} contains a rectangular region of

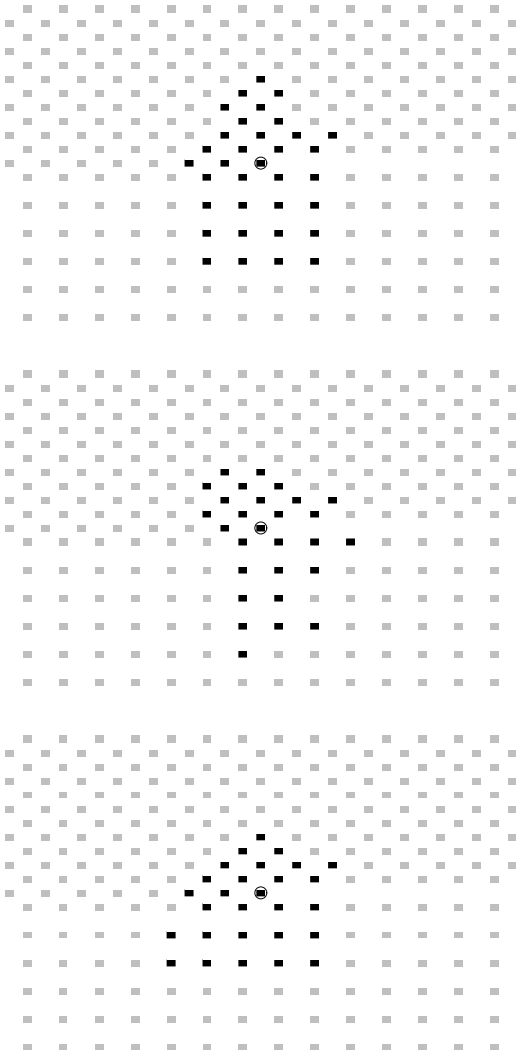

Fig. 8 For a node (i, j) well away from the lattice borders, lower adjacent neighborhoods in the fitted POMMs, $p^{\star}_{\theta}(x)$

nodes in $S_1 \cup S_2 \cup S_3$. Above (i, j), N_{ij} also contains nodes in S_4 and, therefore, it is reasonable that fewer nodes from $S_1 \cup S_2 \cup S_3$ need to be included and a triangle is formed. One should note that N_{ij} of the fitted $p^{\star}_{\theta}(x)$ for training image (c) extends further horizontally than for the other fitted models. However, the differences between N_{ij} obtained for the three training images are reasonably small.

Finally, conditional simulation is examined when the data z is exact observations of two columns in the training image. Specifically, the POMM approximation to $\widetilde{p}_{\theta}(x|z)$ is studied, obtained as described in the last paragraph of Sect. 5.3. Figure 9 shows three realizations from this distribution for each of the three training images, and Fig. 10 shows corresponding estimated marginal probabilities based on 1,000 conditional realizations. The images to the left, in the middle and to the right in these figures correspond to the training images in Fig. 2(a), (b), and (c),

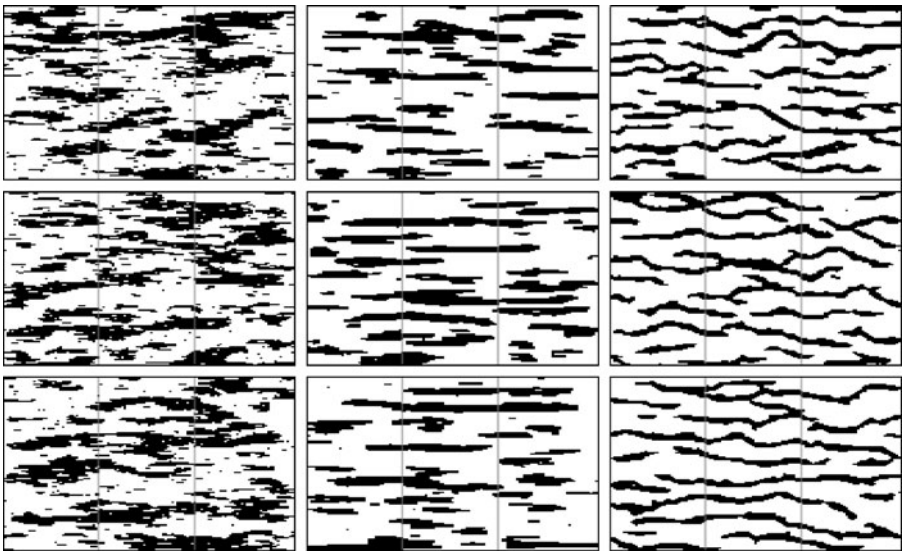

Fig. 9 For each training image, three realizations from the POMM approximation of the fitted $\widetilde{p}_\theta(x|z)$ when z is exact observations of two columns in the training image

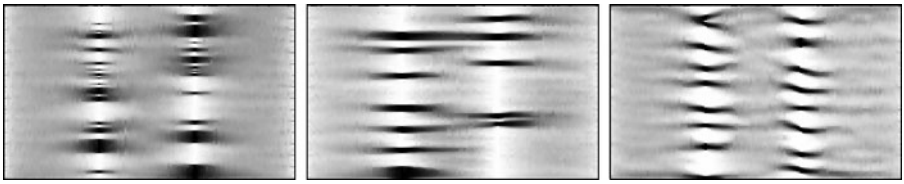

Fig. 10 Estimated marginal probabilities for the POMM approximation of the fitted $\widetilde{p}_\theta(x|z)$ when z is exact observations of two columns in the training image

respectively. The positions of the two observed columns are marked with vertical lines. When simulating from an approximate distribution conditioned to exact data, as is done here, the data may stand out from the simulated data if the approximation is not good enough. One should note that it is impossible to observe such an effect in these realizations. As the conditional realizations are conditioned to data from two vertical wells taken from the training image it should be expected that these are more similar to the training image than the unconditional realizations in Fig. 6. Box and Whisker plots for the conditional distributions are given in Fig. 11, which is organized similar to Fig. 7 described above. We see that in the Box and Whisker plots for the conditional distribution the boxes have moved slightly for all of the models. Some have moved further away from one and some have moved closer to one. In particular, the most significant difference is that the number of white objects in the fitted model to training image (b) has moved much closer to one and that the average extensions and average circumference have increased. This all indicates that there are fewer small white objects within the black

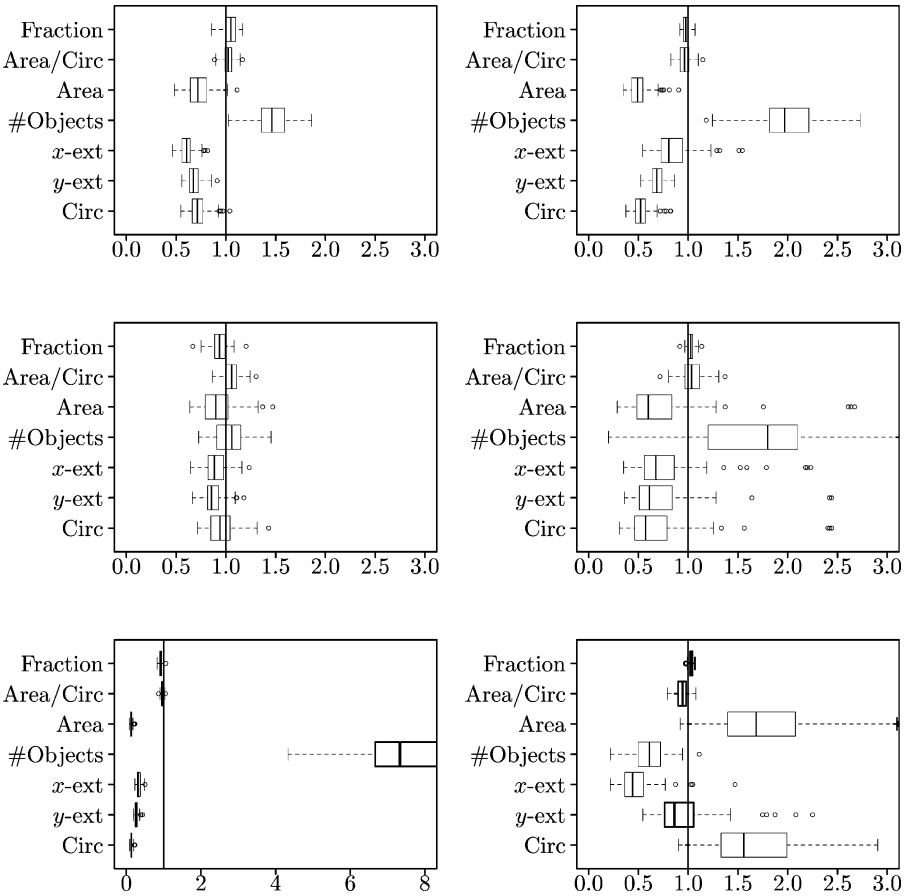

Fig. 11 Box and Whisker plots of the standardized descriptive statistics for the conditional simulation when conditioning on two columns

ones. For the other models, the differences are too small to justify solid conclusions.

An example with observations in every tenth column has also been run. This puts many restrictions on $p_\theta(x|z)$, and could, as discussed above, potentially be problematic for the POMM approximation to $\widetilde{p}_\theta(x|z)$. The results are found in Figs. 14 and 15 in Appendix. These figures are organized similar to Figs. 9 and 10, respectively. Again, it can be observed that the data does not stand out in the realizations. Now it is also apparent that the introduction of more data has resulted in realizations which are much more similar to the training images. The behavior of the approximation is thereby the same as what we would expect from $p_\theta(x|z)$. However, the artifact of the estimated model for the rightmost training image is again clearly seen. The widths of the boxes in the corresponding Box and Whisker plots shown in Fig. 16 in Appendix have decreased, which is also to be expected when the amount of data is increased.

7 Conclusions

In this paper, a new procedure for fitting an MRF to a given binary training image is proposed. The model is defined by a multi-grid approach, which means that the lattice is split into a series of sub-lattices. On each sub-lattice an MRF conditioned on the values in the previous sub-lattices is fitted. The examples should demonstrate the flexibility of the proposed approach, but also its limitations. The MRF multi-grid formulation has an unknown normalizing constant in each lattice level. This complicates the use of the MRF multi-grid model. This problem is resolved by approximating these unknown normalizing constants, ending up with a POMM approximation to the specified MRF. We also define a POMM approximation to the corresponding conditional distribution.

In geostatistics, the most popular strategy by far for constructing a prior model from a training image is to adopt a multi-point statistics algorithm. As discussed in the introduction this modeling strategy has important shortcomings whenever simulation conditioned to observed data is of interest. In this article, an alternative prior modeling strategy is proposed, where conditional simulation from a corresponding conditional distribution is made possible by adopting a Markov chain Monte Carlo algorithm. Faster approximate algorithms for conditional simulation are also given. The focus of this article is on the methodological aspects of the modeling strategy. Therefore, we have limited the attention to two-dimensional binary training images. The procedure is easy to extend to training images with more than two values. The computational cost increases rapidly with the number of values, however, so details in the implementation are crucial in making the algorithms computationally feasible for more than two values. A direct generalization of the approach to three dimensions is possible, but we are of the belief that a better alternative is to model this as a Markov chain of two-dimensional models, where the approach introduced in this paper can be adopted for each of the two-dimensional models.

A major concern in the construction of a prior model from a training image and corresponding conditional simulation is the computational complexity of the algorithms. The most computer intensive part of our algorithms by far is the parameter estimation procedure discussed in Sect. 4.3, which is done via an iterative numerical optimization algorithm. This estimation procedure is clearly computationally more costly than the estimation procedure used in multi-point statistics, where estimation simply consists of counting the observed template configurations. One should note, however, that the clique sizes used in the examples in Sect. 6 are smaller than the template size typically needed to obtain satisfactory results in the multi-point statistics algorithms. When the parameters have been estimated, both our algorithm to simulate from the prior model and the approximate algorithm to simulate from a corresponding conditional distribution are sequential algorithms. The computational complexity for these algorithms are thereby comparable to the corresponding algorithms for the multi-point statistics models.

As opposed to the Markov mesh model defined in Stien and Kolbjørnsen (2011), our MRF model formulation does not include any directionality. The node ordering in our POMM approximation might potentially induce directionality in our final POMM, but we are not able to find any significant such effect in any of our

examples. Note that the multi-point statistics models (Journel and Zhang 2006; Strebelle 2002) avoid this directionality problem by simulating the nodes in a random order. When it comes to conditional simulation, our POMM is comparable to the Markov mesh model of Stien and Kolbjørnsen (2011). As both formulations have explicit formulas for the fitted distributions, conditional realizations can be generated by adopting the Metropolis–Hastings procedure. Alternatively, as we detail for our POMM in Sect. 5.3, realizations from an approximation to the conditional distribution can be generated by feeding the conditional distribution into the approximation procedure of Tjelmeland and Austad (2012). As discussed in Sect. 1, conditional simulation from the multi-point statistics models is more complicated.

Acknowledgements We thank the sponsors of the Uncertainty in Reservoir Evaluation (URE) project at the Norwegian University of Science and Technology (NTNU). We also thank two anonymous journal reviewers for comments to an earlier version of this paper.

Appendix: Additional Plots

In this Appendix, the results of the simulation from $p^\star_\theta(x)$ and from the conditional distribution $\tilde{p}_\theta(x|z)$ are presented, when z is the observations of 11 vertical traces in the training image. In Fig. 12, we show realizations from the fitted $p^\star_\theta(x)$, and in Fig. 13 we show Box and Whisker plots of the corresponding standardized descriptive statistics. In Fig. 14, we show realizations from the POMM approximation of the fitted $\tilde{p}_\theta(x|z)$, when z is 11 vertical traces taken from the training image, and corresponding marginal probabilities are shown in Fig. 15. Box and Whisker plots of the corresponding standardized descriptive statistics are shown in Fig. 16.

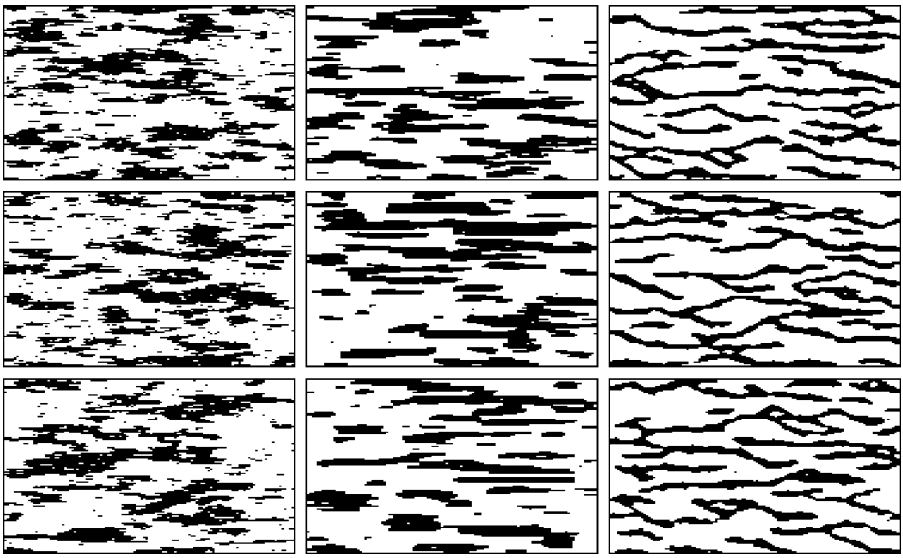

Fig. 12 For each training image, three realizations from the fitted $p^\star_\theta(x)$

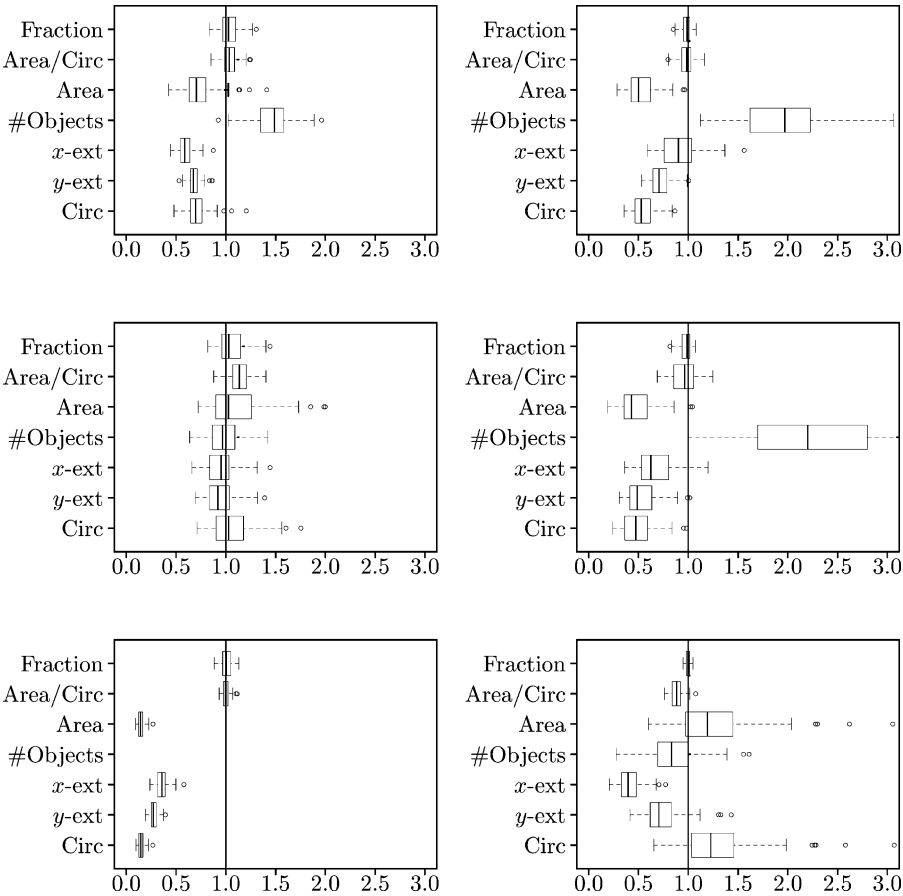

Fig. 13 Box and Whisker plots of the standardized descriptive statistics for unconditional simulations from the fitted model $p^{\star}_{\theta}(x)$

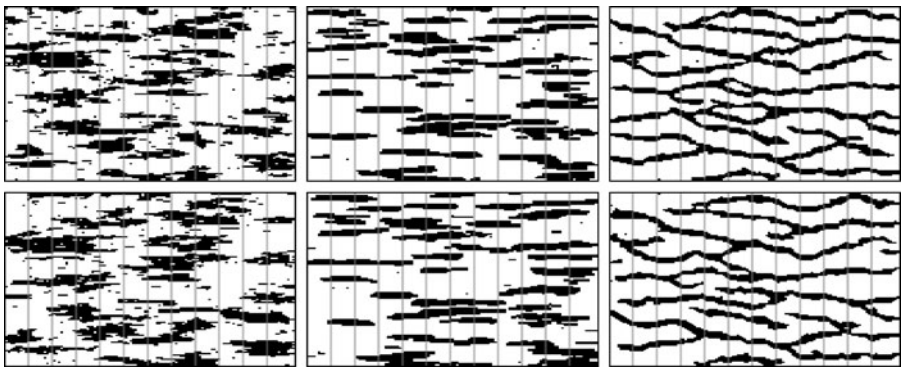

Fig. 14 For each training image, three realizations from the POMM approximation of the fitted $\widetilde{p}_{\theta}(x|z)$ when z contains exact observations of eleven columns in the training image

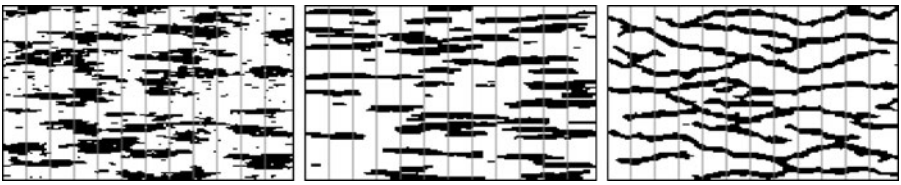

Fig. 14 (*Continued*)

Fig. 15 Estimated marginal probabilities for the POMM approximation of the fitted $\widetilde{p}_\theta(x|z)$ when z is exact observations of eleven columns in the training image

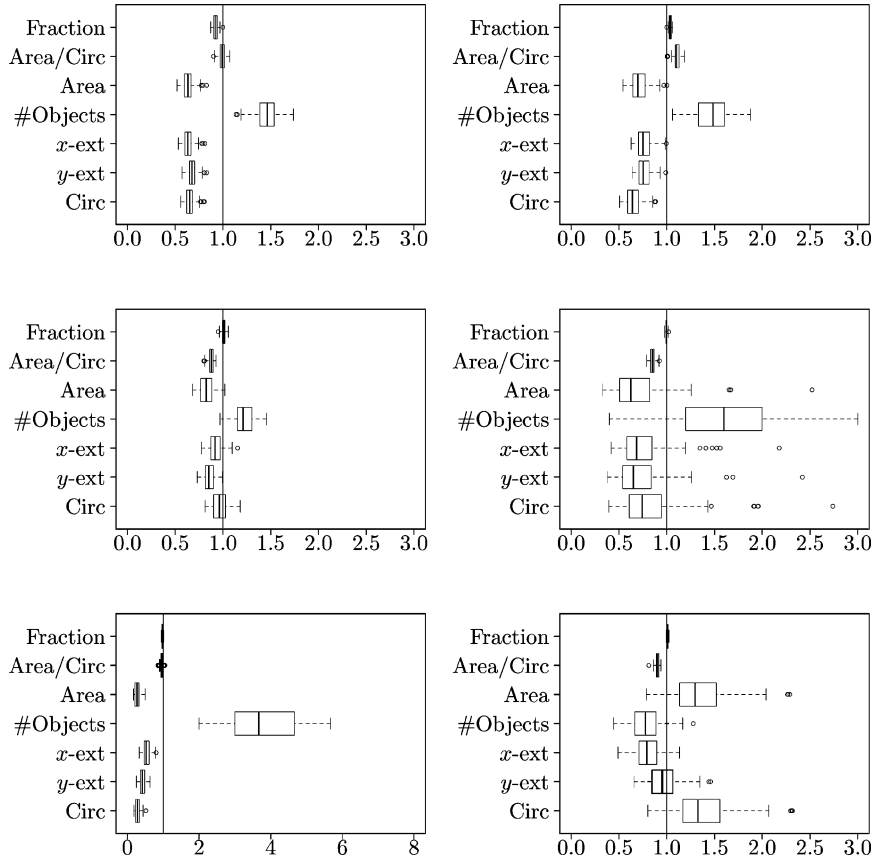

Fig. 16 Box and Whisker plots of the standardized descriptive statistics for the conditional simulation when conditioning on eleven columns

References

Bartolucci F, Besag J (2002) A recursive algorithm for Markov random fields. Biometrika 89:724–730

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 36:192–225

Chatterjee S, Dimitrakopoulos R, Mustapha H (2012) Dimensional reduction of pattern-based simulation using wavelet analysis. Math Geosci 44:343–374

Clifford P (1990) Markov random fields in statistics. In: Grimmett GR, Welsh DJA (eds) Disorder in physical systems. Oxford University Press, London, pp 19–31

Cressie NAC (1993) Statistics for spatial data, 2nd edn. Wiley, New York

Cressie N, Davidson J (1998) Image analysis with partially ordered Markov models. Comput Stat Data Anal 29:1–26

Descombes X, Mangin J, Pechersky E, Sigelle M (1995) Fine structures preserving model for image processing. In: Proc. 9th SCIA 95, Uppsala, Sweden, pp 349–356

Eidsvik J, Avseth P, Omre H, Mukerji T, Mavko G (2004) Stochastic reservoir characterization using prestack seismic data. Geophysics 69:978–993

Friel N, Rue H (2007) Recursive computing and simulation-free inference for general factorizable models. Biometrika 94:661–672

Friel N, Pettitt AN, Reeves R, Wit E (2009) Bayesian inference in hidden Markov random fields for binary data defined on large lattices. J Comput Graph Stat 18:243–261

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6:721–741

Geyer CJ, Thompson EA (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. J Am Stat Assoc 90:909–920

Gonzalez EF, Mukerji T, Mavko G (2008) Seismic inversion combining rock physics and multiple-point geostatistics. Geophysics 73:R11–R21

Hurn M, Husby O, Rue H (2003) A tutorial on image analysis. In: Møller J (ed) Spatial statistics and computational methods. Lecture notes in statistics, vol 173. Springer, New York, pp 87–139

Journel J, Zhang T (2006) The necessity of a multiple-point prior model. Math Geol 38:591–610

Kindermann R, Snell JL (1980) Markov random fields and their applications. Am Math Soc, Providence

Künsch HR (2001) State space and hidden Markov models. In: Barndorff-Nielsen OE, Cox DR, Klüppelberg C (eds) Complex stochastic systems. Chapman & Hall/CRC, London

Li SZ (2009) Markov random field modeling in image analysis, 3rd edn. Springer, London

Pettitt AN, Friel N, Reeves R (2003) Efficient calculation of the normalising constant of the autologistic and related models on the cylinder and lattice. J R Stat Soc B 65:235–247

Scott AL (2002) Bayesian methods for hidden Markov models: recursive computing in the 21st century. J Am Stat Assoc 97:337–351

Stien M, Kolbjørnsen O (2011) Facies modeling using a Markov mesh model specification. Math Geosci 43:611–624

Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Math Geol 34:1–21

Tjelmeland H (1996) Stochastic models in reservoir characterization and Markov random fields for compact objects. PhD thesis, Norwegian University of Science and Technology. Thesis number 44:1996

Tjelmeland H, Austad H (2012) Exact and approximate recursive calculations for binary Markov random fields defined on graphs. J Comput Graph Stat 21:758–780

Tjelmeland H, Besag J (1998) Markov random fields with higher order interactions. Scand J Stat 25:415–433

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

Winkler G (2003) Image analysis, random fields and Markov chain Monte Carlo methods. Springer, London

Zhang T, Pedersen SI, Knudby C, McCormick D (2012) Memory-efficient categorical multi-point statistics algorithms based on compact search trees. Math Geosci 44:863–879