NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Construction of binary multi-grid Markov random field prior models from training images

by

Håkon Toftaker and Håkon Tjelmeland

PREPRINT
STATISTICS NO. 1/2012

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL `http://www.math.ntnu.no/preprint/statistics/2012/S1-2012.pdf`

Håkon Toftaker has homepage: `http://www.math.ntnu.no/∼toftaker`
E-mail: `toftaker@stat.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology,
N-7491 Trondheim, Norway.

Construction of binary multi-grid Markov random field prior models from training images

Håkon Toftaker and Håkon Tjelmeland
Department of Mathematical Sciences
Norwegian University of Science and Technology

Abstract

Bayesian modelling requires prior and likelihood models to be specified. In reservoir characterization it is common practice to estimate the prior from a training image. We consider 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. We illustrate the proposed methods using three frequently used binary training images.

Key words: Markov random field, forward-backward algorithm, multi-grid, facies modelling, maximum likelihood

1 Introduction

Discrete Markov random fields (MRFs) are frequently used as prior models in spatial statistics. Following the seminal paper of Geman and Geman (1984), MRFs are nowadays routinely used as prior distributions in image analysis, see for example Hurn et al. (2003), Winkler (2003) and Li (2009) and references therein. The MRF is typically used only as a token prior. In the binary case, for example, the autologistic model (Besag, 1974) is frequently adopted as a prior despite that this model does not reflect 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, see for example Descombes et al. (1995) and Tjelmeland and Besag (1998) where a Markov chain Monte Carlo maximum likelihood procedure (Geyer and Thompson, 1995) is used in attempts to construct more realistic MRF priors.

Spatial models for discrete variables are also important when modelling the spatial distribution of rock types in petroleum reservoirs, see for example Strebelle (2002), Eidsvik et al. (2004), Gonzalez et al. (2008), Ulvmoen and Omre (2010) and references therein. 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 have 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, see examples in Strebelle (2002) and Journel and Zhang (2006). 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 $p(x_1, \ldots, x_n) \propto \sum_t \left[p(x_{t(1)}) p(x_{t(2)} | x_{t(1)}) \cdot \ldots \cdot p(x_{t(n)} | x_{t(1)}, \ldots, x_{t(n-1)}) \right]$, where the sum is over all possible permutations t, and each of the factors $p(x_{t(i)} | x_{t(1)}, \ldots, x_{t(i-1)})$ are 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 ordered Markov model (POMM), see 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. The problem is that also the posterior distribution becomes a mixture distribution, and the component weights are not analytically available. The multi-point statistics solution to this problem is to modify the posterior expression to get a distribution from which it is easy 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 standard solution is to put zero weights to all mixture components that does not correspond to permutations that start with the observed nodes, and equal weights for the remaining components. This strategy is often quite successful, at least as far as it possible to evaluate from visual inspections of generated realizations. When more complicated likelihoods are of interest, it becomes more difficult to prescribe how to modify the posterior distribution to get an approximation it is possible to sample from. As a result, in these cases one typically sees clear visual artifacts when inspecting the generated multi-point statistics realizations.

An alternative to the multiple-point statistics prior is to pick one of the mixture components in the multiple-points statistics prior and adopt this as a prior. This implies that the prior becomes a POMM and realizations from a corresponding posterior distribution can easily be generated by Markov chain Monte Carlo (Gamerman and Lopes, 2006; Brooks et al., 2011). Unfortunately, the experience is that this often generates clear visual artifacts in the realizations from both the prior and the posterior. When not used within a mixture model the conditional independence structure adopted in the multi-point statistics models seem inconsistent with the typical training images used. An exception from this rule, however, seems to be the POMM constructed in Stien and Kolbjørnsen (2011).

In the present paper our objective is to define a prior distribution that both is 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 we limit the attention to binary fields, but our strategy can be generalized to a situation with more than two possible values. We adopt a POMM as prior distribution, but specify the POMM 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 we adopt the strategy of Austad and Tjelmeland (2011). 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 we found it necessary to adopt a multi-grid approach. The resulting model is thereby a product of POMMs, which is again a POMM.

The paper is organized as follows. In Sections 2 and 3 we review the definition and some basic properties of POMMs and MRFs, respectively. In Section 4 we first discuss how to find a POMM approximation to a given MRF. Thereafter we describe 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. In Section 5 the multi-grid MRF is defined and the POMM approximation is adapted to this situation. In this section we also discuss how to define a POMM approximation to a conditional multi-grid model. Finally, Section 6 presents simulation examples and evaluations of the proposed procedures, and Section 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 we only introduce the basic concepts necessary to understand our POMM approximation to binary MRFs.

Assume 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 to be so that there exist 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 will typically not 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 we also use the standard notations $x_A = (x_{ij}, (i,j) \in A)$ for $A \subseteq S$, $x_{-A} = x_{S \setminus A}$ and $x_{-(i,j)} = x_{-\{(i,j)\}}$. Letting θ denote a vector of model parameters, the joint distribution of the POMM is

$$p_\theta(x) = \prod_{(i,j)\in S} p_\theta(x_{(i,j)}|x_{N_{ij}}). \tag{1}$$

To evaluate the likelihood $p_\theta(x)$ for a given image x is straight forward from (1). 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 for example be found in Besag (1974), Kindermann and Snell (1980) and Cressie (1993). Here we give an introduction to binary MRFs defined on a rectangular lattice.

Assume as in the previous section that we have an $n \times m$ rectangular lattice and denote the set of lattice nodes by $S = \{(i,j), i = 1, \dots, n, j = 1, \dots, m\}$. To each node $(i,j) \in S$ we associate a set of neighbor nodes $\partial(i,j) \subseteq S \setminus \{(i,j)\}$, where we require the neighborhood 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 we say (i,j) and (r,s) are 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 above we associate a binary variable $x_{ij} \in \{0,1\}$ to each $(i,j) \in S$ and 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 fulfil the Markov assumption

$$p_\theta(x_{ij}|x_{-(i,j)}) = p_\theta(x_{ij}|x_{\partial(i,j)}) \tag{2}$$

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

$$p_\theta(x) = c(\theta)\exp\{-U_\theta(x)\}, \tag{3}$$

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 (2) the most general form the energy function can take is

$$U_\theta(x) = \sum_{C\in\mathcal{C}} V_C(x_C, \theta), \tag{4}$$

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

$$c(\theta) = \left[\sum_{x\in\Omega} \exp\{-U_\theta(x)\}\right]^{-1}, \tag{5}$$

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. See, however, Geyer and Thompson (1995) for how to approximate MLE from a set of MCMC samples.

4 Forward-backward algorithm and the POMM approximation

In this section we first describe an exact forward-backward algorithm (Künsch, 2001; Scott, 2002; Pettitt et al., 2003) for a binary MRF $p_\theta(x)$. As this exact procedure is practical only for MRFs defined on small lattices and with small neighborhoods, we next review the corresponding approximate algorithm of Austad and Tjelmeland (2011) which produces a POMM approximation $\widetilde{p}_\theta(x)$ to the MRF. Finally we discuss 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.

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

$$p_\theta(x_{\{\rho^{-1}(l),l=k,\ldots,mn\}}) \text{ for } k = 2,\ldots,mn \tag{6}$$

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 (6). 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 (6) can be decomposed into a product

$$p_\theta(x_{\{\rho^{-1}(l),l=k,\ldots,mn\}}) = g_\theta(x_{\rho^{-1}(k)}, x_{A_k}) h_\theta(x_{\rho^{-1}(l),l=k+1,\ldots,mn\}}). \tag{7}$$

When summing out $x_{\rho^{-1}(k)}$ from (6) 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 (6) the conditional distribution

$$p_\theta(x_{\rho^{-1}(k)}|x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}}) = \frac{p_\theta(x_{\{\rho^{-1}(l),l=k,\ldots,mn\}})}{p_\theta(x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}})} \propto g_\theta(x_{\rho^{-1}(k)}, x_{A_k}) \tag{8}$$

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

$$p_\theta(x) = p_\theta(x_{\rho^{-1}(mn)}) \prod_{k=1}^{mn-1} p_\theta(x_{\rho^{-1}(k)}|x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}}). \tag{9}$$

It should be noted that (9) 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. Austad and Tjelmeland (2011) 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 (6). 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 (7). 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

$$\widetilde{p}_\theta(x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}}) = \sum_{x_{\rho^{-1}(k)}=0}^{1} \widetilde{p}_\theta(x_{\rho^{-1}(l),l=k,\ldots,mn\}}). \tag{10}$$

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, see Austad and Tjelmeland (2011) for details. Thus, we have a decomposition of $\widehat{p}_\theta(x_{\rho^{-1}(l),l=k,\ldots,mn})$ corresponding to (7),

$$\widehat{p}_\theta(x_{\rho^{-1}(l),l=k,\ldots,mn}) = \widehat{g}_\theta(x_{\rho^{-1}(k)}, x_{\widehat{A}_k})\widehat{h}_\theta(x_{\rho^{-1}(l),l=k+1,\ldots,mn}). \tag{11}$$

Thereafter $\widetilde{p}_\theta(x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}})$ is defined as in (10), 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 (8) and (9). Thus,

$$\widetilde{p}_\theta(x) = \widetilde{p}_\theta(x_{\rho^{-1}(mn)}) \prod_{k=1}^{mn-1} \widetilde{p}_\theta(x_{\rho^{-1}(k)} | x_{\rho^{-1}(l),l=k+1,\ldots,mn\}}), \tag{12}$$

where

$$\widetilde{p}_\theta(x_{\rho^{-1}(k)} | x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}}) = \frac{\widetilde{p}_\theta(x_{\{\rho^{-1}(l),l=k,\ldots,mn\}})}{\widetilde{p}_\theta(x_{\{\rho^{-1}(l),l=k+1,\ldots,mn\}})} \propto \widetilde{p}_\theta(x_{\{\rho^{-1}(l),l=k,\ldots,mn\}}) \tag{13}$$

for $k = 1, \ldots, mn-1$. As for the exact decomposition in (9), $\widetilde{p}_\theta(x)$ 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 (12) 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 we consider 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 $p_\theta(x)$.

4.3 Maximum likelihood estimation by importance sampling

To cope with the computationally intractable normalizing constant $c(\theta)$ in $p_\theta(x)$ we use importance sampling. 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 we can generate independent samples from the POMM approximation $\widetilde{p}_\theta(x)$ 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 (5) we get for a fixed parameter vector θ^0 that

$$\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)} = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \sum_{z\in\Omega} \varphi_\theta(z) = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \sum_{z\in\Omega} \left[\frac{\varphi_\theta(z)}{\widetilde{p}_{\theta^0}(z)} \widetilde{p}_{\theta^0}(z)\right] = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \mathrm{E}\left[\frac{\varphi_\theta(z)}{\widetilde{p}_{\theta^0}(z)}\right], \tag{14}$$

where the expectation is 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

$$\widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)}} = \frac{\widetilde{p}_{\theta^0}(x)}{\varphi_\theta(x)} \cdot \frac{1}{R} \sum_{r=1}^{R} \frac{\varphi_\theta(z^r)}{\widetilde{p}_{\theta^0}(z^r)}, \tag{15}$$

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 (15) available, it is tempting to find an approximate maximum likelihood estimate for θ by numerically minimizing (15) 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 (15) may become very large. The numerical optimization of (15) should therefore be stopped whenever the (estimated) variance of the decrease 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, a new POMM approximation should be constructed and new realizations $z_1, \ldots, z_R$ generated to obtain a new estimate (15) 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

$$\widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)}} - \widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_{\theta'}(x)}} = \frac{1}{R} \sum_{r=1}^{R} \left[\frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)} \frac{\varphi_\theta(z^r)}{\varphi_\theta(x)} - \frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)} \frac{\varphi_{\theta'}(z^r)}{\varphi_{\theta'}(x)}\right] \tag{16}$$

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

$$\widehat{\sigma}^2(\theta,\theta') = \frac{1}{R-1} \sum_{r=1}^{R} \left[\frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)} \frac{\varphi_\theta(z^r)}{\varphi_\theta(x)} - \frac{\widetilde{p}_{\theta^0}(x)}{\widetilde{p}_{\theta^0}(z^r)} \frac{\varphi_{\theta'}(z^r)}{\varphi_{\theta'}(x)} - \left(\widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_\theta(x)}} - \widehat{\frac{\widetilde{p}_{\theta^0}(x)}{p_{\theta'}(x)}}\right)\right]^2. \tag{17}$$

Assuming the estimate in (16) 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).

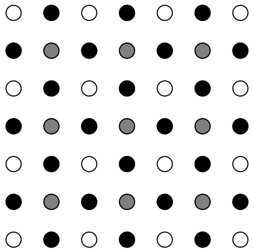


Figure 1: Illustration of the splitting of node set S for a 7×7 lattice into three sub-lattices S_1, S_2 and S_3. The white nodes are in S_1, the gray nodes in S_2 and the black nodes in S_3.

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 Section 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.

5 Multi-grid MRF and POMM approximation

In this section we first define a general multi-grid MRF and adapt the POMM approximation defined in Section 4.2 to this situation. Thereafter we discuss how to construct a POMM approximation to a corresponding conditional distribution, before we adapt the parameter estimation procedure discussed in Section 4.3 to the multi-grid MRF situation.

5.1 Multi-grid MRF

In the multi-grid approach we split the nodes in our rectangular lattice S 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 is placed between the nodes in S_1 as illustrated in Figure 1. The nodes in the sub-lattice S_3 is 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 does 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 is 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.

We specify the joint distribution for $x = (x_{(i,j)}, (i,j) \in S)$ 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 $S_{1:t-1} = S_1 \cup \ldots \cup S_{t-1}$. We adopt a separate parameter vector for each of these T distributions, denoted by $\theta_1, \ldots, \theta_T$, respectively. Thereby we have

$$p_\theta(x) = p_{\theta_1}(x_{S_1}) \prod_{t=2}^{T} p_{\theta_t}(x_{S_t} | x_{S_{1:t-1}}), \tag{18}$$

where $\theta = (\theta_1, \ldots, \theta_T)$. For the marginal distribution $p_{\theta_1}(x_{S_1})$ we adopt an MRF exactly as discussed in Section 3, whereas for each of $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ we adopt an MRF where the conditioning variables are included as covariates. 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 Section 6.1 we specify the neighborhood structure and parametric form of the potential functions that we use in the simulation examples. Now we focus 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 we can adopt the approximation scheme discussed in Section 4 to each of $p_{\theta_1}(x_{S_1})$ 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 Section 3, so the approximation scheme defined in Section 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 we discuss details of the two alternatives in turn.

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 $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$ and thereafter the POMM approximation defined in Section 4 can be directly applied. We denote this approximation 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 procedure 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. We next discuss the second approximation scheme for $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$, which can be used also for conditional simulation.

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 (3) we have

$$p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = c(\theta_t, x_{S_{1:t-1}}) \exp\left\{-U_{\theta_t}(x_{S_{1:t}})\right\}, \tag{19}$$

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 $x_{S_{1:t-1}}$. We now consider the following distribution for $x_{S_{1:t}}$,

$$f_{\theta_t}(x_{S_{1:t}}) \propto \exp\left\{-U_{\theta_t}(x_{S_{1:t}})\right\}, \tag{20}$$

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

$$f_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = p_{\theta}(x_{S_t}|x_{S_{1:t-1}}) \quad \text{and} \quad f_{\theta_t}(x_{S_{1:t-1}}) = \frac{1}{c(\theta_t, x_{S_{1:t-1}})}, \tag{21}$$

respectively. As $f_{\theta_t}(x_{S_{1:t}})$ is an MRF, the approximation scheme defined in Section 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, we get approximations to the two distributions in (21). As detailed in Section 4.2 the approximation to the conditional distribution is given as a product of univariate conditional distributions,

$$\widetilde{p}_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}) = \prod_{k=1}^{|S_t|} \widetilde{f}_{\theta_t}(x_{\rho^{-1}(k)}|x_{\rho^{-1}(l)}, l = k+1, \ldots, |S_{1:t}|), \tag{22}$$

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

$$\widetilde{f}_{\theta_t}(x_{S_{1:t-1}}) = \frac{1}{\widetilde{c}(\theta_t, x_{S_{1:t-1}})}, \tag{23}$$

has no special form.

Performing the procedure discussed above for each $t = 2, \ldots, T$ and combining the results we obtain two alternative approximations of $p_\theta(x)$. By 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 (23) we get the approximation

$$\widetilde{p}_\theta(x) \propto \exp\left\{-U_{\theta_1}(x_{S_1})\right\} \prod_{t=2}^{T} \widetilde{c}(\theta_t, x_{S_{1:t-1}}) \exp\left\{-U_{\theta_t}(x_{S_{1:t}})\right\}, \tag{24}$$

whereas by combining the approximations in (22) we obtain the approximation

$$p^\star_\theta(x) = \widetilde{p}_{\theta_1}(x_{S_1}) \prod_{t=2}^{T} \widetilde{p}_{\theta_t}(x_{S_t}|x_{S_{1:t-1}}), \tag{25}$$

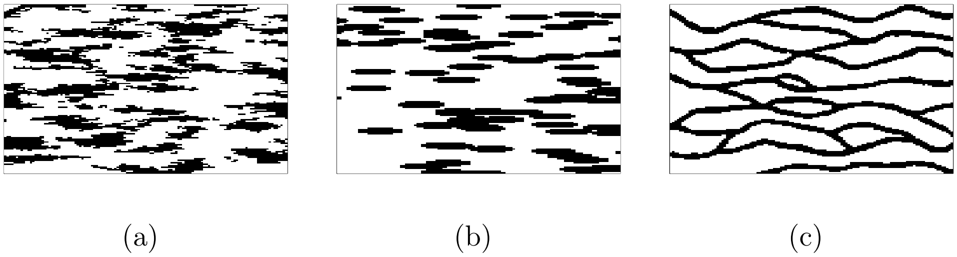
where $\widetilde{p}_{\theta_1}(x_{S_1})$ is the POMM approximation to $p_{\theta_1}(x_{S_1})$. The latter approximation, $p^\star_\theta(x)$, is given as a product of univariate conditional distributions and is then by definition a POMM. Thereby unconditional realizations from $p^\star_\theta(x)$ can 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 we have available an explicit formula for the distribution so 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 Section 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 (23). It is therefore reasonable to consider $\widetilde{p}_\theta(x)$ to be a better approximation than $p^\star_\theta(x)$ to $p_\theta(x)$.

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 (24) and (25), 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 we discuss how to define and simulate from approximations to this conditional distribution.

For each of the two approximations $\widetilde{p}_\theta(x)$ and $p^\star_\theta(x)$ to $p_\theta(x)$ we get 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 neither of these, but for both explicit formulas are available up to a normalizing constant, 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 assume that also $\widetilde{p}_\theta(x|z)$ 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, we recommend to use $\widetilde{p}_\theta(x|z)$ as a 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 Section 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 Section 5.2. In particular this produces an internal consistency in our approximations as it implies that if we have no data, so z is empty, the resulting POMM approximation becomes $p^\star_\theta(x)$.



(a) (b) (c)

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

5.4 Parameter estimation by maximum likelihood

Let x be a given training image to which we want to fit a multi-grid MRF. We adopt the maximum likelihood principle, so to estimate $\theta = (\theta_1, \ldots, \theta_T)$ we need to maximize the likelihood function (18) with respect to θ. As we have specified the multi-grid MRF 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 we can directly apply the optimization procedure specified in Section 4.3. 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 Section 5.2.

6 Examples

To evaluate the performance of our approximation scheme we apply it to the three 121×121 training images shown in Figure 2. For all the 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 we first give details of the parametric form for the multi-grid MRF we are using in our examples. Thereafter we define the node ordering we use for the approximations within each level, and finally we present numerical examples.

6.1 Parametric multi-grid MRF used in the simulation examples

In this section we define the exact neighborhood system and parametric energy functions we are using in the numerical examples presented below. 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 to avoid too many parameters 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 Figure 3(a). The number of neighbors for the nodes on the boundary of the lattice is correspondingly reduced, so that the four corner nodes have only

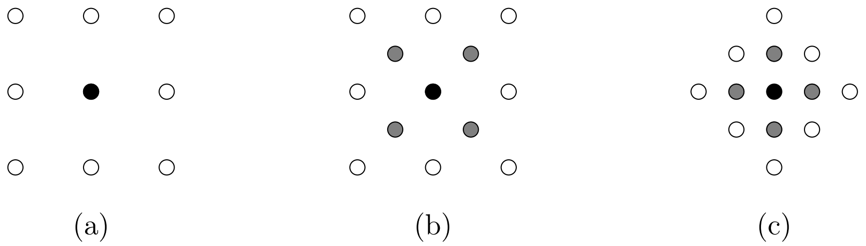


Figure 3: The neighborhoods used in the multi-grid MRF for nodes not located on the boundary of the lattice. The number of neighbors for nodes on the boundary of the lattice is correspondingly reduced. Figure (a) shows the neighborhoods for sub-lattice S_1, (b) the neighborhood for S_t when t is even, and (c) the neighborhood for S_t when $t > 1$ is odd. The white colored nodes are neighbors of the black node. The gray nodes are nodes from $S_{1:t-1}$ which values are used as covariates in the specified MRFs.

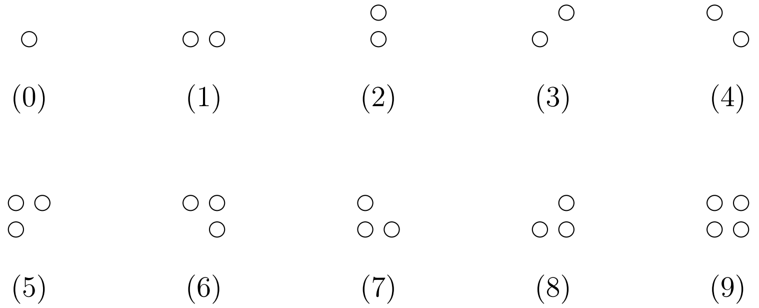


Figure 4: Clique types, numbered from zero to nine, used in the definition of $p_{\theta_1}(x_{S_1})$.

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 that we have to consider, see Figure 4. For each of these clique types we associate a corresponding parameter, $\theta_{1,k}$ for clique type (k), and adopt for clique type (k) the potential function

$$V_C(x_C, \theta_1) = \theta_{1,k} \prod_{(i,j) \in C} x_{ij}, \tag{26}$$

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 we can put one of the ten parameters equal to zero, so for the rest of this paper we fix $\theta_0 = 0$ 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 we again adopt a second order neighborhood model, but in addition we associate to each node $(i,j) \in S_t$ a set B_{ij} that contains the four nodes in $S_{1:t-1}$ that are located closest to (i,j), see the illustration in Figure 3(b). For the energy function we adopt the following parametric form,

$$U_{\theta_t}(x_{S_t}, x_{S_{1:t-1}}) = U^1_{\theta_{t,1:9}}(x_{S_t}) + \sum_{(i,j) \in S_t} U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}}), \tag{27}$$

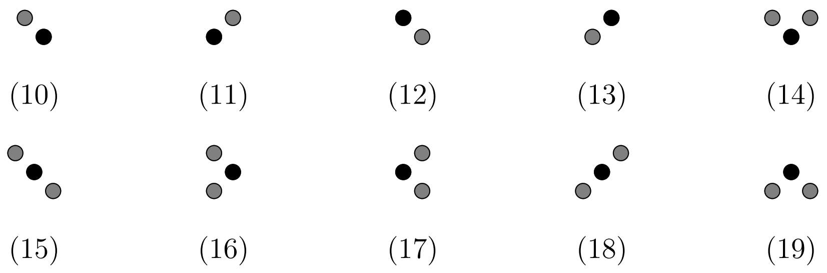


Figure 5: Sets, numbered from ten to nineteen, used to define the potential functions building up the energy function $U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}})$. Node (i, j) and the nodes in B_{ij} are shown in black and gray, respectively.

where the parameter vector θ_t has nineteen elements and is split into $\theta_{t,1:9}$ and $\theta_{t,10:19}$ that contain the first nine and the remaining elements of θ_t, respectively. For $U^1_{\theta_{t,1:9}}(x_{S_t})$ we adopt exactly the same parametric form as for $U_{\theta_1}(x_{S_1})$. For the specification of $U^2_{\theta_{t,10:19}}(x_{ij}, x_{B_{ij}})$ we follow a similar strategy as for the energy function for x_{S_1}, but 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 Figure 5, and the potential function corresponding to node set (k) in that figure is

$$V_C(x_{ij}, x_C, \theta_{t,10:19}) = \theta_{tk}\ x_{ij} \prod_{(r,s)\in C} x_{rs}, \tag{28}$$

where C is the set of gray nodes in the figure. One should note that with (27) the conditioning variables $x_{S_{1:t-1}}$ only affect the first order effects, corresponding to clique type (0) in Figure 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 because this will result in a dramatic increase in the number of parameters.

When $t > 1$ is odd, the nodes in S_t is organized in a lattice that is rotated 45° relative to the lattices making up S_1 and S_t for t even, see Figure 1. We define the energy function for the conditional MRF for x_{S_t} when $t > 1$ is odd exactly as when t is even, except that all cliques and sets B_{ij} are rotated 45° clockwise. Thus, 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 we use the lexicographical ordering of the nodes. 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 we refer to Figure 1. We number the black nodes, i.e. S_3, in this 7×7 lattice 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, we need to specify the values of the parameters γ and R defined in Section 4.3. We start out 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 $R := R/2$. We stop the optimization when $R = 3200$ and the decrease is less than 0.05. In the computations, the POMM approximations depends on the value of κ. In all examples we present we use $\kappa = 12$, which we think is a reasonable trade off between approximation quality and computational complexity. We have also run the same examples for $\kappa = 14$, without detecting significant differences in the fitted models.

6.4 Simulation examples

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

The optimization of the likelihood is done by estimating the likelihood by importance sampling. When estimating θ_t, we need independent samples from the POMM approximation of $p_{\theta_t}(x_{S_t}|x_{S_{1:t-1}})$. 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 needed to compute to reach the MLE was about one hundred.

Each of the training images have some distinct features that put our model fitting procedure to the test. In training image (a) the shape of the black objects are very irregular and we expect to get the best results for this training image. Image (b), on the other hand, has very regular black objects which are mostly convex. We may not be able 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, we generate realizations from $\widehat{p}_{\theta}(x)$ and judge the quality of the model by visual inspection. Figure 6 shows three realizations from $\widehat{p}_{\theta}(x)$ for each of the three training images. 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

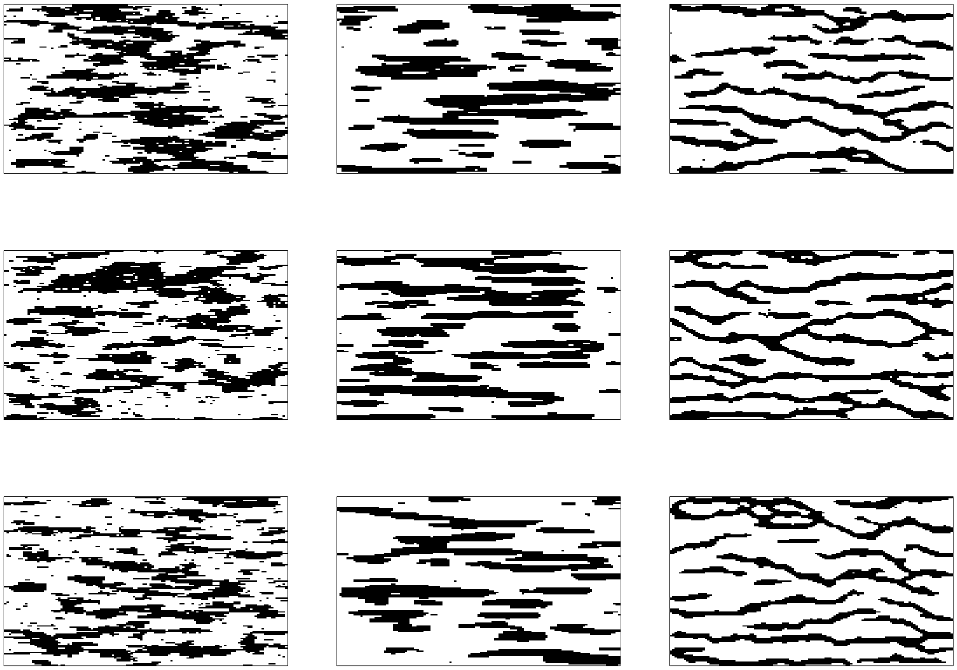


Figure 6: Three realizations from the fitted $\widehat{p}_\theta(x)$, for training images (a) (left column), (b) (middle column) and (c) (right column) in Figure 2.

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 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 modelling a key feature of the reservoir.

To get a better impression we now study descriptive statistics. The statistics we consider are, for each of the two colors, area fraction, average ratio of area and circumference of an object, average area of an object, number of objects, average extension of objects in x- and y-directions, and average circumference of an object. We compute the statistics from 100 realizations from $\widehat{p}_\theta(x)$ and standardize these with respect to the corresponding value from the training image. Box and Whisker plots of the results are shown in Figure 7. 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 we consider. 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 asses the performance also for $p^\star_\theta(x)$ we repeat the same simulation exercise for this POMM approximation. The results of this is found in Figures 11 and 12 in Appendix A. 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, e.g 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 we show in Figure 8 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 around, 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. We see that below node (i,j), N_{ij} contains a rectangular region of 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 we look at conditional simulation, when the data, z, is exact observations of two columns in the training image. Specifically we study the POMM approximation to $\widetilde{p}_\theta(x|z)$, obtained as described in the last paragraph of Section 5.3. Figure 9 shows three realizations from this distribution for each of the three training images. When simulating from an approximate distribution conditioned to exact data, as we do here, the data may 'stand out' from the simulated data if the approximation is too severe. One should note that it is impossible to observe such an effect in our realizations. As the conditional realizations are conditioned

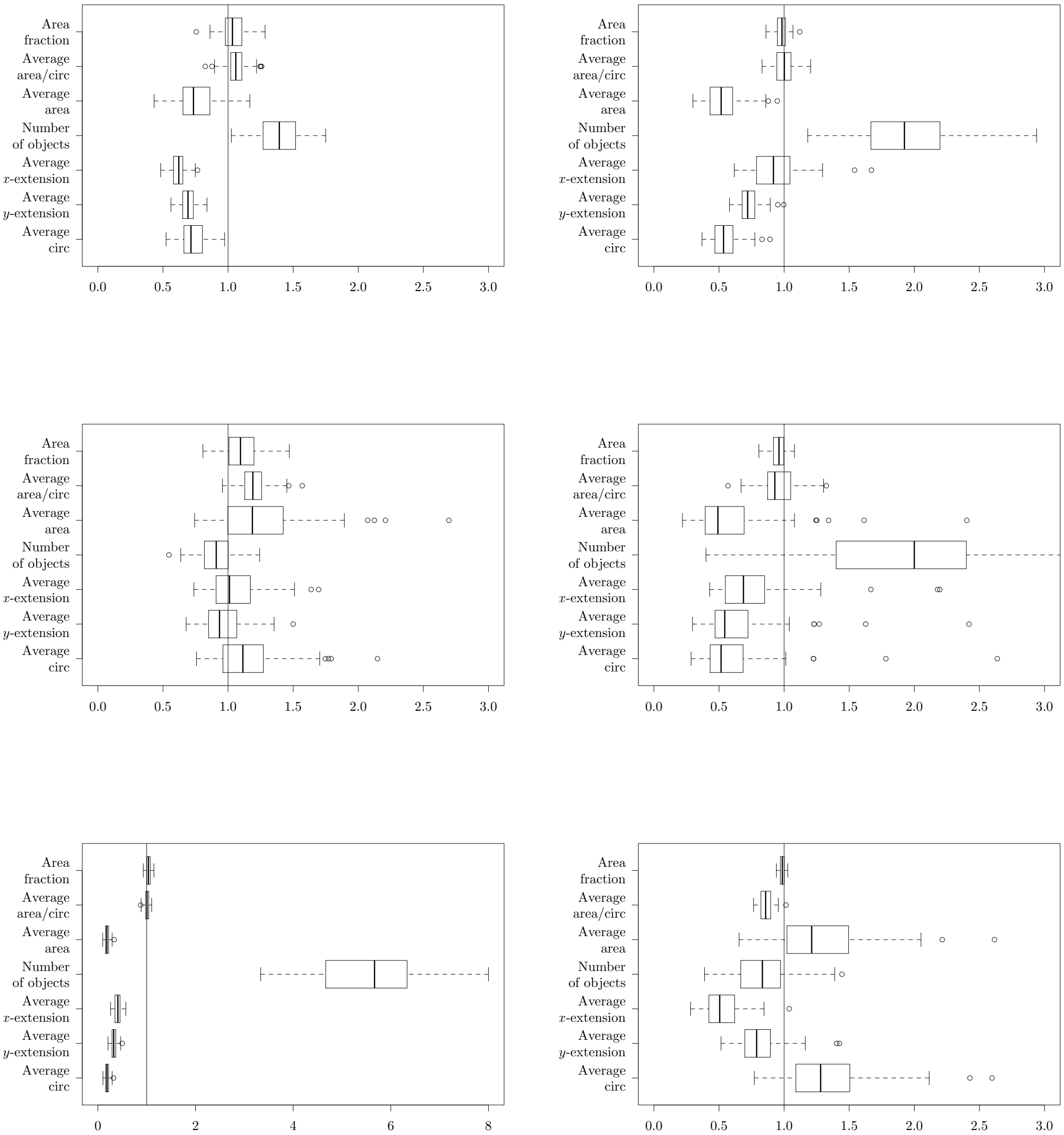


Figure 7: Box and Whisker plots of the standardized descriptive statistics for unconditional simulations from the fitted model $\widehat{p}_\theta(x)$. The value of the training image, which is one, is indicated by a vertical solid line. Plots corresponding to black and white objects are found in the left and right columns, respectively, whereas the upper, middle and lower rows corresponds to training image (a), (b) and (c) in Figure 2, respectively.

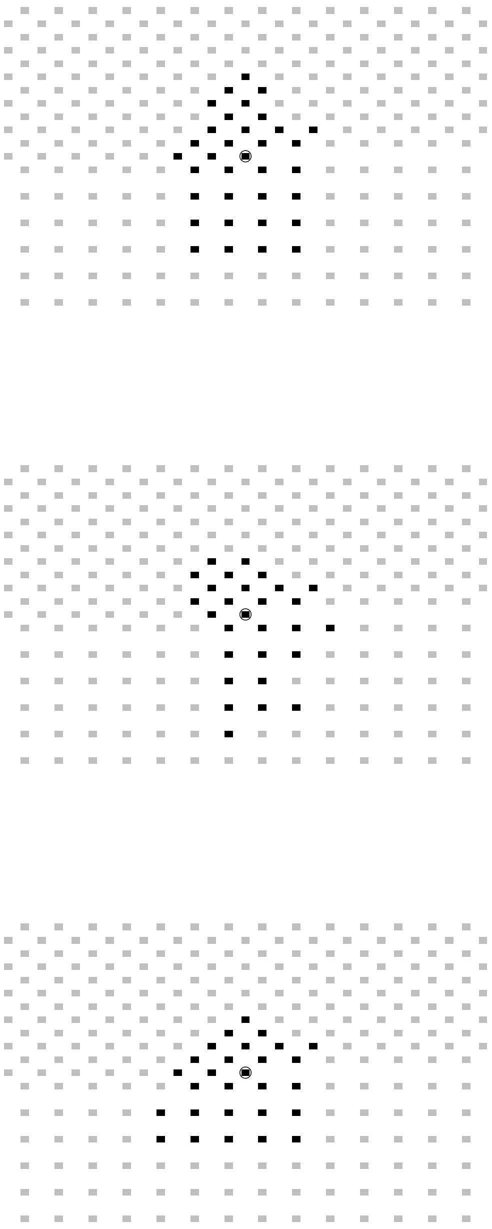


Figure 8: For a node (i, j) well away from the lattice borders, lower adjacent neighborhoods in the fitted POMMs, $p^{\star}_{\theta}(x)$, for each of the three training images in Figure 2. The upper, middle and lower plots corresponds to training image (a), (b) and (c) in Figure 2, respectively. Pixel (i, j) is black with a circle around, the pixels in N_{ij} are black, whereas the pixels in $\{\rho^{-1}(l), l = \rho(i, j) + 1, \ldots, mn\} \setminus N_{ij}$ are colored gray. Note that only a portion of the lattice closest to pixel (i, j) is shown in the figures.

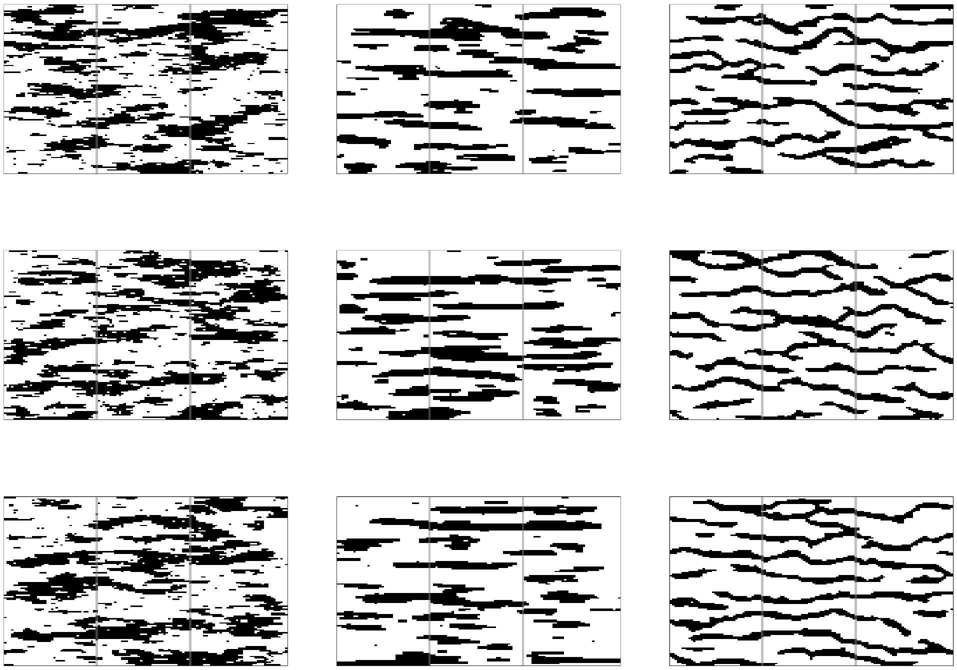


Figure 9: Three realizations from the the POMM approximation of the fitted $\widetilde{p}_\theta(x|z)$ when z contains exact observations of two columns in the training image. In the left, middle and right columns we find realizations from the models fitted to training image (a), (b) and (c) in Figure 2, respectively. The positions of the 'wells' are marked with vertical lines.

to data from two 'wells' taken from the training image we should expect that these are more similar to the training image than the unconditional realizations in Figure 6. Looking at the Box and Whisker plots for the conditional distribution in Figure 10, we see that 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 has increased. This all indicates that we have fewer small white objects within the black ones. For the other models the differences are too small to justify solid conclusions.

We have also run an example with eleven wells located at every tenth column. This puts many restrictions on $p_\theta(x|z)$, and could as discussed above potentially be problematic for the POMM approximation to $\widetilde{p}_\theta(x|z)$. Figure 13 in Appendix A contains three realizations from the POMM approximation to $\widetilde{p}_\theta(x|z)$ for each of the three training images. Again we can observe 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 $p_\theta(x|z)$. The widths of the boxes in the corresponding Box and Whisker plots shown in Figure 14 have decreased which is also to be expected when the amount of data is increased.

7 Closing remarks

In this paper we propose a new procedure for fitting an MRF to a given binary training image. The model is defined by a multi-grid approach, which means that we split the lattice into a series of sublattices. For each sublattice we fit an MRF conditioned on the values in the previous sublattices. Our examples should demonstrate the flexibility of our approach, but also the 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. We resolve this problem 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 this paper we have restricted the attention to binary training images, but our procedure can easily be extended to models with more than two possible values. However, the computational cost increases rapidly with the number of values, so in practice the procedure can only be used for training images with a limited number of values. A direct generalization of our modelling procedure to 3D is also possible, but for this to be computationally feasible in practice the simulation and estimation procedures must be carefully implemented, and the MRF neighborhoods must be chosen small. An alternative, and perhaps better, procedure to handle the 3D situation would be to model it as a Markov chain of 2D models, where our model formulation can be adopted for the 2D 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 (Strebelle, 2002; Journel and Zhang, 2006) avoid this directionality problem by simulating the nodes in a random order. When it comes to conditional simulation our

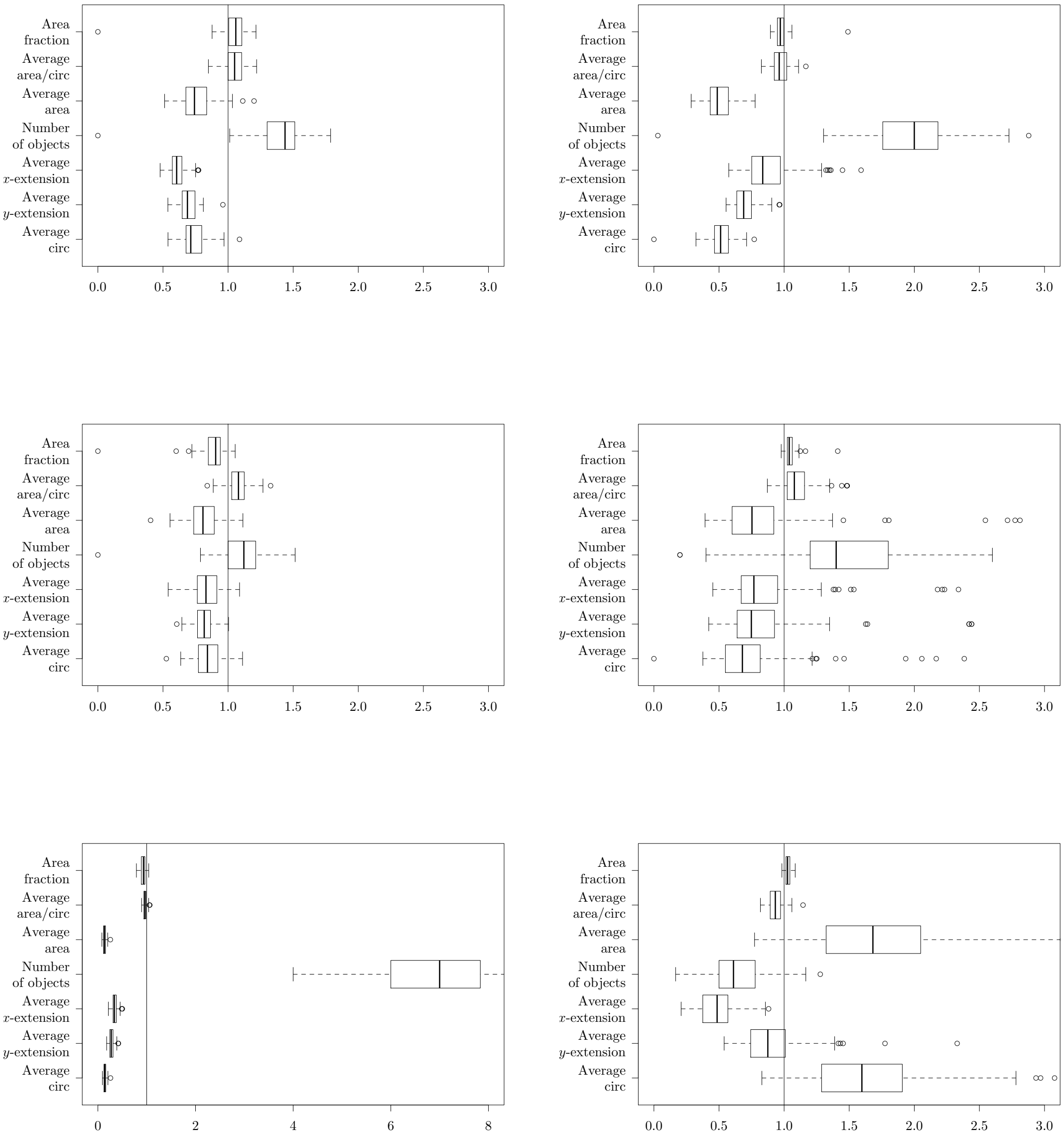


Figure 10: Box and Whisker plots of the standardized descriptive statistics for the conditional simulation when conditioning on two 'wells'. The value of the training image, which is one, is indicated by the solid line. Plots corresponding to black and white objects are found in the left and right columns, respectively, whereas the upper, middle and lower rows corresponds to training image (a), (b) and (c) in Figure 2, respectively.

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 Section 5.3, realizations from an approximation to the conditional distribution can be generated by feeding the conditional distribution into the approximation procedure of Austad and Tjelmeland (2011). As discussed in Section 1, conditional simulation from the multi-point statistics models is more complicated.

Acknowledgments

We thank the sponsors of the Uncertainty in Reservoir Evaluation (URE) project at NTNU.

References

Austad, H. and Tjelmeland, H. (2011). An approximate forward-backward algorithm applied to binary Markov random fields, *Technical report 11/2011*, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Bartolucci, F. and 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, *Journal of the Royal Statistical Society, Series B* **36**: 192–225.

Brooks, S., Gelman, A., Jones, G. and Meng, X. L. (2011). *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC, London.

Clifford, P. (1990). Markov random fields in statistics, *in* G. R. Grimmett and D. J. A. Welsh (eds), *Disorder in Physical Systems*, Oxford University Press, pp. 19–31.

Cressie, N. A. C. (1993). *Statistics for spatial data*, 2 edn, John Wiley, New York.

Cressie, N. and Davidson, J. (1998). Image analysis with partially ordered Markov models, *Computational Statistics and Data Analysis* **29**: 1–26.

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

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

Friel, N., Pettitt, A. N., Reeves, R. and Wit, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices, *Journal of Computational and Graphical Statistics* **18**: 243–261.

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

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*, 2nd edn, Chapman & Hall/CRC, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference, *Journal of American Statistical Association* **90**: 909–920.

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

Hurn, M., Husby, O. and Rue, H. (2003). A tutorial on image analysis, *in* J. Møller (ed.), *Spatial statistics and computational methods*, Vol. 173 of *Lecture Notes in Statistics*, Springer, pp. 87–139.

Journel, J. and Zhang, T. (2006). The necessity of a multiple-point prior model, *Mathematical Geology* **38**: 591–610.

Kindermann, R. and Snell, J. L. (1980). *Markov random fields and their applications*, American Mathematical Society, Providence, R.I.

Künsch, H. R. (2001). State space and hidden Markov models, *in* O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg (eds), *Complex Stochastic Systems*, Chapman & Hall/CRC.

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

Pettitt, A. N., Friel, N. and Reeves, R. (2003). Efficient calculation of the normalising constant of the autologistic and related models on the cylinder and lattice, *Journal of the Royal Statistical Society, Series B* **65**: 235–247.

Scott, A. L. (2002). Bayesian methods for hidden Markov models: Recursive compution in the 21st century, *Journal of the American Statistical Association* **97**: 337–351.

Stien, M. and Kolbjørnsen, O. (2011). Facies modeling using a markov mesh model specification, *Mathematical Geosciences* **43**: 611–624.

Strebelle, S. (2002). Conditional simulation of complex geolgical structures using multiple-point statistics, *Mathematical Geology* **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. and Besag, J. (1998). Markov random fields with higher order interactions, *Scandinavian Journal of Statistics* **25**: 415–433.

Ulvmoen, M. and 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.

A Additional plots

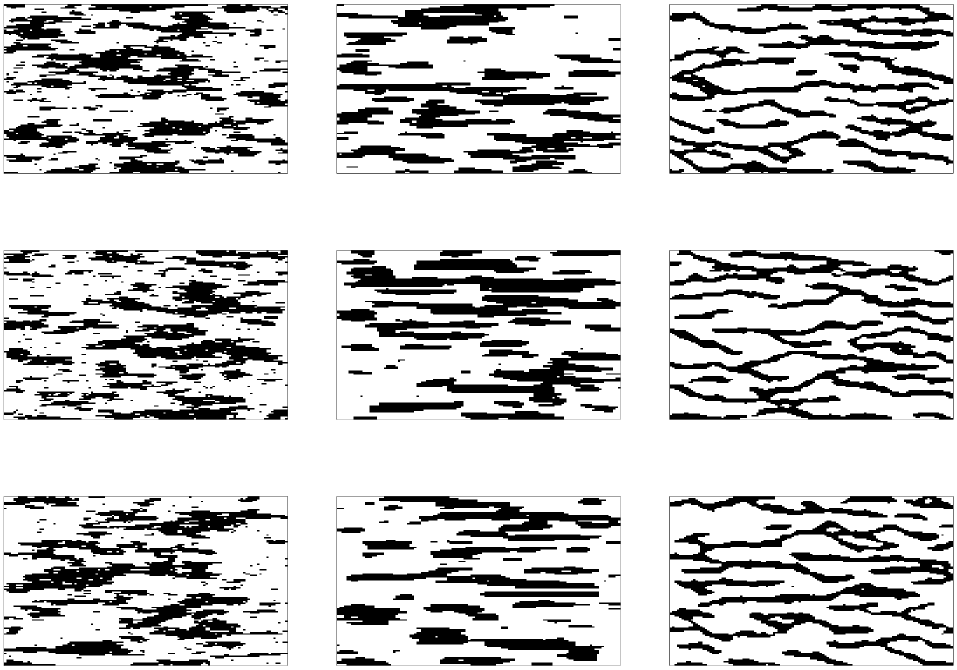


Figure 11: Three realizations from the fitted $p_\theta^\star(x)$, for training images (a) (left column), (b) (middle column) and (c) (right column) in Figure 2.

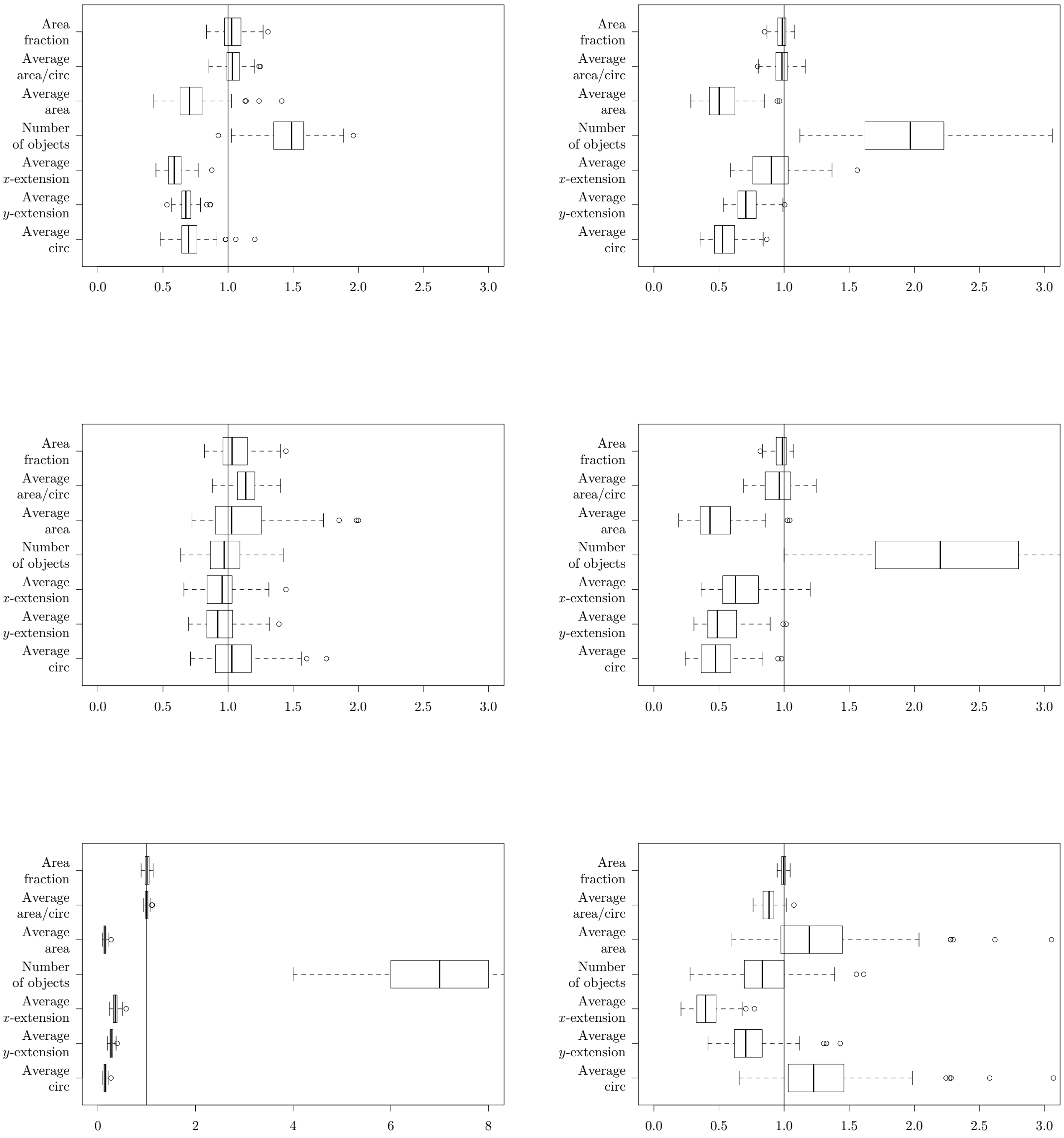


Figure 12: Box and Whisker plots of the standardized descriptive statistics for unconditional simulations from the fitted model $p^{\star}_{\theta}(x)$. The value of the training image, which is one, is indicated by a vertical solid line. Plots corresponding to black and white objects are found in the left and right columns, respectively, whereas the upper, middle and lower rows corresponds to training image (a), (b) and (c) in Figure 2, respectively.

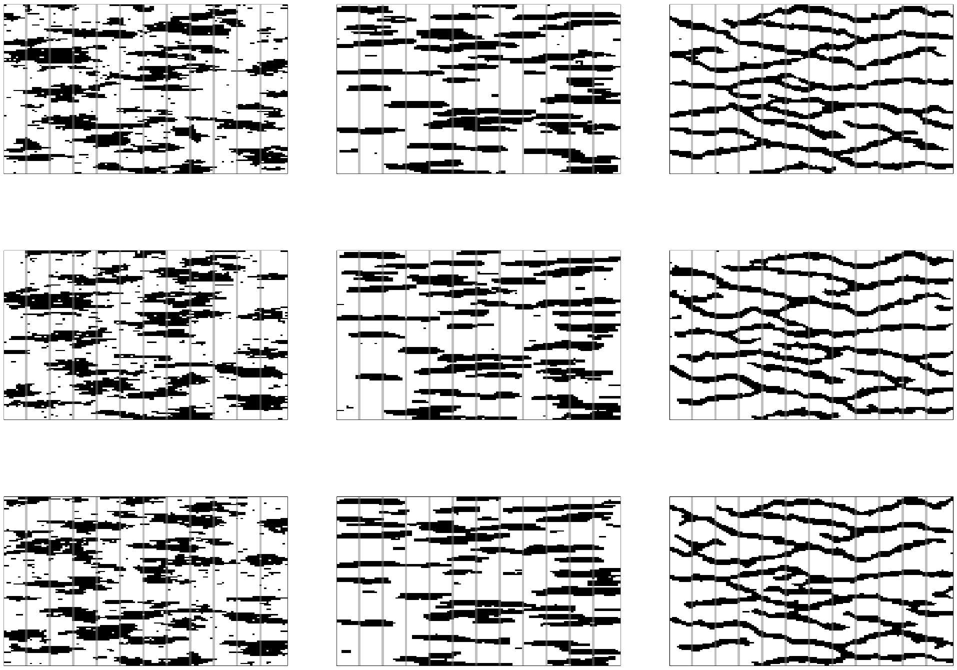


Figure 13: Three realizations from the the POMM approximation of the fitted $\widetilde{p}_\theta(x|z)$ when z contains exact observations of eleven columns in the training image. In the left, middle and right columns we find realizations from the models fitted to training image (a), (b) and (c) in Figure 2, respectively. The positions of the 'wells' are marked with vertical lines.

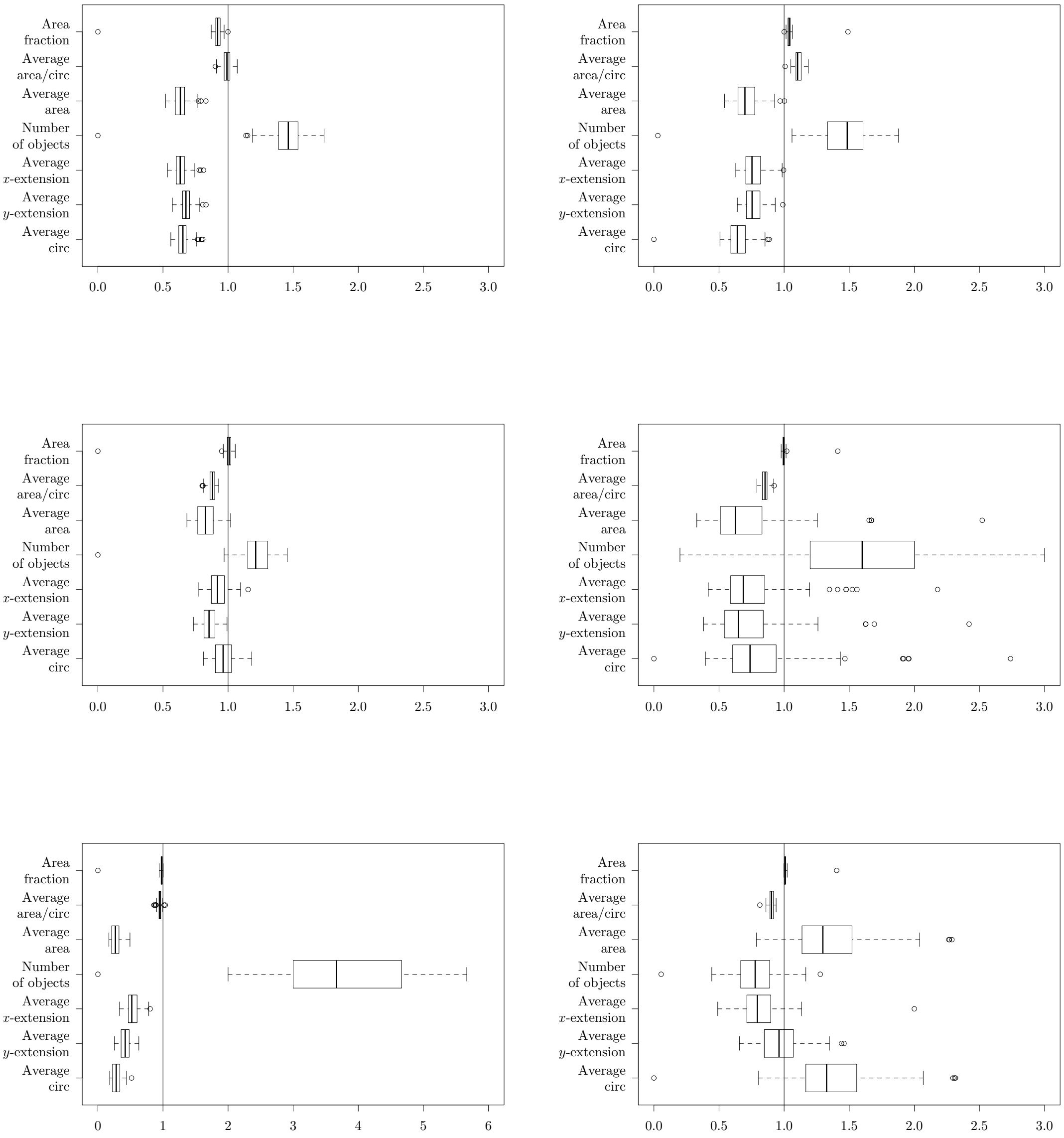


Figure 14: Box and Whisker plots of the standardized descriptive statistics for the conditional simulation when conditioning on eleven 'wells'. The value of the training image, which is one, is indicated by the solid line. Plots corresponding to black and white objects are found in the left and right columns, respectively, whereas the upper, middle and lower rows corresponds to training image (a), (b) and (c) in Figure 2, respectively.