Inference of the Transition Matrix in Convolved Hidden Markov Models and the Generalized Baum–Welch Algorithm

David Volent Lindberg and Henning Omre

***Abstract*—A convolved two-level hidden Markov model is defined, with convolved observations as the top layer, continuous responses as the hidden middle layer, and categorical variables following a Markov chain as the hidden bottom layer. The model parameters include the transition probabilities of the Markov chain, response means and variances, and the convolution kernel and error variance. The focus of this study is on categorical deconvolution of the observations into the hidden bottom categorical layer when the transition probability parameters defining the Markov chain transition matrix are unknown. The inversion is cast in a Bayesian setting where the solution is represented by an approximate posterior model. Three algorithms, all of them based on the approximate model, for inference on the unknown parameters are defined and discussed. The inference techniques are compared in two empirical studies and in a seismic case study from a petroleum reservoir offshore Norway. We conclude that an approximate expectation-maximization algorithm, appearing as a generalized Baum–Welch algorithm, is preferable if point parameter estimates and marginal maximum a posterior class predictions are required. If uncertainty quantifications are also required, a more computationally demanding Bayesian Markov chain Monte Carlo algorithm must be used.**

***Index Terms*—Bayesian inversion, deconvolution, expectation-maximization (EM) algorithm, forward–backward (FB) algorithm, markov chain monte carlo (MCMC), maximum-likelihood (ML) estimation, Monte Carlo expectation maximization (MCEM).**

I. Introduction

CONVOLVED signals appear in many physical phenomena for which the observed signal sequence is captured as a convolution of an underlying hidden process and a convolution kernel. In signal processing, the inverse problem of trying to recapture the sequence of the hidden process from the observations is termed a deconvolution problem. When the hidden process takes categorical states, the processing operation is termed categorical deconvolution.

The main motivation of this work is from seismic inversion. In seismic surveys, signal waves are emitted from a point source, thereafter travel through the subsurface Earth layers from where they are reflected back and measured by signal receivers [1]. An example from a marine seismic survey is shown in Fig. 1. The received signal is captured as a convolution between the contrasts of the Earth's reflection coefficients and a convolution kernel from the point source. Seismic deconvolution refers to removing the convolution effects and observation error to regain the reflection coefficients. The objective here is on categorical deconvolution directly into the categorical subsurface rock-type sediments.

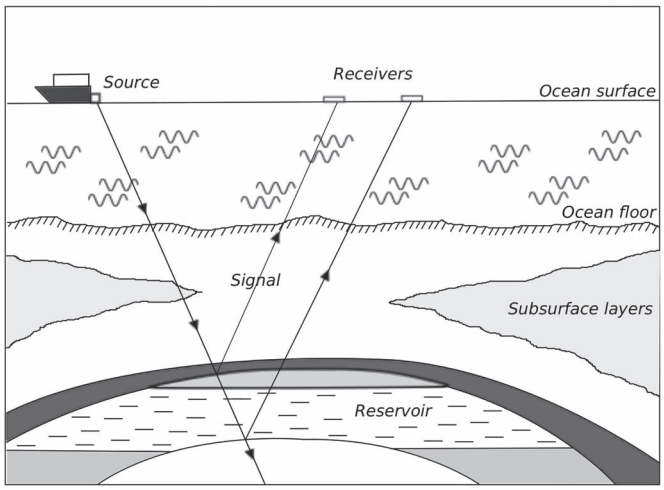

Fig. 1. Cross section of a schematic offshore reservoir subject to a marine seismic survey. Sound waves are emitted from the signal source, reflected by the subsurface layers, and registered by signal receivers.

A standard hidden Markov model (HMM) is defined for two layers: hidden categorical states following a Markov process as the bottom layer and observations that are elementwise conditional independent given the bottom layer as the top layer. A convolved HMM is defined with three layers: a hidden bottom categorical layer following a Markov process, a hidden continuous middle layer, and with convolved observations as the top layer. For the seismic inverse problem in particular, the seismic observations represent the top layer, the reflection coefficients represent the hidden middle layer, and the categorical subsurface rock-type classes represent the hidden bottom level. The model is introduced in Larsen *et al.* [2] for the seismic Bayesian inverse problem of deconvolution into subsurface lithology/fluid (LF) classes, also discussed in [3]–[5] and later generalized in [6]. Blind categorical deconvolution is considered in [7], i.e., categorical deconvolution when the parameters of the

Manuscript received August 15, 2014; revised February 9, 2015; accepted May 23, 2015.

D. V. Lindberg was with the Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway. He is now with DNV GL, 1363 Høvik, Norway (e-mail: davidlin@math.ntnu.no).

H. Omre is with the Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: omre@math.ntnu.no).

Digital Object Identifier 10.1109/TGRS.2015.2440415

convolution kernel and the observation error in addition are unknown. In this paper, we consider categorical deconvolution when the parameters in the hidden bottom categorical layer are unknown, particularly the transition matrix parameter defining the transition probabilities of the Markov process.

For the blind categorical deconvolution problem in [7], the unknown parameters of interest are assessed by approximate maximum-likelihood (ML) estimation and is conducted as a typical direct optimization problem assuming low-parametric convolution kernels. Direct optimization of the parameters defining the Markov process is, however, harder, as the number of unknown parameters is higher. For a standard HMM, Baum *et al.* [8] showed that all model parameters can be assessed simultaneously by maximum-likelihood estimation analytically, the method known as the Baum–Welch algorithm. For the convolved HMM, the common Baum–Welch algorithm cannot be applied because of the convolution effect.

Three methods for inference on the unknown Markov process parameters are defined and discussed: 1) an approximate Baum–Welch algorithm with analytical maximum-likelihood estimation; 2) a sample-based maximum-likelihood estimation approach by the Monte Carlo expectation-maximization (EM) algorithm; and 3) a fully Bayesian sample-based approach by a Markov chain Monte Carlo (MCMC) algorithm. All three approaches involve the approximation of the posterior model for the convolved HMM defined in [6]. The three inference methods are compared in a small example inspired by the standard model for blind seismic deconvolution, then in an empirical test study with three different reference transition matrices, and, finally, in a case study for seismic amplitude versus offset (AVO) inversion into LF classes from a reservoir offshore Norway. The model and algorithms are defined for a 1-D vertical profile, but extensions can be made to 2-D and 3-D settings. These extensions include MCMC inference by Gibbs sampling (see [5] and [9]).

In the notation, let $p(\cdot)$ be the common generic expression for both the probability density function (pdf) and the probability mass function. Moreover, the term $\phi_i(\cdot)$ denotes a Gaussian pdf, where i is the dimension of the variable.

II. Model Description

Suppose we have a stochastic field on the lattice $\mathcal{L} = \{1, \ldots, T\}$ with equal step length δ, with continuous observations discretized into $\mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_T)'$, $\mathbf{d}_t \in \mathbb{R}^m$, $t \in \mathcal{L}$. Associated to the observations are a set of continuous response variables $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_T)$, $\mathbf{r}_t \in \mathbb{R}^n$, $t \in \mathcal{L}$. For notational convenience, we consider 1-D observations and responses for the rest of this section, i.e., $d_t \in \mathbb{R}^1$, $r_t \in \mathbb{R}^1$. The relation between the response variables and the observations captures the convolution effect

$$d_t = \sum_{v=-a}^{a} w_v r_{t+v} + e_t \tag{1}$$

hence, each element d_t is thus gathered as a locally weighted sum of the elements in $\mathbf{r}$. Here, $\mathbf{w} = (w_{-a}, \ldots, w_a)'$ is a centered convolution kernel $(2a \times 1)$ vector corresponding to the weights, and e_t is an additive independent error term. The response variables, in turn, depend elementwise on a set of categorical variables $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)'$, where each state π_t, $t \in \mathcal{L}$, belongs to one of L classes, whose state space is set to integers for notational convenience, i.e., $\pi_t \in \Omega_\pi : \{1, \ldots, L\}$. This model defines the convolved HMM, and the relationship between the variables is presented in Fig. 2(a), where the directed arrows represent conditional dependence. A comparison with the standard HMM is presented in Fig. 2(b), in which the convolution relation is not present. The objective is to assess the hidden categorical layer given the observations $[\boldsymbol{\pi}|\mathbf{d}]$, whose solution in a Bayesian probabilistic setting, given all model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_p, \boldsymbol{\theta}_l\}$, is given by the posterior model

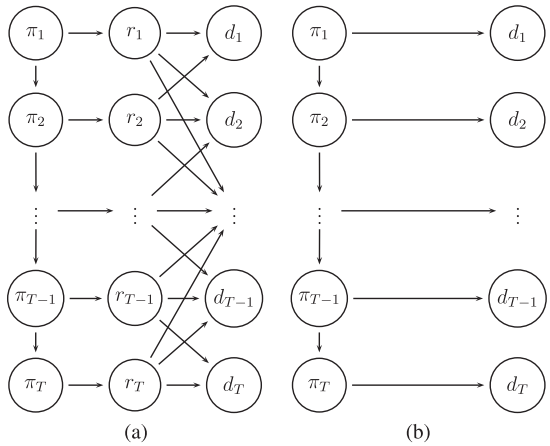

Fig. 2. Directed acyclic graph of a (a) convolved HMM and a (b) standard HMM in which the directed arrows represent dependence between the variables. Here, $\mathbf{d} = (d_1, \ldots, d_T)$ are the observed data, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ is the latent categorical field of interest, and $\mathbf{r} = (r_1, \ldots, r_T)$ is a latent continuous field.

$$p(\boldsymbol{\pi}|\mathbf{d};\boldsymbol{\theta}) = \frac{1}{p(\mathbf{d};\boldsymbol{\theta})} \times p(\mathbf{d}|\boldsymbol{\pi};\boldsymbol{\theta}_l) \times p(\boldsymbol{\pi};\boldsymbol{\theta}_p). \tag{2}$$

Here, $p(\boldsymbol{\pi};\boldsymbol{\theta}_p)$ is termed the prior model, and $p(\mathbf{d}|\boldsymbol{\pi};\boldsymbol{\theta}_l)$ is termed the likelihood model, defined by the model parameters $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_l$, respectively. The term $1/p(\mathbf{d};\boldsymbol{\theta})$ is a model-parameter-dependent normalizing constant.

A. Markov Chain Prior Model

The categorical variables in $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ are assigned a first-order stationary Markov chain prior model. Hence, the class at position t depends on the previous state only, i.e., $p(\pi_t|\pi_1, \ldots, \pi_{t-1}) = p(\pi_t|\pi_{t-1})$ [see Fig. 2(a)]. Denote the probability of a transition from class i to class j by $p(\pi_t = i|\pi_{t-1} = j) = P_{ij}$, $i, j \in \Omega_\pi$; hence, all transition probabilities are captured by the transition $(L \times L)$ matrix $\mathbf{P}$ with entries P_{ij}. The stationary distribution $(L \times 1)$ vector $\mathbf{p}^s$ of the transition matrix is defined by $\mathbf{p}^s = \mathbf{p}^s\mathbf{P}$ and is hence defined from $\mathbf{P}$. The stationary distribution represents the proportions of each class in the long run, i.e., in an infinitely long Markov chain, and is set as initial marginal prior pdf $p(\pi_1 = i) = p_i^s$. The Markov chain prior model is then

$$p(\boldsymbol{\pi};\boldsymbol{\theta}_p) = p(\pi_1;\boldsymbol{\theta}_p) \times \prod_{t=2}^{T} p(\pi_t|\pi_{t-1};\boldsymbol{\theta}_p). \tag{3}$$

The prior model is thus fully defined by the transition matrix, with all the prior model parameters given by $\boldsymbol{\theta}_p = \{\mathbf{P}\}$.

The thickness of the class layers in the categorical field can be shown to be geometric distributed, i.e., the probability that an arbitrary layer of class $i \in \Omega_\pi$ is Δ elements thick is

$$p(\Delta) = \left(1 - \frac{1}{u_i}\right)^{\Delta - 1} \times \frac{1}{u_i}, \qquad \Delta = 1, 2, 3, \ldots$$

with expected thicknesses defined by the diagonal elements of the transition matrix $\{P_{ii}, i = 1, \ldots, L\}$ as $u_i = \delta/(1 - P_{ii})$. High expected thickness thus corresponds to large probability for transitions from a class into the same class.

B. Likelihood Model

The likelihood model given the likelihood parameters $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{\text{rl}}, \boldsymbol{\theta}_{\text{ol}}\}$ is defined as

$$p(\mathbf{d}|\boldsymbol{\pi}; \boldsymbol{\theta}_l) = \int p(\mathbf{d}|\mathbf{r}; \boldsymbol{\theta}_{\text{ol}}) \times p(\mathbf{r}|\boldsymbol{\pi}; \boldsymbol{\theta}_{\text{rl}}) d\mathbf{r}. \tag{4}$$

Here, $p(\mathbf{d}|\mathbf{r}; \boldsymbol{\theta}_{\text{ol}})$ and $p(\mathbf{r}|\boldsymbol{\pi}; \boldsymbol{\theta}_{\text{rl}})$ are termed the observation likelihood model and the response likelihood model, defined by the model parameters $\boldsymbol{\theta}_{\text{ol}}$ and $\boldsymbol{\theta}_{\text{rl}}$, respectively. The two likelihood models define the forward relationship between $\mathbf{r}$ and $\mathbf{d}$ and between $\boldsymbol{\pi}$ and $\mathbf{r}$, respectively [see Fig. 2(a)].

The response likelihood model is defined with conditionally independent marginals; hence, each element r_t is independent of the rest $\mathbf{r}_{-t}$ when conditioned on π_t, i.e.,

$$p(\mathbf{r}|\boldsymbol{\pi}; \boldsymbol{\theta}_{\text{rl}}) = \prod_{t=1}^{T} p(r_t|\pi_t; \boldsymbol{\theta}_{\text{rl}}) \tag{5}$$

[see also Fig. 2(a)]. We choose to assign Gaussian pdfs to the marginal response likelihood models, i.e., $p(r_t|\pi_t; \boldsymbol{\theta}_{\text{rl}}) = \phi_1(r_t; \mu_{r|\pi_t}, \sigma^2_{r|\pi_t})$, whose expectation and variance are dependent on the class of π_t. Hence, the full response likelihood model is Gaussian, i.e., $p(\mathbf{r}|\boldsymbol{\pi}; \boldsymbol{\theta}_{\text{rl}}) = \phi_T(\mathbf{r}; \boldsymbol{\mu}_{r|\pi}, \boldsymbol{\Sigma}_{r|\pi})$, with expectation $(T \times 1)$ vector $\boldsymbol{\mu}_{r|\pi} = (\mu_{r|\pi_1}, \ldots, \mu_{r|\pi_T})'$ and diagonal covariance $(T \times T)$ matrix $\boldsymbol{\Sigma}_{r|\pi}$, with diagonal elements $(\sigma^2_{r|\pi_1}, \ldots, \sigma^2_{r|\pi_T})$ and all off-diagonal elements 0. The response likelihood model is hence fully defined by the model parameters $\boldsymbol{\theta}_{\text{rl}} = \{(\mu_{r|\pi}, \sigma^2_{r|\pi})_{\pi=1,\ldots,L}\}$.

The observation likelihood model captures the convolution effect of (1), for which the error term is assumed to be independent and Gaussian $p(e_t) = \phi_1(e_t; 0, \sigma_d^2)$. Assuming that both the convolution kernel $\mathbf{w}$ and the error variance σ_d^2 are stationary throughout the observation profile $\mathbf{d}$, the observation likelihood model is

$$p(\mathbf{d}|\mathbf{r}; \boldsymbol{\theta}_{\text{ol}}) = \phi_T\left(\mathbf{d}; \mathbf{W}\mathbf{r}, \sigma_d^2 \mathbf{I}\right) \tag{6}$$

with $\mathbf{W}$ a convolution $(T \times T)$ matrix with $\mathbf{w}$ as centered rows. There are border effects present for $t < a$ and $t > (T - a)$ as seen from (1), which we solve by choosing normalized truncated convolution kernels in $\mathbf{W}$. The observation likelihood model is hence fully defined by the model parameters $\boldsymbol{\theta}_{\text{ol}} = \{\mathbf{w}, \sigma_d^2\}$.

With both the response likelihood and the observation likelihood being Gaussian, the full likelihood model in (4) can be shown to be Gaussian as well, i.e.,

$$p(\mathbf{d}|\boldsymbol{\pi}; \boldsymbol{\theta}_l) = \phi_T\left(\mathbf{d}; \mathbf{W}\boldsymbol{\mu}_{r|\pi}, \mathbf{W}\boldsymbol{\Sigma}_{r|\pi}\mathbf{W}' + \sigma_d^2\mathbf{I}\right) \tag{7}$$

and is hence fully defined by the model parameters $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{\text{rl}}, \boldsymbol{\theta}_{\text{ol}}\}$. Notice that this Gaussian model does not have a diagonal covariance matrix; hence, the elements in $[\mathbf{d}|\boldsymbol{\pi}]$ are not elementwise conditionally independent, as shown in Fig. 2(a). Notice also that the variance of the Gaussian likelihood model can be regarded a sum of colored and white errors.

C. Approximate Posterior Model

The posterior model of the categorical variables in (2) is defined with prior and likelihood models given by (3) and (7), respectively, and is fully defined by the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_p, \boldsymbol{\theta}_l\}$. For the rest of this section, the parameter dependence is omitted in the notation. The posterior model is computable up to the constant $1/p(\mathbf{d})$, which is unfeasible to compute as it requires $O(L^T)$ computations. We therefore approximate the posterior model according to Larsen *et al.* [2], Rimstad and Omre [6], and Lindberg and Omre [7] by a kth-order approximate model $\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d})$, constituting a model on factorizable form, which can be exactly assessed by the recursive generalized forward–backward (FB) algorithm [10]. Define a kth-order categorical state by $\boldsymbol{\pi}_t^{(k)} = (\pi_{t-k+1}, \ldots, \pi_t)$, $t = k, \ldots, T$. The kth-order approximate posterior model on factorizable form is then given as

$$\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d}) = C_d^{(k)} \times q_{k-1}\left(\boldsymbol{\pi}_{k-1}^{(k-1)}|\mathbf{d}\right) \times q_T\left(\boldsymbol{\pi}_T^{(k-1)}|\mathbf{d}\right) \times \left[\prod_{t=k}^{T} \hat{p}^{(k)}\left(\mathbf{d}_t^{(k)}|\boldsymbol{\pi}_t^{(k)}\right)^{1/k} \times p\left(\boldsymbol{\pi}_t^{(k)}|\boldsymbol{\pi}_{t-1}^{(k)}\right)\right] \tag{8}$$

where $C_d^{(k)}$ is a normalizing constant, $q_{k-1}(\cdot)$ and $q_T(\cdot)$ are edge factors, $\hat{p}^{(k)}(\mathbf{d}_t^{(k)}|\boldsymbol{\pi}_t^{(k)})$ is a kth-order approximate likelihood model, and $p(\boldsymbol{\pi}_t^{(k)}|\boldsymbol{\pi}_{t-1}^{(k)})$ defines the Markov chain transitions between the kth-order states. The generalized FB algorithm for computing the approximate posterior model is specified in Algorithm A1 in the Appendix. The approximate posterior model is in fact a nonstationary kth-order Markov process with respect to the kth-order states. Under the approximate posterior model, it is hence possible to compute the normalizing constant $C_d^{(k)}$ in (8) exactly by the FB algorithm (for more details on the approximation and its notation, see [7]).

To sample realizations from the exact posterior model $p(\boldsymbol{\pi}|\mathbf{d})$ in (2), the approximate posterior model is used as proposal distribution in an independent proposal MCMC algorithm; the sampling algorithm is given in Algorithm A2 in the Appendix. Higher order approximations provide better accuracy, but at a higher computation cost. In a predictive setting, i.e., with the model parameters assumed known, an approximation order of $k = 2$ is shown to provide a good tradeoff between predictive accuracy and computer demands (see [6]). When doing inference on the convolution parameters $\boldsymbol{\theta}_{\text{ol}}$ only, an approximation order of $k = 4$ is shown to provide a good tradeoff between accuracy and computer demands (see [7]).

III. Inference on the Markov Process Model Parameters

Here, it is assumed that the prior model parameters, i.e., the transition matrix $\boldsymbol{\theta}_p = \{\mathbf{P}\}$, are unknown; all likelihood parameters $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{\mathrm{rl}}, \boldsymbol{\theta}_{\mathrm{ol}}\}$ are assumed known. Three approaches for inference on $\mathbf{P}$ are defined: 1) ML estimation by an approximate EM (aEM) algorithm; 2) ML estimation by a Monte Carlo EM (MCEM) algorithm; and 3) simulation-based Bayesian inference of the posterior parameter pdf by an MCMC algorithm. All three approaches involve the kth-order approximation of the posterior model of the convolved HMM. For a fixed approximation order k, the aEM algorithm provides an approximate ML estimate and is very computationally efficient. The MCEM algorithm provides an ML estimate up to the Monte Carlo sampling precision, but is relatively computer demanding. The MCMC algorithm provides an asymptotically exact assessment of the Bayesian posterior pdf on the cost of considerable computational resources. One major advantage for inference through MCMC compared with ML is that the ensemble of simulated realizations obtained when doing MCMC enables us to do simultaneous inference on the unknown parameters and the hidden categorical profile, whereas ML provides a parameter point estimate only. For the categorical field $\boldsymbol{\pi}$, Bayesian inference by MCMC thus provides samples from the posterior distribution randomized over $\mathbf{P}$, $p(\boldsymbol{\pi}|\mathbf{d})$, whereas for inference by aEM and MCEM, we may only sample from the posterior distribution based on plug-in parameter prior model $p(\boldsymbol{\pi}|\widehat{\mathbf{P}})$. However, ML estimation is, as mentioned, much less computer demanding than MCMC, particularly so for aEM; hence, a tradeoff between precision and computer demands is required. Computer demand is crucial if the inference is to be made in real time or based on 2-D or 3-D observations.

A. ML Estimation by aEM

For estimation by ML, we are interested in finding the parameter $\mathbf{P}$ such that the marginal likelihood $p(\mathbf{d};\mathbf{P})$ is maximized. The EM algorithm was proposed for computing the ML estimate in the case of incomplete data [11]. The term incomplete data refers, in our case, to the latent field $\boldsymbol{\pi}$, with observed data $\mathbf{d}$ that are assessable only through the likelihood function $p(\mathbf{d}|\boldsymbol{\pi})$. The EM algorithm is suitable if maximization of $p(\mathbf{d};\mathbf{P})$ is unfeasible, but maximization of the joint pdf $p(\boldsymbol{\pi},\mathbf{d};\mathbf{P})$ is easier.

For any HMM, suppose the current transition matrix estimate is $\mathbf{P}^*$. For each update in the EM algorithm, the E-step constitutes computing the expected value of the logarithm of the joint pdf under the current posterior pdf, i.e.,

$$\begin{aligned} Q(\mathbf{P},\mathbf{P}^*) &= E_{\boldsymbol{\pi}|\mathbf{d};\mathbf{P}^*}\left[\log p(\boldsymbol{\pi},\mathbf{d};\mathbf{P})\right] \\ &= \sum_{\boldsymbol{\pi}} \log p(\boldsymbol{\pi},\mathbf{d};\mathbf{P}) p(\boldsymbol{\pi}|\mathbf{d};\mathbf{P}^*) \end{aligned} \tag{9}$$

followed by the M-step, which constitutes maximizing this quantity with respect to the unknown parameters:

$$\mathbf{P} = \arg\max_{\mathbf{P}} \left\{ Q(\mathbf{P},\mathbf{P}^*) \right\}. \tag{10}$$

For the standard HMM in Fig. 2(b), Baum *et al.* [8] proved that the parameters maximizing the Q function in the M-step in (10) are analytical tractable (see also [12] for a detailed derivation). For a standard HMM, the likelihood parameters are analytical tractable as well [8]. In fact, the E- and M-steps are performed simultaneously after first computing the marginal posterior probabilities for the hidden states through the FB algorithm. The full algorithm for ML estimation of all parameters in a standard HMM is known as the Baum–Welch algorithm. It can be shown that the transition matrix elements maximizing the Q function in (10) are

$$P_{ij} = \frac{\sum_{t=2}^{T} p(\pi_{t-1}=i, \pi_t = j|\mathbf{d};\mathbf{P}^*)}{\sum_{t=1}^{T} p(\pi_t = i|\mathbf{d};\mathbf{P}^*)}, \qquad i,j = 1,\ldots,L$$

where the probabilities are obtained from the FB algorithm. For the convolved HMM, the probabilities computed in the generalized FB algorithm are kth-order approximations, i.e., $\hat{p}^{(k)}(\boldsymbol{\pi}_t^{(k-1)}|\mathbf{d})$, $t = k-1,\ldots,T$ (see Algorithm A1 in the Appendix). We can thus approximate the given probabilities by summing out for the remaining states

$$\begin{aligned} &\hat{p}\left(\pi_t^{(2)}|\mathbf{d};\mathbf{P}^*\right) \\ &= \begin{cases} \sum\limits_{\pi_1}\cdots\sum\limits_{\pi_{t-2}}\sum\limits_{\pi_{t+1}}\cdots\sum\limits_{\pi_{k-1}} \hat{p}^{(k)}\left(\boldsymbol{\pi}_{k-1}^{(k-1)}|\mathbf{d};\mathbf{P}^*\right), & t<k-1 \\ \sum\limits_{\boldsymbol{\pi}_{t-k+2}}\cdots\sum\limits_{\pi_{t-2}} \hat{p}^{(k)}\left(\boldsymbol{\pi}_t^{(k-1)}|\mathbf{d};\mathbf{P}^*\right), & t\geq k-1 \end{cases} \end{aligned} \tag{11}$$

$$\begin{aligned} &\hat{p}(\pi_t|\mathbf{d};\mathbf{P}^*) \\ &= \begin{cases} \sum\limits_{\pi_1}\cdots\sum\limits_{\pi_{t-1}}\sum\limits_{\pi_{t+1}}\cdots\sum\limits_{\pi_{k-1}} \hat{p}^{(k)}\left(\boldsymbol{\pi}_{k-1}^{(k-1)}|\mathbf{d};\mathbf{P}^*\right), & t<k-1 \\ \sum\limits_{\pi_{t-k+2}}\cdots\sum\limits_{\pi_{t-1}} \hat{p}^{(k)}\left(\boldsymbol{\pi}_t^{(k-1)}|\mathbf{d};\mathbf{P}^*\right), & t\geq k-1. \end{cases} \end{aligned} \tag{12}$$

The aEM algorithm is given in the following.

Algorithm 1 aEM for estimating $\mathbf{P}$

Initialize some $\mathbf{P}^*$.
Iterate until convergence, do

1) **FB-algorithm**: Run the generalized FB algorithm in Algorithm A1 to obtain $\hat{p}(\boldsymbol{\pi}_t^{(k-1)}|\mathbf{d};\mathbf{P}^*)$, $t=k-1,\ldots,T$. Compute $\hat{p}(\pi_{t-1},\pi_t|\mathbf{d};\mathbf{P}^*)$ and $\hat{p}(\pi_t|\mathbf{d};\mathbf{P}^*)$ by (11) and (12).
2) **E and M step**: Compute for $i,j = 1,\ldots,L$

$$P_{ij} = \frac{\sum_{t=2}^{T} \hat{p}(\pi_{t-1}=i,\pi_t=j|\mathbf{d};\mathbf{P}^*)}{\sum_{t=1}^{T} \hat{p}(\pi_t=i|\mathbf{d};\mathbf{P}^*)}.$$

3) Set $\mathbf{P}^* = \mathbf{P} = \{P_{ij}\}_{i,j=1,\ldots,L}$

The aEM algorithm is a generalization of the Baum–Welch algorithm. While the likelihood model parameters are analytically tractable in the Baum–Welch algorithm for the standard HMM, this is not the case in the aEM algorithm for the convolved HMM due to the convolution effect. In the aEM algorithm, activating the FB-step for the kth-order approximation

has runtime $O(R_{\mathrm{FB},k})$ (see the Appendix), and this step is the most computationally demanding one. This step must be activated once per EM-step. Note that zero-probability transitions are reproduced by each EM-step.

The estimate $\mathbf{P}^*$, for a fixed kth order, appears as an approximate ML estimate of the transition $(L \times L)$ matrix $\mathbf{P}$. One may assess the precision of each of the elements of $\mathbf{P}^*$ by computing the corresponding Hessian matrices. The inversion of the convolved HMM (see [6]) can be based on the plug-in prior model $p(\boldsymbol{\pi}; \mathbf{P}^*)$.

B. ML Estimation by MCEM

It is possible to sample the hidden variables $\boldsymbol{\pi}$ from its posterior model given all model parameters; hence, we may proceed with an MCEM algorithm, as proposed by Wei and Tanner [13]. In each iteration with current estimate $\mathbf{P}^*$, S realizations of the categorical field $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{S} \sim p(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^*)$ are sampled (see Algorithms A1 and A2 in the Appendix). Note that this step involves sampling from the kth-order approximation of the posterior model of the convolved HMM. Next, in the E-step, the Q-function integral in (9) is approximated by its Monte Carlo estimate

$$\begin{aligned}\widehat{Q}(\mathbf{P}, \mathbf{P}^*) &= \frac{1}{S}\sum_{s=1}^{S} \log\left[p(\boldsymbol{\pi}^{(s)}, \mathbf{d}; \mathbf{P})\right] \\ &= \frac{1}{S}\sum_{s=1}^{S}\left\{\log\left[p\left(\mathbf{d}|\boldsymbol{\pi}^{(s)}; \mathbf{P}\right)\right] + \log\left[p\left(\boldsymbol{\pi}^{(s)}; \mathbf{P}\right)\right]\right\}\end{aligned} \tag{13}$$

followed by the M-step, maximizing this approximate Q function:

$$\begin{aligned}\mathbf{P} &= \arg\max_{\mathbf{P}}\left\{\widehat{Q}(\mathbf{P}, \mathbf{P}^*)\right\} \\ &= \arg\max_{\mathbf{P}}\left\{\frac{1}{S}\sum_{s=1}^{S}\log\left[p\left(\boldsymbol{\pi}^{(s)}; \mathbf{P}\right)\right]\right\}.\end{aligned} \tag{14}$$

Note that the likelihood model $p(\mathbf{d}|\boldsymbol{\pi}^{(s)}; \mathbf{P})$ is a constant independent of $\mathbf{P}$. The MCEM algorithm for the convolved HMM is described in the following.

Algorithm 2 MCEM for estimating $\mathbf{P}$

Initialize some $\mathbf{P}^*$.
Iterate until convergence do

1) **FB-algorithm**: Run the generalized FB algorithm in Algorithm A1 to obtain the approximate posterior model $\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^*)$.
2) Sample $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{S} \sim p(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^*)$ by the independent proposal MCMC algorithm in Algorithm A2.
3) **MCE step**: Compute $\widehat{Q}(\mathbf{P}, \mathbf{P}^*)$ by (13).
4) **M step**: Maximize numerically $\widehat{Q}(\mathbf{P}, \mathbf{P}^*)$ by (14).
5) Set $\mathbf{P}^* = \mathbf{P}$

In the MCEM algorithm, there are three computationally demanding parts: 1) activating the kth-order FB algorithm with runtime $O(R_{\mathrm{FB},k})$ in step 1; 2) convergence of the MCMC algorithm in step 2; and 3) numerical optimization in step 4. Each part must be activated once per EM-step. Note that increasing order k will increase the runtime in part 1 but decrease the runtime in part 2, and vice versa. It is hence obvious that the computational demands of the MCEM algorithm are much larger than those of the aEM algorithm. This is not surprising since the former provides a consistent ML estimate up to MC sampling error, whereas the latter only provides an approximate ML estimate.

The sample size S may depend on the iteration, with higher MC accuracy for larger sample sizes. Early iterations tolerate smaller sample sizes, whereas the final iterations require larger sample sizes when the need for accuracy increases (see [14] and [15] for expansions to varying-sample-size automated MCEM algorithms). A special case of the MCEM algorithm is for sampling a single realization of the latent field, which is called a stochastic EM algorithm [16]. This is a quite poor approximate EM algorithm, as one sample in each E-step provides a poor approximation to the sum in the original E-step.

One problem occurring when doing inference by the MCEM algorithm is that, while the EM algorithm guarantees an increase in the likelihood function in each update, this is not necessarily so for its stochastic version due to the MC error. A solution proposed by Caffo *et al.* [14] for $N = 1$ and later generalized by Jank [15] is to check in each step that $U_{t,N} = \hat{\delta}_{t,N} + z_\alpha(\hat{\sigma}_{t,N}^2/S_t) \geq 0$, in which

$$\begin{aligned}\hat{\delta}_{t,N} &= \left|\widehat{Q}\left(\mathbf{P}^{(t)}, \mathbf{P}^{(t-N)}\right) - \widehat{Q}\left(\mathbf{P}^{(t-N)}, \mathbf{P}^{(t-N)}\right)\right| \\ &= \frac{1}{S}\sum_{s=1}^{S}\log\frac{p\left(\boldsymbol{\pi}_s, \mathbf{d}; \mathbf{P}^{(t)}\right)}{p\left(\boldsymbol{\pi}_s, \mathbf{d}; \mathbf{P}^{(t-N)}\right)}\end{aligned} \tag{15}$$

and $\hat{\sigma}_{t,N}$ is the standard deviation of $\hat{\delta}_{t,N}$. Hence, this measure is the absolute differences in the $Q(\cdot,\cdot)$ functions between the current estimate and the estimate N steps back. When the MCEM algorithm converges, this measure should thus approach 0.

The estimate $\mathbf{P}^*$ appears as an ML estimate up to MC sampling error. The precision of the elements in $\mathbf{P}^*$ can be assessed by computing the corresponding Hessian matrix. The inversion of the convolved HMM can be based on the plug-in prior model $p(\boldsymbol{\pi}; \mathbf{P}^*)$.

C. Bayesian Estimation by MCMC

For parameter estimation by Bayesian inference, a prior parameter pdf is assigned to the unknown transition matrix $\mathbf{P}$. In an MCMC setting, each iteration s is defined by first sampling the categorical field from its posterior distribution under the current parameter estimate, i.e., $\boldsymbol{\pi}^{(s)} \sim p(\boldsymbol{\pi}|\mathbf{d}, \mathbf{P}^{(s-1)})$ (see Algorithm A1 and Algorithm A2 in the Appendix), followed by sampling the unknown transition matrix from its parameter posterior distribution given the current categorical field $\mathbf{P}^{(s)} \sim p(\mathbf{P}|\boldsymbol{\pi}^{(s)}, \mathbf{d})$. Note that the first step involves sampling from the kth-order approximation of the posterior model of the convolved HMM. After convergence, the samples $\{\boldsymbol{\pi}^{(s)}, \mathbf{P}^{(s)}\}_{s=1}^{S}$ will contain dependent realizations from the joint posterior model $p(\boldsymbol{\pi}, \mathbf{P}|\mathbf{d})$ from which we can derive the desired inference. We must use a Gibbs sampling strategy since the joint

constants, i.e., the normalizing constant in the posterior model in (2) and the normalizing constant in the approximate posterior model in (8), are dependent on the model parameter $\mathbf{P}$, which prevents joint sampling of $[\boldsymbol{\pi}, \mathbf{P}]$.

In the second step, the parameter posterior model can be shown to depend on $\boldsymbol{\pi}$ only, i.e., $p(\mathbf{P}|\boldsymbol{\pi}, \mathbf{d}) = \text{const} \times p(\boldsymbol{\pi}|\mathbf{P}) \times p(\mathbf{P})$. A Dirichlet prior model is assigned to each row of the transition matrix $\mathbf{P}$, with identical hyperparameters $\eta \geq 0$ following Eidsvik *et al.* [17] for $i = 1, \ldots, M$, $p(\mathbf{P}_i) = p(P_{i1}, \ldots, P_{iM})$ given by

$$p(\mathbf{P}_i; \eta) = \frac{1}{B(\eta)} \times \prod_{j=1}^{M} P_{ij}^{\eta-1} \quad B(\eta) = \frac{\Gamma(\eta)^M}{\Gamma(M\eta)} \tag{16}$$

where $P_{i1}, \ldots, P_{iM} \geq 0$, and $\sum_j P_{ij} = 1$; hence, $P_{iM} = 1 - \sum_{j=1}^{M-1} P_{ij}$. Define the $(M \times 1)$ hyperparameter vector $\boldsymbol{\eta} = (\eta, \ldots, \eta)'$. The Dirichlet pdf is chosen as it is the conjugate distribution, i.e., given $\boldsymbol{\pi}$, the parameter posterior distribution is also Dirichlet, i.e.,

$$p(\mathbf{P}|\boldsymbol{\pi}; \boldsymbol{\eta}) = \frac{1}{B(\boldsymbol{\eta} + \mathbf{n}_i)} \prod_{j=1}^{M} P_{ij}^{\eta + n_{ij} - 1}$$
$$B(\boldsymbol{\eta} + \mathbf{n}_i) = \frac{\prod_{j=1}^{M} \Gamma(\eta + n_{ij})}{\Gamma\left(M\eta + \sum_{j=1}^{M} n_{ij}\right)} \tag{17}$$

where $\mathbf{n}_i = (n_{i1}, \ldots, n_{iM})$, and n_{ij} is the total number of transitions from class i into class j throughout $\boldsymbol{\pi}$. Hence, we can sample each row in $\mathbf{P}$ from its posterior distribution in Gibbs steps. The complete MCMC algorithm is given in the following.

Algorithm 3 MCMC for estimating $\mathbf{P}$

Set initial $\mathbf{P}^{(0)}$, $\boldsymbol{\pi}^{(0)}$.
For $s = 1, \ldots$ do
1) **FB-algorithm**: Run the generalized FB algorithm in Algorithm A1 to obtain the approximate posterior model $\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^{(s-1)})$
2) Sample $\boldsymbol{\pi}^{(s)} \sim p(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^{(s-1)})$ by the independent proposal MCMC algorithm in Algorithm A2.
3) Sample $\mathbf{P}^{(s)} \sim p(\mathbf{P}|\boldsymbol{\pi}^{(s)}; \boldsymbol{\eta})$ according to (17).

In the MCMC algorithm, there are two computationally demanding parts: 1) activating the kth-order FB algorithm with runtime $O(R_{\text{FB},k})$ in step 1 and 2) convergence of the MCMC algorithm in step 2. In addition, there is an outer MCMC loop. Consequently, the algorithm appears as a nested MCMC algorithm, which is alarming since nested MCMC algorithms often appear with poor convergence rates. Note that increasing order k increases the demands in part 1, whereas it decreases the demands of part 2, and vice versa. We expect the computational demands of the MCMC algorithm to be considerably larger than those of the MCEM and aEM algorithms, particularly larger than those of the aEM algorithm. Note, however, that the MCMC algorithm provides samples from the joint pdf $p(\boldsymbol{\pi}, \mathbf{P}|\mathbf{d})$, which makes assessment of both $p(\mathbf{P}|\mathbf{d})$ and $p(\boldsymbol{\pi}|\mathbf{d})$ possible, the latter randomized over $\mathbf{P}$.

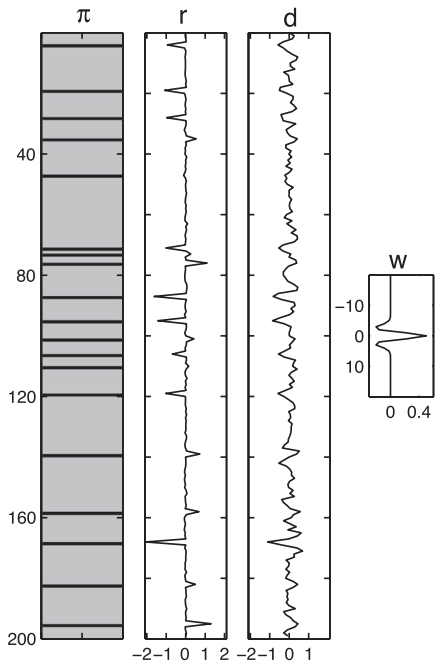

Fig. 3. Reference categorical profile, response profile, observation profile, and reference Ricker wavelet in the seismic inversion toy example.

IV. Toy Example: Seismic Deconvolution With Two Classes

A convolved HMM of length $T = 200$ with $L = 2$ possible classes $\Omega_\pi = \{\text{gray}, \text{black}\}$ is considered. This particular example is inspired by the more conventional Bernoulli–Gaussian (BG) model, with parameters set similar to the synthetic trace example of Rosec *et al.* [18]. The BG model is commonly applied on blind seismic deconvolution problems, with the objective of identifying transitions between different subsurface geological layers from convolved seismic data [18], [19]. There are thus two classes for which black represents a change between two layers denoted a high reflector point and gray represents staying within the same layer.

A reference categorical profile $\boldsymbol{\pi}$ is generated from a prior model with reference transition matrix $\mathbf{P}$ (see Fig. 3). The reference transition matrix and the empirical transition matrix $\mathbf{P}_e$ estimated from $\boldsymbol{\pi}$ by a simple counting process are

$$\mathbf{P} = \begin{pmatrix} 0.90 & 0.10 \\ 0.98 & 0.02 \end{pmatrix} \quad \mathbf{P}_e = \begin{pmatrix} 0.89 & 0.11 \\ 1 & 0 \end{pmatrix} \tag{18}$$

with reference and empirical stationary pdfs $\mathbf{p}^s = (0.91, 0.09)$ and $\mathbf{p}_e^s = (0.90, 0.10)$ and expected thicknesses $\mathbf{u} = (10.00, 1.02)$ and $\mathbf{u}_e = (9.47, 1.00)$. Hence, we expect thicker gray layers with some thin black layer transitions. Notice that the empirical transition matrix based on a finite profile will differ from the reference transition matrix; in particular, there are no consecutive black transitions in $\boldsymbol{\pi}$, resulting in a corresponding zero probability in $\mathbf{P}_e$. A longer trace would better randomize over the model; in the limit for an infinitely long trace, the empirical values would equal the reference values. When making inference on the transition matrix by the three proposed inference techniques, we thus expect the estimated transition matrix

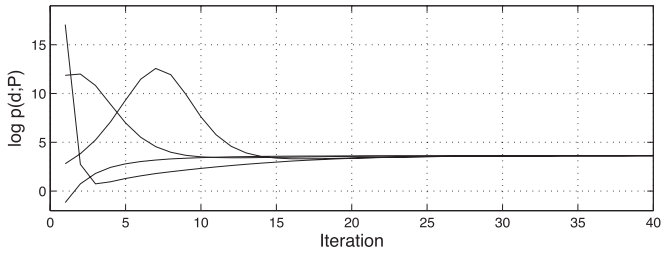

Fig. 4. Trace plots for aEM of the approximate log-likelihood value $\log\{\hat{p}^{(k)}(\mathbf{d};\mathbf{P})\}$ for different initial values.

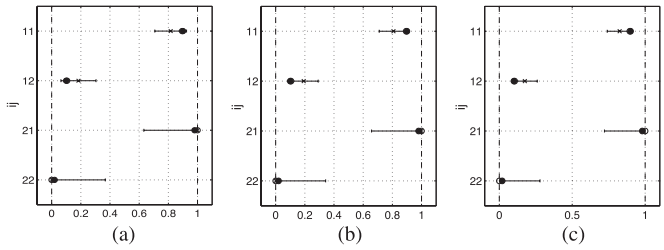

Fig. 5. Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i,j \in \{1,2\}$. The estimates are denoted as crosses, reference values as black dots, and empirical values as white dots. (a) aEM. (b) MCEM. (c) MCMC.

to reproduce the empirical transition matrix better than the reference transition atrix.

The response likelihood model parameters are defined for the two classes as $\mu_{r|\pi} = \{0, 0\}$ and $\sigma^2_{r|\pi} = \{0.001, 1\}$; hence, the response levels are equal. High reflector points have high variance, whereas inhomogeneities within the layers are represented by a much smaller variance. The reference response profile simulated from $\boldsymbol{\pi}$ is presented in Fig. 3. Notice that most high reflector points are registered by spikes in the response. Some high reflector points do not cause a significant deviation in the response due to identical response level, and we cannot expect these to recapture in the inversion. The observation profile $\mathbf{d}$ is simulated from the response profile $\mathbf{r}$ convolved with a Ricker wavelet; $R(t;\lambda,\gamma) = \gamma(1-(t^2/\lambda^2))\exp\{-(t^2/2\lambda^2)\}$, $\lambda > 0$, $\gamma \geq 0$, with $\lambda = 1.5, \gamma = 0.5$, and error variance $\sigma_d^2 = 0.15^2$. The reference observation profile $\mathbf{d}$ and wavelet $\mathbf{w}$ are presented in Fig. 3. The main objective applying the BG model is inversion of $\mathbf{d}$ into $\mathbf{r}$, whereas we are performing inversion from $\mathbf{d}$ directly into $\boldsymbol{\pi}$ when also $\mathbf{P}$ is unknown. We estimate the transition matrix $\mathbf{P}$ by the three proposed inference techniques, setting all other parameters to their reference values.

The aEM algorithm (see Algorithm 1) is the least computer demanding algorithm, and an approximation order of $k = 6$ is hence chosen. The algorithm is run with different initial values, for which all runs seem to have converged after about 25 EM-steps (see Fig. 4). Most runs in Fig. 4 do not display a monotone increase in the approximate log-likelihood value, however, since the approximate likelihood model also depends on the transition matrix. The final transition matrix estimate is

$$\widehat{\mathbf{P}}_{\mathrm{aEM}} = \begin{pmatrix} 0.816 & 0.184 \\ \sim 1 & 10^{-5} \end{pmatrix}$$

with estimated stationary pdf $\hat{\mathbf{p}}^s_{\mathrm{aEM}} = (0.85, 0.15)$ and expected waiting times $\hat{\mathbf{u}}_{\mathrm{aEM}} = (5.43, 1.00)$. The black class of high reflector points is hence slightly overrepresented in the estimate compared with the reference transition matrix in (18). The estimates and the 90% confidence intervals of each element in $\mathbf{P}$ are given in Fig. 5(a), the latter computed by the Hessian matrix and truncated at [0,1]. Notice that both the reference and empirical transition matrix values are within the confidence bounds. Ten thousand realizations of the categorical field are simulated by Algorithm A2 in the Appendix, i.e., $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{10\,000} \sim p(\boldsymbol{\pi}|\mathbf{d};\widehat{\mathbf{P}}_{\mathrm{aEM}})$, with a sixth-order approximate posterior model set as proposal distribution. The marginal posterior probabilities $p(\pi_t|\mathbf{d};\widehat{\mathbf{P}}_{\mathrm{aEM}})$, $t = 1,\ldots,T$ and the marginal maximum a posterior (MMAP) prediction computed from the realizations

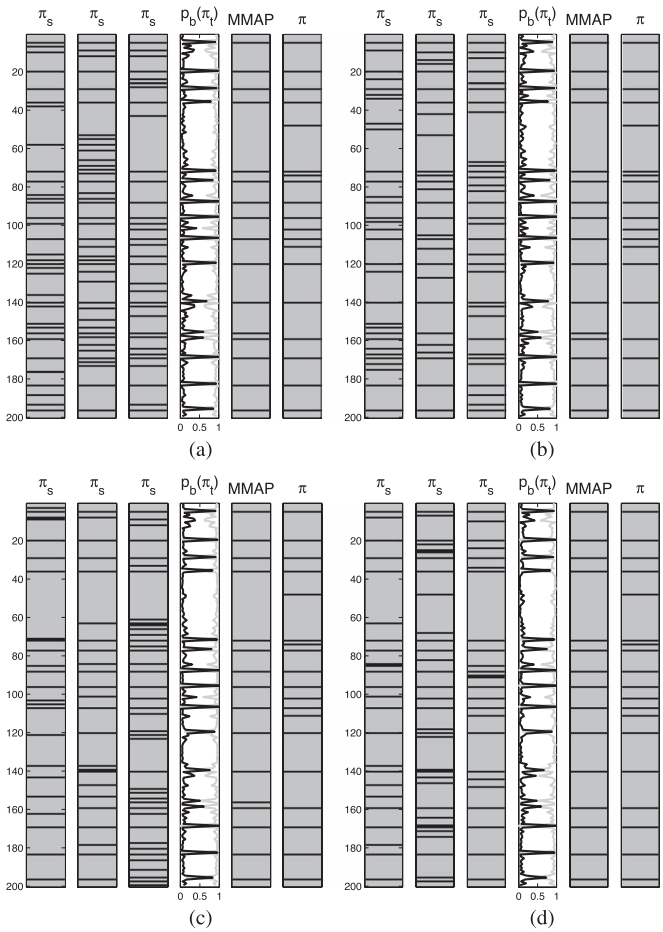

Fig. 6. Prediction of the categorical profile compared (a)–(c) across the inference methods and (d) with the BG model, displaying three random realizations from the posterior distribution $\boldsymbol{\pi}_s$, locationwise marginal posterior probabilities $p_b(\pi_t) = p(\pi_t|\mathbf{d})$ $t = 1,\ldots,T$, and the MMAP predictions computed from the realizations and the reference categorical profile. (a) aEM. (b) MCEM. (c) MCMC. (d) BG: MCMC.

are presented in Fig. 6(a). Notice that the MMAP prediction corresponds very reliably to the reference profile. The realizations display variability in the simulations, in which too many layer transitions are simulated due to the overestimated transition probability $\widehat{P}_{12}$ from gray to black. Some of the high reflector points with smaller deviations in the response are not identified as expected.

For inference by the MCEM algorithm, each EM-step in Algorithm 2 is based on thousand MC samples of the

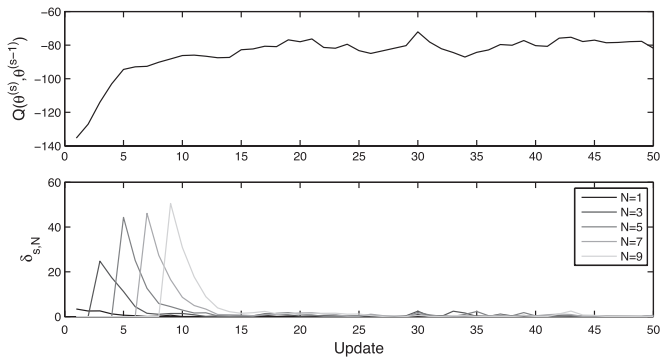

Fig. 7. Trace plots for MCEM, with the top plot displaying the approximate Q functions and the lower plot the values of $\hat{\delta}_{t,N}$ for $N = 1, 3, 5, 7, 9$.

categorical field, which are generated from the posterior model under the current parameter value, $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{1000} \sim p(\boldsymbol{\pi}|\mathbf{d}; \mathbf{P}^*)$, with a fourth-order approximate posterior model set as proposal distribution. The optimization in the M-step is performed by the MATLAB function FMINSEARCH. After about 50 EM-steps, the algorithm seems to have converged (see Fig. 7). The approximate Q functions from (13) increase until convergence as expected, and the values of $\hat{\delta}_{t,N}$ defined in (15) also converge toward 0 as desired, when the differences between the approximate Q functions across iterations diminish. The final transition matrix estimate is

$$\widehat{\mathbf{P}}_{\mathrm{MCEM}} = \begin{pmatrix} 0.808 & 0.192 \\ 1 & 0 \end{pmatrix}$$

with estimated stationary pdf $\hat{\mathbf{p}}^s_{\mathrm{MCEM}} = (0.84, 0.16)$ and expected waiting times $\hat{\mathbf{u}}_{\mathrm{MCEM}} = (5.20, 1)$. The estimates and the 90% confidence intervals on the elements of $\mathbf{P}$ again capture the reference and empirical values [see Fig. 5(b)]. As for aEM, 10 000 realizations of the categorical field are simulated, i.e., $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{10\,000} \sim p(\boldsymbol{\pi}|\mathbf{d}; \widehat{\mathbf{P}}_{\mathrm{MCEM}})$, with a sixth-order approximate posterior model as proposal distribution. The marginal posterior probabilities and the MMAP prediction are presented in Fig. 6(b). Notice again that the MMAP prediction corresponds very reliably to the reference profile, being identical to the aEM mode.

For inference by the MCMC algorithm, the algorithm is run iteratively with 10 000 iterations with a fourth-order approximate posterior model set as proposal distribution as for the MCEM approach. The Dirichlet parameter prior model in (16) is set with hyperparameter $\eta = 1$, which represents a noninformative prior. The MCMC algorithm seems to have converged after about 500 iterations (see Fig. 8), and the burn-in value is set to 2000. The estimated transition matrix is

$$\widehat{\mathbf{P}}_{\mathrm{MCMC}} = \begin{pmatrix} 0.825 & 0.175 \\ 0.983 & 0.017 \end{pmatrix}$$

with estimated stationary pdf $\hat{\mathbf{p}}^s_{\mathrm{MCMC}} = (0.84, 0.16)$ and expected waiting times $\hat{\mathbf{u}}_{\mathrm{MCMC}} = (5.71, 1.02)$. The transition matrix estimate is set as the ensemble mode rather than the expectation, as the expected value will always be pushed further from the borders, i.e., can never be the values 1 or 0, which we

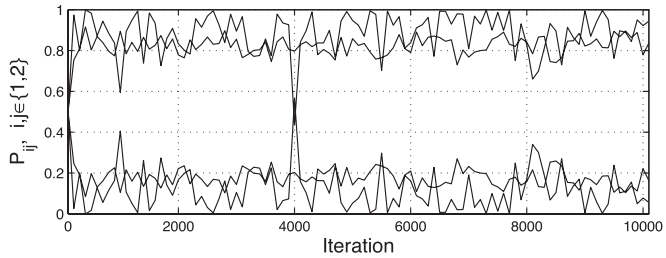

Fig. 8. Trace plots of the four sampled elements of $\mathbf{P}$ for estimation by MCMC.

should be able to capture. The estimates and the 90% confidence intervals for the elements of $\mathbf{P}$, computed from the 0.9 range of the simulated values, are given in Fig. 5(c). The marginal posterior probabilities and MMAP prediction of $\boldsymbol{\pi}$ computed from the ensemble of realizations are presented in Fig. 6(c). The MMAP prediction corresponds very reliably to the reference profile again; in fact, the MMAP predictions for the three approaches are identical. Notice also that consecutive high reflector points are present in some of the realizations, as the MCMC estimate was the only estimate with $\hat{P}_{22}$ clearly different from zero.

For the BG model, we follow Rosec *et al.* [18], doing MCMC with 10 000 iterations, assigning a $Beta(1, 1)$ prior parameter model to λ for which $p(\pi_t = \text{black}) = \lambda$; hence, the prior marginal distribution is $p(\pi_t) = (1 - \lambda, \lambda)$. The estimated value is $\hat{\lambda} = 0.16$ with prior pdf $\hat{p}(\pi_t) = (0.84, 0.16)$ corresponding to $\hat{\mathbf{p}}^s$ for the convolved HMM. The marginal posterior probabilities and MMAP prediction of $\boldsymbol{\pi}$ are presented in Fig. 6(d). These inversion results appear as very similar to the ones from the convolved HMM. Similar to the MCMC inversion from the convolved HMM, some realizations have too thick black layers. Hence, including spatial vertical correlation through the Markov model assumption does not affect the predictions much for this problem; however, it enables us to set the probability for consecutive high reflector points to 0 if wanted.

V. Empirical Test Study

A convolved HMM of length $T = 200$ with $L = 3$ possible classes $\Omega_\pi = \{\text{white}, \text{gray}, \text{black}\}$ is considered. Three reference prior model transition matrices are defined, termed cases 1 to 3. The reference categorical profiles for the three cases simulated from the respective prior models are presented in Fig. 9. The reference and empirical matrices are

$$\mathbf{P}_1 = \begin{pmatrix} 0.50 & 0.50 & 0 \\ 0.25 & 0.50 & 0.25 \\ 0 & 0.50 & 0.50 \end{pmatrix} \quad \mathbf{P}_{1,e} = \begin{pmatrix} 0.59 & 0.41 & 0 \\ 0.22 & 0.53 & 0.25 \\ 0 & 0.55 & 0.45 \end{pmatrix} \tag{19}$$

$$\mathbf{P}_2 = \begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix} \quad \mathbf{P}_{2,e} = \begin{pmatrix} 0.94 & 0.04 & 0.01 \\ 0.02 & 0.91 & 0.07 \\ 0.06 & 0.09 & 0.85 \end{pmatrix} \tag{20}$$

$$\mathbf{P}_3 = \begin{pmatrix} 0.20 & 0 & 0.80 \\ 0.10 & 0.20 & 0.70 \\ 0.15 & 0.15 & 0.70 \end{pmatrix} \quad \mathbf{P}_{3,e} = \begin{pmatrix} 0.10 & 0 & 0.90 \\ 0.09 & 0.13 & 0.78 \\ 0.17 & 0.14 & 0.69 \end{pmatrix} \tag{21}$$

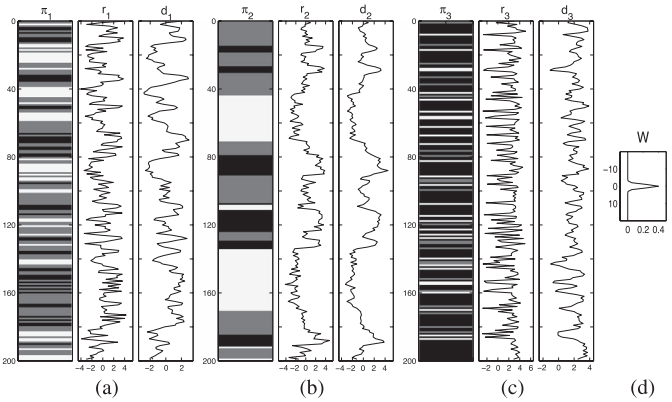

Fig. 9. (a)–(c) Reference categorical profiles, response profiles, and observation profiles for the empirical test study cases and (d) reference convolution kernel. (a) Case 1. (b) Case 2. (c) Case 3. (d) Kernel.

with corresponding reference and empirical stationary pdfs

$$\mathbf{p}_1^s = (0.25, 0.50, 0.25) \quad \mathbf{p}_{1,e}^s = (0.27, 0.51, 0.22)$$
$$\mathbf{p}_2^s = (0.33, 0.34, 0.33) \quad \mathbf{p}_{2,e}^s = (0.39, 0.39, 0.22)$$
$$\mathbf{p}_3^s = (0.15, 0.13, 0.72) \quad \mathbf{p}_{3,e}^s = (0.15, 0.12, 0.73)$$

and reference and empirical expected thicknesses

$$\mathbf{u}_1 = (2.00, 2.00, 2.00) \quad \mathbf{u}_{1,e} = (2.43, 2.15, 1.81)$$
$$\mathbf{u}_2 = (10.0, 10.0, 10.0) \quad \mathbf{u}_{2,e} = (17.0, 10.6, 6.6)$$
$$\mathbf{u}_3 = (1.25, 1.25, 3.33) \quad \mathbf{u}_{3,e} = (1.11, 1.15, 3.24).$$

For case 1, the probability of staying in the current class is equal to the probability of switching class; hence, there will be rapid layer transitions, as shown in the waiting times [see Fig. 9(a)]. We also enforce classes black and white never to be neighbors in the prior model for case 1. For case 2, the probability of staying in the current class is much larger than for switching class, hence thicker layers [see Fig. 9(b)]. For case 3, the transition probabilities into the black class are the largest; hence, the black class will dominate with some thinner gray and white layers [see Fig. 9(c)]. We enforce the white class never to be on top of gray in the prior model in case 3. As for the previous toy example, the simulated reference profiles are of finite length, and we thus expect the transition probability estimates to be closer to the empirical values than the reference values. The response likelihood model parameters are set to $\mu_{r|\pi} \in \{-2, 0, 3\}$ and $\sigma_{r|\pi} = \{0.7, 0.7, 0.7\}$, corresponding to the three classes. The observation likelihood parameters are set to a Gaussian $N(0, 1)$ convolution kernel and error variance $\sigma_d^2 = 0.3^2$ [see Fig. 9(d)]. The response profiles $\mathbf{r}$ simulated from $\boldsymbol{\pi}$ and the observation profiles $\mathbf{d}$ simulated from the corresponding $\mathbf{r}$ are presented in Fig. 9 for the three cases.

We assess the transition matrix $\mathbf{P}$ for each case by the three inference methods aEM, MCEM, and MCMC with subscripts a, b, c in the estimates, respectively. The approximation orders are set to $k = 6$, $k = 4$, and $k = 4$, respectively, as for the previous toy example. Confidence intervals of the nine elements of the transition matrix and the corresponding marginal MMAP predictions of the categorical states are compared.

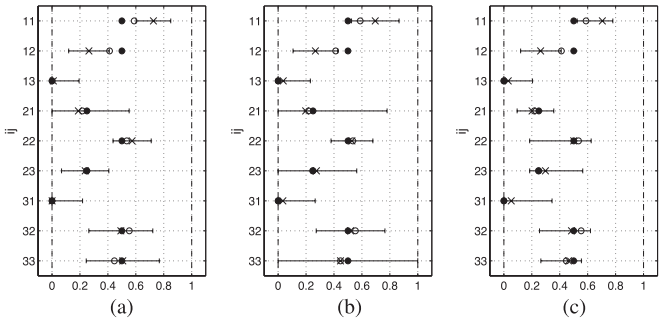

Fig. 10. Case 1: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$. The estimates are denoted as crosses, reference values as black dots, and empirical values as white dots. (a) aEM. (b) MCEM. (c) MCMC.

A. Case 1

For case 1, with reference transition matrix $\mathbf{P}_1$ in (19), the estimated transition matrices are

$$\widehat{\mathbf{P}}_{1a} = \begin{pmatrix} 0.73 & 0.26 & 0.01 \\ 0.19 & 0.57 & 0.24 \\ 10^{-8} & 0.49 & 0.51 \end{pmatrix} \quad \widehat{\mathbf{P}}_{1b} = \begin{pmatrix} 0.69 & 0.27 & 0.04 \\ 0.20 & 0.53 & 0.27 \\ 0.03 & 0.52 & 0.45 \end{pmatrix}$$
$$\widehat{\mathbf{P}}_{1c} = \begin{pmatrix} 0.70 & 0.26 & 0.04 \\ 0.18 & 0.54 & 0.28 \\ 0.06 & 0.49 & 0.45 \end{pmatrix}.$$

All three methods have small estimated probabilities P_{13} and P_{31} as desired, with the aEM estimates clearly the lowest. The diagonal probability estimates are reliable, being close to 0.5, although slightly overestimated for the white class being closer to the empirical values as expected. The estimates and the 90% confidence intervals are presented in Fig. 10, with all intervals capturing the reference values, except for the nonzero transitions from the white class P_{11} and P_{12}; however, all empirical values are captured. Notice in particular that the reference zero probabilities P_{13} and P_{31} are captured. The MCEM algorithm has the widest intervals, hence with largest parameter estimate uncertainty. Correspondingly, the MCMC algorithm provides the narrowest intervals, but still captures the reference and empirical parameter values. The corresponding estimated stationary pdfs and expected thicknesses are

$$\hat{\mathbf{p}}_{1a}^s = (0.31, 0.46, 0.23) \quad \hat{\mathbf{u}}_{1a} = (3.65, 2.34, 2.03)$$
$$\hat{\mathbf{p}}_{1b}^s = (0.32, 0.44, 0.24) \quad \hat{\mathbf{u}}_{1b} = (3.28, 2.12, 1.81)$$
$$\hat{\mathbf{p}}_{1c}^s = (0.35, 0.39, 0.26) \quad \hat{\mathbf{u}}_{1c} = (3.33, 2.16, 1.82)$$

and these estimate the reference values reliably. The white class is estimated with somewhat high expected thickness because of the overestimated probability P_{11}.

The marginal posterior probabilities of the categorical profile $p(\pi_t|\mathbf{d})$ for $t = 1, \ldots, T$ and the MMAP predictions with the reference profile are given in Fig. 11. The predictions are very similar for the three inference algorithms, resembling the reference profile well, however, with somewhat too thick white layers due to the overestimated transition matrix value. Notice that predictions shall contain somewhat fewer transitions than the true profile due to regression toward larger classes. A few noneligible transitions between white and black are present in the MCMC MMAP, due to overestimation of P_{13} and P_{31}.

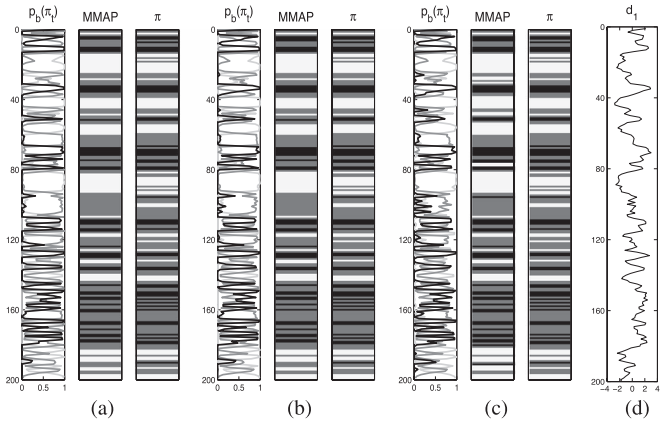

Fig. 11. Case 1: Locationwise marginal posterior probabilities $p(\pi_t|\mathbf{d})$ and the MMAP predictions computed from the realizations and the reference categorical profile and observations. (a) aEM. (b) MCEM. (c) MCMC. (d) Obs.

B. Case 2

For case 2, with reference transition matrix $\mathbf{P}_2$ in (20), the estimated transition matrices are

$$\widehat{\mathbf{P}}_{2a} = \begin{pmatrix} 0.97 & 0.03 & 0 \\ 0.04 & 0.88 & 0.08 \\ 10^{-5} & 0.17 & 0.83 \end{pmatrix} \quad \widehat{\mathbf{P}}_{2b} = \begin{pmatrix} 0.96 & 0.04 & 0 \\ 0.03 & 0.88 & 0.09 \\ 0.01 & 0.20 & 0.79 \end{pmatrix}$$

$$\widehat{\mathbf{P}}_{2c} = \begin{pmatrix} 0.96 & 0.03 & 0.01 \\ 0.03 & 0.87 & 0.10 \\ 0.01 & 0.20 & 0.79 \end{pmatrix}.$$

The estimated diagonal probabilities are reliably large, but slightly underestimated for the black class P_{33} and compensated overestimated for the white class P_{11}, again resembling the empirical values more than the reference values. The transition probabilities from the white to the black class and vice versa are consistently underestimated, with the aEM and MCEM algorithms even estimating zero probabilities for P_{13}. The estimates and the 90% confidence intervals are presented in Fig. 12, again with most intervals capturing the reference and empirical values. Notice the narrow intervals for probabilities close to zero and one. The intervals for the MCEM algorithm appear as overestimated. The corresponding estimated stationary pdfs and expected thicknesses are

$$\hat{\mathbf{p}}^s_{2a} = (0.42, 0.39, 0.19) \quad \hat{\mathbf{u}}_{2a} = (30.9, 8.38, 5.80)$$
$$\hat{\mathbf{p}}^s_{2b} = (0.38, 0.44, 0.18) \quad \hat{\mathbf{u}}_{2b} = (25.5, 8.45, 4.74)$$
$$\hat{\mathbf{p}}^s_{2c} = (0.39, 0.39, 0.22) \quad \hat{\mathbf{u}}_{2c} = (25.5, 7.84, 4.74).$$

The estimated stationary pdfs reproduce the reference values reliably, whereas the expected thicknesses for white are somewhat large and for black somewhat low. The expected thickness is very sensitive to overestimated diagonal probabilities; hence, the estimate for P_{11} of 0.97 for aEM compared with the reference value 0.90 increases the expected thickness to above 30 compared with the reference expected thickness of about 10.

The marginal posterior probabilities of the categorical profile and the MMAP predictions are given in Fig. 13, and they are again almost identical for the three inference algorithms.

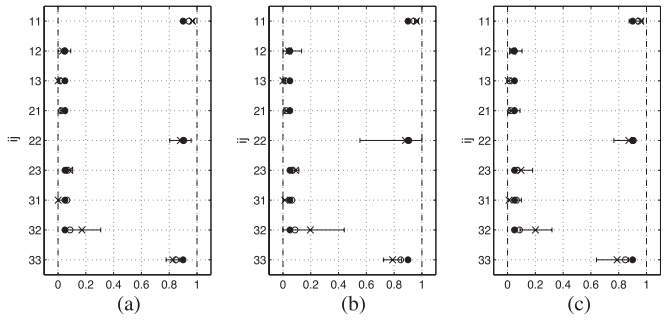

Fig. 12. Case 2: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$. (a) aEM. (b) MCEM. (c) MCMC.

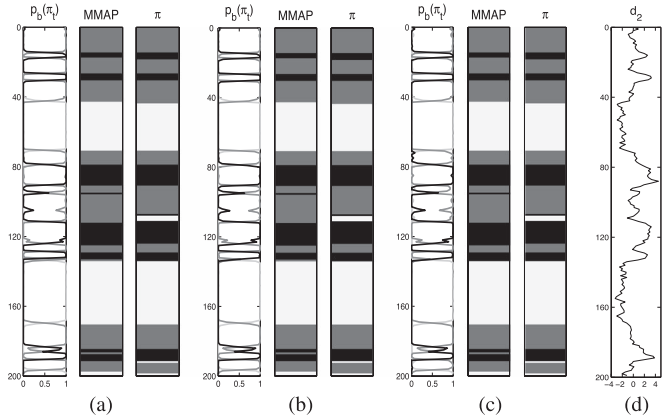

Fig. 13. Case 2: Locationwise marginal posterior probabilities, MMAP predictions, the reference categorical profile, and the observations. (a) aEM. (b) MCEM. (c) MCMC. (d) Obs.

The MMAP predictions resemble the reference profile well, although all of them are unable to recognize the thinner white middle layer.

C. Case 3

For case 3, with reference transition matrix $\mathbf{P}_3$ in (21), the estimated transition matrices are

$$\widehat{\mathbf{P}}_{3a} = \begin{pmatrix} 0.18 & 0.05 & 0.77 \\ 10^{-17} & 0.28 & 0.72 \\ 0.13 & 0.12 & 0.75 \end{pmatrix}$$

$$\widehat{\mathbf{P}}_{3b} = \begin{pmatrix} 0.16 & 10^{-12} & 0.84 \\ 10^{-7} & 10^{-7} & \sim 1 \\ 0.17 & 0.09 & 0.74 \end{pmatrix}$$

$$\widehat{\mathbf{P}}_{3c} = \begin{pmatrix} 0.14 & 0.02 & 0.84 \\ 0.01 & 0.01 & 0.98 \\ 0.15 & 0.11 & 0.74 \end{pmatrix}.$$

All estimated transition matrices have high probability for transitions into the black dominant class as for the reference values. The MCEM estimate is the only one reproducing the zero probability from white to gray in P_{12}. The probabilities from gray to white and from gray to gray in P_{21} and P_{22} are consistently underestimated, except for P_{22} for aEM. The estimates and the 90% confidence intervals are presented in Fig. 14, with all intervals capturing the reference values. The interval widths are larger than for the two previous cases (see, in particular, the

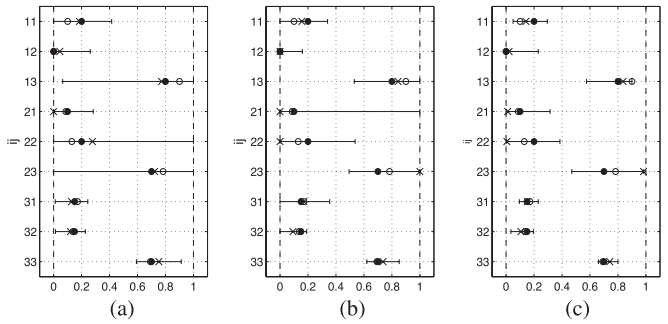

Fig. 14. Case 3: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$. (a) aEM. (b) MCEM. (c) MCMC.

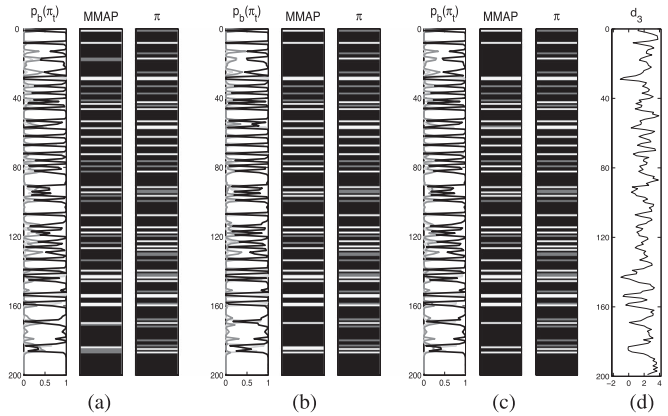

Fig. 15. Case 3: Locationwise marginal posterior probabilities, MMAP predictions, the reference categorical profile, and the observations. (a) aEM. (b) MCEM. (c) MCMC. (d) Obs.

intervals P_{22} and P_{23} for aEM and P_{21} for MCEM that cover the full interval [0,1]). The MCMC algorithm seems to appear with more realistic confidence intervals. The corresponding estimated stationary pdfs and expected thicknesses are

$$\hat{\mathbf{p}}^s_{3a} = (0.12, 0.13, 0.75) \quad \hat{\mathbf{u}}_{3a} = (1.23, 1.39, 4.04)$$
$$\hat{\mathbf{p}}^s_{3b} = (0.16, 0.07, 0.77) \quad \hat{\mathbf{u}}_{3b} = (1.19, 1.00, 3.79)$$
$$\hat{\mathbf{p}}^s_{3c} = (0.15, 0.11, 0.74) \quad \hat{\mathbf{u}}_{3c} = (1.17, 1.01, 3.86)$$

which reproduce the reference values reliably. The gray class thickness will be almost one layered for the MCEM and MCMC estimates, corresponding to the underestimation of P_{22}.

The marginal posterior probabilities on the categorical profile and the MMAP predictions are given in Fig. 15. All predictions reproduce the black class as the dominant class with thinner gray and white layers.

D. Discussion

The tests are run on a regular laptop computer, and for one case, the typical runtime for the aEM algorithm with approximation order $k = 4$ is about 1 min, whereas the MCMC algorithm for the same approximation order requires more than 1 h. The estimates of the transition matrix $\mathbf{P}$ appear as similar for the three algorithms: aEM, MCEM, and MCMC. All estimates are reliable, although estimation errors for probabilities close to 0 and 1 have large impact on the inversion results. The confidence intervals for the MCMC algorithm are consistently narrower, while still capturing the reference and empirical values, and hence, they appear as most reliable. The inversion results for $\boldsymbol{\pi}$ based on the three algorithms are difficult to distinguish, and they all appear as reliable, capturing the main characteristics of the reference profile. The computational demands for the three algorithms vary dramatically, however. If focus is on point estimates for $\mathbf{P}$ and MMAP predictions for $\boldsymbol{\pi}$, the aEM algorithm is preferable since the estimates and predictions are as reliable as for the other algorithms, whereas the computational demands are clearly favorable. The MCMC algorithm can provide more reliable estimation intervals and precision measures on the cost of considerably larger computational demands. There appear to be few arguments for using the MCEM algorithm.

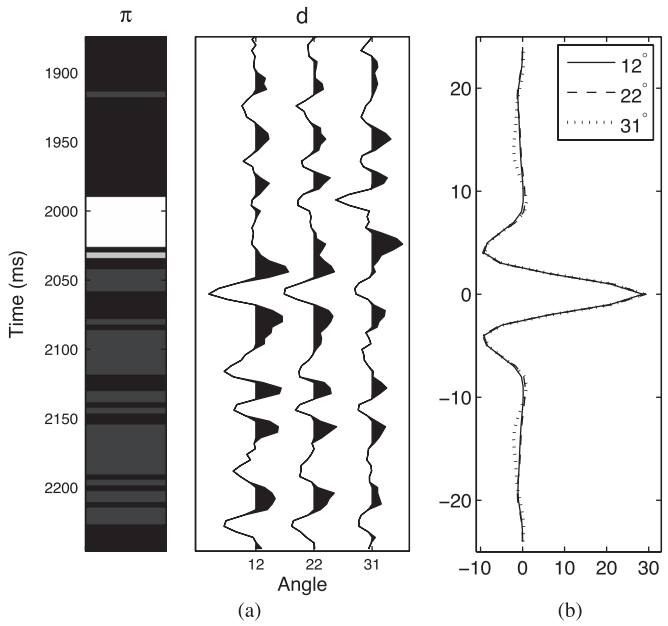

Fig. 16. Seismic case study. (a) Reference LF profile and prestack seismic traces from the angles {12°, 22°, 31°}. (b) Reference wavelets for the three angles.

VI. Case Study

Data from a well in a reservoir offshore Norway are considered (see [6]), with the four LF classes {gas-filled sandstone, oil-filled sandstone, water-filled sandstone, shale} represented by the colors white, light gray, dark gray, and black, respectively. The true LF profile is given in Fig. 16, together with synthetic seismic AVO traces from three angles 12°, 22°, and 31°, following Rimstad and Omre [6]. The focus is on categorical deconvolution of the seismic observations into the LF states when, in addition, the Markov model transition matrix, defining the prior transition probabilities between the LF classes downward along the profile, is unknown.

The seismic observations are captured as a convolution of the contrasts in the elastic properties and a seismic wavelet $\mathbf{w}$ corresponding to the convolution kernel. The reference wavelets for the three angles as given by the data owner are presented in Fig. 16(b). The elastic properties, P-wave velocity v_p, S-wave velocity v_s, and density ρ are assumed to be elementwise conditional independent given the underlying LF states, whose logarithms represent the hidden middle level $\mathbf{r}$. The Gaussian

response means and variances for the elastic properties $\log(v_p), \log(v_s), \log(\rho)$ are by row

$$\mu_{r|\pi} \in \begin{cases} \{8.11, 8.12, 8.16, 7.88\} \\ \{7.61, 7.60, 7.58, 7.08\} \\ \{7.66, 7.69, 7.73, 7.77\} \end{cases} \qquad \sigma_{r|\pi} = \begin{cases} 0.0860 \\ 0.1483 \\ 0.0608 \end{cases} \tag{22}$$

assuming equal variances for the four LF classes. The convolution relation can be reliably modeled by $\mathbf{d} = \mathbf{WADr} + \mathbf{e}$, where $\mathbf{W}$ is the convolution matrix with centralized rows $\mathbf{w}$, $\mathbf{A}$ is a matrix of angle-dependent weak contrast Aki–Richards coefficients [20], $\mathbf{D}$ is a central differential matrix calculating the contrasts, and $\mathbf{e}$ are independent Gaussian errors (see [21] for more details).

For the four LF classes, there are gravity effects present, i.e., gas can never occur directly below oil or water, and oil can never occur directly below water. There is naturally no true transition matrix, and the empirical downward transition matrix estimated from the LF profile is

$$\mathbf{P}_e = \begin{array}{c} \\ g \\ o \\ w \\ s \end{array} \begin{array}{c} \begin{array}{cccc} g & o & w & s \end{array} \\ \begin{pmatrix} 0.89 & 0 & 0 & 0.11 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0.69 & 0.31 \\ 0.01 & 0.01 & 0.11 & 0.87 \end{pmatrix} \end{array} \tag{23}$$

where g is gas-filled sandstone, o is oil-filled sandstone, w is water-filled sandstone, and s is shale. The corresponding reference stationary pdf is $\mathbf{p}_e^s = (0.07, 0.01, 0.23, 0.69)$, and expected thicknesses $\mathbf{u}_e = (9.00, 1, 3.20, 7.92)$. Notice that the gravity effects are honored by P_{21}, P_{31}, and P_{32} being zero; however, the probabilities P_{12}, P_{13}, P_{22}, and P_{23} are also zero because of lacking corresponding transitions in the reference LF profile. In particular, for the two hydrocarbon classes gas-sand and oil-sand, there is only one layer present; hence, the empirical transition probabilities do not capture all possible transitions. We can foresee big challenges in inference of the transition probabilities involving these classes due to lack of repeatability, particularly for oil-filled sandstone, which has thickness one in the reference profile.

The transition matrix is estimated from the seismic observations by the aEM algorithm only, which is a preferred choice for reliable point estimates. The approximation order is set to $k = 6$. As zero probabilities are reproduced when updating in the Baum–Welch algorithm, we enforce the probabilities P_{21}, P_{31}, and P_{32} to be zero by setting them to zero in the initial parameter input to the algorithm to honor the gravity effects. The final transition matrix estimate is

$$\widehat{\mathbf{P}} = \begin{array}{c} \\ g \\ o \\ w \\ s \end{array} \begin{array}{c} \begin{array}{cccc} g & o & w & s \end{array} \\ \begin{pmatrix} 0.84 & \sim 10^{-5} & \sim 10^{-11} & 0.16 \\ 0 & 0.97 & \sim 10^{-5} & 0.03 \\ 0 & 0 & 0.78 & 0.22 \\ 0.02 & 0.01 & 0.08 & 0.89 \end{pmatrix} \end{array} \tag{24}$$

with corresponding estimated stationary pdf $\hat{\mathbf{p}}^s = (0.07, 0.19, 0.20, 0.54)$ and expected thicknesses $\hat{\mathbf{u}} = (25.5, 165.5, 18.6, 36.8)$. The estimated transition matrix resembles the empirical transition matrix in (23) reliably for transitions from the classes gas-filled sandstone, water-filled sandstone, and shale. For the transitions from oil-filled sandstone, which appear only once and with thickness one, the values differ significantly. These results indicate inference problems for rarely occurring very thin classes. The estimates and the 90% confidence intervals are presented in Fig. 17, and most intervals capture the empirical values as desired. The poorly estimated transitions out of oil-filled sand, i.e., P_{22} and P_{24}, are not indicated by wide intervals, and this is alarming.

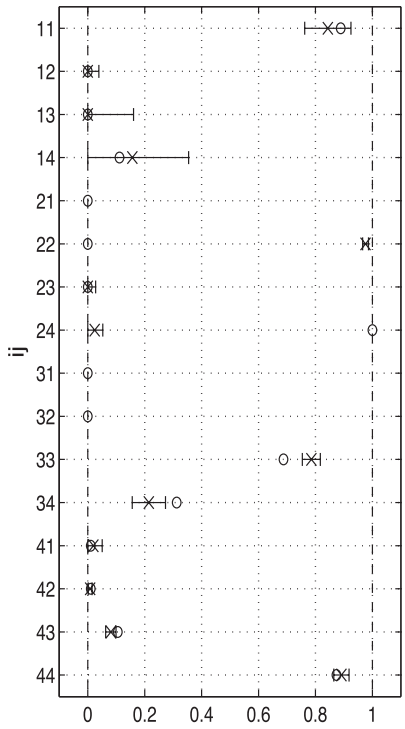

Fig. 17. Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3, 4\}$ representing the LF classes {gas-filled sandstone, oil-filled sandstone, water-filled sandstone, shale}, respectively. The elements $\{P_{21}, P_{31}, P_{32}\}$ are set to 0.

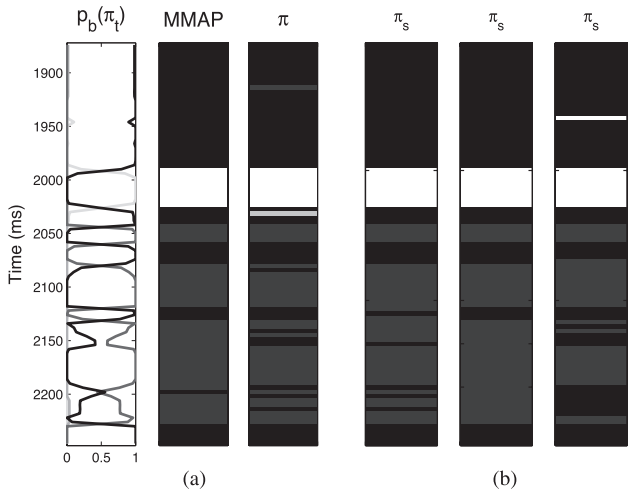

Fig. 18. (a) Locationwise marginal posterior probabilities $p(\pi_t|\mathbf{d})$, the MMAP prediction, and the reference categorical profile and (b) three realizations simulated from the posterior model $\boldsymbol{\pi}_s \sim p(\boldsymbol{\pi}|\mathbf{d}; \widehat{\mathbf{P}})$ in the case study.

The marginal posterior probabilities of the LF states throughout the profile and the marginal MMAP prediction together with three realizations from the posterior model with plug-in estimates are given in Fig. 18. The MMAP prediction is less

heterogeneous than the reference profile due to regression toward water-filled sand, whereas the realizations expose correct heterogeneity. Note that oil-filled sand is not predicted, although there is some probability for it to occur through P_{42}.

To summarize, even for complex likelihood models with response to contrasts in the LF profile, we are able to estimate the transition matrix $\mathbf{P}$ reliably by the aEM algorithm, as long as the classes appear with repeatability or have some thickness in the LF profile. As indicated by the simulation test, the reliability of the confidence intervals assessed by the aEM algorithm is much smaller. The inversion of seismic AVO data into the LF profile can be reliably made. In order to ensure larger repeatability for thin rare classes, one may evaluate several seismic profiles from the reservoir jointly.

VII. Conclusion

The convolved HMM has been defined and discussed. An approximate posterior model for the categorical bottom-level variables given the convolved observations are exactly assessable by a recursive generalized FB algorithm. The posterior model is assessable through simulation-based inference using an MCMC independence sampler with the approximate model as proposal generator. Hence, Bayesian inversion of the convolved HMM is available when all model parameters are specified.

We study Bayesian inversion of the convolved HMM when also the transition matrix in the prior Markov chain model is considered unknown. Three alternative algorithms for inference and inversion of the matrix and the categorical variables, respectively, are defined, discussed, and evaluated. The algorithms, namely, aEM, MCEM, and MCMC, have dramatically different computational demands, in increasing order. The aEM algorithm is based on a generalization of the Baum–Welch algorithm and provides an approximate ML estimate of the transition matrix. The MCEM algorithm contains an MC simulation step and provides an ML estimate up to MC precision. The MCMC algorithm is based on Bayesian inference of the transition matrix and provides asymptotically consistent estimates of the model parameter posterior model. The aEM and MCEM algorithms do provide inversions of the categorical variables based on prior Markov chain models with plug-in estimates of the transition matrix. The MCMC algorithm includes joint assessment of the transition matrix and the categorical variables. Hence, the inversion is made randomized over the parameter posterior model.

The conclusions based on discussions and empirical studies of the algorithms are as follows.

- If focus is on point estimates of the transition matrix and MMAP predictions of the categorical variables, the aEM algorithm is recommended. The estimates and predictions are as good as for the other algorithms while the computational demands are clearly favorable.
- If focus is on uncertainty quantification in addition to point estimates, the MCMC algorithm is recommended. The precision intervals appear as much more reliable than for the two other algorithms, on the expense of much larger computational demands.
- There appears to be few, if any, arguments for using the MCEM algorithm.
- Inference of transition probabilities for thin and very rarely occurring classes can be unreliable.

A natural extension of the study being presented is to include inference of the parameters in the response likelihood model also. The original Baum–Welch algorithm for standard HMM does actually provide estimates for these parameters. The potential of the algorithms in seismic inversion needs to be documented in future publications of full-scale case studies.

Appendix
FB Algorithm

Define backward probabilities as conditioned on the full set of observations, e.g., $p_b(\boldsymbol{\pi}_t^{(k-1)}) = \hat{p}^{(k)}(\boldsymbol{\pi}_t^{(k-1)}|\mathbf{d})$ and $p_b(\pi_t|\boldsymbol{\pi}_{t+k-1}^{(k-1)}) = \hat{p}^{(k)}(\pi_t|\boldsymbol{\pi}_{t+k-1}^{(k-1)}, \mathbf{d})$. The FB algorithm presented follows Reeves and Pettitt [10] and has runtime $O(R_{\mathrm{FB},k})$, with $R_{\mathrm{FB},k} = (T-k+2)L^k$.

Algorithm A1 FB algorithm

Forward:

• Initiate for $t = k$:

$$z_k\left(\boldsymbol{\pi}_k^{(k-1)}\right) = \sum_{\pi_1}\prod_{i=1}^{k} p(\pi_i|\pi_{i-1}) \times \hat{p}^{(i)}\left(\mathbf{d}_i^{(i)}|\boldsymbol{\pi}_i^{(i)}\right)^{1/k}$$

• Iterate for $t = k+1, \ldots, T-1$:

$$z_t\left(\boldsymbol{\pi}_t^{(k-1)}\right) = p(\pi_t|\pi_{t-1}) \times \sum_{\pi_{t-k+1}} \hat{p}^{(k)}\left(\mathbf{d}_t^{(k)}|\boldsymbol{\pi}_t^{(k)}\right)^{1/k} z_{t-1}\left(\boldsymbol{\pi}_{t-1}^{(k-1)}\right)$$

• Last element, $t = T$:

$$z_T\left(\boldsymbol{\pi}_T^{(k-1)}\right) = p(\pi_T|\pi_{T-1}) \times \sum_{\pi_{T-k+1}}\prod_{i=1}^{k} \hat{p}^{(i)}\left(\mathbf{d}_T^{(i)}|\boldsymbol{\pi}_T^{(i)}\right)^{1/k} \times z_{T-1}\left(\boldsymbol{\pi}_{T-1}^{(k-1)}\right)$$

• Normalizing constant:

$$C_d^{(k)} = \left[\sum_{\pi_{T-k+2}} \cdots \sum_{\pi_T} z_T\left(\boldsymbol{\pi}_T^{(k-1)}\right)\right]^{-1}$$

Backward:

• Initiate:

$$p_b\left(\boldsymbol{\pi}_T^{(k-1)}\right) = C_d^{(k)} z_T\left(\boldsymbol{\pi}_T^{(k-1)}\right)$$

$$p_b\left(\pi_{T-k+1}\middle|\boldsymbol{\pi}_T^{(k-1)}\right) = p(\pi_T|\pi_{T-1}) \times \prod_{i=1}^{k} \hat{p}^{(i)}\left(\mathbf{d}_T^{(i)}|\boldsymbol{\pi}_T^{(i)}\right)^{1/k} \times z_{T-1}\left(\boldsymbol{\pi}_{T-1}^{(k-1)}\right) \Big/ z_T\left(\boldsymbol{\pi}_T^{(k-1)}\right)$$

$$p_b\left(\boldsymbol{\pi}_{T-1}^{(k-1)}\right) = \sum_{\pi_T=1}^{L} p_b\left(\pi_{T-k+1}\middle|\boldsymbol{\pi}_T^{(k-1)}\right) \times p_b\left(\boldsymbol{\pi}_T^{(k-1)}\right)$$

• Iterate for $t = T-k, \ldots, 2$:

$$p_b\left(\pi_t \middle| \boldsymbol{\pi}_{t+k-1}^{(k-1)}\right) = p(\pi_{t+k-1}|\pi_{t+k-2})\hat{p}^{(k)}\left(\mathbf{d}_{t+k-1}^{(k)}|\boldsymbol{\pi}_{t+k-1}^{(k)}\right)^{1/k} \times z_{t+k-2}\left(\boldsymbol{\pi}_{t+k-2}^{(k-1)}\right) \Big/ z_{t+k-1}\left(\boldsymbol{\pi}_{t+k-1}^{(k-1)}\right)$$

$$p_b\left(\boldsymbol{\pi}_{t+k-2}^{(k-1)}\right) = \sum_{\pi_{t+k-1}=1}^{L} p_b\left(\pi_t \middle| \boldsymbol{\pi}_{t+k-1}^{(k-1)}\right) \times p_b\left(\boldsymbol{\pi}_{t+k-1}^{(k-1)}\right)$$

• Last element $t = 1$:

$$p_b\left(\pi_1 \middle| \boldsymbol{\pi}_k^{(k-1)}\right) = \prod_{i=1}^{k} p(\pi_i|\pi_{i-1}) \times \hat{p}^{(i)}\left(\mathbf{d}_i^{(i)}|\boldsymbol{\pi}_i^{(i)}\right)^{1/k} \Big/ z_k\left(\boldsymbol{\pi}_k^{(k-1)}\right)$$

$$p_b\left(\boldsymbol{\pi}_{k-1}^{(k-1)}\right) = \sum_{\pi_k=1}^{L} p_b\left(\pi_1 \middle| \boldsymbol{\pi}_k^{(k-1)}\right) \times p_b\left(\boldsymbol{\pi}_k^{(k-1)}\right)$$

The full kth-order approximate posterior model is computed by

$$\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d}) = \left[\prod_{t=1}^{T-k} p_b\left(\pi_t \middle| \boldsymbol{\pi}_{t+k-1}^{(k-1)}\right)\right] \times p_b\left(\boldsymbol{\pi}_T^{(k-1)}\right).$$

Realizations from the exact posterior model $p(\boldsymbol{\pi}|\mathbf{d})$ in (2) can be sampled with the approximate posterior model set as an independent sampler proposal distribution in MH-steps. In each MH-step where $\boldsymbol{\pi}$ is the current value, a new realization $\tilde{\boldsymbol{\pi}}$ is sampled from the approximate posterior model with acceptance probability

$$\alpha(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}) = \min\left\{1, \frac{p(\tilde{\boldsymbol{\pi}}|\mathbf{d})}{p(\boldsymbol{\pi}|\mathbf{d})} \times \frac{\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d})}{\hat{p}^{(k)}(\tilde{\boldsymbol{\pi}}|\mathbf{d})}\right\}. \quad (25)$$

When the parameters are given, the order $k = 2$ is found to be suitable [6]. The algorithm for sampling from the exact posterior model is given in the following.

Algorithm A2 Sample from posterior model

- Sample $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_T)$ from the approximate posterior model:
  - Sample $(\tilde{\pi}_{T-k+2}, \ldots, \tilde{\pi}_T) = \tilde{\boldsymbol{\pi}}_T^{(k-1)} \sim p_b(\boldsymbol{\pi}_T^{(k-1)})$
  - For $t = T-k+1, \ldots, 1$ sample $\tilde{\pi}_t \sim p_b(\pi_t|\tilde{\boldsymbol{\pi}}_{t+k-1}^{(k-1)})$
- Accept/reject $\tilde{\boldsymbol{\pi}}$ with acceptance probability $\alpha(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})$ by (25).

The sampling algorithm has runtime $O(T)$.

ACKNOWLEDGMENT

This work is part of the Uncertainty in Reservoir Evaluation activity-consortium in the Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

REFERENCES

[1] R. E. Sheriff and L. P. Geldart, *Exploration Seismology*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1995.
[2] A. L. Larsen, M. Ulvmoen, H. Omre, and A. Buland, "Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model," *Geophysics*, vol. 71, no. 5, pp. R69–R78, Sep. 2006.
[3] H. Hammer and H. Tjelmeland, "Approximate forward–backward algorithm for a switching linear Gaussian model," *Comput. Stat. Data Anal.*, vol. 55, no. 1, pp. 154–167, Jan. 2011.
[4] M. Ulvmoen and H. Hammer, "Bayesian lithology-fluid inversion—Comparison of two algorithms," *Comput. Geosci.*, vol. 14, no. 2, pp. 357–367, Mar. 2010.
[5] M. Ulvmoen and H. Omre, "Improved resolution in Bayesian lithology/fluid inversion from prestack seismic data and well observations: Part 1—Metodology," *Geophysics*, vol. 75, no. 2, pp. R21–R35, Mar. 2010.
[6] K. Rimstad and H. Omre, "Approximate posterior distributions for convolutional two-level hidden Markov models," *Comput. Stat. Data Anal.*, vol. 58, pp. 187–200, Feb. 2013.
[7] D. V. Lindberg and H. Omre, "Blind categorical deconvolution in two-level hidden Markov models," *IEEE Trans. Geosc. Remote Sens.*, vol. 52, no. 11, pp. 7435–7447, Nov. 2014.
[8] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.*, vol. 41, no. 3, pp. 164–171, Feb. 1970.
[9] K. Rimstad, P. Avseth, and H. Omre, "Hierarchical Bayesian lithology/fluid prediction: A North Sea case study," *Geophysics*, vol. 77, no. 2, pp. B69–B85, Mar. 2012.
[10] R. Reeves and A. N. Pettitt, "Efficient recursions for general factorisable models," *Biometrika*, vol. 91, no. 3, pp. 751–757, Sep. 2004.
[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 39, pp. 1–38, 1977.
[12] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Int. Comput. Sci. Inst., Berkeley, CA, USA Tech. Rep. ICSI TR-97-021, 1998.
[13] G. C. G. Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *J. Amer. Stat. Assoc.*, vol. 85, no. 411, pp. 699–704, Sep. 1990.
[14] B. S. Caffo, W. Jank, and G. L. Jones, "Ascent-based Monte Carlo expectation-maximization," *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 67, no. 2, pp. 235–251, Apr. 2005.
[15] W. Jank, "Implementing and diagnosing the stochastic approximation EM algorithm," *J. Comput. Graph. Stat.*, vol. 15, no. 4, pp. 803–829, Dec. 2006.
[16] G. Celeux and J. Diebolt, "A stochastic approximation type EM algorithm for the mixture problem," *Stoch. Stoch. Rep.*, vol. 41, no. 1/2, pp. 119–134, Oct. 1992.
[17] J. Eidsvik, T. Mukerji, and P. Switzer, "Estimation of geological attributes from a well log: An application of hidden Markov chains," *Math. Geol.*, vol. 36, no. 3, pp. 379–397, Apr. 2004.
[18] O. Rosec, J. M. Boucher, B. Nsiri, and T. Chonavel, "Blind marine seismic deconvolution using statistical MCMC methods," *IEEE J. Ocean. Eng.*, vol. 28, no. 3, pp. 502–512, Jul. 2003.
[19] J. J. Kormylo and J. M. Mendel, "Maximum-likelihood seismic deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. GE-21, no. 1, pp. 72–82, Jan. 1983.
[20] K. Aki and P. G. Richards, *Quantitative Seismology: Theory and Methods*. New York, NY, USA: Freeman, 1980.
[21] A. Buland and H. Omre, "Bayesian linearized AVO inversion," *Geophysics*, vol. 68, no. 1, pp. 185–198, Jan. 2003.

David Volent Lindberg received the M.Sc. degree in industrial mathematics and the Ph.D. degree in geostatistics from the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2010 and 2014, respectively.

He is currently a Researcher with DNV GL, Oslo, Norway. His research interests include Bayesian inverse problems, hidden Markov models, blind seismic deconvolution, state-space models, parameter estimation, uncertainty quantification, and stochastic optimization.

Henning Omre received the M.Sc. degree in statistics from the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1975 and the Ph.D. degree in geostatistics from Stanford University, Stanford, CA, USA, in 1985.

In the period 1976–1992, he was with the Norwegian Computing Center. In 1992, he joined, as a Professor of statistics, the Department of Mathematical Sciences, NTNU, where he is currently the Head of the "Uncertainty in Reservoir Evaluation" Consortium. His major topics of research are spatiotemporal statistics and Bayesian inversion.