Paper II

Inference of the Transition Matrix in Convolved Hidden Markov Models by a Generalized Baum-Welch Algorithm

David V. Lindberg & Henning Omre

Paper submitted

Inference of the Transition Matrix in Convolved Hidden Markov Models by a Generalized Baum-Welch Algorithm

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 process as the hidden bottom layer. The model parameters include the transition probabilities of the Markov process, response means and variances and the convolution kernel and error variance. Focus of the study is on categorical deconvolution of the observations into the hidden bottom categorical layer when the transition probability parameters defining the Markov process transition matrix are unknown. The inversion is cast in a Bayesian setting where the solution is represented by an approximate posterior model. Three methods for inference on the unknown parameters are discussed, for which a generalization of the common Baum-Welch algorithm is found to be preferable. The inference techniques are compared in two empirical studies and in a seismic case study from a petroleum reservoir offshore Norway.

1 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 travels through the subsurface Earth layers from which it is reflected back and measured by signal receivers (Sheriff and Geldart 1995). 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 Earths 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 layers.

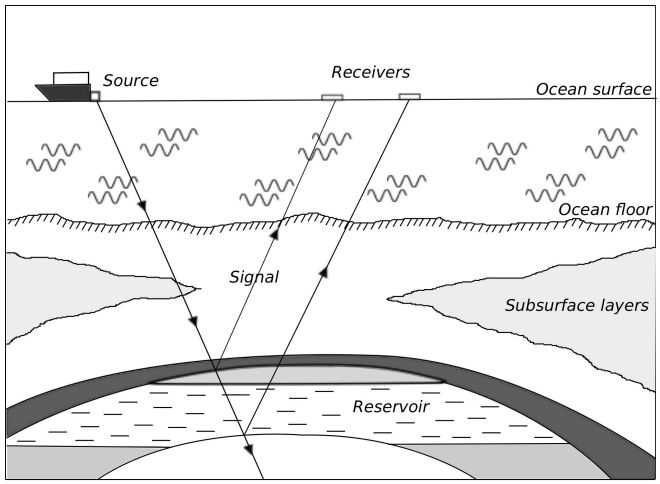


Figure 1: Cross section of a schematic offshore reservoir subject to a marine seismic survey. Sound waves are emitted from the signal source, reflected by the subsurface layers and registered by signal receivers.

A standard hidden Markov model is defined for two layers: hidden categorical states following a Markov process as the bottom layer and observations as the top layer that are elementwise conditional independent given the bottom layer. A convolutional two-level hidden Markov model is defined with three layers: a hidden bottom categorical layer following a Markov process, a hidden 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 the hidden middle layer and the categorical subsurface rock type classes the hidden bottom level. The model is introduced in Larsen et al. (2006) for the seismic inverse problem of deconvolution into subsurface lithology/fluid classes, and later generalized in Rimstad and Omre (2013). Blind categorical deconvolution is considered in Lindberg and Omre (2014), i.e. categorical deconvolution when the parameters of the convolution kernel and the observation error in addition are unknown. In this work, we consider categorical deconvolution when the parameters in the hidden bottom categorical layer are unknown, in particular the transition matrix parameter defining the transition probabilities of the Markov process.

For the blind categorical deconvolution problem in Lindberg and Omre (2014), the unknown parameters of interest are assessed by approximate maximum likelihood 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 standard hidden Markov models, Baum et al. (1970) shows that all model parameters can be assessed simultaneously by maximum likelihood estimation analytically, the method known as the Baum-Welch algorithm. For the two-level 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 described: 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 algorithm and 3) a fully Bayesian sample-based approach by a Markov chain Monte Carlo algorithm. The three inference methods are compared in a small example inspired by the standard model for blind seismic deconvolution, next in an empirical test study with three different reference transition matrices, and finally in a case study for seismic AVO inversion into lithology/fluid classes from a reservoir offshore Norway.

In the notation, let $p(\cdot)$ be the common generic expression for both the prob-

ability density function (pdf) and the probability mass function (pmf). Moreover, the term $\phi_i(\cdot)$ denotes a Gaussian pdf where i is the dimension of the variable.

2 Model Description

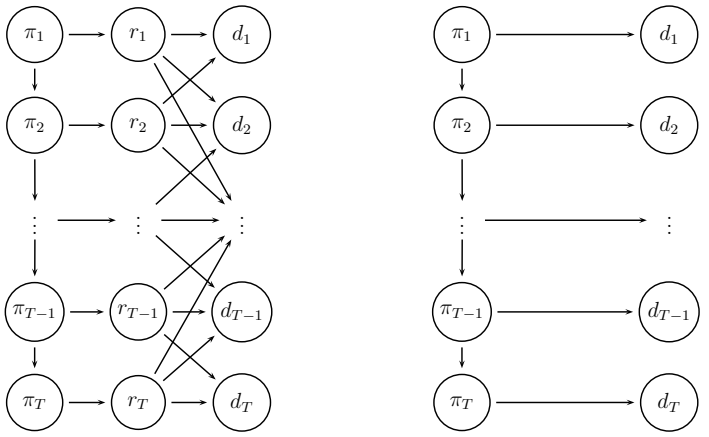
Suppose we have a stochastic field on the lattice $\mathscr{L} = \{1,\dots,T\}$ with equal step length δ, with continuous observations discretized into $\mathbf{d} = (\mathbf{d}_1,\dots,\mathbf{d}_T)'$, $\mathbf{d}_t \in \mathbb{R}^m$, $t \in \mathscr{L}$. Associated to the observations are a set of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

$$d_t = \sum_{\nu=-a}^{a} w_\nu r_{t+\nu} + e_t \,, \tag{1}$$

hence each element d_t is thus gathered as a locally weighted sum of the elements in $\mathbf{r}$. Here, $\mathbf{w} = (w_{-a},\dots,w_a)'$ 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 $\boldsymbol{\pi} = (\pi_1,\dots,\pi_T)'$ where each state π_t, $t \in \mathscr{L}$ belong to one of L classes, which state space is set to integers for notational convenience, $\pi_t \in \Omega_\pi : \{1,\dots,L\}$. The relationship between the variables are displayed in Fig.2a where the directed arrows represent dependencies. A comparison to the standard one-level HMM is displayed in Fig.2b, in which the convolution relation is not present. The objective is to assess the hidden categorical layer given the observations, $[\boldsymbol{\pi}|\mathbf{d}]$, which solution in a Bayesian probabilistic setting, given all model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_p, \boldsymbol{\theta}_l\}$, is given by the posterior model

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

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



(a) Convolved two-level HMM **(b)** Standard HMM

Figure 2: Directed acyclic graph of (a) a convolutional two-level HMM and (b) a standard (one-level) HMM and in which the directed arrows represent dependencies 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.

2.1 Markov Chain Prior Model

The categorical variables in $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ are assigned a first order Markov chain prior model. Hence the class at position t depends on the previous state only, i.e. $p(\pi_t|\pi_1, \ldots, \pi_T) = p(\pi_t|\pi_{t-1})$, see Fig.2a. 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 indices 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 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 d elements thick is

$$p(d) = \left(1 - \frac{1}{u_i}\right)^{d-1} \times \frac{1}{u_i}, \quad d = 1, 2, 3, \dots .$$

Here, $(u_1, \dots, u_L)$ are expected thicknesses defined by the diagonal elements of the transition matrix, $\{P_{ii},\ i = 1, \dots, L\}$ as $u_i = \delta/(1 - P_{ii})$. High expected thickness thus correspond to large probability for transitions from a class into the same class.

2.2 Likelihood Model

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

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

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

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

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

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

The observation likelihood model captures the convolution effect of Exp.(1),

for which the error term is assumed to be 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

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

Here $\mathbf{W}$ is 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 Exp.(1), which we solve by choosing truncated convolution kernels in $\mathbf{W}$. The observation likelihood model is hence fully defined by the model parameters $\boldsymbol{\theta}_{ol} = \{\mathbf{w}, \sigma_d^2\}$.

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

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

and is hence fully defined by the model parameters $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{rl}, \boldsymbol{\theta}_{ol}\}$. 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 can be seen in Fig.2a.

2.3 Approximate Posterior Model

The posterior model of the categorical variables in Exp.(2) is defined with prior and likelihood model given by Exp.(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 dependency 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 therefor approximate the posterior model according to Larsen et al. (2006); Rimstad and Omre (2013);Lindberg and Omre (2014), by a kth order approximate model $\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d})$, gaining a model on factorisable form which can be fully computed by the recursive generalized forward-backward (FB) algorithm (Reeves and Pettitt 2004). Define a kth order categorical state by $\pi_t^{(k)} = (\pi_{t-k+1}, \ldots, \pi_t)$, $t = k, \ldots, T$. The kth order approximate posterior model on factorisable form is then given

as

$$\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d}) = C_d^{(k)} \times q_{k-1}\left(\pi_{k-1}^{(k-1)}|\mathbf{d}\right) \times q_T\left(\pi_T^{(k-1)}|\mathbf{d}\right) \times \left[\prod_{t=k}^{T} \hat{p}^{(k)}\left(d_t^{(k)}|\pi_t^{(k)}\right)^{1/k} \times p\left(\pi_t^{(k)}|\pi_{t-1}^{(k)}\right)\right], \quad (8)$$

where $C_d^{(k)}$ is a normalizing constant, $q_{k-1}(\cdot)$ and $q_T(\cdot)$ are edge factors, $\hat{p}^{(k)}(d_t^{(k)}|\pi_t^{(k)})$ a kth order approximate likelihood model and $p(\pi_t^{(k)}|\pi_{t-1}^{(k)})$ defines the Markov process transitions between the kth order states. The generalized FB-algorithm for computing the full approximate posterior model is given as Algorithm A1 in Appendix A. To sample realizations from the exact posterior model $p(\boldsymbol{\pi}|\mathbf{d})$ in Exp.(2), the approximate posterior model is set as an independent sampler proposal distribution in Metropolis Hastings (MH) steps, the full sampling algorithm given as Algorithm A2 in Appendix A. The approximate posterior model is in fact a non-stationary kth order Markov process with respect to the kth order states. For more details on the approximation and its notation, see Lindberg and Omre (2014).

3 Inference on the Markov Process Model Parameters

In this section, it is assumed that the prior model parameters, i.e. the transition matrix $\boldsymbol{\theta}_p = \{\mathbf{P}\}$, are unknown, all likelihood parameters $\boldsymbol{\theta}_l = \{\boldsymbol{\theta}_{rl}, \boldsymbol{\theta}_{ol}\}$ are assumed known. Three approaches for inference on $\mathbf{P}$ are considered: a) maximum likelihood (ML) estimation by an approximate Expectation-Maximization (aEM) algorithm, b) ML estimation by a Monte Carlo Expectation-Maximization (MCEM) algorithm and c) estimation through simulation based Bayesian inference by a Markov chain Monte Carlo (MCMC) algorithm. The major advantage for inference through MCMC compared to ML is that the ensemble of simulated realizations obtained when doing MCMC enable 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 $\boldsymbol{\pi}$, inference by MCMC thus provides samples from the true 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 with plug-in parameter estimate $p(\boldsymbol{\pi}|\mathbf{d}, \widehat{\mathbf{P}})$. However, ML estimation is much less computer demanding than by MCMC, especially for aEM, hence a trade-off between precision and computer demands is required.

3.1 ML Estimation by Approximate EM

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 (Dempster et al. 1977) for inverse problems. The term incomplete data refers in this 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

$$Q(\mathbf{P},\mathbf{P}^*) = \mathrm{E}_{\boldsymbol{\pi}|\mathbf{d};\mathbf{P}^*}\left[\log p(\boldsymbol{\pi},\mathbf{d};\mathbf{P})\right] = \sum_{\boldsymbol{\pi}} \log p(\boldsymbol{\pi},\mathbf{d};\mathbf{P}) p\left(\boldsymbol{\pi}|\mathbf{d};\mathbf{P}^*\right) \tag{9}$$

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.2b, Baum et al. (1970) proved that the parameters maximizing the Q-function in the M-step in Exp.(10) are analytical tractable, see also Bilmes (1998) for a detailed derivation. The likelihood parameters for a standard HMM are analytical tractable as well (Baum et al. 1970). 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 Exp.(10) are

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

where the probabilities are obtained from the FB-algorithm. For the convolved two-level HMM, the probabilities computed in the FB-algorithm are approximations, i.e. $\hat{p}^{(k)}(\pi_t^{(k-1)}|\mathbf{d})$, $t=k-1,\ldots,T$, see Appendix A. We can thus approximate the given probabilities by summing out for the remaining

states

$$\hat{p}(\pi_{t-1},\pi_t|\mathbf{d};\mathbf{P}^*)=\begin{cases}\sum_{\pi_{k-1}^{(k-1)}\setminus\{\pi_{t-1},\pi_t\}}\hat{p}^{(k)}(\pi_{k-1}^{(k-1)}|\mathbf{d};\mathbf{P}^*) & ,\ t<k-1\\ \sum_{\pi_t^{(k-1)}\setminus\{\pi_{t-1},\pi_t\}}\hat{p}^{(k)}(\pi_t^{(k-1)}|\mathbf{d};\mathbf{P}^*) & ,\ t\geq k-1\end{cases} \tag{11}$$

$$\hat{p}(\pi_t|\mathbf{d};\mathbf{P}^*)=\begin{cases}\sum_{\pi_{k-1}^{(k-1)}\setminus\{\pi_t\}}\hat{p}^{(k)}(\pi_{k-1}^{(k-1)}|\mathbf{d};\mathbf{P}^*) & ,\ t<k-1\\ \sum_{\pi_t^{(k-1)}\setminus\{\pi_t\}}\hat{p}^{(k)}(\pi_t^{(k-1)}|\mathbf{d};\mathbf{P}^*) & ,\ t\geq k-1\end{cases}, \tag{12}$$

where $\pi_t^{(k-1)}\setminus\{\cdot\}$ is the elements of the $(k-1)$th order categorical states except for the elements in the brackets, for example $\pi_t^{(k-1)}\setminus\{\pi_t\}=(\pi_{t-k+2},\ldots,\pi_{t-1})$. The aEM algorithm is given below.

ALGORITHM 1: APPROXIMATE EM FOR ESTIMATING $\mathbf{P}$

Initialize some $\mathbf{P}^*=\{P_{ij}^*\}_{i,j=1,\ldots,L}$.

Iteratively do

1. **FB-algorithm:** Run the FB-algorithm in Algorithm A1 to obtain the marginal approximate posterior probabilities $\hat{p}(\pi_t^{(k-1)}|\mathbf{d};\mathbf{P}^*)$, $t=k-1,\ldots,T$. Compute $\hat{p}(\pi_{t-1},\pi_t|\mathbf{d};\mathbf{P}^*)$ and $\hat{p}(\pi_t|\mathbf{d};\mathbf{P}^*)$ by Exp.(11) and (12).
2. **E and M step:** Solve 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 proposed aEM algorithm is a generalization of the Baum-Welch algorithm applied to a two-level HMM in terms of estimation of the Markov chain parameters in the prior model. The algorithm has runtime $O(V(T-k+2)L^k)$ where V is the number of iterations of the three steps given. An advantageous property of the Baum-Welch algorithm is that zero-probabilities in the transition matrix are kept when updating, hence we can enforce that impossible transitions are honored by setting the corresponding values to 0 in the

initial parameter input to the algorithm. Whereas the likelihood parameters for a standard HMM are analytical tractable, this is not the case for the convolutional two-level HMM because of the convolution effect. Throughout this paper these parameters are assumed to be known.

3.2 ML Estimation by MCEM

As it is possible to sample the hidden variables $\boldsymbol{\pi}$ from its posterior model given all model parameters we may proceed with a Monte Carlo EM (MCEM) algorithm as proposed by Wei and Tanner (1990). In each iteration, S realizations of the categorical field $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{S} \sim p(\boldsymbol{\pi}|\mathbf{d};\mathbf{P}^*)$ are sampled, see Algorithm A2 in Appendix A. Next, in the E-step, the Q-function integral in Exp.(9) is approximated by its Monte Carlo estimate

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

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

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

The full MCEM algorithm for the convolutional two-level HMM is described below.

ALGORITHM 2: MCEM FOR ESTIMATING $\mathbf{P}$

Initialize some $\mathbf{P}^*$.

Iteratively do

1. **FB-algorithm:** Run FB-algorithm in Algorithm A1 with input parameters $\mathbf{P}^*$ 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 Algorithm A2.
3. **Monte Carlo E (MCE) step:** Compute $\widehat{Q}(\mathbf{P},\mathbf{P}^*)$ by Exp.(13).
4. **M step:** Maximize $\widehat{Q}(\mathbf{P},\mathbf{P}^*)$ by Exp.(14).

5. Set $\mathbf{P}^* = \mathbf{P}$

The algorithm has runtime $O(V(ST + (T-k+2)L^k))$ where V is the number of iterations of the five steps given, which is slower than for aEM due to the sampling step. The latent field 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 of accuracy increases. See Caffo et al. (2005), Jank (2006) 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, termed a stochastic EM algorithm (SEM) (Celeux and Diebolt 1992). This is a quite poor approximate EM-algorithm, as one sample in each SE-step provides a poor approximation to the sum in the original E-step.

A problem occurring when doing inference by the MCEM algorithm is that whereas 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 Monte Carlo error. A solution proposed by Caffo et al. (2005) for $K = 1$ and later generalized by Jank (2006) is to check in each step that $U_{t,K} = \hat{\Delta}_{t,K} + z_\alpha(\hat{\sigma}^2_{t,K}/S_t) \geq 0$ in which

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

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

3.3 Estimation by Bayesian Inference Through MCMC

For parameter estimation by Bayesian inference, a prior parameter pdf is assigned to the unknown transition matrix $\mathbf{P}$. In a MCMC setting, in each iteration s we proceed by first sampling the categorical field from its posterior distribution under the current parameter estimate, $\boldsymbol{\pi}^{(s)} \sim p(\boldsymbol{\pi}|\mathbf{d}, \mathbf{P}^{(s-1)})$, followed by sampling the unknown transition matrix from its parameter posterior distribution given the current categorical field $\mathbf{P}^{(s)} \sim p(\mathbf{P}|\boldsymbol{\pi}^{(s)}, \mathbf{d})$. After

some burnin s_b, the samples $\{\boldsymbol{\pi}^{(s)}, \mathbf{P}^{(s)}\}_{s=s_b}^{S}$ will then be realizations from their joint posterior model $p(\boldsymbol{\pi}, \mathbf{P}|\mathbf{d})$ from which we can derive the desired inference.

The parameter posterior model can be shown to depend on $\boldsymbol{\pi}$ only

$$p(\mathbf{P}|\boldsymbol{\pi}, \mathbf{d}) = const \times p(\boldsymbol{\pi}|\mathbf{P}) \times p(\mathbf{P}) .$$

A Dirichlet prior model is assigned to each row of the transition matrix, with identical hyperparameters $\eta \geq 0$ following Eidsvik et al. (2004) 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

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

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 below.

ALGORITHM 3: MCMC FOR ESTIMATING $\mathbf{P}$

Set initial $\mathbf{P}^{(0)}$, $\boldsymbol{\pi}^{(0)}$.

For $s = 1, \ldots$ do

1. **FB-algorithm:** Run FB-algorithm in Algorithm A1 with input parameters $\mathbf{P}^{(s-1)}$ 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 Algorithm A2.
3. Sample $\mathbf{P}^{(s)} \sim p(\mathbf{P}|\boldsymbol{\pi}^{(s)}; \boldsymbol{\eta})$ according to Exp.(17),

---

The algorithm has runtime $O(V(T-k+2)L^k)$ as for aEM, where V is the number of iterations of the three steps given. For inference by MCMC however, the number of iterations V is much higher, to gain a significant number of realizations from the posterior distributions.

4 Toy Example: Seismic Deconvolution with Two Classes

A convolutional two-level hidden Markov model of length $T = 200$ with $L = 2$ possible classes $\Omega_x = \{\text{grey,black}\}$ 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. (2003). The BG model is commonly applied on blind seismic deconvolution problems, with the objective to identify transitions between different subsurface geological layers from convolved seismic data (Kormylo and Mendel 1983, Rosec et al. 2003). There are thus two classes for which black represents a change between two layers denoted a high reflector point and grey 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 $\mathbf{P}_{emp}$ estimated from $\boldsymbol{\pi}$ by a simple counting process are

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

with reference and empirical stationary pdfs $\mathbf{p}^s = (0.91, 0.09)$ and $\mathbf{p}^s_{emp} = (0.90, 0.10)$ and expected thicknesses $\mathbf{u} = (10.00, 1.02)$ and $\mathbf{u}_{emp} = (9.47, 1.00)$. Hence we expect thicker layers with some layer transitions, also the reference probability for consecutive layer transitions is set very low ($P_{22} = 0.02$). Notice that the empirical transition matrix differs from the reference transition matrix, in particular there are no consecutive black transitions in $\boldsymbol{\pi}$ resulting in a corresponding zero-probability in $\mathbf{P}_{emp}$. This is because the categorical profile considered is a finite trace, hence there will be heterogeneities in realizations simulated from the model. 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 to reproduce the empirical transition matrix better

than the reference transition matrix.

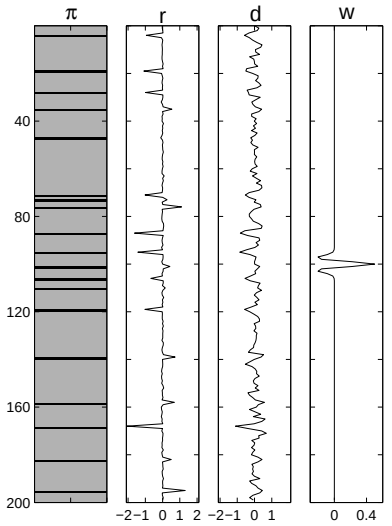


Figure 3: Reference categorical profile, response profile, observation profile and reference Ricker wavelet in the seismic inversion toy example.

The response likelihood model parameters are defined for the two classes as $\mu_{r|\pi} = \{0, 0\}$, $\sigma^2_{r|\pi} = \{0.001, 1\}$, hence the response from high reflector points has high variance whereas inhomogeneities within the layers are represented by a much smaller variance. The reference response profile simulated from $\boldsymbol{\pi}$ is displayed 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 heterogeneities in the response error variance however, hence we cannot expect to recapture these in the inversion. The observation profile $\mathbf{d}$ is simulated from the response profile $\mathbf{r}$ convolved with a Ricker wavelet $R(1.5, 0.5)$ wavelet defined by

$$R(t; \lambda, \gamma) = \gamma \left(1 - \frac{t^2}{\lambda^2}\right) \exp\left\{-\frac{t^2}{2\lambda^2}\right\}, \qquad \lambda > 0, \quad \gamma \geq 0\,. \tag{19}$$

and error variance $\sigma_d^2 = 0.15^2$. The reference observation profile $\mathbf{d}$ and wavelet $\mathbf{w}$ is displayed in Fig.3. The main objective applying the BG model is on inversion from $\mathbf{d}$ 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 in Algorithm 1 is the least computer demanding out of the three inference techniques, 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 iterations, see Fig.4. Some of the runs in Fig.4 however do not display a monotone increase in the approximate log-likelihood value, which occurs because the approximate likelihood model also depends on the transition matrix. The final transition matrix estimate is

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

with estimated stationary pdf $\hat{\mathbf{p}}^{s}_{aEM} = (0.85, 0.15)$ and expected waiting times $\hat{\mathbf{u}}_{aEM} = (5.43, 1)$. The black class of high reflector points is hence slightly overrepresented in the estimate compared to the reference transition matrix in Exp.(18). The 90% confidence intervals of each element in $\mathbf{P}$ are given in Fig.5a, computed by the Hessian matrix and truncated at [0,1]. Notice that the reference transition matrix values are all within the confidence bounds. Ten thousand realizations of the categorical field are simulated by Algorithm A2 in Appendix A, $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{10000} \sim p(\boldsymbol{\pi}|\mathbf{d}; \widehat{\mathbf{P}}_{aEM})$, with a 6th order approximate posterior model set as proposal distribution. The marginal posterior probabilities $p(\pi_t|\mathbf{d}; \widehat{\mathbf{P}}_{aEM})$, $t = 1, \ldots, T$ and the marginal maximum a posterior (MAP) prediction computed from the realizations are displayed in Fig.6a. Notice that the MAP 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 grey to black. Some of the high reflector points with smaller deviations in the response are not identified as expected.

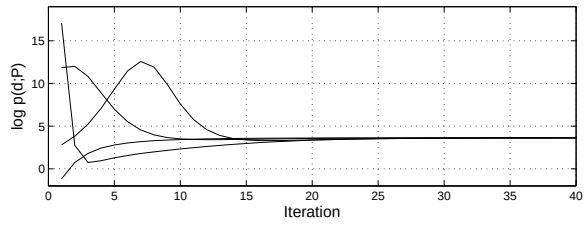
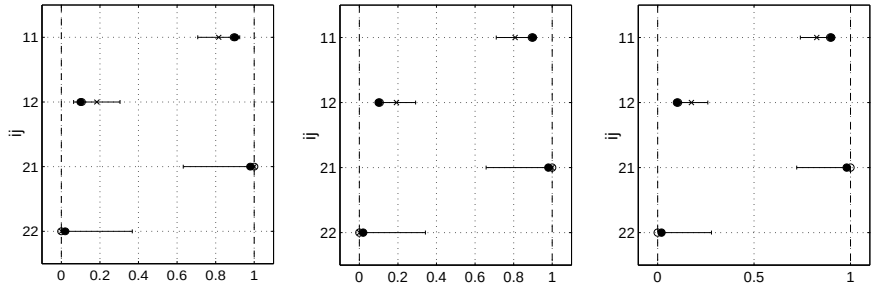


Figure 4: Trace plots for aEM of the approximate log-likelihood value $log\{\hat{p}^{(k)}(\mathbf{d}; \mathbf{P})\}$ for different initial values.

For inference by MCEM, in each iteration in Algorithm 2 thousand realiza-



(a) aEM **(b)** MCEM **(c)** MCMC

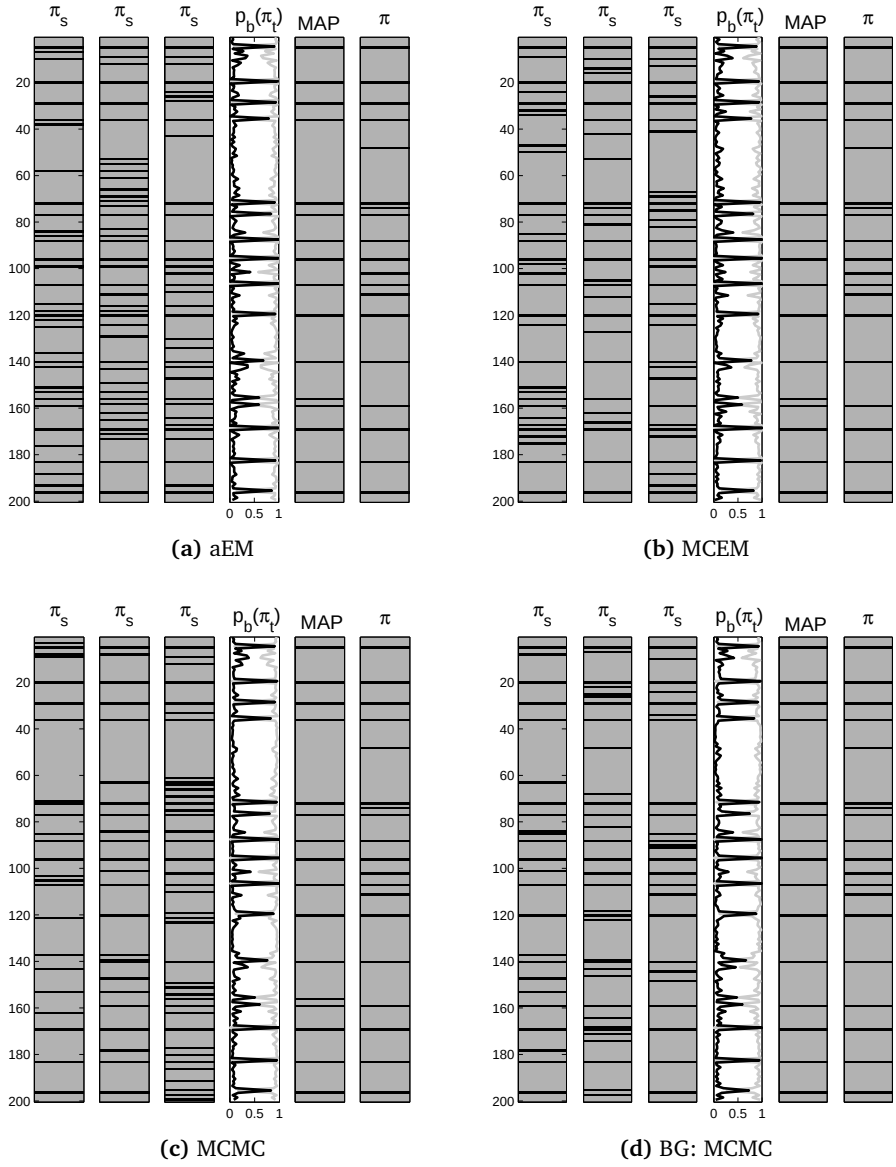
Figure 5: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2\}$. The estimates as crosses, reference values as black dots and empirical values as white dots.

tions of the categorical field are sampled 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 4th order approximate posterior model set as proposal distribution. The optimization in the M-step is performed by the MATLAB function FMINSEARCH. After about 50 iterations, the algorithm seems to have converged, see Fig.7. The approximate Q functions from Exp.(13), increase until convergence as expected and the values of $\hat{\Delta}_{t,K}$ defined in Exp.(15) also converge towards 0 as desired, when the differences between the approximate Q functions across iterations diminishes. The final transition matrix estimate is

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

with estimated stationary pdf $\hat{\mathbf{p}}^s_{MCEM} = (0.84, 0.16)$ and expected waiting times $\hat{\mathbf{u}}_{MCEM} = (5.20, 1)$. The 90% confidence intervals on the elements of $\mathbf{P}$ again capture the reference values, see Fig.12c. As for aEM, ten thousand realizations of the categorical field are simulated for the last iteration, $\{\boldsymbol{\pi}^{(s)}\}_{s=1}^{10000} \sim p(\boldsymbol{\pi}|\mathbf{d}; \widehat{\mathbf{P}}_{MCEM})$, with a 6th order approximate posterior model as proposal distribution. The marginal posterior probabilities and the MAP prediction are displayed in Fig.6b. Notice again that the MAP prediction corresponds very reliably to the reference profile, and that it is almost identical to the aEM mode.

For inference by MCMC, the algorithm is run with ten thousand iterations. The Dirichlet parameter prior model in Exp.(16) is set with hyperparameter



(a) aEM **(b)** MCEM

(c) MCMC **(d)** BG: MCMC

Figure 6: Prediction of the categorical profile compared (a)-(c) across the inference methods and (d) to 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 MAP predictions computed from the realizations and the reference categorical profile.

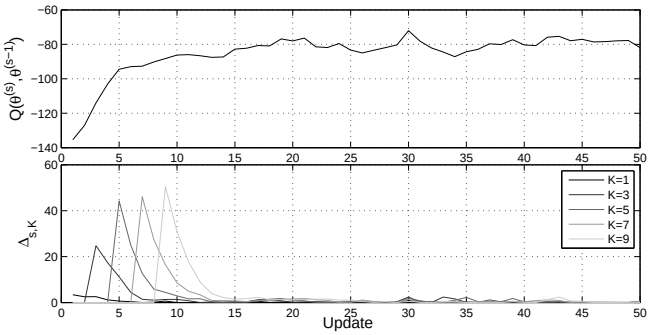


Figure 7: Trace plots for MCEM, the top plot displaying the approximate Q functions, the lower plot the values of $\hat{\Delta}_{t,K}$ for $K = 1, 3, 5, 7, 9$.

$\eta = 1$. The MCMC algorithm seems to have converged after about 500 steps, see Fig.8, and the burn-in value is set to $s_b = 2000$. The estimated transition matrix is

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

with estimated stationary pdf $\hat{\mathbf{p}}^s_{MCMC} = (0.84, 0.16)$ and expected waiting times $\hat{\mathbf{u}}_{MCMC} = (5.71, 1.02)$. 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 should be able to capture. The 90% confidence intervals for the elements of $\mathbf{P}$, computed from the 0.9 range of the simulated values, are given in Fig.5c. The marginal posterior probabilities and MAP prediction on $\boldsymbol{\pi}$ computed from the ensemble of realizations, are displayed in Fig.6c. The MAP prediction corresponds very reliably to the reference profile again, in fact the MAP predictions for the three approaches are almost identical. Notice also that consecutive high reflector points are present in some of the realizations, as the MCMC estimate was the only estimate with a significant corresponding probability $\hat{P}_{22}$.

For the BG model we follow Rosec et al. (2003), doing MCMC with ten thousand updates, 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 HMM. The marginal posterior probabilities and MAP prediction on $\boldsymbol{\pi}$ are displayed in Fig.6d, and are as reliable as the predictions

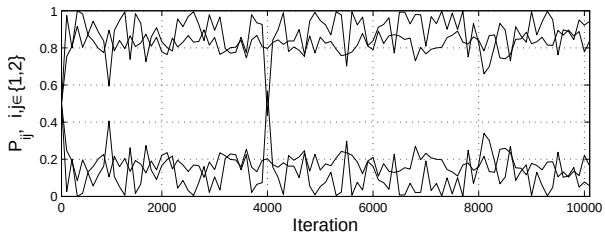


Figure 8: Trace plots of the four sampled elements of **P** for estimation by MCMC.

from the convolved HMM. As for the MCMC estimate for the HMM, some realizations have thicker 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.

5 Empirical Test Study

A convolutional two-level hidden Markov model of length $T = 200$ with $L = 3$ possible classes $\Omega_x = \{\text{white,grey,black}\}$ is considered. Three reference prior model transition matrices are considered, termed case 1 through 3. The reference categorical profiles for the three cases simulated from the respective prior models are displayed in Fig.9. The reference and empirical matrices are

$$\mathbf{P}_1 = \begin{pmatrix} 0.50 & 0.50 & 0 \\ 0.25 & 0.50 & 0.25 \\ 0 & 0.50 & 0.50 \end{pmatrix} \quad , \quad \mathbf{P}_{1,emp} = \begin{pmatrix} 0.59 & 0.41 & 0 \\ 0.22 & 0.53 & 0.25 \\ 0 & 0.55 & 0.45 \end{pmatrix} \tag{20}$$

$$\mathbf{P}_2 = \begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix} \quad , \quad \mathbf{P}_{2,emp} = \begin{pmatrix} 0.94 & 0.04 & 0.01 \\ 0.02 & 0.91 & 0.07 \\ 0.06 & 0.09 & 0.85 \end{pmatrix} \tag{21}$$

$$\mathbf{P}_3 = \begin{pmatrix} 0.20 & 0 & 0.80 \\ 0.10 & 0.20 & 0.70 \\ 0.15 & 0.15 & 0.70 \end{pmatrix} \quad , \quad \mathbf{P}_{3,emp} = \begin{pmatrix} 0.10 & 0 & 0.90 \\ 0.09 & 0.13 & 0.78 \\ 0.17 & 0.14 & 0.69 \end{pmatrix} \tag{22}$$

with corresponding reference and empirical stationary pdfs

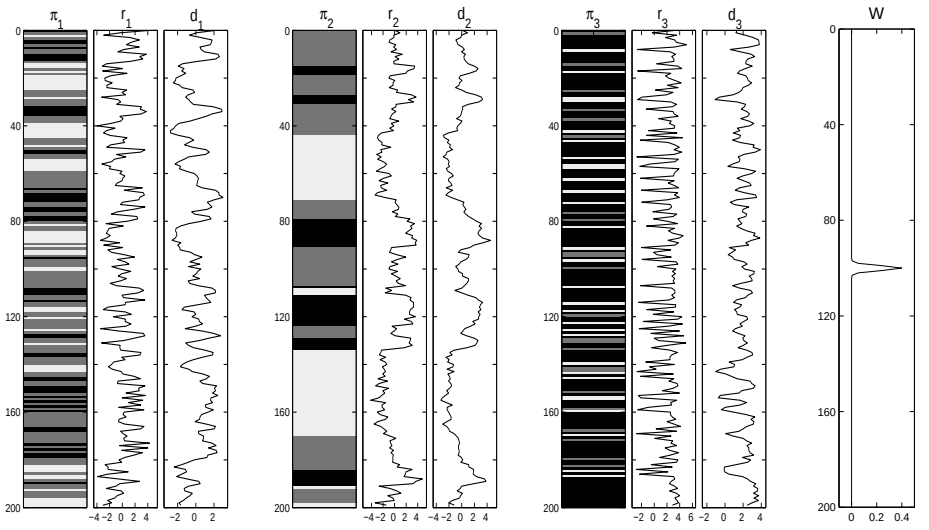
$$\begin{aligned}
\mathbf{p}_1^s &= (0.25, 0.50, 0.25) \ , \ \mathbf{p}_{1,emp}^s = (0.27, 0.51, 0.22) \\
\mathbf{p}_2^s &= (0.33, 0.34, 0.33) \ , \ \mathbf{p}_{2,emp}^s = (0.39, 0.39, 0.22) \\
\mathbf{p}_3^s &= (0.15, 0.13, 0.72) \ , \ \mathbf{p}_{3,emp}^s = (0.15, 0.12, 0.73)
\end{aligned}$$

and reference and empirical expected thicknesses

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

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.9a. We also enforce in the prior model for case 1 that the classes black and white can never be neighbors. For case 2, the probability of staying in the current class is much larger than for switching class, hence thicker layers, see Fig.9b. For case 3, the transition probabilities into the black class is the largest, hence the black class will dominate with some thinner grey and white layers, see Fig.9c. We enforce in the prior model in case 3 that the class white can never be on top of grey. 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 $\mu_{r|\pi} \in \{-2, 0, 3\}$, $\sigma_{r|\pi} = \{0.7, 0.7, 0.7\}$, corresponding to the three classes. The observation likelihood parameters are set to a Gaussian $N(0, 1)$ convolution kernel and error variance $\sigma_d^2 = 0.3^2$. 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. The reference Gaussian $N(0, 1)$ convolution kernel is displayed in Fig.9d.

We assess the transition matrix $\mathbf{P}$ for each case by the three inference methods aEM, MCEM and MCMC with respective subscripts a, b, c in the estimates. Confidence intervals of the nine elements of the transition matrix and the corresponding marginal MAP predictions of the categorical states are compared.



(a) Case 1 **(b)** Case 2 **(c)** Case 3 **(d)** Kernel

Figure 9: (a)-(c) Reference categorical profiles, response profiles and observation profiles for the empirical test study cases and (d) reference convolution kernel.

5.1 Case 1

For case 1, with reference transition matrix $\mathbf{P}_1$ in Exp.(20), the estimated transition matrices are

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

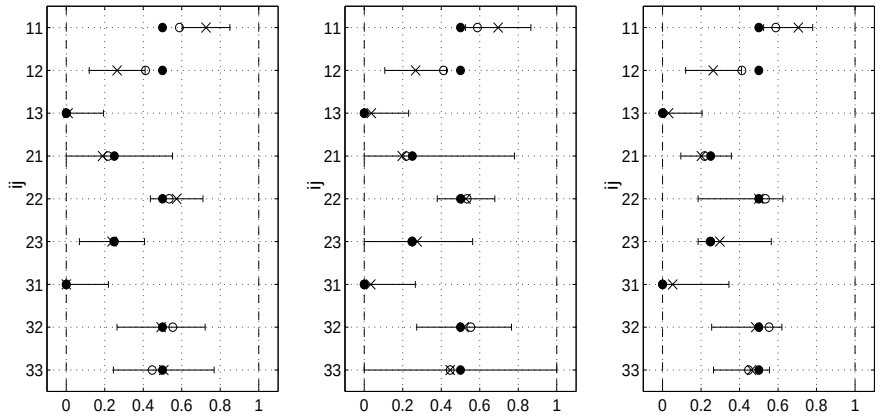
$$\widehat{\mathbf{P}}_{1c} = \begin{pmatrix} 0.70 & 0.26 & 0.04 \\ 0.18 & 0.54 & 0.28 \\ 0.06 & 0.49 & 0.45 \end{pmatrix}.$$

All three methods have small estimated probabilities P_{13} and P_{31} as desired, with the aEM estimates the lowest. The diagonal probability estimates are reliable, being close to 0.5, however slightly overestimated for the white class being closer to the empirical values as expected. The 90% confidence intervals are displayed in Fig.10, with all intervals capturing the reference values

except for the non-zero 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 method has the widest intervals, hence with most parameter estimate uncertainty. The corresponding estimated stationary pdfs and expected thicknesses are

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

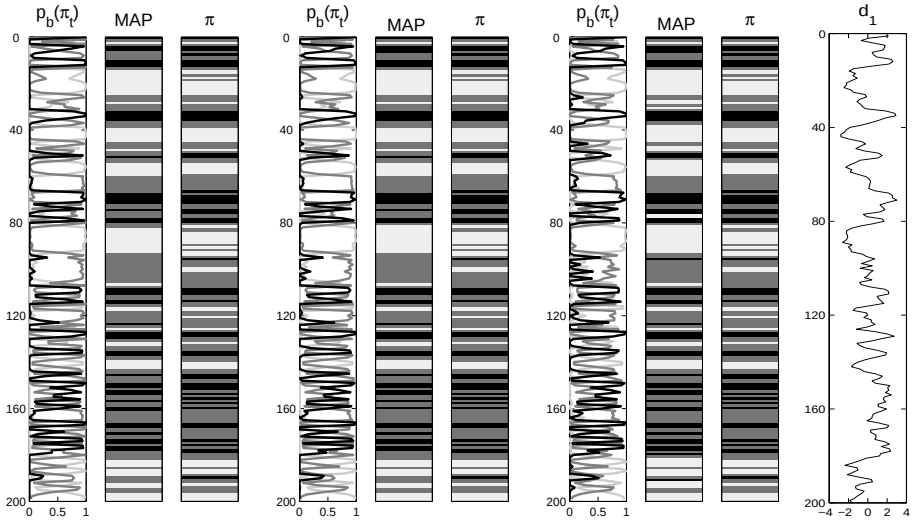
and resemble the reference values reliably. The white class has a bit too high estimated expected thickness because of the overestimated probability P_{11}.



(a) aEM **(b)** MCEM **(c)** MCMC

Figure 10: Case 1: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$. The estimates as crosses, reference values as black dots and empirical values as white dots.

The marginal posterior probabilities on the categorical profile $p(\pi_t|\mathbf{d})$ for $t = 1, \ldots, T$ and the MAP predictions compared to the reference profile are given in Fig.11. The predictions are very similar for the three inference methods, resembling the reference profile well, however with a bit too thick white layers because of the overestimated transition matrix value. A few impossible transitions between white and black are present in the MCMC mode, which estimated transition matrix has the highest corresponding estimates P_{13} and P_{31}.



(a) aEM **(b)** MCEM **(c)** MCMC **(d)** Obs

Figure 11: Case 1: Locationwise marginal posterior probabilities, $p(\pi_t|\mathbf{d})$, and the MAP predictions computed from the realizations and the reference categorical profile and observations.

5.2 Case 2

For case 2, with reference transition matrix $\mathbf{P}_2$ in Exp.(21), the estimated transition matrices are

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

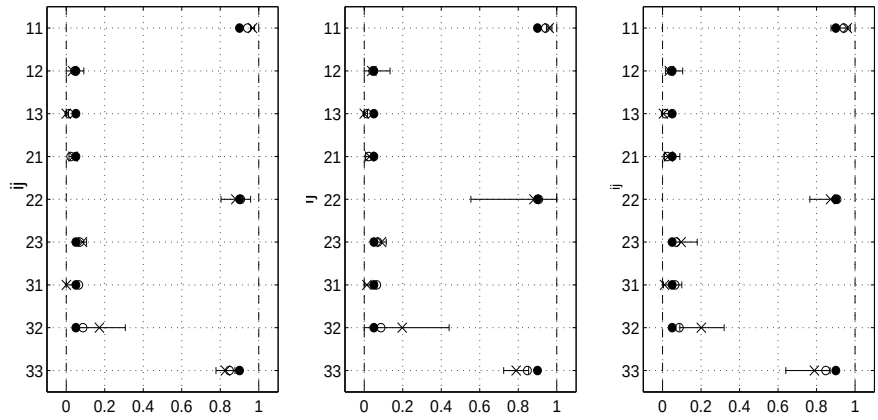
$$\widehat{\mathbf{P}}_{2c} = \begin{pmatrix} 0.96 & 0.03 & 0.01 \\ 0.03 & 0.87 & 0.10 \\ 0.01 & 0.20 & 0.79 \end{pmatrix}.$$

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 the

opposite are consistently underestimated, with the aEM and MCEM methods even estimating zero-probabilities for P_{13}. The 90% confidence intervals are displayed in Fig.12, again with most intervals capturing the reference and empirical values. Notice the narrow intervals for the low probabilities. The corresponding estimated stationary pdfs and expected thicknesses are

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

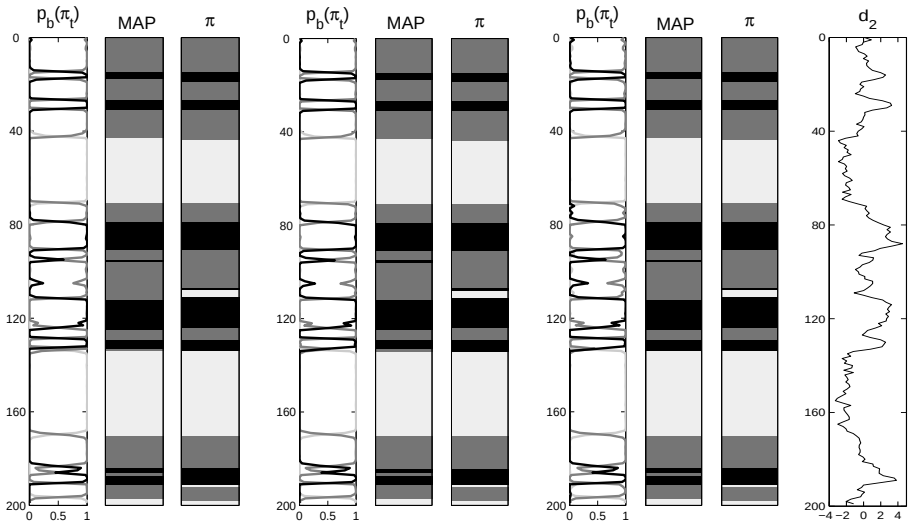
The estimated stationary pdfs recognize the reference values reliably, while the expected thicknesses for white are too large and for black too low corresponding to the estimates of P_{11} and P_{33}. The expected thickness is very sensitive to high diagonal probabilities, as the estimate for P_{11} of 0.97 for aEM compared to the reference value 0.90 increases the expected thickness to above 30 compared to the reference expected thickness of about 10.



(a) aEM **(b)** MCEM **(c)** MCMC

Figure 12: Case 2: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$.

The marginal posterior probabilities on the categorical profile and the MAP predictions are given in Fig.13, and are again almost identical for the three inference methods. The MAP predictions are resembling the reference profile well, but are all unable to recognize the thinner white middle layer. The MAP predictions are not identical to the reference profiles also because of model heterogeneities with a limited reference profile.



(a) aEM **(b)** MCEM **(c)** MCMC **(d)** Obs

Figure 13: Case 2: Locationwise marginal posterior probabilities, MAP predictions, the reference categorical profile and the observations.

5.3 Case 3

For case 3, with reference transition matrix $\mathbf{P}_3$ in Exp.(22), the estimated transition matrices are

$$\widehat{\mathbf{P}}_{3a} = \begin{pmatrix} 0.18 & 0.05 & 0.77 \\ 10^{-17} & 0.28 & 0.72 \\ 0.13 & 0.12 & 0.75 \end{pmatrix}, \widehat{\mathbf{P}}_{3b} = \begin{pmatrix} 0.16 & 10^{-12} & 0.84 \\ 10^{-7} & 10^{-7} & \sim 1 \\ 0.17 & 0.09 & 0.74 \end{pmatrix}$$

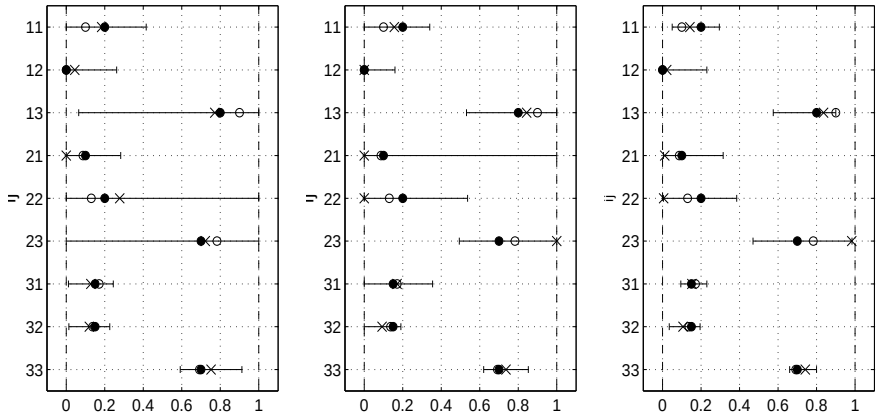
$$\widehat{\mathbf{P}}_{3c} = \begin{pmatrix} 0.14 & 0.02 & 0.84 \\ 0.01 & 0.01 & 0.98 \\ 0.15 & 0.11 & 0.74 \end{pmatrix}.$$

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 almost recognizing the zero-probability from white to grey in P_{12}. The probabilities from grey to white and grey to grey in P_{21} and P_{22} are consistently underestimated, except for P_{22} for aEM. The 90% confidence intervals

are displayed in Fig.14, with all intervals capturing the reference values. The interval widths are larger than for the two previous cases, see especially the intervals for P_{13}, P_{22} and P_{23} for aEM and for P_{21} for MCEM that cover the full interval $(0,1)$, hence with larger parameter estimate uncertainty. The corresponding estimated stationary pdfs and expected thicknesses are

$$\hat{\mathbf{p}}^s_{3a} = (0.12, 0.13, 0.75)\ ,\ \hat{\mathbf{p}}^s_{3b} = (0.16, 0.07, 0.77)\ ,\ \hat{\mathbf{p}}^s_{2c} = (0.15, 0.11, 0.74)$$
$$\hat{\mathbf{u}}_{3a} = (1.23, 1.39, 4.04)\ ,\ \hat{\mathbf{u}}_{3b} = (1.19, 1.00, 3.79)\ ,\ \hat{\mathbf{u}}_{3c} = (1.17, 1.01, 3.86)$$

and resemble the reference values reliably. The grey class will for the MCEM and MCMC estimate almost always be one-layered, corresponding to the small estimates of P_{22}.



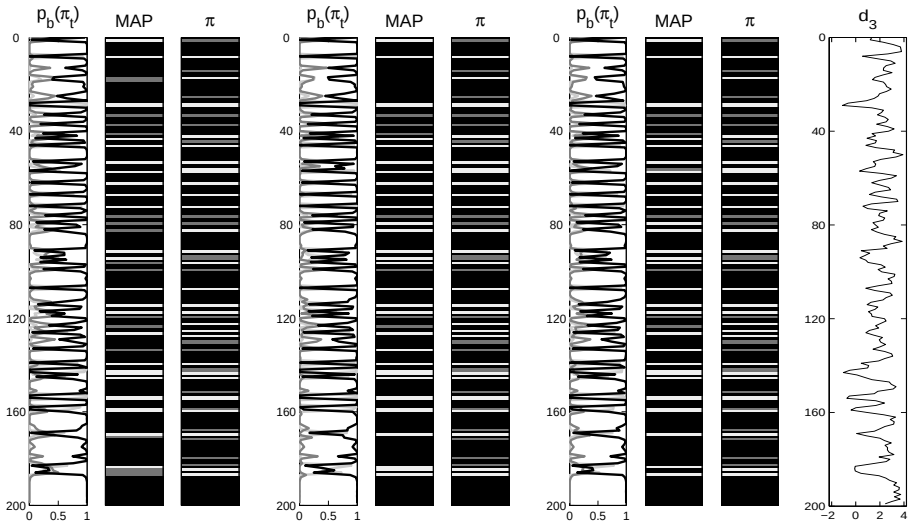
(a) aEM **(b)** MCEM **(c)** MCMC

Figure 14: Case 3: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3\}$.

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

5.4 Discussion

The results from the three test cases, along with the results from the previous toy example with two classes, indicate that all three inference methods estimate the transition matrix reliably. Whereas the estimates for aEM and MCEM capture zero-probabilities better, they also tend to estimate some mis-



(a) aEM **(b)** MCEM **(c)** MCMC **(d)** Obs

Figure 15: Case 3: Locationwise marginal posterior probabilities, MAP predictions, the reference categorical profile and the observations.

specified zero-probabilities. For simulation based inference through MCMC, zero-probabilities are not captured due to the Dirichlet posterior model. Since the reference profiles considered have limited length, there will be model heterogeneities, hence the estimates are expected to resemble the empirical values better than the reference values. The confidence intervals for the transition matrix elements for MCMC are always within the valid region $(0, 1)$ because the sampled values are always within the region. For aEM and MCEM, the confidence intervals are approximated by Hessians, hence the need to truncate the intervals to avoid unfeasible values. The approximated confidence intervals may also be subject to higher uncertainty as seen for e.g. P_{22} and P_{23} for Case 3 by aEM in Fig.14a. The predictions of the categorical field $\boldsymbol{\pi}$ are reliable for all three inference methods. As inference by aEM is computationally superior to the other two methods, it should therefor be preferred.

6 Case Study

Data from a well in a reservoir offshore Norway is considered, see Rimstad and Omre (2013), with the four lithology/fluid (LF) classes {gas-filled sand-

stone, oil-filled sandstone, water-filled sandstone, shale} represented by the colors white, light-grey, dark-grey and black respectively. The true LF-profile is given in Fig.16, to which we assign a Markov process, together with synthetic seismic AVO traces from three angles $12°, 22°, 31°$ following Rimstad and Omre (2013). 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, is unknown.

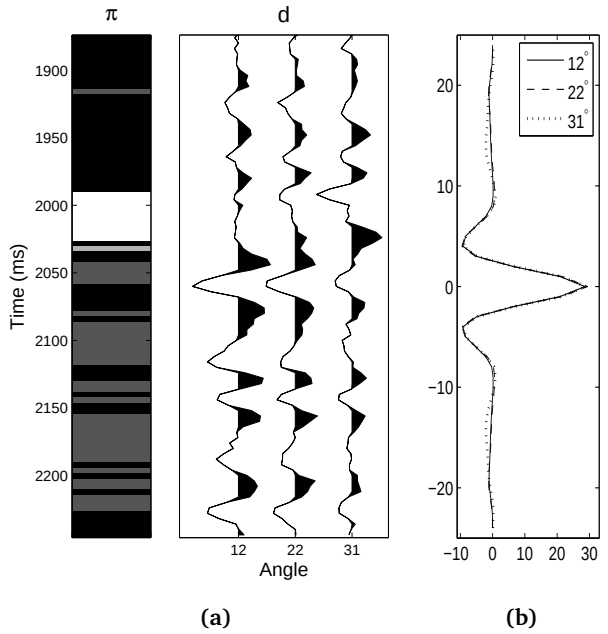


Figure 16: Seismic case study: (a) Reference LF-profile and prestack seismic traces from the angles $\{12°, 22°, 31°\}$ and (b) reference wavelets for the three angles.

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 displayed in Fig.16b. 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, which 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

$$\mu_{r|\pi} \in \begin{cases} \{8.11, 8.12, 8.16, 7.88\} \\ \{7.61, 7.60, 7.58, 7.08\} \\ \{7.66, 7.69, 7.73, 7.77\} \end{cases}, \quad \sigma_{r|\pi} = \begin{cases} 0.0860 \\ 0.1483 \\ 0.0608 \end{cases} \tag{23}$$

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 (Aki and Richards 1980), $\mathbf{D}$ is a central differential matrix calculating the contrasts and $\mathbf{e}$ are independent Gaussian errors, see Buland and Omre (2003) for more details.

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

$$\mathbf{P}_{emp} = \begin{pmatrix} 0.89 & 0 & 0 & 0.11 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0.69 & 0.31 \\ 0.01 & 0.01 & 0.11 & 0.87 \end{pmatrix} \tag{24}$$

with corresponding reference stationary pdf $\mathbf{p}^s_{emp} = (0.07, 0.01, 0.23, 0.69)$ and expected thicknesses $\mathbf{u}_{emp} = (9.00, 1, 3.20, 7.92)$. Notice that the gravity effects are honored, however the probabilities P_{12}, P_{13}, P_{22} and P_{23} are also set to 0 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 only thickness one.

The transition matrix is estimated from the seismic observations by the approximate EM algorithm only, which was shown to be the preferred choice of inference in the previous sections. As zero-probabilities are kept when updating in the Baum-Welch algorithm, we enforce the probabilities P_{21}, P_{31} and P_{32} to be 0 by setting them to 0 in the initial parameter input to the algorithm

to honor the gravity effects. The final transition matrix estimate is

$$\widehat{\mathbf{P}} = \begin{pmatrix} 0.84 & \sim 10^{-5} & \sim 10^{-11} & 0.16 \\ 0 & 0.97 & \sim 10^{-5} & 0.03 \\ 0 & 0 & 0.78 & 0.22 \\ 0.02 & 0.01 & 0.08 & 0.89 \end{pmatrix} \tag{25}$$

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 Exp.(24) 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. This indicates inference problems for rarely occurring classes. The 90% confidence intervals are displayed in Fig.17 for which most intervals capture the empirical values as desired.

The marginal posterior probabilities of the LF states throughout the profile and the marginal MAP prediction together with three random realizations from the posterior model with plug-in estimates are given in Fig.18. In particular the hydrocarbon layer of most interest, i.e. the gas-filled sandstone layer, is reliably recognized, whereas the thin oil-filled sandstone layer is harder to identify because of the convolution effect. The MAP prediction and realizations resemble the reference LF profile reliably, with uncertainty in the posterior model represented by the variability in the realizations.

7 Conclusion

The inverse problem of categorical deconvolution, that is inversion of convolved continuous observations into underlying categorical states, is cast in a Bayesian setting. The underlying states are distributed according to a stationary Markov chain, for which the transition probability parameters that defines the Markov process are assumed to be unknown. These parameters are collected in the transition matrix. The solution to the Bayesian inverse problem is represented by an approximate posterior model in which the categorical states are distributed according to a non-stationary Markov chain.

Three inference methods on the unknown model parameters are considered,

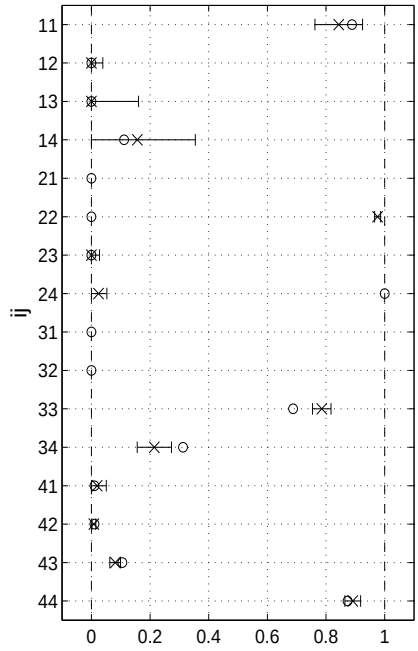


Figure 17: Approximate 90% confidence intervals for the elements of $\widehat{P}_{ij}$ for $i, j \in \{1, 2, 3, 4\}$ representing the LF classes {gas-filled sandstone, oil-filled sandstone, water-filled sandstone, shale} respectively. The estimates as crosses, empirical values as white dots. The elements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

for which a trade-off between computation time and precision is made. Of the inference methods considered, an approximate expectation-maximization algorithm is found best in terms of computation time while also providing as accurate model parameter estimates. Approximate confidence bounds are computable through Hessians. The algorithm is in fact a generalization of the well known Baum-Welch algorithm applied on convolved data, when inference is made on the transition probability parameters only.

A toy example from seismic inversion is presented, in which the objective is to identify layer transitions between subsurface rock type classes. The corresponding transition matrix is reliably estimated, and the layer transitions are reliably predicted under the corresponding model parameter estimates. An empirical test study with various transition matrices is presented, for which the structure of the categorical field differs significantly. The inference meth-

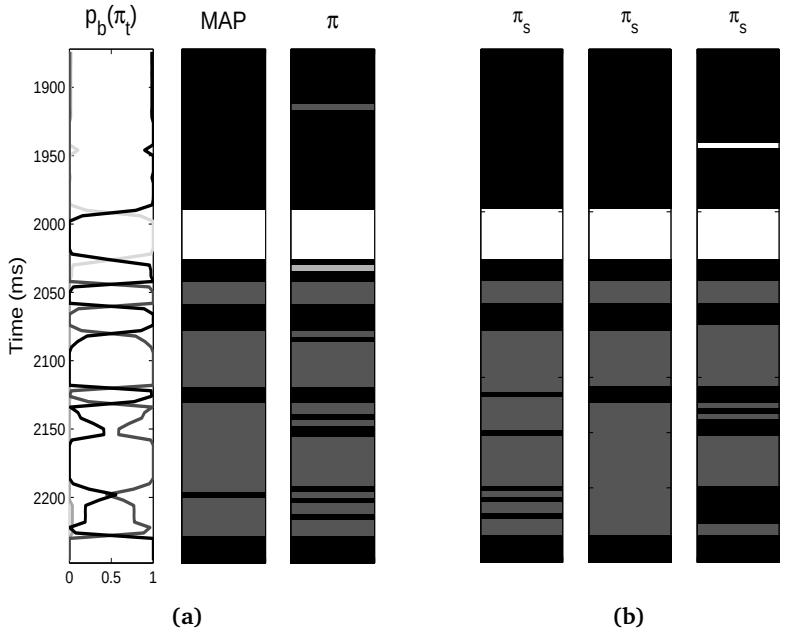


Figure 18: (a) Locationwise marginal posterior probabilities, $p(\pi_t|\mathbf{d})$, the MAP prediction and the reference categorical profile and (b) three realizations simulated from the posterior model, $\boldsymbol{\pi}_s \sim p(\boldsymbol{\pi}|\mathbf{d};\widehat{\mathbf{P}})$ in the case study.

ods consistently provide reliable estimates and corresponding predictions of the categorical field. With limited reference categorical profiles, the transition probability estimates resemble the empirical values estimated directly from the reference profiles better than the reference parameter values. The approximate Baum-Welch algorithm is applied to data from a reservoir offshore Norway, with the objective to identify the subsurface lithology/fluid layers. The empirical transition probabilities between the lithology/fluid states are reliably estimated, and the predictions of the categorical profile reliably identifies the hydrocarbon layer of most interest.

A natural extension of the generalized Baum-Welch algorithm presented is to include inference on all parameters, as is the case for the original Baum-Welch algorithm. This would thus constitute a solution to the blind categorical deconvolution problem, i.e. with unknown convolution kernel and observation errors.

Acknowledgments

This research is a part of the Uncertainty in Reservoir Evaluation (URE) activity at the Norwegian University of Science and Technology (NTNU).

References

Aki, K. & Richards, P. G. (1980), *Quantitative seismology: Theory and methods*, W.H. Freeman and Co., New York.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970), 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *The Ann. of Math. Stat.* **41**(3), 164–171.

Bilmes, J. (1998), A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, Technical report.

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

Caffo, B. S., Jank, W. & Jones, G. L. (2005), 'Ascent-based Monte Carlo expectation-maximization', *Journal of the Royal Statistician Society, ser. B* **67**, 235–251.

Celeux, G. & Diebolt, J. (1992), 'A stochastic approximation type EM algorithm for the mixture problem', *Stochastics and Stochastic Rep.* **41**(1-2), 119–134.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistician Society Ser. B* **39**, 1–38.

Eidsvik, J., Mukerji, T. & Switzer, P. (2004), 'Estimation of geological attributes from a well log: An application of hidden Markov chains', *Math. Geology* **36**(3).

Jank, W. (2006), 'Implementing and diagnosing the stochastic approximation EM algorithm', *Journal of Computational and Graphical Statistics* **15**(4), 803–829.

Kormylo, J. J. & Mendel, J. M. (1983), 'Maximum-likelihood seismic deconvolution', *IEEE Trans. Geosci. Remote Sens.* **GE-21**(1), 72–82.

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

Lindberg, D. V. & Omre, H. (2014), 'Blind categorical deconvolution in two-level hidden Markov models', *Accepted for publication in IEEE Transactions on Geoscience and Remote Sensing* .

Reeves, R. & Pettitt, A. N. (2004), 'Efficient recursions for general factorisable models', *Biometrika* **91**(3), 751–757.

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

Rosec, O., Boucher, J. M., Nsiri, B. & Chonavel, T. (2003), 'Blind marine seismic deconvolution using statistical MCMC methods', *IEEE J. Ocean. Eng.* **28**(3), 502–512.

Sheriff, R. E. & Geldart, L. P. (1995), *Exploration Seismology*, 2 edn, Cambridge University Press.

Wei, G. C. G. & Tanner, M. A. (1990), 'A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms', *Journal of the American Statistical Association* **85**(411), 699–704.

Appendix A: Forward-Backward Algorithm

Define backward probabilities as conditioned on the full set of observations, e.g. $p_b(\pi_t^{(k-1)}) = \hat{p}^{(k)}(\pi_t^{(k-1)}|\mathbf{d})$ and $p_b(\pi_t|\pi_{t+k-1}^{(k-1)}) = \hat{p}^{(k)}(\pi_t|\pi_{t+k-1}^{(k-1)},\mathbf{d})$. The FB algorithm presented follows Reeves and Pettitt (2004) and has runtime $O((T-k+2)L^k)$.

ALGORITHM A1: FORWARD-BACKWARD ALGORITHM

Forward:

- Initiate for $t = k$:
  $z_k\left(\pi_k^{(k-1)}\right) = \sum_{\pi_1}\prod_{i=1}^{k} p(\pi_i|\pi_{i-1}) \times \hat{l}^{(i)}\left(\pi_i^{(i)}\right)^{1/k}$
- Iterate for $t = k+1,\ldots,T-1$:
  $z_t\left(\pi_t^{(k-1)}\right) = p(\pi_t|\pi_{t-1}) \times \sum_{\pi_{t-k+1}} \hat{l}^{(k)}\left(\pi_t^{(k)}\right)^{1/k} \times z_{t-1}\left(\pi_{t-1}^{(k-1)}\right)$

- Last element, $t = T$:

$$z_T\left(\pi_T^{(k-1)}\right) = p(\pi_T|\pi_{T-1}) \times \sum_{\pi_{T-k+1}} \prod_{i=1}^{k} \left[\hat{l}^{(i)}\left(\pi_T^{(i)}\right)^{1/k} \times z_{T-1}\left(\pi_{T-1}^{(k-1)}\right)\right]$$

- Normalizing constant:

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

Backward:

- Initiate:

$$p_b\left(\pi_T^{(k-1)}\right) = C_d^{(k)} \times z_T\left(\pi_T^{(k-1)}\right)$$

$$p_b\left(\pi_{T-k+1}\middle|\pi_T^{(k-1)}\right) = \frac{p(\pi_T|\pi_{T-1}) \times \prod_{i=1}^{k} \hat{l}^{(i)}\left(\pi_T^{(i)}\right)^{1/k} \times z_{T-1}\left(\pi_{T-1}^{(k-1)}\right)}{z_T\left(\pi_T^{(k-1)}\right)}$$

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

- Iterate for $t = T-k, \ldots, 2$:

$$p_b\left(\pi_t\middle|\pi_{t+k-1}^{(k-1)}\right) = \frac{p(\pi_{t+k-1}|\pi_{t+k-2}) \times \hat{l}^{(k)}\left(\pi_{t+k-1}^{(k)}\right)^{1/k} \times z_{t+k-2}\left(\pi_{t+k-2}^{(k-1)}\right)}{z_{t+k-1}\left(\pi_{t+k-1}^{(k-1)}\right)}$$

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

- Last element $t = 1$:

$$p_b\left(\pi_1\middle|\pi_k^{(k-1)}\right) = \frac{\prod_{i=1}^{k} p(\pi_i|\pi_{i-1}) \times \hat{l}^{(i)}\left(\pi_i^{(i)}\right)^{1/k}}{z_k\left(\pi_k^{(k-1)}\right)}$$

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

The full kth order approximate posterior model is computed by

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

Realizations from the exact posterior model $p(\boldsymbol{\pi}|\mathbf{d})$ in Exp.(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

$$\alpha(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}) = \min\left\{1, \frac{p(\tilde{\boldsymbol{\pi}}|\mathbf{d})}{p(\boldsymbol{\pi}|\mathbf{d})} \times \frac{\hat{p}^{(k)}(\boldsymbol{\pi}|\mathbf{d})}{\hat{p}^{(k)}(\tilde{\boldsymbol{\pi}}|\mathbf{d})}\right\}. \tag{26}$$

When the parameters are given, the order $k = 2$ is found to be suitable (Rimstad and Omre 2013). The algorithm for sampling from the exact posterior model is given below.

ALGORITHM A2: SAMPLE FROM THE POSTERIOR MODEL

- Sample $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_T)$ from the approximate posterior model:
  - Sample $(\tilde{\pi}_{T-k+2}, \ldots, \tilde{\pi}_T) = \tilde{\pi}_T^{(k-1)} \sim p_b\left(\pi_T^{(k-1)}\right)$
  - For $t = T-k+1, \ldots, 1$ sample $\tilde{\pi}_t \sim p_b\left(\pi_t \left| \tilde{\pi}_{t+k-1}^{(k-1)}\right.\right)$
- Accept/reject $\tilde{\boldsymbol{\pi}}$ with acceptance probability $\alpha(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})$ by Exp.(26).

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