NTNU – Trondheim
Norwegian University of
Science and Technology

Bayesian Inversion and Inference of Categorical Markov Models with Likelihood Functions Including Dependence and Convolution

Torstein Mæland Fjeldstad

Master of Science in Statistics
Submission date: June 2015
Supervisor: Karl Henning Omre, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

A convolutional two-level Markov model is studied in this thesis. The bottom level contains a latent Markov chain, and given the variables, the middle contains a latent Gaussian random field. We observe the second level through a convolution with additive Gaussian noise. Previously studied models are extended by including additional spatial correlation in the middle layer.

We propose two different approximations of the likelihood function, namely the truncation and projection approximation, of varying order. These approximate models are exactly assessed by the Forward-Backward algorithm.

Properties of various predictors are studied in different approximate posterior models. The predictors are seen to be stable with respect to an increase of the spatial correlation in the response model. An increase of k, being the approximation order, is not seen to have a great effect on the predictors.

The approximate posterior models are used as proposal densities in a Metropolis-Hastings algorithm to assess the correct posterior model, and we quantify the quality of each approximation by the acceptance rate. The acceptance rate is observed to be an increasing function of k. We observed higher acceptance rates when the proportion of the acquisition convolution was high, relative to the spatial correlation. A high class response variance also increased the acceptance rate.

Estimation of the transition matrix, using the EM-algorithm and simulation based inference, is found to be feasible under certain conditions. A univariate maximum marginal likelihood estimation of the model parameter in the Ricker acquisition convolution kernel is considered.

Samandrag

I denne masteroppgåva studerer me ein konvolvert to-nivå Markov modell. Det første nivået er ei ikkje-observerbar Markovkjede, som definerer eit ikkje-observerbart Gaussisk stokastisk felt. Me observerer dette feltet gjennom ein konvolusjon, saman med Gaussiske feil. Modellen vår utvidar tidlegare studerte modellar ved å inkludere romleg korrelasjon på det midterste nivået.

Me føreslår to ulike approksimasjonar for likelihoodfunksjonen. Dei er baserte på høvevis trunkering og projeksjon. Dei approksimative modellane kan evaluerast eksakt med framlengs-baklengs algoritmen.

Ulike prediktorar for den approksimative posteriorifordelinga er samanlikna, og me studerer eigenskapane deira under ulike modellføresetnader. Prediktorane er observert å vere nær uavhengig av romleg korrelasjon i responsmodellen, samt nær uavhengig av approksimasjonsordenen, k.

Dei approksimative modellane er nytta som forslagsfordelingar i ein Metropolis-Hastings algoritme til å generere realisasjonar frå den sanne posteriorifordelinga. Akseptansesannsynet er nytta som eit mål for å kvantifisere approksimasjonen. Akseptansesannsynet er observert å auke saman med k. Approksimasjonane er sett å vere gode når konvolusjon i observasjonsmodellen er stor, samanlikna med den romlege korrelasjonsfunksjonen. Akseptansesannsynet er observert å auke dersom variansen i responsklassane vert auka.

Parameterestimering av overgangsmatrisa ved hjelp av EM-algoritmen og simulering, er studert under visse føresetnader. Estimatet er sett å samsvare med den sanne overgangsmatrisa i gitte tilfelle. A priori kjennskap er sett å vere naudsynt, særskilt dersom dei ulike klassane overlappar kvarandre. Univariat optimalisering av marginal likelihoodfunksjonen er studert for ein Rickerfunksjon.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Professor Henning Omre, for his help and guidance during my work. His inputs and feedback has ensured my progress, and he has been very encouraging and supportive during my effort to complete my thesis.

I would also like to thank Assistant Professor Dario Grana at the University of Laramie, Wyoming, for his hospitality last fall.

Special thanks to my friends and fellow students for five enjoyable years.

I would also like to thank my family for their support during my stay in Trondheim.

Finally, I would like to thank Torill for her continuous support.

Contents

Chapter 1

Introduction

This chapter introduces the necessary notation and defines the variables of interest. We relate our variables of interest to seismic inversion, and introduce briefly the concepts of Bayesian inversion. A short introduction to point predictors and parameter inference is given.

1.1 Outline of Notation

A generic vector of length t is denoted by $\mathbf{a} = (a_1, \ldots, a_t)^\top$, and we define $\mathbf{a}_{-k} = (a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_t)^\top$. We denote a generic $(t \times s)$-matrix by $\mathbf{A}$, where the identity matrix is denoted by $\mathbf{I}$. Element (i, j) in $\mathbf{A}$ is denoted by $[\mathbf{A}]_{ij}$. The indicator function, $\mathbf{1}\{A\}$, is defined to be equal to 1 if A is true, and 0 otherwise.

A random variable $\mathbf{x}$ with sample space $\Omega_{\mathbf{x}}$, is assumed to be distributed according to a generic probability distribution $p(\mathbf{x})$. If $\mathbf{x}$ is discrete we refer to $p(\mathbf{x})$ as a probability mass function, and if $\mathbf{x}$ is continuous we refer to it as a probability density function. Relevant probability distributions are given in Appendix A.

1.2 Problem Description

We consider a random field defined on $\mathcal{D} \in \mathbb{R}$, discretized onto a lattice $\mathcal{L}_{\mathcal{D}} : \{1, \ldots, N\}$. This can for example represent a vertical profile through a geological unit, such as a seismic profile penetrating the subsurface.

Our variable of interest is a vector $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{N_\kappa})^\top$, where we for notational ease let $N_\kappa = N$. For $n = 1, \ldots, N$, each κ_n represents a nominal or ordinal class with $\kappa_n \in \Omega_{\boldsymbol{\kappa}} : \{1, \ldots, K\}$. This could for example represent the lithology/fluid-characteristics, such as {shale, sand/brine, sand/oil, sand/gas}. The set $\Omega_{\boldsymbol{\kappa}}^N$ is defined to be the K^N possible configurations of $\boldsymbol{\kappa}$, which in practice usually is an extremely large set.

We observe a continuous vector $\mathbf{d} = (d_1, \ldots d_{N_d})^\top$, where $N_d \leq N$ in most situations. In for example reservoir modelling, the observations may contain information from seismic data, well-logs or production history data. We only consider one-dimensional observations, i.e. $d_n \in \mathbb{R}$ for $n = 1, \ldots, N_d$, but it is possible to extend to multivariate observations. Buland and Omre (2003) discuss how the latter can be modeled with seismic amplitude versus offset (AVO) data. The elastic properties P-wave velocity, S-wave velocity and density are modeled utilizing the fact that the seismic reflection amplitude depends on the contrast of the material properties and reflection angles at each point of reflection.

Our goal is to assess $[\boldsymbol{\kappa}|\mathbf{d}]$, i.e. classify the latent categorical vector based on the observations. We operate in a probabilistic framework,

$$[\boldsymbol{\kappa}|\mathbf{d}] \sim p(\boldsymbol{\kappa}|\mathbf{d}), \tag{1.1}$$

where the random variable $[\boldsymbol{\kappa}|\mathbf{d}]$ is distributed according to the probability mass function $p(\boldsymbol{\kappa}|\mathbf{d})$. A major benefit with assessing Eq. (1.1) in a probabilistic framework is that we can provide point predictions with uncertainty statements.

We assess Eq. (1.1) in a Bayesian framework, where we assign a prior model, $p(\boldsymbol{\kappa})$, to $\boldsymbol{\kappa}$. The prior model represents a priori knowledge of $\boldsymbol{\kappa}$, for example the expected waiting time in each class. Correspondingly, we define an observation model, $[\mathbf{d}|\boldsymbol{\kappa}] \sim p(\mathbf{d}|\boldsymbol{\kappa})$. Since $\mathbf{d}$ is given, and $\boldsymbol{\kappa}$ is the unknown variable, $p(\mathbf{d}|\boldsymbol{\kappa})$ is in fact a likelihood function as it need not be normalized with respect to $\boldsymbol{\kappa}$. The posterior model for $[\boldsymbol{\kappa}|\mathbf{d}]$ is assessed by using Bayes' theorem,

$$p(\boldsymbol{\kappa}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})}{p(\mathbf{d})}. \tag{1.2}$$

The posterior model $p(\boldsymbol{\kappa}|\mathbf{d})$ is referred to as the solution to a Bayesian inversion problem. Being a function of $\boldsymbol{\kappa}$, the posterior is seen to be proportional to the likelihood times the prior. The probabilistic characteristics of $[\boldsymbol{\kappa}|\mathbf{d}]$ are captured in the posterior. We may generate realizations from the posterior model.

We operate in a predictive setting, and want to make predictions with the associated uncertainty statements. We choose the maximum a posteriori probability (MAP) predictor as our predictor since the predictor is contained in the discrete sample space. This need not be true for the posterior mean or median. The MAP predictor is defined as

$$\hat{\boldsymbol{\kappa}} = \arg\max_{\boldsymbol{\kappa}} \left\{ p(\boldsymbol{\kappa}|\mathbf{d}) \right\}. \tag{1.3}$$

Assessment of the MAP predictor constitutes a hard problem since it requires evaluation of K^N possible configurations of $\boldsymbol{\kappa}$. An alternative is therefore to consider the marginal MAP (MMAP) predictor,

$$\hat{\hat{\boldsymbol{\kappa}}} = \left\{ \hat{\kappa}_n = \arg\max_{\kappa_n} \left\{ p(\kappa_n|\mathbf{d}) \right\}; n \in \mathcal{L}_{\mathcal{D}} \right\}. \tag{1.4}$$

Uncertainty statements can be made by computing the marginal probabilities for each class. In practice the predictors differ from the posterior median, which is dependent on the labeling of $\boldsymbol{\kappa}$.

Both the prior and likelihood models are dependent on unknown model parameters. We denote them $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_l)$, where respectively $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_l$ are the prior and likelihood model parameters. To make the dependence on model parameters clear, we may rewrite Eq. (1.2) as

$$p(\boldsymbol{\kappa}|\mathbf{d};\boldsymbol{\theta}) = \frac{p(\mathbf{d}|\boldsymbol{\kappa};\boldsymbol{\theta}_l)p(\boldsymbol{\kappa};\boldsymbol{\theta}_p)}{p(\mathbf{d};\boldsymbol{\theta})}. \tag{1.5}$$

The maximum marginal likelihood estimator, $\hat{\boldsymbol{\theta}}$, and the normalization constant $p(\mathbf{d};\boldsymbol{\theta})$ are closely related since

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ p(\mathbf{d};\boldsymbol{\theta}) \right\}. \tag{1.6}$$

Eq. (1.6) can for example be maximized using the expectation-maximization (EM) algorithm. Due to the spatial dependency and possible local optima, the optimization might be complex to perform.

It is also possible to impose prior knowledge on $\boldsymbol{\theta}$, by assuming $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. The assessment of $\boldsymbol{\theta}$ is then cast into a Bayesian inference setting. Then we are able to generate posterior realizations from $p(\boldsymbol{\theta}|\mathbf{d})$. The latter can be done using Markov chain Monte Carlo simulation.

In Chapter 2 we introduce the current model in greater detail. We specify a convolutional Markov model through a prior, response and acquisition model, and deduce the posterior model. We study various k-th order approximations of the posterior model in Chapter 3, which can be assessed by the Forward-Backward algorithm. In Chapter 4 we study various model parameter estimation techniques, and discuss how the various model parameters can be assessed efficiently. Chapter 5 contains a thorough study of MAP predictors for various likelihood approximations. We compare various distance measures between the correct posterior model and the approximate posterior model. In Chapter 6 we have included two case studies where we estimate the transition matrix. In Chapter 7 a synthetic seismic test study is included. Finally, a summary of our findings are given in Chapter 8.

Chapter 2

Probabilistic Model

The posterior model,

$$p(\boldsymbol{\kappa}|\mathbf{d};\boldsymbol{\theta}) = \frac{p(\mathbf{d}|\boldsymbol{\kappa};\boldsymbol{\theta}_l)p(\boldsymbol{\kappa};\boldsymbol{\theta}_p)}{p\left(\mathbf{d};\boldsymbol{\theta}\right)}, \tag{2.1}$$

is proportional to the likelihood model times the prior model. These models are presented in greater details in the following chapter. The prior is assumed to follow a first order Markov chain, and we assume that each observation, d_n, depends on $\boldsymbol{\kappa}$. We relate the model assumptions to a hidden Markov model, as defined in Cappe et al. (2005), and Frühwirth-Schnatter (2006). We specify a Gauss-linear acquisition likelihood model, and introduce a latent response likelihood model. The response likelihood model can for example represent the log-physics response in well-log data. From the acquisition and response likelihoods we define the gross likelihood, and study the apparent convolution kernel. In the following chapter we omit the model parameter dependence to ease notation.

2.1 Prior Model

Let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_N)$ be a first order Markov chain, i.e. it satisfies

$$p(\kappa_n|\kappa_{n-1}, \ldots, \kappa_1) = p(\kappa_n|\kappa_{n-1}) \tag{2.2}$$

for $n = 2, \ldots, N$. The transition $(K \times K)$-matrix is defined as $\mathbf{P}_{\boldsymbol{\kappa}} = [p_{ij}]_{i,j\in\Omega_{\boldsymbol{\kappa}}}$, where $p_{ij} = p(\kappa_n = j|\kappa_{n-1} = i)$, is identical for all n. We assume a stationary Markov chain, i.e. the transition probabilities are independent of n, and has a stationary distribution given by

$$p_s(\boldsymbol{\kappa}) = \mathbf{P}_{\boldsymbol{\kappa}}\ p_s(\boldsymbol{\kappa}). \tag{2.3}$$

Since $\kappa_1 \sim p_s(\kappa_1)$, it follows that $\kappa_2 \sim \mathbf{P}_{\boldsymbol{\kappa}}\ p_s(\kappa_1)$, $\kappa_3 \sim \mathbf{P}_{\boldsymbol{\kappa}}^2\ p_s(\kappa_1)$ an so on. Hence, Eq. (2.3) gives the marginal distributions as

$$p(\kappa_n) = p_s(\kappa_n). \tag{2.4}$$

Thus, the marginal probability mass functions are identical for $n = 1, \ldots, N$. We define the prior model as

$$p(\boldsymbol{\kappa}) = \prod_{n=1}^{N} p(\kappa_n|\kappa_{n-1}), \tag{2.5}$$

where $p(\kappa_1|\kappa_0) = p_s(\kappa_1)$ for notational ease. Since

$$\begin{aligned} p\left(\kappa_n|\boldsymbol{\kappa}_{-n}\right) &= \frac{p\left(\boldsymbol{\kappa}\right)}{p\left(\boldsymbol{\kappa}_{-n}\right)} \\ &= \frac{p_s(\kappa_1)p(\kappa_2|\kappa_1)\cdots p(\kappa_N|\kappa_{N-1})}{p_s(\kappa_1)p(\kappa_2|\kappa_1)\cdots p(\kappa_{n+1}|\kappa_{n-1})\cdots p(\kappa_N|\kappa_{N-1})} \\ &= \frac{p(\kappa_n|\kappa_{n-1})p(\kappa_{n+1}|\kappa_n)}{p(\kappa_{n+1}|\kappa_{n-1})} \\ &= \frac{p(\kappa_{n-1})p(\kappa_n|\kappa_{n-1})p(\kappa_{n+1}|\kappa_n)}{p(\kappa_{n-1})p(\kappa_{n+1}|\kappa_{n-1})} \\ &= \frac{p\left(\kappa_{n-1},\kappa_n,\kappa_{n+1}\right)}{p\left(\kappa_{n-1},\kappa_{n+1}\right)} \\ &= p\left(\kappa_n|\kappa_{n-1},\kappa_{n+1}\right) \end{aligned} \quad , \tag{2.6}$$

each κ_n is conditionally independent of $\kappa_1, \ldots, \kappa_{n-2}, \kappa_{n+2}, \ldots, \kappa_N$ given κ_{n-1} and κ_{n+1}. In Fig. 2.1 the correlation structure of a first order Markov chain is given. Indeed, the first order Markov chain is a simple one dimensional Markov random field. Informally, the latter is defined for a random variable $\mathbf{x}$ on a lattice $\mathcal{S}$, with a neighbouhood system δ_s, if for all $s \in \mathcal{S}$

$$p\left(x_s|\mathbf{x}_{-s}\right) = p\left(x_s|x_t; t \in \delta_s\right). \tag{2.7}$$

In our case, $\mathcal{S}$ is one dimensional and identical to $\mathcal{L}_\mathcal{D}$, where for each $s \in \mathcal{S}$, $\delta_s = (s-1, s+1)$, except at the boundary.

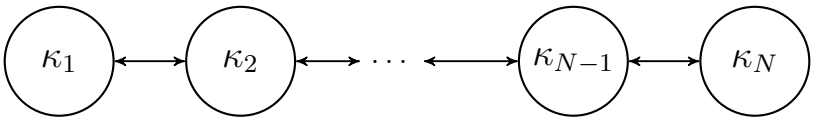


Figure 2.1: Graphical model of the correlation structure of a first order Markov chain.

The first order Markov assumption ensures a forward spatial coupling in the prior model, however also the time-reversed chain defined by

$$p\left(\boldsymbol{\kappa}\right) = p(\kappa_N)p(\kappa_{N-1}|\kappa_N)p(\kappa_{N-2}|\kappa_{N-1},\kappa_N)\ldots p(\kappa_1|\kappa_2,\ldots,\kappa_N), \tag{2.8}$$

is a first order Markov chain since

$$\begin{aligned} p(\kappa_n|\kappa_{n+1},\ldots,\kappa_N) &= \frac{p(\kappa_n)\times\prod_{i=n+1}^{N}p(\kappa_i|\kappa_{i+1})}{p(\kappa_{n+1})\times\prod_{i=n+2}^{N}p(\kappa_i|\kappa_{i+1})} \\ &= \frac{p(\kappa_n)p(\kappa_{n+1}|\kappa_n)}{p(\kappa_{n+1})} \\ &= p(\kappa_n|\kappa_{n+1}) \end{aligned} \quad . \tag{2.9}$$

The prior model for the time-reversed Markov chain is given as

$$p\left(\boldsymbol{\kappa}\right) = p_s(\kappa_N)\times\prod_{n=1}^{N-1}p(\kappa_n|\kappa_{n+1}). \tag{2.10}$$

If the stationary distribution is uniform, then the time-reversed Markov chain and original Markov chain are identically distributed.

The stationary, first-order Markov chain assumption is not critical in our approach, in fact any non-homogeneous higher order Markov chain can be used.

The prior model is completely specified by the transition matrix, $\mathbf{P}_\kappa$, thus the prior model parameters are given as $\boldsymbol{\theta}_p = \{\mathbf{P}_\kappa\}$. There are $K \times (K-1)$ unknown model parameters in the prior model since each row has to sum to unity.

2.2 Likelihood Model

We assume a gross likelihood model by introducing a latent continuous random field $\mathbf{r} = (r_1, \ldots, r_N)^\top$, where $r_n \in \mathbb{R}$ for $n = 1, \ldots, N$, as in Rimstad and Omre (2013), and Lindberg and Omre (2014a). We assume $[\mathbf{d}, \boldsymbol{\kappa}]$ to be conditionally independent given $\mathbf{r}$, i.e. $\mathbf{r}$ can be thought of as a bridge between $\boldsymbol{\kappa}$ and $\mathbf{d}$, since we assume $p(\mathbf{d}, \mathbf{r}|\boldsymbol{\kappa}) = p(\mathbf{d}|\mathbf{r})p(\mathbf{r}|\boldsymbol{\kappa})$. The likelihood models are referred to as the response model, $[\mathbf{r}|\boldsymbol{\kappa}]$, and the acquisition model, $[\mathbf{d}|\mathbf{r}]$. The gross likelihood model is given as

$$p(\mathbf{d}|\boldsymbol{\kappa}) = \int_{\mathbb{R}^N} p(\mathbf{d}|\mathbf{r})p(\mathbf{r}|\boldsymbol{\kappa})\,\mathrm{d}\mathbf{r}. \tag{2.11}$$

The latent field $\mathbf{r}$ can for example represent the logarithm of the elastic material properties, such as pressure wave velocity, shear wave velocity and density. Experience from seismic profiles indicates that $\mathbf{r}$ is a smooth field with spatial correlation. Therefore, we do not assume the elements of $\mathbf{r}$ to be conditionally independent given $\boldsymbol{\kappa}$, as studied in Rimstad and Omre (2013), and Lindberg and Omre (2014a).

We consider only so called Gauss-linear likelihood models, i.e. likelihood models that are linear in the modeling variable with additive Gaussian errors.

The gross likelihood depends on a vector of model parameters $\boldsymbol{\theta}_l = (\boldsymbol{\theta}_{l_r}, \boldsymbol{\theta}_{l_a})$, where $\boldsymbol{\theta}_{l_r}$ and $\boldsymbol{\theta}_{l_a}$ are respectively the model parameters in the response and acquisition likelihood.

2.2.1 Response Likelihood

We define the following response model,

$$[\mathbf{r}|\boldsymbol{\kappa}] = \boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}} + \mathbf{e}_{\mathbf{r}|\boldsymbol{\kappa}}, \tag{2.12}$$

where $\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}$ is a N-vector with the mean and $\mathbf{e}_{\mathbf{r}|\boldsymbol{\kappa}}$ is a N-vector with errors. The error-vector $\mathbf{e}_{\mathbf{r}|\boldsymbol{\kappa}}$ is assumed to be Gaussian with zero mean and covariance $(N \times N)$-matrix $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}$. The response likelihood is thus given as

$$p(\mathbf{r}|\boldsymbol{\kappa}) = \phi_N\left(\mathbf{r}; \boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}, \boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\right). \tag{2.13}$$

We assume the response likelihood to be stationary having mean and variance equal to

$$\begin{aligned} \mu_{r_n|\kappa_n} &= \textstyle\sum_{\kappa' \in \Omega_\kappa} \mu_{r|\kappa'} \times \mathbf{1}\{\kappa' = \kappa_n\} \\ \sigma^2_{r_n|\kappa_n} &= \textstyle\sum_{\kappa' \in \Omega_\kappa} \sigma^2_{r|\kappa'} \times \mathbf{1}\{\kappa' = \kappa_n\} \end{aligned} \qquad \text{for } n = 1, \ldots, N, \tag{2.14}$$

where $\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'} = \left(\mu_{r|\kappa'_1}, \ldots, \mu_{r|\kappa'_K}\right)^\top$ and $\boldsymbol{\sigma}^2_{r|\boldsymbol{\kappa}'} = \left(\sigma^2_{r|\kappa'_1}, \ldots, \sigma^2_{r|\kappa'_K}\right)^\top$. That is, $\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}} = \left(\mu_{r_1|\kappa_1}, \ldots, \mu_{r_N|\kappa_N}\right)^\top$. The covariance matrix is decomposed as

$$\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}} = \boldsymbol{\Sigma}^{\sigma}_{\mathbf{r}|\boldsymbol{\kappa}} \boldsymbol{\Sigma}^{\rho}_{\mathbf{r}|\boldsymbol{\kappa}} \boldsymbol{\Sigma}^{\sigma}_{\mathbf{r}|\boldsymbol{\kappa}}, \tag{2.15}$$

where $\boldsymbol{\Sigma}^{\sigma}_{\mathbf{r}|\boldsymbol{\kappa}} = \text{diag}(\sigma_{r_1|\kappa_1}, \ldots, \sigma_{r_N|\kappa_N})$ is a diagonal standard deviation $(N \times N)$-matrix. The $(N \times N)$-matrix with correlations, $\boldsymbol{\Sigma}^{\rho}_{\mathbf{r}|\boldsymbol{\kappa}}$, is defined from the correlation function, $\rho_{\mathbf{r}|\boldsymbol{\kappa}}(h)$. We propose a correlation model for the random field, $\mathbf{r}$, with a dependent mode process. The dependent mode process represents a common spatial correlation function for all mode processes,

$$[\boldsymbol{\Sigma}^{\rho}_{\mathbf{r}}]_{n,n+h} = \rho_{\mathbf{r}}(h). \tag{2.16}$$

With a dependent mode process the residuals in the Gauss mode processes are correlated. More complicated spatial correlation functions are possible, and include among others a switching process between different independent mode processes defined through an indicator function.

The marginal density of $\mathbf{r}$ is studied in greater detail, since its distributional properties are used to propose an approximation to the response likelihood. Indeed,

$$p(\mathbf{r}) = \sum_{\boldsymbol{\kappa} \in \Omega^N_{\boldsymbol{\kappa}}} \phi_N(\mathbf{r}|\boldsymbol{\kappa}) p(\boldsymbol{\kappa}) \tag{2.17}$$

is a multivariate Gaussian mixture with marginal distributions,

$$p(r_n) = \sum_{\kappa \in \Omega_\kappa} \phi_1(r_n|\kappa) p_s(\kappa) \qquad \text{for } n = 1, \ldots, N, \tag{2.18}$$

being identical Gaussian mixtures.

A graphical representation of the current response model is given in Fig. 2.2, where the arrows show the correlation structure in the prior and response models.

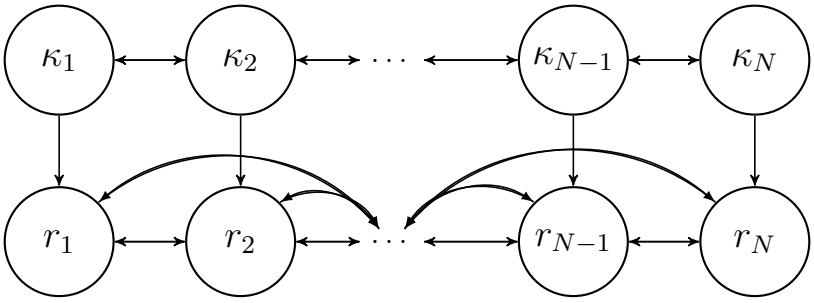


Figure 2.2: Graphical model of the current response likelihood with the spatial correlation structure.

We assume the correlation function, $\rho_{\mathbf{r}}(h)$, to be parametrized by a truncation range, a_ρ, and $\boldsymbol{\psi}_\rho$, being the functional representation of $\rho_{\mathbf{r}}(h)$. Therefore, $\boldsymbol{\Sigma}^{\rho}_{\mathbf{r}}$ is a band-diagonal matrix with bandwidth $2a_\rho + 1$. The response likelihood depends on model parameters $\boldsymbol{\theta}_{l_r} = \left(\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'}, \boldsymbol{\sigma}^2_{r|\boldsymbol{\kappa}'}, a_\rho, \boldsymbol{\psi}_\rho\right)$. Indeed, the marginal Gaussian mixtures in Eq. (2.18) are defined by the conditional their respective conditional mean and variance.

2.2.2 Acquisition Likelihood

The acquisition model represents the observational procedure, describing the data collection procedure. This can for example be either local averages, some exact observations, or relative contrasts. We define the acquisition model to be a linear model,

$$[\mathbf{d}|\mathbf{r}] = \mathbf{H}\mathbf{r} + \mathbf{e}_{\mathbf{d}|\mathbf{r}}, \tag{2.19}$$

where $\mathbf{H}$ is a general acquisition $(N_d \times N)$-matrix, and $\mathbf{e}_{\mathbf{d}|\mathbf{r}}$ is a N_d-vector with independent error. The acquisition matrix may have N_d smaller, larger, or equal to N, but in most cases $N_d \leq N$.

The acquisition likelihood is specified to be Gauss-linear, i.e. we assume $\mathbf{e}_{\mathbf{d}|\mathbf{r}}$ to be additive, independent of $\mathbf{r}$ and Gaussian, more specifically with zero mean and covariance $(N_d \times N_d)$-matrix $\mathbf{\Sigma}_{\mathbf{d}|\mathbf{r}} = \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}$. Hence,

$$p(\mathbf{d}|\mathbf{r}) = \phi_{N_d}\left(\mathbf{d}; \mathbf{H}\mathbf{r}, \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right). \tag{2.20}$$

For a fixed observational matrix $\mathbf{H}$, the acquisition likelihood is assumed to only depend on a parameter $\sigma^2_{\mathbf{d}|\mathbf{r}}$, being the observational error for each observation. The observational matrix, $\mathbf{H}$, is completely general, and may be a convolution, selection, or mixed operator. We will, however, consider only convolution operators.

A convolution arises naturally as a result of the dispersion of, for example, a physical wavelet. A convolution is a local smoothness operator which makes d_n not only dependent on r_n, but also the neighbours of r_n. In signal processing a convolution kernel is often used, since it can represent smooth functions in an efficient way.

We denote our acquisition convolution $(N \times N)$-matrix by $\mathbf{H} = \mathbf{W}$, where the acquisition convolution kernel $\mathbf{w}$ is centered at the diagonal in $\mathbf{W}$. We only consider symmetric and stationary kernels, i.e. acquisition convolution kernels which are identical for all n, except at the boundary. As Lindberg and Omre (2014a), we propose to truncate every element. Thus, each internal-node can be written as a sum,

$$d_n = \sum_{i=-a_w}^{a_w} w_i r_{n+i} + e_n \qquad \text{for } n = 1, \ldots, N. \tag{2.21}$$

Popular choices of acquisition convolution kernels are the Gaussian, the powered exponential, and Ricker wavelet, which we discretize and truncate on a grid. The truncation reduces $\mathbf{W}$ to a band-diagonal matrix with bandwidth $2a_w + 1$.

A graphical representation of the convolved acquisition likelihood, together with the prior and response models, is given in Fig. 2.3. We assume the acquisition convolution kernel to be parametrized by $\boldsymbol{\psi}_w$. Thus, the acquisition likelihood is defined by $\boldsymbol{\theta}_{l_a} = \left(a_w, \boldsymbol{\psi}_w, \sigma^2_{\mathbf{d}|\mathbf{r}}\right)$.

In for example seismic inversion the convolutional matrix $\mathbf{W}$ can be used together with a differential matrix, $\mathbf{D}$, and a matrix $\mathbf{A}$ with the angle-dependent weak Aki-Richards coefficients creating a mixed model, $\mathbf{H} = \mathbf{WAD}$. We refer to Buland and Omre (2003) for a study of such models.

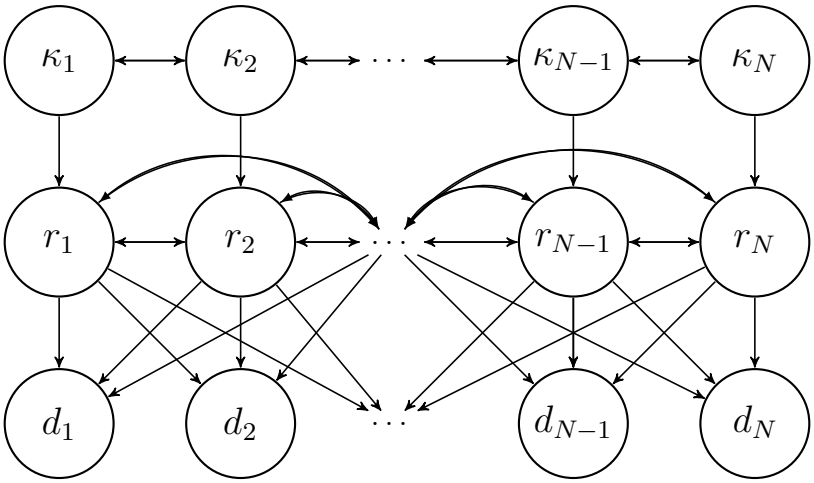


Figure 2.3: Graphical model of the current convolved model.

2.2.3 Gross Likelihood

We study the gross likelihood model, $[\mathbf{d}|\boldsymbol{\kappa}]$, in Eq. (2.11) in greater detail. As both our response and acquisition likelihood models are assumed to be Gauss-linear, the gross model

$$[\mathbf{d}|\boldsymbol{\kappa}] = \mathbf{W}\left(\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}} + \mathbf{e}_{\mathbf{r}|\boldsymbol{\kappa}}\right) + \mathbf{e}_{\mathbf{d}|\mathbf{r}}, \tag{2.22}$$

is also Gauss-linear. Thus, the gross likelihood is

$$\begin{aligned} p(\mathbf{d}|\boldsymbol{\kappa}) &= \phi_{N_d}\left(\mathbf{d}; \mathbf{W}\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}, \mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right) \\ &= \phi_{N_d}\left(\mathbf{d}; \boldsymbol{\mu}_{\mathbf{d}|\boldsymbol{\kappa}}, \boldsymbol{\Sigma}_{\mathbf{d}|\boldsymbol{\kappa}}\right) \end{aligned} . \tag{2.23}$$

As seen in Eq. (2.23), $\boldsymbol{\mu}_{\mathbf{d}|\boldsymbol{\kappa}}$ is only dependent on the acquisition convolution kernel and not the spatial correlation function $\rho_{\mathbf{r}}(h)$. Since each d_n appear as a weighted sum of $\mathbf{r}$, a short range acquisition convolution kernel ensures each d_n to be a good read of r_n. We denote this the 'shoulder effect', since a small a_w ensures that each observation d_n, determined by r_n and its neighbours, appears as a distinct shoulder in $\mathbf{d}$.

In general, the covariance matrix depends on the band matrices $\mathbf{W}$ and $\boldsymbol{\Sigma}^\rho_{\mathbf{r}}$. Therefore, also $\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}$ is a band matrix. It can be verified that $\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top$ in general results in coloured noise. We introduce the concept of an apparent convolution kernel, being the observed convolutional effect. Clearly, it is possible to fix the covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{d}|\boldsymbol{\kappa}}$, and vary $\mathbf{W}$ and $\boldsymbol{\Sigma}^\rho_{\mathbf{r}}$ accordingly. Therefore, the effect is either from the spatial correlation in the response model, or the from the acquisition convolution kernel, or both. Since

$$\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top = \boldsymbol{\Sigma}^\sigma_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}\boldsymbol{\Sigma}^\rho_{\mathbf{r}}\mathbf{W}^\top\boldsymbol{\Sigma}^\sigma_{\mathbf{r}|\boldsymbol{\kappa}}, \tag{2.24}$$

we define the apparent convolution kernel as

$$\mathbf{W}^{\mathrm{A}} = \mathbf{W}\boldsymbol{\Sigma}^{\rho\,1/2}_{\mathbf{r}}. \tag{2.25}$$

The name apparent convolution refers to the observed convolution effect through the data. If $\boldsymbol{\Sigma}^{\rho\,1/2}_{\mathbf{r}}$ and $\mathbf{W}$ are parametrized by second order exponentials, then also the apparent convolution kernel can be parametrized by a second order exponential.

In Fig. 2.4 we have simulated a latent field, $\boldsymbol{\kappa}$, and generated two set of observations from posterior models with identical posterior covariance matrix. If $\mathbf{W}^{\mathrm{A}} = \boldsymbol{\Sigma}_{\mathbf{r}}^{\rho 1/2}$, the observation appears to have distinct shoulders. On the other hand, if $\mathbf{W}^{\mathrm{A}} = \mathbf{W}$ the observations are smoothed, and the small-scale variability is lost. We have therefore reason to expect that classification of the reference profile is an easier problem if most of the apparent convolution kernel results from the spatial correlation function.

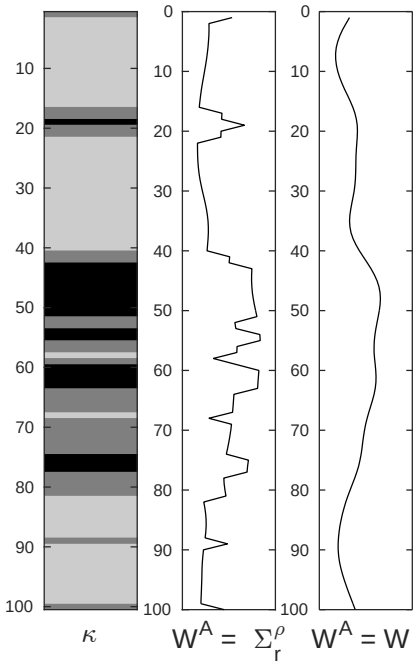


Figure 2.4: Comparison of observed data with fixed apparent convolution. Left: Reference profile. Middle: Apparent convolution kernel equals correlation function, $\boldsymbol{\Sigma}_{\mathbf{r}}^{\rho} = \mathbf{W}^{\mathrm{A}}$. Right: Apparent convolution kernel equals acquisition convolution kernel, $\mathbf{W} = \mathbf{W}^{\mathrm{A}}$.

Finally, the gross likelihood model is defined by the joint set of model parameters, $\boldsymbol{\theta}_l = (\boldsymbol{\theta}_{l_r}, \boldsymbol{\theta}_{l_a}) = \left(\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'}, \boldsymbol{\sigma}^2_{r|\boldsymbol{\kappa}'}, a_\rho, \boldsymbol{\psi}_\rho, \sigma^2_{\mathbf{d}|\mathbf{r}}, a_w, \boldsymbol{\psi}_w\right)$.

2.3 Posterior Model

As we have seen in Eq. (1.2), the posterior model is given as

$$p(\boldsymbol{\kappa}|\mathbf{d}) = \text{const} \times \phi_{N_d}\left(\mathbf{d}; \mathbf{W}\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}, \mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \boldsymbol{\sigma}^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right) \times \prod_{n=1}^{N} p(\kappa_n|\kappa_{n-1}), \tag{2.26}$$

where the normalizing constant is given as

$$\text{const} = \left[\sum_{\boldsymbol{\kappa}' \in \Omega_{\boldsymbol{\kappa}}^N} \phi_{N_d}\left(\mathbf{d}; \mathbf{W}\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}'}, \mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}'}\mathbf{W}^\top + \boldsymbol{\sigma}^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right) \times \prod_{n=1}^{N} p(\kappa'_n|\kappa'_{n-1})\right]^{-1}. \tag{2.27}$$

Calculating the normalization constant, $p(\mathbf{d})$, requires evaluating a sum including K^N permutations of $\boldsymbol{\kappa}$. It is therefore computationally infeasible to evaluate Eq. (2.26) in general. In practice the covariance matrix, $\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}$, is a band matrix with band width at most $4a_w + 2a_\rho + 1$. Note that if $\mathbf{W}$ and $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}$ are diagonal, then also the covariance matrix in Eq. (2.26) is diagonal.

A r-th order factorial form function is defined to be

$$f(x_1, \ldots, x_n) = \prod_{i=r+1}^{n} f_i(x_{i-r}, \ldots, x_i), \tag{2.28}$$

which we denote a lag-r model for $r < n$. In practice f could be a likelihood function, such that f is a product of f_i-s, being likelihood approximations. The factorial form model is related to the conditional independence structure in a model. A lag-r model defines a Markov random field with the neighbourhood determined by $\delta_i = \{i - r, \ldots, i + r\}$ for node i. Independent x_i-s corresponds to a lag-0 model, where one of the most studied lag-0 models is the hidden Markov model.

Our aim is to propose an approximation such that our posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, is on a lower order factorial form, and therefore a Markov random field. The approximate posterior model can then be exactly assessed, using the Forward-Backward algorithm. We need not approximate our prior model since it is already on factorial form. Our approximation extends previously studied models.

2.3.1 Related Models

The spatial coupling in $[\mathbf{r}|\boldsymbol{\kappa}]$ makes our response likelihood model different from the one studied in Rimstad and Omre (2013), and Lindberg and Omre (2014a). They assumed a hidden Markov model for $[\mathbf{r}|\boldsymbol{\kappa}]$, hence their response likelihood is on factorial form

$$p(\mathbf{r}|\boldsymbol{\kappa}) = \prod_{n=1}^{N} \phi_1\left(r_n; \mu_{r|\kappa'_n}, \sigma^2_{r|\kappa'_n}\right). \tag{2.29}$$

The response model studied in Rimstad and Omre (2013) is presented in Fig. 2.5, and should be compared with the current response model in Fig. 2.2. We observe that the current response model also includes spatial coupling. Their model is identical to our if $\rho_{\mathbf{r}}(h) = 0$ for all $h = 1, \ldots, N - 1$.

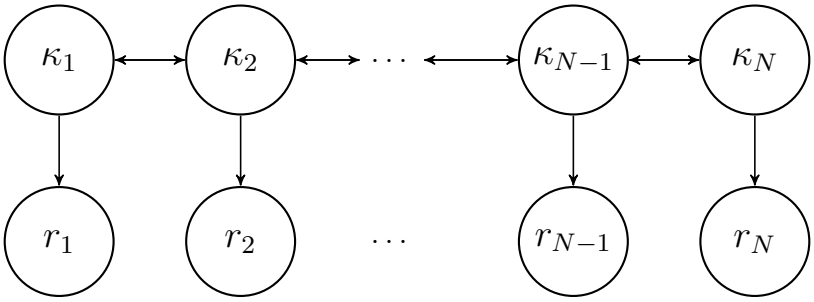


Figure 2.5: Graphical model of the response model presented in Rimstad and Omre (2013).

The current model also extends the Bernoulli-Gaussian model presented in Cheng et al. (1996), which only allows one-sided convolution. Compared to the current model, no spatial dependence is assumed in their prior and response models. A graphical model of the Bernoulli-Gaussian model is given in Fig. 2.6. In the Bernoulli-Gaussian model we can not enforce any prior spatial dependence in $\boldsymbol{\kappa}$.

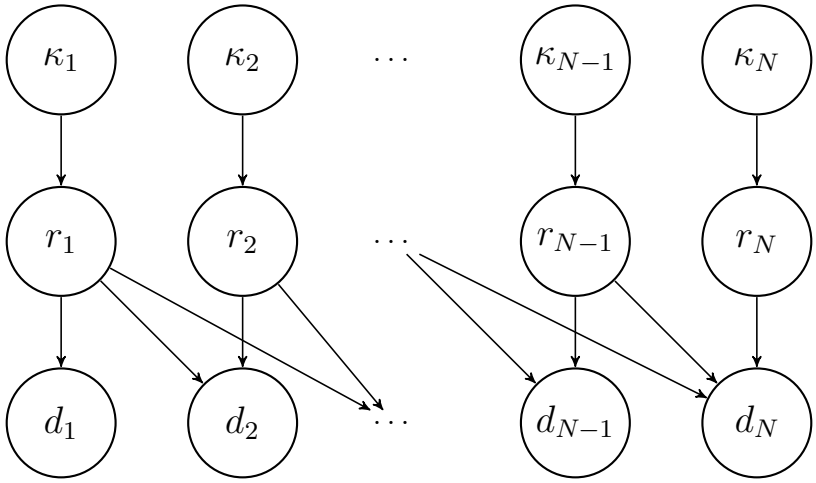


Figure 2.6: Graphical model of the Bernoulli-Gaussian model presented in Cheng et al. (1996).

Finally, we present the Gaussian mixture model by Grana and Della Rossa (2010), and later formalized by Amaliksen (2014). Instead of focusing on $\boldsymbol{\kappa}$, they studied the posterior $p(\mathbf{r}|\mathbf{d})$ by assigning a prior $p(\mathbf{r})$ to $\mathbf{r}$. They studied the continuous elastic material properties, not the hidden categorical field, $\boldsymbol{\kappa}$, representing the lithofacies. A graphical model of their model is shown in Fig. 2.7. Note that they included no spatial dependence in $\boldsymbol{\kappa}$.

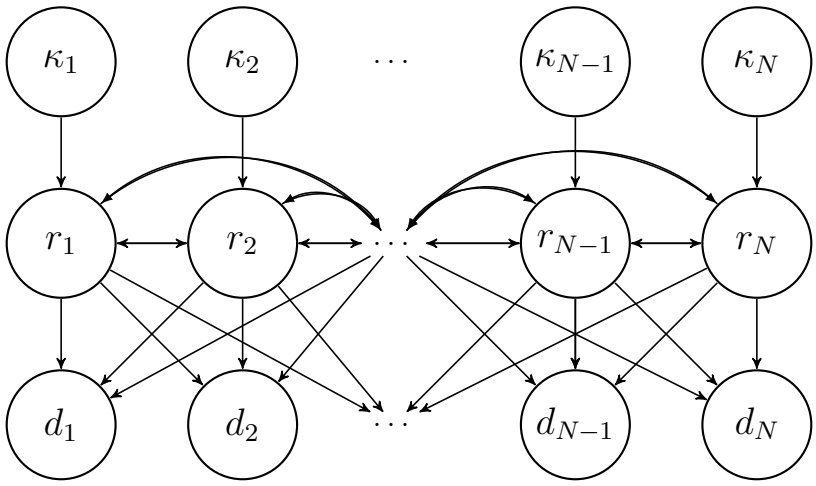


Figure 2.7: Graphical model of the Gaussian mixture model presented in Grana and Della Rossa (2010).

The imposed spatial correlation and multimodality are observed features from drilled vertical wells, see Grana and Della Rossa (2010). As we have seen in Eq. (2.17), $\mathbf{r}$ is a

multivariate Gaussian mixture model in the current model. Moreover, since

$$\begin{aligned} p(\mathbf{r}|\mathbf{d}) &= [p(\mathbf{d})]^{-1} p(\mathbf{d}|\mathbf{r})p(\mathbf{r}) \\ &= [p(\mathbf{d})]^{-1} \times \sum_{\boldsymbol{\kappa}} p(\mathbf{d}|\mathbf{r})p(\mathbf{r}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}) \\ &= \sum_{\boldsymbol{\kappa}\in\Omega_{\boldsymbol{\kappa}}^N} p(\mathbf{d}|\mathbf{r},\boldsymbol{\kappa})p(\mathbf{r}|\boldsymbol{\kappa})p(\boldsymbol{\kappa}) \left[p(\boldsymbol{\kappa},\mathbf{d})\right]^{-1} p(\boldsymbol{\kappa},\mathbf{d}) \left[p(\mathbf{d})\right]^{-1} \quad , \\ &= \sum_{\boldsymbol{\kappa}\in\Omega_{\boldsymbol{\kappa}}^N} p(\mathbf{r}|\mathbf{d},\boldsymbol{\kappa})p(\boldsymbol{\kappa}|\mathbf{d}) \end{aligned} \tag{2.30}$$

also the posterior $[\mathbf{r}|\mathbf{d}]$ is a multivariate Gaussian mixture model. In fact, Eq. (2.30) is a mixture in general for arbitrary densities $p(\boldsymbol{\kappa})$, $p(\mathbf{r}|\boldsymbol{\kappa})$ and $p(\mathbf{d}|\mathbf{r})$. If we use known results for Gaussian models, it follows that

$$p(\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}) = \phi_N\left(\mathbf{r}; \boldsymbol{\mu}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}}, \boldsymbol{\Sigma}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}}\right), \tag{2.31}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}} &= \boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}} + \boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right)^{-1} \left(\mathbf{d} - \mathbf{W}\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}\right) \\ \boldsymbol{\Sigma}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}} &= \boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}} - \boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right)^{-1} \mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}} \end{aligned} \quad . \tag{2.32}$$

If we have the posterior $p(\boldsymbol{\kappa}|\mathbf{d})$, then we also have the posterior $p(\mathbf{r}|\mathbf{d})$. We therefore only focus on assessing $p(\boldsymbol{\kappa}|\mathbf{d})$.

As we have seen, our current model generalizes the models presented here. It is possible to extend our model by assuming coloured noise in the acquisition likelihood. However, the convolution impose coloured noise in the posterior covariance matrix. Therefore, we do not choose to assume a more complicated acquisition likelihood model. The prior model, $p(\boldsymbol{\kappa})$, may also be extended to a higher order Markov chain or a non-stationary Markov chain.

Chapter 3

Posterior Assessment

The posterior model,

$$p(\boldsymbol{\kappa}|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\kappa})p(\boldsymbol{\kappa})}{p(\mathbf{d})}, \tag{3.1}$$

is computationally infeasible because of the normalization constant, $p(\mathbf{d})$. We propose to approximate the posterior model such that it can be written on factorial form, and hence be efficiently evaluated by the Forward-Backward algorithm. The simplest factorial form approximation of Eq. (3.1), corresponding to $k = 1$, is

$$p(\boldsymbol{\kappa}|\mathbf{d}) = \frac{\prod_{n=1}^{N} p(\mathbf{d}|\kappa_n)\, p(\kappa_n|\kappa_{n-1})}{p(\mathbf{d})}, \tag{3.2}$$

where the likelihood is factorized into single-site dependent factors. If we rewrite Eq. (3.2), we have

$$p(\boldsymbol{\kappa}|\mathbf{d}) = p(\kappa_1|\mathbf{d}) \times \prod_{n=2}^{N} p(\kappa_n|\kappa_{n-1}, \ldots, \kappa_1, \mathbf{d}) = p(\kappa_1|\mathbf{d}) \times \prod_{i=2}^{N} p(\kappa_n|\kappa_{n-1}, \mathbf{d}). \tag{3.3}$$

Indeed, the last equality in Eq. (3.3) holds since

$$\begin{aligned} p(\kappa_n|\kappa_{n-1}, \ldots, \kappa_1, \mathbf{d}) &\propto p(\kappa_1, \ldots, \kappa_n|\mathbf{d}) \\ &\propto \prod_{i=1}^{n} p(\mathbf{d}|\kappa_i) p(\kappa_i|\kappa_{i-1}) \\ &\propto p(\mathbf{d}, \kappa_n|\kappa_{n-1}) \\ &\propto p(\kappa_n|\kappa_{n-1}, \mathbf{d}) \end{aligned} \quad . \tag{3.4}$$

Hence, κ_n depend only on $\mathbf{d}, \kappa_1, \ldots, \kappa_{n-1}$ through κ_{n-1} and $\mathbf{d}$. Therefore, Eq. (3.3) constitutes a first order non-stationary Markov chain. The posterior transition probabilites being conditional on the observations are however no longer a homogenous Markov chain.

For higher order k approximations, let $\boldsymbol{\kappa}_n^{(k)} = (\kappa_{n-k+1}, \ldots, \kappa_n)$ be the k-th order state. Our previous first order Markov chain is now rephrased as a k-th order Markov chain,

$$\boldsymbol{\kappa}^{(k)} = \big((\kappa_1, \ldots, \kappa_k), \ldots, (\kappa_{N-k+1}, \ldots, \kappa_N)\big), \tag{3.5}$$

with a transition $(K^k \times K^k)$-matrix, $\mathbf{P}_{\boldsymbol{\kappa}}^{(k)}$. The elements are given as

$$p\left(\boldsymbol{\kappa}_n^{(k)}|\tilde{\boldsymbol{\kappa}}_{n-1}^{(k)}\right) = p(\kappa_n|\tilde{\kappa}_{n-1}) \times \prod_{i=1}^{k-1} \mathbf{1}\left\{\kappa_{n-k+i} = \tilde{\kappa}_{n-k+i}\right\}. \tag{3.6}$$

In order for the model to be consistent, the $(k-1)$ top mode labels in $\boldsymbol{\kappa}_{n-1}^{(k)}$ must equal the $(k-1)$ bottom mode labels in $\boldsymbol{\kappa}_n^{(k)}$. Therefore, we need not store the full transition matrix $\mathbf{P}_{\boldsymbol{\kappa}^{(k)}}$. Similarly,

$$\begin{aligned} p\left(\kappa_n|\tilde{\boldsymbol{\kappa}}_{n-1}^{(k)}\right) &= \sum_{\boldsymbol{\kappa}_{n-1}^{(k-1)}} p\left(\boldsymbol{\kappa}_n^{(k)}|\tilde{\boldsymbol{\kappa}}_{n-1}^{(k-1)}\right) \\ &= p\left(\kappa_n|\tilde{\kappa}_{n-1}\right) \sum_{\boldsymbol{\kappa}_{n-1}^{(k-1)}} \prod_{i=1}^{k-1} \mathbf{1}\left\{\kappa_{n-k+i} = \tilde{\kappa}_{n-k+i}\right\} \\ &= p\left(\kappa_n|\tilde{\kappa}_{n-1}\right) \end{aligned} \quad , \tag{3.7}$$

since there is only one $\boldsymbol{\kappa}_{n-1}^{(k-1)}$ such that $\prod_{i=1}^{k-1} \mathbf{1}\left\{\kappa_{n-k+i} = \tilde{\kappa}_{n-k+i}\right\} = 1$. Indeed, the prior

$$p\left(\boldsymbol{\kappa}^{(k)}\right) = \prod_{n=k}^{N} p\left(\boldsymbol{\kappa}_n^{(k)}|\tilde{\boldsymbol{\kappa}}_{n-1}^{(k)}\right) = \prod_{n=k}^{N} \left(\prod_{i=1}^{k-1} \mathbf{1}\{\kappa_{n-k+i} = \tilde{\kappa}_{n-k+i}\}\right) \times p\left(\kappa_n|\tilde{\kappa}_{n-1}\right), \tag{3.8}$$

is still defined by the transition matrix $\mathbf{P}_{\boldsymbol{\kappa}}$.

Our likelihood approximation is inspired by Rimstad and Omre (2013), i.e. we seek a likelihood approximation on factorial form,

$$p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}^{(k)}\right) = \prod_{n=k}^{N} p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right). \tag{3.9}$$

This is of the same form as for $k = 1$, hence the likelihood approximations presented later are valid for all k. If we combine Eq. (3.8) and Eq. (3.9), we can approximate Eq. (3.1) with

$$p^{(k)}\left(\boldsymbol{\kappa}^{(k)}|\mathbf{d}\right) = \text{const} \times \prod_{n=k}^{N} p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right) p\left(\boldsymbol{\kappa}_n^{(k)}|\boldsymbol{\kappa}_{n-1}^{(k)}\right), \tag{3.10}$$

where $p\left(\boldsymbol{\kappa}_k^{(k)}|\boldsymbol{\kappa}_{k-1}^{(k)}\right) = p_s\left(\boldsymbol{\kappa}_k^{(k)}\right)$ for notational ease. Thus, Eq. (3.10) is a k-th order Markov chain with respect to $\boldsymbol{\kappa}_n^{(k)}$. The approximate posterior model in Eq. (3.10) is on lag-$(k-1)$ factorial form. The approximate posterior model is given as

$$p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d}\right) = \text{const} \times \prod_{n=k}^{N} p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right) p\left(\boldsymbol{\kappa}_n^{(k)}|\boldsymbol{\kappa}_{n-1}^{(k)}\right), \tag{3.11}$$

and is a factorial form model of lag-$(k-1)$ for $k \geq 2$.

We present two different likelihood approximations to $p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right)$, namely the truncation and projection approximation. The Forward-Backward algorithm is derived in Section 3.2. In Section 3.3 the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, is assessed using the approximate posterior model, $p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d}\right)$, in an iterative McMC MH-algorithm.

3.1 Likelihood Approximations

We define two different likelihood approximations to Eq. (3.9), namely the truncation and projection based approximations. Define the k-th order truncations $\mathbf{r}_t^{(k)} = (r_{t-k+1}, \ldots, r_t)^\top$ and $\mathbf{d}_t^{(k)} = (d_{t-k+1}, \ldots, d_t)^\top$ for $n = k, \ldots, N$. In both approximations we need the marginal versions of $p(\mathbf{r}|\boldsymbol{\kappa})$, and we approximate the acquisition likelihood, $p(\mathbf{d}|\mathbf{r})$, by either truncation or projection. We present the marginal response likelihoods, since they are identical for both approximations.

The response model, $[\mathbf{r}|\boldsymbol{\kappa}]$, is Gaussian by assumption, hence from marginalization also $[\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}]$ for $n = k, \ldots, N$ are Gaussian. The mean, $\boldsymbol{\mu}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}}$, and covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}}$, are found by extracting the appropriate rows and columns from $\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}$ and $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}$. By conditional independence it follows that

$$p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}\right) = p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}, \kappa_1, \ldots, \kappa_{n-k}, \kappa_n, \ldots, \kappa_N\right) = p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right), \tag{3.12}$$

see Section 2.2.1, which is an exact expression.

3.1.1 Truncation

We present the truncation approximation for a convolutional acquisition likelihood model. It is, however, possible to generalize our approach to a general acquisition likelihood model. Since

$$p(\mathbf{d}|\mathbf{r}) = \prod_{n=1}^{N} p(d_n|\mathbf{r}), \tag{3.13}$$

we define $\mathbf{w}_n$ to be the n-th row of $\mathbf{W}$. Then,

$$p(d_n|\mathbf{r}) = \phi_1\left(d_n; \mathbf{w}_n\mathbf{r}, \sigma^2_{\mathbf{d}|\mathbf{r}}\right), \tag{3.14}$$

for $n = 1, \ldots, N$. For $k = 2k' + 1$ and $k' = 0, \ldots, N-1$, we define the band diagonal matrix $\mathbf{W}^{(k)}$ as the truncation of $\mathbf{W}$, where every element more than k' away from the diagonal element is truncated to zero. Let $\mathbf{w}_n^{(k)}$ be the n-th row in $\mathbf{W}^{(k)}$. Indeed,

$$p^{(k)}(d_n|\mathbf{r}) = \phi_1\left(d_n; \mathbf{w}_n^{(k)}\mathbf{r}, \sigma^2_{\mathbf{d}|\mathbf{r}}\right) = p^{(k)}\left(d_n|\mathbf{r}_n^{(k)}\right) \tag{3.15}$$

for $n = k+1, \ldots, N-1$. Define $\mathbf{w}_{nn}^{(k)}$ to be the subvector of length k in $\mathbf{w}_n^{(k)}$ that not being truncated, then

$$p^{(k)}\left(d_n|\mathbf{r}_n^{(k)}\right) = \phi_1\left(d_n; \mathbf{w}_{nn}^{(k)}\mathbf{r}_n^{(k)}, \sigma^2_{\mathbf{d}|\mathbf{r}}\right), \tag{3.16}$$

for $n = k+1, \ldots, N-1$, with the additional boundary terms for $n = k$ and $n = N$,

$$\begin{aligned} p^{(k)}\left(\mathbf{d}_k^{(k)}|\mathbf{r}_k^{(k)}\right) &= \phi_k\left(\mathbf{d}_k^{(k)}; \mathbf{W}_k^{(k)}\mathbf{r}_k^{(k)}, \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right) \\ p^{(k)}\left(\mathbf{d}_N^{(k)}|\mathbf{r}_N^{(k)}\right) &= \phi_k\left(\mathbf{d}_N^{(k)}; \mathbf{W}_N^{(k)}\mathbf{r}_N^{(k)}, \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right) \end{aligned} . \tag{3.17}$$

where the matrices $\mathbf{W}_k^{(k)}$ and $\mathbf{W}_N^{(k)}$ are respectively the upper left $((k'+1)\times(2k'+1))$-block matrix and lower right $((k'+1)\times(2k'+1))$-block matrix in $\mathbf{W}^{(k)}$.

Moreover, as shown in Eq. (3.12), $p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right)$ is Gaussian with mean $\boldsymbol{\mu}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}}$ for $n=k,\ldots,N$. Combined with Eq. (3.15), the k-th order marginal truncation approximation is given as

$$p^{(k)}\left(d_n|\boldsymbol{\kappa}_n^{(k)}\right)=\phi_1\left(d_n;\mathbf{w}_{nn}^{(k)}\boldsymbol{\mu}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}},\mathbf{w}_{nn}^{(k)}\boldsymbol{\Sigma}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}}\mathbf{w}_{nn}^{(k)^\top}+\sigma^2_{\mathbf{d}|\mathbf{r}}\right) \tag{3.18}$$

for $n=k+1,\ldots,N-1$. At the boundary it can be verified that

$$p^{(k)}\left(\mathbf{d}_k^{(k)}|\boldsymbol{\kappa}_k^{(k)}\right)=\phi_k\left(\mathbf{W}_k^{(k)}\boldsymbol{\mu}_{\mathbf{r}_k^{(k)}|\boldsymbol{\kappa}_k^{(k)}},\mathbf{W}_k^{(k)}\boldsymbol{\Sigma}_{\mathbf{r}_k^{(k)}|\boldsymbol{\kappa}_k^{(k)}}\mathbf{W}_k^{(k)^\top}+\sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right), \tag{3.19}$$

and similar for $p^{(k)}\left(\mathbf{d}_N^{(k)}|\boldsymbol{\kappa}_N^{(k)}\right)$. The k-th order truncation is then formally defined as

$$p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}^{(k)}\right)=p^{(k)}\left(\mathbf{d}_k^{(k)}|\boldsymbol{\kappa}_k^{(k)}\right)\times\prod_{n=k+1}^{N-1}p^{(k)}\left(d_n|\boldsymbol{\kappa}_n^{(k)}\right)\times p^{(k)}\left(\mathbf{d}_N^{(k)}|\boldsymbol{\kappa}_N^{(k)}\right). \tag{3.20}$$

If $p(\mathbf{d}|\boldsymbol{\kappa})=\prod_{n=1}^N p(d_n|\kappa_n)$, i.e. $\mathbf{W}$ and $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}$ are diagonal matrices, the method is exact for $k=1$ since Eq. (3.20) equals $p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}\right)=\prod_{n=1}^N p(d_n|\kappa_n)$. In fact the truncation approximation is exact if $\mathbf{W}=\mathbf{W}^{(k)}$ and $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}=\boldsymbol{\Sigma}^{(k)}_{\mathbf{r}|\boldsymbol{\kappa}}$, where the latter is the k-band truncation of $\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}$. It is possible to extend the truncation approximation discussed here by introducing a sliding window based on $\mathbf{W}_n^{(k)}$, and then compute $p^{(k)}\left(\mathbf{d}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right)$ for $n=k,\ldots,N$. The latter densities are then multivariate Gaussian, however they have to be scaled to ensure that the observations are used only once.

3.1.2 Projection

Consider $\mathbf{r}$, which is a multivariate Gaussian mixture,

$$p(\mathbf{r})=\sum_{\boldsymbol{\kappa}\in\Omega_\kappa^n}\phi_N\left(\mathbf{r};\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}},\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\right)p(\boldsymbol{\kappa}). \tag{3.21}$$

We propose a Gaussian approximation to $\mathbf{r}$. From the law of total expectation we have

$$\mu_r=\sum_{\kappa'\in\Omega_{\boldsymbol{\kappa}}}\mu_{r|\kappa'}\;p_s(\kappa'), \tag{3.22}$$

and we define $\boldsymbol{\mu}_{\mathbf{r}}=(\mu_r,\ldots,\mu_r)^\top$. The covariance matrix, $\boldsymbol{\Sigma}_{\mathbf{r}}$, for a dependent mode process is given as

$$\begin{aligned}[\boldsymbol{\Sigma}_{\mathbf{r}}]_{m,m+h}=\sum_{\kappa'_m\in\Omega_\kappa}\sum_{\kappa'_{m+h}\in\Omega_\kappa}\Big[&\sigma_{r|\kappa'_m}\sigma_{r|\kappa'_{m+h}}\times\rho_{\mathbf{r}}(h)\\&+\left(\mu_{r|\kappa'_m}-\mu_r\right)\left(\mu_{r|\kappa'_{m+h}}-\mu_r\right)\Big]p(\kappa'_{m+h}|\kappa'_m)\end{aligned}. \tag{3.23}$$

for $m, m+h \in \{1, \ldots, N\}$. Thus, we propose $p_*(\mathbf{r}) = \phi_N(\mathbf{r}; \boldsymbol{\mu}_\mathbf{r}, \boldsymbol{\Sigma}_\mathbf{r})$, where $\boldsymbol{\mu}_\mathbf{r}$ and $\boldsymbol{\Sigma}_\mathbf{r}$ are as given above. Since our acquisition likelihood, $p(\mathbf{d}|\mathbf{r})$, is assumed to be Gauss-linear, the approximate joint density is given as

$$p_*(\mathbf{d}, \mathbf{r}) = p(\mathbf{d}|\mathbf{r}) p_*(\mathbf{r}), \tag{3.24}$$

which is also Gaussian with

$$\begin{aligned} p_*\left(\begin{pmatrix}\mathbf{d}\\ \mathbf{r}\end{pmatrix}\right) &= \phi_{N_d+N}\left(\begin{pmatrix}\mathbf{d}\\ \mathbf{r}\end{pmatrix}; \begin{pmatrix}\mathbf{W}\boldsymbol{\mu}_\mathbf{r}\\ \boldsymbol{\mu}_\mathbf{r}\end{pmatrix}, \begin{pmatrix}\mathbf{W}\boldsymbol{\Sigma}_\mathbf{r}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I} & \mathbf{W}\boldsymbol{\Sigma}_\mathbf{r}\\ \boldsymbol{\Sigma}_\mathbf{r}\mathbf{W}^\top & \boldsymbol{\Sigma}_\mathbf{r}\end{pmatrix}\right) \\ &= \phi_{N_d+N}\left(\begin{pmatrix}\mathbf{d}\\ \mathbf{r}\end{pmatrix}; \begin{pmatrix}\boldsymbol{\mu}_\mathbf{d}\\ \boldsymbol{\mu}_\mathbf{r}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{d},\mathbf{d}} & \boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}}\\ \boldsymbol{\Gamma}^\top_{\mathbf{d},\mathbf{r}} & \boldsymbol{\Sigma}_{\mathbf{r},\mathbf{r}}\end{pmatrix}\right) \end{aligned}. \tag{3.25}$$

The marginal distributions $[\mathbf{d}, \mathbf{r}_n^{(k)}]$ are also Gaussian, and can be found by marginalization. That is, by extracting the appropriate columns and rows from the mean vector and covariance matrix in Eq. (3.25), defining $\boldsymbol{\mu}_{\mathbf{r}_n^{(k)}}$, $\boldsymbol{\Sigma}_{\mathbf{r}_n^{(k)}}$ and $\boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}}$. By conditioning on $\mathbf{r}_n^{(k)}$, we obtain the Gaussian density

$$p_*\left(\mathbf{d}|\mathbf{r}_n^{(k)}\right) = \phi_{N_d}\left(\mathbf{d}; \boldsymbol{\mu}_{\mathbf{d}|\mathbf{r}_n^{(k)}}, \boldsymbol{\Sigma}_{\mathbf{d}|\mathbf{r}_n^{(k)}}\right), \tag{3.26}$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{d}|\mathbf{r}_n^{(k)}} &= \boldsymbol{\mu}_\mathbf{d} + \boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}^{-1}_{\mathbf{r}_n^{(k)}} \left(\mathbf{r}_n^{(k)} - \boldsymbol{\mu}_{r_n^{(k)}}\right) \\ \boldsymbol{\Sigma}_{\mathbf{d}|\mathbf{r}_n^{(k)}} &= \boldsymbol{\Sigma}_{\mathbf{d},\mathbf{d}} - \boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}^{-1}_{\mathbf{r}_n^{(k)}} \boldsymbol{\Gamma}^\top_{\mathbf{d},\mathbf{r}_n^{(k)}} \end{aligned}. \tag{3.27}$$

Moreover, $p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right)$ is Gaussian with mean and covariance as discussed before. We have

$$p_*\left(\mathbf{d}, \mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right) = p_*\left(\mathbf{d}|\mathbf{r}_n^{(k)}\right) p\left(\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}\right). \tag{3.28}$$

Hence, by integrating out $\mathbf{r}_n^{(k)}$, we obtain that $p_*\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right)$ is Gaussian with

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}} &= \boldsymbol{\mu}_\mathbf{d} + \boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}^{-1}_{\mathbf{r}_n^{(k)}} \left(\boldsymbol{\mu}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}} - \boldsymbol{\mu}_{\mathbf{r}_t^{(k)}}\right) \\ \boldsymbol{\Sigma}_{\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}} &= \boldsymbol{\Sigma}_{\mathbf{d}|\mathbf{r}_t^{(k)}} + \boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}^{-1}_{\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}_{\mathbf{r}_n^{(k)}|\boldsymbol{\kappa}_n^{(k)}} \left(\boldsymbol{\Gamma}_{\mathbf{d},\mathbf{r}_n^{(k)}} \boldsymbol{\Sigma}^{-1}_{\mathbf{r}_n^{(k)}}\right)^\top \end{aligned}. \tag{3.29}$$

We therefore propose the following likelihood approximation to Eq. (3.9),

$$p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right) \overset{\text{def}}{=} \begin{cases} \left[p_*\left(\mathbf{d}|\boldsymbol{\kappa}_k^{(k)}\right)\right]^{1/k} \times \prod_{i=1}^{k-1}\left[p_*\left(\mathbf{d}|\boldsymbol{\kappa}_{k-i}^{(k-i)}\right)\right]^{1/k} & \text{if } n = k \\ \left[p_*\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right)\right]^{1/k} & \text{if } n = k+1, \ldots, N-1 \\ \left[p_*\left(\mathbf{d}|\boldsymbol{\kappa}_N^{(k)}\right)\right]^{1/k} \times \prod_{i=1}^{k-1}\left[p_*\left(\mathbf{d}|\boldsymbol{\kappa}_N^{(k-i)}\right)\right]^{1/k} & \text{if } n = N \end{cases}. \tag{3.30}$$

The k-th root in Eq. (3.30) ensures that all observations are used once, and the second terms are boundary corrections. Because of the Gaussian approximation, the projection approximation is not exact, even if $p(\mathbf{d}|\boldsymbol{\kappa}) = \prod_{n=1}^N p(d_n|\kappa_n)$.

3.1.3 Comparison of Approximations

The truncation based approximation is extremely fast for convolutional acquisition likelihood models. Indeed, it only has to extract rows from $\mathbf{W}$ and multiply matrices of low dimension. We have reason to believe that the truncation approximation is poor if a significant part of the weight in $\mathbf{w}_n$ is not covered by $\mathbf{w}_n^{(k)}$. Ideally the truncation should be of order $k = 4a_w + 2a_\rho + 1$ in order to capture the information in the likelihood, but then the assessment of the posterior model is usually computationally infeasible.

Compared to the truncation approximation discussed in Rimstad and Omre (2013), our truncation approximation is valid for models where the class response variances are dependent on $\boldsymbol{\kappa}$. That is, the various classes can be separated by both a change in the conditional variance and a shift in the conditional mean. They studied a truncation of the precision matrix in the gross likelihood,

$$p(\mathbf{d}|\boldsymbol{\kappa}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}^\top \mathbf{A}\boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}} + \boldsymbol{\mu}_{\mathbf{r}|\boldsymbol{\kappa}}^\top \mathbf{b}\right)\right), \tag{3.31}$$

where $\mathbf{A} = \mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right)^{-1}\mathbf{W}$ and $\mathbf{b} = -2\mathbf{W}^\top \left(\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}|\boldsymbol{\kappa}}\mathbf{W}^\top + \sigma^2_{\mathbf{d}|\mathbf{r}}\mathbf{I}\right)^{-1}\mathbf{d}$. They truncated $\mathbf{A}$ to a matrix $\mathbf{A}^{(k)}$ having band width k, and obtained a model on factorial form. Indeed, $p^{(k)}(\mathbf{d}|\boldsymbol{\kappa}) = \prod_{n=k}^N p\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right)$ in their model.

The projection method is inspired by Rimstad and Omre (2013). For lower order k we expect the projection approximation to be superior to the truncation approximation if the Gaussian approximation to $\mathbf{r}$ is good. This follows since more of the correlation structure is preserved in the approximated likelihood. However, the Gaussian approximation, $p_*(\mathbf{r})$, may be poor if the there is a high average or maximum discrepancy between the Gaussian mixture and Gaussian approximation.

3.2 Assessment of the Approximate Posterior Model

We present the Forward-Backward algorithm for a hidden Markov model, inspired by Künsch (2001). These recursions have been applied to switching Gaussian process, see for example Scott (2002), and Frühwirth-Schnatter (2006). Baum et al. (1970) studied parameter inference in a hidden Markov model.

We derive the Forward-Backward algorithm for a hidden Markov model, which correspond to a lag-0 model, with observations $\mathbf{y} = (y_1, \ldots, y_M)$ and a latent variable $\mathbf{x} = (1, \ldots, x_M)$. We assume that $x_m \in \Omega_{\mathbf{x}} = \{1, \ldots, C\}$ for $m = 1, \ldots, M$, and assume $\mathbf{x}$ to satisfy the first order Markov property. Each observation y_m is dependent only on x_m, thus each pair of observations in $\mathbf{y}$ is assumed to be conditionally independent given $\mathbf{x}$. The likelihood model is on factorial form,

$$p(\mathbf{y}|\mathbf{x}) = \prod_{m=1}^M p(y_m|x_m). \tag{3.32}$$

A directed acyclic graph of a hidden Markov model is given in Fig. 3.1.

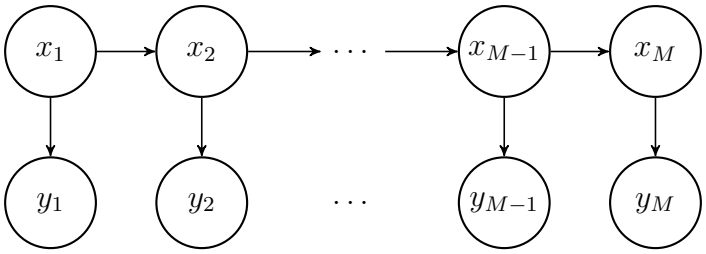


Figure 3.1: Directed acyclic graph of a hidden Markov model.

We refer to $p(x_m|y_m,\ldots,y_1)$, $p(x_{m+s}|y_m,\ldots,y_1)$ and $p(x_m|y_M,\ldots,y_1)$ as respectively the filtering, s-step prediction and smoothing density. At the initial step, the filtering density is given as

$$p(x_1|y_1) \propto p(y_1|x_1)p(x_1). \tag{3.33}$$

For $m = 2$,

$$\begin{aligned} p(x_2|y_2,y_1) &\propto \sum_{x_1} p(x_2,x_1,y_2,y_1) \\ &= \sum_{x_1} p(x_1)p(y_1|x_1)p(x_2|x_1)p(y_2|x_2) \\ &\propto \sum_{x_1} p(x_1|y_1)p(x_2|x_1)p(y_2|x_2) \end{aligned} \quad , \tag{3.34}$$

which depends on the previous filtering density, likelihood and transition probabilities. In general

$$p(x_m|y_m,\ldots,y_1) \propto \sum_{x_{m-1}} p(x_{m-1}|y_{m-1},\ldots,y_1)p(x_m|x_{m-1})p(y_m|x_m), \tag{3.35}$$

for $m = 2,\ldots,M$. Eq. (3.35) only depends on the previous filtering, likelihood and the transition probabilities, hence we can compute it recursively. Since we have to loop over all observations M, and for each $m = 1,\ldots,M$ calculate C^2 sums, the computational cost is $\mathcal{O}(M \times C^2)$. Compared to the brute-force approach, where we sum over C^N permutations, the Forward-Backward algorithm provides a significant improvement.

The one step prediction is derived as following for $m = 2$,

$$\begin{aligned} p(x_2|y_1) &\propto \sum_{x_1} p(x_2,x_1|y_1) \\ &= \sum_{x_1} p(x_1|y_1)p(x_2|x_1,y_1) \\ &= \sum_{x_1} p(x_1|y_1)p(x_2|x_1) \end{aligned} \quad , \tag{3.36}$$

which depends on the filtering density and transition probabilities. The s-step prediction is computed recursively,

$$p(x_{m+s}|y_m,\ldots,y_1) \propto \sum_{x_{m+s-1}} p(x_{m+s-1}|y_m,\ldots,y_1)p(x_{m+s}|x_{m+s-1}), \tag{3.37}$$

for $s \geq 2$. The s-step prediction depends on the $(s-1)$-step prediction and transition probabilities. Evaluation of the $(s-1)$-step predictions has a computational cost of $\mathcal{O}(s \times C^2)$. A forward step, which includes computing the filtering and prediction densities, has a computational cost of $\mathcal{O}\left((M+s) \times C^2\right)$.

The smoothing density, or backward probabilities, are given as

$$\begin{aligned} p(x_m|y_M,\ldots,y_1) &= \sum_{x_{m+1}} p(x_m, x_{m+1}|y_M,\ldots,y_1) \\ &= \sum_{x_{m+1}} p(x_m|x_{m+1}, y_M,\ldots,y_1) p(x_{m+1}|y_M,\ldots,y_1) \\ &= \sum_{x_{m+1}} p(x_m|x_{m+1}, y_m,\ldots,y_1) p(x_{m+1}|y_M,\ldots,y_1) \\ &= \sum_{x_{m+1}} \frac{p(x_m, x_{m+1}|y_m,\ldots,y_1)}{p(x_{m+1}|y_m,\ldots,y_1)} p(x_{m+1}|y_M,\ldots,y_1) \\ &= \sum_{x_{m+1}} \frac{p(x_m|y_m,\ldots,y_1) p(x_{m+1}|x_m, y_m,\ldots,y_1)}{p(x_{m+1}|y_m,\ldots,y_1)} p(x_{m+1}|y_M,\ldots,y_1) \\ &= \sum_{x_{m+1}} \frac{p(x_m|y_m,\ldots,y_1) p(x_{m+1}|x_m)}{p(x_{m+1}|y_m,\ldots,y_1)} p(x_{m+1}|y_M,\ldots,y_1) \end{aligned} \quad , \quad (3.38)$$

for $m = 1,\ldots,M$. Eq. (3.38) depends on the transition probabilities, filtering, one-step prediction and previous smoothing densities. The smoothing probabilities can be computed recursively at a cost of $\mathcal{O}(M \times C^2)$. Thus, the total cost for the Forward-Backward algorithm is $\mathcal{O}(M \times C^2)$. In practice $C << M$, thus the Forward-Backward algorithm is linear in the number of observations. An immediate consequence of Eq. (3.38) is that we can compute the joint density, $p\left(x_m, x_{m+1}|y_M,\ldots,y_1\right)$, as

$$p\left(x_m, x_{m+1}|y_M,\ldots,y_1\right) = \frac{p(x_m|y_m,\ldots,y_1) p(x_{m+1}|x_m)}{p(x_{m+1}|y_m,\ldots,y_1)} p(x_{m+1}|y_M,\ldots,y_1), \quad (3.39)$$

for $m = 1,\ldots,M-1$. Indeed, Eq. (3.39) depends only on the filtering density, transition probabilities, one-step prediction density and previous backward probabilities.

Since,

$$\begin{aligned} p\left(x_M,\ldots,x_1|y_M,\ldots,y_1\right) = p\left(x_M|y_M,\ldots,y_1\right) &\times p\left(x_{M-1}|x_M, y_M,\ldots,y_1\right) \\ &\times \cdots \times p\left(x_1|x_M,\ldots,x_2, y_M,\ldots,y_1\right) \end{aligned} \quad , \quad (3.40)$$

we can simulate sequentially from the posterior $[\mathbf{x}|\mathbf{y}]$. We compute

$$\begin{aligned} p\left(x_m|x_M,\ldots,x_{m+1}, y_M,\ldots,y_1\right) &\propto p\left(x_m|x_M,\ldots,x_{m+1}, y_m,\ldots,y_1\right) \\ &= p\left(x_m|x_{m+1}, y_m,\ldots,y_1\right) \\ &\propto p\left(x_m, x_{m+1}|y_m,\ldots,y_1\right) \\ &= p\left(x_m|y_m,\ldots,y_1\right) p\left(x_{m+1}|x_m, y_m,\ldots y_1\right) \\ &= p\left(x_m|y_m,\ldots,y_1\right) p\left(x_{m+1}|x_m\right) \end{aligned} \quad , \quad (3.41)$$

in reverse index order, i.e. by iterating from $m = M-1$ down to $m = 1$. The simulation is initialized by simulating from the last filtering density, $p\left(x_M|y_M,\ldots,y_1\right)$. Since Eq. (3.41)

only depends on the filtering density and the transition probabilities, we need not compute the smoothing density. Similarly, we can find the mode

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \left\{ p\left(x_1, \ldots, x_M | y_1, \ldots, y_M\right) \right\}, \tag{3.42}$$

sequentially by maximizing

$$\begin{aligned} \hat{x}_M &= \arg\max_{x_M} \left\{ p\left(x_M | y_1, \ldots, y_M\right) \right\} \\ &\vdots \\ \hat{x}_1 &= \arg\max_{x_1} \left\{ p\left(x_1 | \hat{x}_M, \ldots, \hat{x}_2, y_M, \ldots, y_1\right) \right\} \end{aligned} . \tag{3.43}$$

This global maximization procedure is often called the Viterbi algorithm, and utilizes dynamic programming. Similar, it is possible to define a local maximization procedure, where we maximize the marginal smoothing density,

$$\hat{\hat{\mathbf{x}}} = \left\{ \hat{\hat{x}}_m = \arg\max_{x_m} \left\{ p(x_m | y_1, \ldots, y_M) \right\}; m = 1, \ldots, M \right\}. \tag{3.44}$$

Compared to the Viterbi algorithm, only the marginal MAP (MMAP) predictor is obtained by Eq. (3.44).

We have presented the Forward-Backward algorithm for a hidden Markov model, which can adopted to our model. As discussed earlier, $p(\boldsymbol{\kappa}|\mathbf{d})$ is approximated by a non-stationary first-order Markov chain. Therefore, we only have to use the approximated likelihood instead of the exact likelihood, i.e. we use $p(\mathbf{d}|\kappa_n)$ instead of $p(y_m|x_m)$ in the Forward-Backward algorithm. The transition probabilities are given as $p(\kappa_n|\kappa_{n-1})$. For the first order approximation of the likelihood we can evaluate the approximate posterior model in $\mathcal{O}\left(N \times K^2\right)$ operations.

To assess higher order likelihood approximations we redefine the state space, i.e. let $(x_1, \ldots, x_M) = \left(\boldsymbol{\kappa}_k^{(k)}, \ldots, \boldsymbol{\kappa}_N^{(k)}\right)$ be a first order Markov chain of length $M = N - k + 1$. The Forward-Backward algorithm is extended to higher order likelihood approximations by increasing the state space. If we have two different classes, c_1 and c_2, the second order approximation contains four classes, being the four different permutations of c_1 and c_2 of length two. In Appendix B the Forward-Backward algorithm is derived formally for higher order factorial form models.

3.3 Assessment of the Correct Posterior Model

So far we have only assessed the approximate posterior model, $p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d}\right)$, as defined in Eq. (3.11). Our goal is to assess the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$. We propose to assess $p(\boldsymbol{\kappa}|\mathbf{d})$ through Markov chain Monte Carlo (McMC) sampling, using the Metropolis-Hastings (MH) algorithm. In general it is a hard problem to find a good proposal density in the McMC MH-algorithm. We propose to use the k-th order approximate posterior density as the proposal density. Since the proposal density is independent of the previous iteration, we use a so-called independent proposal McMC MH-algorithm. This have been

studied in Rimstad and Omre (2013) for a convolutional model with no spatial correlation in the response model.

The iterative procedure for generating realizations, $\boldsymbol{\kappa}_{(\text{burn-in})}, \ldots, \boldsymbol{\kappa}_{(B)}$, from the correct posterior model is given in Alg. 1. The proposal density has to be assessed through the Forward-Backward algorithm, discussed in Section 3.2. We discard the initial realizations as the burn-in period. Indeed, after the Markov chain generated by Alg. 1 has reached convergence, the realizations are from the correct posterior model. The normalization constant, $p(\mathbf{d})$, which until now has made the correct posterior model computational infeasible, cancels in the first fraction of the acceptance probability. We propose to use the MAP predictor for the approximate posterior model as the initial value in Alg. 1 to reduce the burn-in period.

Algorithm 1: Independent proposal McMC MH-algorithm

Result: Realizations $\boldsymbol{\kappa}$ from the exact posterior $p(\boldsymbol{\kappa}|\mathbf{d})$.
Initialize $\boldsymbol{\kappa}_{(0)}$ with $p\left(\boldsymbol{\kappa}_{(0)}|\mathbf{d}\right) > 0$
for $b = 1$ **to** B **do**
Propose: $\boldsymbol{\kappa}_{\text{prop}} \sim p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$
Accept/reject step:

$$\boldsymbol{\kappa}_{(i)} = \begin{cases} \boldsymbol{\kappa}_{\text{prop}} & \text{with probability } \min\left\{1, \frac{p(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}{p\left(\boldsymbol{\kappa}_{(b-1)}|\mathbf{d}\right)} \times \frac{p^{(k)}\left(\boldsymbol{\kappa}_{(b-1)}|\mathbf{d}\right)}{p^{(k)}(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}\right\} \\ \boldsymbol{\kappa}_{(b-1)} & \text{otherwise} \end{cases}$$

end
return $\left(\boldsymbol{\kappa}_{(\text{burn-in})}, \ldots, \boldsymbol{\kappa}_{(B)}\right)$

Robert and Casella (2005) show certain convergence properties of the chain, $\boldsymbol{\kappa}_{(1)}, \ldots, \boldsymbol{\kappa}_{(B)}$, from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$. The chain is irreducible and aperiodic, hence ergodic and independent of initial conditions, if and only if $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$ is almost everywhere positive on the support of $p(\boldsymbol{\kappa}|\mathbf{d})$. Since both $p(\boldsymbol{\kappa}|\mathbf{d})$ and $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$ are probability mass functions, the set of $\boldsymbol{\kappa}$ where $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}) = 0$ has measure zero. Therefore, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$ is almost everywhere positive on the support of $p(\boldsymbol{\kappa}|\mathbf{d})$. Compared to a rejection sampler, the independent proposal is more efficient in general since it on average accepts more realizations when the chain has reached its stationary distribution, see Robert and Casella (2005). Unfortunately, this comes at cost of not having independent realizations from the posterior model, since the acceptance probability is dependent on the previous realization.

We define the MMAP predictor for the correct posterior model as

$$\hat{\boldsymbol{\kappa}}_{\text{MMAP}} = \left\{ \hat{\kappa}_n^{\text{MH}} = \arg\max_{\kappa_n} \left\{ \sum_{i=1}^{B} \mathbf{1}\left(\kappa_{(i)_n} = \kappa_n\right) \right\}; \quad n = 1, \ldots, N \right\}, \tag{3.45}$$

where, $\kappa_{(i)_n}$ is the i-th realizations at location n for $n = 1, \ldots, N$. The marginal probabilities are given as

$$\hat{p}_n(j) = \left\{ \text{Prob}\{\kappa_n = j\} = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}\left\{\kappa_{n_{(i)}} = \kappa_n\right\}; \quad j \in \Omega_{\boldsymbol{\kappa}} \right\}, \tag{3.46}$$

for $n = 1, \ldots, N$. Given the approximate posterior model, generating realizations from the correct posterior model can be done extremely fast.

We introduce various distance measures to quantify the similarities between the approximate posterior model and the correct posterior model. The acceptance rate in the McMC MH-algorithm, as suggested in Rimstad and Omre (2013), is studied. If the proposal distribution, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, is close to the target distribution, $p(\boldsymbol{\kappa}|\mathbf{d})$, the acceptance probability is close to unity. The acceptance rate is formally defined as

$$\begin{aligned}\alpha &= \mathrm{E}\left\{\min\left\{1, \frac{p(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}{p(\boldsymbol{\kappa}_{\text{prev}}|\mathbf{d})} \times \frac{p^{(k)}(\boldsymbol{\kappa}_{\text{prev}}|\mathbf{d})}{p^{(k)}(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}\right\}\right\} \\ &= \sum_{\boldsymbol{\kappa}_{\text{prop}}} \sum_{\boldsymbol{\kappa}_{\text{prev}}} \min\left\{1, \frac{p^{(k)}(\boldsymbol{\kappa}_{\text{prev}}|\mathbf{d})}{p(\boldsymbol{\kappa}_{\text{prev}}|\mathbf{d})} \times \frac{p(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}{p^{(k)}(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})}\right\} p(\boldsymbol{\kappa}_{\text{prev}}|\mathbf{d})\, p^{(k)}(\boldsymbol{\kappa}_{\text{prop}}|\mathbf{d})\end{aligned}. \quad (3.47)$$

If $p(\boldsymbol{\kappa}|\mathbf{d}) = p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, then the acceptance rate is always 1. An approximation is said to be good if the acceptance rate is close to unity.

We compare the α measure to an approximate distance measure $D[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$, being a criterion to compare $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$ against $p(\boldsymbol{\kappa}|\mathbf{d})$. The approximate distance measure is defined as

$$D\left[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})\right] = \max_{\boldsymbol{\kappa}} \left\{\left|p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}) - p(\boldsymbol{\kappa}|\mathbf{d})\right|\right\}, \quad (3.48)$$

being the maximum difference between the approximate posterior model and the correct posterior model. If our approximation is exact, $D\left[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})\right] = 0$. The approximate distance measure is often referred to as the total variation distance. An approximation is said to be good, with respect to the total variation measure, if it is close to 0. A priori we have reason to believe that a high acceptance rate, α, corresponds to a small distance. However, we find the maximum difference to be a poor measure of difference since we explicitly approximate the normalization constant, which we could have used in a rejection sampler instead. Indeed, the rejection sampler gives independent realizations whereas the independent proposal McMC MH-algorithm produces correlated samples at the cost of a lower acceptance rate.

The Kullback–Leibler divergence (KLD) is defined as

$$D_{\text{KL}}\left[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})\right] = \mathrm{E}_{p^{(k)}}\left\{\log \frac{p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})}{p(\boldsymbol{\kappa}|\mathbf{d})}\right\} = \sum_{\boldsymbol{\kappa}} p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}) \times \log \frac{p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})}{p(\boldsymbol{\kappa}|\mathbf{d})}, \quad (3.49)$$

and is a distance measure often used in probability and information theory to compare distributions. The KLD measures the information lost when $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$ is used to approximate $p(\boldsymbol{\kappa}|\mathbf{d})$. Note that the KLD is defined only when $p(\boldsymbol{\kappa}|\mathbf{d}) = 0$ implies $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}) = 0$, which in our case always holds. It can be proven that the KLD is non-negative with value 0 if and only $p(\boldsymbol{\kappa}|\mathbf{d}) = p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, however the KLD is not a true metric since it is not symmetric. We have to evaluate both Eq. (3.48) and Eq. (3.49) by sampling, since they require to evaluate a sum over K^N elements. As for the total variation measure, we find the KLD measure to be poor since we explicitly estimate the normalization constant. We refer to Levin et al. (2008) for a discussion of different distance measures and their properties.

If the posterior model $[\mathbf{r}|\mathbf{d}]$ is of interest, it is straightforward to generate posterior realizations. Realizations from $[\mathbf{r}|\mathbf{d}]$ can be generated by Alg. 2, where $\boldsymbol{\mu}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}}$ and $\boldsymbol{\Sigma}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}}$ are

given in Eq. (2.32). Given realizations from $[\boldsymbol{\kappa}|\mathbf{d}]$, we can generate realizations from $[\mathbf{r}|\mathbf{d}]$ almost for free.

Algorithm 2: Generate realizations from $p(\mathbf{r}|\mathbf{d})$

Result: Generate realizations from the correct posterior response model $p(\mathbf{r}|\mathbf{d})$
$\mathbf{z}^s \sim \phi_N(\mathbf{0}, \mathbf{I})$
Generate $\boldsymbol{\kappa}^s \sim p(\boldsymbol{\kappa}|\mathbf{d})$
$\mathbf{r}^s = \boldsymbol{\mu}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}^s} + \boldsymbol{\Sigma}^{1/2}_{\mathbf{r}|\mathbf{d},\boldsymbol{\kappa}^s}\mathbf{z}^s$
return $\mathbf{r}^s$

We propose the weighted Viterbi MMAP, given as

$$\widehat{[\mathbf{r}|\mathbf{d}]} = \left\{ \hat{r}_n = \arg\max_{r_n} \left\{ \sum_{\kappa' \in \Omega_{\boldsymbol{\kappa}}} \phi_1\left(r_n; \mu_{r_n|\mathbf{d},\kappa'}, \sigma^2_{r_n|\mathbf{d},\kappa'}\right) \times p(\kappa'|\mathbf{d}) \right\}; n = 1, \ldots, N \right\}. \quad (3.50)$$

This entails N univariate optimizations, which may be done by evaluating the K modes. Similar, the confidence $(1-\alpha)$ confidence limits, $[Q_{n,1-\alpha/2}, Q_{n,\alpha/2}]$, may be found pointwise by numerical integration. They are given are given as

$$\text{Prob}\left\{Q_{n,1-\alpha/2} \leq r_n \leq Q_{n,\alpha/2}|\mathbf{d}\right\} = 1 - \alpha, \quad (3.51)$$

for $n = 1, \ldots, N$.

Chapter 4

Parameter Inference

The approximate and correct posterior models are dependent on a vector of model parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_{l_r}, \boldsymbol{\theta}_{l_a})$, respectively denoting the model parameters in the prior, response and acquisition models. We present various techniques to estimate these model parameters.

4.1 Marginal Likelihood

The maximum marginal likelihood (MML) estimates the model parameters which are most likely to have generated the observations. From Bayes' rule we have

$$p\left(\boldsymbol{\kappa}|\mathbf{d};\boldsymbol{\theta}\right) = \frac{p\left(\mathbf{d}|\boldsymbol{\kappa};\boldsymbol{\theta}\right)p\left(\boldsymbol{\kappa};\boldsymbol{\theta}\right)}{p\left(\mathbf{d};\boldsymbol{\theta}\right)}. \tag{4.1}$$

The marginal likelihood, being the normalization constant, is given as

$$p\left(\mathbf{d};\boldsymbol{\theta}\right) = \sum_{\boldsymbol{\kappa}} p\left(\mathbf{d}|\boldsymbol{\kappa};\boldsymbol{\theta}\right)p\left(\boldsymbol{\kappa};\boldsymbol{\theta}\right). \tag{4.2}$$

Such marginal likelihoods are usually hard to evaluate, and often have to be assessed through numerical methods. These methods include simulation based on McMC or optimization through the expectation-maximization algorithm.

4.1.1 Approximate Maximum Marginal Likelihood

We consider optimization of the marginal likelihood as in Lindberg and Omre (2014a). The maximum marginal likelihood estimate (MMLE) is the model parameter vector, $\boldsymbol{\theta}$, maximizing the likelihood function, randomized over $(\boldsymbol{\kappa}, \mathbf{r})$. The correct posterior model has MMLE given as

$$\hat{\boldsymbol{\theta}}_{\text{mml}} = \arg\max_{\boldsymbol{\theta}} \left\{-\log p\left(\mathbf{d};\boldsymbol{\theta}\right)\right\}. \tag{4.3}$$

In log-scale, the k-th order approximate marginal likelihood is

$$\log p^{(k)}\left(\mathbf{d};\boldsymbol{\theta}\right) = -\log \sum_{\kappa_{N-k+2}} \cdots \sum_{\kappa_N} z_{N+k-1}\left(\boldsymbol{\kappa}_N^{(k-1)}\right) = -\log z_d^{(k)}, \tag{4.4}$$

where the normalization constant is computed in the forward recursion, see Section 3.2. The k-th order approximate MMLE is

$$\hat{\boldsymbol{\theta}}_{\text{mml}}^{(k)} = \arg\max_{\boldsymbol{\theta}} \left\{ -\log p^{(k)}\left(\mathbf{d};\boldsymbol{\theta}\right) \right\}, \tag{4.5}$$

which is exact with respect to the approximate likelihood model. Optimization of Eq. (4.5) is in practice a hard problem, and dependent on the unknown model parameters. The idea is to tailor a maximization scheme such that for a sequence of model parameters, $\{\hat{\boldsymbol{\theta}}_i\}$, the sequence of log-likelihoods, $\log p^{(k)}\left(\mathbf{d};\hat{\boldsymbol{\theta}}_i\right)$, is decreasing. If the number of unknown model parameters is small, a possible procedure is to discretize each model parameter on to a grid. Eq. (4.5) is first maximized on a coarse grid, before we decrease the step size on a smaller grid. The approach is presented in Lindberg (2010), but without any constraints it is infeasible if the number of unknown model parameters is high.

Our workflow is the same as in Lindberg and Omre (2014a), and is given in Fig. 4.1. We start by initiating the unknown model parameters, $\boldsymbol{\theta}$, and then compute $p^{(k)}\left(\mathbf{d};\boldsymbol{\theta}\right)$ using the Forward-Backward algorithm. If the current marginal likelihood is not a maximum value, we update $\boldsymbol{\theta}$. After convergence, we fix $\boldsymbol{\theta}_{\text{mml}}^{(k)}$ equal to the $\boldsymbol{\theta}$ that maximizes $p^{(k)}\left(\mathbf{d};\boldsymbol{\theta}\right)$. The MAP and MMAP predictors are found based on $\hat{\boldsymbol{\theta}}_{\text{mml}}^{(k)}$.

4.1.2 Approximate Maximum Marginal A Posterior

We present a Bayesian alternative to the approximate MMLE in Section. 4.1.1. A prior distribution, $p\left(\boldsymbol{\theta}\right)$, is assigned to the vector of unknown model parameters, $\boldsymbol{\theta}$. The posterior model is then

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{4.6}$$

The mode of Eq. (4.6) is the MAP estimate, $\hat{\boldsymbol{\theta}}_{\text{map}}^{(k)}$. The marginal likelihood, $p(\mathbf{d};\boldsymbol{\theta})$, is given in Eq. 4.2, but now $\boldsymbol{\theta}$ is a random variable. Hence, the posterior $p(\boldsymbol{\theta}|\mathbf{d})$ is proportional to the likelihood, $p(\mathbf{d}|\boldsymbol{\theta})$, times the prior of $\boldsymbol{\theta}$. The k-th order approximate MAP estimate is hence given as

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{\text{map}}^{(k)} &= \arg\max_{\boldsymbol{\theta}} \left\{ p^{(k)}\left(\boldsymbol{\theta}|\mathbf{d}\right) \right\} \\
&= \arg\max_{\boldsymbol{\theta}} \left\{ \log p^{(k)}\left(\mathbf{d}|\boldsymbol{\theta}\right) + \log p(\boldsymbol{\theta}) \right) \\
&= \arg\max_{\boldsymbol{\theta}} \left\{ \log p(\boldsymbol{\theta}) - \log p^{(k)}(\mathbf{d};\boldsymbol{\theta}) \right\}
\end{aligned} \tag{4.7}$$

Optimization of the approximate marginal likelihood and approximate marginal a posterior are therefore similar problems. The approximate maximum MAP estimate should be evaluated similar as the approximate MMLE. The prior, $p(\boldsymbol{\theta})$, may depend on a set of hyperparameter, $\boldsymbol{\tau} = (\boldsymbol{\eta}, \boldsymbol{\tau}_{l_r}, \boldsymbol{\tau}_{l_a})$. These could be known, or dependent on another layer of hyperparameters.

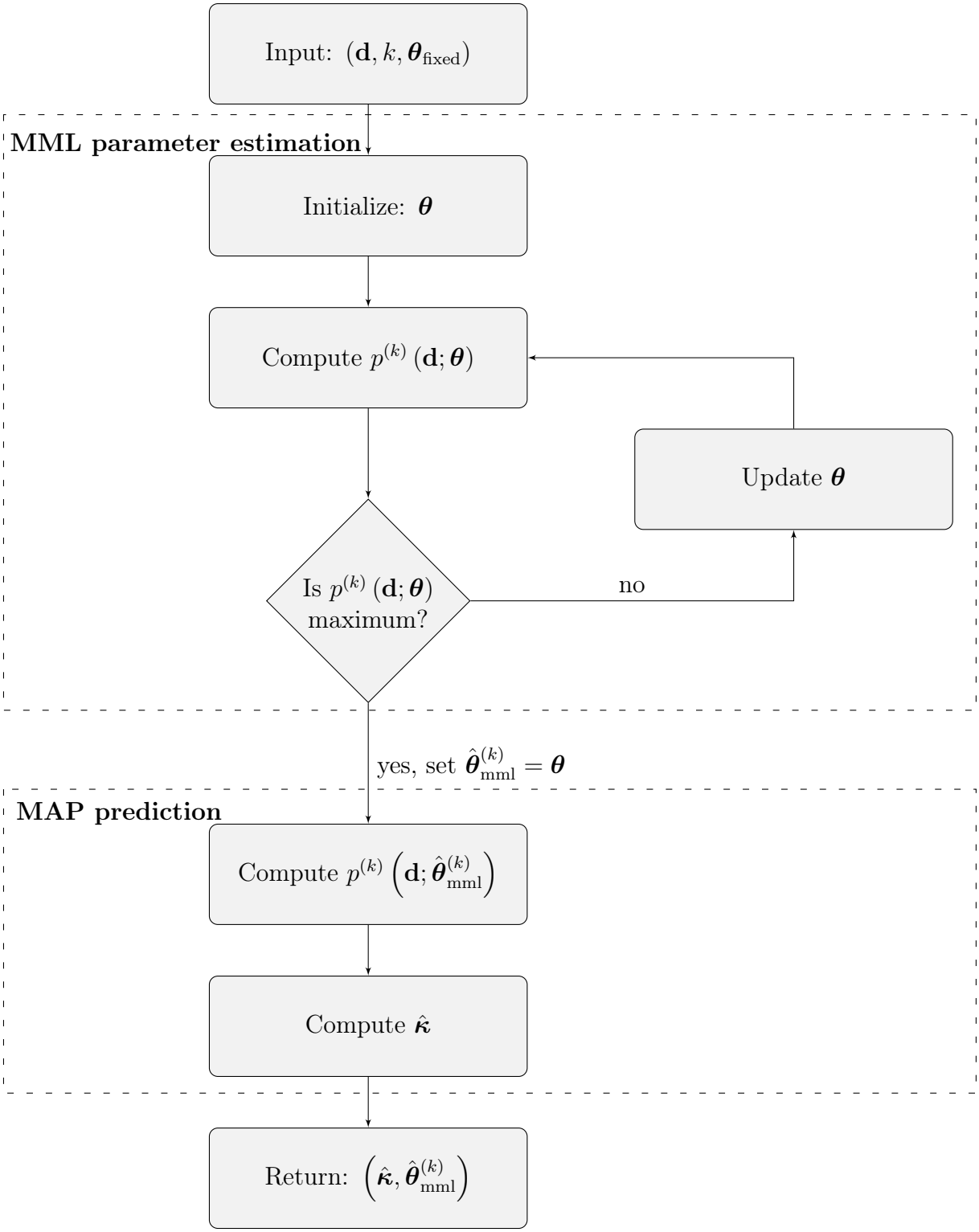


Figure 4.1: Workflow diagram of the MML parameter estimation procedure.

4.2 The Expectation-Maximization Algorithm

In the following chapter we specify densities dependent on a parameter $\boldsymbol{\theta}$ as $p(\mathbf{x}|\boldsymbol{\theta})$ instead of $p(\mathbf{x};\boldsymbol{\theta})$, to clarify the dependence on the parameter. The approximate MML given in Eq. (4.5), or equivalently the maximum MAP in Eq. (4.7), can be optimized by the expectation-maximization (EM) algorithm. The EM-algorithm was first introduced by Dempster et al. (1977) to overcome difficulties in maximizing likelihoods by introducing latent variables. A thorough introduction to the EM-algorithm is found in Hastie et al. (2001), and Robert and Casella (2005). We present the EM-algorithm in general for a univariate parameter ν.

We observe $\mathbf{x}$ independent and identically distributed from $g(x|\nu)$. Assessment of the maximum likelihood estimator $\hat{\nu} = \arg\max L(\nu|\mathbf{x})$ is often a hard problem. We introduce an augmented, or latent, variable, $\mathbf{z}$. Let the joint density be denoted by $f(\mathbf{x}, \mathbf{z}|\nu)$. Define the conditional density of the augmented variable, given the observations, as

$$h(\mathbf{z}|\mathbf{x}, \nu) = \frac{f(\mathbf{x}, \mathbf{z}|\nu)}{g(\mathbf{x}|\nu)}. \tag{4.8}$$

If we rewrite Eq. (4.8) and take the logarithm on both sides, we obtain

$$\log g(\mathbf{x}|\nu) = \log f(\mathbf{x}, \mathbf{z}|\nu) - \log h(\mathbf{z}|\mathbf{x}, \nu). \tag{4.9}$$

We take the expectation on both sides with respect to $h(\mathbf{z}|\mathbf{x}, \nu_0)$ for an arbitrary ν_0, and write the conditional densities in terms of log-likelihoods, then,

$$\begin{aligned} l(\boldsymbol{\nu}|\mathbf{x}) &= \mathrm{E}_{\boldsymbol{\nu}_0}\{l(\boldsymbol{\nu}|\mathbf{x}, \mathbf{z})\} - \mathrm{E}_{\boldsymbol{\nu}_0}\{l(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})\} \\ &= Q(\nu|\nu_0) - R(\nu_0, \nu) \end{aligned}. \tag{4.10}$$

In the E-step we compute $Q(\boldsymbol{\nu}|\boldsymbol{\nu}_0,) = \mathrm{E}_{\boldsymbol{\nu}_0}\{l(\boldsymbol{\nu}|\mathbf{x}, \mathbf{z})\}$, followed by the M-step where $Q(\boldsymbol{\nu}|\boldsymbol{\nu}_0)$ is maximized with respect to $\boldsymbol{\nu}$. We iterate between the E- and M-step until convergence. The sequence of estimators $\{\hat{\boldsymbol{\nu}}_i\}$ satisfies

$$l(\hat{\boldsymbol{\nu}}_{i+1}|\mathbf{x}) \geq l(\hat{\boldsymbol{\nu}}_i|\mathbf{x}), \tag{4.11}$$

with equality if and only if $Q(\hat{\boldsymbol{\nu}}_{i+1}|\hat{\boldsymbol{\nu}}_i) = Q(\hat{\boldsymbol{\nu}}_i|\hat{\boldsymbol{\nu}}_i)$. By definition, $Q(\hat{\boldsymbol{\nu}}_{i+1}|\hat{\boldsymbol{\nu}}_i) \geq Q(\hat{\boldsymbol{\nu}}_i|\hat{\boldsymbol{\nu}}_i)$, since $\hat{\boldsymbol{\nu}}_{i+1}$ is defined as the value $\boldsymbol{\nu}$ maximizing $Q(\boldsymbol{\nu}|\hat{\boldsymbol{\nu}}_i)$. Jensen's inequality implies that

$$\mathrm{E}_{\boldsymbol{\nu}_i}\left\{\log\left(\frac{l(\mathbf{z}|\hat{\boldsymbol{\nu}}_{i+1}, \mathbf{x})}{l(\mathbf{z}|\hat{\boldsymbol{\nu}}_i, \mathbf{x})}\right)\right\} \leq \log \mathrm{E}_{\boldsymbol{\nu}_i}\left\{\frac{p(\mathbf{z}|\hat{\boldsymbol{\nu}}_{i+1}, \mathbf{x})}{p(\mathbf{z}|\hat{\boldsymbol{\nu}}_i, \mathbf{x})}\right\} = 0. \tag{4.12}$$

Therefore, we need only to optimize $Q(\nu|\nu_0)$, not $R(\nu_0, \nu)$.

The EM-algorithm provides an increasing sequence of likelihood-values. However, we are not able to conclude that $\hat{\boldsymbol{\nu}}_i$ converges to the maximum likelihood estimator. In practice it is necessary to run the EM-algorithm multiple times with different initial values to ensure that the global maximum is found. The EM-algorithm is given in Alg. 3. Cappe et al. (2005) discuss the EM-algorithm and its convergence properties for a hidden Markov model.

In practice the E-step requires calculating the expected log-likelihood, which may constitute a hard problem. This can be overcome by estimating the expectation with a Monte Carlo step, see Wei and Tanner (1990). Their method is often referred to as Monte Carlo Expectation Maximization (MCEM).

Algorithm 3: Expectation-Maximization algorithm

Result: Maximum likelihood estimate $\hat{\boldsymbol{\nu}}$
Initialize $\hat{\boldsymbol{\nu}}_0$, $i = 1$
repeat
  E-step: $Q(\boldsymbol{\nu}|\hat{\boldsymbol{\nu}}_{i-1}) = \mathrm{E}_{\hat{\boldsymbol{\nu}}_{i-1}}\{l(\boldsymbol{\nu}|\mathbf{x}, \mathbf{z})\}$
  M-step: $\hat{\boldsymbol{\nu}}_i = \arg\max_{\boldsymbol{\nu}} Q(\boldsymbol{\nu}|\hat{\boldsymbol{\nu}}_{i-1})$
  $i = i + 1$
until convergence;
return $\hat{\boldsymbol{\nu}}$

4.3 Model Parameters

Assessment of the model parameters are in general a hard problem. We present various optimization techniques for the prior, response and acquisition model parameters separately. We refer to Chapter 3 for an introduction of the model parameters. We assign independent hyperparameters, $\boldsymbol{\eta}, \boldsymbol{\tau}_{l_r}$ and $\boldsymbol{\tau}_{l_a}$, to the prior, response and acquisition model parameters, respectively. In Fig. 4.2 the hierarchical structure of the convolutional model is given.

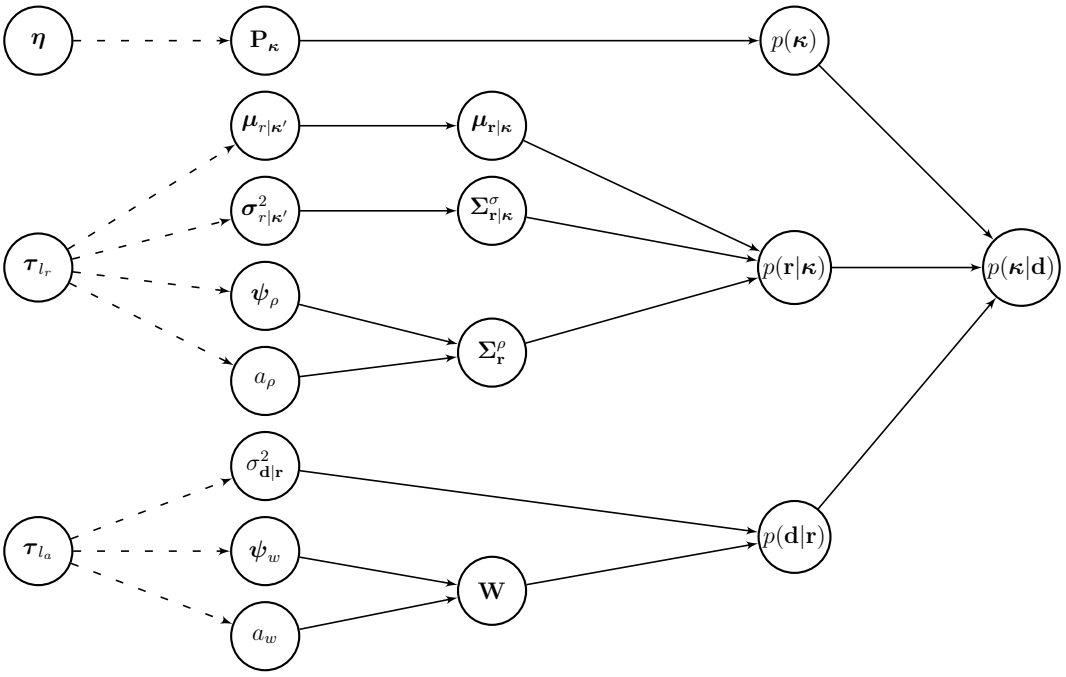


Figure 4.2: Hierarchical structure of the model parameters in the current convolutional model.

4.3.1 Prior Model Parameters

In this section we assume the likelihood model parameters, $(\boldsymbol{\theta}_{l_r}, \boldsymbol{\theta}_{l_a})$, to be known, and that the only unknown quantity is $\mathbf{P}_{\boldsymbol{\kappa}}$, where its dimension is fixed and known. If the dimension is unknown an extension of the reversible jump McMC algorithm by Green

(1995) or a Bayesian variable dimension model as defined in Robert and Casella (2005) may be used. We refer to Scott (2002), and references therein, for a discussion of the various approaches. We define the transition matrix to be

$$\mathbf{P}_{\boldsymbol{\kappa}} = \begin{pmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KK} \end{pmatrix}, \tag{4.13}$$

where the stationary distribution depends on $\mathbf{P}_{\boldsymbol{\kappa}}$. Each row in $\mathbf{P}_{\boldsymbol{\kappa}}$ must sum to unity, thus we have $K \times (K-1)$ unknown model parameters.

The k-th order approximate MMLE, presented in Section. 4.1.1, is given as

$$\hat{\boldsymbol{\theta}}^{(k)}_{\text{mml}} = \arg\max_{\boldsymbol{\theta}_p} \left\{ -\log p^{(k)}\left(\mathbf{d}; \boldsymbol{\theta}_p\right) \right\}. \tag{4.14}$$

The EM-algorithm is suitable since maximization of $p(\mathbf{d}; \boldsymbol{\theta})$ is infeasible, whereas maximization of the joint, $p(\mathbf{d}, \boldsymbol{\kappa}; \boldsymbol{\theta})$, is feasible. The EM algorithm is given with an E-step

$$Q\left(\mathbf{P}_{\boldsymbol{\kappa}} | \mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}\right) = \mathrm{E}_{p\left(\boldsymbol{\kappa}|\mathbf{d};\mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}\right)} \left\{ \log p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d}; \mathbf{P}_{\boldsymbol{\kappa}}\right) \right\}, \tag{4.15}$$

followed by a M-step

$$\mathbf{P}_{\boldsymbol{\kappa}} = \arg\max_{\mathbf{P}_{\boldsymbol{\kappa}}} \left\{ Q\left(\mathbf{P}_{\boldsymbol{\kappa}} | \mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}\right) \right\}. \tag{4.16}$$

It can be shown, see Cappe et al. (2005), that the transition matrix, $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}$, with elements

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}} = \left\{ p_{ij} = \frac{\sum_{n=2}^{N} p\left(\kappa_{n-1} = i, \kappa_n = j | \mathbf{d}; \mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}\right)}{\sum_{n=1}^{N} p\left(\kappa_n = i | \mathbf{d}; \mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}\right)}; \quad i,j = 1, \ldots, K \right\}, \tag{4.17}$$

where $\mathbf{P}^{\text{prev}}_{\boldsymbol{\kappa}}$ is the previous estimator, maximizes the Q-function in Eq. (4.16).

We consider the transition matrix through sampling, and we assign a prior distribution to the transition matrix, $\mathbf{P}_{\boldsymbol{\kappa}}$. A symmetric Dirichlet prior with identical hyperparameters, $\eta \geq 0$, is assumed, following the idea of Eidsvik et al. (2004). The Dirichlet distribution is given in Appendix A. In each iteration i we generate a realization, $\boldsymbol{\kappa}^{(i)} \sim p\left(\boldsymbol{\kappa}|\mathbf{d}, \mathbf{P}^{(i-1)}_{\boldsymbol{\kappa}}\right)$, and then generate a new transition matrix $\mathbf{P}^{(i)}_{\boldsymbol{\kappa}} \sim p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa}^{(i)}, \mathbf{d}\right)$. After the burn-in period, the joint realizations $\{\boldsymbol{\kappa}^{(i)}, \mathbf{P}^{(i)}_{\boldsymbol{\kappa}}\}$ have the correct distribution. The posterior, $p(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa}, \mathbf{d})$, depends only on $\boldsymbol{\kappa}$ since

$$p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa}, \mathbf{d}\right) \propto p(\mathbf{d}|\boldsymbol{\kappa}) p\left(\boldsymbol{\kappa}|\mathbf{P}_{\boldsymbol{\kappa}}\right) p\left(\mathbf{P}_{\boldsymbol{\kappa}}\right) \propto p\left(\boldsymbol{\kappa}|\mathbf{P}_{\boldsymbol{\kappa}}\right) p\left(\mathbf{P}_{\boldsymbol{\kappa}}\right). \tag{4.18}$$

Each row in $\mathbf{P}_{\boldsymbol{\kappa}}$ has a prior defined by

$$p(\mathbf{P}^i_{\boldsymbol{\kappa}}; \eta) = \frac{\Gamma(K\eta)}{\Gamma(\eta)^K} \times \prod_{j=1}^{K} p_{ij}^{\eta-1}, \tag{4.19}$$

where $p(\mathbf{P}^i_{\boldsymbol{\kappa}}) = p\left(p_{i1}, \ldots, p_{iK}\right)$ and $\sum_j p_{ij} = 1$ for all $i = 1, \ldots, K$. Since the Dirichlet distribution is a conjugate prior, we have

$$p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa}; \eta\right) = \frac{\Gamma\left(K\eta + \sum_{j=1}^{K} n_{ij}\right)}{\prod_{j=1}^{K} \Gamma\left(\eta + n_{ij}\right)} \times \prod_{j=1}^{K} p_{ij}^{\eta + n_{ij} - 1}, \tag{4.20}$$

where n_{ij} is the number of transitions from class i to class j in $\boldsymbol{\kappa}$.

Since the posterior $p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa},\mathbf{d}\right)$ is a Dirichlet distribution it is straightforward to generate realizations from $p(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa},\mathbf{d})$. The McMC algorithm is given in Alg. 4, inspired by Lindberg and Omre (2014b). The computational cost of each step in Alg. 4 is $\mathcal{O}\left(T\times(n-k+2)\times K^k\right)$, where T represents the computational cost of generating a realization $\boldsymbol{\kappa}$. Indeed, T may vary with the transition matrix. The computational cost of the McMC approach is infeasible if a large number of iterations are required to obtain realizations from the posterior $p\left(\boldsymbol{\kappa}|\mathbf{d};\mathbf{P}\right)$. From the posterior realizations $\mathbf{P}_{\boldsymbol{\kappa}}^{(\text{burn-in})},\ldots,\mathbf{P}_{\boldsymbol{\kappa}}^{(B)}$, we are able to assess the uncertainty in the estimates. We propose to use the mode of each p_{ij} and not the mean, since the latter is a poor estimator close to zero and one.

Algorithm 4: McMC for $\mathbf{P}_{\boldsymbol{\kappa}}$

Result: Realizations $\left(\mathbf{P}_{\boldsymbol{\kappa}}^{(\text{burn-in})},\ldots,\mathbf{P}_{\boldsymbol{\kappa}}^{(B)}\right)\sim p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa},\mathbf{d}\right)$

Initialize $\mathbf{P}_{\boldsymbol{\kappa}}^{(0)}$ with $p\left(\mathbf{P}_{\boldsymbol{\kappa}}^{(0)}|\boldsymbol{\kappa},\mathbf{d}\right)>0$

for $i=1$ **to** B **do**

Generate $\boldsymbol{\kappa}\sim p\left(\boldsymbol{\kappa}|\mathbf{d};\mathbf{P}_{\boldsymbol{\kappa}}^{(i-1)}\right)$

Generate $\mathbf{P}_{\boldsymbol{\kappa}}^{(i)}\sim p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa};\eta\right)$ according to Eq. (4.20)

end

return $\left(\mathbf{P}_{\boldsymbol{\kappa}}^{(\text{burn-in})},\ldots,\mathbf{P}_{\boldsymbol{\kappa}}^{(B)}\right)$

The k-th order MAP estimate given in Eq. (4.7) is

$$\hat{\boldsymbol{\theta}}_{\text{map}}^{(k)}=\arg\max_{\boldsymbol{\kappa}}\left\{\log p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa};\eta\right)-\log p^{(k)}(\mathbf{d};\boldsymbol{\theta})\right\}. \tag{4.21}$$

4.3.2 Response Model Parameters

In the following section we assume the prior and acquisition likelihood model parameters to be fixed and known. The class response models are Gaussian with in total $2K$ parameters describing the means and variances. The response model, $[\mathbf{r}|\boldsymbol{\kappa}]$, depends on a_ρ and $\boldsymbol{\psi}_\rho$. They describe respectively the truncation width, and model parameters in the spatial correlation function. We assume $\rho_{\mathbf{r}}$ to be parametrized by a powered exponential,

$$\rho_{\mathbf{r}}(h;\xi,\zeta)=\exp\left(-\left(\frac{h}{\xi}\right)^{\zeta}\right), \tag{4.22}$$

where ξ and ζ are the range and smoothness parameter, respectively. Hence, the response likelihood model is defined by $2K+3$ parameters. The MMLE is given as

$$\hat{\boldsymbol{\theta}}_{l_r,\text{mml}}^{(k)}=\arg\max_{\boldsymbol{\theta}_{l_r}}\left\{-\log p^{(k)}\left(\mathbf{d};\boldsymbol{\theta}_{l_r}\right)\right\}, \tag{4.23}$$

and similarly for the MAP estimate. The MAP estimate is dependent on hyperparameters $\boldsymbol{\tau}_{l_r}$. If a difference operator $\mathbf{D}$ is included in the acquisition operator, i.e. we observe

relative contrasts, only the variation and change in magnitude is possible to assess. That is, we can not assess the various means, only their relative difference.

We present optimization of Eq. (4.23) for a univariate parameter $\boldsymbol{\theta}_{l_r}$, but it extends to higher dimensions. The parameter is discretized on a lattice, $\mathcal{L}_{\boldsymbol{\theta}_{l_r}}$, with stepsize $h_{\boldsymbol{\theta}_{l_r}}$. In practice, $\mathcal{L}_{\boldsymbol{\theta}_{l_r}}$ is often restricted by the parameters itself, for example in Eq. (4.22) we restrict ourselves to $\delta \in [1, 2]$. The optimization procedure is given in Alg. 5. The maximum MAP estimate is evaluated similarly.

Algorithm 5: MMLE $\boldsymbol{\theta}_{l_r}$

Result: Maximum marginal likelihood estimate $\boldsymbol{\theta}_{l_r}$.
for $\tilde{\boldsymbol{\theta}} \in \mathcal{L}_{\boldsymbol{\theta}_{l_r}}$ **do**
    Run the Forward-Backward algorithm with model parameters $\tilde{\boldsymbol{\theta}}$
    Evaluate the marginal likelihood $\hat{l}(\mathbf{d}; \tilde{\boldsymbol{\theta}}) = -\log p(\mathbf{d}; \tilde{\boldsymbol{\theta}})$
end
return $\boldsymbol{\theta}_{l_r} = \arg\max_{\tilde{\boldsymbol{\theta}}} \left\{ \hat{l}(\mathbf{d}; \tilde{\boldsymbol{\theta}}) \right\}$

4.3.3 Parameters in the Acquisition Model

Finally, we assume $(\boldsymbol{\theta}_p, \boldsymbol{\theta}_{l_r})$ to be known. The vector of unknown model parameters is given as $\boldsymbol{\theta}_{l_a} = \left(a_w, \boldsymbol{\psi}_w, \sigma^2_{\mathbf{d}|\mathbf{r}} \right)$. In general, the dimension of $\boldsymbol{\theta}_{l_a}$ is $\dim(\boldsymbol{\psi}_w) + 2$. As for the response likelihood model, the dimension of $\boldsymbol{\psi}_w$ is in general unknown, however we assume a parametric acquisition convolution kernel. We assume the convolution kernel to be defined by a normalized powered exponential,

$$\mathbf{w}(h; \chi, \delta) \propto \exp\left\{ -\left(\frac{h}{\chi} \right)^{\delta} \right\}. \tag{4.24}$$

In Eq. (4.24), χ and δ are respectively the range and shape parameter. By assuming a parametric acquisition convolution kernel we reduce the dimensionality of the approximate likelihood function, which entails an optimization of a lower dimension

The k-th order MMLE is given as

$$\hat{\boldsymbol{\theta}}^{(k)}_{l_a,\text{mml}} = \arg\max_{\boldsymbol{\theta}_{l_a}} \left\{ -\log p^{(k)}(\mathbf{d}; \boldsymbol{\theta}_{l_a}) \right\}, \tag{4.25}$$

and similar for the MAP estimate. The MAP estimate depends on a vector of hyperparameters, $\boldsymbol{\tau}_{l_a}$. Eq. (4.25) constitues a hard optimization problem, but in can be evaluated as in Section 4.3.2. A study of parametric acquisition convolution kernels is found in Lindberg and Omre (2014a).

Chapter 5

MAP Case Studies

We compare the truncation and projection based likelihood approximations for various orders of k. The one dimensional reference profile, $\boldsymbol{\kappa}$ is displayed in Fig. 5.1, and it is assumed to be of length $n = 100$. It contains three different classes, {light-grey, dark-grey, black}. From our reference profile we generate various response models $\mathbf{r}$, given $\boldsymbol{\kappa}$. Conditioned on $\mathbf{r}$, we generate observations, $\mathbf{d}$, through the acquisition model. We study different response and acquisition models. In particular, we vary the spatial correlation function and class response variance in the response model. The apparent convolution kernel is assumed to be either a powered exponential, second order exponential, or Ricker function with different kernel widths.

We compare the MAP and MMAP predictors for the likelihood approximations for various k, and estimate the similarity measure, α. The similarity measure is a measure of similarity between the approximate and exact posterior models, see Section 3.3. Higher values of α indicate that the approximate posterior, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, is a good approximation of the correct posterior, $p(\boldsymbol{\kappa}|\mathbf{d})$. The distance measures, $D[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$ and $D_{\mathrm{KL}}[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$ are also estimated.

Sequences of 100,000 realizations from the correct posterior models, $p(\boldsymbol{\kappa}|\mathbf{d})$, are generated, using an independent proposal McMC MH-algorithm. We discard the 10,000 first realizations as a burn-in period. The McMC MH-algorithm is initiated with the MAP predictor of the approximate posterior model.

The model parameters are assumed to be fixed and known in this case study.

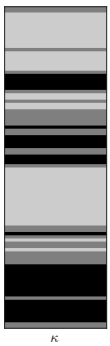


Figure 5.1: Reference profile, $\boldsymbol{\kappa}$.

5.1 Model Specification

The reference profile, $\boldsymbol{\kappa}$, with $K = 3$, is generated from a prior with the symmetric transition matrix

$$\mathbf{P}_{\kappa} = \begin{pmatrix} 0.8 & 0.2 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.2 & 0.8 \end{pmatrix}, \tag{5.1}$$

having stationary distribution $1/3 \times (1, 1, 1)^{\top}$. We see that the light-grey class and black class are not allowed to be neighbours. The time-reversed Markov chain is distributed identically to the original Markov chain since the marginal distribution is uniform, see Eq. (2.8).

The class response means are fixed to $\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'} = (-1, 0, 1)^{\top}$, and remain unchanged throughout this chapter. The variances, $\boldsymbol{\sigma}^2_{r|\boldsymbol{\kappa}'}$, are varied throughout this chapter. The test cases are defined from a spatial correlation function, $\rho_{\mathbf{r}}(h; \xi)$, and either an apparent convolution kernel, $\mathbf{w}^{\mathrm{A}}$, or an acquisition convolution kernel, $\mathbf{w}$.

We sort the various test cases by name dependent on their apparent convolution kernel, apparent kernel width, response model variances, and spatial correlation range. The name conventions are listed in Tab. 5.1. Each test case is uniquely defined by its name, and we define SE/MK/MV/MC to be the reference case. That is, the case with a second order exponential acquisition kernel, medium kernel width, medium variances in the response model, and a medium spatial correlation range.

Table 5.1: Name conventions for the MAP test case studies.

	Name	**Abbreviation**
Apparent convolution kernel type	Powered exponential	PE
	Second order exponential	SE
	Ricker exponential	RE
Apparent convolution kernel width	Short kernel	SK
	Medium kernel	MK
	Long kernel	LK
Class response variance	Low variance	LV
	Medium variance	MV
	High variance	HV
Spatial correlation range	Short correlation	SC
	Medium correlation	MC
	Long correlation	LC

The observational error is assumed to be $\sigma^2_{\mathbf{d}|\mathbf{r}} = 10^{-4}$ throughout this chapter. Since the observational error is assumed to be fixed, we define the associated signal-to-noise ratio to be

$$\mathrm{S/N} \stackrel{\mathrm{def}}{=} \frac{\mathrm{Tr}\left(\mathbf{W}^{\mathrm{A}}\right)}{N \times \sum_{\kappa' \in \Omega_{\boldsymbol{\kappa}}} \sigma^2_{r|\kappa'} p_s(\kappa')}, \tag{5.2}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. A high signal-to-noise ratio assures the observations to be a good read off from the response profile. For each $i = 1, \ldots, K$, we define

the misclassification rates as in Lindberg (2010),

$$\begin{aligned} l_i &= \frac{\sum_{n=1}^{N} \mathbf{1}\{\kappa_n^{\mathrm{r}} = i\} p^{(k)}\left(\kappa_n = i | \mathbf{d}\right)}{\sum_{n=1}^{N} \mathbf{1}\{\kappa_n^{\mathrm{r}} = i\}} \\ u_i &= \frac{\sum_{n=1}^{N} p^{(k)}\left(\kappa_n = i | \mathbf{d}\right)}{\sum_{n=1}^{N} \mathbf{1}\{\kappa_n^{\mathrm{r}} = i\}} \end{aligned}, \tag{5.3}$$

where κ_n^{r} is n-th value of the reference profile. We refer to l_i as the lower part, and it represents the ability for the approximate posterior model to correctly predict the reference profile. Similarly, u_i is the upper part, and it is defined to be the ratio between how much the posterior favors class i, compared to the reference model. Indeed, we have $l_i \leq u_i$. If a predictor is good, both l_i and u_i are close to unity.

5.1.1 Reference Case

The reference case, SE/MK/MV/MC, is studied in detail, and is later compared with the other test cases. The class response standard deviation vector is given as $\boldsymbol{\sigma}_{r|\boldsymbol{\kappa}'} = (0.7, 0.7, 0.7)^\top$. In Fig. 5.2a the class response densities are displayed with solid lines. We observe that the class response densities are overlapping. In specific, the dark-grey class is partly masked by the two other classes. We have therefore reason to believe that we underestimate the proportion of the dark-grey class. The Gaussian mixture is displayed with a dashed line, and the Gaussian approximation with a dotted line. Indeed, the multimodality of the Gaussian mixture is hard to observe. Since the Gaussian approximation is close to the Gaussian mixture, we expect the projection approximation to perform well.

The apparent convolution kernel is defined to be a second order exponential function,

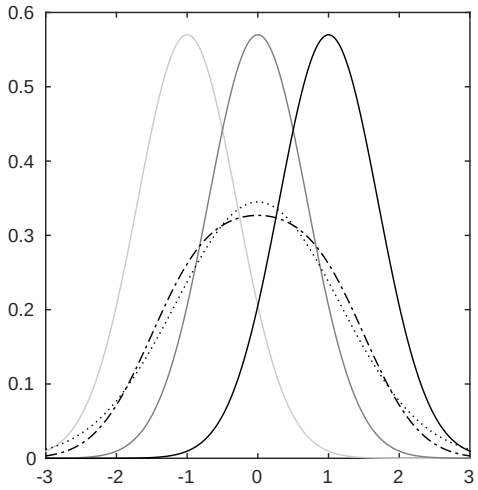
$$\mathbf{w}^{\mathrm{A}}(h) = \mathrm{const} \times \exp\left\{-\frac{1}{2} \times \left(\frac{h}{4}\right)^2\right\}, \tag{5.4}$$

which we normalize. It is displayed in Fig. 5.2b. Each observation is dependent on roughly its 35 closest neighbours in the response model.

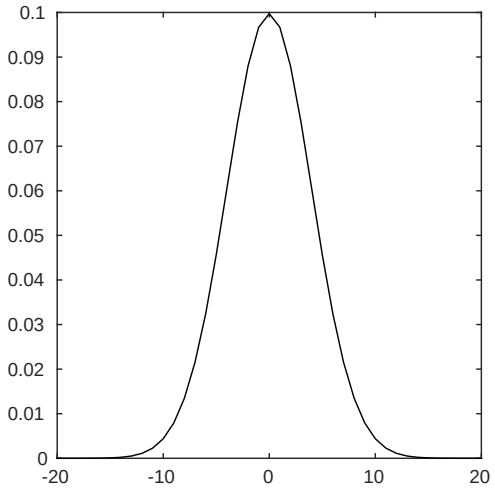
The spatial correlation function, $\rho_{\mathbf{r}}(h)$, is defined to be a powered exponential,

$$\rho_{\mathbf{r}}(h) = \exp\left\{-\frac{1}{2} \times \left(\frac{h}{2}\right)^2\right\}. \tag{5.5}$$

Together, $\mathbf{W}^{\mathrm{A}}$ and $\boldsymbol{\Sigma}_{\mathbf{r}}^{\rho}$ define the acquisition convolution operator, $\mathbf{W}$, since $\mathbf{W}^{\mathrm{A}} = \mathbf{W}\boldsymbol{\Sigma}_{\mathbf{r}}^{\rho 1/2}$. The spatial correlation function is given together with the acquisition convolution kernel in Fig. 5.3. The correlation function has an effective range of 5, and the effective width of the acquisition convolution kernel is close to 35, but the values are reduced relative to the apparent convolution kernel.

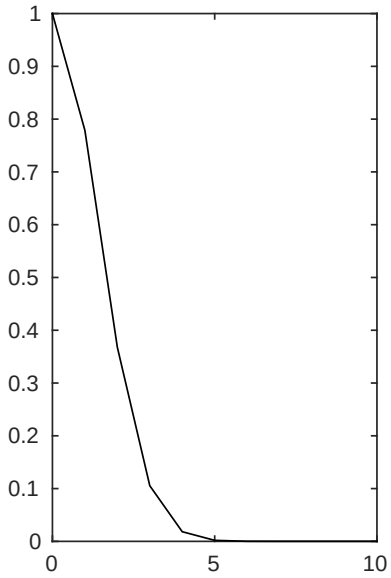


(a) Class response models



(b) Apparent convolution, $\mathbf{w}^{\mathrm{A}}$

Figure 5.2: Left (a): Class response densities with solid lines, Gaussian mixture is displayed with dashed line, and the Gaussian approximation with a dotted line. Right (b): Apparent convolution kernel, $\mathbf{w}^{\mathrm{A}}$.



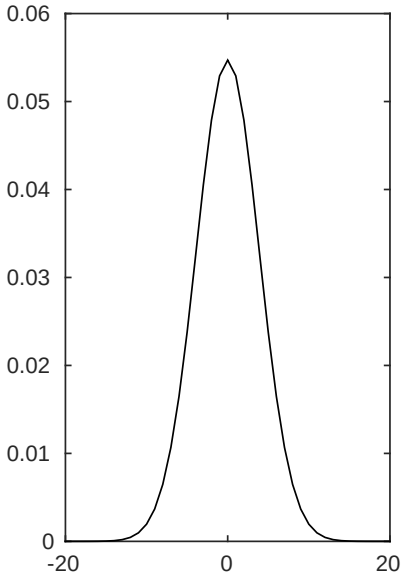


Figure 5.3: SE/MK/MV/MC: From left to right: Spatial correlation function, $\rho_{\mathbf{r}}(h)$, and acquisition convolution kernel, $\mathbf{w}$.

The signal-to-noise ratio is S/N ≈ 0.204. The response profile, $\mathbf{r}$, is shown together with the observations, $\mathbf{d}$, in Fig. 5.4. Indeed, $\mathbf{d}$ is smooth with a poor signal-to-noise ratio, and the mode jumps in $\boldsymbol{\kappa}$ are only partly identifiable through visual inspection. Therefore, we expect only to capture the main characteristics of $\boldsymbol{\kappa}$.

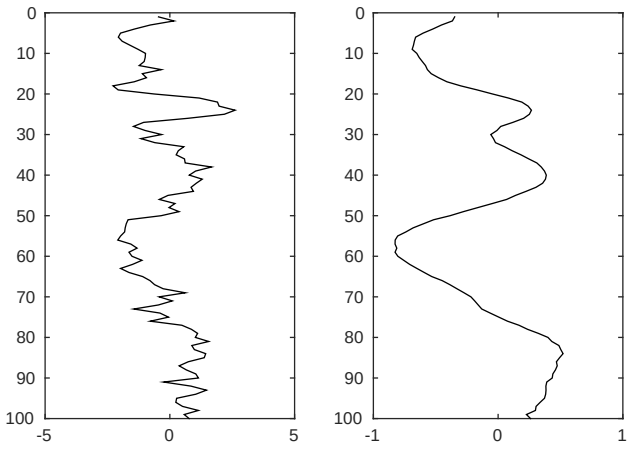


Figure 5.4: SE/MK/MV/MC: From left to right: Response profile, $\mathbf{r}$, and observations, $\mathbf{d}$.

We compare the MAP predictors for the approximate posterior models, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$. The MAP predictors are given in Fig. 5.5. The projection based approximation identifies slightly more of the small-scale variability in the reference profile than the truncation based approximation. Indeed, the MAP predictors are smooth compared to the reference profile, as they will be due to a poor signal-to-noise ratio. Both approximations have fairly stable MAP predictors for increasing values of k. A lower order approximation is observed to be sufficient if assessment of the MAP predictors is of interest.

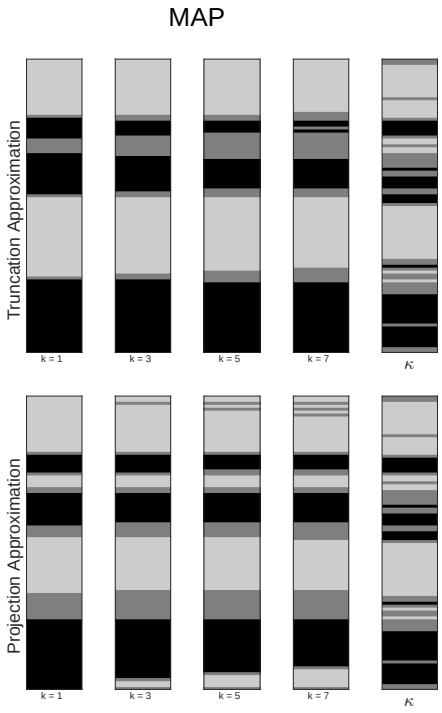


Figure 5.5: SE/MK/MV/MC: MAP predictors for the truncation (top) and projection (bottom) approximations for various order of k, together with the reference profile.

The marginal probabilities and MMAP predictors for the approximate posterior models, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, are compared in Fig. 5.6. We see that the marginal probabilities based on the projection approximation captures more of the heterogeneity than the marginal probabilities based on the truncation approximation. The MMAP predictors share the main characteristics with the MAP predictors. We find the projection approximation to perform better, since it captures rapid transitions in $\boldsymbol{\kappa}$ better than the truncation approximation. Since the MMAP predictor is a pointwise property, it is not guaranteed that the MMAP predictor provides a legal predictor. That is, the MMAP predictor can predict the light-grey and black classes to be neighbours, which has zero probability in the prior model, and hence in the approximate posterior model. This will not occur in the MAP predictor.

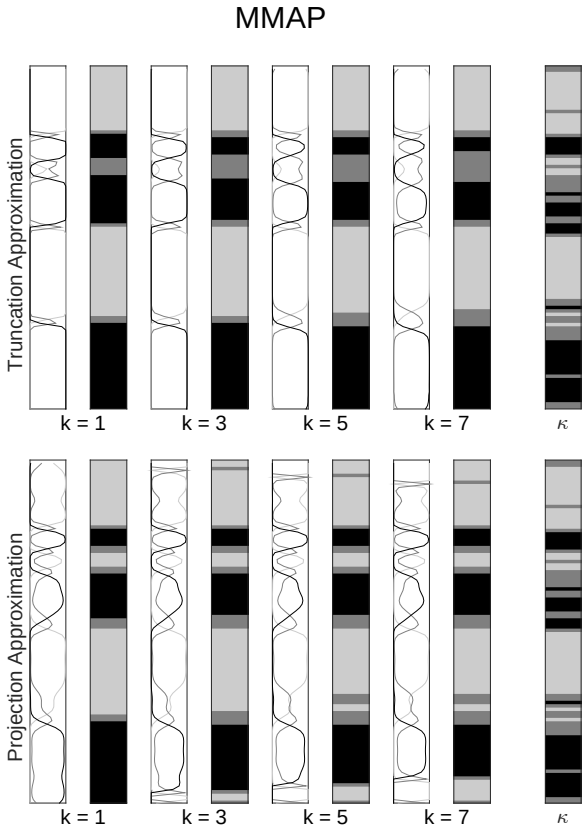


Figure 5.6: SE/MK/MV/MC: MMAP predictors and marginal probabilities for various k for respectively the truncation (top) and projection (bottom) approximations, together with the reference profile.

The misclassification coverage statistics for the approximate posterior models, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$, are given in Fig. 5.7. In the top row of displays, the truncation approximation has be used to approximate the likelihood model. From left to right, we present the misclassification statistics for classes light-grey, dark-grey and black. The misclassification rates are shown as functions of $k \in \{1, 3, 5, 7\}$, with a line segment for the corresponding misclassification coverage statistic. For example, the light-grey class has a misclassification coverage statistic of approximately $[0.87, 1.15]$ for the first order truncation based approximation. The misclassification coverage statistics for the projection approximation are given in the bottom row of displays, in an identical format as for the truncation approximation. Both

the truncation and projection approximation tend to overestimate the occurrences of the light-grey and black classes, whereas the dark-grey is severely underestimated. The coverage bands are fairly wide, which is reasonable since the class responses are chosen to be partly overlapping. We see that an increasing k is preferable for correctly classifying the middle dark-grey class, at the cost of possibly underestimating the occurrence of the black class. We observe that the projection based approximation has slightly shorter misclassification coverage statistics intervals for the light-grey and black classes than the truncation based approximation.

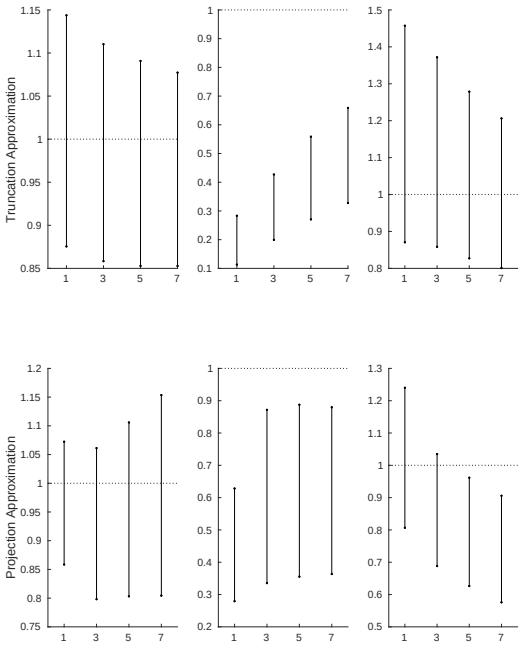


Figure 5.7: SE/MK/MV/MC: Misclassification coverage statistics. Top row: Truncation approximation. From left to right: Light-grey class, dark-grey class, and black class, as functions of k. Bottom row: Projection approximation. From left to right: Light-grey class, dark-grey class, and black class, as functions of k.

We generate 100,000 realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, using the various approximate posterior models as proposal densities in the independent proposal McMC MH-algorithm. The acceptance rates, α, of the independent proposal McMC MH-algorithm, as functions of k, are given in Fig. 5.8. For $k \geq 3$, we get a reasonable acceptance rate for the projection approximation, slightly above 10%. Higher order k is not found to have a great effect on the acceptance rates. If the likelihood is approximated by the truncation method, the acceptance rates appear to be somewhat lower, compared to the ones based on the projection method. Both the truncation and projection based approximation appear to have a slight decrease in the acceptance rate for $k = 7$, compared to $k = 5$.

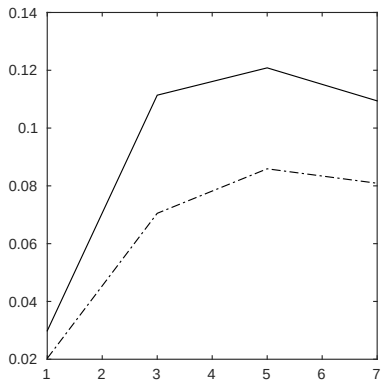
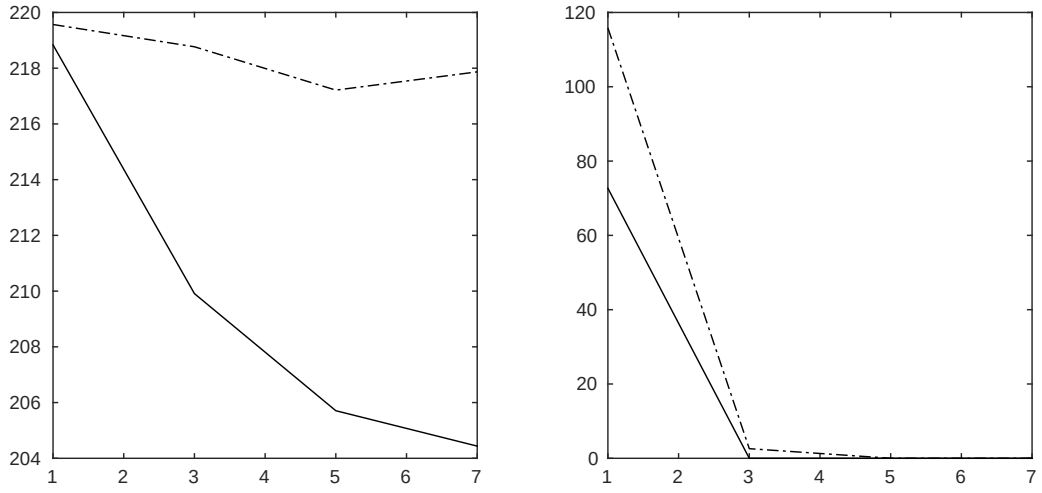


Figure 5.8: SE/MK/MV/MC: Acceptance rates as functions of k. Results based on the projection approximation is given with a solid line, and based on the truncation approximation with a dashed line.

We compare the two other distance measures defined in Section 3.3. Note that these measures are inverted relative to the acceptance rates α, since they decrease to zero when the approximate posterior model get closer to the correct posterior model. In Fig. 5.9a the log-maximum total variation distance measure, $D[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$, is given for the likelihood approximations. The results based on the truncation approximation decrease for k up to 5, and then slightly increase, while the projection approximation is strictly decreasing for increasing k. This is consistent with the acceptance rates in Fig. 5.8. The Kullback-Leibler divergence measure, $D_{\mathrm{KL}}[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$, is given in Fig. 5.9b. The distance is strictly decreasing for increasing k. We observe that by increasing the order of the approximation, the approximate posterior models get closer to the correct posterior model. Since the distance measures appear with similar behaviour, we choose only to consider the acceptance rates in the following case studies.



(a) $D[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$ (b) $D_{\mathrm{KL}}[p^{(k)}(\boldsymbol{\kappa}|\mathbf{d}), p(\boldsymbol{\kappa}|\mathbf{d})]$.

Figure 5.9: SE/MK/MV/MC: Log-maximum total variance and Kullback-Leibler divergence measures. Results for the projection approximation are given with a solid line, and a dashed line for the truncation approximation.

In Fig. 5.10, 5,000 realizations from the correct posterior model $p(\boldsymbol{\kappa}|\mathbf{d})$ are given. The realizations are generated by the independent proposal McMC MH-algorithm. The truncation approximation has been used to approximate the likelihood in the proposal density. Compared to the truncation based MAP and MMAP predictors, the posterior realizations are more heterogeneous, and they are comparable to the reference profile.

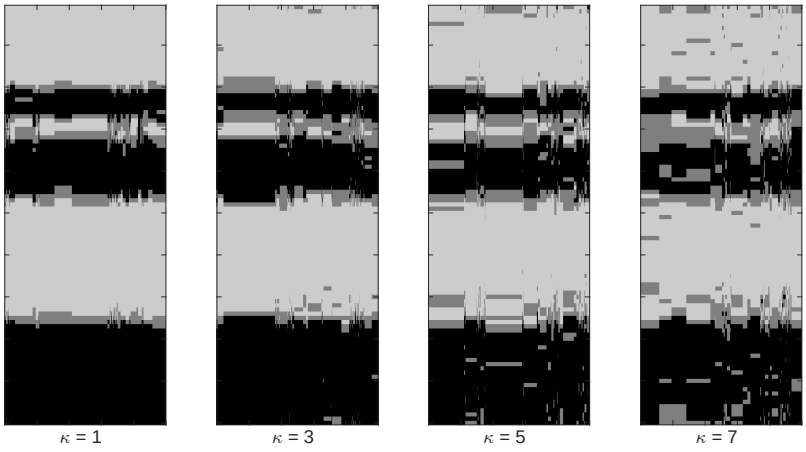


Figure 5.10: SE/MK/MV/MC: 5,000 realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, based on the truncation approximation.

In Fig. 5.11, 5,000 realizations from the correct posterior model are given, using k-th order projection approximations. These realizations share most of the characteristics with their associated MAP and MMAP predictors. Compared to the realizations in Fig. 5.10, the realizations in Fig. 5.11 appear to fluctuate more rapidly.

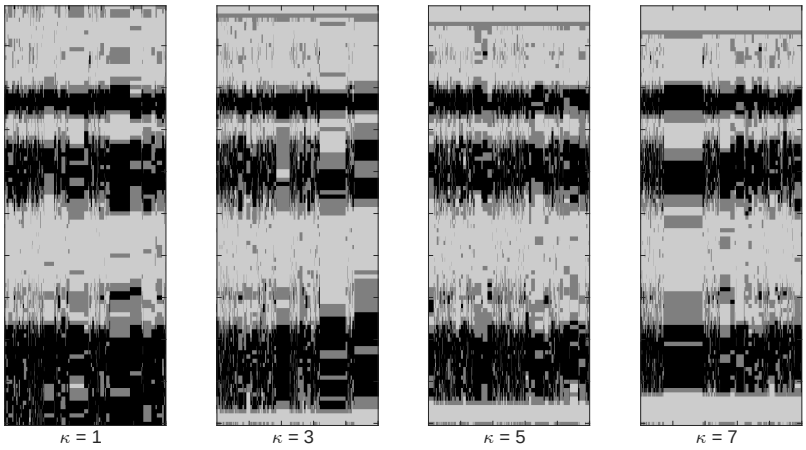


Figure 5.11: SE/MK/MV/MC: 5,000 realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, based on the projection approximation.

In Fig. 5.12 the marginal probabilities and MMAP predictor are estimated based on the realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, together with the reference profile. These realizations are generated using a seventh order approximation, where the acquisition likelihood has been approximated by the projection method. The estimated MMAP predictor based on the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, is almost identical to the corresponding exact MMAP predictor in Fig. 5.6 based on the approximate posterior model, $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d})$. The marginal probabilities appear to be less smooth than the corresponding ones in Fig. 5.6, probably due to estimation error. Indeed, the MMAP predictor is less heterogeneous than the reference profile because of the poor signal-to-noise ratio.

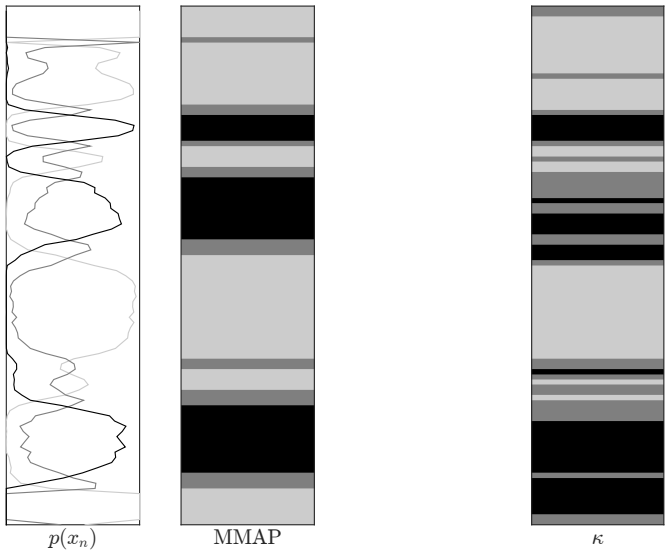


Figure 5.12: SE/MK/MV/MC: Marginal probabilities and MMAP predictor for the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, together with the reference profile, $\boldsymbol{\kappa}$.

5.1.2 Apparent Convolution Kernel

We consider various apparent convolution kernels. The model parameters, except those describing the apparent convolution kernel, are assumed to be as in Section 5.1.1. Cases PE/MK/MV/MC, SE/MK/MV/MC and RE/MK/MV/MC are studied, and we refer to them as cases one, two and three. In case one we assume the apparent convolution kernel to be a powered exponential function,

$$\mathbf{w}^{\mathrm{A}}(h) = \text{const} \times \exp\left\{-\left(\frac{|h|}{4}\right)^{1.2}\right\}. \tag{5.6}$$

Case two is the reference case as defined in Section 5.1.1, while case three is defined from a Ricker kernel,

$$\mathbf{w}^{\mathrm{A}}(h) = \text{const} \times \left(1 - \frac{h^2}{4^2}\right) \times \exp\left\{-\frac{h^2}{2 \times 4^2}\right\}. \tag{5.7}$$

We require each row in $\mathbf{W}^{\mathrm{A}}$ to sum to unity, i.e. we normalize the apparent convolution kernels. The acquisition convolution kernels for cases one, two and three, which are all symmetric, are displayed in Fig. 5.13.

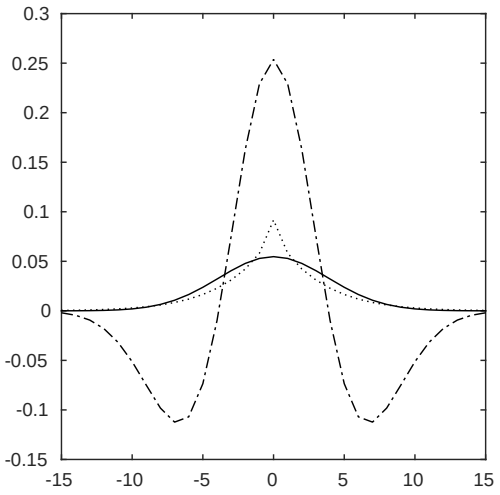


Figure 5.13: Acquisition convolution kernel for cases one to three, displayed with respectively a dotted, solid and dashed line.

The signal-to-noise ratios are 2.04, 0.204 and 0.885 for cases one, two and three, respectively.

In Fig. 5.14 and Fig. 5.15 we display the results based on the three cases. We present the figure layout in great detail. At the top row the model parameters are given. Note that the signal-to-noise ratios are different for the three cases. In the top row of displays, the spatial correlation function and acquisition convolution kernel are given in pairs for the three cases.

At the second to top display, the reference profile, $\boldsymbol{\kappa}$, is given. Each pair of response profiles, $\mathbf{r}_i$, and observations, $\mathbf{d}_i$, are then given together for cases one, two and three.

They are ordered such that each pair of realizations is beneath their respective spatial correlation function and acquisition convolution kernel. The response profiles, $\mathbf{r}_i$, are identical in the three cases. Indeed, $\mathbf{d}_1$ and $\mathbf{d}_2$ have very similar observations, but with a different degree of smoothness. We observe that $\mathbf{d}_3$ appear with the most 'shoulder effect'. Indeed, more of the small-scale variability is evident in $\mathbf{d}_3$ than in $\mathbf{d}_1$ and $\mathbf{d}_2$.

In the third row of display, the MAP predictors for the approximate posterior models are given for the truncation based approximation for various k. They are ordered such that they are beneath their respective response and observational realizations. The MAP predictors for cases one and two are almost identical for all k, and both reproduce the main characteristics of $\boldsymbol{\kappa}$. The MAP predictors in case three also reproduce some of the small-scale variability in $\boldsymbol{\kappa}$. Indeed, the MAP predictors are fairly stable as functions of k. The acceptance rate, α, is specified beneath each MAP predictor. The acceptance rate, α, is estimated from 90,000 iterations of the McMC MH-algorithm. An increase of k is observed to in general increase the acceptance rates, α.

The MAP predictors for the projection based approximation are presented in the same format as for the truncation approximation above. Compared to the truncation based approximation MAP predictors, the projection based approximation MAP predictors reproduce more of the variability in $\boldsymbol{\kappa}$. The MAP predictors are almost identical for increasing order k. In cases two and three, the acceptance rates seem to stabilize for $k \geq 3$ for the projection approximation. In case one, we see that a fifth order approximation is needed in order to obtain a reasonable acceptance rate. In fact, the truncation based approximation performs slightly better than the projection based approximation for $k = 7$ in cases one and three.

At the top row of display in Fig. 5.15, the acceptance rates are plotted as functions of k for cases one, two and three. In general, higher order approximations are observed to be preferable. The acceptance rates have similar behaviour, and the projection approximation is preferable for lower order k. Below 5,000 realizations from the correct posterior model in cases one and three are given. We observe that the realizations are more heterogeneous than the MAP predictors, as they should be. We observe that the MAP predictors in Fig. 5.14 do share most of the main characteristics with the realizations in Fig. 5.15.

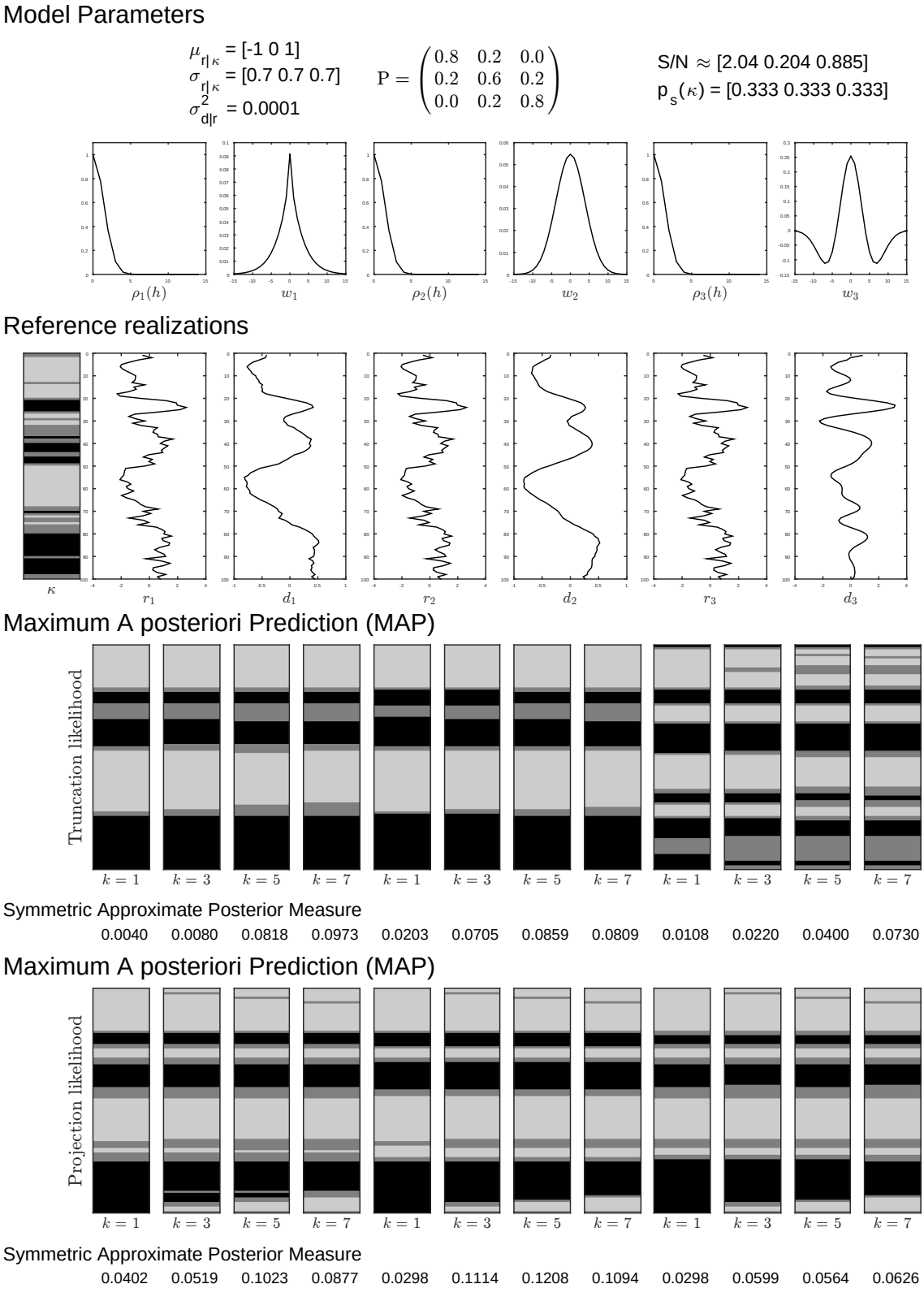


Figure 5.14: PE-SE-RE/MK/MV/MC: Model parameters, reference cases and MAP predictions/α-values for truncation and projection approximation for varying order k. The acceptance rates, α, are estimated from 90,000 iterations from the McMC MH-algorithm.

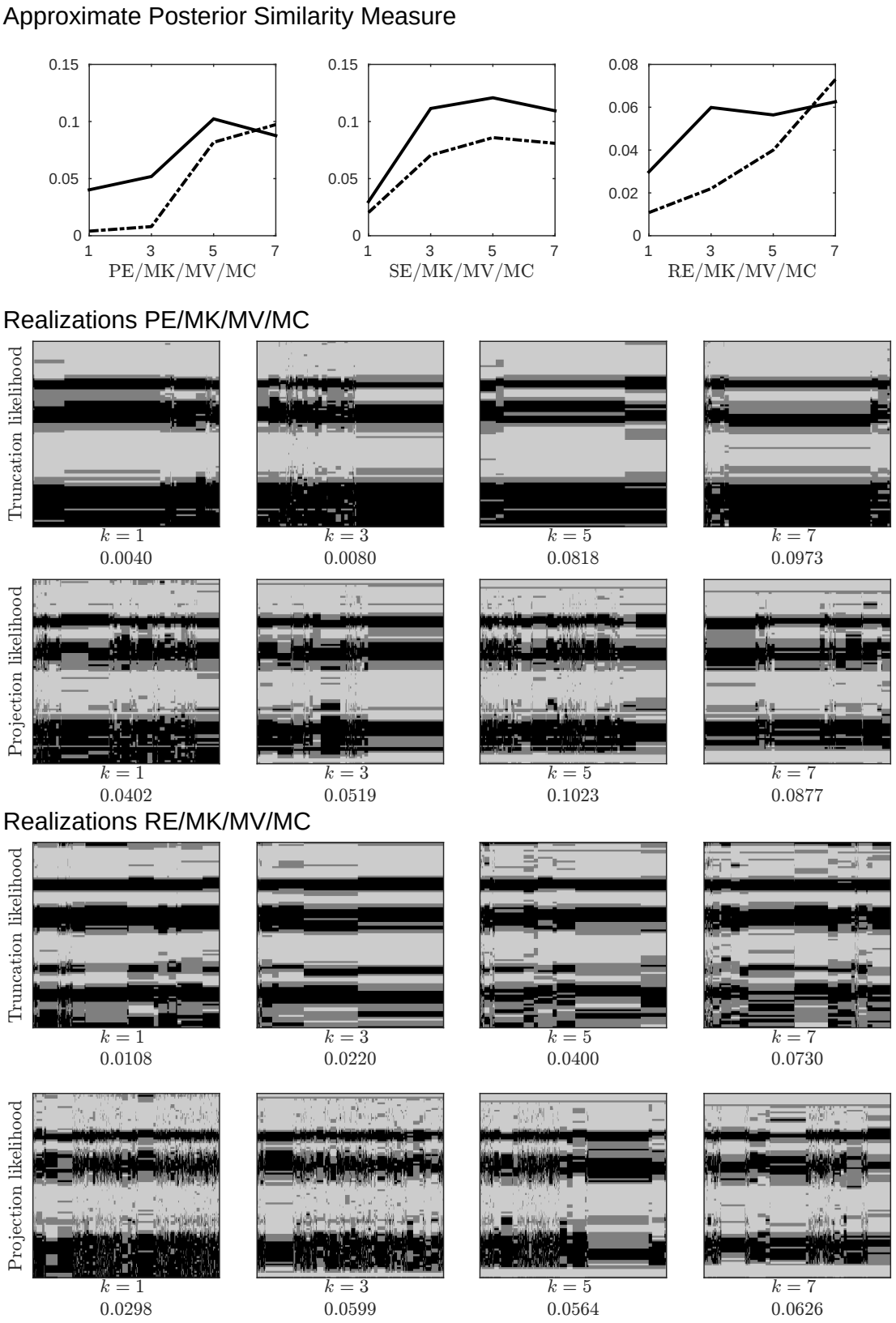


Figure 5.15: PE-SE-RE/MK/MV/MC: Top row: Acceptance rates as function of k. Projection approximation is shown with a solid line and truncation approximation with a dashed line. Bottom: 5,000 realizations from the various models with varying k. Acceptance rates, α, are included.

5.1.3 Apparent Convolution Width

We study a normalized, second order exponential apparent convolution kernel, but assume it to also be dependent on the range,

$$\mathbf{w}^{\mathrm{A}}(h;\epsilon) = \text{const} \times \exp\left\{-\frac{1}{2} \times \left(\frac{h}{\epsilon}\right)^2\right\}. \tag{5.8}$$

The apparent convolution kernel range is assumed to be $\epsilon \in \{2, 4, 6\}$, i.e. they have either a short, medium or long apparent convolution kernel. These cases are respectively SE/SK/MV/MC, SE/MK/MV/MC and SE/LK/MV/MC, and we refer to them as cases one, two and three. The remaining model parameters are assumed to be fixed, and as defined in Section 5.1.1. The resulting acquisition convolution kernels are given in Fig. 5.16.

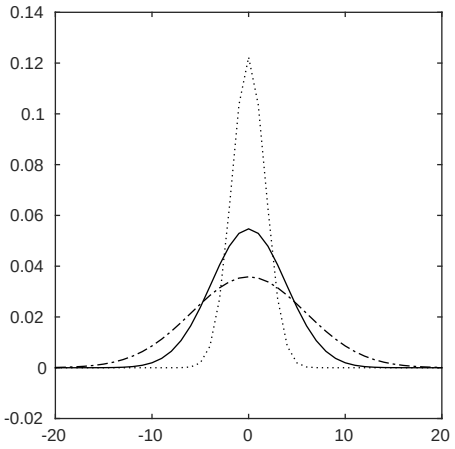


Figure 5.16: SE/SK-MK-LK/MV/MC: Short, medium and long acquisition convolution kernel, given with respectively a dotted, solid and dashed line.

Fig. 5.17 and Fig. 5.18 are in the same format as Fig. 5.14 and Fig. 5.15, respectively. The model parameters are specified at the top row. In the top row of displays, the spatial correlation function and the acquisition convolution kernel, are plotted in pairs. From left to right they are given with an increasing acquisition convolution kernel width. In the second to top row we see that the observations appear with decreasing 'shoulder effects', which is intuitively correct since the apparent convolution kernel increases. Both the truncation and projection based approximation have MAP predictors that appear to capture more of the small-scale variability in case one. As before, we find the projection approximation to be mostly superior to the truncation one. Indeed, a higher order likelihood approximation yields a higher acceptance rate in general.

The acceptance rates are presented at the top row of displays in Fig. 5.18. The acceptance rates increase for an increasing kernel width, i.e. it increases as the influence of the acquisition convolution kernel increases relative to the spatial correlation function. In the low displays of Fig. 5.17, 5,000 realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, are included for cases one and three.

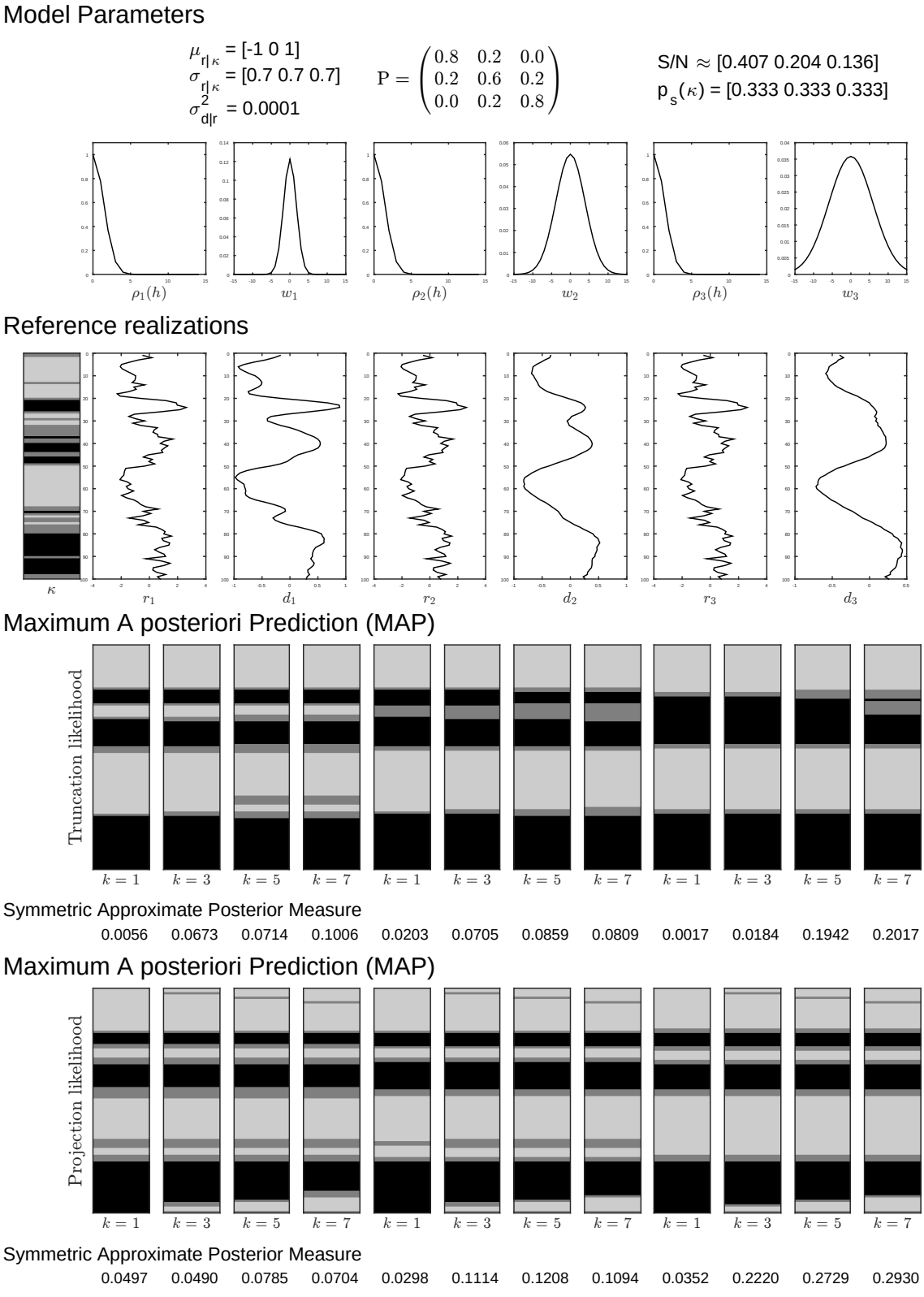


Figure 5.17: SE/SK-MK-LK/MV/MC: Model parameters, reference cases and MAP predictions/α-values for truncation and projection approximation for varying order k. The acceptance rates, α, are estimated from 90,000 iterations from the McMC MH-algorithm.

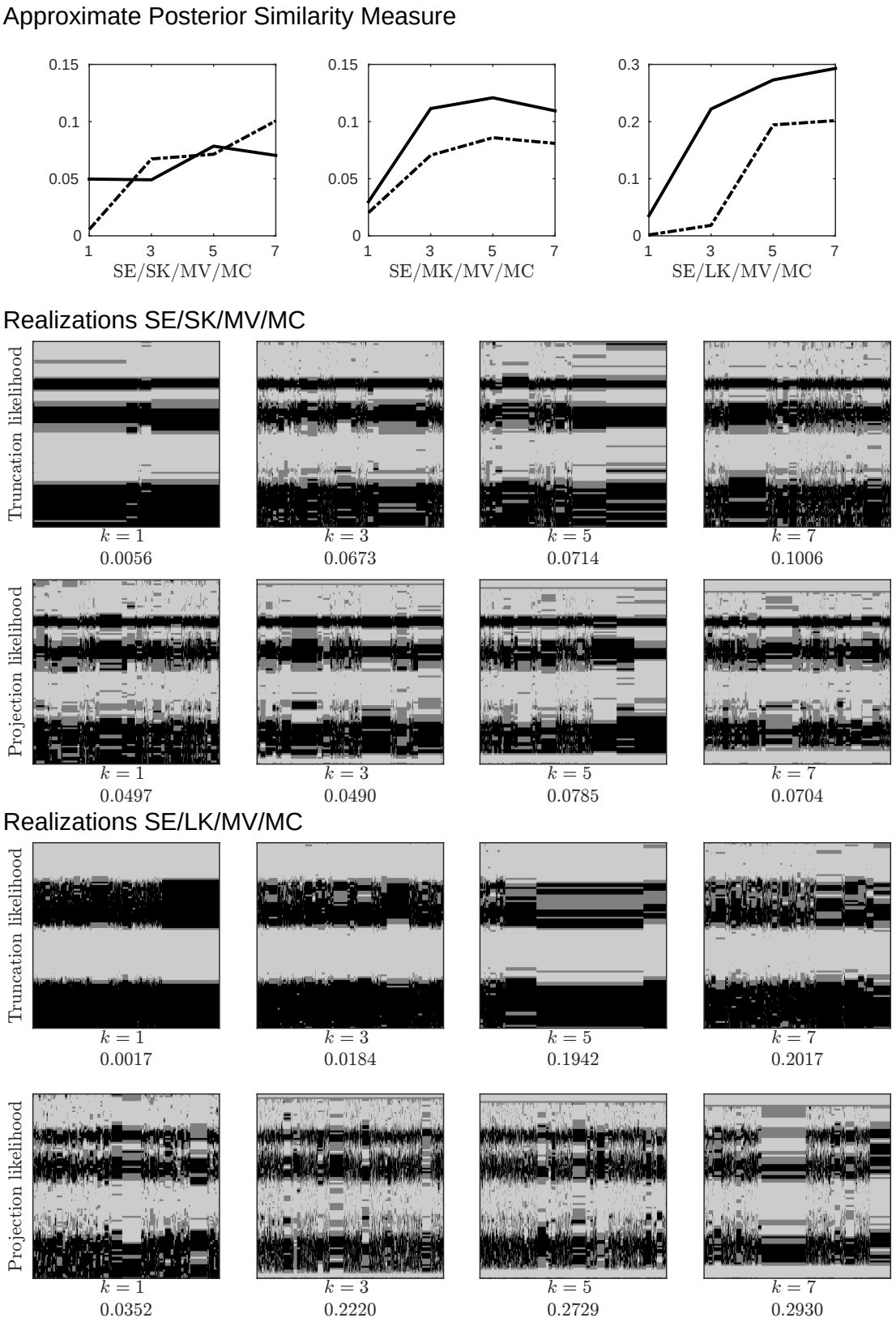
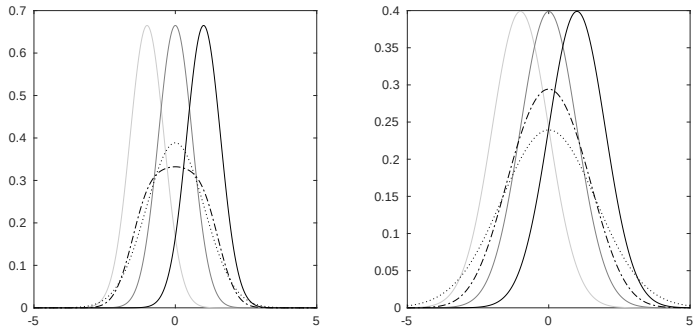


Figure 5.18: SE/SK-MK-LK/MV/MC: Top row: Acceptance rates as function of k. Projection approximation is shown with a solid line and truncation approximation with a dashed line. Bottom: 5,000 realizations from the various models with varying k. Acceptance rates, α, are included.

5.1.4 Variances in Response Model

We study the impact of different variances in the class response models. A low variance and a high variance case are introduced, in addition to the reference case. These cases are SE/MK/LV/MC, SE/MK/MV/MC and SE/MK/HV/MC, which we refer to as cases one, two and three. They are assumed to have $\sigma_{r|\boldsymbol{\kappa}'} = (0.6, 0.6, 0.6)^\top$, $\sigma_{r|\boldsymbol{\kappa}'} = (0.7, 0.7, 0.7)^\top$ and $\sigma_{r|\boldsymbol{\kappa}'} = (1, 1, 1)^\top$, respectively. The remaining model parameters are given in Section 5.1.1. The marginal class response densities are shown in Fig. 5.19, and should be compared with Fig. 5.2a. The spatial correlation function and apparent convolution kernel are fixed, and hence also the acquisition convolution kernel. The approximate posterior models have identical posterior means, but different covariance matrices.



(a) Low variance (b) High variance

Figure 5.19: SE/MK/LV-HV/MC: Class response densities displayed with solid lines. The Gaussian mixture is given with dashed line, and the Gaussian approximation is given with a dotted line.

The results are presented in Fig. 5.20 and Fig. 5.21. The model parameters are given at the top in Fig. 5.20, and we have assumed three different variances. At the top row of display from left to right, the apparent convolution kernel, and in pair the spatial correlation function and acquisition convolution kernel are given. Indeed, the response profiles and observations are almost identical in the three different cases since they have identical spatial correlation and acquisition convolution kernel. The MAP predictors for the truncation approximation are almost identical in the three cases. For the truncation based approximation, an increase in the class response variances entails a loss of small-scale variability in the MAP predictors. The MAP predictors based on the projection approximation are fairly similar in the various cases, however for case three, a small part of the dark-grey class is not identified compared to the other two cases.

The acceptance rates increase as functions of k in Fig. 5.21, and a high class response variance is favourable, as expected. The acceptance rates as functions of k have similar behaviour for the truncation and projection approximation. Based on the acceptance rates, the approximations are observed to favour high response variances.

For cases one and three, 5,000 realizations from the correct posterior model are displayed in Fig. 5.21. Higher heterogeneity than in the MAP predictions are observed, of course.

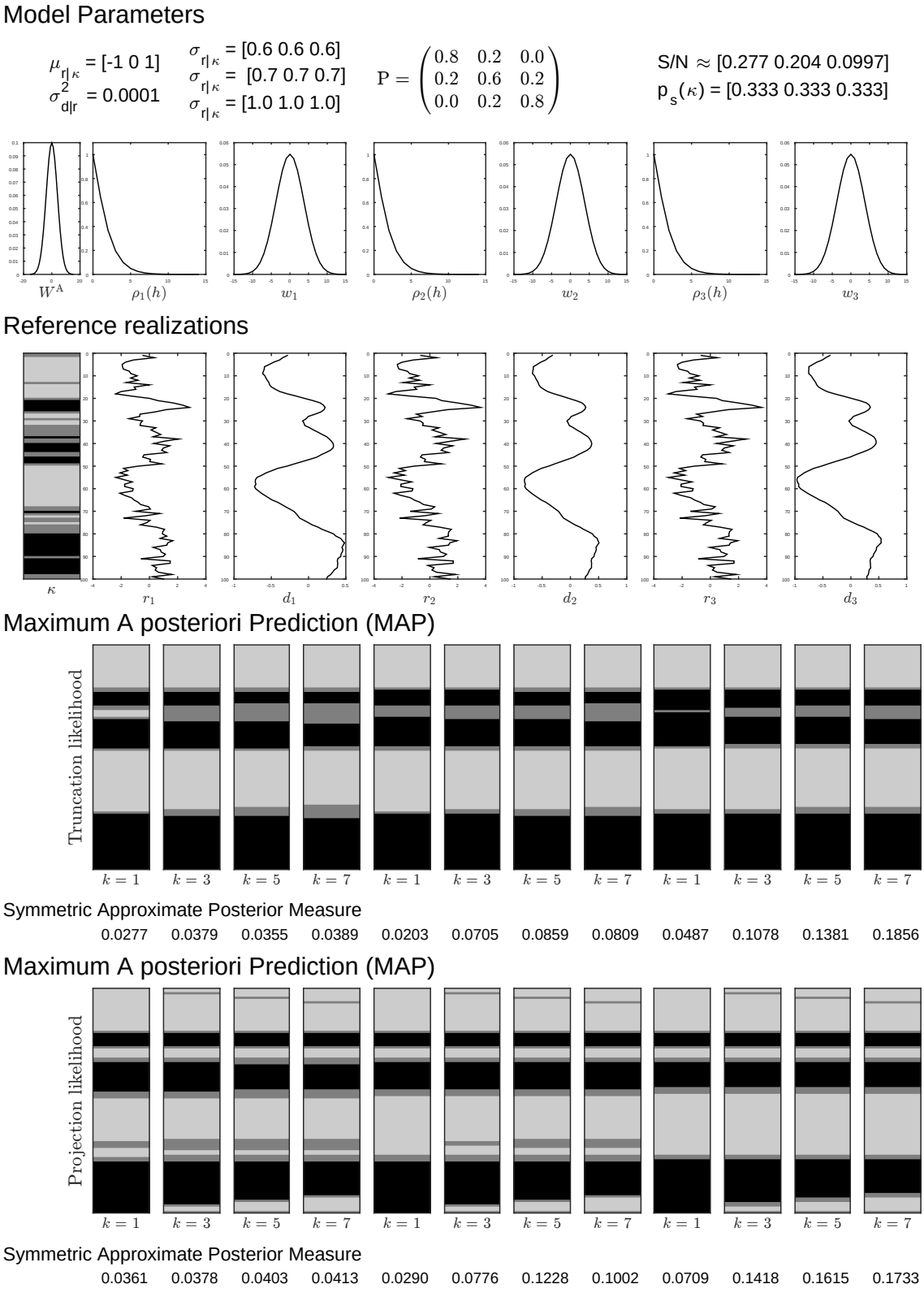


Figure 5.20: SE/MK/LV-MV-HV/MC: Model parameters, reference cases and MAP predictions/α-values for truncation and projection approximation for varying order k. The acceptance rates, α, are estimated from 90,000 iterations from the McMC MH-algorithm.

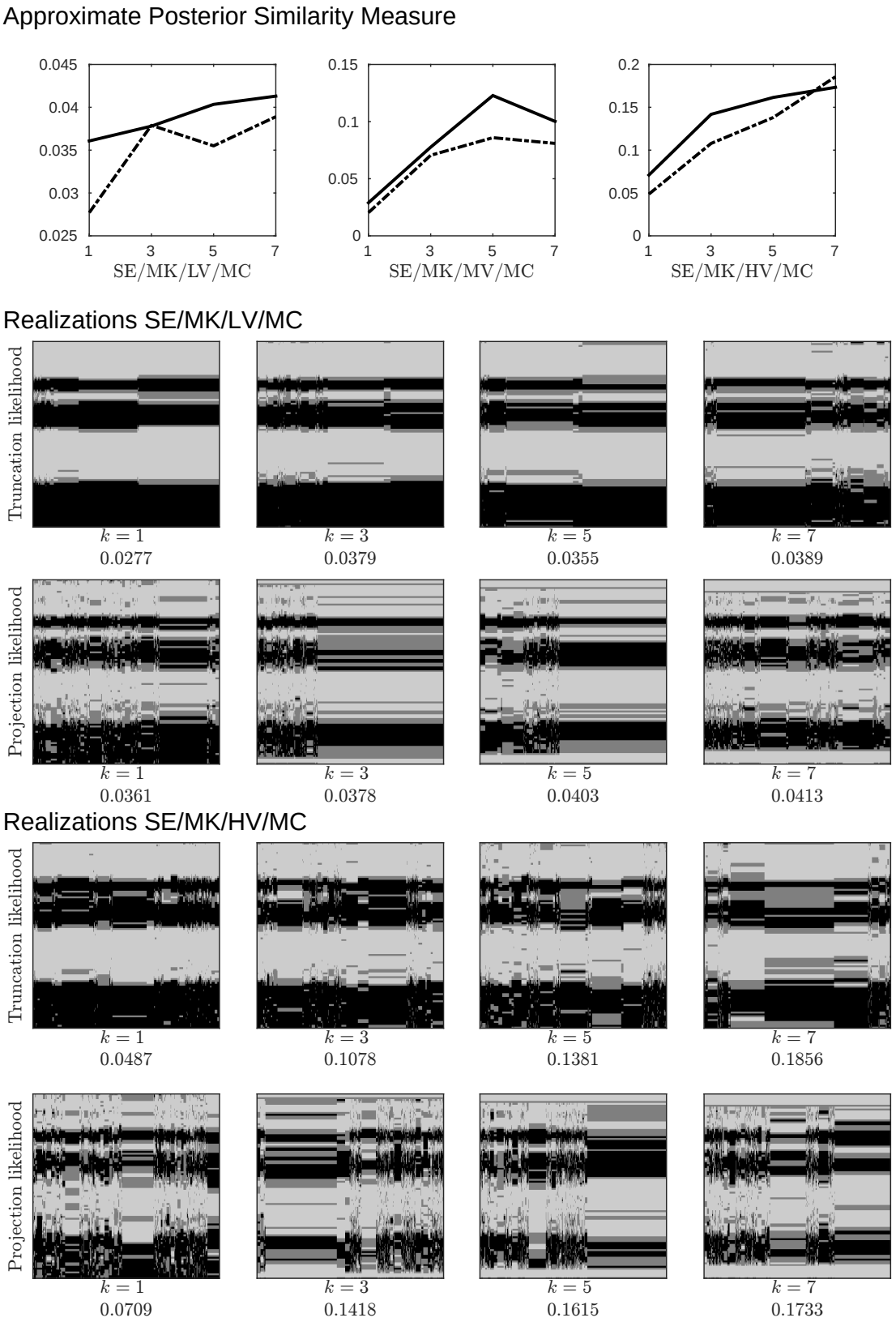


Figure 5.21: SE/MK/LV-MV-HV/MC: Top row: Acceptance rates as function of k. Projection approximation is shown with a solid line and truncation approximation with a dashed line. Bottom: 5,000 realizations from the various models with varying k. Acceptance rates, α, are included.

5.1.5 Spatial Correlation Response Model

A case study where the apparent convolution kernel is fixed, but the spatial correlation function is varied, is now considered. We assume the remaining model parameters to be given as in Section 5.1.1. The spatial correlation function is defined as

$$\rho_{\mathbf{r}}(h;\xi) = \exp\left\{-\frac{1}{2} \times \left(\frac{h}{\xi}\right)^2\right\}, \tag{5.9}$$

where the range parameter, ξ, describes the effective range of the spatial correlation. We consider test cases SE/MK/MV/SC, SE/MK/MV/MC and SE/MK/MV/LC, with respectively $\xi \in \{1, 2, 3\}$. They are referred to as cases one, two and three, respectively. The spatial correlation functions are presented in Fig. 5.22.

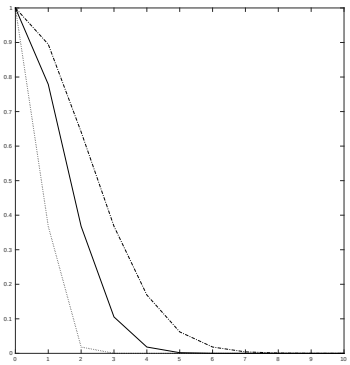


Figure 5.22: SE/MK/MV/SC-MC-LC: Spatial correlation functions, $\rho_{\mathbf{r}}(h)$, with ranges 1, 2 and 3.

The results are presented in Fig. 5.23 and Fig. 5.24. Model parameters are given at the top in Fig. 5.23. The apparent convolution kernel is given in the leftmost plot in the top row of displays. Then, in pair with increasing spatial correlation, the spatial correlation function is given together with the acquisition convolution kernel. The various response profiles, $\mathbf{r}_1$ to $\mathbf{r}_3$, appear with increasing smoothness. We observe that the different observations do share the main characteristics, with slightly more smoothness in $\mathbf{d}_1$.

The MAP predictors for the truncation approximation are almost identical for the various cases, and as functions of k. As before, the projection approximation captures more of the small-scale variability in the reference profile

As we have discussed in Section 5.1.3, the acceptance rates increase when the influence of the acquisition convolution kernel, $\mathbf{w}$, increases relative to the spatial correlation function, $\rho_{\mathbf{r}}(h)$. Indeed, case one has significantly higher acceptance rates than case three in Fig. 5.24. Acceptance rates of approximately 0.5 are obtained, which entails good mixing in the McMC algorithm. This can be observed in the display of the 5,000 realizations for case one.

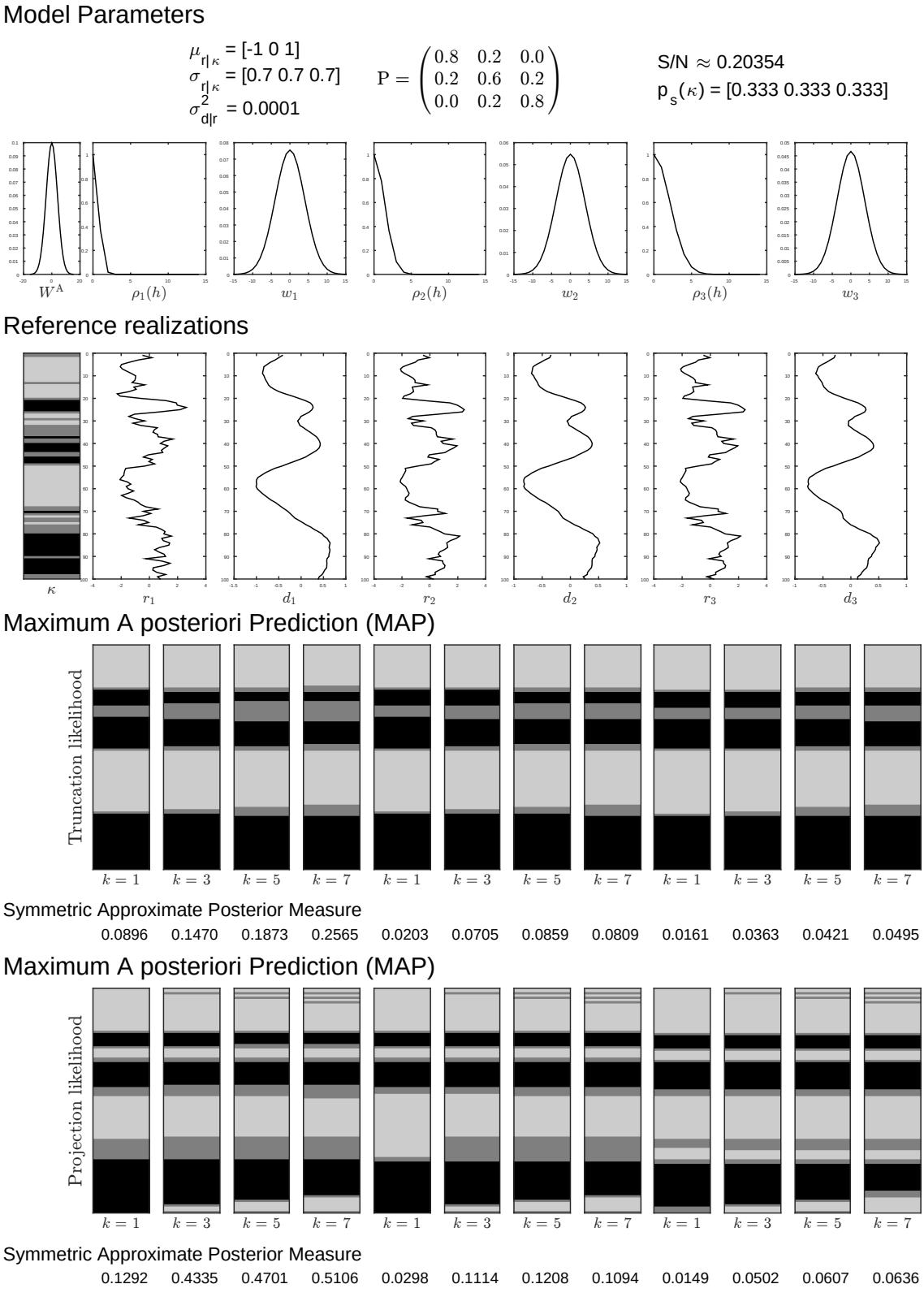


Figure 5.23: SE/MK/MV/SC-MC-LC: Model parameters, reference cases and MAP predictions/α-values for truncation and projection approximation for varying order k. The acceptance rates, α, are estimated from 90,000 iterations from the McMC MH-algorithm.

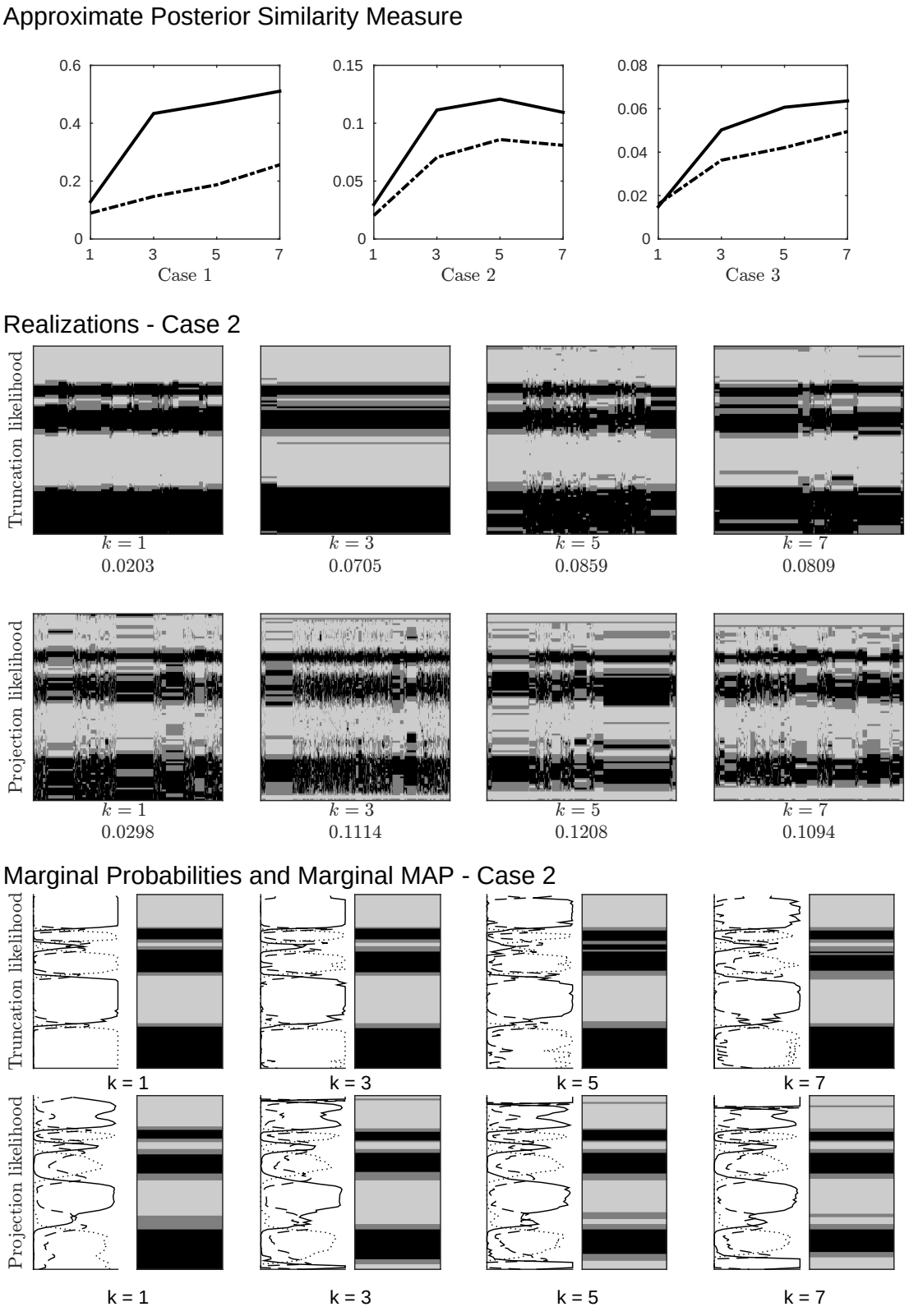


Figure 5.24: SE/MK/MV/SC-MC-LC: Top row: Acceptance rates as function of k. Projection approximation is shown with a solid line and truncation approximation with a dashed line. Bottom: 5,000 realizations from the various models with varying k. Acceptance rates, α, are included.

5.2 Closing Remarks

We have studied in total nine different models. Two different likelihood approximations, of various order, are evaluated.

The MAP predictors based on projection for the approximate posterior models capture more of the model heterogeneity than the MAP predictor for the truncation based approximation. The MAP and MMAP predictors are almost identical for various k, hence a lower order approximation is sufficient if prediction of the approximate posterior model is of interest.

Since the computational cost increases exponentially with increasing k, we seek a low order approximation. The approximate posterior should be sufficiently close to the correct posterior model, so that realizations from the correct posterior model can be generated at a reasonable computational cost. We introduce the acceptance rate as a measure of similarity, i.e. it quantifies the quality of the approximations. The projection based approximate posterior model has generally higher acceptance rates than the truncation one for lower order approximations. In most cases the acceptance rate is shown to be an increasing function of k, which stabilizes for $k \geq 3$. We observe that the truncation based approximation has increasing acceptance rates for increasing k, and we expect it to be better than the projection approximation for a sufficiently large k. Indeed, the truncation approximation is exact if k is chosen sufficiently large.

If the acquisition convolution kernel is wide, the projection based approximation is found to perform significantly better than the truncation based approximation. If the apparent convolution kernel is short, the relative difference between the truncation and projection approximations is small. Response models which are strongly correlated, are more likely to have a low acceptance rate compared to the response processes with less spatial correlation. A high class response variance is also preferable with respect to the acceptance rates.

The realizations from the correct posterior model represent more of the heterogeneity than the predictors, which they should.

We conclude that a third or fifth order projection approximation is favourable since it yields reasonable acceptance rates at a low computational cost.

Chapter 6

Assessment of the Transition Matrix

We study assessment of the transition matrix, $\mathbf{P}_{\boldsymbol{\kappa}}$. A simple model with two different classes, and a more complicated model with ordered classes, are studied. The first case study is inspired by the simple Bernoulli-Gaussian model, where the second class denotes a jump between classes. In the latter test study we assume a structured $\mathbf{P}_{\boldsymbol{\kappa}}$.

Only the third order projection approximation of the likelihood model is considered. This is found to be a reasonable trade-off between computational cost and accuracy in Chapter 5. We compare the approximate EM-algorithm and McMC approach, presented in Chapter 4 to assess the transition matrix, $\mathbf{P}_{\boldsymbol{\kappa}}$. In the latter case we also assess the uncertainty in the estimates. We use the estimated transition matrices as plug-in estimates to assess the MAP and MMAP predictors.

We assume the model parameters, except the transition probabilities, to be fixed and known.

6.1 High Reflector Points

The first test study is an extension of a model studied in Lindberg and Omre (2014b). A convolutional two-level hidden Markov model of length $N = 200$, with $\kappa_n \in \Omega_{\boldsymbol{\kappa}} = \{\text{grey}, \text{black}\}$, is studied. The Bernoulli-Gaussian model only predicts the location of the transitions, and not the distinct classes.

6.1.1 Model Specification

We assume the transition matrix to be defined as

$$\mathbf{P}_{\boldsymbol{\kappa}} = \begin{pmatrix} 0.90 & 0.10 \\ 0.95 & 0.05 \end{pmatrix}, \tag{6.1}$$

having stationary distribution $(0.9048, 0.0952)$. The reference profile is displayed in Fig. 6.1. The corresponding empirical transition matrix, $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{emp}}$, which we obtain by counting the number of transitions in $\boldsymbol{\kappa}$, is

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{emp}} = \begin{pmatrix} 0.8764 & 0.1236 \\ 1.0000 & 0 \end{pmatrix}. \tag{6.2}$$

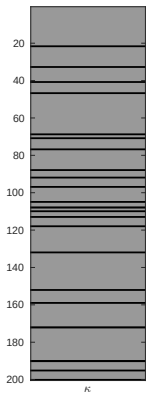


Figure 6.1: Reference profile, $\boldsymbol{\kappa}$.

A longer reference profile would randomize over the model, and with an infinite long profile the correct and empirical transition matrices coincide.

We assume the class response models to be defined by $\boldsymbol{\mu}_{r|\kappa} = (0, 0)^\top$ and $\boldsymbol{\sigma}_{r|\kappa} = (0.1, 3)^\top$. The two classes are chosen to have identical means, but different variances. High reflector points have high variance, and inhomogeneities inside each class are assumed to have a smaller variance. In Fig. 6.2 the conditional response densities, $p(r|\kappa = \text{grey})$ and $p(r|\kappa = \text{black})$, are displayed with solid lines. The corresponding Gaussian mixture is given with a dashed line, and almost overlaps with $p(r|\kappa = \text{grey})$. The Gaussian approximation is displayed with a dotted line.

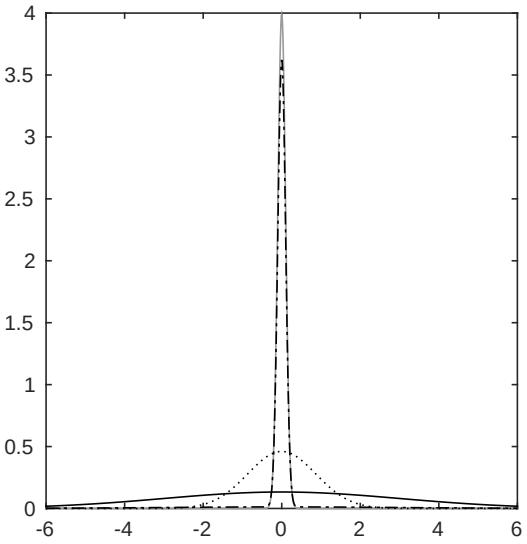


Figure 6.2: Conditional densities displayed with solid lines, the Gaussian mixture with a dashed line, and the Gaussian approximation with a dotted line.

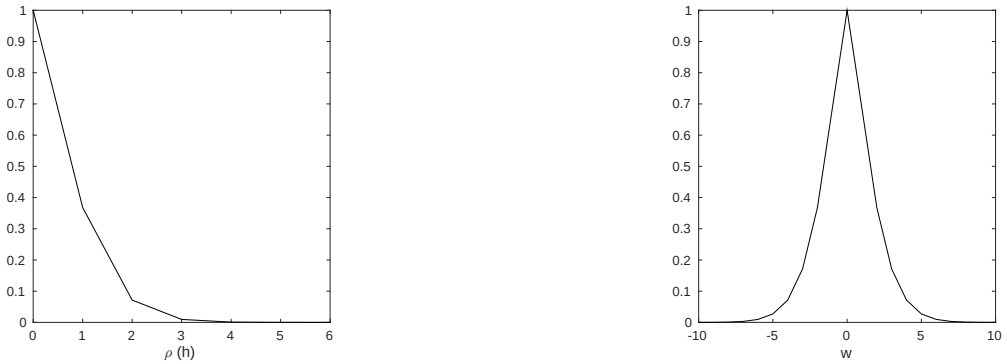
The spatial correlation function

$$\rho(h) = \exp\left\{-h^{1.4}\right\}, \tag{6.3}$$

is displayed in Fig. 6.3a. We truncate $\mathbf{\Sigma}_{\mathbf{r}}^{\rho}$ by assuming $a_\rho = 5$. The acquisition convolution kernel is assumed to be a powered exponential,

$$\mathbf{w}(h) = \exp\left\{-\left(\frac{h}{2}\right)^{1.4}\right\}, \tag{6.4}$$

and it is given in Fig. 6.3b. We truncate $\mathbf{W}$ to be of band width 15, i.e. $a_w = 7$.



(a) Correlation, $\rho_{\mathbf{r}}(h)$ (b) Convolution, $\mathbf{w}$

Figure 6.3: Spatial correlation function, $\rho_{\mathbf{r}}(h)$, and acquisition convolution kernel, $\mathbf{w}$.

An observational error $\sigma^2_{\mathbf{d}|\mathbf{r}} = 0.01$ is assumed, and the associated signal-to-noise ratio is 1.4237. The signal-to-noise ratio is defined in Section 5.1.

In Fig. 6.4 the response profile, $\mathbf{r}$, is given together with the observations, $\mathbf{d}$. Most of the high reflection points appear as spikes in the response profile. Because of identical expectation, not every transition in reference profile is evident in the response profile. Indeed, the main characteristics of reference profile appear as distinct shoulders in the observations.

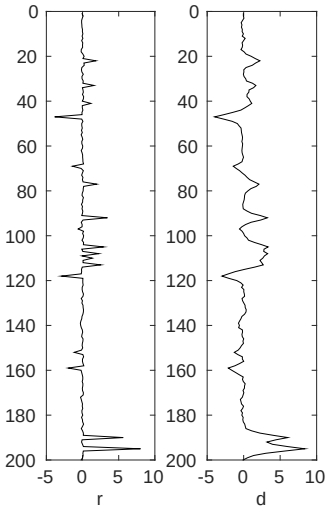


Figure 6.4: Response profile, $\mathbf{r}$, and observations, $\mathbf{d}$.

6.1.2 Results

The approximate EM-algorithm is run in total 25 times with various initial transition matrices, $\mathbf{P}_{\boldsymbol{\kappa}}$. The initial transition matrices are chosen both informative, i.e. close to the correct transition matrix, and uninformative, having uniform probabilities. The 25 versions converge to the same estimate. In Fig. 6.5 we have included trace plots of the log-likelihoods found by the approximate EM-algorithm based on four different initial transition matrices. The log-likelihoods converge within seven iterations. The estimated transition matrix is given as

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{aEM}} = \begin{pmatrix} 0.8803 & 0.1197 \\ 0.9235 & 0.0765 \end{pmatrix}, \tag{6.5}$$

with stationary distribution $(0.8853, 0.1147)$. Compared to the correct transition matrix, given in Eq. (6.1), the approximate EM estimate slightly overestimates the proportion of the black class. As discussed, this is not unreasonable since the Gaussian approximation has a heavier tail than the Gaussian mixture in our response model.

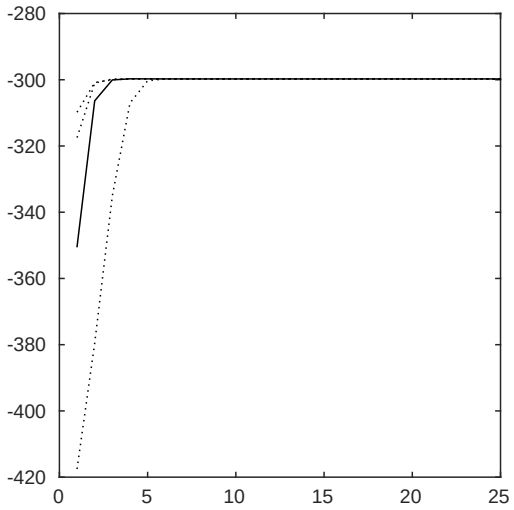


Figure 6.5: Trace plot of the log-likelihoods based on four different initial transition matrices.

The approximate EM-algorithm provides an extremely fast and accurate estimate of the transition matrix. However, the fast computation comes at the cost of not having the associated uncertainty in the parameter estimates. It is possible to construct approximate confidence bands from the Hessian matrix, see Lindberg and Omre (2014b), but these are not very reliable estimates close to zero or one.

In Fig. 6.6 the marginal probabilities, and MAP and MMAP predictors for $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\kappa}^{\text{aEM}}\right)$ are given together with the reference profile. The MAP and MMAP predictors represent the main characteristics of the reference profile. The MMAP predictor is almost identical to the reference profile, whereas the MAP predictor differs slightly more.

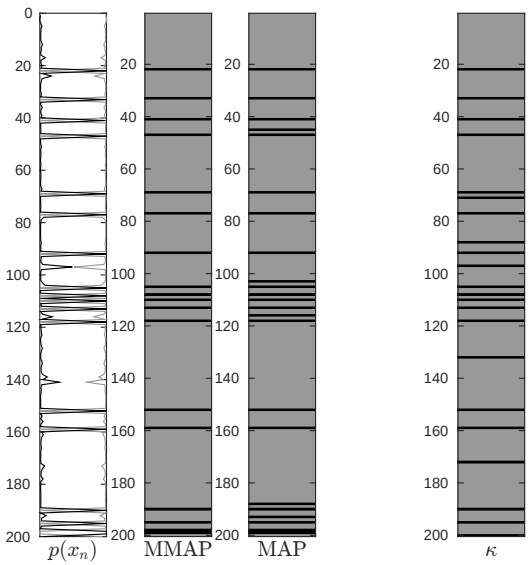


Figure 6.6: Marginal probabilities, and MAP and MMAP predictors for $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\kappa}^{\mathrm{aEM}}\right)$, together with reference profile.

We assess the transition matrix through McMC sampling, see Section. 4.3.1, by assuming a symmetric Dirichlet prior, with $\eta_{ij} = 1$ for all i, j. A sequence of 5,000 realizations is generated from the posterior $p(\mathbf{P}_{\boldsymbol{\kappa}}|\mathbf{d})$. The McMC algorithm is initialized with a transition matrix having elements equal to 1/2. Trace plots of the transition probabilities in the McMC-algorithm are displayed in Fig. 6.7. We observe that the sequence of realizations converges almost instantaneously. The burn-in period is fixed to 1,000 iterations.

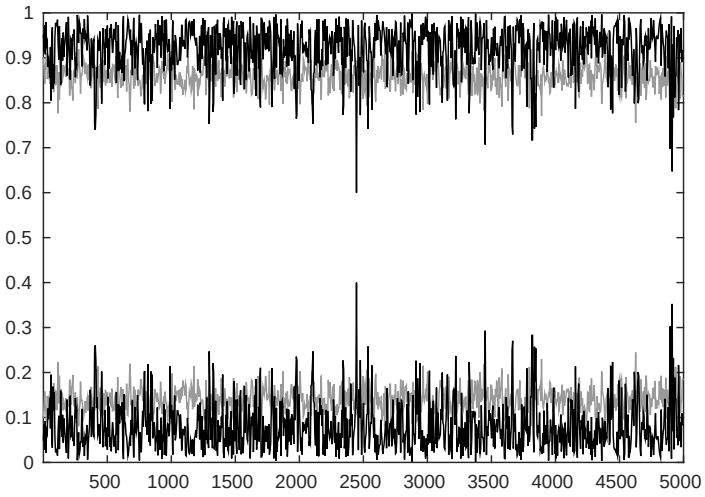


Figure 6.7: Trace plots of the estimated transitions probabilities.

The McMC estimate is given as

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}} = \begin{pmatrix} 0.8608 & 0.1392 \\ 0.9203 & 0.0797 \end{pmatrix}, \tag{6.6}$$

with stationary distribution $(0.8686, 0.1314)$. Each element, $\hat{p}_{ij}$, in Eq. (6.6) is chosen to be the estimated mode of the realizations, since the estimated mean in general is a poor estimate if it is close to zero or one. The estimated transition matrices are found to be almost identical. Both estimates overestimates the proportion of the black class.

In Fig. 6.8 the estimated transition probabilities are given with their associated 95% confidence bands. The confidence bands are defined from the 95% range of the posterior for each parameter. We see that the correct transition probabilities are well inside their respective 95% confidence band.

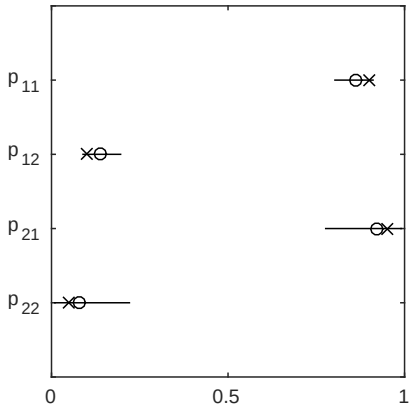


Figure 6.8: Estimators $\left[\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}\right]_{ij}$, with estimated values displayed with 'o', correct values, '×', and 95% confidence range.

In Fig. 6.9 the marginal probabilities, and MAP and MMAP predictors for $p\left(\boldsymbol{\kappa}|\mathbf{d}; \hat{\mathbf{P}}_{\kappa}^{\mathrm{McMC}}\right)$ are given together with the reference profile. The marginal probabilities capture most of the variability in the correct posterior model. Indeed, the MAP and MMAP predictors based on the McMC estimate are very similar to the MAP and MMAP predictors based on the approximate EM.

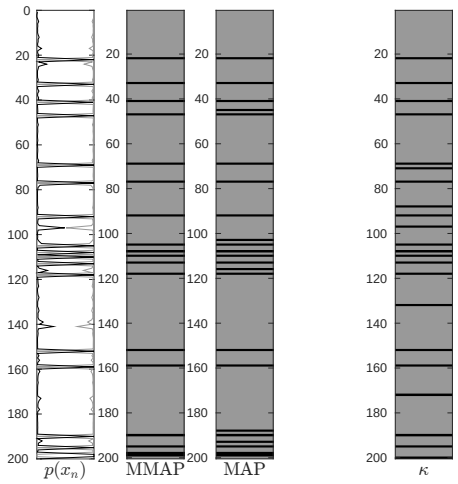


Figure 6.9: Marginal probabilities, and MAP and MMAP predictors for $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\kappa}^{\mathrm{McMC}}\right)$, together with reference profile.

We compare the misclassification coverage statistics as defined in Eq. (5.3). The estimated transition matrices, $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}$ and $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}$, are used as plug-in estimates. In Fig. 6.10 we observe that the approximate posteriors, $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}\right)$ and $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}\right)$, are good approximations for the grey class since the coverage statistics are small and close to unity. We find this to be reasonable since we have more observations from the grey class. Indeed, the black class is slightly underestimated. Indeed, the coverage statistics based on the approximate EM and McMC estimates are almost identical.

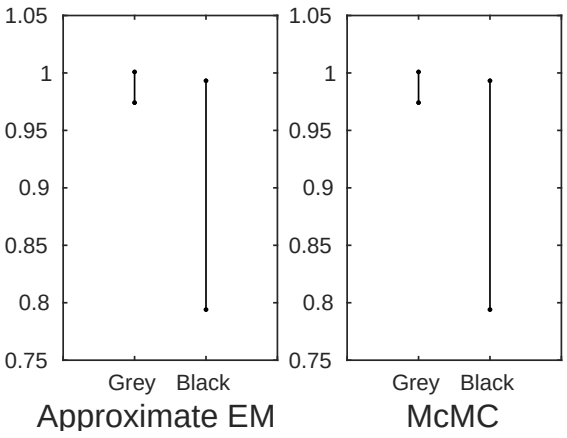


Figure 6.10: Misclassification coverage statistics based on the approximate EM and McMC estimation methods.

The computational requirement of $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}$ is severe compared to $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}$. Indeed, the latter requires only to run the Forward-Backward algorithm to ensure convergence to the correct maximum of the log-likelihood. To generate realizations from the correct posterior model,

we need to run the Forward-Backward algorithm each time, and then generate realizations from $p(\mathbf{P}_{\boldsymbol{\kappa}}|\boldsymbol{\kappa})$, which may be very expensive if the acceptance rate is low.

6.2 Ordered Profile

We study a ordered reference profile, $\boldsymbol{\kappa}$, of length $N = 100$. The reference profile is assumed to be a first order Markov chain, with $\kappa_n \in \Omega_{\boldsymbol{\kappa}} = \{1, \ldots, 4\}$.

6.2.1 Model Specification

We assume the transition matrix to be given as

$$\mathbf{P}_{\boldsymbol{\kappa}} = \begin{pmatrix} 0.8 & 0.2 & 0.0 & 0.0 \\ 0.1 & 0.7 & 0.2 & 0.0 \\ 0.1 & 0.0 & 0.7 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 \end{pmatrix}, \tag{6.7}$$

with stationary distribution $(0.4154, 0.2769, 0.1846, 0.1231)$. The reference profile is displayed in Fig. 6.11, and it appears with a stair-like sequence, which randomly falls back to class one. By counting the number of transitions in $\boldsymbol{\kappa}$, the empirical transition matrix is found to be

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{emp}} = \begin{pmatrix} 0.8298 & 0.1702 & 0 & 0 \\ 0.0800 & 0.6800 & 0.2400 & 0 \\ 0.1667 & 0 & 0.6666 & 0.1667 \\ 0.3333 & 0 & 0 & 0.6667 \end{pmatrix}, \tag{6.8}$$

with stationary distribution $(0.4769, 0.2557, 0.1778, 0.0896)$.

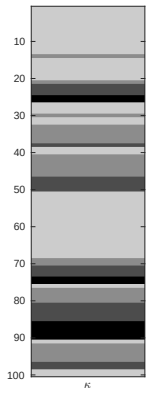


Figure 6.11: Reference profile, $\boldsymbol{\kappa}$.

The model parameters in the class responses are assumed to be given as $\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'} = (0, 1, 2, 3)^\top$ and $\boldsymbol{\sigma}_{r|\boldsymbol{\kappa}'} = (0.5, 0.5, 0.5, 0.5)^\top$. In Fig. 6.12 the class response densities $p(r_n|\kappa_n = i)$ for $i = 1, \ldots 4$ are shown with solid lines, with the appropriate grey-scale. The Gaussian mixture is shown with a dashed line, and it is found to be skewed. The Gaussian approximation, which is symmetric, is displayed with a dotted line.

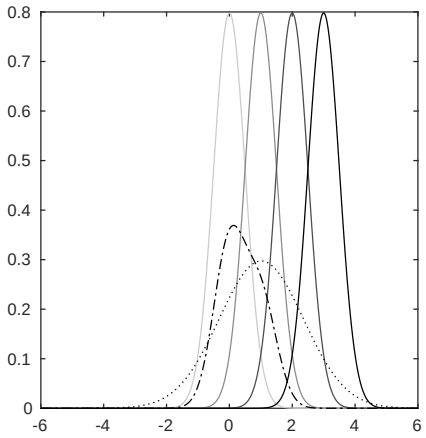


Figure 6.12: Class response densities displayed with solid lines, together with the Gaussian mixture displayed with a dashed line and the Gaussian approximation displayed with a dotted line.

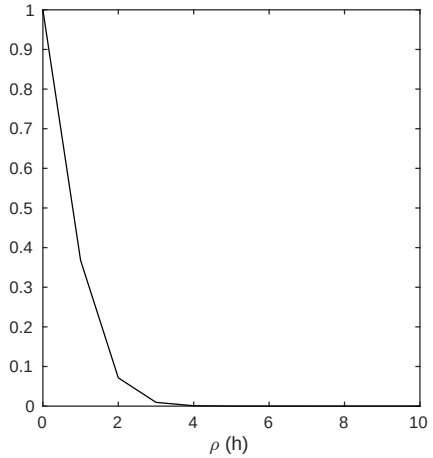
For the response likelihood, a dependent mode process is assumed, with spatial correlation function

$$\rho_{\mathbf{r}}(h) = \exp\left\{-h^{1.4}\right\}. \tag{6.9}$$

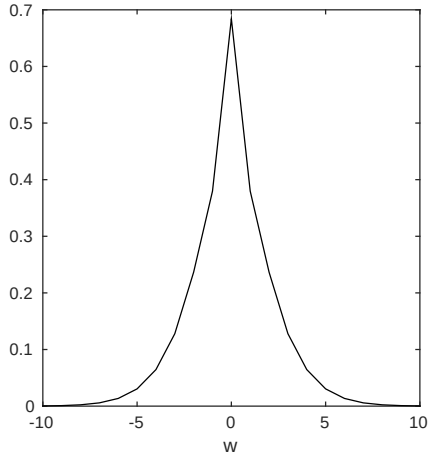
The spatial correlation function is given in Fig. 6.13a, and it is truncated with $a_\rho = 4$. We assume the acquisition convolution kernel to be given as

$$\mathbf{w}^{\mathrm{A}}(h) = \text{const} \times \exp\left\{-\left(\frac{h}{2.5}\right)^{1.4}\right\}. \tag{6.10}$$

In Fig. 6.13b the acquisition convolution kernel is given, and we truncate $\mathbf{W}$ to a band-diagonal matrix of band-width 21, i.e. $a_w = 10$. The acquisition convolution kernel has a fairly high spike.



(a) Correlation, $\rho_{\mathbf{r}}(h)$



(b) Convolution, $\mathbf{w}$

Figure 6.13: Spatial correlation function, $\rho_{\mathbf{r}}(h)$, and acquisition convolution kernel, $\mathbf{w}$.

We assume the observational error to be $\sigma^2_{\mathbf{d}|\mathbf{r}} = 0.01$. The ordering model is specified with a reasonably high signal-to-noise ratio, S/N ≈ 3.25. The response profile and observations are given in Fig. 6.14. We assume prior knowledge about the zero-probabilities in $\mathbf{P}_{\boldsymbol{\kappa}}$. That is, we require $p_{13}, p_{14}, p_{32}, p_{42}$ and p_{43} to be zero in the estimate. Since the likelihood approximations and Forward-Backward algorithm preserve zero-probabilities, we are able to capture this in the estimate. Hence, the number of transition probabilities to be estimated is ten. Numerical experiments without this assumption tend to estimate band diagonal transition matrices, i.e. we estimate each class to move to one of its closest neighbours.

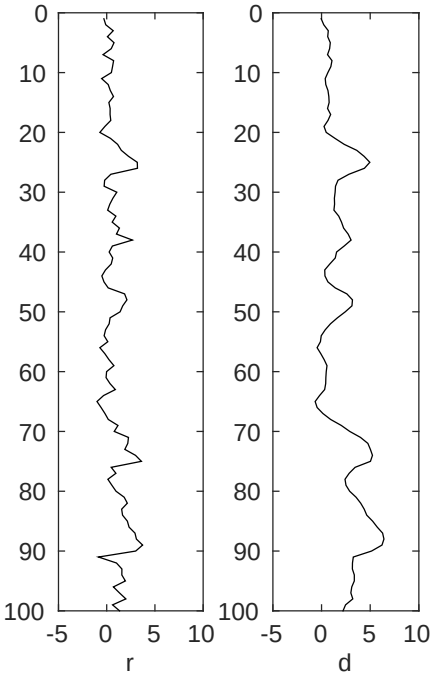


Figure 6.14: Response profile, $\mathbf{r}$, and observations, $\mathbf{d}$.

6.2.2 Results

The approximate EM-algorithm is run 25 times with various initial transition matrices. The estimate of $\mathbf{P}_{\boldsymbol{\kappa}}$ is

$$\hat{\mathbf{P}}^{\text{aEM}}_{\boldsymbol{\kappa}} = \begin{pmatrix} 0.8919 & 0.1080 & 0 & 0 \\ 0.0003 & 0.7032 & 0.2964 & 0 \\ 0.1186 & 0 & 0.6652 & 0.2162 \\ 0.3697 & 0 & 0 & 0.6302 \end{pmatrix}. \tag{6.11}$$

Compared to the correct transition matrix, $\hat{\mathbf{P}}^{\text{aEM}}_{\boldsymbol{\kappa}}$ is found to be reasonable since it captures most of the structure in $\mathbf{P}_{\boldsymbol{\kappa}}$. The stationary distribution of $\hat{\mathbf{P}}^{\text{aEM}}_{\boldsymbol{\kappa}}$ is found to be $(0.5334, 0.1942, 0.1719, 0.1005)$, and it overestimates the proportion of class one, at the cost of class two. Indeed, also p_{23} is overestimated at the cost of p_{21}.

In Fig. 6.15 trace plots of the approximate log-likelihood values are given based on four different initial transition matrices. The log-likelihoods converge within ten iterations, however the log-likelihoods are not necessarily non-decreasing because of the likelihood approximations.

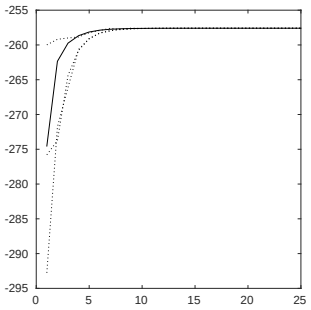


Figure 6.15: Trace plot of the log-likelihoods based on four different initial transition matrices.

For the McMC based inference of $\mathbf{P}_{\boldsymbol{\kappa}}$, we generate 5,000 realizations from $p\left(\mathbf{P}_{\boldsymbol{\kappa}}|\mathbf{d}\right)$, where we assume the hyperparameters to be $\eta_{ij} = 0$ for $(i,j) \in \{(1,3),(1,4),(2,4),(3,2),(4,2),(4,3)\}$, and $\eta_{ij} = 1$ else. We have chosen to discard the first 1,000 realizations as a burn-in period. The estimated transition matrix is given as

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{McMC}} = \begin{pmatrix} 0.8811 & 0.1189 & 0 & 0 \\ 0.0518 & 0.6028 & 0.3454 & 0 \\ 0.1236 & 0 & 0.6414 & 0.2350 \\ 0.4200 & 0 & 0 & 0.5800 \end{pmatrix}, \tag{6.12}$$

with stationary distribution $(0.5718, 0.1712, 0.1649, 0.0922)$. If we compare the elements with the corresponding elements in the correct transition matrix, we see that the main characteristics of the transition matrix are found by the McMC estimate. Compared to $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{aEM}}$, the estimates are fairly similar. Note that $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{aEM}}$ estimates p_{21} to be essentially zero, while $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{McMC}}$ estimates it to be slightly above 0.05.

In Fig. 6.16 the estimated non-zero transition probabilities are presented together with their corresponding 95% range. We see that the correct transition probabilities lie inside their respective 95% confidence band. Compared to the confidence bands in the previous example, the confidence bands given here are fairly wide. We find this to be reasonable since the number of classes is larger and the classes are overlapping.

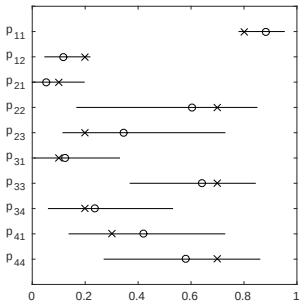


Figure 6.16: Estimators $\left[\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}\right]_{ij}$, with estimated values displayed with 'o', correct values, '×', and 95% range.

In Fig. 6.17 the MAP predictors are given based on the approximate EM and McMC estimates. Compared to the reference profile, the predictors are surprisingly similar. Most of the small-scale variability are in fact captured in both of the predictors, which we find to be encouraging. The resulting MAP predictors are observed to be almost identical.

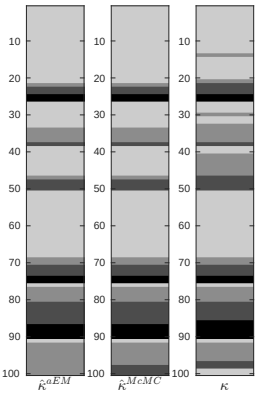


Figure 6.17: MAP predictors for $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}\right)$ and $p\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{McMC}}\right)$, together with the reference profile, $\boldsymbol{\kappa}$.

The misclassification coverage statistics based on the estimated transition matrices are presented in Fig. 6.18. Indeed, the approximate EM-algorithm and McMC method provide slightly different results. Classes one and four, corresponding to the light-grey and black class, are seen to have the best misclassification coverage statistics in both cases. This is as expected, since they are the extreme classes. The misclassification coverage statistic varies in the two middle classes. This is to be expected since they are partly overlapping, and intuitively should be hard to separate.

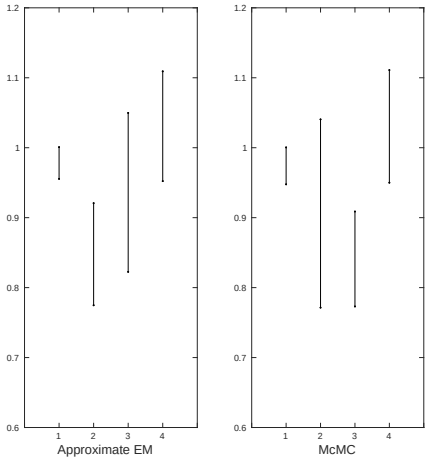


Figure 6.18: Misclassification coverage statistics based on the approximate EM and McMC estimation methods.

6.3 Closing Remarks

Estimation of the transition matrix based on the approximate EM-algorithm and McMC sampling are demonstrated to be feasible. Compared to the approximate EM-algorithm, the McMC based inference is computationally expensive. The McMC method quickly becomes computational infeasible if there is a large number of unknown transition probabilities.

The approximate EM-algorithm only provides point estimates, while the McMC based inference also provides parameter uncertainties. The 95% confidence bands covers the correct transition probabilities well. If we study a approximate posterior model with plug-in transition matrix estimates, the MAP predictors are reliable representations of the reference profile.

The structure of $\mathbf{P}_{\boldsymbol{\kappa}}$ is found to be important. That is, if we have prior knowledge of zero transitions in $\mathbf{P}_{\boldsymbol{\kappa}}$, this should be taken into account in both the approximate EM-algorithm and McMC based inference.

Chapter 7

Case Study: Seismic Inversion

We present a synthetic seismic test study. A reference profile of length $N = 150$, with $\kappa_n \in \Omega_{\boldsymbol{\kappa}} = \{1, \ldots, 4\}$, is studied. The reference profile is displayed in Fig. 7.1. The four classes represent the lithology/fluid classes shale, gas-saturated sandstone, oil-saturated sandstone and brine-saturated sandstone.

The MAP and MMAP predictors based on the approximate posterior model are assessed when the model parameters are assumed to be known. These predictors are compared with the MMAP predictor based on the correct posterior model.

Realizations from the correct posterior response model, $p(\mathbf{r}|\mathbf{d})$, are generated.

We estimate the transition matrix, $\mathbf{P}_{\boldsymbol{\kappa}}$, using the approximate EM-algorithm under two different assumptions. First, we have no restrictions on $\mathbf{P}_{\boldsymbol{\kappa}}$. Secondly, we require some of the elements in $\mathbf{P}_{\boldsymbol{\kappa}}$ to be zero, imposing an optimization of a lower dimension.

Finally, a univariate optimization of the model parameter in the Ricker acquisition convolution kernel is studied.

The model parameters are assumed to be similar to the ones given in Lindberg (2010).

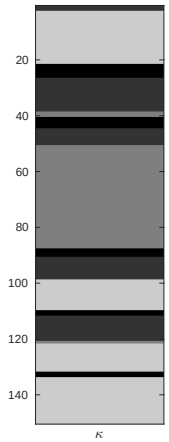


Figure 7.1: Reference profile, $\boldsymbol{\kappa}$.

7.1 Model Specification

We assume the first order Markov chain describing the lithofacies to be defined by the transition matrix

$$\mathbf{P}_{\boldsymbol{\kappa}} = \begin{pmatrix} 0.94 & 0 & 0 & 0.06 \\ 0.04 & 0.91 & 0 & 0.05 \\ 0.01 & 0.02 & 0.95 & 0.02 \\ 0.02 & 0.02 & 0.11 & 0.85 \end{pmatrix}, \tag{7.1}$$

with stationary distribution $(0.2309, 0.1398, 0.4326, 0.1966)$. Classes one to four are assigned various shades of grey. Transitions in the vertical downward direction are described by the Markov chain. By counting the transitions in the reference profile, the empirical transition matrix is found to be

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\text{emp}} = \begin{pmatrix} 0.9464 & 0 & 0 & 0.0536 \\ 0.0250 & 0.9250 & 0 & 0.0500 \\ 0.0541 & 0.0811 & 0.8649 & 0 \\ 0.0625 & 0 & 0.2500 & 0.6875 \end{pmatrix}, \tag{7.2}$$

having stationary distribution $(0.4499, 0.2268, 0.2098, 0.1134)$. The empirical transition matrix overestimates the proportion of class one, at the cost of class three in particular. Except p_{43} and p_{44}, the main characteristics of $\mathbf{P}_{\boldsymbol{\kappa}}$ are observed in empirical transition matrix.

The class response models are defined by $\boldsymbol{\mu}_{r|\boldsymbol{\kappa}'} = (15.74, 15.80, 15.87, 16.01)^\top$ and $\boldsymbol{\sigma}_{r|\boldsymbol{\kappa}'} = (0.0200, 0.0173, 0.0141, 0.0100)^\top$. In Fig. 7.2 we present the class response densities, and they are displayed with the appropriate grey-scale. The Gaussian mixture and Gaussian approximation are displayed with a dashed and dotted line, respectively.

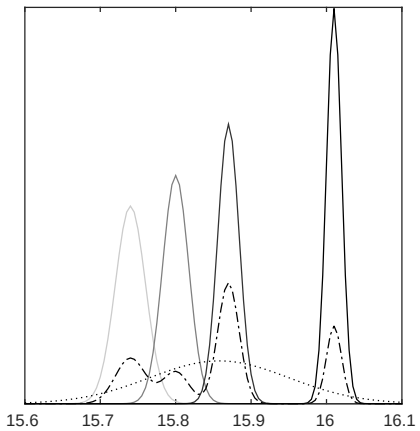


Figure 7.2: Class response densities displayed with solid lines. The Gaussian mixture is displayed with a dashed line, and the Gaussian approximation with a dotted line.

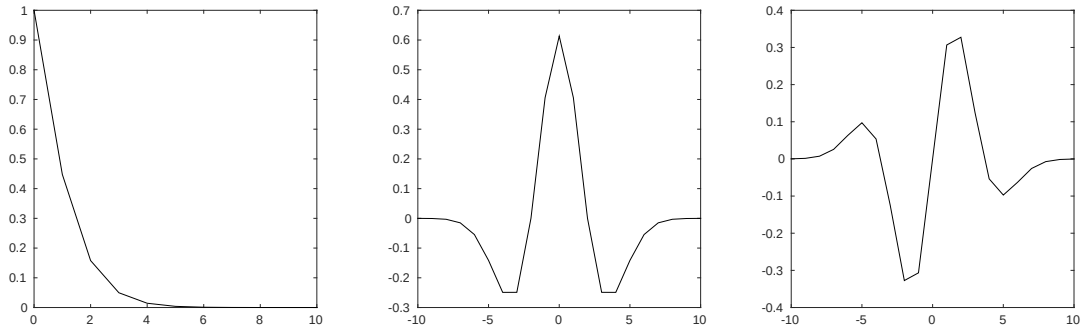
We assume the spatial correlation function to be a powered exponential,

$$\rho_{\mathbf{r}}(h) = \exp\left\{-\left(\frac{h}{1.2}\right)^{1.2}\right\}. \tag{7.3}$$

The spatial correlation function is assumed to be truncated with $a_\rho = 5$, and it is displayed in Fig. 7.3a. We study a Ricker acquisition convolution kernel,

$$\mathbf{w}^{\mathrm{R}}(h;\chi) = \text{const} \times \left(1 - \frac{h^2}{\chi^2}\right) \times \exp\left\{-\frac{n^2}{2 \cdot \chi^2}\right\}, \tag{7.4}$$

where χ is the model parameter. We assume $\chi = 2$, which entails a fairly short Ricker kernel. The discretized Ricker convolution kernel is stored in a matrix, $\mathbf{W}^{\mathrm{R}}$, and normalize each row. The Ricker kernel is shown in Fig. 7.3b. We define a differential operator $\mathbf{D} = \text{tridiag}(-0.5, 0, .5)$, i.e. a band matrix with diagonal band $(-0.5, 0, 0.5)$. The acquisition convolution kernel is defined as $\mathbf{W} = \mathbf{W}^{\mathrm{R}}\mathbf{D}$, having a contrast kernel $\mathbf{w}$. The acquisition convolution kernel, $\mathbf{W}$, is truncated to be of band width 21, i.e. $a_w = 10$. The contrast kernel is presented in Fig. 7.3c.



(a) $\rho_{\mathbf{r}}(h)$ (b) $\mathbf{w}^{\mathrm{R}}$ (c) $\mathbf{w}$

Figure 7.3: Spatial correlation function, $\rho_{\mathbf{r}}(h)$, Ricker acquisition convolution kernel, $\mathbf{w}^{\mathrm{R}}$, and contrast kernel, $\mathbf{w}$.

The observational error is assumed to be $\sigma^2_{\mathbf{d}|\mathbf{r}} = 0.01$. In Fig. 7.4 the response profile, $\mathbf{r}$, and observations, $\mathbf{d}$, are given. Indeed, part of the characteristics from the response profile are preserved in the observations, however the observations appear as highly fluctuating. The class transitions are partly identifiable in the response profile. Since we have included an approximation of the derivative, the observations appear as relative contrasts, and the class response means are unidentifiable.

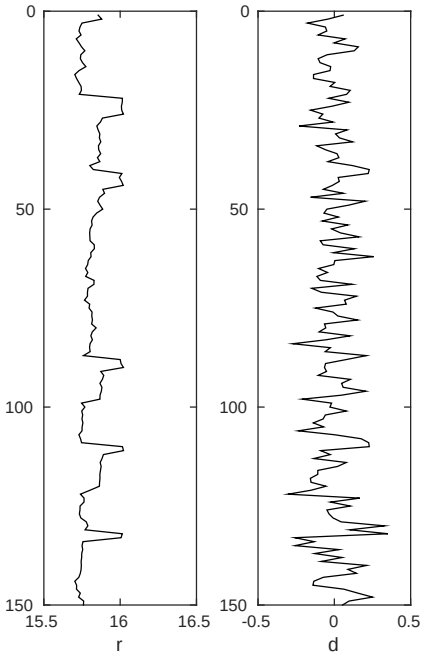


Figure 7.4: Response profile, $\mathbf{r}$, and observations, $\mathbf{d}$.

7.2 Results

We consider only the third order projection approximation since it is found to be a reasonable approximation in Chapter 5.

7.2.1 MAP Prediction

We consider the model parameters to be known and fixed in this section. The MAP and MMAP predictors are displayed in Fig. 7.5, together with the marginal probabilities and reference profile. Both the MAP and MMAP predictors are observed to reproduce the main characteristics of the reference profile. The predictors appear to be too smooth, which is as expected. The MAP predictor is closer to the reference profile than the MMAP predictor is. Indeed, the MMAP predictor is observed to overestimate the proportion of class three.

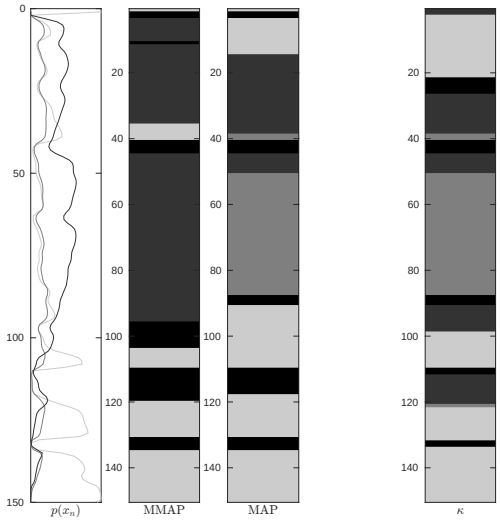


Figure 7.5: Marginal probabilities, and MAP and MMAP predictors for $p^{(k)}(\boldsymbol{\kappa}|\mathbf{d};\mathbf{P}_{\boldsymbol{\kappa}})$, given together with the reference profile.

We generate 50,000 realizations from the correct posterior model, $p(\boldsymbol{\kappa}|\mathbf{d})$, by the independent proposal McMC MH-algorithm. We discard the 10,000 first as a burn-in period. The acceptance rate is found to be 0.0181. Together with the marginal probabilities, MMAP predictor and reference profile, 5,000 realizations from the correct posterior model are given in Fig. 7.6. The marginal probabilities reproduce the variability in the correct posterior model. The MMAP predictor slightly overestimates the proportion of the dark-grey class. Only some of the main characteristics in the reference profile are found by the MMAP predictor based on the realizations.

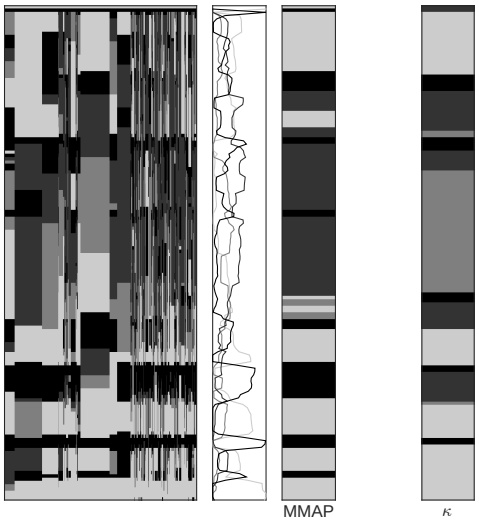
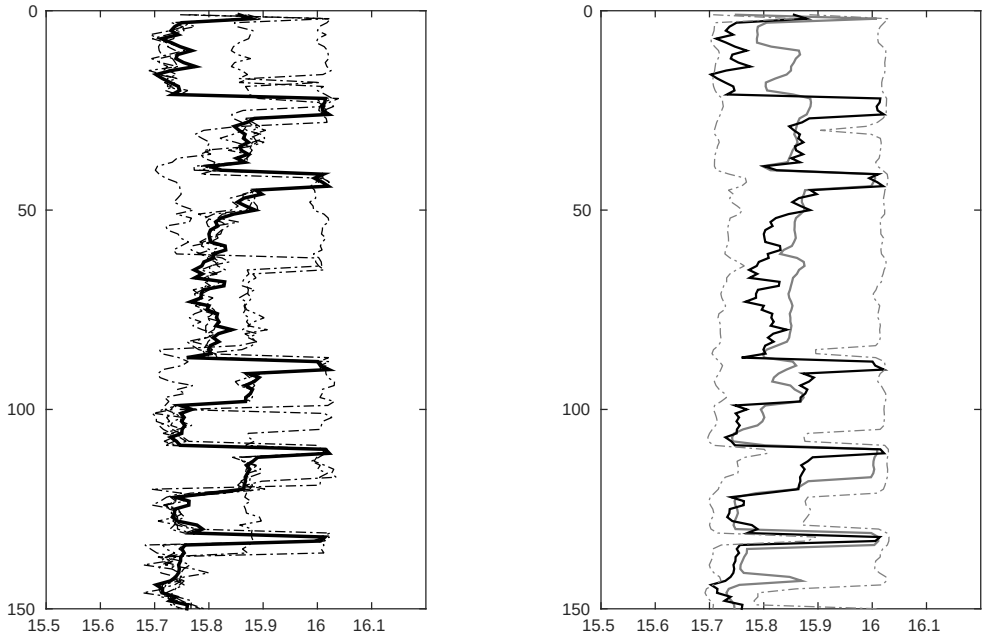


Figure 7.6: Realizations from the correct posterior model $p(\boldsymbol{\kappa}|\mathbf{d})$, together with the marginal probabilities and MMAP predictor, together with the reference profile.

7.2.2 Simulation from the Response Model

We refer to Section 2.3.1 where we have discussed the model studied in Grana and Della Rossa (2010). We generate 5,000 realizations from $p(\mathbf{r}|\mathbf{d})$ using the realizations generated in Section 7.2.1.

In Fig. 7.7a the response profile is shown with a solid line, together with five conditional realizations from $p(\mathbf{r}|\mathbf{d})$, displayed with dashed lines. The conditional realizations reproduce some of the characteristics in the response profile, $\mathbf{r}$. The MMAP predictor $[\widehat{\mathbf{r}|\mathbf{d}}]$ is given together with its approximate 95% posterior range in Fig. 7.7b. Indeed, the uncertainty in $[\boldsymbol{\kappa}|\mathbf{d}]$, entails uncertainty in $[\mathbf{r}|\mathbf{d}]$. The approximate 95% posterior range is found to be fairly wide, however the main characteristics of $\mathbf{r}$ are captured in the MMAP predictor.



(a) Conditional simulations (b) MMAP

Figure 7.7: Response profile, $\mathbf{r}$, and five conditional simulations from $p(\mathbf{r}|\mathbf{d})$, together with the MMAP predictor and approximate 95% posterior range.

7.2.3 Estimation of the Transition Matrix

The transition matrix is now assumed to be unknown. It is assessed using the approximate EM-algorithm. Since the reference profile is of limited length, we expect the estimates to resemble the empirical transition matrix more than the correct transition matrix. Initially, we assume the structure of $\mathbf{P}_{\boldsymbol{\kappa}}$ to be unknown. The approximate EM-algorithm is run in total 25 times, with different initial transition matrices to ensure convergence to the correct maximum log-likelihood value.

In Fig. 7.8 five different trace plots based on the log-likelihood values are given. The log-likelihoods are not necessarily non-decreasing since we consider the approximate log-likelihoods. The transition matrices converging to the highest log-likelihood value are found to be identical, and is given as

$$\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}} = \begin{pmatrix} 0.8379 & 0.0736 & 0.0886 & 0.0000 \\ 0.4642 & 0.5358 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.8603 & 0.1397 \\ 0.0000 & 0.2137 & 0.0000 & 0.7863 \end{pmatrix}, \tag{7.5}$$

with stationary distribution $(0.4171, 0.1457, 0.2644, 0.1728)$. The stationary distribution is observed to be fairly close to the empirical stationary distribution. Indeed, the estimated transition matrix and correct transition matrix have some similarities. In particular, p_{21} and p_{42} are overestimated, while p_{22} is severely underestimated. Some of the smaller transitions probabilities in the transition matrix are estimated to be zero.

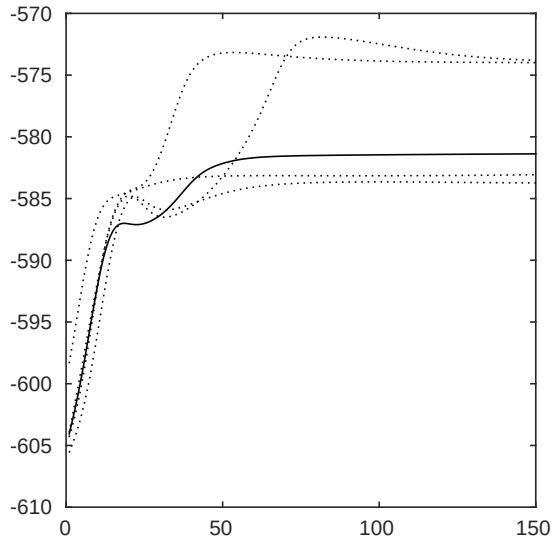


Figure 7.8: Trace plot of the log-likelihoods based on five different initial transition matrices.

The marginal probabilities based on $\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}$ are given together with the MAP and MMAP predictors in Fig. 7.9. Parts of the variability in the reference profile are captured by the MAP predictor. Indeed, the MAP predictor in Fig. 7.9 is fairly close to the MAP predictor in Fig. 7.5. The marginal probabilities, and thereby the MMAP predictor, are found to be poor.

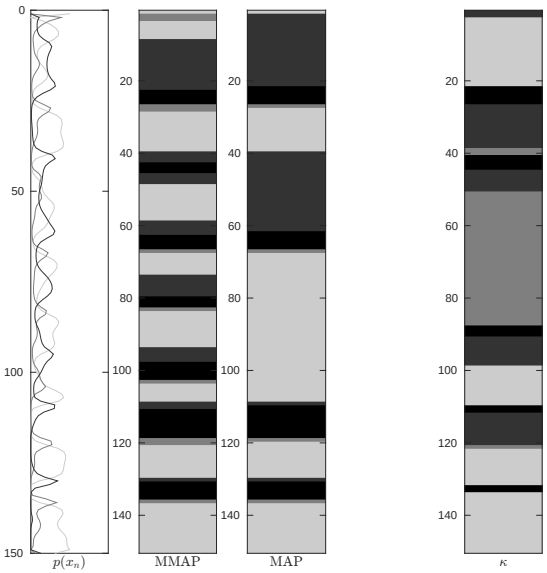


Figure 7.9: Marginal probabilities, and MAP and MMAP predictor based on $p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\mathbf{P}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}\right)$, together with the reference profile.

As discussed in Section 6.2, it is possible to enforce zero probability transitions in the estimate. We assume p_{12}, p_{13} and p_{23} to be equal to zero in the initial transition matrix. Zero transitions in the prior model are preserved in the likelihood approximation, and hence also in the approximate posterior model. As before, the approximate EM-algorithm is run 25 times with different initial conditions. The estimated transition matrix is given as

$$\hat{\hat{\mathbf{P}}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}} = \begin{pmatrix} 0.8713 & 0 & 0 & 0.1287 \\ 0.0305 & 0.8103 & 0 & 0.1592 \\ 0.0000 & 0.0293 & 0.9707 & 0.0000 \\ 0.0000 & 0.0000 & 0.1989 & 0.8011 \end{pmatrix}. \tag{7.6}$$

Its stationary distribution is given as $(0.0273, 0.1154, 0.7472, 0.1101)$, which severely overestimates the proportion of the third class. Compared to the correct transition matrix, the estimates are fairly close.

In Fig. 7.10 trace plots based on various log-likelihoods are given. The log-likelihood appears to have multiple local maximums as in Fig. 7.8. The maximization does not necessarily provide non-decreasing sequences of the log-likelihoods, since we consider likelihood approximations.

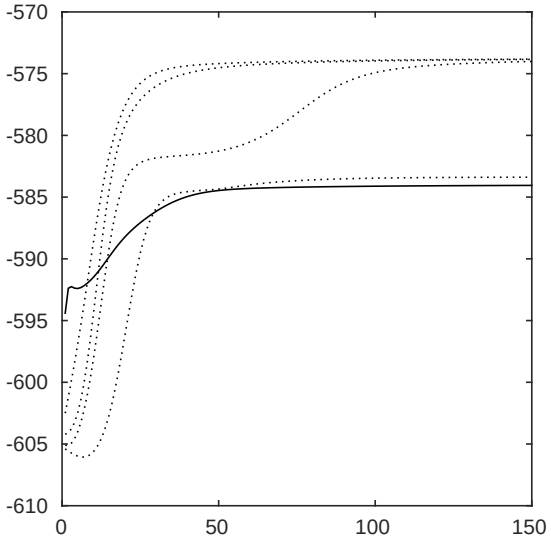


Figure 7.10: Trace plot of the log-likelihoods based on five different initial transition matrices.

The marginal probabilities, and MAP and MMAP predictors for $p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\hat{\mathbf{P}}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}\right)$ are given in Fig. 7.11. The MAP predictor looses small-scale variability compared to Fig. 7.5. Both predictors favour the dark-grey class, i.e. class three. This should come as no surprise since the estimated stationary distribution favours class three, compared to the correct stationary distribution.

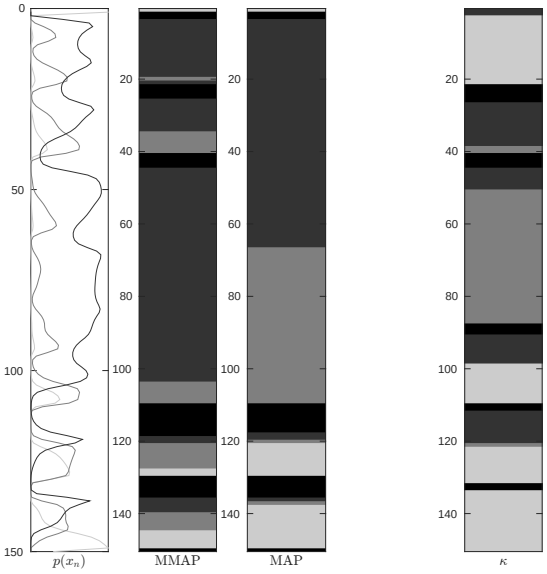


Figure 7.11: Marginal probabilities, and MAP and MMAP predictor based on $p^{(k)}\left(\boldsymbol{\kappa}|\mathbf{d};\hat{\hat{\mathbf{P}}}_{\boldsymbol{\kappa}}^{\mathrm{aEM}}\right)$, together with the reference profile.

7.2.4 Estimation of the Acquisition Convolution Kernel

We consider the model parameter χ in the Ricker acquisition convolution kernel to be unknown. A univariate optimization is considered when the remaining model parameters are assumed to be known. For simplicity, we impose the restriction $\chi \in [1.5, 3.0]$. The optimization procedure is presented in Chapter 4. The correct value is given as $\chi = 2$.

The univariate marginal likelihood is displayed in Fig. 7.12, being maximized for $\hat{\chi} = 2.375$. Compared to the correct value, the estimate is too high. Numerical experiments indicate that the model parameter tends to be overestimated when we increase the range of the spatial correlation function. Compared to Lindberg (2010), we see that the spatial correlation function may lead to overestimation of χ.

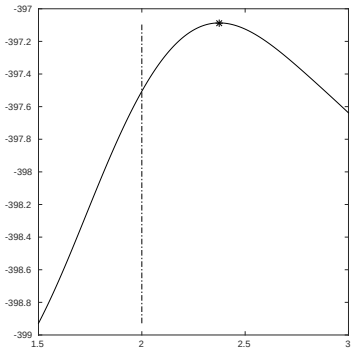


Figure 7.12: Marginal likelihood function. MMLE indicated with an '*', and the correct maximum is shown with a dashed line.

7.3 Closing Remarks

A synthetic seismic test case is studied. We are able to provide fairly reliable predictions if the model parameters are assumed to be known. In particular, the MAP predictor captures the variability in the reference profile. Unfortunately, the acceptance rate is fairly low, slightly below two percent.

Indeed, we are able to generate realizations from $p(\mathbf{r}|\mathbf{d})$. Since the realizations from $p(\boldsymbol{\kappa}|\mathbf{d})$ fluctuate rapidly, each conditional realization from $p(\mathbf{r}|\mathbf{d})$ do not necessarily reproduce the response profile.

Without prior information concerning the zero probability transitions, the estimated transition matrix is found to be a poor estimate. Convergence of the approximate EM-algorithm is dependent on the initial transition matrix. The MAP and MMAP predictors are found to be fairly different.

A univariate optimization of the model parameter in the Ricker acquisition convolution kernel is found to be feasible. We have reason to believe that the spatial correlation leads to overestimation.

Chapter 8

Conclusions and Future Work

In this thesis we study a one dimensional categorical random field, which can for example represent a vertical profile through a geological unit. We consider a convolutional Markov model. The bottom level contains latent categorical variables. Given the categorical variables, the response variables define a latent continuous response model. Convolved observations are collected along the profile, and previously studied models are extended by including spatial correlation in the response model. We assess the correct posterior model and its model parameters. Relevant theory and models are introduced and discussed.

Two different likelihood approximations are proposed, namely the truncation and projection based approximation. The approximations are studied for varying order, k. Evaluation of the truncation approximation has a slightly lower computational cost than evaluation of the projection approximation. The truncation approximation is exact if k is sufficiently large, but then the computational cost is infeasible. We assess the approximate posterior models by the Forward-Backward algorithm. The MAP and MMAP predictors for the approximate posterior models represent the main characteristics of the categorical variables. The predictors are seen to be stable for increasing values of k. The predictors for the approximate posterior model based on the projection approximation are observed to capture more of the small-scale variability in the correct posterior model.

The approximate posterior models are used as proposal densities in an independent proposal McMC algorithm to generate realizations from the correct posterior model. The acceptance rate is defined as a measure to quantify the similarities between the approximate posterior models and the correct posterior model. We obtain higher acceptance rates when we increase k. A response model with high class variances, and a short spatial correlation range, yields the highest acceptance rates. For a higher order projection approximation with a short spatial correlation range we obtained an acceptance rate above 50%. The projection based approximation is in general found to be preferable compared to the truncation approximation, in particular for lower order approximations. If the convolution kernel and the spatial dependencies are short, the relative difference between the approximations is observed to be small.

The realizations from the correct posterior model generated by the McMC algorithm are seen to represent more of the heterogeneity than the predictors based on the approximate posterior models, as expected.

Since the computational cost increases exponentially when increasing k, we conclude that the projection approximation of third or fifth order is favourable.

Estimation of the transition matrix, i.e. the prior model parameters, for the categorical variables based on the approximate EM-algorithm and McMC sampling are found to be feasible. The approximate EM-algorithm provides point estimates at a low computational cost, while uncertainty statements are also provided by the McMC estimates at a higher computational cost. Prior knowledge about zero-probabilities in the transition matrix is seen to be important. A univariate optimization of the MML for the model parameter in the Ricker acquisition convolution kernel is studied, and found feasible. This should be studied in greater detail in the future.

Topics for future research might include a study of the transition matrix, when the dimension is unknown, and model parameter estimation in the response model. The former can for example be done by extending the reversible jump methodology, or comparing models of different dimensions with an appropriate criterion.

Our model may be generalized by assuming the response variables to be dependent on the elastic material properties P-wave velocity, S-wave velocity and density. We may also have angle dependent observations, and extend to two dimensions.

Bibliography

Abrahamsen, P. (1997). A Review of Gaussian Random Fields and Correlation Functions. Technical Report 917, Norwegian Computing Center, Oslo, Norway.

Amaliksen, I. (2014). Bayesian Inversion of Time-lapse Seismic Data using Bimodal Prior Models. Master's thesis, Norwegian University of Science and Technology.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970, 02). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Statist. 41*(1), 164–171.

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

Cappe, O., E. Moulines, and T. Ryden (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Casella, G. and R. Berger (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Celeux, G. and J. Diebolt (1992). A stochastic approximation type EM algorithm for the mixture problem. *Journal of the Royal Statistician Society 67*, 235–51.

Cheng, Q., R. Chen, and T. Li (1996). Simultaneous wavelet estimation and deconvolution of reflection seismic signals. *IEEE T. Geoscience and Remote Sensing 34*(2), 377–384.

Cressie, N. A. C. (1993). *Statistics for Spatial Data* (Revised Edition ed.). Wiley.

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

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Eidsvik, J., T. Mukerji, and P. Switzer (2004). Estimation of geological attributes from a well log: An application of hidden Markov chains. *Math. Geology 36*(3).

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing 16*(2), 203–213.

Fearnhead, P. and P. Clifford (2003). On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(4), 887–899.

Friel, N. and H. Rue (2007). Recursive Computing and Simulation-Free Inference for General Factorizable Models. *Biometrika 94*(3), pp. 661–672.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer Series in Statistics. Springer.

Grana, D. and E. Della Rossa (2010). Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic invervion. *Geophysics* (3), O21–O37.

Grana, D., T. Mukerji, L. Dovera, and E. Della Rossa (2012). Sequential Simulations of Mixed Discrete-Continuous Properties: Sequential Gaussian mixture Simulation. In *Geostatistics Oslo 2012, Quantitative Geology and Geostatistics*, Volume 17, pp. 239–250.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hershey, J. R. and P. A. Olsen (2007). Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.*, Volume 4, pp. IV–317–IV–320. IEEE.

Karr, A. (1993). *Probability*. Springer texts in statistics. Springer-Verlag.

Killick, R., P. Fearnhead, and I. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*(500), 1590–1598.

Künsch, H. (2001). State space and hidden Markov models. In D. R. C. O. E. Barndorff-Nielsen and C. Klüppelberg (Eds.), *Complex Stochastic Systems*, pp. 109–173. CRC Press.

Levin, D., Y. Peres, and E. Wilmer (2008). *Markov Chains and Mixing Times*. American Mathematical Society.

Lindberg, D. (2010). Parameter estimation in convolved categorical models. Master's thesis, Norwegian University of Science and Technology.

Lindberg, D. and H. Omre (2014a). Blind Categorical Deconvolution in Two-Level Hidden Markov Models. *Geoscience and Remote Sensing, IEEE Transactions on 52*(11), 7435–7447.

Lindberg, D. and H. Omre (2014b). Inference of the Transition Matrix in Convolved Hidden Markov Models by a Generalized Baum-Welch Algorithm. *Submitted for publication*.

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

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

Robert, C. P. and G. Casella (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Ross, S. (2006). *Introduction to Probability Models*. Elsevier Science.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Scott, S. L. (2002). Bayesian Methods for Hidden Markov Models: Recursive Computing in the 21st Century. *Journal of the American Statistical Association 97*, 337–351.

Sung, H. G. (2004). *Gaussian Mixture Regression and Classification*. Ph. D. thesis, Rice University.

Viterbi, A. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory 13*(2), 260–269.

Wei, G. and M. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. American Statist. Assoc. 85*, 699–704.

Appendices

Appendix A

Probability Distributions

A.1 Gaussian Distribution

Definition 1 (Multivariate Gaussian Distribution)**.** A random vector $\mathbf{x} = (x_1, \ldots, x_T)^\top$ is said to have the multivariate Gaussian distribution if for every constant vector $\mathbf{a} \in \mathbb{R}^T$, $Y = \mathbf{a}^\top \mathbf{x}$ has a univariate Gaussian distribution. The multivariate Gaussian probability density function is given as

$$\phi_T(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{T}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \tag{A.1}$$

where $\boldsymbol{\mu}$ is the expectation T-vector and $\boldsymbol{\Sigma}$ is the positive definite covariance $(T \times T)$-matrix.

Theorem A.1 (Conditional Gaussian)**.** *Assume* $\mathbf{x} \sim \phi_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where*

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{A.2}$$

Here, $\mathbf{x}_1 = (x_1, \ldots, x_t)^\top$ *and* $\mathbf{x}_2 = (x_{t+1}, \ldots, x_T)^\top$*. The conditional distribution* $[\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{a}]$ *is then also Gaussian,* $\phi_t\left(\boldsymbol{\mu}_{\mathbf{x}_1|\mathbf{x}_2}, \boldsymbol{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2}\right)$ *where*

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}_1|\mathbf{x}_2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{a} - \boldsymbol{\mu}_2), \\ \boldsymbol{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned} \tag{A.3}$$

Theorem A.2 (Linear combination of Gaussian)**.** *Let* $\mathbf{x} \sim \phi_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. Let* $\mathbf{A} \in \mathbb{R}^{M \times T}$ *be of full rank and* $\mathbf{y} \in \mathbb{R}^M$*. Then* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ *is also Gaussian with*

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}} &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \\ \boldsymbol{\Sigma}_{\mathbf{y}} &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top. \end{aligned} \tag{A.4}$$

Corollary A.2.1. *Assume* $\mathbf{y} \sim \phi_T\left(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}}\right)$*,* $\mathbf{M} \in \mathbb{R}^{M \times T}$*,* $\mathbf{b} \in \mathbb{R}^M$ *and* $[\mathbf{x}|\mathbf{y}] = \mathbf{M}\mathbf{y} + \mathbf{b}$*, e.g.* $[\mathbf{x}|\mathbf{y}] = \phi_M\left(\mathbf{M}\mathbf{y} + \mathbf{b}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\right)$*. Then the marginal for* $\mathbf{x}$ *is*

$$p(\mathbf{x}) = \phi_M\left(\mathbf{M}\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{b}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} + \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{M}^\top\right). \tag{A.5}$$

A.2 Dirichlet Distribution

Definition 2 (Dirichlet Distribution)**.** Let $\mathbf{x} = (x_1, \ldots, x_T)^\top$ be a Dirichlet distributed random vector defined for $\{0 < x_i < 1; i = 1, \ldots, T\}$ and $\sum_{i=1}^{T} x_i = 1$. The Dirichlet probability density function, with scale parameter vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)$ where $\alpha_i > 0 \, \forall i$, is defined as

$$p(\mathbf{x}) = \frac{\Gamma\left(\sum_{i=1}^{T} \alpha_i\right)}{\prod_{i=1}^{T} \Gamma(\alpha_i)} \times \prod_{i=1}^{T} x_i^{\alpha_i - 1}, \tag{A.6}$$

where $\Gamma(\cdot)$ is the gamma-function.

Appendix B

Generalized Forward-Backward Algorithm

We extend the Forward-Backward algorithm described in Section 3.2 to a higher order factorial form model, following the lines of Reeves and Pettitt (2004), and Friel and Rue (2007). We consider a model on factorial form, as in Eq. (2.28),

$$p(x_1, \ldots, x_n) = \prod_{i=r+1}^{n} p_i(x_{i-r}, \ldots, x_i), \tag{B.1}$$

where p_i are densities. The normalization constant, z, can be recursively computed since

$$z = \sum_{x_n} \cdots \sum_{x_m} p_m(x_m, \ldots, x_n) \sum_{x_{m-1}} p_{m-1}(x_{m-1}, \ldots, x_{n-1}) \cdots \sum_{x_1} p_1(x_1, \ldots, x_{r+1}). \tag{B.2}$$

The recursive procedure, often referred to as the forward recursion, is given as following

$$\begin{aligned} z_1(x_2, \ldots, x_{r+1}) &= \sum_{x_1} p_1(x_1, \ldots, x_{r+1}) \\ z_i(x_{i+1}, \ldots x_{r+i}) &= \sum_{x_i} p_i(x_i, \ldots, x_{i+r}) z_{i-1}(x_i, \ldots, x_{i+r-1}) \quad \text{for } i = 2, \ldots, m. \\ z &= \sum_{x_{m+1}} \cdots \sum_{x_n} z_m(x_{m+1}, \ldots, x_n) \end{aligned} \tag{B.3}$$

The joint density is computed recursively by a backward step since

$$p(x_1, \ldots, x_n) = p(x_{m+1}, \ldots, x_n) \prod_{i=1}^{m} p(x_i | x_{i+1}, \ldots x_n). \tag{B.4}$$

The backward transitions are given as

$$\begin{aligned} p(x_{m+1}, \ldots, x_n) &= \frac{z_m(x_{m+1}, \ldots, x_n)}{z} \\ p(x_i | x_{i+1}, \ldots x_{i+r}) &= \frac{p_i(x_i, \ldots, x_{i+r}) z_{i-1}(x_i, \ldots, x_{i+r-1})}{z_i(x_{i+1}, \ldots, x_{i+r})} \quad \text{for } i = m, \ldots, 2. \\ p(x_1 | x_2, \ldots, x_{r+1}) &= \frac{p_1(x_1, \ldots, x_{1+r})}{z_1(x_2, \ldots, x_{r+1})} \end{aligned} \tag{B.5}$$

Friel and Rue (2007) noted that by storing the normalization constants found in the forward recursion, it is possible to generate $(x_1, \ldots, x_n)$ in a reverse index order, i.e. from

n down to 1. Similarly, from Eq. (B.4) we can sequentially find the mode from a recursive scheme, similar as in Eq. (3.43). That is,

$$\begin{aligned}(\hat{x}_{m+1},\ldots,\hat{x}_n) &= \underset{(x_{m+1},\ldots,x_n)}{\arg\max}\; p\left(x_{m+1},\ldots,x_n\right)\\ (\hat{x}_i|\hat{x}_{i+1},\ldots,\hat{x}_{i+r}) &= \underset{x_i}{\arg\max}\, p\left(x_i|\hat{x}_{i+1},\ldots,\hat{x}_{i+r}\right) \quad \text{for } i=m,\ldots,1\end{aligned}. \tag{B.6}$$

A possibility is to record ties between two or more different paths, and then sample uniformly between the different paths in the backward step. We refer to Eq. (B.6) as the extended Viterbi algorithm, which is an extension of the algorithm presented in Viterbi (1967), which produces the MAP predictor.

The MMAP predictor is evaluated similarly by maximizing $p(x_i)$ for $i = 1,\ldots,N$. Indeed,

$$p(x_{m+1},\ldots,x_n) = \frac{z_m(x_{m+1},\ldots,x_n)}{z}. \tag{B.7}$$

Since $p\left(x_i|x_{i+1},\ldots x_{i+r}\right)$ is available from the backward recursion, we assess

$$p(x_i,\ldots,x_{i+r-1}) = \sum_{x_{i+r}} p(x_i|x_{i+1},\ldots,x_{i+r})p(x_{i+1},\ldots,x_{i+r}) \tag{B.8}$$

recursively. Finally, the MMAP predictor are obtained by summing out the remaining indices, i.e.

$$p(x_i) = \sum_{x_{i+1}} \cdots \sum_{x_{i+r-1}} p(x_i,\ldots,x_{i+r-1}). \tag{B.9}$$

Friel and Rue (2007) noted that the generalized Forward-Backward algorithm is extremely useful in practice since it amounts to reuse the same algorithm twice. First a forward recursion with increasing indices, and then a forward recursion with decreasing indices.

Until now we have only considered the generalized Forward-Backward algorithm for a general factorial form model. Indeed, the densities p_i in Eq. (B.1) may be likelihoods, which need not be normalized. Therefore, we define

$$p_n\left(\boldsymbol{\kappa}_n^{(k)}\right) \stackrel{\text{def}}{=} p^{(k)}\left(\mathbf{d}|\boldsymbol{\kappa}_n^{(k)}\right) p\left(\kappa_n|\kappa_{n-1}\right) \tag{B.10}$$

for $n = k,\ldots,N$. Since our k-th order approximation in Eq. (3.10) is a lag-$(k-1)$ general factorisable model, we assess it using the generalized Forward-Backward algorithm. Note that this only holds for $k \geq 2$, but for $k = 1$ the forward and backward recursions simplify to the well-known Forward-Backward algorithm presented in Section 3.2. Since the prior model is already on factorial form, we need not approximate it. The prior model could in fact have been a k-th order Markov chain without increasing the lag of the posterior model.

The forward recursion evaluates the normalization constant, which we denote $z_d^{(k)}$ for the k-th order approximation. The normalization constant appears in Eq. (3.10) as $p^{(k)}\left(\mathbf{d}\right) = z_d^{(k)}$. The forward and backward recursions for Eq. (3.10) are presented in Alg. 6 and Alg. 7. Evaluations of the approximate posterior for a general lag-$(k-1)$ factorisable model can be exactly assessed in $\mathcal{O}\left((N-k)K^k\right)$ operations, using the generalized Forward-Backward algorithm. In fact, the approximate posterior model is exact up to the approximation of

Algorithm 6: Forward recursion for a general factorisable model

Result: Normalization constant, z, to Eq. (3.10)
Initial step
$z_1\left(\boldsymbol{\kappa}_k^{(k-1)}\right) = \sum_{\kappa_1} l^{(k)}\left(\boldsymbol{\kappa}_k^{(k)}\right) p\left(\kappa_k|\kappa_{k-1}\right)$
for $t = 2$ **to** $n-k+1$ **do**
$z_t\left(\boldsymbol{\kappa}_{t+k-1}^{(k-1)}\right) = \sum_{\kappa_t} l^{(k)}\left(\boldsymbol{\kappa}_{t+k-1}^{(k)}\right) z_{t-1}\left(\boldsymbol{\kappa}_{t+k-2}^{(k-1)}\right) p\left(\kappa_{t+k-1}|\kappa_{t+k-2}\right)$
end
$z = \left[\sum_{\kappa_{t-k+2}} \cdots \sum_{\kappa_n} z_{n+k-1}\left(\kappa_n^{(k-1)}\right)\right]^{-1}$
return z

Algorithm 7: Backward recursion for a general factorisable model

Result: Backward probabilities $p(\boldsymbol{\kappa}|\mathbf{d})$
Initial step
$p\left(\boldsymbol{\kappa}_n^{(k-1)}|\mathbf{d}\right) = z \times z_{n-k+1}\left(\boldsymbol{\kappa}_n^{(k-1)}\right)$
for $t = n-k+1$ **to** 2 **do**
$p\left(\kappa_t|\boldsymbol{\kappa}_{t+k-1}^{(k-1)}, \mathbf{d}\right) = \frac{p(\kappa_{t+k-1}|\kappa_{t+k-2}) \times l^{(k)}\left(\boldsymbol{\kappa}_{t+k-1}^{(k)}\right) \times z_{t-1}\left(\boldsymbol{\kappa}_{t+k-2}^{(k-1)}\right)}{z_{t+k-1}\left(\boldsymbol{\kappa}_{t+k-1}^{(k)}\right)}$
end
$p\left(\kappa_1|\boldsymbol{\kappa}_k^{(k-1)}, \mathbf{d}\right) = \frac{p(\kappa_k|\kappa_{k-1}) \times l^{(k)}\left(\boldsymbol{\kappa}_k^{(k)}\right)}{z_1\left(\boldsymbol{\kappa}_k^{(k-1)}\right)}$
$p\left(\boldsymbol{\kappa}|\mathbf{d}\right) = p\left(\boldsymbol{\kappa}_n^{(k-1)}|\mathbf{d}\right) \times \prod_{t=1}^{k} p\left(\kappa_t|\boldsymbol{\kappa}_{t+k-1}^{(k-1)}, \mathbf{d}\right)$
return $p\left(\boldsymbol{\kappa}|\mathbf{d}\right)$

the acquisition likelihood, thus we evaluate the exact posterior if $k = 4a_w + 2a_\rho + 1$. This requires a sum over K^k elements.

For a general factorisable model the MAP predictor is assessed using Alg. 8. Finally, define $\boldsymbol{\kappa}_n^{(k-1)} \backslash \kappa_n = (\kappa_{n-k+2}, \ldots, \kappa_{n-1})$. The MMAP predictor algorithm is given in Alg 9. In practice, we expect the MAP and MMAP predictors to have similar characteristics, but they are not necessarily identical. In specific, the MMAP predictors are also given with uncertainty statements, since we evaluate the marginal probabilities $p(\kappa_n|\mathbf{d})$ for $n = 1, \ldots, N$.

Algorithm 8: MAP predictor for a general factorisable model

Result: MAP predictor, $\hat{\boldsymbol{\kappa}}$
Initial step
$\hat{\boldsymbol{\kappa}}_N^{(k-1)} = \arg\max_{\boldsymbol{\kappa}_N^{(k-1)}} p\left(\boldsymbol{\kappa}_N^{(k-1)}\right)$
Iterate
for $n = N-k+1$ **to** 2 **do**
$\hat{\kappa}_n = \arg\max_{\kappa_n} p\left(\kappa_n|\hat{\boldsymbol{\kappa}}_{n+k-1}^{(k-1)}\right)$
end
return $\hat{\boldsymbol{\kappa}}$

Algorithm 9: MMAP predictor for a general factorisable model

Result: MMAP predictor, $\hat{\hat{\boldsymbol{\kappa}}}$

for $n = 1$ **to** N **do**

$$\hat{\hat{\kappa}}_n = \begin{cases} \arg\max_{\kappa_n} \left\{ \sum_{\boldsymbol{\kappa}_n^{(k-1)} \setminus \kappa_n} p^{(k)} \left(\boldsymbol{\kappa}_n^{(k-1)} | \mathbf{d} \right) \right\} & \text{if } n \geq k-1 \\ \arg\max_{\kappa_n} \left\{ \sum_{\boldsymbol{\kappa}_{k-1}^{(k-1)} \setminus \kappa_n} p^{(k)} \left(\boldsymbol{\kappa}_{k-1}^{(k-1)} | \mathbf{d} \right) \right\} & \text{if } n < k-1 \end{cases}$$

end

return $\hat{\hat{\boldsymbol{\kappa}}}$