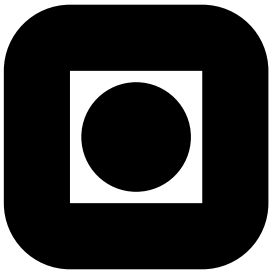
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Collapsed Blocks Approximations

by

Ingelin Steinsland and Håvard Rue

PREPRINT
STATISTICS NO. 9/2005



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL `http://www.math.ntnu.no/preprint/statistics/2005/S9-2005.pdf`
Ingelin Steinsland has homepage: `http://www.math.ntnu.no/~ingelins`
E-mail: `ingelins@math.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Collapsed Blocks Approximations

Ingelin Steinsland & Håvard Rue,
Norwegian University of Science and Technology

October 28, 2005

Abstract

The collapsed blocks approximations (CBAs) are a class of approximations to multivariate distributions based on full conditional distributions for blocks of variables. The CBAs are useful when the full conditional distributions are known and computationally cheap to sample from and evaluate (i.e. for Markov models). In addition knowledge about the dependence structure is needed (e.g. spatial models). The CBAs can then computationally efficiently be sampled from and evaluated. An one-block Metropolis-Hastings algorithm with CBA proposal is set up. This algorithm is used to analyse functional Magnetic Resonance Imaging (fMRI) data modelled as a latent Gaussian Markov random field (GMRF) with Gaussian likelihood This model has more that 100000 variables. The methodology is also extended to include cases where only sampling and evaluation from an approximation to the full conditional distributions are possible. The CBAs aid balancing two often conflicting requirements in MCMC; fast computations and fast mixing.

KEYWORDS: Markov chain Monte Carlo, Gaussian Markov Random field, probability density function approximation, blocking, functional magnetic resonance imaging (fMRI).

1 Introduction

For making inference from spatial and spatio-temporal models we often have data collected at spatial locations which are indirect observations of the phenomena of interest. The phenomena can in a Bayesian setting be modelled as a latent field $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, that is assumed to have a spatial dependence structure described by the prior density $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ with hyper-parameters $\boldsymbol{\theta}$. Popular choices for the latent field are Gaussian random fields (GRFs) and Gaussian Markov random fields (GMRFs). Let $\boldsymbol{y}$ denote the observed data and assume the posterior density reads

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) \prod_i \pi(y_i \mid x_i), \tag{1}$$

where $\pi(\boldsymbol{\theta})$ is the prior of $\boldsymbol{\theta}$, and $\pi(y_i|x_i)$ is the likelihood for one observation. Except in a few special cases where we can do analytical inference on $(\boldsymbol{x}, \boldsymbol{\theta})$, we need to rely on simulation based inference methods using Markov chain Monte Carlo (MCMC).

MCMC methods based on single-site updating schemes have traditionally been most popular, see for example Diggle et al. (1998). In the recent years, there have been an increasing concern about the extremely slow convergence and mixing single-site schemes may have for such models, see for example Christensen et al. (2003), Knorr-Held and Rue (2002) and Rue et al. (2004). A common remedy to improve mixing is to update several variables simultaneously (Liu (1994), Liu et al. (1994) and Carter and Kohn (1996)). This is known as blocking or grouping.

At first sight, a two-block Gibbs-sampler which updates $\boldsymbol{\theta}$ and $\boldsymbol{x}$ in separate blocks,

$$\boldsymbol{\theta} \;\sim\; \pi(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{y}), \quad \text{and} \quad \boldsymbol{x} \;\sim\; \pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}) \tag{2}$$

seems appealing. However, even few hyper-parameters can slow down the convergence considerably, see for example Papaspiliopoulos et al. (2003) and Rue and Held (2005, Ch. 4.1) for theoretical results. Gelfand et al. (1995); Pitt and Shephard (1999); Papaspiliopoulos et al. (2003) suggest using reparameterisation to weaken the dependency between $\boldsymbol{\theta}$ and $\boldsymbol{x}$, an approach which must be specifically tailored to, and is only possible for certain models. A simpler and more general approach was introduced by Knorr-Held and Rue (2002); the acceptance of $\boldsymbol{\theta}$ is simply delayed until $\boldsymbol{x}$ is updated as well, see also Wilkinson (2003). This updating scheme is of the following form where $\boldsymbol{x}^{\text{old}}$ and $\boldsymbol{\theta}^{\text{old}}$ denote the current state of the chain,

$$\boldsymbol{\theta}^{\text{new}} \;\sim\; q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{\text{old}}), \quad \boldsymbol{x}^{\text{new}} \;\sim\; q(\boldsymbol{x} \mid \boldsymbol{\theta}^{\text{new}}, \boldsymbol{y}, \boldsymbol{x}^{\text{old}}) \tag{3}$$

and then $(\boldsymbol{x}^{\text{new}}, \boldsymbol{\theta}^{\text{new}})$ is accepted/rejected jointly with a Metropolis-Hastings (M-H) step. Rue and Held (2005) name such a scheme the *one-block* algorithm. Empirical studies for latent GMRF models (Knorr-Held and Rue (2002) and Gamerman et al. (2003)) have demonstrated that a joint update of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ using the one-block algorithm give appropriate mixing, and is superior to the two-block scheme in (2). This is supported by some theoretical results in Rue and Held (2005, Ch. 4.1). In (3), $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{old})$ is often taken as a simple (log-) random walk proposal, while $\boldsymbol{x}$ is updated from its full-conditional, i.e.

$q(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{x}^{old}) = \pi(\boldsymbol{x}|\boldsymbol{\theta}^{new}, \boldsymbol{y})$, or from some approximation to $\pi(\boldsymbol{x}|\boldsymbol{\theta}^{new}, \boldsymbol{y})$. When the dimension of $\boldsymbol{x}$ is large, sampling from a full dimensional field can be infeasible, and an approximation based on full conditional distributions for blocks seems to be an appealing alternative. Our scope is to construct such approximations.

For many problems we have random variables with known dependence structure and positive dependency. E.g. for most spatial models variables close in space are more dependent than variables further apart. In this paper we utilise our knowledge about the dependence structure, to construct an approximation to $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ that can be sampled from and evaluated. This enables one-block M-H algorithms for problems else too large. In addition to a known dependence structure, our approach requires that sampling from and evaluation of the full conditional distributions for blocks of variables (or an approximation to these) are possible.

To introduce our approximation idea we consider a simple time series $AR(1)$ process of dimension three, $\boldsymbol{x} = (x_1, x_2, x_3)$ with known auto-correlation $0 < \phi < 1$. Our aim is to sample from $\pi(\boldsymbol{x})$ using a M-H algorithm. Assume that only sampling and evaluation from the full conditional for blocks of maximum two variables are possible. Given an initial field $\boldsymbol{x}^{old}$, we can sample $x_1^{new} \sim \pi(x_1|x_2^{old}, x_3^{old})$ and $(x_2^{new}, x_3^{new}) \sim \pi(x_2, x_3|x_1^{new})$. If ϕ is large, $\boldsymbol{x}^{new}$ depends strongly on $\boldsymbol{x}^{old}$ and

$$\pi(x_1|x_2^{old}, x_3^{old}) \cdot \pi(x_2, x_3|x_1) \tag{4}$$

is a poor approximation to $\pi(\boldsymbol{x})$. The dependence on $\boldsymbol{x}^{old}$ comes from $\pi(x_1|x_2^{old}, x_3^{old})$. Because of the time series structure, x_1 depends most on x_2, and by sampling $x_1^{new} \sim \pi(x_1|x_3^{old})$, i.e. integrating out x_2, the interaction is relaxed. We are able to sample from $\pi(x_1|x_3^{old})$ by sampling $(x_1^{new}, x_2^{\text{buffer}}) \sim \pi(x_1, x_2|x_3^{old})$, and simply through away x_2^{buffer}. It can be evaluated by using $\pi(x_1|x_3) = \pi(x_1, x_2|x_3)/\pi(x_2|x_1, x_3)$ for any value of x_2. The dummy variable x_2^{buffer} has a function as a buffer between the new and old field. Note that only sampling and evaluation of full conditional distributions of dimension two are needed. The resulting approximation is

$$\pi(x_1|x_3^{old}) \cdot \pi(x_2, x_3|x_1) \tag{5}$$

and it is a member in the class of approximation we will name *collapsed blocks approximations* (CBAs).

Using (5) as proposal of a M-H algorithm will improve mixing compared to using (4). For the $AR(1)$ process the improvement in convergence can be quantified. Assume $\boldsymbol{x} = (x_1, \ldots, x_n)$ with dimension n instead of three, and consider the approximation

$$\pi(x_1 \mid x_{b+2}^{old}, \ldots, x_n^{old})\pi(x_2, \ldots, x_n \mid x_1) \tag{6}$$

where b determines the length of the buffer. This is a generalisation of (4) (for $b = 0$) and (5) (for $1 \leq b \leq n-2$). It can be shown that the rate of convergence of an MCMC algorithm using (6) as a proposal, is

$$(\xi^2)^{(b+1)/L} \tag{7}$$

where L is the correlation length defined by $|\phi|^L = \xi$, commonly with $\xi = 0.05$. We observe that the effect of collapsing a buffer here depends explicitly on the ratio of the buffer length b and the correlation length L.

The main contribution of this paper is a generalisation of (5). This is our class of CBAs which is introduced in Section 2. A Gibbs sampler approach is taken in Section 3. Further, in Section 4, an one-block M-H algorithm is set up with an alternative to Peskuns acceptance probability, the opposite reverse acceptance probability. In Section 5 the properties of the CBAs are demonstrated through a simulation study. The practical usefulness of CBA is demonstrated in Section 6. A time-series of functional Magnetic Resonance Imaging (fMRI) data are modelled as a hidden GMRF model with Gaussian likelihood. An one-block M-H sampler with CBA as part of the proposal is used for making inference. The paper ends with some closing remarks in Section 7.

2 Collapsed blocks approximation

Consider variables with a spatial structure, e.g. on a lattice, with probability density function $\pi(\boldsymbol{x})$. Assume we can sample from and evaluate the full conditional distribution for blocks of variables, and that the dependency is decreasing with increasing distance. We want to construct an approximation based on the full conditional for blocks and initial values $\boldsymbol{x}^{\text{old}}$, $\pi^*(\boldsymbol{x}; \boldsymbol{x}^{\text{old}}) \approx \pi(\boldsymbol{x})$, that we can sample from and evaluate. We assume there are reasonable initial values available for the field, e.g. the current state of a M-H algorithm. Divide the study area into disjoint contiguous parts, and make blocks of variables corresponding to this. Let $S = \{1, 2, \ldots, n\}$ be the indexes of the variables, and $\{b_1, b_2, \ldots, b_N\}$ the disjoint partition of S that corresponds to the blocks. Define $-b_i = S \backslash b_i$, $b_{<i} = \cup_{j<i} b_j$ and $b_{>i} = \cup_{j>i} b_j$, i.e. all blocks but b_i, all blocks with lower index that i, and all blocks with higher index than i, respectively. We then have;

$$\pi(\boldsymbol{x}) = \pi(\boldsymbol{x}_{b_1})\pi(\boldsymbol{x}_{b_2}|\boldsymbol{x}_{b_1}) \ldots \pi(\boldsymbol{x}_{b_N}|\boldsymbol{x}_{b_{<N}}) = \prod_{i=1}^{N} \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}). \tag{8}$$

Exact sampling from $\pi(\boldsymbol{x})$ is straight forward if these marginal conditional distributions are known and possible to sample from. Though, in the situations considered here, it is only possible to sample from the full conditional for blocks. We will make an approximation to $\pi(\boldsymbol{x})$ by approximating each of the factors in (8), $\pi(\boldsymbol{x}) \approx \pi^*(\boldsymbol{x}; \boldsymbol{x}^{\text{old}}) = \prod_{i=1}^{N} \pi^*(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}; \boldsymbol{x}^{\text{old}})$ where $\pi^*(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}; \boldsymbol{x}^{\text{old}})$ is an approximation to $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}})$ that depends on the initial values $\boldsymbol{x}^{\text{old}}$. If there are any spatial dependence in the field, the full conditional distribution $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{-i}}) = \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}} = \boldsymbol{x}_{b_{<i}}^{new}, \boldsymbol{x}_{b_{>i}} = \boldsymbol{x}_{b_{>i}}^{old})$ depends strongly on $\boldsymbol{x}^{old}$ and is a poor approximation to $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}} = \boldsymbol{x}_{b_{<i}}^{new})$. The approximation will improve if we use distributions that depends less on the initial field. For each block we introduce a buffer containing selected variables in $\boldsymbol{x}_{b_{>i}}$ that the block variables, $\boldsymbol{x}_{b_i}$, depends most on, i.e. those closest in space. If the buffer variables are collapsed, i.e. integrated out of the full conditional distribution, we loosen the connection to the initial values.

Denote the indexes for the buffer variables for block i by β_i ($\subset b_{>i}$), and define $B_i = b_i \cup \beta_i$ and $-B_i = S \backslash B_i$. This collapsed distribution, $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{-B_i}) = \int \pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{-B_i}) d\boldsymbol{x}_{\beta_i}$, depends less on the initial values, and hence is a better approximation to $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}})$. Using this we get an overall approximation;

$$\pi^*(\boldsymbol{x}; \boldsymbol{x}^{\text{old}}) = \prod_{i=1}^{b_N} \pi^*(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}; \boldsymbol{x}^{\text{old}}) = \prod_{i=1}^{b_N} \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{<b_i} = \boldsymbol{x}^{\text{new}}, \boldsymbol{x}_{B_{>i}} = \boldsymbol{x}^{\text{old}}) \tag{9}$$

where $B_{>i} = b_{>i} \backslash \beta_i$. We name this class of approximations the *collapsed blocks approximation* (CBAs). For each block we are able to sample from $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}, \boldsymbol{x}_{B_{>i}})$ by sampling from the full conditional $(\boldsymbol{x}_{b_i}^{\text{new}}, \boldsymbol{x}_{\beta_i}^{\text{buffer}}) \sim \pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|x_{-B_i})$ and only consider $\boldsymbol{x}_{b_i}^{\text{new}}$. Evaluation is possible from $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{-B_i}) = \pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{-B_i}) / \pi(\boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{b_i}, \boldsymbol{x}_{-B_i})$ for any value of $\boldsymbol{x}_{\beta_i}$, e.g. the sampled values. The numerator is the full conditional distribution sampled from, and can in many cases be evaluated without much additional cost. The denominator is the full conditional distribution for the buffer, and it has to be calculated from scratch.

We introduced dependency to $\boldsymbol{x}^{\text{old}}$ because using full conditional distributions give large computational benefits for Markov models. Further we have relaxed the dependency to $\boldsymbol{x}^{\text{old}}$ by collapsing selected variables of $\boldsymbol{x}_{>b_i}$. In this way we are able to balance our conflicting requirements of fast computations and independence of $\boldsymbol{x}^{\text{old}}$ / fast mixing.

For an illustration of the CBA see Figure 1, left. The blocks and buffers are here chosen such that the computational cost is equal when sampling each block, i.e. the number of variables in $\boldsymbol{x}_{B_i}$ is constant. For the first block $b_1 = \{d_1\}$, $\beta_1 = \{d_2, d_4, d_5\}$, $B_1 = \{d_1, d_2, d_4, d_5\}$ and $-B_1 = \{d_3, d_6, d_7, d_8, d_9\}$. Given initial values $\boldsymbol{x}^{\text{old}}$, the first factor of the CBA can be sampled by sampling from $\pi(\boldsymbol{x}_{B_1}|\boldsymbol{x}_{-B_1} = \boldsymbol{x}_{-B_1}^{\text{old}})$ and only update the values for $\boldsymbol{x}_{b_1} = \boldsymbol{x}_{b_1}^{\text{new}}$. Next a sample for $\boldsymbol{x}_{b_2}$, $b_2 = \{d_2, d_3\}$, is obtained from $\pi(\boldsymbol{x}_{b_2}|\boldsymbol{x}_{b_1} = \boldsymbol{x}_{b_1}^{\text{new}}, \boldsymbol{x}_{B_{>2}} = \boldsymbol{x}_{B_{>2}}^{old})$ with $B_{>2} = \{d_4, d_7, d_8, d_9\}$.

If we are only able to sample from an approximation of the full conditional of blocks, the corresponding approximation for collapsed blocks can be used. This can be seen as two levels of approximations; first each factor in (8) is approximated, $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}) \approx \pi^*(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}; \boldsymbol{x}^{\text{old}}) = \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}} = \boldsymbol{x}^{\text{new}}, \boldsymbol{x}_{B_{>i}} = \boldsymbol{x}^{\text{old}})$, then the collapsed blocks are approximated, $\pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}, \boldsymbol{x}_{B_{>i}}) \approx \pi^{**}(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}, \boldsymbol{x}_{B_{>i}}; \boldsymbol{x}^{\text{old}})$. We refer to the corresponding simultaneous approximation $\pi^{**}(\boldsymbol{x}; \boldsymbol{x}^{\text{old}}) = \prod_{i=1}^{N} \pi^{**}(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{b_{<i}}, \boldsymbol{x}_{B_{>i}}; \boldsymbol{x}^{\text{old}})$ as the approximated collapsed blocks approximation (ACBA).

3 Collapsed- and overlapping blocks Gibbs samplers

The main aim of this paper is to construct a class of approximations, the CBAs, and to use these in the one-block M-H algorithm. Though, the same ideas used for CBAs, can also be used to improve mixing of block Gibbs samplers. In addition, insight of the opposite reverse acceptance rate used in Section 4 can be gained from the resulting Gibbs samplers.

As in Section 2, consider a spatial field, e.g. a lattice, divided into contiguous disjoint blocks. The realisation obtained by sampling each block from its full conditional often

depends strongly on the old values for the field, which causes slow mixing. If we sample from a distribution for each block that depends less on the old field, the mixing will improve. We now take the same approach as for the CBA: For each block the variables in $\boldsymbol{x}_{b_{>i}}$ that depend most on $\boldsymbol{x}_{b_i}$, i.e. those closest in space, are included in a buffer that is collapsed from the full conditional distribution. The connection between the old field and the distribution sampled from is now relaxed. This Gibbs sampler we refer to as a *collapsed blocks Gibbs sampler* (CBGS).

Using the same notation as in Section 2, a conventional block Gibbs sampler draws from the full conditional distribution $\boldsymbol{x}_{b_i}^{\text{new}} \sim \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{-b_i})$, while the CBGS uses the collapsed conditional distribution, $\boldsymbol{x}_{b_i}^{\text{new}} \sim \pi(\boldsymbol{x}_{b_i}|\boldsymbol{x}_{-B_i}) = \int \pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{-B_i}) d\boldsymbol{x}_{\beta_i}$. Sampling can be performed by sampling from the full conditional, $(\boldsymbol{x}_{b_i}^{\text{new}}, \boldsymbol{x}_{\beta_i}^{\text{buffer}}) \sim \pi(\boldsymbol{x}_{b_i}, x_{\beta_i}|x_{-B_i})$ and only updating for $\boldsymbol{x}_{b_i}$. Figure 1, left, can also be seen as an illustration of the CBGS. The CBGS is within the framework of *MCMC based on conditional sampling* in Appendix 2 in Besag et al. (1995). The method can also be described as a combination of blocking (sampling more than one variable at the time) and collapsing (integrating out variables), where our knowledge about the dependence structure is used to set up blocks and buffers.

If also samples for the buffers are updated, we have an *overlapping blocks Gibbs sampler* (OBGS); a block Gibbs sampler with blocks that overlap such that buffers are sampled at least twice. See Figure 1, right. The OBGS can be set up as a special case of the general Gibbs sampler in Appendix 1 in Carter and Kohn (1996).

From an algorithmic point of view the difference between OBGS and CBGS is whether the buffer variables $\boldsymbol{x}_{\beta_i}$ are updated or not. While they are not used for the CBGS, also the buffer variables are updated for the OBGS. Further, when the stationary distribution is reached, each step of the OBGS produces a new sample from $\pi(\boldsymbol{x})$, while a full scan is needed for the CBGS. The two Gibbs samplers becomes identical if blocks and buffers are set up such that temporary buffer samples are never conditioned upon, e.g. a time series kind of blocking as in the $AR(1)$ process in Section 1.

In a Gibbs setting it will be hard to argue that one should use CBGS instead of the OBGS. Though, our main aim is to make an approximation, and in most cases the CBGS is then the only feasible option. It is straight forward to construct an approximation based on the OBGS that we can sample from. This approximation will in most cases be hard to evaluated, as

$$\pi^*(\boldsymbol{x}) = \int \dots \int \prod_{i=1}^{N_B} \pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{-B_i}) d\boldsymbol{x}_{\beta_1} \dots d\boldsymbol{x}_{\beta_{Nb}} \tag{10}$$

as in each factor $\pi(\boldsymbol{x}_{b_i}, \boldsymbol{x}_{\beta_i}|\boldsymbol{x}_{-B_i})$ we condition on $\boldsymbol{x}_{-B_i}$ which is a mixture of $\boldsymbol{x}^{\text{new}}$, $\boldsymbol{x}^{\text{old}}$ and temporary buffer variables that are integrated out.

If the full conditional distributions for blocks can be evaluated, but sampling is only possible from an approximation, Metropolised versions of CBGS and OBGS can be used.

4 One-block Metropolis-Hasting with CBA proposal and opposite reverse acceptance rate

We will now set up an one-block M-H algorithm sampling from (1) with some simple proposal for the hyper-parameters (e.g. random walk), and with a CBA as proposal for $\boldsymbol{x}$. We will not use Peskuns optimal acceptance probability. To motivate this, assume fixed hyper-parameters (suppressed when assumed fixed) and that we have a sample from $\pi(\boldsymbol{x}|\boldsymbol{y})$ as initial values. If the CBA or the approximation corresponding to a OBGS is used as proposal, proposed samples are from the target distribution. Even though all proposed samples are from the target distribution, a M-H algorithm with the optimal acceptance probability of Peskun (1973) does not generally give acceptance rate one. This because M-H algorithms are constructed such that a reversible Markov chain is produced. We want to avoid this "unnecessary" low acceptance rate (in many cases the acceptance rate is surprisingly low).

Using one scan of a OBGS as proposal acceptance rate one is achieved by a symmetric scan, e.g. $B_1 \to B_2 \to B_3 \to B_4 \to B_3 \to B_2 \to B_1$ for the blocks in Figure 1 (right). It can also be achieved by using a mixture of proposals as a proposal. Let q_0 be the proposal distribution for updating scheme $B_1 \to B_2 \to B_3 \to B_4$, and q_1 for updating scheme $B_4 \to B_3 \to B_2 \to B_1$. Randomly choose which proposal to use, each with probability 0.5; $P(q_0) = P(q_1) = 0.5$. This gives an overall proposal $q(\boldsymbol{x}^{'}|\boldsymbol{x}) = 0.5q_0(\boldsymbol{x}^{'}|\boldsymbol{x}) + 0.5q_1(\boldsymbol{x}^{'}|\boldsymbol{x})$. Both these proposals give acceptance probability one, but the computational cost is twice the original one. We consider the mixture of proposals, but instead of using Peskun's optimal acceptance probability formulae;

$$\alpha(\boldsymbol{x}'|\boldsymbol{x}) = \min\left\{1, \frac{\pi(\boldsymbol{x}')q(\boldsymbol{x}|\boldsymbol{x}')}{\pi(\boldsymbol{x})q(\boldsymbol{x}'|\boldsymbol{x})}\right\} \tag{11}$$

we use the *opposite reverse acceptance probability* (ORAP) (Tjelmeland and Hegstad (2001))

$$\alpha_{i,1-k}(\boldsymbol{x}'|\boldsymbol{x}) = \min\left\{1, \frac{\pi(\boldsymbol{x}')q_{1-k}(\boldsymbol{x}|\boldsymbol{x}')}{\pi(\boldsymbol{x})q_k(\boldsymbol{x}'|\boldsymbol{x})}\right\} \qquad k \in \{0,1\}. \tag{12}$$

It does not give optimal convergence as a function of M-H iteration, but if the overall proposal is computationally expensive to evaluate it may improve convergence as a function of computation time. Here, calculating (12) is only half the computational cost of calculating (11). It can be shown that for the OBGS $\alpha_{0,1} = \alpha_{1,0} = 1$. Acceptance probability one is not achieved for the CBA since the opposite scans are generally not symmetric. But they are nearly symmetric, and using the opposite reverse strategy gives an acceptance rate close to one.

An one-block M-H with CBA and ORAP is set up in algorithm 1. For proposing $\boldsymbol{x}$ we use CBAs $\pi_k^*(\boldsymbol{x}; \boldsymbol{x}^j, \boldsymbol{\theta}^{\text{new}})$, where $\boldsymbol{x}^j$ and $\boldsymbol{\theta}^{\text{new}}$ indicates which initial field and hyper-parameters the approximation is based upon. The index k gives the direction the blocks $\boldsymbol{x}_{bi}$ are sampled in; increasing is given by $k = 0$ and decreasing by $k = 1$. The ORAP used

Algorithm 1 One-block Metropolis-Hasting sampler with CBA and ORAP

- Given $\boldsymbol{x}^0$ and $\boldsymbol{\theta}^0$
- for $j = 0 : (niter - 1)$
  - Sample $\boldsymbol{\theta}^{\text{new}} \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^j)$
  - Sample $k \sim Unif(\{0, 1\})$
  - Sample $\boldsymbol{x}^{\text{new}} \sim \pi_k^*(\boldsymbol{x}; \boldsymbol{x}^j, \boldsymbol{\theta}^{\text{new}})$.
  - Calculate ORAP and accept / reject
  - if(accept)
    - $\boldsymbol{\theta}^{j+1} = \boldsymbol{\theta}^{\text{new}}$ and $\boldsymbol{x}^{j+1} = \boldsymbol{x}^{\text{new}}$
  - else
    - $\boldsymbol{\theta}^{j+1} = \boldsymbol{\theta}^j$ and $\boldsymbol{x}^{j+1} = \boldsymbol{x}^j$
- Return $(\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^{\text{niter}})$ and $(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^{\text{niter}})$

is

$$\alpha(\boldsymbol{\theta}^{\text{new}}, \boldsymbol{x}^{\text{new}}|\boldsymbol{\theta}^j, \boldsymbol{x}^j) = \min\left\{1, \frac{\pi(\boldsymbol{y}|\boldsymbol{x}^j)\pi(\boldsymbol{x}^j|\boldsymbol{\theta}^j)\pi(\boldsymbol{\theta}^j)q(\boldsymbol{\theta}^{\text{new}}|\boldsymbol{\theta}^j)\pi_{1-k}^*(\boldsymbol{x}^{\text{new}}; \boldsymbol{x}^j, \boldsymbol{\theta}^{\text{new}})}{\pi(\boldsymbol{y}|\boldsymbol{x}^{\text{new}})\pi(\boldsymbol{x}^{\text{new}}|\boldsymbol{\theta}^{\text{new}})\pi(\boldsymbol{\theta}^{\text{new}})q(\boldsymbol{\theta}^j|\boldsymbol{\theta}^{\text{new}})\pi_k^*(\boldsymbol{x}^j; \boldsymbol{x}^{\text{new}}, \boldsymbol{\theta}^j)}\right\}$$

Note that the target distribution $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ must be known up to a normalisation constant not dependent on $\boldsymbol{\theta}$. Hence the prior $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ must be known (and possible to evaluate) also as a function of $\boldsymbol{\theta}$. In Section 6 we find that $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is computable for important classes of GMRF priors.

5 Simulation experiments

Here we report some simulation studies done to explore OBGS, CBA and ACBA. In all these studies our starting point is a GMRF on a lattice of 20×20 pixels that is a proxy to a GRF with precision $\tau = 1.0$ and exponential correlation with range $r = 60$ as described in Rue and Tjelmeland (2002). Hyper-parameters are assumed known and fixed in all these studies.

5.1 Experiment 1, OBGS

We first explore the overlapping blocks Gibbs sampler (OBGS). We sample from the GMRF described above using a OBGS with four square blocks each of size (10×10) (non overlapping), (11×11), (12×12) and (15×15). For comparison we also ran an one-block sampler

which gives independent samples from the GMRF. All samplers had a sample from the target GMRF as initial values, and ran for 10000 iterations. The estimated autocorrelation functions (ACF) for pixel $(10, 10)$ are plotted in Figure 2. We observe that the ACF decreases faster with more overlap, and with blocks of size (15×15) (i.e. 10 pixels overlap) the ACF is close to that of the one-block sampler, i.e. its samples are close to independent.

5.2 Experiment 2, CBA with Peskuns acceptance rate and ORAP

In this experiment we sample from our GMRF using an one-block M-H algorithm with CBA proposals. We explore the CBA proposals corresponding to the OBGS considered in Section 5.1. I.e. the blocks actually sampled from, $\boldsymbol{x}_{B_i}$, $B_i = \{b_i \cap \beta_i\}$, are set up to be squares each of size (10×10) (non overlapping), (11×11), (12×12) and (15×15) (width of buffer 10). This corresponds to the situations illustrated in Figure 1. We also explore two different proposal - acceptance probability configurations: 1) we use CBA as proposal with Peskuns acceptance probability. And 2) we use a mixture of two CBA (with opposite directions) as proposal and the ORAP. Approach 2 corresponds to Algorithm 1 with fixed/known hyper-parameters.

In Figure 3 the estimated ACFs are plotted for these simulation studies as well as the acceptance rates for an extended study with all possible block sizes. From the ACF plots we find that increasing buffers decreases the autocorrelation. Comparing the ACFs of the two proposal-acceptance rate approaches indicates that approach 2 gives lower autocorrelation for small buffer sizes while no difference can be seen for larger buffers. From the acceptance rate plot we see that this is due to very low acceptance rate for approach 1 for small buffers. When the buffers increase, the approximation becomes better, and for both approaches the proposed samples depend less on $\boldsymbol{x}^{old}$. Hence, the direction of the CBA becomes invisible, and the benefit of a symmetric proposal (as in approach 2) disappears.

5.3 Experiment 3, ACBA with Peskuns acceptance rate and ORAP

In our last simulation experiment we explore using the ACBA as proposal in a hidden GMRF model setting. Based on a sample $\boldsymbol{x}$ from our GMRF we simulate data for each pixel from a Poisson distribution; $y_i \sim Po(0.5 \exp(x_i))$. We want to sample from the posterior of the latent field, i.e. from $\pi(\boldsymbol{x}|\boldsymbol{y})$, which takes the form (1) (hyper-parameters are assumed fixed). $\pi(\boldsymbol{x}|\boldsymbol{y})$ is non-Gaussian and to sample we use an one-block M-H algorithm with ACBA as proposal for $\boldsymbol{x}$. As approximated distribution for each block we use the classes of approximations to non-Gaussian latent (or hidden) GMRFs introduced in Rue et al. (2004), see Appendix A.1. for technical details. As in Simulation 2 two proposal - acceptance rate configurations were explored: 1) ACBA as proposal with Peskuns acceptance rate. And 2) a mixture of two ACBA (with opposite directions) as proposal and the ORAP. We use the same blocks and buffers as in Simulation 2, and all simulations ran for 10000 iterations.

The estimated ACFs of the samples from the one-block M-H sampler with ACBA using the $A1$ approximation are plotted in Figure 4. For comparison results from an independence

sampler with an one-block $A1$-approximation as proposal are also included. From the plot we find that approach 1 without overlap gives high autocorrelation. The autocorrelation rapidly decreases using approach 2 and/ or overlapping, and for approach 2 we can not distinguish between the one-block approximation and the ACBA with $\boldsymbol{x}_{B_i}$ of dimension (11×11). For approach 1 we can not distinguish between the one block approximation and the ACBA with $\boldsymbol{x}_{B_i}$ of dimension (12×12). Hence we need smaller buffers in this study that in simulation study 2. There are two reasons for this: First, each data point y_i contain much information about its corresponding latent field variable x_i, and the other latent variables are less important. Second, the M-H sampler with a one-block approximation as proposal does not give independent samples (it is an independence sampler with acceptance rate 0.61), and we therefore compare with something less ambitious than in simulation experiment 1 and 2.

6 Functional magnetic resonance imaging

In this Section we demonstrate the applicability of CBAs in an one-block M-H algorithm setting through an example. The model is so large that $\boldsymbol{x}^{\text{old}}$ independent proposals for the field is computationally impossible. We use GMRFLib for fast sampling of GMRFs, (Rue and Follestad (2002)).

Here we analyse functional magnetic resonance imaging (fMRI) data from a visual stimulation experiment is analysed. These data are previously studied in Göss et al. (2000). The stimulus was a 8 Hz flickering checkerboard, and the experiment lasted for 210 seconds with four periods (of 30 seconds) rest and three periods stimulus. A crossection of the brain $(128 \times 128$ pixels) was observed every third second, hence a time series of 70 images. Images of size 75×67 is enough to cover the brain and the analyses is based on these images. Functional magnetic resonance utilise the different magnetic properties of oxygenated and disoxygenated blood and is useful for observing blood oxygenation level dependent (BOLD) effects. External stimulation is related to the BOLD effect and the aim of the experiment was to detect areas activated by the visual stimulation.

Traditionally temporal and, if at all, spatial effects have been analysed separately. In Göss et al. (2001) Bayesian hierarchic parametric and semi-parametric spatial and spatio-temporal models for this problem are introduced. We introduce a space-time GMRF model similar to the one in Göss et al. (2001). To make robust inference an one-block M-H algorithm with CBA and ORAP is used.

We model the observations of pixel i at time t as $y_{it} = b_i + z_t a_{it} + \epsilon_{it}$, for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, where $\boldsymbol{b}$ is the baseline image (of size N) and a_{it} is the activation effect of pixel i at time t (of size $N \times T$). A transformed stimulus, $z_t > 0$, $t = 1, 2, \ldots, T$, of the original stimulus is used. A common choice is to use a temporal shift of the original stimulus $\boldsymbol{x}$ by a time-delay d and a convolution h, $z_t = \sum_{s=0}^{t-d} h(s, \phi) x_{t-d-s}$ with h either Poisson or gamma density function. The parameters involved here, (d, ϕ), are estimated by least square from similar data. The measurement errors ϵ_{it} are assumed independent identically Gaussian distributed with common precision τ_{Data}. The tempo-spatial

modelling is done through the priors of $\boldsymbol{a}$ and $\boldsymbol{b}$, which both are given intrinsic GMRF priors; $\pi(\boldsymbol{b}) \propto \exp(-\frac{1}{2}\tau_B \sum_{i \overset{s}{\sim} j}(b_i - b_j)^2)$ and $\pi(\boldsymbol{a}) \propto \exp(-\frac{1}{2}\tau_A \sum_{t=1}^{T} \sum_{i \overset{s}{\sim} j}(a_{it} - a_{jt})^2) \cdot \exp(-\frac{1}{2}\tau_T \sum_{i=1}^{N} \sum_{t \overset{t}{\sim} r}(a_{it} - a_{ir})^2)$. Each non-boarder pixel has four neighbours in space (both in $\boldsymbol{a}$ and $\boldsymbol{b}$) and two in time (for $\boldsymbol{a}$ only). The priors for τ_A, τ_B and τ_T were all set to independent Gamma distributed with expected value 5.0 and variance 100.

The posterior distribution of the hyper-parameters $\boldsymbol{\theta} = (\tau_A, \tau_B, \tau_T)$, the baseline image $\boldsymbol{b}$ and the activation effects $\boldsymbol{a}$, $\pi(\boldsymbol{b}, \boldsymbol{a}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{a}|\boldsymbol{\theta})\pi(\boldsymbol{b}|\boldsymbol{\theta}) \prod_i \pi(y_i|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta})$ is our distribution of interest. In Appendix A.2 $\pi(\boldsymbol{b}, \boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{y})$ is found to be multivariate Gaussian with a conditional dependence structure as illustrated with the graph in Figure 5. Visual stimulation effects are only present in the mid-section of the brain, and this part is analysed. This gives a dataset of 75×21 pixels $\times 70$ time steps, or 110250 data points, and $\{\boldsymbol{a}, \boldsymbol{b}\}$ has dimension 111825. We are not able to sample exact from a GMRF of this size. Further, the precision for the data was estimated from the part of the image not used $\tau_{Data} = 0.003$.

We use the one-block M-H algorithm from Section 4 to make inference. A strategy for calculating the normalisation constant of $\pi(\boldsymbol{b}, \boldsymbol{a}|\boldsymbol{\theta})$ is found in In Appendix A.3. Because of the dependence structure (see Figure 5) the CBA is set up such that the baseline image $\boldsymbol{b}$ is in every buffer. All but the last block consist of three activation images images with the two succeeding $\boldsymbol{a}$ images and $\boldsymbol{b}$ as buffers. The last block consists of $\boldsymbol{a}$ and the last five $\boldsymbol{a}$ images, see Figure 5. Further, hyper-parameters are proposed independently of each other; $\tau^{\text{new}} \sim Unif([\frac{1}{f}\tau^{\text{old}}, f\tau^{\text{old}}])$. The one-block M-H algorithm ran for 20000 iterations. Trace plots with cumulative mean after a burn-in of 2000 iterations are found in Figure 6.

We observe quite low spatial dependence ($\overline{\tau_A} = 0.000070$ and $\overline{\tau_B} = 0.000009$), though reasonable values considering the values of the data. The precision in time is much higher ($\overline{\tau_T} = 0.17$). The mixing of the smoothing parameters of $\boldsymbol{a}$ is quite slow, but the cumulated means have stabilised. Figure 7 contains images of the mean estimate of the baseline image and the activation for two stimulus time steps ($t = 18$ and $t = 38$) and one rest time step ($t = 28$). We find activation areas in the upper part of the brain, which agrees with previous studies of the same data. In Figure 8 data and mean estimates are plotted for three pixels; one with high, one with moderate and one without stimulus activation. We observe that the estimates are smooth and appears less noisy, but not smoothed too much, -the stimulus activation is well kept.

7 Closing Remarks

In this paper we have introduced a new method for constructing approximations to high dimensional distributions with known dependency structure that can be sampled from and evaluated. The CBAs are based on full conditional distributions for blocks of variables, and require that sampling from and evaluation of (an approximation to) these distributions are feasible. Our knowledge about the dependence structure is utilised when setting up blocks and buffers. The method is applied on latent GMRF models. Due to recent advantages in computing the normalising constant for discrete Markov random fields (MRFs), Bartolucci

and Besag (2002) and Reeves and Pettitt (2004), it is possible to use our CBA for discrete MRFs as well. However, each block cannot be larger than about 20×20 in the binary case, because of computational requirements.

How large buffers that are needed depends on the variance of and dependency within the field. Smaller variance and stronger dependency between variables imply that more buffer variables are required to achieve the same accuracy. In many applications the variance and within field dependency is determined by hyper-parameters which also are to be estimated. Hence, it depends on the dataset under study, and, if the CBA is used as part of a proposal in an one-block M-H, the proposed value of the hyper-parameters. We have tuned buffer sizes from preliminary runs.

The computational complexity of CBAs depends on how the increased number of variables influence the dependency. Consider a lattice covering a study area: If the number of variables increases because it should cover a larger area and the dependency between neighbouring variables is fixed, the complexity is linear as only the number of blocks increases. On the other hand, if the field gets denser; e.g. more locations in a fixed area, the dependency between neighbouring variables increases. The blocks and buffers needs to cover the same physical area as before to give the same accuracy, and the computational complexity equals the cost of the sampling method.

Through examples the CBA has proved to be a powerful method for constructing proposals for the latent field when evaluating spatial and tempo-spatial latent models using one-block M-H algorithms. The CBAs enable us to apply one-block updating schemes for problems else to large.

Acknowledgements

For the first author the research was founded by grants from the Research Council of Norway (project 133695/432) and parts of the work was done while she visited Department of Statistics at Trinity College Dublin as a part of the IITAC-project (The Institute for Information Technology and Advanced Computation). We are very thankful to Andrea Hennerfeind and Leonhard Held at the Department of Statistics, Ludwig-Maximilians-University Munich, for providing data and information about the fMRI experiment.

A Appendix

A.1 Evaluation of ACBA

In Rue et al. (2004) latent GMRF models with mutually independent likelihoods are considered, and approximations to $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ also known as a hidden GMRF (HGMRF), are constructed. We will use these as approximations for $\pi(\boldsymbol{x}_{Bi}|\boldsymbol{x}_{-Bi},\boldsymbol{y},\boldsymbol{\theta})$, as also the conditional distribution for blocks are HGMRFs. For simplicity we suppress $\boldsymbol{\theta}$ and $\boldsymbol{y}$ in the following. The approximations in Rue et al. (2004) are done in three stages: The first stage is to find a Gaussian approximation in the mode of $\pi(\boldsymbol{x})$. Denote this approximation

$\pi^{A1}(\boldsymbol{x})$. As a first improvement, π^{A2}, non-Gaussian corrections for the likelihood term for the corresponding data point (correct for y_i when x_i is sampled) are done. The third approximation, π^{A3}, also correct for non-Gaussian likelihood terms from other locations through sampling.

We denote approximation Aj to $\pi(\boldsymbol{x}_{B_i}|\boldsymbol{x}_{-B_i})$ $\pi^{Aj}(\boldsymbol{x};\boldsymbol{x}_{-B_i})$, and suppress the method number j in expressions valid for all methods. The ACBA can be written as

$$\pi^A(\boldsymbol{x};\boldsymbol{x}') = \prod_{i=1}^{N} \pi^A(\boldsymbol{x}_{bi};\boldsymbol{x}_{Bi-},\boldsymbol{x}'_{Bi+}) \tag{13}$$

Note that $\pi^A(\boldsymbol{x}_{Bi};\boldsymbol{x}_{Bi-},\boldsymbol{x}'_{Bi+})$ is the marginal distribution of $\boldsymbol{x}_{bi}$ of the approximation; $\pi^A(\boldsymbol{x}_{bi};\boldsymbol{x}_{Bi-},\boldsymbol{x}'_{Bi+}) = \int \pi^A(\boldsymbol{x}_{bi},x_{\beta i};\boldsymbol{x}_{Bi-},\boldsymbol{x}'_{Bi+})d\boldsymbol{x}_{\beta i}$. This proposal requires evaluation of the transition density (13):

Evaluation of $\pi^{A1}(\boldsymbol{x}|\boldsymbol{x}')$

The first approximation, $\pi^{A1}(\boldsymbol{x}_B;\boldsymbol{x}_{-B})$, is made by finding the mode of $\pi(\boldsymbol{x}_B|\boldsymbol{x}_{-B})$ and a Gaussian approximation in the mode. Hence it is Gaussian, and evaluation can be done as described in Section 2.

Evaluation of $\pi^{A2}(\boldsymbol{x}|\boldsymbol{x}')$ and $\pi^{A3}(\boldsymbol{x}|\boldsymbol{x}')$

The approximations $A2$ and $A3$ are refined versions of $A1$ and the refinements are done sequential. It can be written as $\pi^{A2}(\boldsymbol{x}) = \pi^{A2}(x_k)\pi^{A2}(x_{k-1}|x_k)\dots\pi^{A2}(x_1|x_2,x_3,\dots,x_k)$ where $\pi^{A2}(x_k)$ is an approximation to the marginal distribution of x_k, $\pi^{A2}(x_{k-1}|x_k)$ an approximation to x_{k-1} conditioned on the x_k sampled in the previous step and so forth (similar for $A3$). Since the approximation for $\pi(x_i|x_{i+1},\dots,x_k)$ is only done for the sampled $(x_{i+1},x_{i+2},\dots,x_k)$ evaluations of $\pi^{A2}(x)$ and $\pi^{A3}(x)$ are not straight forward and generally not easily obtained.

If the sampling of $\boldsymbol{x}$ is stopped when $p-1$ steps remain the density of the obtained sample is given as $\pi^{A2}(x_k)\pi^{A2}(x_{k-1}|x_k)\dots\pi^{A2}(x_p|x_{p+1}\dots,x_k)$ This is the marginal density for $(x_p,\dots,x_k)$ of $\pi^{A2}(\boldsymbol{x})$. If the ordering within each block is done such that the buffer elements have the lowest indexes, $\boldsymbol{x}_B = (\boldsymbol{x}_\beta,\boldsymbol{x}_b)$, we can both sample from and evaluate $\pi^{A2}(\boldsymbol{x}_{bi};\boldsymbol{x}_{-Bi})$ simply by stopping the sampling process when $\boldsymbol{x}_{bi}$ is sampled.

A.2 Calculating $\pi(\boldsymbol{b},\boldsymbol{a}|\tau_{Data},\tau_A,\tau_B,\tau_T,\boldsymbol{y})$

The distribution is given by the likelihood $\pi(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a},,\tau_{Data})$ and the priors of $\boldsymbol{a}$ and $\boldsymbol{b}$; $\pi(\boldsymbol{b},\boldsymbol{a}|\tau_{Data},\tau_A,\tau_B,\tau_T,\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{a},\tau_{Data})\pi(\boldsymbol{b}|\tau_B)\pi(\boldsymbol{a}|\tau_A,\tau_T))$ The likelihood is multivariate Gaussian, and can be written as: $\boldsymbol{y}|\boldsymbol{a},\boldsymbol{b},\tau_{Data} \sim N(\boldsymbol{1}_T\otimes\boldsymbol{b}+(\text{diag}(\boldsymbol{z})\otimes I_N)\boldsymbol{a},\tau_{Data}I_{NT})$, where $\boldsymbol{1}_T$ is a column vector of size T containing ones, and I_m is an identity matrix of size $m\times m$. The priors for $\boldsymbol{a}$ and $\boldsymbol{b}$ are intrinsic Gaussian as given in Section 4.1, and we denote the corresponding precision matrices Q_a and Q_b. Hence, also $\pi(\boldsymbol{b},\boldsymbol{a}|\tau_{Data},\tau_A,\tau_B,\tau_T,\boldsymbol{y})$ is a

GMRF, and can be written as $\pi(\boldsymbol{b}, \boldsymbol{a}|\tau_{Data}, \tau_A, \tau_B, \tau_T, \boldsymbol{y}) \propto \exp(\frac{1}{2}[\boldsymbol{b}, \boldsymbol{a}]Q[\boldsymbol{b}, \boldsymbol{a}]^T + c^T[\boldsymbol{b}, \boldsymbol{a}]^T)$ with

$$Q = \begin{bmatrix} T\tau_{Data}I_N + Q_b & \tau_{Data}(\boldsymbol{z}^T \otimes I_N) \\ \tau_{Data}(I_N \otimes \boldsymbol{z}) & \tau_{Data}(\text{diag}(\boldsymbol{z}^2) \otimes I_N) + Q_a \end{bmatrix}$$

where $\text{diag}(\boldsymbol{z}^2)$ is a diagonal matrix with elements z_i^2. And

$$\boldsymbol{c}^T = \left[\sum_{t=1}^{T} y_{1t}, \sum_{t=1}^{T} y_{2t}, \ldots, \sum_{t=1}^{T} y_{Nt}, y_{11}z_1, y_{21}z_1, \ldots, y_{N1}z_1, y_{12}z_2, \ldots y_{NT}z_T\right]$$

A.3 The normalisation constant for the prior

The precision matrix for the space-time prior for $\boldsymbol{a}$ in Section 6 is non-positive definite, and its determinant is 0. We still need to know the normalisation constant as a function of τ_A and τ_T. A fruitful approach is to define the determinant det* of a non-negative matrix as the product of its non-zero eigenvalues, $\det^*(Q) = \prod_{i=1}^{m} \lambda_i$, where λ_i, $i = 1, \ldots, m$ is the non-zero eigenvalues of Q. If Q is positive definite is $\det^*(Q) = \det(Q)$. For definitions and proofs of linear algebra results used in this appendix, see Strang (1987) and Harville (1997). First we notice that the precision matrix can be written as $Q = \tau_S Q_S + \tau_T Q_T$, where Q_T is the precision matrix for the time dependence with $\tau_T = 1$, and Q_S for the spatial dependence with $\tau_A = 1$. Furthermore, Q_T and Q_S can be written as Kronecker products. Let R_T be the precision matrix (with $\tau_T = 1$) for one region, and R_S the precision matrix for one time-step (with $\tau_A = 1$); $Q_T = R_T \otimes I_N$ and $Q_S = I_T \otimes R_S$, where I_N and I_T are identity matrices of dimension $N \times N$ and $T \times T$. The spectral theorem gives that symmetric real matrices can be decomposed as $Q_S = V_S \Lambda_S V_S$ and $Q_T = V_T \Lambda_T V_T$ with orthonormal eigenvectors of $Q_\cdot$ in $V_\cdot$ and eigenvalues in $\Lambda_\cdot$. Two diagonalisable matrices A and B share eigenvector matrix V if and only if $AB = BA$. We have that $Q_S Q_T = (I_T \otimes R_S)(R_T \otimes I_N) = (I_T R_T) \otimes (R_S I_N) = R_T \otimes R_S$ and $Q_T Q_S = (R_T \otimes I_N)(I_T \otimes R_S) = (R_T I_T) \otimes (I_N R_S) = R_T \otimes R_S$, hence Q_T and Q_S share eigenvector matrix V, and $Q_S + Q_T = V\Lambda_S V^T + V\Lambda_T V^T = V(\Lambda_S + \Lambda_T)V^T$. The eigenvalues and -vectors of the factors in a Kronecker product gives eigenvalues and -vectors of the product: If A has eigenvalues $(\lambda_{A1}, \lambda_{A2}, \ldots, \lambda_{AN})$ and eigenvectors $(e_{A1}, e_{A2}, \ldots, e_{AN})$, and B has eigenvalues $(\lambda_{B1}, \lambda_{B2}, \ldots, \lambda_{BT})$ and eigenvectors $(e_{B1}, e_{B2}, \ldots, e_{BT})$, $A \otimes B$ has eigenvalues and vectors given by $\lambda_{Ai}\lambda_{Bj}$ and $e_{Ai} \otimes e_{Bj}$ $\forall i \in \{1, 2, \ldots, N\}$ and $j \in \{1, 2, \ldots, T\}$. Since the identity matrix has eigenvalues 1 and eigenvectors equal the standard basis we see that; $\text{diag}(\Lambda_T) = (\lambda_{T1}, \lambda_{T1}, \ldots, \lambda_{T1}, \lambda_{T2}, \ldots, \lambda_{T2}, \ldots, \lambda_{TT}, \ldots, \lambda_{TT})$ and $\text{diag}(\Lambda_S) = (\lambda_{S1}, \lambda_{S2}, \ldots, \lambda_{SN}, \lambda_{S1}, \ldots, \lambda_{SN}, \ldots, \lambda_{S1}, \ldots, \lambda_{SN})$. Further, $\det(Q_S + Q_T) = \det(\Lambda_S + \Lambda_T) = \prod_{i=1}^{N}\prod_{j=1}^{T}(\lambda_{Si} + \lambda_{Tj})$ or for us $\det^*(Q_S + Q_T) = \det(\Lambda_S + \Lambda_T) = \prod_{i=1}^{N}\prod_{j=1}^{T} f(\lambda_{Si} + \lambda_{Tj})$, where $f(x) = x$ for $x > 0$ and $f(x) = 1$ for $x = 0$. We observe that $\det^*(\kappa Q_S + \tau_T Q_T) = \prod_{i=1}^{N}\prod_{j=1}^{T} f(\kappa\lambda_{Si} + \tau_T\lambda_{Tj})$ and hence can be calculated from the eigenvalues of R_T and R_S.

References

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

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–66.

Carter, C. K. and Kohn, R. (1996). Markov Chain Monte Carlo in conditional Gaussian state space models. *Biometrika*, 83(3):589–601.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2003). Robust MCMC for spatial GLMM's. preprints in mathematical sciences. Technical report, Department of Mathematical Sciences, Lund University.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C*, 47(3):299–350.

Gamerman, D., Moreira, A. R. B., and Rue, H. (2003). Space-varying regression models: Specifications and simulations. *Computational Statistics and Data Analysis*, 42(3):513–533.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Göss, C., Auer, D., and Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics*, 57:544–562.

Göss, C., Auer, D. P., and Fahrmeir, L. (2000). Dynamic models in fMRI. *Magnetic Resonance in Medicine*, 43:72–81.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer Verlag New York.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Liu, J. S. (1994). The collapsed Gibbs sampler with application to a gene regulation problem. *Journal of the American Statistical Association*, 89:958–966.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with application to the comparison of estimators and augmentation schemes. *Biometrika*, 81:27–40.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics, 7*, pages 307–326. Oxford Univ. Press, New York.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612.

Pitt, M. K. and Shephard, N. (1999). Analytic convergence rates and parameterization issues for the Gibbs sampler applied to state space models. *Journal of Time Series Analysis*, 20(1):63–85.

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

Rue, H. and Follestad, T. (2002). GMRFLib: A C-library for fast and exact simulation of Gaussian Markov random fields. Statistics Report No. 1, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

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

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 66(4):877–892.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Strang, G. (1987). *Linear algebra and its applications*. Harcourt Brace & Company, 3rd edition.

Tjelmeland, H. and Hegstad, B. K. (2001). Mode jumping proposal in MCMC. *Scandinavian Journal of Statistics*, 28:205–223.

Wilkinson, D. J. (2003). Discussion to "Non-centered parameterizations for hierarchical models and data augmentation" by O. Papaspiliopoulos, G. O. Roberts and M. Sköld. In *Bayesian Statistics, 7*, pages 323–324. Oxford Univ. Press, New York.

Figures

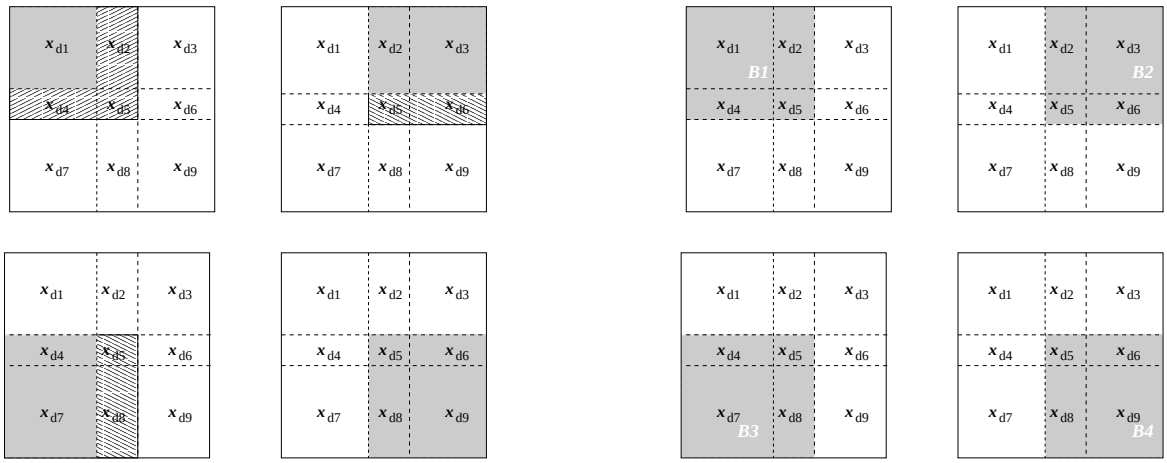


Figure 1: *Left: Illustration of the blocks of CBA and CBGS. In each step the gray area is sampled from its conditional distribution with the hatched area integrated out. Right: Illustration of the blocks of an OBGS. In each step the gray area is sampled from its full conditional distribution.*

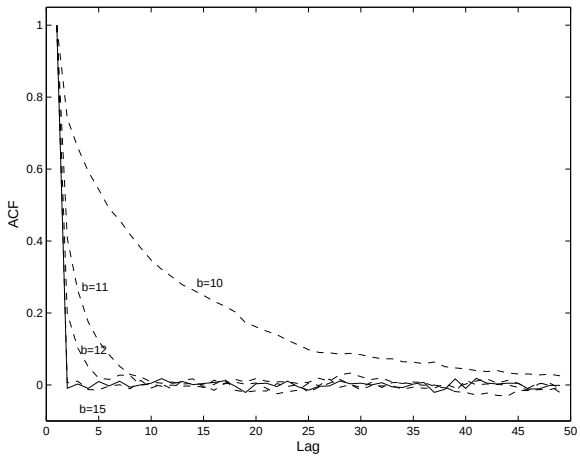


Figure 2: *Simulation experiment 1. Dashed lines; autocorrelation function (ACF) for OBGS with four blocks each of size $b \times b$. Solid line; ACF for independent samples.*

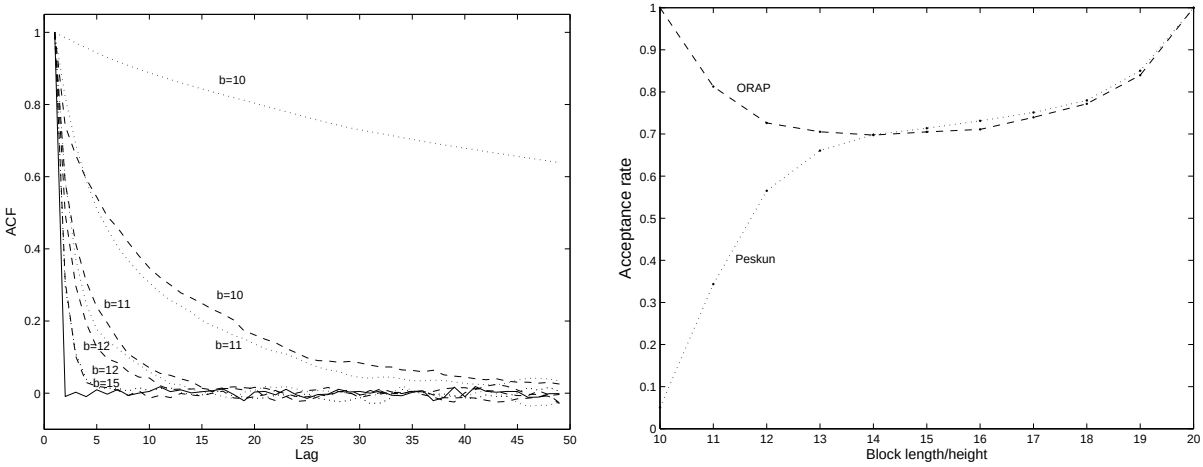


Figure 3: *Simulation experiment 2. Left: ACF for samples from one-block M-H with CBA proposal with four blocks each of size* $(b \times b)$*. Dotted lines; CBA and Peskun. Dashed lines; mixture of CBAs and ORAP. Solid line; ACF for independent samples. Right: Acceptance-rates for M-H samplers with CBAs proposals as a function of block sizes. Dotted line; CBA and Peskun. Dashed line; mixture of CBAs and ORAP.*

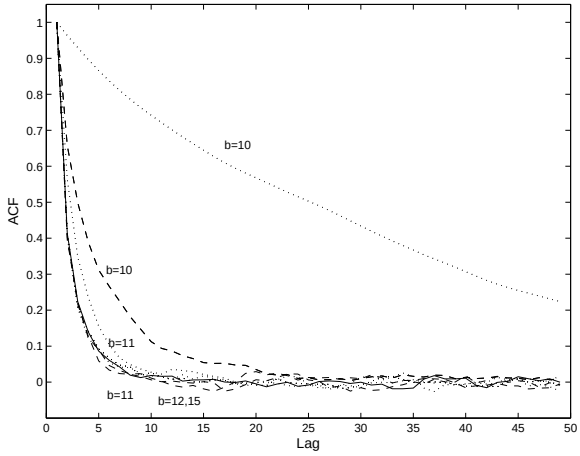


Figure 4: *Simulation experiment 3. ACFs for samples from one-block M-H with ACBA proposal with four blocks each of size* $(b \times b)$ *. Dotted lines; CBA and Peskun. Dashed lines; mixture of CBAs and ORAP. Solid line; ACF for samples from an independence sampler with corresponding one-block approximation.*

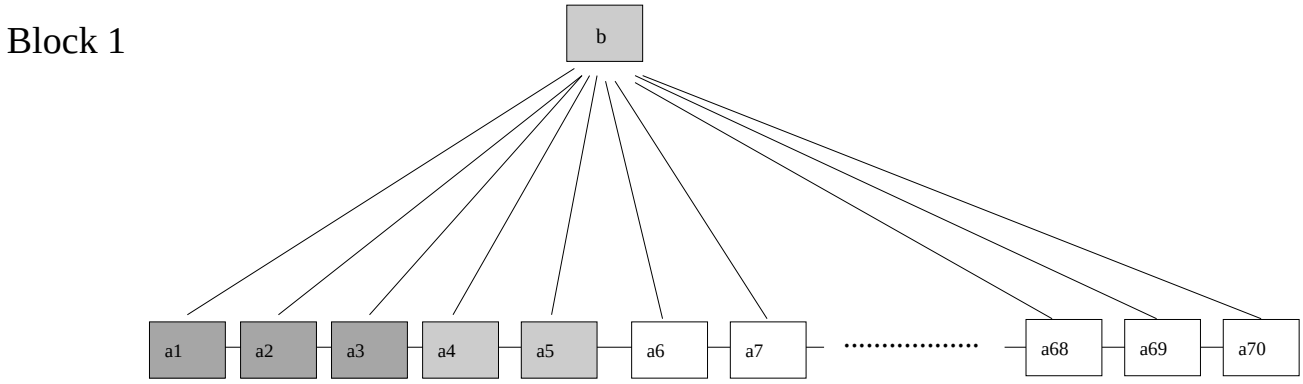


Figure 5: *fMRI example: The dependence structure of* $\pi(\boldsymbol{a}, \boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta})$ *and the first block of the CBA. The darker gray nodes* $b_1 = \{a1, a2, a3\}$ *are sampled from their conditional distribution with the brighter gray nodes,* $\beta_1 = \{b, a4, a5\}$ *integrated out.*

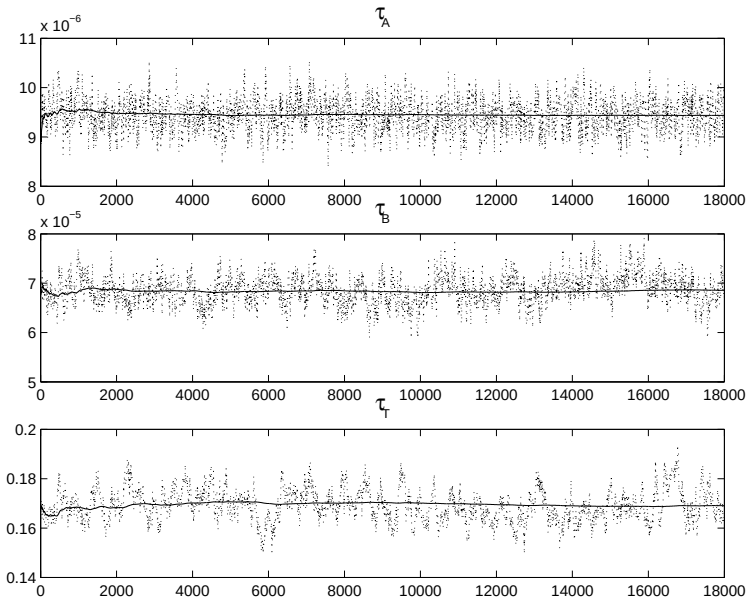


Figure 6: *fMRI example: Trace plots (doted lines) and cumulative means (solid lines) for* τ_A, τ_B *and* τ_T *after a burn-in of* 3000 *samples.*

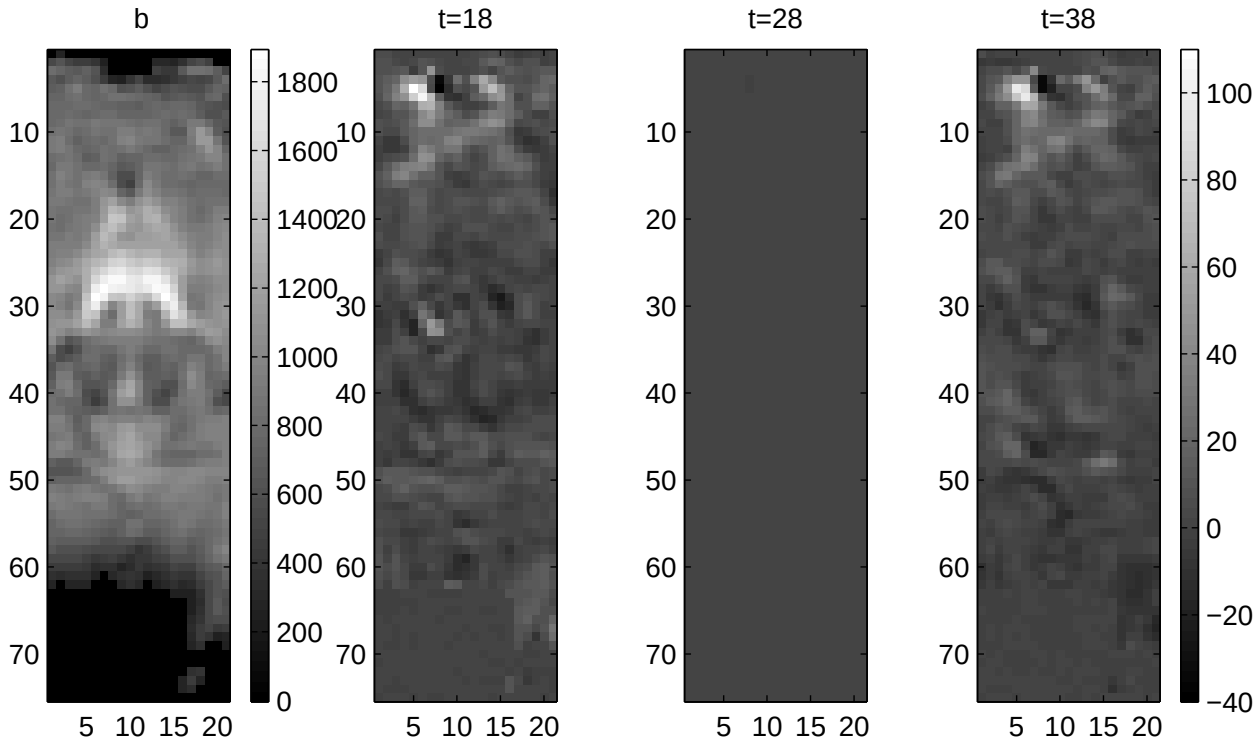


Figure 7: *fMRI example: The posterior mean for the baseline image* $\boldsymbol{b}$ *and the activation effect* $z_t \boldsymbol{a}_{\cdot t}$ *for time step* $t = 18$, $t = 28$ *and* $t = 38$

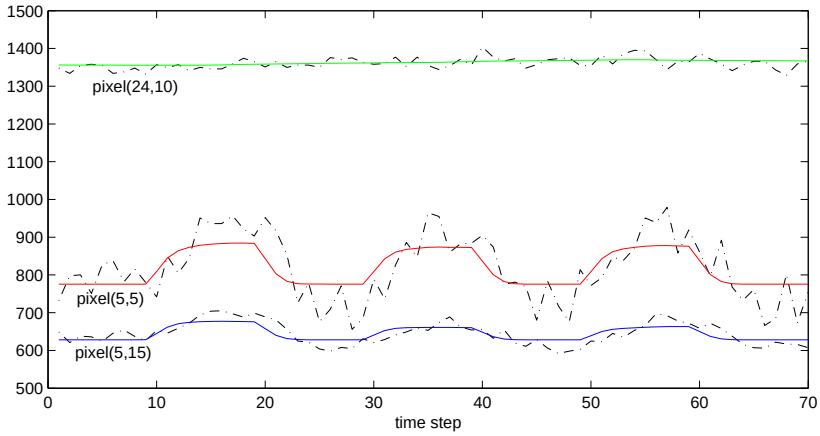


Figure 8: *fMRI example: The observed values and the posterior mean for pixels* $(24, 10)$, $(5, 5)$ *and* $(5, 15)$.