NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Overlapping block proposals for latent Gaussian Markov random fields

by

Ingelin Steinsland and Håvard Rue

PREPRINT
STATISTICS NO. 8/2003

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2003/S8-2003.ps
Ingelin Steinsland has homepage: http://www.math.ntnu.no/∼ingelins
E-mail: ingelins@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and
Technology, N-7491 Trondheim, Norway.

Overlapping block proposals for latent Gaussian Markov random fields

Ingelin Steinsland & Håvard Rue

Norwegian University of Science and Technology

Abstract

In this report we construct a full dimensional proposal distribution for the posterior of latent Gaussian Markov random fields $\pi(x|y,\theta)$, where x denotes the latent field which is of dimension n, y data and θ hyper-parameters. We can both sample from and evaluate these proposal without working directly with an n-dimensional distribution. The key idea in the construction of the proposals is to combine samples from overlapping blocks of the latent field. Each block is sampled from its conditional distribution or an approximation to its conditional distribution. The overlapping block proposals for x are used together with proposals for the hyper-parameters θ and an opposite reverse acceptance probability in one-block updating scheme Metropolis-Hastings algorithms.

Through examples the method prove to work well both when each block is sampled exact and when an approximation is necessary. Overlapping block proposals are successfully applied for a latent GMRF problem of dimension 100000. For some of the problems hyper-parameters with a Gaussian prior are also included in the overlapping blocking scheme.

1 Introduction

Markov chain Monte Carlo (MCMC) techniques are a general and powerful tool for making inference from analytically intractable statistical models, see e.g. Robert and Casella (1999), Gilks et al. (1996) and Liu (2001).

Traditionally single site updating schemes (updating one variable in each iteration) have been the most commonly used methods. In recent years it has been discovered that for highly structured problems these sampling schemes give unsatisfactory slow mixing; the Markov chain explore our target distribution too slowly. Updating several variables simultaneous, known as blocking or grouping, is a remedy for improving mixing, see Liu (1994), Liu et al. (1994), Knorr-Held and Rue (2002) and Gamerman et al. (2003). Knorr-Held and Rue (2002) empirically demonstrates that for some problems updating all variables in one block gives by far the best mixing. We will from now on refer to an updating scheme where all variables are updated simultaneously as an one-block updating scheme. Applying an one-block updating scheme for an n-dimensional model requires both sampling from and evaluation of an n-dimensional distribution. In most cases finding an appropriate distribution to sample from is hard, and both the sampling and the needed evaluations are computationally expensive.

In this report we focus on a much used class of spatial models; spatial latent Gaussian Markov random field models (presented below). For these models an one-block updating scheme not involving direct sampling from an n-dimensional distribution is constructed and tested on several problems.

1.1 Spatial latent GMRF models

In many situations data are collected with a spatial index and is indirect observations of the phenomena of interest. In this report we consider problems where the phenomena of interest is modelled as a latent field x, possible given some hyper-parameters θ. Further the observations y are modelled conditioned on x and as mutually independent; $\pi(y|x) = \prod_{i=1}^{M} \pi(y_i|x)$. The likelihood may has its own hyper-parameter, but they are suppressed here. The dimension of x and y is not necessarily equal, but they share the spatial reference system. The latent field is assumed to have a spatial dependence structure specified through a prior $\pi(x|\theta)$, a popular choice is multivariate Gaussian priors. We will restrict our examples to a special version of these; Gaussian Markov random fields (GMRFs). A GMRF is a multivariate Gaussian distribution, also known as a Gaussian random field (GRF), with a neighbourhood structure and a corresponding Markov property. Let $i \sim j$ denote that element i and j are neighbours and denote i's neighbourhood $\mathcal{N}_i = \{j : j \sim i\}$. The Markov property gives that x_i conditioned on its neighbourhood is independent of all other variables:

$$\pi(x_i|x_{-i}) = \pi(x_i|x_{\mathcal{N}_i}) \qquad \forall i$$

where $-i$ denotes the complement of the whole set $\{1,2,\ldots,n\}$ and i. The non-zero structure of the precision matrix Q (the inverse of the covariance matrix) reflects x's neighbourhood structure. Let q_{ij} denote element (i, j) of Q, then $q_{ij} \neq 0$ if and only if $i \sim j$ or $i = j$. If the neighbourhood is relatively small Q is sparse; a matrix where most elements are zero. We benefit from using GMRF rather then GRF models because computational complexity of both evaluation and sampling of GMRFs is more then an order cheaper then for GRF, see results in Rue (2001), Rue and Follestad (2003) and Steinsland (2003). GMRFs often appear as building blocks in spatial models, see e.g. Cressie (1993), Wikle et al. (1998), Besag and Higdon (1999), Fernandez and Green (2002), Heikkinen and Arjas (1998) and Besag and Kooperberg (1995).

The distribution of interest is the posterior $\pi(x, \theta|y)$ and it is given by

$$\pi(x, \theta|y) \propto \pi(y|x)\pi(x|\theta)\pi(\theta)$$

where $\pi(\theta)$ is the prior of the hyper-parameters. An illustration of the general model setting we work with is given in figure 1.

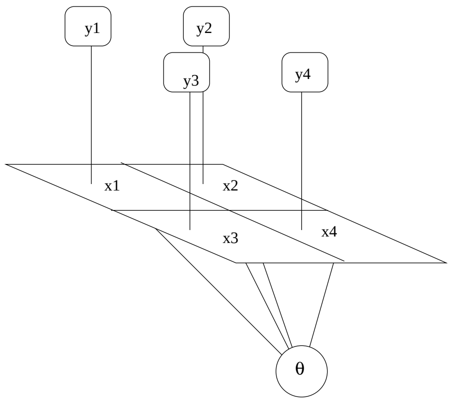


Figure 1: *A typical latent field model. The latent field $x = (x_1, x_2, x_3, x_4)$ depends on hyper-parameter(s) θ. The observations $y = (y_1, y_2, y_3, y_4)$ is modelled as independent given x.*

Image analysis and disease mapping are two areas where latent GMRF models have been popular. To give the reader a better notion of the kind of problems considered we introduce two problems to be analysed in examples later and set up models for them.

Disease mapping in Germany

The German oral cavity cancer dataset consists of all cases of oral cavity cancer mortality for males from 1986 to 1990 in each of Germany's 544 administrative regions. It has previously been studied in Knorr-Held and Raßer (2000) and in this report used in example 10, section 5.2. There are between 1 and 505 cases in each region and the number of cases relative to population size, y_i/c_i, is given in figure 2. Here c_i is the expected number of cases in region i if the individual risks were equal; let h_i be the population in region i,

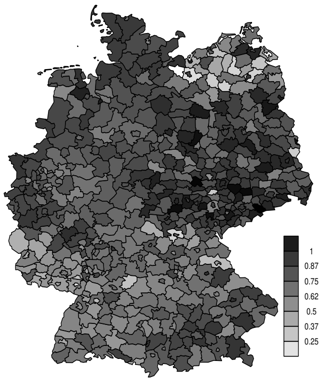


Figure 2: *Administrative map of Germany with oral cavity cancer point-vise relative risk* y_i/c_i.

then $c_i = h_i \sum_{\forall j} y_j / \sum_{\forall j} h_j$. Our interest is the log relative risk and its spatial structure. In disease mapping a common approach is to give the log relative risk an intrinsic GMRF prior, where region i and j are neighbours, $i \sim j$, if they share a boarder. The smoothing parameter is set to κ and

$$\pi(x|\kappa) \propto \kappa^{(n-1)/2} \exp(-\frac{1}{2}\kappa \sum_{i\sim j}(x_i - x_j)^2)$$

The number of cases in the regions, y_i $i = 1, 2, \ldots, n = 544$, is assumed conditionally independent Poisson with expected values $c_i \exp(x_i)$. Our interest is the posterior distribution of the log relative risk x and the smoothing parameter κ, given as

$$\pi(x, \kappa|y) \propto \pi(y|x)\pi(x|\kappa)\pi(\kappa)$$

where $\pi(\kappa)$ is the prior of κ.

Magnetic Resonance Image of the brain

In example 9 (section 4.4) a time series of magnetic resonance (MR) images is analysed, see figure 3 for the first image of the time series. Here we introduce only a part of the problem; estimating the underlying truth from one image.

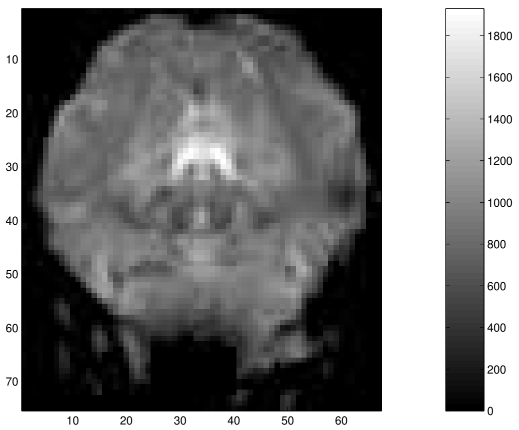


Figure 3: *The first MR image of the image time series analysed in example 9.*

The MR image is a lattice of observed magnetic resonance responses from the corresponding area in the brain. The measurement process is not perfect and observations are often assumed independent Gaussian conditioned on the true responses x, $y_i|x \sim N(x_i, \tau^{-1})$. The true image x is given the same intrinsic GMRF prior (with smoothing parameter θ) as the log relative risk field above. For the image we use a neighbourhood where each non-boarder pixel has four neighbours. Our interest is the posterior distribution of the true image and the hyper-parameters $\theta = (\kappa, \tau)$. It is given by

$$\pi(x, \theta|y) \propto \pi(y|x, \tau)\pi(x|\kappa)\pi(\theta)$$

where $\pi(\theta)$ is the prior of θ.

1.2 One-block updating scheme in MCMC

Only for a few special likelihood and prior choices it is possible to do analytically inference from latent GMRF models. In this report focus is on Markov chain Monte Carlo sampling methods for doing inference. Empirical studies, e.g. Knorr-Held and Rue (2002) and Gamerman et al. (2003), demonstrate that the latent field x and the hyper-parameters θ should be updated simultaneously to get appropriate mixing. In Rue and Follestad (2003) a model with $\mu \sim N(0, \tau_\mu^{-1})$ and $x = (x_1, x_2, \ldots, x_n)$, x_i i.i.d. $N(\mu, \tau_x)$ is consider. The aim is to sample from the posterior of μ (which is $N(0, \tau_\mu)$), and a two-block Gibbs sampler is used sampling μ in one block and x in one block, see algorithm 1. They prove that the correlation length of $\mu^1, \mu^2, \ldots$ is linear in the dimension of x, n. Hence the importance of one-block updating increase with the dimension of the latent field and is essential for high-dimensional problems.

The updating scheme we aim to use is given in algorithm 2. This is an one-block

Algorithm 1 TWO-BLOCK GIBBS-SAMPLER

- Given x^0 and μ^0
- for $j = 0 : (niter - 1)$
  - $\mu^{j+1} \sim \pi(\mu|x^j)$
  - $x^{j+1} \sim \pi(x|\mu^{j+1})$
- Return $(x^1, x^2, \ldots, x^{niter})$ and $(\mu^1, \mu^2, \ldots, \mu^{niter})$

Algorithm 2 GENERAL ONE-BLOCK METROPOLIS-HASTING SAMPLER

- Given x^0 and θ^0
- for $j = 0 : (niter - 1)$
  - Sample $\theta^{new} \sim q(\theta|\theta^j)$
  - Sample $x^{new} \sim q(x|x^j, \theta^{new})$.
  - Calculate the acceptance probability and accept / reject
  - if(accept)
    * $\theta^{j+1} = \theta^{new}$ and $x^{j+1} = x^{new}$
  - else
    * $\theta^{j+1} = \theta^j$ and $x^{j+1} = x^j$
- Return $(x^1, x^2, \ldots, x^{niter})$ and $(\theta^1, \theta^2, \ldots, \theta^{niter})$

updating scheme where the proposal consists of two steps. In the first step a set of new hyper-parameters, θ^{new}, is proposed. This proposal distribution is independent of the current values of the latent field, but may depends on the current values of the hyper-parameter. The dimension of θ is low and in most cases some kind of independent random walk proposal is appropriate. In the second step of the proposal a latent field x^{new} is proposed. The next stage of the algorithm is to accept or reject (θ^{new}, x^{new}) jointly.

The challenging part of this algorithm is to find a good proposal for the latent field; $q(x|x^{old}, \theta^{new})$. Distributions that are approximations to $\pi(x|\theta^{new})$ and that are independent of x^{old} are proposed in Rue et al. (2003) and also used in Steinsland (2003). These kind of x^{old} independent proposals are computationally demanding with a minimum cost of $\mathcal{O}(n^{3/2})$ for a spatial Gaussian Markov random field of dimension n. In this report we construct a proposal for x based on overlapping small blocks. This overlapping blocks proposal enable us to get samples from an distribution that is an approximation of $\pi(x|y, \theta)$ without sampling directly from an n dimensional distribution.

1.3 Outline

The overlapping block Gibbs proposal is presented in chapter 2 when each sub-block is sampled exact and it can be concidered a Gibbs sampler. We also look more into an important subclass with an one-dimensional structure of the overlapping blocks and briefly introduce a partial conditioning block sampler. Since $q(x|x^{old}, \theta^{new})$ is only one part of the overall proposal we need to use it in a Metropolis-Hasting accept/reject step even when it is build up by block Gibbs samplings steps. In chapter 3 we introduce an opposite kernel. Together with an opposite reverse acceptance probability it makes us keep the acceptance rate equal to one for the updates of $x|\theta$ when the blocks are sampled using Gibbs steps. The full one-block updating scheme Metropolis-Hasting algorithm is set up in chapter 4 followed by examples. In chapter 5 an overlapping block proposal with sampling from approximated distributions is introduced and empirically tested. The report is enclosed with a discussion in chapter 6.

2 Proposal from overlapping block Gibbs sampler

In this section we assume we are able to sample exact from the conditional posterior distribution for each sub-block x_B of the latent field i.e. from $\pi(x_B|x_{-B}, y, \theta)$. Further the hyper-parameters θ are assumed fixed and we are in a setting where we are able to run a Gibbs sampler. For evaluation of the samplers we keep in mind our purpose; to construct a proposal for the latent field x when hyper-parameters change. There are three important features we want our proposal to have: We want it to be a distribution close to the exact one, $\pi(x|y, \theta)$, we want succeeding samples to have low dependency, and evaluation and sampling should be computationally affordable. The first property is necessary to get acceptance in the one-block updating scheme sampler, while the second is important for the mixing of the Markov chain. In the following θ and y are suppressed from our target distribution, which is now denoted $\pi(x)$.

2.1 Traditional block Gibbs sampler

One opportunity is to use one scan of a Gibbs sampler as the proposal for the latent field. One scan of a single site Gibbs sampler is not appropriate because of slow mixing. A Gibbs sampler with sampling done on disjunct sets of variables instead of single variables is known as a *block Gibbs sampler*, see algorithm 3.

Algorithm 3 TRADITIONAL BLOCK GIBBS SAMPLER

- Given x^0 and a disjunct partition of x; $x = \{x_{B_1}, x_{B_2}, \ldots, x_{B_K}\}$
- for $j = 0 : (niter - 1)$
  - $x = x^j$
  - for $k = 1 : K$
    - Sample $x^*_{B_k} \sim \pi(x_{B_k}|x_{-B_k})$
    - $x_{B_k} = x^*_{B_k}$
  - $x^{j+1} = x$
- Return $x^1, x^2, \ldots x^{niter}$

It is well known that sampling in blocks can improve mixing within the blocks, but on block borders high correlation is still a problem. An illustration is given in example 1 below. Common solutions to this problem are to select block sizes and locations randomly or to let succeeding scans systematically have different border locations. These solutions do not suite us since we want to use one scan of a Gibbs sampler as the proposal for the latent field and hence only as one part of our overall one-block proposal. In the next section we try to overcome the border problems of block Gibbs samplers and construct a block Gibbs sampler with blocks that are *not* disjunct.

Example 1: Traditional block sampling

We consider a zero-mean Gaussian Markov random field (GMRF) on a lattice of size 100×100 which is a proxy to a GRF with exponential correlation function;

$$\rho(x_i, x_j) = \exp(\frac{-3d(i,j)}{r})$$

where $d(i,j)$ is the distance between element i and j (each pixel is set to be of size 1×1). A 5×5 neighbourhood is used and the coefficients in the precision matrix Q is chosen using the way of approximating a GMRF to a GRF introduced in Rue and Tjelmeland (2002). We keep in mind that we want to use the Gibbs sampler as a proposal for the latent field when also hyper-parameters are changing and set the initial values of the latent field to 3. For range $r = 40$ we have run a single site Gibbs sampler, block Gibbs samplers with 16 blocks (each of size 25×25 pixels) and 4 blocks (each of size 50×50 pixels) and an exact sampler. Images of the samples after one and 200 systematic scans are in figure 4. In figure 5 is the first sample for a block Gibbs sampler with 2×2 blocks for different

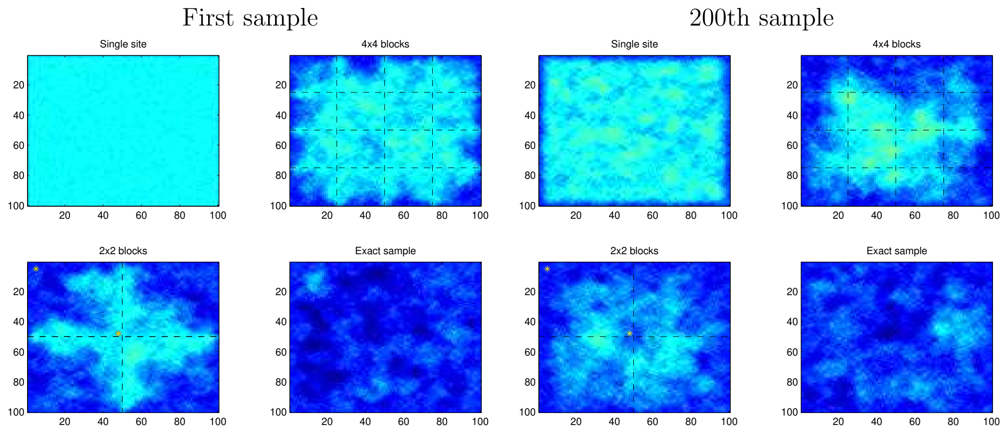


Figure 4: *To example 1: The first (left) and 200th (right) sample form (from upper left) single site Gibbs, block Gibbs with four and eight blocks and an exact sampler. In lower left images is element* $(5,5)$ *and* $(48,48)$ *are marked with stars.*

ranges r. From these figures we see how blocking helps for the convergence and also how the border regions are held back by the initial values. As the internal dependence in the field (here r) increases the slow converging border regions become wider.

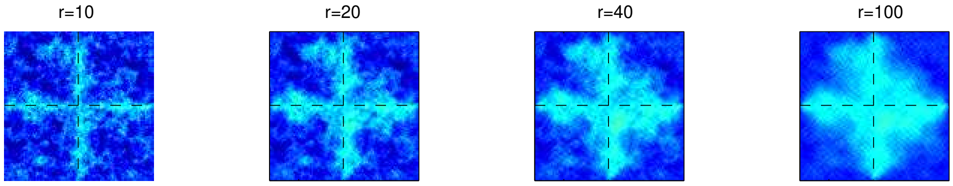


Figure 5: *To example 1: The first sample from block Gibbs samplers for different ranges (r). The initial values for the field is three.*

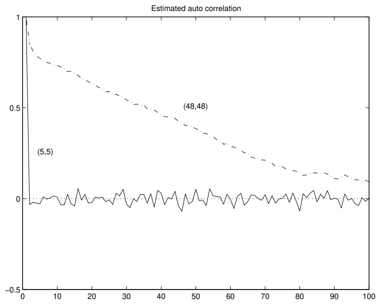


Figure 6: *To example 1: Estimated auto-correlation for element* $(5, 5)$ *and* $(48, 48)$ *for the block Gibbs sampler with* 2×2 *blocks. These pixels are marked with stars in figure 4. Model as in example 1 with* $r = 40$*, and with a sample from the target distribution as initial value.*

In figure 6 the estimated auto-correlations for a point close to the edge of the field and a point close to the border between blocks are shown for $r = 40$. We see that the variable of the border pixel has much poorer mixing then the variable of the edge pixel.

2.2 Overlapping block Gibbs sampler

Our idea is to let the blocks overlap such that the border regions are in (at least) two blocks. This results in a block Gibbs sampler with non-disjunct blocks, see figure 7 and algorithm 4. We will refer to these kind of samplers as overlapping block Gibbs samplers.

Consider a field of variables blocked as in figure 7. The grey block covering x_1, x_2, x_4 and x_5 is the first block of the overlapping block Gibbs sampler; $B_1 = \{1, 2, 4, 5\}$. The second block $B_2 = \{2, 3, 5, 7\}$, the third one $B_3 = \{4, 5, 7, 8\}$ and $B_4 = \{5, 6, 8, 9\}$. Of the variables sampled in B_1 only those in x_1 are part of the sampled returned after a whole scan, the other variables $\{x_2, x_4, x_5\}$ are sampled over later in the scan. The first sample of $\{x_2, x_4, x_5\}$ works as a buffer between the new field being sampled (here $\{x_1\}$) and the part of the old field yet not sampled over (here $\{x_3, x_6, x_7, x_8, x_9\}$). We hope these buffers make the new sample for the field (almost) independent of the old one, also when

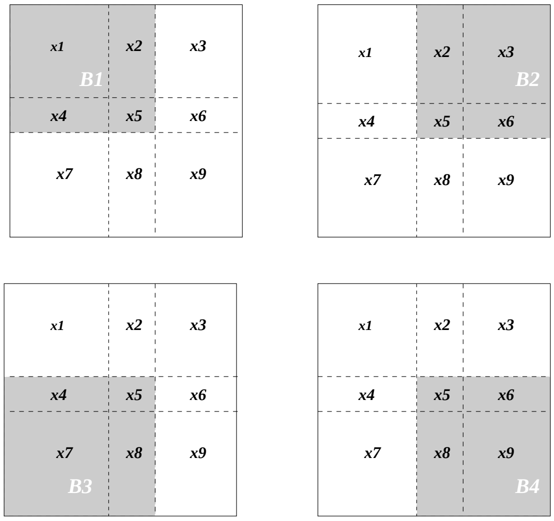


Figure 7: *Illustration of the blocks of an overlapping block Gibbs sampler. Notation used in algorithm 4 and in the proof in appendix A1.*

its hyper-parameters have changed.

Let x be the old sample, $x^{'}$ the new one and index the temporary samples for the buffers with their block numbers. The transition kernel of the sampler is then given by;

$$\begin{aligned} K(x, x^{'}) \;=\; & \int [\pi(x_1^{'}, x_2^{B1}, x_4^{B1}, x_5^{B1} | x_3, x_6, x_7, x_8, x_9) \\ & \pi(x_2^{'}, x_3^{'}, x_5^{B2}, x_6^{B2} | x_1^{'}, x_4^{B1}, x_7, x_8, x_9) \\ & \pi(x_4^{'}, x_5^{B3}, x_7^{'}, x_8^{B3} | x_1^{'}, x_2^{'}, x_3^{'}, x_6^{B2}, x_9) \\ & \pi(x_5^{'}, x_6^{'}, x_8^{'}, x_9^{'} | x_1^{'}, x_2^{'}, x_3^{'}, x_4^{'}, x_7^{'})] dx_2^{B1} dx_4^{B1} dx_5^{B1} dx_5^{B2} dx_5^{B3} dx_6^{B2} dx_8^{B3} \end{aligned}$$

It is intuitive that $\pi(x)$ is the stationary distribution for the corresponding Markov chain and a proof is given in appendix A1. The sampler is also valid with other block and buffer configurations. The only requirement is that each element is updated at least once. A special case is the traditional block Gibbs sampler.

The extra computational cost caused by the buffers depends on the cost of the sampler. To sample from a spatial Gaussian Markov random field (GMRF) of dimension n costs $\mathcal{O}(n^{3/2})$. Let the size of blocks to be sampled be $(l_b\sqrt{n}) \times (l_b\sqrt{n})$ variables in the overlapping case when a traditional block Gibbs sampler would have blocks of $\sqrt{n} \times \sqrt{n}$ variables. This corresponds to buffers of length $2(l_b - 1)\sqrt{n}$ and the computation time per iteration has increased with a factor l_b^3 due to the overlapping blocks for $\pi(x)$ a GMRF.

Algorithm 4 Overlapping blocks Gibbs sampler

- Given x^0
- for $i = 0 : (niter - 1)$
  - Sample $(x_1^{i+1}, x_2^{B1}, x_4^{B1}, x_5^{B1}) \sim \pi(x_{B_1} | x_3^i, x_6^i, x_7^i, x_8^i, x_9^i)$
  - Sample $(x_2^{i+1}, x_3^{i+1}, x_5^{B2}, x_6^{B2}) \sim \pi(x_{B_2} | x_1^{i+1}, x_4^{B1}, x_7^i, x_8^i, x_9^i)$
  - Sample $(x_4^{i+1}, x_5^{B3}, x_7^{i+1}, x_8^{B3}) \sim \pi(x_{B_3} | x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_6^{B2}, x_9^i)$
  - Sample $(x_5^{i+1}, x_6^{i+1}, x_8^{i+1}, x_9^{i+1}) \sim \pi(x_{B_4} | x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_4^{i+1}, x_7^{i+1})$
- Return $((x_1^1, x_2^1, \ldots, x_9^1), (x_1^2, x_2^2, \ldots, x_9^2), \ldots, (x_1^{niter}, x_2^{niter}, \ldots, x_9^{niter}))$

Example 2: Mixing with overlapping block Gibbs samplers

To explore how overlapping blocks influence the mixing we have tested different blocking schemes. We use the same case as in example 1; a zero mean Gaussian Markov random field on a 100×100 lattice, with a 5×5 neighbourhood. It is a proxy of a GRF with an exponential correlation function ρ with range $r = 40$;

$$\rho(x_i, x_j) = \exp(\frac{-3d(i,j)}{r})$$

where $d(i, j)$ is the distance between pixel i and j, and each pixel has size 1×1. We use a sample from the target distribution $\pi(x)$ as out initial field and run all the Gibbs samplers for 200 systematic scans. The Gibbs samplers we test are single site, traditional block Gibbs sampler with 2×2 blocks (i.e. each block is of 50×50 pixels), and overlapping block Gibbs samplers with 2×2 blocks each extended with buffers of one, two and five pixels in both directions (hence blocks of (51×51), (52×52) and (55×55) pixels and overlapping buffer lengths two, four and ten, respectively). An one-block Gibbs sampler is also ran for reference purposes. This is the ideal case; the samples are exact samples from $\pi(x)$, they are independent and the Markov chain converges immediately. Figure 8 shows the initial field and the sample from the 200th scan for the five blocking schemes tested. We can see many similarities between the initial field and the 200th iteration of the single site sampler and believe the 200th sample is highly correlated with the initial field.

Since the same model is used here as the traditional block Gibbs sampler was tested for in example 1 section 2.1, we know from figure 6 that succeeding samples at $(5, 5)$ are almost independent, while they at $(48, 48)$ are very dependent. To explore how different buffer lengths influence the mixing we have made trace plot, cumulative mean plot, cumulative variance plot and plotted the estimated auto-correlation for pixel $(48, 48)$ for each sampler tested, see figure 9. We observe that the auto-correlation decreases faster with larger buffers. For buffers of length four and ten the estimated auto-correlation decreases much faster then in the block sampler without buffers. This example support our hypotheses

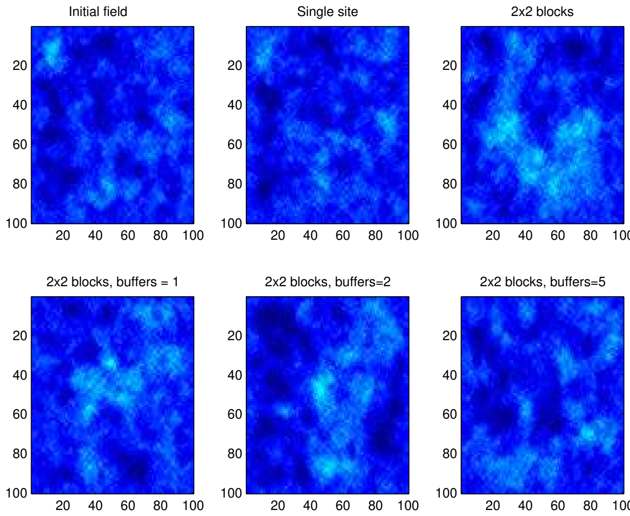


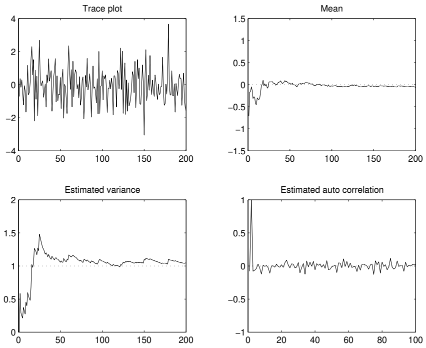
Figure 8: *To example 2: The initial field (top left) and the sample after 200 scans for the different blocking schemes. The bottom row is for the overlapping block Gibbs samplers with* 2×2 *blocks each of (from left)* 51×51, 52×52 *and* 55×55 *pixels.*

that overlapping blocks give better mixing and the pay-off is better then the increase in computation time. In the overlapping block sampler with buffer length ten blocks of 55×55 variables are sampled. Compared with the corresponding traditional block Gibbs sampler where blocks of 50×50 variables are sampled the computational cost has increased about 30%.

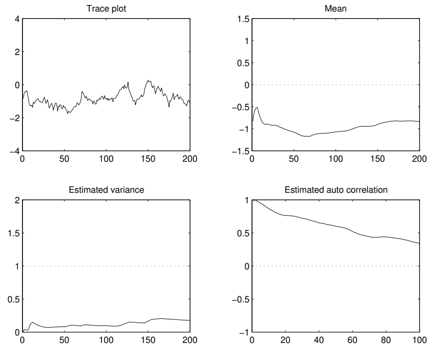
Example 3: Burn-in with different ranges for overlapping block Gibbs samplers

In this example we visually and with estimated auto-correlation inspect a case relevant for how we intend to use the sampler. We want to be able to update the hyper-parameters independent of the field, i.e. the initial field can be quite far away from the high density areas of the distribution we now want to sample from. A sampler with a short burn-in would be helpful. In this example a constant field with value three is used as initial values. The distribution we want to sample from is the same as in example 1, i.e. the initial field is three standard deviations from the expected value zero. We have explored how different values of the range (r) influence the dependence to the initial field and how overlapping blocks change this dependence. Figure 10 shows the first sample from samplers with different buffer sizes and different ranges, and plots of the estimated auto-correlations are in figure 11.

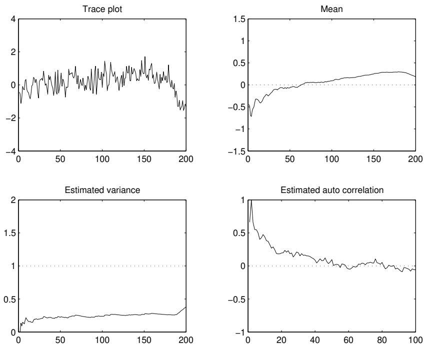
The images in figure 10 suggest what is to be accepted; larger range gives higher dependence between the initial field and the first sample, overlapping blocks decrease



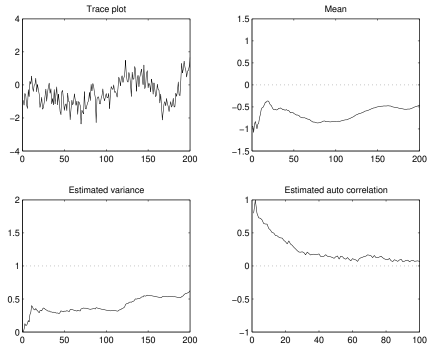
Exact sampling



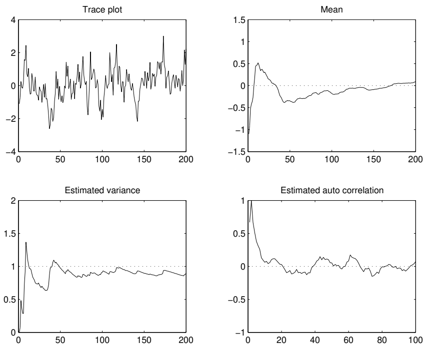
Single site



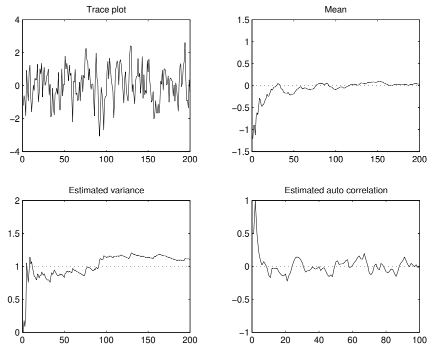
Traditional 2×2 blocks



Overlapping, buffer 1



Overlapping, buffer 2



Overlapping, buffer 5

Figure 9: *To example 2: From top left; an exact sampler, a traditional block Gibbs sampler with 2×2 blocks, and overlapping block Gibbs samplers with 2×2 blocks and blocks extended with one, two and five pixels in both direction. For each sampler, from top left to right: Trace plot, plot of cumulative mean, plot of cumulative estimated variance and plot of the estimated auto-correlation function.*

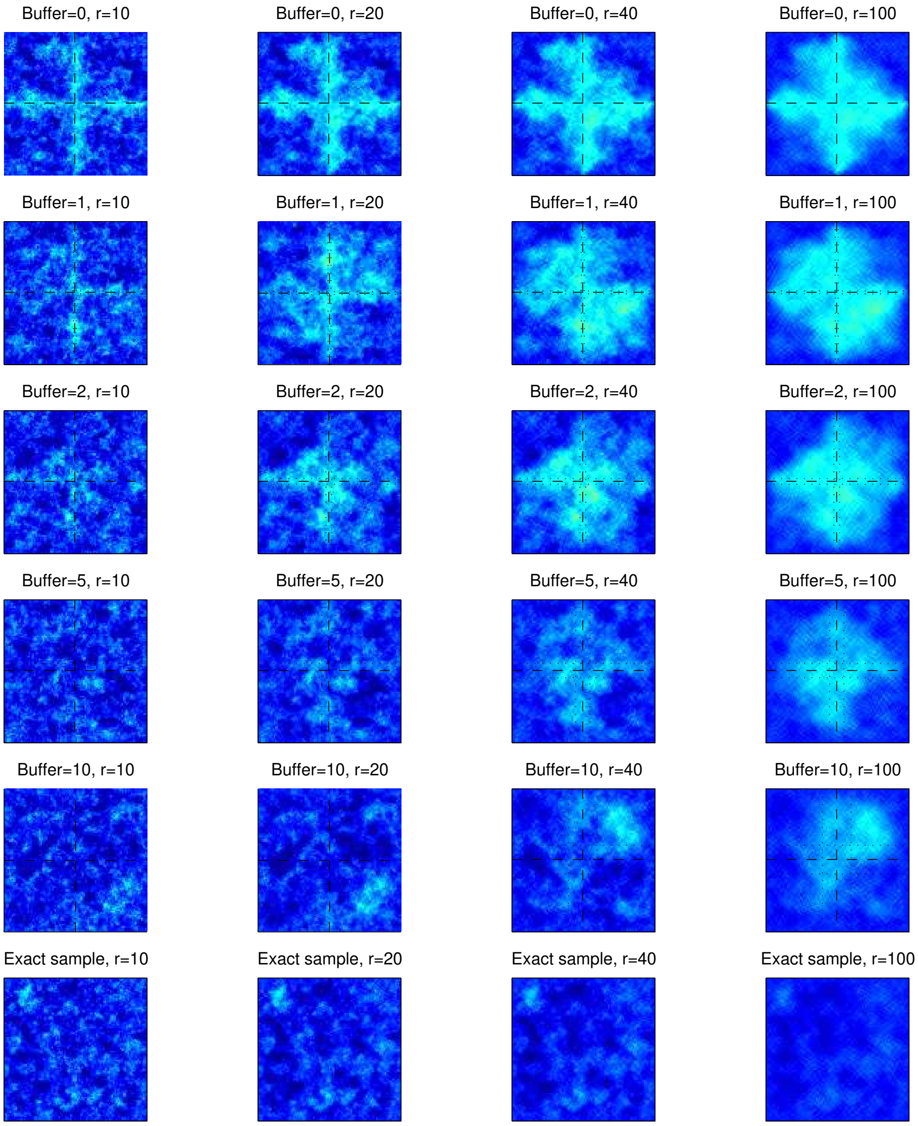


Figure 10: *To example 3: The first sample from overlapping block Gibbs samplers for different ranges (r) and buffer sizes (each block extended with buffer pixels in both direction). The initial values of the field is three.*

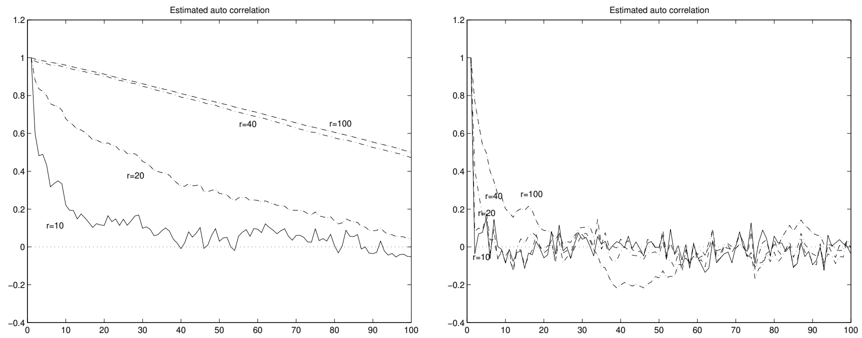


Figure 11: *To example 3: Estimated auto-correlation function for samples at pixel* $(48, 48)$ *for the traditional block Gibbs sampler with* 2×2 *blocks (left) and the overlapping block Gibbs sampler with* 2×2 *blocks, each extended with a buffer of five pixel in both direction (right)*

the dependence, and larger buffers are required to give equal results for increased range. The plots of estimated auto-correlation in figure 11 support this: As the range increases auto-correlation between samples also increase. Further, overlapping blocks decrease the auto-correlation for all the tested ranges, especially for the largest ranges.

2.3 Time series overlapping block Gibbs sampler

A special case of the overlapping block Gibbs sampler is a time series version of the overlapping blocks. If we only let blocks overlap in one direction, or more precise never condition on temporary (buffer) samples, we obtain some nice simplifications, see figure 12 and the corresponding algorithm, algorithm 5. The algorithm looks similar to the general over-

Algorithm 5 Overlapping block Gibbs sampler, time series

- Given x^0
- for $i = 0 : (ninter - 1)$
  - Sample $(x_1^{i+1}, x_2^{B1}) \sim \pi(x_{B_1} | x_3^i, x_4^i, x_5^i)$
  - Sample $(x_2^{i+1}, x_3^{i+1}, x_4^{B2}) \sim \pi(x_{B_2} | x_1^{i+1}, x_4^i)$
  - Sample $(x_4^{i+1}, x_5^{i+1}) \sim \pi(x_{B_3} | x_1^{i+1}, x_2^{i+1}, x_3^{i+1})$
- Return $((x_1^1, x_2^1, \ldots, x_5^1), (x_1^2, x_2^2, \ldots, x_5^2), \ldots, (x_1^{niter}, x_2^{niter}, \ldots, x_5^{niter})$

lapping block Gibbs sampler in algorithm 4. The simplification is found in the transition

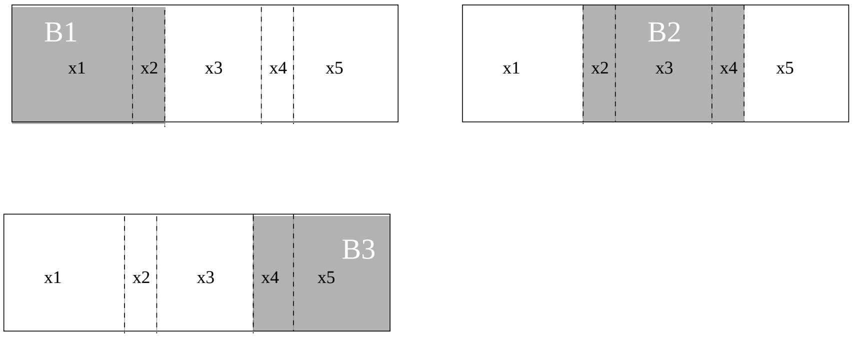


Figure 12: *Illustration of the the time series version of the overlapping blocks Gibbs sampler. Notation as used in algorithm 5.*

kernel: We are now able to integrate out the buffers;

$$
\begin{aligned}
K(x, x') &= \int [\pi_1(x_1', x_2^{B1} | x_3, x_4, x_5) \\
&\quad \pi_2(x_2', x_3', x_4^{B2} | x_1', x_5) \\
&\quad \pi_3(x_4', x_5' | x_1', x_2', x_3')] dx_2^{B1} dx_4^{B2} \\
&= \pi(x_1' | x_3, x_4, x_5) \cdot \pi(x_2', x_3' | x_1', x_5) \cdot \pi(x_4', x_5' | x_1', x_2', x_3')
\end{aligned}
$$

In the time series situation the overlapping block Gibbs sampler gives us partial conditioning sampling, where the buffers in each block are integrated out. Partial conditioning sampling is mentioned in Besag et al. (1995) as a way doing MCMC, but then as an alternative to Gibbs sampling and not a way of constructing a proposal.

If we consider only the variables of a block not sampled later they are sampled from their marginal conditional distribution with the other variables of the block integrated out. This can be thought of as local version of collapsing as defined in Liu (1994). And the time series version of the overlapping Gibbs sampler can be viewed as a traditional block Gibbs sampler with local use of collapsing.

When we later want to calculate the density for the transition a way of doing that is to use (for each block);

$$
\pi(x_1' | x_3^*, x_4^*, x_5^*) = \frac{\pi(x_1', x_2^{B1} | x_3^*, x_4^*, x_5^*)}{\pi(x_2^{B1} | x_1', x_3^*, x_4^*, x_5^*)}
$$

for any x_2^{B1}, e.g. the actual buffer we have sampled. Hence, the time series version of the overlapping block Gibbs sampler enables us to calculate the transition density for the full update of the field. This is essential when we later want to incorporate it into an one-block updating scheme together with hyper-parameters.

While an overlapping block Gibbs sampler and a partial conditional block sampler is equal with a time series version of blocking is this generally not the case. In a partial conditional block sampler there are non temporary buffer samples. In figure 13 and algorithm 6 the partial conditional block sampler corresponding to the overlapping block Gibbs sampler in figure 7 and algorithm 4 is set up.

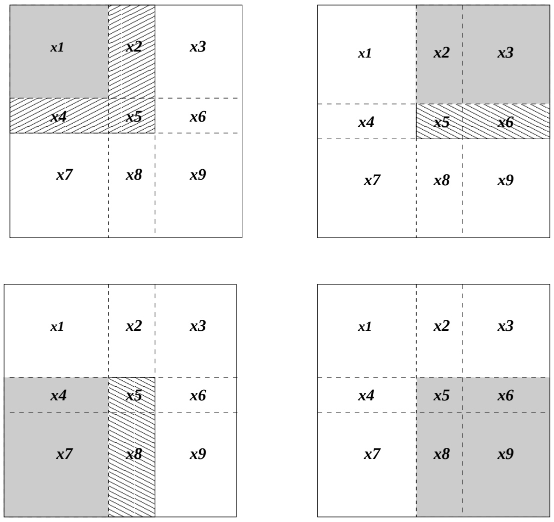


Figure 13: *Illustration of the blocks of a partial conditional block sampler. Notation used in algorithm 6. In each step the gray area is sampled from its partially conditional distribution with the hatched area integrated out.*

The transition kernel is given by;

$$\begin{aligned} K(x,x^{'}) &= \pi(x_1'|x_3,x_6,x_7,x_8,x_9) \\ &\quad \pi(x_2',x_3'|x_1',x_4,x_7,x_8,x_9) \\ &\quad \pi(x_4',x_7'|x_1',x_2',x_3',x_6,x_9) \\ &\quad \pi(x_5',x_6',x_8',x_9'|x_1',x_2',x_3',x_4',x_7') \end{aligned}$$

The actual sampling can be done as in the overlapping block Gibbs sampler, but the samples for the buffers are not kept. The density for the transition can be calculated as in the time series overlapping case.

We will in this report only use the time series overlapping blocks Gibbs sampler when we need to calculate the transition density. Our largest problems are time series and to use the time series overlapping blocks approach is then a natural choice.

Algorithm 6 PARTIAL CONDITIONAL BLOCK SAMPLER

- Given x^0
- for $i = 0 : (niter - 1)$
  - Sample $(x_1^{i+1}) \sim \pi(x_1|x_3^i, x_6^i, x_7^i, x_8^i, x_9^i)$
  - Sample $(x_2^{i+1}, x_3^{i+1}) \sim \pi(x_2, x_3|x_1^{i+1}, x_4^i, x_7^i, x_8^i, x_9^i)$
  - Sample $(x_4^{i+1}, x_7^{i+1}) \sim \pi(x_4, x_7|x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_6^i, x_9^i)$
  - Sample $(x_5^{i+1}, x_6^{i+1}, x_8^{i+1}, x_9^{i+1}) \sim \pi(x_5, x_6, x_8, x_9|x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_4^{i+1}, x_7^{i+1})$
- Return $((x_1^1, x_2^1, \ldots, x_9^1), (x_1^2, x_2^2, \ldots, x_9^2), \ldots, (x_1^{niter}, x_2^{niter}, \ldots, x_9^{niter}))$

Example 4: Sampling variance in an AR(1) process

We want to explore the difference between a traditional block Gibbs sampler and an overlapping one for an AR(1) process. The block border is of special interest. We consider an

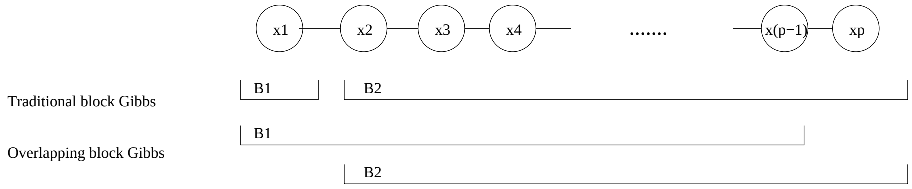


Figure 14: *To example 4: Illustration of the AR(1) blocking example.*

AR(1) process x of length p, block Gibbs samplers with blocks of length $b = p - 1$, and for the overlapping block Gibbs sampler buffer of length $\beta = p - 2 = b - 1$. See figure 14 for an illustration. We choose the variance to be 1 and set the correlation at lag one to ρ. This gives a covariance matrix;

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{p-1} \\ \rho & 1 & \rho & \ldots & \rho^{p-2} \\ \vdots & & \ddots & & \vdots \\ & & & 1 & \rho \\ \rho^{p-1} & & \ldots & \rho & 1 \end{bmatrix}$$

An exact sampler and the transition kernels of the two Gibbs samplers can be written as:

$$\begin{aligned} \text{Exact sampler:} \quad q(x|x') &= \pi(x_1)\pi(x_2, \ldots, x_p|x_1) \\ \text{Traditional block Gibbs:} \quad q(x|x') &= \pi(x_1|x'_2, \ldots, x'_p)\pi(x_2, \ldots, x_p|x_1) \\ \text{Overlapping block Gibbs:} \quad q(x|x') &= \pi(x_1|x'_{2+\beta})\pi(x_2, \ldots, x_p|x_1) \end{aligned}$$

The last block would be exact sampled if x_1 was. We therefore focus on the first part $\pi(x_1|\dots)$. How good the mixing is depends on the variance of this distribution. In the traditional case it is

$$\text{Var}(x_1|x_2,\dots,x_p) = Var(x_1|x_2) = 1-\rho^2$$

while in the overlapping blocks case it is

$$\begin{aligned}\text{Var}(x_1|x_{2+\beta}) &= [\text{Cov}(x_1,x_2,\dots,x_{\beta+1}|x_p)]_{(1,1)} \\ &= [\text{Cov}(x_1,x_2,\dots,x_{p-1}) - [\rho^{\beta+1}\ \rho^{\beta}\dots\rho]^T[\rho^{\beta+1}\ \rho^{\beta}\dots\rho]]_{(1,1)} \\ &= 1-\rho^{2(\beta+1)}\end{aligned}$$

where $[A]_{(i,j)}$ is element (i,j) of matrix A. The result is not surprising: As the buffer-length β increases the variance in x_1's proposal, $1-\rho^{2(\beta+1)}$, approaches x_1's marginal variance (here 1).

Example 5: Acceptance rate versus number of blocks and buffer lengths

For an independent Metropolis-Hastings sampler with the optimal acceptance probability from Peskun (1973) the acceptance rate is a good measure of how close the proposal is the target distribution, see section 6.4.1 in Robert and Casella (1999). If we use one scan of the time series version of an overlapping Gibbs sampler as proposal and a sample from $\pi(x)$ as initial values the proposed sample, x^{new}, is from $\pi(x)$, but it is not independent of the current sample x^{old}. Our Metropolis-Hasting algorithm produces a reversible Markov chain and this motivates us to use the acceptance rate as a measure of how far from reversible samples our proposal gives. We aim to propose x^{new} that is (almost) independent of x^{old}. If we succeed reversibility with respect to π is achieved and we would get acceptance probability one in the Metropolis-Hasting algorithm. We can therefore in this situation use the acceptance rate as a measure of how independent the proposed field, x^{new}, is the current field, x^{old}. We have used different numbers of blocks and different buffer lengths for the same GMRF model as in example 2 on a 100×10 lattice. The initial field was a sample from the target distribution and all the samplers were run for 1000 scans. The acceptance rates are plotted in figure 15.

We find that the acceptance rate decreases as the number of blocks increases and as the buffer length decreases. More blocks means more troublesome borders and less reversibility. Our cure for this is overlapping blocks and as the buffer length increases the border problem decreases; we get more independent samples. For buffer length 10 we get an acceptance rate > 0.99 and we believe the samples are close to independent exact samples.

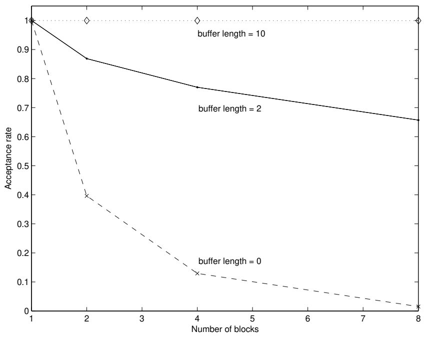


Figure 15: *To example 5: Acceptance rates with fixed hyper-parameters for a* 10×100 *lattice with buffer lengths* 0, 2 *and* 10.

3 From Gibbs to Metropolis-Hastings

The proposal for x, $q(x|x^{old}, \theta^{new})$ is going to be incorporated into an one-block updating scheme Metropolis-Hastings algorithm together with the proposal for the hyper-parameters. If one step of the overlapping Gibbs sampler presented in chapter 2 is used as proposal together with the optimal acceptance probability in Peskun (1973) we do not get acceptance probability 1 even when hyper-parameters are fixed. As discussed in example 5 this is not due to samples not being from $\pi(x)$, but because the Gibbs sampler used does not produce a reversible Markov chain. Getting "any" acceptance is often a problem, and we would like to keep the Gibbs property of acceptance probability 1 when hyper-parameters are fixed. For a given $q(x|x^{old}, \theta^{new})$ the tuning of the hyper-parameters' proposals then control the level of the acceptance probability.

A way of achieving this is to use a symmetric scan Gibbs sampler. A symmetric scan could for the blocks in figure 7 be $B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_3 \rightarrow B_2 \rightarrow B_1$. This proposal would give acceptance probability 1 with fixed hyper-parameter. Though, the computational cost is twice the original one.

It is possible to use a mixture of proposals as a proposal. Let q_0 be the proposal distribution for updating scheme $B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$, and q_1 the proposal distribution for updating scheme $B_4 \rightarrow B_3 \rightarrow B_2 \rightarrow B_1$. We randomly choose which proposal to use, each with probability 0.5; $P(q_0) = P(q_1) = 0.5$. This gives us an overall proposal

$$q(x^{'}|x) = 0.5q_0(x^{'}|x) + 0.5q_1(x^{'}|x)$$

Instead of using the Peskun's optimal acceptance probability formulae;

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

we use the acceptance probability suggested in Tjelmeland and Hegstad (2002):

$$\alpha_{i,1-i}(y|x) = \min\left\{1, \frac{\pi(x')q_{1-i}(x|x')}{\pi(x)q_i(x'|x)}\right\} \qquad i \in \{0,1\}$$

We will refer to this acceptance probability as opposite reverse acceptance probability. It does not give optimal convergence as a function of Metropolis-Hastings steps. But if the overall proposal is computationally expensive to evaluate it may give optimal convergence as function of computation time. We write our proposal distributions as a product of transition kernels:

$$q_0(x|x') = \int q^1(x_1'|y)q^2(x_2'|x_1') \dots q^b(x|x_{b-1}')dx_1'dx_2' \dots dx_{b-1}' \tag{1}$$

and

$$q_1(x|x') = \int q^b(x_{b-1}'|x')q^{b-1}(x_{b-2}'|x_1') \dots q^1(x|x_1^{'})dy_{b-1}dy_{b-2} \dots dy_1 \tag{2}$$

In our setting is

$$q^i(x|x^{'}) = \begin{cases} \pi(x_{B_i}|x^{'}_{-B_i}), & \text{if } x_j = x^{'}_j, \forall j \notin B_i \\ 0, & \text{if } \exists j \notin B_i \text{ such that } x_j \neq x^{'}_j \end{cases}$$

If $x \sim \pi(x)$ will also $x_i \sim \pi(x)$. We first consider $\alpha_{0,1}$ and its denominator $\pi(x)q_0(x'|x)$.

$$\begin{aligned}
\pi(x)q_0(x'|x) = &\quad \pi(x) \int q^1(x_1'|x)q^2(x_2'|x_1') \dots q^b(x'|x_{b-1}')dx_1'dx_2' \dots dx_{b-1}' \\
= &\quad \int q^1(x_1'|x)\pi(x)q^2(x_2'|x_1') \dots q^b(x|x_{b-1}')dx_1'dx_2' \dots dx_{b-1}' \\
= &\quad \int q^1(x_1'|x')\pi(x_1')q^2(x_2'|x_1') \dots q^b(x'|x_{b-1}')dx_1'dx_2' \dots dx_{b-1}' \\
\vdots & \\
= &\quad \int q^1(x_1'|x')q^2(x_1'|x_2') \dots \pi(x_{b-1}')q^b(x'|x_{b-1}')dx_1'dx_2' \dots dx_{b-1}' \\
= &\quad \pi(x') \int q^1(x'|x_1')q^2(x_1'|x_2') \dots q^n(x_{b-1}'|x')dx_1'dx_2' \dots dx_{b-1}' \\
= &\quad \pi(x')q_1(x|x')
\end{aligned}$$

Hence

$$\alpha_{0,1} = 1$$

We get the same result for $\alpha_{1,0}$. By using the opposite reverse acceptance probability for the proposal we have developed a Metropolis-Hastings algorithm with acceptance probability 1 from the overlapping blocks Gibbs sampler, see algorithm 7. We observe that the only

Algorithm 7 Metropolis-Hasting overlapping blocks Gibbs sampler, opposite reverse

- Given x^0
- for $j = 0 : (niter - 1)$
  - Sample i, $P(i = 0) = P(i = 1) = 0.5$
  - Sample $x^j \sim q_i(x|x^{j-1})$.
- Return $(x^0, x^1, \dots, x^{niter})$

calculations we need to do is those involved in the sampling, hence without any extra computational cost. The same acceptance probability could have been achieved using a systematic scan Gibbs sampler, but this would require twice as many calculations.

For the general partial conditional block sampler (as in algorithm 6) we do not achieve acceptance rate one when using opposite reverse acceptance probability.

4 One-block updating scheme Metropolis-Hasting with overlapping block Gibbs proposal

Recall the distribution of our interest is the posterior $\pi(x, \theta|y)$ given by

$$\pi(x, \theta|y) \propto \pi(y|x)\pi(x|\theta)\pi(\theta)$$

As stated earlier, the latent field x and the hyper-parameters θ should be updated simulations to improve mixing. Using the one-block updating scheme in algorithm 2 the ideal proposal for x would be $q(x|x^{old}, \theta^{new}) = \pi(x|\theta^{new})$, but this could be computational too expensive. One scan of the overlapping block Gibbs sampler from chapter 2 seems to be a good alternative, giving algorithm 8. We name this kind of proposals overlapping block Gibbs proposals. This algorithm is set up with opposite reverse acceptance probability as

Algorithm 8 Metropolis-Hastings algorithm with overlapping Gibbs blocks

- Given θ^0 and y^0
- for $j = 0 : (niter - 1)$
  - Sample $\theta^{new} \sim q(\theta|\theta^j)$
  - Sample i: $P(i = 0) = P(i = 1) = 0.5$.
  - Sample from overlapping block Gibbs proposal $x^{new} \sim q(x|x^{old}, \theta^{new})$
  - Calculate acceptance probability

$$\alpha = \min(1, \frac{\pi(y|x^{new})\pi(x^{new}|\theta^{new})\pi(\theta^{new})q(\theta^j|\theta^{new})q_i(x^j|x^{new}, \theta^j)}{\pi(y|x^j)\pi(x^j|\theta^j)\pi(\theta^j)q(\theta^{new}|\theta^j)q_{1-i}(x^{new}|x^j, \theta^{new})})$$

  - Sample $u \sim \text{Unif}(0, 1)$
  - if$(u < \alpha)$
    - $\theta^{j+1} = \theta^{new}$
    - $x^{j+1} = x^{new}$
  - else
    - $\theta^{j+1} = \theta^j$
    - $x^{j+1} = x^j$
- Return $((\theta^1, x^1), (\theta^2, x^2), \ldots, (\theta^n, x^n))$.

introduced in chapter 3. Also a non-randomised proposal with Peskun's acceptance probability will be used. The i sampling step is then omitted, and the acceptance probability is

$$\alpha = \min(1, \frac{\pi(y|x^{new})\pi(x^{new}|\theta^{new})\pi(\theta^{new})q(\theta^j|\theta^{new})q_0(x^j|x^{new},\theta^j)}{\pi(y|x^j)\pi(x^j|\theta^j)\pi(\theta^j)q(\theta^{new}|\theta^j)q_0(x^{new}|x^j,\theta^{new})})$$

The challenge using this algorithm is to evaluate the proposal $q(x^{new}|x^j,\theta^{new})$ (subscript suppressed). It is given by the transition kernel in chapter 2.2 and is generally not known. As demonstrated in chapter 2.3 an exception is the time series version of the overlapping blocks. Before the expressions are set up we introduce some useful notation for the time series version of overlapping blocks: For each block B_i there are (at most) two kind of elements, those sampled over later, the buffer elements, and those finally sampled in this block. We denote the buffer elements β_i and the final sampled ones b_i, so $B_i = \{b_i, \beta_i\}$. We denote the complement of the whole field and a block B_i by $-B_i$. Further this complement is divided into two disjunct parts; those elements in blocks with a lower index are in B_i- while those in a block with higher index are in B_i+. This is formalised below:

$$\begin{aligned}
B_i &= \{b_i, \beta_i\} \\
-B_i &= \{j\} \text{ such that } j \cap B_i = \emptyset \\
B_i- &= \{j\} \text{ such that } j \cap B_i = \emptyset \text{ and } j \in B_k \text{ with } k < i \\
B_i+ &= \{j\} \text{ such that } j \cap B_i = \emptyset \text{ and } j \in B_k \text{ with } k > i
\end{aligned}$$

Note that while β_i and b_i change for q_0 and q_1, $-B_i$, B_i- and B_i+ are fixed. If the setting is as in figure 12 as B_2 is sampled using q_0 $\{B_2-\} = \{1\}$, $\{b_2\} = \{2,3\}$, $\{\beta_2\} = \{4\}$ and $\{B_2+\} = \{5\}$. The transition kernels can be calculated from;

$$q_0(x|x',\theta) = \prod_{i=1}^{n_B} \pi(x_{B_i}|x_{B_i-}, x'_{B_i+},,\theta) = \prod_{i=1}^{n_B} \frac{\pi(x_{B_i}|x_{B_i-}, x'_{B_i+},\theta)}{\pi(x_{\beta_i}|x_{B_i-}, x'_{B_i+}, x_{b_i},\theta)}$$

and

$$q_1(x|x',\theta) = \prod_{i=1}^{n_B} \pi(x_{B_i}|x'_{B_i-}, x_{B_i+},,\theta) = \prod_{i=1}^{n_B} \frac{\pi(x'_{B_i}|x_{B_i-}, x_{B_i+},\theta)}{\pi(x_{\beta_i}|x'_{B_i-}, x_{B_i+}, x_{b_i},\theta)}$$

where n_B is the number of blocks. For each block both $\pi(x^{new}_{B_i}|x^{new}_{B_i-}, x^j_{B_i+}, \theta^{new})$ and $\pi(x^{new}_{\beta_i}|x^{new}_{B_i-}, x^j_{B_i+}, x^{new}_{B_i}, \theta^{new})$ have to be calculated. The calculation of $\pi(x^{new}_{B_i}|x^{new}_{B_i-}, x^j_{B_i+}, \theta^{new})$ is done while sampling without any extra costs. The denominator $\pi(x^{new}_{\beta_i}|x^{new}_{B_i-}, x^j_{B_i+}, x^{new}_{B_i}, \theta^{new})$ has to be calculated from scratch. In the spatial GMRF case this costs $\mathcal{O}(n^{1.5}_{\beta_i})$, where n_{β_i} is the dimension of x_{β_i}.

4.1 Example 6: Toy example with one-block updating scheme

As a first investigation of the one-block updating scheme Metropolis-Hastings algorithm with overlapping block Gibbs proposal for x, we apply it to sample from a known distribution.

We aim to sample from a GMRF on a lattice with expected value β;

$$x \sim N(\beta\mathbf{1}, \tau Q(r))$$

with a 5×5 neighbourhood and with precision matrix $Q(r)$ such that it is a proxy of a GRF with exponential correlation function

$$\rho(x_i, x_j) = \exp(\frac{-3d(i,j)}{r})$$

denoted as described in example 1. The hyper-parameters $\theta = (\beta, \tau, r)$ are given independent priors; $\beta \sim N(0,1)$, $r \sim \text{Unif}(1, 50)$ and $\tau \sim \text{Gamma}(0.25, 0.05)$. The elements of $Q(r)$ approximating a GRF with range r have to be calculated for each r. To decrease the computational cost r is discretised and all the possible proxies $Q(r)$ can be calculated once and in a pilot run. Hence r's prior is here discrete uniform over its discretion. We apply this model for two lattices, one of size 100×10 and one of size 32×32.

To use the one-block updating scheme of algorithm 2 both a proposal for the hyper-parameters θ and the field x are needed. For the hyper-parameters we chose independent random walk proposals; $\beta^{new} \sim N(\beta^{old}, 0.5)$, $\tau^{new} \sim \text{Unif}(\tau^{old}/f, f\tau^{old})$ with $f = 1.5$ and $r^{new} \sim \text{Unif}(r^{old} - \Delta r, r^{old} + \Delta r)$ with $\Delta r = 50$. For the field x we have used different overlapping block Gibbs proposals. For both lattices overlapping block Gibbs proposals with two, four and eight blocks are tested, in the 32×32 case with buffer lengths zero (i.e. traditional block Gibbs sampler), two and four, and in the 100×10 case with buffer lengths zero, two and ten. Remark the buffer of length 10 in the 100×10 case ($n_\beta = 100$) is computationally cheaper then buffer length 4 in the 32×32 case ($n_\beta = 128$). For reference purposes also an exact proposal for x, $q(x|x^{old}, \theta^{new}) = \pi(x|\theta^{new})$, is used. Both opposite reverse acceptance probability (as in algorithm 8) and Peskun's acceptance probability are tested.

A well known problem in MCMC is slow mixing of hyper-parameters, Knorr-Held and Rue (2002). To evaluate the different one-block samplers we have therefore chosen to investigate the Markov chain for the hyper-parameter β.

To see how the number of blocks, buffer lengths and choice of acceptance probability influence the mixing, auto-correlation between samples of β are estimated, see figure 16. In figure 17 and 18 are trace plots, cumulative mean plots and estimated density (adjusted histograms) for β for the two different lattices. The same quantities using the exact proposal for x is plotted in figure 19 and figure 20.

From these plots we first of all observe that correlation between samples is smaller for the opposite reverse acceptance probability, especially with many blocks and small buffers. This is the cases where the acceptance rate is low and the improved mixing seems to come from increased acceptance. As the length of the buffer increases and the number of blocks decreases the difference between the acceptance probability methods becomes invisible. (Buffer length 10 gives the same acceptance rate as using exact proposal for x). For the square lattice the auto-correlation is higher and the mixing poorer then in the rectangular case even for equal buffer lengths. We suspect this is caused by boundary effects; the

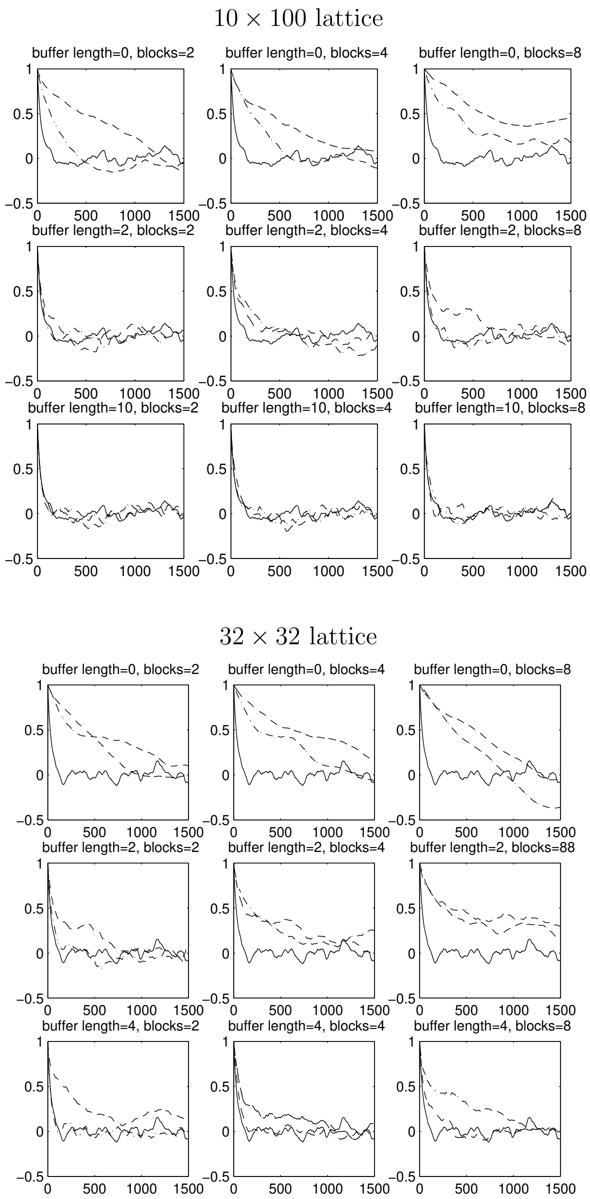


Figure 16: *To example 6: Estimated auto-correlation from samples of* β *for* 2, 4 *and* 8 *blocks with buffers of lengths* 0, 2 *and* 10 *and Peskun's (dashed line) and opposite reverse (dash-dot line) acceptance probability. Estimated auto-correlation for exact proposal for* x *is included in all plots (solid line).*

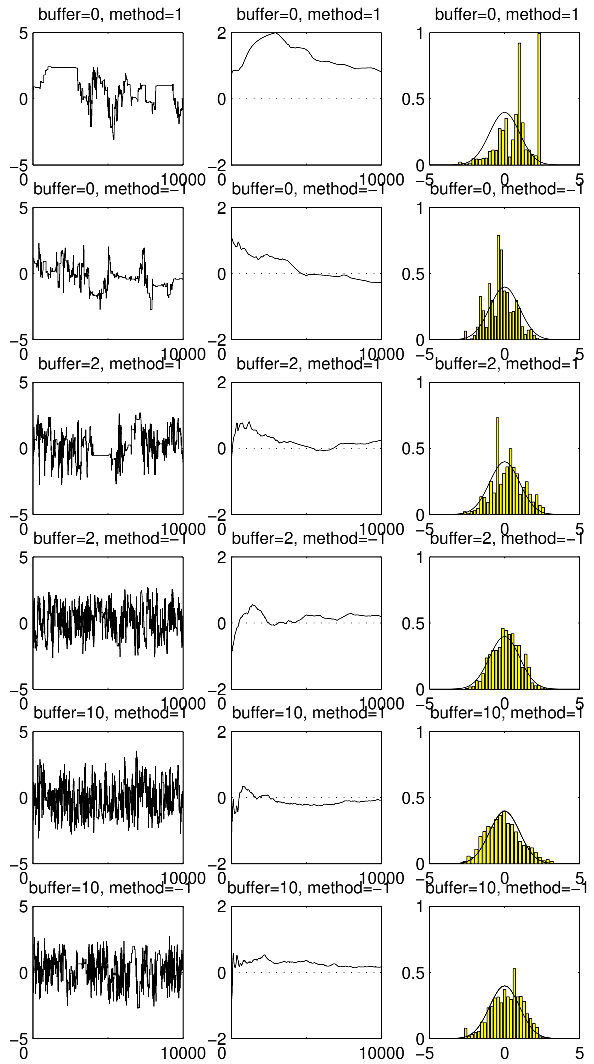


Figure 17: *To example 6: Plots from simulations of β for a 10×100 lattice with buffer lengths 0, 2 and 10 and both acceptance probability alternatives, Peskun's method (1) and opposite reverse method (-1). Left column contains trace plots, middle column cumulated means and right column histograms (adjusted to become densities) and target distribution.*

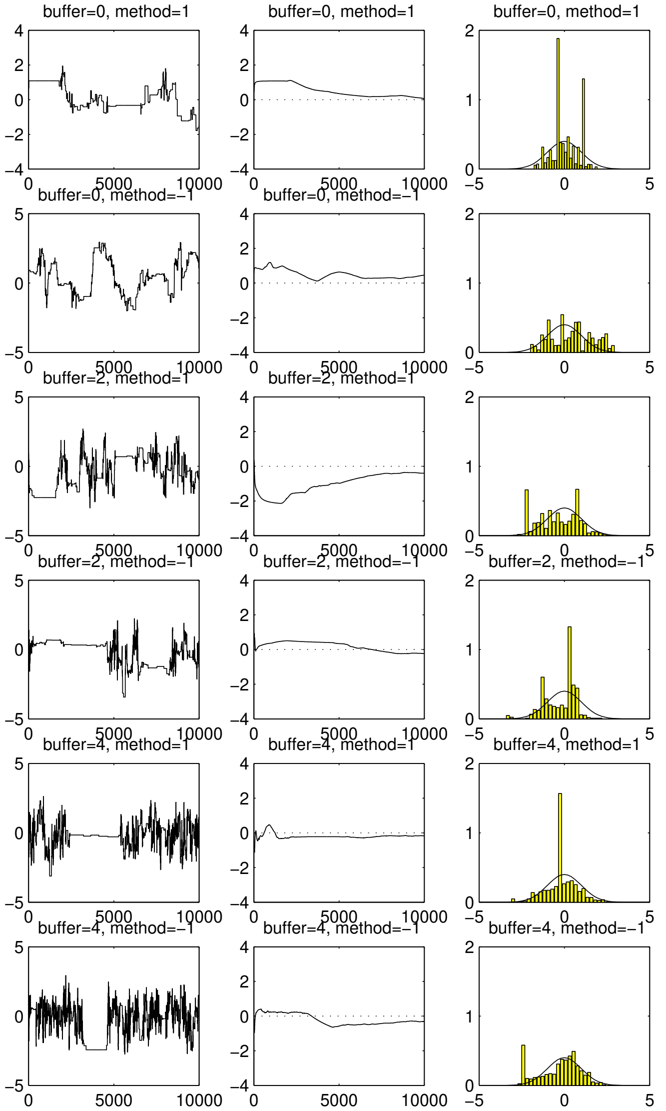


Figure 18: *To example 6: Plots from simulations of β for a 32×32 lattice with buffer lengths* 0, 2 *and* 4 *and both acceptance probability alternatives, Peskun's method (1) and opposite reverse method (-1). Left column contains trace plots, middle column cumulated means and right column histograms (adjusted to become densities) and target distribution.*

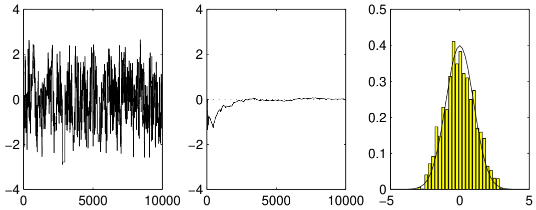


Figure 19: *To example 6: Plots from simulations of* β *for a* 10×100 *lattice using exact proposal for* x*. From left trace plot, cumulated mean and histogram (adjusted to become a density) and target distribution.*

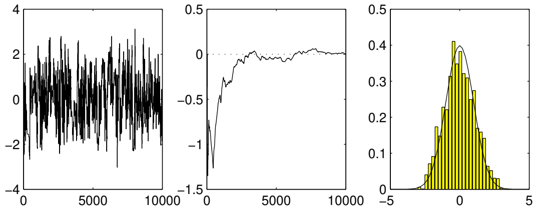


Figure 20: *To example 6: Plots from simulations of* β *for a* 32×32 *lattice using exact proposal for* x*. From left column contains trace plot, middle column cumulated mean and right column histogram (adjusted to become a density) and target distribution.*

GMRF approximation used for the lattice corresponds to condition on a boundary with the expected value. This causes less variance closer to the boundary and less correlation between close to boundary elements.

In figure 17 and 18 we first of all observe that the mixing improves as the buffer lengths increase. Especially is the difference large going from non to a buffer of length two. The convergence of the estimated mean of β and the quality of the density estimate have also improved.

In this example we have used an one-block updating scheme Metropolis-Hasting algorithm with different overlapping block Gibbs samplers as proposals for the field x. The mixing is investigated for a hyper-parameter; the expected value of the field β. For long enough buffers the overlapping block Gibbs proposal gives as good mixing as the exact proposal for x, $\pi(x|\theta)$. The necessary buffer length decreases using opposite reverse acceptance probability.

4.2 Example 7: Graph GMRF model with Gaussian likelihood

In this section overlapping block Gibbs proposals are tested for a GMRF model on an irregular graph. We have used a map of Germany making a graph and neighbourhood structure as described in section 1.1. Our interest is a latent field x and its spatial structure.

It is given a GMRF prior with smoothing parameter κ;

$$\pi(x|\kappa) \propto \kappa^{(n-1)/2} \exp(-\frac{1}{2}\kappa \sum_{i \sim j}(x_i - x_j)^2)$$

The likelihood is consider mutually independent Gaussian:

$$\pi(y_i|x) \sim N(c_i x_i, \tau^{-1})$$

where τ is the precision and c_i is a region specific constant. As data we have used the German oral cavity cancer dataset. The Gaussian likelihood is not appropriate for the data and the study in this example should only be interpreted as a study of the samplers used. Both hyper-parameters are given vague Gamma-priors.

As in example 6 we use an one-block updating scheme with a proposal of two stages. For the hyper-parameters $\theta = (\kappa, \tau)$ we use independent log random walk proposals. For the latent field x we aim to use an overlapping block Gibbs proposal, though how to set up the blocks is not trivial for an irregular graph.

The method for sampling GMRFs described in Rue (2001) reorders the elements of x such that the bandwidth of the precession matrix b_ω is small. To this new ordering there correspond an auto-regressive time series with conditional dependents length b_ω. In the overlapping block Gibbs proposal blocks (and buffers) are set up in this reordered world. If buffers are of length b_ω we are guaranteed that all (not earlier sampled) neighbours are included in the buffer. The bandwidth of the reordered graph of Germany is 44.

The sampler was run for 10000 iterations for different numbers of blocks (1,2,4 and 8), buffer lengths 10, 22, 44 and 66, and with the two different acceptance probabilities. For both hyper-parameters scaled uniform proposals ($\tau^{new} \sim \text{Unif}(\frac{\tau^{old}}{f}, f\tau^{old})$ with $f = 1.3$) are used. The convergence properties are tested for the smoothing parameter κ and for region 29. This is region 200 in the reordered indexes, and hence about $1.5 b_\omega$ from the block border for two blocks.

The acceptance rate is a measure for how close the proposal is the target distribution for a Metropolised independence sampler with Peskun's acceptance probability. As discussed in example 5 and in chapter 3 it could be used as a measure of reversibility and independence of the sampled fields x for an overlapping block Gibbs sampler with fixed hyper-parameters and Peskun's acceptance probability. (Recall that with fixed hyper-parameters opposite reversing acceptance probability gives acceptance rate 1.) In our algorithm only the field part x aim to be independent and we compare the acceptance rate for overlapping block Gibbs proposals for x with exact proposal for x using Peskun's acceptance probability. We use this difference as a measure of how fare $q(x|x^{old}, \theta^{new})$ is from $\pi(x|\theta^{new})$. From figure 21 we see that the decrease in acceptance rate disappears as the buffer length increase. For buffer lengths 44 and 66 are there no notable difference. The mixing and convergence for both hyper-parameters and the latent field are good in this example, see figure 22 for the cumulative mean estimate for some of the samplers. To investigate the mixing further the auto-correlation within the Markov Chain is estimated

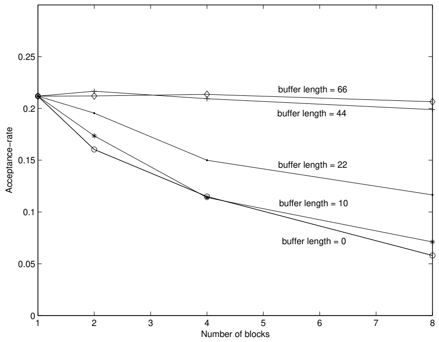


Figure 21: *To example 7: Acceptance-rate for samplers as a function of number of blocks for buffers of length 0 (stars), 10 (circles), 22(dots), 44(plus sign) and 66 (diamonds).*

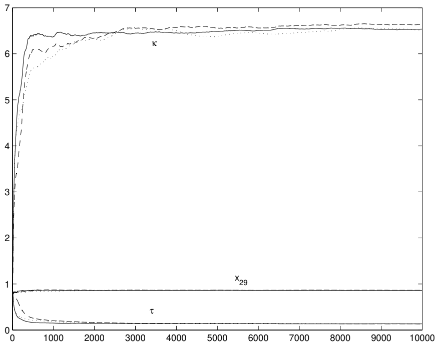


Figure 22: *To example 7. Cumulated mean for τ, κ and x_{29} with the one-block x proposal (solid line) and overlapping block Gibbs proposal for x with 8 blocks without buffers (dotted line) and with buffer length 44 (dashed line). All samplers used Peskun's acceptance probability.*

for κ and x_{29} from samplers with overlapping block Gibbs proposals with 8 blocks, see figure 23 and 24. The auto-correlation is quite good for both variables, even without buffers and with Peskun's acceptance probability. We observe the same trends as earlier. The Markov chain of the hyper-parameter (here κ) has higher autocorrelation then the field variables (here investigated for x_{29}). The opposite reverse acceptance probability gives smaller estimated auto-correlation then Peskun's acceptance probability. For both acceptance probabilities the auto-correlation decreases as the buffer length increases. In this example the auto-correlation is almost equal for the exact proposal for x and overlapping block Gibbs proposal when the buffer length equals the bandwidth.

The simulations in this example suggest that overlapping block Gibbs proposal with

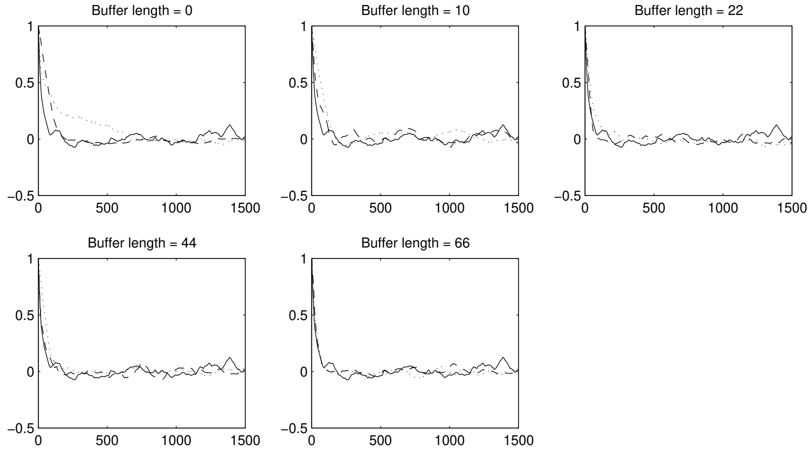


Figure 23: *To example 7: Estimated auto-correlation for κ for different buffer lengths and for Peskun's (dotted line) and opposite (dashed line) reverse acceptance probabilities. Estimated autocorrelation for a exact sampler for $x|\theta^{new}$ included (solid line).*

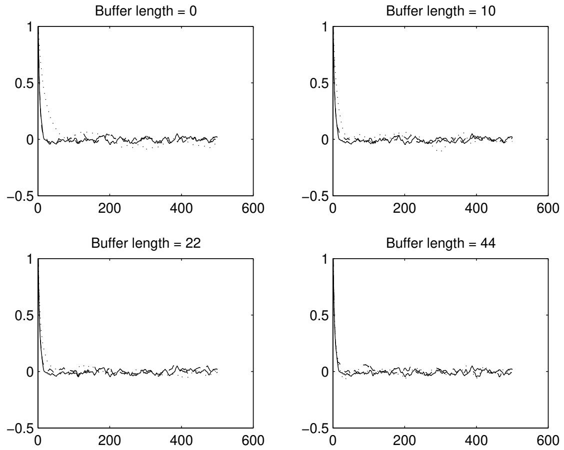


Figure 24: *To example 7: Estimated auto-correlation for x_{29} for different buffer lengths and for Peskun's (dotted line) and opposite reverse (dashed line) acceptance probabilities.*

buffers of length the bandwidth gives samples almost with the same mixing and converting quality as using $\pi(x|\theta^{new})$ as proposal for x. When the buffer length is small, the opposite reverse acceptance probability seems to improve the mixing for the hyper-parameters.

4.3 Example 8: Space-time GMRF model with Gaussian likelihood

In this section we present a space-time GMRF model and sample from it applying an one-block updating scheme Metropolis-Hasting algorithm with an overlapping block Gibbs proposal for the latent field. The spatial graph from example 7 is extended to a space-time graph by making a copy of the graph for each of the T time steps and making the graphs a chain by connecting each node x_{it} with the corresponding node for the time step immediate before ($x_{i,t-1}$) and after ($x_{i,t+1}$). We believe the smoothing is different in time and space, and model it with two parameters, τ_S for space and τ_T for time;

$$\pi(x|\kappa,\tau_T) \propto \prod_{t=1}^{T} \exp(-\frac{1}{2}\tau_S \sum_{i\sim j}(x_{it}-x_{jt})^2) \prod_{i=1}^{N} \exp(-\frac{1}{2}\tau_T \sum_{s\sim t}(x_{is}-x_{it})^2)$$

This corresponds to a multivariate $N \cdot T$ dimensional Gaussian distribution with precision matrix Q:

$$Q_{ij} = \begin{cases} -\tau_S, & \text{if } i \overset{s}{\sim} j \\ -\tau_T, & \text{if } i \overset{t}{\sim} j \\ \kappa \mathrm{nnb}_s(i) + \tau_T \mathrm{nnb}_t(i), & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

where $\overset{s}{\sim}$ denotes neighbours in space and $\overset{t}{\sim}$ in time, $\mathrm{nnb}_s(i)$ is the number of neighbours for element i in space and $\mathrm{nnb}_t(i)$ in time. Also note that $Q = \tau_S Q_S + \tau_T Q_T$ where Q_T is the precision matrix for the time dependence and Q_S for the spatial dependence with $\tau_S = 1$ and $\tau_T = 1$. This prior is intrinsic and we need to know how the normalisation constant depends on τ_T and τ_S. In appendix A.2 the normalisation constant is found as a function of the non-zero eigenvalues of the precision matrix for one region in T time step, and for the N regions for one time step. As in example 7 the likelihood is consider mutually independent Gaussian;

$$\pi(y_i|x) \sim N(c_i x_i, \tau^{-1})$$

Synthetic data were made for this model on the Germany graph with 100 time steps and with parameters $\tau_T = 5.0$, $\tau_S = 5.0$ and $\tau = 20.0$. The latent field has dimension 54400 and we needed 250 seconds to get an exact sample from $\pi(x|\theta^{new})$ while one scan of an overlapping block Gibbs sampler with 20 blocks and a buffer length of 544 variables (i.e. one time step in the space-time graph) needed about 7 seconds. We used this overlapping block Gibbs proposal for x together with log random walk proposals for the hyper-parameters and opposite reverse acceptance probability in an one-block updating scheme Metropolis-Hastings algorithm. Results for the hyper-parameters are plotted in figure 25 (trace plots) and in figure 26 (histograms and cumulative mean). The sampler was run for 25000 iterations and had an acceptance rate of 0.28. The precision parameters of the GMRF seems to have stabilised close to their original values ($\overline{\tau_S} = 5.007$ and $\overline{\tau_T} = 4.94$). The

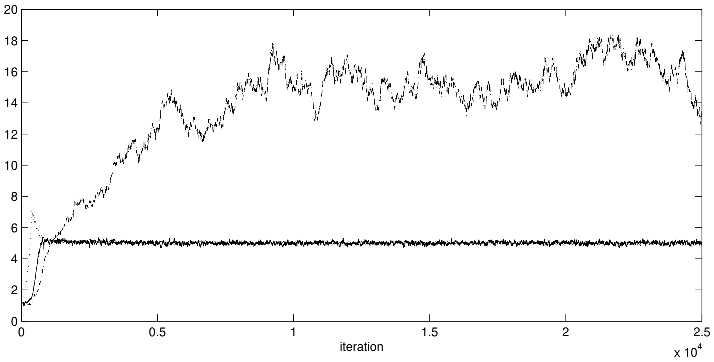


Figure 25: *To example 8: Trace plots for τ_S (dotted line), τ_T (solid line) and τ (dash-dot line).*

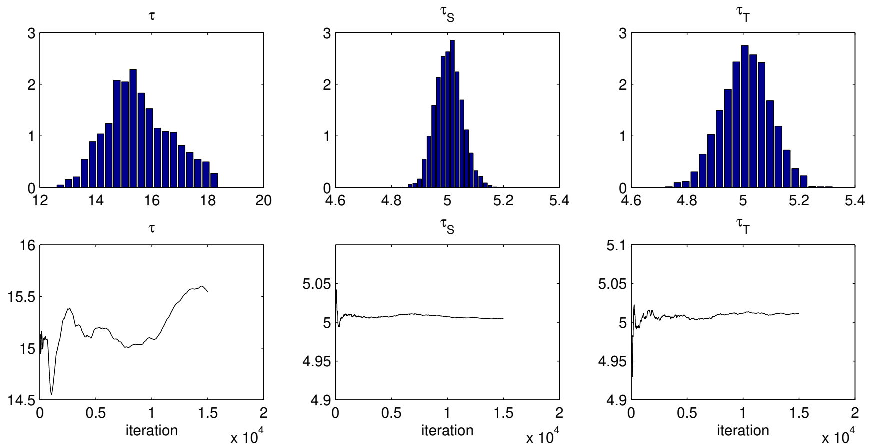


Figure 26: *To example 8: Histograms (adjusted to become densities) and cumulative mean for τ, τ_S and τ_T with a burn-in of 10000 iterations omitted.*

precision of the data, τ, has poorer mixing and the cumulated mean has not stabilised. Simulations from other initial values gave the same level of τ, which indicates that the Markov chains have converged.

In the space-time problem constructed in this example an one-block updating scheme for (x, θ) with exact proposal for $x|\theta^{new}$ is computationally infeasible. Using an overlapping block Gibbs proposal for x support us with appropriate samples and is computationally affordable. It enables us to use an one-block updating scheme.

4.4 Example 9: Functional magnetic resonance imaging

In this example we use functional magnetic resonance imaging (fMRI) data previously studied in Göss et al. (2000). The data are from a visual stimulation experiment. The stimulus was a 8 Hz flickering checkerboard, and the experiment lasted for 210 seconds with four periods (a 30 seconds) rest and three periods stimulus. A crossection of the brain (128×128 pixels) was observed every third second, hence a time series of 70 time steps. Images of size 75×67 is enough to cover the brain and we based our analyses on these images. See figure 27 for images of the mean and standard deviation in time for the data, the first 20 images can be seen in appendix A.3. Functional magnetic resonance utilise

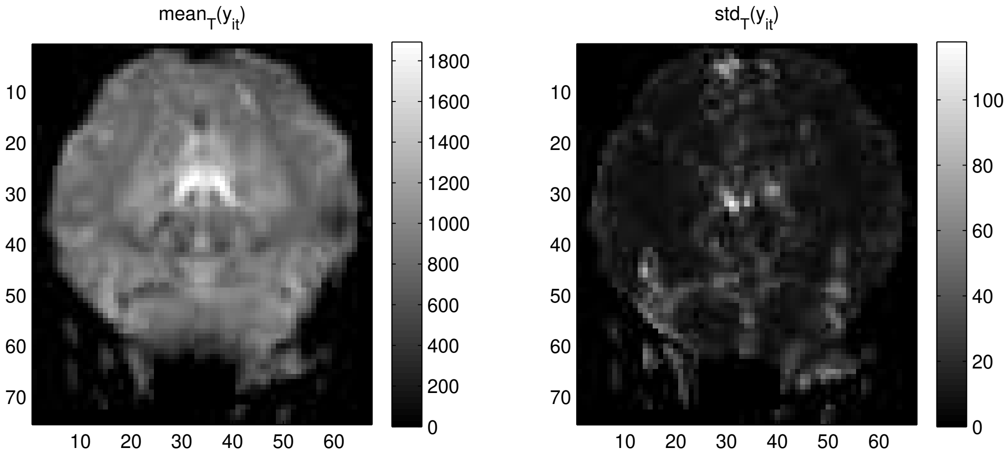


Figure 27: *To example 9: The mean (left) and standard deviation in time for the data.*

the different magnetic properties of oxygenated and disoxygenated blood and is useful for observing BOLD (blood oxygenation level dependent) 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 will now make a space-time GMRF model for this problem with smaller dimension then the space-time model in Göss et al. (2001). To estimate parameters an one-block updating scheme Metropolis-Hastings algorithm with an overlapping block Gibbs proposal is used.

We model the observations of pixel i at time t as:

$$y_{it} = a_i + z_t b_{it} + \epsilon_{it}$$

for $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$. Here a is the baseline image (of size N) and b_{it} is the activation effect of pixel i at time t (of size $N \cdot T$). See figure 28 for the directed acyclic graph of the model. A transformed stimulus, z_t $t = 1, 2, \ldots, T$ of the original stimulus is

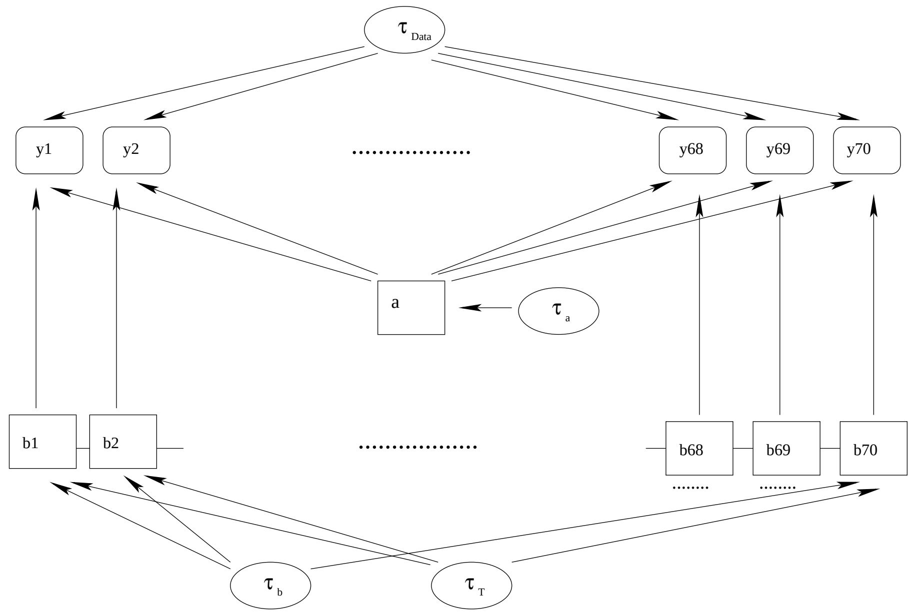


Figure 28: *To example 9: The DAG (directed acyclic graph) of our fMRI model.*

used. A common choice is to use a temporal shift of the original stimulus 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 $\epsilon_{it} \sim N(0, \tau_{Data}^{-1})$ with common precision τ_{Data}. The tempo-spatial modelling is done through the priors of a and b. They are both given intrinsic Gaussian priors:

$$\begin{aligned}
\pi(a) &\propto \exp(-\frac{1}{2}\tau_A \sum_{t=1}^{T} \sum_{i \overset{s}{\sim} j} (a_i - a_j)^2) \\
\pi(b) &\propto \exp(-\frac{1}{2}\tau_B \sum_{t=1}^{T} \sum_{i \overset{s}{\sim} j} (b_{it} - b_{jt})^2) \\
& \quad \exp(-\frac{1}{2}\tau_T \sum_{i=1}^{N} \sum_{t \overset{t}{\sim} r} (b_{it} - b_{ir})^2)
\end{aligned}$$

Each non-boarder pixel has four neighbours in space (both in a and b) and two in time (for b only), see figure 29. The priors for τ_A, τ_B and τ_T were all set to be independent Gamma

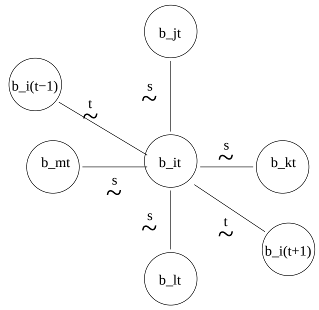


Figure 29: *To example 9: Illustration of the space-time neighbourhood structure for b.*

distributed with expected value 5.0 and variance 100.

The posterior distribution of the hyper-parameters $\theta = (\tau_{Data}, \tau_A, \tau_B, \tau_T)$, the baseline image a and the activation effect b is our distribution of interest. In appendix A.4 $\pi(a, b|\theta, y)$ is found to be multivariate Gaussian with a conditional dependence structure as illustrated with the graph in figure 30. This distribution is only proper if $z_i > 0 \; \forall i$. Our pragmatic solution to this problem is to use $z_i + 1$ instead of z_i. This causes some of the "baseline level" to be moved from a to b. Sampling is conceptually done as in the previous examples: Independent random walk proposals for the hyper-parameters followed by an overlapping block Gibbs proposal for $\pi(a, b|\theta, y)$. All variables are accepted/ rejected in one block.

To decrease the problem size we have used only the mid section of the brain. This gives us a dataset of 75×21 pixels $\times 70$ time steps, or 110250 data points, and $\{a, b\}$ has dimension 111825. We are not be able to sample exact from a GMRF of this size.

The blocks are chosen as illustrated in figure 30. Because of the dependence structure a is sampled in every block together with some time-steps of b.

The precision for the data, τ_{Data}, was estimated from the part of the image not used, and was fixed to this value ($\tau_{Data} = 0.003$). We used an updating scheme with blocks consisting

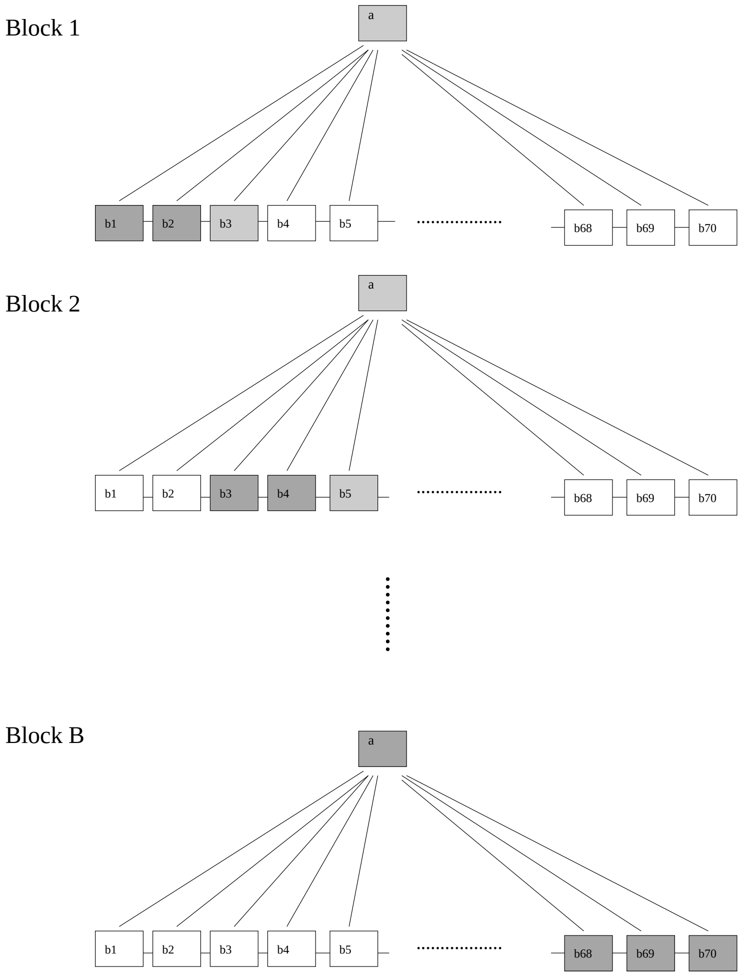


Figure 30: *To example 9: The overlapping blocks used when sampling from a and b. The images with a brighter shade will be sampled over later.*

of a and five b images, and with two b images and a overlap. The hyper-parameters were proposed independently of each other uniformly on $[\frac{1}{f}\tau^{old}, f\tau^{old}]$. The algorithm was run for 20000 iterations with initial values for the hyper-parameters $\tau_A = 1.0$ $\tau_B = 5.0$ and $\tau_T = 1.0$. Trace plots with cumulative mean can be found in figure 31. We observe quite low spatial dependence ($\overline{\tau_A} = 0.000009$ and $\overline{\tau_B} = 0.000056$), though reasonable values considering the values of the data. The precision in time is much higher ($\overline{\tau_T} = 0.26$). The mixing of the smoothing parameters of b is quite slow, but the cumulated mean has

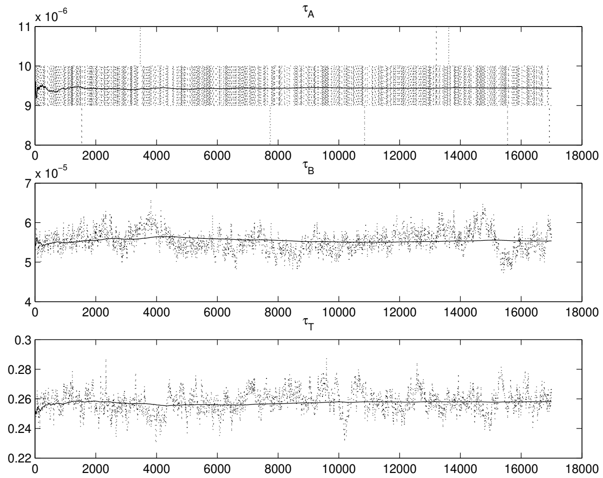


Figure 31: *To example 9: Trace plots (doted line) and cumulative mean (solid line) with burn in of* 3000 *for* τ_A, τ_B *and* τ_T.

stabilised. Figure 32 contains images of the mean estimate of the baseline image and the activation for the time steps. The baseline estimate seems to be a smoother version of the mean in time image in figure 27. The activation estimate images are from two stimulus time steps ($t = 18$ and $t = 38$) and one rest time step ($t = 28$). From these images we see that the activation areas are in the upper part of the brain. This agrees with previous studies of the same data.

Most of the smoothing is done in the time direction. To illustrate this we have in figure 33 plotted the data and our mean estimates 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.

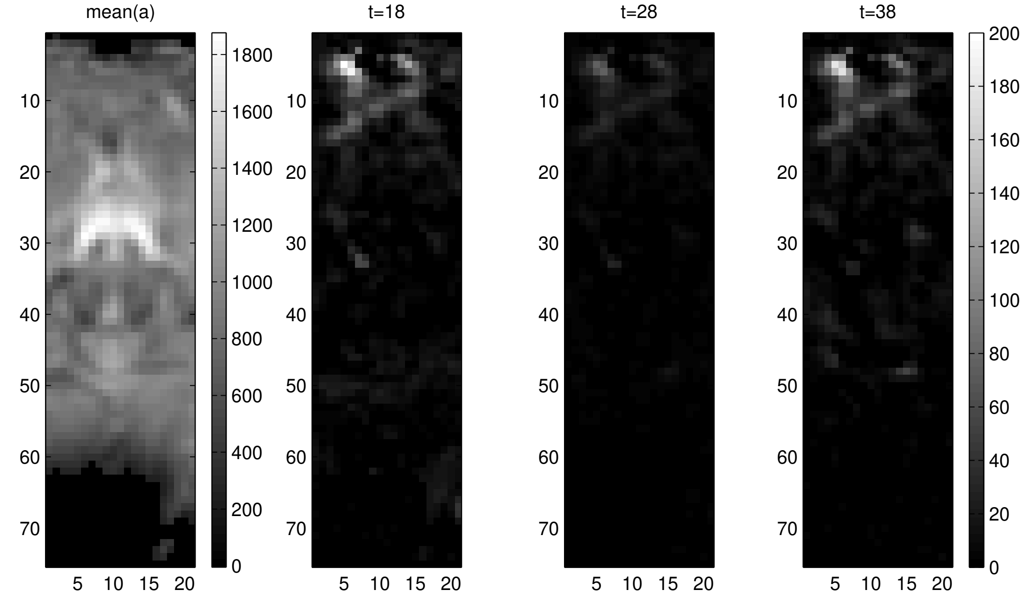


Figure 32: *To example 9: The mean estimate of baseline image(left), and the mean estimated activation effect* $z_t b_i$ *for time step 18, 28 and 38.*

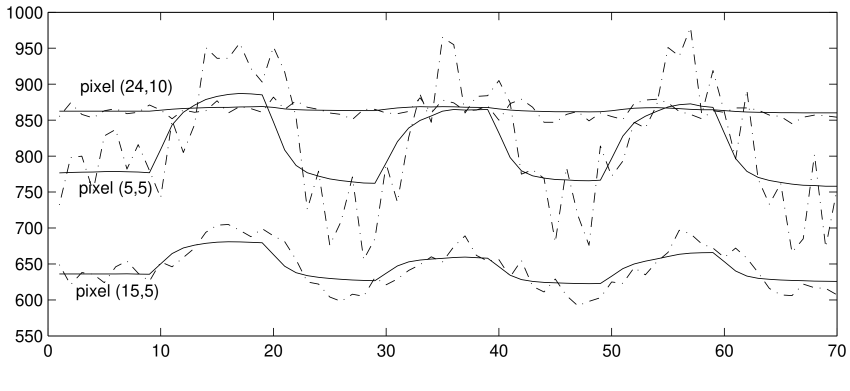


Figure 33: *To example 9: The observed values and the mean estimates for three pixels.*

5 Overlapping approximated blocks proposals

Until now we have only analysed models with posteriors $\pi(x|y,\theta)$ for which we are able to sample from $\pi(x_B|x_{-B},y,\theta)$ exact. In this chapter we demonstrate that the overlapping block approach for constructing a proposal is fruitful even when we are unable to sample exact from $\pi(x_B|x_{-B},y,\theta)$. We construct approximations to $\pi(x_B|x_{-B},y,\theta)$ and use sample from these approximations instead of sampling from the exact distribution of the blocks.

5.1 Constructing a proposal for the latent field from approximated blocks

The models we consider have a latent field x with a GMRF-prior, $x|\theta \sim GMRF$. The observations are conditioned on x mutually independent; $\pi(y|x,\theta) = \prod_{i=1}^{n} \pi(y_i|x_i)$. In the overlapping block setting the distribution we want to sample from is $\pi(x_B|x_{-B},y,\theta)$. For our models we get:

$$\pi(x_B|x_{-B},y,\theta) = \pi(x_B|x_{-B},y_B,\theta) \propto \pi(x_B|x_{-B},\theta)\pi(y_B|x_B)$$

If $\pi(x|\theta)$ is a GMRF is also $\pi(x_B|x_{-B},\theta)$ a GMRF and the conditional independence structure of $\pi(x_B|x_{-B},y,\theta)$ is the same as for $\pi(x|y,\theta)$. In Rue et al. (2003) these kind of densities for $\pi(x|y,\theta)$ are named hidden GMRFs (HGMRFs) and a class of approximations to non-Gaussian HGMRFs is introduced. The approximations are done for full dimensional x, but as we have seen our conditional problem has the same structure and we want to use the approximations for these blocks. The approximations in Rue et al. (2003) are done in three stages: The first stage is to find a Gaussian approximation in the mode of $\pi(x|y,\theta)$. Denote this approximation $\pi^{A1}(x|y,\theta)$. As a first improvement non-Gaussian corrections for the likelihood term for the corresponding data point (correct for y_i when x_i is sampled) are done. We denote this approximation π^{A2}. The third approximation, π^{A3}, also correct for non-Gaussian likelihood terms from other locations through sampling.

We denote approximation Ai to $\pi(x_{B_i}|x_{-B_i},y,\theta)$ $\pi^{Ai}(x;x_{-B_i},y,\theta)$, and suppress the method number i in expressions valid for all methods. For notational convenience y and θ are suppressed when assumed fixed. For a time series version of overlapping as in figure 12 we make a proposal for the latent field from overlapping approximated blocks, see algorithm 9. The corresponding transition density can be written as a product of marginal approximated distributions;

$$q^A(x|x') = \pi^A(x_1;x'_3,x'_4,x'_5)\pi^A(x_2,x_3;x_1,x'_5)\pi^A(x_4,x_5;x_1,x_2,x_3)$$

Remark that $\pi^A(x_1;x'_3,x'_4,x'_5)$ is the marginal distribution of x_1 of the approximation; $\pi^A(x_1;x'_3,x'_4,x'_5) = \int \pi^A(x_1,x_2;x'_3,x'_4,x'_5)dx_2$. In a time series version of overlapping blocks with n_b blocks we get;

$$q^A(x|x') = \pi^A(x_{b_1};x'_{B_1+})\pi^A(x_{b_2};x_{B_2-},x'_{B_2+})\dots\pi^A(x_{b_{nb}};x_{B_{nb}-})$$

Algorithm 9 OVERLAPPING APPROXIMATED BLOCKS PROPOSAL

- Given $x^{'}$
- Sample $(x_1, x_2^{B1}) \sim \pi^A(x_{B_1}; x_3^{'}, x_4^{'}, x_5^{'})$
- Sample $(x_2, x_3, x_4^{B2}) \sim \pi^A(x_{B_2}; x_1, x_4^{'})$
- Sample $(x_4, x_5) \sim \pi^A(x_{B_3}; x_1, x_2, x_3)$
- Return $(x_1, x_2, \ldots, x_5)$

Algorithm 10 METROPOLIS-HASTINGS ALGORITHM, SAMPLING FROM x WITH OVERLAPPING APPROXIMATED BLOCKS PROPOSAL

- Given x^0
- for $j = 0 : (niter - 1)$
  - Sample from overlapping approximated blocks proposal $x^{new} \sim q^A(x|x^{old})$.
  - Calculate
  $$\alpha = \min(1, \frac{\pi(x|x^{new})\pi(x^{new})q^A(x^j|x^{new})}{\pi(x|x^j)\pi(x^j)q^A(x^{new}|x^j)})$$
  - $u \sim \text{Unif}(0, 1)$
  - if$(u < \alpha)$
    - $x^{j+1} = x^{new}$
  - else
    - $x^{j+1} = x^j$
- Return $(x^1, x^2, \ldots, x^{niter})$.

with notation as in section 4. Since each block is sampled from an approximated density we need to include a Metropolis-Hasting step even when hyper-parameters are fixed, see algorithm 10.

This sampler requires evaluation of the transition density $q^A(x|x')$. Below we describe ways of doing this for approximation $A1$, $A2$ and $A3$.

Evaluation of $q^A(x|x')$ when using approximation $A1$

The first approximation, $\pi^{A1}(x_B; x_{-B})$, is made by finding the mode of $\pi(x_B|x_{-B})$ and a Gaussian approximation in the mode. Hence is $\pi^{A1}(x_B; x_{-B})$ Gaussian and for our mutually independent likelihood it is a GMRF with the same precision structure as the prior of x_B. The marginal distribution can therefore be calculated by

$$\pi^{A1}(x_b; x_{-B}) = \frac{\pi^{A1}(x_b, x_\beta^B; x_{-B})}{\pi^{A1}(x_\beta^B|x_b; x_{-B})}$$

for any x_β^{B1}, e.g. the sampled one. Note that $\pi^{A1}(x_\beta^B|x_B; x_{-B})$ is the conditional distribution of x_β^B given x_b of the Gaussian approximation to $x_B = \{x_b, x_\beta\}$ given x_{-B}. Evaluation of $q^{A1}(x|x')$ is therefore very similar to evaluation of $q(x|x')$;

$$q^{A1}(x^{new}|x^j) = \prod_{i=1}^{n_B} \pi^{A1}(x_{b_i}^{new}; x_{B_i-}^{new}, x_{B_i+}^j) = \prod_{i=1}^{n_B} \frac{\pi^{A1}(x_{b_i}^{new}, x_{\beta_i}^{new}; x_{B_i-}^{new}, x_{B_i+}^j)}{\pi^{A1}(x_{\beta_i}^{new}|x_{b_i}^{new}; x_{B_i-}^{new}, x_{B_i+}^j)}$$

Still $\pi^{A1}(x_{b_i}^{new}, x_{\beta_i}^{new}; x_{B_i-}^{new}, x_{B_i+}^j)$ is evaluated while sampling without any extra cost while $\pi^{A1}(x_{\beta_i}^{new}|x_{b_i}^{new}; x_{B_i-}^{new}, x_{B_i+}^j)$ has to be evaluated from scratch. Since the approximation is a GMRF this costs $\mathcal{O}(n_\beta^{3/2})$ where n_β is the dimension of x_β.

Evaluation of $q^A(x|x')$ when using approximation $A2$ or $A3$

The approximations $A2$ and $A3$ are refined versions of $A1$ and the refinements are done sequential. Assume we want to sample $x = (x_1, x_2, \ldots, x_k)$ and $x \sim \pi(x)$. The sampling is inspired of the equality;

$$\pi(x) = \pi(x_k)\pi(x_{k-1}|x_k)\ldots\pi(x_1|x_2, x_3, \ldots, x_k)$$

The refined approximations are done in this sequential way and can for $A2$ (and $A3$) be written as

$$\pi^{A2}(x) = \pi^{A2}(x_k)\pi^{A2}(x_{k-1}|x_k)\ldots\pi^{A2}(x_1|x_2, x_3, \ldots, x_k)$$

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. Since the approximation for $\pi(x_i|x_{i+1}, \ldots, x_k)$ is only done for the sampled $(x_{i+1}, x_{i+2}, \ldots, x_k)$ evaluations of $\pi^{A2}(x)$ and $\pi^{A3}(x)$ is not straight forward and generally not easily obtained. We are able to calculate the density for our sample, $\pi^{A2}(x_B; x_{-B})$ or $\pi^{A3}(x_B; x_{-B})$, while sampling without any extra cost. But it is generally not achievable to evaluate $\pi^{A2}(x_\beta|x_b;x_{-B})$ (or $\pi^{A3}(x_\beta|x_b;x_{-B})$) and we need an other way of evaluating $\pi^{A2}(x_b|x_{-B})$ and $\pi^{A3}(x_b|x_{-B})$.

If the sampling of 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 $\pi^{A2}(x)$. If the ordering within each block is done such that the buffer elements have the lowest indexes, $x_B=(x_\beta,x_b)$, we can both sample from and evaluate $\pi^{A2}(x_b;x_{-B})$ simply by stopping the sampling process when x_b is sampled.

Indeed, we also sample from multivariate Gaussian distributions in this sequential way. And we could also previously sample directly from the partial conditional distributions by reordering. But there is a computational cost of restricting the ordering. Since we do a time series version of blocking restricting the indexes of the buffer x_β to be indexed first forces us to use a bandwidth kind of ordering. If we let the block B be of size k the sampling cost of a k-dimensional spatial GMRF with bandwidth ordering is $\mathcal{O}(k^2)$ while a general optimal ordering would cost $\mathcal{O}(k^{3/2})$. On the other hand we do not have to sample x_β or evaluate $\pi^A(x_\beta|x_b;x_{-B})$, but the cost of this is only $\mathcal{O}(n_\beta^{3/2})$ with x_β of size n_β.

5.2 Approximated blocks and changing hyper-parameters

Incorporating the latent field sampler in algorithm 10 into an one-block Metropolis-Hasting algorithm together with hyper-parameters is now trivial. We still want to use opposite reverse acceptance probability and introduce the two opposite direction transition kernels for x;

$$q_0^A(x|x')=\pi^A(x_{b_1};x'_{B_1+})\pi^A(x_{b_2};x_{B_2-},x'_{B_2+})\dots\pi^A(x_{b_{nb}};x_{B_{nb}-})$$

and

$$q_1^A(x|x')=\pi^A(x_{b_{nb}};x'_{B_{nb}-})\pi^A(x_{b_{(nb-1)}};x'_{B_{(nb-1)}-},x_{B_{(nb-1)}+})\dots\pi^A(x_{b_1};x_{B_1+})$$

with notation as in section 4. The one-block Metropolis-Hastings sampler with hyper-parameters is in algorithm 11. The proposals $q_0^A(x|x')$ and $q_1^A(x|x')$ are sampled from and evaluated as described in section 5.1.

5.3 Example 10: Disease mapping in Germany

To explore how an overlapping approximated blocks proposal works we consider the German oral cavity cancer dataset described and modelled in section 1.1. We use a GMRF prior for the log relative risks and a Poisson likelihood;

$$\pi(y_i|x_i)\sim P_o(c_i\exp(x_i))$$

We are not able to sample exact from $\pi(x|y,\kappa)$ nor from $\pi(x_B|x_{-B,}y,\kappa)$ for a block x_B. The same data and model were analysed in Rue et al. (2003) using full dimensional approximations of $\pi(x|y,\kappa)$ as proposals. We now want to use overlapping approximated

Algorithm 11 ONE-BLOCK METROPOLIS-HASTINGS ALGORITHM WITH OVERLAPPING APPROXIMATED BLOCKS PROPOSAL

- Given θ^0 and y^0
- for $j = 0 : (niter - 1)$
  - Sample $\theta^{new} \sim q(\theta^{new}|\theta^j)$
  - Sample i: $P(i = 0) = P(i = 1) = 0.5$.
  - Sample from overlapping approximated blocks proposal $x^{new} \sim q_i^A(x|x^{old}, \theta^{new})$.
  - Calculate

$$\alpha = \min(1, \frac{\pi(y|x^{new})\pi(x^{new}|\theta^{new})\pi(\theta^{new})q(\theta^j|\theta^{new})q_i^A(x^j|x^{new}, \theta^j)}{\pi(y|x^j)\pi(x^j|\theta^j)\pi(\theta^j)q(\theta^{new}|\theta^j)q_{1-i}^A(x^{new}|x^j, \theta^{new})})$$

  - $u \sim \text{Unif}(0, 1)$
  - if$(u < \alpha)$
    - $\theta^{j+1} = \theta^{new}$
    - $x^{j+1} = x^{new}$
  - else
    - $\theta^{j+1} = \theta^j$
    - $x^{j+1} = x^j$
- Return $\theta = (\theta^1, \theta^2, \ldots, \theta^n)$ and $x = (x^1, x^2, \ldots, x^{niter})$.

blocks proposals. To set up the blocks we use the same bandwidth ordering as described in section 4.2.

Fixed hyper-parameter and approximation $A1$

First the Gaussian approximation ($A1$) was tested for different fixed hyper-parameter; $\kappa = 1.0$, $\kappa = 10.0$ and $\kappa = 25.0$. This was done for 2, 4 and 8 blocks with buffer lengths 0, 22, 44 and 66 (the bandwidth of the graph is 44) using both Peskun's and opposite reverse acceptance probability. For reference purposes also a sampler with full dimensional $A1$ approximation proposal for x was run for each tested value of κ.

Later in this example we find that $\kappa = 1.0$ is a small value for the posterior of $\kappa|y$, $\kappa = 10.0$ is close to its mode and $\kappa = 25.0$ is a large value. In figure 34 the acceptance rates from the samplers with Peskun's acceptance probability are plotted. All samplers were run for 10000 iterations. With $\kappa = 1.0$ we get a posterior far from Gaussian but with small spatial dependence. This causes low acceptance rate for $A1$ also for a full dimensional approximation. Because the elements of x are almost independent blocking does not change the acceptance rate significantly and hence buffer lengths either. For the

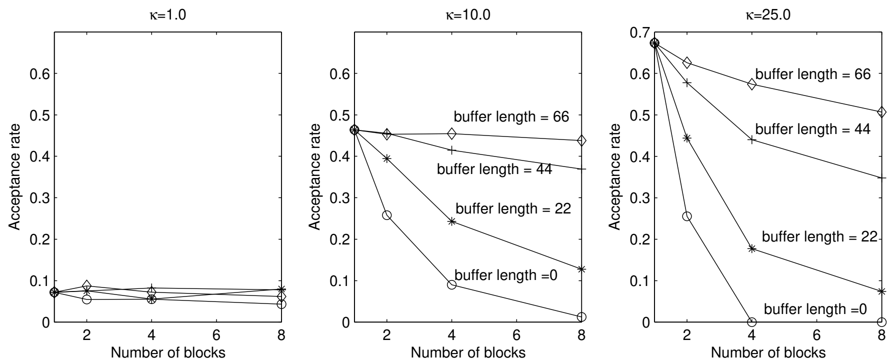


Figure 34: *To example 10,* $A1$ *and fixed hyper-parameters: Acceptance rates for different number of blocks with buffers of length 0 (stars), 22(stars), 44(plus signs) and 66 (diamonds) using Peskun's acceptance probability.*

realistic smoothing, $\kappa = 10.0$, the posterior is closer to Gaussian and has more spatial dependence. The full dimensional approximation is closer to the posterior and gives much higher acceptance rate. Because of the spatial dependence blocking without buffers causes low acceptance and increasing buffer lengths increases the acceptance rate. The overlapping approximated blocks proposal with buffer length 1.5 times the bandwidth has acceptance rate at the same level as the full dimensional approximation proposal. With $\kappa = 25.0$ we get a posterior that is close to Gaussian and with high spatial dependence. The acceptance rate for the full dimensional approximation proposal increases and decreases faster for more blocks of the overlapping approximated blocks proposals. Longer buffers are needed to “hide” the block borders and buffer length 66 is not enough to get acceptance as for the full dimensional approximation proposal.

The samplers with opposite reverse acceptance probability all gave acceptance rates on the same levels as the corresponding full dimensional approximation proposal samplers and results are not further reported.

The results for fixed hyper-parameters and approximation $A1$ are very similar to those in section 2.3, the only difference being that the our reference acceptance rate level is the acceptance rate of a sampler with the corresponding full dimensional approximation proposal rather then 1.

Changing hyper-parameters and approximation $A1$

The one-block Metropolis-Hastings sampler in algorithm 11 was run for 10000 iterations with overlapping $A1$-approximated blocks proposals for eight blocks with overlap 0 and 44 (the bandwidth). For reference purposes also a sampler with full dimensional approximation proposal was run. See figure 35 for estimated auto-correlation for κ and for element x_{410}. The blocks are set up in the bandwidth reordered graph and there x_{410} is element number 270, i.e. close to a block border. In figure 36 is trace plots and cumulative means

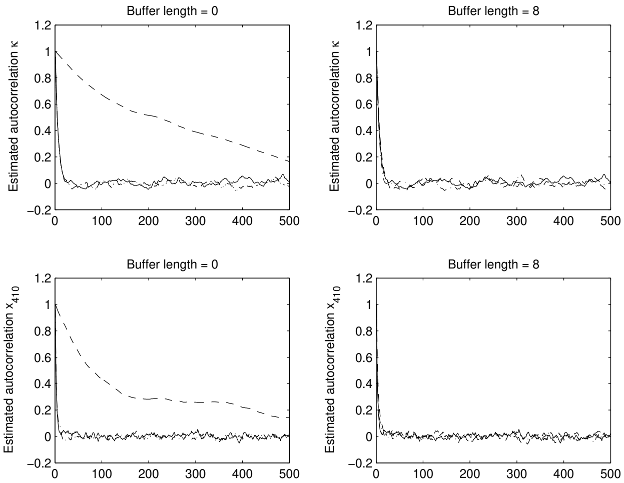


Figure 35: *To example 10, $A1$ and changing hyper-parameters: Estimated auto-correlation for κ and x_{410} with Peskun's (dashed line) and opposite reverse (dashed-dotted line) acceptance probabilities. Estimated auto-correlation for a full dimensional approximated proposal is included in all plots (solid line).*

for the different samplers for 1000 iterations after 1000 iterations burn-in. From the figures we see that blocking with Peskun's acceptance probability and without buffers causes slow mixing for κ and x_{410}. Both using the opposite reverse acceptance probability and buffers

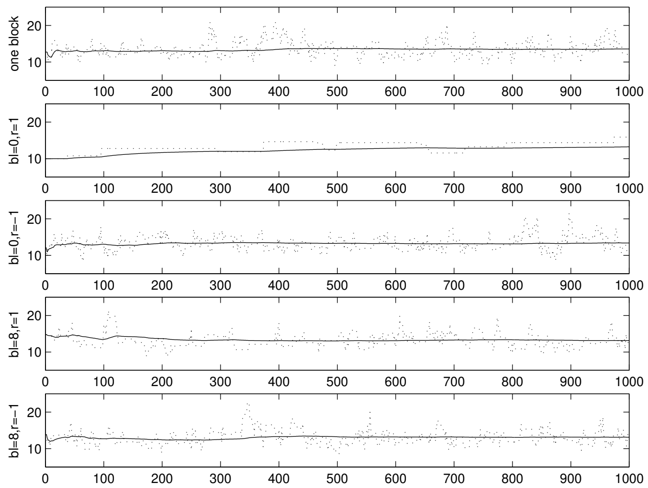


Figure 36: *To example 10, $A1$ and changing hyper-parameters: Trace-plots and cumulative means for κ. Proposals with buffer length (bl) zero and 44 and Peskun's ($r = 1$) and opposite reverse ($r = -1$) acceptance rate. Burn-in of* 1000 *iterations omitted.*

make the mixing almost as fast as for the sampler with the full dimensional approximation proposal for x. This agrees with results for the sampler in section 4.1.

Fixed hyper-parameter and approximation $A2$

A Metropolis-Hastings sampler with an overlapping $A2$-approximated blocks proposal was tested with fixed hyper-parameters for the same hyper-parameters, number of blocks and buffer lengths as for approximation $A1$. The samplers were run for 1000 iterations. Plots of the acceptance rates as function of number of blocks for samplers using Peskun's acceptance probability are in figure 37. The most significant difference from the $A1$ samplers (see figure 34) is the improved acceptance rates for $\kappa = 1.0$, i.e. little spatial dependence and a posterior far from Gaussian. The posterior is then much influenced by the non-Gaussian likelihood. Approximation $A2$ point-wise correct for non-Gaussian parts of the likelihood term. For $\kappa = 1.0$ the acceptance rate does not decrease much when blocking and only a short buffer is needed to reestablish the acceptance rate level of the full dimensional approximation proposal. For the realistic smoothing parameter, $\kappa = 10.0$, the sampler with full dimensional $A2$-approximation proposal has improved compared to $A1$. As in the $A1$ case the acceptance rate decreases as the number of blocks increases and increases with increased buffer length. With a buffer length of 66 (1.5 times the bandwidth) the acceptance rate for samplers using overlapping approximated blocks proposals are at the same level as for the full dimensional approximation proposal sampler. For $\kappa = 25.0$ the posterior is almost Gaussian and the refinements of $A2$ do not increase the acceptance rate

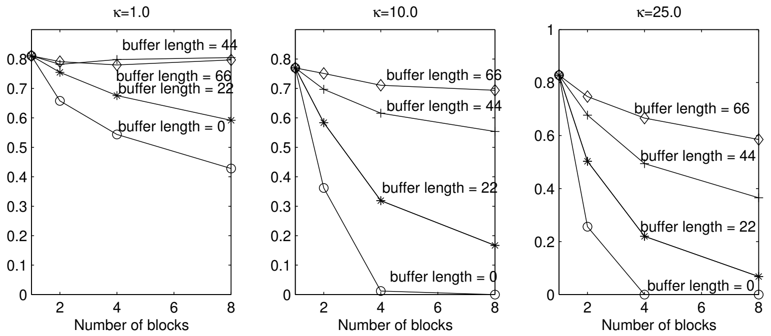


Figure 37: *To example 10, $A2$ and fixed hyper-parameters: Acceptance rates buffers length 0 (stars), 22(stars), 44(plus signs) and 66 (diamonds) using Peskun's acceptance probability.*

much. As a function of the number of blocks and buffer lengths it also behaves as $A1$.

Changing hyper-parameter and approximation $A2$

As for $A1$-approximated proposals the one-block Metropolis-Hastings sampler in algorithm 11 was run with overlapping $A2$-approximated blocks proposals for eight blocks with overlap 0 and 44 (the bandwidth) and with a full dimensional approximation proposal. Due to higher computational cost the $A2$ samplers were run for only 2000 iterations. See figure 38 for estimated auto-correlation for κ and for x_{410}. In figure 39 is trace plots and cumulative means for the different samplers for the last 1000 iterations, i.e. after 1000 iterations burn-in. For the range of κ with high posterior density $\pi(x|y,\kappa)$ is close enough to Gaussian to use approximation $A1$ and we can not observe any improved mixing of either κ nor x_{410} from using approximation $A1$ for this dataset.

5.4 Example 11: Cervical cancer in GDR

In this example data on new incidences of cervical cancer in the former German Democratic Republic (GDR) are analysed. The data are available on a yearly basis from 1961 until 1989 (i.e. for $T = 29$ years) and for each of GDR's $N = 216$ administrative districts. Further are the incidences reported with the age-group of the woman ($J = 15$ age-groups; $0-20, 20-24, 25-29, \ldots, 80-84, 85+$) and which stage the cancer was discovered in (six stages, with stage six as the most severe). We aggregate the data into a premalignant stage (stage 1 and 2), denoted $S1$, and a malignate stage (stage 3-6) denoted $S2$. Our interest is the proportion of cases discovered in $S1$ and its variation in time, age and space. Data from 1975 have previously been analysed in Knorr-Held et al. (2002). Pap smear screening programs were introduced in this period and results in Knorr-Held et al. (2002) as well as earlier results suggest large spatial variability with respect to the time of introduction and effectiveness of the screening.

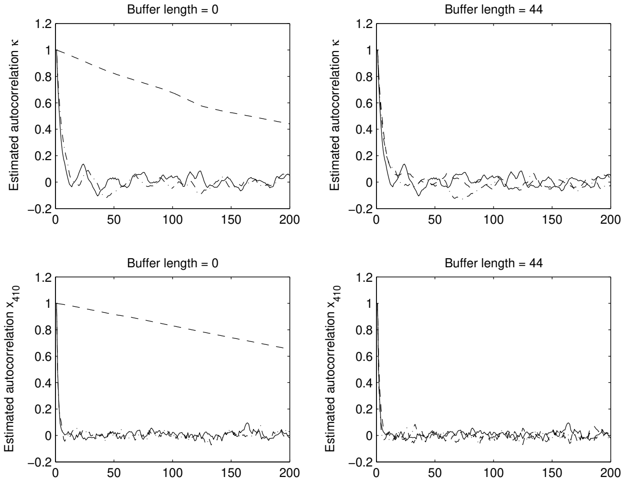


Figure 38: *To example 10, A2 and changing hyper-parameters: Estimated auto-correlation for κ and x_{410} with buffers length 0 and 44 and Peskun's (dashed line) and opposite reverse (dashed-dotted line) acceptance probability and with a full dimensional approximation (solid line).*

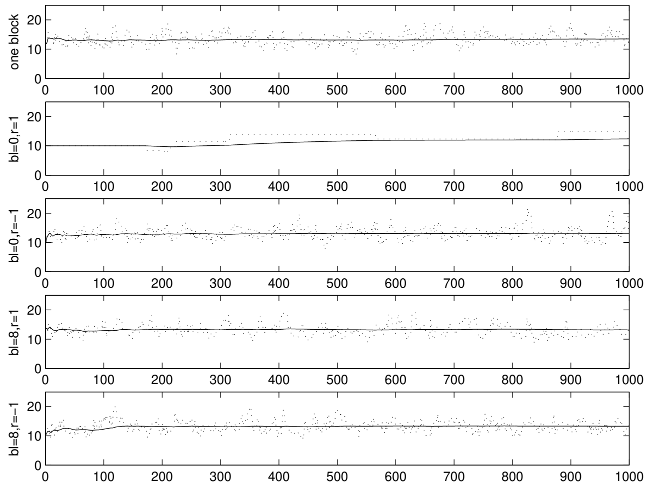


Figure 39: *To example 10, A2 and changing hyper-parameters: Trace-plots and cumulative means for κ, buffer lengths (bl) zero and 44 and Peskun's ($r = 1$) and opposite reverse ($r = -1$) acceptance probability.*

In this example we model the proportion of discoveries in $S1$ for age-space and time-space. An age-time-space model is not yet developed and from a data-analysis point of view the analysis in this example should be considered a pre-study.

Let p_{ijt} denote the probability that a discovered cervical cancer case in year t is in stage $S1$ for a woman in district i and age-group j. We only model either age-space or time-space and the third index is then suppressed. The number of cases discovered in $S1$ for a cell, y_{ijt}, is assumed binomial;

$$y_{ijt} \sim \text{bin}(p_{ijt}, N_{ijt})$$

where N_{ijt} is the total number of discovered cases for region i, age-group j and year t. The likelihood is assumed mutually independent; $\pi(y|p) = \prod_i \prod_j \prod_t \pi(y_{ijt}|p_{ijt})$. We use the logit transform, and further model the log relative success probability;

$$\text{logit}(p_{ijt}) = \log(\frac{p_{ijt}}{1 - p_{ijt}}) = x_{ijt}$$

Space-time-age dependence is introduced through the prior of x. We start off considering data for one year making a space-age model. Later we us the same model for space-time for data aggregated over age-groups.

Space-age model

It is reasonable that there is an overall level for each age-group, β_j $(j = 1, 2, \ldots, J)$, $\text{E}(x_{ij}) = \beta_j$. Further we believe there are dependence both in time- and in age-direction, and we choose to use the time-space prior introduced in section 4.3 with some additional white noise;

$$\pi(x|\beta, \tau_S, \tau_A, \tau) \sim N((I_A \otimes \mathbf{1}_N)\beta, Q(\tau_S, \tau_A) + \tau I)$$

where $\mathbf{1}_N$ is a vector of length N containing ones and element (i, j) of $Q(\tau_S, \tau_A)$ is given by

$$Q_{ij} = \begin{cases} -\tau_S, & \text{if } i \overset{s}{\sim} j \\ -\tau_A, & \text{if } i \overset{a}{\sim} j \\ \tau_S \text{nnb}_s(i) + \tau_A \text{nnb}_a(i), & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

where $\overset{s}{\sim}$ denotes neighbours in space and $\overset{a}{\sim}$ in age, $\text{nnb}_s(i)$ is the number of neighbours for element i in space and $\text{nnb}_a(i)$ in age. Age-group j's neighbours are the age-group below, $i = j - 1$, and above, $i = j + 1$. The overall level β is given an intrinsic Gaussian prior,

$$\pi(\beta|\tau_\beta) \propto \exp(-\frac{1}{2}\tau_\beta \sum_{j \overset{a}{\sim} k} (\beta_k - \beta_j)^2)$$

The prior is illustrated in figure 40. The joint distribution (x, β) is also a GMRF, see Appendix A.5.

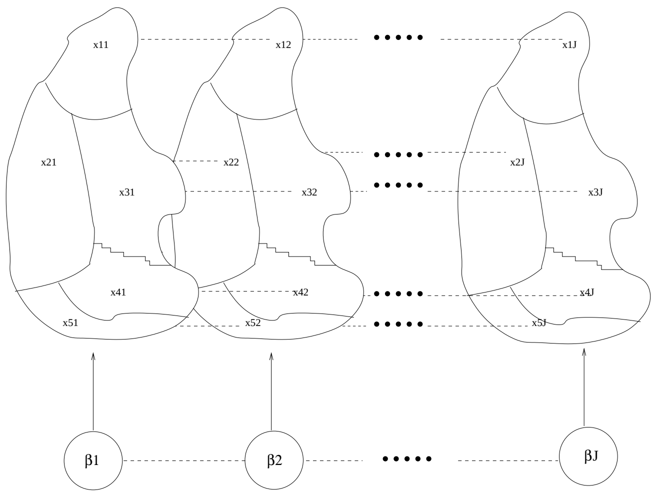


Figure 40: *To example 11: An illustration of the prior model for* (x, β) *(hyper-parameters not included).*

There have appeared four hyper-parameters, $\theta = (\tau_S, \tau_A, \tau_\beta, \tau)$. In addition β can be viewed as hyper-parameters. In our updating scheme θ is first proposed, then x and β jointly using an overlapping block proposal. Since β serves as a hyper-parameter for x it is included in every block. This is similar to what was done with the baseline image a in the fMRI example in section 4.4.

Data from 1961, 1975 and 1989 are analysed using the model suggested above. There are very few cases in the youngest age-group and we therefore only use the 14 oldest age-groups; $J = 14$. The dimension of the problem $(14 \cdot 216 + 14 = 3024)$ is small enough to use a full dimensional approximation as proposal for (x, β). For 1975 and 1989 $A1$-approximations were used while $A2$ was used for 1969 due to low acceptance rate else. All samplers were run for 10000 iterations.

In figure 41, 42 and 43 are trace plots of the hyper-parameters for 1961, 1975 and 1989, respectively, and in figure 44 their estimated auto-correlation functions. All parameters seem to have converged, but for 1961 we should have had a longer simulation. For all years we find a relatively small spatial dependence and a strong age dependence. The spatial dependence is smaller in 1961 then in 1975 and 1989, which can be explained with the screening programs introduced later on district level. Marginal estimated means and standard deviation for β are plotted in figure 45. We see an improvement in cases discovered in stage $S1$ between 1961 and 1989 with its largest step between 1961 and 1975. We further observe that the standard deviation for 1975 is much larger then for the other years. In figure 46, 47 and 48 are maps with the difference between the estimated mean of x_{ij} and the estimated mean of the corresponding age-group level β_j. A first observation is that there is not much difference in the spatial pattern for different age-groups, but

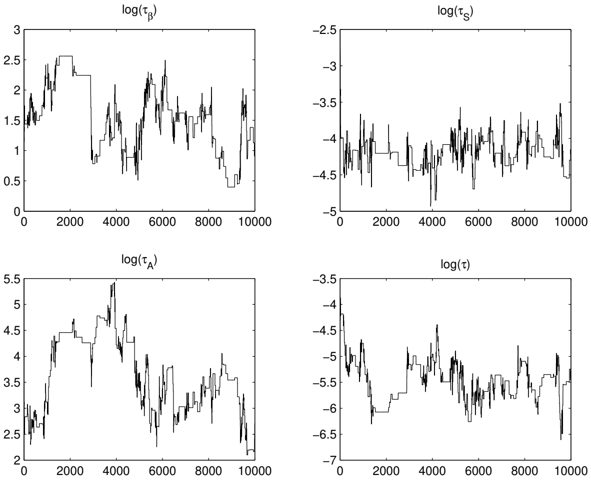


Figure 41: *To example 11, space-age model: Trace plots for hyper-parameters for 1961 data with full dimensional approximation proposal for* (x, β).

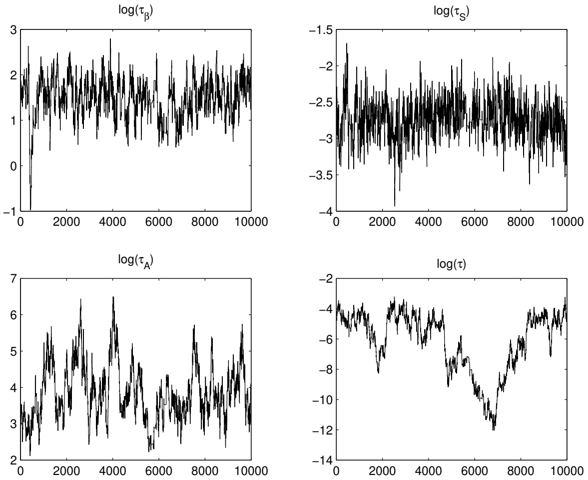


Figure 42: *To example 11, space-age model: Trace plots for hyper-parameters for 1975 data with full dimensional approximation proposal for* (x, β).

the latent field is more homogeneous for the elder age-groups. It could be argued that a two-dimensional model would be adequate. If we compare the different years we see that the spatial pattern has changed. While the north-west of GDR had some of the lowest

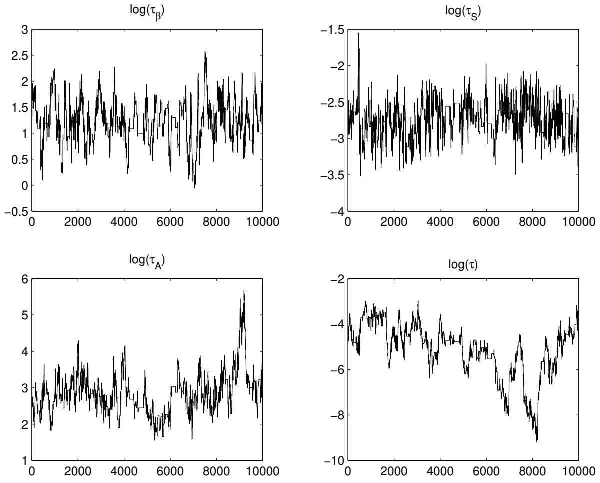


Figure 43: *To example 11, space-age model: Trace plots for hyper-parameters for 1989 data with full dimensional approximation proposal for* (x, β).

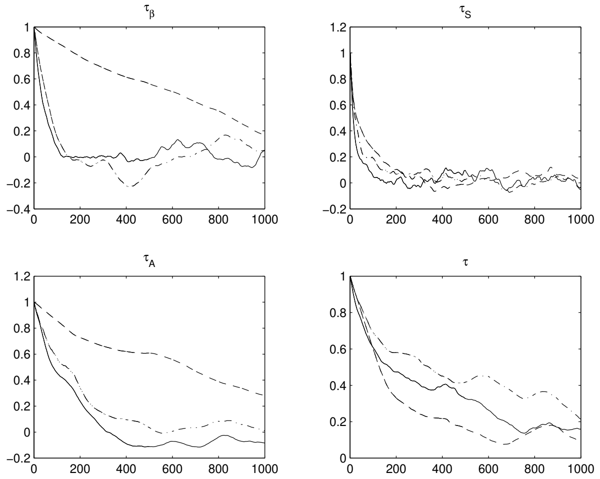


Figure 44: *To example 11, space-age model: Estimated auto-correlation for hyper-parameters for 1961 data (dashed line), 1975 data (solid line) and 1989 data (dash-dotted line) with full dimensional approximation proposal for* (x, β)

proportions in 1961 it had some of the best ones in 1975 and 1989. This indicates that a space-time model should be three dimensional.

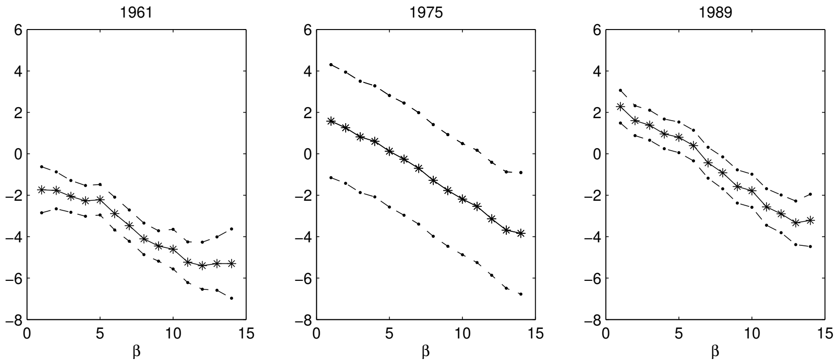


Figure 45: *To example 11, space-age model: Marginally estimated means and standard deviation of the posterior of β from every 10th sample of the one-block sampler with full dimensional approximation proposal for (x, β).*

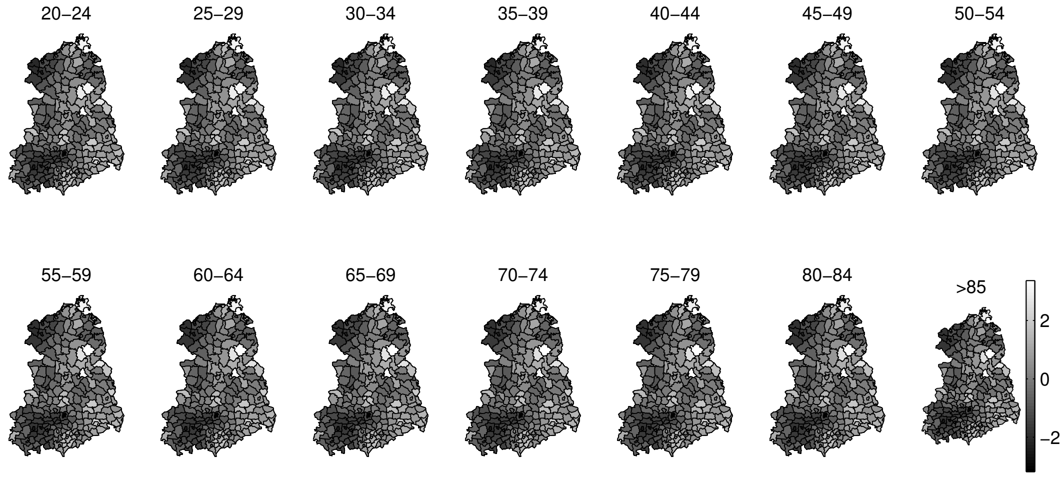


Figure 46: *To example 11, space-age model: The difference of mean estimates, $\bar{x_{ij}} - \bar{\beta}_j$, for 1961.*

For 1975 we have also ran samplers with overlapping block proposals for (x, β). In these samplers we include β in every block. The latent field can be divided either in age direction or in space. Samplers with the two different overlapping approximated blocks proposals were run; one age divided and one space divided. For the age divided proposal blocks consisted of β and nine age-groups with two age-groups and β overlap. Also the space divided proposal had two blocks. The blocks was set up from a bandwidth ordering

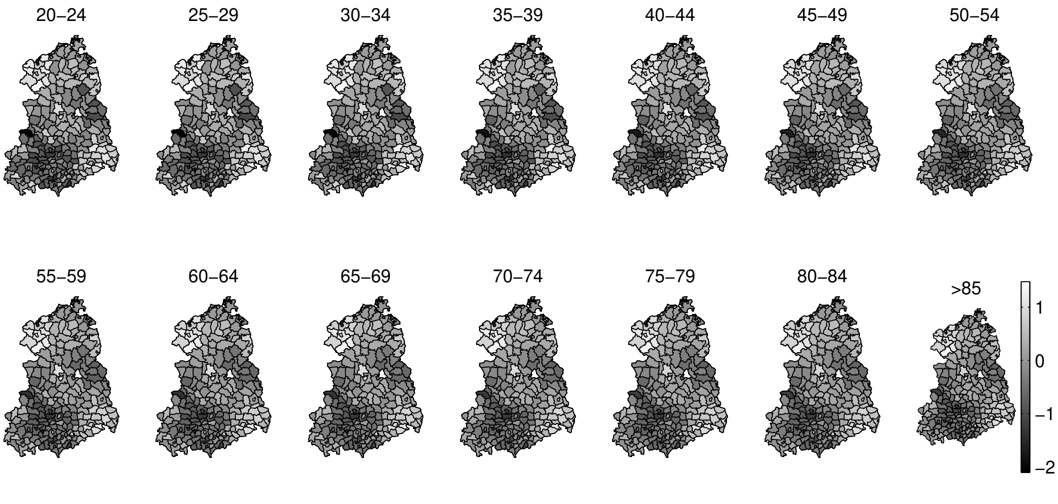


Figure 47: *To example 11, space-age model: The difference of mean estimates,* $\bar{x_{ij}} - \bar{\beta}_j$*, for 1975.*

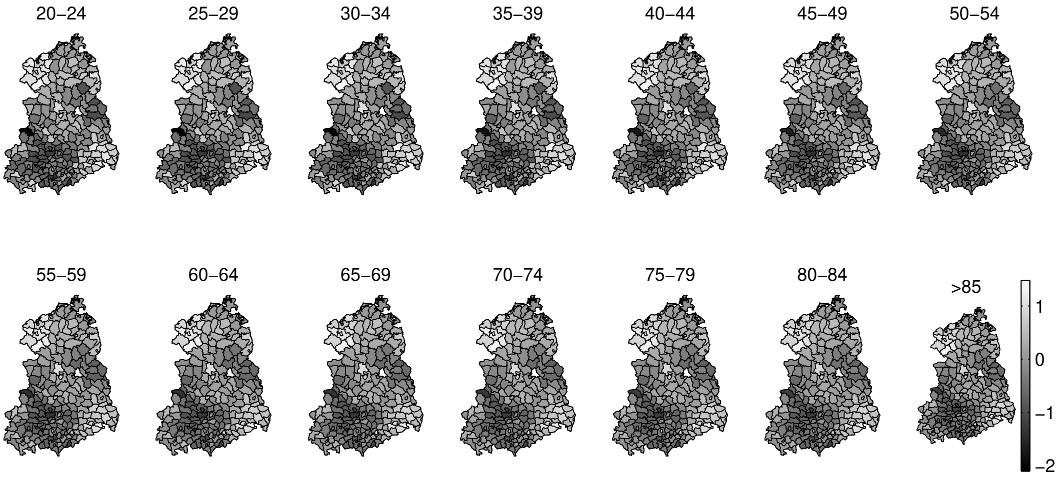


Figure 48: *To example 11, space-age model: The difference of mean estimates,* $\bar{x_{ij}} - \bar{\beta}_j$*, for 1989.*

for one age-group and the overlap corresponded to a bandwidth (for GDR $b_w = 20$, i.e. an overlap in x of 20×13 elements) in addition to β. Trace plots for the age-group divided proposal sampler are in figure 49, and for the space divided proposal sampler in figure 50.

We know from the full dimensional approximation proposal sampler that there are much

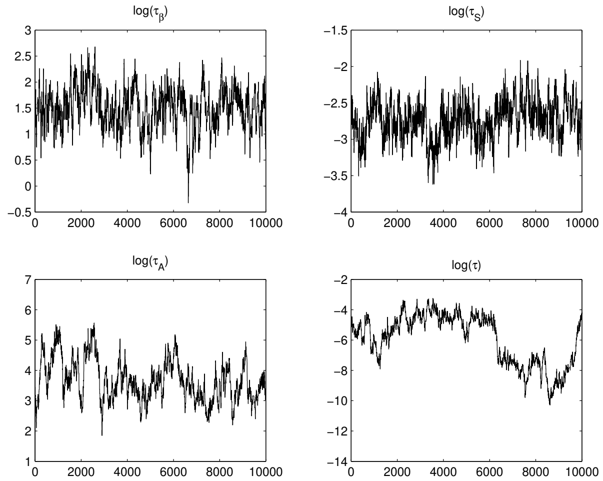


Figure 49: *To example 11, space-age model: Trace plots for hyper-parameters for 1975 data with age divided overlapping approximated blocks proposal.*

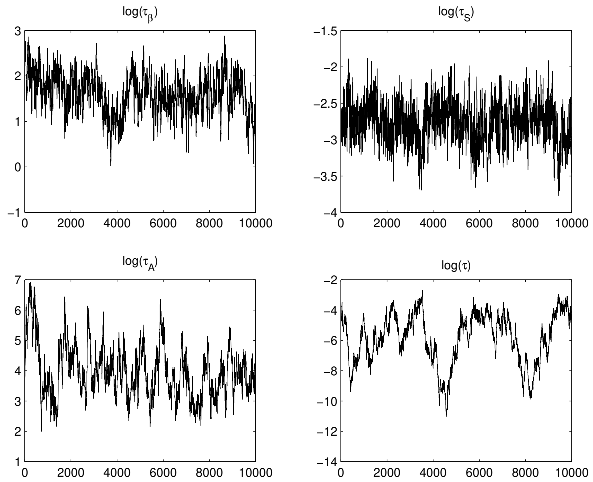


Figure 50: *To example 11, space-age model: Trace plots for hyper-parameters for 1975 data with space divided overlapping approximated blocks proposal.*

stronger dependence in age than in space. The strong dependence in age direction causes problems for the age divided sampler. Comparing plots in figure 42 and 49 we observe that the spread for τ_A is smaller for the overlapping block sampler: With high age dependence an overlap of two age-groups is not enough to propose “acceptable” samples for the latent field for large changes in τ_A.

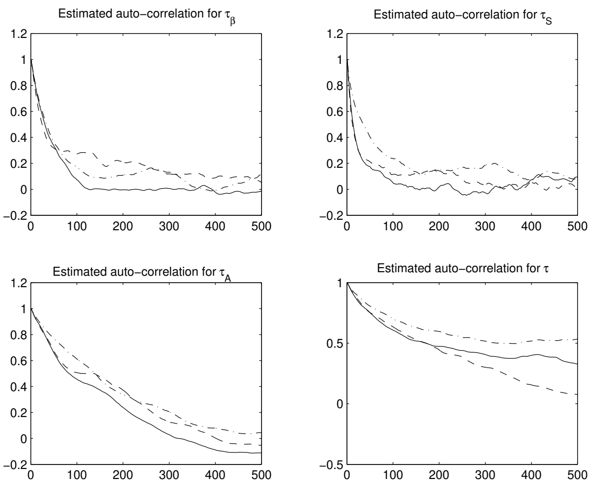


Figure 51: *To example 11, space-age model: Estimated auto-correlation for the hyper-parameters for the different proposals for 1975 data; full dimensional (solid line), age divided overlapping blocks (dashed line) and space divided overlapping blocks (dash-dotted line)*

The low spatial dependence makes the space divided overlapping block sampler look appealing. Though, we most not forget that β_j is the expected value for $x_{\cdot j}$, and not sampling the whole of $x_{\cdot j}$ and β_j together causes the same high auto-correlation problem of expected value (here β_j) as described in Rue and Follestad (2003) (and in section 1.2). The acceptance rate for the full block sampler was 0.35, while the space divided overlapping block sampler gave 0.45. We see from figure 51 that the estimated auto-correlation has increased for the space divided overlapping block proposal, especially for τ_S.

Space-time model

We now aggregate data over age-groups and make a space-time model for the proportion of incidences discovered in $S1$. We use the same model as for space-age: We assume there is an overall level for each year, β_t $(t = 1, 2, \ldots, T)$, $E(x_{it}) = \beta_t$. Further we choose similar prior for x as in the space-age model;

$$\pi(x|\beta, \tau_S, \tau_T, \tau) \sim N((I_T \otimes \mathbf{1}_N)\beta, Q(\tau_S, \tau_T) + \tau I)$$

where $\mathbf{1}_N$ is a vector of length N containing ones and element (i, j) of $Q(\tau_S, \tau_T)$ is given by

$$Q_{ij} = \begin{cases} -\tau_S, & \text{if } i \overset{s}{\sim} j \\ -\tau_T, & \text{if } i \overset{t}{\sim} j \\ \tau_S \text{nnb}_s(i) + \tau_T \text{nnb}_t(i), & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

where $\overset{s}{\sim}$ denotes neighbours in space and $\overset{t}{\sim}$ in time, $\text{nnb}_s(i)$ is the number of neighbours for element i in space and $\text{nnb}_t(i)$ in time. Year ts neighbours are the year before, $t = t-1$, and after, $t = t+1$. The overall level β is given an intrinsic Gaussian prior,

$$\pi(\beta|\tau_\beta) \propto \exp(-\frac{1}{2}\tau_\beta \sum_{j \overset{t}{\sim} k} (\beta_k - \beta_j)^2)$$

The illustrated in figure 40 is also valid for this prior. The latent field (x, β) now has dimension 6293 together with a quite dense dependency structure this makes a full dimensional approximation proposal for (x, β) computationally too expensive. We have run an one-block Metropolis-Hastings sampler with an overlapping $A1$-approximated blocks proposal for this problem. We used time-divided blocks each of length 15 years and with ten years and β as overlap. The sampler was run for 10000 iterations, see trace plots for the hyper-parameters in figure 52 and their estimated auto-correlation in figure 53. The sam-

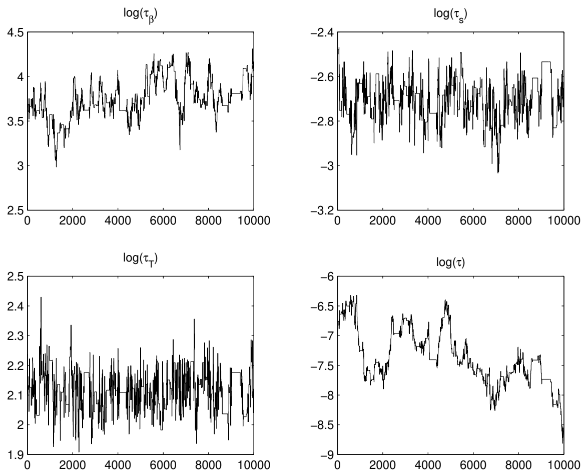


Figure 52: *To example 11, space-time mode: Trace plots for the hyper-parameters.*

pler seems to have converged. Though, both from the trace plots and from the estimated auto-correlation we see that the mixing is not very rapid and the samplers should have

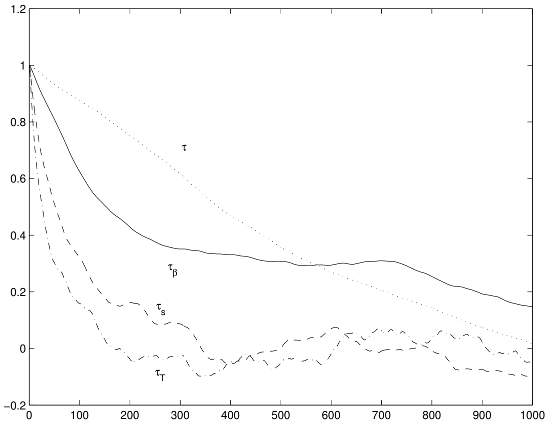


Figure 53: *Example 11, space-time model: Estimated auto-correlation for the hyper-parameters*

been run for more iterations. Comparing the hyper-parameters we find, not surprisingly, that the spatial dependence in the space-time model is at the same level as in the space-age model with data from 1975 and 1989. In figure 54 is estimated mean and standard deviation of the posterior of β. As indicated from the results from 1961, 1975 and 1989

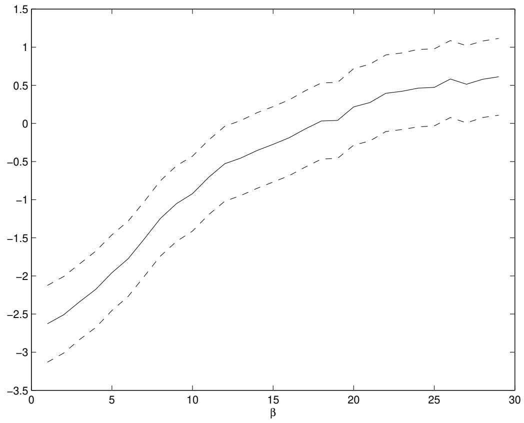


Figure 54: *Example 11, space-time model:Mean and and standard deviation for β estimated from every 10th sample.*

the proportion of cases discovered in stage $S1$ has increased especially in the first 15 years. In figure 55 is maps with the differences of the sample mean of x_{it} and mean of β_t. We see

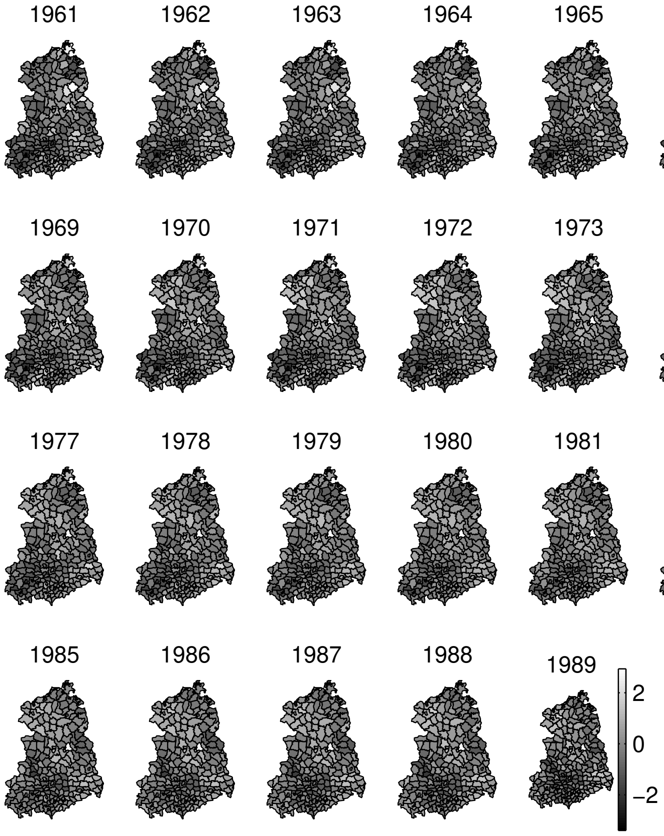


Figure 55: *Example 11, space-time model: The difference of mean estimates;* $\bar{x_{ij}} - \bar{\beta}_j$.

that the spatial structure changes over the years, and then in particular for the north-west of GDR from 1961 to the mid-seventies.

6 Discussion

We have in this report presented a method for constructing proposals for the latent field in spatial latent Gaussian Markov random field models. The key idea of the method is to do conditional sampling from small blocks of the field and to let these blocks overlap. We refer to this class of proposals as overlapping block proposals including both when each block is exact sampled and when an approximation is used. Overlapping block proposals have the appealing properties that they are relatively cheap to sample from and to evaluate, they are in most cases a good approximation to the ideal distribution $\pi(x|\theta, y)$ and they produce samples relatively independent of the previous one. These are all necessary properties when we combine a proposal for the hyper-parameters with an overlapping block proposal to the proposal of an one-block updating scheme Metropolis-Hastings algorithm.

Through examples overlapping block proposals have proved to work well for many problems both when each block is exact sampled and when an approximation is used. But the method has its limitations: As the dependence within the field gets stronger more overlap is needed. Further to evaluate the overlapping blocks proposal the blocks have to be set up such that temporary samples are never conditioned on. We have achieved this by setting up the blocks as a time series. This restricts how small blocks we can use. We have also seen that that including variables in the blocking with hyper-parameters function can cause problems and should be done with care. We can the easily fall back to the situation we want to avoid using an one-block updating scheme; slow mixing because of strong interaction between variables proposed conditioned on each other.

We have presented and named the method after how the sampling is performed. It could also been viewed as block wise partial conditional sampling. This is theoretically an other approach as we then never condition on temporary samples. But in practice this is not done anyway because we are then not able to evaluate the proposal. This approach does not have the limitation when it come to block sizes as the time series approach. A natural extension of the work done here would be to explore the opportunities of partial conditional samplers as proposals.

We believe we have introduced a powerful method for constructing proposals for the latent field when evaluating spatial latent GMRF models using one-block updating scheme Metropolis-Hasting algorithms. Overlapping block proposals can also be used for similar time and space-time models. The method enable us to use our knowledge about the dependence structure of the problem: The blocks are set up such that variable we believe are highly dependent are either sampled together or integrated out. And we achieve appropriate proposals for the latent field without working with the full dimensional distribution directly.

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 problems for the fMRI experiment and cervical cancer dataset, respectivily.

A Appendix

A.1 Proof of the overlapping Gibbs sampler

We will here prove that the overlapping block Gibbs sampler suggested in section 2.2 has $\pi(x)$ as its stationary distribution. We do this for a special case, but the extension is trivial and intuitive.

We consider a field of variables blocked as shown in figure 7. We let $\pi(x)$ be our target distribution, and $\pi(x_{B_i}|x_{B_{-i}})$ be the conditional distribution for block B_i given the rest of the field.

The transition kernel is

$$\begin{aligned} K(x,x^{'}) &= \int [\pi(x_1^{'}, x_2^{B1}, x_4^{B1}, x_5^{B1}|x_3, x_6, x_7, x_8, x_9) \\ &\quad \pi(x_2^{'}, x_3^{'}, x_5^{B2}, x_6^{B2}|x_1^{'}, x_4^{B1}, x_7, x_8, x_9) \\ &\quad \pi(x_4^{'}, x_5^{B3}, x_7^{'}, x_8^{B3}|x_1^{'}, x_2^{'}, x_3^{'}, x_6^{B2}, x_9) \\ &\quad \pi(x_5^{'}, x_6^{'}, x_8^{'}, x_9^{'}|x_1^{'}, x_2^{'}, x_3^{'}, x_4^{'}, x_7^{'})] dx_2^{B1} dx_4^{B1} dx_5^{B1} dx_5^{B2} dx_5^{B3} dx_6^{B2} dx_8^{B3} \end{aligned}$$

where x is the old sample, $x^{'}$ the new one and the extra samples for the buffers are indexed with their block numbers. From Markov chain theory it is known (see e.g. Robert and Casella (1999)) that $\pi(x)$ is the stationary distribution of an ergodic Markov chain with kernel $K(x,x')$ if

$$\pi(A) = \int_A K(x,A)\pi(dx)$$

for any $A \in \mathcal{B}(\chi)$. Here χ is the chains support, and $\mathcal{B}(\chi)$ any Borel set on χ. For convenience we denote all the buffer sample x^B. Let $\pi(x)$ be continuous, and hence the integration order can be changed.

$$
\begin{aligned}
P(X^{'} \in A) = & \int \mathbb{I}_A(x^{'}) & K(x,x^{'})\pi(x)dx^{'}dx \\
= & \int \mathbb{I}_A(x^{'}) & ([\pi(x_1^{'},x_2^{B1},x_4^{B1},x_5^{B1}|x_3,x_6,x_7,x_8,x_9) \\
& & \pi(x_2^{'},x_3^{'},x_5^{B2},x_6^{B2}|x_1^{'},x_4^{B1},x_7,x_8,x_9) \\
& & \pi(x_4^{'},x_5^{B3},x_7^{'},x_8^{B3}|x_1^{'},x_2^{'},x_3^{'},x_6^{B2},x_9) \\
& & \pi(x_5^{'},x_6^{'},x_8^{'},x_9^{'}|x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_7^{'})]dx_2^{B1}dx_4^{B3}dx_5^{B1}dx_5^{B2}dx_5^{B3}dx_6^{B2}dx_8^{B3}) \\
& & \pi(x_3,x_6,x_7,x_8,x_9)\pi(x_1,x_2,x_4,x_5|x_3,x_6,x_7,x_8,x_9)dxdx^{'} \\
= & \int \mathbb{I}_A(x^{'}) & \pi(x_2^{'},x_3^{'},x_5^{B2},x_6^{B2}|x_1^{'},x_4^{B1},x_7,x_8,x_9) \\
& & \pi(x_4^{'},x_5^{B3},x_7^{'},x_8^{B3}|x_1^{'},x_2^{'},x_3^{'},x_6^{B2},x_9) \\
& & \pi(x_5^{'},x_6^{'},x_8^{'},x_9^{'}|x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_7^{'}) \\
& & \pi(x_1^{'},x_2^{B1},x_3,x_4^{B1},x_5^{B1},x_6,x_7,x_8,x_9) \\
& & dx_3dx_6dx_7dx_8dx_9dx_2^{B1}dx_4^{B3}dx_5^{B1}dx_5^{B2}dx_5^{B3}dx_6^{B2}dx_8^{B3}dx^{'} \\
= & \int \mathbb{I}_A(x^{'}) & \pi(x_4^{'},x_5^{B3},x_7^{'},x_8^{B3}|x_1^{'},x_2^{'},x_3^{'},x_6^{B2},x_9) \\
& & \pi(x_5^{'},x_6^{'},x_8^{'},x_9^{'}|x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_7^{'}) \\
& & \pi(x_1^{'},x_2^{'},x_3^{'},x_4^{B1},x_5^{B2},x_6^{B2},x_7,x_8,x_9) \\
& & dx_7dx_8dx_9dx_4^{B1}dx_5^{B2}dx_5^{B3}dx_6^{B2}dx_8^{B3}dx^{'} \\
= & \int \mathbb{I}_A(x^{'}) & \pi(x_5^{'},x_6^{'},x_8^{'},x_9^{'}|x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_7^{'}) \\
& & \pi(x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_5^{B3},x_6^{B2},x_7^{'},x_8^{B3},x_9) \\
& & dx_3dx_6dx_7dx_8dx_9dx_2^{B1}dx_4^{B3}dx_5^{B1}dx_5^{B2}dx_5^{B3}dx_6^{B2}dx_8^{B3}dx^{'} \\
= & \int \mathbb{I}_A(x^{'}) & \pi(x_1^{'},x_2^{'},x_3^{'},x_4^{'},x_5^{'},x_6^{'},x_7^{'},x_8^{'},x_9^{'})dx^{'} \\
= & \int_A \pi(x^{'}) & dx^{'}
\end{aligned}
$$

If we imaging each block as a leaf, in each step of the proof we integrate out those variables just covered by the leaf. This can of course be extended to more blocks, and other configurations. The only requirement is that each element is updated all least once. A special case is the traditional block Gibbs sampler.

A.2 The normalisation constant for the prior

The precision matrix for the space-time model in section 4.3 is non-positive definite, and the determinant is 0. We still need to know the normalisation constant as a function of κ 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 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 $\kappa = 1$):

$$Q_T = R_T \otimes I_N$$
$$Q_S = I_T \otimes R_S$$

where I_N and I_T are identity matrices of dimension $N \times N$ and $T \times T$. The spectral theorem gives that that symmetric real matrices can be decomposed as:

$$\begin{aligned} Q_S &= V_S \Lambda_S V_S \\ Q_T &= V_T \Lambda_T V_T \end{aligned}$$

with orthonormal eigenvectors of $Q_{\cdot}$ in $V_{\cdot}$ and eigenvalues in $\Lambda_{\cdot}$. Two diagonalisable matrixes A and B share eigenvector matrix V if and only if $AB = BA$, and

$$\begin{aligned} Q_S Q_T &= (I_T \otimes R_S)(R_T \otimes I_N) \\ &= (I_T R_T) \otimes (R_S I_N) \\ &= R_T \otimes R_S \\ Q_T Q_S &= (R_T \otimes I_N)(I_T \otimes R_S) \\ &= (R_T I_T) \otimes (I_N R_S) \\ &= R_T \otimes R_S \end{aligned}$$

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 $e_{Ai}e_{Bj}$ $\forall i \in \{1, 2, \ldots, N\}$ and $j \in \{1, 2, \ldots, T\}$. Since the identity matrix has eigenvalues 1 and eigenvectors equal the standard basis we see that;

$$\begin{aligned} \mathrm{diag}(\Lambda_T) &= (\lambda_{T1}, \lambda_{T1}, \ldots, \lambda_{T1}, \lambda_{T2}, \ldots, \lambda_{T2}, \ldots, \lambda_{TT}, \ldots, \lambda_{TT}) \\ \mathrm{diag}(\Lambda_S) &= (\lambda_{S1}, \lambda_{S2}, \ldots, \lambda_{SN}, \lambda_{S1}, \ldots, \lambda_{SN}, \ldots, \lambda_{S1}, \ldots, \lambda_{SN}) \end{aligned}$$

Further we see that

$$\det(Q_S + Q_T) = \det(\Lambda_S + \Lambda_T) = \prod_{i=1}^{N} \prod_{j=1}^{T} (\lambda_{Si} + \lambda_{Tj})$$

or for our determinant det*:

$$\det{}^*(Q_S + Q_T) = \det(\Lambda_S + \Lambda_T) = \prod_{i=1}^{N}\prod_{j=1}^{T} f(\lambda_{Si} + \lambda_{Tj})$$

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.

A.3 Functional magnetic resonance images

In figure 56 are the first 20 images of the fMRI experiment.

A.4 Calculating $\pi(a, b|\tau_{Data}, \tau_A, \tau_B, \tau_T)$

The distribution is given by the likelihood $\pi(y|x, \tau_{Data})$ and the priors of a and b:

$$\pi(a, b|\tau_{Data}, \tau_A, \tau_B, \tau_T) \propto \pi(y|a, b, \tau_{Data})\pi(a|\tau_A)\pi(b|\tau_B, \tau_T))$$

The likelihood term is multivariate Gaussian, and can be written as:

$$y|a, b, \tau_{Data} \sim N(\mathbf{1}_T \otimes a + (\mathrm{diag}(z) \otimes I_N)b, \tau_{Data} I_{NT})$$

where $\mathbf{1}$ is a column vector of size T containing ones, and I_m is an identity matrix of size m. The priors are intrinsic Gaussian as given in section 4.4. Hence is $\pi(a, b|\tau_{Data}, \tau_A, \tau_B, \tau_T)$ multivariate Gaussian, and can be written as.

$$\pi(a, b|\tau_{Data}, \tau_A, \tau_B, \tau_T) \propto \exp(\frac{1}{2}[a, b]Q[a, b]^T + c^T[a, b]^T)$$

with

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

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

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

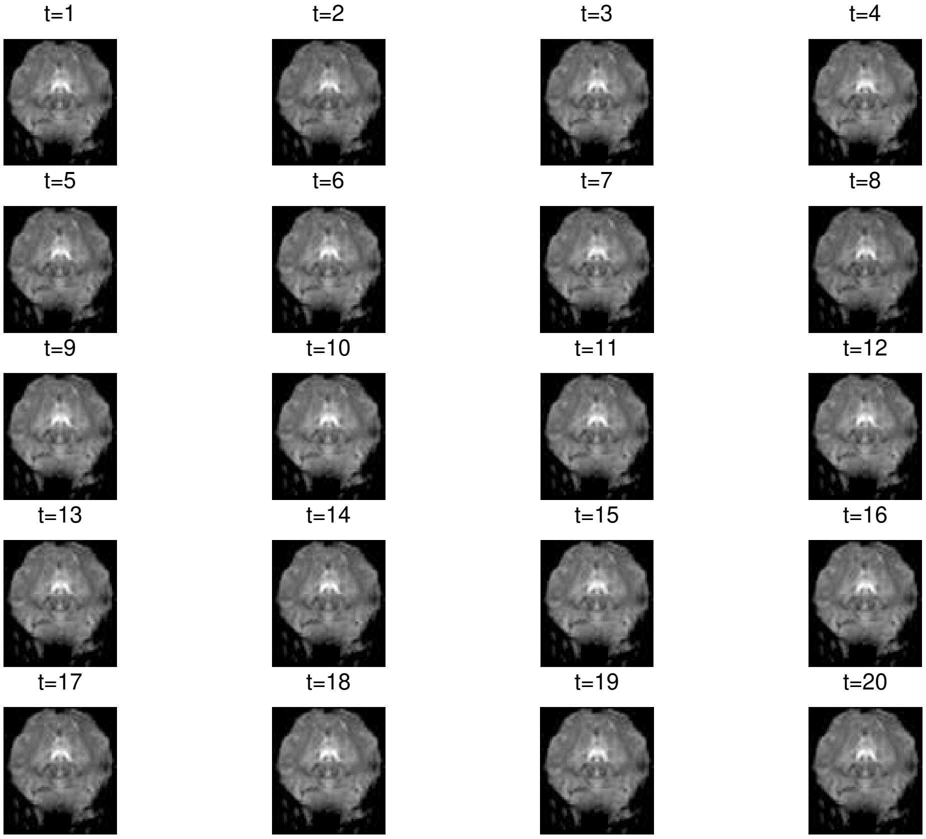


Figure 56: *To example 9: The first 20 images of the fMRI experiment.*

A.5 Calculating $\pi(x, \beta|\tau_S, \tau_A, \tau_\beta, \tau)$

The distribution is given by the priors of a and b:

$$\pi(x, \beta|\tau_S, \tau_A, \tau_\beta, \tau) \propto \pi(x|\beta, \tau_S, \tau_A, \tau)\pi(\beta|\tau_\beta)$$

The precision matrix of $x|\beta$ can be written as

$$Q_x = \tau_A(R_A \otimes I_N) + \tau_S(I_J \otimes R_S) + \tau I_{N \cdot J}$$

where R_A is the precision matrix for the one region over all age-groups and R_S is for one age-group over all regions with $\tau_A = 1$ and $\tau_S = 1$. In the same way we set $Q_\beta = \tau_\beta R_\beta$.

We now get

$$\begin{aligned}
& \pi(x,\beta|\tau_S,\tau_A,\tau_\beta,\tau) \\
\propto\ & \exp(-\frac{1}{2}[x-(I_T\otimes \mathbf{1}_N)\beta]^T(\tau_A(R_A\otimes I_N)+\tau_S(I_J\otimes R_S)+\tau I_{N\cdot J})[x-(I_T\otimes \mathbf{1}_N)\beta]) \\
& \exp(-\frac{1}{2}\tau_\beta\beta^T R_\beta\beta) \\
=\ & \exp(-\frac{1}{2}(x,\beta)^T Q(x,\beta))
\end{aligned}$$

with

$$Q=\begin{bmatrix}\tau_A(R_A\otimes I_N)+\tau_S(I_J\otimes R_S)+\tau I_{N\cdot J} & \tau_T(R_A\otimes \mathbf{1}_N^T)+\tau(I_J\otimes \mathbf{1}_N^T) \\ \tau_T(R_A\otimes \mathbf{1}_N)+\tau(I_J\otimes \mathbf{1}_N) & \tau_\beta R_\beta+N(\tau_A R_A+\tau I_J)\end{bmatrix}.$$

References

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

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B*, 61(4):691–746.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregression. *Biomentrica*, 82(4):733–746.

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

Fernandez, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: A Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 64(4):805–826.

Gamerman, D., Moreira, A. R. B., and Rue, H. (2003). Space-varying regression models: specifications and simulations. *Computational Statistics and Data Analysis.* to appear in Special Issue on Computational Economics.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

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.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25(3):435–450.

Knorr-Held, L., Gasser, G., and Becker, N. (2002). Disease mapping of stage-specific cancer incidence data. *Biometrics*, 58:492–501.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56:13–21.

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. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Verlag New York.

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.

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

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer Verlag, New York.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338.

Rue, H. and Follestad, T. (2003). Gaussian Markov random field models with applications in spatial statistics. Preprint Statistics 5/2003, Norwegian University of Science and Technology.

Rue, H., Steinsland, I., and Erland, S. (2003). Approximating hidden Gaussian Markov random fields. Preprint Statistics 1/2003, Norwegian University of Science and Technology.

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

Steinsland, I. (2003). Parallel sampling of GMRFs and geostatistical GMRF models. Preprint Statistics 7/2003, Norwegian University of Science and Technology.

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

Tjelmeland, H. and Hegstad, B. K. (2002). Mode jumping proposal in MCMC. *Scandinavian Journal of Statistics*, 28:205–223.

Wikle, C. K., Berliner, L. M., and Cressie, N. A. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.