Computational Statistics with Environmental and Remote Sensing Applications

Alexei Teterukovskiy
Centre of Biostochastics
Department of Forest Economics
Umeå

Doctoral thesis
Swedish University of Agricultural Sciences
Umeå 2003

Acta Universitatis Agriculturae Sueciae
Silvestria 277

ISSN 1401-6230
ISBN 91-576-6511-7

Printed by: SLU Grafiska Enheten, Umeå, Sweden, 2003

Abstract

Teterukovskiy A. 2003. *Computational Statistics with Environmental and Remote Sensing Applications.* Doctoral dissertation.
ISSN 1401-6230, ISBN 91-576-6511-7.

This thesis deals with application of several methods of computational statistics to the estimation of the parameters in various models in remote sensing and environmental applications.

The considered methods and applications are the following:

- mapping of the spatial distribution of reindeer in the case of the incomplete ground survey by the Gibbs sampler
- detection of small-sized tracks in aerial photos and satellite images with help of the Gibbs sampler
- contextual classification of multispectral images with spatially correlated noise using Markov chain Monte Carlo methods and Markov random field prior
- maximum likelihood estimation of the parameters of forest growth models with measurement errors
- maximum spacing estimation based on Dirichlet tesselation for univariate and multivariate observations.

In paper I we try to answer the following question. In mapping of animal distributions what is the minimum adequate number of plots one must survey to maintain a high accuracy of prediction? We use the Gibbs sampler to simulate the data for unsurveyed plots, and then use the simulated data to fit the autologistic model.

In paper II we propose an algorithm for extracting small tracks from remotely sensed images. We specify several prior distributions of varying complexity, and calculate a maximum a posteriori estimate of the map of tracks using the Gibbs sampler.

Paper III deals with classification of multispectral imagery in presence of autocorrelated noise. By means of simulation study we show how the classification results of conventional algorithms can be improved by adopting the Markov random field prior model.

In paper IV the forest growth model with measurement errors is introduced. We establish some asymptotic properties for maximum likelihood estimates of the parameters of this model.

Paper V is devoted to maximum spacing estimation based on Dirichlet tesselation. We prove consistency of such maximum spacing estimate in univariate case and conjecture it holds in higher dimensions.

Keywords: spatial autocorrelation; road detection; misclassification rate; consistency; efficiency.

Author's address: Alex Teterukovskiy, Centre of Biostochastics, SLU, SE-901 83 UMEÅ, Sweden, e-mail: Alex.Teterukovskiy@sekon.slu.se

Contents

Appendix

Papers I–V

The present thesis is based on the following papers, which will be referred to by their Roman numerals.

I. Teterukovskiy, A. and Edenius, L. (2001) 'Effective field sampling for predicting the spatial distribution of reindeer (*Rangifer tarandus*) with help of the Gibbs sampler'. Submitted.

II. Teterukovskiy, A. (2003) 'Detection of tracks in aerial photos by the Gibbs sampler'. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 17, No. 1, pp. 1-16.

III. Teterukovskiy, A. and Yu, J. (2002) 'Contextual reclassification of multispectral images: A Markov Random Field approach'. *Information Processes*, Vol. 2, No. 1, pp. 12-21.

IV. Teterukovskiy, A. and Ranneby, B. (2003) 'Maximum likelihood estimation in forest growth models with measurement errors'. Submitted.

V. Teterukovskiy, A. and Ranneby, B. (2003) 'Maximum spacing estimation based on Dirichlet tesselation'. Manuscript.

Papers II and III are reproduced by permission of the journals concerned.

1 Introduction

This thesis is devoted to the application of several methods of computational statistics to the estimation of the parameters in various models in image analysis and environmental applications. In papers I, II and III the approach is in principle Bayesian, and the Markov chain Monte Carlo method called the Gibbs sampler is used to obtain the maximum a posteriori estimate (MAPE) of the parameters. In paper IV the maximum likelihood estimate (MLE) of the parameters of the non-linear regression function is sought, and, finally, in paper V, the maximum spacing estimate (MSP) based on Dirichlet tesselation for univariate and multivariate observations is investigated.

2 Bayesian inference

Since the first three papers of the current thesis are dealing with methods of Bayesian statistics, we start with brief formulation of the problems of Bayesian inference.

From a Bayesian perspective, both the data and the parameters of the model governing the data, are treated as random quantities. Let us denote the data by D, and the parameters by θ. Then the joint distribution of the data and the parameters can be written as $P(D, \theta) = P(D|\theta)P(\theta)$, where $P(\theta)$ is the prior distribution and $P(D|\theta)$ is the distribution of the data given the parameters, or the likelihood. Application of Bayes theorem

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta} \tag{1}$$

allows us to obtain the posterior distribution $P(\theta|D)$ which is the main point of interest in Bayesian statistics.

Most problems with Bayesian analysis stem from the fact that the "real" prior distribution is usually unknown and specified, to a certain degree, subjectively. If the chosen prior distribution considerably deviates from the "real" one, the Bayesian analysis can produce misleading results. In practice, however, the realistic choice of the prior distribution gives way to trustworthy results that would be hard to obtain otherwise.

By sampling from the posterior distribution one can make inference about the unknown parameters of the model. The latter can also include the unobserved part of data (see, for example, paper I of the current thesis).

One way to perform this inference, is to estimate the mode of the posterior distribution. Thus one obtains the MAPE, which is very often used in image analysis applications (see, for example, Besag (1986) or papers II and III of the current thesis).

The integration in (1) is the cause of the practical difficulties for the Bayesian inference in high dimensions. It is possible to tackle this issue by taking one of the dynamic Monte Carlo (i.e. numerical) approaches known as Markov chain Monte Carlo (MCMC) methods (see, e.g. Smith and Roberts (1993)).

3 MCMC methods

MCMC methods provide a framework for analyzing many complicated problems by realistic statistical modeling. As is obvious from the name itself, MCMC is essentially the numerical (Monte Carlo) integration coupled with Markov chains. The original idea, introduced by Metropolis et al. (1953), is as follows. Suppose we need to sample from the given probability distribution. In many real-life situations direct sampling is impossible due to either the high-dimensionality of the data or the complicated structure of the distribution of interest or both. If this is the case, a Markov chain can be constructed in such a way that (i) it is easy to sample from, and (ii) after running the chain for a long time, samples from the chain become approximately the same as the samples from the required distribution. Out of many ways to construct such chains, we have chosen one which is convenient for our applications. Although it was known in statistical physics as the heat bath algorithm, it was introduced into mainstream statistics by Geman and Geman (1984) who called it *the Gibbs sampler.* Below we will describe the Gibbs sampler, but before that we will give some definitions and results from the Markov chains theory. These concepts are required for better understanding of the Gibbs sampler.

3.1 Markov chains with finite state space

Let $\mathbf{X}$ be a finite set called state space. Let us also specify a family $P(x, \cdot), x \in \mathbf{X}$ of probability distributions. This family will be called transition probabilities or a Markov kernel. Suppose that starting with element $X_0 \in \mathbf{X}$ we sample each subsequent state according to the corresponding distribution from a Markov kernel. The sequence $X_0, X_1, \dots$ is called a Markov chain, if $P_n(x,y) = P(X_n = y|X_{n-1} = x, X_k, k < n-1) = P(X_n = y|X_{n-1} = x)$ for all n. A Markov chain on the finite space $\mathbf{X}$ is determined by the initial distribution ν, from which the starting state is sampled, and a collection of transition probabilities $P_1, P_2, \dots$ on $\mathbf{X}$. Consider the following definition.

Definition 1 *Let μ be a probability distribution on $\mathbf{X}$. If distribution μ satisfies $\mu P = P$ for some distribution P then distribution μ is called **invariant** or **stationary** for P.*

Subject to regularity conditions, the Markov chain gradually converges to a unique stationary distribution, regardless of the initial distribution used. Therefore, after a sufficiently long *burn-in* phase, samples from the chain will look like the samples from the stationary distribution. This result will be made exact by the following theorem.

Let us denote by $c(P)$ the following quantity called the *contraction coefficient*:

$$c(P) = \tfrac{1}{2} \max_{x,y} ||P(x,\cdot) - P(y,\cdot)||,$$

where the norm is given by

$$||P(x,\cdot) - P(y,\cdot)|| = \sum_z |P(x,z) - P(y,z)|.$$

The following theorem (see, for example, Winkler (2003)) reads

Theorem 1 *Let $P_n, n \geq 1$, be Markov kernels and assume that each P_n has a stationary distribution μ_n. Assume that the following conditions are satisfied*

$$\sum_n ||\mu_n - \mu_{n+1}|| < \infty,$$

$$\lim_{n\to\infty} c(P_i \dots P_n) = 0 \quad \text{for every } i \geq 1.$$

Then $\mu_\infty = \lim_{n\to\infty} \mu_n$ exists and uniformly in all initial distributions ν,

$$\nu P_1 \dots P_n \to \mu_\infty$$

for $n \longrightarrow \infty$.

3.2 The Gibbs sampler

The question that remains now is how to construct such a chain that its stationary distribution would be the distribution of interest. We will describe one way to do it, namely the Gibbs sampler, which was used in papers I-III of this thesis.

Suppose that each component $x \in \mathbf{X}$ can be decomposed into a number of components $x_i, i = 1, \dots, k$. Let us denote our targeted distribution by $\pi(x) = \pi(x_1, x_2, \dots, x_k)$. In many cases when sampling from distribution $\pi(\cdot)$ is impossible, it is possible to sample from the conditional distributions $\pi(x_i|x_{-i})$, where $x_{-i} = \{x_j, j \neq i, j = 1, \dots, k\}, i = 1, \dots, k$.

First, let us pick an arbitrary starting configuration $x^0 = (x_1^0, \dots, x_k^0)$. The Gibbs sampler operates as follows:

- sample x_1^1 from $\pi(x_1|x_{-1}^0)$
- sample x_2^1 from $\pi(x_2|x_1^1, x_3^0, \dots, x_k^0)$

- ...
- sample x_k^1 from $\pi(x_k|x_{-k}^1)$.

At this moment one *sweep* of the Gibbs sampler, i.e., the transition to configuration $x^1 = (x_1^1, \ldots, x_k^1)$, is completed. Sweeps are repeated until the chain has converged. Recall that there were no restrictions on the visiting scheme. The components to be updated may be chosen randomly or in strict order, provided only that all components are visited infinitely often (for theoretical convergence).

One must note that if the updated components are strongly correlated, the convergence will be slow due to little movement at each step. To avoid this problem joint updating of blocks of correlated components can be considered. Of course in this case one is forced to sample from the *multivariate* conditional distributions.

3.3 Determining the stopping time

After running the chain for sufficiently long time its distribution approaches the stationary distribution. Convergence rates for the Gibbs sampler and other MCMC methods are studied, for example, in Frigessi et al. (1993). There exist several rules that help to determine when to stop the process (Propp and Wilson (1996), Brooks and Roberts (1999), etc.). The method called *coupling from the past* and due to Propp and Wilson (1996) is particularly interesting as it provides perfect sampling, i.e., its output are samples exactly from the desired distribution. The idea behind this method is to consider copies of the Markov chain starting in all possible states at some time in the distant past, and to run them until time 0. If all copies at time 0 are identical, it means that the value at time 0 does not depend on the starting configuration and therefore can be taken as a final output. If not all copies are identical, then an earlier starting time is tried. This technique obviously requires a lot of computer resources both in terms of memory and of time, although if the state space has a partial order which is preserved under the transition of the chain, the efficiency is improved. Various modifications of this method for complicated situations with no partial order in the state space (see, for example, Häggström and Nelander (1999)) as well as the improvements which help reduce the time to coalescence (Meng (2000)) recently appeared.

4 Image as a Gibbs field

Now we shall reformulate the notions defined above in terms appropriate for image analysis. Let S be a finite index set - the set of sites (e.g. pixels). For every $s \in S$ let $\mathbf{X}_s$ be a finite space of states x_s (colors or gray levels).

The finite product $\mathbf{X} = \prod_{s \in S} \mathbf{X_s}$ is the space of finite configurations x of size $|S|$ (all possible images). Let us consider probability measures (distributions) Π on $\mathbf{X}$. A strictly positive probability measure Π on $\mathbf{X}$ is called a random field. Any such measure Π can be written in the following form:

$$\Pi(x) = \frac{\exp(-H(x))}{\sum\limits_{z \in \mathbf{X}} \exp(-H(z))}. \tag{2}$$

This representation of Π is called the Gibbs form or Gibbs field induced by the energy function $H(\cdot)$. It is easy to notice that every random field is a Gibbs field for some energy function.

For a given subset $A \subset S$ let us consider the conditional probabilities of the form

$$\Pi(X_A = x_A | X_{S \setminus A} = x_{S \setminus A}), \quad x_A \in \mathbf{X}_A, x_{S \setminus A} \in \mathbf{X}_{S \setminus A}.$$

These Markov kernels are called the *local characteristics*. Suppose that the chain is currently in state x, and next state is y, that differs from x at most on A. Then the transition probabilities in Gibbs form will be

$$\Pi_A(x, y) = \begin{cases} \dfrac{\exp(-H(y_A x_{S \setminus A}))}{\sum\limits_{z_A} \exp(-H(z_A x_{S \setminus A}))} & \text{if } y_{S \setminus A} = x_{S \setminus A}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

If the local characteristics depend only on a small number of neighbors, then they can be evaluated easily. Note that the summation in denominator of the above formula involves only a small number of terms.

Taking into account the form of local characteristics, it can be easily shown that the Gibbs field Π is invariant for its local characteristics and hence for composition of local characteristics too.

4.1 Simulated annealing

As soon as the stationary distribution of the Gibbs sampler is attained, another problem can emerge. In image analysis applications, the posterior distributions are typically multi-modal. If one is interested in the mode of the posterior distribution, then the main question that arises is how to make the Gibbs sampler to locate the *global* maxima of the posterior distribution. A simple modification of the Gibbs sampler helps to solve this problem. Recall the energy function $H(\cdot)$. Maximization of the distribution in the Gibbs form is equivalent to minimization of its energy function. It is clear that function $H(\cdot)\beta$ for large β will have the same minima as $H(\cdot)$ but deeper. The following result (see, for example, Winkler (2003)) holds.

Lemma 1 *Let $\Pi_{H\beta}$ be a Gibbs field with energy function $H(\cdot)\beta$. Let M denote the set of global minimizers of H. Then*

$$\lim_{\beta \longrightarrow \infty} \Pi_{H\beta} = \begin{cases} \frac{1}{|M|} & \textit{if } x \in M, \\ 0 & \textit{otherwise.} \end{cases} \tag{4}$$

This means that the Gibbs sampler converges for each $\Pi_{H\beta}$, and the limits in turn converge to the uniform distribution on the minima of H. The following theorem states that increasing β in each step of the Gibbs sampler gives an algorithm that minimizes H. Let $\beta(n)$ be an increasing sequence of positive numbers. It is called a *cooling schedule* and its values *the inverse temperature.* After each sweep of the Gibbs sampler, the inverse temperature is increased to the next value according to the cooling schedule. Let us denote by P_i the transition probability of the i-th sweep of the Gibbs sampler. It is in fact a product of the transition probabilities for single pixels (taken at the same temperature). The following result (Winkler, 2003) holds

Theorem 2 *Let $|S|$ be the number of pixels in the image. Let $\beta(n)$ be a cooling schedule increasing to infinity such that*

$$\beta(n) \leq \frac{1}{\Delta |S|}$$

for some constant Δ. Then

$$\lim_{n \longrightarrow \infty} \nu P_1 \ldots P_n(x) = \begin{cases} \frac{1}{|M|} & \textit{if } x \in M, \\ 0 & \textit{otherwise} \end{cases} \tag{5}$$

uniformly in all initial distributions ν.

Remark: Instead of *sampling* from the conditional distributions one can simply choose a *mode* of such distribution. This can be understood as a variant of simulated annealing at infinite inverse temperature. This method is called Iterated Conditional Mode (ICM), and it is analogous to maximal descent algorithm in optimization. The ICM converges much faster than the Gibbs sampler, but only to the local minimum of the energy function. It still can be an appropriate choice in practice if the initial configuration is close to the global minimum. We use the ICM in paper III for reclassification, i.e. when the initial classification is obtained by another method.

4.2 Ising model

Often it is possible to decompose the energy into the sum of individual contributions of the subsets of S. This decomposition is particularly convenient for implementation of such methods as the Gibbs sampler, whose transition probabilities are in form of local characteristics.

Probably, the simplest example of such energy is the Ising field with pairwise interactions. This model often serves as an example due to its simplicity and to the fact that with its help all fundamental issues of Markov fields can be illustrated. Besides, the Ising model and its generalization - the Potts model, are very popular (though, supposedly, not best - see Descombes et al. (1999)) choices for prior distributions in Bayesian image analysis. Below we will describe the Ising model, show how to sample from it using the Gibbs sampler and comment on its usability as a prior distribution.

Consider a binary image with pixels taking values either $+1$ or -1 (in statistical physics, where the model was initially introduced, the pixels were particles, and the pixel values - spins). The Ising field in the simplest case can be defined by an energy function of the form $H(x) = -\beta \sum_{<s,t>} x_s x_t$, where $<s,t>$ indicates that pixels s and t are first-order neighbors. The positive parameter β controls the strength of attraction between the neighboring pixels. The larger the β is, the more neighboring pixels tend to be of the same color. In fact β is very close to the inverse temperature described in Section 4.1. It is easy to guess how the configurations of minimal energy for Ising model will look like. They are the two single-color images with all pixels colored either as -1 or as $+1$. Therefore, the probability distribution function of the Ising field has exactly two equal modes.

The existence of multiple modes brings about several problems. First of all, it is a cause of the phenomenon called the *phase transition*, which is a coexistence of different phases below some inverse temperature (called critical temperature), i.e., when the coupling between neighbors becomes less strong. It means that at some temperature (equal to 0.44 in Ising model) the realization of the random field becomes chaotic, e.g. in terms of image analysis, the image loses all regular patterns. Apparently, such prior distribution is not very informative.

Another problem is that an algorithm searching for a global minimum of energy function, after finding it (or not!), may start searching for others. Moreover, it will try to visit each single minimum again and again, thus complicating the determination of the stopping time.

Now we will demonstrate how to simulate configurations of the Ising field for a given value of β (we take the one well above the critical value) with the help of the Gibbs sampler.

Implementation of the Gibbs sampler for sampling from the Ising field can be done as follows. Start with a completely random configuration. Update each pixel x_s by sampling from the following Bernoulli distribution:

$$\Pi_s(x_s = 1) = \frac{e^{\beta \sum_{s \in \delta s} x_s}}{e^{-\beta \sum_{s \in \delta s} x_s} + e^{\beta \sum_{s \in \delta s} x_s}},$$

$$\Pi_s(x_s = -1) = 1 - \Pi_s(x_s = 1),$$

where δs is a collection of four nearest neighbors of location s. It would be much more efficient to update blocks of pixels simultaneously and in a special order (see, for example, Winkler (2003)), but our example suffices for illustrative purpose.

Figure 1 shows the pepper and salt initial configuration and configurations of the Ising field at inverse temperature 1.33 after a number of sweeps. We note that at high inverse temperatures (such as the one we have chosen), the image becomes degenerate (single-color), which is also useless with respect to the prior distribution. In paper III, we use the Potts prior, which is the generalization for several colors of the Ising model, and choose the value of the coupling parameter β taking into account the considerations described above.

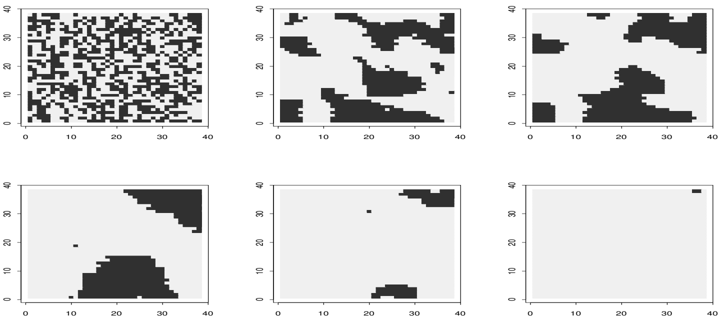

Figure 1: Ising field after 0, 10, 20, 70, 180 and 260 sweeps

4.3 Constrained optimization

Suppose, our goal is to obtain a configuration x such that

$$x = \mathrm{argmin}_{x:\ V(x)=0} U(x, y). \tag{6}$$

One can think of the pair $(U(x, y), V(x))$ as the new energy function.

Since the observed data y is fixed, we will drop it from subsequent formulas. A standard way to solve constrained optimization problems such as (6) is to introduce a Lagrange multiplier λ. Sampling from

$$\Pi(x, \beta, \lambda) = \frac{\exp\{-\beta(U(x) + \lambda V(x))\}}{\sum_{x'} \exp\{-\beta(U(x') + \lambda V(x'))\}},$$

where the summation is performed over all possible configurations of $\mathbf{X}$, and gradually increasing the inverse temperature β and the Lagrange multiplier λ would solve the minimization problem. Whereas this is impossible, it is feasible to evaluate ratios of the form

$$\frac{\exp\{-\beta(U(x_1)+\lambda V(x_1))\}}{\exp\{-\beta(U(x_2)+\lambda V(x_2))\}} \tag{7}$$

and thus conditional probabilities.

Let us denote by Ω^* the set $\{x : V(x) = 0\}$, and by $\Pi^*(x)$ the probability measure concentrated on this set, where

$$\Pi^*(x) = \frac{\exp(-U(x))}{\sum\limits_{x' \in \Omega^*} \exp(-U(x'))} \delta_{\Omega^*}(x),$$

and $\delta(\cdot)$ is the indicator function. Let us also denote by Ω_0^* the subset of Ω^* where $U(x)$ is minimal. So, $\Omega_0^* = \{\omega \in \Omega^* : U(\omega) = \min_{x\in\Omega^*} U(x)\}$. Let us denote by $X(k)$ a state of a Markov chain that we are constructing, at iteration k. The following theorems (Geman et al. (1990)) hold:

Theorem 3 *Let $\beta_k \equiv 1$ and $\lambda_k \nearrow +\infty, \lambda_k \leq c \cdot \log k$. Then*

$$\lim_{k\to\infty} P(X(k) = x | X(0) = \eta) = \Pi^*(x).$$

Here c is a constant, and η is an arbitrary initial state of the process. The second theorem is a convergence result for simulated annealing.

Theorem 4 *Let $\beta_k \nearrow +\infty$ and $\lambda_k \nearrow +\infty, \beta_k\lambda_k \leq c \cdot \log k$. Then*

$$\lim_{k\to\infty} P(X(k) = x | X(0) = \eta) = \begin{cases} |\Omega_0^*|^{-1}, & x \in \Omega_0^*, \\ 0 & \textit{otherwise.} \end{cases}$$

Example 1 We will now give an example how the Gibbs sampler can be used for detection of objects in an image. In fact this example is a simplified case of the problem considered in paper II of the current thesis.

Suppose that the true image depicts a small linear object such as a needle (Figure 2a). Let the image be degraded by noise (Figure 2b). Consider the problem of locating the needle in the degraded image. As an output of an algorithm we expect a binary map of the same size as the original image, where presence and absence of the needle would be denoted as 1 and 0, respectively. We reformulate this problem in Bayesian terms as follows.

Consider the following notations. Let $y = \{y_{ij}, 1 \leq i, j \leq N\}$ be the degraded image. Let us also introduce a dual matrix x of the following structure. $x = \{x_{ij}, 1 \leq i, j \leq N, x_{ij} \in \{0, 1\}\}$. Let us call it a label matrix. Elements of the label matrix are the binary values, representing presence/absence of a needle at the corresponding pixel. We have actually decomposed the data into two parts - observed and unobserved. Our aim

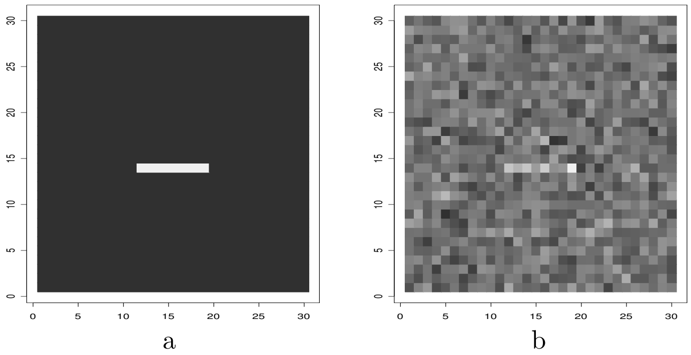

Figure 2: (a) true image (b) degraded image

is to construct a posterior density of unobserved label matrix given the observed image, and then find an MAPE by the Gibbs sampler.

Let us use the Gibbs representation for all considered probabilities. We will treat this problem as one of the constrained optimization. Let the energy function $H(\cdot)$ depend on both the observed image and the label matrix: $H(x, y) = U(x, y) + V(x)$. The first summand $U(x, y)$ is a term that accounts for data-label interaction, and the second summand $V(x)$ is a "pure" label contribution term, called the "penalty function", that will take only positive values. The larger the value of $V(x)$ the less probable is the configuration x. For $U(x, y)$ we adopt the form

$$U(x, y) = -\sum_{s} x_s \times \Phi_s(y), \tag{8}$$

where the summation extends over all pixels in the image. As $\Phi_s(y)$ we take a function that measures the possibility that a given pixel belongs to the needle. $\Phi_s(y)$ will be chosen so that pixels that are the likely candidates for a needle produce positive value of this function, whereas the pixels more likely to belong to the background produce negative values. Let us adopt the following function. For each pixel s we calculate the four mean values $m_s^1, m_s^2, m_s^3, m_s^4$ of the gray levels along the four pieces of possible straight lines passing through this pixel (suppose for simplicity that the needle can be oriented only vertically, horizontally or at angles 45° and 135°). If the pixel belongs to the needle, then one and only one of these mean values will be an "outlier" whereas the remaining three will be grouped together. Otherwise, all four values will approximately be the same. As a measure of possibility for a pixel to belong to the needle we choose

$$\Phi_s = \max_i(|m_s^i - \text{median}(m_s^1, m_s^2, m_s^3, m_s^4)|) - \text{const}.$$

We subtract a constant, so that the function Φ would take both positive and negative values. Otherwise, (8) would always attain its minimum at trivial single color configuration. As a constant we subtract the 98-percentile of $\mathbf{\Phi} = (\Phi_1, \Phi_2, \dots, \Phi_{|S|})$.

The choice of energy above was made out of the following considerations. In states with low energy one expects pixels belonging to the needle to be marked as 1 ($-x_s = -1, \Phi_s > 0$), and pixels from the background to be assigned 0 ($-x_s = 0, \Phi_s < 0$). Recall that we seek the minimum of the energy function. Penalty function (the second summand in the energy function) is a number of forbidden patterns in the image. These are patterns incoherent to our prior beliefs of the structure of the image. Since we know exactly the object we are looking for, we consider crosses, turns, extra-thick lines, loops, etc. as such patterns. Penalty function helps to assign larger probabilities to configurations more resembling the true image. After the energy function is specified, the Gibbs sampler can be straightforwardly implemented.

The results of the detection by the Gibbs sampler coupled with simulated annealing are displayed in Figure 3.

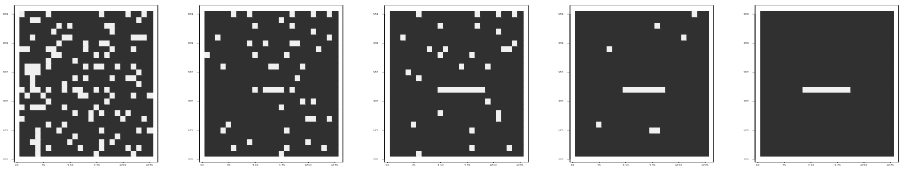

Figure 3: iterations 1 to 5

5 Maximum likelihood estimation

Maximum likelihood estimation was introduced by English statistician R. Fisher in 1912 and quickly gained world-wide recognition. We will formulate this method for independent non-identically distributed (i.n.i.d) observations, since this was the case in the problem motivating our study (heteroscedastic regression).

For a sequence of independent random variables $Y_k, k = 1, 2, \dots, n$, which take values in some measure space and possess densities $f_k(y_k, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$, the likelihood function takes form

$$L_n(\boldsymbol{\theta}) = \prod_{k=1}^{n} f_k(y_k, \boldsymbol{\theta}).$$

The maximum likelihood estimator (MLE) of the true parameter value $\boldsymbol{\theta}^0$

is denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ and is defined to be any point in Θ such that

$$L_n(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) \geq L_n(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Theta.$$

The MLE in i.n.i.d case under some conditions possesses a number of attractive properties, such as consistency (Bradley and Gart (1962), Hoadley (1971) or Borovkov (1998)), asymptotic normality (Roussas (1968), Hoadley (1971) or Borovkov (1998)) and asymptotic efficiency (Borovkov (1998)).

5.1 Estimation of parameters in growth models

Growth functions are important tools for evaluation of the wood production at alternative silvicultural regimes and the sustainability of forestry. The modern multipurpose forestry, which often aims at uneven-aged and mixed-species stands, stresses the need for accurate and generally applicable single-tree growth functions.

The most important factors to consider in a single-tree growth function are the size and condition of the tree itself, its competitive situation and the site conditions. The resolution of growth models is dependent on available data at both construction and application of the models. Natural variation in tree growth and measurement errors at growth estimation by repeated callipering of the trees are large and must be considered.

This motivates the introduction of the following growth model:

$$Y_k = e^{\boldsymbol{\beta}\mathbf{x_k} + U_k} + W_k, \tag{9}$$

where W_k is a normally distributed measurement error, whose variance depends through the known positive function g upon vector of independent variables $\mathbf{x_k}$:

$$W_k \sim N(0, \tau^2 g(\mathbf{x_k})). \tag{10}$$

Let us call the model (9) *the growth model with measurement errors.* The parameter vector $\boldsymbol{\beta}$ of dimension p in (9), the proportion coefficient τ^2 in (10), and parameters of U_k are to be estimated. Suppose that the distribution of the deviation U_k does not depend on k, and is Gaussian $N(\mu, \sigma^2)$. Obviously, the parameters of this distribution must obey the condition of unbiasedness:

$$Ee^{\boldsymbol{\beta}\mathbf{x_k} + U_k} = e^{\boldsymbol{\beta}\mathbf{x_k}}.$$

From this relation we obtain $\mu = -\sigma^2/2$, which reduces by one the number of the parameters to be estimated. Therefore, the total number of such parameters is $m = p + 2$. Let us denote the vector of the parameters for convenience by $\boldsymbol{\theta}$=$(\theta_1, \theta_2, \ldots, \theta_m)$.

It is clear that (9) is a non-linear regression model with an error term distributed according to the convolution of the lognormal and normal distributions (LNN). In paper IV we prove consistency and asymptotic normality for the MLE of the parameters of model (9).

6 Maximum spacing estimation

Consider an image classification problem based on remotely sensed data. In order for parametric classification methods such as linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) to perform well, it is essential that the assumption of normality for the data holds or that the deviations from it are small. For non-normal densities it is nevertheless possible to use maximum likelihood (ML) classifiers but with other models for the class distributions. For example, in Taxt et al. (1991) mixtures of normal distributions are used for classification of images. Their results indicate a substantial increase in correct classification rates compared to classification under normal densities. Mixture models are examples of models when traditional estimation methods such as ML have a tendency to fail. It has been observed that in remotely sensed data feature vectors quite frequently possess bimodal or multimodal empirical distributions. If the training sets are objectively selected the empirical distributions of this kind will become even more common. Therefore general estimation methods for both the univariate and multivariate class distributions, which give efficient and robust estimates also when traditional methods break down, are of fundamental importance for parametric classification.

The maximum spacing method (MSP) is a general method of estimating continuous distributions and is an alternative to ML method. The MSP method was proposed by Cheng and Amin (1983) (in the name "the maximum product of spacings"), and independently by Ranneby (1984). The argument in Cheng and Amin (1983) was that the maximum of $(n+1)^{-1}\sum\log\{(n+1)V_i\}$ (the V_i's representing the spacings $F_{\boldsymbol{\theta}}(X_{(i+1)}) - F_{\boldsymbol{\theta}}(X_{(i)})$), under the constraint $\sum V_i = 1$, is obtained if and only if all the V_i's are equal. Note that by setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the V_i's become identically distributed, i.e., the uniform spacings $F_{\boldsymbol{\theta}_0}(X_{(i+1)}) - F_{\boldsymbol{\theta}_0}(X_{(i)})$ should become "nearly equal". Ranneby (1984) derived the MSP method from an approximation of the Kullback-Leibler information (recall that the ML method also can be derived from an approximation of the Kullback-Leibler information). In Titterington (1985) it was observed that the MSP method can be regarded as an ML approach based on grouped data. Having indicated the similarities between ML and MSP, we note that there are many situations in which the MSP method works better than the ML method (see example below). Moreover, attractive properties such as consistency and asymptotic efficiency of the maximum spacing estimator (MSPE) closely parallel those of the maximum likelihood estimator (MLE).

Let $F_{\boldsymbol{\theta}}(x)$, where the unknown parameter vector $\boldsymbol{\theta}$ is contained in the parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}^q$, denote a family of continuous univariate distribution functions. Let $X_1, ..., X_n$ be independent identically distributed (i.i.d.) random variables with distribution function $F_{\boldsymbol{\theta}_0}(x)$, and denote the corresponding order statistics by $-\infty \equiv X_{(0)} \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} \equiv \infty$. Define

$$S_n(\boldsymbol{\theta}) = \frac{1}{n+1} \sum_{i=0}^{n} \log \left\{ (n+1) \left(F_{\boldsymbol{\theta}}(X_{(i+1)}) - F_{\boldsymbol{\theta}}(X_{(i)}) \right) \right\}.$$

The function $S_n(\boldsymbol{\theta})$ can be understood as an analogue to the log-likelihood function.

Definition 2 *Any $\hat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ which maximizes $S_n(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$ is called a maximum spacing estimator of the unknown true parameter vector $\boldsymbol{\theta}_0$.*

Example 2 (Ranneby (1984)). Let

$$F_{\boldsymbol{\theta}}(x) = \frac{1}{2}\Phi(x) + \frac{1}{2}\Phi\left(\frac{x-\mu}{\sigma}\right), \qquad \boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+,$$

where $\Phi(x)$ is the standard normal distribution function, and let $X_1, ..., X_n$ be i.i.d. from $F_{\boldsymbol{\theta}_0}(x)$, $\boldsymbol{\theta}_0 = (\mu_0, \sigma_0) \in R \times R^+$. Then the MLE of $\boldsymbol{\theta}_0$ does not exist, since the likelihood function of an observed sample $x_1, ..., x_n$ approaches infinity as, for example, $\mu = x_1$ and $\sigma \downarrow 0$. However, any approximate MSPE $\theta_n^* \in \Theta$ defined by

$$S_n(\boldsymbol{\theta}_n^*) \geq -c_n + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} S_n(\boldsymbol{\theta})$$

where $0 < c_n$ and $c_n \to 0$ as $n \to \infty$, is a consistent estimator of $\boldsymbol{\theta}_0$.

General consistency theorems of (approximate) MPSEs are given in Ekström (1996) and Shao and Hahn (1999), among others. These theorems cover many situations in which the ML method fails.

Asymptotic normality theorems for the MSPE $\hat{\boldsymbol{\theta}}_n$, i.e., that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} N(0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1}),$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix, have been given by several authors, among them Cheng and Stephens (1989) and Shao and Hahn (1994). The conditions used in both these papers are similar to those given in Cramér (1946) for the MLE. Because of the form of the asymptotic covariance matrix $\mathbf{I}(\boldsymbol{\theta}_0)^{-1}$, the estimator $\hat{\boldsymbol{\theta}}_n$ is generally regarded as an asymptotically efficient estimator of $\boldsymbol{\theta}_0$. For results on efficiency of MSPE see Ghosh and Jammalamadaka (2001).

In present thesis transition of maximum spacing estimator to dimensions higher than 1 is investigated. Since there is no natural order relation in $\mathbb{R}^d$ when $d > 1$ the approach has to be modified. Essentially, there are two different possibilities for the transition, the geometric and probabilistic counterparts to the univariate case. If we to each observation attach its Dirichlet cell, the geometrical correspondence is obtained. The probabilistic counterpart would be to use the nearest neighbor balls. Let us consider these two possibilities.

6.1 Dirichlet tesselation

Let $\xi_1, \xi_2, \ldots, \xi_n$ be i.i.d. d-dimensional random vectors with an absolutely continuous distribution P_0 with density function $g(x)$ and suppose that we assign a model with density functions $\{f(x, \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta} \subset \mathbb{R}^{\mathbf{q}}$.

Consider the following definitions. Given an open set $\Omega \subset \mathbb{R}^d$, the set $\{V_i\}_{i=1}^n$ is called a *tesselation* of Ω if $V_i \cap V_j = \oslash$ for $i \neq j$ and $\cup_{i=1}^n \overline{V}_i = \overline{\Omega}$. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^d$. Given a set of points $\{\xi_i\}_{i=1}^n$ belonging to $\overline{\Omega}$, the Dirichlet cell $V(\xi_i)$ corresponding to the point ξ_i is defined by

$$V(\xi_i) = \{y \in \Omega : |y - \xi_i| \leq |y - \xi_j|, \text{for } j = 1, \ldots, n, j \neq i\}.$$

The probabilities of Dirichlet cells, of course, always add up to one. Therefore it is possible to consider the following alternative definition of the spacing function based on the Dirichlet tesselation. Let

$$v_i(n, \boldsymbol{\theta}) = nP_{\boldsymbol{\theta}}(V(\xi_i)).$$

$$S_n^v(\boldsymbol{\theta}) \quad = \quad \frac{1}{n} \sum \log(v_i(n, \boldsymbol{\theta})).$$

The MSPE of $\boldsymbol{\theta}$ is defined as the maximizer of $S_n^v(\boldsymbol{\theta})$.

6.2 Nearest neighbor balls

Another way to define the spacing function in dimensions higher than 1 is through the nearest neighbor balls.

Let

$$B(\xi_i, r) = \{y : |\xi_i - y| \leq r\}$$

denote the ball of radius r with center at ξ_i and let

$$R_n(i) = \min_{j \neq i} |\xi_i - \xi_j|.$$

The latter quantity is called the nearest neighbor distance for ξ_i. Define,

$$z_i(n, \boldsymbol{\theta}) \quad = \quad nP_{\boldsymbol{\theta}}(B(\xi_i, R_n(i))).$$

A natural generalization of the univariate definition of the spacing function is to define it in multivariate case as

$$\frac{1}{n}\sum_{i=1}^{n}\log z_i(n,\boldsymbol{\theta}).$$

However, this approach has serious shortcomings, mainly that under some probability measures the sum of the probabilities of the nearest neighbor balls may become too large (as for instance when $P_{\boldsymbol{\theta}}$ has the same location as P_0 but much smaller variance). As a consequence there is no guarantee that the estimator will be consistent. To overcome this problem, the probabilities for the nearest neighbor balls when the sum of their probabilities exceeds one, have to be normalized. When the sum is less than one the remaining probability enters the spacing function in the same way as the probabilities for the nearest neighbor balls. The MSPE of $\boldsymbol{\theta}$ is then defined as the maximizer of $S_n^b(\boldsymbol{\theta})$, where

$$\begin{aligned} S_n^b(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\log(z_i(n,\boldsymbol{\theta})) \\ &\quad + \left(\frac{1}{n}\log(1-\frac{1}{n}\sum_{i=1}^{n}z_i(n,\boldsymbol{\theta}))\right) I(\frac{1}{n}\sum_{i=1}^{n}z_i(n,\boldsymbol{\theta})\le 1) \\ &\quad - I(\frac{1}{n}\sum_{i=1}^{n}z_i(n,\boldsymbol{\theta})>1)\log(\frac{1}{n}\sum_{i=1}^{n}z_i(n,\boldsymbol{\theta})), \end{aligned}$$

where $I(A)$ is the indicator function of the set A.

In paper V we investigate the MSPE based on Dirichlet tesselation. We prove the consistency of such estimator in $\mathbb{R}^1$ and conjecture that the consistency holds in higher dimensions. The simulation studies we conducted in $\mathbb{R}^2$ support our assumption of consistency and also that of asymptotic efficiency. We also compare the efficiency of the MSPE based on Dirichlet tesselation with that of the MSPE based on nearest neighbor balls and note that the former is much closer to being efficient.

7 Random search optimization

Optimization by the random search is the method we used in papers IV and V in order to find the maxima of complicated functions in multidimensional spaces. This methods and its modifications are extensively used in nondifferentiable problems, or when computation of derivatives is prohibitive due to its complexity. The main idea of random search optimization is very simple and is essentially the restatement of the "trial-and-error" method. Suppose, we seek the maximum of the object function F in the parameter space $\mathcal{P}$. The value of F is calculated in a reasonably chosen starting point $P_0 \in \mathcal{P}$. As is usual with optimization problems, the good choice of the starting point significantly speeds up the convergence. Then a random (according to some specified distribution) step in the parameter space is made, producing another point P_1. If $P_1 \notin \mathcal{P}$ another step is made from P_0, otherwise $F(P_1)$ is compared to $F(P_0)$. If $F(P_1) > F(P_0)$, the next step is made from P_1. The next step (also random) can differ from the previous one, making an algorithm adaptive, i.e., dependent on the previous results. The algorithm continues until some sort of stability is obtained. Numerous modifications of this simple idea, aimed to speed up the convergence, exist (see, for example, Lee and Rhinehart (1998) and Lei (2002)).

Obviously, in order for the random optimization procedure to converge in finite time, it is necessary to impose some inequality-type restrictions (i.e., bounds) on the parameters.

In papers IV and V we chose the random steps as follows. The first steps in the parameter space were one-dimensional Gaussian for each parameter. After several thousands Gaussian steps the value of the MLE stabilized. One could suppose that the obtained MLE was close to the real one (given the restrictions imposed on parameter values). Then the uniform distribution was employed to search in the m-dimensional cubic neighborhood of the current MLE (m here is the size of the parameter vector). The range of the uniform distribution was gradually narrowed, until the the required accuracy was achieved. All the computations were carried out using the resources of High Performance Computing Center North (HPC2N).

8 Summary of paper I

In this paper the problem of estimating the spatial distribution of reindeer (*Rangifer tarandus*) from incomplete survey is considered. The standard model used to describe the presence/absence of species is the logistic regression. In order to obtain a better fit of the regression function one must also consider the intrinsic spatial autocorrelation arising in gregarious populations. This can be achieved by including extra covariates, which describe the spatial covariation, explicitly in the model. The logistic model with such spatial covariates included is known as autologistic model. Often it is the case that the ground survey is incomplete, i.e., it has gaps which do not allow us to calculate the spatial covariation. In such situations the Gibbs sampler can be applied to estimate presence/absence of the species at unsurveyed plots. When it is done, the autologistic regression function is fitted in the usual way. We consider the question how much the results of the logistic regression can be improved by simulating the unsurveyed data by the Gibbs sampler. Various choices of training sets for obtaining the initial estimates by logistic regression are also studied.

The results indicate that the Gibbs sampler helped to improve the accuracy of prediction by 2-3% on average. Another conclusion that we have drawn was that the size of the training set was of much less importance compared to its positioning. The best results were obtained when the training set was connected, and chosen perpendicular to the elevation curves.

9 Summary of paper II

In paper II we consider a problem of automatic detection of tracks in aerial photos or satellite images. We adopt a Bayesian approach and base our inference on an *apriori* knowledge of the structure of tracks. The probability of a pixel to belong to a track depends on how the pixel gray level differs from the gray levels of pixels in the neighborhood and on additional prior information about the shape of the tracks. The Gibbs sampler is used to construct the most probable (according to the posterior distribution) configuration of the tracks. We apply the algorithm to aerial photos with resolution of 1 meter. Even for detection of tracks of width which is comparable with or smaller than the resolution, positive results are achieved. We study several examples of varying complexity, and suggest how to construct the Gibbs sampler (i.e., the prior distribution) for each particular case.

10 Summary of paper III

In this paper we analyze how such MCMC methods as the Iterated Conditional Mode (ICM) and the Gibbs sampler can be applied for improving the classification of multispectral images by conventional methods. The paper describes a simulation study which was performed as follows. First a “true” image was simulated from the ground truth. This “true” image was degraded by autocorrelated noise and then classified into three classes by noncontextual methods. The nonparametric rule called k-NN (Fix and Hodges (1951)) for $k = 1$ and 3, as well as the quadratic discriminant analysis (QDA) were considered. The outputs of these algorithms were used as the inputs for the ICM and the Gibbs sampler, which reclassify the map according to the posterior distribution (with Potts prior). It is shown that the use of the Gibbs sampler for reclassification is particularly justified if the initial classification performed poorly. A significant improvement in total misclassification rates is obtained.

11 Summary of paper IV

The forest growth model with measurement errors is introduced. The model is an exponential regression function with a heteroscedastic error term which is distributed as a convolution of normal and lognormal distributions. The model takes into account the fact that the measurement error is affected by the size of the tree. The maximum likelihood estimates (MLE) of the parameters of this model are proven to be consistent and asymptotically normally distributed. The model is applied to the real data from Swedish National Forest Inventory and the MLE of the parameters are calculated.

12 Summary of paper V

A maximum spacing (MSP) estimate based on Dirichlet tesselation is studied. For univariate observations consistency of such estimate is proved. For multivariate observations the asymptotic properties of a maximum spacing estimate based on Dirichlet tesselation are investigated by a simulation study. The results complement those obtained in Ranneby and Jammalamadaka (2002) in the following way. Whereas in Ranneby and Jammalamadaka (2002) the MSP estimation is extended to multivariate observations by means of spacings which are the nearest neighbor balls, in our paper we take an alternative approach to the same problem. By means of the simulation study we directly compare the asymptotic properties of the MSP estimate based on nearest neighbor balls with those of the MSP estimate based on the Dirichlet tesselation. The results indicate that the variance of the estimate is lower when spacings are based on the Dirichlet tesselation.

References

[1] Besag, J. E. (1986). On the statistical analysis of dirty pictures. *J. R. Statist. Soc.*, **B**, 48, pp. 259-302.

[2] Borovkov, A.A., (1998). *Mathematical Statistics.* Gordon and Breach Science Publishers.

[3] Bradley, R.A. and Gart, J.J. (1962). The asymptotic properties of ML estimators when sampling from associated populations *Biometrica*, **49**, pp. 205-214.

[4] Brooks, S.P. and Roberts, G.O. (1999). On quantile estimation and Markov chain Monte Carlo convergence. *Biometrika*, **86**, pp. 710-717.

[5] Cheng, R.C.H. and Amin, N.A.K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *J. R. Statist. Soc.*, **B**, 45, pp. 394-403.

[6] Cheng, R.C.H. and Stephens, M.A. (1989). A goodness-of-fit test using Moran's statistic with estimated parameters. *Biometrika*, **76**, pp. 385-392.

[7] Cramér, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N.J.

[8] Descombes, X., Morris, S.D., Zerubia, J. and Berthod, M. (1999). Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood. *IEEE transactions on image processing* **8**, 7, pp. 954-963.

[9] Ekström, M. (1996). Strong consistency of the maximum spacing estimate. *Theory Probab. Math. Statist.*, **55**.

[10] Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155-160.

[11] Fix, E. and Hodges, J.L. (1951). Discriminatory analysis - nonparametric discrimination: consistency properties. Report no. 4, US Air Force School of Aviation Medicine, Randolph Field, Texas.

[12] Frigessi, A., di Stefano, P., Hwang, C-R. and Sheu, S-J. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. R. Statist. Soc.* **B**, 55, pp. 205-219.

[13] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE transactions on pattern analysis and machine intelligence*, PAMI-6, November, pp. 721-741.

[14] Geman, D., Geman, S., Graffigne, C. and Dong, P. (1990). Boundary detection by constrained optimization. *IEEE transactions on pattern analysis and machine intelligence*, Vol.12, No.7, pp. 609-628.

[15] Ghosh, K. and Jammalamadaka, S. R. (2001). A general estimation method using spacings. *Journal of Statistical Planning and Inference*, **93**, pp. 71-82.

[16] Häggström, O. and Nelander, K. (1999). On exact simulation of Markov random fields using coupling from the past. *Scand. J. Stat.*, Vol.12, pp. 395-411.

[17] Hoadley, B. (1971). Asymptotic properties of ML estimators for independent not identically distributed case. *The Annals of Mathematical Statistics*, **42**, pp. 1977-1991.

[18] Lee, J. and Rhinehart, R. (1998). Heuristic random optimization. *Comput. Chem. Eng.,* **22**, pp. 427-444.

[19] Lei, G. (2002). Adaptive random search in quasi-Monte Carlo methods for global optimization. *Comput. Math. Appl.*, **43**, pp. 747-754.

[20] Meng, X.-L. (2000). Towards a more general Propp-Wilson algorithm: multistage backward coupling. *Monte-Carlo methods. Proceedings of the Workshop on Monte Carlo Methods held at The Fields Institute for Research in Mathematical Sciences.*

[21] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, pp. 1087-1092.

[22] Propp, J.G. and Wilson, D.B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random structures and algorithms*, **9**, pp. 223-252.

[23] Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scand. J. Statist.*, **11**, pp. 93-112.

[24] Ranneby, B. and Jammalamadaka, S.R. (2002). The maximum spacing estimate for multivariate observations. Invited presentation at IISA Fourth Biennial International Conference on Statistics, Probability and Related Areas, Northern Illinois University, DeKalb, Illinois, June 14-16, 2002.

[25] Roussas, G. (1968). Asymptotic normality of the maximum likelihood estimate in Markov processes. *Metrika*, **14**, pp. 62-70.

[26] Shao, Y. and Hahn, M.G. (1994). Maximum spacing estimates: a generalization and improvement of maximum likelihood estimates I. *Probability in Banach Spaces*, **9**, pp. 417-431. Birkhauser, Boston.

[27] Shao, Y. and Hahn, M.G. (1999). Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Ann. Inst. Statist. Math.*, **51**, pp. 31-49.

[28] Smith, A. F. M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.*, **B**, 55, pp. 3-23.

[29] Taxt, T., Lid Hjort, N. and Eikvil, L. (1991). Statistical classification using a linear mixture of two multinormal probability densities. *Pattern Recognition Letters*, Vol. 12, pp. 731-737.

[30] Titterington, D.M. (1985). Comment on "estimating parameters in continuous univariate distribution". *J. R. Statist. Soc.*, **B**, 47, pp. 115-116.

[31] Winkler, G. (2003). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: A Mathematical Introduction*, Springer-Verlag.

Acknowledgements

I am very grateful to Professor Bo Ranneby for excellent supervision and for improving all my drafts. His enthusiasm and scientific experience are the driving force behind this work.

This thesis would not be possible without the careful guidance of my co-supervisor Associate Professor Jun Yu. It is impossible to over-estimate the influence of his personal involvement and his optimism on this work.

I am deeply indebted to my co-supervisor Professor Dmitrii Silvestrov whose professional help was very important for me during the whole course of my studies.

Special thanks go to Associate Professor Lars Edenius and Professor Björn Elfving, collaborations with whom I have greatly enjoyed.

I would like to thank my colleagues and friends from the Centre of Biostochastics - Professor Yuri Belyaev for professional assistance and advice, Associate Professor Magnus Ekström for scientific discussions and for improving my drafts and Fredrik Norström. Working with them was always a pleasure.

This research was partly conducted using the resources of High Performance Computing Center North (HPC2N).

I would like to thank the European Union, the Ariana research team in INRIA (Institut National de Recherche de Informatique et en Automatique), France and personally the head of the Ariana team Professor Josiane Zerubia.

Finally, I thank the Swedish Foundation for Strategic Environmental Research, MISTRA, for financial support.