



LANGEVIN AND HESSIAN WITH FISHER APPROXIMATION STOCHASTIC SAMPLING FOR PARAMETER ESTIMATION OF STRUCTURED COVARIANCE

Cornelia VACAR, Jean-François GIOVANNELLI, Yannick BERTHOUMIEU

Université de Bordeaux, UB1, IPB, ENSEIRB-Matmeca,
Laboratoire IMS UMR 5218 Groupe Signal et Image,
351 cours de la Libération 33405 Talence, France

ABSTRACT

We have studied two efficient sampling methods, Langevin and Hessian adapted Metropolis Hastings (MH), applied to a parameter estimation problem of the mathematical model (Lorentzian, Laplacian, Gaussian) that describes the Power Spectral Density (PSD) of a texture. The novelty brought by this paper consists in the exploration of textured images modeled by centered, stationary Gaussian fields using directional stochastic sampling methods. Our main contribution is the study of the behavior of the previously mentioned two samplers and the improvement of the Hessian MH method by using the Fisher information matrix instead of the Hessian to increase the stability of the algorithm and the computational speed.

The directional methods yield superior performances as compared to the more popular Independent and standard Random Walk MH for the PSD described by the three models, but can easily be adapted to any target law respecting the differentiability constraint. The Fisher MH produces the best results as it combines the advantages of the Hessian, *i.e.*, approaches the most probable regions of the target in a single iteration, and of the Langevin MH, as it requires only first order derivative computations.

Index Terms— Stochastic sampling, Monte Carlo Markov Chains, Hessian, Fisher, texture.

1. INTRODUCTION

The sampling methods prove to be important tools in signal and image processing, as they enable the exploration of intricate target laws and permit their characterization by computing their statistical descriptors based on the drawn samples.

The performances of these algorithms directly influence the performances of the algorithms that embed them and this is why the speed and the accuracy with which the resulting samples describe the target are very important. We have investigated the performances of two stochastic samplers, Langevin and Hessian adapted MH, in the context of a parameter estimation problem for textured images.

First order derivative-based samplers have also been employed in [1] and in the recent work [2], but these approaches explore the Hamiltonian MH algorithm, whilst we have studied the Langevin MH. However, our contribution is more significant in the case of the second order derivative-based methods, where, in the present estimation problem, we have provided an improvement, by modifying the Hessian algorithm presented in [3] in the sense that the concerns regarding the positiveness of the Hessian matrix are eliminated by its replacement with the Fisher information matrix. The results obtained for the proposed problem are encouraging, confirming the stable behavior and speed performances of the second order derivative-based samplers.

2. BAYESIAN SETTING

Although the sampling algorithms are able to explore any type of target, not necessarily a probability density, we will establish from the beginning a Bayesian context of a parameter estimation problem. In this setting, we will use the following notations throughout the paper: \mathbf{x} is the set of observed data, $\boldsymbol{\theta}$ is the set of parameters to be estimated, $f(\mathbf{x}|\boldsymbol{\theta})$ is the *likelihood*, describing the probability of a certain observed data configuration given the parameters, and π represents the target we intend to sample.

An MH algorithm, as described by [4], [5], relies on a transition kernel consisting of two ingredients: an arbitrary transition kernel, $q(\boldsymbol{\theta}, \phi)$, and an acceptance probability:

$$\alpha(\boldsymbol{\theta}, \phi) = \min \left\{ 1, \frac{\pi(\phi)q(\phi, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \phi)} \right\} \quad (1)$$

where $\boldsymbol{\theta}$ is the current and ϕ is the proposed value for the parameters. The process thus consists in formulating a proposal according to the transition kernel, evaluating the acceptance probability of the proposal and deciding to accept or reject it based on the computed probability.

The functioning of an MH algorithm is:

1. Initialize the iteration counter $j = 1$ and set $\boldsymbol{\theta}^{(0)}$.
2. Propose a new value ϕ for the parameter, generated from the density $q(\boldsymbol{\theta}^{(j-1)}, \cdot)$.
3. Evaluate the acceptance probability $\alpha(\boldsymbol{\theta}^{(j-1)}, \phi)$ given by Eq.(1) and according to this value accept the proposal and update the parameter $\boldsymbol{\theta}^{(j)} = \phi$, or reject it and keep the old value for the parameter $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$.
4. Update the counter and return to step 2 until convergence.

All analyzed samplers are convergent and accurately explore the target, the differences between them being given by the different formulation of the transition kernel and, implicitly, by the expression of the acceptance probability. These differences translate into different evolutions of the samples when exploring the target (and how quickly the samples begin to be drawn from the most representative regions) and in differences between the time needed by each sampler to produce a sample.

Independent MH has a proposition law that does not depend on the current position of the chain, Random Walk (RW) MH proposes an isotropic displacement around the current value, while more advanced methods include a directional component for the proposal, meaning that the proposal is built as follows: add to the current value the directional component and then add the stochastic component, *i.e.*, make an isotropic move around this new value. In the class of directional MH methods lie the Hamiltonian, Langevin and Hessian

adapted MH, the last two algorithms being those that were studied here, with an emphasis on the Hessian method, which has been improved by building a new transition kernel based on the Fisher information matrix.

In the following we will present these algorithms, adapted to our parameter estimation problem. In a Bayesian parameter estimation algorithm the *a posteriori* law for the parameters, given the observations, is the law that dictates the probability distribution law in the parameter space. Thus, this is the law we should explore, *i.e.*, the target, and extract the statistics from the yielding samples. According to Bayes' law, the *a posteriori* distribution is proportional to the product between the *a priori* distribution and the likelihood. With the pre-requisite that the *a priori* will be embedded in the proposition law, the likelihood accurately approximates the posterior distribution law, and thus, we will sample, the Co-Log-Posterior (*CLP*), where $CLP = -\log f(\mathbf{x}|\boldsymbol{\theta})$, will be evaluated as being the target in order to achieve the convexity of the target with respect to the parameters.

3. SAMPLING ALGORITHMS

The principle of stochastic methods is the sampling of a distribution by considering it as being the limiting distribution of a Markov chain and simulating this chain until equilibrium is reached, as stated in [6]. Furthermore, the class of MCMC methods has been narrowed to the Metropolis Hastings (MH) family, due to its adaptation to sampling intricate target laws.

3.1. Langevin Adapted Metropolis Hastings

Langevin algorithms are derived from diffusion approximations and rely on the principle of using the information concerning the target density, in the format $\nabla \log \pi$, in order to build a proposal distribution well-adapted to the problem in question [7]. The Langevin-based MH algorithm proposes a random walk-like transition of the form [8]:

$$\boldsymbol{\theta}_p = \boldsymbol{\theta}_c + \frac{\varepsilon^2}{2} \mathbf{g}(\boldsymbol{\theta}_c) + \varepsilon \mathcal{N}(0, I) \quad (2)$$

where $\mathbf{g}(\boldsymbol{\theta}_c) = \nabla CLP(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_c}$. The acceptance probability can be obtained from (1), for:

$$q(\boldsymbol{\theta}_p, \boldsymbol{\theta}_c) = \exp \left[-\frac{1}{2\varepsilon^2} \|\boldsymbol{\theta}_c - \boldsymbol{\theta}_p - \frac{\varepsilon^2}{2} \mathbf{g}(\boldsymbol{\theta}_p)\|^2 \right] \quad (3)$$

As compared to the non-directional MH methods, the complexity of the Langevin MH is increased due to the form of the acceptance probability and to the necessity of evaluating the gradient for every new proposal. However, the increased computation time per iteration is compensated by the much smaller number of iterations needed to reach convergence.

In regions far from the maximum of probability, the gradient is large (the directional component is dominant), thus the algorithm approaches the high probability regions with high amplitude jumps, while, when near the maximum of probability, the gradient is small, thus the stochastic component is dominant and the region of high probability is explored due to the stochastic component.

3.2. Hessian Adapted Metropolis Hastings

This section is dedicated to a sampling method that has seldom been explored and whose presence in the literature is scarce. The directional component of the proposal is in this case formulated using

Newton's direction, which, for a quadratic law, indicates the maximum. In [3] a version of this sampler has been tested and compared to methods such as Gibbs and optimal marginal data augmentation (DA) samplers on a probit regression problem, proving that the performances of this sampler are superior. The transition for the Hessian sampler is of the form:

$$\boldsymbol{\theta}_p = \boldsymbol{\theta}_c + \varepsilon \boldsymbol{\Sigma}(\boldsymbol{\theta}_c) \mathbf{g}(\boldsymbol{\theta}_c) + \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_c)) \quad (4)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_c) = -\mathbf{H}(\boldsymbol{\theta})^{-1}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_c}$. The acceptance probability will be obtained from (1), for:

$$q(\boldsymbol{\theta}_p, \boldsymbol{\theta}_c) = \mathcal{N}(\boldsymbol{\theta}_p - \boldsymbol{\theta}_c + \varepsilon \boldsymbol{\Sigma}(\boldsymbol{\theta}_c) \mathbf{g}(\boldsymbol{\theta}_c), \boldsymbol{\Sigma}(\boldsymbol{\theta}_c)) \quad (5)$$

The advantage of the method is that, for quadratic and quasi-quadratic distributions, the regions of high probability are approached in a very small number of iterations (ideally, a single one) and then explored with the contribution of the stochastic part of the proposition, a quadratic law of variance $\boldsymbol{\Sigma}(\boldsymbol{\theta}_c)$, which is an accurate approximation of the target. However, it is clear that this method is also rather complex, as each iteration translates in the computation of the gradient and the Hessian matrix and the evaluation of the acceptance probability.

3.3. Fisher Adapted Metropolis Hastings

Another concern is the need to perform the inversion of the Hessian and in order to avoid matrix inversion problems that may occur if the Hessian is not positive definite, the idea of replacing the Hessian by the Fisher information matrix has been explored in the present work.

The Fisher matrix quantifies the amount of information that the observations contain regarding the parameter $\boldsymbol{\theta}$. We have been able to apply such an approximation, as in the case of our problem, we have a great amount of independent observations, thus, a scenario close to the asymptotic case. In such a situation, the *a posteriori* distribution is consistent and normal of inverse variance equal to the Fisher matrix [9]. Thus, instead of computing the Hessian, the quantities:

$$\mathcal{I}_{pq}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}|\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial \theta_p \partial \theta_q} CLP(\boldsymbol{\theta}) \middle| \boldsymbol{\theta} \right] \quad (6)$$

are used in order to formulate the proposal and to evaluate the acceptance probability.

This approximation reduces the complexity of computations as, instead of computing the Hessian, when applying the expectation with respect to the data given the parameters, the term containing second order derivatives becomes null. This fact is due to the particular problem we are addressing, *i.e.*, the model chosen for our textures and to the fact that the coefficients in the Fourier domain have the specific distribution detailed in Section 4. The performances of the algorithm are enhanced by this approximation, as the time per iteration is reduced because no second order derivatives of the target must be computed and all is reduced to first order derivatives computations. The major advantage brought by this innovation is that matrix inversions are always possible and the algorithm is stable, no matter if there are particular cases when the Hessian is not positive-definite (case that is mentioned, but not dealt with in [3]).

4. TEXTURE CONTEXT

This section presents the application of the previously presented sampling methods to a parameter estimation problem. The principle is that, starting from texture images such as those in Fig.2, we estimate the parameters that characterize the features of the texture and

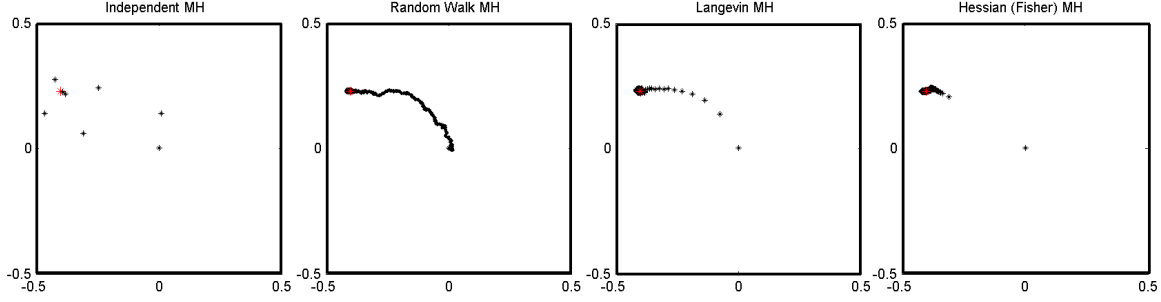


Fig. 1: Samples evolution in the $\nu_x^0 \nu_y^0$ parameter space for the four samplers. To the left, observe the sparseness of the accepted samples for Independent MH, for RW MH the evolution step is very small and undirected, while for Langevin MH the proposal is influenced by the gradient. For Hessian (Fisher) MH, the strong probability regions are approached in a single iteration and then thoroughly sampled, as these are the regions most representative for the target.

thus, reduce its description to a small number of numerical quantities. This approach is possible due to the particular nature of these textures, *i.e.*, to the fact that they can be modeled by zero-mean, stationary Gaussian Random Fields (GRF) of covariance matrix \mathbf{R} [10].

As seen in Fig.2, this method of modeling the textures can describe a broad class of images, *i.e.*, by the simple variation of a relatively reduced set of parameters, various patterns and characteristics appear.

Considering \mathbf{x} as being the N^2 pixel vector of such an image, the model is completely described by its second order statistics and a distribution law of the form:

$$f(\mathbf{x}|\mathbf{R}) = (2\pi)^{-N^2/2} \det(\mathbf{R})^{-1/2} \exp\left[-\frac{1}{2}\mathbf{x}^t \mathbf{R}^{-1} \mathbf{x}\right] \quad (7)$$

Working under the stationarity hypothesis, the \mathbf{R} matrix has a Toeplitz form and can be approximated by a circulant form $\mathbf{R} = \mathbf{F}^\dagger \mathbf{\Delta} \mathbf{F}$, where $\mathbf{\Delta}$ is a diagonal matrix containing the eigenvalues of \mathbf{R} and \mathbf{F} denotes the Fourier basis matrix of size $N \times N$. After further computation, Eq.(7) can be rewritten as:

$$f(\mathbf{x}|\mathbf{R}) \propto \prod_{n,m=1}^N s_{nm}^{-1/2} \exp\left[-\frac{1}{2} \sum_{n,m=1}^N |\hat{x}_{nm}|^2 / s_{nm}\right] \quad (8)$$

In Eq.(8), $|\hat{x}_{nm}|^2$ represent the squared moduli of the image's Fourier coefficients and s_{nm} are the discretized elements of the PSD. The principle of the model does not consist in a discretization of the PSD, which could lead to aliasing problems, but in the use of the PSD values at discrete positions in order to describe the law

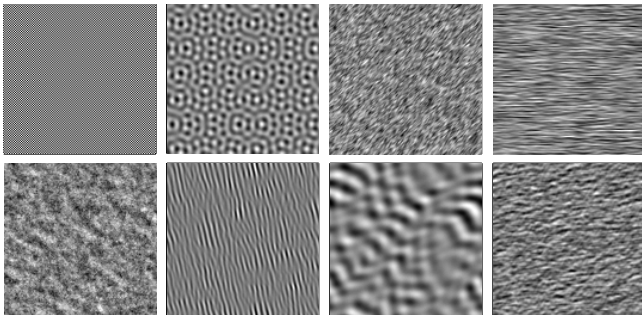


Fig. 2: Texture realizations using GRFs with structured PSDs.

governing the elements of the image in the Fourier domain. The previously mentioned form is achieved using a Whittle approximation that enables the evaluation of the *CLP* in $O(N \log N)$. In fact, the s_{nm} elements dictate the distributions of the image's Fourier transform coefficients in the following manner:

$$\hat{x}_{nm} \sim \mathcal{N}(0, s_{nm}) \quad (9)$$

This means that each Fourier coefficient \hat{x}_{nm} has a zero-mean, normal distribution, of variance s_{nm} , thus we are in a scenario of independent, but not identically distributed observations.

The PSD of the texture can be described by a wide variety of positive-valued mathematical laws and the s_{nm} quantities that intervene in Eq.(8) follow these laws. Our study is focused on the Lorentzian, Laplacian and Gaussian functions and in the following the Lorentzian form will be used to exemplify the computations:

$$S_L(\nu_x, \nu_y, \boldsymbol{\theta}) = \frac{\gamma}{1 + (\nu_x - \nu_x^0)^2 w_x^2 + (\nu_y - \nu_y^0)^2 w_y^2} \quad (10)$$

where $\boldsymbol{\theta} = [\gamma, \nu_x^0, \nu_y^0, w_x, w_y] \in \mathbb{R}^5$ is the parameter set that drives the PSD and $S_L(\nu_x, \nu_y, \boldsymbol{\theta})$ denotes the continuous PSD at position (ν_x, ν_y) for the current value $\boldsymbol{\theta}$ of the parameter set. Furthermore, the previously mentioned s_{nm} elements are functions of $\boldsymbol{\theta}$ and represent the value of the PSD at discrete positions in the frequency domain, as shown by the expression:

$$s_{nm}(\boldsymbol{\theta}) = S_L(n\Delta\nu_x, m\Delta\nu_y, \boldsymbol{\theta}) \quad (11)$$

Eq.(11) illustrates the fact that the s_{nm} elements contain the contribution of the parameter set $\boldsymbol{\theta}$ through the mathematical model for the PSD. Furthermore, as shown by Eq.(9), these elements of the PSD govern the distribution of the texture's Fourier transform. It is thus obvious that the law of dependency of \hat{x}_{nm} on $\boldsymbol{\theta}$ is cumbersome, for which reason stochastic methods were employed.

As mentioned in the previous section, the goal of the sampling algorithms is to properly explore the *a posteriori* law. The directional character of the analyzed samplers would be fully exploited if the *CLP* had specific convexity properties, and this is the reason for which the change of parametrization $\lambda_{nm}(\boldsymbol{\theta}) = s_{nm}(\boldsymbol{\theta})^{-1}$ is proposed in order to render the target law convex with respect to the elements of the PSD. Consequently, the expression of the *CLP*, convex with respect to $\lambda_{nm}(\boldsymbol{\theta})$, is:

$$CLP(\boldsymbol{\theta}) = \frac{1}{2} \sum_{n,m=1}^N \left(\log \frac{1}{\lambda_{nm}(\boldsymbol{\theta})} + |\hat{x}_{nm}|^2 \lambda_{nm}(\boldsymbol{\theta}) \right) \quad (12)$$

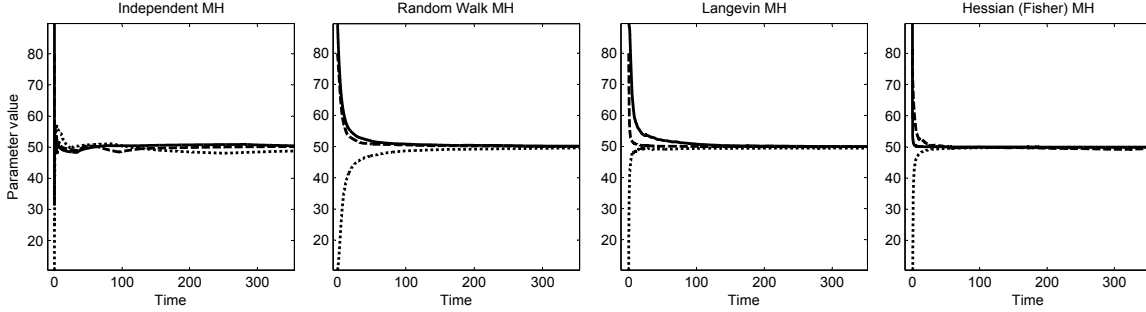


Fig. 3: The stabilization of the recursive means for several chains in the case of each of the four studied methods. As expected, all the algorithms converge to the same value, but the Hessian (Fisher) MH reaches equilibrium the fastest. Although superior in terms of computation speed per iteration, the Independent MH and Random Walk MH require a longer interval to converge.

The algorithms presented in Section 2 employ the first and the second order derivatives of the *CLP*, described by:

$$\frac{\partial CLP}{\partial \theta_p} = \sum_{n,m=1}^N \left(|\hat{x}_{nm}|^2 - \frac{1}{\lambda_{nm}(\theta)} \right) \frac{\partial \lambda_{nm}(\theta)}{\partial \theta_p} \quad (13)$$

$$\begin{aligned} \frac{\partial^2 CLP}{\partial \theta_p \partial \theta_q} &= \sum_{n,m=1}^N \frac{1}{\lambda_{nm}^2(\theta)} \frac{\partial \lambda_{nm}(\theta)}{\partial \theta_p} \frac{\partial \lambda_{nm}(\theta)}{\partial \theta_q} + \\ &\sum_{n,m=1}^N \left(|\hat{x}_{nm}|^2 - \frac{1}{\lambda_{nm}(\theta)} \right) \frac{\partial^2 \lambda_{nm}(\theta)}{\partial \theta_p \partial \theta_q} \end{aligned} \quad (14)$$

5. RESULTS

The sampling methods presented in this paper are all convergent and the use of one or another influences only the speed of convergence and the mixing properties of the algorithm, not the actual equilibrium state.

As the theory announced, all methods yield *posterior* histograms showing that the samplers explore the law and produce samples with the same *posterior* mean and variance. Consequently, in order to compare the methods, the convergence time and the computational cost is analyzed for each sampler. The improvement brought by the directional methods consists in the fact that they approach the region of important probability in a small number of steps, proposing transitions in the direction of probability increase and all at the same time they thoroughly explore the regions most representative for the target.

In Fig.3 it can be seen that, even if it requires the repeated evaluation of gradients, the Fisher RW has speed performances superior to RW MH and better than those of Langevin RW. The parameter ε that intervenes in all the tested methods was tuned in such a manner as to obtain an acceptance rate of approximately 24%, which in [7] is proven as being optimal in the sense that it provides the best compromise between the amplitude of the transitions (that ensures the samples are not too correlated) and the acceptance rate.

The advantage of the Fisher MH is given by the fact that it approaches the regions of high probability in a single iteration, requiring for this only the evaluation of gradients, as opposed to gradients and hessian matrices for Hessian MH. This approximation is possible due to our texture context that positions us in an asymptotic scenario of independent, but not identically distributed observations. This simplification brought by the method is given by the new form of the proposal, as the second term in Eq.(14) becomes null, due to the distribution of the Fourier coefficients in Eq.(9).

6. CONCLUSION AND PERSPECTIVES

The Fisher sampling method proved to be the best suited for the problem of parameter estimation in the case of the texture with a structured PSD, for several complex models. The presented directional sampling methods are also just as efficient in the case of any distribution law, with the condition for it to be twice differentiable.

One of the most interesting directions of pursuit is to further exploit the Fisher approximation and perform a more detailed analysis regarding the gain that such an approximation brings in stochastic sampling applications for other types of targets.

7. ACKNOWLEDGEMENTS

We would like to thank Olivier Féron and François Orieux, with whom we have discussed some of the aspects presented in this paper.

8. REFERENCES

- [1] K. Choo and D.J. Fleet, "People Tracking Using Hybrid Monte Carlo Filtering," *International Conference on Computer Vision*, vol. II, pp. 321–328, 2001.
- [2] K. Kayabol, E.E. Kuruouglu, J.L. Sanz, B. Sankur, E. Salerno, and D. Herranz, "Adaptive Langevin Sampler for Separation of *t*-Distribution Modelled Astrophysical Maps," *IEEE Transactions on Image Processing*, vol. 19, 2010.
- [3] Y. Qi and T.P. Minka, "Hessian-based Markov Chain Monte-Carlo Algorithms," *Workshop on Monte Carlo Methods*, 2002.
- [4] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, pp. 1087–1092, 1953.
- [5] W.K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.
- [6] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.
- [7] A. Gelman, G.O. Roberts, and W.R. Gilks, *Bayesian Statistics 5*, Oxford University Press, 1996.
- [8] J. Besag, P.J. Green, D. Higdon, and K.L. Mengersen, "Bayesian computation and stochastic systems (with discussion)," *Statistical Science*, vol. 10, pp. 3–66, 1995.
- [9] A.N. Philippou and G.G. Roussas, "Asymptotic Distribution of the Likelihood Function in the Independent not Identically Distributed Case," *Annals of Statistics*, vol. 1, pp. 454–471, 1973.
- [10] M. Haindl, "Parameter Estimation in Gaussian Markov Random Fields," *Research Report K335/1997/150*, 1997.