

# Hyperpriors for Matérn fields with applications in Bayesian inversion

January 8, 2018

LASSI ROININEN

Department of Mathematics, Imperial College London  
South Kensington Campus, London SW7 2AZ, United Kingdom

MARK GIROLAMI

Department of Mathematics, Imperial College London, and,  
Alan Turing Institute, London, United Kingdom

SARI LASANEN

University of Oulu  
Oulu, Finland

MARKKU MARKKANEN

Eigenor Corporation  
Lompolontie 1, FI-99600 Sodankylä, Finland

## Abstract

We introduce non-stationary Matérn field priors with stochastic partial differential equations, and construct correlation length-scaling with hyperpriors. We model both the hyperprior and the Matérn prior as continuous-parameter random fields. As hypermodels, we use Cauchy and Gaussian random fields, which we map suitably to a desired correlation length-scaling range. For computations, we discretise the models with finite difference methods. We consider the convergence of the discretised prior and posterior to the discretisation limit. We apply the developed methodology to certain interpolation and numerical differentiation problems, and show numerically that we can make Bayesian inversion which promotes competing constraints of smoothness and edge-preservation. For computing the conditional mean estimator of the posterior distribution, we use a combination of Gibbs and Metropolis-within-Gibbs sampling algorithms.

# 1 Introduction

In many Bayesian statistical estimation algorithms, the objective is to explore the posterior distribution of a continuous-parameter unknown  $v(x)$ ,  $x \in \mathbb{R}^d$ ,  $d = 1, 2, \dots$  given a direct or indirect noisy realisation  $y \in \mathbb{R}^M$  of the unknown  $v$  at fixed locations. Bayesian estimation algorithms are widely used for example in spatial statistics, Machine Learning and Bayesian statistical inverse problems and specific applications include e.g. remote sensing, medical imaging and ground prospecting [1, 4, 13, 16, 21, 25].

A priori information is a key factor in any Bayesian statistical estimation algorithm, as it is used to stabilise the posterior distribution. Gaussian processes and fields are common choices as priors, because of their analytical and computational properties, and because of their easy construction through the mean and covariance functions.

Gaussian priors are known to promote smooth estimates [13]. However, the smoothness of the unknown is inherently problem-specific. For example in many atmospheric remote sensing algorithms, the unknown is often assumed to be continuous [18]. On the other hand, in many subsurface imaging or in medical tomography applications, the unknown may have anisotropies, inhomogeneities, and even discontinuities [8].

A common method for expanding prior models outside the scope of Gaussian distributions is to apply hyperparametric models [5, 6]. In this paper, we study hypermodels for inhomogeneous Matérn fields, and our specific objective is to demonstrate that the proposed priors are flexible enough to promote both smooth and edge-preserving estimates.

Matérn fields, a class of Gaussian Markov random fields [16, 21, 24], are often defined as stationary Gaussian random fields, i.e. with them we can model isotropic unknown. By a simple change of variables, these fields can model also anisotropic features. Via modelling the Matérn fields with stochastic partial differential equations and by locally defining the correlation parameters, we can even construct inhomogeneous random fields [16, 22]. In addition, we can make them computationally very efficient, as the finite-dimensional approximation of the inverse covariance matrix, the precision matrix, is sparse by construction.

Let us denote by  $v^N$  the finite-dimensional approximation of the continuous-parameter random field  $v$ . The solution of a Bayesian estimation problem is a so-called posterior distribution. We give it as an unnormalised probability density

$$D(v^N|y) = \frac{D(v^N) D(y|v^N)}{D(y)} \propto D(v^N) D(y|v^N), \quad (1)$$

where the likelihood density  $D(y|v^N)$  is obtained e.g. through some physical observation system and the a priori density  $D(v^N)$  reflects our information of the unknown object before any actual measurement is done. We take  $v^N$  to be an approximation of a Matérn field.  $D(y)$  is a normalisation constant, which we often can omit from the analysis.

We model the Matérn field length-scaling as a continuous-parameter random

field  $\ell(x)$ , i.e. we construct a prior for certain parameters of the prior  $D(v^N)$ . By denoting the discrete approximation of the continuous-parameter field  $\ell$  by  $\ell^N$ , we may include the hyperprior into the posterior distribution (1), and hence write the posterior probability density as

$$D(v^N, \ell^N | y) \propto D(v^N, \ell^N) D(y | v^N) = D(\ell^N) D(v^N | \ell^N) D(y | v^N),$$

where  $D(v^N, \ell^N)$  is the prior constructed from the hyperprior  $D(\ell^N)$  and the prior itself is  $D(v^N | \ell^N)$ . The idea is similar to the length-scale modelling in [19, 20], but we will use using stochastic partial differential equations and sparse matrices, hence gaining computational advantages.

To simplify the sampling of  $\ell$ , we introduce an auxiliary random field  $u$ , and apply a further parametrisation  $\ell = \ell(x; u)$ . We choose to model  $u$  as a continuous-parameter Cauchy or Gaussian random field, but it could be also something else, for example an  $\alpha$ -stable random field [17, 26]. In this paper, we will put an emphasis on the discretisation of these hypermodels and the convergence of the discrete models to continuous models. From a computational point of view, we need to discuss Markov chain Monte Carlo (MCMC) methods, and in particular we will consider Gibbs and Metropolis-within-Gibbs sampling.

Modelling non-stationary Gaussian random fields and using them in e.g. interpolation is not a new idea, see for example Fuglstad et al. 2015 [9] for a comprehensive reference list. Other constructions of non-stationary Gaussian processes, given lately notable attention, include e.g. so-called deep learning algorithms, where an  $n$ -layer Gaussian process is formed in such a way that the deepest level is a stationary Gaussian process. These methods produce interesting priors from an engineering perspective. These methods, however, are typically lacking rigorous mathematical analysis [19, 20]. Also, many deep learning methods typically rely on full covariance matrices, hence computational burden might become a bottleneck, especially in high-dimensional problems.

We note that discontinuities are best modelled with specific non-Gaussian prior constructions. A common choice is a total variation prior, which promotes edge-preserving estimates. However, total variation priors are well-known to behave as Gaussian smoothness priors when the discretisation is made denser and denser (Lassas and Siltanen 2004) [15]. Constructions of proposed non-Gaussian priors, which do not converge to Gaussian fields in the discretisation limit, include the Besov space priors (Lassas et al. 2009) [14], which are constructed on a wavelet basis, hierarchical Mumford-Shah priors (Helin and Lassas, 2011) [11], and recently applied Cauchy priors (Markkanen et al. 2016) [17]. Construction of a Matérn style random field with non-Gaussian noise has been studied by Bolin 2014 [3]. These algorithms may work suitably well for example for edge-preserving Bayesian inversion, but they may lack the smoothness or limiting properties, which we aim to deploy in the proposed hypermodel construction.

The rest of this paper is organised as follows: In Section 2, we review the basics of linear Bayesian statistical estimation algorithms. In Section 3, we discuss continuous Matérn fields and construct certain hypermodels. In Section 4, we consider discretisation of the hypermodels, and, convergence of the discretised

models to the continuous models. We consider rough hypermodels in Section 5, and discuss Gibbs and Metropolis-within-Gibbs algorithms in Section 6. In Section 7, we show by a number of numerical examples how to use the constructed model in interpolation algorithms.

## 2 Linear Bayesian estimation problems

Let us consider a continuous-parameter linear statistical estimation problem

$$y = Av + e. \quad (2)$$

We assume that we know exactly one realisation of  $y$  and the linear mapping  $\mathcal{A}$  from some function space (e.g. separable Banach space) to a finite-dimensional space  $\mathbb{R}^M$ . We further assume that we know the statistical properties of the noise  $e$ . From now on, we assume that  $e$  is zero-mean Gaussian white noise statistically independent of  $v$ . We emphasise that we do not know the realisation of the noise  $e$ . Given these assumptions, our objective is to estimate the posterior distribution of  $v$ .

For computational Bayesian statistical estimation problems, we discretise Equation (2), and write it as a matrix equation

$$y = Av^N + e. \quad (3)$$

Likelihood probability, a factor of the posterior distribution (see Equation (1)), can then be given as

$$D(y|v^N) \propto \exp\left(-\frac{1}{2}(y - Av^N)^T \Sigma^{-1}(y - Av^N)\right),$$

where the Gaussian measurement noise  $e \sim \mathcal{N}(0, \Sigma)$ , and  $\Sigma$  is noise covariance matrix.

We start the preparations for the numerical sampling of the posterior for the hyperprior. First, we will consider the case when the unknown  $v^N$ , conditioned with the hyperparameter  $\ell^N$ , has Gaussian distribution  $\mathcal{N}(0, C)$ .

We can write the prior as a normalised probability density

$$D(v^N|\ell^N) = \frac{1}{\sqrt{(2\pi)^N |C|}} \exp\left(-\frac{1}{2}(v^N)^T C^{-1} v^N\right).$$

Here we have included the normalisation constant, as we aim to include the hyperparameters of  $v^N$  into the matrix  $C$ , i.e.  $|C| = |C(\ell^N)|$  is not a constant, for the different values of the hyperparameter, unlike the normalisation constant  $|\Sigma|$  in the likelihood density. Our aim is to decompose the prior inverse covariance, i.e. precision matrix, as  $C(\ell^N)^{-1} = (L(\ell^N))^T L(\ell^N)$ . This means that, similarly to Equation (2), we can present the prior also as an observation equation  $L(\ell^N)v^N = -w^N$ , where  $w^N \sim \mathcal{N}(0, I)$ . Hence, we can form a stacked matrix equation

$$\begin{pmatrix} A \\ L(\ell^N) \end{pmatrix} v^N + \begin{pmatrix} e \\ w^N \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}.$$

The benefit of this formulation is that we can easily make Gibbs sampling of the posterior distribution with this formulation, and hence to compute the posterior mean

$$v_{\text{CM}}^N = \frac{\int v^N D(v^N | \ell^N) D(\ell^N) D(y | v^N) dv^N d\ell^N}{\int D(v^N | \ell^N) D(\ell^N) D(y | v^N) dv^N d\ell^N}.$$

We can write similarly the estimate for the variable length-scale  $\ell$

$$\begin{aligned} \ell_{\text{CM}}^N &= \frac{\int \ell^N D(\ell^N) D(v^N | \ell^N) D(y | v^N) d\ell^N dv^N}{\int D(\ell^N) D(v^N | \ell^N) D(y | v^N) d\ell^N dv^N} \\ &= \frac{\int \ell^N(u) D(u) D(v^N | \ell^N(u)) D(y | v^N) du dv^N}{\int D(\ell^N(u)) D(v^N | \ell^N(u)) D(y | v^N) du dv^N}. \end{aligned}$$

For the estimation of  $\ell^N$ , we use the so-called Metropolis-within-Gibbs algorithm [17], where we make Gibbs type sampling, but component-wise, we make Metropolis-Hastings algorithm. Alternatives for Metropolis-within-Gibbs include e.g. pseudo-marginal approach to MCMC studied by Filippone and Girolami 2014 [10]. The estimation of  $v^N$  can be used with standard Gibbs sampling techniques.

### 3 Matérn field priors and hyperpriors

Matérn fields are often defined as stationary Gaussian random field with a covariance function

$$\text{Cov}(x, x') = \text{Cov}(x - x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{|x - x'|}{\ell} \right)^\nu K_\nu \left( \frac{|x - x'|}{\ell} \right), \quad x, x' \in \mathbb{R}^d, \quad (4)$$

where  $\nu > 0$  is the smoothness parameter, and  $K_\nu$  is modified Bessel function of the second kind or order  $\nu$ . The parameter  $\ell$  is called length-scaling. Correlation length, that is to say where correlation is 0.1, corresponds approximately  $\delta = \ell\sqrt{8\nu}$ . From now on, we suppose that the Matérn fields have a zero mean.

Let us recall the stochastic partial differential equation for the Matérn fields [16, 21]. The Fourier transform of the covariance function in Equation (4), gives a power spectrum

$$S(\xi) = \frac{2^d \pi^{d/2} \Gamma(\nu + d/2)}{\Gamma(\nu) \ell^{2\nu}} \left( \frac{1}{\ell^2} + |\xi|^2 \right)^{-(\nu + d/2)}.$$

As first mentioned by Rozanov 1977 [23], only fields with spectral density given by the reciprocal of a polynomial have a Markov representation. For our applications, we fix  $\nu = 2 - d/2$ .

If we let  $w$  be white noise, and by using 'hat'-notation for a Fourier-transformed object, then we may define the basic Matérn field  $v$  through the equation  $\hat{v} = \sigma \sqrt{S(\xi)} \hat{w}$  in the sense of distributions. By using inverse Fourier transforms, we may write a stochastic partial differential equation

$$(1 - \ell^2 \Delta) v = \sigma \sqrt{\ell^d} w.$$

Here we have an elliptic operator equation. We note that the constructed field  $v$  is isotropic, i.e. the field has constant correlation length-scaling  $\ell$  to every coordinate direction.

We modify the isotropic formulation to be inhomogeneous by allowing a spatially variable length-scaling field  $\ell(x)$ , for which we write a stochastic partial differential equation

$$(1 - \ell(x)^2 \Delta) v = \sigma \sqrt{\ell(x)^d} w. \quad (5)$$

In order to have a well-defined elliptic equation, we require that  $\inf_{x \in D} \ell(x) > 0$ . The condition will be fulfilled with the help of an auxiliary transformation. Moreover,  $\ell$  needs to be regular enough. We consider the cases where  $\ell$  has  $L^\infty(D)$ -sample paths. In addition, we consider the problems that arise for rougher sample paths. We will construct the  $\ell(x)$ -model in the following sections.

Let us consider now the discretisation of Equation (5). It is well-known that white noise can be seen as a distributional derivative of the Brownian sheet  $B$ . Moreover, the measurable linear functionals of white noise can be identified with stochastic integrals. Hence, it is natural to discretise white noise as

$$w^N(x) = \sum_{k=1}^{K_n} \left( \frac{1}{|A_k|} \int 1_{A_k}(x) dB_x \right) 1_{A_k}(x), \quad (6)$$

where  $\cup_{k=1}^{K_n} A_k = D$ . The variance of white noise  $w^N$  at  $x \in A_k$ , is

$$w^N|_{x \in A_k} \sim \mathcal{N}(0, |A_k|^{-1}) = \mathcal{N}(0, h^{-d}),$$

where  $h$  is the discretisation step, which we choose to be same along all the coordinate directions.

The only thing we have left to discretise in Equation (5), is the operator part, for which we can use any standard finite difference methods. Hence, we can write e.g. the one-dimensional discretisation of (5) as

$$\begin{aligned} (1 - \ell(x)^2 \Delta) v|_{x=jh} &\approx v_j^N - \ell_j^2 \frac{v_{j-1}^N - 2v_j^N + v_{j+1}^N}{h^2} \\ &= \sigma \sqrt{\ell_j} w_j^N \sim \mathcal{N}(0, \sigma^2 \ell_j h^{-1}), \end{aligned} \quad (7)$$

where  $jh \in h\mathbb{Z}$  is the discretisation lattice, and  $\ell_j := \ell(jh)$ . This model can be given as a matrix equation  $L(\ell^N)v^N = w^N$ , where  $L(\ell^N)$  is a symmetric sparse matrix.

### 3.1 Hypermodels

Let us consider modelling the length-scaling  $\ell(x; u)$  with Gaussian random fields. To achieve an elliptic equation, we apply a log-normal model

$$\ell(x; u) = \exp(u(x)), \quad (8)$$

where  $u$  is a Gaussian random field. For the discrete model on lattice points  $x = jh$ , we set similarly  $\ell_j^N = \exp(u_j^N)$ .

For example, we may choose another Matérn field as a hypermodel for  $u$ . Hence, let  $u$  be a zero-mean Matérn field, with constant length-scaling  $\ell_0$ , and consider its discretisation  $u^N$  by (7). Then  $u^N$  has covariance matrix  $\tilde{C}^N$  and we write the discrete hypermodel as

$$D(u^N) D(v^N | \ell^N(u^N)) \propto \exp\left(-\frac{1}{2} u^T (\tilde{C}^N)^{-1} u^N\right) \times \dots \\ |L(\ell^N)| \exp\left(-\frac{1}{2} (v^N)^T L(\ell^N)^T L(\ell^N) v^N\right).$$

## 4 Discretisation and convergence of the hypermodel

Let us now choose a Matérn field hyperprior for  $u$  as well as Matérn prior for  $v$ , and use Equation (8) for length-scaling  $\ell$ . In this section, we will first consider discretisation and convergence of this hypermodel, and then make some notes, when we modify the hypermodel to have Cauchy walk hyperprior.

As above, let the realisations of the random field  $u$  be tempered distributions that satisfy

$$(1 - \ell_0^2 \Delta) u = \sigma_0 \sqrt{\ell_0^d} \tilde{w}, \quad (9)$$

where  $\tilde{w}$  is  $\mathcal{S}'(\mathbb{R}^d)$ -valued Gaussian white noise on  $\mathbb{R}^d$  and  $\ell_0, \sigma_0 > 0$  are given constants. We assume that  $\tilde{w}$  is statistically independent from  $w$ .

It is easy to verify that  $u$  is a measurable transformation of  $\tilde{w}$  and it has almost surely continuous sample paths for  $d = 1, 2$ . Hence the right hand side of (5) is a well-defined generalised random field with distribution

$$\mu_{\sqrt{\ell(\cdot; u)^d} w}(A) = \int \mu_{\sqrt{\ell(\cdot; f)^d} w}(A) \mu_u(df)$$

on Borel fields  $A$  (with respect to the weak\*-topology) of  $\mathcal{S}'(\mathbb{R}^d)$ .

In the first step, we approximate the random field  $v$  without approximating  $u$  or  $\ell$ . We discretised the Laplacian with finite differences in (5) and, furthermore, discretised the white noise  $w$ .

Let us define a suitable mesh space, where the convergence is studied. We equip the space of all real-valued functions  $v^N$  on the mesh  $h\mathbb{Z}^d \cap \bar{D}$  with norm

$$\|v^N\|_{L^2(D_h)} = \left( \sum_{kh \in \bar{D}} h^2 v^N(kh)^2 \right)^{\frac{1}{2}}.$$

We will use the following notations: we denote by  $B$  a suitable boundary operator that stands e.g. for the periodic or the Dirichlet boundary condition. In the next theorem

$$w^N = T_h w$$

is the Steklov mollified version of the white noise. That is, the radonifying transformation  $T_h$  is defined as

$$T_h f = S_1^2 S_2^2 f,$$

where

$$S_1 f(x_1, x_2) = \frac{1}{h} \int_{x_1-h/2}^{x_1+h/2} f(t, x_2) dt$$

and

$$S_2 f(x_1, x_2) = \frac{1}{h} \int_{x_2-h/2}^{x_2+h/2} f(x_1, t) dt$$

for all  $f \in L^2(D)$ .

**Lemma 4.1.** *Let  $u$  be a Gaussian random field that satisfies (9). Let  $v(x; u)$  satisfy*

$$(\ell(x; u)^{-2} - \Delta) v = \sigma_0 \ell(x; u)^{d/2-2} w \text{ in } D \quad (10)$$

with the periodic boundary condition, where  $\ell(x; u) = g(u(x))$  and

$$g(s) = \exp(s)$$

Let  $v^N(x; u)$  be

$$(\ell(x; u)^{-2} - \Delta_N) v^N(x; u) = \sigma_0 \ell(x; u)^{d/2-2} w^N, \quad (11)$$

on  $h\mathbb{Z}^d \cap D$ , with the boundary condition  $Bv^N = 0$  on  $h\mathbb{Z}^d \cap \partial D$ .

Then  $L^2(L^2(D_h), P)$ -norm of  $v^N - v$  converges to zero as  $h \rightarrow 0$ .

*Proof.* Conditioning with  $u$  inside  $L^2(L^2(D_h), P)$ -norm gives us

$$\mathbb{E} \left[ \|v^N(\cdot; u) - v(\cdot; u)\|_{L^2(D_h)}^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ \|v^N(\cdot; u) - v(\cdot; u)\|_{L^2(D_h)}^2 \mid u \right] \right].$$

Recall that  $u$  is a radonifying transformation of  $\tilde{w}$ , where  $\tilde{w}$  is statistically independent from  $w^N$ . Then also  $u$  and  $w^N$  are statistically independent. Moreover,  $v^N$  is a Carathéodory function of  $(u, w^N)$  (which is radonifying with respect to the second variable). This means that conditioning  $v^N$  with  $u = u_0$ , where  $u_0 \in C(\bar{D})$ , only replaces the random coefficient  $\ell(\cdot; u)$  in (11) with a fixed continuous function  $\ell(\cdot; u_0)$ . The same holds for  $v^N, w^N$  replaced with  $v, w$ .

Let us denote with  $v(\cdot; u, f)$  and  $v^N(\cdot; u, f)$  the solutions of (10) and (11), respectively, when the white noise load  $w$  is replaced with a function  $f \in L^2(D)$ .

By applying adjoints of the solution operators, it is easy to verify that

$$\mathbb{E} \left[ \|v^N - v\|_{L^2(D_h)}^2 \mid u \right] = h^2 \sum_{kh \in D} \sup_{\|f\|_{L^2} \leq 1} (v^N(kh; u, f) - v(kh; u, f))^2$$

due to linearity of the elliptic problem.



By the usual convergence results for the finite-difference scheme (see p. 214 in [12], with straightforward changes for the periodic case), we obtain

$$(v^N(kh; u, f) - v(kh; u, f))^2 \leq Ch^2 \|v(u, f)\|_{W_2^2(D)}^2. \quad (12)$$

By inserting elliptic estimates (see Lemma 4.2 below) into (12), we get the upper bound

$$\mathbb{E} \left[ \|v^N - v\|_{L^2(D_h)}^2 |u\right] \leq C_u \sum_{kh \in D} \sup_{\|f\|_{L^2} \leq 1} h^4 \|v(u, f)\|_{L^2(D)}^2 \leq C_u |D| h^2.$$

where the constant  $C_u \in L^2(P)$ .  $\square$

**Remark 1.** *The above lemma can be easily generalised for the case when  $u$  has almost surely bounded sample paths and  $g$  is bounded from above and below with positive constants. Hence, we obtain similarly the convergence for the Cauchy walk.*

For completeness, we recall the following elliptic estimate.

**Lemma 4.2.**  $\|v(u, f)\|_{W_2^2(D)} \leq C \|f\|_{L^2(D)}$ , where the constant  $C \in L^p(P)$  for all  $p \geq 1$ .

*Proof.* Take  $\sigma_0 = 1$  for simplicity. Let

$$(\ell(x; u)^{-2} - \Delta) v = \ell(x; u)^{d/2-2} f$$

with periodic boundary conditions. Let us write a corresponding integral equation with the help of the operator  $G_{c_0} = (-\Delta + c_0(u))^{-1}$ , where  $c_0(u) = \inf_{x \in D} (\ell(x, u))^2 / 2$ . Then

$$v + G_{c_0}(\ell(\cdot; u) - c_0)v = G_{c_0} \ell(\cdot; u)^{d/2-2} f.$$

With Fourier techniques on the torus, we can show that  $G_{c_0} : L^2(D) \rightarrow H^2(D)$ . Moreover, the norm of the mapping is bounded by the maximum of 1 and  $\sqrt{c_0^{-1}}$ . Therefore,

$$\begin{aligned} \|v\|_{H^2}^2 &\leq \max(1, c_0^{-1}) \sup_{x \in D} (\ell(x; u)^{d-4}) \|f\|_{L^2}^2 + \sup_{x \in D} (\ell(x; u)^2 - c_0)^2 \|v\|_{L^2}^2 \\ &\leq \max(1, c_0^{-1}) \sup_{x \in D} (\ell(x; u)^{d-4}) \|f\|_{L^2}^2 + \frac{\sup(\ell(\cdot; u)^2 - c_0)^2}{\inf(\ell(\cdot; u)^2)} \|f\|_{L^2}^2 \end{aligned}$$

by Babuška-Lax-Milgram theorem. The multipliers of the norm belong to  $L^p(P)$  for all  $p \geq 1$  [7].  $\square$

Next, also Equation (9) is discretised on the equidistant mesh  $h\mathbb{Z}^d$  by finite differences and discretisation of the white noise (6). We further modify the equation

$$(1 - \ell_0^2 \Delta_h) u^N(hk) = \sigma_0 \sqrt{\ell_0^d} w^N(hk), \quad k \in \mathbb{Z}^d,$$

for discrete  $u^N(hk)$  by expressing the discrete white noise  $w^N$  as the measurable transformation  $w^N(hk) = (T_h w)(hk)$  of the continuous-parameter white noise (for details on measurable transformations, see [2]).

**Theorem 4.3.** *Let  $v(x; u)$  satisfy*

$$(1 - \ell(x; u)^2 \Delta) v = \sigma_0 \sqrt{\ell(x; u)^d} w \text{ in } D \quad (13)$$

*with the periodic boundary condition where  $\ell(x; u) = g(u(x))$  and*

$$g(s) = \exp(s)$$

*Let  $v^N(x; u^N)$  satisfy*

$$(1 - \ell(x; u^N)^2 \Delta_N) v^N(x; u^N) = \sigma_0 \sqrt{\ell(x; u^N)^d} w^N,$$

*on  $h\mathbb{Z}^d \cap D$ , with the periodic boundary.*

*Then  $v^N(\cdot; u^N)$  converges to  $v$  in  $L^2(L^2(D_h), P)$  as  $N \rightarrow \infty$ .*

*Proof.* Consider the norm

$$\begin{aligned} \mathbb{E} \left[ \|v^N(\cdot; u^N) - v(\cdot; u)\|_{L^2(D_h)}^2 \right] &\leq 2\mathbb{E} \left[ \|v^N(\cdot; u^N) - v(\cdot; u^N)\|_{L^2(D_h)}^2 \right] \\ &\quad + 2\mathbb{E} \left[ \|v(\cdot; u^N) - v(\cdot; u)\|_{L^2(D_h)}^2 \right], \end{aligned} \quad (14)$$

where  $v(\cdot; u^N)$  solves (13) for some continuous pointwise convergent interpolation of  $u^N$  in place of  $u$ . By Lemma 4.1, the first term of (14) vanishes when  $h \rightarrow 0$ . We show that the second term of (14) vanishes as  $h \rightarrow 0$ . Indeed,

$$\mathbb{E} \left[ \|v(\cdot, u^N) - v(\cdot, u)\|_{L^2(D_h)}^2 | u \right] = h^2 \sum_{kh \in D} \sup_{\|f\|_{L^2} \leq 1} (v(kh; u^N, f) - v(kh; u, f))^2.$$

If we can show that  $v(x; u^N, f) - v(x; u, f)$  converges to zero uniformly with respect to  $x$  and  $f$ , we are done. The term can be considered with the help of Sobolev spaces

$$\begin{aligned} v(x; u^N, f) - v(x; u, f) &= (v(\cdot; u^N, f) - v(\cdot; u, f), \delta_x)_{H^{d/2+\delta, -d/2-\delta}} \\ &\leq C \|v(\cdot; u^N, f) - v(\cdot; u, f)\|_{H^{d/2+\delta}} \end{aligned}$$

We denote the Green's operator for  $\ell(x; u)^{-2} - \Delta$  with  $G_{\ell(\cdot; u)}$  and for  $\ell(x; u^N)^{-2} - \Delta$  with  $G_{\ell(\cdot; u^N)}$ . Then

$$\begin{aligned} \|v(\cdot; u^N, f) - v(\cdot; u, f)\|_{H^{d/2+\delta}} &\leq \left\| (G_{\ell(\cdot; u)} - G_{\ell(\cdot; u^N)}) \sqrt{\ell(\cdot; u)^{d-1}} f \right\|_{H^{d/2+\delta}} \\ &\quad + \left\| G_{\ell(\cdot; u^N)} \left( \sqrt{\ell(\cdot; u)^{d-1}} - \sqrt{\ell(\cdot; u^N)^{d-1}} \right) f \right\|_{H^{d/2+\delta}}. \end{aligned}$$

By using integral equation techniques, we can show that  $\|G_{\ell(\cdot; u^N)}\|_{L^2, H^{d/2+\delta}}$  are uniformly bounded with respect to  $N$ , and the upper bound belongs to  $L^2(P)$ . The uniform boundedness of  $\sup_{x \in D} \ell(x; u^N)$  follows from Sobolev space

estimates and convergence of  $u^N$  to  $u$  in discrete Sobolev norms. Hence, the second term converges. The first term converges similarly since

$$G_{\ell(\cdot;u)} - G_{\ell(\cdot;u^N)} = G_{\ell(\cdot;u)} (\ell(\cdot;u^N) - \ell(\cdot;u)) G_{\ell(\cdot;u^N)}.$$

□

**Remark 2.** *We can show that the Steklov mollification can be replaced with the weaker mollification (6) with similar technique as in the proof of Theorem 4.3.*

**Remark 3.** *Theorem 4.3 holds also for the Cauchy walk, when  $\ell$  is uniformly bounded from above and below.*

## 5 Rough hypermodels

In addition to the models mentioned above, we may wish to use a Cauchy noise, a Cauchy walk, or, white noise, which have rougher sample paths than the previously discussed examples. For the discrete white noise, see Equation (6).

Let us consider a one-dimensional Cauchy walk and its discretisation [17]. The Cauchy walk  $u(x)$  is an  $\alpha$ -stable Lévy motion with  $\alpha = 1$  defined by

$$u(x) = M([0, x]),$$

where  $M(A)$  has Cauchy probability density function

$$f(x) = \frac{|A|}{\pi(|A|^2 + x^2)}$$

for all Borel sets  $A \subset \mathbb{R}_+$ . We denote  $u(x) \sim \text{Cauchy}(|x|, 0)$ . The Cauchy walk is right-continuous and has finite left limits. The discretisation of Cauchy walk is based on independent increments

$$u(hj) - u(h(j-1)) \sim \text{Cauchy}(h, 0).$$

We note that  $\ell(x; u)$  is not stationary. To control the length-scaling in the elliptic equation (5), we make a transformation  $\ell(x; u) = g(u(x))$ , where

$$g(s) = \frac{a}{b + c|s|} + d, \tag{15}$$

and  $d > 0$  is a fixed small constant, and  $a, b, c > 0$  are suitably chosen constants. The corresponding discretised hypermodel can then be constructed as

$$D(u^N)D(v^N|\ell^N(u^N)) \propto \prod_{j=1}^N \frac{h}{h^2 + (u_j^N - u_{j-1}^N)^2} \times \dots \\ |L(\ell^N)| \exp\left(-\frac{1}{2}(v^N)^T L(\ell^N)^T L(\ell^N)v^N\right).$$

We note again that we have included the normalisation constant as  $L(\ell^N)$ -matrix depends on the length-scaling  $\ell^N$ , i.e. it is not a constant.

Similarly, we could use discrete Cauchy noise, which is an iid process with rougher sample paths. For discussion of the Cauchy noise for Bayesian statistical inverse problems, see Sullivan 2016 [26]. Using again Equation (15), discrete Cauchy noise  $u^N$  has typically values around zero, or 'sporadically' some big values. Hence,  $\ell^N$  is typically either long, or 'sporadically' short, i.e. the chosen hypermodel promotes either long or short length-scaling. We note that transformation (15) is constructed to be symmetric with respect to zero.

## 5.1 Realisations

In Figure 1, we have plotted realisations of Cauchy and Gaussian process hyperparameters  $\ell_\omega^N$ , the resulting covariance matrices and realisation of  $v^N$ . The realisations  $v_\omega^N$  clearly have non-stationary features, parts where we have highly oscillatory features and edges, and then smoother parts. In the Bayesian inversion analysis itself, i.e. in the posterior distribution, the  $\ell^N$  is a parameter vector to be estimated, and hence to find where the non-stationarities, like the edges are located.

In Figure 2, we have realisations of the constructed two-dimensional hypermodels. As hyperprior realisations, we have a constant-parameter Matérn field realisation, and, an inhomogeneous Matérn field realisations obtained by using different values of length-scaling and tilt-angle in different parts of the domain. Here we have used an extended domain when constructing the realisation, and in order to remove the boundary effects, we have cropped the images suitably. By varying the Matérn field models, we have three different length-scaling fields. These fields are then used as input for  $g(s)$ , and in the bottom row, we have plotted realisations of the prior. In the bottom panel, second image from the right, we have used  $\ell(u(x)) = \ell_1(x) = 2\ell_2(x)$  and non-zero constant tilt-angle theta. In this way, we can make also anisotropic features.

## 6 MCMC

In order to draw estimates from the posterior distribution, we will use a combination of Gibbs sampling and Metropolis-within-Gibbs algorithms. We summarise the algorithm as follows:

1. Initiate  $v^{N,(0)}$  and  $\ell^{N,(0)}$ .
2. For  $k = 1 \dots K$ 
  - (a) Update  $v^{N,(k)}$  given fixed  $\ell^{N,(k-1)}$  and draw  $\eta \sim \mathcal{N}(0, I)$ , and set

$$v^{N,(k)} = \begin{pmatrix} \sigma^{-1}A \\ L(\ell^{N,(k-1)}) \end{pmatrix}^\dagger \left( \begin{pmatrix} \sigma^{-1}y \\ 0 \end{pmatrix} + \eta \right),$$

where  $\dagger$  denotes the matrix pseudoinverse.

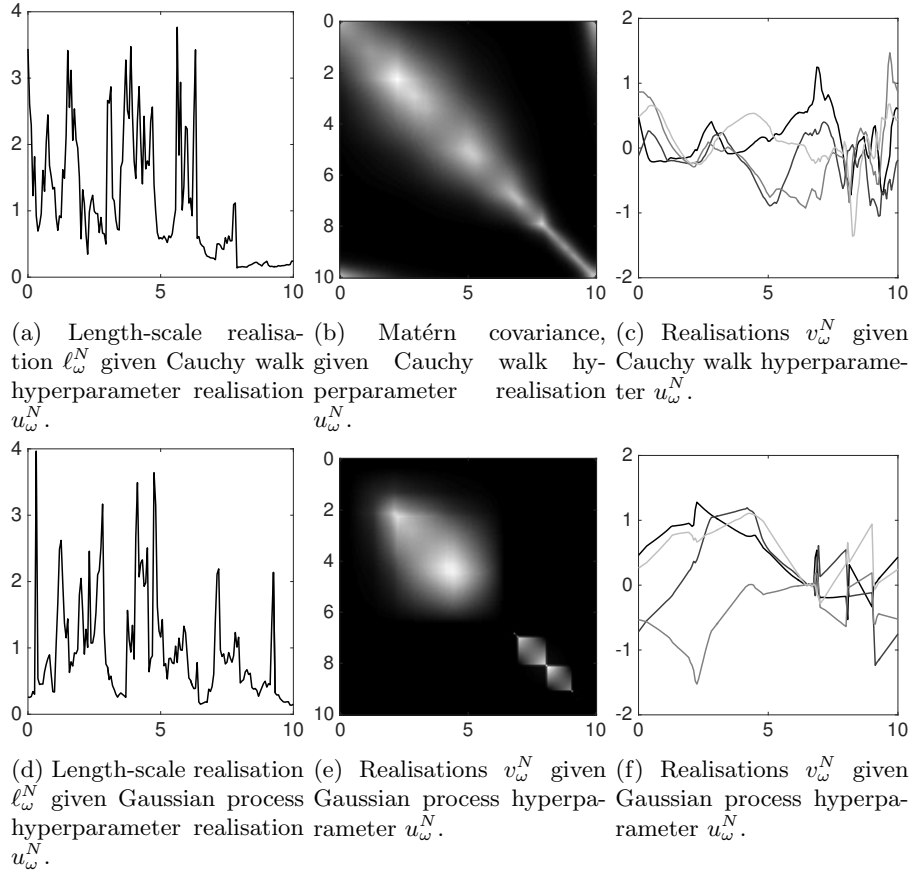


Figure 1: Examples of constructing non-stationary Matérn realisations with hypermodels. Top panel – from left to right: Realisation  $\ell_\omega^N$  given Cauchy walk as  $u_\omega^N$ , resulting covariance matrix, and four realisations. Bottom panel: Same as above, but with a Gaussian process hyperprior.

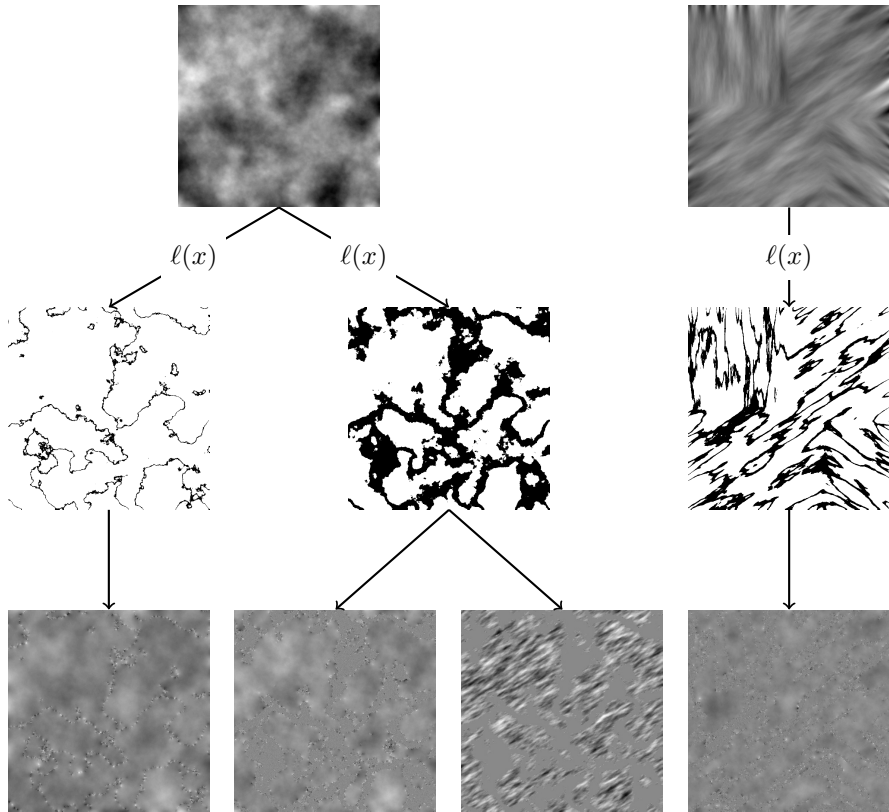


Figure 2: Non-stationary structures obtained by starting from a constant-parameter or inhomogeneous Matérn field realisation (upper panel), after which have been mapped to correlation length-scaling fields (middle). In the bottom panel, we have corresponding realisations with isotropic and anisotropic structures. This kind of structure can detect regions within which the behaviour of the random field is smooth, but the regions are distinct.

- (b) Update  $\ell^{N,(k)}$  using Metropolis-within-Gibbs given fixed  $v^{N,(k)}$ .  
 For  $n = 1 \dots N$
- i. Draw a candidate sample  $\tilde{\ell}_n^N$  from a proposal distribution  $Q(\cdot | \ell_n^{N,(k-1)})$
  - ii. Denote  $\ell_{j \neq n}^{N,(k)} := (\ell_1^{N,(k)}, \dots, \ell_{n-1}^{N,(k)}, \ell_{n+1}^{N,(k-1)}, \dots, \ell_N^{N,(k-1)})^T$ , and accept with probability
 
$$p_n = \min \left( 1, \frac{D(\tilde{\ell}_n^N | v^{N,(k)}, \ell_{j \neq n}^{N,(k)}) Q(\ell_n^{N,(k-1)} | \tilde{\ell}_n^N)}{D(\ell_n^{N,(k-1)} | v^{N,(k)}, \ell_{j \neq n}^{N,(k)}) Q(\tilde{\ell}_n^N | \ell_n^{N,(k-1)})} \right)$$
  - iii. If accepted, we set  $\ell_n^{N,(k)} = \tilde{\ell}_n^N$ . Otherwise we set  $\ell_n^{N,(k)} = \ell_n^{N,(k-1)}$ .
  - iv. Set  $n \leftarrow n + 1$ , and repeat from step (i) until  $n = N$
- (c) Set  $k \leftarrow k + 1$ , and repeat from step (a) until the desired sample size  $K$  is reached.

Metropolis-within-Gibbs is explained for example in [8, 17]. The latter uses the term single-component Metropolis-Hastings to emphasise the fact that we sample every single component separately with the Metropolis-Hasting algorithm. We aim at acceptance ratio between 25-50 per cent, which is obtained by tuning the random walk proposal process.

In computing acceptance probability, it is a common practice, due to numerical reasons, to take logarithms instead of using ratios. For example, in the case of the Gaussian hyperprior, the logarithm of the posterior is

$$\begin{aligned} \log(D(v^N, \ell^N | y)) = & R - \frac{1}{2}(u^N)^T (\tilde{C}^N)^{-1} u^N + \log(|L(\ell^N)|) - \dots \\ & \frac{1}{2}(v^N)^T L(\ell^N)^T L(\ell^N) v^N - \frac{1}{2}(y - Av^N)^T \Sigma^{-1} (y - Av^N), \end{aligned} \quad (16)$$

where  $R$  is some constant, which we may omit from the analysis. We note that the normalisation constant computation, i.e. logarithmic determinant  $\log(|L(\ell^N)|)$ , is a numerically unstable and computationally expensive operation, especially in higher dimensions. We need to compute altogether  $N \times K$  logarithmic determinants in our estimation algorithm, hence we may wish to minimise the log-determinant computation time.

We note that in the Metropolis-Hastings part, when updating  $\ell_n^{N,k}$ , we are actually computing ratio of the proposed and old normalisation constant. Let us denote the proposed and old covariances by  $C_{\text{old}}$  and  $C_{\text{prop}}$ , respectively. Then we should actually calculate the ratio, as originally in Equation (16), and not take logarithms. Then by simple algebra we have

$$\frac{\sqrt{|C_{\text{old}}|}}{\sqrt{|C_{\text{prop}}|}} = \frac{\sqrt{|(L_{\text{old}}^T L_{\text{old}})^{-1}|}}{\sqrt{|(L_{\text{prop}}^T L_{\text{prop}})^{-1}|}} = \frac{|L_{\text{prop}}|}{|L_{\text{old}}|} = |L_{\text{prop}} L_{\text{old}}^{-1}|.$$

Now, we note that as we update only one row at the time, the matrix-inverse-matrix-product is of the form:

$$L_{\text{prop}}L_{\text{old}}^{-1} = \begin{pmatrix} I & 0 \\ \times & \times \\ 0 & I \end{pmatrix}.$$

Hence, the product is diagonal, except for the  $n^{\text{(th)}}$  updated row. This means that we simply need to compute diagonal alue of the updated row, i.e. only one value. It would seem that we would need to invert whole matrix  $L_{\text{old}}^{-1}$ . However, we note that as  $L$ -matrices are sparse, so we will have only a limited amount of non-zero values. Consider e.g. the one-dimensional case. We could have e.g.

$$(0 \quad \dots \quad 0 \quad a \quad b \quad a \quad 0 \quad \dots \quad 0)L_{\text{old}}^{-1} = (0 \quad \dots \quad 0 \quad \times \quad \times \quad \times \quad 0 \quad \dots \quad 0),$$

where  $a$  and  $b$  are constants derived from the approximations in Equation (7). By simple matrix operations and removing the 'zeroes' from the matrix equation, we can rewrite this as

$$(a \quad b \quad a) = (\times \quad \times \quad \times)\tilde{L}_{\text{old}},$$

where  $\tilde{L}_{\text{old}}$  is a  $3 \times 3$  matrix. Hence, the computation of the determinant is then simply inverting a  $3 \times 3$  matrix and making one matrix-vector multiplication. In two-dimensional problems, we need to invert  $5 \times 5$  matrix.

## 7 Numerical examples

Now, we shall apply the developed methodology to one-dimensional interpolation and numerical differentiation, and, to two-dimensional interpolation.

### 7.1 One-dimensional interpolation

We model discrete noise-perturbed observations of a continuous object  $v$  as

$$y(jh) = v(j'h') + e(jh),$$

where  $e(jh)$  is zero-mean white noise with known variance, and  $j \in \mathbb{J} \subset \mathbb{Z}$  is the measurement mesh and  $j' \in \mathbb{J}' \subset \mathbb{Z}$  is the mesh of the discretised unknown  $v^N$ . The discretisation steps  $h, h' > 0$ . This model, can be rewritten in the form given in Equation (3), i.e. as  $y = Av^N + e$ . Hence we can write the whole posterior distribution with the help of hypermodels as discussed earlier.

Let us consider  $v$  consisting of a  $C^\infty$  mollifier function and two boxcar functions

$$v(x) = \begin{cases} \exp\left(4 - \frac{25}{x(5-x)}\right), & x \in (0, 5) \\ 1, & x \in [7, 8] \\ -1, & (8, 9] \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$



This function has smooth parts, edges, and it is also piecewise constant for  $x \in [5, 10]$ . In Figure 3, we have simulation results with three different prior and hypermodels:

1. Constant-parameter Matérn prior, i.e. this model does not have hyperprior.
2. Hypermodel with Cauchy walk  $u$ .
3. Hypermodel with stationary Gaussian zero-mean process  $u$  with exponential covariance

The domain for both the measurements  $y$  and unknown  $v^N$  is  $[0, 10]$ . The measurement mesh is  $j = \{0, 1, \dots, 80\}$ ,  $h = 1/8$  and the unknown mesh  $j' = \{0, 1, \dots, 160\}$ ,  $h' = 1/16$ . Zero-mean measurement noise has standard deviation  $\sigma = 0.1$ . Matérn prior has periodic boundary conditions.

With the constant-parameter Matérn prior, we have plotted estimates with a long length-scaling (D), length-scaling minimising maximum absolute error (G), and length-scaling minimising root mean square error (J). These estimates capture the smoothness or edges, but not both at the same time. With the Cauchy and Gaussian hypermodels, the algorithm finds short and long length-scaling  $\ell^N$  (subfigures (B) and (C)). Also, the corresponding  $v^N$  estimates in (E) and (F) show that we can reconstruct both smooth and edge-preserving parts. In subfigures (H) and (I), we have plotted  $v^N$  on the measurement mesh, and in subfigures (K) and (L) in the interpolated points, i.e. between the measurement grid points. This shows that the interpolated estimates are behaving as expected.

In order to relate this study to Paciorek’s 2003 study [19], we note that the (D,G,J) subfigures correspond to Paciorek’s so-called single-layer model. For deep learning, one should build a series of hyperpriors over hyperpriors. Here, instead, we have a two-layer model, and we may note that it captures different properties with very good precision. Hence, the question remains whether two layers is actually often enough in deep Gaussian processes, and what is the actual gain using deep layers. We will leave this question open, and hope to address that in subsequent studies.

In Figures 4 and 5, we study the behaviour of the  $\ell^N$  and  $v^N$ , when  $N$  changes, i.e. we numerically study discretisation-invariance. We choose  $N = 81, 161, 321$ . The Cauchy and Gaussian hypermodels, as well as the forward theory, are the same as in example in Figure 3. As we have constructed the hypermodels based on continuous-parameter processes, we assume that the finite-dimensional estimates essentially look the same. This was covered theoretically in Section 4. This behaviour can be verified from all the  $v^N$  estimates visually rather easily, but the  $\ell^N$  are not as well behaving. The reason is mostly due to too short chains, but already with the chains here, with  $K = 100,000$ , the essential features are in practice rather similar.

We have also plotted the MCMC chains and cumulative means for 15<sup>(th)</sup> and 66<sup>(th)</sup> elements of  $\ell^N$  and  $v^N$  with  $N = 81$ . The elements are chosen in

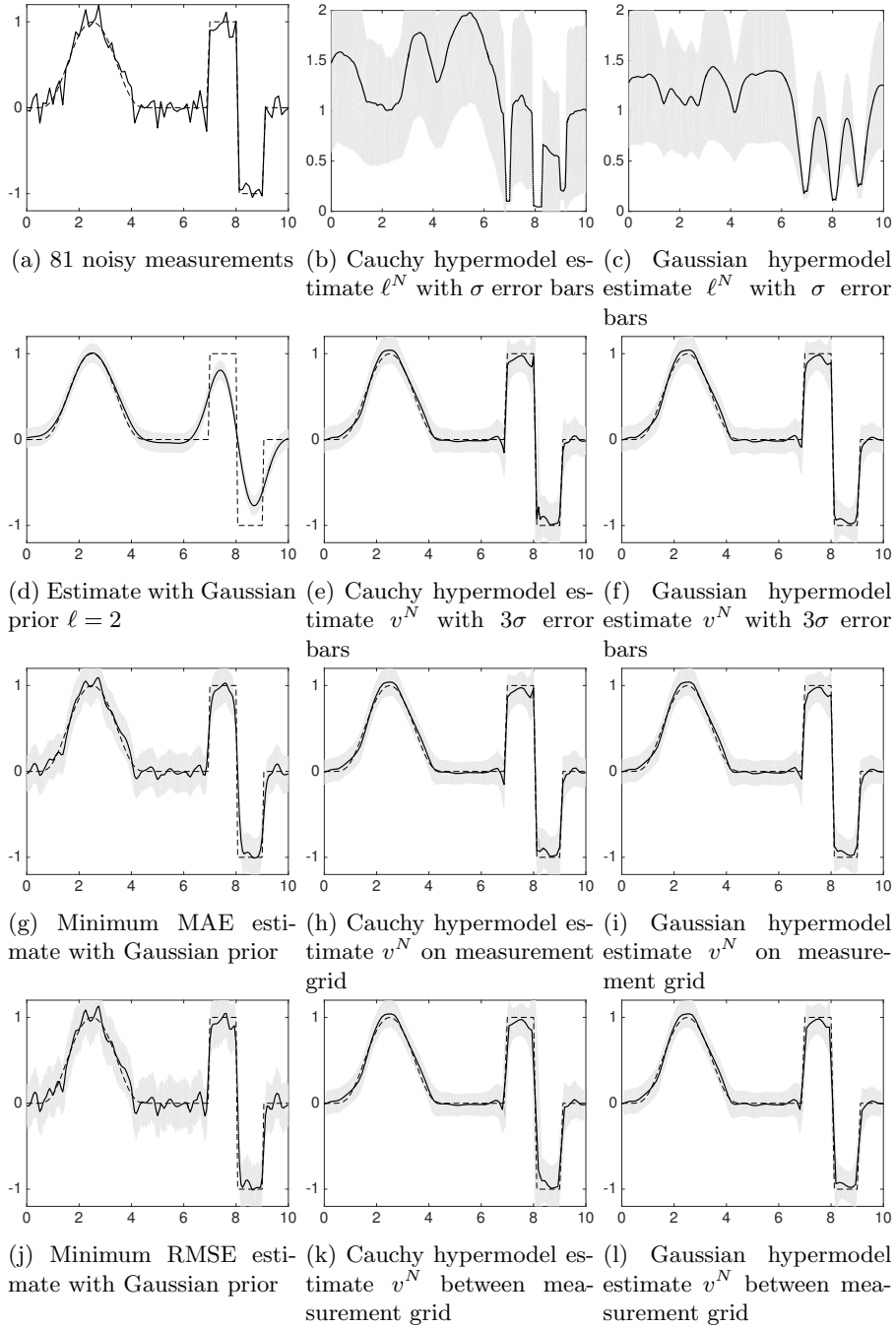


Figure 3: Top panel: 81 noisy measurements and estimated  $\ell^N$  ( $N = 161$ ) with Cauchy noise (B) and Gaussian hyperprior (C). (D,G,J) are conditional mean estimates of  $v^N$  ( $N = 161$ ) with long length-scaling (D),  $\ell^N$  minimising MAE (G),  $\ell^N$  minimising RMSE. (E,H,K) and (F,I,L) are CM-estimates of  $v^N$  on different meshes with Cauchy hypermodel and Gaussian hypermodels, respectively.

such a way the the 15<sup>(th)</sup> element is on a smoothly varying part of the unknown, hence long length-scaling. Around 66<sup>(th)</sup> element, we expect to detect an edge, hence short length-scaling. The chains show both good mixing and convergence to the values expected. We use 50,000 as burning period. As can be seen from the figures, we could use much shorter burning periods and chains, but here our objective is simply to demonstrate the estimators, not optimisation of the chains. Hence, we leave optimisation of MCMC chains for future studies.

## 7.2 Multimodal posterior densities

Let us now consider posterior densities of  $v_j^N$ . If we use the Gaussian hypermodel, and use exponentiating  $u^N$  (Equation (8)), it is easy to show numerically that the posterior densities of  $v^N$  are Gaussian. It would be tempting to model longest and shortest length-scaling, i.e. a priori lower and upper bounds for  $\ell^N$ . One such model, could be given as

$$g(s) = \begin{cases} \gamma, & s < P_{\text{upper}} \\ \exp(a|s|) - b, & \text{otherwise} \\ \lambda, & s > P_{\text{lower}} \end{cases} \quad (18)$$

where  $a > 0$ ,  $b \in (0, 1)$  are some constants, and  $P_{\text{upper}} > P_{\text{lower}} > 0$ . It is is convenient to model the lower bound as  $\ell(0) = 1 - b = P_{\text{lower}}$ . However, using this model leads to multimodal posterior densities due to the max-min cutoff, especially at the jumps.

In Figure 6, we have considered the same Gaussian hypermodel and interpolation problem as in Figure 3. Given the MCMC chains, we compute kernel density estimates of the posterior densities at the jump at  $x \approx 8$ . This corresponds to grid elements at  $j = 129, 130, 131$ , where we have a jump from  $+1$  to  $-1$ . The densities at grid elements 129 and 131 are Gaussian. However, the at grid element 130, the density is trimodal. We have plotted also the Gaussian density estimate with dashed line, and clearly it fails to capture the trimodality. The reason for the trimodal density is, assumably, in the non-linear transformation of Equation (18). Hence, the algorithm does detect edges, but we need to be careful when assessing the multimodality of the posterior densities.

## 7.3 Numerical differentiation

As a second numerical example, we consider numerical differentiation of noisy data. The continuous forward model is given with a first kind Fredholm integral equation

$$y(jh) = \int H(jh - x)v(x)dx + e_j,$$

where the convolving kernel  $H$  is a Heaviside step function. If we had a deterministic equation (i.e. no noise term  $e_j$ ), then  $y' = v$  is the differentiated function. Hence, we can formulate differentiation as a linear Bayesian inverse problem with observations given as  $y = Av^N + e$ .

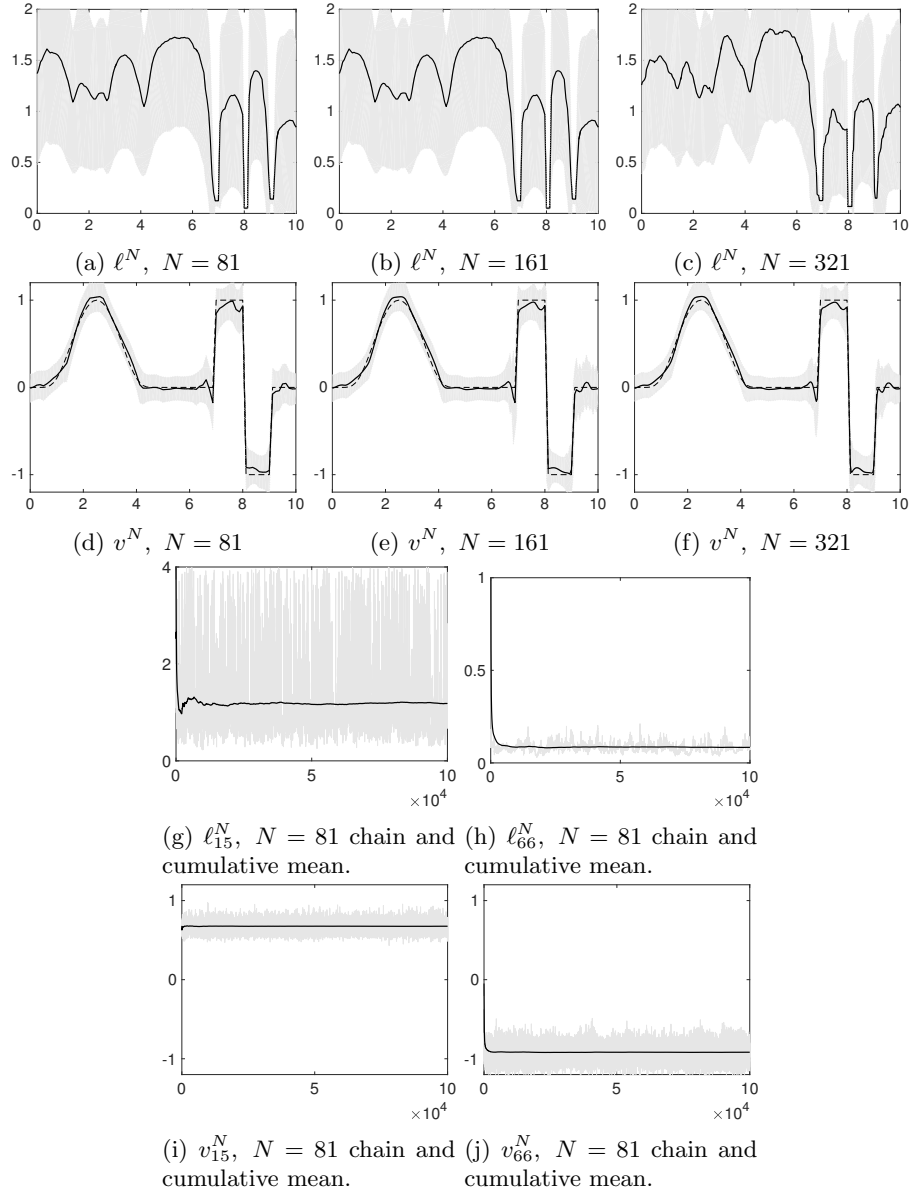


Figure 4: Estimates of  $\ell^N$  and  $v^N$  with a Cauchy walk hypermodel  $u^N$  on different lattices with 81 measurements, with the number of unknowns varying as in figures. Bottom four subfigures (G-J) are chains and cumulative means of certain  $\ell^N$  and  $v^N$  elements.

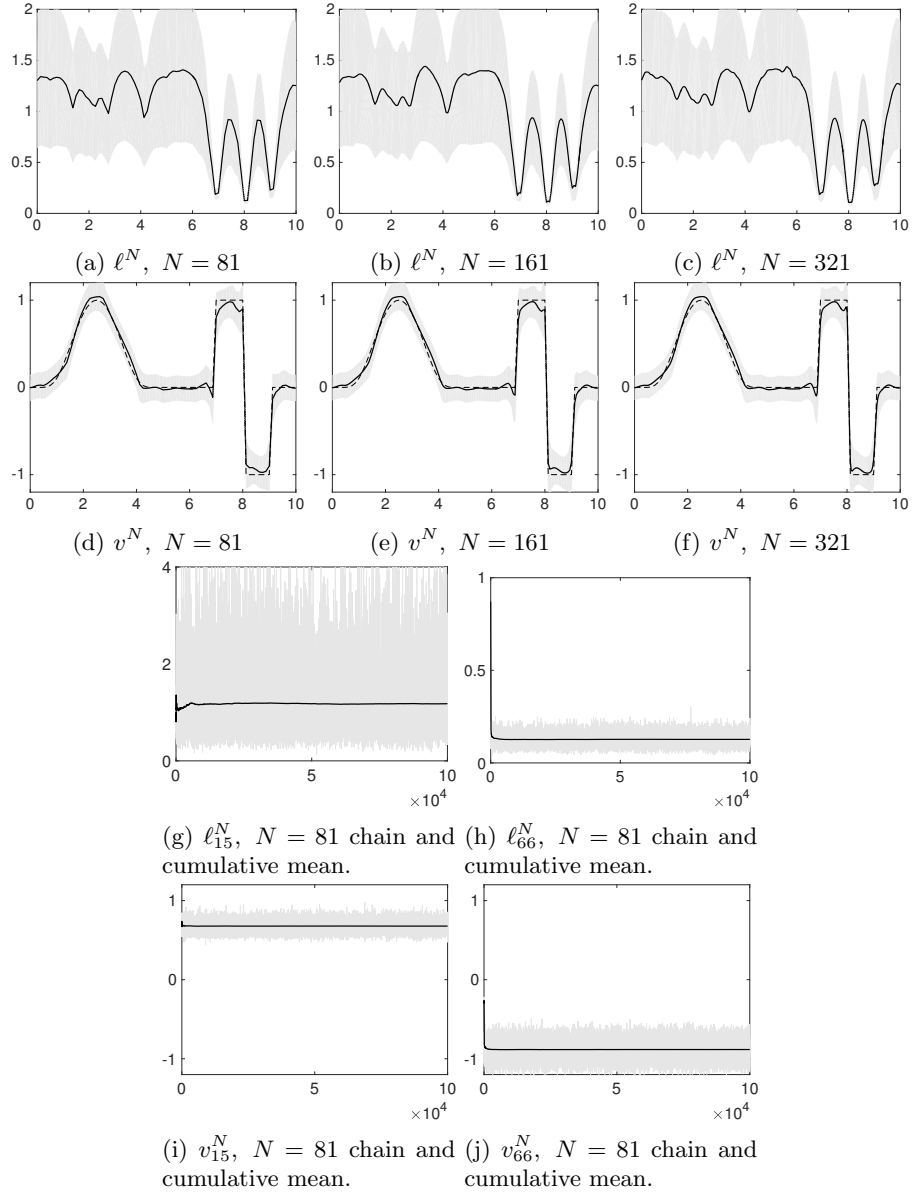


Figure 5: Estimates of  $\ell^N$  and  $v^N$  with a Gaussian hyperprior  $u^N$  on different lattices with 81 measurements, with the number of unknowns varying as in figures. Bottom four subfigures (G-J) are chains and cumulative means of certain  $\ell^N$  and  $v^N$  elements.

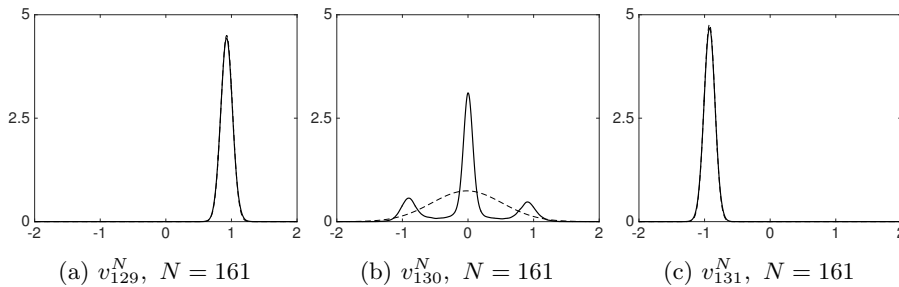


Figure 6: Multimodality of the posterior densities  $v_j^N$  at the edge around  $x \approx 8$ , when using Gaussian hyperprior and Equation (18).

For numerical tests, we choose an unknown consisting of a mollifier and a triangle function

$$\int H(x' - x)v(x)dx = \begin{cases} \exp\left(4 - \frac{25}{x'(5-x')}\right), & x' \in (0, 5) \\ x' - 7, & x' \in [7, 8] \\ -x' + 9, & (8, 9] \\ 0, & \text{otherwise.} \end{cases}$$

Derivative of a triangle function is a piecewise constant function, the same two boxcar functions as in Equation (17). Differentiated mollifier is again a smoothly varying function. We have chosen the mollifier constants in such a way that the differentiated mollifier stays in the same range as the mollifier, simply to avoid any visualisation problems in scaling. The unknown function is then

$$v(x) = \begin{cases} \left(\frac{25}{x^2(5-x)} - \frac{25}{x(5-x)^2}\right) \exp\left(4 - \frac{25}{x(5-x)}\right), & x \in (0, 5) \\ 1, & x \in [7, 8] \\ -1, & (8, 9] \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 7, we have numerical derivatives  $v^N$  on three different meshes, as well as the Gaussian hyperprior process  $\ell^N$ . In the simulations, we have 101 observations  $y$  with measurement noise  $\sigma = 0.03$ . We note that as numerical differentiation is an ill-posed problem, so we cannot use as high noise-levels as in the interpolation examples. However, with the used noise levels, the algorithm finds the edges, as well as the smooth structures.

## 7.4 Two-dimensional interpolation

Similarly to the one-dimensional interpolation examples with Cauchy and Gaussian hypermodels, we can make two-dimensional interpolation. In Figure 8, we have interpolation of noisy observations, originally on a  $41 \times 41$  mesh, and we

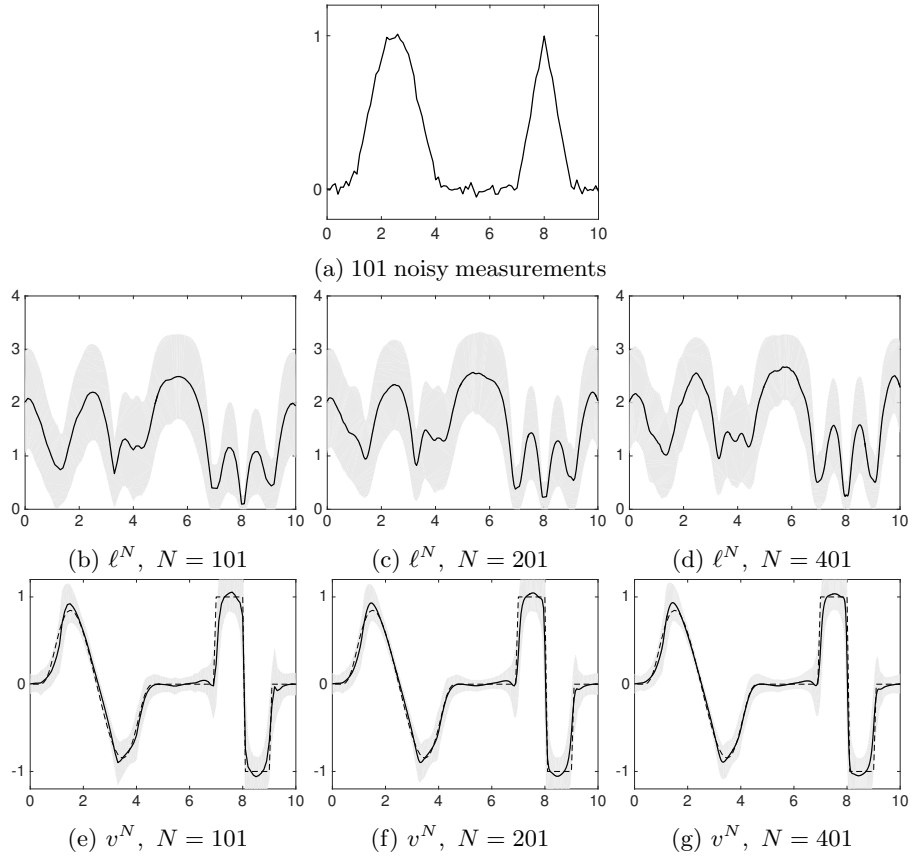


Figure 7: Numerical differentiation of a noisy signal with the developed Gaussian hypermodel. We plot  $v^N$  on different meshes for seeing the discretisation-invariance of the estimates.

estimate the unknown on a  $81 \times 81$  mesh. Measurement noise standard deviation is  $\sigma = 0.025$ . As a hyperprior, we have used a two-dimensional Matérn field with short length-scaling and periodic boundary conditions. The unknown consists of a rectangle-shaped box of height 0.75 and a Gaussian-shaped smooth function of height 1. We can clearly detect both smooth and edgy properties in this case also.

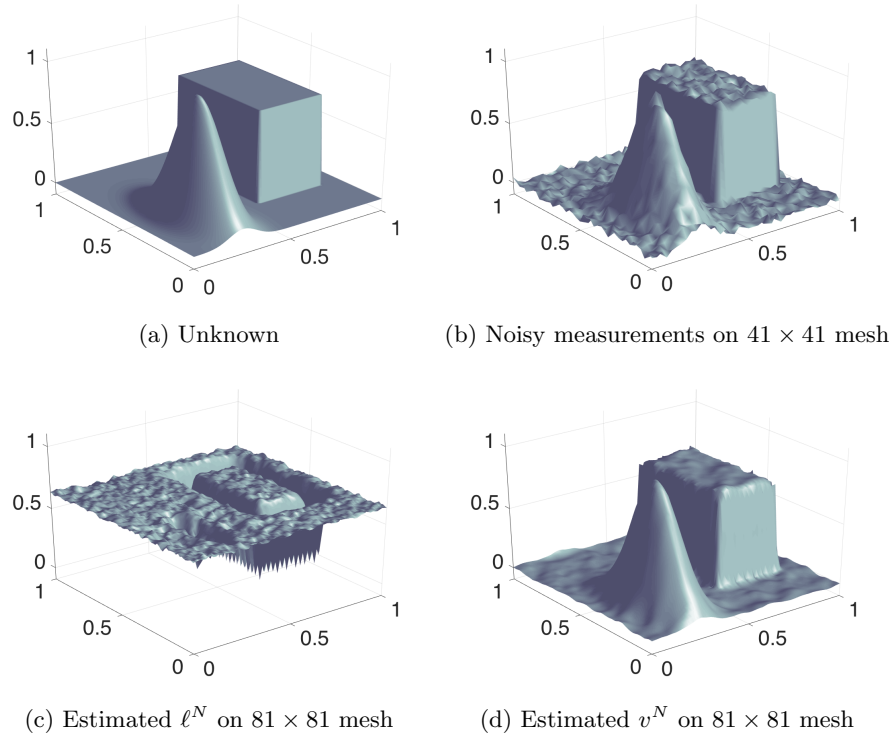


Figure 8: Two-dimensional interpolation of block-shaped and Gaussian-shaped structures from noisy observations. A Matérn hyperprior is used in the analysis.

## 8 Conclusion and discussion

We have considered the construction of hypermodels which promote both smoothness and rapid oscillatory features. The methodology is based on constructing Cauchy and Gaussian hypermodels for Matérn field length-scaling  $\ell^N$ . We constructed a combined Gibbs and Metropolis-within-Gibbs algorithm for computing estimates of the unknown and length-scaling, respectively. In addition, we have shown both analytically and numerically discretisation-invariance of the estimates. The estimates provide significant advances in comparison to



standard constant-parameter Matérn field priors, as we can detect more versatile features. In this study, we did not include all the Matérn field parameters in the hyperprior. In the future studies, e.g. in the two-dimensional problems, we should have hypermodel fields for  $\ell_1, \ell_2, \theta$  and in addition to the variance scaling mask  $\sigma^2$ .

We consider this paper to be a concept paper and hence we have considered simple inversion examples. However, the methodology can be applied to e.g. electrical impedance tomography, Darcy flow models and X-ray tomography. In addition, implementing spatiotemporal models with infinite-dimensional Kalman filter techniques would be an interesting path forwards. In the more theoretical side, we should study the discretisation-invariance issues more rigorously. Also, the computational machinery needs to be developed further, for example by using MCMC algorithms, i.e. the Metropolis-within-Gibbs can be run with multicore computers. Utilisation of GPUs would also be of interest.

## Acknowledgments

The authors thank Professor Andrew Stuart for useful discussions and proposals. This work has been funded by Engineering and Physical Sciences Research Council, United Kingdom (EPSRC Reference: EP/K034154/1 – Enabling Quantification of Uncertainty for Large-Scale Inverse Problems), and, Academy of Finland (application number 250215, Finnish Program for Centers of Excellence in Research 2012-2017).

## References

- [1] (MR3063540) J. M. Bardsley, Gaussian Markov random field priors for inverse problems, *Inverse Problems and Imaging*, **7** (2013), 397–416.
- [2] (MR2267655) V. I. Bogachev, *Measure Theory*, Springer-Verlag, Berlin, 2007.
- [3] (MR3249417) D. Bolin, Spatial matérn fields driven by non-gaussian noise, *Scandinavian Journal of Statistics*, **41** (2014), 557–579.
- [4] (MR2351679) D. Calvetti and E. Somersalo, *Introduction to Bayesian Scientific Computing – Ten Lectures on Subjective Computing*, Springer, New York, 2007.
- [5] (MR2421950) D. Calvetti and E. Somersalo, A Gaussian hypermodel to recover blocky objects, *Inverse Problems*, **23** (2007), 733–754.
- [6] (MR2421950) D. Calvetti and E. Somersalo, Hypermodels in the Bayesian imaging framework, *Inverse Problems*, **24** (2008), 034013.

- [7] (MR2888311) J. Charrier, Strong and weak error estimates for elliptic partial differential equations with random coefficients, *SIAM Journal on Numerical Analysis*, **50** (2012), 216–246.
- [8] M. Dunlop, *Analysis and Computation for Bayesian Inverse Problems*. PhD thesis, University of Warwick, 2016.
- [9] (MR3431054) G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue, Does non-stationary spatial data always require non-stationary random fields?, *Spatial Statistics*, **14** (2015), 505–531.
- [10] M. Filippone and M. Girolami, Pseudo-Marginal Bayesian Inference for Gaussian Processes, *IEEE Transactions Pattern Analysis and Machine Intelligence*, **36** (2014), 2214–2226.
- [11] (MR2746411) T. Helin and M. Lassas, Hierarchical models in statistical inverse problems and the Mumford-Shah functional, *Inverse Problems*, **27** (2011), 015008.
- [12] (MR3136501) B. S. Jovanović and E. Süli, *Analysis of Finite Difference Schemes*, Springer, London, 2014.
- [13] (MR2102218) J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer-Verlag, New York, 2005.
- [14] (MR2558305) M. Lassas, E. Saksman and S. Siltanen, Discretization invariant Bayesian inversion and Besov space priors, *Inverse Problems and Imaging* **3** (2009), 87–122.
- [15] (MR2109134) M. Lassas and S. Siltanen, Can one use total variation prior for edge-preserving Bayesian inversion?, *Inverse Problems* **20** (2004), 1537–1563.
- [16] (MR2853727) F. Lindgren, H. Rue, and J. Lindström, An explicit link between Gaussian Markov random fields: the stochastic partial differential equation approach, *Journal of the Royal Statistical Society: Series B*, **73** (2011), 423–498.
- [17] M. Markkanen, L. Roininen, J. M. J. Huttunen, and S. Lasanen, Cauchy difference priors for edge-preserving Bayesian inversion with an application to x-ray tomography, *ArXiv*, (2016).
- [18] J. Norberg, L. Roininen, J. Vierinen, O. Amm, D. McKay-Bukowski, and M. S Lehtinen, Ionospheric tomography in Bayesian framework with Gaussian Markov random field priors, *Radio Science*, **50** (2015), 138–152.
- [19] C. J. Paciorek, *Nonstationary Gaussian Processes for Regression and Spatial Modelling*, PhD thesis, Carnegie Mellon University, 2003.

- [20] (MR2240939) C. J. Paciorek and M. J. Schervish, Spatial modelling using a new class of nonstationary covariance functions, *Environmetrics*, **17** (2006), 483–506.
- [21] (MR3209311) L. Roininen, J. Huttunen and S. Lasanen, Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography, *Inverse Problems and Imaging*, **8** (2014), 561–586.
- [22] (MR3063550) L. Roininen, P. Piiroinen and M. Lehtinen, Constructing continuous stationary covariances as limits of the second-order stochastic difference equations, *Inverse Problems and Imaging* **7** (2013), 611–647.
- [23] Yu. A. Rozanov, Markov random fields and stochastic partial differential equations, *Mat. Sb. (N.S.)*, **103** (1977), 590–613.
- [24] (MR2130347) H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC, Boca Raton, FL, 2005.
- [25] (MR2652785) A. M. Stuart, Inverse problems: a Bayesian perspective, *Acta Numerica*, **19** (2010), 451–559.
- [26] T. J. Sullivan, Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors, *ArXiv*, (2016).