Stochastic Models Involving Second Order Lévy Motions

Wallin, Jonas

2014

Stochastic models involving second order Lévy motions

estimation and prediction problems

Jonas Wallin

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

Contents

Acknowledgements

First and foremost, I would like to thank my supervisor Krysztof Podgórski, for his help, patience, knowledge and friendship. I am forever grateful.

Secondly, I would like to thank co-author, former PhD-colleague and friend David Bolin, with whom I have had much fun working and not working with.

Also, my master thesis supervisor and co-author Finn Lindgren for introducing me to computational and spatial statistics.

At the department, I would like to thank Johan Lindström for always taking time to assisting me. Also, Magnus Wiktorsson for countless interesting discussions on topics ranging form Lévy processes to Turkey. For assistance with various practical problems, I would like to thank James, Maria, and Mona. Finally, I would like to thank my family. My parents for all help and support throughout my life, Martina for living with, supporting and loving me, and Henning for everything.

Lund, 2014 *Jonas Wallin*

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

http://www.maths.lth.se/

Doctoral Theses in Mathematical Sciences 2014:1
ISSN 1404-0034

ISBN 978-91-7473-842-1
LUTFMS-1041-2014

Printed in Sweden by KFS AB, Lund 2014

List of papers

This thesis is based on the following papers:

A Krzysztof Podgórski and Jonas Wallin: Maximizing leave-one-out likelihood for the location parameter of unbounded densities
Annals of the Institute of Statistical Mathematics (in press)

B Krzysztof Podgórski and Jonas Wallin: Convolution invariant subclasses of generalized hyperbolic distributions
to appear in: *Communications in Statistics - Theory and Methods*

C David Bolin and Jonas Wallin : Non-Gaussian Matérn fields with an application to precipitation modeling.
submitted.

D David Bolin, Finn Lindgren and Jonas Wallin: A Gaussian mixture model for multivariate spatially dependent data using discrete and continuous Markov random fields.
submitted.

E Krzysztof Podgórski, Igor Rychlik and Jonas Wallin : Slepian model for moving averages driven by a non-Gaussian noise.

Introduction

> "If you, like us, have applications in mind, it will take no effort whatsoever to convince you that not all random fields occurring in the 'real world' are Gaussian."
>
> *Robert Adler, Jonathan Taylor*

In continuous spatial modeling, the Gaussian fields are so dominant that all processes not being Gaussian are generally classified into the broad generic category of non-Gaussian fields.

There are several reasons for the underrepresentation of non-Gaussian random fields in spatial modeling. It is not trivial to define a model that corresponds to an actual random field. For a model not producing a random field there is no valid distribution at an arbitrary location, and thus interpretation oft the model and prediction with the model may not be possible. Even if the model defines a valid random field, the posterior distribution of the field when observed at some location is in general unknown, or at least practically impossible to deal with. Another difficulty is that in spatial statistics everything is based, for historical reasons, on the covariance function while, generally a processes is not specified uniquely by its covariance function.

An alternative to defining fields through the covariance function, is formulating a stochastic partial differential equation (SPDE). If a linear PDE is driven by Gaussian white noise, then a solution is a Gaussian field with some covariance function depending on the differential equation; if the field is driven by a Lévy noise, the resulting field will be non-Gaussian yet have the same covariance function as its Gaussian counterpart. However, there might be several differential equations generating different processes with the same covariance function. Thus, formulating the problem as an SPDE provides, in general, more information than the covariance function. The SPDE formulation also facilitates refined numerical methods for PDEs that have been developed in numerical analysis.

In this thesis we introduce two new "non-Gaussian" random field models. In Paper C, we build upon Bolin (2013) to create two different types of fields driven

by Lévy noise, allowing for flexible marginal distribution and having the popular Matérn covariance class.

In Paper E we combine two classical random field models: the Potts-field (a discrete Markov random fields) and the classical latent Gaussian random field model to generate a dependent mixture of, possible multivariate, random fields.

What follows in this section is a brief overview of some of the theory and models used in this thesis and at the end there is a short introduction to each paper.

0.1 Stochastic Process

This section gives a very brief introduction to the theory of stochastic processes, which is the backbone of this thesis. Most of the material in this section is based on the three books Adler & Taylor (2007), Azaïs & Wschebor (2009) and Lindgren (2012) .

A probability space $(\Omega, \mathcal{F}, \mathsf{P})$ consist of: a sample space Ω, a family of events $\mathcal{F}$ in Ω , and a probability measure P on $\mathcal{F}$. Let $(\mathcal{B}, B)$ be a measurable space, then any measurable function

$$Y : \Omega \to B$$

is a random variable with probability distribution given by

$$\mathsf{P}(Y(\omega) \in A)$$

for $A \in \mathcal{B}$. Let $(\mathcal{T}, T)$ be a measurable space, then any function

$$X(\omega, \mathbf{t}) : \Omega \times T \to B$$

which is measurable for fixed $\mathbf{t}$ is a stochastic processes, with parameter space T. Thus, by definition, for each fixed $\mathbf{t}_0 \in T$, $X(\omega, \mathbf{t}_0)$ is a random variable from Ω to B; for each fixed $\omega_0 \in \Omega$, $X(\omega, \mathbf{t})$ is a function from T to B, often denoted *realization*, *sample path*, or *trajectory*.

From above definition it is not easy to define a stochastic process, the Kolmogorov extension theorem gives some assistance with this issue. Basically, the theorem says that every consistent family of finite-dimensional distributions defines a unique probability measure on $(B^T, \mathcal{B}^T)$ and thus a stochastic process.

0.2 Stationarity

One often needs to limit the study of stochastic process to certain subclass in order to handle them both practically and theoretically. One of the most important subclasses are the stationary processes.

Definition 1. A stochastic process $X(\omega, \mathbf{t})$ is strictly stationary if for any choice of the positive integer n and $\{\mathbf{t}_1, \ldots, \mathbf{t}_n\} \in T^n$ the joint distribution of $\{X(\omega, \mathbf{t}_1 + \tau), \ldots, X(\omega, \mathbf{t}_n + \tau)\}$ does not depended on τ.

There also exists a weaker (if the process has finite second moment) property namely that of **weak stationarity** which only requires that the first two moments are invariant to shift transformations. It should be noted that when dealing with actual data, stationarity is often quite unrealistic.

0.3 Gaussian processes

Another common subclass are the Gaussian processes. A process is Gaussian if for all finite sets of locations, the corresponding marginal distribution is multivariate normal. Thus, the Gaussian process is completely specified by its two first moments: its mean function $\mu(\mathbf{t})$ and its covariance function $\Sigma(\mathbf{t}, \mathbf{s})$ which must be a non-negative definite function

Definition 2. A real function $\Sigma(\mathbf{s}, \mathbf{t})$ for $\mathbf{s}, \mathbf{t} \in T$, is non-negative definite if for all finite sets of locations $\{\mathbf{t}_1, \ldots \mathbf{t}_m\} \in T$, the (covariance) matrix:

$$\begin{bmatrix} \Sigma(\mathbf{t}_1, \mathbf{t}_1) & \Sigma(\mathbf{t}_1, \mathbf{t}_2) & \ldots & \Sigma(\mathbf{t}_1, \mathbf{t}_m) \\ \Sigma(\mathbf{t}_2, \mathbf{t}_1) & \Sigma(\mathbf{t}_2, \mathbf{t}_2) & \ldots & \Sigma(\mathbf{t}_2, \mathbf{t}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma(\mathbf{t}_m, \mathbf{t}_1) & \Sigma(\mathbf{t}_m, \mathbf{t}_2) & \ldots & \Sigma(\mathbf{t}_m, \mathbf{t}_m) \end{bmatrix} \tag{1}$$

is non-negative definite.

An important property of the multivariate normal distribution is that its predictive distribution given that one has observed the process at some finite number locations is explicit. More precisely, if one has observed the process $X(\mathbf{t})$ at $\mathbf{t} = \{\mathbf{t}_1, \ldots \mathbf{t}_m\}$, then the distribution of the process at $\mathbf{s} = \{\mathbf{s}_1, \ldots \mathbf{s}_m\}$ is normal with mean

$$\mu(\mathbf{s}) + \mathbf{\Sigma}_{\mathbf{s},\mathbf{t}} \mathbf{\Sigma}_{\mathbf{t}}^{-1} (X(\mathbf{t}) - \mu(\mathbf{t})),$$

where and covariance matrix

$$\mathbf{\Sigma_s} - \mathbf{\Sigma_{s,t}}\mathbf{\Sigma_t}^{-1}\mathbf{\Sigma_{t,s}}.$$

Here, $\mathbf{\Sigma_{s,t}}$ denotes the covariance between the vectors $X(\mathbf{s})$ and $X(\mathbf{t})$.

0.4 Differentiability

Even though, the finite dimensional distribution determine a stochastic process, many sample path properties are hard to investigate by just examining the finite dimensional distribution. An important example of this is continuity and differentiability of sample paths. Fortunately there are results that makes it easier to determine if a process has continuous sample paths:

Theorem 1. Assume that the process $X(\omega, \mathbf{t})$, where $\mathbf{t} \in [0,1]^d$, satisfies

$$\mathsf{E}[|X(\mathbf{t}+\mathbf{h}) - X(\mathbf{t})|^p] \leq \frac{K|\mathbf{h}|^d}{|\log|\mathbf{h}||^{1+r}},$$

where p, r and K are positive constants, $p < r$. Then the processes has a version with continuous sample paths.

If a processes is weakly stationary with covariance function r, then by setting $p = 2$, the theorem states that if

$$r(\mathbf{h}) - r(0) = O\left(\frac{|h|^d}{|\log|h||^{1+r}}\right)$$

for $r > 2$, then the he processes has a version with continuous sample paths.

Note that, the theorem above only states sufficient conditions. In fact if the processes is Gaussian, the requirement can be greatly reduced to the dimension-free condition:

Theorem 2. If a Gaussian process $X(\omega, \mathbf{t})$, where $\mathbf{t} \in [0,1]^d$, with continuous mean function, satisfies

$$\mathsf{V}[X(\mathbf{t}+\mathbf{h}) - X(\mathbf{t})^2] \leq \frac{K}{|\log|\mathbf{h}||^r},$$

where $r > 3$. Then it has a version with continuous sample paths.

1 The Rice distribution

The behavior of a stochastic processes can often be different from what one would expect. A good example of this is the behavior of the derivative for the process at level crossings. For example, consider a stationary ergodic Gaussian processes $X(t)$, where the parameter space is $T = \mathbb{R}$, with continuously differentiable sample paths. Since, for a fixed $t_0 \in T$, $X(t_0)$ and $X'(t_0)$ are independent, one expects that for

$$P(X'(t) \in B | X(t) = u) = P(X'(t) \in B) = \frac{1}{\sqrt{2\pi\lambda}} \int_B e^{-\frac{x^2}{2\lambda}}\, dx,$$

where λ is the variance of $X'(t)$. However, some care need to be taken when conditioning on $A = \{X(t) = u\}$ since $P(A) = 0$ and we are thus conditioning on something that does not occur almost surely. To overcome this problem one can try to give meaning to conditioning on A by defining it as

$$P(\cdot|A) = \lim_{\delta \to 0} \frac{P(\cdot, A_\delta)}{P(A_\delta)},$$

where A_δ has the following two properties: $P(A_\delta) > 0$ for each δ and $\lim_{\delta\to 0} A_\delta \to A$. A set of events belongs to A_δ if it satisfies the two previously mentioned properties. For instance, $A^1_\delta = \{x(t) \in [u, u+\delta]\}$ belongs to A_δ. It can then be shown that

$$P(B|A) = \lim_{\delta \to 0} \frac{P(\{x'(t) \in B\} \cap A^1_\delta)}{P(A^1_\delta)} = \frac{1}{\sqrt{2\pi\lambda}} \int_B e^{-\frac{x^2}{2\lambda}}\, dx.$$

So we get the answer we expected. However, A^1_δ is not the only possible sequence of events in A_δ. Another possible A_δ sequence of events is "there exists a $\hat{t} \in [t, t+\delta]$ such that $x(\hat{t}) = u$" which we denote by A^2_δ. Then

$$P(B|A) = \lim_{\delta \to 0} \frac{P(\{x'(t) \in B\} \cap A^2_\delta)}{P(A^2_\delta)} = \int_B \frac{|x|}{2\lambda} e^{-\frac{x^2}{2\lambda}}\, dx.$$

We now have two different answers (see Figure 1) to the same question, so which of these two distributions, if any, is actually observed on trajectories of the process? To answer that, one needs to study the empirical distribution, $\hat{P}$, of the behavior at u-level crossings in some interval $[a, b]$. Let $N_{[a,b]}(u)$ denote the number of t's

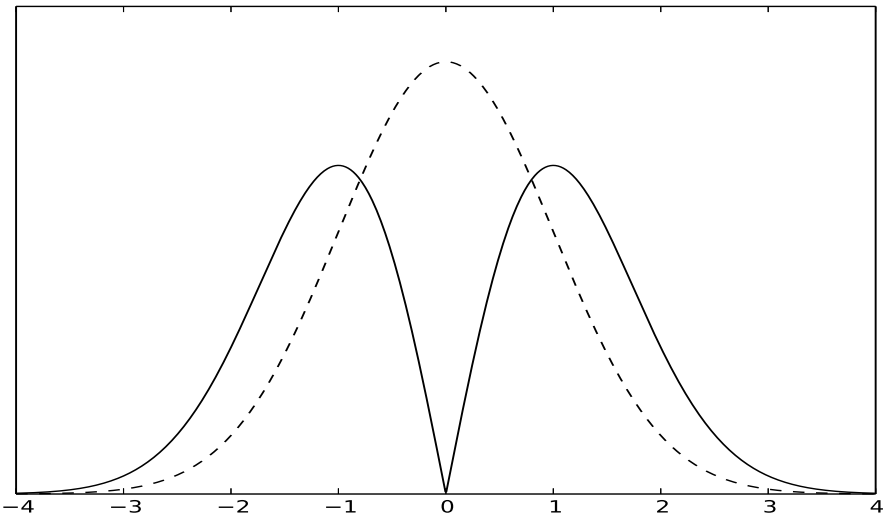


Figure 1: Two possible densities for the probability density function of $X'(t)$ given that $X(t) = u$.

such that $x(t) = u$ for $t \in [a, b]$, and let $N_{[a,b]}(B, u)$ denote the number of t's such that $x(t) = u, x'(t) \in B$ for $t \in [a, b]$. Then

$$\hat{P}(B|A) = \frac{\hat{P}(B, A)}{\hat{P}(A)} = \frac{N_{[a,b]}(B, u)}{N_{[a,b]}(u)}.$$

The limit ($|a-b| \to \infty$) of the above ratio defines (as a function of B) the distribution we want. By the ergodicity of the process, the limit equals $\frac{\mathsf{E}[N_{[a,b]}(B,u)]}{\mathsf{E}[N_{[a,b]}(u)]}$. It is enough to consider $[a, b] = [0, 1]$ since the process is stationary. The following result summarize our discussion

Theorem 3. (Rice's formula). For any stationary process $\{x(t), t \in \mathbb{R}\}$ that has continuously differentiable sample paths and density $f_{x(0)}(u)$, the crossing intensities are given by

$$\mu(u) = \mathsf{E}[N_{[0,1]}(u)] = \int_{-\infty}^{\infty} |z| f_{x(0),x'(0)}(u, z)\, dz. \tag{2}$$

A similar result, often referred to as a generalized Rice formula is giving a similar expression for $\mathsf{E}[N_{[a,b]}(B,u)]$. In the Gaussian case plugging in the formula from the theorem (and some calculations) leads to

$$P(B|A) = \frac{\mathsf{E}[N_{[0,1]}(B,u)]}{\mathsf{E}[N_{[0,1]}(u))]} = \int_B \frac{|x|}{2\lambda} e^{-\frac{x^2}{2\lambda}}\, dx, \tag{3}$$

Thus the empirical distribution will correspond to the limit of $P(B|A^2_\delta)$, and not of, as some (including the author) would expect, $P(B|A^1_\delta)$.

For more details about crossings and Rice's formula see for instance (Azaïs & Wschebor, 2009), (Lindgren, 2012), or the paper Kac & Slepian (1959), on which this section is heavily influenced by.

2 Lévy processes

A Lévy process, L, on $\mathbb{R}^+$ is a process that is stochastic continuous with stationary independent increments and $L(0) = 0$. In $\mathbb{R}^d$ a Lévy process is not as straightforward to define as on $\mathbb{R}^+$. The following definition comes from Dalang & Walsh (1992), a Lévy process (noise) $L = (L(A),\, A \in \mathcal{B}(\mathbb{R}^d))$ where $\mathcal{B}(\mathbb{R}^d)$ are the set of all bounded Borel subset of $\mathbb{R}^d$, is a family of random variables such that

1. $L(A \cup B) = L(A) + L(B)$, if $A \cap B = \emptyset$;
2. $L(A_1), \dots L(A_n)$ are independent if $A_1, \dots, A_n$ are disjoint;
3. if $A_n \downarrow \emptyset$ then $\lim_{n\to\infty} L(A_n) = 0$ in probability (stochastic continuous).

For a Lévy process the log-characteristic function $\varkappa_A(\xi)$ of $L(A)$ has the form

$$\varkappa_A(\xi) = i\gamma(A)\xi - \frac{1}{2}\sigma^2(A)\xi^2 + \int e^{i\xi x} - 1 - i\xi x \mathbb{I}(|x| \le 1)\, \nu_A(dx),$$

where γ is a signed measure, σ^2 is a non-negative measure, and ν_A is a Lévy measure, i.e., a non-negative $\sigma-$ finite measure s.t for all A

$$\nu_A\{x:\ |x| \ge 1\} < \infty,\ \ \nu_A(0) = 0,\ \text{and} \int_{|z|<1} z^2 \nu_A(dx) < \infty.$$

The formula above, is known as the Lévy-Khinchin formula (see Adler *et al.* (1983)). Lévy-Khinchin notes that three measures $(\gamma, \sigma^2, \nu_A)$ uniquely defines

a Lévy processes and are refered to as the Lévy triplet. The first measure, γ, defines is the deterministic part of the process, the second measure, σ^2, represent the Gaussian component of the process, and the third measure, ν_A, represent the jump components of the processes.

The Levy triplet gives some basic understanding of the behavior of a specific Lévy process. Example of very basic Lévy process in $\mathbb{R}^2$ are:

1. $(|A|, 0, 0)$ which represent the regular Lebesgue measure,
2. $(0, |A|\sigma^2, 0)$ which represent a Brownian sheet,
3. $(-|A|\lambda, 0, |A|\lambda\delta_1)$ which represent a Poisson sheet.

2.1 NIG and GAL Lévy processes

The main Lévy processes studied in this thesis are the normal inverse Guassian and the generalized asymmetric Laplace Lévy processes, which means that $L(A)$ is either NIG or GAL distributed. The NIG and GAL distributions are special cases of the generalized Hyperbolic (GH) distribution.

2.1.1 The generalized Hyperbolic distribution

The GH distribution has five parameters $\sigma, \nu \in \mathbb{R}^+$, $\gamma, \mu, \tau \in \mathbb{R}$, and a density function

$$f(x) = c_1 \left(\frac{\sqrt{(\nu\sigma)^2 + (x-\gamma)^2}}{c_2} \right)^{\tau - 1/2} e^{\frac{\mu}{\sigma^2}(x-\gamma)} K_{\tau-1/2}\left(c_2 \sqrt{(\nu\sigma)^2 + (x-\gamma)^2} \right),$$

where $c_1 = \frac{2^{(\tau-1)/2}}{\sqrt{\pi}(\sigma^2\nu)^\tau K_\tau(\sqrt{2}\nu)}$ and $c_2 = \frac{1}{\sigma}\sqrt{2 + \frac{\mu^2}{\sigma^2}}$. The parametrization above is not the regular parametrization for the GH distribution; however, it is an convenient parametrization when formulating a GH random variable (r.v.) as a normal mean-variance mixture r.v., i.e. if X is a r.v. with a GH distribution then

$$X \stackrel{d}{=} \gamma + \mu V + \sigma\sqrt{V}Z,$$

where V is generalized inverse Gaussian (GIG) distributed with parameters (p, a, b) set to $(\tau, 2, \nu^2)$ and $Z \sim N(0, 1)$. The $GIG(p, a, b)$ distribution has the density

function

$$f_V(x) = \frac{\left(a/b\right)^{p/2}}{2K_p\left(\sqrt{ab}\right)} x^{p-1} e^{-\frac{ax+b/x}{2}},$$

where the parameters satisfy $a > 0, b \geq 0$ if $p > 0$, $a > 0, b > 0$ if $p = 0$, and $a \geq 0, b > 0$ if $p < 0$. A useful property of the GH distribution is that given an observation of GH r.v. X the conditional distribution of its random mixture r.v. V is again GIG distributed. This property is used in Paper C, when estimating parameters and making predications, and in Paper E, for sampling of at level-crossings.

The reason for using the NIG and GAL distributions is that both are closed under convolution (if certain parameters are fixed). A class of random variables is closed under convolution if the sum of two random variables from this class also belongs to the class. Ideally one would like a stronger alternative which is that a sum of linear combinations of two independent copies has the same distribution. Unfortunately, this property is very uncommon and, in fact, the only process with that property and finite variance is the Gaussian process.

For moments and other properties of the GH distribution see Schoutens (2003).

2.2 The NIG processes

The NIG distribution has four parameters $(\gamma, \mu, \sigma, \nu^2)$, and is a GH distribution with $\lambda = -1/2$. Its normal mean-variance mixture representation is given by letting V be inverse Gaussian distributed. An example of three symmetric ($\mu = 0$) NIG distributions with varying ν^2 is shown in Figure 2. The Lévy processes comes from letting

$$L(A) \stackrel{d}{=} NIG(\gamma|A|, \mu, \sigma, \nu^2|A|).$$

The Lévy triplet for a NIG process is $(\gamma|A| + \xi, 0, \nu_A)$ where

$$\xi = \frac{2\gamma\alpha}{\pi}|A| \int_0^1 \sinh(\frac{\mu}{\sigma^2}x) K_1(\alpha|x|)\, dx,$$
$$\nu_A(x) = \frac{\gamma\alpha|A|}{\pi|x|} e^{\frac{\mu}{\sigma^2}x} K_1(\alpha|x|),$$

and $\alpha = \sigma^{-2}\sqrt{\mu^2 + 2\sigma^{-2}}$. It may seem counter intuitive that the NIG distribution is a normal mean-variance distribution since it completely lacks Gaussian component in its Lévy triplet.

The tail of the NIG distribution is proportional to

$$|x|^{-3/2} \exp\left(-\alpha|x| + \frac{\mu}{\sigma^2}x\right).$$

2.3 The GAL processes

The GAL distribution has four parameters $(\tau, \gamma, \mu, \sigma)$, and is a GH distribution with $\nu^2 = 0$. An example of three symmetric ($\mu = 0$) GAL distributions with varying τ is shown in Figure 2. Its normal mean-variance mixture representation is given by letting V be Gamma distributed. The Lévy processes comes from letting

$$L(A) \stackrel{d}{=} GAL(\tau|A|, \gamma|A|, \mu, \sigma).$$

The corresponding Lévy triplet is given by $(\gamma|A| + \xi, 0, , \nu_A)$ where

$$\xi = \frac{\tau|A|}{MG}(M(e^{-G} - 1) - G(e^{-M} - 1)),$$
$$\nu_A(x) = \begin{cases} -\tau|A| \exp(Gx)\, x^{-1}, & x < 0, \\ \tau|A| \exp(-Mx)\, x^{-1}, & x > 0, \end{cases}$$

where $G = \left(\sqrt{\frac{1}{4}\mu^2 + \frac{1}{2}\sigma^2} - \frac{1}{2}\mu\right)^{-1}$ and $M = \left(\sqrt{\frac{1}{4}\mu^2 + \frac{1}{2}\sigma^2} + \frac{1}{2}\mu\right)^{-1}$. Like the NIG process the GAL process lacks Gaussian component in the Lévy triplet.

The tail of the GAL distribution is proportional to

$$|x|^{\tau-1} \exp\left(-\alpha|x| + \frac{\mu}{\sigma^2}x\right),$$

where $\alpha = \sigma^{-2}\sqrt{\mu^2 + 2\sigma^{-2}}$.

3 Spatial statistics

This section discuses some topics that, although relevant for any stochastic process, are most prevalent in spatial statistics.

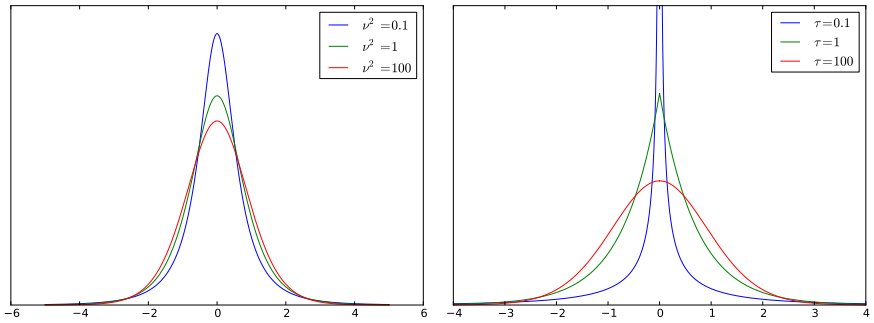


Figure 2: Left: the NIG densities with varying ν^2. Right: the GAL densities with varying τ. In both, σ^2 is chosen so that the variance for the distributions are one, and $\mu = \gamma = 0$. The NIG distribution is always differentiable, whereas the GAL distribution is not differentiable if $\tau \leq 1$ and further it is unbounded if $\tau < 1/2$.

In spatial statistics, stochastic processes are typically refereed to as fields, i.e. a stochastic process for which the dimension of the parameter space T is greater than 1. This section, is by no mean indented to cover with (any) generality spatial statistics. For this we refer to Gaetan & Guyon (2009), Cressie (1993) or Gelfand & Diggle (2010), and for a more theoretical treatment of random fields see Adler & Taylor (2007) and Ibragimov & Rozanov (1978).

3.1 Matérn covariance

This section introduces the most popular covariance function for stationary process in spatial statistics. That is the Matérn covariance family (Matérn, 1960), named after the Swedish statistician Bertil Matérn. The Matérn covariance function is defined by

$$\Sigma(\mathbf{s},\mathbf{t}) = \frac{2^{1-\nu}\sigma^2}{(4\pi)^{d/2}\Gamma(\nu+d/2)\varkappa 2\nu}(\varkappa||\mathbf{h}||)^{\nu}K_{\nu}(\varkappa||\mathbf{h}||),$$

where d is the dimension of the field, $||\mathbf{h}|| = ||\mathbf{t}-\mathbf{s}||$ is the distance between the two points, Γ is the Gamma function and $K_{\nu}(.)$ denotes the modified Bessel function of the second kind. The covariance function uses only the distance between the points thus the function is **isotropic**. Figure 3, displays a few different

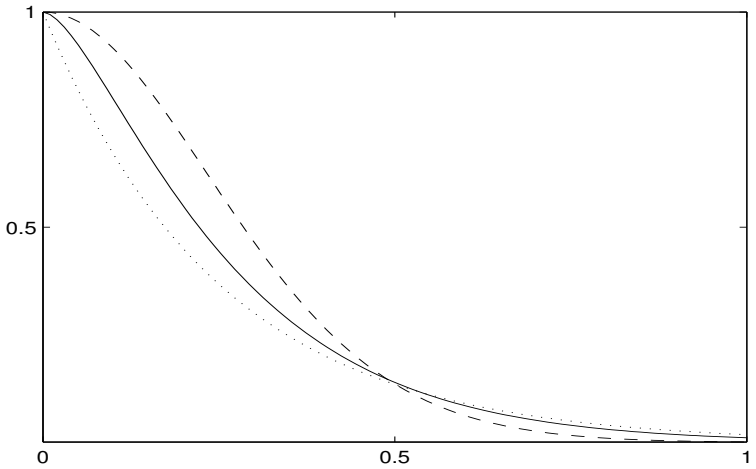


Figure 3: Matérn covariance functions for $\nu = 1$ (solid line), $\nu = 0.5$ (dotted line) and $\nu = 100$ (dashed line). Note the difference of the functions when the argument is close to zero.

covariance function from the Matérn covariance family. The reason for the popularity of the Matérn covariance is its flexible parametric form, which captures the main properties needed for interpolation of stationary spatial processes: that is, $\varkappa$ controls the dependence range, σ controlling the variability of the processes, and finally ν defining the level of differentiability of the covariance function which correspond the smoothness of the stochastic processes.

Both in Paper C and D processes with Matérn covariances are used; it should be noted that the parameter ν is not estimated for these models, due to some difficulties with numerical approximation, and some of the flexibility of the covariance is lost.

3.2 Measurement error

When using stochastic process in practice, it is often unreasonable to assume that one observe the process $X(\mathbf{t})$ directly; usually the observations are noisy or the model for $X(\mathbf{t})$ does not capture the behavior of the true process. To improve the

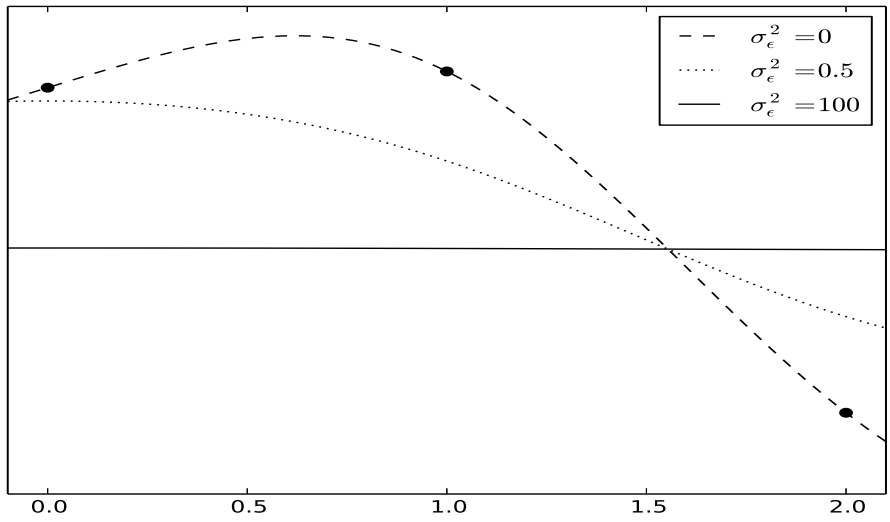


Figure 4: The effect of adding measurement error. The underlying process, $X(t)$, is stationary Gaussian with zero mean, and Matérn covariance function such that $\mathsf{V}[X(t)] = 1$. The three black dots represent the observations. The lines display the posterior mean of the process for the varying variance of the measurement errors.

agreement of the model with reality, one assumes that the measured observations $y(\mathbf{t})$, equals the process $X(\mathbf{t})$ plus some measurement error. The adding of the measurement error avoids over-fitting of the process to the data. Also, if the estimated variance of the measurement error is high, then this is an indication that the process does not represent the true process (or that the measurement really are noisy). Figure 4 describes the effect of the measurement error variance for fitting the process to data.

Modeling with measurement error generates a hierarchical model. At the top-level is the distribution of the data model $\mathsf{P}(y(\mathbf{t})|X(\mathbf{t}), \vartheta)$, that is the distribution of the measurement given the process and some parameters. At the next level is the process model distribution $\mathsf{P}(X(\mathbf{t})|\vartheta)$, that specifies the distribution of the process of interest. By defining another layer with some prior distribution $\mathsf{P}(\vartheta)$ one creates a Bayesian hierarchical model.

3.3 GMRF

An important subclass of Gaussian random fields are the Gaussian Markov random fields (GMRFs). To properly define a GMRF for a discrete vector of random variables, one needs the notation of neighborhood and neighborhood-system: For the random variable $\mathbf{x} = [x_1, \ldots, x_n]$ with distribution f, a neighborhood of x_i is a set of indices $\mathcal{N}_i$ such that $f(x_i|x_{-i}) = f(x_i|x_{\mathcal{N}_i})$; the neighborhood system of $\mathbf{x}$ is the set $\mathcal{N} = \{\mathcal{N}_i\}_{i=1}^N$.

A random variable

$$\mathbf{x} \sim N(0, \mathbf{Q}^{-1}),$$

where $\mathbf{Q}$ is the precision matrix (the inverse of the covariance matrix), is a GMRF for the neighborhood-system $\mathcal{N}$ if

$$Q_{i,j} = 0 \iff j \notin \mathcal{N}_i.$$

This includes any multivariate normal random variables, however the useful case is when most of the neighborhoods are small compared to the size of $\mathbf{x}$. The advantage with formulating a GMRF compared to the regular multivariate distribution is, at least, twofold: firstly, it defines the distribution of the r.v. through its conditional distributions, which typically is much easier to understand and deal with compared to the joint distribution, although care needs to be taken so that the conditional distributions form a valid joint distribution. Secondly, when the neighborhoods are small, the precision matrix is sparse, and efficient numerical methods can be used to simulate $\mathbf{x}$ and compute $f(\mathbf{x})$.

GMRFs have been used especially in spatial statistics on lattice domains, when defining neighborhoods for $X(\mathbf{t})$ is natural. A frequently used GMRF model is the Gaussian conditional autoregressive model (CAR), Besag (1974) is an influential paper that popularized the models. The model can be though of as the lattice counterpart of the autoregressive model that is fundamental in time series analysis. For further details on GMRFs, see Rue & Held (2005).

3.4 The SPDE approach

A long time ago, Whittle (1954) linked the linear stochastic differential equation (SPDE)

$$(\varkappa^2 - \Delta)^{\alpha/2} X = \varphi^2 \mathcal{W}, \tag{4}$$

to a Gaussian random field with Matérn covariance; the solution, X, of the differential equation, where $\mathcal{W}$ is a Brownian sheet, is a Gaussian random field with the Matérn covariance. Not so long ago, in Lindgren *et al.* (2011), the knowledge of the link together with the finite element method Strang & Fix (1973), was utilized to create computationally efficient methods for estimation and prediction of the corresponding Gaussian random field. Basically, the idea is to approximate $X(\mathbf{s})$ with a sum of deterministic basis function with stochastic weights

$$\hat{X}(\mathbf{s}) = \sum_{i=1}^{n} \psi(\mathbf{s}) w_i. \tag{5}$$

To define the weights distribution, one requires that the approximation satisfies a weak formulation of the SPDE with respect to some test functions ψ_i, $i = 1, \ldots, n$:

$$\int \varphi_i(\mathbf{s}) \hat{X}(\mathbf{s}) \, d\mathbf{s} \stackrel{d}{=} \varphi^2 \int \varphi_i(\mathbf{s}) \mathcal{W} \, d\mathbf{s},$$

for $i = 1, \ldots, n$. The equation above can be formulated in matrices form as

$$\mathbf{K}\mathbf{w} \stackrel{d}{=} \mathbf{z}$$

where $K_{i,j} = \int \psi_i(\mathbf{s})(\varkappa^2 - \Delta)^{\alpha/2} \varphi_j(\mathbf{s}) \, d\mathbf{s}$, $\mathbf{w}$ are the weights in equation (5) and $\mathbf{z} \sim N(\mathbf{0}, \mathbf{C}^{-1})$ where $C_{i,j} = \int \psi_i(\mathbf{s}) \varphi_j(\mathbf{s}) \, d\mathbf{s}$. After, some technical tricks, the resulting distribution of $\mathbf{w}$ is a GMRF with sparse precision matrix.

Other advantages of the SPDE approach it is that the PDE formulation is suitable to create non-stationary Gaussian processes, see (Bolin & Lindgren, 2011).

3.5 Fields generated with Lévy noise

A subset of the non-Gaussian random fields are the fields generated by a stochastic integral with respect to a Lévy measure. For example, a stationary the random field $X(\mathbf{t})$ is obtained by

$$X(\mathbf{t}) = \int f(\mathbf{s} - \mathbf{t}) \, dM(\mathbf{s}), \tag{6}$$

where M is a Lévy measure. The function f controls the dependence structure of the process. In Åberg & Podgórski (2010), this type of process was studied where

the convolution above was with respect to a Laplace measure, and the function f was chosen so that the covariance function of the process had Matérn covariance. In the paper all parameters were fitted through the method of moments. In Bolin (2013) it was shown that the SPDE approach, introduced above, can generate random field, of the type defined in equation (6). If $X(\mathbf{t})$ satisfies

$$(\varkappa^2 - \Delta)^{\alpha/2} X(\mathbf{t}) = \varphi^2 \dot{M},$$

where M a Lévy measure, then it is, at least in distribution, equivalent to $X(\mathbf{t})$ in (6) where f is the Green function of the PDE. In Bolin (2013), estimation of the parameters was done through the maximum likelihood method.

Although the distribution of $X(t)$ has no known explicit form, it is easy to see from the characteristic function that $X(t)$ has more flexible marginals than its Gaussian counterpart. Since the finite marginal distributions determines a stochastic process, it is reasonable to expect that the trajectories of $X(\mathbf{t})$ should behave different from the trajectories generated by a Gaussian processes with the same covariance function. This assumption is verified in Adler *et al.* (2013), where it is shown that for high level excursions of a process generated by Lévy field without Gaussian component (recal Lévy-Khinchin formula) the behavior is fundamentally different from a Gaussian field generated by the same moving average function.

3.6 Interpolation

The most common problem in spatial statistics is the problem of interpolation. From a statical perspective the question about interpolation is how to predict the process at some locations, $\mathbf{t}_p$ (usually unobserved) given that the process is observed at $\mathbf{t}$ (possible with measurement error). It is not possible to mention statistical interpolation without mentioning kriging. Kriging, named after the South African mining engineer D. G. Krige (Krige, 1951), is the best linear unbiased predictor (BLUP), where best refers to the prediction with the least mean square error. It is by far the most used method in statistics for preforming interpolation. Given a known mean function, $\mu(\mathbf{t})$, and covariance function, $\Sigma(\mathbf{t}, \mathbf{s})$, kriging is an explicit formula of the observed data $\mathbf{y}(\mathbf{t})$, namely:

$$\hat{\mathbf{y}}(\mathbf{t}_p) = \mu(\mathbf{t}_p) + \mathbf{\Sigma}_{\mathbf{t}_p,\mathbf{t}} \mathbf{\Sigma}_{\mathbf{t},\mathbf{t}}^{-1} (y(\mathbf{t}) - \mu(\mathbf{t})).$$

Note that this equivalent to the conditional mean for a Gaussian process, in fact for a Gaussian process the kriging predictor is the best predictor linear or

not. For non-Gaussian processes this is not the case. An illuminating, although somewhat artificial, example on how poorly the BLUP can preform versus the best predictor is given in Stein (1999).

For the models considered in Papers C and D we preform interpolation using Gibbs sampling, by calculating the posterior mean of the random field at the location $\mathbf{t}_p$. However, for a non symmetric distribution using the mean as a point predictor might not always be what one actual is looking for, both from practical and theoretical perspectives. Often the mode of the distribution is a better alternative, in the sense of what a non-statistician would consider as the best guess, i.e. the most likely value.

4 Parametric inference

This section gives a brief overview of the method used in the preceding papers for estimation of parameters. We do not assume that observations are from processes here and thus use slightly different notation compared to earlier sections. Throughout this section we assume that a random vector Y is observed at y.

Y belongs too parametric family if its distribution $p(.;\vartheta)$ is completely determined by a vector of parameters ϑ. Typically, the parameters are unknown and thus it is necessary to estimate them for making inference about Y. The most popular estimator is the maximum likelihood estimator, i.e. the estimator

$$\hat{\vartheta} = \text{argmax}_{\vartheta}\, p(y;\vartheta).$$

If $\nabla_{\vartheta} p(y;\vartheta)$ is available, a common estimator is the z-estimator $\hat{\vartheta}$, satisfying

$$\nabla_{\vartheta} p(y;\hat{\vartheta}) = 0.$$

Ideally the estimator $\hat{\vartheta}$, if not given analytically, is found using some standard optimization or root finding method like for example the Newton method. However, in many, if not most, situations direct optimization of the likelihood is not possible. Below, two methods, often used when the likelihood is unsuitable for direct optimization, are presented: the EM-algorithm, and the less known Expectation Conjugate Gradient algorithm.

4.1 EM-algorithm

The Expectation Maximization algorithm, first introduced in Dempster *et al.* (1977), is an iterative procedure for maximizing the likelihood function. In many

cases the likelihood for the data y is intractable, however the data augmented with a r.v, X, often produces a tractable likelihood $p(y, X; \vartheta)$. In these situations the EM-algorithm is useful since it does not require evaluation of the data likelihood but it is enough to use $p(y, X; \vartheta)$.

The ith iteration of the algorithm consists of two steps

- $E - step$: Compute the function

$$\mathcal{Q}(\vartheta, \vartheta^{(i-1)}) = \mathsf{E}[\log p(y, X; \vartheta) | Y = y; \vartheta^{(i-1)}]$$

- $M - step$: Preform the maximisation

$$\vartheta^{(i)} = \text{argmax}_{\vartheta}\, \mathcal{Q}(\vartheta, \vartheta^{(i-1)}).$$

Here the $E - step$ denotes the Expectation step and $M - step$ denotes the Maximization step. Under mild conditions the EM-algorithm converges to a stationary point, Wu (1983). The rate of convergence for the EM-algorithm is linear making it an annoyingly slow optimization algorithm.

From a practical perspective the EM-algorithm is often stable and requires no tuning of any parameters by the user, and these are likely two reasons for its popularity.

For many models the $E - step$ is not explicitly available, an alternative is to replace the $\mathcal{Q}$ with an Monte Carlo approximation. The modification to the EM-algorithm is known as the MCEM algorithm,Wei & Tanner (1990). Typically, for the algorithm to be practically useful the non-deterministic information of $\mathcal{Q}$, that needs to obtained through an MC sampler, should be contained in a few sufficient statistics. In paper C, we utilize an MCEM-algorithm to estimate the parameters for the Levy-random fields.

For a thorough introduction to the EM-algorithm we refer to Meng & Van Dyk (1997).

4.2 Expectation Conjugate Gradient

A close relative to the EM-algorithm is the the Expectation Conjugate Graident algorithm (ECG) Lange (1995a). The method is a regular conjugate gradient method split into two steps:

1. $E - step$: Calculate the gradient of the log likelihood using that

$$\nabla_{\vartheta} \log p(y; \vartheta^{(i-1)}) = \mathsf{E}[\nabla_{\vartheta} \log p(y, X; \vartheta) | Y = y; \vartheta^{(i-1)}]|_{\vartheta = \vartheta^{(i-1)}}$$

2. $S-step$: take a CG step:

$$
\begin{aligned}
g_{i-1} &= \nabla_\vartheta \log p(y;\vartheta^{(i-1)}),\\
d_i &= \begin{cases} g_{i-1} & i=1,\\ g_{i-1}+\frac{||g_{i-1}||^2}{||g_{i-2}||^2}d_{i-1} & i>1,\end{cases}\\
\vartheta^{(i)} &= \vartheta^{(i-1)}+\alpha d_i.
\end{aligned}
$$

Here $\alpha > 0$ is the step length.

Like the EM-algortihm, the ECG-algorithm does not require the ability to calculate the full data log likelihood $\log p(y;\vartheta)$, in fact it enough to calculate $\nabla_\vartheta \log p(y;\vartheta^{(i-1)})$ which sometimes, see Paper D, can be easier than calculating $\log p(y;\vartheta^{(i-1)})$. An advantage over the EM-algorithm is that the $S-step$ is often much faster than the $M-step$, and since both algorithms have the same rate of convergence, making the ECG-algorithm much faster in practice. Also there exist methods for improving the rate of convergence of the ECG-algorithm see Lange (1995b).

Approximating the $E-step$ as for MCEM algorithm results in a stochastic gradient descent algorithm, which is an optimization method that has become popular for the popular "big-data" problems, see Bottou (2004). A big advantage over the MCEM algorithm is that one does not need to worry about sufficient statistics since all we need to store is the gradient in each Monte Carlo iteration.

The stochastic gradient descent algorithm is used to estimate parameters in paper D, and in a paper in progress we use the stochastic gradient for more general spde models driven by the same noise as in Paper C.

5 Outline of the papers

Paper A: Maximizing leave-one-out likelihood for the location parameter of unbounded densities

In this paper a new type of estimator is introduced, it is developed to handle estimation of distribution that is unbounded at the mode. For a density f the estimator, $\widehat{\delta}$, is the argument that maximizes

$$
l_n(\delta) = \frac{\prod_{i=1}^n f(X_i-\delta)}{f(X_{k(\delta)}-\delta)},
$$

where $k(\delta) = \text{argmin}_{k\in\{1,\dots,n\}} |X_k - \delta|$. In the paper, it is shown that the estimator can be super-efficient, that is the rate of convergence can be faster then $n^{1/2}$, which is the standard rate of an estimator. The rate of convergence is shown to be nearly the optimal one, however if the estimator is optimal or not remains an open problem. The intended application is to fit parameters of the generalized asymmetric Laplace distribution, for which the distribution is singular in certain section of the parameter space. The main contribution of the paper is an estimator that is almost optimal for singular distribution and also, similarly to the maximum likelihood method, is well suited for a multi-parameter setting.

Paper B: Convolution invariant subclasses of generalized hyperbolic distributions

The invariance under convolution of a class of distribution is an important and desired property, in particular, when dealing with irregular discretization of meshes that are typical for the SPDE methods in section 3.4. Since the Gaussian mean variance mixtures leads to important distributions, determining parametric models of convolution invariant classes is an important problem. In this paper, it is proven that only two subclasses of the GH distributions are closed on the convolution, namely the NIG and GAL distributions. This result has been previously quoted (sometimes mistakenly) in literature, but not proven. The main contribution of this paper is rigorously proving the results.

Paper C: Non-Gaussian Matérn fields with an application to precipitation modeling

This paper deals with random fields generated by Lévy noise introduced in Bolin (2013). The marginal distributions of the fields are flexible, allowing for varying shape, asymmetry and some flexibility in the tails. The main focus of the paper is formulating models and methods so that the processes can be used for real data; important contributions are: accounting for measurement error, stochastic estimation methods for the parameters, and methods for efficiently preforming prediction at unobserved locations. The model for the observed data $\mathbf{y}$ can be formulated as a hierarchical model:

$$\begin{aligned}\mathbf{y} &\sim N(\mathbf{B}\boldsymbol{\beta} + \mathbf{A}\mathbf{w}, \sigma_\varepsilon^2\mathbf{I}),\\ \mathbf{w}|\mathbf{V} &\sim N(\boldsymbol{\mu}_\mathbf{V}, \boldsymbol{\Sigma}_\mathbf{V}),\\ \mathbf{V} &\sim \pi(\mathbf{V})\end{aligned}$$

where $\pi(\mathbf{V})$ is either gamma distribution or inverse Gaussian distribution, and $\mathbf{w}$ are the random weights of the basis approximation to the solution of the SPDE discretization described in section 3.4. Using that the distributions of $\mathbf{w}|\mathbf{V},\mathbf{y}$ and $\mathbf{V}|\mathbf{w},\mathbf{y}$ are explicit allows for efficient generation of the posterior distribution of the random field. The posterior distribution is used in both estimation of the parameter (through MCEM-algorithm) and in prediction at unobserved locations.

Finally, the model is fitted to rain data for a region in Brazil. The predictive ability of the models compared to regular Gaussian random fields and transformed Gaussian random fields using by cross validation.

Paper D: A Gaussian mixture model for multivariate spatially dependent data using discrete and continuous Markov random fields

This paper introduced a novel type of random fields, combining two random fields models: the classical (multivariate) latent Gaussian random fields and a Potts model (or discrete a Markov random field). A Potts model, $\mathbf{x}$, is a discrete valued random field, typically defined on a regular grid. The value of a node, x_i depends on the values of a predefined neighborhood $\mathcal{N}_i$, typically used to cluster nodes with same value. More specifically, the latent model is defined as a mixture of multivariate Gaussian random fields and which field that is observed is determined through the Potts model. A simplified model can be described through the following hierarchical model:

$$\begin{aligned}
\mathbf{y}_i &\sim N(\boldsymbol{\xi}_{i,x_i}, \sigma^2_\varepsilon),\\
\boldsymbol{\xi}_{\cdot,j} &\sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),\ i = 1,\ldots,K,\\
\mathbf{x} &\sim\ mrf(\boldsymbol{\alpha},\boldsymbol{\beta}).
\end{aligned}$$

The conditional distribution of a node in the Markov random field is

$$\mathsf{P}(x_i = k|\mathcal{N}_i,\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\exp\left(\alpha_k + \beta_k f_{i,k}\right)}{\sum_{j=1}^{K}\exp\left(\alpha_j + \beta_j f_{i,j}\right)},$$

where $f_{i,k}$ is the number of neighbourhoods of i with value k.

There are many possible applications for the models, ranging from interpolation of missing pixel values of images to detection of underlying soil types from multivariate spatially dependent measurements of chemicals. Since many of intended applications will have massive data sets, methods for efficient estimation

are developed. Finally, an example where the model is used for smoothing of an MR image is presented.

Paper E: Slepian model for moving averages driven by a non-Gaussian noise

In this paper, a Slepian model of a non-Gaussian moving average model is studied. The main focus is an moving average models driven by symmetric GAL noise. The Slepian model describes the behavior of a processes around a level crossing. The distribution of the Slepian model is typically derived from Rice formula described in section 1. Since a closed form of the Slepian model exists only for Gaussian processes, the problem of finding effective ways to generate Slepian models beyond the Gaussian domain is of great importance. We develop efficient simulation method for moving average models driven by Laplace noise. The simulation methods relies on the fact the Laplace noise can be represented as subordinated Brownian motion, where the subordinate is a Gamma process. More precisely, the moving average model can be written as a hierarchical model:

$$
\begin{aligned}
x_t|\mathbf{L} &= \int f(t-s)\mathbf{L}, \\
\mathbf{L}|\mathbf{K} &\sim N(0,\mathbf{K}), \\
\mathbf{K} &\sim \Gamma(\tau\mathbf{dt},\mathbf{1}),
\end{aligned}
$$

Then discretizing the problem and using that $\mathbf{K}|\mathbf{L}$ and $\mathbf{L}|\mathbf{x},\mathbf{K}$ are known distribution one can, after several steps, generate samples from the Slepian model using a natural Gibbs sampler. The novelty in our approach is that we study Slepian models for noise that is driving the considered stochastic process. By using such the Slepian noise one can simultaneously analyse complex Slepian models at the crossings by simply replacing in the original formulation of the models the original noise by the Slepian noise.

References

Åberg, S. & Podgórski, K. (2010). A Class of Non-Gaussian second order Spatio-Temporal Models. *Extremes* **14**, 187–222.

Adler, R., Monrad, D., Scissors, R. & Wilson, R. (1983). Representations, decompositions and sample function continuity of random fields with independent increments. *Stochastic Processes and their Applications* **15**, 3 – 30.

Adler, R. J., Samorodnitsky, G. & Taylor, J. E. (2013). High level excursion set geometry for non-gaussian infinitely divisible random fields. *The Annals of Probability* **41**, 134–169.

Adler, R. J. & Taylor, J. E. (2007). *Random fields and geometry*, vol. 115. Springer.

Azaïs, J.-M. & Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. John Wiley & Sons.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* , 192–236.

Bolin, D. (2013). Spatial Matérn fields driven by non-Gaussian noise (in press). *SJS* .

Bolin, D. & Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics* **5**, 523–550.

Bottou, L. (2004). Stochastic learning. In O. Bousquet & U. von Luxburg, eds., *Advanced lectures on machine learning*, Lecture Notes in Artificial Intelligence, LNAI 3176. Springer Verlag, Berlin, pp. 146–168.

Cressie, N. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley.

Dalang, R. C. & Walsh, J. B. (1992). The sharp markov property of lévy sheets. *The Annals of Probability* **20**, 591–626.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* , 1–38.

Gaetan, C. & Guyon, X. (2009). *Spatial statistics and modeling*. Springer Series in Statistics. Springer.

Gelfand, A. & Diggle, P. (2010). *Handbook of spatial statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis Group.

Ibragimov, I. & Rozanov, I. (1978). *Gaussian random processes*. Applications of mathematics. Springer-Verlag.

Kac, M. & Slepian, D. (1959). Large excursions of gaussian processes. *The Annals of Mathematical Statistics* **30**, 1215–1228.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Jnl. C'hem. Met. and Min. Soc. S. Afr* .

Lange, K. (1995a). A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* , 425–437.

Lange, K. (1995b). A quasi-newton acceleration of the em algorithm. *Statistica sinica* **5**, 1–18.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 423–498.

Lindgren, G. (2012). *Stationary stochastic processes: Theory and applications*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut* **49**.

Meng, X.-L. & Van Dyk, D. (1997). The em algorithm- an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 511–567.

Rue, H. & Held, L. (2005). *Gaussian Markov random fields; theory and application*, vol. 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Schoutens, W. (2003). *Levy processes in finance: Pricing financial derivatives*. Wiley Series in Probability and Statistics. Wiley.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

Strang, G. & Fix, G. J. (1973). *An analysis of the finite element method*, vol. 212. Prentice-Hall Englewood Cliffs.

Wei, G. C. & Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434–449.

Wu, C. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics* **11**, 95–103.

A

Paper A

Maximizing leave-one-out likelihood for the location parameter of unbounded densities

Krzysztof Podgórski[1] and Jonas Wallin[2]

[1] *Department of Statistics, Lund University, Sweden*
[2] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*

Abstract

We propose an approach to estimation of the location parameter for a density that is unbounded at the mode. The estimator maximizes a modified likelihood in which the singular term in the full likelihood is left out, whenever the parameter value approaches a neighborhood of the singularity location. The consistency and super-efficiency of this maximum leave-one-out likelihood estimator is demonstrated through a direct argument. The importance for estimation in parametric families of distributions is discussed and illustrated by an example involving the gamma mixture of normal distributions.

Key words: unbounded likelihood and location parameter and super-efficiency and generalized asymmetric Laplace distribution

1 Introduction

The classical problem of the location parameter estimation frequently serves as an illustration of how the asymptotic theory can be used to identify an estimator with some optimal properties. In particular, the asymptotics for the maximum likelihood estimators (MLE) has been established not only under the so-called regular conditions but also when the density has a cusp at its mode. The history here goes back to the Ph.D. Thesis of Prakasa Rao, Rao (1966), and the subsequent related paper Rao (1968), where consistency and super-efficiency of the MLE of the location parameter have been demonstrated for a bounded density with a cusp at the mode.

Estimation of the location can be also considered for an unbounded density. This case has been first approached in Ibragimov & Khasminskii (1981a) and later summarized in the influential monograph Ibragimov & Khasminskii (1981b), where, to deal with the unboundedness of the likelihood, Bayesian estimation has been considered. There, as well as in Rao (1966), weak convergence of the log-likelihood ratio process to an appropriately defined Gaussian process has been established yielding the consistency for the MLE, whenever this is well defined, or otherwise for Bayesian-type estimators.

This work also deals with the unbounded density case but instead of resorting to the Bayesian approach we modify the likelihood approach. A modification is needed since the likelihood is unbounded at each data point and the classical MLE is not even properly defined. To remedy this issue, we propose to leave a singular term out from the full likelihood in a neighborhood of the datum location and define an estimator $\hat{\delta}$ that maximizes the *leave-one-out* likelihood function. Under rather natural conditions it is shown that $\hat{\delta}$ is consistent. Moreover, a lower bound for the rate of convergence is established showing, in particular, that the estimator is super-efficient, i.e. its rate is faster than in the classical case of $n^{-1/2}$. The proof presented is completely self-contained, direct, and uses only elementary arguments. Consequently, it is formally independent of any other asymptotic results, including these for the convergence of the likelihood ratio process. Nevertheless, the intuitive reason for the supper-efficiency is the rate of convergence the likelihood ratio process (or its moments as exploited in this work). Namely, for the densities that are unbounded this rate is faster than under the standard regular conditions, see Lemma 4.5 (this faster rate is tied to the asymptotics of the density around the location parameter as presented in Lemma 4.4).

The idea of leaving out a trouble causing factor in the likelihood seems to be

quite natural and, in fact, has been recently proposed in the problem of estimation of parameters for a finite mixtures of normal densities in Seo & Kim (2012). Despite general similarities between the approaches, neither the estimators nor the results of that work translate to the setup of this paper.

The paper is organized as follows. Section 2 motivates the problem and, in particular, points at convenience of the method when used in a general multiparameter setup. In Section 3, we present the assumptions and the main result which is Theorem 3.1. In Section 4, we formulate and prove the lemmas that eventually lead to the proof of Theorem 3.1 presented in Section 5. Finally, in the Appendix, we present an example illustrating how a version of the EM algorithm can be applied to maximize the leave-one-out likelihood.

2 Motivation

Although in this work we concentrate on the location parameter, the applicability of the approach extends to the multiparameter context. The leave-one out likelihood function presents only a slightly modified likelihood and thus the maximizers over other than location parameters would have the asymptotic properties dictated by the classical MLE theory given, of course, that appropriate assumptions of the likelihood are satisfied. For this reason, the proposed estimation of location in the unbounded density case is not only of a theoretical interest but also have important implications for actual estimation problems. In fact, there are natural parametric families for which estimation in the presence of unboundedness becomes an important practical issue. This study was inspired by investigation of applicability of the EM algorithm to parameter estimation for linear models involving the generalized Laplace distributions.

Recall a generalized Laplace random variable X admits the representation $X = \delta + \mu\Gamma + \sigma\sqrt{\Gamma}Z$, where Γ has Gamma distribution with the shape τ and scale one, while Z has the standard normal distribution, see Kotz *et al.* (2001) for details. This class is made of infinitely divisible distributions, is closed under the convolutions and the corresponding Lévy motions are referred to as the Laplace motions (in mathematical finance, specially in the symmetric case, these models are naturally known as the gamma variance processes). The density of X is of the form $p(x)|x|^{\alpha}$, where $\alpha = 2\tau - 1$ and $p(x)$ being a function that is bounded and non-negative around zero.

The explicit form of the density involves one of the Bessel functions so the

distribution is also refered to as the Bessel function distribution. To maximize the likelihood one has to resort to numerical methods and, for example, the EM (expectation-maximization) algorithm can be conveniently employed to evaluate the MLE of the parameters $(\delta, \mu, \sigma, \tau)$. We refer to Protassov (2004) for a presentation of such an approach applied to a subclass of the generalized hyperbolic distributions (the latter were introduced by Barndorff-Nielsen (1978) and include also the generalized Laplace distributions). Since the range of values of τ is a'priori not known, one can not exclude a possibility of an unbounded density, which occurs when $\tau < 1/2$, i.e. $-1 < \alpha < 0$. In fact, the value of τ is tied to the grid of sampling for spatial or temporal models involving the Laplace motion – the finer grid the smaller value of τ which typically leads to an unbounded density.

The EM algorithm can be adopted to the leave-one-out likelihood by not accounting in each loop for the observation that is closest to the evaluated values of the location parameter. This is actually the EM algorithm applied to a penalized log-likelihood where the penalty term is $-\log f(x_{k(\hat{\delta})})$, in which it resembles the method of Chen *et al.* (2008). In these applications, the EM algorithm preserves the fundamental monotonicity property entertained by the original EM method of Dempster *et al.* (1977). In the resulting approximations, the estimate of δ has the same super-efficient asymptotic behavior as demonstrated in this work, while the estimates of μ, σ and τ behave asymptotically in the same way as the MLE under the standard regularity conditions. The formal argument supporting these statements in full generality is left for another occasion. However in the Appendix we do discuss main steps in such an EM approach when applied to the maximizing for the leave-one-out likelihood for the generalized Laplace distributions.

It should be mentioned that the proposed method is useful also in the case when the densities are bounded for all values in the interior of the parameters range but may become unbounded if the parameters reach boundaries of the range. Let us mention two examples when this is of importance. Firstly, for the generalized Laplace distribution, if $\tau \in [1/2, 1)$ and $\sigma > 0$, then the generalized Laplace density is bounded. However, if the parameter value for σ reaches the boundary $\sigma = 0$, then the distribution approaches the gamma distribution with the shape $\tau \in [1/2, 1)$ which constitutes an example of unbounded density. In consequence, using the leave-one-out method allows to avoid ensuing problems. The second case relates to the fact that the generalized Laplace distributions represent a special and the only unbounded density case of the generalized hyperbolic distributions. Here again the leave-one-out method can be applied to deal with

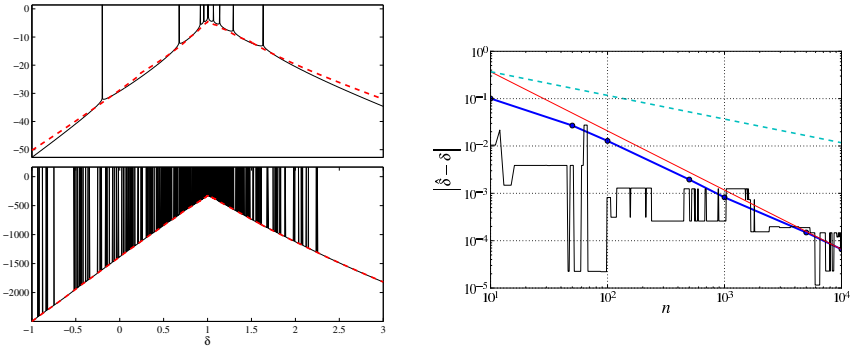


Figure 1: *Left:* The full log-likelihood (solid line) vs. the leave-one-out log-likelihood (dashed line) used for the sample of the size $n = 10$ (*Top*) and $n = 500$ (*Bottom*). In the bottom figure the dashed line cannot be distinguished from the lower envelope of the log-likelihood. *Right:* Asymptotics and super-efficiency of the estimator: the optimal rate – straight thin line, the estimated rate from Monte Carlo simulation – thick line, a trajectory of the absolute estimation error $|\hat{\delta}_n - \delta_0|$ with increasing sample size – thin line. For comparison the rate of MLE under regular assumptions is given by the dashed line.

the unboundedness due to the parameters approaching the values corresponding to a generalized Laplace distribution.

For illustration of the leave-one-out likelihood and the discussed properties of the estimator, we performed a small Monte Carlo (MC) study based on samples generated from an asymmetric generalized Laplace distribution with $(\delta_0, \mu, \sigma, \tau) = (1, -0.5, 1, 0.4)$. In Figure 1 *(Left)*, the full likelihood is compared to the leave-one-out one (dashed line) in a small sample size case ($n = 10$, top) and a large sample size case ($n = 500$, bottom) cases. We can clearly observe the smoothing effect offered by the method.

The asymptotic behavior of the estimator is illustrated in Figure 1 *(Right)*, where, on the logarithmic scale, we see the optimal rate (straight thin line) and the rate for the proposed estimator obtained through MC simulations. The latter is represented here by 90% MC-sample quantiles of $|\hat{\delta}_n - \delta_0|$ computed for 1000 MC samples and for a number of sizes n (thick line). For comparison, a trajectory of $|\hat{\delta}_n - \delta_0|$ evaluated for the subsequently increased n values of a single large

sample is represented by the thin line. Finally, the dashed line on the graph corresponds to the regular rate of convergences $n^{-1/2}$, from which we clearly see a super-efficient rate of the estimator.

3 The maximum leave-one-out likelihood estimator and its supperefficiency

3.1 *Assumptions*

Through the remainder of the paper, let $X_1, \dots, X_n$ be an iid sample from a distribution given by a density $f(x - \delta_0)$ that is differentiable everywhere except for zero. Recall that the Fisher information for a location parameter associated with a density f is defined as $\mathcal{I}_f = \mathbb{E}[((\log f)'(X))^2] = \mathbb{E}[f'^2/f^2(X)]$, where X is a random variable with the distribution defined by f. In our case the Fisher information is not finite due to the assumed unbounded behavior of f around zero so instead we use the incomplete Fisher information defined for $\varepsilon > 0$ as $\mathcal{I}_f(\varepsilon) = \mathbb{E}[f'^2/f^2(X)||X| > \varepsilon]$. We assume that

- A1 $f(x) = p(x)|x|^\alpha,\ \alpha \in (-1, 0)$, p has bounded derivative on $\mathbb{R} \setminus \{0\}$ and, for some $\varepsilon_0 > 0$, is non-zero and continuous either on $[-\varepsilon_0, 0]$ or on $[0, \varepsilon_0]$.
- A2 There exists $b > 0$ such that $f(x) = O(|x|^{-b-1})$ when $|x| \to \infty$.
- A3 For some (and thus for all) $\varepsilon > 0$ the Fisher information $\mathcal{I}_f(\varepsilon)$ is finite.

3.2 *Maximum leave-one-out likelihood estimator*

Here we introduce the estimator and present several convenient representations of the leave-one-out likelihood ratio process.

Let us denote

$$k(\delta) = \operatorname*{argmin}_{k \in \{1,\dots,n\}} |X_k - \delta|,$$

with the convention that if there are two indices we take the one for which corresponding $X_{k(\delta)}$ is on the right hand side of δ. Define the estimator $\hat{\delta} = \hat{\delta}_n$ as the argument that maximizes

$$l(\delta) = l_n(\delta) = \frac{\prod_{i=1}^n f(X_i - \delta)}{f(X_{k(\delta)} - \delta)} \tag{1}$$

Note here that $l(\delta)$ is a cadlag function (the left hand side continuous) and converging to zero at infinity so there is a maximizer (if there are more than one maximizer, we choose, for example, the smallest one). We also observe that $\hat{u}_n = \hat{\delta}_n - \delta_0$ is the maximizer of

$$Z(u) = Z_n(u) = \frac{l(u+\delta_0)}{l(\delta_0)} = \frac{f(X_{k(\delta_0)} - u - \delta_0)}{f(X_{k(u+\delta_0)} - \delta_0)} \prod_{i \neq k(\delta_0), i \neq k(u+\delta_0)} \frac{f(X_i - u - \delta_0)}{f(X_i - \delta_0)}.$$

By introducing the event $C_{i,\delta} = \{k(\delta) \neq i\}$ and its indicator function $I_{C_{i,\delta}}$, we obtain the following convenient representations of the above functions

$$l(\delta) = \prod_{i=1}^{n} f(X_i - \delta)^{I_{C_{i,\delta}}} = \sum_{k=1}^{n} I_{C^c_{k,\delta}} \prod_{i=1, i\neq k}^{n} f(X_i - \delta), \tag{2}$$

and

$$Z(u) = \prod_{i=1}^{n} f(X_i - \delta_0)^{-I_{C_{i,\delta_0}}} \prod_{i=1}^{n} f(X_i - u - \delta_0)^{I_{C_{i,u+\delta_0}}} = \Big(\sum_{l=1}^{n} I_{C^c_{l,\delta_0}} \prod_{i=1, i\neq l}^{n} f(X_i - \delta_0)^{-1} \Big) \cdot \Big(\sum_{k=1}^{n} I_{C^c_{k,\delta_0+u}} \prod_{j=1, j\neq k}^{n} f(X_j - u - \delta_0) \Big). \tag{3}$$

3.3 *The main result*

The purpose of this paper is to establish consistency of $\hat{\delta}_n$ which is done together with getting a super-efficient rate of convergence in the following result.

Theorem 3.1. Let f satisfy the above assumptions and let $\hat{\delta}_n$ be the maximizer of l_n given by (1). Then $\hat{\delta}_n$ is a consistent estimator of δ_0 and for any $\beta < 1/(1+\alpha)$:

$$\lim_{n\to\infty} n^{\beta}(\hat{\delta}_n - \delta_0) \stackrel{p}{=} 0. \tag{4}$$

4 Lemmas and the proof of the theorem

Additionally to the notation and assumptions of the previous section, we also use what follows. For $\lambda > 0$ and $L > 0$:

$$A_\lambda = A_{n,\lambda} = \{\min_{\substack{i,j=1,\dots,n \\ i\neq j}} |X_i - X_j| > \lambda\}, \tag{5}$$

$$B_L = B_{n,L} = \{\max_{i=1,\dots,n} |X_i - \delta_0| < L\}. \tag{6}$$

In our argument the variable L is eventually increasing without bound so whenever the symbol $O(L^\rho)$ is used for some ρ, it means that $\limsup_{L\to\infty} |O(L^\rho)|/L^\rho < \infty$. Finally, for compactness of our formulations, we define $S_r(u_0) = [u_0 - r, u_0 + r]$.

We start with a result about the rate of convergence of the minimal distance between X_i's.

Lemma 4.1. *Assume that a sequence of positive numbers λ_n has the following asymptotics for a certain $c > 0$:*

$$\lambda_n = O\left(n^{-1-\frac{1}{\alpha+1}-c}\right)$$

Then for $A_n = A_{n,\lambda_n}$ defined through (5) we have

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 1.$$

Proof. Since $\lambda_n \leq D \cdot n^{-1-\frac{1}{\alpha+1}-c}$ for some $D > 0$, it is enough to show the result for $\lambda_n = D \cdot n^{-1-\frac{1}{\alpha+1}-c}$. Define

$$C_n = \{X_{n+1} \in \bigcup_{i=1}^{n} [X_i - \lambda_n, X_i + \lambda_n]\}.$$

We first demonstrate that for a proof it is sufficient to show that $C = \limsup_{n\to\infty} C_n$ is of probability zero, which is equivalent to saying that with probability one the number of times that an observation X_{n+1} is inside of $\cup_{i=1}^{n}[X_i - \lambda_n, X_i + \lambda_n]$ is finite.

To see this consider an outcome ω from C^c. Then there exists n_0 such that for $n > n_0$:

$$|X_{n+1}(\omega) - X_i(\omega)| > \lambda_n,\ i = 1, \dots, n.$$

For such n_0, let

$$\varepsilon_0 = \min_{\substack{i,j=1,\dots,n_0 \\ i\neq j}} |X_i(\omega) - X_j(\omega)|$$

while n_1 be such that for $n > n_1 > n_0$ we have $\lambda_n < \varepsilon_0$. Take $n > n_1$ and note that the minimum of $|X_i(\omega) - X_j(\omega)|$ over all pairs (i,j) such that $i,j = 1,\dots,n, i \neq j$ is obtained as the minimum of the numbers standing on the left hand side of the following inequalities

$$\begin{aligned}
\min_{\substack{i,j=1,\dots,n_0 \\ i\neq j}} |X_i(\omega) - X_j(\omega)| &> \lambda_n, \\
\min_{i=1,\dots,n_0} |X_i(\omega) - X_{n_0+1}(\omega)| &> \lambda_{n_0} \geq \lambda_n, \\
\min_{i=1,\dots,n_0+1} |X_i(\omega) - X_{n_0+2}(\omega)| &> \lambda_{n_0+1} \geq \lambda_n, \\
&\vdots \\
\min_{i=1,\dots,n-1} |X_i(\omega) - X_n(\omega)| &> \lambda_{n-1} \geq \lambda_n.
\end{aligned}$$

Consequently the outcome ω has to be in A_n for each $n > n_1$, which proves that

$$C^c \subset \liminf_{n\to\infty} A_n.$$

Thus if A denotes the right hand side event in the above and $\mathbb{P}(C^c) = 1$, then

$$\begin{aligned}
1 =& \mathbb{P}(C^c) \leq \mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k\geq n} A_k\Big) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \leq \\
& \limsup_{n\to\infty} \mathbb{P}(A_n) \leq 1
\end{aligned}$$

and consequently it is indeed enough to show that $\mathbb{P}(C) = 0$.

To prove the latter, by the Borel-Canteli lemma, it is enough to show that $\mathbb{P}(C_n)$'s form a convergent series. To this end notice that by Assumption A1, the density of X_i is bounded except at δ_0. Hence there exists sufficiently small $u > 0$ and an interval neighborhood I of zero and of the diameter not exceeding u such that $f(x) = p(x)|x|^\alpha$ for $x \in I$ is larger than the value $f(y)$ for any $y \notin I$. Thus if

a subset $D \subset \mathbb{R}$ has measure at most u, then

$$\begin{aligned}\mathbb{P}(X_{n+1} \in D) &= \int_D p(x-\delta_0)|x-\delta_0|^\alpha \, dx \\ &\leq \int_I p(x)|x|^\alpha \, dx \leq \mathbb{P}(X \in [-u+\delta_0, u+\delta_0]).\end{aligned}$$

Using this fact, the convergence of $n\lambda_n$ to zero, and independence of X_{n+1} from $\mathbf{X}_n = (X_1, \ldots, X_n)$, we obtain for sufficiently large n:

$$\begin{aligned}\mathbb{P}(C_n) =& \mathbb{P}\big(X_{n+1} \in \bigcup_{i=1}^n [X_i - \lambda_n, X_i + \lambda_n]\big) = \\ & \mathbb{E}\big(\mathbb{P}\big(X_{n+1} \in \bigcup_{i=1}^n [X_i - \lambda_n, X_i + \lambda_n] | \mathbf{X}_n\big)\big) \leq \\ & \mathbb{P}(X \in [-n\lambda_n + \delta_0, n\lambda_n + \delta_0]).\end{aligned}$$

Note that there exists $K > 0$ such that for sufficiently small u we have $\mathbb{P}(X \in [-u+\delta_0, u+\delta_0]) \leq Ku^{\alpha+1}$, so for sufficiently large n:

$$\mathbb{P}(C_n) \leq K(n\lambda_n)^{\alpha+1} \leq K\big(n^{-1/(\alpha+1)-c}\big)^{\alpha+1} = Kn^{-1-c(\alpha+1)}$$

and thus convergence of the series holds. □

The next lemma is a quite obvious consequence of Assumption A2.

Lemma 4.2. *If n/L_n^b converges to zero, then for $B_n = B_{n,L_n}$ given in (6):*

$$\lim_{n\to\infty} \mathbb{P}(B_n) = 1.$$

Proof. By A2, the following inequality holds for some $K > 0$ and sufficiently large L:

$$\mathbb{P}\left(|X - \delta_0| \leq L\right) \leq 1 - KL^{-b}$$

and the result follows immediately from

$$\mathbb{P}(B_n) = \mathbb{P}\{\max_{i=1,\ldots,n} |X_i - \delta_0| < L_n\} \leq \left(1 - KL_n^{-b}\right)^n,$$

which holds for sufficiently large n. □

In the proof of the next result we use Assumption A3, i.e. the finiteness of the partial Fisher information. Let us introduce the following function that is also used in the proof of Lemma 4.4:

$$v(x) = \frac{p'(x)|x|}{2p^{1/2}(x)} + \frac{\alpha}{2}\text{sign}(x)p^{1/2}(x), \tag{7}$$

and note that it is bounded in neighborhood of zero. Moreover for $x \neq 0$:

$$(f^{1/2})'(x) = \frac{f'(x)}{2f^{1/2}(x)} = |x|^{\alpha/2-1}v(x). \tag{8}$$

Lemma 4.3. *There exists $K > 0$ such that for each $x_0 \in \mathbb{R}$, $c < 1$ and $r \in (0, \frac{c}{2})$:*

$$\int_{[-c,c]^c} \sup_{|h|<r} |f^{1/2}(x) - f^{1/2}(x-h)| \cdot f^{1/2}(x+x_0)\, dx \leq Kr^{(\alpha+1)/2}. \tag{9}$$

Proof. First by the Schwartz inequality

$$\begin{aligned}
&\int_{[-c,c]^c} \sup_{|h|<r} |f^{1/2}(x) - f^{1/2}(x-h)| f^{1/2}(x+x_0)\, dx \\
&\qquad \leq \left(\int_{[-c,c]^c} \sup_{|h|<r} (f^{1/2}(x) - f^{1/2}(x-h))^2\, dx \right)^{1/2} \\
&\qquad = \frac{1}{2} \left(\int_{[-c,c]^c} \sup_{|h|<r} \left(\int_0^h \frac{f'}{f^{1/2}}(x-y)\, dy \right)^2 dx \right)^{1/2}.
\end{aligned}$$

By the Jensen inequality and then by the Fubini theorem

$$\begin{aligned}
&\int_{[-c,c]^c} \sup_{|h|<r} \left(\int_0^h \frac{f'}{f^{1/2}}(x-y)\, dy \right)^2 dx \\
&\qquad \leq \int_{[-c,c]^c} \sup_{|h|<r} \left(h \int_0^h \frac{f'^2}{f}(x-y)\, dy \right) dx \\
&\qquad = r \int_0^r 4 \int_{[-c,c]^c} |x-y|^{\alpha-2} v^2(x-y)\, dx\, dy \\
&\qquad = r \int_0^r 4 \int_{[-c+y,y+c]^c} |s|^{\alpha-2} v^2(s)\, ds\, dy.
\end{aligned}$$

Note that for $y \in [0, r]$ we have $-c + y < -c + r < -r$ and $y + c > c > r$. Combining this with the boundedness of v in a neighborhood of zero, we obtain that for some K_0 and for each $\varepsilon > 0$:

$$\begin{aligned}\int_{[-c+y,y+c]^c} 2|s|^{\alpha-2}v^2(s)\,ds &\le \int_{[-r,r]^c} 2|s|^{\alpha-2}v^2(s)\,ds \\ &= \int_r^{\varepsilon} 2|s|^{\alpha-2}v^2(s)\,ds + \int_{-\varepsilon}^{-r} 2|s|^{\alpha-2}v^2(s)\,ds + \frac{1}{2}\mathcal{I}_f(\varepsilon) \\ &\le K_0\left|\int_r^{\varepsilon} s^{\alpha-2}\,ds\right| + \frac{1}{2}\mathcal{I}_f(\varepsilon) \\ &\le \frac{K_0}{2-\alpha}|r^{\alpha-1} - \varepsilon^{\alpha-1}| + \frac{1}{2}\mathcal{I}_f(\varepsilon) \\ &\le K^2 r^{\alpha-1},\end{aligned}$$

where K is some positive constant independent of r and c. From these inequalities we obtain

$$\int_{[-c,c]^c} \sup_{|h|<r} |f^{1/2}(x) - f^{1/2}(x-h)|f^{1/2}(x+x_0)\,dx \le Kr^{(\alpha+1)/2},$$

which concludes the proof. □

The following result stands behind a super-efficient rate of convergence that is eventually obtained in the proof of the main theorem.

Lemma 4.4. *There exist $B > 0$ and $K > 0$ such that for each $s \in \mathbb{R}$:*

$$\mathbb{E}\left[\frac{f^{1/2}(X-s)}{f^{1/2}(X)}\right] \le 1 - K\min(|s|^{\alpha+1}, B). \tag{10}$$

Proof. Let us set $r(x,s) = (f^{1/2}(x+s) - f^{1/2}(x))^2$ and note

$$\begin{aligned}\mathbb{E}\left[\frac{f^{1/2}(X-s)}{f^{1/2}(X)}\right] &= \frac{1}{2}\left(\int f(x)dx + \int f(x-s)dx - \int r(x,s)dx\right) \\ &= 1 - \frac{1}{2}\int r(x,s)\,dx.\end{aligned}$$

Note that $r(s) = \int r(x,s)\,dx$ is a continuous non-negative function taking value 2 at infinity, zero at $s = 0$, which is also its unique global minimum. Consequently, it is enough to show that $r(s)$ is $O(s^{\alpha+1})$.

Consider be a one-sided neighborhood of zero, say $[0, \varepsilon_0]$, where v being negative is separated from zero by, say, $-L$, $L > 0$. Then for positive s and x such that $x + s \in [0, \varepsilon_0]$ we have

$$\begin{aligned} r(x,s) &= \left(\int_0^s (f^{1/2})'(t+x)\,dt\right)^2 = \left(\int_0^s (x+t)^{\alpha/2-1} v(x+t)\,dt\right)^2 \\ &\geq L^2 \left(\int_0^s (x+t)^{\alpha/2-1}\,dt\right)^2 \\ &= \frac{4L^2}{\alpha^2} s^\alpha \left(\left(\frac{x}{s}+1\right)^{\alpha/2} - \left(\frac{x}{s}\right)^{\alpha/2}\right)^2. \end{aligned}$$

Using this we get for positive $s < \varepsilon_0/2$:

$$\begin{aligned} \int r(x,s)\,dx &\geq \int_0^{\varepsilon_0/2} r(x,s)\,dx \\ &\geq \frac{4L^2}{\alpha^2} s^{\alpha+1} \int_0^{\varepsilon_0/(2s)} ((y+1)^{\alpha/2} - y^{\alpha/2})^2\,dy \\ &\geq \frac{4L^2}{\alpha^2} \int_0^1 ((y+1)^{\alpha/2} - y^{\alpha/2})^2\,dy \cdot s^{\alpha+1}. \end{aligned}$$

The argument for negative s follows the same way. □

The preceding result is explicitly used only in the following lemma, which plays a central role in our proof of the main result.

Lemma 4.5. *There exist positive constants K_1, K_2 such that for all $n \in \mathbb{N}$, γ and λ both in $(0,1)$, if $r \in (0, \lambda/6)$ and $|u_0| > \gamma$, then*

$$\mathbb{E}\left[I_{A_\lambda \cap B_L} \sup_{u \in S_r(u_0)} Z^{1/2}(u)\right] \leq O(L^a) r^{\frac{\alpha}{2}} n^2 (1 - K_1 \gamma^{1+\alpha} + K_2 r^{\frac{1+\alpha}{2}})^{n-2}, \tag{11}$$

where $a = \max(0, (1-b)/2)$.

Proof. We note that the left hand side does not depend on δ_0 so let us assume that $\delta_0 = 0$. Let us take arbitrary values λ, r, γ and u_0 that satisfy the required conditions (K_1, K_2 will come later). By (3)

$$\sup_{u\in S_r(u_0)} Z^{1/2}(u) \leq \tag{12}$$
$$\sum_{l=1}^{n} I_{C^c_{l,0}} \prod_{\substack{i=1\\ i\neq l}}^{n} f^{-1/2}(X_i) \cdot \sum_{k=1}^{n} \sup_{u\in S_r(u_0)} I_{C^c_{k,u}} \prod_{\substack{j=1\\ j\neq k}}^{n} f^{1/2}(X_j - u).$$

Let us note that

$$C^c_{k,u} = \left(\bigcup_{i\neq k}^{n} C^c_{i,u}\right)^c = \bigcap_{i\neq k}^{n} C_{i,u}.$$

Moreover, since in A_λ all observations are at least λ apart and in $C_{i,u}$ the value X_i is not the closest to u the distance between X_i and u must be at least $\lambda/2$ which gives

$$\{|X_i - u| \geq \lambda/2\} \supseteq A_\lambda \cap C_{i,u}.$$

For $u \in S_r(u_0)$, by the triangle inequality

$$C_{i,u_0,r} \stackrel{def}{=} \{|X_i - u_0| \geq \lambda/2 - r\} \supseteq \{|X_i - u| \geq \lambda/2\}.$$

Thus for each $k = 1, \ldots, n$:

$$I_{A_\lambda} \sup_{u\in S_r(u_0)} I_{C^c_{k,u}} \prod_{\substack{i=1,\\ i\neq k}}^{n} f^{1/2}(X_i - u) \leq \sup_{u\in S_r(u_0)} \prod_{\substack{i=1,\\ i\neq k}}^{n} f^{1/2}(X_i - u) I_{C_{i,u_0,r}} \tag{13}$$

and for each $l = 1, \ldots, n$ we have

$$I_{A_\lambda} I_{C^c_{k,0}} \prod_{\substack{i=1,\\ i\neq k}}^{n} f^{-1/2}(X_i) \leq \prod_{\substack{i=1,\\ i\neq k}}^{n} \frac{I_{|X_i|>\lambda/2}}{f^{1/2}(X_i)}. \tag{14}$$

Combining (12), (13), and (14) we obtain

$$I_{A_\lambda} \sup_{u\in S_r(u_0)} Z^{1/2}(u) \leq \sum_{k,l=1}^{n} \sup_{u\in S_r(u_0)} \prod_{\substack{i=1,\\ i\neq k}}^{n} f^{1/2}(X_i - u) I_{C_{i,u_0,r}} \prod_{\substack{j=1,\\ j\neq l}}^{n} \frac{I_{|X_j|>\lambda/2}}{f^{1/2}(X_j)}.$$

For $i = 1, \dots, n$ let us define

$$\tilde{Y}_i = \frac{I_{L>|X_i|>\lambda/2}}{f^{1/2}(X_i)},$$
$$\bar{Y}_i(u) = f^{1/2}(X_i - u) I_{C_{i,u_0},r}.$$

Then we obtain

$$I_{A_\lambda \cap B_L} \sup_{u \in S_r(u_0)} Z^{1/2}(u) \le \sum_{k,l=1}^{n} \tilde{Y}_k \sup_{u \in S_r(u_0)} \bar{Y}_l(u) \prod_{\substack{i=1,\\ i\neq k,\\ i\neq l}}^{n} \bar{Y}_i(u) \tilde{Y}_i.$$

As a result and by independence, we obtain

$$\begin{aligned}
\int_{A_\lambda \cap B_L} \sup_{u \in S_r(u_0)} Z^{1/2}(u)\, d\mathbb{P} &\le \sum_{k,l=1}^{n} \mathbb{E}[\tilde{Y}_k] \cdot \mathbb{E}\Big[\sup_{u \in S_r(u_0)} \bar{Y}_l(u) \prod_{\substack{i=1,\\ i\neq k,\\ i\neq l}}^{n} \bar{Y}_i(u) \tilde{Y}_i \Big] \\
&= n^2\, \mathbb{E}[\tilde{Y}_1] \cdot \mathbb{E}\Big[\sup_{u \in S_r(u_0)} \bar{Y}_1(u) \prod_{i=3}^{n} \bar{Y}_i(u) \tilde{Y}_i \Big] \\
&\le n^2\, \mathbb{E}[\tilde{Y}_1] \cdot \mathbb{E}\Big[\sup_{u \in S_r(u_0)} \bar{Y}_1(u) \Big] \cdot \mathbb{E}\Big[\sup_{u \in S_r(u_0)} \bar{Y}_1(u) \tilde{Y}_1 \Big]^{n-2}. \qquad (15)
\end{aligned}$$

In what follows, we bound each of the three expectations on the right hand side of the above inequality.

First, by Assumption A2, $\mathbb{E}[\tilde{Y}_1] \le \int_{-L}^{L} f^{1/2}(x)\, dx = O(L^a)$, where $a = \max(0, (1-b)/2)$. To deal with the second expectation, notice that by Assumption A1 on $f(x)$ there is a constant $K_0 > 0$ such that $f(x) \le K_0 \min(|x^\alpha|, 1) \le K_0(\lambda/2 - 2r)^\alpha$, since $0 < \lambda/2 - 2r < 1$. Therefore, if $|u - u_0| \le r$ and $|x - u_0| \ge \lambda/2$, then $|x - u| \ge \lambda/2 - 2r$ and thus

$$\sup_{u \in S_r(u_0)} \bar{Y}_1(u) \le K_0(\lambda/2 - 2r)^{\alpha/2} \le K_0 r^{\alpha/2},$$

where the last inequality holds since $\lambda > 6r$.

The final expectation requires a few more steps. First, using the triangle inequality yields

$$\tilde{Y}_1 \cdot \sup_{u \in S_r(u_0)} \bar{Y}_1(u) \le \tilde{Y}_1 \cdot \left(\bar{Y}_1(u_0) + \sup_{|h|<r} |\bar{Y}_1(u_0 + h) - \bar{Y}_1(u_0)| \right).$$

Then from Lemma 4.3 there exists K_2 such that

$$\mathbb{E}\Big[\tilde{Y}_1 \cdot \sup_{|h|<r} |\bar{Y}_1(u_0+h) - \bar{Y}_1(u_0)|\Big]$$
$$\leq \int_{[-\lambda/2+r,\lambda/2-r]^c} \sup_{|h|<r} |f^{1/2}(s-h) - f^{1/2}(s)| \cdot f^{1/2}(s+u_0)\, ds$$
$$\leq K_2 r^{(1+\alpha)/2}$$

and from Lemma 4.4:

$$\mathbb{E}[\tilde{Y}_1 \cdot \bar{Y}_1(u_0)] \leq 1 - K_1 \min(\gamma^{1+\alpha}, b).$$

Putting all the three bounds together in (15) completes the proof. □

Chebyshev's inequality combined with the inequality $1 + a \leq e^a$ yields the following corollary to the above lemma.

Corollary 4.1. There exist positive constants K_1 and K_2 such that for all $n \in \mathbb{N}$, γ and λ both in $(0,1)$, if $r \in (0, \lambda/6)$ and $|u| > \gamma$, then

$$\mathbb{P}(I_{A_\lambda \cap B_L} \sup_{u \in S_r(u)} Z(u) \geq 1) \leq O(L^a) r^{\frac{\alpha}{2}} n^2 e^{-(n-2)\left(K_1 \gamma^{\alpha+1} - K_2 r^{(1+\alpha)/2}\right)},$$

where $a = \max(0, (1-b)/2)$.

Lemma 4.5 will enter the proof of the main theorem through the following result, which is a consequence of the above corollary.

Lemma 4.6. *Let $\hat{\delta}_L$ be the maximizer of $l(\delta)$ over $[-L + \delta_0, L + \delta_0]$. There exist positive constants K_1 and K_2 such that for all $n \in \mathbb{N}$, γ and λ both in $(0,1)$, if $r \in (0, \lambda/6)$, then*

$$\mathbb{P}(A_\lambda \cap B_L \cap \{|\hat{\delta}_L - \delta_0| > \gamma\}) \leq \tag{16}$$
$$O(L^{a+1}) r^{\frac{\alpha}{2}-1} n^2 e^{-(n-2)\left(K_1 \gamma^{\alpha+1} - K_2 r^{(1+\alpha)/2}\right)},$$

where $a = \max(0, (1-b)/2)$.

Proof. From the definition of $\hat{\delta}_L$, $\hat{u}_L = \hat{\delta}_L - \delta_0$ maximizes $Z(u)$ over $[-L, L]$ and thus $Z(\hat{u}_L) \geq Z(0) = 1$. Consequently, if $|\hat{u}_L| > \gamma$, then

$$\sup_{u \in [-\gamma,\gamma]^c \cap [-L,L]} Z(u) \geq 1.$$

This leads to

$$\mathbb{P}(A_\lambda \cap B_L \cap \{|\hat{\delta}_L - \delta_0| > \gamma\}) \leq \mathbb{P}(I_{A_\lambda \cap B_L} \sup_{u \in [-\gamma,\gamma]^c \cap [-L,L]} Z(u) \geq 1).$$

Let $S_r(u_k)$, $k = 1, \ldots, 2[L/r] + 1$ be a cover of $[-\gamma, \gamma]^c \cap [-L, L]$, such that $|u_k| > \gamma$. By Corollary 4.1:

$$\begin{aligned}
&\mathbb{P}(A_\lambda \cap B_L \cap \{ \sup_{u \in [-\gamma,\gamma]^c \cap [-L,L]} Z(u) \geq 1\}) \\
&\quad = \mathbb{P}(\bigcup_{k=1}^{2[L/r]+1} \{I_{A_\lambda \cap B_L} \sup_{u \in [-\gamma,\gamma]^c \cap S_r(u_k)} Z(u) \geq 1\}) \\
&\quad \leq \sum_{k=1}^{2[L/r]+1} \mathbb{P}(I_{A_\lambda \cap B_L} \sup_{u \in S_r(u_k)} Z(u) \geq 1) \\
&\quad \leq O(L^a) r^{\frac{\alpha}{2}-1} n^2 e^{-(n-2)\left(K_1 \gamma^{\alpha+1} - K_2 r^{(1+\alpha)/2}\right)}.
\end{aligned}$$

□

5 Proof of Theorem 3.1

Here we present our proof of the main theorem.

Proof. Set $\beta < 1/(1+\alpha)$. Let $L_n = n^{s\frac{2}{1+b}}$, with s being a positive constant that will be set later but at the moment we require only that $L_n > n^{3/b}$. Further, let λ_n be set so that Lemma 4.1 is satisfied.

Because of Lemmas 4.1 and 4.2, the events A_{n,λ_n} and $B_n = B_{n,n^{3/b}}$ that are defined through (5) and (6), respectively, have probabilities converging to one. Consequently, it is sufficient to show that for each $\gamma > 0$:

$$\lim_{n\to\infty} \mathbb{P}(A_{n,\lambda_n} \cap B_n \cap \{n^\beta |\hat{\delta}_n - \delta_0| > \gamma\}) = 0.$$

Let $\gamma_n = \gamma n^{-\beta}$ and note that since $B_n \subseteq B_{n,L_n}$:

$$\begin{aligned}
&\mathbb{P}(A_{n,\lambda_n} \cap B_n \cap \{|\hat{\delta}_n - \delta_0| > \gamma_n\}) \leq \\
&\leq \mathbb{P}(A_{n,\lambda_n} \cap B_{n,L_n} \cap \{\gamma_n < |\hat{\delta}_n - \delta_0| \leq L_n\}) + \mathbb{P}(B_n \cap \{|\hat{\delta}_n - \delta_0| > L_n\}).
\end{aligned} \tag{17}$$

Let us consider the first term on the right hand side and take a sequence r_n so that $r_n \leq \lambda_n/6$. Then, by Lemma 4.6, for $a = \max(0, (1-b)/2)$:

$$\limsup_{n\to\infty} \mathbb{P}(A_{n,\lambda_n} \cap B_{n,L_n} \cap \{\gamma_n < |\hat{\delta}_n - \delta_0| \leq L_n\}) \\ \leq \limsup_{n\to\infty} O(L_n^{\beta+1}) n^2 r_n^{\frac{\alpha}{2}-1} e^{-(n-2)\left(K_1 \gamma_n^{1+\alpha} - K_2 r_n^{(1+\alpha)/2}\right)}.$$

By choosing r_n so that $n r_n^{(1+\alpha)/2} \leq n^{-d}$ for some $d > 0$, we have for suitably chosen $h > 0$, ε, and $K > 0$:

$$\limsup_{n\to\infty} \mathbb{P}(A_{n,\lambda_n} \cap B_{n,L_n} \cap \{\gamma_n < |\hat{\delta}_n - \delta_0| \leq L_n\}) \leq \lim_{n\to\infty} n^h e^{-n^{\varepsilon} + K n^{-d}} \\ = 0.$$

The second term on the right hand side of (17) also converges to zero as shown next. Since $\{|\hat{\delta}_n - \delta_0| > L_n\} \subseteq \{\sup_{|u|>L_n} Z(u) \geq 1\}$ and by a direct application of Chebyshev's inequality it is enough to show

$$\lim_{n\to\infty} \mathbb{E}\left[I_{B_n} \sup_{|u|>L_n} Z^{1/2}(u) \right] = 0. \tag{18}$$

To this end note that on B_n, $|X_i| \leq n^{3/b} + |\delta_0|$, thus for sufficiently large n, for $|u| > L_n$ and on B_n:

$$|X_i - u| \geq |u| - |X_i| \geq L_n - n^{3/b} - |\delta_0| = O(L_n).$$

From this, Assumption A2, and by the choice of L_n:

$$I_{B_n} f^{1/2}(X_i - u) \leq I_{B_n} K |X_i - u|^{-(b+1)/2} \leq O(L_n^{-(b+1)/2}) = O(n^{-s}).$$

In consequence,

$$I_{B_n} \sup_{|u|>L_n} Z^{1/2}(u) \leq O^{n-1}(n^{-s}) I_{B_n} \prod_{i \neq k(\delta_0)} f^{-1/2}(X_i - \delta_0). \tag{19}$$

Using the representation (2), we have

$$\prod_{i \neq k(\delta_0)} f^{-1/2}(X_i - \delta_0) = \sum_{k=1}^{n} I_{C^c_{k,\delta_0}} \prod_{i=1, i\neq k}^{n} f^{-1/2}(X_i - \delta_0) \\ \leq \sum_{k=1}^{n} \prod_{i=1, i\neq k}^{n} f^{-1/2}(X_i - \delta_0).$$

By Assumption A2, we also have

$$\mathbb{E}(I_{|X-\delta_0|<L} f^{-1/2}(X-\delta_0)) = O(L^c),$$

where $c = (1-b)^+/2$, which along with the mutual independence of X_i's yields

$$\begin{aligned}\mathbb{E}\Big[I_{B_n} \prod_{i\neq k(\delta_0)} f^{-1/2}(X_i-\delta_0)\Big] &\leq n\left(\mathbb{E}\left[I_{|X-\delta_0|\leq n^{3/b}} f^{-1/2}(X-\delta_0)\right]\right)^{n-1} \\ &\leq nO^{n-1}(n^{3c/b}).\end{aligned}$$

Putting this together with (19), for sufficiently large n we obtain

$$\mathbb{E}\Big[I_{B_n} \sup_{|u|>L_n} Z^{1/2}(u)\Big] \leq nO^{n-1}(n^{3c/b-s}),$$

where s as of now was not set yet. Thus by taking $s > 3c/b+1$ we make the right hand side converging to zero, which concludes the proof. □

6 Concluding remarks

We have demonstrated that the maximum *leave-one-out likelihood* estimator is consistent and has a superefficient rate of convergence. The rate of convergence does not differ by a power factor from $n^{-1/(1+\alpha)}$ which would be the optimal rate of convergence. In fact, the proof of the main theorem yields a bit stronger result stating that the lower bound on the rate of convergence differs from the optimal rate only by a certain power-of-logarithm factor. However, the presented proof does not yield the optimal rate and an open question is if this rate is actually reached by the estimator. In fact, this rate would be optimal for the minimal variance estimation of the location as discussed in Polfeldt (1970b) and Polfeldt (1970a), where an estimator achieving this rate is constructed. This optimal rate is also obtained in Ibragimov & Khasminskii (1981b) for the Pitman estimators.

It is worth stressing again that the leave-one-out estimator unlike the other estimators has the advantage that it can be easily implemented through the MLE approach in a general multi-parameter setup, for example, when scale or/and shape parameters are present. In the appendix it was demonstrated how the EM algorithm applies when other than the location parameters are present in which case maximizing likelihood is the most natural way to proceed.

A Appendix

Here we present a formalized approach to the maximizing the leave-one-out likelihood by means of the EM algorithm and in the presence of other than location parameter. We focus on an example of a symmetric Laplace distribution while a more complete presentation is left for some future research. Namely, we consider estimation of a vector of parameters $\vartheta = (\delta, \sigma)$ of a symmetric ($\mu = 0$) generalized Laplace distribution with some known shape parameter $\tau < 0.5$. See Section 2 for the definitions and the notation. In our setup, the observed values are $Y_i = \sigma\sqrt{\Gamma_i}Z_i - \delta$, $i = 1, \ldots, n$ and the complete set of variables is $\mathbf{X} = (\Gamma_1, \ldots, \Gamma_n, Y_1, \ldots, Y_n)$. As mentioned before, the density $f_\vartheta(y)$ of Y_i's is having the form $p_\sigma(y-\delta)|y-\delta|^{2\tau-1}$ for some bounded and non-vanishing around zero function p_ϑ.

To precisely formulate our algorithm we need some additional notation and defintions. For a vector $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the permutation of its elements which leads to the order statistics of $(|y_1 - \delta|, \ldots, |y_n - \delta|)$ is denoted by $\pi^\vartheta(\mathbf{y})$, and using this function we define $\mathbf{R}^\vartheta = \pi_\vartheta(\mathbb{R}^n)$. Slightly abusing notation we also write $\pi^\vartheta(\mathbf{x})$ for $(\gamma_1, \ldots, \gamma_n, \pi^\vartheta(\mathbf{y}))$. Further, for $\mathbf{Y}^\vartheta = \pi_\vartheta(\mathbf{Y})$, we consider the conditional distributions of: the vector $\tilde{\mathbf{Y}}^\vartheta = (Y_1^\vartheta, \ldots, Y_{n-1}^\vartheta)$ given $Y_0^\vartheta = y_0 \in \mathbb{R}$, denoted by $g_\vartheta(\tilde{\mathbf{y}}|y_0)$ and defined on $\mathbf{R}_{y_0}^\vartheta = \{\tilde{\mathbf{y}} : (y_0, \tilde{\mathbf{y}}) \in \mathbf{R}^\vartheta\}$; the vector $\mathbf{X}^\vartheta$ conditionally on $\mathbf{Y}^\vartheta = \mathbf{y} \in \mathbf{R}^\vartheta$, denoted by $k_\vartheta(\mathbf{x}|\mathbf{y})$ for $\mathbf{x} \in \mathbb{R}_+^n \times \{\mathbf{y}\}$; and, finally, the distribution of $\mathbf{X}^\vartheta$ given $Y_0^\vartheta = y_0$ for $\mathbf{x} \in \mathbb{R}_+^n \times \{y_0\} \times \mathbf{R}_{y_0}^\vartheta$ and denoted by $h_\vartheta(\mathbf{x}|y_0)$.

We note the relation

$$g_\vartheta(y_1, \ldots, y_{n-1}|y_0) = \frac{(n-1)!}{F(\vartheta, y_0)^{n-1}} f_\vartheta(y_1) \cdots f_\vartheta(y_{n-1}),$$

where $F(\vartheta, y_0) = 1 - \int_{-|y_0-\delta|}^{|y_0-\delta|} f_\vartheta(s)\, ds$. Thus if one wants to treat the leave-one-out likelihood as an actual likelihood it has to be normalized and then it can be viewed as equivalent to $g_\vartheta(y_1, \ldots, y_{n-1}|y_0)$. From now on we consider the maximization of $L^{\mathbf{y}}(\vartheta) = g_\vartheta(y_1, \ldots, y_{n-1}|y_0)$.

We follow a general scheme of the EM algorithm, see for example Wu (1983) and report the following two fundamental facts that hold for any fixed value of

incomplete observations $\mathbf{y} \in \mathbf{R}^{\vartheta'}$:

$$\begin{aligned} L^{\mathbf{y}}(\vartheta) = &\int_{\mathbb{R}_+^n \times \{\mathbf{y}\}} \log h_\vartheta \left(\pi^\vartheta(\mathbf{x}) | y_0^\vartheta \right) k_{\vartheta'}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} + \\ &- \int_{\mathbb{R}_+^n \times \{\mathbf{y}\}} \log k_\vartheta(\pi^\vartheta(\mathbf{x})|\pi^\vartheta(\mathbf{y})) \cdot k_{\vartheta'}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \end{aligned} \tag{20}$$

and

$$\begin{aligned} &\int_{\mathbb{R}_+^n \times \{\mathbf{y}\}} \log k_\vartheta(\pi^\vartheta(\mathbf{x})|\pi^\vartheta(\mathbf{y})) \cdot k_{\vartheta'}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \\ &\qquad \leq \int_{\mathbb{R}_+^n \times \{\mathbf{y}\}} \log k_{\vartheta'}(\pi^{\vartheta'}(\mathbf{x})|\pi^{\vartheta'}(\mathbf{y})) \cdot k_{\vartheta'}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \end{aligned}$$

These two conditions guarantee the monotonicity of $L^{\mathbf{y}}(\hat{\vartheta}_n)$ in n of the algorithm in which the updates $\hat{\vartheta}_n$ are based on the maximizing the first term of the right hand side of (20), which we denote as $Q^{\mathbf{y}}(\vartheta|\vartheta')$.

Let us now discuss how this maximization avoids being trapped in local maxima that are due to the unboundedness of the likelihood. In this discussion, we consider the case of a symmetric generalized Laplace distribution given by f_ϑ. Let $s(\gamma)$ be the density of gamma distribution with the shape parameter $\tau < 0.5$ and the scale equal to one and define

$$\begin{aligned} M(y, y'; \vartheta|\vartheta') &= \frac{\int_0^\infty \left(\log \frac{s(\gamma)}{\sqrt{2\pi\sigma^2\gamma}} - \frac{(y-\delta)^2}{2\sigma^2\gamma} \right) \frac{s(\gamma)}{\sqrt{2\pi\sigma'^2\gamma}} e^{-\frac{(y'-\delta')^2}{2\sigma'^2\gamma}}\, d\gamma}{f_{\vartheta'}\left(y'\right)} \\ &= P(y', \vartheta') - \frac{\log(2\pi\sigma^2)}{2} - \frac{(y-\delta)^2}{2\sigma^2} N(y', \vartheta'), \end{aligned}$$

where

$$\begin{aligned} N(y', \vartheta') &= \frac{\int_0^\infty \frac{1}{\gamma} \frac{s(\gamma)}{\sqrt{2\pi\sigma'^2\gamma}} e^{-\frac{(y'-\delta')^2}{2\sigma'^2\gamma}}\, d\gamma}{f_{\vartheta'}\left(y'\right)}, \\ P(y', \vartheta') &= \frac{\int_0^\infty \log \frac{s(\gamma)}{\sqrt{\gamma}} \cdot \frac{s(\gamma)}{\sqrt{2\pi\sigma'^2\gamma}} e^{-\frac{(y'-\delta')^2}{2\sigma'^2\gamma}}\, d\gamma}{f_{\vartheta'}\left(y'\right)}. \end{aligned}$$

Straight computations lead us to

$$\begin{aligned} Q^{\mathbf{y}}(\vartheta|\vartheta') &= \sum_{i=0}^{n-1} M(y_i^{\vartheta}, y_i^{\vartheta'}; \vartheta|\vartheta') - \log f_{\vartheta}(y_0^{\vartheta}) + \\ &\quad - (n-1)\log\left(F(\vartheta, y_0^{\vartheta})/(n-1)!\right) \\ &= \sum_{i=0}^{n-1} P(y_i^{\vartheta'}, \vartheta') - \frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=0}^{n-1} \frac{(y_i^{\vartheta} - \delta)^2}{2\sigma^2} N(y_i^{\vartheta'}, \vartheta') + \\ &\quad - \log f_{\vartheta}(y_0^{\vartheta}) - (n-1)\log\left(F(\vartheta, y_0^{\vartheta})/(n-1)!\right). \end{aligned} \tag{21}$$

If we would not consider the leave one out algorithm, the maximization would be based on the function of δ that is listed in the second line of the above. This is a simple quadratic function of δ and the maximum is easily found in the explicit form. However, in the unbounded density case, the algorithm would typically be stuck in a value $\hat{\delta}_n = y_0$ and in the next step the solution would favor the same $\hat{\delta}_{n+1} = y_0$. In the leave-one-out version of the algorithm as discussed above, the term in the last line of (21) will punish choosing the value $\hat{\delta}_{n+1}$ close to y_0 as $-\log f_{\vartheta}(y_0^{\vartheta})$ converges to minus infinity at δ approaching $y_0^{\vartheta_n}$. It would effectively be pushing away from taking ϑ_n approaching any particular observation. This would have a similar effect to taking out the term $M(y_0^{\vartheta}, y_0^{\vartheta'}; \vartheta|\vartheta')$ from the second line of (21). In this sense, it would be a leave-one-out EM algorithm in which we protect against sticking with δ in any particular observation.

Acknowledgment

The research of both authors has been supported by the Swedish Research Council Grant 2008-5382.

References

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of statistics* **5**, 151–157.

Chen, J., Tan, X. & Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica* **18**, 443.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

Ibragimov, I. A. & Khasminskii, R. Z. (1981a). Asymptotic behavior of statistical estimates of the shift parameter for samples with unbounded density. *Journal of Mathematical Sciences* **16**, 1035–1041. Translation from Russian, the original date of publication: 1976.

Ibragimov, I. A. & Khasminskii, R. Z. (1981b). *Statistical estimation, asymptotic theory*, vol. 16 of *Applications of Mathematics*. Springer. Translation from Russian, the original date of publication 1979.

Kotz, S., Kozubowski, T. & Podgórski, K. (2001). *The laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Progress in Mathematics Series. Birkhäuser.

Polfeldt, T. (1970a). Minimum variance order when estimating the location of an irregularity in the density. *The Annals of Mathematical Statistics* **41**, 673–679.

Polfeldt, T. (1970b). The order of the minimum variance in a non-regular case. *The Annals of Mathematical Statistics* **41**, 667–672.

Protassov, R. (2004). Em-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ. *Statistics and Computing* **14**, 67–77.

Rao, B. (1966). *Asymptotic distributions in some nonregular statistical problems*. Ph.D. thesis, Michigan State University.

Rao, B. (1968). Estimation of the location of the cusp of a continuous density. *The Annals of Mathematical Statistics* **39**, 76–87.

Seo, B. & Kim, D. (2012). Root selection in normal mixture models. *Computational Statistics & Data Analysis* **56**, 2454–2470.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* **11**, 95–103.

B

Paper B

Convolution invariant subclasses of generalized hyperbolic distributions

Krzysztof Podgórski[1] and Jonas Wallin[2]

[1] *Department of Statistics, Lund University, Sweden*
[2] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*

Abstract

We show that the generalized Laplace distributions and the normal inverse Gaussian distributions together with the corresponding two classes of variance mixing distributions: the gamma distributions and the inverse Gaussian distributions are the only subclasses of the generalized hyperbolic distributions and, respectively, the generalized inverse Gaussian that are closed under convolution.

Key words: variance-mean normal mixture, generalized inverse Gaussian distribution, inverse gamma distribution, generalized asymmetric Laplace distribution, Bessel function distribution, gamma variance normal mixture

1 Preliminaries

In continuous time stochastic modeling, parametric classes of infinitely divisible distributions that are closed under convolution play a central role. Thus for any parametric family it is of interest to identify its convolution invariant subclasses. Within the generalized hyperbolic laws (GH), invariance under convolution of the generalized asymmetric Laplace distributions (also known as Bessel function or variance gamma distributions) and the normal inverse Gaussian (NIG) distributions is well known and frequently quoted in the literature, cf. Barndorff-Nielsen (1978), Kotz *et al.* (2001), and Bibby & Sørensen (2003). To quote from Bibby & Sørensen (2003): "However, in the case of the NIG and VG [variance gamma] distributions, the convolution properties ... imply that the value of the Lévy process will be NIG-distributed, respectively VG-distributed, at all time points. This makes the NIG and VG Lévy processes more natural generalized hyperbolic Lévy processes than the other generalized hyperbolic Lévy processes." Some authors also mention that within the GH distribution, there are no other convolution invariant families, see Fajardo & Farias (2004) (this paper, in fact, reports only the NIG class) or Hammerstein (2010). Despite these few mentions we did not find in the literature any explicit and rigorous argument for the characterizations of the convolution invariant GH subfamilies. The intention of this note is to provide such a one.

A wide range of infinitely divisible distributions can be obtained by mixtures of normal distributions. Such mixtures are distributed according to a density that is represented as a weighted average of normal densities. The most common is the variance mixture of normal densities, the density of which is given through

$$f(x) = \int_0^\infty \frac{1}{\sqrt{\gamma}} \varphi(x/\sqrt{\gamma})\, dF(\gamma),$$

where φ is a standard normal density and $dF(\gamma)$ an arbitrary probability distribution on $[0, \infty)$ that serves as weights with which the densities are mixed together. The variance mixture of normal distributions is equivalently given as the distribution of $X = \sqrt{\Gamma} Z$, where Γ is distributed according to F and independently of a standard normal variable Z.

A further natural (and well-known) extension of mixing of normal densities is given in the following definition.

Definition 3. A random variable X (and also the corresponding distribution) is called a normal variance-mean mixture with a non-negative mixing variable Γ,

variance scale $\sigma > 0$, mean scale $\mu \in \mathbb{R}$ if

$$X = \sigma\sqrt{\Gamma}Z + \mu\Gamma, \tag{1}$$

where Z is a standard normal variable independent of Γ.

This note is dealing with particular subclasses of the normal variance-mean mixtures obtained by restricting the distribution of Γ to some parametric subclass. Among discussed the most general is the class of generalized hyperbolic distributions with the generalized inverse Gaussian distributions as the corresponding class of mixing distributions. For the purpose of settling notation and terminology let us recall formal definitions of the two.

Generalized Inverse Gaussian (GIG) – This class of mixing distributions is given by the density

$$f(x) = \frac{\left(\psi/\chi\right)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} e^{-(\psi x+\chi/x)/2}, \;\; x > 0,$$

where the parameters satisfy

$$\begin{aligned}
&\psi > 0\,, \chi \geq 0\,, \text{ if } \lambda > 0,\\
&\psi > 0\,, \chi > 0\,, \text{ if } \lambda = 0,\\
&\psi \geq 0\,, \chi > 0\,, \text{ if } \lambda < 0.
\end{aligned}$$

The moment generating function of a GIG distribution is given by

$$M(t) = \left(\frac{\psi}{\psi - 2t}\right)^{\lambda/2} \frac{K_\lambda\left(\sqrt{\chi(\psi - 2t)}\right)}{K_\lambda(\sqrt{\psi\chi})}, \; t < \psi/2. \tag{2}$$

Let us mention two special cases.

- The inverse Gaussian distribution (the first passage time by a Brownian motion of a fixed level) is a GIG with $\lambda = -1/2$.
- The gamma distribution is GIG with $\chi = 0$.

Generalized Hyperbolic (GH) – This class is obtained as normal variance-mean mixtures (1) with Γ distributed as a GIG distribution.

The two special cases of GIG mentioned above specify two corresponding classes of the GH distributions, namely the generalized asymmetric Laplace (GAL) distributions with gamma as a mixing distribution and the normal inverse Gaussian (NIG) distributions with inverse Gaussian mixing. For more detailed information we refer to Eberlein & Keller (1995), for the generalized hyperbolic distributions, to Jørgensen (1982) for the generalized inverse Gaussian distributions, and to Kotz *et al.* (2001) for the generalized Laplace distributions.

2 Convolutions of normal variance-mean mixtures

Let us review some fundamental properties of the normal variance-mean mixtures. First, observe that the variance-mean normal mixtures coincide if and only if their mixing distributions are the same, see also Hammerstein (2010) for some related results.

Proposition 1. Let $X_1 = \sqrt{\Gamma_1}Z_1 + \mu\Gamma_1$ and $X_2 = \sqrt{\Gamma_2}Z_2 + \mu\Gamma_2$ be variance-mean normal mixtures. Then they have the same distribution if and only if Γ_1 and Γ_2 are also identically distributed.

Proof. Let $z = z(t,\mu) \in \mathbb{C}$ be a solution to $z^2/2 + \mu z - it = 0$. Then

$$\mathbb{E}\left(e^{zX_1}\right) = \mathbb{E}\left(e^{zX_2}\right), \tag{3}$$

whenever any of the sides is well defined.

However,

$$\begin{aligned}\mathbb{E}\left(e^{zX_1}\right) &= \mathbb{E}\left(\mathbb{E}\left(e^{z\cdot Z\sqrt{\gamma}+\mu\gamma z}|\gamma = \Gamma_1\right)\right)\\ &= \mathbb{E}\left(e^{\Gamma_1(z^2/2+\mu z)}\right)\\ &= \mathbb{E}\left(e^{it\Gamma_1}\right),\end{aligned}$$

which is the characteristic function of X_1. Since the same is true for X_2, it follows from (3) that both the characteristic functions are equal. □

Consider variance-mean mixtures $X_1 = \sqrt{\Gamma_1}Z_1 + \mu\Gamma_1$ and $X_2 = \sqrt{\Gamma_2}Z_2 + \mu\Gamma_2$, where (Γ_1, Z_1) and (Γ_2, Z_2) are independent. Then

$$X_1 + X_2 = \sqrt{\Gamma_1}Z_1 + \sqrt{\Gamma_2}Z_2 + \mu(\Gamma_1+\Gamma_2) \stackrel{d}{=} \sqrt{\Gamma_1+\Gamma_2}Z + \mu(\Gamma_1+\Gamma_2),$$

where Z is a normal random variable and $\stackrel{d}{=}$ stands for the equality of distributions. Hence the sum $X_1 + X_2$ is also a variance-mean mixture with the same scale μ and the mixing variable $\Gamma = \Gamma_1 + \Gamma_2$.

For $\mu \in \mathbb{R}$, let $\mathcal{F}_\mu$ be a sub-family of normal variance-mean mixtures of the form $X = \sqrt{\Gamma}Z + \mu\Gamma$, where $\Gamma \in \mathcal{G}$. From the properties shown above we have the following immediate result.

Proposition 2. For each $\mu \in \mathbb{R}$, $\mathcal{F}_\mu$ is closed under convolution if and only if $\mathcal{G}$ is closed under convolution.

3 Convolution invariance within GH distributions

As we have seen above, for the GH distributions it is sufficient to investigate the closeness under convolution for the corresponding variance mixing distributions, i.e. the GIG distributions. Thus next we investigate what subclasses of the GIG distributions are convolution invariant.

Lemma 1. Let Γ_1 and Γ_2 be two independent GIG distributed variables, with corresponding parameters $(\chi_1, \psi_1, \lambda_1)$ and $(\chi_2, \psi_2, \lambda_2)$, where χ_1 and χ_2 are greater then zero. For $\Gamma = \Gamma_1 + \Gamma_2$ to be again GIG, say, with parameters (χ, ψ, λ), it is necessary that

$$\begin{aligned}
&\psi = \min(\psi_1, \psi_2),\\
&\chi = (\sqrt{\chi_1} + \sqrt{\chi_2})^2,\\
&\lambda = \lambda_1 + \lambda_2 + 1/2
\end{aligned}$$

and, additionally,

$$\left(2^2\chi_1\chi_2\right)^{1/4} \psi^{\lambda/2} K_{\lambda_1}(\sqrt{\chi_1\psi_1}) K_{\lambda_2}(\sqrt{\chi_2\psi_2}) = \left(\pi^2\chi\right)^{1/4} \psi_1^{\lambda_1/2} \psi_2^{\lambda_2/2} K_\lambda(\sqrt{\chi\psi}). \tag{4}$$

Proof. The equality $\psi = \min(\psi_1, \psi_2)$ follows from the domain of the moment generating function given in (2).

If Γ is GIG with parameters (χ, ψ, λ), then for $t < \psi/2$:

$$\sqrt{\frac{\psi^\lambda}{\psi_1^{\lambda_1}\psi_2^{\lambda_2}}}\sqrt{\frac{(\psi_1-2t)^{\lambda_1}(\psi_2-2t)^{\lambda_2}}{(\psi-2t)^\lambda}}\frac{K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})}{K_\lambda(\sqrt{\chi\psi})} \tag{5}$$
$$= \frac{K_{\lambda_1}\left(\sqrt{\chi_1(\psi_1-2t)}\right)K_{\lambda_2}\left(\sqrt{\chi_2(\psi_2-2t)}\right)}{K_\lambda\left(\sqrt{\chi(\psi-2t)}\right)},$$

or, equivalently,

$$\begin{aligned} &A\cdot\exp\left(\sqrt{\chi_1(\psi_1-2t)}+\sqrt{\chi_2(\psi_2-2t)}-\sqrt{\chi(\psi-2t)}\right)\cdot \\ &\cdot\frac{(\psi_1-2t)^{\lambda_1/2+1/4}(\psi_2-2t)^{\lambda_2/2+1/4}}{(\psi-2t)^{\lambda/2+1/4}} \\ =&\frac{F_{\lambda_1}\left(\sqrt{\chi_1(\psi_1-2t)}\right)F_{\lambda_2}\left(\sqrt{\chi_2(\psi_2-2t)}\right)}{F_\lambda\left(\sqrt{\chi(\psi-2t)}\right)}, \end{aligned} \tag{6}$$

where $F_\nu(x) = \sqrt{2x/\pi}e^x K_\nu(x)$ and

$$A = \sqrt{(2/\pi)\sqrt{\chi_1\chi_2/\chi}\cdot\psi^\lambda\psi_1^{-\lambda_1}\psi_2^{-\lambda_2}}\cdot K_{\lambda_1}(\sqrt{\chi_1\psi_1})K_{\lambda_2}(\sqrt{\chi_2\psi_2})/K_\lambda(\sqrt{\chi\psi}).$$

Using the asymptotics

$$\lim_{x\to\infty}\sqrt{\frac{2x}{\pi}}e^x K_\nu(x) = 1$$

and letting t decreasing to negative infinity of the both sides of (6), the right hand side converges to one while the left hand side is converging either to zero or to infinity as long as $\sqrt{2\chi_1}+\sqrt{2\chi_2}\neq\sqrt{2\chi}$.

We can now assume that $\sqrt{\chi_1}+\sqrt{\chi_2} = \sqrt{\chi}$. The left hand side of (6) converges to the same value as

$$A\cdot\frac{(\psi_1-2t)^{\lambda_1/2+1/4}(\psi_2-2t)^{\lambda_2/2+1/4}}{(\psi^2-2t)^{\lambda+1/4}},$$

when $t\to-\infty$, that is zero or infinity unless $\lambda_1+\lambda_2+1/2=\lambda$. Under this constraint one obtains by passing with t to $-\infty$ in (6) the additional relation as given by (4). □

In the previous lemma we have exploited the behaviour of the moment generating function at zero. In the next one, we examine its behaviour at the upper boundary of the domain to derive further restrictions on the parameters.

Lemma 2. Let Γ_1 and Γ_2 be two independent and identically distributed GIG variables, with parameters (χ, ψ, λ), where χ is greater than zero. For $\Gamma = \Gamma_1 + \Gamma_2$ to be again GIG, and thus, by Lemma 1, having parameters $(2^2\chi, \psi, 2\lambda + 1/2)$, it is necessary for λ to be less then -1/4 and, additionally,

$$\sqrt{\pi}\cdot 2^{2\lambda+1/2}\cdot G_{-2\lambda-1/2}\left(2\sqrt{\chi(\psi-2t)}\right) = G^2_{-\lambda}\left(\sqrt{\chi(\psi-2t)}\right),\ t<\psi/2, \quad (7)$$

where $G_\nu(x) = x^\nu K_\nu(x)$.

Proof. We use equation (4) and the relation $K_{-\nu}(x) = K_\nu(x)$, to re-write equation (5) as

$$B\cdot(\psi-2t)^{-1/4+|\lambda|-|\lambda+1/4|} = \frac{G^2_{|-\lambda|}\left(\sqrt{\chi(\psi-2t)}\right)}{G_{|-2\lambda-1/2|}\left(2\sqrt{\chi(\psi-2t)}\right)} \quad (8)$$

where $B = \sqrt{\pi}\cdot 2^{-|2\lambda+1/2|}\cdot\chi^{-1/4+|\lambda|-|\lambda+1/4|}$.

For a positive ν,

$$\lim_{x\to 0^+} G_\nu(x) = \Gamma(\nu)2^{\nu-1} \quad (9)$$

so if t converges from below to $\psi/2$ and λ is greater or equal to $-1/4$, then the left hand side of (8) converges either to zero or is unbounded, while the right hand side convergent to a non-zero constant. □

We are ready for the main result that completely describes the convolution invariant families within the GIG distributions.

Theorem 4. Within the generalized inverse Gaussian distributions there are only two subclasses that are closed under convolution: the gamma distributions and the inverse Gaussian distributions.

Proof. Consider a subfamily of GIG distributions that is closed under convolution and let $(\chi_0, \psi_0, \lambda_0)$ be the parameters of a member of this subfamily.

Assume that $\chi_0 \neq 0$. If $\lambda_0 > -1/2$, then the increasing sequence defined through the recurrence relation $\lambda_n = 2\lambda_{n-1} + 1/2$, is a sequence of parameters

of some members in the family because of Lemma 2. Since λ_n increases without bound the terms will be eventually positive, which is not permitted as shown in Lemma 2.

Now assume that $\lambda_0 < -1/2$, so that for sufficiently large n leads to $-\lambda_n - 1 > 0$. Differentiating both side of (7) with respect to $\sqrt{\chi(\psi - 2t)}$, and using the identity

$$\left[x^{\nu} K_{\nu}(x)\right]' = -x^{\nu-1} K_{\nu-1}(x),\ \nu > 0,$$

we obtain

$$\begin{aligned} &\sqrt{\pi} 2^{2\lambda_n+1/2} G_{-2\lambda_n-3/2}\left(2\sqrt{\chi_n(\psi_n - 2t)}\right) \\ &\qquad = G_{-\lambda_n}\left(\sqrt{\chi_n(\psi_n - 2t)}\right) G_{-\lambda_n-1}\left(\sqrt{\chi_n(\psi_n - 2t)}\right) \end{aligned} \tag{10}$$

Let now consider the limits in the above when $t \to \psi_n/2$. Since $-\lambda_n - 1 > 0$, applying (9) yields

$$\sqrt{\pi} 2^{2\lambda_n+1/2} \Gamma(-2\lambda_n - 3/2) 2^{-2\lambda_n-5/2} = \Gamma(-\lambda_n - 1) 2^{-\lambda_n-2} \Gamma(-\lambda_n) 2^{-\lambda_n-1},$$

or, equivalently,

$$\sqrt{\pi} 2^{2\lambda_n+1} \frac{-\lambda_n - 1}{-2\lambda_n - 3/2} = \frac{\Gamma^2(-\lambda_n)}{\Gamma(-2\lambda_n - 1/2)}. \tag{11}$$

On the other hand applying (9) in (7) yields

$$\sqrt{\pi} 2^{2\lambda_n+1} = \frac{\Gamma^2(-\lambda_n)}{\Gamma(-2\lambda_n - 1/2)},$$

which is only possible when $\lambda_n = -1/2$ contradicting that $\lambda_n < -1$.

We conclude that either $\lambda_0 = -1/2$, which corresponds to a member of the inverse Gaussian distributions that is closed under convolutions, or $\chi_0 = 0$, which corresponds to another convolution closed family, namely that of the gamma distributions. □

An immediate consequence is the following characterization.

Corollary 1. Within the class of the generalized hyperbolical distributions with the support over the entire real line only two classes of distributions are closed under the convolution: the generalized Laplace distributions and the normal inverse Gaussian distributions.

ACKNOWLEDGEMENT

The authors acknowledge the support of the Swedish Research Council Grant 2008-5382.

References

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics* **5**, pp. 151–157.

Bibby, B. & Sørensen, M. (2003). Hyperbolic processes in finance. In S. Rachev, ed., *Handbook of heavy tailed distributions in finance*. Elsevier, pp. 211–248.

Eberlein, E. & Keller, U. (1995). Hyperbolic distributions in finance. *Bernoulli* **1**, 281–299.

Fajardo, J. & Farias, A. (2004). Generalized gyperbolic distributions and Brazilian data. *Brazilian Review of Econometrics* **24**, 249–271.

Hammerstein, E. A. (2010). *Generalized hyperbolic distributions: theory and applications to CDO pricing*. Ph.D. thesis, University of Freiburg, Department of Mathematics and Physics Freiburg im Breisgau.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse gaussian distribution*. Lecture Notes in Statistics. Springer-Verlag.

Kotz, S., Kozubowski, T. & Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Birkhauser.

C

Paper C

Non-Gaussian Matérn fields with an application to precipitation modeling

JONAS WALLIN[1], DAVID BOLIN[2]

[1] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[2] *Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden*

Abstract

The recently proposed non-Gaussian Matérn random field models, generated through stochastic partial differential equations, are extended by considering the class of Generalized Hyperbolic processes as noise forcings. The models are also extended to the standard geostatistical setting where irregularly spaced observations are modeled using measurement errors and covariates. A maximum likelihood estimation technique based on the Monte Carlo Expectation Maximization algorithm is presented, and it is shown how the model can be used to do predictions at unobserved locations. Finally, an application to precipitation data is presented, and the performance of the non-Gaussian models is compared with standard Gaussian and transformed Gaussian models through cross-validation.

Key words: Matérn covariances, SPDE, Markov random fields, Laplace, Normal inverse Gaussian, MCEM algorithm

1 Introduction

Latent Gaussian models are at the heart of modern spatial statistics. The prime reasons for this are that they are both theoretically and practically easy to work with; there exists a well-developed theory for likelihood-based estimation of parameters and the important problem of spatial reconstruction is easily solved using the standard kriging prediction which is optimal for Gaussian models. For non-Gaussian datasets, the standard approach is to try to find some non-linear transformation that enables the use of Gaussian models. This approach is commonly referred to as trans-Gaussian Kriging (Cressie, 1993) and common transformations include the square root transform, (Cressie, 1993, Huerta *et al.*, 2004, Berrocal *et al.*, 2010, Sahu & Mardia, 2005) and the log transform (Cressie, 1993, Cameletti *et al.*, 2013, Bolin & Lindgren, 2011). An effect of using such transforms is that these induce a certain dependence structure between the mean and the covariance for the data in the untransformed scale.

For example, consider the commonly used square root transformed latent Gaussian model $\sqrt{y_i} = X(\mathbf{s}_i) + \varepsilon_i$, where y_i are the observations, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ is measurement noise, and $X(\mathbf{s})$ is a Gaussian field with a stationary covariance function and a mean value $\mathbf{B}(\mathbf{s})\beta$ modeled using some covariates $\mathbf{B}(\mathbf{s})$. According to this model, the mean and covariance of the data in the original scale is given by

$$\begin{aligned}
\mathsf{E}[y_i] &= \mathbb{C}_X(0) + 2(\mathbf{B}(\mathbf{s}_i)\beta)^2, \\
\mathbb{C}[y_i, y_j] &= 2\,\mathbb{C}_X(\mathbf{s}_i - \mathbf{s}_j)^2 + 4(\mathbf{B}(\mathbf{s}_i)\beta)(\mathbf{B}(\mathbf{s}_j)\beta)\,\mathbb{C}_X(\mathbf{s}_i - \mathbf{s}_j),
\end{aligned}$$

where $\mathbb{C}_X$ is the stationary covariance function of $X(\mathbf{s})$ with the measurement variance σ_ε^2 added at $\mathbf{0}$. Viewing the equations above, it is not obvious how to interpret the effect of the measurement error and the the mean field on the observations and the usage of covariates for the mean induces a non-stationary covariance function for the data.

Furthermore, the posterior variance of the process in the same scale as the data is given by

$$\mathsf{V}[X(\mathbf{s})^2|\mathbf{y}] = 2\mathsf{V}[X(\mathbf{s})|\mathbf{y}]^2 + 4\mathsf{E}[X(\mathbf{s})|\mathbf{y}]^2\mathsf{V}[X(\mathbf{s})|\mathbf{y}].$$

Hence, the observations $\mathbf{y}$ and the mean field affects the kriging variance for the transformed Gaussian model, through the term $\mathsf{E}[X(\mathbf{s})|\mathbf{y}]$. This dependence is often not unreasonable for real data, and it has even been used to generate covariance structures (Azaïs *et al.*, 2011). However, as the models grow more complex,

for example by introducing non-stationary covariance functions, spatially varying measurement errors, or covariates, the effects of the transformation methods become less transparent and more stale. In these situations, one would like to use latent non-Gaussian models without resorting to transformations.

Compared to the Gaussian models, very little research has been devoted to latent non-Gaussian models in geostatistics, and the aim of this work is therefore to develop such models. We state three goals: First, we want to find a class of non-Gaussian models that share some of the desirable properties of the Gaussian models while allowing for heavier tails and asymmetry in the data. Secondly, we want to provide tools for fitting these models to real data, assuming a latent structure with covariates and measurement noise. Finally, we want to provide tools for using the models for spatial reconstruction.

We will extend the work of Bolin (2013), where non-Gaussian models with Matérn covariances (Matérn, 1960) formulated as stochastic partial differential equations (SPDEs) driven by non-Gaussian noise were investigated. The work consisted of providing an existence result for such SPDEs, and in some detail study parameter estimation of SPDEs driven by generalized asymmetric Laplace (GAL) noise. Although this is a good starting point for providing the tools we seek, there are some major issues that have to be resolved in order to use those methods for real applications: The estimation procedure proposed in Bolin (2013) was based on using the Expectation Maximization (EM) algorithm, and it works well as long as there is no measurement noise and all nodes in the field are observed. Unfortunately, these requirements are too restrictive for practical applications. However, we will show that these requirements can be avoided, utilizing an Monte-Carlo Expectation Maximization (MCEM) algorithm, and extend the estimation technique to a larger class of non-Gaussian models.

The structure of the paper is as follows. In Section 2, a brief overview of the methodology used for representing the SPDE models is given. This section also introduces the class of models that is considered in this work, namely SPDE models driven by either GAL noise or Normal inverse Gaussian (NIG) noise and we argue that these two cases are the only relevant cases to consider in the class of generalized hyperbolic distributions for non regular sampled observations. In Section 3, we introduce the full hierarchical model that can be used for spatially irregular observations with covariates and measurement error. In Section 4, the MCEM parameter estimation procedure is derived and Section 5 shows how to do spatial prediction and kriging variance estimation using these models. Section

6 contains an application of these models to a real dataset consisting of monthly mean and max precipitation measurements, and results of the non-Gaussian models are compared with results obtained using standard Gaussian models and transformed Gaussian models. Finally, Section 7 contains some concluding remarks and ideas for future work.

2 Non-Gaussian SPDE-based models

The Gaussian Matérn fields are perhaps the most widely used models in spatial statistics. These are stationary and isotropic Gaussian fields with a covariance function on the form

$$C(\mathbf{h}) = \frac{2^{1-\nu}\varphi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})\varkappa^{2\nu}}(\varkappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\varkappa\|\mathbf{h}\|), \quad \mathbf{h}\in\mathbb{R}^d, \nu>0, \tag{1}$$

where d is the dimension of the domain, ν is a shape parameter, $\varkappa$ a scale parameter, φ^2 a variance parameter, and K_ν is a modified Bessel function of the second kind. Since the Matérn-type spatial structure has proven so useful in practice, we want to construct models with this type of spatial structure but with non-Gaussian marginal distributions. In order to do this, we use the fact that a Matérn field $X(\mathbf{s})$ can be viewed as a solution to the SPDE

$$(\varkappa^2-\Delta)^{\frac{\alpha}{2}}X = \dot{M}, \tag{2}$$

where $\Delta = \sum_{i=1}^{d}\frac{\partial^2}{\partial \mathbf{s}_i^2}$ is the Laplacian, and $\alpha = \nu + d/2$ (Whittle, 1963). The Gaussian Matérn fields are recovered by choosing $\dot{M}$ as Gaussian white noise scaled by a variance parameter φ, and the mathematical details of this construction in the case when $\dot{M}$ is non-Gaussian are given in Bolin (2013).

To use these models in practice, we need a method for producing efficient representations of their solutions. One such method is the Hilbert space approximation technique by Lindgren *et al.* (2011) which was extended by Bolin (2013) to the non-Gaussian case when $M(\mathbf{s})$ is a type G Lévy process.

Recall that a Lévy process is of type G if its increments can be represented as a Gaussian variance mixture $V^{1/2}Z$ where Z is a standard Gaussian variable and V is a non-negative infinitely divisible random variable. Rosiński (1991) showed that every type G Lévy process can be represented as a series expansion, and for a compact domain $D\in\mathbb{R}^d$ it can be written as $M(\mathbf{s}) = \sum_{k=1}^{\infty} Z_k g(\gamma_k)^{\frac{1}{2}}\mathbb{I}(\mathbf{s}\geq\mathbf{s}_k)$,

where the function g is the generalized inverse of the tail Lévy measure for V, Z_k are iid $\mathsf{N}(0,1)$ random variables, γ_i are iid standard exponential random variables, $\mathbf{s}_k$ are iid uniform random variables on D, and

$$\mathbb{I}(\mathbf{s} \geq \mathbf{s}_k) = \begin{cases} 1 & \text{if } s_i \geq s_{k,i} \text{ for all } i \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

Since V is infinitely divisible, there exists a non-decreasing Lévy process $V(\mathbf{s})$ with increments distributed the same as V. This process has the series representation $V(\mathbf{s}) = \sum_{k=1}^{\infty} g(\gamma_k)^{\frac{1}{2}} \mathbb{I}(\mathbf{s} \geq \mathbf{s}_k)$.

In the following sections, we briefly describe the Hilbert space approximation technique for the case when M is a type G process, and then introduce a subclass of the type G process that are suitable for the model (2).

2.1 Hilbert space approximations

Assume that M in (2) is a type G Lévy process. The starting point for the Hilbert space approximation method is to consider the stochastic weak formulation of the SPDE,

$$(\varkappa^2 - \Delta)^{\frac{\alpha}{2}} X(\psi) = \dot{M}(\psi), \tag{3}$$

where ψ is in some appropriate space of test functions and $\dot{M}(\psi)$ is defined as the linear functional $\dot{M}(\psi) = \int \psi(\mathbf{s}) M(\mathrm{d}\mathbf{s})$ (see Bolin, 2013, Appendix A for details). A finite element approximation of the solution X is then obtained by representing it as a finite basis expansion $X(\mathbf{s}) = \sum_{i=1}^{n} w_i \varphi_i(\mathbf{s})$, where $\{\varphi_i\}$ is a set of predetermined basis functions and the stochastic weights are calculated by requiring (3) to hold for only a specific set of test functions $\{\psi_i, i = 1, \ldots, n\}$. By assuming that $\{\psi_i\} = \{\varphi_i\}$, one obtains a method which is usually referred to as the Galerkin method and this gives an expression for the distribution of the stochastic weights conditionally on the variance process,

$$\mathbf{w}|V \sim \mathsf{N}\left(\mathbf{K}_\alpha^{-1}\mathbf{m}, \mathbf{K}_\alpha^{-1}\boldsymbol{\Sigma}\mathbf{K}_\alpha^{-1}\right). \tag{4}$$

Here $\mathbf{K}_\alpha = \mathbf{C}(\mathbf{C}^{-1}\mathbf{K})^{\alpha/2}$ and the matrices $\mathbf{K}$, $\mathbf{C}$, and $\boldsymbol{\Sigma}$ have elements given by $C_{ij} = \langle \varphi_i, \varphi_j \rangle$, $K_{ij} = \varkappa^2 \langle \varphi_i, \varphi_j \rangle + \langle \nabla\varphi_i, \nabla\varphi_j \rangle$, $\Sigma_{ij} = \int \varphi_i(\mathbf{s})\varphi_j(\mathbf{s}) V(\mathrm{d}\mathbf{s})$, and $m_i = \int \varphi_i(\mathbf{s}) V(\mathrm{d}\mathbf{s})$.

In order to get a practically useful representation, we need to be able to evaluate the integrals Σ_{ij} and m_i efficiently. Whether this is possible or not depends

on the basis $\{\varphi_i\}$ and the variance process $V(\mathbf{s})$. For the purpose of this work we choose to work with piecewise linear, compactly supported, finite element bases induced by triangulations of the domain of interest. For bases of this type, a mass-lumping procedure gives that $m_i = V_i$ and $\mathbf{\Sigma} = \text{diag}(V_1, V_2, \ldots, V_n)$, where

$$V_i = \int_{h_i} V(\,\mathrm{d}\mathbf{s}) \tag{5}$$

and h_i is the area associated with $\varphi_i(\mathbf{s})$. For further details, see Bolin (2013) and Lindgren *et al.* (2011).

2.2 The generalised hyperbolic processes

The most well known subclass of the type G Lévy process is the class of generalised Hyperbolic processes generated by the Generalized Hyperbolic (GH) distribution (see Barndorff-Nielsen, 1978, Eberlein & von Hammerstein, 2004). The GH distribution covers a wide range of distributions including the NIG distribution, the Normal inverse Gamma distribution, the GAL distribution, and the t-distribution.

The generalised Hyperbolic distribution has five parameters $\sigma, \nu \in \mathbb{R}^+$, γ, μ , $\tau \in \mathbb{R}$, and a density function

$$\begin{aligned} f(x) =& c_1 \left(\frac{\sqrt{(\nu\sigma)^2 + (x-\gamma)^2}}{c_2} \right)^{\tau - 1/2} e^{\frac{\mu}{\sigma^2}(x-\gamma)} \\ & K_{\tau-1/2}\left(c_2 \sqrt{(\nu\sigma)^2 + (x-\gamma)^2} \right), \end{aligned}$$

where $c_1 = \frac{2^{(\tau-1)/2}}{\sqrt{\pi}(\sigma^2\nu)^\tau K_\tau(\sqrt{2}\nu)}$ and $c_2 = \frac{1}{\sigma}\sqrt{2 + \frac{\mu^2}{\sigma^2}}$.. A *GH* r.v. X can be represented as

$$X = \gamma + \mu V + \sigma\sqrt{V}Z, \tag{6}$$

where V is a generalized inverse Gaussian r.v. $V \sim GIG(\tau, 2, \nu^2)$ and $Z \sim N(0,1)$. The $GIG(p, a, b)$ distribution has the density function

$$f(x) = \frac{\left(a/b\right)^{p/2}}{2K_p\left(\sqrt{ab}\right)} x^{p-1} e^{-\frac{ax + b/x}{2}}. \tag{7}$$

where the parameters satisfy $a > 0, b \geq 0$ if $p > 0$, $a > 0, b > 0$ if $p = 0$, and $a \geq 0, b > 0$ if $p < 0$. Two special cases of the GIG distribution are the inverse Gaussian (IG) distribution, obtained when $p = -1/2$, and the Gamma distribution, obtained when $b = 0$. We denote the gamma distribution by $\Gamma(p, a) = GIG(p, a, 0)$ and the inverse Gaussian distribution by $IG(a, b) = GIG(-1/2, a, b)$. For more details of the GIG distribution see Jørgensen (1982).

A property of the GH distribution which is important for likelihood-based parameter estimation is that the variance component V is GIG distributed also conditionally on X. However, integrals of the variance process $V(\mathbf{s})$ of a GH process will in general not have known parametric distributions, and the random variable V_i in equation (5) will therefore not have known parametric distributions in general. Without this property we are not able to derive likelihood-based parameter estimation procedures, nor make spatial predictions, for the models in this work.

The random variables V_i would have known parametric distributions if the variance process belonged to a class of distributions that is closed under convolution. There are only two special cases of the GH distribution for which the variance components are closed under convolution (Podgórski & Wallin, 2013). The first special case is the GAL distribution, in finance is known as the variance gamma distribution, which was studied in the context of the SPDE models in Bolin (2013), and the second is the NIG distribution. Thus, from now on, we focus on the SPDE model (2) driven by either GAL noise or NIG noise.

Examples of marginal distributions for non-Gaussian Matérn fields, generated by the SPDE model (2), are displayed in Figure 1. Compared with Gaussian Matérn fields, the advantage with using NIG or GAL noise is that we can allow for heavier tails and asymmetry in the marginal distributions. The main difference between using GAL noise instead of NIG noise is that the probability density function (pdf) for the NIG case always is differentiable, while the GAL case can allow for sharper peaks at the mode.

For practical implementations of the models, the most important thing to know about the GAL and NIG distributions is how they affect the Hilbert space approximation procedure. For both distributions, we get that $\mathbf{m}$ and $\mathbf{\Sigma}$ in the Hilbert space approximation (4) can be written as $m_i = \gamma\tau h_i + \mu V_i$, and $\Sigma = \text{diag}(V_1, \ldots, V_n)$ respectively. For the GAL distribution $V(s)$ is a gamma process, and the variance components V_i are therefore gamma distributed, $V_i \sim \Gamma(h_i\tau, 1)$. For the NIG distribution the $V(s)$ is a Inverse Gaussian (IG) process, and the

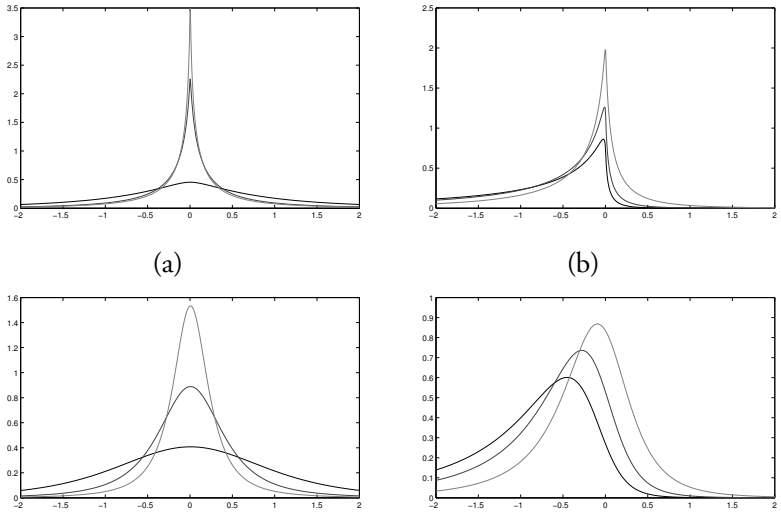


Figure 1: Examples of marginal probability density functions for $X(\mathbf{s})$ from (3) where $\dot{M}$ is either NIG or GAL noise. In Figure a) and b) the pdf is generated by GAL noise with different values of τ for the different curves in a) and different values of μ for the different curves in b). In Figure c) and d) the pdf is generated by NIG noise with varying ν^2 in a) and varying μ in b). For all examples, the random field $X(\mathbf{s})$ has a stationary Matérn covariance function with shape parameter $\alpha = 2$.

variance components are therefore IG distributed, $V_i \sim IG(\nu^2 h_i, 2)$.

Remark 1. If we would work on regular lattices, there are certain distributions in the GH family, such as the t-distribution, where one could imagine fixing the distributions such that the variance process has known distributions for the lattice points; however, having to work on regular lattices is a too strong restriction for us to consider such models any further. Also, even in situations where one has data on a regular lattice and only is interested in predictions to that same lattice, it is not clear what the corresponding continuous model would be if a model of this kind would be used.

3 Model extensions, covariates, and measurement noise

To use the models discussed above for real data, we assume a hierarchical model structure. The field of interest, $X(\mathbf{s})$, is modelled using one of the SPDE models, with observations, $y_1, \ldots, y_N$, at locations $\mathbf{s}_1, \ldots, \mathbf{s}_N$. In practice, these observations are often affected by measurement noise, and we thus need to include this in the model. Furthermore, we allow covariates for the mean value of the field by assuming that $X(\mathbf{s})$ is on the form

$$X(\mathbf{s}) = \sum_{i=1}^{n_x} B_i(\mathbf{s})\beta_i + \xi(\mathbf{s}), \tag{8}$$

where $\xi(\mathbf{s})$ is a SPDE field and $\{B_1, \ldots, B_{n_x}\}$ are known covariates, note that $\xi(\mathbf{s})$ not necessarily has zero mean in the non-Gaussian case. Using the representation (4) for $\xi(\mathbf{s})$, where the noise process is on the form of (6), we obtain the following hierarchical model, expressed in terms of the stochastic weights $\mathbf{w}$ for the basis expansion of $\xi(\mathbf{s})$

$$\begin{aligned} \mathbf{y} &= \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\mathbf{w} + \boldsymbol{\varepsilon}, \\ \mathbf{w} &= \mathbf{K}_\alpha^{-1}\left(\tau\mathbf{a}\gamma + \mathbf{V}\mu + \sigma\sqrt{\mathbf{V}}\mathbf{Z}\right). \end{aligned} \tag{9}$$

Here $\mathbf{A}$ is the observation matrix with elements $A_{ij} = \varphi_i(\mathbf{s}_j)$ linking the measurements to the latent field, $\mathbf{B}$ is a matrix containing the covariates $\{B_i\}$ evaluated at the measurement locations, and $\boldsymbol{\varepsilon}$ is a vector of iid $N(0, \sigma_\varepsilon^2)$ variables representing the measurement noise. The vector $\mathbf{Z}$ contains iid standard Gaussian variables and the distribution of V_i is determined by the noise process, specifically $V_i \sim \Gamma(\tau h_i, 1)$ for GAL noise and $V_i \sim IG(\nu^2 h_i, 2)$ for NIG noise and the V_i are independent, recall that $h_i = \int \varphi_i(\mathbf{s})\, \mathrm{d}\mathbf{s}$. To recover the latent field $X(\mathbf{s})$ at the measurement locations, one has to calculate $\mathbf{X} = \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\mathbf{w}$.

For the SPDE representation of the Gaussian Matérn fields it is easy to introduce non-stationarity in the model by allowing the covariance parameters to vary with space. In practice, this is achieved by representing the covariance parameters as regressions on some smooth covariates, e.g. assuming that $\varkappa(\mathbf{s}) = \exp\left(\sum B_{\varkappa,i}(s)\beta_{\varkappa,i}\right)$ where $\{B_{\varkappa,i}\}$ are known covariates would generate a model with a spatially varying covariance range. In the case of the model above, we have several parameters for the noise process, and it might be of interest to allow for these to vary with space as well, especially in cases when one has covariates that

not only affect the mean value of the field. This can be achieved in the same way as for the covariance parameters, by assuming regressions on some smooth covariates. For example, we can replace γ and μ in (9) by $\gamma(\mathbf{s}) = \sum B_{\gamma,i}(\mathbf{s})\gamma_i$ and $\mu(\mathbf{s}) = \sum B_{\mu,i}(\mathbf{s})\mu_i$ respectively, where $\{B_{\gamma,i}\}$ and $\{B_{\mu,i}\}$ are smooth covariates. Adding the covariates to (9) generates the following hierarchical model:

$$\begin{aligned} \mathbf{y} &= \mathbf{B}\boldsymbol{\beta} + \mathbf{A}\mathbf{w} + \boldsymbol{\varepsilon}, \\ \mathbf{w} &= \mathbf{K}_\alpha^{-1}\left(\tau\mathbf{B}_\gamma\boldsymbol{\gamma} + \mathbf{I}_{\mathbf{V}}\mathbf{B}_\mu\boldsymbol{\mu} + \sigma\sqrt{\mathbf{V}}\mathbf{Z}\right), \end{aligned} \tag{10}$$

where $\mathbf{I}_{\mathbf{V}} = diag(V_1, V_2, \ldots, V_n)$. The matrices $\mathbf{B}_\gamma$ and $\mathbf{B}_\mu$ are respectively given by $\{B_{\gamma,i}\}$ and $\{B_{\mu,i}\}$ evaluated at the node locations. This is a highly flexible model; however, one needs to be careful in defining the model so that the parameters are identifiable. One needs to be especially careful if using location covariates for both $\mathbf{X}$ ($\mathbf{B}$) and $\mathbf{w}$ ($\mathbf{B}_\gamma$) since this easily leads to a non-identifiable model unless the covariates are chosen carefully to avoid this issue.

4 Parameter estimation

Fitting the model above to data requires a parameter estimation method. In this section, we discuss how the parameters $\Theta = \{\varkappa, \boldsymbol{\beta}, \sigma_\varepsilon, \tau, \nu\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma\}$ can be estimated through likelihood methods for the NIG and GAL-driven SPDEs. The idea is to modify the EM-algorithm in Bolin (2013). The modification needed turns out to be the addition of Monte Carlo simulations to estimate the required expectations. We begin with a brief overview of the MCEM-algorithm and then cover the details needed to implement the procedure for our models.

4.1 Monte Carlo EM

The EM-algorithm (Dempster *et al.*, 1977) is convenient to use when the data-likelihood is difficult to work with but there exists some latent variables $\{\mathbf{w}, \mathbf{V}\}$ so that the augmented data $\{\mathbf{y}, \mathbf{w}, \mathbf{V}\}$ has a simpler likelihood (we utilize the same variable names in this subsection as in the rest of the paper for readability, but the result in this subsection is more general then for the models in this paper). The EM-algorithms uses the augmented likelihood $\pi(\mathbf{y}, \mathbf{w}, \mathbf{V}|\Theta)$ instead of the original likelihood $\pi(\mathbf{y}|\Theta)$, but requires the ability to compute expectations of the augmented likelihood.

The pth iteration of the EM-algorithm is done in two steps denoted the E-step and the M-step. In the E-step, one computes the function

$$\mathcal{Q}\left(\Theta, \Theta^{(p)}\right) = \mathsf{E}_{\mathbf{V}}\left[\log \pi(\mathbf{y}, \mathbf{w}, \mathbf{V}|\Theta)|\mathbf{y}, \Theta^{(p)}\right], \tag{11}$$

and in the M-step, one maximizes $\mathcal{Q}(\Theta, \Theta^{(p)})$ and obtains the $(p+1)$th iterate $\Theta^{(p+1)}$. The new iterate has the property $\pi(\mathbf{y}|\Theta^{(\mathbf{p+1})}) \geq \pi(\mathbf{y}|\Theta^{(\mathbf{p})})$ and under quite general conditions the procedure converges to a local maximum of the likelihood (Wu, 1983).

In certain cases when the E-step cannot be calculated analytically, one can use the MCEM algorithm, introduced in Wei & Tanner (1990). The idea of the MCEM algorithm is to replace $\mathcal{Q}$ in the E-step with

$$\mathcal{Q}^{MC}\left(\Theta, \Theta^{(p)}\right) = \frac{1}{k}\sum_{i=1}^{k} \log \pi(\mathbf{y}, \mathbf{V}^{(i)}, \mathbf{w}^{(i)}|\Theta), \tag{12}$$

where $\{\mathbf{w}^{(i)}, \mathbf{V}^{(i)}\}$ is a sample from the distribution $\pi(\mathbf{V}, \mathbf{w}|\mathbf{y}, \Theta^{(\mathbf{p})})$. In situations where it is not possible sample from the joint density for a set of variables $\{\mathbf{w}, \mathbf{V}\}$, but the conditional densities are available one can use the Gibbs sampling algorithm. The algorithm generates k samples from the joint density by sampling sequentially $\mathbf{w}^{(i)}|\mathbf{V}^{(i-1)}$ then $\mathbf{V}^{(i)}|\mathbf{w}^{(i)}$ for $i = 1, \ldots, k$. A downside is that the samples $\{\mathbf{w}^{(i)}, \mathbf{V}^{(i)}\}_{i=1}^{k}$ will not be independent and also a starting point $\mathbf{V}^{(0)}$ is required.

4.2 The E-step

For the model (10), the function $\mathcal{Q}$ in (11) cannot be calculated analytically, and numerical integration is not feasible for the large dimensions of both $\mathbf{w}$ and $\mathbf{V}$. We therefore use the Monte Carlo method described above to evaluate the E step.

Ideally we would simulate from $\pi(\mathbf{V}, \mathbf{w}|\mathbf{y}, \Theta^{(p)})$ in the MC sampler, but the joint distribution for $\{\mathbf{w}, \mathbf{V}\}$ is not known. However, a key observation is that the conditional distributions $\pi(\mathbf{V}|\mathbf{w}, \mathbf{y}, \Theta)$ and $\pi(\mathbf{w}|\mathbf{V}, \mathbf{y}, \Theta)$ are known, so we can use a Gibbs sampler to sample from the joint density.

Note that $\pi(\mathbf{w}|\mathbf{V}, \mathbf{y}, \Theta) \propto \pi(\mathbf{y}|\mathbf{w}, \mathbf{V}, \Theta)\pi(\mathbf{w}|\mathbf{V}, \Theta)$ where, by construction, $\{\mathbf{y}|\mathbf{w}, \mathbf{V}, \Theta\}$ and $\{\mathbf{w}|\mathbf{V}, \Theta\}$ are Gaussian, and $\{\mathbf{w}|\mathbf{V}, \mathbf{y}, \Theta\}$ is therefore also Gaussian. The explicit form of $\pi(\mathbf{w}|\{\mathbf{V}, \mathbf{y}, \Theta\})$ is $N(\hat{\mathbf{m}}, \hat{\mathbf{Q}}^{-1})$ where

$$\hat{\mathbf{m}} = \hat{\mathbf{Q}}^{-1}\left(\mathbf{Q}\mathbf{m} + \frac{1}{\sigma_\varepsilon^2}\mathbf{A}^\top(\mathbf{y} - \mathbf{B}\beta)\right), \hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{\sigma_\varepsilon^2}\mathbf{A}^\top\mathbf{A},$$

	GAL	*NIG*
p	$\mathbf{h}\tau - 1/2$	$-\mathbf{1}$
a	$(\mathbf{B}_\mu\boldsymbol{\mu})^2/\sigma^2 + 2$	$(\mathbf{B}_\mu\boldsymbol{\mu})^2/\sigma^2 + 2$
b	$(\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma})^2/\sigma^2$	$(\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma})^2/\sigma^2 + \mathbf{h}\nu^2$

Table 1: The distribution of $\{\mathbf{V}|\mathbf{w}, \Theta\}$, used in the Gibbs sampler, is $GIG(\mathbf{p}, \mathbf{a}, \mathbf{b})$ with parameters given in the table for the cases of NIG noise and GAL noise. Note that the distribution is independent of $\mathbf{Y}$ in both cases.

$$\mathbf{m} = \mathbf{K}_\alpha^{-1}(\mathbf{B}_\gamma\boldsymbol{\gamma} + \mathbf{I}_\mathbf{V}\mathbf{B}_\mu\boldsymbol{\mu}), \qquad \mathbf{Q} = \frac{1}{\sigma^2}\mathbf{K}_\alpha\mathbf{I}_\mathbf{V}^{-1}\mathbf{K}_\alpha.$$

The density of $\{\mathbf{V}|\mathbf{w}, \mathbf{y}, \Theta\}$ is proportional to

$$\pi(\mathbf{y}|\mathbf{w}, \mathbf{V}, \Theta)\pi(\mathbf{w}|\mathbf{V}, \Theta)\pi(\mathbf{V}|\Theta) \propto \pi(\mathbf{w}|\mathbf{V}, \Theta)\pi(\mathbf{V}|\Theta).$$

For both GAL and NIG processes, $\pi(\mathbf{V}|\Theta)$ can be written as $GIG(\mathbf{p}, \mathbf{a}, \mathbf{b})$ and we therefore get

$$\begin{aligned}
\pi(\mathbf{V}|\mathbf{w}, \mathbf{y}, \Theta) \propto & \left(\prod_j V_j^{p_j-1}\right)\left(\prod_j V_j^{-1/2}\right) \cdot \\
& \cdot \exp\Big(-\frac{1}{2}\Big(\mathbf{1}^\top\mathbf{I}_\mathbf{V}\mathbf{a} - \mathbf{1}^\top\mathbf{I}_\mathbf{V}^{-1}\mathbf{b}\Big) + \\
& - \frac{1}{2\sigma^2}(\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma} - \mathbf{I}_\mathbf{V}\mathbf{B}_\mu\boldsymbol{\mu})^\top\mathbf{I}_\mathbf{V}^{-1}(\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma} - \mathbf{I}_\mathbf{V}\mathbf{B}_\mu\boldsymbol{\mu})\Big) \\
= & \prod_j V_j^{p_j-3/2}\exp\Big(-\frac{1}{2}\left(\frac{(\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma})_j^2}{\sigma^2} + b_j\right)V_j^{-1} \\
& - \frac{1}{2}\left(\frac{(\mathbf{B}_\mu\boldsymbol{\mu})_j^2}{\sigma^2} + a_j\right)V_j\Big),
\end{aligned}$$

which is a GIG distribution with parameters given in Table 1 for the NIG and GAL cases.

	GAL	*NIG*
τ	$\max_\tau \frac{\tau}{k}\mathbf{h}^\top \left(\sum_{i=1}^k \log V^{(i)}\right) - \sum_{j=1}^n \log \Gamma(\tau h_j)$	$-1/2$
ν^2	0	$\left(\frac{\mathbf{1}^\top \mathbf{h}^{1/2} + \sqrt{(\mathbf{1}^\top \mathbf{h}^{1/2})^2 + 2n\mathbf{h}^\top \bar{\mathbf{V}}^{-1}}}{\sqrt{2}\mathbf{h}^\top \bar{\mathbf{V}}^{-1}}\right)^2$

Table 2: The parameter values that maximizes the function $\log \pi(\mathbf{V}|\tau, \nu^2)$ for the cases of GAL and NIG noise. Here $\bar{\mathbf{V}}^{-1} = \frac{1}{k}\sum_{i=1}^k (\mathbf{V}^{(i)})^{-1}$.

4.3 The M-step

To find the updating equations for the parameters, $\mathcal{Q}^{MC}$ should be maximized. The log-likelihood $\log \pi(\mathbf{y}, \mathbf{V}^{(i)}, \mathbf{w}^{(i)}|\boldsymbol{\Theta})$ can be divided into three terms

$$\log \pi(\mathbf{y}|\mathbf{w}^{(i)}, \mathbf{V}^{(i)}, \boldsymbol{\Theta}) + \log \pi(\mathbf{w}^{(i)}|\mathbf{V}^{(i)}, \boldsymbol{\Theta}) + \log \pi(\mathbf{V}^{(i)}|\boldsymbol{\Theta}). \tag{13}$$

The first term on the right hand side is a function of $\boldsymbol{\Theta}$ only through $\{\boldsymbol{\beta}, \sigma_\varepsilon\}$, the second term only through $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa\}$, and the third term only through $\{\tau, \nu\}$, together with that the first term is independent of $\mathbf{V}^{(i)}$ enables us to rewrite equation (13) as

$$\log \pi(\mathbf{y}^{(i)}|\mathbf{w}^{(i)}, \boldsymbol{\beta}, \sigma_\varepsilon) + \log \pi(\mathbf{w}^{(i)}|\mathbf{V}^{(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa) + \log \pi(\mathbf{V}^{(i)}|\tau, \nu). \tag{14}$$

Thus, the joint maximization of (14) for $\boldsymbol{\Theta}$ can be split into three separate steps, where maximization over $\{\tau, \nu\}$, $\{\boldsymbol{\beta}, \sigma_\varepsilon\}$ and $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa\}$ is performed independently.

The part of the log-likelihood depending on $\{\tau, \nu\}$ is

$$\log \pi(\mathbf{V}|\tau, \nu^2) = c + \begin{cases} \tau \mathbf{h}^\top \log \mathbf{V} - \sum_{i=1}^n \log \Gamma(\tau h_i) & \text{for GAL,} \\ n\log(\nu) + \sqrt{2}\mathbf{1}^\top \mathbf{h}^{1/2}\nu - \frac{1}{2}\mathbf{h}^\top \mathbf{V}^{-1}\nu^2 & \text{for NIG,} \end{cases}$$

where c is a constant. The maxima with respect to these parameters are given in Table 2. Note that this is the only part of the M step where the estimation for the NIG and GAL models differ. For the NIG model, the updating equation for ν^2 is given analytically whereas one has to do numerical optimization to update τ in the GAL model.

To update $\{\boldsymbol{\beta}, \sigma_\varepsilon\}$, one should maximize $\sum_i \log \pi(\mathbf{y}, \mathbf{w}^{(i)}|\boldsymbol{\beta}, \sigma_\varepsilon)$, where

$$\log \pi(\mathbf{y}, \mathbf{w}|\boldsymbol{\beta}, \sigma_\varepsilon) = -\frac{1}{2\sigma_\varepsilon^2}\left(\mathbf{y} - \mathbf{A}\mathbf{w} - \mathbf{B}\boldsymbol{\beta}\right)^\top \left(\mathbf{y} - \mathbf{A}\mathbf{w} - \mathbf{B}\boldsymbol{\beta}\right) - n\log(\sigma_\varepsilon) - \frac{n}{2}\log(2\pi).$$

The function is maximized by $\boldsymbol{\beta} = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{b}_x$ and $\sigma_\varepsilon^2 = (H_x - \mathbf{B}^\top\boldsymbol{\beta})$ where

$$\mathbf{b}_x = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{y} - \mathbf{A}\mathbf{w}^{(i)})^\top\mathbf{B}, \qquad H_x = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{y} - \mathbf{A}\mathbf{w}^{(i)})^\top(\mathbf{y} - \mathbf{A}\mathbf{w}^{(i)}).$$

In the third step, we find the maximum of the likelihood for $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa\}$, which only requires maximization of $\sum_i \log \pi(\mathbf{w}^{(i)}|\mathbf{V}^{(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa)$. The estimation needs to be done jointly for these parameters, and in general there is no closed form solution. However, it is possible to split this estimation step into two conditional maximization steps as described in Bolin (2013). An alternative is to use the fact that we can calculate the maximum of the function for a fixed $\varkappa$ by maximizing $\sum_i \log \pi(\mathbf{w}^{(i)}|\mathbf{V}^{(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa)$ over $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma\}$. For a fixed $\varkappa$, this function is maximized by

$$\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\gamma}\end{bmatrix} = \mathbf{Q}_{par}^{-1}\mathbf{b}, \quad \sigma^2 = \frac{1}{n}\left(H - \mathbf{b}^\top\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\gamma}\end{bmatrix}\right),$$

where

$$\mathbf{Q}_{par} = \frac{1}{k}\sum_{i=1}^{k}\begin{bmatrix}\mathbf{B}_\mu^\top\mathbf{I}_{\mathbf{V}^{(i)}}\mathbf{B}_\mu & \mathbf{B}_\mu^\top\mathbf{B}_\gamma\\ \mathbf{B}_\mu\mathbf{B}_\gamma^\top & \mathbf{B}_\gamma^\top\mathbf{I}_{\mathbf{V}^{(i)}}^{-1}\mathbf{B}_\gamma\end{bmatrix}, \ \mathbf{b} = \frac{1}{k}\sum_{i=1}^{k}\begin{bmatrix}(\mathbf{K}_\alpha\mathbf{w}^{(i)})^\top\mathbf{B}_\mu\\ (\mathbf{K}_\alpha\mathbf{w}^{(i)})^\top\mathbf{I}_{\mathbf{V}^{(i)}}^{-1}\mathbf{B}_\gamma\end{bmatrix},$$
$$H = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{K}_\alpha\mathbf{w}^{(i)})^\top\mathbf{I}_{\mathbf{V}^{(i)}}^{-1}\mathbf{K}_\alpha\mathbf{w}^{(i)}.$$

Inserting these expressions for $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma\}$ into $\sum_i \log \pi(\mathbf{w}^{(i)}|\mathbf{V}^{(i)}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa)$ yields an equation which is maximized numerically with respect to $\varkappa$ to find the new values for $\{\boldsymbol{\gamma}, \boldsymbol{\mu}, \sigma, \varkappa\}$. For $\alpha = 2$, this equation is given by

$$-\log(|\mathbf{K}_\alpha|) + \frac{n}{2}\log(H - \mathbf{b}^\top\mathbf{Q}_{par}\mathbf{b})$$

and similar, though more involved expressions can be found for other even values of α since $\mathbf{K}_\alpha$ can be written as a matrix polynomial in these cases.

A potential problem with the MCEM algorithm is that it could require a lot of memory if all values of $\{\mathbf{V}^{(i)}, \mathbf{w}^{(i)}\}$ for $i = 1, \ldots, k$ needed to be stored in order to evaluate the M step. However, as seen above, we only need to store a number of sufficient statistics in order to evaluate the M step. For $\alpha = 2$, these are given by

$$\begin{aligned}
&\sum_{i=1}^{k} (\mathbf{C}\mathbf{w}^{(i)})^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{C}\mathbf{w}^{(i)}, \quad \sum_{i=1}^{k} (\mathbf{C}\mathbf{w}^{(i)})^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{B}_\gamma, \quad \sum_{i=1}^{k} (\mathbf{G}\mathbf{w}^{(i)})^\top \mathbf{B}_\mu, \\
&\sum_{i=1}^{k} (\mathbf{C}\mathbf{w}^{(i)})^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{G}\mathbf{w}^{(i)}, \quad \sum_{i=1}^{k} (\mathbf{G}\mathbf{w}^{(i)})^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{B}_\gamma, \quad \sum_{i=1}^{k} (\mathbf{C}\mathbf{w}^{(i)})^\top \mathbf{B}_\mu, \\
&\sum_{i=1}^{k} (\mathbf{G}\mathbf{w}^{(i)})^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{G}\mathbf{w}^{(i)}, \quad \sum_{i=1}^{k} \mathbf{B}_\gamma^\top \mathbf{I}_{\mathbf{V}^{(i)}}^{-1} \mathbf{B}_\gamma, \quad \sum_{i=1}^{k} \mathbf{B}_\mu^\top \mathbf{I}_{\mathbf{V}^{(i)}} \mathbf{B}_\mu.
\end{aligned}$$

Thus, for $\alpha = 2$, we only need to store nine values to evaluate the M step. As α increases number of sufficient statistics required for storage will increase, but for any reasonable value of α the number of sufficient statistics is much smaller than the number of elements in $\{\mathbf{V}^{(i)}, \mathbf{w}^{(i)}\}$.

Remark 2. For the GIG distribution, $\mathsf{E}(\mathbf{V}^{-1})$ can be unbounded when $|p|$ is small and $b \to 0$. This makes the estimation of $\varkappa$ and γ problematic when $\min\left(|\tau\mathbf{h} - 1/2|\right)$ is small for the GAL model. The same problem exists for the EM algorithm in Bolin (2013), and that work gives some suggestions on how to improve the estimation in this situation.

4.4 Rao-Blackwellization

For each MC sample in the E-step, a sample $\mathbf{w}^{(i)}$, from $\pi(\mathbf{w}|\mathbf{y}, \Theta, \mathbf{V})$, is required. Sampling $\mathbf{w}^{(i)}$ requires a Cholesky decomposition of $\hat{\mathbf{Q}}$ which in general has a computational cost of $O(n^{3/2})$ for the SPDE models on $\mathbb{R}^2$, where n is the number of elements in $\mathbf{w}$. The Cholesky factorization dominates the total computational cost of the E-step, which in turn dominates the total computational cost of the MCEM algorithm. Thus, in order to reduce the computational cost of the estimation it is crucial to reduce the number of MC simulations in the E-step.

A common trick that can be used to reduce the number of required MC simulations to achieve a certain variance of the estimator is to note that for any

function h and any two random variables X and Y, one has that $\mathsf{E}[\mathsf{E}[h(X)|Y]] = \mathsf{E}[h(X)]$ and $\mathsf{V}[\mathsf{E}[h(X)|Y]] \leq \mathsf{V}[h(X)]$. When this is used in estimation, it is usually referred to as Rao-Blackwellization (see Robert & Casella, 2004) due to its association with the Rao-Blackwell Theorem (see Ferguson, 1967).

To apply Rao-Blackwellization to $\mathcal{Q}^{MC}$, we note that $\mathcal{Q}\left(\Theta, \Theta^{(p)}\right)$ can be written as $\mathsf{E}\left[\mathsf{E}\left[\log \pi(\mathbf{y}, \mathbf{V}, \mathbf{w}|\Theta)|\star\right]|\mathbf{y}, \Theta^{(p)}\right]$, where $\star$ denotes $\{\mathbf{y}, \mathbf{w}, \Theta^{(p)}\}$, the inner expectation is taken over $\mathbf{V}$, and the outer expectation is taken over $\mathbf{w}$. Viewing the log likelihood in equation (13) as a function of $\mathbf{V}$, one sees that

$$\begin{aligned}\mathsf{E}[\log \pi(\mathbf{y}, \mathbf{V}, \mathbf{w}|\Theta)|\star] = -\frac{1}{2\sigma^2}\Big((\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma})^\top \mathsf{E}[\mathbf{I}_{\mathbf{V}^{-1}}|\star](\mathbf{K}_\alpha\mathbf{w} - \mathbf{B}_\gamma\boldsymbol{\gamma}) \\ +\boldsymbol{\mu}^\top \mathbf{B}_\mu^\top \mathsf{E}[\mathbf{I}_{\mathbf{V}}|\star]\mathbf{B}_\mu\boldsymbol{\mu}\Big) - \mathsf{E}[\log \pi(\mathbf{V}|\tau, \nu^2)|\star]\end{aligned}$$

up to an additive constant, as a function of $\mathbf{V}$, where the last term is

$$\mathsf{E}[\log \pi(\mathbf{V}|\tau, \nu^2)|\star] = c + \begin{cases} 2^{-1}\mathbf{h}^\top \nu^2 \mathsf{E}[\mathbf{V}^{-1}|\star], & \text{for NIG noise,} \\ \tau\mathbf{h}^\top \mathsf{E}[\log \mathbf{V}|\star], & \text{for GAL noise.} \end{cases}$$

We therefore have the option to replace $\mathcal{Q}^{MC}$ with

$$\mathcal{Q}^{RB}(\Theta, \Theta^{(p)}) = \frac{1}{k}\sum_{i=1}^{k} \mathsf{E}\left[\log \pi(\mathbf{Y}, \mathbf{V}, \mathbf{w}^{(i)}|\Theta)|\mathbf{Y}, \mathbf{w}^{(i)}, \Theta^{(p)}\right],$$

which is a Rao-Blackwelliztion of $\mathcal{Q}^{MC}(\Theta, \Theta^{(p)})$. Here, the expectations $\mathsf{E}[\mathbf{V}|\star]$, $\mathsf{E}[\mathbf{V}^{-1}|\star]$, and $\mathsf{E}[\log \mathbf{V}|\star]$ can be computed numerically using the following formulas for the expectations of a $GIG(p, a, b)$ random variable V

$$\begin{aligned}\mathsf{E}[V^\lambda] &= (b/a)^{\lambda/2}\frac{K_{p+\lambda}\left(\sqrt{ab}\right)}{K_p\left(\sqrt{ab}\right)}, \quad \lambda \in \mathbb{R} \\ \mathsf{E}[\log(V)] &= \log(\sqrt{a/b}) + \frac{\partial}{\partial p}\log K_p\left(\sqrt{ab}\right).\end{aligned}$$

The expectation $\mathsf{E}[\log(V)]$ can be approximated by approximating

$$\frac{\partial}{\partial p}\log K_p\left(\sqrt{ab}\right) \approx \left(\log K_{p+\varepsilon}\left(\sqrt{ab}\right) - \log K_p\left(\sqrt{ab}\right)\right)/\varepsilon$$

for some small $\varepsilon > 0$.

5 Prediction

One of the main problems in spatial statistics is prediction of the latent field at locations where there are no observations. The two main characteristics that are reported in such predictions are the mean and variance of the predictive distribution. In this section, we show how to generate these two quantities for predictions, using the models described previously, at a set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_p$.

Let $\mathbf{A}_p$ be a $p \times n$ observation matrix, constructed the same way as the observation matrix in Section 3, for the locations $\mathbf{s}_1, \ldots, \mathbf{s}_p$. The desired mean values and variances are $\mathsf{E}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta]$ and $\mathsf{V}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta]$ respectively. Since the density of $\mathbf{w}|\mathbf{y}$ is not known, the mean and variance cannot be calculated analytically, and we therefore utilize MC methods to approximate the mean as $\mathsf{E}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta] \approx \frac{1}{k}\sum_{i=1}^{k} \mathbf{A}_p\mathbf{w}^{(i)}$ and the variance as $\mathsf{V}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta] \approx \frac{1}{k}\sum_{i=1}^{k} (\mathbf{A}_p\mathbf{w}^{(i)} - \mathsf{E}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta])^2$, where $\mathbf{w}^{(i)}$ is generated using the Gibbs-sampler described in Section 4.2.

Rao-Blackwelliztion can again be used to reduce the variance of the MC estimates. For the mean, write

$$\begin{aligned}
\mathsf{E}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta] &= \int_{\mathbf{w}} \mathbf{A}_p\mathbf{w}\pi(\mathbf{w}|\mathbf{y}, \Theta)\, \mathrm{d}\mathbf{w} \\
&= \int_{\mathbf{w}}\int_{\mathbf{V}} \mathbf{A}_p\mathbf{w}\pi(\mathbf{w}|\mathbf{V}, \mathbf{y}, \Theta)\pi(\mathbf{V}|\mathbf{y}, \Theta)\, \mathrm{d}\mathbf{V}\, \mathrm{d}\mathbf{w} \\
&= \int_{\mathbf{V}} \mathbf{A}_p\hat{\mathbf{m}}\pi(\mathbf{V}|\mathbf{y}, \Theta) d\mathbf{V} \approx \frac{1}{k}\sum_{i=1}^{k} \mathbf{A}_p\hat{\mathbf{m}}^{(i)},
\end{aligned}$$

which is a Rao-Blackwelliztion of $\mathsf{E}[\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta]$ where $\hat{\mathbf{m}}$ is the conditional mean of $\mathbf{w}$, defined in Section 4.2. Since the Gibbs-sampler uses $\hat{\mathbf{m}}^{(i)}$ to simulate $\mathbf{w}^{(i)}$, the Rao-Blackwelliztion can be produced from the MC sampler in the estimation step with no extra cost. The Rao-Blackwellization for the variance of the prediction is derived similarly as

$$\begin{aligned}
\mathsf{V}[\mathbf{A_p}\mathbf{w}|\mathbf{y}, \Theta] &= \int_{\mathbf{w}} \mathbf{A}_p \left(\mathbf{w} - \hat{\mathbf{w}}\right)\left(\mathbf{w} - \hat{\mathbf{w}}\right)^\top \mathbf{A}_p^\top \pi(\mathbf{A}_p\mathbf{w}|\mathbf{y}, \Theta)\, \mathrm{d}\mathbf{w} \\
&= \int_{\mathbf{V}} \mathbf{A}_p\hat{\mathbf{Q}}^{-1}\mathbf{A}_p^\top \pi(\mathbf{V}|\mathbf{y}, \Theta)\, \mathrm{d}\mathbf{V} \approx \frac{1}{k}\sum_{i=1}^{k} \mathbf{A}_p^\top \left(\hat{\mathbf{Q}}^{(i)}\right)^{-1} \mathbf{A}_p.
\end{aligned} \tag{15}$$

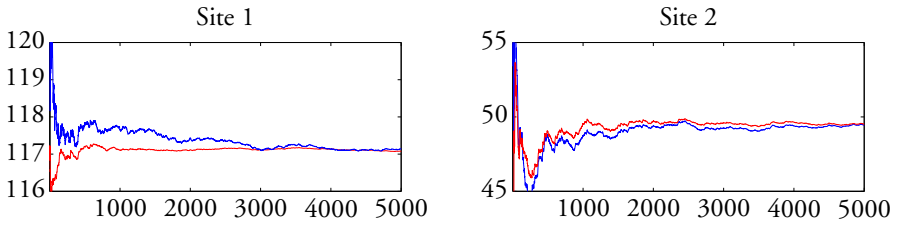


Figure 2: The figure shows the convergence of the Rao-Blackwellization estimator (red lines) and the regular Monte-Carlo estimator (blue lines) when estimating the conditional mean of a field at two distinct locations. For the first location (left panel), the Rao-Blackwellization improves the convergence, whereas the Rao-Blackwellization has no noticeable effect for the second location (right panel).

It would seem as one needs to calculate the inverse of $\hat{\mathbf{Q}}^{(i)}$, which is computationally expensive, to use Rao-Blackwellization of the variances. However, because of the structure of $\mathbf{A}_p$, only the elements of the inverse of $\hat{\mathbf{Q}}^{(i)}$ that corresponds to the non-zero elements in $\hat{\mathbf{Q}}^{(i)}$ are needed to evaluate (15). Using the methods in Campbell & Davis (1995), one can compute these elements at a computational cost of $O(n^{3/2})$, making Rao-Blackwellization for the variances computationally feasible.

To illustrate the effect of the Rao-Blackwellization, we examine the convergence of the Monte-Carlo estimator and the Rao-Blackwellization for the estimation of two conditional means at two distinct locations, $m_1 = \mathsf{E}[\mathbf{A}_1\mathbf{w}|\mathbf{y}, \Theta]$ and $m_2 = \mathsf{E}[\mathbf{A}_2\mathbf{w}|\mathbf{y}, \Theta]$, of the precipitation data used in Section 6. The results can be seen in Figure 2, the convergence of the estimation of m_1 is seen in the left panel and the convergence of the estimation of m_2 is seen in the right panel. As seen in the figure, the Rao-Blackwelliztion has a large effect on the convergence for m_1 whereas it has no visible effect on the convergence for m_2. The reason for this difference is that the largest part of variance of the MC method for m_1 comes from $\mathbf{w}|\mathbf{V}$ whereas the largest part of variance for m_2 comes from the variance of $\mathbf{V}|\mathbf{w}$.

6 An application to precipitation modeling

As an example of how the models presented earlier can be used for real data, we choose a dataset containing precipitation measurement over Parana, Brazil from

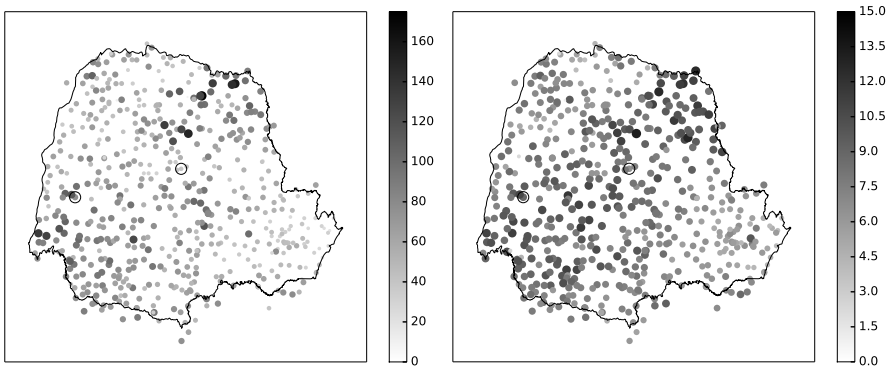


Figure 3: The figures displays the rain measurement for October over Parana, Brazil. To the left is the maximum daily precipitation of the month and to the right is monthly average. The two encircled locations are those locations where the predictive distributions are studied in the Figures 4 and 5; the left location is denoted $\mathbf{s}_2$ and the right $\mathbf{s}_1$.

2012[1]. Rainfall data over Parana has been previously been studied in a statistical framework by Diggle & Ribeiro Jr (2002). We study both the monthly average and the monthly maximum precipitation, for the month October, see Figure 3. The reason for choosing these two datasets is that one would suspect that Gaussian models could fit well for the monthly average, but not for the monthly maximum, and we want to investigate the difference in distribution for two cases and how well the the proposed models can fit these distributions.

6.1 Models

We will compare four different models for the data. The first three are models for the data in the original scale and the fourth is a Gaussian model for square-root transformed data. For the first three models, we assume that the measurements are generated as $y_i = X(\mathbf{s}_i) + \varepsilon_i$, where ε is Gaussian measurement noise with variance σ_ε^2 and $X(\mathbf{s}) = [\mathbf{1}, \mathbf{s}^*]\beta + \xi(\mathbf{s})$ is the latent precipitation field. Here ξ is a mean-zero Matérn field and $\mathbf{s}^*$ denotes longitude and latitude standardized

[1]Avialble in R-package `INLA` (`http://www.r-inla.org/`), see the tutorial available at `http://www.r-inla.org/examples/tutorials/spde-tutorial`

by removing the mean and dividing with standard deviation over the region. We fix the shape parameter α of the Matérn covariance function at two, but estimate the other parameters from the data. The three different models are obtained by choosing the forcing noise in the SPDE (2) as either Gaussian noise, GAL noise, or NIG noise.

The estimated parameters are presented in Table 3. In general, it is difficult to interpret all parameters for the non-Gaussian models. However a few things can be noted: For the maximum case, φ is almost zero for the NIG and GAL models, indicating that forcing noise is almost Gamma and inverse Gaussian noise respectively, and the variance of the measurement noise for the NIG and GAL models is smaller than the measurement noise variance for the Gaussian model. Also, $\varkappa$ for the non-Gaussian models is larger than for the Gaussian model, which indicates a shorter dependence range for the non-Gaussian models. Finally, the estimates for the covariates for the mean, β, are quite different for the different models.

We choose two of the measurement locations, encircled in Figure 3, where we investigate the posterior distribution of the latent precipitation field, $\mathbf{X}$, and the observations, $\mathbf{X} + \boldsymbol{\varepsilon}$. The reason for also studying $\mathbf{X} + \boldsymbol{\varepsilon}$ is that this quantity is what is observed if one removes the measurement at one location and then predicts the process at that location, as done in cross-validation. The posterior pdfs are presented in Figures 4 and 5. Note that we do not display $\mathbf{X}$ for the transformed Gaussian model since the latent field has no clear interpretation in this case. As expected, we observe no large differences in the pdfs for the different models for the monthly average data. However, the results for the monthly maximum data are more interesting. For the location $\mathbf{s}_1$, there is no large difference between the different models, and the reason for this is likely that all measurements near the location $\mathbf{s}_1$ are all similar to each other, which indicates a smooth field that could be well-approximated with a Gaussian field. One the other hand, the measurements are varying much more near the $\mathbf{s}_2$ location. This causes the non-Gaussian pdfs to be highly skewed, as the prediction at the location $\mathbf{s}_2$ will be more uncertain, but this skewness cannot be captured by the Gaussian model. Recall that all models have stationary Matérn covariance functions, so this ability to capture varying smoothness of the latent field is an interesting feature of the non-Gaussian models.

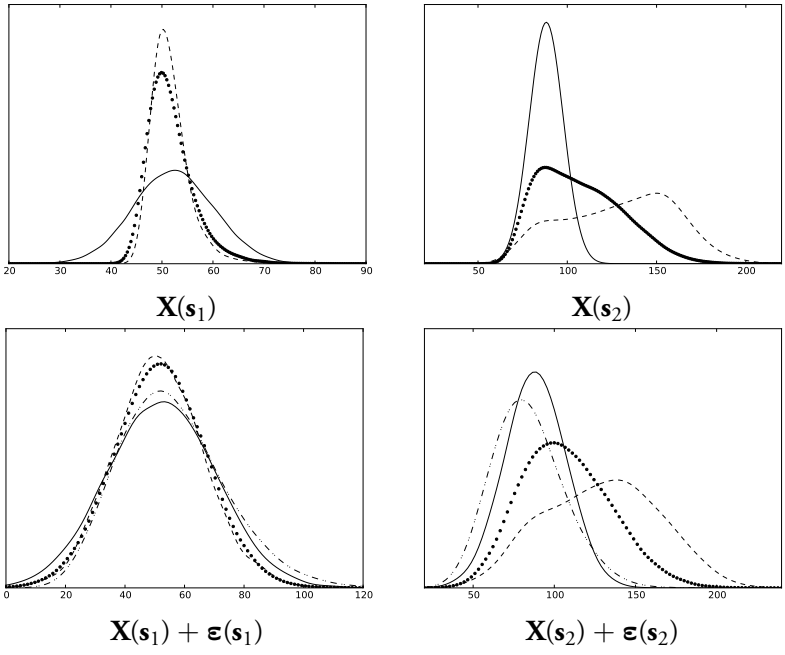


Figure 4: The posterior densities for $\mathbf{X}$ and $\mathbf{X}+\boldsymbol{\varepsilon}$ at the locations $\mathbf{s}_1$ and $\mathbf{s}_2$ for the monthly maxima data for the Gaussian model (solid), the NIG model (dotted), the GAL model (dashed), and the transformed Gaussian model (dash-dotted). The location of the points $\mathbf{s}_1, \mathbf{s}_2$ are displayed in Figure 3.

6.2 Model selection using cross-validation

We can conclude that there is a difference between the different model estimates, and a natural question is therefore which of the models that has the best fit to the data. We focus the model comparison on the accuracy of the spatial predictions and their corresponding error estimates.

To compare the different models ability to do spatial prediction, we use cross-validation. The dataset is divided into ten equally large groups $\mathbf{y}_1, \ldots, \mathbf{y}_{10}$, by doing a random permutation of the dataset and then choosing the first tenth of the dataset as $\mathbf{y}_1$, the second tenth as $\mathbf{y}_2$, etc. For each $k = 1, \ldots, 10$, the expectations $\mathsf{E}(\mathbf{y}_k|\mathbf{y}_{(-k)})$ and the variances $\mathsf{V}(\mathbf{y}_k|\mathbf{y}_{(-k)})$ are calculated, which are the spatial predictions and their variances for the locations in group k using all

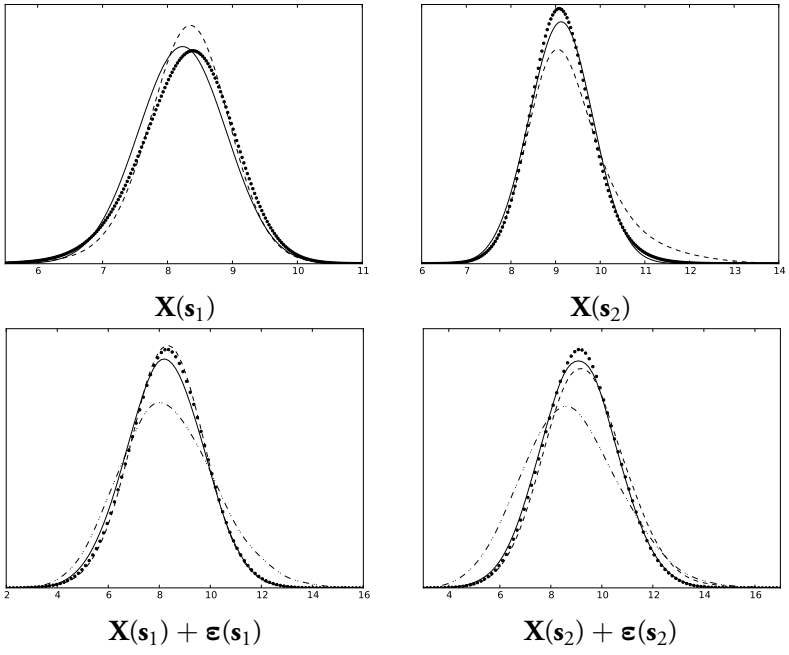


Figure 5: The posterior densities for $\mathbf{X}$ and $\mathbf{X}+\boldsymbol{\varepsilon}$ at the locations $\mathbf{s}_1$ and $\mathbf{s}_2$ for the monthly mean data for the Gaussian model (solid), the NIG model (dotted), the GAL model (dashed), and the transformed Gaussian model (dash-dotted). The location of the points $\mathbf{s}_1, \mathbf{s}_2$ are displayed in Figure 3.

data except the data in that group. By calculating these values for all groups, predictions are performed at all measurement locations, and by subtracting the measurements from these predictions we obtain a complete set of cross-validation residuals $\mathbf{r}$. By dividing each value in $\mathbf{r}$ with the predicted kriging variance for that location, we obtain a set of standardized residuals $\mathbf{r}_s$ which should have variance one if the model is correct.

We displays four statistics from the cross-validation predictions: The variance of the residuals; the variance of the standardised residuals, which should be close to one; the continuous ranked probability score (CRPS) (Matheson & Winkler, 1976), which probably is the most employed scoring role in probabilistic forecasts; and the energy score (ES), which is a multivariate extension of the CRPS.

Define $\hat{Y}_i$, $\hat{Y}_i^{(1)}$, $\hat{Y}_i^{(2)}$ as independent random variables with distribution $F_i = \pi(y_i|\mathbf{y}_{-k(i)})$ where $k(i)$ is the group that observation i belongs to, the CRPS is then given by

$$\text{CRPS} = m^{-1} \sum_{i=1}^{m} \left(\mathsf{E}_{F_i}[|y_i - \hat{Y}_i|] - \frac{1}{2} \mathsf{E}_{F_i}[|\hat{Y}_i^{(1)} - \hat{Y}_i^{(2)}|] \right). \tag{16}$$

where the expected values are approximated using MC approximation.

The results are presented in Table 4. For the monthly maximum data, the GAL model fits best, according to all four statistics, and there are no large differences between the other model, with the possible exception that the transformed Gaussian model overestimates the prediction variance.

For the monthly mean data, it is clear the transformed model does not fit the data very well but the other three models perform similarly. However, the Gaussian model seems to be slightly better than the other models according the the cross-validation statistics, and since it uses fewer parameter than the other models it is the best choice for this case.

7 Conclusions

In this work, we have extended the models of Bolin (2013) to a larger class of non-gaussian models and have shown how to handle missing data, covariates, and measurement noise, which is crucial for practical applications in geostatistics.

The models and the estimation procedure can be extend and improved in several directions. For example, as previously mentioned, for models defined on regular grids the full generalized hyperbolic class could be used and thus give a very large class of non-Gaussian fields on lattices. Also, the estimation procedure was derived assuming that the field was observed under Gaussian measurement error, but it would require only small modification to extend it to Generalised hyperbolic measurement noise, and this could improve the results for the presented application.

It is well-known that the convergence of the EM algorithm is slow, which often means that a large number of iterations are needed to achieve convergence of the parameter estimates, and the algorithm in this article is no exception. The author plans to study other stochastic estimation methods to improve the speed of the estimation. Changing to other estimation methods could also solve the

problem that we are currently only able to estimate the parameters when α is an even integer.

Unlike for Gaussian models, the models described here are not completely determined by the mean and covariance structures. This allows for interesting characteristics when applying other PDEs to the G-type processes. For example, one can create a spatio-temporal model that is not time reversible by considering a spatio-temporal extension of the models discussed here. This work has been a fist step in showing how one can use fully parametric non-Gaussian latent models in geostatistics, and although the results are promising, the main advantages are likely to occur when models like these are used in spatio-temporal applications.

	mean			
	Gauss	tGauss	NIG	GAL
$\varkappa$	1.40	1	2.0	2.0
φ	2.75	1	-	-
σ_ε	1.3	0.3	1.3	1.3
β	$\begin{bmatrix} 6.0 \\ -0.1 \\ -0.16 \end{bmatrix}$	$\begin{bmatrix} 2.7 \\ -0.0 \\ 0.0 \end{bmatrix}$	$\begin{bmatrix} 8 \\ -0.36 \\ 0.36 \end{bmatrix}$	$\begin{bmatrix} 11 \\ -0.2 \\ 0.5 \end{bmatrix}$
μ	-	-	-1.8	-1.0
σ	-	-	8.3	2.3
ν^2	-	-	0.2	-
τ	-	-	-	15

	max			
	Gauss	tGauss	NIG	GAL
$\varkappa$	2.6	1.9	5.8	5.9
φ	11.6	2.5	-	-
σ_ε	16.3	1.1	14 .4	13.5
β	$\begin{bmatrix} -79 \\ -3 \\ 0.3 \end{bmatrix}$	$\begin{bmatrix} 7.8 \\ -0.3 \\ 0.0 \end{bmatrix}$	$\begin{bmatrix} 36 \\ -6 \\ -6 \end{bmatrix}$	$\begin{bmatrix} 28 \\ -5 \\ -3 \end{bmatrix}$
μ	-	-	312	74
σ	-	-	0.0	0.0
ν^2	-	-	0.7	-
τ	-	-	-	17

Table 3: Parameter estimates for the different models for the precipitation max and mean data. Note that the tGauss parameters are for transformed data while Gauss, NIG, and GAL parameters are for raw data and hence should not be compared directly. This is, for example, the reason for the large differences in measurement noise variances.

	max				mean			
	Gauss	tGauss	NIG	GAL	Gauss	tGauss	NIG	GAL
$\mathbb{V}(\mathbf{r}_s)$	0.99	0.89	0.99	1.00	0.99	0.67	1.05	1.02
$\mathbb{V}(\mathbf{r})$	327	334	330	295	2.05	2.12	2.12	2.06
ES	301	304	310	287	24.0	24.5	24.2	24.0
CRPS	9.8	9.7	9.7	9.3	0.76	0.79	0.77	0.76

Table 4: Crossvalidation results for the different models. Here, $\mathbf{r}$ denotes the actual model residuals and $\mathbf{r}_s$ denotes the same residuals standardized by the estimated kriging variances. *CRPS* denotes the continuous ranked probability score of $\mathbf{r}$.

References

Azaïs, J.-M., Déjean, S., León, J. R. & Zwolska, F. (2011). Transformed gaussian stationary models for ocean waves. *Probab. Eng. Mech.* **26**, 342–349.

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scand. J. Statist.* **5**, 151–157.

Berrocal, V. J., Gelfand, A. E. & Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. agr. biol. and environ. statist.* **15**, 176–197.

Bolin, D. (2013). Spatial Matérn fields driven by non-Gaussian noise (in press). *Scand. J. Statist.* .

Bolin, D. & Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Statist.* **5**, 523–550.

Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *Advances in Statistical Analysis* **97**, 1–23.

Campbell, Y. E. & Davis, T. a. (1995). Computing the sparse inverse subset: an inverse multifrontal approach. Tech. Rep. TR-95-021, Computer and Information Sciences Department, University of Florida.

Cressie, N. (1993). *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **39**, 1–38.

Diggle, P. J. & Ribeiro Jr, P. J. (2002). Bayesian inference in gaussian model-based geostatistics. *Geographical and Environmental Modelling* **6**, 129–146.

Eberlein, E. & von Hammerstein, E. (2004). Generalized hyperbolic and inverse gaussian distributions: limiting cases and approximation of processes. In *Seminar on stochastic analysis, random fields and applications iv*, vol. 58. Progress in Probability, Birkhäuser, pp. 221–264.

Ferguson, T. (1967). *Mathematical statistics: a decision theoretic approach*. Probability and mathematical statistics. Academic Press.

Huerta, G., Sansó, B. & Stroud, J. R. (2004). A spatiotemporal model for mexico city ozone levels. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* **53**, 231–248.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse gaussian distribution*. Lecture Notes in Statistics. Springer-Verlag.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **73**, 423–498.

Matérn, B. (1960). Spatial variation. *Meddelanden från statens skogsforskningsinstitut* **49**.

Matheson, J. E. & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science* **22**, 1087–1096.

Podgórski, K. & Wallin, J. (2013). Convolution invariant generalized hyperbolic subclasses. *Preprints in Math. Sci. Lund University* **2013:2**.

Robert, C. & Casella, G. (2004). *Monte carlo statistical methods*. Springer Texts in Statistics. Springer.

Rosiński, J. (1991). On a class of infinitely divisible processes represented as mixtures of Gaussian processes. In *Stable processes and related topics*, vol. 25 of *Progress in Probability*. Birkhauser, Boston, pp. 27–41.

Sahu, S. K. & Mardia, K. V. (2005). A bayesian kriged kalman model for short-term forecasting of air pollution levels. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* **54**, 223–244.

Wei, G. C. & Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699–704.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.* **40**, 974–994.

Wu, C. (1983). On the convergence properties of the em algorithm. *Ann. Statist.* **11**, 95–103.

D

Paper D

A Gaussian mixture model for multivariate spatially dependent data using discrete and continuous Markov random fields

DAVID BOLIN[1], JONAS WALLIN[2] AND FINN LINDGREN[3]

[1] *Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden*
[2] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*
[3] *Mathematical Sciences, University of Bath, Bath, United Kingdom*

Abstract

A novel class of models is introduced with application ranging from land-use classification to brain imaging and geostatistics. The model class, denoted latent Gaussian random filed mixture models (LGFM models), combines the Markov random field mixture model with latent Gaussian random field models. The latent model, which is observed under measurement noise, is defined as a mixture of several, possible multivariate, Gaussian random fields. Which of the fields that is observed at each location is modeled using a discrete Markov random field. In order to use the method for massive data sets that arises in many possible areas of application, such as brain imaging, a computationally efficient parameter estimation method is required. Such an estimation method, based on a stochastic gradient algorithm, is developed and the model is tested on a magnetic resonance imaging application.

Key words: Gaussian mixture; Markov random fields; Random fields; Stochastic gradients

1 Introduction

Gaussian mixture models (GMMs) have successfully been used for classification in several areas of application ranging from video surveillance (Stauffer & Grimson, 1999) to speaker identification (Reynolds & Rose, 1995). Also in geostatistics and statistical image analysis, classification and image segmentation is often performed using GMMs in combination with the Expectation Maximization (EM) algorithm (Dempster *et al.*, 1977) for estimation. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ be observations of some, possibly multivariate, process $\mathbf{Y}(\mathbf{s})$ at locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$. The classical GMM can then be formulated as

$$\pi(\mathbf{Y}_i|\vartheta) = \sum_{k=1}^{K} w_{ik}\pi_k(\mathbf{Y}_i|\vartheta_k), \tag{1}$$

independently for all $i = 1, \ldots m$, where K is the number of classes, w_{ik} denotes the prior probability of $\mathbf{Y}_i$ belonging to class k, and $\pi_k(\mathbf{Y}_i|\vartheta_k)$ denotes the distribution of class k, which is assumed to be Gaussian, $\mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

A drawback with classification based on the classical GMM is that any spatial dependency of the data is ignored. A common strategy to account for spatial dependency in the data is allow for dependency in the allocation variables (w_{ik}), which can be done in several ways. One way is to model the class probabilities, w_{ik}, using a logistic normal model

$$w_{ik} = \frac{\exp(\eta_{ik})}{\sum_j \exp \eta_{ij}}, \tag{2}$$

where $\boldsymbol{\eta}_k$ are assumed to be latent Gaussian fields (Fernández & Green, 2002). Estimation under this model is difficult, and one generally has to resort to computationally expensive MCMC methods. Furthermore, for classification problems, the model is not ideal as the spatial model forces the posterior weights to be smoothly varying, which often can reduce the predictive power of the model.

Another way to allow for dependency in the mixture weights is to note that in the random variable $\mathbf{Y}_i$ defined in (1) equals, in distribution,

$$\sum_{k=1}^{K} z_{ik}\mathbf{G}_{ik}, \tag{3}$$

where $\mathbf{G}_{ik} \sim \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $z_{ik} = 1(x_i = k)$ is an indicator function for the event $x_i = k$, where x_i is a multinomial distributed r.v. defined though the probabilities

$\mathsf{P}(x_i = k) = w_{ik}$. Using this formulation of the GMM, spatial dependency can be introduced by assuming that $\mathbf{x} = \{x_i\}$ is a discrete MRF (see e.g. Held *et al.*, 1997, Zhang *et al.*, 2001, Van Leemput *et al.*, 1999). We refer to this model as a MRF mixture model.

Allowing for spatial dependency in the mixture weights is often reasonable and improves the classification for spatial problems. However, from a modeling perspective the MRF mixture models are not ideal since the data within each class is assumed to be independent observations of the same Gaussian distribution, while one would also like to allow for spatial dependency of the data within each class. Consider, for example, land-use classification from satellite images, where the classes in the mixture are assumed to correspond to distinct land types such as forest, fields, water, etc. For a given class, say forest, the measured values will depend on, for example, vegetation density and vegetation composition which makes the assumption of independent measurements within the class unrealistic.

In geostatistics, the most common approach to model spatially dependent data is to use latent Gaussian random fields (see e.g. Cressie, 1991, Cressie & Wikle, 2011). Collecting all measurements $\{\mathbf{Y}_i\}$ in a vector $\mathbf{Y}$, a latent Gaussian model can be written as

$$\mathbf{Y} = \mathbf{B}\beta + \mathbf{A}\xi + \boldsymbol{\varepsilon}, \tag{4}$$

where ξ is a (multivariate) mean-zero Gaussian random field, $\mathbf{A}$ is a matrix that connects the measurements to the latent field, and ε_i is Gaussian measurement noise. The matrix $\mathbf{B}$ contains covariates for the mean evaluated at the measurement locations, and the latent field evaluated at the measurement locations is given by $\mathbf{X} = \mathbf{B}\beta + \mathbf{A}\xi$. This modeling approach is often preferable if the latent process is smoothly varying and it is highly useful for noise reduction and spatial interpolation in cases of partial observations (Stein, 1999). However, the latent Gaussian random fields are poorly equipped to deal with the discontinuity of both process and covariance common for data in classification problems.

The aim of this work is twofold. First, we want to provide a new class of models that extends the MRF mixture models and can be used for spatial modeling of data that is usually studied in spatial classification problems. The goal is to provide a model class that can be used for classification but also for noise reduction and spatial interpolation. The model class we propose, which we will refer to as the latent Gaussian random field mixture(LGFM) models, combines the MRF mixture models and the latent Gaussian models, by assuming that the latent field

is a MRF mixture of Gaussian random fields. The possible application areas for this model class is much larger than those for the MRF mixture models and ranges from geostatistics and land-use classification problems to brain imaging and MRI modeling and estimation.

The second goal of this work is to provide an efficient estimation method for the LGFM and MRF mixture models that simplifies their usage for applications with massive datasets. The main computational bottle neck for estimation, through likelihood methods, for both the LGFM models and the MRF models is computing the normalizing constants. For the MRF models there exists several ways to handle this issue the two most common method are either gradient estimation or through pseudo likelihood estimation (Guyon, 1995). Recently, gradient methods for large scale GRF models have been developed for likelihood estimation that efficiently deals with the normalizing constants (Anitescu *et al.*, 2012, Stein *et al.*, 2013). We propose a stochastic version of the EM gradient method (Lange, 1995) based on pseudo-likelihoods. The method handles the normalizing constant for both the LGFM and the MRF mixture model.

The structure of this work is as follows. In Section 2, the model class is introduced and connections to other related models are discussed. Section 3 contains an introduction to a particular choice of the model components which is suitable for modeling of large datasets. Section 4 introduces an estimation procedure that is suitable for this model class but also for the standard MRF mixture models and the latent Gaussian models in cases of large datasets. In Section 5, the model class is used for noise reduction in magnetic resonance (MR) imaging. Finally, Section 6 contains a discussion of possible extensions and further work.

2 Latent Gaussian random field mixture models

Let $\mathbf{Y}$ be the vector of, possibly multivariate, observations. The general structure of the LGFM models is then

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}(\mathbf{s}_i) + \boldsymbol{\varepsilon}, \\ \mathbf{X}(s) &= \sum_{k=1}^{K} z_k(\mathbf{s}) \mathbf{X}_k(\mathbf{s}), \\ \mathbf{X}_k(s) &= \sum_{j=1}^{n} \mathbf{B}_{kj}(\mathbf{s}) \beta_{kj} + \boldsymbol{\xi}_k(\mathbf{s}). \end{aligned} \tag{5}$$

Here, $\boldsymbol{\varepsilon}$ is mean-zero Gaussian measurement noise and $\mathbf{X}(\mathbf{s})$ is the latent process. The latent process is described as a mixture of K Gaussian random field models, $\mathbf{X}_1, \ldots, \mathbf{X}_K$, and $\mathbf{z}$ is an indicator field that determines which class that is present at each location. Each Gaussian component is modeled using some covariates $\mathbf{B}_{kj}$ for the mean and a mean-zero Gaussian random field $\boldsymbol{\xi}_k$ with some covariance structure, which may be different for the different classes. This general class contains several interesting models, and some examples with $K = 2$ are shown in Figure 1. In the examples, X_k are independent stationary Gaussian Matérn fields. The indicator field z is obtained as $z_1(s) = \mathbb{I}_{Z(s)>0}(s)$, $z_2(s) = \mathbb{I}_{Z(s)\leq 0}(s)$ where $Z(s)$ is a Gaussian Matérn field, i.e. $z_1(s) = 1$ and $z_2(s) = 0$ for all s where $Z(s) > 0$ and $z_1(s) = 0$ and $z_2(s) = 1$ otherwise. In Panels (a) and (b), $Z(s)$ is independent of X_k. Panel (a) shows an example where X_1 and X_2 have the same covariance function but different mean values and Panel (b) shows an example where X_1 and X_2 have the same mean values but different correlation ranges. One can also imagine that z depends on some of the latent fields. Panels (c) and (d) are the same as Panels (a) and (b) except that $Z = X_1$. Thus, X_1 is only observed if it is positive and otherwise X_2 is observed.

There is a connection to the popular linear coregionalization models (LCM) (Zhang, 2007) in geostatistics. In our notation, an LCM can be written as

$$Y(\mathbf{s}_i) = \boldsymbol{\mu}(\mathbf{s}_i) + \sum_{k=0}^{K} \xi_k(\mathbf{s}_i),$$

and this model is thus a special case of the LGFM models when $z_k(\mathbf{s}) = 1$ for all k and $\mathbf{s}$.

For spatial classification problems, the domain for $\mathbf{s}$ is often discrete, e.g. pixels in satellite images or voxels in MR images. In such situations, the model can be written more compactly as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{z}_k \cdot \left(\mathbf{B}_k\boldsymbol{\beta}_k + \mathbf{A}\boldsymbol{\xi}_k\right) + \boldsymbol{\varepsilon}, \tag{6}$$

where $\cdot$ denotes element-wise multiplication, $\mathbf{B}$ is a matrix containing the covariates evaluated at the measurement locations, and $\mathbf{A}$ is a measurement matrix that determines which components in $\boldsymbol{\xi}_k$ that are observed. The latent field evaluated at the measurement locations is now given by $\mathbf{X} = \sum_{k=1}^{K} \mathbf{z}_k \cdot \left(\mathbf{B}\boldsymbol{\beta}_k + \mathbf{A}\boldsymbol{\xi}_k\right)$, which is a spatially correlated mixture of Gaussian random fields. Thus, there

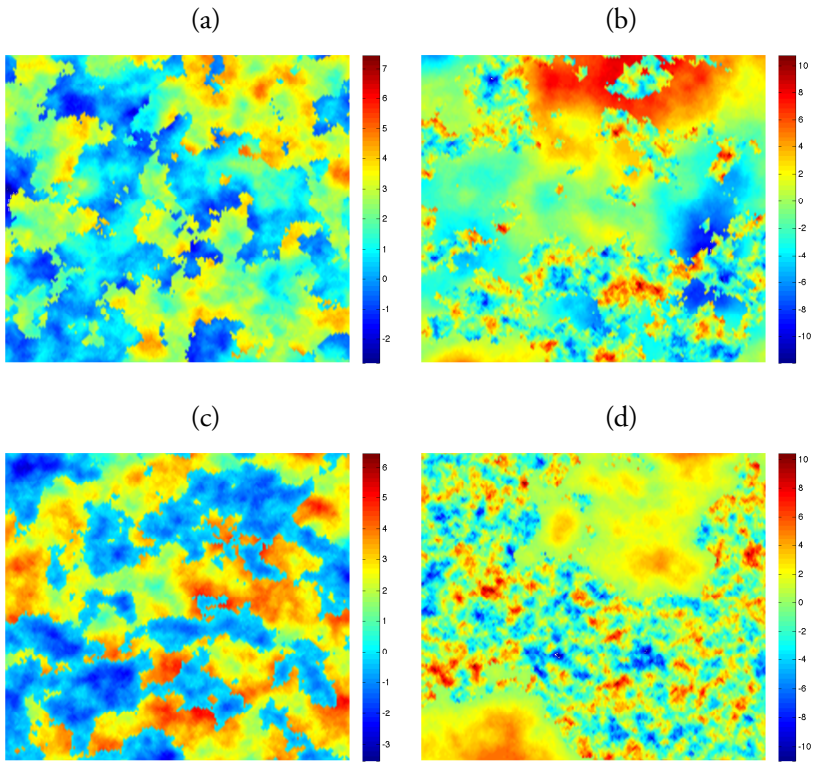


Figure 1: Examples of spatial mixture models with $K = 2$. The latent fields X_1 and X_2 are independent stationary Gaussian Matérn fields and z is obtained as $z_1(s) = Z(s) > 0$, $z_2(s) = Z(s) < 0$ where $Z(s)$ is a Gaussian Matérn field. In Panel (a), X_1 and X_2 have the same covariance function but different mean values and $X(s)$ is independent of X_k. In Panel (b), X_1 and X_2 have the same mean values but different correlation ranges and $X(s)$ is independent of X_k. Panels (c) and (d) are the same as Panels (a) and (b) respectively, except that $Z = X_1$.

is a clear connection between this model and the MRF mixture models; a MRF mixture model with spatially dependent components is obtained by choosing $\mathbf{z}$ as the indicator field of a discrete MRF.

For practical applications of the model one is typically interested in estimates of the latent field $\mathbf{X}$ given the data. For spatial prediction and noise reduction, $\mathsf{E}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is an estimate of the model parameters, is used as a point-estimate of the latent field and $\mathsf{V}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi})$ is used as a measure of the uncertainty in that prediction. To calculate these, we note that

$$\begin{aligned}\mathsf{E}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi}) &= \mathsf{E}[\mathsf{E}(\mathbf{X}|\mathbf{Y}, \mathbf{z}, \boldsymbol{\Psi})|\boldsymbol{\Psi}, \mathbf{Y}],\\ \mathsf{V}(\mathbf{X}|\mathbf{Y}, \boldsymbol{\Psi}) &= \mathsf{E}[\mathsf{V}(\mathbf{X}|\mathbf{Y}, \mathbf{z}, \boldsymbol{\Psi})|\boldsymbol{\Psi}, \mathbf{Y}] + \mathsf{V}[\mathsf{E}(\mathbf{X}|\mathbf{Y}, \mathbf{z}, \boldsymbol{\Psi})|\boldsymbol{\Psi}, \mathbf{Y}].\end{aligned}$$

Here, $\mathsf{E}(\mathbf{X}|\mathbf{Y}, \mathbf{z}, \boldsymbol{\Psi})$ and $\mathsf{V}(\mathbf{X}|\mathbf{Y}, \mathbf{z}, \boldsymbol{\Psi})$ can be calculated analytically since these are posterior means and covariances for Gaussian distributions. The outer expectation and variances, taken over $\mathbf{z}$, are typically not known analytically but can be estimated using Monte Carlo integration by sampling from $\pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi})$. While sampling $\mathbf{z}$, $\mathsf{E}(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi})$ can be estimated and used to classify the data.

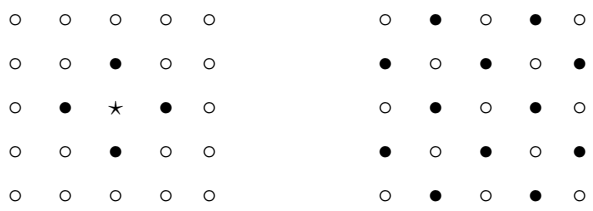
Since the model class is mainly targeted at applications on discrete domains, we choose to study the discrete model in more detail and leave the practical details of the continuous models for further research. In the following section, we outline a reasonable choice for the different components in the model that makes the model applicable to large spatial problems. And in Section 4, an estimation procedure for this particular model is presented.

3 Model components

In this section, we present a particular choice for the model components in (6) which is suitable for modeling of massive multivariate spatial datasets. To increase the computational efficiency of the model, Markov properties are used both for the indicator process $\mathbf{s}$ and for the latent fields $\boldsymbol{\xi}_k$.

3.1 A discrete MRF model for z

A suitable model for the indicator field, $\mathbf{z}$, determining the class belongings for each pixel, is a discrete MRF. We let $\mathbf{x}$ be a discrete MRF taking values in $\{1, \dots, K\}$ and define $z_{ik} = 1(x_i = k)$. The joint distribution of $\mathbf{x}$ can be formulated using the Gibbs distribution $p(\mathbf{x}) = Z^{-1} \exp(-W(\mathbf{x}))$ where $W(\mathbf{x}) = \sum_{\mathcal{C}} V_{\mathcal{C}}(\mathbf{x})$ is the



(a) (b)

Figure 2: A fist order neighborhood structure (a) and corresponding sets of conditionally independent pixels (b).

sum of the potential for all cliques generated by the neighborhood structure and $Z = \sum_{\omega} \exp(-W(\omega))$.

There are many potential choices for the neighborhood structure, but we use a simple first-order neighborhood $\mathcal{N}_\star$, which on a regular lattice in $\mathbb{R}^2$ consists of the four closest nodes, in euclidean distance, and in $\mathbb{R}^3$ consists of the six closest nodes. In $\mathbb{R}^2$, this neighborhood structure is illustrated in Figure 2 (a) where $\bullet$ denotes the neighbors of the pixel $\star$. For this neighborhood structure, there are only first and second-order cliques, and we use the potentials $V_{\{u\}}(\mathbf{x}) = \alpha_k$ when $x_u = k$, and $V_{\{u,v\}}(\mathbf{x}) = \beta_k$ when $x_u = k$ and $x_v = k$.

Hence, the model has parameters $\boldsymbol{\alpha} = \{\alpha_k\}$ and $\boldsymbol{\beta} = \{\beta_k\}$ where $\boldsymbol{\alpha}$ determines the prior probabilities for each class k and $\boldsymbol{\beta}$ are interaction parameters that governs the strength of the spatial dependency. Since only the difference between the αs effect the model, we fix α_1 to zero. Simplified models are obtained by either fixing αs to zero or by assuming that all β_k are equal to some common parameter β.

3.2 A Gaussian random field model for $\boldsymbol{\xi}$

We assume that $\boldsymbol{\xi}_k \sim \mathsf{N}(\mathbf{0}, \mathbf{Q}_k^{-1})$ is a multivariate spatial Gaussian random field with a covariance structure that is separable with respect to space the dimension of the data. This means that $\mathbf{Q}_k$ can be written as $\mathbf{Q}_k = \mathbf{Q}_{kd} \otimes \mathbf{Q}_{ks}$, where $\mathbf{Q}_{ks}$ is determined by a spatial covariance model and $\mathbf{Q}_{kd}$ is the multivariate part. The motivation behind this particular choice is that if there is no spatial dependence

in the data, one can choose $\mathbf{Q}_{ks}$ as the identity matrix and the model reduces to a standard MRF mixture model. Since the precision matrix $\mathbf{Q}_{dk}$ corresponds to the covariance matrices $\mathbf{\Sigma}_k$ in the MRF mixture model, we do not assume any special structure of this matrix. It is therefore parametrized as $\mathbf{Q}_{dk} = \mathbf{R}_{dk}^\top \mathbf{R}_{dk}$ where

$$\mathbf{R}_{dk} = \begin{bmatrix} \exp(\eta_1) & \eta_2 & \eta_4 & \cdots & \eta_x \\ 0 & \exp(\eta_3) & \eta_5 & \cdots & \vdots \\ 0 & 0 & \ddots & & \\ 0 & 0 & 0 & 0 & \exp(\eta_{d(d+1)/2}) \end{bmatrix} \tag{7}$$

is the unique Choesky factor of $\mathbf{Q}_{dk}$ with $d(d+1)/2$ parameters $\boldsymbol{\eta}_k$.

In general, there are no restrictions on the spatial structure of the process, specified through $\mathbf{Q}_s$. However, since we want to use the method for large problems we choose a model so that $\mathbf{Q}_s$ is sparse. For a discrete domain, we can then choose any type of GMRF model, e.g. the popular CAR models (Besag, 1974). The particular choice we use is a CAR model that corresponds to a Gaussian Matérn field. Constructing the spatial precision matrix using the SPDE connection (Lindgren *et al.*, 2011) between the discrete CAR models and the continuous Matérn fields allows us to use separate discretizations for $\mathbf{z}$ and $\boldsymbol{\xi}$, which is desirable if the data is such that the process $\boldsymbol{\xi}$ is smoothly varying compared to the resolution for $\mathbf{z}$. The basic idea is to use a basis expansion $\xi(\mathbf{s}) = \sum_{i=1}^n \varphi_i(\mathbf{s}) w_i$, where $\{\varphi_i\}$ are known compactly supported piecewise linear basis functions and $\mathbf{w} = \{w_i\}$ is a zero mean multivariate normal distribution with precision matrix $\mathbf{Q}_s = c\mathbf{K}\mathbf{C}^{-1}\mathbf{K}$, where $\mathbf{K} = (\mathbf{G} + \varkappa^2\mathbf{C})$ with $G_{ij} = \langle \nabla\varphi_i, \nabla\varphi_j \rangle$, $C_{ii} = \langle \varphi_i, 1 \rangle$ and c as a positive scaling constant. The number of basis functions, n, can be chosen smaller than the number of locations in the domain for $\mathbf{z}$ in order to increase the computational efficiency of the model.

This particular choice of $\mathbf{Q}_s$ corresponds to a Matérn field with shape parameter $\alpha = 2$, which for models in $\mathbb{R}^3$ results in the exponential covariance function. Since the parameter $\varkappa^2$ needs to be positive, we parametrize it as $\varkappa^2 = \exp(\varkappa_0)$. The constant c, in the precision matrix, is chosen so that the spatial part have variance one, which achieved for $c = \Gamma(2 - D/2)(4\pi)^{-D/2}\varkappa^{D-4}$, where D denotes the dimension of the spatial domain. This way, $\mathbf{Q}_s$ determined the spatial correlation and $\mathbf{Q}_d$ controls the variances.

The particular choice of covariance structure presented here is a so called proportional correlation model (Chiles & Delfiner, 1999) as the resulting stationary

covariance function for $\boldsymbol{\xi}$ can be written as $C(\mathbf{h}) = \mathbf{Q}_d^{-1}\rho(\mathbf{h}), \mathbf{h} \in \mathbb{R}^d$. There are several fully parametric alternatives to this model, such multivariate Matérn fields (Hu *et al.*, 2013).

3.3 The measurement noise $\boldsymbol{\varepsilon}$

We assume that the measurement noise $\boldsymbol{\varepsilon}$ is mean-zero Gaussian white noise with a spatially constant variance. One can either assume that the noise is the same for each dimension of the data, $\boldsymbol{\Sigma}_\varepsilon = \sigma^2\mathbf{I}_{nd}$, or one can allow for a separate variance for each dimension of the data, $\boldsymbol{\Sigma}_\varepsilon = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \otimes \mathbf{I}_n$. Here, $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. Since the variance parameters σ_i are positive, we parametrize them as $\sigma_i = \exp(\sigma_{i0})$

4 Parameter estimation

Parameter estimation for MRF mixture models is difficult, and allowing for spatial dependency within each class introduces further complications. Furthermore, we want these models to be useful for massive multivariate problems in $\mathbb{R}^3$, which are common in MR imaging, and this makes computational efficiency of the estimation procedure paramount.

The MRF mixture models are typically either estimated with some modified version of the EM algorithm or through Monte Carlo (MC) methods. Both of these procedures are too computationally demanding to be useful for the LGFM models. Instead, we base our estimation on the EM gradient (EMG) algorithm. The main idea behind this method is that if one can easily calculate the gradient $\nabla \log L(\boldsymbol{\Psi}; \mathbf{z}, \mathbf{Y})$ of the augmented likelihood, then knowing the posterior $\pi(\mathbf{z}|\mathbf{y}, \boldsymbol{\Psi})$ one can compute the exact gradient of the log likelihood $\log L(\boldsymbol{\Psi}; \mathbf{Y})$ as

$$\begin{aligned}
\nabla \log L(\boldsymbol{\Psi}; \mathbf{Y}) =& \nabla \log \pi(\mathbf{Y}|\boldsymbol{\Psi}) = \frac{1}{\pi(\mathbf{Y}|\boldsymbol{\Psi})} \nabla \int \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})\, \mathrm{d}\mathbf{z} \\
&= \int \frac{\pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})}{\pi(\mathbf{Y}|\boldsymbol{\Psi})} \nabla \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})\, \mathrm{d}\mathbf{z} \\
&= \int \pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi}) \nabla \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})\, \mathrm{d}\mathbf{z} = \\
&\quad \mathsf{E}_{\mathbf{z}} \left[\nabla \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi}) | \mathbf{Y}, \boldsymbol{\Psi} \right].
\end{aligned}$$

The idea is then to use the exact gradient in a gradient descent method. At step p in the ECG algorithm, the gradient of the likelihood is calculated and a step

$$\boldsymbol{\Psi}^{(p+1)} = \boldsymbol{\Psi}^{(p)} + \mathbf{S}\nabla \log L(\boldsymbol{\Psi}; \mathbf{Y})$$

where $\mathbf{S}$ is a matrix determining the step size. Taking $\mathbf{S} = \gamma\mathbf{I}$ where $\mathbf{I}$ is the identity matrix, we obtain an ordinary gradient descent method which has linear convergence. Ideally, we would like to take $\mathbf{S}$ as the inverse of the Hessian matrix $\mathbf{H}$ to obtain a Newton method with quadratic convergence. Often, one cannot compute the true Hessian matrix of the log-likelihood, and Lange (1995) instead proposed using

$$\mathbf{S} = \mathsf{E}_{\mathbf{z}}(\Delta \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})|\mathbf{Y}, \boldsymbol{\Psi}). \tag{8}$$

The motivation behind this choice of scaling matrix is that from dealing with spatial data we have experienced that the two first conditional moments often are little affected by changes in the parameters, which would indicate that $\mathbf{S}$ is a good approximation of the true hessian with the advantage of being readily available in most situations.

In the MRF mixture models, we cannot evaluate the gradient of the likelihood analytically, and one can then use MC sampling to estimate the gradient as

$$\nabla \log L(\boldsymbol{\Psi}; \mathbf{Y}) = \mathsf{E}_{\mathbf{z}}\left[\nabla \log \pi(\mathbf{Y}, \mathbf{z}|\boldsymbol{\Psi})|\mathbf{Y}, \boldsymbol{\Psi}\right] \approx \frac{1}{T}\sum_{t=1}^{T} \nabla \log \pi(\mathbf{Y}, \mathbf{z}^{(t)}|\boldsymbol{\Psi}),$$

where $\mathbf{z}^{(t)}$ are draws from $\pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi})$. In a similar fashion, one can use MC sampling to evaluate the approximate Hessian that is used to determine the step size

$$\mathbf{S} \approx \frac{1}{T}\sum_{t=1}^{T} \Delta \log \pi(\mathbf{Y}, \mathbf{z}^{(t)}|\boldsymbol{\Psi}).$$

We refer to this estimation procedure as the MCEMG algorithm.

To simplify the presentation, we split this section in three parts. In the first part, we go through the details of the estimation for the MRF mixture model, presenting a version of the method based on pseudo-likelihoods. In the second part we cover estimation for the latent Gaussian model, and one should note here that the estimation method is an attractive alternative for estimation of latent Gaussian

models for massive datasets since it avoids all calculations of log-determinants, which is usually the computational bottleneck in maximum-likelihood estimation procedures for such problems. Finally, we combine the results for the MRF mixture models and the latent Gaussian models to and estimation procedure for the full LGFM model.

4.1 Estimation of the MRF mixture model

As a first step towards an estimation method for the LGFM models, we in this section discuss parameter estimation of the MRF mixture models. To make the results of this section more easily applicable to the LGFM model, we parametrize the Gaussian distributions using the mean and cholesky factor of the precision matrix. Let $\vartheta_k = \{\mu_k, \mathbf{Q}_k\}$ where $\mathbf{Q}_{dk} = \mathbf{\Sigma}_k^{-1}$ is parametrized as $\mathbf{Q}_{dk} = \mathbf{R}_{dk}^\top \mathbf{R}_{dk}$ and $\mathbf{R}_{dk}$ has the form (7). Thus, the model parameters that need to be estimated are $\mathbf{\Psi} = \{\alpha, \beta, \vartheta\}$, where $\vartheta = \{\vartheta_1, \ldots, \vartheta_K\}$ contains all parameters for the Gaussian distributions, $\vartheta_k = \{\mu_k, \eta_k\}$.

Maximum likelihood estimation for this model is difficult since there is no simple form for the data likelihood. However, if we augment the data with the hidden class belongings, the augmented likelihood has a simpler form, $L(\mathbf{\Psi}; \mathbf{z}, \mathbf{Y}) = \pi(\mathbf{z}|\alpha, \beta)\pi(\mathbf{Y}|\mathbf{z}, \vartheta)$. This suggest that we could use an EM algorithm (Dempster *et al.*, 1977) where one would iterate calculating the function

$$\mathcal{Q}\left(\mathbf{\Psi}, \mathbf{\Psi}^{(p)}\right) = E\left[\log L(\mathbf{\Psi}; \mathbf{z}, \mathbf{Y})|\mathbf{Y}, \mathbf{\Psi}^{(p)}\right], \tag{9}$$

where $\mathbf{\Psi}^{(p)}$ denotes the current estimate of $\mathbf{\Psi}$ at the pth iteration of the algorithm and then maximize $\mathcal{Q}\left(\mathbf{\Psi}, \mathbf{\Psi}^{(p)}\right)$ with respect to $\mathbf{\Psi}$ in order to obtain the next estimate of the parameter vector.

Unfortunately, the normalizing constant Z for the MRF distribution depends on the parameters and is intractable for large problems. Thus, we cannot evaluate $\pi(\mathbf{z}|\alpha, \beta)$. A solution to this problem is to replace $\pi(\mathbf{z}|\alpha, \beta)$ with a pseudo-likelihood, $\pi_p(\mathbf{z}|\alpha, \beta)$, which is a product of the full conditionals of $\mathbf{x}$. Let $f_{ik} = \sum_{j \in \mathcal{N}_i} z_{jk}$ denote the sum of the neighboring pixels to z_{ik}, the conditional class probability of a pixel i can then be written as $\mathsf{P}(x_i = k|f_{ik}, \alpha, \beta) = \mathsf{E}(z_{ik}|f_{ik}, \alpha, \beta) \propto \exp\left(\alpha_k + \beta_k f_{ik}\right)$, and the pseudo-likelihood is

$$\pi_p(\mathbf{z}|\alpha, \beta) = \prod_i \pi(x_i|x_j, j \in \mathcal{N}_i, \alpha, \beta)$$

$$= \prod_i \frac{\exp(\sum_k \alpha_k z_{ik} + \sum_k \beta_k z_{ik} f_{ik})}{\sum_k \exp(\alpha_k + \beta_k f_{ik})}.$$

To avoid bias due to this procedure, only conditionally independent pixels are included in the product simultaneously, and the coding method (Besag, 1974) is used to combine the estimates based on different combinations of conditionally independent sets of pixels. Since the neighborhood structure in Figure 2 (a) is used, two sets of conditionally independent pixels are obtained using the checkerboard pattern shown in Figure 2 (b), where the black nodes are conditionally independent given the white nodes and vice versa.

The function we need to calculate the expectation of to obtain $\mathcal{Q}\left(\vartheta, \vartheta^{(p)}\right)$ is

$$\log PL(\mathbf{\Psi}; \mathbf{z}, \mathbf{Y}) = \log \pi(\mathbf{Y}|\mathbf{z}, \vartheta) + \sum_{k,i} \alpha_k z_{ik} + \sum_{k,i} \beta_k z_{ik} f_{jk} \qquad (10)$$
$$- \sum_i \log \left(\sum_k \exp(\alpha_k + \beta_k f_{ik}) \right).$$

Using the pseudo likelihood for the MRF part of $\mathcal{Q}$, the function can be written as

$$\mathcal{Q}(\mathbf{\Psi}, \mathbf{\Psi}^{(p)}) = \mathsf{E}(\log(\pi_p(\mathbf{z}|\alpha, \beta)) + \sum_i \sum_k \mathsf{E}(z_{ik}|\mathbf{Y}, \vartheta^{(p)}) \log \pi(Y_i|\vartheta_k).$$

We cannot evalute the expectation of the pseudo likelihood analytically, thus we replace it with an Monte Carlo approximation. The MC approximation requires sampling from the posterior distribution. using Bayes formula and the independence assumption, one has

$$\mathsf{E}\left(z_{ik} \mid f_{ik}, \mathbf{Y}, \mathbf{\Psi}\right) \propto p\left(y_i \mid \vartheta_k\right) \exp\left(\alpha_k + \beta_k f_{ik}\right) = \exp\left(\tilde{\alpha}_{ik} + \beta_k f_{ik}\right)$$

with $\tilde{\alpha}_{ik} = \alpha_k + \log \pi(Y_i|\vartheta_k)$. Thus, the posterior distribution is simply a non-stationary extension of the original MRF model. We therefore can use Gibbs sampling to simulate samples $\mathbf{z}^{(t)}, t = 1, \ldots, T$, from the posterior. Dividing the nodes using the checkerboard pattern in Figure 2 (b), and denoting the black nodes $\mathbf{z}_b$ and the white nodes $\mathbf{z}_w$, Gibbs sampling of the joint $\mathbf{z}$ is performed by iterating sampling $\mathbf{z}_w^{(i)}$ from $\pi(\mathbf{z}_w|\mathbf{z}_b^{(i-1)}, \mathbf{Y}, \mathbf{\Psi})$ and sampling $\mathbf{z}_b^{(i)}$ from $\pi(\mathbf{z}_b|\mathbf{z}_w^{(i)}, \mathbf{Y}, \mathbf{\Psi})$.

Now, this is about as far as one gets with the EM algorithm since the M step is highly problematic. Versions of the MRF mixture model has been used several times in tissue classification of magnetic resonance images (Held *et al.*, 1997, Zhang *et al.*, 2001, Van Leemput *et al.*, 1999), and in these situations the model is usually fitted to data using an EM estimator for the Gaussian parameters together with an iterated conditional modes (ICM) estimator for the MRF parameters. Convergence of this mixed estimation procedure is not easy to motivate theoretically, and the method can be computationally demanding.

However, the EM gradient method is straight-forward to implement. The derivatives need to evaluate the gradient are presented in Appendix A. At step p in the EM gradient algorithm, we run the Gibbs sampler to approximate the gradient and the scaling $\mathbf{S}$ and then take a step $\boldsymbol{\Phi}^{(p+1)} = \boldsymbol{\Phi}^{(p)} + \mathbf{S}\nabla \log PL(\boldsymbol{\Psi}; \mathbf{Y})$. Thus, there is no need for numerical optimization or Taylor approximations to calculate the parameter updates, as is needed if an EM algorithm is used. Note that $\nabla \log PL(\boldsymbol{\Psi}'; \mathbf{Y}) = \nabla \mathcal{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}')|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}'}$, thus the function maximized in the gradient algorithm is the same function maximized in the EM-algorithm.

4.2 Estimation of the latent Gaussian model

As a second step towards the estimation procedure for the full LGFM models, we in this section discuss the estimation of the latent Gaussian model (4) where $\boldsymbol{\xi}$ is given introduced in Section 3.2 and $\boldsymbol{\varepsilon}$ is introduced in Section 3.3. To simplify the presentation, we assume that the measurement noise has a common variance for all dimensions of the data, and the extension to separate noise variances is trivial.

Let $\boldsymbol{\Psi} = \{\boldsymbol{\mu}, \boldsymbol{\eta}, \sigma, \varkappa\}$ be the vector containing all model parameters. Since the model is Gaussian, likelihood estimation of all parameters can be performed by numerical optimization of $\log \pi(\boldsymbol{\Psi}|\mathbf{Y})$, which has a closed form (see e.g. Bolin & Lindgren, 2011). Even though this procedure is commonly used and theoretically straight-forward, it is computationally demanding. The problem is that one needs to calculate the determinant of $\hat{\mathbf{Q}} = \mathbf{Q} + \frac{1}{\sigma^2}\mathbf{A}^\top\mathbf{A}$ and solve the quadratic form $\mathbf{Y}^\top\mathbf{A}\hat{\mathbf{Q}}^{-1}\mathbf{A}^\top\mathbf{Y}$ each time the optimizer evaluates the likelihood. This is most efficiently done using sparse Cholesky factorization and backsubstitution; however, even though one has a separable covariance structure, this does not help when calculating the Cholesky factor, which makes the evaluation of the likelihood highly computationally demanding for large multivariate spatial problems.

The need to calculate the determinant of $\hat{\mathbf{Q}}$ is avoided if the EMG method

is used. Hence, the likelihood is augmented with the latent variable $\boldsymbol{\xi}$ and we calculate the gradient and the scaling matrix $\mathbf{S}$ by the procedure described above. The augmented log-likelihood is

$$l = \log \pi(\mathbf{Y}, \boldsymbol{\xi} | \boldsymbol{\Psi}) = - m\sigma_0 - \frac{1}{2e^{2\sigma_0}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\xi})^\top (\mathbf{Y} - \mathbf{B}\boldsymbol{\beta} - \mathbf{A}\boldsymbol{\xi}) + \tag{11}$$

$$+ \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2}\boldsymbol{\xi}^\top \mathbf{Q}\boldsymbol{\xi}, \tag{12}$$

and the derivatives needed to evaluate the gradient of the log-likelihood $L(\boldsymbol{\Psi}; \mathbf{Y})$ are presented in Appendix B.

The gradient method replaces computing $|\hat{\mathbf{Q}}|$ with computing various traces and there are two computational issues that have to be solved for the method to be applicable to large data sets. The first is to solve $\hat{\boldsymbol{\xi}} = \hat{\mathbf{Q}}^{-1}\mathbf{b}$ for a vector $\mathbf{b}$, which can be done using sparse cholesky factorizations and backsubstitution. However, in order to reduce the computationally complexity we instead use the preconditioned conjugate gradient method (PCG) with a robust incomplete Cholesky preconditioner (Ajiz & Jennings, 1984) to solve the equation.

The second issue is to solve the various traces of inverse matrices present in the expressions for the gradients. Recent work in spatial statistics (Anitescu *et al.*, 2012, Stein *et al.*, 2013) has proposed solving this issue using stochastic programming. The basic idea is to note that $\mathsf{E}[\mathbf{u}^\top \mathbf{Q}\mathbf{u}] = tr(\mathbf{Q})$ for any vector $\mathbf{u}$ of independent random variables u_i with mean zero and variance one (Hutchinson, 1990). Thus, we can rewrite all the traces in the gradient ∇l as expectations, which can ben be approximated using Monte Carlo integration. For example $\text{tr}\left(\mathbf{Q}_s^{-1}\frac{\partial \mathbf{Q}_s}{\partial \varphi_j}\right) = \mathsf{E}\left[\mathbf{u}^\top \frac{\partial \mathbf{Q}_s}{\partial \varphi_j}\mathbf{Q}_s^{-1}\mathbf{u}\right]$ is replaced with $k^{-1}\sum_{i=1}^{k}\mathbf{u}_i^\top \frac{\partial \mathbf{Q}_s}{\partial \varphi_j}\mathbf{Q}_s^{-1}\mathbf{u}_i$. The standard choice for $\mathbf{u}_i$ is to use mean-zero Bernoulli random variables but for spatial problems the variance of the estimator can be reduced by for example using the probing vectors proposed by Aune *et al.* (2012). The PCG method is used to efficiently calculate $\mathbf{Q}_s^{-1}\mathbf{u}_i$.

The resulting approximation, ∇l_k, of the gradient ∇l is a random function with $\mathsf{E}[\nabla l_k] = \nabla l$. Shapiro *et al.* (2009) shows that, under mild conditions, the local minimum of ∇l_k converges to a local minimum of ∇l with probability one as $k \to \infty$. Using the iterative methods in combination with the EM gradient method results in a highly computationally efficient method for estimating latent Gaussian models.

4.3 Estimation of the LGFM model

With the estimators for the MRF mixture model and the latent Gaussian model derived, it is now simply a matter of combining these two for making the estimator for the LGFM model. We augment the data-likelihood by both the MRF $\mathbf{z}$ and the GRFs $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k\}$, and let $l = \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ now denotes all model parameters. To calculate the required gradient, we note that

$$\begin{aligned}
\nabla \log L(\boldsymbol{\Psi}; \mathbf{Y}) &= \int \pi(\mathbf{z}, \boldsymbol{\xi}|\mathbf{Y}, \boldsymbol{\Psi}) \nabla \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\boldsymbol{\xi} \\
&= \int \pi(\mathbf{z}|\mathbf{Y}, \boldsymbol{\Psi}) \int \pi(\boldsymbol{\xi}|\mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi}) \nabla \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \, \mathrm{d}\boldsymbol{\xi} \, \mathrm{d}\mathbf{z} \\
&= \mathsf{E}_{\mathbf{z}} \left(\mathsf{E}_{\boldsymbol{\xi}} \left(\nabla \log \pi(\mathbf{Y}, \mathbf{z}, \boldsymbol{\xi}|\boldsymbol{\Psi}) \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi} \right) \mid \mathbf{Y}, \boldsymbol{\Psi} \right) \\
&= \mathsf{E}_{\mathbf{z}} \left(\nabla \log \pi(\mathbf{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) + \right. \\
&\quad \left. + \mathsf{E}_{\boldsymbol{\xi}} \left(\nabla \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\mathbf{z}, \boldsymbol{\sigma}) \mid \mathbf{z}, \mathbf{Y}, \boldsymbol{\Psi} \right) \mid \mathbf{Y}, \boldsymbol{\Psi} \right).
\end{aligned}$$

As in previous section, the expectation with respect to $\mathbf{z}$ must be approximated using MC sampling. However, since the expectation with respect to $\boldsymbol{\xi}$ is known analytically, see Appendix B, we can use Rao-Blackwellization to calculate gradient as

$$\begin{aligned}
\nabla \log L(\boldsymbol{\Psi}; \mathbf{Y}) = & \frac{1}{T} \sum_{t=1}^{T} \left(\nabla \log \pi(\mathbf{z}^{(t)}|\boldsymbol{\alpha}, \boldsymbol{\beta}) + \right. \\
& \left. + \mathsf{E}_{\boldsymbol{\xi}} \left(\nabla \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\mathbf{z}^{(t)}, \boldsymbol{\sigma}) \mid \mathbf{z}^{(t)}, \mathbf{Y}, \boldsymbol{\Psi} \right) \right).
\end{aligned}$$

Thus we can use the gradients calculated in the previous sections with two minor changes.

The first difference is that the Gaussian likelihood (11) for each, independent, field $\boldsymbol{\xi}_k$ is replaced with

$$\begin{aligned}
\log \pi(\mathbf{Y}, \boldsymbol{\xi}_k|\mathbf{z}^{(ti)}, \boldsymbol{\Psi}) = & - \frac{(\mathbf{Y}^{(i)} - \mathbf{B}_k^{(t)}\boldsymbol{\beta} - \mathbf{A}_k^{(i)}\boldsymbol{\xi}_k)^\top (\mathbf{Y}^{(i)} - \mathbf{B}_k^{(i)}\boldsymbol{\beta} - \mathbf{A}_k^{(i)}\boldsymbol{\xi}_k)}{2e^{2\sigma_0}} \\
& + \frac{1}{2} \log |\mathbf{Q}_k| - \frac{1}{2} \boldsymbol{\xi}_k^\top \mathbf{Q}_k \boldsymbol{\xi}_k - m_k^{(i)} \sigma_0
\end{aligned}$$

where $m_k^{(t)} = d \sum_j z_{kj}$ and $\mathbf{Y}^{(t)}$, $\mathbf{A}_k^{(t)}$ and $\mathbf{B}_k^{(t)}$ are constructed by taking $\mathbf{Y}$, $\mathbf{A}$, and $\mathbf{B}$ and only keeping the rows that corresponds to the pixels with $\mathbf{z}_k^{(i)} = 1$. Thus, m,

$\mathbf{A}$ and $\mathbf{B}$ are replaced with $m_k^{(t)}$, $\mathbf{A}_k^{(t)}$ and $\mathbf{B}_k^{(t)}$ respectively in the Gaussian gradients presented in Appendix B.

The second difference is how $\mathbf{z}^{(t)}$ is simulated. Unlike for the regular MRF mixture model, $\mathbf{Y}|\vartheta_k$ is not a vector independent variables and the sampling method for $\mathbf{z}$ in the MRF mixture model therefore has to be modified. To simulate $\mathbf{z}^{(i)}$, we introduce an extra step in the Gibbs sampler for the MRF mixture model as follows

1. Sample the Gaussian fields $\{\boldsymbol{\xi}_k\}^{(i)}$ from their respective distributions $\pi(\boldsymbol{\xi}_k|\mathbf{Y}, \mathbf{z}^{(i-1)}, \boldsymbol{\Psi})$.

2. Sample $\mathbf{z}_w^{(i)}$ from $\pi(\mathbf{z}_w|\mathbf{z}_b^{(i-1)}, \mathbf{Y}, \{\boldsymbol{\xi}_k\}^{(i)}, \boldsymbol{\Psi})$.

3. Sample $\mathbf{z}_b^{(i)}$ from $\pi(\mathbf{z}_b|\mathbf{z}_w^{(i)}, \mathbf{Y}, \{\boldsymbol{\xi}_k\}^{(i)}\}, \boldsymbol{\Psi})$.

Since $\mathbf{Y}|\{\boldsymbol{\xi}_k\}^{(i)}, \boldsymbol{\Psi}$ is a vector of independent variables, the second and third step of the Gibbs sampler are performed in the same way as for the MRF mixture model. It should also be noted that the sampled fields $\{\boldsymbol{\xi}_k\}^{(i)}$ are not used in the optimization other than to generate $\mathbf{z}^{(i)}$.

The simulation from $\pi(\boldsymbol{\xi}_k|\mathbf{Y}, \mathbf{z}^{(i-1)}, \boldsymbol{\Psi})$ is typically solved using Cholesky factorisation of $\hat{\mathbf{Q}}_k = \mathbf{Q}_k + \frac{1}{\sigma^2}(\mathbf{A}_k^{(i)})^\top \mathbf{A}_k^{(i)}$; however, this is not possible for large data sets. We instead use the following method, from (Papandreou & Yuille, 2011), which avoids the calculation of Cholesky factors entirely,

1. Generate $\mathbf{x} = \left((\mathbf{K}\mathbf{C}^{-1/2}) \otimes \mathbf{R}_{dk}\right)\mathbf{x}_1 + \frac{1}{\sigma}(\mathbf{A}_k^{(i)})^\top \mathbf{x}_2$ where $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors of independent $N(0,1)$ random variables.

2. Solve $\hat{\mathbf{Q}}_k \boldsymbol{\xi}_k = \mathbf{x} + \frac{1}{\sigma^2}(\mathbf{A}_k^{(i)})^\top(\mathbf{Y}^{(i)} - \mathbf{B}_k^{(i)}\boldsymbol{\beta})$.

Also here, the PCG method with a robust incomplete Cholesky preconditioner is used to solve the linear equation in the second step.

5 An application to magnetic resonance imaging

There are a number of possible applications to brain imaging that could be considered for this model class. However, in this section we only present a simple application to noise reduction. The MR image we analyze is a subset of data that previously has been used for CT substitute generation and is described in detail

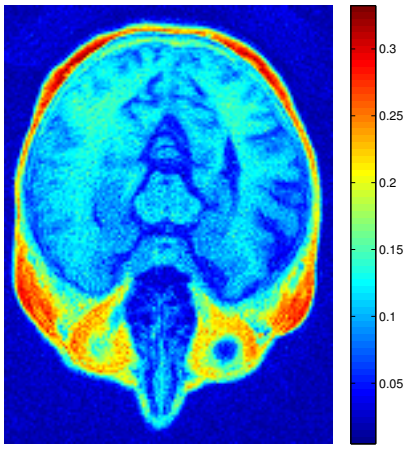


Figure 3: A noisy MR image of size 166×124 pixels.

in Johansson *et al.* (2011). The image is taken with a radial UTE sequence with a 10 degree flip angle, a repetition time of 6 ms, and an echo time of 0.07 ms. The UTE images were reconstructed to a matrix with $192 \times 192 \times 192$ voxels with isotropic resolution and a voxel size of 1.33 mm. For simplicity, we analyse only one slice of this data, which is of size 192×192 pixels. After removing parts of the slice that only contains areas outside the head, we obtain the image shown in Figure 3 which is of size 166×124 pixels.

As seen in the figure, the data is somewhat noisy and the goal is therefore use statistical techniques to reduce the noise in the image. As a first method, we use a standard latent Gaussian model, which can be described as the LGFM model in Section 3 with $K = 1$. The resulting estimate, $\hat{X}$, is shown in Figure 4 (a) and the kriging residuals, $\hat{X} - Y$, are shown in Figure 4 (b). If the model was correct, there should be no spatial structure in the residuals. However, we clearly see the contour of the head in the residuals, which means that this simple latent Gaussian model likely is insufficient for doing noise reduction of this image.

As an alternative to the latent Gaussian model, we fit a LGFM model with three mixture components. The reason for choosing three components is to keep the model simple while being able to separate the air outside the head and the bone from the other tissue types, as these two classes clearly stands out in the image. In order to keep the model simple, the MRF parameters α_k are fixed to

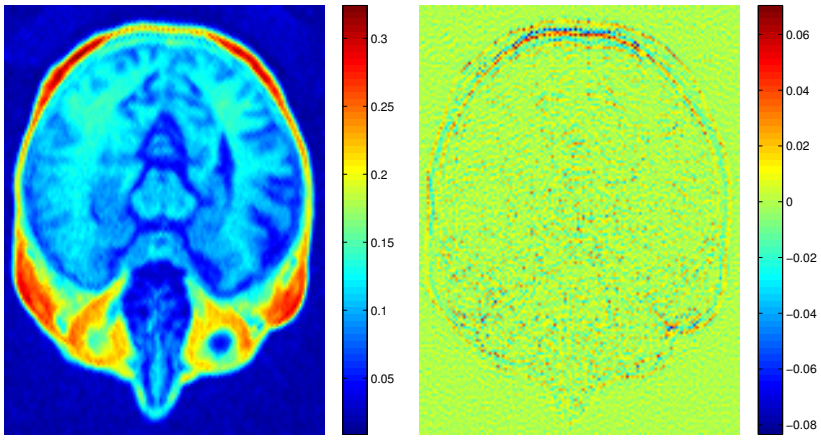
	LGM	$LGFM_1$	$LGFM_2$	$LGFM_3$
κ^2	0.0256	0.0550	0.0132	0.0005
σ^2	0.0396	0.0303	0.0303	0.0303
τ	2.8437	4.1980	194.14	0.0081
μ	1.3568	1.5297	0.3857	3.2251

Table 1: Parameter estimates for the latent Gaussian model (LGM) and the three mixture components of the LGFM model. The spatial dependency parameter β for MRF in the LGFM model was assumed to be the same for all classes, and was estimated to 2.73, and the prior parameters α_k were fixed to zero. The estimation was done on data standardized to have variance one.

zero and a common β parameter is assumed for all classes. Estimates of the other parameters are shown in Table 1, which also shows the parameter estimates for the latent Gaussian model as a reference.

Starting values for the LGFM estimation are obtained by first doing a classification of the data using a standard Gaussian mixture model and then estimating a latent Gaussian model for each class in the estimated mixture. The classification using the LGFM model is shown in Figure 5, Panel (a) and the corresponding classification is show in Panel (b). One should note that this classification is unsupervised and obtained as a byproduct while fitting the LGFM model, and it clearly finds the desired regions in the image. Panel (c) shows the difference between the LGFM estimate and the LGM estimate in Figure 4 (a), and one sees that the difference is quite large, especially near the tissue boundaries. Finally, Panel (d) shows the kriging residuals of the LGFM model in the same color scale as the residuals of the latent Gaussian model in Figure 4 (b), and although there is still some structure in the residuals, the result is much better.

Thus, the LGFM model performs much better than the latent Gaussian model, and one of the reasons for this is that the model parameters are allowed to very between the classes. This behavior could also be obtained by using a non-stationary latent Gaussian model, where the parameters are allowed to vary with space. However, the second important reason for the better behavior of the LGFM model, which is much harder to obtain using a non-stationary latent Gaussian model or an adaptive smoother is that the estimate for each class only uses data that is classified as beloning to that class. This allows for much sharper changes in the resulting estimate, and such behavior cannot be obtained in any



(a) (b)

Figure 4: Estimated MR image using a latent Gaussian model (a) and the kriging residuals (b).

simple way using an ordinary latent Gaussian model.

In this example, the main purpose was noise reduction and using the LGFM model we obtained a classification of the image as a byproduct. If the main objective was segmentation, a method worth mentioning is the popular adaptive segmentation method by Wells III *et al.* (1996). It is worth noting that this method fits into the general LGFM framework. In our notation, their model that is used for classification can be written as

$$\log(\mathbf{Y}_i) = \boldsymbol{\xi}_i + \sum_{k=1}^{K} z_{ik}\mathbf{G}_{ik} \tag{13}$$

where the field $\boldsymbol{\xi}$ is denoted a bias field and the second part is a standard gaussian mixture model (z is not a MRF in this model). This model can be reformulated as a transformed LGFM model, without measurement noise and with dependent mixture fields $\boldsymbol{\xi}_{ki} = \boldsymbol{\xi}_i + \mathbf{G}_{ki}$. An important difference that should be noted is that Wells III *et al.* (1996) assumes that the covariance matrix for $\boldsymbol{\xi}$ is a known band matrix and makes no attempts at estimating it, while we estimate the covariance function for each class.

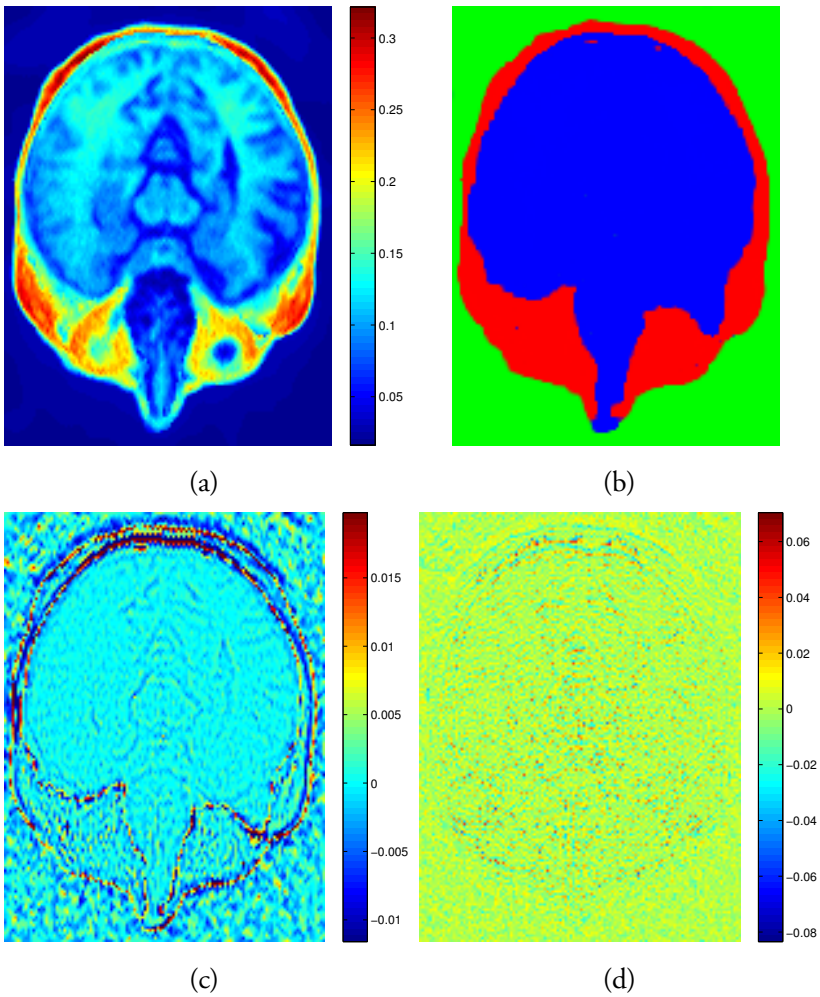


Figure 5: Estimated MR image using a LGFM model with three classes (a) and the corresponding classification (b). Panel (c) shows the difference between this estimate and the estimate using a latent Gaussian model, the color scale has been truncated to the middle 98% of the values to improve the visibility, which means that the largest differences are truncated in the color scale. Panel (d) shows the kriging residuals for the LGFM model, which shows much less spatial structure than the corresponding residuals for the latent Gaussian model. The color scale in Panel (d) has been set to match the color scale in Figure 4 (b).

6 Discussion

This work has introduced the class of LGFM models as well as a computationally efficient stochastic gradient parameter estimation method for the model class.

There are a number of directions in which this work can be extended. The methods were tested on a simple noise reduction application in brain imaging and we are working on more applications, such as substitute CT generation and land-use classification. We focused on a particular model here that is suitable for modeling of massive data sets on regular grids, but it would also be interesting to test the model for more typical geostatistical problems in continuous space. This would not require much work though the particular MRF model for the allocation process would have to be modified.

The proposed estimation method is not only useful for the LGFM models but also for regular MRF mixture models and latent Gaussian models. We have not shown any theoretical properties of the estimator here and to the authors knowledge, there are no applicable results available to show consistency of the estimator for the proposed model class. Comets & Gidas (1992) showed consistency for the maximum likelihood estimator for the MRF mixture models, but the consistency of the maximum likelihood estimator for the LGFM models, the pseudo likelihood estimators for the MRF mixture models, and the pseudo likelihood estimators for the LGFM models are to the authors knowledge unknown, and certainly something for further research.

Finally, the basic estimation method is straightforward to implement. However, we used several sophisticated techniques to reduce the computational cost of the estimation, which increases the complexity of the implementation. We are therefore working on a software package that implements these methods that will simplify the practical usage of the methods.

Acknowledgements

The authors are grateful to Adam Johansson for providing the MR data used in Section 5.

References

Ajiz, M. & Jennings, A. (1984). A robust incomplete choleski-conjugate gradient algorithm. *International Journal for Numerical Methods in Engineering* **20**, 949–966.

Anitescu, M., Chen, J. & Wang, L. (2012). A matrix-free approach for solving the parametric gaussian process maximum likelihood problem. *SIAM Journal on Scientific Computing* **34**, A240–A262.

Aune, E., Simpson, D. P. & Eidsvik, J. (2012). Parameter estimation in high dimensional gaussian distributions. *Statistics and Computing* , 1–17.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **36**, 192–225.

Bolin, D. & Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Ann. Appl. Statist.* **5**, 523–550.

Chiles, J.-P. & Delfiner, P. (1999). *Geostatistics, modeling spatial uncertainty*. Wiley Series in Probability and statistics.

Comets, F. & Gidas, B. (1992). Parameter estimation for gibbs distributions from partially observed data. *The Annals of Applied Probability* **2**, 142–170.

Cressie, N. (1991). *Statistics for spatial data*. John Wiley & Sons Ltd, New York, NY, USA.

Cressie, N. & Wikle, C. (2011). *Statistics for spatio-temporal data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **39**, 1–38.

Fernández, C. & Green, P. J. (2002). Modelling spatially correlated data via mixtures: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 805–826.

Guyon, X. (1995). *Random fields on a network: Modeling, statistics, and applications*. Graduate Texts in Mathematics. Springer.

Held, K., Kops, E. R., Krause, B. J., Wells III, W. M., Kikinis, R. & Muller-Gartner, H.-W. (1997). Markov random field segmentation of brain mr images. *Medical Imaging, IEEE Transactions on* **16**, 878–886.

Hu, X., Lindgren, D. S., Rue, H. *et al.* (2013). Multivariate gaussian random fields using systems of stochastic partial differential equations. *arXiv preprint arXiv:1307.1379* .

Hutchinson, M. (1990). A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* **19**, 433–450.

Johansson, A., Karlsson, M. & Nyholm, T. (2011). Ct substitute derived from mri sequences with ultrashort echo time. *Medical Physics* **38**, 2708.

Lange, K. (1995). A gradient algorithm locally equivalent to the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* , 425–437.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **73**, 423–498.

Papandreou, G. & Yuille, A. L. (2011). Efficient variational inference in large-scale bayesian compressed sensing. In *Computer vision workshops (iccv workshops), 2011 ieee international conference on*. IEEE, pp. 1332–1339.

Reynolds, D. A. & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on* **3**, 72–83.

Shapiro, A., Dentcheva, D. & Ruszczyński, A. P. (2009). *Lectures on stochastic programming: modeling and theory*, vol. 9. SIAM.

Stauffer, C. & Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer vision and pattern recognition, 1999. ieee computer society conference on.*, vol. 2. IEEE.

Stein, M. L. (1999). *Interpolation of spatial data: Some theory for Kriging*. Springer-Verlag, New York.

Stein, M. L., Chen, J. & Anitescu, M. (2013). Stochastic approximation of score functions for gaussian processes. *The Annals of Applied Statistics* **7**, 1162–1191.

Van Leemput, K., Maes, F., Vandermeulen, D. & Suetens, P. (1999). Automated model-based tissue classification of mr images of the brain. *Medical Imaging, IEEE Transactions on* **18**, 897–908.

Wells III, W. M., Grimson, W. E. L., Kikinis, R. & Jolesz, F. A. (1996). Adaptive segmentation of mri data. *Medical Imaging, IEEE Transactions on* **15**, 429–442.

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics* **18**, 125–139.

Zhang, Y., Brady, M. & Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on* **20**, 45–57.

A MRF gradients

Let $l = \log(PL(\Psi; \mathbf{z}, \mathbf{Y}))$, where PL is the pseudo likelihood of the MRF mixture model (10), the derivatives in the gradient are then given by

$$\begin{aligned}
\frac{\partial l}{\partial \alpha_k} &= \sum_i z_{ik} - \sum_i \frac{\exp(\alpha_k + \beta_k f_{ik})}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \\
\frac{\partial l}{\partial \beta_k} &= \sum_i z_{ik} f_{ik} - \sum_i \frac{\exp(\alpha_k + \beta_k f_{ik}) f_{ik}}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \\
\nabla_{\mu_k} l &= \sum_i z_{ik} \mathbf{Q}_{dk} (\mathbf{Y}_i - \mu_k) \\
\frac{\partial l}{\partial \eta_{kj}} &= \sum_i z_{ik} \left(\mathbb{I}_{diag} - \frac{1}{2} \left(Y_i - \mu_k \right)^\top \frac{\partial \mathbf{Q}_{dk}}{\partial \eta_{kj}} \left(Y_i - \mu_k \right) \right).
\end{aligned}$$

Here $\mathbb{I}_{diag}$ is one if η_{kj} is an element on the main diagonal of $\mathbf{R}_{dk}$ and zero otherwise. We have

$$\frac{\partial \mathbf{Q}_{dk}}{\partial \eta_{kj}} = \frac{\partial \mathbf{R}_{dk}^\top}{\partial \eta_{ki}} \mathbf{R}_{dk} + \mathbf{R}_{dk}^\top \frac{\partial \mathbf{R}_{dk}}{\partial \eta_{ki}} \tag{14}$$

where the derivative $\frac{\partial \mathbf{R}_{dk}}{\partial \eta_{ki}}$ is a matrix with all elements zero except the element that corresponds to η_{ki}. These expressions can be obtained with almost no extra cost while running the Gibbs sampler to sample $\mathbf{z}$. The derivatives needed to evaluate the scaling matrix $\mathbf{S}$ are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha_{k_1} \partial \alpha_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1})}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \right. \\
&\qquad \left. + \frac{\exp(\alpha_{k_1} + \beta_{k_1} f_{ik_1}) \exp(\alpha_{k_2} + \beta_{k_2} f_{ik_2})}{(\sum_l \exp(\alpha_l + \beta_l f_{il}))^2} \right) \\
\frac{\partial^2 l}{\partial \beta_{k_1} \partial \beta_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1}) f_{ik_1}^2}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \right. \\
&\qquad \left. + \frac{\exp(\alpha_{k_1} + \beta_{k_1} f_{ik_1}) \exp(\alpha_{k_2} + \beta_{k_2} f_{ik_2}) f_{ik_1} f_{ik_2}}{(\sum_l \exp(\alpha_l + \beta_l f_{il}))^2} \right) \\
\frac{\partial^2 l}{\partial \alpha_{k_1} \partial \beta_{k_2}} &= \sum_i \left(-\mathbb{I}_{k_1 = k_2} \frac{\exp(\alpha_{k_1} + \beta_k f_{ik_1}) f_{ik_1}}{\sum_l \exp(\alpha_l + \beta_l f_{il})} \right.
\end{aligned}$$

$$+\frac{\exp(\alpha_{k_1}+\beta_{k_1}f_{ik_1})\exp(\alpha_{k_2}+\beta_{k_2}f_{ik_2})f_{ik_1}}{(\sum_l \exp(\alpha_l+\beta_l f_{il}))^2}\Bigg)$$

where $\mathbb{I}_{k_1=k_2}$ controls that that factor is only included when $k_1 = k_2$. We also need the the derivatives of the parameters for the independent Gaussian distributions:

$$\begin{aligned}
\Delta_{\boldsymbol{\mu}_k} l &= -\sum_i z_{ik}\mathbf{Q}_{dk} \\
\frac{\partial^2 l}{\partial \eta_{kj_1}\partial \eta_{kj_2}} &= -\frac{1}{2}\sum_i z_{ik}\left(\mathbf{Y}_i-\boldsymbol{\mu}_k\right)^\top \frac{\partial^2 \mathbf{Q}_{dk}}{\partial \eta_{ljj_1}\partial \eta_{kj_2}}\left(\mathbf{Y}_i-\boldsymbol{\mu}_k\right) \\
\frac{\partial}{\partial \eta_{kj}}\nabla_{\boldsymbol{\mu}_k} l &= \sum_i z_{ik}\frac{\partial \mathbf{Q}_{dk}}{\partial \eta_{kj}}(\mathbf{Y}_i-\boldsymbol{\mu}_k)
\end{aligned}$$

where

$$\frac{\partial^2 \mathbf{Q}_{dk}}{\partial \eta_{ljj_1}\partial \eta_{kj_2}} = \mathbb{I}_{j_1=j_2}\mathbb{I}_{diag}\frac{\partial \mathbf{Q}_{dk}}{\partial \eta_{kj}} + \frac{\partial \mathbf{R}_{dk}^\top}{\partial \eta_{kj_1}}\frac{\partial \mathbf{R}_{dk}}{\partial \eta_{kj_2}} + \frac{\partial \mathbf{R}_{dk}^\top}{\partial \eta_{kj_2}}\frac{\partial \mathbf{R}_{dk}}{\partial \eta_{kj_2}} \tag{15}$$

Except for the derivatives with respect to $\boldsymbol{\mu}_k$, all these derivatives are also need for the estimation of the LGFM model.

B Gaussian gradients

Let $l = \log \pi(\mathbf{Y}, \boldsymbol{\xi}|\boldsymbol{\Psi})$, the expectation of the derivatives then needed to evaluate the gradients of the Gaussian likelihood $\left(\log \pi(\mathbf{Y}|\boldsymbol{\Psi})\right)$ are

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial l}{\partial \varkappa_0}|\mathbf{Y},\boldsymbol{\Psi}\right] =& nd(D/4-1) + de^{\varkappa_0}\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{C}\right) + \hat{\boldsymbol{\xi}}^\top \mathbf{Q}_d \otimes \tilde{\mathbf{Q}}_s\hat{\boldsymbol{\xi}} \\
&+ \operatorname{tr}\left(\mathbf{Q}_d \otimes \tilde{\mathbf{Q}}_s\hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial l}{\partial \eta_{lj}}|\mathbf{Y},\boldsymbol{\Psi}\right] =& \mathbb{I}_{diag} - \frac{1}{2}\hat{\boldsymbol{\xi}}^\top \frac{\partial \mathbf{Q}_d}{\partial \eta_{lj}} \otimes \mathbf{Q}_s\hat{\boldsymbol{\xi}} \\
&- \frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{Q}_d}{\partial \eta_{lj}} \otimes \mathbf{Q}_s\hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} l|\mathbf{Y},\boldsymbol{\Psi}\right] =& e^{-2\sigma_0}\mathbf{B}^\top\left(\mathbf{Y}-\mathbf{B}\boldsymbol{\beta}-\mathbf{A}\hat{\boldsymbol{\xi}}\right),
\end{aligned}$$

$$\mathsf{E}\left[\frac{\partial}{\partial \sigma_0} l|\mathbf{Y}, \mathbf{\Psi}\right] = e^{-2\sigma_0}\left(\left(\mathbf{Y} - \mathbf{B}\beta - \mathbf{A}\hat{\xi}\right)^\top \left(\mathbf{Y} - \mathbf{B}\beta - \mathbf{A}\hat{\xi}\right) + \operatorname{tr}\left(\mathbf{A}^\top \mathbf{A}\hat{\mathbf{Q}}^{-1}\right)\right) - m.$$

Here, $\frac{\partial \mathbf{Q}_d}{\partial \eta_j}$ is given by (14), $\hat{\xi}$ is the expected value of ξ given the current parameter estimates, $\hat{\xi} = \hat{\mathbf{Q}}^{-1}\frac{1}{\sigma^2}\mathbf{A}^\top(\mathbf{Y} - \mathbf{B}\beta)$, and $\tilde{\mathbf{Q}}_s = (1 - D/4)\mathbf{Q}_s - ce^{\varkappa_0}\mathbf{K}$. For the scaling **S**, we also need expectation of the second derivatives:

$$\begin{aligned}
\mathsf{E}\left[\frac{\partial^2}{\partial \varkappa_0^2} l|\mathbf{Y}, \mathbf{\Psi}\right] =& e^{\varkappa_0} d \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{C}\right) + e^{2\varkappa_0} d \operatorname{tr}\left(\mathbf{Q}_s^{-1}\mathbf{C}\right) + \\
& \hat{\xi}^\top \mathbf{Q}_d \otimes \frac{\partial \tilde{\mathbf{Q}}_s}{\partial \varkappa_0}\hat{\xi} + \operatorname{tr}\left(\mathbf{Q}_d \otimes \frac{\partial \tilde{\mathbf{Q}}_s}{\partial \varkappa_0}\hat{\mathbf{Q}}^{-1}\right), \\
\mathsf{E}\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} l|\mathbf{Y}, \mathbf{\Psi}\right] =& -\frac{1}{2}\hat{\xi}^\top \frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j} \otimes \mathbf{Q}_s \hat{\xi} + \\
& -\frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j} \otimes \mathbf{Q}_s \hat{\mathbf{Q}}^{-1}\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \varkappa_0 \partial \eta_j} l|\mathbf{Y}, \mathbf{\Psi}\right] =& \hat{\xi}^\top \frac{\partial \mathbf{Q}_d}{\partial \eta_j} \otimes \tilde{\mathbf{Q}}_s \hat{\xi} + \operatorname{tr}\left(\frac{\partial \mathbf{Q}_d}{\partial \vartheta_j} \otimes \tilde{\mathbf{Q}}_s \hat{\mathbf{Q}}^{-1}\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \beta^2} l|\mathbf{Y}, \mathbf{\Psi}\right] =& -e^{-2\sigma_0}\mathbf{B}^\top \mathbf{B} \\
\mathsf{E}\left[\frac{\partial^2}{\partial \sigma^2} l|\mathbf{Y}, \mathbf{\Psi}\right] =& -2e^{-2\sigma_0}\left(\left(\mathbf{Y} - \mathbf{B}\beta - \mathbf{A}\hat{\xi}\right)^\top \left(\mathbf{Y} - \mathbf{B}\beta - \mathbf{A}\hat{\xi}\right) \right. \\
& \left. + \operatorname{tr}\left(\mathbf{A}^\top \mathbf{A}\hat{\mathbf{Q}}^{-1}\right)\right) \\
\mathsf{E}\left[\frac{\partial^2}{\partial \beta \partial \sigma_0} l|\mathbf{Y}, \mathbf{\Psi}\right] =& -2\frac{\partial}{\partial \beta} l.
\end{aligned}$$

Here $\frac{\partial^2 \mathbf{Q}_d}{\partial \eta_i \partial \eta_j}$ is given by (15) and

$$\frac{\partial \tilde{\mathbf{Q}}_s}{\partial \varkappa_0} = -\frac{(D-4)^2}{8}\mathbf{Q}_s + c(3 - D)e^{\varkappa_0}\mathbf{K} - ce^{2\varkappa_0}\mathbf{C}.$$

E

Paper E

Slepian models for moving averages driven by a non-Gaussian noise

KRZYSZTOF PODGÓRSKI[1], IGOR RYCHLIK[2], JONAS WALLIN[3]

[1] *Department of Statistics, Lund University, Sweden*
[2] *Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden*
[3] *Centre for Mathematical Sciences, Lund University, Lund, Sweden*

Abstract

In analysis of extreme behavior of a time series or a stochastic process the Rice formula is often used to obtain the distribution of the process at the instants of high level crossings. For the purpose of simulation or analysis it is convenient to have a Slepian model corresponding to the behavior of the original process sampled at the level crossings. Here a Slepian model is understood as any explicitly defined stochastic process that is distributed according to the crossing level biased sampling distribution. The original Slepian model developed for a stationary Gaussian process is very helpful in analyzing behavior of the process at extreme levels. Here, Slepian models are derived describing the distributional form of a moving average driven by a non-Gaussian noise as observed at level crossings. Our leading non-Gaussian examples are moving averages driven by a Laplace noise. A method of sampling from the corresponding biased sampling distribution of the underlying gamma process is obtained. This is used for efficient simulation of the behavior of a non-Gaussian process at the extreme level crossing. It is observed that the asymptotic behavior of the process at high level crossings that is fundamentally different from that in the Gaussian case.

Key words: Rice formula, level crossings, generalized Laplace distribution, moving average process, extreme episodes, biased sampling distribution.

1 Introduction

In the physical world, a random function is often described as a sequence of local maxima or minima, constituting a series of random waves. In fact, not only the visual impression of the process but also many technologically important implications in such fields as metal fatigue caused by random vibrations, failure caused by excess load on a construction, etc., depend on the character of the process represented in such a random wave form. The basic objects in this theory are level crossings and local extremes, see Podgórski *et al.* (2000) for computation of various crossing distributions, Baxevani *et al.* (2003) for spatial wave characteristics, and Ditlevsen (1985) for an overview of other engineering applications.

A Slepian model is a random function representation of the conditional behavior of a stochastic process after events defined by level or curve crossings. In general, a Slepian model contains one regression term with random coefficients which describe the dependence on initial conditions such as the slope at the crossing, the value of the process at a predetermined point, etc, and one residual term, which describes the deviations from the path set out by the initial conditions. In its classical form, such a model was first introduced in Slepian (1963) to describe the behavior of a stationary Gaussian process after a zero crossing.

The model found applications in more theoretical studies of various asymptotic sample path properties of a Gaussian random process, or a function of a vector valued Gaussian process, see Kac & Slepian (1959), Aronowich & Adler (1988), or Lindgren (1989). Considerable work has been done on studying sample properties of Gaussian or related fields around high local maximum or level set, see Wilson & Adler (1982) and, for more recent work, Azaïs & Wschebor (2009).

Typically, the Slepian model is defined for ergodic processes when the distribution of the model coincides with the long-run empirical distribution of the stochastic model. However, the Slepian model was also defined for a non-stationary Gaussian process to study properties of the process under conditioning that local maximum occurs at time zero, see Gadrich & Adler (1993) and Grigoriu (1989) for an engineering application. It is worth to mention here that in the approach presented in this paper, we analyze non-Gaussian models obtained by a random distortion of the Brownian motion through conditioning on the distortion which leads to (conditionally) non-stationary Gaussian process and through this our work connects with Gadrich & Adler (1993).

In many practical situations, the assumption of Gaussianity is not supported

by empirical data and therefore derivation of a Slepian model for non-Gaussian processes is desirable. This need has driven growing interest in studying level crossing distributions for non-Gaussian models, for example, see Adler *et al.* (2013) for results on the high level crossings and van de Lindt & Niedzwecki (2005) for an example of practical context in which a Slepian model for data exhibiting non-Gaussian features is of interest. In this paper, we present an approach to obtain an effective Slepian models for a class of non-Gaussian models driven by a Laplace motion, i.e. a non-Gaussian Lévy motion obtained by subordination of Brownian motion by a gamma process. This class has proven to be sufficiently flexible to account for most non-Gaussian features observed in practical applications and some work has been done on the level crossing distributions derived from a generalized Rice formula applied to these processes, see Åberg & Podgórski (2010), Åberg *et al.* (2009), Galtier (2011).

The focus of this work is two-fold, firstly, we propose derivation general Slepian models by obtaining Slepian models of the noise that is driving the considered models, secondly, we show how conditioning on a variable or process can help in derivation of a convenient Slepian model for a non-Gaussian model. The novelity of the approach is its focus on the Slepian models of the noise. An alternative approach to building a Slepian model would be through a hierarichical approach to which one could employ a non-stationary Slepian models as discussed in Gadrich & Adler (1993). We prefer to consider a Slepian model of noise for which we found a convenient simulation method through a Gibbs sampler. One advantage of having a Slepian noise is a possibility of simultaneous studies of various random functionals of such a noise without necessity of constructing a separate Slepian model for each of the functionals. These benefits are illustrated by examples and numerical studies. Our interest in simulations of Slepian models is paralleling applied engineering papers on this subject. They are useful, in particular, to study non-linear dynamical systems where such Slepian models can be considered as input to the system to study their responses at particular crossing events. For such application our approach is more direct than the one presented in Adler *et al.* (2013).

2 Preliminaries

We consider a stationary random process X having a.s. absolutely continuous samples and such that the joint probability distribution function (pdf) $f_{X,\dot{X}}$ of

$X(0), \dot{X}(0)$ exists. For $u \in \mathbb{R}$, the u-level upcrossing set within interval $[0,1]$ is defined as

$$\mathcal{C}(u) = \{s \in [0,1] : X(s) = u, \dot{X}(s) > u\}.$$

Let $N(u)$ be the number of elements in $\mathcal{C}(u)$. For a properly defined statement A on trajectories of another stationary stochastic process Y, define $N(A|u)$ to be the number of $s \in \mathcal{C}(u)$ such that $Y(s+\cdot) \in A$. The generalized Rice's formula yields

$$E[N(u)] = \int_0^{+\infty} z f_{\dot{X},X}(z,u)\, dz,$$

where X denotes $X(0)$ and $\dot{X}$ denotes $\dot{X}(0)$. Equivalently

$$E[N(u)] = \int \int_0^{+\infty} z f_{\dot{X},X|K}(z,u|k) f_K(k)\, dz\, dk,$$

where conditioning on the random variable or vector K is used to simplify evaluation of the integral. Here and in what follows whenever the limits of integration are not shown in the notation they are understood to be over the entire set of possible values of the corresponding variable. The focus of this paper is on the Laplace moving average (LMA) processes for which K is a certain, possibly vector valued, functional of the gamma process that serves as the subordinator in the representation of the Laplace motion as a subordinated Brownian motion, Kotz *et al.* (2001).

Similarly, one can consider

$$E[N(A|u)] = \int_0^{+\infty} P(Y \in A|\dot{X} = z, X = u) \cdot z f_{\dot{X},X}(z,u)\, dz =$$
$$= \int \int_0^{+\infty} P(Y \in A|\dot{X} = z, X = u, K = k) \cdot z f_{\dot{X},X|K}(z,u|k) f_K(k)\, dz\, dk$$

and use this to evaluate the u-level upcrossing distribution P^u of Y, see, for example, Zähle (1984) for derivation of this formula for a general class of stochastic processes. The upcrossing distribution P_u is defined as the ratio of the average number of the u-upcrossings at which a trajectory event occurs to the average number of all u-uppcrosings, i.e. for a trajectory event A we consider

$$P^u(A) = \frac{E[N(A|u)]}{E[N(u)]}. \tag{1}$$

Consequently, one has the following representation of u-level upcrossing distributions involving the conditioning on K:

$$P^u(A) = \frac{\int\int_0^{+\infty} P(Y \in A|\dot{X} = z, X = u, K = k) \cdot z f_{\dot{X},X|K}(z,u|k) f_K(k)\, dz\, dk}{\int\int_0^{+\infty} z f_{\dot{X},X|K}(z,u|k) f_K(k)\, dz\, dk} =$$

$$= \frac{\int\int_0^{+\infty} P(Y \in A|\dot{X} = z, X = u, K = k) \cdot z f_{\dot{X}|K,X}(z|k,u) f_{K|X}(k|u)\, dz\, dk}{\int\int_0^{+\infty} z f_{\dot{X}|K,X}(z|k,u) f_{K|X}(k|u)\, dz\, dk}. \qquad (2)$$

A stochastic process Y_u such that its finite dimensional distributions correspond to these given by the upcrossing distribution is referred to as a Slepian model of Y at the u-up-crossings, i.e. for each measurable A in the space of trajectories, Y_u satisfies

$$P(Y_u \in A) = P^u(A).$$

If the Slepian model can be expressed in an explicit form, it can be used for deriving approximations for probabilities of interest as well as simulating trajectories interpreted as sample of the original process observed at instants of the u-level upcrossings. It can also help to analyze asymptotic behavior of the process crossing high level and thus providing an insight into behavior of the process at extremal episodes.

Example 1. One can take $Y = X$ and then a Slepian process X_u describes behavior of X at its own up-crossings of u.

Example 2. Another case is to take $Y = K$ for which a Slepian model K_u has distribution given by the density

$$f_{K_u}(k) \propto \int_0^{+\infty} z f_{\dot{X}|K,X}(z|k,u) f_{K|X}(k|u)\, dz.$$

Example 3. A joint Slepian model for $\dot{X}$ and K is a random vector with the distribution given by

$$f_{K_u,\dot{X}_u}(k,z) \propto z f_{\dot{X}|K,X}(z|k,u) f_{K|X}(k|u).$$

We observe that the distribution of K_u given $\dot{X}_u = z$ is the same as the distribution of K given $\dot{X} = z$ and $X = u$.

Generalizing Example 3 one can conveniently write a scheme for obtaining a Slepian model for a process Y. Namely, let a process $Y(\cdot|k,z,u)$ has distribution equal to that of Y conditionally on $K=k$, $\dot{X}=z$ and $X=u$. Then

$$Y_u(\cdot) = Y(\cdot|K_u, \dot{X}_u(0), u). \tag{3}$$

In this note we discuss a Slepian model for the Laplace moving average (LMA)

$$X(t) = \int g(s-t)\, dL(s)$$

obtained by conditioning on the random time change process $K(t)$ (subordinator) in the representation of the Laplace motion $L(t) = B(K(t))$, where B is the Brownian motion and K is a gamma process, see Kotz *et al.* (2001). As explained above the key to deriving such a Slepian model is obtaining the biased sample distribution of K_u, i.e. that of the gamma process observed at the u level up-crossings of the moving average process X. Our approach provides an alternative to computing crossing level distributions as compared to the results presented in Galtier (2011), Åberg & Podgórski (2010), Åberg *et al.* (2009), where the approximation to the joint distribution of the process and derivative at zero were used for the purpose. The benefits of considering Slepian models is that they provide a unified frame for handling level crossing distributions for a variety of the variables and functionals of stochastic processes.

To achieve this, we proceed as follows. Firstly, we consider a representation $X(t|k,z,u)$ of $X(t)$ conditionally that $K(\cdot)=k(\cdot)$, $X(0)=u$ and $\dot{X}(0)=z$. Here $k(\cdot)$ stands for a realization of the process $K(t)$, $t\in\mathbb{R}$. Secondly, we obtain an effective way of sampling from $K_u(t)$ and $\dot{X}_u(0)$. Then finally, we obtain a Slepian model through $X_u(t) = X(t|K_u(\cdot), \dot{X}_u, u)$.

Practically and for simulation purposes, the actual conditioning will be with respect to a vector valued K having iid Gamma coordinates, which serves as a discretized version of the subordinator K. A Gibbs sampler will be applied to obtain distributions from the so structured Slepian model. This main contribution is presented in Section 5.2. First however, in the next section, we study level crossing distributions of a simpler although related non-Gaussian process obtained by random scaling of a Gaussian process. This can be viewed also as the moving average process with respect to a random time model $B(K\cdot t) = \sqrt{K}B(t)$, where K is a numerical random variable independent of B. In the concluding section of the paper we study the behavior of the Laplace moving average process at the high

level crossings, while the appendix contains all technical details of the presented results.

3 Random scaling model

Here, we discuss the Slepian model for a non-ergodic and non-Gaussian stationary process. Recall that for a stationary Gaussian process Z with variance one and variance of its derivative also equal to one the Slepian model process Z_u around u-upcrossing of Z is given by

$$Z_u(t) = u\,r(t) - R\dot{r}(t) + \Delta(t) = u\,r(t) - \dot{Z}_u(0)\,\dot{r}(t) + \Delta(t),\ t \in \mathbb{R}, \quad (4)$$

where r be the covariance function of Z, R is a standard Rayleigh variable independent from a non-stationary Gaussian process Δ having covariance

$$r(t,s) = r(t-s) - r(t)r(s) - r'(t)r'(s).$$

See Leadbetter *et al.* (1983) for further details.

The above model can be viewed as the one obtained from (3). Namely, it is easy to verify that the process $Z(t|z,u) = u\,r(t) + z\,\dot{r}(t) + \Delta(t)$ has the same distribution as $Z(t)$ conditionally on $Z(0) = u$ and $\dot{Z}(0) = z$. Moreover, for the Gaussian case $\dot{Z}_u$ has the Rayleigh distribution thus following (3), i.e. considering $Z(t|\dot{Z}_u, u)$ yields the right hand side of (4).

Let us consider first a non-random scaling of Z, i.e. a process $X(t) = \sqrt{k}\,Z(t)$, $t \in \mathbb{R}$, where $k > 0$ is a deterministic constant. For X at its own up-crossings of u the following process clearly defines a Slepian model

$$X_u(t) = u\,r(t) - \sqrt{k}Rr'(t) + \sqrt{k}\Delta(t) = u\,r(t) - \dot{X}_u(0)r'(t) + \sqrt{k}\Delta(t).$$

When the non-random scaling $\sqrt{k}$ is replaced by a random one and we consider the process

$$X(t) = \sqrt{K}\,Z(t), \quad (5)$$

where random K is independent of a process X, a simple analogy would suggest that the following process defines a Slepian model for X:

$$X_u(t) = u\,r(t) - \sqrt{K}R\,\dot{r}(t) + \sqrt{K}\Delta(t). \quad (6)$$

However this is not the case because $X(t)$ conditionally on $(K = k, \dot{X} = z, X = u)$ is represented by $u\,r(t) - z\,\dot{r}(t) + \sqrt{k}\Delta(t)$ and in this (k, z) has to be replaced by the Slepian model $(K_u, \dot{X}_u)$. We note that $\dot{X}_u \stackrel{d}{=} \sqrt{K_u}R$, where R is a Rayleigh random variable independent of everything else, and a random variable K_u is 'biased' to account for the fact that the behavior observed at u up-crossings for specific u makes certain scalings more likely than other – the phenomenon frequently referred to as 'sampling bias'.

As shown in the Appendix A, a Slepian model for the pair of variables $(K, \dot{X})$ is given by $(K_u, \dot{X}_u) = (K_u, \sqrt{K_u}R)$, where R is having the Rayleigh distribution and is independent of K_u with the distribution given by

$$f_{K_u}(k) = c_u \cdot e^{-\frac{u^2}{2k}} f_K(k).$$

The distribution of K_u represents the biased sampling distribution of the original random scaling K when the sampling is at the up-crossing of the level u.

Following (3), a Slepian model for X_u is given by

$$X_u(t) = u\,r(t) - \sqrt{K_u}R\ \dot{r}(t) + \sqrt{K_u}\Delta(t) = u\,r(t) - \dot{X}_u(0)\,\dot{r}(t) + \sqrt{K_u}\Delta(t), \quad (7)$$

where $\Delta(t)$ is the non-stationary Gaussian process described above independent of K_u and R. Thus the Slepian model given in (7) corresponds to the application of the conditioning in the Rice formula as presented in (2).

If for the scaling K we assume a gamma distribution, with the shape and scale parameters p and $2/a$, respectively, then its biased sampling distribution is given by

$$f_{K_u}(k) = \frac{a^{p/2}}{2u^p K_p(\sqrt{a}u)} \cdot k^{p-1} \exp\left(-\frac{ak + u^2/k}{2}\right),$$

where the modified Bessel function of the third kind (sometimes also referred to as the second kind) with index p is defined as

$$K_p(u) = \frac{1}{2}\left(\frac{u}{2}\right)^p \int_0^\infty t^{-p-1} \exp\left(-t - \frac{u^2}{4t}\right) dt, \quad u > 0.$$

This means that it belongs to the class of the generalized inverse Gaussian (GIG) distributions, see the Appendix B for basic properties and notation. More generally, the GIG distributions have a convenient invariance property: if we consider the scaling K that is $GIG(p, a, b)$, then its biased sampling distribution is $GIG(p, a, b + u^2)$.

Finally, the distribution of K_u given $\dot{X}_u = z$ is $GIG(p, a, b + u^2 + z^2)$ and the distribution of $\dot{X}_u$ is

$$f_{\dot{X}_u}(z) \propto z \left(\sqrt{b + u^2 + z^2}\right)^p K_p \left(\sqrt{a(b + u^2 + z^2)}\right), \; z > 0. \tag{8}$$

Consequently, we can write an alternative form of the Slepian model

$$X_u(t) = u\, r(t) - \dot{X}_u(0)\, \dot{r}(t) + \sqrt{\tilde{K}_u}\Delta(t), \tag{9}$$

where $\dot{X}_u$ has the distribution given by (8) and $\tilde{K}_u$ is sampled from $GIG(p, a, b + u^2 + \dot{X}_u^2)$, i.e. $\tilde{K}_u$ is a variable sampled from the distribution of K_u given that $\dot{X}_u(0) = z$.

4 Non-ergodicity effect

The random scaling process (5) is not ergodic and thus the derived Slepian model $X_u(t)$ does not represent the distribution of the process as observed at the level up-crossings based on an individual trajectory. This due to the fact that the long sampling distribution for a non-ergodic case is random, i.e. it varies randomly from trajectory to trajectory. Nevertheless, the Slepian model distribution P^u given in (2) still represents the ratio of two expected values to which the following interpretation can be attached.

Consider S independent trajectories of X and let $N_{T,i}(A|u)$ stands for the number of upcrossings in $[0, T]$ marked by the property A for the i-th trajectory while $N_{T,i}(u)$ denotes the total number of upcrossings in $[0, T]$ for this trajectory. With this notation, the proportion of u-upcrossings with property A in interval $[0, T]$ among all u-upcrossings is given by

$$\hat{P}_{T,S}(A|u) = \frac{\sum_{i=1}^{S} N_{T,i}(A|u)/N_{T,i}(u)}{S}.$$

By the pointwise ergodic theorem, which is valid as long as the underlying process is stationary, for each i, $N_{T,i}(A|u)$ and $N_{T,i}(u)$ converge when T increases without bound almost surely to some random variables $\bar{N}_i(A|u)$ and $\bar{N}_i(u)$, respectively. These variables for different i's are independent identically distributed and satisfy

$$E(\bar{N}_i(A|u)) = E(N(A|u)),$$

$$E(\bar{N}_i(u)) = E(N(u)),$$

while, in general,

$$E\left(\frac{\bar{N}_i(A|u)}{\bar{N}_i(u)}\right) \neq \frac{E(N(A|u))}{E(N(u))}.$$

By the Law of Large Numbers

$$\lim_{S\to\infty}\lim_{T\to\infty} \hat{P}_{T,S}(A|u) = E\left(\frac{\bar{N}_1(A|u)}{\bar{N}_1(u)}\right) \tag{10}$$

and

$$\lim_{T\to\infty}\lim_{S\to\infty} \hat{P}_{T,S}(A|u) = \frac{E\left(N(A|u)\right)}{E\left(N(u)\right)}. \tag{11}$$

Due to the lack of ergodicity, the two limits are not the same while the crossing level distribution $P^u(A)$ is defined as the latter one. Thus the Slepian model approximates the proportion of the corresponding number of crossings if they are looked for many independent trajectories over a long but fixed time interval $[0, T]$, where T is chosen *the same* for all trajectories.

On the other hand, it should be remembered that, except for the ergodic case, the Slepian model that represents $P^u(A)$ can not be interpreted as the averaged individual trajectory up-crossings distributions because this would be rather equal to $E\left(\bar{N}_1(A|u)/\bar{N}_1(u)\right)$. The latter is in fact the averaged values of crossing distributions as presented by the model given in (6). In this interpretation and using (10), the model (6) provides the distribution at level crossings observed at independent and complete, i.e. not restricted to a finite interval, sample trajectories.

The moving average model that is discussed in the next section is ergodic and thus these issues with Slepian model interpretation do not longer apply. In fact, in the moving average case, $\bar{N}_i(A|u)$ and $\bar{N}_i(u)$ are non-random and equal to their respective expected values, i.e. $\bar{N}_i(A|u) = E(N(A|u))$ and $\bar{N}_i(u) = E(N(u))$ so that the limits in (10) and (11) becomes the same.

5 Slepian model for the noise

5.1 Slepian noise model for Gaussian moving average

A moving average process is a convolution of a kernel function g, say, with a infinitesimal "white noise" process having variance equal to the discretization step,

say ds. Throughout the paper we assume normalization of the process in its value and its argument so that the variances of the process and of its derivative are equal to one or, equivalently, $\int g^2 = \int \dot{g}^2 = 1$. Here and in what follows, $\int f$ stands for the Lebesgue integral of f, i.e. $\int f = \int f(t)\,dt$

The Gaussian moving average (GMA) is given by

$$X(t) = \int_{-\infty}^{\infty} g(s-t)\,dB(s) \tag{12}$$

and its derivative $\dot{X}$ is given as the moving average with $\dot{g}$ as the kernel. Consider a fixed level $u \in \mathbb{R}$ and the probability distribution P^u on events A in the space of real continuous functions on $\mathbb{R}$ given in (1). For a stationary Gaussian process, a Slepian model is presented in (4). However for the purposes of this presentation we derive another while equivalent Slepian model that explicitly use the moving average form of the underlying process.

We first ask for a Slepian model $dB_u(x)$ for the noise $dB(x)$ at the crossing levels u of X. As argued in the Appendix A.2, the biased sampling distribution of $dB(x)$ is represented by the distribution of the following stochastic process $B_u(t)$, $t \in \mathbb{R}$:

$$B_u(t) = u\int_0^t g + R\int_0^t \dot{g} - \int_0^t g \cdot \int g\,dB - \int_0^t \dot{g}\cdot\int \dot{g}\,dB + B(t),\ \ t\in\mathbb{R}, \tag{13}$$

where random variable R has the Rayleigh distribution and is independent of $dB(t)$, while $B_u(t)$ is understood as a random measure of $[0,t]$, (if $t < 0$ the measure is understood as the minus measure of $[t,0]$). From this representation we can clearly distinguish three component of the Slepian model for the noise: the level and kernel dependent non-random component

$$F_{u,g}(t) = u\int_0^t g,$$

the kernel only dependent random component

$$G_g(t) = \left(R - \int \dot{g}\,dB\right)\cdot\int_0^t \dot{g} - \int g\,dB\cdot\int_0^t g,$$

and purely random noise represented by Brownian motion $B(t)$. We note that G_g and B are stochastically dependent and G_g conditionally on B is a linear combination of non-random functions with one random coefficient distributed according to a Rayleigh distribution.

Example 4 (Gaussian kernel)**.** Let us consider the (normalized) kernel $g(t) = \sqrt{2}e^{-t^2}/(2\pi)^{1/4}$, $t \in \mathbb{R}$. Direct calculations lead to the following form of the Slepian model

$$B_u(t) = F_{u,g}(t) + G_g(t) + B(t),$$

where

$$\begin{aligned} F_{u,g}(t) =& \sqrt[4]{2\pi}\, u\, \Phi_0\left(\sqrt{2}t\right), \\ G_g(t) =& \left(\sqrt[4]{\frac{2}{\pi}}\, R + \sqrt{\frac{8}{\pi}} \int s e^{-s^2}\, dB(u)\right) \cdot \operatorname{sgn}(t) \cdot \\ & \left(e^{-t^2} - 1\right) - \sqrt{2} \int e^{-s^2}\, dB(s) \cdot \Phi_0\left(\sqrt{2}t\right), \end{aligned}$$

where $\Phi_0(s) = (2\pi)^{-1/2} \int_0^s e^{-u^2/2}\, du$.

In Figure 1, we show simulations of samples from this Slepian model for the motion for different levels u and compare them with corresponding samples from a regular Brownian motion. From them we observe that the behavior of $B_u(t)$ depends on the value of a level u. In particular, for a high level u the main contribution to B_u comes from the deterministic part.

There are several benefits of looking at the level crossing distributions through the Slepian model of the noise. Having this biased sampling representation of the noise, the Slepian model for any process that can be obtained as a functional of the Brownian motion B is simply given by replacing B with B_u. Decomposition into three components: level depending, kernel depending, and noise, allows separate studies of different aspect of process behavior at the crossing levels. This is particularly beneficiary if the process under consideration is a linear functional of the noise. More precisely, consider a vector of stochastic processes $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_n(t))$, $t \in \mathbb{R}$ such that they arise as a result of some functionals acting on increments dB of a Brownian motion:

$$Y_i(t) = H_i(t, dB),\ i = 1, \ldots, n,$$

then the joint Slepian model $\mathbf{Y}_u(t)$ for $\mathbf{Y}(t)$ at the instants when the moving average process $X(t)$ up-crosses level u is obtained by considering

$$Y_{u,i}(t) = H_i(t, dB_u),\ i = 1, \ldots, n.$$

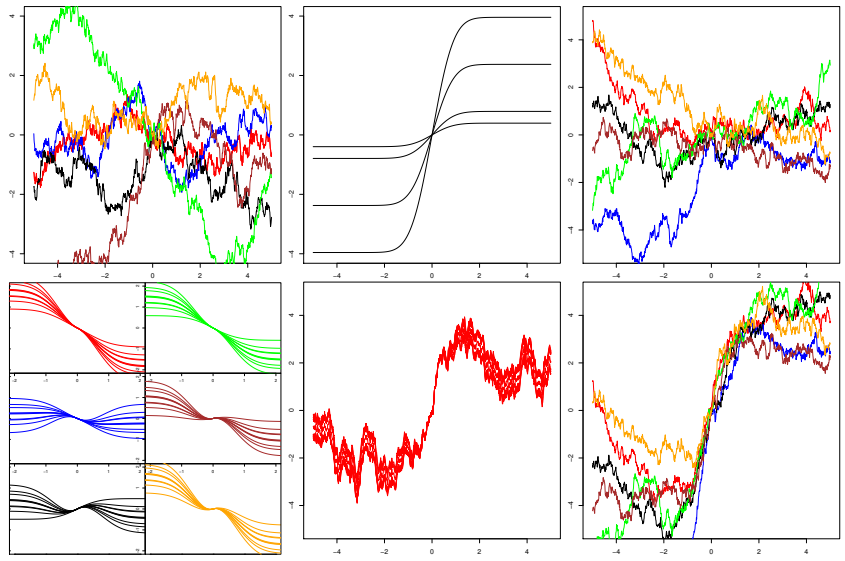


Figure 1: *Top-Left:* Six samples from a regular Brownian motion that were used in computing samples from the Slepian model of $B_u(t)$. *Bottom-Left:* Ten $G_g(t)$'s computed for each of the six Brownian motion samples using ten randomly sampled values R from Rayleigh distributions (the same R's has been used for all six plots). *Middle-Top:* Level dependent deterministic part of the Slepian model obtained for for four levels: $u = 0.5, 1, 3, 5$. *Middle-Bottom:* Samples from the Slepian model for the level $u = 5$, a single path of Brownian motion, and ten different values of the Rayleigh variable. *Right:* Six samples of the Slepian model for B_u corresponding to the samples of Brownian motion. Crossing levels: $u = 0.5$ (top) and $u = 5$ (bottom). A single value for Rayleigh variable is used for all samples.

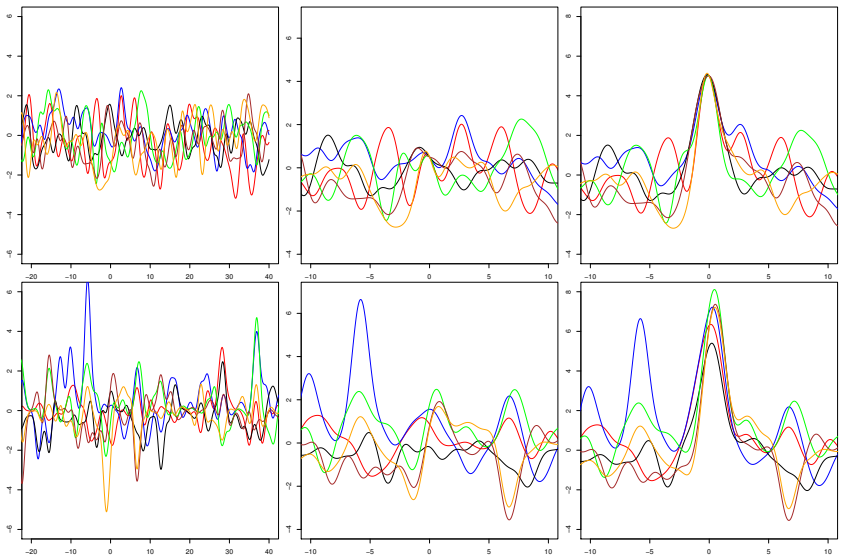


Figure 2: *Left:* Six samples from (X, Y): the Gaussian moving average X – *(top)* and corresponding samples form the Laplace moving average Y – *(bottom)*. Samples are based on five samples of the Brownian motion and a single sample of the gamma process that is used for samples related to Y process. *Middle:* Samples from the joint Slepian model at the crossings of X at level: $u = 0.5$. *Right:* Analogous samples at the crossings of level $u = 5$.

In particular, if functionals $H_i(t, B)$ are linear in B, we obtain joint decomposition

$$Y_{i,u}(t) = u \cdot H_i(t, g\ dt) + \left(R - \int \dot{g} dB\right) \cdot H_i(t, \dot{g}\ dt) + \\ - \int g dB \cdot H_i(t, g\ dt) + Y_i(t),\ \ i = 1, \ldots, n. \tag{14}$$

Example 5. To illustrate the convenience of the approach to level up-crossing distributions through the Slepian model of the underlying noise process, we consider a pair of linear functionals of dB, $\mathbf{Y} = (Y_1, Y_2)$, defined as follows.

The first component $Y_1(t)$, $t \in \mathbb{R}$, is the filtered original process $X(t)$ by

means of a filter $h(t)$, so that

$$Y_1(t) = \int h(s-t)\, dX(s) = \int h * g(s-t)\, dB(s).$$

The second process arises from linear scheme that alter Gaussian distribution of the moving average process. Namely, we consider the moving average driven by a Lévy motion build upon the Laplace distribution – the Laplace motion. The Laplace motion can be obtained through subordination of the Brownian motion by the gamma motion. For a kernel f and the Lévy process K such that $K(1)$ has the gamma distribution with shape τ and scale $1/\tau$ (for negative t, the process $-K(t)$ is an independent copy of $K(t), t \geq 0$), we define the Laplace moving average

$$Y_2(t) = \int f(s-t)\, dB \circ K(s).$$

The direct approach to the joint distribution of (Y_1, Y_2) at up-crossings of X would require analysis of the joint distribution of (Y_1, Y_2) together with the distribution of $X(0)$ and $\dot{X}(0)$. This is not straightforward due to non-Gaussianity of Y_2. However, our approach through the Slepian model of dB as given in (14) yields

$$\begin{aligned}
Y_{1u}(t) =& u \cdot h * r(t) + \left(R - \int \dot{g} dB\right) \cdot h * g * \tilde{\dot{g}}(t) + \\
&- \int g dB \cdot h * r(t) + Y_1(t), \\
Y_{2u}(t) =& u \cdot \int f(s-t)\, dG \circ K(s) + \\
&+ \left(R - \int \dot{g} dB\right) \cdot \int f(s-t)\, dg \circ K(s) + \\
&- \int g dB \cdot \int f(s-t)\, dG \circ K(s) + Y_2(t),
\end{aligned}$$

where $G(t) = \int_0^t g$, $\tilde{\dot{g}}(t) = \dot{g}(-t)$, and $r = g * \tilde{g}$ is the covariance of X. The obtained decomposition clearly reveal more complex dependence structure between processes.

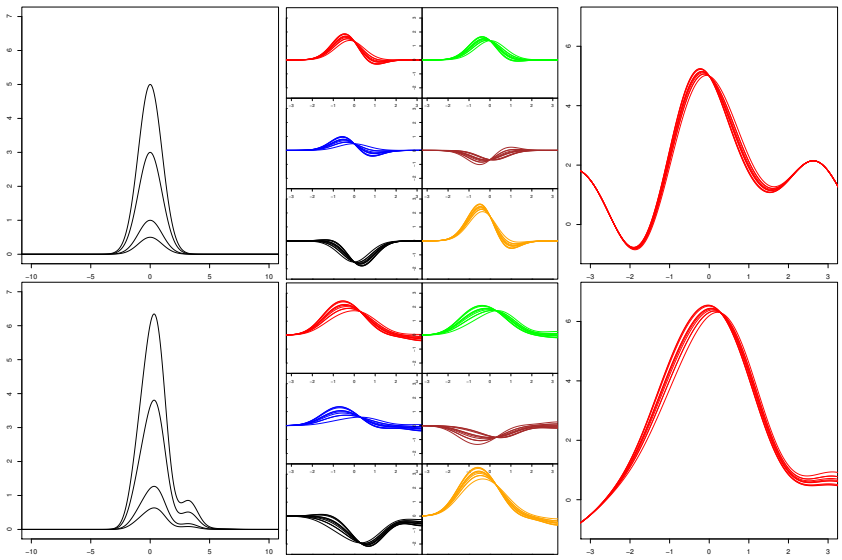


Figure 3: The joint Slepian model for Gaussian (*top*) and Laplace (*bottom*) moving averages. The Laplace case have random dependence on the underlying gamma motion K – here only single sample from this process has been taken for all graphs. *Left:* Level dependent components in the Slepian model, $u = 0.5, 1, 3, 5$. *Middle:* The kernel dependent component for ten randomly chosen values of the Rayleigh variable R). *Right:* Effect of Rayleigh variable on a single sample from the joint Slepian model at level $u = 5$ – randomly sampled ten values of R.

For illustration, we take X as in Example 4 and consider $Y_1 = X$, while $Y_2 = Y$, where

$$Y(t) = \int g(s-t)\, dB \circ K(s), \tag{15}$$

which could be viewed as a modified X by random distortion of time through the gamma process K. We have the following formulas

$$X_u(t) = u \cdot e^{-t^2/2} - \left(R + 2\left(\frac{2}{\pi}\right)^{1/4} \int s e^{-s^2}\, dB(s) \right) \cdot t e^{-t^2/2} +$$

$$
\begin{aligned}
&- \left(\frac{2}{\pi}\right)^{1/4} \int e^{-s^2} dB(s) \cdot e^{-t^2/2} + X(t), \\
Y_u(t) =& 2\, u \cdot \int e^{-(s-t)^2}\, d\Phi_0\left(\sqrt{2}K(s)\right) + \\
&+ \sqrt{\frac{2}{\pi}} \left(R + 2\left(\frac{2}{\pi}\right)^{1/4} \int s e^{-s^2} dB(s) \right) \cdot \int e^{-(s-t)^2}\, de^{-K^2(s)} \\
&- 2\left(\frac{2}{\pi}\right)^{1/4} \int e^{-s^2} dB(s) \cdot \int e^{-(s-t)^2}\, d\Phi_0\left(\sqrt{2}K(s)\right) + Y(t),
\end{aligned}
$$

Using the above relation, we illustrate particular components of the Slepian model for the joint up-crossing distribution of (Y_1, Y_2). We have chosen $\tau = 0.5$ for the shape parameter of the gamma process. The samples of underlying Brownian motion are the same as those in Figure 1.

In Figure 2 *(Top)*, we observe samples simulated from bivariate process $(X(t), Y(t))$ (to facilitate better visual comparison we have used the same sample of the underlying gamma process for all six samples of the Laplace moving average). They reveal complex dependence between processes and leptokurtic behavior of Y, which shows much larger extreme values than X. In the middle and right columns we see sample from the Slepian model at level $u = 0.5$ and $u = 5$, respectively. The level crossing occurs at $t = 0$ as seen at the top middle/right graphs. We observe in the bottom graphs that the random time change introduced by the gamma motion is adding to variability of Y at the crossing instants of X. For large level u the variability relatively to the level is reduced however the process Y still significantly overshoots the crossing value $u = 5$.

Our approach allows for investigating the role of particular components in the model that is illustrated in Figure 3, where the level dependent and the Rayleigh/kernel depend components are presented. We observe that the influence of Rayleigh variable for large level u is not significant and that major contribution to the Slepian model comes from the level dependent component.

5.2 Slepian noise model at crossings of a non-Gaussian moving averages

We have discussed Slepian models for the noise at crossings of a stationary Gaussian moving averages. Our interest will turn now to the case of crossings by a

moving average driven by a non-Gaussian noise $dL(s)$:

$$X(t) = \int g(s-t)\, dL(s) = \int g(s-t)\, dB \circ K(s), \tag{16}$$

where $K(t)$ is a gamma process with shape τ and scale $1/\tau$, i.e. $K(1)$ has the gamma distribution with these parameters. The choice of a gamma process as a subordinator is dictated by convenience of a simple parameterization, but in general one can consider other classes of non-negative second order Lévy processes. In Section 3, we have considered a simple example of random transformation of time given through $dB(Kt) = \sqrt{K}dB(t)$ and we have seen that the Slepian model K_u for K was giving a general way of describing arbitrary Slepian models in that case. The case considered now is more complex and the key to deriving the Slepian model is obtaining the biased sample distribution of $K_u(t)$, i.e. of the gamma process observed at the u level up-crossings of the moving average process (16).

Let us consider an arbitrary process Y and a process $Y(\cdot|k,z,u)$ with the distribution equal to conditional distribution of Y given $K = k$, where k is a trajectory of gamma process K, $\dot{X}(0) = z$, and $X(0) = u$. Then, as it was previously observed, if one have a joint Slepian model $(K_u, \dot{X}_u)$ for $(K, \dot{X})$, then a Slepian model for Y can be obtained through

$$Y_u(t) = Y(t|K_u, \dot{X}_u, u),$$

where for shortness $\dot{X}_u = \dot{X}_u(0)$. This approach splits finding a Slepian model for Y into two separate tasks: firstly, finding $Y(\cdot|k,z,u)$, then, secondly finding a Slepian model $(K_u, \dot{X}_u)$. While finding $Y(\cdot|k,z,u)$ is specific for a given process Y and need to be addressed in each case individually, while obtaining a Slepian model $(K_u, \dot{X}_u)$ is universal and is considered next.

It is easier to consider an extended model $(L_u, K_u, \dot{X}_u)$ and express a Slepian model for sampling from the crossing level distribution of this vector by a convenient Gibbs sampler. Namely, the model will based on alternate Gibbs samples from K_u conditionally on $dL_u, \dot{X}_u$ and $dL_u, \dot{X}_u$ conditionally on K_u. As shown in Appendix A.4, these two conditional distributions are given by

$$f_{K_u|dL_u,\dot{X}_u}(k|l,z) \sim f_{dL_u|dK_u,\dot{X}_u}(l|k,z) f_{\dot{X}_u|dK_u}(z|k)$$
$$f_{dL_u,\dot{X}_u|K_u}(l,z|k) \sim f_{dL_u|dK_u,\dot{X}_u}(l|k,z) f_{\dot{X}_u|dK_u}(z|k)$$

It follows directly by the above conditional arguments and properties stated in Appendix A.3, (19) and Appendix A.4, (20) and (21) that a Slepian model of L at the u-crossings of Y can be written in the form

$$\begin{aligned} L_u(t) = & \frac{G_{K_u}(t) - r_{K_u}\dot{G}_{K_u}(t)}{1-r_{K_u}^2}\frac{u}{\sqrt{\int g^2 dK_u}} + \\ & + \frac{\dot{G}_{K_u}(t) - r_{K_u}G_{K_u}(t)}{1-r_{K_u}^2}\frac{\dot{X}_u}{\sqrt{\int \dot{g}^2 dK_u}} + \\ & + \frac{G_{K_u}(t)}{1-r_{K_u}^2}\int r_{K_u}\,\dot{g}_{K_u} - g_{K_u}\,dB\circ K_u + \\ & \qquad + \frac{\dot{G}_{K_u}(t)}{1-r_{K_u}^2}\int r_{K_u}\,g_{K_u} - \dot{g}_{K_u}\,dB\circ K_u + B\circ K_u(t), \quad (17) \end{aligned}$$

where, for $k = K_u$, $r_k = \int g\dot{g}dk/\sqrt{\int \dot{g}^2 dk \int g^2 dk}$, $g_k = g/\sqrt{\int g^2 dk}$, $\dot{g}_k = \dot{g}/\sqrt{\int \dot{g}^2 dk}$, $G_k(t) = \int_0^t g_k dk$, $\dot{G}_k(t) = \int_0^t \dot{g}_k dk$.

Although the structure of the Slepian model is more complex as compared with the Gaussian case, one can still identify analogous components of the model. More precisely, we have the level and K_u depending component

$$F_{u,g,K_u}(t) = \frac{G_{K_u}(t) - r_{K_u}\dot{G}_{K_u}(t)}{1-r_{K_u}^2}\frac{u}{\sqrt{\int g^2 dK_u}},$$

linear combination of functions depending only on K_u, where the coefficients are random variables

$$\begin{aligned} G_{g,K_u,B}(t) = & \frac{\dot{G}_{K_u}(t) - r_{K_u}G_{K_u}(t)}{1-r_{K_u}^2}\,\frac{\dot{X}_u - \int \dot{g}\,dB\circ K_u}{\sqrt{\int \dot{g}^2 dK_u}} + \\ & + \frac{r_{K_u}\dot{G}_{K_u}(t) - G_{K_u}(t)}{1-r_{K_u}^2}\frac{\int g\,dB\circ K_u}{\sqrt{\int g^2 dK_u}}, \end{aligned}$$

and, finally, time distorted random noise

$$B\circ K_u(t).$$

Using this notation, we can write

$$L_u = F_{u,g,K_u} + G_{g,K_u,B} + B \circ K_u. \tag{18}$$

We observe that in this decomposition, in contrast to the stationary Gaussian case, all terms are dependent on the level u. However, the dependence of the second and third term is only through K_u. The second term is a linear combination of two functions of time t that become deterministic if conditioned on K_u, while the random coefficients of this combination are no longer independent as they were in the Gaussian case. Finally, we observe that our Gibbs sampler simulates at the same time $(\dot{X}_u, K_u, L_u)$ and thus allows for evaluation each of the three components in (18).

6 Asymptotics for Slepian models for large level crossings

The behavior of the process at large level crossings is often of interest for risk assessments at extreme events. Here we show how the derived Slepian models can be utilized to investigate such behavior.

6.1 Behavior at extreme values for the random scaling case

The obtained Slepian model allows for a discussion of asymptotic behavior at a high level crossing of the process. Recall that for a stationary Gaussian process with continuously differentiable trajectories and covariance satisfying $r(t) = 1 - t^2/2 + o(t^2)$ we have the following result, see Leadbetter *et al.* (1983). For each $\tau > 0$, with probability one

$$\lim_{u\to\infty} \sup_{0\le t\le\tau} \left| u\left(X_u(t/u) - u\right) + t^2/2 - R\,t \right| = 0.$$

The obtained Slepian models will be used to extend this result to a more general classes of stochastic processes.

Example 6 (continued)**.** For illustration, let us again consider the Slepian model for $X(t)$ given by (5). In order to obtain the asymptotic behavior in (7) for large u, we need to investigate the asymptotic distribution of K_u. It is verified in Appendix B.2 that the following normalization

$$\frac{K_u - u/\sqrt{a}}{\sqrt{u}/a^{3/4}}$$

yields asymptotic mean zero and variance one while in probability the so standardized variable converges to zero. The same holds for

$$\frac{\sqrt{K_u} - \sqrt{u}/\sqrt[4]{a}}{\sqrt{\frac{2p+1}{a}}\sqrt[4]{u}}$$

This asymptotics for K_u leads to the following behavior after high level crossing for the Slepian model given in (7):

$$\lim_{u\to\infty} \sup_{0\le t\le \tau} \left| X_u(t/\sqrt{u}) - u + t^2/2 - R\, t \right| = 0.$$

To see this let us first note that

$$\sup_{t\le\tau} \left|\Delta(t/\sqrt{u})\right| \le \tau/\sqrt{u} \sup_{0\le t\le\tau} \left|\Delta'(t/\sqrt{u})\right|.$$

By assumed regularity of the underlying Gaussian process, $\Delta'(0) = 0$ and Δ' is continuous with probability one, thus $M_u = \sup_{t\le\tau} |\Delta'(t/\sqrt{u})|$ converges almost surely to zero, when u increases without bound.

Thus

$$\begin{aligned}
&\left| X_u(t/\sqrt{u}) - u + t^2/2 - R\, t \right| \le \\
&\le \left| \left(u\left(r(t/\sqrt{u}) - 1\right) - \sqrt{K_u} R\, \dot{r}(t/\sqrt{u}) \right) + t^2/2 - R\, t \right| + \\
&+ \sqrt{K_u} \sup_{t\le\tau} \left|\Delta(t/\sqrt{u})\right| \\
&\le \left(\frac{\tau^2}{2} + \tau R \right) \sup_{t<\tau} \left| 1 + \ddot{r}(t/\sqrt{u}) \right| + \\
&\qquad + \tau \frac{\sqrt{\frac{2p+1}{a}}}{\sqrt[4]{u}} \left| \frac{\sqrt{K_u} - \frac{\sqrt{u}}{\sqrt[4]{a}}}{\sqrt{\frac{2p+1}{a}}\sqrt[4]{u}} \right| \left(R \sup_{t<\tau} \left|\ddot{r}(t/\sqrt{u})\right| + M_u \right)
\end{aligned}$$

and the right hand side converges to zero because $\ddot{r}(t)$ is continuous at zero and $\ddot{r}(0) = -1$,

6.2 Gaussian noise at high level crossings

Here we use the Slepian model given in (1) to obtain an asymptotic result about the noise behavior at high level crossings. Let us define the process

$$\xi_u(t) = u\left(B_u\left(\frac{t}{u}\right) - B\left(\frac{t}{u}\right) - \frac{\int_0^{t/u} g}{t/u}t\right)$$

and assume that both g and $\dot{g}$ are continuous at zero. Then

$$\sup_{0\le t\le\tau}\left|\xi_u(t) - \dot{g}(0)\left(R - \int \dot{g}\,dB\right)\cdot t - g(0)\int g\,dB\cdot t\right| \le$$
$$\le \tau\left(\frac{\int_0^{\tau/u}|g(s)-g(0)|\,ds}{\tau/u}\left|\int g\,dB\right| + \right.$$
$$\left. + \frac{\int_0^{\tau/u}|\dot{g}(s)-\dot{g}(0)|\,ds}{\tau/u}\left|R - \int \dot{g}\,dB\right|\right).$$

By continuity of g and $\dot{g}$ the right hand side converges to zero when u increases without bound, i.e.

$$\lim_{u\to\infty}\sup_{0\le t\le\tau}\left|\xi_u(t) - \dot{g}(0)\left(R - \int \dot{g}\,dB\right)\cdot t - g(0)\int g\,dB\cdot t\right| = 0.$$

By adding some conditions on the smoothness of g and $\dot{g}$ around zero (for example if g satisfies a Lipschitz condition), this asymptotic result can be made more exact. We will not discuss it here but we point, for a high level u, at a crude approximation of $B_u(t)$ in the neighborhood of 0:

$$B_u(t) \approx g(0)u\cdot t + B(t).$$

References

Åberg, S. & Podgórski, K. (2010). A Class of Non-Gaussian second order Spatio-Temporal Models. *Extremes* **14**, 187–222.

Åberg, S., Podgórski, K. & Rychlik, I. (2009). Fatigue damage assessment for a spectral model of non-Gaussian random loads. *Prob. Eng. Mech.* **24**, 608–617.

Adler, R. J., Samorodnitsky, G. & Taylor, J. E. (2013). High level excursion set geometry for non-gaussian infinitely divisible random fields. *The Annals of Probability* **41**, 134–169.

Aronowich, M. & Adler, R. (1988). Sample path behaviour of χ^2 surfaces at extrema. *Adv. Appl. Probab.* **20**, 719–738.

Azaïs, J.-M. & Wschebor, M. (2009). *Level Sets and Extremes of Random Processes and Fields*. Wiley & Sons.

Baxevani, A., Podgórski, K. & Rychlik, I. (2003). Velocities for moving random surfaces. *Probabilistic Engineering Mechanics* **18**, 251–271.

Ditlevsen, . (1985). Survey on applications of Slepian model processes in structural reliability. In I. Konishi, A. H.-S. A. 1.-S. Ang & M. Shinozuka, eds., *The Proceedings of the 4th international conference on structural safety and reliability, ICOSSAR '85*, vol. 1. pp. 241–250.

Gadrich, T. & Adler, R. (1993). Slepian models for non-stationary Gaussian processes. *J. Appl. Prob.* **30**, 98–111.

Galtier, T. (2011). Note on the estimation of crossing intensity for laplace moving average. *Extremes* **14**, 157–166.

Grigoriu, M. (1989). Reliability of Daniels systems subject to quasistatic and dynamic non-stationary Gaussian load processes. *Prob. Eng. Mech.* **4**, 128–134.

Hörmann, W. & Leydold, J. (2013). Generating generalized inverse gaussian random variates. *Statistics and Computing* , 1–11.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Springer-Verlag.

Kac, M. & Slepian, D. (1959). Large excursions of Gaussian processes. *Ann. Math. Statist.* **30**, 1215–1228.

Kotz, S., Kozubowski, T. & Podgórski, K. (2001). *The laplace distribution and generalizations: A revisit with applications to communications, economics, engineering and finance*. Birkhäuser, Boston.

Leadbetter, M., Lindgren, G. & Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer-Verlag.

Lindgren, G. (1989). Slepian models for χ^2-processes with dependent components with application to envelope upcrossings. *J. Appl. Probab.* **26**, 36–49.

Olver, F. W. J., Lozier, D. W., Boisvert, R. F. & Clark, C. W., eds. (2010). *Nist handbook of mathematical functions*. Cambridge University Press, New York, NY.

Podgórski, K., I., R. & Machado, U. E. B. (2000). Exact distributions for apparent waves in irregular seas. *Ocean Eng.* **27**, 979–1016.

Slepian, D. (1963). On the zeros of Gaussian noise. In M. Rosenblatt, ed., *Time series analysis*. Wiley, New York, pp. 104–115.

van de Lindt, J. W. & Niedzwecki, J. M. (2005). Structural response and reliability estimates: Slepian model approach. *Journal of Structural Engineering* **131**, 1620–1628.

Wilson, R. & Adler, R. (1982). The structure of Gaussian fields near a level crossing. *Adv. Appl. Probab.* **14**, 543–565.

Zähle, U. (1984). A general Rice formula, Palm measures, and horizontal-window conditioning for random fields. *Stochastic Process. Appl.* **17**, 265–283.

A Slepian models – proofs

A.1 Slepian model for random scaling

We derive a Slepian model for the simple case discussed in Example 3. Invoking the generalized Rice formula for the process $Y(t) = \sqrt{K}X(t)$ yields

$$\begin{aligned} P^u(A) &= \frac{E\left(\sqrt{K}\dot{X}^+(0)\{\sqrt{K}X(\cdot) \in A\}|\sqrt{K}X(0) = u\right)}{E\left(\sqrt{K}\dot{X}^+(0)|\sqrt{K}X(0) = u\right)} \\ &= \frac{\int_0^\infty \int_0^\infty \sqrt{k}z \cdot P(A|z,k,u) \cdot f_{\dot{X}|K,\sqrt{K}X}\left(z|k,u\right) f_{K|\sqrt{K}X}(k|u)\, dz\, dk}{\int_0^\infty \int_0^\infty \sqrt{k}z \cdot f_{\dot{X}|K,\sqrt{K}X}\left(z|k,u\right) f_{K|\sqrt{K}X}(k|u)\, dz\, dk}, \end{aligned}$$

where $P(A|z,k,u) = P\left(\sqrt{k}X(\cdot) \in A|\dot{X}(0) = z, K = k, \sqrt{k}X(0) = u\right)$.

We note that $\dot{X} = \dot{X}(0)$ is independent of $(K, \sqrt{K}X(0))$ and

$$f_{K|\sqrt{K}X}(k|u) = c_u \cdot f_K(k) \cdot f_X(u/\sqrt{k})/\sqrt{k} = c_u e^{-\frac{u^2}{2k}} f_K(k)/\sqrt{k}.$$

This yields

$$\begin{aligned} P^u(A) &= \frac{\int_0^\infty \int_0^\infty P\left(\sqrt{k}X(\cdot) \in A|\sqrt{k}X(0) = u, \dot{X}(0) = z, K = k\right) z e^{z^2/2} e^{-\frac{u^2}{2k}} f_K(k)\, dz\, dk}{\int_0^\infty z e^{-z^2/2}\, dz \int_0^\infty e^{-\frac{u^2}{2k}} f_K(k)\, dk} \\ &= c_u \cdot \int_0^\infty \int_0^\infty P\left(ur(\cdot) - \sqrt{k}zr'(\cdot) + \sqrt{k}\Delta(\cdot) \in A\right) z e^{-z^2/2} e^{-\frac{u^2}{2k}} f_K(k)\, dz\, dk, \end{aligned}$$

where $c_u^{-1} = \int_0^\infty e^{-\frac{u^2}{2k}} f_K(k)\, dk$ and the Gaussian process $\Delta(t)$ is as described in Introduction. This leads to the Slepian model

$$Z_u(t) = ur(t) - \sqrt{K_u}Rr'(t) + \sqrt{K_u}\Delta(t),$$

where K_u, R, and $\Delta(t)$ are mutually independent and distributed as follows: the distribution of K_u is given by the density $f_{K_u}(k) = c_u e^{-\frac{u^2}{2k}} f_K(k)$, R is having the Rayleigh distribution.

A.2 Slepian model for homogeneous Gaussian noise

Here we derive a Slepian model for the homogeneous Gaussian noise that is driving a moving average process. Such a noise can be viewed as a stochastic measure defined through a Brownian motion $B(t)$, $t \in \mathbb{R}$, obtained from a regular Brownian motion by reflecting it independently at $t = 0$, so that $B(t)$ represents the measure of $[0, t]$ for $t > 0$ (for negative t it equals to the minus measure of $[t, 0]$). This identification of the measures and processes is kept throughout the paper. The biased sampling distribution for the finite dimensional distributions of the Gaussian process B at the u-level up-crossings of $X(t) = \int g(s-t)\, dB(s)$ are obtained below by considering the conditional distribution of $B(t), B(s)$, for some fixed t and s, $|t| \geq |s|$, given $X(0) = u$ and $\dot{X}(0) = z$.

The covariance matrix of the normally distributed vector $(B(t), B(s), X(0), \dot{X}(0))$ is given by

$$\mathbf{\Sigma} = \begin{bmatrix} |t| & a & G(t) & g(t)-g(0) \\ a & |s| & G(s) & g(s)-g(0) \\ G(t) & G(s) & 1 & 0 \\ g(t)-g(0) & g(s)-g(0) & 0 & 1 \end{bmatrix},$$

where $a = |s|$ if s and t have the same sign and zero otherwise, while $G(t) = \int_0^t g$. The 'ones' on the diagonal are the consequence of the assumption: $\int |g|^2 = \int |\dot{g}|^2 = 1$. Direct verification of the covariances leads to the following representation of the Gaussian noise B given $X(0) = u$, $\dot{X}(0) = z$:

$$B(t|u,z) = u \cdot G(t) + z \cdot (g(t) - g(0)) - G(t) \int g dB - (g(t) - g(0)) \int \dot{g} dB + B(t).$$

From the Rice formula it follows that this is a Slepian process for B at the up-crossing level distribution given give that the derivative $\dot{X}_u$ at the up-crossings is equal to z. Taking into account that the Rayleigh distribution is representing the biased sampling distribution of the derivative and using (3), a Slepian model for the noise is given by

$$\begin{aligned} B_u(t) =& u \cdot G(t) + R \cdot (g(t) - g(0)) - G(t) \int g dB - (g(t) - g(0)) \int \dot{g} dB \\ &+ B(t), \end{aligned}$$

where random variable R has the Rayleigh distribution and is independent of B.

A.3 Conditional Slepian model of noise given subordinator, and derivative

An extension of the representation of the noise from the previous section to a non-homogeneous Gaussian noise is important for derivation of the Slepian model when a moving average with respect to the Laplace noise is crossing a level u. Formally, we are interested in the conditional Slepian model L_u of the non-Gaussian noise L given that $K_u = k$ and $\dot{X}_u = z$, where K_u and $\dot{X}_u$ are some Slepian models for the subordinator K and the derivative $\dot{X}(0)$. Here, the crossing levels are marked by the non-Gaussian moving average given in (16).

As presented in (3), this conditional distribution is equivalent to that of L given $K = k$, $\dot{X} = z$, and $X = u$. This in turn can be presented through conditioning a non-stationary Gaussian moving average process

$$X_k(t) = \int g(s-t)\, dB(k(s)),$$

where $k(s)$ is a non-decreasing function, i.e. we consider a moving-average integral with respect to stochastic measure B_k defined on intervals as

$$B_k(s, s+ds] = B(k(s), k(s+ds)]$$

and as before we use the same notation to denote the corresponding independent increment process B_k. In what follows, we use $\int f dk$ for $\int f(t)dk(t)$.

The joint distribution of $(B_k(t), B_k(s), \dot{X}(0), X(0))$ has the covariance matrix

$$\mathbf{\Sigma}_k = \begin{bmatrix} |k(t)| & a & \int_0^t g dk & \int_0^t \dot{g} dk \\ a & |k(s)| & \int_0^s g dk & \int_0^s \dot{g} dk \\ \int_0^t g dk & \int_0^s g dk & \int g^2 dk & \int g\dot{g} dk \\ \int_0^t \dot{g} dk & \int_0^s \dot{g} dk & \int g\dot{g} dk & \int \dot{g}^2 dk \end{bmatrix},$$

where $a = |k(s)|$ if the signs of $k(t)$ and $k(s)$ are the same and zero otherwise. For compactness of the presentation, let set $r_k = \int g\dot{g} dk / \sqrt{\int \dot{g}^2 dk \int g^2 dk}$, $g_k = g/\sqrt{\int g^2 dk}$, $\dot{g}_k = \dot{g}/\sqrt{\int \dot{g}^2 dk}$, $G_k(t) = \int_0^t g_k dk$, $\dot{G}_k(t) = \int_0^t \dot{g}_k dk$.

Direct verification of the covariances proves the following conditional Slepian model for the non-homogenous noise L_u conditionally on $\dot{X}_u = z, K_u = k$ from which samples can be taken for the Gibbs sampler scheme can be written in the form

$$B_k(t|u,z) = \frac{G_k(t) - r_k\dot{G}_k(t)}{1-r_k^2}\frac{u}{\sqrt{\int g^2 dk}} + \frac{\dot{G}_k(t) - r_k G_k(t)}{1-r_k^2}\frac{z}{\sqrt{\int \dot{g}^2 dk}} +$$
$$+ \frac{G_k(t)}{1-r_k^2}\int r_k\dot{g}_k - g_k\, dB_k + \frac{\dot{G}_k(t)}{1-r_k^2}\int r_k g_k - \dot{g}_k\, dB_k + B_k(t). \quad (19)$$

A.4 Slepian model of noise, subordinator, and derivative based on Gibbs sampler

Here, we discuss a Slepian model $(L_u, K_u, \dot{X}_u)$ that is based on a Gibbs sampler. Let us consider the gamma process K, i.e. the Lévy process such that $K(1)$ has the gamma distribution with shape τ and scale $1/\tau$ (for negative t, the process $-K(t)$ is an independent copy of $K(t), t \geq 0$). For the computational and practical reasons it is more convenient to consider a discretized version of the problem. We consider a uniformly spaced grid dt (for compactness of the notation, we use dt both for the grid and for its diameter) and assume that the stochastic measures are approximated by the Lebesgue measure multiplied by the random increment of a considered stochastic measure over an individual cell of the grid. In particular, dL is a vector of values of the noise increments over this grid, K are random gamma variances distributed with shape τdt and scale $1/\tau$ while Z is a vector of independent standard normal random variables so that $dL = \sqrt{K}Z$, where the multiplication is coordinate-wise. We also use K_u and dL_u as notation for Slepian models of K and dL, respectively.

Further if g is a function, then $\int g$ is a vector of values of the Lebesgue integrals of g over the cells of the grid dt. With the assumed discretization and a slight abuse of the notation, we write $\int g\, dL = \int g \cdot dL$, where $\cdot$ stands for the inner product of the two vectors. Consequently, we write $X = \int f dL = \int f\sqrt{K}Z$ and $\dot{X} = \int \dot{f} dL = \int \dot{f}\sqrt{K}Z$.

We first notice that $(K_u|dL_u = dl, \dot{X}_u = z) \stackrel{d}{=} (K|dL = dl, \dot{X} = z, X = u)$, where each of the sides denotes a conditional distribution. Since, both $\dot{X}$ and X are deterministic functions of dL thus we can assume from now on that $\int \dot{f} dl = z$ and $\int f dl = u$. The Bayes formula yields

$$f_{K_u|dL_u,\dot{X}_u}(k|dl,z) \propto f_{dL|K}(dl|k) f_K(k) \propto \left(\prod_{i=1}^{N} k_i^{\tau dt - 3/2} e^{-(2k_i/\tau - dl_i^2/k_i)/2}\right), \quad (20)$$

which corresponds to the distributed of a vector of independent variables distributed as $GIG(\tau dt - 1/2, 2/\tau, dl_i^2)$.

The Gibbs sampler from the Slepian model $(L_u, K_u, \dot{X}_u)$ will be based on alternate samples from the conditional distributions:

$$k^{(j)} \sim (K_u | dL_u = dl^{(j-1)}, \dot{X}_u = z^{(j-1)})$$
$$\left(dl^{(j)}, z^{(j)}\right) \sim (dL_u, \dot{X}_u | K_u = k^{(j)}).$$

As we have seen above, the first sampling is reduced to sampling from independent GIG distributions for which there exists a uniformly bounded rejection algorithm, see Hörmann & Leydold (2013). Let us next discuss how to sample from $dL_u, \dot{X}_u$ given that $K_u = k$. We note that it is equivalent to sampling z from $\dot{X}_u$ given that $K_u = k$ and then l from dL_u given that $\dot{X}_u = z, K_u = k$ which is the same as sampling from dL given that $\dot{X} = z, K = k, X = u$, which was discussed in the previous section, see (19).

Further extending our notation for the discretized model, for any integrable function g we write $\int g\, dk$ as the inner product between the vector of the Lebesgue integrals of g over the grid cells and the vector of values of the gamma vector $K = k$. Using this convention we define $r = \int g\dot{g} dk / \sqrt{\int \dot{g}^2 dk \int g^2 dk}$ and for a function g we write $G_k(A) = \int_A g\, dk = \int \mathbf{1}_A g\, dk$ and $G_k(t, A) = \int_A g_t\, dk$, where $g_t(s) = g(s - t)$.

It follows directly from the Rice formula that the distribution of $(\dot{X}_u | K_u = k)$ is given by the density

$$f_{\dot{X}_u}(z) \sim z \cdot \exp\left(-\frac{(z-a)^2}{b}\right), \tag{21}$$

where

$$a = u \cdot r \cdot \sqrt{\int \dot{g}^2\, dk} / \sqrt{\int g^2\, dk},$$
$$b = 2(1 - r^2) \int \dot{g}^2\, dk.$$

Sampling from this density can be done through a accept reject algorithm.

A.5 Generation of Slepian noise distribution

The purpose of this section is to approximate the integral for the palm equation (2) by MC approximation for the case when $X(t)$ is the Laplace motion. In general when dealing with the non Gaussian case the evaluation of (2) the distribution $f_{X,\dot{X}}$ is not known and thus evaluating the integral directly is not possible. For models driven by Laplace moving average it is possible to approximate the process with a piecewise constant process; that is we approximate

$$X(t) = \int g(t-s)dL(s)$$

with

$$X(t) = \sum_{i=1}^{N} \mathbb{I}(t < t_i) g(t_i - t_{i-1}/2) \left(dL(t_i) - dL(t_{i-1})\right),$$

where t_i is assumed to be evenly spaced time points in the interval of interest. split the time into intervals and let the process be constant on each interval. Then define K as constant on each interval and taking the value of a Gamma random variable. Now one can in theory evaluate the integrals numerically. But, in general the number of intervals used to discretion of the time will make the numerical integration infeasible.

An natural approach is then to replace the numerical integration with a MC-integration, that is sample gamma r.v and evaluate $\int_0^{+\infty} P(Y \in A|\dot{X} = z, X = u, K = k) \cdot z f_{\dot{X},X|K}(z,u|k)\,dz$ and $\int_0^{+\infty} z f_{\dot{X},X|K}(z,u|k)\,dz$ numerically. However, if u is large then this method will be ineffective as the mass of (2) is where K is large. Instead we sample from a distribution proportional to $f_{X|K}(u|k)f_K(k)$. This avoids the problem of sampling from the regular Gamma distribution. The samples from $f_{X|K}(u|k)f_K(k)$ is generated through the Gibbs sampler described below.

The ith iteration of the Gibbs sampler consists of two steps

$$dl^{(i)} \sim f_{dL|X,K}(\cdot|u, k^{(i-1)})$$
$$k^{(i)} \sim f_{K|dL}(\cdot|dl^{(i)})$$

where dL is the Laplace noise driving the process. Both distributions are explicit namely multivariate normal for dl and Generalised inverse Gaussian for k. We describe both steps in greater detail below

B Generalized inverse Gaussian distribution

B.1 Definition and basic properties

The generalized inverse Gaussian distribution with parameters $p \in \mathbb{R}$, $a \geq 0$, and $b \geq 0$, for shortness $GIG(p, a, b)$, is given by the pdf

$$f(x) = \frac{\left(a/b\right)^{p/2}}{2K_p\left(\sqrt{ab}\right)} x^{p-1} e^{-\frac{ax+b/x}{2}}.$$

The parameters satisfy

$$\begin{aligned} &a > 0\,, b \geq 0\,, \text{ if } p > 0,\\ &a > 0\,, b > 0\,, \text{ if } p = 0,\\ &a \geq 0\,, b > 0\,, \text{ if } p < 0. \end{aligned}$$

The moment generating function of a GIG distribution is given by

$$M(t) = \left(\frac{a}{a-2t}\right)^{p/2} \frac{K_p\left(\sqrt{b(a-2t)}\right)}{K_p(\sqrt{ab})}, \; t < a/2. \tag{22}$$

The following formulas for the expectations of a $GIG(p, a, b)$ random variable X hold

$$\begin{aligned} \mathbb{E}[X^\lambda] &= (b/a)^{\lambda/2} \frac{K_{p+\lambda}\left(\sqrt{ab}\right)}{K_p\left(\sqrt{ab}\right)}, \quad \lambda \in \mathbb{R}\\ \mathbb{E}[\log(X)] &= \log(\sqrt{a/b}) + \frac{\partial \log K_p}{\partial p}\left(\sqrt{ab}\right), \end{aligned} \tag{23}$$

where $\partial \log K_p / \partial p(x)$ is the derivative of the Bessel function $K_p(x)$ with respect of its order p and evaluated at value (p, x), cf. Jørgensen (1982). Consequently, by setting $R_p(x) = K_{p+1}(x)/K_p(x)$ we obtain

$$\begin{aligned} \mathbb{E}[X] &= \sqrt{\frac{b}{a}} \cdot R_p\left(\sqrt{ab}\right),\\ \mathbb{E}[X^{-1}] &= \sqrt{\frac{a}{b}} \cdot \frac{1}{R_{p-1}\left(\sqrt{ab}\right)}. \end{aligned} \tag{24}$$

This together with the following recurrence relation

$$R_p(x) = 2p/x + 1/R_{p-1}(x),$$

see Jørgensen (1982), yields

$$\mathbb{E}[X^{-1}] = \frac{a}{b} \cdot \left(\mathbb{E}[X] - 2p/a\right). \tag{25}$$

The special case of $p = \frac{1}{2}$ corresponds to the reciprocal inverse Gaussian and the following simple forms of expectations hold

$$\mathbb{E}[X] = \sqrt{\frac{b}{a}}\left(1 + \frac{1}{\sqrt{ab}}\right)$$
$$\mathbb{E}[X^{-1}] = \sqrt{\frac{a}{b}},$$

B.2 Asymptotic behavior when b is increasing without bound

In our considerations of high level crossings, we need the asymptotic behavior of $GIG(p, a, b)$ when b is increasing without bound. For this we will use the following property of Bessel function K_p as its arguments increases without bound

$$\lim_{x\to\infty} x^2 \left(\sqrt{\frac{2x}{\pi}} e^x K_p(x) - \left(1 + \frac{a_p}{x}\right)\right) = b_p,$$

where $a_p = (4p^2 - 1)/8$ and $b_p = (4p^2 - 1)(4p^2 - 9)/128$, see Olver *et al.* (2010). We note first that the following behavior of the expected values

$$\begin{aligned} E(X) - \sqrt{\frac{b}{a}} &= \sqrt{\frac{b}{a}}\left(\frac{K_{p+1}\left(\sqrt{ab}\right)}{K_p\left(\sqrt{ab}\right)} - 1\right) \\ &\sim \frac{1}{a}\sqrt{ab}\sqrt{\frac{2\sqrt{ab}}{\pi}} e^{\sqrt{ab}}\left(K_{p+1}\left(\sqrt{ab}\right) - K_p\left(\sqrt{ab}\right)\right) \\ &\sim \frac{a_{p+1} - a_p}{a} = \frac{2p+1}{a}, \end{aligned} \tag{26}$$

$$E(\sqrt{X}) - \sqrt[4]{\frac{b}{a}} = \sqrt[4]{\frac{b}{a}} \left(\frac{K_{p+\frac{1}{2}}\left(\sqrt{ab}\right)}{K_p\left(\sqrt{ab}\right)} - 1 \right) \sim \frac{4p+1}{8\sqrt[4]{a^3 b}}, \tag{27}$$

where the relation $\sim$ means that the ratio of the two expressions converges to one when b converges to infinity.

Similarly, we analyze the variance of X:

$$\begin{aligned} V(X) &= \frac{b}{a} \left(\frac{K_{p+2}\left(\sqrt{ab}\right)}{K_p\left(\sqrt{ab}\right)} - \frac{K_{p+1}^2\left(\sqrt{ab}\right)}{K_p^2\left(\sqrt{ab}\right)} \right) \\ &= \frac{b}{a} R_p(\sqrt{ab}) \left(R_{p+1}\left(\sqrt{ab}\right) - R_p\left(\sqrt{ab}\right) \right) \\ &\sim \frac{b}{a} \left(R_{p+1}\left(\sqrt{ab}\right) - R_p\left(\sqrt{ab}\right) \right). \end{aligned}$$

It remains to investigate the behavior of $R_{p+1}(x) - R_p(x)$ for x increasing without bound.

Consequently,

$$\begin{aligned} R_{p+1}(x) - R_p(x) &\sim \left(1 + \frac{a_{p+2}}{x} + O(x^{-2})\right) \left(1 + \frac{a_{p+1}}{x} + O(x^{-2})\right) + \\ &\quad - \left(1 + \frac{a_{p+1}}{x} + O(x^{-2})\right)^2 \\ &\sim \frac{1}{x}(a_{p+2} + a_p - 2a_{p+1}) = \frac{1}{x}. \end{aligned}$$

This gives the following asymptotics for the variance (when b increases without bound):

$$V(X) \sim \frac{\sqrt{b}}{a^{3/2}}.$$

By examining the moment generating function, it can be verified that the standardized variable $(X - E(X))/\sqrt{V(X)}$ converges in probability to zero. Consequently, in probability,

$$\lim_{b\to\infty} \frac{X - \sqrt{\frac{b}{a}}}{\sqrt[4]{\frac{b}{a^3}}} = 0,$$

while the mean and variance of the so standardized variable converge to zero and one, respectively.

Similarly for $\sqrt{X}$, due to (26) and (27), we have

$$\lim_{b\to\infty}\frac{V(\sqrt{X})}{\sqrt[4]{b}}=\lim_{b\to\infty}\left(E(X)-\sqrt{\frac{b}{a}}+\right.$$
$$\left.+\frac{2}{\sqrt[4]{a}}\left(E(\sqrt{X})-\sqrt[4]{\frac{b}{a}}\right)-\frac{\left(E(\sqrt{X})-\sqrt[4]{\frac{b}{a}}\right)^2}{\sqrt{b}}\right)$$
$$\sim\frac{2p+1}{a}.$$

Thus the variable

$$\frac{\sqrt{X}-\sqrt[4]{\frac{b}{a}}}{\sqrt{\frac{2p+1}{a}}\sqrt[8]{b}}$$

has the asymptotic mean and variance equal to zero and one, respectively, while it is converging to zero in probability.