Posterior Inference for Sparse Hierarchical Non-stationary Models

Lassi Roininen
Lappeenranta University of Technology, Finland
with Karla Monterrubio-Gómez, Sara Wade, Neil Chada, Janne Huttunen, Kenneth Muhumuza, Sari Lasanen, Jarkko Suuronen, Alberto Mendoza, Theo Damoulas, Mark Girolami

Potsdam 2020
Centre of Excellence in Inverse Modelling and Imaging

Finnish Centre of Excellence in Inverse Modelling and Imaging
2018-2025

Universities and research institutes involved:
- University of Oulu
- University of Jyväskylä
- Tampere University of Technology
- Finnish Meteorological Institute
- Aalto University
- LUT Lappeenranta University of Technology
- University of Helsinki

CENTRES OF EXCELLENCE IN RESEARCH

Potsdam 2020 2 / 36
Detect material interfaces, inhomogeneous structures, anisotropies, Gaussian and non-Gaussian features

Gaussian and non-Gaussian hierarchical random field priors, parametric models for structures, noise models etc

Metropolis-within Gibbs, elliptical slice sampling, Hamiltonian Monte Carlo, and optimisation methods

Applications: Subsurface imaging (electrical impedance tomography, Darcy flow models) and near-space remote sensing (High-power radar experiments, satellite tomography and remote sensing)
Non-Gaussian Priors
Lassas and Siltanen 2004 showed that TV are not discretisation-invariant

Lassas, Saksman and Siltanen 2009 constructed Besov space priors

- Often defined via wavelet expansions.
- For edge-preserving inversion the Haar wavelet basis is often used
- However due to the structure of the Haar basis, discontinuities are preferred on an underlying dyadic grid given by the discontinuities of the basis functions. For example, on the domain \((0, 1)\), discontinuity is vastly preferred at \(x = 1/4\) over \(x = 1/3\).
- Thus Besov priors make, in most practical cases, both a strong and unrealistic assumption.
Non-Gaussian models – $\alpha$-stable priors

- Alberto Mendoza, Lassi Roininen, Mark Girolami, Jere Heikkinen and Heikki Haario, Statistical Methods To Enable Practical On-Site Tomographic Imaging of Whole-Core Samples, Geophysics (2019).
Stable random walks

Let $U(t), t \in \mathbb{I} \subset \mathbb{R}^+$ be a stochastic process. We call it a Lévy $\alpha$-stable process starting from zero, or simply as stable process, if $U(0) = 0$, $U$ has independent increments and

$$U(t) - U(s) \sim S_\alpha \left( (t - s)^{1/\alpha}, \beta, 0 \right)$$

for any $0 \leq s < t < \infty$ and for some $0 < \alpha \leq 2, -1 \leq \beta \leq 1$.

For the continuous limit of the Cauchy walk, we apply independently scattered measures. We obtain random walk approximation

$$U_{t_i} - U_{t_{i-1}} \sim S_\alpha (h^{1/\alpha}, \beta, 0)$$

where $t_i - t_{i-1} =: h$. It is easy to see that such random walk approximations converge to the $\alpha$-stable Lévy motion as $h \to 0$ in distribution on the Skorokhod space of functions that are right-continuous and have left limits.
The paper lays the groundwork for Bayesian inverse problems with stable fields, specifically stable stochastic integrals \( U(x) = \int_E f(x, x') M(dx') \)

The paper has expository flavour: We study the very simple stable sheets as an illustrative and easy to follow example. For stable sheets, \( f(x, x') = \begin{cases} 1 \text{ when } x_i' \leq x_i \text{ for all } i = 1, \ldots, d \\ 0 \text{ otherwise} \end{cases} \) (2)

Stable integral is defined like the usual Itô integral, but with stable random measure \( M \) in place of Brownian motion \( B \)/Brownian sheet.

Note: no second moments means that instead of \( L^2 \), the integrals are limits of integrals of simple functions in probability.
Stable Integrals

- Stable integrals $U(x)$ can be presented in many equivalent ways (equivalent in distribution). When it comes to Bayesian inverse problems, the best way seems to be through Lévy-LePage series representation,

- Lévy-LePage series representation is

$$U(x) = (C_\alpha |E|)^{\frac{1}{\alpha}} \sum_{k=1}^{\infty} \rho_k \Gamma_k^{1/\alpha} f(x, V_k),$$  \hspace{1cm} (3)

where $0 < \alpha < 2$,

$$C_\alpha = \left( \int_0^{\infty} x^{-\alpha} \sin(x) dx \right)^{-1},$$  \hspace{1cm} (4)$$

$\rho_k$ is a Rademacher sequence (i.i.d. with values ±1 with equal probabilities), $\Gamma_k$ are arrival times of a Poisson process with arrival rate 1, and $V_k$ are i.i.d. uniformly distributed on $E$. The three sequences $\rho_k, \Gamma_k$ and $V_k$ are mutually independent.
Discretisation

- Lévy-LePage series representation gives
  - 1) Sample path regularity in $L^p$, $1 \leq p < \infty$ (also in the more general Sobolev space $H^s_p$, $s < 1/p$, $p \geq 2$),
  - 2) Convergence in distributions on sample space.

- From 1) and 2), we proceed to posterior convergence in distribution for finite-dimensional data. The discretization of $U$ on $[0,1]^d$ is taken to be

\[
U^N(x) = U(h[x/h]),
\]

(5)

where the ceiling function $\lceil t \rceil = \min\{m \in \mathbb{Z}^d : t_j \leq m_j, j = 1, \ldots, d\}$.

- Computationally, the discretisation is determined from set of difference equations (here in 2D case)

\[
U(hm_1, hm_2) - U(hm_1, h(m_2 - 1)) - U(h(m_1 - 1), hm_2) + U(h(m_1 - 1), h(m_2 - 1)) \sim S_\alpha(|h|^{d/\alpha}, 0, 0)
\]

(6)

with i.i.d. right hand sides and zero boundary values on the coordinate axes.
Theorem

Let $1 \leq p < \infty$. The approximations $U^N(x) = U(h \lceil x/h \rceil)$ converge to $U$ on $L^p((0,1)^d)$ in distribution.

Open questions:
- Can we do the same for infinite-dimensional data (e.g. with Gaussian noise)?
- How to obtain stronger posterior convergence for $\alpha = 1$?
MAP estimates

- Log tomography with different number of projections and prior models

TV

Cauchy

Besov

Gauss

90

30

10
MAP and CM estimates – 30 projections

- Log tomography

TV

MAP

Cauchy

MwG

Besov

HMC
Tomographic imaging of whole-core samples

- 46, 23, 12, 6 projections with 10% noise
Example 4: mixed-wet carbonate reservoir rocks from the Middle-East.*

Waterflood oil-bearing carbonate sample

Born waveform inversion of seismic data

A Bayesian approach to improving acoustic Born waveform inversion of seismic data for viscoelastic media
Sparse Hierarchical Non-stationary Models

Hierachical GP model

- Based on the non-stationary Matérn kernel via varying length-scaling $\ell(x_i)$.
- Hierarchical model for 1-$d$ problems:

\[
y_i \sim \mathcal{N}(z(x_i), \sigma^2_\varepsilon), \quad i = 1, \ldots, m,
\]
\[
z(\cdot) \sim \mathcal{GP} \left(0, C^{\text{NS}}_{\phi}(\cdot, \cdot)\right),
\]
\[
\log \ell(\cdot) \sim \mathcal{GP} \left(\mu_\ell, C^S_{\phi}(\cdot, \cdot)\right),
\]
\[
(\tau^2, \varphi, \sigma^2_\varepsilon, \mu_\ell) \sim \pi(\tau^2)\pi(\varphi)\pi(\sigma^2_\varepsilon)\pi(\mu_\ell),
\]

where $C^{\text{NS}}_{\phi}(\cdot, \cdot)$ denotes a non-stationary kernel, $C^S_{\phi}(\cdot, \cdot)$ is a stationary covariance function with parameters $\varphi$, and $\mu_\ell$ the constant mean of $\log \ell(\cdot)$.

- Extremely flexible, 2-level improves predictive performance
- Fully Bayesian inference challenging:
  - Computationally expensive (2 nested GPs), latent processes and hyperparameters tend to be strongly coupled
  - Model is sensitive to the choice of hyperparameters.
Sparse Hierarchical Non-stationary Models

- **Idea:** Use Gaussian Markov random fields - precision matrix equivalent to \( z \sim \mathcal{N}(0, C_{\phi}^{NS}) \)

- **How to create \( Q \)?**
  - Roininen et al. 2019 derive a SPDE formulation for non-stationary Matérn fields.
  - For \( d = 1 \) and \( \nu = 2 - 1/2 \),
    \[
    \left(1 - \ell(\cdot)^2 \Delta\right) z = \tau \sqrt{\ell(\cdot)} w, \tag{8}
    \]
    where \( \Delta \) is the Laplace operator, \( w \) is white noise on \( \mathbb{R} \), \( \text{Var}(w) = \Gamma(\nu + 1/2)(4\pi)^{1/2}/\Gamma(\nu) \), and \( \ell(\cdot) \) is a spatially varying length-scale.

- A finite-dimensional approximation can be written as
  \[
  L(\ell)z = w,
  \]
  where \( z \in \mathbb{R}^n \) with \( n \) the discretisation size. \( L(\ell) \) is a sparse matrix depending on \( \ell_j := \ell(jh) \), with \( h \) the discretisation step in a chosen finite difference approximation.
The model

- GP regression model: \( y = Az + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_m), \quad A \in \mathbb{R}^{m \times n}, \ z \in \mathbb{R}^n. \)

- Hierarchical formulation

\[
\begin{align*}
  y \mid z, \sigma_\varepsilon^2 &\sim \mathcal{N}(Az, \sigma_\varepsilon^2 I_m), \\
  z \mid \phi &\sim \mathcal{N}(0, Q_\phi^{-1}) \\
  \log \ell &:= u \mid \varphi \sim \mathcal{N}(\mu_\ell, C_\varphi) \\
  (\tau^2, \sigma_\varepsilon^2, \varphi, \mu_\ell) &\sim \pi(\tau^2)\pi(\sigma_\varepsilon^2)\pi(\varphi)\pi(\mu_\ell)
\end{align*}
\]

where \( \mu_\ell \) is the \( n \)-dimensional constant mean vector.

- Key component: \( (C_\phi^{\text{NS}})^{-1} := Q_\phi = L(\phi)^T L(\phi), \) which depends on \( u \) and \( \tau^2. \)

- \( \varphi \) parameters of the covariance that describe properties of the length-scales.
Hyperpriors

- **Stationary** assumption for spatially varying length-scale

- Explore two priors for $u$:
  
  **Squared Exponential:**
  - Strong prior smoothness assumptions on how the correlation of the non-stationary process changes with distance.
  - Precision matrix is **dense** and depends on length-scale $\lambda$ and magnitude $\tau_\ell$.

  **AR(1):**
  - Ornstein-Uhlenbeck covariance
  - Allows quick changes but is smoother than white noise.
  - Precision is **sparse** $Q_\varphi = L(\varphi)^T L(\varphi)$, where $L(\varphi)$ is a banded matrix that depends on $\lambda$ and $\tau_\ell$.

- To improve model identifiability, we fix $\tau$, $\mu_\ell$ and $\tau_\ell$. 
Inference for one-dimensional problems

Posterior of interest:
\[
\pi(z, u, \lambda, \sigma^2 \varepsilon \mid y) \propto \mathcal{N}(y \mid A z, \sigma^2 \varepsilon I_m) \mathcal{N}(z \mid \mu_z, Q_u^{-1}) \mathcal{N}(u \mid \mu_\ell, C_\varphi) \pi(\lambda) \pi(\sigma^2 \varepsilon).
\]

- **Metropolis-within-Gibbs (MWG)**
  - Length scale \( u \) are updated individually.
  - When proposing \( u_k^* \), for \( k = 1, \ldots, n \), log-ratio of acceptance probability simplifies-(\( O(n) \) for SE and \( O(1) \) for AR).
  - When proposing hyperparameter \( \varphi^* \), we require: \( \log \left( \frac{\mathcal{N}(u\mid u_\ell, C_\varphi^s)}{\mathcal{N}(u\mid u_\ell, C_\varphi^s)} \right) \) - (\( O(n^3) \)) for SE and \( O(n) \) for AR).
  - Does not perform well for SE.

- **Whitened Elliptical Slice Sampling (w-ELL-SS)**
  - \( z = L(u)^{-1} \xi \) with \( \xi \sim \mathcal{N}(0, I_n) \) and \( u = R_\varphi \zeta + \mu_\ell \) with \( \zeta \sim \mathcal{N}(0, I_n) \).
  - \( \pi(\zeta, \xi, \lambda, \sigma^2 \varepsilon \mid y) \propto \mathcal{N}(y \mid AL(R_\varphi \zeta + \mu_\ell)^{-1} \xi, \sigma^2 \varepsilon I_m) \mathcal{N}(\xi \mid 0, I_n) \mathcal{N}(\zeta \mid 0, I_n) \pi(\lambda) \pi(\sigma^2 \varepsilon). \)
  - \( u \) updated jointly through \( \zeta \).
  - Likelihood can be evaluated as a product of univariate Gaussians.
  - \( z = L(u)^{-1} \xi \) can be solved in \( O(n) \)
  - \( u = R_\varphi \zeta + \mu_\ell \) - (\( O(n^2) \) for SE and \( O(n) \) for AR)
  - Each iteration may require several likelihood evaluations.
Marginal Elliptical Slice Sampling (m-ELL-SS)

\[ \pi(\zeta, \lambda, \sigma^2_\epsilon \mid y) \propto \mathcal{N}(y \mid 0, AQ_{\zeta,\varphi}^{-1}A^T + \sigma^2_\epsilon l_n)\mathcal{N}(\zeta \mid 0, l_m)\pi(\lambda)\pi(\sigma^2_\epsilon). \]

- \( u \) updated jointly through \( \zeta \).
- \( u = R_{\varphi} \zeta + \mu_\ell - (O(n^2) \text{ for SE and } O(n) \text{ for AR}) \)
- Marginal likelihood computation:

\[
\log \pi(y \mid u, \lambda, \sigma^2_\epsilon) = -\frac{m}{2} \log(2\pi) - \frac{1}{2} \log \det(\Psi) - \frac{1}{2} y^T \Psi^{-1} y
\]

where \( \Psi = AQ_u^{-1}A^T + \sigma^2_\epsilon l_m. \)

- Employ Woodbury identity for \( \Psi^{-1} \)
- Quadratic term: \( \sigma^{-2}_\epsilon \left( y^T y - y^T A \left( L(u)^T L(u) + \sigma^{-2}_\epsilon A^T A \right)^{-1} A^T y \right) \)
- Determinant computation is the dominant term \( (O(m^3) \text{ or } O(nm)) \)
- Improved mixing
Synthetic 1-d experiments

3 synthetic 1-dimensional examples

Figure: (a): 81 observations with domain $[0, 10]$ and $\sigma_\varepsilon^2 = 0.01$. (b): 350 observations with domain $[0, 8]$ and $\sigma_\varepsilon^2 = 0.04$. (c): 512 observations with domain $[0, 1]$ and $\sigma_\varepsilon^2 = 0.04$
Figure: MWG. (a)-(c): Estimated $\ell$ process with 95% credible intervals for AR(1) on different grids. (d)-(f): Estimated $z$ process with 95% credible intervals for AR(1) on different grids with observed data in red. (g)-(i): Estimated $\ell$ process with 95% credible intervals for SE on different grids. (j)-(l): Estimated $z$ process with 95% credible intervals for SE hyperprior on different grids with observed data in red.
Figure: Results for Experiment 1 at the highest resolution (n=253) for SE hyperprior with (left column) w-ELL-SS algorithm and (right column) m-ELL-SS algorithm.
Figure: Example 1: Traceplots with cumulative averages of the chains for SE hyperprior with $n = 253$. (Top row:) element of $\mathbf{u}$ with the lowest ESS. (Bottom row:) the hyperparameter.
OES = ESS/CPUtime

<table>
<thead>
<tr>
<th></th>
<th>MWG</th>
<th>w-ELL-SS</th>
<th>m-ELL-SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 85$</td>
<td>$n = 169$</td>
<td>$n = 253$</td>
</tr>
<tr>
<td>$\sigma^2_{\varepsilon}$</td>
<td>622.76</td>
<td>173.12</td>
<td>65.99</td>
</tr>
<tr>
<td>$\ell_{\min}$</td>
<td><strong>635.36</strong></td>
<td>114.02</td>
<td>41.05</td>
</tr>
<tr>
<td>$z_{\min}$</td>
<td><strong>203.80</strong></td>
<td>42.10</td>
<td>13.91</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>89.84</td>
<td>15.66</td>
<td>6.00</td>
</tr>
<tr>
<td>MAE</td>
<td>0.041</td>
<td>0.051</td>
<td>0.054</td>
</tr>
<tr>
<td>EC</td>
<td>0.988</td>
<td>0.975</td>
<td>0.971</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MWG</th>
<th>w-ELL-SS</th>
<th>m-ELL-SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n = 85$</td>
<td>$n = 169$</td>
<td>$n = 253$</td>
</tr>
<tr>
<td>$\sigma^2_{\varepsilon}$</td>
<td>11.19</td>
<td>4.88</td>
<td>7.49</td>
</tr>
<tr>
<td>$\ell_{\min}$</td>
<td>1.22</td>
<td>0.73</td>
<td>0.64</td>
</tr>
<tr>
<td>$z$</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.59</td>
<td>0.75</td>
<td>0.31</td>
</tr>
<tr>
<td>MAE</td>
<td>0.078</td>
<td>0.100</td>
<td>0.133</td>
</tr>
<tr>
<td>EC</td>
<td>0.889</td>
<td>0.826</td>
<td>0.763</td>
</tr>
</tbody>
</table>

**Table:** Experiment 1: OES with both hyperpriors under various discretisation schemes ($n = 86, 169, 253$) and three different algorithms. $\ell_{\min}$ and $z_{\min}$ report OES for the minimum ESS across all dimensions. Highest values in boldface.
• AR(1) hypermodel adds further computational gains.
• MWG performs poorly for highly correlated hyperprior.
• MWG deteriorates efficiency as the number of observations or discretisation size increase.
• w-ELL-SS for weak likelihoods performs well regardless the hyperprior employed at the price of highly correlated chains.
• Marginal sampler converges to the stationary distribution faster.
• m-ELL-SS good compromise between computational complexity and efficiency of the chains.
Extensions for two-dimensional problems

Employs additive Gaussian process models (AGP)

\[ y = A_1 z_1 + A_2 z_2 + A_3 z_3 + \varepsilon, \]

- \( A_1 \in \mathbb{R}^{m \times n_1} \), \( A_2 \in \mathbb{R}^{m \times n_2} \) and \( A_3 \in \mathbb{R}^{m \times (n_1 n_2)} \) known matrices.
- \( z_1(\cdot) \) and \( z_2(\cdot) \) independent univariate non-stationary processes.
- \( z_3(\cdot) \) is a bivariate, non-stationary, separable process -interaction term.
Hierarchical model:

\[
\begin{align*}
\mathbf{y} | \{\mathbf{z}_r\}_{r=1}^3, \sigma^2_\varepsilon & \sim \mathcal{N}(A_1 \mathbf{z}_1 + A_2 \mathbf{z}_2 + A_3 \mathbf{z}_3, \sigma^2_\varepsilon I_m) \\
\mathbf{z}_r | \phi_r & \sim \mathcal{N}(0, C_{\phi_r}^{\text{NS}}), \quad r = 1, 2, 3 \\
\mathbf{u}_s | \varphi_s & \sim \mathcal{N}(\mu_{\ell_s}, C_{\varphi_s}), \quad s = 1, 2, 3, 4 \\
(\sigma^2_\varepsilon, \varphi) & \sim \pi(\sigma^2_\varepsilon) \pi(\varphi_1) \pi(\varphi_2) \pi(\varphi_3) \pi(\varphi_4),
\end{align*}
\]

with \(\varphi = (\varphi_1, \ldots, \varphi_4)\).

- AGP works based on one-dimensional kernels
- Posterior:

\[
\begin{align*}
\pi(\{\mathbf{z}_r\}_{r=1}^3, \{\mathbf{u}_s, \lambda_s\}_{s=1}^4, \sigma^2_\varepsilon | \mathbf{y}) & \propto \mathcal{N}(\mathbf{y} | A_1 \mathbf{z}_1 + A_2 \mathbf{z}_2 + A_3 \mathbf{z}_3, \sigma^2_\varepsilon I_m) \mathcal{N}(\mathbf{z}_1 | 0, Q_{u_1}^{-1}) \\
& \quad \mathcal{N}(\mathbf{z}_2 | 0, Q_{u_2}^{-1}) \mathcal{N}(\mathbf{z}_3 | 0, Q_{u_3,4}^{-1}) \mathcal{N}(\mathbf{u}_1 | \mu_{\ell_1}, C_{\varphi_1}) \cdots \mathcal{N}(\mathbf{u}_4 | \mu_{\ell_4}, C_{\varphi_4}) \\
& \quad \pi(\lambda_1) \cdots \pi(\lambda_4) \pi(\sigma^2_\varepsilon),
\end{align*}
\]

with \(Q_{u_3,4}^{-1} := Q_{u_3}^{-1} \otimes Q_{u_4}^{-1}\).
Inference for two-dimensional problems

- Blocked Gibbs sampler, that updates the three blocks of parameters \((z_1, u_1, \lambda_1)\); \((z_2, u_2, \lambda_2)\); and \((z_3, u_3, u_4, \lambda_3, \lambda_4)\) from their full conditional distributions.

- **Block marginal elliptical slice sampler (Block-m-ELL-SS)**
  - To sample \((z_1, u_1, \lambda_1)\), the full conditional is factorised:
    \[
    \pi(z_1, \zeta_1, \lambda_1 \mid y, \sigma^2_\varepsilon, z_2, z_3) = \pi(\zeta_1, \lambda_1 \mid y, \sigma^2_\varepsilon, z_2, z_3)\pi(z_1 \mid \zeta_1, \lambda_1, y, \sigma^2_\varepsilon, z_2, z_3),
    \]
  - Interaction term: use eigendecompositions and matrix-vector multiplications for Kronecker matrices!
Figure: 2-dimensional synthetic data. $m = 20,449$ noisy observations in an expanded grid of $n_1 = n_2 = 143$ equally spaced points in $[0, 10]$, employing $z(x_1, x_2) = z(x_1) + z(x_2)$. 
Figure: Posterior mean surface and one-dimensional length-scale processes with 95% credible intervals.

- Capture smooth areas and edges.
- 2-level AGP correctly learns the varying correlation along the surface.
- 99.26 minutes for 50,000 iterations
Comparative evaluation

Figure: Each row shows one of the simulated experiments. Red dots depict observed data, dotted lines show the true signal, solid lines show the posterior mean, and grey areas depict 95% credible intervals. (a)(d)(g)(j): Stationary GP (b)(e)(h)(k): TGP, with blue dotted lines depicting MAP cut-off points. (c)(f)(i)(l): 2-level GP with m-ELL-SS algorithm and the hyperprior with lowest MAE.
Thank you