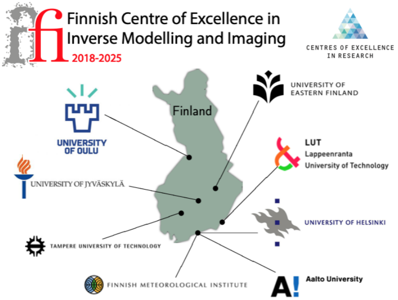
Posterior Inference for Sparse Hierarchical Non-stationary Models

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



Motivation

- 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

TV and Besov space 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 – α-stable priors

- Markku Markkanen, Lassi Roininen, Janne M J Huttunen and Sari Lasanen, Cauchy difference priors for edge-preserving Bayesian inversion, Journal of Ill-posed and Inverse Problems (2019).
- 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).
- Neil Chada, Sari Lasanen and Lassi Roininen, *Posterior Convergence Analysis of α-Stable Sheets*, arXiv (2019).
- Kenneth Muhumuza, Lassi Roininen, Janne M. J. Huttunen, Timo Lähivaara, *A Bayesian-based approach to improving acoustic Born waveform inversion of seismic data for viscoelastic media*, arXiv (2019).

Stable random walks

- Let $U(t), t \in \mathbb{I} \subset \mathbb{R}^+\}$ be a stochastic process. We call it a Lévy α-stable process starting from zero, or simply as stable process, if $U(0) = 0$, U has independent increments and
$$U(t) - U(s) \sim S_\alpha\left((t-s)^{1/\alpha}, \beta, 0\right) \tag{1}$$
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
$$U_{t_i} - U_{t_{i-1}} \sim S_\alpha(h^{\frac{1}{\alpha}}, \beta, 0)$$
where $t_i - t_{i-1} =: h$. It is easy to see that such random walk approximations converge to the α-stable Lévy motion as $h \to 0$ in distribution on the Skorokhod space of functions that are right-continuous and have left limits.

Chada, Lasanen, Roininen: α-stable sheet paper

- 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,

$$f(x,x') = \begin{cases} 1 \text{ when } x_i' \leq x_i \text{ for all } i = 1,\dots,d \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

- Stable integral is defined like the usual Itō 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= 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

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

where $0 < \alpha < 2$,

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

ρ_k is a Rademacher sequence (i.i.d. with values ± 1 with equal probabilities), Γ_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 ρ_k, Γ_k and V_k are mutually independent.

Discretisation

- Lévy-LePage series representation gives
  - 1) Sample path regularity in L^p, $1 \le p < \infty$ (also in the more general Sobolev space H_p^s, $s < 1/p$, $p \ge 2$),
  - 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

$$U^N(x) = U(h\lceil x/h \rceil), \tag{5}$$

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

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

$$\begin{aligned} U(hm_1, hm_2) - U(hm_1, h(m_2-1)) - U(h(m_1-1), hm_2) \\ + U(h(m_1-1), h(m_2-1)) \sim S_\alpha(|h|^{d/\alpha}, 0, 0) \end{aligned} \tag{6}$$

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

Convergence $h \to 0$

Theorem

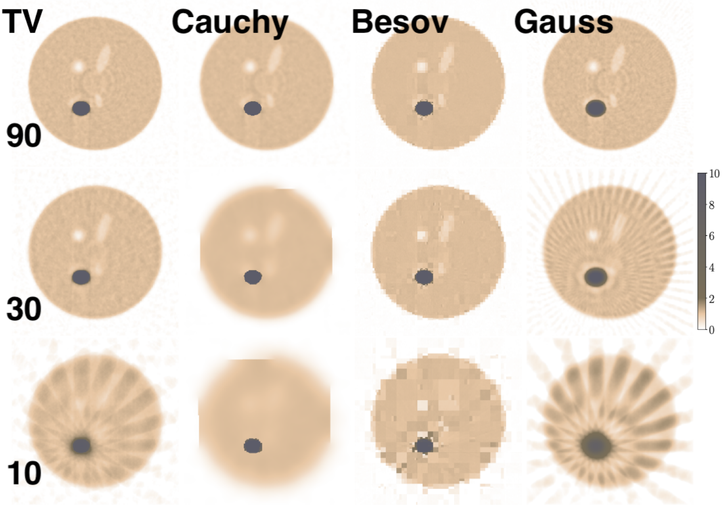
Let $1 \leq p < \infty$. The approximations $U^N(x) = U(h\lceil x/h \rceil)$ converge to U on $L^p((0,1)^d)$ in distribution.

Open questions:

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

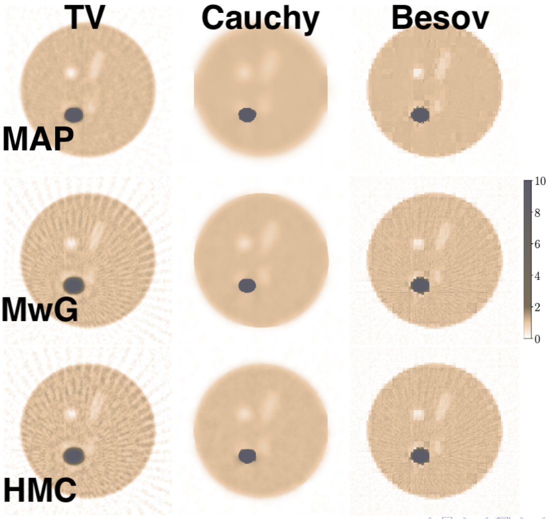
MAP estimates

- Log tomography with different number of projections and prior models



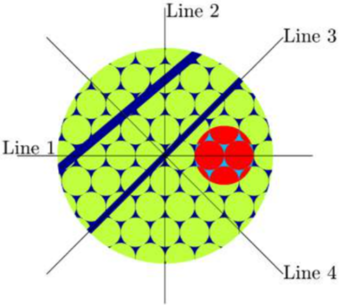
MAP and CM estimates – 30 projections

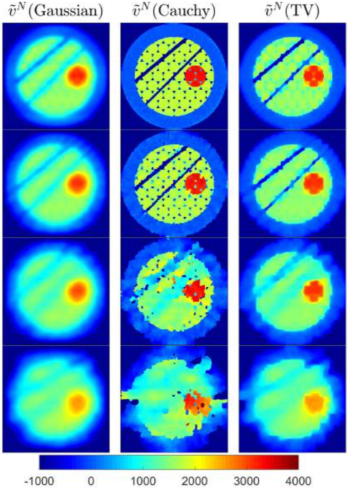
- Log tomography



Tomographic imaging of whole-core samples

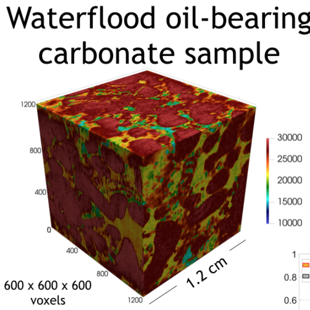
- 46, 23 ,12, 6 projections with 10% noise

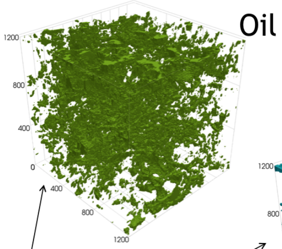


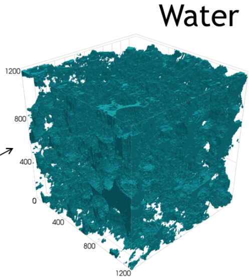


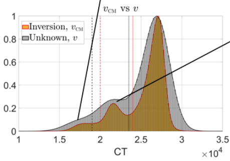
Micro-CT

Example 4: mixed-wet carbonate reservoir rocks from the Middle-East*.





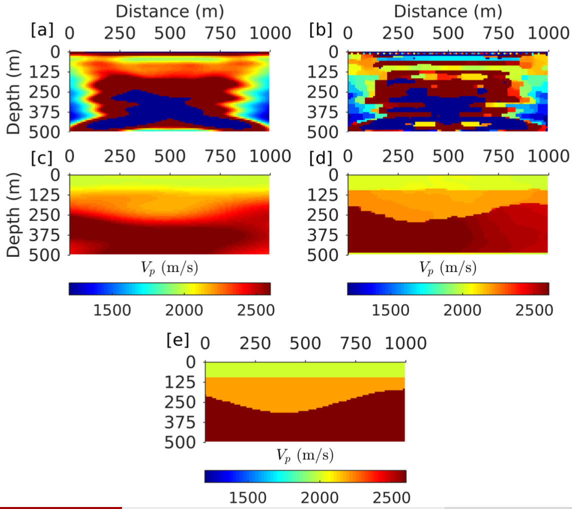




*Alhammadi, A., Alratrout, A., Bijeljic, B., & Blunt, M. (2018, May 17). X-ray micro-tomography datasets of mixed-wet carbonate reservoir rocks for in situ effective contact angle measurements. Retrieved August 16, 2018, from www.digitalrocksportal.org

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

Karla Monterrubio-Gómez, Lassi Roininen, Sara Wade, Theodoros Damoulas, and Mark Girolami, Posterior Inference for Sparse Hierarchical Non-stationary Models, arXiv 2019.

Hierachical GP model

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

$$
\begin{aligned}
y_i &\sim \mathcal{N}(z(x_i), \sigma_\varepsilon^2), \quad i = 1, \ldots, m, \\
z(\cdot) &\sim \mathcal{GP}\left(0, C_\phi^{\mathrm{NS}}(\cdot,\cdot)\right), \\
\log \ell(\cdot) &\sim \mathcal{GP}\left(\mu_\ell, C_\varphi^{\mathrm{S}}(\cdot,\cdot)\right), \\
(\tau^2, \varphi, \sigma_\varepsilon^2, \mu_\ell) &\sim \pi(\tau^2)\pi(\varphi)\pi(\sigma_\varepsilon^2)\pi(\mu_\ell),
\end{aligned} \tag{7}
$$

  where $C_\phi^{\mathrm{NS}}(\cdot,\cdot)$ denotes a non-stationary kernel, $C_\varphi^{\mathrm{S}}(\cdot,\cdot)$ is a stationary covariance function with parameters φ, and μ_ℓ 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 $\mathbf{z} \sim \mathcal{N}\left(0, C_\phi^{\mathrm{NS}}\right)$ equivalent to $\mathbf{z} \sim \mathcal{N}\left(0, (Q_\phi^{\mathrm{NS}})^{-1}\right)$
- 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 Δ is the Laplace operator, w is white noise on $\mathbb{R}$, $\mathrm{Var}(w) = \Gamma(\nu + 1/2)(4\pi)^{1/2} / \Gamma(\nu)$, and $\ell(\cdot)$ is a spatially varying length-scale.

- A finite-dimensional approximation can be written as

$$L(\boldsymbol{\ell})\mathbf{z} = \mathbf{w},$$

where $\mathbf{z} \in \mathbb{R}^n$ with n the discretisation size. $L(\boldsymbol{\ell})$ 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: $\mathbf{y} = A\mathbf{z} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_m), \quad A \in \mathbb{R}^{m\times n}$, $\mathbf{z} \in \mathbb{R}^n$.
- Hierarchical formulation

$$\begin{aligned} \mathbf{y} \mid \mathbf{z}, \sigma_\varepsilon^2 &\sim \mathcal{N}(A\mathbf{z}, \sigma_\varepsilon^2 I_m), \\ \mathbf{z} \mid \phi &\sim \mathcal{N}\left(0, Q_\phi^{-1}\right) \\ \log \boldsymbol{\ell} := \mathbf{u} \mid \boldsymbol{\varphi} &\sim \mathcal{N}\left(\boldsymbol{\mu}_\ell, C_\varphi\right) \\ (\tau^2, \sigma_\varepsilon^2, \boldsymbol{\varphi}, \mu_\ell) &\sim \pi(\tau^2)\pi(\sigma_\varepsilon^2)\pi(\boldsymbol{\varphi})\pi(\mu_\ell) \end{aligned} \tag{9}$$

  where $\boldsymbol{\mu}_\ell$ is the n-dimensional constant mean vector.
- Key component: $(C_\phi^{\mathrm{NS}})^{-1} := Q_\phi = L(\phi)^{\mathsf{T}} L(\phi)$, which depends on $\mathbf{u}$ and τ^2.
- $\boldsymbol{\varphi}$ parameters of the covariance that describe properties of the length-scales.

Hyperpriors

- **Stationary** assumption for spatially varying length-scale
- Explore two priors for $\mathbf{u}$:

  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 λ and magnitude τ_ℓ.

  AR(1):
  - Ornstein-Uhlenbeck covariance
  - Allows quick changes but is smoother than white noise.
  - Precision is **sparse** $Q_\varphi = L(\varphi)^{\mathsf{T}} L(\varphi)$, where $L(\varphi)$ is a banded matrix that depends on λ and τ_ℓ.
- To improve model identifiability, we fix τ, $\boldsymbol{\mu}_\ell$ and τ_ℓ.

Inference for one-dimensional problems

Posterior of intereset:

$$\pi(\mathbf{z}, \mathbf{u}, \lambda, \sigma_\varepsilon^2 \mid \mathbf{y}) \propto \mathcal{N}(\mathbf{y} \mid A\mathbf{z}, \sigma_\varepsilon^2 I_m)\mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_z, Q_\mathbf{u}^{-1})\mathcal{N}(\mathbf{u} \mid \boldsymbol{\mu}_\ell, C_\varphi)\pi(\lambda)\pi(\sigma_\varepsilon^2).$$

- **Metropolis-within-Gibbs (MWG)**
  - Length scale $\mathbf{u}$ are updated individually.
  - When proposing u_k^*, for $k = 1, \dots, n$, log-ratio of acceptance probability simplifies-($O(n)$ for SE and $O(1)$ for AR).
  - When proposing hyperparameter φ^*, we require: $\log\left(\frac{\mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_\ell, C^{\mathrm{S}}_{\varphi^*})}{\mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_\ell, C^{\mathrm{S}}_{\varphi})}\right)$ - ($O(n^3)$ for SE and $O(n)$ for AR).
  - Does not perform well for SE.

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

- **Marginal Elliptical Slice Sampling (m-ELL-SS)**

$$\pi(\zeta, \lambda, \sigma_\varepsilon^2 \mid \mathbf{y}) \propto \mathcal{N}(\mathbf{y} \mid 0, AQ_{\zeta,\varphi}^{-1}A^\mathsf{T} + \sigma_\varepsilon^2 I_n)\mathcal{N}(\zeta \mid 0, I_m)\pi(\lambda)\pi(\sigma_\varepsilon^2).$$

  - $\mathbf{u}$ updated jointly through ζ.
  - $\mathbf{u} = R_\varphi \zeta + \boldsymbol{\mu}_\ell$ - $\left(O(n^2)\right.$ for SE and $\left.O(n)\right.$ for AR$)$
  - Marginal likelihood computation:

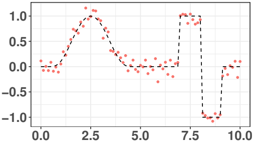
$$\log \pi(\mathbf{y} \mid \mathbf{u}, \lambda, \sigma_\varepsilon^2) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log\det(\Psi) - \frac{1}{2}\mathbf{y}^\mathsf{T}\Psi^{-1}\mathbf{y}$$

  where $\Psi = AQ_{\mathbf{u}}^{-1}A^\mathsf{T} + \sigma_\varepsilon^2 I_m$.

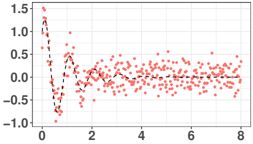
  - Employ Woodbury identity for Ψ^{-1}
  - Quadratic term: $\sigma_\varepsilon^{-2}\left(\mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}A\left(L(\mathbf{u})^\mathsf{T}L(\mathbf{u}) + \sigma_\varepsilon^{-2}A^\mathsf{T}A\right)^{-1}A^\mathsf{T}\mathbf{y}\right)$
  - Determinant computation is the dominant term $\left(O(m^3)\right.$ or $\left.O(nm)\right)$
  - Improved mixing

Synthetic 1-d experiments

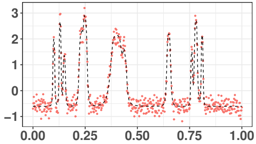
3 synthetic 1-dimensional examples



(a) Experiment 1

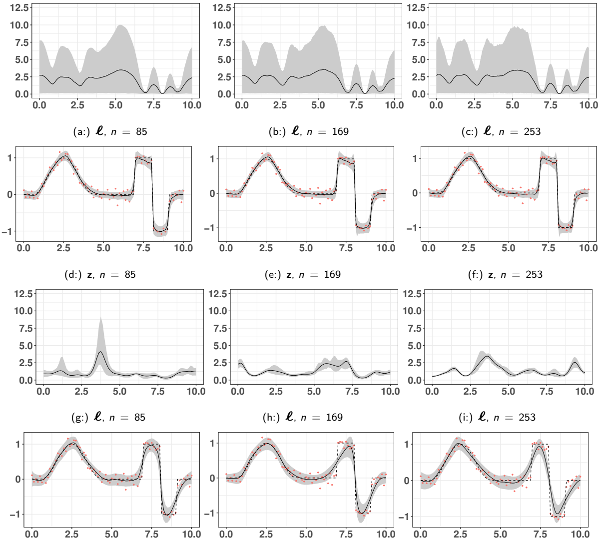


(b) Experiment 2



(c) Experiment 3

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 $\sigma_\varepsilon^2 = 0.04$



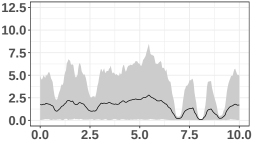
(a:) $\boldsymbol{\ell}$, $n = 85$ (b:) $\boldsymbol{\ell}$, $n = 169$ (c:) $\boldsymbol{\ell}$, $n = 253$

(d:) $\mathbf{z}$, $n = 85$ (e:) $\mathbf{z}$, $n = 169$ (f:) $\mathbf{z}$, $n = 253$

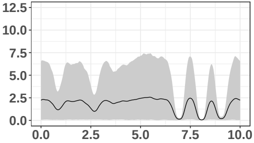
(g:) $\boldsymbol{\ell}$, $n = 85$ (h:) $\boldsymbol{\ell}$, $n = 169$ (i:) $\boldsymbol{\ell}$, $n = 253$

(j:) $\mathbf{z}$, $n = 85$ (k:) $\mathbf{z}$, $n = 169$ (l:) $\mathbf{z}$, $n = 253$

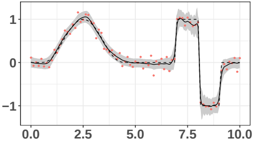
Figure: MWG. (a)-(c): Estimated $\boldsymbol{\ell}$ process with 95% credible intervals for AR(1) on different grids. (d)-(f): Estimated $\mathbf{z}$ process with 95% credible intervals for AR(1) on different grids with observed data in red. (g)-(i): Estimated $\boldsymbol{\ell}$ process with 95% credible intervals for SE on different grids. (j)-(l): Estimated $\mathbf{z}$ process with 95% credible intervals for SE hyperprior on different grids with observed data in red.



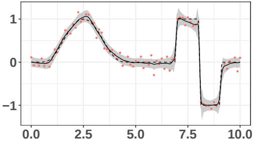
(a) w-ELL-SS $\boldsymbol{\ell}$, $n = 253$



(b) m-ELL-SS $\boldsymbol{\ell}$, $n = 253$

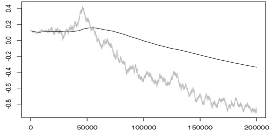


(c) w-ELL-SS $\boldsymbol{\ell}$, $n = 253$

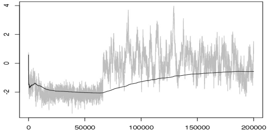


(d) m-ELL-SS $\mathbf{z}$, $n = 253$

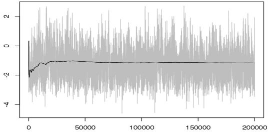
Figure: Results for Experiment 1 at the highest resolution (n=253) for SE hyperprior with (left column) w-ELL-SS algorithm and (right column) m-ELL-SS algorithm.



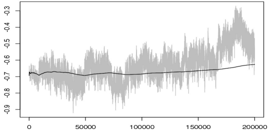
(a) u_{199}, MWG



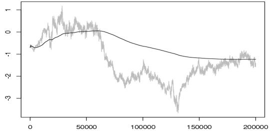
(b) u_{185}, w-ELL-SS



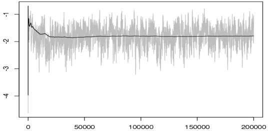
(c) u_{190}, m-ELL-SS



(d) λ, MWG



(e) λ, w-ELL-SS



(f) λ, m-ELL-SS

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

		MWG			w-ELL-SS			m-ELL-SS		
		$n = 85$	$n = 169$	$n = 253$	$n = 85$	$n = 169$	$n = 253$	$n = 85$	$n = 169$	$n = 253$
AR(1)	σ_ε^2	622.76	173.12	65.99	380.89	102.38	38.91	**661.20**	**257.81**	**116.35**
	ℓ_{min}	**635.36**	114.02	41.05	30.90	8.99	2.94	287.16	**114.36**	**59.71**
	z_{min}	**203.80**	42.10	13.91	9.12	2.34	0.86	129.75	**52.16**	**22.30**
	λ	89.84	15.66	6.00	22.77	5.26	2.36	**111.80**	**45.54**	**21.53**
	MAE	0.041	0.051	0.054	0.041	0.051	0.054	0.041	0.051	0.053
	EC	0.988	0.975	0.971	0.988	0.975	0.975	0.988	0.975	0.975
SE	σ_ε^2	11.19	4.88	7.49	246.24	77.72	8.89	**856.15**	**253.91**	**125.97**
	ℓ_{min}	1.22	0.73	0.64	21.69	10.22	2.79	**244.91**	**122.57**	**55.82**
	z	0.06	0.01	0.01	4.71	1.37	0.24	**76.80**	**24.11**	**9.87**
	λ	0.59	0.75	0.31	2.31	0.29	0.01	**16.59**	**4.15**	**2.21**
	MAE	0.078	0.100	0.133	0.040	0.050	0.054	0.039	0.049	0.052
	EC	0.889	0.826	0.763	0.988	0.975	0.971	0.988	0.975	0.979

Table: Experiment 1: OES with both hyperpriors under various discretisation schemes ($n = 86, 169, 253$) and three different algorithms. ℓ_{min} and z_{min} report OES for the minimum ESS across all dimensions. Highest values in boldface.

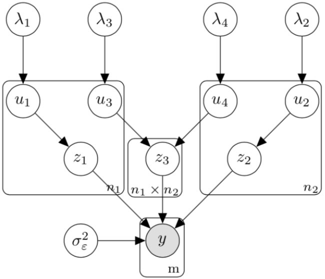
- AR(1) hypermodel adds further computational gains.
- MWG performs poorly for highly correlated hyperprior.
- MWG deteriorates efficiency as the number of observations or discretisation size increase.
- w-ELL-SS for weak likelihoods performs well regardless the hyperprior employed at the price of highly 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)

$$\mathbf{y} = A_1\mathbf{z}_1 + A_2\mathbf{z}_2 + A_3\mathbf{z}_3 + \varepsilon,$$

- $A_1 \in \mathbb{R}^{m\times n_1}$, $A_2 \in \mathbb{R}^{m\times n_2}$ 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:

$$
\begin{aligned}
\mathbf{y} \mid \{\mathbf{z}_r\}_{r=1}^3, \sigma_\varepsilon^2 &\sim \mathcal{N}(A_1\mathbf{z}_1 + A_2\mathbf{z}_2 + A_3\mathbf{z}_3, \sigma_\varepsilon^2 I_m) \\
\mathbf{z}_r \mid \phi_r &\sim \mathcal{N}\left(0, C_{\phi_r}^{\mathrm{NS}}\right), \quad r = 1,2,3 \\
\mathbf{u}_s \mid \varphi_s &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\ell_s}, C_{\varphi_s}^{\mathrm{S}}\right), \quad s = 1,2,3,4 \\
(\sigma_\varepsilon^2, \varphi) &\sim \pi(\sigma_\varepsilon^2)\pi(\varphi_1)\pi(\varphi_2)\pi(\varphi_3)\pi(\varphi_4),
\end{aligned}
$$

with $\varphi = (\varphi_1, \dots, \varphi_4)$.

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

$$
\begin{aligned}
\pi(\{\mathbf{z}_r\}_{r=1}^3, \{\mathbf{u}_s, \lambda_s\}_{s=1}^4, \sigma_\varepsilon^2 \mid \mathbf{y}) \propto \mathcal{N}(\mathbf{y} \mid A_1\mathbf{z}_1 + A_2\mathbf{z}_2 + A_3\mathbf{z}_3, \sigma_\varepsilon^2 I_m)\mathcal{N}(\mathbf{z}_1 \mid 0, Q_{\mathbf{u}_1}^{-1}) \\
\mathcal{N}(\mathbf{z}_2 \mid 0, Q_{\mathbf{u}_2}^{-1})\mathcal{N}(\mathbf{z}_3 \mid 0, Q_{\mathbf{u}_{3,4}}^{-1})\mathcal{N}(\mathbf{u}_1 \mid \boldsymbol{\mu}_{\ell_1}, C_{\varphi_1}) \cdots \mathcal{N}(\mathbf{u}_4 \mid \boldsymbol{\mu}_{\ell_4}, C_{\varphi_4}) \\
\pi(\lambda_1) \cdots \pi(\lambda_4)\pi(\sigma_\varepsilon^2),
\end{aligned}
$$

with $Q_{\mathbf{u}_{3,4}}^{-1} := Q_{\mathbf{u}_3}^{-1} \otimes Q_{\mathbf{u}_4}^{-1}$

Inference for two-dimensional problems

- Blocked Gibbs sampler, that updates the three blocks of parameters $(\mathbf{z}_1, \mathbf{u}_1, \lambda_1)$; $(\mathbf{z}_2, \mathbf{u}_2, \lambda_2)$; and $(\mathbf{z}_3, \mathbf{u}_3, \mathbf{u}_4, \lambda_3, \lambda_4)$ from their full conditional distributions.
- **Block marginal elliptical slice sampler (Block-m-ELL-SS)**
  - To sample $(\mathbf{z}_1, \mathbf{u}_1, \lambda_1)$, the full conditional is factorised:

$$\pi(\mathbf{z}_1, \boldsymbol{\zeta}_1, \lambda_1 \mid \mathbf{y}, \sigma_\varepsilon^2, \mathbf{z}_2, \mathbf{z}_3) = \pi(\boldsymbol{\zeta}_1, \lambda_1 \mid \mathbf{y}, \sigma_\varepsilon^2, \mathbf{z}_2, \mathbf{z}_3)\pi(\mathbf{z}_1 \mid \boldsymbol{\zeta}_1, \lambda_1, \mathbf{y}, \sigma_\varepsilon^2, \mathbf{z}_2, \mathbf{z}_3),$$

  - Interaction term: use eigendecompositions and matrix-vector multiplications for Kronecker matrices!

Synthetic 2-d experiment

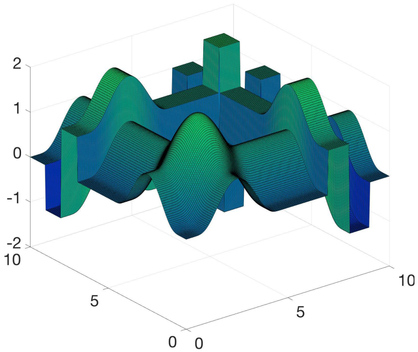


Figure: 2-dimensional synthetic data. $m = 20,449$ noisy observations in an expanded grid of $n_1 = n_2 = 143$ equally spaced points in $[0, 10]$, employing $z(x_1, x_2) = z(x_1) + z(x_2)$.

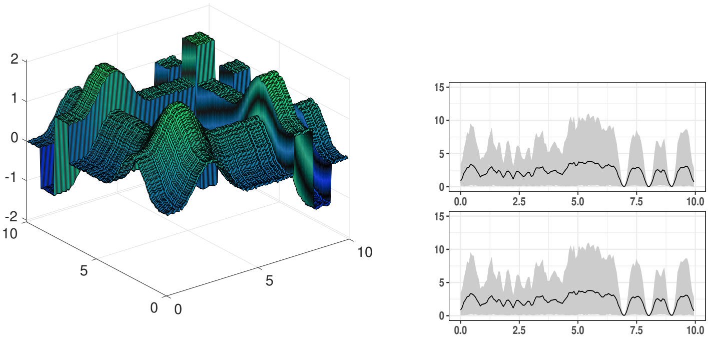
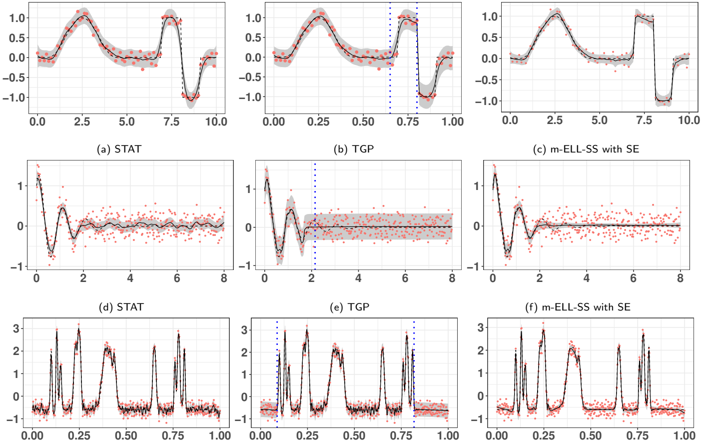


Figure: Posterior mean surface and one-dimensional length-scale processes with 95% credible intervals.

- 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



(a) STAT (b) TGP (c) m-ELL-SS with SE

(d) STAT (e) TGP (f) m-ELL-SS with SE

(g) STAT (h) TGP (i) m-ELL-SS with SE

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