A bivariate Gaussian Markov Random Field for large datasets

M. Molinaro, R. Furrer

Institute for Mathematics, University of Zurich, Switzerland

mattia.molinaro@math.uzh.ch, reinhard.furrer@math.uzh.ch

Universität Zürich[UZH]

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
SNF 143282

Bivariate spatial data on a regular grid

General Circulation Models (GCMs) provide long–term forecasts of Earth's climate (*precipitation* and *temperature*). However, modern datasets are "large". For instance: CMIP5 resolution of 256×128, satellite images are way bigger... Hierarchical linear models account for several sources of uncertainty:

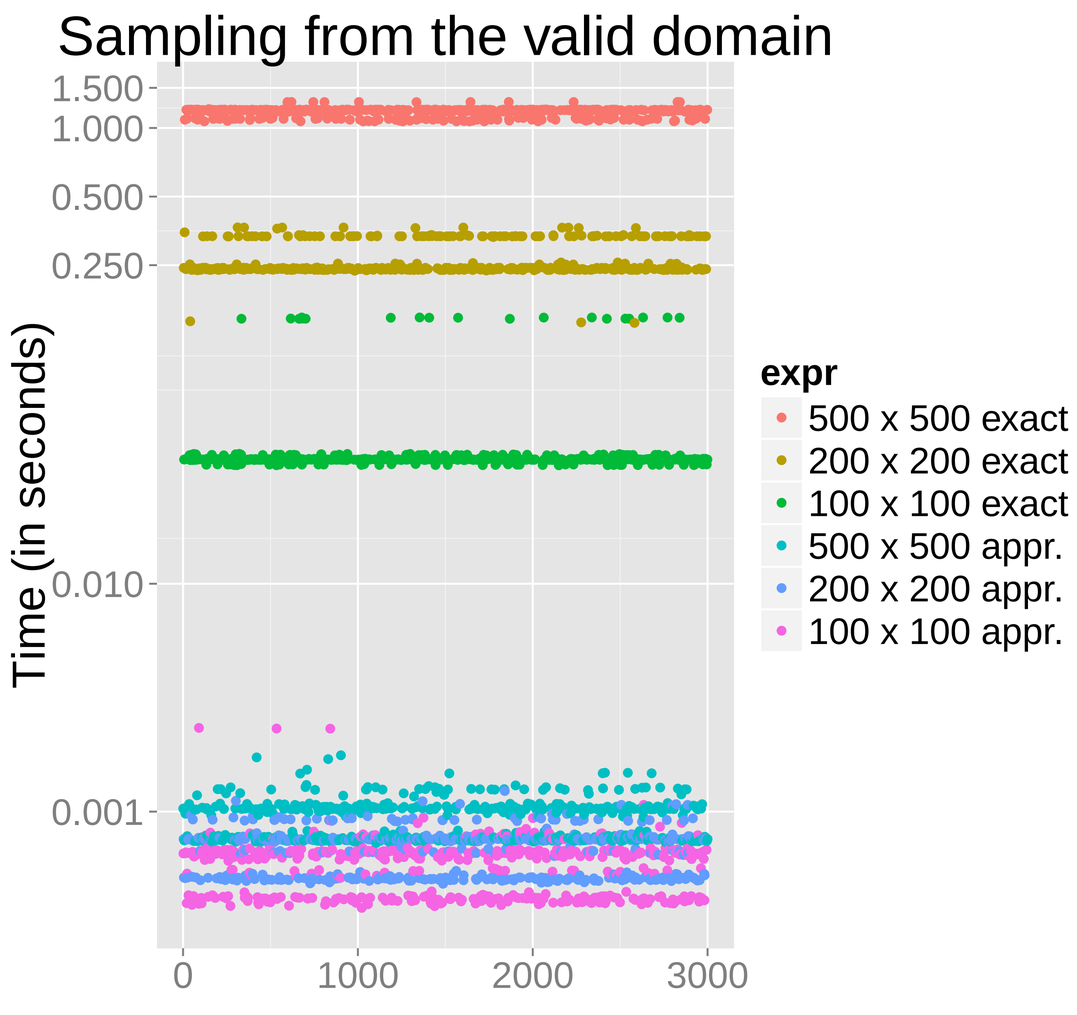
- *dependence structure* between the GCMs
- *complicated spatial dependence* between different locations

We deal with the latter point through a bivariate Gaussian Markov Random Field (GMRF) $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_{2n}^{-1}(\boldsymbol{\theta}))$ over a regular lattice of size n. The field $\boldsymbol{x}$ depends on $\boldsymbol{\theta} \in \mathbb{R}^5$. Computational efficiency is vital in MCMC samplers, where we usually want to run hundreds of thousands of iterations. This is going to be the leitmotiv of what follows.

Challenges in finding the valid parameter space

In order to sample from $\boldsymbol{x}$, $\boldsymbol{Q}_{2n}(\boldsymbol{\theta})$ must be non negative–definite. We find asymptotically closed–form formulas for the valid parameter space through a suitable approximation $\widetilde{\boldsymbol{Q}}_{2n}(\boldsymbol{\theta})$

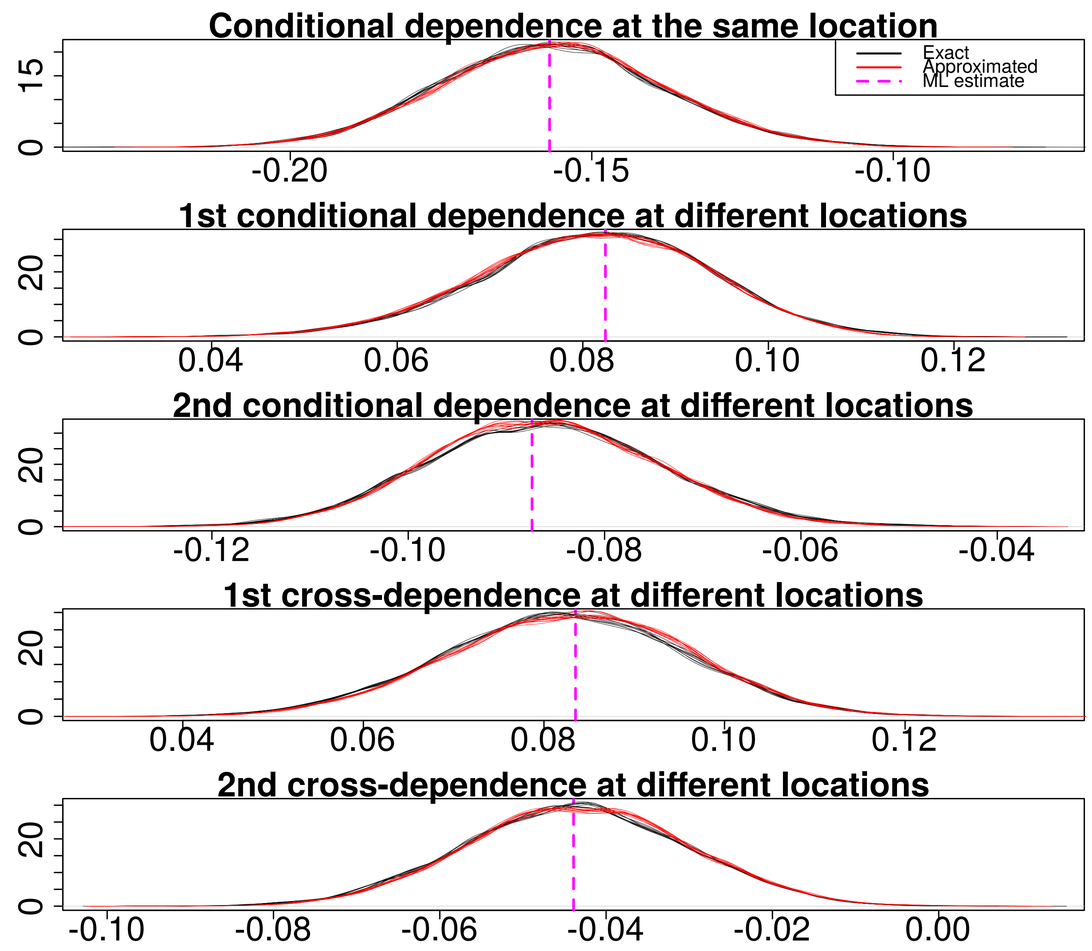
- complicated *geometrical description*
- $\lambda_{\min}(\widetilde{\boldsymbol{Q}}_{2n}(\boldsymbol{\theta})) \leq \lambda_{\min}(\boldsymbol{Q}_{2n}(\boldsymbol{\theta})) \rightsquigarrow$ no "false positives"
- no need to store $\boldsymbol{Q}_{2n}(\boldsymbol{\theta})$
- *computational efficiency* compared to the `spam` routines for sparse matrices



The "walking monkey" and the "running horse"

Efficient approximations to compute $-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x}$ and $\log(\det(\boldsymbol{Q}(\boldsymbol{\theta}))) \rightsquigarrow$ implementation of two MCMC samplers and an efficient ML algorithm.
Here we compare an exact (the "walking monkey") and an approximated sampler (the "running horse").

- Uniform prior on $\boldsymbol{\theta}$. Blocking strategy: three Metropolis–Hastings steps with *truncated* proposal distributions
- grid of size 100×100, same simulated dataset, different starting points
- after burn–in, two chains of length $50,000$. Ten different thinning strategies are considered to assess the within sampler variability



The 'Potential scale reduction factor' is at most 1.01 for each parameter, with a coverage probability of 0.99. The 'Multivariate psrf' is 1. In addition, no evidence of problematic cross-correlation between the different subchains was found.

What's next?

- the field $\boldsymbol{x}$ is "stationary" $\rightsquigarrow$ add a further layer to the considered model
- theoretical challenges given by the Brook's lemma
- MCMC drawbacks $\rightsquigarrow$ use `INLA`?