SpatialGEV: Fast Bayesian inference for spatial extreme value models in R

Meixi Chen [1][¶], **Martin Lysy** [1], **and Reza Ramezan** [1]

1 Department of Statistics and Actuarial Science, University of Waterloo ¶ Corresponding author

Summary

Extreme weather phenomena such as floods and hurricanes are of great concern due to their potential to cause extensive damage. To develop more reliable damage prevention protocols, statistical models are often used to infer the chance of observing an extreme weather event at a given location (Coles & Casson, 1998; Cooley et al., 2007; Sang & Gelfand, 2010). Here we present **SpatialGEV**, an R package providing a fast and convenient toolset for analyzing spatial extreme values using a hierarchical Bayesian modeling framework. In this framework, the marginal behavior of the extremes is given by a generalized extreme value (GEV) distribution, whereas the spatial dependence between locations is captured by modeling the GEV parameters as spatially varying random effects following a Gaussian process (GP). Model inference is carried out using an efficient implementation of the Laplace approximation, which produces highly accurate posterior estimates several orders of magnitude faster than Markov Chain Monte Carlo (MCMC) methods. Users are provided with a streamlined way to build and fit various GEV-GP models in R, which are compiled in C++ under the hood. For downstream analyses, the package offers methods for Bayesian parameter estimation and forecasting of extreme events.

Background

Let y_{ij} denote observation j of an extreme weather event at spatial location i, of which the two-dimensional spatial coordinates are x_i. The general form of the GEV-GP models that can be fit using **SpatialGEV** is

$$
\begin{aligned}
y_{ij} &\overset{\text{iid}}{\sim} \text{GEV}(a(x_i), \exp(b(x_i)), s(x_i)), \\
u(x) &\sim \text{GP}(c_u(x)^T \beta_u(x), \ \mathcal{K}(x, x' \mid \eta_u)),
\end{aligned}
\tag{1}
$$

where $\text{GEV}(a, b, s)$ denotes a GEV distribution with location parameter a, positive scale parameter b and shape parameter s, $\text{GP}(\mu(x), \mathcal{K}(x, x'))$ is a Gaussian process with mean function $\mu(x)$ and covariance kernel $\mathcal{K}(x, x')$, and $u \in \{a, b, s\}$ is any subset of the GEV parameters to be modelled as spatially varying. The GEV-GP model has important applications in meteorological studies. For example, to forecast extreme rainfalls, it is often of interest for meteorologists to estimate the $1/p$-year rainfall return level $z_p(x)$, which is the value above which precipitation levels at location x occur with probability p, i.e.,

$$
\Pr(y(x) > z_p(x)) = 1 - F(z_p(x) \mid a(x), b(x), s(x)) = p,
\tag{2}
$$

where $F(z_p(x) \mid a(x), b(x), s(x))$ is the cumulative density function of the GEV distribution specific to location x.

Statement of need

In a Bayesian context, the posterior distribution $p(z_p(x) \mid Y)$ conditional on all data Y is very useful for forecasting extreme weather events. Traditionally, MCMC methods are used to sample from the posterior distribution of the GEV model (e.g., Cooley et al., 2007; Dyrrdal et al., 2015; Schliep et al., 2010). However, this can be extremely computationally intensive when the number of locations is large. **SpatialGEV** implements an approximate Bayesian inference approach as an alternative to MCMC, making large-scale spatial analyses orders of magnitude faster while achieving roughly the same accuracy as MCMC. We construct a Normal approximation to the joint posterior distribution of both GEV parameters and GP hyperparameters $p(u(x), \eta_u, \beta_u \mid Y)$, which is then used to estimate the return level posterior. This is done via an efficient Laplace approximation to the marginal hyperparameter posterior $p(\eta_u, \beta_u \mid Y)$ transforming a high-dimensional MCMC into a nested optimization problem that is faster to solve (Chen et al., 2024; Kristensen et al., 2016; Tierney & Kadane, 1986). The Laplace approximation is carried out using the R/C++ package **TMB** (Kristensen et al., 2016). Details of the inference method can be found in Chen et al. (2024).

The R package **SpatialExtremes** (Ribatet et al., 2022) is a popular software for fitting spatial extreme value models including GEV-GP. Although it supports a wider range of model classes, its inference for GEV-GP models relies on a basic Gibbs sampler updating each of the hyperparameters and random effects one at a time, which tends to converge very slowly since these variables are highly correlated with each other. Furthermore, GP computation in **SpatialExtremes** scales as $\mathcal{O}(n^3)$ with the number of locations n, whereas **SpatialGEV** offers an option for approximate GP computation scaling as $O(n^{3/2})$ (Lindgren et al., 2011). Coupled with the Laplace approximation, this allows **SpatialGEV** to fit GEV-GP models to several hundreds spatial locations on a personal computer in minutes (Chen et al., 2024). A more efficient MCMC algorithm for hierarchical spatial models is Hamiltonian Monte Carlo and its variants (Hoffman & Gelman, 2014; Neal, 2011), for which a highly efficient and self-tuning implementation is provided by the R/C++ package **RStan** (Stan Development Team, 2020). Chen et al. (2024) compares the speed and accuracy of the Laplace method implemented in **SpatialGEV** to **RStan**. It is found that, while **SpatialGEV** tends to underestimate the posterior variance of the hyperparameters, it accurately estimates the posteriors of both GEV parameters and return levels – and does this three orders of magnitude faster than **RStan**. A well-known alternative to MCMC is the integrated Laplace approximation (INLA) method, whose R implementation is provided in the **R-INLA** package (Lindgren & Rue, 2015). As an extension of the Laplace approximation, INLA is typically more accurate. However, **R-INLA** is inapplicable to GEV-GP models in which two or more GEV parameters are modeled as random effects following different Gaussian processes. In contrast, **SpatialGEV** offers more flexibility as it is straightforward for the user to choose what GEV parameters are spatial random effects. Wood (2023) and Youngman (2022) provide another means for estimating spatially varying GEV parameters via a scalable basis representation reducing the number of random effects in the model. Compared to **SpatialGEV** which keeps all random effects for inference, the basis function expansion approach is less accurate for estimating spatial processes that are not smooth or exhibit short-range correlation (Lindgren et al., 2021; Wood, 2020).

Example

The main functions of the **SpatialGEV** package are `spatialGEV_fit()`, `spatialGEV_sample()`, and `spatialGEV_predict()`. This example shows how to apply these functions to analyze a simulated dataset with 400 locations using the GEV-GP model. The GEV location parameter $a(x)$ and the scale parameter $b(x)$ are generated from surfaces depicted in Figure 1, whereas the shape parameter s is a constant $\exp(-2)$ across space. The simulated data is provided by the package as a list called `simulatedData`, whose values are calibrated to the level of total daily precipitation in mm.
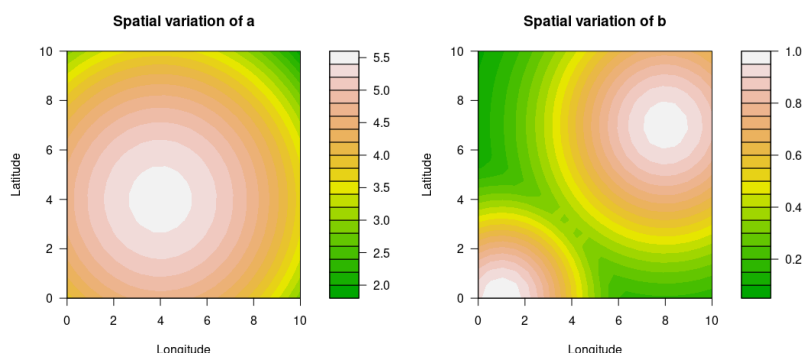
Figure 1: The simulated GEV location parameters $a(x_i)$ and scale parameters $(b(x_i))$ plotted on regular lattices.

The GEV-GP model is fit by calling `spatialGEV_fit()`. By specifying `random="ab"`, only the GEV parameters a and b are considered spatiallying varying. The argument `reparam_s="positive"` means we constrain the shape parameter to be positive. The GP kernel function is chosen using `kernel="exp"`. Other kernel function options are the Matérn kernel and the approximate GP computation method employing an SPDE approximation to the Matérn (Lindgren et al., 2011). Covariates to include in the mean functions can be provided in a matrix form to X_{a/b}. Further details about the arguments are given in the vignette.

```r
set.seed(123)
library(SpatialGEV)
n_loc <- 50                            # number of locations
locs <- simulatedData$locs[1:n_loc,]  # location coordinates
a <- simulatedData$a[1:n_loc]         # true GEV location parameters
logb <- simulatedData$logb[1:n_loc]   # true GEV (log) scale parameters
logs <- simulatedData$logs            # true GEV (log) shape parameter
y <- simulatedData$y[1:n_loc]         # simulated observations
# Model fitting
fit <- spatialGEV_fit(data = y, locs = locs, random = "ab",
                      init_param = list(a = rep(4, n_loc),
                                        log_b = rep(0,n_loc),
                                        s = -2,
                                        beta_a = 4, beta_b = 0,
                                        log_sigma_a = 0, log_ell_a = 1,
                                        log_sigma_b = 0, log_ell_b = 1),
                      reparam_s = "positive", kernel="exp",
                      X_a = matrix(1, nrow=n_loc, ncol=1),
                      X_b = matrix(1, nrow=n_loc, ncol=1),
                      silent=T)
print(fit)
#> Model fitting took 8.78002285957336 seconds
#> The model has reached relative convergence
#> The model uses a exp kernel
#> Number of fixed effects in the model is 7
#> Number of random effects in the model is 100
#> Hessian matrix is positive definite.
#> Use spatialGEV_sample to obtain posterior samples
```

Posterior sampling from the joint posterior distribution of the GEV parameters using the function `spatialGEV_sample()`, which takes in the list output by `spatialGEV_fit()` as an argument.

```
sam <- spatialGEV_sample(model = fit, n_draw = 2000, observation = TRUE)
print(sam)
#> The samples contains 2000 draws of 107 parameters
#> The samples contains 2000 draws of response at 50 locations
#> Use summary() to obtain summary statistics of the samples
pos_summary <- summary(sam)
```

The samples are then used to calculate the posterior mean estimate of the 10-year return level at each location, e.g., using the qgev() function from the **evd** package (Stephenson, 2002). See Figure 2.
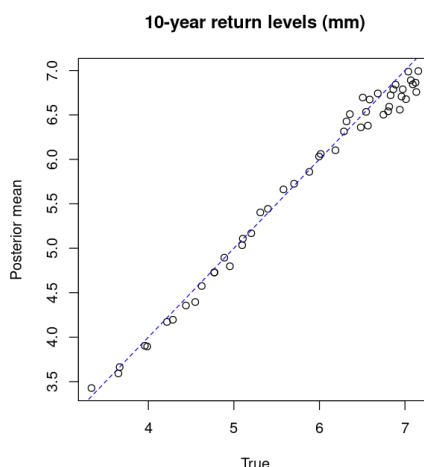
10-year return levels (mm)

Figure 2: Posterior mean estimates of the 10-year return level $z_{10}(x)$ plotted against the true values at different locations.

Conclusion

As shown in Figure 2, and based on other evidence in Chen et al. (2024), **SpatialGEV** provides approximate Bayesian inference methods for fitting a flexible class of spatial extreme value models. Its computational efficiency over MCMC makes it suitable for analyzing large datasets such as extreme weather data.

Acknowledgements

References

Chen, M., Ramezan, R., & Lysy, M. (2024). Fast and scalable approximate inference for spatial extreme value models. *Canadian Journal of Statistics (Accepted)*. https://doi.org/10.1002/cjs.11829

Coles, S. G., & Casson, E. (1998). Extreme value modelling of hurricane wind speeds. *Structural Safety*, *20*, 283–296. https://doi.org/10.1111/1467-9876.00189

Cooley, D., Nychka, D., & Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, *102*, 824–840. https://doi.org/10.1198/016214506000000780

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., & Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in Norway. *Environmetrics*, *26*, 89–106. https://doi.org/10.1002/env.2301

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. https://doi.org/10.5555/2627435.2638586

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*(5), 1–21. https://doi.org/10.18637/jss.v070.i05

Lindgren, F. K., Bolin, D., & Rue, H. (2021). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*. https://doi.org/10.1016/j.spasta.2022.100599

Lindgren, F. K., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*, 1–25. https://doi.org/10.18637/jss.v063.i19

Lindgren, F. K., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B*, *73*, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *The handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press.

Ribatet, M., Singleton, R., & R Core team. (2022). *SpatialExtremes: Modelling spatial extremes*. Version 2.1-0. https://CRAN.R-project.org/package=SpatialExtremes.

Sang, H., & Gelfand, A. E. (2010). Continuous spatial process models for spatial extreme values. *Journal of Agricultural, Biological, and Environmental Statistics*, *15*, 49–56. https://doi.org/10.1007/s13253-009-0010-1

Schliep, E. M., Cooley, D., Sain, S. R., & Hoeting, J. A. (2010). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, *13*, 219–239. https://doi.org/10.1007/s10687-009-0098-2

Stan Development Team. (2020). *RStan: The R interface to Stan*. http://mc-stan.org/

Stephenson, A. G. (2002). Evd: Extreme value distributions. *R News*, *2*(2), 31–32.

Tierney, L., & Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86. https://doi.org/10.1080/01621459.1986.10478240

Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. *Test*, *29*, 307–339. https://doi.org/10.1007/s11749-020-00711-5

Wood, S. N. (2023). *mgcv: Mixed GAM computation vehicle with automatic smoothness estimation*. Version 1.9-1. https://CRAN.R-project.org/package=mgcv.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, *103*, 1–26. https://doi.org/10.18637/jss.v103.i03