Earthquake forecasting

Finn Lindgren

2023-07-03

Project leader: Finn Lindgren
Project helpers: Daniel Paulin, Man Ho Suen (School of Mathematics)
Experts: Francesco Serafini (Newcastle University) and Mark Naylor (School of Geosciences)
Session: 2 (10 July – 11 August)

Background

Even though prediction of individual earthquakes is an elusive unsolved problem, by combining geophysical insight with statistical modelling, it is possible to construct probabilistic forecasts of earthquake activity. When constructed for specific regions, such models may, for example, include information about geological structures, historical information about seismic activity, and a stochastic model for how “mainshocks” can generate an “aftershock sequence”. A classic statistical tool for modelling temporal and spatial point data (in this case times, locations, and magnitudes of earthquakes) is the Poisson process model, where points are assumed to be independent, conditionally on an intensity process. For earthquakes with aftershocks, we need to extend this to model dependent points, via a Hawkes process model, that allows the intensity of future earthquakes depend on the occurrence of previous earthquakes.

A fundamental question for all forecasting models is how good they are at forecasting the future, and how much, and what type, of data is needed to reliably estimate the needed model parameters. The aim of this project is to investigate some aspects of this problem, for some models developed at the University of Edinburgh. To investigate this, simulated data may be used, to compare different scenarios where the true models are known. In a purely temporal setting, one can for example generate predefined mainshocks and stochastic aftershock sequences, and compare model estimates and forecasts under different settings.

Software for specifying and estimating the models is available in R, built on top of the INLA and inlabru packages. For forecast assessment, R can be used as well, but there’s also a possibility of using the pyCSEP Python package for this purpose.

Useful courses

Generalised Regression Models, Bayesian Data Analysis, R skills needed, and optionally some Python knowledge

References

The main papers about estimating Hawkes process models with inlabru:

1. Francesco Serafini et al (2023), Approximation of Bayesian Hawkes process with inlabru, `https://arxiv.org/abs/2206.13360` and `https://onlinelibrary.wiley.com/doi/full/10.1002/env.2798`
2. Mark Naylor et al (2023), Bayesian modeling of the temporal evolution of seismicity using the ETAS.inlabru package, `https://arxiv.org/abs/2212.06077` and `https://www.frontiersin.org/articles/10.3389/fams.2023.1126759/full`

Further reading materials

Papers about related estimation methods and modelling aspects

1. Gordon J. Ross (2021), Bayesian Estimation of the ETAS Model for Earthquake Occurrences, `https://pubs.geoscienceworld.org/ssa/bssa/article-abstract/111/3/1473/597788`
2. Stefanie Seif et al (2017), Estimating ETAS: The effects of truncation, missing data, and model assumptions, `https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2016JB012809`
3. Christian Grimm et al (2022), Solving three major biases of the ETAS model to improve forecasts of the 2019 Ridgecrest sequence, `https://link.springer.com/article/10.1007/s00477-022-02221-2`
4. S. Hainzl et al (2008), Impact of Earthquake Rupture Extensions on Parameter Estimations of Point-Process Models, `Impact-of-Earthquake-Rupture-Extensions-on`
5. About evaluation of forecasts: `Pseudoprospective-Evaluation-of-UCERF3-ETAS`, implemented in pyCSEP, `pyCSEP-A-Python-Toolkit-for-Earthquake-Forecast`

Software documentation and tutorials

0. Software documentation: `https://edinburgh-seismicity-hub.github.io/ETAS.inlabru/`
1. Using ETAS.inlabru on a real data sequence, including an introduction to the ETAS model: `edinburgh-seismicity-hub.github.io/ETAS.inlabru/articles/tutorial_real.html`
2. Testing the basic of the software by generating synthetic data: `edinburgh-seismicity-hub.github.io/ETAS.inlabru/articles/generateSyntheticCatalogues.html`
3. Generating synthetic data and estimating models: `edinburgh-seismicity-hub.github.io/ETAS.inlabru/articles/tutorial_synth.html`
4. How does the initial part of the data sequence affect the analysis? `edinburgh-seismicity-hub.github.io/ETAS.inlabru/articles/sensitivityToStartingPoint.html`

More technical background on how the Hawkes process model is implemented in inlabru can be found in the `how_to_build_hawkes` folder of `github.com/Serra314/Hawkes_process_tutorials`

Software installation

1. Install `INLA` from its special repository:

```r
install.packages(
  "INLA",
  getOption("repos"), INLA="https://inla.r-inla-download.org/R/testing"),
  dep = TRUE
)
```

2. Install `inlabru` from CRAN:

```r
install.packages("inlabru")
```

3. Install `ETAS.inlabru` from github:

```r
remotes::install_github("edinburgh-seismicity-hub/ETAS.inlabru")
```

Points of interest

There are many interesting questions one may address. Here are the main points of interest; it's expected that you'll be able to adress a selection of these in the project.

1. The ability to correctly retrieve α (which regulates how the number of event offspring scales with the magnitude) depends on how many different high magnitude events we have in the sequence. In principle, we expect that an increasing number of different high magnitude events will lead to better estimates of α, as opposed to having the same number of high magnitude events but all with the same magnitude. Investigate how the posterior distribution for α changes with the number of different high magnitude events, and the level of these magnitudes (e.g. can we accurately and/or precisely estimate α, having only one magnitude 4 event?)
2. Compare the posterior distributions when we fit the model on all parameters or when considering some of them fixed to the values used when generating the data. The question would is, which are the parameters that we need "to get right" in order to estimate the others correctly? What is the effect of mis-specifying some of the parameters?
3. Fit the model on a large number of synthetic catalogues and study how the posteriors change with respect to sequence characteristics. Examples are the number of events in the sequence, the number of large events, some measure of clustering.
4. Assess the forecast accuracy and/or precision of the models, e.g. by implementing some forecast assessment methods, such as the "N-test", and compute it for multiple forecasting periods.
5. Investigate how the empirical distribution of "time between events" behaves in synthetic vs real data (or synthetic from estimated vs known parameters).