The Journal of Open Source Software

Bernadette: Bayesian Inference and Model Selection for Stochastic Epidemics in R

Lampros Bouranis [1]

1 Department of Statistics, Athens University of Economics and Business, Athens, Greece

DOI: 10.21105/joss.05612

Software

- Review
- Repository
- Archive

Editor: Samuel Forbes

Reviewers:

- @rowlandseymour
- @strengejacke

Submitted: 04 June 2023
Published: 26 September 2023

Summary

The Coronavirus Disease 2019 (COVID-19) outbreak caused by SARS-CoV-2 has led to developments in Bayesian infectious disease modeling, allowing modelers to assess the impact of mitigation strategies on transmission and to quantify the burden of the pandemic. The Bernadette package (Bouranis, 2023) for the open-source R statistical software (R Core Team, 2023) implements a Bayesian evidence synthesis approach to modeling the age-specific transmission dynamics of a disease based on daily mortality counts. The functionality of Bernadette can be used to reconstruct the epidemic drivers from publicly available data, to estimate key epidemiological quantities like the rate of disease transmission, the latent counts of infections and the reproduction number for a given population over time, and to perform model comparison using information criteria. While Bernadette is motivated by the analysis of healthcare surveillance data related to COVID-19, it provides a template for implementation to a broader range of infectious disease epidemics and outbreaks.

Statement of need

The modeling of disease transmission dynamics is an active research area; a list of R packages dedicated to the analysis of epidemics is presented in the R Epidemics Consortium website (https://www.repidemicsconsortium.org/). Among these packages, EpiEstim (Cori et al., 2013; Cori, 2021) offers a data-driven Bayesian framework for the reconstruction of the time-varying reproduction number of a disease which represents the number of secondary infections generated by a case infected at day t. EpiEstim uses areal time-series in the form of case counts for a given period and population. The incidence of infection is assumed to follow a Poisson process with expectation given by a renewal equation. The epidemia package (Scott et al., 2020) links observed areal data to latent infections, which are in turn modeled as a self-exciting process tempered by time-varying reproduction numbers. It also supports Bayesian multilevel models by pooling information across multiple populations.

Bernadette complements these packages by implementing the Bayesian hierarchical modeling approach described in Bouranis et al. (2022). It links the observed age-stratified mortality counts for the population of a given country over time to latent infections via an over-dispersed count model. The change in infections over time is governed by a deterministic multi-type compartmental model which is driven by potentially non-scalar diffusion processes to adequately capture the temporal evolution of the age-stratified transmission rates. Bernadette relaxes the assumption of a homogeneous population, incorporates age structure and accounts for the presence of social structures via publicly available contact matrices. This allows for further evidence synthesis utilizing information from contact surveys and for sharing statistical strength across age groups and time. Further, the Bayesian evidence synthesis approach implemented in the Bernadette package enables learning the age-stratified virus transmission rates from one group to another, with a particular focus on the transmission rate between and onto vulnerable groups which could support public health authorities in devising interventions targeting these groups. The effective reproduction number is estimated from the daily age-stratified mortality counts, accounting for variations in transmissibility that are not obvious from reported infection counts. While estimation of the uncertainty over the effective reproduction number is itself a challenging problem (Gostic et al., 2020), it is propagated naturally via Markov chain Monte Carlo (Brooks et al., 2011).

Functionality

`Bernadette` features eight main functions for data processing and visualization, a main function for specifying and fitting the Bayesian hierarchical model and five functions for post-processing and visualization of posterior model estimates of important epidemiological quantities. The Github page for this package contains a detailed README file with a description of the modeling framework and the steps involved in the workflow.

Data processing and visualization

1. `age_distribution`: Imports the age distribution of a country for a given year, broken down by 5-year age bands and gender, following the United Nations 2019 Revision of World Population Prospects.
2. `aggregate_age_distribution`: Aggregates the age distribution according to user-defined age groupings.
3. `contact_matrix`: For a given country, it imports a 16 by 16 contact matrix whose row i of a column j corresponds to the number of contacts made by an individual in group i with an individual in group j.
4. `aggregate_contact_matrix`: Aggregates the contact matrix according to user-defined age groupings.
5. `aggregate_ifr_react`: Aggregates the age-specific Infection Fatality Ratio (IFR) estimates reported by the REACT-2 large-scale community study of SARS-CoV-2 seroprevalence in England (Ward et al., 2021) according to age group mappings defined by the user.
6. `itd_distribution`: Parameterizes the time distribution from an infection to an observation. `Bernadette` focuses on observed mortality counts, therefore we refer to this distribution as the *infection-to-death distribution*.
7. `plot_age_distribution`: visualizes a given age distribution using a bar plot.
8. `plot_contact_matrix`: visualizes a given contact matrix using a heatmap.

The age-specific time series of mortality counts and the age-specific time series of cumulative reported infections complete the list of data streams that are integrated together with expert knowledge into a coherent modeling framework via a Bayesian evidence synthesis approach for estimation of key epidemiological quantities. Two example datasets (objects `age_specific_mortality_counts` and `age_specific_cusum_infection_counts`) are provided with the `Bernadette` package to guide the user regarding the format of their input.

Parameter estimation

The main function for Bayesian parameter estimation, `stan_igbm`, allows for specification of a joint distribution for the outcomes (in this case, the age-specific mortality counts) and the unknown quantities, which is expressed by the likelihood for the outcomes conditional on the unknowns multiplied by a marginal prior distribution for the unknowns. This joint distribution is proportional to the posterior distribution of the unknowns conditional on the observed data. Prior beliefs for the unknown model parameters can be expressed by a selection of appropriate distributions available to the end-user. `Bernadette` uses the framework offered by the probabilistic programming language `Stan` (Carpenter et al., 2017) to specify a model and draw from the posterior distribution using MCMC. The user-specified model is internally translated into data that are passed to a precompiled Stan program and then it is fit using sampling methods from rstan (Stan Development Team, 2023).

Post-processing

The output of stan_igbm contains draws from the posterior distribution, which can be post-processed and visualized to extract insights about the mechanism of disease transmission over a given period.

1. plot_posterior_cm: Density plots of the posterior distribution of the random contact matrix.
2. posterior_infections: Summarizes the posterior distribution of the infection counts over time (age-specific and aggregated). It is accompanied by the plotting function plot_posterior_infections.
3. posterior_mortality: Summarizes the posterior distribution of the mortality counts over time (age-specific and aggregated). It is accompanied by the plotting function plot_posterior_mortality.
4. posterior_transmrate: Summarizes the posterior distribution of the age-specific transmission rate. It is accompanied by the plotting function plot_posterior_transmrate.
5. posterior_rt: Summarizes the posterior distribution of the time-varying reproduction number. The posterior trajectory is visualized with plot_posterior_rt.

The output of stan_igbm can additionally be used to compute approximate leave-one-out cross-validation with the loo R package (Vehtari et al., 2023). This enables estimation of information criteria which are considered when comparing among a set of alternative models for the same data.

Licensing and Availability

The Bernadette package is licensed under the GNU General Public License (GPL v3.0). It is available on CRAN, and can be installed using install.packages("Bernadette"). All code is open-source and hosted on GitHub, and bugs can be reported at https://github.com/bernadette-eu/Bernadette/issues/.

Acknowledgements

Lampros Bouranis' research is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101027218.

References

Bouranis, L. (2023). *Bernadette: Bayesian inference and model selection for stochastic epidemics*. https://CRAN.R-project.org/package=Bernadette

Bouranis, L., Demiris, N., Kalogeropoulos, K., & Ntzoufras, I. (2022). *Bayesian analysis of diffusion-driven multi-type epidemic models with application to COVID-19*. arXiv. https://doi.org/10.48550/arXiv.2211.15229

Brooks, S., Gelman, A., Jones, G., & Meng, X. (2011). *Handbook of Markov chain Monte Carlo*. CRC press.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Cori, A. (2021). *EpiEstim: Estimate time varying reproduction numbers from epidemic curves*. https://CRAN.R-project.org/package=EpiEstim

Cori, A., Ferguson, N., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, *178*(9), 1505–1512. https://doi.org/10.1093/aje/kwt133

Gostic, K., McGough, L., Baskerville, E., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J., De Salazar, P., Hellewell, J., Meakin, S., Munday, J., Bosse, N., Sherrat, K. e., Thompson, R., White, L., Huisman, J., Scire, J., … Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, Rt. *PLoS Computational Biology*, *16*(12), 1–21. https://doi.org/10.1371/journal.pcbi.1008409

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Scott, J., Gandy, A., Mishra, S., Unwin, J., Flaxman, S., & Bhatt, S. (2020). *epidemia: Modeling of epidemics using hierarchical Bayesian models*. https://imperialcollegelondon.github.io/epidemia/

Stan Development Team. (2023). *RStan: The R interface to Stan*. https://mc-stan.org/

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P., Paananen, T., & Gelman, A. (2023). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. https://mc-stan.org/loo/

Ward, H., Atchison, C., Whitaker, M., Ainslie, K., Elliott, J., Okell, L., Redd, R., Ashby, D., Donnelly, C., Barclay, W., Darzi, A., Cooke, G., Riley, S., & Elliott, P. (2021). SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. *Nature Communications*, *12*, 905. https://doi.org/10.1038/s41467-021-21237-w