Expand support for spatial models in PyMC

Introduction

This project would implement the Besag, York, Mollie (BYM) model in PyMC. BYM is a highly flexible model for studying spatial data and is used widely in epidemiology, agriculture, and ecology. The strategy behind the BYM model is to treat the outcome of interest as the result of three factors: some set of predictor variables, spatial covariance between neighboring regions, and random effects that represent non-spatial heterogeneity.

Although very flexible, the BYM model can be difficult to specify in a way that is simultaneously computationally efficient, interpretable, and identifiable. Recently, Morris et al (2019) demonstrated an alternative specification of the BYM model that is significantly more efficient, interpretable, and can be identified with Monte Carlo Markov Chain (MCMC) samplers.

Developing an implementation in Python with PyMC would make the model more accessible to a greater variety of users. Furthermore, BYM models are only available in a Bayesian framework. As a leading Bayesian statistics package, PyMC should support this extremely useful model in spatial statistics.

Technical details

Spatial data is typically reported in one of two ways: points and areas. PyMC currently has good support for spatial analysis problems involving point data with its powerful Gaussian process module. However, areal data (such as land tracts, political districts, or neighborhoods) requires a distinct family of statistical models that are not fully developed in the PyMC ecosystem.

In areal data, the distance between observations cannot be represented as a continuous quantity. Instead, what matters is whether two regions are neighboring. Adjacency matrices are often used to represent spatial relationships. Areal data naturally lends itself to a family of statistical models known as conditionally auto-regressive models (CAR). There are several variants of the CAR model which each have strengths and drawbacks.

The standard CAR model uses an adjacency matrix to generate the precision matrix of a multivariate normal distribution. The multivariate normal, in turn, models how neighboring regions covary. Mathematically, CAR might be represented as:

$$\phi \sim MvNormal(0,\ Q^{-1})$$

$$Q\ =\ D\ -\ \alpha W$$

Suppose we have n regions. W is an $n \times n$ adjacency matrix between regions. D is a diagonal matrix where the diagonal represents how many neighbors each of n regions have. α is a parameter which represents how strong spatial correlations are amongst neighbors, with 0 being no spatial correlation and 1 being perfect spatial correlation. The drawback of the standard CAR model is computational inefficiency. Generating the covariance for the multivariate normal requires matrix inversion which grows approximately cubically with the size of the data. Similarly, the log density function requires computing a determinant (Morris et al, 2019).

The intrinsic conditionally auto-regressive model (ICAR) is a special case of the CAR model that assumes complete spatial correlations between regions. In the terms of the previous model, it sets $\alpha = 1$. This simplifies the computations significantly as it is no longer necessary to compute a determinant in the log density calculation (Morris et al, 2019). However, the assumption is unrealistic. The amount that space matters in a given research problem should be an adjustable and, hopefully, identifiable quantity.

The BYM model addresses the limits of the ICAR model by nesting ICAR into a larger regression. This regression treats the outcome of interest as the result of three factors: some set of predictor variables, spatial covariance between neighboring areas, and random effects that represent non-spatial heterogeneity. Typically, BYM models assume a poisson outcome distribution, but in probabilistic programming frameworks other outcome distributions could easily be substituted. A parameter can control the relative balance of spatial covariance and random effects. This model is both more realistic and more computationally efficient. There are two drawbacks. First, the parameter that controls the balance of spatial and random effects is often non-identifiable without informative priors. Second, it is very difficult to design a default prior that is sufficiently informative, interpretable, and transportable across datasets (Morris et al, 2019).

Recently, Riebler et al (2016) developed a new parameterization of the BYM model where the parameters are easier to interpret and interact more effectively with the MCMC family of samplers. Mathematically, the model can be represented as:

$$y = a + b * x + \sigma(\sqrt{\rho / s}\,\phi + \sqrt{1 - \rho}\,\theta)$$

The first part of the expression, $y = a + b * x$, is a standard linear model. σ is the overall standard deviation for the combined effect of spatial covariance and the random effect. ρ controls the balance of spatial covariance and random effects as sources of heterogeneity. s is a scaling factor computed from the neighborhood graph. ϕ is the ICAR model described above. Finally, θ is the random effects components. It is a normal distribution with mean 0 and standard deviation equal to the number of connected subgraphs. Typically, areal data is fully connected and the standard deviation is 1.

Morris et al showed that the model can be implemented in the probabilistic programming language Stan, and illustrated the utility of the model on New York City traffic accident data. This paper is a valuable proof of concept for my proposal. However, the existing work is somewhat restricted by the accessibility of the Stan language. Users more comfortable with Python would benefit from a PyMC implementation. Other popular spatial Bayesian modeling packages have limitations. BayesCAR uses a combination of Gibbs and Metropolis-Hastings samplers which have largely been superseded by the kind of Hamiltonian samplers deployed by PyMC. Similarly, INLA supports BYM models but the package is restricted to the INLA technique for approximating the posterior. It would be helpful to decouple the BYM model from any particular estimation procedure. One goal of probabilistic programming is that users should be able to freely mix and match models with estimation procedures. PyMC offers a broad array of techniques and continuously updates its sampling algorithms so the performance of BYM models will improve over time.

Schedule

Community bonding period

May 4 - 29

I would focus on expanding my knowledge of the PyMC codebase. This includes deepening my knowledge of PyTensor, the computational backend of PyMC. I would need to learn how to debug PyTensor graphs and how PyTensor derives log-probabilities from random graphs. I would also need to learn more about the standards of software engineering associated with the creation of new distributions and random variables. This would require learning more about how the testing regime operates around new distributions.

I would also consult with members of the PyMC community to further refine my project to ensure it aligns with the goals of the community. Specifically, we would need to discuss the best way of making the BYM model available to users. This might require decisions around API design so I would want to understand how PyMC makes its decision around existing distribution APIs.

Phase 1

May 29 - June 26 (4 weeks)

The beginning of the project would focus on updating PyMC's existing CAR capacities.

- Update the current notebook on CAR to version 5
- Finish development on ICAR by adding appropriate tests to the existing but stale pull request
- Develop a prototype BYM model and explore how the computational efficiency of the model interacts with PyTensor. This will help identify whether our implementation can follow Stan's or whether we need modifications.

Phase 2

June 26 - July 31 (5 weeks)

The middle of the project would focus on developing a new distribution

- Create new BYM RV ops. Decide whether it is a special case of CAR of MvNormal or whether it is its own random variable.
- Implement the BYM distribution and associated methods with log-PDF, and moments
- Write for tests for: log-PDF, moments, rng function for the random variable,
- Explore the need for specialized tests of BYM, likely connected to the use of neighborhood graphs and adjacency matrices.
- Write documentation

July 14 - midterm evaluation deadline

Phase 3

July 31 - August 28 (4 weeks)

The end of the project would focus on making the work accessible.

- Draft notebook on BYM adapted from Morris et al with New York pedestrian injuries dataset
- Publish notebook on BYM model to the PyMC examples page

After GSoC

I'll maintain the PyMC spatial distributions (CAR, ICAR and BYM) and support users of spatial models on the PyMC discourse.

Why me?

I'm a PhD student in philosophy at the University of British Columbia. My main area of research is philosophy of science, a field that studies the justification of research methods. I have a particular interest in understanding how computational statistics is transforming behavioral science. I like to take a hands-on approach—that means philosophy of science research happens through close engagement with active scientific communities. Software projects are a wonderful case study for thinking about the decisions and justifications behind the tools researchers use.

As such, I've built up a habit of contributing to open-source software projects. Here are some of my contributions:

- Documentation for AR distribution: https://github.com/PyMC-devs/PyMC/pull/6080
- Documentation for Simulator distribution: https://github.com/PyMC-devs/PyMC/pull/6035
- Documentation for Gaussian process module: https://github.com/PyMC-devs/PyMC/pull/6609

I've used PyMC in research. Here is a blog post and associated code that reanalyzes the data from a recent quantitative philosophy of science paper and argues that the original conclusions become much stronger if we incorporate a multi-level model.

I also support the PyMC community on discourse, helping users with a variety of questions. Larry Dong, a PyMC core developer, generously provided some feedback on this proposal.

Finally, I teach research methods in the cognitive science program at UBC. My course is an integrated philosophy of science, computational modeling, and bayesian statistics course that uses PyMC. So I have an active interest in seeing the package improve in the coming years.

This summer I'll have sufficient time for GSoC. My only other responsibility is dissertation related research. In a typical semester, I work on dissertation research part time while managing teaching responsibilities. If selected, I will opt out of teaching responsibilities and have roughly 30 hours a week to dedicate to GSoC.

Contact:

- email: dsaunders406@gmail.com
- phone: 416-841-8758 (Canada)
- twitter: @CarolBasknRobns

References

- Mitzi Morris, Katherine Wheeler-Martin, Dan Simpson, Stephen J. Mooney, Andrew Gelman, Charles DiMaggio. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. Spatial and Spatio-temporal Epidemiology, Volume 31, https://doi.org/10.1016/j.sste.2019.100301. https://www.sciencedirect.com/science/article/pii/S1877584518301175
- Riebler A, Sørbye SH, Simpson D, Rue H. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. (2016). Statistical Methods in Medical Research. 25(4):1145-1165, doi:10.1177/0962280216660421