Model dispersion with PRISM; an alternative to MCMC for rapid analysis of models

Ellert van der Velden[1, 2]

1 Centre for Astrophysics and Supercomputing, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia **2** ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D)

DOI: 10.21105/joss.01229

Software

- Review
- Repository
- Archive

Submitted: 25 January 2019
Published: 10 June 2019

In partnership with

AMERICAN ASTRONOMICAL SOCIETY

This article and software are linked with research article DOI 10.3847/1538-4365/ab1f7d, published in the Astrophysical Journal Supplement Series.

Summary

Rapid technological advancements allow for both computational resources and observational/experimental instruments to become better, faster and more precise with every passing year. This leads to an ever-increasing amount of scientific data being available and more research questions being raised. As a result, scientific models that attempt to address these questions are becoming more abundant, and are pushing the available resources to the limit as these models incorporate more complex science and more closely resemble reality.

However, as the number of available models increases, they also tend to become more distinct, making it difficult to keep track of their individual qualities. A full analysis of every model would be required in order to recognize these qualities. We commonly employ Markov chain Monte Carlo (MCMC) methods and Bayesian statistics for performing this task. However, as these methods are meant to be used for making approximations of the posterior probability distribution function, we think there is a more efficient way of analyzing them.

Based on the algorithms described by Vernon, Goldstein, & Bower (2010), we have built *PRISM*, a publicly available *Probabilistic Regression Instrument for Simulating Models* for Python 3. *PRISM* uses the Bayes linear approach (Goldstein & Wooff, 2007), emulation technique (Craig, Goldstein, Seheult, & Smith, 1996, 1997) and history matching (Craig et al., 1996, 1997; Raftery, Givens, & Zeh, 1995) to construct an approximation ('emulator') of any given model, by combining limited model evaluations with advanced regression techniques, covariances and probability calculations. It is designed to easily facilitate and enhance existing MCMC methods by restricting plausible regions and exploring parameter space more efficiently. However, *PRISM* can additionally be used as a standalone alternative to MCMC for model analysis, providing insight into the behavior of complex scientific models. These techniques have been used successfully in the past (Bower et al., 2010; Vernon et al., 2010; Vernon et al., 2018) to speed up model analyses, but their algorithms are typically written for a specific application and are not publicly available. With *PRISM*, the time spent on evaluating a model is minimized, providing developers with an advanced model analysis for a fraction of the time required by more traditional methods.

The API for *PRISM* was designed to work well for both simple and advanced projects, with every class being written as a base class, but also almost every user-method solely taking optional arguments. This allows for the user to quickly get started with *PRISM*, while still being able to make adjustments to various routines with minimal effort. Its `Pipeline` class features a user-friendly environment that connects all of *PRISM*'s methods together, whereas the `ModelLink` abstract base class helps users wrapping ('linking') their model to *PRISM*. *PRISM* relies heavily on popular existing Python packages for its expensive computations, like `NumPy` (Oliphant, 2006), `Scikit-learn` (Pedregosa et al., 2011) and `Mlxtend` (Raschka, 2018), making it more robust and future-proof.

Test applications of *PRISM* (see van der Velden, Duffy, Croton, Mutch, & Sinha (2019))

show that *PRISM* can provide a qualitative parameter estimation over 15 times faster than stand-alone MCMC methods, while also being able to give insight into the model's behavior (which MCMC cannot provide). In future work, *PRISM* will be used together with the MCMC package `mhysa` (Mutch et al. in prep.) to analyze and explore the parameter space of the semi-analytic model `Meraxes` (Mutch et al., 2016). Also, several smaller application projects with *PRISM* are currently being planned. The source code for *PRISM* can be found at https://github.com/1313e/PRISM

Acknowledgements

EV would like to thank Alan Duffy, Darren Croton, Simon Mutch and Manodeep Sinha for being a great help with writing the code. EV would also like to thank Chris Blake, Colin Jacobs and Theo Steininger for fruitful discussions and valuable suggestions. Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. Parts of this work were performed on the OzSTAR national facility at Swinburne University of Technology. OzSTAR is funded by Swinburne University of Technology and the National Collaborative Research Infrastructure Strategy (NCRIS).

References

Bower, R. G., Vernon, I., Goldstein, M., Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S., et al. (2010). The parameter space of galaxy formation. *Monthly Notices of the Royal Astronomical Society*, *407*, 2017–2045. doi:10.1111/j.1365-2966.2010.16991.x

Craig, P. S., Goldstein, M., Seheult, A. H., & Smith, J. A. (1996). Bayes linear strategies for history matching of hydrocarbon reservoirs. In J. M. Bernardo, A. P. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5* (pp. 69–95). Oxford, UK: Clarendon Press.

Craig, P. S., Goldstein, M., Seheult, A. H., & Smith, J. A. (1997). Pressure matching for hydrocarbon reservoirs: A case study in the use of bayes linear strategies for large computer experiments. In C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch, P. Rossi, & N. D. Singpurwalla (Eds.), *Case studies in bayesian statistics* (pp. 37–93). New York, NY: Springer New York.

Goldstein, M., & Wooff, D. (2007). *Bayes Linear Statistics: Theory and Methods* (1st ed.). West Sussex: John Wiley & Sons Ltd.

Mutch, S. J., Geil, P. M., Poole, G. B., Angel, P. W., Duffy, A. R., Mesinger, A., & Wyithe, J. S. B. (2016). Dark-ages reionization and galaxy formation simulation - III. Modelling galaxy formation and the epoch of reionization. *Monthly Notices of the Royal Astronomical Society*, *462*, 250–276. doi:10.1093/mnras/stw1506

Oliphant, T. (2006). NumPy: A guide to NumPy. USA: Trelgol Publishing. Retrieved from http://www.numpy.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Raftery, A. E., Givens, G. H., & Zeh, J. E. (1995). Inference from a Deterministic Population Dynamics Model for Bowhead Whales. *Journal of the American Statistical Association*, *90*(430), 402–416. doi:10.1080/01621459.1995.10476529

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, *3*(24). doi:10.21105/joss.00638

van der Velden, E., Duffy, A. R., Croton, D., Mutch, S. J., & Sinha, M. (2019). Model dispersion with PRISM; an alternative to MCMC for rapid analysis of models. *The Astrophysical Journal Supplement Series*. doi:10.3847/1538-4365/ab1f7d

Vernon, I., Goldstein, M., & Bower, R. G. (2010). Galaxy formation: A bayesian uncertainty analysis. *Bayesian Anal.*, *5*(4), 619–669. doi:10.1214/10-BA524

Vernon, I., Liu, J., Goldstein, M., Rowe, J., Topping, J., & Lindsey, K. (2018). Bayesian uncertainty analysis for complex systems biology models: Emulation, global parameter searches and evaluation of gene functions. *BMC Systems Biology*, *12*(1), 1. doi:10.1186/s12918-017-0484-3