The Journal of Open Source Software

pgmuvi: Quick and easy Gaussian Process Regression for multi-wavelength astronomical timeseries

Peter Scicluna [1,2¶], **Stefan Waterval** [3,4], **Diego A. Vasquez-Torres** [5], **Sundar Srinivasan** [5], **and Sara Jamal** [6]

1 European Southern Observatory, Alonso de Córdova 3107, Vitacura, Santiago, Chile **2** Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA **3** New York University Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates **4** Center for Astro, Particle and Planetary Physics (CAP^3), New York University Abu Dhabi, PO Box 129188, Abu Dhabi, United Arab Emirates **5** IRyA, Universidad Nacional Autónoma de México, Morelia, Michoacán, México **6** Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Axel Donath

Reviewers:

- @joshspeagle
- @baptklein

Submitted: 31 July 2023
Published: unpublished

Summary

Time-domain observations are increasingly important in astronomy, and are often the only way to study certain objects. The volume of time-series data is increasing dramatically as new surveys come online - for example, the Vera Rubin Observatory will produce 15 terabytes of data per night, and its Legacy Survey of Space and Time (LSST) is expected to produce five-year lightcurves for $> 10^7$ sources, each consisting of 5 photometric bands. Historically, astronomers have worked with Fourier-based techniques such as the Lomb-Scargle periodogram or information-theoretic approaches; however, in recent years Bayesian and data-driven approaches such as Gaussian Process Regression (GPR) have gained traction. However, the computational complexity and steep learning curve of GPR has limited its adoption. `pgmuvi` makes GPR of multi-band timeseries accessible to astronomers by building on cutting-edge open-source machine-learning libraries, and hence `pgmuvi` retains the speed and flexibility of GPR while being easy to use. It provides easy access to GPU acceleration and Bayesian inference of the hyperparameters (e.g. the periods), and is able to scale to large datasets.

Statement of need

Astronomical objects are in general not static, but vary in brightness over time. This is especially true for objects that are variable by nature, such as pulsating stars, or objects that are variable due to their orbital motion, such as eclipsing binaries. The study of these objects is called time-domain astronomy, and is a rapidly growing field. A wide range of approaches to time-series analysis have been developed, ranging from simple period-finding algorithms to more complex machine learning techniques (e.g. Donoso-Oliva et al., 2023; Friedman, 1984; Huijse et al., 2018; Palmer, 2009 and many more). Perhaps the most popular in astronomy is the Lomb-Scargle periodogram (Lomb, 1976; Scargle, 1982), which is a Fourier-based technique to find periodic signals in unevenly sampled data. However, the handling of unevenly sampled data is not the only challenge in time-series analysis.

A particular challenge in astronomy is handling heterogeneous, multiwavelength data (VanderPlas & Ivezić, 2015). Data must often be combined from a wide variety of instruments, telescopes or surveys, and so the systematics or noise properties of different datasets vary widely. In addition, by combining multiple wavelengths, we gain a better understanding of the physical processes driving the variability of the object. For example, some variability mechanisms differ as a function of wavelength only in amplitude (e.g. eclipsing binaries), while others may vary in

phase (e.g. pulsating stars) or even period (e.g. multiperiodic systems). Thus, it is important to combine data from multiple wavelengths in a way that accounts for these differences.

Gaussian processes (GPs) have recently become a popular tool to handle these challenges. GPs are a flexible way to forward-model arbitrary signals, by assuming that the signal is drawn from a multivariate Gaussian distribution. By constructing a covariance function describing the covariance between any two points in the signal, we can model the signal as a Gaussian process. By doing so, we are freed from any assumptions about sampling, and can model the signal as a continuous function. We are also able to model the noise in the data, and thus account for heteroscedastic noise. We also gain the ability to directly handle multiple wavelengths, by constructing a multi-dimensional covariance function. Hence, Gaussian Process Regression (GPR) is a machine learning technique that is able to model non-periodic signals in unevenly sampled data, and is thus well suited for the analysis of astronomical time-series data.

However, GPR is not without its challenges. The most popular covariance functions are often not tailored to modelling complex signals, which implies that users must construct their own covariance functions. GPs are also computationally expensive, and thus approximations must be used to scale to large datasets. Finally, many GP packages either have very steep learning curves or only provide limited features, and thus are not suitable for the average astronomer.

In this paper we present a new Python package, `pgmuvi`, to perform GPR on multi-wavelength astronomical time-series data. The package is designed to be easy to use, and provide a quick way to perform GPR on multi-wavelength data by exploiting uncommon but powerful kernels and appoximations. The package is designed to be flexible and allow the user to customize the GPR model to their needs. `pgmuvi` exploits multiple strategies to scale regression to large datasets, making it suitable for the current era of large-scale astronomical surveys.

A number of other software packages exist to perform GPR, some of which, such as `celerite` (D. Foreman-Mackey et al., 2017; D. Foreman-Mackey, 2018), `tinygp` (Daniel Foreman-Mackey, 2023) or `george` (Ambikasaran et al., 2015) were developed within the astronomical community with astronomical time-series in mind. However, these each have their own limitations. `celerite` is extremely fast, but is limited to one-dimensional inputs, and thus cannot handle multiwavelength data, except under certain restrictive assumptions. Furthermore, since it achieves its speed through a specific form of kernel decomposition, it is not able to handle arbitrary covariance functions. It is therefore restricted to a small number of kernels with specific forms; while it is able to handle the most common astronomical timeseries by combining these kernels, not all signals can be modelled in this way. `tinygp` is a more general package, able to retain a high degree of flexibility while still being fast thanks to a `JAX`-based implementation, whic makes it feasible to implement models in `tinygp` that are equivalent to those in `pgmuvi`. However, `pgmuvi` is designed to provide an easier learning curve by packaging GPs with data transforms and inference routines. In essence, `tinygp` could in principle be used by `pgmuvi` as a GP backend instead of GPyTorch. For a summary of the state of the art of GPR in astronomy, see the recent review by Aigrain & Foreman-Mackey (2023).

`pgmuvi` is used in two ongoing projects by our group: one of the authors' (DAVT) masters thesis and the paper resulting from this work deals with the analysis of multiwavelength light curves for targets from the Nearby Evolved Stars Survey (NESS; Scicluna et al. (2022), https://evolvedstars.space). This work served as the first test of the code and has analyzed thousands of light curves at optical and infrared wavelengths for over seven hundred dusty stars within 3 kpc of the Solar Neighborhood. The paper will be published in 2023 (Vasquez-Torres et al., in prep.). A different project related to dusty variable stars in M33 has also used `pgmuvi` to estimate the periods of these objects from infrared light curves. This work will be published in 2023 (Srinivasan et al., in prep.).

The Journal of Open Source Software

Method and Features

pgmuvi builds on the popular GPyTorch library. GPyTorch (Gardner et al., 2018) is a Gaussian process library for PyTorch (Paszke et al., 2019), which is a popular machine learning library for Python. By default, pgmuvi exploits the highly-flexible Spectral Mixture kernel (Wilson & Adams, 2013) in GPyTorch, able to model a wide variety of signals. This kernel is particularly interesting for astronomical time-series data, as it can effectively model multi-periodic and quasi-periodic signals. The spectral mixture kernel models the power spectrum of the covariance matrix as Gaussian mixture model (GMM), making it highly flexible, easy to interpret and adaptable to multi-dimensional data. This kernel is known for its ability to extrapolate, and is thus well suited to cases where prediction is important (for example, preparing astronomical observations of variable stars). By modelling the power spectrum in this way, pgmuvi effectively filters out noise in the periodogram, and thus is able to find the dominant periods in noisy data more effectively than, for example, the Lomb-Scargle periodogram.

However, the flexibility of this kernel comes at a cost; for more than one component in the mixture, the solution space becomes highly non-convex, and the optimization of the kernel hyperparameters becomes difficult. pgmuvi addresses this by first exploiting the Lomb-Scargle periodogram to find the dominant periods in the data, and then using these periods as initial guesses for the means of the mixture components.

Multiple options are available to accelerate inference depending on the size of the dataset. For small datasets, the exact GPs can be used to scale to datasets of up to ~ 1000 points. pgmuvi can exploit the Structured Kernel Interpolation (SKI) approximation (Wilson & Nickisch, 2015) to scale to datasets of up to $\sim 10^5$ points. Future work will include implementing approximations for even larger datasets: pgmuvi can in principle exploit the Sparse Variational GP (SVGP) or Variational Nearest Neighbour (VNN) approximations (Hensman et al., 2013; Wu et al., 2022) to scale to datasets of arbitrary size. pgmuvi can employ also GPU computing for both exact and variational GPs.

For exact GPs and SKI, pgmuvi performs maximum a posteriori (MAP) estimation of the hyperparameters, or full Bayesian inference. MAP estimation can exploit any PyTorch optimizer, but by default it uses ADAM (Kingma & Ba, 2014). Bayesian inference uses the pyro (Bingham et al., 2018) implementation of the No-U-Turn Sampler (NUTS) (Hoffman et al., 2014), which is a Hamiltonian Monte Carlo (HMC) sampler.

Finally, by allowing arbitrary GPyTorch likelihoods to be used, pgmuvi can be extended to a wide range of problems. For example, an instance of `gpytorch.likelihoods.StudentTLikelihood` can be dropped in to turn pgmuvi into a T-Process regressor, or missing data can be handled using `gpytorch.likelihoods.GaussianLikelihoodWithMissingObs`.

To summarise, the key features of pgmuvi are:

- Highly flexible kernel, able to model a wide range of multiwavelength signals
- Able to exploit multiple strategies to scale to large datasets
- GPU acceleration for both exact and variational GPs
- Fully Bayesian inference using HMC
- Initialisation of kernel hyperparameters using Lomb-Scargle periodogram
- Automated creation of diagnostic and summary plots
- Automated reporting of kernel hyperparameters and their uncertainties, and summary of MCMC chains.

Acknowledgements

This project was developed in part at the 2022 Astro Hack Week, hosted by the Max Planck Institute for Astronomy and Haus der Astronomie in Heidelberg, Germany. This work was

partially supported by the Max Planck Institute for Astronomy, the European Space Agency, the Gordon and Betty Moore Foundation, the Alfred P. Sloan foundation.

SS and DAVT acknowledge support from UNAM-PAPIIT Program IA104822.

References

Aigrain, S., & Foreman-Mackey, D. (2023). Gaussian process regression for astronomical time series. *Annual Review of Astronomy and Astrophysics*, *61*(1), null. https://doi.org/10.1146/annurev-astro-052920-103508

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. (2015). Fast Direct Methods for Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*, 252. https://doi.org/10.1109/TPAMI.2015.2448083

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*.

Donoso-Oliva, C., Becker, I., Protopapas, P., Cabrera-Vives, G., Vishnu, M., & Vardhan, H. (2023). ASTROMER. A transformer-based embedding for the representation of light curves. *670*, A54. https://doi.org/10.1051/0004-6361/202243928

Foreman-Mackey, D. (2018). Scalable Backpropagation for Gaussian Processes using Celerite. *Research Notes of the American Astronomical Society*, *2*(1), 31. https://doi.org/10.3847/2515-5172/aaaf6c

Foreman-Mackey, Daniel. (2023). *dfm/tinygp: The tiniest of Gaussian Process libraries* (Version v0.2.4rc1). Zenodo. https://doi.org/10.5281/zenodo.7646759

Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. (2017). Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series. *154*, 220. https://doi.org/10.3847/1538-3881/aa9332

Friedman, J. H. (1984). A variable span scatterplot smoother. Laboratory for Computational Statistics, Stanford University Technical Report, *5*.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., & Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in Neural Information Processing Systems*, *31*.

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv Preprint arXiv:1309.6835*.

Hoffman, M. D., Gelman, A., & others. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Huijse, P., Estévez, P. A., Förster, F., Daniel, S. F., Connolly, A. J., Protopapas, P., Carrasco, R., & Príncipe, J. C. (2018). Robust period estimation using mutual information for multiband light curves in the synoptic survey era. *The Astrophysical Journal Supplement Series*, *236*(1), 12. https://doi.org/10.3847/1538-4365/aab77c

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*.

Lomb, N. R. (1976). Least-Squares Frequency Analysis of Unequally Spaced Data. *39*(2), 447–462. https://doi.org/10.1007/BF00648343

Palmer, D. M. (2009). A FAST CHI-SQUARED TECHNIQUE FOR PERIOD SEARCH OF IRREGULARLY SAMPLED DATA. *The Astrophysical Journal*, *695*(1), 496. https://doi.org/10.1088/0004-637X/695/1/496

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd international conference on neural information processing systems*. Curran Associates Inc.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *263*, 835–853. https://doi.org/10.1086/160554

Scicluna, P., Kemper, F., McDonald, I., Srinivasan, S., Trejo, A., Wallström, S. H. J., Wouterloot, J. G. A., Cami, J., Greaves, J., He, J., Hoai, D. T., Kim, H., Jones, O. C., Shinnaga, H., Clark, C. J. R., Dharmawardena, T., Holland, W., Imai, H., van Loon, J. T., … Zijlstra, A. A. (2022). The Nearby Evolved Stars Survey II: Constructing a volume-limited sample and first results from the James Clerk Maxwell Telescope. *512*(1), 1091–1110. https://doi.org/10.1093/mnras/stab2860

VanderPlas, J. T., & Ivezić, Ž. (2015). Periodograms for Multiband Astronomical Time Series. *812*(1), 18. https://doi.org/10.1088/0004-637X/812/1/18

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 1067–1075). PMLR. https://proceedings.mlr.press/v28/wilson13.html

Wilson, A., & Nickisch, H. (2015). Kernel interpolation for scalable structured gaussian processes (KISS-GP). *International Conference on Machine Learning*, 1775–1784.

Wu, L., Pleiss, G., & Cunningham, J. P. (2022). Variational nearest neighbor gaussian process. *International Conference on Machine Learning*, 24114–24130.