Main research interests

I'm primarily interested in the fundamental, methodological and practical aspects of learning statistical models from data. I have also found an application for these models that attracted me since very little and for which my country (Chile) is famous: Astronomy. Nowadays most of **my personal research falls under the relatively new field of Astroinformatics, a combination between astronomy, data science and statistical/machine learning. In particular I'm focused on the development of new methods to automatically analyze astronomical data**, and I'm interested in the scenario posed by near-future synoptic surveys. For a complete list of publications please see my full CV at `http://phuijse.github.io`.

Summary of past research and selected works

I did my doctoral studies at the *Computational Intelligence Laboratory, Universidad de Chile*, under Prof. Dr. Pablo A. Estévez, developing new methods to automatically analyze large volumes of astronomical time series (light curves). For this I specialized myself in the fields of statistical signal processing, information theory and machine learning. I was awarded with two government-funded scholarships for doctoral internships in 2012 and 2013, respectively. In 2012 I spent 8 months at the *Institute for Applied Computational Sciences (IACS), Harvard University*, under Prof. Dr. Pavlos Protopapas[1], where we worked on analyzing the EROS-2 astronomical survey. Then in 2013 I spent 8 months at the *Computational Neuro-Engineering Laboratory (CNEL), University of Florida*, under Prof. Dr. José Principe[2], where we worked on the fundamental limits of the information theoretic methods I've been implementing for astronomical data. I graduated in September 2014 and my doctoral thesis led to several journal publications [1–4].

In 2015-2017 I was a postdoc researcher at the *Millennium Institute of Astrophysics (MAS)*[3], for the development and application of machine learning models with astronomical data. This position was fully-funded by the Chilean ministry of science. Within MAS I collaborated closely with astronomers and astrophysicists, expanding considerably in the practical aspects of dealing with astronomical data from different surveys. I contributed to MAS by developing feature-based classifiers for the rapid detection of Supernovae [5–7] and semi-supervised models for the detection of RR Lyrae [8].

In 2018 I joined the *Informatics Institute at the Universidad Austral de Chile (UACh)* as an assistant professor, where, to this day, I guide student theses and give lectures on topics related to statistics, machine learning and data science. I continued my research on astroinformatics through a research project titled "Efficient methods based on information theory and machine learning for astronomical images and time series analysis", which was

[1] `https://iacs.seas.harvard.edu/people/pavlos-protopapas`
[2] `http://www.cnel.ufl.edu/people/people.php?name=principe`
[3] `https://www.astrofisicamas.cl/en/`

awarded with funding from the Chilean ministry of science (2017-2020). In the context of this project I developed information theoretic methods to detect periodic behavior in light curves [9] and neural-network based representations models for images of transient astronomical objects [10, 11]. I collaborated on projects related to follow-up target selection based on information theory[12] and neural networks to classify transients from image sequences [13]. I also directed research on novel generative-inference models for images in general [14].

I'm currently a young researcher at MAS and I participate in the *Automatic Learning for the Rapid Classification of Events (ALeRCE)*[4] collaboration, where we aim at building an astronomical broker. The ALeRCE system is connected to the stream of large etendue telescopes, processes the data in real-time using machine learning methods and outputs the results to the astronomical community. My role in the ALeRCE project is to develop models for analysis and classification of astronomical images and light curves. The libraries and tools that we have developed are open-source and freely available to the community [15–17].

References

[1] P. **Huijse** et al. 2011. DOI: 10.1109/LSP.2011.2141987. arXiv: 1112.2962.

[2] P. **Huijse** et al. 2012. DOI: 10.1109/TSP.2012.2204260. arXiv: 1212.2398.

[3] P. **Huijse** et al. 2014. DOI: 10.1109/MCI.2014.2326100. arXiv: 1509.07823.

[4] P. Protopapas et al. 2015. DOI: 10.1088/0067-0049/216/2/25. arXiv: 1412.1840.

[5] P. **Huijse** et al. 2015. DOI: 10.1016/j.procs.2015.07.276.

[6] F. Förster et al. 2016. DOI: 10.3847/0004-637X/832/2/155. arXiv: 1609.03567.

[7] J. Martínez-Palomera et al. 2018. DOI: 10.3847/1538-3881/aadfd8. arXiv: 1609.03567.

[8] R. C. Ramos et al. 2018. DOI: 10.3847/1538-4357/aacf90. arXiv: 1807.04303.

[9] P. **Huijse** et al. 2018. DOI: 10.3847/1538-4365/aab77c. arXiv: 1709.03541.

[10] P. **Huijse** et al. 2018. URL: https://www.esann.org/sites/default/files/proceedings/legacy/es2018-130.pdf.

[11] N. Astorga et al. 2018. DOI: 10.1109/IJCNN.2018.8489358.

[12] J. Astudillo et al. 2019. DOI: 10.3847/1538-3881/ab557d. arXiv: 1911.02444.

[13] R. Carrasco-Davis et al. 2019. DOI: 10.1088/1538-3873/aaef12. arXiv: 1807.03869.

[14] N. Astorga et al. 2020. DOI: 10.1007/978-3-030-58592-1_39. arXiv: 2008.09641.

[15] F. Förster et al. 2021. DOI: 10.3847/1538-3881/abe9bc. arXiv: 2008.03303.

[16] P. Sánchez-Sáez et al. 2021. DOI: 10.3847/1538-3881/abd5c1. arXiv: 2008.03311.

[17] A. Sánchez et al. 2021. URL: https://ml4physicalsciences.github.io/2021/files/NeurIPS_ML4PS_2021_10.pdf.

[4] http://alerce.science/

Research plans

New synoptic surveys such as ZTF[5] and LSST[6], which aim to explore the dynamic sky, i.e. transient, variable or moving astronomical phenomena, bring new challenges and also new opportunities. Novel computational methods that are robust and highly efficient are needed to process these new astronomical data streams. **I plan to address these challenges under two lines of research: novel deep learning architectures for astronomical time series and efficient methods to estimate predictive uncertainty of models trained with astronomical data.** These ideas are part of my most recent research project titled "Novel deep learning architectures for astronomical time series" (2021-2023). This project has also been awarded with government funding.

Deep Learning represents the state of the art in classification, prediction and representation of unstructured data and time series in particular. Nonetheless, most of the work related to deep learning with time series is focused on regularly-sampled data, which is not the case of astronomical time series (light curves). A careful review of the literature on machine learning models for light curves shows two trends. The most prevalent strategy is based on training models using features engineered for light curves, e.g. [1, 2]. These features do not scale well (in terms of inference time) to the case of synoptic surveys, and most importantly they were designed for older surveys with very different characteristics. The second trend has been to train deep learning models using established architectures such as convolutional and recurrent neural networks, on raw light curves, e.g. [3, 4]. Inference with these models is faster than feature computation. However, these architectures were not designed for irregularly sampled sequences, and in many practical examples they have not surpassed feature-based classifiers in terms of accuracy, which is contrary to the trends seen in other areas. **My hypothesis is that new neural architectures that are designed to address the irregularity/sparsity in sampling, the multi-dimensionality and the heteroscedasticity will be able to extract more information from synoptic survey light curves and hence improve the state of the art in classification and representation learning, while at the same time being efficient and scalable so as to implement them within synoptic survey alert-processing systems**.

To verify this hypothesis I plan to extend the concept of parametric or continuous convolution, where a kernel is a function instead of an array of fixed values. This idea has been explored as interpolation-layers for irregular medical records [5, 6]. In this case the kernel has a fixed parametric form, e.g. a gaussian function. This facilitates training but limits the capabilities of the learned filters. The filter could instead be a small fully-connected neural network (one hidden layer). This greatly expands in terms of flexibility but may require strong regularization schemes. Interpolation to a a-priori known time-grid might not be a good choice considering the sparse sampling of light curves from synoptic surveys. For this I plan to explore multi-scale time grids that can be learned from data. Another option is to drop the receptive field of convolutions in favor of attention-like mechanisms that take

[5] `https://www.ztf.caltech.edu/`
[6] `https://www.lsst.org/`

appropriate consideration of continuous and irregular time through modulation. The second concept that I plan to explore are continuous-time differential equation models such as neural-ODE or neural processes [7]. These methods could be extended to the heteroscedastic case through appropriate cost functions.

To test the performance of baselines and proposed architectures I will consider simulations and real data from synoptic survey alerts. The LSST collaboration has provided excellent tools (catsim and obsim) to synthetize light curve data following the characteristics of LSST. For the real data we will use the data releases collected and pre-processed by the ALeRCE project. This dataset has been labeled through cross-match with previous surveys and through the use of the current ALeRCE classifiers. Additional challenges to be addressed in the case of real data are the high class-imbalance, the bias towards brighter objects and uncertainty in labels from previous models. Neural networks models will be implemented using the PyTorch and the Python 3 scientific computing suite. The source codes will be open-source, free and distributed via the github platform, zenodo and the PyPI repository. We are currently carrying out experiments with different combinations of the aforementioned neural layers in encoder-decoder architectures for a semi-supervised scenario.

For the second line of research I'm exploring the intersection of Bayesian learning and deep neural networks, i.e. Bayesian neural networks, and in particular how to implement and train these models effectively. In this framework uncertainty of the parameters can be propagated to obtain well-calibrated uncertainty for the model predictions. This is particularly relevant for over-parameterized models such as deep neural networks which tend to be over-confident in their predictions[8]. Accurate uncertainty estimation for the predictions is key for several downstream tasks such as anomaly detection and active learning. The former is particularly important in the case of synoptic surveys, where the serendipitous discovery of new populations of astronomical objects is expected. Interesting out-of-distribution phenomena should be found quickly in order to trigger follow-up observations and inspection by experts. One disadvantage of bayesian neural networks is that inference is not tractable and asymptotically exact methods such as MCMC have prohibitive computational cost. Approximate methods such as variational inference scale better but more research is needed regarding the guarantees of the obtained posteriors. In fact very simple factorized posteriors perform well in practice for deeper models [9], i.e. a rich predictive posterior can be obtained combining complex transformations and simple posteriors for the parameters of the model.

In the particular case of light curves, uncertainty is generally available in the form of an estimated photometric error. I plan on developing a bayesian neural network framework that incorporates uncertainty in the input in addition to uncertainty in the parameters. For this I'll explore functional priors at the intersection of bayesian neural networks and Gaussian processes. For inference we will first consider Amortized Variational Inference (AVI) with simple posteriors. In later stages of research low-rank posteriors and other approximated bayesian methods such as mc-dropout will be compared. Mini-batch implementations of MCMC could also be competitive with smaller neural networks. Simulated and real data will be used to perform the experiments. Bayesian model and inference routines will be

implemented using the JAX framework and the numpyro probabilistic programming library. The correspondence of uncertainties and prediction accuracies will be assessed using the expected calibration error and the negative log-likelihood. The training and inferential time required by each strategy will also be compared. Results will be analyzed as a function of the model complexity (depth), prior variance and the simulated observing conditions, so as to contrast with current theoretical findings regarding posterior quality and model depth. I also plan on carrying out experiments on outlier light curve detection based on the predictive posteriors, where I'll compare against feature-based and non-bayesian anomaly detection methods.

Astronomy has proven to be a fertile ground to develop and test novel computational methods. Thanks to the excellent scientific network at MAS I have not only first-hand access to astronomical data but also the expertise of world-class astronomers and astrophysicists. Participating on the ALeRCE project has been a very fruitful experience but there are still plenty of unsolved challenges. I'm eager to continue contributing to this project through my own research and also by coordinating and supervising final-year projects and thesis. I will also remain open to new collaborations that may benefit from my expertise. **Collaboration between astronomers, mathematicians, computer scientists and engineers is key to solve the near-future astroinformatics challenges.**

References

[1] J. W. Richards et al. 2011. DOI: 10.1088/0004-637X/733/1/10.

[2] G. Narayan et al. 2018. DOI: 10.3847/1538-4365/aab781.

[3] I. Becker et al. 2020. arXiv: 2002.00994.

[4] S. Jamal et al. 2020. DOI: 10.3847/1538-4365/aba8ff.

[5] S. N. Shukla et al. 2019. URL: https://openreview.net/forum?id=r1efr3C9Ym.

[6] S. N. Shukla et al. 2021. URL: https://openreview.net/forum?id=4c0J6lwQ4_.

[7] A. Norcliffe et al. 2021. URL: https://openreview.net/forum?id=27acGyyI1BY.

[8] P. Izmailov et al. 2021. arXiv: 2104.14421.

[9] S. Farquhar et al. 2020. arXiv: 2002.03704.

Appendix

Participation in other activities not related to astroinformatics:

- FUSA system (2021-2022): I participated as senior developer and leader of the machine learning group in a technology-transfer project between UACh and the chilean ministry of environment. The goal of project was to build a system for the automatic classification of urban sound events for the development of noise maps of the city of Valdivia. My responsibilities included proposing, implementing and training artificial neural networks models to fulfill the project requirements, i.e. classifying the audio data streams. I personally developed the data classification pipelines starting from acquisition to inference APIs. I was also in charge of writing the proposal associated to the classification models. The project was funded by the chilean ministry of science through an ANID FONDEF grant.

- Motivus (2019-today): I'm one of the founders of Motivus, a startup that focuses on building an ecosystem to share computing capacity and data processing algorithms. My role in Motivus is as researcher and chief technological advisor for the development and implementation of AI models within the motivus platform. For more details please see `https://motivus.cl/`

- Scientific conference organization (2019-today): In 2019 I was the general chair of the Latin American Summer School on Computational Intelligence (EVIC 2019). The school, which was sponsored by IEEE-CIS and chilean enterprises, was focused on neural networks, evolutionary computing and fuzzy systems. We had more than 300 participants from Chile and neighboring countries who experienced plenary lectures and tutorials by top international and national speakers. In 2021 and 2022 I was the chair of the neural and learning systems track of the Latin American Conference on Computational Intelligence (LA-CCI), where I was in charge of reviewing and managing decisions for 50 papers related to neural networks. I also collaborate