On the Quality of Deep Representations for Unevenly-Sampled Light-Curves using Variational AutoEncoders

Francisco Mena, Patricio Olivares, Margarita Bugueño, Gabriel Molina, Mauricio Araya

Universidad Técnica Federico Santa María, Chile

Abstract

Analyzing light curves nowadays usually involves processing large datasets. Therefore, finding a good representation for them is both, a key and a non trivial task. In this paper we show that variational (stochastic) autoencoder models (VRAE$_t$) can be applied to learn an effective, informative and robust deep representation of transit light curves. In addition, we introduce S-VRAE$_t$, which stands for *re-Scaling Variational Recurrent Auto Encoder*, a technique that embeds the re-scaling preprocessing of a time series into the learning model in order to use the scale information in the detection of transit. The objective is to achieve the most likely low dimensional output of unevenly-sampled time series that matches latent variables to reconstruct it. For assessing our approach we use the largest transit dataset obtained by the Kepler mission in the past four years, and compare our results with similar techniques used for light curves. Our results show that the stochastic models have an improvement on the quality of representation with respect to their deterministic counterparts. Moreover, the S-VRAE$_t$ model is at the same time a deep denoising model, generating light curves similar to the Mandel Agol fit.

Introduction and Problem

In astronomy there exists different type of data that are processed and studied, i.e. catalogs, time series and data cubes. In the exoplanet detection field time series are used, particularly light curves. A *light curve* is the measurement of light intensity of a celestial object or region, which can varies for different factors. For example, variable star varies the intensity depending on its chemical composition. If an orbiting planet is involved, it passes in front of a light source (star) and a fraction of the light is blocked. This phenomenon is called **transit** and is an effective method to find exoplanets.

Transit Detection Methods

The standard method to model transit light curves is the Mandel-Agol (M-A) process [5], which models the transit of a spherical planet around a spherical star assuming an uniform light source which is eclipsed by the planet. The increasing popularity of machine learning methods and the large amount of data generated by observatories, allow to detect exoplanets by reducing effort and time. The common approach to classify variable star is to extract hand-crafted *specialized features* from the light curve and apply classic machine learning algorithms [7].

Time series on astronomy has the challenge that they are quite uneven on their sampling, meaning that it have missing values for several timestamps. To tackle this sampling problem, some approaches [2] represent the light curve as phase aligned sections called "folds". This folded light curve centers the transit stacking all the times on which they occurs. It usually get binned on a window proportional to the transit period. Deep learning has been also applied to transit detection using 1D convolutional neural networks (CNN) using global and local representation of the folded light curves [8].

Some methods focus on self-learned representation for unevenly-sampled light curves without direct human intervention. For example, in [6] a recurrent auto-encoder (RAE) is used for learning embeddings for variable stars. It uses the variations on time as an additional channel, as in [1] with an 1D CNN. Variational (stochastic) auto-encoder (VAE) [4] has been used for representation learning of time series with recurrent neural networks (VRAE).

Extending VRAE to Uneven Samples

We propose two VRAE extension that can properly handle dimensionality reduction of unevenly-sampled light curves. We focus on self-learned representation on the transit-shape domain that has less bias than the human-crafted. Both model adapts VRAE to unevenly-sampled light curves as [6] by using the time intervals (δ) as an extra input channel to the encoder and decoder. The first model, called **VRAE**$_t$ (VRAE *plus* time information), is a natural extension from VRAE that includes times. The standard pre-process on time series re-scale the values to more tractable magnitudes in order to achieve bounded weights and activation on neural net models. In order to avoid loss of information on the original scale, the second model embed the re-scaling process of the time series into the learning as an end-to-end architecture. The scale is also encoded with an auxiliary task of reconstruct it (decoder). The objective is to use the re-scaled version of the data to detect pattern behaviors instead of magnitudes behaviors and also extract the information on the original scale. The architecture of the VRAE$_t$ with *re-Scaling* or **S-VRAE**$_t$ is illustrated in Figure 1.

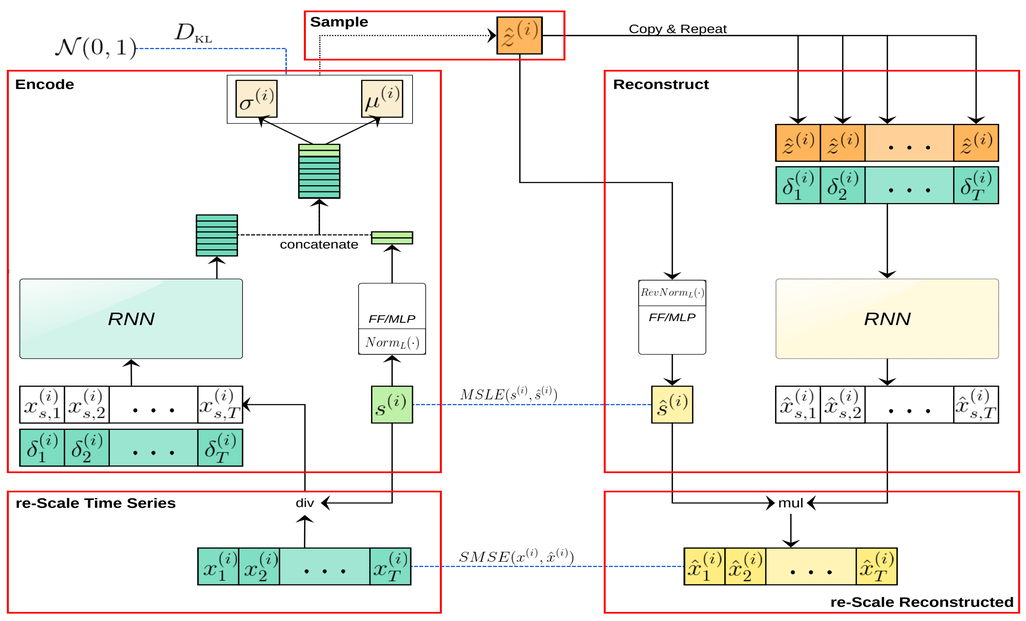

Figure 1: VRAE$_t$ architecture: Raw time series is re-scaled and fed to the RNN encoder in addition to the delta times. In parallel the scale is encoded too. Both embeddings are concatenated to obtain the parameters of the Normal distribution. The sampled value of this latent distribution is repeated and stacked to the delta times to reconstruct the scaled time series. Finally, the raw scale is returned to the time series.

Experiments

We validate on *Kepler Objects of Interest* (KOI) dataset of Kepler mission which classifies an object as *Confirmed* or *False Positive*. For metadata, we selected those features derived from the light curve. For the experiments, we use the Kepler de-trend pipeline to obtain a standard light curve and set the **folded-global** representation with delta times binned by 300 points. We set a mask over the 8054 objects to filter out light curves without a transit behavior, obtaining 4317 objects to train our models. The selection is based on the Kepler metadata: *flags*, *transit score* and the error of a Mandel agol fit. We also duplicate the dataset by mirroring each folded light curve.

We set the dimension of the encoder representation to 16, generating a 95% of compression. For the autoencoder evaluation (Table 1) we compare against the Butterworth passband filter, a moving average filter, a Mandel-Agol [5] fit on the metadata, a Recurrent Auto-Encoder plus time (RAE$_t$) [6]. For the dependence experiment (Table 2) we compare against the representation of Kepler metadata, RAE$_t$ and PCA extracted over the Fourier series (F+PCA) [3]. The performance of an application (classification task) of the learned representation to detect exoplanets is present on Table 3. Here we compare against previously mentioned methods besides 1D CNN, which is a based on [8] and "RNN$_t$", which is the encoder of RAE$_t$/VRAE$_t$ with a classification layer at the top, without decoder phase. Figure 2 shows examples of reconstructed light curves for a better understanding of the results.

		Reconstruction		*Denoising*		*Residual Noise*
Method	**Configuration**	**RMSE**	**MAE**	**AutoC**	**Diff-M**	**Spectral-H**
Original LC		-	-	0.27	0.78	0.84
Passband	1-500	1.08	0.62	0.97	0.05	0.82
	50-1500	1.04	0.65	0.83	0.20	0.84
	50-2500	0.96	0.64	0.67	0.36	0.85
Mov. avg	3	0.72	0.46	0.70	0.27	0.88
	5	0.84	0.51	0.78	0.17	0.86
	10	0.94	0.55	0.84	0.09	0.84
M-A sim.	library: *batman*	2.34	0.63	0.64	0.22	0.89
RAE$_t$	16	0.69	0.48	0.50	0.13	0.90
VRAE$_t$	16	0.69	0.48	0.59	0.07	0.90
S-VRAE$_t$	16	0.72	0.49	0.61	0.06	0.90

Table 1: Autoencoder Evaluation: Reconstruction error and denoising results by different methods. *Autocorr*: Autocorrelation and *Diff-M*: Mean of the differences between consecutive values on the estimated time series. *Spectral-H*: Spectral entropy over the model output residual. The configuration of the passband corresponds to low and high pass filters respectively, for mov. average is the window size, while for AE methods is the dimension of the encoder representation.

Representation	**PC**	**A-PC**	**MI**	**N-MI**
Metadata	0.064	0.162	0.275	0.044
(Raw) F+PCA	0.000	0.000	0.210	0.027
(Fold) F+PCA	0.000	0.000	0.061	0.008
RAE$_t$	0.057	0.277	0.255	0.033
VRAE$_t$	0.012	0.168	0.122	0.016
S-VRAE$_t$	-0.003	0.138	0.072	0.009

Table 2: Feature Dependence: Pearson correlation mean (PC) and absolute mean (A-PC) between the features, this express the linear dependence on the representations. Mutual Information (MI) and normalized value between 0 and 1 (N-MI) between the features, this express an approximation to the information dependence on the representation.

Representation	*input dim*	**Non-Exo**	**Exo**	**F1-Ma**
Metadata	10	90.13	83.85	87.00
Folded-Global	$T = 300$	83.51	69.36	76.45
Unsupervised Methods				
(Raw) F+PCA	16	77.14	71.21	68.66
(Fold) F+PCA	16	84.62	71.34	77.98
RAE$_t$	16	84.42	72.40	78.41
VRAE$_t$	16	85.88	74.84	80.36
S-VRAE$_t$	16	87.39	77.91	82.65
Supervised Methods				
RNN$_t$	$T = 300$	88.67	79.09	83.87
1D CNN	$T = 300$	89.27	81.40	85.33

Table 3: Classification: Performance based on F1 score (macro and desegregated by classes). Exo stands for Exoplanet.

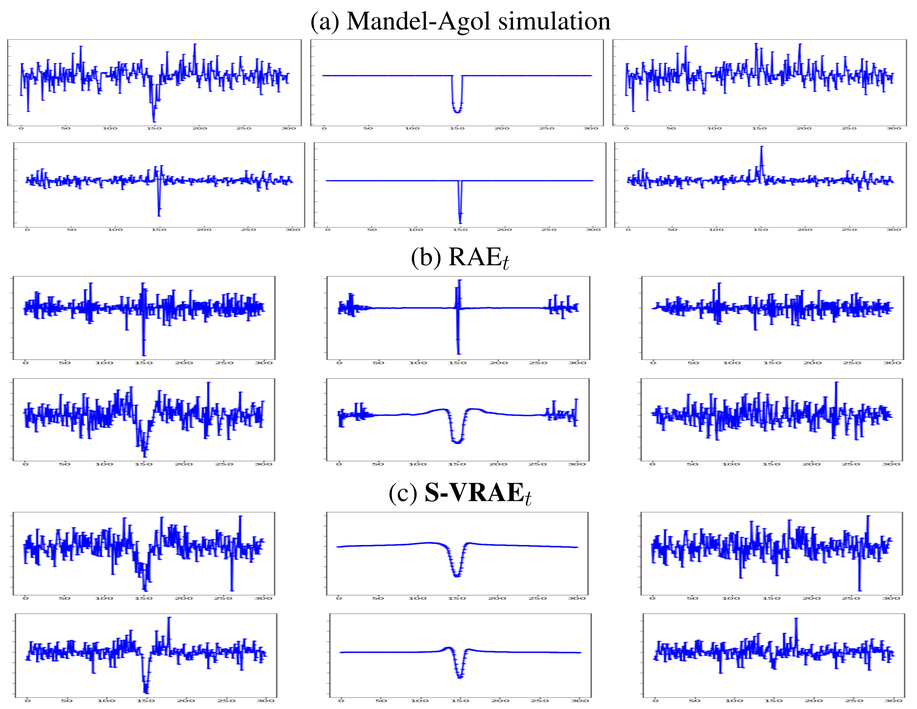

Figure 2: Examples of reconstructed light curves by the methods. At the left, the original folded-global raw light curve, the reconstructed or denoised version in the middle. At the right, the residual output.

Conclusions and Keypoints

- We propose two VAE models that adapts uneven sampled light curves with recurrent operations. These stochastic methods increase the advantages of the discrete counterpart.
- The quality evaluation of the representation indicates that learned representation (via the encoder) is more informative, useful and robust than other methods in the literature.
- The learned representation are more *disentangled* that some compared. The **S-VRAE**$_t$ proposal almost reaches the optimum disentangled representation of PCA.
- The reconstruction results show that **S-VRAE**$_t$ proposal can be seen as a deep denoising model generating denoised time series based on the original raw scale.

Acknowledgments. Funding of ANID-Basal Project FB0008 (AC3E) and ANID PIA/APOYO AFB180002 (CCTVal).

References

[1] Carlos Aguirre, Karim Pichara, and Ignacio Becker. Deep multi-survey classification of variable stars. *Monthly Notices of the Royal Astronomical Society*, 482(4):5078–5092, 2019.

[2] David J Armstrong, Don Pollacco, and Alexandre Santerne. Transit shapes and self organising maps as a tool for ranking planetary candidates: Application to kepler and k2. *Monthly Notices of the RAS*, 2016.

[3] Margarita Bugueno, Francisco Mena, and Mauricio Araya. Refining exoplanet detection using supervised learning and feature engineering. In *2018 XLIV Latin American Computer Conference (CLEI)*, pages 278–287, 2018.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Kaisey Mandel and Eric Agol. Analytic light curves for planetary transit searches. *The ApJL*, 580(2):L171, 2002.

[6] Brett Naul, Joshua S Bloom, Fernando Pérez, and Stéfan van der Walt. A recurrent neural network for classification of unevenly sampled variable stars. *Nature Astronomy*, 2(2):151–155, 2018.

[7] Joseph W Richards, Dan L Starr, Nathaniel R Butler, Joshua S Bloom, John M Brewer, Arien Crellin-Quick, Justin Higgins, Rachel Kennedy, and Maxime Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10, 2011.

[8] Christopher J Shallue and Andrew Vanderburg. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astronomical Journal*, 155(2):94, 2018.