Path Imputation Strategies for Signature Moddels

Michael Moor, Max Horn, Christian Bock, Karsten Borgwardt, Bastian Rieck

ICML 2020 Workshop on the Art of Learning with Missing Values (Artemiss)

Michael_D_Moor

Workshop paper: https://openreview.net/pdf?id=P0DL7M6T57o
(Long) preprint: https://arxiv.org/pdf/2005.12359.pdf

Problem setup

- The signature transform is a 'universal nonlinearity' on the space of continuous vector-valued paths and has gained attention in ML for being a powerful feature extractor which can be easily integrated to neural networks.
- The signature acts on *continuous paths*. However, in real-world applications, temporal data typically appears as a *discretized* collection of observations.
- To apply signature techniques to this data, the data first has to be transformed (or "embedded") into a continuous path.
- This step has been typically glossed over as an unimportant detail, yet, we hypothesize that this step could have a considerable impact on the resulting signature.

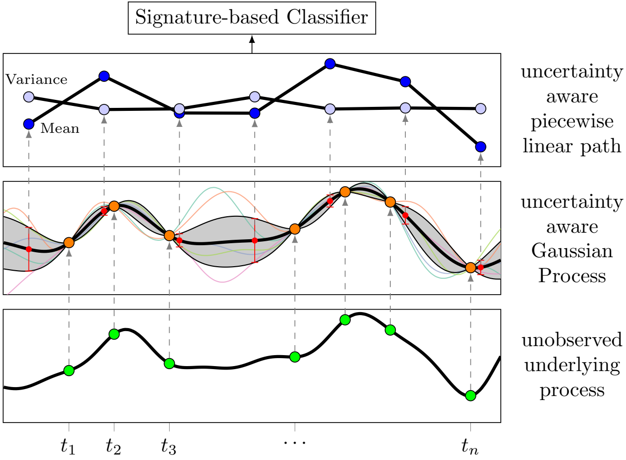
Path Imputation Strategies

Here, we study the effect of the following imputation strategies on time series classifiers (with or without signatures):

1. zero imputation
2. forward filling
3. indicator imputation
4. linear interpolation
5. causal imputation
6. Gaussian process (GP) adapter
7. GP adapter with posterior moments (novel)

GP adapters with posterior moments (PoM)

In addition, we propose the following strategy which is an extension to GP adapters that is uncertainty aware at the *prediction* step as opposed to the training phase:



Results

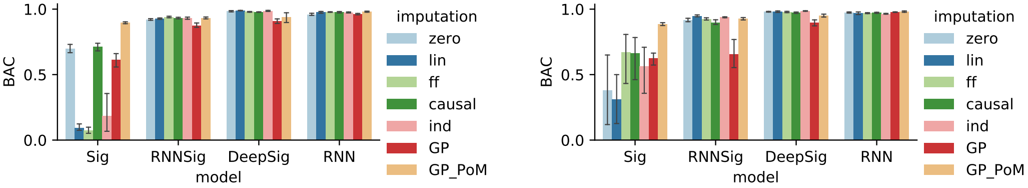


Figure: Results for CharacterTrajectories in terms of balanced accuracy (BAC). Missing 50% at random (left), label-based subsampling 40-60% (right).

- Over various datasets, imputation schemes, and models we observed that signature models can be drastically affected by differing imputations.
- We found that GP-PoM tends to make signature models more robust (especially shallow ones which are more affected).

Supplementary Slides

Paths

Definition

A *path* X in $\mathbb{R}^d$ is a continuous mapping from $[a, b]$ to $\mathbb{R}^d$, i.e.

$$\begin{aligned} X\colon [a, b] &\to \mathbb{R}^d \\ t &\mapsto X(t) \end{aligned} \tag{1}$$

for $t \in \mathbb{R}$. High-dimensional paths can be *decomposed* into a collection of real-valued paths, i.e. $X = \left(X^1, \ldots, X^d\right)$, with $X^i\colon [a, b] \to \mathbb{R}$.

We will write X_t to denote $X(t)$.

Path integrals

Given a function $f\colon \mathbb{R} \to \mathbb{R}$ and a one-dimensional path $X\colon [a, b] \to \mathbb{R}$, the *path integral* of X against f is defined as

$$\int_a^b f(X)\,\mathrm{d}X = \int_a^b f(X(t))\frac{\mathrm{d}X_t}{\mathrm{d}t}\,\mathrm{d}t, \tag{2}$$

which can be seen as a re-parametrised Riemann integral. Intuitively, it measures how f changes as a function of the path X.

Path signatures

Let X be a d-dimensional path. For $i \in \{1, \dots, d\}$, let

$$\mathcal{S}(X)^i_{a,t} := \int_{a<s<t} \mathbb{1} \, \mathrm{d}X^i_s = X^i_t - X^i_a \tag{3}$$

i.e. the increment of the i^{th} coordinate of the path at some point $t \in [a, b]$.

Notably, $\mathcal{S}(X)^i_{a,\cdot} \colon [a, b] \to \mathbb{R}$ is itself a real-valued path!

Path signatures

Therefore, we can iterate this process. For $i, j \in \{1, \ldots, d\}$, we have

$$\mathcal{S}(X)^{i,j}_{a,t} := \int_{a<s<t} \mathcal{S}(X)^{i}_{a,s} \, \mathrm{d}X^{j}_{s} = \int_{a<r<s<t} \mathbb{1} \, \mathrm{d}X^{i}_{r} \, \mathrm{d}X^{j}_{s}. \tag{4}$$

Path signatures

Finally, for a collection of indices $i_1, \ldots i_k \in \{1, \ldots, d\}$, with $k \geq 1$, we can define

$$\mathcal{S}(X)^{i_1,\ldots,i_k}_{a,t} := \int_{a<s<t} \mathcal{S}(X)^{i_1,\ldots,i_{k-1}}_{a,s} \, \mathrm{d}X^{i_k}_s, \tag{5}$$

$$:= \int_{a<t_k<t} \cdots \int_{a<t_1<t_2} \mathrm{d}X^{i_1}_{t_1} \ldots \mathrm{d}X^{i_k}_{t_k}. \tag{6}$$

Path signatures

Definition

The *path signature*, or simply the *signature* is the collection of all the iterated integrals of X, i.e.

$$\mathrm{Sig}(X)_{a,b} := \Big(1, \mathcal{S}(X)^{1}_{a,b}, \ldots \mathcal{S}(X)^{d}_{a,b}, \mathcal{S}(X)^{1,1}_{a,b}, \mathcal{S}(X)^{1,2}_{a,b}, \ldots \mathcal{S}(X)^{d,d}_{a,b}, \ldots\Big), \tag{7}$$

for which all superscripts follow some ordering of multi-indices.

Intuition behind the signature

Analytical

Signatures are (partly) inspired by iterative approaches for solving ODEs, such as Picard iterations. For example [1], consider the ODE

$$\frac{dy}{dx} = y(x), \quad y(0) = 1$$

$$\begin{aligned}
&y_0(x) = 1\\
&y_1(x) = 1 + \textstyle\int_0^x y_0(t)dt = 1 + x\\
&y_2(x) = 1 + \textstyle\int_0^x y_1(t)dt = 1 + x + \frac{1}{2}x^2\\
&y_3(x) = 1 + \textstyle\int_0^x y_2(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3\\
&y_k(x) = 1 + \textstyle\int_0^x y_{k-1}(t)dt\\
&\vdots
\end{aligned}$$

which converges to $y(x) = e^x$ as $k \to \infty$.

Intuition behind the signature

Analytical

Signatures are (partly) inspired by iterative approaches for solving ODEs, such as Picard iterations. For example [1], consider the ODE

$$\frac{dy}{dx} = y(x), \quad y(0) = 1$$

$$\begin{aligned}
y_0(x) &= 1 \\
y_1(x) &= 1 + \int_0^x y_0(t)dt = 1 + x \\
y_2(x) &= 1 + \int_0^x y_1(t)dt = 1 + x + \tfrac{1}{2}x^2 \\
y_3(x) &= 1 + \int_0^x y_2(t)dt = 1 + x + \tfrac{1}{2}x^2 + \tfrac{1}{6}x^3 \\
y_k(x) &= 1 + \int_0^x y_{k-1}(t)dt \\
&\vdots
\end{aligned}$$

which converges to $y(x) = e^x$ as $k \to \infty$.

Intuition behind the signature

Analytical

Signatures are (partly) inspired by iterative approaches for solving ODEs, such as Picard iterations. For example [1], consider the ODE

$$\frac{dy}{dx} = y(x), \quad y(0) = 1$$

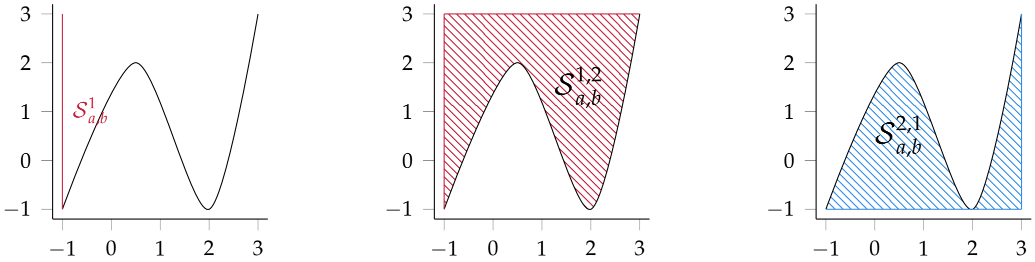
$$\begin{aligned}
&y_0(x) = 1 \\
&y_1(x) = 1 + \textstyle\int_0^x y_0(t)dt = 1 + x \\
&y_2(x) = 1 + \textstyle\int_0^x y_1(t)dt = 1 + x + \frac{1}{2}x^2 \\
&y_3(x) = 1 + \textstyle\int_0^x y_2(t)dt = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 \\
&y_k(x) = 1 + \textstyle\int_0^x y_{k-1}(t)dt \\
&\vdots
\end{aligned}$$

which converges to $y(x) = e^x$ as $k \to \infty$.

Intuition behind the signature

Geometric

Signatures of order 1, i.e. terms of the form $\mathcal{S}^i_{a,b}$ correspond to the *increment* of a path. Terms of order 2, i.e. $\mathcal{S}^{i,j}_{a,b}$, for $i \neq j$, correspond to *signed* areas above and below of a path.



Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Kidger et al.

Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Properties of the signature

Theoretical

- Uniqueness
- Universal nonlinearity
- Factorial decay of higher-order terms

Practical

- The truncated signature (up to order k) captures therefore most of the available information.
- Can be computed with tensor operations alone.
- The signature of concatenated paths can be computed efficiently.

Related Works

Path Signature in Machine Learning

- A Primer on Signatures for Machine Learning [1]
- Gaussian Processes with signature kernels [2]
- Signatures for Sepsis Prediction [3]
- Deep Signature Transforms [4]: ”Signature Layer” inside a neural network.

Further Details on Experimental Setup

Imputation strategies

In total, we have 7 imputation strategies:

1. zero imputation
2. forward filling
3. indicator imputation
4. linear interpolation
5. causal imputation
6. GP adapter (monte carlo)
7. GP adapter (PoM)

Models

We compare the following four models:

1. Sig: a simple MLP which employs one signature layer.
2. RNNSig: a GRU that slides over a window-based stream of signatures
3. RNN: a conventional GRU model [5]
4. DeepSig: a deeper network employing 2 signature layers [4]

Datasets

We make use of four real-world time series datasets: Physionet2012 challenge [6], PenDigits [7], LSST [8], and CharacterTrajectories [7].

Preprocessing

To challenge signature models, we subsample time series which are not irregularly observed in the first place. To this end two subsampling schemes are employed:

- "Random": missing at random. 50% of observations (PenDigits: 30%)
- "Label-based": missing not at random: [40 - 60%], uniformly sampled per class, (PenDigits: [20 - 40%])

Datasets

We make use of four real-world time series datasets: Physionet2012 challenge [6], PenDigits [7], LSST [8], and CharacterTrajectories [7].

Preprocessing

To challenge signature models, we subsample time series which are not irregularly observed in the first place. To this end two subsampling schemes are employed:

- ”Random”: missing at random. 50% of observations (PenDigits: 30%)
- ”Label-based”: missing not at random: [40 - 60%], uniformly sampled per class, (PenDigits: [20 - 40%])

Training

For each setting in [imputations $\times$ models $\times$ datasets ($\times$ subsamplings)]
$\rightarrow$ run hyperparameter search (20 fits in a randomized search)
$\rightarrow$ per fit: train until convergence (patience $= 20$) or at most 100 epochs
Per setting, select the best hyperparameter configuration in terms of performance on the validation split.

Evaluation

- For binary classifcation: average precision
- For multi-class classification: balanced accuracy

For each best setting, we refit 5 repetitions and report the test measures with error bars.

Training

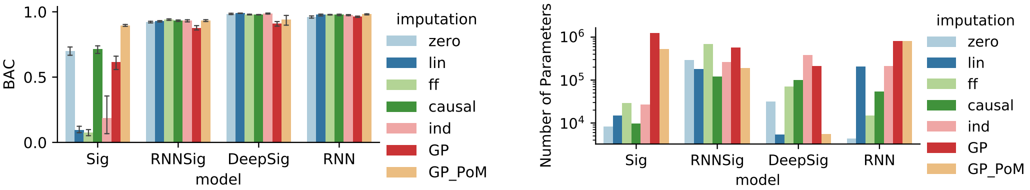
For each setting in [imputations $\times$ models $\times$ datasets ($\times$ subsamplings)]
$\rightarrow$ run hyperparameter search (20 fits in a randomized search)
$\rightarrow$ per fit: train until convergence (patience $= 20$) or at most 100 epochs
Per setting, select the best hyperparameter configuration in terms of performance on the validation split.

Evaluation

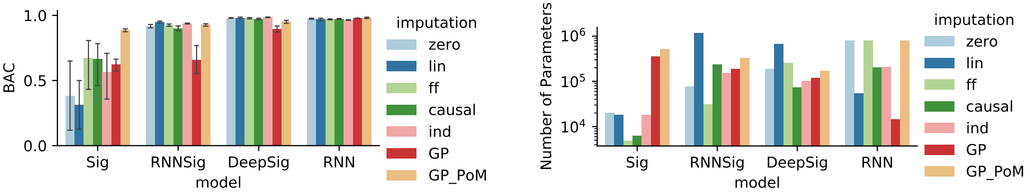
- For binary classifcation: average precision
- For multi-class classification: balanced accuracy

For each best setting, we refit 5 repetitions and report the test measures with error bars.

Results

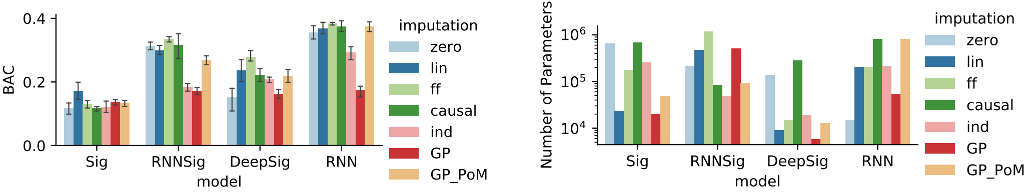


(a) CharacterTrajectories-R

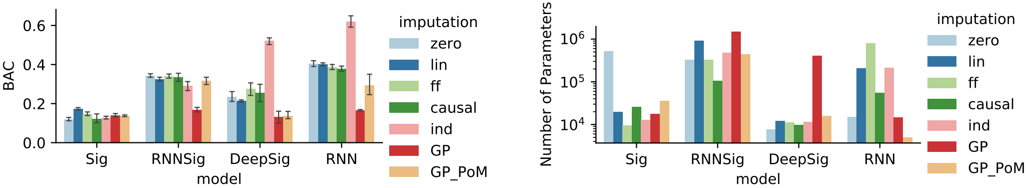


(b) CharacterTrajectories-L

Results II

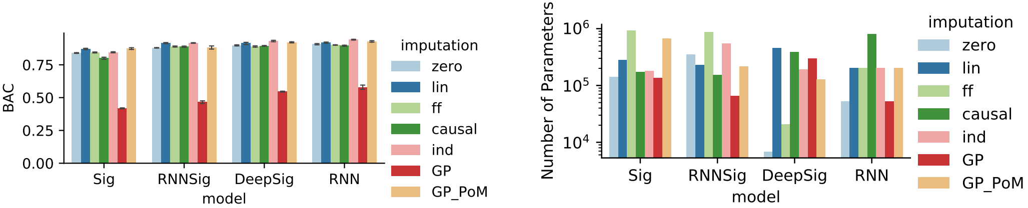


(a) LSST-R

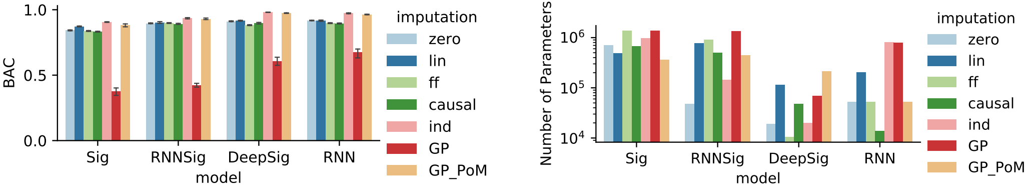


(b) LSST-L

Results III

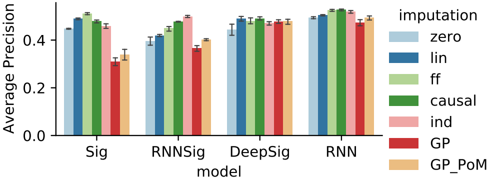


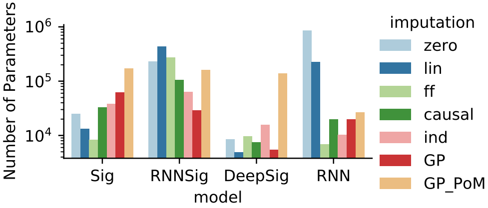
(a) PenDigits-R



(b) PenDigits-L

Results IV





(a) Physionet2012

Conclusions

- Imputation strategies *drastically* affect the performance of signature-based models; we observe this most prominently in shallow signature models.
- Uncertainty-aware approaches tend to fair best, whereas uncertainty information has to be accessible *during prediction*.
- GP-PoM, our proposed end-to-end imputation strategy shows competitive performance, while considerably improving upon the existing monte-carlo approach.
- Among signature models, we observe that deep signature models are most robust in tackling irregular time series over different imputations (comparable to non-signature RNNs, yet more paramater-efficient).

Conclusions

- Imputation strategies *drastically* affect the performance of signature-based models; we observe this most prominently in shallow signature models.
- Uncertainty-aware approaches tend to fair best, whereas uncertainty information has to be accessible *during prediction*.
- GP-PoM, our proposed end-to-end imputation strategy shows competitive performance, while considerably improving upon the existing monte-carlo approach.
- Among signature models, we observe that deep signature models are most robust in tackling irregular time series over different imputations (comparable to non-signature RNNs, yet more paramater-efficient).

Conclusions

- Imputation strategies *drastically* affect the performance of signature-based models; we observe this most prominently in shallow signature models.
- Uncertainty-aware approaches tend to fair best, whereas uncertainty information has to be accessible *during prediction*.
- GP-PoM, our proposed end-to-end imputation strategy shows competitive performance, while considerably improving upon the existing monte-carlo approach.
- Among signature models, we observe that deep signature models are most robust in tackling irregular time series over different imputations (comparable to non-signature RNNs, yet more paramater-efficient).

Conclusions

- Imputation strategies *drastically* affect the performance of signature-based models; we observe this most prominently in shallow signature models.
- Uncertainty-aware approaches tend to fair best, whereas uncertainty information has to be accessible *during prediction*.
- GP-PoM, our proposed end-to-end imputation strategy shows competitive performance, while considerably improving upon the existing monte-carlo approach.
- Among signature models, we observe that deep signature models are most robust in tackling irregular time series over different imputations (comparable to non-signature RNNs, yet more paramater-efficient).

GP adapters I

Let $\mathcal{W}, \mathcal{H}$ refer to the weight space and hyperparameter space, respectively. Let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss function. Let $\mathcal{S}(\mathcal{X}^*)$ be the space of time series over the data space including missing observations. Let $F\colon \mathcal{X}^{[a,b]} \times \mathcal{W} \to \mathcal{Y}$, be some (typically neural network) model. Let

$$\mu\colon [a,b] \times \mathcal{S}(\mathcal{X}^*) \times \mathcal{H} \to \mathcal{X}$$
$$\Sigma\colon [a,b] \times [a,b] \times \mathcal{S}(\mathcal{X}^*) \times \mathcal{H} \to \mathcal{X}$$

be mean and covariance functions. The dependence on $\mathcal{S}(\mathcal{X}^*)$ is to represent conditioning on observed values.
Then the goal is to solve

$$\operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}, \boldsymbol{\eta}\in\mathcal{H}} \sum_{k=1}^{N} \overbrace{\mathbb{E}_{\mathbf{z}_k \sim \mathcal{N}(\mu(\cdot,\mathbf{x}_k,\eta),\Sigma(\cdot,\cdot,\mathbf{x}_k,\eta))}\left[\ell(F(\mathbf{z}_k, \mathbf{w}), y_k)\right]}^{E_k}. \tag{8}$$

GP adapters II

As this expectation is typically not tractable, it is estimated by Monte Carlo (MC) sampling with S samples, i.e.

$$E_k \approx \frac{1}{S}\sum_{s=1}^{S} \ell(F(\mathbf{z}_{s,k}, \mathbf{w}), y_k), \tag{9}$$

where

$$\mathbf{z}_{s,k} \sim \mathcal{N}\left(\mu(\,\cdot\,, \mathbf{x}_k, \eta), \Sigma(\,\cdot\,, \,\cdot\,, \mathbf{x}_k, \eta)\right). \tag{10}$$

Posterior moments GP adapter (GP-PoM) I

We simplify matters by taking the posterior variance at every point, and concatenate it with the posterior mean at every point, to produce a path whose evolution describes the uncertainty at every point:

$$\tau\colon [a,b] \times \mathcal{S}(\mathcal{X}^*) \times \mathcal{H} \to \mathcal{X} \times \mathcal{X}$$
$$\tau\colon t, \mathbf{x}, \eta \mapsto (\mu(t, \mathbf{x}, \eta), \Sigma(t, t, \mathbf{x}, \eta)).$$

This corresponds to solving

$$\operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}, \boldsymbol{\eta}\in\mathcal{H}} \sum_{k=1}^{N} \ell(F(\tau(\,\cdot\,, \mathbf{x}_k, \eta), \mathbf{w}), y_k), \tag{11}$$

where instead now

$$F\colon (\mathcal{X} \times \mathcal{X})^{[a,b]} \times \mathcal{W} \to \mathcal{Y}.$$

Acknowledgements

Bastian Rieck, Max Horn, Christian Bock, Karsten Borgwardt, and Patrick Kidger.

References I

[1] I. Chevyrev and A. Kormilitzin, "A primer on the signature method in machine learning," *arXiv preprint arXiv:1603.03788*, 2016.

[2] C. Toth and H. Oberhauser, "Variational gaussian processes with signature covariances," *arXiv preprint arXiv:1906.08215*, 2019.

[3] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *2019 Computing in Cardiology (CinC)*, pp. Page–1, IEEE, 2019.

[4] P. Kidger, P. Bonnier, I. P. Arribas, C. Salvi, and T. Lyons, "Deep signature transforms," in *Advances in Neural Information Processing Systems*, pp. 3099–3109, 2019.

References II

[5] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” *arXiv preprint arXiv:1409.1259*, 2014.

[6] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals,” *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[7] D. Dua and C. Graff, “UCI machine learning repository,” 2017.

[8] T. Allam Jr, A. Bahmanyar, R. Biswas, M. Dai, L. Galbany, R. Hložek, E. E. Ishida, S. W. Jha, D. O. Jones, R. Kessler, *et al.*, “The photometric lsst astronomical time-series classification challenge (plasticc): Data set,” *arXiv preprint arXiv:1810.00001*, 2018.