Modeling Irregularly Sampled Clinical Time Series

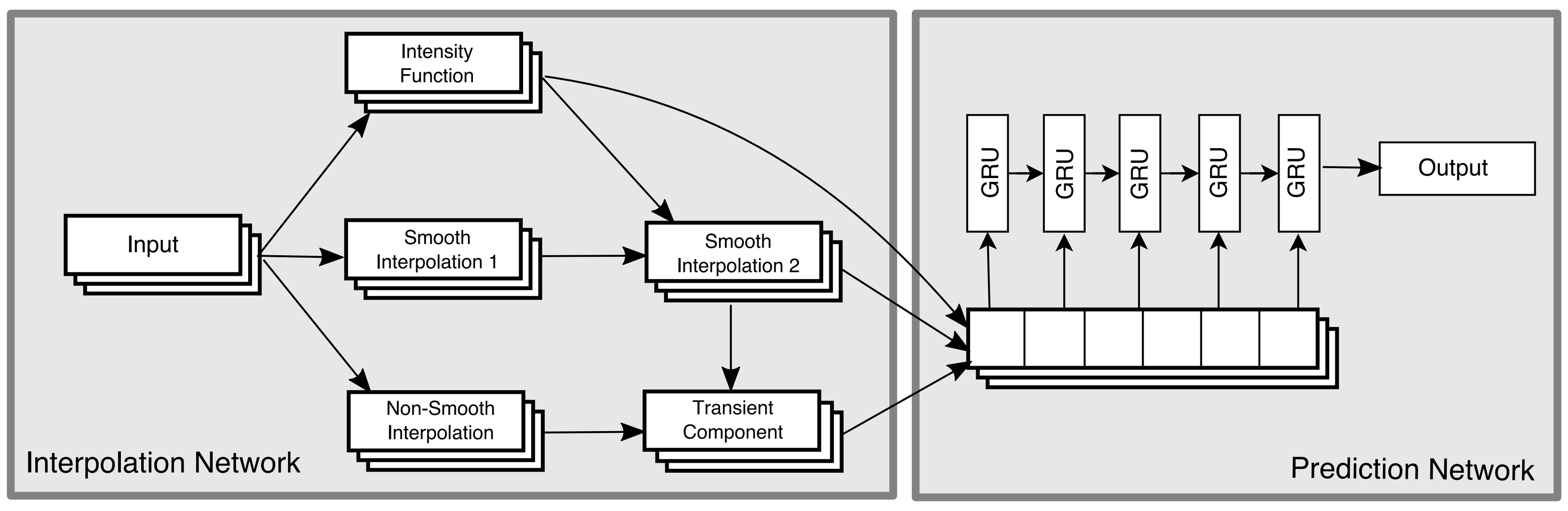
Satya Narayan Shukla and Benjamin Marlin
University of Massachusetts Amherst

Introduction

We present a novel model architecture for supervised learning with multivariate sparse and irregularly sampled data: Interpolation-Prediction Networks. The architecture is based on the use of several semi-parametric interpolation layers organized into an interpolation network, followed by the application of a prediction network that can leverage any standard deep learning model. This work is motivated by problems in the analysis of electronic health records (EHR) where it remains rare for hospital systems to capture dense physiological data streams.

Model Framework

The overall model architecture consists of two main components: an interpolation network and a prediction network.



Architecture of the proposed model

The interpolation network interpolates the multivariate, sparse, and irregularly sampled input time series against a set of reference time points. We propose a two layer interpolation network where the 1st layer performs a semi-parametric univariate interpolation for each of the D time series separately and the 2nd layer merges information from across all of the D time series at each reference time point by taking into account learned correlations $\rho_{dd'}$.

$$\text{Layer 1}\begin{cases}\hat{x}^{1c}_{kdn} = f^{1c}_{\theta}(r_k, \mathbf{s}_{dn}) = \dfrac{\sum_{j=1}^{L_{dn}} w_{dc}(r_k, t_{jdn})\, x_{jdn}}{\sum_{j=1}^{L_{dn}} w_{dc}(r_k, t_{jdn})} \\ w_{dc}(r_k, t_{jdn}) = \exp\left(-\alpha_{dc}\|r_k - t_{jdn}\|^2\right)\end{cases} \tag{1}$$

$$\text{Layer 2}\begin{cases}\hat{x}^{2c}_{kdn} = f^{2c}_{\theta}(r_k, \mathbf{s}_{dn}) = \dfrac{\sum_{d'} \rho_{dd'}\, i^c_{kd'n}\, \hat{x}^{1c}_{kd'n}}{\sum_{d'} i^c_{kd'n}} \\ i^c_{kdn} = f^{3c}_{\theta}(r_k, \mathbf{s}_{dn}) = \sum_j w_{dc}(r_k, t_{jdn})\end{cases} \tag{2}$$

where x_{jdn}, t_{jdn} and L_{dn} represent the observed value, time point for j^{th} observation and length of time series d and data case n, r_k is a reference time point, and $\mathbf{s}_{dn} = (\mathbf{t}_{dn}, \mathbf{x}_{dn})$ represent time series d for data case n as a tuple.

Output of Interpolation Network:

- Smooth Interpolant to capture the trends: $\hat{\mathbf{x}}^{21}_{dn}$, parameters: α_{d1}, $\rho_{dd'}$ (correlation parameters)
- Non-smooth interpolant to capture the transients: $\hat{\mathbf{x}}^{12}_{dn} - \hat{\mathbf{x}}^{21}_{dn}$, parameters: $\alpha_{d2} = \kappa\alpha_{d1}$, $\kappa > 1$
- Intensity function to retain information about where observations occur: $\mathbf{i}^{1}_{dn}$, parameters shared with $\hat{\mathbf{x}}^{21}_{dn}$

The prediction network can consist of any standard supervised neural network architecture (fully-connected feedforward, convolutional, recurrent, etc) thus, making our model fully modular.

Learning

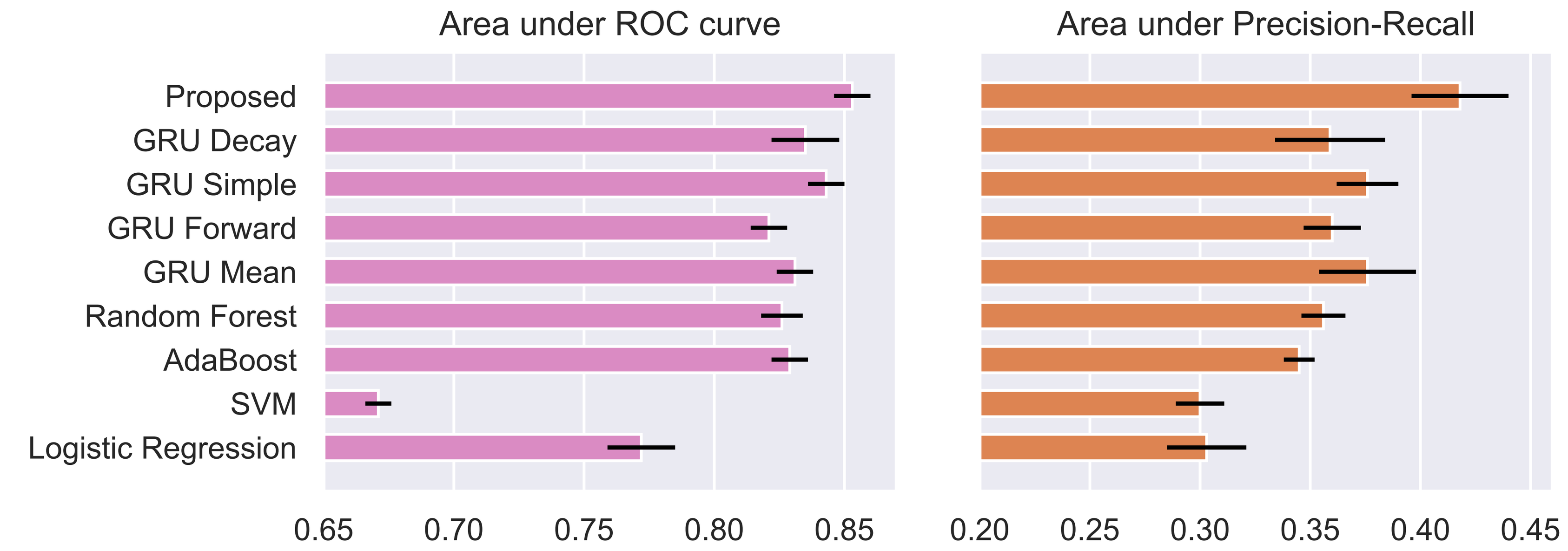
The parameters of the interpolation and prediction networks are learned end-to-end via a composite objective function consisting of supervised and unsupervised components.

$$\begin{aligned}\theta_*, \phi_* = \underset{\theta,\phi}{\operatorname{argmin}} \sum_{n=1}^{N} \ell_P(y_n, g_\phi(f_\theta(\mathbf{s}_n))) + \lambda_I\|\theta\|_2^2 + \lambda_P\|\phi\|_2^2 \\ + \delta \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{j=1}^{L_{dn}} m_{jdn}\,\ell_I(x_{jdn}, f^{21}(t_{jdn}, (1-\mathbf{m}_n)\odot \mathbf{s}_n))\end{aligned} \tag{3}$$

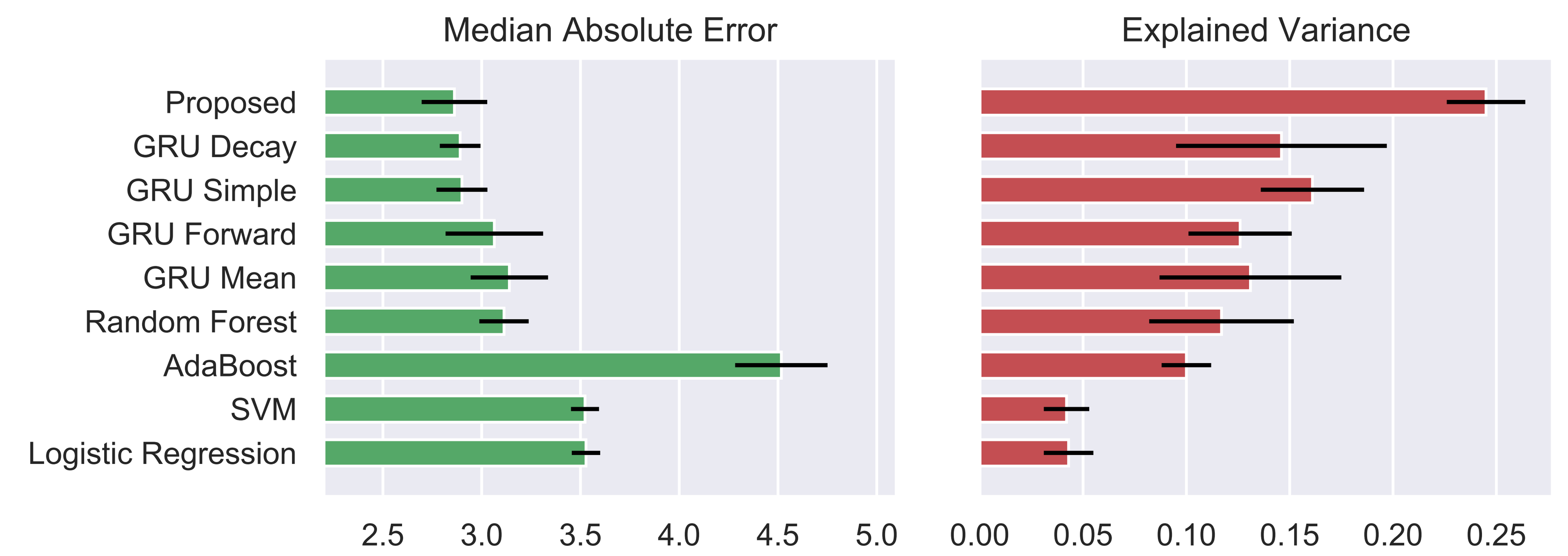
where $\mathbf{m}$ denotes mask, y_n is the label for data case n, f and g are interpolation and prediction network respectively. We use mask $\mathbf{m}$ to hold out some observed data points during learning to compute the reconstruction loss.

Experiments and Results

Our experiments are based on the publicly available MIMIC-III dataset. We focus on predicting in-hospital mortality and length of stay using the first 48 hours of data from 12 features extracted from each of the patient record. We evaluate all models using a five-fold cross-validation estimate of generalization performance. The proposed model achieves statistically significant improvements over all baseline models for both the tasks.

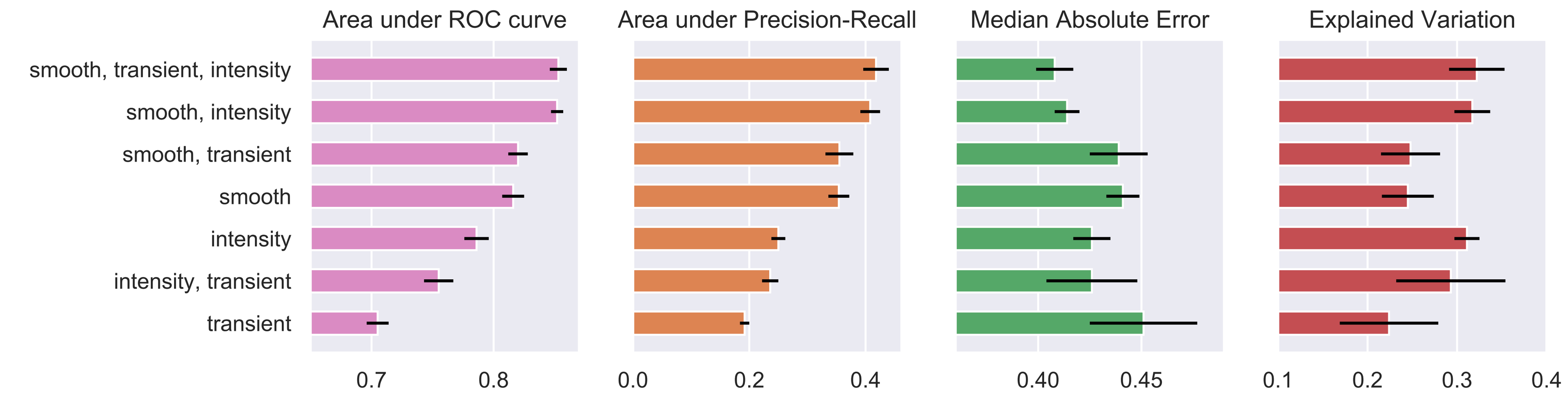


Classification performance for the mortality prediction task. GRU-Mean and GRU-Forward: missing observations replaced with global mean and last observation, GRU-Simple: input concatenated with masking variable to identify missingness and time interval indicating how long the particular variable is missing, GRU-Decay: instead of just replacing the missing value with the last measurement, it is decayed over time towards the empirical mean.



Regression performance for length of stay prediction

To assess the impact of each of the interpolation network outputs, we conduct a set of ablation experiments where we consider using all sub-sets of outputs for both the classification task and the regression task.



Performance of all subsets of the interpolation network outputs on Mortality (classification) and Length of stay prediction (regression) tasks.

{snshukla, marlin}@cs.umass.edu