Temporal-hierarchical VAE for heterogeneous and missing data handling

Daniel Barrejón-Moreno Pablo M. Olmos Antonio Artés-Rodríguez

Dept. Signal Theory and Communications, Universidad Carlos III de Madrid. Spain

Motivation

- The goal of this work is to develop a novel methodology to deal with large-scale data aggregation, where artifacts such as **missing data** or **outliers** appear very often.
- The idea of the model is to use **Deep Generative Models (DGM)** with hierarchical levels in order to find complex hidden correlations that might be present in real-world datasets.
- We evaluate the model on a medical dataset from eB2 [1]. This dataset presents challenging problems, such as very big percentages of missing data, outliers, and heterogeneous variables.
- In our results, we show **that evaluating these kind of models in practical applications is not straight-forward:** Measuring the performance on standard error metrics is not representative enough. To prove the validity of the model we evaluate the **cross-correlation** between predicted and original signals which are not seen by the model.

Which data can the model handle?

We want our model to tackle datasets that can have the following characteristics:

1. **Heterogeneous data**: Datasets present attributes of different nature *i.e.*, real, positive (continuous variables); binary, categorical (discrete values), etc. Training such models are not easy, since different data means different likelihood models or scales and therefore some likelihoods might have more effect on the optimization.
2. **Sequential data**: Following [2], we extend the standard properties of VAEs[3] to tackle temporal data streams.
3. **Corrupted data**: Following [4], we develop a model that can tackle missing data and can actually impute at those positions a value considering other observed values from other time instants, or from other variables thanks to the hidden correlations.

Proposed model

- Following [4], we only evaluate the **ELBO on the observed values**.
- In order to capture temporal dynamics of the data, we use a RNN on a continuous latent variable $\mathbf{z}$.
- A part from the continuous latent variable $\mathbf{z}$, we also use a discrete latent variable $\mathbf{s}$ that serves as a prior for $\mathbf{z}$ and also clusters the input data.
- The latent $\mathbf{z}$ is shared among all the attributes, and then a different likelihood model is considered for each attribute, *i.e.*

$$p(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{s}_t) = \prod_{d=1}^{D} p(x_{td}|\mathbf{z}_{\leq t}, \mathbf{s}_t)$$

- To alleviate the problem of working with different likelihoods, we apply a batch normalization at the encoder and a denormalization at the decoder.
- We optimize the following ELBO

$$\log p(\mathbf{X}^o) \geq \sum_{t=1}^{T} \Bigg[\mathop{\mathbb{E}}_{\substack{q(\mathbf{s}_t|\mathbf{x}_t^o,\mathbf{z}_{<t}) \\ q(\mathbf{z}_t|\mathbf{x}_t^o,\mathbf{z}_{<t},\mathbf{s}_t)}} \left[\log p(\mathbf{x}_t^o|\mathbf{z}_{\leq t}, \mathbf{s}_t)\right] - \mathop{\mathbb{E}}_{q(\mathbf{s}_t|\mathbf{x}_t^o,\mathbf{z}_{<t})} \left[KL(q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{x}_t^o, \mathbf{s}_t)||p(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{s}_t))\right] - KL(q(\mathbf{s}_t|\mathbf{x}_t^o, \mathbf{z}_{<t})||p(\mathbf{s}_t)) \Bigg].$$

HissVAE

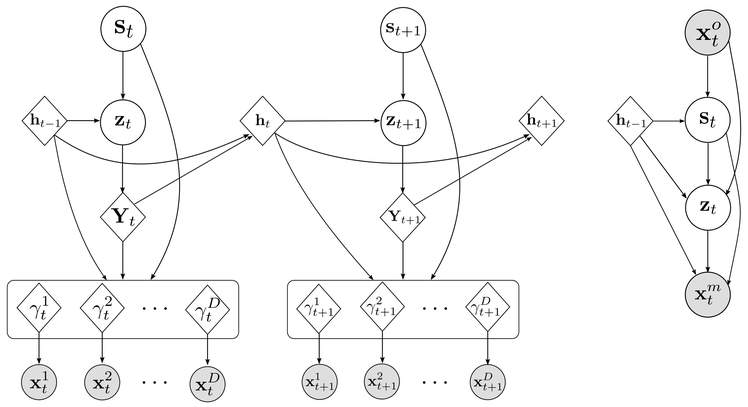


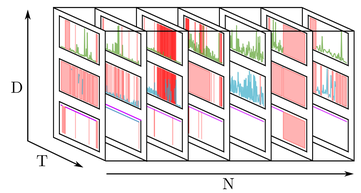
Figure 1. On the left, the generative model. On the right, the recognition model.

Real-world application: Medical, temporal and heterogeneous dataset,

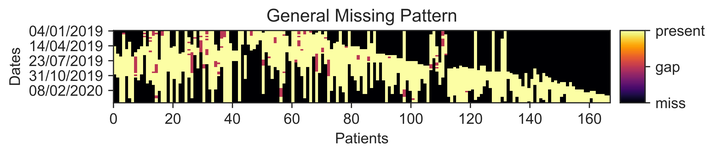
- We use data from [1] to evaluate the model.
- The type of data and the missing data percentages are shown in the table below.

	Type	Missing Rate
Distance	Positive	0.39
Steps Home	Binary	0.64
Steps Total	Positive	0.18
App Usage	Positive	0.36
Sport	Binary	0.6
Sleep	Positive	0.28
Vehicle	Positive	0.42
Emotional	Categorical	0.71

- The figure on the right shows a sketch of how the data is structured, where D refers to the attributes, T to time and N to the number of samples.

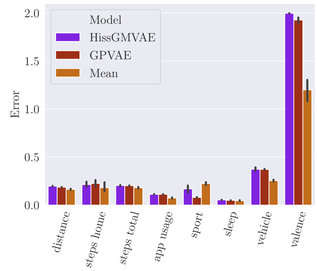


- The figure below shows the streams of data from the dataset. On yellow we show that for a given day there is data, in black there was no record of that day and in magenta a missing gaps within each patient sequence.



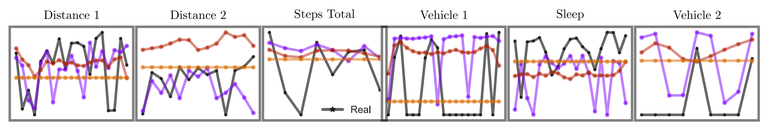
Evaluation metrics are not straight-forward

- We evaluate the imputation performance on bursts of artificial missing manually incorporated to the dataset. Such data is never seen by the model.
- We compare the performance of our model with a simple mean imputation strategy and a more novel model, the GPVAE [5], which merges the capabilities of GPs with VAEs.
- For real and positive variables, we use the NRMSE as error metrics, and categorical and binary data we use the disagreement to measure the error.
- From the error bars we see that our model and the GPVAE in principle have worse performance.



Our model captures more correlation

- The GPVAE tends to impute smoother signals due to the GP prior, thus producing results closer to the mean of the observed values.
- However, our model imputes data more correlated to the nature of the dataset, and exploits the underlying hidden correlations that are shared among the different attributes.



Cross-Correlation	Distance 1	Distance 2	Steps Total	Vehicle 1	Sleep	Vehicle 2
HissVAE	7.98	2.07	4.70	42.10	0.60	29.86
GP-VAE [5]	2.63	0.84	2.92	14.44	0.18	5.91
Mean	0.00	0.00	0.00	0.00	0.00	0.00

Conclusions

This preliminary work shows that, **in really unstable scenarios**, with big missing rates and heterogeneous data streams, **our model is able to find flexible and powerful correlations**.

References

[1] eB2: Evidence-Based Behavior, eB2.

[2] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *NeurIPS 2015*, arXiv:1506.02216.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114.

[4] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition 2020*, arXiv:1807.03653.

[5] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, "Gp-vae: Deep probabilistic time series imputation," *AISTATS 2020*, arXiv:1907.04155.

dbarrejo@ing.uc3m.es, olmos@tsc.uc3m.es, antonio@tsc.uc3m.es