When in Doubt: Neural Non-Parametric Uncertainty Quantification for Epidemic Forecasting

Authors: Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodriguez, Chao Zhang, B. Aditya Prakash.

Presenter: Akash Choudhuri.

27th October, 2021

Contents

Background

Motivation

- Statistical Approaches for the forecasting problem are fairly new.
- Mechanistic models have been the popular choice over time. However, they cannot leverage data from multiple indicators or predict composite signals.
- Deep learning approaches have been promising but they need to handle uncertainty to give reliable forecasts.
- So-called 'point' forecasts are not helpful as they are often not accounting for the uncertainty.

Contributions

- The authors designed a **Deep Generative Neural Gaussian Process Framework** for epidemic forecasting which **automatically learns** stochastic correlations between query sequences and historical data sequences for nonparametric uncertainty quantification.
- The model (EPIFNP) outputs forecast distribution based on similarity between current season's data till current week and data from each of the historical seasons in a **latent space**.
- Rigorous benchmarking on flu forecasting is performed to assert EPIFNP is **clearly superior** to other strong baselines in providing up to **2.5x more accurate and 2.4x better calibrated forecasts** using standard evaluation metrics.

Problem

Epidemic Forecasting Task

- $x_i^{(t)}$: Incidence for season i at week t.
- Current season= N+1 Current Week=t
- The snippet of time series values up to week t is given by:

$$\mathbf{x}_{N+1}^{(1\ldots t)} = \{x_{N+1}^{(1)}, \ldots, x_{N+1}^{(t)}\}$$

- We are also provided with data from past historical seasons 1 to N given by:

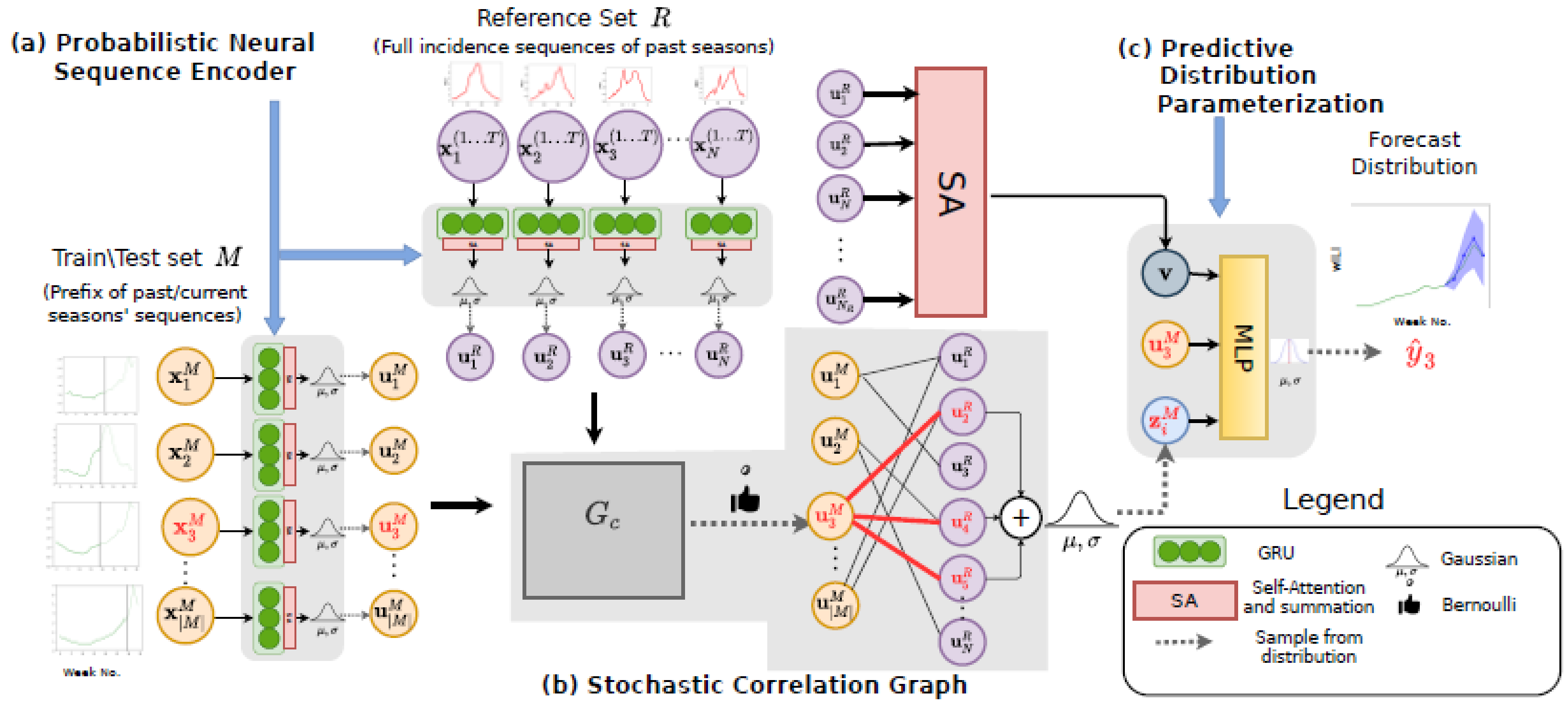
$$H = \{\mathbf{x}_i^{(1\ldots T)}\}_{i=1}^{N} \text{ where } T \text{ is number of weeks per season}$$

- Our goal is, given a dataset of historical incidence sequences H and a snippet of current season N+1 till week t, **estimate an accurate prediction of the next few future values (usually till 4 weeks in future).**
- Predict $y_{N+1}^{(t)} = x_{N+1}^{(t+k)}$ k week in future given H and $X_{N+1}^{(1\ldots t)}$.

Method

Overview

- During Training Phase, EPIFNP is trained to predict $x_i^{(t+k)}$ given $x_i^{(1,..,t)}$ as input for i ≤ N.
- **Training Set (M):** $\{(\mathrm{x}_i^{(1\ldots t)}, y_i^{(t)})$, i ≤ N, t+k ≤ T, $y_i^{(t)} = x_i^{(t+k)}$ $\}$
- **Reference Set (R):** $\{\mathrm{x}_i^{(1\ldots T)}\}_{i=1}^{N_R}$
- Elements of M are $\{\mathrm{x}_i^M, y_i^M\}_{i=1}^{N_M}$ and elements of R are $\{\mathrm{x}_i^M\}_{i=1}^{N_R}$.
- X_D is the union of the reference and training sequences



EPIFNP has 3 key steps:

- **Probabilistic neural sequence encoding:** Uses a DSM to encode the sequence $\mathbf{x}_i \in \mathbf{X}_{\mathcal{D}}$ into a variational latent embedding $\mathbf{u}_i \in \mathbf{U}_{\mathcal{D}}$.
- **Stochastic correlation graph construction:** Captures correlations between reference and training data points in the latent embedding space.
- **Final predictive distribution parameterization:** 3 stochastic latent variables, namely global and local stochastic latent variables and the stochastic sequence embeddings are passed to a MLP to give the distribution.

The Factorized Form of the predictive distribution of the training sequence is written as:

$$p(\mathbf{y}_M|\mathbf{X}_M, R) = \sum_{\mathbf{G}} \int \underbrace{p_\theta(\mathbf{U}_\mathcal{D}|\mathbf{X}_\mathcal{D})}_{\text{(a)}} \underbrace{p(\mathbf{G}|\mathbf{U}_\mathcal{D})}_{\text{(b)}}$$

$$\underbrace{p_\theta(\mathbf{Z}_M, |\mathbf{G}, \mathbf{U}_R) p_\theta(\mathbf{v}|\mathbf{U}_R) p_\theta(\mathbf{y}_M|\mathbf{U}_M, \mathbf{Z}_M, \mathbf{v})}_{\text{(c)}} d\mathbf{U}_\mathcal{D} d\mathbf{Z}_M d\mathbf{v}.$$

Probabilistic Neural Sequence Encoder

- Input Sequence is passed through a Gated Recurrent Unit (GRU):

$$\{\mathbf{h}_i^{(1)}\ldots,\mathbf{h}_i^{(t)}\} = \mathrm{GRU}(\{x_i^{(1)}\ldots,x_i^{(t)}\}).$$

- As ILI Data is delayed, we cannot depend fully on the last hidden state:

$$\{\alpha_i^{(1)}\ldots,\alpha_i^{(t)}\} = \text{Self-Atten}(\{\mathbf{h}_i^{(1)}\ldots,\mathbf{h}_i^{(t)}\}), \qquad \bar{\mathbf{h}}_i = \sum_{t'=1}^{t}\alpha_i^{(t')}\mathbf{h}_i^{(t')},$$

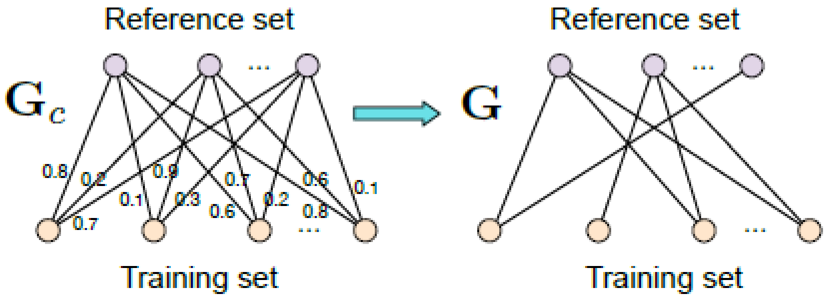
- U_i is then computed as a Gaussian having probability Distribution:

$$p_\theta([\mathbf{u}_i]_k|\mathbf{x}_i) = \mathcal{N}([g_1(\bar{\mathbf{h}}_i)]_k, \exp([g_2(\bar{\mathbf{h}}_i)]_k)),$$

Where g1 and g2 are 2 MLPs and $[.]_k$ is the k-th dimension of the variable.

Stochastic Data Correlation Graph (Most Important part of This Work)

- It is a bipartite graph from the reference set R to the training set M based on the similarity between their sequence embeddings. It models correlation among sequences.
- First we construct complete bipartite graph G_c from R to M where nodes are the sequences. Weight of each edge is the similarity between 2 sequences.
- But we use G (a stochastic binary bipartite Graph) to encode data correlations.



- The Binary Adjacency Matrix is Parameterized by:

$$p(\mathbf{G}|\mathbf{U}_{\mathcal{D}}) = \prod_{i \in R} \prod_{j \in M} \text{Bernoulli}(\mathbf{G}_{i,j}|\kappa(\mathbf{u}_i^R, \mathbf{u}_j^M)).$$

Parameterizing Predictive Distribution

Functional uncertainty is captured from different perspectives:

- **Local latent variable:** Summarizes the information of the correlated reference points for each training point and captures the uncertainty of data correlations. It is given by:

$$\mathbf{z}_{i,k}^{M} \sim \mathcal{N}(C_i \sum_{j:\mathbf{G}_{j,i}=1} h_1(\mathbf{u}_j^R)_k, \exp(C_i \sum_{j:\mathbf{G}_{j,i}=1} h_2(\mathbf{u}_j^R)_k)),$$

- **Global latent variable:** Encodes the information in all the reference points.

$$\beta_1, \ldots, \beta_{N_R} = \text{Self-Atten}(\mathbf{u}_1^R, \ldots, \mathbf{u}_{N_R}^R), \qquad \mathbf{v} = \sum_{i=1}^{N_R} \beta_i \mathbf{u}_i^R.$$

- **Sequence embedding:** A direct path from the latent embedding of the training sequence to the final prediction to enable the neural network to extrapolate beyond the distribution of the reference sequences. This is useful in novel/unprecedented patterns where the input sequence can not rely only on reference sequences from historical data for prediction.

Concatenate the 3 variables into a single vector e_i and obtain the final predictive distribution where d_1 and d_2 are MLPs:

$$\mathbf{e}_i = \mathrm{concat}(\mathbf{z}_i, \mathbf{v}_i, \mathbf{u}_i)$$

$$p(y_i|\mathbf{z}_i^M, \mathbf{v}, \mathbf{u}_i^M) = \mathcal{N}(d_1(\mathbf{e}_i), \exp(d_2(\mathbf{e}_i)))$$

Learning the Distribution

- Amortized variational inference in used to approximate the true posterior by:

$$q_\phi(\mathbf{U}_\mathcal{D}, \mathbf{G}, \mathbf{Z}_M, \mathbf{v}|R, M) = p_\theta(\mathbf{U}_\mathcal{D}|\mathbf{X}_\mathcal{D})p(\mathbf{G}|\mathbf{U}_\mathcal{D})p(\mathbf{v}|\mathbf{U}_R)q_\phi(\mathbf{Z}_M|M)$$

q_ϕ is a single layer of neural network parameterized by ϕ, which outputs mean and variance of the Gaussian distribution.

- Adam is used to maximize the evidence lower bound (ELBO) of the log likelihood. The ELBO is written as:

$$\mathcal{L} = -\mathrm{E}_{\mathbf{Z}_M, \mathbf{G}, \mathbf{U}_\mathcal{D}, \mathbf{v} \sim q_\phi(\mathbf{Z}_M|\mathbf{X}_M)p_\theta(\mathbf{G}, \mathbf{U}_\mathcal{D}, \mathbf{v}|\mathcal{D})}[\log P(\mathbf{y}_M|\mathbf{Z}_M, \mathbf{U}_M, \mathbf{v}) \\ + \log P(\mathbf{Z}_M|\mathbf{G}, \mathbf{U}_R) - q_\phi(\mathbf{Z}_M|\mathbf{X}_M)].$$

- After getting the optimal parameter, the predictive distribution of a new unseen partial sequence on the reference set is formulated by:

$$p(y^*|R, \mathbf{x}^*) = p_{\theta_{\text{opt}}}(\mathbf{U}_R, \mathbf{u}^*|\mathbf{X}_M, \mathbf{x}^*)p(\mathbf{a}^*|\mathbf{U}_R, \mathbf{u}^*) \\ p_{\theta_{\text{opt}}}(\mathbf{z}^*|\mathbf{a}^*, \mathbf{U}_R, \mathbf{u}^*)p_{\theta_{\text{opt}}}(y^*|\mathbf{u}^*, \mathbf{z}^*, \mathbf{v})d\mathbf{U}_R d\mathbf{z}^* d\mathbf{v}$$

Experiments

Results

- **Prediction Accuracy:**

Table 1: Average US National Performance: k week ahead forecasting for seasons 2014/15-2019/20.

	RMSE			MAPE			LS			CS		
Model	k=2	k=3	k=4	k=2	k=3	k=4	k=2	k=3	k=4	k=2	k=3	k=4
ED	0.73	1.13	1.81	0.14	0.23	0.33	4.26	6.37	8.75	0.24	0.15	0.42
GRU	1.72	1.87	2.12	0.28	0.31	0.356	7.98	8.21	8.95	0.16	0.2	0.22
MCDP	2.24	2.41	2.61	0.46	0.51	0.6	9.62	10	10	0.24	0.32	0.34
GP	1.28	1.36	1.45	0.21	0.22	0.26	2.02	2.12	2.27	0.24	0.25	0.28
BNN	1.89	2.05	2.43	0.34	0.46	0.51	6.92	7.56	8.03	0.18	0.22	0.25
SARIMA	1.43	1.81	2.12	0.28	0.35	0.42	3.11	3.4	3.81	0.43	0.38	0.34
RNP	0.61	0.98	1.18	0.13	0.22	0.29	3.34	3.61	3.89	0.43	0.46	0.45
EB	1.21	1.23	1.25	0.57	0.58	0.58	6.92	7	7.12	0.07	0.082	0.085
DD	0.6	0.79	0.94	0.35	0.41	0.45	3.56	3.87	4.02	0.12	0.12	0.13
EpiFNP	**0.48**	**0.79**	**0.78**	**0.089**	**0.128**	**0.123**	**0.56**	**0.84**	**0.89**	**0.068**	**0.081**	**0.035**

EPIFNP significantly outperforms all other baselines for RMSE, MAPE, LS (which measure forecast accuracy). We notice around 13% and 42% improvement over the second-best baseline in RMSE and MAPE respectively. Impressively LS of EPIFNP is 2.5 to 3.5 times less than closest baseline.

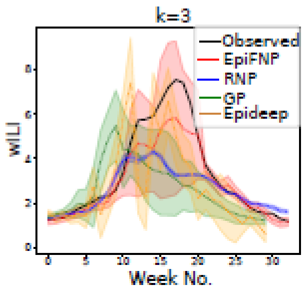


Figure 4: *Forecasts and 95% confidence bounds on 2017/18 season.*

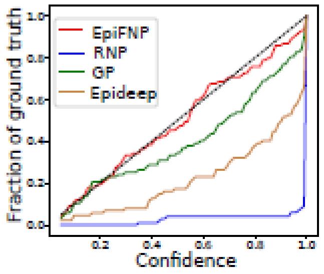
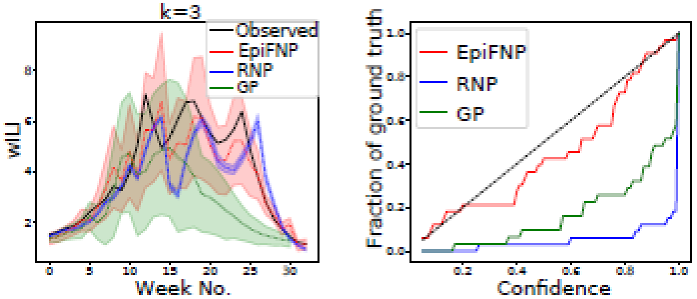


Figure 5: *CPs for* EPIFNP *and next 3 accurate baselines, k=4*



(a) 2019/20 predictions (b) 2019/20 CP

Figure 6: EPIFNP *outperforms top 2 baselines during abnormal COVID-19 season 2019/20.*

We measure how well calibrated EPIFNP's uncertainty bounds (Figure 4) are via CS. Calibration Plots (CPs) (Figure 5) show EPIFNP is much closer to the diagonal line (ideal calibration) compared to even the most competitive baselines.

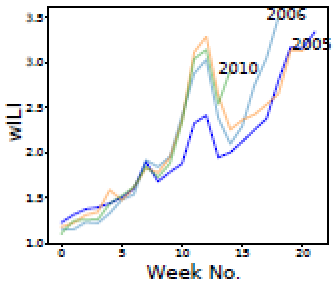


Figure 7: *2015/16 snippet & most similar seasons chosen by* EPIFNP.

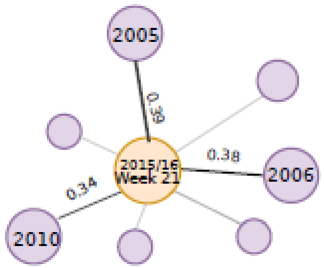


Figure 8: *Average edge probabilities for week 21 of 2015/16 season.*

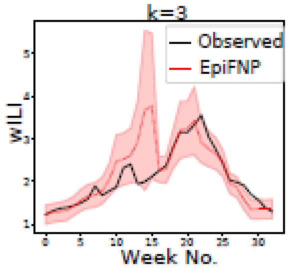


Figure 9: *Higher uncertainty around peaks*

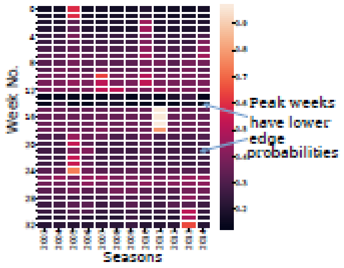


Figure 10: *SDCG Avg. edge probabilities*

Discussions/ Takeaways/ Future Work

Conclusion

- EPIFNP was the only model capable of reliably handling unprecedented scenarios e.g., H1N1 and COVID19 seasons.
- EPIFNP automatically retrieves the most relevant historical sequences matching its current week's predictions.
- EPIFNP can be affected by any systematic biases in data collection (for example, some regions might have poorer surveillance and reporting capabilities).
- In future works, EPIFNP can be extended to handle other diseases and the core technique can be adapted for other general sequence modeling.
- Future variants of EPIFNP can also use heterogeneous data from multiple sources.
- Explore incorporating domain knowledge of prior dependencies between different sources/features.

Thank You