University of Bern

Technology and Diabetes Management

101722-HS2021

Hypoglycemia prediction with ARIMA

Semester project

Students:
Emilía Tinna Sigurðardóttir
Ricardo Andrade

Supervisors:
Stavroula Mougiakakou
Matthias Fontanellaz

Date:
December 15, 2021

1 Introduction

Type 1 diabetes (T1D) is a disease affecting 15 people out of 100,000 worldwide [1]. Recently, the prevalence of diabetes in both developing and developed countries has increased along with raising costs, currently 12% of global healthcare expenses [2]. Hypoglycemia can have severe consequences if left untreated and has serious effect on quality of life [3]. Hypoglycemia is described as lower plasma glucose level than normal and symptoms can be trembling, sweating, nausea and difficulty concentrating [4].
A hypoglycemic event is described as follows [5].

- *Beginning of a CGM event:* readings below the threshold for at least 15 min is considered an event. For example, at least 15 min $<$ 54mg/dL (3.0 mmol/L) to define a clinically significant hypoglycemic event.
- *End of a CGM event:* readings for 15 min at $\geq$ 70 mg/dL (3.9 mmol/L)

Intensive insulin therapy has been proven to be significantly better than conventional insulin therapy when it comes to insulin-dependent patients and normalizing their blood glucose levels [6, 7]. On the other hand, this brings other complications. Severe hypoglycemic incidences have been reported with a threefold increase [6]. Maintaining a narrow normoglycemic range (70-120 mg/dL) while avoiding hypoglycemia is a major challenge for type 1 diabetes patients.

1.1 Model and dataset

Time series forecasting for blood glucose (BG) values can be used to reduce the harm of diabetes complications. This is done by tracking the patient's BG several times a day with continuous glucose monitoring (CGM) [8] which provides valuable information for improving the insulin management of those patients. CGM uses interstitial fluid to measure glucose in the subcutaneous tissue space and provides frequent, automated readings [9]. Predicting future glucose levels is a difficult task due to the complex human body's organisms.

The data used in this project is the OhioT1DM [10] dataset. It contains data for 12 people with type 1 diabetes over 8 weeks period. All participants were on insulin pump therapy with CGM. The data set includes the following: a CGM blood glucose level every 5 minutes, blood glucose self-monitoring (finger sticks), bolus and basal insulin doses, meals reported by participants along with estimates of carbohydrates, exercises, sleep, work, stress, and illness, along with data from fitness bands.

Data driven models can be used for these BG values to predict future hypoglycemic events. Time series forecasting is a powerful tool where past values are collected and evaluated in order to develop a model outlining the underlying relationship. This model is then used to predict future values, *i.e.* based on the past ones [11]. One of the most widely used time series models is the Autoregressive Integrated Moving Average (ARIMA). The model is combined of three different parts,

- AR, *Autoregression.* Regression of a given observation on the previous observation (called lagged observations). An observation is dependent on the previous ones.

- I, *Integrative*. Raw observations are differentiated (two consecutive observations are subtracted) which is used to make the data stationary.
- MA, *Moving Average*. The model indicates that the output variable is linearly dependent on the various past and current values of a stochastic term [12].

The model has three parameters, each representing one part of the model. The parameters are the following,

- p (order), number of the autoregressive model time lags, *i.e.* number of previous observations we consider.
- d (degree of differentiation), the amount of times the raw data has been subtracted from past values.
- q (order of the moving average), moving average window size [13].

1.2 Related work

One of the first attempts to predict future glucose levels based on previous recordings was done in 1999 by Bremer and Gough [14]. Machine learning and statistical methods have been used ever since in order to predict future glucose levels. Reifman *et al.* examined the possibility of data-driven AR models in order to capture glucose time-series data correlations and make predictions accurate enough as a function of prediction horizon. The methods were used on CGM data from nine T1D patients which was collected over a period of 5 days. The results showed for a horizon data driven AR models with a 30 minute prediction give accurate enough estimates of glucose levels for a therapy in timely manner, along with showing that AR models are good models for predictive monitoring diabetes patients [15]. Yang *et al.* used an ARIMA model to predict future BG levels based on continuous glucose monitoring (CGM) focusing on short-term characteristics of the data. Results showed that the model proposed avoids hypoglycemia deterioration by implementation of early alarms with 9.4% false rate alarm and 100% sensitivity [16]. Shanti *et al.* used an ARIMA model with Thikonov regularisation for the prediction of near future glucose concentration on hypoglycemic events. The results showed the maximum RMSE of 0.9, 2.7 and 4.2mg/dL in the horizon of prediction of 10, 20 and 30 minutes, respectively [17].

2 Methods

A Python code was provided with basic functions such as loading the data and translating raw CGM measures into a continuous time series without missing info (NaN values). A binary sequence was generated from the continuous time series in order to encode each CGM time step as no hypoglycemic event (False, 0) or a hypoglycemic event (True, 1). These past values are then used for the ARIMA model in order to create an alarming system. In order to simplify things, the threshold level for the beginning and the end of a hypoglycemic event is set to 70mg/dL, not 54mg/dL for the beginning and 70mg/dL for the end, as stated in the Introduction. Two hypoglycemic events separated with a gap less than 15 minutes are treated as one event and an event needs to be at least 15 minutes to be treated as an event,

otherwise it is neglected. A dataset of 12 patients was used for our model [10]. Each patient's dataset was split in test and training sets, and the model was fitted with only the last 200 observations from the training set, using a moving window method. Statistical tests were done to access each parameter of the model before training.

To access the stationary of the training set (parameter d) an Augmented Dickey-Fuller test was done, utilizing the appropriate function from the statsmodel package. This tests the hypothesis that a unit root (a characteristic of non-stationary time series) is present, with the null hypothesis that the time-series is stationary. This test gives us a p-value, which we can access with different significance levels to decide if the time series needs differencing. If the hypothesis is significant we differentiate the series using the dataframe structure of the pandas package and proceed with the next times for the differentiated series. This was tested for two orders of differencing and for a significance level of 0.05 (95% confidence interval.)

To access the order of the Autoregressive term (p), we can calculate the Partial Autocorrelation function (PACF) of the time series utilizing the adequate function from the statsmodel package. The PACF accesses the correlation between a lag (a past value) and it's time series. The higher the correlation for a given lag the more important it is to consider it in the AR term. For this purpose, we took the first n lags that had a correlation value higher than 0.5. For the moving average parameter (q) a similar method was used but with a pure Auto-correlation function (ACF) plot. Since an MA gives us the error of the lagged forecast, an ACF test gives us the number of lags we have to take into account to remove any autocorrelation. As in the PACF test, we took the first n lags with correlation values higher than 0.5. You can find examples of both the PACF and ACF plots in the appendix of this report.

Having accessed the parameters, the model was trained for each patient, each with their own parameters. The training and predicting of the model follows a simple protocol.

First, the training set is chosen from the dataset. Then we have to append to the training set a number of observations from the test set equal to p, since this parameter gives us the number of past observations we are using as predictors for the future observations. These are removed from the test set and the model is trained and fitted to the training set. The fitted model then predicts the next observation, which we then append to the training set. The first observation is removed from the training set, the next one from the test set is appended to the training set and we repeat the previous process until we have predicted a number of observations equal to a prediction horizon. For the prediction horizon we did tests for 30 minutes into the future (6 observations) and 60 minutes (12 observations). You can find a plot of the targets and predictions for patient 1 in the appendix of this report.

After making the predictions, a root mean square error between the test dataset and the predictions were calculated for each patient as well as a performance indication (RMSE multiplied by 400). Both the test dataset and the predictions were converted to binary sequences where 0 represents no hypoglycemic event and 1 represents an hypoglycemic event. Sensitivity and specificity was also calculated as in equation 1. As per usual, a TP is when both the test and predictions show an event, a TN is when both the test and predictions show that there is not an event. A FN is when the test shows an event but the predictions show no event and FP is when the test shows no event but the prediction shows an event. Sensitivity tells us how good the

model is at predicting events when an event actually occurs. Conversely, specificity tells us how good the model is at predicting no events when no event actually occurs.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP} \tag{1}$$

$$TN = TrueNegative, TP = TruePositive$$

$$FN = FalseNegative, FP = FalsePositive.$$

3 Data and Experimentation Setup

The data from [10] includes blood glucose samples taken every 5 minutes, already split intro training and test datasets, with the timestamps of the observations from the test dataset coming immediately after the ones from the training. Since the blood glucose samples can be as low as 50 mg/dL and well above 200 mg/dL we resize the datasets by dividing each sample by 400, making the values range from 0.1 to 1 and thus easier to manage and examine. Finally, all observations that have missing measurements were removed (due to sensor errors for example) and select only sequences that have correct measurements for at least 12 hours.

4 Results

4.1 Parameter Estimation

Utilizing the methods described before, we get the results in Table 1 for the parameters of each patient and training dataset.

Table 1: Parameter results for each patient. Training set with the last 200 observations, ADF confidence level of 95%, PACF and ACF terms with correlation higher than 0.5.

	p (AR Term)	d (order of differencing)	q (MA Term)
Patient 1	1	0	7
Patient 2	1	1	1
Patient 3	1	1	2
Patient 4	1	1	1
Patient 5	1	1	1
Patient 6	1	1	2
Patient 7	0	1	0
Patient 8	0	1	1
Patient 9	0	1	1
Patient 10	1	1	2
Patient 11	1	0	7
Patient 12	0	1	0

4.2 Observation Prediction

The results for each patient can be seen in Tables 2 and 3. NA values represent datasets where the patient had no hypoglycemic events and has such we cannot calculate values of sensitivity and specificity. These are left out of the averages for these metrics.

Table 2: RMSE, performance, sensitivity and specificity for each patient and average. Dataset of last 200 observations, prediction horizon of 6 observations (30 minutes).

	RMSE	Performance (mg/dl)	Sensitivity	Specificity
Patient 1	0.064	25.69	0.71	1.0
Patient 2	0.047	18.72	1.0	1.0
Patient 3	0.052	20.73	1.0	0.88
Patient 4	0.059	23.44	1.0	0.6
Patient 5	0.045	18.1	1.0	0.75
Patient 6	0.11	44.47	0.8	0.55
Patient 7	0.05	20.02	0.0	0.6
Patient 8	0.069	27.57	0.86	0.88
Patient 9	0.072	28.79	NA	NA
Patient 10	0.048	19.34	0.0	0.4
Patient 11	0.061	24.38	0.25	1.0
Patient 12	0.057	22.61	0.71	0.8
Average	0.061	24.49	0.67	0.77

Table 3: RMSE, performance, sensitivity and specificity for each patient and average. Dataset of last 200 observations, prediction horizon of 12 observations (60 minutes).

	RSME	Performance (mg/dl)	Sensitivity	Specificity
Patient 1	0.11	43.93	0.0	0.9
Patient 2	0.09	35.85	0.0	0.5
Patient 3	0.094	37.69	0.33	0.5
Patient 4	0.098	39.15	1.0	0.5
Patient 5	0.08	32.13	0.5	0.75
Patient 6	0.16	64.54	0.8	0.5
Patient 7	0.082	32.96	0.0	0.6
Patient 8	0.11	43.44	0.57	0.7
Patient 9	0.12	46.41	NA	NA
Patient 10	0.086	34.39	0.0	0.4
Patient 11	0.091	36.3	0.0	1.0
Patient 12	0.093	37.29	0.25	0.64
Average	0.10	40.34	0.31	0.64

5 Discussion

As table 2 and 3 show, the average Performance is 24.32 and 40.34 mg/dL for 30 and 60 minutes, respectively. Searching for papers using both the OhioT1DM dataset and the ARIMA algorithm, no papers were found. However, other algorithms have been used on the same dataset with good results. Cappon *et al.* introduced a new deep learning method for BG prediction. The method was established on a personalised bidirectional long short-term memory (LSTM), an artificial recurrent network (RNN) using the OhioT1DM dataset. The results obtained were rather good showing accuracy of RMSE = 20.20/32.19 mg/dL for prediction horizon 30/60 min, respectively. A novelty for this method is that the algorithm is also interpretable with an obtention of a "transparent model" where each feature's impact on the output of the model is explicitly expressed [18]. Three methods were used by Nemat *et al.*, Multilayer perceptron (MLP), Long short-term memory (LSTM) and Partial least squares regression (PLSR). The data from the three methods was fused using the stacked regression structure, a method used to enhance the blood glucose level performance prediction. Finally, these predictions were used to train a new PLSR model giving the final predictions. Proposing two methods, the first one using the average value of activity data appended to a CGM window to train the first-level models. The second one training the first-level models twice, once using windows of activity data and then once using CGM data. Slightly better performance was obtained from Method 1 with using a history of 30 minutes, providing RMSE = 18.99/33.39 mg/dL with a prediction horizon of 30/60 minutes, respectively. Using Method 1 with a history of 60 minutes, an accuracy of RMSE = 19.09/33.55 mg/dL with a PH 30/60 minutes, respectively [19]. Other methods gave RMSE = 18.34/32.31 mg/dL for PH of 30/60 minutes with an RNN algorithm [20] and RMSE = 19.19/32.61 mg/dL also for PH of 30/60 minutes with RNN that builds on LSTM [21]. A summary of the results can be seen in table 4.

Table 4: Literature results.

Source	Algorithm	Results (30/60, mg/dL)
Our work	ARIMA	24.32/40.34
Cappon et al	LSTM	20.20/32.19
Nemat et al	MLP/LSTM/PLSR 30 min. history	18.99/33.39
Nemat et al	MLP/LSTM/PLSR 60 min. history	19.09/33.55
Zhu et al 2020	RNN	18.34/32.31
Zhu et al 2021	RNN/LSTM	19.19/32.61

Performance values in Tables 2 and 3 show very similar results, except for patient 6. If that patient were to be removed, the performance would be 22.49 and 38.14 mg/dL which give better accuracy results. The reason for such a high performance value for Patient 6 could be due to multiple factors, namely the parameters chosen. In fact, analysing their results for the ADFuller test, we see that patient 6 has the highest p-value, signaling a time series with high non-stationary. As such, it is possible that this time series performance could have benefited from one more order of differentiation. Sensitivity and specificity values in Table 2 are 0.67 and 0.77, respectively. These values show that there are some FN and FP being detected from the predicted values. Ideally, the values need to be as close to 1, meaning that there are as few false positives and negatives. As of Table 3, the values are even lower with a sensitivity of 0.31 and specificity of 0.64. With a sensitivity that low shows that the algorithm is predicting a lot of false negatives, not giving a prediction good enough.

6 Conclusion

In conclusion, our model proved to be satisfactory in predicting blood glucose levels, with errors and performance in line with the results found in literature. In predicting hypoglycemic events, a prediction horizon of 6 showed promising results, with high sensitivities and specificities. These metrics, however, had a significant decrease when we expand the prediction horizon to 12 observations, most notably in regards to sensitivity.

With this in mind, further research can and should be made. First, in order to improve errors and performance to better match or surpass what we found in literature, a finer assessment of the parameters should be made. After the first estimation with the method proposed in this work, an iterative analysis of the model summary (provided by the ARIMA statsmodel package) should be made. This summary gives various metrics to analyze the model, among them the p-value of each coefficient of each parameter. Analyzing this summary with a small range of parameters around the ones predicted initially can, in theory, render better results.

Finally, an analysis of how the number of starting observations affect the quality of the model can give significant insight, helping to achieve an adequate tradeoff between code performance and model quality.

References

[1] M. Mobasseri, M. Shirmohammadi, T. Amiri, N. Vahed, H. H. Fard, and M. Ghojazadeh, "Prevalence and incidence of type 1 diabetes in the world: a systematic review and meta-analysis," *Health promotion perspectives*, vol. 10, no. 2, p. 98, 2020.

[2] N. Cho, J. Shaw, S. Karuranga, Y. d. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "Idf diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes research and clinical practice*, vol. 138, pp. 271–281, 2018.

[3] F. Alvarez-Guisasola, D. D. Yin, G. Nocea, Y. Qiu, and P. Mavros, "Association of hypoglycemic symptoms with patients' rating of their health-related quality of life state: a cross sectional study," *Health and Quality of Life Outcomes*, vol. 8, no. 1, pp. 1–8, 2010.

[4] J.-F. Yale, B. Paty, and P. A. Senior, "Hypoglycemia," *Canadian journal of diabetes*, vol. 42, pp. S104–S108, 2018.

[5] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel *et al.*, "International consensus on use of continuous glucose monitoring," *Diabetes care*, vol. 40, no. 12, pp. 1631–1640, 2017.

[6] D. Control and C. T. R. Group, "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus," *New England journal of medicine*, vol. 329, no. 14, pp. 977–986, 1993.

[7] U. P. D. S. U. Group *et al.*, "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (ukpds 33)," *The lancet*, vol. 352, no. 9131, pp. 837–853, 1998.

[8] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.

[9] D. Rodbard, "Continuous glucose monitoring: a review of recent studies demonstrating improved glycemic outcomes," *Diabetes technology & therapeutics*, vol. 19, no. S3, pp. S–25, 2017.

[10] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *CEUR workshop proceedings*, vol. 2675. NIH Public Access, 2020, p. 71.

[11] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[12] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[13] "Forecasting Process Details - 9.3." [Online]. Available: https://bit.ly/31KhpXD

[14] T. Bremer and D. A. Gough, "Is blood glucose predictable from previous values? A solicitation for data." *Diabetes*, vol. 48, no. 3, pp. 445–451, 1999.

[15] J. Reifman, S. Rajaraman, A. Gribok, and W. K. Ward, "Predictive monitoring for improved management of glucose levels," *Journal of diabetes science and technology*, vol. 1, no. 4, pp. 478–486, 2007.

[16] J. Yang, L. Li, Y. Shi, and X. Xie, "An arima model with adaptive orders for predicting blood glucose concentrations and hypoglycemia," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1251–1260, 2018.

[17] S. Shanthi, D. Kumar, S. Varatharaj, and S. S. Selvi, "Prediction of hypo/hyperglycemia through system identification, modeling and regularization of ill-posed data," *International Journal of Computer Science & Emerging Technologies*, vol. 1, no. 4, pp. 171–176, 2010.

[18] G. Cappon, L. Meneghetti, F. Prendin, J. Pavan, G. Sparacino, S. Del Favero, and A. Facchinetti, "A personalized and interpretable deep learning based approach to predict blood glucose concentration in type 1 diabetes." in *KDH@ ECAI*, 2020, pp. 75–79.

[19] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Data fusion of activity and cgm for predicting blood glucose levels," in *Knowledge Discovery in Healthcare Data 2020*, vol. 2675. CEUR Workshop Proceedings, 2020, pp. 120–124.

[20] T. Zhu, X. Yao, K. Li, P. Herrero, and P. Georgiou, "Blood glucose prediction for type 1 diabetes using generative adversarial networks," in *CEUR Workshop Proceedings*, vol. 2675, 2020, pp. 90–94.

[21] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, "Blood glucose prediction in type 1 diabetes using deep learning on the edge," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

Appendix

Github repository of this project: `https://github.com/Andrade1999/TDM_Project2.git`

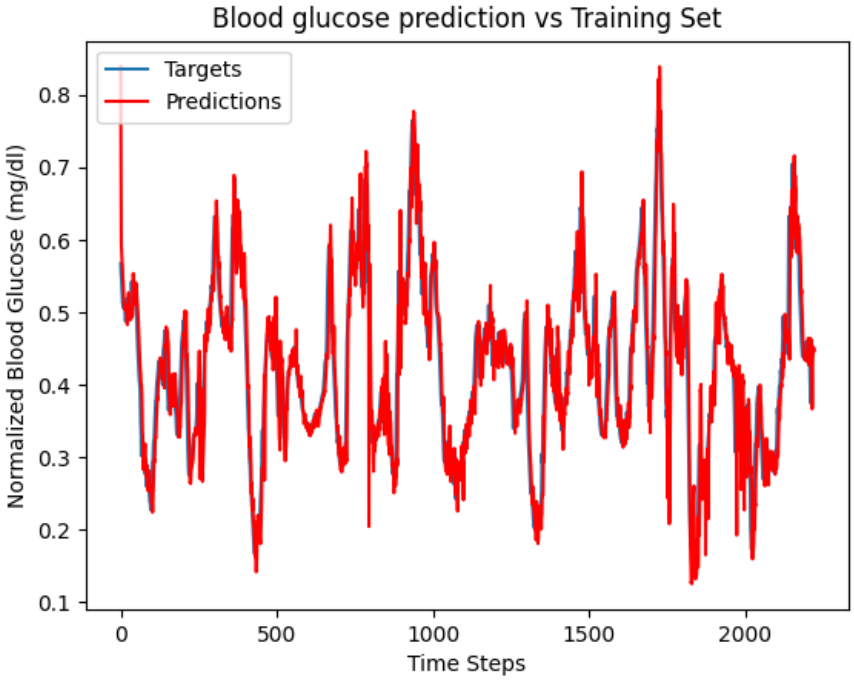

Figure 1: Predictions vs training set of patient 1.

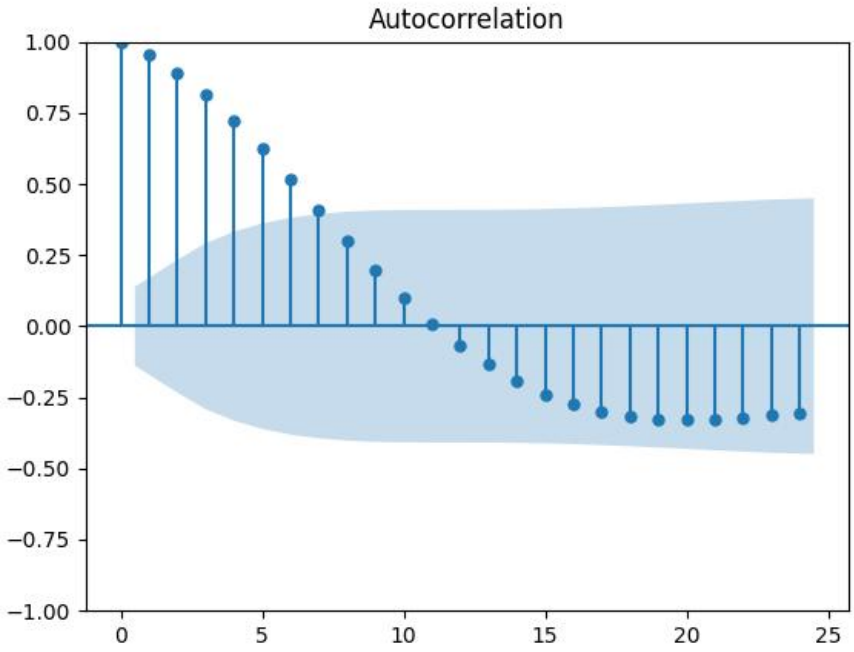

Figure 2: Autocorrelation plot for patient 1

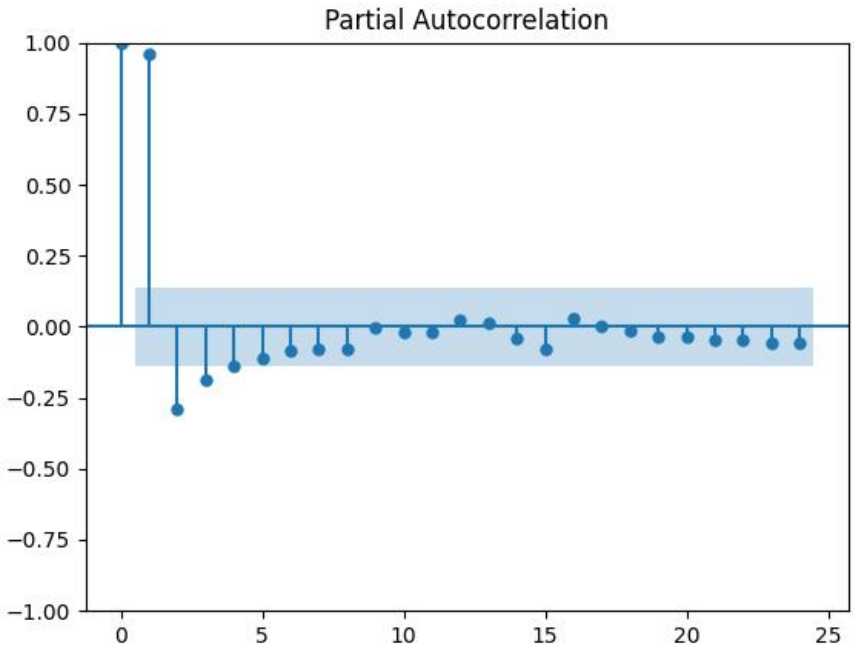

Figure 3: Partial autocorrelation plot for patient 1

Appendix - Contribution

The project was split up the following: Ricardo did most of the coding and parameter estimation research. Emilía did other literature research, writing and interpretation of results. However, we helped each other with all parts.