Time Series Analysis of Biomarkers For Multiple Myeloma

Boyang Kan
Department of Computer Science
University of Toronto
dianna.kan@mail.utoronto.ca

Yang Qu
Department of Computer Science
University of Toronto
yang.qu@mail.utoronto.ca

Shujun Yan
Department of Computer Science
University of Toronto
shujun.yan@mail.utoronto.ca

Qingyang Yu
Department of Computer Science
University of Toronto
qingyangida.yu@mail.utoronto.ca

Abstract

For multiple myeloma patients, a variety of lab test results are tracked and checked by doctors throughout the treatment process. These lab values serve as biomarkers for patients' responses to treatments. Using the longitudinal lab data, we build both auto-regressive moving average (ARMA) model and hidden Markov model (HMM) to predict patients' future lab values. To better utilize patients' baseline and treatment data, we propose changes to the basic ARMA model to account for other external variables. Our results show that this newly proposed model outperforms both ARMA and HMM in predicting future lab value. We hope the predicted future lab values can assist doctors during the treatment process and serve as a reference for doctors to see the potential treatment outcomes.

1 Introduction

Multiple myeloma (MM) is a plasma cell malignant neoplasm, and it constitutes about 10% hematologic malignant disorders [17]. Some common symptoms include bone problems, low blood count, hypercalcemia, and increased serum creatinine level [11]. MM is sensitive to many treatments, but there is no cure for this cancer [2]. Because most of the symptoms mentioned above can be tracked through lab tests, various lab tests are done at multiple time points throughout the treatment process. More importantly, these lab results are checked by doctors to monitor patients' conditions and treatment responses.

Due to the longitudinal nature of these recorded lab values, and how these lab results reflect patients' trajectories, we are motivated to build a time series model to predict future lab values, and we hope the predicted values can serve as a reference to aid the doctors during the treatment process. The doctors could look at the predicted next set of lab values for the patients at a certain time point, and adjust the medical resources allocated to each patients accordingly. Moreover, once a model is established, treatments combination can be toggled to see the potential outcomes of a patient if another set of treatments is given.

To accomplish our goal of predicting future lab values, we decide to explore both univariate and multivariate time series models. For univariate model, an auto-regressive integrated moving average (ARIMA) model is fit for each lab value, while for multivariate model, one hidden Markov model (HMM) is used to predict all the lab values. Although one would imagine that multivariate models having better performance as it has a full picture of all the lab values, we wonder if adding treatment

Preprint. Under review.

and baseline covariates to univariate models will lead to more interpretable models with competitive performance. We refer to the two univariate models with baseline and treatment information added BARMA and BARMAX.

In this report, we will show you our work and results of using both univariate and multivariate time series models to predict future biomarkers for MM patients.

2 Data

Data come from Multiple Myeloma Research Foundation (MMRF) CoMMpass registry [14]. This data set contains longitudinal data for over 1000 newly diagnosed multiple myeloma patients [14]. Researchers record these patients' lab values and treatments from initial diagnosis through their course of treatment over a minimum of 5 years. In addition to these data, demographic data of each patient at baseline and RNA-seq data are also stored in this data set [13].

There are in total 17 baseline covariates including demographics data (i.e. age, gender), multiple myeloma stage and subtype at diagnosis. For patients genetic data, principle component analysis was used to produce a 5-dimensional feature. Missing values are imputed using the mean of available data. In terms of treatment, only the most common 10 treatment combinations are considered across all the treatment, that yields 5 drugs - Dexamethasone, Lenalidomide, Bortezomib, Carfilzomib, and Cyclophosphamide. Besides boolean variables indicating whether a patient is taking these drugs, line of therapy is also recorded at each time point. Lines of therapy after the second line is grouped together into one category.

For our targeted biomarkers, 16 lab values are selected. They reflect the conditions of blood chemistry level (i.e. albumin, blood urea nitrogen (BUN), calcium, creatinine, glucose, and total protein), complete blood counts (i.e. absolute neutrophils (abs neut), hemoglobin, white blood cells (WBC), and plateletes), and serum immunoglobulins (i.e. IgG, IgA, IgM, lambda, kappa, and M-protein). We have taken 33 time points in total with the time interval being 2 months. For an interval with multiple measurements, the last observed value is taken. To make sure all lab values are on the same scale, we standardize lab values by subtracting the maximum healthy values and multiplying the results by a scalar. For a period of time with no lab value observed, we carry the last observed values forward (i.e. last observation carried forward (LOCF)). Due to the amount of missing values in the data set, this method of imputation leads to a problem where HMM only transitioning from a state back to its own state. We solved it by adding noises to the imputed values, and the noises are sampled from the standard normal distribution.

3 Related Work

Many researches have been done with MMRF data set. An R package MMRFBiolinks has been developed to help researchers to identify prognostic markers in gene expression using MMRF data set [19]. Keats et al. have conducted analysis to find molecular predictors of outcomes and drug responses using the same set of data [9]. But these researches have not dug into forecasting of biomarkers using time series models.

For time series analysis, ARIMA model is a regression-based univariate model that has been widely applied in many fields. However, as clinical cases become more and more complex in recent years, ARIMA shows limited power of prediction in the disease management field. Alim et al. have compared ARIMA with the XGBoost model in a human brucellosis study and they have found that ARIMA is not quite suitable since brucellosis is usually affected by uncertain factors and has non-linear characteristics [1]. Sato has also argued that ARIMA cannot replace doctors' experience in disease management and interventions [18]. These might be due to the two main drawbacks of ARIMA. First, it does not consider any external variables other than its own previous value. Secondly, the parameters of ARIMA are obtained from one single time series, so applying the model to predict values for other time series is not ideal.

Besides univariate models, multivariate models have also been used as they are capable of handling multiple features at the same time. HMM as a multivariate model has been used in the task of forecasting. Park et al. have used continuous HMM to predict the direction of next day's closing price in financial time series analysis [16]. HMM has also been used in time series clustering in a paper by Ghassempour et al. [5]. In a recent paper by Hussain et al., a state space model is used to predict high dimensional patients biomarkers [7]. The transition between each state is captured by a attention-based neural architecture that considers patients' genetics, baseline conditions, drugs pharmacokinetics and pharmacodynamics.

One of the objectives of our work is to see if univariate model can perform as well as the multivariate models. To address the first drawback of ARIMA, Hyndman has built an ARIMAX model, which adds an exogenous variable at the last time point t to the original ARIMA model [8]. This method has been used in some economic and financial applications. In a Thailand export study, ARIMAX outperforms ARIMA while considering trade partner's Composite Leading Indicator as an external variable [10]. Whereas, in a study related to the association between electronic cigarette use and cigarette consumption, the ARIMAX model has shown less accuracy with the risk of over-parameterization in some shorter time series [3]. Moreover, there still exists some other limitations of ARIMAX. It only considers external variables at one timestamp and inherits the second problem of ARIMA as it is also trained with one time series.

In this report, we will show univariate models that are trained using all patients time series data. Moreover, external variables (i.e. baseline and treatments) are also taken into consideration at more than one timestamps.

4 Methods

The models are trained and evaluated using a 70/30 train/test split. Hyperparameters are tuned with 5-fold cross validation on the training set, where 25% of the training data is used as the validation set. As mentioned in the data section, the missing lab values are imputed with LOCF and some noises are added to prevent HMM from always going back to its own state. However, this issue does not present in ARMA, BARMA, and BARMAX models. Therefore, univariate models are trained on both LOCF imputed data and LOCF imputed data with noises.

4.1 Univariate Model

Univariate models only consider a single feature at a time, so distinct models are trained for each of the 16 biomarkers.

4.1.1 Background

ARIMA is known as the Box-Jenkins approach proposed by Box and Jenkins, and it takes a one-dimensional time series as input [6]. We propose new univariate models for multiple myeloma biomarkers prediction on top of ARIMA.

In time series analysis, there is an assumption that future data will follow a similar pattern and trend of previous. ARIMA works on the assumption that the time series is stationary, which means constant mean, variance and autocorrelation. If the stationary assumption is not satisfied, the differencing transformations are usually applied before passing the time series to ARIMA model.

The ARIMA model as follows:

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

It consists of two main parts: autoregressive and moving average. Autoregressive ($\phi_1 Y_{t-1} + ... + \phi_p Y_{t-p}$) uses the information of its own lagged values. Moving average ($\theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}$) takes lagged residuals into consideration.

There are three stages in the ARIMA approach: model selection, parameter estimation, and model checking [12]. In the model selection part, three hyperparameters p, q, d are identified using Augmented Dickey-Fuller (ADF) test [15], the autocorrelation function (ACF) and the partial autocorrelation function (PACF). They represent the number of lags of predicted value, the number of lags of residuals, and the minimum number of differencing for stationary respectively. In parameter estimation, coefficients($\phi_1, \phi_2, ..., \phi_p$ and $\theta_1, \theta_2, ..., \theta_q$) in the formula are estimated with the selected model. Finally, residuals independence are usually tested with Ljung–Box test to evaluate the model performance.

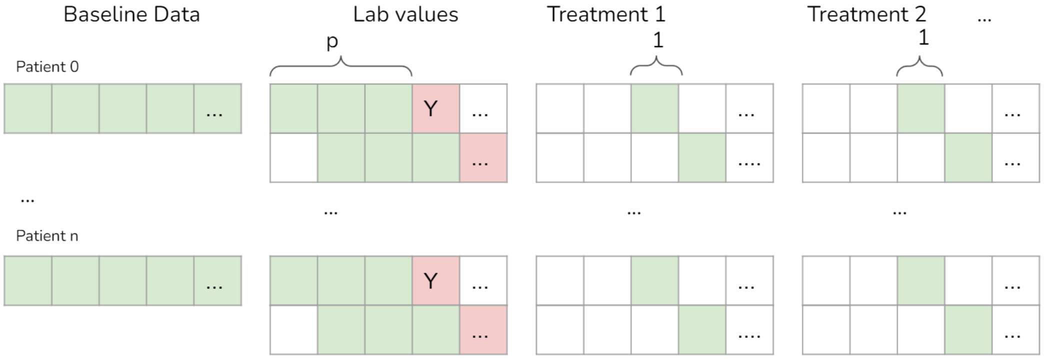


Figure 1: BARMA model data composition; green boxes signal for the values used in this sample from the original data (boxes); red boxes signal for the expected value for this sample

4.1.2 ARMA Model

The ARMA model is based on the ARIMA model, and serves as a baseline for comparing the performance with the improved BARMA and BARMAX model. The differences between ARMA model and the traditional ARIMA model are that in ARMA model, the differencing order is 0, and ARMA model can take multiple time series as input, whereas the original ARIMA model can only take one time series as input.

To determine the differencing order, we have experimented with autocorrelation graphs and Augmented Dickey Fuller test from the statsmodel package. Both methods give a suggested differencing order of zero for this dataset. For further experimentation, we have used a difference order of one to train the model, but the resulting performance is less accurate than with a difference order of zero. Thus, the three univariate models that are built for this paper all use the difference order of 0, taking out the I term in ARIMA model to become ARMA model.

For computation, ARMA model is trained with all of the patients lab values using AR and MA algorithms but in a two dimensional way. The equation is as follows:

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

Y represents the time series of lab values and ϵ represents the residuals that are used in calculating the moving average.This equation is similar to the original ARIMA model, but Y contains the lab values for all patients, so that all samples from all of the patients are concatenated together to form a large sample set.

4.1.3 BARMA Model

Since the baseline data of the patients, including genomics and demographics, and treatments given to the patients both affect the lab value considerably, this BARMA model considers the baseline data for the patients and the most recent treatments given to the patients (Figure 1). The equation is as follows:

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \gamma B + \beta X_{t-1} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

Other than the same part of ARMA, B represents the baseline data for the patients, and X represents the treatments. All variables are included as weighted features in the formula.

4.1.4 BARMAX Model

The BARMA model only considers the most recent treatments, but in reality the treatments may take more time to be reflected and have an effect on the lab values. Thus, the BARMAX model enhances BARMA model by including not only the baseline data and the time series of lab values, but also a time series of the treatments that occur before the prediction time. All of the variables added are

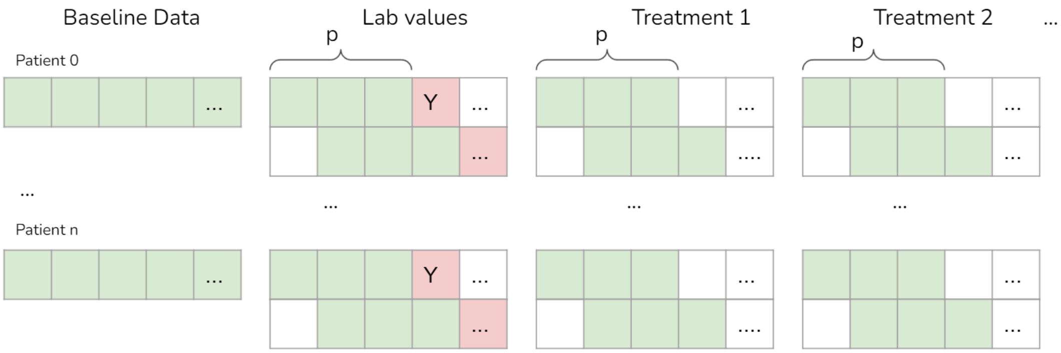


Figure 2: BARMAX model data composition; green boxes are the values used in this sample from the original data (boxes); red boxes are the expected value for this sample

again included on top of the base ARMA model. In this version of the BARMAX model, the lagged treatments to be considered are the ones that occur at the same time as the lab values that are included in the model, so it uses the same hyperparameter p as its window of consideration treatments (Figure 2). The equation is as follows:

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \gamma B + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$$

4.2 Multivariate Model

Multivariate models are able to handle historical information from more than one feature at one time, which makes it more comprehensive compared to univariate models.While there are many choices of multivariate models in time series analysis, we choose to explore discrete HMM.

The relationship of hidden states is modelled by First Order Markov process and the observed lab values are only dependent on the hidden state at the current time point. During the training phase, the model estimates the transition probabilities underlying the Markov process and the emission probabilities of observing lab values at each time given the current hidden state. The number of hidden states is tuned based on the log-likelihood (Appendix A.3). Prediction of the lab value at the next time step is performed using Algorithm 1 below:

Algorithm 1 PREDICTION OF LAB VALUES

```
Input : Transition probability of moving for state $\boldsymbol{z}_i$ to state $\boldsymbol{z}_j$, $\boldsymbol{A}_{\boldsymbol{z}_i \boldsymbol{z}_j}$, for all $i$ and $j$,
        Means of $k^{th}$ lab value for hidden state $\boldsymbol{z}_i$, $\boldsymbol{M}_{\boldsymbol{z}_i k}$, for all $i$ and $k$,
        Variance of $k^{th}$ lab value for hidden state $\boldsymbol{z}_i$, $\boldsymbol{V}_{\boldsymbol{z}_i k}$, for all $i$ and $k$,
        Hidden state at time $t$, $\boldsymbol{z}_t$
Output : Predicted $k^{th}$ lab values at time $t+1$
Predict the most likely hidden state at time $t+1$, $\boldsymbol{z}_{t+1}$ = $\mathrm{argmax}_{\boldsymbol{z}}\ \boldsymbol{A}_{\boldsymbol{z}_t \boldsymbol{z}}$
Find the corresponding mean $\boldsymbol{M}_{\boldsymbol{z}_{t+1} k}$ and $\boldsymbol{V}_{\boldsymbol{z}_{t+1} k}$
Draw a sample $\boldsymbol{X}$ from Gaussian distribution with mean of $\boldsymbol{M}_{\boldsymbol{z}_{t+1} k}$ and variance of $\boldsymbol{V}_{\boldsymbol{z}_{t+1} k}$
Make $\boldsymbol{X}$ be the predicted $k^{th}$ lab values at $t+1$
```

5 Results

Models are evaluated on test set and compared using root mean-squared-error (RMSE).

For univariate models using the LOCF imputed data, the RMSE scores range from 0.025 to 0.865 across the three models (Table 1). The average RMSE of all 16 lab values for ARMA, BARMA, and BARMAX are 0.202, 0.169, and 0.152 respectively, with BARMA about 16% lower than ARMA and BARMAX about 24.6% lower than ARMA. The performances of the three models are to our

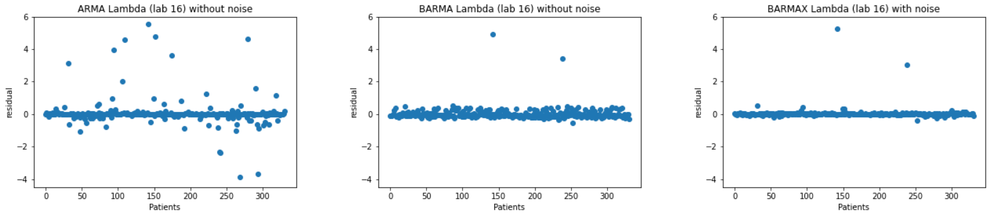


Figure 3: Residual plots for serum Lambda (lab 16) for all patients using LOCF imputed data

expectation that as more information is included in the model, the predictions should become more accurate. Looking at each biomarker individually, BARMAX outperforms the other two models in 13 out of 16 biomarkers, ties with BARMA for Serum IgM (lab 15), and ARMA performs the best in only 2 biomarkers. The four important and most commonly used biomarkers in clinics are serum Kappa, IgA, IgG, and Lambda. According to our results in Table 1, ARMA has the best performance for IgA and BARMAX has the best performance for the rest of them. Overall, our improved models BARMA and BARMAX give a more accurate forecast for the next lab value than the original ARMA model, with BARMAX being the best amongst them.

To better visualize the performances of models, Figure 3 shows the residual scatter plots for all patients on Serum Lambda (lab 16). More points lie close to zero in the BARMAX model than ARMA and BARMA, indicating that for BARMAX more patients have predicted values close to the true values compared the other two models. However, there are several outliers in the plots that significantly affect the overall RMSE for the models. We have examined all of the outliers in all 16 labs. We have found a common property shared among patients with large residuals - the actual last lab value changes dramatically compared to its previous lab values. Since the models are predicting based on the previous trend, drastic changes become hard to capture. Figure 4 shows one of the patients that appear to be an outlier, with the green x being the predicted value and the orange dot being the true value. All the other outliers follow the similar behaviour that the prediction lies along the previous trend but the actual value suddenly increases or decreases causing a large residual.

By including random noise in the LOCF imputed data, the RMSE scores lie between 0.905 and 1.192 (Table 2), which have a significant increase compared to the ones without noise considering the original scale. However, this does not mean the models chosen with noise data are worse than models for LOCF imputed data without noise since the scale of noise matters. The average RMSE across all labs are 1.065, 1.062, 1.054 for ARMA, BARMA, BARMAX accordingly. The difference between three models decreases, where BARMA is 0.3% lower and BARMAX is 1% lower than ARMA (Table 2). When predicting the four biomarkers used for tracking disease progression, BARMA performs the best in IgA and BARMAX is the most accurate model for Kappa, IgG, and Lambda. The reason for this phenomenon could be that IgA is highly related to the most recent treatment, whereas treatments take longer time to reflect on Kappa, IgG, and Lambda. Although BARMAX still has the best performance among the models for most lab values, the baseline ARMA model shows better prediction performance for calcium, creatinine, total protein and WBC. That is because ARMA usually fits a general model without considering the baseline and treatment information. When fitting the LOCF imputed data with noise, ARMA has less risk of overfitting since it has smaller number of parameters (Appendix A.4).

If we look at the residual plots for those four important biomarkers, most of the plots show random patterns (Appendix A.2). This indicates that there is no residual autocorrelation, and residuals for each patient are independent, which satisfies the assumption of a stable univariate process. Overall, the final selected models for biomarkers with noise show stable prediction performance.

By comparing the univariate model and multivariate model (Table 2), we find the average RMSE of univariate models (1.06) is lower than the RMSE of HMM (1.672) at 37%. When we look at the residual plots (Appendix A.2), there are some outliers observed in HMM residuals, which shows evidence against residual independence. The poor prediction outcome of HMM is because HMM model stratifies patients into different states and generates distribution for each state instead of making prediction for each patients separately. Thus, we conclude that univariate models outperform baseline HMM in multiple myeloma biomarker prediction.

Table 1: Univariate models test RMSE results with missing data imputed by carrying the last available data forward. The best results are in bold. CBC, complete blood count.

		Test RMSE		
Lab No.	Lab Name	ARMA	BARMA	BARMAX
1	CBC abs neut	0.196	0.183	**0.144**
2	Chem albumin	0.055	0.097	**0.049**
3	Chem BUN	0.136	0.130	**0.116**
4	Chem calcium	0.832	0.754	**0.752**
5	Chem creatinine	0.075	0.089	**0.068**
6	Chem glucose	0.050	0.082	**0.027**
7	CBC hemoglobin	**0.144**	0.174	0.155
8	Serum Kappa	0.111	0.117	**0.108**
9	Serum M-protein	0.095	0.055	**0.038**
10	CBC platelet	0.148	0.163	**0.146**
11	Chem total protein	0.026	0.026	**0.025**
12	CBC WBC	0.174	0.142	**0.136**
13	Serum IgA	**0.208**	0.213	0.224
14	Serum IgG	0.067	0.070	**0.064**
15	Serum IgM	0.045	**0.043**	**0.043**
16	Serum Lambda	0.865	0.373	**0.340**
Mean		0.202	0.169	**0.152**

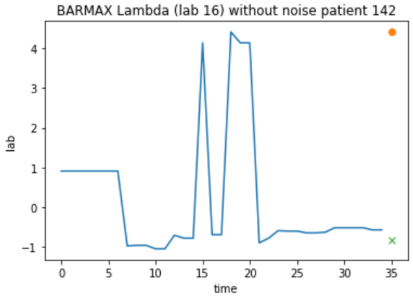


Figure 4: Serum Lambda (lab 16) values for patient 142. Green x: BARMAX (trained with LOCF imputed data) predicted value; Orange dot: actual value.

6 Limitations and Future Work

While BARMAX shows great performance among the models, there are still limitations to our methods. Earlier in the data preprocessing stage, each lab value is standardized by subtracting the maximum healthy value and multiplying the results by a scalar. In this way, a positive number represents an unhealthy lab value. Then, we imputed the missing values using LOCF with a random noise sample from standard normal distribution. However, we might change the sign of the imputed lab values by adding a random noise which potentially changes the health condition of a patient with no supported information. This can be improved in the future by experimenting with various techniques for handling missing data. For example, Che et al. has proposed a recurrent neural network based method for healthcare missing data imputation [4].

Even though we have made some progress with our new univariate models when compared to the baseline ARMA model, there are still some aspects we could improve. Currently in our BARMAX model, we use p as lag order of treatments, which is the same as the lag order of lab values. However, an independent lag order for treatment could be applied to help in achieving a more accurate prediction performance. Secondly, we find the overfitting issue when including a large number of lagged orders and too much baseline and treatment information. Regularization can be applied to deal with that.

Table 2: All models test RMSE results with missing data imputed by carrying the last available data forward and adding noises. The best results are in bold. CBC, complete blood count.

		Test RMSE			
Lab No.	Lab Name	ARMA	BARMA	BARMAX	HMM
1	CBC abs neut	1.115	**1.059**	1.094	1.588
2	Chem albumin	1.090	1.086	**1.077**	1.704
3	Chem BUN	1.068	**1.062**	1.089	1.595
4	Chem calcium	**1.062**	1.066	1.084	1.213
5	Chem creatinine	**0.923**	0.938	0.980	2.104
6	Chem glucose	0.951	0.951	**0.928**	1.403
7	CBC hemoglobin	1.066	1.077	**1.028**	1.920
8	Serum Kappa	0.913	0.924	**0.905**	1.521
9	Serum M-protein	1.144	1.147	**1.070**	1.724
10	CBC platelet	1.142	1.137	**1.093**	2.112
11	Chem total protein	**1.047**	1.072	1.075	1.296
12	CBC WBC	**1.044**	1.047	1.048	1.506
13	Serum IgA	1.078	**1.056**	1.099	1.963
14	Serum IgG	1.013	1.008	**0.982**	1.608
15	Serum IgM	1.192	1.172	**1.161**	1.455
16	Serum Lambda	1.191	1.191	**1.153**	2.047
Mean		1.065	1.062	**1.054**	1.672

While we expect to see HMM performs better than univariate models, it failed to achieve the smallest RMSE in predicting all lab values. One limitation of using HMM is that the output prediction is a hidden state with a corresponding Gaussian distribution, rather than a point prediction. In our algorithm 1, only one random sample is drawn from the corresponding Gaussian distribution. To improve the prediction accuracy and stability, one possible solution is to avoid the randomness at one sample is to draw multiple samples and use average of the obtained samples as the prediction. Moreover, unlike BARMAX, HMM does not account for the baseline covariates and treatment data, possible ways to modify the current implementation are to connect the hidden states and external information via an attention-based neural network [7] or to treat external information as local latent variables [20]. In this way, not only the historical information of observed lab values are considered, but also the dynamic treatments and static baseline covariates are taken into account.

Besides these limitations, some other future work can also be beneficial. With the trained model, one could substitute different treatments combination in to investigate potential future lab values. Doctors might be interested in these values as they reflect potential treatment outcomes. Other chronic diseases such as diabetes, heart diseases can also make use of this model and make time series predictions on their own biomarkers.

7 Conclusion

In this report, we have applied univariate and multivariate time series models to predict future biomarkers for MM patients. The newly proposed model BARMAX, based on the ARMA model, not only incorporates treatments and baseline covariates in the model, but also allows simultaneous training and prediction for multiple time series. The BARMAX model has shown its ability in forecasting by outperforming ARMA, BARMA, and even baseline HMM models. Its linear structure provides excellent interpretability. Moreover, models that predict future biomarkers values create opportunities for the doctors to select better and more personalized treatments for the patients. Last but not least, we believe BARMAX can also be used for forecasting in other fields where external variables present.

8 Team Contributions

Equal contribution. Listing order is random. Boyang and Qingyang designed and implemented the univariate models. Yang and Shujun are in charge of the multivariate model.

Acknowledgement

We would like to thank Dr. Rahul G. Krishnan and our TA Michael Cooper for the valuable feedback and advice they have provided. We would also like to thank the contributors of `https://github.com/clinicalml/ml_mmrf` for providing code base for MMRF data preprocessing [13].

References

[1] Mirxat Alim, Guo-Hua Ye, Peng Guan, De-Sheng Huang, Bao-Sen Zhou, and Wei Wu. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. *BMJ open*, 10(12):e039676, 2020.

[2] Kenneth C Anderson, Melissa Alsina, William Bensinger, J Sybil Biermann, Asher Chanan-Khan, Adam D Cohen, Steven Devine, Benjamin Djulbegovic, Cristina Gasparetto, Carol Ann Huff, et al. Multiple myeloma. *Journal of the National Comprehensive Cancer Network*, 7(9):908–942, 2009.

[3] Emma Beard, Robert West, Susan Michie, and Jamie Brown. Association of prevalence of electronic cigarette use with smoking cessation and cigarette consumption in England: a time–series analysis between 2006 and 2017. *Addiction*, 115(5):961–974, 2020.

[4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 2018.

[5] Shima Ghassempour, Federico Girosi, and Anthony Maeder. Clustering multivariate time series using hidden markov models. *International journal of environmental research and public health*, 11(3):2741–2763, 2014.

[6] Keith William Hipel, Angus Ian McLeod, and William C Lennox. Advances in box-jenkins modeling: 1. model construction. *Water Resources Research*, 13(3):567–575, 1977.

[7] Zeshan Hussain, Rahul G Krishnan, and David Sontag. Neural pharmacodynamic state space modeling. *arXiv preprint arXiv:2102.11218*, 2021.

[8] Rob Hyndman. The arimax model muddle. `https://robjhyndman.com/hyndsight/arimax/`, Oct 2010.

[9] Jonathan J Keats, Gil Speyer, Austin Christofferson, Christophe Legendre, Jessica Aldrich, Megan Russell, Lori Cuyugan, Jonathan Adkins, Alex Blanski, Meghan Hodges, Dan Rohrer, Sundar Jagannath, Ravi Vij, Gregory Orloff, Todd Zimmerman, Ruben Niesvizky, Darla Liles, Joseph W. Fay, Jeffrey L. Wolf, Robert M Rifkin, Norma C Gutierrez, Mmrf CoMMpass Network, Jennifer Yesil, Mary Derome, Seungchan Kim, Winnie Liang, Pamela G. Kidd, Scott Jewell, John David Carpten, Daniel Auclair, and Sagar Lonial. Molecular predictors of outcome and drug response in multiple myeloma: An interim analysis of the Mmrf CoMMpass study. *Blood*, 128(22):194, 2016.

[10] Chaleampong Kongcharoen and Tapanee Kruangpradit. Autoregressive integrated moving average with explanatory variable (ARIMAX) model for Thailand export. In *33rd International Symposium on Forecasting, South Korea*, pages 1–8, 2013.

[11] Robert A Kyle, Morie A Gertz, Thomas E Witzig, John A Lust, Martha Q Lacy, Angela Dispenzieri, Rafael Fonseca, S Vincent Rajkumar, Janice R Offord, Dirk R Larson, et al. Review of 1027 patients with newly diagnosed multiple myeloma. *Mayo Clinic Proceedings*, 78(1):21–33, 2003.

[12] ASC Ling, G Darmesah, KP Chong, and CM Ho. Application of arimax model to forecast weekly cocoa black pod disease incidence. *Mathematics and Statistics*, 7(4A):29–40, 2019.

[13] Machine learning with data from the multiple myeloma research foundation (ML-MMRF). `https://github.com/clinicalml/ml_mmrf`.

[14] MMRF data set. `https://isb-cancer-genomics-cloud.readthedocs.io/en/latest/sections/data/MMRF_about.html#mmrf-data-set`.

[15] Rizwan Mushtaq. Augmented dickey fuller test. 2011.

[16] Sang-Ho Park, Ju-Hong Lee, Jae-Won Song, and Tae-Su Park. Forecasting change directions for financial time series using hidden markov model. In *International Conference on Rough Sets and Knowledge Technology*, pages 184–191. Springer, 2009.

[17] S Vincent Rajkumar and Shaji Kumar. Multiple myeloma: diagnosis and treatment. *Mayo Clinic Proceedings*, 91(1):101–119, 2016.

[18] Renato Cesar Sato. Disease management with ARIMA model in time series. *Einstein (Sao Paulo)*, 11:128–131, 2013.

[19] Marzia Settino, Mariamena Arbitrio, Francesca Scionti, Daniele Caracciolo, Giuseppe Agapito, Pierfrancesco Tassone, Pierosandro Tagliaferri, Maria Teresa Di Martino, and Mario Cannataro. Identifying prognostic markers for multiple myeloma through integration and analysis of mmrf-commpass data. *Journal of Computational Science*, 51:101346, 2021.

[20] Kristen A. Severson, Lana M. Chahine, Luba Smolensky, Kenney Ng, Jianying Hu, and Soumya Ghosh. Personalized input-output hidden markov models for disease progression modeling. *medRxiv*, 2020.

A Appendix

A.1 Residual Comparisons for Kappa, IgG, IgA, and Lambda in univariate models without noise

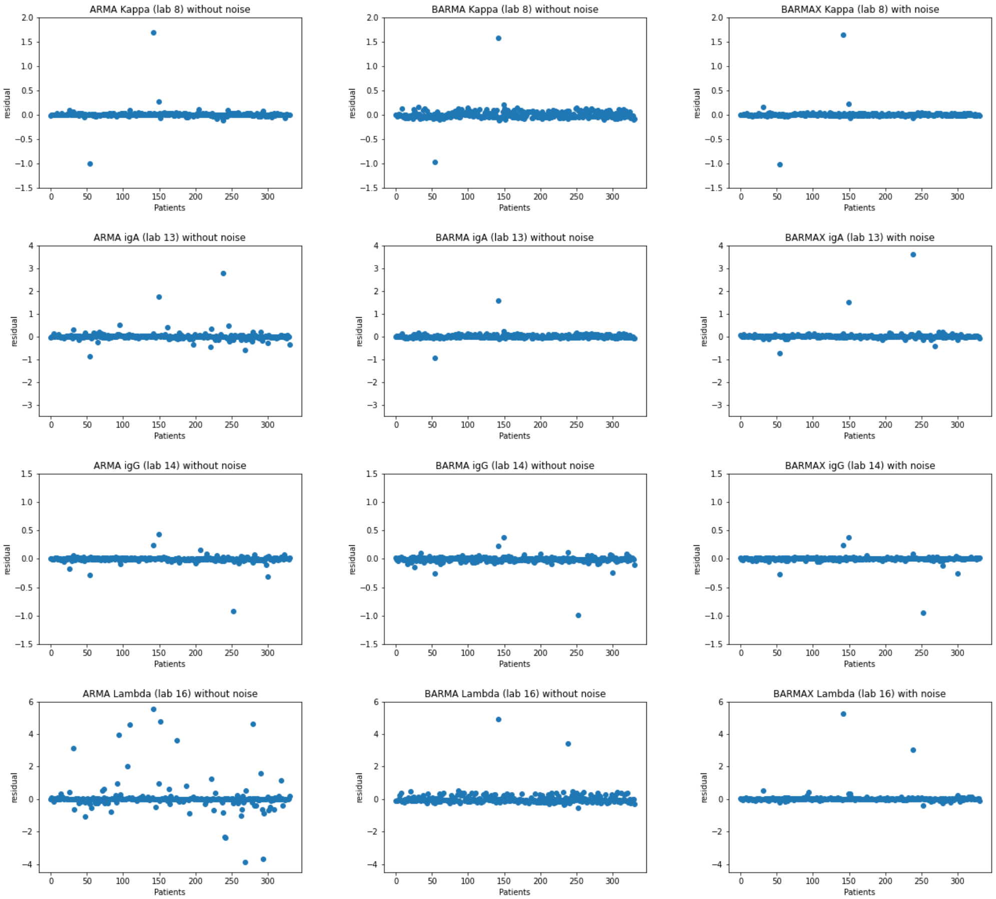
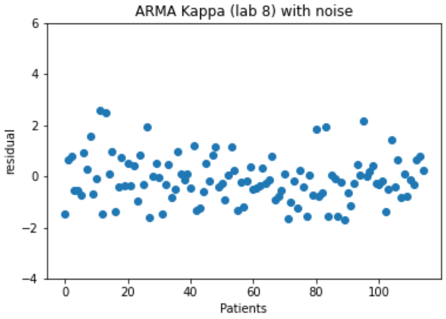
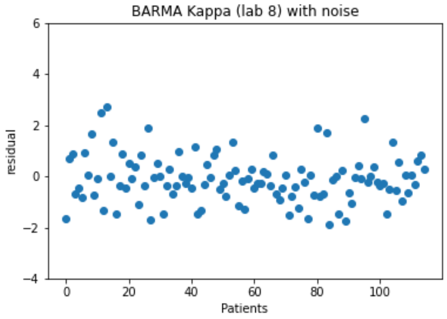
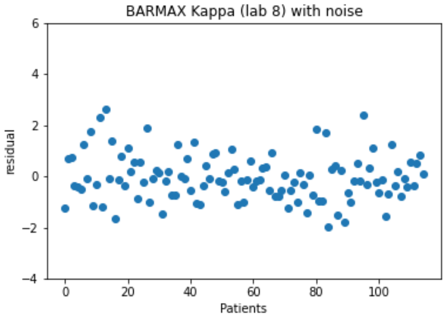


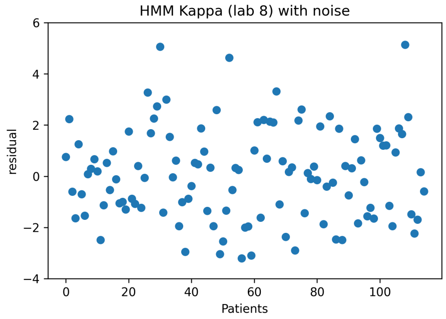
Figure 5: Residual plots for Kappa, IgG, IgA, and Lambda without noise

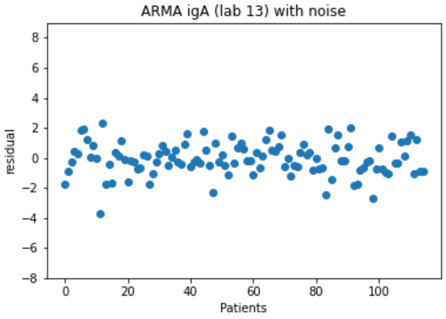
A.2 Residual Comparisons for Kappa, IgG, IgA, and Lambda in all models with noise

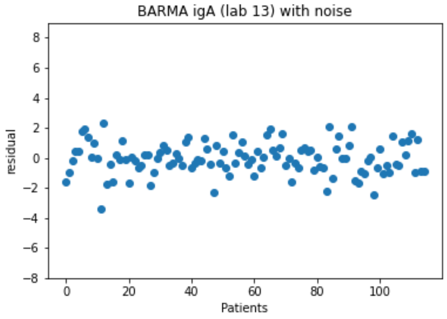

ARMA Kappa (lab 8) with noise

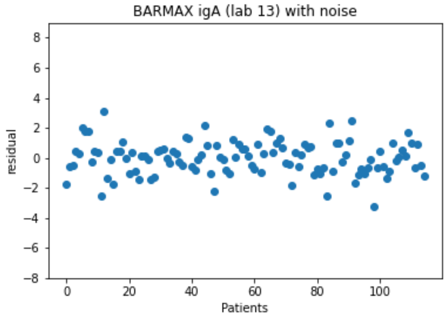

BARMA Kappa (lab 8) with noise

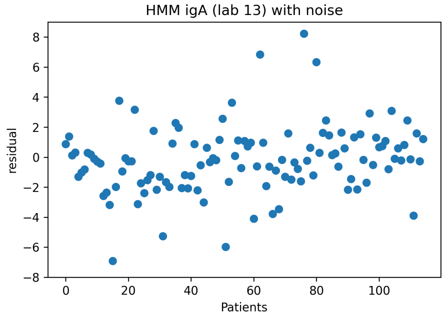

BARMAX Kappa (lab 8) with noise


HMM Kappa (lab 8) with noise


ARMA igA (lab 13) with noise


BARMA igA (lab 13) with noise


BARMAX igA (lab 13) with noise


HMM igA (lab 13) with noise

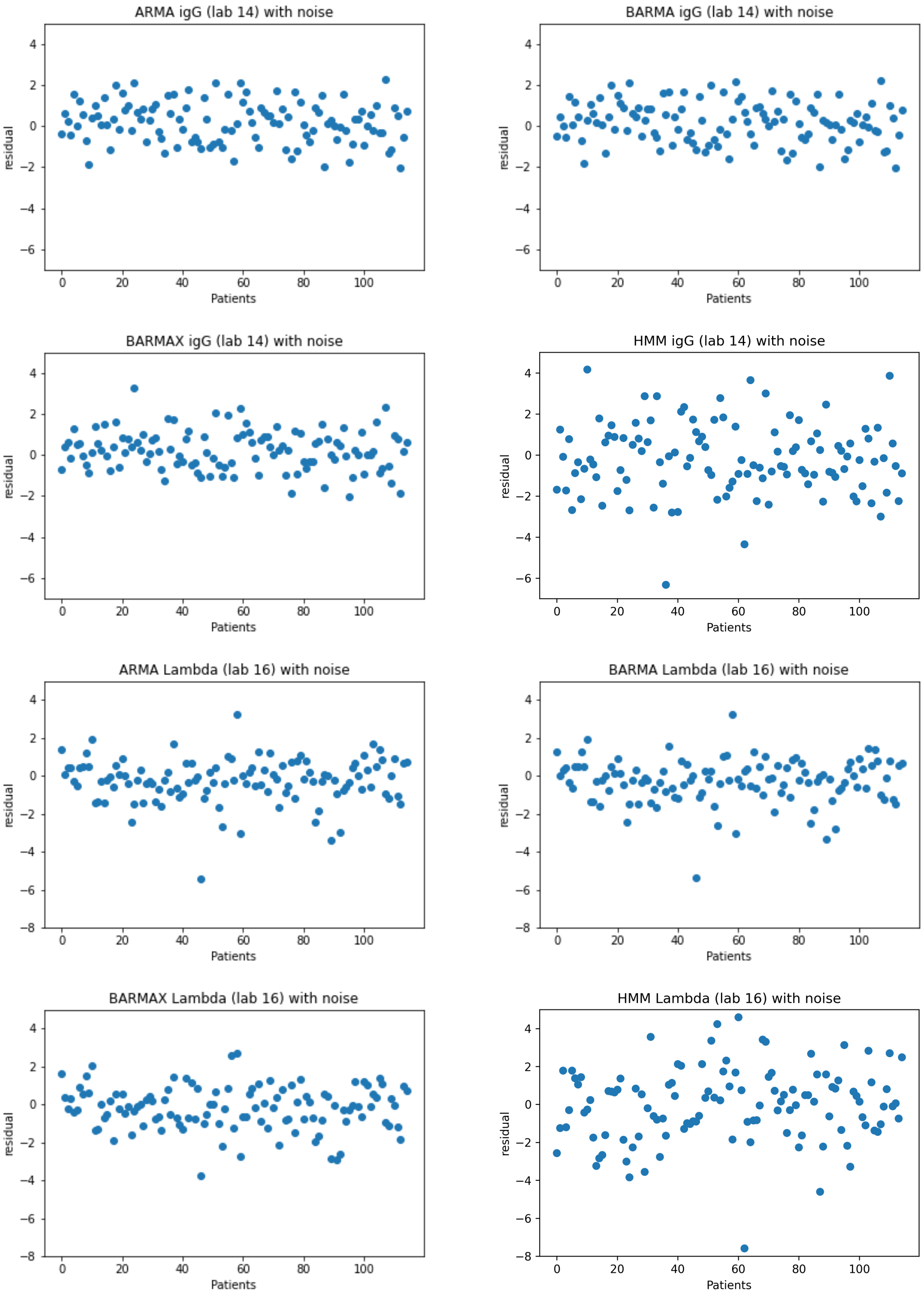


Figure 6: Residual plots for Kappa, IgG, IgA, and Lambda with noise

A.3 Number of hidden states in HMM

To choose the number of state in the HMM model, we have performed grid search in grid $[2, 4, 6, 8, 12, 24]$. The 5-fold cross validation log likelihood result for each number of hidden states is plotted in Figure 7. As the number of hidden states increases, the increase of log likelihood slows down and it reaches a plateau after number of state is 12. Therefore, we have decided to use 12 as the number of hidden state in our HMM.

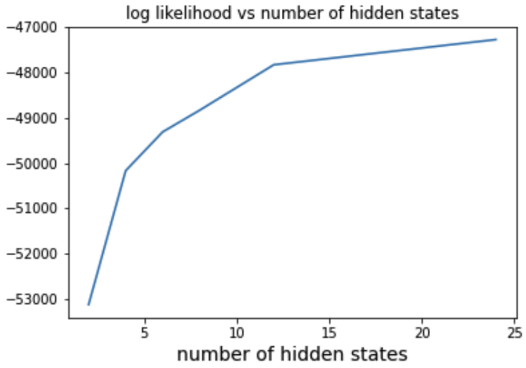


Figure 7: Log-likelihood vs. Number of hidden states in HMM

A.4 Hyperparameters and parameters trained

Table 3: Total number of parameters and hyperparameters trained in each model

Model	hyperparameters	parameters
ARMA	2 (i.e. p & q)	p + q
BARMA	2 (i.e. p & q)	p + q + 8 + 62
BARMAX	2 (i.e. p & q)	p + q + p×8 + 62
HMM	1 (i.e. number of hidden states)	540