High Frequency Inflation Forecasting

Sonan Memon*

This version: August 2022

Abstract

I begin by motivating the utility of high frequency inflation forecasting. I review recent work done at the State Bank of Pakistan for inflation forecasting and now-casting large scale manufacturing growth using machine learning. I also present stylized facts about the structure of historical and especially recent inflation trends in Pakistan and what they imply about forecasting. However, since the available data *and* already tested methods cannot achieve high frequency forecasting, I discuss 3 cutting edge techniques from recent literature including *web scrapping*, *scanner data* and *synthetic data*. Due to lack of access to scanner data and web scrapped data, I generate synthetic data using *generative* machine learning models (Gaussian Copula and PAR models) and *cubic spline interpolation* (numerical analysis) to estimate high frequency inflation (e.g monthly, weekly and daily) and forecasting future short-run (daily, weekly, monthly and quarterly) inflation for Pakistan. I evaluate the accuracy of forecasts using forecast error variance decomposition and VAR's (vector autoregressive models).

Keywords: High Frequency Inflation Forecasting. Web Scrapping. Scanner Data. Synthetic Data. Machine Learning. Hyperinflation. Forecasts of Inflation in Pakistan.
JEL Classification: E30, E31, E32, E37, E47, E52, E58, C53.

*Research Economist, PIDE, Islamabad. smemon@pide.org.pk

CONTENTS

1. MOTIVATION

Accurate forecasting of inflation is a concern for market players, central banks and governments. The market participants want to update their inflation expectations in line with new information revelation so that their investment strategies are optimal. Meanwhile, central banks typically have mandates for price stability and they routinely collect data on inflation expectations and forecasts Cukierman et al. (1992). Of course, hyperinflation dramatically hurts hand to mouth households and this extreme economic turmoil has political consequences for governments, especially when the election period is nearby (see for instance Binder (2021)). Thus, governments have an incentive to make inflation control a priority and interfere with central bank independence, a few months before elections. At the most fundamental level, hyperinflation episodes are humanitarian and social crises which can be addressed to some extent if we develop better inflation forecasting methods, independent central banks and policy interventions.

In fact, there is a well known classic literature on the so called political business cycle, initiated by Nordhaus (1975). For instance, Abrams and Butkiewicz (2012) show that recordings reveal that President *Nixon* of USA manipulated Arthur Burns[1] and the Federal Reserve into creating a political business cycle which helped ensure his reelection victory in 1972. While President Nixon understood the risks that his monetary policy imposed but chose to trade longer-term economic costs to the economy for his own short-term political profits.

While central banks collect data on consumer price indices, the frequency of such data does not allow accounting for sudden swings in inflation *and* inflation expectations. Some examples of standard measures include the HICP (Harmonized Consumer Price Index) data used in the Euro area and the CPI (consumer price index) data from USA. Such data typically tends to be quarterly in worst cases or in best cases monthly, but results are revealed in the next month after collection. However, when for instance, in a matter of few days and weeks, news about the Ukraine and Russian crisis changed the inflation expectations of many products, conventional price indices had little forecasting potential for the following inflation crisis. Similarly, inflation shocks can result from sudden change of central bank's governors or government change, terrorism episodes or political turmoil, especially in developing economies, where inflation tends to more volatile and central banks are less independent (see Vuletin and Zhu (2011)).

[1] Head of the Federal Reserve Bank

2. RESEARCH AT SBP

The State Bank of Pakistan (SBP) has also done some work on inflation forecasting by using machine learning methods (e.g Neural Networks)[2] and monthly year on year (YoY) inflation rates of Pakistan from Jan 1958 to Dec 2017 Hanif et al. (2018). The *Thick ANN* (Artificial Neural Networks) model developed in this paper is found to outperform all the 15 econometric models of Pakistan economy previously developed in forecasting 24 months ahead headline inflation.

Similarly, the SBP has worked on *nowcasting* GDP using large scale manufacturing growth (LSM) in Pakistan Hussain et al. (2018) and LASSO type[3] ML methods. The models are used to extract the unique *information* from a range of variables having close association with LSM in Pakistan. The results displayed in Figure 1 from Hussain et al. (2018) below reveal that the predicted LSM series closely tracks the actual LSM series. Since LSM is available at relatively higher frequency (monthly) relative to the actual GDP (annual), it is a predictor for determinants of economic activity such as key sectors, prices, credit, interest rates and tax collection, external trade and inflows. This is in line with emerging methodologies among central banks worldwide, which are all moving toward big data and machine learning methods (see Doerr et al. (2021)).

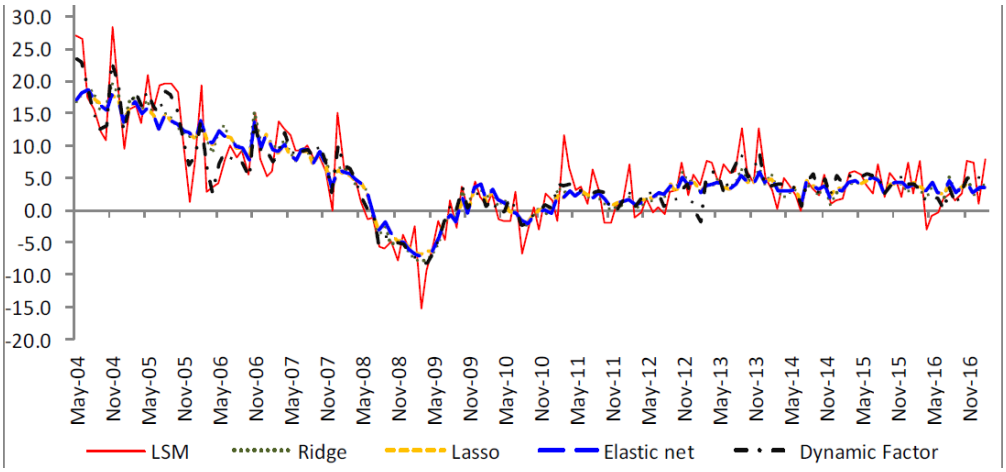

Figure 1: Nowcasting LSM For Pakistan (Source is Hussain et al. (2018))

However, lack of availability of high frequency data on the order of days or weeks poses a limitation in forecasting inflation. Hence, I argue that we need more granular

[2]For a review of machine learning methods see

[3]Least Absolute Shrinkage Operator, Ridge Regressions and Elastic Nets.

data, on the order of days or weeks for enhancing forecasting. Next, I discuss methods for collecting such high frequency data, used at the current frontier of research on inflation.

3. Review of Modern Forecasting

I will briefly review three methods including *web scrapping* online inflation data, using *scanner data* from supermarkets and *synthetic* data for high frequency inflation forecasting. Later, I use synthetic data for a forecasting exercise in this paper due to lack of availability of other data types.

3.1. Web Scrapping

In recent literature, the daily consumer price index (CPI) produced by the Billion Prices Project (BPP CPI) of Cavallo and Rigobon (2016) offers a glimpse of the direction taken by consumer price inflation in *real time*. For instance, Figure 2 is based on web scrapping online inflation data for Argentina Cavallo and Rigobon (2016). It shows that the official CPI significantly under-stated actual inflation, when measured by web scrapping. An added benefit of such data is that it reveals the partisan measurement and particularly disclosure of CPI data in developing economies such as Argentina, where central bank independence is low.

Should we expect a similar lack of correspondence between official inflation data of the SBP (State Bank of Pakistan) and non-partisan research measures? Not much is known about the political business cycle in Pakistan and I believe that independent research, not originating from SBP is needed to address the question. Given the low levels of central bank independence in Pakistan and existing literature Vuletin and Zhu (2011), we should expect that higher levels of price stability can be achieved if governor appointments and turnovers are not manipulated by political powers.

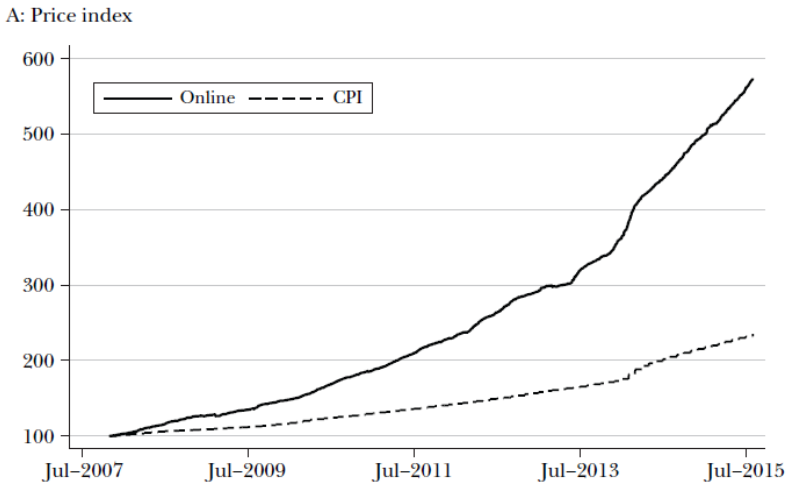

Figure 2: Inflation in Argentina (Source is Cavallo and Rigobon (2016))

With increasing scope of online transactions in Pakistan, *web scrapping* can also be informative, despite absence of Amazon type large scale online transaction services in Pakistan.

3.2. Scanner Data

Meanwhile, another branch of emerging literature uses scanner-based data (see for instance Beck et al. (2020)) on prices rather than web scrapping. In Figure 3 below, recent scanner-based price indices for Germany are disclosed from the work of Beck et al. (2022). The data compares trajectories in 2022 (red and orange solid lines below) with their historical averages from 2019 to 2021 (blue and purple solid lines) along with historic minimum and maximum values (shaded areas). The data indicates a very strong increase in prices for sunflower oil and flour in light of the Ukraine conflict, accompanied by temporarily higher sales. The price increase of sunflower oil was rather gradual and already started as of early February. In contrast, prices for flour increased very sharply, but only more than two months after the invasion. However, in both cases, sales went far beyond their average levels, suggesting increased demand and possibly stockpiling behavior from pessimistic consumers (see Cavallo and Kryvtsov (2021)). Concerning the more recent period up to June 2022, prices for both products seemed to have stabilized at a very high level, whereas quantities have converged back to their average levels.

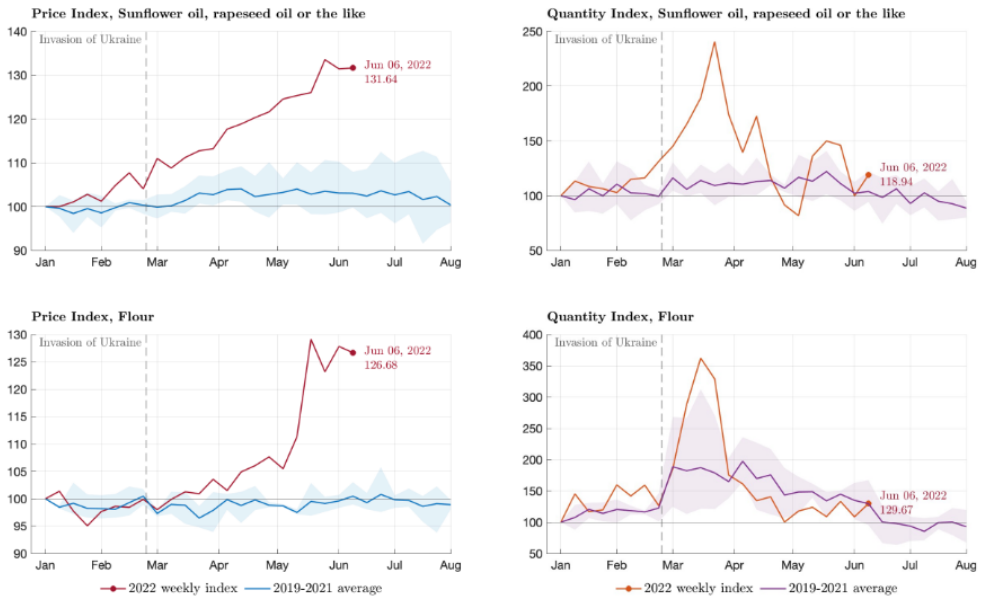

Figure 3: Source is Beck et al. (2022)

I propose that high frequency *scanner data* from super markets in Pakistan can improve high frequency inflation forecasting. In Karachi, Carrefour, Metro, Chase, Chase Up, Imtiaz supermarket and Bin Hashim are some major super markets. Similarly, Al-Fatah, Carrefour and Imtiaz supermarket are some major super market players in Lahore. However, lack of availability for super market scanner data is a constraint which must be overcome.

3.3. Synthetic Data

Synthetic data is artificial data (see Nikolenko (2021)) which is generated to mimic key information of the actual data and provide the ability to draw valid statistical inferences. It allows widespread access to data for analysis while overcoming privacy, confidentiality and cost of data collection concerns (also see Raghunathan (2021)). For instance, Patki et al. (2016) develop the SDV (Synthetic Data Vault) which uses multivariate *Gaussian Copula* (see Chapter 5.1.5 of Stachurski (2016)) to calculate covariances across input columns. The distributions and covariances are sampled from the copula to form synthetic data and as proof of pudding, relational data sets were synthetically generated and used by freelance data scientists to develop predictive models. The researchers found no significant difference between the results produced using the synthetic versus true data Patki et al. (2016).

For review of other generative models for synthetic data and advanced methods such as generative adversarial networks for economists, refer to Koenecke and Varian (2020).

Synthetic data is particularly useful for me since high frequency inflation data is not available for Pakistan. It is also essential to state that even at a private level, State Bank of Pakistan does not have high frequency data, so when I use the term *synthetic*, I mean artificially constructed high frequency data from the available low frequency data. This is in contrast with synthetic methods which are solving an information revelation problem by generating synthetic data from actual data of same frequency (see Patki et al. (2016)). The accuracy of forecasting using *synthetic* data is certainly questionable and using high frequency data through scanner data and web scrapping is part of my long term research agenda. This synthetic data exercise can motivate policy makers and State Bank of Pakistan to initiate collection of high frequency data by providing a glimpse of the utility of high frequency data.

4. Stylized Facts On Pakistan's Inflation

In Figure 4 below, I have the quarterly inflation series for Pakistan from the first quarter of 1970 to the first quarter of 2021. During 1980 to 2008, average annualized, quarterly inflation was below 10%. Pakistan had a severe hyperinflation crisis during the 1970's and other major inflation periods were during 2007-2009 (during the great recession) and around 2020, after the COVID shock. In 2022, Pakistan is again in the middle of an inflation crisis, partially due to the oil price shock and due to fiscal imbalances, reflected by the IMF package.

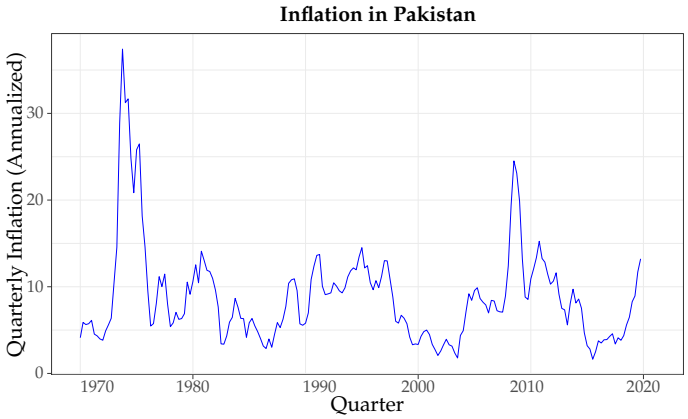

Figure 4: Data is From State Bank of Pakistan

The latest available consumer price index (CPI Inflation with Base Year of $2015-16=100$) data from State Bank of Pakistan recorded inflation at 8 percent on year-on-year basis for Pakistan in December 2020 (see Table 1 below) and 12.3% in December 2021. Moreover, core measure of inflation which excludes food and energy (Non-food, Non-Energy (NFNE)) inflation was recorded at 6.4 % in December 2020 and 8.5% in December 2021. Meanwhile, across products, food (12.9 %) had among the highest inflation rates and inflation in education sector was among the lowest at 1.3% in December 2020 data. In the case of transport prices, we actually observed deflation of -3.5%. in December 2020, which reflects the demand shock due to lower mobility and COVID crisis. On the other hand, particularly transport, followed by clothing had among the highest inflation rates in December 2021, which reflects a massive transformation in the transport sector within one year. Education remained among the sectors with lowest inflation rates in December 2021 data.

Table 1: Annual National Inflation (December 2017 to December 2021)

Year on Year (%)	**Dec 2017**	**Dec 2018**	**Dec 2019**	**Dec 2020**	**Dec 2021**
Categories					
Headline Inflation	4.6	5.4	12.6	8	12.3
Food Inflation	3.8	0.6	17.9	12.9	10.6
Core Measure (NFNE)	5.5	7.64	7.7	6.4	8.5
Clothing	3.6	6.3	9.8	9.7	11.2
Health	10.9	7.1	11.3	8.1	9.4
Transport	4.5	18.4	14.7	-3.5	24.1
Education	12.4	9.8	6	1.3	2.8

Note: Data is from State Bank of Pakistan. Base year is 2015-2016 for all columns, apart from Dec 2017 column for which it is 2007-2008.

In Figure 5, I have plotted the most recent trends in monthly and year on year inflation (headline inflation rate) for Pakistan. The data is from State Bank of Pakistan and covers the months from January 2020 to June 2022. The graph indicates that during 2020, year on year inflation was actually falling despite the COVID shock. In fact, monthly and year on year inflation was close to 5% in the beginning of 2021. However, after 2021 and especially after the debt crisis of 2022, inflation rates have sky rocketed to more than 20%.

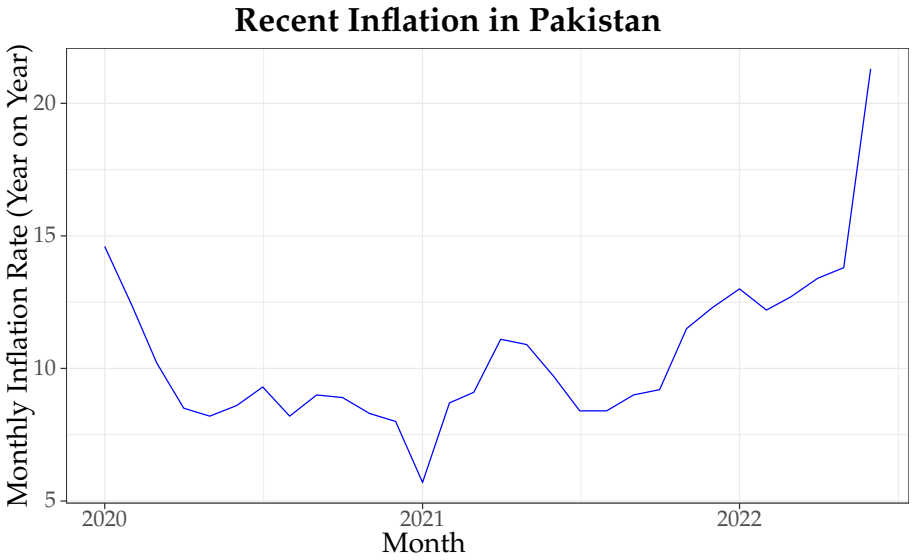

Figure 5: Data is From State Bank of Pakistan

5. METHODOLOGY

In this paper, I will generate higher order synthetic series using quarterly inflation data from State Bank of Pakistan (SBP). I use the multivariate *gaussian copula* (see Patki et al. (2016)), *probability autoregressive models* and *cubic spline interpolation* (numerical analysis) to estimate higher order inflation series for forecasting.

Lastly, I use these high frequency series, along with VAR models to forecast future inflation rates for Pakistan.

5.1. SYNTHETIC DATA FROM COPULA

A copula C in space $\mathbb{R}^n$ is a multivariate CDF (cumulative density function) supported by the unit hyperplane $[0,1]^N$ with the property that all of its marginals are uniformly distributed on $[0,1]$ (see Stachurski (2016)). Formally, C is the function of the form below, where $0 \leq s_n \leq 1$ and $u_n \sim U[0,1], \forall n$.

$$C(s_1, s_2, s_3, ..., s_N) = \mathbb{P}\{u_1 \leq s_1, ..., u_N \leq s_N\} \tag{1}$$

While each u_n has its marginal distribution pinned down, there can be infinitely many ways to specify the joint distribution. For instance, the independence copula, gumbel

copulas and clayton copulas are some different types of joint distributions Stachurski (2016). Figure 6 represents the general structure of a generative model which uses Guassian Copula so that $F(s_1, s_2, ..., s_N) = C(F_1(s_1), ...F_N(s_N))$ and $F_1, F_2, ..., F_N$ are univariate normal distributions.

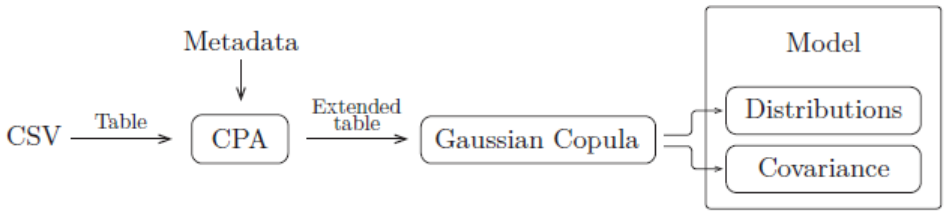

Figure 6: Source is Patki et al. (2016)

In Figure 7 below, I use Gaussian Copula to generate synthetic time series for inflation in Pakistan for 205 quarters from quarter 1 which is 1970Q1 and ending at 2021Q1. A quick comparison with previous Figure 4 can reveal that the series roughly estimates the actual inflation series. The first sub-figure of Figure 6 has a simulation size of 100 and it reveals a particular, rare or unlikely draw which has low volatility and an average inflation rate of 10%. The second graph in Figure 6 below has 500 simulations and outliers are less explicitly visible. The mean, annualized, quarterly inflation in the simulation is close to 10% but with a high degree of volatility. The probability of drawing an inflation rate of less than 5% or above 15% is very low.

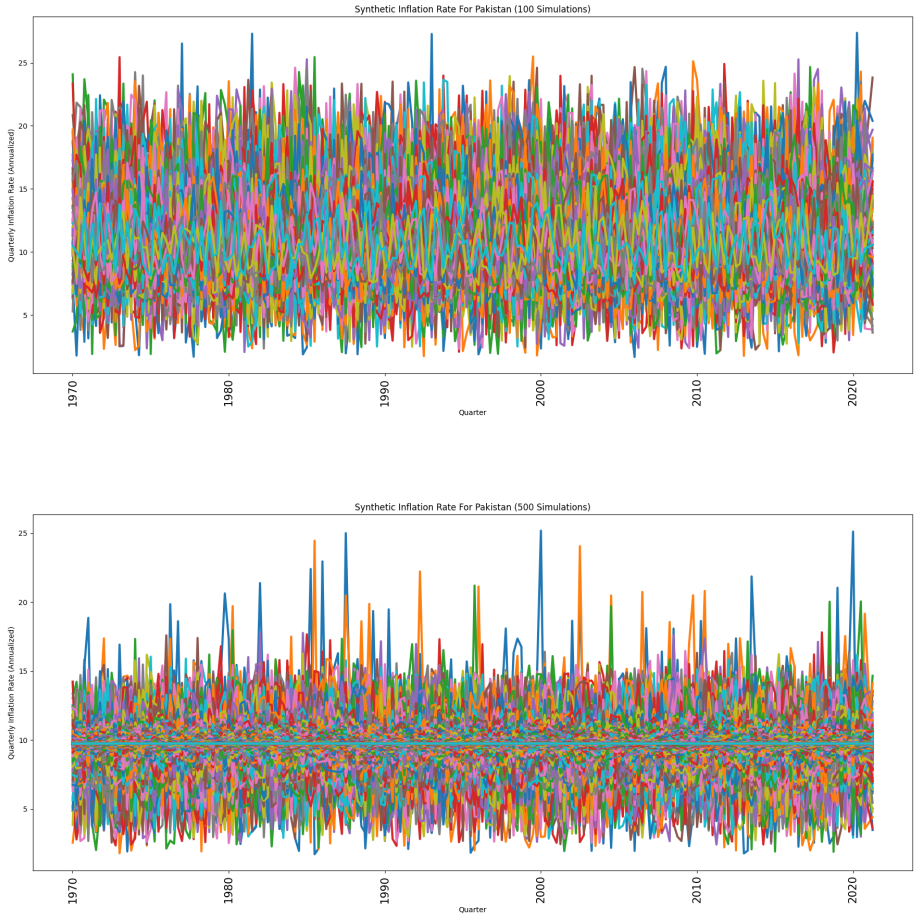

Figure 7: Author's Simulations

5.2. SYNTHETIC DATA FROM PAR MODEL

Probability Autoregessive Model (PAR) is a synthetic data creation methodology which is well suited for time series models and accounts for the autocorrelation structure of time series data. The PAR class allows learning multiple types, multivariate time series data and generating new synthetic data that has the same format and properties as the learned one. Salinas et al. (2020) have done path-breaking work at the frontier by developing probabilistic forecasting models with autoregressive, recurrent *neural networks*.

In Figure 8 below, I reveal my results from applying the PAR model which includes only the inflation time series from 1970Q1 to 2019Q4 for Pakistan. A combination of 100 simulations from the fitted PAR model reveals that this model has roughly the same mean

inflation of around 10% as the output from gaussian copula. However, the volatility of simulated outputs has actually increased since occasionally the PAR model draw values above 30 and even 40% inflation. Even with 500 simulations, the gaussian copula based simulations rarely cross even a 25% inflation level.

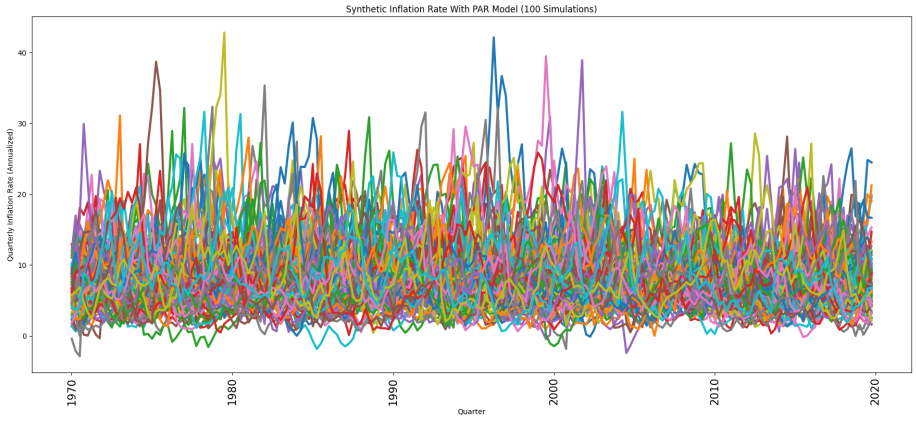

Figure 8: Author's Simulations

5.3. CUBIC SPLINE INTERPOLATION

Cubic spline interpolation is of the common interpolation methods. It uses cubic polynomials to connect the nodes. For a mathematical review of cubic spline interpolation, you can refer to Burden et al. (2015) and the appendix below.

$$(x_0, y_0), (x_1, y_1), ..., (x_n, y_n) \tag{2}$$

where $x_0 < x_1 < ... < x_n$. In equation 3 below, the cubic polynomial's interpolating pairs of data are labeled as $S_0, ..., S_{n-1}$. The polynomial S_i interpolates the nodes (x_i, y_i) and (x_{i+1}, y_{i+1}) in equation 2 above. Let:

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3, \forall i = 0, 1, 2, ..., n-1 \tag{3}$$

In Figure 8 below, I have an example of interpolating the so called Runge function which is $y = \frac{1}{1+x^2}$, in the data range of $[-5, 5]$ by the Natural Cubic Spline (with natural boundary condition) and the Newton Interpolation methods (green line). The graph demonstrates that the natural cubic spline estimates the original function with high pre-

cision. However, the newton interpolation method is noisy and has particularly extreme errors toward the two tale ends of the distribution.

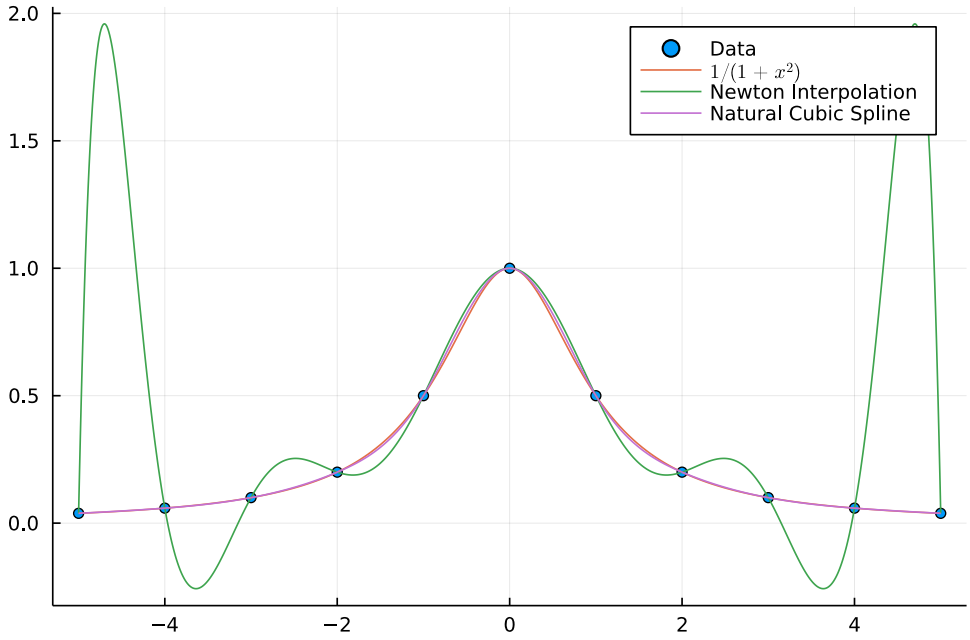

Figure 9: Example is From Ökten (2019)

Based on quarterly inflation series for Pakistan, I carry out a cubic spline interpolation exercise, displayed in Figure 8 below. This gives me a continuous series of inflation data in the $[0, 200]$ range which refers to the data during 1970 to 2020 for Pakistan. Despite having access to only quarterly data (1970Q1 to 2019Q4) visible in the dots of Figure 9, the interpolation allows me access to a higher order approximation for weekly and even daily inflation rates in this period.

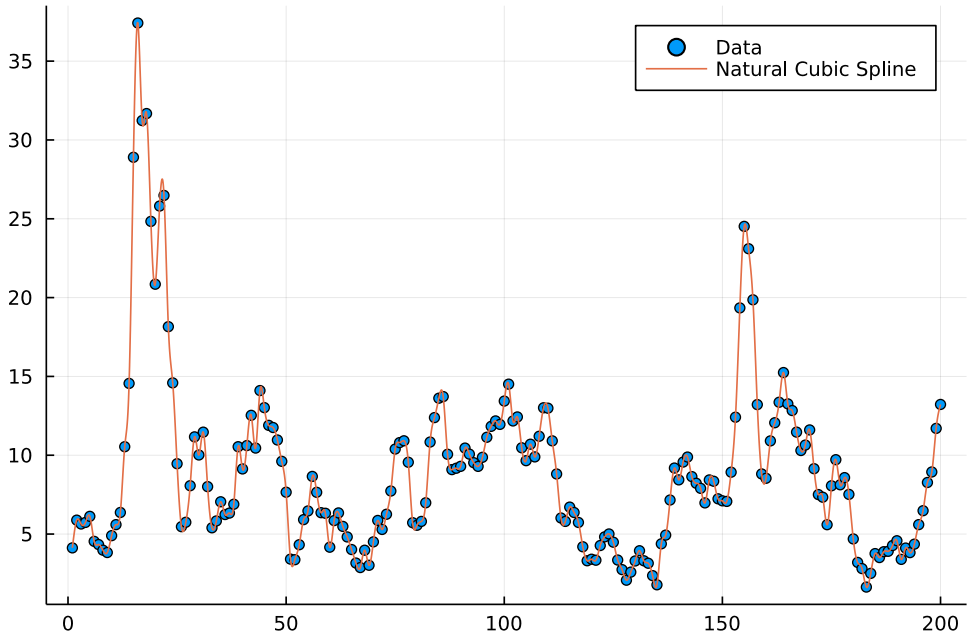

Figure 10: Author's Calculations Using Cubic Splines

6. Conclusion

In this working paper, I review cutting edge methodologies for inflation forecasting, while being motivated by the current inflation crisis of Pakistan. Access to high frequency scanner data, web scrapping and synthetic data can make inflation forecasting and measurement more efficient. In this paper, I use synthetic data and numerical techniques to estimate the unknown high frequency inflation series for Pakistan in the period from 1970 to 2021 due to the lack of access to other data sources.

My empirical analysis uses *gaussian copula*, *probability autoregressive models* and *cubic spline* interpolation methods for high frequency inflation estimation and forecasting. I find that we can approximate monthly and other low order (weekly or daily) inflation series for Pakistan using my methodology. The forecast accuracy can be estimated by using only quarterly inflation series to generate forecast for monthly series before comparing the forecasts with actual monthly series.

APPENDIX

CUBIC SPLINE INTERPOLATION

Cubic spline interpolation is the most common spline interpolation method. It uses cubic polynomials to connect the nodes. Consider the data in equation 4:

$$(x_0, y_0), (x_1, y_1), ..., (x_n, y_n) \tag{4}$$

where $x_0 < x_1 < ... < x_n$. In equation 5 below, the cubic polynomial's interpolating pairs of data are labeled as $S_0, ..., S_{n-1}$. The polynomial S_i interpolates the nodes (x_i, y_i) and (x_{i+1}, y_{i+1}) in equation 4 above.

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3, \forall i = 0, 1, 2, ..., n-1 \tag{5}$$

Under the above formulation, there are $4n$ unknowns to be determined and the following four set of equations (6 to 8) must be satisfied by the interpolating function. The first set of two conditions (sets 6 and 7) are merely consistency conditions with the peer of data inputs. The last two equations are referred to as smoothness or boundary conditions. We have to choose boundary conditions with two possible choices, a free or natural boundary (equation set 8) or a clamped boundary (equation set 9). In sum, there are $4n$ equations since the first set 6 of equations give us $2n$ equations, the second set of conditions gives us $2n-2$ (set 7) and last set (8 or 9) gives us 2 equations ($2n + 2n - 2 + 2 = 4n$) regardless of which we choose. It turns out that this systems of equations has one unique solution (for a proof see Burden et al. (2015)).

$$S_i(x_i) = y_i, S_i(x_{i+1}) = y_{i+1} \tag{6}$$

$$S_{i-1}^{'}(x_i) = S_i^{'}(x_i), S_{i-1}^{''}(x_i) = S_i^{''}(x_i) \tag{7}$$

Free or Natural Boundary:

$$S_0^{''}(x_0) = S_{N-1}^{''}(x_n) = 0 \tag{8}$$

Clamped Boundary:

$$S_0^{'}(x_0) = f'(x_0), S_{N-1}^{'}(x_n) = f'(x_n) \tag{9}$$

REFERENCES

Abrams, Burton A and James L Butkiewicz, "The Political Business Cycle: New Evidence From The Nixon Tapes," *Journal of Money, Credit and Banking*, 2012, *44* (2-3), 385–399.

Beck, Guenter W, Hans-Helmut Kotz, and Natalia Zabelina, "Price Gaps at the Border: Evidence from Multi-Country Household Scanner Data," *Journal of International Economics*, 2020, *127*, 103368.

Beck, Guenter W., Kai Carstensen, and Jan-Oliver Menz, "Real-time Food Price Inflation in Germany in Light of the Russian Invasion of Ukraine," *VOXEU*, 2022.

Binder, Carola Conces, "Political Pressure on Central Banks," *Journal of Money, Credit and Banking*, 2021, *53* (4), 715–744.

Burden, Richard L, J Douglas Faires, and Annette M Burden, *Numerical Analysis*, Cengage Learning, 2015.

Cavallo, Alberto and Oleksiy Kryvtsov, "What can Stockouts tell us about Inflation? Evidence from Online Micro Data," Technical Report, National Bureau of Economic Research 2021.

_ **and Roberto Rigobon**, "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 2016, *30* (2), 151–78.

Cukierman, Alex, Steven B Web, and Bilin Neyapti, "Measuring the Independence of Central Banks and its Effect on Policy Outcomes," *World Bank Economic Review*, 1992, *6* (3), 353–398.

Doerr, Sebastian, Leonardo Gambacorta, José María Serena Garralda et al., "Big Data and Machine Learning in Central Banking," *BIS Working Papers*, 2021, (930).

Hanif, Muhammad Nadim, Khurrum S Mughal, and Javed Iqbal, "A Thick ANN Model for Forecasting Inflation," Technical Report, State Bank of Pakistan, Research Department 2018.

Hussain, Fida, Kalim Hyder, and Muhammad Rehman, "Nowcasting LSM growth in Pakistan," *State Bank of Pakistan*, 2018, p. 1.

Koenecke, Allison and Hal Varian, "Synthetic Data Generation for Economists," *arXiv. arXiv:2011.01374*, 2020.

Nikolenko, Sergey I, *Synthetic Data For Deep Learning*, Vol. 174, Springer, 2021.

Nordhaus, William D, "The Political Business Cycle," *Review of Economic Studies*, 1975, *42* (2), 169–190.

Ökten, Giray, "First Semester in Numerical Analysis with Julia," 2019.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni, "The Synthetic Data Vault," in "2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)" IEEE 2016, pp. 399–410.

Raghunathan, Trivellore E, "Synthetic Data," *Annual Review of Statistics and Its Application*, 2021, *8*, 129–140.

Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski, "DeepAR: Probabilistic Forecasting With Autoregressive Recurrent Networks," *International Journal of Forecasting*, 2020, *36* (3), 1181–1191.

Stachurski, John, *A Primer in Econometric Theory*, MIT Press, 2016.

Vuletin, Guillermo and Ling Zhu, "Replacing a "Disobedient" Central Bank Governor with a "Docile" one: A Novel Measure of Central Bank Independence and its Effect on Inflation," *Journal of Money, Credit and Banking*, 2011, *43* (6), 1185–1215.