Linear Algebra with Applications Final Project

Michael Sarullo; Leo Zhang; Miya Zhao

December 2024

1 Introduction

In 2020, the COVID-19 pandemic irreversibly changed global health, economies, and daily life. Four years later, shifts from rapid response to immediate outbreaks to dealing with the virus's long-term effects led to new important questions: Will COVID-19 infection rates settle down over time, and what factors will influence their long-term behavior? Additionally, can we use sporadic updates to predict daily new cases in order to more accurately track the progression of cumulative cases of COVID through a population? Answering these questions is crucial for future public health strategies and the ever growing impact of airborne infections. This project uses linear algebra methods, specifically Markov chains with steady-state analysis and least-square approximation, to study and predict how COVID-19 infection rates might change over given time periods. Markov chains model the chances of moving between different stages of infection, allowing us to find steady states of infection rates and make important assumptions. Additionally, these matrices allow for individualized probability predictions after a certain time frame (transitions). On the other hand, least-square approximation provides a way to determine COVID-19 cases more accurately in broader data sets, which is useful when COVID-19 data is tracked at larger time intervals. By combining these mathematical tools, this study aims to determine the future of COVID-19 in our population.

2 COVID-19 Steady State Modeling

2.1 Background and Methods

To begin modeling the progression of infection rates, we can first analyze the progression of the beginning of the COVID-19 epidemic when no vaccinations were available. Based on data from Worldometer and the U.S. Census Bureau, the following Susceptible-Infected (SI) table is constructed. For the purposes of this report, natural births and deaths were considered stagnant. The probability of an individual being infected(S to I) was calculated by adding up the number of infected individuals over a 10 day period (the average duration of contagiousness), specifically between the dates December 1, 2020 to December 10, 2020 (right before the vaccine was released), in order to gauge the total number of contagious positive cases on the given day. This was then divided by the total population to give a probability of a person being infected, 1,799,585 positive cases/331,826,033 individuals, or about a 0.5% chance of infection. Due to the model being stochastic, the probability of one not being infected became 1-0.005, or 99.4%, as all who did not become infected stayed susceptible. The probability of recovery was obtained from the number of recovered cases(those who did not die) divided by the number of total cases. 2028 average daily deaths/179,959 daily cases insinuates a death rate of around 1.1% and a recovery rate of 98.9%. This manifested in the stochastic matrix (A) below.

Table 1: SI Model based on 12/1/2020 - 12/10/2020 COVID-19 Infection Data.

	S	I
S	0.9946	0.9887
I	0.0054	0.0113

Theorem 2.9.2

Let P be the transition matrix of a Markov chain and assume that P is regular. Then there is a unique column matrix $\mathbf{s}$ satisfying the following conditions:

1. $P\mathbf{s} = \mathbf{s}$.
2. The entries of $\mathbf{s}$ are positive and sum to 1.

Moreover, condition 1 can be written as

$$(I - P)\mathbf{s} = \mathbf{0}$$

and so gives a homogeneous system of linear equations for $\mathbf{s}$. Finally, the sequence of state vectors $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots$ converges to $\mathbf{s}$ in the sense that if m is large enough, each entry of $\mathbf{s}_m$ is closely approximated by the corresponding entry of $\mathbf{s}$.

2.2 Results and Discussion

This models a Markov Chain as our assumptions allow the next states to only depend on the current probabilities for December 10, 2020, and not rely on any earlier history. Using Markov chains, important information can be deduced about the probability and likelihood of ending in certain states. For example, should a patient be susceptible on day one, what is the probability they are infected on day 10? By raising the matrix A to the 10th power, then reading the (2,1)th entry, we ascertain this probability.

Table 2: SI Model after 10 Days

	S	I
S	0.99456795	0.99456795
I	0.00543205	0.00543205

Computation was performed with Wolfram Mathematica. From this, we can predict that the probability of a person starting as susceptible, by the 10th day, will have a 0.543% chance of being infected, a higher probability than we began on day one. Additionally, this transition matrix to the 10th power allows us to confirm this is a regular matrix (all positive values). Due to it being a regular stochastic matrix, we know by theorem that there must then exist a steady state vector.

A steady-state matrix can be determined by solving $(I - P)s = 0$, where P represents the transition matrix, I represents an identity matrix identical in size to P, and s represents the steady-state matrix.

$$(I - P)s = 0$$

$$\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}0.9946 & 0.9887\\ 0.0054 & 0.0113\end{bmatrix}\right)s = 0$$

$$\begin{bmatrix}0.00543205 & -0.994568\\ -0.00543205 & 0.994568\end{bmatrix}\begin{bmatrix}s_1\\ s_2\end{bmatrix} = 0$$

$$\begin{cases}0.00543205s_1 - 0.994568s_2 = 0\\ -0.00543205s_1 + 0.994568s_2 = 0\end{cases}$$

The values of s_1 and s_2 that comprise the steady-state matrix that satisfy this system are

$$s = \begin{bmatrix}0.999985\\ 0.00546164\end{bmatrix}$$

Computation was performed with Wolfram Mathematica.

After normalization, S = 0.9946 and I = 0.0054. This implies that as time goes to infinity, a person's chance of not being infected is around 99.46% and the chance of being infected will be around 0.54%. While interesting to the regards of the individual, the implication on the population is that given this rate of infection, nearly the entire population remains susceptible, while only an extremely small fraction is infected at equilibrium. Such a minimal steady state infection rate suggests that the disease cannot sustain widespread transmission and is likely to diminish over time. Therefore, our model implies that COVID-19 would become a minimal issue, as the infection rate is too low to maintain significant prevalence within the population. This conclusion aligns with real-world observations, where effective vaccination efforts and public health measures have successfully kept infection rates minimal, preventing the disease from establishing a persistent presence. However, the disease will never be fully eradicated in a population our size as I is not 0. These inferences can be reflected on real-world data, where cases in 2021 are in the thousands, not the hundreds of thousands we observed in 2020.

However, there are limitations to our model, and these predictions are currently contested with endemic disease speculations in the epidemiological world. Firstly, we made many assumptions to the model not reflective of the disease such as assuming a constant population when calculating probabilities even though we suggest death. Notably, we used an SI model, even though an SIRS model would be more reflective of an individual. After they recover, they can become susceptible once again. Furthermore, this model does not account for evolving technologies and policies as not only is this data obtained pre-vaccination, but also pre-treatments strategies, information spread about masking, differing local masking policies, all of which influences the infection and recovery rates.

3 COVID-19 Cumulative Case Modeling

3.1 Background and Methods

The second method of analysis to determine the progression of COVID-19 is approximation via least-squares. Least-squares can be useful in analyzing COVID-19 cases when data is sparse. For example, a small town may only be

able to report the new cases once a week, unable to track the patients that contract the disease and recover in between. As COVID data tracker centers start to close, it becomes increasingly important to predict possible peaks in cases. One of these spikes occurred in January of 2022. By using weekly data points from the World Health Organization fit upon a polynomial fit, we were able to predict day to day cases to decent accuracy. By tracking per day using least-squares, more accurate data of cumulative cases can be determined as people are testing less regularly. To utilize least-square approximation, the following data set is constructed with a time interval of one week (Table 3).

Table 3: COVID-19 Cases Data Split into Two Parts

Date	**Day (x)**	**Cases (y)**
Dec 18	1	125,417
Dec 25	8	190,390
Jan 1	15	354,503
Jan 8	22	614,558
Jan 15	29	800,782

Date	**Day (x)**	**Cases (y)**
Jan 22	36	737,830
Jan 29	43	577,488
Feb 5	50	347,163
Feb 12	57	192,670
Feb 19	64	108,523
Feb 26	71	69,704

Due to not being able to assume any intercepts, we fit along a curve of best fit. An even polynomial to a degree larger than 2 seemed to be the best option.

Theorem 5.6.3

Let n data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be given, and write

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad M = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_m \end{bmatrix}.$$

1. If $\mathbf{z}$ is any solution to the normal equations

$$(M^T M)\mathbf{z} = M^T \mathbf{y},$$

then the polynomial

$$z_0 + z_1 x + z_2 x^2 + \cdots + z_m x^m$$

is a least squares approximating polynomial of degree m for the given data pairs.

2. If at least $m + 1$ of the numbers $x_1, x_2, \ldots, x_n$ are distinct (so $n \geq m + 1$), the matrix $M^T M$ is invertible, and $\mathbf{z}$ is uniquely determined by

$$\mathbf{z} = (M^T M)^{-1} M^T \mathbf{y}.$$

3.2 Results and Discussion

To find the best least-squares approximating polynomial, one follows the formula from Theorem 5.6.3, where

$$z = \begin{bmatrix} z_0 \\ z_1 \\ \cdots \\ z_m \end{bmatrix}; M = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 8 & 64 & 512 & 4096 & 32768 & 262144 \\ 1 & 15 & 225 & 3375 & 50625 & 759375 & 11390625 \\ 1 & 22 & 484 & 10648 & 234256 & 5157952 & 113471024 \\ 1 & 29 & 841 & 24389 & 707281 & 20517009 & 595493261 \\ 1 & 36 & 1296 & 46656 & 1679616 & 60466176 & 2176782336 \\ 1 & 43 & 1849 & 79507 & 3418801 & 147606643 & 6357031589 \\ 1 & 50 & 2500 & 125000 & 6250000 & 312500000 & 15625000000 \\ 1 & 57 & 3249 & 185193 & 10584241 & 603303729 & 34387612557 \\ 1 & 64 & 4096 & 262144 & 16777216 & 1073741824 & 68719476736 \\ 1 & 71 & 5041 & 357911 & 25411681 & 1804289361 & 128104788881 \end{bmatrix}; y = \begin{bmatrix} 125417 \\ 190390 \\ 354503 \\ 614558 \\ 800782 \\ 737830 \\ 577488 \\ 347163 \\ 192670 \\ 108523 \\ 69704 \end{bmatrix}$$

z represents the coefficients of the least squares approximating polynomial such that the least squares approximating polynomial is

$$z_0 + z_1 x + x_2 x^2 + \cdots + z_m x^m$$

M represents a matrix consisting of x values given in a dataset, where the i^{th} column and n^{th} row contains the value x_n^{i-1}. y represents a matrix of y values that correspond to the x values in the dataset. This is the matrix we attempt to approximate using least squares approximating polynomials.
Data points used can be found in Table 3. After creating least-squares approximating polynomials of degrees 2, 4, and 6, we obtain the following approximations. Let $f_j(x)$ represent the least squares approximating polynomial of the j^{th} degree.

$$f_2(x) = 51051.4 + 34212.6x - 508.538x^2$$

$$f_4(x) = 114647 - 7166.04x + 2872.01x^2 - 83.6738x^3 + 0.628446x^4$$

$$f_6(x) = 157679 - 27394.8x + 4072.29x^2 - 68.4019x^3 - 1.58043x^4 + 0.0454741x^5 - 0.00028162x^6$$

Computation was performed with Wolfram Mathematica.

Using this polynomial expression, we can predict the cumulative number of cases throughout the three months, solely using the number of cases reported once a week. Through customized formulas in Excel, the total number of cases were calculated and matched to the actual trend of the time with cumulative cases between December 2020 and February 2021 predicted using our model (2.69 million individuals) only using weekly statistics, being within a margin of error of 0.8% to the actual number of cumulative cases reported (2.76 million cases).

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Actual New Cases	125417	128444	131818	138119	149615	160740	173798	190390	202488	206523	222104	250438	277037	310813
Predicted New Cases	134287	118607	110181	108520	113112	123430	138932	159068	183283	211020	241726	274855	309871	346249
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
	354503	387434	414167	439821	502377	551377	584017	614559	668985	689812	703481	755356	765503	787373
	383481	421077	458571	495515	531492	566108	599002	629840	658322	684181	707183	727130	743857	757237
	29	30	31	32	33	34	35	36	37	38	39	40	41	42
	800782	807276	809735	800049	757613	765149	754375	737831	720413	700785	685958	692055	640190	606584
	767176	773616	776536	775946	771894	764456	753745	739899	723089	703511	681387	656961	630499	602284
	43	44	45	46	47	48	49	50	51	52	53	54	55	56
	577488	541323	526124	503734	441644	413534	378988	347164	312673	295900	287095	248357	234281	218993
	572616	541804	510170	478040	445742	413604	381947	351083	321310	292907	266131	241209	218336	197666
Total	57	58	59	60	61	62	63	64	65	66	67	68	69	70
28213912	192670	174587	169320	162366	143643	130808	117854	108523	101165	97295	94905	76922	77914	73335
27994836	179311	163331	149730	138449	129362	122265	116873	112813	109611	106694	103375	98847	92178	82296

Figure 1: Predicted versus Actual New cases per day between December 2021 to February 2022.

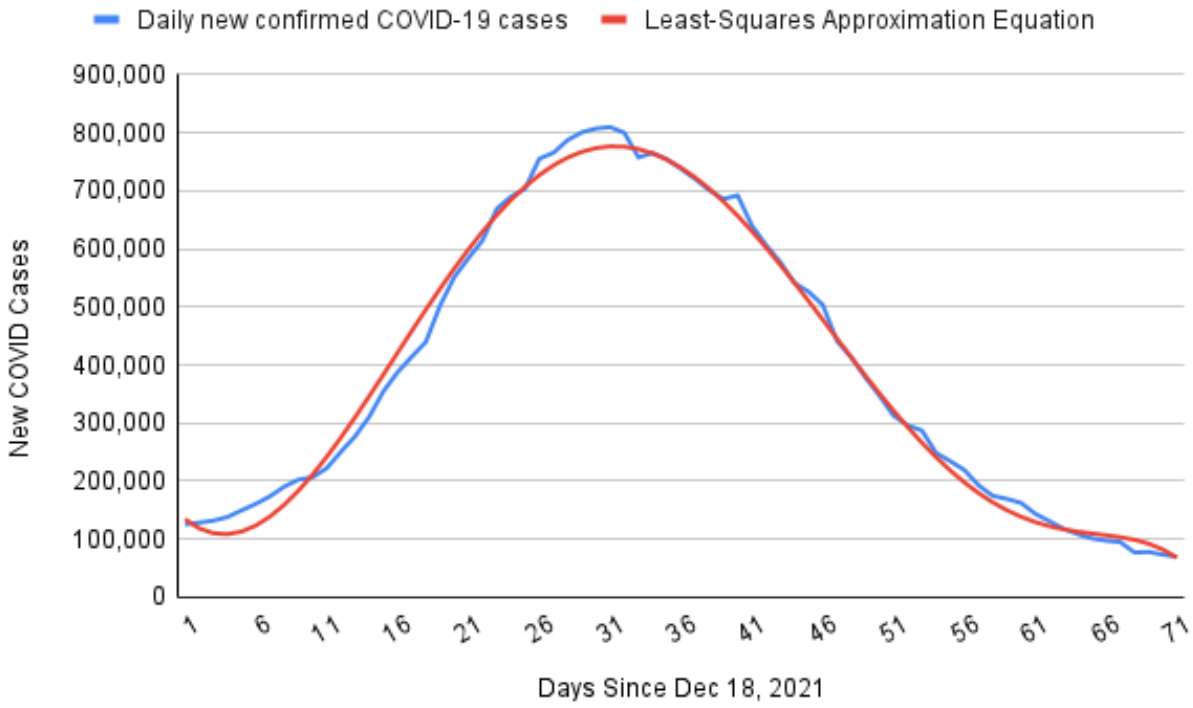

Figure 2: Weekly COVID-19 Deaths from 2020-2024

However, this data will still have variance based on socio-environmental factors. Least squares can only be used to analyze between a time range, thus giving retroactive analysis, not future purposes. Perhaps with machine learning, different magnitudes of this polynomial expression can be used to predict the length and peak of outbreaks.

4 Conclusion

In applying Markov chains and least-square approximations to predict COVID-19 infection rates, it was observed that the SI model was predicted to correctly result in a decrease in COVID-19 cases over time. Least-squares approximation also accurately predicted COVID-19 cases between a time interval of a dataset, highlighting its effectiveness in extracting information between data points. Nevertheless, there are significant limitations that affect the accuracy and reliability of our results. Real-world data and observed events demonstrate that COVID-19 exhibits cyclical patterns with multiple waves of infections driven by various factors such as seasonal changes, public health interventions, and the emergence of new variants. Ultimately, our models' predictions were less precise, highlighting its limitations in handling the complicated dynamics of the pandemic. These limitations underscore the need for more sophisticated modeling approaches that can better accommodate the variability and complexity inherent in COVID-19 infection rates. Overall, they give a good general and retroactive reflection on the dynamics of the pandemic as well as our predictions for the future of COVID-19 in our lives.