Scientific Computation Project 3

Amaury Francou — CID : 01258326

March 25, 2022

Part 1

1

We analyze the temperatures on the surface of Earth over one year. We are provided with a $365 \times 71 \times 144$ hypermatrix filled with temperatures values taken for several latitudes and longitudes each day of the year. We perform dimensionality reduction through principal component analysis of said data. We flatten and stack the geographic values into vectors composed of 10224 entries, which are further standardized. We obtain a corresponding matrix that is processed using the PCA method, diagonalizing the related covariance matrix.

We assess the relative importance of the components obtained through the process. This is done by reordering the eigenvalues and corresponding eigenvectors in an increasing order. We display said eigenvalues in figure 1. We here have an ideal situation in which the first 10 values are substantially higher than the remaining ones. For instance, noting λ_i the i-th greatest eigenvalue, we have that $\lambda_1 \approx 5900$ and $\lambda_{10} \approx 65$. Hence, the first reordered components play a major role in explaining the variance in the data. We compute the total variance V, the variance explained for each component $\frac{\lambda_j}{V}$ and the cumulative variance explained up to component j : $\sum_{k=1}^{j} \frac{\lambda_k}{V}$. We display said computations in figure 2. We observe that the first component accounts for more than 57% of the variance itself. We reach 80% of variance explained for 12 components computed in the PCA process. We further focus on said components.

We plot the top 12 components over time in figure 3. We observe that the variation of temperature compared to the annual average mainly follows a sin-trend. This observation is made more evident displaying only the top 2 components in figure 4. We investigate the temperature data projected on the top 2 components spanned subspace. This is shown in figure 5. Each sample point is labeled with its corresponding day. We observe a cyclicality in the temperature data : the coordinates of the 2 most important temperature descriptors matchs for day 1 and day 365. Moreover those descriptors stay close and vary softly for consecutive days. As seen in figure 6, temperature data projected on the top 3 components spanned subspace shows a hyperbolic paraboloid (pringles shape), which emphasise both cyclicality and seasonality. In particular, we identify 4 extremums : 2 maximums and 2 minimums.

The most important insights come from the reduced data in original variables (using first component), shown in figure 7. The deviation of the temperatures from the mean over the year follows a *stationary wave* pattern. We identify two extremums and two nodes for sin-trend waves in antiphase. The first extremum - which is a maximum for half the data and a minimum for the other half - occurs around day number 40. The second extremum occurs around day 220, which is roughly 6 months later. The nodes are located around day number 120 and day number 300, which also a 6-months difference. The antiphase waves might here describe the temperature data that has been measured in different hemispheres at the same time, the extremum of the wave trend being a maximum deviation from the year average temperature. This would for instance correspond to a given summer in the northern hemisphere and the corresponding winter in the southern hemisphere, or opposite. The nodes describe a minimum deviation from the year average temperature, which might refer to the spring and autumn seasons, depending on the hemisphere. This minimum deviation occurs simultaneously for both hemispheres, as they follow opposite temperature trends that coincide on these dates.

2

2.a

We analyze function `rec1`, which estimates missing or corrupted entries of an input matrix, using a low-rank factorization process. The function computes optimal matrices $A \in \mathcal{M}_{a\times p}(\mathbb{R})$ and $B \in \mathcal{M}_{p\times b}(\mathbb{R})$, such that the relevant entries of the product $\tilde{T} = AB$ give an estimation of the missing data. Matrices A and B are computed such that they minimize an error that is a combination of a modified Frobenius norm-based distance on known entries, plus an l_2-regularization term controlled by a hyperparameter λ.

Our objective is to assess how well the function performs estimation of the missing entries. For this, we are given a ground truth matrix $T^{(0)}$ with no missing data. We further tamper said matrix in order to test `rec1`. Namely, we randomly corrupt a given proportion of entries in T_0, which gives us an incomplete matrix T. We denote K' the set of indices corresponding to the corrupted data. In order to evaluate `rec1`'s performance, we define an error made on the estimations : we consider the average relative error made on all corrupted data. Precisely, our average relative error is defined as : $ARE = \frac{1}{|K'|}\sum_{i,j\in K'}\left|\frac{\tilde{T}_{ij}-T^{(0)}_{ij}}{T^{(0)}_{ij}}\right|$.

We first compute our error for matrices with different proportion of missing data. We expect having in average worse estimations for higher proportions of corrupted data, as less ground truth information is available to perform the predictions. In particular, for fixed values of p and λ, we have an ARE error of 54% for 5% of missing data, and on the opposite an ARE error of 200% for 75% of missing data. We further fix the proportion of missing data to 0.5% and work with the same corrupted matrix for the following assessments.

We secondly evaluate the effects of the hyperparameter λ. The l_2-regularization term penalizes large magnitude entries in matrices A and B. We have here that λ is a *tradeoff* parameter controlling such penalty. Namely, $\lambda = 0$ refers to the standard Frobenius norm-based distance cost, and as $\lambda \to +\infty$, the l_2-regularization becomes dominant in the optimization process, shrinking the matrices to nullity. We observe the effects of λ for $p = 6$ fixed on figure 8. In particular, we have that the ARE is first decreasing then increasing, displaying the existence of an optimal parameter, which is roughly $\lambda = 17$ in this case. Thus, reducing the magnitude of coefficients in the factor matrices first improves the estimations, perhaps by slightly shifting away said entries from fitting too much known coefficients, which would be made at the expense of corrupted data estimations. Furthermore, too large λ coefficients worsen the estimations, as the entries of A and B become too small to effectively produce the awaited estimations.

We finally analyze the evolution of the ARE error for varied p. This parameter gives us the maximum rank of the estimated matrix $\tilde{T} = AB$. The objective is to minimize the rank of said matrix, which corresponds to avoiding the insertion of new trends in the data. We observe the effects of p for $\lambda = 0$ fixed on figure 9. We have that the general trend of the ARE error for varying p is first decreasing then increasing, also displaying the existence of an optimal parameter, which is approximatively $p = 18$ here. Note that the original data $T^{(0)}$ has full rank (71). The optimal rank p corresponds to retrieving the optimal 'amount' of new trends needed to recover the full rank and the total variance given in the initial data set, compensating the information lost during corruption. Not enough or too much new patterns may worsen the approximations.

2.b

We here study a recommender systems that includes a user bias vector β. The cost function is further modified (see subject) and its gradient is to be recomputed.

Following the notations of the lecture notes, we first have that $\frac{\partial c}{\partial A_{kl}} = -2\sum\sum_{i,j\in K}(R_{ij} - \beta_i - \tilde{R}_{ij})\frac{\partial \tilde{R}_{ij}}{\partial A_{kl}} + 2\lambda A_{kl}$. Requiring $\frac{\partial c}{\partial A_{kl}} = 0$ and following the same calculation process as in the notes leads to the new update : $A_{kl}\sum_{j,(k,j)\in K} B^2_{lj} + \lambda A_{kl} = \sum_{j,(k,j)\in K}(R_{kj} - \beta_k - \sum_{s\neq l} A_{kj}B_{sj})B_{lj}$.

The update of matrix B's coefficients is symmetric. We also compute the derivative of the cost with respect to β's coefficients : $\frac{\partial c}{\partial \beta_k} = -2\sum_{j,(k,j)\in K}(R_{kj} - \beta_k - \tilde{R}_{kj}) + 2\eta\beta_k$. Setting $\frac{\partial c}{\partial \beta_k} = 0$ provides the

update : $\beta_k = \frac{\sum_{j,(k,j)\in K} R_{kj} - \tilde{R}_{kj}}{\eta + |K|}$.

We modify said updates in the `rec2` function, first changing the updates in A's coefficients (adding `beta[m]` in `Asum`), then B's coefficients (adding `beta[i]` in `Bsum`), and finally updating the β vector (using vectorized `numpy` methods).

Part 2

1

We study the numerical solutions given by a system two nonlinear partial differential equations. The system implies a parameter α and couples two functions of space and time : $u(x,t)$ and $v(x,t)$. We further focus on the solution u.

For a fixed space coordinate x_0, we study the time variations of $t \to u(x_0,t)$. The function follows an initial transient after which the influence of the initial conditions disappear and leaves an established dynamical regime. We focus on said longer term state. We display the variations of $t \to u(10,t)$ - computed for $\alpha = 0.1$ - in figure 10. We first notice that the obtained signal follows a periodic trend, which seem to be a superposition of sin trends. In particular, we visually identify two sinusoidal patterns : one giving high frequency oscillations and the other inducing lower frequency peak heights, the signal being overall very close to a pure sin function. To confirm this behavior, we perform a Fourier analysis of said signal and we visualize its spectrum in figure 11. We identify two opposite high magnitude thin peaks accounting for the main pure sinusoidal identified trend. A much smaller although still visible peak is located alongside the principal one, which may account for the lower frequency trend.

We display the variations of $t \to u(10,t)$ - this time computed for $\alpha = 3$ - in figure 12. We observe that there is no more periodicity in the signal, which now follows a more chaotic behavior. Indeed, no clear pattern appear in the time variations of the signal, the magnitude of subsequent peaks seeming rather random. This unpredictable behavior arises from a non linear although deterministic system of equations. We investigate further performing a Fourier analysis, shown in figure 13. We now observe the existence of 4 peaks : 2 centered high magnitude ones and two outer low magnitude ones. The peaks are here relatively thick, displaying a spreading in the frequencies corresponding to the more numerous sinusoidal constituents of the signal : it is now far from a pure sin trend.

We uncover the effects of parameter α in the dynamics of the solution function u. Said parameter plays somewhat the same role as the growth rate r in the logistic map, controlling the number of patterns appearing in the periodic signal (for instance see patterns for $\alpha = 1.6$ in figure 14), and having some ranges of value for which it produces a chaotic behavior. In figure 17 we computed a *bifurcation graph*, displaying the different extremums values present in the signal for a given α. In particular, we observe that the chaotic behavior appears first for $\alpha \approx 1.75$.

We finally investigate the variations of u for fixed values of t. We computed $x \to u(x,t_0)$ for $t_0 \in \{50, 100, 150\}$, for $\alpha = 0.1$ and $\alpha = 3$, in figures 15 and 16. We observe that the corresponding spatial trend are more even for $\alpha = 0.1$ and relatively wiggling for $\alpha = 3$, displaying the specific effect that this parameter also has on spatial variations.

2.a

We rewrite the implicit scheme as a matrix equation involving a tridiagonal coefficient matrix and a banded right-hand-side matrix. We build said matrices and compute their solution using `scipy`'s banded solve algorithm. Note that the specific tridiagonal solve TDMA algorithm could have also been used in this case. We effectively store the tridiagonal matrix using the *matrix diagonal ordered form* (storing only the diagonals). The implementation can be found in the `implicitFD`.

2.b

Not attempted.

Figures

Part 1 — Question 1

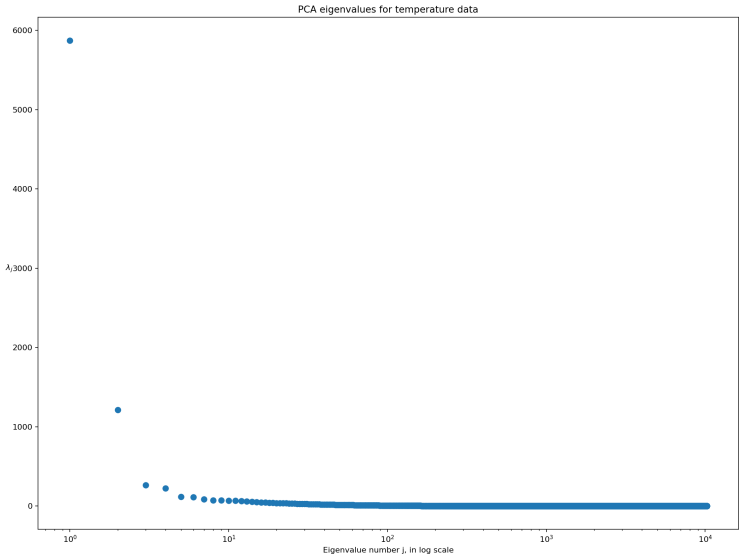

Figure 1: Figure for Part 1 Question 1 — PCA eigenvalues for temperature data

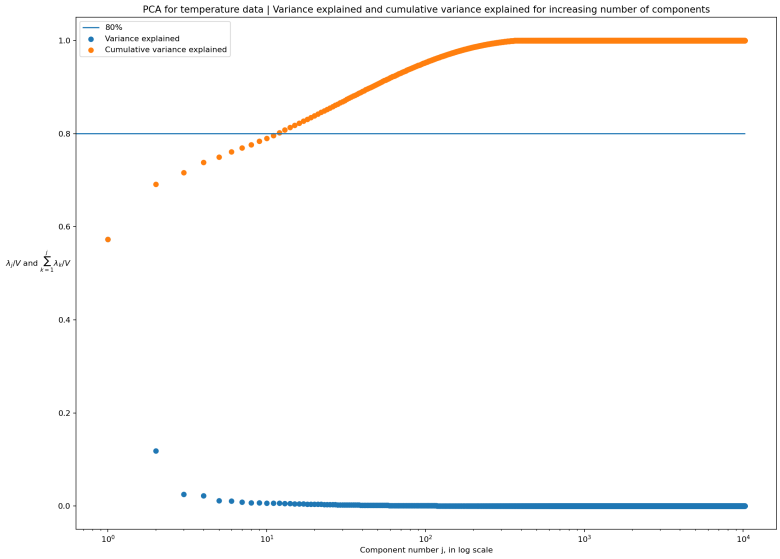

Figure 2: Figure for Part 1 Question 1 — PCA for temperature data : Variance explained and cumulative variance explained for increasing number of components

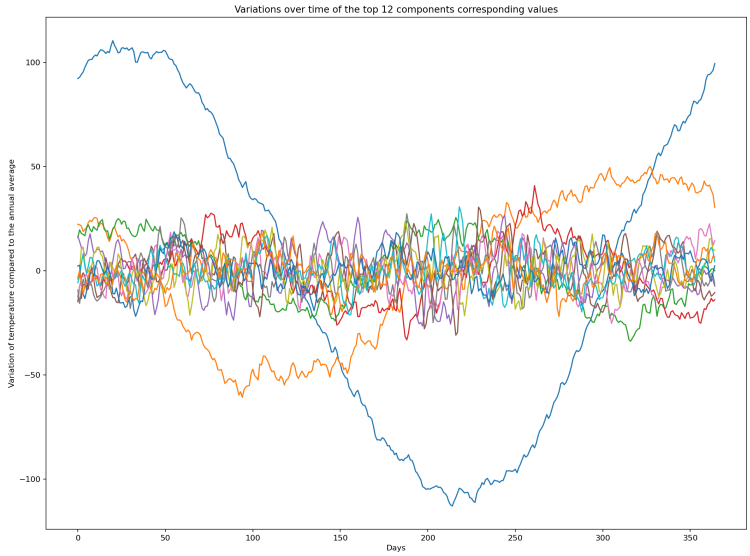

Figure 3: Figure for Part 1 Question 1 — Variations over time of the top 12 components corresponding values

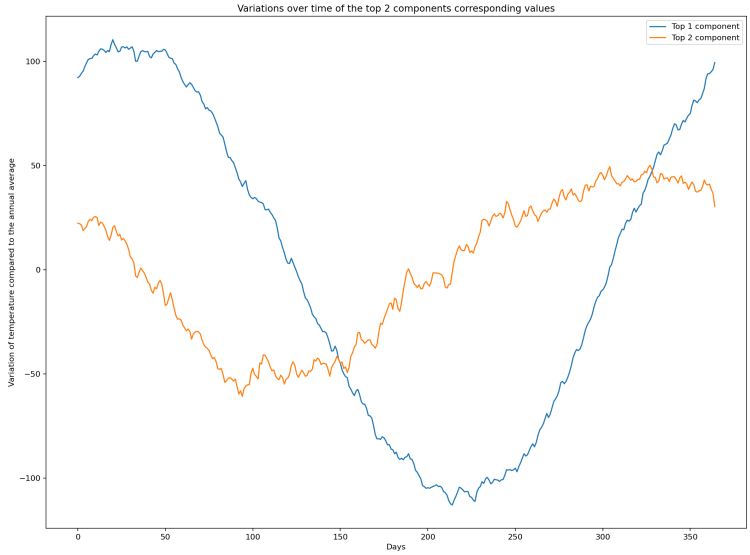

Figure 4: Figure for Part 1 Question 1 — Variations over time of the top 2 components corresponding values

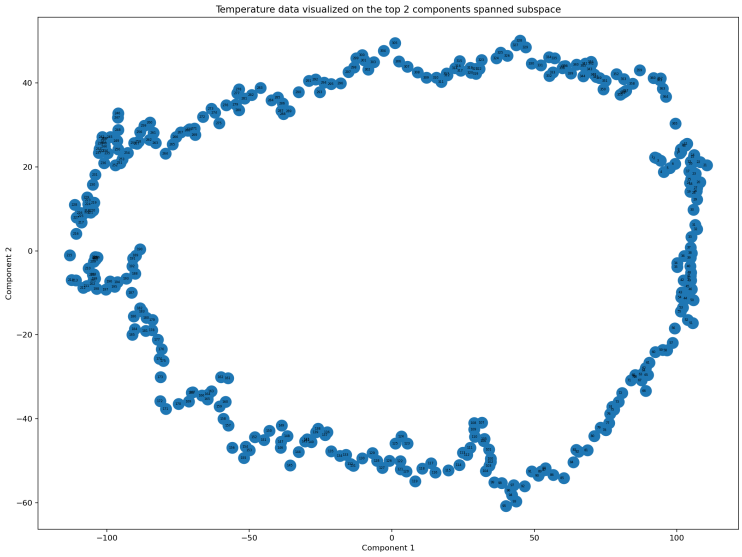

Figure 5: Figure for Part 1 Question 1 — Temperature data visualized on the top 2 components spanned subspace : each sample is labeled with its corresponding day

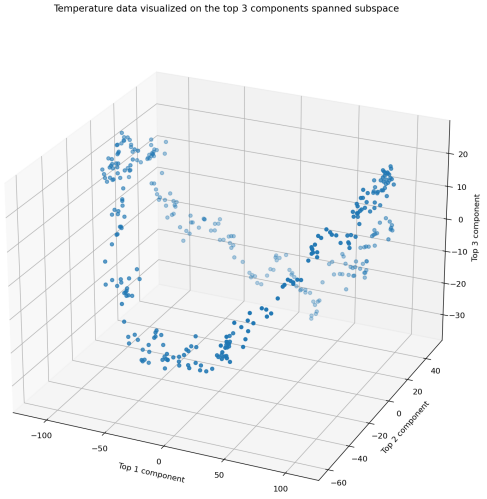

Figure 6: Figure for Part 1 Question 1 — Temperature data visualized on the top 3 components spanned subspace

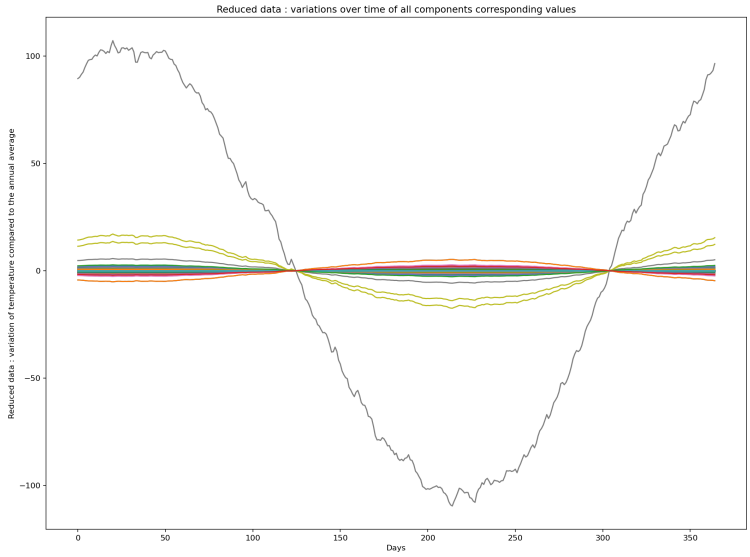

Figure 7: Figure for Part 1 Question 1 — Reduced data : variations over time of all components corresponding values

Part 1 — Question 2.a

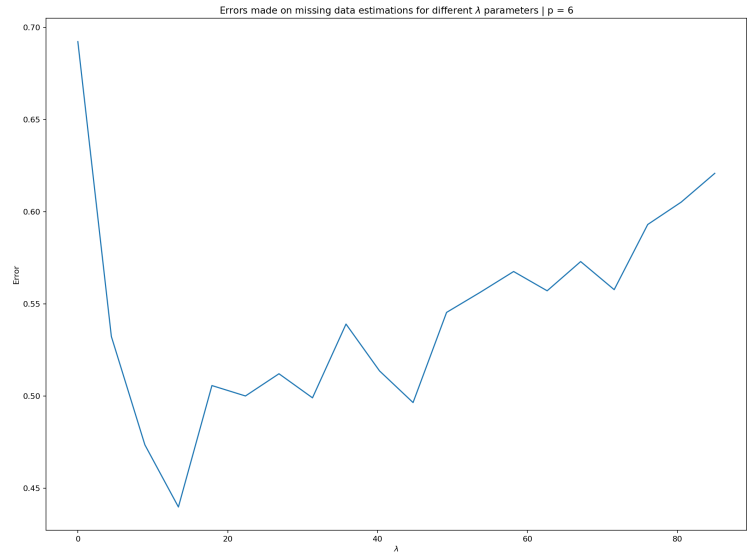

Figure 8: Figure for Part 1 Question 2.a — Errors made on missing data estimations for different λ parameters — p = 6

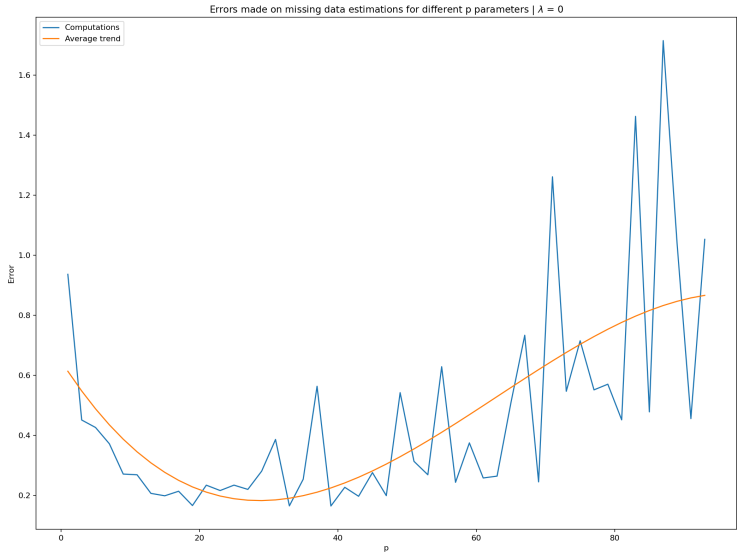

Figure 9: Figure for Part 1 Question 2.a — Errors made on missing data estimations for different p parameters — $\lambda = 0$

Part 2 — Question 1

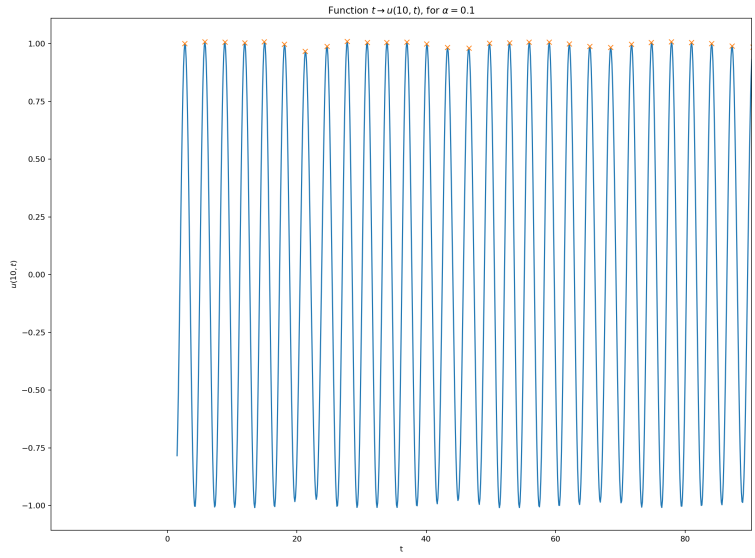

Figure 10: Figure for Part 2 Question 1 — Function $t \to u(10, t)$, for $\alpha = 0.1$

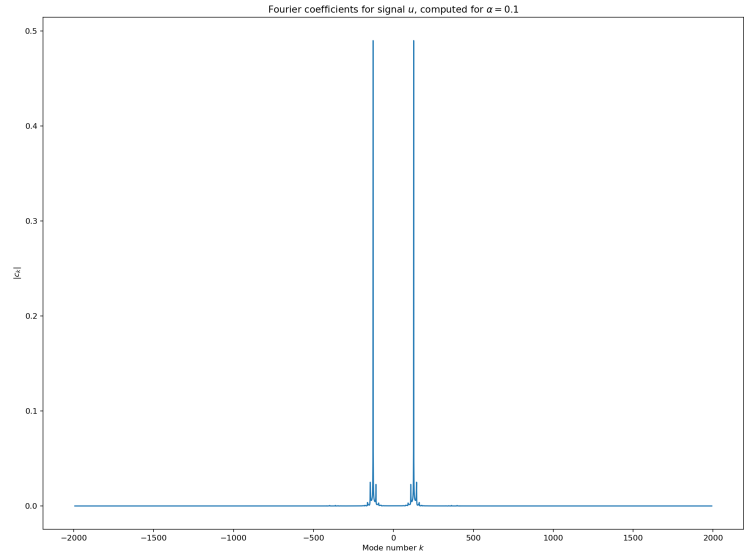

Figure 11: Figure for Part 2 Question 1 — Fourier coefficients for signal u, computed for $\alpha = 0.1$

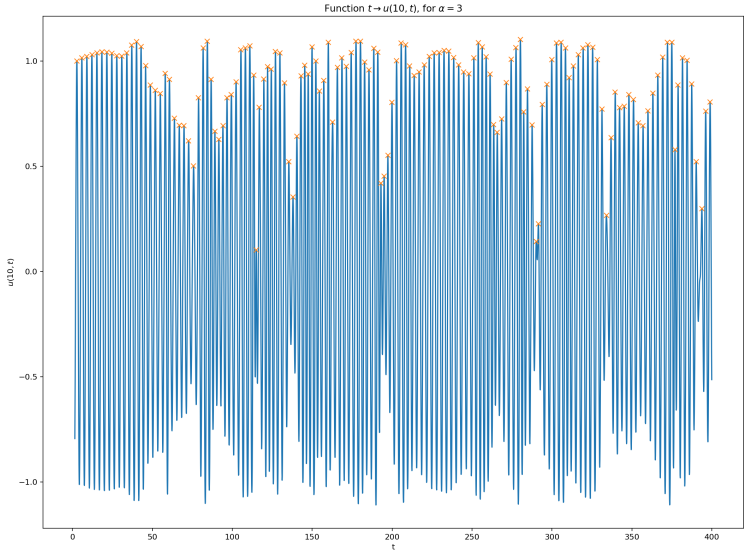

Figure 12: Figure for Part 2 Question 1 — Function $t \to u(10, t)$, for $\alpha = 3$

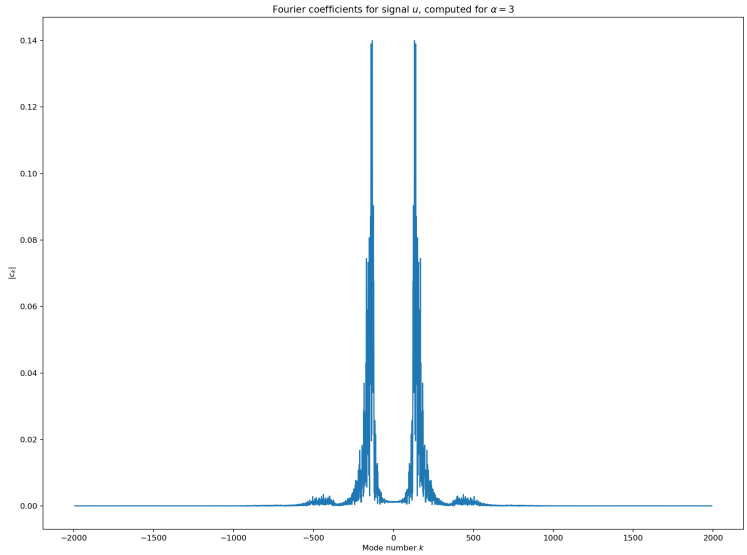

Figure 13: Figure for Part 2 Question 1 — Fourier coefficients for signal u, computed for $\alpha = 3$

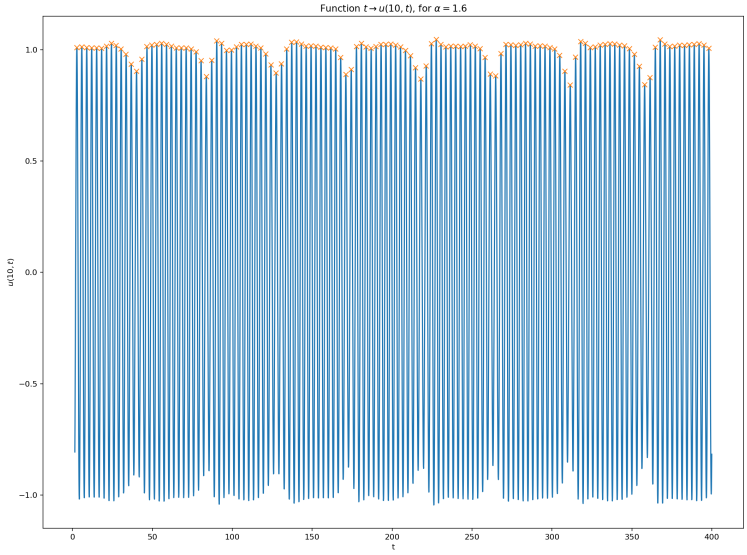

Figure 14: Figure for Part 2 Question 1 — Function $t \to u(10, t)$, for $\alpha = 1.6$

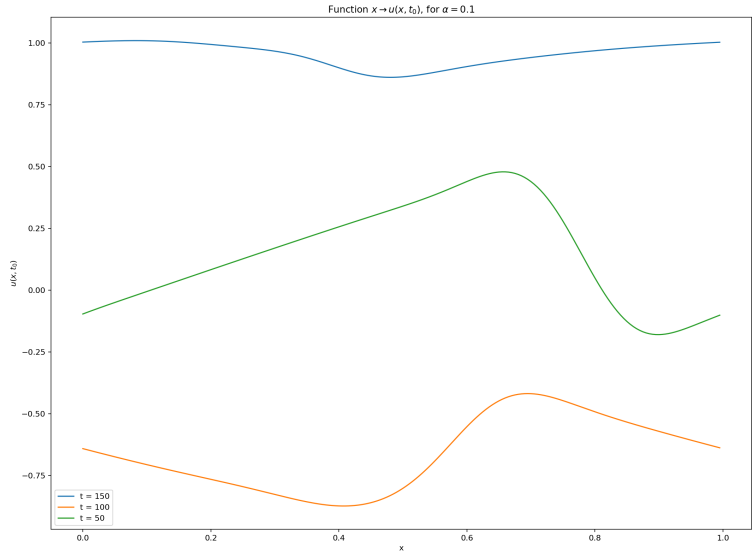

Figure 15: Figure for Part 2 Question 1 — Function $x \to u(x, t_0)$, for $\alpha = 0.1$

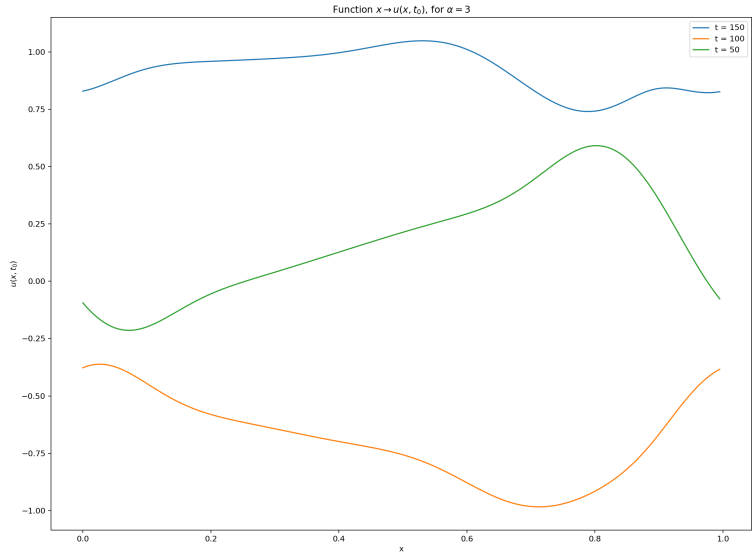

Figure 16: Figure for Part 2 Question 1 — Function $x \to u(x, t_0)$, for $\alpha = 3$

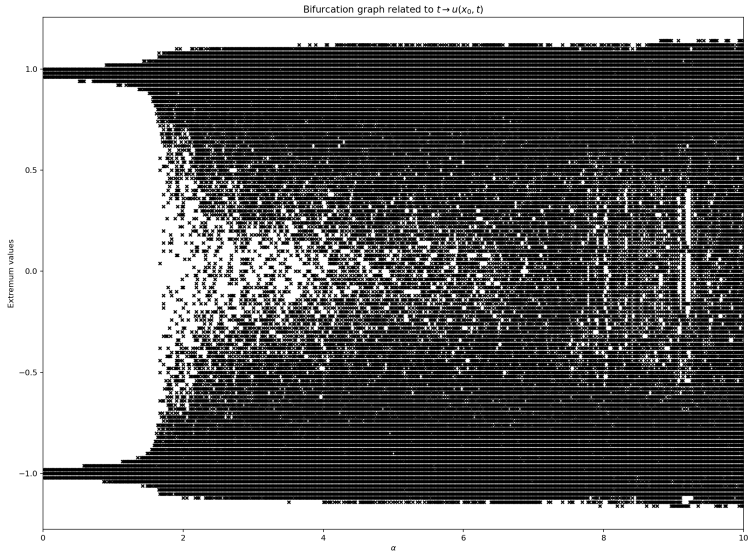

Figure 17: Figure for Part 2 Question 1 — Bifurcation graph related to $t \to u(x_0, t)$