JOURNAL OF MATHEMATICAL SCIENCES AND MODELLING

ISSN: 2636-8692

VOLUME III
ISSUE II

VOLUME III ISSUE II
ISSN 2636-8692

August 2020
http://dergipark.gov.tr/jmsm

JOURNAL OF MATHEMATICAL SCIENCES AND MODELLING

Editors

Editorial Board of Journal of Mathematical Sciences and Modelling

Contents

Journal of Mathematical Sciences and Modelling, 3 (2) (2020) 47-54
Research Article

Journal of Mathematical Sciences and Modelling

Journal Homepage: www.dergipark.gov.tr/jmsm
ISSN 2636-8692
DOI: http://dx.doi.org/10.33187/jmsm.691846

Similarity Among Physical Phenomena Recognized on the Basis of the Classification of Existing Knowledge

Antonios Kanavouras[1] **and Frank A. Coutelieris**[2*]

[1]*Department of Food Science and Human Nutrition, Agricultural University of Athens, 55 IeraOdos Str., GR-11855, Athens, Greece*
[2]*Department of Environmental Engineering, School of Engineering, University of Patras, 2 Seferi Str., GR-30100 Agrinio, Greece*
[*]*Corresponding author E-mail: fcoutelieris@upatras.gr*

Article Info

Keywords: *Classification matrix, Engineering assets, Linear algebra, Physical phenomena, Similarity,*
2010 AMS:
Received: *20 February 2020*
Accepted: *31 August 2020*
Available online: *31 August 2020*

Abstract

This work presents the synthetical mathematical analysis of available knowledge regarding physical phenomena of research interest. The work is not focusing on providing the phenomena according to the physical laws but rather because of them, hence, it is grounded on the philosophically defined concept of "similarity", and progresses to the mathematical treatments of those factors and parameters that are involved into the similarity validation among physical phenomena.A critical validation regarding the effectiveness of such an approach was also performed, in order to conceptualize the relevance of the factors and parameters interactions as a potential control tool against engineering-based hypothesis. Such factors and parameters are generated through the description and delimitation of the system of interest. A "matrix" is used for the classification of the existing knowledge regarding this system. It is consisted of the categorical descriptors of the system in question and the levels of these descriptors.A mathematical analysis of this "matrix" supports that all the existent perceptions of a physical phenomenon constitute a four-dimensional vector space. Within this space, the concept of similarity allows for the definition on which of a specific non-linear mapping that might be applied to strictly classify the existing knowledge about the phenomenon in question.Similarity is used here to define the conditions and the constrains that this mapping must satisfy. In conclusion, the applicability of the suggested approach on an engineering approach regarding a physical problem, was also demonstrated in this study.

1. Introduction

Experimentation is providing the evidently ground base for criticism and rational discussion. The outcome of such a process provides a reason-based scientific knowledge of any world of phenomena. Experiments are the fields of testing theories for their validity, need for changes or corrections. While a new theory may often derive along with a new phenomenon that needs explanation [1]. So far, scientific studies are based on either empirical or transcendental research approaches, aiming towards picturing the identity and/or the differences in time occurring throughout the phenomenon under investigation. Accordingly, research outcome is logically analyzed, and results are combined with empiricism and are all placed against the validity of the scientific hypothesis and its theoretical and practical interpretations. However, any reported result throughout the study is applied instantly and therefore records a series of incidents at certain time intervals. The empirical observations and/or their correlations are only to be further enhanced with an exceeding accumulation of data. In this context, incommensurability is a common situation impacting the objective view of researchers and the meaning they provide and the criteria they implement. The so called, social behavior of the researchers reflected on their commitments as well as on their objectivity is rather rich and diverse as may be realized via the multiplicity of research objects, fields and research subdivisions [2, 3].

***Email addresses and ORCID numbers*: antonios.kanavouras@aua.gr, https://orcid.org/0000- (A. Kanavouras), fcoutelieris@upatras.gr, https://orcid.org/0000- (F. A. Coutelieris)**

Furthermore, the attempt to induct the empirical justification into a theoretical interpretation is additionally present. Although vastly applied, statistical analysis results may contradict to the potential meaning and significance of the independent phenomenological events and individual steps or causes. It is more than obvious that the outcome of such applicational treatments is focusing on the ”general” evolutionary trend, letting away of any valuable incorporation of any data differences that may be eventually lost when the general curve ”trends” or the statistical significance level of the differences are presented. Supplementarytreatments on the curves with geometrical means in order to get additional estimations, such as kinetic parameters and phenomenological rates, are also applied. Therefore, derived conclusions may be highly biased by the selected treatments [4].

It is important to note that any physically executed experiment is actually a reproduction of the physical world under controlled conditions, hence within an artificial environment, namely the lab or an experimental field.What allows for such a transfer is actually, the acceptance that individual phenomena or their classes are referenced on the ground of similarities. Such similarities are constructed or perceived by the lead researcher on the ground of his/her fundamental considerations [5]. Such a reproduction is valid and accepted when it is applied in a standard and repeatable manner [6]. Having said that, the perception of the phenomena under study, is initiated with the essential initial hypothesis and is further build using the empirical observations that eventually lead to a logical understanding fitted well into a theory and practice [3].

Therefore, what is reality may well be perceived via an infinite number of phenomena, reproduced by the researchers in order to approach an objective view of the world, yet with questionable precision level. Moreover, and after having defined that two systems are similar, we may safely interchangeably infer the values of quantities in each of the physical systems from known values of each other. In other words, among two similar systems, it is allowed to draw conclusive outcomes for quantities based on a certain similarity mapping [7].

2. Similarity concept

Any experiment aiming in the collection of data, hence, empirical evidences, aims in identifying that information that may allow for further and broader view of the events well fitting into a theoretical method. The extrapolation of the knowledge gained in such a way, is valid and applicable, on the ground of similarity among the applied phenomena and fields and information may flow interchangeable among them

Nevertheless, the definite similarity of events and/or their classes background remains a question [8]. The answer apparently, may derive when considering that it is allowed to build comparative relationships and draw comparisons on the ground of the objects themselves or the hypothesis under which these objects are investigated [9].

As in any search attempt, the selection of criteria are fundamental as well for identifyingsimilarity. [7] suggested that a classification of the existing knowledge is necessary. Evidently, any inappropriate application of similarity criteria might easily lead to drawing of results lacking solid logical and/or meaningful ground. The role of any researcher is rather apparent in that process in order to avoid research waist and dead ends. Following that, any weak or erroneous used or selected similarity criteria shall guide the researchers into spending money and effort for to gain already existing knowledge, leading to resources’ waste.

A number of knowledge classification attempts have been proposed, based on empirical and arbitrary rules and definitions [4]. In order to fulfil the requirement of similarity criteria selection and application, the work in hand suggests a solid mathematical treatment of identifying similarity and establishing robust similarity criteria through the application of fundamental Linear Algebra concepts (namely, vector spaces and mapping between them) on the relative philosophical aspects already arisen.

3. Knowledge classification

Scientists have aimed in studying any phenomenon in the Newtonian 3-dimensional world with certitude and clearness, completely in the light of principles, and, after having discovered the cause of the doubts and contradictions into which reason fell, to initiate an attempt to have them solved to its highest satisfaction possible.we will call ”judgement” this inner logic that approaches the hypothesis with the thought that a thing is known to have a certain quality or attribute.

Judgement would be herein appointed as either ”analytical”, which when applied will expose the opposite sense to a logical contradictions, or ”synthetical”, in a sense that are additionally supporting that hypothesis, via extra treatments, such as in the case of mathematical laws applied as a judgemental tool. Furthermore, we shall call ”proposals” the valid ways for approaching this logic. The *a-posteriori* proposals are those providing the possibility, while the *a-priori* proposals are those being based on inevitability and universality, distinguishing them as non – revisable. Overall, each and every way of approaching the logic may be either analytical or synthetic

Consistently, by combining the above, the following common general proposals derive: a) verification or the a-posteriori-analytical proposals, b) validation, or a-posteriori-synthetical proposals and c) prediction, or a-priori-analytical proposals. The current work is not focusing on providing the phenomena according to the physical laws but rather because of them, implying that our goal is a knowledge that is simultaneously non-empirical and a priori.

There may be no question regarding the phenomena non-existence before their expression and consequent empirical experience recorded and reported, only constructed through a local and cognitional process. Nevertheless, the level of distinction among the outcomes of the aforementioned process for various researchers, is not easily identified. Furthermore, a general principle of all the analogies rests on the synthetical unity of all phenomena according to their relation in time [10]. In order to overcome such obstacles, the use of mathematical synthetical treatments may allow for a synthetical representation of the perception of a phenomenon, [11].

In order to apply the above, we shall primarily need to define the functionality of the system under investigation and attempt to impend the empirical connections, sense and significance into its hypothesis. A summary of the organizational scheme for this work, is presented in Table 1, were the sequential (left to right and top to bottom) proposed developments, named the ”epistemic structure” of the final engineering judgement, are acknowledged.

In specific, the developmental roadmap proposed, is divided in three (4.2) Stages. Within each development Stage and at the transitional

Table 1: The developmental roadmap for organizing the work

Stages	1	Proposals	2	Proposals	3
Principles		Physical	Intellectual	Mathematical	Engineering
Proctorship		Empirical	Analytical	Formulation	Relativistic
Functions of Judgment		Potential	Synthetical	Necessity	Hypothesis potential conditions
Deliverables	System		Categories		Model

steps in-between, we shall aim in applying a judgemental process on the critical derived milestones (”Deliverables”). In order to do so we shall implement the essential scientific values for this particular system (”Principles”) and the potential ways and methodology that will allow us to identify the interconnections (”Proctorship”) among the systemic parts. This whole procedure will eventually define the outmost interesting deliverable, that may be used as an engineering tool and asset, the ”Model” in Stage 3.

A detailed description of the fundamental activities we acknowledge as per Stage of Table 1, is following. That includes: i) the recognition and classification of the inherent systemic ”ways-of-order” apprehendedas the categorical descriptors of the particular system and ii) the mathematical expressions that are evidently incorporating the principles and the logic and consequently the proof as a straight forward powerful application tool that is capable of assessing the system against the initial hypothesis.

a) **Stage 1**

We shall initially proceed with a validation first action focusing on the physical description of the system. Taken that experience is always proceeding the developed and emerged knowledge, the procedural progress along stages 1 and 2 needs to take int consideration the physical consider both the physical philosophies of the system along with empiricism’s practices. This is only sto be solidly founded on both an appropriate experimental design and incorporate the much-needed monitoring means of the system’s development in time and in space. That will be further discussed later in this work (see Section 5).

As mentioned above, the application of certain methodological, technical and mechanical borders during experimentation practice, place the upcoming expression of the system under the control of the selected border conditions. And therefore, the experimental outcome. Having worked in that manner, allows for a highly trustworthy data collected from the potential combinations within an experimental design hence, allowing us to collect and structure a holistic approach of the system. Through that overall picture, we may be certain of the particular and unique cohesive points of the phenomena occurring in such a well-defined system. From a certain point of view, it is the rationality of any model that may assist an ”economic” experimentation.

b) **Stage 2**

The Deliverable in Stage 2 is the ”Categories”; that are deriving from the research hypothesis itself. The relevant classification scheme shall then necessary and successfully derive the need and aims which when capably applied shall properly complete the description of the system. The description may be perceived as the reflection of the system on humans and that is why it may well be seen asa formal act of human cognition.

Nonetheless, the overall Categories need to contain the hypothesis’ context possibilities for the system in question and regulate the system boundaries (classification frame), via impacting on each and every systemic participant, in-principal. Furthermore, the classification will reflect the theoretical solidness of the upcoming mathematical concept, independent of the empirical experience, in which case the deliverable model of Stage 3, ought to be capable of assessing the shaped categorical imperative of the hypothesis.

For an essential conceptual description of a system under investigation, we may convert the typical ”in-process-out” context into a ”system”, via the incorporation of the finest, yet minimal, but still straightforward, systemic participants of generalized activities. Such activities are adequately described by the following four principal ”categorical descriptors” in an equation form:

$$matter + energy \ \underline{\text{relationships}} > outcome \tag{3.1}$$

In order to advance the above core expression, the ”categorical descriptors” have to be further developed with each and every thing in experience, the properties, the qualities, and the characteristics of any possibility in general.

Accordingly, when the aforementioned four categorical descriptors placed in a classification matrix of three, empirically defined, levels were basically adopted and later on drafted for the hypothesis in particular. Notably, the three levels also firmly satisfy the mathematical principle for the least linearly independent points to follow a non-linear relationship, as presented below.

Looking at each of the descriptors level, we may identify that a set of conditions exist. Moreover, our critical judgement’s degrees of freedom are also present linking the descriptors to the hypothesis in question. Simply stated, all the three levels of each categorical descriptor, condense the impact for this systemic participant against the disclaiming of the hypothesis. Specifically, in Table 2, the four specific categories and their corresponding general levels are presented.

In that sense, the therefore formed 12 classes of the knowledge classification matrix there are the cohesions within the system context. These cohesions may be then expressed and presented with the appropriate mathematical expressions. Last but not least, it is an inter subjective

Table 2: The knowledge classification matrix

CATEGORIES	LEVELS		
Matter (Quantity)	One	many	all
Energy (Quality)	Reality	disallowance	restrictions
Inherent potential (Relationships)	inter-dependent	reasons	intra-dependent
Outcome (Processes)	Potential	existence	necessity

validation of the hypothesis, that is deriving through the cross sections of the classification cells , since the common ground among all classes' interactions may eventually deliver the basic and only physically meaningful and accountable potential conditions of knowledge.

The compulsory - exact analogy - filling-in of the specific hypothesis classes, demarcates the necessitating empirical and theoretical possibility conditions for the phenomena occur. As indicated in Table 1, the physical principles of the system along with the empirical (experimental and knowledge) available proctorship, shall be applied in this transitional step. The conceptual description of the system will, hence, border-lined upon a validation judgement assessment, corresponding to the classification scheme represented by Table 2, filled accordingly to the system, the hypothesis and the phenomena considered.

The validation analysis along with the categorical descriptions,allows the user to structure a system's awareness and consequently may generate the inherent mathematical relativistic expression that exist in the system and are based on their interrelationships. Having an and adequate consideration and performing a proper investigation of all classes, the process guarantees that, in principal, each and every systemic participant will accurately influence the boundaries of the model.

Finally, looking for a theoretical sound base of the system, the Categories may be used to provide the appropriate model, far beyond and above the empirical experience the study of the system may have or will allow. In that way, it is the model that shall assess the shaped categorical imperative of the hypothesis. This classification of the systemic properties will be the contextual guideline in logically conveying the physics in the system, from their mathematical description, to the engineering tool of Stage 3.

c) **Stage 3**

This stage is actually the use of the above two stages for a specific engineering problem.Therefore, a fulfilled 4x3 matrix can be produced, which contains all the necessary information about the knowledge regarding a specific physical phenomenon.

4. The mathematics of classification

Given the goals of this work, let us initially consider a physical and/or chemical phenomenon. What we have actually in mind is the perception of an incident or a series of interconnected incidents that belong to a specific class of physico chemical phenomena. This perception could be described by Equation (3.1) when and only when, the exclusive phenomenon's characteristics may be incorporated within the categorical descriptors specified for this phenomenon. Accordingly, this means that one has to quantify the "matter" and the "energy" involved, to identify a macroscopic magnitude that serves as the "outcome" of the phenomenon as well as to express a valid "relationship" among the parameters involved in the expression of the phenomenon in question. There should be a linear independence among these descriptors because (a)matter is a manifest object, (b) energy is a manifest object, as well, (c) relationship is a mathematical object being valid independently regarding where it is applied for, and (d) outcome is defined subjectively, yet, in-line with the scientific hypothesis, under which the phenomenon is studied.

To the extent that there may exist an infinite number of perceptions of any phenomenon, there are also infinite tetrads, consisted of values for the matter, energy, relationships and outcome categorical descriptors. Each one of these tetrads forms a vector of the form $\underline{u} = \{m, e, R, o\}$, where a *one-by-one* correspondence exists between the values of vector components and a specific perception. Then, we may now determine a set V, consisted of all these vectors $\underline{u}$. There is a non-obvious relation between vectors $\underline{u}$ and time: each vector corresponds to the whole amount of knowledge regarding the phenomenon, which is available at a specific time. This means that every element of V reflects a precise value of time. In that sense, vectors $\underline{u} \in V$ can be considered as functions of time.

Hence, it could now become able to underline that the set $V$is an ordered set where an arrangement, denoted by $\prec$, can be defined as follows:

$$\text{for } \hat{t}_1 < \hat{t}_2 \Leftrightarrow \underline{u}_1(\hat{t}_1) \prec \underline{u}_2(\hat{t}_2) \tag{4.1}$$

Note also that each $\underline{u} \in V$ embeds all the vectors of lower values, i.e. all the knowledge previously obtained and currently available. By considering a physical phenomenon along with a disclaiming hypothesis superimposed (as previously stated), every new perception consists of the current knowledge about the phenomenon plus a new contribution

$$\underline{u}_i(\hat{t}_i) = \underline{u}_{i-1}(\hat{t}_{i-1}) + A \tag{4.2}$$

where $\underline{u}_i(\hat{t}_i)$is the current of knowledge, $\underline{u}_{i-1}(\hat{t}_{i-1})$ is the previously obtained of knowledge and A is the amount of knowledge added to the previous knowledge in order to produce the current one.

Having described the arrangement of the set V, we can define an internal operation $\oplus$ as follows

$$\forall \underline{v}, \underline{w} \in V \; \exists \underline{u} \in V : \underline{u} = \underline{v} \oplus \underline{w} = \begin{cases} \underline{v} & \text{if } \underline{w} \prec \underline{v} \\ \underline{w} & \text{if } \underline{v} \prec \underline{w} \end{cases} \tag{4.3}$$

This operation describes mathematically the knowledge evolution through time, which is obtained by "adding" new contribution on the existing accumulated knowledge.

Regarding its properties, this operation

(a) is commutative

$$\underline{v},\underline{w} \in V \; \underline{v} \oplus \underline{w} = \underline{w} \oplus \underline{v} \tag{4.4}$$

Proof:

If

$$\underline{w} \prec \underline{u} \Longrightarrow \left.\begin{array}{l} \underline{u} \oplus \underline{w} = \underline{u} \\ \underline{w} \oplus \underline{u} = \underline{u} \end{array}\right\} \Rightarrow \underline{u} \oplus \underline{w} = \underline{w} \oplus \underline{u} \tag{4.5}$$

If

$$\underline{u} \prec \underline{w} \Longrightarrow \left.\begin{array}{l} \underline{u} \oplus \underline{w} = w \\ \underline{w} \oplus \underline{u} = \underline{w} \end{array}\right\} \Rightarrow \underline{u} \oplus \underline{w} = \underline{w} \oplus \underline{u} \tag{4.6}$$

(b) is associative

$$\forall \underline{v},\underline{w},\underline{u} \in V \; \underline{v} \oplus (\underline{w} \oplus \underline{u}) = (\underline{w} \oplus \underline{v}) \oplus \underline{u} \tag{4.7}$$

Proof:

If

$$\underline{v} \prec \underline{w} \prec \underline{u} \Longrightarrow \left.\begin{array}{l} \underline{v} \oplus (\underline{u} \oplus \underline{w}) = \underline{v} \oplus \underline{u} = \underline{u} \\ (\underline{w} \oplus \underline{v}) \oplus \underline{u} = \underline{w} \oplus \underline{u} = \underline{u} \end{array}\right\} \Rightarrow \underline{v} \oplus (\underline{u} \oplus \underline{w}) = (\underline{w} \oplus \underline{v}) \oplus \underline{u} \tag{4.8}$$

The other cases for the arrangement ($\underline{v} \prec \underline{w} \prec \underline{u}$, $\underline{v} \prec \underline{u} \prec \underline{w}$, $\underline{u} \prec \underline{v} \prec \underline{w}$, $\underline{w} \prec \underline{v} \prec \underline{u}$, $\underline{u} \prec \underline{w} \prec \underline{v}$ and $\underline{w} \prec \underline{u} \prec \underline{v}$) can be proven in the same way.

(c) has identity element

$$\forall \underline{v} \in V \; \exists \underline{0} \in V\colon \underline{v} \oplus \underline{0} = \underline{0} \oplus \underline{v} = \underline{v} \tag{4.9}$$

Proof:

The element $\underline{0}$stands for the trivial knowledge that tends to zero, thus not actually favoring knowledge evolution.

(d) each vector in V has inverse elements

$$\forall \underline{v} \in V \; \exists (-\underline{v}) \in V\colon \underline{v} \oplus (-\underline{v}) = (-\underline{v}) \oplus \underline{v} = \underline{0} \tag{4.10}$$

Proof:

Vector $(-\underline{v})$ in fact represents a singularity on the knowledge evolution, i.e. one point the knowledge of which tends to collapse all the existing knowledge.

As far as $\underline{u}$ is a vector, its regular norm $\lambda = \|\underline{u}\| \in \mathbb{R}$ is the amount of knowledge that is embedded in the perception of the phenomenon. This norm allows for the quantification of the amount of knowledge between any two elements of V, and given that V is an ordered set, the ratio of such quantities is a realization of knowledge evolution. In terms of mathematics, the above can be expressed as

$$\text{if } \underline{u} \prec \underline{v} \text{ with } \underline{u},\underline{v} \in V, \text{ then } \exists \mu \in (1,+\infty) : \mu = \frac{\|\underline{u}\|}{\|\underline{v}\|} \tag{4.11}$$

We can now define another operation $\bigotimes$ as follows:

$$\forall \underline{v},\underline{w} \in V \; \exists \mu \in R : \; \underline{w} = \mu \bigotimes \underline{v} \Leftrightarrow \mu = \frac{\|\underline{w}\|}{\|\underline{v}\|} \tag{4.12}$$

which describes the relative importance of the difference in knowledge amounts between two particular time periods, i.e. between two perceptions expressed in different times.

Having determined the set V as well the operations $\bigoplus$ by eq. (4.3) and $\bigotimes$ by eq. (4.12), we can now that the group $\{V, \bigoplus, \bigotimes\}$ is a 4^{th} dimensional vector space. It is assured because, on top of the properties of operation $\bigoplus$ presented through eqs. (4.4) – (4.10), it is straightforward to prove that operation $\bigotimes$ satisfies the following relations

$$\lambda\left(\mu\bigotimes\underline{v}\right)=\lambda\mu\bigotimes\underline{v} \tag{4.13}$$

$$(\lambda+\mu)\bigotimes\underline{v}=\lambda\bigotimes\underline{v}\bigoplus\ \mu\bigotimes\underline{v} \tag{4.14}$$

$$\lambda\left(\underline{u}\bigoplus\underline{v}\right)=\lambda\underline{u}\bigoplus\ \mu\underline{v} \tag{4.15}$$

Note also that the dimensions of this vector space are defined by the number of categorical descriptors, as visualized in eq. (3.1). One basis of this space consists of the four vectors $\underline{e}_m=\{m,0,0,0\}$, $\underline{e}_e=\{0,e,0,0\}$,$\underline{e}_R=\{0,0,R,0\}$ and $\underline{e}_o=\{0,0,0,o\}$. We may then report that,

(a) the components of the basis are linearly independent since they reflect the independent categorical descriptors, as previously stated

(b) any vector in the space is described by a unique linear combination of the basis vectors.

Finally, it could be able to identify a non-linear mapping m_p^{In} over the space $\{V,\bigoplus,\bigotimes\}$, as follows

$$\mathbb{R}xV\xrightarrow{m_p^{In}}M_{3x1}(V):\ m_p^{In}(\underline{v})=\{\ \ \lambda_1\underline{v},\quad \lambda_2\underline{v},\quad \lambda_3\underline{v}\ \ \} \tag{4.16}$$

with

$$\lambda_1\to 0 \tag{4.17}$$

$$\forall\lambda_2\in\mathbb{R}\exists M>0:\ \lambda_2>M \tag{4.18}$$

$$\lambda_3\to+\infty \tag{4.19}$$

The first component of the product, given by eq. (4.16) under the constraint (4.17), corresponds to a very low amount of knowledge that tends to zero but it should always be non-zero, meaning that it is even at its minimum level enough to initiate the scientific/research interest for studying the specific phenomenon. The next component describes the currently available knowledge while the last one, along with constraint (4.19) represents the nearly infinite amount of knowledge, needed to be obtained in order to finalize the research on this topic. The above eqs. (4.17)-(4.19) assure the compatibility of "one-many-all", commonly used in modern philosophy as a well-known Kantian wit [12].

The above mapping [eq. (4.16)] produces a matrix with four lines, each one standing for each of the elements $\{m,e,R,o\}$, and three columns, the first for the vector $\lambda_1\underline{v}$, the second for the $\lambda_2\underline{v}$ and the third for the $\lambda_3\underline{v}$. As previously stated, the first column refers to a vector containing the minimum non-zero knowledge of the phenomenon, where only one variable, along with only one mathematical equation produced by one simple conservation law or a relative mass/energy balance, are considered adequate to describe the particular perception. In fact, the first column of the matrix produces a rather primitive ideal outcome, which can roughly represent the phenomenon. The second column refers to the maximum finite knowledge currently available, where a finite number of variables are selected to describe the phenomenon and, therefore, a system of equations is produced, while a single one parameter is again selected to macroscopically describe the phenomenon. Briefly speaking, the second vector is a more accurate and more efficient representation of the phenomenon under consideration. Finally, the third column describes the absolutely holistic perception of the phenomenon, taking into account an infinite number of variables that define a system of equations with infinite dimension. In other words, the third vector describes the overall currently available knowledge about a phenomenon, identifying all the parameters' impact, although not necessarily known in full details.

Following the order of set V, it is

$$\lambda_1\underline{v}\prec\lambda_2\underline{v}\prec\lambda_3\underline{v} \tag{4.20}$$

where each vector contains the knowledge represented by the previous vectors.

We support that the non-linear mapping defined by eqs. (4.16), (4.17)-(4.19), quantifies the similarity between any two perceptions of the phenomenon in question, in terms of the way the evolution of knowledge occurs withing and cross the phenomenon's expressions.

It is also important that the definition of such a mapping is not unique. In order for the researcher to select the most suitable and appropriate mapping, i.e. to select values for λ_1, λ_2, λ_3 as well as m, e, R and o, a complicated methodology must be followed, as described in detail elsewhere [13]. The final aim of the application of this methodology is to identify potential lacks of knowledge, in order to direct the relative research and to avoid resources waste.

Table 3: Classification matrix for adsorption in granular media

	$\lambda_1 \underline{v}$	$\lambda_2 \underline{v}$	$\lambda_3 \underline{v}$
Matter	A	A, solid	A, solid, products B_i
Energy	Diffusion: $\underline{j}_A = -D_A \nabla C_A$	Diffusion: $\underline{j}_A = -D_A \nabla C_A$	Diffusion: $\underline{j}_i = -D_i \nabla C_i$
	Convection: $\underline{j}_A = U_A C_A$	Convection: $\underline{j}_A = U_A C_A$	Convection: $\underline{j}_i = U_i C_i$
	Instantaneous adsorption $C_A(r = R) = 0$	Adsorption by isotherm $D_A \underline{n} \cdot \nabla c_A = \frac{k}{K} c_s$ $\Theta_{eq} = \frac{K c_b}{1 + K c_b}$	Reaction of first order $A \rightarrow B$, with reaction rate $R_n = k_0 e^{\frac{-E_A}{RT}} c_A$
Relation ships	$\frac{dC_A}{dt} + \underline{U}_A \cdot \nabla C_A = D_A \nabla^2 C_A$	$\frac{dC_A}{dt} + \underline{U}_A \cdot \nabla C_A = D_A \nabla^2 C_A$	$\frac{dC_i}{dt} + \underline{U}_i \cdot \nabla C_i = D_i \nabla^2 C_i$
Outcome	$C_A(\underline{r}, t)$	$Sh_o = \frac{k_o \cdot L}{D}$	$\lambda_0 = 1 - \frac{\iint_{S_{outlet}} c_A \underline{U} \cdot \underline{n} dS}{\iint_{S_{inlet}} c_A \underline{U} \cdot \underline{n} dS}$

5. Application & discussion

In order to further support our approach, we shall now provide an application of the above described methodology to the physico-chemical phenomenon of mass transport in porous materials. The geometry where the phenomena occur is described by a swarm of uniform solid spheres. It is assumed that a Newtonian dilution flowsunder laminar flow conditionsthroughout the void space. There is a substance A diluted in the fluid phase, that flows in the pore volume and can be adsorbed on the solid surface of the grains. In general, mass transport of A is driven by diffusion and convection in the fluid phase, and by the considered sorption mechanism on the fluid-solid interfaces. All the above can be mathematically summarized in a generalized form of a vector $\underline{u} \in V$ described as [14]:

$$\begin{aligned} \underline{u} \quad = \quad & \{A, \text{ mass transport mechanisms, } \frac{dC_A}{dt} + \underline{U}_A \cdot \nabla C_A = D_A \nabla^2 C_A \pm R_A \\ & \text{with appropriate BCs, macroscopic quantity described adsorption}\} \end{aligned} \tag{5.1}$$

In this context,the output of the non-linear mapping is presented in the following 4X3 matrix:

The highly significant point to underline for Table 3, is that it unfolds through levels. In terms of mathematics given in eq. (4.16), the mapping can be defined through the identification of vector $\underline{v}$ and real numbers λ_1, λ_2 and λ_3. Precisely,

$$\underline{v} = \{A, \text{convective-diffusion with instantaneous adsorption}, \frac{dC_A}{dt} + \underline{U}_A \cdot \nabla C_A = D_A \nabla^2 C_A \text{with } C_A(r = R) = 0 \, , C_A(\underline{r}, t)\} \tag{5.2}$$

therefore

$$\lambda_1 = 1 \tag{5.3}$$

$$\lambda_2 = \text{ any finite number} \tag{5.4}$$

$$\lambda_3 \rightarrow \infty. \tag{5.5}$$

6. Conclusions

Systemic characteristics, predicates, attributes, qualities, or properties control the progress of the phenomena. A justified selection of them has to be included at certain fine-tuned levels, during the experimental phase, for an ably economic combination at the experimental phase and a comparable outcome data. Such a holistic approach shall adequately reveal the cohesiveness among the phenomena, rather than their simple description. Common ground establishments, under which certain common characteristic are shared among classes, may somehow be related to one another. Creating and applying precise conditions,that shall define the progress of any experimental phase, have a well-defined and significant role in allowing the system in question to unroll and progress on the base of its systemic properties, towards the particular processes' outcome, at empirically controlled environments.

Finally, the matrix may possess an additional added value, which is to confirm the experimental set-up that shall lead researchers to set the out most pragmatic input to an engineering model. The engineering tool's usefulness relies on the fact that it may work far beyond the judgemental capabilities of typical predictions and verifications that a mathematical model typically provides, yet such a model allows for

the identification and management of the experimentation risk options. These should accordingly hold the overall optimum engineering that shall provide the minimum possibility of disclaiming the fundamental disclaiming hypothesis.

Utilizing the herein presented concept of similarity between physical phenomena, when interested in engineering approaches, this work also presents a strict mathematical formulation for the classification of the existing knowledge regarding a physical phenomenon.

This work proves that all the potential perceptions of a phenomenon along with specific operations (analogous to summation and multiplication) with proven properties, constitute a vector space. By taking into account that this space is of four dimensions, a non-linear mapping over this vector space has been also defined to describe mathematically the similarity concept. The value of such a mathematical treatment relies on its capability to provide a deep insight on a specific phenomenon, while at the same time assists the researchers to set up an experimentation plan focusing on the knowledge gaps and work towards filling in these gaps properly. Finally, the proposed model is therefore an engineering tool for avoiding repetition of results and managing waste of research effort, in general.

References

[1] R. Giere, *Scientific Perspectivism*, University of Chicago Press, Chicago, USA, 2006.
[2] H. Douglas, *The irreducible complexity of objectivity*, Synthese, **138** (2004), 453–473.
[3] T. S. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, USA, 1962.
[4] P. Feyerabend, Explanation, *Reduction and Empiricism, Scientific Explanation, Space, and Time*, (Minnesota Studies in the Philosophy of Science, Volume III), H. Feigl, G. Maxwell (editors), University of Minneapolis Press, USA, 1962.
[5] P. Kroes, *Structural analogies between physical systems*, Brit. J. Philos. Sci., **40** (1989), 145-154.
[6] I. Hacking, *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science*, Cambridge University Press, UK, 1983.
[7] S.G. Sterrett, *Models of machines and models of phenomena*, Stud. Philos. Sci., **20** (2006), 69-80.
[8] C. Glymour, *On some patterns of reduction*, Philos. Sci., **37** (1970), 340-353.
[9] S.G. Sterrett, *Physical models and fundamental laws: Using one piece of the world to tell about another*, Mind Soc., **3** (2002), 51-66.
[10] I. Kant, *The Critique of Pure Reason* (Translated by J. M. D. Meiklejoh), University of Adelaide Press, Adelaide, AUS, 2014.
[11] A.S. Troelstra, H. Schwichtenberg, *Basic Proof Theory*, Cambridge University Press, Cambridge, UK, 2000.
[12] K.R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Harper, New York, USA, 1963.
[13] A. Kanavouras, F.A. Coutelieris, *Systematic transition from description to a prediction engineering model for the oxidation in packed edible oils*, J. Food Chem., **229** (2017), 820-827.
[14] R.B. Bird, W.E. Stewart, E.N. Lightfoot, *Transport Phenomena*, Wiley, New York, USA, 1960.

Journal of Mathematical Sciences and Modelling, 3 (2) (2020) 55-75
Research Article

Journal of Mathematical Sciences and Modelling

Journal Homepage: www.dergipark.gov.tr/jmsm
ISSN 2636-8692
DOI: http://dx.doi.org/10.33187/jmsm.669216

Rayleigh-Quotient Representation of the Real Parts, Imaginary Parts, and Moduli of the Eigenvalues of General Matrices

Ludwig Kohaupt[1]

[1]*Department of Mathematics, Beuth University of Technology Berlin, Berlin, Germany*

Article Info

Keywords: *Asymptotic stability of dynamical systems, Circular damped eigenfrequencies, Moduli of eigenvalues, Rayleigh quotient, Real and imaginary parts of eigenvalues, Weighted norm*
2010 AMS: *11E39, 15A18, 15B57, 15B99, 65F35, 65J05*
Received: *8 January 2020*
Accepted: *27 July 2020*
Available online: *31 August 2020*

Abstract

In the present paper, formulas for the Rayleigh-quotient representation of the real parts, imaginary parts, and moduli of the eigenvalues of general matrices are obtained that resemble corresponding formulas for the eigenvalues of self-adjoint and diagonalizable matrices. These formulas are of interest in Linear Algebra and in the theory of linear dynamical systems. The key point is that a weighted scalar product is used that is defined by means of a special positive definite matrix. As applications, one obtains convexity properties of newly-defined numerical ranges of a matrix. A numerical example underpins the theoretical findings.

1. Introduction

For self-adjoint matrices, there are formulas for the eigenvalues in the form of Rayleigh quotients; more precisely, max-, min-, min-max-, and max-min-formulas are known; for this, see, e.g., the book [1, Section 5.4]. Recently, the author has carried over these formulas to the real parts, imaginary parts, and moduli of diagonalizable matrices. The aim of the present paper is to extend these results to general matrices. We mention also that the presentation of this paper parallels that of [2]. So, similarities in the formulation do not happen by accident, but are intended in order to underline the similarities. As a consequence, many verbatim passages in the formulations are taken from there.

As it has already been said in [2], first, the obtained formulas are of interest on their own in Linear Algebra. Second, these are also of potential interest, for example, in the theory of linear dynamical systems. The reason for this is as follows. The real parts of the eigenvalues multiplied by the time are equal to the arguments of the exponential functions that describe the decay behavior of the solution (see, e.g., [3, Section 7.1, p.2011, Formulas (89), (90)]). Further, the system is asymptotically stable if the real parts of all eigenvalues are negative. Moreover, when the eigenvalues are pairwise conjugate-complex, then the moduli of the imaginary parts are the circular damped eigenfrequencies of the system (see, e.g., [3, Section 7.1, p. 2011, (89)]). Third, the paper could be of interest in graduate/undergraduate teaching or research at college level since its style is expository and since its subject can be seen as a supplement of the curriculum in Linear Algebra and Numerical Analysis.

The paper is structured as follows. In Section 2, preliminary materials are assembled on biorthogonality relations for the principal vectors of a general matrix A and the principal vectors of A^* that will be useful in the sequel. Moreover, the construction of positive semi-definite matrices R_j and of the positive definite matrix $R = \sum_{j=1}^{n} R_j$ is reviewed where the last one is employed to define a weighted scalar product $(\cdot,\cdot)_R$ that plays a key role in deriving the new results. In Sections 3, 4, and 5, formulas for the Rayleigh-quotient representation of the real parts, imaginary parts, and moduli of the eigenvalues of a general matrix are given, as the case may be. In Section 6, a connection between the matrices $R^{-1}\frac{A^*R+RA}{2}$, $R^{-1}\frac{RA-A^*R}{2i}$, and $R^{-1}A^*RA$ is established that play a key role in the study of the real parts, imaginary parts, and moduli of the eigenvalues of A, respectively. Section 7 describes the applications and, in Section 8, we give a numerical example. Finally, Section 9 contains the conclusions. The non-cited references [4], [5], [6], [7], [8], [9], [10], and [11] are given because they may be useful to the reader in the context of the present paper.

***Email addresses and ORCID numbers*: kohaupt@bht-berlin.de, https://orcid.org/0000-0003-1781-0600 (L. Kohaupt)**

2. Preliminaries

As a preparation to Theorem 2.1, we formulate the following *conditions*:

($C1'$) $A \in \mathbf{C}^{n\times n}$
($C2'$) $\lambda_i,\ i=1,\cdots,r$ are the eigenvalues of A corresponding to the Jordan blocks $J_i(\lambda_i)\in \mathbf{C}^{m_i\times m_i},\ i=1,\cdots,r$ with the chains of principal vectors $p_1^{(i)},\cdots,p_{m_i}^{(i)},\ i=1,\cdots,r$
($C3'$) $u_1^{(i)^*},\cdots,u_{m_i}^{(i)^*},\ i=1,\cdots,r$ are the principal vectors of A^* corresponding to the eigenvalues $\overline{\lambda_i},\ i=1,\cdots,r$ of the Jordan blocks $J_i(\overline{\lambda_i})\in \mathbf{C}^{m_i\times m_i},\ i=1,\cdots,r$
($C4'$) $\lambda_i\neq\lambda_j,\ i\neq j,\ i,j=1,\cdots,r$

One has the following theorem.

Theorem 2.1. (Biorthogonality relations for principal vectors)
Let the conditions ($C1'$)-($C4'$) be fulfilled. Then, the systems $\{p_1^{(1)},\cdots,p_{m_1}^{(1)};\cdots;p_1^{(r)},\cdots,p_{m_r}^{(r)}\}$ *and* $\{u_1^{(1)^*},\cdots,u_{m_1}^{(1)^*};\cdots;u_1^{(r)^*},\cdots,u_{m_r}^{(r)^*}\}$ *can be constructed such that the following biorthogonality relations hold:*

$$(p_k^{(i)},u_l^{(i)^*})=\begin{cases}1, & l=m_i-k+1\\ 0, & l\neq m_i-k+1\end{cases}$$

$k=1,\cdots,m_i,\ i=1,\cdots,r$ *and*

$$(p_k^{(i)},u_l^{(j)^*})=0,\ i\neq j,$$

$k=1,\cdots,m_i,\ l=1,\cdots,m_j,\ i,j=1,\cdots,r.$
So, with

$$v_l^{(i)^*}:={u_{m_i-l+1}^{(i)}}^*,$$

$l=1,\cdots,m_i,\ i=1,\cdots,r$ *one has the biorthogonality relations*

$$(p_k^{(i)},v_l^{(i)^*})=\delta_{kl},$$

$k,l=1,\cdots,m_i,\ i=1,\cdots,r,$ *and*

$$(p_k^{(i)},v_l^{(j)^*})=0,\ i\neq j,$$

$k=1,\cdots,m_i,\ l=1,\cdots,m_j,\ i,j=1,\cdots,r.$

Proof. See proof of [12, Theorem 2]. □

Remark 2.2. *The hypothesis $\lambda_i\neq\lambda_j,\ i\neq j,\ i,j=1,\cdots,r$ can be omitted, see [12, Theorem 4]. But, since in our example this condition is fulfilled, we preserve it.*

Theorem 2.3. (Construction of positive definite matrix R)
Let the conditions ($C1'$)-($C4'$) be fulfilled. Let $\alpha_j=\lambda_j(A)$ be the eigenvalues and $u_1^{(j)},\ldots,u_{m_j}^{(j)}$ be a chain of associated left principal vectors for $j=1,\ldots,r$. Further, let $A^*\in\mathbf{C}^{n\times n}$ be the adjoint matrix of A so that $u_1^{(j)^*},\ldots,u_{m_j}^{(j)^*}$ is a chain of right principal vectors corresponding to the eigenvalues $\overline{\alpha_j}=\lambda_j(A^*)$ for $j=1,\cdots,r$, i.e.,

$$u_k^{(j)}A=\alpha_j u_k^{(j)}+u_{k-1}^{(j)}$$

with $u_0^{(j)}=0,\ k=1,\ldots,m_j;\ j=1,\ldots,r$ *and*

$$A^*u_k^{(j)^*}=\overline{\alpha_j}u_k^{(j)^*}+{u_{k-1}^{(j)}}^*$$

with $u_0^{(j)^*}=0, j=1,\ldots,r.$
Let

$$\rho_j=\overline{\alpha_j}+\alpha_j=2Re\,\alpha_j=2Re\,\overline{\alpha_j},\ j=1,\cdots,r,$$

$$\sigma_j=\alpha_j-\overline{\alpha_j}=2\,i\,Im\,\alpha_j,\ j=1,\cdots,r,$$

and

$$R_j^{(k,k)}:=u_k^{(j)^*}u_k^{(j)},\ k=1,\cdots,m_j,\ j=1,\cdots,r.$$

Then,

$$A^*R_j^{(1,1)}+R_j^{(1,1)}A=\rho_j R_j^{(1,1)},\ j=1,\cdots,r,$$

$$R_j^{(1,1)}A - A^*R_j^{(1,1)} = \sigma_j R_j^{(1,1)},\ j = 1,\cdots,r.$$

In other word, the matrix eigenvalue problem

$$A^*V + VA = \mu V$$

has the r solution pairs

$$(\mu, V) = (\rho_j, R_j^{(1,1)})$$

with real ρ_j, and the matrix eigenvalue problem

$$VA - A^*V = \mu V$$

has the r solution pairs

$$(\mu, V) = (\sigma_j, R_j^{(1,1)})$$

with purely imaginary σ_j.
The matrices $R_j^{(k,k)} \in \mathbf{C}^{n\times n}$, $k = 1,\ldots,m_j$, $j = 1,\ldots,r$ are positive semi-definite. Further,

$$R := \sum_{j=1}^{r} R_j = \sum_{j=1}^{r}\sum_{k=1}^{m_j} R_j^{(k,k)} \tag{2.1}$$

is positive definite.

Proof. See [13, Theorem 2]. □

Remark 2.4. *Since R in (2.1) is positive definite, by*

$$(u,v)_R := (Ru,v),\ u,\ v \in \mathbf{C}^n,$$

a weighted scalar product $(\cdot,\cdot)_R$ *is defined and by*

$$\|u\|_R := (Ru,u)^{\frac{1}{2}},\ u \in \mathbf{C}^n,$$

a weighted norm $\|\cdot\|_R$.

3. Formulas for the representation of the real parts of the eigenvalues of a general matrix

In this section, we want to derive formulas for the representation of the real parts of the eigenvalues of a general matrix A by Rayleigh quotients. More precisely, max-, min-, min-max-, and max-min-representations are obtained corresponding to associated formulas for the eigenvalues of diagonalizable matrices in [2] or to the eigenvalues of self-adjoint matrices, assembled, for instance, in the book of [1, Section 5.4].
First, we derive a result similar to that of [2, Lemma 3.1]. For this, with the identity matrix E, we introduce the abbreviation

$$N_{\lambda_j(A)} := \{u \in \mathbf{C}^n \,|\, (A - \lambda_j(A)E)u = 0\},\ j = 1,\ldots,r$$

for the geometric eigenspaces so that

$$N_{\lambda_j(A)} = [p_1^{(j)}] = [p_j],\ j = 1,\ldots,r.$$

Herewith, we define

$$N_{\sigma(A)} := \bigoplus_{j=1}^{r} N_{\lambda_j(A)}.$$

We have the following lemma.

Lemma 3.1. *Let the conditions (C1′)-(C4′) be fulfilled and R be defined by (2.1).*
Then, with the denotations of Theorem 2.3,

$$(Au,u)_R = \sum_{j=1}^{r} \lambda_j(A)\,(u,u)_{R_j} + \sum_{j=1}^{r}\sum_{k=2}^{m_j} (u, v_k^{(j)^*})\overline{(u, v_{k-1}^{(j)^*})},\ u \in \mathbf{C}^n \tag{3.1}$$

leading to

$$(Au,u)_R = \sum_{j=1}^{r} \lambda_j(A)\,(u,u)_{R_j},\ u \in N_{\sigma(A)} \tag{3.2}$$

and thus to

$$Re(Au,u)_R = \sum_{j=1}^{r} Re\lambda_j(A)\,(u,u)_{R_j},\ u \in N_{\sigma(A)} \tag{3.3}$$

where

$$R_j = \sum_{k=1}^{m_j} R_j^{(k,k)} = \sum_{k=1}^{m_j} u_k^{(j)^*} u_k^{(j)} = \sum_{k=1}^{m_j} v_k^{(j)^*} v_k^{(j)},$$

$j = 1, \ldots, r$.
If matrix A is, beyond this, asymptotically stable, i.e., if

$$Re\,\lambda_j(A) < 0,\ j = 1, \cdots, r,$$

then

$$Re\,(Au,u)_R = -\sum_{j=1}^{r} |Re\,\lambda_j(A)|\,(u,u)_{R_j},\ u \in N_{\sigma(A)},$$

so that, in this case,

$$Re\,(Au,u)_R < 0,\ 0 \neq u \in N_{\sigma(A)} \tag{3.4}$$

and

$$|Re\,(Au,u)_R| = \sum_{j=1}^{r} |Re\,\lambda_j(A)|\,(u,u)_{R_j},\ u \in N_{\sigma(A)}. \tag{3.5}$$

Proof. First, we prove (3.1). For this, let $u \in C^n$. Then with the denotations of Theorem 2.1,

$$u = \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, p_k^{(j)}$$

leading to

$$\begin{aligned} Au &= \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) A p_k^{(j)} \\ &= \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) [\lambda_j p_k^{(j)} + p_{k-1}^{(j)}] \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, p_k^{(j)} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, p_{k-1}^{(j)} \end{aligned}$$

since $p_0^{(j)} = 0$, $j = 1, \ldots, r$. This implies

$$\begin{aligned} (Au, Ru) &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, (p_k^{(j)}, Ru) + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, (p_{k-1}^{(j)}, Ru) \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, (Rp_k^{(j)}, u) + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, (Rp_{k-1}^{(j)}, u). \end{aligned}$$

Now,

$$\begin{aligned} Rp_{k-1}^{(j)} &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} R_l^{(s,s)} p_{k-1}^{(j)} = \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*} v_s^{(l)}\, p_{k-1}^{(j)} \\ &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*} (p_{k-1}^{(j)}, v_s^{(l)^*}) \\ &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*}\, \delta_{jl}\, \delta_{s,k-1} = \sum_{s=1}^{m_j} v_s^{(j)^*}\, \delta_{s,k-1} = v_{k-1}^{(j)^*}. \end{aligned}$$

This leads to

$$\begin{aligned} (Au, Ru) &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, (p_k^{(j)}, Ru) + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, (v_{k-1}^{(j)^*}, u) \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, (p_k^{(j)}, Ru) + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, \overline{(u, v_{k-1}^{(j)^*})}. \end{aligned}$$

Further,

$$(p_{k-1}^{(j)}, Ru) = (Rp_{k-1}^{(j)}, u)$$

and, as before,

$$Rp_{k-1}^{(j)} = v_{k-1}^{(j)\,*} = R_j p_{k-1}^{(j)}.$$

Therefore,

$$(p_k^{(j)}, Ru) = (Rp_k^{(j)}, u) = (v_k^{(j)*}, u)$$

and thus

$$\begin{aligned}\sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)*})\,(p_k^{(j)}, Ru) &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)*})\,(v_k^{(j)*}, u) \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} |(u, v_k^{(j)*})| = \sum_{j=1}^{r} \lambda_j (u,u)_{R_j}\end{aligned}$$

since

$$(u,u)_{R_j} = (R_j u, u) = \sum_{k=1}^{m_j} (R_j^{(k,k)} u, u) = \sum_{k=1}^{m_j} (v_k^{(j)*} v_k u, u) = \sum_{k=1}^{m_j} (v_k^{(j)} u, v_k^{(j)} u) = \sum_{k=1}^{m_j} |(u, v_k^{(j)*})|.$$

So, we obtain (3.1).
In order to get (3.2), i.e.,

$$(Au, Ru) = \sum_{j=1}^{r} \lambda_j(A)(u,u)_{R_j},$$

we must have

$$\sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)*}) \overline{(u, v_{k-1}^{(j)\,*})} = 0.$$

Sufficient for this is

$$(u, v_k^{(j)*}) = 0, \quad k = 2, \ldots, m_j,\ j = 1, \ldots, r,$$

for example,

$$u \in \bigoplus_{j=1}^{r} [p_1^{(j)}] = \bigoplus_{j=1}^{r} N_{\lambda_j(A)} = N_{\sigma(A)}.$$

The rest of the proof is clear. □

Remark 3.2. *We have shown that*

$$N_{\sigma(A)} \subset \{u \in \mathbb{C}^n \mid \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)*}) \overline{(u, v_{k-1}^{(j)\,*})} = 0\} =: N.$$

But, the set on the right-hand side can be larger than the set $N_{\sigma(A)}$. For, if $m_s > 1$ for some $s \in \{1, \ldots, r\}$, then

$$p_{m_s}^{(s)} \in N,$$

even though $p_{m_s}^{(s)} \notin N_{\sigma(A)}$. Moreover, we have even for all single $p_j^{(k)}$ $j = 1, \cdots, m_k, k = 1, \cdots, r$ the relations

$$p_j^{(k)} \in N.$$

Remark 3.3. *If condition (C4′) is not fulfilled, then the results of Lemma 3.1 remain valid if its formulation is adapted to [14, Theorem 4]. The details are left to the reader.*

Next, we have the following lemma.

Lemma 3.4. *Let the conditions (C1′)-(C4′) be fulfilled and R be defined by (2.1). Further, let matrix A be asymptotically stable. Then, $A^*R + RA$ is negative definite on $N_{\sigma(A)}$.*

Proof. With Lemma 3.1, Formula (3.3), and $\rho_j = 2Re\,\lambda_j(A),\ j = 1,\cdots,n$, we obtain

$$\begin{aligned}(-[A^*R+RA]u,u) &= \sum_{j=1}^{n}(-\rho_j)\,(u,u)_{R_j} = 2\sum_{j=1}^{r} Re\,(-\lambda_j(A))\,(u,u)_{R_j}\\ &= 2\sum_{j=1}^{r}|Re\,\lambda_j(A)|\,(u,u)_{R_j}\\ &\geq 2\min_{j=1,\cdots,r}|Re\,\lambda_j(A)|\sum_{j=1}^{n}(u,u)_{R_j}\\ &= c_0(Ru,u) > 0,\ 0\neq u\in N_{\sigma(A)}\end{aligned}$$

with $c_0 = 2\min\limits_{j=1,\cdots,r}|Re\,\lambda_j(A)| > 0$. □

Similarly as in [2, Formula (3.6)], we define the following vector spaces.

$$\begin{aligned}&M_{1,N_{\sigma(A)}} := N_{\sigma(A)},\\ &M_{k,N_{\sigma(A)}} := \{u\in N_{\sigma(A)}\,|\,(u,u)_{R_i} = 0,\ i = 1,2,\cdots,k-1\},\ k = 2,\cdots,r.\end{aligned} \tag{3.6}$$

The next lemma characterizes these spaces.

Lemma 3.5. *Let the conditions (C1′)-(C4′) be fulfilled.*
Then,

$$M_{k,N_{\sigma(A)}} = [p_k,\ldots,p_r],\ k = 1,2,\ldots,r. \tag{3.7}$$

Proof. The proof is done for $k = 3$. The general case can be made by induction. So, we have to prove

$$M_{3,N_{\sigma(A)}} = \{u\in N_{\sigma(A)}\,|\,(u,u)_{R_1} = 0,\,(u,u)_{R_2} = 0\} = [p_3,\ldots,p_r].$$

(i) $\underline{[p_3,p_4,\cdots,p_r]\subset M_{3,N_{\sigma(A)}}}$:

Let $u\in[p_3,p_4,\cdots,p_r]$. Then, $u = \sum_{j=3}^{r}\beta_j p_j$ with elements $\beta_j\in C$, $j = 3,\cdots,r$. Let $s\in\{1,2\}$. This entails, due to Theorem 2.3, Lemma 3.1, and $(p_j,p_k)_{R_s} = (p_j,p_k)_{R_s^{(1,1)}}$,

$$\begin{aligned}(u,u)_{R_s} &= \sum_{j,k=3}^{r}\beta_j\overline{\beta_k}(p_j,p_k)_{R_s} = \sum_{j,k=3}^{r}\beta_j\overline{\beta_k}(p_j,p_k)_{R_s^{(1,1)}}\\ &= \sum_{j,k=3}^{r}\beta_j\overline{\beta_k}(R_s^{(1,1)}p_1^{(j)},p_1^{(k)}) = \sum_{j,k=3}^{r}\beta_j\overline{\beta_k}(v_1^{(s)*}v_1^{(s)}p_1^{(j)},p_1^{(k)})\\ &= \sum_{j,k=3}^{r}\beta_j\overline{\beta_k}(v_1^{(s)*}\underbrace{(p_1^{(j)},v_1^{(s)*})}_{\delta_{js}=0},p_1^{(k)}) = 0,\end{aligned}$$

$j\in\{3,\ldots,r\}$, $s\in\{1,2\}$. Therefore, $(u,u)_{R_s} = 0$, $s = 1,2$ and thus $u\in M_{3,N_{\sigma(A)}}$ so that $[p_3,p_4,\cdots,p_r]\subset M_{3,N_{\sigma(A)}}$ is proven.

(ii) $\underline{M_{3,N_{\sigma(A)}}\subset[p_3,p_4,\cdots,p_r]}$:

Let $u\in M_{3,N_{\sigma(A)}}$. This implies $u\in N_{\sigma(A)}$ with $(u,u)_{R_s} = 0$, $s = 1,2$ or

$$\begin{aligned}(R_s u,u) &= \textstyle\sum_{l=1}^{m_s}(R_s^{(l,l)}u,u) = \sum_{l=1}^{m_s}(v_l^{(s)*}v_l^{(s)}u,u) = \sum_{l=1}^{m_s}(v_l^{(s)*}(u,v_l^{(s)*}),u)\\ &= \textstyle\sum_{l=1}^{m_s}(v_l^{(s)*},u)\overline{(v_l^{(s)*},u)} = \sum_{l=1}^{m_s}|(v_l^{(s)*},u)|^2 = 0,\ s = 1,2,\end{aligned}$$

that is, in particular,

$$u\in N_{\sigma(A)}\quad\text{with}\quad(u,v_1^{(s)*}) = (u,v_s^*) = 0,\ s = 1,2. \tag{3.8}$$

Since $u\in N_{\sigma(A)}$, we have,

$$u = \sum_{j=1}^{r}(u,v_j^*)p_j$$

so that with (3.8),

$$u = \sum_{j=1}^{r}(u,v_j^*)p_j = \sum_{j=3}^{r}(u,v_j^*)p_j\in[p_3,\ldots,p_r].$$

This completes the proof of the assertion. □

Similarly to [2], we suppose that the eigenvalues $\lambda_1(A),\cdots,\lambda_r(A)$ of matrix A are arranged such that

$$Re\,\lambda_1(A) \geq Re\,\lambda_2(A) \geq \cdots \geq Re\,\lambda_r(A). \tag{3.9}$$

If A is asymptotically stable, (3.9) is replaced by

$$|Re\,\lambda_1(A)| \geq |Re\,\lambda_2(A)| \geq \cdots \geq |Re\,\lambda_r(A)|. \tag{3.10}$$

One has the following theorem.

Theorem 3.6. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (3.9). Moreover, let the vector spaces $M_{k,N_{\sigma(A)}}$, $k=1,\cdots,r$ be defined by (3.6) or (3.7).*
Then,

$$Re\,\lambda_k(A) = \max_{0\neq u\in M_{k,N_{\sigma(A)}}} \frac{Re(Au,u)_R}{(u,u)_R},\ k=1,2,\cdots,r. \tag{3.11}$$

If matrix A is, beyond this, asymptotically stable, and if the eigenvalues are arranged according to (3.10), then also

$$|Re\,\lambda_k(A)| = \max_{0\neq u\in M_{k,N_{\sigma(A)}}} \frac{|Re(Au,u)_R|}{(u,u)_R},\ k=1,2,\cdots,r. \tag{3.12}$$

The maximum is attained for $u=p_k$.

Proof. According to (3.3), one has

$$Re(Au,u)_R = \sum_{j=1}^{r} Re\,\lambda_j(A)(R_j u,u),\ u\in N_{\sigma(A)}.$$

Choosing $k\in\{1,\cdots,r\}$ fixed and $u\in M_{k,N_{\sigma(A)}}$, using (3.6), one obtains

$$\begin{aligned} Re(Au,u)_R &= \sum_{j=k}^{r} Re\,\lambda_j(A)(R_j u,u) \leq \max_{j=k,\cdots,r} Re\,\lambda_j(A) \sum_{j=k}^{r}(R_j u,u) \\ &= Re\,\lambda_k(A)\sum_{j=1}^{r}(R_j u,u) = Re\,\lambda_k(A)\,(u,u)_R, \end{aligned}$$

that is,

$$\frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_k(A),\ 0\neq u\in M_{k,N_{\sigma(A)}}$$

and thus

$$\max_{0\neq u\in M_{k,N_{\sigma(A)}}} \frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_k(A).$$

Now, the maximum is attained for $u=p_k\in M_{k,N_{\sigma(A)}}$, that is,

$$Re\,\lambda_k(A) = \frac{Re(Ap_k,p_k)_R}{(p_k,p_k)_R} \leq \max_{0\neq u\in M_{k,N_{\sigma(A)}}} \frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_k(A)$$

so that the assertion (3.11) is proven.
Relation (3.12) is proven in the same way as (3.11), but based on (3.5) instead of (3.3) and (3.10) instead of (3.9). □

Remark 3.7. *As we have seen, the proof is similar to that of [2, Theorem 3.4]. The essential difference is that the full space $\boldsymbol{C}^n$ is replaced by the geometric eigenspace $N_{\sigma(A)}\subset\boldsymbol{C}^n$. Therefore, in the sequel, we state the results without proofs.*

Next, we want to state a min-max characterization for the real parts of eigenvalues similar to results for diagonalizable matrices in [2, Theorem 3.5].
One has the following theorem.

Theorem 3.8. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (3.9).*
Then, for every $j=1,\cdots,r$ and every subspace $M\subset N_{\sigma(A)}$ with $dim\,M=m=r+1-j$, the following inequalities are valid:

$$Re\,\lambda_j(A) \leq \max_{0\neq u\in M} \frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_1(A), \tag{3.13}$$

and the following representation formulas hold:

$$Re\,\lambda_j(A) = \min_{dim M=m} \max_{0\neq u\in M\subset N_{\sigma(A)}} \frac{Re(Au,u)_R}{(u,u)_R}.$$

If matrix A is, beyond this, asymptotically stable and the eigenvalues are arranged according to (3.10), then also

$$|Re\,\lambda_j(A)| = \min_{dim M=m} \max_{0\neq u\in M\subset N_{\sigma(A)}} \frac{|Re(Au,u)_R|}{(u,u)_R}.$$

Remark 3.9. *From (3.13), it follows*

$$\frac{Re\,(Au,u)_R}{(u,u)_R} \leq \nu[A] = \max_{j=1,\cdots,r} Re\,\lambda_j(A),\ 0 \neq u \in N_{\sigma(A)}.$$

For the following theorem, we need the vector spaces N_k defined by

$$N_k := [p_1, p_2, \cdots, p_k],\ k = 1,2,\cdots,r. \tag{3.14}$$

Then, we have a result similar to that of Theorem 3.6.

Theorem 3.10. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (3.9). Moreover, let the vector spaces N_k, $k = 1,\cdots,r$ be defined by (3.14).*
Then,

$$Re\,\lambda_k(A) = \min_{0\neq u\in N_k} \frac{Re(Au,u)_R}{(u,u)_R},\ k = 1,2,\cdots,r.$$

If matrix A is, beyond this, asymptotically stable and if the eigenvalues are arranged according to (3.10), then also

$$|Re\,\lambda_k(A)| = \min_{0\neq u\in N_k} \frac{|Re(Au,u)_R|}{(u,u)_R},\ k = 1,2,\cdots,r.$$

The minimum is attained for $u = p_k$.

Next, we want to derive a max-min characterization for the real parts of eigenvalues similar to results for diagonalizable matrices in [2]. One has the following theorem.

Theorem 3.11. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (3.9). Then, for every $j = 1,\cdots,r$ and every subspace $N \subset N_{\sigma(A)}$ with $dim N = j$, the following inequalities are valid:*

$$Re\,\lambda_r(A) \leq \min_{0\neq u\in N} \frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_j(A), \tag{3.15}$$

and the following representation formulas hold:

$$Re\,\lambda_j(A) = \max_{dim N=j}\ \min_{0\neq u\in N} \frac{Re(Au,u)_R}{(u,u)_R}.$$

If matrix A is, beyond this, asymptotically stable and the eigenvalues are arranged according to (3.10), then also

$$|Re\,\lambda_j(A)| = \max_{dim N=j}\ \min_{0\neq u\in N} \frac{|Re(Au,u)_R|}{(u,u)_R}.$$

Remark 3.12. *From (3.15), it follows*

$$\frac{Re\,(Au,u)_R}{(u,u)_R} \geq -\nu[-A] = \min_{j=1,\cdots,r} Re\,\lambda_j(A),\ 0 \neq u \in N_{\sigma(A)}.$$

4. Formulas for the representation of the imaginary parts of the eigenvalues of a general matrix

In this section, we want to state formulas for the representation of the imaginary parts of the eigenvalues of a general matrix A by Rayleigh quotients. More precisely, max-, min-, min-max-, and max-min-representations are obtained corresponding to associated formulas for the eigenvalues of self-adjoint matrices in the textbook [1, Section 5.4] resp. corresponding to those for the imaginary parts of eigenvalues of diagonalizable matrices in [2].
First, we state a result similar to that of [1, Section 5.4 (18)]. This is done in the following Formula (4.1).

Lemma 4.1. *Let the conditions (C1′)-(C4′) be fulfilled and R be defined by (2.1).*
Then, with the denotations of Theorem 2.3,

$$Im(A\,u,u)_R = \sum_{j=1}^{r} Im\lambda_j(A)\,(u,u)_{R_j},\ u \in N_{\sigma(A)}. \tag{4.1}$$

Proof. The proof follows immediately from (3.2). □

Similarly to [1, Section 5.4 (22)] or (3.9), we suppose that the eigenvalues $\lambda_1(A),\cdots,\lambda_r(A)$ of matrix A are arranged such that

$$Im\,\lambda_1(A) \geq Im\,\lambda_2(A) \geq \cdots \geq Im\,\lambda_r(A). \tag{4.2}$$

One has the following theorem.

Theorem 4.2. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (4.2). Moreover, let the vector spaces $M_{k,N_{\sigma(A)}}$, $k = 1,\cdots,r$ be defined by (3.6) or (3.7).*
Then,

$$Im\,\lambda_k(A) = \max_{0\neq u\in M_{k,N_{\sigma(A)}}} \frac{Im(Au,u)_R}{(u,u)_R},\ k = 1,2,\cdots,r.$$

The maximum is attained for $u = p_k$.

Next, we want to state a min-max characterization for the imaginary parts of eigenvalues that corresponds to results for diagonalizable matrices in [2] or that corresponds to results for the real parts of eigenvalues of general matrices in Section 3.
One has the following theorem.

Theorem 4.3. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (4.2). Then, for every $j = 1, \cdots, r$ and every subspace $M \subset N_{\sigma(A)}$ with $dim M = m = r + 1 - j$, the following inequalities are valid:*

$$Im\,\lambda_j(A) \leq \max_{0 \neq u \in M} \frac{Im(Au,u)_R}{(u,u)_R} \leq Im\,\lambda_1(A), \tag{4.3}$$

and the following representation formulas hold:

$$Im\,\lambda_j(A) = \min_{dim M = m} \max_{0 \neq u \in M \subset N_{\sigma(A)}} \frac{Im(Au,u)_R}{(u,u)_R}.$$

Remark 4.4. *From (4.3), it follows*

$$\frac{Im\,(Au,u)_R}{(u,u)_R} \leq \max_{j=1,\cdots,r} Im\,\lambda_j(A),\ 0 \neq u \in N_{\sigma(A)}.$$

With the vector spaces N_k, we have the following theorem.

Theorem 4.5. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (4.2). Moreover, let the vector spaces N_k, $k = 1, \cdots, r$ be defined by (3.14).*
Then,

$$Im\,\lambda_k(A) = \min_{0 \neq u \in N_k} \frac{Im(Au,u)_R}{(u,u)_R},\ k = 1, 2, \cdots, r.$$

The minimum is attained for $u = p_k$.

Next, we state a max-min characterization for the imaginary parts of eigenvalues for general matrices similar to results for the real parts in Section 3.
One has the following theorem.

Theorem 4.6. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (4.2). Then, for every $j = 1, \cdots, r$ and every subspace $N \subset N_{\sigma(A)}$ with $dim N = j$, the following inequalities are valid:*

$$Im\,\lambda_r(A) \leq \min_{0 \neq u \in N} \frac{Im(Au,u)_R}{(u,u)_R} \leq Im\,\lambda_j(A), \tag{4.4}$$

and the following representation formulas hold:

$$Im\,\lambda_j(A) = \max_{dim N = j} \min_{0 \neq u \in N} \frac{Im(Au,u)_R}{(u,u)_R}.$$

Remark 4.7. *From (4.4), it follows*

$$\frac{Im\,(Au,u)_R}{(u,u)_R} \geq \min_{j=1,\cdots,r} Im\,\lambda_j(A),\ 0 \neq u \in N_{\sigma(A)}.$$

5. Formulas for the representation of the moduli of the eigenvalues of a general matrix

In this section, we want to derive formulas for the representation of the moduli of the eigenvalues of a general matrix A by Rayleigh quotients. More precisely, max-, min-, min-max-, and max-min-representations are obtained corresponding to associated formulas for the eigenvalues of diagonalizable matrices in [2] and for the real and imaginary parts of eigenvalues of general matrices in Sections 3 and 4.
First, we derive a result similar to that of [2, Lemma 5.1].

Lemma 5.1. *Let the conditions (C1′)-(C4′) be fulfilled and R be defined by (2.1). Then, with the denotations of Theorem 2.1,*

$$\begin{aligned}
\|Au\|_R^2 &= (RAu, Au) = (A^*RAu, u) = (R^{-1}A^*RAu, u)_R \\
&= \sum_{j=1}^{r} |\lambda_j(A)|^2 (u,u)_{R_j} \\
&+ \sum_{j=1}^{r} \lambda_j(A) \sum_{k=1}^{m_j - 1} (u, v_k^{(j)^*}) \overline{(u, v_{k+1}^{(j)^*})} \\
&+ \sum_{j=1}^{r} \overline{\lambda_j(A)} \sum_{k=2}^{m_j} (u, v_k^{(j)^*}) \overline{(u, v_{k-1}^{(j)^*})} \\
&+ \sum_{j=1}^{r} \sum_{k=2}^{m_j} |(u, v_k^{(j)^*})|^2,\ u \in \boldsymbol{C}^n
\end{aligned} \tag{5.1}$$

leading to

$$\begin{aligned} \|Au\|_R^2 &= (RAu,Au) = (A^*RAu,u) = (R^{-1}A^*RAu,u)_R \\ &= \sum_{j=1}^{r} |\lambda_j(A)|^2 (u,u)_{R_j},\ u \in N_{\sigma(A)} \end{aligned} \tag{5.2}$$

where

$$(u,u)_{R_j} = (u,u)_{R_j^{(1,1)}},\ u \in N_{\sigma(A)},$$

$j = 1,\ldots,r$.

Proof. First, we prove (5.1). For this, let $u \in C^n$. Then with the denotations of Theorem 2.1,

$$u = \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, p_k^{(j)} \tag{5.3}$$

leading to

$$\begin{aligned} Au &= \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) A p_k^{(j)} \\ &= \sum_{j=1}^{r} \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) [\lambda_j p_k^{(j)} + p_{k-1}^{(j)}] \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, p_k^{(j)} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, p_{k-1}^{(j)} \end{aligned}$$

since $p_0^{(j)} = 0$, $j = 1,\ldots,r$. This implies

$$RAu = \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, R p_k^{(j)} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, R p_{k-1}^{(j)}.$$

Now,

$$\begin{aligned} Rp_k^{(j)} &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} R_l^{(s,s)} p_k^{(j)} = \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*} v_s^{(l)} p_k^{(j)} \\ &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*} (p_k^{(j)}, v_s^{(l)^*}) \\ &= \sum_{l=1}^{r} \sum_{s=1}^{m_l} v_s^{(l)^*} \delta_{jl}\, \delta_{s,k} = \sum_{s=1}^{m_j} v_s^{(j)^*} \delta_{s,k} = v_k^{(j)^*} \end{aligned}$$

and correspondingly

$$Rp_{k-1}^{(j)} = v_{k-1}^{(j)^*}.$$

This leads to

$$RAu = \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*})\, v_k^{(j)^*} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*})\, v_{k-1}^{(j)^*}. \tag{5.4}$$

Thus,

$$\begin{aligned} A^*RAu &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) A^* v_k^{(j)^*} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*}) A^* v_{k-1}^{(j)^*} \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) A^* u_{m_j-k+1}^{(j)^*} + \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*}) A^* u_{m_j-(k-1)+1}^{(j)^*} \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) [\overline{\lambda_j} u_{m_j-k+1}^{(j)^*} + u_{m_j-k}^{(j)^*}] \\ &+ \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*}) [\overline{\lambda_j} u_{m_j-k+2}^{(j)^*} + u_{m_j-k+1}^{(j)^*}] \\ &= \sum_{j=1}^{r} \lambda_j \sum_{k=1}^{m_j} (u, v_k^{(j)^*}) [\overline{\lambda_j} v_k^{(j)^*} + v_{k+1}^{(j)^*}] \\ &+ \sum_{j=1}^{r} \sum_{k=2}^{m_j} (u, v_k^{(j)^*}) [\overline{\lambda_j} v_{k-1}^{(j)^*} + v_k^{(j)^*}] \end{aligned}$$

and thus

$$\begin{aligned} A^*RAu &= \sum_{j=1}^{r}|\lambda_j|^2\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\,v_k^{(j)^*}+\sum_{j=1}^{r}\lambda_j\sum_{k=1}^{m_j}(u,v_k^{(j)^*})v_{k+1}^{(j)\ ^*} \\ &+ \sum_{j=1}^{r}\overline{\lambda_j}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,v_{k-1}^{(j)\ ^*}+\sum_{j=1}^{r}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,v_k^{(j)^*}. \end{aligned} \tag{5.5}$$

From (5.3) and (5.5), we conclude that the following chain of equations is valid

$$\begin{aligned} (A^*RAu,u) &= (\sum_{j=1}^{r}|\lambda_j|^2\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\,v_k^{(j)^*},\sum_{l=1}^{r}\sum_{s=1}^{m_l}(u,v_s^{(l)^*})\,p_s^{(l)}) \\ &+ (\sum_{j=1}^{r}\lambda_j\sum_{k=1}^{m_j}(u,v_k^{(j)^*})v_{k+1}^{(j)\ ^*},\sum_{l=1}^{r}\sum_{s=1}^{m_l}(u,v_s^{(l)^*})\,p_s^{(l)}) \\ &+ (\sum_{j=1}^{r}\overline{\lambda_j}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,v_{k-1}^{(j)\ ^*},\sum_{l=1}^{r}\sum_{s=1}^{m_l}(u,v_s^{(l)^*})\,p_s^{(l)}) \\ &+ (\sum_{j=1}^{r}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,v_k^{(j)^*},\sum_{l=1}^{r}\sum_{s=1}^{m_l}(u,v_s^{(l)^*})\,p_s^{(l)}) \\ &= \sum_{j=1}^{r}|\lambda_j|^2\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\sum_{l=1}^{r}\sum_{s=1}^{m_l}\overline{(u,v_s^{(l)^*})}\,\underbrace{(v_k^{(j)^*},p_s^{(l)})}_{\delta_{lj}\delta_{sk}} \\ &+ \sum_{j=1}^{r}\lambda_j\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\sum_{l=1}^{r}\sum_{s=1}^{m_l}\overline{(u,v_s^{(l)^*})}\,\underbrace{(v_{k+1}^{(j)\ ^*},p_s^{(l)})}_{\delta_{lj}\delta_{s,k+1}} \\ &+ \sum_{j=1}^{r}\overline{\lambda_j}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\sum_{l=1}^{r}\sum_{s=1}^{m_l}\overline{(u,v_s^{(l)^*})}\,\underbrace{(v_{k-1}^{(j)\ ^*},p_s^{(l)})}_{\delta_{lj}\delta_{s,k-1}} \\ &+ \sum_{j=1}^{r}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\sum_{l=1}^{r}\sum_{s=1}^{m_l}\overline{(u,v_s^{(l)^*})}\,\underbrace{(v_k^{(j)^*},p_s^{(l)})}_{\delta_{lj}\delta_{s,k}} \end{aligned}$$

so that

$$\begin{aligned} (A^*RAu,u) &= \sum_{j=1}^{r}|\lambda_j|^2\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\,\overline{(u,v_k^{(j)^*})} \\ &+ \sum_{j=1}^{r}\lambda_j\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\overline{(u,v_{k+1}^{(j)\ ^*})} \\ &+ \sum_{j=1}^{r}\overline{\lambda_j}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,\overline{(u,v_{k-1}^{(j)\ ^*})} \\ &+ \sum_{j=1}^{r}\sum_{k=2}^{m_j}(u,v_k^{(j)^*})\,\overline{(u,v_k^{(j)^*})}. \end{aligned} \tag{5.6}$$

Now,

$$\begin{aligned} \sum_{k=1}^{m_j}(u,v_k^{(j)^*})\,\overline{(u,v_k^{(j)^*})} &= (u,\sum_{k=1}^{m_j}(u,v_k^{(j)^*})\,v_k^{(j)^*})=(u,\sum_{k=1}^{m_j}v_k^{(j)^*}v_k^{(j)}\,u) \\ &= (u,R_ju)=(u,u)_{R_j}. \end{aligned} \tag{5.7}$$

Further, for $k=m_j$,

$$v_{k+1}^{(j)\ ^*}=u_{m_j-(k+1)+1}^{(j)\ ^*}=v_{m_j-(m_j+1)+1}^{(j)\ ^*}=u_0^{(j)^*}=0. \tag{5.8}$$

With (5.4)-(5.8), relation (5.1) follows. The rest is clear. □

Similarly to [2], we suppose that the eigenvalues $\lambda_1(A),\cdots,\lambda_r(A)$ of matrix A are arranged such that

$$|\lambda_1(A)|\geq|\lambda_2(A)|\geq\cdots\geq|\lambda_r(A)|. \tag{5.9}$$

One has the following theorem.

Theorem 5.2. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9). Moreover, let the vector spaces $M_{k,N_{\sigma(A)}}$, $k=1,\cdots,r$ be defined by (3.6) or (3.7).*
Then,

$$|\lambda_k(A)|=\max_{0\neq u\in M_{k,N_{\sigma(A)}}}\frac{\|Au\|_R}{\|u\|_R},\ k=1,2,\cdots,r.$$

The maximum is attained for $u=p_k$.

Proof. For the proof, one uses (5.2) and proceeds as in the proof of [2] with the only difference that the full space C^n is replaced by the subspace $N_{\sigma(A)}$. □

In the same way, one obtains the following results.

Theorem 5.3. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9). Then, for every $j = 1, \cdots, r$ and every subspace $M \subset N_{\sigma(A)}$ with $dim M = m = r + 1 - j$, the following inequalities are valid:*

$$|\lambda_j(A)| \leq \max_{0 \neq u \in M} \frac{\|Au\|_R}{\|u\|_R} \leq |\lambda_1(A)|, \tag{5.10}$$

and the following representation formulas hold:

$$|\lambda_j(A)| = \min_{dim M = m} \max_{0 \neq u \in M \subset N_{\sigma(A)}} \frac{\|Au\|_R}{\|u\|_R}.$$

Remark 5.4. *From (5.9), it follows*

$$\frac{\|Au\|_R}{\|u\|_R} \leq \max_{j=1,\cdots,r} |\lambda_j(A)| = \rho(A),\ 0 \neq u \in N_{\sigma(A)},$$

where $\rho(A)$ is the spectral radius *of matrix A.*

Further, we have the following theorem.

Theorem 5.5. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9). Moreover, let the vector spaces N_k, $k = 1, \cdots, r$ be defined by (3.14). Then,*

$$|\lambda_k(A)| = \min_{0 \neq u \in N_k} \frac{\|Au\|_R}{\|u\|_R},\ k = 1, 2, \cdots, r.$$

The minimum is attained for $u = p_k$.

Moreover, the following theorem holds.

Theorem 5.6. *Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9). Then, for every $j = 1, \cdots, r$ and every subspace $N \subset N_{\sigma(A)}$ with $dim N = j$, the following inequalities are valid:*

$$|\lambda_r(A)| \leq \min_{0 \neq u \in N} \frac{\|Au\|_R}{\|u\|_R} \leq |\lambda_j(A)|, \tag{5.11}$$

and the following representation formulas hold:

$$|\lambda_j(A)| = \max_{dim N = j} \min_{0 \neq u \in N} \frac{\|Au\|_R}{\|u\|_R}.$$

Remark 5.7. *From (5.11), it follows*

$$\begin{aligned} \frac{\|Au\|_R}{\|u\|_R} \geq |\lambda_r(A)| &= \min_{j=1,\cdots,r} |\lambda_j(A)| = \min_{j=1,\cdots,r} \frac{1}{|\lambda_j(A^{-1})|} \\ &= \frac{1}{\max\limits_{j=1,\cdots,r} |\lambda_j(A^{-1})|} = \frac{1}{\rho(A^{-1})} = (\rho(A^{-1}))^{-1},\ 0 \neq u \in N_{\sigma(A)}. \end{aligned}$$

6. Connection between the matrices $\mathbf{R^{-1}\frac{A^*R+RA}{2}}$, $\mathbf{R^{-1}\frac{RA-A^*R}{2i}}$, and $\mathbf{R^{-1}A^*RA}$

In [2, Section 6], for diagonalizable matrices $A \in C^{n \times n}$, we have shown that the equation

$$\left(R^{-1}\tfrac{A^*R+RA}{2}\right)^2 + \left(R^{-1}\tfrac{RA-A^*R}{2i}\right)^2 = R^{-1}A^*RA \tag{6.1}$$

is valid. In this section, we prove that this equation remains valid in the subspace

$$N'_{\sigma(A)} := \bigoplus_{\substack{m_j=1 \\ j=1}}^{r} N_{\lambda_j(A)} \subset \bigoplus_{j=1}^{r} N_{\lambda_j(A)} = N_{\sigma(A)}. \tag{6.2}$$

One has the following theorem.

Theorem 6.1. *Let the conditions (C1′)-(C4′) be fulfilled, and R be defined by (2.1). Then,*

$$\left[\left(R^{-1}\tfrac{A^*R+RA}{2}\right)^2 + \left(R^{-1}\tfrac{RA-A^*R}{2i}\right)^2\right] u = R^{-1}A^*RAu,\ u \in N'_{\sigma(A)}. \tag{6.3}$$

Proof. Let $j \in \{1,\cdots,r\}$ with $m_j = 1$. According to [13, Theorems 6 and 7], one has

$$\begin{aligned}
&\left[\left(R^{-1}\tfrac{A^*R+RA}{2}\right)^2+\left(R^{-1}\tfrac{RA-A^*R}{2i}\right)^2\right]p_j\\
=\;&\left(R^{-1}\tfrac{A^*R+RA}{2}\right)^2 p_j+\left(R^{-1}\tfrac{RA-A^*R}{2i}\right)^2 p_j\\
=\;&[Re\lambda_j(A)]^2p_j+[Im\lambda_j(A)]^2p_j\\
=\;&|\lambda_j(A)|^2p_j=\lambda_j(R^{-1}A^*RA)p_j=R^{-1}A^*RAp_j.
\end{aligned}$$

This implies (6.3). □

Remark 6.2. *For diagonalizable matrices $A \in \mathbf{C}^{n\times n}$, the equations (6.3) deliver (6.1) since then $N'_{\sigma(A)} = N_{\sigma(A)} = \mathbf{C}^{n\times n}$.*

7. Applications

In this section, we apply the results of Sections 3, 4, and 5 to obtain the convexity of newly-defined numerical ranges of a general matrix A.

7.1. Applications pertinent to Section 3

Let the conditions $(C1')$-$(C4')$ be fulfilled.
The *numerical range of A restricted to $N_{\sigma(A)}$ with respect to the weighted scalar product $(\cdot,\cdot)_R$* is defined by

$$W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)=\left\{z\in\mathbf{C}\;|\,z=\frac{(Au,u)_R}{(u,u)_R},\;0\neq u\in N_{\sigma(A)}\right\}.$$

Further, let

$$Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]:=\left\{x\in\mathbf{R}\,|\,x=\frac{Re(Au,u)_R}{(u,u)_R},\;0\neq u\in N_{\sigma(A)}\right\};$$

we call it *real part of the numerical range* $W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)$.
Let $\sigma(A)=\{\lambda_j(A),\;j=1,\cdots,r\}$ be the *spectrum of* A, i.e., the set of all eigenvalues of A.
Similarly as before, we define

$$Re[\sigma(A)]:=\{Re\,\lambda_j(A),\;j=1,\cdots,r\}$$

and call it the *real part of the spectrum of* A.
Finally, let $co\{Re[\sigma(A)]\}$ be the *convex hull of* $Re[\sigma(A)]$.
Next, we show the following corollary as an application of Theorem 3.8, Formula (3.13), and Theorem 3.11, Formula (3.15).

Corollary 7.1. (Application 1)
Let the conditions $(C1')$-$(C4')$ be fulfilled.
Then, the set $Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$ is convex, and one has the chain of equations

$$\begin{aligned}
Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] &= \left\{x\in\mathbf{R}\,|\,x=\frac{Re(Au,u)_R}{(u,u)_R},\;0\neq u\in N_{\sigma(A)}\right\}\\
&= \left\{x\in\mathbf{R}\,|\,x=\frac{(R^{-1}\frac{A^*R+RA}{2}u,u)_R}{(u,u)_R},\;0\neq u\in N_{\sigma(A)}\right\}\\
&= W_{N_{\sigma(A)},(\cdot,\cdot)_R}(R^{-1}\tfrac{A^*R+RA}{2})\\
&= co\{Re\,[\sigma(A)]\}.
\end{aligned}$$

If the eigenvalues of A are arranged according to (3.9), then

$$Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]=[Re\,\lambda_r(A),Re\,\lambda_1(A)].$$

Proof. Let $0\neq u\in N_{\sigma(A)}$. Then,

$$2\,\frac{Re(Au,u)_R}{(u,u)_R}=\frac{([A^*R+RA]u,u)}{(u,u)_R}=\frac{(R^{-1}[A^*R+RA]u,u)_R}{(u,u)_R}.$$

The convexity follows from the last form with $R^{-1}[A^*R+RA]$ and the scalar product $(\cdot,\cdot)_R$, see the convexity of the numerical range of a matrix due to Hausdorff in [1, Section 5.4]. Since, with (3.9), one has

$$co\{Re\,[\sigma(A)]\}=[Re\,\lambda_r(A),Re\,\lambda_1(A)],$$

it remains to show that

$$Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] = [Re\,\lambda_r(A), Re\,\lambda_1(A)].$$

The proof of this relation is as follows.
(i) $Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] \subset [Re\,\lambda_r(A), Re\,\lambda_1(A)]$
This inclusion can be deduced from (3.13) with $dim M = m = r - j + 1$ for $j = 1$ and (3.15) with $dim N = r$. Namely, from (3.13), for $j = 1$ and $dim M = r$, i.e., $M = N_{\sigma(A)}$, one has

$$\max_{0\neq u\in N_{\sigma(A)}} \frac{Re(Au,u)_R}{(u,u)_R} \leq Re\,\lambda_1(A)$$

and from (3.15), for $j = r$ and $dim N = r$, i.e., $N = N_{\sigma(A)}$,

$$\min_{0\neq u\in N_{\sigma(A)}} \frac{Re(Au,u)_R}{(u,u)_R} \geq Re\,\lambda_r(A).$$

(ii) $[Re\,\lambda_r(A), Re\,\lambda_1(A)] \subset Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$
Let $\beta \in [Re\,\lambda_r(A), Re\,\lambda_1(A)]$. Then, there exists an α in $0 \leq \alpha \leq 1$ with

$$\beta = \alpha\, Re\lambda_r(A) + (1-\alpha)\, Re\lambda_1(A).$$

Now, due to Theorem 2.3, with the eigenvectors p_r and p_1,

$$Re\lambda_r(A) = \frac{Re(Ap_r,p_r)_R}{(p_r,p_r)_R} \in Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$$

and

$$Re\lambda_1(A) = \frac{Re(Ap_1,p_1)_R}{(p_1,p_1)_R} \in Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)].$$

Thus, due to the convexity of $Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$, it follows that $\beta \in Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$.
In other words, the proof is done along the same line as for [2, Section 7.1] with (3.13) for $j = 1$ and (3.15) for $j = r$. □

Remark 7.2. *One has the relations*

$$Re\lambda_1(A) = \max_{j=1,\cdots,r} Re\lambda_j(A) = \nu[A]$$

and

$$Re\lambda_r(A) = \min_{j=1,\cdots,r} Re\lambda_j(A) = -\nu[-A].$$

Corollary 7.3. (Application 2)
Let the conditions (C1′)-(C4′) be fulfilled. Further, let A be asymptotically stable.
Then,

$$Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] \subset \boldsymbol{R}^- = \{x \in \boldsymbol{R}\,|\,x < 0\}.$$

If A is only stable, then

$$Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] \subset \boldsymbol{R}_0^- = \{x \in \boldsymbol{R}\,|\,x \leq 0\}.$$

Proof. The first assertion follows from (3.4). The second assertion follows in a similar way. □

7.2. Applications pertinent to Section 4

In this section, we proceed in a similar way as in 7.1. So, let

$$Im[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] := \left\{ x \in \boldsymbol{R}\,|\,x = \frac{Im(Au,u)_R}{(u,u)_R},\ 0 \neq u \in N_{\sigma(A)} \right\};$$

we call it *imaginary part of the numerical range* $W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)$.
Further, we define

$$Im[\sigma(A)] := \{Im\,\lambda_j(A),\ j = 1,\cdots,r\}$$

and call it the *imaginary part of the spectrum of A*.
Finally, let $co\{Im[\sigma(A)]\}$ be the *convex hull of* $Im[\sigma(A)]$.

Herewith, we obtain the following corollary.

Corollary 7.4. (Application 3)
Let the conditions (C1′)-(C4′) be fulfilled.
Then, the set $Im[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)]$ *is convex, and one has the chain of equations*

$$\begin{aligned} Im[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] &= \left\{x \in \boldsymbol{R} \,|\, x = \frac{Im(Au,u)_R}{(u,u)_R}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= \left\{x \in \boldsymbol{R} \,|\, x = \frac{(R^{-1}\frac{RA-A^*R}{2i}u,u)_R}{(u,u)_R}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= W_{N_{\sigma(A)},(\cdot,\cdot)_R}(R^{-1}\tfrac{RA-A^*R}{2i}) \\ &= co\{Im[\sigma(A)]\}. \end{aligned}$$

Proof. The assertion follows as in [2, Section 7.2] along with (4.3) for $j = 1$ and (4.4) for $j = r$. □

7.3. Applications pertinent to Section 5

In this subsection, we continue along the same lines as in 7.1 and 7.2.
Let

$$W_{N_{\sigma(A)},\|\cdot\|_R}(A) := \left\{x \in \boldsymbol{R}_0^+ \,|\, x = \frac{\|Au\|_R}{\|u\|_R},\ 0 \neq u \in N_{\sigma(A)}\right\}.$$

Further,

$$|\sigma(A)| := \{|\lambda_j(A)|,\ j = 1, \cdots, r\}$$

is called the *modulus of the spectrum of A*.
Moreover, let $co\{|\sigma(A)|\}$ be the *convex hull of* $|\sigma(A)|$.
Finally, let $S \subset \boldsymbol{R}_0^+$ be any subset. We define

$$S^2 := \{y \,|\, y = s^2, s \in S\}.$$

Next, we show the following corollary.

Corollary 7.5. (Application 4)
Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9).
Then, the set $[W_{N_{\sigma(A)},\|\cdot\|_R}(A)]^2$ *is convex, and one has the chain of relations*

$$\begin{aligned} [W_{N_{\sigma(A)},\|\cdot\|_R}(A)]^2 &= \left\{x \in \boldsymbol{R}_0^+ \,|\, x = \left[\frac{\|Au\|_R}{\|u\|_R}\right]^2 = \frac{\|Au\|_R^2}{\|u\|_R^2}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= \left\{x \in \boldsymbol{R}_0^+ \,|\, x = \frac{([R^{-1}A^*RA]u,u)_R}{(u,u)_R}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= [|\lambda_r(A)|^2, |\lambda_1(A)|^2] \\ &= co\{|\sigma(A)|^2\}. \end{aligned}$$

If A is regular, then $R^{-1}A^*RA$ *is positive definite.*

Proof. The assertion follows as for [2, Corollary 7.4] along with (5.10) for $j = 1$ and (5.11) for $j = r$. Further, $R^{-1}A^*RA$ is apparently regular if A is so as well as self-adjoint and positive definite in the weighted scalar product $(\cdot,\cdot)_R$. □

Next, for $S \subset \boldsymbol{R}_0^+$, we define

$$\sqrt{S} := \{y \,|\, y = \sqrt{s},\ s \in S\}.$$

Herewith, one can rewrite Corollary 7.5 in the following form.

Corollary 7.6. (Application 5)
Let the conditions (C1′)-(C4′) be fulfilled. Further, let the eigenvalues of A be arranged according to (5.9).
Then, the set $W_{N_{\sigma(A)},\|\cdot\|_R}(A)$ *is convex, and one has the chain of relations*

$$\begin{aligned} W_{N_{\sigma(A)},\|\cdot\|_R}(A) &= \left\{x \in \boldsymbol{R}_0^+ \,|\, x = \frac{\|Au\|_R}{\|u\|_R}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= \left\{x \in \boldsymbol{R}_0^+ \,|\, x = \sqrt{\frac{([R^{-1}A^*RA]u,u)_R}{(u,u)_R}}, 0 \neq u \in N_{\sigma(A)}\right\} \\ &= [|\lambda_r(A)|, |\lambda_1(A)|] \\ &= co\{|\sigma(A)|\}. \end{aligned}$$

Proof. For any subset $S \subset \boldsymbol{R}_0^+$, one has

$$\sqrt{S^2} = (\sqrt{S})^2 = S.$$

Thus, from Corollary 7.5, the equations of Corollary 7.6 follow. □

8. Numerical example

In this section, we check the results of Subsection 7.1 as well as of Theorem 6.1, Formula (6.3) numerically. The numerical check of the results of Subsections 7.2 and 7.3 is left to the reader.

8.1. A two-mass vibration model

We take up the multi-mass vibration model of [12], shown in Figure 8.1.

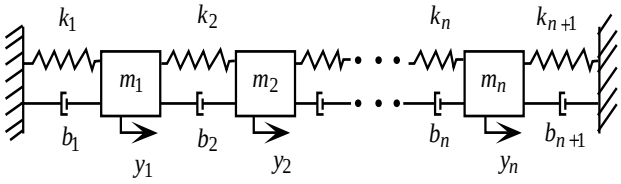


Figure 8.1: Multi-mass vibration model

and study the case $\mathbf{n} = \mathbf{2}$ as in [13]. For the sake of completeness, we give again the details. The associated initial value problem is given by

$$M\ddot{y} + B\dot{y} + Ky = 0,\ y(0) = y_0, \dot{y}(0) = \dot{y}_0,$$

where $y = [y_1, y_2]^T$ and

$$M = \left[\begin{array}{c|c} m_1 & 0 \\ \hline 0 & m_2 \end{array}\right],$$

$$B = \left[\begin{array}{c|c} b_1 + b_2 & -b_2 \\ \hline -b_2 & b_2 + b_3 \end{array}\right],$$

$$K = \left[\begin{array}{c|c} k_1 + k_2 & -k_2 \\ \hline -k_2 & k_2 + k_3 \end{array}\right],$$

with the *mass, damping,* and *stiffness matrices* M, B, and K, as the case may be, and the *displacement vector* y. In *state-space description*, this problem takes the form

$$\dot{x} = Ax, t \geq 0, x(0) = x_0,$$

where $x = [y^T, z^T]^T$, $z = \dot{y}$, and where the *system matrix* A is given by

$$A = \left[\begin{array}{c|c} 0 & E \\ \hline -M^{-1}K & -M^{-1}B \end{array}\right].$$

(i) *Construction of a non-diagonalizable matrix A:*
The pertinent characteristic equation reads

$$|\lambda^2 M + \lambda B + K| = \left|\begin{array}{c|c} \lambda^2 m_1 + \lambda(b_1 + b_2) + (k_1 + k_2) & \lambda(-b_2) - k_2 \\ \hline \lambda(-b_2) - k_2 & \lambda^2 m_2 + \lambda(b_2 + b_3) + (k_2 + k_3) \end{array}\right| = 0.$$

As in [13], for the construction of a case with non-diagonalizable matrix A, we choose

$$b_2 = 0,\ m_2 = m_1 = 1,\ b_3 = b_1,\ k_3 = k_1.$$

Then,

$$\lambda^2 m_1 + \lambda b_1 + (k_1 + k_2) = sk_2 \quad \text{with} \quad s \in \{+1, -1\}.$$

Hence, with $m_1 = 1$,

$$\lambda = -\frac{b_1}{2} \pm \sqrt{(\frac{b_1}{2})^2 - k_1 - k_2 + sk_2}.$$

Now, in order to get one real solution, we set

$$k_1 := (\frac{b_1}{2})^2.$$

This implies

$$\lambda = \begin{cases} -\dfrac{b_1}{2}, & s = +1, \\ -\dfrac{b_1}{2} \pm i\sqrt{2k_2}, & s = -1. \end{cases}$$

(ii) *Data:*
Like in [13], as numerical values for the quantities not yet specified, we choose $b_1 = 1/4$, $k_2 = 2^3 = 8$. On the whole, this delivers the following data:

$$m_1 = m_2 = 1;\ b_1 = 1/4, b_2 = 0, b_3 = 1/4;\ k_1 = 1/64 = 1/2^4, k_2 = 8, k_3 = 1/64 = 1/2^4,$$

which leads to

$$\begin{aligned} M &= \left[\begin{array}{c|c} m_1 & 0 \\ \hline 0 & m_2 \end{array}\right] = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 1 \end{array}\right], \\ B &= \left[\begin{array}{c|c} b_1+b_2 & -b_2 \\ \hline -b_2 & b_2+b_3 \end{array}\right] = \left[\begin{array}{c|c} 0.25 & 0 \\ \hline 0 & 0.25 \end{array}\right], \\ K &= \left[\begin{array}{c|c} k_1+k_2 & -k_2 \\ \hline -k_2 & k_2+k_3 \end{array}\right] = \left[\begin{array}{c|c} 1/64+8 & -1/2 \\ \hline -1/2 & 8+1/64 \end{array}\right] = \left[\begin{array}{c|c} 8.015625 & -0.5 \\ \hline -0.5 & 8.015625 \end{array}\right]. \end{aligned}$$

Further, we choose

$$t_0 = 0$$

as well as

$$y_0 = [-1, 1]^T$$

and

$$\dot{y}_0 = [-1, -1]^T,$$

but y_0 and $\dot{y}_0$ are not needed here.

8.2. Computation of important quantities

Using the Matlab routine *jordan*, one obtains

$$\begin{aligned} \lambda_1(A) &= -0.1250 + 4.0000i, \\ \lambda_2(A) &= -0.1250 - 4.0000i, \\ \lambda_3(A) &= -0.1250, \\ \lambda_4(A) &= \lambda_3(A). \end{aligned}$$

The pertinent eigenvectors and principal vectors are

$$\left[p_1^{(1)}, p_1^{(2)}, p_1^{(3)}, p_2^{(3)}\right] = [p_1, p_2, p_3, p_4]$$

with

$$\begin{aligned} \left[p_1^{(1)}, p_1^{(2)}\right] &= [p_1, p_2] \\ &= \begin{bmatrix} 0.250000000000000 - 0.007812500000000i & 0.250000000000000 + 0.007812500000000i \\ -0.250000000000000 + 0.007812500000000i & -0.250000000000000 - 0.007812500000000i \\ 0 + 1.000976562500000i & 0 - 1.000976562500000i \\ 0 - 1.000976562500000i & 0 + 1.000976562500000i \end{bmatrix}. \end{aligned}$$

and

$$\begin{aligned} \left[p_1^{(3)}, p_2^{(3)}\right] &= [p_3, p_4] \\ &= \begin{bmatrix} 0.062500000000000 & 0.500000000000000 \\ 0.062500000000000 & 0.500000000000000 \\ -0.007812500000000 & 0 \\ -0.007812500000000 & 0 \end{bmatrix}. \end{aligned}$$

They are apparently unnormed. The algebraic multiplicities are thus $m_1 = m_2 = 1$ and $m_3 = 2$.

For the adjoint matrix A^*, we obtain

$$\begin{array}{rcl}\lambda_1(A^*) &=& -0.1250-4.0000i,\\ \lambda_2(A^*) &=& -0.1250+4.0000i,\\ \lambda_3(A^*) &=& -0.1250,\\ \lambda_4(A^*) &=& \lambda_3(A^*).\end{array}$$

The associated eigenvectors and principal vectors are

$$\left[u_1^{(1)^*}, u_1^{(2)^*}, u_1^{(3)^*}, u_2^{(3)^*}\right] = [u_1^*, u_2^*, u_3^*, u_4^*]$$

with

$$\begin{array}{rcl}\left[u_1^{(1)^*}, u_1^{(2)^*}\right] &=& \left[u_1^*, u_2^*\right]\\ &=& \begin{bmatrix} 0.250000000000000+0.007812500000000i & 0.250000000000000-0.007812500000000i\\ -0.250000000000000+0.007812500000000i & -0.250000000000000-0.007812500000000i\\ 0-0.062500000000000i & 0+0.062500000000000i\\ 0+0.062500000000000i & 0-0.062500000000000i\end{bmatrix}.\end{array}$$

and

$$\begin{array}{rcl}\left[u_1^{(3)^*}, u_2^{(3)^*}\right] &=& \left[u_3^*, u_4^*\right]\\ &=& \begin{bmatrix} 0.062500000000000 & 0.500000000000000\\ 0.062500000000000 & 0.500000000000000\\ 0.500000000000000 & 0\\ 0.500000000000000 & 0\end{bmatrix}.\end{array}$$

They are also unnormed.
Now, we biorthogonalize these vectors based on Theorem 2.1 such that the relations

$$(p_k^{(i)}, u_l^{(i)^*}) = \begin{cases} 1, & l = m_i - k + 1\\ 0, & l \neq m_i - k + 1\end{cases}$$

and

$$(p_k^{(i)}, u_l^{(j)^*}) = 0,\ i \neq j.$$

So, with

$$v_l^{(i)^*} = u_{m_i-l+1}^{(i)^*},$$

one has the biorthogonality relations

$$(p_k^{(i)}, v_l^{(j)^*}) = \delta_{kl}\delta_{ij}. \tag{8.1}$$

We give the details. Define

$$\begin{array}{l} v_1^* = v_1^{(1)^*} = u_1^{(1)^*} = u_1^*,\\ v_2^* = v_1^{(2)^*} = u_1^{(2)^*} = u_2^*,\\ v_3^* = v_1^{(3)^*} = u_2^{(3)^*} = u_4^*,\\ v_4^* = v_2^{(3)^*} = u_1^{(3)^*} = u_3^*.\end{array}$$

Then,

$$\alpha_3 := -\frac{(p_4, v_3^*)}{(p_3, v_3^*)} = -8.$$

Define

$$w_4 = p_4 + \alpha_3 p_3$$

and replace p_4 by w_4, i.e., in Matlab set $p_4 = w_4$.
Normalize $v_i^*,\ i = 1,\ldots,4$ by the substitutions

$$v_i^* \to \frac{v_i^*}{\|v_i^*\|_2},\ i = 1,\ldots,4$$

and $p_i,\ i = 1,\ldots,4$ by

$$p_i \to \frac{p_i}{(p_i, v_i^*)}\ i = 1,\ldots,4.$$

Then, we obtain

$$\left[p_1^{(1)}, p_1^{(2)}, p_1^{(3)}, p_2^{(3)}\right] = [p_1, p_2, p_3, p_4]$$

with

$$\begin{aligned}\left[p_1^{(1)}, p_1^{(2)}\right] &= [p_1, p_2] \\ &= \begin{bmatrix} 0.364601934049314 & 0.364601934049314 \\ -0.364601934049314 & -0.364601934049314 \\ -0.045575241756164 + 1.458407736197255i & -0.045575241756164 - 1.458407736197255i \\ 0.045575241756164 - 1.458407736197255i & 0.045575241756164 + 1.458407736197255i \end{bmatrix}.\end{aligned}$$

and

$$\begin{aligned}\left[p_1^{(3)}, p_2^{(3)}\right] &= [p_3, p_4] \\ &= \begin{bmatrix} 0.707106781186548 & 0 \\ 0.707106781186548 & 0 \\ -0.088388347648318 & 0.712609640686961 \\ -0.088388347648318 & 0.712609640686961 \end{bmatrix}.\end{aligned}$$

as well as

$$\left[v_1^{(1)^*}, v_1^{(2)^*}, v_1^{(3)^*}, v_2^{(3)^*}\right] = [v_1^*, v_2^*, v_3^*, v_4^*]$$

with

$$\begin{aligned}\left[v_1^{(1)^*}, v_1^{(2)^*}\right] &= [v_1^*, v_2^*] \\ &= \begin{bmatrix} 0.685679302968773 + 0.021427478217774i & 0.685679302968773 - 0.021427478217774i \\ -0.685679302968773 - 0.021427478217774i & -0.685679302968773 + 0.021427478217774i \\ 0 + 0.171419825742193i & 0 - 0.171419825742193i \\ 0 - 0.171419825742193i & 0 + 0.171419825742193i \end{bmatrix}.\end{aligned}$$

and

$$\begin{aligned}\left[v_1^{(3)^*}, v_2^{(3)^*}\right] &= [v_3^*, v_4^*] \\ &= \begin{bmatrix} 0.707106781186547 & 0.087705801930703 \\ 0.707106781186547 & 0.087705801930703 \\ 0 & 0.701646415445623 \\ 0 & 0.701646415445623 \end{bmatrix}.\end{aligned}$$

With these normed vectors, relations (8.1) are computationally verified.
Further, R in (2.1) can be written as

$$\begin{aligned}R &= u_1^{(1)^*} u_1^{(1)} + u_1^{(2)^*} u_1^{(2)} + u_1^{(3)^*} u_1^{(3)} + u_2^{(3)^*} u_2^{(3)} \\ &= \textstyle\sum_{i=1}^4 u_i^* u_i = \sum_{i=1}^4 v_i^* v_i \\ &= v_1^{(1)^*} v_1^{(1)} + v_1^{(2)^*} v_1^{(2)} + v_1^{(3)^*} v_1^{(3)} + v_2^{(3)^*} v_2^{(3)},\end{aligned}$$

where it goes without saying that the u_i^* and u_i are normed in a similar way as the v_i^* and v_i. Matlab delivers

$$R = \begin{bmatrix} 1.448922794377340 & -0.433538178992724 & 0.068884650702833 & 0.054192272374091 \\ -0.433538178992724 & 1.448922794377340 & 0.054192272374091 & 0.068884650702833 \\ 0.068884650702833 & 0.054192272374091 & 0.551077205622660 & 0.433538178992725 \\ 0.054192272374091 & 0.068884650702833 & 0.433538178992725 & 0.551077205622660 \end{bmatrix}.$$

The eigenvalues of R in (8.5) are given by

$$\begin{aligned}\lambda_1(R) &= 0.117416726023999, \\ \lambda_2(R) &= 0.875965265410791, \\ \lambda_3(R) &= 1.124034734589208, \\ \lambda_4(R) &= 1.882583273976000,\end{aligned}$$

so that R is positive definite.

Remark 8.1. *The vector $p_2^{(3)}$ is a principal vector of stage 2 for matrix A. But, since it is normed such that $(p_2^{(3)}, v_2^{(3)^*}) = 1$ instead of $\|p_2^{(3)}\|_2 = 1$, the equation $Ap_2^{(3)} = \lambda_3 p_2^{(3)} + p_1^{(3)}$ does* **not** *hold, but instead, the equation $Ap_2^{(3)} = \lambda_3 p_2^{(3)} + \gamma_1^{(3)} p_1^{(3)}$ is valid with a factor $\gamma_1^{(3)} \neq 0$, $\gamma_1^{(3)} \neq 1$. Similarly, $u_2^{(3)^*}$ is principal vector of stage 2 for A^*. Similarly, due to the biorthogonalization process, the equation $A^* u_2^{(3)^*} = \overline{\lambda_3} u_2^{(3)^*} + u_1^{(3)^*}$ does* **not** *hold, but instead, the equation $A^* u_2^{(3)^*} = \overline{\lambda_3} u_2^{(3)^*} + \delta_1^{(3)} u_1^{(3)^*}$ is valid with a factor $\delta_1^{(3)} \neq 0$, $\delta_1^{(3)} \neq 1$. We leave it to the reader to check this numerically on our example.*

Remark 8.2. *Due to the foregoing remark, Formula (3.1) looks somewhat different. But, Formula (3.2) remains valid which is the important point since the subsequent findings are based on Formula (3.2), not on Formula (3.1).*

8.3. Numerical check of the validity of Corollary 7.1 (Application 1)

Here, we check the validity of

$$\frac{Re(Au,u)_R}{(u,u)_R} \in Re[W_{N_{\sigma(A)},(\cdot,\cdot)_R}(A)] = [Re\,\lambda_3(A), Re\,\lambda_1(A)],\ 0 \neq u \in N_{\sigma(A)}.$$

or

$$\frac{Re(Au,u)_R}{(u,u)_R} = -0.125,\ 0 \neq u \in N_{\sigma(A)} \subset C^4.$$

We choose $u \in \{p, q, w\}$ where

$$\begin{array}{rcr} p & = & p1 + p2, \\ q & = & 2\,p1 - 3\,p3, \\ w & = & -4\,p2 + 5\,p4. \end{array}$$

We obtain

$$\begin{aligned} \frac{Re(Ap,p)_R}{(p,p)_R} &= -0.125000000000000, \\ \frac{Re(Aq,q)_R}{(q,q)_R} &= -0.125000000000000, \\ \frac{Re(Aw,w)_R}{(w,w)_R} &= -0.616601082213326 \neq -0.125, \end{aligned}$$

where the last result is not surprising since $p_4 = p_2^{(3)} \notin N_{\sigma(A)}$ and thus $w \notin N_{\sigma(A)}$.

8.4. Computational verification of the validity of Theorem 6.1

Here, we check Formula (6.3) of Theorem 6.1 With (3.9), from (6.2) we obtain

$$N'_{\sigma(A)} = N_{\lambda_1(A)} \oplus N_{\lambda_2(A)}$$

and

$$D := \left(R^{-1}\tfrac{A^*R+RA}{2}\right)^2 + \left(R^{-1}\tfrac{RA-A^*R}{2i}\right)^2 - R^{-1}A^*RA =$$

$$\begin{bmatrix} 0.253906250000002 & 0.253906250000000 & 0.000000000000000 & 0.000000000000000 \\ 0.253906249999997 & 0.253906250000002 & 0.000000000000000 & 0.000000000000000 \\ -0.063476562500000 & -0.063476562500000 & -0.253906249999998 & -0.253906250000002 \\ -0.063476562500000 & -0.063476562500000 & -0.253906250000002 & -0.253906250000000 \end{bmatrix}.$$

For

$$p = 2p_1 - 3ip_2 \in N'_{\sigma(A)},$$

we obtain

$$p = \begin{bmatrix} 0.729203868098627 - 1.093805802147941i \\ -0.729203868098627 + 1.093805802147941i \\ -4.466373692104092 + 3.053541197663002i \\ 4.466373692104092 - 3.053541197663002i \end{bmatrix}$$

and

$$Dp = \begin{bmatrix} 0.010908063192107 - 0.018157611026833i \\ -0.030646883054894 + 0.047235291651105i \\ -0.153210777398272 + 0.103250741290140i \\ 0.073274719625260 - 0.045519144009631i \end{bmatrix} \times 10^{-13} \doteq 0$$

so that (6.3) is fulfilled for $u = p$. On the other hand, for

$$q = p_1 + p_3 \notin N'_{\sigma(A)},$$

we obtain

$$q = \begin{bmatrix} 1.071708715235861 \\ 0.342504847137234 \\ -0.133963589404483 + 1.458407736197255i \\ -0.042813105892154 - 1.458407736197255i \end{bmatrix}$$

and

$$Dq = \begin{bmatrix} 0.359077662321295 + 0.000000000000000i \\ 0.359077662321291 - 0.000000000000000i \\ -0.044884707790162 + 0.000000000000005i \\ -0.044884707790162 - 0.000000000000003i \end{bmatrix} \neq 0$$

which is not surprising since $q \notin N'_{\sigma(A)}$.

8.5. Computational aspects

In this subsection, we say something about the used computer equipment and the computation times.
(i) As to the *computer equipment*, the following hardware was available: an Intel Core2 Duo Processor at 3166 GHz, a 500 GB mass storage facility, and two 2048 MB high-speed memories. As software package for the computations, we used MATLAB, Version 7.11.
(ii) The *computation time t* of an operation was determined by the command sequence *t1=clock; operation; t=etime(clock,t1)*. It is put out in seconds, rounded to four decimal places. For the computation of the eigenvalues of matrix A in Subsection 5.3, we used the command [XA,DA]=eig(A); the pertinent computation time was less than 0.0001 s.

9. Conclusion

It has been shown that there exist Rayleigh-quotient representations of the real parts, imaginary parts, and moduli of the eigenvalues of general matrices that parallel those representations known for the eigenvalues of self-adjoint matrices and corresponding to the ones for diagonalizable matrices. The key idea is to use a weighted scalar product defined by a positive definite matrix that is constructed by means of the left principal vectors of the considered matrix and the right principal vectors of its adjoint. As Formulas (3.3), (4.1), and (5.2) show, one essentially obtains the results for general matrices in the same way as for diagonalizable matrices by replacing the full space C^n with the geometric eigenspace $N_{\sigma(A)}$. The results are of interest on their own in Linear Algebra. They are also of potential interest in applications. For example, in the theory of linear dynamical systems, in the study of stability of a vibration problem, the real parts of the eigenvalues of the system matrix are important. Moreover, in systems with conjugate-complex eigenvalues, the moduli of the imaginary parts of the eigenvalues are the circular damped eigenfrequencies of the system. Finally, it could also be of interest for college teaching or research. The relation $(R^{-1}\frac{A^*R+RA}{2})^2 + (R^{-1}\frac{RA-A^*R}{2i})^2 = R^{-1}A^*RA$ derived for diagonalizable matrices A in [2] turns out to be valid only on $N'_{\sigma(A)}$. One feature of the present paper is also that, in the special case of diagonalizable matrices, we get back the results of [2]. On the whole, the results should be of interest to mathematicians as well as engineers.

References

[1] F. Stummel, K. Hainer, Introduction to Numerical Analysis, Scottish Academic Press, Edinburgh, 1980.
[2] L. Kohaupt, *Rayleigh-quotient representation of the real parts, imaginary parts, and moduli of the eigenvalues of diagonalizable matrices*, J. Math. Sci. Model., **2**(2) (2019), 82-98.
[3] L. Kohaupt, *Solution of the vibration problem $M\ddot{y}+B\dot{y}+Ky=0, y(t_0)=y_0, \dot{y}(t_0)=\dot{y}_0$ without the hypothesis $BM^{-1}K = KM^{-1}B$ or $B=\alpha M+\beta K$*, Appl. Math. Sci., **2**(41) (2008), 1989-2024.
[4] A.Czornik, P. Jurgaś, *Some properties of the spectral radius of a set of matrices*, Int. J. Appl. Math. Sci., **16**(2)(2006)183-188.
[5] L. Kohaupt, *Solution of the matrix eigenvalue problem $VA+A^*V=\mu V$ with applications to the study of free linear systems*, J. Comp. Appl. Math., **213**(1) (2008), 142-165.
[6] L. Kohaupt, *Spectral properties of the matrix $C^{-1}B$ with positive definite matrix C and Hermitian B as well as applications*, J. Appl. Math. Comput., **50** (2016), 389-416.
[7] T.J. Laffey, H. Šmigoc, *Nonnegatively realizable spectra with two positive eigenvalues*, Linear Multilinear Algebra, **58**(7-8) (2010), 1053-1069.
[8] P. Lancaster, Theory of Matrices, Academic Press, New York and London, 1969.
[9] P.C. Müller, W.O. Schiehlen, Linear Vibrations, Martinus Nijhoff Publishers, Dordrecht Boston Lancaster, 1985.
[10] S.V. Savchenko, *On the change in the spectral properties of a matrix under perturbations of sufficiently low rank*, Funct. Anal. Appl., **38**(1) (2004), 69-71.
[11] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis, Springer, New York Heidelberg, Third Edition, 2010.
[12] L. Kohaupt, *Construction of a biorthogonal system of principal vectors of the matrices A and A^* with applications to the initial value problem $\dot{x}=Ax, x(t_0)=x_0$*, J. Comput. Math. Optim., **3**(3) (2007), 163-192.
[13] L. Kohaupt, *Further spectral properties of the matrix $C^{-1}B$ with positive definite C and Hermitian B applied to wider classes of matrices C and B*, J. Appl. Math. Comput., **52** (2016), 215-243.
[14] L. Kohaupt, *Biorthogonalization of the principal vectors for the matrices A and A^* with application to the computation of the explicit representation of the solution $x(t)$ of $\dot{x}=Ax, x(t_0)=x_0$*, Appl. Math. Sci., **2**(20) (2008), 961-974.

Journal of Mathematical Sciences and Modelling, 3 (2) (2020) 76-85
Research Article

Journal of Mathematical Sciences and Modelling

Journal Homepage: www.dergipark.gov.tr/jmsm
ISSN 2636-8692
DOI: http://dx.doi.org/10.33187/jmsm.723268

Comparing a Three-Term Perturbation Solution of the Nonlinear ODE of the Jacobi Elliptic SN Function to Its Approximation into Circular Functions

Mohammed Ghazy[1,2]

[1]*Department of Aerospace Engineering, KFUPM University, Dhahran 31261, Saudi Arabia*
[2]*Department of Engineering Mathematics and Physics, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt*

Article Info

Keywords: *Jacobi elliptic function, Lindstedt-Poincaré technique, Perturbation methods*
2010 AMS: *33E05, 34D10, 34E10*
Received: *19 April 2020*
Accepted: *16 August 2020*
Available online: *31 August 2020*

Abstract

In this paper, the nonlinear differential equation of the elliptic sn function is solved analytically using the Lindstedt-Poincaré perturbation method. This differential equation has a cubic nonlinearity and a constant known as the modulus of elliptic integral. This constant takes any value from zero to one and the square of its value is used as a small parameter. Fortunately, there is an exact solution to this differential equation known as the Jacobi sn elliptic function. When the modulus approaches zero, the differential equation becomes linear with the circular sine function as exact solution. The Lindstedt-Poincaré technique is used to render the perturbation solution uniformly valid at larger values of the independent variable and a three-term perturbation solution is obtained. This solution is compared analytically with the approximate expansion of the elliptic function into circular functions in case of a small modulus. Then, it is compared with the exact, numerically calculated, sn elliptic function. The relative percentage error is calculated at certain values of the modulus and for all values of the independent variable. The relative error is reasonably small but increases at larger values of the modulus. In addition, the approximate expansion of the exact solution gives smaller relative error than that of the perturbation solution including the same order of the modulus.

1. Introduction

In some nonlinear problems a perturbation solution may be obtained when a small parameter exists [1]. The obtained perturbation solution depends mainly on the existence of an unperturbed solution i.e. the solution of the same problem when the small parameter vanishes. Through an iterative like procedures, the solution is getting closer to the exact one by adding terms of order of magnitudes less than the base or unperturbed solution. Difficulties arise when a singularity exists in the solution. In this case, it will be non-uniformly valid and some techniques such as those established by Lindstedt-Poincaré or Lighthill can be used to eliminate the non-uniformity in the solution [1, 2, 3]. When applying these techniques the analytical iterations become more difficult as more terms are included in each iteration step. The differential equation of the Jacobi elliptic sn function is an example of a nonlinear ordinary differential equation which includes a cubic nonlinearity and a small parameter. This small parameter has the property of deforming the solution from an initial function to a final one as it goes from zero to unity. The nonlinear differential equation has an exact solution known as Jacobi elliptic sine function or sn function. The value of this function can be obtained from tables or using scientific software such as Matlab or Maple. But when the value of the modulus is close to zero the sn function can be approximated as series expansion of circular functions with different harmonics. This approximation can be calculated without special software [4, 5] and its explicit analytical nature makes it useful in analytical comparison with the perturbation solution.
In this paper the Lindstedt-Poincaré technique will be used to obtain a uniformly valid three-term perturbation solution to the differential equation of the Jacobi elliptic function. In the second section, the perturbation solution is derived and the effect of the modulus on its

Email addresses and ORCID numbers: mohammed.ghazy@kfupm.edu.sa, https://orcid.org/0000-0002-1106-112X (M.Ghazy)

behavior is analytically indicated. In the third section, an approximate series expansion to the exact solution is reviewed. The approximation is derived so that it includes the same order of the small parameter as the perturbation solution. In the fourth section, the perturbation solution is compared with the exact solution and its series expansion in case of small modulus. Solutions in addition to relative errors are tabulated and represented graphically at different values of the modulus. Detailed analysis of behavior of solutions and errors are introduced. Finally, conclusions are drawn in the fifth section.

2. Perturbation solution

Consider the nonlinear differential equation [6]

$$d^2y/dx^2 + (1+k^2)y - 2k^2y^3 = 0, 0 \le k \le 1 \tag{1}$$

where k is constant known as the modulus of elliptic integral and $k \in [0,1]$. When the modulus $k \to 0$, (1) reduces to a simple harmonic oscillator whose solution is a circular function. But when k takes any small positive value less than one i.e. $k \in (0,1)$, a cubic nonlinearity exists. Let define another small parameter $\varepsilon = k^2$, where $\varepsilon < k$ for $k \in (0,1)$. Existence of the small parameter allows using the perturbation technique to solve the above problem. Furthermore, to apply Lindstedt-Poincaré technique, let transform the independent variable from x to u through the following transformation

$$u = \omega x, \omega = \sum_{i=0}^{\infty} \varepsilon^i \omega_i = 1 + \varepsilon\omega_1 + \varepsilon^2\omega_2 + O(\varepsilon^3) \tag{2}$$

substituting (2) into (1) gives

$$\omega^2 y'' + (1+\varepsilon)y - 2\varepsilon y^3 = 0, \tag{3}$$

where $(.)'$ denotes differentiation with respect to u. The next step is expand the dependent variable as series in the small parameter ε

$$y(u;\varepsilon) = \sum_{i=0}^{\infty} \varepsilon^i y_i = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + O(\varepsilon^3). \tag{4}$$

When substituting (4) into (3) then collecting and equating coefficients of equal powers of ε one obtains the following set of linear differential equations

$$y_0'' + y_0 = 0 \tag{5}$$
$$y_1'' + y_1 = (2\omega_1 - 1)y_0 + 2{y_0}^3 \tag{6}$$
$$y_2'' + y_2 = (2\omega_2 + 2\omega_1 - 3{\omega_1}^2)y_0 + (2\omega_1 - 1)y_1 - 4\omega_1{y_0}^3 + 6y_1{y_0}^2. \tag{7}$$

The solution to (5) is the unperturbed solution $y_0(u) = A_0 \sin u$. Knowing from the initial conditions of the sn function and its derivative, i.e. $y(x=0) = sn(0;k) = 0$, $\frac{dy}{dx}(x=0) = cn(0;k)dn(0;k) = 1$, and using the relation in (2), the initial conditions for (5) are $y_0(u=0) = 0, y_0'(u=0) = 1$. After applying these initial conditions, the unperturbed solution takes the form

$$y_0(u) = \sin(u). \tag{8}$$

When substituting (8) into (6) one can solve for $y_1(u)$. But, to make sure that the solution of $y_1(u)$ converges at larger values of u, it is necessary to set $\omega_1 = -1/4$. In this case, the solution to (6) reads

$$y_1(u) = -\frac{3}{16}\sin(u) + \frac{1}{16}\sin(3u). \tag{9}$$

Similarly by substituting $y_0(u), y_1(u)$ from (8), (9) respectively into (7) one can solve for $y_2(u)$. To enforce convergence of $y_2(u)$, it is necessary to set $\omega_2 = 19/64$. Then, by integrating (7) one obtains

$$y_2(u) = \frac{7}{256}\sin(u) - \frac{1}{64}\sin(3u) + \frac{1}{256}\sin(5u). \tag{10}$$

Substituting (8),(9), and (10) into (4), a three term perturbation solution reads

$$y_p(u) = \sin(u) + \varepsilon\left(-\frac{3}{16}\sin(u) + \frac{1}{16}\sin(3u)\right) + \varepsilon^2\left(\frac{7}{256}\sin(u) - \frac{1}{64}\sin(3u) + \frac{1}{256}\sin(5u)\right) + O(\varepsilon^3) \tag{11}$$

or

$$y_p(u) = \left(1 - \frac{3}{16}\varepsilon + \frac{7}{256}\varepsilon^2\right)\sin(u) + \left(\frac{1}{16}\varepsilon - \frac{1}{64}\varepsilon^2\right)\sin(3u) + \frac{1}{256}\varepsilon^2\sin(5u) + O(\varepsilon^3)$$

Rewriting (11) after substituting $\varepsilon = k^2$ gives

$$y_p(u) = \sin(u) + k^2\left(-\frac{3}{16}\sin(u) + \frac{1}{16}\sin(3u)\right) + k^4\left(\frac{7}{256}\sin(u) - \frac{1}{64}\sin(3u) + \frac{1}{256}\sin(5u)\right) + O(k^6)$$

or

$$y_p(u) = \left(1 - \frac{3}{16}k^2 + \frac{7}{256}k^4\right)\sin(u) + \left(\frac{1}{16}k^2 - \frac{1}{64}k^4\right)\sin(3u) + \frac{1}{256}k^4\sin(5u) + O(k^6) \tag{12}$$

where

$$u = \left(1 - \frac{1}{4}k^2 + \frac{19}{64}k^4 + O(k^6)\right)x.$$

2.1. Behavior of the perturbation solution

Equation (12) shows that any term in the obtained solution takes the form of a circular function multiplied by a finite quantity. In addition for the solution to converge the following condition should be satisfied

$$\lim_{n\to\infty} |\frac{\varepsilon^{n+1} y_{n+1}}{\varepsilon^n y_n}| = \varepsilon |\left(\frac{y_{n+1}}{y_n}\right)| = o(1).$$

knowing that $|\left(\frac{y_{n+1}}{y_n}\right)| = O(1)$, the condition of convergence, then, reduces to the condition $\varepsilon = o(1)$, which is known by the definition $\varepsilon = k^2$, where $k \in (0,1)$ as indicated in the second section.

3. Approximate series solution

The Jacobi elliptic function $\mathrm{sn}(x;k)$ is the solution to the differential equation in (1) [6]. For small values of the modulus k this function can be expressed in terms of the circular sine and cosine functions. The derivation of this approximation can be started from knowing that the independent variable in (1), which is the argument of the sn function, is the incomplete elliptic integral of the first kind $F(\phi;k)$;

$$x = F(\phi;k) = \int_0^{\phi} \frac{d\theta}{\sqrt{1-k^2 sin^2\theta}}$$

or

$$x = F(\chi;k) = \int_0^{\chi=sin\phi} \frac{dt}{\sqrt{1-t^2}\sqrt{1-k^2t^2}}$$

where ϕ is known as the amplitude and θ, t are dummy variables. For small values of the modulus k, the sn function can be written as follows [4, 5]

$$sn(x;k) = \sin x - \frac{k^2}{4}\cos x(x - \sin x\cos x).$$

As we derived our perturbation solution to include k^4, it may be reasonable if we compare with approximation of the sn function including k^4 as well. Thus, with some efforts we could derive the following approximation

$$sn(x;k)_{approx} = \sin x - \frac{k^2}{4}\cos x g(x;k) + \frac{k^4}{32}\left(2\cos x g(x;k) - \sin x g^2(x;k)\right), \tag{13}$$

where

$$g(x;k) = (x - \sin x\cos x). \tag{14}$$

4. Results and discussion

The perturbation solution is compared with the exact solution, i.e. the elliptic sn function, and its approximate series expansion in (13). Equations (12), (13) show that when $k \to 0$ the two solutions there reduce to the same base solution $\sin(x)$. Also, the exact sn function reduces to the same solution when $k \to 0$. In this specific case there is no need to compare these solutions numerically. For other values of the modulus, the perturbation solution is expected to be different than the other ones. The three solutions are listed in Table 1 to Table 4 for values of the modulus $k = 0.2, 0.4, 0.6, 0.8$, respectively. In addition, the following relative percentage error form is used to show how close are the explicit perturbation solution and sn_{approx} to the exact, numerically calculated, sn solution.

$$E_{pert} = \frac{sn_{exact} - y_p}{sn_{exact}} \times 100 \tag{15}$$

$$E_{approx} = \frac{sn_{exact} - sn_{app}}{sn_{exact}} \times 100. \tag{16}$$

The perturbation solution, the approximate expansion, and the exact solution at values of the modulus $k = 0.2, 0.4, 0.6, 0.8$ are graphically represented in Figure 1 to Figure 4 respectively, for $x \in [0, K(k)]$, where $K(k)$ is the complete elliptic integral of the first kind. Figure 1 shows that the solutions are very close when $k = 0.2$. In Figure 2 to Figure 4, with increasing k, the Difference between the perturbation solution and the exact solution increases. However, one can notice the small rate of increase of the difference between the approximate expansion and the exact solution.

The relative percentage errors indicated in (15), (16) are graphically represented at the values of the modulus $k = 0.2, 0.4, 0.6, 0.8$ in Figure 5 to Figure 8. It is obvious from Figure 5 to Figure 8, that the relative percentage errors are undefined at $x = 0$ as all solutions are equal to zero at this point. More importantly, the maximum difference between the errors E_{pert} and E_{approx} increases with k. Moreover,one can note that in Figure 5 to Figure 8, this maximum difference occurs at the largest value $x = K(k)$. Actually, such a behavior of the perturbation solution is expected as this solution was not enforced to satisfy the end condition.

The behavior of the error in Figure 5 to Figure 8 can be attributed to the different terms of small and large magnitudes in both solutions y_p and sn_{approx}. The reason can also go back to the different way each solution is mathematically derived, even though, the perturbation

x	y_{pert}	sn_{approx}	sn_{exact}	E_{pert}	E_{approx}
0	0	0	0	NaN	NaN
0.083519	0.082624	0.083418	0.083418	0.95251	-4.6268e-05
0.16704	0.16465	0.16623	0.16623	0.95253	-0.00018229
0.25056	0.24548	0.24785	0.24784	0.95258	-0.00039997
0.33408	0.32455	0.32767	0.32767	0.95268	-0.00068665
0.4176	0.40128	0.40514	0.40514	0.95286	-0.0010262
0.50112	0.47513	0.47971	0.47971	0.95316	-0.0014004
0.58464	0.54561	0.55087	0.55086	0.95361	-0.00179
0.66815	0.61221	0.61812	0.61811	0.95426	-0.0021762
0.75167	0.67451	0.68103	0.68101	0.95517	-0.0025407
0.83519	0.73209	0.73918	0.73916	0.9564	-0.0028665
0.91871	0.78458	0.79219	0.79217	0.95799	-0.0031363
1.0022	0.83165	0.83974	0.83971	0.96003	-0.003332
1.0858	0.87301	0.88153	0.8815	0.96257	-0.0034321
1.1693	0.90841	0.9173	0.91727	0.96568	-0.0034096
1.2528	0.93765	0.94686	0.94683	0.96945	-0.0032287
1.3363	0.96055	0.97002	0.96999	0.97393	-0.0028421
1.4198	0.97697	0.98665	0.98663	0.97922	-0.002188
1.5033	0.98683	0.99667	0.99665	0.98539	-0.0011877
1.5869	0.99007	1	1	0.99254	0.00025687

Table 1: y_{pert}, sn_{approx}, sn, E_{pert}, and E_{approx} at $k = 0.2$

x	y_{pert}	sn_{approx}	sn_{exact}	E_{pert}	E_{approx}
0	0	0	0	NaN	NaN
0.086316	0.083399	0.086192	0.086192	3.2404	-0.00079054
0.17263	0.16608	0.17165	0.17164	3.2416	-0.0031128
0.25895	0.24733	0.25564	0.25562	3.2438	-0.0068231
0.34526	0.32648	0.33748	0.33744	3.2474	-0.011697
0.43158	0.40289	0.4165	0.41643	3.2525	-0.017449
0.51789	0.47597	0.49212	0.492	3.2598	-0.023755
0.60421	0.54518	0.56378	0.56361	3.2698	-0.030275
0.69053	0.61008	0.63102	0.63079	3.2831	-0.036669
0.77684	0.67023	0.6934	0.69311	3.3003	-0.042609
0.86316	0.72532	0.7506	0.75024	3.3222	-0.047775
0.94947	0.77504	0.80232	0.8019	3.3496	-0.051849
1.0358	0.81918	0.84833	0.84787	3.3834	-0.054485
1.1221	0.85755	0.88845	0.88796	3.4246	-0.055282
1.2084	0.89002	0.92254	0.92205	3.4742	-0.053743
1.2947	0.91647	0.95051	0.95004	3.5333	-0.049231
1.3811	0.93685	0.97227	0.97187	3.6031	-0.040921
1.4674	0.9511	0.98776	0.98749	3.6851	-0.027764
1.5537	0.95918	0.99696	0.99687	3.7807	-0.0084412
1.64	0.96109	0.99981	1	3.8914	0.018666

Table 2: y_{pert}, sn_{approx}, sn, E_{pert}, and E_{approx} at $k = 0.4$

x	y_{pert}	sn_{approx}	sn_{exact}	E_{pert}	E_{approx}
0	0	0	0	NaN	NaN
0.092145	0.087225	0.091972	0.091968	5.1574	-0.0045588
0.18429	0.17342	0.18291	0.18288	5.1722	-0.017926
0.27643	0.2576	0.27183	0.27173	5.1976	-0.039204
0.36858	0.33885	0.35781	0.35757	5.2343	-0.066992
0.46072	0.41634	0.44001	0.43957	5.2834	-0.099508
0.55287	0.48938	0.51772	0.51702	5.3462	-0.13473
0.64501	0.5574	0.59037	0.58937	5.4243	-0.17051
0.73716	0.61995	0.65751	0.65617	5.5193	-0.20469
0.8293	0.67673	0.71881	0.71713	5.6334	-0.23514
0.92145	0.72753	0.77407	0.77207	5.7689	-0.25971
1.0136	0.77224	0.82318	0.82091	5.9288	-0.27624
1.1057	0.81084	0.86611	0.86367	6.1169	-0.28239
1.1979	0.84333	0.90287	0.90039	6.3375	-0.27554
1.29	0.86976	0.93353	0.93118	6.5958	-0.25258
1.3822	0.89019	0.95816	0.95615	6.8981	-0.20977
1.4743	0.90469	0.97682	0.97543	7.2513	-0.1425
1.5665	0.91331	0.98955	0.98911	7.6633	-0.045073
1.6586	0.91607	0.99639	0.99728	8.143	0.08948
1.7508	0.913	0.9973	1	8.7005	0.26966

Table 3: y_{pert}, sn_{approx}, sn, E_{pert}, and E_{approx} at $k = 0.6$

x	y_{pert}	sn_{approx}	sn_{exact}	E_{pert}	E_{approx}
0	0	0	0	NaN	NaN
0.10502	0.10064	0.10472	0.1047	3.8767	-0.018692
0.21003	0.19926	0.20769	0.20753	3.9862	-0.073237
0.31505	0.29398	0.30725	0.30676	4.1665	-0.1592
0.42006	0.38318	0.40195	0.40087	4.4146	-0.26971
0.52508	0.46558	0.49061	0.48867	4.7261	-0.39603
0.6301	0.5403	0.57232	0.56931	5.0957	-0.52824
0.73511	0.60684	0.64649	0.64228	5.5183	-0.6557
0.84013	0.66503	0.71283	0.7074	5.9903	-0.76743
0.94514	0.71495	0.77126	0.76474	6.5109	-0.8523
1.0502	0.75687	0.82189	0.81457	7.084	-0.89918
1.1552	0.79113	0.86499	0.8573	7.7188	-0.89703
1.2602	0.8181	0.90089	0.89343	8.431	-0.83504
1.3652	0.83812	0.92995	0.92346	9.2418	-0.70262
1.4702	0.85143	0.95254	0.9479	10.178	-0.48913
1.5752	0.85821	0.96899	0.96722	11.27	-0.18316
1.6803	0.85856	0.97956	0.98181	12.553	0.22877
1.7853	0.85249	0.98442	0.992	14.063	0.7641
1.8903	0.8399	0.98357	0.99801	15.842	1.4465
1.9953	0.82064	0.97692	1	17.936	2.308

Table 4: y_{pert}, sn_{approx}, sn, E_{pert}, and E_{approx} at $k = 0.8$

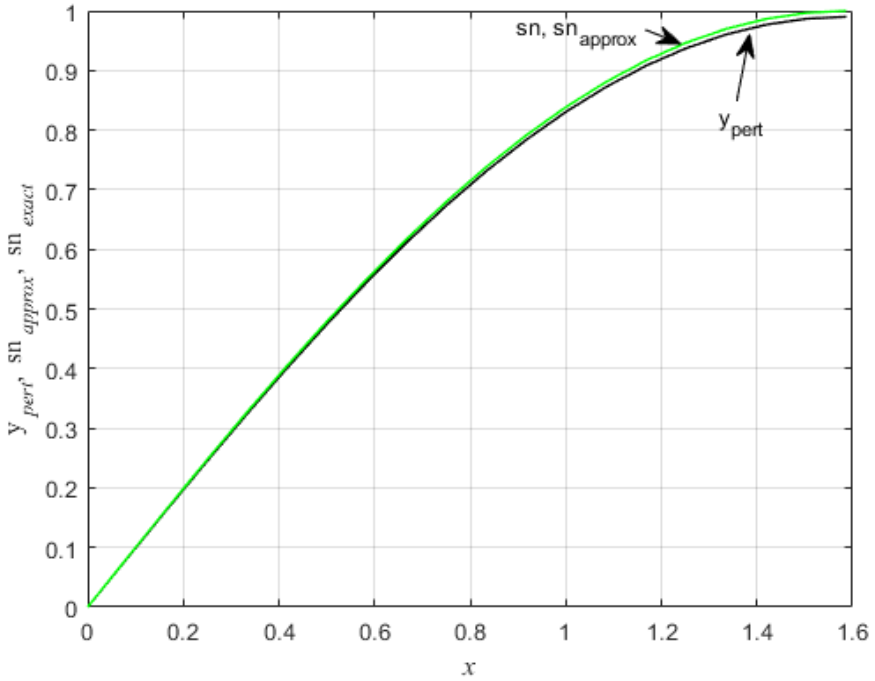


Figure 1: y_p, sn_{approx}, and exact sn at $k = 0.2$.

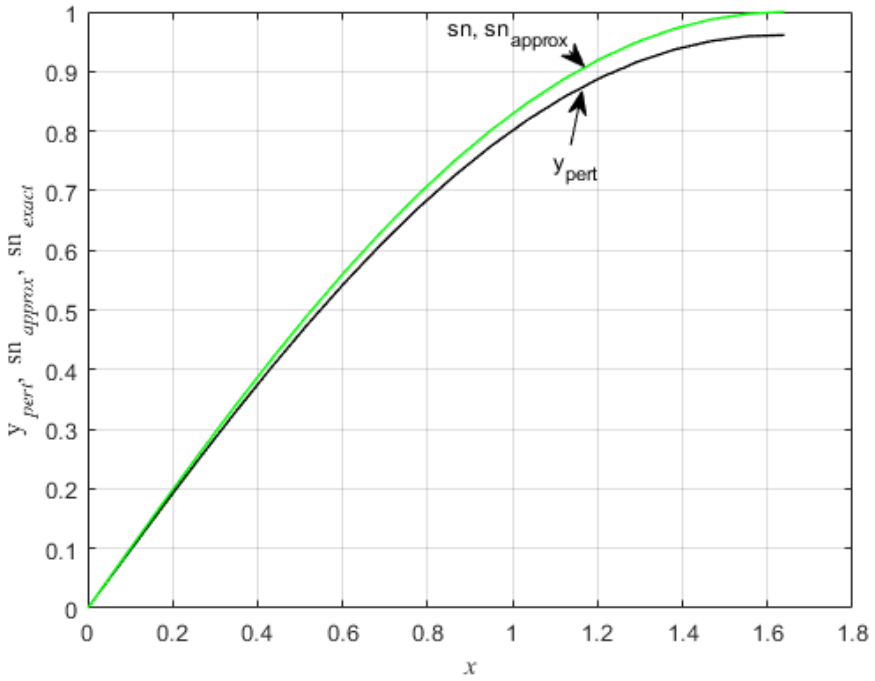


Figure 2: y_p, sn_{approx}, and exact sn at $k = 0.4$.

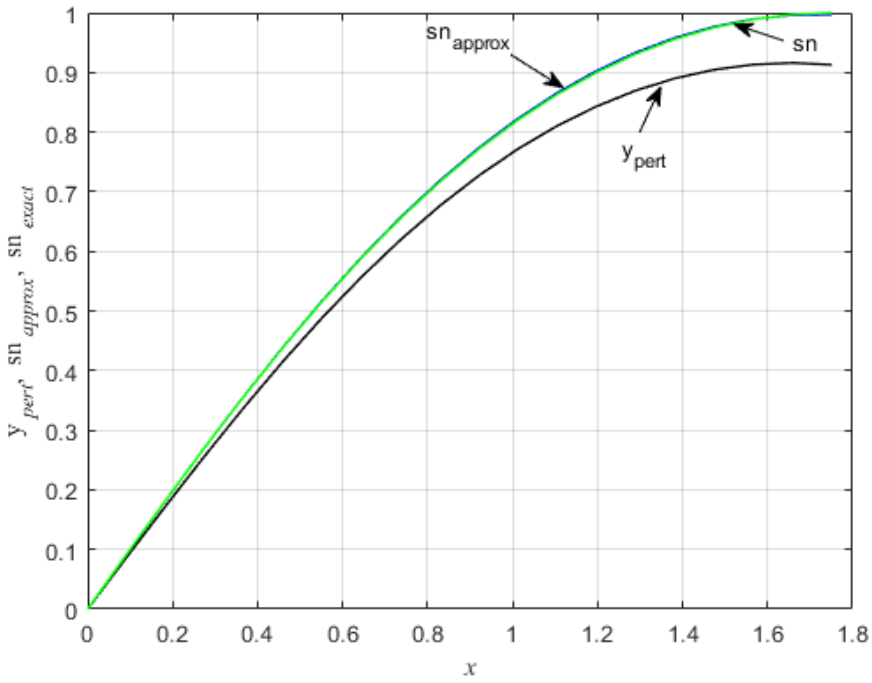


Figure 3: y_p, sn_{approx}, and exact sn at $k = 0.6$.

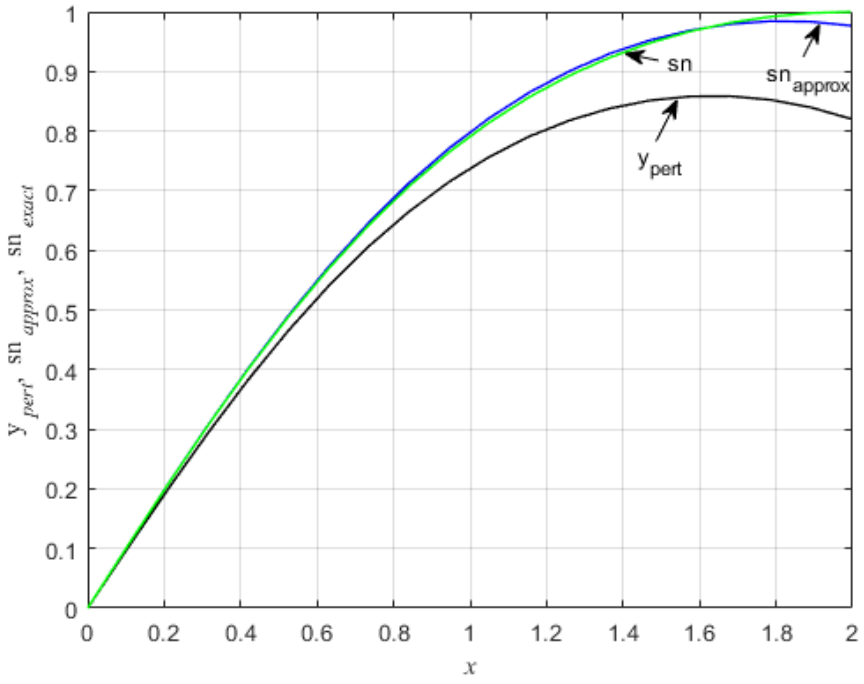


Figure 4: y_p, sn_{approx}, and exact sn at $k = 0.8$.

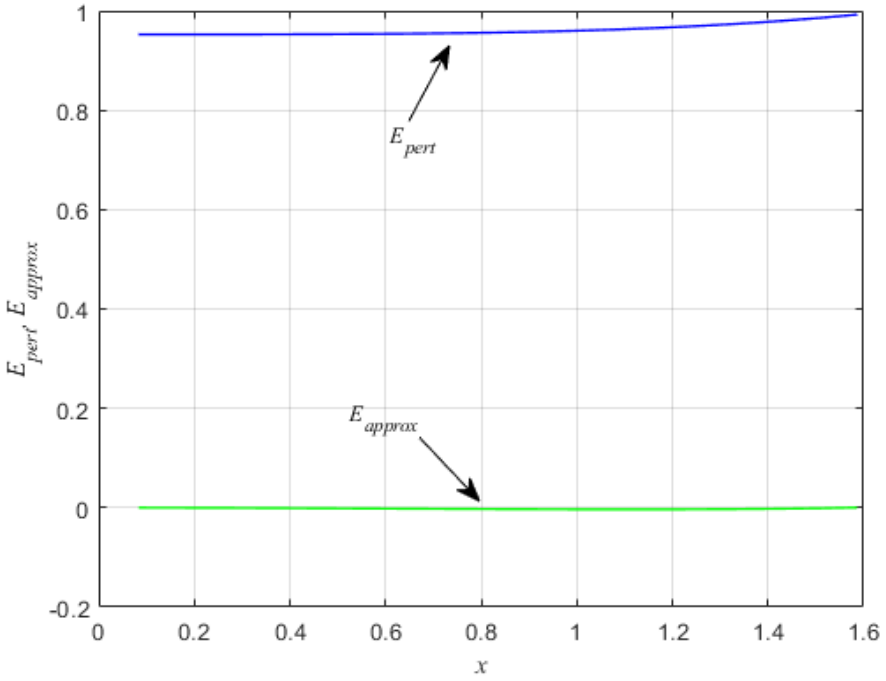


Figure 5: *Relative percentage error at* $k = 0.2$.

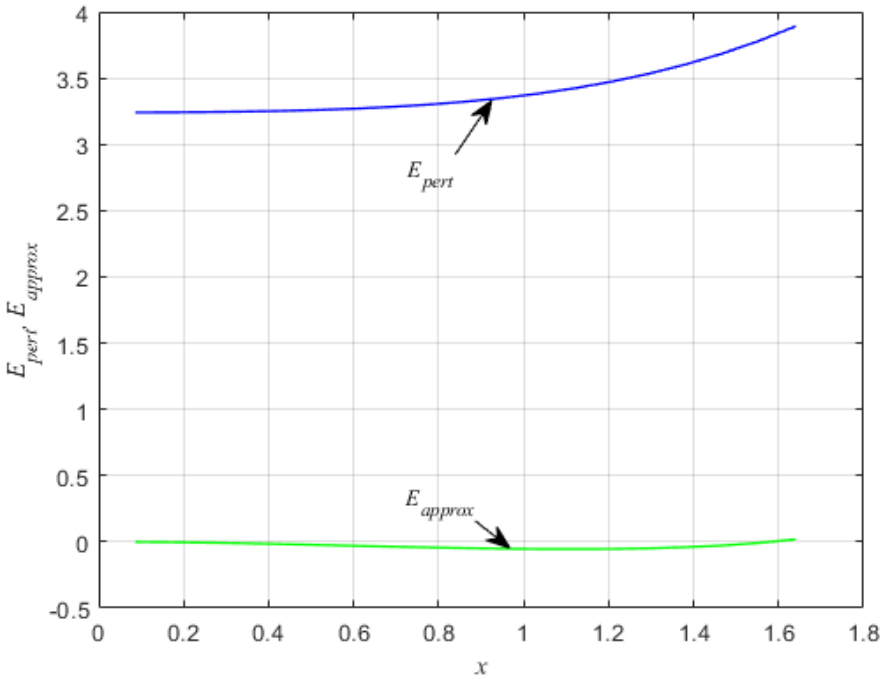


Figure 6: *Relative percentage error at* $k = 0.4$.

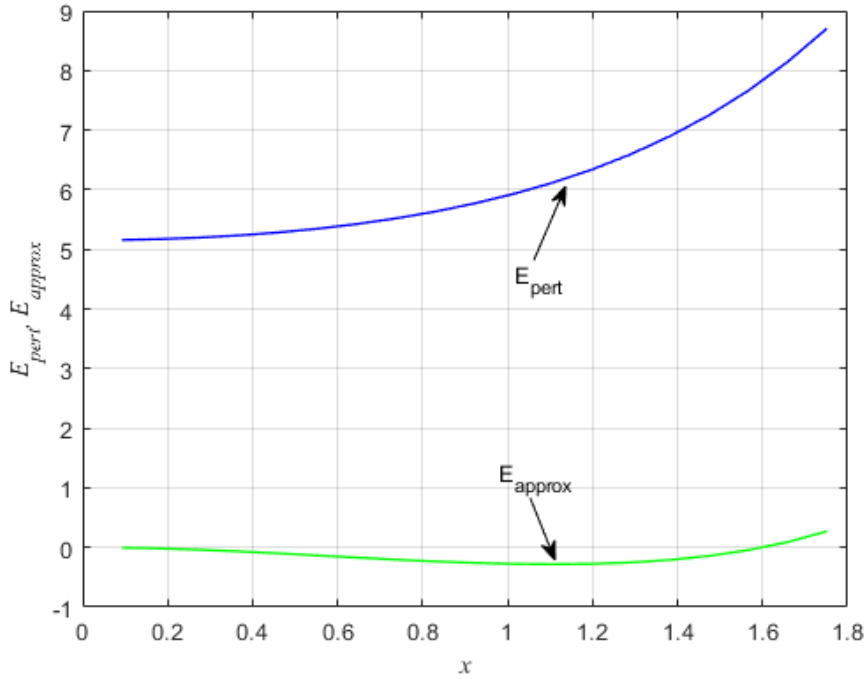


Figure 7: *Relative percentage error at* $k = 0.6$.

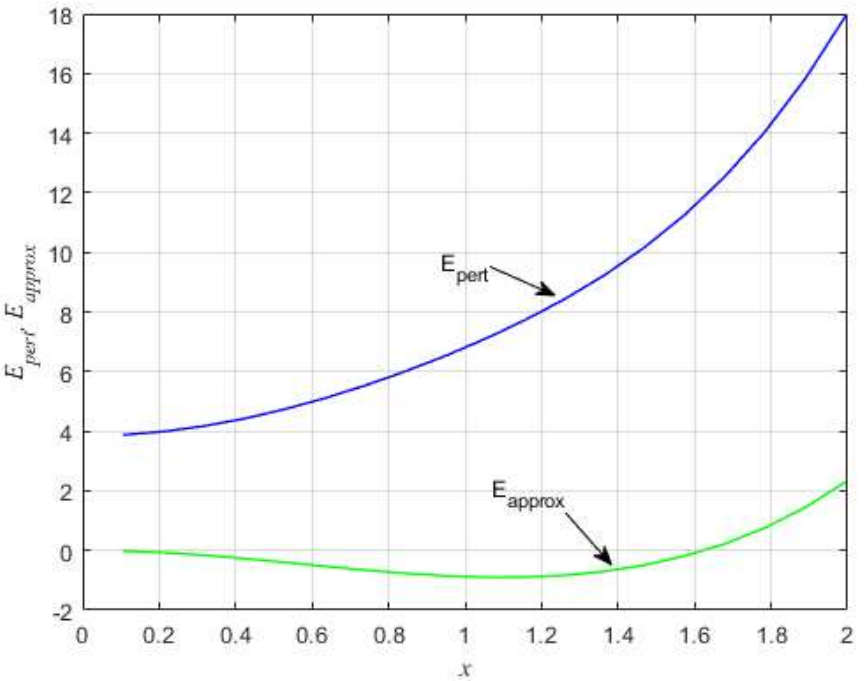


Figure 8: *Relative percentage error at* $k = 0.8$.

solution and the approximate expansion are built on the assumption of a small value of the modulus. The approximate series expansion in (13) includes the function $g(x;k) = (x - \sin x \cos x)$ that does not exist in the perturbation solution in (12). At $k = 0.2, 0.4, 0.6$ the maximum absolute value of E_{pert} is $0.99254\%, 3.8914\%, 8.7005\%$ respectively, while at $k = 0.8$ this value jumps to 17.936%. Thus, for small values of k the relative percentage error is reasonably small and the perturbation solution based on this assumption can be used. Fortunately, in this specific problem we have an exact solution and an approximation of this solution, to compare with the perturbation solution. However, the shown results are indicative of how perturbation solution performs in cases when exact solution doesn't exist.

5. Conclusion

An analytical approximate perturbation solution to the nonlinear ordinary differential equation of the Jacobi elliptic sn function is obtained assuming a small value of the modulus. The relative percentage error between the perturbation solution and the numerical exact one is reasonably small. But, at larger values of the modulus, this error becomes very big. An approximate series expansion of the sn function gives smaller maximum errors than the perturbation solution. However, the magnitude and sign of the error of the series expansion change at different values of the independent variable. Results also give insights into the effect of the mathematical basis of perturbation and approximate series solutions on their accuracy even though they both depend on the small parameter assumption. In future, such results can be considered when applying a Lindstedt-Poincar*é* perturbation solution to nonlinear problems.

Acknowledgement

The author would like to thank the reviewers for their valuable comments that helped in improving the manuscript.

References

[1] A. Nayfeh, *Perturbation Methods*, John Wiley and Sons Inc., New York, 1973.
[2] M. Lighthill, *A technique for rendering approximate solutions to physical problems uniformly valid*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science: Series 7, **40**(311) (1949), 1179-1210.
[3] C. Comstock, *On Lighthill's method of strained coordinates*, SIAM J. Appl. Math., **16**(3) (1986), 596-602.
[4] M. Abramowitz, I. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, Dover Books on Mathematics, 1965.
[5] P. Byrd, M. Friedman, *Handbook of Elliptic Integrals for Engineers and Scientists*, Springer-Verlag, Berlin, 1971.
[6] D. F. Lawden, *Elliptic Functions and Applications*, Springer-Verlag, New York, 1989.

Journal of Mathematical Sciences and Modelling, 3 (2) (2020) 86-94
Research Article

Journal of Mathematical Sciences and Modelling

Journal Homepage: www.dergipark.gov.tr/jmsm
ISSN 2636-8692
DOI: http://dx.doi.org/10.33187/jmsm.634089

A Highly Approximate Pseudo-Spectral Method for the Solution of Convection-Diffusion Equations

M. S. El-Azab[1], R. M. El-Ashwah[2], M. M. Abbas[2*] and G. I. El-Baghdady[1]

[1]*Department of Engineering Physics and Mathematics, Faculty of Engineering, Mansoura University, Egypt.*
[2]*Department of Mathematics, Faculty of Science, Damietta University, New Damietta, Egypt.*
[*]*Corresponding author E-mail: mahy_mam@yahoo.com*

Article Info

Keywords: *Gauss-Legendre polynomials, Legendre differentiation matrices, Legendre pseudo-spectral method, Parabolic advection diffusion equations*
2010 AMS: *65M70. 65M12. 65D25.*
Received: *16 October 2019*
Accepted: *21 July 2020*
Available online: *31 August 2020*

Abstract

The main purpose of this paper is to compute a highly accurate numerical solution of two dimensional convection–diffusion equations with variable coefficients by using Legendre pseudo-spectral method based on Legendre-Gauss-Lobatto nodes. The Kronecker product is used here to formulate a linear system of differentiation matrices; this system was reduced to be more accurate with less memory usage. Error analysis with test problems are introduced to show that the suggested scheme of the spectral method has high accuracy.

1. Introduction

Given a simply connected, domain $\Omega \equiv [a,\ b]\ \times [c,\ d]$ in R^2 , with Lipschitz boundary $\partial\Omega$, a time interval $I \equiv [0,T]$ and the differential operator

$$\mathscr{K}u = -\nabla \cdot (A(\mathbf{x},t)\nabla u) + p(\mathbf{x})u + q \cdot \nabla u, \tag{1.1}$$

where $\mathbf{x} = (x,y)$,∇ and $\nabla\cdot$ denote the gradient and divergence operators, respectively and $A(x)$ represents a matrix-value function. In this paper we are concerned with the numerical approximation of the linear convection- or advection-diffusion problem: Find $u : Q \equiv \Omega \times I \to R$ such that:

$$\partial_t u + \mathscr{K}u = f(\mathbf{x},t) \quad in\ Q, \tag{1.2}$$

$$u = u_B(\mathbf{x}) \quad On\ \partial\Omega \times I, \tag{1.3}$$

$$u(\mathbf{x},0) = u_0(x) \quad in\ \Omega, \tag{1.4}$$

where $q \equiv (\alpha,\beta)$ is the convection vector, u_B is the boundary function and f represents the source or reaction term. Here and elsewhere we shall assume that all coefficients occurring in these equations are uniformly bounded.
For solving this problem, we construct a solution using pseudo-spectral method which may be considered as an extension to the work in [1, 2] There are two steps to form this solution [3, 4, 5]. First, we use polynomial interpolation of the solution based on some suitable nodes as discrete representation of the solution; here we chose Legendre nodes. Secondly, we form a system of algebraic equations from the previous discretization after collocating it; this step replaces differential operators by matrix approximations (see[3, 4, 5, 6]).
A brief outline of the paper is as follows: Section 2 describes some necessary notations about the continuous problem. Section 3 summarizes the steps of solving the problem (1.1)-(1.4) by using Legendre pseudo–spectral method and modifying the resulted system using matrices properties. As a result a system of algebraic linear equations is formed and a solution of the considered problem is discussed. In Section 4 we shall prove error estimates of the scheme given in Section 3. In Section 5, some numerical examples are presented to show the effectiveness of the proposed scheme.

***Email addresses and ORCID numbers*: ms_elazab@hotmail.com, https://orcid.org/0000-0003-2885-107X (M. S. El-Azab), r_elashwah@yahoo.com, https://orcid.org/0000-0002-5490-3745 (R. M. El-Ashwah),mahy_mam@yahoo.com, https://orcid.org/0000-0003-1549-9302 (M. Abbas),amoun1973@yahoo.com, https://orcid.org/0000-0001-5317-5195 (G. I. El-Baghdady)**

2. The continuous problem

Throughout this paper we use standard notations that will be used in the sequel [7]. We use the functional spaces $L_\infty(\Omega), L_2(\Omega), V = H_0^1(\Omega), C(I;L_2(\Omega)), L_2(I;L_2(\Omega))$ (see e.g. [6, 8]). By $(.,.)$ we shall denote either the inner product in $L_2(\Omega)$. We denote by $|\,.\,|, \|.\|_\infty, \|.\|$, the norms in $L_2(\Omega), L_\infty(\Omega)$ and V, respectively. All the constants which occur in the course of this paper will be denoted by C.(ε is small and $C_\varepsilon = C(\varepsilon^{-1})$)
We shall consider the collocation points $\{x_i\}_{i=0}^N$ which is the set of $(N+1)$ Legendre-Gauss-Lobatto nodes [4]. Also we define the family $V_N = span\{\ell_p(s) : 1 \le p \le N-1\}$ of finite dimensional subspace of V where $\{\ell_p\}$ are the Lagrange basis polynomials associated with $\{x_i\}_{i=0}^N$.
Now we list the following assumptions on the coefficients and data in (1.1)-(1.4):
(A1) $A(\mathbf{x},t)$ is symmetric, Lipschitz continuous in t and uniformly positive definite matrix in Q, i.e.,

$$(A\xi,\xi) \ge c|\xi|^2,$$

$$\left|\left|A(\mathbf{x},t) - A(\mathbf{x},t')\right|\right|_M \le c|t-t'| \quad \forall t,\, t' \in I,$$

where $||\cdot||_M$ is a matrix norm.
(A2) $p,\ \nabla\cdot q \in L_\infty(\Omega)$, *with* $p>0, \quad p\cdot n \ge 0$ *on* $\partial\Omega$ and

$$(p\xi,\xi) \ge c|\xi|^2. \tag{2.1}$$

(A3) $p - \frac{1}{2}\nabla\cdot q \ge 0$ *in* Ω.
(A4) $f : Q \to R$ and $u_B(\mathbf{x}) : \partial\Omega\times I \to R$ are Lipschitz continuous in the sense of

$$|f(\mathbf{x},t) - f(\mathbf{x},t')| \le c|t-t'| \quad \forall t,\, t' \in I,$$

and analogously for $u_B(\mathbf{x})$.
(A5) $u_0(\mathbf{x}) \in V$.
Under these assumptions, the weak solution of (1.1)-(1.4) is defined as:
Find $u : Q \to R$ such that

$$(\partial_t u,\chi) + ((u,\chi)) = (f,\chi), \quad \forall \chi \in V, \quad a.e.t \in I, \tag{2.2}$$

$$u(\mathbf{x},0) = u_0(\mathbf{x}) \ \ in\ \Omega.$$

Here we use the notation$((u,\chi))$ to represent the bilinear form corresponding to the differential operator $\mathscr{K}$; i.e.

$$((u,\chi)) = (A(x)\nabla u,\nabla\chi) + (q\cdot\nabla u + p(\mathbf{x})u,\chi). \tag{2.3}$$

Without loss of generality, we may assume a homogeneous Dirichlet boundary condition on $\partial\Omega$, i.e., $u_B(\mathbf{x}) = 0$. Otherwise, we write the bilinear form (2.1) in terms of u and $u_B(\mathbf{x})$ in place of u.
The existence of a weak solution $u \in L_2(I;L_2(\Omega)) \cap C(I;L_2(\Omega))$ is guaranteed by Lax-Milgram theorem. In fact the use of (A2) and (A3) imply that the bilinear form $((.,.))$ is bounded and V-elliptic, i.e.,

$$|((u,\chi))| \le C\|u\|\ \|\chi\| \qquad \forall u,\chi \in L_2(I;L_2(\Omega)) \cap C(I;L_2(\Omega)),$$

$$((u,u)) \ge C\|u\|^2 \qquad \forall u \in L_2(I;L_2(\Omega)) \cap C(I;L_2(\Omega)). \tag{2.4}$$

However, to prove the coercivity condition (2.4), it requires to use (A3) with the application of Gauss's theorem:

$$2(q\cdot\nabla\omega,\omega) = (\omega(q\cdot n),\omega) - (\omega(\nabla\cdot q),\omega).$$

3. The suggested approximation scheme

For a given integer number $N_z \ge 0$, we denote by $z_0,\ z_1,\ \cdots, z_N$, the nodes of the shifted N_z+1-point integration formula of Gauss, Gauss-Radau or Gauss-Lobatto type, and by $w_0,\ w_1,\ \cdots, w_N$ the corresponding weights [4]. In order to solve problem (1.1)-(1.4) by pseudo-spectral method we use the notation

$$\hat{\varphi}(z) = \varphi\left(\frac{2(z-\alpha)}{\beta-\alpha} - 1\right) \qquad \forall z \in [\alpha,\,\beta],$$

to approximate the function $u(\mathbf{x},t)$ in Q by using cubic grid consisting of $(N_x+1)\times(N_y+1)\times(N_t+1)$ nodes as

$$u(\mathbf{x},t) \approx \sum_i^{N_x}\sum_j^{N_y}\sum_k^{N_t} U_{i,j,k}\hat{\varphi}_i(x)\,\hat{\varphi}_j(y)\,\hat{\varphi}_k(t), \tag{3.1}$$

where

$$U_{i,j,k} = u\left(x_i^{(N_x)}, y_j^{(N_y)}, t_k^{(N_t)}\right).$$

Here we have used Lagrange's interpolant functions which are defined by

$$\varphi_i(z) = \prod_{j=0, j\neq i}^{N_z} \frac{z - z_j}{z_i - z_j} \quad , \ \forall 0 \leq i \leq N_z.$$

By using the Kronecker product [9], equation (3.1) can be expressed in the following matrix form:

$$u(x,y,t) \approx \left(\hat{\Phi}^{(N_x)}_{[a,b]}(x) \otimes \hat{\Phi}^{(N_y)}_{[c,d]}(y) \otimes \hat{\Phi}^{(N_t)}_{[0,T]}(t)\right) \bar{U},$$

or simply

$$u(x,y,t) \approx \left(\hat{\Phi}(x) \otimes \hat{\Phi}(y) \otimes \hat{\Phi}(t)\right) \bar{U},$$

where $\bar{U}$ is a column vector given by:

$$\bar{U} = \left[u_{0,0,0}, u_{0,0,1}, u_{0,0,2}, \ldots, u_{0,0,N_t+1}, u_{0,1,0}, \ldots, u_{0,1,N_t+1}, \ldots\ldots, u_{N_x+1,N_y+1,N_t+1}\right]^T.$$

By using a general form of the first differentiation matrix

$$D_{N_z+1} = \left(D_{ij} = \frac{d}{dz}\phi_j\ (z)\Big|_{z=z_i}\right)_{0\leq i,j\leq N_z}$$

from Lagrange interpolation with a modification [6] to improve accuracy

$$D_{im} = \begin{cases} \frac{\prod_{k\neq i,m}^{N_z}(z_i - z_k)}{\prod_{k\neq m}^{N_z}(z_m - z_k)} & i < m, i \neq 0, N_z, \\ -D_{N_z - i, N_z - m} & i > m, i \neq 0, N_z, \\ -\sum_{k\neq i}^{N_z} D_{ik} & i = m \neq 0, N_z, \\ -\frac{N_z(N_z+1)}{4} & i = m = 0, \\ \frac{N_z(N_z+1)}{4} & i = m = N_z, \end{cases}$$

with the classical form of the second order differentiation matrix $D^2_{N_z+1}$ [1]:

$$D^2_{im} = \begin{cases} 2D_{im}\left(D_{ii} - \frac{1}{z_i - z_m}\right) & i \neq m, \\ -\sum_{k=0, k\neq i}^{N_z} D^2_{ik} & i = m, \end{cases}$$

we can rewrite equation (1.2) in discrete form

$$\begin{aligned} &\left[\frac{2}{T}\left(\hat{\Phi}(x) \otimes \hat{\Phi}(y) \otimes \hat{\Phi}(t) D_{N_t+1}\right) + \frac{2\lambda_1(x,y,t)}{b-a}\left(\hat{\Phi}(x) D_{N_x+1} \otimes \hat{\Phi}(y) \otimes \hat{\Phi}(t)\ \right)\right. \\ &+ \frac{2\lambda_2(x,y,t)}{d-c}\left(\hat{\Phi}(x) \otimes \hat{\Phi}(y) D_{N_y+1} \otimes \hat{\Phi}(t)\right) - \frac{4\lambda_3(x,y,t)}{(b-a)^2}\left(\hat{\Phi}(x) D^2_{N_x+1} \otimes \hat{\Phi}(y) \otimes \hat{\Phi}(t)\right) \\ &- \frac{4\lambda_4(x,y,t)}{(d-c)^2}\left(\hat{\Phi}(x) \otimes \hat{\Phi}(y)\ D^2_{N_y+1} \otimes \hat{\Phi}(t)\right) \\ &- \frac{4\lambda_5(x,y,t)}{(b-a)(d-c)}\left(\hat{\Phi}(x) D_{N_x+1} \otimes \hat{\Phi}(y)\ D_{N_y+1} \otimes \hat{\Phi}(t)\right) \\ &\left.+ p(x,y)\left(\hat{\Phi}(x) \otimes \hat{\Phi}(y) \otimes \hat{\Phi}(t)\ \right)\right] \bar{U} = F(x,y,t), \end{aligned} \tag{3.2}$$

where $\lambda_i, i = 1,2,\ldots,5$ are the coefficients of $\partial_x u,\ \partial_y u, \partial^2_x u, \partial^2_y u$ and $\partial_{xy} u$ respectively.
Then, by collocating equation (3.2) at any interior collocation point $\left\{\left(x_i, y_j, t_k\right)\right\}_{i,j,k}$, $1 \leq i \leq N_x - 1, 1 \leq j \leq N_y - 1$, and $1 \leq k \leq N_t$, equation (3.2) becomes:

$$\begin{aligned} &\left[\frac{2}{T}\left(e^{N_x+1}_i \otimes e^{N_y+1}_j \otimes D_{N_t+1} e^{N_t+1}_k\right) + \frac{2\lambda^1_{ijk}}{b-a}\left(D_{N_x+1} e^{N_x+1}_i \otimes e^{N_y+1}_j \otimes e^{N_t+1}_k\right)\right. \\ &+ \frac{2\lambda^2_{ijk}}{d-c}\left(e^{N_x+1}_i \otimes D_{N_y+1} e^{N_y+1}_j \otimes e^{N_t+1}_k\right) - \frac{4\lambda^3_{ijk}}{(b-a)^2}\left(D^2_{N_x+1} e^{N_x+1}_i \otimes e^{N_y+1}_j \otimes e^{N_t+1}_k\right) \\ &- \frac{4\lambda^4_{ijk}}{(d-c)^2}\left(e^{N_x+1}_i \otimes D^2_{N_y+1} e^{N_y+1}_j \otimes e^{N_t+1}_k\right) - \frac{4\lambda^5_{ijk}}{(b-a)(d-c)}\left(D_{N_x+1} e^{N_x+1}_i \otimes D_{N_y+1} e^{N_y+1}_j \otimes e^{N_t+1}_k\right) \\ &\left.+ p_{ij}\left(e^{N_x+1}_i \otimes e^{N_y+1}_j \otimes e^{N_t+1}_k\right)\right] \bar{U} = F_{ijk}, \end{aligned} \tag{3.3}$$

where e^M_r is the r-th row of the identity matrix I_M. Rewriting equation (3.3) in matrix form yields

$$\begin{aligned} &\left[\frac{2}{T}\left([I_{N_x+1}]_{1:N_x-1,:} \otimes [I_{N_y+1}]_{1:N_y-1,:} \otimes [D_{N_t+1}]_{1:N_t,:}\right)\right. \\ &+ \frac{2}{b-a}\tilde{\lambda}_1\left([D_{N_x+1}]_{1:N_x-1,:} \otimes [I_{N_y+1}]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right) \\ &+ \frac{2}{d-c}\tilde{\lambda}_2\left([I_{N_x+1}]_{1:N_x-1,:} \otimes [D_{N_y+1}]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right) \\ &- \frac{4}{(b-a)^2}\tilde{\lambda}_3\left(\left[D^2_{N_x+1}\right]_{1:N_x-1,:} \otimes [I_{N_y+1}]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right) \\ &- \frac{4}{(d-c)^2}\tilde{\lambda}_4\left([I_{N_x+1}]_{1:N_x-1,:} \otimes \left[D^2_{N_y+1}\right]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right) \\ &- \frac{4}{(b-a)(d-c)}\tilde{\lambda}_5\left([D_{N_x+1}]_{1:N_x-1,:} \otimes [D_{N_y+1}]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right) \\ &\left.\tilde{p}\left([I_{N_x+1}]_{1:N_x-1,:} \otimes [I_{N_y+1}]_{1:N_y-1,:} \otimes [I_{N_t+1}]_{2:N_t,:}\right)\right] \bar{U} = \overset{\vee}{F}, \end{aligned}$$

or

$$\mathrm{A}\bar{U} = \check{F}, \tag{3.4}$$

where $\check{F} = \left[f_{1,1,1}, f_{1,1,2}, \ldots, f_{1,1,N_t1}, f_{1,2,1}, \ldots, f_{1,2,N_t}, \ldots\ldots, f_{N_x-1,N_y-1,N_t}\right]^T$, the notation $[A_M]_{i:j,:}$ represents the intersection of rows from i to j and all columns from any matrix $A_{M\times M}$,and the notation $\widetilde{v}$ is a diagonal matrix with $\{v\}_{i,j,k}$ as diagonal elements for interior points $1 \le i \le N_x - 1, 1 \le j \le N_y - 1$, and $1 \le k \le N_t$.
The system (3.4) represents $(N_x - 1) \times \left(N_y - 1\right) \times N_t$ equations in $(N_x + 1) \times \left(N_y + 1\right) \times (N_t + 1)$ variables, which contain variables at both interior and initial-boundary nodes. As we only interested in the unknown interior points $\left\{\left(x_i, y_j, t_k\right)\right\}_{i,j,k}$, $1 \le i \le N_x - 1, 1 \le j \le N_y - 1$, and $1 \le k \le N_t$, the incomplete system (3.4) can be rewritten as

$$\mathrm{A}\left(\bar{V} + \bar{W}\right) = \check{F}, \tag{3.5}$$

with

$$v_{i,j,k} = \begin{cases} u_{i,j,k} & \forall\ 1 \le i \le N_x - 1,\ 1 \le j \le N_y - 1,\ 1 \le k \le N_t, \\ 0 & otherwise, \end{cases}$$

and

$$\bar{W} = \bar{U} - \bar{V}.$$

After separating the unknown variables from the known ones, the system (3.5) could be rewritten as

$$A_1\bar{\bar{\mathrm{V}}} = \check{K} - A_2\bar{\bar{W}},$$

where $\bar{\bar{\mathrm{V}}}, \bar{\bar{W}}$ represent the columns of the nonzero elements of $\bar{V}, \bar{W}$ and the matrices A_1, A_2 represent the remaining matrices of A after eliminating columns corresponding to zeros elements of $\bar{V}$ and $\bar{W}$respectively.

4. Error estimates

This section is devoted to proving the convergence of the proposed scheme and estimating its accuracy. In order to that we will need the following two lemmas:

Lemma 4.1. *[10] Assume that $(N+1)-$ point Gauss, or Gauss-Radau, or Gauss-Lobatto quadrature formula relative to Legendre weight is used to integrate the product uv where $u \in H^m(\Omega)$ and $v \in P_N$. Then there exists a constant C independent of N such that*

$$\left|(u, v) - (u, v)_N\right| \le CN^{-m} \|u\|_{H^m(\Omega)} |v|_{L_2(\Omega)},$$

where

$$(u, v)_N = \sum_{i=0}^{N} u(x_i)\, v(x_i)\, w_i$$

Lemma 4.2. *(Gronwall inequality) [10] If a non-negative integrable function $u(t)$ satisfies*

$$u(t) \le \alpha(t) + C\int_{-1}^{t} u(s)\, ds, -1 \le t \le 1,$$

where $\alpha(t)$ is an integrable function, then

$$\|u(t)\|_{L^p} \le C\|\alpha(t)\|_{L^p}, p \ge 1 \text{ or } p = \infty.$$

Now, we have the following main result.

Theorem 4.3. *Under the assumptions (A1)-(A5), the solution of problem (1.1)-(1.4) is $u \in L_2(I,V) \cap L_\infty(I,V)$ with $\partial_t u \in L_2(I, H^m(\Omega))$, and there exists a positive constant C independent of u, f, N, m and r such that the estimate*

$$\begin{aligned} &|e_u|^2_{L_\infty(I,L_2(\Omega))} + \|e_u\|^2_{L_2(I,V)} + \|e_u\|^2_{L_\infty(I,V)} \le CN^{-r}\|f\|^2_{L_2(I,H^r(\Omega))} \\ &+CN^{-m}\left(|\partial_t u|^2_{L_2(I,H^m(\Omega))} + \|u\|^2_{L_2(I,H^m(\Omega))}\right) \end{aligned}$$

holds uniformly for any integer numbers $r, m \ge 1$.

Proof. On applying (2.2) at x_i, $0 \le i \le N$, we get

$$\left(\partial_t U^N, \chi\right)_N + \left(\left(U^N, \chi\right)\right)_N = (f, \chi)_N, \quad \forall \chi \in V, \quad a.e.\ t \in I \tag{4.1}$$

where $U^N \in L_2(I;V) \cap C(I;L_2(\Omega))$ Subtracting equation(4.1) from (2.2) yields

$$(\partial_t u, \chi) - \left(\partial_t U^N, \chi\right)_N + ((u, \chi)) - \left(\left(U^N, \chi\right)\right)_N = (f, \chi) - (f, \chi)_N$$

Denoting $e_u = u - U^N$, we get

$$(\partial_t e_u, \chi) + ((e_u, \chi)) = (f, \chi) - (f, \chi)_N - \left(\left(\partial_t U^N, \chi\right) - \left(\partial_t U^N, \chi\right)_N\right) - \left(\left(\left(U^N, \chi\right)\right) - \left(\left(U^N, \chi\right)\right)_N\right). \tag{4.2}$$

Putting $\chi = e_u$ and then $\chi = \partial_t e_u$ in equation (4.2) respectively and summing up we get

$$\begin{aligned}
&\tfrac{1}{2}\partial_t |e_u|^2 + \|e_u\|_V^2 + |\partial_t e_u|^2 + \tfrac{1}{2}\partial_t \|e_u\|_V^2 \\
&= (f, e_u) - (f, e_u)_N - \left((\partial_t U^N, e_u) - (\partial_t U^N, e_u)_N\right) \\
&- \left(\left(\left(U^N, e_u\right)\right) - \left(\left(U^N, e_u\right)\right)_N\right) + (f, \partial_t e_u) - (f, \partial_t e_u)_N \\
&- \left((\partial_t U^N, \partial_t e_u) - (\partial_t U^N, \partial_t e_u)_N\right) \\
&- \left(\left(\left(U^N, \partial_t e_u\right)\right) - \left(\left(U^N, \partial_t e_u\right)\right)_N\right).
\end{aligned}$$

From equation (2.3) and Lemma 4.1

$$\begin{aligned}
&\tfrac{1}{2}\partial_t |e_u|^2 + \|e_u\|_V^2 + |\partial_t e_u|^2 + \tfrac{1}{2}\partial_t \|e_u\|_V^2 \leq \\
&CN^{-r} \|f\|_{H^r(\Omega)} |e_u|_{L_2(\Omega)} + CN^{-r} \|f\|_{H^r(\Omega)} |\partial_t e_u|_{L_2(\Omega)} \\
&+CN^{-m} \left\|\partial_t U^N\right\|_{H^m(\Omega)} |e_u|_{L_2(\Omega)} + CN^{-m} \left\|\partial_t U^N\right\|_{H^m(\Omega)} |\partial_t e_u|_{L_2(\Omega)} \\
&+N^{-m} \left\|U^N\right\|_{H^m(\Omega)} |e_u|_{L_2(\Omega)} + CN^{-m} \left\|U^N\right\|_{H^m(\Omega)} |\partial_t e_u|_{L_2(\Omega)}.
\end{aligned} \tag{4.3}$$

The first two terms on are estimated by the use of Peter-Paul inequality [11] as

$$CN^{-r} \|f\|_{H^r(\Omega)} |e_u|_{L_2(\Omega)} \leq CN^{-r} \left(C_\varepsilon \|f\|_{H^r(\Omega)}^2 + \varepsilon |e_u|_{L_2(\Omega)}^2\right),$$

$$CN^{-r} \|f\|_{H^r(\Omega)} |\partial_t e_u|_{L_2(\Omega)} \leq CN^{-r} \left(C_\varepsilon \|f\|_{H^r(\Omega)}^2 + \varepsilon |\partial_t e_u|_{L_2(\Omega)}^2\right),$$

where ε a very small constant. The application of Young's inequality estimates the third term of (4.3)

$$CN^{-m} \left\|\partial_t U^N\right\|_{H^m(\Omega)} |e_u|_{L_2(\Omega)} \leq CN^{-m} \left(C_\varepsilon |\partial_t u|_{H^m(\Omega)}^2 + \frac{1}{2} |\partial_t e_u|_{H^m(\Omega)}^2 + \frac{1}{2} |e_u|_{L_2(\Omega)}^2\right).$$

Similarly, we estimate the last three terms of (4.3). Finally, choosing ε sufficiently small we get

$$\begin{aligned}
&\tfrac{1}{2}\partial_t |e_u|^2 + \|e_u\|_V^2 + \tfrac{1}{2}\partial_t \|e_u\|_V^2 \\
&\leq CN^{-r} C_\varepsilon \|f\|_{H^r(\Omega)}^2 \\
&+CN^{-m} \left(C_\varepsilon |\partial_t u|_{H^m(\Omega)}^2 + C_\varepsilon \|u\|_{H^m(\Omega)}^2 + |e_u|_{L_2(\Omega)}^2\right).
\end{aligned} \tag{4.4}$$

The integration of equation (4.4) over the interval I with the use of Grönwall's inequality yields the asserted estimate. □

5. Numerical examples

To test our proposed method, we apply the proposed scheme for various examples whose exact solutions are provided in each case. For all examples, we take $N_x = N_y = N_t = N$, to study the convergence behavior of the presented method, we also calculated the following norms [8] for errors with different values of N at $t = T$:

1. The L_2-norm (Frobenius norm) defined for any matrix A by:

$$A_2 = \left(\sum_{i,j} |a_{ij}|^2\right)^{\frac{1}{2}}.$$

2. The L_∞-norm defined for A by

$$A_\infty = \max_{i,j} |a_{ij}|.$$

All the computations are carried out in double precision arithmetic using Matlab7.12.0.635 (R2011a). The code was executed on an Intel Core i7-3610QM, 2.3 GHz Laptop.

Example 5.1. *Consider equation (1.2) with constant coefficients.*

$$\partial_t u - \Delta u = e^t \left(x^2 + y^2 - 4\right), \qquad x, y, t \in [0, 1],$$

with initial and boundary conditions as follows:

$$u(x, y, 0) = x^2 + y^2,$$

$$u(x, 0, t) = x^2 e^t,\ u(x, 1, t) = \left(1 + x^2\right) e^t,$$

$$u(0, y, t) = y^2 e^t,\ u(1, y, t) = \left(1 + y^2\right) e^t.$$

and the exact solution

$$u_{exact}(x, y, t) = e^t \left(x^2 + y^2\right).$$

Table 1 *shows the Max error and Frobenius error at different values of N,* Figure 5.1 *shows the similarity in results of the present Method with the exact solution and surface and contour plots for the error* $u_{ij} - u_{ij}^{ex}$ *at* $t = T$ *respectively.*

Example 5.2. *Consider another example for convection diffusion equation (1.2) with constant coefficients*

$$\partial_t u + \nabla \mathrm{u} - \Delta \mathrm{u} = e^t \left(3\cos(x+y) - 2\sin(x+y)\right), \qquad x,y \in [0,\pi],\ t \in [0,1],$$

with initial and boundary conditions as follows

$$u(x,y,0) = \cos(x+y),$$

$$u(x,0,t) = e^t \cos x,\ u(x,\pi,t) = -e^t \cos x,$$

$$u(0,y,t) = e^t \cos y,\ u(\pi,y,t) = -e^t \cos y.$$

and the exact solution

$$u_{exact}(x,y,t) = e^t \cos(x+y)$$

Table 2 *shows the Max error and Frobenius error at different values of N.* Figure 5.2 *shows the similarity in results of the present Method with the exact solution and the surface and contour plots for the error* $u_{ij} - u_{ij}^{ex}$ *at* $t = T$ *respectively.*

Example 5.3. *Now, consider the following differential equation:*

$$\partial_t u + x^2 \partial_x u + y^3 \partial_y u - xye^{-(x+y)} \partial_{xx} u - e^{-(x+y)} \partial_{yy} u = f(x,y,t), \qquad x,y \in [0,\pi],\ t \in [0,1],$$

where

$$f(x,y,t) = e^t \left(\left(1 + (1+xy)e^{-(x+y)}\right)\cos(x+y) - \left(x^2 + y^3\right)\sin(x+y)\right)$$

with initial and boundary conditions as follows:

$$u(x,y,0) = \cos(x+y)$$

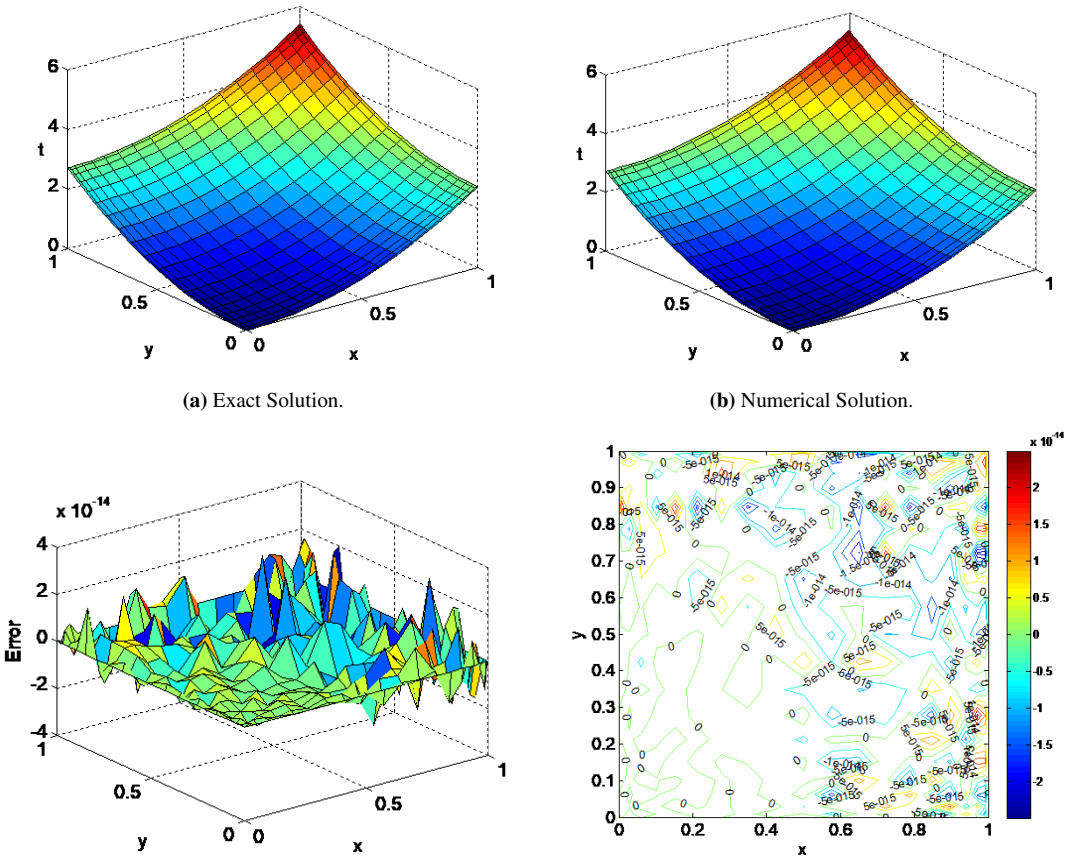
$$u(x,0,t) = e^t \cos x,\ u(x,\pi,t) = -e^t \cos x,$$

$$u(0,y,t) = e^t \cos y,\ u(\pi,y,t) = -e^t \cos y.$$

and the exact solution

$$u_{exact}(x,y,t) = e^t \cos(x+y).$$

then the results are given in Table 3*, again the similarity in results of the present Method with the exact solution, and the surface and contour plots for the error* $u_{ij} - u_{ij}^{ex}$ *at* $t = T$ *are given in* Figure 5.3 *respectively.*



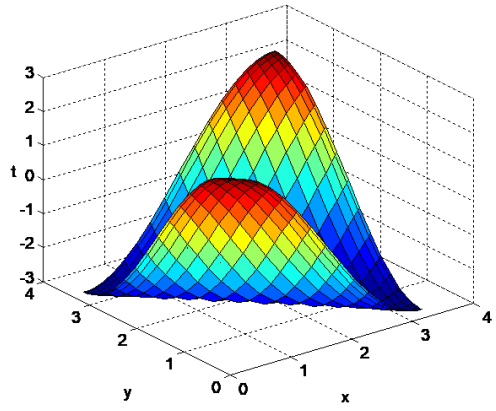
(a) Exact Solution.

(b) Numerical Solution.

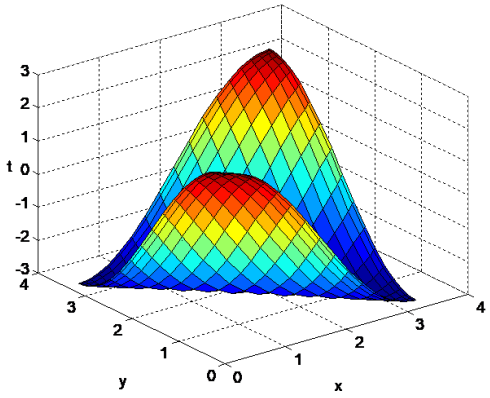
(c) Surface plot for error $u_i j - u_i j^e x$ at $t = T$.

(d) Contour plot for error $u_i j - u_i j^e x$ at $t = T$.

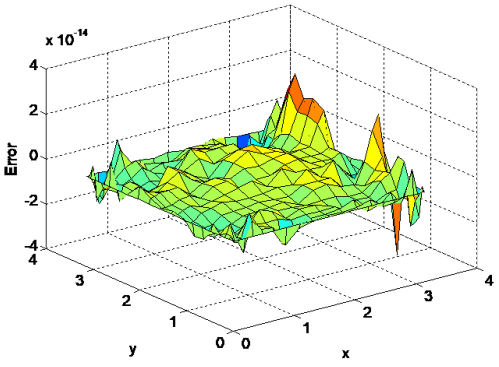
Figure 5.1: Comparison between exact and numerical solutions of Example 5.1 with surface and counter plots of the error.



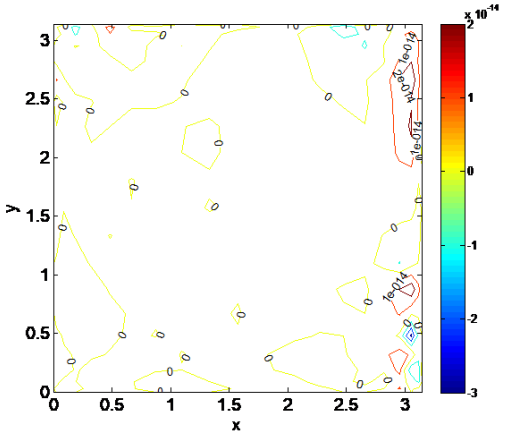
(a) Exact Solution.



(b) Numerical Solution.



(c) Surface plot for error $u_i j - u_i j^e x$ at $t = T$.



(d) Contour plot for error $u_i j - u_i j^e x$ at $t = T$.

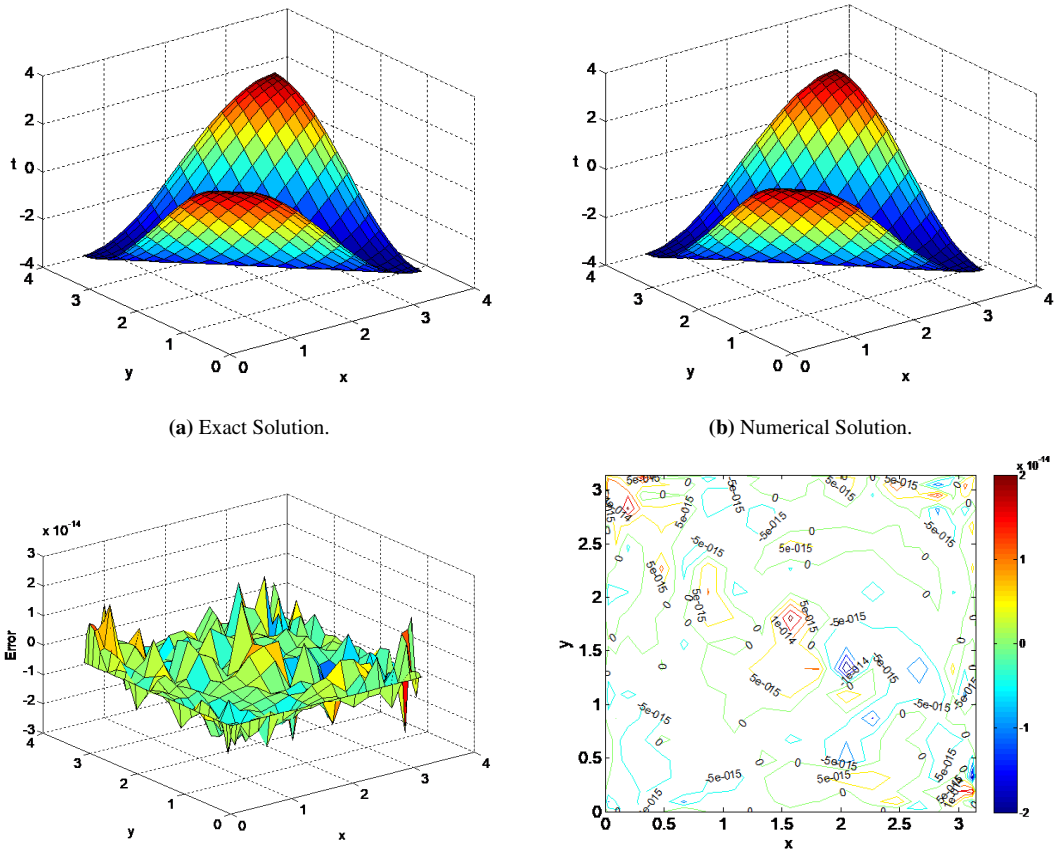
Figure 5.2: Comparison between exact and numerical solutions of Example 5.2 with surface and counter plots of the error.

Table 1: Max error and Frobenius error at different valuse of N for Example 5.1.

N	L_∞	L_2
6	5.1338e-007	1.5710e-006
8	5.7054e-010	2.0571e-009
10	4.2011e-013	1.6188e-012
12	8.8041e-014	3.7847e-013
14	2.1450e-013	7.0891e-013
16	2.5818e-013	1.0988e-012
18	4.8848e-013	1.8580e-012
20	3.7392e-013	2.0660e-012

Table 2: Max error and Frobenius error at different valuse of N for Example 5.2.

N	L_∞	L_2
6	1.2460e-004	2.5773e-004
8	7.5038e-007	2.2982e-006
10	3.4863e-009	1.3228e-008
12	1.1002e-011	5.3229e-011
14	3.0198e-014	1.6745e-013
16	2.9754e-014	9.1839e-014
18	1.9096e-014	7.2120e-014
20	3.1974e-014	1.1562e-013



(a) Exact Solution.

(b) Numerical Solution.

(c) Surface plot for error $u_i j - u_i j^e x$ at $t = T$.

(d) Contour plot for error $u_i j - u_i j^e x$ at $t = T$.

Figure 5.3: Comparison between exact and numerical solutions of Example 5.3 with surface and counter plots of the error.

6. Conclusion

In this paper, Legendre collocation method was applied to calculate approximated solution of 2D Advection-Diffusion Equation with variable Coefficients. By using Kronecker product and modified differentiation matrices with reducing the resulting system, Error analysis and numerical results for this equation show that the suggested method is a high accuracy method.

Table 3: Max error and Frobenius error at different values of N for Example 5.3.

N	L_∞	L_2
6	0.0017	0.0033
8	1.2645e-005	2.9322e-005
10	9.7233e-008	2.9858e-007
12	2.6289e-010	9.5973e-010
14	9.8699e-013	3.5146e-012
16	1.9718e-013	5.7126e-013
18	1.4566e-013	6.6142e-013
20	2.2049e-013	8.2462e-013

References

[1] G.I. El-Baghdady, M.S. El-Azab, *Numerical solution of one dimensional advection-diffusion equation with variable coefficients via Legendre-Gauss-Lobatto time-space pseudo spectral method*, Electron. J. Math. Anal. Appl., **3**(2) (2015), 1-14.
[2] G.I. El-Baghdady, M.S. El-Azab, *Chebyshev-Gauss-Lobatto Pseudo-spectral method for one-dimensional advection-diffusion equation with variable coefficients*, Sohag J. Math., **3**(1) (2016), 1-8.
[3] C. Canuto, A. Quarteroni, *Spectral and pseudo-spectral methods for parabolic problems with non periodic boundary conditions* , Calcolo, **18**(3) (1981), 197-217.
[4] J. Shen, T. Tang, L. Wang, Spectral Methods Algorithms Analysis and Applications, Springer Series in Computational Mathematics, 41, Springer-Verlag Berlin Heidelberg, Germany, 2011.
[5] L. N. Trefethen, Spectral Methods in MATLAB, SIAM, Philadelphia, 2000.
[6] R. Baltensperger, M. R. Trummer, *Spectral differencing with a twist*, SIAM J. Sci. Comput., **24**(5) (2006), 1465-1487.
[7] A. Kufner, O. John, S. Fučik, Function Spaces (Mechanics: Analysis), Noordhoff International Publishing, Netherlands, 1977.
[8] R. A. Horn, Ch. R. Johnson, Matrix Analysis, 2nd ed., Cambridge University Press, UK, 2013.
[9] D. S1. Tracy, R. P. Singh, *A new matrix product and its applications in matrix differentiation*, Stat. Neerl., **26**(4) (1972),143-157.
[10] T. Tang, X. Xu, J. Cheng, *On spectral methods for Volterra integral equations and the convergence analysis*, J. Comput. Math., **26**(6) (2008), 825-837.
[11] J. Rauch, Partial Differential Equations, Springer-Verlag, New York Inc., USA, 1991.

Journal of Mathematical Sciences and Modelling, 3 (2) (2020) 95-101
Research Article

Journal of Mathematical Sciences and Modelling

Journal Homepage: www.dergipark.gov.tr/jmsm
ISSN 2636-8692
DOI: http://dx.doi.org/10.33187/jmsm.680161

Mathematical Determination of the Cultural Interaction between Medieval Groups

Mehmet Erbudak[1]

[1]*Laboratorium für Festkörperphysik, ETHZ, CH-8093 Zurich and, Physics Department, Boğaziçi University, TR-Istanbul*

Article Info

Keywords: *Correlation, Medieval architecture, Ornaments, Symmetry, Tessellation, Wallpaper groups*
2010 AMS: *20H15, 62H20, 82D25.*
Received: *26 January 2020*
Accepted: *5 June 2020*
Available online: *31 August 2020*

Abstract

A mathematical classification of two-dimensional ornaments into 17 plane symmetry groups is presented, which were created by five medieval cultural groups of Middle East. The data are considered representative for the cultural groups. By applying a correlation algorithm on the individual use of symmetry classes by each cultural group, the strength of the interaction between the pairs of groups are quantitatively determined. The analysis shows that the strongest similarity in the creation of periodic ornaments is between Rum Seljuks and Arab Muslims and between Armenians and Byzantium. It is also found that the Rum Seljuks, followed by Armenians, are the most interactive cultures. This report is the first attempt to quantify cultural communication by mathematical means.

1. Introduction

It seems justified to claim that the region of the Middle East, the so-called *fertile crescent*, is the cradle of our civilizations [1]. Evidence to support this claim is constantly being discovered. I find that there is a region of *cultural exchange* somewhere north of the fertile crescent where prominent cultures have interacted. Here I present arguments to substantiate this claim qualitatively and quantitatively.

The symmetry properties of planar ornaments are characteristic of an ethnic group that produced them [2]. We classify the ornaments according to their symmetries using group theoretical methods of mathematics. In analogy to surface crystallography, this results in 17 symmetry groups [3]. To reach our goal, we consider artistic ornaments of five individual cultural groups that lived in and around the geographical region of interest.

The comparison of the ornamental features of two groups shows the degree of their cultural similarity or even cultural interaction. To qualitatively determine the interaction, I apply a correlation algorithm to each pair of symmetry distributions, and I obtain a similarity function for the degree of interaction between two groups under consideration. I apply the procedure to all pairs of groups. This operation is the new idea of the present work to rigorously evaluate cultural similarities.

The first cultural establishment one encounters in this region is Armenia, which extends from modern Armenia to the Van Lake plateau in the south and further southwest to Cappadocia and Cilicia [4]. There are four other groups of people who come and go to the region of cultural exchange after the year 700 AD.

I mention the Eastern Roman Empire [5], later called Byzantium [6]. Emperor Constantine declared Constantinople the capital of his empire after his victory at Milan. The Eastern Roman Empire expanded to the East, conquered Anatolia, today's Syria and its bordering territories. The Byzantine capital is home to at least two masterpieces, the Hagia Sofia and the Chora Church. The Adriatic coast of Croatia and Italy were already in the possession of the Empire. Inspired by Christianity, Byzantium created other breathtaking examples of its culture: the Basilica of San Marco in Venice, the Church of San Vitale in Ravenna, the Cathedral and Monreale Monastery near Palermo, to name but a few. The map in Figure 1 shows schematically the Byzantine territories with their capital (encircled) at the intersection of Europe and Asia.

After the emergence of Islam around 620, the peoples of the Arabian peninsula reached their military and cultural peak during the Umayyad and Abbasid dynasties, followers of the Muslim prophet. They first moved north and northeast to spread their confession by asserting their language and culture [7].

The Great Seljuks are Turkic tribes from the prairies of Central Asia who settled in Iran, especially in Khorasan and later in the southern regions of the Caspian Sea [8, 9, 10]. On their way west they met Byzantine powers to defeat them once and for all. After their own

***Email addresses and ORCID numbers*: erbudak@phys.ethz.ch, https://orcid.org/0000-0002-6316-5115 (M. Erbudak)**

dissolution due to lack of leadership, their descendants founded a culturally high-ranking sultanate, the Seljuks of Anatolia or the Rum Seljuks [11, 12]. In Turkish usage *Rum* stands for *Romans* or simply *West*. The Seljuks are responsible for the spread of the Turkish language in Western Iran, the Caucasus, Iraq, Syria, Anatolia, etc.

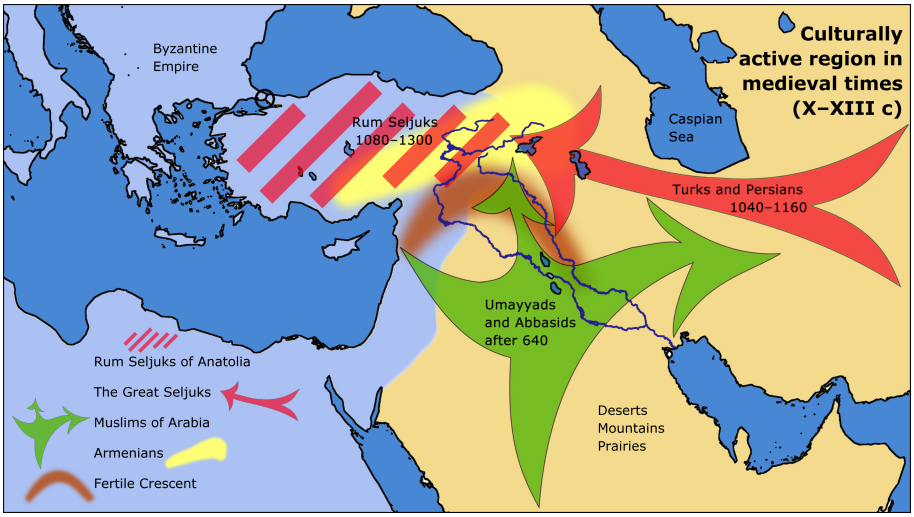


Figure 1: The Middle East, the Balkans and Italy, including parts of Arabia, during the greatest expansion of the Byzantine Empire (light blue) in the 10th to 13the centuries. The circle marks their capital Constantinople. The only stationary civilization is that of the Armenians. Byzantium shrinks under the influence of the Turkish-Persian Seljuk Empire, which moves west. Muslims from Arabia advance north and northeast in their missionary expansion.

It is like a historical play. We have the cast, the stage and the program from about the tenth to the end of the thirteenth century. Below are some details about the cast and their artistic achievements, before I go into the cultural interactions. These cultural groups overlapped geographically and in time, or at least were neighbors. Therefore they influenced each other in their artistic habits.

In analogy to surface crystallography, we classify the ornaments according to their symmetries using group-theoretical methods of mathematics. We obtain 17 symmetry groups [3]. I apply this method to artistic ornaments of five individual cultural groups living in the geographical region under investigation. I record the occurrence of each symmetry group for these five cultures and visually evaluate the similarities and differences.

The comparison of the ornamental features of two groups shows the degree of cultural similarity and possibly their cultural interaction. To determine the degree of interaction, I apply a correlation algorithm to each pair of symmetry distributions, which yields a similarity function. I apply this procedure to all pairs of groups.

The classification of the ornament symmetries into 17 crystallographic groups is not a new achievement. Nevertheless, it is objective, rigorous and easily repeatable, anywhere and at any time. The new achievement of this work consists of the idea and the method of how I compare the results of the classification and obtain a similarity function for the correlation of each pair of cultural groups. Since I have five different groups when I consider the interaction of each group with its four neighbors, I end up with ten correlation functions. This is a mathematical procedure that is applied to the established classification scheme for the first time and allows a quantitative measurement of mutual interactions. I find that cultural groups that interacted most with its neighbors were the Rum Seljuks and Armenians, while the Great Seljuks developed their works of art quite independently.

2. Mathematical classification of ornaments

The creation of two-dimensional ornaments in art and architecture is an artistic achievement. The ornament is a two-dimensional surface decoration that covers the surface completely and periodically. For a given ornament, I first determine the *unit cell*, the smallest surface that is periodically repeated in two dimensions to form a lattice. The *translation* operation to cover the entire surface is consistent with the symmetry properties of the unit cell. I then study the *rotational* symmetries. A given shape is either two-, three-, four- or sixfold symmetric ($n = 2, 3, 4, 6$) or has no such rotational property ($n = 1$). These numbers refer to the fraction of a complete circle around which an object is rotated to be mapped onto itself. After determining the rotational symmetries, I look for reflection symmetries in two main directions. Again, a *mirror* symmetrical object is imaged onto itself when it is reflected at a mirror symmetry line. A *glide* symmetry can practically be compared to foot steps left on sand on both sides of the glide reflection line. This scheme is the basis of the plane symmetry group, a formalism which is exactly the same as that we use to describe the atomic order on a crystal surface [13].

The formalism of crystal groups was proposed at the end of the 19th century, when mathematicians used crystallography to study the symmetry problem and found that, according to group theory, there are 17 different types of periodic distribution of atoms on a surface. These crystal groups are exactly the same as those I obtain for a culturally created artistic ornament. This assumption was made by George Pólya in 1924 [14] and is proved by Edith Müller in her dissertation which examines the ornaments of the Alhambra Palace in Granada, Spain [15]. Since then, there have been several fundamental papers on this subject [3], mainly based on the results of Müller.

The 17 distinct translational planar patterns can be characterized according to the convention of crystallographers and physicists [3]. The ornamental patterns are also called the *wallpaper group*. The designation of each group follows a four-digit recipe: The first letter is a *p* or a *c*, *primitive* or *centered* depending on the type of lattice. The letter *c* is used if there is an additional unit in the centre of the cell. The second digit is the rotational order *n*. The third and fourth digits are a mirror reflection *m* or glide reflection *g* along the two major directions. If there is no mirror or glide, we write 1.

3. Results

I have classified two-dimensional periodic ornaments created by five medieval societies, the Byzantine Empire, the Armenians, the Muslims of Arabia, the Great Seljuks and the Rum Seljuks. The results of this classification are presented below in graphic form.

3.1. Byzantium

Byzantium has created numerous architectural monuments with a generous collection of ornaments. They belonged to the *Eastern Orthodox Nomination* and founded churches wherever the Empire ruled. Nevertheless, I could not find a statistical relevant number of ornaments in their main church Hagia Sofia nor in Chora and San Vitale. Therefore, I concentrated on the floor ornaments of the Basilica San Marco in Venice [16]. On the marble floor of the basilica there are at least 730 different individual areas, which are decorated with periodically arranged marbles. The frequency of the plane groups is shown in Figure 2. In my analysis I found that almost 80 % of the ornaments have a fourfold symmetry. Interestingly there are no ornaments with threefold symmetry.

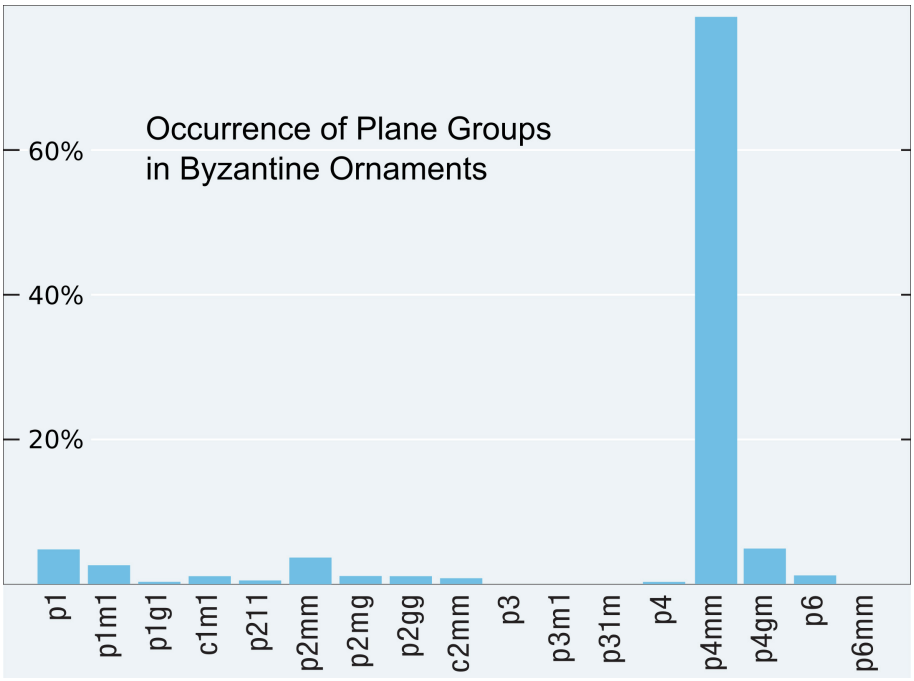


Figure 2: The frequency of the symmetry group in the floor ornaments of the 10th century built St. Marc's Cathedral in Venice. The results are based on 730 individual ornaments and come from Ref. [16].

Byzantium did not last long. In 1071 a battle against Alparslan of the Great Seljuks took place in Manzikert near Lake Van, which initiated the collapse of the Eastern Roman Empire until they lost their capital to the Ottomans in 1453.

3.2. Armenia

In the Middle Ages Armenia was mainly ruled by nobles. Two important families were the Bagratuni in the north and the Artsrunik in the south, from which Vaspurakan emerged as the most powerful kingdom in Armenia. It has suffered from Islamic and Byzantine aggressions and constant civil wars. But the mighty Gagik I was recognized by the Arabs and Byzantium as the *King of Armenia* and proved that their interactions were not always hostile. Armenians were even appointed Byzantine emperors and were occasionally married to Arabs. The traditional Bagratuni capital Duin was destroyed by a devastating earthquake in 893. Trdat was the brilliant architect who built Ani, the next capital. Trdat was even invited by the Byzantine Emperor Basil II to rebuild the dome of Hagia Sofia after it was also destroyed by another devastating earthquake in 989 [4].
The Armenians confess to *Oriental Orthodoxy*. They have created numerous churches with stone carvings on the walls and khachkars, gravestones. We have recently studied Armenian ornaments and classified them [17]. Apart from the fourfold symmetry found in the majority, the Armenian ornaments have a balanced distribution. Figure 3 displays the results.

3.3. Muslims of Arabia

The next group of people I consider are the Muslims from the Arabian Peninsula, Egypt and Syria after they converted to Islam in 622. Around 680, the Umayyad and Abbasid dynasties moved north to the Armenian countries and east to the Iranian territories. In Iran they met the Persian, Turkish and Mongolian peoples and converted them to Islam. They also brought their cultures and together with the new Muslims they created several buildings such as mosques, minarets and madrasas. I found examples of the cultural habits that originated from Syria and Egypt and were brought by the Arab-Muslim occupation [19]. Arab warriors also moved west along the North African coast in the early Umayyad period and established a great culture on the Iberian peninsula. The Umayyads of Andalusia lasted more than eight centuries and shaped the European civilization.

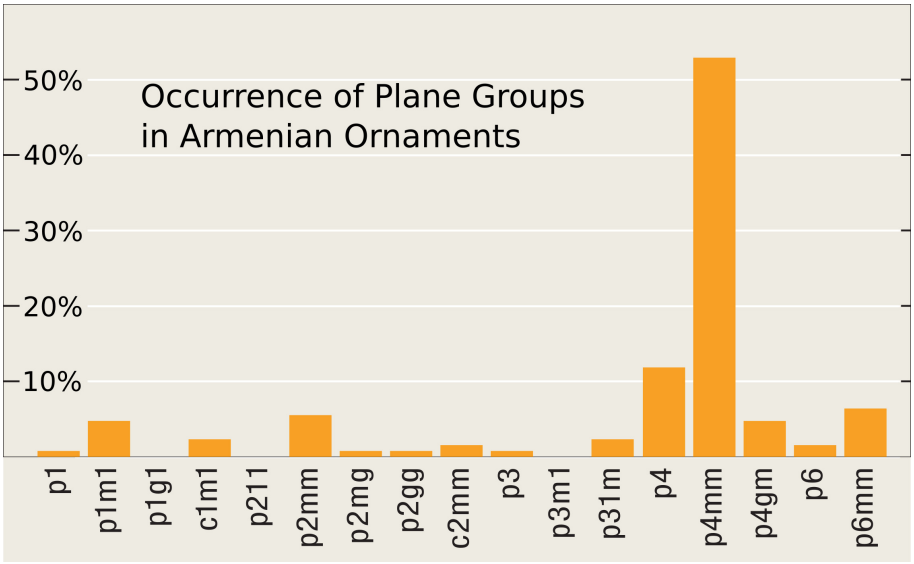


Figure 3: The plane groups of Armenian ornaments from the personal collection of Armen Kyurkchyan [18]. The results are based on 123 individual ornaments.

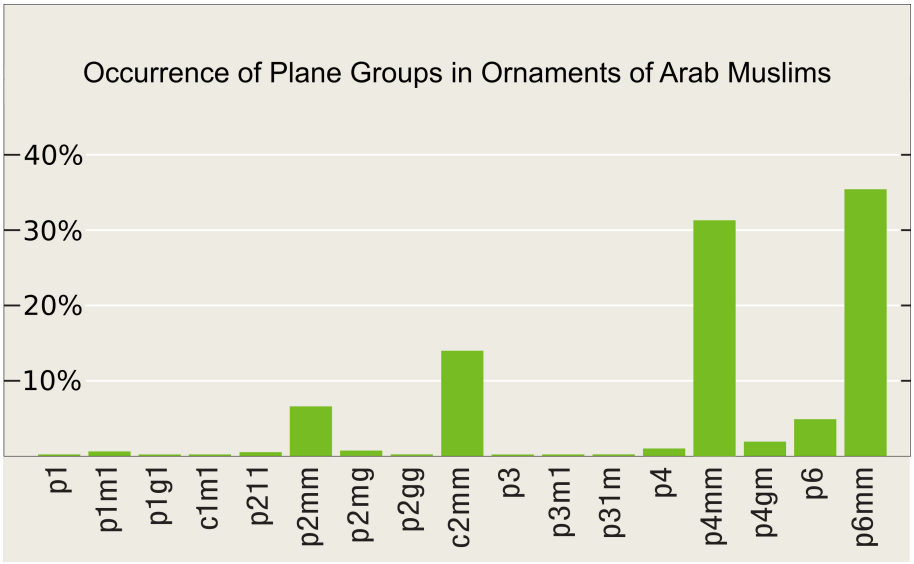


Figure 4: Arab-Muslim ornaments from Syria and Egypt [19, 20]. For the classification 200 drawings are used.

Bourgoin published drawings of more than 200 Arab-Muslim ornaments without mathematical classification [19]. Later in the 1970s two groups, Emil and Milota Makovicky and Syed Jan Abas and Amer Shaker Salman, studied these ornaments according to the group-theoretical recipe [20]. Here I use the average of their results to compensate for smaller differences and show them in Figure 4.

3.4. The Great Seljuks

Another culture that has left its mark in this geography are the Great Seljuks. They are Turkish nomad shepherds from the prairies of Central Asia, descendants of the *Göktürk* tribes. Since the 8th century they spread steadily westward to the Aral Sea and fought against the invading Islamic Arabs until their tribal leader Seljuk Bey took over Islam about 960. His sons first moved from Khorasan to Afghanistan around 1030 and then further west. They conquered Shiraz in 1050 before occupying all of Iran and Baghdad, the seat of the Abbasid caliphs. As Sunni muslims, they fought against the Shiite Fatimids of Cairo. Still as nomads, after 1060 they carried out campaigns with Christians on the borders of the empire in Anatolia and the Caucasus to plunder the territories. Finally, in 1071, Alpaslan ravaged Byzantium in Manzikert and captured Emperor Romanus Diogenes. One year later Alpaslan moved to the east to his residence in Isfahan.

After the assassination of Alpaslan in 1072, fratricidal wars between his descendants in 1150 broke the empire apart into a complete anarchy. The Great Seljuk Empire was a composition of Turks, who were the best fighters, and the Persians, who were administratively the best organizers. It is difficult to distinguish clearly between the Persians and Seljuks in their artistic achievements. They are certainly strongly influenced by Islam, as they built mosques and huge minarets in addition to the madrasas. I have collected the ornamental material from various sources [7, 9, 21, 22]. An earlier analysis was limited to about 30 ornaments from the photographic collection of Emil Makovicky [21]. The results of plane groups shown in Figure 5 contain additional 55 ornaments. The high value of the *c2mm* group is a result of the brick-laying technique used by Seljuks on minarets.

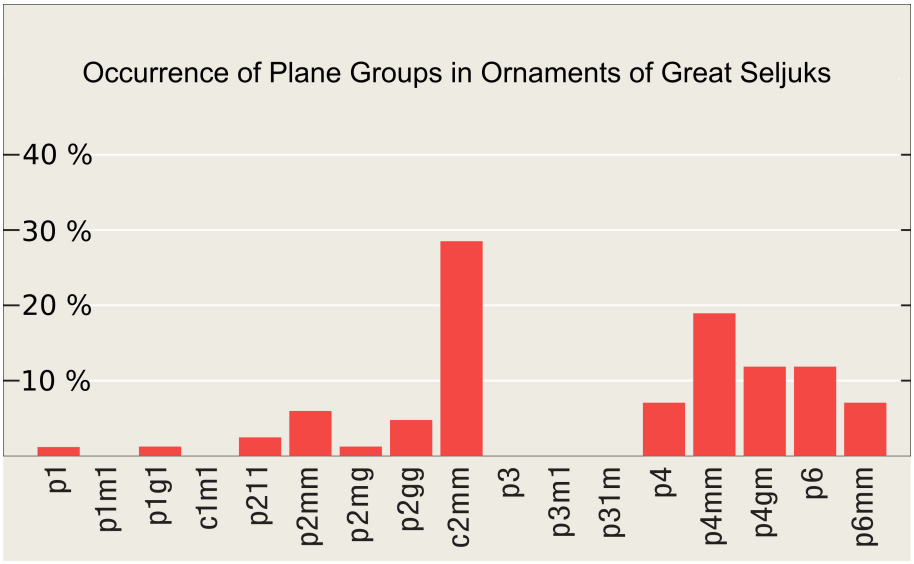


Figure 5: The plane symmetry groups of the ornaments of the Great Seljuk Empire mainly found in Iran. The classification is based on 84 individual objects.

3.5. The Rum Seljuks

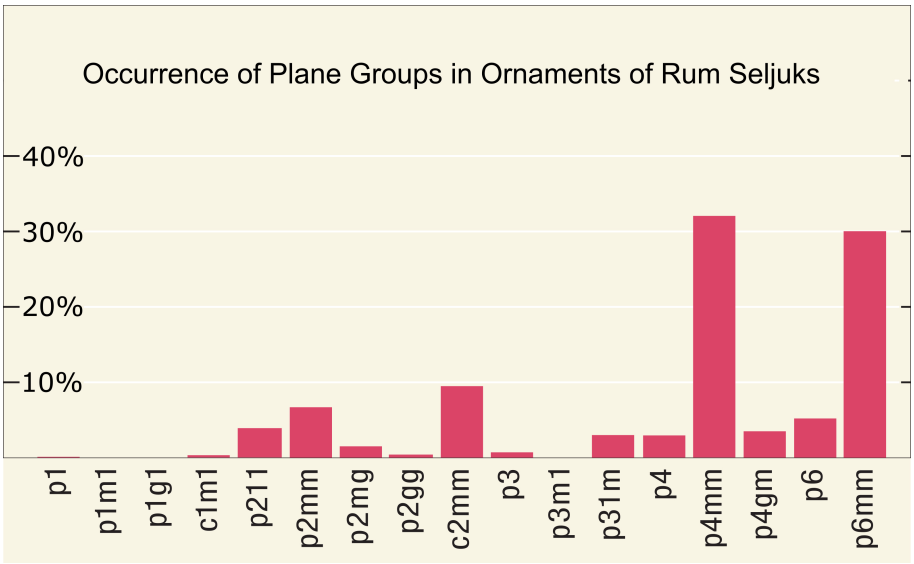


Figure 6: Classification of ornaments of the Sultanate of Rum Seljuks. The results are based on hand drawings by Gerd Schneider [12] and represent 1067 artifacts.

Seljuk warriors, mostly Turkmen tribes, infiltrated Anatolia after 1000. After the Battle at Manzikert in 1071 and the captivity of the Byzantine emperor a civil war broke out on Byzantine territory. The Turkic tribes took advantage of the unrest and founded their own sultanate, the Rum Seljuks. With the exception of the coastal regions in the west, they occupied most of Anatolia. Their first capital was Nicaea in the northwest; which was later conquered by the First Crusades in 1097. Konya became the next capital.
The Rum Seljuks have built great mosques, palaces, caravanserais, mausoleums, madrasas and shipyards. They existed until the impact of Mongol rule during the second half of the 13th century, when it was the Ilkhans, descendants of Genghis Khan, who took over their land.
The best documentation about the Sultanate's work of art can be found in the monumental work of Gerd Schneider, which he created during his several years' stay in Istanbul [12]. Here I present in Figure 6 the results obtained from his meticulous hand drawings.

4. Discussion

4.1. Visual inspection

I first evaluate the results by visual inspection of the Figures 2-6. I find that almost all cultures prefer the fourfold symmetrical *p4*, *p4mm* or *p4gm*. I can only speculate about the origin for this preference. In the case of ornaments carved on building materials, the reason could be the simple manufacture of each masonry block [17]. The reason could also lie in a religious practice, if the ornaments are found in a church, mosque, madrasa or monastery [16]. Threefold symmetry is quite rare in any repertoire. I note that both Arab Muslims and Rum Seljuks have predominantly used ornaments with double mirror reflections on sixfold, fourfold and twofold symmetries. The distribution of the

Groups	Armenia	Byzantium	M. Arabs	G. Seljuks	R. Seljuks
p1	0.8	4.8	0.2	1.2	0.1
p1m1	4.9	2.6	0.6	0	0
p1g1	0	0.3	0.2	1.2	0
c1m1	2.4	1.1	0.2	0	0.3
p211	0	0.5	0.5	2.4	0.5
p2mm	5.7	3.7	6.6	6.0	6.7
p2mg	0.8	1.1	0.7	1.2	1.5
p2gg	0.8	0.1	0.2	4.8	0.4
c2mm	1.6	0.8	14.0	28.6	9.5
p3	0.8	0	0.4	0	0.7
p3m1	0	0	0.4	0	0
p31m	2.4	0	1.4	0	3.0
p4	12.2	0.3	1.0	7.1	2.94
p4mm	54.5	78.5	31.4	19.0	32.1
p4gm	4.9	4.9	1.9	11.9	3.5
p6	1.6	1.2	4.9	11.9	5.2
p6mm	6.6	0	35.5	4.8	30.1

Table 1: Occurrence of the individual symmetry groups (in percentage) in the artworks of Armenia, Byzantium, Muslim Arabs, Great Seljuks and Rum Seljuks. The rotational symmetries are grouped together.

plane groups looks remarkably similar, indicating an intensive interaction of cultures between Arab Muslims and Rum Seljuks.
Table 1 summarizes the occurrence of each symmetry group in five different cultures. We observe that ornaments with a rotational symmetry of 1 and 3 are rare overall. Threefold symmetrical ornaments are completely missing in Byzantium and the Great Seljuks. The fourfold symmetry is by far the most favorable in all groups. It is obvious that the *interlacing* in planar ornaments reduces the symmetry. The published results about the Arab-Muslim group neglect the interlacing [19]. To allow a comparison based on the same assumption, I have used the Armenian and Seljuk results after I have similarly neglected interlacing.

4.2. Mathematical correlation

In order to rigorously determine the efficiency of the interaction and the resulting correlation of the artwork between the five cultural groups considered here, I use a scientific approach. In the social sciences the Pearson correlation coefficient is used for metric variables [23]. To investigate overlapping elements of variables in natural sciences *cosine similarity* is sometimes preferred [24]. In physics and chemistry, *overlap integrals* are used [25].
I determine the correlation of the results considering the product of two *vectors* X and Y with 17 components x_i and y_i each represent the elements of the plane groups. This correlation function is a robust measure of the strength of pairwise interaction between two groups and quantifies the similarity between these cultural groups. I repeat the operation for all the members of the plane group:

$$C = \frac{\vec{X}}{|X|} * \frac{\vec{Y}}{|Y|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}} ,$$

while i runs between 1 and 17, the number of plane groups. The sums over x_i and y_i are normalized to one. This metric is often used to assess the correlation of partially overlapping systems.

C	A	B	AM	GS	RS
A	1	0.965	0.713	0.576	0.783
B	0.965	1	0.636	0.512	0.705
AM	0.713	0.636	1	0.646	0.988
GS	0.576	0.512	0.646	1	0.649
RS	0.783	0.705	0.988	0.649	1

Table 2: Interaction factor C between 1 and 0, based on the occurrence of plane-group symmetries in the artwork from 5 medieval groups. A stands for Armenia, B for Byzantium, AM for Arab Muslims, GS for Great Seljuks and RS for Rum Seljuks.

To find the interaction C in pairs between the two cultures, I use the above formula and obtain ten values for the correlation of Armenia, Byzantium, Arab Muslims, Great Seljuks and Rum Seljuks. The results are presented in Table 2. A high (low) value of C near 1 (near 0) indicates a strong (weak) correlation, i.e., interaction, between the cultures X and Y. There are several ways to interpret the values of C. The easiest way allows us to determine the strong interaction between certain two cultures. We see that the strongest interaction factor is between $0.96 - 0.99$ and is found for Armenia and Byzantium as well as for Muslim Arabs and Rum Seljuks. Similarly, I also add the correlation coefficients of each group with the other four groups and repeat this summation for all five groups. The sum obtained is a measure of the ability of the group to interact culturally or architecturally. A high value stands for a strong creativity of the group, while a lower correlation value C is rather modest. The Great Seljuks were warriors and nomads. In their sultanate, the Persian people were responsible for civil

affairs and administration. According to this analysis, the Great Seljuks were the least interacting group. I find that the groups that interact most with their neighbors are Rum Seljuks, Armenians and the Muslim Arabs.

5. Conclusions

The comparison of the frequent use of particular symmetries in ornaments of ethnic groups could be a useful indication of their interaction and mutual artistic influence. The similarity between two groups of the same religious belief is likely to lead to similar habits. In the present case, however, the Byzantine Empire and the Armenians are both Christians, but from different churches according to the Calcedonian Schism 451. Seljuks and the Umayyads or Abbasids are all Muslims, but with fundamentally different beliefs about the succession of the Prophet Mohammad. Therefore, religion gives us no indication of the artistic habits of the peoples concerned. Nor is the geography of their homeland similar; it differs from the mountainous regions of the Caucasus, the Seljuk prairies, to vast Arab desserts, and the coastal regions of the Byzantine Empire. What they have in common is their campaign through the *region of cultural activity* in the eastern part of present-day Turkey. Yet the exact occurrence of these cultures is geographically and chronologically limited. However, the activity of craftsmen was not limited by national borders. Even if there was a temporal shift in the occurrence of cultures, this fact only determines the direction of the flow of information.

I made several assumptions. Firstly, I have taken the symmetry properties of ornaments as a cultural indication of an ethnic group and neglected all other ornaments that have no symmetry. In order to quantify the symmetries unambiguously, I applied the crystallographic plane-group classification. I express the results as percentage frequencies of 17 plane groups each, which is a scientific way to evaluate the ornaments. I apply this method to five individual cultural groups that were neighbors at some time in the Middle Ages, namely the Armenians, the Byzantine Empire, the Muslim Arabs, the Great Seljuks and the Rum Seljuks.

In order to estimate the mutual interaction of these five groups, I have worked out an interaction coefficient C for each ethnic group with the other four groups. This factor quantitatively describes the artistic similarities by calculating the overlapping of the frequency of the symmetries. The technique is common in many other areas of science; it is used in this context for the first time. Mathematically, a cultural group is a vector and its symmetry groups are the 17 components of the vector. The correlation factor is then the overlapping of the components for two cultural groups under consideration. I compare these values of each pair of groups to determine those two cultures that have influenced each other the most. Here is the strongest interaction between Byzantium and Armenia and between Muslim Arabs and Rum Seljuks. I also add up the C values of each cultural group and find that the Rum Seljuks and Armenians have influenced their neighbors the most and vice versa.

The Byzantine Empire diffused into European territories after the siege of Constantinople by the Ottomans. It is conceivable that Byzantine craftsmen made a positive contribution to the flourishing Renaissance. The Great Seljuks disintegrated into smaller emirates under the influence of the Mongol invasion, while Rum Seljuks survived in Anatolia in the background for several centuries. The art of the Muslim Arabs reached its peak in various places, in Damascus with the Umayyads, in Baghdad among the Abbasids, and on the Iberian peninsula with the Umayyads of Andalusia. The Armenian civilization created artistic masterpieces in the Middle Ages and throughout history until today. They are the only group that has remained on their original Caucasian planes.

Acknowledgements

I would like to thank Prof. Dr. Alphan Sennaroğlu for his expert suggestions for the calculation of correlation functions and Fatma Erbudak for her support at every stage of this study. Both contributions have been very fruitful. This research has not funded by any agency.

References

[1] C. Brinton, J. B. Christopher, R. L. Wolff, *A History of Civilization* Vol. I, Prentice Hall, Englewood, NJ, 1960.
[2] D. K. Washburn, D. W. Crowe, *Symmetries of Culture: Theory and Practice of Plane Pattern Analysis*, University of Washington Press, Washington DC, 1991.
[3] D. Schattschneider, *The plane symmetry groups: Their recognition and notation*, Amer. Math. Monthly, **85**(6) (1978), 439-450.
[4] L. Jones, *Between Islam and Byzantium*, Ashgate, Burlington, VT, 2007.
[5] J. J. Norwich, *Byzantium, The Early Centuries, The Apogee, The Decline and Fall*, Alfred A. Knopf, NY, 1997.
[6] A. E. Kaldellis, *Romanland: Ethnicity and Empire in Byzantium*, Harvard University Press, Cambridge, MA, 2019.
[7] M. Hattstein, P. Delius, *Islam, Kunst und Architektur*, Könemann, Köln, 2000.
[8] S. R. Canby, D. Beyazit, M. Rugiadi, A. C. S. Peacock, *Court and Cosmos, The Great Age of the Seljuqs*, The Metropolitan Museum of Art, NY, 2016.
[9] A. Hutt, L. Harrow, *Iran 1*, Scorpion Publications, London, 1977.
[10] A. Başan, *The Great Seljuks: A History*, Routledge, London, 2010.
[11] S. Mecit, *The Rum Seljuqs: Evolution of a Dynasty*, Routledge, London, 2014.
[12] G. Schneider, *Geometrische Bauornamente der Seldschuken in Kleinasien*, Dr. Ludwig Reichert, Wiesbaden, 1980.
[13] C. Hammond, *The Basics of Crystallography and Diffraction*, 4. Ed., Oxford Univ. Press, Oxford, 2015.
[14] G. Pólya, *Über die analogie der kristallsymmetrie in der ebene*, Z. Kristall., **60** (1924), 278-282.
[15] E. A. Müller, *Gruppentheoretische und strukturanalytische Untersuchung der Maurischen Ornamente aus der Alhambra in Granada*, Ph.D. Thesis, University of Zurich, 1944; *El estudio ornamentos como applicatión de la teoría de los grupos de orden finito*, Euclides (Madrid), **6** (1946), 42-52.
[16] M. Erbudak, *Symmetry analysis of the floor ornaments of the San Marco Cathedral in Venice*, Heliyon, **5** (2019), e01320.
[17] M. Erbudak, A. Kyurkchyan, *Armenian, Byzantine and Islamic Ornaments, Influences Among Neighbors*, Kyurkchyan, Yerevan, 2019, doi.org/10.3929/ethz-b-000394011.
[18] A. Kyurkchyan (Author), H. H. Khatcherian (Photographer), *Armenian Ornamental Art*, Craftology, Yerevan, 2010.
[19] J. Bourgoin, *Les Èlèments de L'Arabe*, Librairie de Firmin-Didot, Paris, 1879; *Arabic Geometrical Pattern and Design*, Dover, New York, 1973.
[20] E. Makovicky, M. Malovicky, *Arabic Geometrical Patterns – A Treasury for Crystallographic Teaching*, Neues Jahrbuch für Mineralogie Monatshefte, **2** (1977), 58-68; S. J. Abas, A. S. Salman, *Symmetries of Islamic Geometrical Patterns*, World Scientific, Singapore, 1995.
[21] E. Makovicky, *Symmetry*, De Gruyter, Berlin, 2016.
[22] J. Bonner, *Islamic Geometric Patterns*, Springer, New York, 2017.
[23] Available at https://en.wikipedia.org:Pearson_correlation_coefficient.
[24] J. Ye, *Cosine similarity measures for intuitionistic fuzzy sets and their applications*, Math. Comput. Modelling, **53** (2011), 91-97.
[25] K. Rudenberg, K. O-Ohata, D. G. Wilson, *Overlap integrals between atomic orbitals*, J. Math. Phys., **7** (3) (1966), 539-546.