ACTA
UNIVERSITATIS PALACKIANAE
OLOMUCENSIS

FACULTAS RERUM NATURALIUM
2009

ACTA
UNIVERSITATIS PALACKIANAE
OLOMUCENSIS

FACULTAS RERUM NATURALIUM
MATHEMATICA 48 (2009)

MATHEMATICA 48

PALACKÝ UNIVERSITY OLOMOUC
OLOMOUC 2009

ISBN 978-80-244-2386-9
ISSN 0231-9721

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009)

CONTENTS

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 7–16

On the Measurability of Sets of Pairs of Intersecting Nonisotropic Straight Lines of Type Beta in the Simply Isotropic Space

Adrijan Varbanov BORISOV [1], Margarita Georgieva SPIROVA [2]

[1] *Department of Mathematics, South-West University,*
"Neofit Rilski" 66, Ivan Mihailov Str., 2700 Blagoevgrad, Bulgaria
e-mail: adribor@aix.swu.bg

[2] *Fakultät für Mathematik, TU Chemnitz*
D-09107 Chemnitz, Germany
e-mail: margarita.spirova@mathematik.tu-chemnitz.de

(Received October 24, 2008)

Abstract

The measurable sets of pairs of intersecting non-isotropic straight lines of type β and the corresponding densities with respect to the group of general similitudes and some its subgroups are described. Also some Crofton-type formulas are presented.

Key words: Simply isotropic space, density, measurability.

2000 Mathematics Subject Classification: 53C65

1 Introduction

The *simply isotropic space* ${I_3}^{(1)}$ (see [8]) is defined as a projective space $\mathbb{P}_3(\mathbb{R})$ in which the absolute consists of a plane ω (the *absolute plane*) and two complex conjugate straight lines f_1, f_2 (the *absolute lines*) within ω. In homogeneous coordinates (x_0, x_1, x_2, x_3) we can choose the plane $x_0 = 0$ as the plane ω, the line $x_0 = 0$, $x_1 + ix_2 = 0$ as the line f_1, and the line $x_0 = 0$, $x_1 - ix_2 = 0$ as the line f_2. Then the intersecting point F of f_1 and f_2, which is called an *absolute point*, has coordinates $(0, 0, 0, 1)$. All regular projectivities transforming the absolute figure into itself form the 8-parametric group G_8 of *general simply*

isotropic similitudes. In affine coordinates (x, y, z) with respect to the affine coordinate system $(O, \overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3})$, any similitude of G_8 can be written in the form ([8, p. 3])

$$\begin{aligned} \overline{x} &= c_1 + c_7(x\cos\varphi - y\sin\varphi), \\ \overline{y} &= c_2 + c_7(x\sin\varphi + y\cos\varphi), \\ \overline{z} &= c_3 + c_4 x + c_5 y + c_6 z, \end{aligned} \tag{1}$$

where $c_1, c_2, c_3, c_4, c_5, c_6, c_7$, and φ are real parameters and $c_7 > 0$.

A plane in ${I_3}^{(1)}$ is said to be *non-isotropic* if its infinite line is not incident with the absolute point F; otherwise the plane is called *isotropic*.

A straight line in ${I_3}^{(1)}$ is said to be *(completely) isotropic* if its infinite point coincides with the absolute point F; otherwise the straight line is said to be *non-isotropic* ([8, p. 5]).

Let G_1 and G_2 be two non-isotropic straight lines and let us denote by U_1 and U_2 their infinite points, respectively. The straight lines G_1 and G_2 are said to be *of type* β if the points U_1, U_2, and F are collinear; otherwise the straight lines are said to be *of type* α ([8, p. 45]).

We will consider also the following subgroups of G_8:

I. $B_7 \subset G_8 \iff c_7 = 1$. This is the group of simply isotropic similitudes of the *δ-distance* ([8, p. 5]).

II. $S_7 \subset G_8 \iff c_6 = 1$. This is the group of simply isotropic similitudes of the *s-distance* ([8, p. 6]).

III. $W_7 \subset G_8 \iff c_6 = c_7$. This is the group of simply isotropic *angular* similitudes ([8, p. 18]).

IV. $G_7 \subset G_8 \iff \varphi = 0$. This is the group of simply isotropic *boundary* similitudes ([8, p. 8]).

V. $V_7 \subset G_8 \iff c_6 c_7^2 = 1$. This is the group of simply isotropic *volume preserving* similitudes ([8, p. 8]).

VI. $G_6 = G_7 \cap V_7$. This is the group of simply isotropic *volume preserving boundary* similitudes ([8, p. 8]).

VII. $B_6 = B_7 \cap G_7$. This is the group of *modular boundary* motions ([8, p. 9]).

VIII. $B_5 = B_7 \cap S_7 \cap G_7$. This is the group of *unimodular boundary* motions ([8, p. 9]).

Basic references on the geometry of the simply isotropic space ${I_3}^{(1)}$ are Sachs' book [8] and Strubecker's papers [8], [11] and [12].

Using some basic concepts from integral geometry in the sense of R. Deltheil [3], M. I. Stoka [10], G. I. Drinfel'd, and A. V. Lucenko [4], [5], [6], we study the measurability of sets of pairs of intersecting nonisotropic straight lines of type β with respect to G_8 and indicated above subgroups. Analogous problems about sets of pairs of intersecting non-isotropic straight lines of type α in ${I_3}^{(1)}$ have been treated in [2].

2 Measurability with respect to $\mathbf{G_8}$

Let (G_1, G_2) be a pair of intersecting non-isotropic straight lines of type β. Let G_i have Plücker coordinates (p_j^i), $i = 1, 2$, $j = 1, \ldots, 6$, which satisfy the relations ([8, p. 38])

$$p_1^i p_4^i + p_2^i p_5^i + p_3^i p_6^i = 0, \quad i = 1, 2. \tag{2}$$

Since G_1 and G_2 are intersecting non-isotropic lines of type β, we have

$$p_1^1 p_4^2 + p_2^1 p_5^2 + p_3^1 p_4^2 + p_4^1 p_1^2 + p_5^1 p_2^2 + p_6^1 p_3^2 = 0, \qquad p_3^1 - p_3^2 \neq 0, \tag{3}$$

$$|p_1^i| + |p_2^i| \neq 0, \quad i = 1, 2, \tag{4}$$

$$p_1^1 p_2^2 - p_2^1 p_1^2 = 0. \tag{5}$$

Having in mind (4), we can assume, without loss of generality, that $p_1^i = 1$. From (2), p_4^i can be expressed by the remaining Plücker coordinates of G_i, and in view if (3) and (5), p_2^2 and p_6^2 also can be expressed by p_2^1, p_3^1, p_5^1, p_6^1, p_3^2 and p_5^2. Thus the pair (G_1, G_2) can be determined by p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2.

Remark 2.1 We note that if G_i, $i = 1, 2$, are represented in the usual way by the equations

$$G_1\colon \begin{cases} x = a_1(z - r) + p \\ y = b_1(z - r) + q \end{cases}, \qquad G_2\colon \begin{cases} x = a_2(z - r) + p \\ y = \dfrac{a_2}{a_1} b_1(z - r) + q \end{cases}, \tag{6}$$

where $P(p, q, r) = G_1 \cap G_2$ and $a_1 \neq 0$, $a_2 \neq 0$, then

$$\begin{gathered} p_2^1 = \frac{b_1}{a_1}, \quad p_3^1 = \frac{1}{a_1}, \quad p_5^1 = r - \frac{p}{a_1}, \quad p_6^1 = p\frac{b_1}{a_1} - q, \\ p_3^2 = \frac{1}{a_2}, \quad p_5^2 = r - \frac{p}{a_2}. \end{gathered} \tag{7}$$

Under the action of (1) the pair $(G_1, G_2)(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$ is transformed into the pair $(\overline{G}_1, \overline{G}_2)(\overline{p}_2^1, \overline{p}_3^1, \overline{p}_5^1, \overline{p}_6^1, \overline{p}_3^2, \overline{p}_5^2)$. Thus we have

$$\begin{aligned} \overline{p}_2^1 &= Kc_7(\sin\varphi + p_2^1 \cos\varphi), \\ \overline{p}_3^1 &= K(c_4 + c_5 p_2^1 + c_6 p_3^1), \\ \overline{p}_5^1 &= K\{(c_3 - c_5 p_6^1 + c_6 p_5^1)c_7 \cos\varphi \\ &\quad - [c_3 + c_4 p_6^1 + c_6(p_2^1 p_5^1 + p_3^1 p_6^1)]c_7 \sin\varphi - c_1(c_4 + c_5 + c_6 p_3^1)\}, \\ \overline{p}_6^1 &= Kc_7[(c_1 p_2^1 - c_2)\cos\varphi + (c_1 + c_2 p_2^1)\sin\varphi + c_7 p_6^1], \\ \overline{p}_3^2 &= K(c_4 + c_5 p_2^1 + c_6 p_3^2), \\ \overline{p}_5^2 &= K\{(c_3 - c_5 p_6^1 + c_6 p_5^2)c_7 \cos\varphi \\ &\quad - [c_3 + c_4 p_6^1 + c_6(p_2^1 p_5^2 + p_3^2 p_6^1)]c_7 \sin\varphi - c_1(c_4 + c_5 + c_6 p_3^2)\}, \end{aligned} \tag{8}$$

where $K = [c_7(\cos\varphi - p_2^1 \sin\varphi)]^{-1}$, $i = 1, 2$. The transformations (8) form the associated group $\overline{G_8}$ of G_8 ([10, p. 34]). The group $\overline{G_8}$ is isomorphic to G_8 and the density with respect to G_8 of the pairs (G_1, G_2) if it exists, coincides with the density with respect to $\overline{G_8}$ of the set of parameters $(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$.

The associated group $\overline{G_8}$ has the infinitesimal operators

$$X_1 = p_3^1\frac{\partial}{\partial p_5^1} - p_2^1\frac{\partial}{\partial p_6^1} - p_3^2\frac{\partial}{\partial p_5^2}, \quad X_2 = \frac{\partial}{\partial p_6^1}, \qquad X_3 = \frac{\partial}{\partial p_5^1} + \frac{\partial}{\partial p_5^2},$$

$$X_4 = \frac{\partial}{\partial p_3^1} + \frac{\partial}{\partial p_3^2}, \quad X_5 = p_2^1\frac{\partial}{\partial p_3^1} - p_6^1\frac{\partial}{\partial p_5^1} + p_2^1\frac{\partial}{\partial p_3^2} - p_6^1\frac{\partial}{\partial p_5^2},$$

$$X_6 = p_3^1\frac{\partial}{\partial p_3^1} + p_5^1\frac{\partial}{\partial p_5^1} + p_3^2\frac{\partial}{\partial p_3^2} + p_5^2\frac{\partial}{\partial p_5^2}, \quad X_7 = p_3^1\frac{\partial}{\partial p_3^1} - p_6^1\frac{\partial}{\partial p_6^1} + p_3^2\frac{\partial}{\partial p_3^2},$$

$$X_8 = [1 + (p_2^1)^2]\frac{\partial}{\partial p_2^1} + p_2^1p_3^1\frac{\partial}{\partial p_3^1} - p_3^1p_6^1\frac{\partial}{\partial p_5^1} + p_2^1p_6^1\frac{\partial}{\partial p_6^1} + p_2^1p_3^2\frac{\partial}{\partial p_3^2} - g_6^1p_3^2\frac{\partial}{\partial p_5^2}, \tag{9}$$

and it acts transitively on the set of parameters $(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$. The infinitesimal operators X_1, X_2, X_3, X_4, X_7, and X_8 are arcwise unconnected and

$$X_6 = \frac{p_5^2 - p_5^1}{p_3^2 - p_3^1}X_1 + p_6^1X_2 + \frac{p_3^1p_5^2 - p_5^1p_3^2}{p_3^2 - p_3^1}X_3 + X_7.$$

Since

$$X_1(\frac{p_5^2 - p_5^1}{p_3^2 - p_3^1}) + X_2(p_6^1) + X_3(\frac{p_3^1p_5^2 - p_5^1p_3^2}{p_3^2 - p_3^1}) + X_7(1) = 3 \neq 0,$$

we can establish the following

Theorem 2.1 *The set of pairs of intersecting non-isotropic straight lines is not measurable with respect to the group G_8, and it has no measurable subsets.*

3 Measurability with respect to S_7

The associated group $\overline{S_7}$ of the group S_7 has the infinitesimal operators X_1, X_2, X_3, X_4, X_5, X_7, and X_8 from (9), and it acts transitively on the set of parameters $(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$. The integral invariant function

$$f = f(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$$

satisfying the so-called system of R. Deltheil (see [3, p. 28]; [10, p. 11])

$$X_1(f) = 0, \quad X_2(f) = 0, \quad X_3(f) = 0, \quad X_4(f) = 0, \quad X_5(f) = 0$$
$$X_7(f) + f = 0, \quad X_8(f) + 5p_2^1f = 0$$

has the form

$$f = \frac{h}{(p_3^1 - p_3^2)[1 + (p_2^1)^2]^2},$$

where $h = const$.

Thus we state the following

Theorem 3.1 *The set of pairs* $(G_1, G_2)(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$ *is measurable with respect to the group* S_7 *and has the density*

$$d(G_1, G_2) = \frac{1}{|p_3^2 - p_3^1|[1 + (p_2^1)^2]^2}\, dp_2^1 \wedge dp_3^1 \wedge dp_5^1 \wedge dp_6^1 \wedge dp_3^2 \wedge dp_5^2. \qquad (10)$$

Differentiating (7) and substituting into (10) we obtain other expression for the density:

Corollary 3.1 *The density (10) for the pairs* (G_1, G_2) *represented by (6) can be written in the form*

$$d(G_1, G_2) = |\frac{a_1}{a_2^2(a_1^2 + b_1^2)^2}|\, da_1 \wedge db_1 \wedge da_2 \wedge dp \wedge dq \wedge dr. \qquad (11)$$

4 Some Crofton-type formulas with respect to S_7

Let us consider the isotropic plane ι, which is determined by the lines G_1 and G_2. The plane ι has the equation

$$\iota\colon b_1 x - a_1 y + a_1 q - b_1 p = 0.$$

If $\widetilde{P}$ is the orthogonal projection of P into Oxy, consider the affine coordinate system $(\widetilde{P}\overrightarrow{e_1}'\overrightarrow{e_2}')$ in the isotropic plane ι, where $\overrightarrow{e_1}' = (a_1, b_1, 1)$, $\overrightarrow{e_2}' = \overrightarrow{e_3}$. It should be noticed, that if $\widetilde{G} = \iota \cap Oxy$ then $\overrightarrow{e_1}' \| \widetilde{G}$. Let $J^1 = Oxz \cap \iota$ and $J^2 = Oyz \cap \iota$. Obviously

$$J^1\colon x = p - \frac{a_1}{b_1} q,\ y = 0, \qquad J^2\colon y = q - \frac{b_1}{a_1} p,\ x = 0,$$

and J^1, J^2 have the equations

$$J^1\colon x = -\frac{q}{b_1}, \qquad J^2\colon x = -\frac{p}{a_1}$$

with respect to $(\widetilde{P}\overrightarrow{e_1}'\overrightarrow{e_2}')$.

Then the density $d(J^1, J^2)$ for the pairs (J^1, J^2) with respect to the group H_4^1, which is the restriction of S_7 into ι, is (see [1, p. 201])

$$d(J^1, J^2) = \left(\frac{p}{a_1} - \frac{q}{b_1}\right)^2 d\frac{p}{a_1} \wedge d\frac{q}{b_1}.$$

Recall that ([8, p. 45])

$$s = \frac{a_1 - a_2}{a_2\sqrt{a_1^2 + b_1^2}} \qquad (12)$$

is the angle from G_1 to G_2, we find

$$d(J^1, J^2) \wedge dP \wedge ds = \frac{(pb_1 - qa_1)pq}{a_1^3 b_1^4 a_2^2 \sqrt{a_1^2 + b_1^2}}\, da_1 \wedge db_1 \wedge dp \wedge dq \wedge dr \wedge da_2.$$

Comparing with (11), we get

$$d(G_1, G_2) = \left| \frac{a_1^4 b_1^4}{pq(pb_1 - qa_1)(a_1^2 + b_1^2)^{\frac{3}{2}}} \right| d(J^1, J^2) \wedge ds \wedge dP. \tag{13}$$

Let φ_i, $i = 1, 2$, b the angle between G_i and Oxy. Then ([8, p. 48])

$$\varphi_1 = \frac{1}{\sqrt{a_1^2 + b_1^2}}, \qquad \varphi_2 = \frac{a_1}{a_2\sqrt{a_1^2 + b_1^2}}, \tag{14}$$

and (13) becomes

$$d(G_1, G_2) = \left| \frac{a_1^4 b_1^4 \varphi_1^3}{pq(pb_1 - qa_1)} \right| d(J^1, J^2) \wedge ds \wedge dP. \tag{15}$$

By differentiation of (14) and by exterior multiplication by (12), we obtain

$$\begin{aligned} d(G_1, G_2) &= \left| \frac{a_1^4 b_1^4}{pq(pb_1 - qa_1)(a_1^2 + b_1^2)^{\frac{3}{2}}} \right| d(J^1, J^2) \wedge d\varphi_2 \wedge dP \\ &= \left| \frac{a_1^4 b_1^4 \varphi_1^3}{pq(pb_1 - qa_1)} \right| d(J^1, J^2) \wedge d\varphi_2 \wedge dP. \end{aligned} \tag{16}$$

If $\widetilde{\varphi}$ is the isotropic distance from J^1 to J^2, then ([7, p. 19])

$$\widetilde{\varphi} = -\frac{p}{a_1} + \frac{q}{b_1}. \tag{17}$$

Putting (17) into (15) and (16), we find

$$d(G_1, G_2) = \left| \frac{a_1^3 b_1^3 \varphi_1^3}{pq\widetilde{\varphi}} \right| d(J^1, J^2) \wedge ds \wedge dP = \left| \frac{a_1^3 b_1^3 \varphi_1^3}{pq\widetilde{\varphi}} \right| d(J^1, J^2) \wedge d\varphi_2 \wedge dP. \tag{18}$$

Let G_i^1 and G_i^2 be now the projections of G_i into Oxz and Oyz obtained in a parallel way to Oy and Ox, respectively. Then

$$\begin{aligned} G_i^1\colon &\quad z = \frac{1}{a_i}x + r - \frac{p}{a_i},\ y = 0,\ i = 1, 2, \\ G_1^2\colon &\quad z = \frac{1}{b_1}y + r - \frac{q}{b_1},\ x = 0, \\ G_2^2\colon &\quad z = \frac{a_1}{a_2 b_1}y + r - \frac{a_1}{a_2 b_1}q,\ x = 0. \end{aligned}$$

Furthermore,

$$d(G_1^1, G_2^1) = \left| \frac{1}{a_1 a_2 (a_2 - a_1)} \right| da_1 \wedge da_2 \wedge dp \wedge dr \tag{19}$$

is the density for the pairs (G_1^1, G_2^1) in the isotropic plane Oxz with respect ${}^1H_4^1$ which is the restriction of S_7 into Oxz and

$$d(G_1^2, G_2^2) = \left| \frac{1}{b_1^2 a_2 (a_2 - a_1)} \right| (a_1 db_1 \wedge da_2 - a_2 db_1 \wedge da_1) \wedge dq \wedge dr$$

is the density for the pairs (G_1^2, G_2^2) in the isotropic plane Oyz with respect ${}^2H_4^1$ which is the restriction of S_7 into Oyz (see [1, p. 177]).

By exterior multiplication of (G_1^1, G_2^1) and $ds \wedge dq$, we get

$$d(G_1, G_2) = \left| \frac{a_1^2 \varphi_1}{b_1} \right| d(G_1^1, G_2^1) \wedge ds \wedge dq, \tag{20}$$

and by exterior multiplication of (19) and $d\varphi_1 \wedge dq$:

$$d(G_1, G_2) = \left| \frac{a_1^2 s}{b_1} \right| d(G_1^1, G_2^1) \wedge d\varphi_1 \wedge dq. \tag{21}$$

If, instead of using $d\varphi_1 \wedge dq$, we multiply by $d\varphi_2 \wedge dq$, we obtain

$$d(G_1, G_2) = \left| \frac{a_1 a_2 s}{b_1} \right| d(G_1^1, G_2^1) \wedge d\varphi_2 \wedge dq. \tag{22}$$

Analogously, we can derive the following formulas:

$$\begin{aligned} d(G_1, G_2) &= \left| \frac{a_1^2 b_1^2 \varphi_1}{a_2^3} \right| d(G_1^2, G_2^2) \wedge ds \wedge dp \\ &= \left| \frac{b_1^2 s}{a_1} \right| d(G_1^2, G_2^2) \wedge d\varphi_1 \wedge dp \\ &= \left| \frac{a_2 b_1^2 s}{a_1^2} \right| d(G_1^2, G_2^2) \wedge d\varphi_2 \wedge dp. \end{aligned} \tag{23}$$

In summary, the following theorem holds.

Theorem 4.1 *The density for the set of pairs (G_1, G_2) of intersecting non-isotropic straight lines of type β, determined by (6), with respect to the group S_7 satisfies the relations (15), (16), (18), (20), (21), (22), and (23).*

5 Measurability with respect to G_6

Now, the corresponding associated group $\overline{G_6}$ has the infinitesimal operators

$$\begin{aligned} &Y_1 = p_3^1 \frac{\partial}{\partial p_5^1} - p_2^1 \frac{\partial}{\partial p_6^1} + p_3^2 \frac{\partial}{\partial p_5^2}, \quad Y_2 = \frac{\partial}{\partial p_6^1}, \\ &Y_3 = \frac{\partial}{\partial p_5^1} + \frac{\partial}{\partial p_5^2}, \quad Y_4 = p_2^1 \frac{\partial}{\partial p_3^1} - p_6^1 \frac{\partial}{\partial p_5^1} + p_2^1 \frac{\partial}{\partial p_3^2} - p_6^1 \frac{\partial}{\partial p_5^2}, \\ &Y_7 = 3p_3^1 \frac{\partial}{\partial p_3^1} + 2p_5^1 \frac{\partial}{\partial p_5^1} - p_6^1 \frac{\partial}{\partial p_6^1} + 3p_3^2 \frac{\partial}{\partial p_3^2} + 2p_5^2 \frac{\partial}{\partial p_5^2}, \quad Y_8 = \frac{\partial}{\partial p_1^1} + \frac{\partial}{\partial p_3^2}. \end{aligned}$$

The group $\overline{G_6}$ acts intransitively on the set of points $(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$ and therefore the set of pairs (G_1, G_2) has not invariant density with respect to G_6. The system

$$Y_1(f) = 0,\ Y_2(f) = 0,\ Y_3(f) = 0,\ Y_4(f) = 0,\ Y_7(f) = 0,\ Y_8(f) = 0$$

has the solution

$$f = p_2^1,$$

and it is an absolute invariant of G_6. Consider the subset of pairs (G_1, G_2) satisfying the condition

$$p_2^1 = h, \tag{24}$$

where $h = const$. The group $\overline{G_6}$ induces on this subset the group G_6^* with the infinitesimal operators

$$Z_1 = p_3^1\frac{\partial}{\partial p_5^1} - h\frac{\partial}{\partial p_6^1} + p_3^2\frac{\partial}{\partial p_5^2}, \quad Z_2 = \frac{\partial}{\partial p_6^1},$$
$$Z_3 = \frac{\partial}{\partial p_5^1} + \frac{\partial}{\partial p_5^2}, \quad Z_4 = p_2^1\frac{\partial}{\partial p_3^1} - p_6^1\frac{\partial}{\partial p_5^1} + p_2^1\frac{\partial}{\partial p_3^2} - p_6^1\frac{\partial}{\partial p_5^2},$$
$$Z_7 = 3p_3^1\frac{\partial}{\partial p_3^1} + 2p_5^1\frac{\partial}{\partial p_5^1} - p_6^1\frac{\partial}{\partial p_6^1} + 3p_3^2\frac{\partial}{\partial p_3^2} + 2p_5^2\frac{\partial}{\partial p_5^2}, \quad Z_8 = \frac{\partial}{\partial p_1^1} + \frac{\partial}{\partial p_3^2}.$$

The integral invariant function $f = f(p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$, which satisfies the Deltheil system

$$Z_1(f) = 0,\ Z_2(f) = 0,\ Z_3(f) = 0,\ Z_4(f) = 0,\ Z_7(f) - 9f = 0,\ Z_8(f) = 0,$$

has the form

$$f = \frac{c}{(p_3^1 - p_3^2)^3},$$

where $c = \text{const}$.

Thus we state the following

Theorem 5.1 *The set of pairs $(G_1, G_2)(p_2^1, p_3^1, p_5^1, p_6^1, p_3^2, p_5^2)$ of intersecting non-isotropic lines of type β is not measurable with respect to G_6, but it has the measurable subset*

$$p_2^1 = h, \quad h = const,$$

with the density

$$d(G_1, G_2) = \frac{1}{|p_3^2 - p_3^1|^3}\, dp_3^1 \wedge dp_5^1 \wedge dp_6^1 \wedge dp_3^2 \wedge dp_5^2. \tag{25}$$

Differentiating (7), (24), and replacing into (25), we establish

Corollary 5.1 *The set of pairs (G_1, G_2) of intersecting non-isotropic lines of type β, determined by (6), is not measurable with respect to the group G_6, but it has the measurable subset*

$$\frac{b_1}{a_1} = h, \quad h = const,$$

with the density

$$d(G_1, G_2) = \frac{1}{(a_1 - a_2)^2}\, da_1 \wedge da_2 \wedge dp \wedge dq \wedge dr.$$

6 Measurability with respect to $\mathbf{B_7}$, $\mathbf{W_7}$, $\mathbf{G_7}$, $\mathbf{V_7}$, $\mathbf{B_6}$, and $\mathbf{B_5}$

By arguments similar to those used in the sections 2, 3, and 5, we investigated the measurability with respect to all the remaining groups. We have the following results:

Theorem 6.1 *The set of pairs (G_1, G_2) of intersecting non-isotropic straight lines of type β, determined by (6), is measurable with respect to the group*

(i) B_7 and it has the density

$$d(G_1, G_2) = \left| \frac{a_1 a_2}{(a_1 - a_2)^3 \sqrt{a_1^2 + b_1^2}} \right| da_1 \wedge db_1 \wedge da_2 \wedge dp \wedge dq \wedge dr;$$

(ii) V_7 and it has the density

$$d(G_1, G_2) = \frac{|a_1|}{(a_1 - a_2)^2 (a_1^2 + b_1^2)}\, da_1 \wedge db_1 \wedge da_2 \wedge dp \wedge dq \wedge dr.$$

Theorem 6.2 *With respect to the groups W_7 and S_7 the set of pairs (G_1, G_2) of intersecting non-isotropic lines of type β is not measurable and it has no measurable subsets.*

Theorem 6.3 *The set of pairs (G_1, G_2) of intersecting non-isotropic straight lines of type β, determined by (6), is not measurable with respect to the group*

(i) B_6, but it has the measurable subset

$$\frac{b_1}{a_1} = h, \quad h = const,$$

with the density

$$d(G_1, G_2) = \left| \frac{a_1 a_2}{(a_1 - a_2)^3} \right| da_1 \wedge da_2 \wedge dp \wedge dq \wedge dr;$$

(ii) B_5, but it has the measurable subset

$$\frac{b_1}{a_1} = h_1, \quad \frac{1}{a_1} - \frac{1}{a_2} = h_2, \quad h_1, h_2 = const,$$

with the density

$$d(G_1, G_2) = \left| \frac{a_2}{a_1 (a_1 - a_2)} \right| da_1 \wedge da_2 \wedge dp \wedge dq \wedge dr.$$

References

[1] Borisov, A.: Integral geometry in the Galilean plane. *Research work qualifying for a degree full-professor, Sofia,* 1998.

[2] Borisov, A. V., Spirova, M. G.: *Crofton-type formulas relating sets of pairs of intersecting nonisotropic straight lines in the simply isotropic space.* Tensor **67** (2006), 243–253.

[3] Deltheil, R.: *Sur la théorie des probabilité géométriques.* Thése Ann. Fac. Sc. Univ. Toulouse **11** (1919), 1–65.

[4] Drinfel'd, G. I.: *On the measure of the Lie groups.* Zap. Mat. Otdel. Fiz. Mat. Fak. Kharkov. Mat. Obsc. **21** (1949), 47–57 (in Russian).

[5] Drinfel'd, G. I., Lucenko, A. V.: *On the measure of sets of geometric elements.* Vest. Kharkov. Univ. **31**, 3 (1964), 34–41 (in Russian).

[6] Lucenko, A. V.: *On the measure of sets of geometric elements and thear subsets.* Ukrain. Geom. Sb. **1**, 3 (1965), 39–57 (in Russian).

[7] Sachs, H.: Ebene isotrope Geometrie. *Friedr. Vieweg Sohn, Braunschweig,* 1987.

[8] Sachs, H.: Isotrope Geometrie des Raumes. *Friedr. Vieweg and Sohn, Braunschweig–Wiesbaden,* 1990.

[9] Santaló, L. A.: Integral Geometry and Geometric Probability. *Addison-Wesley, London,* 1976.

[10] Stoka, M. I.: Geometrie Integrala. *Ed. Acad. RPR, Bucuresti,* 1967.

[11] Strubecker, K.: *Differentialgeometrie des isotropen Raumes I.* Sitzungsber. Österr. Akad. Wiss. Wien **150** (1941), 1–53.

[12] Strubecker, K.: *Differentialgeometrie des isotropen Raumes II, III, IV, V.* Math. Z. **47** (1942), 743–777; **48** (1942), 369–427; **50** (1944), 1–92; **52** (1949), 525–573.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 17–23

Conjugated Algebras*

Ivan CHAJDA

Department of Algebra and Geometry, Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: chajda@inf.upol.cz

(Received February 17, 2008)

Abstract

We generalize the correspondence between basic algebras and lattices with section antitone involutions to a more general case where no lattice properties are assumed. These algebras are called conjugated if this correspondence is one-to-one. We get conditions for the conjugary of such algebras and introduce the induced relation. Necessary and sufficient conditions are given to indicated when the induced relation is a quasiorder which has "nice properties", e.g. the unary operations are antitone involutions on the corresponding intervals.

Key words: Conjugated alegebras, basic algebra, section antitone involution, quasiorder.

2000 Mathematics Subject Classification: 08A40, 06D35, 06A12

Correspondence between MV-algebras and bounded distributive lattices with section antitone involutions is well-known, see e.g. [3] and [5]. This was generalized for basic algebras and general bounded lattices with section antitone involutions, see [2] and [3] for details. Semilattices and lattices with section antitone involutions were treated separately in [1]. If a bounded lattice is replaced by the so-called λ-lattice, the corresponding algebra is called an NMV-algebra, an non-associative generalization of an MV-algebra, see [4]. If a little less is assumed, we get the correspondence between weak basic algebras and directoids with section antitone involutions, see [6]. These attempts motivate us to find a general correspondence between algebras of two sorts. One of them are "MV-like algebras", the other are "semilattice-like algebras" with a set of unary operations. Since in all the aforementioned cases the "semilattice-like algebras" were ordered, we add an assumption that our algebras of the second sort will

*Supported by the Research and Development Council of the Czech Government MSM 6 198 959 214.

be at least quasiordered. If there is a one-to-one correspondence between these algebras, we will say that they are conjugated.

At first, we get precise meaning to mentioned concepts.

We consider two kinds of algebras. The first are algebras $\mathcal{A} = (A; \oplus, \neg, 0)$ of type $(2, 1, 0)$. For the sake of brevity, we will denote $1 := \neg 0$ the algebraic constant of $\mathcal{A}$.

The second are algebras $\mathcal{L} = (A; \sqcup, (^b)_{b\in A}, 0)$ where $\sqcup$ is a binary operation, 0 is a nullary operation and for each $b \in A$, ^{b} is a unary operation on A, i.e. it is a mapping $A \to A$ assigning to $x \in A$ an element x^b. Denote by $1 := 0^0$. To every $\mathcal{A} = (A; \oplus, \neg, 0)$ there can be assigned an algebra $\mathcal{L}(A) = (A; \sqcup, (^b)_{b\in A}, 0)$, where

$$x \sqcup y = \neg(\neg x \oplus y) \oplus y \quad \text{and} \quad x^y = \neg x \oplus y.$$

To every $\mathcal{L} = (L; \sqcup, (^b)_{b\in L}, 0)$ there can be assigned an algebra $\mathcal{A}(L) = (L; \oplus, \neg, 0)$, where

$$x \oplus y = (x^0 \sqcup y)^y \quad \text{and} \quad \neg x = x^0.$$

We call algebras $\mathcal{A} = (A; \oplus, \neg, 0)$ and $\mathcal{L} = (L; \sqcup, (^b)_{b\in A}, 0)$ *conjugated* if

$$\mathcal{L} = \mathcal{L}(A) \quad \text{and} \quad \mathcal{A} = \mathcal{A}(L).$$

This yields $\mathcal{A}(\mathcal{L}(A)) = \mathcal{A}$ and $\mathcal{L}(\mathcal{A}(L)) = \mathcal{L}$, i.e. if they share the same base-set and the aforementioned assignments are one-to-one correspondences.

At first, we can describe the following properties of conjugated algebras.

Theorem 1 *Let $\mathcal{A} = (A; \oplus, \neg, 0)$ satisfy the conditions*

(A1) $\neg\neg x = x$;

(A2) $x \oplus 0 = x$;

(A3) $\neg(\neg(x \oplus y) \oplus y) \oplus y = x \oplus y$.

Then $\mathcal{A}(\mathcal{L}(A)) = \mathcal{A}$ and $\mathcal{L}(A)$ satisfies the conditions

(L1) $(x \sqcup y)^{yy} = x \sqcup y$;

(L2) $x^y = (x \sqcup y)^y$;

(L3) $x \sqcup 0 = x$.

Proof Assume that $\mathcal{A}$ satisfies (A1), (A2) and (A3) and denote by $\boxplus, \sim$ the operations of $\mathcal{A}(\mathcal{L}(A))$. Of course, the nullary operation 0 is the same both in $\mathcal{A}$ and $\mathcal{A}(\mathcal{L}(A))$. We have by (A2)

$$\sim x = x^0 = \neg x \oplus 0 = \neg x.$$

Further, we compute by (A1) and (A3)

$$\begin{aligned} x \boxplus y = (x^0 \sqcup y)^y = (\neg x \sqcup y)^y = (\neg(\neg\neg x \oplus y) \oplus y)^y \\ = \neg(\neg(x \oplus y) \oplus y) \oplus y = x \oplus y \end{aligned}$$

thus $\mathcal{A}(\mathcal{L}(A)) = \mathcal{A}$.

Further, applying (A3), we conclude

$$x^y = \neg x \oplus y = \neg(\neg(\neg x \oplus y) \oplus y) \oplus y = \neg(x \sqcup y) \oplus y = (x \sqcup y)^y$$

proving (L2). Using this we obtain

$$(x \sqcup y)^{yy} = x^{yy} = \neg(\neg x \oplus y) \oplus y = x \sqcup y$$

which is (L1). Using (A1) and (A2) we prove also (L3):

$$x \sqcup 0 = \neg(\neg x \oplus 0) \oplus 0 = \neg\neg x = x. \qquad \square$$

Theorem 2 *Let $\mathcal{L} = (L; \sqcup, (^b)_{b\in L}, 0)$ satisfy (L1), (L2) and (L3). Then $\mathcal{L}(\mathcal{A}(L)) = \mathcal{L}$ and $\mathcal{A}(L)$ satisfies (A1), (A2) and (A3).*

Proof Assume that $\mathcal{L}$ satisfies (L1), (L2) and (L3) and denote by $\vee$ the binary operation and by $(f_b)_{b\in L}$ the set of unary operations of $\mathcal{L}(\mathcal{A}(L))$. Of course, the nullary operation 0 is the same in both the algebras. Then, by (L1),

$$x \vee y = \neg(\neg x \oplus y) \oplus y = (\neg x \oplus y)^y = (x \sqcup y)^{yy} = x \sqcup y.$$

Further, $\neg\neg x = x^{00} = (x \sqcup 0)^{00} = x \sqcup 0 = x$ by (L1), (L2) and (L3). Next, by (L2),

$$f_y(x) = \neg x \oplus y = ((\neg x)^0 \sqcup y)^y = (x^{00} \sqcup y)^y = (x \sqcup y)^y = x^y$$

thus $\mathcal{L}(\mathcal{A}(L)) = \mathcal{L}$ and $\mathcal{A}(L)$ satisfies (A1). Analogously,

$$x \oplus 0 = (x^0 \sqcup 0)^0 = x^{00} = (x \sqcup 0)^{00} = x \sqcup 0 = x$$

thus $\mathcal{A}(L)$ satisfies (A2). Since $\mathcal{A}(L)$ already satisfies (A1), we can easily compute

$$\neg(\neg(x \oplus y) \oplus y) \oplus y = (\neg x \sqcup y)^y = (\neg x)^y = \neg\neg x \oplus y = x \oplus y$$

proving (A3). $\square$

Corollary 1 *Let $\mathcal{A}$ satisfy (A1), (A2) and (A3). Then $\mathcal{A}$ and $\mathcal{L}(A)$ are conjugated. Let $\mathcal{L}$ satisfy (L1), (L2) and (L3). Then $\mathcal{L}$ and $\mathcal{A}(L)$ are conjugated.*

Corollary 2 *Let $\mathcal{A}, \mathcal{L}$ be conjugated algebras. Then $\mathcal{A}$ satisfies (A1), (A2), (A3) if and only if $\mathcal{L}$ satisfies (L1), (L2), (L3).*

Remark 1 As mentioned in the introduction, the correspondence between $\mathcal{A} = (A; \oplus, \neg, 0)$ and $\mathcal{L} = (A; \sqcup, (^b)_{b\in A}, 0)$ was studied for several cases. The results are as follows:

(1) If $\mathcal{A}$ is a basic algebra then $\mathcal{L} = \mathcal{L}(A)$ is a bounded semilattice with section antitone involutions (SAI for short);

(2) If $\mathcal{A}$ is an MV-algebra then $\mathcal{L} = \mathcal{L}(A)$ is a bounded semilattice with SAI satisfying the Exchange Property;

(3) If $\mathcal{A}$ is an NMV-algebra then $\mathcal{L} = \mathcal{L}(A)$ is a commutative directoid with SAI;

(4) If $\mathcal{A}$ is a weak basic algebra then $\mathcal{L} = \mathcal{L}(A)$ is a directoid with SAI (not necessarily commutative).

In all the cases, $\mathcal{A}$ and $\mathcal{L}$ are conjugated and there exists an induced order such that 0 (or 1) is the least (or the greatest) element and $y \leq x \sqcup y$. We are going to study this question concerning some "order-like" relation also on conjugated algebras in general.

Define a binary relation $\leq$ on an algebra $\mathcal{A} = (A; \oplus, \neg, 0)$ as follows

$$x \leq y \quad \text{if and only if} \quad \neg x \oplus y = 1.$$

Call $\leq$ the *induced relation* on $\mathcal{A}$.

Let us note that $1 = \neg 0$. If $\mathcal{A}$ satisfies (A1), then also $\neg 1 = \neg\neg 0 = 0$.

Lemma 1 *The induced relation $\leq$ on $\mathcal{A}$ is reflexive if and only if $\mathcal{A}$ satisfies the identity*

(P) $\neg x \oplus x = 1$.

Let $\mathcal{A}$ satisfy (A1). Then $0 \leq x \leq 1$ for each $x \in A$ if and only if $\mathcal{A}$ satisfies the identity

(A4) $1 \oplus x = 1 = x \oplus 1$.

Proof The first assertion is trivial. For the second one, $0 \leq x$ is equivalent to $1 \oplus x = \neg 0 \oplus x = 1$ and $x \leq 1$ is equivalent to $\neg x \oplus 1 = 1$ for each $x \in A$, i.e. due to (A1), $\mathcal{A}$ satisfies also the identity $x \oplus 1 = 1$. □

Lemma 2 *Let $\mathcal{A}, \mathcal{L}$ be conjugated algebras and $\leq$ be the induced relation on $\mathcal{A}$. Let $\mathcal{A}$ satisfy (A1) and (A4) and $\mathcal{L}$ satisfy (L1). Then the following conditions are equivalent*

(a) $1^x = x$ *and* $x^x = 1$*;*

(b) $x \leq y$ *if and only if* $x \sqcup y = y$.

Proof (a)⇒(b): Let $x \leq y$. Then

$$(x \sqcup y)^y = \neg x \oplus y = 1$$

thus, by (L1),

$$x \sqcup y = (x \sqcup y)^{yy} = 1^y = y.$$

Conversely, if $x \sqcup y = y$ then

$$\neg x \oplus y = (x \sqcup y)^y = y^y = 1,$$

i.e. $x \leq y$.

(b)⇒(a): Applying Lemma 1, (A4) yields $0 \leq x$ and, by the assumption (b), $0 \sqcup x = x$. By (A4) and (L1) we have

$$1^x = (1 \oplus x)^x = (\neg 1 \sqcup x)^{xx} = (0 \sqcup x)^{xx} = 0 \sqcup x = x.$$

Similarly,

$$x^x = (0 \sqcup x)^x = 0^0 \oplus x = 1 \oplus x = 1. \qquad \square$$

Lemma 3 *Let $\mathcal{A}$ and $\mathcal{L}$ be conjugated algebras. Then $x \leq x \sqcup y$ if and only if $\mathcal{A}$ satisfies*

(A5) $\neg x \oplus (\neg(\neg x \oplus y) \oplus y) = 1$.

Proof By the definition of $\leq$ we have that

$$x \leq x \sqcup y \quad \text{if and only if} \quad \neg x \oplus (x \sqcup y) = 1.$$

However, $\mathcal{A}, \mathcal{L}$ are conjugated thus $x \sqcup y = \neg(\neg x \oplus y) \oplus y$. $\square$

A binary relation is called a *quasiorder* if it is reflexive and transitive. We are going to characterize algebras $\mathcal{A} = (A; \oplus, \neg, 0)$ for which the induced relation is a quasiorder which has a special meaning for the assigned algebra $\mathcal{L}$.

Lemma 4 *Let $\mathcal{A} = (A; \oplus, \neg, 0)$ satisfy the identities (A1) and*

(A6) $0 \oplus x = x$;

(A7) $\neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z) = 1$.

Then the induced relation $\leq$ is transitive.

Proof Assume $x \leq y$ and $y \leq z$, i.e. $\neg x \oplus y = 1$ and $\neg y \oplus z = 1$. By (A1), (A7) and (A6) we compute

$$\begin{aligned} 1 &= \neg(\neg(\neg(\neg x \oplus y) \oplus y) \oplus z) \oplus (\neg x \oplus z) \\ &= \neg(\neg(\neg 1 \oplus y) \oplus z) \oplus (\neg x \oplus z) = \neg(\neg(0 \oplus y) \oplus z) \oplus (\neg x \oplus z) \\ &= \neg(\neg y \oplus z) \oplus (\neg x \oplus z) = \neg 1 \oplus (\neg x \oplus z) = 0 \oplus (\neg x \oplus z) = \neg x \oplus z \end{aligned}$$

whence $x \leq z$. $\square$

Let $(A; \leq)$ be a quasiordered set and $f : A \to A$ be a mapping. We say that f is *antitone* if $x \leq y$ yields $f(y) \leq f(x)$ and f is an *involution* if $f(f(x)) = x$ for every $x \in A$. If $a, b \in A$ and $a \leq b$, by an interval $[a, b]$ is meant the subset of A given by $[a, b] = \{x \in A; a \leq x \leq b\}$.

Theorem 3 *Let $\mathcal{A}, \mathcal{L}$ be conjugated algebras, let $\leq$ be the induced relation on $\mathcal{A}$. Let $\mathcal{A}$ satisfy (A1), (A2), (A3), (A4) and (A6). The following conditions are equivalent*

(1) $\mathcal{A}$ satisfies (A7);

(2) $\leq$ is a quasiorder on A such that $x \leq x \sqcup y$ for each $x, y \in A$ and for each $z \in A$ the mapping $x \mapsto x^z$ is an antitone involution on the interval $[z, 1]$.

Proof (1) $\Rightarrow$ (2): Put $y = 0 = z$ in (A7). We get $\neg x \oplus x = 1$ which is (P) of Lemma 1, i.e. $\leq$ is reflexive. Since $\mathcal{A}$ satisfies (A6) and (A7), $\leq$ is transitive by Lemma 4 and hence $(A; \leq)$ is a quasiordered set.

Assume $x \leq y$. Then $\neg x \oplus y = 1$ and, by (A7),

$$1 = \neg(\neg(\neg(\neg x \oplus y) \oplus y) \oplus z) \oplus (\neg x \oplus z) = \neg(\neg y \oplus z) \oplus (\neg x \oplus z)$$

thus

$$\neg y \oplus z \leq \neg x \oplus z. \qquad (*)$$

For $z = 0$ we have $x \leq y \Rightarrow \neg y \leq \neg x$ which is equivalent to

$$\neg x \oplus y = 1 \quad \Rightarrow \quad y \oplus \neg x = 1. \qquad (**)$$

Taking $z = 0$ and replacing x by $\neg x$ in (A7), we obtain

$$(\neg(\neg x \oplus y) \oplus y) \oplus \neg x = 1$$

thus, by $(**)$, we obtain

$$\neg x \oplus (\neg(\neg x \oplus y) \oplus y) = 1$$

which yields

$$x \leq \neg(\neg x \oplus y) \oplus y = x \sqcup y.$$

Let $x, y \in [z, 1]$ and $x \leq y$. By $(*)$ we have $y^z = \neg y \oplus z \leq \neg x \oplus z = x^z$ thus the mapping $x \mapsto x^z$ is antitone. By (A4) we have $x^z \leq 1$. Applying $(*)$ twice and using (A1), we obtain

$$x \leq y \quad \Rightarrow \quad x \oplus z \leq y \oplus z. \qquad (***)$$

Since $\neg x \leq 1$ by (A4), $(*)$ yileds $0 \leq x$ thus, by $(***)$ and (A6), we obtain

$$y = 0 \oplus y \leq x \oplus y.$$

This yields $z \leq \neg x \oplus z = x^z$. We have shown that $x \mapsto x^z$ is really a mapping of the interval $[z, 1]$ into itself. By (L1) and (L2), it is an involution. We have shown (1) $\Rightarrow$ (2).

(2) $\Rightarrow$ (1): By (2) we have $\neg x \leq \neg x \sqcup y$ where the induced relation $\leq$ is a quasiorder on A. By (2),

$$\neg(\neg(x \oplus y) \oplus y) \oplus z) = \neg(\neg x \sqcup y) \oplus z =$$

$$= ((\neg x \sqcup y) \sqcup z)^z = (\neg x \sqcup y)^z \leq (\neg x)^z = (\neg x \sqcup z)^z = x \oplus z$$

thus $\neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z) = 1$ which is just (A7). $\square$

References

[1] Chajda, I.: *Lattices and semilattices having an antitone involution in every upper interval.* Comment. Math. Univ. Carol. **44** (2003), 577–585.

[2] Chajda, I., Emanovský, P.: *Bounded lattices with antitone involutions and properties of MV-algebras.* Discuss. Math., Gener. Algebra and Appl. **24** (2004), 31–42.

[3] Chajda, I., Halaš, R., Kühr, J.: Semilattice Structures. *Heldermann Verlag, Lemgo,* 2007, 228 pp.

[4] Chajda, I., Kühr, J.: *A non-associative generalization of MV-algebras.* Math. Slovaca **57** (2007), 1–12.

[5] Cignoli, R. L. O., D'Ottaviano, M. L., Mundici, D.: Algebraic Foundations of Many-valued Reasoning. *Kluwer Acad. Publ., Dordrecht,* 2000.

[6] Halaš, R., Plojhar, L.: *Weak MV-algebras.* Math. Slovaca **58** (2008), 1–10.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 25–31

Basic Pseudorings*

Ivan CHAJDA [1], Miroslav KOLAŘÍK [2]

[1] *Department of Algebra and Geometry, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: chajda@inf.upol.cz

[2] *Department of Computer Science, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: kolarik@inf.upol.cz

(Received November 3, 2008)

Abstract

The concept of a basic pseudoring is introduced. It is shown that every orthomodular lattice can be converted into a basic pseudoring by using of the term operation called Sasaki projection. It is given a mutual relationship between basic algebras and basic pseudorings. There are characterized basic pseudorings which can be converted into othomodular lattices.

Key words: Basic algebra, basic pseudoring, orthomodular lattice.

2000 Mathematics Subject Classification: 06D35, 06C15, 03G25

It is well-known that every Boolean algebra can be converted into a Boolean ring by using of the symmetrical difference, see e.g. [2]. Also conversely, every Boolean ring can be converted into a Boolean algebra. For orthomodular lattices (instead of Boolean algebras) a similar construction giving a ring-like structure called Boolean quasiring was settled in [6], [7] and generalized for bounded lattices with an antitone involution in [8] and [9]. The natural question is for which algebras used in non-classical logics a similar conversion into a ring-like structure is possible. Of course, Boolean algebras serve as axiomatization of the classical propositional logic and orthomodular lattices play a similar role in the logic of quantum mechanics, see e.g. [1], [7], [8], [9].

In this study we are concentrated in an algebraic counterpart of many-valued logics. This is usually considered to be an MV-algebra for many-valued Łukasiewicz logic. However, it was generalized for more wide class as the concept of basic algebra, see e.g. [3], [4] as sources.

*Supported by the Research and Development Council of the Czech Government MSM 6 198 959 214.

Let us note that a certain ring-like structures corresponding to MV-algebras were investigated by the first author and H. Länger in [5] and analogously, it was done for pseudo MV-algebras by Y. Shang in [10]. We will involve a similar approach which, however, can be used both for MV-algebras and orthomodular lattices.

The concept of basic algebra was introduced in [3] as a common generalization of an MV-algebra and an orthomodular lattice. Recall that a *basic algebra* (see e.g. [3], [4]) is an algebra $\mathcal{A} = (A; \oplus, \neg, 0)$ of type $(2, 1, 0)$ satisfying the following identities

(BA1) $x \oplus 0 = x$;

(BA2) $\neg\neg x = x$ (double negation);

(BA3) $\neg(\neg x \oplus y) \oplus y = \neg(\neg y \oplus x) \oplus x$ (Łukasiewicz axiom);

(BA4) $\neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z) = 1$ (where $1 := \neg 0$).

Let us note that every basic algebra satisfies also the identities $1 \oplus x = 1 = x \oplus 1$, $0 \oplus x = x$, $x \oplus \neg x = \neg x \oplus x = 1$ (see e.g. [3]). In every basic algebra $\mathcal{A} = (A; \oplus, \neg, 0)$, the partial order can be defined by $x \leq y$ if and only if $\neg x \oplus y = 1$. The ordered set $(A; \leq)$ is a bounded lattice where $x \vee y = \neg(\neg x \oplus y) \oplus y$, $x \wedge y = \neg(\neg x \vee \neg y)$ and $1 = \neg 0$. Moreover, it satisfies $y \leq x \oplus y$ and the mapping $x \mapsto \neg x$ is antitone for every $x, y \in A$.

A basic algebra $\mathcal{A} = (A; \oplus, \neg, 0)$ is called *commutative* if it satisfies the identity $x \oplus y = y \oplus x$.

The concept of symmetrical difference can be introduced for basic algebras in a way similar to that of [6] for orthomodular lattices, however, an operation $\oplus$ is considered instead of $\vee$ in orthomodular lattice because $\oplus$ expresses the logical connective disjunction in the corresponding logic.

Searching for an appropriate ring-like structure, we choose the following one from a number of possible ways.

Definition 1 By a *basic pseudoring* we mean an algebra $\mathcal{R} = (R; +, \cdot, 0, 1)$ of type $(2, 2, 0, 0)$ satisfying the identities

(R1) $1 + 0 = 1$;

(R2) $x \cdot 1 = x$;

(R3) $1 + (1 + x) = x$;

(R4) $(1 + x \cdot (1 + y)) \cdot (1 + y) = (1 + y \cdot (1 + x)) \cdot (1 + x)$;

(R5) $1+(1+(1+(1+((1+x)\cdot(1+y)))\cdot(1+y))\cdot(1+z))\cdot((1+x)\cdot(1+z)) = 1$.

One can immediately mention that this concept differs from the concept of a Boolean quasiring or a generalized Boolean quasiring as defined in [7], [8], [9]. From this point it can be of interest that this ring-like structure can be also reached from every orthomodular structure. Of course, this conversion differs due to the fact that instead of a symmetrical difference (see [6]) the *Sasaki*

operation (alias *Sasaki projection*, see [1]) is used. Let us recall that by a Sasaki operation of an orthomodular lattice is meant a term operation

$$(x \vee y') \wedge y.$$

We are ready to state our first result.

Theorem 1 *Let* $\mathcal{L} = (L; \vee, \wedge, ', 0, 1)$ *be an orthomodular lattice. Define*

$$x \cdot y = (x \vee y') \wedge y \quad \text{and} \quad x + y = ((x' \cdot y)' \cdot (x \cdot y')')'.$$

Then $\mathcal{R}(L) = (R; +, \cdot, 0, 1)$ *is a basic pseudoring satisfying the conditions*

(a) $x \cdot x = x$

(b) $x \cdot (1+y) = 0 \Rightarrow 1+(1+((1+(1+y)\cdot(1+x))\cdot(1+x))\cdot(1+x))\cdot(1+x) = y.$

Proof It is an immediate reflexion that

$$x \cdot x = (x \vee x') \wedge x = 1 \wedge x = x$$

proving (a).

Further, $1 \cdot x = (1 \vee x') \wedge x = x$ and $0 \cdot x = (0 \vee x') \wedge x = 0$. Hence,

$$1 + x = ((1' \cdot x)' \cdot (1 \cdot x')')' = (0' \cdot x'')' = (1 \cdot x)' = x'.$$

This yields $1 + 0 = 0' = 1$ proving (R1). Evidently,

$$x \cdot 1 = (x \vee 1') \wedge 1 = x$$

proving (R2) and $1 + (1 + x) = x'' = x$ proving (R3). For (R4) we compute

$$\begin{aligned}(1 + x \cdot (1 + y)) \cdot (1 + y) &= (x \cdot y')' \cdot y' = ((x \vee y) \wedge y')' \cdot y' \\ &= (((x \vee y) \wedge y')' \vee y) \wedge y' = ((x \vee y)' \vee y) \wedge y' = (x \vee y)'\end{aligned}$$

due to the orthomodular law since $(x \vee y)' \leq y'$. By symmetry we obtain (R4).

Since

$$\begin{aligned}1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y) &= 1 + ((x' \vee y) \wedge y')' \cdot y' \\ &= (((x \wedge y') \vee y) \wedge y')' = ((x' \vee y) \wedge y') \vee y = x' \vee y\end{aligned}$$

by the orthomodular law, for (R5) we have

$$\begin{aligned}&1 + (1 + (1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y)) \cdot (1 + z)) \cdot ((1 + x) \cdot (1 + z)) \\ &\quad = 1 + (1 + (x' \vee y) \cdot (1 + z)) \cdot ((1 + x) \cdot (1 + z)) \\ &\quad = 1 + (1 + (x' \vee y) \cdot z') \cdot ((x' \vee z) \wedge z') \\ &\quad = ((((x' \vee y) \cdot z')') \cdot ((x' \vee z) \wedge z'))' \\ &\quad = (((((x' \vee y) \vee z)' \vee z) \vee ((x' \vee z) \wedge z')') \wedge ((x' \vee z) \wedge z'))' \\ &\quad = ((((x' \vee y) \vee z) \wedge z') \wedge ((x' \vee z) \wedge z')) \vee ((x' \vee z) \wedge z')' \\ &\quad = ((x' \vee z) \wedge z') \vee ((x' \vee z) \wedge z')' = 1.\end{aligned}$$

It remains to prove (b). Assume $x \cdot (1+y) = 0$. Then $0 = x \cdot y' = (x \vee y) \wedge y'$ thus $x \vee y = y$ whence $x \leq y$. Thus

$$\begin{aligned} &1 + (1 + ((1 + (1 + y) \cdot (1 + x)) \cdot (1 + x)) \cdot (1 + x)) \cdot (1 + x) \\ &= 1 + (1 + (y \wedge x') \cdot (1 + x)) \cdot (1 + x) = (y \wedge x') \vee x = y \end{aligned}$$

by the orthomodular law. □

Now, we are going to describe a mutual relationship between basic pseudorings and basic algebras.

Theorem 2 *Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring. Define*

$$x \oplus y = 1 + (1 + x) \cdot (1 + y) \quad \text{and} \quad \neg x = 1 + x.$$

Then $\mathcal{A}(R) = (R; \oplus, \neg, 0)$ is a basic algebra.

Proof We will check the axioms of a basic algebra.

(BA1): $x \oplus 0 = 1 + (1 + x) \cdot (1 + 0) = 1 + (1 + x) \cdot 1 = 1 + (1 + x) = x$;

(BA2): $\neg\neg x = 1 + (1 + x) = x$;

(BA3): $\neg(\neg x \oplus y) \oplus y =$
$= 1 + (1 + (1 + (\neg x \oplus y))) \cdot (1 + y) = 1 + (\neg x \oplus y) \cdot (1 + y)$
$= 1 + (1 + (1 + (1 + x)) \cdot (1 + y)) \cdot (1 + y) = 1 + (1 + x \cdot (1 + y)) \cdot (1 + y)$
$= 1 + (1 + y \cdot (1 + x)) \cdot (1 + x) = 1 + (\neg y \oplus x) \cdot (1 + x)$
$= \neg(\neg y \oplus x) \oplus x$;

(BA4): $\neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z) =$
$= \neg(\neg((1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y)) \oplus z) \oplus (1 + (1 + x) \cdot (1 + z))$
$= (1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y)) \cdot (1 + z) \oplus (1 + (1 + x) \cdot (1 + z))$
$= 1 + (1 + (1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y)) \cdot (1 + z)) \cdot ((1 + x) \cdot (1 + z))$
$= 1$. □

We can prove the converse.

Theorem 3 *Let $\mathcal{A} = (A; \oplus, \neg, 0)$ be a basic algebra. Define*

$$x + y = \neg(x \oplus \neg y) \oplus \neg(\neg x \oplus y) \quad \text{and} \quad x \cdot y = \neg(\neg x \oplus \neg y) \quad \text{and} \quad 1 = \neg 0.$$

Then $\mathcal{R}(A) = (A; +, \cdot, 0, 1)$ is a basic pseudoring satisfying the correspondence identity

$$1 + (1 + (1 + x) \cdot y) \cdot (1 + x \cdot (1 + y)) = x + y. \qquad \text{(CI)}$$

Proof First we mention that

$$1 + x = \neg(1 \oplus \neg x) \oplus \neg(0 \oplus x) = \neg 1 \oplus \neg x = 0 \oplus \neg x = \neg x.$$

Now we check the axioms of a basic pseudoring.

(R1): $1 + 0 = \neg(1 \oplus \neg 0) \oplus \neg(\neg 1 \oplus 0) = \neg 1 \oplus \neg 0 = 1$;

(R2): $x \cdot 1 = \neg(\neg x \oplus \neg 1) = \neg(\neg x \oplus 0) = \neg\neg x = x$;

(R3): $1 + (1 + x) = \neg\neg x = x$;

(R4): $(1 + x \cdot (1 + y)) \cdot (1 + y)$
$= (\neg(x \cdot \neg y)) \cdot \neg y = (\neg x \oplus y) \cdot \neg y = \neg(\neg(\neg x \oplus y) \oplus y) = \neg(\neg(\neg y \oplus x) \oplus x)$
$= (1 + y \cdot (1 + x)) \cdot (1 + x)$;

(R5): $1 + (1 + (1 + (1 + ((1 + x) \cdot (1 + y))) \cdot (1 + y)) \cdot (1 + z)) \cdot ((1 + x) \cdot (1 + z))$
$= 1 + (1 + (1 + (x \oplus y) \cdot \neg y) \cdot \neg z) \cdot (\neg x \cdot \neg z)$
$= 1 + (1 + (\neg((x \oplus y) \cdot \neg y)) \cdot \neg z) \cdot \neg(x \oplus z)$
$= 1 + (1 + (\neg(x \oplus y) \oplus y) \cdot \neg z) \cdot \neg(x \oplus z)$
$= 1 + (\neg(\neg(x \oplus y) \oplus y) \oplus z) \cdot \neg(x \oplus z)$
$= 1 + \neg(\neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z))$
$= \neg(\neg(\neg(x \oplus y) \oplus y) \oplus z) \oplus (x \oplus z)$
$= 1$.

Hence, $\mathcal{R}(A) = (A; +, \cdot, 0, 1)$ is a basic pseudoring. It remains to prove (CI). For this, we compute

$$1 + (1 + (1 + x) \cdot y) \cdot (1 + x \cdot (1 + y))$$
$$= \neg(\neg(\neg x \cdot y) \cdot \neg(x \cdot \neg y)) = \neg(x \oplus \neg y) \oplus \neg(\neg x \oplus y) = x + y \qquad \square$$

In what follows we show that this relationship is in fact a one-to-one correspondence if $\mathcal{R}$ satisfies the correspondence identity.

Theorem 4 *(a) Let $\mathcal{A} = (A; \oplus, \neg, 0)$ be a basic algebra and $\mathcal{R}(A)$ the induced basic pseudoring and $\mathcal{A}(\mathcal{R}(A))$ the induced basic algebra. Then $\mathcal{A}(\mathcal{R}(A)) = \mathcal{A}$.*

(b) Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring satisfying the correspondence identity (CI), let $\mathcal{A}(R)$ be the induced basic algebra and $\mathcal{R}(\mathcal{A}(R))$ the induced basic pseudoring. Then $\mathcal{R}(\mathcal{A}(R)) = \mathcal{R}$.

Proof Denote by $\widehat{\oplus}$ and $\widehat{\neg}$ the binary and the unary operation of $\mathcal{A}(\mathcal{R}(A))$. Then clearly,

$$\widehat{\neg}x = 1 + x = \neg(1 \oplus \neg x) \oplus \neg(\neg 1 \oplus x) = 0 \oplus \neg x = \neg x$$

and

$$x \widehat{\oplus} y = 1 + (1 + x) \cdot (1 + y) = \neg(\neg x \cdot \neg y) = \neg(\neg(x \oplus y)) = x \oplus y$$

thus $\mathcal{A}(\mathcal{R}(A)) = \mathcal{A}$.

Denote by $\widehat{+}$ and $\widehat{\cdot}$ the binary operations of $\mathcal{R}(\mathcal{A}(R))$. Then, due to (CI) we compute

$$x \widehat{+} y = \neg(x \oplus \neg y) \oplus \neg(\neg x \oplus y) = (1 + x) \cdot y \oplus x \cdot (1 + y)$$
$$= 1 + (1 + (1 + x) \cdot y) \cdot (1 + x \cdot (1 + y)) = x + y$$

and

$$x \widehat{\cdot} y = \neg(\neg x \oplus \neg y) = 1 + ((1 + x) \oplus (1 + y))$$
$$= 1 + (1 + (1 + (1 + x)) \cdot (1 + (1 + y))) = 1 + (1 + x \cdot y) = x \cdot y$$

thus also $\mathcal{R}(\mathcal{A}(R)) = \mathcal{R}$. $\square$

Several interesting properties of basic pseudorings are described by the following theorem and its corollary.

Theorem 5 *Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring and $a, b \in R$. Then*

$$a + b = 0 \quad \text{if and only if} \quad a = b.$$

Proof Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring and $\mathcal{A}(R) = (R; \oplus, \neg, 0)$ the induced basic algebra. In $\mathcal{A}(R)$ we have $c \leq d$ if and only if $\neg c \oplus d = 1$. Since $x \leq x$ and $\neg x \leq \neg x$, we get $\neg x \oplus x = 1$ and $x \oplus \neg x = \neg\neg x \oplus \neg x = 1$ whence

$$x + x = \neg(x \oplus \neg x) \oplus \neg(\neg x \oplus x) = \neg 1 \oplus \neg 1 = 0 \oplus 0 = 0.$$

Assume now that $c, d \in R$ and $c \oplus d = 0$. Since $d \leq c \oplus d = 0$, we conclude $d = 0$ and hence $c = c \oplus 0 = c \oplus d = 0$, i.e.

$$c \oplus d = 0 \Rightarrow c = d = 0. \qquad (**)$$

Suppose $a, b \in R$ and $a + b = 0$. Then

$$\neg(a \oplus \neg b) \oplus \neg(\neg a \oplus b) = 0$$

and, by $(**)$, $\neg(a \oplus \neg b) = 0 = \neg(\neg a \oplus b)$, i.e. $a \oplus \neg b = 1$ and $\neg a \oplus b = 1$ thus $\neg a \leq \neg b$ and $a \leq b$. However, the first inequality yields $b \leq a$ thus $a = b$. □

Corollary 1 *(a) Every basic pseudoring satisfies the identity $x + x = 0$.*

(b) If a pseudoring $\mathcal{R}$ satisfies the identity $x \cdot y = y \cdot x$ then $\mathcal{A}(R)$ is a commutative basic algebra.

(c) If a basic algebra $\mathcal{A}$ is commutative then $\mathcal{R}(A)$ satisfies the identities $x \cdot y = y \cdot x$ and $x + y = y + x$.

In what follows, we are going to show that not only every basic algebra induces a basic pseudoring and vice versa as shown by Theorems 2 and 3 but also Theorem 1 can be inverted, i.e. every orthomodular lattice induces a basic pseudoring satisfying the conditions (a), (b) but also every such basic pseudoring induces an orthomodular latttice.

Now, we are ready to prove the following

Theorem 6 *Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring satisfying the identities (a) and (b) of Theorem 1. Define a binary relation $\leq$ on R as follows*

$$x \leq y \quad \text{if and only if} \quad x \cdot (1 + y) = 0.$$

Then $\leq$ is an order on R and $(R; \leq)$ is an orthomodular lattice where

$$x \vee y = 1 + (1 + x \cdot (1 + y)) \cdot (1 + y) \quad \text{and} \quad x' = 1 + x.$$

Proof Let $\mathcal{R} = (R; +, \cdot, 0, 1)$ be a basic pseudoring satisfying (a) and (b). Consider the induced basic algebra $\mathcal{A}(R) = (R; \oplus, \neg, 0)$. Then clearly

$$x \cdot (1 + y) = 0 \quad \text{iff} \quad \neg x \oplus y = 1 \quad \text{iff} \quad x \leq y$$

thus $\leq$ is an order on R and $(R; \leq)$ is the lattice induced by the basic algebra $\mathcal{A}(R)$ where $x \vee y = \neg(\neg x \oplus y) \oplus y = 1 + (1 + x \cdot (1 + y)) \cdot (1 + y)$ and $\neg x = 1 + x$ (as already shown by Theorem 2). Hence, for $x \wedge y = (x' \vee y')'$ we have that $(R; \vee, \wedge, ', 0, 1)$ is a bounded lattice with an antitone involution (i.e. $x'' = x$ and $x \leq y \Rightarrow y' \leq x'$).

Further, by (a) we have $x = x \cdot x = \neg(\neg x \oplus \neg x)$, i.e. $\neg x = \neg x \oplus \neg x$ and, due to the double negation law in $\mathcal{A}(R)$, also $x \oplus x = x$ for each $x \in R$. Thus $\neg x \vee x = \neg(x \oplus x) \oplus x = \neg x \oplus x = 1$ and, due to De Morgan law, also $x \wedge \neg x = \neg(\neg x \vee x) = \neg 1 = 0$ thus $x' = \neg x$ is a complement of x, i.e. $(R; \vee, \wedge, ', 0, 1)$ is an ortholattice.

Finally,

$$\begin{aligned} &1 + (1 + ((1 + (1 + y) \cdot (1 + x)) \cdot (1 + x)) \cdot (1 + x)) \cdot (1 + x) \\ &\quad = 1 + (1 + (y \wedge x') \cdot (1 + x)) \cdot (1 + x) = (y \wedge x') \vee x, \end{aligned}$$

thus $x \leq y \Rightarrow x \cdot (1 + y) = 0$ and, by (b) and the previous computation, $x \vee (x' \wedge y) = y$, which is the orthomodular law. Hence, $(R; \vee, \wedge, ', 0, 1)$ is an orthomodular lattice. □

References

[1] Beran, L.: Orthomodular Lattices. *Reidel Publ., Dordrecht,* 1985.

[2] Birkhoff, G.: Lattice Theory. *Publ. AMS, Providence,* 1967.

[3] Chajda, I., Halaš, R., Kühr, J.: Semilattice Structures. *Heldermann Verlag, Lemgo,* 2007.

[4] Chajda, I., Kolařík, M.: *Independence of axiom system of basic algebras.* Soft Computing **13**, 1 (2009), 41–43.

[5] Chajda, I., Länger, H.: *Ring-like structures corresponding to MV-algebras via symmetrical difference.* Sitzungsberichte ÖAW, Math.–Naturw. Kl. Abt. II **213** (2004), 33–41.

[6] Dorfer, G., Dvurečenskij, A., Länger H.: *Symmetric difference in orthomodular lattices.* Math. Slovaca **46** (1996), 435–444.

[7] Dorninger, D., Länger, H., Maczyński, M.: *The logic induced by a system of homomorphisms and its various algebraic characterizations.* Demonstratio Math. **30** (1997), 215–232.

[8] Dorninger, D., Länger, H., Maczyński, M.: *On ring-like structures occuring in axiomatic quantum mechanics.* Sitzungsberichte ÖAW, Math.–Naturw. Kl. Abt. II **206** (1997), 279–289.

[9] Dorninger, D., Länger, H., Maczyński, M.: *Lattice properties of ring-like quantum logics.* Intern. J. of Theor. Physics **39** (2000), 1015–1026.

[10] Shang, Y.: *Ring-like structures corresponding to pseudo MV-algebras.* Soft Computing **13**, 1 (2009), 71–76.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 33–52

A Visual Approach to Test Lattices*

GÁBOR CZÉDLI

Bolyai Institute, University of Szeged,
Szeged, Aradi vértanúk tere 1, Hungary 6720
e-mail: czedli@math.u-szeged.hu

(Received October 12, 2008)

Abstract

Let p be a k-ary lattice term. A k-pointed lattice $L = (L; \vee, \wedge, d_1, \ldots, d_k)$ will be called a *p-lattice* (or a *test lattice* if p is not specified), if $(L; \vee, \wedge)$ is generated by $\{d_1, \ldots, d_k\}$ and, in addition, for any k-ary lattice term q satisfying $p(d_1, \ldots, d_k) \leq q(d_1, \ldots, d_k)$ in L, the lattice identity $p \leq q$ holds in all lattices.

In an elementary visual way, we construct a finite p-lattice $L(p)$ for each p. If p is a canonical lattice term, then $L(p)$ coincides with the optimal p-lattice of Freese, Ježek and Nation [6]. Some results on test lattices and short proofs for known facts on free lattices indicate that our approach is useful.

Key words: Free lattice, test lattice, lattice identity, Whitman's condition.

2000 Mathematics Subject Classification: 06B25

1 Introduction

For a fixed natural number k, by a *k-pointed lattice* we mean a lattice L with k distinguished elements $d_1, \ldots, d_k$. For $\vec{d} = (d_1, \ldots, d_k) \in L^k$, the "$k$-pointed lattice" $(L; \vee, \wedge, d_1, \ldots, d_k)$ will be denoted by $(L; \vec{d})$. If p and q are k-ary lattice terms, then both $p = q$ and $p \leq q$ are called *lattice identities*. A lattice identity is said to be *trivial*, if it holds in all lattices.

We introduce a new concept. Given a k-ary lattice term $p = p(\alpha_1, \ldots, \alpha_k)$, we will call a k-pointed lattice $(L; \vec{d})$ a *p-lattice*, if

- $\{d_1, \ldots, d_k\}$ generates L, and
- for any k-ary lattice term q, $p(d_1, \ldots, d_k) \leq q(d_1, \ldots, d_k)$ in L if and only if $p \leq q$ is a trivial lattice identity.

*This research was partially supported by the NFSR of Hungary (OTKA), grant no. T 049433 and K 60148

We use the terminology *"test lattice"* if we do not want to specify p. That is, if $(L; \vec{d})$ is a p-lattice for some p, then it is also called a test lattice.

For example, if L is freely generated by $\{d_1, \ldots, d_k\}$, then it is obviously a p-lattice for every k-ary lattice term p. Beside other aims, we are going to give a new proof for the following result, which is not so obvious.

Proposition 1 [Freese and Nation [7], Freese, Ježek and Nation [6])] *For each lattice term p, there exists a* finite *p-lattice.*

Our first goal is to point out that test lattices *deserve some attention* independently from the well-developed theory of free lattices (see Freese, Ježek, Nation [6]). Hence we present Theorem 3, soon, and give new proofs for two more or less known properties of test lattices, see Theorems 5 and 6. Further, we give two easy applications. Namely, we demonstrate the usefulness of test lattices by giving a very short, new proof that free lattices satisfy Whitman's condition, see Corollaries 12 and 13, and also by solving (and generalizing) the following (not very difficult) exercise.

Exercise 2 *Let $p^\diamondsuit = (\alpha_1 \vee \alpha_2) \wedge (\alpha_1 \vee \alpha_3)$. Is there a non-trivial lattice identity $p^\diamondsuit \leq q$ that holds in the five-element non-modular lattice?*

Our second goal is to *construct* a finite k-pointed lattice $L(p)$, for each k-ary lattice term p, in a *conceptually simple way*, and to give an *elementary proof* that it is a p-lattice. To follow the rest of the paper until the "Historical remarks" section, the reader is assumed to be familiar only with the rudiments of lattice theory, that is, with a small fraction of, say, G. Grätzer [8]. The only outer reference used in our proof is Jónsson's type 3 representation theorem, see [10], and see also Theorem IV.4.4 in Grätzer [8].

Our third goal is to give a new approach that is *visual*, not just elementary. We *develop a visual toolkit* consisting of purely lattice theoretical results from this section and several statements (Lemmas 7, 8, 9, 14, 15 and Corollaries 10, 11) from Section 3. Although this toolkit is applied to prove some known or easy results only, the geometric perspective may serve a better understanding of the underlining reasons, and it may lead to further useful observations in the future.

Notice at this point that powerful tools from the theory of free lattices, see Freese, Ježek and Nation [6] and its references, have already given or may easily give shorter "standard" proofs to several of our statements. Hence, in the last section, our results will be related to [6]. However, if the necessary previous pages of [6] are also counted, then some of the standard proofs are lengthier than ours. Although we will give some hints to a standard proof in the last section, many readers will probably find easier to follow our approach.

Notice also that, opposed to the present paper, free lattices are hard to imagine visually. For example, $FL(\omega)$ is a sublattice of $FL(3)$ by Whitman [13], and this fact is an obstacle to a proper visual understanding of $FL(3)$, the free lattice on three generators. Hence we hope that our pictorial approach with

graphical background makes sense and contributes to a better understanding of free lattices.

Finally, notice at this point that the only outer reference, Jónsson's type 3 representation theorem, see [10] or Theorem IV.4.4 in [8], is also visual.

From now on, let $p = p(\alpha_1, \ldots, \alpha_k)$ be a fixed k-ary lattice term. We are going to construct a k-pointed lattice $L(p) = (L(p); d_1, \ldots, d_k)$ such that the following theorem holds.

Theorem 3 *$L(p) = (L(p); d_1, \ldots, d_k)$ is a finite p-lattice.*

By an *optimal p-lattice*, we mean a p-lattice that is a k-pointed lattice homomorphic image of any other p-lattice. The following corollary of Theorem 3 is straightforward and more or less evident.

Corollary 4 *For each lattice term p, there exists an optimal p-lattice $K(p)$. It is finite and it is unique up to k-pointed lattice isomorphism.*

The length of a lattice term q, to be defined in the usual syntactical way later, will be denoted by $\mathrm{length}(q)$. We say that p is a *canonical lattice term* if for every k-ary lattice terms q, $p =_{\mathrm{triv}} q$ implies $\mathrm{length}(p) \leq \mathrm{length}(q)$. Like every term, each canonical lattice term p is

- either a variable,
- or the meet of at least two terms,
- or the join of at least two terms.

In the first two cases we say that p is a *join-irreducible canonical term*. (This means that p represent a join-irreducible element of the free lattice generated by $\{\alpha_1, \ldots, \alpha_k\}$.)

Unfortunately, $L(p)$ is usually not an *optimal* p-lattice in general. For example, for $p^\natural = \big(\big((\alpha_1 \vee \alpha_2) \wedge (\alpha_1 \vee \alpha_2 \vee \alpha_3)\big) \vee \alpha_2\big) \wedge \alpha_4$, the $p^\natural$-lattice $L(p^\natural)$ is not optimal. As a compensation, we have the following two theorems.

Theorem 5 [essentially in Freese, Ježek and Nation [6]] *If p is a canonical lattice term, then $L(p)$ equals $K(p)$, the optimal p-lattice.*

Theorem 6 [Freese, Ježek and Nation [6]] *If p is a join-irreducible canonical lattice term, then $K(p) = L(p)$ is subdirectly irreducible.*

Notice that the assumption of join-irreducibility in Theorem 6 cannot be avoided. For example, $L(\alpha_1 \vee \alpha_2) = K(\alpha_1 \vee \alpha_2)$ is the four-element boolean lattice, which is subdirectly (and even directly) reducible. On the other hand, this assumption is not so restrictive. Indeed, if p is the join of its subterms $p_1, \ldots, p_n$, then, evidently, $p \leq_{\mathrm{triv}} q$ iff $p_i \leq_{\mathrm{triv}} q$ for $i = 1, \ldots, n$. Hence, to investigate if $p \leq_{\mathrm{triv}} q$, we can use the subdirectly irreducible $L(p_1), \ldots, L(p_n)$ instead of $L(p)$.

2 The construction of $L(p)$

We fix a set $X = \{\alpha_1, \ldots, \alpha_k\}$ of variables. Since we do not want to make a distinction between lattice terms that differ only modulo commutativity, associativity and idempotency, we give the following inductive definition of $T(X)$, the *set of lattice terms over X*.

- Every $\alpha_i \in X$ is a *doubly irreducible* member of $T(X)$ with $\mathrm{length}(\alpha_i) = 1$.
- Each element of $T(X) \setminus X$ is of length > 1, and it is either join-irreducible and meet-reducible, or meet-irreducible and join-reducible.
- If $q_1, \ldots, q_n$, $n \geq 2$, are *distinct* meet-irreducible members of $T(X)$ then $q = \bigwedge_{i=1}^{n} q_i$ belongs to $T(X)$. It is join-irreducible and meet-reducible, and we have $\mathrm{length}(q) = 1 + \sum_{i=1}^{n} \mathrm{length}(q_i)$. The terms $q_1, \ldots, q_n$ are called the *meetands* of p.
- If $q_1, \ldots, q_n$, $n \geq 2$, are *distinct* join-irreducible members of $T(X)$ then $q = \bigvee_{i=1}^{n} q_i$ belongs to $T(X)$. It is meet-irreducible and join-reducible, and we have $\mathrm{length}(q) = 1 + \sum_{i=1}^{n} \mathrm{length}(q_i)$. The terms $q_1, \ldots, q_n$ are called the *joinands* of p.
- Each member of $T(X)$ is obtained by the previous rules in a finite number of steps.

Notice that for each $q \in T(X)$, either q has no meetand or it has at least two meetands. Dually, the same holds for the joinands of q. For concrete terms in examples, we will write $q_1 \vee \cdots \vee q_n$ rather than $\bigvee_{i=1}^{n} q_i$, and similarly for the meet. By a *join-free* term we mean a variable or a meet of variables.

Our definition of terms is only slightly different from that in page 10 of Freese, Ježek and Nation [6]. Namely, $x \vee y \vee z$ and $x \vee (y \vee z)$ are different terms in [6] but $x \vee (y \vee z)$ is *not* a term in the present paper. Notice also that the (ir)reducibility of a term has not much to do with the (ir)reducibility of the corresponding element of the free lattice $FL(X)$. For example, $(\alpha_1 \vee \alpha_2) \wedge (\alpha_1 \vee \alpha_2 \vee \alpha_3)$ is a join-irreducible and meet-reducible term, but it represents a join-reducible and meet-irreducible element of $FL(X)$.

The *color set* $C(p)$ of p is defined by the following induction. (The terminology "color" will be clear soon.)

- $C(\alpha_i) = \{\alpha_i\}$
- If p is join-reducible with joinands $p_1, \ldots, p_n$, then $C(p) = C(p_1) \cup \cdots \cup C(p_n)$.
- If p is meet-reducible, then let

$$\begin{aligned} M(p) &= \{s : s \text{ is a meetand of } p \text{ with } \mathrm{length}(s) > 1\} \\ &= \{s : s \text{ is a meetand of } p \text{ and } s \text{ is join-reducible}\}, \end{aligned} \tag{1}$$

and define

$$C(p) = \{p\} \cup \bigcup_{s \in M(p)} C(s)\,.$$

Notice that all elements of $C(p)$ are join-irreducible terms. For an example of $C(p)$, see the set of colors of $H(p^\sharp)$ in Figure 3.

Given a relation E, let E^* denote its transitive closure. Throughout the paper, by a *p-graph* or, shortly, *graph* we mean a structure $G = (V, E, \mathrm{col})$ such that

- $(V, E) = (V(G), E(G))$ is a directed graph without loops and multiple edges. That is, V is a nonempty set, the vertex set, and $E \subseteq V^2$, the edge set, is an irreflexive and antisymmetric relation;
- $\mathrm{col}\colon E \to C(p)$, that is, each edge $e \in V$ has a unique color $\mathrm{col}(e) \in C(p)$;
- E^*, also denoted by $\sqsubset$, is a partial ordering of V with least element, called *the left endpoint of* G, and greatest element, called *the right endpoint*.

Unless otherwise specified, the left and right endpoints of our graphs are denoted by x_0 and x_1, respectively. The subgraphs we are going to consider are also graphs in the above sense. However, a proper subgraph of a p-graph G is (isomorphic with) a q-graph for some term q distinct from p.

In figures, the edges are directed from left to right by convention, so the orientation of edges is not indicated. An edge $(a, b) \in E$ is called a *covering edge* of G, if there is no $c \in V$ with $a \sqsubset c \sqsubset b$. To ease our notations, we will say that (a, r, b) is an "edge of G" to express that $(a, b) \in E$ and $r = \mathrm{col}((a, b))$.

If $\{G_1, G_2\}$ is a two-element set of graphs, then a *4-series connection* of this set is obtained from two copies of G_1 and two copies of G_2, all the four copies being pairwise disjoint, via identifying some endpoints as depicted in Figure 1. Of course, this depends on the order of G_1 and G_2, whence $\{G_1, G_2\}$ has two 4-series connections.

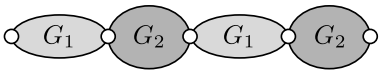


Figure 1: A 4-series connection of $\{G_1, G_2\}$

If $\{G_1, \ldots, G_n\}$ is an n-element set of graphs, then each 4-series connection of this set is obtained in the following way: for some $i \in \{1, \ldots, n\}$ and a 4-series connection H of $\{G_1, \ldots, G_{i-1}, G_{i+1}, \ldots, G_n\}$, we form a 4-series connection of H and G_i. Notice that $\{G_1, \ldots, G_n\}$ has exactly $n!$ many 4-series connections; for $n = 3$ one of them is depicted in Figure 2

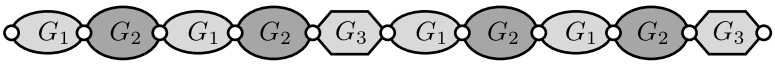


Figure 2: A 4-series connection of $\{G_1, G_2, G_3\}$

Next, we define a sequence $\mathbf{G}_j(p)$ of sets of p-graphs associated with p via induction on j as follows. A particular case,

$$p^\sharp = \alpha_1 \wedge \Big(\alpha_2 \vee \big(\alpha_3 \wedge (\alpha_4 \vee \alpha_5)\big)\Big) \wedge \big(\alpha_2 \vee (\alpha_3 \wedge \alpha_5)\big)$$

is depicted in Figure 3. The reader is advised to look at this figure often while reading the following definition. In Figure 3, $H_j(p^\sharp)$ is just one member of $\mathbf{G}_j(p^\sharp)$.

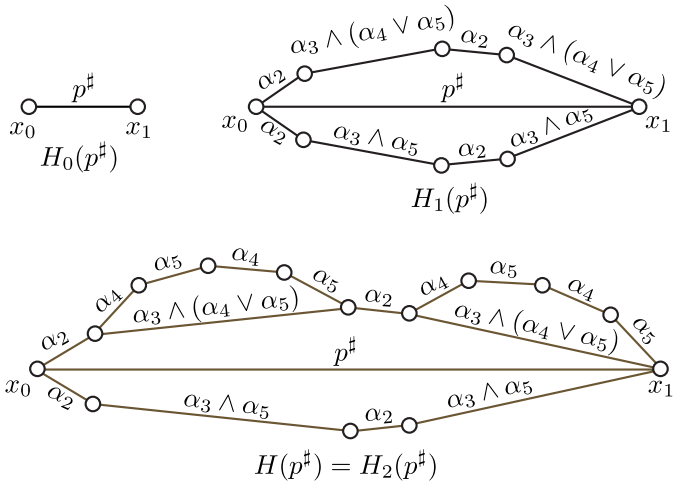


Figure 3: Constructing a member of $\mathbf{G}(p^\sharp)$

If p is join-irreducible, then $\mathbf{G}_0(p)$ consists of a single graph $H_0(p)$. This graph has only two vertices, x_0 and x_1, and only one edge, (x_0, x_1). This edge is colored by p.

If $s = \bigvee_{i=1}^n t_i$ is a join-reducible lattice term with joinands $t_1, \ldots, t_n$, then any 4-series connection of the set $\{H_0(t_1), \ldots, H_0(t_n)\}$ is called an *s-arc*; for $n = 3$ see Figure 4.

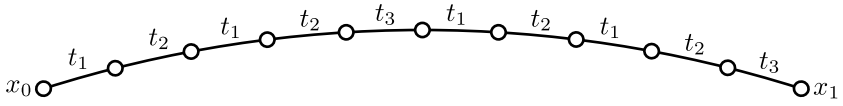


Figure 4: An s-arc, if $s = \bigvee_{i=1}^3 t_i$

If $p = \bigvee_{i=1}^n p_i$ is join-reducible, then let $\mathbf{G}_0(p)$ be the set of all p-arcs.

If $j \geq 1$ and each *covering* edge of every member of $\mathbf{G}_{j-1}(p)$ is colored by a join-free term (variable or meet of variables), then let $\mathbf{G}_j(p) = \mathbf{G}_{j-1}(p)$.

In the opposite case we obtain $\mathbf{G}_j(p)$ from $\mathbf{G}_{j-1}(p)$ in the following way. Take a member $H = H_{j-1}(p) \in \mathbf{G}_{j-1}(p)$. Consider each covering edge (a, r, b) of H whose color r is not join-free. Then r is meet-reducible. For each $s \in M(r)$, see formula (1), we glue an s-arc to H by identifying the left and right endpoints of this arc with a and b, respectively, but keeping other vertices of this arc disjoint from the vertices of H and that of any other arc glued to H. We glue all the necessary arcs to all covering edges with not join-free colors at the same time such that these arcs should be disjoint from each other and from H as

much as possible and, in addition,

$$\text{we must use isomorphic } s\text{-arcs for all } r\text{-colored covering edges.} \qquad (2)$$

This way we obtain H^+. Finally, let $\mathbf{G}_j(p) = \{H^+ : H \in \mathbf{G}_{j-1}(p)\}$.

If $\mathbf{G}_j(p)$ is different from $\mathbf{G}_{j-1}(p)$, then the maximal length of not join-free colors on covering edges in members of $\mathbf{G}_{j-1}(p)$ decreases when we pass from $\mathbf{G}_{j-1}(p)$ to a $\mathbf{G}_j(p)$. Hence there is a smallest $n \in \mathbf{N}$ with $\mathbf{G}_n(p) = \mathbf{G}_{n-1}(p)$. Let $\mathbf{G}(p) = \mathbf{G}_{n-1}(p)$ for this n. Clearly, the colors of *covering* edges of any member of $\mathbf{G}(p)$ are join-free.

Let us agree on the following convention: $H(p)$ will always denote an arbitrarily *fixed* graph in $\mathbf{G}(p)$. Then $H_j(p)$ will stand for the unique graph in $\mathbf{G}_j(p)$ that occurs in the inductive definition leading to $H(p)$. For technical reasons, $H_{-1}(p)$ will denote the empty graph with no edge.

It is evident from the construction that the set of colors occurring on edges of each $H(p) \in \mathbf{G}(p)$ is exactly $C(p)$.

An edge (a, r, b) of a p-graph $H(p) \in \mathbf{G}(p)$ is called an *α_i-edge* if $r = \alpha_i$ or α_i is a meetand of r. (Notice that an α_i-edge is not necessarily α_i-colored!) Let $V(p)$ and $E(p)$ denote the vertex set and the edge set of $H(p)$, respectively, and let $\mathrm{Equ}(V(p))$ stand for the lattice of equivalences on $V(p)$. The smallest member of $\mathrm{Equ}(V(p))$ collapsing the endpoints of each α_i-edge will be denoted by $\alpha_i|_{H(p)}$. In other words, for $a, b \in V(p)$ we have $(a, b) \in \alpha_i|_{H(p)}$ iff there are vertices $c_0 = a, c_1, \ldots, c_n = b$, $n \geq 0$, such that for all $i = 0, 1, \ldots, n-1$ either (c_i, c_{i+1}) or (c_{i+1}, c_i) is an α_i-edge. Still in other words: if there is an *undirected path* from a to b whose edges are α_i-edges. Such a path will be called an *α_i-path*.

Finally, the p-lattice we wanted to construct is

$$L(p) = (L(p); d_1, \ldots, d_k) := \big([\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}]; \alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}\big) \qquad (3)$$

where $H(p) \in \mathbf{G}(p)$ and $[\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}]$ is the sublattice of $\mathrm{Equ}(V(p))$ generated by $\{\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}\}$. Since $L(p)$ will be appropriate for any choice of $H(p)$ in $\mathbf{G}(p)$, we will not investigate if $L(p)$ depends on $H(p)$ in the abstract sense or not.

3 Visual statements and proofs

In forthcoming computations, $\leq^{(n)}$, $\leq^{(\mathrm{T}n)}$, $\leq^{(\mathrm{C}n)}$ and $\leq^{(\mathrm{L}n)}$ will indicate that Formula (n), Theorem n, Corollary n and Lemma n is applied, respectively. Analogous superscript are used with $=$, $\leq_{\mathrm{triv}}$ and $=_{\mathrm{triv}}$. Let $H(p) \in \mathbf{G}(p)$. For a k-ary lattice term t, the equivalence relation $t(\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}) \in L(p) \subseteq \mathrm{Equ}(V(p))$ will be denoted by $t|_{H(p)}$. For $t \in X$, a variable, $t|_{H(p)}$ has its previous meaning. By an (undirected) *$t|_{H(p)}$-path* we mean an (undirected) path U such that for every (undirected) edge (a, b) of U, $(a, b) \in t|_{H(p)}$. Similarly, for $n \geq 1$ and $\mu_1, \ldots, \mu_n \in \mathrm{Equ}(V(p))$, an (undirected) path U is said to be an (undirected) *$\mu_1 \cup \cdots \cup \mu_n$-path*, if $(a, b) \in \mu_1 \cup \cdots \cup \mu_n$ for every (undirected) edge $(a, b) \in U$.

In what follows, the graph $H(p) = (V(p), E(p), \mathrm{col}) \in \mathbf{G}(p)$ is fixed. Let (a, r, b) be an edge of $H(p)$. Then the set $\{c : a \sqsubseteq c \sqsubseteq b\}$ of vertices determines a full subgraph denoted by $S(a, r, b)$. The left and right endpoint of $S(a, r, b)$ are a and b, respectively. If the color r is irrelevant, then we write $S(a, , b)$ instead of $S(a, r, b)$. Notice that $S(x_0, p, x_1)$ is $H(p)$, provided p is a join-irreducible term. It is clear from the construction that $S(a, r, b)$ is a graph. Moreover,

$$S(a, r, b) \cong H(r) \text{ for a (unique) } H(r) \in \mathbf{G}(r). \tag{4}$$

Notice that there is exactly one isomorphism between $H(r)$ and $S(a, r, b)$.

The following lemma is evident by the construction; we formulate it for later reference only.

Lemma 7 *Suppose that (a, r, b) is an edge of $H(p)$. Let $x, y \in V(p)$ such that x belongs to $S(a, r, b)$ but y does not. Let U be an undirected path in $H(p)$ from x to y. Then U goes through at least one of a and b.*

The following lemma is the heart our paper. Roughly saying, its first part states that the "outer world" does not disturb our equivalences inside $S(a, r, b)$.

Lemma 8 *Let t be a k-ary lattice term.*

(a) *If (a, r, b) is an edge of $H(p)$ and x and y are vertices of $S(a, r, b)$ then*

$$(x, y) \in t|_{H(p)} \quad \textit{iff} \quad (x, y) \in t|_{S(a,r,b)}.$$

(b) *Let x and y be vertices of $H(p)$. Then $(x, y) \in t|_{H(p)}$ iff there is an undirected $t|_{H(p)}$-path from x to y. In other words, $t|_{H(p)}$ is the equivalence generated by $t|_{H(p)} \cap E(p)$.*

Proof The proof is an induction on the length of t. The induction hypothesis is the *conjunction* of (a) and (b) for all terms t' shorter than t and for any p. (Notice that the induction would not work for (a) or (b) separately.) We assume that $x \neq y$. The "if" part of (a) and that of (b) are trivial (and, implicitly, will be used in the proof). So we will focus on the "only if" parts. Let $H_0(p)$, $H_1(p)$, $H_2(p)$, ... be the series of graphs that leads to $H(p)$ according to its inductive definition. We have to fix some notations according to p.

If p is join-irreducible, then let $m = \ell = c(1) = 1$, let $z_0 = x_0$, the left endpoint, $z_1 = x_1$, the right endpoint, and let $p_1 = p_{c(\ell)}$ stand for p.

If $p = \bigvee_{i \in F} p_i$ is join-reducible, then let $\{z_0 = x_0, z_1, \ldots, z_{m-1}, z_m = x_1\}$ be the vertex set and $\{(z_{i-1}, p_{c(i)}, z_i) : i = 1, 2, \ldots, m\}$ be the edge set of $H_0(p)$. Here all the $c(i)$ belong to F. If we wrote p_i in Figure 4 instead of t_i, then we would obtain an illustration for the case $F = \{1, 2, 3\}$. Clearly, there is a unique $\ell \in \{1, \ldots, m\}$ such that both a and b are vertices of $S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$. Therefore, $S(a, r, b)$ is a full subgraph of $S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$.

Case 1: $t = \beta \in X$ is a variable. Part (b) is evident. To prove (the "if" part of) (a), let us assume that $(x, y) \in \beta|_{H(p)}$. We also assume that $(a, b) \neq (x_0, x_1) = (z_0, z_n)$, because otherwise $S(a, r, b) = H(p)$, and there is nothing to prove.

Next, we assume that $(a, b) = (z_{\ell-1}, z_\ell)$. By the definition of $\beta|_{H(p)}$, there is a *shortest* undirected β-path in $H(p)$ that connects x and y. It follows from the structure of $H_0(p)$ (even without invoking Lemma 7) that any path exiting $S(a, r, b) = S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$ at a can enter $S(a, r, b)$ again only at a, and the same holds for b. Hence our shortest β-path cannot exit $S(a, r, b)$ at all, and we conclude that $(x, y) \in \beta|_{S(a,r,b)}$.

Now that we have settled the easier subcases, we assume that $\{a, b\} \not\subseteq \{z_{\ell-1}, z_\ell\}$. Then there is a $j \geq 1$ such that a and b belong to $H_j(p)$, in fact to $S(z_{\ell-1}, p_{c(\ell)}, z_\ell) \cong H_j(p_{c(\ell)})$, but at least one of a and b is not in $H_{j-1}(p)$. Hence there is an edge (e, q, f) in $H_{j-1}(p)$, in fact in $S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$, and there is an $s \in M(q)$ such that the edge (a, r, b) belongs to the s-arc glued to the edge (e, q, f) when $H_j(p)$ was obtained from $H_{j-1}(p)$, see Figure 5. This uniquely determined s-arc will be called the *supporting arc* of $S(a, r, b)$.

From the definition of an arc it follows that there is another r-colored edge of our s-arc, say (c, r, d). Notice that, opposed to Figure 5, $\{a, b, c, d\} \cap \{e, f\}$ is not necessarily empty. However, $\{a, b\} \cap \{c, d\} = \emptyset$ by the construction.

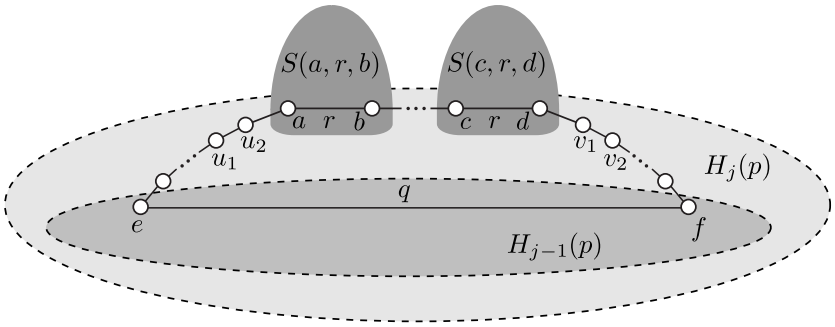


Figure 5: $S(a, r, b)$ and its supporting arc

Since $(x, y) \in \beta|_{H(p)}$, there is a *shortest* undirected β-path U in $H(p)$ from x to y. If U goes entirely in $S(a, r, b)$, then $(x, y) \in \beta|_{S(a,r,b)}$ and we are ready with this subcase. So assume that U leaves $S(a, r, b)$. Since U is a shortest path, we can assume by Lemma 7 that U leaves $S(a, r, b)$ at a and enters it again at b. (Interchanging a and b would make no difference in what follows.) Then, in the order given below, U must go through the vertices x, a, u_1, u_2, …, e of the supporting arc, then through f, …, v_2, v_1, d, c, …, b, y, see Figure 5. (Notice that these vertices are not necessarily consecutive vertices of U.)

Let W denote the segment of U between d and c. Since every path from d to c outside $S(c, r, d)$ should go through a, which would contradict to the assumption that U is the shortest path, we conclude that W goes entirely in $S(c, r, d)$. By stipulation (2), there is a graph isomorphism from $S(c, r, d)$ to $S(a, r, b)$. Replacing the "outer" a, …, e, f, …, b segment of U by the image of W, we obtain a shorter β-path from x to y, a contradiction. Hence $(x, y) \in \beta|_{S(a,r,b)}$, completing the case where t is a variable.

Case 2: t is meet-reducible with meetands $t_1, \dots, t_v$. We assume that the lemma is valid for the meetands $t_1, \dots, t_v$. Then part (a) of the lemma is clearly valid for t. To prove part (b), suppose that $(x, y) \in t|_{H(p)}$ and $x \neq y$.

Subcase 2.1: p is join-irreducible, that is, $m = 1$. Let j denote the smallest subscript such that both x and y belongs to $H_j(p)$; we will prove (b) for t by induction on j.

If $j = 0$, then $\{x, y\} = \{x_0, x_1\}$, whence (x, y) is an undirected edge, which is an undirected $t|_{H(p)}$-path. This settles the case $j = 0$.

Next, let $j > 0$, and assume that (b) holds for t and any two vertices from $H_{j-1}(p)$. We can assume that (x, y) is not an edge of $H(p)$. Let, say, x do not belong to $H_{j-1}(p)$. Then x belongs to an arc glued to $H_{j-1}(p)$, cf. Figure 5 with $x = a$. Suppose that e, the left endpoint of this arc, is *nearer* the edge $(x, r, b) = (a, r, b)$ than f. (The supporting arc consists of an even number of edges, so either e or f is strictly nearer.) According to the position of y, we have to distinguish two possibilities.

Sub-subcase 2.1.1: y is not on this arc. Let $i \in \{1, \dots, v\}$ be an arbitrary subscript. By the induction hypothesis, there is an undirected $t_i|_{H(p)}$-path U_i from x to y. This path leaves the arc at e or f.

We claim that there is an undirected $t_i|_{H(p)}$-path V_i from x to e. This is clear if U_i leaves the arc at e, so assume that it leaves the arc at f. Since e is nearer the edge (a, r, b) than f (in short, e is *near* and f is *far from* the edge (a, r, b)), each color on the arc between e and $a = x$ occurs between a and f. For example, let r' be the color of the edge (u_2, u_1), and also of the edge (v_1, v_2). Since U_i goes through v_1 and v_2, $(v_1, v_2) \in t_i|_{H(p)}$. Since part (a) is already valid for t_i, we get $(v_1, v_2) \in t_i|_{S(v_1, r', v_2)}$. It follows from stipulation (2) that

$$S(v_1, r', v_2) \cong S(u_2, r', u_1), \tag{5}$$

so $(u_2, u_1) \in t_i|_{S(u_2, r', u_1)}$, whence $(u_2, u_1) \in t_i|_{H(p)}$. This argument shows that the segment of the arc between e and $x = a$ is an undirected $t_i|_{H(p)}$-path, indeed.

This holds for all $i \in \{1, \dots, v\}$, and we conclude that there is an (undirected) $t|_{H(p)}$-path from x to $e \in H_{j-1}(p)$. Similarly, there is a $t|_{H(p)}$-path from y to a vertex $y' \in H_{j-1}(p)$. (Possibly, $y' = y$.) Since $(x, x'), (y, y') \in t|_{H(p)}$, the transitivity of $t|_{H(p)}$ implies that $(x', y') \in t|_{H(p)}$. By the induction hypothesis on j, there is an undirected $t|_{H(p)}$-path between x' and y'. Composing the three paths mentioned we obtain an undirected $t|_{H(p)}$-path from x to y, as requested.

Sub-subcase 2.1.2: y is on the same arc as x. Let $i \in \{1, \dots, v\}$, and consider a shortest (undirected) $t_i|_{H(p)}$-path U_i that connects x and y. Related to the arc, there are two possibilities for U_i. We say that it is a *detour*, if it consists of e, f, and all vertices of the arc that are not strictly between x and y. On the other hand, if U_i consists of all edges of the arc that are between x and y, then we say that U_i is a *straight path*. Clearly, U_i is either a detour or a straight path (but not both).

Similarly, there are two possibilities for the position of x and y; note that both possibilities can hold simultaneously. Namely, either x and y are *far* in

the sense that each color occurring on the arc occurs between x and y, or x and y are *near* in the sense that each such color occurs not only between x and y.

Now assume that x and y are far. We claim that there is a $t_i|_{H(p)}$ detour connecting x and y. We have to investigate only the case when U_i is a straight path. Then, similarly to the argument above with (5), part (a) for t_i gives that every edge of the arc is a $t_i|_{H(p)}$-edge. By transitivity, $(e, f) \in t_i|_{H(p)}$. Hence the (unique) detour from x to y is an undirected $t_i|_{H(p)}$-path, indeed. This holds for all i, whence this detour is a $t|_{H(p)}$-path connecting x and y.

If x and y are near, then a straightforward analogous argument shows that the (unique) straight path from x to y is an undirected $t|_{H(p)}$-path.

Subcase 2.2: p is join-reducible, that is, $m \geq 2$. Firstly, assume that x and y belong to the same subgraph $S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$. For all $i \in \{1, \ldots, v\}$, $(x, y) \in t_i|_{S(z_{\ell-1}, p_{c(\ell)}, z_\ell)}$ by part (a) of the lemma. Since $p_{c(\ell)}$ is join-irreducible and we have $S(z_{\ell-1}, p_{c(\ell)}, z_\ell) \cong H(p_{c(\ell)})$ for an appropriate $H(p_{c(\ell)}) \in \mathbf{G}(p_{c(\ell)})$, the previous case implies the existence of a $t|_{S(z_{\ell-1}, p_{c(\ell)}, z_\ell)}$ path from x to y. It is clearly a $t|_{H(p)}$-path.

Secondly, assume that x belongs to the subgraph $S(z_{\ell-1}, p_{c(\ell)}, z_\ell)$ and y belongs to $S(z_{h-1}, p_{c(h)}, z_h)$. Let, say, $\ell < h$. We know that there are shortest $t_i|_{H(p)}$-paths U_i from x to y for $i \in \{1, \ldots, v\}$. There can be no detours now, so all these paths go through $z_\ell, z_{\ell+1}, \ldots, z_{h-1}$. This holds for all $i \in \{1, \ldots, v\}$, whence (x, z_ℓ), $(z_\ell, z_{\ell+1})$, $\ldots$, (z_{h-1}, y) belong to $t|_{H(p)}$. Since the components of each of these pairs belong to the same subgraph, the previous case yields that these components can be connected by $t|_{H(p)}$-paths. Putting these paths together, we obtain a $t|_{H(p)}$-path from x to y.

Case 3: t is join-reducible with joinands $t_1, \ldots, t_v$. Suppose that the lemma is valid for these joinands. Since $t_i|_{H(p)}$-paths are $t|_{H(p)}$-paths as well, part (b) of the lemma is evident.

The argument for part (a) is similar to the case when t was a variable, so we will use the notations introduced in connection with Figure 5. In particular, x and y are vertices of $S(a, r, b)$ and $(x, y) \in t|_{H(p)} = t_1|_{H(p)} \vee \cdots \vee t_v|_{H(p)}$. Using the description of joins in $\mathrm{Equ}(V(p))$ and then the induction hypothesis for the t_i, we obtain a shortest undirected $t_1|_{H(p)} \cup \cdots \cup t_v|_{H(p)}$-path U connecting x and y. We want to show that U goes entirely in $S(a, r, b)$.

This is evident if (a, b) is an edge of $H_0(p)$, that is, it is of the form $(z_{\ell-1}, z_\ell)$. So, assume that (a, b) is not an edge of $H_0(p)$ and, by way of contradiction, assume that U exits $S(a, r, b)$. Then a segment of U connects c and d within $S(c, r, d)$. Each edge of this segment is collapsed by some $t_i|_{H(p)}$, whence by $t_i|_{S(c,r,d)}$ according to the induction hypothesis. Using the isomorphism between $S(c, r, d)$ and $S(a, r, b)$, we obtain a *shorter* path from a to b within $S(a, r, b)$ whose edges are collapsed by appropriate $t_i|_{S(a,r,b)}$, whence by $t_i|_{H(p)}$.

This contradiction shows that U goes in $S(a, r, b)$, indeed. By the induction hypothesis, if an edge of U is collapsed by $t_i|_{H(p)}$ then it is collapsed by $t_i|_{S(a,r,b)}$, and therefore by $t|_{S(a,r,b)}$. Finally, $(x, y) \in t|_{S(a,r,b)}$ follows by transitivity. □

The next lemma will obviously imply Theorem 3 and Proposition 1.

Lemma 9 *The k-pointed lattice $L(p)$ defined by formula (3) is a p-lattice. Moreover, the following three conditions are equivalent for any k-ary lattice term q:*

(a) $p \leq_{\mathrm{triv}} q$;

(b) $p(\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}) \leq q(\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)})$ *in* $L(p)$;

(c) $(x_0, x_1) \in q(\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)})$.

Proof (a) implies (b) trivially. An easy induction on the length of p gives $(x_0, x_1) \in p(\alpha_1|_{H(p)}, \ldots, \alpha_k|_{H(p)}) = p|_{H(p)}$, whence (b) implies (c).

Next, suppose (c), let L be an arbitrary lattice, and let $\beta_1, \ldots, \beta_k \in L$. We know from Jónsson [10] that each lattice has a type 3 representation, see also Theorem IV.4.4 in Grätzer [8]. Hence we can assume that L is a sublattice of some $\mathrm{Equ}(Y)$ and $\gamma \vee \delta = \gamma \circ \delta \circ \gamma \circ \delta$ holds for any $\gamma, \delta \in L$. Let $(y_0, y_1) \in p(\beta_1, \ldots, \beta_k)$. A straightforward induction on the length of p shows the existence of a map $\varphi\colon V(p) \to Y$ such that $x_0 \mapsto y_0$, $x_1 \mapsto y_1$, and for each α_i-edge (u, α_i, v) of $H(p)$, we have $(u\varphi, v\varphi) \in \beta_i$. The same kind of induction on the length of q shows that, for $a, b \in V(p)$, if $(a, b) \in q|_{H(p)}$, then $(a\varphi, b\varphi) \in q(\beta_1, \ldots, \beta_k)$. In particular, $(y_0, y_1) \in (x_0\varphi, x_1\varphi) \in q(\beta_1, \ldots, \beta_k)$. Hence $p \leq q$ holds in L, so $p \leq_{\mathrm{triv}} q$. □

Corollary 10 *Let (a, r, b) be an edge of $H(p) \in \mathbf{G}(p)$, and let t be an arbitrary k-ary lattice term. Then*

(a) $r|_{H(p)}$ *is the smallest element of $L(p)$ that collapses a and b;*

(b) $(a, b) \in t|_{H(p)}$ *if and only if* $r \leq_{\mathrm{triv}} t$;

(c) $r|_{H(p)} \leq t|_{H(p)}$ *if and only if* $r \leq_{\mathrm{triv}} t$.

Proof By (4), there is a (unique) graph $H(r) \in \mathbf{G}(r)$ such that $H(r) \cong S(a, r, b)$. Since $r \leq_{\mathrm{triv}} r$, we conclude $(x_{0,H(r)}, x_{1,H(r)}) \in r|_{H(r)}$ by Lemma 9. Hence $(a, b) \in r|_{S(a,r,b)}$, and Lemma 8(a) gives $(a, b) \in r|_{H(p)}$.

Next, assume that $(a, b) \in t|_{H(p)}$. We obtain from Lemma 8(a) that $(a, b) \in t|_{S(a,r,b)}$. Hence $(x_{0,H(r)}, x_{1,H(r)}) \in t|_{H(r)}$, so $r \leq_{\mathrm{triv}} t$ by Lemma 9. This proves part (b) and completes the proof of part (a). Finally, (c) is an evident consequence of (a) and (b). □

Corollary 11 *Suppose $\mu \in L(p)$, (a, r, b) is an edge of $H(p)$, and t is a k-ary lattice term. Then*

(a) *it depends only on r if $(a, b) \in \mu$;*

(b) $\mu = \bigvee\{s|_{H(p)} : s \in C(p)$ *and all s-colored edges are collapsed by* $\mu\}$.

(c) $t|_{H(p)} = \bigvee\{s|_{H(p)} : s \in C(p)$ *and* $s \leq_{\mathrm{triv}} t\}$.

Proof Since μ is of the form $t|_{H(p)}$, part (a) follows from Corollary 10(b).

Let $B = \{s \in C(p) : \text{all } s\text{-colored edges are collapsed by } \mu\}$ and $\nu = \bigvee_{s\in B} s|_{H(p)}$. Suppose (c, s, d) is an edge with $(c, d) \in \mu = t|_{H(p)}$. Then $s \in B$ by part (a), and $(c, d) \in s|_{H(p)}$ by Corollary 10(a). Hence $(c, d) \in \nu$. Therefore, Lemma 8(b) implies $\mu = t|_{H(p)} \leq \nu$. Conversely, Corollary 10(b) yields that $s \leq_{\text{triv}} t$ for every $s \in B$. Hence $\nu \leq t|_{H(p)} = \mu$, proving part (b).

Finally, part (c) is a consequence of part (b) and Corollary 10(b). □

The following two corollaries (and the dual of the second one) say that free lattices satisfy Whitman's condition. Their original proof in [13] is a bit lengthy. Based on A. Day [5], the approach of Freese, Ježek, Nation [6] to Whitman's condition is shorter. Now, armed with the basic properties of $L(p)$, we are going to give an even shorter proof. Since it is visual, it reveals some new ingredients from the underlying reasons.

Corollary 12 (Whitman [13]) *Let p be a meet-reducible lattice term with with meetands $p_1, \ldots, p_u$, and let q be a join-reducible lattice term with joinands $q_1, \ldots, q_v$. Assume that $p \leq_{\text{triv}} q$. Then either $p_i \leq_{\text{triv}} q$ for some $i \in \{1, \ldots, u\}$ or $p \leq_{\text{triv}} q_j$ for some $j \in \{1, \ldots, v\}$.*

Proof Lemma 9 yields that $(x_0, x_1) \in q|_{H(p)} = q_1|_{H(p)} \vee \cdots \vee q_v|_{H(p)}$. Hence there exists a *shortest* undirected $q_1|_{H(p)} \cup \cdots \cup q_v|_{H(p)}$-path U that connects x_0 and x_1.

Firstly, if U is of length 1, then $p \leq_{\text{triv}} q_j$ for some j by Lemma 9.

Secondly, if length$(U) \geq 2$, then U goes through all vertices of a unique p_i-arc glued to $H_0(p)$. Hence U goes within $H(p_i)$. Let (c, s, d) be an edge of U. Then $(c, d) \in q|_{H(p)}$. Using Lemma 8(a) twice, we get $(c, d) \in q|_{S(c,s,d)}$ and $(c, d) \in q|_{H(p_i)}$. By transitivity, $(x_0, x_1) \in q|_{H(p_i)}$. Hence $p_i \leq_{\text{triv}} q$ by Lemma 9. □

Corollary 13 *Let $p = \bigwedge_{i=1}^{u} p_i$ and $q = \bigvee_{i=1}^{v} q_i$ as in the previous corollary, and let α_i be a variable. Then*

- *if $\alpha_i \leq_{\text{triv}} q$ then $\alpha_i \leq_{\text{triv}} q_j$ for some $j \in \{1, \ldots, v\}$;*
- *if $p \leq_{\text{triv}} \alpha_i$ then $p_j \leq_{\text{triv}} \alpha_i$ for some $j \in \{1, \ldots, u\}$.*

To demonstrate the usefulness of test lattices, we prove the two parts of this corollary separately even if each of them implies the other by the duality principle.

Proof For the first part, let $p' = \alpha_i$ and $H(p') \in \mathbf{G}(p')$. Since $|V(H(p'))| = |L(p')| = 2$ and $1_{L(p')}$ is join-irreducible, we obtain from $(x_0, x_1) \in 1_{L(p')} = p'|_{H(p')} \leq q|_{H(p')} = q_1|_{H(p')} \vee \cdots \vee q_v|_{H(p')}$ that $(x_0, x_1) \in q_j|_{H(p')}$ for some j. Hence $\alpha_i \leq_{\text{triv}} q_j$ by Lemma 9.

For the second part, take a shortest α_i-path U connecting x_0 and x_1 in $H(p)$. If length$(U) = 1$, then α_i equals a meetand p_j of p, whence $p_j \leq_{\text{triv}} \alpha_i$. If length$(U) \geq 2$, then U goes within some $H(p_j)$, and $p_j \leq_{\text{triv}} \alpha_i$ by Lemma 9. □

For a congruence Θ of a k-pointed lattice $(L;\vec{d})$, we will use the notation $\vec{d}/\Theta = (d_1/\Theta, \ldots, d_k/\Theta)$. Let us call Θ a *p-preserving congruence*, if $(c, p(\vec{d})) \in \Theta$ holds for no $c < p(\vec{d})$. The following lemma implies Corollary 4; we formulate this lemma for a later reference. By homomorphisms we still mean k-pointed lattice homomorphisms, and isomorphisms are particular cases.

Lemma 14 *Let $(L, \vec{d})$ be a p-lattice, and let Θ be a congruence of $(L;\vec{d})$.*

- *$(L/\Theta; \vec{d}/\Theta)$ is a p-lattice iff Θ is p-preserving.*
- *There exists an optimal p-lattice. It is finite, and it is unique up to isomorphism.*
- *$(L;\vec{d})$ is an optimal p-lattice iff 0 is the only p-preserving congruence of $L(p)$.*

Proof Assume that Θ is not p-preserving, and choose an element $c = q(\vec{d})$ such that $c < p(\vec{d})$ and $(c, p(\vec{d})) \in \Theta$. Then $p(\vec{d}/\Theta) \leq q(\vec{d}/\Theta)$, for they are equal, but $p \not\leq_{\mathrm{triv}} q$, so L/Θ is not a p-lattice. Conversely, suppose that Θ is p-preserving and $p(\vec{d}/\Theta) \leq q(\vec{d}/\Theta)$. Then $p(\vec{d}/\Theta) \wedge q(\vec{d}/\Theta) = p(\vec{d}/\Theta)$ gives $(p(\vec{d}) \wedge q(\vec{d}), p(\vec{d})) \in \Theta$. Using that Θ is p-preserving, we get $p(\vec{d}) \wedge q(\vec{d}) = p(\vec{d})$. This means that $p(\vec{d}) \leq q(\vec{d})$ in L, whence $p \leq_{\mathrm{triv}} q$, proving the first part.

Let $F = [d_1, \ldots, d_k]$ be the free lattice generated by $\{d_1, \ldots d_k\}$. Then $(F;\vec{d})$ is a p-lattice, whence its smallest congruence is p-preserving. Since the (nonempty) join of all p-preserving congruences of $(F;\vec{d})$ is clearly p-preserving by Lemma III.1.3 from Grätzer [8], $(F;\vec{d})$ has a largest p-preserving congruence Ψ. By the first part of the Lemma, $(K, \vec{d}) := (F/\Psi; \vec{d}/\Psi)$ is a p-lattice.

Let $(M;\vec{d})$ be another p-lattice. Let φ denote the surjective lattice homomorphism $\varphi\colon F \to M$, $d_1 \mapsto d_1$, $\ldots$, $d_k \mapsto d_k$, that is, the unique k-pointed lattice homomorphism from $(F;\vec{d})$ to $(M;\vec{d})$. Clearly, $\operatorname{Ker}\varphi \subseteq \Psi$, whence $(K;\vec{d}) \cong (F/\Psi;\vec{d}/\Psi)$ is a homomorphic image of $(M,\vec{d}) \cong (F/\operatorname{Ker}\varphi;\vec{d}/\operatorname{Ker}\varphi)$. Hence $(K;\vec{d})$ is an optimal p-lattice. It is finite by Theorem 3. Its uniqueness is an evident consequence of finiteness. This proves the second part.

To prove the third part, let Θ be a p-preserving congruence of an *optimal* p-lattice $(L;\vec{d})$. By the first part, $(L/\Theta;\vec{d}/\Theta)$ is again a p-lattice. So, $(L;\vec{d})$ is a homomorphic image of $(L/\Theta;\vec{d}/\Theta)$, and the finiteness of L implies $\Theta = 0$.

Conversely, assume that 0 is the only p-preserving congruence of a p-lattice $(L;\vec{d})$. Consider the (unique) homomorphism $\varphi\colon (L;\vec{d}) \to (K;\vec{d})$. Since

$$(L;\vec{d})/\operatorname{Ker}\varphi \cong (K;\vec{d})$$

is a test lattice, $\operatorname{Ker}\varphi$ is p-preserving by the first part. Hence $\operatorname{Ker}\varphi = 0$ yields that φ is an isomorphism. This implies that $(L;\vec{d})$ is an optimal p-lattice. □

Let $H(p) \in \mathbf{G}(p)$, and let $U = (x_0 = a_0, a_1, a_2, \ldots, a_n = x_1)$ be a *directed* path in $H(p)$. We say that U is a *uniform path*, if the following condition holds: for any $0 \leq i_1 < i_2 < i_3 < i_4 \leq n$ such that (a_{i_1}, a_{i_2}) and (a_{i_3}, a_{i_4}) are edges of

the same color r, the unique isomorphism $S(a_{i_1}, r, a_{i_2}) \to S(a_{i_3}, r, a_{i_4})$, see (4), sends the segment of U between a_{i_1} and a_{i_2} onto the segment of U between a_{i_3} and a_{i_4}.

Lemma 15 *Let U be a uniform path as above and let $\{r_1, \ldots, r_m\}$ be the set of colors of edges of U. Then* $\mathrm{length}(r_1 \vee \cdots \vee r_m) \leq \mathrm{length}(p)$.

Proof We use induction on $\mathrm{length}(p)$. If p is a variable or $n = \mathrm{length}(U) = 1$, then the statement is evident. If p is join-reducible, then $n > 1$ and the induction step is straightforward. If p is meet-reducible and $n > 1$, then there is an $s \in M(p)$, see (1), such that U includes the vertices of the s-arc glued to $H_0(p)$, and the induction step is straightforward again. □

Proof of Theorem 5 According to Lemma 14, it suffices to show that Θ is not p-preserving for any nontrivial congruence Θ of $L(p)$. Since Θ is nontrivial, $\mu < \nu$ and $(\mu, \nu) \in \Theta$ hold for some $\mu, \nu \in L(p)$. In virtue of Lemma 8(b), there is an edge (a, r, b) with $(a, b) \in \nu \setminus \mu$. Let $\eta = \mu \cap r|_{H(p)}$. Corollary 11 implies that $r|_{H(p)} = r|_{H(p)} \wedge \nu$. Hence

$$\eta < r|_{H(p)}, \ \ (\eta, r|_{H(p)}) \in \Theta \ \text{ and } \ (a, b) \in r|_{H(p)} \setminus \eta. \tag{6}$$

Let us fix an $r \in C(p)$ with *maximal length* such that (6) holds with an appropriate edge (a, r, b) and an $\eta \in L(p)$. According to Corollary 11(b), there are $t_1, \ldots, t_u \in C(p)$ such that

$$\eta = t_1|_{H(p)} \vee \cdots \vee t_u|_{H(p)}. \tag{7}$$

Let j denote the unique subscript from $\mathbf{N}_0 = \{0, 1, 2, \ldots\}$ such that (a, r, b) is an edge of $H_j(p)$ but not of $H_{j-1}(p)$.

We have to consider several cases.

Case 1: $j > 0$. Then there is a meet-reducible $q \in C(p)$, an edge (e, q, f) of $H_{j-1}(p)$, and a meetand s of q such that

$$s = r \vee t_{u+1} \vee \cdots \vee t_{u+v} \in M(q). \tag{8}$$

In particular,

$$s|_{H(p)} = r|_{H(p)} \vee t_{u+1}|_{H(p)} \vee \cdots \vee t_{u+v}|_{H(p)}. \tag{9}$$

Notice that $v \geq 1$, and the situation is similar to that of Figure 5. Let

$$\delta := \eta \vee t_{u+1}|_{H(p)} \vee \cdots \vee t_{u+v}|_{H(p)} = t_1|_{H(p)} \vee \cdots \vee t_{u+v}|_{H(p)}. \tag{10}$$

Then (6), (9) and (10) yield that $(\delta, s|_{H(p)}) \in \Theta$ and $\delta \leq s|_{H(p)}$. Since

$$\delta \not< s|_{H(p)} \text{ by } \mathrm{length}(r) < \mathrm{length}(s), \tag{11}$$

we conclude that

$$\delta = s|_{H(p)}.$$

Since $(e,f) \in q|_{H(p)}$ by Corollary 10(a) and, clearly, $q \leq_{\mathrm{triv}} s$, we obtain that

$$(e,f) \in s|_{H(p)} = t_1|_{H(p)} \vee \cdots \vee t_{u+v}|_{H(p)} = (t_1 \vee \cdots \vee t_{u+v})|_{H(p)}. \tag{12}$$

Since $t_1, \ldots, t_{u+v} \in C(p)$, (12) and Corollary 10(b) imply

$$t_1 \vee \cdots \vee t_{u+v} \leq_{\mathrm{triv}} s. \tag{13}$$

Let $H(q) := S(e,q,f)$. Then $H(q) \in \mathbf{G}(q)$ by (4), and

$$(e,f) \in (t_1 \vee \cdots \vee t_{u+v})|_{H(q)} = t_1|_{H(q)} \vee \cdots \vee t_{u+v}|_{H(q)} \tag{14}$$

follows from (12) and Lemma 8(a). Hence, by Lemma 8(b), there is a $t_1|_{H(q)} \cup \cdots \cup t_{u+v}|_{H(q)}$-path U in $H(q) = S(e,q,f)$ that connects e and f. We can assume that U goes through each vertex of $H(q)$ at most once. Then a trivial induction on length(q) shows that U is a *directed* path. Another trivial induction on length(q), based on (4), yields that U can be chosen to be uniform. Let (x,c,y) be an edge of U. Then $(x,y) \in t_i|_{H(q)}$ for some i. Hence $(x,y) \in t_i|_{H(p)}$ by the (trivial direction of) Lemma 8(a).

This shows that U is a uniform $t_1|_{H(p)} \cup \cdots \cup t_{u+v}|_{H(p)}$-path from e to f; in fact, we assume that U is the shortest uniform path with this property.

Subcase 1.1: U consists of a single edge. Then $(e,f) \in t_i|_{H(p)}$ and Corollary 10(b) yield that $q|_{\leq_{\mathrm{triv}}} t_i$ for some $i \in \{1, \ldots, u+v\}$.

Firstly, assume that $i \leq u$. Then $q|_{H(p)} \leq t_i|_{H(p)} \leq \eta \leq r|_{H(p)}$. Hence Corollary 10(c) implies $q \leq_{\mathrm{triv}} r$. Let q' denote the lattice term that we obtain from q by replacing its meetand s with r. Then $q \leq_{\mathrm{triv}} r \leq_{\mathrm{triv}} s$ implies $q' =_{\mathrm{triv}} q$. Since length$(r) <$ length(s), we see that length$(q') <$ length(q). So, q is not a canonical term. This is a contradiction, for all subterms of the canonical p are canonical.

Secondly, assume that $u < i \leq u+v$. Then $q \leq_{\mathrm{triv}} t_i \leq_{\mathrm{triv}} s$, like above. Hence, using t_i instead of r, we can derive the same contradiction.

Subcase 1.2: U consists of at least two edges. Then there is an $s' \in M(q)$, see (1), such that U goes through all the vertices of the s'-arc that was glued to $H_{j-1}(p)$.

Sub-subcase 1.2.1: s' and s are distinct. Let

$$z_0 = e, z_1, \ldots, z_{n-1}, z_n = f \quad \text{and} \quad (z_0, t'_1, z_1), \ldots, (z_{n-1}, t'_n, z_n)$$

be the vertices and the edges of the s'-arc, respectively. Since U goes through z_{i-1} and z_i,

$$(z_{i-1}, z_i) \in t_1|_{H(p)} \vee \cdots \vee t_{u+v}|_{H(p)} = \delta = s|_{H(p)}$$

holds for $i \in \{1, \ldots, n\}$. By Corollary 10(b), $t'_i \leq_{\mathrm{triv}} s$ for all i, which yields that $s' = t'_1 \vee \cdots \vee t'_n \leq_{\mathrm{triv}} s$. This is a contradiction, for the canonical term q cannot have two trivially comparable meetands.

Sub-subcase 1.2.2: s' and s are the same. Then a section W of U, which is a uniform path again, connects a and b. Let $t'_1, \dots, t'_w$ be the colors of the edges of W. By Corollary 10(b),

$$\forall j \in \{1, \dots, w\} \ \exists i \in \{1, \dots, u+v\} \text{ such that } t'_j \leq_{\text{triv}} t_i. \tag{15}$$

Corollary 10(a), applied to the edges of W, and transitivity imply

$$(a, b) \in t'_1|_{H(p)} \vee \dots \vee t'_w|_{H(p)} = (t'_1 \vee \dots \vee t'_w)|_{H(p)}.$$

Hence Corollary 10(b) implies

$$r \leq_{\text{triv}} t'_1 \vee \dots \vee t'_w. \tag{16}$$

This together with (8) yields that $s \leq_{\text{triv}} t'_1 \vee \dots \vee t'_w \vee t_{u+1} \vee \dots \vee t_{u+v}$. Conversely, $t'_1 \vee \dots \vee t'_w \vee t_{u+1} \vee \dots \vee t_{u+v} \leq^{(15)}_{\text{triv}} t_1 \vee \dots \vee t_{u+v} \leq^{(13)}_{\text{triv}} s$. Hence

$$s =_{\text{triv}} t'_1 \vee \dots \vee t'_w \vee t_{u+1} \vee \dots \vee t_{u+v}. \tag{17}$$

If $i = i(j)$ belonged to $\{1, \dots, u\}$ in (15) for each j, then

$$r|_{H(p)} \leq^{(16)} (t'_1 \vee \dots \vee t'_w)|_{H(p)} \leq (t_1 \vee \dots \vee t_u)|_{H(p)} =^{(7)} \eta$$

would contradict (6). Hence, by (15), there is a j, say $j = 1$, such that $t'_1 \leq_{\text{triv}} t_i$ holds for some $i \in \{u+1, \dots, u+v\}$. Let $g = t'_2 \vee \dots \vee t'_w \vee t_{u+1} \vee \dots \vee t_{u+v}$. We see by (17) that $s =_{\text{triv}} g$. However, (8) together with $\text{length}(t'_1 \vee \dots \vee t'_w) \leq^{(\text{L15})} \text{length}(r)$ yields that $\text{length}(g) < \text{length}(s)$. This is a contradiction, because s, as a subterm of p, is canonical.

Case 2: $j = 0$. Firstly, assume p is join-irreducible. Then $H_0(p)$ consists of a single p-colored edge, r coincides with p, whence Θ is not p-preserving, indeed.

Secondly, assume that p is join-reducible. With the temporary notations $s' = s := p$, $e := x_0$ and $f := x_1$, the argument for Sub-subcase 1.2.2 works almost the same way as previously. The only difference is that, instead of (11), we say that

- either $\delta \not< s|_{H(p)}$ and we derive a contradiction the same way as before,
- or $\delta < s|_{H(p)} = p|_{H(p)}$, whence Θ is not p-preserving, indeed.

(Since (e, f) is not an edge now, (11) in itself would not work.) □

Proof of Theorem 6 We can assume that a p is not a join-free term, because otherwise $|L(p)| = 2$ and there is nothing to prove.

We claim that $p|_{H(p)}$ is a join-irreducible element of $L(p)$. By way of contradiction, suppose that there are terms h_1 and h_2 such that $h_1|_{H(p)} < p|_{H(p)}$, $h_2|_{H(p)} < p|_{H(p)}$ but $h_1|_{H(p)} \vee h_1|_{H(p)} = p|_{H(p)}$. Similarly to (and even easier than) the argument right after (14), we conclude that there is a uniform $h_1|_{H(p)} \cup h_2|_{H(p)}$-path U connecting x_0 and x_1. Since $\text{length}(U) = 1$ would imply $p \leq_{\text{triv}} h_i$ for some $i \in \{1, 2\}$ by Corollary 10(b), we obtain that $\text{length}(U) > 1$.

Hence there is an $s \in M(p)$ such that U goes through the vertices of the s-arc glued to $H_0(p)$. Let $s = \bigvee_{i=1}^{n} t_i$. Since U is also a $p|_{H(p)}$-path, we get $t_i \leq_{\mathrm{triv}}^{(\mathrm{C10})} p$ for $i = 1, \ldots, n$. Hence $s \leq_{\mathrm{triv}} p$. Since s is a meetand of p, we have $p \leq_{\mathrm{triv}} s$. Since p is canonical, p coincides with s, which contradicts the assumption that p is a join-irreducible term.

This proves that $p|_{H(p)}$ is join-irreducible in $L(p)$. It is not the 0 of $L(p)$, since $(x_0, x_1) \in p|_{H(p)}$ by Corollary 10(a). Hence $p|_{H(p)}$ has a unique lower cover $p_*|_{H(p)}$. Since congruence classes are intervals and $L(p)$ is optimal by Theorem 5, it follows by Lemma 14 that each non-zero congruence of $L(p)$ collapses $p|_{H(p)}$ and $p_*|_{H(p)}$. Thus, $L(p)$ is subdirectly irreducible. □

It is trivial to check that, for any ternary term q, if q is shorter than $p^{\diamondsuit}$ of Exercise 2, then the identity $p = q$ fails even in the free modular lattice on three generators. Hence $p^{\diamondsuit}$ is a join-irreducible canonical term. It is also trivial to verify that $|L(p^{\diamondsuit})| > 5$. Notice that even Figure 6, which is a useful illustration for test lattices, was very easy to construct. Hence the following proposition clearly solves Exercise 2. In Proposition 16, K will be a lattice in the usual sense while $(L(p); \vec{d})$, the p-lattice, is a k-pointed lattice.

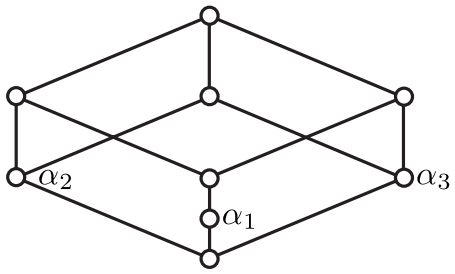


Figure 6: The test lattice $L(p^{\diamondsuit})$

Proposition 16 *Let p be a join-irreducible canonical k-ary lattice term, and let K be a lattice with $|K| < |L(p)|$. Then there exists a k-ary lattice term q such that $p \leq q$ is a nontrivial lattice identity that holds in K.*

Proof Let $n = |K|$. There are n^k ways to make K into a k-pointed lattice $(K; \vec{d})$ by selecting k elements in K. Let $(G; \vec{d})$ be the direct product of all these (n^k many) k-pointed lattices.

Assume that the proposition fails for K. Then $(G; \vec{d})$ is a p-lattice. The (unique) optimal p-lattice is a homomorphic image of $(G; \vec{d})$ by Lemma 14. But, by Theorem 5, the optimal p-lattice is $(L(p); \vec{d})$. Therefore, $L(p)$, as a lattice without constants, belongs to the variety generated by K. Since $L(p)$ is subdirectly irreducible by Theorem 6, the famous $\mathbf{HSP} = \mathbf{P}_s\mathbf{HSP}_u$ theorem of B. Jónsson [11] gives that $L(p)$ is a homomorphic image of a sublattice of K. This contradicts $|K| < |L(p)|$. □

4 Historical remarks

Graphs similar to those here were formerly useful in [1], [2], [4], M. Haiman [9] and P. Lipparini [12]. Even one of the efficient known algorithms for the word problem of lattices is due to graphs, see [3]. (For other algorithms, see also Section XI.8 of Freese, Ježek, Nation [6]). In fact, [3] gives the main motivation to the present work: if graphs are appropriate to solve the word problem, then why not use them for other purposes? However, the mentioned similarity is limited, because our graphs here have more edges than their precursors. In fact, finding the right amount of edges was the main step towards the present approach.

The results of this paper were presented at the conferences organized by the University of Nov Sad and the Technical University of Košice, respectively. It has appeared since then that our approach overlaps Freese, Ježek and Nation [6] more than previously recognized. Since the concepts and the methods of [6] are very different from ours and the counterparts of our results are sometimes only implicitly given in [6], it is reasonable to give a short comparison below.

If we do not assume that p is canonical, then, generally, $L(p)$ does not occur in the book [6]. So, in what follows, let as assume that p' is a *canonical* lattice term.

Using Theorem 3.12 of [6] (in short, Thm. [6].3.12), it is easy to see that $J(p') = J^*(p')$ of [6] is the same as our $C(p')$. Then Cor. [6].3.18 together with Corollary 11(c) gives that $L^\vee(p')$ coincides with our $L(p')$, whence it is our $K(p')$ by Theorem 5. This shows that each optimal test lattice $K(p')$ has been constructed in [6]. This shows also that Theorem 6 is included in Thm. [6].3.24.

The result that $L^\vee(p')$ is a p'-lattice can be easily extracted from [6] in the following way. By the second and third sentences in the proof of Thm. [6].3.15, f in Cor. [6].3.18 is a contraction that acts identically on $L^\vee(p')$, which is a join-subsemilattice of the free lattice $FL(\alpha_1, \ldots, \alpha_k)$. Hence $p' \in FL(\alpha_1, \ldots, \alpha_k)$ is the least preimage of $p' \in L^\vee(p')$. So, $f(p') \leq f(q)$ implies $p' \leq_{\mathrm{triv}} q$, whence $L^\vee(p')$ is a p'-lattice.

It is also possible to extract from [6] that $L^\vee(p')$ is an *optimal* p'-lattice; however, this would require a deeper look into the book, so the details are omitted.

In connection with Theorems 5 and 6, we notice that the name "canonical term" in the present paper means only a *shortest* term, which trivially exists. Opposed to [6], we do not use Whitman's non-trivial theorem on its uniqueness, see [13].

References

[1] Czédli, G.: *On properties of rings that can be characterized by infinite lattice identitites.* Studia Sci. Math. Hungar. **16** (1981), 45–60.

[2] Czédli, G.: *Mal'cev conditions for Horn sentences with congruence permutability.* Acta Math. Hungar. **44** (1984), 115–124.

[3] Czédli, G.: *On the word problem of lattices with the help of graphs.* Periodica Math. Hungar. **23** (1991), 49–58.

[4] Czédli, G., Day, A.: *Horn sentences with (W) and weak Mal'cev conditions.* Algebra Universalis **19** (1984), 217–230.

[5] Day, A.: *Doubling constructions in lattice theory.* Canad. J. Math. **44** (1992), 252–269.

[6] Freese, R., Ježek, J., Nation, J. B.: Free lattices. *American Mathematical Society, Providence, RI,* Mathematical Surveys and Monographs **42**, 1995, viii+293 pp.

[7] Freese, R., Nation, J. B.: *Congruence lattices of semilattices,* Pacific J. Math. **49** (1973), 51–58.

[8] Grätzer, G.: General Lattice Theory. *Birkhäuser Verlag, Basel–Stuttgart,* 1978; *Birkhäuser Verlag,* 1998 (second editon).

[9] Haiman, M.: *Proof theory for linear lattices.* Adv. in Math. **58** (1985), 209–242.

[10] Jónsson, B.: *On the representation of lattices.* Math. Scandinavica **1** (1953), 193–206.

[11] Jónsson, B.: *Algebras whose congruence lattices are distributive.* Math. Scandinavica **21** (1967), 110–121.

[12] Lipparini, P.: *From congruence identities to tolerance identities.* Acta Sci. Math. (Szeged) **73** (2007), 31–51.

[13] Whitman, Ph. M.: *Free lattices.* Ann. of Math. **42** (1941), 325–330.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 53–60

Analytical Representation of Ellipses in the Aitchison Geometry and Its Application

Karel HRON

Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

(Received January 30, 2009)

Abstract

Compositional data, multivariate observations that hold only relative information, need a special treatment while performing statistical analysis, with respect to the simplex as their sample space ([1], [2], [3], [8], [9], [10], [11], [18]). For the logratio approach to the statistical analysis of compositional data the so called Aitchison geometry was introduced and confirmed to be the meaningful one. It was shown in [7], [17] that it is quite easy to express simple geometric objects like compositional lines, this is however not the case for ellipses, although they play a fundamental role within most statistical methods, for example in outlier detection ([8]). The aim of the paper is to introduce a way, based on coordinate representations of compositions, how to obtain an analytical representation of ellipses in the Aitchison geometry.

Key words: Aitchison geometry on the simplex, coordinates, ellipse.

2000 Mathematics Subject Classification: 14P99, 15A03, 15A63, 62H99, 62J05

1 Compositional data

At first, we briefly summarize all the basic properties of compositional data as well as the geometry on the simplex, called in the following Aitchison geometry. More detailed insight is available e.g. in [7]:

Definition 1 A row vector $\boldsymbol{x} = (x_1, \ldots, x_D)$, is called *D-parts composition* when all its components are strictly positive real numbers and they carry only relative information.

The assertion that D-parts compositions (or only compositions in short) carry only relative information means that all the relevant information is contained in the ratios among the parts, i.e. if c is a nonzero real number, $(x_1, \ldots, x_D)$ and $(cx_1, \ldots, cx_D)$ convey essentially the same information. A way to simplify the use of compositions is to represent them in closed form, i.e. as positive vectors with constant sum κ (usually 1 or 100 in case of percentages) of the parts ([7]). As a consequence, D-parts compositions can be identified with the following vector:

Definition 2 For any composition $\boldsymbol{x}$, the *closure operation of* $\boldsymbol{x}$ *to the constant* κ is defined as

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

Proposition 1 *The sample space of compositional data is the simplex, defined as*

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa\}.$$

The basics of the Aitchison geometry on the simplex are mentioned below:

Definition 3 *Perturbation* of a composition $\boldsymbol{x} = \mathcal{C}(x_1, \ldots, x_D) \in \mathcal{S}^D$ by a composition $\boldsymbol{y} = \mathcal{C}(y_1, \ldots, y_D) \in \mathcal{S}^D$ is a composition

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D).$$

Power transformation of a composition $\boldsymbol{x} \in \mathcal{S}^D$ by a constant $\alpha \in \mathbb{R}$ is a composition

$$\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha).$$

The inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ can be expressed as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Proposition 2 *The simplex with the perturbation operation and the power transformation,* $(\mathcal{S}^D, \oplus, \odot)$*, is a linear vector space. Moreover, the Aitchison inner product induces a* $(D-1)$*-dimensional Hilbert space.*

Definition 4 If compositions $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}$ are independent (in terms of the Aitchison geometry), they constitute a *(simplicial) basis* of $\mathcal{S}^D$, i.e. each composition $\boldsymbol{x} \in \mathcal{S}^D$ can be expressed as

$$\boldsymbol{x} = (\alpha_1 \odot \boldsymbol{e}_1) \oplus \cdots \oplus (\alpha_{D-1} \odot \boldsymbol{e}_{D-1})$$

for some coefficients α_i, $i = 1, \ldots, D-1$, that are termed *coordinates* with respect to the basis.

Obviously, using orthonormal bases on the simplex, all operations and metric concepts like perturbation, power transformation, inner product and norm are translated into coordinates as ordinary vector operations (sum of two vectors and multiplication of a vector by a scalar), see [6], [7] and [17] for details. For a composition $\boldsymbol{x}$, we denote $h(\boldsymbol{x})$ its representation in coordinates. Thus, for $\alpha, \beta \in \mathbb{R}$ it holds that

$$h(\alpha \odot \boldsymbol{x} \oplus \beta \odot \boldsymbol{y}) = \alpha \cdot h(\boldsymbol{x}) + \beta \cdot h(\boldsymbol{y});$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \langle h(\boldsymbol{x}), h(\boldsymbol{y}) \rangle_E, \quad \|\boldsymbol{x}\|_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A} = \|h(\boldsymbol{x})\|_E. \tag{1}$$

Example 1 Let us denote the coordinate representation of $\boldsymbol{x}$ as $\boldsymbol{z} = (z_1, \ldots, z_{D-1})$. Coefficients for a chosen simplicial basis ([5]) can be expressed as

$$z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \quad \text{for } i = 1, \ldots, D-1.$$

The inverse transformation, i.e. $h^{-1}(\boldsymbol{z}) = \boldsymbol{x} = \mathcal{C}(x_1, \ldots, x_D)$, is then obtained using

$$x_i = \exp\left(\sum_{j=i}^{D} \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} z_{i-1} \right) \quad \text{with } z_0 = z_D = 0 \text{ for } i = 1, \ldots, D.$$

2 Simplicial ellipses

A $(D-1)$-dimensional real vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{D-1})$ and a positive definite real matrix $\boldsymbol{\Sigma} = (s_{ij})$ determine an ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ with centre $\boldsymbol{\mu}$,

$$\mathcal{E}_{D-1}(\boldsymbol{z})\colon (\boldsymbol{z} - \boldsymbol{\mu})\, \boldsymbol{\Sigma}\, (\boldsymbol{z} - \boldsymbol{\mu})^T = c^2, \quad c > 0. \tag{2}$$

The ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ can be equivalently expressed in analytical form

$$\sum_{i=1}^{D-1} \sum_{j=1}^{D-1} s_{ij} z_i z_j - 2 \sum_{i=1}^{D-1} \sum_{j=1}^{D-1} s_{ij} \mu_i z_j + k = 0$$

with $k = \boldsymbol{\mu} \boldsymbol{\Sigma} \boldsymbol{\mu}^T - c^2$. Using (2) and spectral decomposition of the matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D-1} \lambda_i \boldsymbol{f}_i^T \boldsymbol{f}_i,$$

where λ_i and $\boldsymbol{f}_i$ denote eigenvalues (in descending order) and orthonormal eigenvectors of $\boldsymbol{\Sigma}$, respectively, the ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ can also be expressed in terms of the Euclidean inner product as

$$\sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{f}_i, \boldsymbol{z} \rangle_E)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle \boldsymbol{f}_i, \boldsymbol{\mu} \rangle_E \langle \boldsymbol{f}_i, \boldsymbol{z} \rangle_E + k = 0. \tag{3}$$

It is easy to see that the spectral decomposition of $\mathbf{\Sigma}$ represents only one chosen, nevertheless the most convenient, decomposition of $\mathbf{\Sigma}$ in order to obtain (3). Namely, the vectors $\boldsymbol{f}_i$ determine the ellipse axes' directions, and their lengths are functions of the eigenvalues λ_i.

Let $h(\boldsymbol{x}) = \boldsymbol{z}$, $h(\boldsymbol{\gamma}) = \boldsymbol{\mu}$ and $h(\boldsymbol{e}_i) = \boldsymbol{f}_i$, $i = 1, \ldots, D-1$. Considering (1), the simplicial counterpart to $\mathcal{E}_{D-1}(\boldsymbol{z})$, denoted in the following as $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$, is given by

$$\sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle_A)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle \boldsymbol{e}_i, \boldsymbol{\gamma} \rangle_A \langle \boldsymbol{e}_i, \boldsymbol{x} \rangle_A + k = 0. \tag{4}$$

The following theorem is thus a simple consequence of the above mentioned considerations and definition of the Aitchison inner product:

Theorem 1 *The analytical form of the simplicial ellipse $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$ is uniquely determined as*

$$\sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \sum_{k=1}^{D-1} \sum_{l=k+1}^{D} a_{ijkl} \ln \frac{x_i}{x_j} \ln \frac{x_k}{x_l} + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} b_{ij} \ln \frac{x_i}{x_j} + k = 0,$$

where

$$a_{ijkl} = \frac{1}{D^2} \sum_{m=1}^{D-1} \lambda_m \ln \frac{e_{mi}}{e_{mj}} \ln \frac{e_{mk}}{e_{ml}},\ b_{ij} = -\frac{2}{D} \sum_{m=1}^{D-1} \lambda_m \langle \boldsymbol{e}_i, \boldsymbol{\gamma} \rangle_A \ln \frac{e_{mi}}{e_{mj}}$$

and

$$k = \sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{e}_m, \boldsymbol{\gamma} \rangle_A)^2 - c^2.$$

The compositions $\boldsymbol{\gamma}$ and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iD})$ represent centre of $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$ and the ellipse axes' directions, respectively. Theorem 1 provides a procedure how to construct an analytical representation of an ellipse on the simplex, obtained as a result of statistical computations in coordinates.

Example 2 A simplicial ellipse in coordinates (see Example 1) is given by

$$\boldsymbol{\mu} = (1, 1), \quad \mathbf{\Sigma} = \begin{pmatrix} 2.5, & 1.5 \\ 1.5, & 2.5 \end{pmatrix}, \quad c = 1,$$

i.e. with centre $\boldsymbol{\mu} = (1, 1)$, eigenvalues $\lambda_1 = 4$, $\lambda_2 = 1$ and axis directions

$$\boldsymbol{f}_1 = \frac{1}{\sqrt{2}}(1, 1) \quad \text{and} \quad \boldsymbol{f}_2 = \frac{1}{\sqrt{2}}(1, -1).$$

The analytical form of the ellipse $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$ in coordinates, i.e. $\mathcal{E}_2(\boldsymbol{z})$, is thus

$$2.5z_1^2 + 2.5z_2^2 + 3z_1z_2 - 8z_1 - 8z_2 + 7 = 0.$$

Using (4) and Theorem 1 and after an adjustment we obtain the analytical form of $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$ as

$$0.56\ln^2\frac{x_1}{x_2}+0.84\ln^2\frac{x_1}{x_3}+0.27\ln^2\frac{x_2}{x_3}+1.13\ln\frac{x_1}{x_2}\ln\frac{x_1}{x_3}+0.02\ln\frac{x_1}{x_2}\ln\frac{x_2}{x_3}$$
$$+0.56\ln\frac{x_1}{x_3}\ln\frac{x_2}{x_3}-3.77\ln\frac{x_1}{x_2}-5.15\ln\frac{x_1}{x_3}-1.38\ln\frac{x_2}{x_3}+7=0.$$

Here, the centre $\boldsymbol{\gamma}=(0.72,0.18,0.10)$ and axis directions are $\boldsymbol{e}_1=(0.61,0.23,0.16)$ and $\boldsymbol{e}_2=(0.36,0.13,0.51)$, respectively. Fig. 1 shows the simplicial ellipse displayed in a ternary diagram as well as in coordinates.

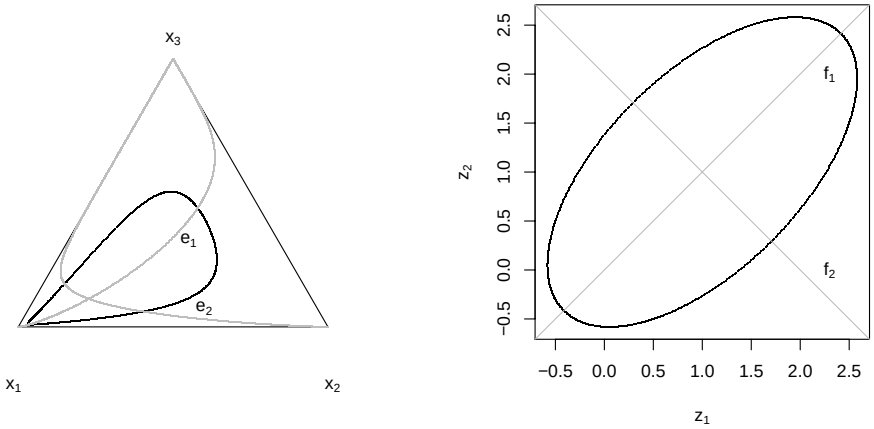


Fig. 1. The simplicial ellipse displayed in a ternary diagram (left) and in coordinates (right) together with the directions of the ellipse axes.

Let us remark that the existence of an analytical expression for ellipses on the simplex opens also a possibility for further generalizations in many directions, e.g. [13], [14].

3 Application in a statistical method

Ellipses frequently occur as a result of many statistical methods. In the case of compositional data one has to be careful to check whether the given problem is solvable in coordinates and how the results can be interpreted back on the simplex. One such problem is to find a regression line (in the compositional sense) that represents the main trend in the data, e.g. using the first principal component or equivalently the total least squares problem, computed in coordinates ([4], [15], [19]). In its simplest form it attempts to fit a line that explains the set of n two-dimensional data points (e.g. three-part compositions in coordinates) in such a way that the sum of squared distances from data points to the estimated line is minimal. In [12] it was shown that in this case, the problem is also solvable iteratively using the theory of linear regression models

with type-II constraints (in the constraints not only parameters of the model but also additional parameters occur, [16]), see [12] for details. Moreover, this approach enables to perform deeper statistical analysis like confidence regions and hypotheses testing. Considering the first mentioned possibility, under the assumption of normality we can construct confidence ellipses for locations of the unknown errorless results of the measurement, i.e. for the locations of each of n points $\boldsymbol{z}_i = h(\boldsymbol{x}_i) = (z_{i1}, z_{i2})$, $i = 1, \ldots, n$. The numerical results in the text below correspond to the statistical analysis of the well known Aphyric Skye Lavas data set that comes from [1, p. 360] and represents percentages of three variables ($Na_2O + K_2O$, Fe_2O_3 and MgO) related to the chemistry of 23 lava samples.

The confidence ellipses for the single errorless results of the measurement (true concentrations of the chemical compounds) in coordinates are constructed in such a way that their centers $\boldsymbol{\mu}_i$ (and $\boldsymbol{\gamma}_i$ in the Aitchison geometry) lie on the regression line $z_2 = \beta_1 + \beta_2 z_1$, where the parameters β_1, β_2 are estimated using the iterative algorithm described in [12]. Thus, we can assert that the unknown errorless results lie in the ellipses with the prescribed probability $1 - \alpha$. The directions $\boldsymbol{f}_1$ of the main half-axes of such ellipses follow the direction given by the estimated line, $\boldsymbol{f}_1 = (0.8903, -0.4554)$, thus $\boldsymbol{f}_2 = (0.4554, 0.8903)$ for the adjacent half-axes.

Although it might not to be visible from the ternary diagram, the unitary directions of the ellipses' main and adjacent half-axes are also the same and for all of them we have $\boldsymbol{e}_1 = (0.4515, 0.1282, 0.4203)$, $\boldsymbol{e}_2 = (0.5654, 0.2969, 0.1377)$; note that, of course, $\langle \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle_A = 0$. Concretely, for a 95%-confidence ellipse, belonging to $\boldsymbol{x}_1$, we obtain the center of this ellipse in coordinates and on the simplex as

$$\boldsymbol{\mu}_1 = (0.0122, 1.7471), \quad \boldsymbol{\gamma}_1 = (0.4763, 0.4682, 0.0556),$$

respectively. Here c^2 equals $2\mathrm{F}_{2,21}(0.95)$, 95%-quantile of the F distribution with 2 and 21 degrees of freedom, see again [12] for details. The analytical representation of the ellipse in coordinates equals to

$$570.53 z_1^2 + 233.99 z_2^2 - 466.44 z_1 z_2 + 801.04 z_1 - 811.93 z_2 + 697.45 = 0$$

(the matrix $\boldsymbol{\Sigma}$ was obtained as inverse of the covariance matrix of the centre $\boldsymbol{\mu}_1$, [12]) and back-transformed to the simplex we obtain for $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$

$$126.79 \ln^2 \frac{x_1}{x_2} + 25.81 \ln^2 \frac{x_1}{x_3} + 115.58 \ln^2 \frac{x_2}{x_3} + 37.02 \ln \frac{x_1}{x_2} \ln \frac{x_1}{x_3} - 216.55 \ln \frac{x_1}{x_2} \ln \frac{x_2}{x_3}$$
$$+ 14.61 \ln \frac{x_1}{x_3} \ln \frac{x_2}{x_3} + 377.61 \ln \frac{x_1}{x_2} - 142.66 \ln \frac{x_1}{x_3} - 520.28 \ln \frac{x_2}{x_3} + 697.45 = 0.$$

Here, the composition $\boldsymbol{x}_1 = (0.52, 0.42, 0.06)$ is not contained in the corresponding confidence ellipse, because $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_1) = 9.85 > 0$. The corresponding results of all compositions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are collected in Table 1. Note that there are many positive values, meaning that the data point is outside the ellipse.

This indicates a poor fit of the model to the data. As a consequence, a more complex model could be selected.

obs. number i	1	2	3	4	5	6
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	9.85	−6.40	−4.20	−6.05	3.70	−3.69
obs. number i	7	8	9	10	11	12
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	1.38	13.92	7.42	6.81	21.94	31.44
obs. number i	13	14	15	16	17	18
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	13.26	−0.33	−5.71	−1.95	67.19	−3.80
obs. number i	19	20	21	22	23	
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	−6.85	−3.61	−6.44	22.36	−5.20	

Tab. 1. Overview of results for the Aphyric Skye Lavas data. The values correspond to the observed compositions $\boldsymbol{x}_i$, $i = 1, \ldots, 23$, substituted in the corresponding confidence ellipses. A value less than zero indicates that the data point is contained inside the ellipse and for values greater than zero outside. Exact zero values would mean that the composition lies on the boundary.

Detailed interpretation of the logratios' coefficients in the analytical representation of ellipses on the simplex is the topic of the author's research and will be presented in the future.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. *Chapman and Hall, London,* 1986.

[2] Aitchison, J., Greenacre, M.: *Biplots of compositional data.* Applied Statistics **51** (2002), 375–392.

[3] Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London*, Special Publications **264**, 2006.

[4] Daunis-i-Estadella, J., Barceló-Vidal, C., Buccianti, A.: *Exploratory compositional data analysis.* In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London,* Special Publications **264** (2006), 161–174.

[5] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C.: *Isometric logratio transformations for compositional data analysis.* Math. Geol. **35** (2003), 279–300.

[6] Egozcue, J. J., Pawlowsky-Glahn, V.: *Groups of parts and their balances in compositional data analysis.* Math. Geol. **37** (2005), 795–828.

[7] Egozcue, J. J., Pawlowsky-Glahn, V.: *Simplicial geometry for compositional data.* In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London,* Special Publications **264** (2006), 145–160.

[8] Filzmoser, P., Hron, K.: *Outlier detection for compositional data using robust methods.* Math. Geosci. **40** (2008), 233–248.

[9] Filzmoser, P., Hron, K.: *Correlation analysis for compositional data.* Math. Geosci. (to appear).

[10] Filzmoser, P., Hron, K., Reimann, C.: *Principal component analysis for compositional data with outliers.* Environmetrics (to appear).

[11] Filzmoser, P., Hron, K., Reimann, C., Garrett, R.: *Robust factor analysis for compositional data.* Computers & Geosciences (to appear).

[12] Fišerová, E., Hron, K.: *Total least squares solution for compositional data using linear models.* Journal of Applied Statistics (to appear).

[13] Jukl, M.: *Linear forms on free modules over certain local ring.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **110** (1993), 49–62.

[14] Jukl, M.: *Inertial law of quadratic forms on modules over plural algebra.* Mathematica Bohemica **3** (1995), 255–263.

[15] Kendall, M. G., Stuart, A.: The advanced theory of statistics, vol 2. *Charles Griffin, London,* 1967.

[16] Kubáček, L., Kubáčková, L.: *One of the calibration problems.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **36** (1997), 117–130.

[17] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, J.: *Lecture notes on compositional data analysis.* http://hdl.handle.net/10256/297, 2007.

[18] Pearson, K.: *Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London **60** (1897), 489–502.

[19] Schuermans, M., Markovsky, I., Wentzell, P. D., Van Huffel, S.: *On the equivalence between total least squares and maximum likelihood PCA.* Analytica Chimica Acta **544** (2005), 254–267.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 61–71

Uncertainty of the design and covariance matrices in linear statistical model*

Lubomír KUBÁČEK [1], Jaroslav MAREK [2]

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
[1] *e-mail: kubacekl@inf.upol.cz*
[2] *e-mail: marek@inf.upol.cz*

(Received January 15, 2009)

Abstract

The aim of the paper is to determine an influence of uncertainties in design and covariance matrices on estimators in linear regression model.

Key words: Linear statistical model, uncertainty, design matrix, covariance matrix.

2000 Mathematics Subject Classification: 62J05

1 Introduction

Uncertainties in entries of design and covariance matrices influence the variance of estimators and cause their bias. A problem occurs mainly in a linearization of nonlinear regression models, where the design matrix is created by derivatives of some functions. The question is how precise must these derivatives be. Uncertainties of covariance matrices must be suppressed under some reasonable bound as well.

The aim of the paper is to give the simple rules which enables us to decide how many ciphers an entry of the mentioned matrices must be consisted of.

*Supported by the Council of Czech Government MSM 6 198 959 214.

2 Symbols used

In the following text a linear regression model (in more detail cf. [2]) is denoted as

$$\mathbf{Y} \sim_n (\mathbf{F}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad \boldsymbol{\beta} \in R^k, \tag{1}$$

where $\mathbf{Y}$ is an n-dimensional random vector with the mean value $E(\mathbf{Y})$ equal to $\mathbf{F}\boldsymbol{\beta}$ and with the covariance matrix $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$. The symbol R^k means the k-dimensional linear vector space. The $n \times k$ matrix $\mathbf{F}$ is given. It is assumed that the rank $r(\mathbf{F})$ of the matrix $\mathbf{F}$ is $r(\mathbf{F}) = k < n$ and the given matrix $\boldsymbol{\Sigma}$ is positive definite. The k-dimensional unknown vector parameter $\boldsymbol{\beta}$ must be estimated on the basis of the realization $\mathbf{y}$ of the random vector $\mathbf{Y}$. Symbol $\mathbf{e}_i^{(n)}$ means n-dimensional vector with the entry 1 at the i-th position; other entries are zero. The matrix of the normal equation $\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$ is denoted as $\mathbf{C}$; its (i,j)-th entry is $\{\mathbf{C}\}_{i,j}$ and the (i,j)-th entry of $\mathbf{C}^{-1}$ is $\{\mathbf{C}\}^{i,j}$. $\mathbf{F}'$ means the transpose of the matrix $\mathbf{F}$. The (i,j)-th entry of the matrix $\boldsymbol{\Sigma}$ is $\sigma_{i,j} = \{\boldsymbol{\Sigma}\}_{i,j}$ and the i-th component of the vector $\mathbf{v}$ is $\{\mathbf{v}\}_i$.

The symbol $\partial \mathbf{l}_h'\mathbf{Y}/\partial\mathbf{F}$ means

$$\frac{\partial \mathbf{l}_h'\mathbf{Y}}{\partial \mathbf{F}} = \begin{pmatrix} \frac{\partial \mathbf{l}_h'\mathbf{Y}}{\partial F_{1,1}}, & \cdots & \frac{\partial \mathbf{l}_h'\mathbf{Y}}{\partial F_{1,k}} \\ \cdots\cdots\cdots\cdots & & \\ \frac{\partial \mathbf{l}_h'\mathbf{Y}}{\partial F_{n,1}}, & \cdots & \frac{\partial \mathbf{l}_h'\mathbf{Y}}{\partial F_{n,k}} \end{pmatrix}, \tag{2}$$

where $F_{i,j} = \{\mathbf{F}\}_{i,j}, i = 1,\ldots,n, j = 1,\ldots,k$, and $\mathbf{l}_h' = \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}$ for an arbitrary $\mathbf{h} \in \mathbb{R}^k$, $\mathbf{h} \neq \mathbf{0}$.

The Kronecker multiplication of matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted as $\mathbf{A} \otimes \mathbf{B}$ (in more detail cf. [3]). If $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$, then $\mathrm{vec}(\mathbf{A}) = (\mathbf{a}_1', \ldots, \mathbf{a}_m')'$. The identity matrix is denoted as $\mathbf{I}$.

3 Uncertainty in the design matrix

In the following text a sensitivity approach is used, i.e. the influence of uncertainty in the design matrix is judged according to the linear term of the Taylor series (cf. also in [1], chpt. VI). The Taylor series of the quantity $\mathbf{l}_h'\mathbf{Y} = \mathbf{h}'\widehat{\boldsymbol{\beta}}$ will be considered.

Lemma 3.1 *Let $\mathbf{h}' \in R^k$ be an arbitrary vector. It is valid that*

$$\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}} = -\mathbf{l}_h'\widehat{\boldsymbol{\beta}}' + \boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}, \quad \mathbf{l}_h = \boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}\mathbf{h}, \tag{3}$$

$$\widehat{\boldsymbol{\beta}} = \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{4}$$

$$\mathbf{v} = \mathbf{Y} - \mathbf{F}\widehat{\boldsymbol{\beta}}. \tag{5}$$

Proof The BLUE (best linear unbiased eastimator) of the linear function $h(\boldsymbol{\beta}) = \mathbf{h}'\boldsymbol{\beta}$, $\boldsymbol{\beta} \in R^k$, is $\mathbf{h}'\widehat{\boldsymbol{\beta}} = \mathbf{l}_h'\mathbf{Y} = \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$. Thus

$$\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial F_{i,j}} = \mathbf{h}'\frac{\partial \mathbf{C}^{-1}}{\partial F_{i,j}}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} + \mathbf{h}'\mathbf{C}^{-1}\frac{\partial \mathbf{F}'}{\partial F_{i,j}}\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

and

$$\begin{aligned}\frac{\partial \mathbf{C}^{-1}}{\partial F_{i,j}} &= \frac{\partial (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}}{\partial F_{i,j}} = -\mathbf{C}^{-1}\left(\frac{\partial \mathbf{F}'}{\partial F_{i,j}}\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{F}'\boldsymbol{\Sigma}^{-1}\frac{\partial \mathbf{F}}{\partial F_{i,j}}\right)\mathbf{C}^{-1} \\ &= -\mathbf{C}^{-1}\left\{\left[\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\right]'\boldsymbol{\Sigma}^{-1}\mathbf{F} + \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\right\}\mathbf{C}^{-1} \\ &= -\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}.\end{aligned}$$

It implies

$$\begin{aligned}\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial F_{i,j}} &= -\mathbf{l}_h'\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \mathbf{h}'\mathbf{C}^{-1}\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{Y} \\ &\quad + \mathbf{h}'\mathbf{C}^{-1}\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \\ &= -\mathbf{l}_h'\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\widehat{\boldsymbol{\beta}} + \mathbf{h}'\mathbf{C}^{-1}\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{v} \\ &= \left\{-\mathbf{l}_h\widehat{\boldsymbol{\beta}}' + \boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}\right\}_{i,j}, \quad i = 1,\dots,n,\ j = 1,\dots,k. \qquad \square\end{aligned}$$

Lemma 3.2 *Let in the model from Lemma 3.1 the symbol $\delta\mathbf{F}$ denote the matrix of uncertainties in the design matrix $\mathbf{F}$. Then*

$$(i) \quad E\left[\operatorname{Tr}\left(\delta\mathbf{F}'\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}}\right)\right] = -\operatorname{Tr}(\delta\mathbf{F}'\mathbf{l}_h\boldsymbol{\beta}'), \tag{6}$$

$$(ii) \quad \operatorname{Var}\left[\operatorname{Tr}\left(\delta\mathbf{F}'\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}}\right)\right] = \mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}\delta\mathbf{F}'\mathbf{l}_h + \mathbf{h}'\mathbf{C}^{-1}\delta\mathbf{F}'(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+ \times \delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}, \tag{7}$$

where

$$(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+ = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}$$

is the Moore–Penrose generalized inverse of the matrix $\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F$ (in more detail cf. [3]).

Proof The statement (i) is obvious. As far as (ii) is concerned, it is valid that

$$\begin{aligned}\operatorname{Var}\left[\operatorname{Tr}\left(\delta\mathbf{F}'\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}}\right)\right] &= \operatorname{Var}\left\{[\operatorname{vec}(\delta\mathbf{F})]'\operatorname{vec}\left(\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}}\right)\right\} \\ &= \operatorname{Var}\left([\operatorname{vec}(\delta\mathbf{F})]'\left\{-(\mathbf{I}\otimes\mathbf{l}_h)\widehat{\boldsymbol{\beta}} + \left[(\mathbf{C}^{-1}\mathbf{h})\otimes\boldsymbol{\Sigma}^{-1}\right]\mathbf{v}\right\}\right).\end{aligned}$$

Since $\widehat{\boldsymbol{\beta}}$ and $\mathbf{v}$ are noncorrelated, $\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \mathbf{C}^{-1}$ and $\mathrm{Var}(\mathbf{v}) = \boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}'$, we have

$$\mathrm{Var}\left[\mathrm{Tr}\left(\delta\mathbf{F}'\frac{\partial \mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial \mathbf{F}}\right)\right] = [\mathrm{vec}(\delta\mathbf{F})]'(\mathbf{I}\otimes \mathbf{l}_h)\mathbf{C}^{-1}(\mathbf{I}\otimes \mathbf{l}_h')\,\mathrm{vec}(\delta\mathbf{F})$$

$$+ [\mathrm{vec}(\delta\mathbf{F})]'[(\mathbf{C}^{-1}\mathbf{h})\otimes\boldsymbol{\Sigma}^{-1}](\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}')[(\mathbf{h}'\mathbf{C}^{-1})\otimes\boldsymbol{\Sigma}^{-1}]\,\mathrm{vec}(\delta\mathbf{F})$$

$$= [\mathrm{vec}(\delta\mathbf{F})]'[\mathbf{C}^{-1}\otimes(\mathbf{l}_h\mathbf{l}_h')]\,\mathrm{vec}(\delta\mathbf{F}) + [\mathrm{vec}(\delta\mathbf{F})]'[(\mathbf{C}^{-1}\mathbf{h}\mathbf{h}'\mathbf{C}^{-1})\otimes(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+]$$

$$\times\,\mathrm{vec}(\delta\mathbf{F}) = \mathrm{Tr}[(\delta\mathbf{F})'\mathbf{l}_h\mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}] + \mathrm{Tr}[(\delta\mathbf{F})'(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}\mathbf{h}'\mathbf{C}^{-1}]$$

$$= \mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}(\delta\mathbf{F})'\mathbf{l}_h + \mathbf{h}'\mathbf{C}^{-1}(\delta\mathbf{F})'(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}. \qquad \square$$

Remark 3.1 Regarding Lemma 3.1 the influence of $\delta\mathbf{F}$ on the estimate of the function $\mathbf{h}'\boldsymbol{\beta}$, $\boldsymbol{\beta}\in R^k$, can be evaluated. If $\delta\mathbf{F}\neq\mathbf{0}$, then instead of $\mathbf{h}'\widehat{\boldsymbol{\beta}} = \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$ ($\mathbf{y}$ is a realization of $\mathbf{Y}$) we obtain

$$\mathbf{h}'\tilde{\boldsymbol{\beta}} \approx \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \mathrm{Tr}[(\delta\mathbf{F})'\mathbf{l}_h\widehat{\boldsymbol{\beta}}'] + \mathrm{Tr}[(\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}] \tag{8}$$

(for practical purposes the values $\tilde{\boldsymbol{\beta}}$ and $\mathbf{y}-\mathbf{F}\tilde{\boldsymbol{\beta}}$ can be used on the right hand side of the last approximate equality instead of $\widehat{\boldsymbol{\beta}}$ and $\mathbf{v}$).

In an actual case we can judge whether uncertainty $\delta\mathbf{F}$ in the used matrix $\mathbf{F}$ satisfy the inequality

$$|-\mathrm{Tr}[(\delta\mathbf{F})'\mathbf{l}_h\widehat{\boldsymbol{\beta}}'] + \mathrm{Tr}[(\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}]| < \varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}},$$

where $\varepsilon > 0$ is sufficiently small (according to an opinion of a statistician) number.

If $\delta\mathbf{F} = \mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\Delta$, then

$$-\mathrm{Tr}[(\delta\mathbf{F})'\mathbf{l}_h\widehat{\boldsymbol{\beta}}'] + \mathrm{Tr}[(\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}] =$$

$$= -\mathrm{Tr}\left[\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\mathbf{l}_h\widehat{\boldsymbol{\beta}}'\right] + \mathrm{Tr}\left[\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}\right]$$

$$= -\{\mathbf{l}_h\}_i\{\widehat{\boldsymbol{\beta}}\}_j + \{\boldsymbol{\Sigma}^{-1}\mathbf{v}\}_i\left\{\mathbf{C}^{-1}\mathbf{h}\right\}_j.$$

Remark 3.2 According to Lemma 3.2 the influnce of $\delta\mathbf{F}$ on the estimator of the function $\mathbf{h}'\boldsymbol{\beta}$, $\boldsymbol{\beta}\in R^k$, can be evaluated. As far as the bias of the estimator $\mathbf{h}'\widehat{\boldsymbol{\beta}}$ is concerned, if

$$\tilde{\boldsymbol{\beta}} = \left[(\mathbf{F}+\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}(\mathbf{F}+\delta\mathbf{F})\right]^{-1}(\mathbf{F}+\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

then

$$E(\mathbf{h}'\tilde{\boldsymbol{\beta}}) \approx \mathbf{h}'\boldsymbol{\beta} - \mathrm{Tr}\left[(\delta\mathbf{F})'\mathbf{l}_h\boldsymbol{\beta}'\right],$$

i.e. the bias of the estimator is $-\mathrm{Tr}\left[(\delta\mathbf{F})'\mathbf{l}_h\boldsymbol{\beta}'\right]$. It must be suppressed under some reasonable bound, i.e. it must be

$$|\,\mathrm{Tr}\left[(\delta\mathbf{F})'\mathbf{l}_h\boldsymbol{\beta}'\right]| < \varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}.$$

(Instead of $\boldsymbol{\beta}$ the estimator of it can be used what could be sufficient for practical purposes.)

For the sake of simplicity let $\delta\mathbf{F} = \mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\Delta$. Then

$$\mathrm{Tr}\left[(\delta\mathbf{F})'\mathbf{l}_h\boldsymbol{\beta}'\right] = \Delta\,\mathrm{Tr}\left[\mathbf{e}_j^{(k)}(\mathbf{e}_i^{(n)})'\mathbf{l}_h\boldsymbol{\beta}'\right] = \Delta\{\mathbf{l}_h\}_i\{\boldsymbol{\beta}\}_j;$$

thus it should be valid

$$\Delta \ll \varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\frac{1}{\{\mathbf{l}_h\}_i\{\boldsymbol{\beta}\}_j}.$$

The value

$$\triangle_{\boldsymbol{crit,i,j}}^{(\boldsymbol{F})} = \varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\frac{1}{\{\mathbf{l}_h\}_i\{\boldsymbol{\beta}\}_j} \tag{9}$$

is the maximum admissible contamination of the (i,j)-th entry of the design matrix $\mathbf{F}$. It causes a bias of the estimator $\mathbf{h}'\tilde{\boldsymbol{\beta}}$ not larger than $\varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}$.

As far as the variance of the estimator $\mathbf{h}'\widehat{\boldsymbol{\beta}}$ is concerned, we have

$$\begin{aligned}\mathbf{h}'\tilde{\boldsymbol{\beta}} &= \mathbf{h}'\widehat{\boldsymbol{\beta}} + \Big\{\mathrm{Tr}\left[-(\delta\mathbf{F})'\mathbf{l}_h\widehat{\boldsymbol{\beta}}'\right] + \mathrm{Tr}\left[(\delta\mathbf{F})'\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}\right]\Big\}\\ &= (\mathbf{h}' - \mathbf{l}_h'\delta\mathbf{F})\widehat{\boldsymbol{\beta}} + \mathbf{h}'\mathbf{C}^{-1}\delta\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{v}\end{aligned}$$

and thus

$$\begin{aligned}\mathrm{Var}(\mathbf{h}'\tilde{\boldsymbol{\beta}}) &= (\mathbf{h}' - \mathbf{l}_h'\delta\mathbf{F})\mathbf{C}^{-1}\left[\mathbf{h} - (\delta\mathbf{F})'\mathbf{l}_h\right] + \mathbf{h}'\mathbf{C}^{-1}(\delta\mathbf{F})'(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^{+}\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}\\ &= \mathrm{Var}(\mathbf{h}'\widehat{\boldsymbol{\beta}}) - 2\mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h} + \mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}(\delta\mathbf{F})'\mathbf{l}_h + \mathbf{h}'\mathbf{C}^{-1}(\delta\mathbf{F})'\\ &\quad\times(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^{+}\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}.\end{aligned}$$

The variance of the estimator with an uncertain design matrix differs from the variance of the estimator with the proper design matrix. The difference is

$$-2\mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h} + \mathbf{l}_h'\delta\mathbf{F}\mathbf{C}^{-1}(\delta\mathbf{F})'\mathbf{l}_h + \mathbf{h}'\mathbf{C}^{-1}(\delta\mathbf{F})'(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^{+}\delta\mathbf{F}\mathbf{C}^{-1}\mathbf{h}.$$

For the sake of simplicity let $\delta\mathbf{F} = \mathbf{e}_i^{(n)}(\mathbf{e}_j^{(k)})'\Delta$. Then the difference is

$$\begin{aligned}&\gamma_{h,(i,j)} =\\ &= -2\Delta\{\mathbf{l}_h\}_i\{\mathbf{C}^{-1}\mathbf{h}\}_j + \Delta^2\Big[\{\mathbf{C}\}^{j,j}(\{\mathbf{l}_h\}_i)^2 + (\{\mathbf{C}^{-1}\mathbf{h}\}_j)^2\{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^{+}\}_{i,i}\Big].\end{aligned}$$

It can be assumed that $\gamma_{h,(i,j)} \ll \mathbf{h}'\mathbf{C}^{-1}\mathbf{h}$ and thus

$$\begin{aligned}\sqrt{\mathrm{Var}(\mathbf{h}'\tilde{\boldsymbol{\beta}})} &= \sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h} + \gamma_{h,(i,j)}} = \sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\left(1 + \frac{\gamma_{h,(i,j)}}{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\right)^{1/2}\\ &\approx \sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\left(1 + \frac{1}{2}\frac{\gamma_{h,(i,j)}}{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\right).\end{aligned}$$

The solution $\triangle_{\boldsymbol{crit,i,j}}^{(\boldsymbol{V})}$ of the quadratic equation

$$\begin{aligned}&\Delta^2\Big[\{\mathbf{C}\}^{j,j}(\{\mathbf{l}_h\}_i)^2 + (\{\mathbf{C}^{-1}\mathbf{h}\}_j)^2\{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^{+}\}_{i,i}\Big]\\ &\quad - 2\Delta\{\mathbf{l}_h\}_i\{\mathbf{C}^{-1}\mathbf{h}\}_j - 2\varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}} = 0\end{aligned} \tag{10}$$

is the maximum admissible contamination of the (i,j)-th entry of the design matrix $\mathbf{F}$. It causes an enlargement of the standard deviation $\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}$ not larger than $\varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}$. The value of the quantity $\gamma_{h,(i,j)}$ is the same for both roots of the quadratic equation.

It is useful to arrange tables of the values $\triangle^{(\boldsymbol{F})}_{\boldsymbol{crit,i,j}}$ (cf. (9)) and $\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit,i,j}}$ (cf. (10)) for all $i = 1,\dots,n$ and $j = 1,\dots,k$, cf. section 5 Numerical examples.

Remark 3.3 The most dangerous shift $\delta\mathbf{F}$ of the matrix $\mathbf{F}$ with respect to the bias of the estimator is in the direction of the gradient, i.e.

$$\delta\mathbf{F}^* = kE\left(\frac{\partial\mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial\mathbf{F}}\right) = -k\mathbf{l}_h\boldsymbol{\beta}'.$$

(The number k will be determined later.) The bias of the estimator caused by $\delta\mathbf{F}^*$ is

$$-\operatorname{Tr}\left[(\delta\mathbf{F}^*)'\mathbf{l}_h\boldsymbol{\beta}'\right] = k\boldsymbol{\beta}'\boldsymbol{\beta}\mathbf{l}_h'\mathbf{l}_h.$$

The number k now can be bounded according to the condition

$$k\boldsymbol{\beta}'\boldsymbol{\beta}\mathbf{l}_h'\mathbf{l}_h < \varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}.$$

The matrix

$$\delta\mathbf{F}^* = \frac{\varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}}{\boldsymbol{\beta}'\boldsymbol{\beta}\mathbf{l}_h'\mathbf{l}_h}\mathbf{l}_h\boldsymbol{\beta}' \tag{11}$$

can serve as a good information on the necessary accuracy of the matrix $\mathbf{F}$ in connection with the bias of the estimator $\mathbf{h}'\widehat{\boldsymbol{\beta}}$.

It is to be remarked that in the case $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, the number k must satisfy the inequality $k < \sigma\varepsilon/\left(\boldsymbol{\beta}'\boldsymbol{\beta}\sqrt{\mathbf{h}'(\mathbf{F}'\mathbf{F})^{-1}\mathbf{h}}\right)$.

4 Uncertainty in the covariance matrix

Lemma 4.1 *In the regular linear model* $\mathbf{Y} \sim_n (\mathbf{F}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, $\boldsymbol{\beta} \in R^k$, *for a given linear function* $\mathbf{h}'\boldsymbol{\beta}$, $\boldsymbol{\beta} \in R^k$, *it is valid that*

$$\frac{\partial\mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial\sigma_{i,j}} = -\{\mathbf{l}_h\}_i\{\boldsymbol{\Sigma}^{-1}\mathbf{v}\}_j - \{\mathbf{l}_h\}_j\{\boldsymbol{\Sigma}^{-1}\mathbf{v}\}_i,\ i,j = 1,\dots,n.$$

Proof Since $\mathbf{h}'\widehat{\boldsymbol{\beta}} = \mathbf{h}'(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, it is valid that

$$\begin{aligned}
\frac{\partial\mathbf{h}'\widehat{\boldsymbol{\beta}}}{\partial\sigma_{i,j}} &= \mathbf{h}'\frac{\partial(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}}{\partial\sigma_{i,j}}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} + \mathbf{h}'(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}'\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\sigma_{i,j}}\mathbf{Y} \\
&= \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\left[\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(n)})' + \mathbf{e}_j^{(n)}(\mathbf{e}_i^{(n)})'\right]\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \\
&\quad - \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\left[\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(n)})' + \mathbf{e}_j^{(n)}(\mathbf{e}_i^{(n)})'\right]\boldsymbol{\Sigma}^{-1}\mathbf{Y} \\
&= \left\{-\left[\mathbf{l}_h(\boldsymbol{\Sigma}^{-1}\mathbf{v})' + \boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{l}_h'\right]\right\}_{i,j}, \quad i,j = 1,\dots,n.
\end{aligned}$$

□

Remark 4.1 Since uncertainty in the covariance matrix does not cause the bias of the estimator, only a change of the variance of the estimator must be taken into account. Since it is valid that

$$\mathbf{h}'\big[\mathbf{F}'(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^{-1}\mathbf{F}\big]^{-1}\mathbf{F}'(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^{-1}\mathbf{Y} \approx \mathbf{h}'\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$
$$-\operatorname{Tr}\left[\delta\boldsymbol{\Sigma}(\mathbf{l}_h\mathbf{v}'\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{v}\mathbf{l}_h')\right] = \mathbf{h}'\widehat{\boldsymbol{\beta}} - 2\mathbf{l}'\delta\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{v},$$

we have

$$\operatorname{Var}_\Sigma\left\{\mathbf{h}'\big[\mathbf{F}'(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^{-1}\mathbf{F}\big]^{-1}\mathbf{F}'(\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma})^{-1}\mathbf{Y}\right\}$$
$$\approx \mathbf{h}'\mathbf{C}^{-1}\mathbf{h} + 4\mathbf{l}_h'\delta\boldsymbol{\Sigma}(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\delta\boldsymbol{\Sigma}\mathbf{l}_h.$$

If

$$\delta\boldsymbol{\Sigma} = \begin{cases} \big[\mathbf{e}_i^{(n)}(\mathbf{e}_j^{(n)})' + \mathbf{e}_j^{(n)}(\mathbf{e}_i^{(n)})'\big]\Delta, & i \neq j \\ \big[\mathbf{e}_i^{(n)}(\mathbf{e}_i^{(n)})'\big]\Delta, & i = j \end{cases}$$

then if $i \neq j$

$$d_{h,(i,j)} = 4\mathbf{l}_h'\delta\boldsymbol{\Sigma}(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\delta\boldsymbol{\Sigma}\mathbf{l}_h$$
$$= 4\big(\{\mathbf{l}_h\}_j, \{\mathbf{l}_h\}_i\big)\begin{pmatrix} \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,i}, & \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,j} \\ \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{j,i}, & \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{j,j} \end{pmatrix}\begin{pmatrix} \{\mathbf{l}_h\}_j \\ \{\mathbf{l}_h\}_i \end{pmatrix}\Delta^2,$$

if $i = j$

$$d_{h,(i,i)} = 4(\{\mathbf{l}_h\}_i)^2\{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,i}\Delta^2.$$

Since we can assume that $d_{h,(i,j)} \ll \mathbf{h}'\mathbf{C}^{-1}\mathbf{h}$, we can write

$$\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h} + d_{h,(i,j)}} \approx \sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\left(1 + \frac{1}{2}\frac{d_{h,(i,j)}}{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\right).$$

The matrix $\mathbf{D}_h$ with the (i,j)-th entry

$$\{\mathbf{D}_h\}_{i,j} = \left(1 + \frac{1}{2}\frac{d_{h,(i,j)}}{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}\right), \quad i,j = 1,\ldots,n,$$

can help to analyze the influence of $\delta\boldsymbol{\Sigma}$ on the standard deviation of the estimator $\mathbf{h}'\widehat{\boldsymbol{\beta}}$. The value $\{\mathbf{D}_h\}_{i,j}$ means the ratio of the standard deviation of the estimator calculated with the covariance matrix $\boldsymbol{\Sigma}+\delta\boldsymbol{\Sigma}$ to the standard deviation of the estimator calculated with proper covariance matrix $\boldsymbol{\Sigma}$.

The solution $\triangle^{(\boldsymbol{\Sigma})}_{\boldsymbol{crit,i,j}}$ of the equation (for $i \neq j$)

$$2\Delta^2\big(\{\mathbf{l}_h\}_j, \{\mathbf{l}_h\}_i\big)\begin{pmatrix} \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,i}, & \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,j} \\ \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{j,i}, & \{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{j,j} \end{pmatrix}\begin{pmatrix} \{\mathbf{l}_h\}_j \\ \{\mathbf{l}_h\}_i \end{pmatrix} = \varepsilon\mathbf{h}'\mathbf{C}^{-1}\mathbf{h} \quad (12)$$

and the equation (for $i = j$)

$$2\Delta^2(\{\mathbf{l}_h\}_i)^2\{(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)^+\}_{i,i} = \varepsilon\mathbf{h}'\mathbf{C}^{-1}\mathbf{h} \quad (13)$$

is the maximum admissible contamination of the (i,j)-th entry of the variance matrix Σ. It causes an enlargement of the standard deviation $\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}$ not greater than $\varepsilon\sqrt{\mathbf{h}'\mathbf{C}^{-1}\mathbf{h}}$.

5 Numerical examples

Example 5.1 Let the regression model be

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} \sim_n \left[\begin{pmatrix} 1, & 1 \\ 1, & 2 \\ 1, & 3 \\ 1, & 4 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 \mathbf{I} \right], \ \sigma = 0.1$$

and $\mathbf{y} = (1.6, 1.9, 2.6, 3.1)'$.

Then

$$(\mathbf{F}'\mathbf{F})^{-1} = \begin{pmatrix} 1.5, & -0.5 \\ -0.5, & 0.2 \end{pmatrix}, \quad \sigma^2(\mathbf{F}'\mathbf{F})^{-1} = \begin{pmatrix} 0.0150, & -0.0050 \\ -0.0050, & 0.0020 \end{pmatrix},$$

$$(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}' = \begin{pmatrix} 1.0, & 0.5, & 0.0, & -0.5 \\ -0.3, & -0.1, & 0.1, & 0.3 \end{pmatrix},$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y} = \begin{pmatrix} 1.00 \\ 0.52 \end{pmatrix}, \quad \mathbf{v} = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}} = (0.08, -0.14, 0.04, 0.02)'.$$

Let $\mathbf{h}_1 = (1, 0)'$ in situation A, $\mathbf{h}_2 = (0, 1)'$ in situation B and $\varepsilon = 0.2$.

Then in situation A according Remark 3.4 formulas (9) and (10) we will determine:

$$\triangle^{(\boldsymbol{F})}_{\boldsymbol{crit}} = \begin{pmatrix} 0.0245 & 0.0471 \\ 0.0490 & 0.0942 \\ \infty & \infty \\ -0.0490 & -0.0942 \end{pmatrix}, \quad \delta\mathbf{F}^* = \begin{pmatrix} 0.0129 & 0.0067 \\ 0.0064 & 0.0033 \\ 0 & 0 \\ -0.0064 & -0.0033 \end{pmatrix},$$

from (10) two solution ${}_1\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}}$ and ${}_2\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}}$ are obtained

$${}_1\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}} = \begin{pmatrix} -0.9620 & -6.4139 \\ -1.2464 & -5.9078 \\ -1.7637 & -5.2910 \\ -2.9893 & -4.5720 \end{pmatrix}, \quad {}_2\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}} = \begin{pmatrix} 2.3413 & 2.7775 \\ 2.0156 & 3.6855 \\ 1.7637 & 5.2910 \\ 1.5608 & 8.5720 \end{pmatrix}.$$

These two matrices cause an enlargement of standard deviation not more ε-times. As a criterion the value

$$\min\left\{|{}_1\triangle^{(\boldsymbol{V})}{}_{crit,i,j}|, |{}_2\triangle^{(\boldsymbol{V})}{}_{crit,i,j}|\right\}$$

must be choosen in practice.

$$\triangle^{(\boldsymbol{\Sigma})}_{\boldsymbol{crit}} = \begin{pmatrix} 0.0071 & 0.0063 & 0.0046 & 0.0093 \\ 0.0063 & 0.0093 & 0.0093 & 0.0071 \\ 0.0046 & 0.0093 & \infty & 0.0093 \\ 0.0093 & 0.0071 & 0.0093 & 0.0141 \end{pmatrix},$$

For example the value $\triangle^{(\boldsymbol{F})}_{crit,(3,1)}$ and $\triangle^{(\boldsymbol{F})}_{crit,(3,2)}$ for $\mathbf{h} = (1, 0)'$ cannot be determined, since $\{\mathbf{l}_h\}_3\{\boldsymbol{\beta}\}_1$ and $\{\mathbf{l}_h\}_3\{\boldsymbol{\beta}\}_2$, respectively are zero. Ever it

seems that the contamination of the design matrix $\mathbf{F}$ in the third row can be any larger number, it is not so. An aproach to determination of the value $\triangle^{(\boldsymbol{F})}_{\boldsymbol{crit,i,j}}$ is infinitesimal and therefore some carefulness it necessary. If, e.g. $\triangle^{(\boldsymbol{F})}_{crit,(3,1)} = 0.1$, then the bias of the estimator $\widehat{(1,0)\boldsymbol{\beta}}$ is $(0.0096, 0.0064)'$, what is admissible. However the value $\triangle^{(\boldsymbol{F})}_{crit,(3,1)} = 1$ leads to a non-admissible bias.

In situation B according Remark 3.4 formulas (9), (10) and from the Remark 3.5 formula (11) we will determine:

$$\triangle^{(\boldsymbol{F})}_{\boldsymbol{crit}} = \begin{pmatrix} -0.0298 & -0.0573 \\ -0.0894 & -0.1720 \\ 0.0894 & 0.1720 \\ 0.0298 & 0.0573 \end{pmatrix}, \quad \delta\mathbf{F}^* = \begin{pmatrix} -0.0106 & -0.0055 \\ -0.0035 & -0.0018 \\ 0.0035 & 0.0018 \\ 0.0106 & 0.0055 \end{pmatrix},$$

$${}_1\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}} = \begin{pmatrix} -2.2905 & -9.9767 \\ -2.8165 & -8.4173 \\ -3.3428 & -7.0840 \\ -3.7190 & -5.9767 \end{pmatrix}, \quad {}_2\triangle^{(\boldsymbol{V})}_{\boldsymbol{crit}} = \begin{pmatrix} 3.7190 & 5.9767 \\ 3.3428 & 7.0840 \\ 2.8165 & 8.4173 \\ 2.2905 & 9.9767 \end{pmatrix},$$

and from the Remark 4.2 formulas (12) and (13) we will determine

$$\triangle^{(\boldsymbol{\Sigma})}_{\boldsymbol{crit}} = \begin{pmatrix} 0.0086 & 0.0069 & 0.0053 & 0.0105 \\ 0.0069 & 0.0169 & 0.0105 & 0.0053 \\ 0.0053 & 0.0105 & 0.0169 & 0.0069 \\ 0.0105 & 0.0053 & 0.0069 & 0.0086 \end{pmatrix}.$$

Let for $\delta\mathbf{F} = \delta\mathbf{F}^*$ the value of the estimator (8) from Remark 3.3 be compared with $\mathbf{h}'\hat{\boldsymbol{\beta}} = \mathbf{h}' \begin{pmatrix} 1.00 \\ 0.52 \end{pmatrix}$; $\mathbf{h}'\tilde{\boldsymbol{\beta}} = \mathbf{h}'\hat{\boldsymbol{\beta}} - \mathrm{Tr}[(\delta\mathbf{F}^*)'\mathbf{l}_h\hat{\boldsymbol{\beta}}'] + \mathrm{Tr}[(\delta\mathbf{F}^*)'\mathbf{v}\mathbf{h}'\mathbf{C}^{-1}]$.

If $\mathbf{h} = (1,0)'$, then $\mathbf{h}'\tilde{\boldsymbol{\beta}} - \mathbf{h}'\hat{\boldsymbol{\beta}} = 0.9755 - 1.0000 = -0.0245$.

If $\mathbf{h} = (0,1)'$, then $\mathbf{h}'\tilde{\boldsymbol{\beta}} - \mathbf{h}'\hat{\boldsymbol{\beta}} = 0.5111 - 0.5200 = -0.0089$.

Example 5.2 Let the regression model be

$$y_i = \frac{\beta_1 x_i}{\beta_2 + x_i}, \ i = 1,2,3,4,5 \tag{14}$$

and results of measurement of y at points $x_1, \ldots, x_5$ be

x	1	2	3	4	5
y	3.2	4.9	6.2	6.5	7.3

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.1^2, & 0, & 0, & 0, & 0 \\ 0, & 0.1^2, & 0, & 0, & 0 \\ 0, & 0, & 0.2^2, & 0, & 0 \\ 0, & 0, & 0, & 0.2^2, & 0 \\ 0, & 0, & 0, & 0, & 0.2^2 \end{pmatrix}.$$

Equations (14) enable us to obtain an approximate values $\boldsymbol{\beta}^{(0)}$.

For 1^{st} and 5^{nd} measurement two equations for unknown parameters lead to an approximate values $\boldsymbol{\beta}^{(0)} = (10, 2)'$.

The linear version of the functions (14) obtained by the using the Taylor expansion at the approximate point $\boldsymbol{\beta}^{(0)}$ is in the form $\mathbf{Y} - \mathbf{g}(\boldsymbol{\beta}^{(0)}) = \mathbf{F}\delta\boldsymbol{\beta}$, where $\mathbf{F} = \frac{\partial g(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}'}$ and $\mathbf{g}(\boldsymbol{\beta}^{(0)}) = (g_1(\boldsymbol{\beta}^{(0)}), \ldots, g_5(\boldsymbol{\beta}^{(0)}))'$, $g_i(\boldsymbol{\beta}^{(0)}) = \frac{\beta_1^{(0)} x_i}{\beta_2^{(0)} + x_i}$, $i = 1, 2, 3, 4, 5$.

In our case we will determine

$$\mathbf{F} = \begin{pmatrix} 0.3333 & -1.1111 \\ 0.5000 & -1.2500 \\ 0.6000 & -1.2000 \\ 0.6667 & -1.1111 \\ 0.7143 & -1.0204 \end{pmatrix}, \quad \mathbf{y}_i^0 = \frac{\beta_1^0 x_i}{\beta_2^0 + x_i},\ i = 1, 2, 3, 4, 5$$

$$\mathbf{y}^0 = (3.3333, 5.0000, 6.0000, 6.6667, 7.1429)',$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(0)} + \delta\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(0)} + (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{y}^0) = \begin{pmatrix} 10.5230 \\ 2.2754 \end{pmatrix},$$

$$\mathbf{v} = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}} = (-0.0127, -0.0226, 0.2158, -0.2075, 0.0681)'.$$

Let $\mathbf{h} = (1, 0)'$, $\sigma = 0.1$, $\varepsilon = 0.2$. Then in our linearized model we will determine numerically from the Remark 3.4 formula (9) and from the Remark 3.5 formula (11)

$$\triangle_{\boldsymbol{crit}}^{(\boldsymbol{F})} = \begin{pmatrix} -0.0681 & -0.1294 \\ -0.4915 & -0.9333 \\ 0.3349 & 0.6359 \\ 0.1672 & 0.3174 \\ 0.1184 & 0.2248 \end{pmatrix}, \quad \delta\mathbf{F}^* = \begin{pmatrix} -0.0342 & -0.0180 \\ -0.0047 & -0.0025 \\ 0.0070 & 0.0037 \\ 0.0140 & 0.0073 \\ 0.0197 & 0.0104 \end{pmatrix},$$

and from the Remark 3.4 formulas (9), (10) and from the Remark 3.5 formula (11)

$${}_1\triangle_{\boldsymbol{crit}}^{(\boldsymbol{V})} = \begin{pmatrix} -0.5132 & -1.2038 \\ -0.3135 & -0.7568 \\ -0.3516 & -0.8476 \\ -0.2763 & -0.6640 \\ -0.2308 & -0.5531 \end{pmatrix}, \quad {}_2\triangle_{\boldsymbol{crit}}^{(\boldsymbol{V})} = \begin{pmatrix} 0.1342 & 0.3216 \\ 0.2530 & 0.6107 \\ 0.5799 & 1.3966 \\ 0.7268 & 1.7326 \\ 0.8581 & 2.0160 \end{pmatrix},$$

and from the Remark 4.2 formulas (12) and (13)

$$\triangle_{\boldsymbol{crit}}^{(\boldsymbol{\Sigma})} = \begin{pmatrix} 0.0095 & 0.0083 & 0.0116 & 0.0126 & 0.0160 \\ 0.0083 & 0.0536 & 0.0310 & 0.0172 & 0.0125 \\ 0.0116 & 0.0310 & 0.0600 & 0.0303 & 0.0226 \\ 0.0126 & 0.0172 & 0.0303 & 0.0329 & 0.0296 \\ 0.0160 & 0.0125 & 0.0226 & 0.0296 & 0.0273 \end{pmatrix}.$$

Let $\mathbf{h} = (0,1)'$, $\varepsilon = 0.2$. Then

$$\triangle_{\boldsymbol{crit}}^{(\boldsymbol{F})} = \begin{pmatrix} 0.0027 & -0.1200 \\ 0.0043 & -0.1912 \\ 0.0217 & -0.9609 \\ -0.0083 & 0.3674 \\ -0.0038 & 0.1699 \end{pmatrix}, \quad \delta\mathbf{F}^* = \begin{pmatrix} -0.0315 & -0.0166 \\ -0.0114 & -0.0060 \\ 0.0031 & 0.0016 \\ 0.0082 & 0.0043 \\ 0.0125 & 0.0066 \end{pmatrix},$$

$${}_1\triangle_{\boldsymbol{crit}}^{(\boldsymbol{V})} = \begin{pmatrix} -0.5941 & -1.3733 \\ -0.5255 & -1.1809 \\ -0.6593 & -1.4756 \\ -0.5643 & -1.2680 \\ -0.4958 & -1.1190 \end{pmatrix}, \quad {}_2\triangle_{\boldsymbol{crit}}^{(\boldsymbol{V})} = \begin{pmatrix} 0.2216 & 0.5018 \\ 0.3422 & 0.7678 \\ 0.8050 & 1.8022 \\ 0.9644 & 2.1739 \\ 1.1108 & 2.5353 \end{pmatrix},$$

$$\triangle_{\boldsymbol{crit}}^{(\boldsymbol{\Sigma})} = \begin{pmatrix} 0.0076 & 0.0080 & 0.0097 & 0.0108 & 0.0143 \\ 0.0080 & 0.0166 & 0.0229 & 0.0165 & 0.0129 \\ 0.0097 & 0.0229 & 0.1013 & 0.0389 & 0.0262 \\ 0.0108 & 0.0165 & 0.0389 & 0.0415 & 0.0347 \\ 0.0143 & 0.0129 & 0.0262 & 0.0347 & 0.0321 \end{pmatrix}.$$

6 Concluding remarks

The aim in linear statistical models is to determine an estimator of the parameter $\boldsymbol{\beta}$ on the basis of the observation vector $\mathbf{Y}$.

In this article we concentrated on a fundamental questions – how uncertainty of the design and covariance matrices influence the bias and the variance of estimators.

The quantities $\triangle_{\boldsymbol{crit}}^{(\boldsymbol{F})}$, $\delta\mathbf{F}^*$, $\triangle_{\boldsymbol{crit}}^{(\boldsymbol{V})}$, $\triangle_{\boldsymbol{crit}}^{(\boldsymbol{\Sigma})}$ enables to judge how precise the record of the design matrix and the covariance matrix must be.

In the last example it can be seen that in the situation B for $\varepsilon = 0.2$ the record of the design matrix must take into account the values 0.001 and that record of the covariance matrix must take into account the values 0.01.

References

[1] Kubáček, L., Kubáčková, L.: Statistics and Metrology. *Vyd. Univ. Palackého, Olomouc,* 2000 (in Czech).

[2] Rao, C. R.: Linear Statistical Inference and Its Applications. *J. Wiley, New York–London–Sydney,* 1973 (second editon).

[3] Rao, C. R., Mitra, S. K.: Generalized Inverse of Matrices and its Applications. *J. Wiley, New York–London–Sydney–Toronto*, 1971.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 73–82

Linearization Regions for a Confidence Ellipsoid in Singular Nonlinear Regression Models*

Lubomír KUBÁČEK [1], Eva TESAŘÍKOVÁ [2]

[1] *Department of Mathematical Analysis and Applications of Mathematics*
Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: kubacekl@inf.upol.cz

[2] *Department of Algebra and Geometry, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: tesariko@inf.upol.cz

(Received March 24, 2009)

Abstract

A construction of confidence regions in nonlinear regression models is difficult mainly in the case that the dimension of an estimated vector parameter is large. A singularity is also a problem. Therefore some simple approximation of an exact confidence region is welcome. The aim of the paper is to give a small modification of a confidence ellipsoid constructed in a linearized model which is sufficient under some conditions for an approximation of the exact confidence region.

Key words: Nonlinear regression model, confidence region, singularity.

2000 Mathematics Subject Classification: 62F10, 62J05

1 Introduction

A construction of a confidence region for unbiasedly estimable functions of nonlinear singular regression model parameters can be a difficult numerical problem (for more detail on nonlinear models cf. [6]). Mainly the case of a large dimension of a vector parameter is unwelcome. If a confidence region can be

*Supported by the Council of the Czech Government MSM 6 198 959 214.

approximated by a confidence ellipsoid (in the case of normally distributed observation vector), then a numerical calculation and an interpretation of results are much more easier and simpler.

Therefore an attempt to find a simple measure of nonlinearity which enable us to decide whether an approximate confidence ellipsoid can be used instead of exact confidence region, is the aim of the paper.

2 Notation and some useful statements

The following notation is used.

$$\mathbf{Y} \sim N_n(\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Sigma}) \tag{1}$$

means that $\mathbf{Y}$ is an n-dimensional normally distributed random vector with the mean value $E(\mathbf{Y})$ equal to $\mathbf{f}(\boldsymbol{\beta})$ and with the covariance matrix $\mathrm{Var}(\mathbf{Y}) = \boldsymbol{\Sigma}$. Let the function $\mathbf{f}(\cdot)\colon R^k \to R^n$ (R^n is the n-dimensional real linear vector space) can be expressed by the Taylor series of the second order, i.e.

$$\begin{aligned}
\mathbf{f}(\boldsymbol{\beta}) &= \mathbf{f}_0 + \mathbf{F}\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}), \quad \delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0,\\
\mathbf{f}_0 &= \mathbf{f}(\boldsymbol{\beta}_0), \quad \boldsymbol{\beta}_0 \text{ is an approximate value of } \boldsymbol{\beta},\\
\mathbf{F} &= \frac{\partial \mathbf{f}(\mathbf{u})}{\partial \mathbf{u}}\Big|_{u=\beta_0}, \quad \boldsymbol{\kappa}(\delta\boldsymbol{\beta}) = \big[\kappa_1(\delta\boldsymbol{\beta}), \ldots, \kappa_n(\delta\boldsymbol{\beta})\big]',\\
\kappa_i(\delta\boldsymbol{\beta}) &= \delta\boldsymbol{\beta}' \frac{\partial^2 f_i(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}'}\Big|_{u=\beta_0} \delta\boldsymbol{\beta}, \quad i = 1, \ldots, n.
\end{aligned}$$

The matrix $\mathbf{F}$ need not be of the full rank in columns and $\boldsymbol{\Sigma}$ need not be positive definite.

The linearized version of the model (1) is

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n(\mathbf{F}\delta\boldsymbol{\beta}, \boldsymbol{\Sigma}) \tag{2}$$

and the quadratized version is

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n\left(\mathbf{F}\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}), \boldsymbol{\Sigma}\right). \tag{3}$$

In the following text the notations

$\mathbf{A}^-$ ... g-inverse (generalized inverse) of the matrix $\mathbf{A}$,

$\mathbf{A}^+$ ... the Moore–Penrose g-inverse of the matrix $\mathbf{A}$,

$\mathbf{A}^-_{m(W)}$ ... minimum $\mathbf{W}$-seminorm g-inverse of the matrix $\mathbf{A}$, ($\mathbf{W}$ is positive semidefinite matrix),

$\mathcal{M}(\mathbf{A}_{m,n}) = \{\mathbf{A}\mathbf{u}\colon \mathbf{u} \in R^n\}$ (column space of the matrix) $\mathbf{A}$,

$\mathbf{I}$ ... identity matrix,

$\mathbf{P}_{F'} = \mathbf{F}'(\mathbf{F}\mathbf{F}')^{-}\mathbf{F}$ the projection matrix on the space $\mathcal{M}(\mathbf{F}')$ in the Euclidean norm,

$r(\mathbf{A})$ ... the rank of the matrix $\mathbf{A}$,

$\mathbf{U} = \mathrm{Var}(\widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}})$,

$\mathbf{T} = \boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}'$,

will be used. More on a g-inverse of a matrix cf. [7].

In the model (2) a representative of all unbiasedly estimable linear functions of the parameter $\boldsymbol{\beta}$ is the vector

$$\boldsymbol{\gamma} = \mathbf{P}_{F'}\boldsymbol{\beta} = \mathbf{P}_{F'}\boldsymbol{\beta}_0 + \mathbf{P}_{F'}\delta\boldsymbol{\beta} = \boldsymbol{\gamma}_0 + \delta\boldsymbol{\gamma}.$$

Lemma 1 *In the model (2) the* $(1-\alpha)$*-confidence ellipsoid of the vector* $\mathbf{P}_{F'}\delta\boldsymbol{\beta}$ *is*

$$\begin{aligned}\mathcal{E}_{P_{F'}\delta\beta} = \Big\{\mathbf{P}_{F'}\mathbf{u}\colon \mathbf{P}_{F'}\mathbf{u} - \widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}} \in \mathcal{M}\big[\mathrm{Var}(\widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}})\big], (\mathbf{P}_{F'}\mathbf{u} - \widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}})' \\ \times \big[\mathrm{Var}(\widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}})\big]^{-}(\mathbf{P}_{F'}\mathbf{u} - \widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}}) \le \chi^2_{r[F'(\Sigma+\mathbf{F}\mathbf{F}')^{-}\Sigma]}(0;1-\alpha)\Big\},\end{aligned}$$

where

$$\begin{aligned}\widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}} &= \mathbf{P}_{F'}\big[(\mathbf{F}')^{-}_{m(\Sigma)}\big]'(\mathbf{Y} - \mathbf{f}_0), \\ \mathrm{Var}(\widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}}) &= \mathbf{P}_{F'}\big[(\mathbf{F}'\mathbf{T}^{-}\mathbf{F})^{-} - \mathbf{I}\big]\mathbf{P}_{F'}, \quad \mathbf{T} = \boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}'.\end{aligned}$$

Proof is given in [2]. □

In the following text it is necessary to take into account the fact that even $\boldsymbol{\beta}_0$ can be considered to be known, only $\mathbf{P}_{F'}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \mathbf{P}_{F'}\delta\boldsymbol{\beta}$ can be unbiasedly estimated. Let

$$\boldsymbol{\beta}_0 = \boldsymbol{\gamma}_0 + \boldsymbol{\omega}_0, \quad \boldsymbol{\gamma}_0 = \mathbf{P}_{F'}\boldsymbol{\beta}_0, \quad \boldsymbol{\omega}_0 = \mathbf{M}_{F'}\boldsymbol{\beta}_0;$$

the parameter $\delta\boldsymbol{\gamma} = \mathbf{P}_{F'}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is unbiasedly estimable in the model (2), however $\delta\boldsymbol{\omega} = \mathbf{M}_{F'}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ is not. Therefore the model

$$\mathbf{Y} \sim N_n\left[\mathbf{f}(\boldsymbol{\beta}_0) + \mathbf{F}\delta\boldsymbol{\gamma} + \frac{1}{2}\boldsymbol{\kappa}_{\omega_0}(\delta\boldsymbol{\gamma}), \boldsymbol{\Sigma}\right] \tag{4}$$

will be considered instead the model (3). Here

$$\begin{aligned}\boldsymbol{\kappa}_{\omega_0} &= (\kappa_{\omega_0,1}, \ldots, \kappa_{\omega_0,n})', \\ \kappa_{\omega_0,i} &= \delta\boldsymbol{\gamma}'\frac{\partial^2 f_i(\boldsymbol{\gamma}_0 + \boldsymbol{\omega}_0)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\delta\boldsymbol{\gamma}, \quad i = 1, \ldots, n, \\ \mathbf{F} &= \frac{\partial\mathbf{f}(\boldsymbol{\gamma}_0 + \boldsymbol{\omega}_0)}{\partial\boldsymbol{\gamma}'}.\end{aligned}$$

Lemma 2 *The bias* $\mathbf{b}$ *of the estimator*

$$\widehat{\delta\boldsymbol{\gamma}} = \widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}} = \mathbf{P}_{F'}\big[(\mathbf{F}')^{-}_{m(\Sigma)}\big]'(\mathbf{Y} - \mathbf{f}_0)$$

in the model (4) is

$$\begin{aligned}\mathbf{b} \;&=\; E(\widehat{\delta\boldsymbol{\gamma}}) - \delta\boldsymbol{\gamma} = \frac{1}{2}\mathbf{P}_{F'}\big[(\mathbf{F}')^{-}_{m(\Sigma)}\big]'\boldsymbol{\kappa}_{\omega_0}(\delta\boldsymbol{\gamma})\\ &= \frac{1}{2}\mathbf{P}_{F'}(\mathbf{F}'\mathbf{T}^{-}\mathbf{F})^{-}\mathbf{F}'\mathbf{T}^{-}\boldsymbol{\kappa}_{\omega_0}(\delta\boldsymbol{\gamma}).\end{aligned}$$

Proof is implied by the definition of the bias. □

Lemma 3 *Let* $\mathbf{Y} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *Then*

$$\mathbf{Y}'\boldsymbol{\Sigma}^{+}\mathbf{Y} \sim \chi^2_{r(\Sigma)}(\delta),$$

where $\delta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}\boldsymbol{\mu} = \boldsymbol{\mu}'\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}\boldsymbol{\mu}$.

Proof Let $\mathbf{J}$ be a $k \times r(\boldsymbol{\Sigma})$ matrix such that $\mathbf{J}\mathbf{J}' = \boldsymbol{\Sigma}$ and $\mathbf{K}$ such a $k \times r(\boldsymbol{\Sigma})$ matrix that $\mathbf{K}\mathbf{K}' = \boldsymbol{\Sigma}^{+}$ (i.e. $\mathbf{J}'\mathbf{K} = \mathbf{I}$). Then $\mathbf{K}'\mathbf{Y} = \mathbf{K}'\boldsymbol{\mu} + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim N_{r(\Sigma)}(\mathbf{0}, \mathbf{I})$. Thus

$$\mathbf{Y}'\mathbf{K}\mathbf{K}'\mathbf{Y} \;=\; \mathbf{Y}'\boldsymbol{\Sigma}^{+}\mathbf{Y} = \boldsymbol{\eta}'\boldsymbol{\eta} + 2\boldsymbol{\eta}'\mathbf{K}'\boldsymbol{\mu} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}\boldsymbol{\mu} \sim \chi^2_{r(\Sigma)}(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}\boldsymbol{\mu}).$$

However $\boldsymbol{\Sigma}^{+} = \mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}$, since

$$\begin{gathered}\boldsymbol{\Sigma}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}, \quad \mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma} = \mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma},\\ \boldsymbol{\Sigma}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma} = \mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma} = \mathbf{P}_{\Sigma}, \quad \mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{P}_{\Sigma}\boldsymbol{\Sigma}^{-}\mathbf{P}_{\Sigma} = \mathbf{P}_{\Sigma}.\end{gathered}$$

(in more detail cf. [7]). □

3 A linearization region for a confidence ellipsoid

Since $r\big[\operatorname{Var}(\widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}})\big] = r[\mathbf{F}'(\boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}')^{-}\boldsymbol{\Sigma}]$, it can happen that

$$r\big[\operatorname{Var}(\widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}})\big] = r\big[\operatorname{Var}(\widehat{\delta\boldsymbol{\gamma}})\big] < r(\mathbf{F}').$$

Therefore the vector $\mathbf{b}$ need not be an element of $\mathcal{M}\big[\operatorname{Var}(\widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}})\big]$.

The relation

$$\delta\boldsymbol{\gamma} = \mathbf{P}_{F'\delta}\boldsymbol{\beta} = E(\widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}}) - \mathbf{b} = E(\widehat{\delta\boldsymbol{\gamma}}) - \mathbf{b},$$

valid in the model (3) and (4), respectively, implies that in general case the vector $\mathbf{P}_{F'\delta}\boldsymbol{\beta}$ need not be an element of $\mathcal{E}_{P_{F'\delta\beta}}$ from Lemma 1. Thus it seems to be reasonable to enlarge the ellipsoid $\mathcal{E}_{P_{F'\delta\beta}}$ to $\overline{\mathcal{E}}$ in such a way that $\mathbf{P}_{F'\delta}\boldsymbol{\beta} \in \overline{\mathcal{E}}$ with sufficiently high probability.

In the following text the notation $\mathbf{U} = \operatorname{Var}(\widehat{\mathbf{P}_{F'\delta}\boldsymbol{\beta}})$ will be used.

Definition 1 Let a set $\overline{\mathcal{E}}$ be defined as

$$\overline{\mathcal{E}} = \Big\{ \mathbf{P}_{F'}\mathbf{u} \colon \mathbf{u} \in R^k, (\mathbf{P}_{F'}\mathbf{u} - \widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}})' \big[\mathbf{U} + c^2(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^+ \times (\mathbf{P}_{F'}\mathbf{u} - \widehat{\mathbf{P}_{F'}\delta\boldsymbol{\beta}}) \le \chi^2_{r(F'T^-\Sigma)}(0; 1-\alpha) \Big\},$$

where $\mathbf{T} = \boldsymbol{\Sigma} + \mathbf{F}\mathbf{F}'$ and the choice c^2 depends on the opinion of the user (cf. the following remark).

Remark 1 The number c^2 should be comparable with the spectral numbers of the matrix $\mathbf{U}$. The semiaxes of $\overline{\mathcal{E}}$ in the space $\mathcal{M}(\mathbf{P}_{F'} - \mathbf{P}_U)$ have the same size equal to

$$a = c\sqrt{\chi^2_{r(F'T^-\Sigma)}(0; 1-\alpha)}.$$

The smaller is c, the smaller is the probability $P\{\mathbf{P}_{F'}\delta\boldsymbol{\beta} \in \overline{\mathcal{E}}\}$. Thus c cannot be smaller than some reasonable bound. If $\mathbf{b} \in \mathcal{M}(\mathbf{U})$, then it can be tolerated in the case $\mathbf{b}'\mathbf{U}^-\mathbf{b} \le \varepsilon$. Let

$$\mathbf{U} = \sum_{i=1}^{f} \lambda_i \mathbf{f}_i \mathbf{f}_i', \quad f = r(\mathbf{F}'\mathbf{T}^-\boldsymbol{\Sigma}),$$

be spectral decomposition of the matrix $\mathbf{U}$ and

$$\lambda_{\max} = \max\{\lambda_i \colon i = 1, \dots, r(\mathbf{F}'\mathbf{T}^-\boldsymbol{\Sigma})\}.$$

If $\mathbf{h} = s\mathbf{f}_{\max}$ (the vector $\mathbf{f}_{\max}$ corresponds to $\lambda_{\max}$), then, regarding the Scheffé theorem [8] ($\mathbf{b}'\mathbf{U}^-\mathbf{b} \le \varepsilon \Leftrightarrow \forall\{\mathbf{h} \in \mathcal{M}(\mathbf{U})\}|\mathbf{h}'\mathbf{b}| \le \varepsilon\sqrt{\mathbf{h}'\mathbf{U}\mathbf{h}}$),

$$|\mathbf{h}'\mathbf{b}| = s|\mathbf{f}'_{\max}\mathbf{b}| \le s\varepsilon\sqrt{\lambda_{\max}}.$$

In the worst case (i.e. $\mathbf{b} = t\mathbf{f}_{\max}$) $\|\mathbf{b}\| = t < \varepsilon\sqrt{\lambda_{\max}}$. It implies that the bias $\mathbf{b}$ with the norm smaller than $\varepsilon\sqrt{\lambda_{\max}}$ can be tolerated and thus the choice $c^2 = \lambda_{\max}$ is reasonable.

Definition 2 Let the measure of nonlinearity for the confidence ellipsoid be

$$C^{(ell)} = \sup \left\{ \frac{2\sqrt{\mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^+\mathbf{b}(\delta\boldsymbol{\gamma})}}{\delta\boldsymbol{\gamma}'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^+\delta\boldsymbol{\gamma}} \colon \delta\boldsymbol{\gamma} \in R^{r(F)} \right\},$$

where

$$\mathbf{b}(\delta\boldsymbol{\gamma}) = \frac{1}{2}\mathbf{P}_{F'}(\mathbf{F}'\mathbf{T}^-\mathbf{F})^-\mathbf{F}'\mathbf{T}^-\boldsymbol{\kappa}(\delta\boldsymbol{\gamma}).$$

Theorem 1 *If* $\delta\boldsymbol{\beta} \in \mathcal{L}^{(ell)}_{\delta\gamma}$, *where*

$$\mathcal{L}^{(ell)}_{\delta\gamma} = \left\{ \delta\boldsymbol{\gamma} \colon \delta\boldsymbol{\gamma} \in \mathcal{M}(\mathbf{F}'), \delta\boldsymbol{\gamma}'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^-\delta\boldsymbol{\gamma} \le \frac{2\sqrt{\delta_{\max}}}{C^{(ell)}} \right\},$$

then

$$P\Big\{\delta\boldsymbol{\gamma} \in \overline{\mathcal{E}}\Big\} \geq 1 - \alpha - \varepsilon.$$

Here $\delta_{\max}$ *is a solution of the equation*

$$P\Big\{\chi^2_f(\delta_{\max}) \leq \chi^2_f(0; 1-\alpha)\Big\} = 1 - \alpha - \varepsilon$$

and $f = r(\mathbf{F}'\mathbf{T}^{-}\boldsymbol{\Sigma})$.

Proof Regarding Definition 6

$$2\sqrt{\mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{+}\mathbf{b}(\delta\boldsymbol{\gamma})} \leq \delta\boldsymbol{\gamma}'\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{-}\delta\boldsymbol{\gamma}C^{(ell)}.$$

Let

$$\delta\boldsymbol{\gamma}'\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{-}\delta\boldsymbol{\gamma} \leq \frac{2\sqrt{\delta_{\max}}}{C^{(ell)}}.$$

Further

$$\begin{aligned}
(\widehat{\delta\boldsymbol{\gamma}} - \delta\boldsymbol{\gamma})'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}(\widehat{\delta\boldsymbol{\gamma}} - \delta\boldsymbol{\gamma}) = \\
= \big[\widehat{\delta\boldsymbol{\gamma}} - E(\widehat{\delta\boldsymbol{\gamma}}) + E(\widehat{\delta\boldsymbol{\gamma}}) - \delta\boldsymbol{\gamma}\big]'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+} \\
\quad \times \big[\widehat{\delta\boldsymbol{\gamma}} - E(\widehat{\delta\boldsymbol{\gamma}}) + E(\widehat{\delta\boldsymbol{\gamma}}) - \delta\boldsymbol{\gamma}\big] \\
= \big[\widehat{\delta\boldsymbol{\gamma}} - E(\widehat{\delta\boldsymbol{\gamma}})\big]'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}\big[\widehat{\delta\boldsymbol{\gamma}} - E(\widehat{\delta\boldsymbol{\gamma}})\big] \\
\quad + 2\mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}\big[\widehat{\delta\boldsymbol{\gamma}} - E(\widehat{\delta\boldsymbol{\gamma}})\big] \\
\quad + \mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}\mathbf{b}(\delta\boldsymbol{\gamma}) = \chi^2_f(\delta),
\end{aligned}$$

where

$$\delta = \mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}\mathbf{b}(\delta\boldsymbol{\gamma}),$$

what is implied by Lemma 3. The relation

$$\begin{aligned}
&\big[(\mathbf{Y}-\boldsymbol{\mu})+\boldsymbol{\mu}\big]'\boldsymbol{\Sigma}^{+}\big[(\mathbf{Y}-\boldsymbol{\mu})+\boldsymbol{\mu}\big] = \\
&= (\mathbf{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-}(\mathbf{Y}-\boldsymbol{\mu}) + 2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}(\mathbf{Y}-\boldsymbol{\mu}) + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}\boldsymbol{\mu} = \chi^2_{r(\Sigma)}(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{+}\boldsymbol{\mu}),
\end{aligned}$$

based on Lemma 3 is used as well.

Thus

$$(\widehat{\delta\boldsymbol{\gamma}} - \delta\boldsymbol{\gamma})'\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}(\widehat{\delta\boldsymbol{\gamma}} - \delta\boldsymbol{\gamma}) = \chi^2_f(\delta),$$

where

$$\delta = \mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U} + \lambda_{\max}(\mathbf{P}_{F'} - \mathbf{P}_U)\big]^{+}\mathbf{b}(\delta\boldsymbol{\gamma}).$$

If $\delta \leq \delta_{\max}$, then

$$P\Big\{\chi^2_f(\delta) \leq \chi^2_f(0; 1-\alpha)\Big\} \geq P\Big\{\chi^2_f(\delta_{\max}) \leq \chi^2_f(0; 1-\alpha)\Big\} = 1 - \alpha - \varepsilon,$$

what means $P\Big\{\delta\boldsymbol{\gamma} \in \overline{\mathcal{E}}\Big\} \geq 1 - \alpha - \varepsilon$. □

Remark 2 Let us apply Theorem 1 on the regular linearized model. Then $\mathbf{P}_{F'} = \mathbf{P}_U = \mathbf{I}$, $\overline{\mathcal{E}} = \mathcal{E}_{\delta\gamma}$ and $C^{(ell)} = K^{(par)}$, where $K^{(par)}$ is the Bates and Watts parametric curvature

$$K^{(par)} = \sup\left\{\frac{\sqrt{\boldsymbol{\kappa}'(\delta\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}\mathbf{P}_F^{\Sigma^{-1}}\boldsymbol{\kappa}(\delta\boldsymbol{\beta})}}{\delta\boldsymbol{\beta}'\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\delta\boldsymbol{\beta}}\colon \delta\boldsymbol{\beta}\in R^k\right\}$$

(in more detail cf. [1]). In this case the statement

$$\delta\boldsymbol{\beta}\in\left\{\mathbf{u}\colon \mathbf{u}'\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{u}\le\frac{2\sqrt{\delta_{\max}}}{K^{(par)}}\right\}\Rightarrow P\left\{\mathbf{P}_{F'}\delta\boldsymbol{\beta}\in\mathcal{E}_{P_{F'}\delta\beta}\right\}$$
$$= P\{\delta\boldsymbol{\beta}\in\mathcal{E}_{\delta\beta}\}\ge 1-\alpha-\varepsilon$$

is true (cf. also [4]). Thus Theorem 7 is a reasonable generalization suitable for the singular model.

Remark 3 In the case that only one function of the parameter $\boldsymbol{\beta}$, i.e. $h(\boldsymbol{\gamma}) = \mathbf{h}'\boldsymbol{\gamma}_0 + \mathbf{h}'\delta\boldsymbol{\gamma}$, $\delta\boldsymbol{\gamma}\in\mathcal{M}(\mathbf{F}')$, is important, a very simple procedure can be used. Let in the first case $\mathbf{h}'\mathbf{P}_{F'}\left[(\mathbf{F}'\mathbf{T}^-\mathbf{F})^- - \mathbf{I}\right]\mathbf{P}_{F'}\mathbf{h} > 0$.

Since

$$b_h = E(\widehat{\mathbf{h}'\delta\boldsymbol{\gamma}}) - \mathbf{h}'\delta\boldsymbol{\gamma} = \frac{1}{2}\mathbf{h}'\mathbf{P}_{F'}\left[(\mathbf{F}')^-_{m(\Sigma)}\right]'\boldsymbol{\kappa}_{\omega_0}(\delta\boldsymbol{\gamma}) = \delta\boldsymbol{\gamma}'\mathbf{A}_h\delta\boldsymbol{\gamma},$$

where

$$\mathbf{A}_h = \sum_{i=1}^{n}\left\{\frac{1}{2}\mathbf{h}'\mathbf{P}_{F'}\left[(\mathbf{F}')^-_{m(\Sigma)}\right]'\right\}_i\frac{\partial^2 f_i(\mathbf{u}+\boldsymbol{\omega}_0)}{\partial\mathbf{u}\partial\mathbf{u}'}\Big|_{u=\gamma_0},$$

we obtain

$$\delta\boldsymbol{\gamma}\in\mathcal{L}_{h'\delta\gamma} = \left\{\mathbf{u}\colon \mathbf{u}\in\mathcal{M}(\mathbf{F}'), |\mathbf{u}'\mathbf{A}_{h'\delta\beta}\mathbf{u}|\le\sqrt{\delta_{1,\max}}\right\}$$
$$\Rightarrow P\left\{|\mathbf{h}'\delta\boldsymbol{\gamma} - \widehat{\delta\boldsymbol{\gamma}}|\le\sqrt{\chi_1^2(0;1-\alpha)}\sqrt{\mathbf{h}'\mathbf{P}_{F'}\mathbf{U}\mathbf{P}_{F'}\mathbf{h}}\right\}\ge 1-\alpha-\varepsilon.$$

Here $\delta_{1,\max}$ is a solution of the equation

$$P\left\{\chi_1^2(\delta_{1,\max})\le\chi_1^2(0;1-\alpha)\right\} = 1-\alpha-\varepsilon.$$

If $\mathbf{h}'\mathbf{U}\mathbf{h} = 0$, then

$$P\left\{\mathbf{h}'\widehat{\delta\boldsymbol{\gamma}} - E(\mathbf{h}'\widehat{\delta\boldsymbol{\gamma}}) = 0\right\} = 1$$

and thus

$$P\left\{\mathbf{h}'\widehat{\delta\boldsymbol{\gamma}} = \mathbf{h}'\delta\boldsymbol{\gamma} + \mathbf{h}'\mathbf{b}(\delta\boldsymbol{\gamma})\right\} = 1.$$

Thus

$$\delta\boldsymbol{\gamma}\in\mathcal{L}_{h'\delta\gamma} = \left\{\mathbf{u}\colon \mathbf{u}\in\mathcal{M}(\mathbf{F}'), |\mathbf{u}'\mathbf{A}_h\mathbf{u}|\le\Delta\right\}$$
$$\Rightarrow P\left\{\mathbf{h}'\delta\boldsymbol{\gamma}\in\left\{u\colon u\in R^1, |u-\widehat{\mathbf{h}'\delta\boldsymbol{\gamma}}|\le\Delta\right\} = 1.\right.$$

It is interesting to compare the linearization regions $\mathcal{L}_{\delta\gamma}$ and $\mathcal{L}_{h'\delta\gamma}$.

4 Numerical example

Let us consider the regression model

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{pmatrix} \sim N_6 \left[\begin{pmatrix} \beta_1 \exp(-\beta_3) \\ \beta_1 \exp(-\beta_3) \\ \beta_1 \exp(-\beta_3) \\ \beta_2 \exp(-\beta_3) \\ \beta_2 \exp(-\beta_3) \\ \beta_2 \exp(-\beta_3) \end{pmatrix}, \mathbf{\Sigma}_{6,6} \right], \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \in R^3,$$

$$\mathbf{\Sigma}_{6,6} = \sigma^2 \mathbf{I}_{6,6}, \quad \sigma^2 = (0.5)^2.$$

Then

$$\mathbf{F} = \frac{\partial \mathbf{f}(\mathbf{u} + \boldsymbol{\omega}_0)}{\partial \mathbf{u}'}\Big|_{u=\gamma_0} = \begin{pmatrix} \mathbf{1}_3, & \mathbf{0}, & -\mathbf{1}_3 \\ 0, & \mathbf{1}_3, & -\mathbf{1}_3 \end{pmatrix}, \quad \mathbf{1}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\mathbf{F}_1 = \mathbf{F}_2 = \mathbf{F}_3 = \begin{pmatrix} 0, & 0, & -1 \\ 0, & 0, & 0 \\ -1, & 0, & 1 \end{pmatrix}, \quad \mathbf{F}_4 = \mathbf{F}_5 = \mathbf{F}_6 = \begin{pmatrix} 0, & 0, & 0 \\ 0, & 0, & -1 \\ 0, & -1, & 1 \end{pmatrix}.$$

Here

$$\mathbf{F}_i = \frac{\partial^2 f_i(\mathbf{u} + \boldsymbol{\omega}_0)}{\partial \mathbf{u} \partial \mathbf{u}'}\Big|_{u=\gamma_0}, \quad i = 1, \ldots, 6,$$

$$\mathbf{P}_{F'} = \mathbf{F}'(\mathbf{F}\mathbf{F}')^{-}\mathbf{F} = \frac{1}{3} \begin{pmatrix} 2, & -1, & -1 \\ -1, & 2, & -1 \\ -1, & -1 & 2 \end{pmatrix},$$

$$\widehat{\mathrm{Var}(\mathbf{P}_{F'}\delta\boldsymbol{\beta})} = \mathbf{U} = \mathbf{P}_{F'} \Big\{ \big[\mathbf{F}'(\mathbf{\Sigma} + \mathbf{F}\mathbf{F}')^{-}\mathbf{F}\big]^{-} - \mathbf{I} \Big\} \mathbf{P}_{F'} = \frac{\sigma^2}{54} \begin{pmatrix} 10, & -8, & -2 \\ -8, & 10, & -2 \\ -2, & -2, & 4 \end{pmatrix},$$

$$\mathbf{P}_U = \mathbf{U}(\mathbf{U}^2)^{-}\mathbf{U} = \frac{1}{3} \begin{pmatrix} 2, & -1, & -1 \\ -1, & 2, & -1 \\ -1, & -1 & 2 \end{pmatrix},$$

$$\mathbf{U} = \sum_{i=1}^{r[\mathbf{F}'(\Sigma + \mathbf{F}\mathbf{F}')^{-}\Sigma]} \lambda_i \mathbf{f}_i \mathbf{f}_i' = \sum_{i=1}^{2} \lambda_i \mathbf{f}_i \mathbf{f}_i', \quad \lambda_1 = \frac{1}{3}\sigma^2, \ \lambda_2 = \frac{1}{9}\sigma^2, \ \lambda_{\max} = \frac{1}{3}\sigma^2,$$

$\delta_{\max} = 0.48$ is a solution of the equation

$$P\left\{\chi^2_f(0; 1-\alpha)\right\} = 1 - \alpha - \varepsilon,$$

and $f = r\big[\mathbf{F}'(\mathbf{\Sigma} + \mathbf{F}\mathbf{F}')^{-}\mathbf{\Sigma}\big] = 2$, $\alpha = 0.05$, $\varepsilon = 0.04$.

Further

$$C^{(ell)} = \sup\left\{\frac{2\sqrt{\mathbf{b}'(\delta\boldsymbol{\gamma})\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{+}\mathbf{b}(\delta\boldsymbol{\gamma})}}{\delta\boldsymbol{\gamma}'\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{+}\delta\boldsymbol{\gamma}}\colon \delta\boldsymbol{\gamma}\in R^2\right\}$$
$$= \sigma\cdot 0.191273,$$

where

$$\mathbf{b} = \frac{1}{2}\mathbf{P}_{F'}(\mathbf{F}'\mathbf{T}^{-}\mathbf{F})^{-}\mathbf{F}'\mathbf{T}^{-}\boldsymbol{\kappa}_{\omega_0}(\delta\boldsymbol{\gamma}).$$

The linearization region for $\delta\boldsymbol{\gamma} = \mathbf{P}_{F'}\delta\boldsymbol{\beta}$ is

$$\mathcal{L}_{\delta\gamma} = \left\{\mathbf{u}\colon \mathbf{u}\in\mathcal{M}(\mathbf{F}'), \mathbf{u}'\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{+}\mathbf{u} \le \frac{2\sqrt{\delta_{\max}}}{C^{(ell)}}\right\}$$

and the set $\overline{\mathcal{E}_{\delta\gamma}}$ is

$$\overline{\mathcal{E}_{\delta\gamma}} = \Big\{\mathbf{u}\colon \mathbf{u}\in\mathcal{M}(\mathbf{F}'), (\mathbf{u}-\widehat{\delta\boldsymbol{\gamma}})'\big[\mathbf{U}+\lambda_{\max}(\mathbf{P}_{F'}-\mathbf{P}_U)\big]^{+}$$
$$\times(\mathbf{u}-\widehat{\delta\boldsymbol{\gamma}}) \le \chi^2_{r(F'T^{-}\Sigma)}(0;1-\alpha)\Big\}$$

The linearization region $\mathcal{L}_{\delta\boldsymbol{\gamma}}$ is the ellipse in the subspace $\mathcal{M}(\mathbf{F}')$ with the semi-axes

$$a_{\mathcal{L},1} = 1.5539\,\sqrt{\sigma}, \quad a_{\mathcal{L},2} = 0.8972\,\sqrt{\sigma}$$

and $\overline{\mathcal{E}_{\delta\gamma}}$ is the ellipse in $\mathcal{M}(\mathbf{F}')$ with the semi-axes

$$a_{\mathcal{E},1} = 0.2359\,\sigma, \quad a_{\mathcal{E},2} = 0.1362\,\sigma.$$

For $\sigma = 0.5$ it means

$$a_{\mathcal{L},1} = 1.099, \quad a_{\mathcal{L},2} = 0.634$$

and

$$a_{\mathcal{E},1} = 0.118, \quad a_{\mathcal{E},2} = 0.068.$$

Thus the linearization is possible.

As far as the single function of $\boldsymbol{\beta}$ is concerned let us consider $\mathbf{h} = (1,0,0)'$.

$$\mathbf{A}_h = \sum_{s=1}^{6}\left\{\frac{1}{2}\mathbf{h}'\mathbf{P}_{F'}\big[\mathbf{F}'(\boldsymbol{\Sigma}+\mathbf{F}\mathbf{F}')^{-}\mathbf{F}\big]^{-}\mathbf{F}'(\boldsymbol{\Sigma}+\mathbf{F}\mathbf{F}')^{-}\right\}_s \mathbf{F}_s$$
$$= \frac{1}{18}\begin{pmatrix} 0, & 0, & -6 \\ 0, & 0, & 3 \\ -6, & 3, & 3 \end{pmatrix}$$

and

$$\mathcal{L}_{h'\delta\gamma} = \{\mathbf{u}\colon \mathbf{u}\in\mathcal{M}(\mathbf{F}'), \mathbf{u}'\mathbf{A}_h\mathbf{u} \le \delta_{1,\max}\}$$

where $\delta_{1,\max} = 0.339$ is a solution of the equation

$$P\Big\{\chi_1^2(\delta_{1,\max}) \leq \chi_1^2(0; 0.95)\Big\} = 1 - 0.05 - 0.04.$$

The linearization region $\mathcal{L}_{h'\delta\gamma}$ is the hyperbola in $\mathcal{M}(\mathbf{F}')$ with the real semi-axis $a = 1.1768$ and the imaginar bi, $b = 1.714$. Thus the linearization region for the confidence interval for $\delta\gamma_1$ is essentially larger (in the case $\sigma = 0.5$) than the linearization region for the whole vector $\delta\boldsymbol{\gamma}$.

References

[1] Bates, D. M., Watts, D. G.: *Relative curvature measures of nonlinearity.* J. Roy. Stat. Soc. **B 42** (1980), 1–25.

[2] Fišerová, E., Kubáček, L., Kunderová, P.: Linear Statistical Models, Regularity and Singularities. *Academia, Praha,* 2007.

[3] Kubáček, L., Kubáčková, L.: *Regression models with a weak nonlinearity.* Technical report Nr. 1998.1, Universität Stuttgart, 1998 1–67.

[4] Kubáček, L., Kubáčková, L.: Statistics and Metrology. *Vyd. Univ. Palackého, Olomouc,* 2000 (in Czech).

[5] Kubáček, L., Tesaříková, E.: *Linearization region for confidence ellispoids.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **47** (2008), 101–113.

[6] Pázman, A.: Nonlinear Statistical Models. *Kluwer Academic Publisher, Dordrecht–Boston–London and Ister Science Press, Bratislava,* 1993.

[7] Rao, C. R., Mitra, S. K.: Generalized Inverse of Matrices and its Applications. *J. Wiley, New York–London–Sydney–Toronto*, 1971.

[8] Scheffé, H.: The Analysis of Variance. *J. Wiley, New York–London–Sydney,* 1967 (fifth printing).

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 83–92

Some Stability Results in Complete Metric Space

Memudu Olaposi OLATINWO

Department of Mathematics, Obafemi Awolowo University,
Ile-Ife, Nigeria
e-mail: polatinwo@oauife.edu.ng

(Received April 26, 2008)

Abstract

In this paper, we obtain some stability results for the Picard iteration process for one and two metrics in complete metric space by using different contractive definitions which are more general than those of Berinde [1], Imoru and Olatinwo [5] some others listed in the reference section. The results generalize and unify some of the results of Harder and Hicks [4], Rhoades [10, 12], Osilike [8], Berinde [1], Imoru and Olatinwo [5] as well as Imoru et al [6].

Key words: Stability results, Picard and Mann iteration processes.

2000 Mathematics Subject Classification: 47H06, 54H25

1 Preliminaries and Introduction

Let (E, d) be a complete metric space, $T : E \to E$ a selfmap of E.

Definition 1.1 [Harder and Hicks [4]]: Suppose that $F_T = \{p \in E \mid Tp = p\}$ is the set of fixed points of T. Let $\{x_n\}_{n=0}^{\infty} \subset E$ be the sequence generated by an iteration procedure involving T which is defined by

$$x_{n+1} = f(T, x_n), \quad n = 0, 1, \ldots, \tag{1.1}$$

where $x_0 \in E$ is the initial approximation and f is some function. Suppose $\{x_n\}_{n=0}^{\infty}$ converges to a fixed point p of T. Let $\{y_n\}_{n=0}^{\infty} \subset E$ and set $\epsilon_n = d(y_{n+1}, f(T, y_n))$, $n = 0, 1, 2, \ldots$ Then, the iteration procedure (1.1) is said to be $T-$stable or stable with respect to T if and only if $\lim_{n\to\infty} \epsilon_n = 0$ implies $\lim_{n\to\infty} y_n = p$.

Definition 1.2 [Singh et al [13]]: Let $S, T: Y \to E$, $T(Y) \subseteq S(Y)$ and z a coincidence point of S and T, that is, $Sz = Tz = p$ (say). For any $x_0 \in Y$, let the sequence $\{Sx_n\}_{n=0}^{\infty}$, generated by the iteration procedure

$$Sx_{n+1} = f(T, x_n), \quad n = 0, 1, \ldots \tag{1.2}$$

converge to p. Let $\{Sy_n\}_{n=0}^{\infty} \subset E$ be an arbitrary sequence, and set $\epsilon_n = d(Sy_{n+1}, f(T, y_n))$, $n = 0, 1, \ldots$ Then, the iteration procedure (1.2) will be called (S, T)-stable if and only if $\lim_{n\to\infty} \epsilon_n = 0$ implies that $\lim_{n\to\infty} Sy_n = p$.

This definition reduces to that of the stability of iteration procedure due to Harder and Hicks [4] when $Y = E$ and $S = I$ (identity operator).

If in (1.1),

$$f(T, x_n) = Tx_n, \quad n = 0, 1, \ldots,$$

then we have the Picard iteration process, while we obtain the Jungck-type iteration if in (1.2)

$$f(T, x_n) = Tx_n, \quad n = 0, 1, \ldots$$

Definition 1.3 [Berinde [2]]: A function $\psi: \mathbb{R}_+ \to \mathbb{R}_+$ is called a comparison function if:

(i) ψ is monotone increasing;

(ii) $\lim\limits_{n\to\infty} \psi^n(t) = 0, \ \forall t \geq 0.$

We remark here that every comparison function satisfies the condition $\psi(0) = 0$.

Several stability results have been obtained by various authors using different contractive definitions. Harder and Hicks [4] obtained interesting stability results for some iteration procedures using various contractive definitions. Rhoades [10, 12] generalized the results of Harder and Hicks [4] to a more general contractive mapping. In Osilike [8], a generalization of some of the results of Harder and Hicks [4] and Rhoades [12] was obtained by employing the following contractive definition: there exist a constant $L \geq 0$ and $a \in [0, 1)$ such $\forall x, y \in E$,

$$d(Tx, Ty) \leq Ld(x, Tx) + ad(x, y). \tag{1.3}$$

Condition (1.3) is more general than those of Rhoades [12] and Harder and Hicks [4]. As in Harder and Hicks [4], Berinde [1] obtained the same stability results for the same iteration procedures using the same contractive definitions, but applied a different method. The method of Berinde [1] is similar to that employed in Osilike and Udomene [9].

Recently, Imoru and Olatinwo [5] obtained some stability results for Picard and Mann iteration procedures by using a more general contractive condition than those of Harder and Hicks [4], Rhoades [12], Osilike [8], Osilike and Udomene [9] and Berinde [1]. In the paper [5], the following contractive definition was employed: there exist $a \in [0, 1)$ and a monotone increasing function $\varphi: \mathbb{R}_+ \to \mathbb{R}_+$, with $\varphi(0) = 0$, such that $\forall x, y \in E$,

$$d(Tx, Ty) \leq \varphi(d(x, Tx)) + ad(x, y). \tag{1.4}$$

It is our purpose in this paper to obtain several stability results in metric space by applying different contractive definitions. However, we shall employ the following lemmas in the sequel.

Lemma 1.4 [Imoru et al [6]]: *If* $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ *is a subadditive comparison function and* $\{\epsilon_n\}_{n=0}^{\infty}$ *is a sequence of positive numbers such that* $\lim_{n\to\infty} \epsilon_n = 0$, *then for any sequence of positive numbers* $\{u_n\}_{n=0}^{\infty}$ *satisfying*

$$u_{n+1} \leq \sum_{m=0}^{s} \delta_m \psi^m(u_n) + \epsilon_n, \quad n = 0, 1, 2, \ldots,$$

where $\sum_{m=0}^{s} \delta_m = 1$, $\delta_0, \delta_1, \cdots, \delta_s \in [0, 1]$, *we have* $\lim_{n\to\infty} u_n = 0$.

Lemma 1.5 [Imoru et al [6]]: *Let* $\{\psi^k(t)\}_{k=0}^{n}$ *be a sequence of comparison functions. Then, any convex linear combination* $\sum_{j=0}^{n} c_j \psi^j(t)$ *of the comparison functions is also a comparison function, where* $\sum_{j=0}^{n} c_j = 1$ *and* $c_o, c_1, \ldots, c_n$ *are positive constants.*

Lemma 1.6 [Imoru et al [6]]: *Let* $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ *be a comparison function and* $\{v_n\}_{n=0}^{\infty}$ *a sequence of positive numbers such that* $\lim_{n\to\infty} v_n = 0$. *Then, we have*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \psi^{n-k}(v_k) = 0, \quad \textit{for each } k.$$

Lemma 1.7 *If* $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ *is a subadditive comparison function and* $\{\epsilon_n\}_{n=0}^{\infty}$ *is a sequence of positive numbers such that* $\lim_{n\to\infty} \epsilon_n = 0$. *Suppose that* $\epsilon > 0$ *is an arbirarily small given number. Then, for any sequence of positive numbers* $\{u_n\}_{n=0}^{\infty}$ *satisfying*

$$u_{n+1} \leq \sum_{k=0}^{m} \delta_k \psi^k(u_n) + \epsilon_n + \epsilon, \quad n = 0, 1, \ldots, \tag{1.5}$$

where $\delta_k \in [0, 1]$, $k = 0, 1, \ldots, m$, $0 \leq \sum_{k=0}^{m} \delta_k \leq 1$, *we have*

$$\lim_{n\to\infty} u_n = 0$$

Proof By putting $\bar{\psi}(u_n) = \sum_{k=0}^{m} \delta_k \psi^k(u_n)$ in (1.5), then we have

$$u_{n+1} \leq \bar{\psi}(u_n) + \epsilon_n + \epsilon, \quad n = 0, 1, \ldots, \tag{1.6}$$

and also by Lemma 1.5, we have that $\bar{\psi}(u_n)$ is a comparison function. It follows from (1.6) that

$$u_1 \leq \bar{\psi}(u_0) + \epsilon_0 + \epsilon,$$

$$u_2 \leq \bar{\psi}(u_1) + \epsilon_1 + \epsilon \leq \bar{\psi}(\bar{\psi}(u_0) + \epsilon_0 + \epsilon) + \epsilon_1 + \epsilon$$
$$\leq [\bar{\psi}^2(u_0) + \bar{\psi}(\epsilon_0) + \epsilon_1] + [\bar{\psi}(\epsilon) + \epsilon],$$

$$u_3 \leq \bar{\psi}(u_2) + \epsilon_2 + \epsilon \leq \bar{\psi}^3(u_0) + \bar{\psi}^2(\epsilon_0) + \bar{\psi}(\epsilon_1) + \bar{\psi}^2(\epsilon) + \bar{\psi}(\epsilon) + \epsilon_2 + \epsilon$$
$$= [\bar{\psi}^3(u_0) + \bar{\psi}^2(\epsilon_0) + \bar{\psi}(\epsilon_1) + \epsilon_2] + [\bar{\psi}^2(\epsilon) + \bar{\psi}(\epsilon) + \epsilon]$$

In general,

$$u_{n+1} \leq \bar{\psi}^{n+1}(u_0) + \sum_{k=0}^{n} \bar{\psi}^{n-k}(\epsilon_k) + \sum_{k=0}^{n} \bar{\psi}^{k}(\epsilon). \tag{1.7}$$

Since $\bar{\psi}$ is a comparison function, then $\lim_{n\to\infty} \bar{\psi}^{n+1}(u_0) = 0$. □

Using Lemma 1.6, we obtain that

$$\lim_{n\to\infty} \sum_{k=0}^{n} \bar{\psi}^{n-k}(\epsilon_k) = 0 \quad \text{and} \quad \lim_{n\to\infty} \sum_{k=0}^{n} \bar{\psi}^{k}(\epsilon) = 0$$

since $\epsilon > 0$ is arbitrary. Hence, (1.7) leads to $\lim_{n\to\infty} u_n = 0$.

We shall establish our main results in the next two sections. Section 2 deals with some stability results involving one metric, while stability results involving two metrics are proved in section 3.

2 Stability results involving one metric in complete metric space

Theorem 2.1 *Let (E, d) be a complete metric space and $T\colon E \to E$ a selfmap of E satisfying*

$$d(Tx, Ty) \leq \frac{\varphi_1(d(x, Tx)) + \psi(d(x, y))}{\varphi_2(d(x, Tx))}, \quad \forall x, y \in E, \tag{2.1}$$

where $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous comparison function and $\varphi_1, \varphi_2\colon \mathbb{R}_+ \to \mathbb{R}_+$ are monotone increasing functions such that $\varphi_1(0) = 0$ and $\varphi_2(0) = 1$. Suppose T has a fixed point p. Let $x_0 \in E$ and let $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Picard iteration associated to T. Then, the Picard iteration process is T-stable.

Proof Let $\{y_n\}_{n=0}^{\infty} \subset E$ and $\epsilon_n = d(y_{n+1}, Ty_n)$. Assume $\lim_{n\to\infty} \epsilon_n = 0$. Then, we shall establish that $\lim_{n\to\infty} y_n = p$ by using the contractive condition and the triangle inequality:

$$d(y_{n+1}, p) \leq d(Tp, Ty_n) + \epsilon_n \leq \psi(d(y_n, p)) + \epsilon_n. \tag{2.2}$$

Using Lemma 1.4 in (2.2) yields $\lim_{n\to\infty} d(y_n, p) = 0$, that is, $\lim_{n\to\infty} y_n = p$. Conversely, let $\lim_{n\to\infty} y_n = p$. Then, by the contractive condition and the triangle inequality, we have

$$\epsilon_n = d(y_{n+1}, Ty_n) \leq d(y_{n+1}, p) + \psi(d(y_n, p)) \to 0 \quad \text{as } n \to \infty. \qquad \square$$

Corollary 2.2 *Let (E, d) be a complete metric space and $T\colon E \to E$ a selfmap of E satisfying*

$$d(Tx, Ty) \leq \frac{\varphi(d(x, Tx)) + ad(x, y)}{1 + Ld(x, Tx)}, \quad \forall x, y \in E,$$

where $a \in [0,1)$, $L \geq 0$ and $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a monotone increasing function such that $\varphi(0) = 0$. Suppose T has a fixed point p. Let $x_0 \in E$ and let $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Picard iteration associated to T. Then, the Picard iteration process is T-stable.

Corollary 2.3 *Let (E, d) be a complete metric space and $T\colon E \to E$ a selfmap of E satisfying*

$$d(Tx, Ty) \leq \varphi_1(d(x, Tx)) + \frac{\psi(d(x, y))}{\varphi_2(d(x, Tx))}, \quad \forall x, y \in E,$$

where $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous comparison function and $\varphi_1, \varphi_2\colon \mathbb{R}_+ \to \mathbb{R}_+$ are monotone increasing functions such that $\varphi_1(0) = 0$ and $\varphi_2(0) = 1$. Suppose T has a fixed point p. Let $x_0 \in E$ and let $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Picard iteration associated to T. Then, the Picard iteration process is T-stable.

Remark 2.4 Theorem 2.1 and its corollaries generalize and unify Theorem 3.1 of Imoru and Olatinwo [5] and several others in the literature. In particular, see Berinde [1], Imoru and Olatinwo [5], Rhoades [10, 11, 12] and some other references in the reference section of this paper for detail.

We now establish the following stability results for uniform convergence of sequences of operators:

Theorem 2.5 *Let (E, d) be a complete metric space and $\{T_n\}_{n=0}^{\infty}$ a sequence of operators $T_n\colon E \to E$. Let $\{x_n\}_{n=0}^{\infty}$ be the Picard iteration process. If the sequence $\{T_n\}_{n=0}^{\infty}$ converges uniformly to an operator $T\colon E \to E$ satisfying*

$$d(Tx, Ty) \leq \varphi(d(x, Tx)) + \psi(d(x, y)), \quad \forall x, y \in E, \tag{2.3}$$

where $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a monotone increasing function such that $\varphi(0) = 0$ and $\psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous, subadditive comparison function. Suppose also that T has the fixed point p. Then, the Picard iteration process is T-stable.

Proof Let $\{y_n\}_{n=0}^{\infty} \subset E$ and let $\epsilon_n = d(y_{n+1}, T_n y_n)$, $d(T_n x, Tx) < \epsilon$, $\forall x \in E$, $\forall n \geq N$. Assume $\lim_{n\to\infty} \epsilon_n = 0$. Then, we shall establish that $\lim_{n\to\infty} y_n = p$ by using the contraction condition (2.3) for T and the triangle inequality:

$$\begin{aligned} d(y_{n+1}, p) &\leq d(y_{n+1}, T_n y_n) + d(T_n y_n, p) \leq d(Tp, Ty_n) + d(Ty_n, T_n y_n) + \epsilon_n \\ &\leq \psi(d(p, y_n)) + \epsilon_n + \epsilon. \end{aligned} \tag{2.4}$$

Using Lemma 1.7 in (2.4) yields

$$d(y_{n+1}, p) \to 0 \quad \text{as } n \to \infty$$

That is, since $\epsilon > 0$ is arbitrary, then $\lim_{n\to\infty} y_n = p$.

Conversely, let $\lim_{n\to\infty} y_n = p$. Then, we have

$$\epsilon_n = d(y_{n+1}, T_n y_n) \leq d(y_{n+1}, p) + \psi(d(p, y_n)) + \epsilon \to 0 \quad \text{as } n \to \infty,$$

since $\epsilon > 0$ is arbitrary. □

Corollary 2.6 *Let (E,d) be a complete metric space and $\{T_n\}_{n=0}^{\infty}$ a sequence of operators $T_n\colon E\to E$. Let $\{x_n\}_{n=0}^{\infty}$ be the Picard iteration process. If the sequence $\{T_n\}_{n=0}^{\infty}$ converges uniformly to an operator $T\colon E\to E$ satisfying*

$$d(Tx,Ty)\leq\varphi(d(x,Tx))+ad(x,y),\quad \forall x,y\in E,\ a\in[0,1),$$

where $\varphi\colon\mathbb{R}_+\to\mathbb{R}_+$ is a monotone increasing function such that $\varphi(0)=0$. Suppose also that T has the fixed point p. Then, the Picard iteration process is T-stable.

Remark 2.7 We remark that this theorem holds if $\{T_n\}$ converges pointwise to T since uniform convergence is more general than pointwise convergence.

Corollary 2.8 *Let (E,d) be a complete metric space and $\{T_n\}_{n=0}^{\infty}$ a sequence of operators $T_n\colon E\to E$. Let $\{x_n\}_{n=0}^{\infty}$ be the Picard iteration process. If the sequence $\{T_n\}_{n=0}^{\infty}$ converges pointwise to an operator $T\colon E\to E$ satisfying*

$$d(Tx,Ty)\leq\varphi(d(x,Tx))+\psi(d(x,y)),\quad \forall x,y\in E,$$

where $\varphi\colon\mathbb{R}_+\to\mathbb{R}_+$ is a monotone increasing function such that $\varphi(0)=0$ and $\psi\colon\mathbb{R}_+\to\mathbb{R}_+$ is a continuous, subadditive comparison function. Suppose also that T has the fixed point p. Then, the Picard iteration process is T-stable.

Remark 2.9 To the best of our knowledge, this is the first time that stability results are being considered using the concepts of uniform and pointwise convergence of sequences of operators.

Theorem 2.10 *Let (E,d) be a complete metric space and Y an arbitrary set. Suppose that $S,T\colon Y\to E$ are nonselfoperators such that $T(Y)\subseteq S(Y)$, $S(Y)$ a complete subspace of E. Let z be a coincidence point of S and T (that is, $Sz=Tz=p$). Suppose that S and T satisfy the contractive condition*

$$d(Tx,Ty)\leq\frac{\psi(d(Sx,Sy))}{1+Md(Sx,Tx)},\quad M\geq 0,\ \forall x,y\in Y, \tag{2.5}$$

where $\psi\colon\mathbb{R}_+\to\mathbb{R}_+$ is a continuous subadditive comparison function. For $x_0\in Y$, let $\{Sx_n\}_{n=0}^{\infty}$ be the Jungck-type iteration process defined by $Sx_{n+1}=Tx_n$, $n=0,1,\ldots$, converging to p. Then, the Jungck-type iteration process is (S,T)-stable.

Proof We now assume that $\lim_{n\to\infty}\epsilon_n=0$ and establish that $\lim_{n\to\infty}Sy_n=p$, using the contractive condition and triangle inequality. Therefore, we have

$$d(Sy_{n+1},p)\leq d(Sy_{n+1},Ty_n)+d(Ty_n,p)\leq\psi(d(p,Sy_n))+\epsilon_n \tag{2.6}$$

By using Lemma 1.4 in (2.6), we get $\lim_{n\to\infty}d(Sy_n,p)=0$, that is,

$$\lim_{n\to\infty}Sy_n=p.$$

Conversely, let $\lim_{n\to\infty} Sy_n = p$. Then, by the contractive condition on S and T as well as the triangle inequality, we have

$$\begin{aligned}\epsilon_n &= d(Sy_{n+1}, Ty_n) \le d(Sy_{n+1}, p) + d(p, Ty_n)\\ &\le d(Sy_{n+1}, p) + \psi(d(p, Sy_n)) \to 0 \quad \text{as } n \to \infty. \qquad \square\end{aligned}$$

Theorem 2.11 *Let S and T be operators on an arbitrary set Y with values in E such that $T(Y) \subseteq S(Y)$ and $S(Y)$ or $T(Y)$ is a complete subspace of E. Let z be a coincidence point of S and T (i.e. $S(z) = T(z) = p$ (say)). Let $x_0 \in Y$ and let $\{Sx_n\}_{n=0}^{\infty} \subset E$ defined by $Sx_{n+1} = Tx_n$, $n = 0, 1, \cdots$, be the Jungck iteration process converging to p. Suppose that $\{Sy_n\}_{n=0}^{\infty} \subset E$ and $\epsilon_n = d(Sy_{n+1}, Ty_n)$, $n = 0, 1, \cdots$ Suppose that S and T satisfy the contractive condition*

$$d(Tx, Ty) \le \frac{\psi(d(Sx, Sy)) + \varphi(d(Sx, Tx))}{1 + Md(Sx, Tx)}, \quad M \ge 0, \ \forall x, y \in Y, \tag{2.7}$$

where $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous subadditive comparison function and $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotone increasing function such that $\varphi(0) = 0$. Then, the Jungck iteration process is (S, T)-stable.

Proof The proof of this theorem follows a similar argument as in that of Theorem 2.10. $\square$

Remark 2.12 Theorem 2.10 and others extend some celebrated results of [1, 4, 8, 9, 12] and some results due to the author [5, 6]. Infact, Theorem 2.10 is also a generalization and extension of Theorem 3.1 of Singh et al [13].

3 Stability results involving two metrics d and ρ on a nonempty set E

Theorem 3.1 *Let E be a nonempty set, d and ρ two metrics on E and $T\colon E \to E$ a mapping. Suppose that:*

(i) T has a fixed point p;

(ii) there exist $c > 0$, and a monotone increasing function $\varphi_1\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi_1(0) = 0$ such that

$$d(Tx, Ty) \le \varphi_1(\rho(x, Tx)) + c\rho(x, y), \quad \forall x, y \in E;$$

(iii) (E, d) is a complete metric space;

(iv) $T\colon (E, \rho) \to (E, \rho)$ satisfies the contractive condition

$$\rho(Tx, Ty) \le \varphi_2(\rho(x, Tx)) + \psi(\rho(x, y)), \quad \forall x, y \in E,$$

where $\psi^k\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, are continuous comparison functions (ψ^k is the k-th iterate of ψ) and $\varphi_2\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, is a monotone increasing function such that $\varphi_2(0) = 0$.

Let $x_0 \in E$ and $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Picard iteration associated to T. Then, the Picard iteration process with $T\colon (E, d) \to (E, d)$ is T-stable.

Proof Let $\{y_n\}_{n=0}^{\infty} \subset E$, $\epsilon_n = d(y_{n+1}, Ty_n)$, $n = 0, 1, \ldots$, and suppose that $\lim_{n\to\infty} \epsilon_n = 0$. Then, we shall establish that $\lim_{n\to\infty} y_n = p$, using conditions (i)-(iv) and the triangle inequality: Therefore, using (i), (ii) and triangle inequality lead to

$$\begin{aligned} d(y_{n+1}, p) &\leq d(Ty_n, Tp) + \epsilon_n \leq \varphi_1(\rho(p, Tp)) + c\rho(p, y_n) + \epsilon_n \\ &= c\rho(y_n, p) + \epsilon_n. \end{aligned} \tag{3.1}$$

Using (iii), we have that $p \in E$. Condition (iv) shows that T has a unique fixed point. Also by condition (iv), we get

$$\begin{aligned} &\rho(y_n, p) = \rho(Ty_{n-1}, Tp) = \rho(Tp, Ty_{n-1}) \leq \psi(\rho(y_{n-1}, p)) \\ &\leq \psi^2(\rho(y_{n-2}, p)) \leq \cdots \leq \psi^n(\rho(y_0, p)) \to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{3.2}$$

Using (3.2) in (3.1), we have

$$d(y_{n+1}, p) \leq c\psi^n(\rho(y_0, p)) + \epsilon_n. \tag{3.3}$$

Taking limits of both sides in (3.3) yields

$$\lim_{n\to\infty} d(y_{n+1}, p) \leq c \lim_{n\to\infty} \psi^n(\rho(y_0, p)) + \lim_{n\to\infty} \epsilon_n \to 0 \quad \text{as } n \to \infty$$

That is, $\lim_{n\to\infty} y_n = p$.

Conversely, let $\lim_{n\to\infty} y_n = p$. Then, by condition (ii) and (3.2) we have

$$\begin{aligned} \epsilon_n &= d(y_{n+1}, Ty_n) \leq d(y_{n+1}, p) + d(Tp, Ty_n) \\ &\leq d(y_{n+1}, p) + c\psi^n(\rho(p, y_0)) \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

□

Corollary 3.2 *Let E be a nonempty set, d and ρ two metrics on E and $T\colon E \to E$ a mapping. Suppose that:*

(i) T has a fixed point p;

(ii) there exist $c > 0$, $M \geq 0$ such that

$$d(Tx, Ty) \leq M\rho(x, Tx) + c\rho(x, y), \quad \forall x, y \in E;$$

(iii) (E, d) is a complete metric space;

(iv) $T\colon (E, \rho) \to (E, \rho)$ satisfies the contractive condition

$$\rho(Tx, Ty) \leq \varphi(\rho(x, Tx)) + \psi(\rho(x, y)), \quad \forall x, y \in E,$$

where $\psi^k\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, are continuous comparison functions (ψ^k is the k-th iterate of ψ) and $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, monotone increasing functions such that $\varphi(0) = 0$.

Let $x_0 \in E$ and $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Picard iteration associated to T. Then, the Picard iteration process with $T\colon (E, d) \to (E, d)$ is T-stable.

Theorem 3.3 *Let E be a nonempty set and Y an arbitrary set. Let d and ρ two metrics on Y and $S, T\colon Y \to E$ nonselfmappings such that $T(Y) \subseteq S(Y)$ and $S(Y)$ is a complete subspace of E. Suppose that:*

(i) S and T have a coincidence point z (that is $Tz = Sz = p$);

(ii) there exist $c > 0$, and a monotone increasing function $\varphi_1\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi_1(0) = 0$ such that

$$d(Tx, Ty) \leq \varphi_1(\rho(Sx, Tx)) + c\rho(Sx, Sy), \quad \forall x, y \in Y;$$

(iii) (E, d) is a complete metric space;

(iv) $T\colon (Y, \rho) \to (E, \rho)$ satisfies the contractive condition

$$\rho(Tx, Ty) \leq \varphi_2(\rho(Sx, Tx)) + \psi(\rho(Sx, Sy)), \quad \forall x, y \in Y,$$

where $\psi^k\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, are continuous comparison functions (ψ^k is the k-th iterate of ψ) and $\varphi_2\colon \mathbb{R}_+ \to \mathbb{R}_+$, $k = 1, 2, \ldots$, is a monotone increasing function such that $\varphi_2(0) = 0$.

Let $x_0 \in E$ and $x_{n+1} = Tx_n$, $n = 0, 1, \ldots$, be the Jungck-type iteration associated to S and T. Then, the Jungck-type iteration process with $T\colon (Y, d) \to (E, d)$ is (S, T)-stable.

Proof Let $\{Sy_n\}_{n=0}^{\infty} \subset E$, $\epsilon_n = d(Sy_{n+1}, Ty_n)$, $n = 0, 1, \ldots$, and suppose that $\lim_{n\to\infty} \epsilon_n = 0$. Then, we shall establish that $\lim_{n\to\infty} Sy_n = p$, using conditions (i)–(iv) and the triangle inequality: Therefore, using (i), (ii) and triangle inequality lead to

$$\begin{aligned} d(Sy_{n+1}, p) &\leq d(Sy_{n+1}, Ty_n) + d(Ty_n, p) = d(Tz, Ty_n) + \epsilon_n \\ &\leq \varphi_1(\rho(Sz, Tz)) + c\rho(Sz, Sy_n) + \epsilon_n = c\rho(p, Sy_n) + \epsilon_n. \end{aligned} \tag{3.4}$$

Using (iii), we have that $p \in E$. Condition (iv) shows that T has a unique fixed point. Also by condition (iv), we get

$$\begin{aligned} \rho(p, Sy_n) = \rho(Tz, Ty_{n-1}) &\leq \psi(\rho(Sy_{n-1}, p)) \\ \leq \psi^2(\rho(Sy_{n-2}, p)) \leq \cdots \leq \psi^n(\rho(Sy_0, p)) &\to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{3.5}$$

Using (3.5) in (3.4), we have

$$d(Sy_{n+1}, p) \leq c\psi^n(\rho(Sy_0, p)) + \epsilon_n. \tag{3.6}$$

Taking limits of both sides in (3.6) yields

$$\lim_{n\to\infty} d(Sy_{n+1}, p) \leq c \lim_{n\to\infty} \psi^n(\rho(Sy_0, p)) + \lim_{n\to\infty} \epsilon_n = 0$$

That is, $\lim_{n\to\infty} Sy_n = p$.

Conversely, let $\lim_{n\to\infty} Sy_n = p$. Then, by condition (ii) and (3.5) we have

$$\begin{aligned} \epsilon_n = d(Sy_{n+1}, Ty_n) &\leq d(Sy_{n+1}, p) + d(Tz, Ty_n) \\ &\leq d(Sy_{n+1}, p) + c\psi^n(\rho(p, Sy_0)) \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

□

Remark 3.4 Theorem 3.1 and Theorem 3.3 as well as the corollary generalize and extend the well-known stability results in the literature. In particular, see Singh et al [13], Berinde [1], Imoru and Olatinwo [5], Rhoades [10, 11, 12] and some other references in the reference section of this paper for detail. Indeed, Theorem 3.1 and Theorem 3.3 are generalizations and extensions of Theorem 3.1 and Theorem 3.4 of Singh et al [13].

Remark 3.5 To the best of our knowledge, this is the first time the stability of the Picard and Jungck-type iteration processes is being investigated for the case of two metrics.

References

[1] Berinde, V.: *On the stability of some fixed point procedures.* Bul. Stiint. Univ. Baia Mare, Ser. B, Matematica–Informatica **18**, 1 (2002), 7–14.

[2] Berinde, V.: Iterative Approximation of Fixed Points. *Editura Efemeride, Baia Mare, Romania,* 2002.

[3] Berinde, V.: *A priori and a posteriori error estimates for a class of φ-contractions.* Bulletins for Applied Mathematics **90-B** (1999), 183–192.

[4] Harder, A. M., Hicks, T. L.: *Stability results for fixed point iteration procedures.* Math. Japonica **33**, 5 (1988), 693–706.

[5] Imoru, C. O., Olatinwo, M. O.: *On the stability of Picard and Mann iteration processes.* Carpathian J. Math. **19**, 2 (2003), 155–160.

[6] Imoru, C. O., Olatinwo, M. O., Owojori, O. O.: *On the stability of Picard and Mann iteration procedures.* J. Appl. Func. Diff. Eqns. **1**, 1 (2006), 71–80.

[7] Jachymski, J. R.: *An extension of A. Ostrowski's theorem on the round-off stability of iterations.* Aequationes Math. **53** (1997), 242–253.

[8] Osilike, M. O.: *Some stability results for fixed point iteration procedures.* J. Nigerian Math. Soc. Vol. **14/15** (1995), 17–29.

[9] Osilike, M. O., Udomene, A.: *Short proofs of stability results for fixed point iteration procedures for a class of contractive-type mappings.* Indian J. Pure Appl. Math. **30**, 12 (1999), 1229–1234.

[10] Rhoades, B. E.: *Fixed point theorems and stability results for fixed point iteration procedures.* Indian J. Pure Appl. Math. **21**, 1 (1990), 1–9.

[11] Rhoades, B. E.: *Some fixed point iteration procedures.* Internat. J. Math. and Math. Sci. **14**, 1 (1991), 1–16.

[12] Rhoades, B. E.: *Fixed point theorems and stability results for fixed point iteration procedures II.* Indian J. Pure Appl. Math. **24**, 11 (1993), 691–703.

[13] Singh, S. L., Bhatnagar, C., Mishra, S. N.: *Stability of Jungck-type iterative procedures.* Internat. J. Math. & Math. Sc. **19** (2005), 3035–3043.

[14] Zeidler, E.: Nonlinear Functional Analysis and its Applications, Fixed-Point Theorems I. *Springer-Verlag, New York,* 1986.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 93–107

Classes of Filters in Generalizations of Commutative Fuzzy Structures*

Jiří RACHŮNEK [1], Dana ŠALOUNOVÁ [2]

[1] *Department of Algebra and Geometry, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: rachunek@inf.upol.cz

[2] *Department of Mathematical Methods in Economy,*
Faculty of Economy, VŠB–Technical University Ostrava,
Sokolská 33, 701 21 Ostrava, Czech Republic
e-mail: dana.salounova@vsb.cz

(Received August 8, 2008)

Abstract

Bounded commutative residuated lattice ordered monoids ($R\ell$-monoids) are a common generalization of BL-algebras and Heyting algebras, i.e. algebras of basic fuzzy logic and intuitionistic logic, respectively. In the paper we develop the theory of filters of bounded commutative $R\ell$-monoids.

Key words: Residuated ℓ-monoid, deductive system, BL-algebra, MV-algebra, Heyting algebra, filter.

2000 Mathematics Subject Classification: 03G25, 06D35, 06F05

1 Introduction

BL-algebras have been introduced by P. Hájek as an algebraic counterpart of the basic fuzzy logic BL [5]. Omitting the requirement of pre-linearity in the definition of a BL-algebra, one obtains the definition of a bounded commutative residuated lattice ordered monoid ($R\ell$-monoid). Nevertheless, bounded commutative $R\ell$-monoids are a generalization not only of BL-algebras but also of Heyting algebras which are an algebraic counterpart of the intuitionistic propositional logic. Therefore, bounded commutative $R\ell$-monoids could be taken as an algebraic semantics of a more general logic than Hájek's fuzzy logic. It is

*The first author was supported by the Council of Czech Government, MSM 6198959214.

known that every BL-algebra (and consequently every MV-algebra [2], or equivalently, every Wajsberg algebra [4]) is a subdirect product of linearly ordered BL-algebras. Moreover, a bounded commutative $R\ell$-monoid is a subdirect product of linearly ordered $R\ell$-monoids if and only if it is a BL-algebra [13]. On the other side, bounded commutative $R\ell$-monoids which need not be BL-algebras can be constructed from BL-algebras by means of other natural operations, e.g. by means of pasting, i.e. ordinal sums. For example, the pasting of Wajsberg algebras which are not linearly ordered gives bounded commutative $R\ell$-monoids which are not BL-algebras [8, 9].

In both BL-algebras and bounded commutative $R\ell$-monoids, filters coincide with deductive systems of those algebras and are exactly the kernels of their congruences. Various types of filters of BL-algebras were studied in [19], [7] and [11]. Boolean filters of bounded commutative $R\ell$-monoids were investigated in [14].

In this paper we further develop the theory of filters of bounded commutative $R\ell$-monoids and among others, we generalize some results of [7] and [11].

For concepts and results concerning MV-algebras, BL-algebras and Heyting algebras see for instance [2], [5], [1].

2 Preliminaries

A *bounded commutative $R\ell$-monoid* is an algebra $M = (M;\, \odot, \vee, \wedge, \to, 0, 1)$ of type $\langle 2, 2, 2, 2, 0, 0\rangle$ satisfying the following conditions:

- $(R\ell 1)$ $(M; \odot, 1)$ is a commutative monoid.
- $(R\ell 2)$ $(M; \vee, \wedge, 0, 1)$ is a bounded lattice.
- $(R\ell 3)$ $x \odot y \leq z$ if and only if $x \leq y \to z$, for any $x, y, z \in M$.
- $(R\ell 4)$ $x \odot (x \to y) = x \wedge y$, for any $x, y \in M$.

In the sequel, by an *$R\ell$-monoid* we will mean a *bounded commutative $R\ell$-monoid.*

On any $R\ell$-monoid M let us define a unary operation negation $^-$ by $x^- := x \to 0$ for any $x \in M$.

Bounded commutative $R\ell$-monoids are special cases of residuated lattices, more precisely (see for instance [3]), they are exactly commutative integral generalized BL-algebras in the sense of [10].

The above mentioned algebras can be characterized in the class of all $R\ell$-monoids as follows: An $R\ell$-monoid M is

a) a BL-algebra if and only if M satisfies the identity of pre-linearity $(x \to y) \vee (y \to x) = 1$;

b) an MV-algebra if and only if M fulfills the double negation law $x^{--} = x$;

c) a Heyting algebra if and only if the operation "$\odot$" is idempotent.

Lemma 2.1 *See [15] and [16]. In any bounded commutative $R\ell$-monoid M we have for any $x, y, z \in M$:*

(1) $1 \to x = x$.

(2) $x \leq y \iff x \to y = 1$.

(3) $x \odot y \leq x \wedge y$.

(4) $x \leq y \to x$.

(5) $(x \odot y) \to z = x \to (y \to z) = y \to (x \to z)$.

(6) $(x \vee y) \to z = (x \to z) \wedge (y \to z)$.

(7) $x \to (y \wedge z) = (x \to y) \wedge (x \to z)$.

(8) $x \leq x^{--},\ x^{-} = x^{---}$.

(9) $x \leq y \Longrightarrow y^{-} \leq x^{-}$.

(10) $(x \odot y)^{-} = y \to x^{-} = y^{--} \to x^{-} = x \to y^{-} = x^{--} \to y^{-}$.

(11) $x \leq y \Longrightarrow z \to x \leq z \to y,\ y \to z \leq x \to z$.

(12) $x \to y \leq y^{-} \to x^{-}$.

(13) $x \vee y \leq ((x \to y) \to y) \wedge ((y \to x) \to x)$.

(14) $x \to y \leq (y \to z) \to (x \to z)$.

(15) $x \to y \leq (z \to x) \to (z \to y)$.

A non-empty subset F of an $R\ell$-monoid M is called a *filter* of M if

(F1) $x, y \in F$ imply $x \odot y \in F$;

(F2) $x \in F,\ y \in M,\ x \leq y$ imply $y \in F$.

A subset D of an $R\ell$-monoid M is called a *deductive system* of M if

(i) $1 \in D$;

(ii) $x \in D,\ x \to y \in D$ imply $y \in D$.

Proposition 2.2 *[3]. Let H be a non-empty subset of M. Then H is a filter of M if and only if H is a deductive system of M.*

By [18], filters of commutative $R\ell$-monoids are exactly the kernels of their congruences. If F is a filter of M, then F is the kernel of the unique congruence $\Theta(F)$ such that $\langle x, y\rangle \in \Theta(F)$ if and only if $(x \to y) \wedge (y \to x) \in F$, for any $x, y \in M$. Hence we will consider quotient $R\ell$-monoids M/F of $R\ell$-monoids M by their filters F.

A filter F of M is called *maximal* if F is a proper filter of M and is not a proper subset of any proper filter of M.

3 Implicative filters

Let M be an $R\ell$-monoid and F a subset of M. Then F is called an *implicative filter* of M if

(1) $1 \in F$;

(2) $x \to (y \to z) \in F$, $x \to y \in F$ imply $x \to z \in F$.

Proposition 3.1 *Every implicative filter of an $R\ell$-monoid M is a filter of M.*

Proof Let $\emptyset \neq F \subseteq M$ satisfy conditions (1) and (2) and let $x, y \in M$ be such that x, $x \to y \in F$. Then $1 \to (x \to y) \in F$, $1 \to x \in F$, hence $y = 1 \to y \in F$. □

If F is a filter of an $R\ell$-monoid M and $a \in M$, put

$$M_a := \{x \in M : a \to x \in F\}.$$

Theorem 3.2 *Let M be an $R\ell$-monoid and F be a filter of M. Then F is an implicative filter of M if and only if M_a is a filter of M for every $a \in M$.*

Proof Let F be an implicative filter of M and $a \in M$. Then $1 = a \to 1 \in M$, thus $1 \in M_a$. Further, suppose that x, $x \to y \in M_a$, i.e. $a \to x \in F$ and $a \to (x \to y) \in F$. Then we get $a \to y \in F$, and hence $y \in M_a$. That means, M_a is a filter of M for arbitrary $a \in M$.

Conversely, let M_a be a filter of M for each $a \in M$. Suppose that $x \to (y \to z) \in F$ and $x \to y \in F$. Then $y \to z \in M_x$ and $y \in M_x$, hence $z \in M_x$ and therefore $x \to z \in F$. That means, F is implicative. □

Theorem 3.3 *Let F be a filter of an $R\ell$-monoid M. Then the following conditions are equivalent:*

(a) F is an implicative filter of M.

(b) $y \to (y \to x) \in F$ implies $y \to x \in F$, for any $x, y \in M$.

(c) $z \to (y \to x) \in F$ implies $(z \to y) \to (z \to x) \in F$, for any $x, y, z \in M$.

(d) $z \to (y \to (y \to x)) \in F$ and $z \in F$ imply $y \to x \in F$, for any $x, y, z \in M$.

(e) $x \to (x \odot x) \in F$, for any $x \in M$.

Proof (a) ⇒ (b): Suppose that F is an implicative filter of M, $x, y \in M$ and $y \to (y \to x) \in F$. Then since $y \to y = 1 \in F$, we obtain $y \to x \in F$.

(b) ⇒ (c): Let F be a filter of M satisfying the condition (b), $x, y, z \in M$ and $z \to (y \to x) \in F$. Then $z \to (z \to ((z \to y) \to x)) = z \to ((z \to y) \to (z \to x)) \geq z \to (y \to x) \in F$, thus $z \to (z \to ((z \to y) \to x)) \in F$. From this we have $z \to ((z \to y) \to x) \in F$, that means $(z \to y) \to (z \to x) \in F$.

(c) ⇒ (d): Suppose that a filter F satisfies the condition (c). Let $z \to (y \to (y \to x)) \in F$ and $z \in F$. Then also $y \to (y \to x) \in F$. At the same time, $y \to x = (y \to y) \to (y \to x)$, thus $y \to x \in F$.

(d) $\Rightarrow$ (a): Let a filter F fulfill the condition (d). Let $x \to (y \to z) \in F$ and $x \to y \in F$. Then $x \to (y \to z) = y \to (x \to z) \leq (x \to y) \to (x \to (x \to z))$, hence $(x \to y) \to (x \to (x \to z)) \in F$, and therefore $x \to z \in F$.

(a) $\Rightarrow$ (e): Let F be an implicative filter of M. Then $x \to (x \to (x \odot x)) = (x \odot x) \to (x \odot x) = 1 \in F$. Further, $x \to x = 1 \in F$, and hence we obtain $x \to (x \odot x) \in F$.

(e) $\Rightarrow$ (a): Let a filter F satisfy the condition (e) and let $x \to (y \to z) \in F$ and $x \to y \in F$. Then $(x \to (y \to z)) \odot (x \to y) \odot x \odot x \leq (y \to z) \odot y \leq z$, hence $(x \to (y \to z)) \odot (x \to y) \leq (x \odot x) \to z$, and thus $(x \odot x) \to z \in F$. Further, $x \to (x \odot x) \in F$, $(x \odot x) \to x = 1 \in F$, therefore from $(x \odot x) \to z \in F$, we obtain $x \to z \in F$. □

Using the proof (a) $\Rightarrow$ (e) in the preceding theorem, we have as an immediate consequence:

Theorem 3.4 *If F is a filter of an $R\ell$-monoid M, then F is an implicative filter if and only if the quotient $R\ell$-monoid M/F is a Heyting algebra.*

Proposition 3.5 *If F_1 and F_2 are filters of an $R\ell$-monoid M, $F_1 \subseteq F_2$ and F_1 is an implicative filter of M, then F_2 is also an implicative filter of M.*

Proof Suppose that F_1 and F_2 are filters of an $R\ell$-monoid M, $F_1 \subseteq F_2$ and F_1 is implicative. Then, by Theorem 3.3, $x \to x \odot x \in F_1 \subseteq F_2$ for any $x \in M$, and therefore F_2 is also implicative. □

Let M be an $R\ell$-monoid and F a subset of M. Then F is called a *positive implicative filter* of M if

(1) $1 \in F$;

(3) $x \to ((y \to z) \to y) \in F$ and $x \in F$ imply $y \in F$, for any $x, y, z \in M$.

Proposition 3.6 *Every positive implicative filter of an $R\ell$-monoid M is a filter of M.*

Proof Let $x \in F$ and $x \to y \in F$. Then $x \to ((y \to 1) \to y) = x \to (1 \to y) = x \to y$, hence $x \to ((y \to 1) \to y) \in F$, and thus $y \in F$. □

Proposition 3.7 *Every positive implicative filter of M is an implicative filter of M.*

Proof Let F be a positive implicative filter of M, $x, y, z \in M$, $x \to (y \to z) \in F$ and $x \to y \in F$. We have $(x \to y) \to (x \to (x \to z)) \geq y \to (x \to z) = x \to (y \to z)$, hence $(x \to y) \to (x \to (x \to z)) \in F$, and thus also $x \to (x \to z) \in F$.

Since $((x \to z) \to z) \to (x \to z) \geq x \to (x \to z)$, then we get $((x \to z) \to z) \to (x \to z) \in F$. Further, $1 \to (((x \to z) \to z) \to (x \to z)) = ((x \to z) \to z) \to (x \to z)$, and since $1 \to (((x \to z) \to z) \to (x \to z)) \in F$ and $1 \in F$, we obtain $x \to z \in F$.

Therefore F is an implicative filter. □

Theorem 3.8 *Let F be a filter of an $R\ell$-monoid M. Then the following conditions are equivalent:*

(a) F is a positive implicative filter of M.

(b) $(x \to y) \to x \in F$ implies $x \in F$, for any $x, y \in M$.

(c) $(x^- \to x) \to x \in F$, for any $x \in M$.

Proof (a) $\Rightarrow$ (b): Let F be a positive implicative filter of M and $(x \to y) \to x \in F$. Then since $1 \to ((x \to y) \to x) = (x \to y) \to x \in F$ and $1 \in F$, we get $x \in F$.

(b) $\Rightarrow$ (a): Let a filter F satisfy the condition (b) and let $x \to ((y \to z) \to y) \in F$ and $x \in F$. Then $(y \to z) \to y \in F$, and therefore $y \in F$. Hence F is a positive implicative filter of M.

(b) $\Rightarrow$ (c): Let F be a filter of M and $x \in M$. Then $(((x^- \to x) \to x) \to 0) \to ((x^- \to x) \to x) = (x^- \to x) \to ((((x^- \to x) \to x) \to 0) \to x) \geq (((x^- \to x) \to x) \to 0) \to x^- = ((x^- \to x) \to x) \to 0) \to (x \to 0) \geq x \to ((x^- \to x) \to x) = 1 \in F$, thus $(((x^- \to x) \to x) \to 0) \to ((x^- \to x) \to x) \in F$, and hence $(x^- \to x) \to x \in F$.

(c) $\Rightarrow$ (b): Let a filter F satisfy condition (c). Let $(x \to y) \to x \in F$. We have $(x \to y) \to x \leq (x \to 0) \to x = x^- \to x$, hence $x^- \to x \in F$. By the assumption, $(x^- \to x) \to x \in F$, thus $x \in F$. Therefore F satisfies the condition (b). □

Proposition 3.9 *If F_1 and F_2 are filters of an $R\ell$-monoid M, F_1 is a positive implicative filter and $F_1 \subseteq F_2$, then F_2 is also a positive implicative filter of M.*

Proof Let $F_1 \subseteq F_2$ and F_1 be positive implicative. Then for any $x \in M$ we get $(x^- \to x) \to x \in F_1$, thus $(x^- \to x) \to x \in F_2$. Therefore, by Theorem 3.8, F_2 is a positive implicative filter of M. □

Theorem 3.10 *Let M be an $R\ell$-monoid. Then the following conditions are equivalent:*

(a) M is a Heyting algebra.

(b) Every filter of M is implicative.

(c) $\{1\}$ is an implicative filter of M.

Proof (a) $\Rightarrow$ (c): It follows from Theorem 3.4.

(a) $\Rightarrow$ (b): Let M be an idempotent $R\ell$-monoid, F be a filter of M, and $x \in M$. Then $x \to (x \odot x) = x \to x = 1 \in F$, hence by Theorem 3.3, F is an implicative filter.

(b) $\Rightarrow$ (c): It is obvious. □

Proposition 3.11 *Let F be an implicative filter of an Rℓ-monoid M. Then the following conditions are equivalent:*

(a) F is a positive implicative filter of M.

(b) $(x \to y) \to y \in F$ implies $(y \to x) \to x \in F$, for any $x, y \in M$.

Proof (a) $\Rightarrow$ (b): Let F be a positive implicative filter of M and $(x \to y) \to y \in F$. Since $x \leq (y \to x) \to x$, we get $((y \to x) \to x) \to y \leq x \to y$. Hence $(x \to y) \to y \leq (y \to x) \to ((x \to y) \to x) = (x \to y) \to ((y \to x) \to x) \leq (((y \to x) \to x) \to y) \to ((y \to x) \to x)$, and thus $(((y \to x) \to x) \to y) \to ((y \to x) \to x) \in F$. Consequently, also $1 \to ((((y \to x) \to x) \to y) \to ((y \to x) \to x)) \in F$, and since F is a positive implicative filter, we get $(y \to x) \to x \in F$.

(b) $\Rightarrow$ (a): Let an implicative filter F satisfy the condition (b) and let $x \in F$ and $x \to ((y \to z) \to y) \in F$. Then also $(y \to z) \to y \in F$. Further, $(y \to z) \to y \leq (y \to z) \to ((y \to z) \to z)$, hence $(y \to z) \to ((y \to z) \to z) \in F$. Since F is implicative, $(y \to z) \to z \in F$. Then, by the assumption, also $(z \to y) \to y \in F$. Further, $z \leq y \to z$, hence $(y \to z) \to y \leq z \to y$, thus $z \to y \in F$. We have shown $(z \to y) \to y \in F$, therefore $y \in F$. □

Theorem 3.12 *Let M be an Rℓ-monoid. Then the following conditions are equivalent:*

(a) $\{1\}$ is a positive implicative filter.

(b) Every filter of M is positive implicative.

(c) $M(a) := \{x \in M : a \leq x\}$ is a positive implicative filter of M, for every $a \in M$.

(d) $(x \to y) \to x = x$, for any x, $y \in M$.

(e) M is a Boolean algebra.

Proof (a) $\Rightarrow$ (b): It follows from Proposition 3.9.

(b) $\Rightarrow$ (c): Let $a \in M$. Then $1 \in M(a)$. Assume that x, $x \to y \in M(a)$, i.e. $a \to x = 1$, $a \to (x \to y) = 1$. Since by the assumption, $\{1\}$ is a positive implicative filter of M, we obtain $a \to y = 1$, hence $y \in M(a)$. That means $M(a)$ is a filter of M which is also positive implicative.

(c) $\Rightarrow$ (d): If $x, y \in M$, then $(x \to y) \to x \in M((x \to y) \to x)$, therefore $(x \to y) \to x \leq x$ by Theorem 3.8. Moreover, $x \leq (x \to y) \to x$, i.e. $(x \to y) \to x = x$.

(d) $\Rightarrow$ (a): It follows from Theorem 3.8.

(d) $\Rightarrow$ (e): Since $(x \to y) \to x = x$, we obtain $(y \to x) \to x = (y \to x) \to ((x \to y) \to x) \geq (x \to y) \to y$, and similarly, $(x \to y) \to y \geq (y \to x) \to x$. Hence $x^{--} = (x \to 0) \to 0 = (0 \to x) \to x = 1 \to x = x$ and therefore by [12], M is an MV-algebra. Then by [7, Lemma 3.16], furthermore M is a Boolean algebra.

(e) ⇒ (d): Since M is a Boolean algebra, x^- is the lattice complement of x in M, and so $x \vee x^- = 1$. This implies, by [7, Lemma 3.16], $(x \to y) \to x = x$ for any $x, y \in M$. □

Theorem 3.13 *If F is a filter of an $R\ell$-monoid M, then the following conditions are equivalent:*

(a) F is a maximal and positive implicative filter of M.

(b) F is a maximal and implicative filter of M.

(c) If $x, y \in M \setminus F$, then $x \to y \in F$ and $y \to x \in F$.

(d) M/F is a two-element Boolean algebra.

Proof (a) ⇒ (b): It is obvious.

(b) ⇒ (c): Let F be a maximal and implicative filter of M. By Theorem 3.2, $M_y = \{a \in M : y \to a \in F\}$ is a filter of M. If $b \in F$, then from $b \leq y \to b$ it follows that $y \to b \in F$, thus $b \in M_y$. Hence $F \subseteq M_y$. Since F is a maximal filter of M and $y \notin F$, we have $M_y = M$. Therefore $y \to x \in F$. The assumption $x \notin F$ analogously implies $x \to y \in F$.

(c) ⇒ (a): Let a filter F satisfy the condition (c). Suppose that F is not positive implicative. Then by Theorem 3.8, there are $x, y \in M$ such that $x \notin F$ and $(x \to y) \to x \in F$. If $y \in F$, then $x \to y \in F$, and hence $x \in F$, a contradiction. If $y \notin F$, then by (c), $x \to y \in F$, a contradiction. Hence F is a positive implicative filter of M. We will prove that F is also a maximal filter of M. If $a \notin F$, then by the preceding part of the proof, $F \cup \{a\} \subseteq M_a$. We will show that M_a is the least filter of M containing $F \cup \{a\}$. Let G be a filter of M such that $F \cup \{a\} \subseteq G$. If $x \in M_a$, then $a \to x \in F \subseteq G$, and since $a \in G$, we have $x \in G$. Therefore $M_a \subseteq G$. Consider any element $z \in M$. If $z \in F$, then $z \in M_a$. If $z \notin F$, then since also $a \notin F$, the assumption (c) gives $a \to z \in M_a$. Hence $M_a = M$, and therefore F is a maximal filter of M.

(c) ⇒ (d): It is obvious. □

A filter F of an $R\ell$-monoid M is called

a) *Boolean* if $x \vee x^- \in F$ for every $x \in M$;

b) *semi-Boolean* if $(x \wedge x^-)^- \in F$ for every $x \in M$.

Proposition 3.14 *[14, Theorem 3.2]. If F is a filter of an $R\ell$-monoid M, then F is Boolean if and only if M/F is a Boolean algebra.*

Proposition 3.15 *Every Boolean filter of M is semi-Boolean.*

Proof Let $x \in M$. Then $x^- \leq (x \wedge x^-)^-$ and $x \leq x^{--} \leq (x \wedge x^-)^-$, hence $x \vee x^- \leq (x \wedge x^-)^-$. □

Example 3.16 Let $M = \{0, a, b, c, 1\}$ be the lattice with the diagram in Fig. 1, and let $\odot = \wedge$ and $\to$ be defined in the corresponding table in Fig. 1.

$\rightarrow$	0	a	b	c	1
0	1	1	1	1	1
a	b	1	b	1	1
b	a	a	1	1	1
c	0	a	b	1	1
1	0	a	b	c	1

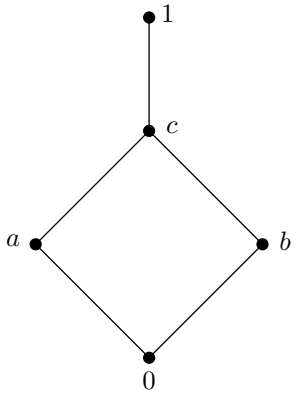


Fig. 1

Then $M = (M; \vee, \wedge, \odot, \rightarrow, 0, 1)$ is an $R\ell$-monoid (which is not a BL-algebra). The filter $F = \{1\}$ is semi-Boolean, but it is not Boolean.

Theorem 3.17 *a) Let M be an $R\ell$-monoid. Then every Boolean filter of M is positive implicative and every positive implicative filter of M is semi-Boolean.*

b) If an $R\ell$-monoid M satisfies condition

$$((x \rightarrow x^-) \rightarrow x^-) \wedge ((x^- \rightarrow x) \rightarrow x) = x \vee x^-, \text{ for any } x \in M, \qquad (*)$$

then Boolean and positive implicative filters of M coincide.

Proof a) Let M be an $R\ell$-monoid, let F be a Boolean filter of M and let $x \in M$. Then by Lemma 2.1, $x \vee x^- \leq ((x \rightarrow x^-) \rightarrow x^-) \wedge ((x^- \rightarrow x) \rightarrow x)$, hence $((x \rightarrow x^-) \rightarrow x^-) \wedge ((x^- \rightarrow x) \rightarrow x) \in F$, and therefore $(x^- \rightarrow x) \rightarrow x \in F$. That means F is positive implicative.

Let now F be an arbitrary positive implicative filter of M and $x \in M$. Then $(x^{--} \rightarrow x^-) \rightarrow x^- \in F$ and by Lemma 2.1, $(x^{--} \rightarrow x^-) \rightarrow x^- = (x \rightarrow x^-) \rightarrow x^- = ((x \rightarrow x^-) \odot x)^- = (x \wedge x^-)^-$. Thus F is a semi-Boolean filter.

b) Let an $R\ell$-monoid M satisfy condition $(*)$ and let F be a positive implicative filter of M. Then a fortiori F is also implicative, hence $x \rightarrow (x \odot x) \in F$ for every $x \in M$. We have $(x \rightarrow x^-) \rightarrow x^- = (x \rightarrow (x \rightarrow 0)) \rightarrow (x \rightarrow 0) = ((x \odot x) \rightarrow 0) \rightarrow (x \rightarrow 0) \geq x \rightarrow (x \odot x)$, hence $(x \rightarrow x^-) \rightarrow x^- \in F$, and thus also $x \vee x^- = ((x \rightarrow x^-) \rightarrow x^-) \wedge ((x^- \rightarrow x) \rightarrow x) \in F$. Therefore F is a Boolean filter. □

As an immediate consequence we get the following theorem.

Theorem 3.18 *[11, Theorem 2]. Boolean and positive implicative filters of any BL-algebra coincide.*

Proof If M is a BL-algebra, then by [5, Lemma 2.3.4(8)], $((x \rightarrow y) \rightarrow y) \wedge ((y \rightarrow x) \rightarrow x) = x \vee y$, for every $x, y \in M$. □

Let F be a filter of an $R\ell$-monoid M. Then F is called an *implicative deductive system* if $x \rightarrow (z^- \rightarrow y) \in F$ and $y \rightarrow z \in F$ imply $x \rightarrow z \in F$, for any x, y, $z \in M$.

Theorem 3.19 *[14, Theorem 3.2]. Let F be a filter of an $R\ell$-monoid M. Then F is an implicative deductive system if and only if F is a Boolean filter.*

Remark 3.20 Now we can rephrase Theorem 3.17 in this way. Let M be an $R\ell$-monoid. Then every implicative deductive system of M is a positive implicative filter and every positive implicative filter of M is semi-Boolean. If M satisfies the condition $(*)$, then implicative deductive systems and positive implicative filters of M coincide.

Theorem 3.21 *If F is a maximal and (positive) implicative filter of an $R\ell$-monoid M, then F is Boolean.*

Proof Let F be a maximal and (positive) implicative filter of M. Then by Theorem 3.13, M/F is a two element $R\ell$-monoid, hence a two element Boolean algebra. Consequently, by Proposition 3.14, F is a Boolean filter. □

Theorem 3.22 *If F is a maximal filter of an $R\ell$-monoid M, then the following conditions are equivalent:*

(a) F is a Boolean filter.

(b) F is a positive implicative filter.

(c) F is an implicative filter.

(d) F is an implicative deductive system.

Proof It follows from Theorems 3.17 and 3.21 and from Remark 3.20. □

Let M be an $R\ell$-monoid. If F is a proper filter of M, denote

$$F^- := \{x \in M : x \leq y^- \text{ for some } y \in F\}.$$

By [14, Proposition 3.4], $F \cup F^-$ is a subalgebra of M for every proper filter F of M.

An $R\ell$-monoid M is called *bipartite* if $M = F \cup F^-$ for some maximal filter F of M.

By [14, Theorem 3.6], M is bipartite if and only if M contains a proper Boolean filter.

An $R\ell$-monoid M is said to be *strongly bipartite* if $M = F \cup F^-$ for every maximal filter F of M.

If M is an $R\ell$-monoid, denote by $B(M)$ the intersection of all Boolean filters of M. Obviously $B(M)$ is the least Boolean filter of M.

Further, denote by $\mathrm{Rad}(M)$ the *radical of* M, i.e. the intersection of all maximal filters of M.

Theorem 3.23 *[14, Theorem 3.8]. If M is an $R\ell$-monoid, then the following conditions are equivalent:*

(a) M is strongly bipartite.

(b) Every maximal filter of M is Boolean.

(c) $B(M) \subseteq \mathrm{Rad}(M)$.

The following theorem is an immediate consequence of Theorems 3.22 and 3.23.

Theorem 3.24 *If M is an $R\ell$-monoid, then the following conditions are equivalent:*

(a) M is strongly bipartite.

(b) $B(M) \subseteq \mathrm{Rad}(M)$.

(c) Every maximal filter of M is Boolean.

(d) Every maximal filter of M is positive implicative.

(e) Every maximal filter of M is implicative.

4 Fantastic filters

Let M be an $R\ell$-monoid and F a subset of M. Then F is called a *fantastic filter* of M if

(1) $1 \in F$;

(4) $z \to (y \to x) \in F$ and $z \in F$ imply $((x \to y) \to y) \to x \in F$, for any $x, y, z \in M$.

Proposition 4.1 *Every fantastic filter of M is a filter of M.*

Proof Let F be a fantastic filter of M and $x, y \in M$. If x, $x \to y \in F$, then also $x \in F$ and $x \to (1 \to y) = x \to y \in F$, and thus by (4), $y \in F$. □

Theorem 4.2 *A filter F of an $R\ell$-monoid M is fantastic if and only if*

(5) *$y \to x \in F$ implies $((x \to y) \to y) \to x \in F$, for every $x, y \in M$.*

Proof Let F be a fantastic filter of M, $x, y \in M$ and $y \to x \in F$. Then $1 \to (y \to x) = y \to x \in F$ and $1 \in F$, hence $((x \to y) \to y) \to x \in F$.

Conversely, let a filter F satisfy the condition (5) and let $z \to (y \to x) \in F$ and $z \in F$. Then $y \to x \in F$, therefore also $((x \to y) \to y) \to x \in F$. □

Theorem 4.3 *Every positive implicative filter of an $R\ell$-monoid M is a fantastic filter of M.*

Proof Suppose F is a positive implicative filter of M and $x, y \in M$ are such that $y \to x \in F$. We have $x \leq ((x \to y) \to y) \to x$, thus

$$(((x \to y) \to y) \to x) \to y \leq x \to y.$$

Further, $((((x \to y) \to y) \to x) \to y) \to (((x \to y) \to y) \to x) \geq (x \to y) \to (((x \to y) \to y) \to x) = ((x \to y) \to y) \to ((x \to y) \to x) \geq y \to x$.

By the assumption $y \to x \in F$, hence also

$$((((x \to y) \to y) \to x) \to y) \to (((x \to y) \to y) \to x) \in F.$$

Since F is positive implicative, we get $((x \to y) \to y) \to x \in F$, and hence F is a fantastic filter. □

Theorem 4.4 *If F is a filter of an $R\ell$-monoid M, then the following conditions are equivalent:*

(a) F is a fantastic filter of M.

(b) $x^{--} \to x \in F$, for every $x \in M$.

(c) $x \to u \in F$ and $y \to u \in F$ imply $((x \to y) \to y) \to u \in F$, for every $x, y, u \in M$.

Proof (a) $\Rightarrow$ (b): Let F be a fantastic filter of M and $x \in M$. Since $0 \to x = 1 \in F$, we obtain from (5) that $x^{--} \to x = ((x \to 0) \to 0) \to x \in F$.

(b) $\Rightarrow$ (c): Suppose that F is a filter of M such that $x^{--} \to x \in F$ for every $x \in M$. Let $x, y, u \in M$, $x \to u \in F$ and $y \to u \in F$. Since $x \to u \leq u^- \to x^-$ and $y \to u \leq u^- \to y^-$, we get $u^- \to x^- \in F$ and $u^- \to y^- \in F$, and thus $(u^- \to x^-) \wedge (u^- \to y^-) \in F$.

Moreover,

$$\begin{aligned}(u^- \to x^-) \wedge (u^- \to y^-) &= u^- \to (x^- \wedge y^-)\\ &= u^- \to (y^- \odot (y^- \to x^-)) = u^- \to (y^- \odot (y^- \to (x \to 0))\\ &= u^- \to (y^- \odot (x \to (y^- \to 0)) = u^- \to (y^- \odot (x \to y^{--})).\end{aligned}$$

Further,

$$\begin{aligned}&(u^- \to (y^- \odot (x \to y^{--}))) \to (u^- \to (y^- \odot (x \to y)))\\ &\quad \geq (y^- \odot (x \to y^{--})) \to (y^- \odot (x \to y)))\\ &\quad \geq (x \to y^{--}) \to (x \to y) \geq y^{--} \to y \in F,\end{aligned}$$

therefore also $u^- \to (y^- \odot (x \to y)) \in F$.

Moreover,

$$u^- \to (y^- \odot (x \to y)) \leq (y^- \odot (x \to y))^- \to u^{--} = ((x \to y) \to y^{--}) \to u^{--},$$

hence $((x \to y) \to y^{--} \to u^{--} \in F$. Further we have

$$\begin{aligned}&(((x \to y) \to y^{--}) \to u^{--}) \to (((x \to y) \to y) \to u^{--})\\ &\quad \geq ((x \to y) \to y) \to ((x \to y) \to y^{--}) \geq y \to y^{--} = 1 \in F,\end{aligned}$$

thus $((x \to y) \to y) \to u^{--} \in F$.

Moreover,

$$(((x \to y) \to y) \to u^{--}) \to (((x \to y) \to y) \to u) \geq u^{--} \to u \in F,$$

therefore also $((x \to y) \to y) \to u \in F$.

(c) $\Rightarrow$ (a): If F satisfies the condition (c), then for $u = x$ we get that whether $y \to x \in F$ then $((x \to y) \to y) \to x \in F$, for every x, $y \in M$, hence F is a fantastic filter of M. $\square$

Theorem 4.5 *If F_1 and F_2 are filters of an $R\ell$-monoid M, $F_1 \subseteq F_2$ and F_1 is fantastic in M, then F_2 is also a fantastic filter of M.*

Proof Let F_1 and F_2 be filters of M, $F_1 \subseteq F_2$, and let F_1 be fantastic. Then by Theorem 4.4, $x^{--} \to x \in F_1 \subseteq F_2$, for every $x \in M$, hence F_2 is also fantastic. □

Theorem 4.6 *A filter F of an $R\ell$-monoid M is fantastic if and only if M/F is an MV-algebra.*

Proof Let F be a filter of M. Then F is fantastic if and only if $x^{--} \to x \in F$ for every $x \in M$, which is equivalent to the following conditions in M/F:

$$x^{--}/F \to x/F = F, \;\; x^{--}/F \leq x/F \;\text{ and }\; x^{--}/F = x/F,$$

for every $x/F \in M/F$, and this is equivalent to M/F is an MV-algebra. □

Proposition 4.7 *If F is a maximal filter of an $R\ell$-monoid M, then F is fantastic.*

Proof It follows from [3, Proposition 3.5], where it is proved that M/F is an MV-algebra for every maximal filter F of M. □

Remark 4.8 The *MV-filters* of $R\ell$-monoids, i.e. filters such that the corresponding quotient $R\ell$-monoids are MV-algebras, were investigated in [16], [17] and [3]. By Theorem 4.6, MV-filters of $R\ell$-monoids are exactly their fantastic filters. If M is an $R\ell$-monoid, denote by $D(M) := \{x \in M \colon x^{--} = 1\}$ the set of all dense elements in M. Then $D(M)$ is a proper filter of M and a filter F of M is an MV-filter if and only if $D(M) \subseteq F$. Therefore we get as a consequence the following proposition.

Proposition 4.9 *A filter F of an $R\ell$-monoid M is fantastic if and only if $D(M) \subseteq F$.*

Proposition 4.10 *Let M be an $R\ell$-monoid. Then the following conditions are equivalent:*

(1) M is an MV-algebra.

(2) Every filter of M is fantastic.

(3) $\{1\}$ is a fantastic filter of M.

Proof (1) ⇒ (2): Let M be an MV-algebra and F be a filter of M. Since the class of MV-algebras is a subvariety of the variety of $R\ell$-monoids, the quotient $R\ell$-monoid M/F is also an MV-algebra. Therefore by Theorem 4.6, F is a fantastic filter.

(2) ⇒ (3): It is obvious.

(3) ⇒ (1): Let $\{1\}$ be a fantastic filter of M. Then $M \cong M/\{1\}$ is an MV-algebra. □

Theorem 4.11 *If F is a filter of an $R\ell$-monoid M, then the following conditions are equivalent.*

(a) F is a Boolean filter.

(b) F is an implicative and fantastic filter.

Proof By Proposition 3.14, a filter F is Boolean if and only if M/F is a Boolean algebra. Moreover, an $R\ell$-monoid M/F is a Boolean algebra if and only if M/F is an MV-algebra and $(x/F) \odot (x/F) = x/F$ for every $x/F \in M/F$. This is equivalent to $(x/F)^{--} = x/F$ and $(x/F) \odot (x/F) = x/F$, and it holds, by Theorems 4.6 and 3.4, if and only if F is a fantastic and implicative filter of M. □

We have characterized filters of $R\ell$-monoids such that the corresponding quotient $R\ell$-monoids are Heyting algebras, Boolean algebras and MV-algebras, respectively. (See e.g. Theorem 3.4, Proposition 3.14 and Theorem 4.6.) Now we will complete it for the case when the quotient $R\ell$-monoid is a BL-algebra.

A filter F of an $R\ell$-monoid M is called a *BL-filter* of M if

$$(x \to y) \vee (y \to x) \in F,$$

for every $x, y \in M$.

Theorem 4.12 *A filter F of an $R\ell$-monoid M is a BL-filter of M if and only if M/F is a BL-algebra.*

Proof We know that an $R\ell$-monoid is a BL-algebra if and only if it satisfies the identity of pre-linearity.

Let M be an $R\ell$-monoid and F be a filter of M. If $x, y \in M$, then

$$(x/F \to y/F) \vee (y/F \to x/F) = ((x \to y) \vee (y \to x))/F.$$

Hence $(x/F \to y/F) \vee (y/F \to x/F) = F$ if and only if $(x \to y) \vee (y \to x) \in F$. □

References

[1] Balbes, R., Dwinger, P.: Distributive Lattices. *Univ. of Missouri Press, Columbia, Missouri,* 1974.

[2] Cignoli, R., D'Ottaviano, I. M. L., Mundici, D.: Algebraic Foundations of Many-valued Reasoning. *Kluwer Acad. Publ., Dordrecht,* 2000.

[3] Dvurečenskij, A., Rachůnek, J.: *Probabilistic averaging in bounded commutative residuated ℓ-monoids.* Discrete Mathematics **306** (2006), 1317–1326.

[4] Font, J. M., Rodriguez, A. J., Torrens, A.: *Wajsberg algebras.* Stochastica **8** (1984), 5–31.

[5] Hájek, P.: Metamathematics of Fuzzy Logic. *Kluwer Acad. Publ., Dordrecht,* 1998.

[6] Hájek, P.: *Basic fuzzy logic and BL-algebras.* Soft Comput. **2** (1998), 124–128.

[7] Haveshki, M., Saeid, A. B., Eslami, E.: *Some types of filters in BL-algebras.* Soft Comput **10** (2006), 657–664.

[8] Iorgulescu, A.: *Classes of BCK algebras – Part I.* Preprint Series of the Institute of Mathematics of the Romanian Academy, preprint nr. 1/2004, 1–33.

[9] Iorgulescu, A.: *Classes of BCK algebras – Part III.* Preprint Series of the Institute of Mathematics of the Romanian Academy, preprint nr. 3/2004, 1–37.

[10] Jipsen, P., Tsinakis, C.: *A survey of residuated lattices.* In: J. Martinez (ed.): Ordered algebraic structures. *Kluwer Acad. Publ. Dordrecht,* 2002, 19–56.

[11] Kondo, M., Dudek, W. A.: *Filter theory of BL-algebras.* Soft Comput. **12** (2008), 419–423.

[12] Rachůnek, J.: *MV-algebras are categorically equivalent to a class of $DR\ell_{1(i)}$-semigroups.* Math. Bohemica **123** (1998), 437–441.

[13] Rachůnek, J.: *A duality between algebras of basic logic and bounded representable $DR\ell$-monoids.* Math. Bohemica **126** (2001), 561–569.

[14] Rachůnek, J., Šalounová, D.: *Boolean deductive systems of bounded commutative residuated ℓ-monoids.* Contrib. Gen. Algebra **16** (2005), 199–208.

[15] Rachůnek, J., Šalounová, D.: *Local bounded commutative residuated ℓ-monoids.* Czechoslovak Math. J. **57** (2007), 395–406.

[16] Rachůnek, J., Slezák, V.: *Negation in bounded commutative $DR\ell$-monoids.* Czechoslovak Math. J. **56** (2006), 755–763.

[17] Rachůnek, J., Slezák, V.: *Bounded dually residuated lattice ordered monoids as a generalization of fuzzy structures.* Math. Slovaca **56** (2006), 223–233.

[18] Swamy, K. L. N.: *Dually residuated lattice ordered semigroups III.* Math. Ann. **167** (1966), 71–74.

[19] Turunen, E.: *Boolean deductive systems of BL-algebras.* Arch. Math. Logic **40** (2001), 467-473.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 109–128

Singular Problems on the Half-line*

Irena RACHŮNKOVÁ [1], Jan TOMEČEK [2]

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
[1] *e-mail: rachunko@inf.upol.cz*
[2] *e-mail: jan_tomecek@seznam.cz*

(Received November 24, 2008)

Abstract

The paper investigates singular nonlinear problems arising in hydrodynamics. In particular, it deals with the problem on the half–line of the form

$$(p(t)u'(t))' = p(t)f(u(t)),$$
$$u'(0) = 0, \quad u(\infty) = L.$$

The existence of a strictly increasing solution (a homoclinic solution) of this problem is proved by the dynamical systems approach and the lower and upper functions method.

Key words: Singular ordinary differential equation of the second order, lower and upper functions, time singularities, unbounded domain, homoclinic solution.

2000 Mathematics Subject Classification: 34B16, 34B40

1 Introduction

In the Cahn–Hilliard theory used in hydrodynamics to study the behaviour of nonhomogenous fluids the following system of PDE's was derived

$$\rho_t + \operatorname{div}(\rho v) = 0, \qquad \frac{dv}{dt} + \nabla(\mu(\rho) - \gamma\triangle\rho) = 0$$

*Supported by the Council of Czech Government MSM 6 198 959 214.

with the density ρ and the velocity v of the fluid, μ is its chemical potential, γ is a constant. In the simplest model, this system can be reduced into the boundary value problem for the ODE of the second order (see [5] or [7])

$$(t^k u')' = 4\lambda^2 t^k (u+1)u(u-\xi), \quad t \in (0,\infty),$$
$$u'(0) = 0, \quad u(\infty) = \xi,$$

where $k \in \mathbb{N}$, $\xi \in (0,1)$, $\lambda \in (0,\infty)$ are parameters. The function $u(t) \equiv \xi$ is a solution of this problem and it corresponds to the case of homogenous fluid (without bubbles). But only the existence of a strictly increasing solution of this problem and the solution itself has a great physical significance. We call it a homoclinic solution. We refer to [1] and [2], where an equivalent problem was investigated. The numerical treatment was done in papers [5], [7].

Here, we study the generalized problem

$$(p(t)u'(t))' = p(t)f(u(t)), \tag{1}$$

$$u'(0) = 0, \quad u(\infty) = L, \tag{2}$$

where $L > 0$.

2 Autonomous equation

The investigation of autonomous equations corresponding to (1) turned out to be quite useful, because some solutions of the perturbed autonomous equation (14) can serve as upper functions to (1).

Let $h\colon \mathbb{R} \to \mathbb{R}$ and x_1, x_2, $x_3 \in \mathbb{R}$ be such that $x_1 < x_2 < x_3$ and

$$h \text{ is lipschitzian on } [x_1, x_3], \tag{3}$$

$$h(x_i) = 0 \qquad \text{for } i = 1,2,3, \tag{4}$$

$$\left.\begin{array}{l}\text{there exists } \delta > 0 \text{ such that } h \in C^1((x_2-\delta, x_2)) \\ \text{and } \lim_{x\to x_2^-} h'(x) = h'_-(x_2) < 0,\end{array}\right\} \tag{5}$$

$$(x-x_2)h(x) < 0 \quad \text{for } x \in (x_1,x_3)\setminus\{x_2\}, \tag{6}$$

$$H(x_1) > H(x_3), \tag{7}$$

where

$$H(x) = -\int_{x_2}^{x} h(z)\,\mathrm{d}z \quad \text{for } x \in \mathbb{R}.$$

Moreover we will assume that

$$\begin{cases} h(x) = 0 & \text{for } x \le x_1, \\ h(x) = x - x_3 & \text{for } x \ge x_3. \end{cases} \tag{8}$$

Let us consider equation

$$u'' = h(u) \tag{9}$$

and the initial condition

$$u(0) = B,\ u'(0) = 0 \tag{10}$$

for $B \in (x_1, x_2)$. Equation (9) is equivalent with the gradient system

$$u_1' = u_2,\ u_2' = h(u_1). \tag{11}$$

An energy function of the system (11) has the form

$$E(u_1, u_2) = \frac{u_2^2}{2} + H(u_1), \quad u_1, u_2 \in \mathbb{R}.$$

Lemma 1 *Let (3), (4), (6), (7) be satisfied. The function H has following properties*

1. *$H(x) > 0$ for $x \in [x_1, x_2) \cup (x_2, x_3]$,*
2. *H is decreasing on (x_1, x_2) and increasing on (x_2, x_3),*
3. *there exists unique $\bar{B} \in (x_1, x_2)$ such that*

$$H(\bar{B}) = H(x_3),$$

4. *if (8) is satisfied, then*

$$\begin{cases} H(x) = H(x_1) & \text{for } x \le x_1, \\ H(x) = H(x_3) - (x - x_3)^2/2 & \text{for } x \ge x_3. \end{cases}$$

Proof The first two properties follow from the definition of H and (6). The third property is a consequence of (6) and (7). The fourth one can be obtained by simple computation. □

Lemma 2 *Let (3), (4), (6)–(8) be satisfied. Let (v_1, v_2) be a solution of problem (11),*

$$u_1(0) = B, u_2(0) = 0, \tag{12}$$

where $B \in (x_1, \bar{B})$, $\bar{B}$ is from Lemma 1. Then there exists $b > 0$ such that

$$v_1(b) = x_3$$

and

$$0 < v_2(t) \le \sqrt{2H(x_1)}$$

for $t \in (0, b]$.

Proof It is well known that the level sets of the energy function E consist of the orbits of the second-order conservative system (11), in particular, the orbit $\gamma((B, 0))$ of system (11) passing the point $(B, 0)$ in the phase plane is a subset of

$$\{(u_1, u_2) \in \mathbb{R}^2 \colon u_2 = \pm\sqrt{2(H(B) - H(u_1))} \wedge H(u_1) \le H(B)\}.$$

From the properties of the function H, we can see that this set can be expressed in the form

$$\{(u_1, u_2) \in \mathbb{R}^2 \colon u_2 = \pm\sqrt{2(H(B) - H(u_1))} \wedge u_1 \geq B\}.$$

This set contains no equilibrium point and hence it is the orbit $\gamma((B, 0))$. Consider the function

$$u_2 = \Phi(u_1) = \sqrt{2(H(B) - H(u_1))} \qquad \text{for } u_1 \geq B.$$

Simple computation yields

$$0 < \Phi(u_1) \leq \Phi(x_2) \qquad \text{for } u_1 \in (B, x_3].$$

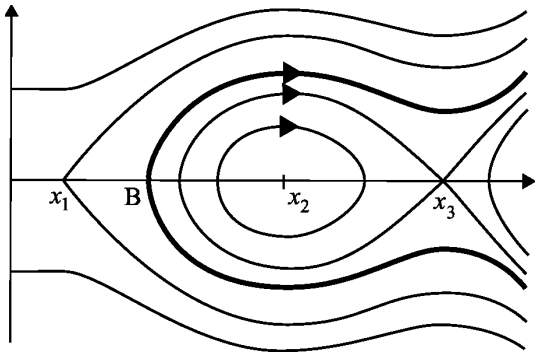


Fig. 1. The escape orbit.

Therefore the orbit $\gamma((B, 0))$ belonging to the solution (v_1, v_2) of (11), (12) has the form on the Figure 1. The direction of the flow on $\gamma((B, 0))$ is determined by the equalities

$$v_1'(0) = v_2(0) = 0 \quad \text{and} \quad v_2'(0) = h(v_1(0)) > 0,$$

see Fig. 1.

Hence there exists $b > 0$ such that

$$(v_1(b), v_2(b)) = (x_3, \Phi(x_3)) = (x_3, \sqrt{2(H(B) - H(x_3))})$$

and

$$0 < v_2(t) \leq \Phi(x_2) \leq \sqrt{2H(x_1)} \quad \text{for } t \in (0, b].$$

The proof is complete. □

As an immediate consequence of Lemma 2 we get Lemma 3.

Lemma 3 *(On escape solution) Let (3), (4), (6)–(8) be satisfied and u be a solution of problem (9), (10) with $B \in (x_1, \bar{B})$. Then there exists $b > 0$ such that*

$$u(b) = x_3, \quad u'(t) > 0 \quad \textit{for } t \in (0, b]. \tag{13}$$

Choose $\epsilon > 0$ and consider the perturbed equation

$$u'' = h(u) - \epsilon. \tag{14}$$

Lemma 4 *(On the perturbed equation) Let (3)–(8) be satisfied. There exists $\epsilon_0 > 0$ such that for $\epsilon \in (0, \epsilon_0)$ the function $h - \epsilon$ has roots $x_i(\epsilon)$ for $i = 1, 2, 3$, such that*

$$h - \epsilon \text{ is lipschitzian on } [x_1(\epsilon), x_3(\epsilon)], \tag{15}$$

$$h(x_i(\epsilon)) = \epsilon \quad \text{for } i = 1, 2, 3, \tag{16}$$

$$\left.\begin{array}{l}\text{there exists } \delta > 0 \text{ such that } h - \epsilon \in C^1((x_2(\epsilon) - \delta, x_2(\epsilon))) \\ \text{and } \lim_{x \to x_2(\epsilon)^-} (h - \epsilon)'(x) = (h - \epsilon)'_-(x_2(\epsilon)) < 0,\end{array}\right\} \tag{17}$$

$$(x - x_2(\epsilon))(h(x) - \epsilon) < 0 \quad \text{for } x \in (x_1(\epsilon), x_3(\epsilon)) \setminus \{x_2(\epsilon)\}, \tag{18}$$

$$H_\epsilon(x_1(\epsilon)) > H_\epsilon(x_3(\epsilon)), \tag{19}$$

where

$$H_\epsilon(x) = -\int_{x_2(\epsilon)}^{x} (h(z) - \epsilon)\,\mathrm{d}z$$

for $x \in \mathbb{R}$.

Proof From (4), (5), (6) and the Implicit function theorem, it follows that there exists $\bar{\epsilon}_0 > 0$ and a continuous function $x_2 \colon [0, \bar{\epsilon}_0) \to (x_1, x_2]$ such that

$$h(x_2(\epsilon)) = \epsilon \text{ for } \epsilon \in [0, \bar{\epsilon}_0),\ x_2(\epsilon) \text{ is decreasing, } x_2(0) = x_2. \tag{20}$$

We define

$$x_1(\epsilon) = \sup\{x \in [x_1, x_2(\bar{\epsilon}_0)] \colon h(x) \le \epsilon\} \tag{21}$$

for $\epsilon \in (0, \bar{\epsilon}_0)$. From the continuity of the function h, the definition of $x_1(\epsilon)$ and the supremum it follows that

$$x_1(\epsilon) \in [x_1, x_2(\bar{\epsilon}_0)) \qquad \text{for } \epsilon \in (0, \bar{\epsilon}_0)$$

and

$$h(x_1(\epsilon)) = \epsilon, \quad \epsilon \in (0, \bar{\epsilon}_0). \tag{22}$$

We will prove that

$$\lim_{\epsilon \to 0+} x_1(\epsilon) = x_1, \tag{23}$$

by contradiction. If (23) does not hold, then there exists a decreasing sequence $\{\epsilon_n\}$, $\epsilon_n \to 0$ such that $x_1(\epsilon_n) \to \bar{x}_1 \in (x_1, x_2(\bar{\epsilon}_0)]$ as $n \to \infty$. From (20) it follows that

$$h(x_1(\epsilon_n)) = \epsilon_n \to 0.$$

From the continuity of h and (4), (6), we get a contradiction.

We put

$$x_3(\epsilon) = x_3 + \epsilon, \quad \epsilon \in (0, \bar{\epsilon}_0).$$

Then, for $\epsilon \in (0, \bar{\epsilon}_0)$, relations (15)–(18) are satisfied. For $\epsilon \in (0, \bar{\epsilon}_0)$, $x \in [x_1, x_1 + \epsilon_0]$ it is valid

$$H_\epsilon(x) = -\int_{x_2(\epsilon)}^{x} (h(z) - \epsilon)\,\mathrm{d}z = -\int_{x_2(\epsilon)}^{x} h(z)\,\mathrm{d}z + \epsilon(x - x_2(\epsilon))$$
$$= H(x) + \int_{x_2}^{x_2(\epsilon)} h(z)\,\mathrm{d}z + \epsilon(x - x_2(\epsilon)).$$

Then

$$|H_\epsilon(x) - H(x)| \le |x_2(\epsilon) - x_2| \max\{|h(z)|\colon z \in [x_1, x_3 + \bar{\epsilon}_0]\} + \epsilon|x_3 + \bar{\epsilon}_0 - x_1|$$

for $\epsilon \in (0, \bar{\epsilon}_0)$ and $x \in [x_1, x_3 + \bar{\epsilon}_0]$. Since the terms on the right-hand side of the inequality converges to zero as $\epsilon \to 0+$ independently on x, we can write

$$H_\epsilon(x) \rightrightarrows H(x) \quad \text{on } [x_1, x_3 + \bar{\epsilon}_0] \text{ as } \epsilon \to 0+ .$$

From this fact and the relations

$$\lim_{\epsilon \to 0+} x_i(\epsilon) = x_i \quad \text{for } i = 1, 3,$$

it follows that

$$\lim_{\epsilon \to 0+} H_\epsilon(x_i(\epsilon)) = H(x_i) \quad \text{for } i = 1, 3.$$

From these facts and (7) it follows that there exists $\epsilon_0 \in (0, \bar{\epsilon}_0)$ such that (19) is valid for $\epsilon \in (0, \epsilon_0)$, together with (15)–(18), as well. □

Lemma 5 *Let (3)–(8) be satisfied. Let $\epsilon \in (0, \epsilon_0)$, where ϵ_0 is from Lemma 4. Then there exist $B \in (x_1, x_2)$ and $b > 0$ such that the corresponding solution u of problem (14), (10) satisfies (13) and*

$$0 \le u'(t) \le \sqrt{2H(x_1)} \quad \textit{for } t \in [0, b]. \tag{24}$$

Proof Let ϵ_0 be from Lemma 4 and $\epsilon \in (0, \epsilon_0)$ be arbitrary. Then relations (15)–(19) hold. From Lemma 1 (with H_ϵ in place of H) it follows that there exists the unique $\bar{B}(\epsilon) \in (x_1(\epsilon), x_2(\epsilon))$ such that $H_\epsilon(\bar{B}(\epsilon)) = H_\epsilon(x_3(\epsilon))$. Let $B(\epsilon) \in (x_1(\epsilon), \bar{B}(\epsilon))$ and u be the solution of problem (14), (10) with $B = B(\epsilon)$. According to Lemma 3 there exists $b(\epsilon) > 0$ such that

$$u(b(\epsilon)) = x_3(\epsilon) \quad \text{and} \quad u' > 0 \quad \text{on } (0, b(\epsilon)]. \tag{25}$$

In particular, $u(t) \in (x_1(\epsilon), x_3(\epsilon)]$ for every $t \in [0, b(\epsilon)]$. Multiplying the perturbed equation (14) by u' and integrating it over interval $(0, t)$ for $t \in [0, b(\epsilon)]$, we get

$$\frac{u'^2(t)}{2} - \frac{u'^2(0)}{2} = -H_\epsilon(u(t)) + H_\epsilon(u(0)),$$

that is

$$u'(t) = \sqrt{2(H_\epsilon(B(\epsilon)) - H_\epsilon(u(t)))}$$

for $t \in [0, b(\epsilon)]$. Since $H_\epsilon(x_1(\epsilon))$ is the maximum of the function H_ϵ in $[x_1(\epsilon), x_3(\epsilon)]$ and H_ϵ is nonnegative, we get

$$u'(t) \le \sqrt{2H_\epsilon(x_1(\epsilon))}$$

for $t \in [0, b(\epsilon)]$. In view of the fact

$$H_\epsilon(x_1(\epsilon)) = \int_{x_1(\epsilon)}^{x_2(\epsilon)} (h(z) - \epsilon)\,\mathrm{d}z \le \int_{x_1(\epsilon)}^{x_2(\epsilon)} h(z)\,\mathrm{d}z \le \int_{x_1}^{x_2} h(z)\,\mathrm{d}z = H(x_1)$$

and (25), it follows that

$$0 \le u'(t) \le \sqrt{2H(x_1)}$$

for $t \in [0, b(\epsilon)]$. By $B(\epsilon) < x_3 < x_3(\epsilon)$ and (25), there exists $b \in (0, b(\epsilon))$ such that (13) and (24) are valid. □

3 Nonautonomous equation

Let us consider equation (1), where

$$f \text{ is locally lipschitzian on } \mathbb{R}, \tag{26}$$

$$\text{there exist } L_0 < 0 < L \text{ such that } f(L_0) = f(0) = f(L) = 0, \tag{27}$$

$$\left.\begin{array}{l}\text{there exists } \delta > 0 \text{ such that } f \in C^1((-\delta, 0)) \\ \text{and } \lim_{x\to 0^-} f'(x) = f'_-(0) < 0,\end{array}\right\} \tag{28}$$

$$xf(x) < 0 \qquad \text{for } x \in (L_0, L) \setminus \{0\}, \tag{29}$$

$$F(L_0) > F(L), \tag{30}$$

where

$$F(x) = -\int_0^x f(z)\,\mathrm{d}z, \qquad x \in \mathbb{R}.$$

Further we assume that

$$p \in C^2((0,\infty)) \cap C([0,\infty)), \tag{31}$$

$$p(0) = 0, \quad p'(t) > 0 \quad \text{for } t \in (0,\infty), \tag{32}$$

$$\lim_{t\to\infty} \frac{p'(t)}{p(t)} = 0, \tag{33}$$

$$\lim_{t\to\infty} \frac{p''(t)}{p(t)} = 0. \tag{34}$$

Moreover, in some lemmas, we will assume that

$$f(x) = 0 \qquad \text{for } x \in (-\infty, L_0] \cup [L, \infty). \tag{35}$$

If (35) is valid, then

$$\begin{cases} F(x) = F(L_0) \text{ for } x \le L_0, \\ F(x) = F(L) \quad \text{for } x \ge L. \end{cases}$$

The following classical result for non–singular initial value problems will be useful in the proofs.

Lemma 6 *Let (26), (31), (32), (35) be satisfied, $a > 0$, $B_0, B_1 \in \mathbb{R}$. Then there exists the unique solution on $[a, \infty)$ of the initial value problem (1),*

$$u(a) = B_0, \ u'(a) = B_1. \tag{36}$$

Proof It is well known that the problem (1), (36) is equivalent to the IVP

$$\begin{cases} u_1' = \frac{u_2}{p(t)}, \ u_2' = p(t)f(u_1), \\ u_1(a) = B_0, \ u_2(a) = B_1. \end{cases}$$

From (26), (31), (32) it follows the unique solvability of this problem and of the problem (1), (36), as well. □

We will study the singular initial value problem (1),

$$u(0) = B, \quad u'(0) = 0 \tag{37}$$

with $B \in (L_0, 0)$.

Definition 7 Let $[a, c) \subset [0, \infty)$. A function $u \in C^1([a, c)) \cap C^2((a, c))$ satisfying equation (1) on $[a, c)$ and fulfilling conditions (37) is a solution of problem (1), (37) on $[a, c)$.

First we state several lemmas.

Lemma 8 *Let us assume that (26)–(29), (31)–(34) be satisfied. Let u be a solution of the initial value problem (1),*

$$u(a) = B, \quad u'(a) = 0 \tag{38}$$

on $[a, \infty)$, where $a \geq 0$ and $B \in (L_0, 0)$. Then there exists $\theta > a$ such that

$$u(\theta) = 0 \quad \textit{and} \quad u'(t) > 0 \textit{ for } t \in (a, \theta]. \tag{39}$$

Moreover, for every $b > \theta$ satisfying

$$u(b) \in (0, L) \quad \textit{and} \quad u'(t) > 0 \textit{ for } t \in [\theta, b), \tag{40}$$

there exist $\alpha \in (a, \theta)$, $\beta \in (\theta, b)$ such that

$$p^2(b)u'^2(b) = 2[p^2(\alpha)F(B) - p^2(\beta)F(u(b))]. \tag{41}$$

Proof Let u be a solution of problem (1), (38), where $a \geq 0$ and $B \in (L_0, 0)$. From (1) and (29) it follows that there exists $\xi > a$ such that $u(t) \in (L_0, 0)$ and $u'(t) > 0$ for $t \in (a, \xi)$. Let us assume that $\xi = \infty$. Then there exists $l \in (B, 0]$ such that

$$\lim_{t \to \infty} u(t) = l. \tag{42}$$

From (1) and (38), it follows that

$$\frac{u'^2(t)}{2} + \int_a^t \frac{p'(s)}{p(s)} u'^2(s) \, \mathrm{d}s = F(B) - F(u(t)). \tag{43}$$

Since the right–hand side of the equation (43) has a finite nonnegative limit $F(B) - F(l)$ as $t \to \infty$ and the function $\int_a^t \frac{p'(s)}{p(s)} u'^2(s)\,\mathrm{d}s$ is positive and monotone, it follows that there exists finite nonnegative limit $\lim_{t\to\infty} u'^2(t)/2$. Since $u' > 0$ on $(0,\infty)$, there exists nonnegative $\lim_{t\to\infty} u'(t)$. If $\lim\limits_{t\to\infty} u'(t) > 0$, then $\lim_{t\to\infty} u(t) = \infty$, which contradicts (42). Consequently,

$$\lim_{t\to\infty} u'(t) = 0. \tag{44}$$

From (1) it follows that

$$u''(t) = -\frac{p'(t)}{p(t)} u'(t) + f(u(t)) \quad \text{for } t \in (0,\infty).$$

This, together with (42), (44), (26) and (33) implies

$$\lim_{t\to\infty} u''(t) = f(l).$$

Using (44), (27) and (29) we can check that $l = 0$.

We define a function

$$v(t) = \sqrt{p(t)}u(t) \quad \text{for } t \in [0,\infty).$$

By virtue of (31) and (32) we see that v is well defined, negative and there exist finite derivatives

$$v'(t) = \frac{p'(t)u(t)}{2\sqrt{p(t)}} + \sqrt{p(t)}u'(t)$$

and

$$v''(t) = v(t)\left[\frac{1}{2}\frac{p''(t)}{p(t)} - \frac{1}{4}\left(\frac{p'(t)}{p(t)}\right)^2 + \frac{f(u(t))}{u(t)}\right]$$

for $t > a$. In view of (33), (34), from the fact that $\lim_{t\to\infty} u(t) = 0$, u is negative and from (28), it follows that there exist $\omega > 0$ and $R > 0$ such that

$$\frac{1}{2}\frac{p''(t)}{p(t)} - \frac{1}{4}\left(\frac{p'(t)}{p(t)}\right)^2 + \frac{f(u(t))}{u(t)} < -\omega \quad \text{for } t \geq R.$$

Then

$$v''(t) > -\omega v(t) > 0 \quad \text{for } t \geq R. \tag{45}$$

Thus, v' is increasing on $[R,\infty)$ and has the limit

$$\lim_{t\to\infty} v'(t) = V.$$

If $V > 0$, then $\lim_{t\to\infty} v(t) = +\infty$, which contradicts the negativity of v. If $V \leq 0$, then $v'(t) < 0$ for every $t \in (R,\infty)$ and therefore

$$0 > v(R) \geq v(t) \quad \text{for } t \geq R.$$

In view of (45) we can see that

$$0 < -\omega v(R) \le -\omega v(t) < v''(t) \quad \text{for } t \ge R.$$

We get $\lim_{t\to\infty} v'(t) = \infty$, which implies $\lim_{t\to\infty} v(t) = \infty$, again. These contradictions imply the existence of $\theta > a$ such that $u(\theta) = 0$ and $u'(t) > 0$ for $t \in (a, \theta)$. Let us assume that $u'(\theta) = 0$. Since $u(\theta) = 0$ we get from Lemma 6, (1) and (27) that $u(t) = 0$ for $t \in (0, \infty)$, which is a contradiction. Thus (39) holds.

Let us consider $b > \theta$ such that (40) is satisfied. Multiplying equation (1) by pu' and integrating it over (a, θ) and (θ, b) we get

$$(pu')^2(\theta) - (pu')^2(a) = 2\int_a^\theta p^2(s) f(u(s)) u'(s)\,\mathrm{d}s,$$

$$(pu')^2(b) - (pu')^2(\theta) = 2\int_\theta^b p^2(s) f(u(s)) u'(s)\,\mathrm{d}s.$$

Using the Mean value theorem, we get $\alpha \in (a, \theta)$ and $\beta \in (\theta, b)$ such that

$$(pu')^2(\theta) = 2p^2(\alpha)\int_a^\theta f(u(s))u'(s)\,\mathrm{d}s,$$

$$(pu')^2(b) - (pu')^2(\theta) = 2p^2(\beta)\int_\theta^b f(u(s))u'(s)\,\mathrm{d}s$$

and substituing $\tau = u(s)$ we get

$$(pu')^2(\theta) = 2p^2(\alpha)(F(u(a)) - F(u(\theta))),$$

$$(pu')^2(b) - (pu')^2(\theta) = 2p^2(\beta)(F(u(\theta)) - F(u(b))).$$

From these two equations, using the fact that $F(u(\theta)) = 0$, we have (41). □

Lemma 9 *Let us assume that (26)–(34) be satisfied. Let u be a solution of the initial value problem (1), (37) on $[0, \infty)$ and let $b > 0$, $\bar{L} \in (0, L)$ be such that*

$$u(b) = \bar{L}, \quad u'(b) = 0. \tag{46}$$

Then there exists $\theta > b$ such that

$$u(\theta) = 0 \quad \textit{and} \quad u'(t) < 0 \qquad \textit{for } t \in (b, \theta]. \tag{47}$$

Moreover, for every $c > \theta$ satisfying

$$u(c) \in (L_0, 0) \quad \textit{and} \quad u'(t) < 0 \quad \textit{for } t \in (\theta, c), \tag{48}$$

there exist $\alpha \in (b, \theta)$ and $\beta \in (\theta, c)$ such that

$$(pu')^2(c) = 2[p^2(\alpha)F(\bar{L}) - p^2(\beta)F(u(c))]. \tag{49}$$

Proof First of all we will prove the existence of θ satisfying (47). By (29) and (46) there exists $b_1 > b$ such that $f(u(t)) < 0$ for $t \in (b, b_1)$. Thus $p(t)u'(t)$ and $u'(t)$ are decreasing and negative on (b, b_1) and $u(t)$ is decreasing and positive on (b, b_1). Assume that $\theta > b$ satisfying (47) does not exist. Then $b_1 = \infty$ and $\lim_{t\to\infty} u(t) \in [0, \bar{L})$. On the other hand, $\lim_{t\to\infty} u'(t) < 0$, which gives $\lim_{t\to\infty} u(t) = -\infty$.
Let us consider $c > \theta$ such that (48) is satisfied. Multiplying equation (1) by pu' and integrating it over (b, θ) and (θ, c) we get $\alpha \in (b, \theta)$ and $\beta \in (\theta, c)$ such that

$$(pu')^2(\theta) - (pu')^2(b) = 2p^2(\alpha)(F(u(b)) - F(u(\theta))),$$

$$(pu')^2(c) - (pu')^2(\theta) = 2p^2(\beta)(F(u(\theta)) - F(u(c))).$$

From these two equations we get (49). □

Lemma 10 *(On three types of solutions) Let (26)–(35) be satisfied, $B \in (L_0, 0)$. Then there exists a unique solution u of problem (1), (37) and it is defined on $[0, \infty)$. There are just three types of solutions:*

- an escape solution *if there exists $b > 0$ such that $u(b) = L$ and $u' > 0$ on $(0, b]$,*
- a homoclinic solution *if $u' > 0$ on $(0, \infty)$ and $\lim_{t\to\infty} u(t) = L$,*
- an oscillatory solution *if u has infinitely many roots and $u(t) \in (B, L)$ for $t \in (0, \infty)$.*

Moreover, for $t \in (0, \infty)$ it is valid

$$|u'(t)| \le \max_{L_0 \le x \le L} |f(x)| \cdot t, \quad |u(t)| \le L_0 + \max_{L_0 \le x \le L} |f(x)| \cdot \frac{t^2}{2}.$$

Proof Step 1. (On the existence of a solution on some neighbourhood of $t = 0$) From (26) and (35) it follows that there exists $\bar{L} > 0$ such that

$$|f(x_1) - f(x_2)| \le \bar{L}|x_1 - x_2| \tag{50}$$

for $x_1, x_2 \in \mathbb{R}$. Let us take $\eta > 0$ such that

$$\frac{\bar{L}\eta^2}{2} < 1. \tag{51}$$

Consider the Banach space $C([0, \eta])$ with the maximum norm $\|\cdot\|_\infty$ and using (32), define an operator $\mathcal{F}\colon C([0, \eta]) \to C([0, \eta])$

$$(\mathcal{F}u)(t) = B + \int_0^t \frac{1}{p(s)} \int_0^s p(\tau) f(u(\tau))\, \mathrm{d}\tau\, \mathrm{d}s.$$

From (50), (32) it follows that for u_1, $u_2 \in C([0,\eta])$, $t \in [0,\eta]$

$$\begin{aligned}|(\mathcal{F}u_1)(t) - (\mathcal{F}u_2)(t)| &\le \left| \int_0^t \frac{1}{p(s)} \int_0^s p(\tau)(f(u_1(\tau)) - f(u_2(\tau)))\,\mathrm{d}\tau\,\mathrm{d}s \right| \\ &\le \bar{L}\|u_1 - u_2\|_\infty \int_0^t \frac{1}{p(s)} \int_0^s p(\tau)\,\mathrm{d}\tau\,\mathrm{d}s \\ &\le \bar{L}\|u_1 - u_2\|_\infty \int_0^t \int_0^s \mathrm{d}\tau\,\mathrm{d}s \le \frac{\bar{L}\eta^2}{2}\|u_1 - u_2\|_\infty.\end{aligned}$$

Inequality (51) implies that $\mathcal{F}$ is a contraction. From the Banach fixed point theorem it follows that there exists a unique fixed point u of the operator $\mathcal{F}$. Then

$$u(t) = B + \int_0^t \frac{1}{p(s)} \int_0^s p(\tau) f(u(\tau))\,\mathrm{d}\tau\,\mathrm{d}s \quad \text{for } t \in [0,\eta].$$

We have $u(0) = B$ and deriving the equality we get

$$u'(t) = \frac{1}{p(t)} \int_0^t p(s) f(u(s))\,\mathrm{d}s, \quad \text{for } t \in (0,\eta). \tag{52}$$

From (52), (26), (35) and (32) we have

$$|u'(t)| \le \max_{L_0 \le x \le L} |f(x)| \frac{1}{p(t)} \int_0^t p(s)\,\mathrm{d}s \le \max_{L_0 \le x \le L} |f(x)| \cdot t, \quad \text{for } t \in (0,\eta).$$

This fact implies $u'(0) = 0$. Moreover, multiplying equation (52) by $p(t)$ and deriving it we get (1). So, the fixed point u is a solution of problem (1), (37). Analogously, every solution of (1), (37) defined on $[0,\eta]$ is a fixed point of the operator $\mathcal{F}$. We conclude that there exists a unique solution of problem (1), (37).

Step 2. (Global solution) From Lemma 6 it follows, that the solution u can be extended onto every interval, where it is bounded. Lemma 8 gives $\theta > 0$ such that

$$u(\theta) = 0 \quad \text{and} \quad u'(t) > 0 \quad \text{for } (0,\theta]. \tag{53}$$

If u is defined on $[0,\omega)$, where $\omega \in (\theta,\infty]$, then

$$u'(t) = \frac{p(\theta)}{p(t)} u'(\theta) + \frac{1}{p(t)} \int_\theta^t p(s) f(u(s))\,\mathrm{d}s$$

for $t \in (\theta,\omega)$. From (29), (53) and the last equation we get three possibilities:
Case A. There exists $b > \theta$ such that

$$u(b) = L \quad \text{and} \quad u'(t) > 0 \quad \text{for } t \in [\theta,b).$$

Case B. For $t \in (\theta,\infty)$ it is valid $u(t) \in (0,L)$ and $u'(t) > 0$.
Case C. There exists $b > \theta$ such that

$$u'(b) = 0, \quad u(b) \in (0,L) \quad \text{and} \quad u'(t) > 0 \quad \text{for } t \in (\theta,b). \tag{54}$$

Let us consider Case A. Since $\tilde{u} \equiv L$ is the solution of the equation (1) and it satisfies $\tilde{u}(b) = L$, $\tilde{u}'(b) = 0$, then from Lemma 6 we get

$$u'(b) > 0.$$

It follows that there exists $\delta > 0$ such that

$$u'(t) > 0 \quad \text{and} \quad u(t) > L \quad \text{for } t \in (b, b+\delta).$$

In view of (35) the solution u satisfies

$$(p(t)u'(t))' = 0 \quad \text{for } t \in (b, b+\delta)$$

and consequently

$$u'(t) = \frac{p(b)u'(b)}{p(t)} > 0 \quad \text{and} \quad u(t) = L + p(b)u'(b)\int_b^t \frac{\mathrm{d}s}{p(s)},$$

for $t \in (b, b+\delta)$. From (31) and (32) it follows that u can be extended onto $[0, \infty)$. This solution is an escape solution.
Let us consider Case B. The monotonicity of u implies the existence of $\tilde{L} \in (0, L]$ such that

$$\lim_{t\to\infty} u(t) = \tilde{L}. \tag{55}$$

We will prove that $\tilde{L} = L$. Since $f(u(t)) < 0$ for $t > \theta$, from (1) it follows, that pu' is decreasing on (θ, ∞). The inequality $u'(t) > 0$ for $t \in (\theta, \infty)$ implies that $u'' < 0$ and hence u' is decreasing on (θ, ∞). That yields the existence of $\lim_{t\to\infty} u'$. Since u is bounded, necessarily

$$\lim_{t\to\infty} u'(t) = 0.$$

From (1) it follows that

$$u''(t) = -\frac{p'(t)}{p(t)}u'(t) + f(u(t))$$

for $t \in (0, \infty)$. In view of (33) we get

$$\lim_{t\to\infty} u''(t) = f(\tilde{L}).$$

According to (27) and (29) we get $\tilde{L} = L$. This solution satisfies the conditions (2) and so it is a homoclinic solution.

Let us consider Case C. From the second part of Lemma 8 we get $\alpha \in (0, \theta)$ and $\beta \in (\theta, b)$ such that (41) holds. In view of (54) we get

$$F(u(b)) = \left(\frac{p(\alpha)}{p(\beta)}\right)^2 F(B). \tag{56}$$

Using Lemma 9 we get the existence of $\theta_1 > b$ such that $u(\theta_1) = 0$ and $u'(t) < 0$ for $t \in (b, \theta_1]$. Let us suppose that there exists $\bar{b}_1 \in (\theta_1, \infty)$ such that

$$u(\bar{b}_1) = B \quad \text{and} \quad u'(t) < 0, \quad \text{for } t \in [\theta_1, \bar{b}_1).$$

Using the second part of Lemma 9, we get $\bar{\alpha}_1 \in (b, \theta_1)$ and $\bar{\beta}_1 \in (\theta_1, \bar{b}_1)$ such that

$$(pu')^2(\bar{b}_1) = 2[p^2(\bar{\alpha}_1)F(u(b)) - p^2(\bar{\beta}_1)F(B)],$$

and together with (56) we obtain

$$\begin{aligned}(pu')^2(\bar{b}_1) &= 2F(B)\left[p^2(\bar{\alpha}_1)\left(\frac{p(\alpha)}{p(\beta)}\right)^2 - p^2(\bar{\beta}_1)\right] \\ &= 2F(B)p^2(\bar{\beta}_1)\left[\left(\frac{p(\bar{\alpha}_1)p(\alpha)}{p(\bar{\beta}_1)p(\beta)}\right)^2 - 1\right] < 0.\end{aligned}$$

This is a contradiction. Hence, by Lemma 8, there exists $b_1 > \theta_1$ such that

$$u(b_1) \in (B, 0), \quad u'(b_1) = 0 \quad \text{and} \quad u'(t) < 0 \quad \text{for } t \in (\theta_1, b_1).$$

From the second part of Lemma 9 we get $\alpha_1 \in (b, \theta_1)$ and $\beta_1 \in (\theta_1, b_1)$ such that

$$0 = 2[p^2(\alpha_1)F(u(b)) - p^2(\beta_1)F(u(b_1))].$$

By (56), we get

$$F(u(b_1)) = \left(\frac{p(\alpha_1)}{p(\beta_1)}\right)^2 F(u(b)) = \left(\frac{p(\alpha_1)p(\alpha)}{p(\beta_1)p(\beta)}\right)^2 F(B). \tag{57}$$

Using Lemma 8 we get $\theta_2 > b_1$ such that $u(\theta_2) = 0$ and $u'(t) > 0$ for $t \in (b_1, \theta_2]$. Let us suppose that there exists $\bar{b}_2 \in (\theta_2, \infty)$ such that

$$u(\bar{b}_2) = u(b) \quad \text{and} \quad u'(t) > 0 \quad \text{for } t \in [\theta_2, \bar{b}_2).$$

By virtue of the second part of Lemma 8, we can find $\bar{\alpha}_2 \in (b_1, \theta_2)$ and $\bar{\beta}_2 \in (\theta_2, \bar{b}_2)$ such that

$$(pu')^2(\bar{b}_2) = 2[p^2(\bar{\alpha}_2)F(u(b_1)) - p^2(\bar{\beta}_2)F(u(b))],$$

and together with (57) we obtain

$$(pu')^2(\bar{b}_2) = 2F(u(b))p^2(\bar{\beta}_2)\left[\left(\frac{p(\bar{\alpha}_2)p(\alpha_1)}{p(\bar{\beta}_2)p(\beta_1)}\right)^2 - 1\right] < 0$$

a contradiction. Hence there exists $b_2 > \theta_2$ such that

$$u(b_2) \in (0, u(b_1)), \quad u'(b_2) = 0 \quad \text{and} \quad u'(t) < 0 \quad \text{for } (\theta_2, b_2).$$

Repeating this procedure we get a sequence $\{\theta_n\}_{n=1}^{\infty}$ of roots of the solution u and a sequence $\{b_n\}_{n=1}^{\infty}$ of roots of the derivative u' such that $\{|u(b_n)|\}_{n=1}^{\infty}$ is decreasing. This solution corresponds to an oscillatory solution.

STEP 3. (Estimations) Let u be a solution of problem (1), (37) with $B \in (L_0, 0)$. Then from (1) it follows that

$$u'(t) = \frac{1}{p(t)} \int_0^t p(s) f(u(s)) \, \mathrm{d}s, \quad \text{for } t \in (0, \infty). \tag{58}$$

Then, in view of (26) and (35)

$$|u'(t)| \leq \max_{L_0 \leq x \leq L} |f(x)| \cdot \int_0^t \mathrm{d}s = \max_{L_0 \leq x \leq L} |f(x)| \cdot t \quad \text{for } t \in (0, \infty).$$

Integrating (58) we get

$$|u(t)| \leq |u(0)| + \left| \int_0^t \frac{1}{p(s)} \int_0^s p(\tau) f(u(\tau)) \, \mathrm{d}\tau \, \mathrm{d}s \right| \leq B + \max_{L_0 \leq x \leq L} |f(x)| \cdot \frac{t^2}{2}.$$

The proof is complete. □

Lemma 11 *(On oscillatory solutions) Let (26)–(34) be satisfied, $B \in (L_0, 0)$ be such that*

$$F(B) < F(L). \tag{59}$$

Then the corresponding solution of problem (1), (37) is oscillatory.

Proof Let u be a solution of problem (1), (37) with $B \in (L_0, 0)$ satisfying (59). STEP 1. Let us assume that u is an escape solution. Then there exist $b > 0$, $\theta \in (0, b)$ such that

$$u(\theta) = 0, \quad u(b) = L \quad \text{and} \quad u'(t) > 0 \quad \text{for } t \in (0, b].$$

From Lemma 8 we get $\alpha \in (0, \theta)$, $\beta \in (\theta, b)$ such that (41) holds. Then

$$p^2(b) u'^2(b) = 2F(L) p^2(\beta) \left[\left(\frac{p(\alpha)}{p(\beta)} \right)^2 \frac{F(B)}{F(L)} - 1 \right] < 0.$$

This contradicts the fact that $u'(b) > 0$.

STEP 2. Let us assume that u is a homoclinic solution. Let $\theta > 0$ be the root of u and $b > \theta$ be arbitrary. Then, by Lemma 8, there exist $\alpha \in (0, \theta)$, $\beta \in (\theta, b)$ such that (41) holds. From (41), the fact $(pu')^2(b) > 0$ and (32) we get

$$F(B) > \left(\frac{p(\beta)}{p(\alpha)} \right)^2 F(u(b)) > F(u(b)).$$

Letting $b \to \infty$ we get $F(B) \geq F(L)$, which contradicts (59). □

Actually, the homoclinic solution is the desired strictly increasing solution of the problem (1), (2). In order to prove the existence of such solution we need the lower and upper functions method for the singular mixed problem

$$(p(t)u')' = p(t) f(u), \qquad u'(a) = 0, \ u(b) = L, \tag{60}$$

where $a, b \in \mathbb{R}$, $a \geq 0$, $b > a$.

Definition 12 A function $\sigma \in C([a,b])$ is called a lower function of problem (60), if there exists a finite set $\Sigma \subset (a,b)$ such that $\sigma \in C^2((a,b] \setminus \Sigma)$, $\sigma'(\tau^+)$, $\sigma'(\tau^-) \in \mathbb{R}$ for $\tau \in \Sigma$,

$$(p(t)\sigma'(t))' \geq p(t)f(\sigma(t)) \quad \text{for } t \in (a,b] \setminus \Sigma,$$

$$\sigma'(a^+) \geq 0,\ \sigma(b) \leq L,\ \sigma'(\tau^-) < \sigma'(\tau^+) \quad \text{for } \tau \in \Sigma.$$

If all inequalities are reversed, then σ is called an upper function of problem (60).

Note that $\sigma'(a^+)$ need not be bounded if $a = 0$.

Theorem 13 *Let p satisfy (31), (32), $f \in C(\mathbb{R})$, σ_1 and σ_2 be a lower function and an upper function of problem (60) and let $\sigma_1(t) \leq \sigma_2(t)$ for $t \in [a,b]$. Then problem (60) has a solution $u \in C^1([a,b]) \cap C^2((a,b])$ such that $\sigma_1(t) \leq u(t) \leq \sigma_2(t)$ for $t \in [a,b]$.*

Proof See [8] Theorem 2.3 for $a = 0$. For $a > 0$ problem (60) is regular and therefore we can use a simplified form of the proof in [8]. □

The next assertion is based on Lemma 4 and Theorem 13.

Lemma 14 *(On escape solutions) Let (26)–(35) be satisfied. There exist $B_* \in (L_0, 0)$ and $c_* \in (0,\infty)$ such that a solution u_* of problem (1), (37) with $B = B_*$ satisfies the condition*

$$u_*(c_*) = L, \qquad u_*'(t) > 0 \quad \text{on } (0, c_*].$$

Proof Let us put

$$\tilde{f}(x) = \begin{cases} f(x) & \text{for } x \leq L, \\ x - L & \text{for } x \geq L. \end{cases} \tag{61}$$

Let $\epsilon_0 \in \mathbb{R}$ be from Lemma 4 for L_0, 0, L, $\tilde{f}$, $\tilde{F}$ in place of x_1, x_2, x_3, h, H, respectively. Here, $\tilde{F}(x) = -\int_0^x \tilde{f}(z)\,\mathrm{d}z$, $x \in \mathbb{R}$. The assumptions of Lemma 4 are satisfied due to (26)–(30), (61). Consider the perturbed equation

$$u'' = \tilde{f}(u) - \epsilon \tag{62}$$

with $\epsilon \in (0, \epsilon_0)$. From Lemma 5 it follows that there exists $B_L \in (L_0, 0)$ such that for the corresponding solution u_L of problem (62), (37) with $B = B_L$, there exists $b > 0$ such that $u_L(b) = L$ and

$$0 < u_L'(t) \leq \sqrt{2\tilde{F}(L_0)} \quad \text{for } t \in [0,b]. \tag{63}$$

From (33) it follows that there exists $a > 0$ such that

$$\frac{p'(t)}{p(t)} < \frac{\epsilon}{\sqrt{2\tilde{F}(L_0)}} \qquad \text{for } t > a.$$

Put $v(t) = u_L(t-a)$ for $t \in [a, a+b]$. Then v satisfies equation (62) on $[a, a+b]$ and fulfils the initial conditions

$$v(a) = B_L, \quad v'(a) = 0.$$

Moreover, $v(a+b) = L$, $\tilde{f}(v(t)) = f(v(t))$ and

$$0 < \frac{p'(t)}{p(t)} v'(t) < \frac{\epsilon}{\sqrt{2F(L_0)}} \sqrt{2F(L_0)} = \epsilon$$

for $t \in [a, a+b]$. Therefore

$$v''(t) = f(v(t)) - \epsilon < f(v(t)) - \frac{p'(t)}{p(t)} v'(t)$$

for $t \in (a, a+b]$. We can see that v is an upper function of the problem

$$u'' + \frac{p'(t)}{p(t)} u' = f(u), \qquad u'(a) = 0, \ u(a+b) = L. \tag{64}$$

Since L_0 is a lower function of problem (64), by Theorem 13 and Lemma 6 there exists a solution u_0 of (64) such that

$$L_0 < u_0(t) \leq v(t) \qquad \text{for } t \in [a, a+b]. \tag{65}$$

By (63), (64), (65) we have $v'(a+b) > 0$, $u_0(a+b) = v(a+b)$ and $u_0(t) \leq v(t)$ for $t \in [a, a+b]$. Therefore

$$u_0'(a+b) > 0. \tag{66}$$

Since $u_0''(a) = f(u_0(a)) > 0$ there exists a minimal $a_0 \in [0, a)$ such that $u_0'(t) < 0$ for $t \in (a_0, a)$ and $u_0(t) < 0$ for $t \in (a_0, a]$. There are two possibilities.

(i) $a_0 > 0$, $u_0(a_0) = 0$,

(ii) $a_0 = 0$, $u_0(t) \leq 0$ for $t \in [0, a]$.

Assume that (i) holds. Then we put

$$\beta(t) = \begin{cases} 0 & \text{for } t \in [0, a_0], \\ u_0(t) & \text{for } t \in (a_0, a+b]. \end{cases}$$

Assume that (ii) holds. Then $u_0''(t) > 0$ for $t \in [0, a]$ and

$$\lim_{t \to 0+} u_0'(t) < 0$$

and we put

$$\beta(t) = u_0(t) \quad \text{for } t \in [0, a+b].$$

Denote $c_* = a+b$. In both cases (i) and (ii) the function β is an upper function of the problem

$$u'' + \frac{p'(t)}{p(t)} u' = f(u), \qquad u'(0) = 0, \quad u(c_*) = L. \tag{67}$$

Since the constant L_0 is a lower function of problem (67), then, by Theorem 13 and Lemma 6, there exists a solution u_* of the problem (67) such that

$$L_0 < u_*(t) \le \beta(t) \qquad \text{for } t \in [0, c_*]. \tag{68}$$

We put $B_* = u_*(0)$. Then u_* is a solution of (1), (37) with $B = B_*$. Finally, by (64) and (66) we have

$$\beta(c_*) = L, \quad \beta'(c_*) > 0.$$

This, together with (68) gives $u_*'(c_*) > 0$. Hence, by Lemma 10, $u_*'(t) > 0$ for $t \in (0, c_*]$. □

Theorem 15 *(On homoclinic solutions) Let (26)–(34) be satisfied. Then there exists at least one strictly increasing solution of problem (1), (2).*

Proof First, we will assume that (35) is satisfied. Let us define

$$\mathcal{M} = \{B_0 \in (L_0, 0) \colon \text{each solution of (1), (37) with } B \in [B_0, 0) \text{ is oscillatory}\},$$

and $\tilde{B} = \inf \mathcal{M}$. Lemma 11 guarantees that $\mathcal{M} \neq \emptyset$ and from Lemma 14 it follows that $\tilde{B} > L_0$. We will prove that there exists $B_{\text{hom}} \in (L_0, \tilde{B}]$ such that the corresponding solution of the problem (1), (37) with $B = B_{\text{hom}}$ is a homoclinic solution. Assume that B_{hom} does not exist.

Case A. Let $\tilde{u}$ be an oscillatory solution of (1), (37) with $B = \tilde{B}$. Then, according to the definition of $\tilde{B}$, we can find a sequence $\{B_n\} \subset (L_0, \tilde{B})$ such that $\lim_{n\to\infty} B_n = \tilde{B}$ and the corresponding solutions u_n of (1), (37) with $B = B_n$ are escape solutions. Let θ_1 be the second zero of $\tilde{u}$, that is, θ_1 fulfils

$$\tilde{u}(\theta_1) = 0, \quad \tilde{u}'(\theta_1) < 0.$$

From Lemma 10 we can see that

$$|u_n(t)| \le L_0 + \frac{\theta_1^2}{2} \max_{L_0 \le x \le L} |f(x)|, \quad |u_n'(t)| \le \theta_1 \cdot \max_{L_0 \le x \le L} |f(x)|$$

for $t \in [0, \theta_1]$, $n \in \mathbb{N}$. Hence the sequence $\{u_n\}$ is bounded and equicontinuous on $[0, \theta_1]$. Therefore we can choose a subsequence $\{u_m\}$, which is uniformly convergent on $[0, \theta_1]$ to a function $v \in C([0, \theta_1])$. Obviously,

$$u_m(t) = B_m + \int_0^t \frac{1}{p(s)} \int_0^s p(\tau) f(u_m(\tau)) \,\mathrm{d}\tau \,\mathrm{d}s$$

for $t \in [0, \theta_1]$, $m \in \mathbb{N}$, and consequently

$$v(t) = \tilde{B} + \int_0^t \frac{1}{p(s)} \int_0^s p(\tau) f(v(\tau)) \,\mathrm{d}\tau \,\mathrm{d}s$$

for $t \in [0, \theta_1]$. We can check that v is a solution of problem (1), (37) and therefore

$$v = \tilde{u} \qquad \text{on } [0, \theta_1].$$

Since u_m are increasing, it follows that v is nondecreasing on $[0, \theta_1]$. This contradicts the fact that $v'(\theta_1) < 0$.

CASE B. Let $\tilde{u}$ be an escape solution of (1), (37) with $B = \tilde{B}$. Then there exists $b > 0$ such that

$$\tilde{u}(b) = L, \quad \tilde{u}'(t) > 0 \quad \text{for } t \in (0, \infty). \tag{69}$$

From the definition of $\tilde{B}$ we get a sequence $\{B_n\} \subset (\tilde{B}, 0)$ such that $\lim_{n\to\infty} B_n = \tilde{B}$ and the corresponding solutions u_n of (1), (37), with $B = B_n$, are oscillatory. Therefore

$$L_0 \le u_n(t) \le L, \qquad |u_n'(t)| \le t \cdot \max_{L_0 \le x \le L} |f(x)| \quad \text{for } t \in [0, \infty), n \in \mathbb{N},$$

and there exist $b_n > 0$ such that $u_n(b_n) = L_n \in (0, L)$, $u_n'(b_n) = 0$ for $n \in \mathbb{N}$. Then there exist $\theta_n > b_n$ such that

$$u_n(\theta_n) = 0, \quad u_n'(\theta_n) < 0, \ n \in \mathbb{N}. \tag{70}$$

The sequence $\{u_n\}$ is bounded and equicontinuous on every $[0, K] \subset [0, \infty)$ and so we can choose a subsequence $\{u_m\}$ which is uniformly convergent on $[0, K]$ to a function $w \in C([0, K])$. As in CASE A we conclude that $w = \tilde{u}$ on $[0, K]$.

Now, we have two possibilities.

(i) Let $\lim_{m\to\infty} \theta_m = \theta_0 < \infty$. Put $K = \max\{\theta_0, b\} + 1$. By (70), each u_m is decreasing at a neighbourhood of θ_m and hence $\tilde{u}$ is nonincreasing at θ_0, which contradicts (69).

(ii) Let $\lim_{m\to\infty} \theta_m = \infty$. Put $K = b + 1$. Since $u_m(b + 1) < L$ for $m \in \mathbb{N}$, it follows that $\tilde{u}(b + 1) \le L$, which is a contradiction.

We have proved that the function $\tilde{u}$ can be neither an escape solution nor an oscillatory solution. Lemma 10 yields that $\tilde{u}$ is a homoclinic solution of problem (1), (2). Since $\tilde{u}(t) \in [L_0, L]$ for $t \in [0, \infty)$ we see that assumption (35) can be omitted. □

Acknowledgements The authors were supported by the Council of Czech Government MSM 6198959214.

References

[1] Berestycki, H., Lions, P. L., Peletier, L. A.: *An ODE approach to the existence of positive solutions for semilinear problems in* $\mathbb{R}^N$. Indiana University Mathematics Journal **30**, 1 (1981), 141–157.

[2] Bonheure, D., Gomes, J. M., Sanchez, L.: *Positive solutions of a second–order singular ordinary differential equation.* Nonlinear Analysis **61** (2005), 1383–1399.

[3] Dell'Isola, F., Gouin, H., Rotoli, G.: *Nucleation of spherical shell-like interfaces by second gradient theory: numerical simulations.* Eur. J. Mech B/Fluids **15** (1996), 545–568.

[4] Gouin, H., Rotoli, G.: *An analytical approximation of density profile and surface tension of microscopic bubbles for Van der Waals fluids.* Mech. Research Communic. **24** (1997), 255–260.

[5] Kitzhofer, G., Koch, O., Lima, P., Weinmüller, E.: *Efficient numerical solution of the density profile equation in hydrodynamics.* J. Sci. Comput. **32**, 3 (2007), 411–424.

[6] Koch, O., Kofler, P., Weinmüller, E.: *Initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind.* Analysis **21** (2001), 373–389.

[7] Lima, P. M., Chemetov, N. V., Konyukhova, N. B., Sukov, A. I.: *Analytical–numerical investigation of bubble-type solutions of nonlinear singular problems.* J. Comp. Appl. Math. **189** (2006), 260–273.

[8] Rachůnková, I., Koch, O., Pulverer, G., Weinmüller, E.: *On a singular boundary value problem arising in the theory of shallow membrane caps.* J. Math. Anal. Appl. **332** (2007), 532–541.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 129–137

Integral Presentations of Deviations of de la Vallee Poussin Right-Angled Sums

Vladimir I. RUKASOV, Olga G. ROVENSKA

Department of Mathematical Analysis, Slavyansk State Pedagogical University,
Batyuka 19, Slavyansk, Ukraine
e-mail: o.rovenskaya@mail.ru

(Received January 10, 2009)

Abstract

We investigate approximation properties of de la Vallee Poussin right-angled sums on the classes of periodic functions of several variables with a high smoothness. We obtain integral presentations of deviations of de la Vallee Poussin sums on the classes $C^{m\alpha}_{\beta,\infty}$.

Key words: Right-angled sums of Vallee Poussin, integral presentations, Fourier series.

2000 Mathematics Subject Classification: 42A10

1 Introduction

Considering [1] we define $\overline{\psi}$-integral classes of periodic functions of several variables in the following way.

Let R^m be an Euclidean space with elements $\vec{x} = (x_1, x_2, \ldots, x_m)$, and let $T^m = \prod_{i=1}^{m}[-\pi;\pi]$ be an m-dimensional cube with the side 2π,

$$\begin{aligned} N^m &= \big\{\vec{x} \in R^m \mid x_i \in N,\ i = 1, 2, \ldots, m\big\},\\ N^m_* &= \big\{\vec{x} \in R^m \mid x_i \in N_* = N \cup \{0\},\ i = 1, 2, \ldots, m\big\},\\ N^m_i &= \big\{\vec{x} \in R^m \mid x_i \in N, x_j \in N_*,\ i \neq j\big\},\\ E^m &= \big\{\vec{x} \in R^m \mid x_i \in \{0; 1\},\ i = 1, 2\big\}. \end{aligned}$$

We denote by $L(T^m)$ the set of summable on a cube T^m functions $f(\vec{x}) = f(x_1, x_2, \ldots, x_m)$ which are 2π-periodic on every variable.

Let $f \in L(T^m)$. Then for every pair of points $\vec{s} \in E^m$, $\vec{k} \in N_*^m$ we have a corresponding value

$$a_{\vec{k}}^{\vec{s}}(f) = \frac{1}{\pi^m} \int\limits_{T^m} f(\vec{x}) \prod_{i=1}^{m} \cos\left(k_i x_i - \frac{s_i \pi}{2}\right) dx_i. \tag{1}$$

Values $a_{\vec{k}}^{\vec{s}}(f)$, $\vec{s} \in E^m$, $\vec{k} \in N_*^m$ are the Fourier coefficients of the function $f(\vec{x})$ [1, p. 546].

For every vector $\vec{k} \in N_*^m$ we have the major harmonic of the function $f(\vec{x})$

$$A_{\vec{k}}(f;\vec{x}) = \sum_{\vec{s} \in E^m} a_{\vec{k}}^{\vec{s}}(f) \prod_{i=1}^{m} \cos\left(k_i x_i - \frac{s_i \pi}{2}\right) \tag{2}$$

and on the variable x_i conjugated harmonic

$$A_{\vec{k}}^{\vec{e}_i}(f;\vec{x}) = \sum_{\vec{s} \in E^m} a_{\vec{k}}^{\vec{s}}(f) \prod_{j \in \overline{m} \setminus \{i\}} \cos\left(k_j x_j - \frac{s_j \pi}{2}\right) \cos\left(k_i x_i - \frac{(s_i+1)\pi}{2}\right).$$

Using [1, p. 545] we define Fourier series of the function $f(\vec{x})$ by the following relation

$$S[f] = \sum_{\vec{k} \in N_*^m} \frac{1}{2^{q(\vec{k})}} A_{\vec{k}}(f, \vec{x}), \tag{3}$$

where $q(\vec{k})$ is a number of zero coordinates of the vector $\vec{k}$.

Let $f \in L(T^m)$ and systems of numbers $\psi_{ij}(k)$, $\Psi_{ij}(k)$, $i = 1, 2, \ldots, m$; $j = 1, 2$, $k \in N_*$ be given.

Let us put

$$\overline{\psi}_i(k) = \sqrt{\psi_{i1}^2(k) + \psi_{i2}^2(k)}, \overline{\Psi}_i(k) = \sqrt{\Psi_{i1}^2(k) + \Psi_{i2}^2(k)}$$

and consider the following conditions be fulfilled: $\overline{\psi}_i(k) \neq 0$, $\overline{\Psi}_i(k) \neq 0$, $k \in N_*$, $\psi_{i1}(0) = 1$, $\Psi_{i1}(0) = 1$, $\psi_{i2}(0) = 0$, $\Psi_{i2}(0) = 0$, $i = 1, 2, \ldots, m$.

Furthermore, let

$$\sum_{\vec{k} \in N_i^m} \frac{1}{2^{q(\vec{k})} \overline{\psi}_i^2(k_i)} [\psi_{i1}(k_i) A_{\vec{k}}(f, \vec{x}) - \psi_{i2}(k_i) A_{\vec{k}}^{\vec{e}_i}(f, \vec{x})] \tag{4}$$

be the Fourier series of some function of $L(T^m)$. It will be denoted by

$$f^{\overline{\psi}_i}(\vec{x}) = \frac{\partial^{\overline{\psi}_i} f(\vec{x})}{\partial x_i}$$

and called $\overline{\psi}_i$-derivative of the function f with respect to the x_i, $i \in \overline{m}$.

Let $\overline{m} = \{1, 2, \dots, m\}$. For a fixed r-elemental set $\mu(r) \subset \overline{m}$, $\mu(r) = \{i_1, i_2, \dots, i_r\}$, we define a function $f^{\overline{\Psi}_\mu}(\vec{x})$ by

$$f^{\overline{\Psi}_\mu}(\vec{x}) = \frac{\partial^{\overline{\Psi}_{i_r}} \partial^{\overline{\Psi}_{i_{r-1}}} \dots \partial^{\overline{\Psi}_{i_1}} f(\vec{x})}{\partial x_{i_r} \partial x_{i_{r-1}} \dots \partial x_{i_1}}$$

and call it mixed $\overline{\Psi}_\mu$-derivative with respect to variables x_i, $i \in \mu(r)$.

Let a set of functions ψ_{ij}, Ψ_{ij}, $i = 1, 2, \dots, m$; $j = 1, 2$ be given. The set of continuous functions $f \in L(T^m)$ having the essentially bounded $\overline{\Psi}_\mu$- and $\overline{\psi}_i$-derivatives, i.e.

$$\operatorname{ess\,sup} \left| f^{\overline{\Psi}_\mu}(\vec{x}) \right| \le 1, \quad \operatorname{ess\,sup} \left| f^{\overline{\psi}_i}(\vec{x}) \right| \le 1, \quad i = 1, 2, \dots, m; \ \mu \subset \overline{m}; \ \vec{x} \in T^m \tag{5}$$

will be denoted by the symbol $C_\infty^{m\overline{\psi}}$.

If for the sets of functions $\psi_{ij}(k)$ and $\Psi_{ij}(k)$, $i = 1, 2, \dots, m$; $j = 1, 2$, the functions $\psi_i(k)$, $\Psi_i(k)$ and numbers β_i, β_i^*, $i = 1, 2, \dots, m$, fulfil

$$\psi_{i1}(k) = \psi_i(k) \cos \frac{\beta_i \pi}{2}; \quad \psi_{i2}(k) = \psi_i(k) \sin \frac{\beta_i \pi}{2};$$

$$\Psi_{i1}(k) = \Psi_i(k) \cos \frac{\beta_i^* \pi}{2}, \quad \Psi_{i2}(k) = \Psi_i(k) \sin \frac{\beta_i^* \pi}{2}, \quad i = 1, 2, \dots, m,$$

then the class $C_\infty^{m\overline{\psi}}$ is the class of (ψ, β)-differentiable periodic functions of m variables (see [2]) and it is denoted by $C_{\beta,\infty}^{m\psi}$. For $m = 2$ these classes are the classes of (ψ, β)-differentiable periodic functions of two variables which are defined in [3] (see also [1]). In the case when the conditions $\Psi_1(k) = k^{-r}$, $\Psi_2(k) = k^{-s}$, $\psi_1(k) = k^{-r_1}$, $\psi_2(k) = k^{-s_1}$, $\beta_1 = r$, $\beta_1^* = s$, $\beta_2 = r_1$, $\beta_2^* = s_1$ for the $r > 0$, $s > 0$, $r_1 \ge r$, $s_1 \ge s$ are also fulfilled the classes $C_{\beta,\infty}^{2\psi}$ and $W_{r_1,s_1}^{r,s}$ are equal(see for example [4]). In [4] (see [5], too) there is proved the asymptotic equality of upper bounds of deviations of Fourier right-angled sums $S_{\vec{n}}(f, \vec{x})$ (taking at the classes $W_{r_1,s_1}^{r,s}$) for $n_i \to \infty$, $i = 1, 2$:

$$\mathcal{E}(W_{r_1,s_1}^{r,s}; S_{\vec{n}}) = \frac{4 \ln n_1}{\pi^2 n_1^{r_1}} + \frac{4 \ln n_2}{\pi^2 n_2^{s_1}} + O(1) \left(\frac{\ln n_1 \ln n_2}{n_1^r n_2^s} + \frac{1}{n_1^{r_1}} + \frac{1}{n_2^{s_1}} \right).$$

Let us put $G_{\vec{n},\vec{p}} = \prod_{i=1}^m [n_i - p_i; n_i - 1]$ for $\vec{n} \in N^m$, $\vec{p} \in N^m$, $p_i < n_i$, $i = 1, 2, \dots, m$,. Then trigonometric polynomials of the type

$$V_{\vec{n},\vec{p}}(f; \vec{x}) = \frac{1}{\prod_{i=1}^m p_i} \sum_{\vec{k} \in G_{\vec{n},\vec{p}}} S_{\vec{k}}(f; \vec{x}), \tag{6}$$

(where $S_{\vec{k}}(f; \vec{x})$ are partial sums of Fourier series defined (2), $\vec{n} \in N^m$, $p_i \in N$, $p_i < n_i$, $i = 1, 2, \dots, m$) are called Vallee Poussin right-angled sums.

In this work the problems of approximation of classes $C_{\beta,\infty}^{m\psi}$ by polynomials $V_{\vec{n},\vec{p}}(f; \vec{x})$ are investigated. The functions which determine these classes are defined in the following way:

$$\psi_i(x) = e^{-\alpha_i x}, \quad \Psi_i(x) = e^{-\alpha_i^* x}, \quad \alpha_i > 0, \ \alpha_i^* > 0, \ i = 1, 2, \dots, m.$$

We denote such classes by $C^{m\alpha}_{\beta,\infty}$ (analagously to the classes of functions of a single variable).

It is proved by S. M. Nikol'skii in [6] (see also [7], [8]) that for upper bounds of the deviations of Fourier sums on the corresponding classes $C^{\alpha}_{\beta,\infty}$ functions of one variabl we obtain the following asymptotic equality for $n \to \infty$:

$$\mathcal{E}\left(C^{\alpha}_{\beta,\infty}; S_n\right) = \frac{8q^n}{\pi^2}K(q) + O(1)\frac{q^n}{n}, \quad q = e^{-\alpha}, \tag{7}$$

where

$$K(q) = \int\limits_0^{\frac{\pi}{2}} \frac{du}{\sqrt{1 - q^2\sin^2 u}}$$

is the total elliptic integral of the first kind.

Asymptotic equalities for upper bounds of the deviations of de la Vallee Poussin sums on the classes $C^{\alpha}_{\beta,\infty}$ may be found in the [9], [10] (see also [11], [12, p. 217]):

$$\mathcal{E}\left(C^{\alpha}_{\beta,\infty}; V_{n,p}\right) = \frac{4q^{n-p+1}}{\pi p(1-q^2)} + O(1)\left(\frac{q^{n-p+1}}{p(n-p)(1-q)^3} + \frac{q^n}{p(1-q^2)}\right), \ 1 < p < n. \tag{8}$$

The 2-dimensional and m-dimensional analogies of equality (7) for the classes $C^{m\alpha}_{\beta,\infty}$ are in the works [13], [14].

2 Main Results

Let $\Lambda = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_m\}$ be a fixed set of infinite triangle numeric matrices, $\Lambda_i = \{\lambda^{(n_i)}_{k_i}\}$, $i = 1, 2, \ldots, m$, $\lambda^{(n_i)}_0 = 1$, $\lambda^{(n_i)}_{k_i} = 0$ for $k_i \geq n_i$.

Further let $\lambda^{(\vec{n})}_{\vec{k}} = \prod_{i=1}^m \lambda^{(n_i)}_{k_i}$ and let $G_{\vec{n}} = \prod_{i=1}^m [0; n_i - 1]$ be an right-angled parallelepiped corresponding to the vector $\vec{n} \in N^m$.

For every function with Fourier series (1) we have trigonometric polynomial

$$U_{\vec{n}}(f; \vec{x}; \Lambda) = \sum_{\vec{k} \in G_{\vec{n}}} 2^{-q(\vec{k})}\lambda^{(\vec{n})}_{\vec{k}} A_{\vec{k}}(f; \vec{x}).$$

Values $\delta_{\vec{n}}(f; \vec{x}; \Lambda) = f(\vec{x}) - U_{\vec{n}}(f; \vec{x}; \Lambda)$ are the deviations of such polynomials of the function $f(\vec{x})$.

In this work there are found the integral presentations of the deviations

$$\delta_{\vec{n},\vec{p}}(f, \vec{x}) = f(\vec{x}) - V_{\vec{n},\vec{p}}(f, \vec{x})$$

of sums $V_{\vec{n},\vec{p}}(f, \vec{x})$ from function $f(\vec{x})$ out of classes $C^{m\alpha}_{\beta,\infty}$.

The following theorem is the main result of this work.

Theorem 1 *If* $\alpha_i > 0$, $\alpha_i^* > 0$, $q_i = e^{-\alpha_i}$, $Q_i = e^{-\alpha_i^*}$, $\beta_i \in R$, $\beta_i^* \in R$, $p_i \in N$, $1 < p_i < n_i$; $i = 1, 2, \ldots, m$,

then for every function $f \in C^{m\alpha}_{\beta,\infty}$ *the following equality is fulfilled*

$$\delta_{\vec{n},\vec{p}}(f,\vec{x}) = \sum_{i=1}^{m} \frac{q_i^{n_i-p_i+1}}{p_i\pi} \int_{-\pi}^{\pi} f_{\beta_i}^{\psi_i}(\vec{x}+t_i\vec{e}_i) b_{n_i-p_i}^{\beta_i}(t_i)dt_i$$

$$- \sum_{i=1}^{m} \frac{q_i^{n_i+1}}{p_i\pi} \int_{-\pi}^{\pi} f_{\beta_i}^{\psi_i}(\vec{x}+t_i\vec{e}_i) b_{n_i}^{\beta_i}(t_i)dt_i$$

$$+ O(1) \sum_{r=2}^{m} \sum_{\mu(r)\in\overline{m}} \prod_{j\in\mu(r)} Q_j^{n_j-p_j+1} \int_{T^r} \left| B_{n_j-p_j}^{\beta_j^\star}(t_j) \right| dt_j, \tag{9}$$

where

$$b_{n_i}^{\beta_i}(t_i) = \frac{(q_i^2\cos t_i - 2q_i + \cos t_i)}{(1-2q_i\cos t_i + q_i^2)^2} \cos\left(n_i t_i + \frac{\beta_i\pi}{2}\right)$$

$$+ \frac{(q_i^2\sin t_i - \sin t_i)}{(1-2q_i\cos t_i + q_i^2)^2} \sin\left(n_i t_i + \frac{\beta_i\pi}{2}\right),$$

$$B_{n_i}^{\beta_i^\star}(t_i) = \frac{(Q_i^2\cos t_i - 2Q_i + \cos t_i)}{(1-2Q_i\cos t_i + Q_i^2)^2} \cos\left(n_i t_i + \frac{\beta_i^\star\pi}{2}\right)$$

$$+ \frac{(Q_i^2\sin t_i - \sin t_i)}{(1-2Q_i\cos t_i + Q_i^2)^2} \sin\left(n_i t_i + \frac{\beta_i^\star\pi}{2}\right).$$

Proof It is clear that

$$\delta_{\vec{n},\vec{p}}(f;\vec{x}) = \frac{1}{\prod_{i=1}^m p_i} \sum_{\vec{k}\in G_{\vec{n},\vec{p}}} \rho_{\vec{k}}(f;\vec{x}) = \frac{1}{\prod_{i=1}^m p_i} \sum_{i=1}^{m} \sum_{k_i=n_i-p_i}^{n_i-1} \rho_{\vec{k}}(f;\vec{x}), \tag{10}$$

where

$$\rho_{\vec{k}}(f;\vec{x}) = f(\vec{x}) - S_{\vec{k}}(f;\vec{x}), \quad \vec{k} = (k_1;k_2;\ldots;k_m).$$

Let us investigate $\rho_{\vec{k}}(f;\vec{x})$. Using theorem 1 in [13] for $f \in C^{m\alpha}_{\beta,\infty}$ we have

$$\rho_{\vec{n}}(f,\vec{x}) = \sum_{i=1}^{m} \frac{1}{\pi} \int_{-\pi}^{\pi} f_{\beta_i}^{\psi_i}(\vec{x}+t_i\vec{e}_i) \sum_{k=n_i+1}^{\infty} \exp(-\alpha_i k)\cos\left(kt_i + \frac{\beta_i\pi}{2}\right) dt_i$$

$$+ \sum_{r=2}^{m} (-1)^{r+1} \sum_{\mu(r)\in\overline{m}} \frac{1}{\pi^r} \int_{T^r} f_{\beta_\mu^\star}^{\Psi_\mu}\left(\vec{x} + \sum_{i\in\mu(r)} t_i\vec{e}_i\right)$$

$$\times \prod_{j\in\mu(r)} \sum_{k_j=n_j+1}^{\infty} \exp(-\alpha_i^\star k_j)\cos\left(k_j t_j + \frac{\beta_j^\star\pi}{2}\right) dt_j.$$

Denote $q_i = \exp(-\alpha_i)$, $Q_i = \exp(-\alpha_i^\star)$. Using [15, p. 123–124] we obtain

$$\sum_{k=n_i+1}^{\infty} \exp(-\alpha_i k)\cos\left(kt_i + \frac{\beta_i\pi}{2}\right)$$

$$= q_i^{n_i}\left[\frac{q_i\cos t_i - q_i^2}{1-2q_i\cos t_i + q_i^2}\cos\left(n_i t_i + \frac{\beta_i\pi}{2}\right) - \frac{q_i\sin t_i}{1-2q_i\cos t_i+q_i^2}\sin\left(n_i t_i + \frac{\beta_i\pi}{2}\right)\right].$$

If

$$h_{n_i}^{\beta_i}(t_i) = \frac{(q_i\cos t_i - q_i^2)\cos\left(n_i t_i + \frac{\beta_i\pi}{2}\right) - q_i\sin t_i\sin\left(n_i t_i + \frac{\beta_i\pi}{2}\right)}{1-2q_i\cos t_i + q_i^2},$$

$$H_{n_i}^{\beta_i^\star}(t_i) = \frac{(Q_i\cos t_i - Q_i^2)\cos\left(n_i t_i + \frac{\beta_i\pi}{2}\right) - Q_i\sin t_i\sin\left(n_i t_i + \frac{\beta_i\pi}{2}\right)}{1-2Q_i\cos t_i + Q_i^2}$$

then

$$\rho_{\vec{n}}(f,\vec{x}) = \sum_{i=1}^{m}\frac{1}{\pi}\int_{-\pi}^{\pi} f_{\beta_i}^{\psi_i}(\vec{x}+t_i\vec{e}_i)q_i^{n_i}h_{n_i}^{\beta_i}(t_i)dt_i$$

$$+\sum_{r=2}^{m}(-1)^{r+1}\sum_{\mu(r)\in\overline{m}}\frac{1}{\pi^r}\int\limits_{T^r} f_{\beta_\mu^\star}^{\Psi_\mu}\left(\vec{x}+\sum_{i\in\mu(r)}t_i\vec{e}_i\right)\prod_{j\in\mu(r)}Q_j^{n_j}H_{n_j}^{\beta_j^\star}(t_j)\,dt_j.$$

According to (10) we obtain

$$\delta_{\vec{n},\vec{p}}(f,\vec{x}) = \sum_{i=1}^{m}\frac{1}{p_i\pi}\sum_{k_i=n_i-p_i}^{n_i-1} q_i^{k_i}\int_{-\pi}^{\pi} f_{\beta_i}^{\psi_i}(\vec{x}+t_i\vec{e}_i)h_{k_i}^{\beta_i}(t_i)dt_i$$

$$+\sum_{r=2}^{m}(-1)^{r+1}\sum_{\mu(r)\in\overline{m}}\frac{1}{\pi^r}\int\limits_{T^r} f_{\beta_\mu^\star}^{\Psi_\mu}\left(\vec{x}+\sum_{i\in\mu(r)}t_i\vec{e}_i\right)$$

$$\times\prod_{j\in\mu(r)}\frac{1}{p_j}\sum_{\nu_j=n_j-p_j}^{n_j-1}Q_j^{\nu_j}H_{\nu_j}^{\beta_j^\star}(t_j)\,dt_j. \tag{11}$$

Let us use [11, p. 232–234]. Applying elementary transformations we obtain

$$\sum_{k_i=n_i-p_i}^{n_i-1} q_i^{k_i}h_{k_i}^{\beta_i}(t) = \sum_{k_i=n_i-p_i}^{n_i-1} q_i^{k_i+1}\left[(\cos(k_i+1)t - q_i\cos k_i t)\cos\frac{\beta_i\pi}{2}\right.$$

$$\left.-(\sin(k_i+1)t - q_i\sin k_i t)\sin\frac{\beta_i\pi}{2}\right](1-2q_i\cos t+q_i^2)^{-1}$$

$$\stackrel{\text{df}}{=}\frac{\Sigma_{i,1}(t)\cos\frac{\beta_i\pi}{2} - \Sigma_{i,2}(t)\sin\frac{\beta_i\pi}{2}}{1-2q_i\cos t+q_i^2}. \tag{12}$$

Let us investigate $\Sigma_{i,1}(t)$ and $\Sigma_{i,2}(t)$. We may write

$$\Sigma_1(t) = \sum_{k=n-p}^{n-1} q^{k+1}(\cos(k+1)t - q\cos kt) = \frac{1}{2}\Big[\sum_{k=0}^{n}(qe^{it})^k - \sum_{k=0}^{n-p}(qe^{it})^k\Big]$$

$$+ \frac{1}{2}\Big[\sum_{k=0}^{n}(qe^{-it})^k - \sum_{k=0}^{n-p}(qe^{-it})^k\Big] - \frac{q^2}{2}\Big[\sum_{k=0}^{n-1}(qe^{it})^k - \sum_{k=0}^{n-p-1}(qe^{it})^k\Big]$$

$$- \frac{q^2}{2}\Big[\sum_{k=0}^{n-1}(qe^{-it})^k - \sum_{k=0}^{n-p-1}(qe^{-it})^k\Big]$$

$$= \frac{1}{2}\Big[\frac{(qe^{it})^{n+1}-1}{qe^{it}-1} - \frac{(qe^{it})^{n-p+1}-1}{qe^{it}-1}\Big] + \frac{1}{2}\Big[\frac{(qe^{-it})^{n+1}-1}{qe^{-it}-1} - \frac{(qe^{-it})^{n-p+1}-1}{qe^{-it}-1}\Big]$$

$$- \frac{q^2}{2}\Big[\frac{(qe^{it})^{n}-1}{qe^{it}-1} - \frac{(qe^{it})^{n-p}-1}{qe^{it}-1}\Big] - \frac{q^2}{2}\Big[\frac{(qe^{-it})^{n}-1}{qe^{-it}-1} - \frac{(qe^{-it})^{n-p}-1}{qe^{-it}-1}\Big].$$

According to [15, p. 124] we denote

$$\Gamma(t) = (1 - 2q\cos t + q^2)^{-1}. \tag{13}$$

Now we have

$$\Sigma_1(t) = \big(q^{n+2}\cos nt - q^{n+1}\cos(n+1)t - q^{n-p+2}\cos(n-p)t + q^{n-p+1}\cos(n-p+1)t$$

$$- q^2(q^{n+1}\cos(n-1)t - q^n\cos nt - q^{n-p+1}\cos(n-p-1)t + q^{n-p}\cos(n-p)t)\big)\Gamma(t)$$

$$= \big(2q^{n+2}\cos nt - 2q^{n-p+2}\cos(n-p)t - q^{n+1}\cos(n+1)t + q^{n-p+1}\cos(n-p+1)t$$

$$- q^{n+3}\cos(n-1)t + q^{n-p+3}\cos(n-p-1)t\big)\Gamma(t)$$

$$= \big((2q^{n+2}\cos nt - q^{n+1}\cos(n+1)t - q^{n+3}\cos(n-1)t) - (2q^{n-p+2}\cos(n-p)t$$

$$- q^{n-p+1}\cos(n-p+1)t - q^{n-p+3}\cos(n-p-1)t)\big)\Gamma(t). \tag{14}$$

Doing elementary transformation of the term in brackets on the right part of equality (14) we have

$$2q^{n+2}\cos nt - q^{n+1}\cos(n+1)t - q^{n+3}\cos(n-1)t$$

$$= q^{n+1}\big((2q - \cos t - q^2\cos t)\cos nt + (\sin t - q^2\sin t)\sin nt\big), \tag{15}$$

$$2q^{n-p+2}\cos(n-p)t - q^{n-p+1}\cos(n-p+1)t - q^{n-p+3}\cos(n-p-1)t$$

$$= q^{n-p+1}\big((2q - \cos t - q^2\cos t)\cos(n-p)t + (\sin t - q^2\sin t)\sin(n-p)t\big). \tag{16}$$

Comparing (13)–(16) we obtain

$$\Sigma_1(t) = \Big[q^{n+1}\big((2q-\cos t-q^2\cos t)\cos nt+(\sin t-q^2\sin t)\sin nt\big)$$
$$-q^{n-p+1}\big((2q-\cos t-q^2\cos t)\cos(n-p)t$$
$$+(\sin t-q^2\sin t)\sin(n-p)t\big)\Big](1-2q\cos t+q^2)^{-1}. \tag{17}$$

Analogously, we may find

$$\Sigma_2(t) = \Big[q^{n+1}\big((q^2\sin t-\sin t)\cos nt+(2q-\cos t-q^2\cos t)\sin nt\big)$$
$$-q^{n-p+1}\big((q^2\sin t-\sin t)\cos(n-p)t$$
$$+(2q-\cos t-q^2\cos t)\sin(n-p)t\big)\Big](1-2q\cos t+q^2)^{-1}. \tag{18}$$

Respecting the last relation we may the equality (12) write in the following way

$$\frac{1}{p_i}\sum_{k_i=n_i-p_i}^{n_i-1} q_i^{k_i}h_{k_i}^{\beta_i}(t_i) = \frac{q_i^{n_i-p_i+1}}{p_i}\Big[(q_i^2\cos t_i-2q_i+\cos t_i)\cos\Big((n_i-p_i)t_i+\frac{\beta_i\pi}{2}\Big)$$
$$+(q_i^2\sin t_i-\sin t_i)\sin\Big((n_i-p_i)t_i+\frac{\beta_i\pi}{2}\Big)\Big](1-2q_i\cos t_i+q_i^2)^{-2}$$
$$-\frac{q_i^{n_i+1}}{p_i}\Big[(q_i^2\cos t_i-2q_i+\cos t_i)\cos\Big(n_it_i+\frac{\beta_i\pi}{2}\Big)$$
$$+(q_i^2\sin t_i-\sin t_i)\sin\Big(n_it_i+\frac{\beta_i\pi}{2}\Big)\Big](1-2q_i\cos t_i+q_i^2)^{-2}. \tag{19}$$

Analogously,

$$\frac{1}{p_i}\sum_{k_i=n_i-p_i}^{n_i-1} Q_i^{k_i}H_{k_i}^{\beta_i^\star}(t_i) =$$
$$=\frac{Q_i^{n_i-p_i+1}}{p_i}\Big[(Q_i^2\cos t_i-2Q_i+\cos t_i)\cos\Big((n_i-p_i)t_i+\frac{\beta_i^\star\pi}{2}\Big)$$
$$+(Q_i^2\sin t_i-\sin t_i)\sin\Big((n_i-p_i)t_i+\frac{\beta_i^\star\pi}{2}\Big)\Big](1-2Q_i\cos t_i+Q_i^2)^{-2}$$
$$-\frac{Q_i^{n_i+1}}{p_i}\Big[(Q_i^2\cos t_i-2Q_i+\cos t_i)\cos\Big(n_it_i+\frac{\beta_i^\star\pi}{2}\Big)$$
$$+(Q_i^2\sin t_i-\sin t_i)\sin\Big(n_it_i+\frac{\beta_i^\star\pi}{2}\Big)\Big](1-2Q_i\cos t_i+Q_i^2)^{-2}. \tag{20}$$

Considering the condition

$$\operatorname*{ess\,sup}_{\vec{x}\in T^m}|f^{\overline{\Psi}_\mu}(\vec{x})|\le 1,\quad \mu\subset\overline{m},\quad f\in C^{m\alpha}_{\beta,\infty}$$

and equalities (11), (19), (20) we have the coretness the theorem. □

3 Conclusion

Using the relation (9) we can obtain an asymptotic equality for upper bounds of the deviations of the de la Vallee Poussin right-angled sums taken over classes of periodic functions of several variables with a high smoothness.

References

[1] Stepanec, A. I., Pachulia, N. L.: *Multiple Fourier sums on the sets of* (ψ, β)*-differentiable functions.* Ukrainian Math. J. **43**, 4 (1991), 545–555 (in Russian).

[2] Lassuria, R. A.: *Multiple Fourier sums on the sets of* $\overline{\psi}$*-differentiable functions.* Ukrainian Math. J. **55**, 7 (2003), 911–918 (in Russian).

[3] Zaderey, P. V.: *Integral presentations of deviations of linear means of Fourier series on the classes of differentiable periodic functions of two variables.* Some problems of the theory of functions: collection of scientific works, Institute of Mathematics, Ukrainian Academy of Sciences, Kiev, 1985, 16–28 (in Russian).

[4] Stepanec, A. I.: Uniform approximation by trigonometric polynomials. *Nauk. Dumka, Kiev,* 1981 (in Russian).

[5] Stepanec, A. I.: *Approximation of some classes of periodic functions two variables by Fourier sums.* Ukrainian Math. J. **25**, 5 (1973), 599–609 (in Russian).

[6] Nikol'skii, S. M.: *Approximation of the functions by trigonometric polynomials in the mean.* News of Acad. of Sc. USSR **10**, 3 (1946), 207–256 (in Russian).

[7] Stechkin, S. B.: *Estimation of the remainder of Fourier series for the differentiable functions.* Works of Math. Inst. Acad. of Sc. USSR **145** (1980), 126–151 (in Russian).

[8] Stepanec, A. I.: *Approximation by Fourier sums of de la Poussin integrals of continuous functions.* Lect. of Rus. Acad. of Sc. **373**, 2 (2000), 171–173 (in Russian).

[9] Rukasov, V. I., Chaichenko, S. O.: *Approximation of the classes of analytical functions by de la Vallee-Poussin sums.* Ukrainian Math. J. **55**, 6 (2003), 575–590.

[10] Rukasov, V. I., Chaichenko, S. O.: *Approximation of continuous functions by de la Vallee-Poussin operators.* Ukrainian Math. J. **55**, 3 (2003), 498–511.

[11] Rukasov, V. I., Novikov, O. A.: *Approximation of analytical functions by de la Vallee Poussin sums. Fourier series: Theory and Applications.* Works of the Institute of Mathematics, Ukrainian Academy of Sciences, Kiev, 1998, 228–241 (in Russian).

[12] Stepanec A. I., Rukasov V. I., Chaichehko S. O.: *Approximation by de la Vallee Poussin sums.* Works of the Institute of Mathematics, Ukrainian Academy of Sciences **68**, 2007, 368 pp. (in Russian).

[13] Rukasov, V. R., Novikov, O. A., Velichko, V. E., Rovenska, O. G., Bodraya, V. I.: *Approximation of the periodic functions of many variables with a high smoothness by Fourier right-angled sums.* Works of the Institute of Mathematics and Mechanics, Ukrainian Academy of Sciences, 2008, 163–170 (in Russian).

[14] Rukasov, V. I., Novikov, O. A., Bodraya, V. I.: *Approximation of the classes of functions of two variables with a high smoothness by the right-angled linear means of Fourier series. Problems of the aproximation of the functions theory and closely related concepts.* Works of the Institute of Mathematics, Ukrainian Academy of Sciences **4**, 1 (2007), 270–283 (in Russian).

[15] Stepanec, A. I.: Classification and approximation of periodic functions. *Nauk. Dumka, Kiev,* 1987 (in Russian).

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 139–152

Convergence Theorems for a Finite Family of Nonexpansive and Asymptotically Nonexpansive Mappings*

Kittipong SITTHIKUL [1], Satit SAEJUNG [2] **

Department of Mathematics, Khon Kaen University,
Khon Kaen 40002, Thailand
[1] *e-mail: sittikul_n@hotmail.com*
[2] *e-mail: saejung@kku.ac.th*

(Received June 3, 2008)

Abstract

In this paper, weak and strong convergence of finite step iteration sequences to a common fixed point for a pair of a finite family of nonexpansive mappings and a finite family of asymptotically nonexpansive mappings in a nonempty closed convex subset of uniformly convex Banach spaces are presented.

Key words: Nonexpansive mapping, asymptotically nonexpansive mapping, common fixed point, finite-step iterative sequence.

2000 Mathematics Subject Classification: 47H09, 47H10

1 Introduction

The class of asymptotically nonexpansive mappings which is an important generalization of that of nonexpansive mappings was introduced by Goebel and Kirk [6]. Iteration processes for nonexpansive and asymptotically nonexpansive mappings in Banach spaces including Mann [11] and Ishikawa [8] iteration processes have been studied extensively by many authors (see [2, 7, 14, 15, 16, 17]).

Recently, Xu and Noor [19] introduced and studied a three-step scheme to approximate fixed points of asymptotically nonexpansive mappings in Banach space. Cho et al. [3] extended the work of Xu and Noor [19] to the three-step

*Supported by the Faculty of Science, Khon Kaen University.
**Correspondence should be addressed to Satit Saejung, saejung@kku.ac.th.

iterative scheme with errors in a Banach space and gave weak and strong convergence theorems for asymptotically nonexpansive mappings. Chidume and Ali [2] considered the multi-step scheme for a finite family of asymptotically nonexpansive mappings and gave weak convergence theorems for this scheme in a uniformly convex Banach space whose the dual space has the Kadec–Klee property. They also proved a strong convergence theorem under some appropriate conditions on a finite family of asymptotically nonexpansive mappings. Liu et al. (see [9] and [10]) established a new method with respect to a pair of nonexpansive and asymptotically nonexpansive mappings. The results in [9] and [10] generalize, improve and unify many known results due to many authors. Moreover, they also gave an example to demonstrate that their results are substantial generalizations and many previous known results are not applicable in this case.

Inspired by the above works, in this paper, a multi-step iteration scheme for a finite family of nonexpansive and asymptotically nonexpansive mappings is introduced and strong and weak convergence theorems of this scheme to common fixed point of nonexpansive and asymptotically nonexpansive mappings are proved. The weak convergence theorem is proved in a uniformly convex Banach space whose dual has the Kadec–Klee property. It is worth mentioning that there are uniformly convex Banach spaces, which have neither a Fréchet differentiable norm nor Opial property; however, their dual does have the Kadec–Klee property (see [5, Example 3.1]). Hence our results are different from [9] and [10] and the proofs are of independent interest.

2 Preliminaries

Let K be a nonempty subset of a real Banach space E and $T\colon K \to K$ be a mapping with the fixed point set $F(T)$, i.e., $F(T) = \{x \in K\colon x = Tx\}$. In this paper, we write $x_n \to x$ (resp. $x_n \rightharpoonup x$) if x_n converges strongly (resp. weakly) to x.

Definition 1 A mapping $T\colon K \to K$ is said to be

1. *asymptotically nonexpansive* if there exists a sequence $\{k_n\} \subset [1, \infty)$ with $\lim_{n\to\infty} k_n = 1$ such that $\|T^n x - T^n y\| \leq k_n \|x - y\|$ for all $x, y \in K$ and $n \geq 1$;

2. *nonexpansive* if $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in K$;

3. *Lipschitzian (with a Lipschitz constant L)* if $\|Tx - Ty\| \leq L\|x - y\|$ for all $x, y \in K$;

4. *demi-closed at a point $p \in K$* if whenever $\{x_n\}$ is a sequence in K which converges weakly to a point $x \in K$ and $\{Tx_n\}$ converges strongly to p, it follows that $Tx = p$.

Definition 2 [4] A norm on a Banach space E is *uniformly convex* (or simply, E is uniformly convex) if for all $\{x_n\}, \{y_n\} \subset \{z \in E\colon \|z\| = 1\}$ such that $\left\|\frac{x_n+y_n}{2}\right\| \to 1$, we have $\|x_n - y_n\| \to 0$.

Let K be a nonempty subset of a Banach space E. Let $S_1, S_2, \ldots, S_N\colon K \to K$ be N nonexpansive mappings, $T_1, T_2, \ldots, T_N\colon K \to K$ be N asymptotically nonexpansive mappings. Then the sequence $\{x_n\}$ defined by

$$\left.\begin{aligned} x_1 &\in K, \\ x_n^{(0)} &= x_n, \\ x_n^{(1)} &= a_n^{(1)} T_1^n x_n^{(0)} + (1 - a_n^{(1)}) S_1 x_n, \\ x_n^{(2)} &= a_n^{(2)} T_2^n x_n^{(1)} + (1 - a_n^{(2)}) S_2 x_n, \\ &\vdots \\ x_n^{(N-1)} &= a_n^{(N-1)} T_{N-1}^n x_n^{(N-2)} + (1 - a_n^{(N-1)}) S_{N-1} x_n, \\ x_n^{(N)} &= a_n^{(N)} T_N^n x_n^{(N-1)} + (1 - a_n^{(N)}) S_N x_n, \\ x_{n+1} &= x_n^{(N)}, \quad n \geq 1, \end{aligned}\right\} \tag{1}$$

where $\{a_n^{(i)}\}_{n=1}^{\infty} \subset [0,1]$, $i = 1, 2, \ldots, N$. An example of such iterations can be found in [9] and [10].

The purpose of this paper is to study the weak and strong convergences of finite-step iteration sequence $\{x_n\}$ defined by (1) to a common fixed point of a finite family of nonexpansive mappings and a finite family of asymptotically nonexpansive mappings in a uniformly convex Banach space.

The following lemmas are our main tool for proving the results.

Lemma 1 ([7]) *Let E be a uniformly convex Banach space and K be a nonempty closed convex subset of E. If $T\colon K \to K$ is an asymptotically nonexpansive mapping, then $I - T$ is demiclosed at zero.*

Lemma 2 *Let E be a uniformly convex Banach space, $\{x_n\}$ and $\{y_n\}$ be sequences in E. Suppose that there is $\delta > 0$ such that $\delta \leq t_n \leq 1 - \delta$ for all $n \in \mathbb{N}$. If $\limsup_{n\to\infty} \|x_n\| \leq a$, $\limsup_{n\to\infty} \|y_n\| \leq a$ and $\lim_{n\to\infty} \|t_n x_n + (1 - t_n) y_n\| = a$ for some $a \geq 0$, then $\lim_{n\to\infty} \|x_n - y_n\| = 0$. Moreover, $\lim_{n\to\infty} \|x_n\| = \lim_{n\to\infty} \|y_n\| = a$.*

Proof The first assertion follows from [15]. It suffices to prove that

$$\liminf_{n\to\infty} \|x_n\| \geq a.$$

In fact, this follows since

$$a = \lim_{n\to\infty} \|t_n x_n + (1 - t_n) y_n\| = \lim_{n\to\infty} \|x_n + (1 - t_n)(y_n - x_n)\|.$$

This finishes the proof. □

Lemma 3 ([13]) *Let $\{a_n\}$ and $\{b_n\}$ be sequences of nonnegative numbers satisfying the inequality $a_{n+1} \leq (1+b_n)a_n$, for all $n \geq 1$. If $\sum_{n=1}^{\infty} b_n < \infty$, then $\lim_{n\to\infty} a_n$ exists. In particular, if $\{a_n\}$ has a subsequence which converges to zero, then $\lim_{n\to\infty} a_n = 0$.*

Lemma 4 ([5]) *Let E be a reflexive Banach space such that its dual E^* has the Kadec–Klee property. Let $\{x_n\}$ be a bounded sequence in E and $p, q \in \omega_w(x_n)$, where $\omega_w(x_n)$ denotes the set of all weak cluster points of the sequence $\{x_n\}$. Suppose that $\lim_{n\to\infty} \|tx_n + (1-t)p - q\|$ exists for all $t \in [0,1]$. Then $p = q$.*

Lemma 5 ([5]) *Let K be a convex subset of a uniformly convex Banach space E. Then there exists a strictly continuous convex function $\phi\colon \mathbb{R}^+ \to \mathbb{R}^+$ with $\phi(0) = 0$ such that for each Lipschitzian mapping $T\colon K \to K$ with a Lipschitz constant L,*

$$\|tTx + (1-t)Ty - T(tx + (1-t)y)\| \leq L\phi^{-1}(\|x-y\| - \frac{1}{L}\|Tx - Ty\|)$$

for all $x, y \in K$ and all $0 < t < 1$.

Proposition 1 ([20]) *Let K be a nonempty subset of a Banach space E and $T_1, T_2 \ldots, T_N\colon K \to K$ be N asymptotically nonexpansive mappings. Then there exists a sequence $\{k_n\} \subset [1,\infty)$ such that $\lim_{n\to\infty} k_n = 1$ and*

$$\|T_i^n x - T_i^n y\| \leq k_n \|x - y\| \tag{2}$$

for all $x, y \in K$, $n \geq 1$ and $i = 1, 2, \ldots, N$.

From now on, we will assume that N asymptotically nonexpansive mappings $T_1, T_2, \ldots, T_N\colon K \to K$ share the same sequence $\{k_n\} \subset [1,\infty)$ as mentioned in the preceding proposition.

3 Technical Lemmas

Lemma 6 *Let K be a nonempty convex subset of a real Banach space E. Let $S_1, S_2, \ldots, S_N\colon K \to K$ be nonexpansive mappings, $T_1, T_2, \ldots, T_N\colon K \to K$ be asymptotically nonexpansive mappings with the sequence $\{k_n\}$ and suppose that $F = \bigcap_{i=1}^{N} F(S_i) \cap F(T_i) \neq \varnothing$. If*

$$\sum_{n=1}^{\infty} (k_n - 1) < \infty, \tag{3}$$

then $\lim_{n\to\infty} \|x_n - q\|$ exists for any $q \in F$, where $\{x_n\}$ is defined by the iterative scheme (1).

Proof Let $q \in F$. It follows from (1) that

$$\begin{aligned} \|x_n^{(1)} - q\| &\leq a_n^{(1)}\|T_1^n x_n - q\| + (1 - a_n^{(1)})\|S_1 x_n - q\| \\ &\leq a_n^{(1)} k_n \|x_n - q\| + (1 - a_n^{(1)})\|x_n - q\| \\ &\leq a_n^{(1)} k_n \|x_n - q\| + (1 - a_n^{(1)}) k_n \|x_n - q\| \\ &= k_n \|x_n - q\| \end{aligned} \tag{4}$$

and from (4), we have

$$\begin{aligned}\|x_n^{(2)} - q\| &\le a_n^{(2)}\|T_2^n x_n^{(1)} - q\| + (1 - a_n^{(2)})\|S_2 x_n - q\| \\ &\le a_n^{(2)} k_n \|x_n^{(1)} - q\| + (1 - a_n^{(2)})\|x_n - q\| \\ &\le a_n^{(2)} k_n^2 \|x_n - q\| + (1 - a_n^{(2)}) k_n^2 \|x_n - q\| \\ &= k_n^2 \|x_n - q\|. \end{aligned} \tag{5}$$

Continuing the above process, we get

$$\|x_n^{(i)} - q\| \le k_n^i \|x_n - q\| \quad \text{for all } n \ge 1,\ i = 1, 2, \ldots, N. \tag{6}$$

In particular,

$$\|x_{n+1} - q\| = \|x_n^{(N)} - q\| \le k_n^N \|x_n - q\| = (1 + (k_n^N - 1))\|x_n - q\|.$$

Notice that (3) holds (if and) only if

$$\sum_{n=1}^{\infty} (k_n^N - 1) < \infty. \tag{7}$$

By Lemma 3, we have $\lim_{n\to\infty} \|x_n - q\|$ exists. This completes the proof. □

Lemma 7 *Under the assumptions of Lemma 6 and suppose that there is $\delta > 0$ such that*

$$\delta \le a_n^{(i)} \le 1 - \delta \quad \textit{for all } n \ge 1, \quad i = 1, 2, \ldots, N. \tag{8}$$

If $\{x_n\}$ is defined by the iterative scheme (1), then

$$\lim_{n\to\infty} \|S_i x_n - T_i^n x_n^{(i-1)}\| = 0 \quad \textit{for all } i = 1, 2, \ldots, N. \tag{9}$$

Proof Let $q \in F$. By Lemma 6, we have

$$d = \lim_{n\to\infty} \|x_n - q\| \text{ exists.} \tag{10}$$

It follows from (6), (10) and $\lim_{n\to\infty} k_n = 1$ that

$$\limsup_{n\to\infty} \|x_n^{(N-1)} - q\| \le d, \tag{11}$$

and so

$$\limsup_{n\to\infty} \|T_N^n x_n^{(N-1)} - q\| \le d.$$

Also,

$$\limsup_{n\to\infty} \|S_N x_n - q\| \le d.$$

Further, from (10) and (1) we have

$$d = \lim_{n\to\infty} \|x_n^{(N)} - q\| = \lim_{n\to\infty} \|a_n^{(N)}(T_N^n x_n^{(N-1)} - q) + (1 - a_n^{(N)})(S_N x_n - q)\|.$$

Then, by Lemma 2, we get

$$\lim_{n\to\infty} \|S_N x_n - T_N^n x_n^{(N-1)}\| = \lim_{n\to\infty} \|(S_N x_n - q) - (T_N^n x_n^{(N-1)} - q)\| = 0,$$

and

$$\lim_{n\to\infty} \|T_N^n x_n^{(N-1)} - q\| = d.$$

Therefore,

$$\begin{aligned} d &= \liminf_{n\to\infty} \|T_N^n x_n^{(N-1)} - q\| \le \liminf_{n\to\infty} k_n \|x_n^{(N-1)} - q\| \\ &= \liminf_{n\to\infty} \|x_n^{(N-1)} - q\| \le \limsup_{n\to\infty} \|x_n^{(N-1)} - q\| \le d. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} \|x_n^{(N-1)} - q\| = d.$$

It follows from (6), (10) and $\lim_{n\to\infty} k_n = 1$ that

$$\limsup_{n\to\infty} \|x_n^{(N-2)} - q\| \le d, \tag{12}$$

and so

$$\limsup_{n\to\infty} \|T_{N-1}^n x_n^{(N-2)} - q\| \le d.$$

Also,

$$\limsup_{n\to\infty} \|S_{N-1} x_n - q\| \le d.$$

Further, from (10) and (1) we have

$$d = \lim_{n\to\infty} \|x_n^{(N-1)} - q\| = \lim_{n\to\infty} \|a_n^{(N-1)}(T_{N-1}^n x_n^{(N-2)} - q) + (1 - a_n^{(N-1)})(S_{N-1} x_n - q)\|.$$

Applying Lemma 2, we have

$$\lim_{n\to\infty} \|S_{N-1} x_n - T_{N-1}^n x_n^{(N-2)}\| = \lim_{n\to\infty} \|(S_{N-1} x_n - q) - (T_{N-1}^n x_n^{(N-2)} - q)\| = 0.$$

Continuing this in an obvious manner, we get (9) and this completes the proof. □

Lemma 8 *Under the assumptions of Lemma 6 and suppose that (8) holds. If* $\{x_n\}$ *is defined by the iterative scheme (1) and*

$$\lim_{n\to\infty} \|x_n - S_i x_n\| = 0 \quad \textit{for all } i = 1, 2, \ldots, N, \tag{13}$$

then $\lim_{n\to\infty} \|x_n - T_i x_n\| = 0$ *for all* $i = 1, 2, \ldots, N$.

Proof By Lemma 7, we have

$$\lim_{n\to\infty} \|S_i x_n - T_i^n x_n^{(i-1)}\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{14}$$

It follows from (13) that,

$$\lim_{n\to\infty} \|x_n - T_i^n x_n^{(i-1)}\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{15}$$

Next, from (1) we have

$$\|x_n - x_{n+1}\| \leq a_n^{(N)} \|x_n - T_N^n x_n^{(N-1)}\| + (1 - a_n^{(N)}) \|x_n - S_N x_n\|.$$

From (13) and (15), we have

$$\lim_{n\to\infty} \|x_n - x_{n+1}\| = 0. \tag{16}$$

Thus, we can estimate, using (1),

$$\begin{aligned} \|x_n - T_i^n x_n\| &\leq \|x_n - T_i^n x_n^{(i-1)}\| + \|T_i^n x_n^{(i-1)} - T_i^n x_n\| \\ &\leq \|x_n - T_i^n x_n^{(i-1)}\| + k_n \|x_n^{(i-1)} - x_n\| \\ &\leq \|x_n - T_i^n x_n^{(i-1)}\| + k_n a_n^{(i-1)} \|T_{i-1}^n x_n^{(i-2)} - x_n\| \\ &\quad + k_n (1 - a_n^{(i-1)}) \|S_{i-1} x_n - x_n\|. \end{aligned}$$

Hence,

$$\lim_{n\to\infty} \|x_n - T_i^n x_n\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{17}$$

It then follows from (16) and (17) that

$$\begin{aligned} \|x_n - T_i x_n\| &\leq \|x_n - x_{n+1}\| + \|x_{n+1} - T_i^{n+1} x_{n+1}\| + \|T_i^{n+1} x_{n+1} - T_i^{n+1} x_n\| \\ &\quad + \|T_i^{n+1} x_n - T_i x_n\| \\ &\leq \|x_n - x_{n+1}\| + \|x_{n+1} - T_i^{n+1} x_{n+1}\| + k_{n+1} \|x_{n+1} - x_n\| \\ &\quad + k_1 \|T_i^n x_n - x_n\| \\ &\leq (1 + k_{n+1}) \|x_n - x_{n+1}\| + \|x_{n+1} - T_i^{n+1} x_{n+1}\| + k_1 \|T_i^n x_n - x_n\| \end{aligned}$$

for $i = 1, 2, \ldots, N$. This implies that

$$\lim_{n\to\infty} \|x_n - T_i x_n\| = 0, \quad \text{for all } i = 1, 2, \ldots, N.$$

This completes the proof. □

Lemma 9 *Under the assumptions of Lemma 6 and suppose that (8) holds and that*

$$\|x - T_i y\| \leq \|S_i x - T_i y\| \quad \textit{for all } x, y \in K \textit{ and } i = 1, 2, \ldots, N. \tag{18}$$

If the sequence $\{x_n\}$ is defined by the iterative scheme (1), then

$$\lim_{n\to\infty} \|x_n - S_i x_n\| = \lim_{n\to\infty} \|x_n - T_i x_n\| = 0,$$

for all $i = 1, 2, \ldots, N$.

Proof We shall show that

$$\lim_{n\to\infty} \|x_n - S_i x_n\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{19}$$

By Lemma 7, we have

$$\lim_{n\to\infty} \|S_i x_n - T_i^n x_n^{(i-1)}\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{20}$$

It follows from (18) that

$$\lim_{n\to\infty} \|x_n - T_i^n x_n^{(i-1)}\| = 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{21}$$

Thus (19) follows. And Lemma 8 guarantees the second equality. □

4 Strong convergence theorems

A finite family $\{T_1, \ldots, T_N\}$ of mappings of K with

$$F = \bigcap_{i=1}^{N} F(T_i) \neq \emptyset$$

is said to satisfy condition (B) [2] if there exists a nondecreasing function $f\colon [0,\infty) \to [0,\infty)$ with $f(0) = 0$ and $f(r) > r$ for all $r \in (0,\infty)$ such that for all $x \in K$

$$\max_{1\le i\le N} \|x - T_i x\| \ge f(d(x, F)),$$

where $d(x, F) = \inf\{\|x - p\| : p \in F\}$.

Theorem 1 *Let K be a nonempty closed convex subset of a uniformly convex Banach space E. Let $S_1, S_2, \ldots, S_N\colon K \to K$ be nonexpansive mappings, $T_1, T_2, \ldots, T_N\colon K \to K$ be asymptotically nonexpansive mappings with the sequence $\{k_n\}$ and suppose that $F = \bigcap_{i=1}^{N} F(S_i) \cap F(T_i) \neq \varnothing$. Suppose that the family $\{S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N\}$ satisfies condition (B) and (3), (8), (18) hold. Then the sequence $\{x_n\}$ defined by (1) converges strongly to a common fixed point of $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N$.*

Proof We have

$$\|x_{n+1} - q\| = \|x_n^{(N)} - q\| \le (1 + (k_n^N - 1))\|x_n - q\| \quad \text{for all } q \in F.$$

Consequently,

$$d(x_{n+1}, F) \le (1 + (k_n^N - 1))d(x_n, F).$$

Applying Lemma 3 to the above inequality, we obtain that $\lim_{n\to\infty} d(x_n, F)$ exists. Also, by Lemma 9,

$$\lim_{n\to\infty} \|x_n - S_i x_n\| = \lim_{n\to\infty} \|x_n - T_i x_n\| = 0 \quad \text{for all } i = 1, 2, \ldots, N. \tag{22}$$

Since $\{S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N\}$ satisfies condition (B), we conclude that

$$\lim_{n\to\infty} d(x_n, F) = 0.$$

We now prove that $\{x_n\}$ is a Cauchy sequence in K. Let $\varepsilon > 0$. Then there exists a positive integer n_0 such that $d(x_{n_0}, F) < \frac{\varepsilon}{4}$. Find $p \in F$ such that $\|x_{n_0} - p\| < \frac{\varepsilon}{4}$. By Lemma 6, we see that $\lim_{n\to\infty} \|x_n - p\|$ exists and so $\{x_n - p\}$ is bounded. Then there is a constant $M > 0$ such that

$$\|x_n - p\| \le M \quad \text{for all } n \ge 1.$$

We now choose a positive integer n_1 such that

$$\sum_{j=n_1}^{\infty} (k_j^N - 1) < \frac{\varepsilon}{4M}.$$

Moreover, we have

$$\|x_{n+1} - p\| \le \|x_n - p\| + M(k_n^N - 1).$$

This implies that

$$\begin{aligned}
\|x_{n+m} - p\| &\le \|x_{n+m-1} - p\| + M(k_{n+m-1}^N - 1) \\
&\le \|x_{n+m-2} - p\| + M(k_{n+m-2}^N - 1) + M(k_{n+m-1}^N - 1) \\
&\le \|x_n - p\| + M \sum_{j=n}^{n+m-1} (k_j^N - 1) \qquad (23)
\end{aligned}$$

for all $n, m \ge 1$. From (23) it follows that, for all $n > n_1$ and $m \ge 1$,

$$\begin{aligned}
\|x_{n+m} - x_n\| &\le \|x_{n+m} - p\| + \|x_n - p\| \\
&\le 2\|x_{n_1} - p\| + M \sum_{j=n_1}^{n+m-1} (k_j^N - 1) + M \sum_{j=n_1}^{n-1} (k_j^N - 1) \\
&\le 2\|x_{n_1} - p\| + 2M \sum_{j=n_1}^{n+m-1} (k_j^N - 1) \\
&\le 2\|x_{n_1} - p\| + 2M \sum_{j=n_1}^{\infty} (k_j^N - 1) \\
&< 2\frac{\varepsilon}{4} + 2M\frac{\varepsilon}{4M} = \varepsilon.
\end{aligned}$$

Hence $\{x_n\}$ is a Cauchy sequence in K. In virtue of the completeness of K, we assume that $x_n \to p' \in K$ as $n \to \infty$. By the continuities of S_i and T_i and (22), we have $S_i p' = p' = T_i p'$ for all $i = 1, 2, \ldots, N$, so $p' \in F$. This completes the proof. □

5 Weak convergence theorems

Lemma 10 *Under the assumptions of Lemma 6 and suppose that (8) holds. Let $\{x_n\}$ be the sequence defined by (1). Then for all $u, v \in F$, the limit $\lim_{n\to\infty} \|tx_n + (1-t)u - v\|$ exists for all $t \in [0,1]$.*

Proof Since $\{x_n\}$ is bounded, there exists $R > 0$ such that $\{x_n\} \subset C := B_R(0) \cap K$. Then C is a nonempty closed convex bounded subset of E. Basically, we shall follow the idea of [17]. Let

$$a_n(t) = \|tx_n + (1-t)u - v\|, \quad \text{where } t \in [0,1].$$

Then $a_n(0) = \|u - v\|$, and from Lemma 6, $\lim_{n\to\infty} a_n(1) = \lim_{n\to\infty} \|x_n - v\|$ exists. We now assume that $t \in (0,1)$. Define $U_n \colon C \to C$ by

$$\begin{aligned}
x^{(1)} &= a_n^{(1)} T_1^n x + (1 - a_n^{(1)}) S_1 x, \quad x \in K \\
x^{(2)} &= a_n^{(2)} T_2^n x^{(1)} + (1 - a_n^{(2)}) S_2 x, \\
x^{(3)} &= a_n^{(3)} T_3^n x^{(2)} + (1 - a_n^{(3)}) S_3 x, \\
&\vdots \\
x^{(N-1)} &= a_n^{(N-1)} T_{N-1}^n x^{(N-2)} + (1 - a_n^{(N-1)}) S_{N-1} x, \\
U_n x &= a_n^{(N)} T_N^n x^{(N-1)} + (1 - a_n^{(N)}) S_N x.
\end{aligned}$$

Then

$$\|U_n x - U_n y\| \le k_n^N \|x - y\|.$$

Set

$$\begin{aligned}
W_{n,m} &= U_{n+m-1} \circ U_{n+m-2} \circ \cdots \circ U_n, \quad m \ge 1, \\
b_{n,m} &= \|W_{n,m}(tx_n + (1-t)u) - (tW_{n,m}x_n + (1-t)W_{n,m}u)\|.
\end{aligned}$$

Then observing that $W_{n,m}x_n = x_{n+m}$, we get

$$\begin{aligned}
a_{n+m}(t) &= \|tx_{n+m} + (1-t)u - v\| \\
&\le b_{n,m} + \|W_{n,m}(tx_n + (1-t)u) - v\| \\
&\le b_{n,m} + \left(\prod_{j=n}^{n+m-1} k_j^N \right) a_n(t) \\
&\le b_{n,m} + L_n a_n(t),
\end{aligned}$$

where $L_n = \prod_{j=n}^{\infty} k_j^N$. By Lemma 5 we have

$$\begin{aligned}
b_{n,m} &\le L_n \phi^{-1}(\|x_n - u\| - L_n^{-1} \|W_{n,m}x_n - u\|) \\
&\le L_n \phi^{-1}(\|x_n - u\| - \|x_{n+m} - u\| + (1 - L_n^{-1})d),
\end{aligned}$$

where $\phi\colon [0,\infty) \to [0,\infty)$ is a strictly increasing continuous function depending only on the diameter of K and $\phi(0) = 0$. Since $\lim_{n\to\infty} L_n = 1$, it follows from Lemma 6 that $\lim_{n,m\to\infty} b_{n,m} = 0$. Therefore,

$$\limsup_{m\to\infty} a_m(t) \le \lim_{n,m\to\infty} b_{n,m} + \liminf_{n\to\infty} L_n a_n(t) = \liminf_{n\to\infty} a_n(t).$$

This completes the proof. □

Recall that a Banach space E has the *Kadec–Klee property* if for every sequence $\{x_n\}$ in E, $x_n \rightharpoonup x$ and $\|x_n\| \to \|x\|$ it follows that $\|x_n - x\| \to 0$.

Theorem 2 *Let K be a nonempty closed convex subset of a uniformly convex Banach space E such that its dual E^* has the Kadec–Klee property. Let $S_1, S_2, \ldots, S_N\colon K \to K$ be nonexpansive mappings, $T_1, T_2, \ldots, T_N\colon K \to K$ be asymptotically nonexpansive mappings with the sequence $\{k_n\}$ and suppose that*

$$F = \bigcap_{i=1}^{N} F(S_i) \cap F(T_i) \neq \varnothing.$$

If (3), (8) and (18) hold, then the sequence $\{x_n\}$ defined by (1) converges weakly to a common fixed point of $S_1, S_2, \ldots, S_N$, $T_1, T_2, \ldots, T_N$.

Proof Let $q \in F$. Then by Lemma 6, $\lim_{n\to\infty} \|x_n - q\|$ exists. Since E is reflexive and $\{x_n\}$ is bounded sequence in K, there exists a subsequence $\{x_{n_j}\}$ of $\{x_n\}$ which converges weakly to some $p \in K$. Moreover $\lim_{j\to\infty} \|x_{n_j} - S_i x_{n_j}\| = 0$ and $\lim_{j\to\infty} \|x_{n_j} - T_i x_{n_j}\| = 0$ for all $i = 1, 2, \ldots, N$, by Lemma 9. From Lemma 1, we have that $(I - S_i)p = (I - T_i)p = 0$ for all $i = 1, 2, \ldots, N$. Thus, $p \in F$.

Now, we show that $\{x_n\}$ converges weakly to p. Suppose that $\{x_{n_k}\}$ is another subsequence of $\{x_n\}$ which converges weakly to some $p' \in K$. By the same method as above, we have $p' \in F$ and so $p, p' \in \omega_w(x_n)$. Then by Lemma 10,

$$\lim_{n\to\infty} \|tx_n + (1-t)p - p'\|$$

exists for all $t \in [0,1]$. Now, Lemma 4 guarantees that $p = p'$. As a result, the whole sequence $\{x_n\}$ converges weakly to p. This completes the proof. □

6 Some analogues and corollaries

With a little effort, we have the following analogues to Theorems 1 and 2.

Theorem 3 *Let K be a nonempty closed convex subset of a uniformly convex Banach space E. Let $S_1, S_2, \ldots, S_N\colon K \to K$ be nonexpansive mappings,*

$T_1, T_2, \ldots, T_N \colon K \to K$ be asymptotically nonexpansive mappings with the sequence $\{k_n\}$ and suppose that $\bigcap_{i=1}^{N} F(S_i) \cap F(T_i) \neq \varnothing$. Let $\{x_n\}$ be the sequence defined by

$$\left.\begin{aligned} x_1 &\in K, \\ x_n^{(0)} &= x_n, \\ x_n^{(1)} &= a_n^{(1)} T_1^n x_n^{(0)} + b_n^{(1)} S_1 x_n + c_n^{(1)} u_n^{(1)}, \\ x_n^{(2)} &= a_n^{(2)} T_2^n x_n^{(1)} + b_n^{(2)} S_2 x_n + c_n^{(2)} u_n^{(2)}, \\ &\vdots \\ x_n^{(N-1)} &= a_n^{(N-1)} T_{N-1}^n x_n^{(N-2)} + b_n^{(N-1)} S_{N-1} x_n + c_n^{(N-1)} u_n^{(N-1)}, \\ x_n^{(N)} &= a_n^{(N)} T_N^n x_n^{(N-1)} + b_n^{(N)} S_N x_n + c_n^{(N)} u_n^{(N)}, \\ x_{n+1} &= x_n^{(N)}, \quad n \geq 1, \end{aligned}\right\} \tag{24}$$

where $\{u_n^{(i)}\}$ are bounded sequences in K and $\{a_n^{(i)}\}_{n=1}^{\infty}$, $\{b_n^{(i)}\}_{n=1}^{\infty}$, $\{c_n^{(i)}\}_{n=1}^{\infty} \subset [0,1]$ such that $a_n^{(i)} + b_n^{(i)} + c_n^{(i)} = 1$ for all $i = 1, 2, \ldots, N$.

Suppose that $\sum_{n=1}^{\infty} c_n^{(i)} < \infty$ for all $i = 1, 2, \ldots, N$,

(i) $\sum_{n=1}^{\infty} (k_n - 1) < \infty$,

(ii) there is $\delta > 0$ such that $\delta \leq a_n^{(i)} \leq 1 - \delta$ for all $n \geq 1, i = 1, 2, \ldots, N$,

(iii) $\|x - T_i y\| \leq \|S_i x - T_i y\|$ *for all $x, y \in K$ and $i = 1, 2, \ldots, N$.*

(a) If the family $\{S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N\}$ satisfies condition (B), then $\{x_n\}$ converges strongly to a common fixed point of $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N$.

(b) If the dual E^* has the Kadec–Klee property, then $\{x_n\}$ converges weakly to a common fixed point of $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N$.

Remark

1. If, moreover $S_1 = S_2 = \cdots = S_N = S$, then by Lemma 7

$$\lim_{n \to \infty} \|S x_n - T_i x_n^{(i-1)}\| = 0 \quad \text{for all } i = 1, 2, \ldots, N.$$

   The assumption (iii) in Theorem 3 can be weakened by assuming that there is $i_0 \in \{1, 2, \ldots, N\}$ such that

$$\|x - T_{i_0} y\| \leq \|Sx - T_{i_0} y\| \quad \text{for all } x, y \in K.$$

2. If, moreover $S_1 = S_2 = \cdots = S_N = I$, then Theorems 2.3 and 2.9 of [18] become a corollary of Theorem 3.
3. Theorem 3 is not only an extension of [9] and [10] but also obtained under the different assumptions.

Theorem 4 *Let K be a nonempty closed convex subset of a uniformly convex Banach space E. Let $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N \colon K \to K$ be nonexpansive*

mappings and suppose that $\bigcap_{i=1}^{N} F(S_i) \cap F(T_i) \neq \varnothing$. *Let* $\{x_n\}$ *be the sequence defined by*

$$\left.\begin{aligned} x_1 &\in K, \\ x_n^{(0)} &= x_n, \\ x_n^{(1)} &= a_n^{(1)} T_1 x_n^{(0)} + b_n^{(1)} S_1 x_n + c_n^{(1)} u_n^{(1)}, \\ x_n^{(2)} &= a_n^{(2)} T_2 x_n^{(1)} + b_n^{(2)} S_2 x_n + c_n^{(2)} u_n^{(2)}, \\ &\vdots \\ x_n^{(N-1)} &= a_n^{(N-1)} T_{N-1} x_n^{(N-2)} + b_n^{(N-1)} S_{N-1} x_n + c_n^{(N-1)} u_n^{(N-1)}, \\ x_n^{(N)} &= a_n^{(N)} T_N x_n^{(N-1)} + b_n^{(N)} S_N x_n + c_n^{(N)} u_n^{(N)}, \\ x_{n+1} &= x_n^{(N)}, \quad n \geq 1, \end{aligned}\right\} \tag{25}$$

where $\{u_n^{(i)}\}$ *are bounded sequences in* K *and* $\{a_n^{(i)}\}_{n=1}^{\infty}$, $\{b_n^{(i)}\}_{n=1}^{\infty}$, $\{c_n^{(i)}\}_{n=1}^{\infty} \subset [0,1]$ *such that* $a_n^{(i)} + b_n^{(i)} + c_n^{(i)} = 1$ *for all* $i = 1, 2, \ldots, N$.

Suppose that $\sum_{n=1}^{\infty} c_n^{(i)} < \infty$ *for all* $i = 1, 2, \ldots, N$,

(i) there is $\delta > 0$ *such that* $\delta \leq a_n^{(i)} \leq 1 - \delta$ *for all* $n \geq 1, i = 1, 2, \ldots, N$,

(ii) $\|x - T_i y\| \leq \|S_i x - T_i y\|$ *for all* $x, y \in K$ *and* $i = 1, 2, \ldots, N$.

(a) If the family $\{S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N\}$ *satisfies condition (B), then* $\{x_n\}$ *converges strongly to a common fixed point of* $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N$.

(b) If the dual E^* *has the Kadec–Klee property, then* $\{x_n\}$ *converges weakly to a common fixed point of* $S_1, S_2, \ldots, S_N, T_1, T_2, \ldots, T_N$.

Acknowledgement The authors would like to thank the referee for the suggestions and comments on the manuscript.

References

[1] Chidume, C. E., Ali, B.: *Approximation of common fixed points for finite families of nonself asymptotically nonexpansive mappings in Banach spaces.* J. Math. Anal. Appl. **326**, 2 (2007), 960–973.

[2] Chidume, C. E., Ali, B.: *Weak and strong convergence theorems for finite families of asymptotically nonexpansive mappings in Banach spaces.* J. Math. Anal. Appl. **330**, 1 (2007), 377–387.

[3] Cho, Y. J., Zhou, H., Guo, G.: *Weak and strong convergence theorems for three-step iterations with errors for asymptotically nonexpansive mappings.* Comput. Math. Appl. **47**, 4-5 (2004), 707–717.

[4] Clarkson, J. A.: *Uniformly convex spaces.* Trans. Amer. Math. Soc. **40**, 3 (1936), 396–414.

[5] García-Falset, J., Kaczor, W., Kuczumow, T., Reich, S.: *Weak convergence theorems for asymptotically nonexpansive mappings and semigroups.* Nonlinear Anal., Ser. A: Theory Methods **43**, 3 (2001), 377–401.

[6] Goebel, K., Kirk, W. A.: *A fixed point theorem for asymptotically nonexpansive mappings.* Proc. Amer. Math. Soc. **35** (1972), 171–174.

[7] Gornicki, J.: *Weak and strong convergence theorems for asymptotically nonexpansive mappings in uniformly convex Banach spaces.* Comment. Math. Univ. Carolin **30** (1989), 249–252.

[8] Ishikawa, S.: *Fixed points by new iteration method.* Proc. Amer. Math. Soc. **44** (1974), 147–150.

[9] Liu, Z., Agarwal, R. P., Feng, C., Kang, S. M.: *Weak and strong convergence theorems of common fixed points for a pair of nonexpansive and asymptotically nonexpansive mappings.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **44** (2005), 83–96.

[10] Liu, Z., Feng, C., Ume, J. S., Kang, S. M.: *Weak and strong convergence for common fixed points of a pair of nonexpansive and asymptotically nonexpansive mappings.* Taiwanese J. Math. **11**, 1 (2007), 27–42.

[11] Mann, W. R.: *Mean value methods in iteration.* Proc. Amer. Math. Soc. **4** (1953), 506–510.

[12] Opial, Z.: *Weak convergence of the sequence of successive approximations for nonexpansive mappings.* Bull. Amer. Math. Soc. **73** (1967), 591–597.

[13] Osilike, M. O., Aniagbosor, S. C.: *Weak and strong convergence theorems for fixed points of asymptotically nonexpansive mappings.* Math. Comput. Modelling **32**, 10 (2000), 1181–1191.

[14] Schu, J.: *Iterative construction of fixed points of asymptotically nonexpansive mappings.* J. Math. Anal. Appl. **158**, 2 (1991), 407–413.

[15] Schu, J.: *Weak and strong convergence to fixed points of asymptotically nonexpansive mappings.* Bull. Austral. Math. Soc. **43**, 1 (1991), 153–159.

[16] Senter, H. F., Dotson, W. G., Jr: *Approximating fixed points of nonexpansive mappings.* Proc. Amer. Math. Soc. **44** (1974), 375–380.

[17] Tan, K. K., Xu, H. K.: *Fixed point iteration processes for asymptotically nonexpansive mappings.* Proc. Amer. Math. Soc. **122** (1994), 733–739.

[18] Thakur, B. S., Jung, J. S.: *Weak and strong convergence of multistep iteration for finite family of asymptotically nonexpansive mappings.* Fixed Point Theory Appl., Art. ID 31056 (2007), 15 pp.

[19] Xu, B., Noor, M. A.: *Fixed-point iterations for asymptotically nonexpansive mappings in Banach spaces.* J. Math. Anal. Appl. **267** (2002), 444–453.

[20] Zhou, Y. Y., Chang, S. S.: *Convergence of implicit iterative process for a finite family of asymptotically nonexpansive mappings in Banach spaces.* Numer. Funct. Anal. and Optimiz. **23** (2002), 911–921.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 153–156

Congruences in Ordered Sets and LU Compatible Equivalences

VÁCLAV SNÁŠEL [1], MAREK JUKL [2]

[1] *Department of Computer Science, VŠB – Technical University of Ostrava*
Ostrava, Czech Republic
e-mail: vaclav.snasel@vsb.cz

[2] *Department of Algebra and Geometry, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: jukl@upol.cz

(Received October 8, 2008)

Abstract

A concept of equivalence preserving upper and lower bounds in a poset P is introduced. If P is a lattice, this concept coincides with the notion of lattice congruence.

Key words: Ordered set, morphism, *LU* compatible equivalence.

2000 Mathematics Subject Classification: 06A06, 06B10

There are various concepts of a congruence relation in ordered sets. All of them define a congruence as an equivalence relation whose classes are convex subsets. However, this concept is too weak, namely the quotient set by such an equivalence need not be an ordered set. Hence, in the definitions additional conditions are usually required. We can mention e.g. the approaches by M. Kolibiar [2, 3], I. Chajda, V. Snášel [1], J. Lihová, A. Haviar [4] and R. Halaš [5], [6]. A natural assumption for a congruence on an ordered set is that if this set is a lattice then the notion of a congruence has to coincide with the lattice one. The aim of our paper is to introduce a concept of *LU* compatible equivalence in an ordered set satisfying all the foregoing assumptions which, moreover, corresponds to the concept of morphism preserving upper and lower bounds.

Let $A \neq \emptyset$ be a set and let $\leq$ be a partial order on A. For a subset $B \subseteq A$, we denote the set of all lower or upper bounds of B in A with respect to $\leq$ by $L_A(B)$ or $U_A(B)$, respectively, i.e.:

$$L_A(B) = \{x \in A;\ x \leq a \text{ for all } a \in B\}$$
$$U_A(B) = \{x \in A;\ a \leq x \text{ for all } a \in B\}.$$

If there is no danger of misunderstanding, the subscript A will be omitted and we will write $U(B)$ or $L(B)$ only.

We adopt the notation $U(B,C) = U(B \cup C)$ and $L(B,C) = L(B \cup C)$. If $B = \{b_1, b_2, \ldots, b_n\}$ is finite, we will write briefly $U(B) = U(b_1, b_2, \ldots, b_n)$, dually for $L(B)$.

Remark that if $B \subseteq C \subseteq A$ then $U(B) \supseteq U(C)$ and $L(B) \supseteq L(C)$.

Definition 1 [1] An equivalence Θ on an ordered set P is called a *congruence* if either $\Theta = P \times P$ or it satisfies:

(i) $[a]_\Theta$ is a convex subset of P for all $a \in P$;

(ii) for every $x, y \in [a]_\Theta$ there exist $u, v \in [a]_\Theta$ such that $u \leq x \leq v$ and $u \leq y \leq v$;

(iii) if $u \leq x$, $u \leq y$ and $u\ \Theta\ x$ then there exists $v \in P$ with $x \leq v$, $y \leq v$ and $y\ \Theta\ v$; if $x \leq v$, $y \leq v$ and $v\ \Theta\ y$ then there exists $u \in P$ with $u \leq x$, $u \leq y$ and $u\ \Theta\ x$.

Of course, the identity relation on P is a congruence on P.

It was already shown in [1] that the quotient set by a congruence is an ordered set again.

Proposition 1 [1] *Let P be an ordered set and Θ be a congruence on P. Then the quotient relation defined on P/Θ by setting $[a]_\Theta \leq_{/\Theta} [b]_\Theta$ iff there exist $x \in [a]_\Theta, y \in [b]_\Theta$ with $x \leq y$ is an order on P/Θ.*

In the following, for any $A \subseteq P$ denote $[A]_\Theta = \{[a]_\Theta;\ a \in A\}$.

Corollary 1 [1] *Let P be an ordered set and Θ be an equivalence on P. Then Θ is a congruence on P if and only if*

(1) P/Θ is an ordered set (with the order $\leq_{/\Theta}$);

(2) $[L_P(x,y)]_\Theta = L_{P/\Theta}([x]_\Theta, [y]_\Theta)$ and $[U_P(x,y)]_\Theta = U_{P/\Theta}([x]_\Theta, [y]_\Theta)$

for every x, y of P.

Definition 2 Let $(P, \leq)$ be a an ordered set. An equivalence Θ on P is called $LU-compatible$ if it satisfies the condition (2) of Corollary 1.

Lemma 1 *If Θ is an LU compatible equivalence then the following holds:*

(1) each block is directed,

(2) the condition (iii) of definition 1 is satisfied,

(3) $\leq_{/\Theta}$ is transitive.

Proof (1) Let $a, b \in [x]_\Theta$. Then $[a]_\Theta = [x]_\Theta = [b]_\Theta$ whence

$$[L(a,b)]_\Theta = L([a]_\Theta, [b]_\Theta) = L([x]_\Theta) \neq \emptyset.$$

(2) Let $u \leq x$, $u \leq y$ and $u\ \Theta\ x$, i.e. $[y]_\Theta = [v]_\Theta$ for some $v \in U(x,y)$. Since Θ is an LU-compatible equivalence, we have

$$[U(x,y)]_\Theta = U([x]_\Theta, [y]_\Theta) = U([u]_\Theta, [y]_\Theta) = [U(u,y)]_\Theta = [U(y)]_\Theta.$$

This shows the first part of (iii) of Definition 1. The rest can be shown analogously.

(3) The proof is analogous to that used in [1]. □

Theorem 1 *Let $(P, \leq)$ be a an ordered set and Θ be an LU compatible equivalence. Then $(P/\Theta, \leq_{/\Theta})$ is an ordered set if and only if each block of Θ is convex.*

Proof It is easy to show that $\leq_{/\Theta}$ is reflexive. The transitivity of $\leq_{/\Theta}$ follows directly from Lemma 1.

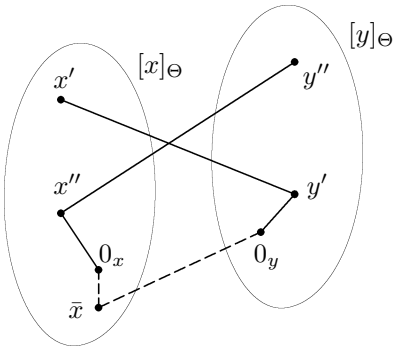
We show that $\leq_{/\Theta}$ is antisymmetric. Let $[x]_\Theta \leq_{/\Theta} [y]_\Theta$ and $[y]_\Theta \leq_{/\Theta} [x]_\Theta$. Then there exist $x', x'' \in [x]_\Theta$ and $y', y'' \in [y]_\Theta$ such that $x' \geq y'$ and $x'' \leq y''$. From Lemma 1 there exist $\bar{y} \in [y]_\Theta$ and $\bar{x} \in [x]_\Theta$ such that $y', y'' \leq \bar{y}$, $\bar{x} \leq x''$ and $\bar{x} \leq y'$. We have $\bar{x} \leq y' \leq x'$. Applying convexity we conclude $[x]_\Theta = [y]_\Theta$.

Conversely, let $(P/\Theta, \leq_{/\Theta})$ be a an ordered set and assume $x \leq y \leq z$ with $[x]_\Theta = [z]_\Theta$. Then $[x]_\Theta \leq_{/\Theta} [y]_\Theta$, $[y]_\Theta \leq_{/\Theta} [z]_\Theta = [x]_\Theta$, thus due to antisymmetry of $\leq_{/\Theta}$ we have $[x]_\Theta = [y]_\Theta$. □

Theorem 2 *Let $(P, \leq)$ be a ordered set and let Θ be an LU compatible equivalence. If every equivalence class of Θ has the least element, then $(P/\Theta, \leq_{/\Theta})$ is an ordered set.*

Proof It is easy to see that $\leq_{/\Theta}$ is reflexive, transitivity of $\leq_{/\Theta}$ follows directly by Lemma 1.

To prove its antisymmetry, denote by 0_x the least element of an arbitrary block $[x]_\Theta$. Let $[x]_\Theta \leq_{/\Theta} [y]_\Theta$ and $[y]_\Theta \leq_{/\Theta} [x]_\Theta$. Then there exist $x', x'' \in [x]_\Theta$ and $y', y'' \in [y]_\Theta$ such that $x' \geq y'$ and $x'' \leq y''$.



Now $0_x \leq y''$ and $0_y \leq y''$, hence by Lemma 1 there exists $\bar{x} \in [x]_\Theta$ with $\bar{x} \leq 0_x$ and $\bar{x} \leq 0_y$. Since 0_x is the least element of $[x]_\Theta$, we have $\bar{x} = 0_x$. This shows $0_x \leq 0_y$. Analogously we prove $0_y \leq 0_x$, consequently $0_x = 0_y$, which finally yields $[x]_\Theta = [y]_\Theta$. □

References

[1] Chajda, I., Snášel, V.: *Congruences in ordered sets.* Math. Bohem. **123**, 1 (1998), 95–100.

[2] Kolibiar, M.: *Congruence relations and direct decomposition of ordered sets.* Acta Sci. Math. (Szeged) **51** (1987), 129–135.

[3] Kolibiar, M.: *Congruence relations and direct decomposition of ordered sets II.* Contributions to General Algebra **6** (1988), 167–171.

[4] Haviar, A., Lihová, J.: *Varieties of posets.* Order **22**, 4 (2005), 343–356.

[5] Halaš, R.: *Congruences on posets.* Contributions to General Algebra **12** (2000), 195–210.

[6] Halaš, R., Hort, D.: *A characterization of 1,2,3,4-endomorphisms of posets.* Czech. Math. J. **53**, 128 (2003), 213–221.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 157–170

Metrizability of Connections on Two-Manifolds*

Alena VANŽUROVÁ [1], Petra ŽÁČKOVÁ [2]

[1] *Department of Algebra and Geometry, Faculty of Science, Palacký University*
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: alena.vanzurova@upol.cz

[2] *Department of Mathematics and Didactics of Mathematics*
Technical University Liberec, Studentská 2, 46117 Liberec, Czech Republic
e-mail: petrazacek@centrum.cz

(Received May 6, 2009)

Abstract

We contribute to the reverse of the Fundamental Theorem of Riemannian geometry: if a symmetric linear connection on a manifold is given, find non-degenerate metrics compatible with the connection (locally or globally) if there are any. The problem is not easy in general. For nowhere flat 2-manifolds, we formulate necessary and sufficient metrizability conditions. In the favourable case, we describe all compatible metrics in terms of the Ricci tensor. We propose an application in the calculus of variations.

Key words: Manifold, linear connection, metric connection, pseudo-Riemannian geometry.

2000 Mathematics Subject Classification: 53B05, 53B20

1 Preliminaries—affine differential geometry

Recall briefly some well-known facts from affine and metric differential geometry. Let M be an n-dimensional smooth manifold ("smooth" always means of the class C^∞), T_xM the tangent space at $x \in M$, and let $\pi\colon TM \to M$ denote the tangent vector bundle of M. $\mathcal{F}(M) = C^\infty(M)$ denotes the ring of all smooth functions on M, $\mathcal{X}(M)$ the $C^\infty(M)$-module of all smooth vector fields on M (which can be viewed as sections of the projection π), and $\Lambda(M)$ the exterior algebra over M. $\pi^1\colon J^1TM \to M$ is the first jet prolongation of the tangent

*Supported by the Research and Development Council of the Czech Government MSM 6 198 959 214.

vector bundle $\pi\colon TM \to M$, that is, the fibred manifold of 1-jets in $J^1(M, TM)$ which may be represented by local sections of the projection π. We have also a canonical projection $\pi_0^1\colon J^1TM \to TM = J^0TM$. Given an n-dimensional smooth manifold M, a (generalized) *connection*[1] on TM is a (smooth) section $\Gamma\colon TM \to J^1TM$ of π_0^1. A section Γ of π_0^1 which is linear as a fibred morphism of vector bundles is called a *linear connection* on TM, [10], [8]. Any linear connection Γ on TM induces the so-called covariant derivative on M, and vice versa. Recall that a *covariant derivative* on M is a mapping $(X, Y) \mapsto \nabla_X Y$, $\nabla\colon \mathcal{X}(M) \times \mathcal{X}(M) \to \mathcal{X}(M)$, such that

$$\begin{gathered}\nabla_X(Y+Z) = \nabla_X Y + \nabla_X Z, \qquad \nabla_X(fY) = f\nabla_X Y + (Xf)Y, \\ \nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z\end{gathered} \tag{1}$$

for any vector fields X, Y, Z on M and functions $f, g \in \mathcal{F}(M)$ on M; often, under a linear connection on M we mean just ∇. To emphasise that ∇ arises from a linear connection Γ we can write ∇^Γ. In what follows, (M, ∇) will denote a manifold with linear[2] connection in the above sense.

If (U, φ), $U \subset M$ open, $\varphi = (x^1, \dots, x^n)$ is a local chart on M denote by (x^i, v^i) the induced adapted coordinates on $V = \pi^{-1}(U) \subset TM$ and by (x^i, v^i, v^i_j) the corresponding fibre coordinates on $(\pi_0^1)^{-1}(V) \subset J^1TM$. A connection Γ on TM can be locally given by functions $v^i_j \circ \Gamma = \Gamma^i_j(x, v)$ called *components* of Γ. A connection is linear if and only if its components are just linear functions in v^k, that is, there exist functions Γ^i_{jk} of coordinates on $U \subset M$ such that $\Gamma^i_j(x, v) = \Gamma^i_{jk}(x)v^k$ holds.

If (x^i) are local coordinates on $U \subset M$, we can introduce *components* (Christoffel symbols) of ∇ relative to the chart under consideration directly as the functions $\Gamma^k_{ij}(x)$ given on U by[3] $\nabla_i \frac{\partial}{\partial x^j} := \nabla_{\frac{\partial}{\partial x^i}} \frac{\partial}{\partial x^j} = \Gamma^k_{ij} \frac{\partial}{\partial x^k}$. Note that the linear connection Γ (or ∇, respectively) is fully determined by components Γ^k_{ij} provided they satisfy the well-known transformation law on overlappings of neighborhoods, [9, I, Ch. 3, Th. 7.2, Th. 7.3]; recall that Γ^k_{ij} are not components of a tensor.

Covariant derivation extends to tensor fields, [9, I]: if F is of type (r, s) then $\nabla_X F$ is of the same type, and ∇F is of type $(r, s+1)$.

The *torsion* of a manifold (M, ∇) with linear connection is a type $(0, 2)$ tensor field T given by $T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y]$ for $X, Y \in \mathcal{X}(M)$. Here $[\,,\,]$ is the Lie bracket, $[X, Y]f = X(Yf) - Y(Xf)$ for $f \in \mathcal{F}(M)$; T is skew-symmetric. The *curvature* of (M, ∇) is a type $(0, 3)$ tensor field R defined by $R(X, Y)Z = [\nabla_X, \nabla_Y]Z - \nabla_{[X,Y]}Z = \nabla_X(\nabla_Y Z) - \nabla_Y(\nabla_X Z) - \nabla_{[X,Y]}Z$.

The map $R(X_x, Y_x)\colon T_xM \to T_xM$ is linear and skew-symmetric, $R(Y, X) = -R(X, Y)$. A connection ∇ is called *torsion-free* (*torsion-less*, or *symmetric*) if

[1] In the sense of Ehresmann

[2] Many authors still use the term "affine connection" instead, from historical reasons; note that affine connection or affine manifold may have a different meaning: each tangent space T_xM is considered as an affine space, and $TM \to M$ as an affine bundle, similarly for morphisms etc., [9, I, Ch. 3].

[3] As usually, $\langle \frac{d}{dx^1}, \dots, \frac{d}{dx^n} \rangle$ is a basis of coordinate vector fields.

$T \equiv 0$ (in local coordinates, $\Gamma^i_{jk} = \Gamma^i_{kj}$), and *flat* if $T \equiv 0$ and $R \equiv 0$. ∇ is flat if and only if around any point, there are local coordinates such that $\Gamma^i_{jk} = 0$ holds. We introduce the *Ricci tensor* Ric of type $(0,2)$ as a trace of a linear map, namely $\mathrm{Ric}(Y,Z) = \mathrm{Tr}\{X \mapsto R(X,Y)Z\}$ (the other possibility differs up to a sign). Components of torsion $T = T^i_{jk}\frac{\partial}{\partial x^i} \otimes dx^j \otimes dx^k$, of curvature $R = R^i_{hjk}\frac{\partial}{\partial x^i} \otimes dx^j \otimes dx^k \otimes dx^h$ and of Ricci tensor $\mathrm{Ric} = R_{jk}dx^j \otimes dx^k$ in terms of components of connection are $T^i_{jk} = \Gamma^i_{jk} - \Gamma^i_{jk}$,

$$R^i_{hjk} = \frac{\partial \Gamma^i_{kh}}{\partial x^j} - \frac{\partial \Gamma^i_{jh}}{\partial x^k} + \sum_s \left(\Gamma^i_{js}\Gamma^s_{kh} - \Gamma^i_{ks}\Gamma^s_{jh}\right), \tag{2}$$

$$R_{jk} = \sum\nolimits_i R^i_{kij} = \sum_i \left(\frac{\partial \Gamma^i_{jk}}{\partial x^i} - \frac{\partial \Gamma^i_{ik}}{\partial x^j}\right) + \sum_{i,s} \left(\Gamma^i_{is}\Gamma^s_{jk} - \Gamma^i_{js}\Gamma^s_{ik}\right). \tag{3}$$

Due to the co-called first Bianchi Identity ($R^i_{[hjk]} = 0$)

$$R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0 \tag{4}$$

and antisymmetry of the curvature we get

$$R_{jk} - R_{kj} = \sum_i \left(R^i_{kij} + R^i_{jki}\right) = R^i_{ikj} = \mathrm{Tr}\ \mathrm{R}_{kj} = \sum_s \frac{\partial \Gamma^s_{sj}}{\partial x^k} - \frac{\partial \Gamma^s_{sk}}{\partial x^j}. \tag{5}$$

Hence in general, the Ricci tensor is not necessarily symmetric, even for a symmetric connection. We can see the following:

Lemma 1 *The Ricci tensor satisfies [14, p. 14]*

$$\mathrm{Ric}(Z,Y) - \mathrm{Ric}(Y,Z) = \mathrm{Tr}\, R(Y,Z).$$

In general, the functions $\psi_i = \sum_s \Gamma^s_{is}$ ("traces") that appear in (5) do not transform as components of a tensor (1-form) since Γ^i_{jk} do not, either. Nevertheless, they play the following role:

Lemma 2 (Local necessary and sufficient condition for symmetry of Ric) *The following conditions are equivalent for (M,∇):*

(i) The Ricci tensor Ric *is symmetric on M.*

(ii) The curvature tensor R is trace-less, $\mathrm{Tr}\, R = 0$.

(iii) In each coordinate neighborhood the components of connection satisfy

$$\frac{\partial \Gamma^s_{is}}{\partial x^j} - \frac{\partial \Gamma^s_{js}}{\partial x^i} = 0, \qquad i,j = 1,\dots,n. \tag{6}$$

The equations (6) in fact tell that there is a function f^U on U such that $\psi_i = \sum_s \Gamma^s_{is} = \frac{df^U}{dx^i}$, $i = 1,\dots,n$; ψ_i is a "gradient vector". That is, if we introduce a one-form on a coordinate nbd U by $\psi^U = \sum_i \psi_i dx_i = \Gamma^s_{is}dx^i$ then (6) is a necessary and sufficient condition for ψ^U be closed on U, $d\psi^U = 0$.

Recall that an *exterior q-form* ω on M is a totally antisymmetric type $(0,q)$ field; ω is *closed* if $d\omega = 0$, and *exact* if $\omega = d\alpha$ for some $(q-1)$-form α. Since $d^2 = 0$, exact forms are obviously closed, but not vice versa. The so-called Poincaré lemma guarantees that any closed form is locally exact. Obviously, a form α from the above formula is not determined by ω uniquely (in fact, there are many $(q-1)$-forms with the same differential).

Symmetry of the Ricci tensor is closely related to the concept of parallel volume element. We say that (M,∇), $\dim M = n$, is *locally equiaffine*, or *volume preserving* if locally, around each point $x \in M$, there exists a non-vanishing and covariantly constant n-form ω; $\nabla\omega = 0$. If this is the case, ω is called a (local) *volume element*. The following holds, [14]:

Lemma 3 *(M,∇) with $T \equiv 0$ is locally equiaffine if and only if the Ricci tensor is symmetric.*

(M,∇) with $T \equiv 0$ is called *equiaffine* if it admits a parallel volume element. If M is simply connected and (M,∇) is locally equiaffine then it is equiaffine [14, p. 15]. Hence a symmetric linear connection with a trace-less curvature tensor (equivalently, with symmetric Ric) on a simply connected manifold is equiaffine.

1.1 Parallelism and recurrency

If $c\colon I \to M$, $t \mapsto c(t)$ is a curve, let $\zeta(t) = (c(t), c'(t))$ denote the corresponding tangent vector field along the curve c; $c'(t) = \frac{dc}{dt}$. Let Y be a vector field along c. Then the covariant derivative $\nabla_\zeta Y$ along c is defined; in terms of local coordinates, if $Y = Y^k(t)(\frac{\partial}{\partial x^k})_{c(t)}$ then

$$\nabla_\zeta Y = \sum_k \left(\frac{dY^k}{dt} + \sum_{i,j} \Gamma^k_{ij}(c(t)) \frac{dc^i}{dt} Y^j \right) \frac{\partial}{\partial x^k}.$$

A regular[4] differentiable curve $t \mapsto c(t)$ is an *unparametrized geodesic*[5], [13], or *pregeodesic*, [14], if there is a real function $\phi(t)\colon I \to \mathbb{R}$ along c such that $\nabla_\zeta\zeta = \phi\zeta$. Equations of (pre)geodesics read $x''^i + \Gamma^i_{jk} x'^j x'^k = \phi\, x'^i$. If the tangent vector field is parallel along the curve, $\nabla_\zeta\zeta = 0$, we speak on *canonically parametrized geodesics*; the so-called *canonical affine parameter* s is determined uniquely up to affine transformations $s \mapsto as+b$ with $a \neq 0$. In local coordinates, canonically parametrized geodesics are described by the well-known system of differential equations

$$\ddot{x}^i + \Gamma^i_{jk} \dot{x}^j \dot{x}^k = 0, \qquad j,k = 1,\dots,n. \tag{7}$$

Connections with the same "symmetric part" ∇_s, $\nabla_s(X,Y) = \frac{1}{2}(\nabla(X,Y) + \nabla(Y,X))$, have the same geodesics, and pregeodesics, too.

[4] in the sense that $\dot{c}(s) = \frac{dc}{ds} \neq 0$ for all $s \in I$

[5] to emphasize that the particular parametrization is unrelevant for actual considerations

A diffeomorphism $f\colon (M,\nabla) \to (\hat{M},\hat{\nabla})$ is called a *geodesic mapping* if all geodesics of (M,∇) are mapped into unparametrized geodesics of $(\hat{M},\hat{\nabla})$.

A non-vanishing tensor field F on (M,∇) is called *parallel*, or *covariantly constant* (with respect to ∇) if $\nabla F = 0$; equivalently[6], $\nabla_X F = 0$ for any $X \in \mathcal{X}(M)$. A non-vanishing tensor field F on M is *recurrent* if there is a one-form ω such that

$$\nabla F = \omega \otimes F. \tag{8}$$

Lemma 4 *Let a type (r,s) tensor field F on (M,∇) be recurrent; $\nabla F = \omega \otimes F$ for some 1-form. Let F be non-vanishing on M. Then the 1-form ω is closed.*

Proof Recurrency means that for arbitrary vector fields $Y_1,\dots,Y_s$ and one-forms $\omega^1,\dots,\omega^r$ on M,

$$(\nabla_X F)(Y_1,\dots,Y_s,\omega^1,\dots,\omega^r) = \omega(X)\cdot F(Y_1,\dots,Y_s,\omega^1,\dots,\omega^r).$$

In local coordinates about any point $p \in M$, let $\omega = \omega_k\,dx^k$, and $\nabla_k = \nabla_{\frac{\partial}{\partial x^k}}$. It follows that $\nabla_k F^{i_1\dots i_r}_{j_1\dots j_s} = \omega_k \cdot F^{i_1\dots i_r}_{j_1\dots j_s}$ for any $k = 1,\dots,n$; $n = \dim M$. Let the component $F^{i_1\dots i_r}_{j_1\dots j_s}$ (for fixed indices) be non-zero at p, and due to continuity, in some nbd U of p (from continuity again, the component is either positive, or negative around the point). Then the components of the 1-form can be expressed in U as

$$\omega_k = \frac{1}{F^{i_1\dots i_r}_{j_1\dots j_s}} \cdot \nabla_k F^{i_1\dots i_r}_{j_1\dots j_s} = \nabla_k (\ln|F^{i_1\dots i_r}_{j_1\dots j_s}|) = \frac{\partial}{\partial x^k}(\ln|F^{i_1\dots i_r}_{j_1\dots j_s}|),\ \ k = 1,\dots,n.$$

That is, about any point $p \in M$, $\omega = d(\ln|F^{i_1\dots i_r}_{j_1\dots j_s}|)$; i.e. ω is locally exact, and $d\omega = d(df) = 0$. □

Lemma 5 *Let F be a type (r,s) tensor field on (M,∇). Let $\alpha \in \mathcal{F}(M)$ be a non-vanishing real function; $\alpha(x) \neq 0$ for $x \in M$. Then the following conditions are equivalent:*

- *$\alpha \otimes F$ is parallel with respect to ∇,*
- *$\nabla F = d(-\ln|\alpha|) \otimes F$.*

Proof Since $\nabla(\alpha\otimes F) = (\nabla\alpha)\otimes F + \alpha\otimes(\nabla F)$ and $\alpha \neq 0$, we have: $\nabla(\alpha\otimes F) = 0$ iff $\nabla F = -(\frac{1}{\alpha}\cdot\nabla\alpha)\otimes F = -d(\ln|\alpha|)\otimes F$. Hence $\alpha\otimes F$ is parallel if and only if $\nabla F = df\otimes F$ where $f = -\ln|\alpha|$. □

Lemma 6 *If a tensor field F of type (r,s) on (M,∇) is recurrent, $\nabla F = \omega\otimes F$, and the 1-form is exact, $\omega = df$, then $e^{-f}\otimes F$ is parallel w.r.t. ∇.*

Proof If $\nabla F = df\otimes F$ denote $\alpha = e^{-f}$. Then $f = -\ln\alpha$, and $\nabla(\alpha\otimes F) = d\alpha\otimes F + \alpha\cdot d(-\ln\alpha)\otimes F = d\alpha\otimes F + \alpha\cdot(-\frac{1}{\alpha})\cdot d\alpha\otimes F = 0$. Hence $\alpha\otimes F = e^{-f}\otimes F$ is parallel. □

[6] In more geometric language, the condition tells that the field is preserved under parallel transport along all curves in M.

1.2 Compatible metrics

Recall that a *pseudo-Riemannian metric* on a smooth manifold M is a (smooth) type $(0,2)$ tensor field on M such that in any point $x \in M$, the corresponding bilinear form g_x defined on T_xM is symmetric and non-degenerate; (M,g) is called a *pseudo-Riemannian* manifold. If g_x is moreover positive definite for all $x \in M$, (M,g) is called the *Riemannian space*. A linear connection ∇ (may be non-symmetric in general) on (M,g) is *compatible* with g if g is parallel with respect to ∇, $\nabla g = 0$.

The Fundamental Theorem of Riemannian geometry states that any pseudo-Riemannian manifold (M,g) admits a unique linear connection $\tilde{\nabla}$, called the Riemannian (or Levi-Civita) connection, or metric connection, of (M,g), characterized by the pair of conditions $T \equiv 0$, $\tilde{\nabla} g = 0$ (the parallel transport with respect to $\tilde{\nabla}$ along any curve preserves the scalar product of tangent vectors defined by g). On (M,g), components Γ^i_{jk} of the Levi-Civita connection are related to components g_{ij} of the metric by the well-known formula $\Gamma^\ell_{ik} = \frac{1}{2} g^{\ell j} \left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{jk}}{\partial x^i} - \frac{\partial g_{ki}}{\partial x^j} \right)$.

On the other hand, given a manifold equipped with a linear connection, (M,∇), we might be interested in metrics the given connection is compatible with. If ∇ is torsion-free, it means to find a metric g on M such that ∇ is just the Levi-Civita connection of (M,g). We say that a manifold (M,∇) is *metrizable*, or *locally metrizable*, respectively, if there exists a metric (or exists locally, respectively) compatible with the connection (metrization problem, MP).

Essentially the same problem can be formulated in a bit more general setting as follows, [13] (the answer is formulated in Corollary 1): *If (M,∇) is given find all geodesic mappings (i.e. diffeomorphisms which map geodesics onto unparametrized geodesics) of (M,∇) onto (all possible) pseudo-Riemannian manifolds $(\bar{M},g)$ (due to diffeomorphisms, we can in fact suppose $\bar{M} = M$).*

In local coordinates, the formula $\nabla g = 0$ reads[7]

$$\frac{\partial g_{ij}}{\partial x^k} = g_{sj}\Gamma^s_{ik} + g_{is}\Gamma^s_{jk}. \tag{9}$$

In principle, to answer the question on (local) metrizability of a connection means to solve the system[8] (9). Employing the curvature, necessary integrability conditions for metrizability can be given in the form of an infinite system of linear equations in $\frac{1}{2}n(n+1)$ functions g_{ij} (with coefficients which are functions in $\Gamma's$ and their partial derivatives), [7]; the coordinate-free form reads

$$g(R(X,Y)Z,W) + g(Z,R(X,Y)W) = 0, \tag{10}$$

$$g(\nabla^r R(X,Y;Z_1;\ldots;Z_r)(Z),W) + g(Z,\nabla^r R(X,Y;Z_1;\ldots;Z_r)(W)) = 0 \tag{11}$$

for all $X,Y,Z,W,Z_1,\ldots,Z_r \in \mathcal{X}(M)$, $1 \le r < \infty$. Flat connections are locally metrizable[9]. If (10) has at least a 1-dimensional solution space containing a

[7] In components, $g_{ij;k} =: \nabla g(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}; \frac{\partial}{\partial x^k}) = \frac{\partial g_{ij}}{\partial x^k} - g_{sj}\Gamma^s_{ik} - g_{is}\Gamma^s_{jk}$.

[8] Which can be done directly in simple cases.

[9] For the detailed theory of flat affine manifolds, cf. [9, I], flat Riemannian manifolds are discussed e.g. in [23].

non-degenerate metric and any solution of (10) satisfies also (11) for $r = 1$ then (M, ∇) is metrizable, [7].

Corollary 1 *If there exist* $\frac{1}{2}n(n+1)$ *(differentiable) functions* g_{ij} *which solve the system*

$$g_{sj}R^s_{ik\ell} + g_{is}R^s_{jk\ell} = 0 \tag{12}$$

and satisfy $g_{ij} = g_{ji}$, $\det(g_{ij}) \neq 0$, *and any solution of (12) solves the system*

$$g_{sj}R^s_{ik\ell;m} + g_{is}R^s_{jk\ell;m} = 0 \tag{13}$$

then (locally) there exist geodesic mappings of (M, ∇) *onto pseudo-Riemannian spaces.*

On a (pseudo-)Riemannian manifold (M, g) with the metric tensor g besides the curvature tensor R in type $(1, 3)$, we can consider the type $(0, 4)$ tensor $\tilde{R}(X, Y, Z, W) = g(R(X, Y)Z, W)$, usually also called curvature tensor; the relations $\tilde{R}(X, Y, Z, W) = \tilde{R}(Z, W, X, Y) = -\tilde{R}(Y, X, Z, W) = -\tilde{R}(X, Y, W, Z)$ hold. In a coordinate system $(U, \varphi = (x^i))$ based at a point $x \in M$, components R^ℓ_{ijk} of R and R_{hijk} of $\tilde{R} = R_{hijk}dx^j \otimes dx^k \otimes dx^i \otimes dx^h$ are related by $R_{hijk} = g_{hs}R^s_{ijk}$, and $g^{\ell h}R_{hijk} = R^\ell_{ijk}$[10].

Lemma 7 *The Ricci tensor of the Levi-Civita connection of a (pseudo-)Riemannian manifold* (M, g) *is always symmetric, [6, p. 331].*

The sectional curvature of a two-space P given by the linearly independent tangent vectors $X, Y \in T_xM$ is given by

$$K(X \wedge Y) = \frac{g(R(X, Y)Y, X)}{g(X, X)g(Y, Y) - g(X, Y)^2} = \frac{\tilde{R}(X, Y, Y, X)}{||X \wedge Y||^2} \tag{14}$$

where $||X \wedge Y||$ is the area of a parallelogram determined by X and Y, [3, p. 94], [6, p. 327] etc. The sectional curvature determines the whole curvature tensor $\tilde{R}$, [8, p. 137].

On (M, g), the Ricci tensor in type $(1, 1)$ is introduced with components $R^i_j = g^i_s R_{sj}$, and the scalar curvature ϱ as its trace, $\varrho = \operatorname{Tr} \operatorname{Ric} = R^s_s = g^{ij}R_{ij}$.

A Riemannian manifold (M, g) is called *isotropic at a point* $x \in M$ if the curvature is the same constant, $K(x)$, on every (two-plane) section, and *isotropic* if it is isotropic at every point, [1]. If x is an isotropic point of (M, g) then the following formula holds at x in any local coordinates around x:

$$R_{hijk} = K(x)(g_{hj}g_{ik} - g_{hi}g_{jk}). \tag{15}$$

A two-dimensional manifold is (trivially) isotropic, therefore it satisfies (15).

Pseudo-Riemannian manifolds with symmetric Ricci tensor for which the Ricci tensor is proportional to the metric tensor, $\operatorname{Ric} = \lambda g$, are called *Einstein spaces*, [12, p. 263], [15], [17]. In the Loretzian case, they are important in

[10] As already mentioned, $R_{hijk} = R_{jkhi} = -R_{ihjk} = -R_{hikj}$.

Einstein's theory of general relativity (the Einstein's field equation is a dynamical equation which describes how matter changes the geometry of spacetime; in vacuum, it is given by the condition $\mathrm{Ric} = 0$). The factor of proportionality can be calculated[11], $\lambda = \frac{1}{n}\varrho$, hence for Einstein spaces,

$$\mathrm{Ric} = \frac{1}{n}\varrho g. \tag{16}$$

Particularly, all two-dimensional pseudo-Riemannian manifolds are Einstein spaces as we check below, cf. [12, p. 263], [15, p. 101].

2 Metrizability of 2-manifolds

Let us pay attention to existence of compatible metrics in the simplest case $n = \dim M = 2$. Let (x^1, x^2) denote local coordinates on a coordinate neighborhood U of a manifold M_2. In dimension two, the curvature is simply given by $R_{hijk} = K(x)(g_{hj}g_{ik} - g_{hi}g_{jk})$ [8, p. 137], and the function $K(x)$ is called the *Gauss curvature*. The Riemann curvature R in type $(1,3)$ and the Ricci tensor Ric are related by [12], [15]

$$R^i_{hjk} = \delta^i_j R_{kh} - \delta^i_k R_{jh}. \tag{17}$$

As far as $R^i_{hjj} = 0$ and $R^i_{hij} = R_{jh}$ holds for $j \neq i$, the curvature tensor of a linear connection ∇ on M_2 is completely determined by its Ricci tensor; explicitely,

$$\begin{aligned} R_{11} &= -R^2_{112} = R^2_{121}, \quad & R_{21} &= -R^1_{121} = R^1_{112}, \\ R_{12} &= -R^2_{212} = R^2_{221}, \quad & R_{22} &= -R^1_{221} = R^1_{212}. \end{aligned} \tag{18}$$

Particularly, $R = 0$ if and only if $\mathrm{Ric} = 0$, and recurrency is also inherited:

Lemma 8 *For (M_2, ∇),* Ric *is recurrent if and only if R is recurrent.*

Proof Let Ric be recurrent, $\nabla\mathrm{Ric} = \omega \otimes \mathrm{Ric}$. In local coordinates, if $\omega = \omega_j dx^j$ then $\nabla_\ell R^i_{hjk} = \delta^i_j \nabla_\ell R_{kh} - \delta^i_k \nabla_\ell R_{jh} = \delta^i_j \omega_\ell R_{kh} - \delta^i_k \omega_\ell R_{jh} = \omega_\ell R^i_{hjk}$, hence $\nabla R = \omega \otimes R$. Vice versa, if $\nabla R = \omega \otimes R$ holds then $\nabla_\ell R_{jk} = \omega_\ell R^i_{kij} = \omega_\ell R_{jk}$, and $\nabla\mathrm{Ric} = \omega \otimes \mathrm{Ric}$. □

On (M_2, g), non-zero components of type $(0,4)$ curvature $\tilde{R}$ are (up to a sign) equal just R_{1212}, and (15) reads ([15, p. 62], [8, p. 137])

$$R_{hijk} = K(g_{hj}g_{ik} - g_{hk}g_{ij}) \tag{19}$$

where $K = K(x)$ is the Gauss curvature, $K = \frac{R_{1212}}{\det(g_{ij})}$.

Lemma 9 *The curvature tensor of a two-dimensional pseudo-Riemannian manifold (M_2, g) satisfies*

$$R^i_{hjk} = K(\delta^i_k g_{hj} - \delta^i_j g_{hk}), \tag{20}$$

and the Ricci tensor is proportional to the metric tensor,

$$\mathrm{Ric} = K \cdot g = \frac{1}{2}\varrho \cdot g. \tag{21}$$

[11] In fact, $\varrho = R_{ij}g^{ij} = \lambda g_{ij}g^{ij} = n\lambda$.

Proof We can either use the fact that M_2 is trivially isotropic, [1, p. 374], and (16) holds, or proceed by direct evaluation: $R^t_{hij} = R^s_{hij}\delta^t_s = R^s_{hij}g_{sk}g^{kt} = R_{hijk}g^{kt} = K(g_{hj}g_{ik} - g_{hk}g_{ij})g^{kt} = K(\delta^t_i g_{hj} - \delta^t_h g_{ij})$. It follows immediately for the Ricci tensor that $R_{hj} = \Sigma_i R^i_{hij} = K \cdot \Sigma_i(\delta^i_i g_{hj} - \delta^i_h g_{ij}) = K \cdot g_{hj}$, hence $\mathrm{Ric} = Kg$, and $\varrho = R_{hj}g^{hj} = 2K$. □

Corollary 2 *(M_2, g) is always an Einstein space. For a nowhere flat (M_2, g), the Ricci tensor is symmetric and non-degenerate.*

Note that according to [9, I, p. 280], any non-flat Riemannian 2-manifold has a recurrent curvature provided its sectional curvature does not vanish. We can check:

Lemma 10 *The Ricci tensor of a nowhere flat pseudo-Riemannian manifold (M_2, g) is recurrent, and the corresponding 1-form is exact*[12]*.*

Proof $R \neq 0$ is equivalent with $K(x) \neq 0$ on M (from continuity, K is either positive, or negative). Since by (21), $g = \alpha(x) \cdot \mathrm{Ric}$ with $\alpha(x) = \frac{1}{K(x)} \neq 0$, and $\nabla g = 0$, we get easily that $\alpha(x) \cdot \mathrm{Ric}$ is parallel. According to Lemma 5, $\nabla\mathrm{Ric} = d(-\ln|\alpha|) \otimes \mathrm{Ric}$ holds. □

It follows from the above discussion on pseudo-Riemannian manifolds that two conditions are necessary for local metrizability of a (symmetric) connection on a 2-manifold: the Ricci tensor must be symmetric, and must be also recurrent, with the corresponding 1-form being closed; Ric may be degenerate only in the case $R = 0$, and then $\mathrm{Ric} = 0$ holds. Furthermore, for global metrizability, the 1-form from the recurrency condition must be even exact. A flat connection is always (globally) metrizable, with $\frac{1}{2}n(n+1)$-parameter solution space; even the signature can be prescribed. So let us pay attention to the situation when the curvature tensor (or equivalently, the Ricci tensor) is non-zero in one point $x_0 \in M$, and due to continuity, in some neighborhood of x_0[13].

Theorem 1 (Existence of local metrics on two-manifolds) *Let a 2-dimensional manifold (M_2, ∇) with a symmetric linear connection be given such that the Ricci tensor is regular, $|R_{ij}| \neq 0$, symmetric, $R_{ij} = R_{ji}$, and recurrent, $\nabla Ric = \varrho \otimes Ric$ for some 1-form ϱ. Then locally, there is a metric compatible with the connection.*

Proof Let $x_0 \in M$. $|R_{ij}| \neq 0$ implies existence of a pair (i, j) of indices such that $R_{ij} \neq 0$ about[14] x_0. Recurrency together with regularity guarantee that $d\varrho = 0$ (Lemma 4). Hence about x_0, there is a function f such that $\varrho = df$. Consequently, $e^{-f} \cdot \mathrm{Ric}$ is parallel about x_0. Therefore $g = e^{-f} \cdot \mathrm{Ric}$ is a local metric on a nbd of x_0 compatible with ∇. □

[12] and consequently closed

[13] The subset of non-flat points is open.

[14] Under "about x" we mean on some neighborhood of x.

Of course, the function f from the proof is not unique. Any function $\tilde{f}$ with the same differential, $d\tilde{f} = df$, also gives a metric; such a function differs up to a constant, $\tilde{f} = f + a$, $a \in \mathbb{R}$.

If R is nowhere zero, a similar proof quarantees existence of global metrizability of a nowhere flat affine manifold:

Proposition 1 *Let (M_2, ∇) be a two-dimensional manifold with a symmetric linear connection. If the Ricci tensor of ∇ is regular, symmetric, and recurrent, $\nabla Ric = \varrho \otimes Ric$, and the 1-form ϱ is exact, i.e. $\varrho = df$ for some function $f \in \mathcal{F}(M)$, then $g = e^{-f} \cdot \mathrm{Ric}$ is a (global) metric tensor compatible with ∇.*

Theorem 2 (Global metrizability of no-where flat connections on 2-manifolds) *A nowhere flat symmetric linear connection on M_2 is metrizable if and only if its Ricci tensor is regular, symmetric, recurrent, and the corresponding 1-form is exact. If this is the case, and $\nabla Ric = df \otimes Ric$ holds for some smooth function $f \in \mathcal{F}(M)$, then all global metrics compatible with ∇ form a 1-parameter family described by the formula*

$$g_b = \exp(-f + b) \cdot \mathrm{Ric}, \quad b \in \mathbb{R}, \tag{22}$$

that is, any of them arises from the Ricci tensor as a multiple by a smooth function. Moreover, any two compatible metrics differ up to a scalar multiple.

Proof The main statement has been already proved - the "if" part in Theorem 1 and Proposition 1, and the "only if" part in Corollary 2 and Lemma 10. As to the rest, let $g = e^{-f} \cdot \mathrm{Ric}$, $\tilde{g} = e^{-\tilde{f}} \cdot \mathrm{Ric}$ be two compatible metrics, then $\tilde{f} - f = a$, $\mathrm{Ric} = e^{\tilde{f}} g$, and $g = e^a \tilde{g}$. We get $\tilde{g} = e^{-f-a} \cdot \mathrm{Ric}$; i.e. (22) holds. □

As an immediate consequence of Theorem 2 we obtain:

Corollary 3 *Two pseudo-Riemannian metrics g_1, g_2 compatible with the same nowhere flat (symmetric) linear connection on M_2 are homothetic.*

Unicity of g declared in [18, p. 532] must be understood in this way.

For positive-definite metrics, this result is a special case of the Theorem 1 of O. Kowalski from [11, p.131] (recall that two metrics g_1, g_2 on a manifold are called *conformally equivalent* if there is a function κ on M such that $g_2 = \kappa g_1$, [23, p. 99]): *Let g, g' be two Riemann metrics on a smooth manifold M with the same Riemann curvature tensor R. Then g, g' are conformally equivalent on the closure of the set of all regular points of R.*

3 Application in the calculus of variations

Let us mention the relationship of our problem to the Calculus of Variations. The so-called Inverse Problem (IP) of the calculus of variations is: if a system $\ddot{x}^i = f^i(t, x^k, \dot{x}^k)$, $i, k = 1, \ldots, n$ of second order differential equations (SODEs)

is given, find—sufficiently differentiable—Lagrangian functions $L(t, x^k, \dot{x}^k)$ and a multiplier matrix $g_{ij}(t, x^k, \dot{x}^k)$ such that

$$g_{ij}(\ddot{x}^i - f^i) \equiv \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i}.$$

Given a system of second order ODEs of a particular type

$$\ddot{x}^i + \Gamma^i_{jk}(x)\dot{x}^j\dot{x}^k = 0, \qquad k = 1, \dots, n, \tag{23}$$

that is, second derivatives can be expressed as quadratic forms in first derivatives, we can use the above theory for deciding whether the system (23) is derivable from a Lagrangian. In fact, provided $\det(g_{ij}) \neq 0$, the system (23) is equivalent to the system

$$g_{mi}(\ddot{x}^i + \Gamma^i_{jk}(x)\dot{x}^j\dot{x}^k) = 0, \quad i, m = 1, \dots, n. \tag{24}$$

Another speaking, MP can be viewed as a particular case of IP, where $f^i = -\Gamma^i_{jk}(x)\dot{x}^j\dot{x}^k$ (that is, f^i are quadratic forms in components of velocities, with coefficients depending only on components of positions) in the particular case when the multipliers are time- and velocities-independent. We can assume that the coefficients in (23), the functions $\Gamma^k_{rs}(x)$, are components of a symmetric linear connection ∇ on some neighborhood $U \subset \mathbb{R}^n$. If ∇ is (locally) metrizable, and $g_{ij}(x)$ (with $\det(g_{ij}(x)) \neq 0$ at any $x \in U$) are components of some non-degenerate metric g compatible with ∇ on U, then (23) and (24) are equivalent, hence the functions $g_{ik}(x)$ can be taken as the desired variational multipliers. One of particular Lagrangians comming from MP (and solving IP) is

$$L = T = \frac{1}{2}g_{ij}(x)\dot{x}^i\dot{x}^j, \tag{25}$$

the kinetic energy. There might exist multipliers of a more general form $g_{ik}(t, x, \dot{x})$, depending on "time, positions and velocities", which might bring more complicated Lagrangians, [5].

4 Examples

Example 1 ([7, p. 122]) On $\mathbb{R}^2$ with coordinates $x = (x^1, x^2)$, assume the system of ODEs

$$(\ddot{x}^1)^2 + (x^1 - x^2)(\dot{x}^1)^2 = 0, \quad (\ddot{x}^2)^2 + (x^1 - x^2)(\dot{x}^2)^2 = 0. \tag{26}$$

Curves $c(s)\colon I \to \mathbb{R}^2$ (parametrized by arcs length), which are solutions of the system, represent the family of geodesics of a (symmetric) linear connection ∇ with components $\Gamma^1_{11} = \Gamma^2_{22} = x^1 - x^2$, $\Gamma^i_{jk} = 0$ otherwise. We ask if the (torsion-free linear) connection is metrizable, i.e. we wish to find type $(0, 2)$ symmetric tensor field g with $\nabla g = 0$. The corresponding system

$$\begin{aligned} &\partial_1 g_{11} = (x^1 - x^2)g_{11}, \quad \partial_1 g_{12} = 0, \quad \partial_1 g_{22} = 0, \\ &\partial_2 g_{11} = 0, \quad \partial_2 g_{12} = (x^1 - x^2)g_{12}, \quad \partial_2 g_{22} = (x^1 - x^2)g_{22} \end{aligned}$$

can be solved directly, but the only solution is trivial, $g_{ij} = 0$ for all i, j. Or, argumentation using the Ricci (or curvature) tensor can be used: $R_{11} = R^2_{121} = 0$, $R_{12} = R^1_{112} = 1$, $R_{21} = R^2_{221} = -1$, $R_{22} = R^1_{212} = 0$, hence the Ricci tensor is not symmetric, our linear connection is not metrizable (even locally).

It appears that in this particular case, the quickest and most comfortable way is to use the criterion from Lemma 2 (iii): we check that $\psi_1 = \psi_2 = x^1 - x^2$, $\partial_1\psi_2 = 1$ while $\partial_2\psi_1 = -1$.

Example 2 The system of equations

$$\ddot{x}^1 = -(\dot{x}^1)^2 - (\dot{x}^2)^2, \quad \ddot{x}^2 = -4\dot{x}^1\dot{x}^2 \tag{27}$$

corresponds to a torsion-free linear connection on $\mathbb{R}^2$ with components

$$\Gamma^1_{11} = \Gamma^1_{22} = 1, \quad \Gamma^1_{12} = \Gamma^1_{21} = 0, \quad \Gamma^2_{11} = \Gamma^2_{22} = 0, \quad \Gamma^2_{12} = \Gamma^2_{21} = 2.$$

Now our "quick" criterion fails, the connection determined by (27) has symmetric Ricci tensor: $\psi_1 = 3$, $\psi_2 = 0$, $\mathrm{Ric} = (R_{hk}) = \begin{pmatrix} -2 & 0 \\ 0 & -1 \end{pmatrix}$. But the connection is not metrizable, either, since Ricci is not recurrent: system of linear equations for functions $\alpha_1(x)$, $\alpha_2(x)$ such that $R_{ij;k} = \alpha_k R_{ij}$

$$\begin{aligned} 4 = R_{11;1} = \alpha_1 R_{11} = -2\alpha_1, \quad & 0 = R_{11;2} = \alpha_2 R_{11} = -2\alpha_2, \\ 4 = R_{22;1} = \alpha_1 R_{22} = -\alpha_1, \quad & 4 = R_{22;2} = \alpha_2 R_{22} = -\alpha_2 \ \text{ etc.} \end{aligned}$$

is inconsistent in our case. The connection is a non-metrizable one. There are no time- and velocities-independent multipliers g_{ij}.

Example 3 ([2]) The system

$$\ddot{x}^1 = 0, \quad \ddot{x}^2 = -2\dot{x}^1\dot{x}^2 \tag{28}$$

defines on $\mathbb{R}^2$ (or on $\mathbb{R} \times \mathbb{S}^1$, or on the torus $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$) a symmetric linear connection ∇ with Christoffel symbols $\Gamma^2_{12} = \Gamma^2_{21} = 1$, $\Gamma^k_{ij} = 0$ otherwise. We can easily check that Ric is symmetric, since $\psi_1 = \Gamma^1_{11} + \Gamma^2_{12} = 1$, and $\psi_2 = \Gamma^1_{21} + \Gamma^2_{22} = 0$. But it is degenerate, evaluation of the components brings $(R_{ij}) = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$. Therefore ∇ is not metrizable (even locally). If we try to solve directly the system corresponding to $\nabla g = 0$,

$$\begin{aligned} &\partial_1 g_{11} = 0, && \partial_1 g_{12} = g_{12}, && \partial_1 g_{22} = 2g_{22}, \\ &\partial_2 g_{11} = 2g_{12}, && \partial_2 g_{12} = g_{22}, && \partial_2 g_{22} = 0, \end{aligned}$$

we get a similar answer, $G = (g_{ij}) = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$.

Example 4 ([2]) Equations

$$\ddot{x}^1 = -(\dot{x}^1)^2, \qquad \ddot{x}^2 = -(\dot{x}^2)^2 \tag{29}$$

determine on $M_2 = \mathbb{R}^2$ a symmetric linear connection $\nabla_{X_1} X_1 = X_1 = 0$, $\nabla_{X_2} X_2 = X_2$, $\nabla_{X_i} X_j = 0$ otherwise, $X_i = \frac{\partial}{\partial x^i}$, with Christoffels

$$\Gamma^1_{11} = \Gamma^2_{22} = 1, \qquad \Gamma^k_{ij} = 0 \quad \text{otherwise.}$$

The curvature tensor R vanishes, equivalently, $\mathrm{Ric} = 0$, the connection ∇ is flat, hence (locally) metrizable, and the system (29) is variational. To find out components of the metric, or another speaking, variational multipliers g_{ij}, we can solve the system of PDEs

$$\begin{aligned} &\partial_1 g_{11} = 2g_{11}, \quad &&\partial_1 g_{12} = g_{12}, \quad &&\partial_1 g_{22} = 0, \\ &\partial_2 g_{11} = 0, \quad &&\partial_2 g_{12} = g_{12}, \quad &&\partial_2 g_{22} = 2g_{22}. \end{aligned}$$

Given $x_0 \in M$, a non-singular 2×2 matrix (g^0_{ij}) and initial data $g_{ij}(x_0) = g^0_{ij}$ the solution is $g_{11} = g^0_{11} e^{2x^1}$, $g_{12} = g^0_{12} e^{x^1 + x^2}$, $g_{22} = g^0_{22} e^{2x^2}$, hence we get a (global) metric on $\mathbb{R}^2$ and the corresponding Lagrangian,

$$g_{ij} = g^0_{ij} \cdot e^{x^i + x^j}, \qquad L = \frac{1}{2} g^0_{ij} e^{x^i + x^j} \dot{x}^i \dot{x}^j$$

(remark that direct search for solution of the corresponding system of PDEs need not be easy in most cases). The Ricci tensor brings the same answer.

Note that if we introduce essentially the same connection on the "infinite cylindr" $\mathbb{S}^1 \times \mathbb{R}$, or on the torus $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$, such a connection is not globally metrizable. Indeed, consider the (continuous, even smooth) function $f(t) = |X_1(\gamma(t))|$, $t \in (0, 1)$, the length of the (smooth and globally defined) coordinate vector field X_1 along the "flow line" (which is the circle without one point): it satisfies $f' = 2f$; the metric behaves "exponentially". We must expect problems with successful "taping" of the metric on the overlap of coordinate neighborhoods.

Another example of C^∞-connection which is metrizable locally but not globally is given in [16], cf. [22].

Example 5 For the system

$$\ddot{x}^1 + \dot{x}^1 \dot{x}^2 = 0, \qquad \ddot{x}^2 - \frac{1}{2} \exp(x^2)(\dot{x}^1)^2 = 0, \tag{30}$$

non-zero components are $\Gamma^1_{12} = \Gamma^1_{21} = \frac{1}{2}$, $\Gamma^2_{11} = -\frac{1}{2} e^{x^2}$. The Ricci tensor with components $\mathrm{Ric} = -\frac{1}{4} e^{x^2} dx^1 \otimes dx^1 - \frac{1}{4} dx^2 \otimes dx^2$ is covariant constant, $\nabla \mathrm{Ric} = 0$, therefore recurrent with vanishing (and consequently exact) 1-form $\omega = 0 = d\,(\mathrm{const})$ entirely on $\mathbb{R}^2$. All (global) compatible metrics on $\mathbb{R}^2$ form a one-parameter family

$$g_b = \exp(x^2 + b)\, dx^1 \otimes dx^1 + \exp(b)\, dx^2 \otimes dx^2, \quad b \in \mathbb{R}, \tag{31}$$

which yields Lagrangians $L = \frac{1}{2} e^{x^2 + b} (\dot{x}^1)^2 + \frac{1}{2} e^b (\dot{x}^2)^2$.

References

[1] Boothby, W. M.: An Introduction to Differentiable Manifolds and Riemannian Geometry. *Academic Press, Amsterdam–London–New York–Oxford–Paris–Tokyo,* 2003 (revised second editon).

[2] Cocos, M.: *A note on symmetric connections.* J. Geom. Phys. **56** (2006), 337–343.

[3] do Carmo, M. P.: Riemannian Geometry. *Birkhäuser, Boston–Basel–Berlin,* 1992.

[4] Cheng, K. S, Ni, W. T.: *Necessary and sufficient conditions for the existence of metrics in two-dimensional affine manifolds.* Chinese J. Phys. **16** (1978), 228–232.

[5] Douglas, J.: *Solution of the inverse problem of the calculus of variations.* Trans. AMS **50** (1941), 71–128.

[6] Dodson, C. T. J., Poston, T.: Tensor Geometry. The Geometric Viewpoint and its Uses. *Spriger, New York–Berlin–Heidelberg,* 1991 (second editon).

[7] Eisenhart, L. P., Veblen, O.: *The Riemann geometry and its generalization.* Proc. London Math. Soc. **8** (1922), 19–23.

[8] Jost, J.: Riemannian Geometry and Geometric Analysis. *Springer, Berlin–Heidelberg–New York,* 2005.

[9] Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry I, II. *Wiley, New York–Chichester–Brisbane–Toronto–Singapore,* 1991.

[10] Kolář, I., Slovák, J., Michor, P. W.: Natural Operations in Differential Geometry. *Springer, Berlin–Heidelberg–New York,* 1993.

[11] Kowalski, O.: *On regular curvature structures.* Math. Z. **125** (1972), 129–138.

[12] Lovelock, D., Rund, H.: Tensors, Differential Forms, and Variational Principle. *Wiley, New York–London–Sydney,* 1975.

[13] Mikeš, J., Kiosak, V., Vanžurová, A.: Geodesic Mappings of Manifolds with Affine Connection. *Palacký Univ. Publ., Olomouc,* 2008.

[14] Nomizu, K., Sasaki, T.: Affine Differential Geometry. Geometry of Affine Immersions. *Cambridge Univ. Press, Cambridge,* 1994.

[15] Petrov, A. Z.: Einstein Spaces. *Moscow,* 1961 (in Russian).

[16] Schmidt, B. G.: *Conditions on a connection to be a metric connection.* Commun. Math. Phys. **29** (1973), 55–59.

[17] Sinyukov, N. S.: Geodesic Mappings of Riemannian Spaces. *Moscow,* 1979 (in Russian).

[18] Thompson, G.: *Local and global existence of metrics in two-dimensional affine manifolds.* Chinese J. Phys. **19**, 6 (1991), 529–532.

[19] Vanžurová, A.: *Linear connections on two-manifolds and SODEs.* Proc. Conf. Aplimat 2007 (Bratislava, Slov. Rep.), Part II (2007), 325–332.

[20] Vanžurová, A.: *Metrization problem for linear connections and holonomy algebras.* Archivum Mathematicum (Brno) **44** (2008), 339–348.

[21] Vanžurová, A.: *Metrization of linear connections, holonomy groups and holonomy algebras.* Acta Physica Debrecina **42** (2008), 39–48.

[22] Vanžurová, A., Žáčková, P.: *Metrization of linear connections.* Aplimat, J. of Applied Math. (Bratislava) **2**, 1 (2009), 151–163.

[23] Wolf, J. A.: Spaces of Constant Curvature. *Berkley, California,* 1972.

Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 171–188

Suitability of Linearization of Nonlinear Problems not only in Biology and Medicine*

Jana VRBKOVÁ

Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: vrbkova@inf.upol.cz

(Received January 30, 2009)

Abstract

Biology and medicine are not the only fields that present problems unsolvable through a linear models approach. One way to overcome this obstacle is to use nonlinear methods, even though these are not as thoroughly explored. Another possibility is to linearize and transform the originally nonlinear task to make it accessible to linear methods. In this aricle I investigate an easy and quick criterion to verify suitability of linearization of nonlinear problems via Taylor series expansion so that linear models with type II constraints could be used.

Key words: Linear models with constraints, compartmental analysis, nonlinear models, linearization via a Taylor series.

2000 Mathematics Subject Classification: 62J05

1 Used symbols

$h(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\mathbf{M_A}$	a matrix $\mathbf{M_A} = \mathbf{I} - \mathbf{P_A}$
$\mathbf{P_A}$	a projector on the space $\mathcal{M}(\mathbf{A})$ in Euclidean norm
$\mathcal{M}(\mathbf{A})$	range space of the matrix $\mathbf{A}$
$\mathbb{R}^k$	k-dimensional linear vector space
$\chi^2_f(0; 1-\alpha)$	$(1-\alpha)$-quantile of the random variable with $\chi^2_f(0)$ distribution
$\mathbf{X}^-$	generalized inverse of the matrix $\mathbf{X}$
$\mathbf{X}^+$	Moore-Penrose g-inverse of the matrix $\mathbf{X}$
$(\mathbf{X})^-_{m(\mathbf{\Sigma})}$	minimum $\mathbf{\Sigma}$-norm (seminorm) g-inverse of the matrix $\mathbf{X}$

*Supported by the Council of the Czech Government MSM 6 198 959 214.

2 Linearization via a Taylor series

Let us consider a general nonlinear model

$$\mathbf{Y} \sim_n (\mathbf{f}(\boldsymbol{\beta}_1), \boldsymbol{\Sigma}), \quad \boldsymbol{\beta}_1 \in \mathbb{R}^{k_1}, \quad \boldsymbol{\beta}_2 \in \mathbb{R}^{k_2},$$

where the parameter $\boldsymbol{\beta}_2$ occurs only in a constraint $\mathbf{g}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{0}$, the function

$$\mathbf{f} : \mathcal{V} \to \mathbb{R}^n, \quad \mathcal{V} = \left\{ \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix} : \mathbf{g}(\boldsymbol{\beta_1}, \boldsymbol{\beta_2}) = \mathbf{0} \right\},$$

has continuous second derivatives, and $\mathbf{g}(\cdot)$ is a q-dimensional function with continuous second derivatives.

If we know approximate values $\boldsymbol{\beta}_1^0$, $\boldsymbol{\beta}_2^0$ of the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ we can linearize functions $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ via Taylor series

$$\mathbf{f}(\boldsymbol{\beta}_1) = \mathbf{f}\left(\boldsymbol{\beta}_1^0\right) + \mathbf{F}\left(\boldsymbol{\beta}_1^0\right) \delta\boldsymbol{\beta_1} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1) + \ldots,$$

where

$$\mathbf{F}\left(\boldsymbol{\beta}_1^0\right) = \partial \mathbf{f}(\boldsymbol{\beta}_1)/\partial \boldsymbol{\beta}_1'|_{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0}, \quad \boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1) = (\delta\boldsymbol{\beta}_1'\mathbf{F}_1\delta\boldsymbol{\beta}_1, , \ldots, \delta\boldsymbol{\beta}_1'\mathbf{F}_n\delta\boldsymbol{\beta}_1)',$$

$$\mathbf{F}_i = \partial^2 f_i(\boldsymbol{\beta}_1)/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'|_{\boldsymbol{\beta}_1 = \boldsymbol{\beta_1^0}}, \quad i = 1, \ldots, n,$$

and

$$\mathbf{g}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2 + \frac{1}{2}\boldsymbol{\omega}(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2) + \ldots,$$

where

$$\mathbf{b} = \mathbf{g}(\boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2^0), \qquad \mathbf{B}_1 = \frac{\partial \mathbf{g}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_1}\Big|_{\substack{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0 \\ \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0}}, \qquad \mathbf{B}_2 = \frac{\partial \mathbf{g}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}_2}\Big|_{\substack{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0 \\ \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0}}$$

and

$$\{\boldsymbol{\omega}(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2)\}_i = (\delta\boldsymbol{\beta}_1', \delta\boldsymbol{\beta}_2') \begin{pmatrix} \mathbf{A}, & \mathbf{B} \\ \mathbf{B}', & \mathbf{D} \end{pmatrix} \begin{pmatrix} \delta\boldsymbol{\beta}_1 \\ \delta\boldsymbol{\beta}_2 \end{pmatrix},$$

$$\mathbf{A} = \partial^2 g_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1'\Big|_{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0},$$

$$\mathbf{B} = \partial^2 g_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_2'\Big|_{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0},$$

$$\mathbf{D} = \partial^2 g_i(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)/\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2'\Big|_{\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^0, \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0},$$

$i = 1, \ldots, q$, $\delta\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^0$, $\delta\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1^0$.

After ommitting terms of the second and higher orders we get a linearized model

$$\mathbf{Y} - \mathbf{f}\left(\boldsymbol{\beta}_1^0\right) \sim_n \left(\mathbf{F}\left(\boldsymbol{\beta}_1^0\right)\delta\boldsymbol{\beta}_1, \boldsymbol{\Sigma}\right), \begin{pmatrix} \delta\boldsymbol{\beta}_1 \\ \delta\boldsymbol{\beta}_2 \end{pmatrix} \in \left\{ \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} : \mathbf{b} + \mathbf{B}_1\mathbf{u} + \mathbf{B}_2\mathbf{v} = \mathbf{0} \right\}.$$

If $h\left(\mathbf{F}\left(\boldsymbol{\beta}_1^0\right)\right) = k_1 < n$, $h\left(\mathbf{B}_1, \mathbf{B}_2\right) = q < k_1 + k_2$, $h\left(\mathbf{B}_2\right) = k_2 < q$, and $\boldsymbol{\Sigma}$ is a positive definite matrix we say that the model is regular. It is a linear model with type II constraints.

Let us denote shortly $\mathbf{f}_0 = \mathbf{f}\left(\boldsymbol{\beta}_1^0\right)$, $\mathbf{F} = \mathbf{F}\left(\boldsymbol{\beta}_1^0\right)$.

Lemma 2.1 *The best linear unbiased estimators (BLUE) of the parameters $\delta\boldsymbol{\beta}_1$, $\delta\boldsymbol{\beta}_2$ in the regular linearized model*

$$\mathbf{Y} - \mathbf{f}_0 \sim_n \left(\mathbf{F}\delta\boldsymbol{\beta}_1, \boldsymbol{\Sigma}\right), \quad \mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2 = \mathbf{0},$$

are

$$\widehat{\widehat{\delta\boldsymbol{\beta}_1}} = \widehat{\delta\boldsymbol{\beta}_1} - \mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^{+}\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right), \tag{1}$$

$$\widehat{\widehat{\delta\boldsymbol{\beta}_2}} = -\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right), \tag{2}$$

and their variance matrices are

$$\operatorname{var}\left(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right) = \left(\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{C}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right)^{+}, \tag{3}$$

$$\operatorname{var}\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) = \left[\mathbf{B}_2'\left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1}\mathbf{B}_2\right]^{-1} - \mathbf{I}, \tag{4}$$

where $\widehat{\delta\boldsymbol{\beta}_1} = \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{f}_0)$ *and* $\mathbf{C} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$.

Proof First we find a constrained extreme of the function

$$\left(\mathbf{Y} - \mathbf{f}_0 - \mathbf{F}\delta\boldsymbol{\beta}_1\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y} - \mathbf{f}_0 - \mathbf{F}\delta\boldsymbol{\beta}_1\right)$$

with a constraint $\mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2 = \mathbf{0}$. Derivatives of the Lagrange function

$$\Phi\left(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2\right) = \left(\mathbf{Y} - \mathbf{f}_0 - \mathbf{F}\delta\boldsymbol{\beta}_1\right)'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{f}_0 - \mathbf{F}\delta\boldsymbol{\beta}_1) - 2\boldsymbol{\lambda}'(\mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2)$$

are

$$\frac{\partial\Phi\left(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2\right)}{\partial\delta\boldsymbol{\beta}_1} = -2\mathbf{F}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y} - \mathbf{f}_0\right) + 2\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\delta\boldsymbol{\beta}_1 - 2\mathbf{B}_1'\boldsymbol{\lambda},$$

$$\frac{\partial\Phi\left(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2\right)}{\partial\delta\boldsymbol{\beta}_1} = -2\mathbf{B}_2'\boldsymbol{\lambda}.$$

We put both derivatives equal to a null vector and solve the ensuing system of equations. By first calculating an estimator of $\delta\boldsymbol{\beta}_1$ from the first equation for the model without constraints, i.e. for $\boldsymbol{\lambda} = \mathbf{0}$, we obtain

$$\widehat{\delta\boldsymbol{\beta}_1} = \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{f}_0),$$

where $\mathbf{C} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$, and therefore $\widehat{\widehat{\delta\boldsymbol{\beta}_1}} = \widehat{\delta\boldsymbol{\beta}_1} + \mathbf{C}^{-1}\mathbf{B}_1'\boldsymbol{\lambda}$. After substituting in the model the constraints $\mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2 = \mathbf{0}$ we solve, together with the second equation, a system

$$\begin{pmatrix} \mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1', & \mathbf{B}_2 \\ \mathbf{B}_2', & \mathbf{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{\lambda} \\ \widehat{\widehat{\delta\boldsymbol{\beta}_2}} \end{pmatrix} = \begin{pmatrix} -\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right) \\ \mathbf{0} \end{pmatrix}.$$

Using the Pandora-box matrix ([2, Lemma A.7.23]) in its special form ([2, Lemma A.7.24]) we obtain a solution

$$\begin{pmatrix} \boldsymbol{\lambda} \\ \widehat{\widehat{\delta\boldsymbol{\beta}_2}} \end{pmatrix} = \begin{pmatrix} \boxed{1} & \boxed{2} \\ \boxed{3} & \boxed{4} \end{pmatrix} \begin{pmatrix} -\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right) \\ \mathbf{0} \end{pmatrix},$$

where

$$\begin{aligned} \boxed{1} &= \left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+, \\ \boxed{2} &= (\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^-\mathbf{B}_2\left[\mathbf{B}_2'(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^-\mathbf{B}_2\right]^-, \\ \boxed{3} &= \boxed{2}', \\ \boxed{4} &= \left[\mathbf{B}_2'(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^-\mathbf{B}_2\right]^- - \mathbf{I}, \end{aligned}$$

and since ([2, Lemma A.7.9])

$$(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')} = (\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^-\mathbf{B}_2\left[\mathbf{B}_2'(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^-\mathbf{B}_2\right]^-$$

we can write

$$\begin{aligned} \boldsymbol{\lambda} &= -\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right), \\ \widehat{\widehat{\delta\boldsymbol{\beta}_2}} &= -\left[(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right), \\ \widehat{\widehat{\delta\boldsymbol{\beta}_1}} &= \widehat{\delta\boldsymbol{\beta}_1} + \mathbf{C}^{-1}\mathbf{B}_1'\boldsymbol{\lambda} = \widehat{\delta\boldsymbol{\beta}_1} - \mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\left(\mathbf{b} + \mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right). \end{aligned}$$

Variance matrices can be obtained as

$$\begin{aligned} \operatorname{var}\left[\begin{pmatrix} \widehat{\widehat{\delta\boldsymbol{\beta}_1}} \\ \widehat{\widehat{\delta\boldsymbol{\beta}_2}} \end{pmatrix}\right] &= \begin{pmatrix} \mathbf{I} - \mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1 \\ -\left[(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{B}_1 \end{pmatrix} \operatorname{var}(\widehat{\delta\boldsymbol{\beta}_1})\times \\ &\quad\times\left(\mathbf{I} - \mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1}, -\mathbf{B}_1'(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right). \end{aligned}$$

Since $\operatorname{var}(\widehat{\delta\boldsymbol{\beta}_1}) = \mathbf{C}^{-1}$ and using [2, Lemmas A.8.4 and A.8.5]

$$\begin{aligned} \operatorname{var}\left(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right) &= \left[\mathbf{I} - \mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\right]\mathbf{C}^{-1} \\ &\quad\times\left[\mathbf{I} - \mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1}\right] \\ &= \mathbf{C}^{-1} - 2\mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{B}_1' \\ &\quad\times\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1} \\ &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}_1'\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{B}_1\mathbf{C}^{-1} \\ &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1} \\ &= \left(\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{C}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right)^+, \end{aligned}$$

and similarly, when we denote $\mathbf{H} = \mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^- \mathbf{B}_2$,

$$\begin{aligned}
\operatorname{var}\left(\widehat{\widehat{\boldsymbol{\delta\beta}_2}}\right) &= \left[(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]' \mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')} \\
&= \mathbf{H}^-\mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \mathbf{B}_2\mathbf{H}^- \\
&= \mathbf{H}^-\mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2' - \mathbf{B}_2\mathbf{B}_2'\right) \\
&\quad \times \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \mathbf{B}_2\mathbf{H}^- \\
&= \mathbf{H}^{-1}\mathbf{H}\mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{H}\mathbf{H}\mathbf{H}^{-1} \\
&= \left[\mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \mathbf{B}_2\right]^{-1} - \mathbf{I},
\end{aligned}$$

because the matrix $\left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)$ can be expressed as multiplication of regular matrices (due to a model regularity)

$$\mathbf{B_1}\mathbf{C}^{-1}\mathbf{B_1'} + \mathbf{B_2}\mathbf{B_2'} = (\mathbf{B_1}, \mathbf{B_2}) \begin{pmatrix} \mathbf{C}^{-1} & ,\mathbf{0} \\ \mathbf{0} & ,\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{B_1'} \\ \mathbf{B_2'} \end{pmatrix},$$

and since we can use common inverse matrices instead of g-inverse matrices $\left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^-$ and $\left[\mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^- \mathbf{B}_2\right]^-$. □

Remark 2.1 Since (see [2, Lemmas A.7.24 and A.7.9])

$$\begin{aligned}
\left(\mathbf{M}_{\mathbf{B_2}}\mathbf{B_1}\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right)^+ &= \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} - \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \\
&\quad \times \mathbf{B}_2 \left[\mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1} \mathbf{B}_2\right]^{-1} \mathbf{B}_2' \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1},
\end{aligned}$$

and

$$(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')} = (\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^{-1}\mathbf{B}_2 \left[\mathbf{B}_2'(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2')^{-1}\mathbf{B}_2\right]^{-1},$$

the estimators of $\delta\boldsymbol{\beta}_1$ and $\delta\boldsymbol{\beta}_2$ in (1) and (2) can be expressed in equivalent forms without generalized inverse matrices

$$\widehat{\widehat{\boldsymbol{\delta\beta}_1}} = \widehat{\boldsymbol{\delta\beta}_1} - \mathbf{C}^{-1}\mathbf{B}_1' \left[\mathbf{T} - \mathbf{T}\mathbf{B}_2 \left(\mathbf{B}_2'\mathbf{T}\mathbf{B}_2\right)^{-1} \mathbf{B}_1'\mathbf{T}\right] \left(\mathbf{b} + \mathbf{B}_1\widehat{\boldsymbol{\delta\beta}_1}\right), \tag{5}$$

$$\widehat{\widehat{\boldsymbol{\delta\beta}_2}} = -\left(\mathbf{B}_2'\mathbf{T}\mathbf{B}_2\right)^{-1} \mathbf{B}_2'\mathbf{T} \left(\mathbf{b} + \mathbf{B}_1\widehat{\boldsymbol{\delta\beta}_1}\right), \tag{6}$$

where $\mathbf{T} = \left(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1}$.

Now we turn back to the model with quadratic terms and explore the properties (1)–(4) of the estimators.

Lemma 2.2 *If*

$$\mathbf{Y} - \mathbf{f}_0 \sim_n \left(\mathbf{F}\delta\boldsymbol{\beta}_1 + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1), \boldsymbol{\Sigma}\right), \quad \mathbf{b} + \mathbf{B}_1\delta\boldsymbol{\beta}_1 + \mathbf{B}_2\delta\boldsymbol{\beta}_2 + \frac{1}{2}\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1, \delta\boldsymbol{\beta}_2\right) = \mathbf{0}, \tag{7}$$

then biases of the estimators (1) and (2) are

$$\begin{aligned}\boldsymbol{b}_1 &= E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right) - \delta\boldsymbol{\beta}_1 = \frac{1}{2}\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)\\ &\quad+\frac{1}{2}\left[\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{C}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right]^+\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1),\\ \boldsymbol{b}_2 &= E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) - \delta\boldsymbol{\beta}_2\\ &= \frac{1}{2}\left[(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left(\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)-\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\right),\end{aligned}$$

where $\mathbf{C}=\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}$.

Proof By [2, Lemmas A.7.24 and A.8.4] and due to $\mathbf{M}_{\mathbf{B_2}}\mathbf{B_2}=\mathbf{0}$, we can write

$$\begin{aligned}&E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right) = E\left(\widehat{\delta\boldsymbol{\beta}_1} - \mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\left[\mathbf{b}+\mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right]\right)\\ &= -\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{b}\\ &\quad+\left[\mathbf{I}-\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{B}_1\right]E\left(\widehat{\delta\boldsymbol{\beta}_1}\right)\\ &= -\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{b}\\ &\quad+\left[\mathbf{I}-\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{B}_1\right]\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{F}\delta\boldsymbol{\beta}_1+\frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\right)\\ &= \delta\boldsymbol{\beta}_1-\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\left(\mathbf{b}+\mathbf{B}_1\delta\boldsymbol{\beta}_1\right)\\ &\quad+\frac{1}{2}\left[\mathbf{I}-\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{B}_1\right]\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\\ &= \delta\boldsymbol{\beta}_1+\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{M}_{\mathbf{B_2}}\left(\mathbf{B}_2\delta\boldsymbol{\beta}_2+\frac{1}{2}\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)\right)\\ &\quad+\frac{1}{2}\left[\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\right]\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\\ &= \delta\boldsymbol{\beta}_1+\frac{1}{2}\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)\\ &\quad+\frac{1}{2}\left[\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{C}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right]^+\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1).\end{aligned}$$

Then

$$\begin{aligned}\boldsymbol{b}_1 &= E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right) - \delta\boldsymbol{\beta}_1 = \frac{1}{2}\mathbf{C}^{-1}\mathbf{B}_1'\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^+\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)\\ &\quad+\frac{1}{2}\left[\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{C}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right]^+\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1).\end{aligned}$$

Similarly by [2, Lemma A.7.20] and due to

$$\left[(\mathbf{B}_2')^-_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{B}_2=\mathbf{I}$$

we obtain

$$
\begin{aligned}
E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) &= E\left(-\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left(\mathbf{b}+\mathbf{B}_1\widehat{\delta\boldsymbol{\beta}_1}\right)\right)\\
&= -\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{b}-\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}E\left(\mathbf{Y}-\mathbf{f}_0\right)\\
&= -\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{b}-\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\\
&\quad\times\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{F}\delta\boldsymbol{\beta}_1+\frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\right)\\
&= -\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'(\mathbf{b}+\mathbf{B}_1\delta\boldsymbol{\beta}_1)-\frac{1}{2}\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\\
&\quad\times\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\\
&= \left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left(\mathbf{B}_2\delta\boldsymbol{\beta}_2+\frac{1}{2}\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)\right)\\
&\quad-\frac{1}{2}\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\\
&= \delta\boldsymbol{\beta}_2+\frac{1}{2}\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left[\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)-\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\right],
\end{aligned}
$$

and therefore

$$
\begin{aligned}
\boldsymbol{b}_2 &= E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right)-\delta\boldsymbol{\beta}_2\\
&= \frac{1}{2}\left[(\mathbf{B}_2')^{-}_{m(\mathbf{B}_1\mathbf{C}^{-1}\mathbf{B}_1')}\right]'\left[\boldsymbol{\omega}\left(\delta\boldsymbol{\beta}_1,\delta\boldsymbol{\beta}_2\right)-\mathbf{B}_1\mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}_1)\right]. \qquad \square
\end{aligned}
$$

3 Measures of nonlinearity and areas of linearization

In this section we suppose the observation vector to be normally distributed. Bias of an estimator of $\delta\boldsymbol{\beta}_2$ can be split into components, i.e.

$$
\boldsymbol{b}_2 = E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right)-\delta\boldsymbol{\beta}_2 = \boldsymbol{b}_{2,0}+\boldsymbol{b}_{2,1},
$$

where

$$
\boldsymbol{b}_{2,0}\in\mathcal{M}\left(\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})\right) \quad\text{and}\quad \boldsymbol{b}_{2,1}\in\mathcal{M}\left(\mathbf{M}_{\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}\right),
$$

as can be seen in Fig. 1.

Let a symbol $\lambda_{\max}$ denote the biggest eigenvalue of the matrix $\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})$. By Theorem 9.2.1 in [3] it is easy to prove that for

$$
\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\sim N_{k_2}(\delta\boldsymbol{\beta}_2+\boldsymbol{b}_2,\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}))
$$

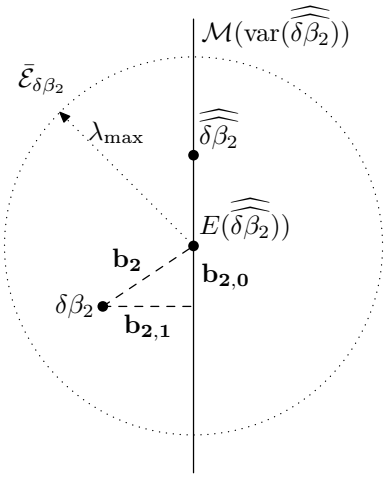


Figure 1: The components of bias.

the random variable

$$T = \left[\widehat{\widehat{\delta\boldsymbol{\beta}_2}} - E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) + \boldsymbol{b}_{2,0}\right]' \left(\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) + \lambda_{\max}\mathbf{M}_{\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}\right)^+ \times \left[\widehat{\widehat{\delta\boldsymbol{\beta}_2}} - E\left(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) + \boldsymbol{b}_{2,0}\right]$$

has a noncentral χ^2 distribution with $f = h(\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}))$ degrees of freedom and a parameter of noncentrality

$$\delta = \boldsymbol{b}_{2,0}' \left(\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) + \lambda_{\max}\mathbf{M}_{\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}\right)^+ \boldsymbol{b}_{2,0}. \tag{8}$$

A random variable

$$\bar{T} = \left(\delta\boldsymbol{\beta}_2 - \widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right)' \left(\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) + \lambda_{\max}\mathbf{M}_{\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}\right)^+ \left(\delta\boldsymbol{\beta}_2 - \widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right)$$

can be then rewritten in the form

$$\bar{T} = T + \frac{\boldsymbol{b}_{2,1}'\boldsymbol{b}_{2,1}}{\lambda_{\max}},$$

because by [2, Lemmas A.7.22 and A.7.2] it holds that

$$\left[\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) + \lambda_{\max}\mathbf{M}_{\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}\right]^+ = \left[\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})\right]^+ + \frac{1}{\lambda_{\max}}\mathbf{M}_{\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})}$$

and

$$\boldsymbol{b}_{2,1}' \left[\mathrm{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})\right]^+ \boldsymbol{b}_{2,1} = \mathbf{0}.$$

This consideration leads us to a modified confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_2$.

Definition 3.1 A modified confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_2$ in the model (7) is defined as

$$\bar{\mathcal{E}}_{\delta\boldsymbol{\beta}_2} = \left\{ \boldsymbol{u} \in \mathbb{R}^{k_2} \colon \left(\boldsymbol{u} - \widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right)' \left\{ \left[\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) \right]^+ + \frac{1}{\lambda_{\max}} \mathbf{M}_{\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})} \right\} \left(\boldsymbol{u} - \widehat{\widehat{\delta\boldsymbol{\beta}_2}}\right) \leq \chi^2_f(0; 1-\alpha) \right\},$$

where $f = h(\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}))$.

As a certain analogy of the Bates-Wats measure of curvature, a measure of nonlinearity for a confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_2$ can be defined.

Definition 3.2 For a linear model with type II constraints in the form (7), we define a measure of nonlinearity of confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_2$ as

$$\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2} = \sup \left\{ \frac{\sqrt{\boldsymbol{b}_2' \left\{ \left[\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}}) \right]^+ + \frac{1}{\lambda_{\max}} \mathbf{M}_{\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_2}})} \right\} \boldsymbol{b}_2}}{\delta\boldsymbol{s}' \mathbf{K}_1' \left\{ \left[\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}) \right]^+ + \frac{1}{\kappa_{\max}} \mathbf{M}_{\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_1}})} \right\} \mathbf{K}_1 \delta\boldsymbol{s}} \colon \delta\boldsymbol{s} \in \mathbb{R}^{k_1+k_2-q} \right\}, \tag{9}$$

where $\kappa_{\max}$ is the biggest eigenvalue of the matrix $\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_1}})$ and $\mathbf{K}_1$ is a matrix of type $k_1 \times (k_1 + k_2 - q)$ satisfying $\mathcal{M}(\mathbf{K}_1) = \mathcal{M}(\mathbf{M}_{\mathbf{B}_1' \mathbf{M}_{\mathbf{B}_2}})$.

It is obvious that

$$\mathcal{P}\left\{ \bar{T} \leq \chi^2_f(0; 1-\alpha) \right\} = \mathcal{P}\left\{ \chi^2_f(\delta) + \frac{\boldsymbol{b}_{2,1}' \boldsymbol{b}_{2,1}}{\lambda_{\max}} \leq \chi^2_f(0; 1-\alpha) \right\}$$

and certainly such $\delta_0 > 0$ exists which satisfies the equality

$$\mathcal{P}\left\{ \chi^2_f(\delta_0) + \delta_0 \leq \chi^2_f(0; 1-\alpha) \right\} = 1 - \alpha - \epsilon \tag{10}$$

for a sufficiently small $\epsilon > 0$. Now we define an area of linearization of the parameter $\delta\boldsymbol{\beta}_2$ for this δ_0.

Definition 3.3 An area of linearization of the parameter $\delta\boldsymbol{\beta}_2$ for the model (7) is

$$\mathcal{L}_{\delta\boldsymbol{\beta}_2} = \left\{ \mathbf{K}_1 \delta\boldsymbol{s} \colon \delta\boldsymbol{s}' \mathbf{K}_1' \left\{ \left[\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_1}}) \right]^+ + \frac{1}{\kappa_{\max}} \mathbf{M}_{\operatorname{var}(\widehat{\widehat{\delta\boldsymbol{\beta}_1}})} \right\} \mathbf{K}_1 \delta\boldsymbol{s} \leq \frac{\sqrt{\delta_0}}{\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2}}, \delta\boldsymbol{s} \in \mathbb{R}^{k_1+k_2-q} \right\},$$

where the matrix $\mathbf{K}_1$ has properties mentioned in Definition 3.2.

Lemma 3.1 *If* $\mathbf{K}_1\delta\boldsymbol{s} \in \mathcal{L}_{\delta\boldsymbol{\beta}_2}$, *then*

$$\mathcal{P}\left\{\delta\boldsymbol{\beta}_2 \in \bar{\mathcal{E}}_{\delta\boldsymbol{\beta}_2}\right\} \geq 1-\alpha-\epsilon.$$

Proof By the definition of $\mathcal{L}_{\delta\boldsymbol{\beta}_2}$ and $C^{II}_{ell,\delta\boldsymbol{\beta}_2}$, we can write

$$\sqrt{\boldsymbol{b}_2'\left\{\left[\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_2)}}\right]^+ + \frac{1}{\lambda_{\max}}\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_2)}}}\right\}\boldsymbol{b}_2}$$
$$\leq \mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2}\delta\boldsymbol{s}'\mathbf{K}_1'\left\{\left[\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}\right]^+ + \frac{1}{\kappa_{\max}}\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}}\right\}\mathbf{K}_1\delta\boldsymbol{s} \leq \sqrt{\delta_0}.$$

Since, with respect to $\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_2)}}}\boldsymbol{b}_{2,0} = \mathbf{0}$,

$$\mathcal{P}\left\{\delta\boldsymbol{\beta}_2 \in \bar{\mathcal{E}}_{\delta\boldsymbol{\beta}_2}\right\} = \mathcal{P}\left\{\bar{T} \leq \chi^2_f(0;1-\alpha)\right\} = \mathcal{P}\left\{\chi^2_f(\delta) + \frac{\boldsymbol{b}_{2,1}'\boldsymbol{b}_{2,1}}{\lambda_{\max}} \leq \chi^2_f(0;1-\alpha)\right\}$$
$$\geq \mathcal{P}\left\{\chi^2_f(\delta_0) + \delta_0 \leq \chi^2_f(0;1-\alpha)\right\} = 1-\alpha-\epsilon. \qquad \square$$

Because the parameter $\delta\boldsymbol{\beta}_2$ is a function of the parameter $\delta\boldsymbol{\beta}_1$ we must, in order to verify of the property $\delta\boldsymbol{\beta}_1 \approx \mathbf{K}_1\delta\boldsymbol{s} \in \mathcal{L}_{\delta\boldsymbol{\beta}_2}$, construct also a modified confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_1$.

Definition 3.4 A modified confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_1$ in the model (7) is

$$\bar{\mathcal{E}}_{\delta\boldsymbol{\beta}_1} = \left\{\boldsymbol{u} \in \mathbb{R}^{k_1}:\ \left(\boldsymbol{u} - \widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right)'\left\{\left[\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}\right]^+ + \frac{1}{\kappa_{\max}}\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}}\right\}\left(\boldsymbol{u} - \widehat{\widehat{\delta\boldsymbol{\beta}_1}}\right)\right.$$
$$\left. \leq \chi^2_{f_1}(0;1-\alpha)\right\},$$

where $f_1 = h(\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)})}$ and $\kappa_{\max}$ is the biggest eigenvalue of the matrix $\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}$.

Similarly as for $\delta\boldsymbol{\beta}_2$, it is also possible to define a measure of nonlinearity for $\delta\boldsymbol{\beta}_1$.

Definition 3.5 For the linear model (7), we define a measure of nonlinearity of a confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_1$ as

$$\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_1} = \sup\left\{\frac{\sqrt{\boldsymbol{b}_1'\left\{\left[\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}\right]^+ + \frac{1}{\kappa_{\max}}\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}}\right\}\boldsymbol{b}_1}}{\delta\boldsymbol{s}'\mathbf{K}_1'\left\{\left[\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}\right]^+ + \frac{1}{\kappa_{\max}}\mathbf{M}_{\widehat{\widehat{\operatorname{var}(\delta\boldsymbol{\beta}_1)}}}\right\}\mathbf{K}_1\delta\boldsymbol{s}}: \delta\boldsymbol{s} \in \mathbb{R}^{k_1+k_2-q}\right\}. \tag{11}$$

A sufficient condition for linearization regarding the confidence ellipsoid for the parameter $\delta\boldsymbol{\beta}_2$ is

$$\mathcal{E}_{\delta\boldsymbol{\beta}_1} \subset\subset \mathcal{L}_{\delta\boldsymbol{\beta}_2} \Rightarrow \frac{\sqrt{\delta_0}}{\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2}} \gg \chi^2_{f_1}(0; 1-\alpha), \tag{12}$$

(cf. Fig. 2).

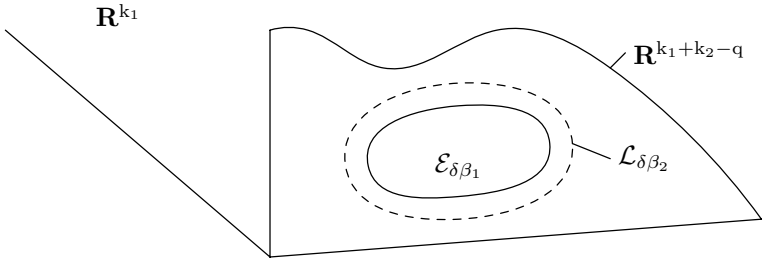


Figure 2: The confidence ellipsoid $\mathcal{E}_{\delta\boldsymbol{\beta}_1}$ and the area of linearization $\mathcal{L}_{\delta\boldsymbol{\beta}_2}$.

4 Numerical example

Tracer kinetics of liver blood flow can be described by a compartmental model (Fig. 3) and an ordinary differential equation

$$\frac{dC_L(t)}{dt} = k_{1a}C_a(t) + k_{1p}C_p(t) - k_2C_L(t). \tag{13}$$

We obtained the values of tracer concentration $C_L(t_i)$ in liver, $C_a(t_i)$ in a liver artery and $C_p(t_i)$ in a portal vein by measuring times t_i, $i = 1, 2, \ldots, n$.

To the equation (13) we can add a delay, in the liver artery or in the portal vein or both. So overall, we can obtain three different equations for our compartmental model (included the one without any delay):

$$\text{(KMI)} \quad \frac{dC_L(t)}{dt} = k_{1a}C_a(t) + k_{1p}C_p(t) - k_2C_L(t),$$

$$\text{(KMII)} \quad \frac{dC_L(t)}{dt} = k_{1a}C_a(t-\tau_a) + k_{1p}C_p(t) - k_2C_L(t),$$

$$\text{(KMIII)} \quad \frac{dC_L(t)}{dt} = k_{1a}C_a(t-\tau_a) + k_{1p}C_p(t-\tau_p) - k_2C_L(t).$$

For the sake of simplicity, let us consider only the model without any delay, denoted as (KMI). A vector of observations of tracer concentrations for this model is in the form

$$\mathbf{Y} = (C_a(t_1), \ldots, C_a(t_{n-1}), C_p(t_1), \ldots, C_p(t_{n-1}), C_L(t_1), \ldots, C_L(t_n))',$$

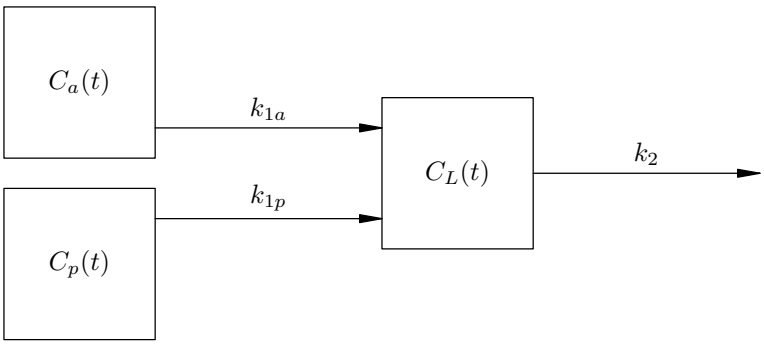


Figure 3: Dual-input one-compartmental model of blood flow in liver.

and a statistical model

$$\mathbf{Y} \sim N_{3n-2}\left(\mathbf{I}\boldsymbol{\beta_1}, \sigma^2\mathbf{I}\right), \tag{14}$$

where $\boldsymbol{\beta_1} = (\mu_1, \dots \mu_{n-1}, \nu_1, \dots, \nu_{n-1}, \zeta_1, \dots, \zeta_n)'$, with constraints

$$\frac{\zeta_{i+1} - \zeta_i}{t_{i+1} - t_i} = k_{1a}\mu_i + k_{1p}\nu_i - k_2\zeta_i, \quad i = 1, 2, \dots, n-1.$$

Let for $i = 1, 2, \dots, n-1$

$$\mu_i = \mu_i^{(0)} + \delta\mu_i, \quad \nu_i = \nu_i^{(0)} + \delta\nu_i, \quad \zeta_i = \zeta_i^{(0)} + \delta\zeta_i,$$

then for

$$\mathbf{Z} = \mathbf{Y} - \left(\mu_1^{(0)}, \dots \mu_{n-1}^{(0)}, \nu_1^{(0)}, \dots, \nu_{n-1}^{(0)}, \zeta_1^{(0)}, \dots, \zeta_n^{(0)}\right)'$$

we have a model

$$\mathbf{Z} \sim N_{3n-2}\left(\mathbf{I}\boldsymbol{\delta\beta_1}, \sigma^2\mathbf{I}\right),$$

where

$$\boldsymbol{\delta\beta_1} = (\delta\mu_1, \dots \delta\mu_{n-1}, \delta\nu_1, \dots, \delta\nu_{n-1}, \delta\zeta_1, \dots, \delta\zeta_n)'.$$

Then for $k_{1a} = k_{1a}^{(0)} + \delta k_{1a}$, $k_{1p} = k_{1p}^{(0)} + \delta k_{1p}$, $k_2 = k_2^{(0)} + \delta k_2$ and

$$\boldsymbol{\beta_2} = \begin{pmatrix} k_{1a} \\ k_{1p} \\ k_2 \end{pmatrix}, \quad \boldsymbol{\delta\beta_2} = \begin{pmatrix} \delta k_{1a} \\ \delta k_{1p} \\ \delta k_2 \end{pmatrix},$$

the model constraints

$$g_i\left(\beta_1, \beta_2\right) = -k_{1a}\mu_i - k_{1p}\nu_i + \left(k_2 - \frac{1}{t_{i+1} - t_i}\right)\zeta_i + \frac{1}{t_{i+1} - t_i}\zeta_{i+1} = 0,$$

$i = 1, 2, \dots, n-1$, can be rewritten in the form

$$g_i(\beta_1,\beta_2) = -\left(k_{1a}^{(0)}+\delta k_{1a}\right)\left(\mu_i^{(0)}+\delta\mu_i\right) - \left(k_{1p}^{(0)}+\delta k_{1p}\right)\left(\nu_i^{(0)}+\delta\nu_i\right)$$
$$+\left(k_2^{(0)}+\delta k_2-\frac{1}{t_{i+1}-t_i}\right)\left(\zeta_i^{(0)}+\delta\zeta_i\right)+\frac{1}{t_{i+1}-t_i}\left(\zeta_{i+1}^{(0)}+\delta\zeta_{i+1}\right)=0,$$
$$i=1,\ldots,n-1.$$

In a matrix form we can write

$$\boldsymbol{g}(\beta_1,\beta_2) = \mathbf{b}+(\mathbf{B}_1,\mathbf{B}_2)\begin{pmatrix}\boldsymbol{\delta\beta_1}\\ \boldsymbol{\delta\beta_2}\end{pmatrix}+\frac{1}{2}\begin{bmatrix}\left(\boldsymbol{\delta\beta_1}',\boldsymbol{\delta\beta_2}'\right)\frac{\partial^2 g_1(\boldsymbol{\beta_1},\boldsymbol{\beta_2})}{\partial\binom{\boldsymbol{\beta_1}}{\boldsymbol{\beta_2}}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\begin{pmatrix}\boldsymbol{\delta\beta_1}\\ \boldsymbol{\delta\beta_2}\end{pmatrix}\\ \vdots\\ \left(\boldsymbol{\delta\beta_1}',\boldsymbol{\delta\beta_2}'\right)\frac{\partial^2 g_{n-1}(\boldsymbol{\beta_1},\boldsymbol{\beta_2})}{\partial\binom{\boldsymbol{\beta_1}}{\boldsymbol{\beta_2}}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\begin{pmatrix}\boldsymbol{\delta\beta_1}\\ \boldsymbol{\delta\beta_2}\end{pmatrix}\end{bmatrix},$$

where for $\Delta t_i = t_{i+1}-t_i$, $i=1,\ldots,n-1$,

$$b_i = -k_{1a}^{(0)}\mu_i^{(0)} - k_{1p}^{(0)}\nu_i^{(0)} + \left(k_2^{(0)}-\frac{1}{\Delta t_i}\right)\zeta_i^{(0)}+\frac{1}{\Delta t_i}\zeta_{i+1}^{(0)},\quad i=1,\ldots,n-1,$$

a matrix $\mathbf{B}_1$ is of type $(n-1)\times(3n-2)$ and it should be divided into three blocks

$$\mathbf{B}_1 = \left[-k_{1a}^{(0)}\mathbf{I}_{n-1}, -k_{1p}^{(0)}\mathbf{I}_{n-1}, \boxed{1}\right],$$

where

$$\boxed{1} = \begin{bmatrix} k_2^{(0)}-\frac{1}{\Delta t_1}, & \frac{1}{\Delta t_1}, & 0, & 0,\ldots, & 0, & 0\\ 0, & k_2^{(0)}-\frac{1}{\Delta t_2}, & \frac{1}{\Delta t_2}, & 0,\ldots, & 0, & 0\\ \vdots & \vdots & \vdots & \vdots\ \vdots & \vdots & \vdots\\ 0, & 0, & 0, & 0,\ldots, & k_2^{(0)}-\frac{1}{\Delta t_{n-1}}, & \frac{1}{\Delta t_{n-1}}\end{bmatrix},$$

$$\mathbf{B}_2 = \begin{pmatrix} -\mu_1^{(0)}, & -\nu_1^{(0)}, & \zeta_1^{(0)}\\ -\mu_2^{(0)}, & -\nu_2^{(0)}, & \zeta_2^{(0)}\\ \vdots & \vdots & \vdots\\ -\mu_{n-1}^{(0)}, & -\nu_{n-1}^{(0)}, & \zeta_{n-1}^{(0)}\end{pmatrix}.$$

For $i=1,\ldots,n-1$ the $(3n+1)\times(3n+1)$ matrix

$$\frac{\partial^2 g_i(\boldsymbol{\beta_1},\boldsymbol{\beta_2})}{\partial\begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}$$

has almost all elements equal to zero except for

$$\left\{\frac{\partial^2 g_i\left(\boldsymbol{\beta_1},\boldsymbol{\beta_2}\right)}{\partial\begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\right\}_{3n-1,i} = -\frac{1}{2}$$

$$\left\{\frac{\partial^2 g_i\left(\boldsymbol{\beta_1},\boldsymbol{\beta_2}\right)}{\partial\begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\right\}_{3n,n+i-1} = -\frac{1}{2}$$

$$\left\{\frac{\partial^2 g_i\left(\boldsymbol{\beta_1},\boldsymbol{\beta_2}\right)}{\partial\begin{pmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{pmatrix}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\right\}_{3n+1,2n+i-2} = +\frac{1}{2}$$

and the corresponding symmetric elements.

Calculation of estimators of $\delta\boldsymbol{\beta}_1$ and $\delta\boldsymbol{\beta}_2$ is iterative. For initiative iteration we put

$$\mu_i^{(1)} = C_a(t_i), \quad \nu_i^{(1)} = C_p(t_i), \quad \zeta_i^{(1)} = C_L(t_i), \quad i = 1,\ldots,n-1,$$

and $k_{1a}^{(1)}$, $k_{1p}^{(1)}$, $k_2^{(1)}$ are calculated as a solution to a system

$$\mathbf{B}_2\begin{pmatrix}k_{1a}^{(1)}\\ k_{1p}^{(1)}\\ k_2^{(1)}\end{pmatrix} = \begin{pmatrix}\frac{\zeta_1^{(1)}-\zeta_2^{(1)}}{\Delta t_1}\\ \vdots\\ \frac{\zeta_{n-1}^{(1)}-\zeta_n^{(1)}}{\Delta t_{n-1}}\end{pmatrix},$$

i.e. from the model constraints for $\delta\boldsymbol{\beta}_1 = \mathbf{0}$ and $\delta\boldsymbol{\beta}_2 = \mathbf{0}$.

From (5), (6) we calculate the $(k+1)$-th iteration of estimators of $\delta\boldsymbol{\beta}_1$ and $\delta\boldsymbol{\beta}_2$, i.e. in this case

$$\begin{aligned}\delta\boldsymbol{\beta}_1^{(k+1)} &= \mathbf{Z}^{(k)} - \mathbf{B}_1'\left[\mathbf{T} - \mathbf{T}\mathbf{B}_2\left[\mathbf{B}_2'\mathbf{T}\mathbf{B}_2\right]^{-1}\mathbf{B}_2'\mathbf{T}\right]\left(\mathbf{b}^{(k)} + \mathbf{B}_1\mathbf{Z}^{(k)}\right)\\ \delta\boldsymbol{\beta}_2^{(k+1)} &= -\left[\mathbf{B}_2'\mathbf{T}\mathbf{B}_2\right]^{-1}\mathbf{B}_2'\mathbf{T}\left(\mathbf{b}^{(k)} + \mathbf{B}_1\mathbf{Z}^{(k)}\right),\end{aligned}$$

where $\mathbf{T} = \left(\mathbf{B}_1\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1}$, $\mathbf{Z}^{(k)} = \mathbf{Y} - \boldsymbol{\beta_1}^{(k)}$, and

$$\mathbf{b}^{(k)} = \mathbf{B_1}\boldsymbol{\beta_1}^{(k)} + \frac{1}{2}\begin{bmatrix}\left(\boldsymbol{\delta\beta_1}^{(k)\prime}, \boldsymbol{\delta\beta_2}^{(k)\prime}\right)\frac{\partial^2 g_1\left(\boldsymbol{\beta_1}^{(k)},\boldsymbol{\beta_2}^{(k)}\right)}{\partial\binom{\boldsymbol{\beta_1}}{\boldsymbol{\beta_2}}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\begin{pmatrix}\boldsymbol{\delta\beta_1}^{(k)}\\ \boldsymbol{\delta\beta_2}^{(k)}\end{pmatrix}\\ \vdots\\ \left(\boldsymbol{\delta\beta_1}^{(k)\prime}, \boldsymbol{\delta\beta_2}^{(k)\prime}\right)\frac{\partial^2 g_{n-1}\left(\boldsymbol{\beta_1}^{(k)},\boldsymbol{\beta_2}^{(k)}\right)}{\partial\binom{\boldsymbol{\beta_1}}{\boldsymbol{\beta_2}}\partial\left(\boldsymbol{\beta_1}',\boldsymbol{\beta_2}'\right)}\begin{pmatrix}\boldsymbol{\delta\beta_1}^{(k)}\\ \boldsymbol{\delta\beta_2}^{(k)}\end{pmatrix}\end{bmatrix}.$$

The matrices $\mathbf{B}_1$, $\mathbf{B}_2$ are constructed with the k-th iteration of the parameters $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$ obtained from

$$\begin{aligned}\boldsymbol{\beta_1}^{(k)} &= \boldsymbol{\beta_1}^{(k-1)} + \boldsymbol{\delta\beta_1}^{(k)},\\ \boldsymbol{\beta_2}^{(k)} &= \boldsymbol{\beta_2}^{(k-1)} + \boldsymbol{\delta\beta_2}^{(k)}.\end{aligned}$$

Estimators of covariance matrices of the final estimators $\widehat{\widehat{\boldsymbol{\delta\beta_1}}}$, $\widehat{\widehat{\boldsymbol{\delta\beta_2}}}$ are calculated from (3), (4), i.e. in this case ($\mathbf{C} = \sigma^{-2}\mathbf{I}$)

$$\begin{aligned}\widehat{\operatorname{var}(\widehat{\widehat{\boldsymbol{\delta\beta_1}}})} = \widehat{\operatorname{var}(\widehat{\widehat{\boldsymbol{\beta_1}}})} &= \widehat{\sigma^2}\left(\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right)^{+},\\ \widehat{\operatorname{var}(\widehat{\widehat{\boldsymbol{\delta\beta_2}}})} = \widehat{\operatorname{var}(\widehat{\widehat{\boldsymbol{\beta_2}}})} &= \widehat{\sigma^2}\left(\left[\mathbf{B}_2'\left(\mathbf{B}_1\mathbf{B}_1' + \mathbf{B}_2\mathbf{B}_2'\right)^{-1}\mathbf{B}_2\right]^{-1} - \mathbf{I}\right),\end{aligned}$$

where

$$\widehat{\sigma^2} = \frac{\left(\mathbf{Y} - \widehat{\widehat{\boldsymbol{\beta_1}}}\right)'\left(\mathbf{Y} - \widehat{\widehat{\boldsymbol{\beta_1}}}\right)}{n + q - (k_1 + k_2)}$$

and

$$\left(\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\mathbf{M}_{\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}}\right)^{+} = \left(\mathbf{I} - \mathbf{B}_1\left[\mathbf{M}_{\mathbf{B_2}}\mathbf{B}_1\mathbf{B}_1'\mathbf{M}_{\mathbf{B_2}}\right]^{+}\mathbf{B}_1\right).$$

For data from the graphic example in [4] (values of tracer concentration in liver, artery and portal vein measures at 23 times—see Table 1 and Fig. 4), i.e. for $n = 23$, $q = n - 1 = 22$, $k_1 = 3n - 2 = 67$ and $k_2 = 3$, we get these results after 4 iterations:

$$\widehat{\widehat{\boldsymbol{\beta_2}}} = \begin{pmatrix}0.002431475\\ 0.009413782\\ 0.039506253\end{pmatrix},$$

$$\widehat{\sigma^2} = 0.001130171,$$

$$\operatorname{var}\left(\widehat{\widehat{\boldsymbol{\beta_2}}}\right) = \operatorname{var}\left(\widehat{\widehat{\boldsymbol{\delta\beta_2}}}\right)$$

$$= \begin{pmatrix}3.238255e-07 & -6.991068e-07 & -2.103772e-06\\ -6.991068e-07 & 3.001722e-06 & 1.255561e-05\\ -2.103772e-06 & 1.255561e-05 & 5.826697e-05\end{pmatrix}.$$

Among the results we were interested only in the vector of kinetics parameters $\boldsymbol{\beta_2}$, because they seem to be important for early diagnosis of substantional liver diseases.

In Fig. 5 there are discrete points of measured tracer concentration in liver and a curve of the tracer concentration in liver estimated from the model (i.e. $\zeta_1, \ldots, \zeta_n$ values).

Now we calculate the measure of nonlinearity $\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2}$ by algorithm mentioned in [1] (pp. 230–231) with the value of δ_0 from (10) set at $\epsilon = 0.04$. The value of

$$\frac{\sqrt{\delta_0}}{\mathcal{C}^{II}_{ell,\delta\boldsymbol{\beta}_2}} = \frac{\sqrt{1.570312}}{0.04511495} = 27.77618$$

is compared with the value of $\chi^2_{48}(0; 0.95) = 65.17077$. From the numerical results it is obvious that the condition mentioned in (12) is not satisfied, i.e. for our data set it is not suitable to linearize the original nonlinear model and work with the estimators of kinetics coefficients obtained from the linearized model, although these estimators seem to be very accurate. If the estimated parameter $\widehat{\sigma}$ was three times lower, which might be accomplished by more accurate measurement or by measurement in shorter time intervals, the condition would be satisfied and linearization would be appropriate.

i	t_i [s]	$C_L(t_i)$ [mmol/l]	$C_a(t_i)$ [mmol/l]	$C_p(t_i)$ [mmol/l]
1	0.00	0.000	0.000	0.00
2	3.30	0.000	0.000	0.00
3	6.75	0.000	2.350	0.00
4	10.00	0.000	4.230	0.07
5	13.25	0.030	4.350	0.19
6	16.75	0.111	3.620	0.68
7	20.00	0.156	2.440	1.36
8	23.50	0.126	1.600	1.88
9	26.75	0.204	1.220	2.11
10	30.00	0.309	1.220	2.49
11	33.50	0.294	1.500	2.30
12	36.75	0.360	2.000	2.21
13	40.50	0.378	2.230	2.26
14	43.50	0.411	2.162	2.21
15	47.00	0.489	1.970	2.40
16	50.50	0.519	1.790	2.28
17	54.00	0.561	1.600	2.35
18	57.00	0.516	1.480	2.26
19	60.50	0.618	1.580	2.23
20	64.00	0.543	1.530	2.16
21	67.00	0.561	1.620	2.26
22	70.50	0.510	1.430	2.16
23	74.00	0.600	1.430	2.07

Table 1: Measured data of tracer concentration.

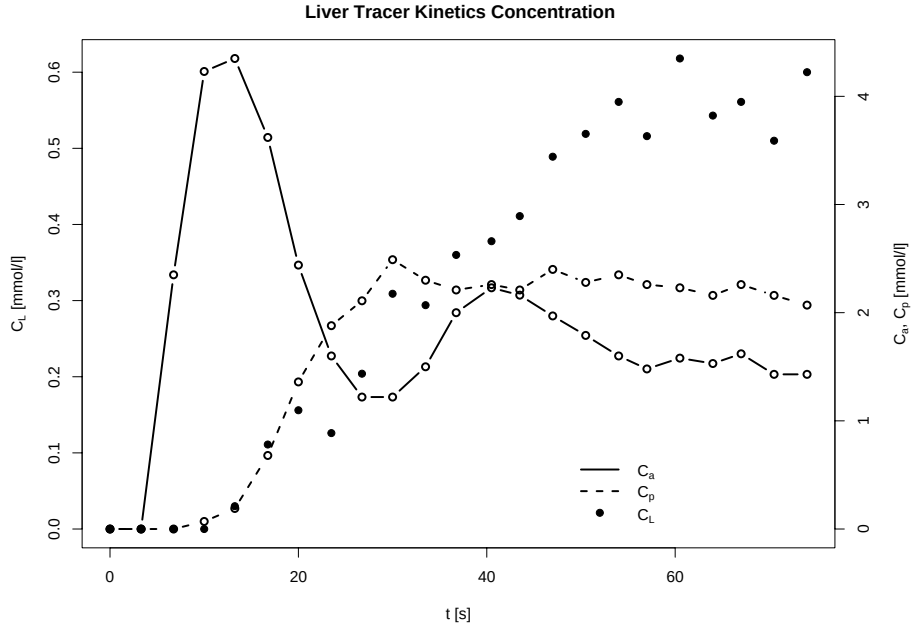


Figure 4: Curves of measured tracer concentration in a liver artery $C_a(t)$ and a portal vein $C_p(t)$ and points of measured tracer concentration in liver $C_L(t)$.

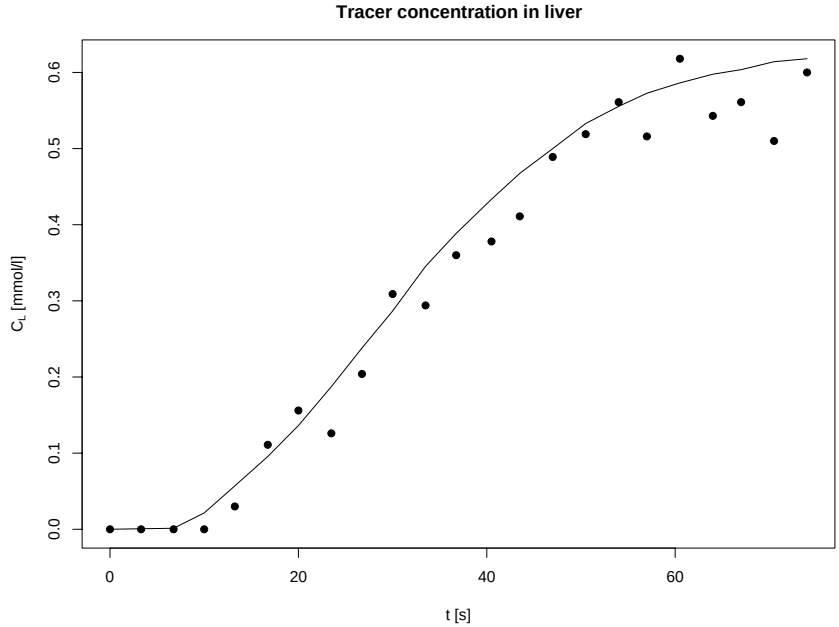


Figure 5: Measured (points) and estimated (a curve) tracer concentration in liver.

5 Conclusions

Many real-life systems are basicaly nonlinear. Particularly in biology and medicine we meet nonlinear problems very often. By treating them as linear we employ a very rough and limited approximation [5]. There are many methods that solve nonlinear problems, mostly numerical methods, but these usualy suppose accurate measurements, and they do not take into consideration inaccuracy and uncertainty inherent in biology and medicine settings (subjective examination, inter- or intraobjective variability and so on). One way out is to apply linearization of nonlinear problems, for example the above-mentioned linearization via Taylor series, to use the well-known and well-explored theory of linear models. We know how to estimate parameters and their variability in the linearized models [1]. However, we should check whether the type of problem and measured data allow for treating the nonlinear problem in this way.

The aim of this article was to find a condition which would guarantee for linear models with type II constraints that the true values of estimated parameters are covered by a modified confidence ellipsoid (with probability no less than $1 - \alpha - \epsilon$ for a preset small $\epsilon > 0$), and to verify in this manner that the usage of linearization is appropriate. As can be seen in the numerical example, this condition is not easy to satisfy, although calculated estimators (and their variances) in the linearized model look very good. When solving a nonlinear problem by linearization we should proove that the linearization is safe. In case of linear models with type II constraints a method of such verification was presented here.

References

[1] Kubáček, L., Kubáčková, L.: Statistics and Metrology. *Vyd. Univ. Palackého, Olomouc,* 2000 (in Czech).

[2] Fišerová, E., Kubáček, L., Kunderová, P.: Linear Statistical Models, Regularity and Singularities. *Academia, Praha,* 2007.

[3] Rao, C. R., Mitra, S. K.: Generalized Inverse of Matrices and its Applications. *J. Wiley, New York–London–Sydney–Toronto*, 1971.

[4] Hagiwara, M., Rusinek, H., Lee, V. S., Losada, M., Bannan, M. A., Krinsky, G. A., Taouli, B.: *Advanced Liver Fibrosis: Diagnosis with 3D Whole-Liver Perfusion MR ImagingInitial Experience.* Radiology **246** (2008), 926–934.

[5] Holčík, J.: Modelování a simulace biologických systémů. *ČVUT, Praha,* 2006 (in Czech).

ACTA
UNIVERSITATIS PALACKIANAE
OLOMUCENSIS

FACULTAS RERUM NATURALIUM
MATHEMATICA 48 (2009)

Published by the Palacký University Olomouc, Křížkovského 8, 771 47 Olomouc

First edition

ISBN 978-80-244-2386-9
ISSN 0231-9721