# TECHNICAL TRANSACTIONS
# FUNDAMENTAL SCIENCES

# CZASOPISMO TECHNICZNE
# NAUKI PODSTAWOWE


ISSUE 2-NP (16)
YEAR 2014 (111)

ZESZYT 2-NP (16)
ROK 2014 (111)






WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ

# TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE

## FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE

ISSUE 2-NP (16) | ZESZYT 2-NP (16)
YEAR 2014 (111) | ROK 2014 (111)

**Fundamential Sciences Series**
**2-NP/2014**

## Editorial Board

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

LUDWIK BYSZEWSKI AND TERESA WINIARSKA*

# THE EXISTENCE AND UNIQUENESS OF SOLUTIONS OF THE DIRICHLET NONLOCAL PROBLEM WITH A NONLOCAL INITIAL CONDITION

# ISTNIENIE I JEDNOZNACZNOŚĆ ROZWIĄZAŃ NIELOKALNEGO ZAGADNIENIA DIRICHLETA Z NIELOKALNYM WRUNKIEM POCZĄTKOWYM

Abstract

The aim of this paper is to prove the existence and uniqueness of solutions of the Dirichlet nonlocal problem with nonlocal initial condition. The considerations are extensions of results by E. Andreu-Vaillo, J. M. Mazón, J. D. Rossi and J. J. Toledo-Melero [1].

*Keywords:existence and uniqueness of solutions, Dirichlet problem, nonlocal problem, nonlocal initial condition*

Streszczenie

W artykule udowodniono istnienie i jednoznaczność rozwiązań nielokalnego zagadnienia Dirichleta z nielokalnym warunkiem początkowym. Rozważania są rozszerzeniami rezultatów otrzymanych przez E. Andreu-Vaillo, J. M. Mazóna, J. D. Rossi i J. J. Toledo-Melero [1].

*Słowa kluczowe: istnienie i jednoznaczność rozwiązań, zagadnienie Dirichleta, zagadnienie nielokalne, nielokalny warunek początkowy*

*Institute of Mathematics, Cracow University of Technology, Poland; lbyszews@usk.pk.edu.pl, twiniars@usk.pk.edu.pl

## 1. Preliminaries

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Moreover, let $T$ be a fixed positive number and $k \in \mathbb{R} \setminus \{0\}$.

We will need the following assumption:

**Assumption** $(H)$ (see: [1]). $J \in C(\mathbb{R}^n, \mathbb{R})$ *is a nonnegative radial function with* $J(0) > 0$ *and*

$$\int_{\mathbb{R}^n} J(x)dx = 1.$$

In [1], the existence and uniqueness of a solution of the following nonlocal Dirichlet boundary value problem

$$\begin{cases} u_t(x,t) = \int_{\mathbb{R}^n} J(x-y)\big(u(y,t) - u(x,t)\big)dy, \ \ x \in \Omega, \ \ t > 0, \\ u(x,t) = g(x,t), \ \ x \notin \Omega, \ \ t > 0, \\ u(x,0) = u_0(x), \ \ x \in \Omega \end{cases}$$

is studied.

For this purpose the Banach fixed point theorem is applied in [1].

The existence and uniqueness of solutions of differential problems were, also, studied using the Banach fixed point theorem, by Kamont [2], Muszyński and Myszkis [3], and Pelczar and Szarski [4].

The aim of the paper is to give a theorem on the existence and uniqueness of a solution of the following nonlocal Dirichlet boundary value problem together with the nonlocal initial condition

$$\begin{cases} u_t(x,t) = \int_{\mathbb{R}^n} J(x-y)\big(u(y,t) - u(x,t)\big)dy, \ \ x \in \Omega, \ \ t \in (0,T), \\ u(x,t) = g(x,t), \ \ x \notin \Omega, \ \ t \in (0,T), \\ u(x,0) + kTu(x,T) = u_0(x), \ x \in \Omega. \end{cases} \tag{1.1}$$

For this purpose we will also apply the Banach fixed point theorem.

We will need the assumption:

**Assumption** $(F)$. $u_0 \in L^1(\Omega)$ and $g \in C\big((0,T); L^1(\mathbb{R}^n \setminus \Omega)\big)$.

## 2. Existence and uniqueness of Solutions

Let Assumptions $(H)$ and $(F)$ be satisfied in this section.

**Definition 2.1.** *A function* $u \in C\Big([0,T]; L^1(\mathbb{R}^n)\Big)$ *is said to be a solution of nonlocal problem (1.1) if*

$$u(x,t) = u_0(x) - kTu(x,T)$$
$$+ \int_0^t \int_{\mathbb{R}^n} J(x-y)\big(u(y,s) - u(x,s)\big)dyds, \quad x \in \Omega,\ t \in (0,T),$$

*and*

$$u(x,t) = g(x,t) \ \textit{for}\ x \notin \Omega, \ \ t \in (0,T).$$

Consider the Banach space

$$X_T = \big\{w \in C([0,T]; L^1(\Omega))\big\}$$

with the norm

$$|||w||| = \max_{0\le t\le T} \parallel w(\cdot,t) \parallel_{L^1(\Omega)} .$$

The solution of problem (1.1) will be obtained as a fixed point of the operator

$$\mathcal{T}_{w_0} : X_T \longrightarrow X_T$$

defined by the formula

$$\mathcal{T}_{w_0}(w)(x,t) = w_0(x) - kTw(x,T)$$
$$+ \int_0^t \int_{\mathbb{R}_n} J(x-y)\big(w(y,s) - w(x,s)\big)dyds, \ \ x \in \Omega, \ \ t \in (0,T),$$

where

$$w(x,t) = g(x,t) \ \text{ for } \ x \notin \Omega, \ \ t \in (0,T).$$

To prove the existence and uniqueness of the solution of problem (1.1), we will need the following lemma:

**Lemma 2.1.** *Let $w_0, z_0 \in L^1(\Omega)$. Then there is a constant*

$$C = \mid k \mid + \tilde{k}, \ \ \textit{where}\ \ \tilde{k} > 0, \tag{2.2}$$

*depending on $J$ and $\Omega$ such that*

$$|||\mathcal{T}_{w_0}(w) - \mathcal{T}_{z_0}(z)||| \le \parallel w_0 - z_0 \parallel_{L^1(\Omega)} + CT|||w - z|||$$

*for all $w, z \in X_T$.*

*Proof.* Observe that

$$\int_\Omega \mid \mathcal{T}_{w_0}(w)(x,t) - \mathcal{T}_{z_0}(z)(x,t) \mid dx \le$$

$$\leq \int_\Omega \mid w_0 - z_0 \mid (x)dx + \mid k \mid T \int_\Omega \mid w(x,T) - z(x,T) \mid dx$$

$$+ \int_\Omega \mid \int_0^t \int_{\mathbb{R}^n} J(x-y)[(w(y,s) - z(y,s)) - \big(w(x,s) - z(x,s)\big)]dyds \mid dx$$

$$\leq \| w_0 - z_0 \|_{L^1(\Omega)} + \mid k \mid T|||w - z||| + \tilde{k}T|||w - z|||$$

$$= \| w_0 - z_0 \|_{L^1(\Omega)} + (\mid k \mid + \tilde{k})T|||w - z|||, \ w, z \in X_T,$$

where $\tilde{k}$ is a positive constant depending on $J$ and $\Omega$.

Consequently, since $w - z$ vanishes outside of $\Omega$ then

$$|||\mathcal{T}_{w_0}(w) - \mathcal{T}_{z_0}(z)||| \leq$$

$$\leq \| w_0 - z_0 \|_{L^1(\Omega)} + CT|||w - z||| \quad \text{for} \quad w, z \in X_T.$$

The proof of Lemma 2.1 is complete.

Applying Lemma 2.1 we will prove the existence and uniqueness of the solution of problem (1.1).

**Theorem 2.1** *Let Assumptions* $(H)$ *and* $(F)$ *be satisfied. Moreover, let* $CT < 1$, *where* $C$ *is given by (2.2).*

*Then there is a unique solution of problem (1.1) on the interval* $[0, T]$.

*Proof.* Firstly, we will show that $\mathcal{T}_{u_0}$ maps $X_T$ into $X_T$. Let $z_0 \equiv 0$, $z \equiv 0$ and $w_0 \equiv u_0$ in Lemma 2.1. Then

$$\mathcal{T}_{u_0}(w) \in C\big([0,T]; L^1(\Omega)\big)$$

for $w \in X_T$.

Since $CT < 1$ then taking $z_0 \equiv w_0 \equiv u_0$ in Lemma 2.1 we get that $\mathcal{T}_{u_0}$ is a strict contraction in $X_T$ and the existence and uniqueness of the solution of problem (1.1) follows from the Banach fixed point theorem on the interval $[0, T]$.

The proof of Theorem 2.1 is complete.

## References


[1] Andreu-Vaillo F., Mazón J. M., Rossi J. D., Toledo-Melero J. J., *Nonlocal Diffusion Problems,* American Mathematical Society, Providence, Rhode Island 2010.

[2] Kamont Z., *Ordinary Differential Equations,* Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 1999 [in Polish].

[3] Muszyński J., Myszkis A. D., *Ordinary Differential Equations,* Państwowe Wydawnictwo Naukowe, Warszawa 1984 [in Polish].

[4] Pelczar A., Szarski J., *Introduction to the Theory of Differential Equations,* Państwowe Wydawnictwo Naukowe, Warszawa 1987 [in Polish].

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

LUDWIK BYSZEWSKI AND TERESA WINIARSKA*

# INTEGRO-DIFFERENTIAL EVOLUTION NONLOCAL PROBLEM FOR THE FIRST ORDER EQUATION (II)

# CAŁKOWO-RÓŻNICZKOWE EWOLUCYJNE ZAGADNIENIE NIELOKALNE DLA RÓWNANIA PIERWSZEGO RZĘDU (II)

Abstract

The aim of this paper is to give two theorems on the existence and uniqueness of mild and classical solutions of a nonlocal semilinear integro-differential evolution Cauchy problem for the first order equation. The method of semigroups, the Banach fixed-point theorem and the Bochenek theorem are applied to prove the existence and uniqueness of the solutions of the considered problem.

*Keywords: nonlocal problem, integro-differential evolution problem, abstract Cauchy problem*

Streszczenie

W artykule udowodniono dwa twierdzenia o istnieniu i jednoznaczności rozwiązań całkowych i klasycznych nielokalnego semiliniowego całkowo-różniczkowego ewolucyjnego zagadnienia Cauchy'ego dla równania rzędu pierwszego. W tym celu zastosowano metodę półgrup, twierdzenie Banacha o punkcie stałym i twierdzenie Bochenka.

*Słowa kluczowe: zagadnienie nielokalne, ewolucyjne zagadnienie całkowo-różniczkowe, abstrakcyjne zagadnienie Cauchy'ego*

*Institute of Mathematics, Cracow University of Technology, Poland; lbyszews@usk.pk.edu.pl, twiniars@usk.pk.edu.pl

## 1. Introduction

In this paper, we give two theorems on the existence and uniqueness of mild and classical solutions of semilinear integro-differential evolution nonlocal Cauchy problem for the first order equation. To achieve this, the method of semigroups, the Banach fixed point theorem and the Bochenek theorem will be used.

Let $E$ be a real Banach space with norm $\|\cdot\|$ and let $A : E \to E$ be a closed densely defined linear operator. For the operator $A$, let $\mathcal{D}(A)$, $\rho(A)$ and $A^*$ denote its domain, resolvent set and adjoint, respectively.

For the Banach space $E$, $\mathcal{C}(E)$ denotes the set of closed linear operators from $E$ into itself.

We will need the class $G(\tilde{M}, \beta)$ of operators $A$ satisfying the conditions:

There exist constants $\tilde{M} > 0$ and $\beta \in \mathbb{R}$ such that

$(C_1)$ $A \in \mathcal{C}(E)$, $\overline{\mathcal{D}(A)} = E$ and $(\beta, +\infty) \subset \rho(-A)$,

$(C_2)$ $\left\|(A+\xi)^{-k}\right\| \leqslant \tilde{M}(\xi-\beta)^{-k}$ for each $\xi > \beta$ and $k = 1, 2, \ldots$

It is known (see [4], p. 485 and [5], p. 20) that for $A \in G(\tilde{M}, \beta)$, there exists exactly one strongly continuous semigroup $T(t) : E \to E$ for $t \geqslant 0$ such that $-A$ is its infinitesimal generator and

$$\|T(t)\| \leqslant \tilde{M} e^{\beta t} \quad \text{for } t \geqslant 0.$$

Throughout this paper, we shall use the notation:

$$\begin{aligned} \mathcal{J} &:= [t_0, t_0 + a], \quad \text{where } t_0 \geqslant 0 \text{ and } a > 0, \\ \Delta &:= \{(t, s) : t_0 \leqslant s \leqslant t \leqslant t_0 + a\}, \\ M &:= \sup\{\|T(t)\|, t \in [0, a]\} \end{aligned}$$

and

$$X := \mathcal{C}(\mathcal{J}, E).$$

The Cauchy problem considered here is of the form:

$$\begin{aligned} u'(t) + Au(t) &= f(t, u(t), u(b(t))) + \int_{t_0}^{t} f_1(t, s, u(s))ds + \\ &+ \int_{t_0}^{t_0+a} f_2(t, s, u(s))ds, \ t \in (t_0, t_0 + a], \qquad (1) \\ u(t_0) + g(u) &= u_0, \qquad (2) \end{aligned}$$

where $f$, $f_i$ $(i = 1, 2)$, $g$ and $b$ are given functions satisfying some assumptions and $u_0 \in E$.

The results obtained in the paper are a continuation of those given in [3] and they are based on those from [1] – [6].

## 2. The Bochenek theorem

The results of this section were obtained by J. Bochenek (see [2]).

Let us consider the Cauchy problem

$$u'(t) + Au(t) = k(t),\ t \in \mathcal{J} \setminus \{t_0\}, \tag{3}$$

$$u(t_0) = x. \tag{4}$$

A function $u : \mathcal{J} \to E$ is said to be a classical solution of problem (3)–(4) if

(i) $u$ is continuous and continuously differentiable on $\mathcal{J} \setminus \{t_0\}$,

(ii) $u'(t) + Au(t) = k(t)$ for $t \in \mathcal{J} \setminus \{t_0\}$,

(iii) $u(t_0) = x$.

**Assumption** (Z). The adjoint operator $A^*$ is densely defined in $E^*$, i.e. $\overline{\mathcal{D}(A^*)} = E^*$.

**Theorem 2.1.** *Let conditions $(C_1)$, $(C_2)$ and Assumption (Z) be satisfied. Moreover, let $k : \mathcal{J} \to E$ be Lipshitz continuous on $\mathcal{J}$ and $x \in \mathcal{D}(A)$.*

*Then $u$ given by the formula*

$$u(t) = T(t - t_0)x + \int_{t_0}^{t} T(t - s)k(s)ds,\ t \in \mathcal{J} \tag{5}$$

*is the unique classical solution of the Cauchy problem* (3)–(4).

## 3. Theorem about a mild solution

A function $u : \mathcal{J} \to E$ satisfying the integral equation

$$\begin{aligned} u(t) &= T(t - t_0)u_0 - T(t - t_0)g(u) + \int_{t_0}^{t} T(t - s)\Big(f(s, u(s)), u(b(s))) + \\ &+ \int_{t_0}^{s} f_1(s, \tau, u(\tau))d\tau + \int_{t_0}^{t_0+a} f_2(s, \tau, u(\tau))d\tau\Big)ds,\ t \in \mathcal{J} \end{aligned}$$

is said to be a mild solution of the integrodifferential evolution nonlocal Cauchy problem (1)–(2).

Arguing analogously as in [3] we can obtain, by the Banach fixed point theorem, the following theorem:

**Theorem 3.1.** *Assume that:*

(i) *the operator $A : E \to E$ satisfies conditions $(C_1)$ and $(C_2)$,*

(ii) $f : \mathcal{J} \times E^2 \to E$ *is continuous with respect to the first variable in* $\mathcal{J}$, $f_i : \Delta \times E \to E$ $(i = 1, 2)$ *are continuous with respect to the variables in* $\Delta$, $g : X \to E$, $b : \mathcal{J} \to \mathcal{J}$ *are continuous and there exist positive constants* $L$, $L_i$ $(i = 1, 2)$ *and* $K$ *such that*

$$\|f(s, z_1, z_2) - f(s, \tilde{z}_1, \tilde{z}_2)\| \leqslant L \sum_{i=1}^{2} \|z_i - \tilde{z}_i\|$$

*for* $s \in \mathcal{J}$, $z_i, \tilde{z}_i \in E$ $(i = 1, 2)$,

$$\|f_i(s, \tau, z) - f_i(s, \tau, \tilde{z})\| \leqslant L_i \|z - \tilde{z}\| \ \ (i = 1, 2)$$

*for* $(s, \tau) \in \Delta$, $z, \tilde{z} \in E$
*and*

$$\|g(w) - g(\tilde{w})\| \leqslant K \|w - \tilde{w}\|_X \quad \text{for } w, \tilde{w} \in X.$$

(iii) $M[a(2L + aL_1 + aL_2) + K] < 1$.

(iv) $u_0 \in E$.

*Then the integrodifferential evolution nonlocal Cauchy problem* (1)–(2) *has a unique mild solution.*

## 4. Theorem about a classical solution

A function $u : \mathcal{J} \to E$ is said to be a classical solution of the nonlocal Cauchy problem (1)–(2) on $\mathcal{J}$ if :

(*i*) $u$ is continuous on $\mathcal{J}$ and continuously differentiable on $\mathcal{J} \setminus \{t_0\}$,

(*ii*) $u'(t) + Au(t) = f\big(t, u(t), u(b(t))\big) + \int_{t_0}^{t} f_1(t, s, u(s))ds +$
$+ \int_{t_0}^{t_0+a} f_2(t, s, u(s))ds \quad$ for $t \in \mathcal{J} \setminus \{t_0\}$,

(*iii*) $u(t_0) + g(u) = u_0$.

**Theorem 4.1.** *Assume that:*

(i) *the operator* $A : E \to E$ *satisfies conditions* $(C_1)$ *and* $(C_2)$, *and Assumption* (Z).

(ii) $f : \mathcal{J} \times E^2 \to E$, $g : X \to E$, *for any* $(s, z) \in \mathcal{J} \times E$ *and* $i = 1, 2$ *functions* $f_i(s, \cdot, z) : \mathcal{J} \ni \tau \mapsto f(s, \tau, z) \in E$ *are continuous,* $b : \mathcal{J} \to \mathcal{J}$ *is continuous on* $\mathcal{J}$ *and there exist positive constants* $C$, $C_i$ $(i = 1, 2)$ *and* $K$ *such that:*

$$\|f(s, z_1, z_2) - f(\tilde{s}, \tilde{z}_1, \tilde{z}_2)\| \leqslant C\Big( |s - \tilde{s}| + \sum_{i=1}^{2} \|z_i - \tilde{z}_i\| \Big)$$

*for* $s, \tilde{s} \in \mathcal{J}$, $z_i, \tilde{z}_i \in E$ $(i = 1, 2)$,

$$\|f_i(s, \tau, z) - f_i(\tilde{s}, \tau, \tilde{z})\| \leqslant C_i(|s - \tilde{s}| + \|z - \tilde{z}\|)$$

*for* $(s, \tau)$, $(\tilde{s}, \tau) \in \Delta$, $z, \tilde{z} \in E$
*and*

$$\|g(w) - g(\tilde{w})\| \leqslant K \|w - \tilde{w}\|_X \quad \text{for } w, \tilde{w} \in X.$$

(iii) $$M\Big(a(2C + aC_1 + aC_2) + K\Big) < 1.$$

*Then the integrodifferential evolution nonlocal Cauchy problem* (1)–(2) *has a unique mild solution (which is denoted by)* $u$. *Moreover, if* $u_0 \in \mathcal{D}(A)$, $g(u) \in \mathcal{D}(A)$ *and if there exists a positive constant* $\mathcal{H}$ *such that*

$$\|u(b(s)) - u(b(\tilde{s}))\| \leqslant \mathcal{H} \|u(s) - u(\tilde{s})\| \quad \text{for } s, \tilde{s} \in \mathcal{J}$$

*then* $u$ *is the unique classical solution of the problem* (1)–(2).

*Proof.* Since all the assumptions of Theorem 3.1 are satisfied, it is easy to see that problem (1)–(2) possesses a unique mild solution which according to the last assumption is denoted by $u$.

Now we shall show that $u$ is the classical solution of the problem (1)–(2). To this end, observe that as in [3] $u$ is Lipschitz continuous on $\mathcal{J}$.

The Lipschitz continuity of $u$ on $\mathcal{J}$ combined with the Lipschitz continuity of $f$ on $\mathcal{J} \times E^2$ and $f_i$ $(i = 1, 2)$ with respect to the first variables imply that the function

$$\mathcal{J} \ni t \mapsto f\big(t, u(t), u(b(t))\big) + \int_{t_0}^{t} f_1(t, s, u(s))ds + \int_{t_0}^{t_0+a} f_2(t, s, u(s))ds$$

is Lipschitz continuous on $\mathcal{J}$. This property of $f$ together with the assumptions of Theorem 4.1 imply, by Theorem 2.1 and Theorem 3.1, that the linear Cauchy problem:

$$\begin{aligned} v'(t) + Av(t) &= f\big(t, u(t), u(b(t))\big) + \int_{t_0}^{t} f_1(t, s, u(s))ds + \\ &+ \int_{t_0}^{t_0+a} f_2(t, s, u(s))ds, \ t \in \mathcal{J} \setminus \{t_0\}, \\ v(t_0) &= u_0 - g(u) \end{aligned}$$

has a unique classical solution $v$ and it is given by

$$\begin{aligned} v(t) &= T(t - t_0)u_0 - T(t - t_0)g(u) + \int_{t_0}^{t} T(t - s)\Big(f\big(s, u(s), u(b(s))\big) + \\ &+ \int_{t_0}^{s} f_1(s, \tau, u(\tau))d\tau + \int_{t_0}^{t_0+a} f_2(s, \tau, u(\tau))d\tau\Big)ds = u(t), \ t \in \mathcal{J}. \end{aligned}$$

Consequently, u is the unique classical solution of the integrodifferential evolution Cauchy problem (1)–(2) and, therefore, the proof of Theorem 4.1 is complete.

□

## References


[1] Balachandran K. , Ilamaran S., *Existence and uniqueness of mild and strong solutions of a semilinear evolution equation with nonlocal conditions*, Indian J. Pure Appl. Math., **25.4**, 1994, 411—418.

[2] Bochenek J., *The existence of a solution of a semilinear first–order differential equation in a Banach space*, *Univ. Iag. Acta Math.*, **31**, 1994, 61—68.

[3] Byszewski L., Winiarska T., *Integrodifferential evolution nonlocal problem for the first order equations*, Technical Transactions, Fundamental Sciences, 1 – NP/2011, 15—21.

[4] Kato T., *Perturbation Theory for Linear Operators*, *Springer - Verlag, New York, Berlin, Heidelberg* 1966.

[5] Pazy A., *Semigroups of Linear Operators and Applications to Partial Differential Equations*, *Springer–Verlag, New York, Berlin, Heidelberg, Tokyo* 1983.

[6] Tidke H. L., *Existence of solutions of nonlinear mixed integrodifferential equations of Sobolev type*, *Nonlinear Functional Analysis and Applications*, **14.4**, 2009, 605—618.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

LUDWIK BYSZEWSKI AND TERESA WINIARSKA*

# ON NONLOCAL EVOLUTION FUNCTIONAL-DIFFERENTIAL PROBLEM IN A BANACH SPACE

# NIELOKALNE EWOLUCYJNE FUNKCJONALNO-RÓŻNICZKOWE ZAGADNIENIE W PRZESTRZENI BANACHA

Abstract

The aim of this paper is to prove two theorems on the existence and uniqueness of mild and classical solutions of a nonlocal semilinear functional-differential evolution Cauchy problem in a Banach space. The method of semigroups, the Banach fixed-point theorem and the Bochenek theorem (see [3]) about the existence and uniqueness of the classical solution of the first order differential evolution problem in a not necessarily reflexive Banach space are used to prove the existence and uniqueness of the solutions of the considered problem. The results are based on publications [1 — 8].
*Keywords: evolution problem, functional-differential problem, nonlocal problem*

Streszczenie

W artykule udowodniono dwa twierdzenia o istnieniu i jednoznaczności rozwiązań całkowych i klasycznych nielokalnego semiliniowego funkcjonalno-różniczkowego ewolucyjnego zagadnienia Cauchy'ego w dowolnej przestrzeni Banacha. W tym celu zastosowano metodę półgrup, twierdzenie Banacha o punkcie stałym i twierdzenie Bochenka [3] o istnieniu i jednoznaczności klasycznego rozwiązania ewolucyjnego zagadnienia różniczkowego pierwszego rzędu w niekoniecznie refleksywnej przestrzeni Banacha. Artykuł bazuje na publikacjach [1 — 8].
*Słowa kluczowe: zagadnienie ewolucyjne, zagadnienie funkcjonalno-różniczkowe, zagadnienie nielokalne*

*Institute of Mathematics, Cracow University of Technology, Poland; lbyszews@usk.pk.edu.pl, twiniars@usk.pk.edu.pl

## 1. Preliminaries

In this paper, we prove two theorems on the existence and uniqueness of mild and classical solutions of a semilinear functional-differential evolution nonlocal Cauchy problem using the method of semigroups, the Banach fixed-point theorem and the Bochenek theorem (see [3]) about the existence and uniqueness of the classical solution of the linear first-order differential evolution problem in a not necessarily reflexive Banach space.

Let $E$ be a real Banach space with norm $\|\cdot\|$ and let $A : E \to E$ be a closed densely defined linear operator. For an operator $A$, let $\mathcal{D}(A)$, $\rho(A)$ and $A^*$ denote its domain, resolvent set and adjoint, respectively.

For Banach space $E$, $\mathcal{C}(E)$ denote the set of closed linear operators from $E$ into itself.

We will need the class $G(\tilde{M}, \beta)$ of operators $A$ satisfying the conditions:

There exist constants $\tilde{M} > 0$ and $\beta \in \mathbb{R}$ such that

$(C_1)$ $A \in \mathcal{C}(E)$, $\overline{\mathcal{D}(A)} = E$ and $(\beta, +\infty) \subset \rho(-A)$,

$(C_2)$ $\left\|(A+\xi)^{-k}\right\| \leqslant \tilde{M}(\xi - \beta)^{-k}$ for each $\xi > \beta$ and $k = 1, 2, \ldots$

We will use the assumption:

**Assumption** (Z). The adjoint operator $A^*$ is densely defined in $E^*$, i.e. $\overline{\mathcal{D}(A^*)} = E^*$.

It is known (see [5], p. 485 and [7], p. 20) that for $A \in G(\tilde{M}, \beta)$ there exists exactly one strongly continuous semigroup $T(t) : E \to E$ for $t \geqslant 0$ such that $-A$ is its infinitesimal generator and

$$\|T(t)\| \leqslant \tilde{M} e^{\beta t} \quad \text{for } t \geqslant 0.$$

Throughout this paper, we assume $(C_1)$, $(C_2)$ and assumption (Z).

In this paper, we assume that $t_0 > 0$, $a > 0$,

$$\begin{aligned} \mathcal{J} := [t_0, t_0 + a],\ \Delta &:= \{(t,s) : t_0 \leqslant s \leqslant t \leqslant t_0 + a\}, \\ M &:= \sup_{t \in [0,a]} \|T(t)\|, \\ X &:= \mathcal{C}(\mathcal{J}, E) \end{aligned} \tag{1.1}$$

and

$$\begin{aligned} F_1 &: \mathcal{J} \times E^{m+1} \to E,\ F_2 : \Delta \times E^2 \to E,\ \tilde{G} : X \to E, \\ f &: \Delta \times E \to E,\ \sigma_i : \mathcal{J} \to \mathcal{J}\ (i = 1, \ldots, m) \end{aligned}$$

are given functions satisfying some assumptions.

The functional-differential evolution nonlocal Cauchy problem considered here is of the form

$$\begin{aligned} u'(t) &+ Au(t) = F_1(t, u(t), u(\tilde{\sigma}_1(t)), \dots, u(\tilde{\sigma}_m(t))) + \\ &+ \int_{t_0}^{t} F_2(t, s, u(s), \int_{t_0}^{s} f(s, \tau, u(\tau))d\tau)ds,\ t \in \mathcal{J} \setminus \{t_0\}, \qquad (1.2) \\ & u(t_0) + \tilde{G}(u) = u_0, \qquad (1.3) \end{aligned}$$

where $u_0 \in E$.

To study problem (1.2)–(1.3) we will need some information related to the following linear problem:

$$\begin{aligned} u'(t) + Au(t) &= k(t), \quad t \in \mathcal{J} \setminus \{t_0\}, \qquad (1.4) \\ u(t_0) &= x \qquad (1.5) \end{aligned}$$

and the following definition:

A function $u : \mathcal{J} \to E$ is said to be a classical solution of problem (1.4)–(1.5) if

(i) $u$ is continuous and continuously differentiable on $\mathcal{J} \setminus \{t_0\}$,

(ii) $u'(t) + Au(t) = k(t)$ for $t \in \mathcal{J} \setminus \{t_0\}$,

(iii) $u(t_0) = x$.

To study problem (1.2)–(1.3) we will also need the following theorem:

**Theorem 1.1** (see [3]). *Let $k : \mathcal{J} \to E$ be Lipschitz continuous on $\mathcal{J}$ and $x \in \mathcal{D}(A)$.*

*Then $u$ given by the formula*

$$u(t) = T(t - t_0)x + \int_{t_0}^{t} T(t - s)k(s)ds, \quad t \in \mathcal{J} \qquad (1.6)$$

*is the unique classical solution of the Cauchy problem* (1.4)–(1.5).

## 2. On mild solution

A function $u : \mathcal{J} \to X$ satisfying the integral equation

$$\begin{aligned} u(t) &= T(t - t_0)u_0 - T(t - t_0)\tilde{G}(u) + \\ &+ \int_{t_0}^{t} T(t - s)F_1(s, u(s), u(\sigma_1(s)), \dots, u(\sigma_m(s)))ds + \\ &+ \int_{t_0}^{t} T(t - s)\Big(\int_{t_0}^{s} F_2(s, \tau, u(\tau), \int_{t_0}^{\tau} f(\tau, \mu, u(\mu))d\mu)d\tau\Big)ds,\ t \in \mathcal{J}, \end{aligned}$$

is said to be a mild solution of the nonlocal Cauchy problem (1.2)–(1.3).

**Theorem 2.1.** *Assume that*

(*i*) *for all* $z_i \in E$ $(i = 0, 1, \dots, m)$, *the function*

$$\mathcal{J} \ni t \mapsto F_1(t, z_0, z_1, \dots, z_m) \in E \quad \textit{is continuous},$$

*for all* $z_i \in E$ $(i = 1, 2)$, *the function*

$$\Delta \ni (t, s) \mapsto F_2(t, s, z_1, z_2) \in E \quad \textit{is continuous},$$

*for all* $z \in E$, *the function*

$$\Delta \ni (t, s) \mapsto f(t, s, z) \quad \textit{is continuous},$$

$$\tilde{G} : X \to E,\ \sigma_i \in \mathcal{C}(\mathcal{J}, \mathcal{J})\ (i = 1, \dots, m) \textit{ and } u_0 \in E.$$

(*ii*) *there are constants* $L_i > 0$ $(i = 1, 2, 3, 4)$ *such that*

$$\|F_1(t, z_0, z_1, \dots, z_m) - F_1(t, \tilde{z}_0, \tilde{z}_1, \dots, \tilde{z}_m)\| \leqslant$$
$$\leqslant L_1 \sum_{i=0}^{m} \|z_i - \tilde{z}_i\| \quad \textit{for } t \in \mathcal{J},\ z_i,\ \tilde{z}_i \in E\ (i = 1, \dots, m); \tag{2.1}$$

$$\|F_2(t, s, z_1, z_2) - F_2(t, s, \tilde{z}_1, \tilde{z}_2)\| \leqslant L_2 \sum_{i=1}^{2} \|z_i - \tilde{z}_i\|$$
$$\textit{for } (t, s) \in \Delta,\ z_i, \tilde{z}_i \in E,\ (i = 1, 2); \tag{2.2}$$

$$\|f(t, s, z) - f(t, s, \tilde{z})\| \leqslant L_3 \|z - \tilde{z}\|$$
$$\textit{for } (t, s) \in \Delta,\ z, \tilde{z} \in E; \tag{2.3}$$

$$\left\|\tilde{G}(w) - \tilde{G}(\tilde{w})\right\| \leqslant L_4 \|w - \tilde{w}\| \quad \textit{for } w, \tilde{w} \in X; \tag{2.4}$$

(*iii*) $M[L_1 a(m + 1) + L_2 a^2(1 + L_3 a) + L_4] < 1$.

*Then the nonlocal problem* (1.2)–(1.3) *has a unique mild solution in* $\mathcal{J}$.

*Proof.* Introduce an operator $\mathfrak{F}$ on $X$ by the formula

$$\begin{aligned}(\mathfrak{F}w)(t) \;&:=\; T(t - t_0)u_0 - T(t - t_0)\tilde{G}(w) + \\ &+\; \int_{t_0}^{t} T(t - s)F_1(s, w(s), w(\sigma_1(s)), \dots, w(\sigma_m(s)))ds + \\ &+\; \int_{t_o}^{t} T(t - s)\Big(\int_{t_0}^{s} F_2(s, \tau, w(\tau), \int_{t_0}^{\tau} f(\tau, \mu, w(\mu))d\mu)d\tau\Big)ds\end{aligned} \tag{2.5}$$

for $w \in X$ and $t \in \mathcal{J}$.

It is easy to see that

$$\mathfrak{F} : X \to X. \tag{2.6}$$

Now, we will show that $\mathfrak{F}$ is a contraction on $X$. For this purpose, observe that from (2.5), (1.1) and (2.1)–(2.4),

$$\begin{aligned}
&\|(\mathfrak{F}w)(t) - (\mathfrak{F}\tilde{w})(t)\| \leqslant ML_4 \|w - \tilde{w}\| + \qquad (2.7)\\
+ \quad & ML_1 \int_{t_0}^{t} \Big( \|w(s) - w(\tilde{s})\| + \sum_{i=1}^{m} \|w(\sigma_i(s)) - \tilde{w}(\sigma_i(s))\| \Big) ds + \\
+ \quad & ML_2 \int_0^t \Big( \int_0^s \big( \|w(\tau) - \tilde{w}(\tau)\| + \\
+ \quad & \int_{t_0}^{\tau} \|f(\tau, \mu, w(\mu)) - f(\tau, \mu, \tilde{w}(\mu))\| \, d\mu \big) d\tau \Big) ds \leqslant \\
\leqslant \quad & ML_4 \|w - \tilde{w}\| + ML_1 a(m+1) \|w - \tilde{w}\| + \\
+ \quad & ML_2 \int_0^t \Big( \int_{t_0}^s [\|w(\tau) - \tilde{w}(\tau)\| + L_3 \int_{t_0}^{\tau} \|w(\mu) - \tilde{w}(\mu)\| \, d\mu] d\tau \Big) ds \leqslant \\
\leqslant \quad & q \|w - \tilde{w}\| \quad \text{for } w, \tilde{w} \in X,
\end{aligned}$$

where

$$q := M\big(L_1 a(m+1) + L_2 a^2 (1 + L_3 a) + L_4\big).$$

Then, by (2.7) and by assumption $(iii)$,

$$\|\mathfrak{F}w - \mathfrak{F}\tilde{w}\| \leqslant q \|w - \tilde{w}\| \quad \text{for } w, \tilde{w} \in X \text{ with } 0 < q < 1. \tag{2.8}$$

Consequently, from (2.6) and (2.8), operator $\mathfrak{F}$ satisfies all the assumptions of the Banach contraction theorem. Therefore, in space $X$ there is only one fixed point of $\mathfrak{F}$ and this point is the mild solution of the nonlocal Cauchy problem (1.2)–(1.3). So, the proof of Theorem 2.1 is complete. □

## 3. On classical solution

A function $u : \mathcal{J} \to E$ is said to be a classical solution of the nonlocal Cauchy problem (1.2)–(1.3) on $\mathcal{J}$ if :

$(i)$ $u$ is continuous on $\mathcal{J}$ and continuously differentiable on $\mathcal{J} \setminus \{t_0\}$,

$(ii)$ $u'(t) + Au(t) = F_1\big(t, u(t), u(\sigma_1(t)), \ldots, u(\sigma_m(t))\big)) +$
$+ \int_{t_0}^{t} F_2(t, s, u(s), \int_{t_0}^{s} f(s, \tau, u(\tau)) d\tau) ds, \ t \in \mathcal{J} \setminus \{t_0\}$,

$(iii)$ $u(t_0) + \tilde{G}(u) = u_0$.

**Theorem 3.1.** *Suppose that assumptions (i)–(iii) of Theorem 2.1 are satisfied. Then the nonlocal Cauchy problem* (1.2)–(1.3) *has a unique mild solution on* $\mathcal{J}$*, denoted by* $u$*. Assume, additionally, that:*

(*i*) $u_0 \in \mathcal{D}(A)$ *and* $\tilde{G}(u) \in \mathcal{D}(A)$*;*

(*ii*) *there are constants* $C_i > 0$ $(i = 1, 2)$ *such that*

$$\begin{aligned} \big\|F_1(t, z_0, z_1, \dots, z_m) - F_1(\tilde{t}, z_0, z_1, \dots, z_m)\big\| &\leqslant C_1 \left|t - \tilde{t}\right| \\ \text{for } t, \tilde{t} \in \mathcal{J},\ z_i \in E\ (i &= 0, 1, \dots, m) \end{aligned} \tag{3.1}$$

*and*

$$\begin{aligned} \big\|F_2(t, s, z_1, z_2) - F_2(\tilde{t}, s, z_1, z_2)\big\| &\leqslant C_2 \left|t - \tilde{t}\right| \\ \text{for } (t, s) \in \Delta,\ (\tilde{t}, s) \in \Delta,\ z_i \in E\ (i &= 1, 2); \end{aligned} \tag{3.2}$$

(*iii*) *there is a constant* $c > 0$ *such that*

$$\begin{aligned} \big\|u(\sigma_i(t)) - u(\sigma_i(\tilde{t}))\big\| &\leqslant c \big\|u(t) - u(\tilde{t})\big\| \\ \text{for } t, \tilde{t} \in \mathcal{J}\ (i &= 0, 1, \dots, m). \end{aligned} \tag{3.3}$$

*Then* $u$ *is the unique classical solution of the nonlocal Cauchy problem* (1.2)–(1.3) *on* $\mathcal{J}$*.*

*Proof.* Since all the assumptions of Theorem 2.1 are satisfied, the nonlocal Cauchy problem (1.2)–(1.3) possesses a unique mild solution which, according to the assumption, is denoted by $u$.

Now we will show that $u$ is the unique classical solution of the problem (1.2)–(1.3) on $\mathcal{J}$. To this end, introduce

$$N_1 := \max_{s \in \mathcal{J}} \big\|F_1\big(s, u(s), u(\sigma_1(s)), \dots, u(\sigma_m(s))\big)\big\| \tag{3.4}$$

and

$$N_2 := \max_{(\xi, \eta) \in \Delta} \left\|F_2(\xi, \eta, u(\eta), \int_{t_0}^{\eta} f(\eta, \mu, u(\mu)) d\mu\right\|, \tag{3.5}$$

and observe that

$$
\begin{aligned}
& u(t+h)-u(t)= \qquad (3.6)\\
= \;& T(t-t_0)\big(T(h)-I\big)u_o - T(t-t_0)\big(T(h)-I\big)\tilde{G}(u)+\\
+ \;& \int_{t_0}^{t_0+h} T(t+h-s)F_1\big(s,u(s),u(\sigma_1(s)),\dots,u(\sigma_m(s))\big)ds+\\
+ \;& \int_{t_0}^{t} T(t-s)\Big(F_1\big(s+h,u(s+h),u(\sigma_1(s+h)),\dots,u(\sigma_m(s+h))\big)-\\
& \qquad - F_1\big(s,u(s),u(\sigma_1(s)),\dots,u(\sigma_m(s))\big)\Big)ds+\\
+ \;& \int_{t_0}^{t_0+h} T(t+h-s)\Big(\int_{t_0}^{s} F_2\big(s,\tau,u(\tau),\int_{t_0}^{\tau} f(\tau,\mu,u(\mu)d\mu\big)d\tau\Big)ds+\\
+ \;& \int_{t_0}^{t} T(t-s)\Big(\int_{t_0}^{s}\big(F_2\big(s+h,\tau,u(\tau),\int_{t_0}^{\tau} f(\tau,\mu,u(\mu))d\mu\big)-\\
& \qquad - F_2\big(s,\tau,u(\tau),\int_{t_0}^{\tau} f(\tau,\mu,u(\mu))d\mu\big)d\tau\Big)ds+\\
+ \;& \int_{t_0}^{t} T(t-s)\Big(\int_{s}^{s+h} F_2\big(s+h,\tau,u(\tau),\int_{t_0}^{\tau} f(\tau,\mu,u(\mu))d\mu\big)d\tau\Big)ds
\end{aligned}
$$

for $t\in[t_0,t_0+a)$, $h>0$ and $t+h\in(t_0,t_0+a]$.

Consequently by (3.6), (1.1) and (3.1)–(3.5),

$$
\begin{aligned}
& \|u(t+h)-u(t)\|\leqslant hM\,\|Au_0\|+hM\,\Big\|A\tilde{G}(u)\Big\|+ \qquad (3.7)\\
+ \;& hMN_1+ahML_1+ML_1\int_{t_0}^{t}\Big(\,\|u(s+h)-u(s)\|+\\
+ \;& \sum_{i=1}^{m}\|u(\sigma_i(s+h))-u(\sigma_i(s))\|\,\Big)ds+a^2ML_2h+2aMN_2h\leqslant\\
\leqslant \;& Ch+ML_1(1+mc)\int_{t_0}^{t}\|u(s+h)-u(s)\|\,ds
\end{aligned}
$$

for $t\in[t_0,t_0+a)$, $h>0$ and $t+h\in(t_0,t_0+h]$, where

$$C:=M\Big(\,\|Au_0\|+\Big\|A\tilde{G}(u)\Big\|+N_1+aL_1+a^2L_2+2aN_2\Big).$$

From (3.7) and Gronwall's inequality,

$$\|u(t+h)-u(t)\|\leqslant Ce^{aML_1(1+mc)}h$$

for $t\in[t_0,t_0+h]$, $h>0$ and $t+h\in(t_0,t_0+a]$.

Hence $u$ is Lipschitz continuous on $\mathcal{J}$.

The Lipschitz continuity of $u$ on $\mathcal{J}$ and inequalities (3.1), (2.1), (3.2) imply that the function

$$\begin{aligned}\mathcal{J} \ni t \mapsto k(t) &:= F_1\big(t, u(t), u(\sigma_1(t)), \dots, \sigma_m(t))\big) + \\ &+ \int_{t_0}^{t} F_2\big(t, s, u(s), \int_{t_0}^{s} f(s, \tau, u(\tau))d\tau\big)ds \in E\end{aligned}$$

is Lipschitz continuous on $\mathcal{J}$. This property of $t \mapsto k(t)$ together with assumptions of Theorem 3.1 imply, by Theorem 1.1, by Theorem 2.1 and by the definition of the mild solution from Section 2, that the linear Cauchy problem

$$\begin{aligned}v'(t) + Av(t) &= k(t), \quad t \in \mathcal{J} \setminus \{t_0\}, \\ v(t_0) &= u_0 - \tilde{G}(u)\end{aligned}$$

has a unique classical solution $v$ such that

$$\begin{aligned}v(t) &= T(t-t_0)u_0 - T(t-t_0)\tilde{G}(u) + \int_{t_0}^{t} T(t-s)k(s)ds = \\ &= T(t-t_0)u_0 - T(t-t_0)\tilde{G}(u) + \\ &+ \int_{t_0}^{t} T(t-s)F_1\big(s, u(s), u(\sigma_1(s)), \dots, u(\sigma_m(s))\big)ds + \\ &+ \int_{t_0}^{t} T(t-s)\Big(\int_{t_0}^{s} F_2\big(s, \tau, u(\tau), \int_{t_0}^{\tau} f(\tau, \mu, u(\mu))d\mu\big)d\tau\Big)ds = \\ &= u(t), \quad t \in \mathcal{J}.\end{aligned}$$

Consequently, $u$ is the unique classical solution of the nonlocal Cauchy problem (1.2)–(1.3) on$\mathcal{J}$. Therefore, the proof of Theorem 3.1 is complete.

□

## References


[1] Balachandran K., Ilamaran S., *Existence and uniqueness of mild and strong solutions of a semilinear evolution equation with nonlocal conditions*, *Indian J. Pure Appl. Math.*, **25.4**, 1994, 411—418.

[2] Balasubramaniam, P. Chandrasekaran, M. *Existence of solutions of nonlinear integrodifferential equation with nonlocal boundary conditions in Banach space*, *Atti Sem. Mat. Fis. Univ. Modena*, **46**, 1998, 1—13.

[3] Bochenek J., *The existence of a solution of a semilinear first–order differential equation in a Banach space*, *Univ. Iag. Acta Math.*, **31** 1994, 61—68.

[4] Byszewski L., *Theorems about the existence and uniqueness of solutions of a semilinear evolution nonlocal Cauchy problem*, *J. Math. Anal. Appl.*, **162.2** 1991, 494—505.

[5] Kato T., *Perturbation Theory for Linear Operators*, *Springer-Verlag, New York, Berlin, Heidelberg* 1966.

[6] Kołodziej K., *Existence and uniqueness of solutions of a semilinear functional-differential evolution nonlocal Cauchy problem*, *JAMSA*, **13.2** 2000, 171–179.

[7] Pazy A., *Semigroups of Linear Operators and Applications to Partial Differential Equations*, *Springer–Verlag, New York, Berlin, Heidelberg, Tokyo*, 1983.

[8] Winiarska T., *Differential Equations with Parameters*, *Monograph 68, Cracow University of Technology* 1988.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

CHRISTIANA DRAKE*, OSKAR KNAPIK**, JACEK LEŚKOW***

# MISSING DATA ANALYSIS IN CYCLOSTATIONARY MODELS

# ANALIZA BRAKUJĄCYCH OBSERWACJI W MODELU CYKLOSTACJONARNYM

Abstract

In recent years, there has been a growing interest in modeling cyclostationary time series. The survey of Gardner and others [5] is quoting over 1500 different recently published papers that are dedicated to this topic. Data that can be reasonable modeled with such time series is often incomplete. To our knowledge, no systematic research has been conducted on that problem. This paper attempts to fill this gap. In this paper we propose to use EM algorithms to extend estimation for situation when some observations are missing.

*Keywords: (almost) periodically correlated time series, cyclostationary signals, EM algorithm, missing data*

Streszczenie

W ostatnim czasie wzrasta zainteresowanie modelowaniem cyklostacjonarnych szeregów czasowych. W pracy Gardner i inni [5] cytowane jest ponad 1500 publikacji poświęconych temu zagadnieniu. Jednakże dane, dla których model cyklostacjonarny jest zasadny, są często niekompletne. Zgodnie z naszą wiedzą nie było do tej pory systematycznego omówienia tego problemu. Celem niniejszego artykułu jest uzupełnienie tej luki. W artykule proponujemy wykorzystanie algorytmu EM w celu estymacji parametrów modelu w sytuacji brakujących obserwacji.

*Słowa kluczowe: (prawie) okresowe szeregi czasowe, sygnały cyklostacjonarne, algorytm EM, brakujące obserwacje*

*Department of Statistics, University of California, Davis; cmdrake@ucdavis.edu
**Oskar Knapik, CREATES, Department of Economics and Business, Aarhus University; knapiko@uek.krakow.pl
***Institute of Mathematics, Cracow University of Technology, Poland; jleskow@riad.pk.edu.pl

## 1. Cyclostationary time series

The starting point of this research is the analysis of nonstationary time series. Let us assume that a time series $\{y_t, t \in N\}$ where $N$ represents the integers, is observed. The cyclostationarity of $\{y_t, t \in N\}$ means repeatable behavior of the first and second order characteristic of such a time series. Let us denote the mean $\mu_Y(t) = E(y_t)$ and the autocovariance function $B_Y(t,\tau) = Cov(y_t, y_{t+\tau})$. The time series will be called *cyclostationary* or *periodically correlated* if the mean $\mu_Y(t)$ and the autocovariance function $B_Y(t,\tau)$ are periodic functions of $t$ (see [9]). A classical statistical inference for analysis of such time series was presented in paper [4] among others. In our research, we will focus on second-order properties of the time series $\{y_t, t \in N\}$. Therefore, we assume that the time series under study is zero-mean.

The aim of this paper is to provide statistical inference procedures in situations where complete observation of the cyclostationary time series is impossible. Missing data analysis of cyclostationary time series frequently arises in the context of economic time series, mechanical signals and also ocean signals [12]. So far no systematic research has been presented in the case of missing data for periodically correlated time series. This work attempts to fill this gap using likelihood based inference and EM algorithms. This paper is divided into six sections. Section 2 presents state of the art results of likelihood-based inference for cyclostationary time series. Section 3 presents inferential methods in the case of missing data. Section 4 describes the EM algorithm in our cyclostationary model with missing data. The original results of this paper are presented in Section 2 where the full likelihood approach in cyclostationary model is presented. In Section 4, the original results are concerned with conditional distributions of our model, thus enabling applications of EM algorithm. In Section 5 the original results of this paper are illustrated with the help of a simulation study. Finally, Section 6 describes further directions of our research.

## 2. Likelihood-based inference for cyclostationary time series

A special class of cyclostationary time series (amplitude-modulated time series) is studied. Such time series can be represented as:

$$y_t = x_t \cdot c_t, \tag{2.1}$$

where

$\{x_t\}$ −a stationary time series (e.g. Gaussian $AR(p)$) ,

$c_t$ −a (deterministic) periodic function.

In model (2.1) it is assumed that:

**(AS1)** The deterministic sequence $c_t \neq 0 \ \forall t$ to prevent from deterministic zeros.

**(AS2)** $\{x_t\}$ is a zero-mean stationary Gaussian sequence with a bounded and continuous spectral density.

We consider $\{x_t\}$to be a Gaussian autoregressive process of order $p$, $AR(p)$, given by

$$x_t - \sum_{i=1}^{p} \varphi_i x_{t-i} = \varepsilon_t \tag{2.2}$$

where

$\{\varepsilon_t\}$ −a sequence of independent Gaussian zero-mean random variables with finite variance $\sigma^2$.

The deterministic sequence $c_t$ is assumed to be a known, periodic function of a finite dimensional unknown vector $\lambda$. The periodic function $c_t$ with $k$ periodicities can be expressed in the following way:

$$c_t = \exp\left(\sum_{j=1}^{k} (\lambda_{1j} \cos \omega_j t + \lambda_{2j} \sin \omega_j t)\right), \tag{2.3}$$

We assume that $k$ is known. We assume furthermore that all frequencies are of the form $\omega_i \in (0, \pi], i = 1, \ldots, k : \omega_i = 2\pi r/P$ for some $r = 1, \ldots, (P-1)/2$, where $P$ is a known period and also that $\lambda_{1j}^2 + \lambda_{2j}^2 > 0$ for $j = 1, \ldots, k$ .

Following approach presented in [7] to ensure the identifiability of the model parameters, $\lambda$ and the sequence $\{c_t\}_{t=1}^{T}$ are assumed to be linked via a one-to-one transformation, and it is assumed that there is no scale ambiguity in $y_t$.

Let $\theta = (\varphi, \lambda, \sigma^2)^T$ be the vector of all unknown parameters and suppose that we have $T = n + p$ observations from the model. The full likelihood for a vector of observations $y = (y_1, \ldots, y_T)^T$ corresponding to the model (2.1) is represented as:

$$L(y, \theta) = \frac{1}{(2\pi)^{T/2}} \cdot \det(R_Y)^{-1/2} \cdot \exp\left\{-\frac{1}{2} y^T R_Y^{-1} y\right\}, \tag{2.4}$$

where

$$R_Y = C_T R_X C_T^T,$$

$R_X$ −the covariance matrix of $AR(p)$ process,

$C_T$ −the $T \times T$diagonal matrix whose diagonal vector is $(c_1, \ldots, c_T)$ and $c_t = c_{i+mP} = c_i$.

Using [13] the inverse of the covariance matrix of an autoregressive process can be represented as follows

$$R_X^{-1} = \frac{1}{\sigma^2}\left(I_T + \sum_{i=1}^{p} \varphi_i^2 E_i - \sum_{i=1}^{p} \varphi_i F_i + \sum_{h=1}^{p-1} \sum_{i=1}^{p-h} \varphi_i \varphi_{i+h} G_{i,i+h}\right), \tag{2.5}$$

where

$I_T$ −the identity matrix of the order $T$,

$E_i$ −the identity with the first and last $i$ ones set to zero,

$F_i$ −the matrix which has ones along the upper and lower $i$th minor diagonals and zeros elsewhere,

$G_{i,i+h} = E_h F_i E_h$,thereby equaling $F_i$ except the top and bottom $h$ ones along the $i$th minor diagonals are replaced by zeros.

The sums in formula above are defined as zero if the upper limit of the summation is zero.

On the other hand, consider a vector of observations$y = (y_T, \ldots, y_1)^T$. From the model (2.1) it is obtained that

$$y = C \cdot x, \tag{2.6}$$

where

$C$ −an $T \times T$ diagonal matrix with diagonal vector $c = (c_T, \ldots, c_1)$,

$x = (x_T, \ldots, x_1)^T$−a vector of observations from $AR(p)$ model.

Under the assumptions above, we have the following

**Theorem 2.1.**

Assume that (AS1) and (AS2) hold and that the time series $\{y_t\}$ follows model (2.1). Then, the log-likelihood function for the complete sample has the form

$$\begin{aligned} &l(y;\theta) = \log f_y(y;\theta) = -\log(|\det C|) + \log f_x\left(C^{-1}y;\theta\right) \\ &= -\log(|\det C|) - \tfrac{T}{2}\log(2\pi) - \tfrac{T}{2}\log\left(\sigma^2\right) + \tfrac{1}{2}\log\left|V_p^{-1}\right| - \tfrac{1}{2\sigma^2}\left(C_p\mathbf{y}_p\right)^T V_p^{-1}\left(C_p\mathbf{y}_p\right) \\ &- \sum_{t=p+1}^{T} \tfrac{1}{2\sigma^2}\left(\tfrac{y_t}{c_t} - \sum_{i=1}^{p}\tfrac{y_{t-i}}{c_{t-i}}\right)^2, \end{aligned} \tag{2.7}$$

where

$\theta = (\varphi, \lambda, \sigma^2)^T$ −the vector of all unknown parameters

**Proof.**

The starting point is the observation that

$$f_y(y) = |\det C|^{-1} f_x\left(C^{-1}y\right). \tag{2.8}$$

Using that fact and properties of $AR(p)$ process (see [8]) the joint density for the complete data set can be written as

$$f_y(y;\theta) = |\det C|^{-1} \cdot f_{x_p,\ldots,x_1}\left(\frac{y_p}{c_p}, \ldots, \frac{y_1}{c_1}\right) \cdot \prod_{t=p+1}^{T} f_{x_t|x_{t-1},\ldots,x_{t-p}}\left(\frac{y_t}{c_t}\,\middle|\,\frac{y_{t-1}}{c_{t-1}}, \ldots, \frac{y_{t-p}}{c_{t-p}}\right). \tag{2.9}$$

The density of the first $p$ observations $f_{x_p,\ldots,x_1}(\cdot)$ is of a $N(0,\sigma^2 V_p)$ variable:

$$f_{x_p,\ldots,x_1}\left(\frac{y_p}{c_p},\ldots,\frac{y_1}{c_1};\theta\right)=(2\pi)^{-p/2}\left(\sigma^{-2}\right)^{p/2}\left|V_p^{-1}\right|^{1/2}\exp\left[-\frac{1}{2\sigma^2}\left(C_p^{-1}\mathbf{y}_p\right)^T V_p^{-1}\left(C_p^{-1}\mathbf{y}_p\right)\right], \tag{2.10}$$

where $\mathbf{y}_p=(y_p,\ldots,y_1)^T$ and $C_p$ is a $p\times p$ diagonal matrix with diagonal vector $(c_p,\ldots,c_1)$ and $V_p^{-1}$ is the inverse covariance matrix given by [3] .

$$v^{ij}(p)=v^{ji}(p)=\left[\sum_{k=0}^{i-1}\varphi_k\varphi_{k+j-i}-\sum_{k=p+1-j}^{p+i-j}\varphi_k\varphi_{k+j-i}\right] \tag{2.11}$$

for $1,,i,,j,,p$, where $\varphi_0=-1$ .

For the remaining observations in the sample the prediction-error decomposition can be used as in $AR(p)$ case. Conditional on the first $t-1$ observations, the density of $t$th observation $f_{x_t|x_{t-1},\ldots,x_{t-p}}(\cdot)$ is

$$f_{x_t|x_{t-1},\ldots,x_{t-p}}\left(\frac{y_t}{c_t}\left|\frac{y_{t-1}}{c_{t-1}},\ldots,\frac{y_{t-p}}{c_{t-p}}\right.\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{y_t}{c_t}-\sum_{i=1}^{p}\varphi_i\frac{y_{t-i}}{c_{t-i}}\right)^2\right] \tag{2.12}$$

The log-likelihood function for the sample has the form

$$\begin{array}{l} l(y;\theta)=\log f_y(y;\theta)==-\log\left(|\det C|\right)+\log f_x\left(C^{-1}y;\theta\right)= \\ =-\log\left(|\det C|\right)-\frac{T}{2}\log(2\pi)-\frac{T}{2}\log\left(\sigma^2\right)+\frac{1}{2}\log\left|V_p^{-1}\right|-\frac{1}{2\sigma^2}\left(C_p\mathbf{y}_p\right)^T V_p^{-1}\left(C_p\mathbf{y}_p\right) \\ -\sum\limits_{t=p+1}^{T}\frac{1}{2\sigma^2}\left(\frac{y_t}{c_t}-\sum\limits_{i=1}^{p}\frac{y_{t-i}}{c_{t-i}}\right)^2, \end{array} \tag{2.13}$$

This completes the proof of Theorem 2.1.

We have the following result

**Corrollary 2.2.** Let us consider the model $y_t=c_t x_t$, where $\{x_t\}$ is $AR(1)$ process. For the $AR(1)$ process $V_p^{-1}$ is a scalar whose value is found by taking $i=j=p=1$:

$$V_1^{-1}=\left(1-\varphi^2\right). \tag{2.14}$$

Thus $\sigma^2 V_1=\sigma^2\left(1-\varphi^2\right)$ which indeed reproduces the formula for the unconditional variance of the $AR(1)$ process.

The exact likelihood for the vector of observations $y$ from the above model is given as

$$l(y;\theta) = -\log(|\det C|) - \frac{T}{2}\log(2\pi) - \frac{T}{2}\log\left(\sigma^2\right) + \frac{1}{2}\log\left|\left(1-\varphi^2\right)\right| \\ -\frac{1}{2\sigma^2}\frac{y_1^2}{c_1^2}\left(1-\varphi^2\right) - \sum_{t=2}^{T}\frac{1}{2\sigma^2}\left(\frac{y_t}{c_t} - \varphi\frac{y_{t-1}}{c_{t-1}}\right)^2. \tag{2.15}$$

It is possible to give an exact formula for the ML estimate of $\sigma^2$ conditional on $\varphi$ and $\lambda$. There is no analytical form of the ML estimates of $\varphi$ and $\lambda$ even conditionally on the other parameters. Because of this the maximization of the exact log likelihood function must be accomplished numerically. It is a high dimensional nonlinear optimization problem. It is worth mentionning that parameters of the deterministic and stochastic parts of this model are linked via a nonlinear relationship.

## 3. Missing Data Mechanisms and Inference

Data that may reasonably be modeled by cyclostationary time series are often incomplete for a number of reasons, e.g. the interruption of measurements due to instrument failure or extreme natural phenomena, the accidental loss of data or erroneous measurements, among others ([12]). Similar problems are encountered in vibroacoustic and economic time series.

Missing data present some potentially serious problems in drawing inferences from time series. The degree to which conclusions under such circumstances are affected, depends on the mechanism by which the data is missing. Missing data mechanisms can be divided into roughly three categories (see [10] and [11]): *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR).

For the purpose of formal description of missingness mechanism let us define the complete data $y = (y_t)$, where observation $y_t$ comes from the data generating process $P_\theta$ parameterized by the unknown vector of parameters $\theta$ for $t = 1, \ldots, T$. Let us also define the missing-data indicator vector $m = (m_t)$, where a random variable $m_t$

$$m_t = \begin{cases} 1, \text{ if } y_t \text{ is missing} \\ 0, \text{ if } y_t \text{ is present} \end{cases}, \tag{3.1}$$

has a distribution $P_\psi$. Let us also assume that $\theta$ and $\Psi$ are distinct. Denote the joint distribution of $(y, m)$ by $P(y, m; \theta, \psi)$, where $y$ consists of two parts. The observed part is denoted by $y_{obs}$ and the missing part by $y_{mis}$. The missing-data mechanism is defined by means of conditional distribution of $m$ given $y$, which is$P(m|y)$.

If data is *missing completely at random* (MCAR) we have

$$P(m_t = 1|y;\theta,\psi) = P(m_t = 1|\psi). \tag{3.2}$$

When data is *missing at random* (MAR), we have

$$P(m_t = 1|y;\theta,\psi) = P(m_t = 1|y_{obs};\theta,\psi). \tag{3.3}$$

When neither MCAR nor MAR hold, we say the data are missing not at random, abbreviated MNAR. Missingness depends on the unobserved data.

Missing completely at random is easiest to deal with. The missing observations constitute a random sub-sample of all observations. Estimates of population parameters are unbiased. Precision of estimates will be affected, particularly if the original sample size was modest and the numbers of missing observations constitute a substantial fraction of all observations.

MAR allows likelihood based and Bayesian inference and does not require modeling of the missingness mechanism. Inference can be based on the observed data likelihood. When data is MNAR unbiased inference is not possible without further assumptions and additional information. Finally, MAR and/or MCAR cannot be established on the basis of observed data alone and require additional information.

Often in cyclostationary data, values are missing, because of instrument failure. This instrumental failure may or may not depend on the missing values.

Instruments sometimes have a limited range for signal detection. This can cause data to be missing, because missing values are below or above the detection limits of the instrument. This is a special type of informative missingness.

The type of missingness mechanism considered in this paper will be MCAR or MAR. The first missingness mechanism (MCAR) assumes that failure to observe the data does not depend on the data. The second missingness mechanism (MAR) is more general than the first: here it is possible that the missingness mechanism depends on observed data.

Within this research we consider only the case of missing at random (MAR)

## 4. EM algorithm in likelihood-based inference for cyclostationary time series with missing observations

The expectation-maximization (EM) algorithm is an iterative procedure for computing the maximum likelihood estimator for data set which is not complete. [2] showed the wide applicability of the EM algorithm in statistics. The convergence and performance of the EM algorithm was proved by [14].

Let $y$ be a complete data vector which consists of $y_{mis}$ missing data and $y_{obs}$ "observed" data. The EM algorithm is an iterative procedure for computing the maximum likelihood estimator only on the basis of the observed data$y_{obs}$. Each iteration of EM algorithm consists of two steps. If $\theta^{(i)}$ denotes the estimated value of the parameter $\theta$ after $i$ iterations, then the two steps in the $(i+1)$th iteration are

E-step: Calculate $Q\left(\theta|\theta^{(i)}\right) = E_{\theta^{(i)}}\left[l\left(\theta; y_{obs}, y_{mis}\right)|y_{obs}\right]$

M-step: Maximize $Q\left(\theta|\theta^{(i)}\right)$ with respect to$\theta$.

Then $\theta^{(i+1)}$ is set to the maximizer of $Q$ in the M-step.

For our model, we have two possible realizations of EM algorithm: one on the basis of conditional likelihood function which leads to modified EM algorithm for normal linear regression settings and on the basis of full likelihood function which

is well-known problem of inference for incomplete data within multivariate normal distribution.

Before description of EM algorithm, let us take a look at properties of considered model.

Many properties of the cyclostationary model (2.1) follow from the autoregressive structure of $\{x_t\}$. It is clear that if $\{x_t\}$ follows the zero-mean Gaussian $AR(p)$ process then

$$x_t|x_{t-1},\ldots,x_{t-p} \sim N\left(\sum_{j=1}^{p}\varphi_i x_{t-i},\sigma^2\right). \tag{4.1}$$

When $y_t = c_t \cdot x_t$, where $\{x_t\}$ is $AR(p)$ process and $c_t$ is a deterministic periodic function then assuming that $c_t \neq 0$ for all $t$ we can write $x_t = y_t/c_t$.

The following result establishes the form of the conditional distribution, the conditional expectation and the conditional variance given the past.

**Theorem 4.1.** Under the assumptions (AS1) and (AS2) and the model equation (2.1) one obtains

$$y_t|y_{t-1},\ldots,y_{t-p} \sim N\left(c_t\left(\sum_{i=1}^{p}\varphi_i\frac{y_{t-i}}{c_{t-i}}\right), c_t^2\sigma^2\right), \tag{4.2}$$

$$E\left(y_t|y_{t-1},\ldots,y_{t-p}\right) = c_t\left(\sum_{i=1}^{p}\varphi_i\frac{y_{t-i}}{c_{t-i}}\right), \tag{4.3}$$

and

$$Var\left(y_t|y_{t-1},\ldots,y_{t-p}\right) = c_t^2\sigma^2. \tag{4.4}$$

**Proof** is straightforward and will be omitted.

In missing data analysis one frequently confronts the situation of 'filling the gaps' that is calculating the conditional expectation given the past and future. For that purpose consider the distribution $p(y_{mis}|y_{obs})$. Due to $AR(p)$ structure of $\{x_t\}$ one obtains

$$x_t = \varphi_1 x_{t-1} + \ldots + \varphi_p x_{t-p} + \varepsilon_t,\ \{\varepsilon_t\} \sim N\left(0,\sigma^2\right) \tag{4.5}$$

taking $x_t = y_t/c_t$, we have

$$\frac{y_t}{c_t} = \varphi_1\frac{y_{t-1}}{c_{t-1}} + \ldots + \varphi_p\frac{y_{t-p}}{c_{t-p}} + \varepsilon_t. \tag{4.6}$$

Denote observed values $y = (y_{i_1}, \ldots, y_{i_r})^T$, with $1?i_1 < \ldots < i_r?T$. If there are no missing observations in the first $p$ observations, then the best estimates of the missing values are found by minimizing

$$\sum_{t=p+1}^{T} \left( \frac{y_t}{c_t} - \varphi_1 \frac{y_{t-1}}{c_{t-1}} - \ldots - \varphi_p \frac{y_{t-p}}{c_{t-p}} \right)^2 \tag{4.7}$$

with respect to the missing values. The minimization of the sum above gives the form of the conditional expectation.

**Theorem 4.2.**
Consider the following sequence $(y_j, y_{j+1}, y_{j+2})$ from the process $y_t = c_t \cdot x_t$, where $c_t$ is (deterministic) periodic function and $\{x_t\}$ is $AR(1)$ process defined by

$$x_t = \varphi x_{t-1} + \varepsilon_t, \, \varepsilon_t \sim N\left(0, \sigma^2\right). \tag{4.8}$$

The conditional expectation of $y_{j+1}$ given $y_j$ (past) and $y_{j+2}$ (future) has the following form:

$$E\left(y_{j+1} | y_j, y_{j+2}\right) = c_{j+1}\varphi \left( \frac{y_j}{c_j} + \frac{y_{j+2}}{c_{j+2}} \right) \Big/ \left(1 + \varphi^2\right). \tag{4.9}$$

**Proof.**
The form of the conditional expectation can be obtained as the minimization of the above sum

$$\left( \frac{y_{j+1}}{c_{j+1}} - \varphi \frac{y_j}{c_j} \right)^2 + \left( \frac{y_{j+2}}{c_{j+2}} - \varphi \frac{y_{j+1}}{c_{j+1}} \right)^2, \tag{4.10}$$

with respect to $y_{j+1}$. Setting the derivative of this expression with respect to $y_{j+1}$ equal to 0 and solving for $y_{j+1}$ and using properties of conditional expectation.

Suppose $y = (y_1, \ldots, y_T)^T$ be the "complete" data vector of which $r$ are observed and $T - r$ are missing. Denote the "observed" data by $y_{obs} = (y_{i_1}, \ldots, y_{i_r})$ (called "incomplete" data) and missing data by $y_{mis} = (y_{j_1}, \ldots, y_{j_{T-r}})$.

If we work with full likelihood function, we have that $y = (y_{obs}, y_{mis})$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $R_Y$ which depends on the parameter $\theta$, the log-likelihood of the complete data is given by

$$l\left(\theta, y\right) = -\frac{T}{2} \ln\left(2\pi\right) - \frac{1}{2} \ln \det\left(R_Y\right) - \frac{1}{2} y^T R_Y^{-1} y \tag{4.11}$$

The E step requires that we compute the expectation of $l(\theta, y)$ with respect to the conditional distribution of $y$ given $y_{obs}$ with $\theta = \theta^i$. Following the approach presented in [1] let us consider $R_Y(\theta)$ as the block matrix

$$R_Y = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{4.12}$$

which is conformable with $y_{mis}$ and $y_{obs}$, the conditional distribution of $y$ given $y_{obs}$ is multivariate normal with mean
and covariance matrix $\begin{pmatrix} \Sigma_{11|2}(\theta) & 0 \\ 0 & 0 \end{pmatrix}$, where $\hat{y}_{mis} = E_\theta(y_{mis}|y_{obs}) = \Sigma_{12}\Sigma_{22}^{-1}y_{obs}$ and $\Sigma_{11|2}(\theta) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. It can be shown that

$$E_{\theta^{(i)}}\left(\left(y_{mis}^T, y_{obs}^T\right)^T R_Y^{-1}(\theta)\left(y_{mis}^T, y_{obs}^T\right)|y_{obs}\right) = trace\left(\Sigma_{11|2}\left(\theta^{(i)}\right)\Sigma_{11|2}^{-1}(\theta)\right) + \hat{y}^T R_Y^{-1}(\theta)\,\hat{y}, \tag{4.13}$$

where $\hat{y} = (\hat{y}_{mis}, y_{obs})$.

As a consequence

$$Q\left(\theta|\theta^{(i)}\right) = l(\theta;\hat{y}) - \frac{1}{2}trace\left(\Sigma_{11|2}\left(\theta^{(i)}\right)\Sigma_{11|2}^{-1}(\theta)\right) \tag{4.14}$$

The first term on the right is the log-likelihood based on the complete data, but with $y_{mis}$ replaced by its "best estimate" $\hat{y}_{mis}$ calculated from the previous iteration. If the increments $\theta^{(i+1)} - \theta^{(i)}$ are small, then the second term on the right is nearly constant ($\approx T - r$) and can be ignored. To make computation easier we can use the modified version

$$\tilde{Q}\left(\theta|\theta^{(i)}\right) = l(\theta;\hat{y}). \tag{4.15}$$

With this modification, the steps in the EM algorithm are as follows:

E-step: Calculate $E_{\theta^{(i)}}(y_{mis}|y_{obs})$ and form $\tilde{Q}\left(\theta|\theta^{(i)}\right)$.

M-step: Find the maximum likelihood estimator for "complete" data problem, i.e. maximize $l(\theta, \hat{y})$.

The best linear predictor of a missing observation $y_{j_k}$ from the vector $y_{mis}$ is $E(y_{j_k}|y_{obs})$, so within E-step we reconstruct "complete" observations in the following way:

$$\hat{y}_t^{(i)} = \begin{cases} y_t, \text{ if } y_t \text{ is present} \\ E\left(y_t|y_{obs};\theta^{(i)}\right), \text{ if } y_t \text{ is missing} \end{cases} \tag{4.16}$$

In M-step, we maximize likelihood function of vector $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_T)$ with respect to $\theta$.

Tab. 1: Results of ML and EM algorithm-based estimation

| Method | $\hat{\varphi}$ | $\hat{\sigma}^2$ | $\hat{\lambda}_1$ | $\hat{\lambda}_2$ |
|---|---|---|---|---|
| *ML estimation* | 0,4976 | 1,2946 | 0,4077 | 0,5387 |
| *EM algorithm* | 0,5599 | 1,3329 | 0,3877 | 0,5523 |

## 5. Simulation study

To see how approximation of the EM algorithm works in practice, we restrict our attention to the class of PC time series of the form

$$y_t = c_t x_t, \tag{5.1}$$

where $x_t = \varphi x_{t-1} + ?_t$, $\{?_t\} \sim N(0, \sigma^2)$ and $c_t = \exp\left(\lambda_1 \cos\left(\frac{2\pi}{20}t\right) + \lambda_2 \sin\left(\frac{2\pi}{20}t\right)\right)$. The following values of parameters were chosen $\varphi = 0.5$, $\sigma^2 = 1$, $\lambda_1 = 0.4$ and $\lambda_2 = 0.5$.

Firstly we simulate $T = 100$ observations from our model and estimate unknown vector of parameters $\theta$ on the basis of complete sample. Then we randomly choose 10% of the data to be missing.

The results are presented in the Table 1.

It can be seen that the approximation of the EM algorithm gives reasonable estimates of unknown parameters $\theta$ in situation when some observations are missing. The question of convergence, however, needs to be further explored.

*Oskar Knapik gratefully acknowledges the Kosciuszko Foundation for financial support for this research*

## References

[1] Brockwell P.J., Davis R.A., *Introduction to time series and forecasting*, Springer, 2002.

[2] Dempster A.P., Laird N.M. and Rubin D.B., *Maximum Likelihood from Incomplete Data via The EM Algorithm (with discussion)*, J. Roy. Statist. Soc. B, Vol. 39, 1977, 1—38.

[3] Galbraith R.F., Galbraith J.I., *On the inverses of some patterned matrices arising in the theory of stationary time series*, Journal of Applied Probability, Vol. 11, 1974, 63—71.

[4] Gardner W.A., *Representation and estimation of cyclostationary processes*, IEEE Transactions on Information Theory, Vol. 19, No. 3, 1973, 375—376.

[5] Gardner W.A., Napolitano A., Paura L., *Cyclostationarity: Half a century of research*, Signal Processing, Vol. 86, No. 4, 2006, 639—697.

[6] Giannakis,G.B, Dandawate, A.V., *Consistent Kth-order time-frequency representations for (almost) cyclostationary signals*, in: IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Victoria, BC, Canada, 79 October 1992, 123—126.

[7] Ghogho, M., Garel, B., *Maximum likelihood estimation of amplitude-modulated time series*, Signal Processing, Vol. 75, 1999 99—116.

[8] Hamilton J.D., *Time Series Analysis*, Princeton University Press, 1994.

[9] Hurd H. L., Miamee A., *Periodically Correlated Random Sequences: Spectral Theory and Practice (Wiley Series in Probability and Statistics),* Wiley-Interscience, 2007.

[10] Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 2nd edn., 2002.

[11] Schaffer J.L., *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC, 1997.

[12] Stefanakos Ch.N., Athanassoulis G.A., *A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data*, Applied Ocean Research, Vol. 23, Issue 4, 2001, 207—220.

[13] Verbyla A.P., *A note on the inverse covariance matrix of the autoregressive process*, Australian Journal of Statistics, Vol. 27, Issue 2, 1985, 221—224.

[14] Wu C.F.J., *On the convergence properties of the EM algorithm*, Annals of Statistics, Vol. 11, 1983, 95—103.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

GRZEGORZ GANCARZEWICZ*

# CYCLES CONTAINING SPECIFIED EDGES IN A GRAPH

# CYKLE ZAWIERAJĄCE WYBRANE KRAWĘDZIE GRAFU

Abstract

The aim of this paper is to prove that if $s \geqslant 1$ and $G$ is a graph of order $n \geqslant 4s + 6$ satisfying

$$\sigma_2 \geqslant \frac{4n - 4s - 3}{3},$$

then every matching of $G$ lies on a cycle of length at least $n - s$ and hence, in a path of length at least $n - s + 1$.

*Keywords: cycle, graph, hamiltonian cycle, hamiltonian path, matching, path*

Streszczenie

W pracy udowodniono, że dla $s \geqslant 1$ w dowolnym grafie $G$ rzędu $n \geqslant 4s + 6$ spełniającym

$$\sigma_2 \geqslant \frac{4n - 4s - 3}{3},$$

każde skojarzenie jest zawarte w cyklu długości co najmniej $n - s$ i stąd w ścieżce długości co najmniej $n - s + 1$.

*Słowa kluczowe: cykl, cykl hamiltonowski, graf, skojarzenie, ścieżka, ścieżka hamiltonowska*

*Institute of Mathematics, Cracow University of Technology, Poland; ggancarzewicz@gmail.com

## 1. Introduction

We consider only finite graphs without loops and multiple edges. By V or V($G$) we denote the vertex set of graph $G$ and respectively, by E or E($G$), the edge set of $G$. By $\mathrm{d}_G(x)$ or $\mathrm{d}(x)$ we denote *the degree of a vertex $x$ in the graph $G$.*

In the proof we will only use oriented cycles and paths. Let $C$ be a cycle and $x \in V(C)$, then $x^-$ is *the predecessor of* $x$ and $x^+$ is its *successor.*

Let us introduce the $\sigma_k$.

**Definition 1.1.** *Let $G$ be a graph and $k \geqslant 0$.*

$$\sigma_k = \min\{\sum_{i=1}^{k} \mathrm{d}(x_i) : \{x_1, \ldots, x_k\} \subset \mathrm{V}(G) \text{ and independent } \}$$

In 1960, O. Ore [5] proved the following:

**Theorem 1.1.** *Let $G$ be a graph on $n \geqslant 3$ vertices. If $G$ satisfies*

$$\sigma_2 \geqslant n$$

*then $G$ is hamiltonian.*

The condition for degree sum in Theorem 1.1 is called *an Ore condition* or *a Ore type condition for graph $G$.*

The Ore condition for a graph $G$ :

$$\sigma_2 \geqslant l$$

can also be written as:

$$\text{If } x, y \in \mathrm{V}(G),\ xy \notin \mathrm{E}(G), \quad \text{then:} \quad \mathrm{d}(x) + \mathrm{d}(y) \geqslant l.$$

There is also a similar condition, proved by V. Chvátal in [3], under which graph $G$ has a hamiltonian path.

**Theorem 1.2.** *If $G$ is a graph on $n \geqslant 3$ vertices satisfying*

$$\sigma_2 \geqslant n - 1, \tag{1.1}$$

*then $G$ has a hamiltonian path.*

We shall call a set of $k$ independent edges of graph $G$ a *$k$-matching* or simply a *matching.*

About graphs with every $k$-matching in a hamiltonian cycle or path Las Vergnas obtained the following two results:

**Theorem 1.3.** *Let $G$ be a graph on $n \geqslant 3$ vertices and let $k$ be an integer such that $0 \leqslant k \leqslant \frac{n}{2}$. If $G$ satisfies*

$$\sigma_2 \geqslant n + k - 1\,,$$

*then every $k$-matching of $G$ lies in a hamiltonian path.*

**Theorem 1.4.** *Let $G$ be a graph on $n \geqslant 3$ vertices and let $k$ be an integer such that $0 \leqslant k \leqslant \frac{n}{2}$. If $G$ satisfies*

$$\sigma_2 \geqslant n + k\,,$$

*then every $k$-matching of $G$ lies in a hamiltonian cycle.*

K.A. Berman proved in [1] the following result conjectured by R. Häggkvist in [4].

**Theorem 1.5.** *Let $G$ be graph of order $n$. If $G$ satisfies*

$$\sigma_2 \geqslant n + 1,$$

*then every matching lies in a cycle.*

Now we shall define a family of graphs $\mathcal{G}_n$. If $\frac{n+2}{3}$ is an integer, $\mathcal{G}_n$ is a family of graphs:

$$G = \frac{n+2}{3}\, K_1 \star H,$$

where $\star$ denotes join and $H$ is a graph of order $\frac{2n-2}{3}$ containing a perfect matching. Otherwise, $\mathcal{G}_n$ is empty.

In 1983 Wojda [6] proved the following Ore type theorem:

**Theorem 1.6.** *Let $G$ be a graph on $n \geqslant 3$ vertices.If $G$ satisfies*

$$\sigma_2 \geqslant \frac{4n-4}{3}\,.$$

*Then every matching of $G$ lies in a hamiltonian cycle or $G \in \mathcal{G}_n$.*

In this paper, we shall find an Ore type condition under which every matching in a graph $G$ lies in a cycle of length at least $n - s$ and hence, in a path of length at least $n - s + 1$.

For the notation and terminology not defined above, a good reference should be [2].

## 2. Results

We proved the following improvement of Theorem 1.6 for matchings.

**Theorem 2.1.** *Let $s \geqslant 1$ and let $G$ be a graph of order $n \geqslant 4s+6$ satisfying*

$$\sigma_2 \geqslant \frac{4n-4s-3}{3}. \tag{2.1}$$

*Then, every matching of $G$ lies on a cycle of length at least $n-s$ and hence in a path of length at least $n-s+1$.*

The special case of this theorem for $s=1$ is:

**Corollary 2.2.** *Let $G$ be a graph of order $n \geqslant 10$, satisfying*

$$\sigma_2 \geqslant \frac{4n-7}{3}.$$

*Then, every matching of $G$ lies on a cycle of length at least $n-1$ and hence in a hamiltonian path.*

Obviously for $k \geqslant \frac{n-2}{3}$, the bound for $\sigma_2$ is lower in Corollary 2.2 than the bound from Theorem 1.3.

Suppose that $s \geqslant 1$ is such that $n \geqslant 4s+6$ and $\frac{n+2s}{3} \geqslant 2$ is an integer.

Now consider the graph $G' = (\frac{n+2s}{3} - 1)\, K_1 * K_{\frac{2n-2s}{3}}$, where $*$ denotes the join of graphs.

We shall define a graph $G''$ as a graph obtained from $G'$ by adding an external vertex $x$ adjacent only to $\frac{2n-2s}{3} - 1$ vertices from $K_{\frac{2n-2s}{3}}$ i.e. we take $\mathrm{V}(G'') = \mathrm{V}(G') \cup \{x\}$, next we choose an arbitrary vertex $h_0 \in \mathrm{V}(K_{\frac{2n-2s}{3}})$ and we put $\mathrm{E}(G'') = \mathrm{E}(G') \cup \{xh : h \in \mathrm{V}(K_{\frac{2n-2s}{3}}) \setminus \{h_0\}\}$. Note theat $G''$ is a graph of order $n$.

Let $u \in \mathrm{V}(K_1)$, then $\mathrm{d}_{G''}(u) = \frac{2n-2s}{3} - 1$ and $\mathrm{d}_{G''}(x) + \mathrm{d}_{G''}(h_0) = \frac{4n-4s-3}{3}$, the graph $G''$ satisfies the assumptions of Theorem 2.1, but violates those of Theorem 1.6. So, Theorem 1.6 and Theorem 2.1 are independent.

It is easy to check that even Corollary 2.2 cannot be obtained as a corollary of Theorem 1.6 by adding to the graph $G$ an external vertex $x$ adjacent to all vertices and removing an edge from the hamiltonian cycle in $G \cup \{x\}$. In this case, $G \cup \{x\}$ does not satisfy the assumptions of Theorem 1.6.

Obviously, for $k \geqslant \frac{n-2}{3}$, the bound for $\sigma_2$ is lower in Theorem 2.2 than the bound from Theorem 1.3.

## 3. Proof

**Proof of Theorem 2.1:**

Take any matching $M$ of $G$. Without loss of generality we can assume that $M$ is maximal, i.e. for any matching $M'$ of $G$ if $M \subset M'$, then $M = M'$.

Observe that since $n \geqslant 4s + 6$, we have $\frac{4n-4s-3}{3} \geqslant n + 1$. If the graph $G$ satisfies the assumptions of Theorem 2.1 then it also satisfies the assumptions of Theorem 1.5. From Theorem 1.5, we know that there is a cycle containing $M$.

Consider a cycle $C$ containing $M$ of maximal length. If $|\mathrm{V}(C)| \geqslant n - s$ the proof is finished. We suppose now that $|\mathrm{V}(C)| \leqslant n - s - 1$ and we give an arbitrary orientation to $C$.

Since $M$ is maximal, the set $\mathrm{V}(G \setminus C)$ is independent.

Since $s \geqslant 1$ we have $|\mathrm{V}(G \setminus C)| \geqslant 2$ and therefore we have two vertices $x$ and $y \in \mathrm{V}(G \setminus C)$ such that $xy \notin \mathrm{E}(G)$ and from (2.1) we have:

$$\mathrm{d}(x) + \mathrm{d}(y) \geqslant \frac{4n - 4s - 3}{3}. \tag{3.1}$$

Note that since $\mathrm{V}(G \setminus C)$ is independent we have:

$$\mathrm{d}_{G \setminus C}(x) = \mathrm{d}_{G \setminus C}(y) = 0. \tag{3.2}$$

On cycle $C$, consider a family of paths $Q_i$, $i \in \{1, \ldots, k\}$, obtained from $C$ by the removal of the edges of matching $M$.

Note that:

$$\sum_{i=1}^{k} |\mathrm{V}(Q_i)| = |\mathrm{V}(C)|. \tag{3.3}$$

$$\text{Since } M \text{ is maximal, } |\mathrm{V}(Q_i)| = 2 \text{ or } |\mathrm{V}(Q_i)| = 3. \tag{3.4}$$

**Remark 3.1.** *If $w \in \mathrm{V}(G \setminus C)$, then $w$ cannot be adjacent to two consecutive vertices on any path $Q_i$, for $i \in \{1, \ldots, k\}$.*

Suppose that $w \in \mathrm{V}(G \setminus C)$ and we have a vertex $u \in \mathrm{V}(Q_i)$ such that $u+ \in \mathrm{V}(Q_i)$ and $wu$, $wu+ \in \mathrm{E}$. In this case, the cycle:

$$C' : \quad uwu^+ \ldots u^- u$$

contains $M$ and is longer than $C$, contradiction with the choice of $C$.

**Case 1:** $|\mathrm{V}(Q_i)| = 2$

From Remark 3.1, we know that $\mathrm{d}_{Q_i}(x) \leqslant 1$ and $\mathrm{d}_{Q_i}(y) \leqslant 1$ and so:

$$\mathrm{d}_{Q_i}(x) + \mathrm{d}_{Q_i}(y) \leqslant 2 = |\mathrm{V}(Q_i)|. \tag{3.5}$$

**Case 2:** $|\mathrm{V}(Q_i)| = 3$

In this case, from Remark 3.1 we know that $\mathrm{d}_{Q_i}(x) \leqslant 2$ and $\mathrm{d}_{Q_i}(y) \leqslant 2$. Note that if $Q_i : q_1^i q_2^i q_3^i$ it is possible that $x$ and $y$ are adjacent at the same time to $q_1^i$ and $q_2^i$.

From the above we have:

$$\mathrm{d}_{Q_i}(x) + \mathrm{d}_{Q_i}(y) \leqslant 4 = |\mathrm{V}(Q_i)| + 1. \tag{3.6}$$

Consider now the set $I = \{Q_i : \ |\mathrm{V}(Q_i)| = 3 \text{ and } Q_i : \ q_1^i q_2^i q_3^i\}$, $l = |I|$. Observe that since $M$ is maximal we have:

1. The set $V_I = \{q_2^i : Q_i \in I\}$ is independent and $|V_I| = l$.
2. If $\mathrm{d}_{Q_i}(x) = 2$, then $xq_2^i \notin E$ and if $\mathrm{d}_{Q_i}(y) = 2$, then $yq_2^i \notin E$.

We have $l + 2$ independent vertices in $G : V_I \cup \{x, y\}$.

Thus:

$$\begin{aligned}\mathrm{d}_C(x) + \mathrm{d}_C(y) = \sum_{i=1}^{k} \mathrm{d}_{Q_i}(x) + \mathrm{d}_{Q_i}(y) \\ \leqslant \sum_{i=1}^{k} |\mathrm{V}(Q_i)| + l = |\mathrm{V}(C)| + l.\end{aligned} \tag{3.7}$$

From (3.2) and (3.7), we have:

$$\mathrm{d}(x) + \mathrm{d}(y) = \sum_{i=1}^{k} \mathrm{d}_{Q_i}(x) + \mathrm{d}_{Q_i}(y) \leqslant |\mathrm{V}(C)| + l. \tag{3.8}$$

Since $|\mathrm{V}(C)| \leqslant n - s - 1$ and $M$ is maximal, we have $l \leqslant \frac{n-s-1}{3}$ and so:

$$\mathrm{d}(x) + \mathrm{d}(y) \leqslant |\mathrm{V}(C)| + l \leqslant n - s - 1 + \frac{n-s-1}{3} = \frac{4n-4s-4}{3}. \tag{3.9}$$

Since $\mathrm{d}(x) + \mathrm{d}(y)$ is an integer, from (3.9), we have:

$$\mathrm{d}(x) + \mathrm{d}(y) \leqslant \left\lfloor \frac{4n-4s-4}{3} \right\rfloor < \frac{4n-4s-3}{3}, \tag{3.10}$$

a contradiction with (3.1) and the proof is finished.

□

## References

[1] Berman K.A., *Proof of a conjecture of Haggkvist on cycles and independent edges*, Discrete Mathematics 46, 1983, 9—13.

[2] Bondy J.A. and Murty U.S.R., *Graph theory with applications,* The Macmillan Press LTD, London 1976.

[3] Chvátal V., *On Hamilton's ideals*, J. Combin. Theory B **12**, 1972, 163—168.

[4] Häggkvist R., *On F-hamiltonian graphs* in Graph Theory and Related Topics, ed. J.A. Bondy and U.S.R. Murty, Academic Press N.Y. 1979 219—231.

[5] Ore O., *Note on hamiltonian circuits,* Amer. Math. Monthly 67, 1960, 55.

[6] Wojda, A.P. , *Hamiltonian cycles through matchings,* Demonstratio Mathematica XXI 2, 1983, 547—553.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

GRZEGORZ GANCARZEWICZ*

# GRAPHS WITH EVERY PATH OF LENGTH $k$ IN A HAMILTONIAN CYCLE

# GRAFY Z DOWOLNĄ ŚCIEŻKĄ DŁUGOŚCI $k$ ZAWARTĄ W PEWNYM CYKLU HAMILTONOWSKIM

Abstract

In this paper we prove that if $G$ is a $(k+2)$-connected graph on $n \geqslant 3$ vertices satisfying $\mathrm{P}(n+k)$ :

$$\mathrm{d}_G(x,y) = 2 \;\Rightarrow\; \max\{\mathrm{d}(x), \mathrm{d}(y)\} \geqslant \frac{n+k}{2}$$

for each pair of vertices $x$ and $y$ in $G$, then any path $S \subset G$ of length $k$ is contained in a hamiltonian cycle of $G$.

*Keywords: cycle, graph, hamiltonian cycle, matching, path*

Streszczenie

W pracy udowodniono, że W $(k+2)$-spójnym grafie $G$ o $n \geqslant 3$ wierzchołkach, który spełnia warunek $\mathrm{P}(n+k)$ :

$$\mathrm{d}(x,y) = 2 \;\Rightarrow\; \max\{\mathrm{d}(x), \mathrm{d}(y)\} \geqslant \frac{n+k}{2}$$

dla dowolnej pary wierzchołków $x$ i $y$, każda ścieżka $S \subset G$ długości $k$ jest zawrta w pewnym cyklu hamiltonowskim grafu $G$.

*Słowa kluczowe: cykl, cykl hamiltonowski, graf, skojarzenie, ścieżka*

*Institute of Mathematics, Cracow University of Technology; ggancarzewicz@gmail.com

## 1. Introduction

We consider only finite graphs without loops and multiple edges. By V or V($G$) we denote the vertex set of the graph $G$ and respectively by E or E($G$), the edge set of $G$. By $\mathrm{d}_x(G)$ or $\mathrm{d}(x)$, we denote *the degree of a vertex $x$ in the graph $G$* and by $\mathrm{d}(x,y)$ or $\mathrm{d}_G(x,y)$, *the distance between $x$ and $y$ in $G$.*

**Definition 1.1** (cf [10])**.** *Let $k, s_1, \dots s_\ell$ be positive integers. We call $S$ a path system of length $k$, if the connected components of $S$ are paths:*

$$\begin{aligned} P^1: &\quad x_0^1 x_1^1 \dots x_{s_1}^1, \\ &\quad\ \vdots \\ P^l: &\quad x_0^\ell x_1^l \dots x_{s_\ell}^\ell \end{aligned}$$

*and* $\sum_{i=1}^{\ell} s_i = k$.

Let $S$ be a path system of length $k$ and let $x \in \mathrm{V}(S)$. We shall call $x$ *an internal vertex* if $x$ is an internal vertex (cf [3]) in one of the paths $P^1, \dots, P^\ell$.

If $q$ denotes the number of internal vertices in a path system $S$ of length $k$ then $0 \leqslant q \leqslant k-1$. If $q = 0$, then $S$ is *a $k$-matching* (i.e. a set of $k$ independent edges).

Let $H$ be a subgraph of $G$. By $G \setminus H$ we denote the graph obtained from $G$ by the deletion of the edges of $H$.

**Definition 1.2.** *The graph $F$ is said to be an $H$-edge cut-set of $G$ if $F \subset \mathrm{E}(H)$ and $G \setminus F$ is not connected.*

**Definition 1.3.** *The graph $F$ is said to be a minimal $H$-edge cut-set of $G$ if $F$ is an $H$-edge cut-set of $G$ which has no proper subset being an edge cut-set of $G$.*

**Definition 1.4** (cf [7])**.** *Let $G$ be a graph on $n \geqslant 3$ vertices and $k \geqslant 0$. $G$ is $k$-edge-hamiltonian if for every path system $P$ of length at most $k$ there exists a hamiltonian cycle of $G$ containing $P$.*

Let $G$ be a graph and $H \subset G$ a subgraph of $G$. For a vertex $x \in \mathrm{V}(G)$, we define the set $N_H(x) = \{y \in \mathrm{V}(H) : xy \in \mathrm{E}(G)\}$. Let $H$ and $D$ be two subgraphs of $G$. $\mathrm{E}(D,H) = \{xy \in \mathrm{E}(G) : x \in \mathrm{V}(D) \text{ and } y \in \mathrm{V}(H)\}$. For a set of vertices $A$ of a graph $G$, we define the graph $G(A)$ as the subgraph induced in $G$ by $A$. In the proof, we will only use oriented cycles and paths. Let $C$ be a cycle and $x \in V(C)$, then $x^-$ is *the predecessor of* $x$ and $x^+$ is its *successor.*

**Definition 1.5** (cf [2])**.** *Let $W$ be a property defined for all graphs of order $n$ and let $k$ be a non-negative integer. The property $W$ is said to be $k$-stable if whenever $G + xy$ has property $W$ and $\mathrm{d}(x) + \mathrm{d}(y) \geqslant k$ then $G$ itself has property $W$.*

J.A. Bondy and V. Chvátal [2] proved the following theorem, which we shall need in the proof of our main result:

**Theorem 1.1.** *Let $n$ and $k$ be positive integers with $k \leqslant n-3$. Then the property of being $k$-edge-hamiltonian is $(n+k)$-stable.*

In 1960, O. Ore [9] proved the following:

**Theorem 1.2.** *Let $G$ be a graph on $n \geqslant 3$ vertices. If for all nonadjacent vertices $x, y \in \mathrm{V}(G)$ we have*

$$\mathrm{d}(x) + \mathrm{d}(y) \geqslant n$$

*then $G$ is hamiltonian.*

Geng-Hua Fan [4] has shown:

**Theorem 1.3.** *Let $G$ be a 2-connected graph on $n \geqslant 3$ vertices. If $G$ satisfies*

$$\mathrm{P}(n): \quad \mathrm{d}(x,y) = 2 \Rightarrow \max\{\mathrm{d}(x), \mathrm{d}(y)\} \geqslant \frac{n}{2}$$

*for each pair of vertices $x$ and $y$ in $G$, then $G$ is hamiltonian.*

The condition for degree sum in Theorem 1.2 is called *an Ore condition* or *an Ore type condition for graph $G$* and the condition $\mathrm{P}(k)$ is called *a Fan condition* or *a Fan type condition for graph $G$*.

Later, many *Fan type theorems* and *Ore type theorems* are shown.

Now we shall present Las Vergnas [8] condition $\mathcal{L}_{n,s}$.

**Definition 1.6.** *Let $G$ be graph on $n \geqslant 2$ vertices and let $s$ be an integer such that $0 \leqslant s \leqslant n$. $G$ satisfies Las Vergnas condition $\mathcal{L}_{n,s}$ if there is an arrangement $x_1, \ldots, x_n$ of vertices of $G$ such that for all $i, j$ if*

$$1 \leqslant i < j \leqslant n\,,\ i + j \geqslant n - s\,,\ x_i x_j \notin \mathrm{E}(G)\,,$$
$$\mathrm{d}(x_i) \leqslant i + s \text{ and } \mathrm{d}(x_j) \leqslant j + s - 1$$

*then $\mathrm{d}(x_i) + \mathrm{d}(x_j) \geqslant n + s$.*

Las Vergnas [8] proved the following theorem:

**Theorem 1.4.** *Let $G$ be a graph on $n \geqslant 3$ vertices and let $0 \leqslant s \leqslant n-1$. If $G$ satisfies $\mathcal{L}_{n,s}$ then $G$ is $s$-edge hamiltonian.*

Note that condition $\mathcal{L}_{n,s}$ is weaker than Ore condition.

Later Skupień and Wojda proved that the condition $\mathcal{L}_{n,s}$ is sufficient for a graph to have a stronger property (for details see [10]). Wojda [11] proved the following Ore type theorem:

**Theorem 1.5.** *Let $G$ be a graph on $n \geqslant 3$ vertices, such that for every pair of nonadjacent vertices $x$ and $y$*

$$\mathrm{d}(x) + \mathrm{d}(y) > \frac{4n-4}{3}\,.$$

*Then every matching of $G$ lies in a hamiltonian cycle.*

In 1996, G. Gancarzewicz and A. P. Wojda proved the following Fan type theorem:

**Theorem 1.6.** *Let $G$ be a 3-connected graph of order $n \geqslant 3$ and let $M$ be a $k$-matching in $G$. If $G$ satisfies* $\mathrm{P}(n+k)$ :

$$\mathrm{d}(x,y) = 2 \;\Rightarrow\; \max\{\mathrm{d}(x), \mathrm{d}(y)\} \geqslant \frac{n+k}{2}$$

*for each pair of vertices $x$ and $y$ in $G$, then $M$ lies in a hamiltonian cycle of $G$ or $G$ has a minimal odd $M$-edge cut-set.*

In this paper we find a Fan type condition under which every path of length $k$ in a graph $G$ lies in a hamiltonian cycle.

For notation and terminology not defined above a good reference should be [3].

## 2. Result

**Theorem 2.1.** *Let $G$ be a graph on $n \geqslant 3$ vertices and let $S$ be a path of length $k$ in $G$. If the graph $G$ is $l$-connected, where $l = \min\{k+2\,,\; n-1\}$ and satisfies* $\mathrm{P}(n+k)$ :

$$\mathrm{d}(x,y) = 2 \;\Rightarrow\; \max\{\mathrm{d}(x), \mathrm{d}(y)\} \geqslant \frac{n+k}{2} \tag{2.1}$$

*for each pair of vertices $x$ and $y \in \mathrm{V}(G)$, then $S$ lies in a hamiltonian cycle of $G$.*

Note that $1 \leqslant k \leqslant n-1$ and since an $(n-1)$-connected graph of order $n$ is a complete graph $K_n$, which is obviously $k$-edge hamiltonian for any $k$ the result is interesting when $k < n-3$.

For $k = 1$, the path $S$ is a 1-matching and we have a special case of Theorem 1.6 (the graph is 3-connected, so in this case we can not have a minimal $S$-edge cut set).

Unfortunately, in Theorem 2.1 we can not decrease the connectivity of graph $G$. We can consider a vertex $x$ and a complete graph $K_m$ , $m \geqslant 3$ . In the complete graph $K_m$ we choose a path $S : s_1 \ldots s_{k+1}$ of length $k = m-2$ . There is only one vertex $y \in K_m$ not contained in $S$.

Let $G$ be a graph of order $n = m+1$ obtained from two complete graphs $K_1 = \{x\}$ and $K_m$ by adding edges $xs_i$ , for $i \in \{1, \ldots, k+1\}$ The path $S$ is a path of length $k$ contained in $G$ which is not contained in any hamiltonian cycle of the $(k+1)$-connected graph $G$, see Figure (1).

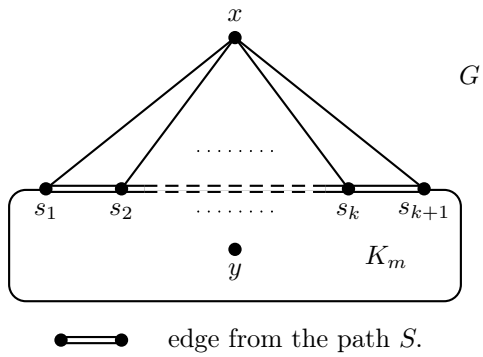


Fig. 1: A $(k+1)$-connected graph $G$ with no hamiltonian cycle through the path $S$ : $s_1 \dots s_{k+1}$.

Note that we can replace the vertex $x$ by a complete graph $K_\ell$ , $\ell \geqslant k+1$ . Let $\{x_1, \dots, x_{k+1}\} \subset \mathrm{V}(K_\ell)$ and let $G$ be a graph of order $n = m + l$ obtained from two complete graphs $K_\ell$ and $K_m$ by adding edges $x_i s_i$ , for $i \in \{1, \dots, k+1\}$ The path $S$ is a path of length $k$ contained in $G$ which is not contained in any hamiltonian cycle of the $(k+1)$-connected graph $G$, see Figure (2).

## 3. Proof

**Proof of Theorem 2.1:**

Take $G$ and $S$ as in the assumptions of Theorem 2.1.

Consider the nonempty set

$$A = \{x \in V(G) : \mathrm{d}_x(G) \geqslant \frac{n+k}{2}\} .$$

Note that if $x$ and $y$ are nonadjacent vertices of $A$, then the graph obtained from $G$ by the addition of the edge $xy$ also satisfies the assumptions of the theorem. Therefore, and by Theorem 1.1 we may assume that:

$$xy \in E(G) \quad \text{for any} \quad x, y \in A \quad \text{and } x \neq y . \tag{3.1}$$

By (3.1), $A$ induces a complete subgraph $G(A)$ of the graph $G$.

In fact, since the property of being $k$-edge-hamiltonian is $(n+k)$-stable, we can replace $G$ with its $(n+k)$-closure.

Let $G_A$ be a graph obtained from $G$ by deletion of vertices of the graph $G(A)$ (i.e. vertices from the set $A$).

Now take $D$, a connected component of the graph $G_A$.

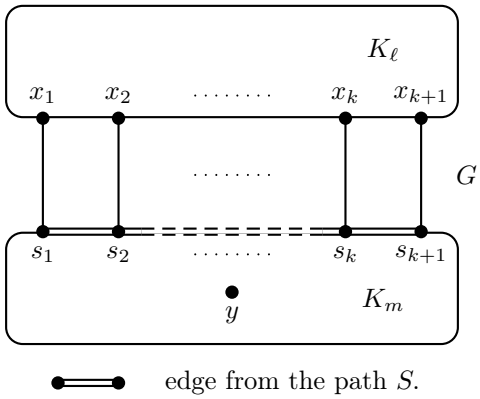


Fig. 2: A $(k+1)$-connected graph $G$ with no hamiltonian cycle through the path $S$ : $s_1 \dots s_{k+1}$.

Suppose that there exist two nonadjacent vertices in $D$. Since $D$ is connected, we have two vertices $x$ and $y$ in $D$ such that $\mathrm{d}_G(x,y) = 2$ and by the assumption that $G$ satisfies $\mathrm{P}(n+k)$, we have $x \in A$ or $y \in A$, which is a contradiction.

We have proved that every component of $G_A$ is a complete graph $K_\iota$ , $\iota \in I$ , joined with $G(A)$ by at least $k+2$ edges.

**Claim 3.1.** *If $K_{\iota_0}, K_{\iota_1} \in \{K_\iota\}_{\iota \in I}$ are such that $\iota_0 \neq \iota_1$, then:*

$$N(K_{\iota_0}) \cap N(K_{\iota_1}) = \emptyset. \tag{3.2}$$

**Proof of Claim 3.1:**

Suppose that $N(K_{\iota_0}) \cap N(K_{\iota_1}) \neq \emptyset$. Then we have a vertex $y \in K_{\iota_0}$ and a vertex $y' \in K_{\iota_1}$ such that $\mathrm{d}_G(y,y') = 2$ and by $\mathrm{P}(n+k)$ either $y \in A$ or $y' \in A$ . This contradicts the fact that $K_{\iota_0}$ and $K_{\iota_1}$ are two connected components of $G_A$ .

□

We have shown that the graph $G$ consists of a complete graph $G(A)$ and of a family of complete components $\{K_\iota\}_{\iota \in I}$ , of $G_A$ , which do not have common neighbors in $G(A)$.

Since $G$ is $(k+2)$-connected, we have the following:

**Claim 3.2.** *Every component $\{K_\iota\}_{\iota \in I}$ , is joined with $G(A)$ by at least three edges such that end vertices of these edges are not internal vertices of the path $S$.*

We label vertices of path $S : s_1 s_2 \dots s_k s_{k+1}$ .

Graph $G$ consists of complete graph $G(A)$ and disjoined complete graphs $\{K_\iota\}_{\iota \in I}$ , joined with $G(A)$ by at least three edges such that end vertices of these edges are not internal vertices of path $S$.

Firstly we consider the case when path $S$ is contained in one complete graph (i.e. $G(A)$ or one graph $K_{\iota_0} \in \{K_\iota\}_{\iota \in I}$). In this case, by Claim3.2 we have a hamiltonian cycle through $S$.

Now we assume that $S$ is not contained in the complete graph $G(A)$ or one graph $K_{\iota_0} \in \{K_\iota\}_{\iota \in I}$ and we can now define a cycle $C \subset G$ containing the path $S$ and all vertices of $G(A)$.

We shall consider four cases:

1. Both end vertices of $S$ are in $G(A)$ i.e. $s_1$ , $s_{k+1} \in G(A)$.

2. Both end vertices of $S$ are in the same component $K_\iota$ of $G_A$ i.e. $s_1$ , $s_{k+1} \in K_\iota$.

3. End vertices of $S$ are in different components of $G_A$ i.e. $s_1 \in K_{\iota_1}$ , $s_{k+1} \in K_{\iota_2}$ , $K_{\iota_1}$ , $K_{\iota_2} \in \{K_\iota\}_{\iota \in I}$ are such that $\iota_1 \neq \iota_2$.

4. One end vertex of $S$ is in $G(A)$ and the other end vertex is in a component $K_\iota$ of $G_A$. In this case, we can assume without loss of generality that $s_1 \in G(A)$ and $s_{k+1} \in K_\iota$.

If $C \subset G$ is a cycle in $G$, then by $G_{V \setminus C}$ we denote a graph obtained from $G$ by deletion of vertices of cycle $C$.

**Case 1: Both end vertices of $S$ are in $G(A)$ i.e. $s_1$ , $s_{k+1} \in G(A)$.**

Note that even in this case, path $S$ may pass through some components $K_\iota$ creating a kind of ears of the complete graph $G(A)$, on every incident graph $K_\iota$. We can find an example of such ears on Figure (3).

Since $G(A)$ is a complete graph we have a cycle $C$ containing the path $S$ and all vertices of $G(A)$ performing the following conditions:

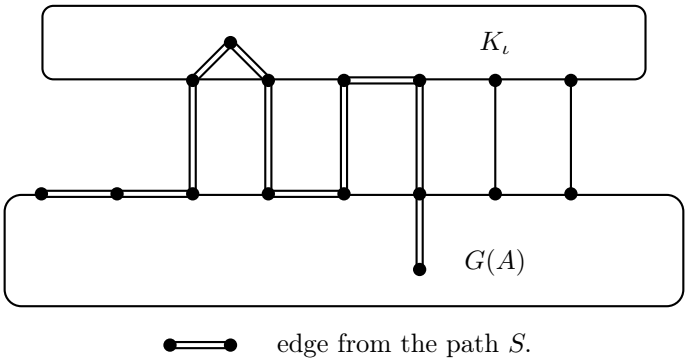


Fig. 3: Example of ears of the graph $G(A)$.

- $C$ contains all edges of $\mathrm{E}(S) \cap \mathrm{E}(G(A))$ and all vertices of $A$. (3.3)
- If $K_{\iota_0}$ and $K_{\iota_1}$ are two different components of $G_{V \setminus C}$ then (3.4)
  $N(K_{\iota_0}) \cap N(K_{\iota_1}) = \emptyset$.
- Let $x \notin V(C)$, $y \in V(C)$ and $xy \in E(G)$ then: (3.5)
  if $y$ is not an internal vertex of $S$, then $y \in A$,
  if $y^-$ is not an internal vertex of $S$, then $y^- \in A$,
  if $y^+$ is not an internal vertex of $S$, then $y^+ \in A$.

Properties (3.3 — 3.5) will allow us to extend $C$ to a hamiltonian cycle. Note that this cycle $C$ may not be contained in $G(A)$.

**Subcase 1.1: Extending the cycle $C$ through components $K_\iota$ incident with ears**

Since graph $G$ is $(k+2)$-connected, we have at least $k+2$ edges joining $K_\iota$ with $G(A)$, so at least one of these edges say $uc_i$, $u \in \mathrm{V}(K_\iota)$, $c_i \in \mathrm{V}(G(A)) \setminus S$, is not incident with $S$. If a component $K_\iota$ is incident with an ear, at least one interior vertex of path $S$ is contained in this ear, and we have an additional edge say $u'c_j$, $u' \in \mathrm{V}(K_\iota)$, $c_j \in \mathrm{V}(G(A)) \setminus S$, not incident with $S$ joining $K_\iota$ with $G(A)$. Using these two edges $uc_i$ and $u'c_j$, we can extend the cycle $C$ through the remaining vertices of $K_\iota$.

Without loss of generality, we can assume that on cycle $C$, the vertices are ordered in the following way: $s_1 \dots s_{k+1} c_{k+2} \dots c_i \dots c_j \dots s_1$.

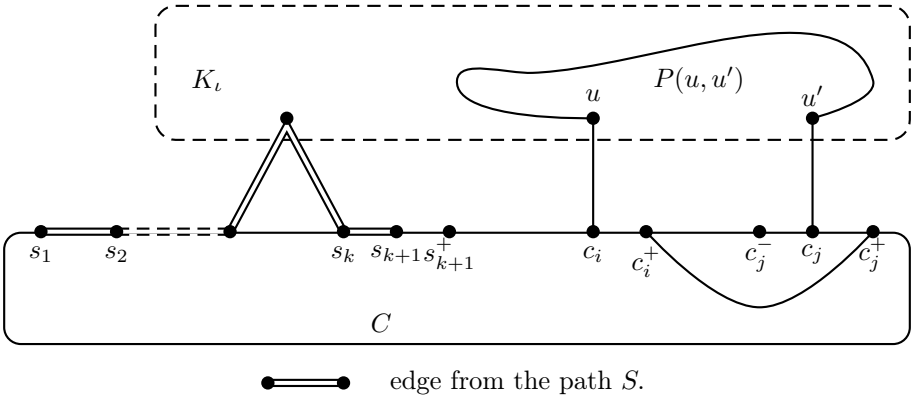


Fig. 4: Extension of the cycle $C$ through component $K_\iota$ incident with an ear.

We can replace cycle $C$, by the following cycle $C'$

$$C': \quad s_1 \dots s_{k+1} s^+_{k+1} \dots c_i P(u,u') c_j c^-_j \dots c^+_i c^+_j \dots s_1 \,,$$

where $P(u,u') \subset K_\iota$ is a path joining $u$ with $u'$ containing all vertices of $K_\iota \setminus S$. Note that this cycle $C'$ satisfies conditions (3.3 — 3.5). We can find an example of the cycle $C'$ on Figure 4.

**Case 2: Both end vertices of $S$ are in the same component $K_\iota$ of $G_A$ i.e.** $s_1\,,\ s_{k+1} \in K_\iota$.

Since the graph $G$ is $(k+2)$-connected, we have at least $k+2$ edges joining $K_\iota$ with $G(A)$. In this case, at least two edges from the path $S : s_i s_{i+1}$ and $s_j s_{j+1}$ are joining $K_\iota$ with $G(A)$, so at least two independent edges $uv$, $u'v'$, $u,\ u' \in \mathrm{V}(K_\iota)$, $v$, $v' \in \mathrm{V}(G(A))$, not incident with $S$ joining $K_\iota$ with $G(A)$.

Consider the following path:

$$P: \quad v u s_1 \dots s_k P(s_{k+1}, u') v' \,,$$

where $P(s_{k+1}, u') \subset K_\iota$ is a path joining $s_{k+1}$ with $u'$ containing all vertices of $K_\iota \setminus \{\mathrm{V}(S) \cup \{u\}\}$. We can find an example of the path $P$ on Figure 5.

The graph $G(A)$ is complete, so we can extend $P$ to a cycle $C$ containing all vertices of $G(A)$ and satisfying (3.3 — 3.5).

Note that as in Case 1 the path $S$ may pass through some components $K_i$ creating the kind of ears of the complete graph $G(A)$. Using the same argument as in the

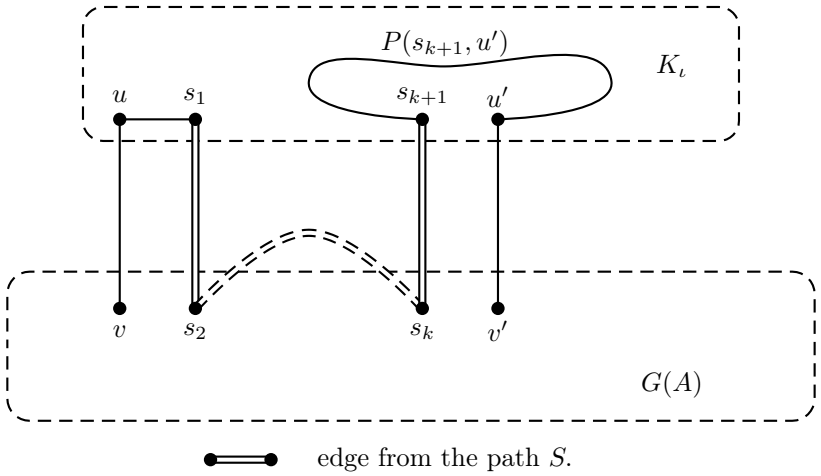


Fig. 5: Path $P$ containing $S$ with both end vertices in $G(A)$.

Subcase 1.1 we can extend the cycle $C$ through components $K_i$ incident with ears preserving the properties (3.3 — 3.5).

**Case 3: End vertices of $S$ are in different components of $G_A$ i.e. $s_1 \in K_{\iota_1}$, $s_{k+1} \in K_{\iota_2}$, $K_{\iota_1}$, $K_{\iota_2} \in \{K_\iota\}_{\iota \in I}$ are such that $\iota_1 \neq \iota_2$.**

Again, since graph $G$ is $(k+2)$-connected we have at least $k+2$ edges joining every component $K_i$ with $G(A)$. In this case, for $i = \iota_1$ and $i = \iota_2$ at least one edge from the path $S$ is joining $K_i$ with $G(A)$, so we have at least two independent edges $uv$, $u \in \mathrm{V}(K_{\iota_1})$, $v \in \mathrm{V}(G(A))$, $u'v'$, $u' \in \mathrm{V}(K_{\iota_2})$, $v' \in \mathrm{V}(G(A))$, not incident with $S$ joining respectively $K_{\iota_1}$ and $K_{\iota_2}$ with $G(A)$.

Consider the following path:

$$P: \quad vP_1(u, s_1)s_2 \dots s_k P_2(s_{k+1}, u')v',$$

where $P_1(u, s_1) \subset K_{\iota_1}$ is a path joining $u$ with $s_1$ containing all vertices of $K_{\iota_1} \setminus \{\mathrm{V}(S) \cup \{u\}\}$ and $P_2(s_{k+1}, u') \subset K_{\iota_2}$ is a path joining $s_{k+1}$ with $u'$ containing all vertices of $K_{\iota_2} \setminus \{\mathrm{V}(S) \cup \{u'\}\}$. See Figure 6.

The graph $G(A)$ is complete so we can extend $P$ to a cycle $C$ containing all vertices of $G(A)$ and satisfying (3.3 — 3.5).

Note that as in Case 1 path $S$ may pass through several components $K_i$ creating the kind of ears of the complete graph $G(A)$. Using the same argument as in Subcase

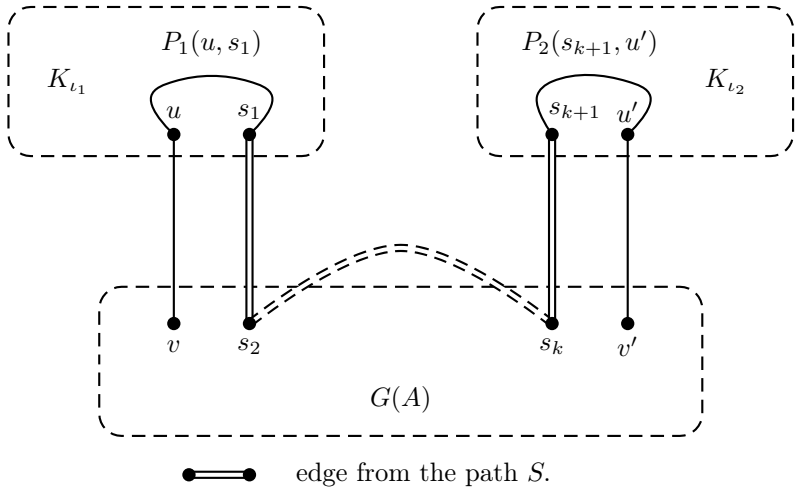


Fig. 6: Path $P$ containing $S$ with both end vertices in $G(A)$.

1.1, we can extend cycle $C$ through components $K_i$ incident with ears preserving the properties (3.3 — 3.5).

**Case 4: One end vertex of $s$ is in $G(A)$ and the other end vertex is in a component $K_\iota$ of $G_A$. In this case, we can assume without loss of generality, that $s_1 \in G(A)$ and $s_{k+1} \in K_\iota$.**

Since graph $G$ is $(k+2)$-connected, we have at least $k+2$ edges joining the component $K_\iota$ with $G(A)$. In this case, at least one edge from the path $S$ is joining $K_\iota$ with $G(A)$, so we have at least one edge $uv$, $u \in \mathrm{V}(K_\iota)$, $v \in \mathrm{V}(G(A))$, not incident with $S$ joining $K_\iota$ with $G(A)$.

Consider the following path:

$$P: \quad s_1 s_2 \dots s_k P(s_{k+1}, u) v\,,$$

where $P(s_{k+1}, u) \subset K_{\iota_0}$ is a path joining $s_{k+1}$ with $u$ containing all vertices of $K_\iota \setminus \{\mathrm{V}(S) \cup \{u\}\}$, see Figure 7.

Both $v$ and $s_1$ are in $G(A)$ and the graph $G(A)$ is complete, so we can extend $P$ to a cycle $C$ containing all vertices of $G(A)$ and satisfying (3.3 — 3.5).

Note that as in Case 1, path $S$ may pass through several components $K_i$ creating the kind of ears of the complete graph $G(A)$. Using the same argument as in Subcase

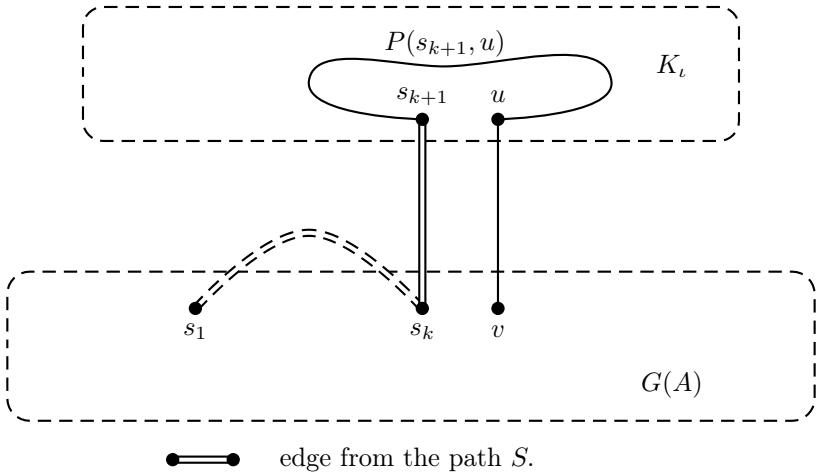


Fig. 7: Path $P$ containing $S$ with both end vertices in $G(A)$.

1.1, we can extend the cycle $C$ through components $K_i$ incident with ears preserving the properties (3.3 — 3.5).

In all cases we have defined a cycle $C$ containing $S$ and now we shall extend this cycle to a hamiltonian cycle.

**Extending the cycle $C$ to a hamiltonian cycle**

We have already a cycle $C$ satisfying conditions (3.3 — 3.5) and containing the path $S$, all vertices of $G(A)$, all vertices from components $K_\iota$ containing vertices of the path $S$.

Consider component $K_\iota$ not included in cycle $C$. This component does not contain any edge from $S$ and since the graph $G$ is $(k+2)$-connected we have at least $k+2$ edges joining $K_\iota$ with $G(A)$, so at least one of these edges say $uc_i$ , $u \in \mathrm{V}(K_\iota)$ , $c_i \in \mathrm{V}(G(A))$ , is not incident with $S$ and at least one edge say $u'c_j$ , $u' \in \mathrm{V}(K_\iota)$ , $c_j \in \mathrm{V}(G(A))\setminus$ , not incident with internal vertices of $S$ , joining $K_\iota$ with $G(A)$. In the worst case $c_j = s_1$ or $c_j = s_{k+1}$.

Using these two edges $uc_i$ and $u'c_j$ , we can extend cycle $C$ through the remaining vertices of $K_\iota$.

We consider the case $c_j = s_1$ and without loss of generality we can, assume that on cycle $C$, the vertices are ordered in the following way:
$s_1 \dots s_{k+1} c_{k+2} \dots c_i \dots s_1$ .

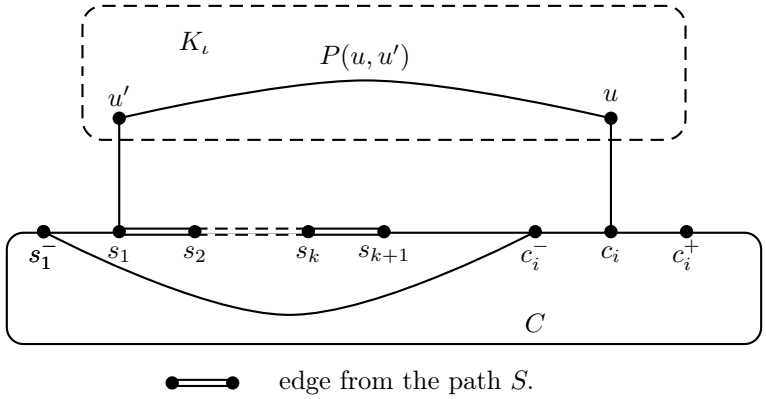


Fig. 8: Extension of cycle $C$ through a component $K_\iota$ not incident with $S$.

Note that since $uc_i$ is not incident with $S$ we have $c_i^- c_i\,,\ c_i c_i^+ \notin \mathrm{E}(S)$ and we can replace cycle $C$ with the following cycle $C'$

$$C': \quad u's_1 \dots s_{k+1}s_{k+1}^+ \dots c_i^- s_1^- \dots c_i^+ c_i P(u,u')\,,$$

where $P(u,u') \subset K_\iota$ is a path joining $u$ with $u'$ containing all vertices of $K_\iota$, see Figure 8.

Note that this cycle $C'$ satisfies conditions (3.3 — 3.5), $\mathrm{V}(C) \subset \mathrm{V}(C')$ and $\mathrm{E}(S) \subset \mathrm{E}(C')$.

The case $c_j = s_{k+1}$ is similar.

Applying this argument for all other components $K_\iota$ we can extend $C$ to a hamiltonian cycle containing the path $S$ and the proof is complete.

□

## References

[1] A. Benhocine and A.P. Wojda, *The Geng-Hua Fan conditions for pancyclic or hamiltonian-connected graphs,* J. Combin. Theory Ser. B 42, 1985, 167—180.

[2] J.A. Bondy and V. Chvátal, *A method in graph theory,* Discrete Math. 15, 1976, 111—135.

[3] J.A. Bondy and U.S.R. Murty, *Graph theory with applications,* The Macmillan Press LTD London 1976.

[4] G. Fan, *New sufficient conditions for cycles in graphs,* J. Combin. Theory Ser. B 37, 1984, 221—227.

[5] G. Gancarzewicz and A.P. Wojda, *Graphs with every k-matching in a hamiltonian cycle,* Discrete Math. 213, 2000, 141 — 151.

[6] R. Hăggkvist, *On F-hamiltonian graphs* in Graph Theory and Related Topics, ed. J.A. Bondy and U.S.R. Murty, Academic Press N.Y. 1979, 219—231.

[7] H.V. Kronk, *Variations of a theorem of Pósa* in The Many Facets of Graph Theory, ed. G. Chartrand and S.F. Kapoor, Lect. Notes Math. 110, Springer Verlag 1969, 193—197.

[8] M. Las Vergnas, *Sur une propriété des arbres maximaux dans un graphe,* C. R. Axad. Sci. Paris, Sér. A, 272, 1971, 1297—1300.

[9] O. Ore, *Note on hamiltonian circuits,* Amer. Math. Monthly 67, 1960, 55.

[10] Z. Skupień and A.P. Wojda, *On highly hamiltonian graphs,* Bull. Acad. Polon. Sci., Sér. Sci. Math. Astronom. Phys. 22, 1974, 463—471.

[11] A.P. Wojda, *Hamiltonian cycles through matchings,* Demonstratio Mathematica XXI 2, 1983, 547—553.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

MARIUSZ JUŻYNIEC*

# THE EXISTENCE OF A WEAK SOLUTION OF THE SEMILINEAR FIRST-ORDER DIFFERENTIAL EQUATION IN A BANACH SPACE

# ISTNIENIE SŁABEGO ROZWIĄZANIA SEMILINIOWEGO RÓWNANIA RÓŻNICZKOWEGO PIERWSZEGO RZĘDU W PRZESTRZENI BANACHA

Abstract

This paper is devoted to the investigation of the existence and uniqueness of a suitably defined weak solution of the abstract semilinear value problem $\dot{u}(t) = Au(t) + f(t, u(t)),\ u(0) = x$ with $x \in X$, where $X$ is a Banach space. We are concerned with two types of solutions: weak and mild. Under the assumption that $A$ is the generator of a strongly continuous semigroup of linear, bounded operators, we also establish sufficient conditions such that if $u$ is a weak (mild) solution of the initial value problem, then $u$ is a mild (weak) solution of that problem.

*Keywords: operator, semigroup, weak solution*

Streszczenie

Celem pracy jest przedstawienie twierdzenia o jednoznaczności i istnieniu słabego rozwiązania abstrakcyjnego semiliniowego równania różniczkowego $\dot{u}(t) = Au(t) + f(t, u(t)),\ u(0) = x$, gdzie $x \in X$, w przestrzeni Banacha $X$. W pracy rozważane są dwa typy rozwiązań: *weak* oraz *mild*. Przy założeniu, że operator $A$ jest generatorem silnie ciągłej półgrupy operatorów liniowych i ograniczonych, podane zostały również warunki wystarczające na to aby rozwiązanie *weak (mild)* było rozwiązaniem *mild (weak)* tego zagadnienia.

*Słowa kluczowe: operator, półgrupa, słabe rozwiązanie*

*Institute of Mathematics, Cracow University of Technology, Poland; juzyniec@pk.edu.pl

## 1. Introduction

For a real or complex Banach space $X$, $X^*$ will denote its dual space. Let $\langle\cdot,\cdot\rangle$ be the duality pairing between $X$ and its dual space $X^*$. For an operator $A$, $D(A)$ and $A^*$ will denote its domain and the adjoint, respectively. We consider the abstract first-order initial value problem

$$\frac{d}{dt}u(t) = Au(t) + f(t, u(t)) \quad \text{for} \quad t \in (0, T], \tag{1.1}$$

$$u(0) = x, \tag{1.2}$$

where $A$ is a densely defined, closed linear operator on the Banach space $X$, $x \in X$ and $f : [0, T] \times X \to X$.

Definition 1. *A function $u \in C([0, T]; X)$ is a weak solution of (1) on $[0, T]$ if for every $v \in D(A^*)$, the function $[0, T] \ni t \to \langle u(t), v\rangle$ is absolutely continuous on $[0, T]$ and*

$$\frac{d}{dt}\langle u(t), v\rangle = \langle u(t), A^*v\rangle + \langle f(t, u(t)), v\rangle \text{ a.e. on } [0, T]. \tag{1.3}$$

## 2. Preliminaries

Let $A$ be a densely defined linear operator on a real or complex Banach space $X$, let $T > 0$ and let $g \in L^1(0, T; X)$. It is well known (see [1]) that

**Theorem 2.** *If $A$ is the generator of a strongly continuous semigroup of bounded linear operators $\{S(t)\}_{t\geq 0}$ on $X$, and if $x \in X$, then the first order linear equation*

$$\dot{w}(t) = Aw(t) + g(t), \quad t \in (0, T], \tag{2.1}$$

*has a unique weak solution (see Definition 3) satisfying $w(0) = x$, and in this case, $w$ is given by*

$$w(t) = S(t)x + \int_0^t S(t - s)g(s)ds, \quad t \in [0, T]. \tag{2.2}$$

Definition 3. *A function $w \in C([0, T]; X)$ is a weak solution of (2.1) on $[0, T]$ if for every $v \in D(A^*)$, the function $\langle w(t), v\rangle$ is absolutely continuous on $[0, T]$ and*

$$\frac{d}{dt}\langle w(t), v\rangle = \langle w(t), A^*v\rangle + \langle g(t), v\rangle \text{ a.e. on } [0, T]. \tag{2.3}$$

When $x \in X$ is arbitrary, then unless $\{S(t)\}_{t\geq 0}$ and $f$ have special properties, $w$ given by (2.2) will not, in general, belong to $D(A)$ for $t \in (0, T]$, so that (2.1) does not even make sense.

## 3. Existence and uniqueness of a weak solution of the problem (1)—(2)

We start with the following

**Theorem 4.** *Let $A$ be the infinitesimal generator of a $C_0$ semigroup $\{S(t)\}_{t\geq 0}$ of bounded linear operators on $X$, $u \in C([0,T];\ X)$ and $f(\cdot, u(\cdot)) \in L^1(0,T;\ X)$. If $u$ is a weak solution of the equation (1) and $u(0) = x$, then $u$ is a solution of the integral equation*

$$u(t) = S(t)x + \int_0^t S(t-s)f(s,u(s))ds, \quad t \in [0,T]. \tag{3.1}$$

*A continuous solution $u$ of the integral equation (3.1) will be called a mild solution if the initial value problem (1)–(2).*

**Proof.** Let $u$ be a weak solution of (1) satisfying $u(0) = x$. This implies that for any $v \in D(A^*)$

$$\frac{d}{dt}\langle u(t), v\rangle = \langle u(t), A^*v\rangle + \langle f(t,u(t)), v\rangle \text{ a.e. on } [0,T]. \tag{3.2}$$

Let us put $g(t) := f(t,u(t))$ and $w(t) := S(t)x + \int_0^t S(t-s)g(s)ds$ for $t \in [0,T]$. Cleary, by Theorem 2, $w$ is a unique weak solution of the problem

$$\begin{cases} \dot{w}(t) = Aw(t) + g(t), \ \ t \in (0,T], \\ w(0) = x. \end{cases} \tag{3.3}$$

By Definition 3,

$$\frac{d}{dt}\langle w(t), v\rangle = \langle w(t), A^*v\rangle + \langle g(t), v\rangle \text{ a.e. on } [0,T]. \tag{3.4}$$

Hence, by (3.2), the function $u$ satisfies (3.4). By the uniqueness of the weak solution of the initial value problem (3.3)

$$u = w,$$

so

$$u(t) = w(t) = S(t)x + \int_0^t S(t-s)g(s)ds = S(t)x + \int_0^t S(t-s)f(s,u(s))ds$$

The proof of Theorem 4 is complete. □

The integral equation (3.1) does not necessarily admit a solution of any kind. However, if it has a continuous solution, then that function is a weak solution of the problem (1)–(2).

**Theorem 5.** *Let $A$ be the infinitesimal generator of a $C_0$ semigroup $\{S(t)\}_{t\geq 0}$ of bounded linear operators on $X$, $u \in C([0,T];\ X)$ and $f(\cdot, u(\cdot)) \in L^1(0,T;\ X)$. If $u$ is a solution of the integral equation (3.1), then $u$ is a weak solution of the equation (1).*

**Proof.** By Theorem 2 the initial value problem

$$\begin{cases} \dot{w}(t) = Aw(t) + f(t, u(t)), \ \ t \in (0,T], \\ w(0) = x \end{cases} \tag{3.5}$$

has exactly one weak solution given by $w(t) := S(t)x + \int_0^t S(t-s)f(s,u(s))ds$ for $t \in [0,T]$. By the assumption

$$u(t) = S(t)x + \int_0^t S(t-s)f(s,u(s))ds$$

for $t \in [0,T]$, so $w = u$ and $u$ is the weak solution of (3.5). This completes the proof. □

The main result of this paper is the following theorem

**Theorem 6.** *Let $f : [0,T] \times X \to X$ be continuous in $t$ on $[0,T]$ and uniformly Lipschitz continuous on $X$. If $A$ is the infinitesimal generator of a $C_0$ semigroup $\{S(t)\}_{t\geq 0}$ of bounded linear operators on $X$, then there exists for each $x \in X$ a unique weak solution $u$ of (1) satisfying $u(0) = x$.*

**Proof.** By Theorem 6.1.2 [4] (page 184) (see [2], p. 77, [3], p. 87) the integral equation (3.1) has a unique solution $u \in C([0,T];\ X)$. From this, by Theorem 5, $u$ is a weak solution of the equation (1) and $u(0) = x$. The uniqueness of $u$ is a consequence of Theorem 4.

The proof is complete. □

## References


[1] Ball J.M., *Strongly continuous semigroups, weak solutions, and the variation of constant formula*, Proc. Amer. Math. Soc., 1977, 370–373.

[2] Hundertmark D., Meyries M., Machinek L., Schnaubelt R., *Operator Semigroups and Dispersive Equations*, 16th Internet Seminar on Evolution Equations, 2013.

[3] Goldstein J., *Semigroups of Linear Operators and Applications*, Oxford U. Press, New York 1985.

[4] Pazy A., *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer–Verlag, 1983.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

JAN KUCWAJ*

# ADAPTIVE UNSTRUCTURED SOLUTION TO THE PROBLEM OF ELASTIC-PLASTIC HARDENING TWIST OF PRISMATIC BARS

# ADAPTACYJNE NIESTRUKTURALNE ROZWIĄZANIE ZAGADNIENIA SKRĘCANIA PRĘTA PRYZMATYCZNEGO W ZAKRESIE SPRĘŻYSTO-PLASTYCZNYM ZE WZMOCNIENIEM

Abstract

This paper presents the appliction of a remeshing algorithm to solution of elastic-plastic torsion of bars with isotropic strain hardening. The remeshing algorithm uses a grid generator with mesh size function [7]. The method of grid generation is based on a coupling of the advancing front method and the Delaunay triangulation. The optimal mesh size for the posed problem is obtained iteratively. For the consequtive steps of the adaptation algorithm error indicators at nodes and in elements are used for mesh size modification. The discretized system of nonlinear algebraic equations is solved by the application of the Newton-Raphson method.

*Keywords: torsion, plasticity, elasticity, grid adaptation, Delaunay triangulation, nonlinearity*

Streszczenie

Praca przedstawia zastosowanie algorytmu typu *remeshing* do rozwiązania zagadnienia sprężysto-plastycznego skręcania prętów pryzmatycznych ze wzmocnieniem. Algorytm typu *remeshing* wykorzystuje generator siatek uwzględniający funkcję rozmiaru siatki [7]. Metoda generowania siatek oparta jest na połączeniu metody postępującego frontu z triangulacją Delaunaya. Optymalny rozmiar siatki dla postawionego problemu otrzymany jest iteracyjnie. W kolejnych krokach adaptacji indykatory błędów w węzłach i elementach są wykorzystane do modyfikacji rozmiaru siatki. Zdyskretyzowany układ nieliniowych równań algebraicznych jest rozwiązywany poprzez zastosowanie metody Newtona-Raphsona.

*Słowa kluczowe: skręcanie, plastyczność, sprężystość, generowanie siatek, adaptacja, triangulacja Delaunaya, nieliniowość*

*Institute of Computer Science, Cracow University of Technology; jkucwaj@usk.pk.edu.pl

## 1. Introduction

The purpose of this paper is to apply the adaptive remeshing [7, 8] to the elastic-plastic torsion of bars with isotropic strain hardening [5]. The problem is posed in a form of the search for the stationary point of the functional defined on function space of infinite dimension satisfying homogenous boundary conditions. In the case of twisted bars, this means that their length is great enough. For the sake of a numerical solution the infinite space is approximated by finite dimensional space spanned by a given set of basis functions [2, 11, 12]. The approximated solution to the problem is equal to a linear combination of the basis functions. The coeffiecients of the linear combination are found from the set of nonlinear algebraic equations. The latter are obtained from the stationarity conditions. These equations are then solved using the Newton-Raphson method.

The grid generation code GRADMESH [6, 7] developed in [7] can be used to generate the FE model. The generator takes into account the mesh size function defined in the whole domain. The mesh size function is iteratively modified in order to give a uniform error distribution. In practice it is rather difficult to obtain uniform error distribution. As a matter of fact the values of mesh size function are defined only at the nodes of the current mesh. Values of the mesh size function are calculated by the linear interpolation in every element. Therefore the modification of the mesh size function leads to the modification of their values at the nodes in the current mesh. These values are multiplied by a value from the interval $[0, 1]$. The values are smaller for nodes with a greater value of error indicator than for nodes with a smaller value of the indicator. Having modified values at the nodes of the current mesh, the new mesh size function is defined in the whole domain. The consecutive mesh is generated next. For the sake of the consecutive mesh generation the whole information about the previous mesh must be stored.

In the current paper, two methods of error evaluation are applied. The first adopts standard methods used in [4, 10, 14] which give values of error estimates in elements. The error estimation at nodes is more involved. A given node is shared by adjacent elements, each in turn, carries a different error. One needs to accommodate these different errors when assigning the final value to the node. The weighted average is used here. The precise equation is given later (eq. 6.5) but the process is as follows. The partial error in a given element is multiplied by its area and is divided by the area of all adjacent elements. Finally, the node error is the sum of partial errors (coming from all elements surrounding the node).

The error indicator at a given node can be evaluated in a different way, for example, by considering first derivatives with respect to all variables, these are taken at the node. In this approach the error indicator is formed summing squares all derivatives. The final value is obtained by taking the square root of the sum. Detailed expression is given later (eq. 6.4). Yet another approach which is frequently and widely known is given by equation [2].

Once the magnitude of the final error indicator at a given node is established, it serves as a size indicator for the new, remeshed model.

The above information is used for remeshing, i.e., for the generation of a new grid of nodes. The grid in the FE model utilizes the advancing front technique (AFT) [9] in conjunction with the Delaunay approach [6, 7, 9].

The adaptation method based on points enrichment [1, 10, 11] leads to unnecessary mesh enrichment or bad quality elements [15]. Remeshing on the other hand [8, 14] seems to be one of the best adaptation techniques, especially in the case of the elastic-plastic twisting of bars. It almost exactly densifies the mesh only at the yield points. In short, the method focuses on adaptation (remeshing) of the grid at places where it is necessary as dictated by local error indicators. It does not alter mesh size where it is not necessary. It is clear that this approach is effiecient and superior to other approaches.

## 2. Problem formulation

In this section, the elastic-plastic torsion of bars with isotropic strain hardening is formulated. According to [5] the problem leads to the search for the extremum of the following functional:

$$I(u) = \iint_{\Omega} [\int_0^T sg(s)ds - 2\omega u] d\Omega, \tag{2.1}$$

where T is the stress intensity

$$T = \sqrt{\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}, \qquad \tau_{13} = \frac{\partial u}{\partial y}, \qquad \tau_{23} = -\frac{\partial u}{\partial x}$$

.

The function $g$ defines the dependence between the effective stress and the effective strain: $T = g(\Gamma)\Gamma$ (Fig. 1), where $\Gamma = \sqrt{\epsilon_{ij}\epsilon_{ij}}$, $\epsilon_{ij}$ is the strain tensor and $\omega$ is the torsion angle (see Ref. [5]).

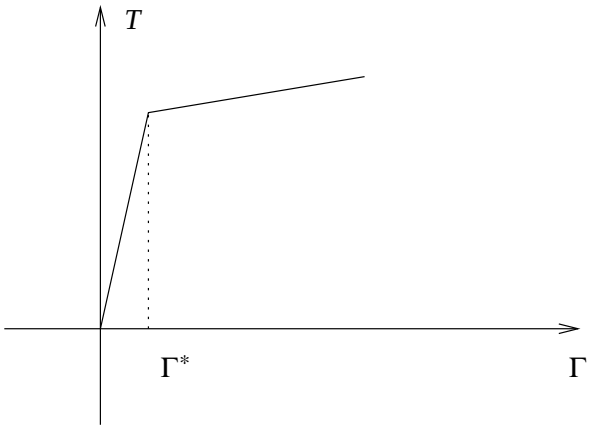


Fig. 1: The dependence between the strain intesity of strain, F and stress intesity ,T

After the substitution $s = \sqrt{r}$, the equation (2.1) is:

$$I(u) = \iint_{\Omega} \left[ \int_0^{T^2} g(\sqrt{r}) \frac{1}{2} dr - 2\omega u \right] d\Omega. \tag{2.2}$$

### 3. Discretization approach and tangent matrix

In this section, a class of problems leading to the search for a stationary point of a functional of the following form is addressed:

$$I(u) = \iint_{\Omega} F(u, u_x, u_y) d\Omega, \text{ where} \tag{3.1}$$

$\Omega$ - is a two-dimensional domain,

$F$ - is a double diffentiable function of several variables $F : \mathbb{R}^M \mapsto \mathbb{R}$.

Let us introduce finite element basis $\{U_i\}_{i=1}^N$ in the space $V^0$ of functions satisfying homogenous boundary conditions. Then, the discretization of the problem leads to the solution of the following system of nonlinear algebraic equations:

$$g_k(\lambda_1, \lambda_2, \ldots, \lambda_N) = \frac{\partial I(\sum_{i=1}^N \lambda_i U_i)}{\partial \lambda_k} = 0, \quad \text{for } k = 1, \ldots, N, \tag{3.2}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N$ are coefficients of the linear combination of the approximate solution.

The chain formula for the calculations of partial derivatives leads to:

$$g_k(\lambda_1, \lambda_2, \ldots, \lambda_N) = \tag{3.3}$$

$$= \frac{\partial}{\partial \lambda_k} \iint_\Omega F(x, y, \sum_{i=1}^{N} \lambda_i U_i, \sum_{i=1}^{N} \lambda_i U_{i,x}, \sum_{i=1}^{N} \lambda_i U_{i,y}) d\Omega = \tag{3.4}$$

$$= \iint_\Omega D_F^T \Psi_k d\Omega, \quad \text{and} \tag{3.5}$$

$$\frac{\partial g_k}{\partial \lambda_j} = \iint_\Omega \Psi_k^T D_{FF} \Psi_j d\Omega. \tag{3.6}$$

In the expressions (3.4), (3.6)

$$\mathbf{D}_F = \left[ \frac{\partial F}{\partial u}, \frac{\partial F}{\partial u_x}, \frac{\partial F}{\partial u_y} \right]^T, \tag{3.7}$$

$$\mathbf{D}_{FF} = \begin{bmatrix} \frac{\partial^2 F}{\partial u \partial u}, & \frac{\partial^2 F}{\partial u \partial u_x}, & \frac{\partial^2 F}{\partial u \partial u_y} \\ \frac{\partial^2 F}{\partial u_x \partial u}, & \frac{\partial^2 F}{\partial u_x \partial u_x}, & \frac{\partial^2 F}{\partial u_x \partial u_y} \\ \frac{\partial^2 F}{\partial u_y \partial u}, & \frac{\partial^2 F}{\partial u_y \partial u_x}, & \frac{\partial^2 F}{\partial u_y \partial u_y} \end{bmatrix}, \tag{3.8}$$

$$\text{and} \qquad \mathbf{U} = \begin{bmatrix} U_1, & \ldots, & U_N \\ U_{1,x}, & \ldots, & U_{N,x} \\ U_{1,y}, & \ldots, & U_{N,y} \end{bmatrix}, \tag{3.9}$$

The above formulas are valid for arbitrary basis functions (for example, polynomial) $\{U_1, U_2, \ldots, U_N\}$ defined in $\Omega$.

In the case of FEM one has the following equations:

$$g_j = \iint_\Omega \mathbf{D}_F^T \psi_j \, dx \, dy = \sum_{e=1}^{N_T} \iint_{T_e} \mathbf{D}_F^T \psi_j \, dx \, dy, \tag{3.10}$$

$$\frac{D(g_1, g_2, \ldots, g_N)}{D(\lambda_1, \lambda_2, \ldots, \lambda_N)} = \left[ \frac{\partial g_i}{\partial \lambda_j} \right] = \left[ \sum_{e=1}^{N_T} \iint_{T_e} \psi_i^T \mathbf{D}_{FF} \psi_j \, dx \, dy \right] \tag{3.11}$$

where $\psi_k$ is the *k-th* column of matrix $\mathbf{U}$.

Introduce the following vector:

$$\mathbf{G}(\mathbf{\Lambda}) = [g_1(\mathbf{\Lambda}), \ldots, g_N(\mathbf{\Lambda})]^T, \text{ and matrix} \tag{3.12}$$

$$\mathbf{J_G} = \left[ \begin{array}{c} \frac{\partial g_i}{\partial \lambda_j} \end{array} \right]. \tag{3.13}$$

Let the set $\{u_1^e, u_2^e, \ldots, u_{n_e}^e\}$ be a set of shape functions of the element with number e, for $e = 1, \ldots, N_T$. The matrix $\mathbf{u}_e$ is then defined as follows:

$$\mathbf{u}_e = \left[ \begin{array}{ccc} u_1, & \ldots, & u_{n_e} \\ u_{1,x}, & \ldots, & u_{n_e,x} \\ u_{1,y}, & \ldots, & u_{n_e,y} \end{array} \right], \tag{3.14}$$

where $n_e$ is number of the shape functions of *e-th* element. Introduce the vector $\mathbf{g}_e$ and the matrix $\mathbf{A}_e$

$$\mathbf{g}_e := \mathbf{D}_F^T \mathbf{u}_e, \quad \mathbf{A}_e := \mathbf{u}_e^T \mathbf{D}_{FF} \mathbf{u}_e. \tag{3.15}$$

The solution of equation (3.6) will normally determine a numerical approach. The latter can be applied to a different engineering problem of the same mathematical class. The only replacement routines would be routines calculating vector $\mathbf{D}_F$, matrix $\mathbf{D}_{FF}$ and starting vector $\mathbf{\Lambda}$ for the Newton-Raphson iteration.

For the elastic-plastic torsion of bars, after substitution of the function:

$$h(r) = \frac{1}{2} g(\sqrt(r)), \tag{3.16}$$

into equation (2.2) matrices $\mathbf{D}_F$ and $\mathbf{D}_{FF}$ can be evaluated using equations (3.4) and (3.6). For the current case they are:

$$\mathbf{D}_F = \left[ -2\omega, 2h(T)\frac{\partial u}{\partial x}, 2h(T)\frac{\partial u}{\partial y} \right]^T, \tag{3.17}$$

$$\mathbf{D}_{FF} = \left[ \begin{array}{ccc} -2, & 0, & 0 \\ 0, & 2h(T) + 4\frac{dh}{dT}(\frac{\partial u}{\partial x})^2, & 4\frac{dh}{dT}\frac{\partial u}{\partial x}\frac{\partial u}{\partial y} \\ 0, & 4\frac{dh}{dT}\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}, & 2h(T) + 4\frac{dh}{dT}(\frac{\partial u}{\partial y})^2 \end{array} \right]. \tag{3.18}$$

## 4. Application of the Newton-Raphson method to the solution of nonlinear algebraic equations

For the solution of the system of nonlinear equations, the Newton-Raphson method is applied. The vector $\mathbf{G}$ and matrix $\mathbf{J_G}$ depend upon $\mathbf{\Lambda}$. The algorithm of Newton-Raphson can be divided into the following steps:

1. Set initial vector $\mathbf{\Lambda}_0$, set i=0;

2. Repeat points 3, 4, 5 until $||\mathbf{G}(\mathbf{\Lambda_i})|| < \epsilon\, ||\mathbf{\Lambda_i}||$;

3. Solve the following system of linear equations:

$$\mathbf{J_G}(\mathbf{\Lambda}_i)\mathbf{\Delta\Lambda}_{i+1} = -\mathbf{G}(\mathbf{\Lambda}_i); \tag{4.1}$$

4. $\mathbf{\Lambda}_{i+1} = \mathbf{\Lambda}_i + \mathbf{\Delta\Lambda}_{i+1}$;

5. $i := i + 1$.

It is assumed that the norm in $\mathbb{R}^N$ is defined as:

$$||\mathbf{x}|| = \max_{i=1,\ldots,N} |x_i|, \text{ where } \mathbf{x} = (x_1, \ldots, x_N)^T \in \mathbb{R}^N. \tag{4.2}$$

The sequence of vectors $\mathbf{\Lambda}_0, \mathbf{\Lambda}_1, \ldots$ tends to the solution. At every iteration step the Jacobi matrix must be assembled. In the presented examples, usually $10 - 15$ iterations were sufficient to obtain the value $||\mathbf{G}(\mathbf{\Lambda}_i)||$ of residuum norm of order $10^{-9}$.

## 5. Unstructured grid generation with mesh size function in arbitrary domains

The generation of a grid with arbitrary size is performed by 2D generator [6, 7]. The main idea of grid generation is based upon the algorithm of the advancing front technique and generalization of Delaunay triangulation for a wide class of 2D domains. It is assumed, that the domain is multiconnected with arbitrary numbers of internal loops. The boundary of the domain may be composed of the following curves:

- a straight line segment,
- an arc of a circle,
- a B-spline curve.

In the case of the advancing front technique combined with Delaunay triangulation, the point insertion and triangulation can be divided into the following steps:

1. Generation of points on the boundary components of the boundary of the domain,
2. Generation of internal points by the advancing front technique,
3. Delaunay triangulation of the previously obtained set of points,
4. Laplacian smoothing of the obtained mesh.

The algorithm for generation of boundary points depends upon the type of boundary segment.

## 6. Algorithm for remeshing

The whole adaptation algorithm consists of the successive generation of meshes $\{\mathbf{T}_\nu\}$, where $\nu = 0, 1, 2, \ldots$, which are based on a fresh mesh size function. By using every mesh of the sequence the problem is solved and, next appropriate error indicators in every element are calculated. The values of the error indicator are reduced to the nodes by an averaging method. Having the values of errors at nodes a continuous error function in the whole domain is constructed by using a piecewise linear interpolation. Clearly this is a simple plane for each element. Each error function spans three nodes. When this is extended to all nodes, one obtains the error function for the whole domain. The error function is appropriately transformed to obtain a multiplier for the mesh size function. The mesh size function decides how big the newly generated elements are.

The proposed approach gives ones the possibility to solve the considered problem on well-conditioned meshes and to obtain optimally graded meshes.

### 6.1. Remeshing scheme

The algorithm for remeshing can be divided into the following steps:

1. Preparation of the information about the geometry and boundary conditions of the problem to be solved,
2. Fixing an initial mesh size function,
3. mMesh generation with the mesh size function,
4. Solution of the problem given by equation 3.2 on the generated mesh,
5. Evaluation of error indicator in every element,
6. Calculation of nodal error indicator by using the averaging method,
7. Definition of the new mesh size function by using the errors found at every point,
8. If the error is not small enough, go to the point 3,
9. end of computations.

In the examples given later, it was sufficient to make 3 to 7 steps of adaptation.

### 6.2. Error indicators

The applied indicators are calculated for every element or directly at the nodes [7, 14]. Let $e_i$ for $i = 1, \ldots, n_0$ be an error indicator at *i-th* apex of the grid $\mathcal{T}_0$, and $\mathcal{P}_0$={ $P_i$,

$i = 1, \ldots, n_P\}$ − set of nodes. We define a patch of elements incidental for a given node $P_i$ as:

$$L_i = \{k : P_i \in \overline{T}_k\} \text{ for } i = 1, \ldots, n_P. \tag{6.1}$$

1. Let $N_i$ be a set of neighbours of *i-th* element:

$$N_i = \{k : \ T_k \ \textit{has a common edge with} \ T_i\}, \tag{6.2}$$

$$\text{then} \qquad \tilde{e}_i = \sqrt{\sum_{k \in N_i} \left( \frac{\partial u_i}{\partial n_k} - \frac{\partial u_k}{\partial n_k} \right)^2}, \tag{6.3}$$

   where $u_i$ is the restriction of the solution to *i-th* element and $n_k$ is unit normal to the the edge common of the *k-th* and the *i-th* element.

2. In this case, it is suggested to directly introduce the values of error indicator at every node of the mesh. The following error indicator is being adopted in the current program. From the numerical analyses it follows that the usage of this error indicator generates similar meshes to the one firstly defined.

$$e_i = \sqrt{\sum_{k \in L_i, \ l \in L_i, \ l \neq k} \left( \frac{\partial u_i}{\partial x} - \frac{\partial u_k}{\partial x} \right)^2 + \left( \frac{\partial u_i}{\partial y} - \frac{\partial u_k}{\partial y} \right)^2}, \tag{6.4}$$

   where $L_i$ is the set of numbers of elements meeting at *i-th* node.

### 6.3. Modification of the mesh size function

The modification of the mesh size function is performed at every adaptation step for the realization of the next one. The main idea of this part of the algorithm relies on the multiplication of the values of the mesh size function by an appropriately chosen function. The chosen function is continuous, linear and has the smallest value at the node where the value of the error indicator is maximum and the greatest where the value of the error is minimum. The value increases when the error decreases. To describe the algorithm of the mesh size function modification, it is necessary to reduce the values of the error indicators to nodes. For every node $P_i$ the weighted averaged value of the indicator is defined as follows:

$$\tilde{e}_i = \frac{\sum_{k \in L_i} area(T_k) e_k}{\sum_{k \in L_i} area(T_k)}, \tag{6.5}$$

where

$$L_i = \{k : \ P_i \in T_k\} \ \text{ and } T_k \text{ is the } k\textit{-th} \text{ element.} \tag{6.6}$$

In such a manner, a set of values of the error at every nodal point is obtained.

$$\alpha = \min_{k=1,2,\ldots,N_{NOD}} \tilde{e}_k, \qquad \beta = \max_{k=1,2,\ldots,N_{NOD}} \tilde{e}_k, \tag{6.7}$$

where $N_{NOD}$ is the number of nodes. Obviously, $\alpha \leq \tilde{e}_k \leq \beta$ for $k = 1, \ldots, N_{NOD}$.

The following new values are introduced:

$\lambda$ – a value indicating the greatest mesh size function reduction,

$\mu$ – a value indicating the smallest mesh size function reduction.

Usually $\lambda$ and $\mu$ have positive values and usually are one smaller than 1, additionally $\mu < \lambda$. The following transformation is defined

$$l : [\alpha, \beta] \mapsto [\mu, \lambda] \tag{6.8}$$

which satisfies the conditions: $l(\alpha) = \lambda$ and $l(\beta) = \mu$. By these assumptions it can be observed that $\mu \leq l(x) \leq \lambda$.

Provided that

$$Q_i = l(\tilde{e}_i) \ for \ i = 1, \ldots, N_{NOD}, \tag{6.9}$$

then one has: $\min_{i=1,2,\ldots,N_{NOD}} Q_i = \mu, \quad \max_{i=1,2,\ldots,N_{NOD}} Q_i = \lambda$.

Introducing the function $r : \overline{D} \mapsto \mathbb{R}$ as follows: $r(\overline{x}) = \Pi(\overline{x}), \ if \ \overline{x} \in \overline{T}_s$, where $\Pi$ is an affine mapping of two variables, satisfying the following equalities:

$$\Pi(P_i) = Q_i \ for \ i = 1, 2, 3, \tag{6.10}$$

where $P_1$, $P_2$, $P_3$ are the vertices of the triangle $T_s$ of the triangulation of $\Omega$, and appropriately $Q_1$, $Q_2$, $Q_3$ are the values defined by the formula (6.9). The function $r(\overline{x})$ is defined in the whole domain because the triangles $\{\overline{T}_s\}_{s=1}^{n_e}$ cover it. The new mesh size function is defined as follows:

$$\gamma_{i+1}(\overline{x}) = \gamma_i(\overline{x}) r(\overline{x}). \tag{6.11}$$

As $\mu \leq r(\overline{x}) \leq \lambda$ then $\mu\gamma_i(\overline{x}) \leq \gamma_{i+1}(\overline{x}) \leq \lambda\gamma_i(\overline{x})$.

It can be checked that:

$$\exists \overline{x}, \overline{y} \in \overline{\Omega} \text{such, that:} \mu\gamma_i(\overline{x}) = \gamma_{i+1}(\overline{x}), \ and \ \gamma_{i+1}(\overline{y}) = \lambda\gamma_i(\overline{y}).$$

It can be shown that

$$||\gamma_{i+1} - \gamma_i||_{\overline{\Omega},max} \leq ||\gamma_i||_{\Omega,max} \max\{|1-\mu|, |1-\lambda|\} \tag{6.12}$$

$$\text{where} \quad ||\gamma||_{\Omega,max} := \max_{\overline{x} \in \Omega}\{|\gamma(\overline{x})|\}. \tag{6.13}$$

# 7. Numerical Examples

## 7.1. Numerical test for squared domain

The size function modification depends on an error indicator and on the coefficients $\lambda$, $\mu$, which determine the value of mesh size reduction. If the values of the coefficients $\lambda$, $\mu$ are small then a smaller number of adaptation steps is necessary. How quickly an adapted grid will be close enough to an optimal mesh, next to the error indicator, it depends on an initial mesh too. In the problems solved here, it was arbitrarily assumed that $\lambda = 0.6$ and $\mu = 0.9$. The proposed algorithm was benchmarked against a known problem of elastic-plastic torsion of a bar with a square cross-section, Figures 2 and 3 provides results. The distribution of elements is seen in Fig. 2. The isolines of the Prandtl-Reuss function are given in Fig. 3. The results given in Figures 2 - 4 were obtained for the angle of twist, $\omega = 0.015 \ \ rad/m^2$. The dependence between stress and strain intensities was (see Fig. 1):

$$T = \begin{cases} 8 \times 10^5 \Gamma & \text{if} = \Gamma \leq 0,0025, \\ 1940 + 24 \times 10^3 \Gamma & \text{otherwise.} \end{cases} \tag{7.1}$$

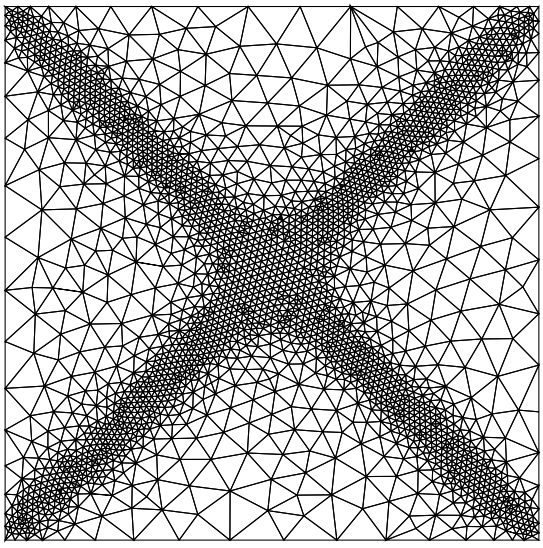

Fig. 2: Mesh after 7 adaptation steps

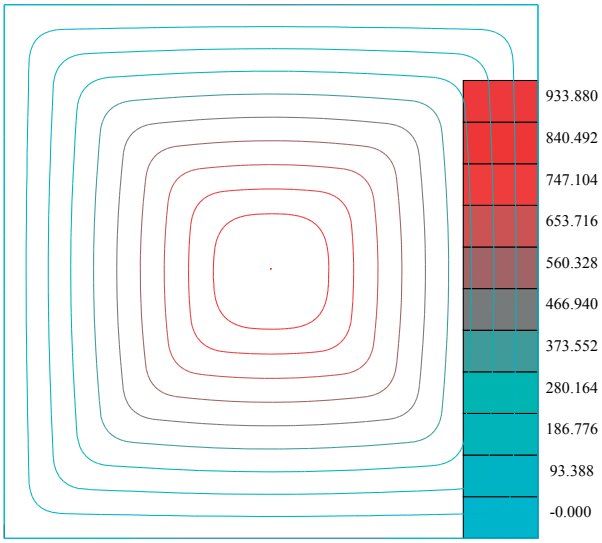


Fig. 3: Prandtl-Reuss function

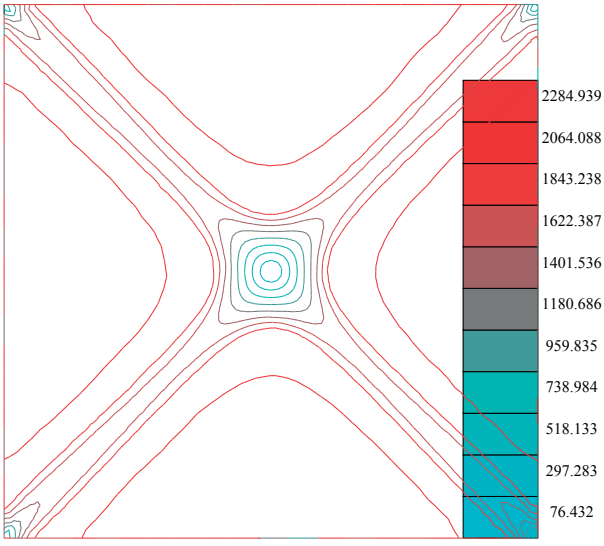


Fig. 4: Stress intensity, T

In nonlinear problems, it is sometimes useful to perform a greater number of adaptation steps. The success of the method is best illustrated in Figures 5 and 6, where a similar problem is solved for an L-shaped domain. It is seen here that when starting from a coarse mesh (Fig. 5) the algorithm quickly converges to a denser mesh (see Fig. 6) where large variations in stress take place. At the the same time the region where plastic staining took place there is no need for a dense grid (outer edges). Fig. 7 shows the Prandtl-Reuss function, and Fig. 4 depicts the stress contours.

It would be rather impossible to obtain the same effect by the methods based on mesh enrichment [1, 3, 10].

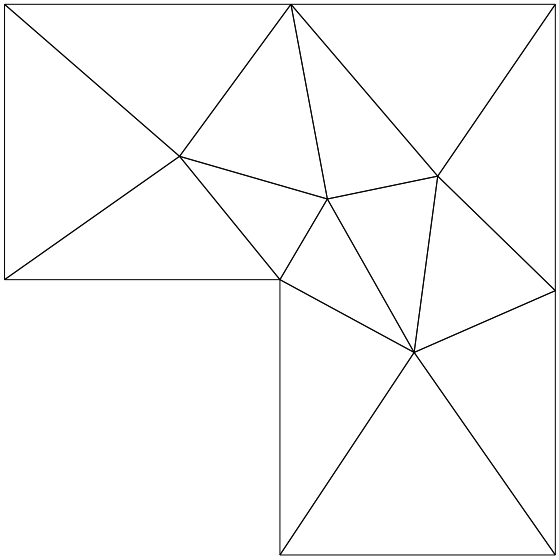

Fig. 5: Initial mesh

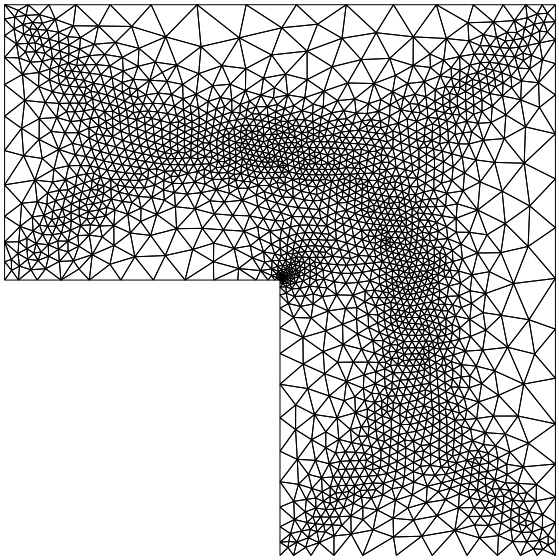

Fig. 6: Final mesh after 7 adaptation steps

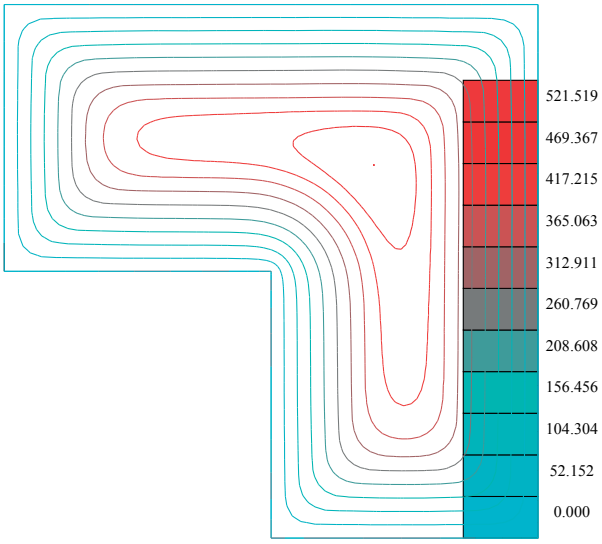


Fig. 7: Prandtl-Reuss function

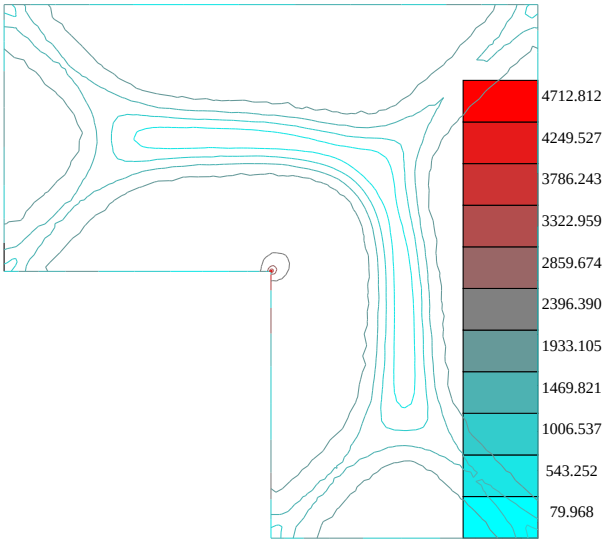


Fig. 8: Stress intensity, T

## 8. Summary and closure

- This paper presents the application of the grid generator with mesh size function of the adapted solution to the elastic-plastic twisting of bars with strain hardening.
- The mesh generator based on the Delaunay condition and the advancing front technique seems very suitable to the class of problems where zones of different in the domain are to be represented.
- After the discretization, the problem led to the solution of nonlinear systems of algebraic equations. For the considered problems, the applied Newton-Raphson method was always convergent giving the residual error of an order $10^{-10}$ in $10 - 14$ iterations.
- The whole algorithm has two loops, the external over adaptation steps and the internal over the Newton-Raphson iterations.
- The effiecient mesh size function is obtained iteratively and it is linked to the values of error indicators at nodes.
- The presented method, with its superiority above other known approaches, makes it a suitable tool for nonlinear model computations.
- An investigation of the anisotropic mesh generation would be interesting.

## References


[1] Bank R., Sherman A. and Weiser A., *Some Refinement Algorithms And Data Structures for Regular Local Mesh Refinement. Scientific Computing IMACS,* 1983.

[2] Bieterman M.B., Busseletti J.E., Hilmes C.L., Johnson F.T., Melvin R.G., Young D.P., *An adaptive grid method for analysis of 3D aircraft configurations, The Boeing Company Seattle*, Technical Report, Washington 1991.

[3] Borouchaki H., Hecht F., Frey P.J., *Mesh gradation control,* Int. J. Num. Meth. in Engng, **43**, 1998 1143—1165.

[4] Huerta A., Diez P., *Error estimation including pollution assessment for nonlinear finite element analysis,* Comp. Meth. Appl. Mech. Engng, **181**, 2000 21—41,.

[5] Kachanov L. M., *Fundamentals of plasticity theory of plasticity, Dover Publications Inc.,* 2004, 479p. (ISBN 0-486-43583-0), Mineola, NY, USA. Moscow 1968.

[6] Thompson J.F., Soni B. K., Weatherwill N.P., *Handbook of Grid Generation,* CRC Press, Boca Raton, 1999.

[7] Kucwaj J., *The Algorithm of Adaptation by Using Graded Meshes Generator,* Computer Assisted Mechanics and Engineering Sciences, **7**, 2000, 615—624.

[8] Kucwaj J., *Numerical Investigations of the Covergence of a Remeshing Algorithm on an Example of Subsonic Flow,* Computer Assisted Mechanics and Engineering Sciences, **17**, 2010, 147—160.

[9] Lo S. H., *Finite element mesh generation and adaptive meshing* , Progress in Structural Engineering and Materials, **4** 2002, 381—399.

[10] Oden J.T., Demkowicz L., Rachowicz W., Westermann T.A., *Towards a universal h-p adaptive finite element strategy, part 2, Aposteriori error estimation,* Comp. Meth. Appl. Mech. Engng., **77**, 1989, 113—180.

[11] Rachowicz W., *An anisotropic h-type mesh-refinement strategy,* Comp. Meth. Appl. Mech. Engng, **109** 1993, 169—181.

[12] Zienkiewicz O.C., Taylor R.L., *The Finite Element Method*, 4-th edition, vol. 1, Basic Formulation and Linear Problems, McGraw-Hill Book Company, London, Washington 1989.

[13] Zienkiewicz O.C.,*Achievements and some unsolved problems of the finite element method,* Int. J. Num. Meth. Engng, **47**, 2000, 9—28.

[14] Zienkiewicz O.C., Zhu J.Z., *Adaptivity and mesh generation,* Int. J. Num. Meth. Engng., **32**, 1991, 783—810.

[15] MAdLib: an open source Mesh Adaptation Library, http://sites.uclouvain.be/madlib/ 2010.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

MATEUSZ MATAN*

# APPLICATION OF THE 3*D* OBJECTS SURFACE SHAPE ANALYSIS ALGORITHMS IN BIOMEDICAL ENGINEERING

# WYKORZYSTANIE ALGORYTMÓW ANALIZY KSZTAŁTÓW POWIERZCHNI OBIEKTÓW 3*d* W INŻYNIERII BIOMEDYCZNEJ

Abstract

The paper describes a concept of measuring possible error estimation and later the decomposition of the predefined model object into convex areas ECD (Exact Convex Decomposition) in order to find a solution to the problem of cavities location with the use of three-dimensional $\mu$–tomography image of the tooth. Such an approach will enable the improvement of automatic cavities detection methods in the future. The paper is also concerned with the problem of a precise object acquisition and estimation of the error value during execution of automatic detection methods.

*Keywords: convex hull, error estimation, data acquisition, biomedical engineering, shape analysis, image processing*

Streszczenie

Publikacja opisuje koncepcję pomiaru możliwego błędu wraz z dekompozycją predefiniowanego modelu obiektu do obszaru wypukłego z użyciem metod ECD (precyzyjnej dekompozycji wypukłej) w celu znalezienia rozwiązania problemu ubytków w zębach z wykorzystaniem trójwymiarowej tomogrfii obrazu zęba. Wykorzystanie tomografu i metod ECD pozwoli w przyszłości poprawić metody automatycznej detekcji ubytków. W publikacji również rozważany jest problem precyzyjnej akwizycji obiektu z szacowaniem błędu podczas wykonywania metod detekcji automatycznej.

*Słowa kluczowe: otoczka wypukła, estymacja błędu, akwizycja danych, inżynieria biomedyczna, analiza kształtu, przetwarzanie obrazów*

*Comarch S.A.; mateuszmatan@gmail.com

## 1. Introduction

The need for algorithms for the three-dimensional objects surface shape analysis arises directly from a desire to automatize real tasks based on extensive domain knowledge concerning issues in the field of medicine, biomedical engineering or physics. However, the complexity and irregularity of real structures (such as $3D$ images of teeth and bones) is a significant obstacle to the effective - in terms of both time and quality - processing and postprocessing of measurement data. One of the possible applications of shape detection algorithms in biomedical engineering is to solve the problem of cavities location with the use of three-dimensional $\mu$–tomograpy image of the tooth [6]. In this paper there is described real (medical) object acquisition process with most common problems in this field (chapter 1). In the next chapter it is given a short survey of concepts in automatic cavities detection (chapter 2). In chapters 3 and 4 is is discussion about possible error estimation methods and expectations that will be shared with described methods. Chapters 5 and 6 are focused on a new convex hull algorithm, its implementation, efficiency and comparison with a commonly used programming library (CGAL). For more details about CGAL please refer to [2].

Precision, of both the measurement and each of the individual processing steps, is a very important aspect in biomedical engineering. Errors made in the early processing stages, accumulated during further processing, may result in falsified test results. Depending on the measured physical quantities, different methods, devices and data processing algorithms are used in the processing. The problem discussed here concerns the method of error estimation, and reducing the differences between the real object, its measurement using predefined model structure used for further processing.

### 1.1. Examined objects

A set of 20 previously prepared mandibular premolars were the examined objects. Examined objects were selected in terms of similarity of basic physical properties and features such as

- size,
- root size,
- damage degree.

All of the studied objects were previously adapted so that the measurement was in accordance with current dental procedures (removal of dental crowns, clogging of root canals, hardening of sealant).

### 1.2. Real object acquisition problems

This is the first of the performed measurements - the measurement of the current tooth structure [7]. It allows the acquisition of digital data in the form of $3D$ images of a

previously prepared object (i.e., the tooth). At this stage, the first errors which may affect the quality of further work with the object may occur. *Measuring uncertainties (errors)*, being an integral part of the measurement process, are generally being understood as deviations from the reference value. The main factors affecting the measurement uncertainty include, among others,

- environmental influence,
- imperfect measurement of environmental conditions,
- accuracy class of measuring instruments,
- imperfection of the measurement method.

One of the most visible aspects affecting the quality of measurement may include environmental influence and accuracy class of measuring instruments. It is very important that the measurement at this stage (sampling, quantization) has been performed using a device of the lowest possible quantization noise. While converting a continuous analogue signal (image) to a digital form, each value is rounded to the nearest integer. This is linked to the inevitable and irreversible loss of information. The input signal is approximated by the quantization values and the difference between quantized value and actual signal value appears as the quantization error. The reduction of the quantization error is possible by increasing the number of bits required to store the quantization levels. For example, increasing the number of bits of 1 doubles the number of quantization levels and causes quantization noise reduction by a factor of

$$20\log(2) = 6.02dB(SNR).$$

In the above formula and throughout the paper *log* denotes the decimal logarithm.

### 1.3. Acquisition and preprocessing of the real object (*ex vivo*)

High-resolution $\mu$–tomograph *X-tek (Nikon) Benchtop CT160Xi* has been used for the measurement. Scanner settings were identical for all of the objects. The original image was recorded using 32-bit floating-point numbers in accordance with *IEEE*–754 and stored as

$$x = S * M * B^E$$

where

- $S$ – sign of a number – 1 bit,
- $M$ – normalized mantissa – 23 bits,
- $B$ – base of the number system (2),

- $E$ – exponent – 8 bits.

The next processing step is to convert each element of the set of measurements to the 8–bit greyscale. This procedure reduces the size of the measurement from 32 bits per point, to 8 bits per point. Conducting a colour image conversion (RGB) to the greyscale for a picture with vibrant colours can have a significant impact on the result obtained. There are many variants of algorithms for image conversion from RGB to greyscale. The quality of the conversion depends on the accuracy of mapping luminance (value) in greyscale, using the luminance of the colour image. The two most commonly used mapping standards are *YUV* (also known as *YIQ*) and *HDTV*, in which the luminance is described as

$$Y_1 = 0.299R + 0.587G + 0.114B \quad \textit{(YUV, YIQ)} \tag{1.1}$$

$$Y_2 = 0.2126R + 0.7152G + 0.0722B \quad \textit{(HDTV)} \tag{1.2}$$

where

- $Y_1$ – greyscale luminance according to *PAL, SECAM, NTSC*,
- $Y_2$ – greyscale luminance according to *ATSC* as a standard one for *HDTV*,
- $R, G, B$ – luminances of the red, the green and the blue component.

In the case of colour images with high red, green and blue component intensity, conversion of the image to greyscale may result in a significant readability decrease, as well as the loss of information about the edges.

For photos taken with $\mu$–CT, the colours contained in the image do not carry significant information, as far as the object processing is concerned. These colours are caused by ambient noise or artefacts (noise spots) caused by the device. Hence, the information about colours can be safely ignored here as one of the steps of measurement uncertainties filtering.

During the tests, the luminance model given by (1.1) was used as the most common one. Also the choice of a luminance model for teeth measurements with the use of $\mu$–CT does not have much influence in the context of object shape processing algorithms and the construction of a model.

By using greyscale other information can be interpreted as the intensity of each pixel of the image (values from 0 - the lowest - black, to 255 - the highest - white). This procedure allowed to reduce the size of the measurement by three times.

Fig. 1: Visualization of the three-dimensional structure of a tooth extracted from a list of sequential measurements of cross–sections

$3D$ object structure obtained in this way is a set of two-dimensional images following one another from the top to the crown of the tooth (see Fig. 1). In the measurements set the dimensions of individual objects vary in the following ranges:

- length – from $210px$ to $430px$ ($px$ – pixel),
- width – from $242px$ to $528px$,
- number of cross-sections (height) – from $1187px$ to $1338px$.

The measurement obtained in such a way is the basis for further measurements and hence, its quality is extremely important. Any uncertainties and measurement errors

at this stage remain without any possibility of their reduction or removal at the later processing stages.

## 2. The problem of automatic cavities detection and availability of solutions

The solution to the problem of automatic cavities detection with the use of $\mu$–CT described in [6, 5] shows that the problem, from a technical point of view, is far from trivial. Solution to this problem requires the application of a series of transformations to simplify the model and to adapt the measurement data to the form in which calculations on the computer will be feasible.

Cavities in the structure of a tooth can be identified in a number of ways. A large amount of thematically varied algorithms (from shape analysis and analysis of the intensity of the pixels in areas with cavities to the construction of mathematical models to simplify the calculation) does not guarantee proper detection of the cavity with an acceptable error value. It is therefore a very important issue to be able to estimate detection error, and to develop complex detection methods by using a new approach and setting it with the current ones in order to compare the quality of algorithms. It is likely that the high-quality automatic detection requires developing a combined approach method based on a segmented object model and its preprocessing and then correcting the result from the previous step by applying, e.g., the pixel intensity analysis.

In this paper, it is written "concave area" instead of "non-convex (plane or $3D$) area". The trivial observation that an area is concave if and only if it is different from its convex hull allows to identify such areas by means of convex hull computer algorithms. There is proposed the solution to the problem of automatic cavities detection which consists of finding all concave areas of the volume greater than an appropriate $\delta$, between the filling of the root canal and the dentin of the tooth. Sample cavities shown in Fig. 2.2, 2.4 and 2.6 can be seen as the dimming between the filling of the root (white colour in the middle of the structure) and the dentin (grey colour) in the form of black or dark grey spots.

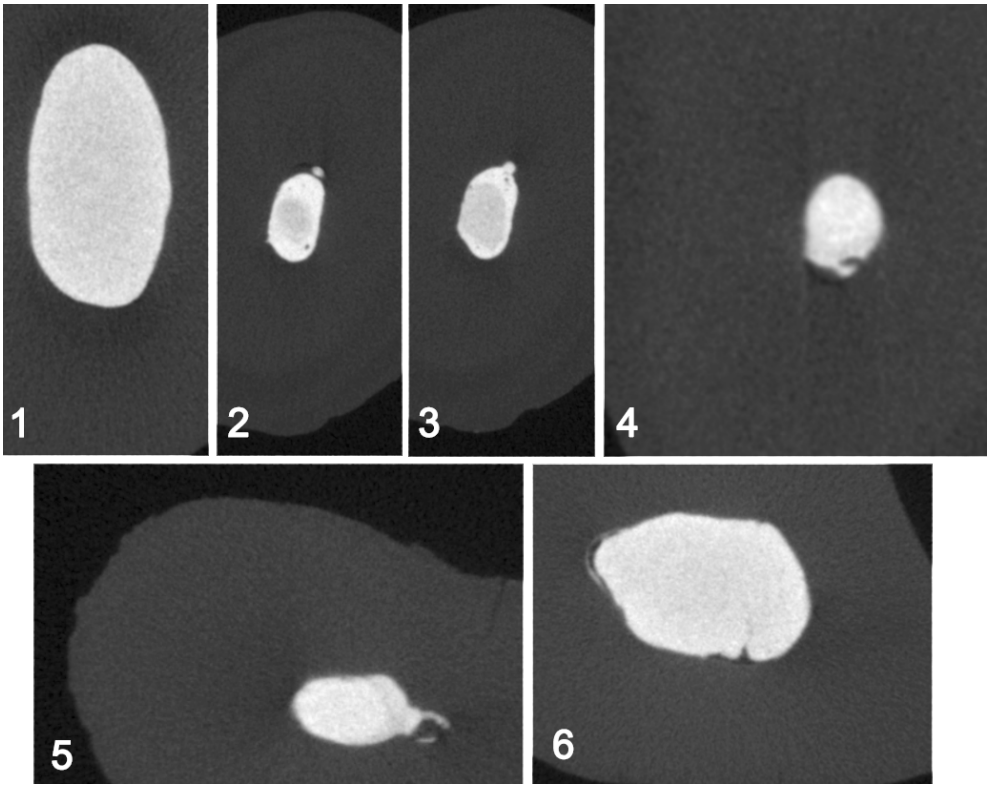


Fig. 2: Images of cross–sections through the tooth structure taken with $\mu$–CT

As one can easily see, inherent limitations (due to the volatility of the real objects) which are not cavities in the filling of the root are among the basic problems of automatic cavities detection. Those include

- natural curvature and concavity (Fig. 2.1, 2.5),
- natural extensions or isolated areas (Fig. 2.3).

Currently studied automatic cavities detection methods with the use of $\mu$–CT [6, 5] focus on the object shape analysis by using numerical techniques and image segmentation tools. The efficiency of these methods still does not allow for their practical use in dentistry, so work on this area focuses on improving the effectiveness of detection methods comparing the effects achieved using dedicated algorithms with results achieved using manual detection.

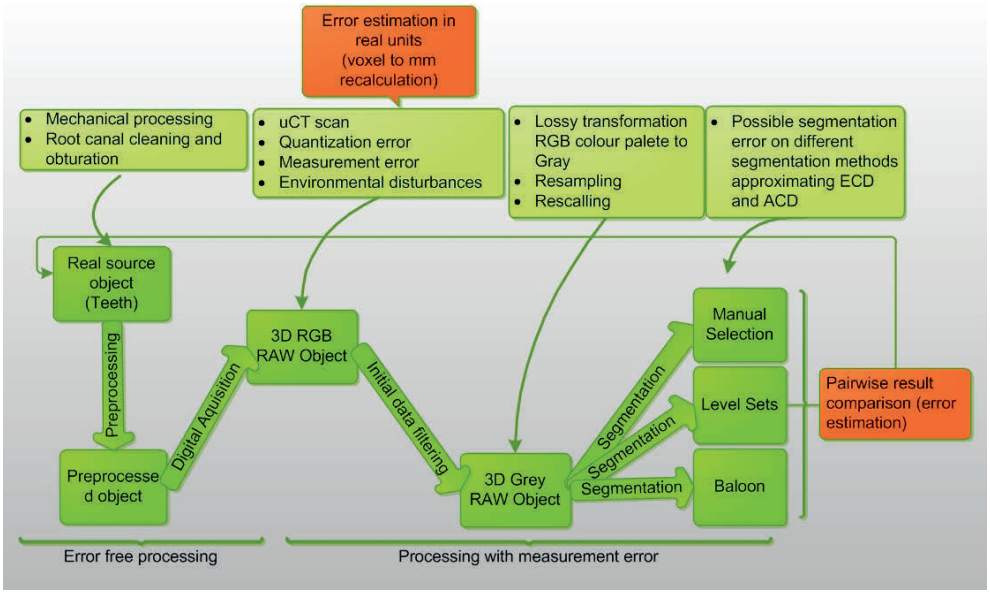


Fig. 3: Stages in the acquisition of medical images using $\mu$–CT in terms of estimating measurement errors

While comparing the effects of algorithms in biomedical engineering it is worth considering the choice of an appropriate comparison method. *MSE* (*Mean Squared Error*), the widely used method for image comparison, gives less than satisfactory results, mainly due to the even distribution of energy (error) on the entire structure of the measured image. Medical image processing algorithms are often aimed at extraction (e.g. an edge of the area), strengthening or segmentation of particular features or areas of an image while leaving the majority of the image in its original form. The use of standard mechanisms for error estimation becomes not only inefficient but can also lead to erroneous conclusions. For this reason, in most medical imaging algorithms, dedicated algorithms are used both to the problem itself and to estimate the accuracy of the measurement. Also for this reason, The latter part of this paper will contain an analysis of the algorithms comparing the results of the segmentation of images made with $\mu$–CT. Error estimation algorithms will be selected in terms of specificity of the current study, which includes all the previously described steps as each of them affects the total error of the test.

## 3. PSNR - Peak Signal to Noise Ratio

PSNR is expressed as the ratio of the maximum possible signal power to the measured power of the signal disturbed by noise and measurement uncertainties. Since the PSNR is a standard and universal error measure applied to the wide variety of signals, a number of measurement methods (including the methods of measuring the similarity of images created during transformation of a measure signal to a $2D$ or $3D$ structural representation) can be compared with PSNR. More detailed description of PSNR and other error estimation standards is available in [8]. Because of the specification of the measured signals and a wide range of their changeability, PSNR is often in a logarithmic scale. PSNR is most commonly used to compare the effects of the reconstruction of images and sounds by means of lossy compression algorithms such as MP3, JPEG. In this case, the initial measurement is the image or sound before compression and the measured object is the compressed image or sound. In the case of such an application, what is interesting is an information about the general average signal distortion across the studied object.

PSNR can be expressed most simply using MSE. Being given a monochrome noiseless image $I$ of an $m \times n$ size and its approximation $K$ (e.g., reconstruction using lossy compression), MSE is defined as follows.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i,j) - K(i,j))^2 \tag{3.1}$$

PSNR can be then expressed as

$$PSNR = 10 \log \left( \frac{M^2}{MSE} \right) \tag{3.2}$$

where $M = \max I(i,j)$.

In practical use, both PSNR and MSE are powerful tools for the study of various processing parameters of different objects, but using the same algorithm, such as JPEG. As it turns out, comparing outputs of different algorithms applied to the same image can sometimes provide subjectively better results for the object whose PSNR had lower value. This is mainly because PSNR returns information about an average error value between source and measured object, in contrast to the human eye, for which features such as preserved edges and pixel intensity are more important than their average values. The situation is very similar in the problem discussed in this paper. As far as cavities detection is concerned, information about the correct edge detection and contour retention is more important than the average accuracy (error) of the total measurement. In the case of a three-dimensional object model, it starts with a fully specified set of points [1, 3]. This collection will be a link between the root filling and the cavity. The error estimation using PSNR and MSE should therefore be

limited to the stage of data acquisition and conversion of the object to greyscale (8 bits per pixel). These stages (Fig. 3) can be treated as an acquisition of the original signal and the preprocessing stage of fixed noise resulting from a specific character of devices and conversion algorithms used.

While modeling the processing stage in terms of data accuracy control one can specify the following stages:

- stage initial - which consists of the measurement accuracy and precision of the preprocessing stage,
- stage processing - which consists of all the algorithms of detection, segmentation and modeling of the digital structure,
- stage feedback loop - in which the results obtained in the processing stage are converted from the pixel value to metric units together with error estimation at each stage.

## 4. Structural Similarity – SSIM

SSIM (Structural Similarity) just like PSNR and MSE is a method of similarity measurement between two images. Measurement of image quality uses as a reference an uncompressed and undisturbed (by any noise) image. This method is an improvement of traditional signal comparison methods which, as mentioned, do not fully reflect the perception of the human eye. Compared to other error measurement methods, such as MSE or PSNR, which estimate the size of the error, SSIM describes the degradation of the image as a preserved change in the information about the structure. The information about the structure is a concept in which the individual pixels or voxels (in $3D$ space) possess mutual dependencies. These dependencies bring important information, from a global perspective, about the examined object.

The original form of SSIM algorithm described by Z. Wang in [9] and [4] starts with the decomposition of the study area and reference area into $N \times N$ squares. Let us denote such a square by $I$. The following procedure is applied to (every) $I$:

$$SSIM(i,j) = \frac{(2\mu_i\mu_j + c_1)(2\sigma_{ij} + c_2)}{(\mu_i^2 + \mu_j^2 + c_1)(\sigma_i^2 + \sigma_j^2 + c_2)} \tag{4.1}$$

where

- $\mu_i$ – the average value of pixels in the examined square with the ordinate $i$,
- $\mu_j$ – the average value of pixels in the examined square with the abscissa $j$,
- $\sigma_i^2$ – the variance of pixels in the examined square with the ordinate $i$,
- $\sigma_j^2$ – the variance of pixels in the examined square with the abscissa $j$,

- $\sigma_{ij}$ – the covariance of pixels in the examined square defined by

$$\sigma_{ij} = \frac{1}{N}\sum_{k}(I(i,k) - \mu_i)(I(k,j) - \mu_j),$$

- $K_1 = 0.01$, $K_2 = 0.03$ – default values *a priori*,
- $L$ – dynamics of the variables used, expressed as their maximum possible value minus one (for the 8-bit greyscale it is $2^8 - 1$),
- $c_1 = (K_1 L)^2$, $c_2 = (K_2 L)^2$ – two constants to avoid the division by 0.

Computed values can be seen from two different perspectives due to this algorithm. As an example, SSIM algorithm with cluster size $10px \times 10px$ will divide the image of size $1000px \times 1000px$ into $2D$ structures of size $100px \times 100px$. These structures will be called "error boxes". Each pixel in error box will contain SSIM value computed based on different area of size $10px \times 10px$. Now, to compare two $2D$ images there can be used two approaches.

- Compare independently precomputed error boxes (each pair of corresponding pixels on both images). This approach is more precise because it will give an information about "area similarity" which is much more detailed than one summary value from one image.
- Treat the first error box as the first stage of SSIM computing. Repeat SSIM computing until the error box will contain only one value. In this case, two images can be compared using summary computed SSIM values.

Compared to the MSE this measure is more consistent with the perception of the human eye, which can be easily seen in Fig. 4 (From a) to f)). With identical values of MSE (144), SSIM values vary depending on the distortions of the image. The least readable image (bottom right corner) has the lowest SSIM value.

Similarly to SSIM, it can be defined a measure of lack of similarity referred to as *DSSIM* (*Structural DisSimilarity*) [4]. It is derived from SSIM in the following way.

$$DSSIM(i,j) = \frac{1}{1 - SSIM(i,j)} \tag{4.2}$$

DSSIM can be also used as an image distance metric which corresponds better to the human perception than MSE and PSNR.

Error estimation methods are the very beginning part of real objects surface estimation. Algorithms such as SSIM and DSSIM can be effectively applied to problems mentioned in this paper, in contrast to more common MSE and PSNR.

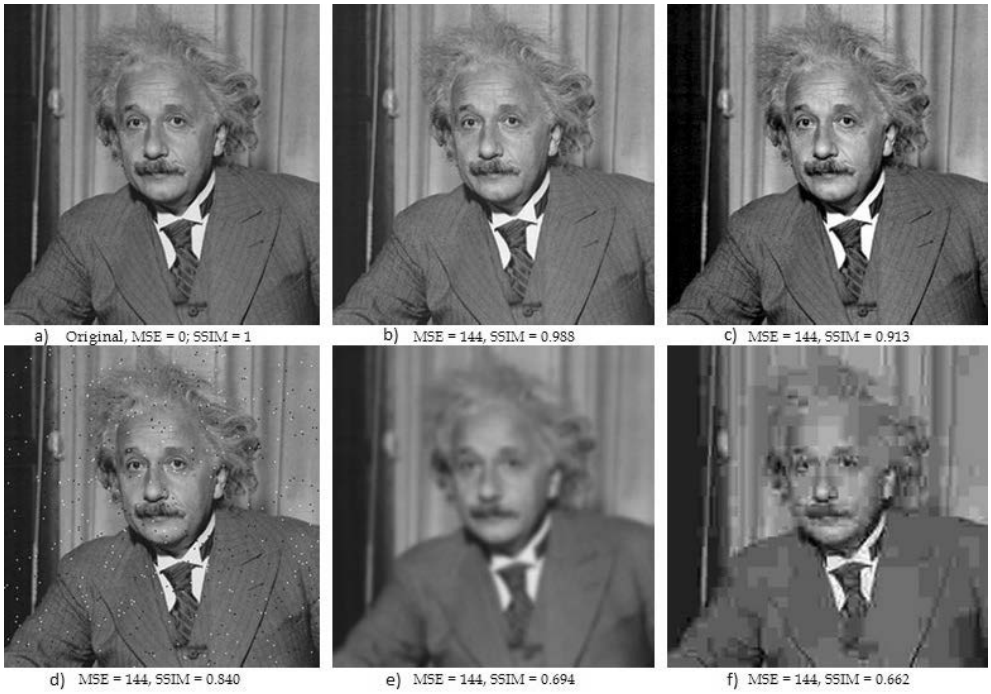


Fig. 4: Juxtaposition of MSE and SSIM with images of different noise levels

## 5. An algorithm for constructing a segmented object model based on the performed measurement

The most important topic in this paper is focused on ECD (Exact Convex Decomposition). In our problem finding a segmented object model allows to create processing feasible object (models based on RAW data are not feasible). This model can be created using a current measurement $3D$ image ($\mu$–tomography scan). To create a computationally desirable model there should be analyzed some convex hull algorithms, because of a few important features:

- less computationally demanding (in postprocessing stage),
- stable (of known polynomial or logarithmic complexity),
- convex surface is the expected shape for a tooth root,
- convex hull algorithms can be easily adapted to new approaches (surfaces, point selection algorithms, parallelization).

As it was shown can see above, our goal is to design an efficient and problem adapted convex hull algorithm, which will produce the convex hull of a tooth root cavity. Optimal parameters for a convex hull algorithm in the $3D$ space, which is worth considering, are:

- computational complexity (both memory and time - the time of the CPU) and its dependence on points (their order, density and number),
- possibility to parallelize,
- the dependence on the current relation between the number of points in the hull and the total number of points (output sensitive algorithms).

A natural consequence of the parallelized convex hull algorithm design for a given problem is incorporation of elements of the problem into the formula of the algorithm. This considerably increases the time for processing, but simultaneously, it allows the optimization of the time necessary to construct a cavity model based on the measurement data. The starting point was therefore a $3D$ version of the commonly used incremental algorithm that was merged into the algorithm for searching qualifying points. The final algorithm version was parallelized at the end to speed up calculations. The algorithm for searching qualifying points is based on a high pass pixel intensity filter. For each pixel in the image algorithm decides if intensity is above or below a hardcoded intensity value. If intensity is above, the pixel is qualified for further processing. This simple algorithm can be easily parallelized. Using constant thread pool (i.e. 8 threads, depending on the number of processor cores), each thread will simply analyse a surface in the $3D$ image structure.

**ParallelIncrementalConvexHull( *R*, *alpha*, *thd* )**

**ParallelIncrementalConvexHull( *R*, *alpha*, *thd* )**
**Input**:
*R* – $3D$ picture of a tooth (vertical set of $2D$ pictures)
*alpha* – value from $[0, 255]$ of pixel brightness
*thd* – number of threads used to run algorithm, optimally as many as processor cores
**Output**:
The convex hull $C = CH(R)$ of a cavity in $R$.

```
1  Run the qualification algorithm with thd on R with alpha.
   This will produce a set P of qualified points (decreased
   global number of points from 50 to 5 mln) ;
2  Find 4 points that form the initial tetrahedron,
   C = CH({p_1, p_2, p_3, p_4});
3  Compute a random permutation {p_5, p_6, ..., p_n} of the remaining
   points;
4  foreach thread in thd do
5  |  Split P into thd sets such that every thread contains
   |  1/thd part of P;
6  |  Each thread will compute a different hull, ALL hulls
   |  will be glued together at the end.;
7  |  Initialize the conflict graph G with all visible pairs
   |  (p_t, f), where f is a facet of C and p_t, t > 4, is a
   |  non-processed point;
8  |  foreach r = 1 to n - insert p_r into C_p - where p_r is a
   |  point from the set P_thd (created at step 6) and C_p is a
   |  part of the convex hull - different for each thread -
   |  which will be glued at the end do
9  |  |  if F_conflict(p_r) is not empty - p_r is outside, any facet
   |  |  is visible then
10 |  |  |  delete all facets in F_conflict(p_r) from C_p - only
   |  |  |  from hull C_p, not from G;
11 |  |  |  walk around the visible region boundary, create
   |  |  |  the list L of horizontal edges;
12 |  |  |  foreach all e ∈ L do
13 |  |  |  |  connect e to p_r by a new triangular facet f;
14 |  |  |  |  if f is coplanar with its neighbor facet f'
   |  |  |  |  along e then
15 |  |  |  |  |  merge f and f', take conflict list from f';
16 |  |  |  |  else
17 |  |  |  |  |  - not coplanar ⟹ determine conflicts for
   |  |  |  |  |  the new facet f;
18 |  |  |  |  |  create a node for f in G;
19 |  |  |  |  |  let f_1 and f_2 be the facets incident to e
   |  |  |  |  |  in the old CH(C_{p_{r-1}});
20 |  |  |  |  |  P(e) = P_conflict(f_1) ∪ P_conflict(f_2) ;
21 |  |  |  |  |  foreach point p ∈ P(e) do
22 |  |  |  |  |  |  if f is visible from p then
23 |  |  |  |  |  |  |  add (p, f) to G - new edges;
24 |  |  |  |  |  |  end
25 |  |  |  |  |  end
26 |  |  |  |  end
27 |  |  |  end
28 |  |  |  Delete the node corresponding to p_r and nodes
   |  |  |  corresponding to facets in F_conflict(p_r) from G,
   |  |  |  together with their incident edges;
29 |  |  end
30 |  end
31 |  Merge C_p into C;
32 end
```

**Algorithm 1:** Parallel Incremental Convex Hull Algorithm

It will be shortly described the above version of Parallelized Random Incremental Convex Hull algorithm. Let $R = \{p_1, \ldots, p_n\}$ be a set of $n$ points that form a $3D$ image. Each thread will execute the same code, so to simplify this description, parallelizing part was skipped. For $i = 5, \ldots, n$ define $R_i = \{p_1, \ldots, p_i\}$. Suppose that the convex hull $CH(R_{i-1})$ is already computed. The goal is to compute $CH(R_i)$. If $p_i$ is inside $CH(R_{i-1})$, then result is simply $CH(R_i) = CH(R_{i-1})$. Otherwise, $p_i$ is on the boundary of $CH(R_i)$. All edges of $CH(R_{i-1})$ between two straight line segments joining $p_i$ to vertices of $CH(R_{i-1})$ should be deleted, and these two straight line segments should be added into $CH(R_i)$. Those few steps are executed sequentially for each point, and in the parallelized version the original problem is divided into *thd* number of subproblems executed by each thread independently.

This method can be parallelied but as an sequential algorithm it is nor the fastest one neither the less resources expensive. This is a good point to start affecting development of image processing in medical areas with deterministic and computationally feasible algorithms designed specially for these purposes.

The time complexity of the algorithm is $O(n \log n)$, and the expected memory usage is $O(n^2)$, where $n$ stands for the number of points processed (i.e., the size of the source problem). What is more important, it is expected a very good time efficiency with results comparable to other popular $3D$ convex hull algorithms like D&C and Gift Wrapping.

## 6. The algorithm tests and comparison with CGAL

CGAL (The Computational Geometry Algorithms Library) is a programming library for C++ which provides a wide range of algorithms, including triangulation, Voronoi diagrams, mesh generation, the convex hull shape analysis and many others. During the measurement it was used to verify the accuracy of the proposed algorithm and in timing tests.

All the tests were performed on Windows 7. The computer used to compare the algorithms was equipped with $6GB$ of memory and 2 core (4 threading) Intel i7 – $2.2 - 2.8$ GHz. The source code was written in C++ and compiled with the GCC compiler (within the MinGW environment).

Test 1. The calculation of the convex hull of a million random points distributed in the $3D$ space.
CGAL static approach – 1,67s
CGAL dynamic approach – 9,76s
CGAL incremental algorithm – 12,01s
The suggested algorithm – 4,33s

Test 2. The calculation of the convex hull of three measurements of mandibular premolar (20 mln. points ±100000).
CGAL static approach – 38,43s
CGAL dynamic approach – 44,25s
CGAL incremental algorithm – 79,22s
The suggested algorithm – 36,94s

Test 3. The calculation of the convex hull of three measurements of cavities in mandibular premolars (5 mln. points ±50000), taking into account the time for qualifying points calculation.
CGAL static approach – 38,65s
CGAL dynamic approach – 81,34s
CGAL incremental algorithm – 93,64s
The suggested algorithm – 33,44s

The performance of the algorithm searching qualifying points takes 28, 32s. The idea of this algorithm is to find all the points of brightness above a specified threshold (inside the cavity). This algorithm is not the subject of this problem and will not be discussed.

From the above results of the measurements, it can be concluded that the proposed algorithm in two of the three generated cases is not optimal in terms of execution time (based on our experience, the difference of approx. 5% of the time size is negligible, because the testing environment influence can be bigger despite few test trials). For a real-life case (described in Test 3) it turns out to be better then the static approach (over 10% of static time). Due to its flexibility (multi-threading, integration of the algorithm searching qualifying points) and the proposed use (analysis of medical structures) it can be a good alternative to the commonly used methods.

## 7. Concluding remarks

The present paper highlights the need for graphics processing for the requirements of dentistry in biomedical engineering applications and the possibilities which are available this area. The first important aspect is the need to verify the quality of the measurement, or the reconstruction, by selecting the appropriate method of error verification and measurement uncertainty. Error indicators such as MSE and PSNR only provide information about the energy distribution of all components and properties of the compared images. In the case of biomedical issues, a more efficient method would be the comparison of images on the basis of contours (edges) and the brightness of each pixel (voxel), mainly due to the analogy with the human eye perception. Applying algorithms such as SSIM or DSSIM in biomedical engineering is not common solution and still has not been tested nor described in other science literature. On the basis of the performed tests juxtaposing MSE error indicators and SSIM it can be

concluded that SSIM algorithm considered by VQEG (video quality experts group) as one of the standard algorithms for image quality assessment, despite the much greater computational complexity, is a strong competitor for the PSNR and MSE in image processing for medical purposes. The second considered aspect is algorithms for the construction of models of areas of interesting features from the point of view of a given field (algorithm qualifying points on the basis of, e.g., their brightness). Here convex hull algorithms were considered. They are able to generate convex area, which meets the initial limitations, on the basis of $3D$ space analysis. The proposed algorithm was a parallelized version of the incremental algorithm. All the tests were compared (in terms of calculation speed and accuracy) to the algorithms implemented within CGAL library which is commonly used for scientific calculations. Modifications of incremental algorithm were limited to calculations scattering and preliminary points classification. The resulting code has proven to be competitive in terms of time for algorithms within CGAL library (QuickHull3D, DC3D, GiftWrapping3D, incremental static and dynamic). This allows to claim that in the future the application of a similar approach to the optimization of calculations for medical structures analysis will allow to create algorithms enabling to generate $3D$ models in real time with an acceptable resemblance to the original structure of the design model. As a result, the operations performed on the segmented object model will be less computationally complex, which will allow broader analysis of structures with computer-assisted methods.

## References

[1] Borgefors G., Sanniti di Baja G., *Analyzing Nonconvex 2D and 3D Patterns*, Computer Vision And Image Understanding, Vol. 63, No. 1, 1996, 145—157.

[2] CGAL Open Source Project – *http://www.cgal.org/*.

[3] Lien J., *Approximate convex decomposition and its applications* (PhD dissertation), Texas A&M University 2006.

[4] Loza et al. A., *Structural Similarity-Based Object Tracking in Video Sequences*, Proc. of the 9th International Conf. on Information Fusion, Florence 2006.

[5] Metzger Z., Zary R., Cohen R. , Teperovich E., Paque F., *The Quality of Root Canal Preparation and Root Canal Obturation in Canals Treated with Rotary versus Self-adjusting Files: A Three-dimensional Micro-computed Tomographic Study*, JOE, Vol. 36, No. 9, 2010, 1569—1573.

[6] Petryniak R., Tabor Z., Kierklo A., Jaworska M., *Detection of voids of dental root canal obturation using micro–CT*, Computer Vision and Graphics, Lecture Notes in Computer Science, Vol. 7594, Springer, 2012, 549—556.

[7] Rhodes J. S., Pitt Ford T. R., Lunch J. A., Liepins P. J., Curtis R. V., *Micro-computed tomography: a new tool for experimental endodontology*, International Endodontic Journal, Vol. 32, 1999, 165—170.

[8] Salomon D., *Data Compression: The Complete Reference (4 ed.)*, Springer 2007.

[9] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P., *Image quality assessment: From error visibility to structural similarity*, IEEE Transactions on Image Processing, Vol. 13, No. 4, 2004, 600—612.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

ZBIGNIEW MROZEK*

# TEACHING, MODELING AND VISUALISATION OF ORDINARY DIFFERENTIAL EQUATIONS

# NAUCZANIE, MODELOWANIE I WIZUALIZACJA RÓWNAŃ RÓŻNICZKOWYCH ZWYCZAJNYCH

Abstract

Advances in computer technology and increased interest in dynamical systems influence the way of teaching ordinary differential equations. The paper presents inquiry oriented teaching, usage of modeling, visualisation and interactive web services. Last chapter describes the ways of using MATLAB or public domain software (e.g. Octave) to solve ordinary differential equations.

*Keywords: dynamical system, ordinary differential equation*

Streszczenie

Postęp w technologii komputerowej oraz wzrost zainteresowania modelowaniem systemów dynamicznych wpływa na sposób nauczania matematyki, w tym równań różniczkowych zwyczajnych. Przedstawiono podejścia: nauczania przez zadawanie pytań, wykorzystanie modelowania, wizualizacji i interaktywnych usług sieci web. Ostatni rozdział opisuje sposoby wykorzystania środowiska MATLAB lub oprogramowania dostępnego jako public domain (np. Octave) do rozwiązywania równań różniczkowych zwyczajnych.

*Słowa kluczowe: równanie różniczkowe zwyczajne, system dynamiczny*

*Faculty of Electrical and Computer Engineering, Cracow University of Technology; zbigniew.mrozek@pk.edu.pl

## 1. Introduction

In most cases teaching of linear second order ODE (ordinary differential equation), as seen in major textbooks, is done in a very procedural manner where the emphasis lies in identifying the type of equation and then apply a number of well established steps that yield to the solution [14].

Need to improve effectivity of teaching differential equations is not new. The importance of visualization of solutions was known to Karl Menger. In his paper [8] he noted that "the simplest introduction to the theory of differential equations is offered by nature". He suggested sprinkle fine iron splinters in magnetic field to visualize solution of the 1-st order ODE (Fig.1).

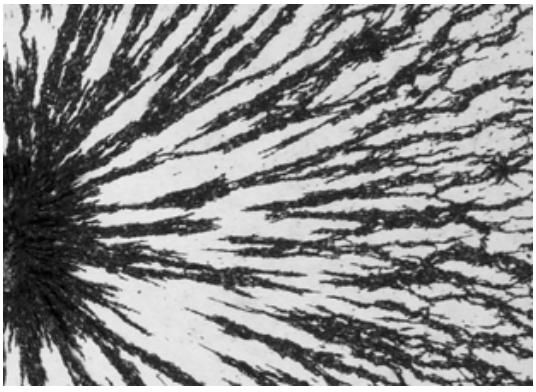

Fig. 1: Magnet and fine iron splinters, ODE visualization proposed by Menger (see also Fig.2)

Gian-Carlo Rota [17] shows how little the content of the ODE courses has changed since Cauchy. Even the order of presentation of the topics has not been altered. Rota was invited to prepare several addresses for Mathematical Association of America. At the 1997 meeting, he described lecture notes for introductory course in differential equations by Cauchy and the book on Differential Equations by Boole. Both were published at about the same time in the nineteenth century. Cauchy notes are written in an attractive, flowing style. Half of the book by Boole describes the solution of the first order differential equations. But the Boole's techniques are not very useful now. Rota claims that only two of them have survived, separation of variables and changes of variables.

## 2. Elementary course of differential equations

In an elementary course of differential equations, students should learn [17] a few basic concepts that they will remember for the rest of their lives, such as the universal occurrence of the exponential function, stability, the relationship between trajectories and integrals of systems, phase plane analysis, manipulation of the Laplace transform, perhaps even the fascinating relationship between partial fraction decompositions and convolutions via Laplace transforms. Students should also learn the basic theorems.

### 2.1. The normal system of first order ODEs

The normal system of first order ODEs will be considered.

$$\begin{pmatrix} \dot{y}_1 \\ \vdots \\ \dot{y}_n \end{pmatrix} = \begin{pmatrix} Y_1(y_1, ...y_n; t) \\ \vdots \\ Y_n(y_1, ...y_n; t) \end{pmatrix} \tag{2.1}$$

Many kinds of ODEs can be reduced to the normal system, e.g. nth order ODE

$$\frac{d^n y}{dt^n} = F(t, y, \frac{dy}{dt}, ..., \frac{d^{(n)} y}{dt^{(n)}}) \tag{2.2}$$

can be changed to the normal system of first order ODE (3) substituting $y_1 = y, y_2 = \frac{dy}{dt}, ..., y_n = \frac{d^n y}{dt^n}$

$$\begin{pmatrix} \frac{dy}{dt} \\ \vdots \\ \frac{d^n y}{dt^n} \end{pmatrix} = \begin{pmatrix} y_2 \\ \vdots \\ F(t, y_1, ..., y_n). \end{pmatrix} \tag{2.3}$$

### 2.2. Theorems on finding solutions of normal system of first order ODEs

The Picard–Lindelöf theorem, Picard's existence theorem, Cauchy-Picard or Cauchy–Lipschitz theorem is an important theorem on the existence and uniqueness of solutions to first-order equations with given initial conditions.

**Theorem 1. The Cauchy-Picard existence and uniqueness theorem**
Assume $f : R^n \to R^n$ has a continuous derivative. For every initial condition $x_0$ there exists $r > 0$ such that on the time interval $[0; r)$ there exists exactly one solution of the initial value problem

$$\dot{x}(t) = f(x(t)), \quad x(0) = x_0$$

Since there is no restriction on $F$ to be linear, this also applies to **non-linear equations**, and it can also be respectively applied to **systems of first order ODE equations** [1].

### 2.3. Inquiry oriented teaching of differential equations

Inquiry oriented teaching is based on guiding students with questioning, prompting and useful feedback [11, 19, 20]. Students are asked for reasoning and justification. The main results emerge through whole class discussion and oral presentation done by students.

Kwon and Rasmussen [5, 6, 15] engaged students in the Inquiry Oriented Differential Equations (IO-DE) project. This was a collaborative effort to explore the prospects and possibilities for improving undergraduate mathematics education, using differential equations as a case example.

Rasmussen, at all [16] conducted an evaluation study between students taking part in IO-DE project and traditionally taught classes. They compared students' skills and conceptual understandings of central ideas and analytic methods for solving differential equations. Students taking part in project, regardless of academic backgrounds and gender differences, outperformed traditionally taught students on the post-test. Thus, the IO-DE approach could enhance long-term retention of mathematics for all students. The drawback is that the IO-DE needs much more work for botch, students and professors [11]. More mathematical problems useful for the ODE class may be found in [18].

### 2.4. Team-based labs of differential equations

The course which covers traditional topics such as first-order equations, second-order linear equations with constant coefficients, Laplace transforms, and systems of first-order equations has been team-taught by two instructors; one from the School of Mathematics and Computer Science and one from the School of Engineering and Technology [13]. The emphasis was on the real-life modeling applications of differential equations.

Examples of teaching a mathematics course with an emphasis on applications have been reported in literature numerous times [4]. Students are assigned challenging real-life problems which are typically open-ended and unstructured. The process of linking the mathematical theory with applications has also led to introduce team-based labs into a differential equations course in which students are exposed to extensive numerical experimentation. Another approach is to provide the students with easy-to-use technological tools such as MATLAB programs or interactive web sites.

## 3. Modeling and simulation in teaching ODE

Computer Algebra Systems (CAS) such as Maple, MATLAB or Mathematica are capable of yielding the numerical or analytical solution. One may choose also public domain software such as COPASI. ECLlect2, FuncDesigner, GNU Octave, Julia, Modelica, Scicoslab, Scilab. Using user friendly software, the student's focus is shifted

from applying the well defined pattern of steps, to that of understanding the derivation of these steps [14].

Computer software plays a central role supporting modeling and simulation in all engineering areas. Thus numerical methods become an important tool to support teaching. Numerical methods and symbolic manipulation are implemented as computer programs, and the results are immediately visualized using powerful graphic software [2]. Even final reports can be automatically generated and printed using e.g. MATLAB Report Generator or menu option Publish (in MATLAB [10]).

### 3.1. Interactive web resources

A good example of Interactive web resources is *Interactive Differential Equations* page *http://www.aw-bc.com/ide* arranged and supported by Pearson Addison-Wesley. The team of authors, from three universities (Cornell, California Polytechnic and St. Louis) prepared 31 problems and many additional sub-problems in area of linear algebra, systems of differential equations, 1st and 2nd order ODE, series solutions, chaos and bifurcation.

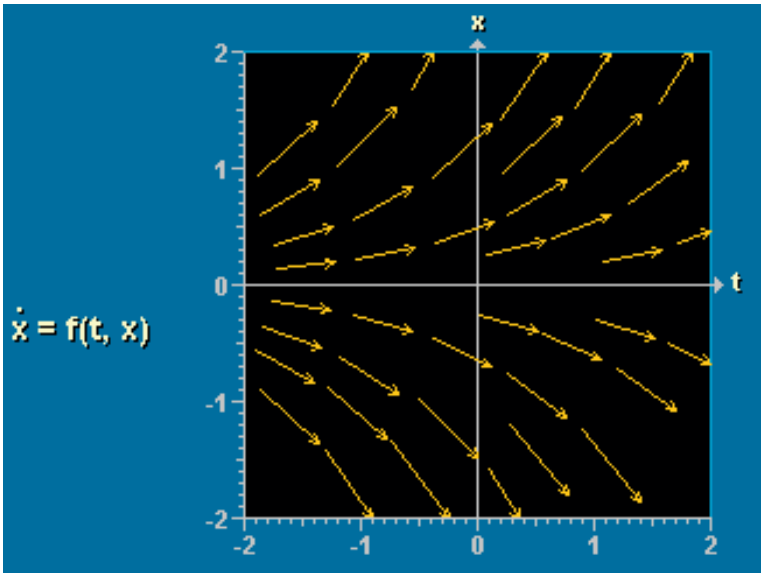


Fig. 2: Interactive web resource [3]: visualization of slope field for $\dot{x} = x$

Each problem starts with few pages of pdf file describing the given problem (may be printed). The tool instruction page describes user interface: how to use buttons and sliders in actual problem. To start simulation one has to setup the initial conditions (IC). This is done by clicking the plot area with mouse at the point chosen to be new

IC. To get family of plots, one has to use several starting points. Operating interactive web is very easy, intuitive and user friendly.

A nice example from this web is *Graphing Differential Equations>Slope Fields* (Fig.2). It may be found on the site:
*http://www.aw-bc.com/ide/idefiles/media/JavaTools/slopefld.html*. To start, one has to click inside any plot area to setup the initial condition for the equations. All problems are ready to use in class or as homework for students and are completely free.

## 4. MATLAB®, Simulink® and other tools

MATLAB is a general high level programming language and programming environment. It is user friendly and integrates reliable algorithms of applied mathematics and numerous expansion modules focused on specific fields of application. MATLAB enables to easily solve a variety of problems in science, industry, medicine, economy and many other areas by facilitating access to efficient computational algorithms and the ability to visualize the results of computation [10, 13]. Many advanced tools such as toolbox libraries, Simulink and blockset libraries and many other) are available for purchase

MATLAB successfully replaces universal programming languages (Fortran, C, C#, C++) in the area of scientific and technical calculations. Its professional math library is based on optimized packages: **LAPACK** ( *Linear Algebra Package*) and **BLAS** ( *Basic Linear Algebra Subroutines*).

### 4.1. ODE algorithms in MATLAB

MATLAB uses effective variable order and variable step solvers for normal system of first order ODEs. The ODE solvers are [7]:

- `ode45` formula, the Dormand-Prince pair is used as a "first try" for most problems;
- `ode23` solver is based on an explicit Runge-Kutta (2,3) pair of Bogacki and Shampine. It may be more efficient than ode45 at crude tolerances and in the presence of mild stiffness;
- `ode113` solver is Adams-Bashforth-Moulton PECE solver. It is a multistep solver and needs solutions at several preceding time points to compute the current solution.

For stiff problems there are four solvers

- `ode15s` is a multistep solver based on the numerical differentiation formulas (NDFs). Optionally, it uses the backward differentiation formulas (BDFs, the

Gear method). The ode15s and ode23t can solve some DAEs ( *Differential Algebraic Equations*) of the form $M(t,y)y' = f(t,y)$, where $M(t,y)$ is singular. The DAEs must be of index 1.

- `ode23s` is based on a modified Rosenbrock formula of order 2. It is a one-step solver and may be more efficient than ode15s
- `ode23t` is an implementation of the trapezoidal rule using a "free" interpolant. It may be used for moderately stiff problems and DAEs.
- `ode23tb` is an implementation of TR-BDF2, an implicit Runge-Kutta formula using a trapezoidal rule step and then a backward differentiation formula of order 2.

Fully implicit differential equations of the form $f(t,y,\dot{y}) = 0$ may be solved with solver `ode15i` using the variable order BDF method.

### 4.2. Other problems related to differential equations

MATLAB may be used to solve many problems related to differential equations, as

- boundary value problem for ODE
- delay differential equation initial value problem
- partial differential equations (PDE): 1-D initial-boundary value problem for parabolic or elliptic PDE. To solve PDE in two-space dimensions (2-D) and time, one should buy *the Partial Differential Equations Toolbox*$^{\text{TM}}$, which extends MATLAB.
- numerical integration and differentiation may be done in MATLAB with quadratures, double and triple integrals, and multidimensional derivatives

### 4.3. ODE example

Given 3rd order Cauchy problem:

$$\dddot{y} + 6\ddot{y} + 11\dot{y} + 6y = 0, \quad y(0) = 0, \quad \dot{y}(0) = 0, \quad \ddot{y}(0) = 1 \tag{4.1}$$

#### 4.3.1. Solving ODE example with odeXX solver

MATLAB ODE solvers accept only set of first-order differential equations. To solve higher-order ODEs, one has to rewrite equation (1) as an equivalent system $\dot{y} = f(t,y)$ of first-order ODE

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} y_2 \\ y_3 \\ -6y_3 - 11y_2 - 6y_1 \end{pmatrix}; \quad \begin{pmatrix} \dot{y}_1(0) \\ \dot{y}_2(0) \\ \dot{y}_3(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{4.2}$$

The next step is to prepare few lines of MATLAB code (M-function) that evaluates the right hand side of (2). This is a very flexible approach as any nonlinearities may be easily computed in MATLAB. There is even more flexibility using `ode15i` solver, which accepts fully implicit set of differential equations of the form $f(t, y, \dot{y}) = 0$. The MATLAB function for problem (2) is

```
function [ dydt ] = odefun3(t,y) % as needed by odeXX solvers
% computes vector of derivatives of equivalent ODE system
dydt=[y(2); y(3); -6*y(3)-11*y(2)-6*y(1)];
%vector dydt=[dydt(1); dydt(2); dydt(3)]
```

The above function (% means start of comment) will be used by solver (e.g. `ode23`) to compute new values of derivatives for all needed `t` and `y`. The solver may be called from the MATLAB **Command Window** as

```
>>[t,y]=ode23('odefun3',[0,10],[0,0,1]);  plot(t,y(:,1))
%compute and plot solution
```

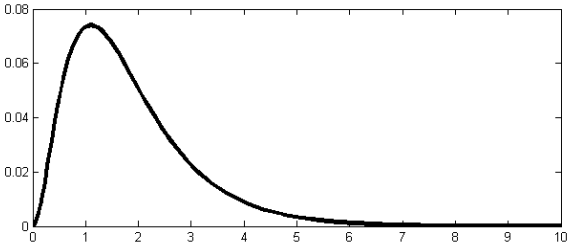


Fig. 3: Numerical solution of 3-rd order ordinary differential equation (1)

Solution may be examined for any change done in equation or its initial conditions, as presented on Fig.4. On slider movement the $\beta$ parameter in ODE $\ddot{x} + \beta\dot{x} + x = 0$ is changed and new plots are immediately generated. Visualisation is supported with GUI ( *graphical user interface*), already included in MATLAB.

### 4.3.2. Analytical solution of ODE

The analytical solution of ODE may be obtained using the `dsolve` command if additional MATLAB library *Symbolic Math Toolbox* is available. The code needed for $\dddot{y} + 6\ddot{y} + 11\dot{y} + 6y = 0$ is:

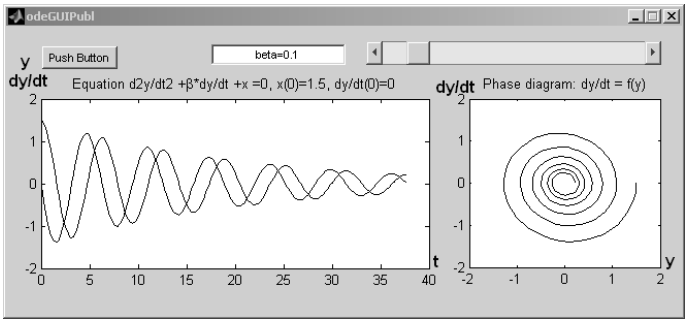


Fig. 4: Graphical User Interface (GUI) and visualisation in MATLAB.

```
>> y=dsolve('D3y+6*D2y+11*Dy+6*y = 0','D2y(0)=1, Dy(0)=0, y(0)=0')
y = exp(-t)/2 - exp(-2*t) + exp(-3*t)/2
```

Solution may be plotted using `ezplot(y,[0,10])`. where [0, 10] is the time span for plot. If initial conditions are not given, the general solution will be computed.

```
>> y=dsolve('D3y+6*D2y+11*Dy+6*y = 0')
y = C1*exp(-2*t) + C2*exp(-t) + C3*exp(-3*t)
```

where `C1, C2, C3` are arbitrary constants used to satisfy the initial conditions.

### 4.4. Solving ODE using MATLAB/Simulink

Simulink is integrated with MATLAB. It provides a graphical editor, customizable block libraries (blocksets) and solvers for modeling and simulation of dynamic systems.

Using Simulink, there is no need to prepare any lines of program code, as blocks from Simulink library are already linked to executable code. There are integrators, amplifiers, adders and other useful blocks in the Simulink library. For example, the integrator block computes the value of the integral of its input signal with respect to time and initial condition. Icon of the integrator block shows $\frac{1}{s}$, the Laplace transform of the unit step function 1(t).

Figure 5 shows Simulink model of 3rd order ODE (1). There are three integrators (icon is $\frac{1}{s}$), three amplifiers (its gain is equal to the coefficients in ODE equation) and one adder. The scope block is used to visualize the ODE solution as a function of time.

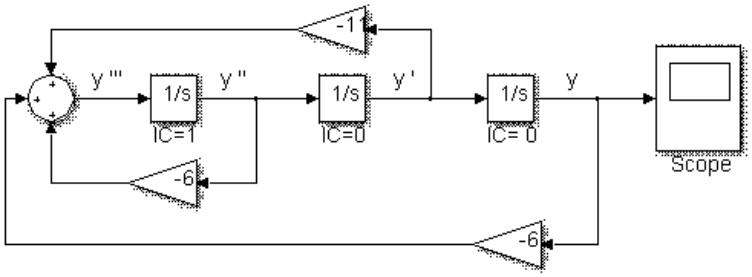


Fig. 5: Graphical model of 3rd order ODE using three integrators (1/s)

## 5. Comparision between modeling and visualisation in MATLAB and its GNU/GPL alternative

### 5.1. GNU Octave

Matlab-compatible ODE functions are provided by the `odepkg` package included in standard distribution of Octave [12]. The `lsode` function uses Hindmarsh's ODE solver and is replacement for ode23 and other ode solvers from MATLAB. Here is an example of solving ode problem (4) in Octave. The `lsode` uses the same parameters as `ode23`, but order of parameters is changed.

```
t=[0:0.1:10]';
[X, ISTATE, MSG] = lsode ('odefun3',[0,0,1],t);
plot(t,X(:,1))
```

The symbolic computation package is not included in Octave.

### 5.2. Modelica® language and Modelica libraries

Modelica [9] is a non-proprietary, object-oriented, equation based language to conveniently model complex physical systems. It has free libraries with elements used in mechanical, electrical, electronic, hydraulic, thermal, control, electric power or process-oriented models. To use Modelica, a user environment as Dymola, MapleSim, SimulationX or Vertex is needed. Another option is SimForge, entirely free graphical Modelica editor released under the terms of the GPL license.

Modelica models can be imported into Simulink using the export features of Dymola, MapleSim, SimulationX and Vertex. Industry is increasingly using Modelica for model based development. Many automotive companies, such as Audi, BMW, Daimler, Ford, Toyota, VW use Modelica to design energy efficient vehicles and/or improved air conditioning systems. Also power plant providers (ABB, EDF, Siemens) and many

other companies use Modelica. Modelica is a more advanced package and should be widely used in education. Figure 6 shows models of the same 22 kW DC motor prepared with Simulink and Modelica.

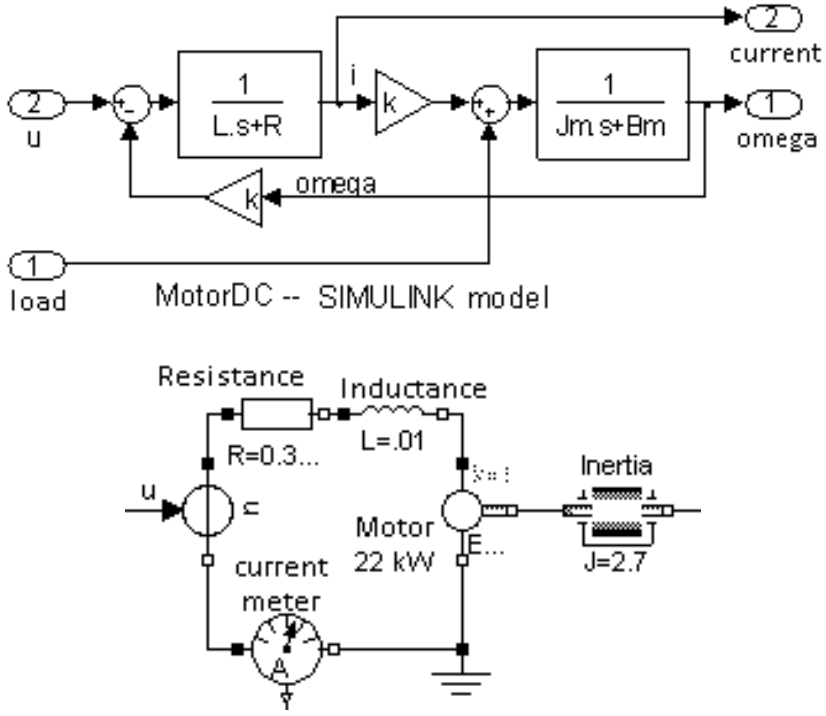


Fig. 6: Models of the 22 kW DC motor built in Simulink and Modelica. Adder, transfer functions e.g $\frac{1}{Ls+1}$ and amplifiers $k$ are used in the Simulink model. Modelica uses models of physical elements of real system.

## 6. Conclusion

Most students take the differential equations course in order to master techniques to be later applied in solving the real world problems in their profession. They should learn a few basic concepts that they will remember for the rest of their lives and they should learn how to use the available computer software to model and simulate the real-life applications of differential equations.

## References


[1] Boyce W.E, Diprima R.C, *Elementary Differential Equations and Boundary Value Problems*, (4th Edition), Wiley International, John Wiley & Sons, ISBN 0-471-83824-1, 1986.

[2] Cariaga E. A., Nualart M. C., *Teaching and learning iterative methods for solving linear systems using symbolic and numeric software*, Comput Appl Eng Educ 10, 2002, 51—58.

[3] *Interactive Differential Equations*, Pearson Addison-Wesley, *http://www.aw-bc.com/ide*, 2013.

[4] Kadijevich D., Haapasalo L., Hvorecky J., *Using technology in applications and modelling*, Teaching Mathematics and its Applications, 24(2-3), 2005, 114—122.

[5] Kwon O., Rasmussen C., *Towards an inquiry approach to undergraduate mathematics* Proceedings of the 4th East Asia Regional Conference on Mathematics Education, Penang, Malayasia, 2007, 39—46.

[6] Kwon, O., at all. *Students' Retention of Mathematical Knowledge and Skills in Differential Equations*, School Science and Mathematics, 105(5), 2005, 1—13.

[7] *MATLAB, Simulink*, *http://www.mathworks.com/products/*, 2013.

[8] Menger K., *On the Teaching of Differential Equations*, The American Mathematical Monthly, Vol. 51, No. 7 1944, 392—395.

[9] Modelica and the Modelica Association, *http://www.modelica.org*, 2014.

[10] Mrozek B., Mrozek Z., *MATLAB i Simulink. Poradnik uzytkownika*, Wydanie III, Helion, 2010.

[11] Nabb K., *Inquiry in Differential Equations: A Teacher's Reflections*, 38th AMATYC Annual Conference, *http://www.amatyc.org/Events/conferences/2012 Jacksonville/proceedings.html*, Jacksonville, Florida, 2012.

[12] Octave, *http://www.octave.org* and *http://octave.sourceforge.net/odepkg*

[13] Padir Taskin, et al., *Teaching differential equations in a diverse classroom*, 2008 Annual Conference and Exposition, AC 2008-1548, American Society for Engineering Education, 2008.

[14] Paraskakis I., *Rethinking the Teaching of Differential Equations through the Constructivism Paradigm*, In: ICALT 2003 - 2003 IEEE International Conference on Advanced Learning Technologies 9-11 July, 2003, Athens, Greece, 2003, 506—510.

[15] Rasmussen Chris, Whitehead Karen, *Learning and Teaching Ordinary Differential Equations*, http://www.maa.org/t-and-l/sampler/research-sampler.html, The Mathematical Association of America, RESEARCH SAMPLER, No. 7, January 2003.

[16] Rasmussen, at all. *Capitalizing on advances in mathematics and K-12 mathematics education in undergraduate mathematics: An inquiry-oriented approach differential equations*, Asia Pacific Education Review, 7(1), 2006, 85—93.

[17] Rota Gian-Carlo *Ten Lessons I Wish I Had Learned Before I Started Teaching Differential Equations*, *http://euler.slu.edu/Dept/Faculty/marks/Pedagogy/RotaOnTeachingDifflEqns.pdf*, 1997.

[18] *Teacher package: Differential equations, Plus Magazine*, University of Cambridge, *http://plus.maths.org/content/teacher-package-differential-equations*.

[19] Wagner J., Speer N., Rossa B., *Beyond mathematical content knowledge: A mathematician's knowledge needed for teaching an inquiry-oriented differential equations course*, Journal of Mathematical Behavior 26, 2007, 247—266.

[20] Zadawanie dobrych pytań
*http://www.ceo.org.pl/pl/cyfrowaszkola/kurs/zadawanie-dobrych-pytan*.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

JERZY RASZKA AND LECH JAMROŻ*

# MAX-PLUS LINEAR SYSTEM IN CONTROL OF DATA PROCESSING

# LINIOWY SYSTEM MAX-PLUS W STEROWANIU PROCESAMI PRZETWARZANIA DANYCH

Abstract

The increasing complexity of information processing in distributed computer systems and microprocessors requires the use of time-saving devices and extended capacities of transmission channels. Processes in computers systems need effective processing time. This article describes an application of the theory of the Max Plus Linear System (MPLS) to controlling digital information processing and transmission time in information systems. System processes are described by an MPLS state equation and an MPLS output equation. The MPLS model makes use of formal mathematical methods of max-plus algebra which include maximization and addition operations in the domain of non-negative real numbers with the addition of minus infinity. The input data and the structure of the processes under consideration are represented by the Timed Event Graph (TEG) formalism constituting a special case of Timed Petri Nets. The suggested MPLS methods are useful for investigating selected properties of network models. They may be applied, among others, to evaluate performance criteria, cycle time, predictive control etc. This article presents the theoretical considerations used to determine the input signals controlling discrete processes, which are then illustrated with examples of numerical computations.

*Keywords: max-plus algebra, Petri nets, data processing, discrete processes*

Streszczenie

Zwiększająca się złożoność procesów przetwarzania informacji w rozproszonych systemach komputerowych i mikroprocesorowych wymaga oszczędnego wykorzystania czasu pracy urządzeń i zwiększonej pojemności kanałów transmisyjnych. Procesy w systemach komputerowych potrzebują efektywnego czasu przetwarzania W niniejszym opracowaniu przedstawiono zastosowanie teorii max-plus liniowego systemu (MPLS) w sterowaniu czasem przetwarzania informacji i czasem transmisji informacji cyfrowej w systemach informatycznych. System procesów opisany jest w przestrzeni MPLS równaniem stanu i równaniem wyjścia. Model MPLS jest oparty na formalnych matematycznych metodach max plus algebry, które są wyposażone w operacje maksymalizacji i dodawania w dziedzinie nieujemnych liczb rzeczywistych rozszerzonych o minus nieskończoność. Dane wejściowe i struktura rozważanych procesów są określone przez formalizm czasowych sieci zdarzeń jako szczególnego przypadku czasowych sieci Petri'ego. Zaproponowane metody MPLS są użyteczne w badania wybranych właściwości sieci. Między innymi mogą być one zastosowane do oceny wydajności, czasu cyklu, sterowania predykcyjnego, itp. W artykule zastosowano teoretyczne rozważania określające wejściowe sygnały sterujące procesem dyskretnym oraz przedstawiono przykładowy wyniki z numerycznych obliczeń.

*Słowa kluczowe: max-plus algebra, sieć Petri'ego, przetwarzanie danych, procesy dyskretne*

*Institute of Computer Science, Cracow University of Technology; jraszka@pk.edu.pl, ljamroz@pk.edu.pl

## 1. Introduction

Technological advances in the field of computer information processing require the use of more effective methods to analyze and model these processes. The Internet evolved from a simple store-and-send network into a more complex communications infrastructure. The security and flexibility of ICT connections require extra processing (e.g. filtering, encryption and prediction) as well as resources, all of which delay signal data transmission. Furthermore, the amount of information in distributed microprocessors and computer systems is growing rapidly (as exemplified by weather information systems or communication networks). The time of both digital information processing and its transmission is increasing (computer networks might soon reach out into the interplanetary space). Consequently, processes in computer systems require effective time control. Measurements of event occurrence times are generally not as susceptible to noise as those of continuous signals, namely variables such as the temperature, speed, pressure, etc. The suggested methods are used to control and minimize the delay of output results compared to selected values by varying the starts of processing tasks and moments of data entry at different points of the studied system.

Discrete systems, and particularly discrete-event dynamic systems, often appear in the context of parallel computing, manufacturing [6] or project management systems, railway [5] or telecommunication networks etc. Recent years have seen a quantitative growth of research on discrete systems that can be modelled as max-plus linear systems. Most of the earlier literature concerning this class of systems discussed modelling, performance and properties analysis rather than control [1, 6] and so did many other authors e.g. J. Bernd Heidergott, Geert Jan Olsder, Didier Dubois, Jean-Pierre Quadrat and Jacob van der Woude. However, articles have recently been published on the control of max plus systems. The authors in [7] designed a feedback controller to guarantee that the system evolves without violating time restrictions . In [10], authors describe a bus network as a non-stationary linear state model and determine eigenvalues as well as eigenvectors used to evaluate the cycle time. Most of existing publications about controlling general discrete systems are presented by B. De Schutter [11].

The work reported in this paper is but a part of a more general study aimed at evaluating and controlling computation on multiprocessor platforms, using some classical analysis and control methods from systems theory utilizing (max, +) algebra. Mikel Cordovilla [12] has worked on the real-time execution of dependent periodic task sets on multiprocessor platforms and I. Elmahi [4] has used max-plus algebra and Petri Nets for supply chain logistics.

Petri nets (PN) [8, 14] are very popular formal mathematical methods of analyzing and representing parallel and distributed computing in concurrent systems, and they draw much attention to modelling and verifying these types of systems. P systems, also referred to as membrane systems [9], are a class of parallel and distributed computing

models [6]. The interest in linking P systems with the PN computation model has produced several important results in simulation and decidability issues. Some efforts have been made to simulate P systems with Petri nets and thus to verify many useful behavioral properties such as reachability, boundedness, liveness, terminating, etc.

With regard to manufacturing or chemical engineering processes, their behavior can often be adequately represented by a discrete event model accounting for the usually discrete sensor and the actual equipment for these processes. In addition, the behavior of these processes is often adequately described by a sequence of transitions between discrete process states. This contribution focuses on a particular class of such discrete event systems in which computational processes are synchronized and controlled. This system class has attracted significant interest in recent years because sequences of event times for such processes can be described by equations that are linear in one specific algebra, namely maxplus algebra [1]. The resulting equations are structurally equivalent to system descriptions from conventional control engineering such as transfer functions or state space models.

Hence, a system theory for these max–plus linear systems has been developed, and various concepts known well from control engineering have been adapted to this system class in the control design and diagnosis.

Timed Event Graphs (TEG's) are a subclass of timed Petri nets which can be used for modelling Discrete Event Dynamic Systems (DEDS) in which synchronization phenomena occur, such as manufacturing, multiprocessor and transportation systems in particular.

In this paper, we have introduced a deterministic Petri net model of a computational system that can be considered to be a discrete event system. Moreover, as such DEDS can be easily modelled with a subclass of Petri nets for evaluation purposes, we subsequently suggest a TEG approach to the model, analyze and control the computational process from this TEG model, and formulate a mathematical model based on max-plus algebra. In this algebra, the behavior of the DEDS can be easily described by a linear system. In the second part, we introduce a computational process model. The third part presents an overview of the max algebra theory and an analysis of a specific model. The last part contains simulation results.

## 2. Data Processing

Let us consider a data process that allows event-driven applications to take advantage of multiprocessors by running the code for event handlers in parallel. To achieve high performance, servers must overlap computation with the I/O. Programs typically achieve this overlap by using threads or events. Threaded programs usually process every request in a separate thread; while one thread block is waiting for the I/O, another thread can run. Event-based programs are structured as a collection of call-back functions which are called by the main loop when I/O events occur. Threads provide an intuitive programming model but require coordinating the access of different threads

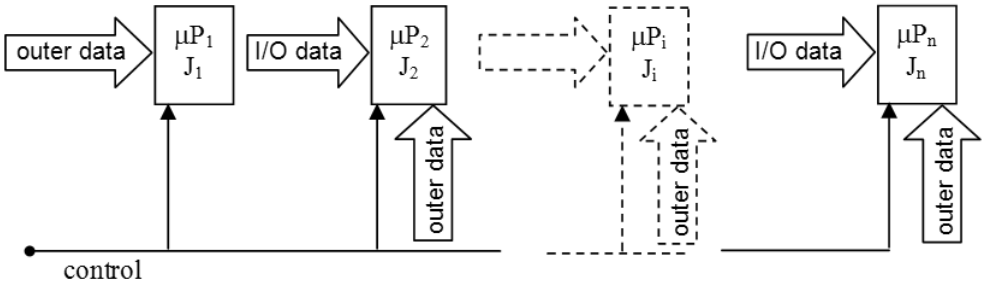


Fig. 1: The structure of the process.

to the shared state, even on a uniprocessor. Event-based programs execute call-backs sequentially so the programmer need not worry about concurrency control; however, event-based programs have so far been unable to make good use of multiprocessors. Much of the effort required to make existing event-driven programs take advantage of multiprocessors is in specifying which events can be handled in parallel.

This paper exemplifies the simple problem of designing the control of a system in which the cost is chosen so that it provides a trade-off between minimizing the delays of the end time of computational process operations (the real time to complete all the tasks in a cyclic computational process, times of final results of one cycle) and the periodicity of the desired output (the time desired or needed) to complete the process

Simple data processing consists of several tasks linked by the waiting for I/O data (Fig. 1). To illustrate our approach, let us consider a process that consists of some tasks (jobs): $J_i$ which runs on microprocessors: $\mu P_i$ for $i$=1..$n$. Each of these tasks is executed on a dedicated microprocessor. In this process, the digital information flows as input/output processing data and a control signal. Outer input data is processed as the first task on the $\mu P_1$ and its output data has to be saved to memory while it waits to be processed. The other microprocessors operate in the same way, but their input data simultaneously constitutes the output result from the previous microprocessor and may need extra outer data.

The aim of this modelling exercise is to evaluate the command of the process according to pre-established criteria. For instance, to ensure continuous computation, we have to input data at a certain specified time and have enough memory to save input, output and temporary data. A good schedule will keep the costs to a minimum and optimize the size of the memory required.

## 3. Net Representation

Petri nets, – graph-oriented formalism – facilitate modelling and analyzing systems which feature such properties as concurrency and synchronization. The Petri net model of a dynamical system consists of two parts: the net structure and marking with tokens. The net structure is a weighed-bipartite directed graph representing the static part of the system. The tokens represent the distributed overall state of the structure. This separation allows one to reason on a net based model on two levels – structural and behavioral.

The net structure is built of two disjoint sets of objects - places and transitions - which are connected by arcs. In a graphic representation, places are drawn as circles, transitions are drawn as thin bars, and arcs as arrows. Places may contain tokens which are drawn as dots. The vector representing the number of tokens in every place is the state of the Petri net and is referred to as its marking. This marking can be changed by the firing of the transitions. Petri nets do not include any notion of time and are aimed at modelling only the logical behavior of systems. The introduction of a timing specification is essential if we want to use this model class to consider the performance problem.

More formally, timed Petri nets (TPN) are 5-tuples: TPN= ($\mathbf{P,T,F,M}_0,\tau$), where $\mathbf{P}=(p_1, p_2, ..., p_n)$, $|\mathbf{P}|\neq 0$; $\mathbf{T}=(t_1, t_2, ..., t_m)$, $|\mathbf{T}|\neq 0$ is a finite, disjoint set of suitable places and transitions; $\mathbf{M}_0$**:** $\mathbf{P}\rightarrow N$ is the initial marking function which defines the initial number of tokens for every place. ($N$={0,1,...}); $\tau$ : T$\rightarrow R$ is the firing time function; and

$\mathbf{F} \subset (\mathbf{P}\text{x}\mathbf{T}) \cup (\mathbf{T}\text{x}\mathbf{P})$ is the set of arcs.

Many different models of timed Petri nets are evolved from the fundamental definition. Its include models in which the parameter of time is also assigned to the places and edges ([15] and others). In this article, the type of DEDS is represented graphically using a Timed Event Graph (TEG), is a class of timed Petri nets in which all places have only one upstream and only one downstream transition and as also mentioned, time delays. The model proposed in chapter 2 has the three microprocessors and time parameters which are represented by $\tau$ in the basic definition of Petri nets.

Fig. 2 show the TEG in which time parameters have been moved to the places and practically are described as ti, - mathematically in the next chapters that will be expressed as $t_i$. This net has three inputs $u_1$, $u_2$, and $u_3$, and one output $y$. The firing time units $u_1$, $u_2$, and $u_3$ are the start times of jobs $J_1$, $J_2$, $J_3$, respectively. Finally, the end time of the process is represented by the firing time of transition $y$.

## 4. Max-Plus Linear System

In recent years, the concept of a max-plus linear system (MPLS) has been increasingly frequently used in the literature [11]. It is based on a mathematical formalism, namely max-plus algebra. The basic operations of max-plus algebra [1] are maximization and

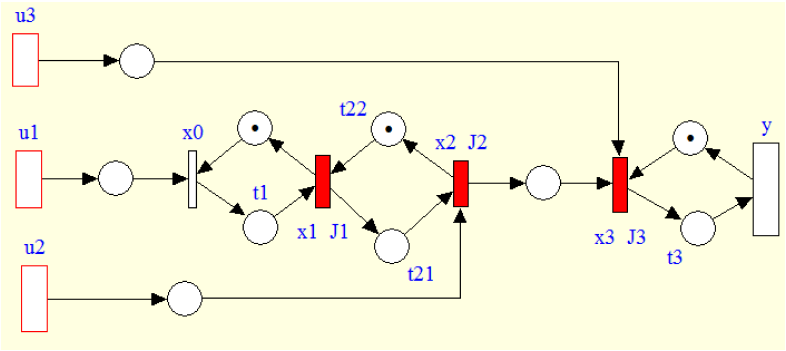


Fig. 2: Timed Event Graphs - a process model.

addition, which will be represented, respectively, by $\oplus$ and $\otimes$:$x \oplus y = \max(x, y)$ and $x \otimes y = x + y$ for

$$x, y \in R_\varepsilon\,, \quad R_\varepsilon \;=^{def}\; R \cup \{-\infty\}$$

The reason for using these symbols is that there is a remarkable analogy between $\oplus$ and conventional addition, and between $\otimes$ and conventional multiplication: many concepts and properties from linear algebra (such as the Cayley-Hamilton theorem, eigenvectors and eigenvalues, Cramer's rule,...) can be translated to max-plus algebra by replacing $+$ with $\oplus$ and $\times$ with $\otimes$. Hence we also call $\oplus$ the max-plus algebraic addition, and $\otimes$ the max-plus algebraic multiplication. Note, however, that a major difference between conventional algebra and max-plus algebra is that, in general, there are no inverse elements with respect to $\oplus$ in $R_\varepsilon$. The zero element for $\oplus$ is $\varepsilon =^{def} -\infty$ and we have $a \oplus \varepsilon = a = \varepsilon + a$ for all $a \in R_\varepsilon$. The structure $(R_\varepsilon, \oplus, \otimes)$ is referred to as max-plus algebra.

Let $r \in R$. The $r^{th}$ max-plus algebraic power of $x \in R$ is denoted by $x^{\otimes r}$ and corresponds to $rx$ in conventional algebra. If $r \in R$, then $x^{\otimes 0} = 0$ and the inverse element of $x$ w.r.t.$\otimes$ is $x^{\otimes -1} = -x$. There is no inverse element for $\varepsilon$ since $\varepsilon$ is absorbing for $\otimes$. If $r > 0$, then $\varepsilon^{\otimes r} = \varepsilon$, and if $r < 0$, then $\varepsilon^{\otimes r}$ is not defined. In this paper, we have $\varepsilon^{\otimes 0} = 0$ by definition.

The rules for the order of evaluation of max-plus algebraic operators correspond to those of conventional algebra. So the max-plus algebraic power has the highest priority, and max-plus algebraic multiplication has a higher priority than max-plus algebraic addition.

The basic max-plus algebraic operations are extended to matrices as follows. If $\mathbf{A}, \mathbf{B} \in R_\varepsilon^{m \times n}$ and $\mathbf{C} \in R_\varepsilon^{m \times p}$, then:

$$(\mathbf{A} \oplus \mathbf{B})_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij})$$
$$(\mathbf{A} \otimes \mathbf{C})_{ij} = \bigoplus_{k=1}^{n} a_{ik} \otimes c_{kj} = \max_{k=1\ldots n} (a_{ik} \otimes c_{ki})$$

for all $i, j$. Note the analogy with the definitions of the matrix sum and the product in conventional linear algebra.

The matrix $\boldsymbol{\mathcal{E}}_{m \times n}$ is the $m \times n$ max-plus algebraic zero matrix: $(\boldsymbol{\mathcal{E}}_{m \times n})_{i,j} = \varepsilon$ for all $i$, $j$; and the matrix $\mathbf{E}_n$ is the $n \times n$ max-plus algebraic identity matrix: $(\mathbf{E}_n)_{i\,i} = 0$ for all $i$ and $\mathbf{E}\,(\mathbf{E}_n)_{i\,j} = \varepsilon \quad i, j$ with $i \quad \neq \quad j$. If the size of the max-plus algebraic identity matrix or the max-plus algebraic zero matrix is not specified, it should be clear from the context. The max-plus algebraic matrix power of $\mathbf{A} \in R_{\varepsilon}^{n \times n}$ is defined as follows: $\mathbf{A}^{\otimes 0} = \mathbf{E}_n$ and $\mathbf{A}^{\otimes k} = \mathbf{A} \otimes \mathbf{A}^{\otimes(k-1)}$for $k = 1,2, \ldots$

Discrete event systems with synchronization only and no concurrency can be modeled by a max-plus algebraic model. The state equation of a max-plus linear system has the following form [1]:

$$\mathbf{x}(k) = \mathbf{A} \otimes \mathbf{x}(k-1) \quad \oplus \quad \mathbf{B} \otimes \mathbf{u}(k) \tag{1}$$
$$\mathbf{y}(k) = \mathbf{C} \otimes \mathbf{x}(k) \tag{2}$$

with $\mathbf{A} \in R_{\varepsilon}^{n \times n}$, $\mathbf{B} \in R_{\varepsilon}^{n \times m}$ , $\mathbf{C} \in R_{\varepsilon}^{l \times n}$ and

$\mathbf{x} \in R_{\varepsilon}^{m}$ −represents the state vector with the initial value $\mathbf{x}(0) = \mathbf{x}_0$,

$\mathbf{u} \in R_{\varepsilon}^{n}$ −the input vector,

$\mathbf{y} \in R_{\varepsilon}^{l}$ −the output vector,

where

$m$ - the number of inputs,

$l$−the number of outputs.

## 5. Process model

### 5.1. Investigation

The purpose of this study is to show that the process satisfying the above assumptions can be modelled in max-plus algebra to determine the input vector $\mathbf{u}$(k) for known values of $\boldsymbol{y}$(k) and to evaluate the error between the actual and the desired output.

Let $\mathbf{u}(k) = [u_l,\ u_2,\ u_3]^T$ be the input vector, $\mathbf{x}(k) = [x_l\ ,\ x_2\ ,\ x_3]^T$ the state vector and $\mathbf{y}(k)$ the output vector coinciding with the model output $y(k)$. For each transition, $x_i$ and $u_i$ are associated with the indicators $x_i(k)$ and $u_i(k)$ respectively, which correspond to the steps of the $k^{th}$ firing of transition $x_i$ (resp. $u_i$) and $y(k)$ is obtained by analogy. The state system in the max plus algebra is as follows:

For example, the $k^{th}$ firing of transition $x_1$(when $J_1$ starts) must wait for $t_1$ time units until the $k^{th}$ input data $u_1$ for task $J_1$ is ready and the $k^{th}$ input data $u_2$ is

$$
\begin{aligned}
x_1(k) &= t_1 \otimes u_1(k) \quad \oplus \quad u_2(k) \quad \oplus \quad t_{22} \otimes x_2(k-1) \quad \oplus \quad t_1 \otimes x_1(k-1) \\
x_2(k) &= t_{21} \otimes x_1(k) \\
x_3(k) &= x_2(k) \quad \oplus \quad u_3(k) \quad \oplus \quad y(k-1)
\end{aligned}
$$

provided. Thus, the linear equation of evolution in$R_\varepsilon$ of this discrete event dynamic system is as follows:

$$\mathbf{x}(k) = \mathbf{A_0} \otimes \mathbf{x}(k) \quad \oplus \quad \mathbf{A_1} \otimes \mathbf{x}(k-1) \quad \oplus \quad \mathbf{B_0} \otimes \mathbf{u}(k) \quad \oplus \quad \mathbf{D}_0 \otimes y(k-1) \tag{3}$$

where

$$\mathbf{A}_0 = \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon \\ t_{21} & \varepsilon & \varepsilon \\ \varepsilon & e & \varepsilon \end{bmatrix}, \ \mathbf{A}_1 = \begin{bmatrix} t_1 & t_{22} & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \end{bmatrix}, \ \mathbf{B}_0 = \begin{bmatrix} t_1 & e & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & e \end{bmatrix}, \ \mathbf{D}_0 = \begin{bmatrix} \varepsilon \\ \varepsilon \\ e \end{bmatrix}.$$

The solution of (3) is

$$\mathbf{x}(k) = \mathbf{A_0^*} \otimes (\mathbf{A_1} \otimes \mathbf{x}(k-1) \quad \oplus \quad \mathbf{B_0} \otimes \mathbf{u}(k) \quad \oplus \quad \mathbf{D_0} \otimes \mathbf{y}(k-1))$$

where

$$\mathbf{A}_0^* = \mathbf{E} \oplus \mathbf{A}_0^{\oplus} \mathbf{A}_0^2 \oplus \mathbf{A}_0^3 \ldots$$

$$\mathbf{A}_0^* = \begin{bmatrix} e & \varepsilon & \varepsilon \\ t_{21} & e & \varepsilon \\ t_{21} & e & e \end{bmatrix}, \qquad \mathbf{A} = \mathbf{A}_0^* \otimes \mathbf{A}_1 = \begin{bmatrix} \varepsilon & t_{22} & \varepsilon \\ \varepsilon & t_{22} \otimes t_{21} & \varepsilon \\ \varepsilon & t_{22} \otimes t_{21} & \varepsilon \end{bmatrix}.$$

$$\mathbf{B} = \mathbf{A}_0^* \otimes \mathbf{B}_0 = \begin{bmatrix} t_1 & e & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & e \end{bmatrix}, \qquad \mathbf{D} = \mathbf{A}_0^* \otimes \mathbf{D}_0 = \begin{bmatrix} \varepsilon \\ \varepsilon \\ e \end{bmatrix}.$$

Then, the model turns into:

$$\mathbf{x}(k) = \mathbf{A} \otimes \mathbf{x}(k-1) \quad \oplus \quad \mathbf{B} \otimes \mathbf{u}(k) \quad \oplus \quad \mathbf{D} \otimes y(k-1)) \tag{4}$$

Recurrence gives us the expression:

$$\mathbf{x}(k) = \mathbf{A}^{k-1} \otimes \mathbf{x}(1) \quad \oplus \quad \sum_{i=2}^{k} \mathbf{A}^{k-i} \otimes \mathbf{B} \otimes \mathbf{u}(i) \quad \oplus \quad \sum_{i=3}^{k-1} \mathbf{A}^{k-i-1} \otimes \mathbf{D} \otimes y(i) \tag{5}$$

To determine the command of the process, first we have to define the whole order-l model which describes its global behavior:

$$\begin{cases} \mathbf{x}(k) = \mathbf{A} \otimes \mathbf{x}(k-1) \ \oplus \ \ \mathbf{B} \otimes \mathbf{u}(k) \ \oplus \mathbf{D} \otimes y(k-1)\,) \\ \mathbf{y}(k) = \mathbf{C} \otimes \mathbf{x}(k) \\ \mathbf{u}(1) = \left[ \begin{array}{ccc} u_1(1) & u_2(1) & u_3(1) \end{array} \right]^{\mathbf{T}} \\ \mathbf{x}(1) = \left[ \begin{array}{ccc} x_1(1) & x_2(1) & x_3(1) \end{array} \right]^{\mathbf{T}} \end{cases} \tag{6}$$

Where $\mathbf{C} = \left[ \begin{array}{ccc} \varepsilon & \varepsilon & t_3 \end{array} \right]$, $\mathbf{u}(1)$ and $\mathbf{x}(1)$ are the initial conditions that we are going to determine below.

### 5.2. Initial conditions

To comply with the previous assumptions, we determine the initial values of $\mathbf{u}(1)$ and $\mathbf{x}(1)$ according to the end process $y(1)$.

$$y\,(1) = \ t_1 \otimes t_{21} \otimes t_3 \otimes u_1\,(1) = t_{21} \otimes t_3 \otimes u_2\,(1) = t_3 \otimes u_3\,(1)$$

Then, the initial control is:

$$\left[ \begin{array}{c} u_1(1) \\ u_2(1) \\ u_3(1) \end{array} \right] = \left[ \begin{array}{l} y(1)\,\phi\,(t_1 \otimes t_{21} \otimes t_3) \\ y(1)\,\phi\,(t_{21} \otimes t_3) \\ y(1)\,\phi\,t_3 \end{array} \right] \tag{7}$$

where "$\phi$" represents the conventional subtraction.

Consequently, the initial value of the state vector is

$$\left[ \begin{array}{c} x_1(1) \\ x_2(1) \\ x_3(1) \end{array} \right] = \left[ \begin{array}{l} t_1 \otimes u_1(1) \\ t_{21} \otimes u_2(1) \\ u_3(1) \end{array} \right]. \tag{8}$$

These two initial vectors mean that if task $\mathrm{J}_1$ begins e.g. at $t = 0$, $t_1$ time units later the data for task $\mathrm{J}_2$ ($u_2$) must be prepared and that task can begin. This ensures a good start-up without the signal of the model being delayed.

### 5.3. Procedure

As indicated before, the network operates under a schedule defined for the final result; this schedule is used to find the suitable inputs of the model. We formulate the model output more explicitly as:

$$\mathbf{y}(k) = \mathbf{C}\otimes\mathbf{A}^{k-1}\otimes\mathbf{x}(1) \oplus \sum_{i=2}^{k} \mathbf{C}\otimes\mathbf{A}^{k-i}\otimes B\otimes\mathbf{u}(i) \oplus \sum_{i=3}^{k-1} \mathbf{C}\otimes\mathbf{A}^{k-i-1}\otimes\mathbf{D}\otimes y(i) \quad (9)$$

or

$$\mathbf{y}(k) \geqslant \max\{\, \mathbf{C}\otimes\mathbf{A}^{k-1}\otimes\mathbf{x}(1)\,,\ \sum_{i=2}^{k} \mathbf{C}\otimes\mathbf{A}^{k-i}\otimes\mathbf{B}\otimes\mathbf{u}(i),\ \sum_{i=3}^{k-1} \mathbf{C}\otimes\mathbf{A}^{k-i-1}\otimes \mathbf{D}\otimes y(i)\,\}$$

which is equivalent to:

$$\mathbf{y}(k) \geqslant \mathbf{C}\otimes\mathbf{A}^{k-1}\otimes\mathbf{x}(1) \quad (10)$$

$$\mathbf{y}(k) \geqslant \sum_{i=2}^{k} \mathbf{C}\otimes\mathbf{A}^{k-i}\otimes\mathbf{B}\otimes\mathbf{u}(i) \quad (11)$$

$$\mathbf{y}(k) \geqslant \sum_{i=3}^{k-1} \mathbf{C}\otimes\mathbf{A}^{k-i-1}\otimes\mathbf{D}\otimes\mathbf{y}(i) \quad (12)$$

We are more interested in the second equation (11) which is transformed into the form below:

$$\mathbf{y}(k) = \sum_{i=2}^{k} \mathbf{C}\otimes\mathbf{A}^{k-1}\otimes\mathbf{B}\otimes\mathbf{u}(i) \quad (13)$$

It is now straightforward that from (13), we can formulate the command $\mathbf{u}(k)$ for the process if the values of the output $\mathbf{y}(k)$ are known. For all of the rest $\mathbf{y}(k)$, there will be the need to find the final result.

For $k$= 2, 3, 4, ...

$$\begin{aligned}
&\mathbf{y}(2) = \mathbf{C}\otimes\mathbf{B}\otimes\mathbf{u}(2)\\
&\mathbf{y}(3) = \mathbf{C}\otimes\mathbf{A}^1\otimes\mathbf{B}\otimes\mathbf{u}(2) \quad \oplus \quad \mathbf{C}\otimes\mathbf{B}\otimes\mathbf{u}(3)\\
&\mathbf{y}(4) = \mathbf{C}\otimes\mathbf{A}^2\otimes\mathbf{B}\otimes\mathbf{u}(2) \quad \oplus \quad \mathbf{C}\otimes\mathbf{A}^1\otimes\mathbf{B}\otimes\mathbf{u}(3) \quad \oplus \quad \mathbf{C}\otimes\mathbf{B}\otimes\mathbf{u}(4)\\
&\ldots
\end{aligned} \quad (14)$$

Since our aim is to compute $\mathbf{u}(k)$ for the specified $\mathbf{y}(k)$, we have to solve the following equation:

$$\mathbf{y}(k) = \mathbf{C}\otimes\mathbf{B}\otimes\mathbf{u}(k) \quad (15)$$

For example, to calculate $\mathbf{u}(2)$ which is the solution of $\mathbf{y}(2) \geqslant \mathbf{C}\otimes\mathbf{B}\otimes\mathbf{u}(2)$, we solve its equation form (14) and keep its smallest solution to be sure that it is also

verified in the equation. Note that we proceed by simplifying terms such as the ones in the expression of $\mathbf{y}(k)$. Indeed, these terms constitute the first condition for the desired output $y(k)$. More explicitly, if we consider that all expressions $\mathbf{y}(k)$ are equal to $\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{u}(k)$, then :

$$\mathbf{C} \otimes \mathbf{A}^2 \otimes \mathbf{B} \otimes \mathbf{u}(2) \ \oplus \ ... \ \oplus \ \mathbf{C} \otimes \mathbf{A}^1 \otimes \mathbf{B} \otimes \mathbf{u}(k-1) \ \leqslant \ \mathbf{C} \otimes \mathbf{B} \otimes \mathbf{u}(k)$$

Which means that we must have

$$y\,(k) \geqslant \ t_1 \otimes t_{21} \otimes t_3 \otimes u_1\,(k-1) \ \oplus \ t_{21} \otimes t_3 \otimes u_2\,(k-1) \ \oplus \ t_3 \otimes u_3\,(k-1)$$

### 5.4. Cyclic processing

We assume that the output of the process, i.e. the final results, will be produced at regular intervals $T_C$. We shall later see how the network reacts depending on various values of $T_C$.

Let $y(k) \geqslant T_C^k \otimes y(0)$, where $T_C$ is the periodicity of the desired output. We apply this condition in our computations in the following form:

$$T_C \geqslant \ \max(\ t_1 + t_{21} + t_3 + u_1\,(k-1)\,,\ t_{21} + t_3 + u_2\,(k-1)\ ,\ \ t_3 + u_3\,(k-1)\,)/k$$

We solve equation (13) and determine the control vector as follows:

$$u_j(k) = y(k)\,\phi\,(\mathbf{C} \otimes \mathbf{B})_{i,j}; \qquad\qquad k = 2, 3, ...; \quad i = 1 \ \text{ and } \ j = 1, 2, 3\,.$$

Where:

$$\mathbf{C}\otimes\mathbf{B} = \left[\ \varepsilon \quad \varepsilon \quad t_3\ \right] \otimes \begin{bmatrix} t_1 & e & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & \varepsilon \\ t_1 \otimes t_{21} & t_{21} & e \end{bmatrix} = \left[\ t_1 \otimes t_{21} \otimes t_3 \quad t_{21} \otimes t_3 \quad t_3\ \right]$$

More explicitly, for every $k > 1$, the general solutions are:

$$\begin{aligned} u_1(k) &= y(k)\,\phi\,(\mathbf{C} \otimes \mathbf{B})_{1,1} \\ u_2(k) &= y(k)\,\phi\,(\mathbf{C} \otimes \mathbf{B})_{1,2} \\ u_3(k) &= y(k)\,\phi\,(\mathbf{C} \otimes \mathbf{B})_{1,3} \end{aligned} \tag{16}$$

These equations determine appropriate control of the modelled process. On the other hand, (10) and (12) contain two constraints of the desired outputs of the

system. At the same time, $T_C$ is periodic in relation to these outputs, $y(k) \geqslant T_C^{k-1} \otimes y(1) = T_C^k$. On the basis of (10) and this assumption, we conclude that:

$$y(k) = T_C^k \geqslant \mathbf{C} \otimes \mathbf{A}^{k-1} \otimes \mathbf{x}(1) \quad \text{or} \quad T_C^k \geqslant t_3 \otimes (t_1 \otimes t_{21})^{k-1} \otimes x_2(1)$$

In practice, $T_C \geqslant (\, t_3 + (t_1 + t_{21})(k-1) + x_2(1)\,)/k$ means that the periodicity must be superior at a certain value in order to have good control.

The subsequent constraint of $T_C$ results from (14):

$y(k) = T_C^k \geqslant \sum_{i=3}^{k-1} \mathbf{C} \otimes \mathbf{A}^{k-i-1} \otimes \mathbf{D} \otimes y(i)$, this constraint is continuously verified since the product $\mathbf{C} \otimes \mathbf{A}^{k-i-1} \otimes \mathbf{D}$ is null.

To conclude this section, we have introduced the following steps:

1- Determine state equations in max-plus algebra as (6).

2- Calculate the global recurrence equation of the linear evolution of the system.

3- Determine initial conditions (7,8).

4- Calculate constraints of the desired model output.

5- Calculate the control vector (16).

## 6. Simulation and Results

In order to simulate the process model (Fig. 2) for $t_1 = 3$, $t_{21} = 5$, $t_{22} = 5$ and $t_3 = 2$, we apply a fixed interval $T_C$ to the desired outputs, which are: $y'(k) = y'(k-1) \otimes T_C$; $k = 2, 3, \ldots$. Using (16), we can compute vectors $u(k)$ and consider the initial values. Figures 3 - 4 show the time evolution of the interpolated values of activity counts - particularly the desired $y'(t)$ and real $y(t)$ outputs represented by the signs ▼ and ▲, respectively.

All example results are obtained for various values of $T_C$. In the first examples, the periodicity of computations equals 10 and 15 ( Fig. 3 and Fig. 4, respectively). The $T_C$ value is large enough to contain all tasks of the process and no error occurs between the desired and actual outputs.

## 7. Conclusion

Engineers who build discrete event systems have to confront dynamic problems as a matter of fact. In particular, even though dynamical systems are well understood based on classical methods, they have had little mathematical support for this. This article suggests and introduces a max-plus linear system: a methodology applied to modeling and simulating discrete event processes. The control of process tasks in a simple multi-processor computational system is considered as an example. This, however, is only a small part of the studies carried out, which require additional testing and even broader-ranging experience, especially in terms of practical applications.

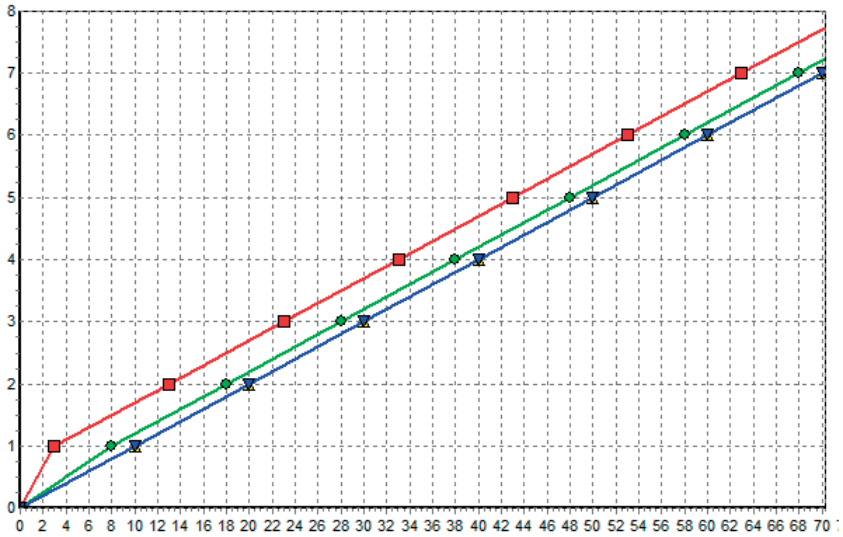


Fig. 3: Activity count of: y' ▼, y ▲, $x_1$ ■, $x_2$ ● for $T_C = 10$.

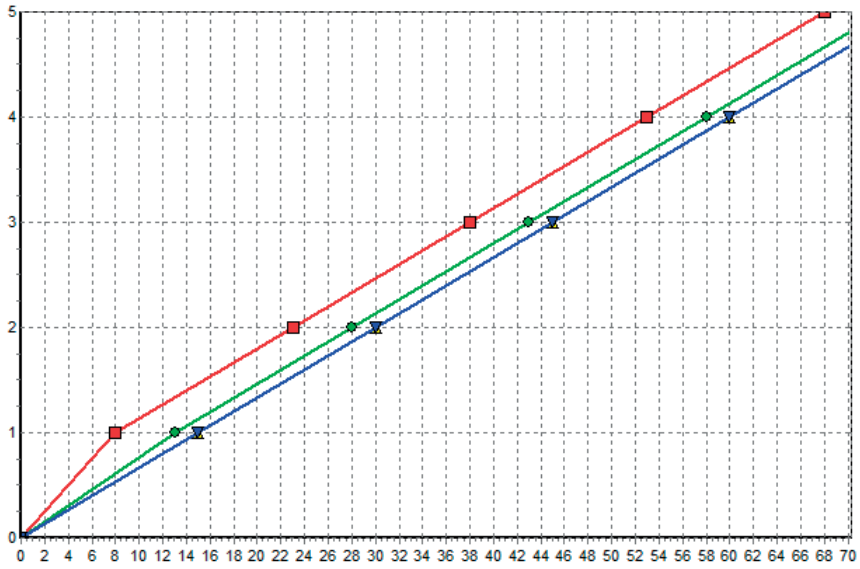


Fig. 4: Activity count of: y' ▼, y ▲, $x_1$ ■, $x_2$ ● for $T_C = 15$.

Furthermore, there is a need to develop a way of controlling these processes and the analyzed conditions for the periodicity of the required results. Further research will include applying the system to larger models and improving the efficiency of the optimization procedure. Future research will focus on design methods, synthesizing process control with output and/or state feedback and using models for predictive and adaptive control.

## References


[1] Bacceli F., Cohen G., Olsder G., Quadrat J., *Synchronization and Linearity. An Algebra for Discrete Event Systems*, London, John Wiley & Sons Ltd, 1992.

[2] Balduzzi F., Giua Has., Menga G., *First-order hybrid Petri net: In model for optimisation and control*, IEEE Trans. On Rob. And Aut. 1 6 (4), 2000 382—399.

[3] Cassandras Ch., Lafortune St., *Introduction to Discrete Event Systems*, Springer, Kluwer Academic Publishers, 2008.

[4] Elmahi I., Grunder O., Elmoudni A., *A max plus algebra approach for modeling and control of lots delivery.* Industrial Technology, IEEE ICIT '04 Vol. 2, 8-10 Dec. 2004, 926—931.

[5] Goverde Rob M.P., *Railway timetable stability analysis using max-plus system theory*, Elsevier, Transportation Research Part B 41, 2007, p. 179–201

[6] Jamroż L., Raszka J., *Simulation method for the performance evaluation of system of discrete cyclic processes.* 16-th IASTED International Conference on Modelling, Identification and Control, Innsbruck, Austria, 17-19.02.1997, 190—193

[7] Maia C.A., Andrade C.R., Hardouin L., *On the control of max-plus linear system subject to state restriction*, Automatica, Volume 47, Issue 5, May 2011, 988—992.

[8] Murata T., *Petri nets: properties, analysis and applications.* Proceedings of the IEEE, vol. 77, no. 4, p.541-580, 1989.

[9] Guţuleac E., Balmuş I. et alt., *Descriptive Timed Membrane Petri Nets for Modelling of Parallel Computing,* International Journal of Computers, Communications & Control, Vol. I No. 3, 2006, 33—39.

[10] Nait-Sidi-Moh A., Manier M.-A., El Moudni A., *Spectral analysis for performance evaluation in a bus network*, European Journal of Operational Research Volume 193: Issue 1, 16 February 2009, 289—302.

[11] De Schutter B., van den Boom T., *Model predictive control for max-plus linear discrete event systems.* Automatica 37(7), July 2001, 1049—1056.

[12] Cordovilla M., Boniol F., et alt., *Off-line Optimal Multiprocessor Scheduling of Dependent Periodic Tasks,* Journées FAC'2011, Formalisation des Activités Concurrentes, 6 et 7 avril 2011- LAAS – CNRS, 2011.

[13] Iwaniak M., Khadzhynov W., *Usage of Petri nets for distributed transactions modeling* Studia Informatica vol. 33, no 105, Silesian University of Technology, 2012.

[14] Szpyrka M., *Sieci Petriego w modelowaniu i analizie systemów współbieżnych*, ISBN 9789788304333, WNT Warszawa, 2008.

[15] Coolahan J.E.jr, Roussopoulos N., *Timing Requirements for Time-Driven Systems Using Augmented Petri Nets*, IEEE Trans. on Software Eng. v. SE-9, no. 5, 1983, 603—616.

TECHNICAL TRANSACTIONS | CZASOPISMO TECHNICZNE
FUNDAMENTAL SCIENCES | NAUKI PODSTAWOWE
2-NP/2014

BARTOSZ STAWIARSKI*

# DATA-DRIVEN SCORE TEST OF FIT FOR CLASS OF GARCH MODELS

# ADAPTACYJNY TEST ZGODNOŚCI DLA KLASY MODELI GARCH

## Abstract

A data-driven score test of fit for testing the conditional distribution within the class of stationary GARCH($p$,$q$) models is presented. In this paper extension of the complete results obtained by Inglot and Stawiarski in [7], as well as in Stawiarski [15] for the parsimonious GARCH(1,1) case is proposed. The null (composite) hypothesis subject to testing asserts that the innovations distribution, determining the GARCH conditional distribution, belongs to the specified parametric family. Generalized Error Distribution (called also Exponential Power) seems of special practical value.

Applying the pioneer idea of Neyman [13] dating back to 1937, in combination with dimension selection device proposed by Ledwina [10] in 1994, lead to derivation uf the efficient score statistic and its data-driven version for this testing problem. In the case of GARCH(1,1) model both the asymptotic null distribution of the score statistic has been already established in [7] and [15], together with the asymptotics of the data-driven test statistic with appropriately regular estimators plugged in place of nuisance parameters. Main results are only stated herewith, while for detailed proofs inspection and power simulations, ample reference to these papers is provided. We show that the test derivation and asymptotic results carry over to stationary ARCH($q$) models for any $q \in N$. Moreover, thanks to ARCH($\infty$) representation of the GARCH($p$,$q$) model, the test can asymptotically encompass the full GARCH family, which as a final result provides the flexible testing tool in the GARCH($p$, $q$) framework.

*Keywords: GARCH model, Neyman smooth test, data-driven score test of fit, Generalized Error Distribution*

## Streszczenie

W pracy przedstawiono adaptacyjny test zgodności dla testowania warunkowego rozkładu w klasie stacjonarnych modeli GARCH($p$,$q$). Jest to rozszerzenie kompletnych wyników uzyskanych w pracach [7] oraz [15] dla przypadku mniej rozbudowanego modelu GARCH(1,1). Podlegająca testowaniu (złożona) hipoteza zerowa postuluje, że rozkład szumu determinujący warunkowy rozkład szeregu GARCH, należy do określonej rodziny parametrycznej. Szczególne znaczenie w kontekście zastosowań ma klasa rozkładów GED.

Zastosowanie pionierskiego pomysłu Neymana z 1937 r. [13] w połączeniu z kryterium wyboru wymiaru, zaproponowanym przez Ledwinę w 1994 r. w pracy [10] pozwala wyprowadzić dla omawianego zagadnienia testowego efektywną statystykę wynikową i jej adaptacyjną wersję. W przypadku modelu GARCH(1,1) zarówno asymptotyczny rozkład statystyki przy hipotezie zerowej, jak i asymptotyka jej adaptacyjnej wersji z odpowiednio regularnymi estymatorami parametrów zakłócających zostały już uzyskane w [7] oraz [15]. Główne wyniki są tu tylko przywołane z przywołaniem licznych referencji do tych prac. Pokazano, że konstrukcja testu i wyniki asymptotyczne przenoszą się na stacjonarne modele ARCH($q$) dla dowolnego $q \in N$. Ponadto dzięki reprezentacji modeli GARCH poprzez ARCH($\infty$) test pozwala asymptotycznie objąć całą klasę GARCH, co w ostateczności daje elastyczne narzędzie testowe dla modeli GARCH($p$,$q$).

*Słowa kluczowe: Model GARCH, gładki test Neymana, adaptacyjny test zgodności, rozkład GED*

*Institute of Mathematics, Cracow University of Technology; bstawiarski@pk.edu.pl

## 1. Introduction

The Autoregressive Conditionally Heteroscedastic class of time series models, ARCH, was introduced by Engle in 1982 and four years later extended to GARCH by Bollerslev in [3]. Modelling financial time series was main empirical motivation standing behind introduction of such models, allowing for fluctuation of their conditional variance. Throughout more than two decades of research, vast theoretical and computational research concerning GARCH models has been done. A good deal of summaries or even books dedicated specifically to this class of time series has appeared, e.g. Francq and Zakoïan (2010), [5].

Widely exploited in the early days, conditionally normal GARCH models were soon questioned because of unsatisfactory fitting to econometric data. Therefore, the problem of checking conditional distribution assumptions in nonlinear time series has become vital. Apart from ad-hoc imposed innovations distribution ($t$, $\alpha$-stable or Laplace), several constructive tests have been proposed only in early 2000's. Specifically, Chen in 2002 [4] proposed a characteristic function based test, while Bai in 2003 [1] presented a martingale transform approach to testing conditional distribution of some dynamic models. Inglot and Stawiarski [7] reconsidered - in the GARCH(1,1) time series context - the idea of smooth tests conceived by Neyman [13] and their data-driven versions devised by Ledwina in [10]. A data-driven score test of fit for conditional distribution for the simple hypothesis i.e. when the null density is fully specified, was derived. Just as the composite hypothesis case for *i.i.d.* random variables was considered in [6] and [8], Stawiarski [15] extended that result to the composite hypothesis case in the GARCH(1,1) framework. This was also desirable from the applicational point of view as testing conditional distribution for financial time series should naturally allow for flexible, parametric families.

In Sections 2 and 3 of this paper we rather succinctly refer the test construction and quote main theoretical results determining the asymptotic behavior of the test statistic. We also briefly report the test power simulation study in Section 4. Detailed theoretical derivations and proofs and simulation studies, due to capacity constraints, can be found in [7] and [15]. The substantially new results are reported in Section 5, which contains extension of the proposed test to the case of ARCH($q$), then GARCH($p$,$q$) models. Main theoretical results carry over from the two papers due to general properties such as imposed stationarity (strict and weak), but the present general GARCH context results in some differences of technical nature (ampler parametrization influencing the final test statistic). Therefore the derivation of complete theoretical results leading to fully-fledged data-driven score test of fit for the whole class of stationary GARCH models calls for some future refinements at several stages. Some power simulation study can be a potential subject of further interest. We conclude the paper with Appendix and discussion concerning the above-mentioned possible paths of future research.

## 2. Score test of fit for composite hypothesis in the GARCH(1,1) model

In [7] and [15] the following GARCH(1,1) time series model $\{X_t\}_{t\in Z}$, introduced by Bollerslev [3], was the object of research:

$$\begin{cases} X_t = \sqrt{h_t}\varepsilon_t \\ h_t = \omega + \alpha X_{t-1}^2 + \beta h_{t-1} \end{cases} \tag{2.1}$$

where $\vartheta = [\omega, \alpha, \beta]^T$ is a column vector of model parameters and $\{\varepsilon_t\}_{t\in Z}$ is a sequence of *i.i.d.* random variables, satisfying $E\varepsilon_t = 0$, $Var\varepsilon_t = 1$. Denoting by $\mathcal{F}_t = \sigma\{..., \varepsilon_{t-1}, \varepsilon_t\}$ the process filtration up to time $t$, it is evident that the conditional variance $h_t$ is $\mathcal{F}_{t-1}$–measurable. It is assumed hereafter that $\vartheta \in \Theta = \{[\omega, \alpha, \beta]^T : \omega, \alpha, \beta > 0;\ \alpha+\beta < 1\}$, which ensures weak and strict stationarity as well as ergodicity of $X_t$ (see e.g. [3]).

The null hypothesis in [7] asserted that the zero-mean and unit-variance innovations $\varepsilon_t$ have a fully specified unknown density $f(x)$ on $R$. Stawiarski [15] allowed for extension to a parametric family of noise distribution densities, namely $\mathcal{G} = \{f(x, \lambda) : \lambda \in \Lambda,\ \Lambda \subset R^m\}$ ($\Lambda$ - an open set), satisfying $\int\limits_R xf(x, \lambda)dx = 0$, $\int\limits_R x^2 f(x, \lambda)dx = 1$ for all $\lambda \in \Lambda$.

Given a finite "data set" of observations $X_{(n)} = (X_1, ..., X_n)$ of the process $\{X_t\}_{t\in Z}$ obeying (2.1), the hypothesis we consider is as follows:

$$H_0\text{: } \varepsilon_t \text{ 's have a density } f(x, \lambda) \text{ belonging to } \mathcal{G}\ ;\ \vartheta \in \Theta.$$

Exploiting the concept of "smooth tests" conceived in [13], we restate the above null hypothesis in the equivalent, parametric form, subject to testing. To this end, choose a natural number $k$ and $\Phi(x) = [\Phi_1(x), ..., \Phi_k(x)]^T$– a vector of bounded orthonormal functions in $L_2[0, 1]$ satisfying $\int\limits_0^1 \Phi(x)dx = 0$. Denoting by$F(x, \lambda)$ the *cdf* of $f(x, \lambda)$, we immerse the hypothetical density into a $k-$parametric exponential family

$$\exp\{\tau^T \Phi(F(x, \lambda)) - C_k(\tau)\} \cdot f(x, \lambda) \tag{2.2}$$

where $\tau = [\tau_1, ..., \tau_k]^T \in R^k$ and $C_k(\tau)$ - the normalizing constant.

Let $\eta$=$[\vartheta^T$,$\lambda^T]^T \in \Theta \times \Lambda \subset R^{3+m}$ denote a vector of all nuisance parameters appearing in this testing problem, stemming from the GARCH(1,1) model and density $f$, respectively. Hence we get the parametric reformulation of the above hypothesis:

$$H_0^* : \tau = 0;\ \eta \in \Theta \times \Lambda. \tag{2.3}$$

For notational simplicity, we shall denote by $\varepsilon$ the r.v. distributed as $\varepsilon_t$, i.e. under $H_0^*$ with true values of all parameters. Densities (2.2) no longer have zero mean or unit variance. Still, small values of $||\tau||$ accounting for moderate departures from the null density are of major interest from the perspective of the test sensitivity.

Let us quote basic assumptions – being rather standard regularity conditions – imposed on the family $\mathcal{G}$ in our testing problem. For every $\lambda \in \Lambda$:

(A1) $\frac{\partial \log f(x,\lambda)}{\partial \lambda}$ exists and is continuous with respect to $\lambda$ for almost all $x \in R$;

Fisher information matrix $I(\lambda) = E\left\{\frac{\partial \log f(\varepsilon,\lambda)}{\partial \lambda}\frac{\partial \log f(\varepsilon,\lambda)}{\partial \lambda^T}\right\}$ is well-defined, continuous w.r.t. $\lambda$ and nonsingular;

(A2) $f(\bullet,\lambda)$ is absolutely continuous on $R$ and the function $\varsigma(x,\lambda) = x\frac{\partial \log f(x,\lambda)}{\partial x} + 1$ defined on the set $\{x : f(x,\lambda) > 0\}$ is not almost everywhere constant;

(A3) the following set of functions is linearly independent:

$$\left\{\Phi_1(F(x,\lambda)), ..., \Phi_k(F(x,\lambda)), \frac{\partial \log f(x,\lambda)}{\partial \lambda_1}, ..., \frac{\partial \log f(x,\lambda)}{\partial \lambda_m}\right\}.$$

(A4) $E\left|\varsigma(\varepsilon,\lambda)\right|^3 < \infty$ and $E\left\|\frac{\partial \log f(\varepsilon,\lambda)}{\partial \lambda}\right\|^3 < \infty$.

Fundamental for our derivation will be the following representation of the conditional variance obtained from (2.1) by successive iteration for $t = 1, 2, ..., n$

$$h_t = \omega\sum_{s=0}^{t-2}\beta^s + \alpha\sum_{s=0}^{t-2}\beta^s X_{t-1-s}^2 + \beta^{t-1}h_1. \tag{2.4}$$

Thus the conditional variance $h_t$ is expressed in terms of the observed sample path $X_{(n)}$. From now on, all calculations will be carried out conditionally on $h_1 = h$ with $h > \omega(1-\beta)^{-1}$, which establishes the link between the infinite past and the present of the process. The constant $h$ is assumed given and its influence asymptotically vanishes due to the exponentially decaying memory of the series. Let $P_h$ be the probability on the $\sigma$-field $\sigma(X_1, X_2, ...)$ induced by the family of conditional distributions of $(X_1, ..., X_n)$, $n = 1, 2, ...$, conditionally on $h_1 = h$ and under the true value of $\eta$. Accordingly, $E_h$ will stand for respective expectation.

Following the lines in [15] we obtain the log-likelihood $L_k(X_1, ..., X_n; \tau, \vartheta, \lambda)$ of $X_{(n)}$, which further is employed in calculating a score vector $\ell = \ell(\eta)$, defined as derivatives of $L_k$ with respect to all parameters involved, evaluated under $H_0^*$:

$$\ell(\eta) = \left[\left(\frac{\partial L_k(X_{(n)};\tau,\vartheta,\lambda)}{\partial \tau}\right)^T \left(\frac{\partial L_k(X_{(n)};\tau,\vartheta,\lambda)}{\partial \vartheta}\right)^T \left(\frac{\partial L_k(X_{(n)};\tau,\vartheta,\lambda)}{\partial \lambda}\right)^T\right]^T\Bigg|_{\tau=0}. \tag{2.5}$$

More explicitly, the constituent vectors – respectively: $k$-, 3-, and $m$-dimensional – are as follows

$$\begin{aligned}
\ell_\tau &= \tfrac{\partial L_k}{\partial \tau}\big|_{\tau=0} = \sum_{t=1}^{n} \Phi(F(X_t/\sqrt{Q_t},\lambda)),\\
\ell_\vartheta &= \tfrac{\partial L_k}{\partial \vartheta}\big|_{\tau=0} = -\tfrac{1}{2}\sum_{t=2}^{n} \tfrac{\varsigma(X_t/\sqrt{Q_t},\lambda)}{Q_t}\tfrac{\partial Q_t}{\partial \vartheta},\\
\ell_\lambda &= \tfrac{\partial L_k}{\partial \lambda}\big|_{\tau=0} = \sum_{t=1}^{n} \tfrac{\partial \log f(X_t/\sqrt{Q_t},\lambda)}{\partial \lambda},
\end{aligned} \tag{2.6}$$

where $Q_1 = h$, and for $t \geqslant 2$

$$\begin{aligned}
Q_t &= Q_t(X_1,...,X_{t-1};h,\vartheta) = \omega\sum_{s=0}^{t-2}\beta^s + \alpha\sum_{s=0}^{t-2}\beta^s X_{t-1-s}^2 + \beta^{t-1}h\ ,\\
\tfrac{\partial Q_t}{\partial \vartheta} &= \left[\sum_{s=0}^{t-2}\beta^s \ \sum_{s=0}^{t-2}\beta^s X_{t-1-s}^2 \ \sum_{s=1}^{t-2} s\beta^{s-1}(\omega+\alpha X_{t-1-s}^2) + (t-1)\beta^{t-2}h\right]^T .
\end{aligned} \tag{2.7}$$

Now, in the form of brief remarks we cite three results concerning the elementary properties of $\ell(\eta)$, holding true for $h > \omega(1-\beta)^{-1}$ and $t \geqslant 1$, cf. Proposition 3.1 – 3.3 in [15].

For $\{X_t\}_{t\in Z}$ following (2.1) random variables $\tilde{\varepsilon}_t = X_t/\sqrt{Q_t}$ are *i.i.d.* under $P_h$ and have the same distribution as $\varepsilon_t$'s. Under former assumptions (A1) – (A4), $E_h\ell(\vartheta,\lambda) = 0$ and $E_h||\ell(\vartheta,\lambda)||^2 < \infty$, where $||\cdot||$ denotes the Euclidean norm; moreover, the components of the score vector given by (2.6) are linearly independent random variables in $L_2(P_h)$ for any $\eta \in \Theta\times\Lambda$.

Determining the covariance matrix $\tilde{B}^{(n)}(\eta)$ of the normalized score vector $n^{-1/2}\ell(\eta)$ is another step. Thanks to (2.6) and by orthonormality of the system $\Phi$, $\tilde{B}^{(n)}(\eta)$ admits the block representation

$$\tilde{B}^{(n)}(\eta) = n^{-1}E_h\{\ell(\eta)(\ell(\eta))^T\} = n^{-1}E_h\begin{bmatrix} \ell_\tau\ell_\tau^T & \ell_\tau\ell_\eta^T \\ \ell_\eta\ell_\tau^T & \ell_\eta\ell_\eta^T \end{bmatrix} = \begin{bmatrix} I_k & \tilde{B}_{12}^{(n)}(\eta) \\ \tilde{B}_{21}^{(n)}(\eta) & \tilde{B}_{22}^{(n)}(\eta) \end{bmatrix}. \tag{2.8}$$

Obviously, $\tilde{B}_{21}^{(n)}(\eta) = [\tilde{B}_{12}^{(n)}(\eta)]^T$and the block matrix $\tilde{B}_{22}^{(n)}(\eta)$ is invertible for any $n$ and $\eta \in \Theta\times\Lambda$, by Proposition 3.3 in [15]. Explicit forms of $\tilde{B}_{12}^{(n)}(\eta)$ and $\tilde{B}_{22}^{(n)}(\eta)$ are given there in (6.7) and (6.12), respectively.

An efficient score vector $\ell^*(\eta)$ is defined as the residual of the orthogonal projection in $L_2(P_h)$ of $n^{-1/2}\ell_\tau$ upon the subspace generated by components of $\ell_\eta$. Standard theorems, cf. [14], imply that $\ell^*(\eta) = n^{-1/2}(\ell_\tau - \tilde{B}_{12}^{(n)}(\eta)[\tilde{B}_{22}^{(n)}(\eta)]^{-1}\ell_\eta)$. This leads to the explicit formula for $\ell^*(\eta)$, which we state in the following proposition for the sake of convenient reference.

**Proposition 2.1.** *[Proposition 3.4 in [15]]. Suppose $\{X_t\}_{t\in Z}$ obeys (2.1) and (A1)–(A4) are satisfied. Then for any $h > \omega(1-\beta)^{-1}$ the efficient score vector $\ell^*(\eta)$ for testing $H_0^*$ has the form*

$$\ell^*(\eta) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \Phi(F(\tilde{\varepsilon}_t, \lambda)) - \frac{1}{\sqrt{n}} \tilde{B}_{12}^{(n)}(\eta)[\tilde{B}_{22}^{(n)}(\eta)]^{-1} \begin{bmatrix} -\frac{1}{2} \sum_{t=2}^{n} \varsigma(\tilde{\varepsilon}_t, \lambda) \frac{1}{Q_t} \frac{\partial Q_t}{\partial \vartheta} \\ \sum_{t=1}^{n} \frac{\partial \log f(\tilde{\varepsilon}_t, \lambda)}{\partial \lambda} \end{bmatrix} \tag{2.9}$$

*and its covariance matrix under $P_h$ is given by*

$$\tilde{M} = \tilde{M}^{(n)}(\eta) = I_k - \tilde{B}_{12}^{(n)}(\eta)[\tilde{B}_{22}^{(n)}(\eta)]^{-1} \tilde{B}_{21}^{(n)}(\eta). \tag{2.10}$$

The martingale-difference array structure of the vector $\ell^*(\eta)$ in (2.9) paves the way for main limit result, leading in the sequel to establishing the asymptotics of our score test statistic. We quote verbatim Theorem 3.5 from [15] and in Appendix below we provide the outline of its proof.

**Theorem 2.2.** *Suppose $\{X_t\}_{t \in Z}$ obeys (2.1) and (A1)–(A4) are satisfied. Then for almost every $h > \omega(1-\beta)^{-1}$ (with respect to the Lebesgue measure) it holds*

$$[\tilde{M}^{(n)}(\eta)]^{-1/2} \ell^*(\eta) \xrightarrow{D} N(0, I_k) \tag{2.11}$$

*under $P_h$ in $R^k$ as $n \to \infty$.*

Now, for fixed natural $k$ introduce a score statistic being a quadratic form

$$W_k = W_k(\eta) = \; ||[\tilde{M}^{(n)}(\eta)]^{-1/2} \ell^*(\eta)||^2 = (\ell^*(\eta))^T [\tilde{M}^{(n)}(\eta)]^{-1} \ell^*(\eta). \tag{2.12}$$

Hence, as a direct implication of the above theorem we get, under its assumptions,

$$W_k \xrightarrow{D} \chi_k^2 \tag{2.13}$$

under $P_h$ as $n \to \infty$, where $\chi_k^2$ is a central chi-square random variable with $k$ degrees of freedom.

Now, let us proceed to define a dimension selection rule. Choose fixed $K \geqslant 1$ - a maximal dimension of the exponential family (2.2) built on $\mathcal{G}$. Define a selection rule $S(\eta)$ as

$$S(\eta) = \min\{k : 1 \leqslant k \leqslant K, \; W_k(\eta) - k \log n \geqslant W_j(\eta) - j \log n \; \forall \; j = 1, ..., K\}. \tag{2.14}$$

Thus, the resulting data-driven score statistic is $W_{S(\eta)}(\eta)$. Since $K$ is fixed, (2.13) implies that $P_h(S(\eta) = 1) \to 1$ as $n \to \infty$ (cf. Section 3.3 in [7]). Accordingly, we obtain the asymptotic behaviour of $W_{S(\eta)}(\eta)$ under $H_0^*$. Therefore, under the assumptions of Theorem 2.2 it holds

$$W_{S(\eta)} \xrightarrow{D} \chi^2_1 \tag{2.15}$$

under $P_h$ as $n \to \infty$, where $S(\eta)$ is given by (2.14).

Still, $W_{S(\eta)}$ is not suitable for testing as it depends on an unknown nuisance parameter. In the following subsection we will focus on the estimated test statistic $\hat{W}_{\hat{S}}$ with a square-root consistent estimator $\hat{\eta}$ instead of $\eta$, briefly recalling main results.

## 3. Data-driven test statistic and its asymptotics

Let $\hat{\eta} = [\hat{\vartheta}^T,\ \hat{\lambda}^T]^T$ be a square-root consistent estimator of the nuisance parameter $\eta = [\vartheta^T,\ \lambda^T]^T$. For $t \geqslant 2$ we set $\hat{\varepsilon}_t = X_t \Big/ \sqrt{\hat{Q}_t}$, $\hat{Q}_t = Q_t|_{\vartheta=\hat{\vartheta}}$ and $\frac{\partial \hat{Q}_t}{\partial \vartheta} = \frac{\partial Q_t}{\partial \vartheta}\Big|_{\vartheta=\hat{\vartheta}}$. Suppose that $\hat{B}$is a consistent estimator of $\tilde{B}^{(n)}(\eta)$, which naturally implies consistency of the block matrices estimators $\hat{B}_{12}, \hat{B}_{22}$ for $\tilde{B}^{(n)}_{12}(\eta)$ and $\tilde{B}^{(n)}_{22}(\eta)$, respectively. Then the estimated efficient score vector $\hat{\ell}^*(\hat{\eta})$, with $\hat{\eta}$ and $\hat{B}$ plugged into it, is as follows

$$\hat{\ell}^*(\hat{\eta}) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \Phi(F(\hat{\varepsilon}_t, \hat{\lambda})) - \frac{1}{\sqrt{n}} \hat{B}_{12} \hat{B}_{22}^{-1} \begin{bmatrix} -\frac{1}{2} \sum\limits_{t=2}^{n} \varsigma(\hat{\varepsilon}_t, \hat{\lambda}) \frac{1}{\hat{Q}_t} \frac{\partial \hat{Q}_t}{\partial \vartheta} \\ \sum\limits_{t=1}^{n} \frac{\partial \log f(\hat{\varepsilon}_t, \hat{\lambda})}{\partial \lambda} \end{bmatrix} \tag{3.1}$$

while

$$\hat{M} = I_k - \hat{B}_{12} \hat{B}_{22}^{-1} \hat{B}_{21} \tag{3.2}$$

is a consistent estimator of the covariance matrix $\tilde{M}$.

In order to proceed, some further theoretical assumptions in addition to (A1)–(A4) on our model have to be imposed. These are listed precisely as (A6)–(A13) in [15] and, besides the consistency of nuisance estimators, they concern the distribution of $\varepsilon$ - r.v. with the density $f(x, \lambda)$ and appropriate regularity conditions for $f$ itself, as well as the system $\Phi$. Specifically it is worth mentioning that (A5) requires $E|\varepsilon|^{2\kappa} < \infty$ for some $\kappa > 2$.

The main limit theorem being a counterpart of Theorem 2.2 is stated below, but for its lengthy proof we refer to Section 8 in [15].

**Theorem 3.1.** *[Theorem 4.1 in [15]] Under assumptions (A1)–(A13) in [15] for almost every $h > \omega(1-\beta)^{-1}$ the following asymptotic result holds*

$$\hat{M}^{-1/2} \hat{\ell}^*(\hat{\eta}_*) \xrightarrow{D} N(0, I_k) \tag{3.3}$$

*under $P_h$ in $R^k$ as $n \to \infty$, where $\hat{\ell}^*(\hat{\eta}_*)$ is parallel to $\hat{\ell}^*(\hat{\eta})$ from (3.1) with discretized version $\hat{\eta}_*$ of the estimator $\hat{\eta}$ and $\hat{M}$ as in (3.2).*

Consequently, the estimated score statistic corresponding to (2.12) takes the form

$$\hat{W}_k = \hat{W}_k(\hat{\eta}_*) = \ ||\hat{M}^{-1/2}\hat{\ell}^*(\hat{\eta}_*)||^2 = (\hat{\ell}^*(\hat{\eta}_*))^T\hat{M}^{-1}\hat{\ell}^*(\hat{\eta}_*), \tag{3.4}$$

while the ready-to-perform in practice data-driven score test statistic is just $\hat{W}_{\hat{S}}$ with

$$\hat{S} = \hat{S}(\hat{\eta}_*) = \min\{k : 1 \leqslant k \leqslant K,\ \hat{W}_k - k\log n \geqslant \hat{W}_j - j\log n\ \forall\ j = 1, ..., K\}.$$

Hence, as far as the testing is concerned, we directly obtain the result of major importance:

**Proposition 3.2.** *Under the assumptions of Theorem 3.1 it holds*

$$\hat{W}_k(\hat{\eta}_*) \xrightarrow{D} \chi^2_k \ \textit{and}\ \hat{W}_{\hat{S}} \xrightarrow{D} \chi^2_1 \tag{3.5}$$

*under* $P_h$ *as* $n \to \infty$.

Detailed remarks concerning the LeCam's discretization method can be found in [15]. Here we just mention its basic concept. Suppose the parameter space $\Theta \times \Lambda$ is partitioned into cubes with edges of length $O(n^{-1/2})$. The estimator $\hat{\eta}_* = [\hat{\vartheta}_*^T,\ \hat{\lambda}_*^T]^T$ is defined to be a discretized version of $\hat{\eta}$ as the center of the cube to which $\hat{\eta}$ belongs (see e.g. [2], p. 44). The origination of the vector $\hat{\ell}^*(\hat{\eta}_*)$ in (3.1) and the test statistic (3.5) follow accordingly.

## 4. Simulation study

The test performance in practice was also examined, both for simple and composite hypothesis in [7] and [15], respectively. In the simple case, conditional normality and standard Laplace distribution were considered as null hypotheses, whereas in the composite framework the class $\mathcal{G}$ of standardized Generalized Error Distributions (GED) was taken as a hypothetical family. Recall that the *p.d.f.* of this (standardized, therefore one-parameter) class describing the random behavior of our innovation sequence $\{\varepsilon_t\}$ has the form

$$f(x, \lambda) = \frac{\lambda C_\lambda}{2\Gamma(\lambda^{-1})}\exp(-|C_\lambda x|^\lambda), \tag{4.1}$$

where $x \in R$, $\lambda \in \Lambda = R_+$ is a parameter indexing the family, $C_\lambda = \sqrt{\Gamma(3\lambda^{-1})/\Gamma(\lambda^{-1})}$ is a normalizing constant and $\Gamma$ denotes the Euler gamma function. This symmetric class is flexible enough to encompass the Laplace ($\lambda = 1$) and normal ($\lambda = 2$) distributions as special cases, see e.g. [12]. Validity of the formerly mentioned assumptions from among (A1)–(A13) concerning the GED family has been checked in Section 9 of Stawiarski [15].

The issue of estimating the nuisance parameter $\eta$ is vital. As for its GARCH(1,1) part $\vartheta$, the Quasi Maximum Likelihood Estimator $\hat{\vartheta}$, QMLE, well-described in literature, was computed by iterative methods. Under mild conditions (cf. [11]) the QMLE $\hat{\vartheta}$ is square-root consistent and asymptotically normal with the asymptotic covariance matrix $\Sigma = Var(\varepsilon^2)U_\infty^{-1}$, where $U_\infty$ appears in (6.11) below. Hence, for fixed $h > \omega\ (1-\beta)^{-1}$ we get $\hat{Q}_t$ and the estimated innovations $\hat{\varepsilon}_t = X_t \Big/ \sqrt{\hat{Q}_t}$, $t = 1, .., n$ appearing in (3.1). This estimated sequence, $\{\hat{\varepsilon}_t\}_{t=1,...,n}$, serves in turn to obtain the $\sqrt{n}$-consistent estimator $\hat{\lambda}$ of $\lambda$ stemming from the hypothetical family $\mathcal{G}$, given by (4.1), subject to our testing procedure. Specifically, $\hat{\lambda}$ is calculated numerically by common method of moments. Its $\sqrt{n}$-consistency is not trivial and was proved in Section 9 of [15].

Accordingly, all other matrices and quantities appearing in the data-driven score statistic $W_S$had to be estimated, yielding finally the data-driven test statistic $\hat{W}_{\hat{S}}(\hat{\eta})$, see (3.4). Details can be found in Section 5 of [15], together with versatile discussion concerning numerical issues emerging during estimating critical values and performing the test power study against vast scope of alternatives. Especially, using the asymptotic, nominal $\chi_1$ quantiles is largely disputable due to present time series framework implying slower asymptotic convergence for moderate path lengths.

We set the quadratic scale coefficient $\omega = 0.001$ and the starting value $h = 0.1$ throughout our simulations. The maximal embedding dimension $K$ was fixed at 10, and the cosine orthonormal system $\Phi_j(x) = \sqrt{2}\cos(j\pi x)$, $j = 1, 2, ...$, on [0,1] was chosen. The significance level is fixed at 0.05. To show the influence of nuisance parameters on the empirical critical values, we considered various combinations of $\vartheta$ and $\lambda$. Theoretical results imply the stability of c.v.'s with respect to changing $\eta$, hence the critical value was proposed as a common average.

Focusing upon $1 \leqslant \lambda \leqslant 2$ is motivated by empirical research in modelling real stock and commodities returns, see e.g. [4], [16]. Changing $\lambda$ from 2 down to 1 results in heavier tails of the distribution. Various values of $\lambda$ were paired with several combinations of the GARCH(1,1) parameters $(\alpha, \beta)$. Again, in accordance with results found in papers concerning econometric time series modelling, we deliberately focused on $\vartheta$'s such that $\beta > \alpha$ and $\alpha + \beta \geqslant 0.8$. In such cases, the influence of past conditional variances $h_s$ on the present $h_t$, is stronger, providing more pronounced autocorrelation structure ("memory") within the squares of the time series. The results are reported in Table 1 as empirical .95-quantiles from 5000 Monte Carlo runs for sample size $n = 1000$.

The observed stability of estimated critical values justifies using the global average 4.501 as the critical value of our test for $n = 1000$. Under the null hypothesis, the Schwarz selection rule picks out $\hat{S} = 1$ with frequency about 95-97%. Slow convergence rate of the estimators, especially $\hat{\lambda}$, suggests caution with time series of length e.g. $n = 500$ in the composite hypothesis case, while for simple hypotheses considered in

Tab. 1: Simulated critical values for $\hat{W}_{\hat{S}}$ in case of GED family of null distributions with $\lambda \in (0.99; 2.01)$. Significance level 0.05, $\omega = 0.001$, $h = 0.1$, $K = 10$, $n = 1000$, $MC = 5000$ Monte Carlo loops

| $(\alpha\ ,\ \beta)$ | $\lambda$ | | | | |
|---|---|---|---|---|---|
| | 1 | 1.25 | 1.5 | 1.75 | 2 |
| (0.3; 0.5) | 4.543 | 4.334 | 4.525 | 4.464 | 4.306 |
| (0.2; 0.7) | 5.025 | 4.770 | 4.871 | 4.280 | 4.195 |
| (0.4; 0.5) | 4.378 | 4.492 | 4.770 | 4.324 | 4.311 |
| (0.25; 0.65) | 4.860 | 4.474 | 4.769 | 4.255 | 4.069 |
| Column-wise averages | 4.701 | 4.517 | 4.734 | 4.331 | 4.220 |
| Average critical value : 4.501 | | | | | |

[7] such length was large enough.

Proceeding now to checking the test performance, we consider $n = 1000$ and take the critical value $\hat{CV} = 4.501$ at 0.05 significance level (the asymptotic CV being 3.84). Following alternative distributions (centered and scaled when necessary) have been considered:

-$t$-Student with $a$ degrees of freedom, referred to as $t(a)$;

-chi-square with 5 degrees of freedom, $\chi^2_5$;

-normal mixture (bimodal) with $\mu > 0$ and $\sigma = 1$, $SN(\mu)$, with the density $f_1(x, \mu) = \frac{1}{2\sqrt{2\pi}} \left(\exp\{-(x-\mu)^2/2\} + \exp\{-(x+\mu)^2/2\}\right), x \in R$;

-first-type beta with $a, b > 0$ including uniform $(a, b = 1)$, $B(a, b)$;

-symmetric Pareto-type with shape $b > 2$, $PR(b)$, given by the density

$$f_1(x, b) = \frac{b}{\sqrt{2(b-1)(b-2)}} \left(1 + \frac{|x|\sqrt{2}}{\sqrt{(b-1)(b-2)}}\right)^{-b-1}, x \in R\ ;$$

-GED($\lambda$) with $\lambda > 2$.

To check the test performance and sensitivity when necessary, the "contaminated" alternatives $f_\rho$ were considered, namely

$f_\rho(x, \lambda) = (1-\rho) f(x, \lambda) + \rho\ f_1(x)$,where $0 \leqslant \rho \leqslant 1$, $f(x, \lambda)$ is the null GED density (4.1) and $f_1(x)$ is taken from the above-listed alternatives. The simulations were run for the specific GARCH(1,1) model, namely with $\vartheta = (0.001;\ 0.3;\ 0.5)$, and the power was estimated as percentage of $H_0^*$ rejections out of 2000 Monte Carlo loops. The results are collected in Tables 2 and 4.

Naturally, the power is generally weaker than that reported in [7] for the simple hypothesis. However, such deviations from GED as bimodality, skewness, excess kurtosis are detected satisfactorily well. Some light-tailed distributions are hardly distinguished from the GED null family, while ultra-thin tails, as those of GED($\lambda$) with $\lambda > 6$ are detected quite well. Heavy Pareto-type tails are easily distinguished,

Tab. 2: Simulated powers of $\hat{W}_{\hat{S}}(\hat{\eta})$ under symmetric, unimodal alternatives. Significance level 0.05, $h = 0.1$, $K = 10$, $n = 1000$, $MC = 2000$ Monte Carlo loops; contaminations of conditionally GED GARCH(1,1) model with $\vartheta = (0.001;\ 0.3;\ 0.5)$ and three values of $\lambda$. $H_0$: $f$ - GED($\lambda$); $\lambda \in [1, 2]$

| $f_1$ | $\rho$ | Simulated power (% of rejections) | | |
|---|---|---|---|---|
| | | $\lambda = 1$ | $\lambda = 1.5$ | $\lambda = 2$ |
| $PR(2.5)$ | 1 | 41 | | |
| | 0.8 | 19 | 19 | 20 |
| $PR(5.5)$ | 1 | 19 | | |
| $PR(5)$ | 1 | 11 | | |
| $t(3)$ | 1 | 59 | | |
| | 0.8 | 39 | 43 | 44 |
| | 0.6 | 30 | 33 | 34 |
| $t(5)$ | 1 | 22 | | |
| GED(4) | 1 | 5 | | |
| GED(6) | 1 | 7 | | |
| GED(8) | 1 | 7 | | |
| $B$(1,1) | 0.8 | 97 | 81 | 62 |
| | 0.6 | 90 | 72 | 43 |
| | 0.4 | 34 | 30 | 28 |
| $B$(1.2; 1.2) | 1 | 25 | | |
| $B$(2,2) | 1 | 18 | | |

Tab. 3: Simulated powers of $\hat{W}_{\hat{S}}(\hat{\eta})$ under skew and bimodal alternatives. Significance level 0.05, $h = 0.1$ , $K = 10$ , $n = 1000$ , $MC = 2000$ Monte Carlo loops; contaminations of conditionally GED GARCH(1,1) model with $\vartheta = (0.001;\ 0.3;\ 0.5)$ and three values of $\lambda$. $H_0 : f-$ GED($\lambda$); $\lambda \in [1, 2]$

| $f_1$ | $\rho$ | Simulated power (% of rejections) | | |
|---|---|---|---|---|
| | | $\lambda = 1$ | $\lambda = 1.5$ | $\lambda = 2$ |
| $B(1;\ 1.5)$ | 1 | 100 | | |
| | 0.8 | 98 | 93 | 87 |
| | 0.6 | 72 | 57 | 46 |
| $\chi^2_5$ | 0.6 | 99 | 100 | 100 |
| | 0.4 | 68 | 74 | 80 |
| | 0.2 | 16 | 21 | 23 |
| $SN(1.6)$ | 0.8 | 99 | 98 | 98 |
| | 0.6 | 86 | 87 | 82 |
| | 0.4 | 50 | 43 | 41 |
| $SN(1.2)$ | 1 | 19 | | |

provided that $2 < b < 3$. Overall, the sensitivity of the test against wide range of alternatives is satisfactory.

The proposed data-driven methodology works well and can be used as an omnibus testing tool against various type alternatives. Other null hypotheses also can be considered upon checking the distributional assumptions imposed in our paper. More accurate and reliable conditional distribution testing procedure can substantially improve the quality of empirical time series modeling and resulting inference, providing fundamentals to better deal with financial engineering, including risk management, hedging, option pricing issues, especially in the face of lingering economic instability, market anomalies like asset bubbles, crashes translating into non-gaussian, heavy-tail and often skew, asymmetric indices, stocks or commodities returns.

## 5. Score test of fit in general GARCH($p$,$q$) case

Now, we aim at extension of the previously obtained data-driven score test of fit, valid for GARCH(1,1) case to the general, finite dimensional GARCH class. This will provide us with useful, more flexible testing tool within the whole GARCH family, but at the price of some more complex derivations. Specifically, the technical lemmas stated in [7] and [15] carry over to the present situation under appropriate reformulations handling the higher dimensionality of the problem. Here we provide main outline of the extension construction, while some minor work might be an object of future research.

According to the pioneer definition in Bollerslev (1986), the class of symmetric GARCH($p$,$q$) models allows $q$-step backward dependency upon squared series values, as well as $p$-step backward dependency upon its past conditional variances, namely:

$$\begin{cases} X_t = \sqrt{h_t}\varepsilon_t \\ h_t = \omega + \sum\limits_{i=1}^{q} \alpha_i X_{t-i}^2 + \sum\limits_{j=1}^{p} \beta_j h_{t-j} \end{cases} \tag{5.1}$$

with zero-mean, unit variance white noise $\{\varepsilon_t\}$, and $\omega > 0$, $\alpha_i$, $\beta_j \geqslant 0$ but $\alpha_q$, $\beta_p > 0$.

As we employ the data-driven score test methodology to strictly stationary series with finite second unconditional moment, let us quote some already known theorems establishing necessary and sufficient conditions for stationarity. To this end, it is convenient to express the GARCH($p$,$q$) model as a vector Markov process $Z_t = B_t + A_t Z_{t-1}$, where

$$B_t = \left(\omega\varepsilon_t^2, 0, ..., \omega, 0, ..., 0\right)^T \in R^{p+q},\ Z_t = \left(X_t^2, ..., X_{t-q+1}^2, h_t, ..., h_{t-p+1}\right)^T \in R^{p+q},$$

$$A_t = \begin{bmatrix} \alpha_1\varepsilon_t^2 & & ... & & \alpha_q\varepsilon_t^2 & \beta_1\varepsilon_t^2 & & ... & & \beta_p\varepsilon_t^2 \\ 1 & 0 & ... & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 & 0 & 0 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & & ... & 1 & 0 & 0 & 0 & ... & 0 & 0 \\ \alpha_1 & & ... & & \alpha_q & \beta_1 & & ... & & \beta_p \\ 0 & 0 & ... & 0 & 0 & 1 & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & 0 & 0 & 1 & ... & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & ... & 0 & 0 & 0 & 0 & ... & 1 & 0 \end{bmatrix} \tag{5.2}$$

Thus $A_t$ is a square, sparse $(p+q) \times (p+q)$ matrix with stochastic first row. Now, successively iterating the formula for $Z_t$ we get the strictly stationary solution

$$Z_t = B_t + \sum_{i=1}^{\infty} \left( \prod_{j=0}^{i-1} A_{t-j} \right) B_{t-i} \tag{5.3}$$

provided the almost sure existence of the series. The examination of strict stationarity can be carried out by means of top Lyapunov exponent $\gamma$ associated with the sequence of (strictly stationary and ergodic) matrices $A_t, t \in Z$. Assuming that $E \log^+ || A_t|| < \infty$ holds, (here $x^+ = \max\{x; 0\}$) the top Lyapunov exponent can be derived as [cf. [5]]

$$\gamma = \lim_{t\to\infty} t^{-1} E(\log ||A_t A_{t-1} ... A_1||) = a.s. \lim_{t\to\infty} t^{-1} \log ||A_t A_{t-1} ... A_1||.$$

Now we cite two theorems concerning strict and weak (second-order) stationarity of the GARCH process, which are the objects of our interest.

**Theorem 5.1.** *[Theorem 2.4 in [5]]*
*Let $\gamma$ be the top Lyapunov exponent of the sequence $\{A_t\}$ given by (5.2). The process (5.1) admits strictly stationary solution if and only if $\gamma < 0$. Such a solution is also nonanticipative (with respect to process filtration) and ergodic.*

**Theorem 5.2.** *[Theorem 2.5 in [5]]*
*If the process obeying (5.1) is weakly stationary and nonanticipative, then*

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1. \tag{5.4}$$

*If, conversely, (5.4) holds, the unique strictly stationary solution of (5.1) is weakly stationary.*

*Remark.* Under the conditions of Theorem 5.1 the unconditional variance of GARCH($p$,$q$) model equals

$$EX_t^2 = v^2 = \frac{\omega}{1 - \sum\limits_{i=1}^{q} \alpha_i - \sum\limits_{j=1}^{p} \beta_j}. \tag{5.5}$$

Moreover, (5.4) implies that $\gamma < 0$ as the weakly stationary solution stated in Theorem 5.1 is also strictly stationary. In the case of GARCH(1,1) checking the Lyapunov exponent in Theorem 5.1 simplifies greatly: the stationarity condition reads as $\gamma = E \log(\alpha\varepsilon_t^2 + \beta) < 0$.

Further vast and versatile theoretical results concerning not only the GARCH models but their various modifications (xARCH), together with broad empirical applications, simulations etc. can be found in [5] and in numerous preceding papers. Now, we proceed to extending our former results derived in [7] and [15] to finite dimensional ARCH family, and then finally for the whole GARCH($p$,$q$) class.

### 5.1. ARCH($\boldsymbol{q}$) case

For ARCH($q$) submodel with no dependence on past conditional variances,

$$\begin{cases} X_t = \sqrt{h_t}\varepsilon_t \\ h_t = \omega + \sum\limits_{j=1}^{q} \alpha_j X_{t-j}^2 \end{cases} \tag{5.6}$$

with $\omega > 0$ and non-negative $\alpha_j$'s but $\alpha_q > 0$, we shall test the conditional distribution of $X_t$ by addressing the distribution of innovations $\varepsilon_t$. With the observed strictly (and weakly) stationary series $(X_1,\ldots, X_n)$ obeying (5.6), and its induced joint distribution taken, as before, conditionally on $h_1 = h > \omega(1-\alpha_1-\ldots-\alpha_q)^{-1}$, we aim at deriving our data-driven test statistic. This can be done by repeating vast part of calculus done for GARCH(1,1) model along the lines from Section 2 and 3, but there are indeed some differences in the score vector and the efficient score vector as now the dynamics of $h_t$ is somewhat different than that of GARCH(1,1). The nuisance parameter in our testing problem becomes now $(q+1+m)$-dimensional, namely $\eta = (\omega, \alpha_1, \ldots, \alpha_q, \lambda_1, \ldots, \lambda_m)^T$ $\in \Theta \times \Lambda$, where the set $\Theta$ of ARCH parameter values ensures finite variance of $X_t$.

In this case, too, we have to derive successively: score vector, efficient score vector and score-statistic with its data-driven version, culminating in the test statistic with nuisance parameters, estimated regularly enough. We put into closer inspection the stages in which derivation differs significantly from the previously considered GARCH(1,1) model. Such a first notable difference is obviously the conditional variance $h_t$ in (5.6), compare (2.1). Successive iteration in a manner leading to formula (2.4) does not carry over which is not cumbersome until some details in proofs of the limit theorem paralleling Theorems 2.2 and 3.1.

The second part of the score vector $\ell(\eta)$ given in (2.6), namely $\ell_\vartheta$, changes accordingly to difference between dynamics of conditional variance in GARCH(1,1) and ARCH($q$) models, compare (2.1) and (5.6). Now, instead of (2.7) we have, conditionally on $h_1$ for $t \geqslant 2$

$$Q_t = Q_t(X_1, ..., X_{t-1}; h, \vartheta) = \omega + \sum_{j=1}^{\min\{t-1;q\}} \alpha_j X_{t-j}^2 \tag{5.7}$$

with $Q_1 = h$, while

$$\frac{\partial Q_t}{\partial \vartheta} = \left[\frac{\partial Q_t}{\partial \omega},\ \frac{\partial Q_t}{\partial \alpha_1},\ ...,\ \frac{\partial Q_t}{\partial \alpha_q}\right]^T = \begin{bmatrix} 1 \\ X_{t-1}^2\,(\mathrm{I}\{t > q\} + \mathrm{I}\{1 < t \leqslant q\}) \\ \vdots \\ X_{t-j}^2\,(\mathrm{I}\{t > q\} + \mathrm{I}\{j < t \leqslant q\}) \\ \vdots \\ X_{t-q}^2\,(\mathrm{I}\{t > q\} + \mathrm{I}\{q < t \leqslant q\}) \end{bmatrix}^T \tag{5.8}$$

except for the first row always being 1, contains $X_{t-j}^2$'s or just zeroes depending on whether $t > q$ or whether $t$ falls between $j+1$ and $q$; $j = 1, \ldots, q$. The formula (2.6) retains its shape but it must be underlined that $\ell_\vartheta$ is now $(q+1)$ dimensional in accordance with (5.8), therefore the whole $\ell(\eta)$ is now $(k+q+1+m)$-dimensional.

The present ARCH($q$) setup implies generally larger (for $q > 2$) dimensions of the block matrices constituting $\tilde{B}^{(n)}(\eta)$ in (2.8). The whole score statistic construction goes along the same lines as in Section 2. The main proofs, stating the asymptotic result concerning $W_k$ and $W_S$ apply, too, but with one important remark. Namely, some stages of the proofs derivation call for a bit modified re-writing whenever we deal with blocks containing the "subvector" $\ell_\vartheta$. Main obstacle to overcome is the fact that it is by a long chalk harder to descend down to $h_1$ in (5.7), than it was possible to do outright in (2.7) by successive iterations. Obviously, it is attainable via vector representation like (5.2), but in that case the crucial Lemmas 5.2 and 5.3 in [7] providing the limit behaviors of $\bar{V}_n$, $E_h\bar{V}_n$, $\bar{U}_n$ and $E_h\bar{U}_n$ (cf. (6.1) and (6.2)) under $P_h$ as $n \to \infty$ deal with some other expressions. Specifically, in the present context, it remains to show that

$$E_h\left\{\frac{1}{n}\sum_{t=2}^{n}\frac{1}{Q_t}\right\} \stackrel{n\to\infty}{\longrightarrow} \Psi_0 \quad , \quad E_h\left\{\frac{1}{n}\sum_{t=2}^{n}\frac{X_{t-j}^2}{Q_t}\right\} \stackrel{n\to\infty}{\longrightarrow} \Psi_j \tag{5.9}$$

under $P_h$ for $j = 1, \ldots, q$. Obviously, the same argument as in Appendix of [7] can be used here, namely the Birkhoff's ergodic theorem applied for appropriately defined $L_1$-integrable functional operating on the sequence of squared innovations $\{\varepsilon_t^2\}$ driving our model (5.6). In the GARCH(1,1) context, conditionally on $h_1 = h$, $Q_t$ could be expressed outright as

$$Q_t = \omega\left(1 + \sum_{i=1}^{t-2}\prod_{j=0}^{i-1}(\beta + \alpha\tilde{\varepsilon}_t^2)\right) + h\prod_{j=0}^{t-1}(\beta + \alpha\tilde{\varepsilon}_t^2) \tag{5.10}$$

which in Section 6 of [7] served to prove the crucial Lemmas 5.2 and 5.3, with the aid of the fact that $E\tilde{e}_t^2 = E(\alpha\tilde{\varepsilon}_t^2 + \beta) < 1$. Now, to prove (5.9) for the ARCH($q$) case, one has to express $Q_t$ tracking back down to $h$ similarly as in (5.10) but taking the model difference into account. The analogue of (5.10) in this context will contain cumulative products of variables being now $\tilde{e}_{t,j}^2 = \alpha_j\tilde{\varepsilon}_{t-j}^2$, with $E\tilde{e}_{t,j}^2 < 1$ for $j = 1, \ldots, q$ and any $t$. Formal proof of (5.9) calls for readjustment in calculus and is the object of current research. Thank to the general properties of ARCH model resulting from our assumptions, the validity of (5.9) can be ascertained with next to zero risk of failure. This will imply the validity of the whole Section 4, as the martingale-difference-array structure of the efficient score vector $\ell^*(\eta)$ in (2.9) remains intact and the Kundu et al. [9] limit theorem applies accordingly.

Repeating the proof of Theorem 3.1 for the data-driven score statistic with QMLE $\hat{\vartheta}$ of the ARCH nuisance parameter, and the estimator $\hat{\lambda}$ as before follows directly. Obviously, detailed estimations and derivations in the auxiliary lemmas in Section 7 of [15] will change accordingly, but with no detriment to the validity of the main limit theorem.

For the whole GARCH($p$,$q$) model (5.1) with $r = \max\{p,\, q\} \geqslant 2$ we obviously have

$$h_t = \omega + \sum_{j=1}^{r} (\alpha_j \varepsilon_{t-j}^2 + \beta_j) h_{t-j}, \tag{5.11}$$

but successive iteration in (5.11) for each of $h_{t-j}$ down to $h_1$ and conditionally on $h_1 = h$ exceeding the right-hand side of (5.5) will not yield any constructive form of $Q_t$ from our test derivation standpoint. Indeed, the products of expressions $(\alpha_j \varepsilon_{t-j}^2 + \beta_j)$ will proliferate so amply that the derivation of the score vector $\ell(\eta)$ defined in (2.5), (2.6) in the closed form is unattainable due to the cumbersome derivative $\frac{\partial Q_t}{\partial \vartheta}$ in (2.7). The same remark applies to the block covariance matrix $\tilde{B}^{(n)}(\eta)$ in (2.8) and the efficient score vector $\ell^*(\eta)$ in (2.9), due to the indisposable component $\ell_\vartheta$. The technically intricate and computationally hardly tractable matrix representation (5.2), using higher-order Kronecker products, does not facilitate the task, either.

However, we can propose a circumvention of these major technical obstacles, transforming the GARCH$(p, q)$ model into the reparametrized ARCH$(\infty)$ model, then consequently truncating it at some fixed lag, say $Q$. This comes at a minor expense of disposing the exponentially vanishing process history tracking back beyond $(t - Q)$ time scale, but allows us to employ the results from the above subsection, dealing again with ARCH($Q$) model. Thus our data-driven score test of fit (asymptotically) encompasses the whole GARCH($p$,$q$) family. Such a shortcut comes with no harm to practical applications, since it is common to consider lower or moderate model sizes, so that the model can successfully serve for applicational purposes imposed by financial time modeling objectives.

Formalizing our concept, let us quote a theorem providing the needed transformation. Such an "inversion" of GARCH model to ARCH$(\infty)$ goes in a similar spirit to the classical case of causal and invertible (linear) ARMA models. The result has already been stated in the pioneer paper of Bollerslev [3], pp. 309-310, refined later on by other authors.

**Theorem 5.3.** *For a strictly stationary GARCH(p,q) model $\{X_t\}$ following (5.1), with $E\varepsilon_t^2 = \sigma^2 < \infty$ and $\sigma^2 \sum\limits_{i=1}^{q} \alpha_i + \sum\limits_{j=1}^{p} \beta_j < 1$, there exists a non-negative, summable sequence $\{\psi_j\}_{j \geqslant 0}$ such that*

$$h_t = \psi_0 + \sum_{j=1}^{\infty} \psi_j X_{t-j}^2 \tag{5.12}$$

*with expansion coefficients given as*

$$\begin{aligned} &\psi_0 = \frac{\omega}{1-\sum\limits_{r=1}^{p}\beta_r}, \\ &\sum_{j=1}^{\infty}\psi_j z^j = \frac{\sum\limits_{i=1}^{q}\alpha_i z^i}{1-\sum\limits_{r=1}^{p}\beta_r z^r},\ j \geqslant 1,\ z \in C; \mid z \mid\ < 1 \end{aligned} \tag{5.13}$$

With this result, having the GARCH($p$,$q$) model with $\sigma^2 = 1$ and original parameters $\omega$, $\alpha_i$, $\beta_j$ we can reparametrize it into ARCH($\infty$), solving the system (5.13) to find the new coefficients $\psi_0$, $\psi_1$,.... Next, the resulting conditional variance $h_t$ in (5.12) can be truncated to ARCH($Q$) representation with $Q$ picked so that an arbitrarily chosen accuracy measured e.g. by $L_1$ norm can be obtained. Define

$$h_t^{(Q)} = \psi_0 + \sum_{j=1}^{Q}\psi_j X_{t-j}^2 \tag{5.14}$$

and for any fixed accuracy $\delta > 0$ choose $Q$ so that (cf. (5.5))

$$E\left\{h_t - h_t^{(Q)}\right\} = \sum_{r=Q+1}^{\infty}\psi_r E X_{t-r}^2 = \nu^2 \sum_{r=Q+1}^{\infty}\psi_r < \delta. \tag{5.15}$$

Finally, we can employ the former derivation of our data-driven score test of fit for the ARCH($Q$) model parallel to (5.1):

$$\begin{cases} X_t = \sqrt{h_t^{(Q)}}\varepsilon_t \\ h_t^{(Q)} = \psi_0 + \sum\limits_{i=1}^{Q}\psi_i X_{t-i}^2 \end{cases} \tag{5.16}$$

where after reparametrization now $\vartheta = [\psi_0, \psi_1, ..., \psi_Q]^T$. Thus, our data-driven score test of fit, designed primarily to verify the hypothetical conditional distribution in the GARCH(1,1) model has been "asymptotically" extended over the whole GARCH($p$,$q$) class of stationary time series, and its performance also can be checked by means of computer simulations.

## 6. Appendix

We provide a brief proof of Theorem 2.2, in which Theorem 1.3 from Kundu et al. [9] for martingale-difference arrays is very helpful. The present arguments mimic the ones used in the proof of Theorem 3.6 in [7]. Reasoning goes along the lines of Subsection 6.2 in [15].

Let us first examine limiting behaviour of the matrices $\tilde{B}_{12}^{(n)}(\eta)$ and $\tilde{B}_{22}^{(n)}(\eta)$ appearing in (2.8), (2.9), (2.10) – originally in (6.7) and (6.12) in [15]. For $t \geqslant 2$ introduce random matrices

$$V_t = \frac{1}{Q_t}\frac{\partial Q_t}{\partial \vartheta}, \quad U_t = V_t V_t^T, \tag{6.1}$$

for notational convenience, set $V_1 = 0$, $U_1 = 0$ and define two corresponding random averages

$$\bar{V}_n = \frac{1}{n}\sum_{t=1}^{n} V_t, \quad \bar{U}_n = \frac{1}{n}\sum_{t=1}^{n} U_t. \tag{6.2}$$

By mutual independence of $\tilde{\varepsilon}_t$'s under $P_h$, straightforward calculation gives

$$n^{-1} E_h \ell_\tau \ell_\vartheta^T = -\Delta_1 (E_h \bar{V}_n^T)/2, \tag{6.3}$$

where

$$\Delta_1 = \Delta_1(\lambda) = E\Phi(F(\varepsilon,\lambda))\varsigma(\varepsilon,\lambda) \tag{6.4}$$

with $\varsigma$ defined in (A2). Similarly we get

$$n^{-1} E_h \ell_\tau \ell_\lambda^T = E\Phi(F(\varepsilon,\lambda))\frac{\partial \log f(\varepsilon,\lambda)}{\partial \lambda^T} = \Delta_2(\lambda) = \Delta_2. \tag{6.5}$$

Now, (6.3) together with (6.5) yield the following block representation of $\tilde{B}_{12}^{(n)}(\eta)$:

$$\tilde{B}_{12}^{(n)}(\eta) = [-\Delta_1 (E_h \bar{V}_n^T)/2 \ \Delta_2]. \tag{6.6}$$

By similar derivation we obtain

$$n^{-1} E_h \ell_\vartheta \ell_\vartheta^T = J_\varsigma (E_h \bar{U}_n)/4 \tag{6.7}$$

with $J_\varsigma = J_\varsigma(\lambda) = E\varsigma^2(\varepsilon,\lambda)$. We further have

$$n^{-1} E_h \ell_\vartheta \ell_\lambda^T = -(E_h \bar{V}_n)\Delta_3^T/2, \tag{6.8}$$

$$\Delta_3 = \Delta_3(\lambda) = E\left\{\varsigma(\varepsilon,\lambda)\frac{\partial \log f(\varepsilon,\lambda)}{\partial \lambda}\right\}. \tag{6.9}$$

Finally, $n^{-1} E_h \ell_\lambda \ell_\lambda^T = I(\lambda)$ - the Fisher information matrix defined in (A1). This and (6.7)-(6.9) lead to the following block form of $\tilde{B}_{22}^{(n)}(\eta)$

$$\tilde{B}_{22}^{(n)}(\eta) = \begin{bmatrix} J_\varsigma (E_h \bar{U}_n)/4 & -(E_h \bar{V}_n)\Delta_3^T/2 \\ -\Delta_3 (E_h \bar{V}_n^T)/2 & I(\lambda) \end{bmatrix}. \tag{6.10}$$

By Lemma 5.2 and 5.3 from [7] it follows that

$$\bar{V}_n \xrightarrow{P_h} V_\infty, \ E_h \bar{V}_n \to V_\infty,$$

$$\bar{U}_n \xrightarrow{P_h} U_\infty, E_h\bar{U}_n \to U_\infty \tag{6.11}$$

as $n \to \infty$, where $V_\infty$ and $U_\infty$ are deterministic matrices given explicitly by (A.12) and (A.13) in Appendix of [7]. Hence

$$\tilde{B}_{12}^{(n)}(\eta) \underset{n\to\infty}{\longrightarrow} \tilde{B}_{12}^{\infty}(\eta) = [\ -\Delta_1 V_\infty^T/2 \quad \Delta_2\ ] \tag{6.12}$$

and

$$\tilde{B}_{22}^{(n)}(\eta) \underset{n\to\infty}{\longrightarrow} \tilde{B}_{22}^{\infty}(\eta) = \begin{bmatrix} J_\varsigma U_\infty/4 & -V_\infty\Delta_3^T/2 \\ -\Delta_3 V_\infty^T/2 & I(\lambda) \end{bmatrix}. \tag{6.13}$$

For further convenience, let us introduce the abbreviated notation for the projection matrix in (2.9),

$$\tilde{A} = \tilde{A}^{(n)}(\eta) = \tilde{B}_{12}^{(n)}(\eta)[\tilde{B}_{22}^{(n)}(\eta)]^{-1}. \tag{6.14}$$

Thus, (6.12) and (6.13) directly imply that for $n \to \infty$
$\tilde{A}^{(n)}(\eta) \to \tilde{A}^{\infty}(\eta) = \tilde{B}_{12}^{\infty}(\eta)[\tilde{B}_{22}^{\infty}(\eta)]^{-1}$and $\tilde{M}^{(n)}(\eta) \to \tilde{M}^{\infty}(\eta) =$
$I_k - \tilde{B}_{12}^{\infty}(\eta)[\tilde{B}_{22}^{\infty}(\eta)]^{-1}\tilde{B}_{21}^{\infty}(\eta)$.

Now, we can view $\ell^*(\eta)$ as a sum composed of martingale difference array summands. To this end, define for $t \geqslant 1$ $\sigma$-fields $\sigma_t = \sigma(X_1, ..., X_t)$ and

$$X_{tn} = \tilde{M}^{-1/2}\frac{1}{\sqrt{n}}\left(\Phi(F(\tilde{\varepsilon}_t, \lambda)) - \tilde{A}\begin{bmatrix} -\varsigma(\tilde{\varepsilon}_t, \lambda)V_t/2 \\ \frac{\partial \log f(\tilde{\varepsilon}_t, \lambda)}{\partial \lambda} \end{bmatrix}\right), \tag{6.15}$$

with $\tilde{M}$ given by (2.10).

It is straightforward to show that the sequence $\{X_{1n}, ..., X_{nn}\}$ is, under $P_h$, a martingale difference array adapted to $\{\sigma_1, ..., \sigma_n\}$ (cf. Proposition 3.5 in [7]) and $\sum_{t=1}^{n} X_{tn} = \tilde{M}^{-1/2}\ell^*(\eta)$. This observation allows us to apply Theorem 1.3 from [9].

Since (A4) is satisfied, checking the Lindeberg-type condition (ii) of that theorem can be replaced with the following stronger, Lyapunov-type one
$\sum_{t=1}^{n} E_h\{||X_{tn}||^3|\sigma_{t-1}\} \xrightarrow{P_h} 0$ as $n \to \infty$.

However, this is carried out along exactly the same lines as in the proof of Theorem 3.6 in [7], therefore we omit the details citing just this reference.

Turning now to condition (i) in the aforementioned theorem of Kundu et al. it suffices to show that

$$\sum_{t=1}^{n} E_h\{X_{tn}X_{tn}^T|\sigma_{t-1}\} \xrightarrow{P_h} I_k \tag{6.16}$$

in $R^k$ as $n \to \infty$, with $X_{tn}$ as in (6.15). According to the form of $X_{tn}$ the left-hand side of (6.16) can be decomposed into $S_{1n} + S_{2n} + S_{3n}$. By Proposition 3.1 in [15] and orthonormality of $\Phi$,

$$S_{1n} = \tilde{M}^{-1/2} E\Phi(F(\varepsilon,\lambda))\Phi^T(F(\varepsilon,\lambda))\tilde{M}^{-1/2} = \tilde{M}^{-1} \to (\tilde{M}^{\infty})^{-1} \text{as } n \to \infty. \quad (6.17)$$

Recalling that $\tilde{\varepsilon}_t$ is independent of $\sigma_{t-1}$ and $V_t$ is $\sigma_{t-1}-$measurable, from (6.2), (6.4) and (6.5) we immediately get

$$S_{2n} = \tilde{M}^{-1/2}\{[\ \Delta_1 \bar{V}_n^T/2 \quad -\Delta_2\ ]\tilde{A}^T + \tilde{A}[\ \bar{V}_n\Delta_1^T/2 \quad -\Delta_2^T\ ]^T\}\tilde{M}^{-1/2} \quad (6.18)$$

and consequently, by (6.11)-(6.14) and the forms of $\tilde{A}^{\infty}(\eta)$ and $\tilde{M}^{\infty}(\eta)$,

$$S_{2n} \xrightarrow{P_h} -2(\tilde{M}^{\infty})^{-1/2}\tilde{B}_{12}^{\infty}(\tilde{B}_{22}^{\infty})^{-1}\tilde{B}_{21}^{\infty}(\tilde{M}^{\infty})^{-1/2} = 2I_k - 2(\tilde{M}^{\infty})^{-1} \text{ as } n \to \infty. \quad (6.19)$$

Exploiting similar arguments and from (6.2), (6.7)-(6.14) we obtain

$$S_{3n} = \tilde{M}^{-1/2}\tilde{A}\begin{bmatrix} J_{\varsigma}\bar{U}_n/4 & -\bar{V}_n\Delta_3^T/2 \\ -\Delta_3\bar{V}_n^T/2 & I(\lambda) \end{bmatrix}\tilde{A}^T\tilde{M}^{-1/2} \xrightarrow{P_h} (\tilde{M}^{\infty})^{-1} - I_k \text{ as } n \to \infty. \quad (6.20)$$

Finally, we conclude (6.16) from (6.17)-(6.20), which completes the proof. *Q.E.D.*

## 7. Final remarks and acknowledgements

The paper essentially solves the problem of testing the conditional distribution in the framework of general (symmetric) GARCH($p$,$q$) models via the omnibus data-driven score test of fit methodology. Composite hypothesis case naturally embraces the simple one by including additional nuisance parameter $\lambda$. Obviously, some minor technical readjustments and restatements of the lemmas and propositions proved in [7] and [15] are welcome to absolutely fully complete the testing problem, whereas the simulation study and practical test performance for empirical data could prove the usefulness of the proposed tool. The possible future computational paper can also contain the primary model diagnostics, while the implementation of the test statistic itself goes exactly along the same lines as described and practiced in the above-mentioned former papers, but with taking into account the presented extension which numerically can be handled outright. However, these issues fall beyond the scope of this paper, due to volume constraints, too.

The author would like to thank two referees for their constructive comments and remarks, improving the clarity and transparency of this paper.

## References


[1] Bai J., *Testing parametric conditional distributions of dynamic models.* Rev. Econ. Stat. 85, 2003, 531—549.

[2] Bickel P.J., Klaassen C.A.J., Ritov Y., Wellner J.A. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Press, London, 1993.

[3] Bollerslev T., *Generalized autoregressive conditional heteroskedasticity.* J. Econometrics 31, 1986, 307—327.

[4] Chen Y.-T., *A new class of characteristic-function-based distribution tests and its application to GARCH Model.* Working paper, Institute for Social Sciences and Philosophy, Academia Sinica, 2002.

[5] Francq C., Zakoïan J.M., *GARCH models. Structure, statistical inference and financial applications.* Wiley, UK. 2010.

[6] Inglot, T., Kallenberg, W.C.M., Ledwina, T. (1997). *Data driven smooth tests for composite hypotheses.* Ann. Stat. 25(3), 1222—1250.

[7] Inglot T., Stawiarski B., *Data-driven score test of fit for conditional distribution in the GARCH(1,1) model.* Prob. Math. Stat. 25, 2005, 331—362.

[8] Kallenberg W.C.M., Ledwina T, *Data driven smooth tests for composite hypotheses: comparison of powers.* J. Stat. Comp. Simul. 59, 1997, 101—121.

[9] Kundu S., Majumdar S., Mukherjee K., *Central Limit Theorems revisited.* Stat. & Prob. Letters 47, 2000, 265—275.

[10] Ledwina, T., *Data driven version of Neyman's smooth test of fit.* J. Amer. Stat. Assoc. 89, 1994, 1000–1005.

[11] Lumsdaine R.L., *Consistency and asymptotic normality of the Quasi Maximum Likelihood Estimator in IGARCH(1,1) and covariance stationary models.* Econometrica 64, 1996, 575—596.

[12] Mittnik S., Paolella M.S., Rachev S., *Unconditional and conditional distributional models for the Nikkei index.* Asia-Pacific Fin. Markets 5, 1998, 99—128.

[13] Neyman J., *Smooth test for goodness of fit.* Skand. Aktuarietidskr. 20, 1937, 149—199.

[14] Rao C. R., *Linear statistical inference and its applications.* Wiley, New York (973.

[15] Stawiarski B., *Score test of fit for composite hypothesis in the GARCH(1,1) model.* JSPI 139, 2009, 593—616.

[16] Varma J.R., *Value at Risk models in the Indian stock market*. Working paper, IIMA, 1999.

## Contents