Implementarea și Validarea unui Model Box-Jenkins pentru Monitorizarea Continuă a Glucozei

Călugăr Mihnea-Gheorghe, subgrupa 2.2

1. Introducere

Tema abordată: Prezenta lucrare se concentrează pe identificarea sistemelor biologice complexe, utilizând ca studiu de caz dinamica glucozei în sânge la pacienții cu diabet zaharat de tip 1. Se utilizează setul de date multivariat HUPA-UCM Diabetes Dataset pentru a construi un model matematic de tip Box-Jenkins (BJ).

Scopul proiectului: Obiectivul principal este dezvoltarea și validarea unui model predictiv capabil să estimeze nivelul glicemiei pe un orizont de timp de 60 de minute. Din punct de vedere tehnic, proiectul urmărește optimizarea structurii polinomiale pentru a elimina decalajul temporal (eroarea de fază) și pentru a asigura independența stochastică a reziduurilor.

Importanța lucrării: Anticiparea variațiilor glicemice este critică pentru prevenirea episoadelor de hipoglicemie și hiperglicemie. Integrarea datelor provenite de la senzori purtabili (ritm cardiac, pași, calorii) în modele matematice permite o gestionare proactivă a afecțiunii, reducând riscurile pe termen lung și oferind un suport decizional digital pentru pacienți.

2. Dezvoltare Teoretică

2.0 Fundament: *„Abordările moderne presupun modelarea perturbațiilor ca procese stochastice și exploatarea proprietăților lor." [1]*

2.1. Modelarea Stochastică: Conform Cap. 14.1 din suportul de laborator, sistemele reale sunt influențate de perturbații modelate ca procese stochastice discrete:

$$z(t) = H(q^{-1})\, e(t)$$

2.2. Modelul Box-Jenkins (BJ):

„Structura de tip Box-Jenkins este considerată cea mai flexibilă formă de modelare parametrică, deoarece permite descrierea independentă a dinamicii procesului și a proprietăților statistice ale perturbatoarelor.” [2]

Aceasta structură separă complet dinamica sistemului de cea a zgomotului, fiind definită de relația:

$$y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t - n_k) + \frac{C(q^{-1})}{D(q^{-1})}e(t)$$

unde polinoamele B,F modelează procesul determinist, iar C,D modelează perturbarea.

2.3. Estimarea și Validarea Identificarea parametrilor $\hat{\theta}$ se realizează prin metoda CMMP, minimizând eroarea de predicție:

$$J(\theta) = \epsilon\epsilon^T$$

Validarea constă în analiza reziduurilor pentru a verifica dacă acestea sunt necorelate (zgomot alb).

3. Dezvoltare practică

3.1. Configurarea experimentului și prezentarea datelor Implementarea modelului predictiv a fost realizată în mediul MATLAB, utilizând *System Identification Toolbox*. Studiul de caz se bazează pe setul de date HUPA-UCM Diabetes Dataset, care cuprinde înregistrări fiziologice multivariate colectate de la pacienți cu diabet zaharat de tip 1 prin intermediul senzorilor de monitorizare continuă a glicemiei (CGM) și a dispozitivelor purtabile.

Obiectivul principal constă în identificarea unei structuri de tip Box-Jenkins care să coreleze următoarele variabile:

- Ieșirea sistemului y(t): Nivelul de glucoză în sânge (mg/dL).

- Intrările sistemului u(t): Ritmul cardiac (Heart Rate), Rata bazală de insulină (Basal Rate), Consumul caloric (Calories) și numărul de pași (Steps).

3.2. Preprocesarea datelor experimentale: Pentru a asigura o identificare corectă și a evita fenomenele de instabilitate numerică sau overfitting, datele au fost supuse următoarelor operațiuni de prelucrare:

1. **Filtrarea semnalului**: S-a utilizat funcția smoothdata (filtru de medie alunecătoare) pentru a elimina zgomotul de înaltă frecvență generat de senzorul de glucoză, prevenind astfel modelarea eronată a artefactelor de măsură.
2. **Centrarea datelor**: Aplicarea funcției detrend pentru a elimina mediile și trendurile liniare, aducând seriile temporale într-o formă staționară, necesară estimării parametrilor polinoamelor de zgomot.
3. **Eșantionarea**: Datele au fost prelucrate la un interval de eșantionare T_s= 300 secunde (5 minute), corespunzător frecvenței de raportare a senzorilor medicali.

3.3. Identificarea modelului Box-Jenkins Procesul de identificare a fost realizat prin parcurgerea următoarelor etape:

1. **Divizarea datelor:** Setul de date a fost segmentat în date de identificare (70%) și date de validare (30%).
2. **Determinarea întârzierilor n_k :** S-a utilizat funcția proprie *CorCalculator* pentru a analiza corelația încrucișată între fiecare intrare și ieșire, stabilind lag-ul optim pentru fiecare canal.
3. **Configurarea ordinelor:** S-au selectat ordinele polinoamelor. Având în vedere complexitatea proceselor biologice am optat pentru un ordin mare al acestora.

4. **Estimarea parametrilor:** Utilizarea funcției *bj* cu opțiunea Focus: *prediction* și aplicarea regularizării (λ) pentru a asigura stabilitatea numerică a vectorului de parametri .

3.4. Analiza rezultatelor și validarea modelului Validarea a fost efectuată pe setul de date neutilizat la antrenare, urmărind două criterii fundamentale:

1. **Indicatorul de Fit (Precizia):** Evaluarea performanței prin funcția compare, vizând un orizont de predicție de 12 pași (60 de minute). S-a urmărit minimizarea decalajului de fază între ieșirea modelului și datele reale.

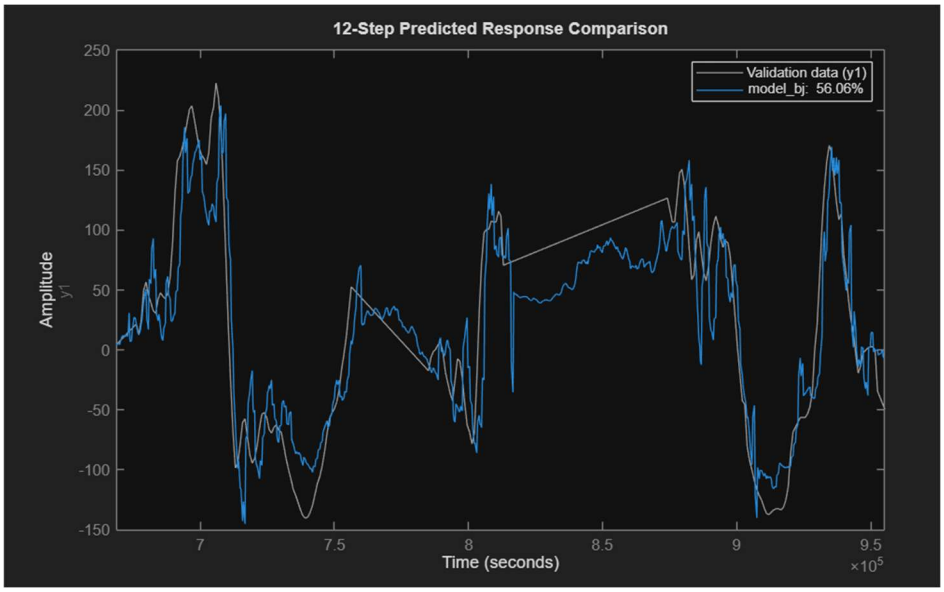

Figura 1 Rezultatul predicției

2. **Analiza reziduurilor (resid):** Verificarea calității modelării prin testarea caracterului de zgomot alb al erorii.
    - **Autocorelația reziduurilor:** S-a urmărit ca valorile să rămână în interiorul intervalului de încredere.
    - **Corelația încrucișată intrare-reziduu:** S-a verificat independența erorii față de variabilele de intrare (Heart Rate, Basal Rate etc.).

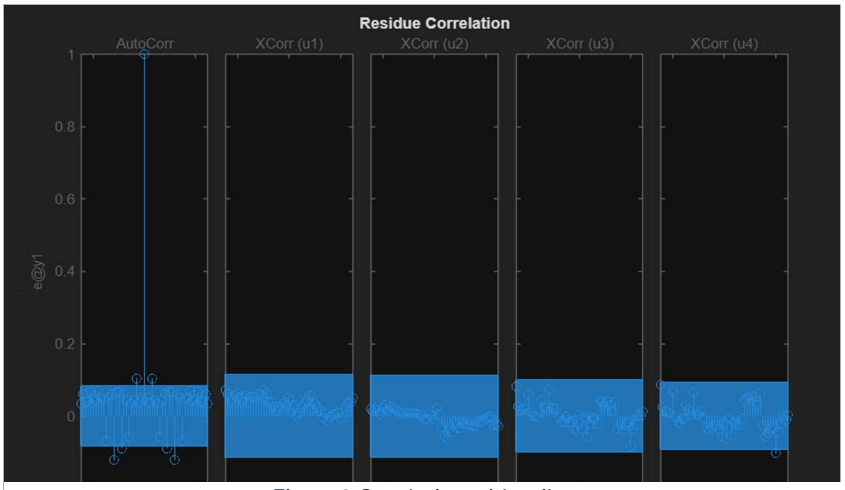

Figura 2 Corelația reziduurilor

4. Concluzie

4.1. Analiza variabilității și limitările modelării liniare Variațiile considerabile ale rezultatelor obținute între diferitele seturi de date testate confirmă faptul că un model universal este dificil de obținut prin metode de identificare liniară clasică. Deși structura Box-Jenkins s-a dovedit utilă pentru filtrarea zgomotului stochastic și captarea trendurilor generale pe termen scurt, performanța fluctuantă demonstrează că modelele de ordin fix întâmpină dificultăți majore în a se adapta la dinamica individuală și non-staționară a fiecărui subiect. Această variabilitate subliniază faptul că parametrii identificați pentru un pacient pot fi complet ineficienți pentru altul, limitând astfel generalizarea modelului.

4.2. Complexitatea fiziologică și perspective viitoare Complexitatea inerentă a sistemelor biologice necesită mai mult decât un model parametric liniar, deoarece mecanismele de reglare ale organismului implică interacțiuni hormonale și metabolice profunde care depășesc capacitatea de reprezentare a polinoamelor de ordin fix. Fundamentarea acestei dificultăți este susținută de literatura de specialitate:

„Biological systems are characterized by their extreme complexity and non-linearity, arising from the intricate interactions between their components at multiple levels of organization, which makes their identification using simple linear approximations a challenging task.” [3]

În concluzie, lucrarea demonstrează utilitatea fundamentului teoretic stochastic în procesarea semnalelor biomedicale, dar punctează în același timp necesitatea evoluției către arhitecturi de modelare neliniară (precum rețelele neuronale) pentru a capta fidel realitatea fiziologică.

Bibliografie:

1. Pagina 188 din suportul de curs (laboratorul 5)

2. Bratu, V., „Identificarea sistemelor dinamice”, Editura Universității Transilvania din Brașov, 2005

3. Kitano H., *Systems Biology: A Brief Overview*, Science, 2002