Uporaba LSE, PCA in PCR metode za identifikacijo modela

Gašper Leskoveca

aFaculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, 1000 Ljubljana, Slovenia.

|  |  |  |
| --- | --- | --- |
| KEYWORDS |  | ABSTRACT |
| LSE, PCA, PCR, standardizacija, kolinearnost |  | Raziskava se osredotoča na identifikacijo modela za proces oksidacije amoniaka v dušikovo kislino s pomočjo PCA, LSE in PCR metod. Posebej analiziramo vpliv standardizacije podatkov na rezultate modeliranja in se ukvarjamo s problematiko kolinearnosti. |

# Uvod

Identifikacija modelov je ključnega pomena za razumevanje in optimizacijo različnih procesov. V tej raziskavi se osredotočamo na proces oksidacije amoniaka v dušikovo kislino in proučujemo vpliv različnih metod modeliranja in standardizacije podatkov.

# Metodologija

## Standardizacija podatkov in vpliv na rezultate

Izvedli smo standardizacijo podatkov na tri različne načine:

a) Na interval [0 1] (range)

b) S standardno deviacijo in odštetim povprečjem (z-score)

c) Uporaba nestandardiziranih podatkov

Nato smo s pomočjo LSE in PCA metod izračunali parametre modela. Za primerjavo modelov smo rezultate standardiziranih spremenljivk transformirali nazaj v dejanski prostor.

### ***LSE opis in enačbe***

V metodi LSE (najmanjših kvadratnih napak) se standardizirani vhodni podatki ​ najprej razširijo z vektorjem enic za vključitev prostega člena, kar dobimo v matriki . Izračun parameterov modela se nato izvede z uporabo matrične inverzne operacije na produktu transponirane  in same , ki je nato pomnožena z odvisno spremenljivko , da pridobimo vektor . Regresijski koeficienti se izluščijo iz vektorja , z izjemo zadnje vrednosti, ki predstavlja prosti člen ​ regresijske enačbe.

### ***PCA opis in enačbe***

V izračunu PCA z uporabo MATLAB, se standardizirani vhodni podatki ​ združijo z odvisno spremenljivko v matriko . Podatki se nato centralizirajo s pomočjo povprečja, ki ga predstavlja , da dobimo . Kovariančna matrika podatkov, oznaka, se nato izračuna iz centraliziranih podatkov. S pomočjo singularne vrednostne dekompozicije (SVD) te matrike pridobimo lastne vektorje . Izberemo prvo glavno komponento, oznaka . Regresijski koeficienti ​ in prosti člen se nato izračunata iz te glavne komponente, ki skupaj predstavljata regresijski model, pridobljen z metodo PCA.

## Problem kolinearnosti

Dodali smo spremenljivko, ki je odvisna od temperature. Parametre modela smo izračunali z LSE in PCR metodami. Proučili smo vpliv kolinearnosti na dobljene parametre.

V vaši kodi se najprej ustvari nova meritev kot funkcija ​, nato se vse spremenljivke standardizirajo z odštevanjem povprečja in deljenjem s standardnim odklonom. Za določitev parametrov se uporabi metoda najmanjših kvadratov (LSE), ki se izračuna kot ​. Glavne komponente so pridobljene s pomočjo analize glavnih komponent (PCA), pri čemer se upoštevajo le tiste, ki pojasnijo vsaj 95% variance. Te glavne komponente se nato uporabijo za transformacijo originalnih podatkov in določitev parametrov z regresijo glavnih komponent (PCR). Končno se oceni kakovost pridobljenih parametrov z izračunom rezidualov in določitvijo njihove variance, ter preveri kolinearnost med spremenljivkami s pomočjo korelacijske matrike.

# Rezultati

## Vpliv standardizacije na rezultate

Primerjava parametrov modelov, pridobljenih s PCA in LSE metodami, kaže na občutljivost PCA metode na način standardizacije. LSE metoda je dala bolj konsistentne rezultate ne glede na standardizacijo. Analiza napake je potrdila te ugotovitve.

***Parametri za PCA in Transformirani PCA***

Pri metodi Range in Z-Score, je absolutna vrednost parametrov pri PCA in Transformiranem PCA metodi manjša kot pri metodi brez standardizacije (None). To kaže na to, da standardizacija zmanjšuje vpliv izrazitih vrednosti in posledično izboljšuje stabilnost in zanesljivost parametrov.

Metoda "None" pri Transformiranem PCA rezultira v enakih parametrih kot pri običajnem PCA, kar kaže na to, da transformacija pri tej metodi nima dodatnega vpliva.

***Parametri za LSE***

LSE metoda prikazuje razlike v parametrih med različnimi metodami standardizacije. Na splošno pa so vrednosti parametrov pri metodah Range in Z-Score višje v primerjavi z metodo None.

|  |
| --- |
|  |
| Slika 1 Tabela vsebuje parametre za LSE metodo. Tudi tukaj so koeficienti 'a1', 'a2', 'a3' in presečišče predstavljeni za metode predobdelave "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje parametre za PCA metodo. Vključuje koeficiente 'a1', 'a2', 'a3' in presečišče oz. prosti člen ('Intercept') za tri različne metode predobdelave podatkov: "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 Tabela vsebuje parametre za transformirano PCA metodo. Ponovno so prikazani koeficienti 'a1', 'a2', 'a3' in presečišče, vendar za transformirane vrednosti, in to za metode "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 prikazuje tri podgrafe, ki primerjajo dejanske podatke Y z različnimi napovedmi pri metodi "none". Prvi podgraf prikaže PCA napoved, drugi transformirano PCA napoved, tretji pa LSE napoved. |

|  |
| --- |
|  |
| Slika 1 prikazuje tri podgrafe, ki primerjajo dejanske podatke Y z različnimi napovedmi pri metodi "range". Prvi podgraf prikaže PCA napoved, drugi transformirano PCA napoved, tretji pa LSE napoved. |

|  |
| --- |
|  |
| Slika 1 prikazuje tri podgrafe, ki primerjajo dejanske podatke Y z različnimi napovedmi pri metodi "z-score". Prvi podgraf prikaže PCA napoved, drugi transformirano PCA napoved, tretji pa LSE napoved. |

***Napake modelov***

Za vse metode je srednja absolutna napaka (MAE) najnižja pri metodi Z-Score, kar pomeni, da je ta metoda najbolj zanesljiva pri napovedovanju dejanskih vrednosti. Prav tako je srednja kvadratna napaka (MSE) najnižja pri Z-Score metodi, kar potrjuje njeno zanesljivost.

Srednja kvadratna napaka pri metodi None je bistveno višja, kar pomeni, da se modeli, pridobljeni brez standardizacije, močno razlikujejo od dejanskih vrednosti.

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje primerjavo povprečnih napak za različne metode. Stolpci tabeli predstavljajo metode "PCA", "PCA\_Transformed" in "LSE", medtem ko vrstice označujejo metode predobdelave podatkov "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje povprečne absolutne napake (MAE) za različne metode. Metode "PCA", "PCA Transformed" in "LSE" so predstavljene v stolpcih, medtem ko vrstice razvrščajo rezultate glede na metode standardizacije: "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje povprečne kvadratne napake (MSE) za različne metode. Metode "PCA", "PCA Transformed" in "LSE" so predstavljene v stolpcih, medtem ko vrstice razvrščajo rezultate glede na metode standardizacije: "Range", "Z-Score" in "None". |

|  |
| --- |
|  |
| Slika 1 Graf v obliki stolpčnice prikazuje primerjavo povprečnih napak za različne metode. Metode "PCA", "PCA Transformed" in "LSE" so predstavljene kot skupine stolpcev za vsako od metod standardizacije, ki so označene kot "Range", "Z-Score" in "None" na x-osi. Y-os prikazuje velikost povprečne napake. |

|  |
| --- |
|  |
| Slika 1 Graf v obliki stolpcev prikazuje primerjavo povprečnih absolutnih napak (MAE) za različne metode. Na osi X so predstavljene metode standardizacije: "Range", "Z-Score" in "None". Na osi Y so prikazane vrednosti MAE. Trije stolpčni sklopi za vsako metodo standardizacije predstavljajo različne metode: "PCA", "PCA Transformed" in "LSE". |

|  |
| --- |
|  |
| Slika 1 Graf v obliki stolpcev prikazuje primerjavo povprečnih kvadratnih napak (MSE) za različne metode. Na osi X so predstavljene metode standardizacije: "Range", "Z-Score" in "None". Na osi Y so prikazane vrednosti MSE. Trije stolpčni sklopi za vsako metodo standardizacije predstavljajo različne metode: "PCA", "PCA Transformed" in "LSE". |

***Varianca parametrov za LSE***

Varianca parametrov nam pove o zanesljivosti ocenjenih parametrov. Nižja varianca pomeni, da so ocene parametrov bolj zanesljive. Pri LSE metodi opazimo naslednje: Varianca parametrov je najnižja pri metodi None. Vendar je treba upoštevati, da čeprav so ocene parametrov zanesljive, lahko model kot celota ni nujno dober, če je srednja napaka visoka, kot smo videli zgoraj.

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje variance za parametre metode LSE. Za vsako metodo standardizacije: "Range", "Z-Score" in "None", so predstavljene variance za parametre 'a1', 'a2', 'a3' in 'Intercept'. |

## Kolinearnost

Rezultati kažejo, da sta LSE in PCR dali različne parametre, še posebej pri spremenljivkah, ki so kolinearne. Ta opažanja kažejo na pomembnost upoštevanja kolinearnosti pri modeliranju.

|  |
| --- |
|  |
| Slika 1 prikazuje izračunane parametre z uporabo metode PCR (Principal Component Regression) in LSE (Least Squares Estimation). Po izračunu parametrov s PCR se ti primerjajo s parametri pridobljenimi z LSE. |

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje prispevke glavnih komponent k posameznim spremenljivkam, izražene v odstotkih. Vsaka vrstica predstavlja spremenljivko, ki je označena z oznako "Xn", pri čemer je "n" zaporedna številka spremenljivke. Vsak stolpec predstavlja eno od glavnih komponent in je označen z oznako "P\_n", kjer je "n" zaporedna številka komponente. Vrednosti v tabeli odražajo delež vsake komponente, ki prispeva k določeni spremenljivki. |

|  |
| --- |
|  |
| Slika 1 Tabela prikazuje variance parametrov, izračunane z metodo najmanjših kvadratov (LSE). Vsaka vrstica predstavlja določeno spremenljivko, ki je označena s specifičnim imenom iz seznama variableNames. Stolpec "LSE" pa prikazuje pripadajoče variance za vsako od teh spremenljivk. |

|  |
| --- |
|  |
| Slika 1 prikazuje korelacijsko matriko R. V tej matriki vsak element predstavlja korelacijski koeficient med dvema spremenljivkama. Diagonalni elementi matrike (od zgornjega levega do spodnjega desnega kota) vedno znašajo 1, saj vsaka spremenljivka popolnoma korelira sama s sabo. |

# Razprava

Standardizacija podatkov ima pomemben vpliv na dobljene modele, zlasti pri uporabi PCA metode. Kolinearnost prav tako vpliva na rezultate modeliranja, kar kaže na potrebo po skrbnem preučevanju in izbiri vhodnih spremenljivk pri modeliranju.

# Zaključek

Identifikacija modela za proces oksidacije amoniaka v dušikovo kislino je ključnega pomena za optimizacijo procesa. Naša raziskava kaže, da je izbira prave metode in pravilna standardizacija podatkov ključna za pridobitev natančnih in uporabnih modelov. Upoštevanje kolinearnosti je prav tako ključnega pomena pri izboru vhodnih spremenljivk za modeliranje.