

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ



Ана Петровић

ТЕСТИРАЊЕ НЕЗАВИСНОСТИ У
ПРИСУСТВУ ГРЕШАКА МЕРЕЊА

мастер рад

Београд, 2026.

Ментор:

проф. др Бојана МИЛОШЕВИЋ, ванредни професор
Универзитет у Београду, Математички факултет

Чланови комисије:

др Марија ЦУПARIЋ, доцент
Универзитет у Београду, Математички факултет

проф. др Марко ОБРАДОВИЋ, ванредни професор
Универзитет у Београду, Математички факултет

Датум одбране: . 2026.

Мојој сесџри

Наслов мастер рада: Тестирање независности у присуству грешака мерења

Резиме: Тема овог рада је тестирање независности случајних променљивих у присуству грешака мерења. Анализира се утицај грешке мерења на могућност детектовања зависности између латентних променљивих. Уводи се однос шума и сигнала који мери утицај грешке. Посматрају се подаци који задовољавају претпоставке класичног адитивног модела грешке мерења.

У раду се обрађују и упоређују различити тестови засновани на $dCov$ мери зависности, као и њихова проширења различитим језгрима. Разматра се и коришћење поновљених мерења ради ублажавања утицаја грешке мерења. Перформансе предложених тестова се разматрају при различитим нивоима зависности и различитој јачини шума, за класичне и усмерене податке. Након примене поступка тестирања предложеног у овом раду, описују се отворена питања и могућа проширења представљене методологије, са посебним освртом на податке који су истовремено оптерећени грешкама у мерењу и цензурисањем.

Кључне речи: статистика, тестирање, независност, грешке, латентне променљиве, цензурисање, карактеристичне функције, корелација, коваријација

Садржај

1	Увод	1
1.1	Класичан адитивни модел грешке мерења	2
1.2	Однос шума и сигнала	5
1.3	Дефиниција и особине $dCov$ мере	8
2	Тестови независности засновани на $dCov$ мери у случају комплетних података	12
2.1	Тест заснован на $dCov$ мери	12
2.2	Алтернативан поступак тестирања	16
2.3	Проширења заснована на језгрима	20
2.4	Емпиријска студија	23
3	Примена на реалним подацима	35
3.1	Зависност крвног притиска и индекса телесне масе	36
3.2	Зависност крвног притиска и укупног холестерола у крви . . .	38
4	Могућност проширење тестова независности на цензурисане податке	42
5	Закључак	49
	Библиографија	51

Глава 1

Увод

Тестирање независности случајних променљивих је један од основних проблема савремене статистичке анализе и има примену у различитим научним областима, као што су медицина, економија и машинско учење. Оно има значајну улогу у изградњи статистичких модела и избору променљивих. Класичне мере зависности, попут Пирсоновог коефицијента корелације, су ефикасне у откривању линеарне зависности, али у присуству нелинеарне или неког сложенијег облика зависности често греше. Због тога је развијена алтернативна метода заснована на $dCor$ мери зависности, која се може применити на различите типове података због чега чини изузетно погодан алат.

У практичним применама анализе података често се јавља значајан проблем грешака у мерењу. У многим реалним ситуацијама променљиве од интереса није могуће непосредно посматрати, већ су доступне само њихове измерене вредности, које су оптерећене грешкама. Оне могу настати из различитих разлога, као што су мерни инструменти који нису довољно прецизни, утицај окружења, начин прикупљања или складиштења података и људски фактор. У чланку [9] је описан проблем грешака у мерењу у медицинским истраживањима.

Присуство оваквих грешака може нарушити структуру зависности између променљивих од значаја, што доводи до пристрасних оцена и непоузданих закључака. Због тога је потребно модификовати стандардне тестове независности. Са циљем превазилажења овог проблема, предложене су модификације класичне $dCor$ мере.

Грешка у мерењу се може дефинисати као разлика стварне и измерене вредности променљиве. Постоје две врсте грешака у мерењу:

- систематска грешка,
- случајна грешка.

За разлику од случајне грешке систематска грешка представља предвидиво одступање од стварне вредности. Овај рад ће бити посвећен случајним грешкама у мерењу, и то класичном адитивном моделу грешке мерења, разматраном у [11] и [1].

Осим проблема грешака мерења, у анализи података се често јавља и проблем цензурисаних података. Због тога постоји потреба за тестом који у обзир узима и грешке мерења и цензурисане податке. Такви случајеви се могу јавити у медицинским истраживањима.

У раду [6] изложен је проблем тестирања независности случајних величина када постоји случајна грешка у мерењу и описан је поступак тестирања заснован на $dCov$ мери зависности. Аутори уводе претпоставку о доступности поновљеног мерења. Проширења $dCov$ мере различитим језгрима су обрађена у [4]. Тестирање у случају када податке истовремено погађају цензурисање и грешке у мерењу још увек није детаљно обрађено у литератури. У [3] је дат модел који описује такве податке. Примена $dCov$ мере за тестирање независности цензурисаних података је обрађена у раду [5].

1.1 Класичан адитивни модел грешке мерења

Нека су $X \in \mathbb{R}^p$ и $Y \in \mathbb{R}^q$ променљиве од интереса, а $U \in \mathbb{R}^p$ и $V \in \mathbb{R}^q$, редом, измерене вредности тих променљивих. Ако важи

$$U = X + \varepsilon,$$

$$V = Y + \delta,$$

где је $\varepsilon \in \mathbb{R}^p$ независна од X , а $\delta \in \mathbb{R}^q$ независна од Y и $E(\varepsilon) = E(\delta) = 0$, онда се ради о класичном адитивном моделу грешке мерења, где су ε и δ грешке.

Нека даље важе следеће претпоставке:

- 1) ε и δ су независне од (X, Y) ;
- 2) ε и δ су независне;
- 3) $E(\varepsilon) = E(\delta) = 0$;
- 4) грешке у мерењу није могуће непосредно посматрати;
- 5) X и Y не морају бити независне.

У готово свим емпиријским истраживањима посматране променљиве подлежу одређеном степену грешке мерења. Уколико посматрана променљива узима вредности из непрекидног скупа она не може бити измерена са апсолутном тачношћу. Узрок томе су мерни инструменти који имају ограничену прецизност. У пракси свако мерење представља апроксимацију стварне вредности, а део информација се губи. Код многих инструмената је присутно заокруживање, односно дискретизација, резултата. Тада податке не описује класичан адитивни модел грешке мерења већ модел облика $U = R(X + \varepsilon)$, где је $R(\cdot)$ функција дискретизације. На пример, висина човека може узети било коју реалну вредност из одређеног интервала, док мерни инструменти омогућавају мерење само са ограниченом прецизношћу. Слично важи за телесну тежину, крвни притисак, ниво глукозе у крви, концентрацију холестерола или температуру. Мерења висине, тежине и крвног притиска представљају комбинацију дискретизације и адитивне грешке, случајне грешке могу настати због положаја тела и главе, померања испитаника током мерења, различитог читавања скале од стране мерног лица, нестабилности подлоге код мерења висине и тежине и дисање током мерења, положај руке, као и тренутног психофизичког стања испитаника приликом мерења притиска.

Осим у медицинским и биолошким истраживањима грешке мерења се могу јавити и у другим областима. У економији се јављају при процени дохотка домаћинства, потрошње, тржишне вредности некретнина, приноса финансијских инструмената или инфлације. Грешке настају у физици при мерењу времена, брзине или електричног напона; навигацији при одређивању координата помоћу *GPS* уређаја; метеорологији при мерењу температуре, влажности ваздуха и брзине ветра.

Свако мерење има ограничену прецизност и подложно је случајним осцилацијама инструмента, утицају околине и људском фактору приликом мерења или уноса података. Латентна (стварна) вредност променљиве остаје непозната, а позната је посматрана вредност која садржи грешку. Уколико се може сматрати да је ефекат заокруживања мали у поређењу са осталим изворима варијабилности подаци се могу уклопити у класичан адитивни модел грешке мерења. Дакле, класични адитивни модел представља добру апроксимацију за мерења код којих је резолуција инструмента довољно фина.

Усмерени типови података

Усмерени подаци представљају посебну врсту података чије су вредности правци или оријентације. Они се јављају у областима бројних природних наука.

У случају дводимензионалних усмерених података (циркуларних података) правац се често описује углом θ , који се може представити јединичним вектором $(\cos \theta, \sin \theta)$. Анализа усмерених података се заснива на посебним статистичким моделима и методама, због њихове природе. Поступци тестирања независности описани у овом поглављу се могу применити и на усмерене типове података.

Класичан адитивни модел грешке мерења се може уопштити и на усмерене типове података. Нека су X и Y латентне променљиве циркуларног типа, а ε и δ одговарајуће грешке у мерењу, које су такође циркуларног типа. Адитивни модел се дефинише са

$$U = (X + \varepsilon) \mod 2\pi,$$

$$V = (Y + \delta) \mod 2\pi,$$

где су U и V посматране променљиве, а сабирање се врши по модулу 2π .

Више о усмереним типовима података се може прочитати у [7], а у наставку је дат преглед неких од најзначајнијих циркуларних расподела које ће бити коришћене у симулацијама.

Фон Мизесова расподела

Једна од најпознатијих циркуларних расподела је фон Мизесова расподела са параметрима $\mu \in \mathbb{R}$ и $\kappa > 0$, $VM(\mu, \kappa)$. Њена функција густине је дата са:

$$f(z) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(z-\mu)}, \quad z \in [a - \pi, a + \pi),$$

за било коју вредност $a \in \mathbb{R}$, где $I_\nu(x) = \frac{1}{\pi} \int_0^\pi e^{x \cos \theta} \cos(\nu \theta) d\theta$ представља модификовану Беселову функцију прве врсте реда ν , када је ν цео број.

Обавијене расподеле

Нека је Z^* случајна променљива која узима вредности из скупа $\mathbb{R}$ и чија је функција густине f^* . Тада је са

$$Z = Z^* \bmod 2\pi$$

дефинисана случајна променљива која настаје обавијањем Z^* око јединичне кружнице. Густина расподеле променљиве Z је дата са:

$$f(\theta) = \sum_{k=-\infty}^{+\infty} f^*(\theta + 2k\pi), \quad \theta \in [0, 2\pi).$$

Неки од примера су обавијена нормална $WN(\mu, \sigma^2)$ и обавијена Лапласова расподела $WL(\mu, b)$.

1.2 Однос шума и сигнала

Однос шума и сигнала NSR (од енгл. *Noise-to-Signal Ratio*) представља однос јачине шума и јачине сигнала. Нека је $U = X + \varepsilon$, где X представља сигнал, а ε представља шум.

Уколико се ради о класичним (линеарним) подацима NSR се рачуна као:

$$NSR_U = \frac{E(\varepsilon^2)}{E(X^2)}.$$

Уколико се посматрају усмерени (кружни) подаци NSR се рачуна као:

$$NSR_U = \frac{1 - R_\varepsilon}{1 - R_X}.$$

Величина R_Z се назива средња дужина резултанте (или концентрациони параметар) и дефинисана је са $R_Z = |\mathbb{E}(e^{iZ})|$ за случајну величину усмереног типа Z . Кружна дисперзија случајне величине Z се дефинише као $1 - R_Z$. Употребом Ојлерове формуле средња дужина резултанте се може изразити у следећем облику:

$$|E(e^{iZ})| = |E(\cos Z + i \sin Z)| = \sqrt{(E \cos Z)^2 + (E \sin Z)^2}.$$

Случајна величина Z дефинисана на следећи начин

$$Z = \begin{cases} X, p, \\ Y, 1 - p, \end{cases}$$

представља мешавину случајних величина X и Y са вероватноћом $p \in [0, 1]$. Оваква расподела се означава са $Z \sim Mix(X, Y, p)$. Уколико Z има расподелу класичног типа важи $EZ = pEX + (1 - p)EY$ и $EZ^2 = pEX^2 + (1 - p)EY^2$, а уколико има расподелу усмереног типа важи:

$$\begin{aligned} |E(e^{iZ})| &= |pE(e^{iX}) + (1 - p)E(e^{iY})| \\ &= |pE(\cos X) + (1 - p)E(\cos Y) + i(pE(\sin X) + (1 - p)E(\sin Y))| \\ &= \sqrt{(pE(\cos X) + (1 - p)E(\cos Y))^2 + (pE(\sin X) + (1 - p)E(\sin Y))^2}. \end{aligned}$$

У наставку се налазе примери расподела кружног типа које ће се јављати у симулацијама са формулама за рачунање средње дужине резултанте.

1) Нека $Z \sim VM(\mu, \kappa)$. Тада важи:

$$E(\cos Z) = \frac{I_1(\kappa)}{I_0(\kappa)} \cos \mu,$$

$$E(\sin Z) = \frac{I_1(\kappa)}{I_0(\kappa)} \sin \mu.$$

Величина R_Z се рачуна као:

$$R_Z = \sqrt{\left(\frac{I_1(\kappa)}{I_0(\kappa)} \cos \mu\right)^2 + \left(\frac{I_1(\kappa)}{I_0(\kappa)} \sin \mu\right)^2} = \frac{I_1(\kappa)}{I_0(\kappa)} \sqrt{\cos^2 \mu + \sin^2 \mu} = \frac{I_1(\kappa)}{I_0(\kappa)},$$

где се за задати параметар κ вредности $I_0(\kappa)$ и $I_1(\kappa)$ могу израчунати позивом функције $besselI(x, nu)$ у R-у.

2) Нека $Z \sim \text{Mix}(X, Y, p)$, где $X \sim VM(\mu_1, \kappa_1)$, $Y \sim VM(\mu_2, \kappa_2)$. Тада

$$R_Z =$$

$$\sqrt{\left(p \frac{I_1(\kappa_1)}{I_0(\kappa_1)} \cos \mu_1 + (1-p) \frac{I_1(\kappa_2)}{I_0(\kappa_2)} \cos \mu_2\right)^2 + \left(p \frac{I_1(\kappa_1)}{I_0(\kappa_1)} \sin \mu_1 + (1-p) \frac{I_1(\kappa_2)}{I_0(\kappa_2)} \sin \mu_2\right)^2}.$$

За $\mu_1 = \mu_2 = 0$ је

$$R_Z = p \frac{I_1(\kappa_1)}{I_0(\kappa_1)} + (1-p) \frac{I_1(\kappa_2)}{I_0(\kappa_2)} = pR_X + (1-p)R_Y.$$

За $\mu_1 = 0, \mu_2 = \frac{\pi}{2}$:

$$R_Z = \sqrt{\left(p \frac{I_1(\kappa_1)}{I_0(\kappa_1)}\right)^2 + \left((1-p) \frac{I_1(\kappa_2)}{I_0(\kappa_2)}\right)^2} = \sqrt{(pR_X)^2 + ((1-p)R_Y)^2}.$$

За $\mu_1 = 0, \mu_2 = \pi$:

$$R_Z = \left| p \frac{I_1(\kappa_1)}{I_0(\kappa_1)} - (1-p) \frac{I_1(\kappa_2)}{I_0(\kappa_2)} \right| = |pR_X - (1-p)R_Y|.$$

3) Нека $Z \sim WN(\mu, \sigma^2)$ и нека је $\rho = e^{-\frac{\sigma^2}{2}}$. Тада се показује да важи:

$$E(\cos Z) = \rho \cos \mu,$$

$$E(\sin Z) = \rho \sin \mu,$$

одакле следи да је :

$$R_Z = \sqrt{\rho^2 \cos^2 \mu + \rho^2 \sin^2 \mu} = \rho.$$

Некада се у литератури уместо ознаке $Z \sim WN(\mu, \sigma^2)$ користи и ознака $Z \sim WN(\mu, \rho)$.

4) Нека $Z \sim \text{Mix}(X, Y, p)$, где $X \sim WN(\mu_1, \rho_1)$ $Y \sim WN(\mu_2, \rho_2)$. Тада

$$R_Z = \sqrt{(p\rho_1 \cos \mu_1 + (1-p)\rho_2 \cos \mu_2)^2 + (p\rho_1 \sin \mu_1 + (1-p)\rho_2 \sin \mu_2)^2}.$$

За $\mu_1 = \mu_2 = 0$ је

$$R_Z = p\rho_1 + (1-p)\rho_2 = pR_X + (1-p)R_Y.$$

За $\mu_1 = 0, \mu_2 = \frac{\pi}{2}$:

$$R_Z = \sqrt{(p\rho_1)^2 + ((1-p)\rho_2)^2} = \sqrt{(pR_X)^2 + ((1-p)R_Y)^2}.$$

За $\mu_1 = 0$, $\mu_2 = \pi$:

$$R_Z = |p\rho_1 - (1-p)\rho_2| = |pR_X - (1-p)R_Y|.$$

5) Нека $Z^* \sim \mathcal{L}(\mu, b)$ и нека је $Z = Z^* \bmod 2\pi$. Тада $Z \sim W\mathcal{L}(\mu, b)$. Ако је $z = z^* \bmod 2\pi$ тада за свако $z^* \in \mathbb{R}$ постоји јединствено $k \in \mathbb{Z}$ тако да важи $z^* = z + 2k\pi$. Из $e^{i(Z+2k\pi)} = e^{iZ}e^{2ik\pi} = e^{iZ}(\cos 2k\pi + i \sin 2k\pi) = e^{iZ}$ следи

$$R_Z = |E(e^{iZ})| = |E(e^{i(Z+2k\pi)})| = |E(e^{iZ^*})|,$$

па се рачунање средње дужине резултанте за обавијену Лапласову расподелу своди на рачунање карактеристичне функције у тачки $t = 1$ за одговарајућу Лапласову расподелу. Функција густине расподеле случајне Z^* величине је дата са: $f^*(x) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$, $x \in \mathbb{R}$. Тада је:

$$\begin{aligned} \phi_{Z^*}(1) &= E(e^{iZ^*}) = \int_{-\infty}^{+\infty} \frac{e^{ix}}{2b} e^{-\frac{|x-\mu|}{b}} dx = \frac{1}{2b} \left(\int_{-\infty}^{\mu} e^{ix+\frac{x-\mu}{b}} dx + \int_{\mu}^{+\infty} e^{ix-\frac{x-\mu}{b}} dx \right) \\ &= \frac{e^{i\mu}}{2b} \left(\int_{-\infty}^0 e^{ix+\frac{x}{b}} dx + \int_0^{+\infty} e^{ix-\frac{x}{b}} dx \right) = \frac{e^{i\mu}}{2b} \left(\int_{-\infty}^0 e^{x(i+\frac{1}{b})} dx + \int_0^{+\infty} e^{x(i-\frac{1}{b})} dx \right) \\ &= \frac{e^{i\mu}}{2b} \left(\frac{1}{i+\frac{1}{b}} - \frac{1}{i-\frac{1}{b}} \right) = \frac{e^{i\mu}}{2b} \left(\frac{\frac{1}{b}-i}{\frac{1}{b^2}+1} - \frac{\frac{1}{b}+i}{\frac{1}{b^2}+1} \right) = \frac{e^{i\mu}}{2} \frac{\frac{2}{b}}{\frac{1}{b}+b} = e^{i\mu} \frac{1}{1+b^2}, \\ \text{а } R_Z &= |e^{i\mu} \frac{1}{1+b^2}| = \frac{1}{1+b^2} |e^{i\mu}| = \frac{1}{1+b^2}. \end{aligned}$$

Дефиниција односа шума и сигнала се може уопштити на вишедимензионе случајне величине. Нека $X = (X_1, \dots, X_s)$, $\epsilon = (\epsilon_1, \dots, \epsilon_s) \in \mathbb{R}^s$ где је $s > 1$. Тада је

$$NSR_U = [NSR_{U_i}]_{i=1}^s,$$

где је $U_i = X_i + \epsilon_i$, за $i = 1, \dots, s$.

1.3 Дефиниција и особине $dCov$ мере

У статистици $dCov$ (скраћено од енгл. Distance covariance) представља меру зависности случајних вектора $X \in \mathbb{R}^p$ и $Y \in \mathbb{R}^q$, који не морају бити истих димензија. Пирсонов коефицијент корелације дефинисан са:

$$\rho_{(X,Y)} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sigma_X \sigma_Y},$$

где су са σ_X и σ_Y означена стандардна одступања одговарајућих случајних величина, представља класичну меру зависности. Ова мера је осетљива на линеарну зависност две случајне променљиве, али није ефикасна у откривању неког сложенијег облика зависности. Због тога је настала потреба за мером која ће превазићи овај недостатак. Као решење тог проблема у раду [10] је године предложена $dCov$ мера.

Дефиниција 1 Нека су X и Y променљиве од интереса и нека су (X_1, Y_1) и (X_2, Y_2) независне копије пара (X, Y) . Тада је са:

$$\begin{aligned} dCov^2(X, Y) = & \mathbb{E}(\|X - X_1\| \|Y - Y_1\|) + \mathbb{E}(\|X - X_1\|)\mathbb{E}(\|Y - Y_1\|) \\ & - 2\mathbb{E}(\|X - X_1\| \|Y - Y_2\|) \end{aligned}$$

дефинисана $dCov$ **мера зависности** између случајних променљивих X и Y .

Помоћу једнакости $\mathbb{E}(\|Y - Y_1\|) = \mathbb{E}(\|Y - Y_2\|)$ се може извести веза између класичне Пирсонове коваријације и $dCov$ мере:

$$dCov^2(X, Y) = Cov(\|X - X_1\|, \|Y - Y_1\|) - 2Cov(\|X - X_1\|, \|Y - Y_2\|).$$

Алтернативно, $dCov$ мера се може дефинисати као:

$$dCov^2(X, Y) = \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} |\varphi_{(X,Y)}(t, s) - \varphi_X(t)\varphi_Y(s)|^2 \omega_1(t) \omega_2(s) ds dt, \quad (1.1)$$

где су $\varphi_{(X,Y)}(t, s) = \mathbb{E}e^{i(t^\top X + s^\top Y)}$, $\varphi_X(t) = \mathbb{E}e^{it^\top X}$ и $\varphi_Y(s) = \mathbb{E}e^{is^\top Y}$ карактеристичне функције а $\omega_1(t) = \Gamma\left(\frac{1+p}{2}\right) (\pi^{\frac{1+p}{2}} \|t\|^{1+p})^{-1}$ и $\omega_2(s) = \Gamma\left(\frac{1+q}{2}\right) (\pi^{\frac{1+q}{2}} \|s\|^{1+q})^{-1}$ тежинске функције.

У наставку је показано да су дате дефиниције еквивалентне. Еуклидска норма се помоћу Фуријеове трансформације може записати у облику:

$$\|x\| = \int_{\mathbb{R}^p} c_p \frac{1 - \cos(t^\top x)}{\|t\|^{p+1}} dt, \quad x \in \mathbb{R}^p,$$

где је $c_p = \Gamma\left(\frac{1+p}{2}\right) \pi^{-\frac{1+p}{2}}$. Како је $c_p \|t\|^{-p-1} \sin(t^\top x)$ непарна функција на $\mathbb{R}^p$ важи:

$$\|x\| = \int_{\mathbb{R}^p} c_p \frac{1 - e^{it^\top x}}{\|t\|^{p+1}} dt, \quad x \in \mathbb{R}^p.$$

Одатле следи да је:

$$\begin{aligned}
\mathbb{E}(\|X - X_1\| \|Y - Y_1\|) &= \mathbb{E} \left(\int_{\mathbb{R}^p} c_p \frac{1 - e^{it^\top(X-X_1)}}{\|t\|^{p+1}} dt \int_{\mathbb{R}^q} c_q \frac{1 - e^{is^\top(Y-Y_1)}}{\|s\|^{q+1}} ds \right) \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} \mathbb{E} \left((1 - e^{it^\top(X-X_1)})(1 - e^{is^\top(Y-Y_1)}) \right) \omega_1(t) \omega_2(s) ds dt \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} \mathbb{E} \left(1 - e^{it^\top(X-X_1)} - e^{is^\top(Y-Y_1)} + e^{it^\top(X-X_1)} e^{is^\top(Y-Y_1)} \right) \omega_1(t) \omega_2(s) ds dt \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} (1 - |\varphi_X(t)|^2 - |\varphi_Y(s)|^2 + |\varphi_{(X,Y)}(t, s)|^2) ds dt,
\end{aligned}$$

где је $\omega_1(t) = \frac{c_p}{\|t\|^{1+p}}$ и $\omega_2(s) = \frac{c_q}{\|s\|^{1+q}}$, а друга једнакост је последица Фуби-нијево теореме.

Слично,

$$\begin{aligned}
&\mathbb{E}(\|X - X_1\|) \mathbb{E}(\|Y - Y_1\|) \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} (1 - |\varphi_X(t)|^2) (1 - |\varphi_Y(s)|^2) \omega_1(t) \omega_2(s) ds dt, \\
&\mathbb{E}(\|X - X_1\| \|Y - Y_2\|) \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} \left(1 - |\varphi_X(t)|^2 - |\varphi_Y(s)|^2 + \mathcal{R} \left(\varphi_{(X,Y)}(t, s) \overline{\varphi_X(t) \varphi_Y(s)} \right) \right) \\
&\quad \omega_1(t) \omega_2(s) ds dt,
\end{aligned}$$

па је

$$\begin{aligned}
&\text{dCov}^2(X, Y) \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} (1 - |\varphi_X(t)|^2 - |\varphi_Y(s)|^2 + |\varphi_{(X,Y)}(t, s)|^2 \\
&\quad + 1 - |\varphi_X(t)|^2 - |\varphi_Y(s)|^2 + |\varphi_X(t)|^2 |\varphi_Y(s)|^2 \\
&\quad - 2 + 2|\varphi_X(t)|^2 + 2|\varphi_Y(s)|^2 - 2\mathcal{R} \left(\varphi_{(X,Y)}(t, s) \overline{\varphi_X(t) \varphi_Y(s)} \right)) \omega_1(t) \omega_2(s) ds dt \\
&= \int_{\mathbb{R}^p} \int_{\mathbb{R}^q} |\varphi_{(X,Y)}(t, s) - \varphi_X(t) \varphi_Y(s)|^2 \omega_1(t) \omega_2(s) ds dt.
\end{aligned}$$

Дефиниција 2 Нормализована верзија dCov се може дефинисати са:

$$\text{dCor}^2(X, Y) = \frac{\text{dCov}^2(X, Y)}{\text{dCov}(X, X) \text{dCov}(Y, Y)}.$$

Неке важне особине $dCov$ мере и њене нормализације су:

1. $dCov^2(X, Y) \geq 0$;
2. $dCov(X, Y) = 0 \iff X$ и Y су независне;
3. $dCor(X, Y) \in [0, 1]$;
4. $dCor(X, Y) = 0 \iff X$ и Y су независне.

Друга особина је најзначајнија. Она следи из облика (1.1).

Такође, она може открити и линеарну и нелинеарну зависност, а не захтева да X и Y буду исте димензије.

Уколико су X и Y променљиве циркуларног типа оне се применом трансформације

$$\theta \rightarrow (\cos \theta, \sin \theta)$$

могу представити као двокомпонентни вектори у $\mathbb{R}^2$. На тим векторским презентацијама се може применити $dCov$, па се због тога циркуларни подаци могу укључити у исти оквир као и класични.

Глава 2

Тестови независности засновани на $dCov$ мери у случају комплетних података

У овом поглављу ће бити представљене методе тестирања независности засноване на $dCov$ мери, које се могу применити на класичне и усмерене типове података. Осим тога, биће представљена и проширења различитим језгрима. Перформансе предложених тестова се на крају анализирају и пореде кроз симулационе студије. За имплементацију теста се користи пермутациона бутстреп метода (енг. *bootstrap*) која ће у даљем тексту бити детаљније објашњена.

2.1 Тест заснован на $dCov$ мери

Нека су $X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ и $Y = (Y_1, \dots, Y_q)^\top \in \mathbb{R}^q$ латентни случајни вектори који су контаминирани грешкама приликом мерења. У том случају се не може директно посматрати (X, Y) , већ (U, V) , где је

$$U = X + \varepsilon, \quad V = Y + \delta,$$

а случајни вектори $\varepsilon \in \mathbb{R}^p$ и $\delta \in \mathbb{R}^q$ представљају грешке у мерењу, које се не могу директно посматрати. Латентне променљиве се називају сигналом, док се грешке називају шумом. Нека важе претпоставке класичног адитивног модела грешке мерења.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Циљ је тестирање хипотезе

$$H_0 : X \text{ и } Y \text{ су независне,}$$

против

$$H_1 : X \text{ и } Y \text{ нису независне,}$$

на основу простог случајног узорка $\{(U_i, V_i)\}_{i=1, \dots, n}$.

Нека су φ_X , φ_Y и $\varphi_{(X,Y)}$ карактеристичне функције од X , Y и (X, Y) , редом. Аналогно, φ_U , φ_V и $\varphi_{(U,V)}$ су карактеристичне функције од U , V и (U, V) , а φ_ε и φ_δ су карактеристичне функције од ε и δ . Под нултом хипотезом важи

$$\varphi_{(X,Y)}(t, s) = \varphi_X(t)\varphi_Y(s), \text{ за } \forall t \in \mathbb{R}^p, s \in \mathbb{R}^q,$$

односно

$$\int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(X,Y)}(t, s) - \varphi_X(t)\varphi_Y(s)|^2 \omega_1(t)\omega_2(s) dt ds = 0, \quad (2.1)$$

где су $\omega_1(\cdot)$ и $\omega_2(\cdot)$ позитивне тежинске функције, такве да је интеграл из (2.1) добро дефинисан.

Из особина карактеристичне функције, за претпостављени модел важи следеће:

$$\begin{aligned} \varphi_U(t) &= \varphi_X(t) \varphi_\varepsilon(t), \text{ за } \forall t \in \mathbb{R}^p, \\ \varphi_V(s) &= \varphi_Y(s) \varphi_\delta(s), \text{ за } \forall s \in \mathbb{R}^q, \\ \varphi_{(U,V)}(t, s) &= \varphi_{(X,Y)}(t, s) \varphi_\varepsilon(t) \varphi_\delta(s), \text{ за } \forall t \in \mathbb{R}^p, s \in \mathbb{R}^q. \end{aligned}$$

Нека даље важе претпоставке да је $\varphi_\varepsilon(t) \neq 0$ за $\forall t \in \mathbb{R}^p$ и $\varphi_\delta(s) \neq 0$ за $\forall s \in \mathbb{R}^q$. Тада се интеграл са леве стране једнакости (2.1) може трансформисати на следећи начин:

$$\begin{aligned} & \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(X,Y)}(t, s) - \varphi_X(t)\varphi_Y(s)|^2 \omega_1(t)\omega_2(s) dt ds \\ &= \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(U,V)}(t, s) \varphi_\varepsilon^{-1}(t) \varphi_\delta^{-1}(s) - \varphi_U(t) \varphi_\varepsilon^{-1}(t) \varphi_V(s) \varphi_\delta^{-1}(s)|^2 \omega_1(t)\omega_2(s) dt ds \\ &= \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(U,V)}(t, s) - \varphi_U(t)\varphi_V(s)|^2 (\omega_1(t) |\varphi_\varepsilon^{-1}(t)|^2) (\omega_2(s) |\varphi_\delta^{-1}(s)|^2) dt ds. \end{aligned}$$

Ако се $\tilde{\omega}_1(\cdot)$ и $\tilde{\omega}_2(\cdot)$ дефинишу са

$$\tilde{\omega}_1(t) = \omega_1(t) |\varphi_\varepsilon^{-1}(t)|^2 \quad \text{и} \quad \tilde{\omega}_2(s) = \omega_2(s) |\varphi_\delta^{-1}(s)|^2$$

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

једнакост (2.1) се може записати у облику

$$\int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(U,V)}(t, s) - \varphi_U(t)\varphi_V(s)|^2 \tilde{\omega}_1(t) \tilde{\omega}_2(s) dt ds = 0. \quad (2.2)$$

Како интеграли из (2.1) и (2.2) имају исти облик, а разликују се само по аргументима, важи закон инваријантности. Суштина закона инваријантности, у овом случају, је у томе да се за тестирање независности не морају реконструисати латентне променљиве, већ су довољне посматране променљиве, које садрже и грешке мерења.

Може се одабрати специјалан облик тежинских функција

$$\tilde{\omega}_1(t) = \Gamma\left(\frac{1+p}{2}\right) (\pi^{\frac{1+p}{2}} \|t\|^{1+p})^{-1}, \quad \tilde{\omega}_2(s) = \Gamma\left(\frac{1+q}{2}\right) (\pi^{\frac{1+q}{2}} \|s\|^{1+q})^{-1}$$

тако да важи:

$$\text{dCov}^2(U, V) = \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} |\varphi_{(U,V)}(t, s) - \varphi_U(t)\varphi_V(s)|^2 \tilde{\omega}_1(t) \tilde{\omega}_2(s) dt ds. \quad (2.3)$$

Следећа теорема показује да се информација о независности латентних променљивих не губи ни у присуству грешака мерења.

Теорема 1 *Нека је $E(\|X\| + \|Y\| + \|\varepsilon\| + \|\delta\|) < +\infty$ у класичном адитивном моделу грешке мерења:*

$$U = X + \varepsilon, \quad V = Y + \delta,$$

где грешке мерења ε и δ задовољавају услове:

$$\varphi_\varepsilon(t) \neq 0, \quad \forall t \in \mathbb{R}^p,$$

$$\varphi_\delta(s) \neq 0, \quad \forall s \in \mathbb{R}^q.$$

Тада су случајне променљиве X и Y независне ако и само ако је

$$\text{dCov}^2(U, V) = 0.$$

Доказ. Познато је да важи следећа еквиваленција:

$$\text{dCov}^2(U, V) = 0 \quad \Longleftrightarrow \quad U \text{ и } V \text{ су независне случајне величине.}$$

Због тога је довољно показати да су X и Y независне ако су U и V независне случајне променљиве. Како су у класичном адитивном моделу грешке мерења

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

међусобно независне, из независности X и Y следи независност U и V . Остаје показати да из независности U и V следи независност X и Y .

Нека су U и V независне случајне величине. Тада важи:

$$\begin{aligned}\varphi_{(X,Y)}(t, s) &= \varphi_{(U,V)}(t, s) \varphi_\varepsilon^{-1}(t) \varphi_\delta^{-1}(s) \\ &= \varphi_U(t) \varphi_V(s) \varphi_\varepsilon^{-1}(t) \varphi_\delta^{-1}(s) \\ &= \varphi_U(t) \varphi_\varepsilon^{-1}(t) \varphi_V(s) \varphi_\delta^{-1}(s) \\ &= \varphi_X(t) \varphi_Y(s),\end{aligned}$$

одакле следи да су X и Y независне случајне величине, чиме је теорема доказана. $\square$

Пожељно је дефинисати меру степена зависности која узима вредности из интервала $[0, 1]$. Из неједнакости Коши-Шварца следи:

$$\text{dCov}^2(U, V) \leq \text{dCov}(U, U) \text{dCov}(V, V),$$

па би природно било посматрати нормализацију

$$\frac{\text{dCov}^2(U, V)}{\text{dCov}(U, U) \text{dCov}(V, V)} \in [0, 1], \quad (2.4)$$

која достиже вредност 1 када су променљиве U и V линеарно зависне. Ипак циљ није мерење независности између U и V , већ између X и Y . Чак и када су X и Y потпуно линеарно зависне, нормализована величина (2.4) не мора достићи вредност 1, јер грешке утичу на дисперзију променљивих. Вредности $\text{dCov}(U, U)$ и $\text{dCov}(V, V)$ не описују дисперзију латентних променљивих. Тај проблем се може решити увођењем поновљених мерења.

Претпоставља се да је доступно поновљено мерење пара (U, V) које означавамо са $(\tilde{U}, \tilde{V})$ и које задовољава претпоставке класичног адитивног модела грешке мерења:

$$\begin{aligned}\tilde{U} &= X + \tilde{\varepsilon}, \\ \tilde{V} &= Y + \tilde{\delta},\end{aligned}$$

где су $\tilde{\varepsilon} \in \mathbb{R}^p$ и $\tilde{\delta} \in \mathbb{R}^q$ грешке настале приликом другог мерења. Претпоставља се и да су грешке ε и $\tilde{\varepsilon}$, као и δ и $\tilde{\delta}$ исто расподељене, односно да при поновљеном мерењу грешка има исту расподелу.

У раду [6] предложена је следећа мера:

$$\rho^2(X, Y) = \frac{\text{dCov}^2(U, V)}{\text{dCov}(U, \tilde{U}) \text{dCov}(V, \tilde{V})}. \quad (2.5)$$

Овако дефинисана мера омогућава да се степен зависности између латентних променљивих процењује на основу посматраних података који су контаминирани грешкама мерења.

Нека је доступан узорак $\{U_i, V_i\}_{i=1, \dots, n}$ обима n . Величина $\text{dCov}^2(U, V)$ се може апроксимирати V-статистиком

$$\begin{aligned} \widehat{\text{dCov}}^2(U, V) = & \frac{1}{n^2} \sum_{i,j} \|U_i - U_j\| \|V_i - V_j\| - \frac{2}{n^3} \sum_{i,j,k} \|U_i - U_j\| \|V_i - V_k\| \\ & + \frac{1}{n^4} \sum_{i,j,k,l} \|U_i - U_j\| \|V_k - V_l\|. \end{aligned} \quad (2.6)$$

Напомена. У раду [6] је уместо (2.6) предложена одговарајућа U-статистика. Измена у овом раду је направљена због ефикасније имплементације и краћег времена извршавања програма. Приликом репродуковања нумеричких примера из рада исти закључци су добијени и при коришћењу V-статистике. Због тога се у наставку рада користи V-статистика.

На основу простог случајног узорка $\{(U_i, V_i, \tilde{U}_i, \tilde{V}_i)\}_{i=1, \dots, n}$, где $(\tilde{U}_i, \tilde{V}_i)$ представља поновљено мерење одговарајућих величина (U_i, V_i) , се може израчунати узорачка верзија величине $\rho^2(X, Y)$:

$$\hat{\rho}^2(X, Y) = \frac{\widehat{\text{dCov}}^2(U, V) + \widehat{\text{dCov}}^2(U, \tilde{V}) + \widehat{\text{dCov}}^2(\tilde{U}, V) + \widehat{\text{dCov}}^2(\tilde{U}, \tilde{V})}{4\widehat{\text{dCov}}(U, \tilde{U})\widehat{\text{dCov}}(V, \tilde{V})}. \quad (2.7)$$

У (2.7) су $\widehat{\text{dCov}}^2(U, V)$, $\widehat{\text{dCov}}^2(U, \tilde{V})$, $\widehat{\text{dCov}}^2(\tilde{U}, V)$ и $\widehat{\text{dCov}}^2(\tilde{U}, \tilde{V})$ узорачке оцене исте теоријске величине $\text{dCov}^2(U, V)$. Аутори користе њихову аритметичку средину да би повећали ефикасност оцене и смањили дисперзију.

2.2 Алтернативан поступак тестирања

У одељку 2.1 претпостављено је да је за сваку посматрану јединицу доступно једно поновљено мерење. Та претпоставка омогућава конструкцију тест статистике (2.7) за испитивање независности латентних променљивих у присуству грешака у мерењу.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Међутим, у практичним применама може бити доступно и више од једног поновљеног мерења, а природно је очекивати да ће већи број поновљених мерења пружити више информација о латентним променљивама и довести до прецизнијег тестирања. Проширење статистике (2.7) за већи број поновљених мерења би подразумевало рачунање $dCov$ за све комбинације поновљених мерења. Са бројем поновљених мерења расте и број комбинација, па у симулационим студијама тај приступ постаје рачунски веома захтеван и спор.

У наставку се предлаже нова тест статистика, чије је увођење засновано на истој идеји као и (2.7), али која омогућава већу ефикасност.

Нека су X и Y латентне променљиве од значаја а U и V измерене променљиве, контаминирани грешком мерења, које задовољавају претпоставке класичног адитивног модела грешке мерења.

Нека је доступан узорак

$$\begin{aligned} U^j &= (U_1^j, U_2^j, \dots, U_n^j), j \in \{1, 2, \dots, m_u\}, \\ V^j &= (V_1^j, V_2^j, \dots, V_n^j), j \in \{1, 2, \dots, m_v\}, \end{aligned} \quad (2.8)$$

где је m_u број поновљених мерења променљиве X , m_v број поновљених мерења променљиве Y , n обим узорка, а U_i^j и V_i^j означавају j -то поновљено мерење i -те опсервације променљивих X и Y , редом, односно:

$$\begin{aligned} U_i^j &= X_i + \varepsilon_i^j, j \in \{1, 2, \dots, m_u\} \\ V_i^j &= Y_i + \delta_i^j, j \in \{1, 2, \dots, m_v\}, \end{aligned}$$

за $i \in \{1, 2, \dots, n\}$, где су $X_i \in \mathbb{R}^p$ и $Y_i \in \mathbb{R}^q$ стварне вредности обележја X и Y за i -ту опсервацију, а $\varepsilon_i^j \in \mathbb{R}^p$ и $\delta_i^j \in \mathbb{R}^q$ грешке настале при одговарајућем мерењу за одговарајућу опсервацију. Допуштено је да се број поновљених мерења за X и Y разликује, зато што је такав случај могућ у пракси, поготово ако се за истраживање користе претходно прикупљени подаци из доступних база података.

Као и у одељку 2.1 потребно је конструисати узорачку верзију величине:

$$\frac{dCov^2(U, V)}{dCov(U, U) dCov(V, V)},$$

која ће омогућити тестирање независности латентних променљивих на основу узорка (2.8).

Уместо нормализоване статистике ρ^2 , предлагемо алтернативу засновану на $dCov$, која се примењује на аритметичким срединама поновљених мерења. Конструкција је мотивисана истом интуицијом као и статистика ρ^2 , односно идејом да аритметичка средина више поновљених мерења прецизније описује латентне променљиве и смањује утицај грешака мерења. За разлику од статистике ρ^2 , чије би проширење захтевало рачунање $dCov$ велики број пута, код предложеног приступа се она рачуна исти број пута, независно од бројева m_u и m_v , што је веома важно у симулационим студијама.

Алтернативну тест статистику означавамо са T . Њена вредност се рачуна на следећи начин:

$$T = \widehat{dCor}^2(\bar{U}, \bar{V}) = \frac{\widehat{dCov}^2(\bar{U}, \bar{V})}{\widehat{dCov}(\bar{U}, \bar{U}) \widehat{dCov}(\bar{V}, \bar{V})}, \quad (2.9)$$

где је:

$$\begin{aligned} \widehat{dCov}^2(\bar{U}, \bar{V}) &= \frac{1}{n^2} \sum_{i,j} \|\bar{U}_i - \bar{U}_j\| \|\bar{V}_i - \bar{V}_j\| - \frac{2}{n^3} \sum_{i,j,k} \|\bar{U}_i - \bar{U}_j\| \|\bar{V}_i - \bar{V}_k\| \\ &\quad + \frac{1}{n^4} \sum_{i,j,k,l} \|\bar{U}_i - \bar{U}_j\| \|\bar{V}_k - \bar{V}_l\|, \end{aligned} \quad (2.10)$$

за прост случајан узорак:

$$\bar{U}_i = \frac{1}{m_u} \sum_{j=1}^{m_u} U_i^j, \quad \bar{V}_i = \frac{1}{m_v} \sum_{j=1}^{m_v} V_i^j,$$

где $i \in \{1, 2, \dots, n\}$. Ако латентне променљиве имају расподеле циркуларног типа, операција аритметичке средине се врши након трансформације

$$\theta \rightarrow (\cos \theta, \sin \theta).$$

Испоставља се да овако конструисан узорак задовољава претпоставке класичног адитивног модела грешке мерења. Важи:

$$\bar{U}_i = \frac{1}{m_u} \sum_{j=1}^{m_u} U_i^j = \frac{1}{m_u} \sum_{j=1}^{m_u} (X_i + \varepsilon_i^j) = X_i + \frac{1}{m_u} \sum_{j=1}^{m_u} \varepsilon_i^j = X_i + \bar{\varepsilon}_i,$$

$$\bar{V}_i = \frac{1}{m_v} \sum_{j=1}^{m_v} V_i^j = \frac{1}{m_v} \sum_{j=1}^{m_v} (Y_i + \delta_i^j) = Y_i + \frac{1}{m_v} \sum_{j=1}^{m_v} \delta_i^j = Y_i + \bar{\delta}_i,$$

за $i \in \{1, 2, \dots, n\}$. Величине $\bar{\varepsilon}_i$ и $\bar{\delta}_i$ су међусобно независне и независне од X_i и Y_i , редом, за $i \in \{1, 2, \dots, n\}$. Важи:

$$\mathbb{E}(\bar{\varepsilon}_i) = \mathbb{E}\left(\frac{1}{m_u} \sum_{j=1}^{m_u} \varepsilon_i^j\right) = \frac{1}{m_u} \sum_{j=1}^{m_u} \mathbb{E}(\varepsilon_i^j) = \frac{1}{m_u} \sum_{j=1}^{m_u} 0 = 0,$$

и аналогно $\mathbb{E}(\bar{\delta}_i) = 0$, за $i \in \{1, 2, \dots, n\}$. Очигледно је да се грешке не могу непосредно посматрати, па је показано да важе све претпоставке класичног адитивног модела грешке мерења.

Осим тога, може се показати и да је дисперзија грешке умањена овом трансформацијом. Нека су X, ε и U једнодимензионе случајне величине и нека је $\sigma_\varepsilon^2 = \mathbb{D}(\varepsilon_i^j)$, за свако $i \in \{1, 2, \dots, n\}$ и $j \in \{1, 2, \dots, m_u\}$. Под додатном претпоставком о међусобној независности грешака различитих мерења за исту опсервацију важи:

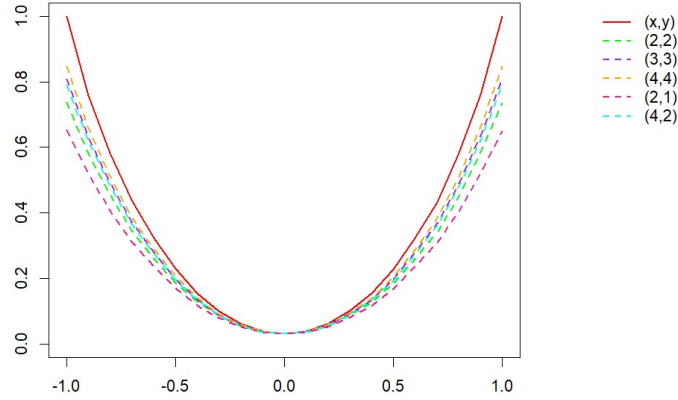
$$\mathbb{D}(\bar{\varepsilon}_i) = \mathbb{D}\left(\frac{1}{m_u} \sum_{j=1}^{m_u} \varepsilon_i^j\right) = \frac{1}{m_u^2} \sum_{j=1}^{m_u} \mathbb{D}(\varepsilon_i^j) = \frac{1}{m_u^2} \sum_{j=1}^{m_u} \sigma_\varepsilon^2 = \frac{1}{m_u^2} m_u \sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{m_u} < \sigma_\varepsilon^2.$$

Аналогно претходном, у једнодимензионом случају важи:

$$\mathbb{D}(\bar{\delta}_i) < \mathbb{D}(\delta_i^j).$$

Ови резултати се могу уопштити и на вишедимензионе податке.

Коришћење аритметичких средина поновљених мерења задржава структуру модела, а смањује дисперзију грешке мерења. Што је број поновљених мерења већи добија се прецизнија апроксимација латентне променљиве, чиме је оправдано апроксимирање величина $\text{dCov}(U, U)$ $\text{dCov}(V, V)$ са $\widehat{\text{dCov}}(\bar{U}, \bar{U})$ $\widehat{\text{dCov}}(\bar{V}, \bar{V})$ у статистици (2.9)



Слика 2.1: Поређење вредности тест статистике T за различите вредности m_u и m_v са емпиријском $dCor^2(X, Y)$, за обим узорка $n = 100$. На y оси су приказане вредности статистика, док је на x оси Пирсонов коефицијент корелације. Пуном линијом су приказане вредности добијене на основу узорка који није погођен грешком мерења, док су испрекиданим линијама приказане вредности добијене на основу поновљених мерења, код којих су присутне грешке. У заградама се налазе бројеви поновљених мерења, (m_u, m_v) .

На слици 2.1 се може уочити како се са повећањем броја поновљених мерења вредности тест статистике приближавају вредностима узорачке $dCor$ за латентне променљиве.

2.3 Проширења заснована на језгрима

За имплементирање емпиријске $dCov$ се у претходним одељцима користи еуклидска норма. Дефиниција $dCov$ мере се може проширити различитим језгрима. Тако се тестирање независности засновано на $dCov$ мери може прилагодити различитим типовима података, као што су на пример усмерени.

Нека је са $\|\cdot\|$ означена еуклидска норма и нека су $\mathcal{S}^{d_1} = \{x \in \mathbb{R}^{d_1+1} \mid \|x\| = 1\}$ и $\mathcal{S}^{d_2} = \{x \in \mathbb{R}^{d_2+1} \mid \|x\| = 1\}$ сфере димензија d_1 и d_2 , редом. Нека пар случајних величина (X, Y) узима вредности из простора $\mathcal{X} \times \mathcal{Y}$, где $\mathcal{X} \in \{\mathcal{S}^{d_1}, \mathbb{R}^p\}$ и $\mathcal{Y} \in \{\mathcal{S}^{d_2}, \mathbb{R}^q\}$. Нека су (X_1, Y_1) и (X_2, Y_2) независне копије пара (X, Y) .

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Дефиниција 3 Нека је $(\mathcal{X}, \mathcal{A})$ мерљив простор, а $\mathcal{M}$ скуп вероватносна мера на $(\mathcal{X}, \mathcal{A})$. Нека за симетричну функцију $K_1 : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ и за произвољан број $n \in \mathbb{N}$ важи:

$$\sum_{i,j=1}^n c_i c_j K_1(x_i, x_j) \leq 0, \quad (2.11)$$

за све $x_1, x_2, \dots, x_n \in \mathcal{X}$, $y_1, y_2, \dots, y_n \in \mathcal{Y}$ и произвољне реалне бројеве $c_1, c_2, \dots, c_n$ који задовољавају једнакост $\sum_{i=1}^n c_i = 0$. Тада се она назива **условно нејативно дефинишним језјром**. Уколико у (2.11) важи једнакост само ако је $c_1 = c_2 = \dots = c_n = 0$, онда се функције K_1 назива **сјројо нејативно дефинишним језјрима** (енг. *strictly negative definite kernels*). Нека $Q \in \mathcal{M}$ и нека је h функција интеграбилна по Q , таква да важи

$$\int_{\mathcal{X}} h(x) dQ(x) = 0.$$

Језјро K_1 је **јакко нејативно дефинишно** (енг. *strongly negative definite kernel*), ако је нејативно дефинишно и из једнакости:

$$\int_{\mathcal{X}} \int_{\mathcal{X}} K_1(x, y) h(x) h(y) dQ(x) dQ(y) = 0$$

слиди да је $h \equiv 0$ скоро свуда за $Q \in \mathcal{M}$.

Слично важи и за функцију $K_2 : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$.

Дефиниција 4 Следећим изразом:

$$\begin{aligned} d\text{Cov}_K^2(X, Y) &= \mathbb{E}(K_1(X, X_1)K_2(Y, Y_1)) + \mathbb{E}(K_1(X, X_1))\mathbb{E}(K_2(Y, Y_1)) \\ &\quad - 2\mathbb{E}(K_1(X, X_1)K_2(Y, Y_2)), \end{aligned}$$

где је $K = (K_1, K_2)$, а K_1 и K_2 су јакко нејативно дефинишна језјра, се дефинише $d\text{Cov}$ **индукована језјрима**.

Дефиниција 5 Уколико се прате ознаке из претходне дефиниције, изразом:

$$d\text{Cor}_K^2(X, Y) = \frac{d\text{Cov}_K^2(X, Y)}{d\text{Cov}_K(X, X) d\text{Cov}_K(Y, Y)}$$

се дефинише $d\text{Cor}$ **индукована језјрима**.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Уколико K_1 и K_2 имају исти функционални облик оба се означавају само са K .

За $K(x, y) = E(x, y) = \|x - y\|$ се добија облик (2.3).

Језгра која ће бити коришћена у остатку рада су

$$\begin{aligned} E(x, y) &= \|x - y\|, \\ K_k(x, y) &= \frac{\|x - y\|}{1 + \|x - y\|}, \\ K_l(x, y) &= \log(1 + \|x - y\|^2). \end{aligned}$$

Узорачки аналогон величине $\text{dCov}_K^2(X, Y)$ је:

$$\begin{aligned} \widehat{\text{dCov}}_K^2(X, Y) &= \frac{1}{n^2} \sum_{i,j} K_1(X_i, X_j) K_2(Y_i, Y_j) - \frac{2}{n^3} \sum_{i,j,k} K_1(X_i, X_j) K_2(Y_i, Y_k) \\ &\quad + \frac{1}{n^4} \sum_{i,j,k,l} K_1(X_i, X_j) K_2(Y_k, Y_l), \end{aligned}$$

где је $\{(X_i, Y_i)\}_{i=1,\dots,n}$ прост случајан узорак обима $n \in \mathbb{N}$.

Одатле следи да је узорачка $\text{dCor}_K^2(X, Y)$ облика:

$$\widehat{\text{dCor}}^2(X, Y) = \frac{\widehat{\text{dCov}}_K^2(X, Y)}{\widehat{\text{dCov}}_K(X, X) \widehat{\text{dCov}}_K(Y, Y)}.$$

У радовима [4] и [6] се за одређивање критичне области приликом тестирања независности користи пермутациони бутстреп приступ, који је оправдан асимптотским резултатима. Асимптотска својства статистика су теоријски значајна, али како је примарни циљ овог рада разумевање и примена саме идеје тестирања независности заснованог на dCov мери, асимптотска теорија неће бити детаљно разматрана. Бутстреп процедура, која ће бити детаљно обрађена у делу 2.4, се користи за апроксимацију расподеле тест статистике. Више о асимптотским својствима $\widehat{\text{dCov}}_K^2$ и $\hat{\rho}^2$ се може прочитати у [6], док су у [4] изложена асимптотска својства $\widehat{\text{dCov}}_K$ и $\hat{\rho}_K^2$.

Нека су латентне променљиве X и Y оптерећене грешком мерења и нека су задовољени услови класичног адитивног модела грешке мерења. Уколико је доступан узорак (2.8) тест статистика $T = \widehat{\text{dCor}}^2(U, V)$ се може проширити на следећи начин:

$$T_K = \widehat{\text{dCor}}_K^2(U, V) = \frac{\widehat{\text{dCov}}_K^2(\bar{U}, \bar{V})}{\widehat{\text{dCov}}_K(\bar{U}, \bar{U}) \widehat{\text{dCov}}_K(\bar{V}, \bar{V})},$$

за $K = (K_1, K_2)$, где су K_1 и K_2 строго негативно дефинитна језгра.

Ради поређења, предлаже се и следеће проширење статистике $\hat{\rho}^2(X, Y)$ језгрима:

$$\hat{\rho}_K^2(X, Y) = \frac{\widehat{\text{dCov}}_K^2(U, V) + \widehat{\text{dCov}}_K^2(U, \tilde{V}) + \widehat{\text{dCov}}_K^2(\tilde{U}, V) + \widehat{\text{dCov}}_K^2(\tilde{U}, \tilde{V})}{4\widehat{\text{dCov}}_K(U, \tilde{U}) \widehat{\text{dCov}}_K(V, \tilde{V})},$$

где је $m_u = m_v = 2$, а са $(\tilde{U}, \tilde{V})$ је означено друго мерење пара (X, Y) .

Из теореме 1 и резултата рада [4] следи:

Теорема 2 *Нека су сва математичка очекивања која се јављају у дефиницији $\text{dCov}_K^2(U, V)$ коначна и нека важе претпоставке класичног адитивног модела грешке мерења, где је:*

$$U = X + \varepsilon, V = Y + \delta.$$

Тада су случајне променљиве X и Y независне ако и само ако је

$$\text{dCov}_K^2(U, V) = 0,$$

где је $K = (K_1, K_2)$, за јако нелинеарно дефинирана језгра K_1 и K_2 .

2.4 Емпиријска студија

Циљ симулација које следе је поређење предложених статистика

$$T_K \text{ и } \hat{\rho}_K^2,$$

за $m_u = m_v = 2$, као и њихово поређење са вредностима емпиријске $dcor$ за латентне променљиве ($\text{DC}_K = \widehat{\text{dCor}}_K^2(X, Y)$) и посматране променљиве код којих постоје грешке ($\hat{\rho}_K^2 = \widehat{\text{dCor}}_K^2(U, V)$), али поновљено мерење није доступно, за различита језгра E , K_k и K_l . Такође, биће размотрено и како однос шума и сигнала утиче на тестирање.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Основни критеријум за поређење ће бити моћ теста, односно вероватноћа да тест одбаци нулту хипотезу када латентне променљиве нису независне, односно када је нулта хипотеза нетачна.

Емпиријска моћ теста се добија помоћу *warp-speed* бутстреп алгоритма, као удео симулационих понављања у којима је H_0 одбачена. Већа вредност моћи указује на већу способност теста да детектује постојећу зависност између променљивих.

Бутстреп метода

Алгоритам 1 Пермутациони бутстреп алгоритам

- 1: На основу поновљених узорака $U^j = (U_1^j, U_2^j, \dots, U_n^j)$, $j = 1, 2, \dots, m_u$, $V^j = (V_1^j, V_2^j, \dots, V_n^j)$, $j = 1, 2, \dots, m_v$, се рачуна вредност тест статистике $\mathcal{T}$;
 - 2: Конструирају се бутстреп узорци $U_b^j = (U_1^j, U_2^j, \dots, U_n^j)$, $j = 1, 2, \dots, m_u$, $V_b^j = (V_{\pi(1)}^j, V_{\pi(2)}^j, \dots, V_{\pi(n)}^j)$, $j = 1, 2, \dots, m_v$, за неку случајну пермутацију $\pi : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$;
 - 3: На основу бутстреп узорка се рачуна вредност тест статистике $\mathcal{T}_b$;
 - 4: Кораци 2 и 3 се понављају B пута и добија се низ $\{\mathcal{T}_{b,1}, \mathcal{T}_{b,2}, \dots, \mathcal{T}_{b,B}\}$;
 - 5: На основу низа $\{\mathcal{T}_{b,1}, \mathcal{T}_{b,2}, \dots, \mathcal{T}_{b,B}\}$ се конструира критична област $W_{n,\alpha}$, за ниво значајности теста α ;
Ако $\mathcal{T} \in W_{n,\alpha}$ H_0 се одбацује.
-

Уколико се користи овакав алгоритам симулационе студије могу бити временски веома захтевне. Уколико Монте Карло симулација има N итерација тест статистика се рачуна приближно NB пута.

Због тога се као алтернатива предлаже *warp-speed* бутстреп алгоритам.

Алгоритам 2 *Warp-speed* бутстреп алгоритам

- 1: На основу поновљених узорака $U^j = (U_1^j, U_2^j, \dots, U_n^j)$, $j = 1, 2, \dots, m_u$, $V^j = (V_1^j, V_2^j, \dots, V_n^j)$, $j = 1, 2, \dots, m_v$, се рачуна вредност тест статистике $\mathcal{T}$;
 - 2: Конструирају се бутстреп узорци $U_b^j = (U_1^j, U_2^j, \dots, U_n^j)$, $j = 1, 2, \dots, m_u$, $V_b^j = (V_{\pi(1)}^j, V_{\pi(2)}^j, \dots, V_{\pi(n)}^j)$, $j = 1, 2, \dots, m_v$, за неку случајну пермутацију $\pi : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$;
 - 3: На основу бутстреп узорка се рачуна вредност тест статистике $\mathcal{T}_b$;
 - 4: Кораци 1, 2 и 3 се понављају N пута и добијају се низови $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_N\}$ и $\{\mathcal{T}_{b,1}, \mathcal{T}_{b,2}, \dots, \mathcal{T}_{b,N}\}$;
 - 5: Нулта хипотеза H_0 се одбацује за i -ту итерацију, где $i \in \{1, 2, \dots, N\}$, ако $\mathcal{T}_i > c_\alpha$, где је c_α $(1 - \alpha)\%$ квантил емпиријске расподеле бутстреп тест статистика $\{\mathcal{T}_{b,1}, \mathcal{T}_{b,2}, \dots, \mathcal{T}_{b,N}\}$.
-

Warp-speed бутстреп алгоритам је обрађен у раду [8]. Код њега се у свакој Монте Карло итерацији генерише само један узорак, док се код класичног пермутационог бутстреп алгоритма у свакој итерацији генерише B узорака. Ако је N број итерација статистика се код *warp-speed* алгоритма рачуна око $2N$ пута, што може бити значајно побољшање у односу на BN . Због тога се у овом раду користи *warp-speed* бутстреп алгоритам који задржава основну идеју пермутационог бутстреп алгоритма, али уз мањи број израчунавања. Овај приступ је погодан ако је примарни циљ поређење понашања различитих тест статистика.

Дизајн студије

Нека су X и Y случајне величине и нека се тестира њихова независност, при чему су мерења оптерећена грешкама. Даље се претпоставља да су доступна поновљена мерења, односно да су $m_u, m_v \geq 2$ и да грешке задовољавају претпоставке класичног адитивног модела грешке мерења:

$$\begin{aligned}U &= X + \varepsilon, \\V &= Y + \delta.\end{aligned}$$

У овом раду се разматрају тест статистике засноване на $dCov$ мери. Нека је са $\mathcal{T}$ означена нека од претходно дефинисаних тест статистика. Асимптотска расподела $\mathcal{T}$ под нултом хипотезом зависи од непознатих маргиналних расподела и због тога она није слободна од расподеле. Пермутациони бутстреп алгоритам омогућава апроксимацију расподеле тест статистике под нултом хипотезом, без експлицитног одређивања њене граничне расподеле.

Под нултом хипотезом заједничка расподела узорка остаје непромењена након случајног пермутовања једне од компоненти. Зато се бутстреп узорци могу генерисати пермутовањем компоненте V , док компонента U остаје непромењена. Тако се веза између променљивих уклања, док се маргиналне расподеле чувају.

Нека је N број понављања Монте Карло симулације. Тада је са

$$\hat{\beta} = 100 \frac{1}{N} \sum_{i=1}^N I\{\mathcal{T}_i > c_\alpha\}$$

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

дата процена моћи теста, у процентима, где је $I\{\mathcal{T}_i > c_\alpha\}$ индикатор догађаја да је нулта хипотеза одбачена у i -тој итерацији симулације.

Приликом процене моћи теста узорак се у свакој итерацији симулације генерише тако да важи алтернативна хипотеза, односно

$$\begin{aligned} U_i^j &= X_i + \varepsilon_i^j, \quad i \in \{1, 2, \dots, n\}, \quad j \in 1, 2, \dots, m_u, \\ V_i^j &= Y_i + \delta_i^j, \quad i \in \{1, 2, \dots, n\}, \quad j \in 1, 2, \dots, m_v, \end{aligned}$$

где између X_i и Y_i постоји зависност за свако $i \in \{1, 2, \dots, n\}$.

Уколико се узорак узме тако да су X_i и Y_i независне за свако $i \in \{1, 2, \dots, n\}$, може се добити емпиријска грешка прве врсте:

$$\hat{\alpha} = 100 \frac{1}{N} \sum_{i=1}^N I\{\mathcal{T}_i > c_\alpha\},$$

где је ниво значајности теста дефинисан као вероватноћа одбацивања тачне нулте хипотезе.

У наставку се разматрају различити модели за класичне и усмерене податке, са циљем да се испита ефикасност тестова независности дизајнираних за податке оптерећене грешкама мерења. Разматрани модели обухватају различите расподеле и димензије података. Биће илустровано како степен зависности латентних променљивих, однос шума и сигнала и избор језгра утичу на моћи различитих тестова.

За класичне податке разматрају се следећи модели:

- Модел K1: $X, Y \sim \mathcal{N}_d(0, \Sigma)$, $\varepsilon, \delta \sim \mathcal{N}(0, \sigma^2 I_d)$, где:

$$\Sigma = (0.5^{|i-j|})_{i,j \in \{1, \dots, d\}}, \quad \sigma^2 > 0,$$

а са I_d је означена идентичка матрица реда d ;

- Модел K2: $(X, Y)^\top \sim \mathcal{N}_2(0, \Sigma)$, $\varepsilon, \delta \sim \mathcal{N}(0, \sigma^2)$, где је

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad \sigma^2 > 0$$

а са $r \in (-1, 1)$ је означен Пирсонов коефицијент корелације између X и Y ;

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

- Модел К3: $X \sim \mathcal{N}_d(0, \Sigma)$, $Y \sim \text{Mix}(X, \tilde{X}, p)$, $\varepsilon, \delta \sim \mathcal{N}(0, \sigma^2 I_d)$, где $\tilde{X} \sim \mathcal{N}_d(0, \Sigma_1)$, $p \in [0, 1]$,

$$\Sigma = (0.5^{|i-j|})_{i,j \in \{1, \dots, d\}},$$

$$\Sigma_1 = (0.8^{|i-j|})_{i,j \in \{1, \dots, d\}},$$

а са I_d је означена идентичка матрица реда d ;

- Модел К4: $X \sim t_\nu$, $Y \sim \text{Mix}(X, \tilde{X}, p)$, $\varepsilon, \delta \sim \mathcal{N}(0, \sigma^2)$, где $\tilde{X} \sim t_{\nu_1}$.

У моделу К1 су латентне променљиве независне, односно важи нулта хипотеза. У моделу К2 се степен линеарне зависности између X и Y контролише Пирсоновим коефицијентом корелације. За различите вредности r је могуће испитати понашање тестова при различитим нивоима зависности. У моделима К3 и К4 се јављају мешавине случајних величина, дефинисане као у уводном поглављу. Између латентних променљивих се јавља сложенији облик зависности. У моделу К3 се разматрају различите димензије променљивих X и Y .

Поред класичних, разматрају се и модели за усмерене податке:

- Модел С1: $X, Y \sim VM(0, 1)$, $\varepsilon, \delta \sim VM(0, \kappa)$;
- Модел С2: $X \sim VM(0, 1)$, $Y \sim \text{Mix}(X, \tilde{X}, p)$, $\varepsilon, \delta \sim VM(0, \kappa)$, где $\tilde{X} \sim VM(0, 2)$ и $p \in [0, 1]$;
- Модел С3: $X \sim WN(0, 0.9)$, $Y \sim \text{Mix}(X, \tilde{X}, p)$, $\varepsilon, \delta \sim WL(0, 1.04^{-1})$, где $\tilde{X} \sim WN(\mu, 0.9)$ и $p \in [0, 1]$.

У моделу С1 латентне променљиве су независне, док се у моделима С2 и С3 јавља сложени облик зависности.

У свим наведеним моделима пажња је посвећена утицају грешке мерења. Ниво грешке се контролише преко односа шума и сигнала. Што је NSR мањи и утицај грешке мерења на сигнал је мањи. Веће вредности NSR указују на већи утицај грешке мерења. Конкретно, ако је $NSR < 1$ снага сигнала је већа од снаге шума, за $NSR = 1$ су снаге једнаке, а за $NSR > 1$ је снага шума већа. Када је снага шума већа од снаге сигнала, очекује се теже откривање зависности. То значи да при повећању NSR услови тестирања постају неповољнији.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

Перформансе тестова се процењују на основу емпиријског нивоа значајности за моделе K1 и C1 и на основу емпиријске моћи за преостале моделе.

На основу добијених резултата испитује се у којој мери предложени приступ ублажава негативан утицај грешке мерења и како на то утичу однос шума и сигнала, јачина зависности и структура података.

Резултати се изводе помоћу *warp-speed* бутстреп алгоритма, где је број итерација једнак $N = 5000$ и за обим узорка $n = 100$. Ниво значајности теста је $\alpha = 5\%$.

Резултати

Резултати емпиријске студије показују да однос шума и сигнала знатно утиче на моћ тестова. У свим моделима и код сваког теста се уочава тренд смањења моћи са повећањем интензитета грешке мерења, поготово ако је зависност између латентних променљивих слабијег интензитета.

Уколико је зависност између латентних променљивих линеарна, тестови засновани на различитим језгрима се слично понашају када је зависност јака, док се при слабијој зависности тест заснован на језгру K_k понаша лошије.

Слично, код сложенијих облика зависности, у моделима K3 и K4, разлике у избору језгара су мање изражене при већем нивоу зависности латентних променљивих.

Код модела са циркуларним подацима могу се уочити разлике између језгара када је ниво шума висок, односно када је вредност параметра κ мала. Језгро K_k показује најбоље резултате у моделу C3 када компоненте имају различите правце.

Резултати указују на то да ниво шума, односно однос шума и сигнала, представља један од најзначајнијих фактора који утичу на перформансе тестова. Очигледно је и да се зависност успешније детектује са њеним повећањем, односно да моћ тестова расте са јачином зависности између променљивих. Очекивано, најбоље перформансе постижу се у случају када променљиве нису контаминирани грешкама мерења, док присуство грешке доводи до смањења моћи тестова, посебно при већем односу шума и сигнала.

Тестови који узимају у обзир претпоставке класичног адитивног модела грешке мерења показују добре перформансе и у присуству грешке, а при јачој зависности њихова моћ се приближава моћи теста који се примењује на неконтаминирани податке. При томе су разлике између разматраних тест ста-

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

тестистика релативно мале. Имајући у виду да тест статистика T у појединим случајевима постиже чак и нешто боље резултате од $\hat{\rho}^2$, уз истовремено краће време извршавања, њено коришћење као алтернативе представља оправдан избор. Коначно, треба имати у виду да је у овој симулационој студији разматран случај са само два поновљена мерења. Будући да се прецизнија процена латентних променљивих може очекивати са повећањем броја поновљених мерења, очекивано је да би тест статистика T у случају већег броја мерења показала још боље перформансе.

$d = 5$

σ	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\rho_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	5	5	5	5	5	5	6	5	5	5	6	5
1.0	5	5	6	5	5	5	5	5	5	5	5	5
1.2	5	5	6	5	5	5	6	5	5	5	5	5

$d = 10$

σ	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\rho_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	5	5	5	5	5	5	5	5	5	5	5	5
1.0	5	5	5	5	5	5	5	5	5	5	5	5
1.2	5	5	5	5	5	5	5	5	5	6	5	5

Табела 2.1: Емпиријски ниво значајности за различите параметре у моделу K1. Параметар σ узима вредности из скупа $\{0.5, 1.0, 1.2\}$, чиме се разматрају различити нивои грешке мерења, од релативно малог до израженог шума. У моделу K1 је $NSR_X = NSR_Y = [\sigma^2]_{i=1, \dots, d}$ за све различите вредности параметара σ и d . Однос шума и сигнала расте од 0.25 до 1.44

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

r	σ	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC_E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC_{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC_{K_l}
-1	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	1.0	100	100	100	100	100	100	97	100	100	100	100	100
	1.2	100	100	99	100	90	100	81	100	100	100	96	100
-0.7	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	1.0	99	99	90	100	70	93	63	100	98	99	87	100
	1.2	91	98	76	100	40	80	42	100	84	96	67	100
-0.5	0.5	99	99	98	100	92	92	84	99	99	99	97	100
	1.0	80	87	66	100	32	57	34	98	75	83	59	100
	1.2	52	76	46	100	18	46	23	98	44	70	41	100
-0.3	0.5	72	70	62	80	37	46	25	48	68	64	60	76
	1.0	34	44	27	82	10	22	15	57	29	39	24	80
	1.2	20	35	18	82	7	19	11	57	17	31	16	80
-0.2	0.5	38	40	33	44	20	25	9	15	36	38	31	41
	1.0	15	20	16	43	8	10	8	24	12	17	13	40
	1.2	10	14	12	43	6	8	7	24	9	11	10	40
-0.1	0.5	13	13	10	14	8	8	1	3	12	11	11	13
	1.0	9	8	7	18	5	6	6	11	8	7	7	17
	1.2	7	8	7	18	6	6	6	11	6	8	7	17
0.1	0.5	13	14	11	13	8	9	2	3	12	13	11	13
	1.0	10	11	10	17	5	7	7	12	9	11	10	17
	1.2	8	9	9	17	5	6	6	12	7	9	8	17
0.2	0.5	35	35	29	44	18	19	8	16	31	33	26	41
	1.0	15	22	16	46	6	12	9	28	14	19	13	44
	1.2	10	18	12	46	6	11	7	28	8	15	11	44
0.3	0.5	70	71	66	84	42	43	27	51	68	67	60	82
	1.0	31	49	26	86	11	23	16	57	28	44	23	81
	1.2	20	35	17	86	8	19	11	57	16	32	17	81
0.5	0.5	100	100	98	100	94	94	85	99	99	99	97	100
	1.0	79	90	59	100	30	63	34	98	71	85	53	100
	1.2	55	78	44	100	19	48	20	98	43	72	37	100
0.7	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	1.0	99	100	90	100	69	94	68	100	97	99	86	100
	1.2	92	98	77	100	44	84	46	100	84	96	71	100
1	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	1.0	100	100	100	100	100	100	97	100	100	100	100	100
	1.2	100	100	98	100	87	100	81	100	100	100	97	100

Табела 2.2: Емпиријска моћ теста за различите параметре у моделу K2. Параметар σ узима вредности из скупа $\{0.5, 1.0, 1.2\}$. Пирсонов коефицијент корелације, r , узима вредности од -1 до 1 , чиме се разматрају различити нивои зависности између латентних променљивих. У моделу K2 је $NSR_X = NSR_Y = \sigma^2$.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

$d = 5$													
σ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	0.5	100	100	100	100	93	97	92	100	100	100	100	100
	0.3	78	83	77	88	38	52	40	66	74	83	74	90
	0.1	13	14	13	15	7	8	8	11	12	13	12	15
1	0.5	90	99	92	100	50	72	40	100	77	98	85	100
	0.3	37	66	45	88	14	26	15	66	25	58	37	90
	0.1	8	11	9	15	6	6	6	11	6	9	8	15
1.2	0.5	73	97	80	100	36	55	25	100	55	94	67	100
	0.3	24	56	34	88	12	19	11	66	16	47	27	90
	0.1	7	10	8	15	5	6	6	11	6	8	8	15

$d = 10$													
σ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	0.5	100	100	100	100	90	91	85	95	99	100	99	100
	0.3	73	81	75	87	35	39	32	46	59	73	63	82
	0.1	12	13	11	15	7	8	7	8	9	11	10	14
1	0.5	87	99	94	100	63	73	46	95	73	96	83	100
	0.3	35	63	44	87	18	24	14	46	23	49	31	82
	0.1	7	9	9	15	6	7	6	8	6	8	8	14
1.2	0.5	71	97	83	100	49	61	30	95	57	91	66	100
	0.3	24	53	32	87	15	19	11	46	17	39	21	82
	0.1	6	8	7	15	6	6	6	8	6	7	7	14

Табела 2.3: Емпиријска моћ теста за различите параметре у моделу К3. Параметар σ узима вредности из скупа $\{0.5, 1.0, 1.2\}$. Параметар p узима вредности 0.5, 0.3 и 0.1. Он одређује ниво зависности између X и Y . У моделу К3 је $NSR_X = NSR_Y = [\sigma^2]_{i=1, \dots, d}$ за све вредности параметара σ и d .

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

$\nu_1 = 7$													
σ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	0.5	100	100	99	100	99	100	97	100	100	100	99	100
	0.3	82	83	75	92	68	75	61	95	83	84	74	93
	0.1	21	20	17	23	12	13	11	19	19	19	16	22
1.0	0.5	93	96	82	100	50	83	55	100	90	95	76	100
	0.3	54	64	41	92	17	39	22	95	46	59	36	93
	0.1	12	14	10	23	7	9	6	19	10	13	8	22
1.2	0.5	81	91	68	100	30	67	37	100	71	87	58	100
	0.3	38	54	30	92	11	27	14	95	30	47	25	93
	0.1	9	13	8	23	6	8	6	19	8	11	7	22

$\nu_1 = 15$													
σ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0.5	0.5	100	100	99	100	99	100	97	100	100	100	99	100
	0.3	82	83	74	92	69	76	59	95	82	83	72	92
	0.1	19	19	16	23	12	14	11	20	18	17	14	22
1.0	0.5	95	97	83	100	48	83	50	100	90	95	76	100
	0.3	54	62	41	92	18	40	21	95	45	56	35	92
	0.1	11	13	11	23	7	9	8	20	10	11	9	22
1.2	0.5	82	91	68	100	30	68	35	100	71	87	57	100
	0.3	39	53	32	92	12	28	16	95	30	45	25	92
	0.1	9	11	9	23	6	7	7	20	7	10	8	22

σ	0.5	1	1.2
NSR_X	0.15	0.6	0.86

$\nu_1 = 7$										
σ	0.5			1.0			1.2			
p	0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1	
NSR_Y	0.16	0.17	0.18	0.65	0.68	0.7	0.94	0.97	1.01	

$\nu_1 = 15$										
σ	0.5			1.0			1.2			
p	0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1	
NSR_Y	0.18	0.19	0.21	0.71	0.77	0.83	1.02	1.1	1.2	

Табела 2.4: Емпиријска моћ теста за различите параметре у моделу К4. Параметар σ узима вредности из скупа $\{0.5, 1.0, 1.2\}$. Параметар p узима вредности 0.5, 0.3 и 0.1. Он одређује ниво зависности између X и Y , ν_1 узима вредности 7 и 15, да би се испитало како тежина репа утиче на перформансе теста. Однос шума и сигнала за различите вредности параметара је приказан у табели.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

κ	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\rho_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
1	5	5	6	5	5	5	6	5	5	5	6	5
2	5	6	6	5	5	6	7	5	5	5	6	5
5	5	5	5	5	4	5	5	4	5	5	5	5
7	5	5	5	6	6	6	5	5	5	5	5	6

κ	1	2	5	7
$NSR_{\mathbf{X}}$	1	0.55	0.19	0.13
$NSR_{\mathbf{Y}}$	1	0.55	0.19	0.13

Табела 2.5: Емпиријски ниво значајности за различите параметре у моделу С1. Параметар κ узима вредности из скупа $\{1, 2, 5, 7\}$, чиме се разматрају различити нивои концентрације грешке мерења. Однос шума и сигнала је дат у табели.

κ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\rho_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
1	0.5	11	30	13	100	11	25	11	100	11	31	13	100
	0.3	8	13	8	96	7	11	7	96	7	13	7	96
	0.1	6	6	5	23	6	6	5	21	6	6	5	24
2	0.5	81	94	67	100	56	89	56	100	86	94	68	100
	0.3	30	54	27	96	18	44	22	96	35	55	28	96
	0.1	7	10	7	24	6	9	7	22	7	10	7	25
5	0.5	100	100	100	100	100	100	99	100	100	100	100	100
	0.3	86	88	76	96	73	81	64	96	87	88	77	96
	0.1	17	18	14	23	11	14	11	20	18	19	14	24
7	0.5	100	100	100	100	100	100	100	100	100	100	100	100
	0.3	91	92	84	97	83	86	74	97	92	92	84	97
	0.1	18	18	16	24	14	15	12	20	18	18	16	24

κ	1	2	5	7
$NSR_{\mathbf{X}}$	1	0.55	0.19	0.13

κ	1			2			5			7		
p	0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1
$NSR_{\mathbf{Y}}$	1.29	1.47	1.69	0.71	0.8	0.92	0.25	0.28	0.33	0.17	0.2	0.23

Табела 2.6: Емпиријска моћ теста за различите параметре у моделу С2. Параметар κ узима вредности из скупа $\{1, 2, 5, 7\}$, чиме се разматрају различити нивои концентрације грешке мерења. Параметар p узима вредности 0.5, 0.3 и 0.1. Однос шума и сигнала за све вредности параметара је приказан у табели.

Глава 2. ТЕСТОВИ НЕЗАВИСНОСТИ ЗАСНОВАНИ НА DCOV МЕРИ У СЛУЧАЈУ КОМПЛЕТНИХ ПОДАТАКА

μ	p	$\hat{\rho}_E^2$	T_E	$\tilde{\rho}_E$	DC _E	$\hat{\rho}_{K_k}^2$	T_{K_k}	$\tilde{\rho}_{K_k}$	DC _{K_k}	$\hat{\rho}_{K_l}^2$	T_{K_l}	$\tilde{\rho}_{K_l}$	DC _{K_l}
0	0.5	99	99	96	100	98	99	94	100	99	99	96	100
	0.3	71	74	60	90	62	69	54	94	72	73	61	87
	0.1	13	13	12	20	10	12	11	19	14	14	13	19
$\frac{\pi}{2}$	0.5	97	98	90	100	98	99	93	100	96	96	88	100
	0.3	47	55	42	82	55	65	45	96	47	50	41	71
	0.1	8	8	8	12	7	8	8	15	8	8	8	12
π	0.5	96	97	89	100	98	99	93	100	94	95	87	100
	0.3	44	51	39	76	52	61	44	94	43	46	37	66
	0.1	8	8	8	10	7	8	8	13	8	8	8	10

μ	0			$\frac{\pi}{2}$			π		
p	0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1
NSR_Y	0.38	0.38	0.38	0.54	0.5	0.42	0.04	0.06	0.14

Табела 2.7: Емпиријска моћ теста за различите параметре у моделу С3. Параметар μ , који одређује правац друге компоненте мешавине у односу на правац прве компоненте, чиме се омогућава разматрање различитих конфигурација усмерених података, узима вредности из скупа $\{0, \frac{\pi}{2}, \pi\}$. Параметар p узима вредности 0.5, 0.3 и 0.1. Однос шума и сигнала NSR_Y за све вредности параметара је приказан у табели, а $NSR_x = 0.38$.

Глава 3

Примена на реалним подацима

За илустрацију претходно разматраних тестова независности биће коришћени подаци из истраживања *NHANES* (скраћено од енг. *National Health and Nutrition Examination Survey*), [2]. У питању је велико истраживање здравственог и нутритивног стања становништва Сједињених Америчких Држава. Подаци су добијени путем упитника, прегледа и лабораторијских анализа, а организовани су у више засебних датотека. За повезивање података из различитих датотека се користи идентификатор *SEQN* (скраћено од енг. *Respondent Sequence Number*) који је додељен сваком од испитаника.

У примеру се користи циклус из 2017-2018 године. Он садржи поновљена мерења неких карактеристика, због чега је погодан за илустрацију методологије приказане у претходном поглављу.

Пре примене теста, биће одабране одговарајуће променљиве, а затим ће испитаници са недостајућим вредностима бити уклоњени. Повезивање се врши према променљивој *SEQN*. На тако формираном узорку примениће се тест статистика T , дефинисана у претходном поглављу. Циљ примене је да се илуструје употреба предложеног поступка на стварним подацима и испита постојање зависности између посматраних променљивих. Резултати добијени на реалним подацима представљају допуну симулационој студији и омогућавају практичну илустрацију примене разматране методологије.

3.1 Зависност крвног притиска и индекса телесне масе

У овом одељку биће испитана зависност систолног крвног притиска и индекса телесне масе. Како је повећана телесна маса повезана са повећаним ризиком од повишеног крвног притиска и развоја кардиоваскуларних обољења, очекује се постојање зависности између ове две променљиве.

Индекс телесне масе представља меру која се користи за процену телесне масе у односу на висину. Он је дат у фајлу `BMX_J.xpt` и представљен променљивом `BMI`. Вредности су изражене у јединици $\frac{kg}{m^2}$. Индекс телесне масе се користи као индикатор ухрањености, као и један од показатеља у процени ризика од различитих кардиоваскуларних и метаболичких обољења.

Систолни крвни притисак представља највишу вредност крвног притиска током једног срчаног циклуса, односно притисак у артеријама у тренутку контракције срчаног мишића. Изражава се у милиметрима живиног стуба $mmHg$. У *NHANES* бази, систолни крвни притисак је мерен у више узастопних мерења, при чему су за прво, друго и треће мерење доступне променљиве `BPXSY1`, `BPXSY2` и `BPXSY3`, у фајлу `BPX_J.xpt`.

У складу са протоколом прикупљања података, мерење крвног притиска врши се код испитаника старости осам и више година, након што испитаник проведе најмање пет минута у мирном седећем положају. Врше се три узастопна мерења крвног притиска, док се у случају прекинутог или непотпуног мерења може извршити и четврто мерење.

Овакав протокол омогућава да се за истог испитаника добије више мерења исте физиолошке карактеристике.

У анализи се посматрају случајне променљиве:

$$X = BMI \text{ и } Y = BP_S.$$

Очекивано је да при мерењу обе карактеристике настају грешке, па су доступне реализације следећих променљивих:

$$U = BMI + \epsilon,$$

$$V_1 = BP_S + \delta_1,$$

$$V_2 = BP_S + \delta_2,$$

$$V_3 = BP_S + \delta_3,$$

где ε , δ_1 , δ_2 и δ_3 представљају грешке мерења. Може се претпоставити да грешке задовољавају претпоставке класичног адитивног модела грешке мерења. Тада се хипотеза

$$H_0 : X \text{ и } Y \text{ су независне случајне променљиве}$$

може тестирати против алтернативе

$$H_1 : X \text{ и } Y \text{ су зависне случајне променљиве,}$$

помоћу тест статистике T дефинисане у претходном поглављу. Како је $m_u = 1$ и $m_v = 3$ тест статистика се примењује на

$$\{(U_i, \bar{V}_i)\}_{i=1,2,\dots,n},$$

где је $\bar{V}_i = \frac{1}{3}(V_i^1 + V_i^2 + V_i^3)$, $i \in \{1, 2, \dots, n\}$, а $n \in \mathbb{N}$ је обим узорка. Након припреме података (упаривања на основу $SEQN$ и уклањања испитаника са недостајућим вредностима) узорак је обима $n = 6008$.

У анализи се користи тест статистика T , дефинисана у претходном поглављу. Одлука о одбацивању нулте хипотезе доноси се на основу критичне области теста, уз избрани ниво значајности.

Очекује се да ће анализа показати зависност између променљивих.

Тестирање независности се врши уз помоћ пермутационог бутстреп алгорита. За задати ниво значајности $\alpha = 0.05$ се генерише $B = 500$ случајних пермутација измерених вредности променљиве $\bar{V}$, док вредности променљиве U остају непромењене. За сваку од пермутација $\bar{V}_b$ и U се рачуна вредност тест статистике T_E , чиме се формира емпиријска расподела тест статистике која одговара нултој хипотези о независности латентних променљивих. Тестирање се може вршити преко критичне области или p -вредности. Добијени су следећи резултати:

- граница критичне области: 0.032,
- вредност тест статистике: 0.356,
- p -вредност: 0.002.

Како је $0.03176833 < 0.3559207$ нулта хипотеза о независности се одбацује. Исти закључак се добија и на основу p -вредности, јер је она мања од задатог нивоа значајности.

Резултати указују на статистички значајну зависност између систолног крвног притиска и индекса телесне масе.

Резултати добијени применом предложене тест статистике се могу упоредити са резултатима добијеним применом Пирсоновог коефицијента корелације, ρ , који се често користи за тестирање независности случајних променљивих. Како примена Пирсоновог коефицијента корелације има ограничења у детектовању нелинеарне зависности, а он није инваријантан на грешке мерења, добијени резултати се у овом контексту користе само као допуна анализи.

Узорак	$\{(U_i, V_i^1)\}_i$	$\{(U_i, V_i^2)\}_i$	$\{(U_i, V_i^3)\}_i$	$\{(U_i, \bar{V}_i)\}_i$
$\hat{\rho}$	0.307	0.314	0.321	0.318

Табела 3.1: Пирсонов коефицијент корелације крвног притиска и индекса телесне тежине, на основу различитих узорка.

Резултати указују на умерену линеарну зависност променљивих, што је у складу са претходно добијеним резултатима тестирања, према којима се одбацује хипотеза о независности променљивих. При томе, претходно тестирање се не ограничава на линеарну независност, па се резултати Пирсоновог коефицијента могу посматрати као потврда постојања линеарне компоненте зависности у контаминираним узорку.

3.2 Зависност крвног притиска и укупног холестерола у крви

Према подацима из 2018. године, болести срца су у том периоду представљале водећи узрок смрти у Сједињеним Америчким Државама. Нивои липида у крви представљају један од основних показатеља који се користе при процени ризика од настанка кардиоваскуларних болести. Липиди су супстанце које, између осталог, имају важну улогу у функционисању организма, а њихов поремећен ниво у крви може бити повезан са повећаним ризиком од одређених обољења срца и крвних судова. Мерења липида у крви у оквиру истраживања *NHANES* обухватају укупни холестерол, холестерол липопротеина високе густине, холестерол липопротеина ниске густине и триглицериде.

У овом примеру посматра се укупни холестерол. Холестерол је супстан-

ца слична масти која има важну улогу у организму и неопходна је, између осталог, за изградњу ћелија и производњу одређених хормона. Организам сам производи највећи део холестерола, пре свега у јетри, док мањи део потиче из хране. Према доступној документацији, око 75% холестерола настаје синтезом у организму, док преостали део потиче из исхране. Мерење његовог нивоа у крви значајно је јер повишене вредности могу бити повезане са повећаним ризиком од настанка атеросклерозе, односно постепеног накупљања масних наслага у зидовима крвних судова, што може отежати нормалан проток крви.

Ниво укупног холестерола у крви у оквиру истраживања *NHANES* одређиван је лабораторијском методом заснованом на ензимском тесту. Ензими су супстанце које у организму и лабораторијским поступцима убрзавају одређене хемијске реакције. У овом случају, низом хемијских реакција холестерол се претвара у једињење које производи обојени раствор. Интензитет те боје затим се мери посебним лабораторијским инструментом и на основу њега се одређује концентрација укупног холестерола у крви. На тај начин се хемијска особина узорка преводи у бројчану вредност која представља измерени ниво холестерола.

Током циклуса истраживања *NHANES* 2017–2018 није било промена у лабораторијској методи која је коришћена за мерење укупног холестерола, опреми на којој су анализе вршене, нити лабораторији у којој су обављене. Узорци крви су након обраде чувани на температури од приближно -30°C до слања на лабораторијску анализу. Чување на ниској температури омогућава очување својстава узорка и смањује могућност његове промене пре самог мерења.

Поред тога, током прикупљања и анализе података примењивани су поступци контроле квалитета. То подразумева проверу рада лабораторијске опреме, начина обраде узорака и тачности добијених мерења. У оквиру контроле квалитета, за део узорака (2%) вршена су и поновљена мерења. Како је у датотеци доступно само једно мерење, важно је нагласити да је оно добијено у строго контролисаним лабораторијским условима и да не представља нужно резултат само једног непосредно извршеног лабораторијског мерења.

У овом примеру се разматра зависност између нивоа укупног холестерола у крви и систолног крвног притиска. За сваког испитаника подаци о крвном притиску доступни су у датотеци `BPX_J`, док су подаци о укупном холестеролу садржани у датотеци `TCHOL_J`. Датотеке се, као у претходном примеру,

могу повезати помоћу јединственог идентификатора испитаника $SEQN$.

Укупни холестерол представља један од значајних показатеља кардиоваскуларног ризика. Холестерол је липидна супстанца која има важну физиолошку улогу, али његове повишене концентрације могу бити повезане са развојем различитих кардиоваскуларних обољења. У посматраном циклусу укупни холестерол је одређиван директним лабораторијским мерењем ензимском методом, а променљива $LBXTC$ садржи његову концентрацију у mg/dL .

Са друге стране, систолни крвни притисак представља притисак крви на зидове артерија у тренутку контракције срца и један је од основних показатеља функционисања кардиоваскуларног система. Повишен крвни притисак је значајан фактор ризика за различита кардиоваскуларна обољења, па је испитивање његове повезаности са другим факторима ризика од великог медицинског значаја. Три узастопна мерења систолног притиска забележена су променљивама $BPXSY1$, $BPXSY2$ и $BPXSY3$.

Узорак чине реализације променљивих:

$$\begin{aligned}U &= X + \varepsilon, \\V_1 &= Y + \delta_1, \\V_2 &= Y + \delta_2, \\V_3 &= Y + \delta_3,\end{aligned}$$

где је $X = TCHOL$ и $Y = BPS$, а ε , δ_1 , δ_2 и δ_3 представљају грешке мерења. Може се претпоставити да грешке задовољавају претпоставке класичног адитивног модела грешке мерења. Хипотеза

$$H_0 : X \text{ и } Y \text{ су независне случајне променљиве}$$

се може тестирати против алтернативе

$$H_1 : X \text{ и } Y \text{ су зависне случајне променљиве,}$$

помоћу тест статистике T . Како је $m_u = 1$ и $m_v = 3$ тест статистика се примењује на

$$\{(U_i, \bar{V}_i)\}_{i=1,2,\dots,n},$$

где је $\bar{V}_i = \frac{1}{3}(V_i^1 + V_i^2 + V_i^3)$, $i \in \{1, 2, \dots, n\}$, а $n \in \mathbb{N}$ је обим узорка. Након припреме података узорак је обима $n = 5578$.

Као и у првом примеру, очекује се да ће анализа показати зависност између променљивих.

Тестирање независности се поново врши уз помоћ пермутационог бутстреп алгоритма, за задати ниво значајности $\alpha = 0.05$ и $B = 500$ итерација бутстреп алгоритма. Добијени су следећи резултати:

- граница критичне области: 0.032,
- вредност тест статистике: 0.243,
- p -вредност: 0.002,

на основу којих се нулта хипотеза о независности одбацује. Закључује се да између систолног крвног притиска и нивоа укупног холестерола у крви постоји зависност.

Узорак	$\{(U_i, V_i^1)\}_i$	$\{(U_i, V_i^2)\}_i$	$\{(U_i, V_i^3)\}_i$	$\{(U_i, \bar{V}_i)\}_i$
$\hat{\rho}$	0.233	0.231	0.227	0.233

Табела 3.2: Пирсонов коефицијент корелације крвног притиска и укупног холестерола у крви, на основу различитих узорака.

Као и у претходном примеру, резултати указују на умерену линеарну зависност променљивих, што је у складу са резултатима тестирања.

Глава 4

Могућност проширење тестова независности на цензурисане податке

У овом поглављу ће бити представљена потенцијална проширења претходно обрађене методологије, у случају цензурисаних података. Правци даљег истраживања могу бити засновани на тесту датом у раду [5], који користи $dCov$ меру као основу за тестирање независности променљивих код којих се јавља цензурисање, и на деконволуционом приступу из рада [3], у коме се уводе модели који описују податке који су истовремено погођени цензуром и грешкама у мерењу.

У клиничким истраживањима често се ефикасност неке интервенције процењује праћењем броја испитаника који су преживели одређени временски период, односно код којих је интервенција довела до повољног исхода. Осим смрти неки од неповољних исхода могу бити и појава инфекције или повратак болести након лечења. Може се пратити и време потребно за остварење неког специфичног исхода, као што је време до зачећа. Време протекло од дефинисаног почетног тренутка до настанка догађаја од интереса се назива **време преживљавања**.

Експерименти у којима се посматра време преживљавања често су отежани ситуацијама у којима испитаници не желе или не могу да учествују у њима до краја периода праћења. Такође, може се десити и да испитаник не доживи смрт до краја периода праћења. Те ситуације се називају **десно цензуриса-**

Глава 4. МОГУЋНОСТ ПРОШИРЕЊЕ ТЕСТОВА НЕЗАВИСНОСТИ НА ЦЕНЗУРИСАНЕ ПОДАТКЕ

ним посматрањима. За овакве испитанике је доступна непотпуна информација. Познато је да се праћени догађај није догодио до последњег тренутка праћења испитаника, и да ће се, уколико до њега дође, догодити након тог тренутка, али није познато када. Потпуно искључивање таквих испитаника из истраживања није оправдано, јер носе информације о преживљавању.

У претходном поглављу је разматрано тестирање независности случајних променљивих у присуству грешака мерења, које задовољавају класичан адитивни модел грешке мерења. Претпостављено је да су посматране вредности доступне за цео узорак. Ипак у многим истраживањима се уз проблем грешака у мерењу јавља и проблем цензурисаних података. У наставку су наведени примери података код којих могу настати и цензура и грешке истовремено.

- **Код времена до смрти пацијента** датум дијагнозе или почетка лечења не мора бити потпуно прецизан, па настаје грешка при мерењу времена од дијагнозе до смрти. Уколико је пацијент жив на крају студије или је напустио студију пре краја настаје цензурисање са десне стране.
- **Код променљивих попут времена до појаве компликације након операције и времена до поновног јављања болести** грешка настаје јер се догађај не региструје у тренутку дешавања већ на следећем контролном прегледу. Десна цензура настаје уколико се не јави компликација, односно поновно јављање болести до краја истраживања, или ако пацијент напусти истраживање пре краја.
- У економији грешке могу настати приликом мерења времена до банкрота, уколико је датум стварног престанка активности непрецизан. Десно цензурисање настаје уколико фирма још увек послује на крају посматраног периода или напусти истраживање.
- Пример из физике је време до распада нестабилне честице.

Због оваквих података се јавља потреба за тестом који истовремено у обзир узима нетачност мерења и непотпуност посматраних података.

Како је ова област још увек недовољно истражена, остаје низ отворених питања која захтевају даља теоријска и емпиријска испитивања. У овом поглављу ће бити представљене методе које су до сада разматране у литератури,

као и могућности њиховог даљег развоја. Пажња ће бити посвећена могућим правцима будућих истраживања и проширењима постојећих методолошких приступа.

Време преживљавања

Анализа преживљавања је грана статистике која се бави мерењем времена до настанка догађаја од интереса. Уколико је проценат цензурисаних елемената мали, они се могу занемарити, али се у супротном тестови морају прилагодити новој структури података.

Нека је са X означено време преживљавања. У питању је позитивна непрекидна случајна променљива. Нека је $F_X(\cdot)$ функција расподеле случајне величине X . Функција дефинисана са

$$S_X(t) = P\{X > t\} = \int_t^{+\infty} dF_X(u),$$

се назива функцијом преживљавања, а функција дефинисана са

$$\lambda_X(t) = \lim_{dt \rightarrow 0} \frac{P\{t \leq X \leq t + dt \mid X \geq t\}}{dt},$$

се назива стопом хазарда.

Због непотпуних информација се функција $F_X(\cdot)$ не може апроксимирати на класичан начин, због чега су развијене посебне методе анализе преживљавања. У наставку ће бити приказана Каплан-Мајерова оцена функције преживљавања.

Нека је $\{(T_j, \Delta_j)\}_{j=1, \dots, n}$ прост случајан узорак обима $n \in \mathbb{N}$, где је

$$T_j = \min\{X_j, C_j\} \quad \text{и} \quad \Delta_j = I\{X_j \leq C_j\},$$

за $j = 1, 2, \dots, r$. Нека су $t_1, t_2, \dots, t_n$ нецензурисане реализације случајне величине T . Нека је са r означен број различитих доступних реализација и нека је $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ одговарајући варијациони низ. Са $d_{(j)}$ се означава колико је пута забележена вредност $t_{(j)}$, за $j = 1, 2, \dots, r$. Важи $d_{(1)} + d_{(2)} + \dots + d_{(r)} = n$.

Са $R(t_{(j)})$ је означен број посматрања која се још нису реализовала или нису цензурисана, непосредно пре тренутка $t_{(j)}$, за $j = 1, 2, \dots, r$. Нека је:

$$\begin{aligned} t_{(0)} &= t_0, & R(t_{(0)}) &= n, \\ t_{(r+1)} &= +\infty, & R(t_{(r+1)}) &= 0. \end{aligned}$$

Глава 4. МОГУЋНОСТ ПРОШИРЕЊЕ ТЕСТОВА НЕЗАВИСНОСТИ НА ЦЕНЗУРИСАНЕ ПОДАТКЕ

Нека је $t \in [t_{(j)}, t_{(j+1)})$. Тада важи:

$$S_X(t) = P\{X > t_{(1)} | X > t_{(0)}\} P\{X > t_{(2)} | X > t_{(1)}\} \\ \dots P\{X > t_{(j)} | X > t_{(j-1)}\}.$$

Каплан-Мајерова оцена функције преживљавања $S_X(t)$ је дата са:

$$\hat{S}_X^{KM}(t) = \prod_{j | t_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{R(t_{(j)})}\right) \\ = \prod_{j | t_{(j)} \leq t} \left(1 - d_{(j)} \left(\sum_{k=1}^n I\{t_k \geq t_{(j)}\}\right)^{-1}\right).$$

Тестови независности засновани на $dCov$ мери у случају цензурисаних података

У раду [5] је описано како се $dCov$ мера може користити за тестирање независности уколико се јавља десно цензурисање података. Аутори уводе следећи модел:

- $X \in [0, +\infty)$ је време преживљавања;
- $C \in [0, +\infty)$ је цензуришућа променљива;
- $Y \in \mathbb{R}^q$ је вектор мерених карактеристика;
- C је независна од пара (X, Y) ;
- X и Y могу бити зависне.

Нека је $\{(Y_j, T_j, \Delta_j)\}_{j=1, \dots, n}$, где је $T_j = \min\{X_j, C_j\}$ и $\Delta_j = I\{X_j \leq C_j\}$, за $j = 1, 2, \dots, n$, прост случајан узорак обима $n \in \mathbb{N}$. Циљ је тестирање хипотезе

$$H_0 : X \text{ и } Y \text{ су независне случајне величине,}$$

против алтернативе

$$H_1 : X \text{ и } Y \text{ нису независне случајне величине,}$$

на основу датог узорака.

Нека је $F_C(\cdot)$ функција расподеле цензуришуће променљиве C и нека $\hat{S}_C^{KM}(\cdot)$ Каплан-Мајерова оцена функције преживљавања $S_C(\cdot)$ дате са:

$$S_C(t) = P\{C > t\} = \int_t^{+\infty} dF_C(u).$$

Нека су $W_1, W_2, \dots, W_n$ случајне величине дефинисане са: $W_i = \frac{\Delta_i}{\hat{S}_C^{KM}(T_i-)}$, за $i \in \{1, 2, \dots, n\}$, где је са $\hat{S}_C^{KM}(T_i-)$ означен леви лимес функције $\hat{S}_C^{KM}(\cdot)$ у тачки T_i и нека је:

$$\begin{aligned} a_{ij} &= \|Y_i - Y_j\|, \\ d_{ij} &= \|T_i - T_j\|, \\ m &= \sum_{i=1}^n W_i, \\ M &= \sum_{i=1}^n W_i^2, \end{aligned}$$

за $i, j \in \{1, 2, \dots, n\}$.

У раду [5] је представљена верзија оцено величине $d\text{Cov}^2(X, Y)$ добијена скалирањем инверзном вероватноћом цензурисања $\tilde{\Omega}$. Она је задата једнакошћу:

$$\begin{aligned} n(n-1)(n-2)(n-3)\tilde{\Omega} &= \sum_{i,j=1}^n W_i W_j a_{ij} d_{ij} (m^2 - m(W_i + W_j) - M + 2W_i W_j) \\ &\quad - 2 \sum_{i=1}^n W_i (m + W_i) \left(\sum_{j=1}^n W_j a_{ij} \right) \left(\sum_{j=1}^n W_j d_{ij} \right) \\ &\quad + 2 \sum_{i=1}^n W_i \left(\sum_{j=1}^n W_j^2 a_{ij} \right) \left(\sum_{j=1}^n W_j d_{ij} \right) \\ &\quad + 2 \sum_{i=1}^n W_i \left(\sum_{j=1}^n W_j a_{ij} \right) \left(\sum_{j=1}^n W_j^2 d_{ij} \right) \\ &\quad + \sum_{i,j=1}^n W_i W_j a_{ij} \sum_{i,j=1}^n W_i W_j d_{ij}. \end{aligned}$$

Скалирање инверзном вероватноћом цензурисања се врши да би се смањила пристрасност. Тај приступ смањује утицај цензурисаних података у узорку цензурисаном са десне стране. Случајне величине W_i , $i \in \{1, 2, \dots, n\}$ представљају тежине које се добијају као инверзне оцено вероватноће да посматрање није цензурисано до времена T_i .

Глава 4. МОГУЋНОСТ ПРОШИРЕЊЕ ТЕСТОВА НЕЗАВИСНОСТИ НА ЦЕНЗУРИСАНЕ ПОДАТКЕ

У раду [5] су изведена асимптотска својства тест статистике $\tilde{\Omega}$ под посебним претпоставкама.

Приликом тестирања, нулта хипотеза се одбацује на нивоу значајности α ако је вредност тест статистике $\tilde{\Omega}$ израчуната на оригиналном узорку већа од одговарајуће критичне вредности c_α , која се може израчунати пермутационом бутстреп методом.

Питање могућности прилагођавања статистике $\tilde{\Omega}$ подацима погођеним грешкама мерења остаје отворено. Својства статистике у случају таквих података још увек нису испитана. У наставку ће бити наведен предлог могућег проширења.

Цензурисани подаци и грешке мерења

У литератури није истражен проблем тестирања независности уколико су цензурисани подаци контаминирани грешкама мерења. У наставку ће бити представљен приступ заснован на деконволуцији за анализу цензурисаних података у присуству грешака мерења. У раду [3] су изложене основне идеје деконволуционог приступа, његова примена у условима цензурисања, као и претпоставке неопходне за његову примену.

Изложена идеја се у будућим истраживањима може применити на тестирање хипотеза. Конкретно, она се може искористити за конструкцију модификоване оцене $dCov$, која би омогућила тестирање независности цензурисаних података у присуству грешака мерења.

Нека су у наставку цензурисани подаци измерени са грешком. Разликују се два случаја:

- грешка утиче и на X и на C ;
- грешка утиче само на X .

У првом моделу се посматрају променљиве χ и ζ , дате са

$$\begin{aligned}\chi &= X + \varepsilon, \\ \zeta &= C + \varepsilon,\end{aligned}$$

где је ε грешка у мерењу. Овај случај настаје када је почетак времена непознат, на пример ако се мери време до смрти, али није познато када се тачно јавио први симптом болести.

Глава 4. МОГУЋНОСТ ПРОШИРЕЊЕ ТЕСТОВА НЕЗАВИСНОСТИ НА ЦЕНЗУРИСАНЕ ПОДАТКЕ

У другом моделу се посматрају χ и C , јер није познат тачан тренутак настанка догађаја од интереса. Овде је пример настанак компликације након операције.

Уколико се подаци уклапају у први модел узорак је облика

$$\{(U_j^{(1)}, \Upsilon_j^{(1)})\}_{j=1, \dots, n},$$

а уколико се уклапају у други облика

$$\{(U_j^{(2)}, \Upsilon_j^{(2)})\}_{j=1, \dots, n},$$

где је:

$$\begin{aligned} U_j^{(1)} &= \min\{X_j, C_j\} + \epsilon_j = \min\{X_j + \epsilon_j, C_j + \epsilon_j\} = \min\{\chi_j, \zeta_j\}, \\ \Upsilon_j^{(1)} &= I\{X_j \leq C_j\}, \\ U_j^{(1)} &= \min\{X_j + \epsilon_j, C_j\} = \min\{\chi_j, C_j\}, \quad \Upsilon_j^{(1)} = I\{\chi_j \leq C_j\}. \end{aligned}$$

Први модел се уклапа у претпоставке класичног адитивног модела грешке мерења.

Аутори у раду [3] дају оцену стопе хазарда $\lambda_X(\cdot)$, на основу које се, из везе

$$S_X(t) = e^{-\int_0^t \lambda_X(u) du},$$

може извести оцена функције преживљавања која се може применити у конструкцији нових статистичких тестова.

Приступ из [3] може послужити као основа за прилагођавање тест статистике из [5] подацима код којих, поред десног цензурисања, постоје и грешке мерења. Класична Каплан-Мајерова оцена би могла бити замењена одговарајућом деконволуционом оценом. Тако би могле да се модификују тежине W_i , $i \in \{1, 2, \dots, n\}$, дефинисане у претходном одељку. У зависности од модела грешка мерења може погодити време преживљавања, цензуришућу променљиву или вектор мерених карактеристика. Посебно је интересантан случај када грешка погађа све три променљиве у моделу.

Такође, методологија се може проширити и на општији модел, код кога време преживљавања и цензуришућа променљива не би биле погођене истом грешком:

$$\begin{aligned} \chi &= X + \varepsilon_X, \\ \zeta &= C + \varepsilon_C. \end{aligned}$$

Глава 5

Закључак

У овом раду је размотрен проблем тестирања независности случајних променљивих у присуству грешака мерења. Грешке мерења настају као последица различитих фактора, а у овом раду је посебна пажња посвећена случајним грешкама које задовољавају претпоставке класичног адитивног модела грешке мерења. Посматране променљиве представљају контаминирани верзије латентних променљивих од интереса. Присуство грешака у мерењу може изменити структуру података, због чега је потребно развити тест који на основу посматраних променљивих детектује зависност између латентних променљивих. Поступци тестирања обрађени у овом раду се заснивају на $dCov$ мери зависности.

У првом делу рада су дефинисани основни појмови неопходни за анализу овог проблема. То су класични адитивни модел грешке мерења, циркуларни подаци, однос шума и сигнала и $dCov$ мера. Приказане су и најважније особине $dCov$ мере, због којих она представља добру основу за конструкцију тестова чији је циљ детектовање било ког облика зависности за различите типове података.

У наставку су приказане и упоређене различите тест статистике засноване на $dCov$ мери. Поред стандардног поступка тестирања, већ познатог у литератури, предложен је и нови поступак заснован на средњим вредностима поновљених мерења. Поновљена мерења пружају додатну информацију о латентним променљивама и делимично ублажавају дејство грешке мерења.

Емпиријска студија обухватила је више модела за класичне и усмерене податке. На тај начин испитано је понашање тестова у условима независности,

различитих степена зависности, различитих нивоа грешке мерења и различитих односа шума и сигнала. Посебно су разматрани модели са линеарном зависношћу, сложенијим облицима зависности и различитим расподелама латентних променљивих. Размотрени су модели класичних и усмерених типова.

Добијени резултати указују на то да је интензитет грешке мерења, односно однос шума и сигнала, један од кључних фактора који одређују перформансе тестова независности. Са повећањем односа шума и сигнала уочава се смањење моћи тестова, док се зависност успешније детектује у ситуацијама у којима је сама зависност између латентних променљивих израженија. Овај резултат је у складу са интуицијом: што је грешка мерења већа у односу на сигнал, то посматрани подаци садрже мање информације о стварној структури зависности. У студији је потврђена очекивана претпоставка, да се најбоље перформансе постижу у случају када подаци нису оптерећени грешкама мерења. Уколико се подаци контаминирају грешкама тест статистике T и $\hat{\rho}^2$, за $m_u = m_v = 2$, показују сличне перформансе, а моћи тестова се у случају јаке зависности приближавају моћи теста који се примењује на неконтаминирани податке. Када је однос шума и сигнала већи од један моћи тестова који користе контаминирани узорак нису толико високе. Показано је да је коришћење тест статистике T , која је представљена у оквиру овог рада, оправдано. Њене перформансе су на нивоу перформанси тест статистике $\hat{\rho}^2$, која је већ обрађена у литератури [6], а истовремено је рачунски мање захтевна. Такође, уколико је могуће поновити мерење више од два пута статистика T се може боље показати, јер је дефинисана за произвољне m_u и m_v , па, за разлику од $\hat{\rho}^2$ не захтева додатна проширења.

За усмерене податке добијени су слични закључци као за класичне. Резултати показују да предложени приступ није ограничен само на класичне податке, већ да се основне идеје могу проширити и на различите структуре података. То је посебно важно јер се усмерени подаци јављају у разним областима.

Сви тестови показују сличне перформансе за различите језгра.

На крају су представљена и нека отворена питања и правци за даље истраживање ове области. Природно је добијене резултате проширити и на цензурисане податке, како се у пракси често јављају подаци које погађа и цензурисање и грешке у мерењу.

Библиографија

- [1] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, 2 edition, 2006.
- [2] Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. Приступљено 21. августа 2026.
- [3] Fabienne Comte, Adeline Samson, and Julien J. Stirnemann. Hazard estimation with censoring and measurement error: application to length of pregnancy. *TEST*, 27:338–359, 2018.
- [4] Marija Cuparić, Bruno Ebner, and Bojana Milošević. Flexible independence testing for hyperspherical data: a kernel approach for vectors of different dimensions. *Journal of Statistical Computation and Simulation*, 96(3):513–531, 2026.
- [5] Dominic Edelmann, Thomas Welchowski, and Axel Benner. A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics*, 78:867–879, 2022.
- [6] Jinlin Fan, Yaowu Zhang, and Liping Zhu. Independence tests in the presence of measurement errors: An invariance law. *Journal of Multivariate Analysis*, 188:104818, 2022.
- [7] Riccardo Gatto and S. Rao Jammalamadaka. Directional statistics: Introduction. In *Wiley StatsRef: Statistics Reference Online*, pages 1–8. John Wiley & Sons, Ltd., 2015.
- [8] Raffaella Giacomini, Dimitris N. Politis, and Halbert White. A Warp-Speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators. *Econometric Theory*, 29(3):567–589, 2013.

- [9] Gabriel K. Innes, Fiona Bhondoeckhan, Bryan Lau, Alden L. Gross, Derek K. Ng, and Alison G. Abraham. The Measurement Error Elephant in the Room: Challenges and Solutions to Measurement Error in Epidemiology. *Epidemiologic Reviews*, 43(1):94–105, 2021.
- [10] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.
- [11] Grace Y. Yi, Aurore Delaigle, and Paul Gustafson, editors. *Handbook of Measurement Error Models*. Chapman & Hall/CRC, Boca Raton, 2021.

Биографија аутора

Ана Петровић је рођена 12.01.2002. године у Београду. Математички факултет Универзитета у Београду, смер Статистика, актуарска и финансијска математика, уписала је 2020/2021. године. Дипломирала је 2025. године, са просечном оценом 9,13. Исте године уписала је мастер студије на Математичком факултету Универзитета у Београду, смер Статистика, актуарска и финансијска математика.

У јесењем семестру школске 2025/2026. године, била је ангажована као демонстратор на Економском факултету универзитета у Београду, на предмету Математика и на Математичком факултету, на предметима Увод у финансијску математику и Теорија вероватноћа, док је у пролећном семестру исте школске године била ангажована на Економском факултету, на предмету Теоријска статистика.